Διαχείριση Δεδομένων και Επιχειρηματική Ευφυΐα

Θεωρία και Εφαρμογές για Στελέχη Επιχειρήσεων

ΓΕΩΡΓΙΟΣ ΣΤΑΛΙΔΗΣ

ΔΗΜΗΤΡΙΟΣ ΚΑΡΔΑΡΑΣ

ΓΕΩΡΓΙΟΣ ΣΤΑΛΙΔΗΣ
Επίκουρος Καθηγητής ΤΕΙ Θεσσαλονίκης

ΔΗΜΗΤΡΙΟΣ ΚΑΡΔΑΡΑΣ
Επίκουρος Καθηγητής Οικονομικού Πανεπιστημίου Αθηνών

Διαχείριση Δεδομένων και Επιχειρηματική Ευφυΐα

Θεωρία και Εφαρμογές για Στελέχη Επιχειρήσεων

Διαχείριση Δεδομένων και Επιχειρηματική Ευφυΐα

Συγγραφή

Γεώργιος Σταλίδης

Δημήτριος Καρδαράς

Κριτικός αναγνώστης

Κωνσταντίνος Διαμαντάρας

Συντελεστές έκδοσης

Γραφιστική Επιμέλεια: Σωτήριος Τριανταφύλλου

ISBN: 978-960-603-398-8

ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

Contents

Πίνακας συντομεύσεων-ακρωνύμια

ASCII	American Standard Code for Information Interchange
BI	Business Intelligence
CD	Compact Disc
CRM	Customer Relationship Management
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DBMS	DataBase Management System
ERD	Entity-relationship diagram
ERP	Enterprise Resource Planning
KBS	Knowledge-Based System
KDD	Knowledge discovery in Data Bases
KM	Knowledge model
KMS	Knowledge Management System
OLAP	Online AnaLytical Processing
OWL	Ontology Web Language
POS	Point of Sale
SWRL	Semantic Web Query Language
UML	Unified Modeling Language
ΒΔ	Βάση Δεδομένων
Η/Υ	Ηλεκτρονικός Υπολογιστής
ΚΜ	Κανονική μορφή
ΣΔΒΔ	Σύστημα Διαχείρισης Βάσεων Δεδομένων

Ευρετήριο Ελληνικών-Αγγλικών

Ακεραιότητα δεδομένων	Data integrity
Άμεση Αναλυτική Επεξεργασία	OnLine Analytical Processing
Ανακάλυψη γνώσης σε Βάσεις Δεδομένων	Knowledge discovery in Data Bases
Ανάλυση πρόβλεψης	Predictive analytics
Αναλυτική επεξεργασία δεδομένων	Data analytics
Αντιφατικότητα	Inconsistency
Αξιολόγηση	Evaluation
Αποθετήριο	Repository
Αρχείο	File
Ασαφής Λογική	Fuzzy Logic
Βάση Δεδομένων	Data Base
Βασισμένο σε γνώση	Knowledge-based
Βοηθητικά εργαλεία	Utilities
Γραφήματα	Charts
Δεδομένα	Data
Δείγμα	Sample
Δέντρο Απόφασης	Decision Tree
Διάγραμμα οντοτήτων-συσχετίσεων	Entity-relationship diagram
Διάγραμμα τάξεων	Class diagram
Διαδικασία	Process
Δοσοληψία	Transaction
Διάσπαρτα δεδομένα	Data sparsity
Δυαδικός	Binary
Εγγραφή	Record
Ερώτημα	Query
Εξόρυξη δεδομένων	Data mining
Ιδιότητα	Property, Attribute
Κατηγορηματική γνώση	Explicit knowledge
Κύβος	Cube
Λανθάνουσα μνήμη	Cache memory
Μεγάλα δεδομένα	Big data
Μετα-δεδομένα	Meta-data
Μετασχηματισμός δεδομένων	Data transformation
Μηχανική μάθηση	Machine learning
Μηχανική της γνώσης	Knowledge engineering
Μοντέλο γνώσης	Knowledge model
Μοντέλο δεδομένων	Data model

Μοντέλο κανόνων	Rule-based model
Μοντέλο πληροφορίας	Information model
Μοντελοποίηση	Modeling
Ομοιότητα	Similarity
Οντότητα	Entity
Παλινδρόμηση	Regression
Παράδειγμα	Example
Πεδίο	Field
Πίνακας ισοδυναμίας	Equivalence matrix
Πίνακας σύγχυσης	Confusion matrix
Πλεονασμός Δεδομένων	Data redundancy
Περίπτωση	Case
Προβολή Αποτελεσμάτων	Results Perspective
Προβολή Σχεδίασης	Design Perspective
Σημασιολογικά δίκτυα	Semantic networks
Σταθμισμένες έννοιες	Weighted concepts
Στατιστικά	Statistics
Στιγμιότυπο	Instance
Συμπερασματική μηχανή Ή Συλλογιστική μηχανή	Inference engine
Σύνδεσμοι	Connectors
Συνεργατικό φιλτράρισμα	Collaborative filtering
Συσταδοποίηση	Clustering
Σύστημα Διαχείρισης Βάσεων Δεδομένων	Data Base Management System
Σύστημα Συστάσεων	Recommender system
Σχέση	Relationship
Τάξη, Κλάση	Class
Τελεστής	Operator
Υπονοούμενη γνώση	Tacit knowledge
Χαρακτηριστικό	Attribute
Χαρακτηριστικό-στόχος	Target attribute

Εισαγωγή

Η υιοθέτηση σύγχρονων τεχνολογιών πληροφορικής σε μια επιχείρηση δεν είναι πια στις μέρες μας συγκριτικό πλεονέκτημα, αλλά απαραίτητο στοιχείο επιβίωσης. Με την ίδια λογική, η εξοικείωση με τις τεχνολογίες διαχείρισης πληροφορίας είναι απαραίτητο στοιχείο επιβίωσης για ένα στέλεχος επιχείρησης. Επιπλέον, όμως, της εξοικείωσης με τις συνήθεις τεχνολογίες πληροφορικής, η ικανότητα αξιοποίησης των νέων ευκαιριών που δίνουν οι τεχνολογίες αιχμής στο χώρο των επιχειρηματικών δεδομένων φαίνεται να αποτελεί ένα εξαιρετικά πολύτιμο εφόδιο σε χώρους όπως η διοίκηση και το μάρκετινγκ. Είναι πλέον αποδεκτό ότι οι αποφάσεις σε όλους σχεδόν τους χώρους πρέπει να λαμβάνονται με βάση στοιχεία και όχι τη διαίσθηση ή την τύχη. Ακόμα και οι επιλογές που είναι αδύνατο να βασιστούν σε ξεκάθαρα κριτήρια έχουν καλύτερες προοπτικές επιτυχίας όταν πραγματοποιούνται μετά από έγκυρη και ολοκληρωμένη πληροφόρηση.

Ο όρος Επιχειρηματική Ευφυΐα (Business Intelligence) αναφέρεται σε μεθόδους και διαδικασίες που έχουν ως σκοπό τη μετατροπή των δεδομένων σε πληροφορίες και στη συνέχεια σε γνώση και χρησιμοποιείται για την υποστήριξη της λήψης αποφάσεων σε μια επιχείρηση. Ο σκοπός της Επιχειρηματικής Ευφυΐας είναι η καλύτερη κατανόηση των δραστηριοτήτων μιας επιχείρησης και του περιβάλλοντός της μέσω της συλλογής και ανάλυσης δεδομένων, έτσι ώστε να υποστηριχθεί η διοίκηση της επιχείρησης στη λήψη αποφάσεων και το σχεδιασμό.

Το βιβλίο καλύπτει δύο βασικές, αλληλένδετες μεταξύ τους, γνωστικές περιοχές της πληροφορικής, την Ηλεκτρονική Διαχείριση Δεδομένων και την Επιχειρηματική Ευφυΐα (Business Intelligence). Η πρώτη ενότητα περιλαμβάνει τις βασικές αρχές βάσεων δεδομένων και ειδικότερα θέματα συλλογής και αξιοποίησης δεδομένων σε περιβάλλον επιχειρήσεων. Δίνονται παραδείγματα και οδηγίες εφαρμογής που αφορούν τα συνηθέστερα, απλούστερα αλλά και ουσιαστικά εργαλεία για διαχείριση δεδομένων με ευρέως διαδεδομένο λογισμικό γραφείου, πληροφορημένη λήψη αποφάσεων και σχεδιασμό βασισμένο σε στοιχεία. Η δεύτερη ενότητα εισάγει τους αναγνώστες στα θέματα Επιχειρηματικής Ευφυΐας και εισέρχεται βαθύτερα στο χώρο αυτό, παρουσιάζοντας μεθόδους και εφαρμογές με υψηλότερο βαθμό ευφυΐας, που βασίζονται σε ειδικό λογισμικό.

Οι τεχνικές διαχείρισης δεδομένων και απλές εφαρμογές εξαγωγής πληροφοριών από τα δεδομένα αυτά, όπως διοικητικές αναφορές και συγκεντρωτικά στοιχεία, παρουσιάζονται έτσι ώστε να μπορούν να υλοποιηθούν στο ευρέως διαδεδομένο λογισμικό Βάσεων Δεδομένων Microsoft Access. Ο αναγνώστης θα είναι σε θέση να γνωρίσει και να δημιουργήσει ο ίδιος εφαρμογές για την αποτελεσματικότερη διαχείριση των δεδομένων του, αλλά και την καλύτερη αξιοποίησή τους, παράγοντας χρήσιμη πληροφορία.

Στη συνέχεια, παρουσιάζονται μέθοδοι εξαγωγής γνώσης από δεδομένα, που εφαρμόζονται σε τυπικά προβλήματα επιχειρηματικής ευφυΐας, όπως πρόβλεψη συμπεριφοράς καταναλωτών, επιλογή αγοράς στόχου, τοποθέτηση προϊόντων, μέτρηση της αποτελεσματικότητας ενεργειών προώθησης, εντοπισμός ευκαιριών διασταυρωμένων πωλήσεων, κ.ά. Με χρήση του ειδικού λογισμικού ανάλυσης δεδομένων RapidMiner, ο αναγνώστης θα μπορεί να επιλύει προβλήματα που μπορούν να αναβαθμίσουν δραστικά την αποτελεσματικότητά του στη λήψη αποφάσεων και το σχεδιασμό στο μάρκετινγκ και τη διοίκηση.

Στα τελευταία κεφάλαια, παρουσιάζονται ειδικότερα θέματα, όπως αναπαράσταση γνώσης και Βάσεις Γνώσεις, καθώς και ειδικά συστήματα συστάσεων και στήριξης αποφάσεων. Εξετάζονται μέθοδοι υποστήριξης αποφάσεων βασισμένες σε γνώση και παρουσιάζονται τεχνολογίες δημιουργίας μοντέλων επεξήγησης ή πρόβλεψης για την επίλυση επιχειρηματικών προβλημάτων, όπως εξεύρεση καταναλωτικών προτύπων, εκτίμηση πιστότητας πελατών και προσδιορισμός των παραγόντων που την επηρεάζουν.

Σκοπός του βιβλίου συνολικά είναι η εξοικείωση, μέσω εύληπτης θεωρίας και πρακτικών εφαρμογών, με τα σύγχρονα εργαλεία συλλογής και ανάλυσης δεδομένων, της εξαγωγής γνώσης και της αποτελεσματικής χρήσης της γνώσης αυτής στην επίλυση προβλημάτων. Το σύγγραμμα είναι προσαρμοσμένο στις ανάγκες ενός φοιτητή ή στελέχους διοίκησης επιχειρήσεων και διαφέρει τόσο από συγγράμματα που προορίζονται για πληροφορικούς ή μηχανικούς, όσο και από πρακτικούς οδηγούς χρήσης λογισμικού. Παρέχονται στοιχεία θεωρίας που επαρκούν στην κατανόηση των βασικών αρχών και τρόπου σκέψης που απαιτείται για την επίλυση τυπικών προβλημάτων, αλλά περιλαμβάνονται και εφόδια πρακτικής εφαρμογής. Επίσης, το σύγγραμμα δεν περιορίζεται στην εξωτερική παρουσίαση συστημάτων ή στη θεωρητική συζήτηση περιπτώσεων αλλά καθοδηγεί τον αναγνώστη στο να επιλύει το ίδιος πραγματικά προβλήματα, ανακαλύπτοντας την πληροφορία και γνώση που κρύβεται στα δεδομένα που έχει στη διάθεσή του και λαμβάνοντας τεκμηριωμένες «ευφυείς» αποφάσεις.

Κεφάλαιο1. Εισαγωγή στη βασισμένη σε δεδομένα επιχειρηματική ευφυΐα

Σύνοψη
Εισάγονται η έννοια και οι στόχοι της επιχειρηματικής ευφυΐας και παρουσιάζεται ο ρόλος της ηλεκτρονικής διαχείρισης δεδομένων ως θεμέλιο των ευφυών πληροφοριακών συστημάτων για επιχειρήσεις. Γίνεται σύντομη αναφορά στην ιστορία της διαχείρισης δεδομένων με Η/Υ και περιγράφονται οι δυνατότητες που προσφέρουν τα σύγχρονα εργαλεία, καθώς και οι τάσεις της διαθέσιμης σήμερα τεχνολογίας.

Προαπαιτούμενη γνώση
Εισαγωγή στην Πληροφορική

1.1 Η βασισμένη στην πληροφορία διοίκηση, η επιχειρηματική ευφυΐα και ο ρόλος των δεδομένων

Οι δραστηριότητες μιας σύγχρονης επιχείρησης, από τις καθημερινές της λειτουργίες, τη λήψη διοικητικών αποφάσεων σε όλους τους τομείς, όπως παραγωγή, μάρκετινγκ και χρηματο-οικονομικά, ως και τη διαμόρφωση της στρατηγικής της, συμπεριλαμβάνουν τη διακίνηση δεδομένων και πληροφοριών. Η αποτελεσματική λειτουργία της επιχείρησης πολύ συχνά σχετίζεται με σύνθετες διαδικασίες που απαιτούν τη συνεργασία πολλών τμημάτων και την επίλυση λογικών προβλημάτων. Πολλές από τις λειτουργίες εμπεριέχουν τη διαχείριση μεγάλων όγκων δεδομένων, όπως για παράδειγμα η καταγραφή των στοιχείων των πελατών, των παραγγελιών, των προμηθειών και των χρηματοοικονομικών συναλλαγών. Άλλες λειτουργίες απαιτούν λήψη αποφάσεων, κρίση και δημιουργικότητα, όπως η επιλογή των πιο κερδοφόρων πελατών στους οποίους θα θέλαμε να εστιάσουμε τις προωθητικές ενέργειες της επιχείρησης ή η βέλτιστη τοποθέτηση ενός νέου προϊόντος. Επιπρόσθετα, μια επιχείρηση οφείλει να επικοινωνεί με το περιβάλλον της, τόσο για τις λειτουργικές ανάγκες των δοσοληψιών της όσο και για την άντληση πολύτιμων πληροφοριών, κάτι που συνεπάγεται επιπλέον ανάγκες διαχείρισης μεγάλων όγκων πληροφορίας.

Η ανάγκη υποστήριξης των λειτουργιών αυτών από κατάλληλα συστήματα πληροφορικής είναι πλέον αδιαμφισβήτητη και μπορούμε να πούμε ότι η υιοθέτηση σύγχρονων τεχνολογιών πληροφορικής σε μια επιχείρηση δεν είναι πια συγκριτικό πλεονέκτημα αλλά απαραίτητο στοιχείο επιβίωσης. Με την ίδια λογική, η εξοικείωση με τις τεχνολογίες διαχείρισης πληροφορίας είναι απαραίτητο στοιχείο επιβίωσης για ένα στέλεχος επιχείρησης. Επιπλέον, όμως, της εξοικείωσης με τις συνήθεις τεχνολογίες πληροφορικής, η ικανότητα αξιοποίησης των νέων ευκαιριών που δίνουν οι τεχνολογίες αιχμής στο χώρο των επιχειρηματικών δεδομένων φαίνεται να αποτελεί ένα εξαιρετικά πολύτιμο εφόδιο σε χώρους όπως η διοίκηση και το μάρκετινγκ.

Είναι πλέον αποδεκτό ότι οι αποφάσεις σε όλους σχεδόν τους χώρους πρέπει να λαμβάνονται με βάση στοιχεία (και όχι τη διαίσθηση ή την τύχη). Ακόμα και οι επιλογές που είναι αδύνατον να βασιστούν σε ξεκάθαρα κριτήρια έχουν καλύτερες προοπτικές επιτυχίας, όταν πραγματοποιούνται μετά από έγκυρη και ολοκληρωμένη πληροφόρηση. Μια βασική αρχή που έχει περάσει από τον χώρο των μηχανικών στη διοίκηση είναι ότι, για να ελέγξεις κάτι αποτελεσματικά, πρέπει αρχικά να μπορείς να το μετρήσεις, αλλά και να εκτιμήσεις τους μηχανισμούς που το επηρεάζουν. Έτσι, ένα στέλεχος μάρκετινγκ, για να εκπονήσει ένα σχέδιο που αποσκοπεί στη βελτίωση της κερδοφορίας, πρέπει πρώτα να μπορεί να μετρήσει την κερδοφορία αυτή και τις πηγές της, δηλαδή να γνωρίζει πού βρίσκεται και πού θέλει να πάει. Επίσης, πρέπει να μελετήσει τις παραμέτρους που την επηρεάζουν, βασιζόμενος σε στοιχεία για το προϊόν, την αγορά στην οποία απευθύνεται και πιθανότατα σε πολλά άλλα. Δεν είναι, λοιπόν, παράξενο το ότι κύριο στοιχείο της δουλειάς ενός στελέχους μάρκετινγκ είναι η έρευνα και η ανάλυση στοιχείων, δηλαδή η συλλογή και αξιοποίηση πληροφορίας.

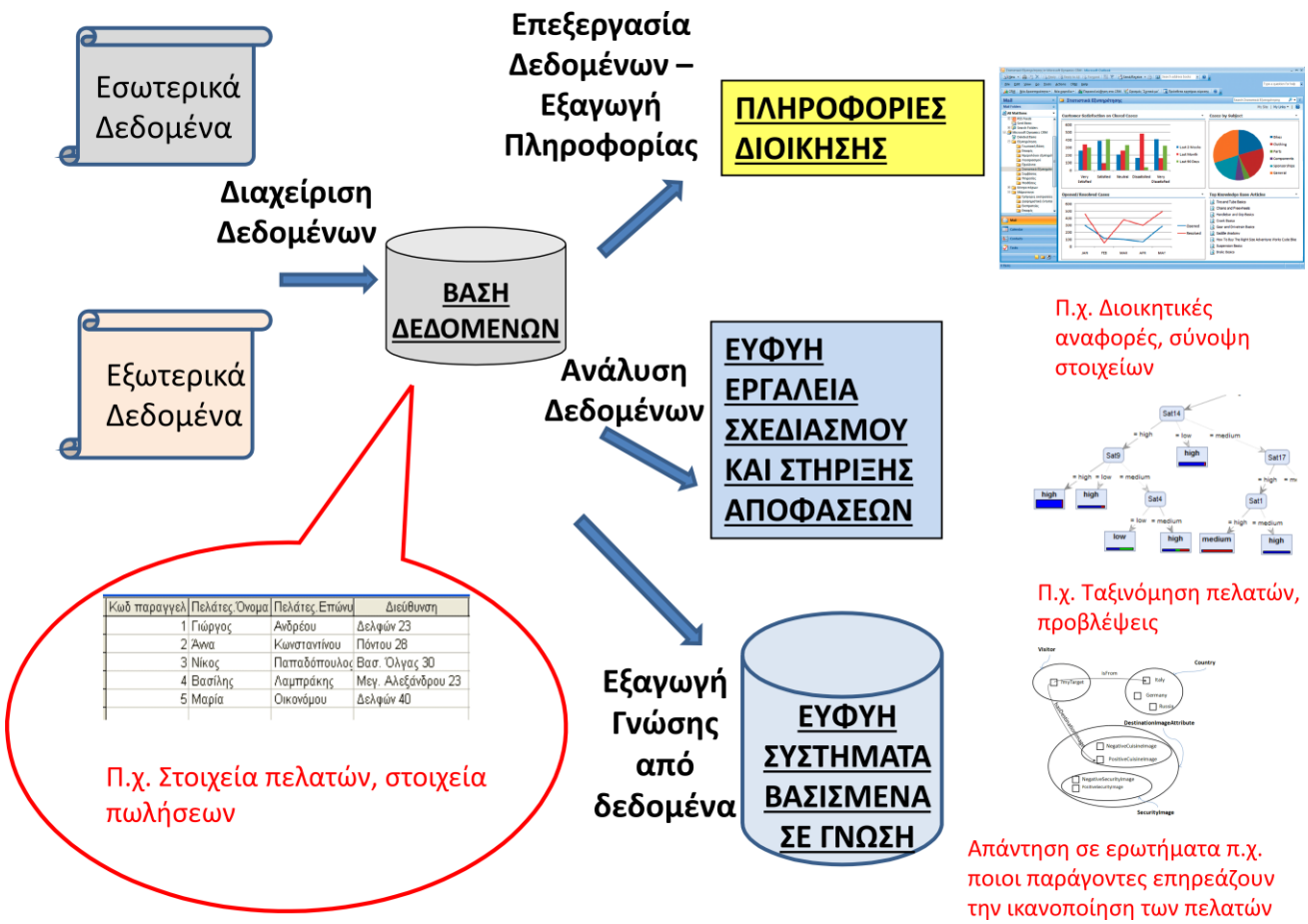
Η πληροφορία ή οι μετρήσεις τα οποία καλείται να αξιοποιήσει ένα στέλεχος επιχείρησης μπορεί να είναι διαθέσιμα σε διάφορες μορφές ή να πρέπει να παραχθούν, ενώ μπορεί να αφορούν το εσωτερικό της επιχείρησης ή το εξωτερικό της περιβάλλον. Τα εργαλεία και η μεθοδολογία εύρεσης και αξιοποίησης της πληροφορίας μπορεί να διαφέρουν και συνήθως εντάσσονται σε κάποια από τις παρακάτω περιπτώσεις:

- Η πληροφορία μπορεί να είναι πρωτογενής, που μπορεί να συλλεχθεί για συγκεκριμένο σκοπό με χρήση κατάλληλων εργαλείων, όπως ερωτηματολόγια και στατιστικές έρευνες. Η πληροφορία αυτή αναλύεται με κατάλληλα εργαλεία, συνήθως στατιστικά, ώστε να μας οδηγήσει σε συμπεράσματα χρήσιμα στο σχεδιασμό και τη λήψη αποφάσεων.
- Η πληροφορία μπορεί να καταγράφεται κατά τις καθημερινές διεκπεραιωτικές λειτουργίες μιας επιχείρησης, όπως π.χ. η λήψη μιας παραγγελίας ή τα στοιχεία σχετικά με την εκτέλεση μιας πληρωμής. Οι πληροφορίες αυτές χρησιμοποιούνται στην αυτοματοποίηση απλών λειτουργιών της επιχείρησης και, μετά από τυποποιημένη επεξεργασία, στη λογιστική παρακολούθηση. Στην αρχική τους μορφή, έχουν μικρή χρησιμότητα στη διοίκηση και το μάρκετινγκ.
- Πληροφορία μπορεί να εξαχθεί από αυτήν της προηγούμενης κατηγορίας μετά από επιλογή και επεξεργασία π.χ. υπολογισμός των κερδών του έτους ανά προϊόν ή εύρεση των μεγαλύτερων σε τζίρο πελατών της επιχείρησης. Πληροφορία αυτής της μορφής παράγεται με υπολογιστικές μεθόδους, απαντάει σε προκαθορισμένα ερωτήματα και χρησιμοποιείται ευρύτατα στη διοίκηση. Η δημιουργία και ο χειρισμός της πληροφορίας αυτού του είδους είναι δυνατά με χρήση ευρέως διαθέσιμου λογισμικού.
- Πληροφορία (ή γνώση, όπως θα αναφερθεί στο επόμενο κεφάλαιο) που δεν είναι ορατή, μπορεί να αναδυθεί μέσα από πληροφορία των προηγούμενων κατηγοριών με ειδικές μεθόδους ανάλυσης που διαθέτουν στοιχεία «ευφυΐας». Π.χ. από τη μελέτη των πωλήσεων ενός σούπερ μάρκετ είναι δυνατή η εύρεση προτύπων στη συμπεριφορά των καταναλωτών, όπως το δημογραφικό προφίλ των πελατών που ανταποκρίνονται περισσότερο σε κάποιο τύπο προσφορών.

Όλα τα είδη πληροφορίας που αναφέρθηκαν στις προηγούμενες παραγράφους βασίζονται στα δεδομένα (data) και ειδικότερα στα ηλεκτρονικά δεδομένα, δηλαδή αυτά που μπορούμε να χειριστούμε σε συστήματα βασισμένα σε Ηλεκτρονικούς Υπολογιστές (Η/Υ). Ο ορισμός των δεδομένων και η σχέση τους με αυτό που ονομάζουμε πληροφορία θα εξεταστεί ειδικότερα στο επόμενο κεφάλαιο, μπορούμε όμως να πούμε με απλά λόγια ότι οτιδήποτε μετριέται και καταγράφεται στον πραγματικό κόσμο μεταφέρεται στον Η/Υ με τη μορφή δεδομένων, που μπορούμε να τα αντιληφθούμε σαν αριθμητικά στοιχεία, λέξεις/κείμενα, εικόνες, βίντεο και οτιδήποτε άλλο μπορεί να καταγραφεί. Η πληροφορία που μπορεί να αξιοποιηθεί για την επίλυση ενός προβλήματος, είτε πρόκειται για μια απλή αναφορά είτε για ένα εύρημα που ανακαλύφθηκε με ευφυείς τεχνικές, βασίζεται στα δεδομένα. Επομένως, **τα δεδομένα είναι τα θεμέλια της επιχειρηματικής ευφυΐας**.

Ο όρος **Επιχειρηματική Ευφυΐα (Business Intelligence)** αναφέρεται σε μεθόδους και διαδικασίες που έχουν ως σκοπό τη μετατροπή των δεδομένων σε πληροφορίες και στη συνέχεια σε γνώση και χρησιμοποιείται για την υποστήριξη της λήψης αποφάσεων σε έναν οργανισμό. Ο σκοπός της Επιχειρηματικής Ευφυΐας είναι η καλύτερη κατανόηση των δραστηριοτήτων μιας επιχείρησης και του περιβάλλοντός της μέσω της συλλογής και ανάλυσης δεδομένων, έτσι ώστε να υποστηριχθεί η διοίκηση της επιχείρησης στη λήψη αποφάσεων και τον σχεδιασμό.

Ο όρος Επιχειρηματική Ευφυΐα χρησιμοποιήθηκε από τον Howard Dresner το 1989 για να περιγράψει τις μεθόδους που μπορούν να βελτιώσουν τη λήψη αποφάσεων με την υποστήριξη συστημάτων που βασίζονται σε στοιχεία (Power, 2007). Στη βιβλιογραφία συναντούμε ανάλογους ορισμούς για την Επιχειρηματική Ευφυΐα, όπως των Elbashir et al (2008), σύμφωνα με τους οποίους τα συστήματα επιχειρηματικής ευφυΐας είναι εξειδικευμένα εργαλεία που χρησιμοποιούνται για την ανάλυση δεδομένων, την υποβολή ερωτημάτων και λήψη εκθέσεων, που υποστηρίζουν τη λήψη αποφάσεων ενός οργανισμού που ενδεχομένως βελτιώνουν την απόδοση μιας σειράς επιχειρηματικών διαδικασιών.



Σχήμα 1.1. *Τα δεδομένα ως θεμέλια της επιχειρηματικής ευφυΐας.*

Η σημασία της Επιχειρηματικής Ευφυΐας στην επιτυχία μιας επιχείρησης γίνεται όλο και μεγαλύτερη όσο το μέγεθός της, οι διαδικασίες της και ο αντίστοιχος όγκος της πληροφορίας που διακινείται αυξάνονται. Ταυτόχρονα, οι ευκαιρίες που δίνονται για τη βελτίωση μιας επιχείρησης και η κρισιμότητα της συμβολής της Επιχειρηματικής Ευφυΐας προς αυτήν την κατεύθυνση, αυξάνονται όσο εντείνεται ο ανταγωνισμός, το περιβάλλον και η αγορά διευρύνονται και γίνονται πιο περίπλοκα, οι εξελίξεις επιταχύνονται και οι απαιτήσεις βαθιάς γνώσης της πραγματικότητας αυξάνονται. Η διοίκηση μιας επιχείρησης έχει ανάγκη από πληροφορίες προκειμένου να είναι δυνατή η παρακολούθηση των δραστηριοτήτων της, αλλά και από ισχυρά και αποτελεσματικά «έξυπνα εργαλεία» για την καλύτερη αξιοποίηση της πληροφορίας αυτής. Η διαχείριση και κατανόηση των πληροφοριών προς όφελος της διοίκησης και του σχεδιασμού μπορεί να είναι ιδιαίτερα δύσκολη, καθώς σήμερα τα πληροφοριακά συστήματα συλλέγουν και επεξεργάζονται τεράστιες ποσότητες δεδομένων σε διάφορες μορφές (Laudon & Laudon, 2009).

Ως τυπικές εφαρμογές της Επιχειρηματικής Ευφυΐας, χωρίς να είναι οι μόνες, αναφέρονται οι εξής:

- η ανάλυση συμπεριφοράς καταναλωτών και ο προσδιορισμός αγοραστικών προτύπων και τάσεων στην αγορά
- η μέτρηση, παρακολούθηση και πρόβλεψη των οικονομικών επιδόσεων της επιχείρησης
- η παρακολούθηση της απόδοσης των εκστρατειών μάρκετινγκ
- η εύρεση ευκαιριών cross selling
- η ομαδοποίηση πελατών και εύρεση των χαρακτηριστικών και προτιμήσεών τους
- η διαχείριση πελατειακών σχέσεων
- η βελτίωση της αποτελεσματικότητας της παραγωγής και της αλυσίδας εφοδιασμού

Για να γίνει καλύτερα αντιληπτό ποιο είναι το εύρος των δυνατοτήτων της Επιχειρηματικής Ευφυΐας και τι μπορεί να προσφέρει πρακτικά σε μια μικρή ή μεγάλη επιχείρηση, αναφέρουμε ένα παράδειγμα από το

χώρο των λιανικών πωλήσεων. Ας θεωρήσουμε μια αλυσίδα συνοικιακών καταστημάτων τροφίμων και ειδών νοικοκυριού (μίνι μάρκετ). Τι μπορεί να προσφέρει η Επιχειρηματική Ευφυΐα σε μια τέτοια επιχείρηση; Τα δεδομένα συναλλαγών που διακινούνται στα πληροφοριακά συστήματα που διαθέτει η επιχείρηση για την υποστήριξη των λειτουργικών και διαχειριστικών της αναγκών είναι ένα χαρακτηριστικό παράδειγμα δεδομένων μεγάλου όγκου, που μπορούν να αξιοποιηθούν πέρα από το λειτουργικό τους σκοπό. Με τεχνικές επιχειρηματικής ευφυΐας μπορεί να αποκτηθεί πολύτιμη γνώση, όπως για παράδειγμα:

- **Γνώση σχετική με τα προϊόντα:** Ποιος είναι ο όγκος πωλήσεων κάθε προϊόντος ανά χρονική περίοδο; Ποιες ώρες της ημέρας, ποιες ημέρες της εβδομάδας, ποιους μήνες και σε ποιες ειδικές περιστάσεις έχουμε έντονες μεταβολές στις πωλήσεις κάθε προϊόντος; Ποια προϊόντα συνεισφέρουν περισσότερο στην κερδοφορία της επιχείρησης, λαμβάνοντας υπόψη χρηματο-οικονομικούς παράγοντες και το κόστος διαχείρισης; Ποια προϊόντα έχουν μεγάλο ποσοστό επιστροφών και παραπόνων; Με τα στοιχεία αυτά, ο υπεύθυνος μάρκετινγκ μπορεί να προσδιορίσει καλύτερα τους στόχους του όσον αφορά την επιλογή των προϊόντων και ο υπεύθυνος προμηθειών/διακίνησης μπορεί να κάνει αποτελεσματικότερο προγραμματισμό.
- **Γνώση σχετική με την τιμολογιακή πολιτική και τις προωθητικές ενέργειες:** Πόσο αποτελεσματικές είναι οι προσφορές κάθε προϊόντος; Πόσο αυξάνονται οι πωλήσεις του όταν η έκπτωση είναι π.χ. 20% και πόσο όταν είναι 50%; Πόσο συμβάλλει κάθε προσφορά στην αύξηση ή στην απώλεια κερδών του καταστήματος; Ποια η πιθανότητα, κάποιος που αγόρασε π.χ. μακαρόνια σε προσφορά να αγοράσει και τη νέα σάλτσα ζυμαρικών που επιθυμούμε να προωθήσουμε; Πόσα και ποια από τα εκπτωτικά κουπόνια που μοιράστηκαν χρησιμοποιήθηκαν τελικά από τους πελάτες; Μετρώντας συστηματικά τα αποτελέσματα των προωθητικών ενεργειών και συσχετίζοντάς τα με χρηματο-οικονομικές παραμέτρους, ο υπεύθυνος μάρκετινγκ μπορεί να σχεδιάσει με ακρίβεια τις ενέργειες της επιχείρησης.
- **Γνώση σχετική με το προφίλ των πελατών:** Πόσοι πελάτες κάνουν μικρές αγορές και πόσοι μεγάλες; Πόσοι από τους πελάτες έρχονται τακτικά και πόσοι ευκαιριακά; Τι προϊόντα αγοράζουν συχνότερα και ποιες κατηγορίες προτιμούν; Πόσο επηρεάζονται από τις μεταβολές στις τιμές; Με μεθόδους ανάλυσης δεδομένων είναι δυνατή η εύρεση, όχι μόνο του επικρατέστερου προφίλ των πελατών, αλλά και η ανακάλυψη τάσεων και παραγόντων, η κατηγοριοποίηση των πελατών και ο συσχετισμός κάθε κατηγορίας με ιδιαίτερα χαρακτηριστικά. Π.χ. η ανάλυση μπορεί να οδηγήσει σε τύπους, όπως ο πελάτης του Σαββάτου, που αγοράζει κυρίως κατεψυγμένα, είδη νοικοκυριού/καθαρισμού και σνακ/σοκολατοειδή, η αξία των αγορών του είναι μεταξύ 30 και 50€ και προτιμάει τα επώνυμα προϊόντα, ο πελάτης της προσφοράς που πραγματοποιεί μικρές εστιασμένες αγορές, κλπ.

Τα παραπάνω παραδείγματα είναι ενδεικτικά και εστιάζουν στα δεδομένα των πωλήσεων. Ο αναγνώστης μπορεί να φανταστεί ότι η Επιχειρηματική Ευφυΐα μπορεί να προσφέρει ακόμα περισσότερες δυνατότητες όταν εφαρμόζεται σε όλους τους τομείς της επιχείρησης, όπως ανθρώπινους πόρους, επικοινωνία, ηλεκτρονικές πωλήσεις, προμήθειες, όπως επίσης όταν ενσωματώνει πληροφορίες από το εξωτερικό περιβάλλον (π.χ. πληθωρισμός, ανταγωνισμός, στατιστικά στοιχεία κλπ) και διαθέσιμη συσσωρευμένη γνώση (π.χ. μοντέλα πρόβλεψης, διαδικασίες/μέθοδοι διοίκησης, κλπ).

1.2 Από τα δυαδικά δεδομένα στα ευφυή συστήματα

Τα συστήματα πληροφορικής έχουν γνωρίσει τεράστια διάδοση σε όλους τους τομείς των επιχειρηματικών και επιστημονικών εφαρμογών και σε μεγάλο εύρος χρηστών. Επίσης, έχει επιτευχθεί αξιόλογη αποτελεσματικότητα σε σύνθετες λειτουργίες, όπως η στήριξη επιχειρηματικών αποφάσεων, η ανάλυση οικονομικών φαινομένων και η αναζήτηση χρήσιμης πληροφορίας σε μεγάλους όγκους ακατέργαστων δεδομένων. Στη βάση όλων των συστημάτων βρίσκεται ο Η/Υ που δεν είναι τίποτα άλλο από ένα σύνολο ηλεκτρονικών κυκλωμάτων που χειρίζονται δεδομένα στη μορφή ακολουθιών από δυαδικούς αριθμούς, με βάση προγράμματα, δηλαδή σειρές από μεγάλο αριθμό απλών εντολών.

Οι δυνατότητες που μας προσφέρουν τα πληροφοριακά συστήματα για την αποτελεσματική εκτέλεση λειτουργιών όπως οι παραπάνω είναι τεράστιες, σημαντικοί όμως είναι και οι περιορισμοί που προκύπτουν

από τη δομή και τη φιλοσοφία της ίδιας της τεχνολογίας της πληροφορικής. Για να είναι κατανοητή η φύση των ηλεκτρονικών δεδομένων, είναι χρήσιμο να αναφερθούμε στον τρόπο με τον οποίο αναπαριστώνται σε ένα πληροφοριακό σύστημα. Οι δυνατότητες των Ηλεκτρονικών Υπολογιστών (Η/Υ) τους καθιστούν ένα πανίσχυρο εργαλείο, που για να είναι όμως τελικά χρήσιμο, πρέπει να πάρει τη μορφή ενός πλήρους πληροφοριακού συστήματος. Ένα πληροφοριακό σύστημα είναι ένα σύστημα που περιλαμβάνει όλα τα επιμέρους τμήματα που απαιτούνται ώστε να εκτελεί αποτελεσματικά και αξιόπιστα μια συγκεκριμένη λειτουργία. Ένα σύστημα πληροφορικής δεν είναι απλά ένας Η/Υ αλλά είναι σχεδιασμένο έτσι ώστε να δέχεται μια είσοδο (π.χ. δεδομένα) και να παράγει ένα αποτέλεσμα, καθώς επίσης και να προσφέρει δυνατότητες χειρισμού στους χρήστες στους οποίους απευθύνεται.

Αντίστοιχα, δύο είναι και τα «ισχυρά» χαρακτηριστικά ενός συμβατικού συστήματος πληροφορικής βασισμένου σε Η/Υ: (α) η ικανότητά του να εκτελεί αριθμητικές πράξεις με μεγάλη ταχύτητα και (β) η ικανότητά του να αποθηκεύει και να διαχειρίζεται μεγάλο όγκο δεδομένων. Σε αυτά τα δύο χαρακτηριστικά βασίστηκαν οι κυριότερες εφαρμογές από την αρχή της ιστορίας των Η/Υ και γύρω από αυτά αναπτύχθηκαν οι σημαντικότερες τεχνολογίες. Από τη μια πλευρά, με τη βοήθεια Η/Υ έγινε δυνατή η επίλυση πολλών μαθηματικών προβλημάτων εφαρμόζοντας τεχνικές που βασίζονται στην υπολογιστική ισχύ και από την άλλη αναπτύχθηκαν εφαρμογές χειρισμού μεγάλου όγκου δεδομένων. Στις παραπάνω δυνατότητες των Η/Υ έχει προστεθεί τα τελευταία χρόνια η δυνατότητα επικοινωνίας μεταξύ συστημάτων μέσω της ραγδαίας ανάπτυξης των δικτύων υπολογιστών και του Διαδικτύου, που επέκτεινε τις εφαρμογές της πληροφορικής και στο χώρο των επικοινωνιών. Σε αντιδιαστολή όμως με αυτά τα ισχυρά στοιχεία, είναι χαρακτηριστική η αδυναμία των ηλεκτρονικών συστημάτων να επιλύσουν αδόμητα και ασαφή προβλήματα που απαιτούν ευφυΐα, δημιουργικότητα και που γενικά δεν ακολουθούν κάποια προκαθορισμένη διαδικασία επίλυσης. Για το λόγο αυτό και οι Η/Υ συχνά αποκαλούνται ως «κουτά» μηχανήματα που δεν αντιλαμβάνονται εύκολα την ανθρώπινη λογική.

Στην επιστήμη της πληροφορικής γίνεται μια διαρκής προσπάθεια να αναπτυχθούν και να τελειοποιηθούν μέθοδοι ώστε τα πληροφοριακά συστήματα να προσφέρουν ολοένα και πιο αποτελεσματικές εφαρμογές, όσο γίνεται πιο κοντά στον τρόπο σκέψης και τις πραγματικές ανάγκες του ανθρώπου-χρήστη, δηλαδή όσο γίνεται πιο έξυπνες. Εφαρμογές που απαιτούν διαχείριση μεγάλου όγκου δεδομένων και εκτέλεση πολλών πράξεων επεξεργασίας, αλλά που η διαδικασία επεξεργασίας είναι ξεκάθαρη και σταθερή, προσφέρονται ιδιαίτερα για υλοποίηση με πληροφοριακά συστήματα. Για παράδειγμα, η συγκέντρωση και αποθήκευση μεγάλου όγκου στοιχείων πωλήσεων και δοσοληψιών και ο υπολογισμός στατιστικών δεικτών με βάση αυτά τα στοιχεία, είναι κάτι που μπορεί σχετικά εύκολα να πραγματοποιηθεί αποτελεσματικά από ένα πληροφοριακό σύστημα. Επίσης, τα πληροφοριακά συστήματα προσφέρονται λόγω της δομής τους για την υποστήριξη καλά καθορισμένων λειτουργικών διαδικασιών, όπως για παράδειγμα η αυτοματοποίηση της διαδικασίας λήψης παραγγελιών, τιμολόγησης και αποστολής παραστατικών. Αντίθετα, εφαρμογές όπου απαιτείται ευφυΐα και διαχείριση ανθρώπινης γνώσης – πράγματα που δεν είναι χειροπιαστά δεδομένα – μπορούν να επιτευχθούν μόνο με τη συνεισφορά ειδικών επιστημονικών τομέων, όπως η εξόρυξη γνώσης από δεδομένα, η τεχνητή νοημοσύνη και η μοντελοποίηση γνώσης, μερικά από τα οποία θα γνωρίσουμε στα τελευταία κεφάλαια του βιβλίου αυτού. Έτσι, στα πιο υψηλά επίπεδα ανάπτυξης πληροφοριακών συστημάτων βλέπουμε «ευφυή» υπολογιστικά συστήματα, ικανά π.χ. να ελέγξουν όλα τα στάδια μιας αλυσίδας σε μια μονάδα παραγωγής ή να αναλύσουν ένα σύνολο από δεδομένα μιας αγοράς, ώστε να εξαχθεί συμπέρασμα σχετικά με τη βιωσιμότητα της εμπορευματοποίησης ενός νέου προϊόντος.

Όλες οι λειτουργίες ενός συστήματος ελέγχονται από λογισμικό, δηλαδή προγράμματα. Όταν επιθυμούμε να εκτελέσουμε μια αυτοματοποιημένη διαδικασία ή να επιλύσουμε ένα πρόβλημα που απαιτεί συγκεκριμένα και προκαθορισμένα βήματα, αυτό που απαιτείται είναι ουσιαστικά η μετάφραση μιας σύνθετης ανθρώπινης εντολής σε ένα σύνολο απλούστερων εντολών που να μπορεί να εκτελέσει ο Η/Υ. Κάτι τέτοιο υλοποιείται με τον προγραμματισμό της διαδικασίας σε κάποια κατάλληλη γλώσσα προγραμματισμού Η/Υ. Η μεγαλύτερη πρόκληση είναι όμως να μπορέσει ένα πληροφοριακό σύστημα να αποτυπώσει την ανθρώπινη λογική, ώστε να εξάγει συμπεράσματα, εκτιμήσεις και προβλέψεις που απαιτούν ευφυΐα. Στην περίπτωση αυτή, απαιτούνται ειδικές τεχνικές ευφυούς προγραμματισμού και σύνθετες τεχνολογίες που η καθεμιά τους είναι κατάλληλη για περιορισμένους μόνο τύπους προβλημάτων, ενώ δεν είναι πάντα αποτελεσματικές. Πηγαίνοντας ακόμα παραπέρα, θα θέλαμε, σε ορισμένες εφαρμογές, ένα πληροφοριακό σύστημα να αντιδρά με επιτυχία σε δεδομένα και καταστάσεις που δεν έχουν προβλεφτεί από τον κατασκευαστή του. Κάτι τέτοιο είναι ιδιαίτερα δύσκολο να επιτευχθεί σε ένα συμβατικό σύστημα Η/Υ.

1.3 Στόχοι και διάρθρωση του βιβλίου

Το βιβλίο απευθύνεται σε προπτυχιακούς ή μεταπτυχιακούς φοιτητές διοίκησης επιχειρήσεων ή/και μάρκετινγκ, όπως επίσης και σε στελέχη επιχειρήσεων που επιθυμούν να ενισχύσουν την ικανότητά τους στην εφαρμογή επιχειρηματικής ευφυΐας. Το σύγγραμμα καλύπτει δύο βασικές αλληλένδετες μεταξύ τους γνωστικές περιοχές της πληροφορικής. Στο πρώτο μέρος του, το βιβλίο περιλαμβάνει τις βασικές αρχές βάσεων δεδομένων και ειδικότερα θέματα συλλογής και οργάνωσης δεδομένων σε περιβάλλον επιχειρήσεων. Στο δεύτερο μέρος εισάγει τους αναγνώστες στα θέματα επιχειρηματικής ευφυΐας όπως εξόρυξη γνώσης από δεδομένα, επεξεργασία και ανάλυση δεδομένων, μοντέλα πρόβλεψης, αναπαράσταση γνώσης και στήριξη αποφάσεων.

Σκοπός του βιβλίου συνολικά είναι η εξοικείωση, μέσω εύληπτης θεωρίας και πρακτικών εφαρμογών, με τα σύγχρονα εργαλεία συλλογής και ανάλυσης δεδομένων, της εξαγωγής χρήσιμης πληροφορίας και της αποτελεσματικής χρήσης της πληροφορίας αυτής στην επίλυση προβλημάτων. Παρουσιάζονται εφαρμογές όπως πρόβλεψη συμπεριφοράς καταναλωτών, επιλογή αγοράς στόχου, τοποθέτηση προϊόντων, μέτρηση της αποτελεσματικότητας ενεργειών προώθησης, εντοπισμός ευκαιριών διασταυρωμένων πωλήσεων, κ.α. Εξετάζονται επίσης μέθοδοι υποστήριξης αποφάσεων βασισμένες σε γνώση και παρουσιάζονται τεχνολογίες δημιουργίας μοντέλων επεξήγησης ή πρόβλεψης για την επίλυση επιχειρηματικών προβλημάτων, όπως εξεύρεση καταναλωτικών προτύπων, μέτρηση πιστότητας πελατών και προσδιορισμός των παραγόντων που την επηρεάζουν. Το σύγγραμμα είναι προσαρμοσμένο στις ανάγκες ενός φοιτητή ή στελέχους διοίκησης επιχειρήσεων και διαφέρει τόσο από συγγράμματα για πληροφορικούς ή μηχανικούς (που εμβαθύνουν στη θεωρητική θεμελίωση), όσο και από πρακτικούς οδηγούς χρήσης λογισμικού. Παρέχονται στοιχεία θεωρίας που επαρκούν στην κατανόηση των βασικών αρχών και τρόπου σκέψης που απαιτείται για την επίλυση τυπικών προβλημάτων, αλλά περιλαμβάνονται και εφόδια πρακτικής εφαρμογής. Επίσης το σύγγραμμα δεν περιορίζεται στην περιγραφική παρουσίαση συστημάτων ή στη συζήτηση περιπτώσεων, αλλά καθοδηγεί τον αναγνώστη στο να επιλύει το ίδιος πραγματικά προβλήματα με χρήση κατάλληλου λογισμικού.

Στο επόμενο κεφάλαιο (2) περιγράφονται οι τρόποι αναπαράστασης και οργάνωσης δεδομένων σε ένα πληροφοριακό σύστημα και γίνεται αναφορά στις Βάσεις Δεδομένων. Στο κεφάλαιο 3, γίνεται εκτενής αναφορά στο σχεσιακό μοντέλο, που αποτελεί τον κυριότερο τρόπο οργάνωσης δεδομένων σε μια Βάση Δεδομένων, ενώ στο κεφάλαιο 4 παρουσιάζεται η πλήρης διαδικασία υλοποίησης και αξιοποίησης μιας Βάσης Δεδομένων. Στόχος των τριών αυτών κεφαλαίων είναι να κατανοήσει ο αναγνώστης τη φύση των δεδομένων και τις σχετικές δυνατότητες και περιορισμούς, έτσι ώστε να εργάζεται αποτελεσματικά με επιχειρηματικά δεδομένα και να μπορεί με χρήση ευρέως διαδεδομένου λογισμικού (Microsoft Access) να διαχειρίζεται σωστά τα δεδομένα που αφορούν την εργασία του. Στη συνέχεια, με θεμέλιο τη σωστή οργάνωση και κατανόηση των δεδομένων, στο κεφάλαιο 5 παρουσιάζονται οι τρόποι αξιοποίησής τους, μετατρέποντάς τα σε πληροφορία χρήσιμη στην επίλυση προβλημάτων. Δίνονται παραδείγματα και οδηγίες εφαρμογής που αφορούν τα συνηθέστερα, απλούστερα αλλά και ουσιαστικά εργαλεία για πληροφορημένη λήψη αποφάσεων και σχεδιασμό βασισμένο σε στοιχεία. Τα κεφάλαια 6, 7, 8 και 9 εισέρχονται βαθύτερα στην Επιχειρηματική Ευφυΐα, παρουσιάζοντας μεθόδους και εφαρμογές με υψηλότερο βαθμό ευφυΐας, που βασίζονται σε ειδικό λογισμικό. Ο αναγνώστης θα είναι σε θέση να γνωρίσει και να δημιουργήσει ο ίδιος εφαρμογές που δίνουν μεγαλύτερη αξία στα δεδομένα του, επιλύοντας προβλήματα που μπορούν να αναβαθμίσουν δραστικά την αποτελεσματικότητά του στη λήψη αποφάσεων και το σχεδιασμό στο μάρκετινγκ και τη διοίκηση.

Βιβλιογραφία/Αναφορές

Elbashir, M. Z., Collier, P. A., Davern, M. J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. International Journal of Accounting Information Systems, 9(3), 135-153.

Laudon K.C & Laudon J.P. (2009). *Πληροφοριακά Συστήματα Διοίκησης* (8η έκδοση). Αθήνα: Κλειδάριθμος.

Power, D.J. (2007). *A Brief History of Decision Support Systems*. Retrieved from http://DSSResources.COM/history/dsshistory.html, version 4.0.

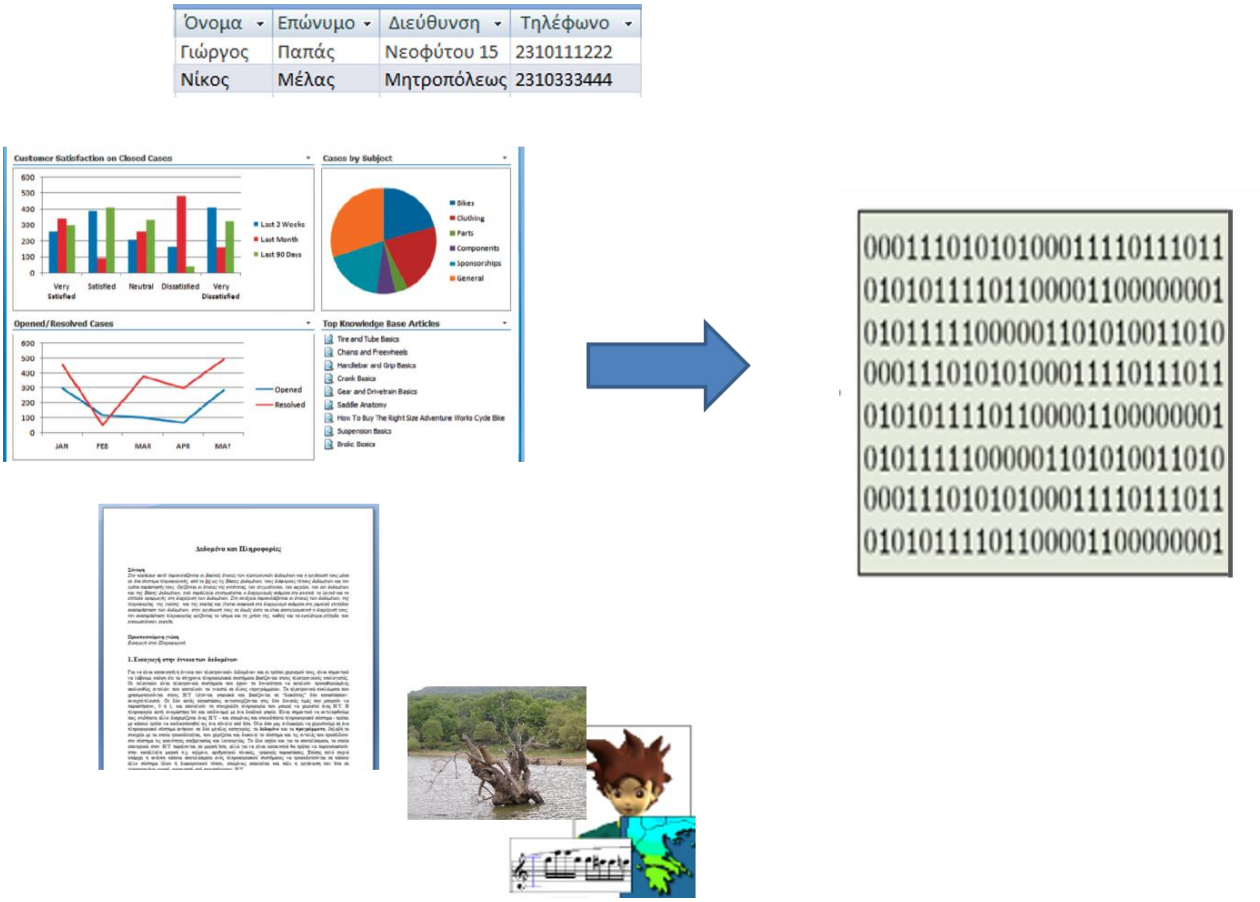
Κεφάλαιο 2. Δεδομένα και Πληροφορίες

Σύνοψη
Στο κεφάλαιο αυτό παρουσιάζονται οι βασικές έννοιες των ηλεκτρονικών δεδομένων και η οργάνωσή τους μέσα σε ένα σύστημα πληροφορικής, από το bit ως τις Βάσεις Δεδομένων, τους διάφορους τύπους δεδομένων και τον τρόπο παράστασής τους. Ορίζονται οι έννοιες της οντότητας, του στιγμιότυπου, του αρχείου, του σετ δεδομένων και της Βάσης Δεδομένων, ενώ παράλληλα επισημαίνεται ο διαχωρισμός ανάμεσα στο φυσικό, το λογικό και το επίπεδο εφαρμογής στη διαχείριση των δεδομένων. Στη συνέχεια παρουσιάζονται οι έννοιες των δεδομένων, της πληροφορίας, της γνώσης και της σοφίας και γίνεται αναφορά στο διαχωρισμό ανάμεσα στη χαμηλού επιπέδου αναπαράσταση των δεδομένων, στην οργάνωσή τους σε δομές ώστε να είναι αποτελεσματική η διαχείρισή τους, την αναπαράσταση πληροφορίας ορίζοντας το νόημα και τη χρήση της, καθώς και τα υψηλότερα επίπεδα που ενσωματώνουν ευφυΐα.

Προαπαιτούμενη γνώση
Εισαγωγή στην Πληροφορική

2.1 Εισαγωγή στην έννοια των δεδομένων

Για να είναι κατανοητή η έννοια των ηλεκτρονικών δεδομένων και οι τρόποι χειρισμού τους, είναι σημαντικό να λάβουμε υπόψη ότι τα σύγχρονα πληροφοριακά συστήματα βασίζονται στους ηλεκτρονικούς υπολογιστές. Οι τελευταίοι είναι ηλεκτρονικά συστήματα που έχουν τη δυνατότητα να εκτελούν προκαθορισμένες ακολουθίες εντολών που αποτελούν τα γνωστά σε όλους «προγράμματα». Τα ψηφιακά ηλεκτρονικά κυκλώματα που χρησιμοποιούνται στους Η/Υ βασίζονται σε "διακόπτες" δύο καταστάσεων: ανοιχτό/κλειστό. Οι δύο αυτές καταστάσεις αντιστοιχίζονται στις δύο δυνατές τιμές που μπορούν να παραστήσουν, 0 ή 1, και αποτελούν τη στοιχειώδη πληροφορία που μπορεί να χειριστεί ένας Η/Υ. Η πληροφορία αυτή ονομάστηκε bit (binary digit) και ισοδυναμεί με ένα δυαδικό ψηφίο. Είναι σημαντικό να αντιληφθούμε πως οτιδήποτε άλλο διαχειρίζεται ένας Η/Υ - και επομένως και οποιοδήποτε πληροφοριακό σύστημα - πρέπει με κάποιο τρόπο να κωδικοποιηθεί ως ένα σύνολο από bits (Σχήμα 2.1). Όλα όσα μας ενδιαφέρει να χειριστούμε σε ένα πληροφοριακό σύστημα ανήκουν σε δύο μεγάλες κατηγορίες: τα **δεδομένα** και τα **προγράμματα**, δηλαδή τα στοιχεία με τα οποία τροφοδοτείται, που χειρίζεται και διακινεί το σύστημα και τις εντολές που προσδίδουν στο σύστημα τις ικανότητες επεξεργασίας και λειτουργίας. Το ίδιο ισχύει και για τα αποτελέσματα, τα οποία εσωτερικά στον Η/Υ παράγονται σε μορφή bits, αλλά για να είναι κατανοητά θα πρέπει να παρουσιαστούν στην κατάλληλη μορφή π.χ. κείμενο, αριθμητικοί πίνακες, γραφικές παραστάσεις. Επίσης πολύ συχνά υπάρχει η ανάγκη κάποια αποτελέσματα ενός πληροφοριακού συστήματος να τροφοδοτούνται σε κάποιο άλλο σύστημα ίδιου ή διαφορετικού τύπου, επομένως απαιτείται και πάλι η οργάνωση των bits σε τυποποιημένη μορφή, κατανοητή από περισσότερους Η/Υ.



Σχήμα 2.1. *Αναπαράσταση δεδομένων ως δυαδικά ψηφία*

Γίνεται λοιπόν φανερό ότι για την επίλυση ενός προβλήματος του πραγματικού κόσμου αλλά και για να επικοινωνήσει με τον άνθρωπο ένα σύστημα που στη βάση του αναγνωρίζει μόνο δυαδικούς αριθμούς, απαιτούνται:

- Τεχνικές αναπαράστασης της πληροφορίας που αναγνωρίζει ο άνθρωπος σε μορφή που να μπορεί να χειριστεί ο Η/Υ και αντίστροφα. Αυτό επιτυγχάνεται με κατάλληλες τεχνικές κωδικοποίησης και οργάνωσης, τις οποίες θα γνωρίσουμε στα επόμενα κεφάλαια αυτού του βιβλίου. Οι τεχνικές αυτές θεωρούνται ώριμες, δηλαδή οι τεχνολογίες που έχουν ήδη αναπτυχθεί, επιτρέπουν σήμερα την αποτελεσματική αναπαράσταση και διαχείριση της πληροφορίας, μέσω πολλών εργαλείων που είναι διαθέσιμα, δοκιμασμένα και ευρέως διαδεδομένα.
- Κωδικοποίηση της λογικής επίλυσης ενός προβλήματος, δηλαδή των βημάτων που θα πρέπει να ακολουθήσει ο Η/Υ για να πραγματοποιήσει σωστά μια λειτουργία ή να παράγει ένα αποτέλεσμα. Αυτό επιτυγχάνεται με τις κατάλληλες τεχνολογίες λογισμικού, σχετικά εύκολα όταν ο τρόπος επίλυσης του προβλήματος μπορεί να προκαθοριστεί και να διατυπωθεί ξεκάθαρα, αλλά πολύ πιο δύσκολα όταν το πρόβλημα είναι ασαφές και δεν είναι ξεκάθαρος ο τρόπος λύσης του.
- Κωδικοποίηση αυτού που θα ονομάζαμε γνώση, δηλαδή της ικανότητας που θα θέλαμε να εισάγουμε σε μια μηχανή του να καταλαβαίνει τη σημασία της κάθε πληροφορίας, να κατανοεί ένα πρόβλημα και να μπορεί να επιλέξει ή και να χρησιμοποιήσει η ίδια η μηχανή τη διαθέσιμη πληροφορία για την επίλυση του προβλήματος. Η γνώση είναι πιο δύσκολο να κωδικοποιηθεί και να εισαχθεί σε ένα υπολογιστικό σύστημα σε σχέση με την πληροφορία και απαιτούνται για αυτό πιο σύνθετες τεχνικές. Οι τεχνολογίες αναπαράστασης γνώσης βρίσκονται σε εξέλιξη και έχουν μεγάλο ερευνητικό ενδιαφέρον, αφού υπάρχουν πολλά προβλήματα που δεν έχουν ακόμα επιλυθεί.

Στη συνέχεια του κεφαλαίου αυτού, θα μας απασχολήσει το πρώτο από τα παραπάνω στοιχεία, δηλαδή η αναπαράσταση της πληροφορίας, η οποία, όπως θα εξηγηθεί, βασίζεται στην αναπαράσταση και κατάλληλη οργάνωση των ηλεκτρονικών δεδομένων.

2.2 Σχέση δεδομένων και πληροφορίας

Βασικός σκοπός κάθε πληροφοριακού συστήματος είναι η κάλυψη των πληροφοριακών αναγκών των χρηστών του, δηλαδή των αναγκών για συγκέντρωση, οργάνωση, αποθήκευση, αναζήτηση και διάχυση πληροφορίας. Συχνότερα όμως ίσως από τη λέξη πληροφορία, συναντάμε τη λέξη δεδομένα. Τι είναι δεδομένα και τι πληροφορία;

Μπορούμε να ορίσουμε τέσσερα επίπεδα «πληροφορίας» τα οποία, ανάλογα με το πρόβλημά μας, μπορεί να θέλουμε να διαχειριστούμε με τη βοήθεια ενός πληροφοριακού συστήματος. Ως χαμηλότερο επίπεδο θεωρούμε αυτό που βρίσκεται πιο κοντά στη μηχανή και που είναι προσαρμοσμένο στο δικό της τρόπο λειτουργίας, ενώ υψηλότερο επίπεδο είναι αυτό που βρίσκεται πιο κοντά στη λογική του ανθρώπου - ως χρήστη του τελικού αποτελέσματος - και που περιλαμβάνει εντονότερα στοιχεία οργάνωσης και εξυπνάδας (Σχήμα 2.2). Τα τέσσερα αυτά επίπεδα, από το χαμηλότερο προς το υψηλότερο είναι:

1. **Δεδομένα**. Αποτελούνται από αριθμούς, κείμενο ή σήματα όπως εικόνες, ήχος και βίντεο, που μπορούν να καταγραφούν σε ένα σύστημα. Τα δεδομένα είναι οτιδήποτε μπορεί να καταγραφεί στον πραγματικό κόσμο και να εισαχθεί σε έναν Η/Υ για αποθήκευση και επεξεργασία, χωρίς απαραίτητα να είναι ξεκάθαρη η σημασία του. Εσωτερικά στον Η/Υ, όλα τα δεδομένα παριστάνονται σε μορφή bits, όμως με την κατάλληλη οργάνωση και το αντίστοιχο λογισμικό, ο χρήστης μπορεί να τα βλέπει σε μορφή αντιληπτή από τον άνθρωπο, όπως κείμενο, εικόνα κλπ. Για παράδειγμα, ως δεδομένα θεωρούμε έναν πίνακα με κείμενα και αριθμούς, όπου τα κείμενα είναι ονόματα πελατών και οι αριθμοί είναι ένας κωδικός για τον κάθε πελάτη.
2. **Πληροφορία**. Είναι τα δεδομένα που συνοδεύονται από μια ερμηνεία, έχουν νόημα για τον άνθρωπο και συγκεκριμένη χρησιμότητα. Ο παραπάνω πίνακας δεδομένων θεωρείται πληροφορία αν γνωρίζουμε ότι περιέχει τα στοιχεία των πελατών μας που έκαναν αγορές μέσα στην προηγούμενη εβδομάδα. Συχνά η πληροφορία είναι αποτέλεσμα επεξεργασίας δεδομένων, όπως π.χ. η συνολική αξία των πωλήσεων του μήνα ενός καταστήματος, που προκύπτει από την άθροιση της αξίας όλων των συναλλαγών.
3. **Γνώση**. Αποτελείται από επιλεγμένες πληροφορίες, μαζί με την ικανότητα χρήσης τους στην επίλυση προβλημάτων. Η Γνώση μπορεί να βασίζεται σε συνδυασμό από πληροφορίες, αφορά συγκεκριμένο θέμα και μπορεί να οδηγήσει στη λήψη απόφασης. Παράδειγμα γνώσης είναι ο κανόνας ότι αν ένας πελάτης σούπερ μάρκετ αγοράσει ζυμαρικά, με μεγάλη πιθανότητα θα αγοράσει και τυρί.
4. **Σοφία.** Σύνολο από γνώσεις και εμπειρία που συνδυάζεται με δυνατότητα κρίσης και μπορεί να εφαρμοστεί για τη λήψη αποφάσεων σε απρόβλεπτες περιστάσεις.

Σχήμα 2.2. *Τα τέσσερα επίπεδα πληροφορίας*

Από τα παραπάνω τέσσερα επίπεδα, θα μας απασχολήσουν τα δεδομένα, η πληροφορία και η γνώση. Η έννοια της «Σοφίας» δε συμπεριλαμβάνεται στα πλαίσια του βιβλίου, επειδή αποτελεί εξαιρετικά ειδικευμένο πεδίο που δε συναντάται σε εφαρμογές που αφορούν τις επιχειρήσεις.

Τα επίπεδα πληροφορίας είναι αλληλένδετα, αφού η πληροφορία προκύπτει από την επεξεργασία των δεδομένων, ενώ η γνώση προκύπτει είτε από την ανάλυση των δεδομένων, είτε από την κωδικοποίηση και γενίκευση επιλεγμένης πληροφορίας. Τα δεδομένα αποτελούν το πρωτογενές υλικό από το οποίο προκύπτουν όλα τα υπόλοιπα και είναι αυτό στο οποίο βασίζονται όλα τα πληροφοριακά συστήματα. Για την καλύτερη κατανόηση της σχέσης δεδομένων, πληροφορίας και γνώσης, αναφέρονται οι διαδικασίες που μπορούν να πραγματοποιηθούν πάνω στα δεδομένα και την πληροφορία, οι οποίες από τις πιο βασικές ως τις πιο ευφυείς είναι οι ακόλουθες:

- **Διαχείριση δεδομένων** είναι η αποθήκευσή τους, η χρήση, η συντήρηση και η διάθεσή τους, με τρόπο αποτελεσματικό, αξιόπιστο και ασφαλή. Σχετική είναι και η **μοντελοποίηση δεδομένων (Data Modeling)**, που αφορά την τυποποίηση και οργάνωση των δεδομένων.
- **Επεξεργασία δεδομένων** είναι η εκτέλεση υπολογισμών, η τροποποίηση της μορφής τους και η οργάνωσή τους με τρόπο που να επιτρέπει την καλύτερη αξιοποίησή τους, την εξαγωγή συμπερασμάτων και τη μεταφορά τους. Επεξεργασία είναι για παράδειγμα η καταγραφή της βαθμολογίας κάθε τριμήνου στα μαθήματα του σχολείου για τους μαθητές μιας τάξης, ώστε να υπολογίζονται αυτόματα οι μέσοι όροι κάθε μαθητή, ανά μάθημα και συνολικά.
- **Ανάλυση δεδομένων** είναι η εφαρμογή ειδικών στατιστικών ή υπολογιστικών μεθόδων με σκοπό την ανάδυση υψηλού επιπέδου πληροφορίας, που δεν είναι ορατή με επισκόπηση ή επεξεργασία των δεδομένων, για την επίλυση συγκεκριμένων προβλημάτων. Τυπικά προβλήματα ανάλυσης είναι η εύρεση συσχετίσεων, η ταξινόμηση ατόμων/αντικειμένων και η επιβεβαίωση θεωρητικών μοντέλων για τη μελέτη νόμων που διέπουν τα δεδομένα.
- **Εξαγωγή πληροφορίας από δεδομένα** είναι η διαδικασία παραγωγής με αυτόματο τρόπο χρήσιμης δομημένης πληροφορίας που βρίσκεται κρυμμένη σε σύνολα αδόμητων, όχι κατάλληλα οργανωμένων ή ακατέργαστων δεδομένων. Μπορούμε να πούμε ότι είναι η εύρεση μέσα από δεδομένα χωρίς ξεκάθαρη σημασία, λογικών συμπερασμάτων που μπορούν να χρησιμοποιηθούν στην επίλυση προβλημάτων. Η εξαγωγή πληροφορίας επιτυγχάνεται με συνδυασμό μεθόδων επεξεργασίας και ανάλυσης δεδομένων και έχει συνήθως ως στόχο την αυτοματοποιημένη εξαγωγή πληροφορίας από δεδομένα που δεν προορίζονταν για αυτόν το σκοπό. Παραδείγματα εξαγωγής πληροφορίας από δεδομένα είναι η επεξεργασία κειμένων σε φυσική ανθρώπινη γλώσσα, η επεξεργασία δεδομένων από συναλλαγές πώλησης για τον υπολογισμό π.χ. της κερδοφορίας των προϊόντων ανά κατηγορία και η αναζήτηση στο διαδίκτυο από αυτόματους μηχανισμούς (ρομπότ, όπως αποκαλούνται) για π.χ. παρόχους κάποιου προϊόντος.
- **Εξαγωγή γνώσης από δεδομένα** είναι διαδικασία αντίστοιχη με την εξαγωγή πληροφορίας, με τη διαφορά ότι αυτό που εξάγεται είναι γνώση, δηλαδή ευρήματα που μπορεί να χρησιμοποιήσει η ίδια η μηχανή για την επίλυση προβλημάτων. Εξαγωγή γνώσης από δεδομένα συναλλαγών πώλησης είναι π.χ. η εύρεση κανόνων για το πώς επηρεάζει η ηλικία των πελατών την επιλογή τύπου προϊόντος. Στο χώρο αυτό ανήκει η **μηχανική μάθηση (Machine Learning)**, ενώ στο κομμάτι της καταγραφής και αξιοποίησης της εξαχθείσας γνώσης αναφέρεται η **μηχανική της γνώσης (Knowledge Engineering).** Επίσης, στα στενά όρια ανάμεσα στην εξαγωγή πληροφορίας και την εξαγωγή γνώσης τοποθετούνται η αναλυτική επεξεργασία δεδομένων (Data Analytics) και η προβλεπτική ανάλυση (Predictive Analytics).
- **Εξόρυξη δεδομένων (Data mining)** ή εξόρυξη πληροφορίας από δεδομένα ή εξόρυξη γνώσης από δεδομένα είναι έννοια συγγενική με αυτήν της εξαγωγής από δεδομένα, με τη διαφορά ότι αναφέρεται σε μεγάλο όγκο ανομοιόμορφων δεδομένων, μειωμένης αξιοπιστίας, από πολλαπλές πηγές και με ανύπαρκτη ή μη ελεγχόμενη δόμηση. Η διαδικασία αυτή είναι καθαρά εξερευνητική, χωρίς εγγυημένο αποτέλεσμα και απαιτεί ειδικές τεχνολογίες ικανές να αντιμετωπίσουν τις αυξημένες απαιτήσεις των τεράστιων όγκων δεδομένων. Συναφής όρος, ιδιαίτερα δημοφιλής τα τελευταία χρόνια, είναι αυτός των **μεγάλων δεδομένων (Big Data).**

Στις ενότητες που ακολουθούν στη συνέχεια αυτού του κεφαλαίου, περιγράφονται οι τρόποι αναπαράστασης των δεδομένων σε ένα σύστημα Η/Υ, δηλαδή το πώς οργανώνονται τα δεδομένα που αφορούν τον πραγματικό κόσμο έτσι ώστε να μπορεί να τα διαχειριστεί ένα ψηφιακό σύστημα. Η παράθεση πραγματοποιείται ξεκινώντας από τα χαμηλότερα επίπεδα (αυτά που αφορούν τη σωστή λειτουργία της μηχανής) και οδεύοντας προς τα ανώτερα επίπεδα (αυτά που είναι πιο κοντά στον τρόπο σκέψης του ανθρώπου). Παρόλο που η έννοια των δεδομένων διαχωρίζεται από την έννοια της πληροφορίας, σύμφωνα με τους παραπάνω ορισμούς, πολύ συχνά στην πράξη η διαχείριση και οργάνωση των δεδομένων τα φέρνει πολύ κοντά στην έννοια της πληροφορίας έτσι ώστε να είναι δύσκολο να καθορίσουμε με απόλυτη σαφήνεια τα όρια ανάμεσα στη διαχείριση δεδομένων και τη διαχείριση πληροφορίας. Στη συνέχεια του κεφαλαίου γίνεται εκτενής αναφορά στην αναπαράσταση δεδομένων, σημειώνεται όμως ότι αυτή αφορά σε κάποιο βαθμό και την αναπαράσταση πληροφορίας. Η αναπαράσταση γνώσης απαιτεί ακόμα υψηλότερα επίπεδα οργάνωσης και ειδικές τεχνολογίες. Στο χειρισμό της Γνώσης από ένα ευφυές σύστημα πληροφορικής γίνεται εκτενής αναφορά στο Κεφάλαιο 8.

2.3 Αναπαράσταση Δεδομένων

Όλα τα δεδομένα που διαχειρίζεται ένας Η/Υ (κείμενο, αριθμητικά στοιχεία, εικόνες, κλπ.) παριστάνονται τελικά σαν ακολουθίες δυαδικών αριθμών. Επειδή ένα δυαδικό ψηφίο περιέχει ελάχιστη πληροφορία, χρησιμοποιούνται στην πράξη ομάδες δυαδικών ψηφίων που σχηματίζουν πιο σύνθετες δομές δεδομένων. Είναι ευθύνη των προγραμμάτων/εφαρμογών να κωδικοποιήσουν τις πληροφορίες που διαχειρίζονται σε κατάλληλη μορφή για τον Η/Υ και, αντίστροφα, να ερμηνεύσουν τα δυαδικά ψηφία ως χρήσιμη πληροφορία.

2.3.1 Τύποι δεδομένων

Το πρώτο και πιο στοιχειώδες επίπεδο οργάνωσης των δεδομένων, ώστε να μπορούν αυτά να παρασταθούν σε έναν Η/Υ, είναι ο ορισμός τύπων δεδομένων και προτύπων κωδικοποίησης (Πίνακας 2.1).

Τύπος	**Μέγεθος σε Bytes**	**Περιγραφή**
Byte	1	Ομάδα 8 δυαδικών ψηφίων που μπορεί να παραστήσει ως 2^8=256 διαφορετικές τιμές. Μπορεί να ερμηνευτεί ως ένας θετικός ακέραιος από 0 ως 255 ή προσημασμένος ακέραιος από -127 ως 128 ή να χρησιμοποιηθεί ως κωδικός.
Ακέραιος	2	Περιλαμβάνει ακέραιες αριθμητικές τιμές από -32.768 ως 32.767
Μεγάλος Ακέραιος	4	Περιλαμβάνει ακέραιες αριθμητικές τιμές με μεγαλύτερο εύρος, από -2,147,483,648 ως 2,147,483,647
Πραγματικός αριθμός απλής ακρίβειας	4	Περιλαμβάνει πραγματικούς αριθμούς με απλή ακρίβεια μέσα στο διάστημα -3.402823E38 ως -1.401298E-45 για αρνητικές τιμές και από 1.401298E-45 ως 3.402823E38 για θετικές τιμές.
Πραγματικός διπλής ακρίβειας	8	Περιλαμβάνει πραγματικούς αριθμούς με διπλή ακρίβεια μέσα στο διάστημα από -1.79769313486231E308 ως -4.94065645841247E-324για αρνητικές τιμές και από 4.94065645841247E-324 ως 1.79769313486232E308 για θετικές τιμές.
Κείμενο	1 χαρακτήρας = 1 Byte. Μπορεί να καθοριστεί σύμφωνα με το μέγιστο αριθμό χαρακτήρων που απαιτούνται	Ένας αριθμός από 0-255 (1 Byte) αντιστοιχεί σε ένα χαρακτήρα. Περιλαμβάνονται σύμβολα και ειδικοί χαρακτήρες

Πίνακας 2.1. *Βασικοί τύποι δεδομένων*

2.3.1.1. Αριθμητικά δεδομένα

Η αναπαράσταση των αριθμητικών δεδομένων σε Bytes (ή ουσιαστικά σε bits) γίνεται κωδικοποιώντας τα σε ομάδες από Bytes, ανάλογα με την επιθυμητή ακρίβεια π.χ. οι απλοί ακέραιοι απαιτούν 2 Bytes (ή ισοδύναμα 16 bits) και μπορούν να παραστήσουν ακέραιες τιμές από -32.768 ως 32767 ή από 0 ως 65536, ανάλογα με το αν τους θεωρούμε θετικούς ή προσημασμένους, πραγματικοί αριθμοί απλής ακρίβειας αποθηκεύονται σε 4 Bytes, κλπ. Η αντιστοίχιση αυτή γίνεται με τρόπο προκαθορισμένο από το λογισμικό του συστήματος ή της εφαρμογής και δεν απασχολεί το χρήστη μιας εφαρμογής. Αυτό που ενδιαφέρει το χρήστη είναι η **επιλογή του σωστού τύπου δεδομένων ανάλογα με τις ανάγκες και η κατανόηση των περιορισμών του κάθε τύπου**. Π.χ. η τιμή ενός προϊόντος σε € πρέπει να αποθηκεύεται ως πραγματικός απλής ακρίβειας, επειδή αν αποθηκευτεί ως ακέραιος, θα χάνονται τα δεκαδικά ψηφία που αντιστοιχούν στα λεπτά. Ο αριθμός τεμαχίων μιας παραγγελίας θα πρέπει να είναι ακέραιος, αφού τιμές με δεκαδικά ψηφία δεν έχουν έννοια.

2.3.1.2 Κείμενο

Η μετατροπή του κειμένου σε Bytes γίνεται εφαρμόζοντας κάποιο πρότυπο κωδικοποίησης. Ένα από τα πρώτα και ίσως το πιο ευρέως διαδεδομένο πρότυπο κωδικοποίησης κειμένου είναι ο κώδικας ASCII **(**American Standard Code for Information Interchange) (ASCII, 1968), που αποτελεί την πιο απλή κωδικοποίηση κειμένου για αποθήκευση σε Η/Υ και δεν περιλαμβάνει πληροφορία μορφοποίησης Σύμφωνα με αυτόν, η τιμή ενός Byte παριστάνει ένα χαρακτήρα ή σύμβολο έτσι ώστε να αντιστοιχεί ένα Byte για κάθε χαρακτήρα. Ο κώδικας περιλαμβάνει όλους τους χαρακτήρες του αλφαβήτου, τα αριθμητικά ψηφία, ειδικούς χαρακτήρες και σύμβολα, ενώ υπάρχει πρόβλεψη και για χαρακτήρες άλλων γλωσσών εκτός από αυτούς του λατινικού αλφαβήτου (π.χ. ελληνικό αλφάβητο). Ο κώδικας ASCII, που παρουσιάζεται στον Πίνακα 2.2, αναγνωρίζεται από όλες σχεδόν τις εφαρμογές ως τύπος «απλού κειμένου» και χρησιμοποιείται ως εσωτερικός τρόπος αναπαράστασης κειμένου στις περισσότερες εφαρμογές διαχείρισης δεδομένων. Πιο σύνθετοι τρόποι κωδικοποίησης κειμένου ενσωματώνουν στοιχεία μορφοποίησης και πιο σύνθετα στοιχεία (π.χ. πίνακες, μαθηματικές εκφράσεις, κλπ.). Τέτοιοι τύποι κειμένου αντιστοιχούν σε συγκεκριμένες εφαρμογές και έχουν σχεδιαστεί ώστε να εξυπηρετούν τις ειδικές ανάγκες των εφαρμογών αυτών.

Δυαδ.	Οκτ.	Δεκ.	Δεκαεξ.	Γραφ.
010 0000	040	32	20	SP
010 0001	041	33	21	!
010 0010	042	34	22	"
010 0011	043	35	23	#
010 0100	044	36	24	$
010 0101	045	37	25	%
010 0110	046	38	26	&
010 0111	047	39	27	'
010 1000	050	40	28	(
010 1001	051	41	29	)
010 1010	052	42	2A	*
010 1011	053	43	2B	+
010 1100	054	44	2C	,
010 1101	055	45	2D	-
010 1110	056	46	2E	.
010 1111	057	47	2F	/
011 0000	060	48	30	0
011 0001	061	49	31	1
011 0010	062	50	32	2
011 0011	063	51	33	3
011 0100	064	52	34	4
011 0101	065	53	35	5
011 0110	066	54	36	6
011 0111	067	55	37	7
011 1000	070	56	38	8
011 1001	071	57	39	9
011 1010	072	58	3A	:
011 1011	073	59	3B	;
011 1100	074	60	3C	<
011 1101	075	61	3D	=
011 1110	076	62	3E	>
011 1111	077	63	3F	?

Δυαδ.	Οκτ.	Δεκ.	Δεκαεξ.	Γραφ.
100 0000	100	64	40	@
100 0001	101	65	41	A
100 0010	102	66	42	B
100 0011	103	67	43	C
100 0100	104	68	44	D
100 0101	105	69	45	E
100 0110	106	70	46	F
100 0111	107	71	47	G
100 1000	110	72	48	H
100 1001	111	73	49	I
100 1010	112	74	4A	J
100 1011	113	75	4B	K
100 1100	114	76	4C	L
100 1101	115	77	4D	M
100 1110	116	78	4E	N
100 1111	117	79	4F	O
101 0000	120	80	50	P
101 0001	121	81	51	Q
101 0010	122	82	52	R
101 0011	123	83	53	S
101 0100	124	84	54	T
101 0101	125	85	55	U
101 0110	126	86	56	V
101 0111	127	87	57	W
101 1000	130	88	58	X
101 1001	131	89	59	Y
101 1010	132	90	5A	Z
101 1011	133	91	5B	[
101 1100	134	92	5C	\
101 1101	135	93	5D	]
101 1110	136	94	5E	^
101 1111	137	95	5F	_

Δυαδ.	Οκτ.	Δεκ.	Δεκαεξ.	Γραφ.
110 0000	140	96	60	`
110 0001	141	97	61	a
110 0010	142	98	62	b
110 0011	143	99	63	c
110 0100	144	100	64	d
110 0101	145	101	65	e
110 0110	146	102	66	f
110 0111	147	103	67	g
110 1000	150	104	68	h
110 1001	151	105	69	i
110 1010	152	106	6A	j
110 1011	153	107	6B	k
110 1100	154	108	6C	l
110 1101	155	109	6D	m
110 1110	156	110	6E	n
110 1111	157	111	6F	o
111 0000	160	112	70	p
111 0001	161	113	71	q
111 0010	162	114	72	r
111 0011	163	115	73	s
111 0100	164	116	74	t
111 0101	165	117	75	u
111 0110	166	118	76	v
111 0111	167	119	77	w
111 1000	170	120	78	x
111 1001	171	121	79	y
111 1010	172	122	7A	z
111 1011	173	123	7B	{
111 1100	174	124	7C	|
111 1101	175	125	7D	}
111 1110	176	126	7E	~

Πίνακας 2.2. *Ο κώδικας αντιστοίχισης Bytes με χαρακτήρες κειμένου ASCII.*

2.3.1.3 Ειδικοί τύποι

Εκτός από τους αριθμούς και τα κείμενα, είναι χρήσιμος ο ορισμός ειδικών τύπων που να τους αναγνωρίζει μια εφαρμογή και να τους χειρίζεται κατάλληλα. Οι συνηθέστεροι τέτοιοι τύποι είναι:

Ημερομηνία/ώρα: Μπορεί να παραστήσει μια ημερομηνία του ημερολογίου ή/και ώρα. Στον τύπο αυτό μπορούν να εφαρμοστούν περιορισμοί εγκυρότητας (π.χ. απαγορεύεται η ημερομηνία 32 Ιανουαρίου), μπορεί το ίδιο δεδομένο να προβληθεί αυτόματα σε διαφορετικές μορφές και μπορούν να εκτελεστούν πράξεις με ειδικό τρόπο π.χ. η αφαίρεση δύο ημερομηνιών έχει ως αποτέλεσμα τον αριθμό ημερών του χρονικού διαστήματος ανάμεσα στις ημερομηνίες αυτές.

Ναι/Οχι ή δυαδικός τύπος (binary): μπορεί να πάρει μόνο δύο τιμές (0/1) που ερμηνεύονται ως Ναι/Οχι και μπορούν να χρησιμοποιηθούν σε λογικές πράξεις, ελέγχους συνθηκών και σήμανση (flag) επιλογών.

2.4 Οργάνωση δεδομένων σε δομές

Τα επιχειρηματικά δεδομένα αποτελούνται κυρίως από κείμενα (π.χ. ονόματα, περιγραφές προϊόντων), αριθμούς π.χ. τιμές, πωλήσεις κλπ. και ειδικούς τύπους όπως ημερομηνίες και επιλογές τύπου Ναι/Όχι. Τα δεδομένα αυτά αντιστοιχούν σε κάποια πληροφορία του πραγματικού κόσμου, δηλαδή έχουν κάποιο νόημα. Επομένως, εκτός από το να κωδικοποιηθούν σωστά ώστε να μπορεί να τα χειριστεί ο Η/Υ, πρέπει και να οργανωθούν σύμφωνα με το νόημά τους, ώστε να είναι χρήσιμα. Ο επικρατέστερος τρόπος οργάνωσης επιχειρηματικών δεδομένων είναι αυτός που παρουσιάζεται στο Σχήμα 2.3 και περιλαμβάνει τις παρακάτω έννοιες, οι οποίες βασίζονται η μία στην άλλη, από την απλούστερη προς την πιο σύνθετη μορφή:

bit: Είναι ένα δυαδικό ψηφίο που περιέχει στοιχειώδη πληροφορία π.χ. «1».

Byte: Σύνολο από 8 bits που μπορεί να παραστήσει ένα μικρό αριθμό ή ένα χαρακτήρα π.χ. «Π».

Πεδίο (Field): Αποτελείται από ένα σύνολο από Bytes, έχει συγκεκριμένο μέγεθος και συγκεκριμένο τύπο δεδομένων (π.χ. αριθμό, κείμενο, κλπ.) και μπορεί να παραστήσει μια στοιχειώδη πληροφορία όπως την τιμή μιας μεταβλητής ή ένα χαρακτηριστικό. Τα πεδία έχουν κάποιο λογικό νόημα και συνήθως τους δίνεται κατάλληλο όνομα ώστε να αναδεικνύεται το νόημα των τιμών που περιέχουν. Π.χ. το πεδίο με όνομα «Επώνυμο πελάτη» μπορεί να περιέχει το δεδομένο «Παπάς». Το πεδίο αυτό πρέπει να είναι τύπου Κείμενο, που σημαίνει ότι τα Bytes που περιέχει αντιστοιχίζονται σε αλφαριθμητικούς χαρακτήρες.

Εγγραφή (Record): Είναι μία ομάδα πεδίων που σχετίζονται μεταξύ τους. Τα πεδία μπορεί να είναι διαφορετικά μεταξύ τους και συνήθως το καθένα από αυτά αφορά ένα διαφορετικό στοιχείο πληροφορίας για κάποιο αντικείμενο, πρόσωπο ή γεγονός. Π.χ. μια εγγραφή μπορεί να αποτελείται από τις τιμές «Γιώργος», «Παπαδόπουλος», «2310001122», που αποτελούν αντίστοιχα το περιεχόμενο των πεδίων «Όνομα Πελάτη», «Επώνυμο πελάτη» και «Τηλέφωνο».

Αρχείο (File). Είναι ένα σύνολο εγγραφών ίδιου τύπου. Το αρχείο έχει σειριακή μορφή, δηλαδή οι εγγραφές ακολουθούν η μία την άλλη και έχει συγκεκριμένη αρχή και τέλος. Όλες οι εγγραφές καταλαμβάνουν τον ίδιο αριθμό Bytes, ανάλογα με τον τύπο και το μέγεθος των πεδίων που περιλαμβάνονται στη δομή του, ανεξάρτητα από το συγκεκριμένο περιεχόμενο, επομένως, το μέγεθος του αρχείου σε αποθηκευτικό χώρο καθορίζεται από τον αριθμό των εγγραφών του. Παράδειγμα αρχείου είναι ένας τηλεφωνικός κατάλογος, όπου κάθε εγγραφή αντιστοιχεί σε ένα συνδρομητή. Ο κυριότερος τρόπος οργάνωσης των δεδομένων για εγγραφή σε οποιοδήποτε μέσο είναι το αρχείο δηλαδή ως μια σειριακή αποθήκευση ομοειδών δεδομένων. (**Διευκρίνιση:** Ο όρος Αρχείο (File) χρησιμοποιείται ευρέως στη σύγχρονη γλώσσα των Η/Υ και με μια παρεμφερή έννοια, που είναι ίσως περισσότερο οικεία στους αναγνώστες. Με τη λέξη Αρχείο Η/Υ εννοούμε συχνά το αντικείμενο ή «δοχείο» που περιέχει ένα σύνολο από πληροφορίες, δεδομένα ή έναν πόρο π.χ. εικόνα, πρόγραμμα, κλπ. και βρίσκεται σε κάποιο αποθηκευτικό μέσο Η/Υ. Κάθε εφαρμογή μπορεί να διατηρεί με το δικό της τρόπο τα δεδομένα που χειρίζεται σε ένα δικού της τύπου αρχείο, του οποίου τη δομή αναγνωρίζει η ίδια η εφαρμογή και μπορεί να είναι πιο σύνθετη από αυτήν του απλού σειριακού αρχείου).

Βάση δεδομένων (Data Base): Είναι οργανωμένη ομάδα αρχείων που αφορούν το ίδιο θέμα, έχουν καθορισμένη δομή, σχετίζονται μεταξύ τους και ελέγχονται από ειδικό σύστημα διαχείρισης. Ο γνωστότερος τύπος Βάσης Δεδομένων είναι η Σχεσιακή Βάση Δεδομένων που θα αναπτυχθεί στο επόμενο κεφάλαιο.

Επίσης σχετικοί είναι και οι παρακάτω ορισμοί:

Οντότητα (Entity): Είναι μια κατηγορία προσώπων, πραγμάτων ή γεγονότων για τα οποία τηρούνται πληροφορίες, π.χ. πελάτης, προϊόν. Με τη λέξη οντότητα αναφερόμαστε στην κατηγορία, ενώ κάθε συγκεκριμένο πρόσωπο, πράγμα ή γεγονός που εκπροσωπεί μια οντότητα (δηλαδή ανήκει στην κατηγορία) λέγεται **στιγμιότυπο (instance)** της οντότητας π.χ. ο πελάτης Γιώργος Παπαδόπουλος.

Ιδιότητα (Attribute): Είναι μια πληροφορία που αφορά ένα ιδιαίτερο χαρακτηριστικό μιας συγκεκριμένης οντότητας, π.χ. όνομα πελάτη, τιμή προϊόντος, περιγραφή προϊόντος.

Συνήθως στην πράξη, τα δεδομένα που αφορούν μια ιδιότητα μιας οντότητας αποθηκεύονται σε ένα πεδίο, ενώ το σύνολο των δεδομένων που αφορούν ένα στιγμιότυπο μιας οντότητας αποθηκεύονται σε μια εγγραφή.

Δομή οργάνωσης	Περιγραφή	Παράδειγμα
bit	0 ή 1	0
Byte	8 bits = ένας αριθμός από 0 ως 255	01000001 -> το γράμμα Α σε ASCII
Πεδίο (Field)	Ένα σύνολο από Bytes που παριστάνει μια τιμή με νόημα	Αλέξανδρος -> η τιμή που περιέχεται στο πεδίο Όνομα πελάτη
Εγγραφή (Record)	Σύνολο από πεδία διαφόρων τύπων που συγκεντρώνουν πληροφορία για ένα αντικείμενο	Όνομα | Επώνυμο | Διεύθυνση | Τηλέφωνο Αλέξανδρος | Νίκου | Μελά 23 | 6911222333
Αρχείο (File)	Σύνολο από εγγραφές ίδιου τύπου	Όνομα | Επώνυμο | Διεύθυνση | Τηλέφωνο Γιώργος | Παπάς | Νεοφύτου 15 | 2310111222 Νίκος | Μέλας | Μητροπόλεως | 2310333444 Αλέξανδρος | Νίκου | Μελά 23 | 6911222333
Βάση Δεδομένων (Data Base)	Σύνολο από συνδεδεμένα αρχεία που καλύπτουν τις πληροφοριακές ανάγκες σχετικά με ένα θέμα	ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ ΠΩΛΗΣΕΩΝ ΠΕΛΑΤΕΣ — ΠΑΡΑΓΓΕΛΙΕΣ — ΠΡΟΪΟΝΤΑ
Οντότητα (Entity)	Πρόσωπο, αντικείμενο ή γεγονός για το οποίο τηρούνται δεδομένα	ΠΕΛΑΤΗΣ
Χαρακτηριστικό (Attribute)	Πληροφορία που αφορά ένα χαρακτηριστικό μιας οντότητας	ΟΝΟΜΑ ΠΕΛΑΤΗ

Σχήμα 2.3. *Η οργάνωση των δεδομένων*

Επίσης υπάρχει η περίπτωση κάποια εφαρμογή να χειρίζεται ειδικά σύνθετα αντικείμενα, τα οποία συνήθως είναι πολυμεσικά αντικείμενα όπως εικόνες, ήχος, βίντεο ή γραφήματα. Τέτοιου είδους αντικείμενα κωδικοποιούνται επίσης ως ακολουθίες από bits ακολουθώντας κάποια πρότυπα και η επεξεργασία/προβολή τους γίνεται με τη βοήθεια ειδικών εργαλείων που είναι εκτός του αντικειμένου αυτού του βιβλίου. Αυτό που πρέπει να τονιστεί είναι ότι οι τύποι δεδομένων που παρουσιάστηκαν παραπάνω (δηλ αριθμοί, κείμενο, κλπ.) είναι επεξεργάσιμοι από ένα σύστημα διαχείρισης δεδομένων μέσω τυποποιημένων πράξεων και διαδικασιών, όπως εκτέλεση αναζητήσεων, τροποποιήσεων, αριθμητικών υπολογισμών, κλπ. Αντίθετα, δεδομένα σε μορφή εικόνων, βίντεο, κλπ. μπορούν μόνο να αποθηκευτούν αυτούσια και να περιγραφούν, αλλά όχι και να γίνει επεξεργασία του περιεχομένου τους. Επομένως π.χ. σε ένα αρχείο πελατών όπου τα δεδομένα είναι οργανωμένα σε πεδία και εγγραφές, μπορούμε να αναζητήσουμε τους πελάτες με το συγκεκριμένο όνομα «Γιώργος» και να πληροφορηθούμε τη διεύθυνσή τους. Αν το αρχείο περιλαμβάνει και φωτογραφίες των πελατών, θα μπορούσαμε να ανασύρουμε τη φωτογραφία του πελάτη «Γιώργου». Αν όμως ο πίνακας με τα στοιχεία των πελατών ήταν σε μορφή εικόνας, δηλαδή φωτογραφημένος από ένα έντυπο, δε θα μπορούσε το πληροφοριακό σύστημα να αναγνωρίσει το περιεχόμενο και να πραγματοποιήσει αναζήτηση.

2.5 Το Φυσικό και το λογικό επίπεδο χειρισμού των δεδομένων

Η οργάνωση και ο χειρισμός των δεδομένων, από τις φυσικές λειτουργίες που πραγματοποιούνται εσωτερικά σε έναν Η/Υ μέχρι την τελική αξιοποίησή τους στην επίλυση προβλημάτων, διακρίνεται σε 3 βασικά επίπεδα, όπως φαίνεται στο Σχήμα 2.4.

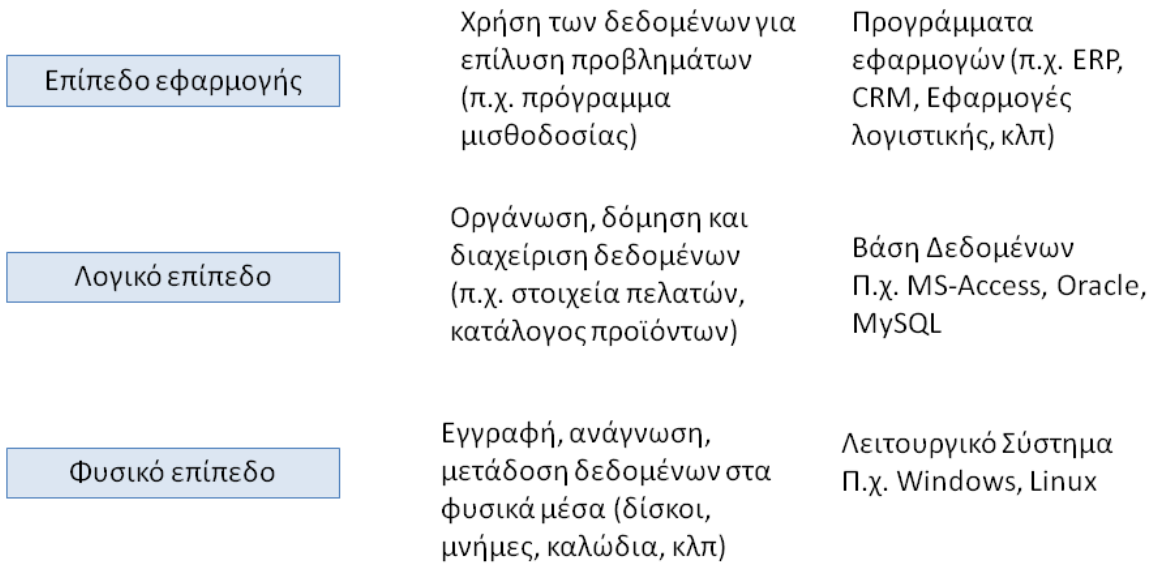
Τα χαμηλότερο επίπεδο, δηλαδή αυτό που βρίσκεται πιο κοντά στη μηχανή, είναι το **Φυσικό** επίπεδο που αφορά τις τεχνολογίες αποθήκευσης, ανάγνωσης ή μετάδοσης των δεδομένων στα φυσικά μέσα. Στο επίπεδο αυτό, τα δεδομένα είναι ακολουθίες από bits των οποίων το νόημα δεν αφορά τις σχετικές

τεχνολογίες. Τυπικές λειτουργίες είναι η εγγραφή, ανάγνωση ή μεταφορά ενός αρχείου σε κάποιο μέσο αποθήκευσης ή μέσω ενός δικτύου.

Ένα επίπεδο υψηλότερα βρίσκεται το **Λογικό** επίπεδο, όπου τα δεδομένα είναι οργανωμένα με βάση το νόημά τους. Τυπικές λειτουργίες είναι η εισαγωγή στοιχείων για το στιγμιότυπο μιας οντότητας π.χ. εισαγωγή των στοιχείων επικοινωνίας του πελάτη Γιώργου, ή η αναζήτηση στοιχείων όπως οι παραγγελίες που πραγματοποιήθηκαν σήμερα.

Το υψηλότερο επίπεδο, δηλαδή αυτό που βρίσκεται πιο κοντά στον άνθρωπο/χρήστη είναι το επίπεδο **Εφαρμογής**. Στο επίπεδο αυτό βρίσκονται οι μηχανισμοί που χειρίζονται τα δεδομένα για την επίλυση προβλημάτων και είναι αυτοί που τελικά ενδιαφέρουν τον τελικό χρήστη ώστε να αξιοποιήσει τα δεδομένα. Τυπικές λειτουργίες είναι η επεξεργασία των δεδομένων και η παρουσίαση των αποτελεσμάτων π.χ. έκδοση ενός τιμολογίου, η εκτέλεση μισθοδοσίας ή η δημιουργία μιας αναφοράς.

Ο διαχωρισμός στα παραπάνω τρία επίπεδα δεν είναι θεωρητικός αλλά ουσιαστικός. Η ανεξαρτησία των επιπέδων είναι θεμελιώδης αρχή, με τεράστια πλεονεκτήματα για την ανάπτυξη των σχετικών τεχνολογιών. Οι λειτουργίες που αφορούν κάποιο επίπεδο πρέπει να υλοποιούνται χωρίς να επηρεάζουν τα υπόλοιπα επίπεδα και η επικοινωνία ανάμεσα στα απομονωμένα μεταξύ τους επίπεδα γίνεται μέσω αυστηρά τυποποιημένων μηχανισμών που διαβιβάζουν αιτήματα και επιστρέφουν αποτελέσματα. Με τον τρόπο αυτό, κάθε επίπεδο συνεχίζει να λειτουργεί σωστά χωρίς καμία ανάγκη τροποποίησης, ακόμα και αν αλλάξουν δραστικά οι τεχνολογίες των άλλων επιπέδων. Με απλά λόγια, το πρόγραμμα που χρησιμοποιεί ένας πωλητής για να πληροφορηθεί για το ιστορικό αγορών ενός πελάτη δε θα πρέπει να επηρεάζεται από τον τρόπο με τον οποίο είναι οργανωμένα τα στοιχεία αυτά σε πεδία, εγγραφές κλπ. αλλά ένας κατάλληλος και αποτελεσματικός μηχανισμός θα φροντίζει να τηρεί τα δεδομένα με αξιοπιστία και οικονομία και να παρέχει αυτά ακριβώς που χρειάζεται η εφαρμογή τη στιγμή που τα χρειάζεται. Αντίστοιχα, μια Βάση Δεδομένων δε θα πρέπει να επηρεάζεται από τον τρόπο με τον οποίο γράφονται τα δεδομένα στο σκληρό δίσκο ενός υπολογιστή και θα πρέπει να λειτουργεί σωστά ακόμα και αν αλλάξει το φυσικό μέσο αποθήκευσης.



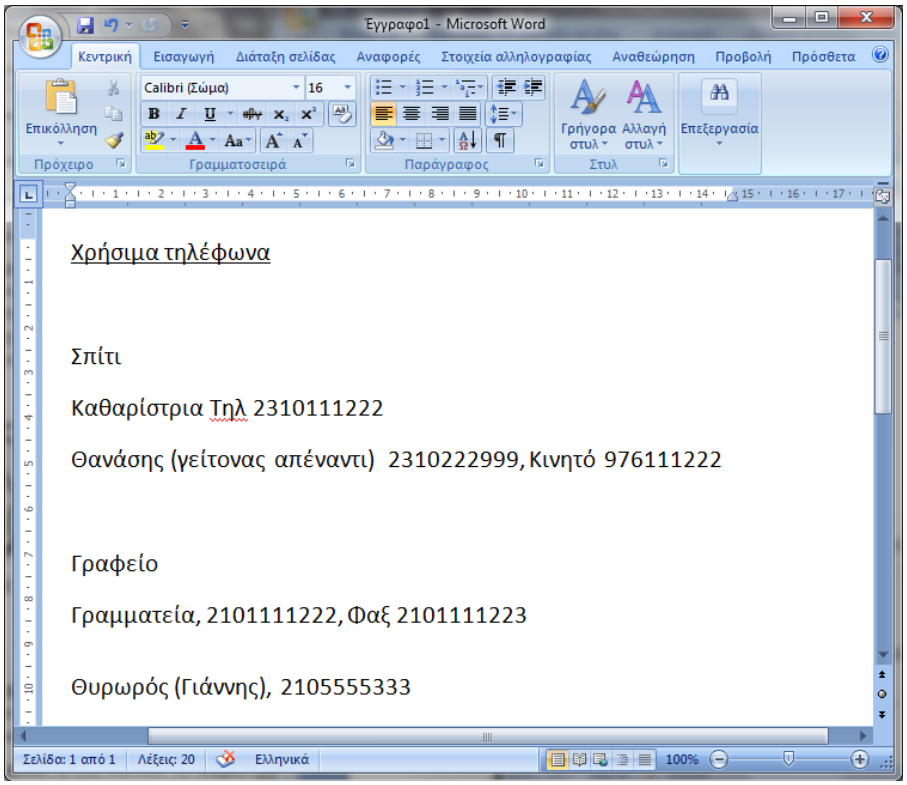
Σχήμα 2.4. *Τα λογικά επίπεδα χειρισμού των δεδομένων*

2.6 Συνήθεις τρόποι οργάνωσης και αξιοποίησης δεδομένων

Στην παράγραφο αυτή παρουσιάζονται οι διαφορετικοί τρόποι οργάνωσης των δεδομένων, από την οπτική γωνία του τελικού χρήστη και επιχειρείται μια διασύνδεση ανάμεσα στον τρόπο με τον οποίο βλέπει πρακτικά ένα στέλεχος τα δεδομένα του και στα επίπεδα οργάνωσης που αναφέρθηκαν παραπάνω, καθώς και τις συνηθέστερες εφαρμογές. Οι τρόποι οργάνωσης αυτοί, αναφέρονται από τους πιο απλούς και διαισθητικούς ως τους πιο ισχυρούς και αποτελεσματικούς.

2.6.1 Άτυπη και αδόμητη αποθήκευση σε κάποιο ηλεκτρονικό μέσο.

Οι χειρόγραφες σημειώσεις μπορούν με απλό τρόπο να αποθηκευτούν σε ηλεκτρονικά αρχεία, όπως η εγγραφή σημειώσεων σε μορφή ελεύθερου κειμένου σε κάποιο πρόγραμμα επεξεργασίας κειμένου (π.χ. MS-Word), λογιστικό φύλλο (π.χ. MS-Excel) ή ακόμα και ως ψηφιοποίηση ή φωτογράφηση κάποιου εντύπου. Στη μορφή αυτή, η πληροφορία μπορεί να αποθηκευτεί σε ηλεκτρονική μορφή, να αναπαραχθεί και να αποσταλεί μέσω δικτύου και να αναζητηθεί με κάποιον υποτυπώδη τρόπο π.χ. στα περιεχόμενα του δίσκου ενός Η/Υ ή ενός CD. Επομένως ισχύουν ορισμένα από τα πλεονεκτήματα της ηλεκτρονικής διαχείρισης δεδομένων. Ωστόσο, η απουσία οργάνωσης καθιστά τον τρόπο αυτό απολύτως αναποτελεσματικό και ακατάλληλο για ουσιαστική αξιοποίηση των δεδομένων. Ο κύριος λόγος είναι ότι η έλλειψη δομής κάνει τα δεδομένα κατανοητά μόνο από τον άνθρωπο-χρήστη και είναι δύσκολη η επεξεργασία, η επιβολή ελέγχων και η αξιοποίησή τους με αυτοματοποιημένες μεθόδους.



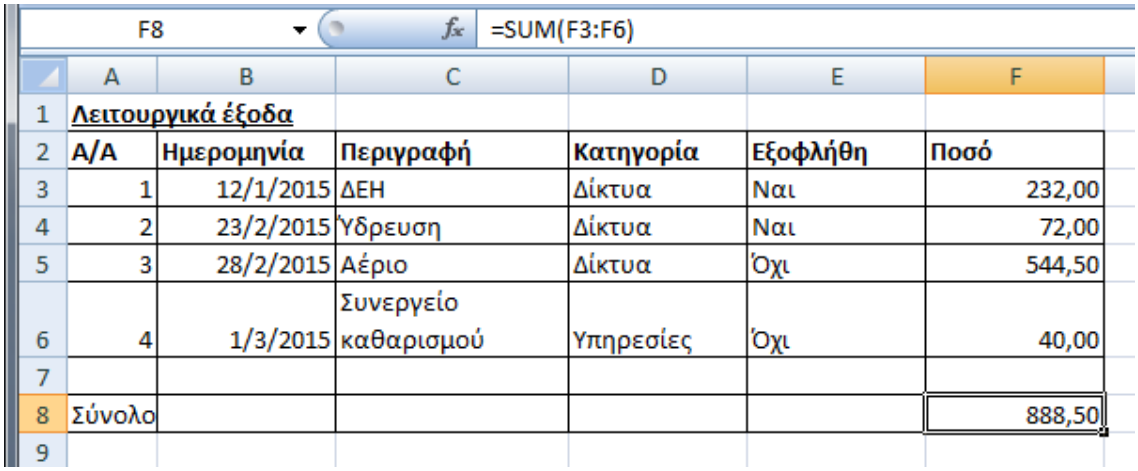
Σχήμα 2.5. *Αδόμητη αποθήκευση δεδομένων σε ηλεκτρονικό αρχείο*

2.6.2 Λογιστικά φύλλα ή πίνακες.

Τα δεδομένα οργανώνονται πολύ συχνά σε μορφή πινάκων ή γενικότερα σε γραμμές και στήλες. Οι γραμμές και οι στήλες μπορούν να έχουν τίτλους που να κάνουν σαφή την έννοιά τους και επίσης μπορεί να καθοριστεί ο τύπος δεδομένων του κύριου περιεχομένου (δηλαδή των κελιών). Η μορφή αυτή είναι απλή και κατανοητή, τόσο από τον άνθρωπο όσο και από ένα πρόγραμμα Η/Υ και είναι ιδιαίτερα αποτελεσματική και χρήσιμη για ένα μεγάλο εύρος εφαρμογών. Ο χειρισμός των δεδομένων σε αυτήν την «πινακοποιημένη» μορφή είναι ο καταλληλότερος για πολλές κατηγορίες εφαρμογών και υποστηρίζεται από πλήθος εργαλείων και προγραμμάτων. Χαρακτηριστικές περιπτώσεις είναι οι παρακάτω:

Λογιστικά φύλλα. Είναι τα γνωστά και ευρέως χρησιμοποιούμενα προγράμματα λογιστικών φύλλων όπως το Excel της Microsoft. Τα δεδομένα είναι οργανωμένα σε γραμμές και στήλες, έτσι ώστε η κάθε γραμμή να αντιστοιχεί σε κάποια περίπτωση, κάποιο γεγονός, αντικείμενο κλπ. και η κάθε στήλη επίσης σε κάποια περίπτωση ή κάποιο στοιχείο πληροφορίας που διασταυρώνεται με τις γραμμές. Τα δεδομένα μπορεί να είναι αριθμητικά, κείμενο, ημερομηνίες και γενικά όλοι οι γνωστοί τύποι που αναφέρθηκαν. Ο τύπος δεδομένων και το μέγεθος μπορεί να καθοριστεί για κάθε στήλη, γραμμή ή ομάδα κελιών, ανάλογα με τα δεδομένα που αναμένεται να εισαχθούν σε αυτές και επίσης μπορούν να επιβληθούν περιορισμοί στις τιμές των κελιών και να επιλεγούν διάφοροι τρόποι προβολής. Επιπλέον, η επεξεργασία των δεδομένων μπορεί να

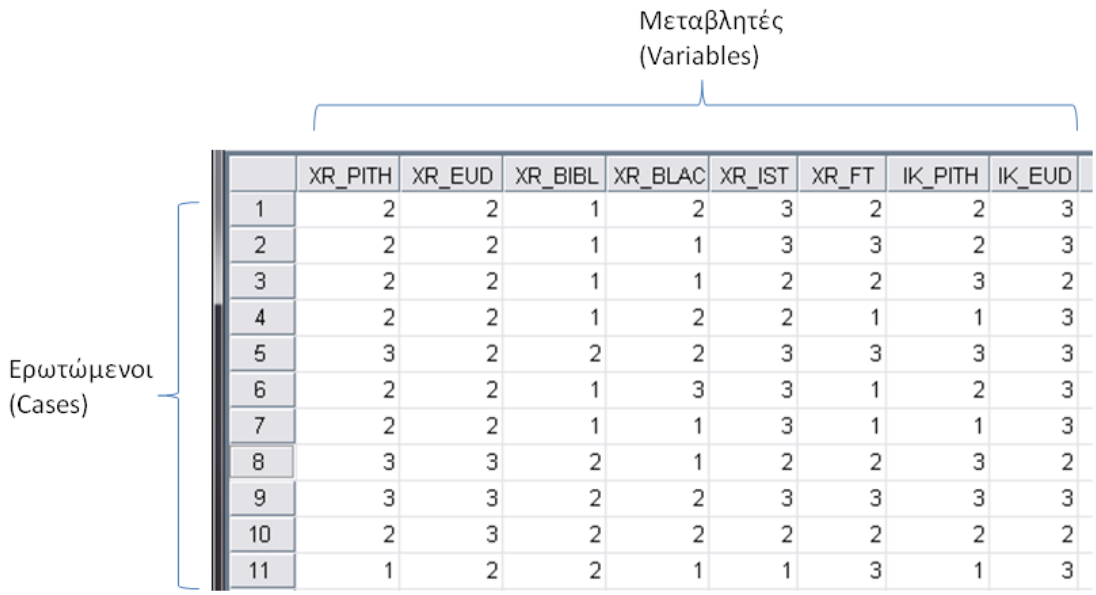
ενσωματωθεί στους πίνακες και να εκτελείται άμεσα, εισάγοντας συναρτήσεις και μαθηματικές παραστάσεις. Στο παράδειγμα του Σχήματος 2.6, σε ένα λογιστικό φύλλο τηρούνται τα λειτουργικά έξοδα μιας επιχείρησης. Κάθε γραμμή αντιστοιχεί σε μια πληρωμή και κάθε στήλη στα στοιχεία της πληρωμής αυτής (περιγραφή, ημερομηνία, ποσό, κατηγορία, κλπ.). Η στήλη της περιγραφής περιέχει κείμενα, ενώ η στήλη του ποσού περιέχει πραγματικούς αριθμούς με 2 δεκαδικά ψηφία. Το σύνολο υπολογίζεται αυτόματα ως το άθροισμα των ποσών.



	A	B	C	D	E	F
1	**Λειτουργικά έξοδα**					
2	**Α/Α**	**Ημερομηνία**	**Περιγραφή**	**Κατηγορία**	**Εξοφλήθη**	**Ποσό**
3	1	12/1/2015	ΔΕΗ	Δίκτυα	Ναι	232,00
4	2	23/2/2015	Υδρευση	Δίκτυα	Ναι	72,00
5	3	28/2/2015	Αέριο	Δίκτυα	Όχι	544,50
6	4	1/3/2015	Συνεργείο καθαρισμού	Υπηρεσίες	Όχι	40,00
7						
8	Σύνολο					888,50
9						

Σχήμα 2.6. *Αποθήκευση και επεξεργασία δεδομένων σε φύλλο δεδομένων.*

Δεδομένα έρευνας κατάλληλα για προγράμματα στατιστικής ανάλυσης. Τα δεδομένα πρωτογενών ερευνών που συλλέγονται με χρήση ερωτηματολογίων ή με κάποιον αυτοματοποιημένο μηχανισμό καταγραφής (datasets) οργανώνονται σε πίνακες. Κάθε γραμμή αντιστοιχεί σε μια περίπτωση, αντικείμενο ή ερωτώμενο και κάθε στήλη σε μια ερώτηση ή στοιχείο. Με άλλα λόγια, κάθε γραμμή είναι μια στατιστική μονάδα και κάθε στήλη μια μεταβλητή. Συνήθως η πρώτη στήλη περιέχει μια αύξουσα αρίθμηση των στατιστικών μονάδων και η πρώτη γραμμή τα ονόματα των μεταβλητών. Το περιεχόμενο κάθε κελιού είναι η τιμή της μεταβλητής της στήλης για τη στατιστική μονάδα της γραμμής. Ο τύπος δεδομένων που μπορεί να δεχθεί η κάθε στήλη καθορίζεται ανάλογα με τον τύπο της αντίστοιχης μεταβλητής π.χ. πραγματικός αριθμός για συνεχή ποσοτική μεταβλητή, ακέραιος για κωδικοποιημένη κατηγορική μεταβλητή κλπ. Αυτή είναι η βασική δομή που αναγνωρίζουν τα τυπικά προγράμματα στατιστικής ανάλυσης και με τη μορφή αυτή προβάλλουν στο χρήστη τα δεδομένα και τα αποθηκεύουν εσωτερικά στα αρχεία του δικού τους τύπου. Επίσης σε αυτήν τη μορφή μπορούν τα δεδομένα να τοποθετηθούν σε ένα γενικής χρήσης φύλλο δεδομένων (π.χ. Excel) και μέσω απλών αυτόματων μηχανισμών να εισαχθούν ή να εξαχθούν από/προς το πρόγραμμα ανάλυσης. Στο παράδειγμα του Σχήματος 3.5 παρουσιάζονται τα δεδομένα μιας έρευνας σχετικά με τη χρήση ηλεκτρονικών υπηρεσιών από τους φοιτητές, έτσι όπως έχουν εισαχθεί στο πρόγραμμα στατιστικής ανάλυσης SPSS. Κάθε γραμμή του πίνακα αντιστοιχεί σε έναν ερωτώμενο φοιτητή και περιέχει την απάντησή του σε κάθε ερώτηση. Επειδή στο παράδειγμα αυτό, οι απαντήσεις αφορούν ερωτήσεις κλειστού τύπου, τα δεδομένα είναι κατάλληλα κωδικοποιημένα σε ακέραιες αριθμητικές τιμές, έτσι ώστε κάθε τιμή να αντιστοιχεί σε συγκεκριμένη απάντηση. Κάθε στήλη αντιστοιχεί σε μια ερώτηση/μεταβλητή και έχει καθοριστεί, όπως φαίνεται στο Σχήμα 3.5 (β), ο τύπος των δεδομένων που θα περιέχει και η σημασία της κάθε αναμενόμενης τιμής.



	XR_PITH	XR_EUD	XR_BIBL	XR_BLAC	XR_IST	XR_FT	IK_PITH	IK_EUD
1	2	2	1	2	3	2	2	3
2	2	2	1	1	3	3	2	3
3	2	2	1	1	2	2	3	2
4	2	2	1	2	2	1	1	3
5	3	2	2	2	3	3	3	3
6	2	2	1	3	3	1	2	3
7	2	2	1	1	3	1	1	3
8	3	3	2	1	2	2	3	2
9	3	3	2	2	3	3	3	3
10	2	3	2	2	2	2	2	2
11	1	2	2	1	1	3	1	3

(α)

	Name	Type	Width	Deci...	Label	Values	Missing	Col...	Align	Measure
1	XR_PITH	Numeric	3	0	Χρησιμοποιείτε την ηλεκτρονική σελίδα...	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
2	XR_EUD	Numeric	3	0	Χρησιμοποιείτε το eudoxus;	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
3	XR_BIBL	Numeric	3	0	Χρησιμοποιείτε την σελίδα της βιβλιοθ...	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
4	XR_BLAC	Numeric	3	0	Χρησιμοποιείτε το blackboard;	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
5	XR_IST	Numeric	3	0	Χρησιμοποιείτε την ιστοσελίδα του τμή...	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
6	XR_FT	Numeric	3	0	Χρησιμοποιείτε τα προφίλ του τμήματο...	{1, Δεν την χρησιμοποιώ...	None	12	Right	Nominal
7	IK_PITH	Numeric	3	0	Είστε ικανοποιημένος από το Πυθία;	{1, Καθόλου}...	None	12	Right	Nominal
8	IK_EUD	Numeric	3	0	Είστε ικανοποιημένος από το eudoxus;	{1, Καθόλου}...	None	12	Right	Nominal

(β)

Σχήμα 2.7. *(α) Αποτελέσματα έρευνας κατάλληλα για στατιστική ανάλυση (β) Η οθόνη ορισμού των μεταβλητών, όπου καθορίζεται ο τύπος δεδομένων και η ερμηνεία των τιμών.*

Δεδομένα έρευνας ή δοσοληψιών κατάλληλα για ειδικά προγράμματα αναλυτικής επεξεργασίας (data analytics) και εξόρυξης πληροφορίας/γνώσης (Data mining). Τα προγράμματα αναλυτικής επεξεργασίας διαθέτουν ισχυρά και ευέλικτα εργαλεία εισαγωγής δεδομένων από οποιαδήποτε μορφή και μετασχηματισμού τους σε οποιαδήποτε άλλη μορφή εξυπηρετεί την επισκόπηση και επεξεργασία τους. Ωστόσο, η βασική δομή οργάνωσης των δεδομένων με την οποία συνήθως καταγράφονται και διακινούνται είναι αυτή των απλών πινάκων, όπως ακριβώς και στις παραπάνω περιπτώσεις. Συγκεκριμένα, η τυποποίηση που προβλέπεται από όλα σχεδόν τα πακέτα αναλυτικής επεξεργασίας καθορίζει ότι τα δεδομένα είναι τοποθετημένα σε πίνακα, όπου κάθε γραμμή αντιστοιχεί σε **μια δοσοληψία, άτομο** ή **παράδειγμα** από αυτά που μελετώνται **(transaction, case ή example)** και κάθε στήλη σε ένα **χαρακτηριστικό ή τιμή μεταβλητής (attribute)**. Στο παράδειγμα του Σχήματος 2.7, εμφανίζονται τα δεδομένα που αφορούν τους πελάτες μιας επιχείρησης, που καταγράφηκαν με σκοπό την ταξινόμηση των πελατών σε «σταθερούς» ή «με κίνδυνο απώλειας». Κάθε γραμμή αντιστοιχεί σε έναν πελάτη (case) και κάθε στήλη σε ένα χαρακτηριστικό του πελάτη (attribute) όπως ύψος αγορών, ηλικία, κλπ. Στην περίπτωση που τα δεδομένα προορίζονται για εφαρμογή εκπαιδευόμενης εκμάθησης (βλέπε κεφάλαιο 6), η τελευταία στήλη περιλαμβάνει το χαρακτηριστικό-στόχο (target attribute) δηλαδή το χαρακτηριστικό που θεωρείται εξαρτημένο από τα υπόλοιπα και του οποίου την τιμή προσπαθούμε να προβλέπουμε ή αυτό του οποίου τη συμπεριφορά θέλουμε να μοντελοποιήσουμε.

Χαρακτηριστικά (Attributes)

Χαρακτηριστικό-στόχος (Target Attribute)

Πελάτες-περιπτώσεις (Cases)

Κωδικός πελάτη	Ηλικία	Οικογενειακή κατάσταση	Αξία αγορών ανά έτος	Πλήθος αγορών ανά έτος	Παράπονα/ επιστροφές	Καθυστέρηση πληρωμών	Σταθερός Ν/Ο
17-001	21	1	1206,30	4	0	Όχι	1
17-004	44	2	25,00	1	1	Ναι	2
22-302	43	1	234,10	5	2	Όχι	1
14-101	27	1	450,20	2	0	Όχι	1

Σχήμα 2.8. *Δεδομένα πελατών για την αυτόματη κατάταξή τους σε σταθερούς ή όχι*

Η οργάνωση των δεδομένων σε λογιστικά φύλλα ή γενικότερα σε απλούς πίνακες, όπως περιγράφηκε παραπάνω, είναι συνήθης και κατάλληλη όταν τα δεδομένα έχουν την απλούστερη δομή, η οποία μπορεί να ονομαστεί «επίπεδη». Η δομή θεωρείται επίπεδη όταν δεν έχουμε περισσότερες έννοιες από αυτές της περίπτωσης/αντικειμένου και του χαρακτηριστικού/μεταβλητής, δεν έχουμε συσχετίσεις ανάμεσα σε αντικείμενα διαφορετικού τύπου, ούτε περισσότερα επίπεδα οργάνωσης, όπως π.χ. γενικότερο-ειδικότερο επίπεδο, μικρή –μεγάλη κλίμακα, σύνολα ή ιεραρχία. Αντίθετα, όταν στα δεδομένα εμπεριέχονται πολλαπλές έννοιες και σχέσεις μεταξύ τους, ο τρόπος αποθήκευσης που περιγράφηκε σε αυτήν την παράγραφο είναι ακατάλληλος και υπάρχει ανάγκη υιοθέτησης πιο σύνθετων μοντέλων.

2.6.3 Αποθήκευση σε αρχεία - δεδομένα ελεγχόμενα από προγράμματα

Ο παλαιότερος και πιο «παραδοσιακός» τρόπος αποθήκευσης δεδομένων είναι αυτός των σειριακών αρχείων, των οποίων ο χειρισμός ελέγχεται απευθείας από τα προγράμματα εφαρμογής. Τα αρχεία είναι της μορφής που περιγράφηκε στην παράγραφο 3.2, δηλαδή αποτελούνται από ένα σύνολο ομοειδών εγγραφών. Οι εγγραφές αντιστοιχούν σε αντικείμενα συγκεκριμένου τύπου π.χ. η εφαρμογή μπορεί να διαχειρίζεται ένα αρχείο «πελατών» όπου να αποθηκεύονται στη σειρά τα στοιχεία για τους πελάτες της. Αυτός ο τρόπος αποθήκευσης δεδομένων σε παραδοσιακά αρχεία θεωρείται ξεπερασμένος και δε χρησιμοποιείται σε σύγχρονα πληροφοριακά συστήματα επειδή παρουσιάζει τουλάχιστον δύο ουσιαστικές αδυναμίες: (α) είναι αναποτελεσματικός όταν υπάρχουν διαφορετικές οντότητες που σχετίζονται μεταξύ τους, κάτι που είναι το συνηθέστερο σε επιχειρηματικά δεδομένα και (β) ο χειρισμός των δεδομένων απευθείας από το πρόγραμμα εφαρμογής είναι ασύμφορος, ανασφαλής και αναποτελεσματικός και στερείται της ευελιξίας που προσφέρει ο διαχωρισμός του επιπέδου εφαρμογής από το λογικό και φυσικό επίπεδο χειρισμού δεδομένων.

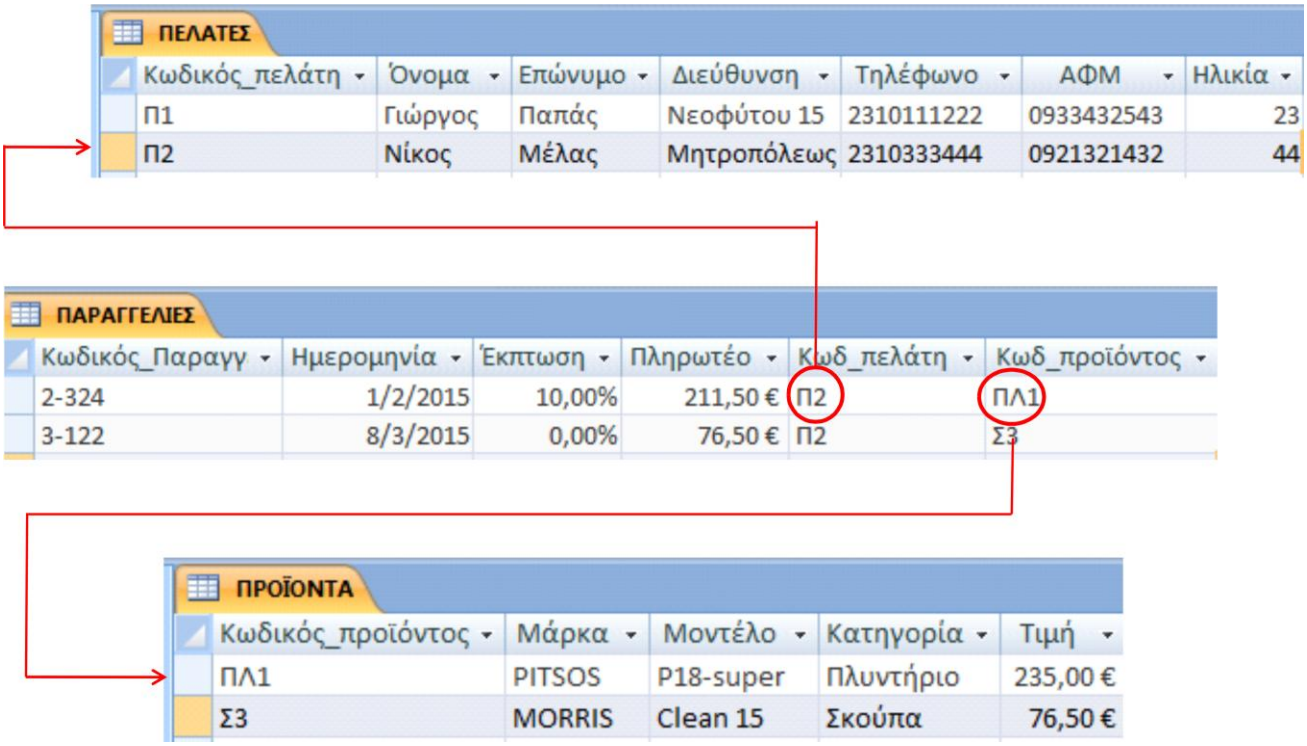
2.6.4 Βάσεις Δεδομένων

Ο συνηθέστερος και καταλληλότερος για επιχειρηματικά δεδομένα τρόπος αποθήκευσης είναι η Βάση Δεδομένων, η οποία με ειδικό λογισμικό οργανώνει τα δεδομένα σε ένα σύνολο αντικειμένων αποθήκευσης (π.χ. αρχείων ή πινάκων), έτσι ώστε να μπορούν να παρασταθούν δεδομένα με σύνθετη δομή. Η Βάση Δεδομένων διαθέτει το λεγόμενο Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) ή Data Base Management System (DBMS), που είναι υπεύθυνο για την αποθήκευση, ανάγνωση, τροποποίηση και την ασφάλεια των δεδομένων. Εκτός από το βασικό χειρισμό των δεδομένων, το ΣΔΒΔ κυρίως επιτρέπει την οργάνωση των δεδομένων ώστε να είναι αποτελεσματική η αποθήκευση, η χρήση και η συντήρησή τους. Το ΣΔΒΔ επιτρέπει τη διασύνδεση μεταξύ των προγραμμάτων εφαρμογών και των φυσικών αρχείων δεδομένων. Μια εφαρμογή (π.χ. ένα πρόγραμμα μισθοδοσίας) μπορεί να διαβάζει δεδομένα από μια Βάση Δεδομένων με τυποποιημένες εντολές ειδικά για αυτό το σκοπό (π.χ. να ζητάει την «ημερομηνία πληρωμής» και το «ποσό» του εργαζόμενου με συγκεκριμένο «όνομα») χωρίς να ενδιαφέρεται για το ποια bytes και ποιού αρχείου αντιστοιχούν σε αυτά τα δεδομένα.

Η Βάση Δεδομένων ακολουθεί κάποιο «μοντέλο», δηλαδή κάποιον τρόπο οργάνωσης των δεδομένων. Τα βασικότερα μοντέλα ΒΔ είναι:

1. Σχεσιακό μοντέλο δεδομένων: λογικό μοντέλο που χειρίζεται τα δεδομένα σαν πίνακες και σχέσεις.
2. Ιεραρχικά: τα δεδομένα είναι οργανωμένα σε δομή τύπου δένδρου.
3. Δικτυακά: λογικό μοντέλο που χρησιμεύει στην αποτύπωση πολυσήμαντων σχέσεων.

Η έννοια του μοντέλου δεδομένων και ειδικότερα το συνηθέστερο μοντέλο Βάσεων Δεδομένων, που είναι το Σχεσιακό, θα αναπτυχθούν στο Κεφάλαιο 3, ενώ στο Κεφάλαιο 4 θα παρουσιαστεί ο τρόπος σχεδιασμού και υλοποίησης μιας Βάσης Δεδομένων. Στο παράδειγμα του Σχήματος 2.9 φαίνεται η δυνατότητα των Βάσεων Δεδομένων να διαχειρίζονται όχι μόνο χαρακτηριστικά αντικειμένων όπως π.χ. τα στοιχεία των πελατών, αλλά και σχέσεις ανάμεσα σε εγγραφές διαφορετικού τύπου, όπως το ποιος πελάτης έδωσε μια παραγγελία και το ποιο προϊόν περιλαμβάνεται σε μια παραγγελία.



Σχήμα 2.9. *Η Βάση Δεδομένων είναι απαραίτητη όταν τα δεδομένα περιλαμβάνουν και σχέσεις ανάμεσα σε διαφορετικού τύπου αντικείμενα, πρόσωπα ή γεγονότα.*

Τα πλεονεκτήματα του χειρισμού των δεδομένων με χρήση Βάσης Δεδομένων, αντί των παλαιότερων μεμονωμένων αρχείων ή των πιο σύγχρονων φύλλων δεδομένων είναι τα ακόλουθα:

- Παρέχεται η δυνατότητα διαχείρισης όλων των δεδομένων που αφορούν ένα θέμα, ακόμα και αν αυτά είναι ποικίλων ειδών, σε έναν ενιαίο χώρο. Τα δεδομένα ελέγχονται κεντρικά και πιο αποτελεσματικά και είναι δυνατή η διασύνδεση μεταξύ τους.
- Περιορισμός πλεονασμού δεδομένων (data redundancy).
- Περιορισμός αντιφατικότητας (inconsistency) δεδομένων.
- Περιορισμός της πολυπλοκότητας του πληροφοριακού συστήματος και μείωση του κόστους ανάπτυξης και συντήρησής του.
- Καλύτερος κεντρικός έλεγχος της δημιουργίας και του προσδιορισμού των δεδομένων.
- Καλύτερη πρόσβαση και διαθεσιμότητα πληροφορίας.
- Περιορισμός της εξάρτησης προγραμμάτων εφαρμογών και δεδομένων.
- Αύξηση ευελιξίας συστήματος.

Τα παραπάνω πλεονεκτήματα είναι πολύ ισχυρά, ώστε η χρήση Βάσης Δεδομένων να θεωρείται επιβεβλημένη για οποιαδήποτε δεδομένα διαθέτουν δομή πέρα από την υποτυπώδη «επίπεδη» μορφή.

2.6.5 Μοντέλα άμεσης αναλυτικής επεξεργασίας και άλλα εξειδικευμένα μοντέλα

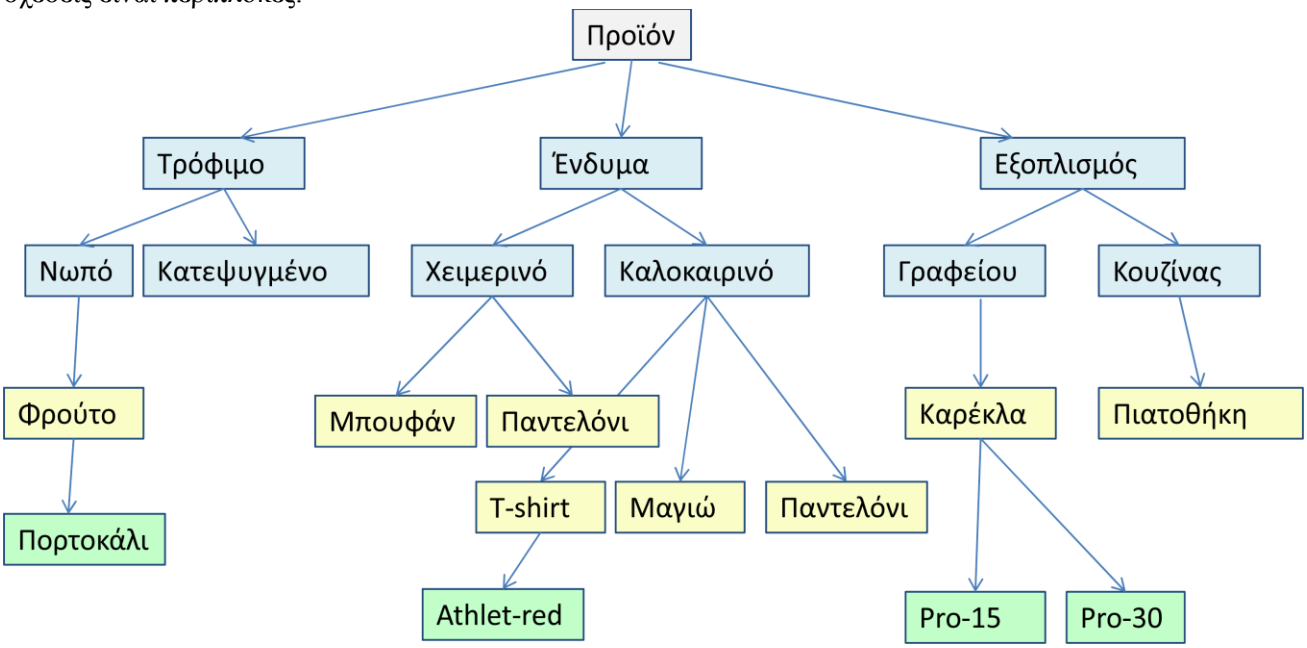
Σε ορισμένες περιπτώσεις, οι ιδιαίτερες ανάγκες κάποιου προβλήματος δεν μπορούν να καλυφθούν από κάποιον από τους βασικούς τρόπους οργάνωσης δεδομένων και εφαρμόζονται εξειδικευμένες λύσεις που είναι προσαρμοσμένες στις ανάγκες αυτές. Ο χειρισμός των δεδομένων που βρίσκονται σε τέτοιες μορφές γίνεται από ειδικό λογισμικό, είτε στο επίπεδο εφαρμογής (δηλαδή η ίδια η εφαρμογή γνωρίζει το νόημα και την οργάνωση των δεδομένων ώστε να μπορεί να τα χειριστεί), είτε σε κατάλληλο λογικό επίπεδο (δηλαδή υπάρχει ειδικό Σύστημα Διαχείρισης Βάσεων Δεδομένων που αναλαμβάνει να χειριστεί τα δεδομένα και να τα παρέχει προς επεξεργασία στην εφαρμογή).

Στην κατηγορία αυτή ανήκουν οι Βάσεις **Άμεσης Αναλυτικής Επεξεργασίας (OnLine Analytical Processing Databases)**, γνωστές με το ακρωνύμιο **OLAP**, που προορίζονται για την αποτελεσματικότερη επεξεργασία και ανάλυση των δεδομένων. Οι Βάσεις OLAP υποστηρίζουν την προβολή της πληροφορίας με τη βοήθεια σχημάτων πολλών διαστάσεων, που ονομάζονται κύβοι (cubes) και αποτελούν σημαντικό εργαλείο επιχειρηματικής ευφυΐας, που θα παρουσιαστεί στο Κεφάλαιο 5.

Στο χώρο της διαχείρισης δεδομένων, υπάρχουν και άλλα εξειδικευμένα μοντέλα δεδομένων, που είναι εκτός των σκοπών του βιβλίου αυτού, αλλά για να έχει ο αναγνώστης πληρέστερη εικόνα, αναφέρονται τα παρακάτω δύο χαρακτηριστικά παραδείγματα:

Ιεραρχικά δεδομένα. Τέτοιου είδους δεδομένα έχουμε όταν υπάρχει η έννοια του ανώτερου/κατώτερου ή του γενικότερου/ειδικότερου. Τα ιεραρχικά δεδομένα μπορούν να οργανωθούν σε μορφή δέντρου, όπου κάθε εγγραφή μπορεί να συνδέεται με μια εγγραφή ανώτερου επιπέδου (π.χ. την κατηγορία στην οποία ανήκει) και πολλές εγγραφές κατώτερου επιπέδου (π.χ. τις υποκατηγορίες που περιλαμβάνει). Στο παράδειγμα του Σχήματος 2.10 εμφανίζονται τα ιεραρχικά δεδομένα που αφορούν την κατηγοριοποίηση των προϊόντων μιας εταιρείας.

Γράφοι. Χρησιμοποιούνται για να περιγράψουν τις σχέσεις ανάμεσα σε αντικείμενα του πραγματικού κόσμου και αποτελούνται από κόμβους, που παριστάνουν τα αντικείμενα, και βέλη ή ευθύγραμμα τμήματα που παριστάνουν τις σχέσεις. Οι γράφοι είναι γενικά και ευέλικτα μοντέλα, που μπορούν να παραστήσουν είτε κατευθυνόμενες σχέσεις, είτε αμφίπλευρες και είναι ιδιαίτερα χρήσιμοι όταν οι σχέσεις είναι περίπλοκες.



Σχήμα 2.10. *Παράδειγμα ιεραρχικών δεδομένων που αφορούν την κατηγοριοποίηση των προϊόντων μιας εταιρείας.*

2.7 Αναπαράσταση πληροφορίας και γνώσης

Στην προηγούμενη ενότητα περιγράφηκαν οι τρόποι με τους οποίους αναπαριστώνται τα δεδομένα σε ένα πληροφοριακό σύστημα, έτσι ώστε να είναι αποδοτική η διαχείρισή τους. Αναδείχθηκε το γεγονός ότι οι μηχανισμοί διαχείρισης των δεδομένων είναι ένα σύνθετο οικοδόμημα που ξεκινάει από την κωδικοποίηση

αριθμών, κειμένου κλπ. ως δυαδικά ψηφία και ακολουθείται από την οργάνωσή τους σε σύνθετες δομές που αντανακλούν το νόημα που έχουν τα δεδομένα αυτά στον πραγματικό κόσμο. Παρόλο που τα υψηλότερα επίπεδα από αυτά που αναφέρθηκαν (ουσιαστικά οι Βάσεις Δεδομένων) ενσωματώνουν τη λογική και το νόημα των δεδομένων, σκοπός τους παραμένει η διαχείριση των δεδομένων ως ακατανόητα Bytes και όχι με βάση τη σημασία ή τη χρησιμότητά τους. Η διαχείριση πληροφορίας, δηλαδή η οργάνωσή της σε ένα σύστημα πληροφορικής με βάση τη σημασία της, απαιτεί επιπλέον μηχανισμούς αναπαράστασης.

Η οργάνωση της πληροφορίας επιτυγχάνεται με την κατασκευή κατάλληλων μοντέλων πληροφορίας (information models) που περιγράφουν το νόημα και τη χρησιμότητα των δεδομένων. Τα μοντέλα πληροφορίας εστιάζουν στον ορισμό των εννοιών και στις σχέσεις ανάμεσα στα αντικείμενα, χωρίς να υπεισέρχονται σε θέματα σχετικά με το πώς θα αποθηκευτούν ή θα ανασυρθούν τα σχετικά δεδομένα. Επίσης μπορούμε να πούμε ότι εστιάζουν στην περιγραφή της πληροφορίας από την οπτική γωνία του λογικού προβλήματος και κρύβουν τις τεχνικές λεπτομέρειες που αφορούν το πώς η πληροφορία αυτή θα μεταφραστεί σε δεδομένα. Ένα μοντέλο πληροφορίας μπορεί να περιγράφει τις έννοιες που συμπεριλαμβάνονται σε μια εφαρμογή, μαζί με τη λογική με την οποία συνδέονται μεταξύ τους και τις ενέργειες που αφορούν την καθεμιά. Επίσης μπορεί να περιγράφει μια δομή δεδομένων (π.χ. ποια στοιχεία περιλαμβάνει μια μισθοδοτική κατάσταση) ή μια επιχειρηματική διαδικασία (π.χ. με ποια βήματα εξυπηρετείται μια παραγγελία).

Η αναπαράσταση της πληροφορίας γίνεται με χρήση εργαλείων υψηλότερου επιπέδου από αυτά της αναπαράστασης δεδομένων. Σημαντικά τέτοια εργαλεία για την κατασκευή μοντέλων πληροφορίας είναι τα μετα-δεδομένα (Metadata), καθώς και κατάλληλα διαγράμματα σε τυποποιημένες γλώσσες μοντελοποίησης, με κυριότερο το διάγραμμα τάξεων (Class diagram), που περιλαμβάνεται στη γλώσσα UML (Unified Modeling Language).

2.7.1 Μετα-δεδομένα (Metadata)

Ο ορισμός των μετα-δεδομένων είναι **δεδομένα σχετικά με τα δεδομένα** (Guenther & Radebaugh, 2004) και χρησιμοποιούνται για να καταγράψουν πληροφορία που αφορά κάποιο αντικείμενο που χειριζόμαστε ως δεδομένο. Για παράδειγμα, αν σε μια Βάση Δεδομένων περιλαμβάνονται δεδομένα σχετικά με προϊόντα μιας εταιρείας, περιμένουμε τα δεδομένα αυτά να είναι οργανωμένα σε πίνακες με πεδία όπως «όνομα προϊόντος», «τιμή», «περιθώριο κέρδους» κλπ. Τα μετα-δεδομένα που περιγράφουν αυτά τα δεδομένα μπορεί να είναι στοιχεία όπως: η γλώσσα στην οποία εκφράζεται το όνομα προϊόντος, το πώς εννοούμε την τιμή (π.χ. τιμή ραφιού, χονδρική, με μεταφορικά, κλπ.), ή το πώς υπολογίζεται το περιθώριο κέρδους. Ένα δεύτερο παράδειγμα είναι μια Βάση Δεδομένων με επιστημονικά άρθρα. Ενώ τα δεδομένα είναι τα ίδια τα άρθρα, τα μετα-δεδομένα είναι στοιχεία για αυτά, όπως ο συγγραφέας, η επιστημονική περιοχή και το έτος έκδοσης. Τα μετα-δεδομένα χρησιμοποιούνται λοιπόν για να περιγράψουν πληροφοριακούς πόρους, επιτρέποντας την κατηγοριοποίησή τους, την αποθήκευση, αναζήτηση και σωστή ανάσυρση και χρήση τους.

Τα μετα-δεδομένα μπορεί να είναι περιγραφικά, δηλαδή να περιγράφουν ένα αντικείμενο ώστε να το αναγνωρίσουμε και να το ανακτήσουμε ή δομικά, που περιγράφουν την οργάνωση σύνθετων αντικειμένων και τη χρήση των συστατικών τους. Στο χώρο των επιχειρησιακών δεδομένων, τα μετα-δεδομένα περιλαμβάνουν βοηθητικές πληροφορίες για την προέλευση, το νόημα και τη χρήση των κύριων δεδομένων, όπως για παράδειγμα πότε πραγματοποιήθηκε μια εγγραφή ή τι περιορισμοί ασφαλείας ισχύουν για κάποια στοιχεία.

Διευκρινίζεται ότι τα ίδια τα μετα-δεδομένα, ως περιεχόμενο, είναι στην πραγματικότητα και αυτά δεδομένα, ενώ μοντέλο πληροφορίας είναι το λεγόμενο «σχήμα» των μετα-δεδομένων, δηλαδή ο σχεδιασμός του ποια μετα-δεδομένα περιγράφουν ποια δεδομένα και ποιο είναι το νόημά τους. Ένα μοντέλο μετα-δεδομένων (metadata model) μπορεί να οριστεί ως μια δομημένη περιγραφή των χαρακτηριστικών και των ιδιοτήτων ενός τύπου πληροφορίας. Η περιγραφή αυτή επιτρέπει τον ακριβέστερο προσδιορισμό ανομοιόμορφων πηγών πληροφόρησης, τη δημιουργία καταλόγων αναζήτησης που λειτουργούν σωστά σε πολλές διαφορετικές πηγές, τη σύνδεση ανομοιόμορφων δεδομένων και την αποτελεσματικότερη αυτοματοποιημένη αναζήτηση πληροφοριακών πόρων. Ένας μεγάλος αριθμός από μοντέλα μετα-δεδομένων έχει προταθεί από διεθνείς οργανισμούς τυποποίησης, ώστε να είναι δυνατή η κοινή χρήση περιεχομένου σε συγκεκριμένα πεδία. Παράδειγμα ενός τέτοιου προτύπου μετα-δεδομένων είναι το Dublin Core (DublinCore, 2015). Το Dublin Core αποτελεί ένα διεθνές πρότυπο γενικού σκοπού για την περιγραφή ψηφιακών τεκμηρίων, το οποίο έχει βρει ευρεία αποδοχή επειδή τα στοιχεία του είναι αρκετά απλά και γενικά, ώστε να εφαρμόζουν σε ένα μεγάλο πεδίο εφαρμογών. Χρησιμοποιεί 15 στοιχεία (elements) προτυποποιημένων μετα-

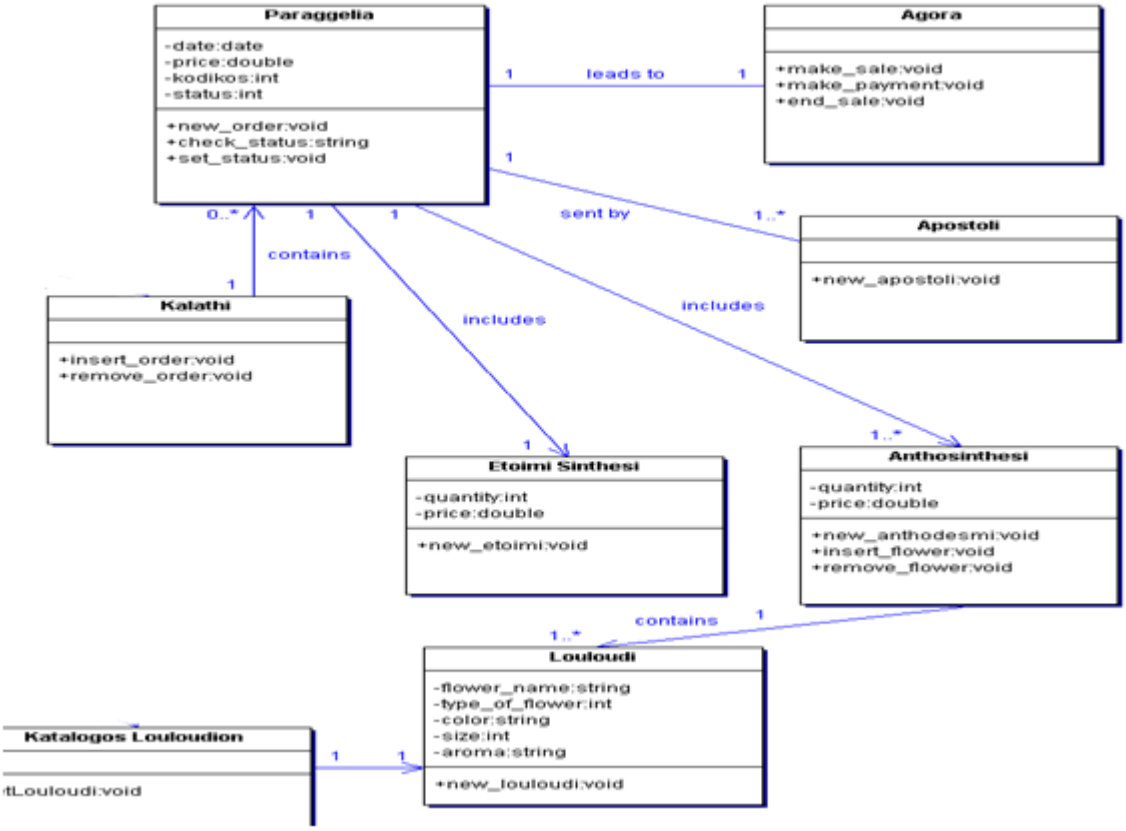
δεδομένων για την περιγραφή ψηφιακών αντικειμένων που σχετίζονται με βιβλιογραφικά τεκμήρια (άρθρα, βιβλία, κλπ.), με σκοπό τον εύκολο εντοπισμό και ανάκτησή τους. Στον Πίνακα 3.3 παρουσιάζεται ένα απόσπασμα του Dublin Core.

Α/Α	**ΣΤΟΙΧΕΙΟ (ELEMENT)**	**ΠΕΡΙΓΡΑΦΗ**	**ΥΠΟΧΡΕΩ-ΤΙΚΟ**	**ΠΟΛΛΑΠΛΟΤΗΤΑ**
1	Τίτλος (Title)	*Το όνομα που δίνεται στο τεκμήριο, συνήθως από τον δημιουργό ή τον εκδότη.*	ΝΑΙ	Μία τιμή
2	Περιγραφή (Description)	*Μία αυτούσια περιγραφή του περιεχομένου του τεκμηρίου*	ΝΑΙ	Μία τιμή
3	Λέξεις-κλειδιά (Keywords)	*Το θέμα αυτού του τεκμηρίου εκφρασμένο σαν λέξεις-κλειδιά ή φράσεις που περιγράφουν το θέμα ή το περιεχόμενο του τεκμηρίου.*	ΝΑΙ	Πολλές τιμές
4	Γλώσσα (Language)	*Η γλώσσα του περιεχομένου του τεκμηρίου.*	ΝΑΙ	Πολλές τιμές
5	Πηγή (Source)	*Αναφορά σε μία πηγή από την οποία το παρόν τεκμήριο απορρέει.*	ΟΧΙ	Πολλές τιμές
6	Σχέση (Relation)	*Ένα αναγνωριστικό ενός δεύτερου πόρου και η σχέση του με το παρόν τεκμήριο. Αυτό το στοιχείο επιτρέπει τη δήλωση συνδέσμων μεταξύ σχετικών πόρων.*	ΟΧΙ	Πολλές τιμές
	. . .			

Πίνακας 2.3. *Απόσπασμα του μοντέλου μετα-δεδομένων Dublin Core που χρησιμοποιείται για την περιγραφή βιβλιογραφικών τεκμηρίων.*

2.7.2 Διαγράμματα τάξεων (class diagram)

Οι έννοιες που περιλαμβάνονται στα μοντέλα πληροφορίας μπορούν να περιγραφούν από τάξεις (classes), σχέσεις (relationships) και ιδιότητες (properties), τα οποία μπορούν να αποτυπωθούν σε ένα τυποποιημένο διάγραμμα που λέγεται διάγραμμα τάξεων (class diagram). Η τάξη είναι ένα πρότυπο που καθορίζει τα χαρακτηριστικά μιας συγκεκριμένης κατηγορίας αντικειμένων για τα οποία τηρούνται δεδομένα σε μια εφαρμογή και που συμμετέχει σε κάποιες διαδικασίες. Η τάξη αντιπροσωπεύει μια οντότητα σε ένα μοντέλο πληροφορίας, π.χ. μια παραγγελία και οι μέθοδοι εκτέλεσής της. Μια σχέση είναι ένα πρότυπο του τρόπου με τον οποίο συνδέονται μεταξύ τους τάξεις και καθορίζει κριτήρια για το πώς μπορούν να συνδεθούν δύο αντικείμενα. Η ιδιότητα είναι ένα πρότυπο με το οποίο πρέπει να συμμορφώνονται οι τιμές που αποθηκεύονται και αφορούν κάποιο χαρακτηριστικό μιας οντότητας, π.χ. η τιμή του προϊόντος είναι ένας πραγματικός αριθμός με 2 δεκαδικά ψηφία που δεν μπορεί να είναι αρνητικός. Στο παράδειγμα του Σχήματος 2.11 παρουσιάζεται ένα απόσπασμα από το διάγραμμα τάξεων που περιγράφει το σχεδιασμό ενός ηλεκτρονικού ανθοπωλείου. Φαίνεται π.χ. ότι το «Καλάθι» μπορεί να περιλαμβάνει καμία, μία ή περισσότερες «Παραγγελίες», μία «Παραγγελία» οδηγεί σε μία «Αγορά», ενώ η «Παραγγελία» μπορεί να περιλαμβάνει μια ή περισσότερες «Ανθοσυνθέσεις», για τις οποίες πρέπει να γνωρίζουμε την ποσότητα και την τιμή.



Σχήμα 2.11. *Απόσπασμα του διαγράμματος τάξεων ενός ηλεκτρονικού ανθοπωλείου.*

2.7.3 Από την πληροφορία στη γνώση

Η γνώση παράγεται από την πληροφορία με ειδικές μεθόδους ανάλυσης και «εξαγωγής γνώσης». Είναι κάτι που εμπεριέχει ευφυΐα, δηλαδή τρόπο σκέψης και ικανότητα λήψης αποφάσεων και επίλυσης προβλημάτων. Για την παραγωγή γνώσης από πληροφορία απαιτείται επιλογή της κατάλληλης πληροφορίας, δόμηση της πληροφορίας ώστε να αντιστοιχεί σε κάποιο μοντέλο και γενίκευση της πληροφορίας. Γενίκευση είναι όταν από έναν αριθμό περιπτώσεων (π.χ. τις απαντήσεις σε μια έρευνα) εξάγεται ένας γενικός κανόνας που ισχύει σε όλες τις αντίστοιχες περιπτώσεις (π.χ. η αποδοχή ενός προϊόντος από μια μερίδα καταναλωτών).

Όπως θα μπορεί να αντιληφθεί ο αναγνώστης, η γνώση δεν είναι δυνατό να αποθηκευτεί σε πίνακες όπως στην περίπτωση των δεδομένων (π.χ. τιμή προϊόντος = 12,30), ούτε ως πληροφορία (π.χ. η τιμή προϊόντος είναι χωρίς ΦΠΑ και μεταφορικά), αλλά απαιτούνται ειδικοί τρόποι παράστασης που ονομάζονται Μοντέλα Γνώσης. Για παράδειγμα, με τη βοήθεια ενός μοντέλου κανόνων (Rule-Based model), που αποτελεί ειδικό τύπο μοντέλου γνώσης, θα μπορούμε να καταγράψουμε ότι «αν η τιμή του προϊόντος αυξηθεί κατά 2€, προβλέπεται μείωση των πωλήσεων κατά 10%». Εννοείται ότι ο κανόνας αυτός θα πρέπει να εισαχθεί σε ένα πληροφοριακό σύστημα με τρόπο που να τον κατανοεί το σύστημα και να μπορεί να τον χρησιμοποιήσει για τον αυτόματο υπολογισμό προβλέψεων και όχι σε μορφή ακατανόητου από το σύστημα κειμένου. Η μοντελοποίηση γνώσης είναι σημαντικό συστατικό των σύγχρονων ευφυών συστημάτων και θα μας απασχολήσει εκτενώς στο κεφάλαιο 8 του βιβλίου.

Βιβλιογραφία/Αναφορές

American Standard Code for Information Interchange (1963). ASA X3.4-1963, American Standards Association, June 17, 1963.

Dublin Core.org. Dublin Core Metadata initiative. Web. May 2015.

Guenther, R. & Radebaugh, J. (2004). *Understanding Metadata*. Bethesda, MD: National Information Standards Organization Press. ISBN 1-880124-62-9.

Κεφάλαιο 3. Το σχεσιακό μοντέλο Βάσεων Δεδομένων

Σύνοψη
Το σχεσιακό μοντέλο είναι το συνηθέστερο και πιο ευρέως διαδεδομένο μοντέλο διαχείρισης δεδομένων, που βασίζεται στην οργάνωση των δεδομένων σε πίνακες που συσχετίζονται μεταξύ τους μέσω κοινών χαρακτηριστικών. Το κεφάλαιο αυτό παρουσιάζει τις βασικές έννοιες του σχεσιακού μοντέλου ώστε ο αναγνώστης να κατανοήσει τη φιλοσοφία οργάνωσης των δεδομένων και τα βασικά στοιχεία του μοντέλου, όπως οι πίνακες, οι συνδέσεις και τα κλειδιά Παρουσιάζεται το Διάγραμμα Οντοτήτων-Συσχετίσεων, ως ένα εργαλείο αποτύπωσης των αναγκών μιας Βάσης Δεδομένων και αναλύεται βήμα προς βήμα η διαδικασία σχεδίασης. Γίνεται επίσης αναφορά σε θέματα κανονικοποίησης πινάκων, πλεονασμού δεδομένων και εγκυρότητας.

Προαπαιτούμενη γνώση
Κεφάλαιο 2. Δεδομένα και Πληροφορίες.

3.1 Γενικά για τη μοντελοποίηση δεδομένων

Η Βάση Δεδομένων είναι μια συλλογή πληροφοριών που σχετίζονται με ένα συγκεκριμένο θέμα ή σκοπό π.χ. την παρακολούθηση των παραγγελιών των πελατών ή την οργάνωση των στοιχείων μιας έρευνας αγοράς (Date, 1981). Μια Βάση Δεδομένων πρέπει να είναι προσεκτικά σχεδιασμένη, ακολουθώντας μερικές βασικές αρχές και κανόνες, ώστε τα δεδομένα που περιέχει να είναι σωστά οργανωμένα και δομημένα (Hawryszkiewycz, 1991). Με τον τρόπο αυτό περιορίζονται στο ελάχιστο οι περιττές επαναλήψεις, οι κίνδυνοι λαθών και αυξάνεται η ασφάλεια των δεδομένων απέναντι σε αλλοιώσεις. Επίσης, πρωταρχικό πλεονέκτημα της σωστής οργάνωσης των δεδομένων είναι η δυνατότητα να αναζητούμε γρήγορα και αποτελεσματικά τα δεδομένα που χρειαζόμαστε και να βρίσκουμε με ακρίβεια και χωρίς απροσδιοριστία τα σωστά στοιχεία. Τέλος, εκτός από απλές αναζητήσεις, όπως π.χ. να βρούμε τα στοιχεία ενός πελάτη, μια καλά οργανωμένη Βάση Δεδομένων μας δίνει εργαλεία σύνθετων αναζητήσεων, όπως π.χ. η επιλογή των πελατών που κάνουν τακτικά αγορές αυτών των προϊόντων που είναι για την επιχείρηση τα πιο επικερδή.

Η οργάνωση των δεδομένων είναι δυνατόν να πραγματοποιηθεί με πολλούς διαφορετικούς τρόπους, ανάλογα με τη φύση τους και κυρίως σύμφωνα με την λογική με την οποία θα αξιοποιηθούν κατά τη χρήση τους. Το σύστημα που εφαρμόζουμε για την οργάνωση των δεδομένων με τυποποιημένο και αποτελεσματικό τρόπο, ονομάζεται **μοντέλο δεδομένων**. Τα κρίσιμα στοιχεία του μοντέλου αυτού είναι:

- το πώς αποθηκεύονται τα δεδομένα με ορθότητα, ασφάλεια και οικονομία
- το πώς ανασύρουμε αυτό που χρειαζόμαστε σε μορφή χρήσιμης πληροφορίας και
- το πώς επικαιροποιούμε τα δεδομένα χωρίς κίνδυνο απώλειας.

Με άλλα λόγια, μπορούμε να πούμε ότι μοντέλο δεδομένων είναι ένα σύνολο από **κανόνες οργάνωσης** που καθορίζουν τη δομή της Βάσης Δεδομένων, ένα σύνολο από **πράξεις** για τη διαχείριση των δεδομένων και ένα σύνολο από **περιορισμούς** που εξασφαλίζουν την ορθότητά της.

Μοντελοποίηση δεδομένων ονομάζουμε τη διαδικασία καθορισμού των αναγκών σχετικά με την αποθήκευση και διαχείριση δεδομένων, ώστε να υποστηρίζονται οι επιχειρηματικές διαδικασίες ενός οργανισμού. Η μοντελοποίηση δεδομένων πρέπει να γίνεται ακολουθώντας συγκεκριμένη τυπική μεθοδολογία, ώστε το αποτέλεσμα να είναι τυποποιημένο, συνεπές και εύκολα αξιοποιήσιμο στην υλοποίηση του πληροφοριακού συστήματος που θα αναλάβει το χειρισμό των δεδομένων. Το μοντέλο που προκύπτει ως αποτέλεσμα της διαδικασίας είναι οι προδιαγραφές αλλά και ένας σχεδιασμός για την υλοποίηση της κατάλληλης Βάσης Δεδομένων.

Στη μοντελοποίηση δεδομένων οφείλει να συμμετέχει ο χρήστης των δεδομένων, δηλαδή κάποιο στέλεχος της επιχείρησης που γνωρίζει τον τρόπο λειτουργίας της και μπορεί να υποδείξει τις πληροφοριακές ανάγκες της επιχείρησης που προκύπτουν από τις διαδικασίες της, το περιβάλλον στο οποίο λειτουργεί και τους κανόνες της. Οι προδιαγραφές που θα θέσει το στέλεχος της επιχείρησης θα μεταφραστούν στη συνέχεια

σε ένα ορθά σχεδιασμένο σύστημα διαχείρισης δεδομένων. Είναι λοιπόν σημαντικό για ένα στέλεχος επιχείρησης που εργάζεται με σύγχρονα συστήματα πληροφορικής να γνωρίζει τις βασικές αρχές μοντελοποίησης δεδομένων, ακόμα και αν δεν πρόκειται να υλοποιήσει ο ίδιος την εφαρμογή που χρειάζεται.

Ένα γνώρισμα της μοντελοποίησης δεδομένων είναι ότι πρόκειται για προοδευτική και επαναληπτική διαδικασία. Ξεκινάει συνήθως από ένα **εννοιολογικό μοντέλο**, που αποτελεί αφηρημένη καταγραφή των αναγκών σε δεδομένα, χωρίς να αναφέρεται σε συγκεκριμένο τρόπο υλοποίησης ή τεχνολογία. Στην συνέχεια δημιουργείται ένα **λογικό μοντέλο** που περιέχει συγκεκριμένες δομές δεδομένων που μπορούν να υλοποιηθούν ως βάση δεδομένων. Τέλος, το μοντέλο μετεξελίσσεται σε **φυσικό μοντέλο** δεδομένων, που περιέχει λύσεις και περιορισμούς σε μεγάλο επίπεδο λεπτομέρειας. Η διαδικασία χαρακτηρίζεται ως επαναληπτική επειδή πολύ συχνά απαιτείται επαναπροσδιορισμός των προδιαγραφών, λόγω αλλαγών στο περιβάλλον της επιχείρησης ή εντοπισμού προβλημάτων στη λειτουργία και κατά συνέπεια ακολουθεί επανασχεδιασμός της Βάσης.

Στα πλαίσια του παρόντος βιβλίου θα περιοριστούμε στην παρουσίαση μιας σχετικά απλής και εύκολα εφαρμόσιμης διαδικασίας μοντελοποίησης δεδομένων, σχεδιασμού και υλοποίησης μιας Βάσης Δεδομένων, ώστε ο αναγνώστης, χωρίς ειδικές γνώσεις πληροφορικής, να μπορεί:

- να καταγράψει τις ανάγκες του για μια μικρής έκτασης εφαρμογή διαχείρισης δεδομένων,
- να αποτυπώσει με ακρίβεια το πρόβλημα του πραγματικού κόσμου σε τυποποιημένη μορφή ώστε να είναι κατανοητό και επιλύσιμο,
- να μετατρέψει τις προδιαγραφές σε σχεδιαστική λύση, ακολουθώντας τυποποιημένη διαδικασία που εγγυάται την ορθότητα του αποτελέσματος
- να υλοποιήσει τη λύση με χρήση λογισμικού βάσεων δεδομένων (MS-Access).

Για το σκοπό αυτό, στα επόμενα υποκεφάλαια παρουσιάζεται το σχεσιακό μοντέλο ως το συνηθέστερο και πρακτικότερο μοντέλο δεδομένων, το διάγραμμα οντοτήτων-συσχετίσεων (στην απλούστερη δυνατή μορφή του), ως ένα εργαλείο μοντελοποίησης του προβλήματος και καταγραφής των προδιαγραφών με τυποποιημένο τρόπο και επίσης μια τυποποιημένη και πρακτική διαδικασία σχεδιασμού μιας Βάσης Δεδομένων που βασίζεται στο διάγραμμα οντοτήτων-συσχετίσεων και καταλήγει στην τελική υλοποίηση της Βάσης.

3.2 Οι βασικές έννοιες του σχεσιακού μοντέλου δεδομένων

3.2.1 Τι είναι το σχεσιακό μοντέλο

.

Το συνηθέστερο και πιο ευρέως διαδεδομένο μοντέλο διαχείρισης δεδομένων είναι το **Σχεσιακό Μοντέλο** (Κεχρής, 2015), που προτάθηκε κατά τη δεκαετία του 70 και συναντιέται σήμερα στις περισσότερες Βάσεις Δεδομένων και σχεδόν σε όλες όσες αφορούν τις εφαρμογές που ενδιαφέρουν μια επιχείρηση. Οι Βάσεις Δεδομένων που ακολουθούν το μοντέλο αυτό ονομάζονται **Σχεσιακές Βάσεις Δεδομένων.** Το σχεσιακό μοντέλο έχει απλή δομή και είναι αρκετά ευέλικτο ώστε να μπορεί να χρησιμοποιηθεί στις περισσότερες εφαρμογές για τις οποίες δεν υπάρχουν ειδικές απαιτήσεις (Παπαθανασίου, 2008).

Η βασική αρχή μιας Σχεσιακής Βάσης Δεδομένων είναι η οργάνωση των δεδομένων σε πίνακες, οι οποίοι συνδέονται μεταξύ τους μέσω κοινών χαρακτηριστικών. Σύμφωνα με την αρχή αυτή, τα δεδομένα διαιρούνται σε ξεχωριστούς χώρους αποθήκευσης, οι οποίοι λέγονται πίνακες. Ο κάθε πίνακας περιλαμβάνει δεδομένα που αφορούν μόνο μία έννοια, για παράδειγμα, τα στοιχεία ενός πελάτη, όπως όνομα και τηλέφωνο, θα πρέπει να τοποθετηθούν σε έναν πίνακα που αντιστοιχεί στην έννοια «Πελάτες», ενώ τα στοιχεία των παραγγελιών που δίνει (π.χ. ημερομηνία παραγγελίας και κόστος) θα πρέπει να τοποθετηθούν σε άλλον πίνακα στον οποίο καταγράφονται οι «Παραγγελίες». Οι δύο αυτοί πίνακες συνδέονται μεταξύ τους έτσι ώστε να γνωρίζουμε ποιος πελάτης έδωσε ποια παραγγελία. Στη συνέχεια, όταν ανασύρουμε κάποια δεδομένα, όπως για παράδειγμα τα ονόματα των πελατών που έδωσαν παραγγελίες σήμερα, ο κατάλληλος μηχανισμός της Βάσης Δεδομένων καλείται να συνδυάσει τους δύο πίνακες και να συνθέσει το αποτέλεσμα, το οποίο θα περιλαμβάνει τόσο στοιχεία από τον πίνακα των πελατών, όσο και από τον πίνακα των παραγγελιών. Σύμφωνα με το σχεσιακό μοντέλο, δεν επιτρέπεται να αποθηκευτούν στον ίδιο χώρο αποθήκευσης (δηλαδή πίνακα) δεδομένα που αφορούν διαφορετική έννοια. Εκτός από τις ελάχιστες

περιπτώσεις όπου τα δεδομένα είναι πολύ απλά σε δομή και δεν περιέχουν συσχετίσεις ανάμεσα σε διαφορετικές έννοιες, συνήθως επιβάλλεται:

- Οργάνωση των δεδομένων σε μια δομή που περιλαμβάνει πολλούς διαφορετικούς πίνακες και συσχετίσεις μεταξύ των πινάκων.
- Κατά την αποθήκευση, γίνεται διαχωρισμός των δεδομένων στους κατάλληλους πίνακες.
- Κατά την ανάσυρση, αξιοποιείται η σύνδεση των πινάκων και πραγματοποιείται σύνθεση του αποτελέσματος από τους κατάλληλους πίνακες.

Η παραπάνω βασική αρχή του σχεσιακού μοντέλου είναι σημαντικό να γίνει κατανοητή γιατί καθορίζει απόλυτα τον τρόπο με τον οποίο γίνεται ο σχεδιασμός, η υλοποίηση και η χρήση μιας σχεσιακής βάσης δεδομένων. Στις επόμενες παραγράφους παρουσιάζονται οι βασικές έννοιες και τα δομικά στοιχεία του σχεσιακού μοντέλου, ενώ στη συνέχεια του κεφαλαίου καλύπτονται όσα πρέπει να γνωρίζει ο αναγνώστης για να σχεδιάσει σωστά μια απλή Βάση Δεδομένων. Επειδή το σχεσιακό μοντέλο είναι το μόνο που μας αφορά, από το σημείο αυτό και σε όλη την έκταση του βιβλίου, σε οποιαδήποτε αναφορά σε Βάση Δεδομένων, όταν δεν επισημαίνεται κάτι διαφορετικό, θα εννοείται ότι αναφερόμαστε σε Σχεσιακή Βάση Δεδομένων.

3.2.2 Τα σημαντικότερα στοιχεία του μοντέλου

3.2.2.1 Οντότητα

Οντότητα (Entity) ονομάζουμε το πρόσωπο, αντικείμενο ή γεγονός για το οποίο χειριζόμαστε δεδομένα. Κάθε οντότητα αντιστοιχεί και σε μια έννοια που εμφανίζεται στο πεδίο της εφαρμογής μας, η οποία έχει κάποια συγκεκριμένα χαρακτηριστικά και που εκπροσωπείται από συγκεκριμένα πρόσωπα ή αντικείμενα. Οντότητα είναι για παράδειγμα ο *Πελάτης*, ο *Προμηθευτής*, το *Προϊόν* και η *Παραγγελία*, δηλαδή κάτι που «υπάρχει» στο πρόβλημά μας και μας ενδιαφέρει να καταγράψουμε στοιχεία για αυτό. Μπορεί, εκτός από κάτι υλικό, όπως το προϊόν, να πρόκειται για κάτι άυλο, όπως η παραγγελία, η οποία παρόλο που δεν «πιάνεται» είναι και αυτή κάτι που υπάρχει και συγκεντρώνει κάποια στοιχεία που χρειάζεται να χειριστούμε για αυτήν (π.χ. πραγματοποιήθηκε σε κάποια ημερομηνία, ακυρώθηκε, εξοφλήθηκε, κλπ).

Η έννοια της οντότητας είναι σημαντική στη σχεδίαση μιας Βάσης Δεδομένων γιατί τη χρησιμοποιούμε ως οδηγό για να δομήσουμε τη Βάση και να προσδιορίσουμε τα σημαντικότερα στοιχεία της, που είναι οι πίνακες. Τα πραγματικά δεδομένα που θα αποθηκευτούν στη Βάση Δεδομένων θα είναι εκπρόσωποι ή «στιγμιότυπα», όπως ονομάζονται, των οντοτήτων. Αν υποθέσουμε ότι έχουμε να χειριστούμε τα στοιχεία 2 πελατών της επιχείρησής μας, οι 2 αυτοί πελάτες, ο Γιώργος Παπαδόπουλος και ο Νίκος Μέλας, αποτελούν στιγμιότυπα της οντότητας *Πελάτης*.

3.2.2.2 Χαρακτηριστικό

Ονομάζουμε χαρακτηριστικά (Attributes) τα επιμέρους στοιχεία που περιγράφουν ή αφορούν μια οντότητα, δηλαδή τα δεδομένα που μας ενδιαφέρουν για κάθε οντότητα. Τα χαρακτηριστικά της οντότητας *Πελάτης* είναι για παράδειγμα, το Όνομα Πελάτη, το Επώνυμο Πελάτη, το Τηλέφωνο, η Ηλικία και το ΑΦΜ. Για κάθε στιγμιότυπο μιας οντότητας, τα δεδομένα που μπορούμε να αποθηκεύσουμε στη Βάση είναι οι τιμές (ή το περιεχόμενο) των χαρακτηριστικών αυτών για το συγκεκριμένο στιγμιότυπο. Με απλά λόγια, στο παράδειγμά μας, τα δεδομένα που κρατάμε για κάποιον πελάτη είναι το όνομα, επώνυμο, τηλέφωνο, η ηλικία και το ΑΦΜ του, που για τους δυο πελάτες μας είναι:

- Γιώργος, Παπαδόπουλος, 2310111222, 23, 0933432543
- Νίκος, Μέλας, 2310333444, 35, 0921321432

Σε μια Βάση Δεδομένων, η έννοια του χαρακτηριστικού ταυτίζεται με το Πεδίο και η έννοια του στιγμιότυπου με την Εγγραφή, όπως αυτά αναφέρθηκαν στο Κεφάλαιο 2. Κάθε χαρακτηριστικό μιας οντότητας μπορεί να περιέχει συγκεκριμένου τύπου δεδομένα, όπως κείμενο ή αριθμούς, επομένως για την αποθήκευσή του πρέπει να προβλέψουμε ένα κατάλληλο πεδίο, που το μέγεθός του να επαρκεί για τις

ανάγκες αποθήκευσης του χαρακτηριστικού. Για κάθε χαρακτηριστικό που ορίζουμε, πρέπει να καθορίσουμε και τον τύπο δεδομένων του αντίστοιχου πεδίου, ώστε το λογισμικό της Βάσης Δεδομένων να γνωρίζει πώς οι τιμές που θα εισάγουμε, θα μεταφράζονται σε ένα σύνολο από Bytes, ώστε να αποθηκεύονται αποτελεσματικά και πώς θα γίνεται με τον καλύτερο τρόπο η προβολή και ο χειρισμός του χαρακτηριστικού αυτού. Το Όνομα και το Επώνυμο μπορούν να οριστούν ως ένα πεδίο κειμένου με μέγιστο μέγεθος 50 χαρακτήρες, ενώ η Ηλικία θα είναι ένα αριθμητικό πεδίο, κατάλληλο για την αποθήκευση μικρών ακεραίων.

Μεταξύ των χαρακτηριστικών μιας οντότητας, μπορούμε να ορίσουμε κάποιο ως πρωτεύον χαρακτηριστικό. **Πρωτεύον χαρακτηριστικό** είναι αυτό που προσδιορίζει μονοσήμαντα το κάθε στιγμιότυπο, δηλαδή δεν μπορεί να έχει την ίδια τιμή για δύο διαφορετικά στιγμιότυπα. Το Όνομα Πελάτη δεν μπορεί να είναι πρωτεύον χαρακτηριστικό γιατί δεν αποκλείεται η περίπτωση συνωνυμίας και επομένως, μόνο από το Όνομα δεν μπορούμε να είμαστε βέβαιοι για το ποιος είναι ο πελάτης. Για την οντότητα *Πελάτης*, πρωτεύον χαρακτηριστικό μπορεί να είναι ο ΑΦΜ, εφόσον αποκλείεται δύο πελάτες να έχουν τον ίδιο ΑΦΜ. Το πρωτεύον χαρακτηριστικό μπορεί να είναι σύνθετο, όταν αντί ενός χαρακτηριστικού αποτελείται από ομάδα περισσότερων χαρακτηριστικών. Στην περίπτωση αυτή, ενώ το κάθε χαρακτηριστικό από μόνο του δεν προσδιορίζει μοναδικά κάποιο στιγμιότυπο, ο συνδυασμός τιμών όλων των πεδίων που αποτελούν το σύνθετο πρωτεύον χαρακτηριστικό είναι μοναδικός και προσδιορίζει μονοσήμαντα κάποιο στιγμιότυπο.

3.2.2.3 Πίνακας

Ένας πίνακας είναι μια συλλογή δεδομένων σχετικών με μια συγκεκριμένη οντότητα. Χρησιμοποιώντας διαφορετικό πίνακα για κάθε οντότητα αποφεύγεται ο πλεονασμός δεδομένων, η Βάση Δεδομένων γίνεται πιο αποδοτική και μειώνονται τα σφάλματα καταχώρισης δεδομένων. Όπως θα περιγραφεί αναλυτικά σε επόμενο κεφάλαιο, η πρώτη και ουσιαστικότερη εργασία στην υλοποίηση μιας βάσης δεδομένων είναι η σχεδίαση των πινάκων. Μελετώντας την εφαρμογή που θέλουμε να υλοποιήσουμε, εντοπίζουμε τις «οντότητες» για τις οποίες πρέπει να κρατήσουμε δεδομένα π.χ. *Πελάτης*, *Παραγγελία*, *Προϊόν*, κλπ. και στη συνέχεια σχεδιάζουμε έναν πίνακα για την καθεμιά από αυτές τις οντότητες.

Οι πίνακες οργανώνουν τα δεδομένα σε στήλες που λέγονται πεδία και σειρές που λέγονται εγγραφές. Κάθε στήλη ή πεδίο αντιστοιχεί σε ένα χαρακτηριστικό της οντότητας που αφορά ο πίνακας, ενώ κάθε γραμμή ή εγγραφή αντιστοιχεί σε ένα στιγμιότυπο της οντότητας.

- Στήλη πίνακα = Πεδίο = χαρακτηριστικό οντότητας
- Γραμμή πίνακα = Εγγραφή = στιγμιότυπο οντότητας

Κατά τη σχεδίαση του κάθε πίνακα, αποφασίζουμε ποια πεδία (δηλ. ποιες στήλες) θα περιλαμβάνει και τι τύπου δεδομένα θα περιέχονται σε κάθε πεδίο. Κάθε πίνακας περιλαμβάνει τα πεδία που χρειάζονται για να κρατάμε όλες τις πληροφορίες που επιθυμούμε για την κάθε οντότητα ή με άλλα λόγια, όλα τα χαρακτηριστικά της οντότητας (π.χ. στον πίνακα ΠΕΛΑΤΕΣ θα θέλουμε πεδία όπως το όνομα του πελάτη, τηλέφωνο, διεύθυνση, κλπ.). Επισημαίνεται κάτι που θα επαναληφθεί αρκετές φορές στη συνέχεια, ότι είναι σοβαρότατο σφάλμα να συμπεριλάβουμε σε έναν πίνακα πεδία που αφορούν πληροφορία που δεν αποτελεί χαρακτηριστικό της συγκεκριμένης οντότητας.

Παράδειγμα 1

Παρακάτω δίνονται παραδείγματα τριών πινάκων με ενδεικτικό περιεχόμενο που αντιστοιχούν σε μια απλή Βάση Δεδομένων όπου τηρούμε τις παραγγελίες των πελατών μας για ένα κατάστημα ηλεκτρικών ειδών. Η πληροφορία που πρέπει να αποθηκεύσουμε αφορά τα στοιχεία των ίδιων των παραγγελιών, δηλαδή την αξία τους, το πότε δόθηκαν, κλπ, καθώς και τα στοιχεία των πελατών που έδωσαν την κάθε παραγγελία και τα στοιχεία των προϊόντων που περιλαμβάνονται στην παραγγελία.

Επίσης ας θεωρήσουμε για λόγους απλότητας ότι κάθε παραγγελία μπορεί να περιλαμβάνει μόνο ένα προϊόν (παρόλο που ο περιορισμός αυτός δε φαίνεται ρεαλιστικός, είναι ιδιαίτερα σημαντικός για τη διατήρηση της απλότητας του παραδείγματος).

Σύμφωνα με τις βασικές αρχές του σχεσιακού μοντέλου, θα ήταν σφάλμα να αποθηκεύσουμε σε έναν μόνο πίνακα όλη την πληροφορία που μας ενδιαφέρει. Όπως θα αναπτυχθεί σε επόμενο υποκεφάλαιο, το

σωστό σκεπτικό είναι να εντοπίσουμε τις οντότητες του προβλήματος και στη συνέχεια να μοιράσουμε την πληροφορία σε ξεχωριστούς πίνακες για κάθε οντότητα. Σύμφωνα με αυτό το σκεπτικό, θα χρειαστούμε τους πίνακες ΠΕΛΑΤΕΣ, ΠΡΟΪΟΝΤΑ και ΠΑΡΑΓΓΕΛΙΕΣ (Σχήμα 3.1), που ο καθένας τους θα περιλαμβάνει πεδία που αφορούν μόνο την οντότητα στην οποία αντιστοιχεί. (Σημείωση: ο κάθε πίνακας περιλαμβάνει επίσης πεδία που χρησιμεύουν στη σύνδεσή του με άλλους πίνακες, όπως θα εξηγηθεί παρακάτω).

ΠΕΛΑΤΕΣ

Κωδικός Πελάτη	**Όνομα**	**Επώνυμο**	**Διεύθυνση**	**Τηλ**	**Ηλικία**	**ΑΦΜ**
Π1	Γιώργος	Παπαδόπουλος	Νεοφύτου 15	2310111222	23	0933432543
Π2	Νίκος	Μέλας	Μητροπόλεως 2	2310333444	*Null*	0921321432

ΠΡΟΪΟΝΤΑ

Κωδικός Προϊόντος	**Μάρκα**	**Μοντέλο**	**Κατηγορία**	**Τιμή**
A1	PITSOS	P18-super	Πλυντήριο	235,00
A2	MORRIS	Clean 15	Πλυντήριο	332,50
A3	MORRIS	SC43	Σκούπα	76,70

ΠΑΡΑΓΓΕΛΙΕΣ

Κωδικός Παραγγελίας	**Ημερομηνία παραγγελίας**	**Ημερομηνία παράδοσης**	**Έκπτωση (%)**	**Πληρωτέο**	**Παραδόθηκε**	**Κωδ_πελάτη**	**Κωδ προϊόντος**
1	1/2/2015	5/2/2015	0	23	Ναι	Π1	A2
2	8/3/2015	8/3/2015	10	35	Ναι	Π2	A2
3	31/1/2015	31/1/2015	5		Ναι	Π1	A3

Σχήμα 3.1. *Οι πίνακες Πελάτες, Προϊόντα και Παραγγελίες στους οποίους κατανέμονται τα δεδομένα για τις παραγγελίες μιας επιχείρησης.*

Με αναφορά στους πίνακες του παραπάνω παραδείγματος επισημαίνονται τα εξής, που θεωρούνται πολύ σημαντικά για την κατανόηση των πινάκων και του τρόπου χρήσης τους:

1. Κάθε πίνακας αντιστοιχεί σε μια διαφορετική οντότητα. Ορίσαμε τρεις διαφορετικούς πίνακες, έναν για καθεμιά από τις οντότητες *Πελάτες*, *Προϊόντα* και *Παραγγελίες* και για να είναι κατανοητή η Βάση Δεδομένων μας, τους δώσαμε το αντίστοιχο όνομα.
2. Η κάθε στήλη ενός πίνακα ονομάζεται πεδίο και αντιστοιχεί σε ένα χαρακτηριστικό της οντότητας του πίνακα π.χ. το πεδίο (ή αλλιώς στήλη) «Όνομα» του πίνακα ΠΕΛΑΤΕΣ είναι η στήλη όπου θα αποθηκεύουμε το όνομα του κάθε πελάτη, θεωρώντας ότι το όνομα είναι ένα από τα χαρακτηριστικά του πελάτη που θα πρέπει να αποθηκεύουμε.
3. Δεν επιτρέπεται κάποιο πεδίο ενός πίνακα να περιλαμβάνει χαρακτηριστικό της οντότητας άλλου πίνακα. Αν π.χ. στον πίνακα ΠΕΛΑΤΕΣ υπήρχε το πεδίο «Μάρκα» για την αποθήκευση της μάρκας του προϊόντος που αγόρασε ο πελάτης, θα ήταν σοβαρό λάθος. Το πεδίο «Μάρκα» αφορά χαρακτηριστικό του Προϊόντος, επομένως μπορεί να βρίσκεται μόνο στον πίνακα ΠΡΟΪΟΝΤΑ. Σημείωση: Εκτός από τα πεδία που αποτελούν χαρακτηριστικά της οντότητας, κάποιος πίνακας μπορεί να περιλαμβάνει και πεδία που χρησιμεύουν στη σύνδεσή του με κάποιον άλλο πίνακα, όταν οι οντότητες των δύο πινάκων σχετίζονται μεταξύ τους π.χ. το πεδίο «Κωδ_Πελάτη» του πίνακα ΠΑΡΑΓΓΕΛΙΕΣ, παρόλο που ο κωδικός πελάτη είναι χαρακτηριστικό του Πελάτη, συμπεριλαμβάνεται και στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ επειδή επιτρέπει τη σύνδεση των δύο οντοτήτων, ώστε να γνωρίζουμε ποιος είναι ο πελάτης που έδωσε την κάθε παραγγελία.
4. Η κάθε γραμμή ενός πίνακα λέγεται εγγραφή και αντιστοιχεί σε ένα στιγμιότυπο της οντότητας του πίνακα. Με απλά λόγια, η κάθε γραμμή του πίνακα ΠΕΛΑΤΕΣ αντιστοιχεί και σε έναν πελάτη και περιλαμβάνει στοιχεία για τον πελάτη αυτόν. Οι τιμές όλων των πεδίων

μιας γραμμής αφορούν αποκλειστικά και μόνο τον πελάτη στον οποίο αναφέρεται η γραμμή. Για την εισαγωγή ενός νέου πελάτη, θα πρέπει να δημιουργηθεί μια νέα εγγραφή (γραμμή), ενώ για τη διαγραφή ενός πελάτη θα πρέπει να διαγραφεί ολόκληρη η εγγραφή (γραμμή) που του αντιστοιχεί και δεν επιτρέπεται να κρατηθούν επιμέρους πεδία. Μπορεί επίσης να γίνει τροποποίηση των στοιχείων κάποιου πελάτη, μεταβάλλοντας την τιμή κάποιου πεδίου στην εγγραφή που του αντιστοιχεί.

5. Η τιμή ενός πεδίου για κάποια συγκεκριμένη εγγραφή (δηλαδή το περιεχόμενο ενός «κελιού» του πίνακα) μπορεί να είναι μόνο μία. Απαγορεύεται αυστηρά η εισαγωγή πολλαπλών τιμών στο ίδιο πεδίο, κάτι που θα ήταν αντίθετο με τις θεμελιώδεις αρχές του σχεσιακού μοντέλου και θα δημιουργούσε καταστροφική απροσδιοριστία στο χειρισμό των δεδομένων. Για παράδειγμα, στο πεδίο «Τηλ» δεν μπορούμε να εισάγουμε δύο ή περισσότερους αριθμούς τηλεφώνου για έναν πελάτη. Αν ένας πελάτης έχει περισσότερα από ένα τηλέφωνα, θα πρέπει να υπάρχει πρόβλεψη, μετά από κατάλληλο σχεδιασμό, ώστε οι αριθμοί αυτοί να αποθηκευτούν σε διαφορετικά πεδία, π.χ. να υπάρχει πεδίο «Τηλέφωνο Γραφείου», «Κινητό», «Τηλέφωνο σπιτιού».
6. Οι πίνακες των Βάσεων δεδομένων διαφέρουν από τους μαθηματικούς πίνακες, όπως διαφέρουν και από τους πίνακες που σχηματίζουμε σε ένα λογιστικό φύλλο π.χ. στο MS-Excel. Η κάθε γραμμή ενός πίνακα είναι για τη Βάση Δεδομένων μια εγγραφή ή αλλιώς ένα σύνολο τιμών πεδίων που προσδιορίζουν ένα στιγμιότυπο μιας οντότητας. Το σύνολο του περιεχομένου μιας γραμμής, δηλαδή των τιμών όλων των πεδίων σε μια γραμμή ονομάζεται «πλειάδα». Η σειρά με την οποία είναι τοποθετημένη η κάθε εγγραφή (πλειάδα) σε έναν πίνακα δεν έχει καμία σημασία. Το μόνο που μας ενδιαφέρει είναι το ποιες εγγραφές περιλαμβάνει ένας πίνακας, οι οποίες μπορούν να προβληθούν με τυχαίο τρόπο ή να ταξινομηθούν όπως επιθυμούμε. Π.χ. η σειρά με την οποία έχουν αποθηκευτεί οι πελάτες μας στον πίνακα ΠΕΛΑΤΕΣ δεν παίζει ρόλο και η προβολή των πελατών που περιλαμβάνονται στον πίνακα γίνεται είτε με τυχαία σειρά, είτε επιλέγοντας κατά την προβολή κάποιο τρόπο ταξινόμησης (όπως π.χ. αλφαβητικά) που δεν επηρεάζει και δεν εξαρτάται από τον τρόπο αποθήκευσης στον πίνακα.
7. Κάποιο πεδίο μπορεί να είναι κενό, αν για κάποιο λόγο δε γνωρίζουμε την αντίστοιχη πληροφορία ή αν δεν μας ενδιαφέρει για τη συγκεκριμένη εγγραφή. Π.χ. για τον πελάτη Νίκο Μέλα δε γνωρίζουμε την ηλικία, επομένως στον πίνακα ΠΕΛΑΤΕΣ, στη γραμμή που αντιστοιχεί σε αυτόν τον πελάτη, το πεδίο «Ηλικία» θα είναι κενό. Στη γλώσσα των Βάσεων Δεδομένων, αναφερόμαστε στο κενό με τη λέξη Null, εννοώντας το απολύτως άγνωστο ή ασυμπλήρωτο, ώστε να το ξεχωρίζουμε από το χαρακτήρα «κενό» που μπορεί να συναντήσουμε σε ένα κείμενο ή την αριθμητική τιμή 0. Κατά τη σχεδίαση ενός πίνακα, μπορούμε να καθορίσουμε ποια πεδία επιτρέπεται να παραμείνουν κενά και ποια είναι υποχρεωτικό να περιέχουν κάποια τιμή για να είναι έγκυρη η εγγραφή.

Πρωτεύον κλειδί

Σε κάθε πίνακα ορίζουμε ένα πεδίο ως **«πρωτεύον κλειδί»**. Οι τιμές που θα παίρνει το πεδίο αυτό πρέπει να είναι μοναδικές για κάθε εγγραφή, έτσι ώστε να μας εξασφαλίζει ότι όλες οι εγγραφές του πίνακα θα διαφέρουν μεταξύ τους τουλάχιστον ως προς αυτό το πεδίο. Π.χ. ο αριθμός ταυτότητας μπορεί να οριστεί ως πρωτεύον κλειδί στον πίνακα ΠΕΛΑΤΕΣ γιατί είναι μοναδικός για κάθε πελάτη και έτσι δε θα βρεθούν ποτέ σε αυτόν τον πίνακα δύο απολύτως ίδιες εγγραφές, ακόμα και αν υπάρχουν πελάτες με ακριβώς ίδια όλα τα υπόλοιπα στοιχεία τους. Η έννοια του πρωτεύοντος κλειδιού ενός πίνακα είναι η ίδια με την έννοια του πρωτεύοντος χαρακτηριστικού της οντότητας που αφορά ο πίνακας. Ως πρωτεύον κλειδί ενός πίνακα μπορεί να οριστεί το πεδίο που αντιστοιχεί στο πρωτεύον χαρακτηριστικό της οντότητας.

Ο ορισμός κάποιου πεδίου ως πρωτεύον κλειδί είναι σημαντικός και δε θα πρέπει να παραλείπεται. Θεωρητικά, δεν είναι υποχρεωτικό να ορίζεται πρωτεύον κλειδί για όλους τους πίνακες. Στην πράξη όμως, και για όλες σχεδόν τις περιπτώσεις που είναι πιθανό να συναντήσει κάποιος μη ειδικός στην ανάπτυξη Βάσεων Δεδομένων, το πρωτεύον κλειδί είναι απαραίτητο. Αρκετά συχνά, σε έναν πίνακα μπορεί να υπάρχουν περισσότερα πεδία που να προσδιορίζουν μοναδικά κάθε εγγραφή και επομένως το καθένα από αυτά μπορεί να αποτελέσει πρωτεύον κλειδί του πίνακα. Σε αυτήν την περίπτωση, ο σχεδιαστής ορίζει ένα από αυτά ως πρωτεύον κλειδί, ενώ τα υπόλοιπα ονομάζονται εναλλακτικά πρωτεύοντα κλειδιά. Για παράδειγμα, στον πίνακα ΠΕΛΑΤΕΣ, τα πεδία «Αριθμός ταυτότητας», «ΑΦΜ» και «Κωδικός πελάτη»,

μπορούν το καθένα να αποτελέσει πρωτεύον κλειδί, οφείλουμε όμως να ορίσουμε μόνο το ένα από αυτά, ενώ τα υπόλοιπα παραμένουν εναλλακτικά πρωτεύοντα κλειδιά.

Στην πράξη συνηθίζεται να συμπεριλαμβάνουμε σε κάθε πίνακα ένα πεδίο που να χρησιμεύει ως κωδικός ή προσδιοριστικό και να ορίζουμε αυτό το πεδίο ως πρωτεύον κλειδί. Στο πίνακα ΠΕΛΑΤΕΣ μπορούμε να συμπεριλάβουμε το πεδίο «Κωδικός Πελάτη», στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ το «Κωδικός Παραγγελίας», στον πίνακα ΠΡΟΙΟΝΤΑ το «Κωδικός Προϊόντος». Ο λόγος για τον οποίο γίνεται αυτό είναι πρακτικός. Παρόλο που θα μπορούσαμε να προσδιορίσουμε τον κάθε πελάτη μας από τον ΑΦΜ του ή τον αριθμό ταυτότητάς του, προτιμούμε να ορίσουμε εμείς έναν κωδικό που να ελέγχεται από εμάς και να παραμένει αμετάβλητος, έτσι ώστε να μην παρουσιάζεται πρόβλημα αν π.χ. δε γνωρίζουμε τον ΑΦΜ του πελάτη ή αν μετά από κάποια χρόνια ο πελάτης αυτός αλλάξει ταυτότητα. Επίσης ο κωδικός μπορεί να περιλαμβάνει κάποια πληροφορία χρήσιμη για εμάς, π.χ. οι κωδικοί των πελατών εσωτερικού να αρχίζουν από τον αριθμό 100 και οι κωδικοί των πελατών εξωτερικού από το 900. Σε κάποιες περιπτώσεις, όπως π.χ. στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ, μπορεί να μην υπάρχει κάποιο πεδίο με φυσική έννοια που να μπορεί να χρησιμοποιηθεί ως πρωτεύον κλειδί, επομένως είναι απαραίτητος ο ορισμός από εμάς του κωδικού. Οι τιμές που θα παίρνει ένα τέτοιο πεδίο, π.χ. οι κωδικοί παραγγελιών, δεν περιέχουν κάποια χρήσιμη πληροφορία, αλλά εξυπηρετούν την ανάγκη να μπορούμε να προσδιορίσουμε μοναδικά την κάθε παραγγελία.

Το σχεσιακό μοντέλο επιτρέπει να ορίσουμε σε κάποιον πίνακα **σύνθετο πρωτεύον κλειδί**, το οποίο (κατά αναλογία με το σύνθετο πρωτεύον χαρακτηριστικό) αποτελείται από ομάδα περισσότερων από ένα πεδίων που συνολικά προσδιορίζουν μοναδικά τις εγγραφές του πίνακα. Στην περίπτωση του σύνθετου πρωτεύοντος κλειδιού, η τιμή κάποιου ή κάποιων από τα πεδία που συνθέτουν το πρωτεύον κλειδί μπορεί να επαναλαμβάνεται σε περισσότερες εγγραφές. Όμως ο συνδυασμός τιμών όλων των πεδίων που συνθέτουν το πρωτεύον κλειδί είναι μοναδικός, δεν επιτρέπεται να επαναλαμβάνεται και προσδιορίζει μία μόνο εγγραφή. Για παράδειγμα, μπορούμε στον πίνακα ΠΕΛΑΤΕΣ να θεωρήσουμε ως σύνθετο πρωτεύον κλειδί το συνδυασμό πεδίων «Όνομα», «Επώνυμο» και «Ημερομηνία Γέννησης». Ενώ μπορεί περισσότεροι από ένας πελάτες να έχουν το ίδιο όνομα ή ακόμα και το ίδιο επώνυμο ή κάποιοι πελάτες να έχουν την ίδια ημερομηνία γέννησης, θεωρούμε ότι αποκλείεται δύο πελάτες να έχουν ταυτόχρονα ίδιο όνομα, επώνυμο και ημερομηνία γέννησης, επομένως η τριάδα τιμών προσδιορίζει μοναδικά έναν πελάτη.

Σημείωση: Επισημαίνεται ότι ο περιορισμός του να παίρνει το πρωτεύον κλειδί οπωσδήποτε μοναδικές τιμές για κάθε εγγραφή (είτε είναι απλό είτε σύνθετο πρωτεύον κλειδί), πρέπει να επιβάλλεται από τη λογική της σχεδίασης και να είναι εξασφαλισμένο για όλα τα δεδομένα που είναι δυνατόν να προκύψουν κατά τη λειτουργία της Βάσης Δεδομένων. Δεν αρκεί το να ισχύει η μοναδικότητα των τιμών του πρωτεύοντος κλειδιού μόνο περιστασιακά για ένα σύνολο δεδομένων που έχουμε διαθέσιμα κατά τη δημιουργία της Βάσης Δεδομένων, αφού κάτι τέτοιο μπορεί να ανατραπεί όταν θα εισαχθούν νέα δεδομένα.

3.2.2.4 Συσχέτιση πινάκων

Η συσχέτιση ή σύνδεση πινάκων είναι θεμελιώδες στοιχείο στη λογική μιας Σχεσιακής Βάσης Δεδομένων, εξίσου σημαντικό με τους ίδιους τους πίνακες. Εφόσον η πληροφορία κατανέμεται σε πολλούς πίνακες, έναν για κάθε οντότητα, πρέπει να υπάρχει και ένας τρόπος να μπορεί η πληροφορία αυτή να ανακτηθεί, συσχετίζοντας κατάλληλα τους πίνακες μεταξύ τους. Η βασική αρχή που εφαρμόζεται είναι ότι ένα κοινό πεδίο συνδέει δύο πίνακες ώστε η Βάση Δεδομένων να μπορεί να συγκεντρώσει τα δεδομένα από τους δύο πίνακες για προβολή ή επεξεργασία.

Το κοινό πεδίο δεν μπορεί να είναι τυχαίο αλλά να περιέχει κοινή πληροφορία και να έχει προβλεφθεί για να συνδέει λογικά τους δύο πίνακες. Επιπλέον, το κοινό πεδίο πρέπει να προσδιορίζει ακριβώς και μοναδικά το ποια εγγραφή του ενός πίνακα συνδέεται με κάποια εγγραφή του άλλου πίνακα. Εκτός από ελάχιστες ειδικές περιπτώσεις, το κοινό πεδίο που χρησιμοποιούμε για να συνδέσουμε δύο πίνακες είναι το πρωτεύον κλειδί του ενός πίνακα, το οποίο προστίθεται και στον άλλο πίνακα ως «ξένο» κλειδί (foreign key).

Στο παραπάνω παράδειγμα, ο πίνακας ΠΕΛΑΤΕΣ και ο πίνακας ΠΑΡΑΓΓΕΛΙΕΣ πρέπει να ενωθούν με βάση το πεδίο «Κωδικός πελάτη», που είναι το πρωτεύον κλειδί στον πίνακα ΠΕΛΑΤΕΣ και υπάρχει και στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ ως ξένο κλειδί, με το παρεμφερές όνομα «Κωδ πελάτη». Γνωρίζοντας από τη λογική του προβλήματός μας ότι μια παραγγελία πρέπει να συνδέεται με έναν πελάτη, προβλέψαμε στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ το πεδίο «Κωδ πελάτη», όπου για κάθε εγγραφή παραγγελίας καταγράφουμε τον κωδικό του πελάτη που έδωσε την παραγγελία αυτή. Οι επιτρεπτοί κωδικοί που θα εισάγουμε στο «Κωδ Πελάτη» για να προσδιορίσουμε τον πελάτη θα πρέπει προφανώς να είναι αντίστοιχοι με τους κωδικούς που

έχουμε προσδιορίσει για τους πελάτες στο πεδίο «Κωδικός Πελάτη» του πίνακα ΠΕΛΑΤΕΣ, ώστε να γίνεται σωστά η σύνδεση.

Στους πίνακες του παραπάνω παραδείγματος παρατηρούμε ότι η πρώτη παραγγελία του πίνακα (με κωδικό παραγγελίας «1» συνδέεται με τον πελάτη με κωδικό Π1. Στον πίνακα ΠΕΛΑΤΕΣ βλέπουμε ότι ο πελάτης με κωδικό Π1 είναι ο «Γιώργος Παπαδόπουλος». Εφόσον ο κωδικός πελάτη είναι πρωτεύον κλειδί στον πίνακα ΠΕΛΑΤΕΣ, είναι εξασφαλισμένο ότι ο κωδικός Π1 προσδιορίζει με ακρίβεια έναν και μόνο συγκεκριμένο πελάτη. Όλα τα στοιχεία του πελάτη αυτού περιέχονται στην αντίστοιχη εγγραφή (γραμμή) του πίνακα ΠΕΛΑΤΕΣ.

Για την καλύτερη κατανόηση της τρόπου με τον οποίο εφαρμόζεται η σύνδεση πινάκων, και αναφορικά με το παραπάνω παράδειγμα, διατυπώνονται τα παρακάτω ερωτήματα:

- **Πώς μπορούν να βρεθούν τα στοιχεία του πελάτη που έδωσε μια παραγγελία;** Στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ καταγράφεται μόνο ο κωδικός του πελάτη, ο οποίος είναι αρκετός για να προσδιορίσει τον πελάτη. Όλα τα στοιχεία του πελάτη μπορούν στη συνέχεια να βρεθούν στην κατάλληλη γραμμή του πίνακα ΠΕΛΑΤΕΣ.
- **Μπορώ αντί για τον κωδικό πελάτη να χρησιμοποιήσω το όνομα του πελάτη, το τηλέφωνό του ή κάποιο άλλο πεδίο για να συνδέσω τους δύο πίνακες;** Όχι, γιατί αν π.χ. δύο πελάτες έχουν το ίδιο όνομα θα προκληθεί απροσδιοριστία σχετικά με το ποιος είναι ο πελάτης στον οποίο αναφέρεται η παραγγελία. Το πεδίο που μας εξασφαλίζει ότι δε θα υπάρξει τέτοιο πρόβλημα είναι το πρωτεύον κλειδί στον πίνακα ΠΕΛΑΤΕΣ, δηλαδή στην περίπτωσή μας ο «Κωδικός Πελάτη».
- **Μπορώ εκτός από τον κωδικό πελάτη, στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ να συμπεριλάβω επιπρόσθετα και το όνομα του πελάτη ή και κάποια άλλα στοιχεία του πελάτη;** Όχι, γιατί όλα τα στοιχεία του πελάτη υπάρχουν στον πίνακα ΠΕΛΑΤΕΣ. Αν τα ίδια στοιχεία αποθηκευτούν και στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ, έχουμε πλεονασμό δεδομένων και δημιουργείται κίνδυνος σοβαρών προβλημάτων.
- **Μπορεί μια παραγγελία να συνδέεται με περισσότερους από έναν πελάτες;** Όχι. Αυτό καταρχήν αντιβαίνει στη λογική του προβλήματός μας γιατί είναι σίγουρο ότι μια παραγγελία δίνεται από έναν μόνο πελάτη και δε θα χρειαστεί ποτέ να τη συνδέσουμε με περισσότερους πελάτες. Επιπρόσθετα, η σχεδίαση των πινάκων του παραδείγματος δεν επιτρέπει τη σύνδεση μιας παραγγελίας με περισσότερους πελάτες γιατί το πεδίο «Κωδ Πελάτη» μπορεί να πάρει μόνο μία τιμή.
- **Μπορεί ένας πελάτης να συνδέεται με περισσότερες από μία παραγγελίες;** Ναι, αυτό επιτρέπεται από τη λογική του προβλήματός μας και μπορεί να αποτυπωθεί στους πίνακες, αφού δεν υπάρχει περιορισμός στο πόσες παραγγελίες θα έχουν ως κωδικό πελάτη τον ίδιο κωδικό. Στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ του παραδείγματος, οι δύο παραγγελίες με κωδικούς «1» και «3» συνδέονται με τον ίδιο πελάτη, συγκεκριμένα τον πελάτη με κωδικό «Π1» δηλαδή τον «Γιώργο Παπαδόπουλο».

Η σύνδεση δύο πινάκων με τη χρήση ξένου κλειδιού ως κοινό πεδίο, δε γίνεται πάντα με τον ίδιο τρόπο, αλλά εξαρτάται από τον τύπο της σχέσης ανάμεσα στις οντότητες των πινάκων. Η σχέση μεταξύ δύο οντοτήτων μπορεί να είναι Ένα-προς-ένα, Ένα-προς-πολλά ή Πολλά-προς-πολλά, ανάλογα με το πόσα στιγμιότυπα της μιας οντότητας μπορούν να συνδέονται με στιγμιότυπα της άλλης οντότητας. Θα μπορούσαμε για παράδειγμα, αντί της χρήσης του κωδικού πελάτη ως ξένο κλειδί στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ, να συνδέαμε πελάτες με παραγγελίες μεταφέροντας τον Κωδικό Παραγγελίας από τον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ ως ξένο κλειδί στον πίνακα ΠΕΛΑΤΕΣ; Με τον τρόπο αυτό, αντί να καθορίζουμε για κάθε παραγγελία τον πελάτη που την έδωσε, θα μπορούσαμε να καθορίσουμε για κάθε πελάτη την παραγγελία που έδωσε. Ο λόγος για τον οποίο δεν μπορεί βέβαια να γίνει αυτό είναι ότι δεν καλύπτεται η ανάγκη να απεικονίσουμε τη σχέση κάποιου πελάτη με πολλές παραγγελίες. Σε κάθε εγγραφή του πίνακα ΠΕΛΑΤΕΣ, που αντιστοιχεί και σε έναν πελάτη, θα μπορούσαμε να εισάγουμε μόνο έναν κωδικό παραγγελίας. Φαίνεται λοιπόν, ότι για να γίνει σωστά η σύνδεση των πινάκων, πρέπει πρώτα να μελετήσουμε ποιες σχέσεις υπάρχουν ανάμεσα στις οντότητες του προβλήματός μας και τι τύπου είναι οι σχέσεις αυτές. Κάτι τέτοιο μπορεί να γίνει μεθοδικά, όπως θα παρουσιαστεί στις επόμενες δύο παραγράφους, χρησιμοποιώντας ως εργαλείο το διάγραμμα οντοτήτων-συσχετίσεων.

(Το παράδειγμα διατίθεται υλοποιημένο σε Access 2007 μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/Example1.accdb)

3.3 Διάγραμμα οντοτήτων-συσχετίσεων

3.3.1 Γενικά για το διάγραμμα οντοτήτων-συσχετίσεων (Entity-Relationship Diagram – ERD)

Το διάγραμμα οντοτήτων-συσχετίσεων (ERD) είναι ένα διάγραμμα όπου απεικονίζονται γραφικά οι οντότητες που συμμετέχουν στην εφαρμογή που εξετάζουμε και οι συσχετίσεις μεταξύ τους, σύμφωνα με τη λογική της εφαρμογής αυτής. Το ERD είναι ένα σχετικά απλό και ιδιαίτερα χρήσιμο εργαλείο για να αποτυπώσουμε με τυποποιημένο και κατανοητό τρόπο τις πληροφοριακές ανάγκες του χρήστη, δηλαδή για το ποια δεδομένα πρέπει να αποθηκευτούν και ποιος θα είναι ο τρόπος χρήσης τους ως πληροφορία.

Το ERD παρουσιάζεται στο βιβλίο αυτό (σε μια απλουστευμένη του μορφή) επειδή μπορεί να χρησιμεύσει σε ένα στέλεχος επιχείρησης για την οργανωμένη καταγραφή των αναγκών του σχετικά με μια εφαρμογή πληροφορικής που αφορά διαχείριση δεδομένων. Στη συνέχεια, μπορεί ο ίδιος να αναπτύξει μια σχετικά απλή εφαρμογή, ακολουθώντας μια τυποποιημένη διαδικασία, ή εναλλακτικά να συνεργαστεί με έναν ειδικό πληροφορικής, ο οποίος θα αναπτύξει την εφαρμογή καθοδηγούμενος από τη σαφή και τεκμηριωμένη πληροφορία που περιέχει το ERD.

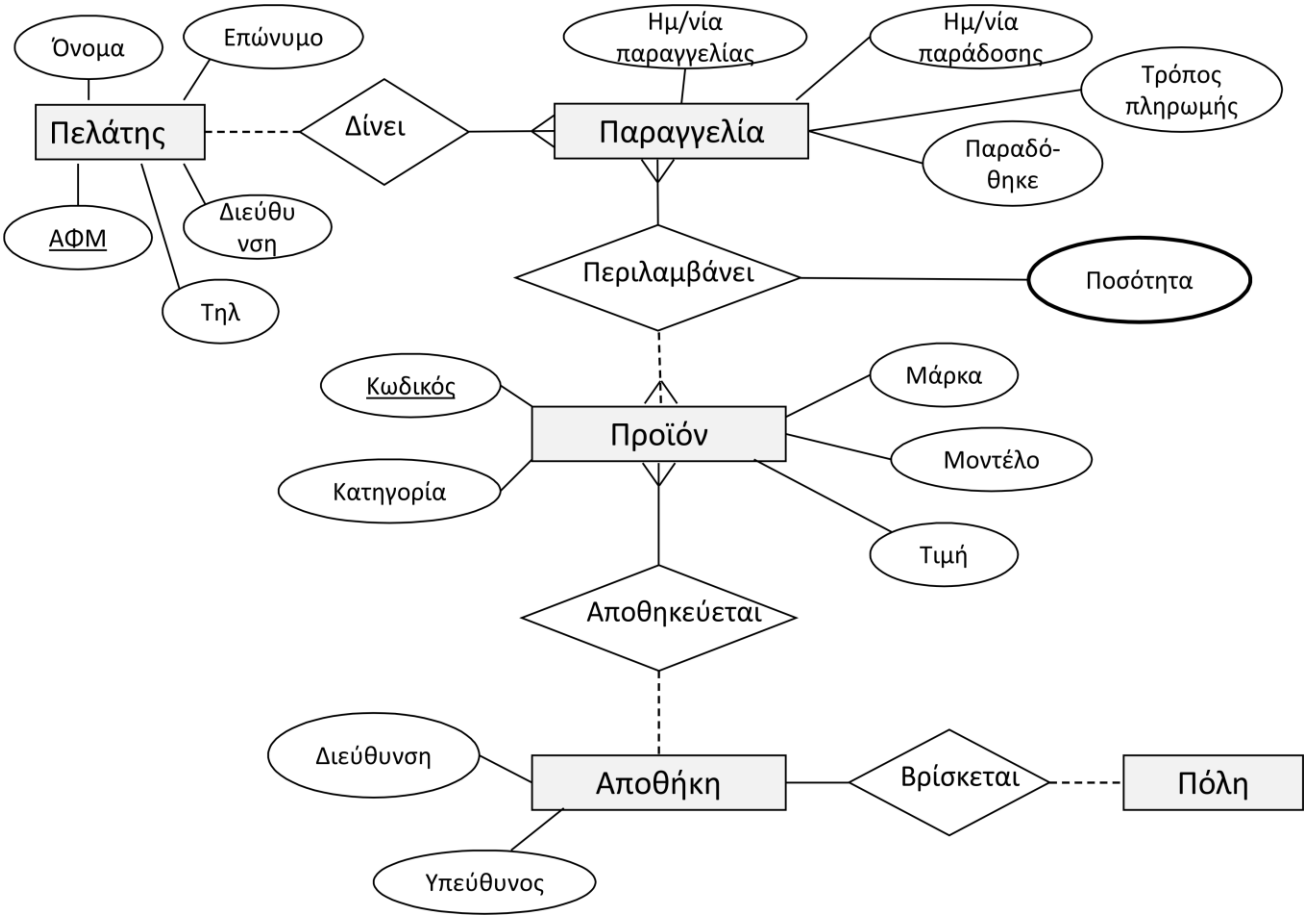
Για την παρουσίαση του ERD, επεκτείνουμε το παράδειγμα για το κατάστημα πώλησης ηλεκτρικών ειδών ως εξής:

Παράδειγμα 2

Για ένα κατάστημα ηλεκτρικών ειδών επιθυμούμε να καταγράφουμε τις παραγγελίες των πελατών και να τηρούμε τα στοιχεία των πελατών και των προϊόντων. Επίσης γνωρίζουμε ότι η επιχείρηση διαθέτει πολλές αποθήκες σε διαφορετικές πόλεις της περιφέρειας και είναι έτσι οργανωμένη ώστε το κάθε προϊόν να είναι συγκεντρωμένο σε μια μόνο αποθήκη. Θεωρούμε ότι η κάθε παραγγελία ενός πελάτη μπορεί να περιλαμβάνει οποιοδήποτε αριθμό προϊόντων και ότι δεν υπάρχουν περισσότερες από μία αποθήκες στην ίδια πόλη.

Στο Σχήμα 3.2 παρατίθεται το διάγραμμα οντοτήτων-συσχετίσεων του παραδείγματος, του οποίου τα στοιχεία και ο τρόπος δημιουργίας επεξηγούνται στις επόμενες παραγράφους, ενώ στο επόμενο υποκεφάλαιο (4) παρουσιάζεται η διαδικασία μετατροπής του διαγράμματος σε πίνακες Βάσης Δεδομένων που καλύπτουν πλήρως τις ανάγκες της εφαρμογής και μπορούν να υλοποιηθούν άμεσα σε οποιοδήποτε λογισμικό Βάσης Δεδομένων.

Τονίζεται ότι το διάγραμμα του Σχήματος 3.2 είναι μόνο μία από τις λύσεις που μπορούν να δοθούν και θα μπορούσε να είναι εξίσου αποδεκτό σε πολλές παραλλαγές του. Εφόσον το ERD είναι εργαλείο σχεδιασμού και κυρίως τρόπος αποτύπωσης των αναγκών του πραγματικού κόσμου, είναι αναμενόμενο να υπάρχει σχετικότητα και ευελιξία στην κατασκευή του. Πολλά στοιχεία του διαγράμματος είναι επιλογές που εμπεριέχουν κρίση και γνώση του τρόπου με τον οποίο θα χρησιμοποιηθούν τα δεδομένα.



Σχήμα 3.2. *Διάγραμμα οντοτήτων-συσχετίσεων καταστήματος ηλεκτρικών ειδών*

3.3.2 Καθορισμός οντοτήτων

Το πρώτο και σημαντικότερο βήμα στη μοντελοποίηση των δεδομένων είναι ο καθορισμός των οντοτήτων που συμμετέχουν στην εφαρμογή που εξετάζουμε. Σύμφωνα με τον ορισμό που έχει δοθεί, οντότητα είναι το πρόσωπο, πράγμα ή γεγονός που έχει κάποια ύπαρξη και για το οποίο κρατάμε κάποια δεδομένα. Κάθε οντότητα που συμμετέχει στο μοντέλο, παριστάνεται στο ERD ως ένα παραλληλόγραμμο και το όνομα της οντότητας συνήθως σημειώνεται στον ενικό αριθμό (π.χ. *Πελάτης*).

Στο ERD μπορούμε να σημειώσουμε και τα χαρακτηριστικά (properties) της κάθε οντότητας. Αυτά συμβολίζονται με ένα ελλειπτικό σχήμα που συνδέεται με την οντότητα που χαρακτηρίζει. Το πρωτεύον χαρακτηριστικό σημειώνεται υπογραμμίζοντας το όνομά του. Επιπλέον, συμβολίζουμε με έλλειψη διπλής γραμμής τα χαρακτηριστικά που μπορούν να πάρουν πολλαπλές τιμές. Η καταγραφή των χαρακτηριστικών όλων των οντοτήτων δεν είναι απαραίτητο να είναι διεξοδική. Ανάλογα με το βαθμό λεπτομέρειας, πολύ συχνά στην πράξη σημειώνουμε τα ιδιαίτερα χαρακτηριστικά στα οποία θεωρούμε ότι πρέπει να δοθεί προσοχή, ενώ μπορούμε να παραλείψουμε αυτά που θεωρούμε αυτονόητα.

Οι οντότητες που συμμετέχουν στο παράδειγμα του καταστήματος ηλεκτρικών ειδών είναι:

- Πελάτης
- Παραγγελία
- Προϊόν
- Αποθήκη
- Πόλη

Στο διάγραμμα του Σχήματος 3.1 φαίνονται οι παραπάνω οντότητες, καθώς και τα χαρακτηριστικά τους.

Σχολιασμός και παρατηρήσεις για την αποφυγή πιθανών λαθών.

- Πρέπει να είναι σαφές ότι οι οντότητες ορίζονται ως έννοιες και όχι ως συγκεκριμένα πρόσωπα ή πράγματα. Μπορούμε να τα σκεφτόμαστε ως τύπους ή ομάδες προσώπων ή πραγμάτων που θέλουμε να καταγράψουμε σε πίνακες. Π.χ. Πελάτης είναι πράγματι μια οντότητα που περιγράφει ένα σύνολο προσώπων. Ο «Γιώργος» είναι ένας συγκεκριμένος πελάτης, δηλαδή στιγμιότυπο ή εκπρόσωπος οντότητας και όχι η ίδια η οντότητα.
- Δε θα πρέπει να γίνεται σύγχυση ανάμεσα στην οντότητα και τα χαρακτηριστικά κάποιας οντότητας. Αν στην εφαρμογή του παραδείγματος θεωρούμε σημαντικό να κρατούμε τους αριθμούς τηλεφώνου των πελατών, ώστε να μπορεί το κατάστημα να τους καλέσει για επιβεβαίωση της παράδοσης, δε θα πρέπει να ξεχνούμε ότι ο «αριθμός τηλεφώνου» δεν είναι οντότητα, αλλά χαρακτηριστικό της οντότητας *Πελάτης*. Πιο πρακτικά, δε θα πρέπει να σκεφτούμε ότι θα χρειαστούμε έναν πίνακα με ΤΗΛΕΦΩΝΑ αλλά έναν πίνακα με ΠΕΛΑΤΕΣ όπου θα αποθηκεύουμε – πιθανότατα μεταξύ άλλων στοιχείων – τα τηλέφωνα των πελατών. Ωστόσο όμως, ο αναγνώστης μπορεί να παρατηρήσει ότι στο σχήμα έχει συμπεριληφθεί ως οντότητα η *Πόλη*. Δε θα ήταν πιο σωστό να θεωρηθεί η «Πόλη» ως χαρακτηριστικό της Αποθήκης (όπως π.χ. η διεύθυνσή της) και όχι ως ξεχωριστή οντότητα; Η απάντηση είναι ότι στέκουν και τα δύο και η επιλογή γίνεται από το σχεδιαστή σύμφωνα με τις ανάγκες της εφαρμογής. Αν θεωρήσουμε ότι η πόλη στην οποία βρίσκεται η αποθήκη είναι απλά ένα στοιχείο της *Αποθήκης* και δεν πρόκειται να χρησιμοποιηθεί ως έννοια οπουδήποτε αλλού, θα αρκούσε ίσως να συμπεριληφθεί ως χαρακτηριστικό της Αποθήκης. Αν θέλουμε όμως να έχουμε τις πόλεις καταγραμμένες με οργανωμένο τρόπο, κωδικοποιημένες και ενδεχομένως να συμπεριλάβουμε επιπλέον στοιχεία για αυτές (π.χ. πληθυσμός) ή/και να στις συσχετίσουμε στην εφαρμογή μας με άλλη οντότητα (π.χ. να διαχωρίζουμε τους πελάτες ανά πόλη ώστε να γνωρίζουμε ποιους από αυτούς εξυπηρετεί καλύτερα η κάθε αποθήκη), τότε είναι απαραίτητο να συμπεριληφθεί η *Πόλη* ως οντότητα.

3.3.3 Καθορισμός συσχετίσεων

Οι συσχετίσεις ανάμεσα στις οντότητες προκύπτουν από τη λογική της εφαρμογής και αποτυπώνουν τον τρόπο με τον οποίο συνδέονται εννοιολογικά οι οντότητες στον πραγματικό κόσμο. Σε κάθε συσχέτιση δίνεται ένα όνομα που είναι συνήθως ρήμα, ώστε η συσχέτιση να μπορεί να διατυπωθεί σε μορφή πρότασης που να συνδέει τις δύο οντότητες, π.χ. Ο *Πελάτης* και η *Παραγγελία* συνδέονται με τη συσχέτιση «δίνει» ώστε να μπορούμε να πούμε: Ο *Πελάτης* δίνει *Παραγγελία*. Η συσχέτιση μπορεί να διαβαστεί και αντίστροφα, αλλάζοντας τη φωνή του ρήματος ή το όνομα της συσχέτισης, διατηρώντας το ίδιο νόημα π.χ. Η *Παραγγελία* δίνεται από τον *Πελάτη*.

Οι συσχετίσεις συμβολίζονται στο ERD ως ρόμβοι που συνδέονται σχηματικά με τις οντότητες που συσχετίζουν.

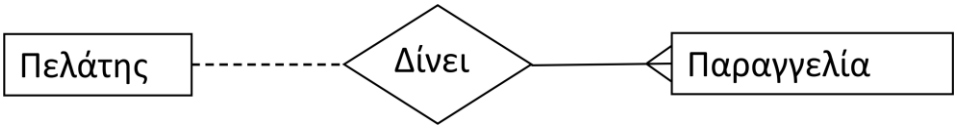
Πολύ σημαντικό είναι να καθορίσουμε τον τύπο της συσχέτισης ως προς την πολλαπλότητα (Multiplicity), δηλαδή τον αριθμό των στιγμιότυπων κάθε οντότητας που μπορούν να συνδέονται με ένα στιγμιότυπο της άλλης οντότητας. Μπορούμε να διακρίνουμε τρεις τύπους συσχέτισης:

- **Ένα-προς-ένα**, όταν ένα στιγμιότυπο της μιας οντότητας συνδέεται μόνο με ένα στιγμιότυπο της άλλης οντότητας και αντιστρόφως.
- **Ένα-προς-πολλά**, όταν ένα στιγμιότυπο της μιας οντότητας μπορεί να συνδέεται με πολλά στιγμιότυπα της άλλης οντότητας, αλλά κάθε στιγμιότυπο της δεύτερης οντότητας συνδέεται μόνο με ένα στιγμιότυπο της πρώτης οντότητας.
- **Πολλά-προς-πολλά**, όταν ένα στιγμιότυπο της μιας οντότητας μπορεί να συνδέεται με πολλά στιγμιότυπα της άλλης οντότητας και επίσης κάθε στιγμιότυπο της δεύτερης οντότητας μπορεί να συνδέεται με πολλά στιγμιότυπα της πρώτης οντότητας.

Ο τύπος της συσχέτισης προκύπτει αφού καθορίσουμε την πολλαπλότητά της και από τις δύο πλευρές. Σε κάθε πλευρά της συσχέτισης σημειώνουμε την πολλαπλότητα της συσχέτισης ως προς την αντίστοιχη οντότητα: η σύνδεση της συσχέτισης με κάθε οντότητα σημειώνεται με απλή γραμμή όταν η σύνδεση είναι προς ένα στιγμιότυπο και με πολλαπλή γραμμή όταν η σύνδεση μπορεί να είναι προς πολλά

στιγμιότυπα. Π.χ. η συσχέτιση «δίνει» που συνδέει τον Πελάτη με την Παραγγελία (Σχήμα 3.3) είναι ένα-προς-πολλά με το «ένα» από την πλευρά του *Πελάτη* και το «πολλά» από την πλευρά της *Παραγγελίας* επειδή:

- Ένας πελάτης μπορεί να δώσει πολλές παραγγελίες
- Μία παραγγελία μπορεί να δοθεί μόνο από έναν πελάτη



Σχήμα 3.3. *Η συσχέτιση του πελάτη με την παραγγελία είναι ένα-προς-πολλά.*

Ένα επιπλέον στοιχείο που μπορούμε να αποτυπώσουμε στο ERD είναι το αν είναι υποχρεωτική ή όχι η συμμετοχή της κάθε οντότητας σε μια συσχέτιση. Η συμμετοχή της οντότητας *Πελάτης* στην συσχέτιση «Δίνει» δεν είναι υποχρεωτική γιατί δεν αποκλείεται να υπάρχει κάποιος πελάτης που να μη συνδέεται με καμία παραγγελία και ωστόσο να αποτελεί έγκυρο στιγμιότυπο της οντότητας (π.χ. έχουν εισαχθεί τα στοιχεία του πελάτη ή είναι γνωστά από συναλλαγή σε παλαιότερη διαχειριστική περίοδο αλλά στο διάστημα που εξετάζουμε δεν έχει δώσει κάποια παραγγελία). Η συμμετοχή της *Παραγγελίας* στη συσχέτιση είναι υποχρεωτική γιατί δε νοείται παραγγελία για την οποία να μην υπάρχει πελάτης που την έδωσε. Η υποχρεωτική συμμετοχή συμβολίζεται με συνεχή γραμμή ενώ η μη υποχρεωτική με διακεκομμένη γραμμή.

Σε κάποιες περιπτώσεις χρειάζεται να συμπεριλάβουμε κάποιο δεδομένο που δεν μπορεί να χαρακτηριστεί ως οντότητα, ούτε αποτελεί χαρακτηριστικό μίας μόνο οντότητας, αλλά είναι χαρακτηριστικό που αφορά τη συσχέτιση δύο οντοτήτων μεταξύ τους. Στην περίπτωση αυτή, το χαρακτηριστικό προσάπτεται στη συσχέτιση και όχι σε κάποια οντότητα, δηλαδή, σχηματικά, η έλλειψη που παριστάνει το χαρακτηριστικό συνδέεται με το ρόμβο που παριστάνει τη συσχέτιση. Παράδειγμα τέτοιας περίπτωσης είναι η ποσότητα (ή αριθμός τεμαχίων) ενός προϊόντος που περιλαμβάνεται σε μια παραγγελία. Το χαρακτηριστικό «Ποσότητα» δεν αφορά αποκλειστικά το προϊόν αφού η ποσότητα αναμένεται να είναι διαφορετική σε κάθε παραγγελία, ούτε αφορά αποκλειστικά μια παραγγελία, αφού είναι διαφορετική για κάθε προϊόν της παραγγελίας. Η ποσότητα ενός συγκεκριμένου προϊόντος που περιλαμβάνεται σε μια συγκεκριμένη παραγγελία αφορά το γεγονός ότι το προϊόν περιλαμβάνεται στην παραγγελία ή με άλλα λόγια τη συσχέτιση των οντοτήτων *Προϊόν* και *Παραγγελία*.

Για την ολοκλήρωση του παραδείγματος, η πολλαπλότητα και η συμμετοχή για τις άλλες τρεις συσχετίσεις που εμφανίζονται στο Σχήμα 3.1 εξηγούνται ως εξής:

Η *Παραγγελία* περιλαμβάνει *Προϊόν*: Μία παραγγελία μπορεί να περιλαμβάνει πολλά προϊόντα και ένα προϊόν μπορεί να περιλαμβάνεται σε πολλές παραγγελίες άρα η σχέση είναι πολλά-προς-πολλά. Δεν υπάρχει παραγγελία που να μη περιλαμβάνει κανένα προϊόν άρα υποχρεωτική συμμετοχή της παραγγελίας στη συσχέτιση, όμως μπορεί να υπάρχει προϊόν που να μην περιλαμβάνεται σε καμία παραγγελία, άρα όχι υποχρεωτική συμμετοχή του προϊόντος στη συσχέτιση.

Το Προϊόν αποθηκεύεται σε Αποθήκη: Ένα προϊόν αποθηκεύεται σε μία μόνο αποθήκη (είναι δεδομένο του παραδείγματος) αλλά μια σε μια αποθήκη μπορεί να αποθηκεύονται πολλά προϊόντα, άρα η σχέση είναι ένα-προς-πολλά με το *Προϊόν* από την πλευρά του πολλά. Δεν υπάρχει προϊόν που να μην αποθηκεύεται σε καμία αποθήκη αλλά μπορεί να υπάρχει αποθήκη στην οποία να μην αποθηκεύεται κανένα προϊόν άρα η συμμετοχή στη συσχέτιση είναι υποχρεωτική μόνο για το *Προϊόν*.

Η *Αποθήκη* βρίσκεται σε *Πόλη*: Μια αποθήκη βρίσκεται μόνο σε μια πόλη και σε μια πόλη βρίσκεται μόνο μία αποθήκη, άρα η συσχέτιση είναι ένα-προς-ένα. Δεν υπάρχει αποθήκη που να μη βρίσκεται σε καμία πόλη αλλά σε μια πόλη είναι δυνατόν να μη βρίσκεται καμία αποθήκη, άρα η συμμετοχή στη συσχέτιση είναι υποχρεωτική μόνο για την *Αποθήκη*.

Παρατήρηση: το ότι σε μια πόλη βρίσκεται μόνο μία αποθήκη δεν είναι κάτι που επιβάλει η λογική του προβλήματος, αλλά κάτι που γνωρίζουμε εμείς και θα μπορούσε να είναι διαφορετικό. Για την ακρίβεια είναι κάτι που γνωρίζει ο υπεύθυνος του καταστήματος και μέσω του ERD γίνεται σαφές σε αυτόν που θα υλοποιήσει τη Βάση Δεδομένων. Γίνεται λοιπόν αντιληπτή η χρησιμότητα του ERD και του ότι εμπεριέχει εξειδικευμένες πληροφορίες για το πρόβλημα.

3.4 Σχεδίαση πινάκων μιας Βάσης Δεδομένων

Το ουσιαστικό μέρος του σχεδιασμού μιας Βάσης Δεδομένων είναι η σχεδίαση των πινάκων της, δηλαδή ο καθορισμός του ποιοι πίνακες θα δημιουργηθούν και ποια πεδία θα περιλαμβάνει ο κάθε πίνακας, ώστε να καλύπτονται οι ανάγκες για τη σωστή αποθήκευση όλων των στοιχείων που χρειαζόμαστε, καθώς και των συσχετίσεων μεταξύ τους. Το διάγραμμα οντοτήτων-συσχετίσεων περιέχει σε σχηματική μορφή τις σημαντικότερες πληροφορίες που θα χρειαστούμε για να υλοποιήσουμε σωστά τους πίνακες της Βάσης Δεδομένων.

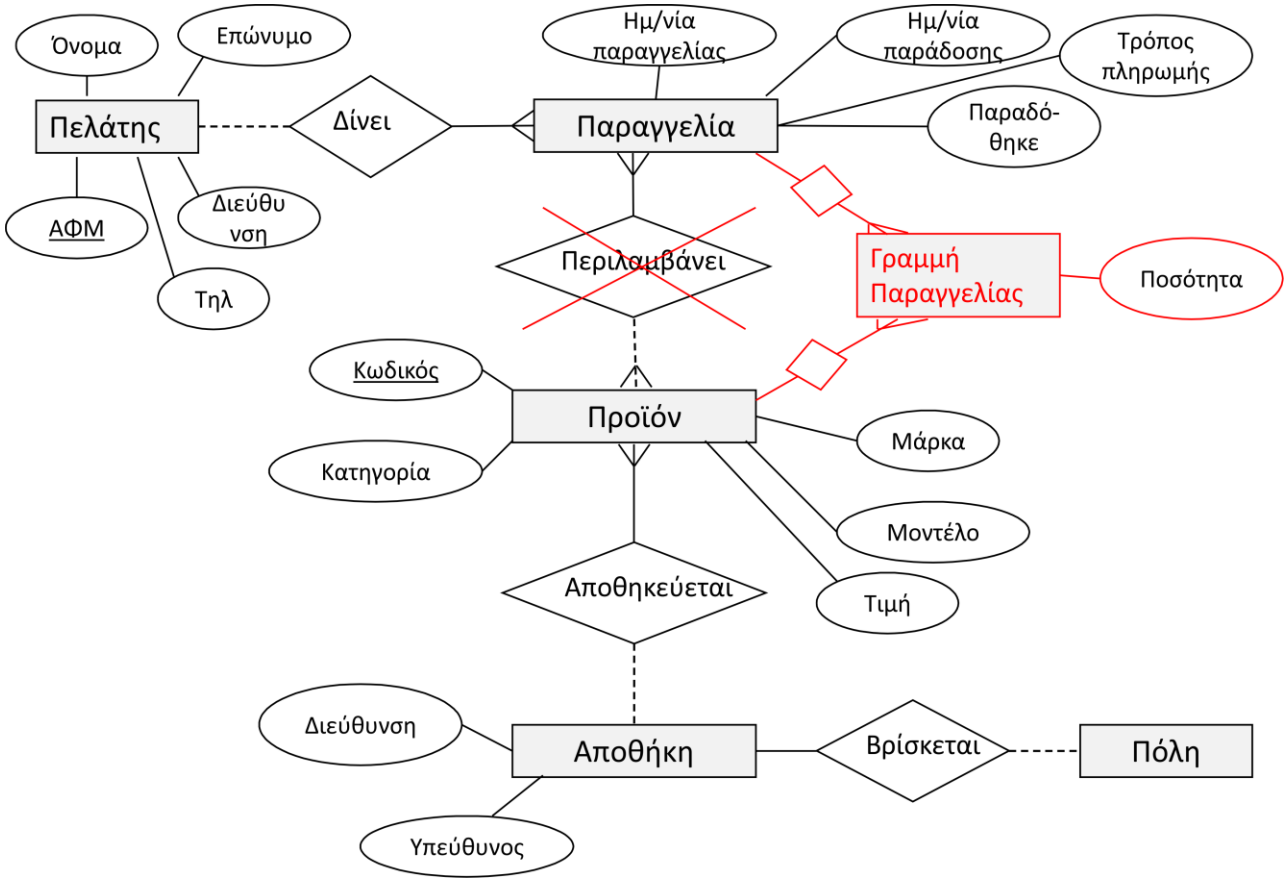
Ακολουθώντας τα παρακάτω βήματα μπορούμε να μετατρέψουμε την πληροφορία που περιέχει το διάγραμμα οντοτήτων-συσχετίσεων σε ένα σύνολο πινάκων που ικανοποιεί τις απαιτήσεις που έχουν αποτυπωθεί στο διάγραμμα:

1. Εντοπίζουμε στο διάγραμμα τις τυχόν συσχετίσεις πολλά-προς-πολλά (M:M). Επειδή οι συσχετίσεις αυτές δεν μπορούν να υλοποιηθούν με χρήση ξένου κλειδιού, δημιουργούμε για καθεμία από αυτές μια νέα βοηθητική οντότητα που συνδέει τις αρχικές οντότητες με σχέση ένα-προς-πολλά. Επομένως η αρχική σχέση πολλά-προς-πολλά αντικαθιστάται από δύο σχέσεις ένα-προς πολλά. Η νέα οντότητα που δημιουργείται εκφράζει τον τρόπο με τον οποίο συνδέονται οι δύο αρχικές οντότητες και έχει ως στιγμιότυπα συνδυασμούς στιγμιότυπων των οντοτήτων αυτών. Μπορούμε να πούμε ότι μια συσχέτιση πολλά-προς-πολλά υλοποιείται με μια νέα οντότητα. Η νέα οντότητα μπορεί να ονομαστεί όπως επιθυμεί ο σχεδιαστής. Σε αρκετές περιπτώσεις έχει κάποια φυσική έννοια οπότε είναι προτιμότερο να έχει ένα όνομα από να υποδηλώνει την έννοια αυτή. Αν υπάρχει χαρακτηριστικό που συνδέεται με την αρχική συσχέτιση πολλά-προς-πολλά, τότε αυτό γίνεται χαρακτηριστικό της νέας οντότητας.
2. Δημιουργούμε έναν πίνακα για κάθε οντότητα. Το όνομα του πίνακα μπορεί να είναι το όνομα της οντότητας.
3. Δημιουργούμε σε κάθε πίνακα ένα πεδίο που θα αποτελέσει το πρωτεύον κλειδί του πίνακα. Το πεδίο αυτό μπορεί να αντιστοιχεί στο πρωτεύον χαρακτηριστικό της οντότητας. Προτιμότερο είναι στην πράξη να δημιουργούμε ένα πεδίο «Κωδικός».
4. Σε κάθε πίνακα δημιουργούμε κατάλληλα πεδία για την αποθήκευση όλων των χαρακτηριστικών της αντίστοιχης οντότητας. Μπορούμε επίσης να προσθέσουμε πεδία για χαρακτηριστικά της κάθε οντότητας που πιθανώς παραλείφθηκαν από το ERD ως αυτονόητα. <u>Προσοχή:</u> δε δημιουργούμε σε έναν πίνακα πεδία για χαρακτηριστικά άλλης οντότητας ούτε πεδία που αφορούν συσχέτιση με άλλες οντότητες.
5. Υλοποιούμε τις συσχετίσεις των οντοτήτων με προσθήκη ξένων κλειδιών ως εξής:
   α) Αν η συσχέτιση των οντοτήτων Α και Β είναι ένα-προς-ένα, έχουμε τρεις επιλογές: προσθέτουμε το πρωτεύον κλειδί του πίνακα Α ως ξένο κλειδί στον πίνακα Β ή προσθέτουμε το πρωτεύον κλειδί του πίνακα Β ως ξένο κλειδί στον πίνακα Α ή αν θέλουμε μπορούμε να συγχωνεύσουμε τους δύο πίνακες σε έναν που θα περιέχει το σύνολο των πεδίων των δύο πινάκων Α και Β. (Σημείωση: η συγχώνευση των δύο πινάκων σε έναν επιτρέπεται όσον αφορά την ορθότητα του σχεδιασμού, αλλά συνήθως δεν είναι επιθυμητή γιατί δημιουργεί ασάφεια σχετικά με τη φυσική έννοια των πινάκων.)
   β) Αν η συσχέτιση των πινάκων Α και Β είναι ένα-προς-πολλά με τον Β να είναι από την πλευρά του πολλά (δηλαδή μια εγγραφή του Α μπορεί να συνδέεται με πολλές εγγραφές του Β), προσθέτουμε το πρωτεύον κλειδί του Α ως ξένο κλειδί στον πίνακα Β.
   γ) Η προηγούμενη περίπτωση (β) εφαρμόζεται ακριβώς με τον ίδιο τρόπο και για τις συσχετίσεις των βοηθητικών οντοτήτων που προέκυψαν από τη διάσπαση των σχέσεων πολλά-προς-πολλά.

Παράδειγμα

Η παραπάνω διαδικασία θα εφαρμοστεί στο παράδειγμα 2, χρησιμοποιώντας ως αφετηρία το ERD του σχήματος 3.2.

1. Στο διάγραμμα υπάρχει μια συσχέτιση πολλά-προς-πολλά, η συσχέτιση «η *Παραγγελία* περιλαμβάνει *Προϊόν*». Αυτή θα αντικατασταθεί από δύο συσχετίσεις ένα-προς-πολλά, αφού προστεθεί μια νέα βοηθητική οντότητα, όπως φαίνεται στο Σχήμα 3.4. Επειδή υπάρχει και το χαρακτηριστικό «Ποσότητα» που συνδέεται με τη συσχέτιση Περιλαμβάνει, η «Ποσότητα» μεταφέρεται ως χαρακτηριστικό της νέας οντότητας. Η νέα οντότητα ονομάστηκε «Γραμμή παραγγελίας» που υποδηλώνει ότι τα στιγμιότυπά της είναι τα επιμέρους στοιχεία των παραγγελιών που αφορούν το κάθε προϊόν (φανταστείτε μια έντυπη παραγγελία όπου η κάθε γραμμή αφορά και ένα διαφορετικό προϊόν, την ποσότητά του και την αξία του). Είναι εύκολο να διαπιστώσει κανείς την πολλαπλότητα των συσχετίσεων που προκύπτουν: Μια παραγγελία έχει πολλές γραμμές παραγγελίας, αλλά κάθε γραμμή παραγγελίας ανήκει σε μία μόνο παραγγελία. Αντίστοιχα, ένα προϊόν περιλαμβάνεται σε πολλές γραμμές παραγγελίας αλλά μια γραμμή παραγγελίας περιλαμβάνει μόνο ένα προϊόν.



Σχήμα 3.4. *Το ERD μετά την αντικατάσταση των συσχετίσεων πολλά-προς-πολλά.*

2. Δημιουργούμε έναν πίνακα για κάθε οντότητα, δηλαδή τους πίνακες ΠΕΛΑΤΕΣ, ΠΑΡΑΓΓΕΛΙΕΣ, ΠΡΟΪΟΝΤΑ, ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ, ΑΠΟΘΗΚΕΣ, ΠΟΛΕΙΣ.
3. Για κάθε πίνακα ξεκινάμε με τη δημιουργία του πεδίου που θα χρησιμεύσει ως πρωτεύον κλειδί. Στο Σχήμα 3.5 παρουσιάζονται οι πίνακες σε αυτήν την πρώτη φάση δημιουργίας τους. Προβάλλονται σε μορφή «Σχεδίασης» (η προβολή Σχεδίασης χρησιμεύει κατά τη φάση δημιουργίας των πινάκων όπου βλέπουμε για έναν πίνακα ποια είναι τα πεδία του και ποιος ο τύπος δεδομένων του κάθε πεδίου, χωρίς να μας ενδιαφέρει το περιεχόμενο – διαφέρει από την προβολή «Δεδομένων» όπου βλέπουμε τον πίνακα στην τελική του μορφή με γραμμές και στήλες μαζί με τις εγγραφές που περιέχει). Προτιμήθηκε η δημιουργία πεδίων τύπου

Κωδικός ακόμα και στις περιπτώσεις όπου είχε καθοριστεί κάποιο πρωτεύον χαρακτηριστικό π.χ. προτιμήσαμε τη δημιουργία κωδικού πελάτη αντί της χρήσης του ΑΦΜ ως πρωτεύον κλειδί. Οι κωδικοί μπορεί να είναι τύπου «Αριθμός» αν επιθυμούμε να αποτελούνται από καθαρούς αριθμούς ή τύπου «Κείμενο» αν θέλουμε να περιλαμβάνουν και χαρακτήρες του αλφαβήτου. Στα παραδείγματα του βιβλίου αυτού επιλέγουμε κωδικούς τύπου Κείμενο.

ΠΕΛΑΤΕΣ
*Κωδικός_Πελάτη: Κείμενο

ΠΑΡΑΓΓΕΛΙΕΣ
*Κωδικός_Παραγγελίας: Κείμενο

ΠΡΟΙΟΝΤΑ
*Κωδικός_Προϊόντος: Κείμενο

ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ
*Κωδικός_Γραμμής: Κείμενο

ΑΠΟΘΗΚΕΣ
*Κωδικός_Αποθήκης: Κείμενο

ΠΟΛΕΙΣ
*Κωδικός_Πόλης: Κείμενο

Σχήμα 3.5. *Οι πίνακες σε προβολή σχεδίασης κατά την πρώτη φάση σχεδιασμού τους.*

4. Συμπληρώνουμε στους πίνακες τα πεδία που απαιτούνται για την αποθήκευση όλων των χαρακτηριστικών που έχουν προσδιοριστεί στο ERD και καθορίζουμε τον τύπο τους όπως φαίνεται παρακάτω. Σχόλια: (α) το Τηλέφωνο και ο ΑΦΜ ορίζονται ως πεδία κειμένου και όχι ως αριθμοί, όπως θα περίμενε κανείς, επειδή το περιεχόμενό τους δεν εκφράζει ποσότητα αλλά είναι ένα σύνολο από αριθμητικά ψηφία που τα χρησιμοποιούμε ακριβώς όπως το κείμενο – δεν κάνουμε πράξεις, δεν έχουμε υποδιαστολές, δεν διώχνουμε τα αρχικά μηδενικά. (β) Χρησιμοποιούμε τον ειδικό τύπο που υπάρχει για τις ημερομηνίες και όχι το κείμενο ώστε να γίνεται πιο αποτελεσματικά ο χειρισμός τους π.χ. υπολογισμός διαστήματος, εντοπισμός άκυρων δεδομένων κλπ.

ΠΕΛΑΤΕΣ	
*Κωδικός_Πελάτη: Κείμενο	
Όνομα:	Κείμενο
Επώνυμο:	Κείμενο
Διεύθυνση:	Κείμενο
Τηλ:	Κείμενο
ΑΦΜ:	Κείμενο

ΠΑΡΑΓΓΕΛΙΕΣ	
*Κωδικός_Παραγγελίας: Κείμενο	
Ημ/νία_Παραγγελίας:	Ημ/νία
Ημ/νία_Παράδοσης:	Ημ/νία
Τρόπος_πληρωμής:	Κείμενο
Κόστος:	Αριθμός
Παραδόθηκε:	Ναι/Όχι
Εξοφλήθηκε	Ναι/Όχι

ΠΡΟΙΟΝΤΑ	
*Κωδικός_Προϊόντος: Κείμενο	
Κατηγορία:	Κείμενο
Μάρκα: Κείμενο	
Μοντέλο:	Κείμενο
Τιμή:	Αριθμός

ΓΡΑΜΜΕΣ **ΠΑΡΑΓΓΕΛΙΑΣ**
*Κωδικός_Γραμμής: Κείμενο
Ποσότητα: Αριθμός

ΑΠΟΘΗΚΕΣ	
*Κωδικός_Αποθήκης: Κείμενο	
Διεύθυνση:	Κείμενο
Υπεύθυνος:	Κείμενο

ΠΟΛΕΙΣ	
*Κωδικός_Πόλης: Κείμενο	
Όνομα:	Κείμενο

Σχήμα 3.6. *Οι πίνακες με τα κατάλληλα πεδία για την αποθήκευση των χαρακτηριστικών.*

5. Για κάθε συσχέτιση του διαγράμματος, προσθέτουμε τα πεδία που απαιτούνται ως ξένα κλειδιά για την υλοποίηση των συσχετίσεων. Για την υλοποίηση της συσχέτισης *Πελάτης* δίνει *Παραγγελία*, μεταφέρουμε αντίγραφο του Κωδικός_πελάτη (πρωτεύον κλειδί του πίνακα από την πλευρά «ένα» της συσχέτισης) στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ (από την πλευρά «πολλά» της συσχέτισης) ως ξένο κλειδί, με το ίδιο ή παρεμφερές όνομα (π.χ. Κωδ_πελάτη). Ομοίως για όλες τις συσχετίσεις ένα-προς-πολλά. Η συσχέτιση *Αποθήκη* βρίσκεται σε *Πόλη* είναι ένα-προς-ένα επομένως μπορούμε να αντιγράψουμε το πρωτεύον κλειδί όποιου πίνακα προτιμούμε ως ξένο κλειδί στον άλλο πίνακα ή ακόμα και να συγχωνεύσουμε τους δύο πίνακες. Στο παράδειγμα αντιγράψαμε τον κωδικό πόλης στον πίνακα ΑΠΟΘΗΚΕΣ (Σχήμα 3.7).

ΠΕΛΑΤΕΣ	
*Κωδικός_Πελάτη: Κείμενο	
Όνομα:	Κείμενο
Επώνυμο:	Κείμενο
Διεύθυνση:	Κείμενο
Τηλ:	Κείμενο
ΑΦΜ:	Κείμενο

ΠΑΡΑΓΓΕΛΙΕΣ	
*Κωδικός_Παραγγελίας: Κείμενο	
Ημ/νία_Παραγγελίας:	Ημ/νία
Ημ/νία_Παράδοσης:	Ημ/νία
Τρόπος_πληρωμής:	Κείμενο
Κόστος:	Αριθμός
Παραδόθηκε:	Ναι/Όχι
Κωδ_πελάτη:	Κείμενο

ΠΡΟΙΟΝΤΑ	
*Κωδικός_Προϊόντος: Κείμενο	
Κατηγορία:	Κείμενο
Μάρκα: Κείμενο	
Μοντέλο:	Κείμενο
Τιμή:	Αριθμός
Κωδ_αποθήκης: Κείμενο	

ΓΡΑΜΜΕΣ **ΠΑΡΑΓΓΕΛΙΑΣ**	
*Κωδικός_Γραμμής: Κείμενο	
Ποσότητα:	Αριθμός
Κωδ_παραγγελίας:	Κείμενο
Κωδ_προϊόντος: Κείμενο	

ΑΠΟΘΗΚΕΣ	
*Κωδικός_Αποθήκης: Κείμενο	
Διεύθυνση:	Κείμενο
Υπεύθυνος:	Κείμενο
Κωδ_Πόλης:	Κείμενο

ΠΟΛΕΙΣ	
*Κωδικός_Πόλης: Κείμενο	
Όνομα:	Κείμενο

Σχήμα 3.7. *Οι τελικοί πίνακες όπου τα πεδία που υλοποιούν τις συσχετίσεις σημειώνονται με μπλε χρώμα και το πρωτεύον κλειδί του κάθε πίνακα με *.*

Η σχεδίαση των πινάκων, που είναι το ουσιαστικότερο μέρος μιας Βάσης Δεδομένων, ολοκληρώθηκε και μπορεί να μεταφερθεί απευθείας σε λογισμικό Βάσεων Δεδομένων όπως η MS-Access. Στη συνέχεια μπορεί να ξεκινήσει η χρήση της Βάσης Δεδομένων με την εισαγωγή πραγματικών δεδομένων. Για την επίδειξη του τρόπου χρήσης των πινάκων που σχεδιάστηκαν, παρουσιάζονται σε προβολή δεδομένων με ενδεικτικό φανταστικό περιεχόμενο (Σχήμα 3.8).

ΠΕΛΑΤΕΣ

Κωδικός Πελάτη	**Όνομα**	**Επώνυμο**	**Διεύθυνση**	**Τηλ**	**ΑΦΜ**
Π1	Γιώργος	Παπαδόπουλος	Νεοφύτου 15	2310111222	0933432543
Π2	Νίκος	Μέλας	Μητροπόλεως 2	2310333444	0921321432
Π3	Μάριος	Καλής	Κορομηλά 3	2310222111	0911121232

ΠΡΟΪΟΝΤΑ

Κωδικός Προϊόντος	**Κατηγορία**	**Μάρκα**	**Μοντέλο**	**Τιμή**	**Κωδ_αποθήκης**
Α1	Πλυντήριο	PITSOS	P18-super	235,00	1
Α2	Πλυντήριο	MORRIS	Clean 15	332,50	1
Α3	Σκούπα	MORRIS	SC43	76,70	2
Α4	Αναλώσιμα	FIRST	G1	13,10	2

ΠΑΡΑΓΓΕΛΙΕΣ

Κωδικός Παραγγελίας	**Ημ/νία_ παραγγελίας**	**Ημ/νία_ παράδοσης**	**Τρόπος_ πληρωμής**	**Κόστος**	**Παραδόθηκε**	**Κωδ_πελάτη**
1	1/2/2015	5/2/2015	Μετρητοίς	235,00	Ναι	Π1
2	8/3/2015	8/3/2015	Μετρητοίς	332,50	Ναι	Π2
3	31/1/2015	31/1/2015	Πίστωση	102,90	Ναι	Π3
4	5/3/2015	8/3/2015	Πίστωση	76,70	Όχι	Π2

ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ

Κωδικός_ Γραμμής	**Κωδ_ παραγγελίας**	**Κωδ_ προϊόντος**	**Ποσότητα**
1	1	Α1	1
2	2	Α2	1
3	3	Α3	1
4	3	Α4	2
5	4	Α3	1

ΑΠΟΘΗΚΕΣ

Κωδικός_ Αποθήκης	**Διεύθυνση**	**Υπεύθυνος**	**Κωδ_πόλης**
ΑΠ1	Γεωλάτου 15	Γιώργος	1
ΑΠ2	Μαρίνας 8	Γιάννης	2

ΠΟΛΕΙΣ

Κωδικός_ Πόλης	**Όνομα**
1	Θεσσαλονίκη
2	Αλεξάνδρεια
3	Μουδανιά

Σχήμα 3.8. *Οι τελικοί πίνακες σε προβολή δεδομένων με ενδεικτικό περιεχόμενο.*

3.5 Περιορισμοί και αρχές αποφυγής προβλημάτων

3.5.1 Πλεονασμός δεδομένων

Ως πλεονασμό δεδομένων ονομάζουμε την κατάσταση κατά την οποία η ίδια πληροφορία είναι αποθηκευμένη περισσότερες φορές από όσες χρειάζεται στη Βάση Δεδομένων. Αυτό μπορεί να συμβεί αν στον ίδιο ή διαφορετικούς πίνακες έχουμε πεδία που περιλαμβάνουν το ίδιο δεδομένο ή κάποιο ισοδύναμο, με την έννοια ότι η τιμή του ενός πεδίου προκύπτει μοναδικά από την τιμή του άλλου πεδίου, χωρίς να δίνει περισσότερη πληροφορία. Π.χ. αν στον πίνακα ΠΕΛΑΤΕΣ υπάρχει πεδίο που αφορά την Ημερομηνία γέννησης του πελάτη, θα ήταν πλεονασμός δεδομένων αν υπήρχε και πεδίο για την Ηλικία του πελάτη, αφού γνωρίζοντας την ημερομηνία γέννησης γνωρίζουμε και την ηλικία του. Ομοίως, αν στον πίνακα ΠΕΛΑΤΕΣ περιλαμβάνεται το τηλέφωνο του πελάτη, είναι πλεονασμός δεδομένων να υπάρχει το τηλέφωνο του πελάτη σε οποιονδήποτε άλλο πίνακα, όπως στις παραγγελίες, το ιστορικό επικοινωνίας, κλπ. Πλεονασμό δεδομένων μπορεί να έχουμε (α) αν κάποιο πεδίο αντιστοιχεί στην ίδια πληροφορία με κάποιο άλλο με βάση την έννοιά του, όπως στο αμέσως παραπάνω παράδειγμα ή (β) αν λόγω κακής σχεδίασης των πινάκων είμαστε αναγκασμένοι να επαναλαμβάνουμε τις ίδιες τιμές σε περισσότερες από μια εγγραφές ενός πίνακα π.χ. αν στον πίνακα ΠΑΡΑΓΓΕΛΙΕΣ υπήρχε και το πεδίο ΑΦΜ πελάτη, θα έπρεπε για κάθε παραγγελία του ίδιου πελάτη να επαναλαμβάνουμε στην αντίστοιχη σειρά του πίνακα το ΑΦΜ του πελάτη.

Όταν υπάρχει πλεονασμός δεδομένων, εκτός από την άσκοπη επανάληψη, δημιουργούνται σοβαρά προβλήματα διαχείρισης των δεδομένων και κίνδυνος σφαλμάτων και ασυνέπειας, σε βαθμό που να απειλείται η Βάση Δεδομένων ακόμα και με αχρήστευση. Με απλά λόγια, θέλουμε το κάθε δεδομένο να βρίσκεται αποθηκευμένο σε ένα μόνο καλά καθορισμένο σημείο, ώστε να μπορούμε να το αναζητήσουμε σωστά, να το τροποποιήσουμε ή να το διαγράψουμε. Ο πλεονασμός δεδομένων θεωρείται σφάλμα στη σχεδίαση μιας Βάσης Δεδομένων.

Σημείωση: Το να «τύχει» να επαναληφθεί κάποια τιμή σε κάποιο πεδίο ή και σε διαφορετικά πεδία ενός πίνακα επειδή αυτή είναι η σωστή πληροφορία, δεν είναι φυσικά πλεονασμός δεδομένων, αλλά κάτι απολύτως αναμενόμενο π.χ. πολλοί πελάτες μπορεί να ονομάζονται «Γιώργος», αλλά το ότι η τιμή «Γιώργος υπάρχει στο πεδίο Όνομα πελάτη σε πολλές εγγραφές δεν έχει καμία σχέση με πλεονασμό δεδομένων. Επίσης η επανάληψη κάποιου πεδίου ενός πίνακα σε άλλο πίνακα ως ξένο κλειδί, ώστε να πραγματοποιηθεί η σύνδεσή τους, δεν είναι πλεονασμός δεδομένων.

3.5.2 Ακεραιότητα των δεδομένων

Ακεραιότητα των δεδομένων (Data integrity) ονομάζουμε την εξασφάλιση ότι τα δεδομένα μπορούν να αποθηκευτούν και να διαβαστούν σωστά, χωρίς αλλοίωση, απροσδιοριστία, σφάλματα ή αστάθεια. Για να επιτευχθεί η ακεραιότητα των δεδομένων, το σχεσιακό μοντέλο επιβάλλει ορισμένους κανόνες σχετικά με το τι είδους τιμές μπορούν να αποθηκευτούν στα πεδία ενός πίνακα. Οι κανόνες αυτοί επιβάλλονται από ειδικούς μηχανισμούς κατά τη λειτουργία της Βάσης Δεδομένων αλλά είναι σημαντικό να τους γνωρίζουμε και κατά τη φάση της σχεδίασης.

- **Ατομικές και αδιαίρετες τιμές**. Βασικός περιορισμός του μοντέλου είναι ότι κάθε πεδίο οποιουδήποτε πίνακα μπορεί για μια συγκεκριμένη εγγραφή να περιέχει μόνο μία απλή τιμή και σε καμία περίπτωση πολλαπλές τιμές. Αν η τιμή ενός πεδίου για κάποια εγγραφή είναι άγνωστη, τότε λέμε ότι η τιμή του πεδίου αυτού είναι το Null, που σημαίνει ότι το πεδίο είναι κενό.
- Δεν επιτρέπεται σε κάποιον πίνακα να υπάρχουν εγγραφές που να είναι απολύτως ίδιες μεταξύ τους σε όλα τους τα πεδία. Για να αποφευχθεί αυτό, είναι υποχρεωτική η χρήση του πρωτεύοντος κλειδιού.
- **Ακεραιότητα της οντότητας (entity integrity)**. Η τιμή του πρωτεύοντος κλειδιού για κάθε εγγραφή δεν μπορεί να είναι κενή (Null) και πρέπει να είναι μοναδική (δηλαδή να μην μπορεί να επαναληφθεί η ίδια τιμή δεύτερη φορά). Με τον τρόπο αυτό, η τιμή του πρωτεύοντος κλειδιού καθορίζει μοναδικά και χωρίς απροσδιοριστία μια εγγραφή, δηλαδή ένα στιγμιότυπο της οντότητας.

- **Ακεραιότητα αναφοράς (referential integrity).** Αν ένας πίνακας περιέχει ένα πεδίο που αποτελεί ξένο κλειδί με το οποίο συσχετίζεται με ένα δεύτερο πίνακα, πρέπει η τιμή που περιέχει το πεδίο αυτό σε κάθε εγγραφή του πίνακα, να ταιριάζει με κάποια τιμή του πρωτεύοντος κλειδιού στον πίνακα στον οποίο αναφέρεται. Με άλλα λόγια, δεν μπορεί κάποια εγγραφή ενός πίνακα να συνδέεται με ανύπαρκτη εγγραφή άλλου πίνακα. Π.χ. οι τιμές του πεδίου «κωδικός πελάτη» για όλες τις εγγραφές του πίνακα ΠΑΡΑΓΓΕΛΙΕΣ πρέπει να είναι τιμές που υπάρχουν στον πίνακα ΠΕΛΑΤΕΣ. Η ακεραιότητα αναφοράς παραβιάζεται αν π.χ. διαγραφεί κάποιος πελάτης αλλά παραμείνουν στη Βάση Δεδομένων οι παραγγελίες που συνδέονται με τον πελάτη αυτόν, οι οποίες συσχετίζονται πλέον με ανύπαρκτο πελάτη.

Εκτός από τους παραπάνω περιορισμούς, που είναι θεμελιώδεις για τη σωστή λειτουργία της Βάσης Δεδομένων, υπάρχουν επιπλέον μηχανισμοί με τους οποίους μπορούμε, μέσω κατάλληλων περιορισμών, να ενισχύσουμε την εγκυρότητα των δεδομένων. Οι περιορισμοί αυτοί σχετίζονται με τη λογική της εφαρμογής και όχι με το ίδιο το μοντέλο.

- **Τύπος δεδομένων**. Ο καθορισμός του τύπου δεδομένων για κάθε πεδίο είναι σημαντικός για την αποτελεσματική αποθήκευση και χειρισμό των δεδομένων, παράλληλα όμως αποτελεί και στοιχειώδη μηχανισμό επιβολής ενίσχυσης της εγκυρότητάς τους. Αν π.χ. για το πεδίο Ημερομηνία γέννησης καθοριστεί ότι είναι τύπου Ημερομηνία (και όχι κείμενο), το σύστημα μπορεί να αναγνωρίσει λανθασμένες ημερομηνίες (π.χ. 32/1/2010).
- **Κανόνες επικύρωσης**. Με τους κανόνες αυτούς μπορούν να αποτραπούν σφάλματα που αντιβαίνουν τη λογική του περιεχομένου των πεδίων π.χ. η ηλικία ενός ατόμου μπορεί να περιοριστεί στο διάστημα 0-120, να απαγορευτούν αρνητικοί αριθμοί στις τιμές των προϊόντων ή μια πιθανότητα να περιοριστεί στο διάστημα 0 ως 1.

3.5.3 Κανονικοποίηση πινάκων

Η κανονικοποίηση (normalization) είναι μια τυπική διαδικασία πολλαπλών βημάτων που μπορεί να εφαρμοστεί στους πίνακες μιας Βάσης Δεδομένων ώστε η σχεδίασή τους να είναι σωστή και αποτελεσματική. Εφαρμόζοντας την κανονικοποίηση, μπορεί να γίνει εντοπισμός πιθανών λαθών στη σχεδίαση των πινάκων και να καθοδηγηθούμε στον επανασχεδιασμό τους, μέσω της διάσπασής τους σε μικρότερους «κανονικούς» πίνακες.

Η διαδικασία μπορεί να εφαρμοστεί είτε (α) μετά τον αρχικό σχεδιασμό ή τροποποίηση κάποιων πινάκων ώστε να ελεγχθεί η ορθότητά τους ή να διορθωθεί η σχεδίασή τους, είτε (β) ως εναλλακτική μέθοδος σχεδιασμού μιας Βάσης Δεδομένων. Στη δεύτερη περίπτωση, αντί του ορισμού των οντοτήτων και τη διαδικασία σχεδιασμού που περιγράφηκε παραπάνω, μπορεί κάποιος να ξεκινήσει με τη δημιουργία ενός πίνακα (ή μικρού αριθμού πινάκων) που να περιέχει στα πεδία του όλη την πληροφορία που θέλουμε να αποθηκεύσουμε. Στην συνέχεια, εφαρμόζοντας τα βήματα της κανονικοποίησης, ο πιθανότατα προβληματικός αρχικός πίνακας μετατρέπεται σταδιακά σε μια ορθή συνολική σχεδίαση που προκύπτει από τη διάσπασή του σε ένα σύνολο κανονικών πινάκων. Η εναλλακτική αυτή μέθοδος σχεδιασμού ονομάζεται **σχεδίαση με διάσπαση**.

Κάθε βήμα της διαδικασίας κανονικοποίησης έχει σαν στόχο να εξασφαλίσει ότι ο πίνακας που εξετάζουμε ακολουθεί κάποιον κανόνα και επομένως βρίσκεται στην αντίστοιχη **κανονική μορφή**. Επειδή η διαδικασία είναι σταδιακή, κάθε κανονική μορφή είναι αριθμημένη και προϋποθέτει τη συμμόρφωση με την προηγούμενη κανονική μορφή. Έτσι, ξεκινάμε με το να εξασφαλίσουμε ότι ένας πίνακας βρίσκεται σε **$1^{η}$ κανονική μορφή**. Στη συνέχεια, με δεδομένο ότι είναι ήδη σε $1^{η}$ κανονική μορφή, εφαρμόζουμε τον κανόνα που θα τον φέρει σε $2^{η}$ κανονική μορφή και συνεχίζουμε με το ίδιο τρόπο. Έχουν οριστεί συνολικά 5 κανονικές μορφές (Codd, 1972), από τις οποίες οι δύο πρώτες είναι ιδιαίτερα σημαντικές, ενώ οι τελευταίες έχουν ενδιαφέρον μόνο σε σύνθετες Βάσεις Δεδομένων με ιδιαίτερες απαιτήσεις. Οι κανονικές μορφές ορίζονται με αυστηρό τρόπο με χρήση μαθηματικών εννοιών, ώστε να εξασφαλίζεται η ορθότητα και αποτελεσματικότητα της υλοποίησης. Με απλά λόγια, οι δύο πρώτες κανονικές μορφές είναι οι εξής:

1η κανονική μορφή. Ένας πίνακας βρίσκεται σε 1η κανονική μορφή (1ΚΜ) αν κάθε τιμή που αποθηκεύεται σε αυτόν είναι απλή και αδιαίρετη. Μας εξασφαλίζει τη στοιχειώδη απαίτηση του σχεσιακού μοντέλου να μην μπορούν να εισαχθούν σε κανένα πεδίο περισσότερες από μία τιμές για την ίδια εγγραφή.

2η κανονική μορφή. Ο τυπικός ορισμός είναι ότι ένας πίνακας βρίσκεται σε 2η κανονική μορφή αν είναι σε 1ΚΜ και επιπλέον όλα τα πεδία του εξαρτώνται συναρτησιακά από το πλήρες πρωτεύον χαρακτηριστικό του. Μας εξασφαλίζει ουσιαστικά ότι κάθε πίνακας αντιστοιχεί σε μια μόνο οντότητα και δεν περιλαμβάνει πεδία που θα μπορούσαν να είναι χαρακτηριστικά άλλης οντότητας. Τα πεδία που δεν πληρούν τον κανόνα αφαιρούνται από τον αρχικό πίνακα και τοποθετούνται σε νέο πίνακα μαζί με το πρωτεύον χαρακτηριστικό που τα προσδιορίζει. Έτσι, στη θέση του προβληματικού πίνακα δημιουργείται ένα σύνολο νέων πινάκων σε 2η κανονική μορφή, ώστε να εξασφαλίζεται ορθότερη κατανομή της πληροφορίας και να αποφεύγεται ο πλεονασμός δεδομένων.

Η πλήρης παρουσίαση της διαδικασίας κανονικοποίησης είναι εκτός των πλαισίων του βιβλίου αυτού και ο ενδιαφερόμενος αναγνώστης παραπέμπεται στη βιβλιογραφία που αναφέρεται στο τέλος του κεφαλαίου.

Βιβλιογραφία/Αναφορές

Κεχρής Ε. (2015). *Σχεσιακές Βάσεις Δεδομένων*, 2η έκδοση. Αθήνα: Εκδόσεις Κριτική.

Codd E.F. (1972). *Further Normalization of the Data Base Relational Model*, New Jersey: Prentice-Hall.

Date C.J. (1981). *An Introduction to Database Systems*, Boston: Addison-Wesley.

Hawryszkiewycz I.T. (1991). *Database Analysis and Design*, London: Maxwell Macmillan International Editions.

Παπαθανασίου Ε.Α. (2008). *Αρχεία και Βάσεις Δεδομένων για Διοικητικά Στελέχη*, Αθήνα: Γκιούρδας Εκδοτική.

Κεφάλαιο 4. Δημιουργία και χρήση μιας σχεσιακής Βάσης Δεδομένων

Σύνοψη
Στο κεφάλαιο αυτό παρουσιάζεται η διαδικασία δημιουργίας μιας Βάσης Δεδομένων σε Access. Έχοντας γνωρίσει τον τρόπο με τον οποίο οργανώνονται τα δεδομένα σε μια Σχεσιακή Βάση Δεδομένων και το βασικό εργαλείο σχεδιασμού πινάκων, περνάμε στην παρουσίαση της ολοκληρωμένης διαδικασίας δημιουργίας μιας Βάσης Δεδομένων, από το σχεδιασμό μέχρι την τελική υλοποίηση και χρήση σε μορφή λογισμικού. Περιγράφονται οι κύριες φάσεις υλοποίησης της Βάσης Δεδομένων, που περιλαμβάνουν τη δημιουργία πινάκων και τον ορισμό συσχετίσεων, ώστε ένα σχέδιο ERD να μετατραπεί σε πραγματική Βάση Δεδομένων. Στη συνέχεια, επεξηγείται η διαδικασία δημιουργίας ερωτημάτων, φορμών και εκθέσεων, με τα οποία η Βάση Δεδομένων γίνεται μια πλήρης λειτουργική και χρηστική εφαρμογή. Στο κεφάλαιο αυτό περιλαμβάνεται η παρουσίαση του περιβάλλοντος σχεδίασης της Access με χρήση απλών παραδειγμάτων και δίνονται σχήματα που καθοδηγούν τον αναγνώστη στην υλοποίηση του παραδείγματος στην πράξη.

Προαπαιτούμενη γνώση
Κεφάλαιο 2. Δεδομένα και Πληροφορίες, Κεφάλαιο 3. Το σχεσιακό μοντέλο Βάσεων Δεδομένων

4.1 Γενικά για την υλοποίηση και χρήση μιας Βάσης Δεδομένων

4.1.1 Ο ρόλος της Βάσης Δεδομένων

Η Βάση Δεδομένων είναι το συνηθέστερο και το πιο ευρέως διαδεδομένο εργαλείο οργάνωσης και αξιοποίησης μιας συλλογής δεδομένων που αφορά μια συγκεκριμένη εφαρμογή πληροφορικής. Όλες σχεδόν οι εφαρμογές πληροφορικής που αφορούν επιχειρήσεις, περιλαμβάνουν και μια Βάση Δεδομένων, η οποία μπορεί να ποικίλει σε μέγεθος και πολυπλοκότητα, από μια ΒΔ λίγων πινάκων που μπορεί να χρησιμοποιείται σε μια προσωπική εφαρμογή π.χ. «οι επαφές μου», ως ΒΔ με χιλιάδες πίνακες και δυνατότητες για εκατομμύρια εγγραφές, υψηλές ταχύτητες και ειδική πρόβλεψη για ασφάλεια και αξιοπιστία, όπως π.χ. αυτή που εξυπηρετεί ένα αεροδρόμιο ή μια αλυσίδα τραπεζικών καταστημάτων. Μια Βάση Δεδομένων δημιουργείται και λειτουργεί με τη βοήθεια ειδικού λογισμικού που και αυτό ποικίλει σε δυνατότητες και κόστος.

Το λογισμικό αυτό ονομάζεται Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) ή DataBase Management System (DBMS). Διαθέτει μηχανισμούς εισαγωγής-εξαγωγής δεδομένων, αναζήτησης, τροποποίησης, αλλά κυρίως επιτρέπει την οργάνωση των δεδομένων ώστε να είναι αποτελεσματική η αποθήκευση, η χρήση και η συντήρησή τους. Το ΣΔΒΔ επιτρέπει τη διασύνδεση μεταξύ των προγραμμάτων εφαρμογών και των φυσικών αρχείων δεδομένων. Μια εφαρμογή (π.χ. ένα πρόγραμμα μισθοδοσίας) μπορεί να διαβάζει δεδομένα από μια Βάση Δεδομένων με τυποποιημένες εντολές ειδικά για αυτό το σκοπό (π.χ. να ζητάει την «ημερομηνία πληρωμής» και το «ποσό» του εργαζόμενου με συγκεκριμένο «όνομα») χωρίς να ενδιαφέρεται για το ποια bytes και ποιού αρχείου αντιστοιχούν σε αυτά τα δεδομένα.

Η μορφή στην οποία συναντιέται στην πράξη μια ΒΔ και ο ρόλος που παίζει το ΣΔΒΔ μπορεί να διακριθεί στις εξής περιπτώσεις:

1. Η ΒΔ πολύ συχνά βρίσκεται στο λογικό επίπεδο χειρισμού των δεδομένων (βλέπε ενότητα 3.3. του κεφαλαίου 2) δηλαδή δεν αποτελεί την κύρια εφαρμογή που βλέπει ο χρήστης, αλλά το ΣΔΒΔ εξυπηρετεί μια άλλη τελική εφαρμογή, χωρίς αυτό να γίνεται αντιληπτό από τον άνθρωπο-χρήστη. Για παράδειγμα, μια εφαρμογή μισθοδοσίας βασίζεται για το χειρισμό των δεδομένων της σε μια ΒΔ, η οποία μπορεί να είναι ανεξάρτητη και να εξυπηρετεί τις ανάγκες της εφαρμογής σε αποθήκευση και ανάκτηση δεδομένων μέσω του ΣΔΒΔ. Ο χρήστης της εφαρμογής μισθοδοσίας έχει πρόσβαση στα δεδομένα που χρειάζεται μέσω των λειτουργιών, οθονών, επιλογών κλπ. της εφαρμογής αυτής, χωρίς να ενδιαφέρεται για την οργάνωση της ΒΔ που κρύβεται από κάτω.

2. Η ΒΔ μπορεί να αποτελεί μια ολοκληρωμένη τελική εφαρμογή, δηλαδή δεν τοποθετείται μόνο στο Λογικό, αλλά και στο Επίπεδο Εφαρμογής. Αυτό συμβαίνει όταν η εφαρμογή αφορά κυρίως την αποθήκευση δεδομένων και την αναζήτηση χρήσιμων πληροφοριών μέσα από τα δεδομένα αυτά. Σε αυτήν την περίπτωση, οι λειτουργίες διαχείρισης των δεδομένων, όπως αποθήκευση, ενημέρωση, αναζήτηση και υποβολή ερωτημάτων, αφορούν απευθείας τον τελικό χρήστη και όχι κάποια πιο σύνθετη εφαρμογή. Για την εξυπηρέτηση τέτοιων αναγκών, το λογισμικό του ΣΔΒΔ διαθέτει ένα ολοκληρωμένο και εύχρηστο περιβάλλον δημιουργίας και χρήσης της ΒΔ και προσφέρει εργαλεία για τη διαμόρφωση διεπαφών χρήστη (π.χ. φορμών, εντύπων, χειριστηρίων) και εφαρμογών που λειτουργούν αυτόνομα. Χαρακτηριστικότερο παράδειγμα είναι η Microsoft Access, που θα γνωρίσουμε στη συνέχεια, η οποία δίνει τη δυνατότητα δημιουργίας ολοκληρωμένων εφαρμογών διαχείρισης δεδομένων, που μπορεί να επιλύσουν ένα μικρής πολυπλοκότητας πρόβλημα σχετικό με δεδομένα.
3. Με τον όρο ΒΔ αναφερόμαστε επίσης σε μια «πηγή» διαθέσιμων δεδομένων που παρέχονται προς αξιοποίηση, είτε μέσω Διαδικτύου είτε μέσω κάποιου ηλεκτρονικού αποθηκευτικού μέσου. Π.χ. μια εταιρεία προώθησης μπορεί να αγοράσει μια «Βάση Δεδομένων» με στοιχεία (ονόματα, διευθύνσεις, κλπ.) πιθανών πελατών-στόχων. Στην περίπτωση αυτή, η ΒΔ είναι μια ήδη ολοκληρωμένη τελική εφαρμογή, που περιλαμβάνει το περιεχόμενο. Η αξία της εστιάζεται στο περιεχόμενο και οι λειτουργίες που μας ενδιαφέρουν είναι μόνο η ανάγνωση και η αναζήτηση, που συνήθως υποστηρίζονται από κάποια εύχρηστη διεπαφή χρήστη.

Από τις παραπάνω 3 περιπτώσεις, στο κεφάλαιο αυτό επικεντρωνόμαστε στην περίπτωση (2), με σκοπό να καλυφθούν οι ανάγκες ενός στελέχους επιχείρησης, ερευνητή ή μαρκετίστα που εργάζεται με δεδομένα και επιθυμεί να μπορεί να τα χειριστεί αποτελεσματικά. Στη συνέχεια του κεφαλαίου αυτού θα παρουσιαστεί η συνολική διαδικασία δημιουργίας μιας εφαρμογής Βάσης Δεδομένων σε περιβάλλον Microsof-Access, από τον αρχικό σχεδιασμό μέχρι τη δημιουργία εύχρηστων εργαλείων για την αξιοποίησή της.

4.1.2 Η χρησιμότητα της Βάσης Δεδομένων με λίγα λόγια

Στις παρακάτω γραμμές συνοψίζεται τι είναι η Βάση Δεδομένων, σε τι χρειάζεται και πώς δημιουργείται:

- Η Βάση Δεδομένων είναι μια συλλογή πληροφοριών που σχετίζονται με ένα συγκεκριμένο θέμα ή σκοπό π.χ. την παρακολούθηση των παραγγελιών των πελατών ή την οργάνωση των στοιχείων μιας έρευνας αγοράς. Μια απλή εφαρμογή ΒΔ μπορεί να δημιουργηθεί σε περιβάλλον Access από ένα στέλεχος επιχείρησης για να καλύψει τις ιδιαίτερες ανάγκες του. Για το σκοπό αυτό αρκούν βασικές γνώσεις πληροφορικής και κατανόηση της διαδικασίας που παρουσιάζεται στις επόμενες ενότητες.
- Όταν τα δεδομένα περιλαμβάνουν πολλές έννοιες και σχέσεις ανάμεσα σε αντικείμενα διαφορετικού τύπου ή όταν επιθυμούμε να κάνουμε τροποποιήσεις και σύνθετες αναζητήσεις, απαιτείται η χρήση Βάσης Δεδομένων. Η τήρηση δεδομένων σε απλούστερες μορφές π.χ. λογιστικά φύλλα, είναι δυνατή μόνο όταν η δομή τους είναι πολύ απλή και οι απαιτήσεις χειρισμού τους είναι ελάχιστες.
- Μια Βάση Δεδομένων ακολουθεί αυστηρούς κανόνες οργάνωσης ώστε να περιορίζονται στο ελάχιστο οι περιττές επαναλήψεις και οι κίνδυνοι λαθών, να αυξάνεται η ασφάλεια των δεδομένων και να δίνεται η δυνατότητα να αναζητούμε γρήγορα και αποτελεσματικά τα δεδομένα που χρειαζόμαστε, όσο σύνθετη και αν είναι η αναζήτηση.
- Δημιουργώντας μια ΒΔ, ένα στέλεχος επιχείρησης μπορεί να συγκεντρώνει όλες τις πληροφορίες που αφορούν ένα θέμα του σε ένα ειδικό αρχείο βάσης δεδομένων. Μέσα στο αρχείο αυτό, τα δεδομένα διαιρούνται σε ξεχωριστούς χώρους αποθήκευσης οι οποίοι λέγονται **πίνακες**. Στη συνέχεια, ο χρήστης μπορεί να αναζητά ή και να τροποποιεί αυτόματα τα δεδομένα που τον ενδιαφέρουν με τη χρήση **ερωτημάτων**. Μπορεί επίσης να εισάγει, να προβάλει και να ενημερώνει τα δεδομένα των πινάκων με τη χρήση ηλεκτρονικών **φορμών.**

Ακόμα, είναι δυνατή η παρουσίαση, εκτύπωση ή δημοσίευση στο Διαδίκτυο δεδομένων σε ευπαρουσίαστη μορφή με τη χρήση **εκθέσεων**. Τέλος, με τη χρήση των εργαλείων που προσφέρονται για σύνθετες αναζητήσεις, επεξεργασία και επισκόπηση, είναι δυνατή η καλύτερη αξιοποίηση των δεδομένων, μετατρέποντας τα ακατέργαστα δεδομένα σε χρήσιμη πληροφορία.

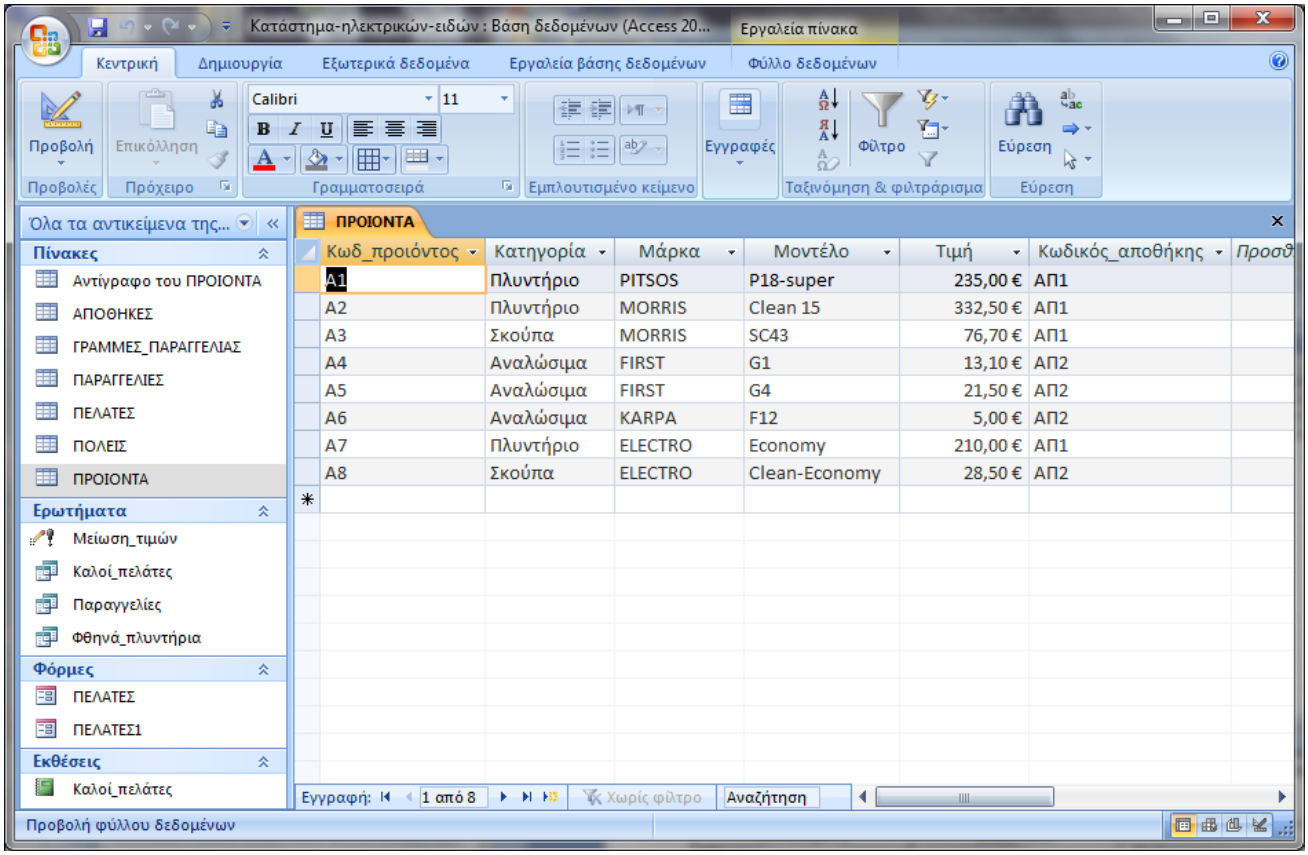
4.1.3 Το περιβάλλον της MS-Access 2007

Στο βιβλίο αυτό, παρουσιάζεται, ως εργαλείο δημιουργίας και χρήσης μιας ΒΔ, η Access της εταιρείας Microsoft, που περιλαμβάνεται στο πακέτο Office (Ξαρχάκος & Καρολίδης, 2010). Ειδικότερα, παρουσιάζεται η έκδοση Office 2007, ως η πιο ευρέως διαδεδομένη κατά το χρόνο συγγραφής του βιβλίου, θεωρείται όμως ότι με τις γνώσεις που θα αποκτήσει ο αναγνώστης θα μπορέσει εύκολα να μεταβεί σε οποιαδήποτε νεότερη έκδοση, ή και να εργαστεί σε κάποια παλαιότερη.

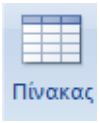
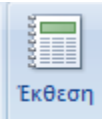
Η Access είναι ένα εύχρηστο Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) που απευθύνεται σε χρήστες χωρίς βαθιά γνώση Βάσεων Δεδομένων και διαθέτει χρήσιμα εργαλεία για τη δημιουργία ολοκληρωμένων εφαρμογών χειρισμού δεδομένων. Το γεγονός ότι συμπεριλαμβάνεται στο πακέτο λογισμικού MS-Office, δείχνει ότι δεν αποτελεί εξειδικευμένο λογισμικό για εφαρμογές υψηλών απαιτήσεων, αλλά ένα προσιτό εργαλείο για οποιονδήποτε επιθυμεί να αποθηκεύσει τα δεδομένα του με τρόπο οργανωμένο, ασφαλή, αποτελεσματικό και με τη δυνατότητα να εξάγει χρήσιμες πληροφορίες. Η Access δίνει τη δυνατότητα δημιουργίας αντικειμένων, όπως πίνακες για την αποθήκευση δεδομένων και ερωτήματα για την ανάκτησή τους.

Τα αντικείμενα της Access μπορούν να δημιουργηθούν και να διαμορφωθούν μέσω ενός γραφικού περιβάλλοντος, που είναι γνώριμο στους περισσότερους αναγνώστες μέσω άλλων ευρέως διαδεδομένων προγραμμάτων της Microsoft, όπως το Word και το Excel. Μια σημαντική παρατήρηση είναι όμως το ότι, ενώ προγράμματα όπως επεξεργαστές κειμένου ή λογιστικών φύλλων μπορούν να χρησιμοποιηθούν σωστά ακόμα και από κάποιον αρχάριο, εμβαθύνοντας σιγά-σιγά στις πιο σύνθετες λειτουργίες, η χρήση της Access απαιτεί την κατανόηση ορισμένων βασικών εννοιών και διαδικασιών πριν μπορέσει ο χρήστης να κάνει την πρώτη του εφαρμογή. Η εξοικείωση με το περιβάλλον, τα κουμπιά, τα μενού κλπ. της Access δεν αρκούν για να φτιάξει κάποιος μια εφαρμογή Βάσης Δεδομένων και αν κάποιος επιχειρήσει να το κάνει «πρακτικά» είναι αναμενόμενο να δημιουργήσει κάτι που θα λειτουργεί λανθασμένα και αναξιόπιστα.

Στο Σχήμα 4.1 παρουσιάζεται μια εικόνα του περιβάλλοντος της Access, ενώ στον Πίνακα 4.1 μια σύνοψη των αντικειμένων της. Μέσω του παραθύρου περιήγησης «Όλα τα αντικείμενα της Access», που βρίσκεται στο αριστερό μέρος του κεντρικού παραθύρου, ο χρήστης έχει πρόσβαση σε όλα τα αντικείμενα σε μια βάση δεδομένων της Access, είτε για να τα προβάλει, είτε για να εργαστεί σε αυτά. Μπορεί να επιλεγεί ο τρόπος με τον οποίο παρουσιάζονται και ταξινομούνται τα αντικείμενα π.χ. ανά τύπο αντικειμένου, κατά ημερομηνία δημιουργίας, κλπ. Συνιστάται η προβολή κατά τύπο αντικειμένου. Με διπλό κλικ σε οποιοδήποτε αντικείμενο, αυτό ανοίγει για προβολή και επεξεργασία στο κύριο μέρος της περιοχής εργασία με τη μορφή μιας καρτέλας (tab). Μπορούμε να έχουμε πολλά αντικείμενα ταυτόχρονα ανοιχτά και να φέρνουμε στο προσκήνιο όποιο επιθυμούμε, επιλέγοντας της αντίστοιχη καρτέλα. Στο Σχήμα 4.1 φαίνεται ως ενεργή η καρτέλα εντολών «Κεντρική», η οποία περιλαμβάνει τις βασικές επιλογές επεξεργασίας και μορφοποίησης περιεχομένου. Σημαντικό είναι το κουμπί επιλογών «Προβολές» που βρίσκεται πρώτο από αριστερά και καθορίζει τον τρόπο προβολής του ανοιγμένου αντικειμένου (η χρησιμότητά του θα φανεί στις επόμενες ενότητες). Η δημιουργία ενός καινούριου αντικειμένου οποιουδήποτε τύπου γίνεται από την καρτέλα εντολών «Δημιουργία».



Σχήμα 4.1. *Το περιβάλλον της Access, όπου φαίνονται όλα τα Αντικείμενα και τα περιεχόμενα του πίνακα ΠΡΟΙΟΝΤΑ.*

Αντικείμενο	Εικονίδιο	Περιγραφή
Πίνακας		Το βασικότερο αντικείμενο της ΒΔ που χρησιμεύει στην αποθήκευση δεδομένων. Τα δεδομένα μοιράζονται σε πολλούς διαφορετικούς πίνακες και οργανώνονται σε γραμμές που λέγονται εγγραφές και στήλες που λέγονται πεδία.
Ερώτημα		Χρησιμεύει στην ανάσυρση δεδομένων ή στην αυτοματοποιημένη τροποποίηση δεδομένων. Δίνει τη δυνατότητα συνδυασμού δεδομένων από πολλούς πίνακες, αναζήτησης με βάση κριτήρια, επιλογής και προβολής των πληροφοριών που μας ενδιαφέρουν.
Φόρμα		Χρησιμεύει στη μορφοποιημένη προβολή ή την εύχρηστη εισαγωγή δεδομένων. Λειτουργεί ως οθόνη στον Η/Υ και μπορεί να διαθέτει κουμπιά, πεδία εισαγωγής, μενού και άλλα ενεργά στοιχεία.
Έκθεση		Χρησιμεύει στη μορφοποιημένη προβολή επιλεγμένων δεδομένων ή πληροφοριών. Προσφέρεται για τη δημιουργία εντύπων κατάλληλων για εκτύπωση.

Σχήμα 4.2. *Τα αντικείμενα της Access*

Τα βασικά βήματα που ακολουθούμε για τη δημιουργία και χρήση μιας εφαρμογής ΒΔ είναι τα ακόλουθα:

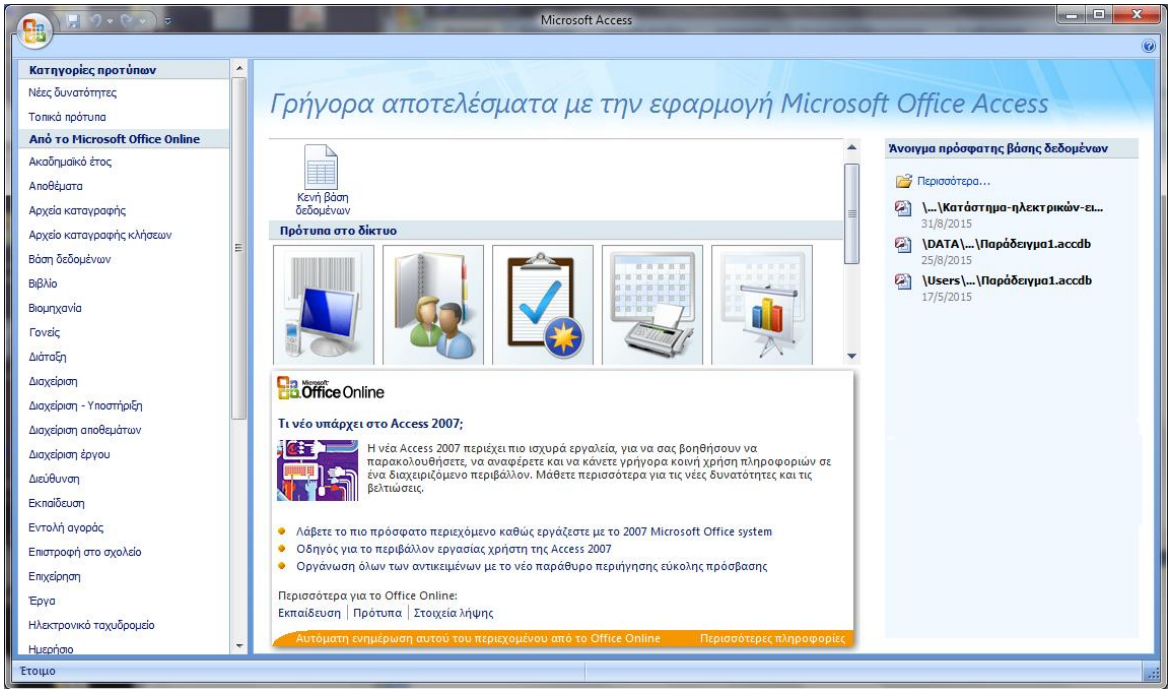
1. Δημιουργία νέας κενής Βάσης Δεδομένων
2. Δημιουργία πινάκων. Δημιουργούμε το χώρο και την οργάνωση που απαιτείται για τη σωστή αποθήκευση των δεδομένων.
3. Ορισμός σχέσεων (προαιρετικό). Δηλώνουμε τον τρόπο σύνδεσης των πινάκων (σε απλές περιπτώσεις μπορεί να παραληφθεί και η σύνδεση των πινάκων να γίνεται κατά την ανάκτηση δεδομένων στα ερωτήματα).
4. Εισαγωγή αρχικών δεδομένων (προαιρετικό). Στη φάση αυτή, ο χρήστης μπορεί να αρχίσει να εισάγει δεδομένα που είναι διαθέσιμα (εναλλακτικά, μια βάση δεδομένων μπορεί να δημιουργηθεί και να είναι πλήρως λειτουργική ακόμα και αν δεν περιέχει δεδομένα – τα δεδομένα εισάγονται αργότερα, κατά τη χρήση).
5. Δημιουργία Ερωτημάτων. Προετοιμάζουμε εργαλεία που αναζητούν, συνδυάζουν, και επεξεργάζονται τα δεδομένα των πινάκων ώστε να παρέχουν τις πληροφορίες που μας ενδιαφέρουν.
6. Δημιουργία φορμών και εκθέσεων. Κατασκευάζουμε εργαλεία που βοηθούν στη μορφοποιημένη και ευπαρουσίαστη προβολή και χειρισμό των δεδομένων.
7. Χρήση. Η εφαρμογή είναι έτοιμη για χρήση στο τελικό περιβάλλον εργασίας. Κατά τη χρήση είναι δυνατή η εισαγωγή δεδομένων, ενημέρωση δεδομένων, ανάκτηση/αναζήτηση δεδομένων, εκτέλεση ερωτημάτων, προβολή φορμών και εκθέσεων.
8. Αναπροσαρμογή. Προσθήκες ή/και τροποποιήσεις στη σχεδίαση των αντικειμένων για την εξυπηρέτηση νέων αναγκών ή τη διόρθωση λαθών. Οι αλλαγές στη σχεδίαση μετά τη χρήση πρέπει να γίνονται με προσοχή επειδή υπάρχει κίνδυνος απώλειας δεδομένων.

Στη συνέχεια του κεφαλαίου αυτού, θα παρουσιαστεί η παραπάνω διαδικασία στο περιβάλλον της Access, χρησιμοποιώντας ως παράδειγμα την περίπτωση του καταστήματος ηλεκτρικών ειδών, όπως περιγράφηκε στο Κεφάλαιο 3. (Το παράδειγμα διατίθεται υλοποιημένο σε Access 2007 μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/Store_electric_appliances_v1.accdb)

4.2 Δημιουργία, άνοιγμα και αποθήκευση μιας Βάσης Δεδομένων Access

Ξεκινώντας την MS-Access 2007, εμφανίζεται η οθόνη υποδοχής του Σχήματος 4.3. Στην οθόνη αυτήν δίνεται η δυνατότητα να δημιουργηθεί «Κενή Βάση Δεδομένων», επιλέγοντας το αντίστοιχο εικονίδιο στο επάνω μέρος της οθόνης, να ανοιχτεί ένα υπάρχον αρχείο της Access, επιλέγοντας ένα από τα πρόσφατα στα οποία έχουμε εργαστεί από τη λίστα στο δεξιό μέρος της οθόνης ή να γίνει λήψη ενός από τα έτοιμα πρότυπα που προσφέρει η Microsoft. Επιλέγοντας Κενή Βάση Δεδομένων, εμφανίζεται στο κάτω δεξιά μέρος της οθόνης ένα πεδίο όπου καθορίζουμε το όνομα που θα δοθεί στο νέο αρχείο και ο φάκελος μέσα στον οποίο θα δημιουργηθεί. Η κατάληξη στο όνομα αρχείου που χρησιμοποιεί η Access 2007 είναι η **.accdb**, που εμφανίζεται αυτόματα στο παραπάνω πεδίο και καλό είναι να μην αλλαχτεί.

Σημείωση: Συνιστάται να γίνει εξαρχής η εισαγωγή κατάλληλα επιλεγμένου ονόματος αρχείου και του επιθυμητού φακέλου στον οποίο θα δημιουργηθεί η νέα Βάση Δεδομένων και όχι κάποιου πρόχειρου ονόματος. Ο λόγος είναι ότι από τη στιγμή της δημιουργίας ή ανοίγματος ενός αρχείου, η Access διατηρεί το αρχείο ανοιχτό ώστε να διαβάζει και να γράφει δεδομένα οποιαδήποτε στιγμή. Επομένως, για να γίνει μετονομασία ή μεταφορά σε άλλη θέση στο δίσκο του Η/Υ, πρέπει πρώτα να κλείσει η ΒΔ, ώστε η Access να αποδεσμεύσει το αρχείο.



Σχήμα 4.3. *Η οθόνη υποδοχής της Access.*

Η αποθήκευση όλων των στοιχείων της Βάσης Δεδομένων γίνεται καθόλη τη διάρκεια κατά την οποία ένα αρχείο είναι ανοιχτό στην Access και αναφέρεται στα συγκεκριμένα αντικείμενα στα οποία εργαζόμαστε και όχι σε ολόκληρη τη Βάση Δεδομένων. Σημειώνεται ότι η γνωστή από άλλα προγράμματα του MS-Office «Αποθήκευση» ή «Αποθήκευση ως», στην Access λειτουργεί διαφορετικά, δηλαδή η αποθήκευση ή μετονομασία γίνεται ανά αντικείμενο εντός της Βάσης και όχι συνολικά. Τελειώνοντας την εργασία μας και αφού έχουν αποθηκευτεί οι τελικές προσθήκες που θέλουμε να κρατήσουμε σε όλα τα επιμέρους αντικείμενα, αρκεί να κλείσουμε το αρχείο της Access.

Σημείωση: αν η νέα Βάση δημιουργηθεί σε φορητό μέσο αποθήκευσης (π.χ. usb stick) δε θα πρέπει το μέσο αυτό να αφαιρεθεί όσο εργαζόμαστε στη ΒΔ (μέχρι να κλείσει επιτυχώς το αρχείο Access). Αν γίνει αυτό κατά λάθος, η Access θα «κολλήσει» και θα χαθούν όλες οι αναποθήκευτες αλλαγές.

Αφού επιλεγεί το όνομα και η επιθυμητή θέση του αρχείου, πατώντας «Δημιουργία», η Access ξεκινά μια κενή Βάση Δεδομένων και μας μεταφέρει αμέσως στη δημιουργία του πρώτου πίνακα, στον οποίο έχει δοθεί το πρόχειρο όνομα «Πίνακας1».

4.3 Δημιουργία Πινάκων

4.3.1. Αρχικός Σχεδιασμός

Η πρώτη και ουσιαστικότερη εργασία στην υλοποίηση μιας Βάσης Δεδομένων είναι η σχεδίαση των πινάκων. Αν η σχεδίαση αυτή δε γίνει σωστά από την αρχή, η εφαρμογή δε θα είναι αξιόπιστη και αποδοτική αλλά πιθανότατα ακόμα και άχρηστη.

Ο Πίνακας, οι σχετικές με αυτόν έννοιες και οι αρχές σχεδίασής του παρουσιάζονται στο Κεφάλαιο 3, παρ 2.2.3. Στην παρ 4 του ίδιου κεφαλαίου παρουσιάζεται αναλυτικά η διαδικασία σχεδίασης πινάκων, χρησιμοποιώντας ως εργαλείο σχεδιασμού το διάγραμμα οντοτήτων-συσχετίσεων (Κεχρής, 2015). Η διαδικασία που παρουσιάζεται στο Κεφάλαιο 3 είναι η συνιστώμενη μέθοδος που προτείνεται για τον αρχικό σχεδιασμό των πινάκων, η οποία ξεκινάει από την οργανωμένη καταγραφή των αναγκών της εφαρμογής, ακολουθεί αξιόπιστη τυποποιημένη διαδικασία και καταλήγει σε έναν «καλό» σχεδιασμό. Το αποτέλεσμα της διαδικασίας αυτής είναι το σύνολο των πινάκων, με τα πεδία που πρέπει να περιλαμβάνει ο καθένας, τον τύπο δεδομένων του κάθε πεδίου, καθώς και τον τρόπο σύνδεσης των πινάκων με χρήση κλειδιών, όπως στο παράδειγμα του Σχήματος 3.7 του Κεφαλαίου 3. Ακολουθώντας τη σχεδίαση αυτή, μπορούν να

δημιουργηθούν οι πίνακες στην Access μέσω απλών πρακτικών βημάτων που παρουσιάζονται στην επόμενη ενότητα.

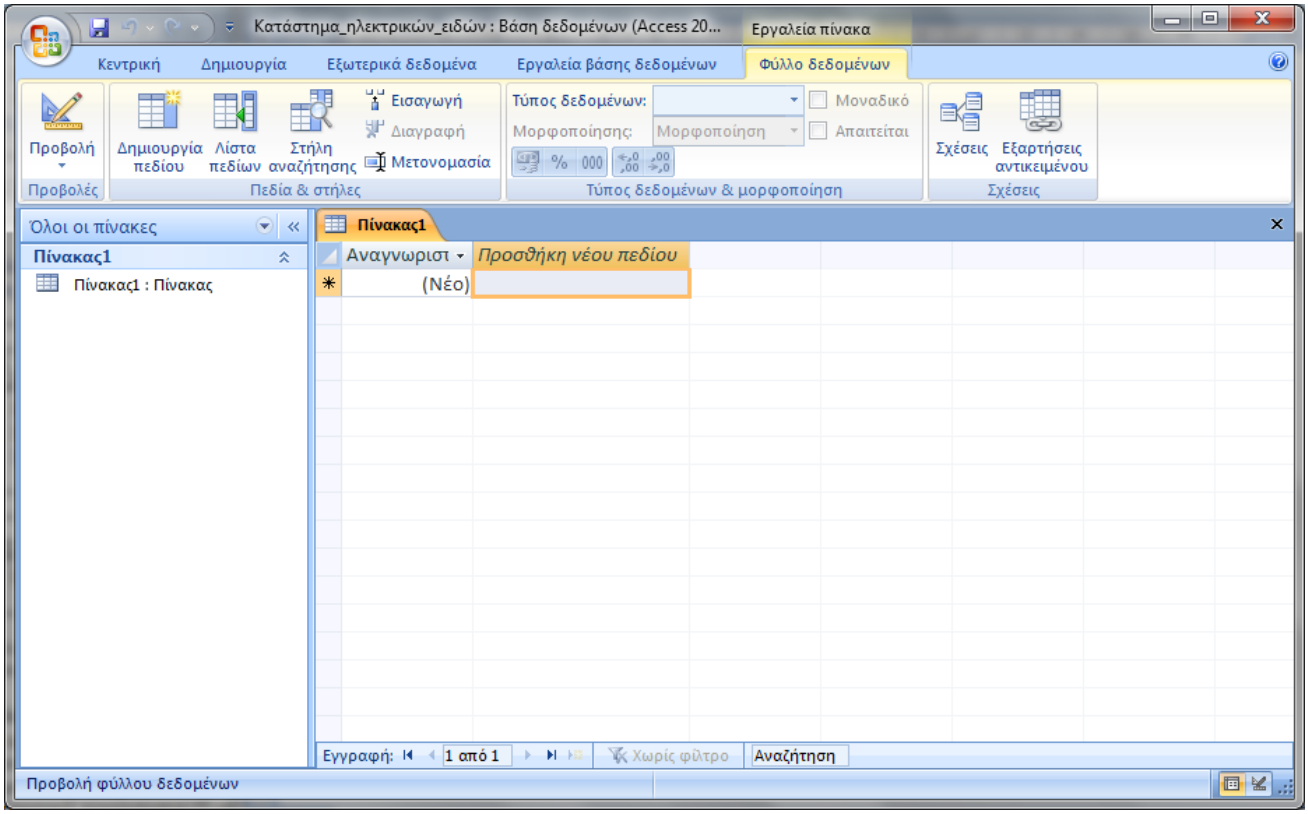
Σε περιπτώσεις μικρών και απλών εφαρμογών, και αφού ο δημιουργός της εφαρμογής έχει κατανοήσει τις έννοιες σχετικά με τους πίνακες και τη σχεσιακή βάση δεδομένων γενικότερα, είναι δυνατή η δημιουργία των πινάκων απευθείας στο περιβάλλον της Access, χωρίς να έχει προηγηθεί ιδιαίτερη φάση σχεδιασμού. Παρόλο που η «εμπειρική» δημιουργία μιας Βάσης Δεδομένων δε συνιστάται, στη συνέχεια παρουσιάζεται μια απλοποιημένη πρακτική διαδικασία υλοποίησης πινάκων, που συνοψίζει τις θεμελιώδεις αρχές που πρέπει να ληφθούν υπόψη:

- Ένας πίνακας είναι μια συλλογή δεδομένων σχετικών με μια συγκεκριμένη οντότητα, δηλ. αντικείμενο, έννοια ή γεγονός (π.χ. προϊόν, πελάτης, παραγγελία). Χρησιμοποιώντας διαφορετικό πίνακα για κάθε οντότητα αποφεύγεται ο πλεονασμός δεδομένων, η βάση δεδομένων γίνεται πιο αποδοτική και μειώνονται τα σφάλματα καταχώρισης δεδομένων.
- Μελετώντας την εφαρμογή που θέλουμε να υλοποιήσουμε, εντοπίζουμε τις «οντότητες» για τις οποίες πρέπει να κρατήσουμε δεδομένα π.χ. «πελάτης», «παραγγελία», «προϊόν», κλπ. Στη συνέχεια σχεδιάζουμε έναν πίνακα για την καθεμιά από αυτές τις οντότητες.
- Οι πίνακες οργανώνουν τα δεδομένα σε στήλες που λέγονται πεδία και σειρές που λέγονται εγγραφές. Κατά τη σχεδίαση του κάθε πίνακα, αποφασίζουμε ποια πεδία (δηλ. ποιες στήλες) θα περιλαμβάνει και τι τύπου δεδομένα θα περιέχονται σε κάθε πεδίο. Κάθε πίνακας περιλαμβάνει τα πεδία που χρειάζονται για να κρατάμε όλες τις πληροφορίες που επιθυμούμε για την κάθε οντότητα (π.χ. στον πίνακα Πελάτες θα θέλουμε τα πεδία: όνομα του πελάτη, τηλέφωνο, διεύθυνση, κλπ.). Είναι σφάλμα να συμπεριλάβουμε σε ένα πίνακα πεδία που αφορούν πληροφορία άσχετη με τη συγκεκριμένη οντότητα.
- Σε κάθε πίνακα ορίζουμε ένα πεδίο ως «πρωτεύον κλειδί». Οι τιμές που θα παίρνει το πεδίο αυτό πρέπει να είναι μοναδικές για κάθε εγγραφή, έτσι ώστε να μας εξασφαλίζει ότι όλες οι εγγραφές του πίνακα θα διαφέρουν μεταξύ τους τουλάχιστον ως προς αυτό το πεδίο. Π.χ. ο αριθμός ταυτότητας μπορεί να οριστεί ως πρωτεύον κλειδί στον πίνακα πελάτες γιατί είναι μοναδικός για κάθε πελάτη και έτσι δε θα βρεθούν ποτέ σε αυτόν τον πίνακα δύο απολύτως ίδιες εγγραφές, ακόμα και αν υπάρχουν πελάτες με ακριβώς ίδια όλα τα υπόλοιπα στοιχεία τους. Συνήθως χρησιμοποιούμε ως πρωτεύον κλειδί έναν κωδικό που ορίζουμε εμείς.
- Ένα κοινό πεδίο συσχετίζει δύο πίνακες ώστε η Access να μπορεί να συγκεντρώσει τα δεδομένα από τους δύο πίνακες για προβολή, επεξεργασία ή εκτύπωση.
- Η σύνδεση δύο πινάκων γίνεται αντιγράφοντας το πρωτεύον κλειδί του ενός πίνακα ως ξένο κλειδί στο δεύτερο πίνακα. Προσοχή: Σε σχέσεις «1 προς πολλά», χρησιμοποιούμε το πρωτεύον κλειδί του πίνακα από την πλευρά του «1» ως ξένο κλειδί στον πίνακα από την πλευρά του «πολλά». Π.χ. για να συνδεθούν οι παραγγελίες με τους πελάτες, επειδή ένας πελάτης δίνει πολλές παραγγελίες, πρέπει να χρησιμοποιηθεί ο κωδικός πελάτη στον πίνακα παραγγελίες (και όχι το αντίστροφο), ώστε για κάθε παραγγελία να γνωρίζουμε τον κωδικό του μοναδικού πελάτη που την έδωσε. Σχέσεις «πολλά-προς-πολλά» υλοποιούνται μόνο με χρήση βοηθητικού πίνακα.

Σχήμα 4.4. *Παράδειγμα ενός πίνακα Access*

4.3.2 Δημιουργία πινάκων

Μετά τον αρχικό σχεδιασμό ακολουθεί η δημιουργία των πινάκων στην Access. Υπάρχουν δύο βασικοί τρόποι να δούμε και να εργαστούμε με έναν πίνακα. Σε **προβολή σχεδίασης** μπορούμε να δημιουργήσουμε έναν ολόκληρο πίνακα από την αρχή ή να προσθέσουμε, να διαγράψουμε ή να προσαρμόσουμε τα πεδία ενός πίνακα που υπάρχει ήδη. Σε **προβολή φύλλου δεδομένων** μπορούμε να προβάλλουμε τα δεδομένα ενός πίνακα, να προσθέσουμε ή να επεξεργαστούμε δεδομένα. Με άλλα λόγια, σε προβολή σχεδίασης βλέπουμε τον ορισμό της μορφής του πίνακα, δηλαδή τους τύπους των δεδομένων που θα μπορεί να δεχθεί, ενώ σε προβολή δεδομένων βλέπουμε τα ίδια τα δεδομένα που περιέχει. Η δημιουργία ενός νέου πίνακα συνιστάται να γίνεται πάντα σε προβολή σχεδίασης, γιατί έτσι έχουμε περισσότερες δυνατότητες και πλήρη έλεγχο στην κατασκευή του πίνακα.



Σχήμα 4.5. *Η αρχική οθόνη αμέσως μετά τη δημιουργία νέας Βάσης Δεδομένων σε Access*

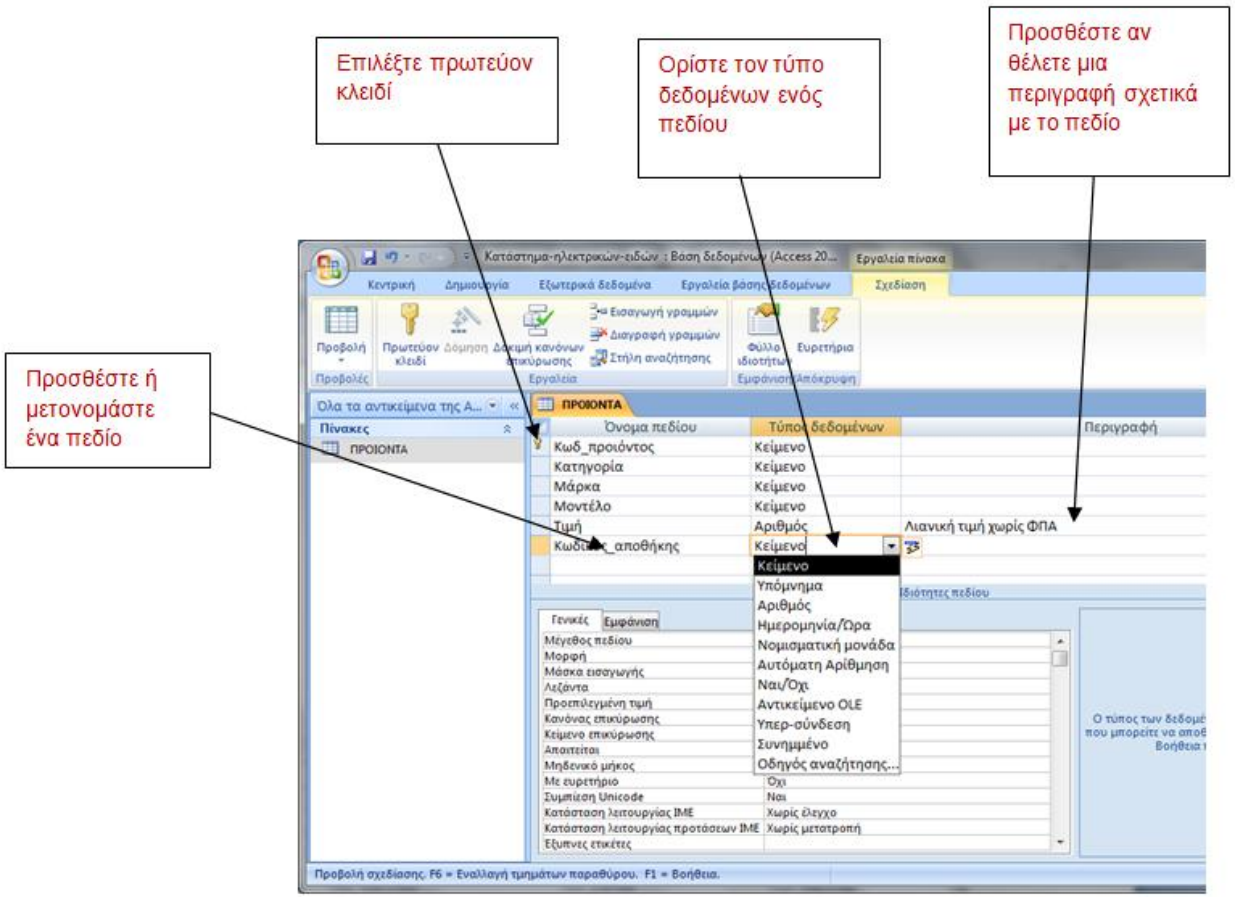
Η διαδικασία δημιουργίας πινάκων παρουσιάζεται χρησιμοποιώντας ως παράδειγμα τη Βάση Δεδομένων που σχεδιάστηκε στο Κεφάλαιο 3 για ένα κατάστημα ηλεκτρικών ειδών. Έστω ότι προηγήθηκε η διαδικασία σχεδιασμού του κεφαλαίου 3, με βάση την οποία προέκυψε η ανάγκη δημιουργίας των πινάκων ΠΕΛΑΤΕΣ, ΠΑΡΑΓΓΕΛΙΕΣ, ΠΡΟΪΟΝΤΑ, ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ, ΑΠΟΘΗΚΕΣ και ΠΟΛΕΙΣ, και επιθυμούμε να ξεκινήσουμε υλοποιώντας τον πίνακα ΠΡΟΪΟΝΤΑ. Εναλλακτικά, μπορεί να έχουμε αποφασίσει την ίδια σχεδίαση για αυτόν το πίνακα ακολουθώντας τα πρακτικά βήματα της προηγούμενης ενότητας.

ΠΡΟΪΟΝΤΑ
*Κωδικός_Προϊόντος: Κείμενο
Κατηγορία: Κείμενο
Μάρκα: Κείμενο
Μοντέλο: Κείμενο
Τιμή: Αριθμός
Κωδ_αποθήκης: Κείμενο

Μετά τη δημιουργία της νέας ΒΔ, η Access μας οδηγεί απευθείας στη δημιουργία του πρώτου πίνακα, όπως φαίνεται στο Σχήμα 4.5. Στον πίνακα αυτόν έχει δοθεί αυτόματα το πρόχειρο όνομα «Πίνακας1», εμφανίζεται σε προβολή φύλλου δεδομένων και έχει δημιουργηθεί αυτόματα ένα πρώτο πεδίο με όνομα «Αναγνωριστικό». Τα βήματα που ακολουθούμε είναι τα παρακάτω:

1. Επιλέγουμε να μεταφερθούμε σε προβολή σχεδίασης από το κουμπί Προβολή, στο επάνω αριστερά σημείο της κεντρικής καρτέλας (παρόλο που μπορούμε να συνεχίσουμε σε προβολή φύλλου δεδομένων, υπενθυμίζεται ότι συνιστάται να εργαζόμαστε σε προβολή σχεδίασης).
2. Πριν μεταβούμε σε προβολή σχεδίασης, η Access ζητάει να αποθηκεύσουμε τον πίνακα. Δίνουμε λοιπόν το όνομα ΠΡΟΪΟΝΤΑ και αποθηκεύουμε.

3. Στο περιβάλλον σχεδίασης καθορίζουμε ένα-ένα τα πεδία δίνοντας το όνομά τους, επιλέγοντας τον τύπο δεδομένων τους και καθορίζοντας (αν χρειάζεται) κάποια ειδικότερη ιδιότητα. Επίσης στη στήλη Περιγραφή μπορούμε προαιρετικά να εισάγουμε μια σημείωση σε φυσική γλώσσα, που αγνοείται από την Access και χρησιμεύει μόνο ως υπενθύμιση σε εμάς ή σε κάποιον άλλο χρήστη. Αρχικά δημιουργούμε το πεδίο «Κωδικός_προϊόντος» και επιλέγουμε ως τύπο δεδομένων «Κείμενο». (Επειδή στον πρώτο πίνακα η Access δημιουργεί αυτόματα ένα πεδίο «Αναγνωριστικό», το οποίο δε θέλουμε, είτε το διαγράφουμε είτε κάνουμε κλικ στο όνομα το πεδίου αυτού και το μετονομάζουμε στο επιθυμητό «Κωδικός_προϊόντος» και επίσης αλλάζουμε τον τύπο δεδομένων από Αυτόματη αρίθμηση σε κείμενο). Αφού εισάγουμε όλα τα πεδία, ορίζουμε το κατάλληλο πεδίο ως πρωτεύον κλειδί (π.χ. τον κωδικό προϊόντος). Αυτό γίνεται είτε κάνοντας διπλό κλικ στο επιθυμητό πεδίο και επιλέγοντας Πρωτεύον κλειδί από το μενού επιλογών που θα εμφανιστεί, είτε επιλέγοντας το πεδίο και πατώντας το κουμπί «Πρωτεύον κλειδί» στο αριστερό μέρος της καρτέλας εργαλείων «Σχεδίαση».



Σχήμα 4.6. *Κατασκευή ενός πίνακα σε προβολή σχεδίασης*

Οι **ιδιότητες πεδίου**, που φαίνονται στο κάτω μέρος της οθόνης προβολής σχεδίασης, είναι επιπλέον χαρακτηριστικά που μπορούμε να προσδιορίζουμε για κάθε πεδίο, ανάλογα με τις ανάγκες μας. Οι σημαντικότερες από αυτές είναι:

- **Μέγεθος πεδίου**. Προσδιορίζει το μέγιστο μέγεθος των δεδομένων που θα πρέπει να χωρέσουν σε ένα τέτοιο πεδίο. Η Access φροντίζει για την κατάλληλη πρόβλεψη σε αριθμό Bytes που θα αφιερώσει στην αποθήκευση του πεδίου για κάθε εγγραφή. Στην περίπτωση πεδίου κειμένου, το μέγεθος καθορίζει το μέγιστο αριθμό χαρακτήρων (προεπιλεγμένη τιμή

255 χαρακτήρες), ενώ σε περίπτωση αριθμητικού πεδίου καθορίζεται η ακρίβεια του αριθμού που θα μπορεί να παρασταθεί (προεπιλεγμένη τιμή Μεγάλος ακέραιος).

- **Μορφή.** Καθορίζει τον τρόπο προβολής του περιεχομένου ενός πεδίου. Στην περίπτωση αριθμού, μπορεί να επιλεγεί ο αριθμό ψηφίων, η εμφάνιση συμβόλου νομίσματος ή ποσοστού, κλπ.
- **Προεπιλεγμένη τιμή**. Μπορεί να καθοριστεί μια τιμή για το πεδίο που για λόγους ευκολίας να συμπληρώνεται αυτόματα όταν δημιουργείται μια νέα εγγραφή. Π.χ. αν γνωρίζουμε ότι σχεδόν όλοι οι πελάτες μας είναι από τη Θεσσαλονίκη, κατά τη δημιουργία ενός νέου πελάτη, στο πεδίο Πόλη μπορεί να συμπληρώνεται αυτόματα η τιμή «Θεσσαλονίκη» (φυσικά ο χρήστης μπορεί να τροποποιήσει αυτήν την τιμή).
- **Κανόνας επικύρωσης**. Είναι ένας κανόνας που μπορεί να περιορίζει τις τιμές που θα μπορούν να εισαχθούν σε ένα πεδίο, σύμφωνα με τη λογική της εφαρμογής, π.χ. ότι το ποσοστό έκπτωσης δε θα πρέπει να είναι μεγαλύτερο του 100% ή ότι η τιμή ενός προϊόντος θα πρέπει να μην είναι αρνητική. Ο κανόνας μπορεί να είναι μια έκφραση που συντάσσεται με τη βοήθεια ενός ειδικού οδηγού. Ο κανόνας επικύρωσης είναι σημαντικός μηχανισμός που βελτιώνει την εγκυρότητα των δεδομένων. Το **Κείμενο επικύρωσης** είναι το μήνυμα που εμφανίζει η Access αν παραβιαστεί ο κανόνας επικύρωσης.
- **Απαιτείται.** Μπορεί να επιλεγούν οι τιμές Ναι ή Όχι ανάλογα με το αν είναι απαραίτητο ή όχι να συμπληρωθεί οπωσδήποτε κάποια τιμή στο πεδίο, ώστε να είναι έγκυρη μια εγγραφή. Π.χ. αν επιλεγεί το Ναι στο πεδίο Επώνυμο του πίνακα ΠΕΛΑΤΕΣ, δε θα επιτρέπεται να εισαχθεί κάποιος πελάτης αν δε συμπληρωθεί το επώνυμό του.

Σημειώνεται ότι οι ιδιότητες που είναι διαθέσιμες για κάθε πεδίο εξαρτώνται από τον τύπο δεδομένων που έχει επιλεγεί για το πεδίο αυτό, π.χ. για τα πεδία τύπου Αριθμός εμφανίζεται η ιδιότητα Δεκαδικές θέσεις, η οποία δε διατίθεται σε πεδία τύπου κειμένου.

Σε απλές εφαρμογές, συνήθως δε χρειάζεται να καταβάλουμε ιδιαίτερη προσπάθεια στον καθορισμό όλων των ιδιοτήτων για όλα τα πεδία, αφού οι προεπιλεγμένες τιμές είναι συνήθως κατάλληλες. Οφείλουμε όμως να γνωρίζουμε το ρόλο τους γιατί σε ορισμένες περιπτώσεις η λάθος επιλογή ιδιότητας δημιουργεί προβλήματα. Π.χ. στο πεδίο **Τιμή** του πίνακα **ΠΡΟΪΟΝΤΑ** πρέπει η ιδιότητα μέγεθος πεδίου να αλλάξει από **Μεγάλος ακέραιος** σε **Πραγματικός απλής ακρίβειας**, αλλιώς οι τιμές που θα εισάγονται θα μετατρέπονται σε ακέραιους και θα χάνονται τα δεκαδικά ψηφία.

Μετά την ολοκλήρωση της δημιουργίας του πίνακα **ΠΡΟΪΟΝΤΑ**, μπορούμε με την ίδια ακριβώς διαδικασία να δημιουργήσουμε και τους υπόλοιπους πίνακες του σχεδίου, συγκεκριμένα τους πίνακες **ΠΕΛΑΤΕΣ, ΠΑΡΑΓΓΕΛΙΕΣ, ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ, ΑΠΟΘΗΚΕΣ, ΠΟΛΕΙΣ.**

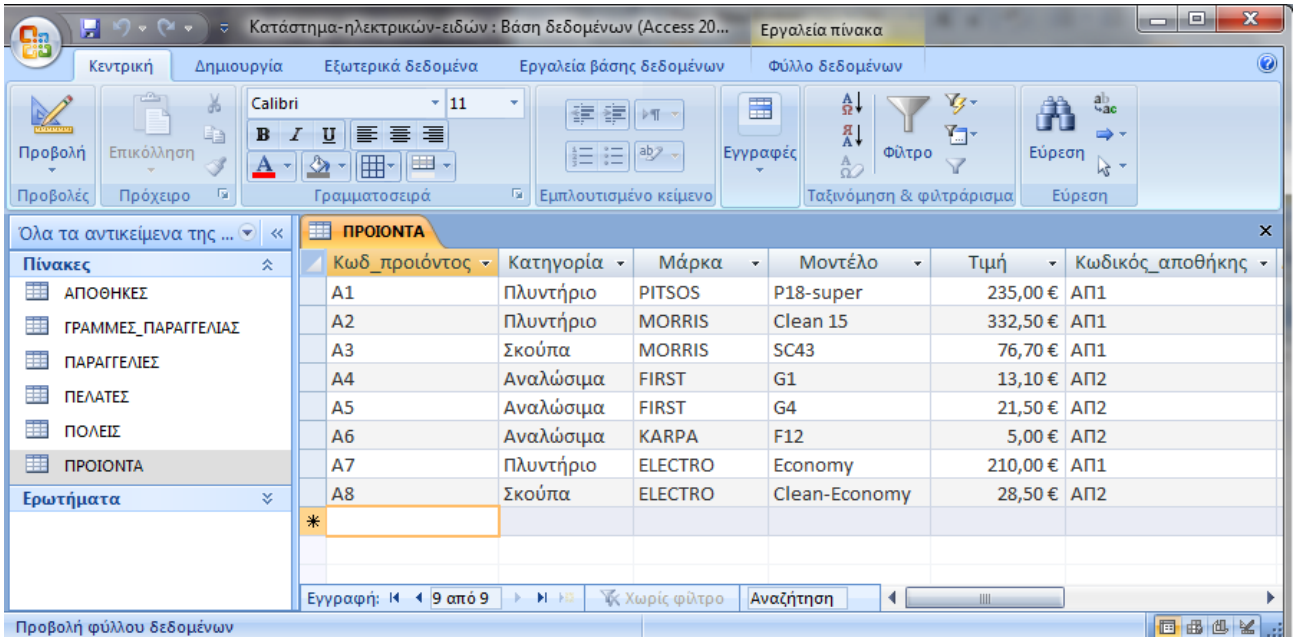
4.3.3 Εισαγωγή δεδομένων

Μετά την ολοκλήρωση της δημιουργίας των πινάκων, ο χρήστης μπορεί να εισάγει, να τροποποιήσει ή να διαγράψει δεδομένα. Στο Σχήμα 4.7 φαίνεται ο πίνακας **ΠΡΟΪΟΝΤΑ** σε προβολή φύλλου δεδομένων. Παρατηρήστε ότι όλες οι στήλες του πίνακα αντιστοιχούν ακριβώς στα πεδία που έχουν οριστεί κατά τη σχεδίαση του πίνακα.

Κάθε γραμμή του πίνακα αντιστοιχεί σε μια εγγραφή δηλαδή σε ένα στιγμιότυπο της οντότητας του πίνακα (π.χ. σε ένα συγκεκριμένο προϊόν). Ένας κενός πίνακας δεν έχει καμία εγγραφή, παρά μόνο μία γραμμή εισαγωγής, που σημειώνεται με έναν αστερίσκο (*), όπου ο χρήστης μπορεί να δημιουργήσει μια νέα εγγραφή.

Κατά την εισαγωγή ή τροποποίηση δεδομένων, η Access ελέγχει τη συμβατότητα με τους τύπους δεδομένων που έχουν οριστεί κατά τη σχεδίαση και επιβάλλει την εφαρμογή των κανόνων εγκυρότητας.

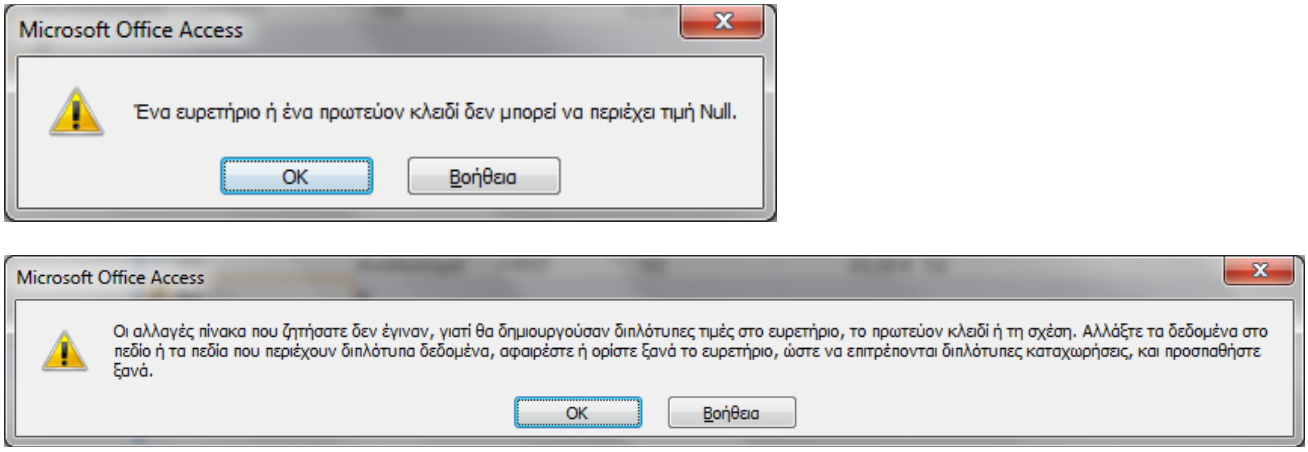
Έτσι, σε πεδίο τύπου π.χ. **Αριθμός** δεν επιτρέπεται η εισαγωγή διαφορετικών δεδομένων από αριθμητικά. Επίσης δεν επιτρέπεται να εισαχθεί κάτι που ξεπερνάει το μέγεθος πεδίου ούτε να μείνει κενό κάποιο πεδίο για το οποίο Απαιτείται να πάρει κάποια τιμή.



Σχήμα 4.7. *Ο πίνακας ΠΡΟΪΟΝΤΑ σε προβολή φύλλου δεδομένων, όπου είναι δυνατή η εισαγωγή/τροποποίηση του περιεχομένου του.*

Προσοχή: Η Access ελέγχει αυστηρά την εγκυρότητα του πρωτεύοντος κλειδιού. Αν κατά την εισαγωγή μιας εγγραφής, το πεδίο που έχει οριστεί ως πρωτεύον κλειδί παραμείνει κενό ή δεν περιέχει έγκυρη τιμή (π.χ. εισαχθεί μια τιμή που υπάρχει ήδη σε άλλη εγγραφή), η Access εμφανίζει αντίστοιχο μήνυμα λάθους (Σχήμα 4.8) και δεν επιτρέπει στο χρήστη να συνεχίσει με οποιαδήποτε άλλη ενέργεια αν προηγουμένως δεν εισάγει έγκυρη τιμή στο πρωτεύον κλειδί για την εγγραφή στην οποία βρίσκεται.

Συνιστάται η εισαγωγή δεδομένων να πραγματοποιείται ολοκληρώνοντας μία-μία τις εγγραφές (δηλαδή κατά σειρές) και όχι συμπληρώνοντας ένα-ένα τα πεδία (δηλαδή κατά στήλες) γιατί στη δεύτερη περίπτωση δημιουργούνται, έστω και πρόσκαιρα, πολλές ελλιπείς εγγραφές, που μπορεί να δημιουργήσουν προβλήματα εγκυρότητας.



Σχήμα 4.8. *Αν ο χρήστης επιχειρήσει να δημιουργήσει εγγραφή χωρίς έγκυρη τιμή στο πρωτεύον κλειδί δεν μπορεί να προχωρήσει σε καμία άλλη ενέργεια αν δε διορθώσει την εγγραφή στην οποία βρίσκεται.*

Σημείωση: Η σειρά με την οποία εισάγονται και προβάλλονται οι εγγραφές ενός πίνακα δεν παίζει κανένα ρόλο. Για αυτόν το λόγο, δεν έχει νόημα και δε γίνεται η μετακίνηση σειρών σε άλλη θέση. Σε έναν πίνακα

Βάσης Δεδομένων αυτό που μας ενδιαφέρει είναι το αν περιλαμβάνεται ή όχι μια εγγραφή και ποιο είναι το περιεχόμενό της.

Σημείωση: Η πλήρης διαγραφή μιας εγγραφής επιτρέπεται από την Access, αλλά κατά τη χρήση της Βάσης Δεδομένων σε μια εφαρμογή πρέπει κάτι τέτοιο να γίνεται με προσοχή και μόνο σε ειδικές περιπτώσεις. Όταν επιθυμούμε να αφαιρέσουμε ένα στιγμιότυπο από έναν πίνακα, η συνήθης πρακτική είναι να σημειώνεται ως διαγραμμένο (ή άκυρο) και να παραμένει η εγγραφή στον πίνακα, χωρίς να διαγράφεται πραγματικά.

Π.χ. αν η εταιρεία σταματήσει να διακινεί ένα προϊόν, η αντίστοιχη εγγραφή στον πίνακα προϊόντα δε θα πρέπει να διαγραφεί τελείως σαν να μην υπήρξε ποτέ αυτό το προϊόν, αλλά μέσω του κατάλληλου πεδίου, να σημειωθεί ότι το προϊόν έχει καταργηθεί.

4.4 Δημιουργία Ερωτημάτων

4.4.1 Χρησιμότητα και τύποι ερωτημάτων

Τα ερωτήματα χρησιμοποιούνται για να ανασύρουμε ή να τροποποιήσουμε με αυτοματοποιημένο τρόπο δεδομένα, αναζητώντας και επιλέγοντάς τα από έναν πίνακα ή συγκεντρώνοντάς τα από περισσότερους πίνακες. Με τα ερωτήματα μπορούμε επίσης να κάνουμε υπολογισμούς πάνω στα κύρια δεδομένα και να παράγουμε δευτερογενείς πληροφορίες.

Τα ερωτήματα είναι ο μηχανισμός που διαθέτει η Access για να έχουμε εύκολη πρόσβαση στα δεδομένα και να αντλούμε χρήσιμες πληροφορίες από τα δεδομένα αυτά. Για να γίνει απολύτως κατανοητή η χρησιμότητα των ερωτημάτων, αρκεί να θυμηθούμε από το προηγούμενο κεφάλαιο ότι όλα τα δεδομένα είναι κατανεμημένα σε πίνακες που αντιστοιχούν σε διαφορετική οντότητα ο καθένας και σχετίζονται μεταξύ τους με τη χρήση κλειδιών.

Έτσι, π.χ. τα στοιχεία ενός πελάτη βρίσκονται στον πίνακα **ΠΕΛΑΤΕΣ**, ενώ τα στοιχεία των παραγγελιών του βρίσκονται στον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ**. Πώς θα προβάλουμε όλα τα στοιχεία που μας ενδιαφέρουν για μια συγκεκριμένη παραγγελία που δόθηκε π.χ. σήμερα από ένα συγκεκριμένο πελάτη; Είναι βέβαιο ότι ο χρήστης δεν είναι δυνατόν να ψάχνει μέσα στους πίνακες για να βρει αυτό που θέλει, αλλά χρειαζόμαστε ένα μηχανισμό που:

- να επιλέγει τις κατάλληλες εγγραφές με κάποια κριτήρια (π.χ. τη συγκεκριμένη παραγγελία που δόθηκε σήμερα)
- να προβάλλει μόνο τα πεδία που μας ενδιαφέρουν από συγκεκριμένους πίνακες (π.χ. μόνο το όνομα και τηλέφωνο του πελάτη που έδωσε την παραγγελία και όχι όλα τα στοιχεία που κρατάμε για αυτόν)
- να συνδέει σωστά κατάλληλες εγγραφές από διαφορετικούς πίνακες (π.χ. να βρίσκει τα στοιχεία του συγκεκριμένου πελάτη που έδωσε μια συγκεκριμένη παραγγελία και όχι οποιουδήποτε άλλου)

Ο πιο κοινός τύπος ερωτήματος είναι το ερώτημα επιλογής. Είναι ένα ερώτημα που αναζητά δεδομένα από έναν ή περισσότερους πίνακες, χρησιμοποιώντας κριτήρια που μπορούν να οριστούν, και τα εμφανίζει με τη μορφή που επιθυμούμε.

Το αποτέλεσμα που παρέχει ένα ερώτημα επιλογής είναι και αυτό ένας πίνακας που περιλαμβάνει, όπως και οι κανονικοί πίνακες, κάποια πεδία (στήλες) και κάποιες εγγραφές (γραμμές).

Επίσης είναι δυνατή η χρήση ενός ερωτήματος επιλογής για να ομαδοποιήσουμε εγγραφές και να υπολογίσουμε αθροίσματα, πλήθη, μέσους όρους και άλλους τύπους σύνοψης δεδομένων.

Ένας άλλος τύπος ερωτήματος που χρησιμοποιείται συχνά είναι το ερώτημα Ενημέρωσης. Ένα τέτοιο ερώτημα χρησιμεύει στην αυτόματη τροποποίηση των περιεχομένων ενός πίνακα. Π.χ. αν θέλουμε να αυξήσουμε την τιμή όλων των προϊόντων μιας συγκεκριμένης εταιρείας κατά 10%, ένα κατάλληλο ερώτημα ενημέρωσης μπορεί να επιλέξει τις εγγραφές που αντιστοιχούν στα προϊόντα της συγκεκριμένης εταιρείας και σε αυτά να τροποποιήσει το περιεχόμενο του πεδίου **Τιμή** με βάση την επιθυμητή αριθμητική έκφραση.

Οι κύριοι τύποι ερωτημάτων που διαθέτει η Access απαριθμούνται στον Πίνακα 4.1

Τύπος ερωτήματος	**Περιγραφή**
Επιλογής	Επιλογή και εμφάνιση εγγραφών από έναν ή περισσότερους πίνακες της ΒΔ.
Ενημέρωσης	Ενημέρωση/τροποποίηση του περιεχομένου ενός πίνακα.
Δημιουργίας πίνακα	Επιλογή εγγραφών ενός πίνακα και αποθήκευση των εγγραφών αυτών σε νέο πίνακα.
Προσάρτησης	Επιλογή εγγραφών ενός πίνακα και προσάρτηση των εγγραφών αυτών σε έναν άλλο υπάρχοντα πίνακα.
Διαγραφής	Διαγραφή επιλεγμένων δεδομένων από έναν υπάρχοντα πίνακα με βάση κριτήρια.
Διασταύρωσης	Συγκέντρωση δεδομένων και χρήση των συγκεντρωτικών στοιχείων για τη δημιουργία των γραμμών και των στηλών ενός νέου πίνακα διασταύρωσης.

Πίνακας 4.1 *Τύποι ερωτημάτων της Access.*

4.4.2 Δημιουργία ερωτήματος επιλογής

Η δημιουργία ενός ερωτήματος γίνεται από την καρτέλα Δημιουργία, επιλέγοντας Οδηγός Ερωτημάτων ή Σχεδίαση ερωτήματος. Συνιστάται η Σχεδίαση Ερωτήματος επειδή δίνει καλύτερο έλεγχο και προσφέρει πλήρεις δυνατότητες σχεδιασμού. Η διαδικασία που ακολουθούμε στη συνέχεια είναι η ακόλουθη:

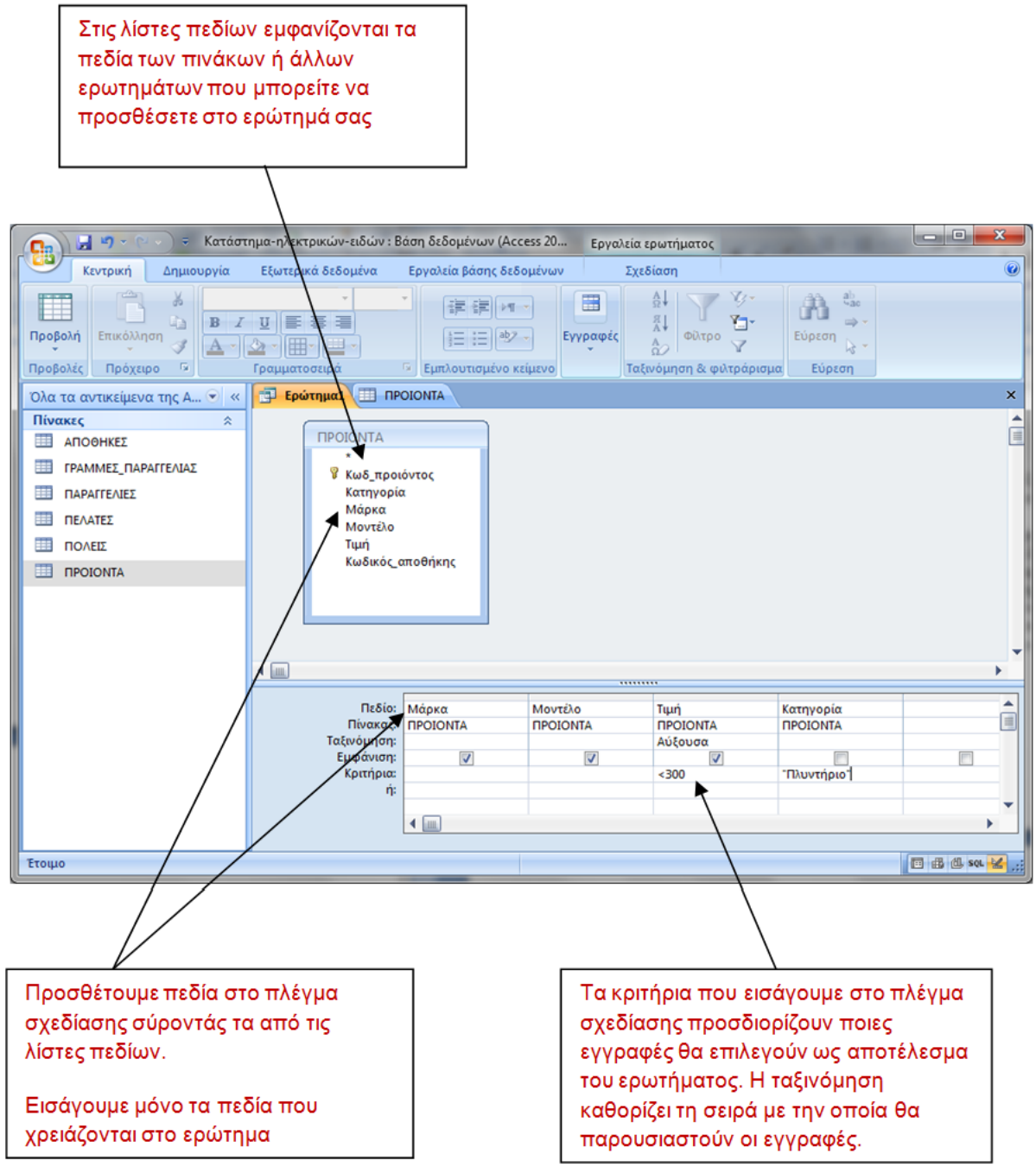
1. Προσθήκη πίνακα σε ένα ερώτημα. Επιλέγουμε τον πίνακα ή τους πίνακες από τους οποίους θα αντλήσουμε τα δεδομένα που χρειαζόμαστε.
2. Προσθήκη πεδίων σε ένα ερώτημα. Εισάγουμε ένα-ένα στις στήλες του πλέγματος σχεδίασης τα πεδία των πινάκων που αφορούν το ερώτημα, δηλαδή τα πεδία που θέλουμε να εμφανίζονται στο αποτέλεσμα και αυτά που είναι απαραίτητα για να γίνει η επιλογή των εγγραφών.
3. Ορισμός κριτηρίων ερωτήματος. Συμπληρώνουμε το κριτήριο ή τα κριτήρια που επιθυμούμε στα κατάλληλα πεδία, ώστε να γίνεται η επιλογή συγκεκριμένων πληροφοριών (εγγραφών).
4. Εκτέλεση υπολογισμών σε ερώτημα. Μπορούμε να προβάλουμε σύνθετα δεδομένα, εκτελώντας πράξεις στις τιμές που περιέχονται στους πίνακες.
5. Έλεγχος και βελτίωση ερωτήματος

Η διαδικασία παρουσιάζεται στη συνέχεια μέσω 2 παραδειγμάτων, ενός απλού ερωτήματος που βασίζεται σε έναν μόνο πίνακα και ενός πιο σύνθετου.

4.4.2.1 Απλά ερωτήματα ενός πίνακα

Έστω ότι θέλουμε να δημιουργήσουμε ένα ερώτημα που να εμφανίζει όλα τα προϊόντα κατηγορίας «Πλυντήριο» που διαθέτει το κατάστημά μας και έχουν τιμή μικρότερη από 300€. Για τα προϊόντα αυτά μας ενδιαφέρει η μάρκα, το μοντέλο και η τιμή τους.

Ξεκινώντας τη σχεδίαση ερωτήματος, η Access εμφανίζει ένα παράθυρο με τίτλο Εμφάνιση Πίνακα μέσω του οποίου επιλέγουμε τον πίνακα από τον οποίο θα αντληθούν τα δεδομένα. Επιλέγουμε τον πίνακα ΠΡΟΪΟΝΤΑ, πατάμε Προσθήκη και κλείνουμε το παράθυρο. Έτσι, δημιουργείται ένα νέο ερώτημα το οποίο μπορούμε να διαμορφώσουμε. Βρισκόμαστε σε προβολή σχεδίασης (Σχήμα 4.9) που περιλαμβάνει ένα χώρο εμφάνισης των πινάκων που συμμετέχουν στο ερώτημα (επάνω) και ένα πλέγμα σχεδίασης (κάτω).



Σχήμα 4.9. *Δημιουργία απλού ερωτήματος*

Για την υλοποίηση του παραπάνω παραδείγματος, κάνουμε τις εξής ενέργειες:

- Στις στήλες του πλέγματος σχεδίασης εισάγουμε ένα-ένα τα πεδία του πίνακα που αφορούν το ερώτημα, δηλαδή τα πεδία που θέλουμε να εμφανίζονται στο αποτέλεσμα και αυτά που είναι απαραίτητα για να γίνει η επιλογή των εγγραφών. Εισάγουμε τα πεδία του πίνακα **ΠΡΟΪΟΝΤΑ: Μάρκα, Μοντέλο, Τιμή** και **Κατηγορία**. Τα 3 πρώτα πεδία είναι αυτά που μας δίνουν τις πληροφορίες που χρειαζόμαστε στο αποτέλεσμα, ενώ το πεδίο **Κατηγορία** συμπεριλαμβάνεται γιατί με βάση αυτό θα επιλεγούν τα προϊόντα που αφορούν την κατηγορία «Πλυντήριο».

- Στο πεδίο **Τιμή** εισάγουμε ως κριτήριο την έκφραση **<300.** Με τον τρόπο αυτό θα επιλεγούν μόνο οι εγγραφές για τις οποίες το πεδίο **Τιμή** περιέχει έναν αριθμό για τον οποίο ισχύει **τιμή < 300.** Επίσης στο πεδίο **Κατηγορία** εισάγουμε ως κριτήριο το κείμενο **"Πλυντήριο"**, ώστε να επιλεγούν οι εγγραφές για τις οποίες το περιεχόμενο του πεδίου Κατηγορία είναι η λέξη «Πλυντήριο».
- Αν επιθυμούμε τα αποτελέσματα να εμφανίζονται ταξινομημένα σύμφωνα με την τιμή τους, με πρώτα τα φθηνότερα, στη στήλη του πεδίου **Τιμή** και στη γραμμή **Ταξινόμηση** επιλέγουμε **Αύξουσα.**
- Επειδή το πεδίο **Κατηγορία** χρησιμοποιήθηκε για την εφαρμογή του κριτηρίου επιλογής και δε μας ενδιαφέρει να δούμε το περιεχόμενό του στο αποτέλεσμα του ερωτήματος (γνωρίζουμε από πριν ότι όλες οι εγγραφές που θα προκύψουν στο αποτέλεσμα θα είναι κατηγορίας Πλυντήριο), μπορούμε να από-επιλέξουμε την **Εμφάνιση** για αυτό το πεδίο.
- Αποθηκεύουμε το ερώτημα με το όνομα που επιθυμούμε π.χ. «Φθηνά πλυντήρια». Μπορούμε επίσης να το εκτελέσουμε ώστε να βεβαιωθούμε ότι λειτουργεί σωστά, επιλέγοντας προβολή φύλλου δεδομένων ή πατώντας Εκτέλεση (επάνω αριστερά στην καρτέλα εντολών Σχεδίαση).

Κατά τη δοκιμαστική εκτέλεση του ερωτήματος (Σχήμα 4.10) παρατηρούμε ότι:

- Το αποτέλεσμα είναι σε μορφή πίνακα με στήλες **Μάρκα, Μοντέλο, Τιμή**, δηλαδή εμφανίζονται τα πεδία που εισάγαμε στο πλέγμα σχεδίασης, εκτός από το πεδίο **Κατηγορία** για το οποίο απενεργοποιήσαμε την εμφάνιση.
- Προκύπτει ένα μόνο προϊόν (PITSOS, P18-super, 235€). Αν παρατηρήσουμε το περιεχόμενο του πίνακα **ΠΡΟΪΟΝΤΑ** (Σχήμα 4.7) διαπιστώνουμε ότι το αποτέλεσμα του ερωτήματος είναι αυτό που αναμενόταν, δηλαδή επιλέχθηκε το μόνο προϊόν της κατηγορίας Πλυντήριο με τιμή μικρότερη από 300€.

Σημείωση: Τα αποτελέσματα του ερωτήματος εξαρτώνται από τα δεδομένα που περιέχονται τη στιγμή της εκτέλεσής του στη Βάση Δεδομένων. Επομένως το γεγονός ότι εμφανίστηκαν τα αποτελέσματα που περιμέναμε δεν είναι απόδειξη ότι όλα είναι σωστά, αλλά μια επαλήθευση. Αντίστοιχα, αν δεν εμφανιστεί κανένα αποτέλεσμα επειδή απλά καμία εγγραφή δεν πληρούσε τα κριτήρια ή ήταν κενός ο πίνακας, δε θα πρέπει να θεωρηθεί ως πρόβλημα.

Η χρήση των κριτηρίων στα ερωτήματα

Τα κριτήρια είναι περιορισμοί που θέτουμε σε ένα ερώτημα ή σε ένα σύνθετο φίλτρο για να προσδιοριστούν οι συγκεκριμένες εγγραφές που θέλουμε να επιλεγούν. Για παράδειγμα, αντί να εμφανίζονται όλα τα προϊόντα της εταιρείας, θέλουμε να περιοριστούμε σε αυτά που ανήκουν στην κατηγορία «Πλυντήριο». Στο παραπάνω ερώτημα συμπεριλάβαμε δύο κριτήρια σε δύο διαφορετικά πεδία (τιμή μικρότερη από **300€** και κατηγορία **Πλυντήριο**). Τα κριτήρια αυτά επιβλήθηκαν από την Access συζευκτικά, θεωρώντας ότι είναι δύο συνθήκες που συνδέονται με το λογικό ΚΑΙ (AND). Με απλά λόγια, για να επιλεγεί ένα προϊόν πρέπει και η τιμή του να είναι μικρότερη από 300€ και η κατηγορία του να είναι Πλυντήριο. Η λογική σύνδεση των κριτηρίων με το ΚΑΙ ισχύει γενικά για όλα τα κριτήρια που θα εισαχθούν σε διαφορετικά πεδία στο πλέγμα σχεδίασης. Αν θέλαμε να συμπεριλάβουμε κάποιο κριτήριο που να συνδέεται διαζευκτικά, δηλαδή με το λογικό Η (OR), θα πρέπει να το εισάγουμε στη γραμμή με τίτλο **ή:** (ακριβώς κάτω από τη γραμμή **Κριτήρια:**). Π.χ. αν επιθυμούμε να εμφανίζονται είτε τα παραπάνω προϊόντα είτε αυτά που το μοντέλο τους είναι τύπου «Economy», μπορούμε να προσθέσουμε ως επιπλέον κριτήριο τη λέξη **"Economy"** στη θέση που αντιστοιχεί στο πεδίο **Μοντέλο** και στη γραμμή **ή:.**

Τα κριτήρια χρειάζεται συχνά να είναι πιο σύνθετα από τις παραπάνω απλές περιπτώσεις, όπως να περιλαμβάνουν υπολογισμούς ή λογικές εκφράσεις. Για να καθοριστούν κριτήρια για ένα πεδίο στο πλέγμα σχεδίασης, καταχωρούμε μια παράσταση στο κελί **Κριτήρια** για το συγκεκριμένο πεδίο. Η παράσταση στο προηγούμενο παράδειγμα είναι "Πλυντήριο" και αφορά μια συγκεκριμένη τιμή για την οποία εννοείται το = (ίσον). Μπορούμε όμως να χρησιμοποιήσουμε και πιο σύνθετες παραστάσεις που να περιλαμβάνουν σύνολα τιμών, διαστήματα και ελέγχους.

Μπορούμε να εισάγουμε επιπλέον κριτήρια για το ίδιο πεδίο ή για διαφορετικά πεδία. Όταν εισάγονται παραστάσεις σε περισσότερα από ένα κελιά **Κριτήρια**, η Access τις συνδυάζει χρησιμοποιώντας τους τελεστές **And** ή **Or**. Εάν οι παραστάσεις βρίσκονται σε διαφορετικά κελιά στην ίδια γραμμή, η Access χρησιμοποιεί τον τελεστή **And**, που σημαίνει ότι θα επιστραφούν μόνο οι εγγραφές που ανταποκρίνονται σε όλα τα κριτήρια ταυτόχρονα. Εάν οι παραστάσεις είναι σε διαφορετικές σειρές του πλέγματος σχεδίασης, η Access χρησιμοποιεί τον τελεστή **Or**, που σημαίνει ότι θα επιστραφούν οι εγγραφές που ανταποκρίνονται σε τουλάχιστον ένα από τα κριτήρια.

Σε μια παράσταση στο κελί **Κριτήρια** του κατάλληλου πεδίου, μπορεί να προσδιοριστεί ένα διάστημα χρησιμοποιώντας τον τελεστή **Between...And** ή τους τελεστές σύγκρισης (<, >, <>, <= και >=). Για παράδειγμα, μπορείτε να βρείτε τις παραγγελίες που έγιναν μέσα στο έτος 2015 εισάγοντας την έκφραση

Between #1/1/2015# And #31/12/2015#

ή εναλλακτικά την έκφραση

>= #1/1/2015# And <= #31/12/2015#

Παράσταση	**Έννοια**
>234	Αριθμοί μεγαλύτεροι από 234
Between #2/2/93# And #12/1/93#	Ημερομηνίες από 2-Φεβ-93 έως 1-Δεκ-93
<1200.45	Αριθμοί μικρότεροι από το 1200.45
>="K"	Όλα τα ονόματα που ταξινομούνται αλφαβητικά από το γράμμα K και μετά, μέχρι το τελευταίο γράμμα του αλφαβήτου
"P18*"	Όλα τα ονόματα που ξεκινούν από τα γράμματα P18. Το σύμβολο * λέγεται χαρακτήρας μπαλαντέρ και ταιριάζει με οποιαδήποτε σειρά χαρακτήρων οποιουδήποτε μήκους.
"G??-super"	Όλα τα ονόματα που ξεκινούν από G, στη συνέχεια περιέχουν ακριβώς 2 οποιουσδήποτε χαρακτήρες και μετά τους χαρακτήρες –super. Το σύμβολο ? είναι χαρακτήρας μπαλαντέρ που ταιριάζει με οποιονδήποτε χαρακτήρα, αλλά μόνο έναν.
"ΚΟΚΚΙΝΟ" Or "ΠΡΑΣΙΝΟ"	Είτε ΚΟΚΚΙΝΟ είτε ΠΡΑΣΙΝΟ

Πίνακας 4.2. *Παραδείγματα παραστάσεων και τελεστών που χρησιμοποιούνται ως κριτήρια.*

4.4.2.2 Ερωτήματα που συνδυάζουν δεδομένα από περισσότερους πίνακες ή ερωτήματα

Η ισχύς των ερωτημάτων βρίσκεται στη δυνατότητα που έχουν να συγκεντρώνουν ή να εκτελούν ενέργειες με δεδομένα από περισσότερους από έναν πίνακες ή άλλα ερωτήματα. Για παράδειγμα, μπορούμε να εμφανίσουμε τις πληροφορίες ενός πελάτη μαζί με τις παραγγελίες που έθεσε ο πελάτης. Για να εμφανιστούν αυτές οι πληροφορίες, χρειάζονται δεδομένα από ένα πίνακα Πελατών και ένα πίνακα Παραγγελιών.

Όταν προσθέτουμε περισσότερους από έναν πίνακες ή ερωτήματα σε ένα ερώτημα, πρέπει να βεβαιωθούμε ότι είναι συνδεδεμένες οι λίστες πεδίων τους με μια γραμμή σύνδεσης, έτσι ώστε η Access να γνωρίζει με ποιο τρόπο θα συνδυάσει τις πληροφορίες.

Προσοχή: Η σύνδεση (ή σχέση) δύο πινάκων πρέπει να γίνει με βάση το σωστό πεδίο δηλ. το πεδίο που περιέχει κοινή πληροφορία και έχει προβλεφθεί για να συνδέει λογικά τους δύο πίνακες. Π.χ. ο πίνακας **ΠΕΛΑΤΕΣ** και ο πίνακας **ΠΑΡΑΓΓΕΛΙΕΣ** πρέπει να ενωθούν με βάση το πεδίο **Κωδ_πελάτη**, που είναι το πρωτεύον κλειδί στον πίνακα **ΠΕΛΑΤΕΣ** και το πεδίο **Κωδικός_Πελάτη** που υπάρχει στον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ** ως «ξένο» κλειδί για να προσδιορίζει τον πελάτη που έδωσε την κάθε παραγγελία.

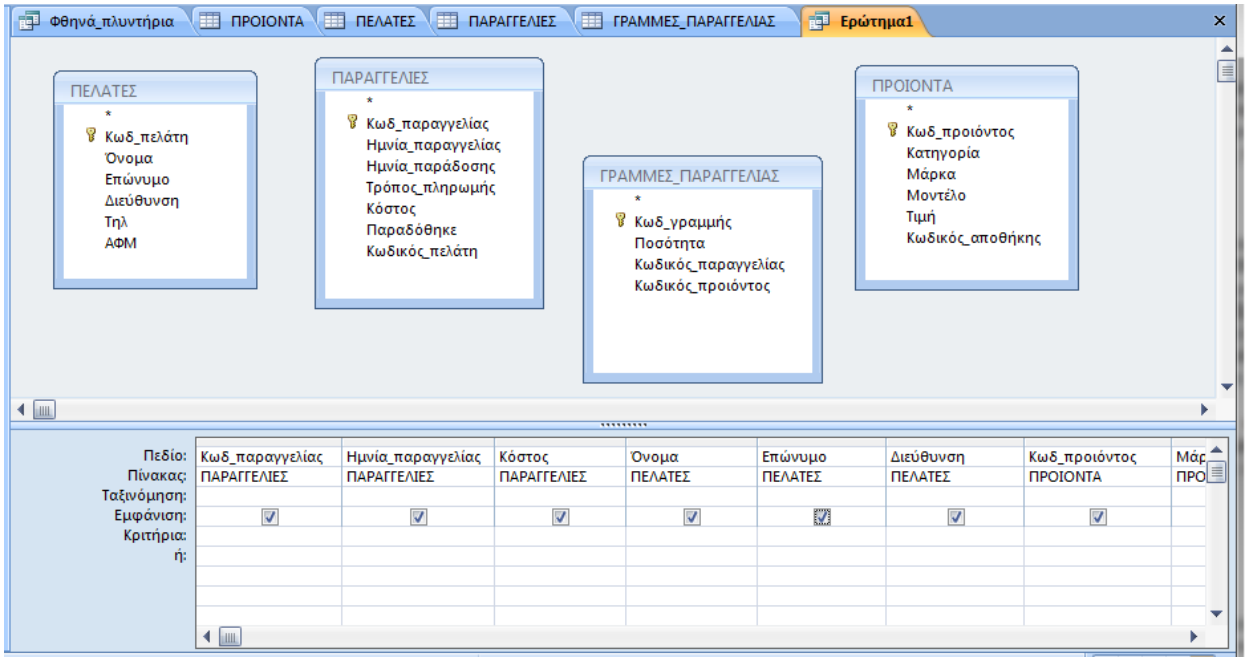
Σημείωση: Εάν οι πίνακες σε ένα ερώτημα δεν είναι ενωμένοι μεταξύ τους, άμεσα ή έμμεσα, η Access δε γνωρίζει τις σχέσεις μεταξύ των εγγραφών και συνεπώς εμφανίζει όλους τους συνδυασμούς

εγγραφών μεταξύ των δύο πινάκων (αυτό ονομάζεται «γινόμενο διασταύρωσης» ή «Καρτεσιανό γινόμενο»), το οποίο είναι εμφανώς λανθασμένο αποτέλεσμα.

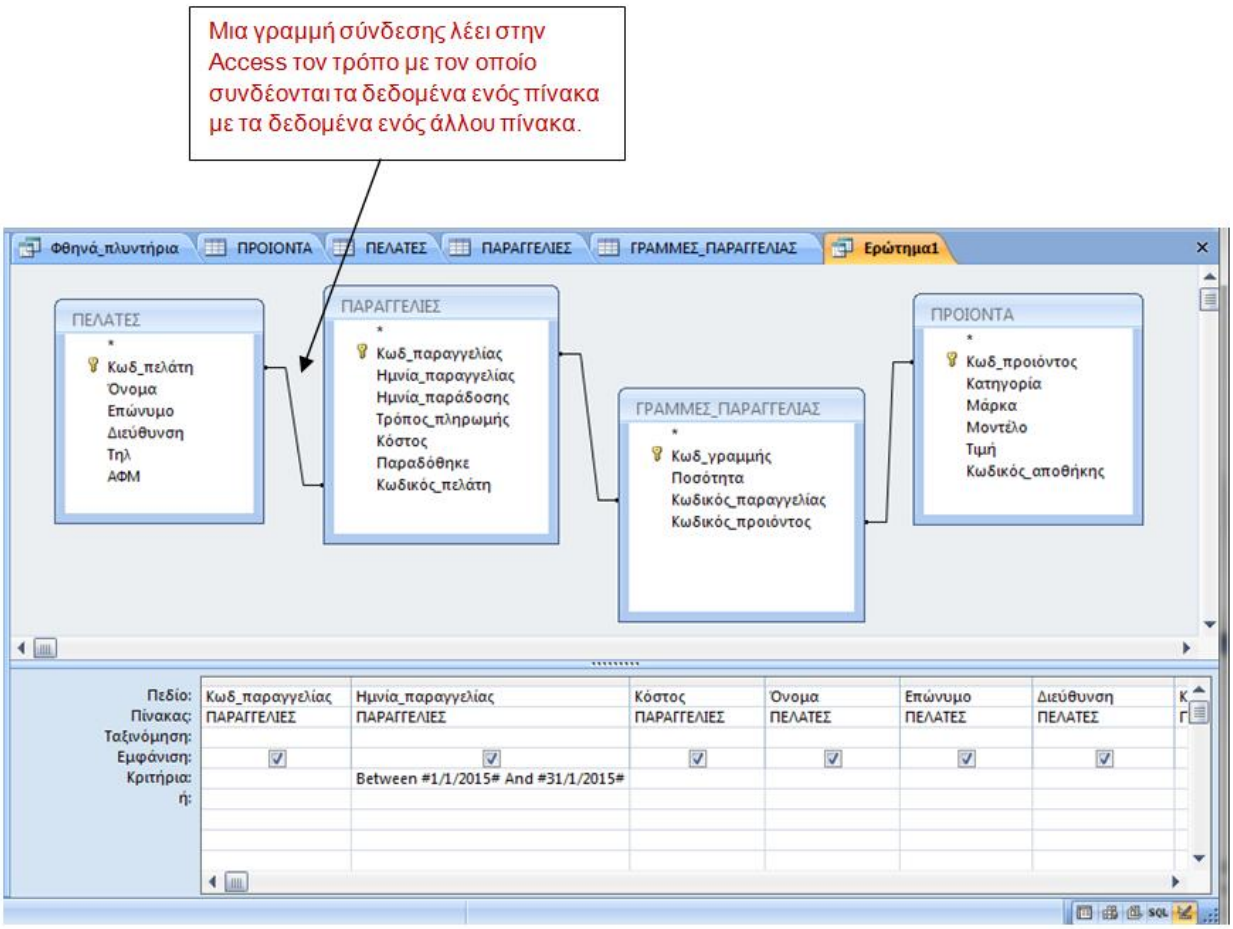
Για την παρουσίαση της δημιουργίας ερωτήματος πολλαπλών πινάκων χρησιμοποιούμε το εξής παράδειγμα: Για το κατάστημα ηλεκτρικών ειδών θα δημιουργήσουμε ένα ερώτημα που να αναζητά τις παραγγελίες που δόθηκαν τον Ιανουάριο 2015 και να μας εμφανίζει για αυτές το ονοματεπώνυμο και διεύθυνση του πελάτη, το συνολικό κόστος της παραγγελίας καθώς και τα είδη και τις ποσότητες τους που περιλαμβάνονται σε κάθε παραγγελία.

Για την υλοποίηση του παραπάνω παραδείγματος, κάνουμε τις εξής ενέργειες:

- Επιλέγουμε την καρτέλα **Δημιουργία** και από εκεί **Σχεδίαση ερωτήματος**. Από το παράθυρο **Εμφάνιση Πίνακα** επιλέγουμε και προσθέτουμε τους πίνακες **ΠΑΡΑΓΓΕΛΙΕΣ, ΠΕΛΑΤΕΣ, ΠΡΟΪΟΝΤΑ** και **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ**. Οι πίνακες αυτοί επιλέχθηκαν επειδή στις ΠΑΡΑΓΓΕΛΙΕΣ περιέχεται η πληροφορία σχετικά με το ποιες παραγγελίες δόθηκαν και το συνολικό κόστος τους, στους ΠΕΛΑΤΕΣ βρίσκονται τα στοιχεία των πελατών (ονοματεπώνυμο, διεύθυνση), στα ΠΡΟΪΟΝΤΑ βρίσκονται τα ονόματα των προϊόντων και στις ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ βρίσκεται μέσω των κατάλληλων κωδικών η πληροφορία σχετικά με τα επιμέρους προϊόντα και τις ποσότητές τους που περιλαμβάνονται σε κάθε παραγγελία.
- Προσθέτουμε στο πλέγμα σχεδίασης τα πεδία που χρειάζονται από κάθε πίνακα: από τον **ΠΑΡΑΓΓΕΛΙΕΣ** τα πεδία **Κωδ_Παραγγελίας**, **Ημνια_Παραγγελίας** και **Κόστος**, από τον **ΠΕΛΑΤΕΣ** τα **Όνομα, Επώνυμο**, **Διεύθυνση**, από τον **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ** το **Ποσότητα** και από τον **ΠΡΟΪΟΝΤΑ** τα **Κωδ_προιόντος**, **Μάρκα**, **Μοντέλο** (Σχήμα 4.10).
- Συνδέουμε τους πίνακες σύμφωνα με τη λογική της σχεδίασης που έχουμε κάνει. Δύο πίνακες συνδέονται μεταξύ τους όταν σύμφωνα με τη λογική μας υπάρχει σχέση ανάμεσά τους και έχει γίνει πρόβλεψη για την ύπαρξη ενός κοινού πεδίου σύνδεσης. Για τη σύνδεση σύρουμε ένα πεδίο από τη λίστα πεδίων του ενός πίνακα στο αντίστοιχο πεδίο στη λίστα πεδίων του άλλου πίνακα. Με τον τρόπο αυτό συνδέουμε τα ζευγάρια πεδίων: **ΠΕΛΑΤΕΣ.Κωδ_πελάτη** με **ΠΑΡΑΓΓΕΛΙΕΣ.Κωδικός_πελάτη**, **ΠΑΡΑΓΓΕΛΙΕΣ.Κωδ_παραγγελίας** με **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ.Κωδικός_παραγγελίας** και **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ.Κωδικός_προϊόντος** με **ΠΡΟΪΟΝΤΑ.Κωδ_προιόντος** (Σχήμα 4.11).
- Εισάγουμε τα κριτήρια. Το μοναδικό κριτήριο στο παράδειγμα είναι αυτό για την επιλογή των παραγγελιών του μήνα Ιανουαρίου 2015. Για το σκοπό αυτό εισάγουμε στο κελί **Κριτήρια** του πεδίου **Ημνια_παραγγελίας** την έκφραση: **Between #1/1/2015# And #31/1/2015# .**



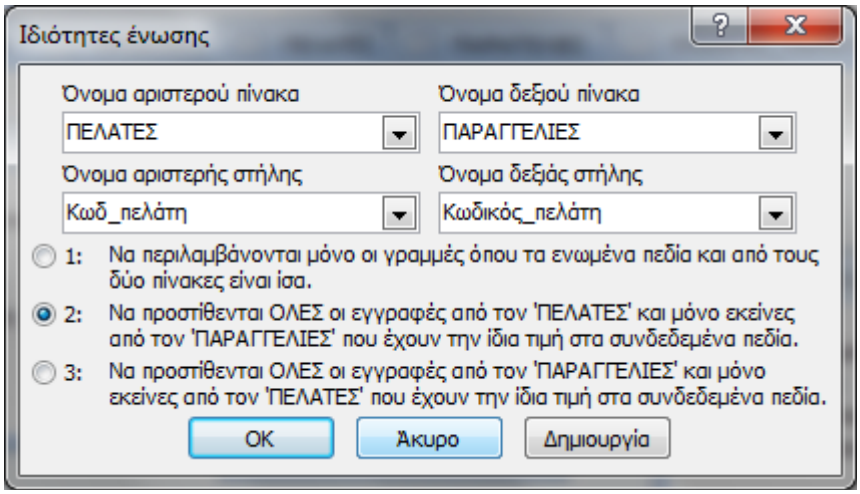
Σχήμα 4.10. *Εισαγωγή των απαραίτητων πεδίων στο πλέγμα σχεδίασης από 4 διαφορετικούς πίνακες*



Σχήμα 4.11. *Σύνδεση πινάκων και εισαγωγή κριτηρίων*

Από τη στιγμή που είναι συνδεδεμένοι δύο ή περισσότεροι πίνακες και έχουν προστεθεί πεδία και από τους πίνακες αυτούς στο πλέγμα σχεδίασης, η σύνδεση καθοδηγεί το ερώτημα να αναζητήσει τιμές που ταιριάζουν στα συνδεδεμένα πεδία. (Αυτό ονομάζεται εσωτερική ένωση στην ορολογία των Βάσεων Δεδομένων.) Όταν βρεθούν τιμές που ταιριάζουν, συνδυάζονται οι αντίστοιχες εγγραφές των διαφορετικών πινάκων και εμφανίζονται ως μια εγγραφή στα αποτελέσματα του ερωτήματος. Έτσι, οι γραμμές που εμφανίζονται κατά την εκτέλεση του ερωτήματος αποτελούνται από τμήματα εγγραφών που έχουν επιλεγεί από τους πίνακες έτσι ώστε ανά δύο να έχουν την ίδια τιμή στα συνδεδεμένα τους πεδία.

Σύμφωνα με αυτόν το μηχανισμό, εάν κάποιες εγγραφές του ενός πίνακα δεν ταιριάζουν με καμία εγγραφή του συνδεδεμένου πίνακα, δεν εμφανίζονται στα αποτελέσματα του ερωτήματος. Π.χ. στο αποτέλεσμα του ερωτήματος δε θα εμφανιστούν τα στοιχεία κάποιου πελάτη αν αυτός δε συνδέεται με καμία παραγγελία (κάτι που είναι αναμενόμενο να συμβεί και είναι επιθυμητό) αλλά επίσης τα στοιχεία κάποιας παραγγελίας δε θα εμφανιστούν αν δε συνδέεται με κάποιον πελάτη (κάτι τέτοιο δεν περιμένουμε βέβαια να προκύψει υπό φυσιολογικές συνθήκες). Εάν θέλουμε το ερώτημα να εμφανίζει όλες τις εγγραφές από έναν πίνακα, ανεξάρτητα από το αν υπάρχουν εγγραφές που ταιριάζουν στον άλλο πίνακα, πρέπει να επιλεγεί ένας ειδικός τύπος ένωσης από τις Ιδιότητες συνδέσμου (δεξί κλικ στη γραμμή της σύνδεσης και **Ιδιότητες συνδέσμου**). Π.χ. αν θέλουμε μια λίστα όλων των πελατών μας και το κόστος της τελευταίας τους παραγγελίας μέσα στο τρέχον έτος, το ερώτημα θα πρέπει να περιλαμβάνει τους πίνακες ΠΕΛΑΤΕΣ και ΠΑΡΑΓΓΕΛΙΕΣ, συνδεδεμένους στον κωδικό πελάτη. Η σύνδεση των πινάκων όμως επιβάλει οι πελάτες που δεν έχουν δώσει καμία παραγγελία μέσα στο τελευταίο έτος να μην εμφανιστούν καθόλου (σαν να μην υπάρχουν). Αν λοιπόν θέλουμε να εμφανιστούν ακόμα και οι πελάτες που δε συνδέονται με καμία παραγγελία, πρέπει να αλλάξει ο προκαθορισμένος τύπος ένωση με την κατάλληλη επιλογή στις Ιδιότητες συνδέσμου (Σχήμα 4.12).



Σχήμα 4.12. *Στις ιδιότητες συνδέσμου μπορεί να επιλεγεί η εμφάνιση όλων των εγγραφών του ενός πίνακα, ακόμα και αν κάποιες δεν ταιριάζουν με κάποια εγγραφή του άλλου πίνακα.*

Σημείωση: Εάν έχουν ήδη δημιουργηθεί συνδέσεις μεταξύ πινάκων στο παράθυρο **Σχέσεις,** η Access εμφανίζει αυτόματα τις γραμμές σύνδεσης όταν προσθέτουμε σχετιζόμενους πίνακες σε Προβολή σχεδίασης ερωτήματος. Επίσης, εάν είναι επιβεβλημένη η ακεραιότητα αναφορών (βλέπε κεφάλαιο 3), η Access εμφανίζει το σύμβολο "1", επάνω από τη γραμμή ένωσης, για να δείξει τον πίνακα που βρίσκεται στην πλευρά "ένα" μιας σχέσης ένα-προς-πολλά και το σύμβολο του απείρου για να δείξει τον πίνακα που βρίσκεται στην πλευρά "πολλά". Ακόμη και αν δεν έχουν δημιουργηθεί σχέσεις, η Access μπορεί να δημιουργήσει αυτόματα συνδέσεις, αν προστεθούν σε ένα ερώτημα δύο πίνακες οι οποίοι έχουν από ένα πεδίο με ίδιο όνομα και τύπο δεδομένων και αν ένα από τα πεδία είναι πρωτεύον κλειδί. Τις συνδέσεις αυτές που δημιουργούνται αυτόματα για τη διευκόλυνση του χρήστη, μπορούμε να τις κρατήσουμε ή να τις διαγράψουμε, οπωσδήποτε όμως δεν μπορούμε να βασιστούμε στον αυτοματισμό της Access αλλά θα πρέπει να ελέγξουμε αν είναι σωστές.

Πρακτικές συμβουλές για αποφυγή λαθών στη σύνδεση πινάκων:

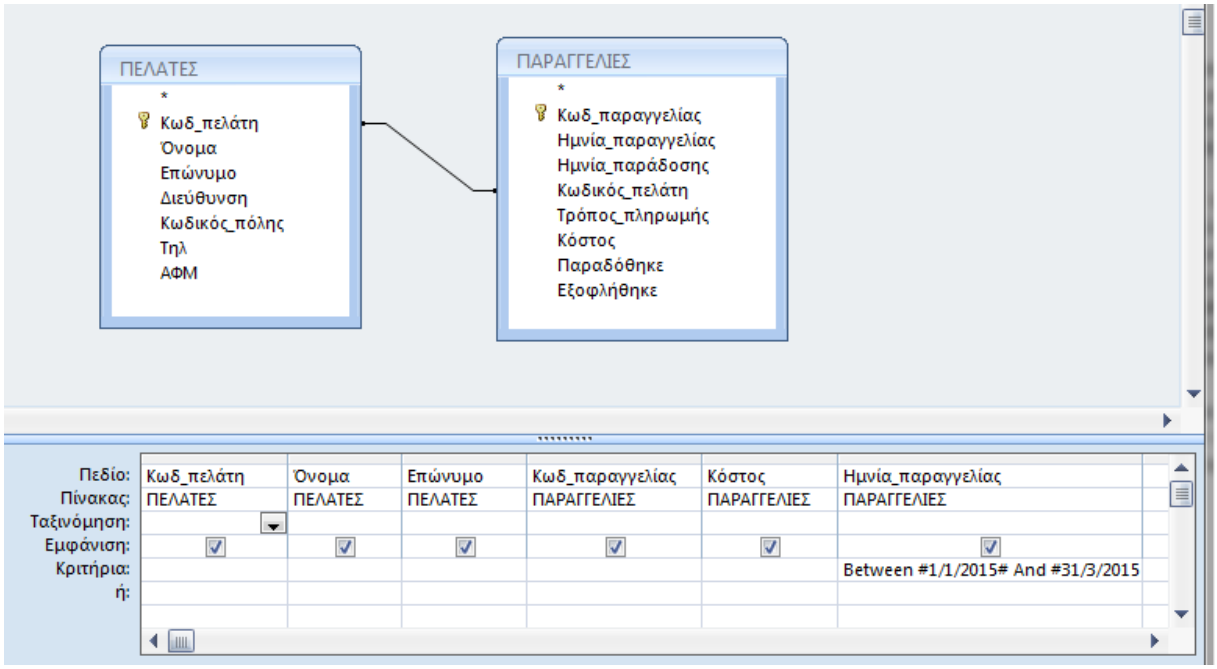
- Κανένας πίνακας δεν πρέπει να είναι τελείως ασύνδετος
- Η σύνδεση μεταξύ δύο πινάκων γίνεται με βάση κοινό πεδίο που συνήθως έχει το ίδιο όνομα και στους δύο πίνακες, είναι το πρωτεύον κλειδί του ενός πίνακα και αποτελεί ξένο κλειδί στον άλλο πίνακα. Τα δύο πεδία που θα συνδεθούν πρέπει να είναι ίδιου τύπου και τα δεδομένα που θα εισαχθούν να είναι συμβατά μεταξύ τους.
- Περισσότερες από μία συνδέσεις ανάμεσα στους ίδιους δύο πίνακες είναι πιθανότατα λάθος.

4.4.2.3 Ερωτήματα που υπολογίζουν συγκεντρωτικά στοιχεία

Μια σημαντική δυνατότητα που προσφέρουν τα ερωτήματα επιλογής είναι να υπολογίζουν συγκεντρωτικά στοιχεία, όπως αθροίσματα, μέσους όρους και άλλα, σε ένα σύνολο εγγραφών. Οι υπολογισμοί αυτοί πραγματοποιούνται κατηγοριοποιώντας τις γραμμές που παράγει ένα ερώτημα με βάση ένα ή περισσότερα πεδία ομαδοποίησης και συγκεντρώνοντας τα δεδομένα από πολλές γραμμές της ίδιας ομάδας σε μία μόνο γραμμή συγκεντρωτικών στοιχείων ανά ομάδα.

Για να γίνει κατανοητός ο τρόπος χρήσης των συγκεντρωτικών στοιχείων, χρησιμοποιούμε το παρακάτω παράδειγμα: Στη Βάση Δεδομένων του καταστήματος ηλεκτρικών ειδών θέλουμε να δημιουργήσουμε ένα ερώτημα που να μας εμφανίζει τον αριθμό των παραγγελιών και τη συνολική αξία των παραγγελιών που έχει δώσει ο κάθε πελάτης στο διάστημα Ιανουαρίου-Μαρτίου 2015, έτσι ώστε να μπορούμε να αξιολογήσουμε ποιοι ήταν οι «καλύτεροι» πελάτες στο διάστημα αυτό. Θα θέλαμε μάλιστα να εμφανίζονται οι πελάτες ταξινομημένοι κατά φθίνουσα σειρά της συνολικής αξίας των αγορών τους και επίσης να εμφανίζονται όλοι ανεξαιρέτως οι πελάτες μας, ακόμα και αυτοί που δεν έκαναν καμία αγορά στο συγκεκριμένο διάστημα.

Το πρώτα βήματα για τη δημιουργία ενός τέτοιου ερωτήματος είναι αυτά που περιγράφηκαν στην προηγούμενη υποενότητα, δηλαδή η εισαγωγή των κατάλληλων πινάκων και η σύνδεση μεταξύ τους, η εισαγωγή στο πλέγμα σχεδίασης των κατάλληλων πεδίων και η εισαγωγή των κατάλληλων κριτηρίων. Το ερώτημα μέχρι αυτό το στάδιο φαίνεται στο Σχήμα 4.13. Χρειαζόμαστε τους πίνακες **ΠΕΛΑΤΕΣ** που περιέχει τα στοιχεία των πελατών (όνομα, κλπ.) και **ΠΑΡΑΓΓΕΛΙΕΣ** που περιέχει τα στοιχεία που χρειαζόμαστε σχετικά με τις παραγγελίες του καθενός. Από τον **ΠΕΛΑΤΕΣ** έχουμε εισάγει τα πεδία **Κωδ_πελάτη**, **Όνομα** και **Επώνυμο** και από τον **ΠΑΡΑΓΓΕΛΙΕΣ** τα πεδία **Κωδ_παραγγελίας** (ώστε να ξεχωρίζουν μεταξύ τους και να μπορούν να μετρηθούν οι παραγγελίες), το **Κόστος** (εφόσον μας ενδιαφέρει το κόστος των παραγγελιών κάθε πελάτη) και το **Ημνία_παραγγελίας** (ώστε να μπορούμε να επιλέξουμε τις παραγγελίες εντός του επιθυμητού χρονικού διαστήματος).



Σχήμα 4.13. *Το πρώτο στάδιο δημιουργίας ερωτήματος για την εύρεση των καλύτερων πελατών.*

Στο Σχήμα 4.14 φαίνεται το αποτέλεσμα της εκτέλεσης του ερωτήματος, χωρίς να έχουμε προβλέψει τον υπολογισμό συγκεντρωτικών στοιχείων. Παρατηρούμε ότι εμφανίζονται σωστά τα στοιχεία των πελατών μαζί με όλες τις παραγγελίες τους εντός του διαστήματος Ιανουάριος-Μάρτιος 2015. Η διαφορά με το αποτέλεσμα που θα θέλαμε είναι ότι για κάθε πελάτη εμφανίζεται ξεχωριστά η κάθε παραγγελία του και το όνομά του επαναλαμβάνεται για κάθε παραγγελία. Θα θέλαμε να εμφανίζεται μία μόνο γραμμή ανά πελάτη, όπου να προβάλλεται το όνομά του, να μετριέται το πλήθος όλων των παραγγελιών του και να υπολογίζεται το άθροισμα του κόστους των παραγγελιών αυτών.

Καλοί_πελάτες

Κωδ_πελάτη	Όνομα	Επώνυμο	Κωδ_παραγγελίας	Κόστος	Ημνία_παραγγελίας
Π1	Γιώργος	Παπαδόπουλο	1	235,00 €	1/2/2015
Π2	Νίκος	Μέλας	4	76,70 €	5/3/2015
Π2	Νίκος	Μέλας	2	332,50 €	8/3/2015
Π3	Μάριος	Καλής	3	109,90 €	31/1/2015
Π3	Μάριος	Καλής	6	278,00 €	17/3/2015
Π9	Πελάτης Λιανι		5	53,50 €	6/3/2015
*					

Σχήμα 4.14. *Το αποτέλεσμα του ερωτήματος χωρίς πρόβλεψη για συγκεντρωτικά στοιχεία.*

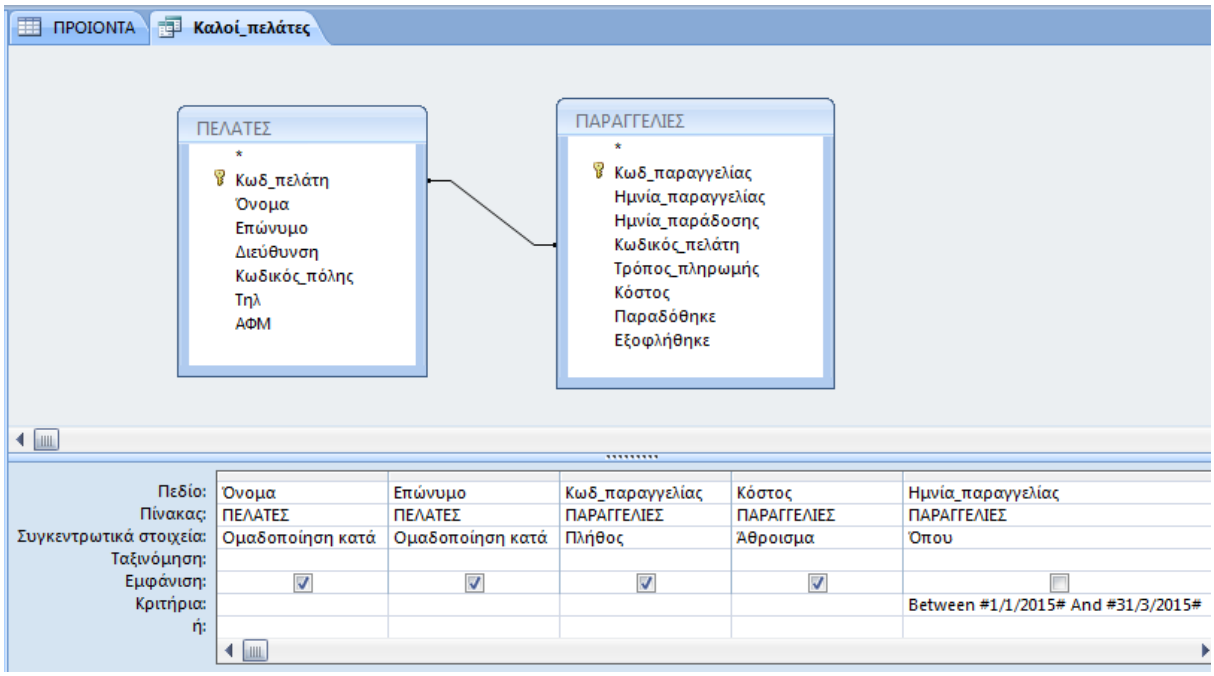
Η ενεργοποίηση των συγκεντρωτικών στοιχείων γίνεται όταν βρισκόμαστε στην προβολή σχεδίασης ενός ερωτήματος, από το σχετικό κουμπί με το σύμβολο **Σ**, που βρίσκεται στην καρτέλα **Σχεδίαση**. Ενεργοποιώντας τα συγκεντρωτικά στοιχεία, παρατηρούμε ότι στο πλέγμα σχεδίασης εμφανίζεται μια επιπλέον γραμμή με όνομα **Συγκεντρωτικά στοιχεία**. Στη γραμμή αυτή μπορούμε να καθορίσουμε, για κάθε πεδίο χωριστά, τον τρόπο με τον οποίο επιθυμούμε να πραγματοποιείται η συγκέντρωση των τιμών του πεδίου αυτού. Προεπιλεγμένη τιμή για όλα τα πεδία είναι το Ομαδοποίηση κατά. Οι διαθέσιμες επιλογές και ο τρόπος λειτουργίας τους παρουσιάζονται στον παρακάτω Πίνακα 4.3.

Τύπος συγκέντρωσης στοιχείων	**Λειτουργία**
Ομαδοποίηση κατά	Οι εγγραφές ομαδοποιούνται σύμφωνα με τις τιμές του συγκεκριμένου πεδίου. Όταν κάποιες εγγραφές περιέχουν την ίδια τιμή, μπορούν να συγχωνευθούν σε μια ομάδα που εκπροσωπείται από μια νέα συγκεντρωτική εγγραφή. Η κοινή αυτή τιμή παραμένει ως τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή που θα προκύψει. Εγγραφές που περιέχουν διαφορετική τιμή στο συγκεκριμένο πεδίο, δε συγχωνεύονται αλλά παραμένουν σε διαφορετικές ομάδες και η καθεμιά κρατάει το περιεχόμενό της.
Άθροισμα	Κατά τη συγχώνευση εγγραφών, η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι το άθροισμα των τιμών του πεδίου των υπό συγχώνευση εγγραφών.
Μέσος_όρος	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι ο μέσος όρος των τιμών του πεδίου των υπό συγχώνευση εγγραφών.
Μικρότερη_τιμή	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η μικρότερη από τις τιμές που θα βρεθούν στο πεδίο αυτό στις υπό συγχώνευση εγγραφές.
Μεγαλύτερη_τιμή	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η μεγαλύτερη από τις τιμές που θα βρεθούν στο πεδίο αυτό στις υπό συγχώνευση εγγραφές.
Πλήθος	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι το πλήθος των διαφορετικών τιμών που θα βρεθούν στο πεδίο αυτό στις υπό συγχώνευση εγγραφές.
Τυπική_απόκλιση	Κατά τη συγχώνευση εγγραφών, η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η τυπική απόκλιση τιμών του πεδίου των υπό συγχώνευση εγγραφών. (Η τυπική απόκλιση ποσοτικών δεδομένων ορίζεται από τη στατιστική ως η τετραγωνική ρίζα της διακύμανσης και συμβολίζεται συνήθως με **σ**)
Διακύμανση	Κατά τη συγχώνευση εγγραφών, η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η διακύμανση των τιμών του πεδίου των υπό συγχώνευση εγγραφών. (Η διακύμανση ποσοτικών δεδομένων μετρά τη στατιστική διακύμανση όλων των τιμών της στήλης γύρω από το μέσο όρο τους και συμβολίζεται συνήθως με $\boldsymbol{\sigma^2}$)
Πρώτο	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η πρώτη από τις τιμές που θα βρεθούν στο πεδίο αυτό στις υπό συγχώνευση εγγραφές. (Δεν πρέπει να συγχέεται με τη μικρότερη τιμή)
Τελευταίο	Η τιμή του συγκεκριμένου πεδίου στη συγκεντρωτική εγγραφή είναι η τελευταία από τις τιμές που θα βρεθούν στο πεδίο αυτό στις υπό συγχώνευση εγγραφές.
Έκφραση	Δίνει τη δυνατότητα σύνταξης ειδικής έκφρασης όπου να ορίζεται ο υπολογισμός της τιμής του πεδίου στη συγκεντρωτική εγγραφή.
Όπου	Επιλέγεται όταν το συγκεκριμένο πεδίο δε θέλουμε να επηρεάσει την ομαδοποίηση και δε μας ενδιαφέρει η τιμή που θα προκύψει στη συγκεντρωτική εγγραφή. Χρησιμοποιείται στα πεδία που συμμετέχουν στο ερώτημα επειδή περιέχουν κάποιο κριτήριο επιλογής εγγραφών. Επειδή η τιμή που θα προκύψει στα συγκεντρωτικά στοιχεία είναι απροσδιόριστη, η Access απενεργοποιεί αυτόματα την Εμφάνιση του πεδίου αυτού.

Πίνακας 4.3 *Οι διαθέσιμες επιλογές συγκέντρωσης στοιχείων* **.**

Μετά την ενεργοποίηση των συγκεντρωτικών στοιχείων, πρέπει να επιλεγεί ο κατάλληλος τρόπος υπολογισμού για κάθε πεδίο στο πλέγμα του ερωτήματος. Στο παραπάνω παράδειγμα, οι κατάλληλες επιλογές είναι οι ακόλουθες:

- **Κωδ_πελάτη:** επιλέγουμε **Ομαδοποίηση κατά** επειδή επιθυμούμε ομαδοποίηση των γραμμών ανά πελάτη, δηλαδή: ίδιος κωδικός σημαίνει ίδιος πελάτης, επομένως συγχώνευση εγγραφών, διαφορετικός κωδικός σημαίνει διαφορετικός πελάτης, επομένως διαφορετική ομάδα εγγραφών.
- **Όνομα** και **Επώνυμο:** επιλέγουμε **Ομαδοποίηση κατά** με το ίδιο σκεπτικό με αυτό του κωδικού πελάτη. Εφόσον ομαδοποιούμε ανά πελάτη, σε όλα τα πεδία που αντιστοιχούν σε χαρακτηριστικό του πελάτη επιλέγουμε Ομαδοποίηση κατά.
- **Κωδ_παραγγελίας**: επιλέγουμε **Πλήθος** επειδή ο κωδικός παραγγελίας είναι πρωτεύον κλειδί στις παραγγελίες, επομένως το πλήθος των κωδικών που θα μετρηθούν σε μια ομάδα εγγραφών ενός πελάτη, μας δίνει το πλήθος των παραγγελιών που συνδέονται με τον πελάτη αυτόν.
- **Κόστος**: επιλέγουμε **Άθροισμα** έτσι ώστε να αθροίζεται το κόστος κάθε επιμέρους παραγγελίας του ίδιου πελάτη.
- **Ημνία_παραγγελίας:** επιλέγουμε **Όπου** επειδή η ημερομηνία παραγγελίας συμμετέχει στο ερώτημα ώστε να είναι δυνατή η επιλογή των παραγγελιών του επιθυμητού διαστήματος, αλλά δε θέλουμε να επηρεάσει τον τρόπο συγκέντρωσης στοιχείων. Η τιμή της ημερομηνίας παραγγελίας είναι απροσδιόριστη στις συγκεντρωτικές εγγραφές, αφού για κάθε πελάτη συγκεντρώνονται στοιχεία από πολλές παραγγελίες με διαφορετικές ημερομηνίες η καθεμιά και για αυτό η εμφάνιση του πεδίου αυτού είναι απενεργοποιημένη. Σημείωση: αν αμελήσουμε αν επιλέξουμε Όπου και αφήσουμε το προεπιλεγμένο Ομαδοποίηση κατά, δε θα πραγματοποιηθεί σωστά η συγκέντρωση στοιχείων, αφού οι παραγγελίες του ίδιου πελάτη με διαφορετικές ημερομηνίες θα θεωρηθούν ως διαφορετική ομάδα.



Σχήμα 4.15. *Ερώτημα με συγκεντρωτικά στοιχεία.*

Στο Σχήμα 4.16 φαίνεται το αποτέλεσμα του ερωτήματος. Παρατηρούμε ότι εμφανίζεται μία μόνο γραμμή ανά πελάτη, όπου προβάλλονται τα στοιχεία του και - ακριβώς όπως θέλαμε - η Access μετράει το πλήθος των παραγγελιών του και υπολογίζει το άθροισμα του κόστους τους. Η ορθότητα των αποτελεσμάτων μπορεί

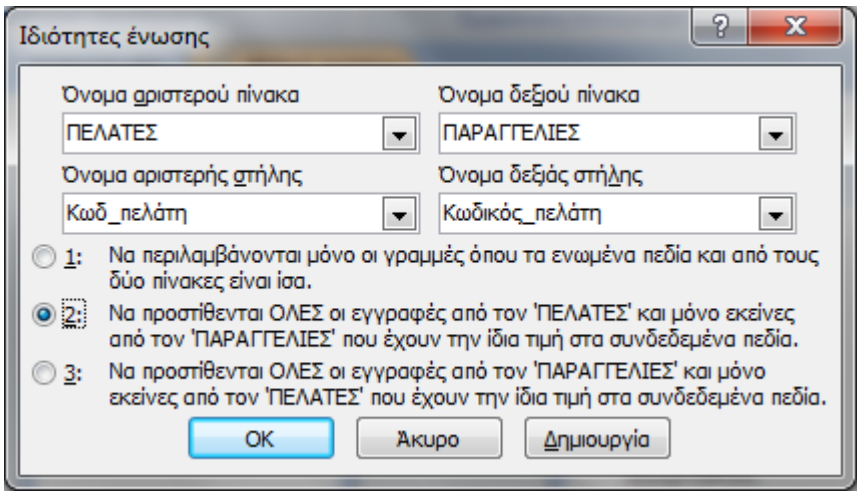
να επαληθευτεί αν συγκριθεί με αυτά του Σχήματος 4.14. Ενώ χωρίς τη χρήση συγκεντρωτικών στοιχείων μπορούσαμε να δούμε ότι π.χ. ο πελάτης Νίκος Μέλας με κωδικό Π2 είχε 2 παραγγελίες με κόστος 76,70€ και 332,50€ αντίστοιχα, το αποτέλεσμα της συγκέντρωσης στοιχείων μας πληροφορεί σε μια γραμμή ότι το πλήθος των παραγγελιών του πελάτη αυτού είναι 2 και το συνολικό κόστος 409,20€. Παρατηρούμε επίσης τις αλλαγές των ονομάτων των στηλών από **Κωδ_παραγγελίας** σε **ΠλήθοςΤουΚωδ_παραγγελίας** και από **Κόστος** σε **ΆθροισμαΤουΚόστος**, που πραγματοποιήθηκαν αυτόματα από την Access, ώστε να φανερώνουν τη λειτουργία συγκέντρωσης στοιχείων που έχει εκτελεστεί στα πεδία αυτά.

ΠΡΟΙΟΝΤΑ | ΠΑΡΑΓΓΕΛΙΕΣ | Καλοί_πελάτες

Κωδ_πελάτη	Όνομα	Επώνυμο	ΠλήθοςΤουΚωδ_παραγγελίας	ΆθροισμαΤουΚόστος
Π1	Γιώργος	Παπαδόπουλος	1	235
Π2	Νίκος	Μέλας	2	409,199996948242
Π3	Μάριος	Καλής	2	387,900001525879
Π9	Πελάτης Λιανικής		1	53,5

Σχήμα 4.16. *Το αποτέλεσμα του ερωτήματος με συγκεντρωτικά στοιχεία.*

Για να ολοκληρωθεί το ερώτημα, απομένει να φροντίσουμε για δύο ακόμα απαιτήσεις. Πρώτον, θέλουμε οι πελάτες να εμφανίζονται ταξινομημένοι σύμφωνα με τη συνολική αξία των παραγγελιών τους. Αυτό επιτυγχάνεται αν στο πλέγμα σχεδίασης επιλέξουμε **Φθίνουσα** στο κελί **Ταξινόμηση** για το πεδίο **Κόστος**. Η δεύτερη απαίτηση ήταν να εμφανίζονται τα στοιχεία ακόμα και των πελατών που δεν έχουν δώσει καμία παραγγελία κατά το επιλεγμένο διάστημα. Στην παρούσα μορφή του ερωτήματος, λόγω της σύνδεσης των 2 πινάκων, εμφανίζονται μόνο συνδυασμοί εγγραφών που ταιριάζουν στο πεδίο σύνδεσης, οπότε δεν εμφανίζονται καθόλου οι εγγραφές των πελατών που δε συνδέονται με καμία παραγγελία. Για να επιτύχουμε το επιθυμητό αποτέλεσμα, κάνουμε δεξί κλικ στη γραμμή σύνδεσης και επιλέγουμε ιδιότητες συνδέσμου. Όπως φαίνεται στο Σχήμα 4.17, ενώ ήταν προεπιλεγμένη η επιλογή **1: Να περιλαμβάνονται μόνο οι γραμμές όπου τα ενωμένα πεδία και από τους δύο πίνακες είναι ίσα**, η επιλογή που ταιριάζει στο ζητούμενο είναι η **2: Να προστίθενται ΟΛΕΣ οι εγγραφές από τον «ΠΕΛΑΤΕΣ» …** Το τελικό αποτέλεσμα φαίνεται στο Σχήμα 4.18, όπου παρατηρούμε ότι οι πελάτες είναι ταξινομημένοι σύμφωνα με τη συνολική αξία των παραγγελιών τους εμφανίζονται τα στοιχεία ακόμα και του πελάτη **Γιώργου Νίκου** με κωδικό **Π4**, ο οποίος δεν έχει καμία παραγγελία.



Σχήμα 4.17. *Στις ιδιότητες συνδέσμου μπορούμε να επιτρέψουμε την εμφάνιση εγγραφών του ενός πίνακα ακόμα και αν δε συνδέονται με τον άλλο πίνακα.*

Κωδ_πελάτη	Όνομα	Επώνυμο	ΠλήθοςΤουΚωδ_παραγγελίας	ΆθροισμαΤουΚόστος
Π2	Νίκος	Μέλας	2	409,199996948242
Π3	Μάριος	Καλής	2	387,900001525879
Π1	Γιώργος	Παπαδόπουλο	1	235
Π9	Πελάτης Λιανι		1	53,5
Π4	Γιώργος	Νίκου	0	

Σχήμα 4.18. *Το αποτέλεσμα του τελικού ερωτήματος όπου έχει συμπεριληφθεί η ταξινόμηση και έχει επιλεγεί η κατάλληλη ιδιότητα συνδέσμου.*

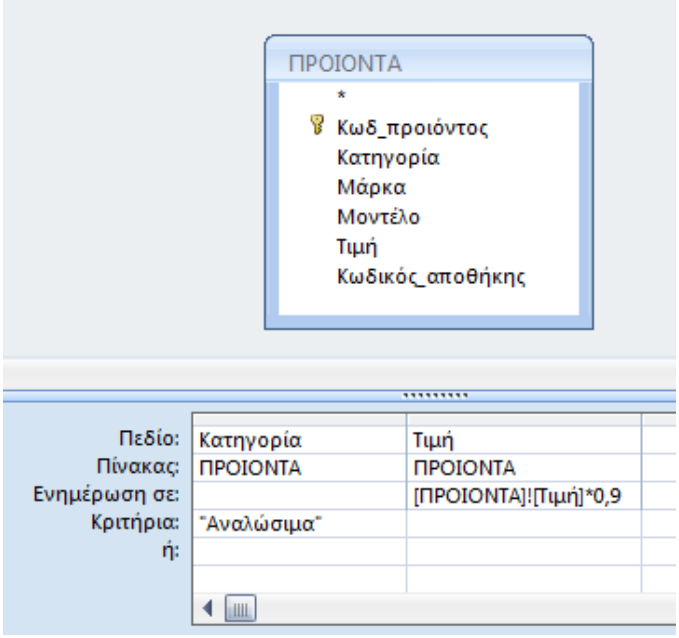
4.4.2.4 Άλλα είδη ερωτημάτων

Η αναζήτηση και ανάκτηση δεδομένων πραγματοποιείται με τα ερωτήματα επιλογής, όπως παρουσιάστηκε στις προηγούμενες υποενότητες. Μέσω των δυνατοτήτων που προσφέρουν οι μηχανισμοί επιλογής και συγκεντρωτικών υπολογισμών των ερωτημάτων αυτών, επιτυγχάνεται επίσης η εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα, κάτι που θα παρουσιαστεί αναλυτικότερα στο Κεφάλαιο 5. Για άλλου είδους ενέργειες διαχείρισης δεδομένων, απαιτείται η χρήση διαφορετικών τύπων ερωτημάτων:

Με ένα **ερώτημα ενημέρωσης**, μπορούμε να εισάγουμε τιμές σε συγκεκριμένα πεδία επιλεγμένων εγγραφών ενός πίνακα. Οι τιμές που θα εισαχθούν μπορεί να είναι σταθερές ή να προκύπτουν από άλλες τιμές που περιέχονται στη Βάση Δεδομένων μέσω υπολογισμών. Π.χ. μπορούμε να ενημερώσουμε μαζικά τις τιμές όλων των προϊόντων της κατηγορίας «Αναλώσιμα» αν έχει αποφασιστεί η μείωσή τους κατά 10%. Στο Σχήμα 4.19 παρουσιάζεται η σχεδίαση ενός τέτοιου ερωτήματος. Για τη δημιουργία του ξεκινάμε, όπως και με το ερώτημα επιλογής, από **Δημιουργία** -> **Σχεδίαση Ερωτήματος**, εισάγουμε τον πίνακα που επιθυμούμε να τροποποιήσουμε και στο **Τύπο ερωτήματος** επιλέγουμε **Ενημέρωση**. Στο πλέγμα σχεδίασης εμφανίζεται η σειρά **Ενημέρωση σε**, στη θέση των **Ταξινόμηση** και **Εμφάνιση**. Εισάγουμε στο πλέγμα το πεδίο **Κατηγορία**, ώστε να μπορούμε να επιλέξουμε τα προϊόντα κατηγορίας «Αναλώσιμα» και το πεδίο **Τιμή**, του οποίου την τιμή θέλουμε να ενημερώσουμε. Για να καθοριστεί ο τρόπος υπολογισμού της νέας τιμής, κάνουμε δεξί κλικ στο κελί **Ενημέρωση σε** του πεδίου **Τιμή** και επιλέγουμε **Δόμηση**. Η επιλογή αυτή ενεργοποιεί έναν οδηγό σύνταξης μαθηματικής έκφρασης, με τη βοήθεια του οποίου καθορίζουμε ότι η νέα τιμή θα ισούται με το τρέχον περιεχόμενο του πεδίου [Τιμή] * 0,9. Μετά την εκτέλεση του ερωτήματος παρατηρούμε ότι έχουν τροποποιηθεί οι τιμές των προϊόντων της κατηγορίας «Αναλώσιμα».

Προσοχή: το ερώτημα ενημέρωσης μπορεί να προκαλέσει μη αναστρέψιμη αλλοίωση των περιεχομένων ενός πίνακα.

Σημείωση: ένα ερώτημα ενημέρωσης μπορεί να εισάγει, να τροποποιήσει ή να διαγράψει τιμές σε πεδία ενός πίνακα που ήδη υπάρχουν, και μόνο σε υπάρχουσες εγγραφές. Δε δίνει τη δυνατότητα δημιουργίας ή διαγραφής ολόκληρων εγγραφών, ούτε τη μεταβολή της σχεδίασης του πίνακα.



Σχήμα 4.19. *Σχεδίαση ερωτήματος ενημέρωσης*

Με ένα ερώτημα **Δημιουργίας Πίνακα**, μπορούμε να δημιουργήσουμε ένα νέο πίνακα στον οποίο θα αποθηκευτούν τα αποτελέσματα του ερωτήματος. Το ερώτημα μπορεί να αντλεί, να επιλέγει και να επεξεργάζεται δεδομένα από έναν ή περισσότερους πίνακες ή άλλα ερωτήματα, ακριβώς όπως και ένα ερώτημα επιλογής, και στη συνέχεια να εισάγει τα αποτελέσματα στο νέο πίνακα, δημιουργώντας τα κατάλληλα πεδία και εισάγοντας τις κατάλληλες εγγραφές.

Ένα ερώτημα **Προσάρτησης** λειτουργεί περίπου όπως και ένα ερώτημα **Δημιουργίας**, με τη διαφορά ότι αντί της δημιουργίας ενός νέου πίνακα, τα αποτελέσματα του ερωτήματος προστίθενται σε έναν υπάρχοντα πίνακα, σε συνέχεια των εγγραφών που ήδη περιέχει.

Με ένα ερώτημα **Διαγραφής** μπορούμε να διαγράψουμε επιλεγμένες εγγραφές από έναν πίνακα. Σημειώνεται ότι διαγράφονται πάντα ολόκληρες εγγραφές.

4.5 Δημιουργία Φορμών και Εκθέσεων

4.5.1 Φόρμες

Οι φόρμες αποτελούν ένα εύχρηστο και εύκολα προσαρμοζόμενο εργαλείο για την εισαγωγή ή προβολή των δεδομένων. Εξυπηρετούν την ανάγκη να δημιουργούμε στην Access εύχρηστες εφαρμογές που να απευθύνονται σε χρήστες που δε χρειάζεται να είναι εξοικειωμένοι με το περιβάλλον σχεδιασμού και προβολής δεδομένων της Access. Οι φόρμες μεσολαβούν ανάμεσα στο χρήστη και τα δεδομένα των πινάκων και λειτουργούν ως μια προβολή επιλεγμένων πεδίων από πίνακες ή ερωτήματα, όπου μπορούμε να μορφοποιήσουμε την εμφάνιση, να ταξινομήσουμε τα δεδομένα, να προσθέσουμε ετικέτες και επιπλέον στοιχεία όπως λογότυπα, υποσέλιδα, κλπ., καθώς και πρότυπα για την ορθή διαχείριση των δεδομένων.

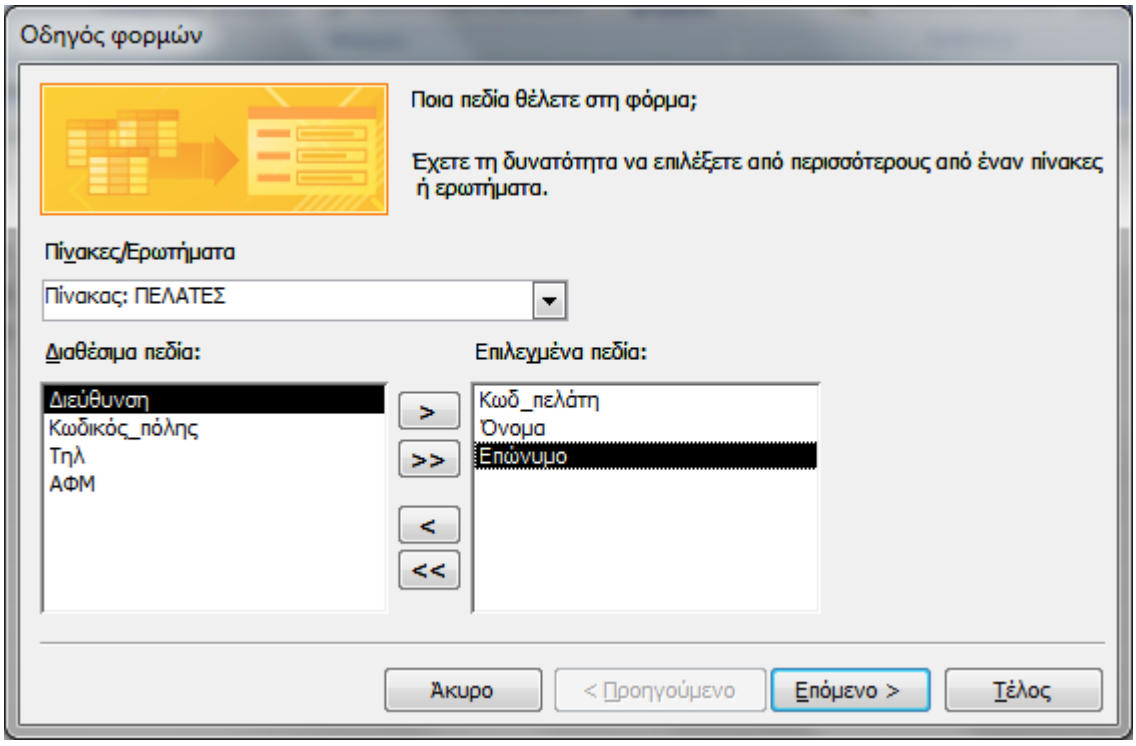
Μπορούμε να θεωρήσουμε ότι μια φόρμα είναι μια «οθόνη» ενός προγράμματος εφαρμογής που μπορεί να χρησιμοποιήσει ένα στέλεχος επιχείρησης για να εκτελέσει κάποια εργασία που σχετίζεται με δεδομένα π.χ. να συμπληρώσει τα στοιχεία ενός πελάτη ή να προβάλει μια παραγγελία. Η φόρμα περιέχει ενεργά στοιχεία που λειτουργούν σε έναν Η/Υ, όπως κουμπιά ή μενού επιλογής, που ελέγχουν τις επιθυμητές ενέργειες.

Η σύνθεση της φόρμας γίνεται με την προσθήκη γραφικών αντικειμένων στον ενεργό χώρο της φόρμας. Τα στοιχεία που μπορούν να προστεθούν σε μια φόρμα λέγονται **Στοιχεία ελέγχου** και μπορούν να επιλεγούν από μια πληθώρα εργαλείων που προσφέρει η Access. Με τα στοιχεία ελέγχου επιτυγχάνεται η σύνδεση μεταξύ μιας φόρμας και της προέλευσης δεδομένων της, καθώς και η δομή και εμφάνισή της.

Ο πιο κοινός τύπος στοιχείου ελέγχου που χρησιμοποιείται για την εμφάνιση και καταχώριση δεδομένων είναι το **Πλαίσιο κειμένου**. Το πλαίσιο κειμένου συνδέεται με συγκεκριμένο πεδίο κάποιου πίνακα ή ερωτήματος και μπορεί είτε να διαβάζει και να εμφανίζει μορφοποιημένα τα περιεχόμενα του πεδίου αυτού, είτε να επιτρέπει στο χρήστη να εισάγει ή να τροποποιήσει τα δεδομένα του πεδίου. Χρήσιμο στοιχείο ελέγχου είναι επίσης η **Ετικέτα**, το οποίο δίνει τη δυνατότητα να εμφανίζουμε στη φόρμα σταθερά στοιχεία όπως τίτλους ή σημάνσεις.

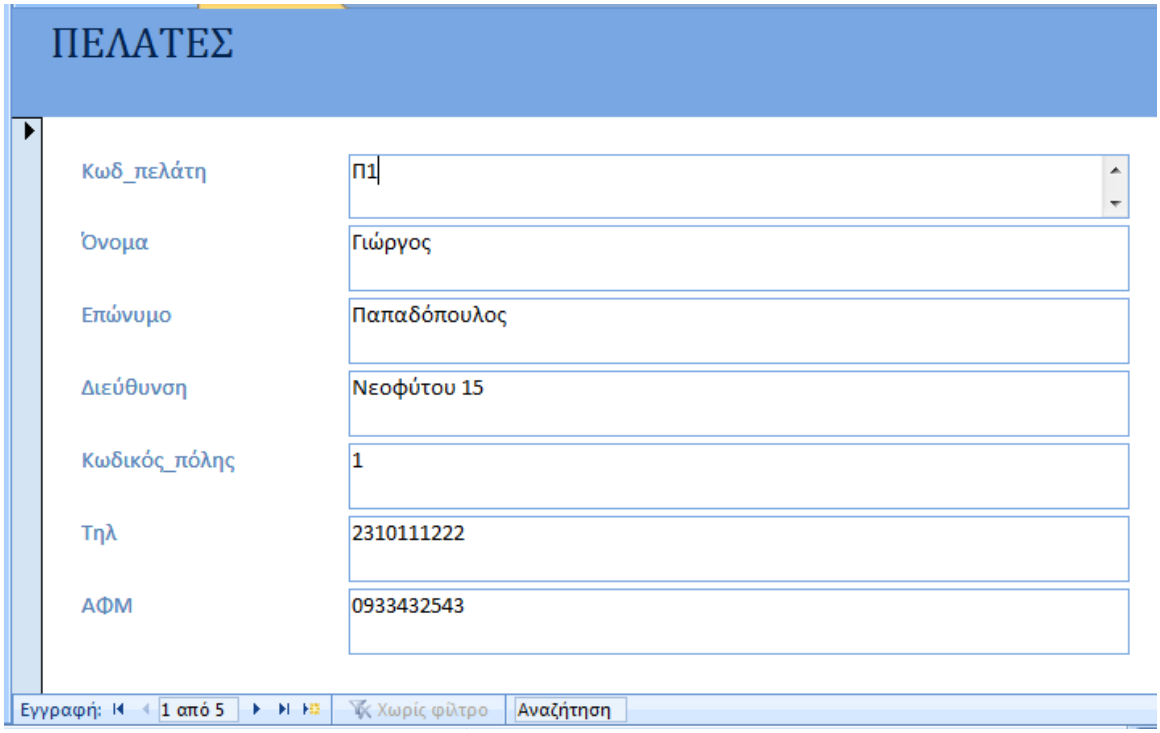
Για την αρχική δημιουργία της φόρμας συνιστάται η χρήση του προσφερόμενου από την Access οδηγού, ο οποίος μας επιτρέπει μέσω ενός μικρού αριθμού απλών βημάτων να δημιουργήσουμε τη βασική φόρμα που χρειαζόμαστε. Στη συνέχεια, συνιστάται η μετάβαση σε προβολή σχεδίασης, όπου μπορούμε να τελειοποιήσουμε τη φόρμα.

Για την επίδειξη της διαδικασίας δημιουργίας μιας φόρμας θα χρησιμοποιήσουμε ως παράδειγμα μια φόρμα για την ενημέρωση των στοιχείων των πελατών ή την εισαγωγή των στοιχείων ενός νέου πελάτη. Ξεκινάμε επιλέγοντας **Δημιουργία** και **Οδηγός φορμών**. Στο πρώτο βήμα του οδηγού επιλέγουμε τον πίνακα ή ερώτημα από τον οποίο προέρχονται τα δεδομένα και επιλέγουμε τα πεδία που επιθυμούμε να εμφανιστούν στο ερώτημα (Σχήμα 4.20). Επιλέγουμε όλα τα πεδία του πίνακα **ΠΕΛΑΤΕΣ**. Στο επόμενο βήμα επιλέγουμε τη διάταξη Στήλης (ώστε να προβάλλονται στο διαθέσιμο όλα τα στοιχεία μίας μόνο εγγραφής) , στη συνέχεια ένα από τα διαθέσιμα στυλ και ολοκληρώνουμε τη διαδικασία με την επιλογή ονόματος.



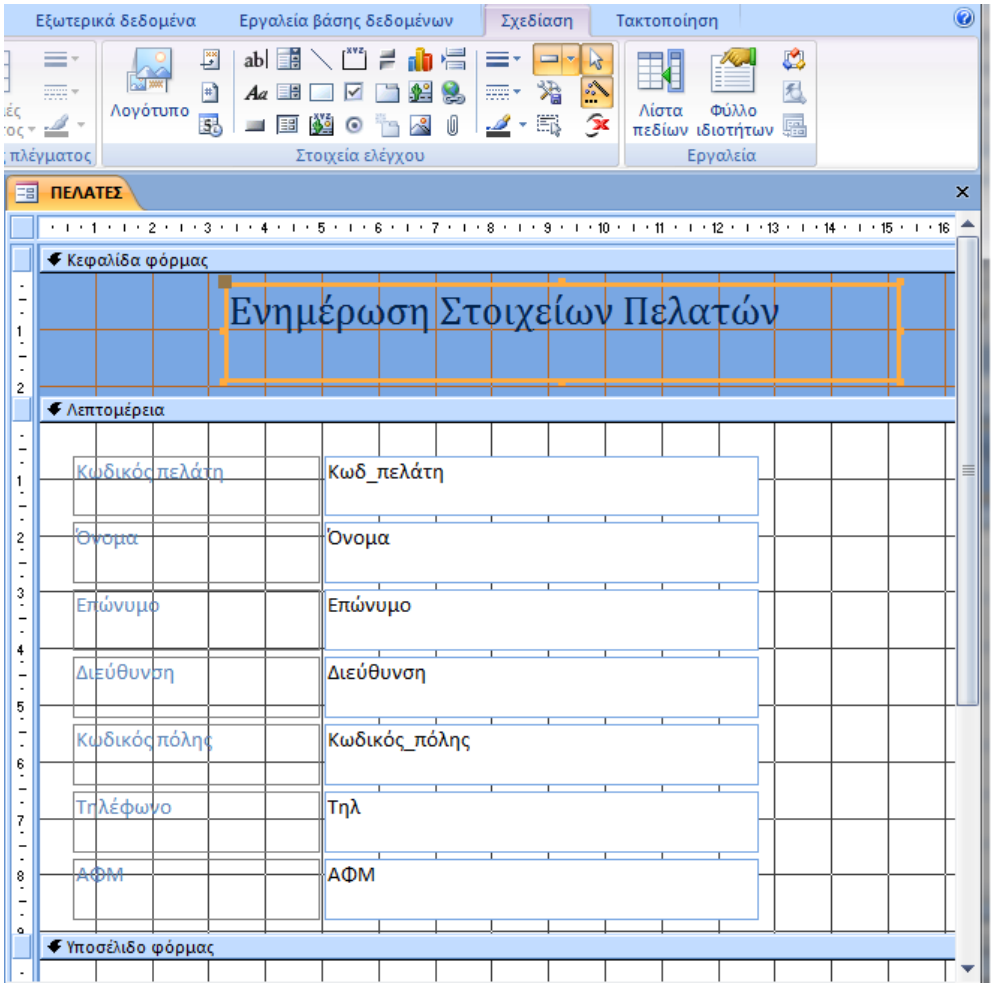
Σχήμα 4.20. *Το πρώτο βήμα του οδηγού φορμών.*

Το αποτέλεσμα του οδηγού φαίνεται στο Σχήμα 4.21. Η φόρμα είναι σε λειτουργία και προβάλει τα στοιχεία της 1ης εγγραφής του πίνακα ΠΕΛΑΤΕΣ. Ο χρήστης μπορεί να μετακινηθεί σε οποιαδήποτε εγγραφή επιθυμεί και να δημιουργήσει νέα κενή εγγραφή, όπου μπορεί να εισάγει τα στοιχεία ενός νέου πελάτη.



Σχήμα 4.21. *Η δημιουργημένη φόρμα συνδέεται άμεσα με τον πίνακα ΠΕΛΑΤΕΣ και δίνει πρόσβαση στο περιεχόμενό του*

Μεταβαίνοντας σε Προβολή σχεδίασης, μπορούμε να τελειοποιήσουμε τη φόρμα, προσθέτοντας, τροποποιώντας ή απλά μετακινώντας Στοιχεία ελέγχου. Π.χ. μπορούμε να αλλάξουμε τον τίτλο σε «Ενημέρωση στοιχείων πελατών» και να τον μετακινήσουμε στο κέντρο, να αλλάξουμε τα ονόματα των πεδίων σε πιο ευανάγνωστα (από **Κωδ_πελάτη** σε **Κωδικός πελάτη**, κλπ.) και να μειώσουμε το πλάτος της στήλης των δεδομένων (Σχήμα 4.22). Παρατηρούμε ότι στο χώρο σχεδίασης εμφανίζονται ζευγάρια από στοιχεία ελέγχου με το ίδιο όνομα π.χ. **Κωδ_πελάτη**. Διευκρινίζεται ότι το ένα από αυτά (αριστερά στην περίπτωσή μας) είναι στοιχείο **Ετικέτα**, δηλαδή το περιεχόμενό του είναι στατικό, εμφανίζεται στη φόρμα ακριβώς όπως το βλέπουμε και μπορούμε να το τροποποιήσουμε όπως μας εξυπηρετεί. Το δεύτερο στοιχείο με το ίδιο όνομα είναι **Πλαίσιο κειμένου**, που σημαίνει ότι το περιεχόμενό του δεν είναι ελεύθερο κείμενο αλλά το όνομα του πεδίου με το οποίο είναι συνδεδεμένο. Στην περίπτωση αυτή, στην προβαλλόμενη φόρμα δεν εμφανίζεται το κείμενο **Κωδ_πελάτη** αλλά τα δεδομένα που περιέχει ο πίνακας **ΠΕΛΑΤΕΣ** στο πεδίο **Κωδ_πελάτη**. Στο Σχήμα 4.22 παρατηρούμε επίσης σε μορφή εικονιδίων τα Στοιχεία ελέγχου που διαθέτει η Access. Περισσότερες λεπτομέρειες για τη χρήση τους μπορεί να βρει ο αναγνώστης στη **Βοήθεια** της Access ή στη βιβλιογραφία που αναφέρεται στο τέλος του κεφαλαίου αυτού.



Σχήμα 4.22. *Η προβολή σχεδίασης της φόρμας απ'όπου μπορούμε να την τελειοποιήσουμε.*

Σημείωση: Η φόρμα προορίζεται για την προβολή των δεδομένων και όχι για την αναζήτηση και το χειρισμό τους. Επομένως όταν απαιτούνται ενέργειες επιλογής, συνδυασμού δεδομένων από περισσότερους από έναν σχετιζόμενους πίνακες, συγκέντρωσης στοιχείων και οποιασδήποτε μορφής επεξεργασίας, είναι απαραίτητο να προηγείται η δημιουργία κατάλληλου ερωτήματος που να παράγει τα δεδομένα ή την πληροφορία που χρειαζόμαστε. Στη συνέχεια, μπορούμε να δημιουργήσουμε τη φόρμα που θα συνδεθεί με το ερώτημα και θα προβάλει τα δεδομένα με τον κατάλληλο τρόπο.

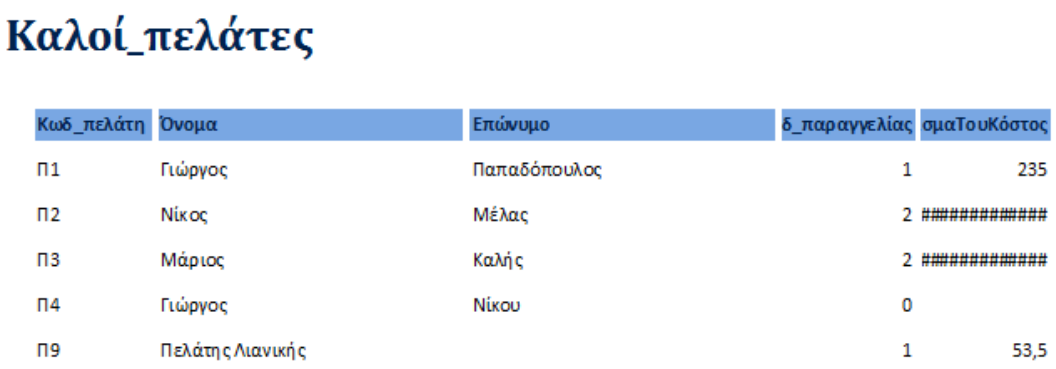
4.5.2 Εκθέσεις

Οι Εκθέσεις χρησιμεύουν στην ευπαρουσίαστη και μορφοποιημένη προβολή των δεδομένων, έτσι ώστε να μπορούν να συμπεριληφθούν σε ένα έντυπο (είτε εκτυπώσιμο είτε ηλεκτρονικό) όπως π.χ. μια αναφορά, μια παρουσίαση ή ένα παραστατικό. Ακριβώς όπως και οι φόρμες, οι Εκθέσεις συνδέονται με πίνακες ή ερωτήματα, από τα οποία αντλούνται τα δεδομένα και δίνουν τη δυνατότητα μορφοποίησης και διάταξής τους πάνω στον ενεργό χώρο του εντύπου, καθώς και προσθήκης ετικετών, λογότυπων, υποσέλιδων και άλλων στοιχείων. Η διαδικασία δημιουργίας της έκθεσης και τα διαθέσιμα εργαλεία είναι παρόμοια με αυτά της φόρμας. Η διαφορά τους είναι ότι η έκθεση δημιουργεί ένα έντυπο, το οποίο αποτελεί στατικό αποτέλεσμα, που δεν μπορεί να περιέχει ενεργά στοιχεία, όπως κουμπιά και μενού επιλογών, και φυσικά δεν μπορεί να χρησιμοποιηθεί για εισαγωγή στοιχείων.

Για την επίδειξη της δημιουργίας μιας έκθεσης, χρησιμοποιήθηκε το παράδειγμα της υποενότητας 4.2.3: Επιθυμούμε να δημιουργήσουμε μια έκθεση όπου να παρουσιάζονται οι πελάτες του καταστήματος ηλεκτρικών ειδών, ο αριθμός των παραγγελιών και τη συνολική αξία των παραγγελιών που έχει δώσει ο κάθε πελάτης στο διάστημα Ιανουαρίου-Μαρτίου 2015, ταξινομημένοι κατά φθίνουσα σειρά της συνολικής αξίας

των αγορών τους, έτσι ώστε να μπορούμε να αξιολογήσουμε ποιοι ήταν οι «καλύτεροι» πελάτες στο διάστημα αυτό. Εφόσον τα δεδομένα που θα προβληθούν στην έκθεση δεν είναι διαθέσιμα αυτούσια σε κάποιον πίνακα, απαιτείται να δημιουργηθεί πρώτα κατάλληλο ερώτημα για την άντληση των δεδομένων και στη συνέχεια να συνδεθεί η έκθεση με το ερώτημα αυτό. Στην περίπτωσή μας, το κατάλληλο ερώτημα είναι αυτό που παρουσιάστηκε στην υποενότητα 4.2.3, το οποίο αποθηκεύτηκε με το όνομα «Καλοί_πελάτες».

Η δημιουργία της έκθεσης ξεκινάει από την καρτέλα **Δημιουργία**, επιλέγοντας **Οδηγός έκθεσης**. Στο πρώτο βήμα του οδηγού επιλέγουμε το **Ερώτημα: Καλοί_πελάτες** και εισάγουμε όλα τα διαθέσιμα πεδία του στα επιλεγμένα πεδία. Στο 2^{ο} βήμα μπορούμε να προσθέσουμε επίπεδα ομαδοποίησης, δηλαδή να επιλέξουμε κάποιο πεδίο, με βάση τις τιμές του οποίου δημιουργούνται ομάδες εγγραφών. Στο συγκεκριμένο παράδειγμα δεν έχει νόημα κάποια ομαδοποίηση. Στη συνέχεια μπορούμε να ορίσουμε τρόπο ταξινόμησης, επειδή όμως το ερώτημα που δημιουργήσαμε παρέχει το αποτέλεσμα ήδη ταξινομημένο σύμφωνα με το συνολικό κόστος των παραγγελιών κάθε πελάτη, δε χρησιμεύει η δυνατότητα ταξινόμησης που προσφέρει η έκθεση. Τέλος, επιλέγουμε διάταξη **Πίνακα**, τον προσανατολισμό εντύπου που μας εξυπηρετεί π.χ. Κατακόρυφο και το στυλ της αρεσκείας μας. Το αποτέλεσμα της έκθεσης που δημιουργήθηκε μέσω της παραπάνω διαδικασίας φαίνεται στο Σχήμα 4.23, και αποτελεί καλό παράδειγμα έκθεσης κακής ποιότητας που απαιτεί τελειοποίηση με χρήση της προβολής σχεδίασης.



Καλοί_πελάτες

Κωδ_πελάτη	Όνομα	Επώνυμο	δ_παραγγελίας	σμαΤουΚόστος
Π1	Γιώργος	Παπαδόπουλος	1	235
Π2	Νίκος	Μέλας	2	############
Π3	Μάριος	Καλής	2	############
Π4	Γιώργος	Νίκου	0	
Π9	Πελάτης Λιανικής		1	53,5

Σχήμα 4.23. *Η προεπισκόπηση της έκθεσης που δημιουργήθηκε από τον οδηγό.*

Οι βελτιώσεις που κρίνεται ότι απαιτούνται είναι (α) ένα καλύτερος τίτλος αντί του Καλοί_πελάτες (είναι το όνομα του ερωτήματος που δόθηκε αυτόματα και στην έκθεση) (β) καλύτεροι τίτλοι στηλών π.χ. το ακατανόητο ΠλήθοςΤουΚωδ_παραγγελίας να δίνει «Πλήθος παραγγελιών» και (γ) να βελτιωθεί η διάταξη έτσι ώστε το πλάτος των στηλών να αναλογεί στο περιεχόμενο, να γίνει καλύτερη εκμετάλλευση του χώρου και το περιεχόμενο να είναι πιο ευανάγνωστο.

Σχήμα 4.24. *Η προβολή σχεδίασης της έκθεσης, όπου έχουν πραγματοποιηθεί σχεδιαστικές βελτιώσεις*

Ταξινόμηση πελατών κατά συνολικό κόστος παραγγελιών

Κωδικός πελάτη	Όνομα	Επώνυμο	Πλήθος παραγγελιών	Συνολικό κόστος παραγγελιών
Π2	Νίκος	Μέλας	2	409,20 €
Π3	Μάριος	Καλής	2	387,90 €
Π1	Γιώργος	Παπαδόπουλος	1	235,00 €
Π9	Πελάτης Λιανικής		1	53,50 €
Π4	Γιώργος	Νίκου	0	

Τρίτη, 13 Οκτωβρίου 2015 Σελίδα 1 από 1

Σχήμα 4.25. *Η έκθεση μετά τις επεμβάσεις στη σχεδίαση*

Βιβλιογραφία/Αναφορές

Κεχρής Ε. (2015). *Σχεσιακές Βάσεις Δεδομένων*, 2η έκδοση, Αθήνα: Εκδόσεις Κριτική.

Ξαρχάκος Κ. & Καρολίδης Δ. (2010). *Μαθαίνετε εύκολα Microsoft Office 2007*. Αθήνα: Εκδόσεις Άβακας.

Κεφάλαιο 5. Μετατροπή των δεδομένων σε πληροφορία

Σύνοψη
Στο κεφάλαιο αυτό παρουσιάζεται πώς τα δεδομένα μπορούν να μετατραπούν σε χρήσιμη πληροφορία, όπως διοικητικές αναφορές και υπολογισμοί δεικτών, με την αξιοποίηση σύνθετων ερωτημάτων, συναρτήσεων και συγκεντρωτικών στοιχείων. Περιγράφονται τα χρησιμότερα εργαλεία και τεχνικές μετασχηματισμού δεδομένων σε πληροφορία με απλή επεξεργασία δεδομένων, όπως υπολογισμός στατιστικών δεικτών, συνδυαστικές αναζητήσεις και πίνακες συγκεντρωτικών στοιχείων. Περιγράφονται επίσης οι κύβοι OLAP, ως εργαλείο επιχειρηματικής ευφυΐας. Περιλαμβάνονται αντιπροσωπευτικά παραδείγματα από το χώρο της διοίκησης επιχειρήσεων και του μάρκετινγκ όπως π.χ. ταξινόμηση των προϊόντων με βάση την κερδοφορία και υπολογισμός τζίρου ανά χρονική περίοδο. Τα παραδείγματα υλοποιούνται σε Access και παρουσιάζονται με τη βοήθεια σχημάτων.

Προαπαιτούμενη γνώση
Κεφάλαιο 4. Δημιουργία και χρήση μιας σχεσιακής Βάσης Δεδομένων

5.1 Εισαγωγή

Ακόμα και αν ο πρωταρχικός ρόλος μιας Βάσης Δεδομένων είναι η διαχείριση των δεδομένων που απαιτούνται για τις λειτουργικές ανάγκες μιας επιχείρησης, τα δεδομένα αυτά μπορούν να αξιοποιηθούν για την εξαγωγή χρήσιμης πληροφορίας, που μπορεί να χρησιμοποιηθεί ως εργαλείο επιχειρηματικής ευφυΐας στη λήψη αποφάσεων διοίκησης (Roiger & Geatz, 2008). Η απλούστερη μορφή επιχειρηματικής ευφυΐας που μπορεί να ενσωματωθεί σε μια εφαρμογή Βάσεων Δεδομένων, χωρίς την ανάγκη εξειδικευμένων εργαλείων, βασίζεται σε δύο απλές τακτικές:

- τη χρήση μηχανισμών αναζήτησης, διασταύρωσης και επεξεργασίας των ακατέργαστων δεδομένων, ώστε να αναδειχθεί χρήσιμη πληροφορία, όπως π.χ. η άθροιση του κέρδους ανά προϊόν, που προκύπτει από όλες τις πωλήσεις του έτους, ώστε να αναδειχθεί ποια προϊόντα συμβάλουν περισσότερο στην κερδοφορία της επιχείρησης
- την ενσωμάτωση στη Βάση Δεδομένων, επιπρόσθετα των λειτουργικών δεδομένων, επιπλέον στοιχείων που αποσκοπούν σε χρήσεις επιχειρηματικής ευφυΐας, όπως θα ήταν π.χ. η προσθήκη πεδίων στον πίνακα ΠΕΛΑΤΕΣ για την πιστότητα και τις προτιμήσεις τους, η προσθήκη πίνακα ΠΑΡΑΠΟΝΑ ή η προσθήκη πίνακα ΣΤΟΧΟΙ_ΠΩΛΗΣΕΩΝ, όπου να εισάγονται οι στόχοι πωλήσεων έτους ανά προϊόν.

Η πληροφορία που μπορεί να εξαχθεί από τα δεδομένα είναι δομημένη και απαντάει μόνο σε τυποποιημένα ερωτήματα, που θα πρέπει να μπορούν να διατυπωθούν με σαφήνεια. Τα ερωτήματα αυτά συνήθως είναι προκαθορισμένα και θέλουμε ο χρήστης της εφαρμογής να μπορεί να λάβει τη σχετική πληροφορία αυτοματοποιημένα, χωρίς να πρέπει να γνωρίζει λεπτομέρειες σχετικά με το εσωτερικό της Βάσης Δεδομένων. Πληροφορίες τέτοιου τύπου είναι οι τυποποιημένες διοικητικές αναφορές, όπως η κερδοφορία του έτους ανά κατάστημα, οι πωλήσεις ανά προϊόν σε σύγκριση με το στόχο, κλπ. Επίσης, σημαντικός τρόπος εξαγωγής και αξιοποίησης πληροφορίας από δεδομένα είναι ο υπολογισμός τυποποιημένων αριθμητικών δεικτών και η εκτίμηση, με τη βοήθεια των δεικτών αυτών, της απόδοσης της επιχείρησης, με βάση τις διαθέσιμες τεχνικές της επιχειρησιακής έρευνας.

Επιπρόσθετα, ένας πιο έμπειρος χρήστης με γνώσεις χειρισμού δεδομένων, μπορεί να αναζητήσει εξειδικευμένες πληροφορίες, σύμφωνα με τις ιδιαίτερες ανάγκες του. Για το σκοπό αυτό, μπορεί να διαμορφώσει κατάλληλα τη Βάση Δεδομένων και να δημιουργήσει «έξυπνα» ερωτήματα που να τον βοηθούν να λαμβάνει τεκμηριωμένες αποφάσεις.

Σε μεγαλύτερο επίπεδο εμβάθυνσης στην Επιχειρηματική Ευφυΐα, διατίθενται εξειδικευμένα εργαλεία μετατροπής δεδομένων σε πληροφορία, όπως οι Κύβοι Άμεσης Αναλυτικής Επεξεργασίας (Microsoft Office support, 2015), που παρουσιάζονται στην ενότητα 4 του κεφαλαίου αυτού και δίνουν τη δυνατότητα στο χρήστη να «εξερευνήσει» τα δεδομένα με εύχρηστο, γρήγορο και αποτελεσματικό τρόπο,

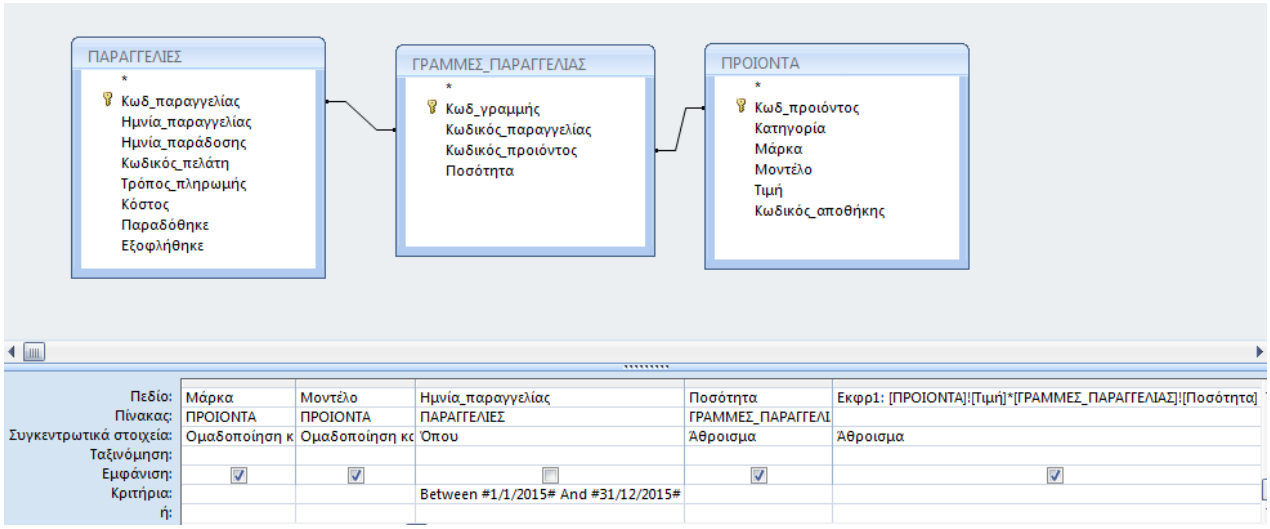
χωρίς την ανάγκη δημιουργίας ερωτημάτων. Η τεχνολογία αυτή απαιτεί τη χρήση επιπλέον λογισμικού, που μπορεί να προσαρτηθεί σε μια απλή Βάση Δεδομένων και να της προσδώσει δυνατότητες επιχειρηματικής ευφυΐας.

Στη συνέχεια του κεφαλαίου αυτού, παρουσιάζονται με τη βοήθεια παραδειγμάτων αντιπροσωπευτικοί τρόποι μετατροπής δεδομένων σε πληροφορία στο περιβάλλον της Access, χρησιμοποιώντας ως εργαλείο τα Ερωτήματα, τις Εκθέσεις και τις Φόρμες. Παρουσιάζεται επίσης ο τρόπος λειτουργίας και η χρησιμότητα των κύβων OLAP. (Τα παραδείγματα που θα παρουσιαστούν διατίθενται υλοποιημένα σε Access 2007 μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/Store_electric_appliances_v2.accdb)

5.2 Διοικητικές αναφορές

Οι διοικητικές αναφορές πληροφορούν για τις οικονομικές και λειτουργικές πτυχές των δραστηριοτήτων μιας επιχείρησης, προσφέρουν σε ένα στέλεχος επιχείρησης μια αντικειμενική και σε βάθος γνώση της επιχείρησης και αποτελούν πολύτιμο εργαλείο για την τεκμηριωμένη λήψη αποφάσεων. Οι συνηθέστερες αναφορές βασίζονται στη συγκέντρωση ποσοτικών στοιχείων για την καταγραφή και σύγκριση των μεγεθών που σχετίζονται με διάφορα τμήματα της επιχείρησης, όπως πωλήσεις, χρηματο-οικονομική διοίκηση, παραγωγή, προμήθειες, προσωπικό, λογιστική και μάρκετινγκ.

Χρησιμοποιώντας το παράδειγμα του καταστήματος ηλεκτρικών ειδών, θα δημιουργήσουμε απλές αναφορές για την παρουσίαση των ετήσιων πωλήσεων σε τεμάχια και αξία, ανά προϊόν και ανά πελάτη.



Σχήμα 5.1. *Η σχεδίαση του ερωτήματος υπολογισμού των τεμαχίων και της συνολικής αξίας πωλήσεων ανά προϊόν*

Για τον υπολογισμό των πωλήσεων ανά προϊόν, δημιουργούμε αρχικά το ερώτημα του Σχήματος 5.1. Συμπεριλαμβάνονται οι πίνακες **ΠΑΡΑΓΓΕΛΙΕΣ, ΓΡΑΜΜΕΣ ΠΑΡΑΓΓΕΛΙΑΣ** και **ΠΡΟΪΟΝΤΑ** και εισάγονται στο πλέγμα σχεδίασης τα πεδία **Κωδ_προϊόντος, Μάρκα, Μοντέλο** από τον πίνακα **ΠΡΟΪΟΝΤΑ** (ως τα βασικά στοιχεία ταυτοποίησης του προϊόντος), η **Ημνία_παραγγελίας** από τον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ** (ώστε να μπορούν να επιλεγούν οι παραγγελίες του έτους), η **Ποσότητα** από τον πίνακα **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ** (για να υπολογιστεί ο αριθμός τεμαχίων που πωλήθηκαν) και μια δομημένη έκφραση για τον υπολογισμό της αξίας των πωληθέντων τεμαχίων του προϊόντος. Σημειώνεται ότι το πεδίο **Κόστος** της παραγγελίας περιέχει το συνολικό κόστος μιας παραγγελίας, που μπορεί να αφορά πολλά διαφορετικά προϊόντα. Επομένως η αξία από την πώληση ενός συγκεκριμένου προϊόντος πρέπει να υπολογιστεί πολλαπλασιάζοντας τον αριθμό τεμαχίων (που είναι διαθέσιμος στη γραμμή παραγγελίας) με την τιμή του προϊόντος (που είναι διαθέσιμη στα στοιχεία του προϊόντος). Για το σκοπό αυτό, στην τελευταία στήλη του πλέγματος, στη θέση **Πεδίο**, κάνουμε δεξί κλικ και επιλέγουμε **Δόμηση**, ώστε με τη βοήθεια του

οδηγού **Δόμησης Εκφράσεων**, να γίνει η σύνταξη της παρακάτω έκφρασης υπολογισμού της αξίας πωλήσεων του προϊόντος:

Εκφρ1: [ΠΡΟΙΟΝΤΑ]![Τιμή]*[ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ]![Ποσότητα]

Για την επιλογή των παραγγελιών του έτους 2015, εισάγουμε ως κριτήριο στο πεδίο **Ημνία_παραγγελίας** την έκφραση **Between #1/1/2015# And #31/12/2015#**. Στη συνέχεια, ενεργοποιούμε τα **Συγκεντρωτικά στοιχεία** και καθορίζουμε τις κατάλληλες επιλογές συγκέντρωσης στοιχείων ανά πεδίο. Εφόσον ομαδοποιούμε ανά προϊόν, στα πεδία **Κωδ_προϊόντος, Μάρκα** και **Μοντέλο**, αφήνουμε την επιλογή **Ομαδοποίηση κατά**, στην **Ημνία_παραγγελίας** επιλέγουμε **Όπου**, ώστε να μην επηρεαστεί η ομαδοποίηση, ενώ στην **Ποσότητα** και στην έκφραση **Εκφρ1** επιλέγουμε Άθροισμα, ώστε να προστεθούν τα τεμάχια και οι αξίες κάθε γραμμής παραγγελίας που αφορά το ίδιο προϊόν.

Κωδ_προιόντος	Μάρκα	Μοντέλο	ΆθροισμαΤουΠοσότητα	Εκφρ1
A1	PITSOS	P18-super	2	470
A2	MORRIS	Clean 15	1	332,5
A3	MORRIS	SC43	2	153,399993896484
A4	FIRST	G1	4	52,4000015258789
A6	KARPA	F12	5	25
A8	ELECTRO	Clean-Economy	1	28,5

Σχήμα 5.2. *Το αποτέλεσμα του ερωτήματος για την αναφορά πωλήσεων ανά προϊόν*

Στο Σχήμα 5.2 φαίνεται το αποτέλεσμα της εκτέλεσης του ερωτήματος. Για να προβάλλεται το αποτέλεσμα αυτό σε ευπαρουσίαστη μορφή, ώστε να είναι κατάλληλο για μια διοικητική αναφορά, μπορούμε να δημιουργήσουμε κατάλληλη **Έκθεση**. Το αποτέλεσμα της Έκθεσης, όπου έχει οριστεί τίτλος και έχουν προσαρμοστεί τα ονόματα και το μέγεθος των στηλών, φαίνεται στο Σχήμα 5.3.

Πωλήσεις ανά προϊόν έτους 2015

Παρασκευή, 16 Οκτωβρίου 2015 3:21:55 μμ

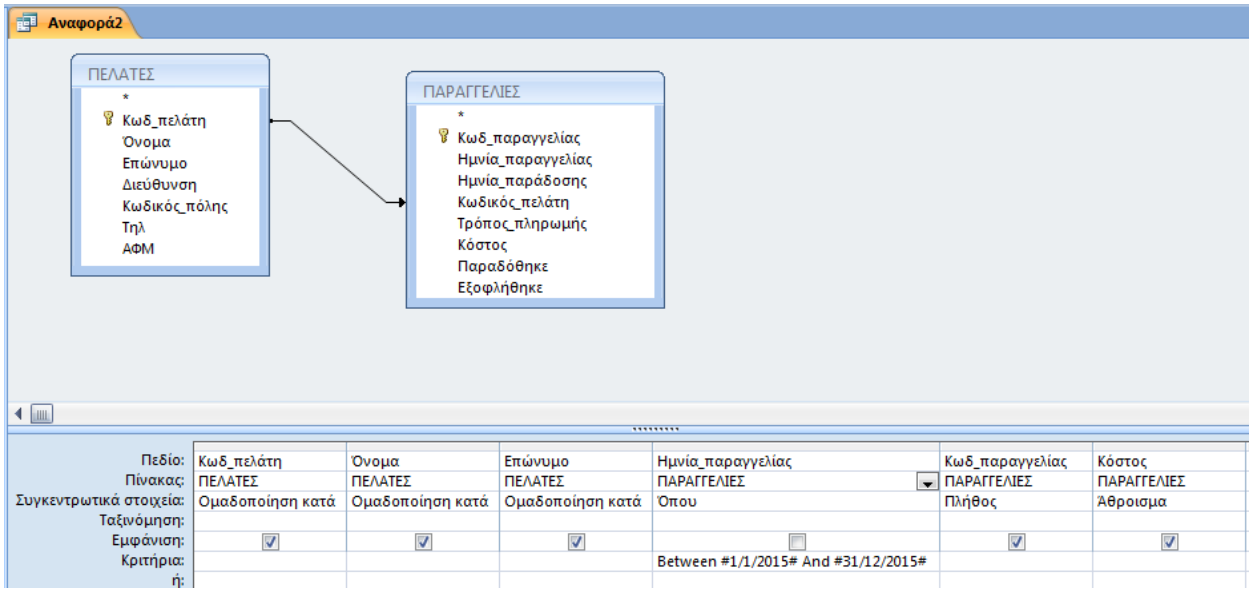
Κωδικός	Μάρκα	Μοντέλο	Τεμάχια	Αξία
A1	PITSOS	P18-super	2	470,00 €
A2	MORRIS	Clean 15	1	332,50 €
A3	MORRIS	SC43	2	153,40 €
A4	FIRST	G1	4	52,40 €
A6	KARPA	F12	5	25,00 €
A8	ELECTRO	Clean-Economy	1	28,50 €

6

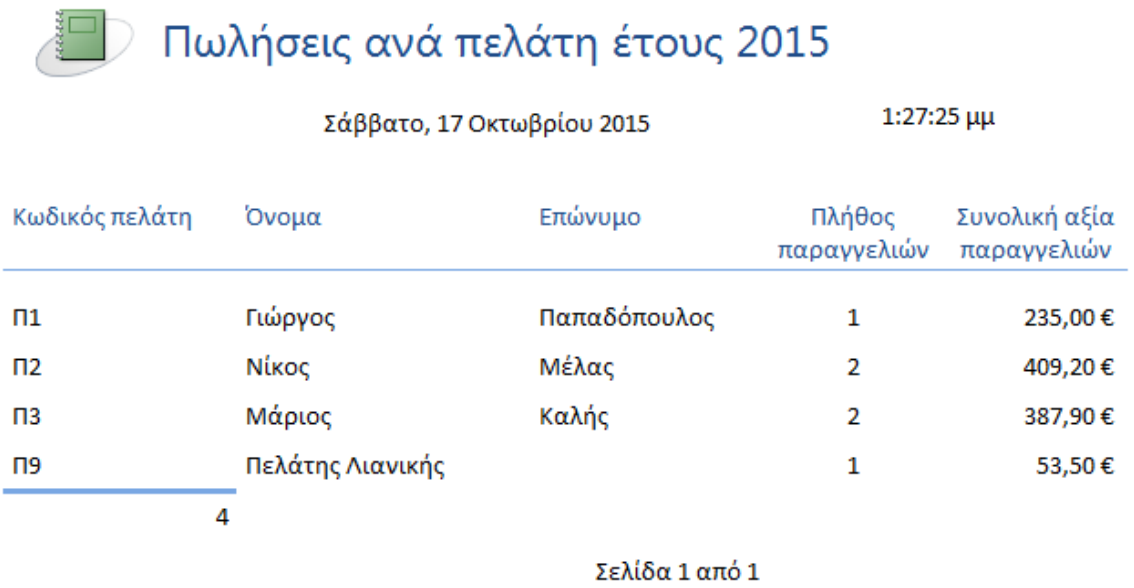
Σελίδα 1 από 1

Σχήμα 5.3. *Η αναφορά των πωλήσεων του έτους ανά προϊόν στην τελική της μορφή.*

Για τον υπολογισμό των πωλήσεων ανά πελάτη, απαιτείται τροποποίηση στο ερώτημα, ώστε να εμφανίζονται τα στοιχεία του πελάτη, και τα σύνολα των πωλήσεων να υπολογίζονται ομαδοποιημένα ανά πελάτη. Στο Σχήμα 5.4 παρουσιάζεται το κατάλληλο ερώτημα, που περιλαμβάνει τους πίνακες **ΠΕΛΑΤΕΣ** και **ΠΑΡΑΓΓΕΛΙΕΣ**. Εκτός από τα βασικά στοιχεία του πελάτη, από τον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ** εισάγουμε τα πεδία **Κωδ_παραγγελίας** και **Κόστος**. Ο κωδικός παραγγελίας, εφόσον είναι πρωτεύον κλειδί για τις παραγγελίες, μπορεί να χρησιμοποιηθεί για τη μέτρηση του αριθμού των παραγγελιών, επιλέγοντας **Πλήθος** στα Συγκεντρωτικά στοιχεία. Για το πεδίο **Κόστος** επιλέγουμε **Άθροισμα** ώστε να αθροίζονται οι αξίες των παραγγελιών για κάθε πελάτη. Μετά τη δημιουργία και έλεγχο του ερωτήματος, δημιουργούμε την κατάλληλη έκθεση για την παρουσίαση των αποτελεσμάτων. Στο Σχήμα 5.5 παρουσιάζεται το τελικό αποτέλεσμα που δίνει η έκθεση των πωλήσεων ανά πελάτη.



Σχήμα 5.4. *Η σχεδίαση του ερωτήματος για τον υπολογισμό των πωλήσεων ανά πελάτη*



Σχήμα 5.5. *Το τελικό αποτέλεσμα της αναφοράς*

Παρατηρήσεις: Τα παραπάνω παραδείγματα είναι απλουστευμένα σε σχέση με τις ανάγκες μιας πραγματικής εφαρμογής, ώστε να είναι κατανοητά και πιο εύκολα παρουσιάσιμα. Σε πιο ρεαλιστικές περιπτώσεις, θα περιμέναμε να υπάρχουν περισσότερα στοιχεία που θα έπρεπε να ληφθούν υπόψη, όπως: Ενώ στο παράδειγμα έχουμε ταυτίσει τις πωλήσεις με τις παραγγελίες, είναι πιθανό κάποιες παραγγελίες να μην έχουν ακόμα εκτελεστεί ή και να έχουν ακυρωθεί, επομένως δε θα έπρεπε να προσμετρηθούν. Επίσης, σε κάποιες παραγγελίες, θα μπορούσε να είχε εφαρμοστεί έκπτωση, της οποίας το ποσοστό να έχει καθοριστεί ανά πελάτη (π.χ. ο καλός πελάτης δικαιούται έκπτωση 10%), ανά προϊόν (π.χ. αν για το συγκεκριμένο προϊόν ισχύει προωθητική ενέργεια) ή ανά παραγγελία (π.χ. λόγω πληρωμής μετρητοίς εφαρμόζεται έκπτωση 1%). Ανάλογα με την περίπτωση, οι εκπτώσεις αυτές θα ήταν εμφανείς σε διαφορετικά πεδία διαφορετικών πινάκων και θα απαιτούνταν πιο σύνθετο ερώτημα για τον υπολογισμό της αξίας των πωλήσεων ανά προϊόν.

5.3 Αναζήτηση και αξιοποίηση πληροφορίας

Οι διοικητικές αναφορές είναι συνήθως τυποποιημένες και η χρήση τους δεν περιλαμβάνει παρέμβαση στα δεδομένα, παρά μόνο συγκέντρωση και παρουσίασή τους στην επιθυμητή μορφή, ώστε να αναδειχθεί η ζητούμενη πληροφορία. Αξιοποιώντας τα ίδια βασικά εργαλεία, που διαθέτει η Access (και οποιαδήποτε άλλη Βάση Δεδομένων) για την αναζήτηση και επεξεργασία δεδομένων, υπάρχει η δυνατότητα να εξάγουμε επιπλέον πληροφορία, προσαρμοσμένη στις ανάγκες των προβλημάτων που επιθυμούμε να επιλύσουμε.

Ένας στέλεχος διοίκησης ή μάρκετινγκ μπορεί να ενισχύσει τις δυνατότητες μιας εφαρμογής σε εξαγωγή πληροφορίας, επεκτείνοντας το σχήμα της Βάσης Δεδομένων, ώστε να τηρούνται επιπρόσθετα πρωτογενή δεδομένα, προσθέτοντας υπολογιζόμενα πεδία που εκφράζουν δευτερογενή δεδομένα και δημιουργώντας ειδικά σχεδιασμένα ερωτήματα που παράγουν την πληροφορία που απαιτείται για τη λήψη συγκεκριμένων αποφάσεων. Τα ειδικά αυτά ερωτήματα μπορεί να είναι προκατασκευασμένα, ώστε να είναι διαθέσιμα για τακτική χρήση, επιπρόσθετα όμως, ένας έμπειρος χρήστης μπορεί να εκμεταλλευτεί μια Βάση Δεδομένων «εξερευνητικά», δημιουργώντας και εκτελώντας κατά περίπτωση «έξυπνα» ερωτήματα, που μπορεί να οδηγήσουν σε ενδιαφέροντα και αξιοποιήσιμα συμπεράσματα. Με τους τρόπους αυτούς μπορούμε να ενσωματώσουμε σε μια εφαρμογή Βάσης Δεδομένων επιπλέον στοιχεία προς την κατεύθυνση της Επιχειρηματικής Ευφυΐας. Πληροφορία που μπορεί να προστεθεί σε μια Βάση Δεδομένων για τις ανάγκες της επιχειρηματικής ευφυΐας είναι π.χ. η πιστότητα/φερεγγυότητα του πελάτη, η επιβάρυνση και τα προβλήματα που μπορεί να προκαλεί ένας πελάτης (επιστροφές, μεταφορικά, κλπ.), χρόνος απαξίωσης προϊόντων, πολιτικές προώθησης που σχετίζονται με μη χρηματοοικονομικούς λόγους, κ.ά.. Πιθανά σενάρια αναζήτησης πληροφορίας στα δεδομένα είναι η εύρεση των καλύτερων και πιστότερων πελατών, ευκαιρίες διασταυρωμένων πωλήσεων, ο φόρτος εργασίας της κάθε αποθήκης και του κάθε εργαζομένου, το μερίδιο αγοράς σε κάθε πόλη, οι ημέρες του έτους με τη μεγαλύτερη δραστηριότητα και κερδοφορία, και πολλά άλλα. Επίσης, τα δεδομένα συναλλαγών, που κατά κανόνα καταγράφονται σε ένα πληροφοριακό σύστημα που υποστηρίζει τη λειτουργία μιας επιχείρησης, μπορούν να χρησιμεύσουν στον υπολογισμό δεικτών αποδοτικότητας της επιχείρησης σε διάφορους τομείς όπως ο χρηματοοικονομικός, η διαχείριση αποθεμάτων, το προσωπικό και η διοίκηση.

Στη συνέχεια της ενότητας αυτής, θα παρουσιαστούν δύο απλά παραδείγματα, που μπορούν να αποτελέσουν αφετηρία για πληθώρα αντίστοιχων εφαρμογών που θα μπορεί να υλοποιήσει ο αναγνώστης σύμφωνα με τις δικές του ανάγκες. Τα παραδείγματα που έχουν επιλεγεί βασίζονται στο κατάστημα ηλεκτρικών ειδών και είναι (α) η εύρεση των πιο επικερδών προϊόντων (β) η εύρεση των πελατών με πρόβλημα πίστωσης.

5.3.1 Τα πιο επικερδή προϊόντα

Για να διαπιστώσουμε ποια είναι τα πιο επικερδή προϊόντα της εταιρείας, επιθυμούμε μια λίστα των προϊόντων όπου να εμφανίζεται για καθένα από αυτά ο αριθμός τεμαχίων, η συνολική αξία και το συνολικό κέρδος, για όλες τις πωλήσεις του έτους 2015, ταξινομημένη κατά κέρδος. Επιπλέον, θα ήταν χρήσιμο να έχουμε την πληροφορία αυτή και ανά πόλη, ώστε να προσαρμόσουμε ανάλογα τις ενέργειες μάρκετινγκ.

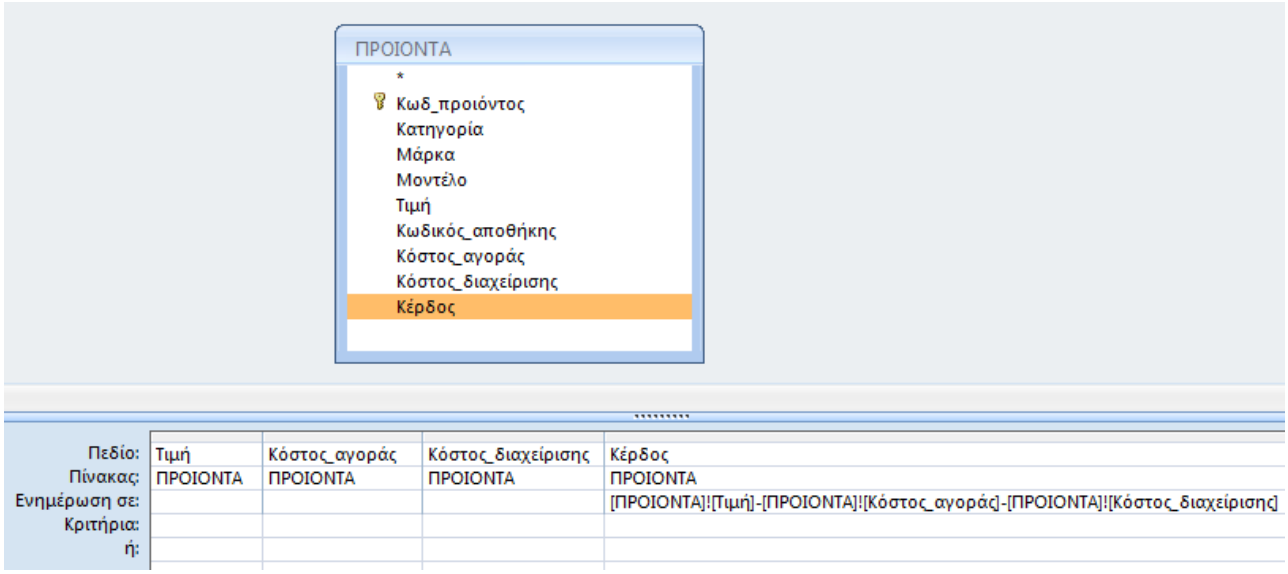
Διαπιστώνουμε ότι στη μέχρι τώρα υλοποίηση της Βάσης Δεδομένων για το κατάστημα ηλεκτρικών ειδών, δεν έχει συμπεριληφθεί η έννοια του κέρδους, επομένως για τον υπολογισμό των πιο επικερδών προϊόντων, είναι απαραίτητο να συμπληρωθεί πληροφορία για το κόστος και το κέρδος που αντιστοιχεί σε κάθε προϊόν. Σε μια σχετικά απλή περίπτωση, θεωρούμε ότι η τιμή πώλησης κάθε προϊόντος είναι σταθερή

και τηρείται στον πίνακα ΠΡΟΪΟΝΤΑ, μπορεί όμως σε κάποια παραγγελία να εφαρμόζεται έκπτωση, η οποία προσδιορίζεται ανά προϊόν (θα μπορούσε να υπάρχει έκπτωση, εναλλακτικά ή επιπρόσθετα, και γενικά στη συνολική παραγγελία). Επίσης θεωρούμε ότι το κόστος προσδιορίζεται από την τιμή αγοράς και το κόστος διαχείρισης, που εκλαμβάνονται ως σταθερά και γνωστά ανά προϊόν. Για την προσαρμογή της εφαρμογής απαιτούνται τα εξής βήματα:

Προσαρμογή της σχεδίασης της Βάσης Δεδομένων. Στον πίνακα **ΠΡΟΪΟΝΤΑ** προσθέτουμε τα πεδία **Κόστος_αγοράς**, **Κόστος_διαχείρισης** και **Κέρδος**. Στον πίνακα **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ** προσθέτουμε τα πεδία **Έκπτωση** και **Ποσό.**

Εισαγωγή επιπρόσθετων δεδομένων. Στα πεδία που δημιουργήθηκαν θα πρέπει να εισαχθούν τα αντίστοιχα δεδομένα. Θεωρούμε ότι τα δεδομένα αυτά είναι διαθέσιμα στην επιχείρηση και μπορούν να αντληθούν από τις αντίστοιχες πηγές δεδομένων (π.χ. από την εφαρμογή με την οποία διαχειριζόμαστε τους προμηθευτές και το δίκτυο διανομής μας ή τα σχετικά χειρόγραφα αρχεία). Οι τιμές των πεδίων **Κέρδος** και **Ποσό** μπορούν να υπολογιστούν και να εισαχθούν αυτόματα με χρήση ενός ερωτήματος ενημέρωσης. Στο Σχήμα 5.6 φαίνεται το κατάλληλο ερώτημα ενημέρωσης, που δημιουργήθηκε αφού προστέθηκαν στη Βάση Δεδομένων τα παραπάνω πεδία και συμπληρώθηκαν με τα σχετικά δεδομένα. Μετά την εκτέλεση του ερωτήματος, η τιμή του πεδίου Κέρδος ενημερώνεται με βάση τις τιμές των πεδίων **Τιμή**, **Κόστος_αγοράς** και **Κόστος_διαχείρισης** (Σχήμα 5.7), χρησιμοποιώντας την έκφραση:

[ΠΡΟΙΟΝΤΑ]![Τιμή]-[ΠΡΟΙΟΝΤΑ]![Κόστος_αγοράς]-[ΠΡΟΙΟΝΤΑ]![Κόστος_διαχείρισης]



Σχήμα 5.6. *Η σχεδίαση του ερωτήματος ενημέρωσης που υπολογίζει το κέρδος ανά τεμάχιο για κάθε προϊόν.*

Μάρκα	Μοντέλο	Τιμή	Κωδικός_αι	Κόστος_αγς	Κόστος_διαχείρισης	Κέρδος
PITSOS	P18-super	235,00 €	ΑΠ1	180,00 €	10,00 €	45,00 €
MORRIS	Clean 15	332,50 €	ΑΠ1	200,00 €	10,00 €	122,50 €
MORRIS	SC43	76,70 €	ΑΠ1	51,00 €	4,00 €	21,70 €
FIRST	G1	13,10 €	ΑΠ2	8,60 €	0,20 €	4,30 €
FIRST	G4	21,50 €	ΑΠ2	14,50 €	0,20 €	6,80 €
KARPA	F12	5,00 €	ΑΠ2	3,10 €	0,20 €	1,70 €
ELECTRO	Economy	210,00 €	ΑΠ1	155,00 €	10,00 €	45,00 €
ELECTRO	Clean-Economy	28,50 €	ΑΠ2	17,00 €	3,00 €	8,50 €

Σχήμα 5.7. *Το περιεχόμενο του πίνακα ΠΡΟΪΟΝΤΑ μετά την προσθήκη των κατάλληλων πεδίων και την ενημέρωση των τιμών τους.*

Αντίστοιχη διαδικασία πρέπει να πραγματοποιηθεί για να ληφθεί υπόψη η πιθανή έκπτωση που εφαρμόστηκε κατά την πώληση κάθε είδους. Εφόσον η έκπτωση αφορά συγκεκριμένο προϊόν σε συγκεκριμένη παραγγελία, αποτελεί χαρακτηριστικό της γραμμής παραγγελίας. Επομένως, για να αποθηκεύσουμε την πληροφορία σχετικά με την έκπτωση αυτή, προσθέτουμε στον πίνακα **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ** το πεδίο **Έκπτωση**, όπου ορίζεται το ποσοστό έκπτωσης.

Ερώτημα επιλογής. Εφόσον τώρα είναι διαθέσιμο το ποσό του κέρδους ανά τεμάχιο που προκύπτει από την πώληση κάθε προϊόντος (μη συμπεριλαμβανομένης τυχόν έκπτωσης), καθώς και το ποσοστό έκπτωσης που εφαρμόστηκε κατά περίπτωση σε συγκεκριμένες παραγγελίες, δημιουργούμε ερώτημα για τον υπολογισμό των συγκεντρωτικών μεγεθών τεμαχίων, αξίας και κέρδους ανά προϊόν για τη διάρκεια του έτους. Στο ερώτημα πρέπει να συμπεριληφθούν: από τον πίνακα **ΠΡΟΪΟΝΤΑ**, τα πεδία **Κωδ_προϊόντος, Μάρκα, Μοντέλο, Τιμή** και **Κέρδος**, από τον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ**, η **Ημνία_παραγγελίας** (ώστε να επιλεγούν οι παραγγελίες του έτους) και το **Ακυρώθηκε** (ώστε να μη ληφθούν υπόψη οι ακυρωμένες παραγγελίες), και από τον πίνακα **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ**, τα πεδία **Ποσότητα** και **Έκπτωση.**

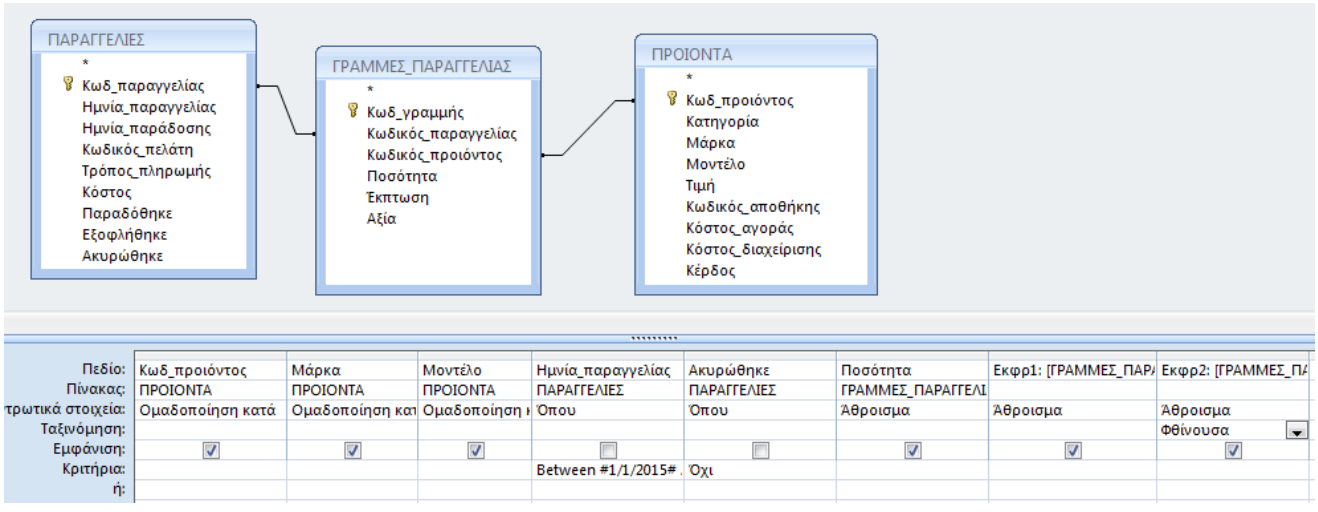
Για τον υπολογισμό των συνολικών τεμαχίων αρκεί η άθροιση της ποσότητας κάθε γραμμής παραγγελίας, που είναι διαθέσιμη στον **ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ**. Για τον υπολογισμό της συνολικής αξίας των πωλήσεων ανά προϊόν, πρέπει να αθροιστεί η αξία κάθε γραμμής πώλησης, αφού συνυπολογιστεί η τυχόν έκπτωση, όπως υπολογίζεται από την έκφραση:

[ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ]![Ποσότητα] * [ΠΡΟΙΟΝΤΑ]![Τιμή] * (1-[ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ]![Έκπτωση] /100)

Για τον υπολογισμό της συνολικού κέρδους των πωλήσεων ανά προϊόν, πρέπει να αθροιστεί το κέρδος ανά γραμμή παραγγελίας, που μπορεί να υπολογιστεί από την έκφραση:

[ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ]![Ποσότητα] * ([ΠΡΟΙΟΝΤΑ]![Κέρδος] - [ΠΡΟΙΟΝΤΑ]![Τιμή] * [ΓΡΑΜΜΕΣ_ΠΑΡΑΓΓΕΛΙΑΣ]![Έκπτωση] /100)

Στο Σχήμα 5.8 φαίνεται η σχεδίαση του τελικού ερωτήματος επιλογής, όπου έχουν εισαχθεί τα πεδία που αναφέρθηκαν παραπάνω, οι εκφράσεις για τον υπολογισμό της αξίας και του κέρδους, καθώς και τα κριτήρια σχετικά με την ημερομηνία (**Between #1/1/2015# And #31/12/2015#)** και την τιμή **Όχι** στο πεδίο **Ακυρώθηκε**. Όσον αφορά τα συγκεντρωτικά στοιχεία, στο **Κωδ_προϊόντος, Μάρκα** και **Μοντέλο** αφήνουμε το **Ομαδοποίηση κατά** (εφόσον ομαδοποιούμε κατά προϊόν), στο **Ημνία_παραγγελίας** και το **Ακύρωση** επιλέγουμε **Όπου** (ώστε να μην επηρεάσει την ομαδοποίηση), ενώ στο **Ποσότητα** και τις εκφράσεις για την αξία και το κέρδος επιλέγουμε **Άθροισμα**. Τέλος, επιλέγουμε φθίνουσα ταξινόμηση ως προς την Έκφραση 2, ώστε να εμφανίζονται τα προϊόντα ταξινομημένα σύμφωνα με το συνολικό κέρδος που μας έχουν αποφέρει.



Σχήμα 5.8. *Το ερώτημα επιλογής για τον υπολογισμό των τελικών αποτελεσμάτων*

Στο Σχήμα 5.9 φαίνεται το αποτέλεσμα της εκτέλεσης του ερωτήματος, σύμφωνα με το οποίο το πιο επικερδές προϊόν είναι αυτό με κωδικό Α2 (MORRIS Clean15), από το οποίο έχουν πωληθεί 2 τεμάχια, έχουν εισπραχθεί 615,13€ και μας έχει αποφέρει κέρδος 195,13€.

Κωδ_προιόντος	Μάρκα	Μοντέλο	ΆθροισμαΤουΠοσότητα	Εκφρ1	Εκφρ2
A2	MORRIS	Clean 15	2	615,125	195,125
A1	PITSOS	P18-super	2	470	90
A3	MORRIS	SC43	2	153,399993896484	43,3999938964844
A6	KARPA	F12	22	102,5	29,9000010490417
A4	FIRST	G1	4	52,4000015258789	17,2000007629395
A8	ELECTRO	Clean-Economy	1	28,5	8,5

Σχήμα 5.9. *Τα αποτελέσματα του ερωτήματος για τα πιο επικερδή προϊόντα.*

Έκθεση παρουσίασης αποτελεσμάτων. Το τελευταίο βήμα είναι η δημιουργία μιας έκθεσης (ή φόρμας) για την προβολή των αποτελεσμάτων. Στο Σχήμα 5.10 παρουσιάζεται το αποτέλεσμα μιας τέτοιας έκθεσης, που μπορεί να δημιουργηθεί πολύ εύκολα με χρήση του Οδηγού εκθέσεων και στη συνέχεια τελειοποίηση των τίτλων και της διάρθρωσης σε προβολή σχεδίασης.

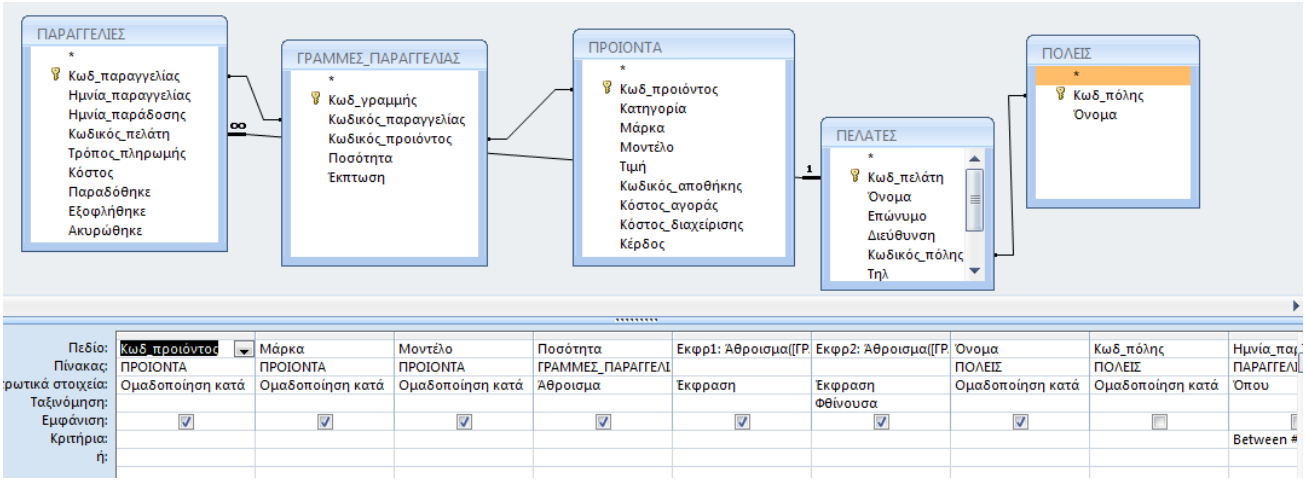
Τα πιο Επικερδή προϊόντα του 2015

Κωδικός προιόντος	Μάρκα	Μοντέλο	Συνολικά τεμάχια	Συνολική Αξία	Συνολικό κέρδος
A2	MORRIS	Clean 15	2	615,13 €	195,13 €
A1	PITSOS	P18-super	2	470,00 €	90,00 €
A3	MORRIS	SC43	2	153,40 €	43,40 €
A6	KARPA	F12	22	102,50 €	29,90 €
A4	FIRST	G1	4	52,40 €	17,20 €
A8	ELECTRO	Clean-Economy	1	28,50 €	8,50 €

Τρίτη, 20 Οκτωβρίου 2015 Σελίδα 1 από 1

Σχήμα 5.10. *Το τελικό αποτέλεσμα της αναφοράς.*

Για την εύρεση των πιο επικερδών προϊόντων ανά πόλη, θα πρέπει στο προηγούμενο ερώτημα να προσθέσουμε τον πίνακα **ΠΟΛΕΙΣ** και να εισάγουμε στο πλέγμα σχεδίασης τα πεδία **Κωδ_πόλης** (για να γίνει η ομαδοποίηση ανά πόλη) και **Όνομα** (ώστε να προβληθεί στο χρήστη το όνομα της πόλης, αντί του κωδικού). Επίσης πρέπει να προσθέσουμε τον πίνακα **ΠΕΛΑΤΕΣ**, επειδή η πληροφορία σχετικά με το από ποια πόλη δόθηκε κάποια παραγγελία προκύπτει από την πόλη στην οποία βρίσκεται ο πελάτης που την έδωσε, δηλαδή από την τιμή του πεδίου **Κωδικός_πόλης** του πίνακα **ΠΕΛΑΤΕΣ**. Οι δύο πίνακες που προστέθηκαν στο ερώτημα πρέπει να συνδεθούν όπως φαίνεται στο Σχήμα 5.11: οι **ΠΕΛΑΤΕΣ** με τις **ΠΑΡΑΓΓΕΛΙΕΣ** με βάση τον κωδικό πελάτη και οι **ΠΟΛΕΙΣ** με τους **ΠΕΛΑΤΕΣ** με βάση τον κωδικό πόλης. Σημειώνεται ότι στο ερώτημα δε συμμετέχει κανένα πεδίο του πίνακα **ΠΕΛΑΤΕΣ**, ωστόσο η συμμετοχή του στο ερώτημα είναι απαραίτητη για να συνδεθεί η πόλη με την παραγγελία.



Σχήμα 5.11. *Το ερώτημα για την εύρεση των πιο επικερδών προϊόντων ανά πόλη.*

Στο Σχήμα 5.12 φαίνεται η έκθεση που παρουσιάζει το αποτέλεσμα του παραπάνω ερωτήματος. Στη δημιουργία της έκθεσης αυτής με χρήση του οδηγού εκθέσεων, στο βήμα Επίπεδα ομαδοποίησης, επιλέχθηκε ως επίπεδο ομαδοποίησης το πεδίο **Όνομα** του πίνακα **ΠΟΛΕΙΣ**, ώστε τα προϊόντα να εμφανιστούν ομαδοποιημένα ανά πόλη. Από το αποτέλεσμα πληροφορούμαστε ότι στην Αθήνα το πιο επικερδές προϊόν είναι το A2 MORRIS Clean 15, που μας έχει αποφέρει κέρδος 122,5€, ενώ στη Θεσσαλονίκη το A1 PITSOS P18-super, που μας έχει αποφέρει 90€.

Τα πιο επικερδή προϊόντα ανά πόλη

Όνομα	Κέρδος	Κωδικός προιόντος	Μάρκα	Μοντέλο	Τεμάχια	Αξία
Αθήνα						
	122,50 €	A2	MORRIS	Clean 15	1	332,50 €
Θεσσαλονίκη						
	90,00 €	A1	PITSOS	P18-super	2	470,00 €
	72,63 €	A2	MORRIS	Clean 15	1	282,63 €
	43,40 €	A3	MORRIS	SC43	2	153,40 €
	18,80 €	A6	KARPA	F12	14	65,00 €
	17,20 €	A4	FIRST	G1	4	52,40 €

Τρίτη, 20 Οκτωβρίου 2015

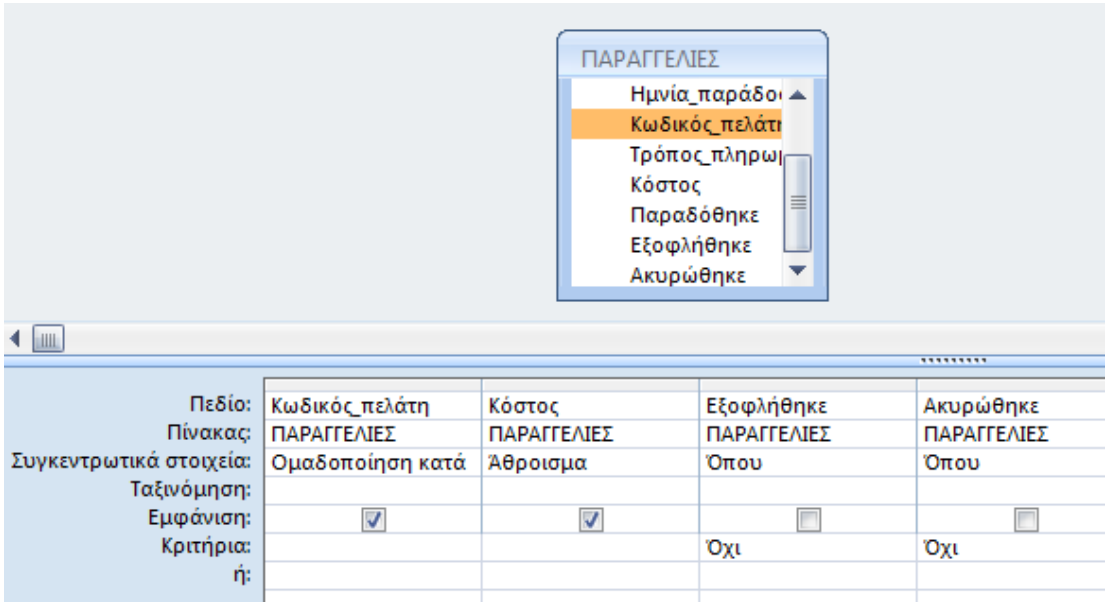
Σχήμα 5.12. *Η αναφορά σχετικά με τα πιο επικερδή προϊόντα ανά πόλη*

5.3.2 Πελάτες με πρόβλημα πίστωσης

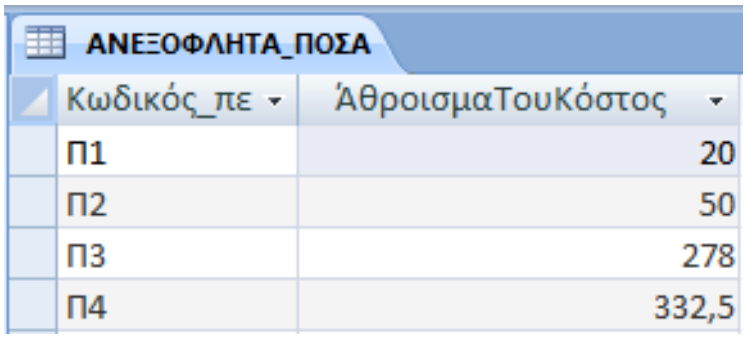
Στην επιχείρηση πώλησης ηλεκτρικών ειδών, θεωρούμε χάριν παραδείγματος, ότι μπορούμε να παρέχουμε πίστωση σε όποιους πελάτες επιθυμούμε, μέχρι ένα χρηματικό όριο που καθορίζει ο διευθυντής του οικονομικού. Επιθυμούμε να διαπιστώσουμε ποιοι πελάτες έχουν πρόβλημα πίστωσης, ώστε να επισημανθεί το γεγονός αυτό και επομένως να μπορεί ο πωλητής να αναστείλει την εκτέλεση των εκκρεμών παραγγελιών του πελάτη αυτού ή και να απορρίψει τη λήψη νέων παραγγελιών. Ως πρόβλημα πίστωσης θεωρούμε το γεγονός η συνολική αξία των παραγγελιών του πελάτη, οι οποίες είναι ανεξόφλητες, να υπερβαίνει το πιστωτικό του όριο.

Για τη λειτουργία αυτή, απαιτείται η ύπαρξη στον πίνακα **ΠΕΛΑΤΕΣ** ενός αριθμητικού πεδίου για την τήρηση του πιστωτικού του ορίου και ενός πεδίου **Πιστωτικό_πρόβλημα**, τύπου Ναι/Όχι, όπου να σημαίνεται το αν κάποιος πελάτης έχει πρόβλημα πίστωσης. Η τιμή του πεδίου **Πιστωτικό_πρόβλημα** θέλουμε να υπολογίζεται από την Access και να ενημερώνεται αυτόματα, σύμφωνα με τις τρέχουσες παραγγελίες του κάθε πελάτη. Ως πρώτο βήμα σε αυτό το παράδειγμα, προσαρμόζουμε τη σχεδίαση της Βάσης Δεδομένων, προσθέτοντας τα παραπάνω πεδία και εισάγοντας τις τιμές που αντιστοιχούν στο επιθυμητό πιστωτικό όριο του κάθε πελάτη.

Η ενημέρωση του πεδίου **Πιστωτικό_πρόβλημα**, ώστε να σημαίνονται οι πελάτες που έχουν ξεπεράσει το πιστωτικό τους όριο, μπορεί να γίνεται με ένα ερώτημα ενημέρωσης. Επειδή στα ερωτήματα ενημέρωσης δεν μπορούν να υπολογιστούν συγκεντρωτικά στοιχεία, απαιτείται να προηγηθεί η δημιουργία ενός βοηθητικού ερωτήματος, που να υπολογίζει τη συνολική αξία των ανεξόφλητων παραγγελιών κάθε πελάτη και να την αποθηκεύει σε κατάλληλο πίνακα, στον οποίο μπορεί να δοθεί το όνομα **ΑΝΕΞΟΦΛΗΤΑ_ΠΟΣΑ**. Το ερώτημα αυτό είναι τύπου **Δημιουργίας πίνακα**, και συμπεριλαμβάνει τον πίνακα **ΠΑΡΑΓΓΕΛΙΕΣ**, από τον οποίο εισάγονται τα πεδία **Κωδικός_πελάτη** (ώστε να γίνει η συγκέντρωση των ανεξόφλητων ποσών ανά πελάτη και να μπορεί να συνδεθεί με τους πελάτες), **Κόστος**, **Εξοφλήθηκε** και **Ακυρώθηκε** (ώστε να επιλεγούν μόνο οι μη ακυρωμένες παραγγελίες). Η σχεδίαση του ερωτήματος φαίνεται στο Σχήμα 5.13 και το αποτέλεσμα της εκτέλεσής του είναι ο πίνακας του Σχήματος 5.14.



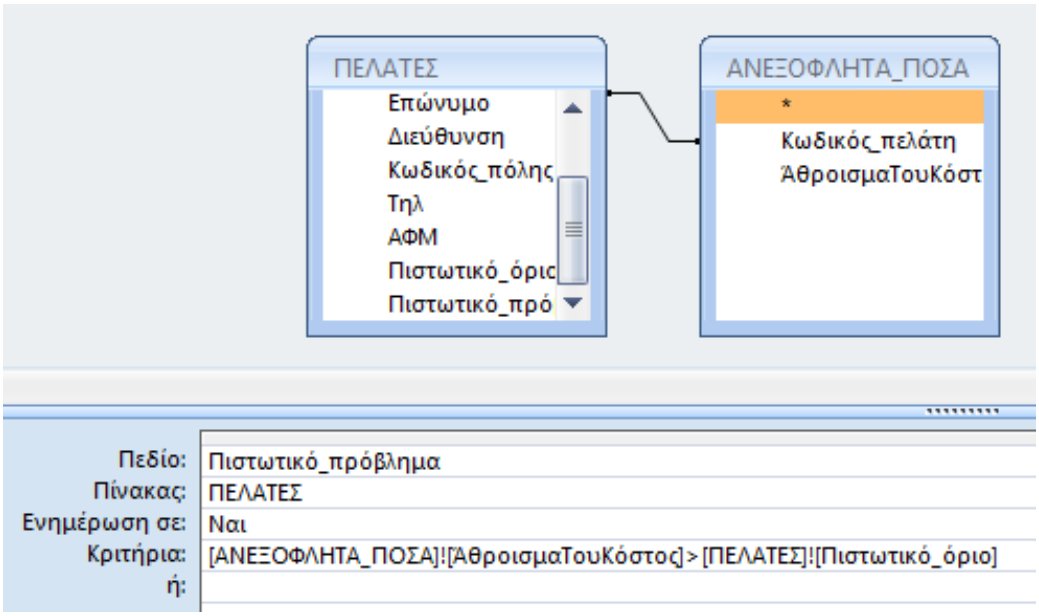
Σχήμα 5.13. *Το ερώτημα δημιουργίας πίνακα για τον υπολογισμό των ανεξόφλητων ποσών κάθε πελάτη.*



Σχήμα 5.14. *Το αποτέλεσμα υπολογισμού των ανεξόφλητων ποσών ανά πελάτη.*

Ένα ερώτημα ενημέρωσης (Σχήμα 5.15) μπορεί στη συνέχεια να ενημερώνει το πεδίο **Πιστωτικό_πρόβλημα** του πίνακα **ΠΕΛΑΤΕΣ**, ελέγχοντας για κάθε πελάτη αν το ανεξόφλητο ποσό είναι μεγαλύτερο από το πιστωτικό του όριο, εισάγοντας ως κριτήριο την έκφραση:

[ΑΝΕΞΟΦΛΗΤΑ_ΠΟΣΑ]![ΆθροισμαΤουΚόστος]>[ΠΕΛΑΤΕΣ]![Πιστωτικό_όριο]



Σχήμα 5.15. Ερώτημα ενημέρωσης που σημειώνει τους πελάτες με πιστωτικό πρόβλημα.

Ως τελευταίο βήμα, μπορούμε να δημιουργήσουμε ένα ερώτημα που να επιλέγει και να προβάλλει τα στοιχεία των πελατών με πιστωτικό πρόβλημα, καθώς και μια φόρμα, μέσω της οποίας να μπορεί ο χρήστης να εκτελέσει οποιαδήποτε στιγμή τη συνολική διαδικασία και να πληροφορηθεί για το αποτέλεσμα. Το ερώτημα επιλογής αρκεί να περιλαμβάνει τα πεδία του πίνακα **ΠΕΛΑΤΕΣ** που αντιστοιχούν στα βασικά στοιχεία του πελάτη και επίσης το πεδίο **Πιστωτικό_πρόβλημα** με κριτήριο την τιμή **Ναι**. Η φόρμα μπορεί να δημιουργηθεί με τη βοήθεια του οδηγού φορμών, στο πρώτο βήμα του οποίου επιλέγουμε το παραπάνω ερώτημα **Πελάτες_πιστωτικό_πρόβλημα** και εισάγουμε τα πεδία **Κωδ_πελάτη**, **Όνομα** και **Επώνυμο** (και όποιο άλλο στοιχείο του πελάτη επιθυμούμε να προβάλουμε. Η τιμή του πεδίου **Πιστωτικό_πρόβλημα** δε χρειάζεται να προβληθεί, επειδή είναι γνωστό ότι θα είναι **Ναι**. Μετά την ολοκλήρωση του οδηγού, πρέπει να μεταβούμε σε προβολή σχεδίασης για να τελειοποιήσουμε τη φόρμα. Εκτός από τις αισθητικές βελτιώσεις, είναι ιδιαίτερα χρήσιμο να προσθέσουμε ένα Κουμπί με το οποίο να μπορεί ο χρήστης να εκτελέσει ενημέρωση των υπολογισμών, ώστε να επικαιροποιείται το αποτέλεσμα πριν προβληθεί.

Η προσθήκη κουμπιού ενημέρωσης, με το οποίο θα πρέπει να εκτελούνται τα ερωτήματα ενημέρωσης που παρουσιάστηκαν στις προηγούμενες παραγράφους, μπορεί να δημιουργηθεί ως εξής:

- Δημιουργούμε χώρο υποσέλιδου, σύροντας τη γραμμή που ορίζει το σχετικό όριο.
- Επιλέγουμε στην καρτέλα των στοιχείων ελέγχου το **Κουμπί,** ενεργοποιούμε τον Οδηγό (το κουμπάκι με το μαγικό ραβδί) και δημιουργούμε ένα νέο κουμπί στο σημείο του υποσέλιδου και στο μέγεθος που επιθυμούμε.
- Στο πρώτο βήμα του οδηγού κουμπιών εντολής, επιλέγουμε **Διάφορα** και **Εκτέλεση ερωτήματος.** Στη συνέχεια επιλέγουμε το επιθυμητό κείμενο που θα τοποθετηθεί ως λεζάντα πάνω στο κουμπί π.χ. «Ενημέρωση».
- Μετά την ολοκλήρωση του οδηγού, κάνουμε δεξί κλικ πάνω στο κουμπί και επιλέγουμε **Δόμηση συμβάντος**, ώστε να συμπληρώσουμε τις ενέργειες που πρέπει να εκτελεστούν. Η πρώτη ενέργεια, που έχει ήδη εισαχθεί, είναι η εκτέλεση του βοηθητικού ερωτήματος υπολογισμού των ανεξόφλητων ποσών ανά πελάτη. Στην επόμενη γραμμή, επιλέγουμε ως

ενέργεια **Άνοιγμα ερωτήματος** και στα ορίσματα επιλέγουμε το ερώτημα **Ενημέρωση_πιστωτικού_προβλήματος**. Αποθηκεύοντας, έχουμε ολοκληρώσει την προσθήκη ενός κουμπιού με το οποίο εκτελούνται τα ερωτήματα υπολογισμού και ενημέρωσης που απαιτούνται για να είναι έγκυρη η πληροφορία που θα προβάλει η φόρμα εύρεσης των πελατών με πιστωτικό πρόβλημα.

- Η ολοκληρωμένη φόρμα σε λειτουργία φαίνεται στο Σχήμα 5.16.

Σχήμα 5.16. *Η τελική φόρμα αναζήτησης και προβολής των πελατών με πιστωτικό πρόβλημα.*

5.4 Οι κύβοι Άμεσης Αναλυτικής Επεξεργασίας (OnLine Analytical Processing)

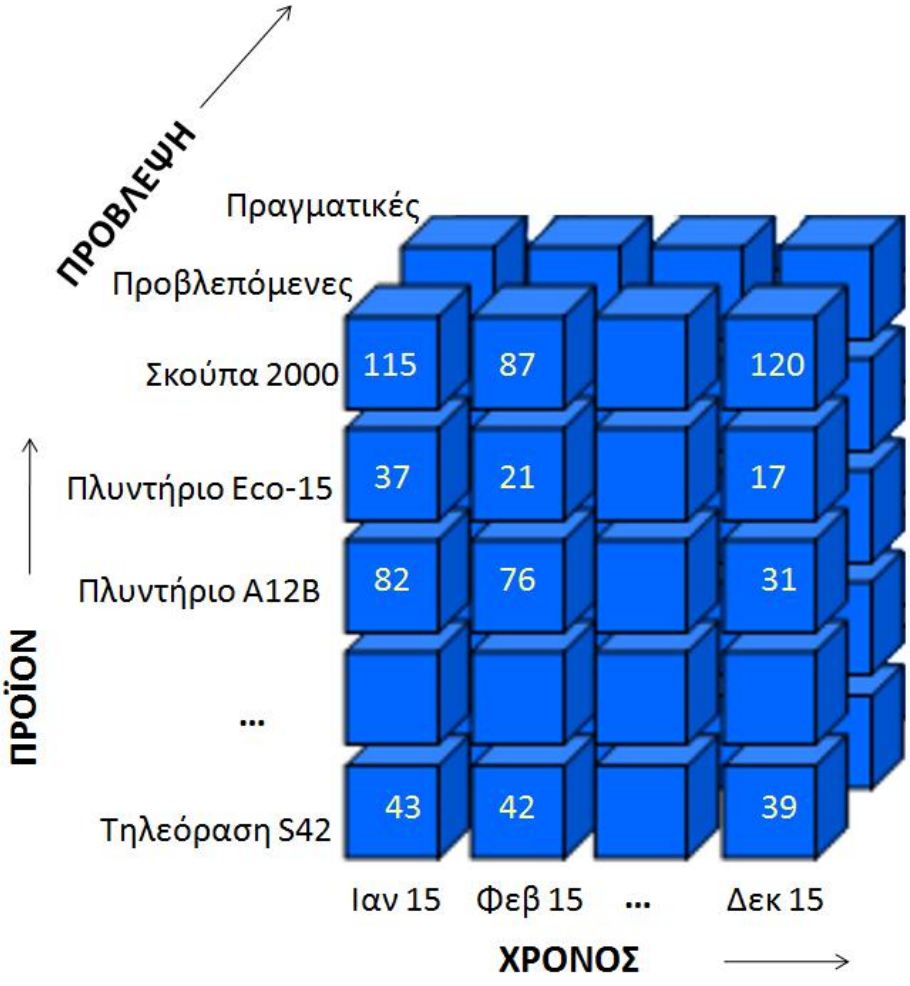
Οι σχεσιακές Βάσεις Δεδομένων είναι το βασικό εργαλείο χειρισμού δεδομένων σε μια επιχείρηση. Επειδή όμως έχουν σχεδιαστεί κυρίως για τη διαχείριση των δεδομένων, δηλαδή αποθήκευση, ενημέρωση και αναζήτηση, παρουσιάζουν κάποια μειονεκτήματα όταν επιθυμούμε να επεξεργαστούμε μεγάλους όγκους δεδομένων, ώστε να εξάγουμε υψηλού επιπέδου πληροφορία. Για το σκοπό αυτό, έχουν αναπτυχθεί Βάσεις Δεδομένων με ιδιαίτερη δομή, κατάλληλες για την αποτελεσματική επεξεργασία και ανάλυση των δεδομένων. Κύριος εκπρόσωπος της κατηγορίας αυτής είναι οι Βάσεις Άμεσης Αναλυτικής Επεξεργασίας (OnLine Analytical Processing Databases), γνωστές με το ακρωνύμιο OLAP. Οι Βάσεις OLAP υποστηρίζουν την προβολή της πληροφορίας με τη βοήθεια σχημάτων πολλών διαστάσεων, που ονομάζονται κύβοι (cubes).

Οι κύβοι OLAP αποτελούν σημαντικό εργαλείο επιχειρηματικής ευφυΐας, που επιτρέπει την «εξερεύνηση» των δεδομένων με άμεσο τρόπο, αντί της προετοιμασίας και εκτέλεσης ερωτημάτων, όπως θα απαιτούνταν σε μια σχεσιακή Βάση Δεδομένων. Οι κύβοι είναι διαδραστικά εργαλεία που επιτρέπουν στους χρήστες να βλέπουν την πληροφορία με τμηματικό τρόπο και με βαθμό λεπτομέρειας που εξελίσσεται όσο προχωρά η ανάλυση της πληροφορίας σε βάθος. Βασικό στοιχείο στην προσέγγιση OLAP είναι η δυνατότητα καθορισμού των διαστάσεων που έχουν ενδιαφέρον, καθώς και της κλίμακας κατά την οποία πραγματοποιείται συγκέντρωση των δεδομένων, κατά μήκος κάθε διάστασης.

Ας θεωρήσουμε ως παράδειγμα μια από τις πιο δημοφιλείς εφαρμογές των κύβων OLAP, που είναι η μελέτη των πωλήσεων της επιχείρησης. Η πληροφορία που θα ενδιέφερε ένα στέλεχος της επιχείρησης, και που μπορούμε με μεθόδους Επιχειρηματικής Ευφυΐας να εξάγουμε από τα δεδομένα δοσοληψιών, είναι απαντήσεις σε προβλήματα όπως τα ακόλουθα:

- Ποιες είναι οι συνολικές πωλήσεις του έτους ανά προϊόν και ποιες ανά κατηγορία προϊόντος;
- Πώς μεταβάλλεται το κέρδος ανά έτος, ανά μήνα ή ανά εβδομάδα;
- Ποια είναι η απόκλιση ανάμεσα στους στόχους και τα αποτελέσματα σε κάθε υποκατάστημα;
- Πώς σχετίζεται το ύψος των πωλήσεων με την ηλικία των πελατών;

Μπορούμε να σκεφτούμε τις απαντήσεις στα παραπάνω προβλήματα ως πίνακες που παρουσιάζουν συγκεντρωτικά στοιχεία ομαδοποιημένα ως προς μία μεταβλητή (π.χ. οι συνολικές πωλήσεις ενός έτους ανά προϊόν) ή που διασταυρώνουν 2 μεταβλητές (όπως π.χ. ένας δυσδιάστατος πίνακας όπου σε κάθε κελί του παρουσιάζει τις πωλήσεις ανά προϊόν και ανά έτος). Κάθε τέτοιος πίνακας αφορά 2 διαστάσεις, που αντιστοιχούν στις μεταβλητές που διασταυρώνονται. Οι διαφορετικές διαστάσεις (ή με άλλα λόγια οι μεταβλητές) που μπορεί να ενδιαφέρουν έναν αναλυτή, είναι περισσότερες, όπως είναι τα προϊόντα, ο χρόνος, ο χώρος, πρόβλεψη και αποτέλεσμα, και πολλές άλλες που μπορούν να προκύψουν από οποιοδήποτε πεδίο οποιασδήποτε οντότητας π.χ. η ηλικία των πελατών, η τιμή του προϊόντος, ο μισθός του υπαλλήλου, κ.ά. Επίσης, η ποσότητα που υπολογίζεται, μπορεί να αφορά ένα άθροισμα, μια καταμέτρηση, ένα στατιστικό στοιχείο ή οποιονδήποτε άλλον υπολογισμό.



Σχήμα 5.17. *Ένας νοητός κύβος OLAP των συνολικών πωλήσεων, όπου φαίνονται 3 από τις διαστάσεις του.*

Ένας κύβος OLAP διαθέτει προ-υπολογισμένα συγκεντρωτικά δεδομένα που έχουν υπολογιστεί με διαφορετικούς τρόπους ομαδοποίησης, με βάση όλες τις διαστάσεις που έχουν οριστεί. Είναι ένας πολυδιάστατος πίνακας, που ο χρήστης μπορεί να «περιστρέψει» νοητά και να επιλέξει για επισκόπηση οποιαδήποτε δυσδιάστατη προβολή του («φέτα»). Ο χρήστης μπορεί επίσης να μεταβάλει την κλίμακα, ή με άλλα λόγια, το βαθμό λεπτομέρειας που επιθυμεί για κάθε διάσταση. Στο Σχήμα 5.17 παρουσιάζεται ένας

κύβος που αναφέρεται στο ύψος των πωλήσεων, με 3 διαστάσεις: προϊόν, χρόνος και προβλεπόμενες-πραγματικές (σε μια πραγματική εφαρμογή θα είχε περισσότερες από 3 διαστάσεις, ασχέτως του ότι δε θα μπορούσαν να παρουσιαστούν όλες μαζί σε ένα τέτοιο σχήμα). Επιλέγοντας την όψη που αντιστοιχεί στο επίπεδο που ορίζεται από τους άξονες **Προϊόν** και **Χρόνος**, και τη συγκεκριμένη «φέτα» που αντιστοιχεί στην τιμή **Πραγματικές** του τρίτου άξονα, προβάλλονται οι πραγματικές πωλήσεις ανά προϊόν και ανά μήνα. Επιλέγοντας το ίδιο επίπεδο, αλλά τη φέτα που αντιστοιχεί στις **Προβλεπόμενες** πωλήσεις, έχουμε πρόσβαση στις προβλέψεις ανά προϊόν και ανά μήνα. Περιστρέφοντας τον κύβο ώστε να κοιτάζουμε το επίπεδο που ορίζεται από τους άξονες **Προϊόν** και **Προβλεπόμενες-πραγματικές**, μπορούμε να δούμε τη σύγκριση προβλεπόμενων-πραγματικών πωλήσεων για κάθε προϊόν και ειδικά για το συγκεκριμένο μήνα που αντιστοιχεί στη φέτα που θα επιλέξουμε. Αντίστοιχα, περιστρέφοντας τον κύβο, μπορούμε να βλέπουμε τη σύγκριση προβλεπόμενων-πραγματικών πωλήσεων ανά μήνα και για ένα συγκεκριμένο προϊόν της επιλογής μας. Σημαντική είναι επίσης η δυνατότητα μεταβολής της κλίμακας σε κάθε διάσταση, ώστε η πληροφορία να αντιστοιχεί στον επιθυμητό βαθμό λεπτομέρειας. Η διάσταση του χρόνου, που στο παράδειγμα του σχήματος κλιμακώνεται ανά μήνα, μπορεί να οριστεί έτσι ώστε να συγκεντρώνει τις πωλήσεις ανά έτος ή ανά ημέρα, αναλόγως του αν επιθυμούμε μεγαλύτερο ή μικρότερο βαθμό λεπτομέρειας. Αντίστοιχα, κατά τη διάσταση του προϊόντος, μπορούμε να ομαδοποιήσουμε τα προϊόντα ανά είδος ή ανά κατηγορία είδους.

Η πληροφορία που παρέχει ένας κύβος OLAP βασίζεται στη συγκέντρωση δεδομένων και τη συγκριτική παρουσίαση επεξεργασμένης πληροφορίας. Όπως μπορεί όμως να παρατήρησε κάποιος αναγνώστης, η πληροφορία αυτή είναι παρόμοια με αυτήν που παράγεται με τη χρήση ερωτημάτων σε μια σχεσιακή Βάση Δεδομένων, σύμφωνα με την προηγούμενη ενότητα του κεφαλαίου αυτού. Επομένως ποια είναι η αξία των κύβων OLAP και γιατί να μην αρκεστεί ένας αναλυτής στη χρήση ερωτημάτων και διοικητικών αναφορών; Η απάντηση είναι ότι η τεχνική OLAP έχει ως κύριο πλεονέκτημα τη δυνατότητα άμεσης ανταπόκρισης σε οποιοδήποτε ερώτημα, που συνδυάζει οποιεσδήποτε παραμέτρους σε οποιοδήποτε επίπεδο λεπτομέρειας, χωρίς να απαιτείται δημιουργία και εκτέλεση ερωτήματος. Τα δεδομένα έχουν ήδη μετατραπεί σε μια δομή όπου όλα τα στοιχεία έχουν συγκεντρωθεί ως προς όλες τις διαστάσεις και ως προς όλα τα δυνατά επίπεδα. Έτσι, είναι σαν να λέμε ότι ένας κύβος «περιέχει ήδη όλες τις απαντήσεις από πριν» και αρκεί να επιλέξουμε τον τρόπο με τον οποίο επιθυμούμε να τις δούμε. Αντίθετα, σε μια σχεσιακή Βάση Δεδομένων, η δομή των δεδομένων είναι τέτοια, που η παραγωγή αντίστοιχης πληροφορίας απαιτεί σύνθετα ερωτήματα με χρονοβόρα και δαπανηρή εκτέλεση. Η εξέταση π.χ. όλων των πωλήσεων και των αντίστοιχων προβλέψεων ανά προϊόν και ανά χρονική περίοδο, αλλά και ανά περιοχή και ανά ποσοστό έκπτωσης που προσφέρθηκε σε τυχόν προωθητική ενέργεια, θα απαιτούσε, για κάθε προβολή, την επεξεργασία χιλιάδων εγγραφών, δηλαδή μεγάλη υπολογιστική ισχύ και χρονική καθυστέρηση, ενώ ένας κύβος θα παρείχε την απάντηση αστραπιαία. Επιπλέον, στα θετικά στοιχεία των κύβων OLAP συγκαταλέγεται και το ότι η προβολή της πληροφορίας μπορεί να γίνει μέσω γνωστών και εύχρηστων εργαλείων, όπως λογιστικά φύλλα (τύπου Excel), προγραμμάτων πλοήγησης διαδικτύου ή γραφικών προγραμμάτων επισκόπησης δεδομένων, όπως το Microsoft Data Analyzer.

Οι Βάσεις Δεδομένων OLAP έχουν διαφορετική δομή από το σχεσιακό μοντέλο. Συγκριτικά με το τελευταίο, έχουν λιγότερους πίνακες, συνδεδεμένους μεταξύ τους σε σχήμα αστέρα. Στο κέντρο του αστέρα βρίσκεται ένας πίνακας «γεγονότων» (Facts), που στο παράδειγμα των πωλήσεων, θα έχει ως πεδία την αξία των πωλήσεων, την ποσότητα, το βάρος και την έκπτωση. Οι υπόλοιποι πίνακες δεν αντιστοιχούν στις γνωστές οντότητες, αλλά στις διαστάσεις που συμμετέχουν στο μοντέλο, όπως ο χρόνος, το προϊόν, ο πελάτης και ο εργαζόμενος.

Οι Βάσεις Δεδομένων OLAP δεν μπορούν να χρησιμοποιηθούν για τη διαχείριση δεδομένων συναλλαγών και τυπικά λειτουργούν μόνο για ανάγνωση. Επομένως, δεν μπορούν να αντικαταστήσουν τη σχεσιακή Βάση Δεδομένων, αλλά να τη συμπληρώσουν. Πρακτικά, μια επιχείρηση διαθέτει μία ή περισσότερες σχεσιακές Βάσεις για τις λειτουργικές της ανάγκες. Εφόσον επιθυμεί ένα ισχυρό εργαλείο επιχειρηματικής ευφυΐας, εγκαθιστά επιπλέον μια Βάση OLAP και το αντίστοιχο λογισμικό προβολής του κύβου. Η Βάση OLAP τροφοδοτείται από τη σχεσιακή Βάση και, μέσω ειδικών μηχανισμών, αντιγράφει το περιεχόμενο, εκτελεί τους απαραίτητους υπολογισμούς και το προσαρμόζει στη δική της δομή, χωρίς να μπορεί να αλλοιώσει το περιεχόμενο της σχεσιακής Βάσης.

Η παρουσίαση πρακτικών οδηγιών και εκτελέσιμων παραδειγμάτων χρήσης κύβων OLAP είναι εκτός των πλαισίων αυτού του βιβλίου. Σημειώνεται όμως ότι ο αναγνώστης που ενδιαφέρεται να χρησιμοποιήσει το εργαλείο αυτό, θα βρει πληθώρα εμπορικού ή και ελεύθερου λογισμικού, κατάλληλου για να αναπτύξει τις δικές του εφαρμογές επιχειρηματικής ευφυΐας (π.χ. http://olap.com/ , Pentaho, Jasper Reports Server, Mondrian, κ.ά.), συμπεριλαμβανομένων πακέτων μεγάλων εταιρειών, όπως Microsoft και IBM. Πολλά από

τα διαθέσιμα προγράμματα είναι σε μορφή προσαρτημάτων, έτσι ώστε να λειτουργούν σε συνεργασία με γνώριμα περιβάλλοντα επεξεργασίας δεδομένων, όπως το Microsoft Excel ή το Google docs. Αναφέρεται ενδεικτικά ότι Microsoft διαθέτει λειτουργίες OLAP ενσωματωμένες σε κάποιες εκδόσεις του γνωστού πακέτου Office, ως επέκταση του Excel, ενώ επίσης διαθέτει το λογισμικό για υποδομή OLAP, ως τμήμα της Βάσης Δεδομένων SQL server, με το όνομα Analysis Services (τελευταία έκδοση Analysis Services 2014).

Βιβλιογραφία/Αναφορές

Roiger R. J., & Geatz M. W. (2008). *Εξόρυξη Πληροφορίας – Ένας εισαγωγικός οδηγός με παραδείγματα*, Αθήνα: Εκδόσεις Κλειδάριθμος.

Microsoft Office support (2015, October 10). *Overview of Online Analytical Processing (OLAP)*. Retrieved from https://support.office.com/en-us/article/Overview-of-Online-Analytical-Processing-OLAP-15d2cdde-f70b-4277-b009-ed732b75fdd6

Κεφάλαιο 6. Μέθοδοι εξόρυξης γνώσης από δεδομένα

Σύνοψη
Στο κεφάλαιο αυτό, ο αναγνώστης εισάγεται στις μεθόδους εξόρυξης γνώσης από δεδομένα, που βασίζονται στις αρχές της στατιστικής και της μηχανικής μάθησης, αλλά και στην ισχύ των σύγχρονων υπολογιστικών συστημάτων. Παρουσιάζονται αντιπροσωπευτικές τεχνικές εξόρυξης γνώσης από δεδομένα, όπως εύρεση κανόνων συσχέτισης και δέντρα αποφάσεων, ώστε να κατανοηθούν ποιοτικά και να εντυπωθούν στον αναγνώστη τα πεδία εφαρμογής τους, οι δυνατότητες και αδυναμίες τους, καθώς και ο ρόλος των βασικών παραμέτρων προς ρύθμιση. Στο τρέχον κεφάλαιο, οι μέθοδοι επεξηγούνται θεωρητικά με χρήση απλών παραδειγμάτων, ώστε ο αναγνώστης να μπορεί, προχωρώντας στο Κεφάλαιο7, να τις εφαρμόσει σε δικά του πραγματικά προβλήματα. Ταυτόχρονα, ο αναγνώστης εισάγεται σε τεχνικές ανάλυσης δεδομένων που μπορούν να χρησιμοποιηθούν στη λήψη αποφάσεων γενικότερα, αλλά και ειδικότερα στα συστήματα συστάσεων (recommender systems) και στην εξατομίκευση των υπηρεσιών των συστημάτων web (web personalization). Οι τεχνικές αυτές βασίζονται στις αρχές της των μαθηματικών, της επιχειρησιακής έρευνας και της τεχνητής νοημοσύνης. Στο κεφάλαιο αυτό, παρουσιάζονται ενδεικτικές τεχνικές και αριθμητικά παραδείγματα που δείχνουν τα βήματα εφαρμογής τους, όπως και παραδείγματα εφαρμογών στην πράξη με χρήση κατάλληλων συστημάτων λογισμικού, όπως το Excel, το Expert Choice.

Προαπαιτούμενη γνώση
Κεφάλαιο 1. Εισαγωγή στη βασισμένη σε δεδομένα επιχειρηματική ευφυΐα, Κεφάλαιο 2. Δεδομένα και Πληροφορίες.

6.1 Εισαγωγή στις ευφυείς μεθόδους λήψης αποφάσεων

Η εφαρμογή τεχνικών ανάλυσης δεδομένων αποσκοπεί στη βελτίωση της λήψης των αποφάσεων, η οποία μπορεί να επιτευχθεί μέσα από την πληρέστερη κατανόηση των θεμάτων που πρέπει να διερευνηθούν πριν ληφθεί μια απόφαση. Η βαθύτερη κατανόηση ενός θέματος προϋποθέτει την όσο το δυνατόν πιο ολοκληρωμένη θεώρηση και μελέτη των μεταβλητών που το επηρεάζουν και κατά συνέπεια επηρεάζουν την/τις αποφάσεις σχετικές με το θέμα αυτό. Στο κεφάλαιο αυτό, θα παρουσιαστούν οι τεχνικές Analytic Hierarchy Process (AHP) και η Fuzzy Analytic Hierarchy Process (FAHP), ως ενδεικτικές τεχνικές πολυκριτηριακής ανάλυσης, τεχνικές ομοιότητας (similarity methods) και τεχνικές βασισμένες σε κανόνες (rule based). Οι τεχνικές ανάλυσης δεδομένων εφαρμόζονται στα πλαίσια της διαδικασίας λήψης αποφάσεων. Η διαδικασία λήψης αποφάσεων ολοκληρώνεται, σύμφωνα με τον Simon (1977), σε τρεις φάσεις, ενώ μία τέταρτη προστέθηκε αργότερα. Η απεικόνιση των φάσεων και της διαδικασίας λήψης αποφάσεων παρουσιάζεται στον παρακάτω πίνακα.

Φάση I (Intelligence Phase)
• Σύγκριση Επιχειρηματικών Στόχων σχετικών με μία περιοχή επιχειρηματικής δραστηριότητας και προσδιορισμός τυχόντων αποκλίσεων. • Έρευνα και συλλογή δεδομένων. • Προσδιορισμός του προβλήματος ή και ευκαιριών και της/των αποφάσεων που πρέπει να ληφθούν. • Ταξινόμηση του προβλήματος-απόφασης . • Διατύπωση του προβλήματος και της/των αποφάσεων που πρέπει να αντιμετωπισθούν.
Φάση II (Design Phase)
• Κατανόηση του προβλήματος. • Ανάπτυξη ενός ή περισσότερων μοντέλων για την μελέτη του προβλήματος-απόφασης. • Αξιολόγηση των μοντέλων. • Καθορισμός των κριτηρίων με βάση τα οποία θα επιλεγούν οι εναλλακτικές λύσεις. • Έρευνα για τον προσδιορισμό των εναλλακτικών λύσεων. • Εκτίμηση και εάν είναι δυνατόν Μέτρηση των αναμενόμενων αποτελεσμάτων.
Φάση III (Choice Phase)
• Επίλυση του/των μοντέλων. • Ανάλυση ευαισθησίας επί των λύσεων από το/τα μοντέλα. • Επιλογή της καταλληλότερης λύσης σύμφωνα με τα κριτήρια που έχουν προσδιορισθεί στη φάση του Σχεδιασμού. • Προγραμματισμός για την υλοποίηση των επιλογών. • Δημιουργία μηχανισμού ελέγχου για τον έλεγχο της υλοποίησης.
Φάση IV(Implementation Phase)
• Υλοποίηση των αποφάσεων.

Πίνακας 6.1 *Η Διαδικασία Επίλυσης Προβλημάτων και Λήψης Αποφάσεων.*

Κατά τη Intelligence Phase, οι υπεύθυνοι για τη λήψη της απόφασης, πρέπει να ανιχνεύουν συνεχώς ή κατά τακτά χρονικά διαστήματα, το περιβάλλον (επιχειρηματικό, τεχνολογικό, κοινωνικό, πολιτικό, νομικό) για τον εντοπισμό προβλημάτων αλλά και ευκαιριών. Η φάση αυτή αρχίζει με την μελέτη των επιχειρηματικών στόχων και σκοπών της απόφασης. Σκοπός της φάσης αυτής είναι να εντοπισθούν και να ορισθούν τα προβλήματα που οφείλονται από τις αποκλίσεις της πραγματικής απόδοσης της επιχείρησης από την επιθυμητή. Προσοχή χρειάζεται κατά τον προσδιορισμό ενός προβλήματος, ώστε να ορισθεί επ' ακριβώς το πρόβλημα αυτό καθ' αυτό και να μην περιγραφούν απλά τα συμπτώματα του προβλήματος. Στη διάρκεια της δεύτερης φάσης (Design Phase), δημιουργούνται και αναλύονται εναλλακτικά σχέδια δράσης για την επίλυση του προβλήματος, όπως αυτό έχει διατυπωθεί κατά την προηγούμενη φάση. Στη φάση αυτή αναπτύσσονται μοντέλα, (μαθηματικά, στατιστικής, τεχνητής νοημοσύνης, επιχειρησιακής έρευνας, κλπ.) και εφαρμόζονται τεχνικές και μέθοδοι ανάλυσης δεδομένων με τις οποίες διερευνώνται οι επιδράσεις διαφόρων παραγόντων (που είναι οι μεταβλητές στα διάφορα μοντέλα) πάνω στη λύση ή τις πιθανές εναλλακτικές λύσεις του προβλήματος. Ακολουθεί η Choice Phase κατά την οποία εξετάζονται και αξιολογούνται οι εναλλακτικές λύσεις του προβλήματος. Τα όρια και ο διαχωρισμός της Design Phase από την Choice Phase δεν είναι πάντα ευδιάκριτα. Αυτό συμβαίνει γιατί μπορεί να χρειαστεί η αναθεώρηση των μοντέλων αλλά και των κριτηρίων επιλογής των αποφάσεων που έχουν καθορισθεί στη διάρκεια της Design Phase, ενώ έχει ήδη αρχίσει η φάση της επιλογής της λύσης στην Choice Phase. Η επιλογή της «άριστης» λύσης δεν είναι πάντα εφικτή. Σημαντικά είναι τα παρακάτω ερωτήματα.

- Ποια είναι η άριστη λύση και για ποιον; Διαφορετικά στελέχη ή ομάδες ανθρώπων έχουν διαφορετικές προσδοκίες, πολύ δε περισσότερο διαφορετική αντίληψη όσον αναφορά στο πρόβλημα και τις πιθανές του λύσεις.
- Υπάρχει άριστη λύση; Μπορεί να υποστηριχθεί ότι υπάρχει πλήρης γνώση του προβλήματος και όλων των συναφών με αυτό επιπτώσεων;
- Κάτω από ποιες προϋποθέσεις είναι άριστη μια λύση;

Θα πρέπει εδώ να διαχωριστούν δύο έννοιες που προτάθηκαν αρχικώς από τον Simon (1977). Η έννοια της ορθολογικής (rational) και της ικανοποιητικής (satisfying) απόφασης, η οποία θα πρέπει να

απολαμβάνει υψηλό βαθμό αποδοχής (π.χ. από τα ενδιαφερόμενα στελέχη και άλλες ομάδες ατόμων) και πραγματοποίησης του στόχου. Στα πλαίσια της ορθολογικής απόφασης γίνονται οι παρακάτω παραδοχές:

- Το στέλεχος που επιφορτίζεται με την λήψη της απόφασης ξέρει επακριβώς το στόχο της
- απόφασης και πώς αυτός θα μετρηθεί.
- Είναι πλήρως γνωστές όλες οι εναλλακτικές λύσεις και οι ακριβείς συσχετίσεις τους με
- τον/τους στόχους που επιδιώκονται.
- Είναι επιθυμητή η όποια λύση μεγιστοποιεί το βαθμό ικανοποίησης του/των στόχων.

Τα περισσότερα μοντέλα επιχειρησιακής έρευνας έχουν αναπτυχθεί με βάση τις παραπάνω παραδοχές. Στην πραγματικότητα όμως, οι αποφάσεις λαμβάνονται με βάση κάποιο βαθμό ικανοποίησης του επιλεγμένου στόχου που επιδιώκεται. Ο Simon υποστήριξε ότι είναι αδύνατον να ληφθούν πλήρως ορθολογικές αποφάσεις, δεδομένου ότι ο ανθρώπινος ορθολογισμός είναι περιορισμένος (bounded rationality), που οφείλεται στην έλλειψη ικανότητας απόκτησης όλων των πληροφοριών που χρειάζονται, λόγω αβεβαιότητας αλλά και λόγω των οικονομικών, τεχνολογικών, πολιτικών και χρονικών πιέσεων. Η επίλυση ενός προβλήματος ολοκληρώνεται με την επιτυχημένη εφαρμογή μιας αποδεκτής λύσης. Το ποια λύση είναι αποδεκτή έχει ήδη προσδιοριστεί στη φάση της επιλογής. Η Implementation Phase είναι αρκετά δύσκολη λόγω κυρίως παραγόντων όπως η αντίδραση σε οποιαδήποτε αλλαγή (resistance to change), εξασφάλιση της υποστήριξης της ανώτατης διοίκησης, η ενημέρωση και εκπαίδευση του προσωπικού, κλπ.

6.2 Ιεραρχική Ανάλυση Αποφάσεων (Analytic Hierarchy Process-AHP)

Η μέθοδος AHP είναι μία μέθοδος πολυκριτηριακής ανάλυσης. Είναι δηλαδή μία μέθοδος η οποία θεωρεί ένα αριθμό κριτηρίων με βάση τα οποία λαμβάνεται μία απόφαση. Είναι μία τεχνική που έχει χρησιμοποιηθεί σε πολλές περιπτώσεις, όπως επιλογή ιστοχώρου για διαφήμιση (Ngai, 2003), στην επιλογή προμηθευτών (Mani, et al., 2014) κ.ά. Οι εργασίες των (Subramanian & Ramanathan, 2012; Vaidya & Kumar, 2006) δίνουν μία εκτενή περιγραφή των εφαρμογών της AHP. Η μέθοδος AHP αναπτύχθηκε από τον Thomas Saaty στη δεκαετία του 1970 (Saaty, 1977) και αποτελεί μια δομημένη τεχνική για την οργάνωση και την ανάλυση πολύπλοκων και αδόμητων αποφάσεων, με βάση τα μαθηματικά και την ανθρώπινη κρίση - ψυχολογία. Η μέθοδος αποτελείται από τα εξής βήματα:

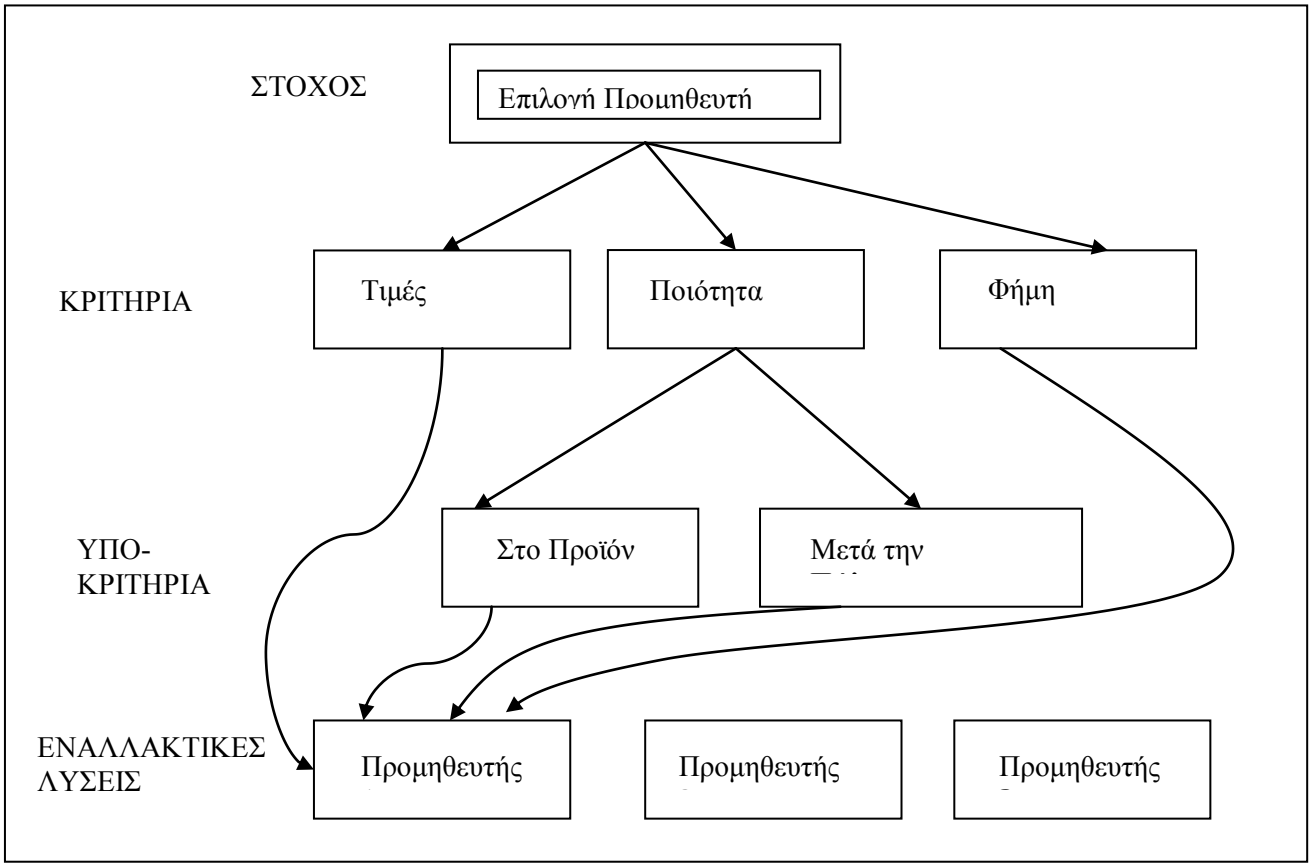
1. Καθορισμός στόχου.
2. Προσδιορισμός των κριτηρίων και των υπο-κριτηρίων με βάση τα οποία γίνεται η αξιολόγηση και η λήψη της απόφασης. Προσδιορισμός των εναλλακτικών λύσεων που θα εξετασθούν σχετικά με την απόφαση.
3. Οργάνωση του στόχου, των κριτηρίων και των εναλλακτικών σε μία ιεραρχία, στην κορυφή της οποίας είναι ο στόχος και στη συνέχεια στα επόμενα επίπεδα βρίσκονται τα κριτήρια, τα υπο-κριτήρια και οι εναλλακτικές λύσεις.
4. Συλλογή δεδομένων.
5. Εφαρμογή του μαθηματικού μοντέλου της AHP.
6. Έλεγχος αξιοπιστίας μοντέλου και αποτελεσμάτων.
7. Εξαγωγή συμπερασμάτων.

Βήμα 1. Καθορισμός στόχου. Ο καθορισμός του στόχου αναφέρεται στη διατύπωση της απόφασης η οποία επιδιώκεται να απαντηθεί. Παραδείγματα αποτελούν οι στόχοι να αποφασιστεί «ποιος είναι ο προμηθευτής που θα πρέπει να επιλεγεί;», «ποια η καλύτερη επενδυτική πρόταση;», «ποιο αυτοκίνητο να αγοράσουμε;», κλπ. Ο προσδιορισμός και η διατύπωση του στόχου θα πρέπει, ιδανικά, να εκφράζει όλους όσους συμμετέχουν στη λήψη της απόφασης. Θα μπορούσε να θεωρηθεί ότι ο προσδιορισμός του στόχου εντάσσεται στην Intelligence Phase, του μοντέλου λήψης αποφάσεων του Simon στον πίνακα 1.1.

Βήμα 2. Προσδιορισμός των κριτηρίων και των υπο-κριτηρίων με βάση τα οποία γίνεται η αξιολόγηση και η λήψη της απόφασης. Προσδιορισμός των εναλλακτικών λύσεων που θα εξετασθούν σχετικά με την απόφαση. Για την επίτευξη του στόχου και την αξιολόγηση μιας κατάστασης είναι απαραίτητο τα ορισθούν κάποια κριτήρια και υπο-κριτήρια. Για παράδειγμα, η επιλογή ενός προμηθευτή μπορεί να αξιολογηθεί με βάση τις *τιμές* που προσφέρει, την *ποιότητα* αυτών που προσφέρει, την *αξιοπιστία* του αλλά

και τη *φήμη* του. Όπως γίνεται φανερό, τα κριτήρια μπορεί να είναι ποσοτικά όπως οι τιμές, αλλά και ποιοτικά όπως η ποιότητα, αλλά ακόμα και υποκειμενικά όπως η φήμη. Είναι χαρακτηριστικό πλεονέκτημα της AHP η δυνατότητα ενσωμάτωσης στο μοντέλο και ποσοτικών, αλλά και ποιοτικών κριτηρίων. Οι εναλλακτικές λύσεις προσδιορίζονται για παράδειγμα: «Προμηθευτής 1», «Προμηθευτής 2», «Προμηθευτής 3», κλπ.

Βήμα 3. Οργάνωση του στόχου, των κριτηρίων και των εναλλακτικών σε μία ιεραρχία, στην κορυφή της οποίας είναι ο στόχος και στη συνέχεια στα επόμενα επίπεδα βρίσκονται τα κριτήρια, τα υπο-κριτήρια και οι εναλλακτικές λύσεις. Η εφαρμογή της μεθόδου συνεχίζεται με την διαμόρφωση του στόχου, των κριτηρίων και των υπο-κριτηρίων σε μία ιεραρχία όπως η παρακάτω:



Πίνακας 6.2 *Η Ιεραρχία Στόχου και Κριτηρίων στην AHP.*

1. *Συλλογή δεδομένων*.

Τα δεδομένα συλλέγονται με τη διαμόρφωση του ερωτηματολογίου της AHP. Η μορφή του είναι ιδιαίτερη και παράδειγμα παρατίθεται πιο κάτω. Οι επιλογές γίνονται μετά από κάθε σύγκριση των κριτηρίων σε ζεύγη (pairwise comparison).

Σε σχέση με τον στόχο της επιλογής Προμηθευτή, ποιο κριτήριο θεωρείτε πιο σημαντικό και πόσο;

Τιμές	9	7	5	3	1	3	5	7	9	**Φήμη**
		V								

Πίνακας 6.3 *Παράδειγμα ερωτηματολογίου AHP.*

Οι δυνατές επιλογές κωδικοποιούνται με βάση τους αριθμούς 1,3,5,7,9 αλλά και τους 2,4,6,8, ως ενδιάμεσες απαντήσεις, ως εξής:

Βαθμός προτίμησης κριτηρίου (και εναλλακτικής λύσης)	**Ορισμός/Σημασία**
1	Τα δύο κριτήρια είναι *ίσης* σημασίας/προτίμησης
3	Σχετικά πιο σημαντικό
5	Έντονα πιο σημαντικό
7	Πολύ έντονα πιο σημαντικό
9	Εξαιρετικά πιο σημαντικό
2,4,6,8	Ενδιάμεσα επίπεδα προτίμησης

Πίνακας 6.4 Κλίμακα προτιμήσεων στην AHP.

Πρώτα ο ερωτώμενος απαντά σχετικά με το ποιο κριτήριο θεωρεί προτιμότερο και μετακινείται προς την κατεύθυνση του κριτηρίου αυτού στο ερωτηματολόγιο. Δηλαδή, εάν είναι προτιμότερο (ή πιο σημαντικό) το κριτήριο «τιμές» μετακινείται στην επιλογή προς τις «τιμές». Στη συνέχεια, επιλέγει το βαθμό προτίμησης. Στον πίνακα 1.3 η απάντηση είναι: Το κριτήριο «τιμές» είναι «πολύ έντονα ποιο σημαντικό σε σχέση με τη φήμη». Η μέθοδος AHP δεν έχει ιδιαίτερες απαιτήσεις ως προς το μέγεθος του δείγματος. Δεν είναι μια μέθοδος στατιστικής.

2. *Εφαρμογή του μαθηματικού μοντέλου της AHP.*

Η εφαρμογή του μαθηματικού μοντέλου της AHP αποσκοπεί στο να προσδιοριστούν οι συντελεστές βαρύτητας των κριτηρίων, των υπο-κριτηρίων με βάση τα οποία λαμβάνεται μια απόφαση, αλλά και της καταλληλότητας των εναλλακτικών λύσεων. Έστω ο πίνακας (A) με τις ανά ζεύγη κριτηρίων συγκρίσεις:

$$A=\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{matrix}$$

Τα στοιχεία a_{ij} του πίνακα φανερώνουν τη σχετική σημαντικότητα (προτίμηση) του κριτηρίου (i) έναντι του κριτηρίου (j). Στην περίπτωση που το δείγμα αφορά περισσότερα από ένα ερωτηματολόγια, τότε τα στοιχεία υπολογίζονται χρησιμοποιώντας το μέσο αριθμητικό των απαντήσεων από κάθε ερωτηματολόγιο ή (όπως προτείνεται στη βιβλιογραφία), υπολογίζεται ο γεωμετρικός μέσος m_{ij} των απαντήσεων με τον τύπο που ακολουθεί: $m_{i,j} = \sqrt{\prod_{i=1}^{n} a_{i,j}}$. Για κάθε στοιχείο του πίνακα συγκρίσεων A, ισχύουν οι παρακάτω ιδιότητες:

a_{ij} = 1, αφού γίνεται σύγκριση με το ίδιο στοιχείο.

a_{ij} > 1, όταν το στοιχείο (i) είναι σημαντικότερο από το στοιχείο (j), για την επίτευξη του στόχου.

a_{ij} < 1, όταν το στοιχείο (j) είναι σημαντικότερο από το στοιχείο (i), για την επίτευξη του στόχου.

$a_{ji} = \frac{1}{a_{ij}}$, δηλαδή ο πίνακας είναι κάτω από τη διαγώνιο έχει τις αντίστροφες τιμές.

Αν τα στοιχεία που συγκρίνονται ανά ζεύγη είναι (n), τότε απαιτούνται (n-1)/2, συγκρίσεις για τα κριτήρια στον πίνακα A (Saaty, 1980). Οι σημαντικότητες των κριτηρίων (w) υπολογίζονται από την σχέση:

$Aw = \lambda_{\max} w$, όπου (A) ο με τις ανά ζεύγη κριτηρίων συγκρίσεις, και το $\lambda_{\max}$, είναι η μέγιστη ιδιοτιμή (eigenvalue) του πίνακα (A). Πιο συγκεκριμένα:

a) Υπολογίζεται το άθροισμα της κάθε γραμμής του πίνακα A $s_i = \sum_j a_{ij}$.

b) Για κάθε γραμμή του πίνακα A υπολογίζονται οι σημαντικότητες ως το παρακάτω πηλίκο: $w_i = \frac{s_i}{\sum_j a_{ij}}$, με τα βάρη w_i που προκύπτουν να είναι κανονικοποιημένα δηλαδή: $\sum_i w_i = 1.$

c) Στη συνέχεια υψώνεται υπολογίζεται το γινόμενο $A^1 = A * A$, και υπολογίζονται πάλι τα βάρη $w^1{}_i$. Συγκρίνονται τα βάρη $w^1{}_i$ και w_i. Εάν διαφέρουν τότε επαναλαμβάνεται η διαδικασία από το βήμα (a). Εάν δεν διαφέρουν σημαντικά τότε τα τελευταία βάρη που υπολογίσθηκαν αποτελούν τις τελικές προτιμήσεις των κριτηρίων.

3. *Έλεγχος αξιοπιστίας μοντέλου και αποτελεσμάτων*.

Η *αξιοπιστία* ή η *συνέπεια* της μεθόδου εξαρτάται από τη συνέπεια του πίνακα που προκύπτει από την ανά ζεύγη σύγκριση των κριτηρίων (consistency of the pairwise matrix), με την οποία συνέπεια εννοείται ότι, όταν γνωρίζουμε ένα βασικό αριθμό των στοιχείων μιας σειράς του πίνακα, τα υπόλοιπα στοιχεία μπορούν να εξαχθούν λογικά από αυτό. Το παραγόμενο μοντέλο αξιολογείται ως προς την αξιοπιστία με βάση του δείκτη αξιοπιστίας ο οποίος υπολογίζεται ως εξής:

$CI = \frac{\lambda_{\max} - n}{n-1}$, όπου CI (Consistency Index), το (n) είναι η διάσταση του πίνακα συγκρίσεων (A), δηλαδή είναι το πλήθος των προς θεώρηση κριτηρίων. Το $\lambda_{\max}$, είναι η μέγιστη ιδιοτιμή (eigenvalue) του πίνακα (A).

$CR = \frac{CI}{RI} * 100$, όπου το CR (Consistency Ratio) είναι ο δείκτης συνέπειας, το RI είναι ο τυχαίος δείκτης του οποίου οι τιμές έχουν υπολογισθεί για διαφορετικά μεγέθη δειγμάτων και τις βρίσκουμε στον πιο κάτω πίνακα (Saaty, 1977).

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0,5	0,89	1,11	1,25	1,35	1,4	1,45	1,49	1,51	1,54	1,56	1,57	1,58

Όπου (n) είναι η διάσταση του πίνακα.

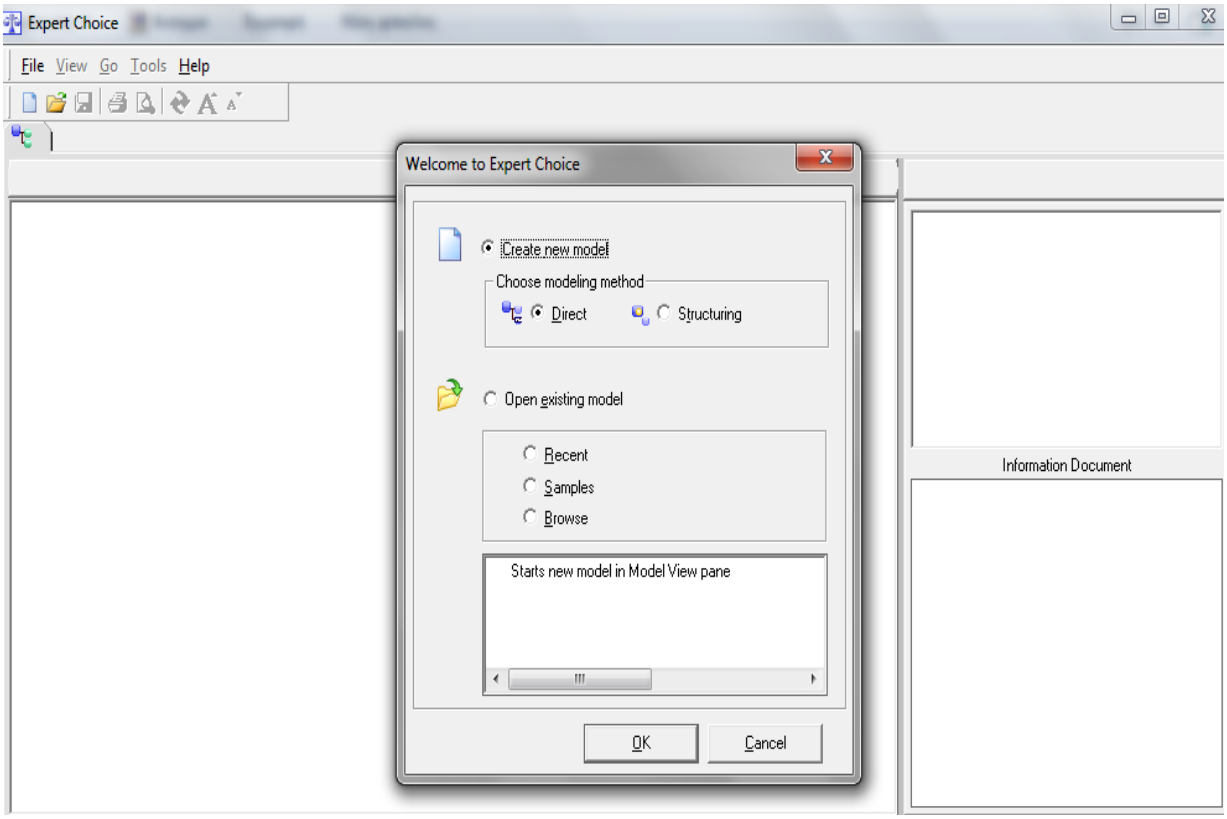
Τα αποτελέσματα είναι συνεπή και αξιόπιστα όταν $CR < 0.1$

4. *Εξαγωγή συμπερασμάτων*.

Τα συμπεράσματα αφορούν στη λήψη της απόφασης. Δηλαδή στον προσδιορισμό της βαρύτητας των κριτηρίων και της αξιολόγησης των προς εξέταση εναλλακτικών λύσεων.

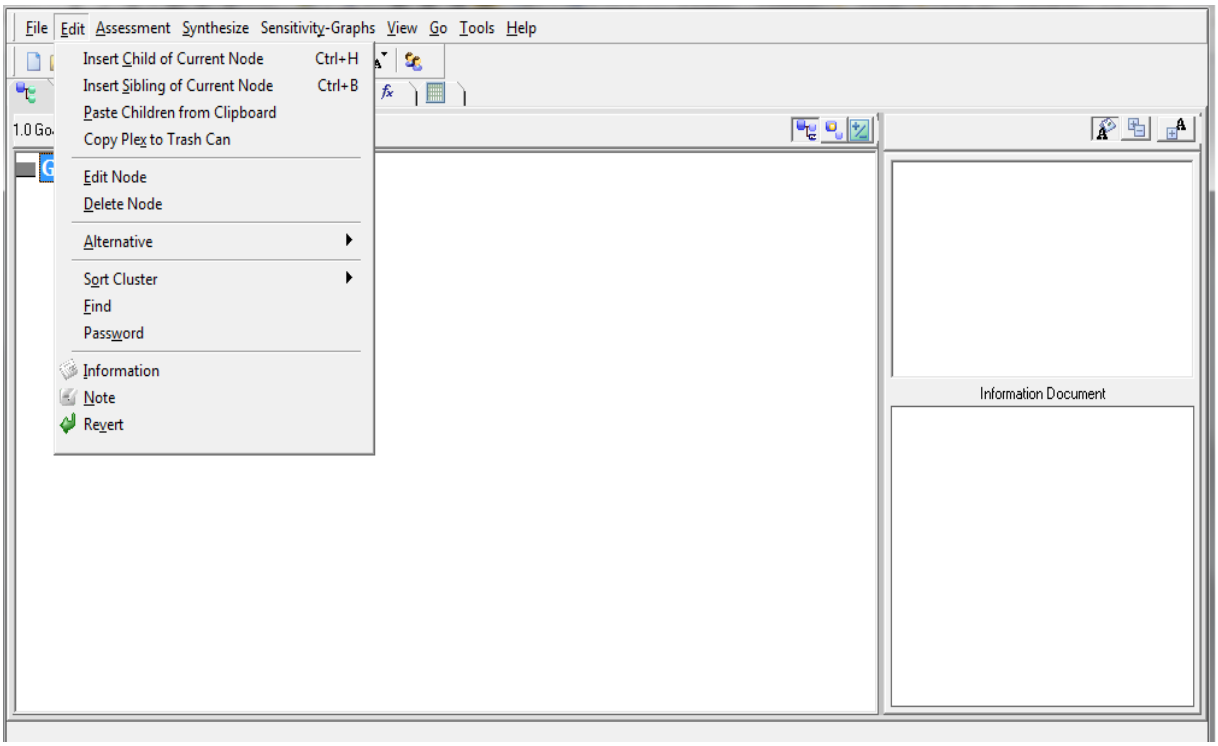
Η μέθοδος AHP μπορεί να υλοποιηθεί με τη χρήση του Excel αλλά υπάρχουν και λογισμικά τα οποία έχουν ήδη αναπτυχθεί όπως το PriEsT (Priority Estimation Tool (AHP)). Για περισσότερες πληροφορίες, δείτε τη διεύθυνση http://sourceforge.net/projects/priority/, όπως επίσης και στη διεύθυνση http://bpmsg.com/ahp-online-calculator/, όπου υπάρχει online υπολογιστικό σύστημα της μεθόδου. Ένα άλλο εξαιρετικά δημοφιλές σύστημα είναι το Expert Choice, http://expertchoice.com/, του οποίου παράδειγμα ακολουθεί.

Η έναρξη του λογισμικού expert choice φαίνεται στην πιο κάτω εικόνα:



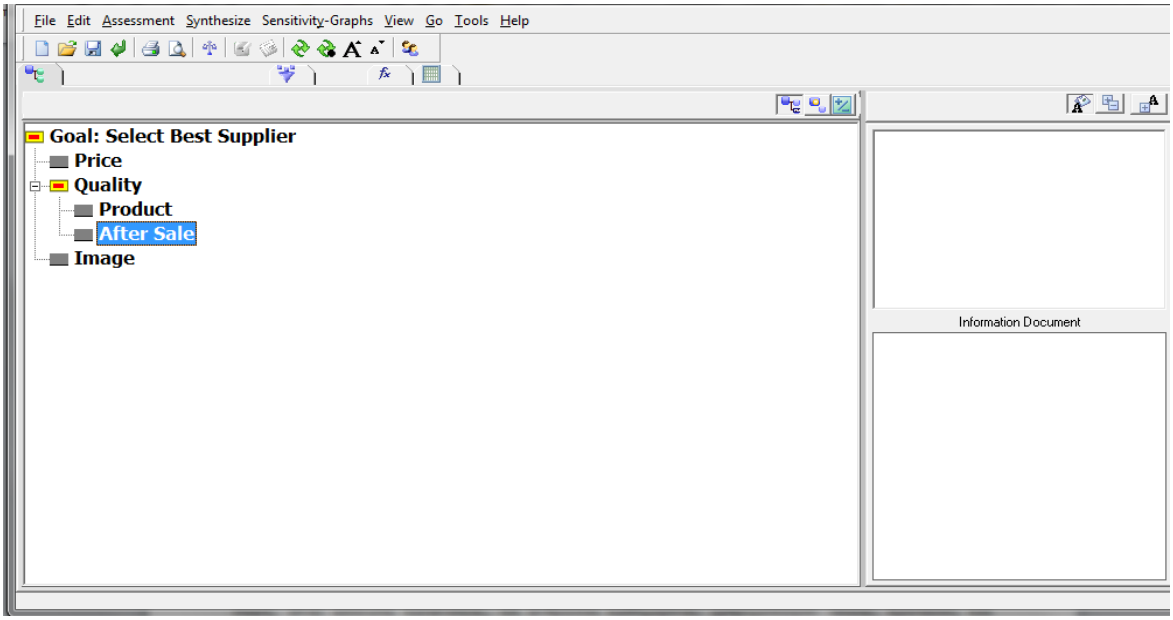
Εικόνα 6.1. *Η έναρξη του expert choice ζητά την δημιουργία νέου μοντέλου*

Την επιλογή του στόχου ως «Select Best Supplier» ακολουθεί ο προσδιορισμός των κριτηρίων (insert child), όπως φαίνεται στην παρακάτω εικόνα.



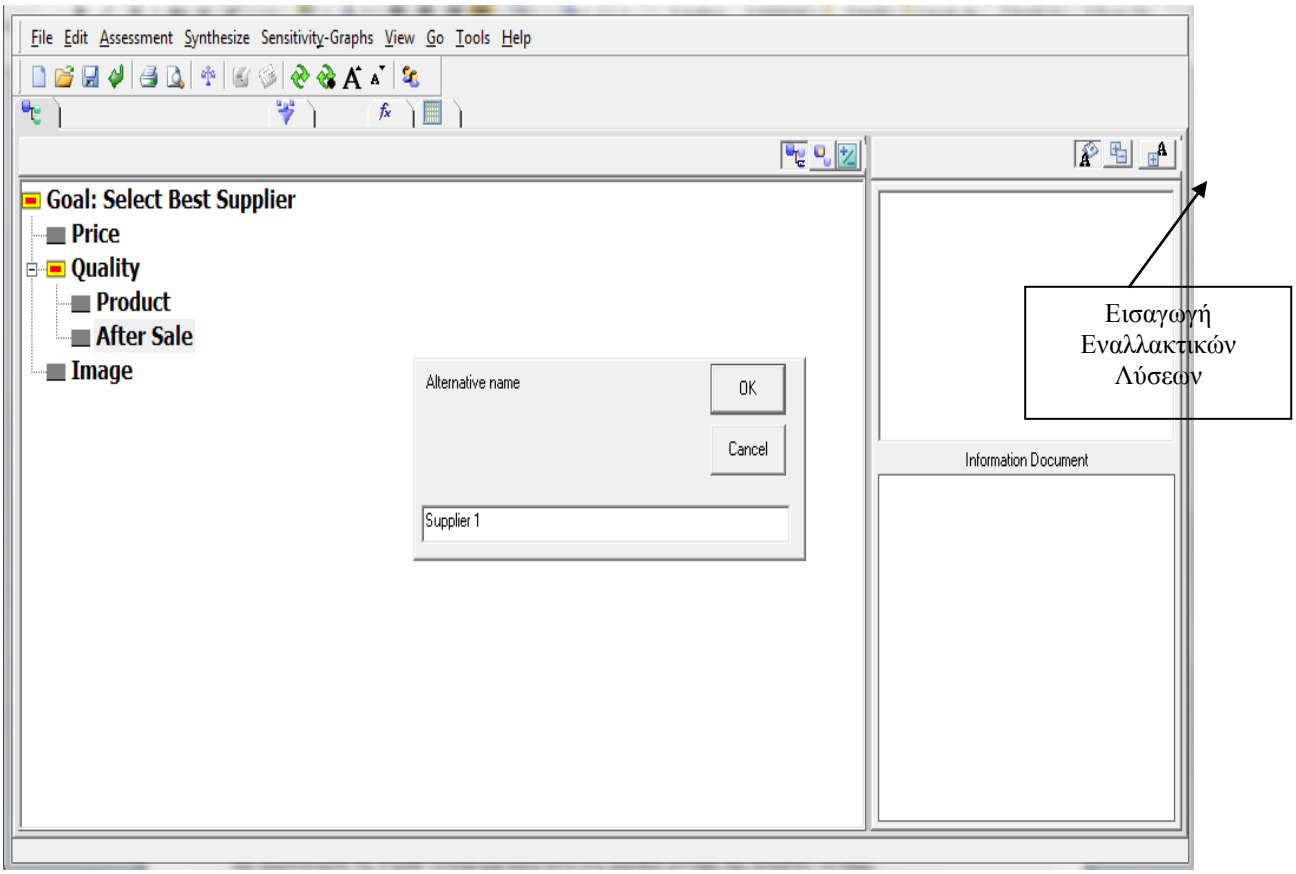
Εικόνα 6.2*. Η εισαγωγή κριτηρίων*

Η ιεραρχία συμπληρώνεται με τα κριτήρια και τα υπο-κριτήρια. Τα υπο-κριτήρια είναι στη ουσία κριτήρια για την επίτευξη του στόχου, που είναι το αντίστοιχο κριτήριο στο πιο πάνω επίπεδο.



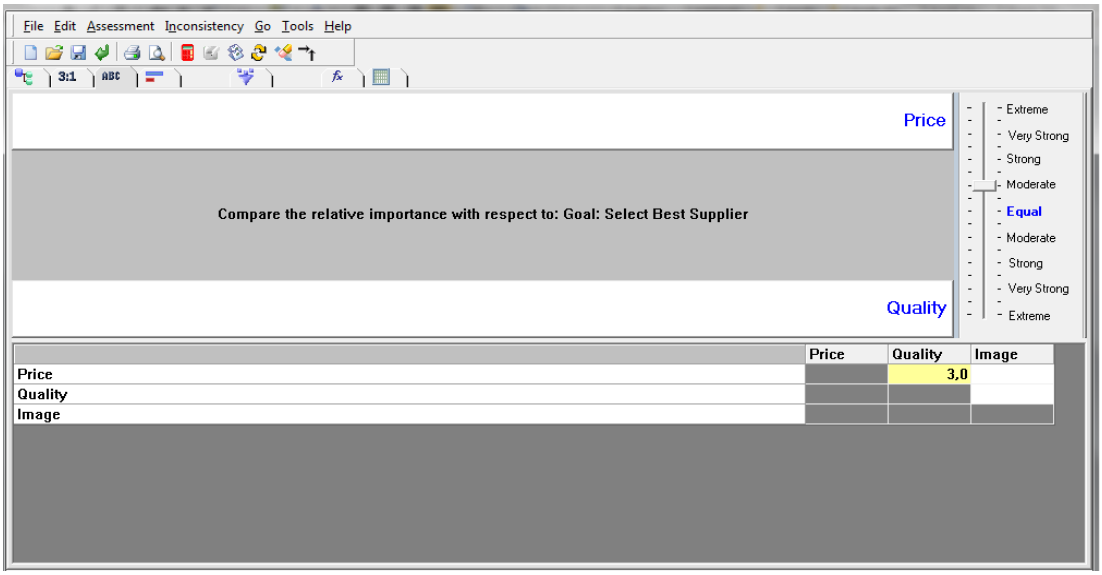
Εικόνα 6.3. *Η εισαγωγή κριτηρίων και υπο-κριτηρίων*

Η εισαγωγή των εναλλακτικών λύσεων ακολουθεί.



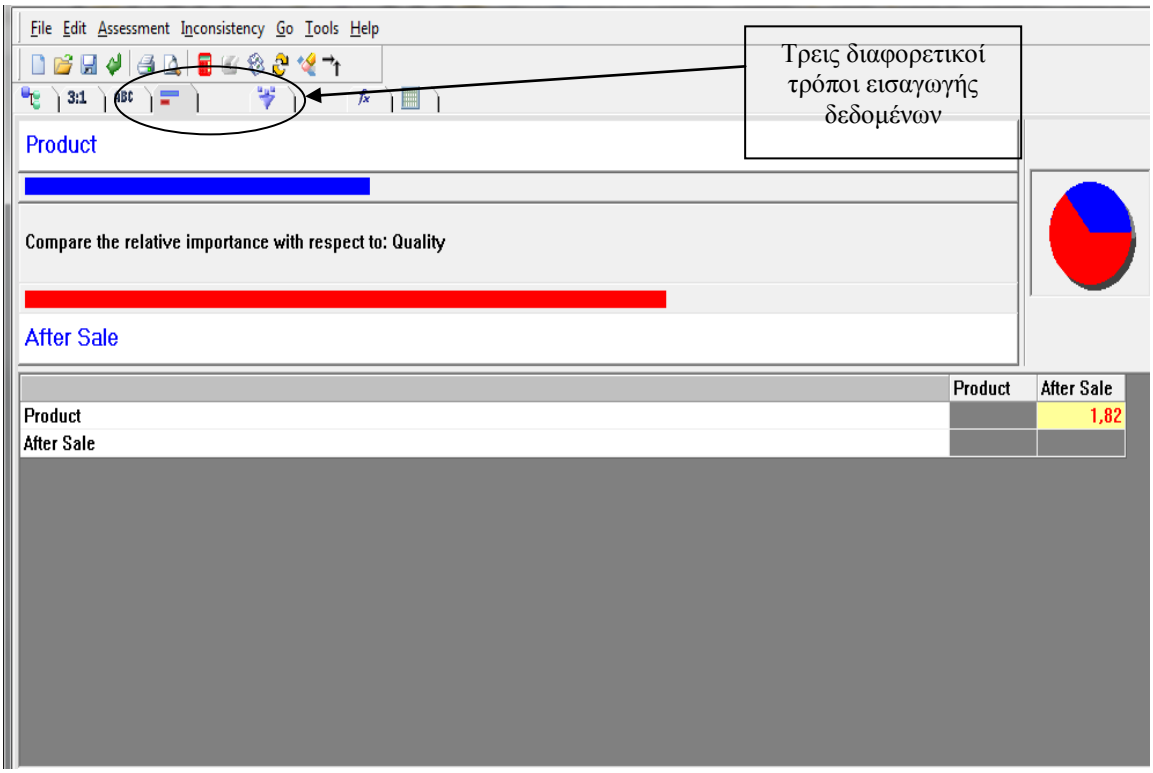
Εικόνα 6.4. Η εισαγωγή δύο ή περισσότερων εναλλακτικών λύσεων

Η εισαγωγή των δεδομένων γίνεται επιλέγοντας την επιλογή «Assessment»>>> «Pairwise».



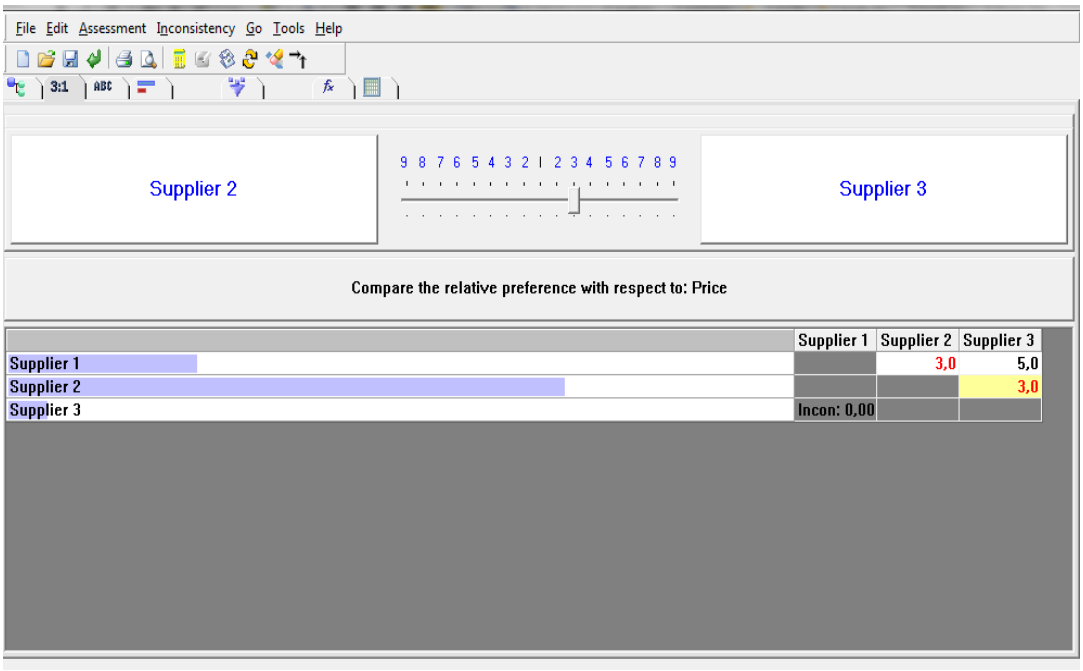
Εικόνα 6.5. *Η εισαγωγή δεδομένων με την ανά ζεύγη σύγκριση κριτηρίων*

Ακολουθεί η εισαγωγή δεδομένων σχετικά με τα υπο-κριτήρια.



Εικόνα 6.6. *Η εισαγωγή δεδομένων με την ανά ζεύγη σύγκριση υπο-κριτηρίων*

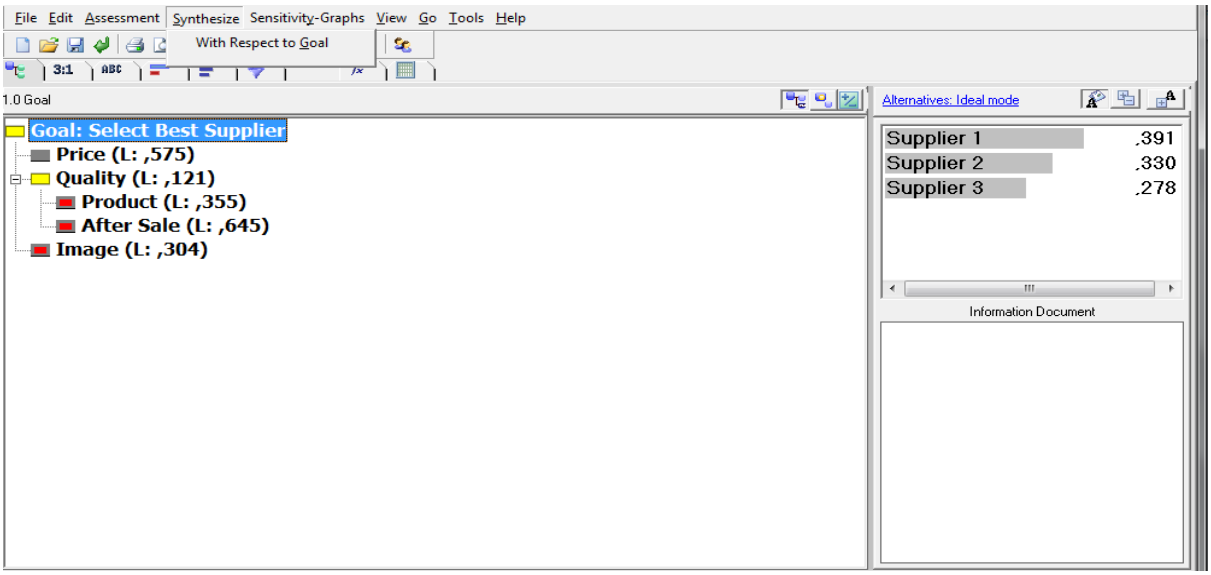
Στην πιο πάνω εικόνα φαίνεται ένας διαφορετικός τρόπος εισαγωγής δεδομένων, από τους συνολικά τρεις που προσφέρει το expert choice.



Εικόνα 6.7. *Η εισαγωγή δεδομένων σχετικά με τις προτιμήσεις των εναλλακτικών λύσεων ως προς τα κριτήρια*

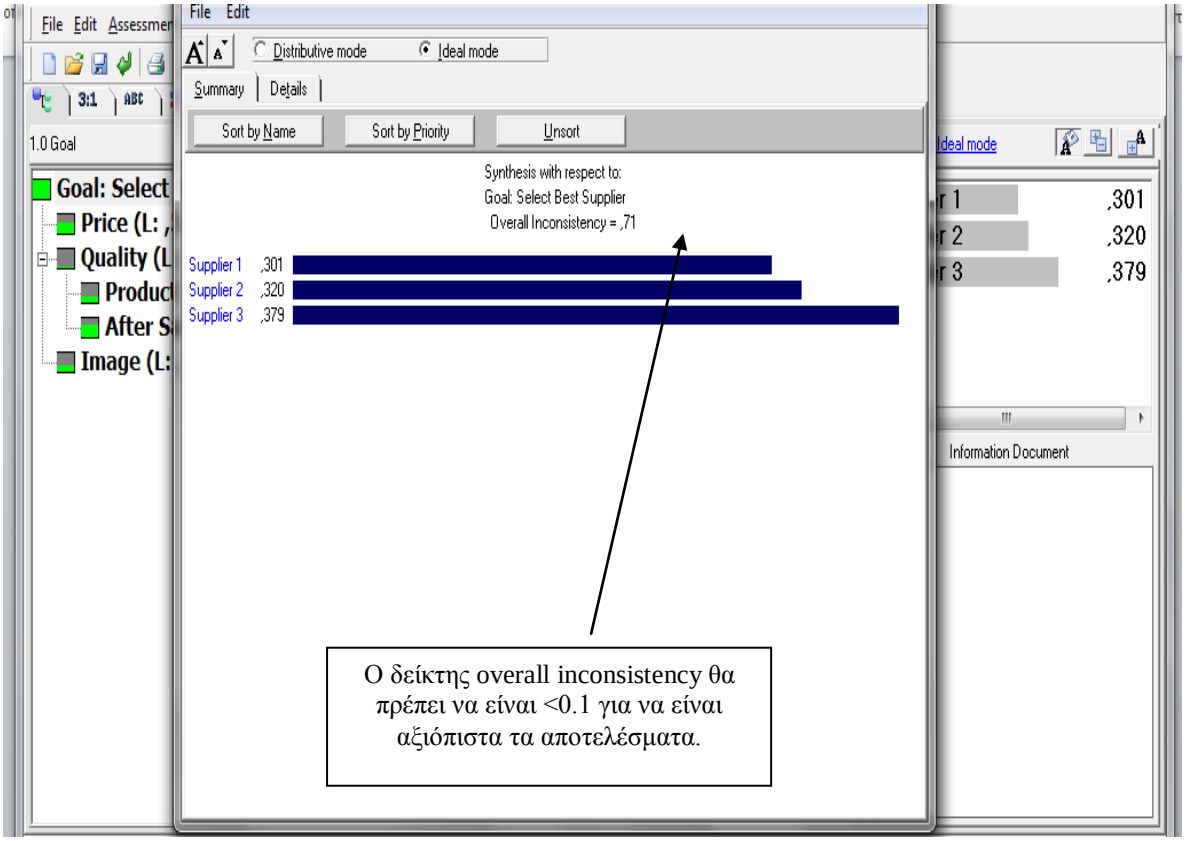
Ακολουθώντας την ίδια διαδικασία με αυτήν της αξιολόγησης των κριτηρίων, η Εικόνα 6.7 δείχνει τις απαντήσεις του ερωτώμενου σχετικά με το ποιος είναι ο προτεινόμενος προμηθευτής, ως προς το κριτήριο «price». Με την διατύπωση των κριτηρίων και των υπο-κριτηρίων, των εναλλακτικών λύσεων και των δεδομένων σχετικά με τις προτιμήσεις, ολοκληρώνεται η ιεραρχία AHP.

Για την εξαγωγή των αποτελεσμάτων ακολουθούμε τις επιλογές όπως φαίνονται στην επόμενη εικόνα.



Εικόνα 6.8. *Η εισαγωγή δεδομένων σχετικά με τις προτιμήσεις των εναλλακτικών λύσεων ως προς τα κριτήρια*

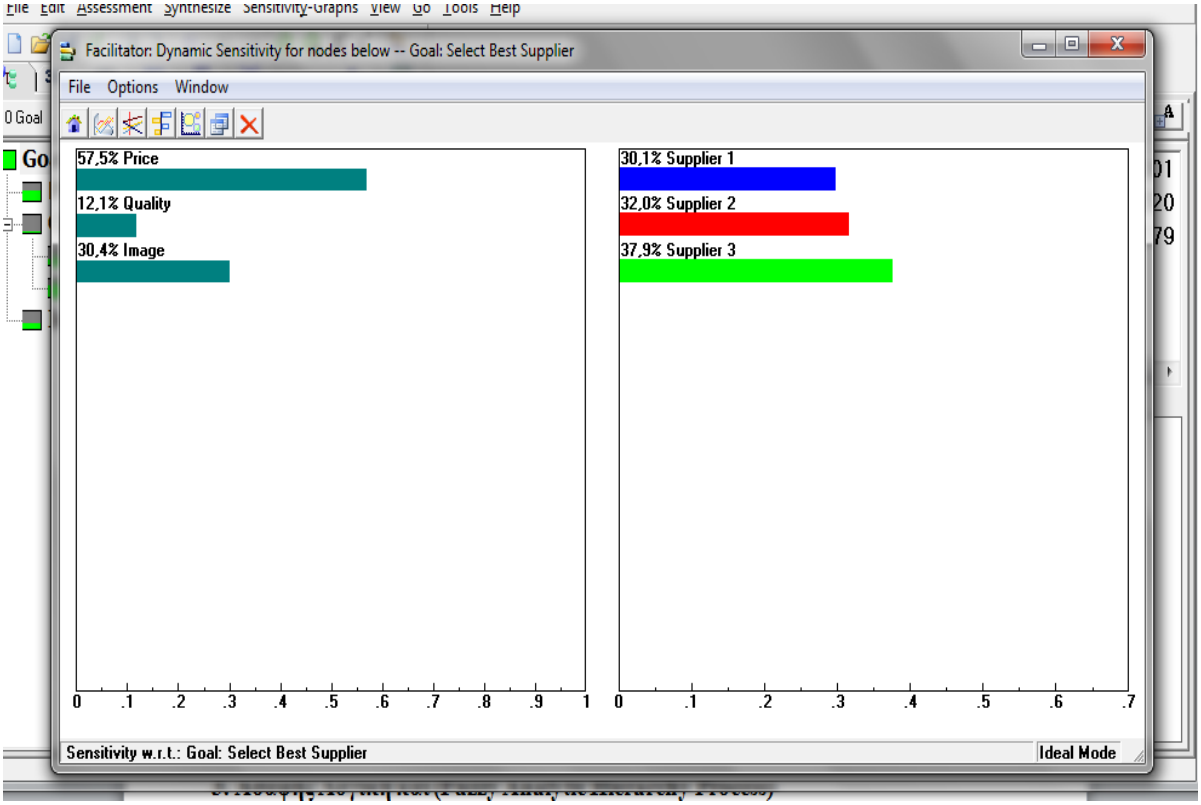
Τα αποτελέσματα φαίνονται στην παρακάτω Εικόνα 6.9.



Εικόνα 6.9. *Η εισαγωγή δεδομένων σχετικά με τις προτιμήσεις των εναλλακτικών λύσεων ως προς τα κριτήρια*

Τα αποτελέσματα δηλώνουν ότι ο προτιμότερος προμηθευτής είναι ο «supplier3» με αξιολόγηση 0.379 και ακολουθούν οι «supplier2» με αξιολόγηση 0.320 και «supplier1» με βαθμό 0.301. Όπως φαίνεται στην Εικόνα 6.9, ο δείκτης overall inconsistency είναι >0.1 που σημαίνει ότι τα αποτελέσματα **δεν** είναι αξιόπιστα. Θα πρέπει να επανεξετασθούν οι προτιμήσεις που δηλώθηκαν και συμπληρώθηκαν στα ερωτηματολόγια και να επαναληφτεί η ανάλυση.

Το λογισμικό expert choice παρέχει και δυνατότητες ανάλυσης ευαισθησίας.



Εικόνα 6.10. *Η ανάλυση ευαισθησίας, για την μελέτη των επιπτώσεων στις λύσεις από τις μεταβολές των βαρών των κριτηρίων*

Η ανάλυση ευαισθησίας χρησιμεύει στην ανάλυση των επιπτώσεων στις λύσεις από μεταβολές στις προτεραιότητες των κριτηρίων, αλλά και την ανάλυση του κατά πόσο αλλάζουν οι επιλογές των λύσεων, εξετάζοντας την σταθερότητα των επιλογών.

6.3 Ασαφής Λογική και Ιεραρχική Ανάλυση Αποφάσεων (Fuzzy Analytic Hierarchy Process)

Η μέθοδος FAHP αποτελεί μια έκδοση της AHP η οποία χρησιμοποιεί έννοιες από την ασαφή λογική. Πριν την παρουσίαση της FAHP γίνεται μια εισαγωγή στις βασικές έννοιες της ασαφούς λογικής.

6.3.1 Εισαγωγή στις έννοιες της Ασαφούς Λογικής (Fuzzy Logic)

Στην καθημερινή μας ζωή χρησιμοποιούμε ασαφείς εκφράσεις, όπως «το ευρώ γίνεται ιδιαίτερα δυνατό», «ο καιρός είναι γλυκός». Οι εκφράσεις αυτές, παρά το γεγονός ότι δεν είναι ποσοτικοποιημένες και σαφείς, είναι απόλυτα κατανοητές. Οι άνθρωποι καθημερινώς χρησιμοποιούν τέτοιες προτάσεις και επικοινωνούν χωρίς πρόβλημα. Με αυτόν τον τρόπο δηλαδή, εκφράζουμε έναν άλλον τρόπο σκέψης του ανθρώπου. Συμπερασματικά, μπορούμε να πούμε ότι η κλασσική προτασιακή λογική που δέχεται μόνο δύο τιμές (ναι ή όχι) για μια κατάσταση δεν είναι αρκετή για να εκφραστούμε πλήρως. Δηλαδή, ο καιρός δεν είναι ή βροχερός ή ξηρός. Υπάρχει ανάγκη διατύπωσης και άλλων καταστάσεων (καταρρακτώδης, υγρός, σχετικά υγρός, με ξαφνικές καταιγίδες, με ψιχάλισμα, κλπ.). Υπάρχουν αρκετοί χαρακτηρισμοί για να αποδώσουν τις καιρικές συνθήκες. Η ανάγκη για την ανάπτυξη μιας μαθηματικής λογικής που θα λαμβάνει υπ' όψιν περισσότερες τιμές από τις κλασσικές δύο είχε διατυπωθεί από την αρχαιότητα στην Ελλάδα από τον Ηράκλειτο και τον Αριστοτέλη. Πέρασαν όμως πολλά χρόνια έως ότου διατυπωθεί επαρκώς η Ασαφής Λογική το 1965 από τον Lofti Zadeh του πανεπιστημίου της California, Berkeley.

Η Ασαφής Λογική (Ross, 2009) δίνει τη δυνατότητα αναπαράστασης και μελέτης της ασάφειας και της αβεβαιότητας σε πολλά φαινόμενα και καταστάσεις. Διαφέρει από τη Θεωρία των Πιθανοτήτων, γιατί η τελευταία μελετά τη συχνότητα εμφάνισης γνωστών και αναμενόμενων αποτελεσμάτων. Για παράδειγμα,

όταν ρίχνουμε ένα ζάρι ξέρουμε ποια είναι τα πιθανά αποτελέσματα. Αυτά και οι πιθανότητες εμφάνισής τους μπορούν να ορισθούν αντικειμενικά. Όμως όταν ένας άνθρωπος 30 ετών χαρακτηρίζεται ως «νέος» σε μια κρουαζιέρα, δεν χαρακτηρίζεται το ίδιο σε ένα σχολείο. Δηλαδή, ο χαρακτηρισμός «νέος» ορίζεται διαφορετικά σε διαφορετικά περιβάλλοντα αλλά και από διαφορετικούς ανθρώπους. Αυτή η διαφορετικότητα μπορεί να αποδοθεί από την Ασαφή Λογική με μαθηματικό τρόπο.

Οι εφαρμογές αρχικά περιορίζονταν στον αυτόματο έλεγχο συστημάτων, αλλά αργότερα επεκτάθηκαν στην επεξεργασία γυαλιού, σε κινηματογραφικές μηχανές, πλυντήρια, αυτόματα κιβώτια μετάδοσης αυτοκινήτων, κ.ά. Στα παραδείγματα εφαρμογής της Ασαφούς Λογικής μπορούν να αναφερθούν ο υπόγειος σιδηρόδρομος της πόλης Σεντάι της Ιαπωνίας. Η επιτάχυνση ή η επιβράδυνση των συρμών ελέγχεται με τέτοιο τρόπο ώστε να είναι όσο το δυνατόν πιο ομαλή, με αποτέλεσμα οι επιβάτες να μη χρειάζεται να κρατιούνται από τις χειρολαβές. Επίσης, η εταιρεία Mitsubishi Heavy Industries ενσωμάτωσε Ασαφή Λογική στα κλιματιστικά της, με αποτέλεσμα να επιτύχει μείωση της κατανάλωσης κατά 20%.

Έχουν επίσης αναπτυχθεί αρκετά Συστήματα Στήριξης Απόφασης βασισμένα σε Ασαφή Λογική, όπως στους χώρους της ιατρικής διάγνωσης, της ανάλυσης επενδύσεων, του χρονοπρογραμματισμού λεωφορείων, της ανάπτυξης επιχειρηματικής στρατηγικής, της διοίκησης πληροφοριακών συστημάτων κ.ά. Αν και οι εφαρμογές της Ασαφούς Λογικής σε βιομηχανικές κατασκευές είναι αρκετά επιτυχημένες, οι εφαρμογές της σε θέματα διοίκησης επιχειρήσεων και λήψης αποφάσεων έπονται. Χαρακτηριστικό σ' αυτήν την περιοχή είναι το σύστημα που έχει αναπτυχθεί σε εταιρεία των ΗΠΑ για την αξιολόγηση σεναρίων εξαγορών επιχειρήσεων. Τα τελευταία χρόνια, η έρευνα και ανάπτυξη τέτοιων συστημάτων επεκτείνεται γοργά και σε καθαρά επιχειρηματικά θέματα.

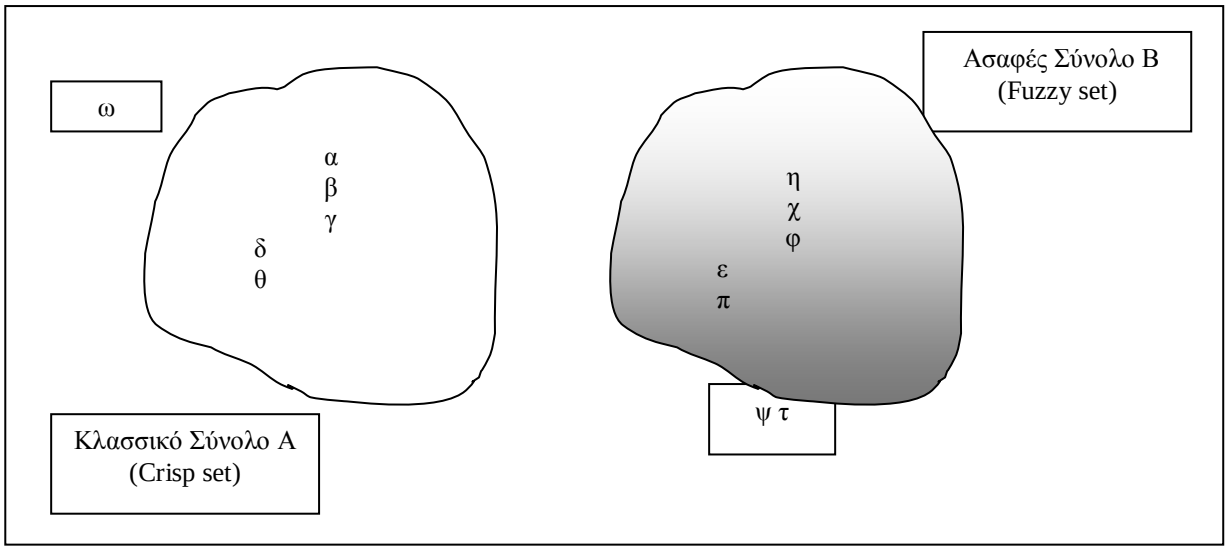
Η Ασαφής Λογική χρησιμοποιείται σε προβλήματα με τα παρακάτω χαρακτηριστικά:

- **Όπου δεν υπάρχει λύση ή είναι πολύ δύσκολη η επίλυση του προβλήματος.** Υπάρχουν προβλήματα για τα οποία η αλγεβρική λύση δεν υπάρχει ή είναι εξαιρετικά δύσκολο να προσδιοριστεί αναγκάζοντας υπολογιστικά συστήματα σε εξαιρετικά χρονοβόρους ατέρμονους υπολογισμούς.
- **Όπου δεν ενδιαφέρει η ακριβής λύση του προβλήματος.** Για παράδειγμα μας αρκεί η πληροφόρηση ότι βρέχει για να πάρουμε την ομπρέλα μας μαζί. Δε μας ενδιαφέρει ο ρυθμός με τον οποίο πέφτουν οι σταγόνες!!!
- **Μη-δομημένα προβλήματα**, όπως αυτά που συχνά αντιμετωπίζουν οι σύγχρονες επιχειρήσεις.

Βασική έννοια στην Ασαφή Λογική είναι το ασαφές σύνολο. Ένα σύνολο περιλαμβάνει μια σειρά από έννοιες, αντικείμενα, κλπ. Στην κλασσική προτασιακή λογική ένα σύνολο είναι σαφώς προσδιορισμένο, δηλαδή είναι γνωστό μετά βεβαιότητας ποια στοιχεία του συνόλου ανήκουν στο σύνολο και ποια στοιχεία δεν ανήκουν. Η χρήση της Ασαφούς Λογικής βοηθά στην ανάπτυξη κανόνων για την αντιμετώπιση προβλημάτων που είναι πιο κοντά στον τρόπο σκέψης των ανθρώπων από ότι η παραδοσιακή δομή «IF-THEN-ELSE» που χρησιμοποιείται στους Η/Υ. Έτσι, μπορούμε να αναπτύξουμε κανόνες όπως ο παρακάτω.

IF κάνει *πολύ κρύο* THEN παίρνουμε μαζί γάντια και παλτό.

Η Ασαφής Λογική είναι χρήσιμη γιατί μπορούμε να περιγράψουμε προβλήματα αλλά και τις λύσεις τους με τρόπο με τον οποίο σκέπτονται και οι άνθρωποι. Η έννοια του Ασαφούς Συνόλου είναι βασική για την περαιτέρω κατανόηση της Ασαφούς Λογικής. Στο παρακάτω σχεδιάγραμμα έχουμε δύο σύνολα. Το Κλασικό σύνολο Α και το Ασαφές Σύνολο Β.



Σχήμα 6.1 Κλασσικά και Ασαφή Σύνολα.

Η διαφορά των δύο συνόλων έγκειται στο γεγονός ότι τα όρια του *Κλασικού Συνόλου* είναι σαφώς ορισμένα. Με άλλα λόγια, γνωρίζουμε ακριβώς ποια στοιχεία ανήκουν σ' αυτό και ποια όχι. Δηλαδή στην ερώτηση "Ανήκει το στοιχείο «α» στο σύνολο Α;", η απάντηση είναι σαφώς «ναι». Στην ερώτηση "Ανήκει το στοιχείο «ω» στο σύνολο Α;", η απάντηση τώρα είναι σαφώς «όχι». Ένα στοιχείο ή ανήκει ή δεν ανήκει σ' ένα κλασικό σύνολο.

Στο *Ασαφές Σύνολο* Β όμως, ένα στοιχείο μπορεί να ανήκει στο σύνολο Β, μπορεί να μην ανήκει αλλά μπορεί να «ανήκει και να μην ανήκει», δηλαδή δεν είναι απόλυτα σαφής η απάντηση. Στο πιο πάνω σχήμα, τα στοιχεία «ε» και «φ» ανήκουν σαφώς στο Β. Επίσης σαφώς το στοιχείο «ω» δεν ανήκει στο Β. Υπάρχουν όμως και τα στοιχεία «ψ» και «τ» τα οποία βρίσκονται στο μεταίχμιο, στα όρια του συνόλου Β. Για αυτά τα στοιχεία δεν μπορούμε να πούμε ότι σαφώς ανήκουν ή δεν ανήκουν στο σύνολο Β. Πώς μπορούμε λοιπόν να αναπαραστήσουμε ένα Ασαφές Σύνολο; Με μία συνάρτηση που ονομάζεται *Συνάρτηση Συμμετοχής*, και η οποία δείχνει το βαθμό με τον οποίο ανήκει ένα στοιχείο σ' ένα ασαφές σύνολο. Εάν σαφώς ανήκει στο σύνολο, τότε ο βαθμός συμμετοχής αυτού του στοιχείου στο εν' λόγω ασαφές σύνολο είναι «1». Εάν σαφώς δεν ανήκει τότε ο βαθμός συμμετοχής του στοιχείου είναι «0». Οι ενδιάμεσοι βαθμοί, δηλαδή οι αριθμοί στο κλειστό διάστημα [0,1], δείχνουν το κατά πόσο ένα στοιχείο ανήκει σ' ένα σαφές σύνολο. Βασιζόμενοι στην παραπάνω εισαγωγή, ο ορισμός της έννοιας του ασαφούς συνόλου δίνεται ως εξής:

Ένα ασαφές υποσύνολο Α που ανήκει στο χώρο U χαρακτηρίζεται από μία συνάρτηση συμμετοχής $\mu_A: U \to [0,1]$, η οποία αντιστοιχεί κάθε στοιχείο υ του U με έναν αριθμό $\mu_A(x)$ στο κλειστό διάστημα [0,1] οποίος αναπαριστά το βαθμό με τον οποίο ανήκει το στοιχείο υ στο ασαφές σύνολο Α.

Για παράδειγμα, θεωρείστε τα παρακάτω στοιχεία και σύνολα.

Όνομα	**Ύψος (μέτρα)**
Άννα	1,60
Γιώργος	1,75
Χριστίνα	1,80
Παναγιώτα	1,77
Θανάσης	1,77
Μαρία	1,82
Βασίλης	2,01
Μανώλης	1,69

Πίνακας 6.5. *Δεδομένα χαρακτηρισμού ύψους ανθρώπων*

Έστω ότι θεωρούμε την παρακάτω σχέση:

Σχέση 1: Όταν το ύψος του ανθρώπου είναι μέχρι το 1,70 μέτρα τότε ο άνθρωπος θεωρείται "χαμηλού αναστήματος", όταν είναι μεταξύ του 1,70 και του 1,80 θεωρείται "μεσαίου αναστήματος" και όταν είναι μεγαλύτερο του 1,80 θεωρείται "υψηλού αναστήματος". Δηλαδή η σχέση αυτή καθορίζει μονοδιάστατα το ύψος των προσώπων στο παράδειγμα. Εάν δηλαδή κάποιος είναι μεσαίου ύψους αποκλείεται να είναι και υψηλού.

Άρα μπορούμε να διακρίνουμε το σύνολο των ανθρώπων μεσαίου αναστήματος ως εξής:

Μεσαίο Ανάστημα = {Γιώργος, Χριστίνα, Παναγιώτα, Θανάσης}.

Δηλαδή ο Μανώλης, ή η Μαρία δεν ανήκουν στο σύνολο αυτό για ένα ή δύο εκατοστά αντίστοιχα. Είναι όμως σημαντική αυτή η διαφορά ύψους; Μήπως θα μπορούσαν να αποκλειστούν από κάποια διαδικασία χωρίς ουσιαστικό λόγο; Αλλά εάν αλλάξουμε λίγο τα δεδομένα και υποθέσουμε ότι οι παραπάνω είναι υποψήφιοι παίκτες καλαθόσφαιρας, μήπως θα έπρεπε να αλλάξουμε τους χαρακτηρισμούς για το ποιος είναι ψηλός και ποιός όχι; Εάν πάλι αλλάξουμε τα ονόματα με Γιαπωνέζικα, μήπως θα έπρεπε πάλι να αλλάξουμε τους χαρακτηρισμούς; Σε τέτοιες περιπτώσεις, όπου οι έννοιες **έχουν διαφορετική σημασία ή ορισμό, ή προτεραιότητα ή αξιολόγηση**, ανάλογα με την άποψη διαφορετικών ανθρώπων ή διαφορετικών καταστάσεων, η ασαφής λογική μας δίνει λύση.

Σύμφωνα λοιπόν με τη σχέση 1, η συμμετοχή ενός προσώπου στο αντίστοιχο σύνολο χαμηλού, μεσαίου ή υψηλού αναστήματος συμβολίζεται στον παρακάτω πίνακα με 0 ή 1.

Όνομα	**Ύψος (μέτρα)**	**Χαμηλού Αναστήματος**	**Μεσαίου Αναστήματος**	**Υψηλού Αναστήματος**
Άννα	1,60	1	0	0
Γιώργος	1,75	0	1	0
Χριστίνα	1,80	0	1	0
Παναγιώτα	1,77	0	1	0
Θανάσης	1,77	0	1	0
Μαρία	1,82	0	0	1
Βασίλης	2,01	0	0	1
Μανώλης	1,69	1	0	0

Πίνακας 6.6. *Τα κλασσικά σύνολα χαμηλού, μεσαίου ή υψηλού αναστήματος ύψους ανθρώπων*

Στην Ασαφή Λογική, όπως έχει ήδη προαναφερθεί, οι τιμές δεν είναι μόνο 0 ή 1. Δηλαδή ένα στοιχείο μπορεί να ανήκει σε δύο σύνολα (μιας συγκεκριμένης έννοιας, π.χ. του αναστήματος, εισοδήματος, αποτελεσματικότητας, επικινδυνότητας, ικανοποίησης πελάτη, κλπ.) αλλά με διαφορετικούς βαθμούς συμμετοχής (membership degree). Ο παρακάτω πίνακας δείχνει πώς θα μπορούσαν να ορισθούν τα **Ασαφή Σύνολα** του χαμηλού, μεσαίου και υψηλού αναστήματος.

Όνομα	**Ύψος (μέτρα)**	**Χαμηλού Αναστήματος**	**Μεσαίου Αναστήματος**	**Υψηλού Αναστήματος**
Άννα	1,60	1	0,4	0,1
Γιώργος	1,75	0,3	1	0,4
Χριστίνα	1,80	0,1	1	0,6
Παναγιώτα	1,77	0,1	1	0,5
Θανάσης	1,77	0,1	1	0,5
Μαρία	1,82	0,1	0,6	0,7
Βασίλης	2,01	0	0	1
Μανώλης	1,69	0,8	0,5	0,1

Πίνακας 6.7. *Τα ασαφή σύνολα χαμηλού, μεσαίου ή υψηλού αναστήματος ύψους ανθρώπων*

Παρατηρούμε από το παραπάνω πίνακα ότι οι τιμές που χαρακτηρίζουν τη συμμετοχή ενός προσώπου στο ένα ή το άλλο σύνολο αναστήματος ανήκουν στο κλειστό διάστημα [0,1]. Δηλαδή μπορούν να πάρουν τιμές από το 0 έως και το 1. Ο κάθε αριθμός υποδηλώνει το βαθμό συμμετοχής μιας τιμής

αναστήματος σε κάποιο από τα τρία παραπάνω σύνολα. Για παράδειγμα, το ύψος του Βασίλη (2,01) είναι σίγουρα ένα υψηλού αναστήματος και σίγουρα όχι χαμηλού ή μεσαίου. Το ύψος της Μαρίας βέβαια, μάλλον δεν είναι χαμηλού και είναι ίσως κάτι μεταξύ μεσαίου και υψηλού προς το υψηλό. Οι χαρακτηρισμοί αυτοί που δίνονται περιγραφικά βλέπουμε πως εκφράζουν με ποιοτικό τρόπο το ανάστημα κάθε προσώπου. Κάθε ασαφές σύνολο λοιπόν, ορίζεται από μια συνάρτηση που συνδέει τις τιμές μιας μεταβλητής (π.χ. ύψος, ταχύτητα, ικανοποίηση πελάτη, κλπ.), με ένα αριθμό στο διάστημα [0,1] που δηλώνει το βαθμό συμμετοχής μιας τιμής στο εν λόγω σύνολο. Το ασαφές σύνολο **μεσαίο ανάστημα,** για παράδειγμα, μπορεί να ορισθεί ως η παρακάτω **Συνάρτηση Συμμετοχής (membership function)**:

Ασαφές Σύνολο (ΜΑ) «Μεσαίο Ανάστημα»

Τιμές Αναστήματος (χ) (τιμές μεταβλητής)	**Τιμές Συμμετοχής μ(χ)**
1,00	0
1,20	0
1,50	0,1
1,60	0,2
1,65	0,3
1,70	0,6
1,75	1
1,78	1
1,80	1
1,85	0,7
1,88	0,6
1,90	0,3
2,00	0
2,08	0

Πίνακας 6.8. *Το ασαφές σύνολο μεσαίου ύψους ανθρώπων*

Η παραπάνω συνάρτηση συμμετοχής μ(χ) μπορεί να αναπαρασταθεί και ως πίνακας όπως ο παραπάνω και αλγεβρικά ως:

$$\mathbf{MA= \{\mu 1/\chi 1 + \mu 2/\chi 2 + \mu 3/\chi 3 + \ldots \mu\omega/\chi\omega\}}$$

Δηλαδή, κάθε τιμή της μεταβλητής (χ) συνδέεται με την αντίστοιχη τιμή μ(χ) της συνάρτησης συμμετοχής. Τα σύμβολα «/» και «+» δεν δηλώνουν διαίρεση και πρόσθεση στις αναπαραστάσεις των ασαφών συνόλων, αλλά τη συμμετοχή των στοιχείων με την αντίστοιχη τιμή συμμετοχής τους (membership degree). Σε συνεχείς συναρτήσεις το «άθροισμα» αντικαθίσταται με «ολοκλήρωμα». Έτσι, το ασαφές σύνολο «μεσαίου ύψους» του παραδείγματός μας αναπαρίσταται όπως πιο κάτω:

ΜΑ= {1,00/0 + 1,20/0 + 1,50/1 + 1,60/0,2 + 1,65/0,3 + 1,70/0,6 + 1,75/1 + 1,78/1 + 1,80/1 +1,85/0,7 + 1,88/0,6 + 1,90/0,3 + 2,00/0 + 2,08/0}.

Θεωρώντας ως πεδίο τιμών το σύνολο Χ, οι πράξεις των συνόλων, *ένωση*, *τομή*, και *συμπληρωματικό*, ορίζονται με βάση τις τιμές των συναρτήσεων συμμετοχής των ασαφών συνόλων ως εξής (Ross, 2009):

1. Ένωση D των Α και Β: $\mu D(\chi)=\max\{(\mu A(\chi), \mu B(\chi)\}, \chi \in X.$
2. Τομή C των Α και Β: $\mu C(\chi)=\min\{(\mu A(\chi), \mu B(\chi)\}, \chi \in X.$
3. Συμπληρωματικό S του Α: $\mu sA(\chi)=(1- (\mu A(\chi)), \chi \in X.$

6.3.2 Ασαφής Ιεραρχική Ανάλυση Αποφάσεων (Fuzzy Analytic Hierarchy Process-FAHP)

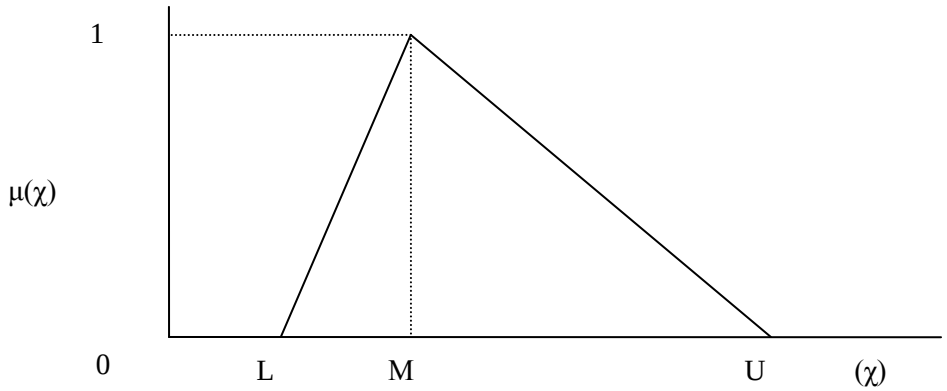
Η ασαφής λογική παρέχει το μαθηματικό υπόβαθρο για την ανάπτυξη τεχνικών, όπως ασαφής παλινδρόμηση, ασαφής βελτιστοποίηση, ασαφής πολυκριτηριακή ανάλυση, παράδειγμα της οποίας αποτελεί η Fuzzy Analytic Hierarchy Process-FAHP. Η FAHP (Ho, 2012), η οποία αποτελεί παραλλαγή της AHP, κάνοντας χρήση των αρχών και των μεθόδων της ασαφούς λογικής. Το επιχείρημα για τη χρήση της FAHP έναντι της

ΑΗΡ είναι ότι οι άνθρωποι μπορούν να εκφραστούν καλύτερα και πιο φυσιολογικά σε όρους λεκτικούς παρά σε αριθμητικούς. Δηλαδή είναι πιο εύκολο και φυσιολογικό για έναν άνθρωπο να εκφράσει την προτίμηση σε όρους «πολύ προτιμότερο» παρά με την έκφραση «2 φορές προτιμότερο». Στη μέθοδο FAHP, οι προτιμήσεις σε σχέση με τα κριτήρια εκφράζονται πάλι σε κλίμακα όπως στην περίπτωση της ΑΗΡ. Για παράδειγμα δίδεται η κλίμακα στον επόμενο Πίνακα 6.9.

Λεκτική Κλίμακα (Linguistic scale)	**Τριγωνικά Ασαφή Σύνολα Triangular Fuzzy Number (TFN)**	**Αντίστροφα Τριγωνικά Ασαφή Σύνολα**
Ίσης σημασίας	(1, 1, 1)	(1, 1, 1)
Ασθενώς πιο σημαντικό	(2/3, 1, 3/2)	(2/3, 1, 3/2)
Αρκετά πιο σημαντικό	(3/2, 2, 5/2)	(2/5, 1/2, 2/3)
Έντονα πιο σημαντικό	(5/2, 3, 7/2)	(2/7, 1/3, 2/5)
Εξαιρετικά πιο σημαντικό	(7/2, 4, 9/2)	(2/9, 1/ 4, 2/7)

Πίνακας 6.9. *Η κλίμακα προτιμήσεων εκφρασμένη σε ασαφή σύνολα*

Παρατηρείται ότι οι προτιμήσεις εκφράζονται σε τριάδες αριθμών. Κάθε μια τριάδα εκφράζει ένα τριγωνικό ασαφές σύνολο. Δηλαδή ένα τριγωνικό ασαφές σύνολο χαρακτηρίζεται ως εξής: $TFN = (L, M, U)$, όπου τα *L, M, U*, εκφράζουν τα τρία σημεία που ορίζουν το TFN ως εξής:



Σχήμα 6.2. *Γραφική παράσταση τριγωνικού ασαφούς συνόλου* $TFN = (L, M, U)$

Επομένως, ο πίνακας (Α) των ανά ζεύγη συγκρίσεων των κριτηρίων αποτελείται από στοιχεία τα οποία είναι TFN. Δηλαδή $a_{ij} = \{l, m, u\}$. Τα βήματα εφαρμογής της FAHP ακολουθούν.

Βήμα 1. Υπολογισμός των , S_i όπου i = 1,2,…,n και S_i δηλώνουν τα αθροίσματα των γραμμών των TFN στον πίνακα Α. Ο υπολογισμός των S_i γίνεται ως εξής.

$$S_i = \sum_{j=1}^{m} M_{g_i}^{j} \otimes \left[\sum_{i=1}^{n} \sum_{j=1}^{m} M_{g_i}^{j} \right]^{-1} \qquad (1)$$

όπου,

$$\sum_{j=1}^{m} M_{g_i}^{j} = \left(\sum_{j=1}^{m} l_j, \sum_{j=1}^{m} m_j, \sum_{j=1}^{m} u_j \right) \tag{2}$$

και

$$\sum_{i=1}^{n} \sum_{j=1}^{m} M_{g_i}^{j} = \left(\sum_{i=1}^{n} l_i, \sum_{i=1}^{n} m_i, \sum_{i=1}^{n} u_i \right) \tag{3}$$

Και $M_{g_i}^{j} (j = 1,2, \dots, m)$ είναι TFN.
Στη συνέχεια υπολογίζεται το

$$\left[\sum_{i=1}^{n} \sum_{j=1}^{m} M_{g_i}^{j} \right]^{-1} = \left(\frac{1}{\sum_{i=1}^{n} u_i}, \frac{1}{\sum_{i=1}^{n} m_i}, \frac{1}{\sum_{i=1}^{n} l_i} \right) \tag{4}$$

Βήμα 2. Θεωρώντας τα TFN $S_i = (l_i, m_i, u_i)$ υπολογίζεται η πιθανότητα ώστε $S_j = (l_j, m_j, u_j) \geq S_i = (l_i, m_i, u_i)$
Ως εξής

$$V(S_j > S_i) = \begin{Bmatrix} 1, & if\ m_j \geq m_i \\ 0, & if\ l_i \geq u_j \\ \dfrac{l_i - u_j}{(m_j - u_j) - (m_i - l_i)}, & otherwise \end{Bmatrix} \tag{5}$$

Βήμα 3. Υπολογισμός του ελάχιστου βαθμού πιθανότητας ώστε ένα κυρτό ασαφές σύνολο να είναι μεγαλύτερο από κ- άλλα κυρτά ασαφή σύνολα εφαρμόζοντας τη σχέση:

$$\begin{aligned} &V(S \geq S_1, S_2, ..., S_k) = V\left[(S \geq S_1) and (S \geq S_2) and \dots and (S \geq S_k) \right] \\ &= \min V(S \geq S_i),\ i = 1,2,3,...,k. \end{aligned} \tag{6}$$

Υποθέτοντας ότι

$$d'(A_i) = \min V(S_i \geq S_k), \quad for\ k = 1,2,...,n\ and\ k \neq i$$

τότε τα βάρη (βαθμοί σημαντικότητας) των κριτηρίων είναι:

$$W' = (d'(A_1), d'(A_2), \dots, d'(A_n))^T$$

Βήμα 4. Κανονικοποίηση των συντελεστών βαρύτητας.

Ακολουθεί ένα αριθμητικό παράδειγμα με τη χρήση του excel. Έστω ότι έχουμε τον πίνακα (Α) με τις ανά ζεύγη προτιμήσεις συνολικά 8 κριτηρίων. Για λόγους ευκρίνειας παρουσιάζεται ο πίνακας με τα πρώτα 3 κριτήρια.

Criteria	FACTOR 1			FACTOR 2			FACTOR 3		
TFN	L	M	U	L	M	U	L	M	U
FACTOR 1	1,000	1,000	1,000	0,667	1,619	3,500	0,667	1,122	2,500
FACTOR 2	0,286	0,618	1,500	1,000	1,000	1,000	0,667	1,260	2,500
FACTOR 3	0,400	0,891	1,500	0,400	0,794	1,500	1,000	1,000	1,000
FACTOR 4	0,400	0,618	1,000	0,400	0,661	1,000	0,667	0,794	1,500
FACTOR 5	0,286	0,618	1,500	0,400	0,794	1,500	0,286	0,408	0,667
FACTOR 6	0,400	0,550	1,000	0,400	0,661	1,500	0,400	0,707	1,000
FACTOR 7	0,286	0,400	1,500	0,286	0,833	1,500	0,400	0,707	1,500
FACTOR 8	0,286	0,500	1,500	0,400	0,794	1,500	0,286	0,589	1,000

Πίνακας 6.10. *Ο πίνακας Α με 3 κριτήρια και τις εκφρασμένες προτιμήσεις σε όρους TFN*

Παρατηρούμε ότι το κριτήριο1 (factor 1) έχει ίδιο βαθμό προτίμησης (λογικά) με τον εαυτό του και αναπαρίσταται με το (1,1,1) TFN, στο κελί (1,1) του πίνακα. Εφαρμόζοντας τους τύπους (1)-(4), υπολογίζουμε τα αθροίσματα των TFN, δηλαδή σε κάθε γραμμή του πίνακα υπολογίζονται τα $S_i = (l_i, m_i, u_i)$, δηλαδή τα αθροίσματα των (l), των (m) και των (u), παράγοντας τα πιο κάτω αποτελέσματα. Πιο συγκεκριμένα, με την εφαρμογή των σχέσεων (2) και (3) παίρνουμε τα παρακάτω αποτελέσματα.

Si		
L	M	U
6,333	12,495	22,500
5,619	9,625	18,500
6,633	10,921	17,500
5,133	9,446	17,500
4,257	7,756	13,667
4,667	7,729	14,500
3,610	6,090	11,000
2,943	5,557	10,500

Στη συνέχεια, γίνεται εφαρμογή της σχέσης (4) με τα παρακάτω αποτελέσματα.

SUM (L,M,U)	39,195	69,619	125,667
1/SUM (L,M,U)	0,008	0,014	0,026

Και με την εφαρμογή της σχέσης (1) λαμβάνουμε τα $S_i = (l_i, m_i, u_i)$.

S1 (F1)=	0,050	0,179	0,574
S2 (F2)=	0,045	0,138	0,472
S3 (F3)=	0,053	0,157	0,446
S4 (F4)=	0,041	0,136	0,446
S5 (F5)=	0,034	0,111	0,349
S6 (F6)=	0,037	0,111	0,370
S7 (F7)=	0,029	0,087	0,281
S8 (F8)=	0,023	0,080	0,268

Πίνακας 6.11. *Υπολογισμός των* $S_i = (l_i, m_i, u_i)$

Ακολουθεί η εφαρμογή της σχέσης (5).

V(S1>=Si)		V(S2>=Si)		V(S3>=Si)		V(S4>=Si)		V(S5>=Si)		V(S6>=Si)		V(S7>=Si)		V(S8>=Si)	
V(S1>=S2)	1,000	V(S2>=S1)	0,911	V(S3>=S1)	0,946	V(S4>=S1)	0,900	V(S5>=S1)	0,814	V(S6>=S1)	0,824	V(S7>=S1)	0,715	V(S8>=S1)	0,686
V(S1>=S3)	1,000	V(S2>=S3)	0,957	V(S3>=S2)	1,000	V(S4>=S2)	0,994	V(S5>=S2)	0,919	V(S6>=S2)	0,923	V(S7>=S2)	0,823	V(S8>=S2)	0,793
V(S1>=S4)	1,000	V(S2>=S4)	1,000	V(S3>=S4)	1,000	V(S4>=S3)	0,949	V(S5>=S3)	0,867	V(S6>=S3)	0,874	V(S7>=S3)	0,767	V(S8>=S3)	0,736
V(S1>=S5)	1,000	V(S2>=S5)	1,000	V(S3>=S5)	1,000	V(S4>=S5)	1,000	V(S5>=S4)	0,927	V(S6>=S4)	0,930	V(S7>=S4)	0,833	V(S8>=S4)	0,803
V(S1>=S6)	1,000	V(S2>=S6)	1,000	V(S3>=S6)	1,000	V(S4>=S6)	1,000	V(S5>=S6)	1,000	V(S6>=S5)	0,999	V(S7>=S5)	0,912	V(S8>=S5)	0,881
V(S1>=S7)	1,000	V(S2>=S7)	1,000	V(S3>=S7)	1,000	V(S4>=S7)	1,000	V(S5>=S7)	1,000	V(S6>=S7)	1,000	V(S7>=S6)	0,912	V(S8>=S6)	0,881
V(S1>=S8)	1,000	V(S2>=S8)	1,000	V(S3>=S8)	1,000	V(S4>=S8)	1,000	V(S5>=S8)	1,000	V(S6>=S8)	1,000	V(S7>=S8)	1,000	V(S8>=S7)	0,969

Πίνακας 6.12. *Υπολογισμοί του βήματος 2*

Η εφαρμογή της σχέσης (6) δίνει τα παρακάτω αποτελέσματα.

D'(1)=minV(S1>=S2,S3,S4,S5,S6,S7,S8	1,000
D'(2)=minV(S2>=S1,S3,S4,S5,S6,S7,S8	0,911
D'(3)=minV(S3>=S1,S2,S4,S5,S6,S7,S8	0,946
D'(4)=minV(S4>=S1,S2,S3,S5,S6,S7,S8	0,900
D'(5)=minV(S5>=S1,S2,S3,S4,S6,S7,S8	0,814
D'(6)=minV(S6>=S1,S2,S3,S4,S5,S7,S8	0,824
D'(7)=minV(S7>=S1,S2,S3,S4,S5,S6,S8	0,715
D'(8)=minV(S8>=S1,S2,S3,S4,S5,S6,S7	0,686

Πίνακας 6.13. *Υπολογισμοί του βήματος 3*

Από τα αποτελέσματα του πίνακα 6.19 προκύπτουν οι συντελεστές βαρύτητας των κριτηρίων.

W'=	1,000	0,910928482	0,946	0,900	0,814	0,824	0,715	0,686	5,395

Το άθροισμα των βαρών ισούται με 5,395. Στη συνέχεια, γίνεται κανονικοποίηση (διαίρεση κάθε βάρους με το άθροισμα των βαρών) και λαμβάνουμε τα τελικά κανονικοποιημένα βάρη των κριτηρίων.

W=	0,185	0,169	0,175	0,167	0,151	0,15265302	0,132	0,127

Πίνακας 6.14. *Κανονικοποιημένοι συντελεστές σημαντικότητας των κριτηρίων*

Με βάση αυτά τα κριτήρια και τη σημαντικότητά τους, αξιολογούνται οι εναλλακτικές λύσεις.

6.4 Μέθοδοι Ομοιότητας (Similarity Methods)

Οι μέθοδοι ομοιότητας (ΜΟ) αποσκοπούν στο να προσδιορίσουν κάποιο είδος ή πρότυπο ομοιότητας και τον αντίστοιχο βαθμό της ομοιότητας μεταξύ δύο ή περισσότερων στοιχείων. Οι μέθοδοι αυτές έχουν βρει αρκετές εφαρμογές στη λήψη των αποφάσεων, στη μηχανική μάθηση (machine learning) όπως για παράδειγμα έγκριση υπογραφών (Doroz, et al, 2016) και βεβαίως στα συστήματα συστάσεων, όπου το ζητούμενο είναι ο προσδιορισμός των όμοιων πελατών, των όμοιων προϊόντων και υπηρεσιών, έτσι ώστε να μπορούν να προταθούν (Dessì & Pes, 2015).

Ας υποθέσουμε λοιπόν ότι είναι διαθέσιμο ένα δείγμα δεδομένων X, το οποίο αναπαρίσταται με τη μορφή διανύσματος (array), ως εξής:

$$X = \{x_1, x_2, ..., x_n\}.$$

Κάθε στοιχείο x_i στο διάνυσμα X, είναι και αυτό με τη σειρά του ένα διάνυσμα *m*-διάστασης. Δηλαδή,

$$x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$$

Το ζητούμενο είναι να διαμορφωθούν σχέσεις (relations) μεταξύ των στοιχείων του διανύσματος X, δηλαδή μεταξύ των στοιχείων-διανυσμάτων x_i, x_j,…, x_m. Οι (MO), δηλαδή, υπολογίζουν την «ομοιότητα» δύο διανυσμάτων x_i και x_j, τα οποία διανύσματα αναπαριστούν τα προς σύγκριση στοιχεία, π.χ. χρήστες, προϊόντα, κλπ. Δηλαδή, επιδιώκεται ο υπολογισμός της ομοιότητας μεταξύ χρηστών, η ομοιότητα μεταξύ υπηρεσιών, κλπ. Η ένταση της σχέσης υποδηλώνει το βαθμό ομοιότητας. Οι βαθμοί ομοιότητας παίρνουν τιμές στο διάστημα [0,1]. Στη περίπτωση που χρησιμοποιείται ασαφής λογική, η ομοιότητα δύο στοιχείων ορίζεται ως η ασαφής συνάρτηση $\tilde{S}$ και ο βαθμός ομοιότητας είναι ουσιαστικά ο βαθμός ένταξης (membership degree) της συνάρτηση συμμετοχής (membership function) της ομοιότητας των δύο στοιχείων $\mu_S(x_i, x_j) \in [0,1]$. Η σχέση ομοιότητας έχει τη μορφή ενός πίνακα S(nXn). Ο πίνακας ομοιότητας έχει δύο ιδιότητες:

- Τα στοιχεία του στη διαγώνιο ισούνται με ένα (1) (reflectivity), δηλαδή $S(i, j) = 1$. Η ιδιότητα αυτή υποδηλώνει πως ένα στοιχείο έχει βαθμό ομοιότητας με τον εαυτό του το ένα (1), δηλαδή το μέγιστο.
- Ο πίνακας S(nXn) είναι συμμετρικός (symmetric) δηλαδή $S(i, j) = S(j, i)$. Η ιδιότητα αυτή δηλώνει πως ένα στοιχείο x_i και ένα στοιχείο x_j έχουν λογικά τον ίδιο βαθμό ομοιότητας με τα x_j και x_i.

Υπάρχουν αρκετές μέθοδοι ομοιότητας. Με την υπόθεση ότι έχουμε ένα δείγμα δεδομένων $X = \{x_1, x_2, ..., x_n\}$, όπου $x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$, πιο κάτω αναφέρονται ενδεικτικά μερικές από τις πιο διαδεδομένες μεθόδους (Ross, 2009).

Η Μέθοδος συνημίτονου (Cosine Amplitude)

$$s_{i,j} = \frac{\left|\sum_{k=1}^{m} x_{ik} x_{jk}\right|}{\sqrt{(\sum_{k=1}^{m} x_{ik}^2)(\sum_{k=1}^{m} x_{jk}^2}},$$

όπου $i, j = 1,2,..., n$.

Η Μέθοδος Max-Min

$$s_{i,j} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \max(x_{ik}, x_{jk})},$$

όπου $i, j = 1,2,...,n$.

Η Μέθοδος geometric Average minimum

$$s_{i,j} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \left[x_{ik}, x_{jk}\right]^{1/2}}$$

όπου $i, j = 1,2,...,n$.

Άλλες μέθοδοι είναι η Chi-squared (X^2) (Liu & Setiono, 1995), η Information Gain (Hall & Holmes, 2003), η Symmetrical Uncertainty (Witten, Frank, & Hall, 2011), κλπ.

Ακολουθεί αριθμητικό παράδειγμα εφαρμογής των μεθόδων Cosine Amplitude και Max-Min. Ας υποθέσουμε τα παρακάτω δεδομένα:

	Cosine Amlitude				
	Ταινία 1	Ταινία 2	Ταινία 3	Ταινία 4	Ταινία 5
Πολύ καλή	0,2	0,5	0,7	0,9	0,2
Μέτρια	0,5	0,3	0,2	0,1	0,7
Χαμηλού επιπέδου	0,3	0,2	0,1	0	0,1

Πίνακας 6.15. *Δεδομένα αξιολόγησης ταινιών*

Στον παραπάνω πίνακα, υποθέτουμε ότι έχουμε δεδομένα αξιολόγησης ταινιών από έναν αριθμό ανθρώπων. Οι αριθμοί στα κελιά του πίνακα δηλώνουν το ποσοστό των ανθρώπων που αξιολόγησαν μία ταινία ως «πολύ καλή», «μέτρια», κλπ. Τα ποσοστά ανά ταινία είναι κανονικοποιημένα, δηλάδη έχουν άθροισμα που ισούται με 1. Τα δεδομένα πρέπει να είναι κανονικοποιημένα ή να κανονικοποιούνται, ώστε να είναι δυνατή η σύγκριση των αξιολογήσεων των ταινιών στην ίδια κλίμακα, ακόμα και αν οι χρήστες που αξιολογούν χρησιμοποιούν διαφορετικές κλίμακες αξιολόγησης, κάτι που θα καθιστούσε τη σύγκριση των ταινιών μη δυνατή. Εφαρμόζοντας την μέθοδο *Cosine Amplitude*, χρησιμοποιώντας το MS EXCEL, παίρνουμε τον πιο κάτω πίνακα ομοιότητας S.

	Ταινία 1	Ταινία 2	Ταινία 3	Ταινία 4	Ταινία 5
Πίνακας Ομοιότητας	1	0,815789	0,306227	0,23	0,456227
	0,81578947	1	0,454151	0,48	0,354151
	0,30622662	0,454151	1	0,19	0,385185
	0,23	0,48	0,19	1	0,74
	0,45622662	0,354151	0,385185	0,74	1

Πίνακας 6.16. *Πίνακας Ομοιότητας Ταινιών (Similarity ή Proximity ή Tolerance Matrix)*

Ο υπολογισμός, για παράδειγμα, του S(1,2) γίνεται ως εξής:

$$S_{1,2} = \frac{(0,2*0,5)+(0,5*0,3)+(0,3*0,2)}{\sqrt{((0,2)^2+(0,5)^2+(0,3)^2)}*((0,5)^2+(0,3)^2+(0,2)^2)} = 0,81578947$$

Για τον υπολογισμό του S(1,2) με τη μέθοδο Max-Min, πραγματοποιούνται οι παρακάτω υπολογισμοί:

$$S_{1,2} = \frac{\min(0,2/0,5)+\min(0,5/0,3)+\min(0,3/0,2)}{\max(0,2/0,5)+\max(0,5/0,3)+\max(0,3/0,2)} = \frac{0,2+0,3+0,2}{0,5+0,5+0,3} = \frac{0,7}{1,3} = 0,538462$$

Γίνεται φανερό ότι διαφορετικές μέθοδοι ομοιότητας υπολογίζουν διαφορετικούς βαθμούς ομοιότητας. Σημασία βεβαίως έχει η σχετική ομοιότητα και όχι ο απόλυτος βαθμός ομοιότητας που έχουν υπολογισθεί από διαφορετικές μεθόδους.

Ομοίως, εφαρμόζοντας τη μέθοδο Cosine Amplitude, υπολογίζονται όλα τα στοιχεία του πίνακα ομοιότητας S(1,2). Από τον πίνακα ομοιότητας βλέπουμε το βαθμό ομοιότητας κάθε ταινίας με οποιαδήποτε άλλη για την οποία έχουμε αξιολογήσεις και βρίσκεται στον πίνακα. Όσο πιο κοντά στο ένα (1) ο βαθμός ομοιότητας, τόσο πιο όμοιες είναι δύο ταινίες. Έτσι, η ταινία 1 είναι πολύ «κοντά» στην ταινία 2, εφόσον ο βαθμός ομοιότητας είναι S(1,2)=0,815 ενώ είναι λίγο «όμοια» με την ταινία 4, εφόσον ο βαθμός ομοιότητας είναι S(1,4)=0,23.

Παρακάτω παρατίθεται ακόμα ένα παράδειγμα εφαρμογής της μεθόδου Cosine Amplitude, με διαφορετική παρουσίαση δεδομένων.

Ας υποθέσουμε ότι είναι διαθέσιμα τα δεδομένα αξιολόγησης ταινιών ανά χρήστη.

	Cosine Amlitude				
	Ταινία 1	Ταινία 2	Ταινία 3	Ταινία 4	Ταινία 5
Χρήστης 1	3	2	5	2	1
Χρήστης 2	5	2	3	3	5
Χρήστης 3	1	2	5	5	3

Πίνακας 6.17 *Πίνακας Αξιολόγησης Ταινιών ανά χρήστη*

Ο πίνακας 6.17 παρουσιάζει τις αξιολογήσεις των χρηστών για κάθε μια ταινία που παρακολούθησαν. Δηλαδή διακρίνουμε στον παραπάνω πίνακα τα διανύσματα των ταινιών (οι αντίστοιχες στήλες) και τα διανύσματα των χρηστών (οι αντίστοιχες γραμμές). Έτσι, εφαρμόζοντας μια (ΜΟ), ουσιαστικά συγκρίνουμε

την ομοιότητα δύο διανυσμάτων. Δηλαδή η σύγκριση της ταινίας 1 και της ταινίας 2 ανάγεται στη σύγκριση των παρακάτω διανυσμάτων:

Ταινία-1={3,5,1} και Ταινία-2={2,2,2}. Παρομοίως, εάν θέλουμε τη σύγκριση δύο χρηστών τότε θεωρούμε τα αντίστοιχα διανύσματα. Για παράδειγμα έστω τα διανύσματα Χρήστης-1={3,2,5,2,1} και Χρήστης-2={5,2,3,3,5}.

Μετά τους υπολογισμούς για όλα τα ζεύγη ταινιών και όλα τα ζεύγη χρηστών, έχουμε τα παρακάτω:

	Cosine based similarity				
	Ταινία 1	Ταινία 2	Ταινία 3	Ταινία 4	Ταινία 5
Ταινία 1	1	0,87	0,77	0,71	0,88
Ταινία 2		1	0,97	0,93	0,87
Ταινία 3			1	0,92	0,77
Ταινία 4				1	0,87
Ταινία 5					1

Πίνακας 6.18. *Πίνακας Ομοιότητας Ταινιών*

Ο πίνακας 6.18, αναφέρεται στο βαθμό ομοιότητας των ταινιών. Με ίδιο τρόπο λαμβάνουμε τον πίνακα ομοιότητας χρηστών στον πίνακα 6.19.

	Cosine based similarity		
	Χρήστης 1	Χρήστης 2	Χρήστης 3
Χρήστης 1	1	0,8	0,85
Χρήστης 2		1	0,79
Χρήστης 3			1

Πίνακας 6.19. *Πίνακας Ομοιότητας Χρηστών*

Οι υπολογισμοί, όπως αναφέρθηκε πιο πάνω, μπορούν να πραγματοποιηθούν με ευκολία στο MS EXCEL. Όμως ως προς το cosine similarity, υπάρχει ενδεικτικά λογισμικό στο διαδίκτυο, στη διεύθυνση (*http://www.appliedsoftwaredesign.com/archives/cosine-similarity-calculator*), που πραγματοποιεί τους υπολογισμούς.

Χρησιμοποιώντας (ΜΟ) μπορούμε να προσδιορίσουμε ομοιότητες μεταξύ χρηστών, πελατών, αλλά και μεταξύ προϊόντων και υπηρεσιών, στα πλαίσια ενός συστήματος προτάσεων, τα οποία εξετάζονται σε άλλο κεφάλαιο.

6.5 Συσταδοποίηση με βάση τον πίνακα Equivalence (Clustering with Equivalence Matrix)

Οι μέθοδοι ομοιότητας χρησιμοποιούνται επίσης και για τη δημιουργία του πίνακα equivalence (πίνακας ισοδυναμίας). Ο πίνακας equivalence έχει μία επιπρόσθετη ιδιότητα από αυτές που έχει ο πίνακας tolerance. Δηλαδή εκτός των γνωστών reflective και symmetric ο equivalence πίνακας είναι και transitive. Δηλαδή έστω

ότι $\mu_R(x_i, x_j) = \lambda_1$ και $\mu_R(x_j, x_i) = \lambda_2$ τότε $\mu_R(x_i, x_k) = \lambda$, όπου $\lambda \geq \min[\lambda_1, \lambda_2]$. Ο equivalence πίνακας παράγεται με τις διαδοχικές συνθέσεις (compositions) του tolerance πίνακα. Λογισμικά όπως το MATLAB υποστηρίζουν με την συγγραφή κώδικα τον υπολογισμό ενός equivalence πίνακα. Στο σημείο αυτό θα παρουσιαστεί ένα παράδειγμα για το πώς ένας equivalence πίνακας μπορεί να χρησιμοποιηθεί για να ομαδοποιηθούν δεδομένα με βάση κοινά χαρακτηριστικά, ή συμπεριφορές. Έστω ο παρακάτω equivalence πίνακας (R).

$$R = \begin{matrix} 1 & 0.8 & 0.4 & 0.5 & 0.8 \\ 0.8 & 1 & 0.4 & 0.5 & 0.9 \\ 0.4 & 0.4 & 1 & 0.4 & 0.4 \\ 0.5 & 0.5 & 0.4 & 1 & 0.5 \\ 0.8 & 0.9 & 0.4 & 0.5 & 1 \end{matrix}$$

Κάθε γραμμή ή στήλη του πίνακα (R) αναφέρεται σε κάποιο χαρακτηριστικό ή έννοια ή οντότητα x_i. Στο σημείο αυτό παρουσιάζουμε την έννοια λ_{cut}, που δηλώνει ένα όριο. Όσα στοιχεία του πίνακα είναι μεγαλύτερα ή ίσα με το λ_{cut} τότε αυτά αντικαθιστώνται από τη μονάδα και εάν είναι μικρότερο του λ_{cut} τότε γίνεται μηδέν. Δηλαδή,

$a_{ij} = 1,$ εάν $\alpha_{ij} \geq \lambda_{cut}$, ενώ $a_{ij} = 0,$ εάν $\alpha_{ij} < \lambda_{cut}$

Έτσι για τον πιο πάνω πίνακα για λ_{cut}=0.9 έχουμε.

$$R_{0.9} = \begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{matrix}$$

Παρομοίως για λ_{cut}=0.5 έχουμε.

$$R_{0.5} = \begin{matrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{matrix}$$

Παρατηρώντας τους πίνακες $R_{0.9}$ και $R_{0.5}$ μπορούμε να ομαδοποιήσουμε τα χαρακτηριστικά (π.χ. πελάτες, προϊόντα, κλπ.) σε ομάδες ως εξής.

Επίπεδο λ_{cut}	**Ομάδες**
λ_{cut} =0.9	$\{x_1\},\{x_2, x_5\},\{x_3\},\{x_4\}$
λ_{cut} =0.5	$\{x_1, x_2, x_4, x_5\},\{x_3\}$

Το πλεονέκτημα της μεθόδου συσταδοποίησης με βάση τον equivalence πίνακα είναι ότι σε αντίθεση με άλλες τεχνικές clustering (όπως k-means, c-means), δεν απαιτεί να γίνει κάποια αρχική υπόθεση για τον αριθμό των ομάδων που θα προκύψουν. Περισσότερα στοιχεία για τη μέθοδο και τον τρόπο εφαρμογής της, μπορεί να βρει ο αναγνώστης στους Ross (2009) και Kardaras, et al. (2011).

6.6 Μέθοδοι εξόρυξης κανόνων

Με τον όρο εξόρυξη κανόνων από δεδομένα αναφερόμαστε σε μια κατηγορία τεχνικών που προορίζονται για την αντιμετώπιση μεγάλου (ως τεράστιου) όγκου δεδομένων, συνήθως μη ελέγξιμης δομής και αμφίβολης ποιότητας. Οι μέθοδοι της κατηγορίας αυτής βασίζονται κυρίως στην υπολογιστική ισχύ των μοντέρνων συστημάτων Η/Υ και σε «έξυπνους αλγορίθμους», ικανούς να ανακαλύπτουν πρότυπα χωρίς οι απαιτήσεις σε επεξεργασία και μνήμη να εκτοξεύονται, λόγω του όγκου των δεδομένων, σε μη πρακτικά επίπεδα. Στη συνέχεια της ενότητας αυτής, περιγράφονται δύο τυπικές οικογένειες τέτοιων μεθόδων, η εξόρυξη κανόνων συσχέτισης (Mining of Association Rules) και η κατάταξη με δέντρα αποφάσεων (Decision Tree classifiers).

6.6.1 Εξόρυξη κανόνων συσχέτισης

Οι κανόνες συσχέτισης είναι δηλώσεις του τύπου **Αν … τότε … (if …. then …)** που εκφράζουν σχέσεις ανάμεσα σε ομοειδή αντικείμενα, γεγονότα ή ιδιότητες, με την έννοια της ταυτόχρονης εμφάνισης ή της πρόβλεψης κάποιων από αυτά, εφόσον γνωρίζουμε ότι έχουν εμφανιστεί κάποια άλλα. Σε τέτοιου είδους κανόνες βασίζονται π.χ. συστήματα προτάσεων που ενσωματώνονται σε πολλά ηλεκτρονικά καταστήματα, όπου αν επιλέξεις κάποιο προϊόν, σου προτείνουν και κάποιο άλλο για το οποίο προβλέπεται ότι ενδιαφέρεσαι. Από τις πιο χαρακτηριστικές εφαρμογές των κανόνων συσχέτισης είναι η μελέτη των αγορών των πελατών (π.χ. ενός σούπερ μάρκετ), όπως αυτές καταγράφονται σε ένα σύστημα δοσοληψιών (π.χ. στην ταμειακή μηχανή), για την ανακάλυψη ειδών που τείνουν οι πελάτες να αγοράζουν μαζί με άλλα είδη (π.χ. αν αγοράσεις δημητριακά, προβλέπεται ότι θα αγοράσεις και γάλα). Η εφαρμογή αυτή είναι γνωστή με το όνομα ανάλυση καλαθιού αγορών (market basket analysis), βρίσκει όμως χρησιμότητα σε πληθώρα άλλων προβλημάτων, όπου θέλουμε να ανακαλύψουμε πρότυπα συσχέτισης (π.χ. αν εμφανιστεί μια βλάβη/σύμπτωμα σε κάποιο σύστημα, θα εμφανιστεί και μια άλλη). Σημειώνεται ότι η προσέγγιση βασίζεται καθαρά στην καταμέτρηση συχνών εμφανίσεων, χωρίς να αναζητούνται αιτιακές σχέσεις ή να γίνεται προσπάθεια ανάλυσης/επεξήγησης του φαινομένου.

Σε αντίθεση με τις μεθόδους κατάταξης, όπου υπάρχει συγκεκριμένη μεταβλητή-στόχος, που θα πρέπει να προβλεφθεί με βάση κάποιες «ανεξάρτητες» μεταβλητές, στην εξόρυξη κανόνων συσχέτισης δεν είναι γνωστός ο στόχος, καθώς οποιοδήποτε αντικείμενο μπορεί να συμμετέχει σε κάποιον κανόνα, είτε ως συνθήκη, είτε ως αποτέλεσμα. Επίσης, σε αντίθεση με επιβλεπόμενες μεθόδους εκμάθησης, όπου διαθέτουμε ένα σύνολο «σωστών» παραδειγμάτων από τα οποία καλείται η μέθοδος να «μάθει», η εκμάθηση είναι απολύτως μη επιβλεπόμενη.

Για την κατανόηση του σκεπτικού και των βασικών εννοιών της εξόρυξης κανόνων συσχέτισης, χρησιμοποιείται ένα απλουστευμένο παράδειγμα ανάλυσης καλαθιού αγορών, ορίζεται η βασική ορολογία και παρουσιάζονται οι μέθοδοι συνοπτικά και περιγραφικά. Τα δεδομένα του προβλήματος είναι της μορφής του Πίνακα 6.1 και αποτελούνται από στοιχεία για διαφορετικές αγορές που πραγματοποιήθηκαν σε ένα κατάστημα τροφίμων. Κάθε γραμμή του πίνακα είναι μια **δοσοληψία (transaction)**, που περιλαμβάνει ένα σύνολο ειδών που αγοράστηκαν ταυτόχρονα.

Δοσοληψία	**Είδη**
1	{Ψωμί, Γάλα}
2	{Ψωμί, Πατατάκια, Μπύρα, Αυγά}
3	{Γάλα, Πατατάκια, Μπύρα, Χυμός}
4	{Ψωμί, Γάλα, Πατατάκια, Μπύρα}
5	{Ψωμί, Γάλα, Πατατάκια, Χυμός}

Πίνακας 6.20. *Παράδειγμα δεδομένων δοσοληψιών*

Τα δεδομένα του Πίνακα 6.20. μπορούν εύκολα να μετατραπούν σε μορφή δυαδικού πίνακα, όπως φαίνεται στον Πίνακα 6.21. Στη μορφή αυτή, κάθε στήλη του πίνακα αντιστοιχεί σε ένα προϊόν, κάθε γραμμή σε μια δοσοληψία, ενώ το περιεχόμενο του πίνακα είναι η τιμή «1» όταν το προϊόν συμμετέχει στη δοσοληψία, διαφορετικά η τιμή «0». Σε πραγματικά προβλήματα, αναμένεται ότι ένας τέτοιος πίνακας μπορεί να έχει πολλές στήλες, αφού ο αριθμός των ειδών μπορεί να είναι μεγάλος (π.χ. σε ένα σούπερ μάρκετ μπορεί να ισχύουν 30.000 κωδικοί προϊόντων). Ο αριθμός των γραμμών μπορεί να είναι πολύ μεγαλύτερος, αφού ο αριθμός των αγορών που πραγματοποιούνται σε ένα κατάστημα λιανικής μπορεί να είναι πολλά εκατομμύρια ανά έτος. Επίσης αναμένεται ότι το μεγαλύτερο μέρος του πίνακα θα περιέχει «0», αφού από τα χιλιάδες προσφερόμενα προϊόντα, κάθε καλάθι θα περιέχει μερικές δεκάδες.

Δοσοληψία	**Ψωμί**	**Γάλα**	**Πατατάκια**	**Μπύρα**	**Αυγά**	**Χυμός**
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Πίνακας 6.21. *Τα δεδομένα των δοσοληψιών σε μορφή δυαδικού πίνακα*

Από τα δεδομένα των δοσοληψιών, θα θέλαμε να εξάγουμε κανόνες που δείχνουν συσχέτιση ανάμεσα σε είδη, όπως ο κανόνας

Αν Πατατάκια **τότε** Μπύρα

ο οποίος υποδεικνύει ότι αν κάποιος αγοράσει πατατάκια, θα αγοράσει και μπύρα, βασιζόμενος στη συχνή εμφάνιση του συνδυασμού {Μπύρα, Πατατάκια} (στο παράδειγμα, από τις 4 δοσοληψίες στις οποίες περιλαμβάνονται τα Πατατάκια, στις 3 περιλαμβάνεται και η Μπύρα). Το μέρος του κανόνα που αντιστοιχεί στο «Αν» είναι η συνθήκη ή κεφαλή του κανόνα, ενώ το μέρος του «τότε» είναι το αποτέλεσμα ή σώμα του κανόνα.

Τους συνδυασμούς ειδών που εμφανίζονται ταυτόχρονα σε μια δοσοληψία τους ονομάζουμε **σύνολα αντικειμένων (itemsets)** και ως μέγεθος των συνόλων αντικειμένων ορίζουμε τον αριθμό των αντικειμένων που περιλαμβάνουν. Αντίστοιχα, ως μέγεθος μιας δοσοληψίας ορίζουμε τον αριθμό των αντικειμένων που περιλαμβάνονται σε μια δοσοληψία. Μια δοσοληψία λέμε ότι περιλαμβάνει ένα σύνολο στοιχείων, αν το σύνολο στοιχείων είναι υποσύνολο της δοσοληψίας π.χ. η δοσοληψία 2 του παραδείγματος περιλαμβάνει το σύνολο στοιχείων {Ψωμί, Μπύρα}.

Σημαντικό χαρακτηριστικό ενός συνόλου στοιχείων είναι η **Υποστήριξή** του **(Support)**, που ορίζεται ως το ποσοστό των δοσοληψιών στις οποίες περιλαμβάνεται το συγκεκριμένο σύνολο στοιχείων. Π.χ. το σύνολο {Ψωμί, Γάλα} περιλαμβάνεται σε 3 από τις 5 δοσοληψίες, επομένως η υποστήριξή του είναι 0.6.

Η ικανότητα πρόβλεψης ενός κανόνα μετριέται με τη βοήθεια δύο ποσοτήτων, της **Υποστήριξης (Support)** και της **Εμπιστοσύνης (Confidence)**, που ορίζονται ως εξής:

- **Υποστήριξη** ενός κανόνα είναι το ποσοστό των δοσοληψιών για τις οποίες ισχύει η συνθήκη του κανόνα, δηλαδή που περιλαμβάνουν το σύνολο στοιχείων της κεφαλής του κανόνα. Η υποστήριξη εκφράζει το πόσο συχνά ή σπάνια μπορεί να χρησιμοποιηθεί ο κανόνας.
- **Εμπιστοσύνη** ενός κανόνα είναι το ποσοστό των δοσοληψιών στις οποίες ισχύει το προβλεπόμενο αποτέλεσμα προς αυτές στις οποίες είναι εφαρμόσιμος. Υπολογίζεται διαιρώντας τον αριθμό των δοσοληψιών που περιλαμβάνουν το σύνολο στοιχείων και της κεφαλής και του σώματος προς τον αριθμό των δοσοληψιών που περιλαμβάνουν το σύνολο στοιχείων της κεφαλής. Η εμπιστοσύνη εκφράζει το πόσο ακριβής είναι ο κανόνας δηλαδή το πόσο συχνά ισχύει πράγματι αυτό που προβλέπει.

Οι κανόνες συσχέτισης εξάγονται σε δύο στάδια: (α) εξόρυξη συχνών συνόλων και (β) κατασκευή κανόνων.

6.6.1.1 Εξόρυξη συχνών συνόλων

Κατά το στάδιο αυτό, αναλύεται ένας μεγάλος αριθμός από παραδείγματα, ώστε να ανιχνευθούν συνδυασμοί από αντικείμενα που εμφανίζονται συχνά στην ίδια δοσοληψία. Οι συνδυασμοί αυτοί ονομάζονται **συχνά σύνολα αντικειμένων (frequent itemsets)** και επιλέγονται με βάση ένα ελάχιστο όριο στη συχνότητα εμφάνισης. Το πρόβλημα της εύρεσης συχνών συνόλων φαίνεται απλό, αφού αρκεί να καταγραφούν όλα τα σύνολα αντικειμένων που εμφανίζονται στα παραδείγματα και να μετρηθεί η συχνότητα εμφάνισης του καθενός, ώστε, στη συνέχεια, να επιλεγούν αυτά που θεωρούνται συχνά. Η δυσκολία βρίσκεται στο ότι ο αριθμός των δυνατών συνόλων που μπορεί να δημιουργηθούν από τα αντικείμενα των δοσοληψιών ενός πραγματικού προβλήματος είναι κολοσσιαίος. Λόγω αυτού, απαιτούνται έξυπνοι αλγόριθμοι, ικανοί να εξάγουν τα συχνά σύνολα με αποτελεσματικό τρόπο. Το πρόβλημα έχει διατυπωθεί εδώ και μερικές δεκαετίες (Agrawal & Srikant, 1994) και από τότε έχουν αναφερθεί πολλές προσπάθειες ανάπτυξης αποτελεσματικότερων και εξυπνότερων μεθόδων (Wu, 2010). Οι δύο πιο γνωστοί αλγόριθμοι είναι ο **Apriori** και ο **FP-Growth,** οι οποίοι έχουν αναπτυχθεί σε διάφορες παραλλαγές.

Ο Apriori βασίζεται στην αρχή ότι αν ένα σύνολο αντικειμένων μεγέθους *k* δεν είναι συχνό, οποιοδήποτε υπερσύνολό του μεγέθους μεγαλύτερου από *k* δε θα είναι επίσης συχνό. Επομένως, σταματούμε να εξετάζουμε ως υποψήφια όλα τα σύνολα που περιέχουν κάποιο αντικείμενο ή σύνολο αντικειμένων που έχει απορριφθεί, περιορίζοντας από νωρίς την περιοχή αναζήτησης. Η λογική του Apriori, σε συντομία, είναι η εξής:

Ο αλγόριθμος ξεκινάει με την εύρεση των συχνών συνόλων μεγέθους 1, δηλαδή των μεμονωμένων αντικειμένων που ικανοποιούν ένα όριο ελάχιστης υποστήριξης (έστω *minsup)*. Στη συνέχεια, μια επαναληπτική διαδικασία βρίσκει διαδοχικά τα συχνά σύνολα μεγέθους 2, 3, κλπ., μέχρι να βρεθούν όλα τα συχνά σύνολα που ικανοποιούν τον περιορισμό ελάχιστης υποστήριξης.

Ο αλγόριθμος διατηρεί συνεχώς μια λίστα με τα συχνά σύνολα οποιουδήποτε μεγέθους που έχουν βρεθεί μέχρι στιγμής, και μια λίστα με τα έγκυρα αντικείμενα (αυτά που ικανοποιούν το όριο ελάχιστης υποστήριξης). Αφού βρεθούν τα συχνά σύνολα μεγέθους π.χ. 2, για καθένα από αυτά δημιουργούνται όλα τα υποψήφια συχνά σύνολα μεγέθους 3, δοκιμάζοντας ένα-ένα όλα τα υπόλοιπα έγκυρα αντικείμενα. Μετριέται η υποστήριξη των υποψήφιων συνόλων και επιλέγονται για προσθήκη στη λίστα συχνών συνόλων αυτά που ικανοποιούν τη συνθήκη ελάχιστης υποστήριξης.

Στο παράδειγμα του Πίνακα 6.1, ας επιλέξουμε ως όριο ελάχιστης υποστήριξης *minsup*=60% δηλαδή 3 από τις 5 δοσοληψίες. Τα αντικείμενα που εμφανίζονται σε τουλάχιστον 3 δοσοληψίες και, επομένως είναι τα έγκυρα αντικείμενα από τα οποία δημιουργούνται τα συχνά σύνολα μεγέθους 1 είναι:

{Ψωμί}, {Γάλα}, {Πατατάκια}, {Μπύρα}

Παρατηρούμε ότι τα Αυγά και ο Χυμός εμφανίζονται μόνο 1 ή 2 φορές και επομένως απορρίπτονται. Για την εύρεση των συχνών συνόλων μεγέθους 2, ξεκινάμε από ένα σύνολο μεγέθους 1 και δοκιμάζουμε όλους τους συνδυασμούς του με τα έγκυρα αντικείμενα. Π.χ. αν ξεκινήσουμε από το {Ψωμί}, τα υποψήφια σύνολα είναι:

{Ψωμί, Γάλα}, {Ψωμί, Πατατάκια}, {Ψωμί, Μπύρα}

Από αυτά, το {Ψωμί, Γάλα} περιλαμβάνεται σε 3 δοσοληψίες και το {Ψωμί, Πατατάκια} σε 4, επομένως προστίθενται στα συχνά σύνολα, ενώ το {Ψωμί, Μπύρα} μόνο σε 2 δοσοληψίες, επομένως απορρίπτεται. Συνεχίζουμε εξετάζοντας τα υποψήφια σύνολα που δημιουργούνται από καθένα από τα υπόλοιπα συχνά σύνολα μεγέθους 1, μέχρι να βρεθούν όλα τα συχνά σύνολα μεγέθους 2. Ομοίως, η διαδικασία επαναλαμβάνεται για τα μεγαλύτερα μεγέθη.

Ο Apriori μειώνει δραστικά τους απαιτούμενους υπολογισμούς σε σχέση με την «κουτή» διαδικασία του να δημιουργήσουμε και να εξετάσουμε όλους τους δυνατούς συνδυασμούς αντικειμένων. Ακόμα μεγαλύτερη βελτίωση στην ταχύτητα εκτέλεσης και στον περιορισμό της απαιτούμενης μνήμης επιτυγχάνεται με το αλγόριθμο FP-Growth (Frequency Pattern = πρότυπα συχνότητας). Το κύριο χαρακτηριστικό του τελευταίου είναι ότι αντί της τήρησης των ευρεθέντων συχνών συνόλων σε μορφή λίστας, αυτά κωδικοποιούνται κατάλληλα σε μορφή δέντρου. Με τον τρόπο αυτό, η επιλογή των συχνών συνόλων γίνεται με εξυπνότερο τρόπο, ώστε να απαιτούνται από τη διαδικασία να διατρέχει τα δεδομένα μόνο δύο φορές, αντί των πολλαπλών αναζητήσεων που πραγματοποιεί ο Apriori. Επιπλέον, η αποθήκευση των ενδιάμεσων αποτελεσμάτων για τη λειτουργία της διαδικασίας είναι περισσότερο συμπαγής, κάτι που μειώνει τις

απαιτήσεις σε μνήμη και κάνει δυνατή την επεξεργασία ακόμα μεγαλύτερων συνόλων δεδομένων. Πληρέστερη περιγραφή των μεθόδων είναι διαθέσιμη στην πλούσια σχετική βιβλιογραφία (Νανόπουλος, 2008).

Το όριο *minsup* που θα οριστεί για την ελάχιστη υποστήριξη είναι καθοριστικό για τον αριθμό των συχνών συνόλων που θα βρεθούν. Η επιλογή του γίνεται λαμβάνοντας υπόψη τα χαρακτηριστικά του προβλήματος και των δεδομένων, αλλά και με πειραματισμό. Ένα υψηλό όριο θα περιορίσει τα ευρήματα σε αντικείμενα που εμφανίζονται πολύ συχνά π.χ. υποστήριξη 80% στην ανάλυση καλαθιού αγορών σημαίνει ότι θα αναζητηθούν συνδυασμοί προϊόντων που θα πρέπει να βρίσκονται σε τουλάχιστον 8 από τα 10 καλάθια, κάτι που μπορεί να συμβαίνει μόνο για πολύ δημοφιλή καθημερινά προϊόντα.

6.6.1.2 Κατασκευή κανόνων

Μετά την εύρεση των συχνών συνόλων, ακολουθεί η διαδικασία δημιουργίας κανόνων συσχέτισης του τύπου $X \rightarrow Y$.

Ένα συχνό σύνολο μπορεί να οδηγήσει στη δημιουργία πολλών διαφορετικών κανόνων, χωρίζοντάς το σε 2 μη κενά υποσύνολα, από τα οποία το ένα χρησιμοποιείται ως συνθήκη (κεφαλή) και το άλλο ως αποτέλεσμα (σώμα). Π.χ. από το συχνό σύνολο

{Ψωμί , Γάλα, Χυμός}

μπορούν να δημιουργηθούν οι κανόνες:

Ψωμί , Γάλα → Χυμός
Ψωμί , Χυμός → Γάλα
Γάλα, Χυμός → Ψωμί

ενώ τα ίδια στοιχεία συσχετίζονται και μέσω των γενικότερων κανόνων:

Ψωμί → Χυμός
Ψωμί → Γάλα
Γάλα → Χυμός
Χυμός → Ψωμί
Γάλα → Ψωμί
Χυμός → Γάλα

Από όλους τους δυνατούς κανόνες, επιλέγονται αυτοί που πληρούν μια ελάχιστη απαίτηση εμπιστοσύνης. Το όριο εμπιστοσύνης (έστω *minconf*) χρησιμοποιείται για την επιλογή των κανόνων που είναι συχνά σωστοί, δηλαδή έχουν ικανοποιητική ακρίβεια πρόβλεψης. Επίσης, όλοι οι κανόνες πληρούν την απαίτηση για ελάχιστη υποστήριξη, εφόσον προέρχονται από συχνά σύνολα. Περισσότερες λεπτομέρειες σχετικά με τους αλγόριθμους εύρεσης των κανόνων από συχνά σύνολα δεν έχουν ενδιαφέρον για τον αναγνώστη. Ωστόσο, καίριας σημασίας για την επιλογή και αξιοποίηση των κανόνων σε πρακτικά προβλήματα είναι οι παρακάτω παρατηρήσεις:

- Ο συνδυασμός υψηλής υποστήριξης και υψηλής εμπιστοσύνης για έναν κανόνα είναι σημαντικό προαπαιτούμενο για να είναι αυτός χρήσιμος. Η υψηλή υποστήριξη σημαίνει ότι είναι συχνές οι περιπτώσεις στις οποίες μπορεί να χρησιμοποιηθεί και η υψηλή εμπιστοσύνη σημαίνει ότι όταν χρησιμοποιείται είναι συνήθως σωστός.
- Η χρήση του συνδυασμού υποστήριξης/εμπιστοσύνης ως μοναδικού κριτηρίου για την επιλογή των κανόνων έχει αδυναμίες, με σημαντικότερη το ότι δε λαμβάνονται υπόψη οι εκ των προτέρων πιθανότητες παρουσίας των αντικειμένων στις δοσοληψίες. Αν π.χ. ένα αντικείμενο i_1 έχει συχνότητα εμφάνισης 70% και ένα άλλο i_2 έχει συχνότητα 80%, η θεωρητική συχνότητα εμφάνισης του συνδυασμού τους $\{i_1, i_2\}$, αν είναι ανεξάρτητα, είναι 56%. Ενώ λοιπόν η υποστήριξη 56% του συνόλου αυτού φαίνεται μεγάλη, στην πραγματικότητα είναι η τυχαία υποστήριξη που θα περιμέναμε χωρίς τα αντικείμενα να έχουν καμία συσχέτιση μεταξύ τους. Επίσης, ο κανόνας $i_1 \rightarrow i_2$ έχει εμπιστοσύνη 80%, που είναι

ένα υψηλό ποσοστό, στην πραγματικότητα όμως το ποσοστό αυτό είναι ακριβώς η πιθανότητα να εμφανιστεί το i_2 ανεξάρτητα της παρουσίας ή όχι του i_1.

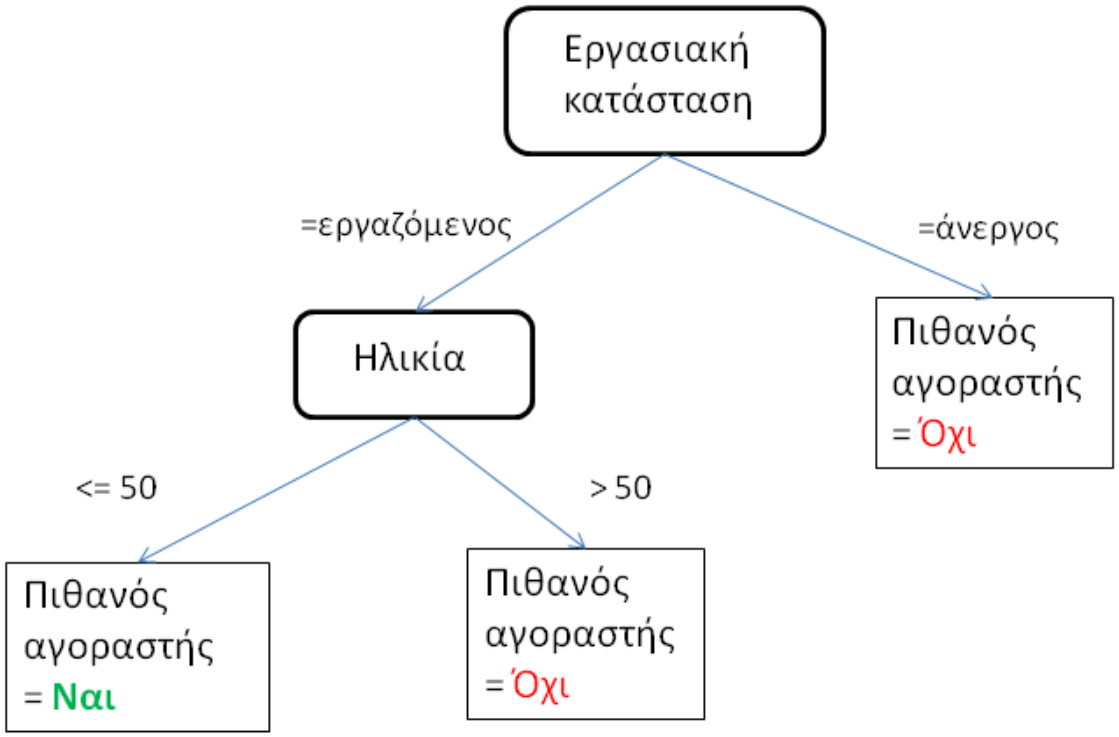
- Για την καλύτερη εκτίμηση της αξίας των κανόνων και την αντιμετώπιση προβλημάτων όπως τα παραπάνω, έχουν προταθεί διάφορες μετρικές, που μπορούν να αντικαταστήσουν την εμπιστοσύνη ως κριτήριο επιλογής κανόνων ή να παρέχουν επιπλέον πληροφόρηση για τις δυνατότητες αξιοποίησης των κανόνων. (Wu, 2010). Οι κυριότερες είναι:
    - **Lift.** Υπολογίζεται ως ο λόγος της υποστήριξης του κανόνα προς την υποστήριξη που θα είχαμε αν τα αντικείμενα που συνδέει ήταν ανεξάρτητα. Εκφράζει το πόσο απέχουν τα συσχετιζόμενα αντικείμενα από την ανεξαρτησία και επομένως το πόσο πραγματικά επηρεάζει η υπόθεση το αποτέλεσμα. Τιμές κοντά στο 1 δείχνουν ανεξαρτησία των αντικειμένων και χαμηλό ενδιαφέρον του κανόνα Π.χ. ας υποθέσουμε ότι πάρα πολλοί πελάτες αγοράζουν συχνά ψωμί, ανεξάρτητα από τις άλλες αγορές τους, με αποτέλεσμα το ψωμί να εμφανίζεται στο π.χ. 80% των καλαθιών. Τότε, ένας κανόνας που θα έλεγε ότι **όταν ένας πελάτης αγοράζει μουστάρδα, έχει πιθανότητα 81% να αγοράσει και ψωμί** θα είχε σχετικά μεγάλη εμπιστοσύνη (81%), αλλά καμία αξία, αφού και οποιοσδήποτε πελάτης, θα είχε πιθανότητα 80% να αγοράσει ψωμί, είτε αγόρασε μουστάρδα είτε όχι.
    - **Conviction.** Δείχνει το κατά πόσο περιέχεται χρήσιμη πληροφορία στη φορά του κανόνα, ώστε να εκτιμήσουμε ποιο είδος μέσα σε ένα σύνολο είναι αυτό που οδηγεί στην αγορά του άλλου. Π.χ. αν ο κανόνας **δημητριακά → γάλα** έχει υψηλότερη εμπιστοσύνη από τον αντίστροφο **γάλα → δημητριακά,** τότε έχει υψηλό conviction, που σημαίνει ότι η φορά του κανόνα εμπεριέχει πραγματική πληροφορία.
    - **Gain, Laplace και ps**. Ειδικότερα κριτήρια που σχετίζονται με εκτίμηση της ποσότητας πληροφορίας που περιέχεται στον κανόνα.

6.6.2 Κατάταξη με Δέντρα Αποφάσεων

6.6.2.1 Ορισμός του δέντρου αποφάσεων

Τα δέντρα αποφάσεων είναι δομές που παίρνουν το όνομά τους λόγω του σχήματός τους. Σκοπός τους είναι η μοντελοποίηση της κατάταξης αντικειμένων με βάση διάφορα χαρακτηριστικά τους σε γνωστές κατηγορίες ή, με άλλα λόγια, η πρόβλεψη ενός ονομαστικού χαρακτηριστικού στόχου με βάση άλλα χαρακτηριστικά. Σημαντικό πλεονέκτημα των δέντρων αποφάσεων ως μοντέλα κατάταξης/πρόβλεψης είναι ότι αποτελούν ιδιαίτερα παραστατικές και εύληπτες απεικονίσεις, που παρέχουν άμεσα κατανοητή από τον άνθρωπο ερμηνεία της εξαχθείσας γνώσης.

Τα δέντρα αποφάσεων αποτελούνται έναν αρχικό κόμβο που αντιστοιχεί στη ρίζα, τους εσωτερικούς κόμβους που αποτελούν τις διακλαδώσεις και τους τελικούς κόμβους ή φύλλα. Τα δέντρα εκφράζουν μια ιεραρχία, δηλαδή εμπεριέχουν την έννοια του ανώτερου και του κατώτερου, και έχουν ορισμένο βάθος. Ανώτερος κόμβος είναι η ρίζα, η οποία διασπάται και συνδέεται μέσω κλάδων (συνδέσμων που συμβολίζονται με ακμές) με κατώτερους κόμβους. Κάθε κόμβος αποτελεί διακλάδωση ενός ανώτερου κόμβου, με τον οποίο συνδέεται μέσω ενός εισερχόμενου συνδέσμου. Κάθε εσωτερικός κόμβος διακλαδίζεται σε 2 ή περισσότερους κατώτερους κόμβους, στους οποίους οδηγούν αντίστοιχοι εξερχόμενοι σύνδεσμοι. Οι τελικοί κόμβοι (ή φύλλα του δέντρου) αποτελούν την κατάληξη μιας σειράς διακλαδώσεων και δεν διασπώνται περεταίρω. Η ρίζα και κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα χαρακτηριστικό (με άλλα λόγια μια μεταβλητή) με βάση το οποίο μπορεί να διαχωριστεί το δείγμα σε υποομάδες. Ο διαχωρισμός του δείγματος (ή διάσπαση του κόμβου) γίνεται σύμφωνα με έναν κανόνα σύγκρισης της τιμής του χαρακτηριστικού με κάποιο όριο ή ονομαστική τιμή, ανάλογα με το αν το χαρακτηριστικό είναι ποσοτικό ή ονομαστικό. Κάθε φύλλο αντιστοιχεί σε μια κατηγορία, που εκφράζεται ως η τιμή του χαρακτηριστικού-στόχου σύμφωνα με το οποίο θέλουμε να κατατάξουμε τον πληθυσμό.



Σχήμα 6.3.. *Δέντρο αποφάσεων για την κατάταξη ενός πελάτη σε πιθανό ή όχι αγοραστή.*

Στο Σχήμα 6.3 παρουσιάζεται ένα απλό παράδειγμα δέντρου αποφάσεων για την κατάταξη των υποψηφίων πελατών σε πιθανούς ή μη πιθανούς αγοραστές ενός νέου προϊόντος.

6.6.2.2. Λειτουργία του δέντρου ως μοντέλο κατάταξης

Η ρίζα του δέντρου αντιστοιχεί στο χαρακτηριστικό **Εργασιακή κατάσταση**. Αν η τιμή του χαρακτηριστικού αυτού είναι «άνεργος», η κατάταξη οδηγείται στον τελικό κόμβο «Όχι», που αποτελεί και την τελική απόφαση. Αν η τιμή είναι «εργαζόμενος», η κατάταξη οδηγείται στον εσωτερικό κόμβο **Ηλικία** και, στη συνέχεια, διακλαδίζεται ανάλογα με την τιμή του τελευταίου, προς τον τελικό κόμβο «Ναι», αν η ηλικία είναι μικρότερη ή ίση του 50 και στο «Όχι» αν είναι μεγαλύτερη. Το δέντρο του παραδείγματος έχει βάθος 2 επειδή φτάνουμε στο πιο μακρινό φύλλο μετά από 2 διακλαδώσεις. Κάθε μονοπάτι που καταλήγει σε ένα φύλλο μπορεί να θεωρηθεί ως ένας κανόνας που στην κεφαλή του περιλαμβάνει, ως συνθήκες συνδεδεμένες με το λογικό **και**, όλες τις διακλαδώσεις που ακολουθεί, π.χ. το μονοπάτι που ξεκινάει από τη ρίζα και καταλήγει στο πρώτο από αριστερά φύλλο «Ναι» του σχήματος, ισοδυναμεί με τον κανόνα:

Αν Εργασιακή κατάσταση = *Εργαζόμενος* **και** Ηλικία <= *50* **Τότε** Πιθανός αγοραστής = *Ναι*

Οι συνολικοί κανόνες που μπορούν να προκύψουν από το δέντρο είναι όσοι και τα φύλλα (στο παραπάνω απλό παράδειγμα είναι μόλις 3) και ο μέγιστος αριθμός συνθηκών που βρίσκονται σε σύζευξη σε έναν κανόνα είναι όσο το βάθος του δέντρου (στο παραπάνω παράδειγμα είναι 2). Παρατηρούμε ακόμα ότι μπορεί πολλά φύλλα να αντιστοιχούν στην ίδια κατηγορία κατάταξης (π.χ. 2 φύλλα αντιστοιχούν στο «Όχι»). Αν λοιπόν θέλουμε έναν συνολικό κανόνα που να καθορίζει το πότε κατατάσσεται ένας πελάτης στην κατηγορία «Όχι», αυτός θα αποτελείται από όλους του κανόνες που καταλήγουν σε φύλλο «Όχι», συνδεδεμένους με το λογικό **ή,** π.χ.

Αν (Εργασιακή κατάσταση = *Άνεργος)* ***ή*** (Εργασιακή κατάσταση = *Εργαζόμενος* **και** Ηλικία > *50*) **Τότε** Πιθανός αγοραστής = *Όχι*

6.6.2.3 Βασικοί τύποι δέντρων αποφάσεων

Ένα δέντρο αποφάσεων μπορεί να επιτρέπει σε κάθε εσωτερικό κόμβο να έχει απεριόριστο αριθμό διακλαδώσεων π.χ. η Ηλικία να διαχωρίζει σε πολλές ηλικιακές κατηγορίες <= 20 ετών, 21 ως 30 ετών, κλπ. Ορισμένα δέντρα, που ονομάζονται δυαδικά (binary) επιτρέπουν μόνο δύο διακλαδώσεις σε κάθε εσωτερικό κόμβο. Τα δυαδικά δέντρα δέχονται ως χαρακτηριστικά των εσωτερικών κόμβων είτε μεταβλητές που είναι εξαρχής δυαδικές, δηλαδή που οι τιμές τους είναι τύπου Ναι/Όχι, είτε ονομαστικές που μετατρέπονται σε δυαδικές με χωρισμό των κατηγοριών σε δύο συμπληρωματικά υποσύνολα, είτε αριθμητικές που διαχωρίζονται σε δύο μέρη με βάση κάποιο όριο.

Τα δέντρα αποφάσεων διαχωρίζονται επίσης ανάλογα με τον τύπο των μεταβλητών που δέχονται ως χαρακτηριστικά διακλάδωσης. Ορισμένα δέντρα μπορούν να περιλαμβάνουν οποιοδήποτε συνδυασμό ποσοτικών (συνεχών και ασυνεχών), ονομαστικών και δυαδικών μεταβλητών. Το κριτήριο διακλάδωσης για τις ποσοτικές μεταβλητές μπορεί να είναι μια ανισότητα ή ένα διάστημα, ενώ για τις ονομαστικές θα είναι η ισότητα με κάποια από τις πιθανές ονομαστικές τιμές. Ορισμένοι τύποι δέντρων δέχονται αποκλειστικά ονομαστικές μεταβλητές, π.χ. δέντρα CHAID (βλέπε Κεφάλαιο 7).

6.6.2.4 Κατασκευή του δέντρου

Η κατασκευή ενός δέντρου αποφάσεων είναι και το κύριο πρόβλημα μοντελοποίησης, που πραγματοποιείται με μεθόδους εξόρυξης γνώσης από δεδομένα. Υπάρχουν πολλοί αλγόριθμοι κατασκευής δέντρων αποφάσεων, με διαφορετικά χαρακτηριστικά και πλεονεκτήματα ο καθένας. Αναφέρονται ως πιο διαδεδομένοι οι ID3, CART, CHAID (Νανόπουλος, 2008).

Η μοντελοποίηση γίνεται με βάση ένα εκπαιδευτικό σύνολο δεδομένων (παραδείγματα), για το οποίο είναι γνωστά όλα τα χαρακτηριστικά των παραδειγμάτων, συμπεριλαμβανομένης της επιθυμητής τιμής του χαρακτηριστικού-στόχου. Η ανάπτυξη του δέντρου πραγματοποιείται με επαναληπτικούς διαμερισμούς του εκπαιδευτικού συνόλου, με βάση τις τιμές των χαρακτηριστικών. Σε κάθε επανάληψη, ο αλγόριθμος ακολουθεί τα εξής βήματα:

- Ένα χαρακτηριστικό Α επιλέγεται για διαχωρισμό του δείγματος. Η επιλογή του καταλληλότερου χαρακτηριστικού είναι κρίσιμη για τη δημιουργία επιτυχημένου δέντρου και

βασίζεται σε κάποιο κριτήριο, που εξαρτάται από τον αλγόριθμο και σε κάποιες περιπτώσεις μπορεί να ρυθμιστεί από τον αναλυτή.

- Τα παραδείγματα διαμερίζονται σε υποσύνολα, ένα για κάθε τιμή του χαρακτηριστικού Α (για ονομαστικά χαρακτηριστικά) ή για κάθε διάστημα από αυτά στα οποία χωρίζεται η περιοχή τιμών ενός ποσοτικού χαρακτηριστικού.
- Για κάθε υποσύνολο δημιουργείται ένα κλάδος που οδηγεί είτε σε ένα κατώτερο υπο-δέντρο, είτε σε ένα φύλλο. Το υπο-δέντρο αναπτύσσεται επαναλαμβάνοντας επαναληπτικά τα προηγούμενα βήματα.

Η διάσπαση ενός κόμβου, ώστε να συνεχιστεί η ανάπτυξη του δέντρου, σταματάει και ο κόμβος χαρακτηρίζεται ως φύλλο, όταν όλα τα παραδείγματα που περιλαμβάνει έχουν την ίδια τιμή για το χαρακτηριστικό-στόχο. Σε αυτήν την περίπτωση, που είναι η πιο επιθυμητή, το φύλλο είναι «καθαρό» και η κατάταξη η ακριβέστερη δυνατή. Ωστόσο, μπορεί να επιτραπεί ένα περιθώριο σφάλματος, ώστε η διάσπαση να μπορεί να σταματήσει και όταν τα περισσότερα παραδείγματα έχουν την ίδια τιμή. Η ανάπτυξη ενός υπο-δέντρου επίσης σταματάει όταν:

- Τα παραδείγματα που περιλαμβάνονται σε ένα υπο-δέντρο είναι κάτω από ένα ελάχιστο όριο. Το όριο ρυθμίζεται από τον αναλυτή.
- Κανένα χαρακτηριστικό δε θεωρείται κατάλληλο για περεταίρω διάσπαση, σύμφωνα με κάποιο κριτήριο βελτίωσης της πληροφορίας του δέντρου.
- Το υπο-δέντρο έχει φτάσει στο μέγιστο επιτρεπτό βάθος.

Παράλληλα εφαρμόζεται μια διαδικασία «κλαδέματος», κατά την οποία τα φύλλα που δεν προσθέτουν στην ικανότητα κατάταξης του δέντρου αφαιρούνται. Με το κλάδεμα αποφεύγεται η λεγόμενη υπερ-προσαρμογή (over fitting) του δέντρου, δηλαδή η φωτογραφική εκμάθηση μεμονωμένων περιπτώσεων, και η επιτυγχάνεται η δημιουργία ενός δέντρου που μαθαίνει γενικεύοντας, με καλύτερη ικανότητα πρόβλεψης σε άγνωστα παραδείγματα. Ο αλγόριθμος μπορεί να εφαρμόζει προ-κλάδεμα (pre-pruning), δηλαδή να εντοπίζει και να κόβει κλαδιά κατά την ανάπτυξη ή μετά-κλάδεμα (post-pruning), δηλαδή να κόβει τα ακατάλληλα κλαδιά αφότου έχει ολοκληρωθεί η ανάπτυξη του δέντρου.

Οι σημαντικότερες επιλογές που καλείται να ορίσει ο αναλυτής, ώστε να επιτύχει την κατασκευή ενός αποτελεσματικού δέντρου είναι:

- Επιλογή του κατάλληλου τύπου δέντρου. Το δέντρο θα πρέπει να είναι κατάλληλο για τον τύπο δεδομένων που είναι διαθέσιμα π.χ. αν περιλαμβάνονται ποσοτικές μεταβλητές, αποκλείονται οι αλγόριθμοι που λειτουργούν μόνο με ονομαστικές, ενώ αν δεν υπάρχουν μεταβλητές με πολλές κατηγορίες, μπορεί να επιλεγεί ένα δυαδικό δέντρο, όπως το CART.
- Επιλογή του ειδικότερου αλγορίθμου που καθορίζει την εύρεση του αποτελεσματικότερου χαρακτηριστικού και τον τρόπο διάσπασης των κόμβων, καθώς και η ρύθμιση των σχετικών παραμέτρων. Τα κριτήρια μπορεί να είναι:
  - Κέρδος πληροφορίας (information gain). Επιλέγεται το χαρακτηριστικό που οδηγεί στη διάσπαση με τη μικρότερη εντροπία (η εντροπία είναι μέτρο της «αταξίας» των δεδομένων). Το κριτήριο αυτό έχει την τάση να ευνοεί τα χαρακτηριστικά που παίρνουν πολλές διαφορετικές τιμές.
  - Λόγος κέρδους πληροφορίας (gain ratio). Είναι παραλλαγή του κέρδους πληροφορίας που κανονικοποιεί την πληροφορία ανά χαρακτηριστικό ώστε να είναι αντιπροσωπευτικότερη η σύγκριση.
  - gini index. Είναι ένας δείκτης της μη καθαρότητας των δεδομένων και οδηγεί στην επιλογή του χαρακτηριστικού που επιτυγχάνει τη μεγαλύτερη βελτίωση στην καθαρότητα.
  - Ακρίβεια (accuracy). Επιλέγεται το χαρακτηριστικό που βελτιώνει περισσότερο τη συνολική ακρίβεια του δέντρου.
  - Απόσταση χ^2 (Chi-square). Είναι το κριτήριο που εφαρμόζεται στα δέντρα τύπου CHAID και είναι ιδιαίτερα κατάλληλο σε ονομαστικά δεδομένα.
- Επιλογή του μέγιστου βάθους. Ο περιορισμός του βάθους βοηθάει στη δημιουργία πιο συμπαγών και γενικών δέντρων, σε βάρος της συνολικής ακρίβειας.

- Επιλογή της ενεργοποίησης ή όχι κλαδέματος και του ορίου εμπιστοσύνης που θα χρησιμοποιηθεί στην εφαρμογή του κλαδέματος. Το κλάδεμα είναι σχεδόν πάντα επιθυμητό και απενεργοποιείται σε ειδικές περιπτώσεις.
- Ρύθμιση του ελάχιστου μεγέθους του φύλλου. Πολύ μικρά φύλλα δεν έχουν αξία και δεν πρέπει να επιτρέπονται.
- Ρύθμιση του ελάχιστου μεγέθους κόμβου προς διάσπαση. Κόμβοι που έχουν μικρότερο αριθμό παραδειγμάτων από το όριο αυτό, δεν διασπώνται περεταίρω, επειδή οδηγούν σε υπερ-προσαρμογή και μη αξιόπιστα υπο-δέντρα.

Μετά την κατασκευή του δέντρου, ακολουθεί η αξιολόγησή του, για να διαπιστωθεί το αν είναι αρκετά ακριβές ώστε να μην είναι παραπλανητικό (και επομένως άχρηστο), καθώς και αν τα ευρήματα που αναδεικνύει είναι χρήσιμα και αξιοποιήσιμα στην επίλυση του επιχειρηματικού προβλήματος. Η αξιολόγηση γίνεται με χρήση παραδειγμάτων που δε χρησιμοποιήθηκαν στην εκπαίδευση με σκοπό να ελεγχθεί η ικανότητα γενίκευσης του μοντέλου σε άγνωστα δεδομένα. Στο Κεφάλαιο 7 παρουσιάζονται παραδείγματα εφαρμογής των δέντρων αποφάσεων σε εκπαιδευτικά και πραγματικά προβλήματα με χρήση ειδικού λογισμικού. Στα παραδείγματα αυτά αναδεικνύονται στην πράξη οι δυνατότητες, οι περιορισμοί και οι επιλογές παραμετροποίησης των δέντρων αποφάσεων, καθώς και η αξία τους στην επίλυση μεγάλου εύρους προβλημάτων.

6.7 Συμπεράσματα

Υπάρχουν πολλές και διαφορετικές μέθοδοι και τεχνικές εξόρυξης γνώσης και ανάλυσης δεδομένων. Το βέβαιο είναι ότι καμία δεν μπορεί να δώσει λύση στα πολύπλοκα προβλήματα της πραγματικότητας χωρίς τη συνεργασία και το συνδυασμό περισσότερων από μίας τεχνικών. Οι διαφορετικές τεχνικές όμως δε χρησιμοποιούνται μόνο για να ταιριάσουν σε διαφορετικού τύπου υπολογισμούς και δεδομένα, αλλά συχνά αποτελούν εναλλακτικές προσεγγίσεις σε σύνθετα προβλήματα. Είναι ιδιαίτερα σημαντικό να μπορούμε να αποκτούμε μια σφαιρική άποψη των προβλημάτων και των προεκτάσεών τους, ώστε να ακολουθήσουμε, συχνά χωρίς εγγύηση για το αποτέλεσμα, την καταλληλότερη προσέγγιση . Είναι σημαντικό να μπορούμε να εμβαθύνουμε στα προβλήματα, να εντοπίζουμε τις ιδιαιτερότητές τους και να τεκμηριώνουμε τις λύσεις τους. Στο Κεφάλαιο αυτό, παρουσιάστηκε μια σειρά μεθόδων εξαγωγής γνώσης από δεδομένα, μέσα από ένα μεγάλο εύρος προσεγγίσεων, ώστε ο αναγνώστης να αποκτήσει μια σφαιρική εικόνα για το χώρο αυτό. Στο επόμενο κεφάλαιο, παρουσιάζονται σε πιο πρακτική μορφή, μια σειρά επιλεγμένων τεχνικών, μαζί με τα κατάλληλα εργαλεία για την αξιοποίησή τους σε πραγματικά προβλήματα.

Βιβλιογραφία/Αναφορές

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules in large databases*. In *20th VLDB* (pp. 487–499). Chile: Morgan Kaufmann.

Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications*, 42, 4632–4642

Doroz R., Porwik, P., & Orczyk, T. (2016). Dynamic signature verification method based on association of features with similarity measures. *Neurocomputing*, 171, 921-931.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6), 1437–1447.

Ho, C.C. (2012). Construct factor evaluation model of Health Management Center selected by customers with Fuzzy Analytic Hierarchy Process. *Expert Systems with Applications*, 39(1), 954-959.

Kardaras D., Mamakou X., and V. Karakostas (2011). *Adaptive web site design based on fuzzy user profiles, usability rules and design patterns*. Published in the proceedings of the 1st European Workshop on HCI Design and Evaluation: The influence of domains, Cyprus University of Technology, the

European University Cyprus in collaboration with SIGCHI Cyprus, 8 April, Limassol, Cyprus, Printed by: IRIT Press, Toulouse, France, ISBN: 978-2-917490-13-6.

Liu, H., & Setiono, R. (1995). Chi2: *Feature selection and discretization of numeric attributes*. In Proceedings of the 7th international conference on tools with artificial intelligence, ICTAI'95 (pp. 338–391).

Mani, V., Agrawal, R., & Sharma, V. (2014). Supplier selection using social sustainability: AHP based approach. *International Strategic Management Review*, 2(2), 98-112.

Ngai, E. W. T. (2003). Selection of web sites for online advertising using the AHP. *Information & Management*, 40(4), 233–242.

Ross, T. (2009). *Fuzzy Logic with Engineering Applications*, 3rd Edition, Wiley Publ., ISBN: 978-0-470-74376-8.

Saaty,T.L.(1977), A scaling method for priorities in hierarchical structures, *Journal of Mathematical Psychology.* 15(3): 234-281.

Saaty,T.L.(1980). *The Analytic Hierarchy Process*, New York: McGraw-Hill.

Simon, H.A. (1977). *The New Science of Management Decision* (3rd revised edition; first edition 1960) Prentice-Hall, Englewood Cliffs, NJ.

Subramanian, N., & Ramanathan, R. (2012). A review of applications of Analytic Hierarchy Process in operations management. *International Journal of Production Economics*, 138(2), 215-241.

Vaidya, O., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, 169(1), 1-29.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco: Morgan Kaufmann.

Wu, T., Chen, Y., Han, J. (2010). Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, *21*(3), 371–397.

Νανόπουλος Α., & Μανωλόπουλος Ι. (2008). *Εισαγωγή στην εξόρυξη και τις αποθήκες δεδομένων*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Κεφάλαιο 7. Εφαρμογές επιχειρηματικής ευφυΐας

Σύνοψη
Στο κεφάλαιο αυτό παρουσιάζονται μια σειρά από εφαρμογές στο χώρο της διοίκησης επιχειρήσεων και του μάρκετινγκ, συμπεριλαμβανομένων των δημοφιλέστερων προβλημάτων όπως η ανάλυση καλαθιού αγορών, πρόβλεψη ανταπόκρισης πελατών στις ενέργειες προώθησης, μελέτη και πρόβλεψη των προτιμήσεων και της ικανοποίησης των πελατών. Οι εφαρμογές πραγματοποιούνται στο λογισμικό εξόρυξης δεδομένων RapidMiner, το οποίο διατίθεται σε δωρεάν έκδοση και υποστηρίζεται από πλούσια τεκμηρίωση. Για κάθε εφαρμογή, παρουσιάζονται οι στόχοι και εξηγείται το σκεπτικό επιλογής των κατάλληλων μεθόδων. Στη συνέχεια παρουσιάζεται αναλυτικά η διαδικασία δημιουργίας του μοντέλου ανάλυσης με χρήση του γραφικού περιβάλλοντος του πακέτου, η διαδικασία εκτέλεσης της ανάλυσης σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου, καθώς και η διαδικασία λήψης και ερμηνείας των αποτελεσμάτων.

Προαπαιτούμενη γνώση
Κεφάλαιο 2. Δεδομένα και Πληροφορίες, Κεφάλαιο 6. Μέθοδοι εξόρυξης γνώσης από δεδομένα

7.1 Λογισμικό εξόρυξης γνώσης από δεδομένα

7.1.1 Σκοπός και διαδικασίες

Η εφαρμογή των μεθόδων που παρουσιάστηκαν στο Κεφάλαιο 6 γίνονται με χρήση κατάλληλου λογισμικού που, είτε διατίθεται ως ανεξάρτητη εφαρμογή, είτε αποτελεί ειδικό τμήμα ανάλυσης που επεκτείνει σε δυνατότητες επιχειρηματικής ευφυΐας κάποιο πακέτο λογισμικού για επιχειρήσεις. Επίσης, το λογισμικό αυτό μπορεί να αποτελεί μια πλατφόρμα που να διαθέτει εργαλεία και ένα περιβάλλον ανάπτυξης, πάνω στα οποία μπορεί ένας χρήστης να χτίσει τις εφαρμογές του, να τις τελειοποιήσει και να προχωρήσει σε μεγαλύτερο βάθος, αναπτύσσοντας δικές του τεχνικές. Στην κατηγορία αυτή ανήκει το RapidMiner που παρουσιάζεται στο κεφάλαιο αυτό, καθώς και το WEKA (WEKA, n.d.), που αποτελεί δημοφιλές πακέτο ελεύθερου λογισμικού. Το λογισμικό μπορεί όμως να έχει και τη μορφή μιας λύσης «με το κλειδί στο χέρι», που απευθύνεται αποκλειστικά στους τελικούς χρήστες, δηλαδή τα στελέχη επιχειρήσεων που δεν έχουν γνώσεις αναλυτή, αλλά επιθυμούν την άμεση πρόσβαση σε αποτελέσματα (όλες σχεδόν οι εταιρείες πληροφοριακών συστημάτων ERP, CRM και Βάσεων δεδομένων – μεγάλες και μικρές- διαθέτουν σήμερα κάποιες επεκτάσεις για business intelligence (Laudon, 2009). Τέλος, αναφέρουμε ότι υπάρχουν περιβάλλοντα ανάπτυξης «χαμηλού επιπέδου», που συγκαταλέγονται στις γλώσσες προγραμματισμού και όχι στο έτοιμο λογισμικό, αλλά ωστόσο είναι δημοφιλή εργαλεία εξόρυξης γνώσης για έμπειρους αναλυτές. Αναφέρονται ως παράδειγμα η γλώσσα στατιστικής ανάλυσης R (R, n.d.), η γλώσσα Python (Python, n.d.) και το μαθηματικό πακέτο Matlab (Matlab, n.d.).

Το λογισμικό εξόρυξης γνώσης μπορεί να συναντηθεί με μια πλειάδα ονομάτων, που το καθένα προσδιορίζει τους ιδιαίτερους στόχους στους οποίους τοποθετείται και τις μεθόδους στις οποίες δίνεται έμφαση. Συνηθισμένα ονόματα για κατηγορίες λογισμικού στον ευρύτερο χώρο είναι (Νανόπουλος, 2008):

- Εξόρυξη δεδομένων (ή πιο σωστά, εξόρυξη πληροφορίας/γνώσης από δεδομένα) – Data mining
- Ανακάλυψη γνώσης σε Βάσεις Δεδομένων – Knowledge Discovery in Databases (KDD)
- Αναλυτική επεξεργασία δεδομένων – Data Analytics
- Ανάλυση πρόβλεψης – Predictive Analytics
- Μηχανική μάθηση – Machine Learning
- Μεγάλα δεδομένα – Big Data
- Ανάλυση δεδομένων – Data analysis

Το όνομα κάθε κατηγορίας φανερώνει τις ιδιαιτερότητές της, όπως η λέξη εξόρυξη (mining) δείχνει την προσπάθεια να ανακαλύψουμε κάτι με αξία μέσα σε μεγάλους όγκους ανεξερεύνητων δεδομένων, ο όρος πρόβλεψη (predictive) δείχνει ότι δίνεται έμφαση σε μοντέλα πρόβλεψης και η ονομασία μεγάλα δεδομένα (big data) παραπέμπει στην πρόκληση του χειρισμού των τεράστιων όγκων αδόμητων δεδομένων που διακινούνται πλέον ηλεκτρονικά και είναι αδύνατον να αντιμετωπίσει ένα συμβατικό σύστημα ανάλυσης. Σημειώνεται ότι υπάρχουν επίσης άλλες πιο ειδικές κατηγορίες που αφορούν συγκεκριμένου τύπου δεδομένα, όπως η εξόρυξη από κείμενο (text mining) και η αναγνώριση προτύπων (pattern recognition), που αναφέρεται συνήθως σε σήματα, εικόνα ή πολυμέσα. Υπάρχει μεγάλη αλληλοεπικάλυψη ανάμεσα στις μεθόδους που επικρατούν σε κάθε κατηγορία και ίσως και κάποια σύγχυση όσον αφορά το διαχωρισμό κάποιων κατηγοριών από άλλες. Σημαντικό είναι επίσης να τονιστεί ότι υπάρχει μεγάλη συνάφεια ανάμεσα σε όλα τα παραπάνω και τη στατιστική ανάλυση, αφού μια μεγάλη κατηγορία μεθόδων βασίζονται στη στατιστική. Επομένως, πολλές πλατφόρμες data mining ενσωματώνουν εργαλεία στατιστικής ανάλυσης και πολλά πακέτα στατιστικής ανάλυσης χρησιμοποιούνται ευρέως για εξόρυξη γνώσης από δεδομένα.

Στο βιβλίο αυτό εστιάζουμε σε ένα βασικό σύνολο μεθόδων και εργαλείων «γενικής χρήσης», κατάλληλων για την υποστήριξη της επιχειρηματικής ευφυΐας μιας επιχείρησης. Θα χρησιμοποιούμε τον όρο εξαγωγή γνώσης από δεδομένα, ο οποίος είναι συναφής αλλά ελαφρά πιο γενικός από την εξόρυξη δεδομένων (Data mining). Στη πραγματικότητα οι δύο αυτοί όροι απεικονίζουν εξίσου καλά τις μεθόδους και τεχνικές που παρουσιάζονται σε αυτό το κεφάλαιο.

Για τις ανάγκες του βιβλίου αυτού, επιλέχθηκε η χρήση του RapidMiner Studio, της εταιρείας RapidMiner GmbH (http://www.rapidminer.com) (Rapidminer, n.d.), το οποίο καλύπτει με εξαιρετικό τρόπο τις βασικές ανάγκες ανάπτυξης διαδικασιών ανάλυσης, διαθέτει ιδιαίτερα φιλικό και εύχρηστο γραφικό περιβάλλον και όλα τα απαραίτητα βοηθητικά εργαλεία για την εισαγωγή, προεπισκόπηση και προεπεξεργασία των δεδομένων, καθώς και την παρουσίαση των αποτελεσμάτων. Η εταιρεία που υποστηρίζει το RapidMiner προέρχεται από μετεξέλιξη ερευνητικής κοινοπραξίας και το ίδιο το πρόγραμμα αποτελεί τελειοποίηση αποτελέσματος ερευνητικού έργου. Το RapidMiner διατίθεται στην έκδοση Basic εντελώς δωρεάν, με αξιοπρεπή όμως υποστήριξη από την εταιρεία και ευρεία κοινότητα χρηστών, ενώ διατίθεται επίσης σε ακαδημαϊκή και σε πλήρως εμπορική έκδοση. Στο βιβλίο αυτό, χρησιμοποιήθηκε η έκδοση RapidMiner Studio 6.5 με παραχωρημένη ακαδημαϊκή άδεια, ενώ σημειώνεται ότι όλα τα στοιχεία που παρουσιάζονται είναι διαθέσιμα και στην έκδοση Basic.

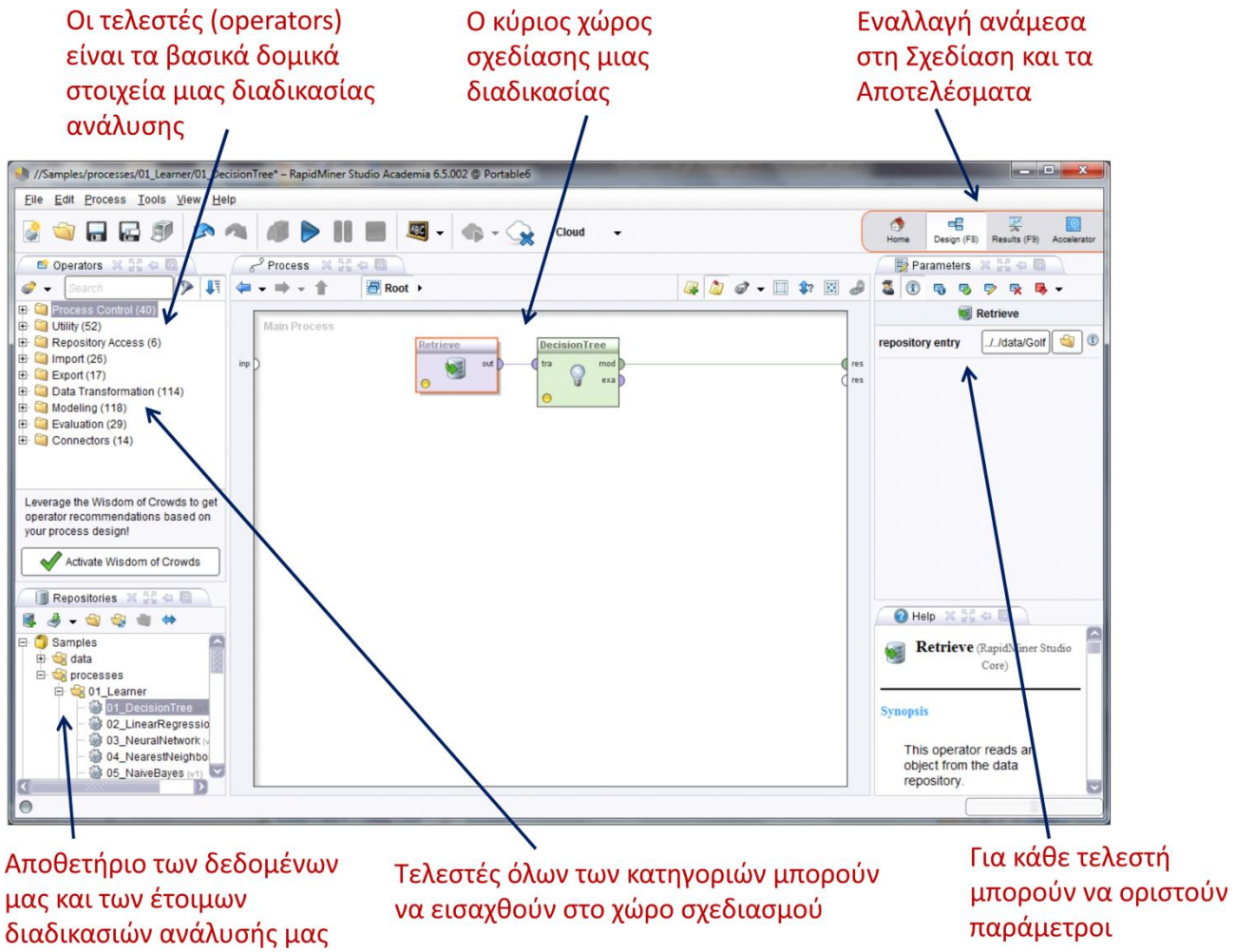
7.1.2 Το περιβάλλον του RapidMiner

Το RapidMiner προφέρει ένα ολοκληρωμένο και εύχρηστο γραφικό περιβάλλον, μέσα από το οποίο ο χρήστης μπορεί να εισάγει και να επισκοπήσει τα δεδομένα του, να εκτελέσει απλές αναλύσεις ή να συνθέσει περισσότερο περίπλοκες και να περιηγηθεί στα αποτελέσματα. Το περιβάλλον του προγράμματος παρουσιάζεται στο Σχήμα 7.1.

Βασικές οπτικές (perspectives) του χώρου εργασίας είναι η προβολή **Σχεδίασης (Design)** και η προβολή **Αποτελεσμάτων (Results)**. Η Σχεδίαση είναι ο χώρος «δημιουργίας», όπου καθορίζονται και παραμετροποιούνται τα βήματα της ανάλυσης, τα δεδομένα εισόδου και η επιθυμητή έξοδος. Η προβολή Αποτελεσμάτων δεν περιορίζεται στην παράθεση των αποτελεσμάτων με τη μορφή π.χ. ενός πίνακα και μιας αναφοράς, αλλά διαθέτει δυνατότητες «εξερεύνησης» των αποτελεσμάτων και προβολής παραμετροποιήσιμων γραφικών, διαγραμμάτων και πινάκων.

Βασική έννοια στο RapidMiner είναι η **Διαδικασία (Process)**. Είναι μια αυτοτελής διαδικασία ανάλυσης, που έχει κάποιες εισόδους δεδομένων, κάποιες εξόδους και, ενδιάμεσα, διαδοχικά βήματα επεξεργασίας και ανάλυσης δεδομένων. Μια διαδικασία αντιστοιχεί σε μια εφαρμογή και αποτελεί τη λύση ενός επιμέρους προβλήματος. Κάθε βήμα της διαδικασίας παριστάνεται με ένα γραφικό αντικείμενο που ονομάζεται **Τελεστής (Operator)**, το οποίο συνδέεται με άλλους τελεστές με τη βοήθεια μιας γραμμής σύνδεσης.

Όλες οι λειτουργίες διαχείρισης, προβολής, επεξεργασίας και ανάλυσης εκτελούνται μέσω τελεστών, οι οποίοι μπορούν να δεχθούν είσοδο από κάποιον άλλο τελεστή ή να έχουν πρόσβαση στο εξωτερικό περιβάλλον (π.χ. ανάγνωση αρχείου) και να τροφοδοτήσουν άλλους τελεστές ή να παρέχουν την τελική έξοδο της διαδικασίας. Οι τελεστές είναι παραμετροποιήσιμοι, έτσι ώστε να μπορεί ο χρήστης να τους προσαρμόσει στο πρόβλημα και να βελτιστοποιήσει την απόδοσή τους.



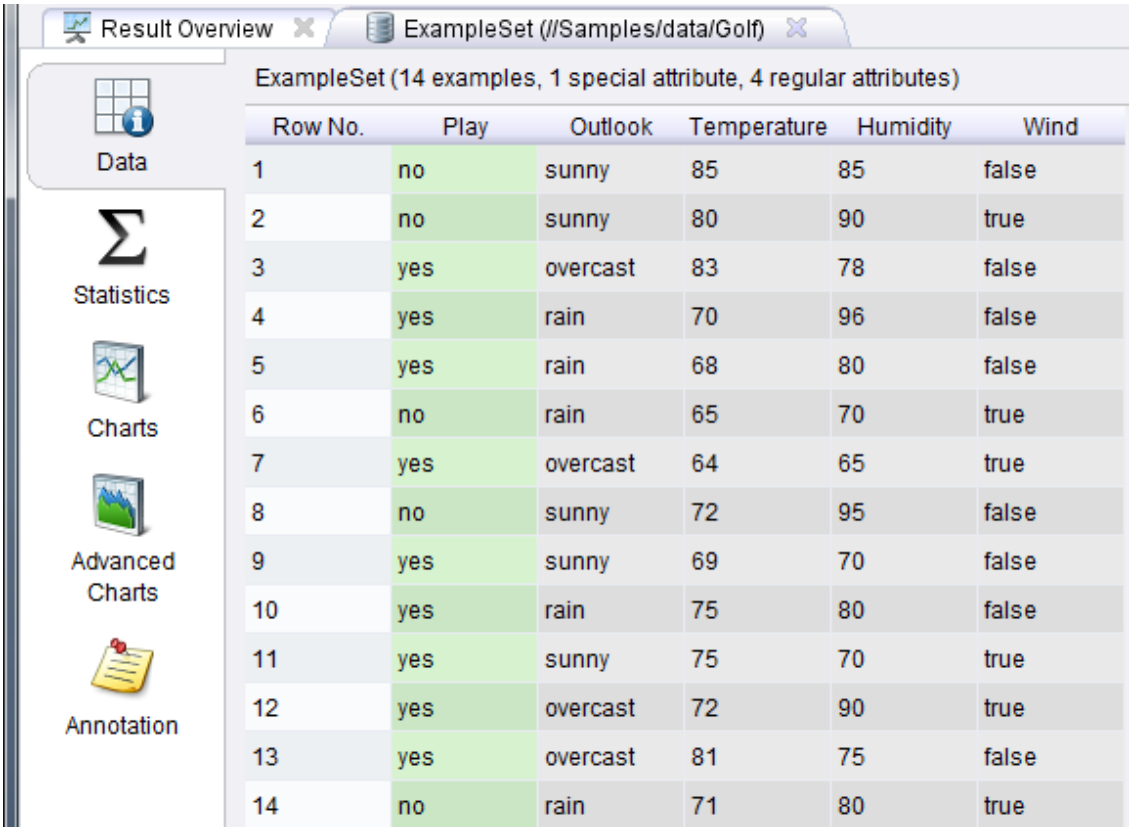
Σχήμα 7.1. *Το περιβάλλον εργασίας του RapidMiner σε προβολή Σχεδίασης*

Τα πρώτα βήματα που μπορεί να κάνει ένας αναγνώστης, ακόμα και χωρίς καμία εμπειρία, ώστε να γνωριστεί με το πρόγραμμα είναι τα εξής:

Άνοιγμα του προγράμματος: Αφού εγκατασταθεί το πρόγραμμα (οδηγίες περιλαμβάνονται στον οδηγό εγκατάστασης http://docs.rapidminer.com/studio/installation/), το ανοίγουμε σε προβολή Σχεδίασης (Design perspective) επιλέγοντας **Design**.

Επισκόπηση δεδομένων: Η πρώτη σημαντική ενέργεια, πριν ξεκινήσει η διαδικασία της ανάλυσης, είναι η επισκόπηση των αρχικών δεδομένων. Στο πρώτο αυτό παράδειγμα, θα ανοίξουμε ένα από τα σετ δεδομένων που υπάρχουν στην εγκατάσταση του προγράμματος ως δείγματα (Samples) στο τοπικό αποθετήριο (Local repository). Το σετ δεδομένων με όνομα **Golf** αφορά μετεωρολογικά στοιχεία, με βάση τα οποία θέλουμε να προβλέψουμε το αν μπορεί να διεξαχθεί ή όχι ένας αγώνας golf. Για την επισκόπηση των δεδομένων, ανοίγουμε το φάκελο **Data** και επιλέγουμε με διπλό κλικ το σετ με όνομα **Golf**. Το RapidMiner ανοίγει μια νέα καρτέλα, όπου προβάλει τα δεδομένα σε μορφή πίνακα (Σχήμα 7.2). Η δομή του πίνακα είναι εμφανής: κάθε στήλη του πίνακα είναι ένα χαρακτηριστικό (attribute) και κάθε γραμμή μια περίπτωση (transaction, case ή example). Σημειώνεται ότι για τις περιπτώσεις (γραμμές του πίνακα), στο RapidMiner χρησιμοποιείται ο όρος **παράδειγμα (example)** και το σετ δεδομένων που προορίζεται για εξαγωγή γνώσης ονομάζεται **example set**. Όταν σκοπεύουμε να χρησιμοποιήσουμε τα δεδομένα για εκπαιδευόμενη εκμάθηση, ένα από τα χαρακτηριστικά πρέπει να δηλωθεί ως **χαρακτηριστικό-στόχος (target attribute)** που στο RapidMiner ονομάζεται **Label**. Σημειώνεται ότι τα χαρακτηριστικά είναι αντίστοιχα με τα πεδία των

πινάκων των βάσεων δεδομένων και τα παραδείγματα είναι το αντίστοιχο των εγγραφών (βλέπε Κεφάλαιο 2, παρ. 3.4.2).



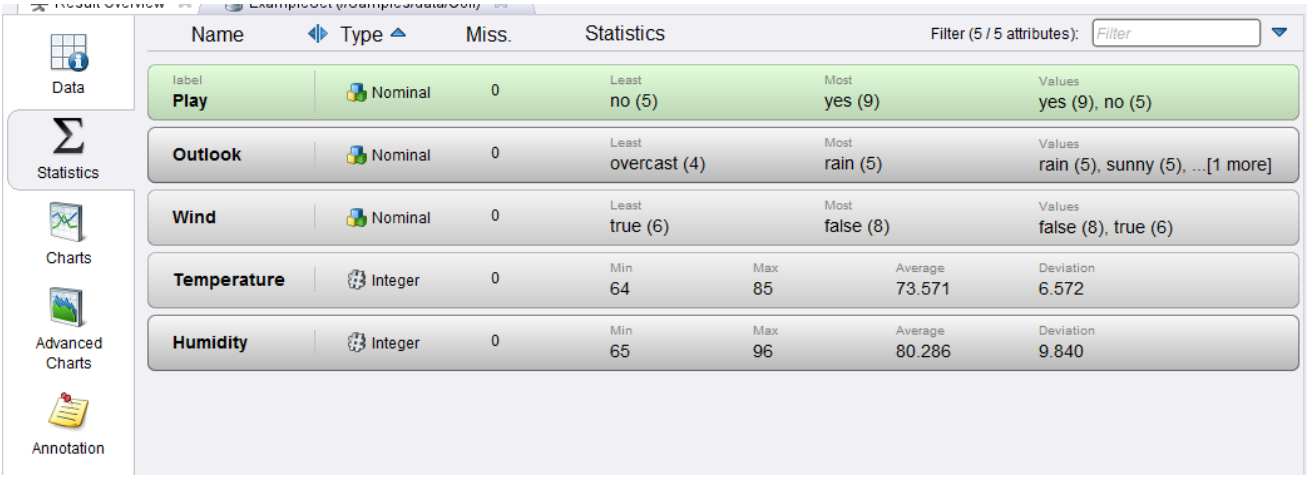
ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

Σχήμα 7.2. *Επισκόπηση δεδομένων σε μορφή πίνακα*

Παρατηρούμε ότι στο σετ δεδομένων περιέχονται 14 παραδείγματα, για τα οποία έχουν συλλεγεί τα χαρακτηριστικά: **Play**, που παίρνει τις τιμές *no, yes* (διεξαγωγή ή όχι ενός παιχνιδιού golf), **Outlook**, που παίρνει τις τιμές *sunny, overcast, rain* (πρόβλεψη καιρού), **Temperature**, που παίρνει αριθμητικές τιμές (θερμοκρασία), **Humidity**, που παίρνει αριθμητικές τιμές 0-100 (υγρασία) και **Wind**, που παίρνει τιμές *false, true* (αν υπάρχει ισχυρός άνεμος ή όχι). Το χαρακτηριστικό **Play** έχει οριστεί ως label, δηλαδή αποτελεί το χαρακτηριστικό που θέλουμε να μάθουμε να προβλέπουμε με βάση τις τιμές των υπολοίπων.

Μεταβαίνοντας στην υποκαρτέλα **Statistics** (Σχήμα 7.2), μπορούμε να δούμε τον τύπο δεδομένων κάθε χαρακτηριστικού και τα βασικά περιγραφικά του στατιστικά. Για τα ονομαστικά χαρακτηριστικά δίνονται οι συχνότητες των τιμών, ενώ για τα ποσοτικά χαρακτηριστικά δίνονται το εύρος τιμών, η μέση τιμή και η τυπική απόκλιση , π.χ. το **Play** παίρνει ονομαστικές τιμές, από τις οποίες η λιγότερο συχνή είναι η *no*, που εμφανίζεται 5 φορές, και συχνότερη η *yes*, που εμφανίζεται 9 φορές. Στις υποκαρτέλες **Charts** και **Advanced Charts** μπορούμε να δούμε τα δεδομένα με τη βοήθεια απλών ή σύνθετων γραφημάτων όπως ιστογράμματα, scatter plots, ραβδογράμματα, κ.ά.



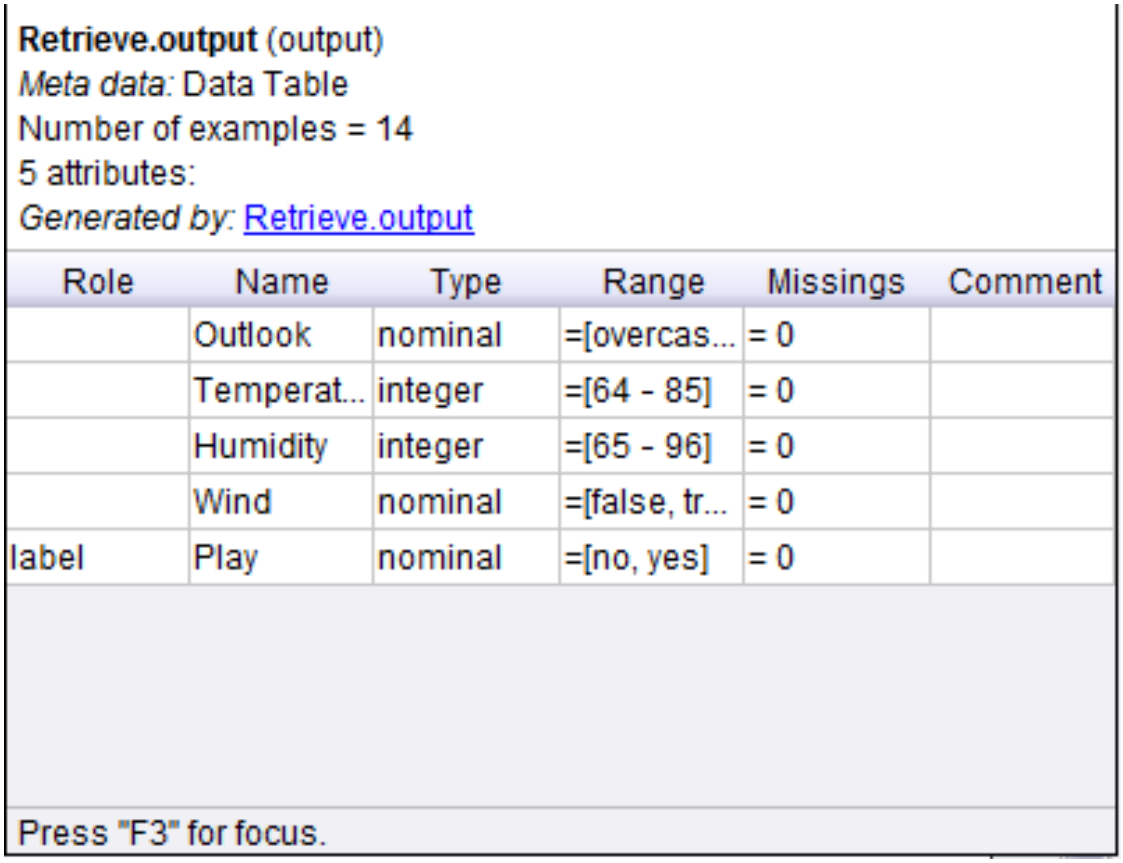
Σχήμα 7.3. *Επισκόπηση στατιστικών στοιχείων για όλα τα χαρακτηριστικά.*

Σύνθεση διαδικασίας: Για να δημιουργήσουμε μια διαδικασία, σύρουμε στο χώρο σχεδίασης **Main process** τους τελεστές που χρειαζόμαστε και τους συνδέουμε σύροντας κατάλληλες γραμμές σύνδεσης. Στην υποτυπώδη (αλλά απολύτως έγκυρη) διαδικασία που εικονίζεται στο Σχήμα 7.1, ο τελεστής **Retrieve** χρησιμοποιείται για την ανάγνωση ενός σετ δεδομένων από το Αποθετήριο. Στο πάνω δεξιά μέρος του σχήματος μπορούμε να διακρίνουμε ότι η μοναδική παράμετρος που έχει οριστεί για τον τελεστή αυτόν είναι το όνομα του σετ δεδομένων που θα αναγνωστεί. Η έξοδος του **Retrieve** οδηγείται στον τελεστή **Decision Tree**, που αποτελεί κλασικό αλγόριθμο εκμάθησης μοντέλου κατάταξης, μέσω δημιουργίας δέντρου αποφάσεων, ικανού να χειριστεί τόσο ποσοτικά, όσο και ονομαστικά χαρακτηριστικά. Η έξοδος του **Decision Tree** οδηγείται στη θύρα **res**, που συμβολίζει την έξοδο της διαδικασίας προς το εξωτερικό περιβάλλον.

Κάθε τελεστής μπορεί να απαιτεί μια ή περισσότερες εισόδους, που θα πρέπει να είναι συμβατές με τις προδιαγραφές του, και παρέχει κάποια έξοδο, της οποίας η δομή εξαρτάται από τον τελεστή. Στο παράδειγμα του σχήματος, ο τελεστής **Retrieve** δεν απαιτεί είσοδο από άλλον τελεστή επειδή διαβάζει τα δεδομένα εισόδου από το αποθετήριο. Η έξοδός του κατευθύνεται προς τον **Decision Tree**, ο οποίος, κατά την εκτέλεσή του, μαθαίνει από τα δεδομένα και παρέχει στην έξοδο της διαδικασίας το τελικό μοντέλο που έχει αναπτυχθεί.

Σημαντική παρατήρηση: Τα πραγματικά δεδομένα θα διαβαστούν από το αποθετήριο (ή οποιαδήποτε άλλη πηγή δεδομένων) και θα περάσουν από τελεστή σε τελεστή, αφού μετασχηματιστούν κατάλληλα από τον καθέναν από αυτούς, καταλήγοντας στο τελικό αποτέλεσμα, μόνο όταν ενεργοποιήσουμε την **Εκτέλεση** της διαδικασίας. (Σε μεγάλους όγκους δεδομένων, η εκτέλεση μπορεί να διαρκέσει μεγάλα χρονικά διαστήματα). Το RapidMiner όμως, διαθέτει μια ισχυρή δυνατότητα: η περιγραφή των δεδομένων (για την οποία ο ακριβής όρος είναι μετα-δεδομένα) είναι διαθέσιμη στην προβολή Σχεδίασης πριν γίνει η εκτέλεση, και μάλιστα ο χρήστης μπορεί να δει την εξέλιξή της σε όλα τα στάδια της διαδικασίας.

Τοποθετώντας το δείκτη του ποντικιού στην έξοδο του **Retrieve**, μπορούμε να δούμε τη δομή των δεδομένων που θα διαβαστούν από το αποθετήριο και θα περάσουν στον τελεστή εκμάθησης (Σχήμα 7.4). Στο παράδειγμα του σχήματος, βλέπουμε στα μετα-δεδομένα ότι στο σετ δεδομένων που θα διαβαστεί περιέχονται 14 παραδείγματα (δηλ. εγγραφές) και περιλαμβάνονται 5 χαρακτηριστικά (attributes) (δηλ. πεδία), των οποίων δίνονται οι περιγραφές (ρόλος, όνομα, τύπος, κλίμακα, ελλιπή στοιχεία, σχόλια).



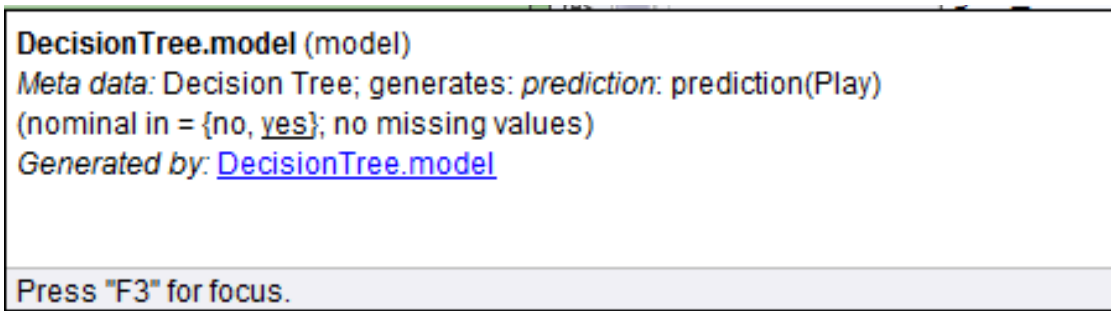
Retrieve.output (output)
Meta data: Data Table
Number of examples = 14
5 attributes:
Generated by: Retrieve.output

Role	Name	Type	Range	Missings	Comment
	Outlook	nominal	=[overcas...	= 0	
	Temperat...	integer	=[64 - 85]	= 0	
	Humidity	integer	=[65 - 96]	= 0	
	Wind	nominal	=[false, tr...	= 0	
label	Play	nominal	=[no, yes]	= 0	

Press "F3" for focus.

Σχήμα 7.4. *Τα μετα-δεδομένα που μας πληροφορούν για τη δομή του σετ δεδομένων πριν διαβαστούν τα ίδια τα δεδομένα.*

Τοποθετώντας το δείκτη του ποντικιού στην έξοδο του **Decision Tree**, βλέπουμε ότι, στο σημείο αυτό, τα δεδομένα έχουν μετασχηματιστεί σε ένα δέντρο αποφάσεων που προβλέπει το χαρακτηριστικό Play (Σχήμα 7.5).



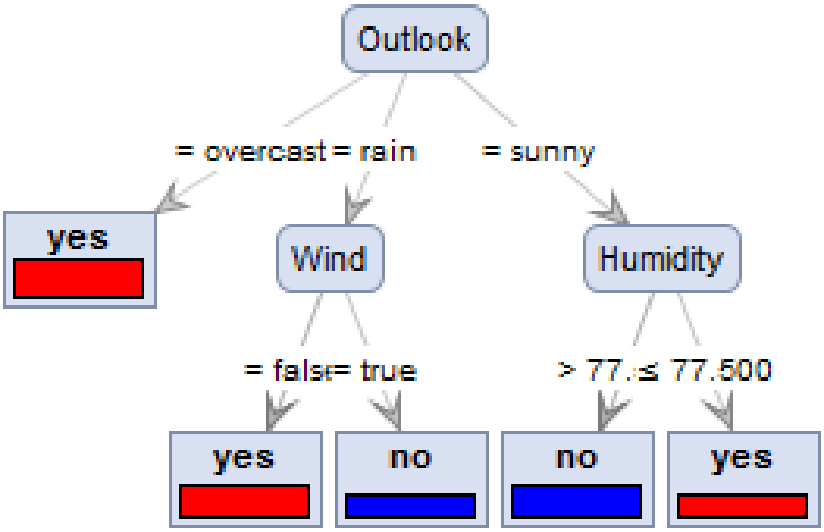
DecisionTree.model (model)
Meta data: Decision Tree; generates: *prediction*: prediction(Play)
(nominal in = {no, yes}; no missing values)
Generated by: DecisionTree.model

Press "F3" for focus.

Σχήμα 7.5. *Τα μετα-δεδομένα στην έξοδο του Decision Tree μας πληροφορούν ότι τα δεδομένα εισόδου έχουν μετασχηματιστεί σε ένα μοντέλο πρόβλεψης τύπου δέντρου αποφάσεων.*

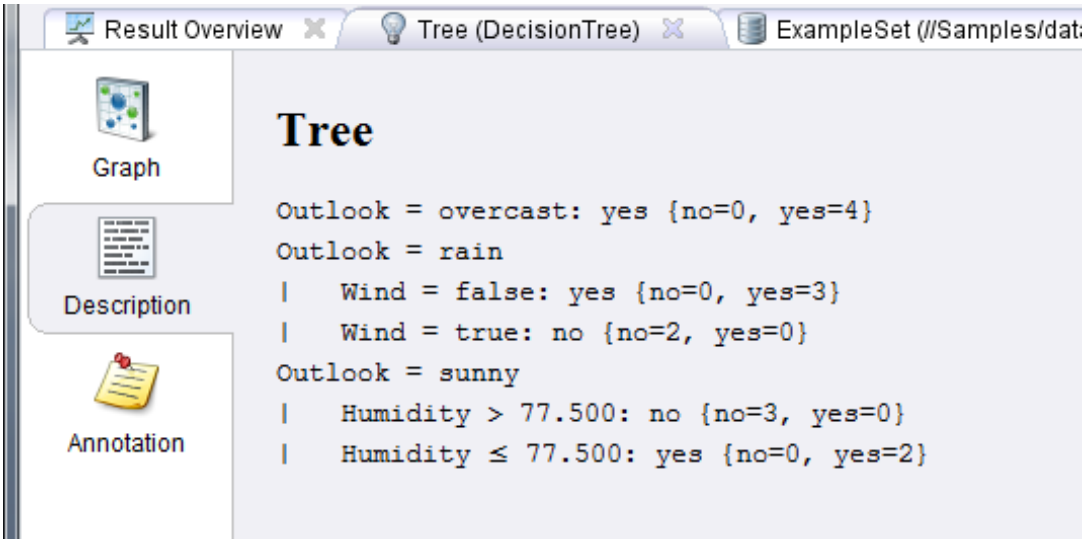
Αποτέλεσμα: Κάνοντας κλικ στο κουμπί της εκτέλεσης (Run or resume the current process) ή πιέζοντας F11, το RapidMiner εκτελεί τη διαδικασία και προβάλει τα αποτελέσματα σε μια νέα καρτέλα με το όνομα της διαδικασίας, μεταβαίνοντας σε προβολή Αποτελεσμάτων (Results perspective). Η καρτέλα των αποτελεσμάτων περιλαμβάνει 3 υποκαρτέλες: **Γράφημα (Graph)**, **Περιγραφή (Description)** και **Σημείωση (Annotation)**. Στο Σχήμα 7.6. φαίνεται το αποτέλεσμα της διαδικασίας σε μορφή γραφήματος. Δημιουργήθηκε ένα δέντρο αποφάσεων για την πρόβλεψη του χαρακτηριστικού **Play** από τα χαρακτηριστικά **Outlook**, **Wind** και **Humidity**. Το γράφημα βοηθάει το χρήστη να κατανοήσει το αποτέλεσμα, να εκτιμήσει

την επιτυχία της εκμάθησης και να ερμηνεύσει το φαινόμενο. Π.χ. στο σχήμα του συγκεκριμένου παραδείγματος, φαίνεται ότι η καλύτερη πρόβλεψη για τη διεξαγωγή του παιχνιδιού γίνεται διαχωρίζοντας τις περιπτώσεις με βάση τη μεταβλητή **Outlook**. Αν η τιμή της είναι *overcast* (συννεφιά), προβλέπεται με βεβαιότητα **Play**=*yes*. Αν η πρόβλεψη είναι βροχή, τότε, αν δεν υπάρχει άνεμος, προβλέπεται ότι θα διεξαχθεί το παιχνίδι αλλά, αν υπάρχει άνεμος, προβλέπεται ότι δε θα διεξαχθεί.



Σχήμα 7.6*. Το εκπαιδευμένο δέντρο αποφάσεων που προκύπτει ως αποτέλεσμα της διαδικασίας.*

Στην καρτέλα **Περιγραφή (Description)**, εμφανίζεται το ίδιο αποτέλεσμα σε μορφή κειμένου (Σχήμα 7.7). Πατώντας **Design**, μπορούμε να μεταβούμε και πάλι στην προβολή Σχεδίασης, για την πιθανή τροποποίηση της διαδικασίας, αποθήκευση ή κλείσιμο.



Σχήμα 7.7*. Το μοντέλο πρόβλεψης σε μορφή κειμένου.*

7.1.3 Οι διαθέσιμοι τελεστές

Όλες οι ενέργειες που μπορεί να περιλαμβάνονται σε μια διαδικασία ανάλυσης ή μοντελοποίησης διατίθενται με τη μορφή τελεστών (operators), ώστε να μπορεί ο χρήστης με εύχρηστο τρόπο να συνθέσει οποιαδήποτε

διαδικασία. Το RapidMiner Studio διαθέτει περισσότερους από 400 τελεστές, οι οποίοι ανήκουν στις ακόλουθες κατηγορίες:

8. **Πρόσβαση στο αποθετήριο (Repository access)**. Ανάγνωση/εγγραφή δεδομένων στο αποθετήριο και διαχείριση εγγραφών.
9. **Εισαγωγή (Import)**. Εισαγωγή από εξωτερικά αρχεία ή Βάσεις Δεδομένων διαφόρων τύπων δεδομένων, μοντέλων, παραμέτρων, κλπ.
10. **Εξαγωγή (Export)**. Εξαγωγή σε εξωτερικά αρχεία ή Βάσεις Δεδομένων.
11. **Μετασχηματισμός δεδομένων (Data transformation)**. Πολλές και σημαντικές δυνατότητες προσαρμογής και προεπεξεργασίας δεδομένων, όπως φιλτράρισμα, μετατροπή τύπων, υπολογισμός συγκεντρωτικών στοιχείων, ταξινόμηση και τροποποίηση τιμών.
12. **Μοντελοποίηση (Modeling)**. Οι κύριες μέθοδοι ανάλυσης και εξαγωγής γνώσης (μοντελοποίησης/εκμάθησης). Περιλαμβάνονται οι σημαντικότερες κατηγορίες μεθόδων, όπως ταξινόμησης, συσταδοποίησης/τμηματοποίησης, παλινδρόμησης, εύρεσης συσχέτισης και συχνών συνόλων, νευρωνικών δικτύων, κ.ά.
13. **Έλεγχος διαδικασίας (Process control)**. Περιλαμβάνει λειτουργίες ελέγχου ροής της διαδικασίας, όπως διακλαδώσεις και βρόγχους, δυνατότητες συνένωσης ροών δεδομένων, προσωρινής αποθήκευσης/ανάκλησης ενδιάμεσων αποτελεσμάτων, κ.ά.
14. **Αξιολόγηση (Evaluation)**. Λειτουργίες μέτρησης αποτελεσματικότητας και ακρίβειας των μοντέλων, με βάση πληθώρα μεθόδων και δεικτών.
15. **Σύνδεσμοι (Connectors)**. Δυνατότητες σύνδεσης μέσω διαδικτύου με εξωτερικές αποθήκες δεδομένων όπως Dropbox και Twitter.
16. **Βοηθητικά εργαλεία (Utilities)**. Μακροεντολές, έλεγχος εξωτερικών αρχείων, υποσημειώσεις, εκτέλεση εξωτερικών διαδικασιών, κ.ά.

Μια απλή τυπική διαδικασία περιλαμβάνει τουλάχιστον τα ακόλουθα:

- έναν τελεστή για την εισαγωγή των δεδομένων εισόδου
- έναν ή περισσότερους τελεστές για την προετοιμασία των δεδομένων π.χ. επιλογή των χαρακτηριστικών που μας ενδιαφέρουν, τροποποίηση κάποιων τύπων δεδομένων ώστε να είναι συμβατοί με την ανάλυση που θα ακολουθήσει, φιλτράρισμα εγγραφών με ελλιπή στοιχεία, κ.ά.
- τον κύριο τελεστή μοντελοποίησης, που θα μάθει από τα δεδομένα και θα εξάγει τα στοιχεία που μας ενδιαφέρουν.

7.2 Η συνολική διαδικασία εφαρμογής της εξόρυξης γνώσης από δεδομένα

Η εξαγωγή γνώσης από δεδομένα, ως στοιχείο επιχειρηματικής ευφυΐας για την ενίσχυση των λειτουργιών μιας επιχείρησης, πρέπει να εντάσσεται σε με μια συνολική διαδικασία, που περιλαμβάνει τόσο τεχνικά όσο και οργανωτικά βήματα. Από την πλευρά του στελέχους, που χειρίζεται τα επιχειρηματικά δεδομένα και πραγματοποιεί την ανάλυση ή εξόρυξη, δεν αρκεί η καλή γνώση χειρισμού των προγραμμάτων, ούτε ακόμα και η καλή κατανόηση των μεθόδων. Η συνολική διαδικασία που πρέπει να ακολουθείται για την πραγματοποίηση μιας επιτυχημένης εφαρμογής βασισμένης σε δεδομένα έχει τα χαρακτηριστικά ενός έργου (project), δηλαδή οφείλει να έχει στόχους, προγραμματισμένα βήματα και να ακολουθεί κάποια πρότυπα οργάνωσης, δηλαδή να ακολουθεί κάποιο μοντέλο. Έχουν προταθεί διάφορα μοντέλα διαδικασιών για εξαγωγή γνώσης από δεδομένα, τόσο στον ακαδημαϊκό χώρο, όσο και στη βιομηχανία. Ένα από τα επικρατέστερα τέτοια μοντέλα είναι το CRISP-DM (CRoss-Industry Standard Process for Data Mining) (CRISP, 2015), που αναπτύχθηκε και υποστηρίζεται από μια μεγάλη κοινοπραξία ευρωπαϊκών εταιρειών και αποτελεί το δημοφιλέστερο βιομηχανικό μοντέλο.

Το μοντέλο εξαγωγής γνώσης CRISP-DM αποτελείται από τα παρακάτω έξι βήματα:

1. **Κατανόηση του επιχειρηματικού προβλήματος (Business understanding)**. Το πρώτο βήμα εστιάζει στην κατανόηση των στόχων και των αναγκών από την οπτική γωνία της

επιχείρησης. Τα στοιχεία αυτά μετατρέπονται σε στόχους και ανάγκες του έργου εξαγωγής γνώσης και οδηγούν σε έναν πρώτο σχεδιασμό για τις μεθόδους που θα ακολουθηθούν.

2. **Κατανόηση των δεδομένων (Data understanding).** Το βήμα αυτό ξεκινάει από τη συλλογή δεδομένων και περιλαμβάνει την εξοικείωση με τα χαρακτηριστικά των δεδομένων, τον προσδιορισμό της ποιότητας και των ενδεχόμενων προβλημάτων τους. Οι έμπειροι ερευνητές του χώρου τονίζουν τη σπουδαιότητα αυτού του βήματος και συστήνουν την αφιέρωση μεγάλου χρόνου και προσπάθειας σε αυτό. Βασικά θέματα που πρέπει να εξεταστούν είναι το αν τα διαθέσιμα δεδομένα είναι αρκετά για την επίλυση του προβλήματος, αν εκπροσωπούν επαρκώς όλο το εύρος των περιπτώσεων που θέλουμε να μελετήσουμε και αν εμπεριέχουν σφάλματα.
3. **Προετοιμασία δεδομένων (Data preparation)**. Το βήμα αυτό έχει ως σκοπό τη μετατροπή του αρχικού σετ δεδομένων σε μορφή κατάλληλη να τροφοδοτηθεί στην κύρια μονάδα ανάλυσης/εκμάθησης. Τυπικές ενέργειες που περιλαμβάνονται είναι η επιλογή χαρακτηριστικών, η τροποποίηση του τύπου κάποιων χαρακτηριστικών (π.χ. από ποσοτικό σε ονομαστικό χωρίζοντας σε τάξεις), η σύνθεση νέων χαρακτηριστικών και το φιλτράρισμα παραδειγμάτων (π.χ. η απόρριψη από το σετ δεδομένων εκμάθησης όλων των συναλλαγών που αφορούν επιστροφές).
4. **Μοντελοποίηση (Modeling)**. Αφορά την κύρια λειτουργία κατασκευής του μοντέλου γνώσης από τα δεδομένα. Συχνά απαιτεί την εφαρμογή περισσότερων από μία μεθόδων ή την εκτέλεση επαναληπτικών διαδικασιών (δηλαδή την εκτέλεση, εκτίμηση του λάθους, διόρθωση παραμέτρων και επανεκτέλεση ώστε να συγκλίνουμε προς τη λύση). Εκτός από την επιλογή των μεθόδων, απαιτείται και η ρύθμιση των παραμέτρων τους ώστε να προσαρμοστούν στο πρόβλημα και να βελτιστοποιηθούν τα αποτελέσματα. Τα επιμέρους βήματα είναι (α) η επιλογή των τεχνικών, (β) κατασκευή του δείγματος δεδομένων εκμάθησης, (γ) δημιουργία του μοντέλου και (δ) εκτίμηση της ορθότητας του προκύπτοντος μοντέλου.
5. **Αξιολόγηση (Evaluation)**. Μετά την κατασκευή ενός μοντέλου που φαίνεται ακριβές από τη σκοπιά της ανάλυσης δεδομένων, το αποτέλεσμα αξιολογείται και από τη σκοπιά των επιχειρηματικών στόχων. Σκοπός του βήματος αυτού είναι να καθοριστεί η χρησιμότητα της εξαχθείσας γνώσης και η ικανότητά της να λύνει πραγματικά προβλήματα. Ένα μοντέλο πρόβλεψης μπορεί να περιλαμβάνει κανόνες που προβλέπουν το προφανές ή που είναι αβέβαιοι και, επομένως, να πρέπει να απορριφθεί, είτε στο σύνολό του, είτε κατά ένα μέρος του περιεχομένου του. Ένα αποτέλεσμα όπως π.χ. ότι ένας πελάτης καταστήματος καλλυντικών που είναι γυναίκα και έχει ηλικία μεγαλύτερη των 20 ετών έχει μεγάλη πιθανότητα να αγοράσει κραγιόν, αποτελεί κάτι προφανές για ένα στέλεχος μάρκετινγκ και ένα μοντέλο που επιδίδεται μόνο σε τέτοιου είδους προβλέψεις δεν έχει καμία αξία. Πρέπει να τονιστεί ότι είναι φυσικό κάποια προβλήματα να μη λύνονται ικανοποιητικά, επειδή, απλούστατα, τα δεδομένα μπορεί να μην περιέχουν κάποια χρήσιμη γνώση σχετική με αυτά που μας ενδιαφέρουν ή το φαινόμενο που προσπαθούμε να εξηγήσουμε με τη βοήθεια κανόνων να είναι από τη φύση του απρόβλεπτο.
6. **Ανάπτυξη και Εφαρμογή (Deployment)**. Η γνώση που καταφέραμε να εξάγουμε πρέπει να οργανωθεί σε μορφή κατάλληλη για την αξιοποίησή της στην επιχείρηση. Η απαίτηση αυτή μπορεί στην απλούστερη περίπτωση να αφορά τη δημιουργία μας αναφοράς ή φόρμας προβολής αποτελεσμάτων. Σε κάποιες περιπτώσεις απαιτείται το εκπαιδευμένο μοντέλο να είναι διαθέσιμο σε λειτουργική μορφή, έτσι ώστε να μπορούμε να το τροφοδοτούμε με δεδομένα εισόδου και, μετά από εκτέλεση κάποιων λειτουργιών, να παίρνουμε ένα αποτέλεσμα, π.χ. για να αξιοποιηθεί ένα μοντέλο πρόβλεψης της πιστότητας των πελατών, πρέπει να μπορούμε να εισάγουμε τα στοιχεία ενός πελάτη και κάποιος μηχανισμός να μας επιστρέφει το αποτέλεσμα πρόβλεψης του μοντέλου για τον πελάτη αυτόν.

Τα παραπάνω βήματα δίνουν μια γενική εικόνα για τη συνολική διαδικασία μιας εφαρμογής επιχειρηματικής ευφυΐας με εξαγωγή γνώσης. Έχοντας γνωρίσει τις βασικές μεθόδους στο Κεφάλαιο 6, καθώς και το περιβάλλον του λογισμικού RapidMiner στο τρέχον κεφάλαιο, παρουσιάζουμε στη συνέχεια σε ποιο πρακτική μορφή τα αντίστοιχα βήματα που πρέπει να ακολουθήσει ο αναγνώστης για να ολοκληρώσει μια εφαρμογή επιχειρηματικής ευφυΐας.

- **Σχεδιασμός.** Ξεκινάμε προσδιορίζοντας τον ειδικό επιχειρηματικό στόχο και διατυπώνουμε με σαφήνεια το πρόβλημα. Δεν έχει νόημα να τροφοδοτήσουμε ένα μηχανισμό εξαγωγής γνώσης με ό,τι δεδομένα διαθέτουμε, χωρίς να ξέρουμε τι ψάχνουμε. Ο στόχος πρέπει να είναι εφικτός, κάτι που μπορούμε να κρίνουμε με βάση την αντίληψη που έχουμε για τα διαθέσιμα δεδομένα, τις δυνατότητες των προσφερόμενων μεθόδων και τη λογική του προβλήματος. Γνωρίζοντας ποιες μέθοδοι υπάρχουν και για τι είδους πρόβλημα προσφέρεται η καθεμιά, μπορεί να γίνει ένας σχεδιασμός για τον καταλληλότερο τρόπο μοντελοποίησης (π.χ. αν θέλω μάθω πώς να ξεχωρίζω τους πελάτες σε «καλούς-κακούς», χρειάζομαι ένα μοντέλο κατάταξης, ενώ αν θέλω να μάθω τι προφίλ πελατών υπάρχουν και τι προτιμάει ο καθένας, κατευθύνομαι προς μια μέθοδο συσταδοποίησης).
- **Εισαγωγή και προσαρμογή των δεδομένων**. Για το σκοπό αυτό, μπορεί να δημιουργηθεί στο RapidMiner μια προκαταρκτική διαδικασία (process), που να περιλαμβάνει έναν τελεστή (operator) εισαγωγής δεδομένων, είτε εισαγωγής από εξωτερικό αρχείο/βάση δεδομένων, είτε ανάγνωσης από το αποθετήριο. Ένας ή περισσότεροι τελεστές μπορούν στη συνέχεια να προστεθούν για την επιλογή και διαμόρφωση των δεδομένων στην επιθυμητή μορφή. Το βασικότερο, που χρειάζεται σχεδόν πάντοτε, είναι να επιλεγούν τα χαρακτηριστικά που είναι χρήσιμα στην ανάλυση και να «κοπούν» όλα τα υπόλοιπα, ενώ ταυτόχρονα μπορεί να απαιτείται να προσδιοριστεί ο ρόλος κάποιων χαρακτηριστικών. Η λογική του προβλήματος και ο τύπος του τελεστή εκμάθησης απαιτούν να προσδιοριστούν ορισμένα χαρακτηριστικά με συγκεκριμένους ρόλους, με συνηθέστερους τον **id** (περιέχει τιμές που προσδιορίζουν μοναδικά το κάθε παράδειγμα), τον **regular** (απλό χαρακτηριστικό που περιγράφει ένα παράδειγμα), τον **label** (καθορίζει το χαρακτηριστικό ως στόχο εκμάθησης) και τον **weight** (καθορίζει το «βάρος» ή σημαντικότητα του κάθε παραδείγματος). Επίσης, συχνά απαιτείται η μετατροπή του τύπου δεδομένων ενός χαρακτηριστικού, π.χ. από ονομαστικό σε αριθμητικό ή συνεχές ποσοτικό σε διακριτό. Τέλος, σημαντική είναι η επιλογή παραδειγμάτων με βάση κάποιο κριτήριο (π.χ. μόνο οι άντρες πελάτες) και το φιλτράρισμα ανεπιθύμητων παραδειγμάτων, όπως αυτά που περιέχουν πολλές κενές τιμές. Οι τελεστές που αφορούν αυτό το βήμα είναι των κατηγοριών **Repository Access**, **Import** και **Data Transformation**.
- **Επισκόπηση των δεδομένων.** Συνιστάται η προσεκτική επισκόπηση των δεδομένων που πρόκειται να οδηγηθούν στον κύριο τελεστή μοντελοποίησης. Πολύ χρήσιμα εργαλεία για αυτόν το σκοπό, πέρα από την προβολή του πίνακα δεδομένων (Σχήμα 7.2), είναι η καρτέλα των στατιστικών του RapidMiner (Σχήμα 7.3) και οι καρτέλες γραφημάτων. Η «γνωριμία» με τα δεδομένα θεωρείται πολύ σημαντική γιατί δίνει τη δυνατότητα να εντοπιστούν οι αδυναμίες τους και να μετασχηματιστούν έτσι ώστε να βελτιωθεί η απόδοση της μοντελοποίησης. Π.χ. θεωρήστε μια προσπάθεια μοντελοποίησης του πώς επηρεάζεται η συνολική ικανοποίηση του πελάτη από τους παράγοντες εξυπηρέτηση, τιμή, ποιότητα προϊόντος. Παρατηρώντας τα δεδομένα εισόδου, μπορεί να διαπιστωθούν προβλήματα όπως το εξής: Όλοι οι πελάτες που καταγράφηκαν ως παραδείγματα ήταν πολύ ικανοποιημένοι και κανένας ουδέτερος ή δυσαρεστημένος. Πώς είναι δυνατόν να εκτιμηθεί το ποιος παράγοντας συνέβαλε περισσότερο στην ικανοποίηση αν όλοι οι πελάτες είναι πάντα ικανοποιημένοι; Μπορεί επίσης να παρατηρηθούν ελλιπή στοιχεία που δυσχεραίνουν την ανάλυση ή «ύποπτα» ασυνεπή στοιχεία (π.χ. κάποιος πελάτης δηλώνει δυσαρεστημένος από όλα τα επιμέρους στοιχεία, αλλά συνολικά είναι απόλυτα ικανοποιημένος).
- **Μοντελοποίηση.** Το βήμα αυτό πραγματοποιείται χρησιμοποιώντας τους κατάλληλους τελεστές της κατηγορίας **Modeling**. Οι τελεστές που διαθέτει το RapidMiner καλύπτουν μεγάλο εύρος αναγκών και όλες τις βασικές κατηγορίες μεθόδων, συμπεριλαμβανομένων των καθαρά στατιστικών (π.χ. παλινδρόμηση, K-means, ANOVA), των αλγοριθμικών (π.χ. Decision Tree, Rule Induction), καθώς και ειδικών μεθόδων, όπως τα νευρωνικά δίκτυα. Η υλοποίηση της διαδικασίας γίνεται: (α) Επιλέγοντας και συνδέοντας κατάλληλα τους απαραίτητους τελεστές, λαμβάνοντας υπόψη ότι πολλές τεχνικές απαιτούν περισσότερα από ένα βήματα και ότι πολλοί τελεστές απαιτούν ειδική προετοιμασία στα δεδομένα πριν τροφοδοτηθούν σε αυτούς (π.χ. ενώ ο τελεστής **Decision Tree** δέχεται μείγμα ονομαστικών

και ποσοτικών χαρακτηριστικών, τα δέντρα **CHAID** απαιτούν αποκλειστικά ονομαστικά χαρακτηριστικά, επομένως πρέπει να προηγηθεί διαδικασία χωρισμού σε τάξεις για όλα τα ποσοτικά χαρακτηριστικά). (β) Ρυθμίζοντας τις παραμέτρους του κάθε τελεστή μοντελοποίησης, ώστε η μέθοδος να προσαρμοστεί στα δεδομένα του προβλήματος. Παρόλο που το RapidMiner προσφέρει εύστοχα καθορισμένες προεπιλεγμένες τιμές για όλες τις παραμέτρους, για να επιτευχθούν άριστα αποτελέσματα, απαιτείται διαδικασία βελτιστοποίησης, που περιλαμβάνει πειραματισμό και απαιτεί κατανόηση της λειτουργίας των μεθόδων. Σημειώνεται ότι το RapidMiner διαθέτει εξαιρετικό ενσωματωμένο **Help**, που επεξηγεί με κατανοητό τρόπο το σκοπό του κάθε τελεστή και την έννοια της κάθε παραμέτρου του.

- **Εκτέλεση της διαδικασίας και κατανόηση των αποτελεσμάτων**. Το κύριο αποτέλεσμα της διαδικασίας εξαγωγής γνώσης είναι ένα «εκπαιδευμένο» ή «υπολογισμένο» μοντέλο γνώσης (π.χ. ένα κατασκευασμένο δέντρο αποφάσεων, οι συντελεστές μιας παλινδρόμησης ή ένα εκπαιδευμένο νευρωνικό δίκτυο). Ανάλογα με τον τύπο του μοντέλου και τους στόχους μας, μπορεί το κύριο ενδιαφέρον μας να είναι το ίδιο το περιεχόμενο του μοντέλου ή μπορεί να μας ενδιαφέρει κυρίως η μελλοντική εφαρμογή του μοντέλου σε άγνωστα παραδείγματα (ή φυσικά και τα δύο). Ένα δέντρο αποφάσεων μας πληροφορεί για το ποια χαρακτηριστικά παίζουν τον κυριότερο ρόλο στην πρόβλεψη του αποτελέσματος και δίνουν μια εξήγηση για το φαινόμενο που μελετούμε. Ομοίως, οι κανόνες συσχέτισης που προκύπτουν από μια ανάλυση καλαθιού αγορών είναι το κύριο (και μάλλον το τελικό) αποτέλεσμα. Αντίθετα, ένα εκπαιδευμένο νευρωνικό δίκτυο δε μας λέει τίποτα αυτό καθεαυτό, αλλά χρησιμεύει μόνο δίνοντας απαντήσεις όταν θα τροφοδοτηθεί με άγνωστες περιπτώσεις. Στο RapidMiner, έχοντας ολοκληρώσει τη σχεδίαση της διαδικασίας, πατούμε **Run or resume the current process**, ώστε να εκτελεστεί η διαδικασία και να δούμε το αποτέλεσμα. (Σημείωση: αυτό είναι το κύριο βήμα υπολογισμών και απαιτεί μεγάλη υπολογιστική ισχύ και αντίστοιχο χρόνο, ιδιαίτερα αν εφαρμόζεται μια περίπλοκη διαδικασία σε μεγάλο όγκο δεδομένων) Ανάλογα με τη φύση της διαδικασίας, το αποτέλεσμα μπορεί να προβληθεί ως γράφημα (π.χ. δέντρο), κείμενο (π.χ. κανόνες) ή σε πινακοποιημένη μορφή. Η διαδικασία της κατανόησης των αποτελεσμάτων και η εξέταση της ορθότητας και της ακρίβειας του μοντέλου είναι σημαντικά και συνήθως είναι μέρος μιας επαναληπτικής διαδικασίας βελτίωσης της ανάλυσης, όπου, ανάλογα με τα προβλήματα που εντοπίζουμε στο αποτέλεσμα, τροποποιούμε τις παραμέτρους ή και όλη τη διαδικασία (π.χ. αν ο τύπος δέντρου που επιλέξαμε δεν μπορεί να «μάθει» με ακρίβεια τη σωστή κατάταξη, μπορούμε να αλλάξουμε το κριτήριο διάσπασης κόμβων, το μέγιστο βάθος, το επίπεδο εμπιστοσύνης που εφαρμόζεται στο κλάδεμα, ή ακόμα και να αλλάξουμε τον τύπο του δέντρου). Στο πρόγραμμα προσφέρονται χρήσιμα εργαλεία για την επισκόπηση των αποτελεσμάτων, όπως επιλογή των κανόνων που αφορούν συγκεκριμένο χαρακτηριστικό, φιλτράρισμα των πινάκων με βάση κάποια κριτήρια, αναζήτηση, κ.ά.
- **Εφαρμογή του μοντέλου**. Το εκπαιδευμένο μοντέλο, ανάλογα με τη μορφή του, μπορεί να εφαρμοστεί σε άγνωστα δεδομένα (δηλαδή παραδείγματα όπου το χαρακτηριστικό-στόχος είναι άγνωστο και πρέπει να προβλεφθεί), με άλλα λόγια, η γνώση που εξάγαμε για ένα φαινόμενο αναλύοντας ένα σύνολο παραδειγμάτων μπορεί να εφαρμοστεί για να εκτιμήσουμε μια νέα άγνωστη περίπτωση. Η ενέργεια αυτή πραγματοποιείται στο RapidMiner χρησιμοποιώντας τον τελεστή **Apply Model**. Η εφαρμογή του μοντέλου μπορεί να γίνει κατά τη φάση ανάπτυξης σε ένα σετ δεδομένων ελέγχου, ώστε να πραγματοποιηθεί αξιολόγηση της διαδικασίας ή κατά την τελική χρήση από ένα στέλεχος επιχείρησης για τη λήψη απόφασης σε νέες πραγματικά άγνωστες περιπτώσεις.

Η υλοποίηση των παραπάνω βημάτων στην πράξη επεξηγείται με περισσότερη λεπτομέρεια με τη βοήθεια των παραδειγμάτων τυπικών εφαρμογών επιχειρηματικής ευφυΐας που παρουσιάζονται στην επόμενη ενότητα.

7.3 Εφαρμογές επιχειρηματικής ευφυΐας με χρήση εξαγωγής γνώσης από δεδομένα

7.3.1 Πρόβλεψη απώλειας πελάτη

7.3.1.1 Ορισμός προβλήματος

Ένα απλό και τυπικό παράδειγμα εφαρμογής επιχειρηματικής ευφυΐας είναι η πρόβλεψη της απώλειας πελάτη για μια εταιρεία τηλεπικοινωνιών. Οι πληροφορίες που είναι συνήθως διαθέσιμες για τους πελάτες μιας τέτοιας εταιρείας περιλαμβάνουν τα δημογραφικά του στοιχεία (ηλικία, φύλο, επάγγελμα, κλπ.), δεδομένα για τη συμπεριφορά (π.χ. με τι τρόπο πληρώνει, τι υπηρεσίες χρησιμοποιεί, πόσο διάστημα είναι πελάτης της εταιρείας) και πληροφορίες σχετικά με τα έσοδα από αυτόν. Κατά την ανανέωση συμβολαίου, κάποιοι πελάτες ανανεώνουν το συμβόλαιό τους και κάποιοι μεταπηδούν σε άλλη εταιρεία (η αλλαγή εταιρείας αποδίδεται με τη λέξη churn). Είναι χρήσιμο για την εταιρεία να γνωρίζει ποιοι πελάτες έχουν μεγάλη πιθανότητα να αλλάξουν εταιρεία, ώστε να προσπαθήσει να το αποτρέψει, ειδικά στην περίπτωση πελατών που αποφέρουν μεγάλα έσοδα/κέρδος. Μελετώντας την πληροφορία που έχουμε για τους πελάτες μας και τη συμπεριφορά τους στο παρελθόν, ο στόχος μας είναι να μάθουμε να προβλέπουμε για κάθε πελάτη τον κίνδυνο να αλλάξει εταιρεία στο μέλλον.

Το πρόβλημα που περιγράφηκε είναι πρόβλημα πρόβλεψης. Ξεκινάμε από ένα σετ δεδομένων που περιλαμβάνει μεγάλο αριθμό παραδειγμάτων πελατών, για τους οποίους γνωρίζουμε τα χαρακτηριστικά τους και το αν άλλαξαν τελικά εταιρεία ή όχι. Κατασκευάζουμε μια διαδικασία με την οποία προσπαθούμε να μοντελοποιήσουμε (ή αλλιώς να «μάθουμε») ποιοι συνδυασμοί χαρακτηριστικών μπορούν να προβλέψουν την απόφαση του πελάτη. Εκτελώντας τη διαδικασία, παίρνουμε ως αποτέλεσμα ένα μοντέλο γνώσης που έχει εκπαιδευτεί να προβλέπει το αν κάποιος πελάτης μας θα αλλάξει εταιρεία ή όχι, με βάση συγκεκριμένα χαρακτηριστικά που είναι διαθέσιμα για αυτόν. Το μοντέλο μπορεί να μελετηθεί από ένα στέλεχος μάρκετινγκ για να βγάλει συμπεράσματα για τον τρόπο απόφασης των πελατών. Ο σκοπός όμως στον οποίο εστιάζουμε, είναι να μπορούμε να τροφοδοτούμε το εκπαιδευμένο μοντέλο με δεδομένα πελατών, των οποίων το συμβόλαιο πλησιάζει στη λήξη τους, ώστε το μοντέλο να εντοπίζει για εμάς αυτούς για τους οποίους υπάρχει κίνδυνος απώλειας.

Το σετ δεδομένων που θα χρησιμοποιηθεί για την επίδειξη της εφαρμογής είναι τεχνητό και διατίθεται από την RapidMiner μαζί με το πρόγραμμα, ως εκπαιδευτικό δείγμα (http://docs.rapidminer.com/studio/getting-started/customer-churn-data.xlsx). Περιέχει υποθετικά στοιχεία για 996 πελάτες, για τους οποίους γνωρίζουμε τα χαρακτηριστικά (attributes): **Φύλο (Gender)**, **Ηλικία (Age)**, **Τρόπος πληρωμής (Payment Method)**, **Αλλαγή εταιρείας (Churn)** και **Τελευταία δοσοληψία (LastTransaction)**. Σημειώνεται ότι τα χαρακτηριστικά είναι λίγα και απλοϊκά σε σχέση με μια πραγματική εφαρμογή, είναι όμως αντιπροσωπευτικά και κατάλληλα για εκπαιδευτικούς σκοπούς. Στη συνέχεια θα παρουσιαστεί ο τρόπος εκτέλεσης των βασικών βημάτων που αναφέρονται στην παραπάνω Ενότητα 2.

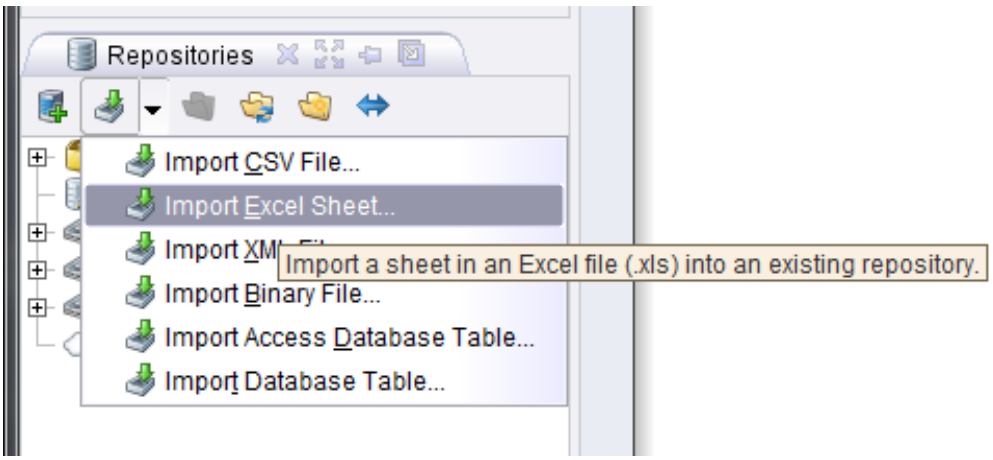
7.3.1.2 Σχεδιασμός

Στο πρόβλημα που έχει οριστεί, επιθυμούμε την κατάταξη των πελατών σε μια από τις δύο πιθανές περιπτώσεις: να αλλάξουν εταιρεία (**Churn**=*yes*) ή να μην αλλάξουν (**Churn**=*no*). Επομένως, η μέθοδος μοντελοποίησης πρέπει να ανήκει στην κατηγορία των μεθόδων κατάταξης (Classification). Στην κατηγορία αυτή ανήκει πληθώρα μεθόδων, συμπεριλαμβανομένων των απλών στατιστικών (Bayesian modeling, παλινδρόμηση), των δέντρων κατάταξης και των νευρωνικών δικτύων. Επειδή στο πρόβλημά μας συμπεριλαμβάνονται ονομαστικά χαρακτηριστικά (φύλο, τρόπος πληρωμής) και ποσοτικά (ηλικία, τελευταία συναλλαγή), αποκλείουμε τις μεθόδους που δεν μπορούν να χειριστούν κάποιο από αυτά και επίσης αποφεύγουμε τα νευρωνικά δίκτυα, επειδή είναι επιθυμητό να παράγουμε κάποιο μοντέλο του οποίου η λογική να είναι κατανοητή και ελέγξιμη. Επομένως, μια καλή επιλογή είναι ένα δέντρο κατάταξης, συγκεκριμένα ο τελεστής **Decision Tree**, που χειρίζεται ταυτόχρονα ποσοτικά και ονομαστικά χαρακτηριστικά.

7.3.1.3 Εισαγωγή και προσαρμογή των δεδομένων

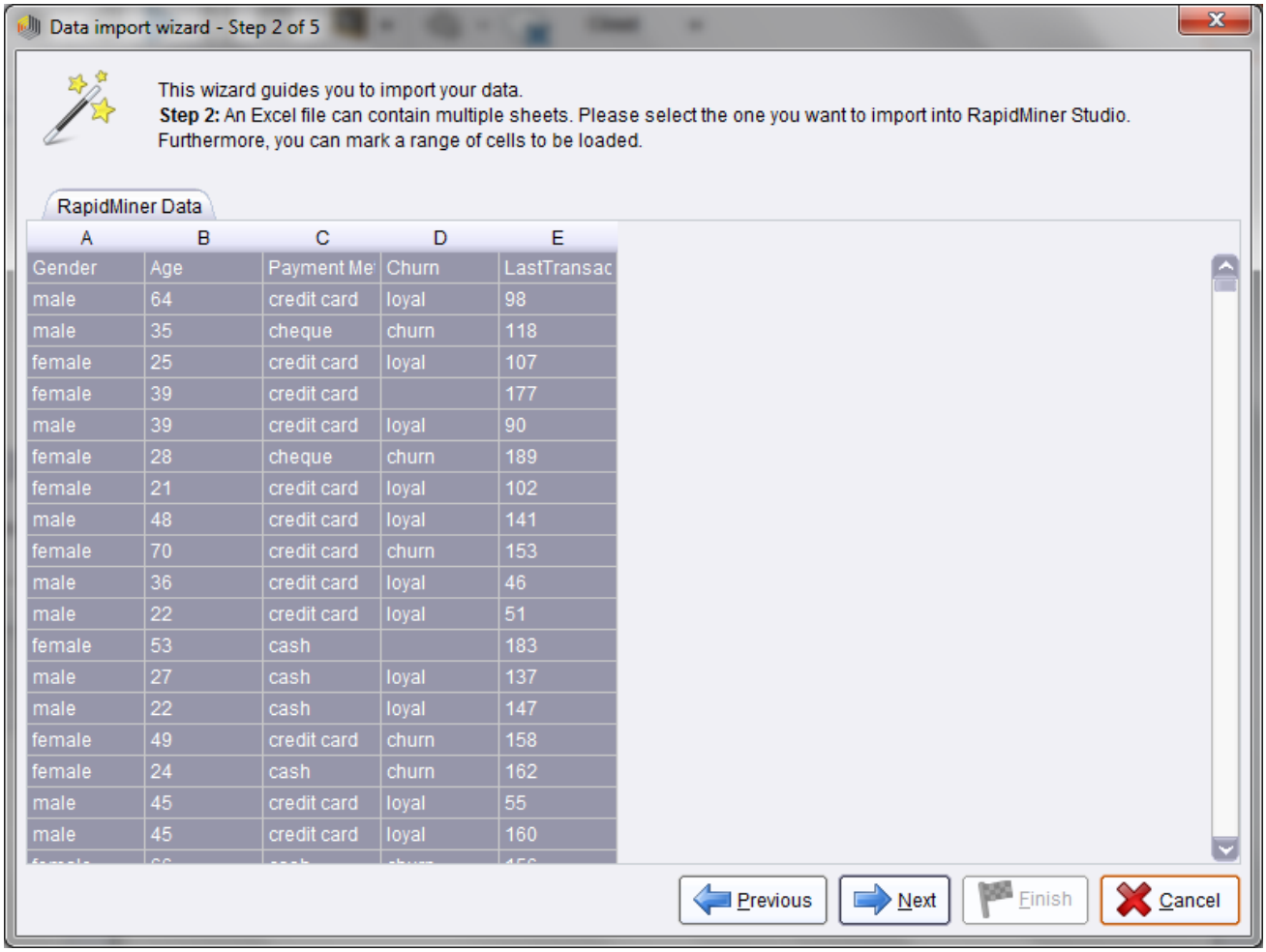
Η εισαγωγή των εκπαιδευτικών δεδομένων απευθείας από ένα εξωτερικό αρχείο Excel μπορεί να ενσωματωθεί εύκολα σε μια διαδικασία με την βοήθεια του τελεστή **Read Excel**. Εναλλακτικά, μπορεί να γίνει εισαγωγή των δεδομένων στο αποθετήριο του RapidMiner, μέσω του κατάλληλου οδηγού, και αφού γίνει ένας προκαταρκτικός έλεγχος. Σε αυτήν την περίπτωση, τα δεδομένα είναι διαθέσιμα για ανάγνωση από τη διαδικασία μέσω του τελεστή **Retrieve**. Στη συνέχεια θα επιδειχθεί ο δεύτερος τρόπος.

Από το παράθυρο Repositories, επιλέγουμε Import Excel Sheet (Σχήμα 7.8). Στο πρώτο βήμα του οδηγού εισαγωγής δεδομένων (data import wizard), επιλέγουμε το αρχείο που περιέχει τα δεδομένα και πατάμε Next.



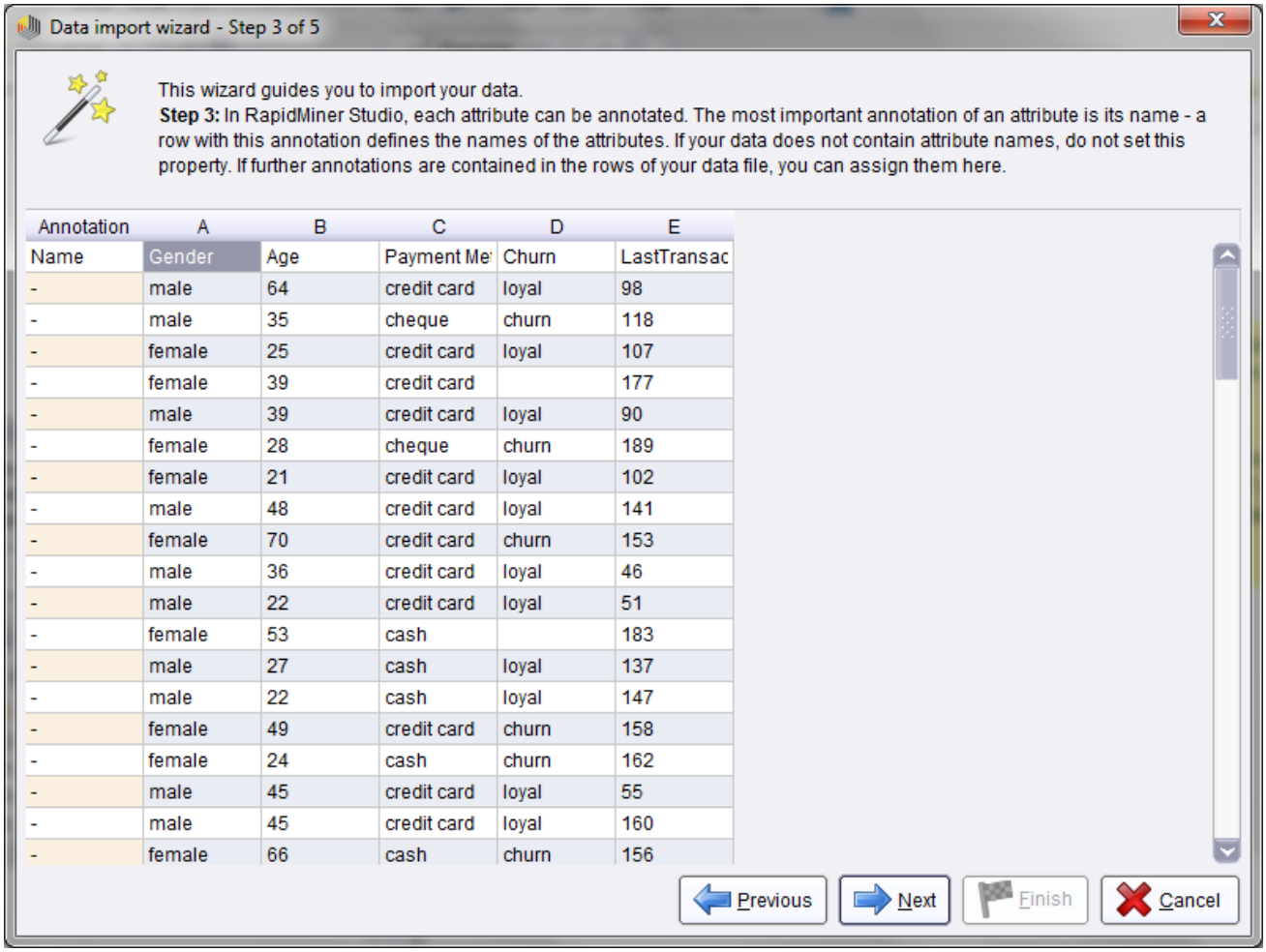
Σχήμα 7.8. *Ενεργοποίηση του οδηγού εισαγωγής δεδομένων*

Στο 2^{ο} βήμα, ο οδηγός προβάλει το περιεχόμενο του αρχείου. Μπορούμε να επιλέξουμε για εισαγωγή μια συγκεκριμένη περιοχή κελιών ή όλο το περιεχόμενο. Αν στο αρχείο υπάρχουν περισσότερα από ένα φύλλα, μπορούμε επίσης να επιλέξουμε το φύλλο.



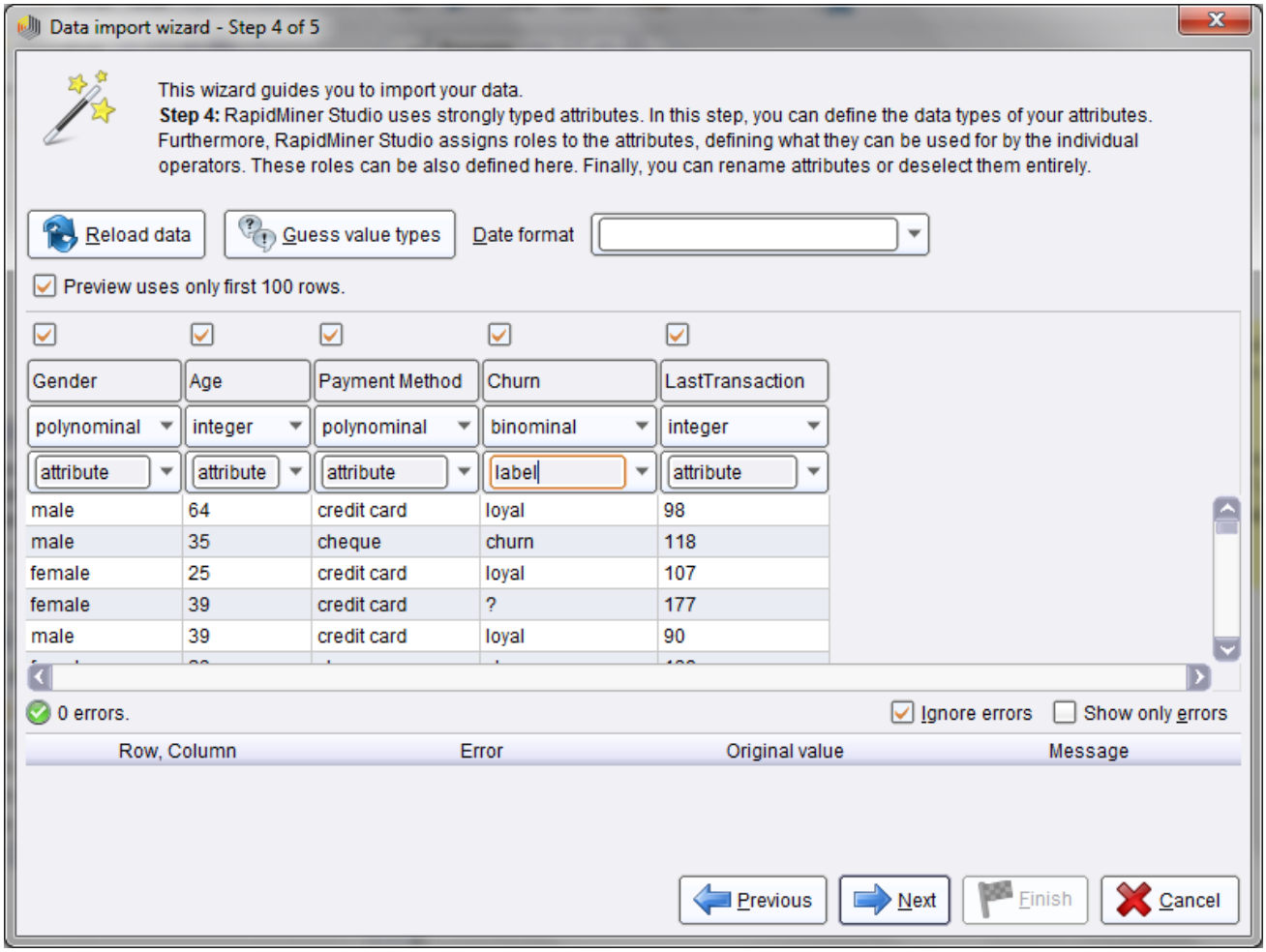
Σχήμα 7.9. *Επιλογή φύλλου και κελιών προς εισαγωγή.*

Στο 3^{ο} βήμα, μπορούμε να καθορίσουμε τα ονόματα των χαρακτηριστικών ή να προσθέσουμε σημειώσεις για αυτά. Στο Σχήμα 7.10 φαίνεται ότι το RapidMiner ερμήνευσε αυτόματα την πρώτη γραμμή του φύλλου Excel ως τα ονόματα των χαρακτηριστικών, κάτι που δε χρειάζεται να αλλάξουμε.



Σχήμα 7.10. *Καθορισμός των ονομάτων των χαρακτηριστικών*

Στο 4ο βήμα, μπορούμε να καθορίσουμε τον τύπο του κάθε χαρακτηριστικού και το ρόλο του. Στο Σχήμα 7.11 βλέπουμε ότι το RapidMiner μάντεψε τους τύπους των χαρακτηριστικών ως εξής: τα **Gender, Payment method** και **Churn** είναι τύπου polynominal (ονομαστικά που παίρνουν πολλαπλές τιμές) και τα **Age** και **LastTransaction** τύπου integer (ποσοτικά με ακέραιες τιμές). Οι ορισμοί είναι σωστοί, με μόνη παρατήρηση ότι το **Churn** παίρνει μόνο δύο τιμές *(loyal* και *churn*), επομένως είναι προς όφελός μας να οριστεί ως binominal (λογικό χαρακτηριστικό τύπου Ναι/όχι). Στο βήμα αυτό καθορίζουμε επίσης το ρόλο του κάθε χαρακτηριστικού. Αφήνουμε την επιλογή attribute σε όλα τα χαρακτηριστικά (απλό χαρακτηριστικό που περιέχει πληροφορία για το παράδειγμα), όμως αλλάζουμε το χαρακτηριστικό **Churn** σε label (ο στόχος που θα πρέπει το μοντέλο να μάθει να προβλέπει ή με άλλα λόγια η μεταβλητή που θεωρούμε εξαρτημένη από τις υπόλοιπες). Τέλος, δίνουμε ένα όνομα στο νεοεισαχθέν σετ δεδομένων και επιλέγουμε το φάκελο του αποθετηρίου στον οποίο θα αποθηκευτεί.

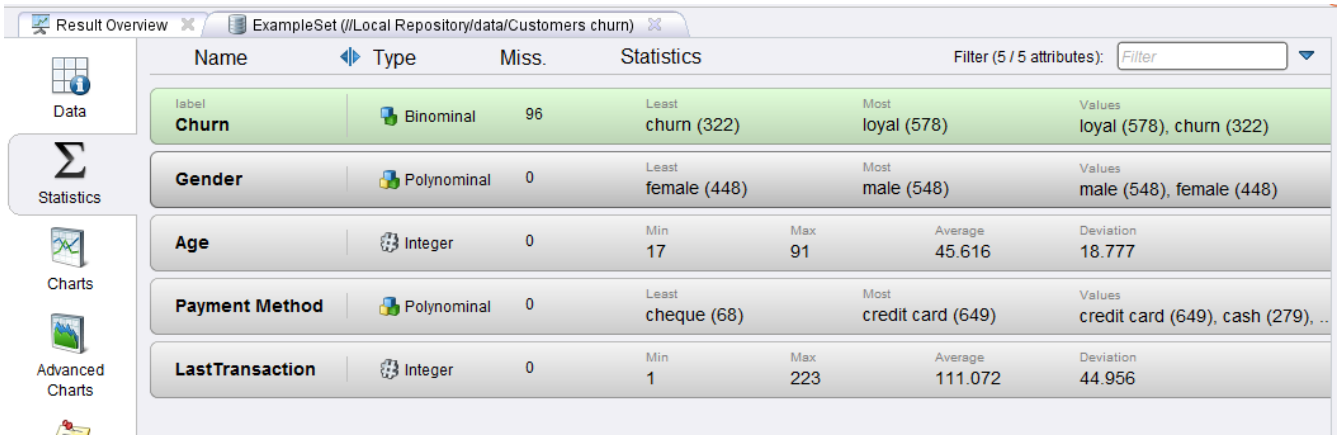


Σχήμα 7.11. *Το βήμα καθορισμού του τύπου δεδομένων και του ρόλου κάθε χαρακτηριστικού.*

7.3.1.4 Επισκόπηση των δεδομένων

Αμέσως μετά το κλείσιμο του οδηγού εισαγωγής, το RapidMiner μας μεταφέρει αυτόματα σε προβολή Αποτελεσμάτων και ανοίγει για επισκόπηση τα δεδομένα που εισήχθησαν. Το άνοιγμα των δεδομένων για επισκόπηση μπορεί να γίνει οποιαδήποτε στιγμή, μεταβαίνοντας στο παράθυρο του Αποθετηρίου και κάνοντας διπλό κλικ στο αντίστοιχο αντικείμενο.

Στην καρτέλα Data μπορούμε να παρατηρήσουμε τα δεδομένα, ώστε να έχουμε μια αντίληψη για το περιεχόμενό τους. Παρατηρούμε ότι υπάρχουν παραδείγματα για τα οποία είναι κενή η τιμή του χαρακτηριστικού **Churn** (η απούσα τιμή παριστάνεται με το σύμβολο του ερωτηματικού **?**). Τα παραδείγματα αυτά είναι προφανώς άχρηστα για εκμάθηση και θα πρέπει να φιλτραριστούν. Στην καρτέλα Statistics μπορούμε να δούμε τις τιμές που παίρνουν τα χαρακτηριστικά και τις συχνότητές τους. Παρατηρούμε ότι το δείγμα περιλαμβάνει 448 γυναίκες και 548 άντρες, επομένως είναι σχετικά ισορροπημένο όσον αφορά το φύλο, ενώ αντίθετα περιλαμβάνει, ως προς τον τρόπο πληρωμής, μόνο 68 παραδείγματα πληρωμής με επιταγή, σε σύγκριση με 649 πελάτες που πληρώνουν με πιστωτική κάρτα. Στο παράδειγμα αυτό δεν απαιτείται να κάνουμε κάποια ενέργεια, θα μπορούσαμε όμως να είχαμε εντοπίσει προβλήματα, όπως άκυρες τιμές ή περιπτώσεις που εκπροσωπούνται ελάχιστα.

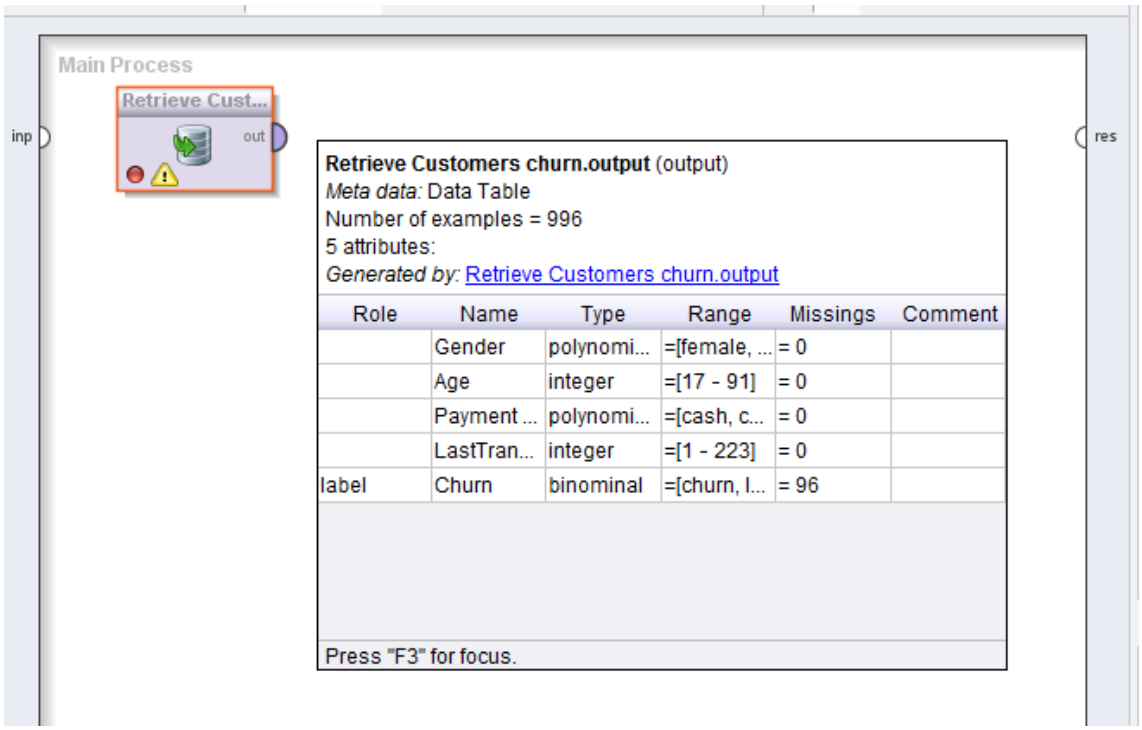


Σχήμα 7.12. *Επισκόπηση των δεδομένων στην καρτέλα στατιστικών για έλεγχο της ποιότητάς τους.*

7.3.1.5 Μοντελοποίηση

Δημιουργούμε βήμα-βήμα τη διαδικασία μοντελοποίησης ως εξής:

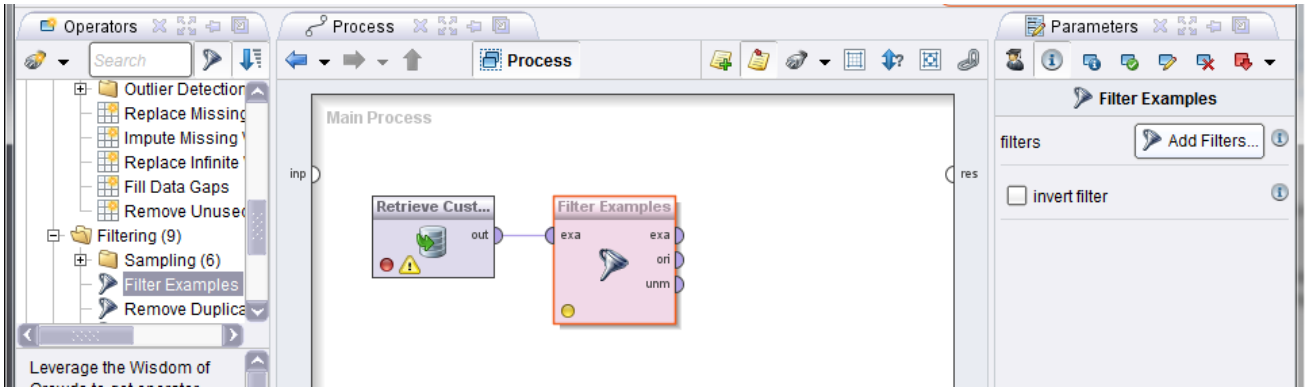
Εισάγουμε τα δεδομένα Customers churn σύροντας το αντίστοιχο εικονίδιο από το Αποθετήριο στο χώρο σχεδίασης της διαδικασίας. Δημιουργείται αυτόματα ο κατάλληλος τελεστής **Retrieve**. Τοποθετώντας το ποντίκι πάνω από τη θύρα out του τελεστή αυτού, εμφανίζονται τα μετα-δεδομένα (δηλαδή η περιγραφή των δεδομένων) (Σχήμα 7.13).



Σχήμα 7.13. *Εισάγοντας τα δεδομένα στο χώρο σχεδίασης της διαδικασίας, μπορούμε να δούμε την περιγραφή τους.*

Εισάγουμε τον τελεστή Filter Examples της κατηγορίας Data Transformation για να απορρίψουμε τα ανεπιθύμητα παραδείγματα για τα οποία δεν υπάρχει τιμή για το χαρακτηριστικό **Churn**. Συνδέουμε την έξοδο του **Retrieve** με την είσοδο του **Filter Examples** και επιλέγουμε το τελευταίο ώστε να έχουμε

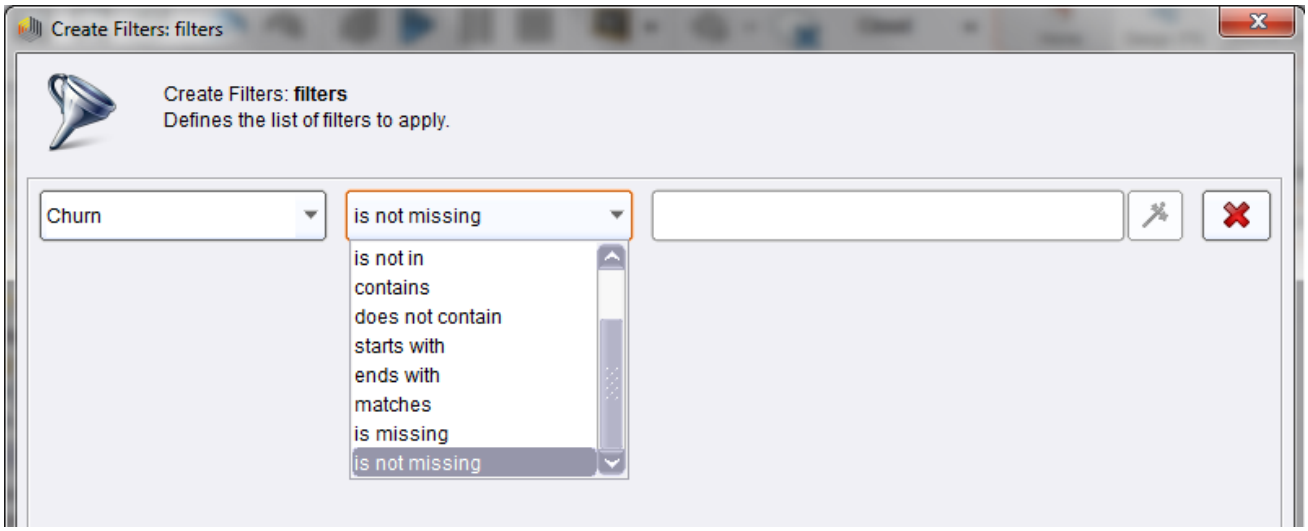
πρόσβαση στις παραμέτρους του. Από την καρτέλα Parameters πατάμε το κουμπί Add Filters, ώστε να ανοίξει ο οδηγός δημιουργίας φίλτρων.



Σχήμα 7.14. *Η εισαγωγή φίλτρου για την απομάκρυνση των ακατάλληλων παραδειγμάτων.*

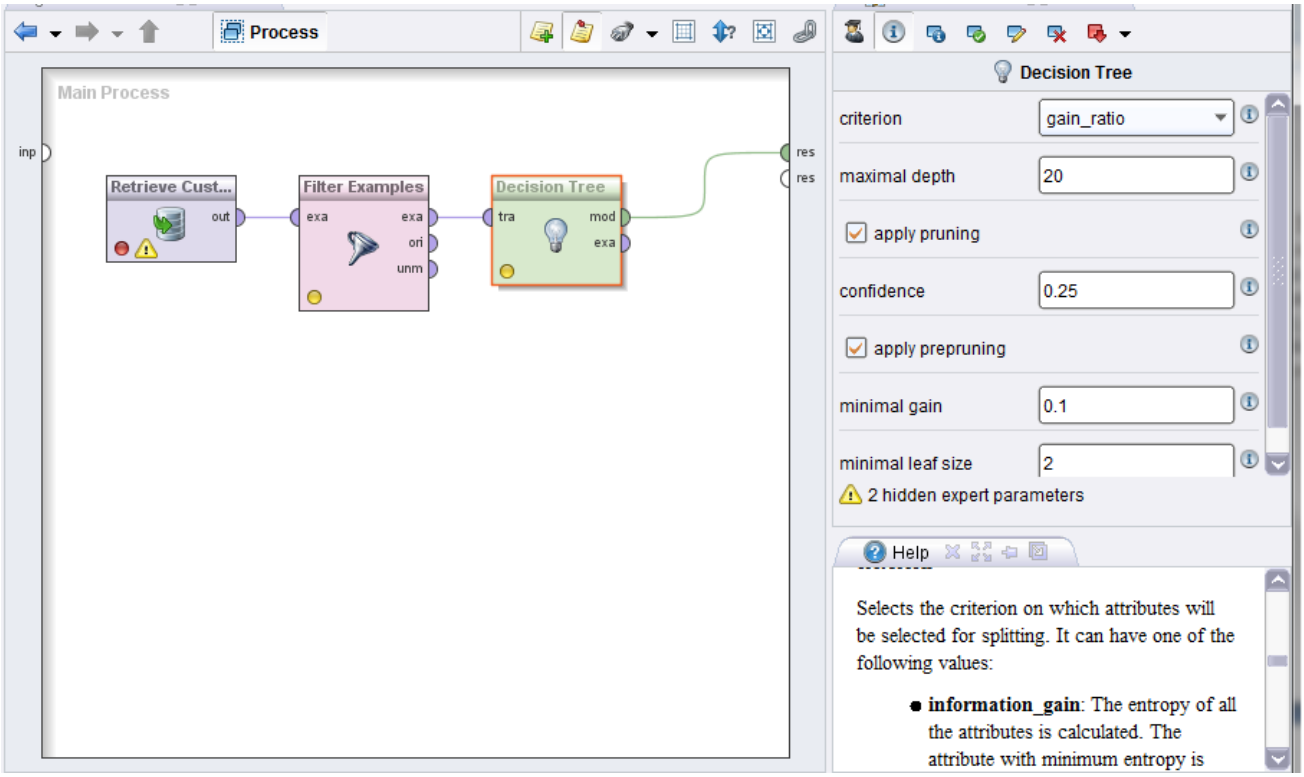
Στον οδηγό δημιουργίας φίλτρων επιλέγουμε το πεδίο Churn και τη συνθήκη is Not Missing (Σχήμα 7.15).

.



Σχήμα 7.15. *Ο οδηγός δημιουργίας φίλτρων.*

Στη συνέχεια, εισάγουμε τον τελεστή που θα πραγματοποιήσει τη μοντελοποίηση. Στο δέντρο επιλογής τελεστών, ανοίγουμε την κατηγορία **Modeling**, μετά την **Classification and Regression** και στη συνέχεια την **Tree Induction**. Εκεί βρίσκουμε τον τελεστή **Decision Tree**, τον οποίο σύρουμε στη διαδικασία και συνδέουμε με την έξοδο του **Filter Examples**. Συνδέουμε επίσης την έξοδο *mod* (δηλ model) με την έξοδο της διαδικασίας (θύρα *res*) (Σχήμα 7.16). Επιλέγοντας τον **Decision Tree**, εμφανίζονται οι παράμετροι της μεθόδου εκμάθησης. Για κάθε παράμετρο, εμφανίζεται στη δεξιά του άκρη ένα εικονίδιο βοήθειας, που προβάλλει πληροφορίες για την έννοια της παραμέτρου, ενώ επίσης, στο κάτω δεξιά άκρο της οθόνης εμφανίζεται καρτέλα βοήθειας με αναλυτική περιγραφή του σκοπού και της λειτουργίας του τελεστή.

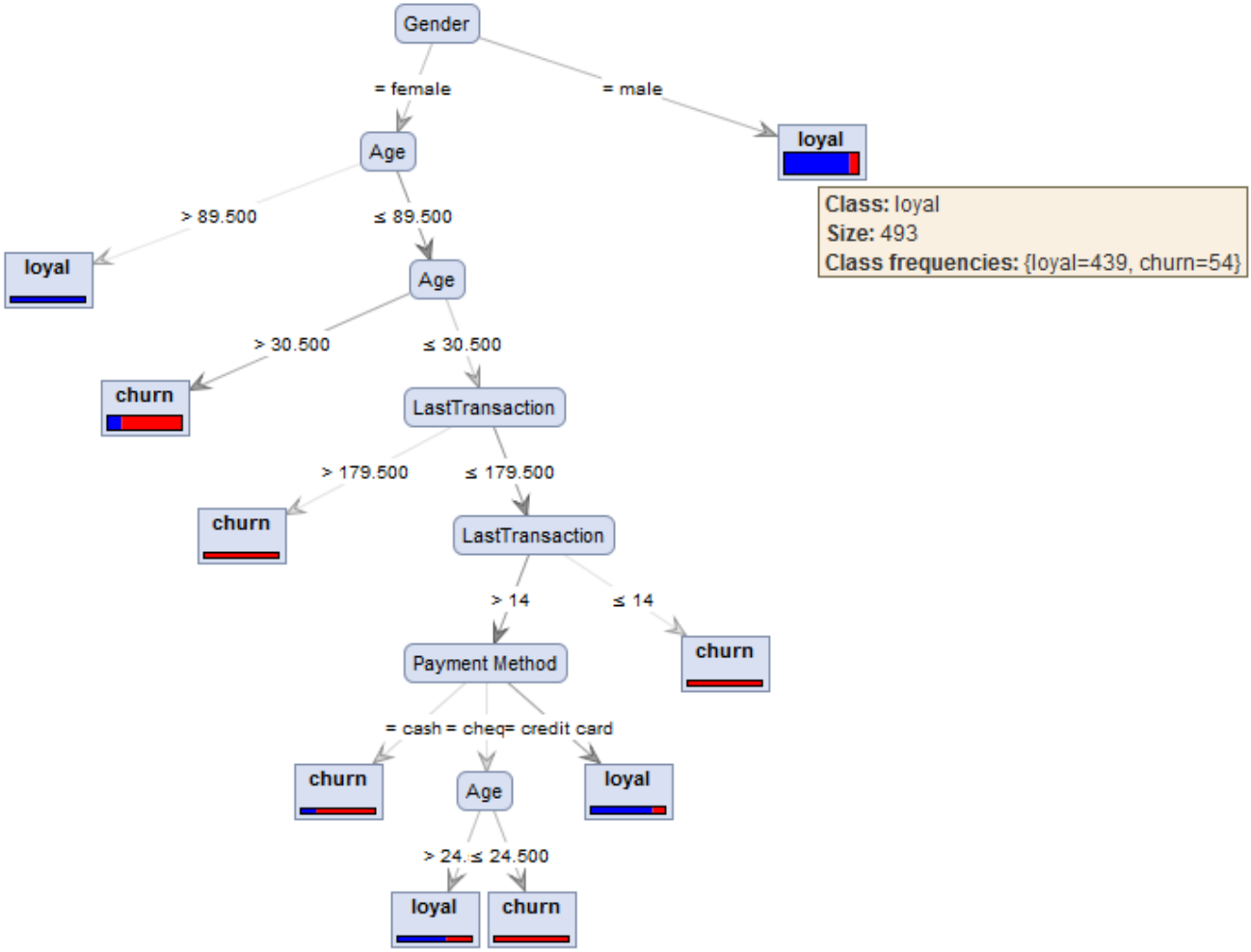


Σχήμα 7.16. *Η εισαγωγή και ρύθμιση του τελεστή μοντελοποίησης Decision Tree.*

Μια από τις σημαντικότερες παραμέτρους είναι το κριτήριο με βάση το οποίο θα αποφασίζει ο αλγόριθμος ποιος διαχωρισμός (splitting) θεωρείται καλύτερος (δηλαδή ποιο χαρακτηριστικό πρέπει να τοποθετηθεί στις υψηλότερες διακλαδώσεις ώστε να αποφέρει τον αποτελεσματικότερο διαχωρισμό σε κλαδιά. Άλλη παράμετρος είναι το μέγιστο βάθος μέχρι το οποίο μπορεί να προχωρήσει η ανάπτυξη του δέντρου. Για την κατανόηση του τρόπου επιλογής των παραμέτρων συνιστάται στον αναγνώστη να αναφερθεί στο Κεφάλαιο 6 και επίσης στην καρτέλα βοήθειας του RapidMiner.

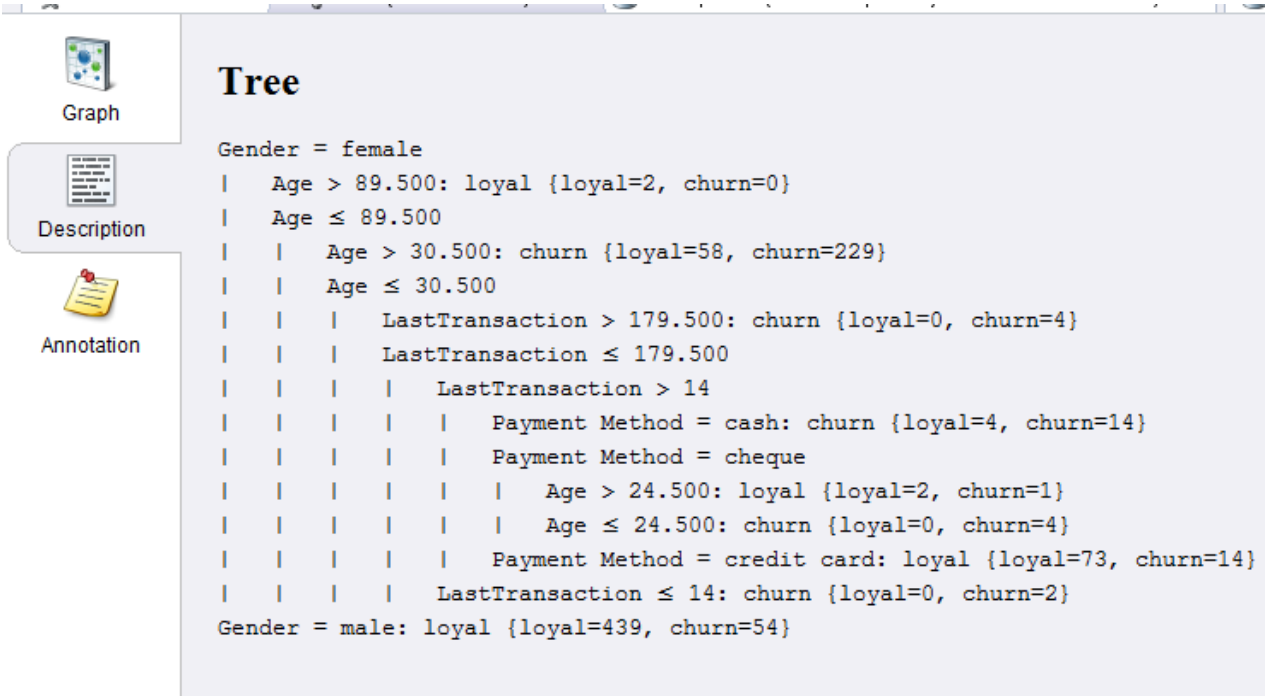
Στο παράδειγμα αυτό, κρατάμε αρχικά τις προεπιλεγμένες τιμές για όλες τις παραμέτρους και προχωράμε στην εκτέλεση. Πατώντας το κουμπί εκτέλεσης, το RapidMiner διαβάζει τα δεδομένα, εκτελεί τους απαραίτητους υπολογισμούς και μας μεταφέρει στην προβολή Αποτελεσμάτων. Το αποτέλεσμα είναι το δέντρο του Σχήματος 7.17. Παρατηρούμε ότι ο πρώτος καλύτερος διαχωρισμός των παραδειγμάτων γίνεται με βάση το φύλο. Η τιμή *male* για το χαρακτηριστικό **Gender** οδηγεί σε έναν τελικό κόμβο (φύλλο), στον οποίο κατατάσσονται 493 πελάτες, από τους οποίους οι 439 (89%) είναι πιστοί και οι 54 (11%) με πιθανότητα απώλειας. Ο κόμβος αυτός θεωρείται αρκετά ομοιόμορφος και για αυτό δε διασπάται περεταίρω, αλλά χαρακτηρίζεται ως φύλλο που αντιστοιχεί στο αποτέλεσμα **πιστός πελάτης (loyal)**. Η τιμή *female*, αντιθέτως, δεν οδηγεί σε τελικό αποτέλεσμα αλλά διασπάται με βάση την ηλικία. Το χαρακτηριστικό **Ηλικία (Age)** είναι ποσοτικό, που σημαίνει ότι ο διαχωρισμός δε γίνεται με βάση προκαθορισμένες τιμές, αλλά ο αλγόριθμος ανάπτυξης του δέντρου αναζητά το όριο που δίνει τον καλύτερο διαχωρισμό. Στο δέντρο του σχήματος παρατηρούμε ότι αν η ηλικία είναι μεγαλύτερη από 89.5 τότε ο πελάτης είναι πιστός, ενώ αν είναι μικρότερη, οδηγούμαστε σε νέα διάσπαση. Σχολιάζοντας αυτό το σημείο, σημειώνουμε ότι ο κόμβος στον οποίο οδηγεί η συνθήκη **ηλικία μεγαλύτερη από 89.5**, συμπεριλαμβάνει μόνο 2 παραδείγματα και, επομένως, δεν έχει ιδιαίτερη αξία. Με άλλα λόγια, παρόλο που ξέρουμε ότι όλες οι γυναίκες πελάτες ηλικίας μεγαλύτερης από 89.5 ετών είναι πιστές, αυτές είναι μόνο 2, που αντιστοιχούν σε μια μάλλον ασυνήθιστη περίπτωση. Από τεχνική άποψη θα λέγαμε ότι ένας τέτοιος κανόνας δεν είναι αξιόπιστος γιατί βασίζεται σε πολύ λίγες περιπτώσεις, ενώ από πρακτική άποψη θα λέγαμε ότι δεν είναι ιδιαίτερα χρήσιμος, γιατί είναι απίθανο να ενδιαφερόμαστε για τόσο ηλικιωμένους πελάτες. Η επόμενη διάσπαση είναι και πάλι με βάση την ηλικία και δείχνει ότι οι γυναίκες ηλικίας μεγαλύτερης από 30.5 (αλλά μικρότερης από 89.5) είναι πιθανό να αλλάξουν εταιρεία, επειδή από τις 287 αυτής της κατηγορίας, οι 229 (80%) έχουν αλλάξει.

Η οπτική ερμηνεία του δέντρου υποβοηθείται από τις χρωματιστές ράβδους των τελικών κόμβων (φύλλων). Κάθε χρώμα αντιστοιχεί σε καθεμιά από τις τιμές του προβλεπόμενου χαρακτηριστικού (μπλε = πιστός, κόκκινο = αλλαγή). Οι αναλογίες των χρωμάτων στις ράβδους δείχνουν την ομοιομορφία ή όχι των κόμβων: όταν επικρατεί καθαρά (ή απόλυτα) κάποιο χρώμα, η κατάταξη ενός παραδείγματος σε αυτό το φύλλο αντιστοιχεί σε σημαντική πιθανότητα να είναι κάποιος πελάτης πιστός ή όχι, ενώ όταν έχουμε ανάμικτο αποτέλεσμα, η κατάταξη σε αυτό το φύλλο δεν έχει ιδιαίτερη αξία. Το πάχος της ράβδου δείχνει τον αριθμό των παραδειγμάτων που καταλήγουν σε αυτό το φύλλο. Όσο μεγαλύτερος είναι αυτός, τόσο πιο αξιόπιστος είναι ο αντίστοιχος κανόνας. Επίσης, η κατασκευή του δέντρου μας πληροφορεί ότι η επίδραση των χαρακτηριστικών στην κατάταξη ενός πελάτη σε πιστό ή μη είναι μεγαλύτερη για το φύλο, στη συνέχεια για την ηλικία, την τελευταία συναλλαγή και μικρότερη για τον τρόπο πληρωμής.



Σχήμα 7.17*. Το δέντρο κατάταξης που προκύπτει από την εκτέλεση της διαδικασίας*

Το μοντέλο μπορεί να προβληθεί και σε μορφή περιγραφής, μέσω της καρτέλας Description, όπως στο Σχήμα 7.18. Τέλος, η διαδικασία μπορεί να αποθηκευτεί, ώστε να μπορεί να εφαρμοστεί σε νέα δεδομένα.



Σχήμα 7.18. *Το δέντρο κατάταξης σε μορφή κειμένου.*

7.3.1.6 Εφαρμογή του μοντέλου σε άγνωστα δεδομένα

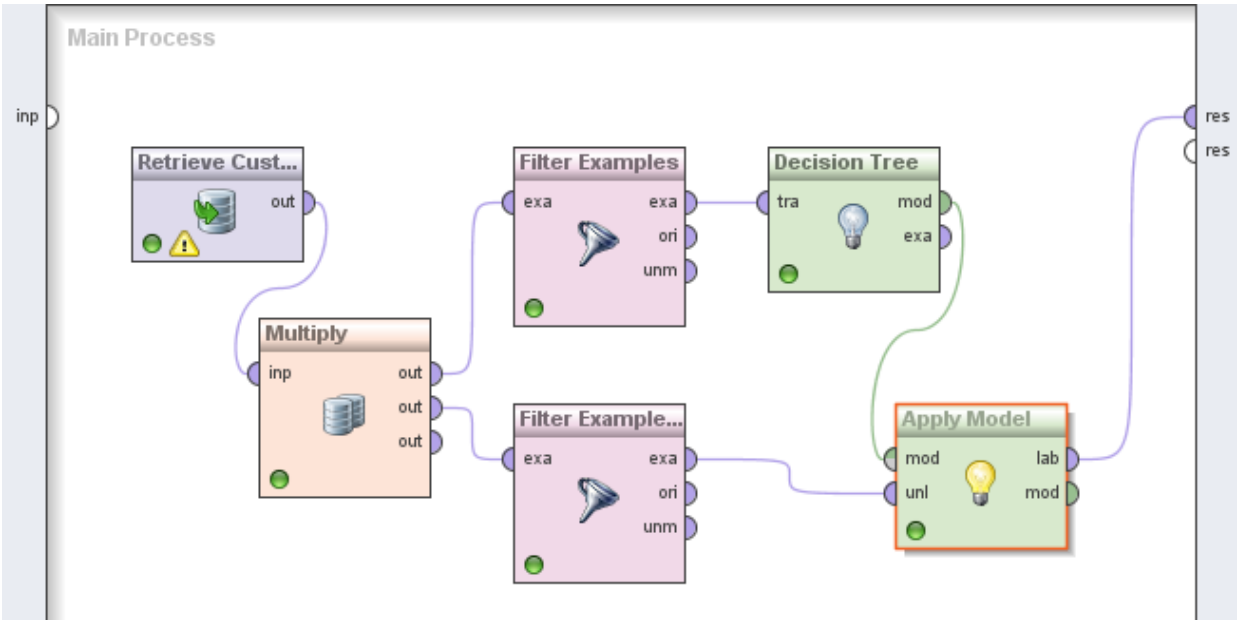
Το δέντρο μπορεί να χρησιμοποιηθεί για την κατάταξη ενός άγνωστου πελάτη, αν το τροφοδοτήσουμε με τα χαρακτηριστικά του πελάτη και ακολουθήσουμε τις διακλαδώσεις του δέντρου μέχρι κάποιο φύλλο. Η λειτουργία αυτή μπορεί να γίνει για μεγάλο αριθμό παραδειγμάτων με τη βοήθεια μιας διαδικασίας.

Ο τελεστής που χρησιμοποιείται για την εφαρμογή του μοντέλου είναι ο **Apply Model**, που βρίσκεται στην κατηγορία **Model Application**. Ο τελεστής αυτός δέχεται ως είσοδο στη θύρα *mod* (model) το μοντέλο που παράγει ο **Decision Tree** και τα άγνωστα δεδομένα, δηλ. χωρίς το χαρακτηριστικό με ρόλο label, στη θύρα *unl* (unlabeled).

Τα άγνωστα δεδομένα μπορεί να βρίσκονται σε ένα εξωτερικό αρχείο ή κάπου αποθηκευμένα στο Αποθετήριο. Στο παράδειγμα αυτό, θα χρησιμοποιήσουμε για την επίδειξη της εφαρμογής του μοντέλου τα παραδείγματα του αρχικού σετ δεδομένων για τα οποία είναι άγνωστη η τιμή του χαρακτηριστικού Churn (είναι αυτά που αποκλείστηκαν από την εκπαίδευση στη φάση της μοντελοποίησης). Για το σκοπό αυτό, μπορούμε να δημιουργήσουμε ένα αντίγραφο των δεδομένων εισόδου και να επιλέξουμε με το κατάλληλο φίλτρο τα παραδείγματα για τα οποία η τιμή **Churn** είναι κενή.

Για την υλοποίηση της εφαρμογής του μοντέλου, πραγματοποιούμε τις ακόλουθες προσθήκες στη διαδικασία μοντελοποίησης: Εισάγουμε τον τελεστή **Multiply**, και συνδέουμε στην είσοδό του τα δεδομένα που παρέχει ο **Retrieve**. Ο **Multiply** παρέχει στην έξοδό του αναλλοίωτα τα δεδομένα που δέχεται στην είσοδό του, αλλά προσφέρει περισσότερες από μια θύρες εξόδου, ώστε να δημιουργεί ανεξάρτητα αντίγραφα των δεδομένων. Η μία έξοδος συνδέεται με την είσοδο του **Filter Examples**, ακριβώς όπως συνδεόταν προηγουμένως ο **Retrieve**, ώστε τα δεδομένα αυτά να χρησιμοποιηθούν για την ανάπτυξη του δέντρου. Η δεύτερη έξοδος του **Multiply** οδηγείται σε ένα νέο τελεστή **Filter Examples**, που προσθέτουμε στη διαδικασία, με σκοπό την επιλογή των άγνωστων δεδομένων στα οποία θα εφαρμοστεί το μοντέλο.

Η συνολική διαδικασία φαίνεται στο Σχήμα 7.19. Η θύρα εξόδου *mod* (model) του **Decision Tree** συνδέεται με τη θύρα *mod* του **Apply Model**, ενώ η είσοδος *unl* του **Apply Model** συνδέεται με την έξοδο του 2ου **Filter Examples** (αυτού που επιλέγει τα άγνωστα παραδείγματα). Η θύρα εξόδου *lab* (labels) του **Apply Model** οδηγείται στην έξοδο της διαδικασίας.



Σχήμα 7.19. *Η διαδικασία εφαρμογής του μοντέλου στην πρόβλεψη άγνωστων παραδειγμάτων*

Πατώντας εκτέλεση, μπορούμε να δούμε τα αποτελέσματα της εφαρμογής του δέντρου κατάταξης στα άγνωστα παραδείγματα του αρχείου customers churn (Σχήμα 7.20). Τα αποτελέσματα της πρόβλεψης εμφανίζονται στη στήλη prediction (Churn). Οι τιμές στη στήλη confidence (Churn) δείχνουν το ποσοστό εμπιστοσύνης της κάθε κατάταξης (είναι η πιθανότητα να είναι σωστή η πρόβλεψη, όπως εκτιμήθηκε κατά τη φάση εκπαίδευσης).

ExampleSet (96 examples, 4 special attributes, 4 regular attributes) Filter (96 /

Row No.	Churn	prediction(Churn)	confidence(loyal)	confidence(churn)	Gender	Age	Payment M...	LastTransa...
1	?	churn	0.202	0.798	female	39	credit card	177
2	?	churn	0.202	0.798	female	53	cash	183
3	?	churn	0.202	0.798	female	33	credit card	194
4	?	churn	0.202	0.798	female	71	credit card	27
5	?	loyal	0.890	0.110	male	81	cash	153
6	?	churn	0.202	0.798	female	54	cheque	146
7	?	loyal	0.890	0.110	male	63	credit card	102
8	?	churn	0.202	0.798	female	58	credit card	176
9	?	churn	0.202	0.798	female	45	credit card	150
10	?	churn	0.202	0.798	female	33	credit card	144
11	?	loyal	0.890	0.110	male	40	credit card	82
12	?	loyal	0.890	0.110	male	36	credit card	91
13	?	churn	0.202	0.798	female	72	credit card	158
14	?	churn	0.202	0.798	female	66	cash	199
15	?	loyal	0.890	0.110	male	17	cheque	138
16	?	loyal	0.839	0.161	female	30	credit card	137
17	?	loyal	0.890	0.110	male	55	cheque	128
18	?	loyal	0.839	0.161	female	18	credit card	117
19	?	loyal	0.890	0.110	male	27	cash	130
20	?	churn	0.202	0.798	female	75	credit card	207
21	?	churn	0.202	0.798	female	51	credit card	142

Σχήμα 7.20. *Τα αποτελέσματα της πρόβλεψης εμφανίζονται στη στήλη prediction(Churn)*

7.3.1.7 Αξιολόγηση του μοντέλου

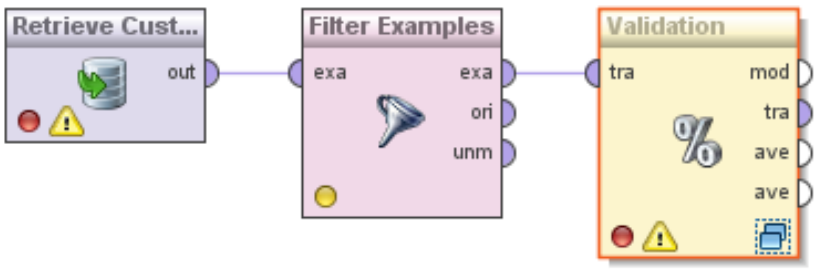
Μετά τη δημιουργία του μοντέλου, είναι σκόπιμο να εκτιμηθεί η ακρίβειά του, δοκιμάζοντας τις ικανότητές του στην κατάταξη πελατών. Για το σκοπό αυτό, μπορούμε να εφαρμόσουμε το μοντέλο σε ένα σετ δεδομένων παρόμοιο με αυτό που χρησιμοποιήθηκε στην εκπαίδευση. Το χαρακτηριστικό-στόχος **Churn** δε θα δοθεί ως είσοδος στο μοντέλο, αλλά θα προβλεφθεί αυτόματα από το τελευταίο. Η τιμή που θα προβλεφθεί για το **Churn** μπορεί, στη συνέχεια, να συγκριθεί με την πραγματική τιμή, ώστε να μετρηθεί η ακρίβεια της πρόβλεψης.

Σημείωση: τα σετ δεδομένων που χρησιμοποιούνται για τον έλεγχο της ακρίβειας του μοντέλου ονομάζονται σετ δεδομένων ελέγχου, σε αντιδιαστολή με τα δεδομένα εκπαίδευσης. Και στις δύο περιπτώσεις πρέπει να γνωρίζουμε το σωστό αποτέλεσμα, ώστε να μπορούμε να κατευθύνουμε την εκπαίδευση ή να ελέγξουμε το αποτέλεσμα, αντίστοιχα. Επομένως, τα δεδομένα αυτά πρέπει να αποτελούνται από γνωστά παραδείγματα.

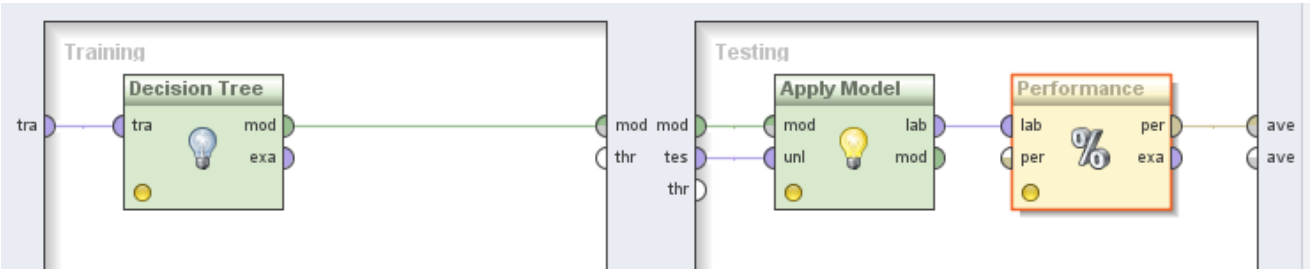
Για να είναι αξιόπιστη η αξιολόγηση του μοντέλου, θα πρέπει αυτό να δοκιμαστεί σε δεδομένα άγνωστα σε αυτό, δηλαδή που δε χρησιμοποιήθηκαν κατά τη φάση εκπαίδευσης. Η τυποποιημένη και γενικά αποδεκτή ως αξιόπιστη μέθοδος για μια τέτοια αξιολόγηση είναι η μέθοδος cross validation. Η βασική αρχή της είναι: το αρχικό σετ δεδομένων χωρίζεται τυχαία σε μικρά τμήματα (συνήθως μεγέθους 10% του συνόλου), η εκπαίδευση πραγματοποιείται χρησιμοποιώντας όλο το σετ εκτός από ένα τέτοιο τμήμα και στη συνέχεια αξιολογείται στο τμήμα αυτό. Η διαδικασία εκπαίδευσης στο 90% των δεδομένων και αξιολόγησης στο 10% επαναλαμβάνεται για όλα τα τμήματα του σετ δεδομένων και τα αποτελέσματα όλων των δοκιμών συνεκτιμούνται.

Η αξιολόγηση πραγματοποιείται ως ακολούθως:

- Δημιουργούμε μια νέα διαδικασία.
- Εισάγουμε, ρυθμίζουμε και συνδέουμε τους τελεστές **Retrieve** και **Filter Examples**, ακριβώς όπως και στη διαδικασία μοντελοποίησης, ώστε να πραγματοποιείται ανάγνωση των δεδομένων customer churn και να αφαιρούνται τα παραδείγματα με άγνωστο το χαρακτηριστικό **Churn.**
- Εισάγουμε τον τελεστή **X-Validation** και τον συνδέουμε στην έξοδο *exa* του **Filter Examples**. Αφήνουμε τις προκαθορισμένες ρυθμίσεις των παραμέτρων (Σχήμα 7.21).
- Κάνουμε διπλό κλικ στο γαλάζιο εικονίδιο στο κάτω δεξιά μέρος του **Validation**. Η παρουσία του εικονιδίου αυτού δείχνει ότι ο τελεστής περιέχει στο εσωτερικό του μια υποδιαδικασία που μπορούμε να καθορίσουμε. Στη συγκεκριμένη περίπτωση, πρέπει να καθοριστούν δύο υποδιαδικασίες, αυτή της εκμάθησης και αυτή του ελέγχου.
- Αφού ανοίξει η υποδιαδικασία του τελεστή **Validation**, στο αριστερό μέρος (**Training**) εισάγουμε τον τελεστή **Decision Tree** και στο δεξί μέρος (**Testing**) τους τελεστές **Apply Model** και **Performance** (Σχήμα 7.22). Συνδέουμε όπως φαίνεται στο Σχήμα. Φαίνεται ότι η έξοδος του Validation είναι η μέτρηση της απόδοσης από τον τελεστή **Performance.** Ρυθμίζοντας τις παραμέτρους του τελευταίου, μπορούμε να επιλέξουμε τους στατιστικούς δείκτες που μας ενδιαφέρουν.
- Εκτελώντας τη διαδικασία, παίρνουμε το αποτέλεσμα του Σχήματος 7.23. Η συνολική ακρίβεια πρόβλεψης (μετρημένη για παραδείγματα που ήταν άγνωστα κατά την εκπαίδευση) εκτιμήθηκε ως 83,89%. Ο πίνακας σύγχυσης (confusion matrix) δείχνει χωριστά για κάθε κατηγορία τον αριθμό των παραδειγμάτων που προβλέφθηκαν σωστά.



Σχήμα 7.21. Η πρώτη φάση σχεδιασμού της διαδικασίας αξιολόγησης.



Σχήμα 7.22. *Οι υποδιαδικασίες του τελεστή Validation.*

accuracy: 83.89% +/- 3.49% (mikro: 83.89%)

	true loyal	true churn	class precision
pred. loyal	509	76	87.01%
pred. churn	69	246	78.10%
class recall	88.06%	76.40%	

Σχήμα 7.23. *Τα αποτελέσματα της αξιολόγησης του μοντέλου με τη μέθοδο cross validation.*

Ο πίνακας σύγχυσης του Σχήματος 7.23 μας δίνει την ευαισθησία και ακρίβεια της πρόβλεψης για κάθε τύπο πελάτη. Από τους πελάτες που πραγματικά άλλαξαν εταιρεία (322), προβλέφθηκαν σωστά οι 246, ενώ οι υπόλοιποι 76 προβλέφθηκαν λανθασμένα ως πιστοί. Το ποσοστό 246/322=76,4% δείχνει την **ευαισθησία** της πρόβλεψης των πελατών που μπορεί να αλλάξουν εταιρεία. Από τους 315 πελάτες για τους οποίους προβλέφθηκε ότι μπορεί να αλλάξουν εταιρεία, για τους 246 η πρόβλεψη ήταν σωστή, ενώ για τους 69 η πρόβλεψη ήταν λάθος, επειδή στην πραγματικότητα ήταν πιστοί. Το ποσοστό 246/315=78,1% δείχνει την **ακρίβεια** της πρόβλεψης. Η ευαισθησία μετράει το πόσοι από τους «επίφοβους» πελάτες τελικά εντοπίζονται, ενώ η ακρίβεια δείχνει το πόσοι από αυτούς που εντοπίσαμε ήταν πραγματικά επίφοβοι να αλλάξουν εταιρεία. Οι δύο αυτοί δείκτες δείχνουν συνολικά την απόδοση του μοντέλου. Συχνά μπορούμε με κατάλληλη ρύθμιση των παραμέτρων να βελτιώσουμε τον έναν δείκτη σε βάρος του άλλου π.χ. να κάνουμε την πρόβλεψη πιο «ευαίσθητη» ώστε να ξεφεύγουν ελάχιστοι «επίφοβοι» πελάτες, με συνέπεια όμως να χαρακτηρίζουμε ως «επίφοβους» και πολλούς πιστούς πελάτες. Αντίστροφα, μπορούμε να ρυθμίσουμε το μοντέλο ώστε να είναι πιο ακριβές, δηλαδή να είμαστε περισσότερο σίγουροι ότι οι πελάτες που εντοπίστηκαν ως επίφοβοι είναι πράγματι έτσι, ακόμα και αν είναι πολλοί αυτοί που δεν εντοπίζονται. Στο συγκεκριμένο παράδειγμα, δεν είναι έκδηλο το αν είναι προτιμότερο για την εταιρεία να δοθεί έμφαση στην ευαισθησία ή την ακρίβεια, κάτι που εξαρτάται από το αν είναι μεγαλύτερο το κόστος των ενεργειών διατήρησης ενός πελάτη που θα ήταν έτσι και αλλιώς πιστός ή αυτό της απώλειας ενός πελάτη.

7.3.2 Ανάλυση καλαθιού αγορών

7.3.2.1 Ορισμός προβλήματος

Το πρόβλημα της ανάλυσης καλαθιού αγορών είναι ένα τυπικό πρόβλημα εξαγωγής γνώσης από δεδομένα για την υποβοήθηση της διοίκησης και του μάρκετινγκ. Μπορεί να εφαρμοστεί σε πολλές παραλλαγές και σε διαφορετικά πεδία, όμως στην ενότητα αυτή θα παρουσιάσουμε το πρόβλημα στην πιο αντιπροσωπευτική μορφή του, που είναι και αυτή από την οποία πήρε η συγκεκριμένη ανάλυση το όνομά της.

Κάθε φορά που ένας πελάτης κάποιου καταστήματος λιανικής ολοκληρώνει τις αγορές του, στο ταμείο καταγράφονται τα είδη που αγόρασε, οι αντίστοιχες ποσότητες και αξίες, καθώς και άλλα στοιχεία που αφορούν τη συναλλαγή, όπως η ημερομηνία/ώρα, το κατάστημα, τυχόν «πόντοι» έκπτωσης, κ.ά. Αν μάλιστα ο πελάτης χρησιμοποιεί και κάρτα πιστότητας, καταγράφεται και ο κωδικός του πελάτη, που σημαίνει ότι όλες οι αγορές του μπορούν να συνδεθούν με αυτόν. Ο συνολικός όγκος των δεδομένων που προκύπτει κατά τη λειτουργία π.χ. ενός σούπερ μάρκετ είναι συνήθως τεράστιος. Μέσα σε αυτά τα δεδομένα, έχει ενδιαφέρον να εντοπίσουμε ποια είδη εμφανίζονται στις αγορές των πελατών συχνά μαζί π.χ. αυτοί που αγοράζουν δημητριακά για πρωινό, αγοράζουν μαζί και γάλα; Η γνώση αυτή μπορεί να χρησιμοποιηθεί με πολλούς τρόπους από έναν μαρκετίστα, τόσο στο σχεδιασμό προωθητικών ενεργειών ή την τιμολογιακή πολιτική, όσο και στη χωρική τοποθέτηση ή προβολή του κάθε προϊόντος. Τα σύνολα από είδη που εμφανίζονται συχνά μαζί ονομάζονται συχνά σύνολα (frequent itemsets) και μπορούν να οδηγήσουν σε κανόνες του τύπου «Αν ο πελάτης αγοράσει δημητριακά, έχει πιθανότητες 70% να αγοράσει και γάλα», που ονομάζονται κανόνες συσχέτισης (association rules).

Στο πρόβλημα που θα εξετάσουμε, θεωρούμε ένα συνοικιακό κατάστημα λιανικής πώλησης τροφίμων (μίνι μάρκετ). Διαθέτουμε ως δεδομένα τα στοιχεία που καταγράφηκαν για κάθε δοσοληψία του τελευταίου μήνα. Αγνοούμε οποιαδήποτε επιπλέον πληροφορία και επικεντρωνόμαστε στο ποια είδη εμφανίστηκαν στο κάθε «καλάθι» αγορών, με στόχο τον εντοπισμό κανόνων συσχέτισης, που θα μας δείξουν ποια είδη τείνουν οι πελάτες μας να αγοράζουν ταυτόχρονα. Το σετ δεδομένων που θα χρησιμοποιηθεί είναι

εκπαιδευτικού χαρακτήρα και αποτελείται από ~10.000 συναλλαγές λιανικών πωλήσεων. Τα δεδομένα που διατίθενται για κάθε συναλλαγή είναι το ποια προϊόντα αγοράστηκαν στη συναλλαγή αυτή. Εμφανίζονται συνολικά περίπου 150 διαφορετικά προϊόντα, τα οποία αναφέρονται με ένα απλό περιγραφικό όνομα στα Αγγλικά (π.χ. coffee, butter, beef, κλπ.). Σημειώνεται ότι σε μια πραγματική εφαρμογή θα περιμέναμε πολύ μεγαλύτερο αριθμό προϊόντων, το καθένα από τα οποία θα αντιστοιχούσε σε κάτι πολύ πιο συγκεκριμένο (συγκεκριμένη μάρκα, τύπος και συσκευασία).

Το σετ δεδομένων διατίθεται σε μορφή αρχείου MS-Excel μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/groceries.xlsx)

7.3.2.2 Σχεδιασμός

Το πρόβλημα που περιγράφηκε παραπάνω, διαφέρει αρκετά από το πρόβλημα κατάταξης της ενότητας 3.1, αφού δεν υπάρχουν παραδείγματα με βάση τα οποία θα πρέπει να μάθει το μοντέλο μας να προβλέπει, ούτε γνωστές κατηγορίες στις οποίες θέλουμε να εντάξουμε τα παραδείγματα. Επίσης δεν πρόκειται για πρόβλημα συσταδοποίησης/τμηματοποίησης, αφού δε μας ενδιαφέρει να εντοπίσουμε ομάδες παραδειγμάτων. Το πρόβλημα ανήκει στην κατηγορία της εξαγωγής κανόνων συσχέτισης, κάτι που εμπεριέχει το πρόβλημα της εύρεσης συνόλων αντικειμένων (και όχι παραδειγμάτων). Το πρόβλημα είναι σχετικά απλό στη λογική και λύνεται σε 2 βήματα:

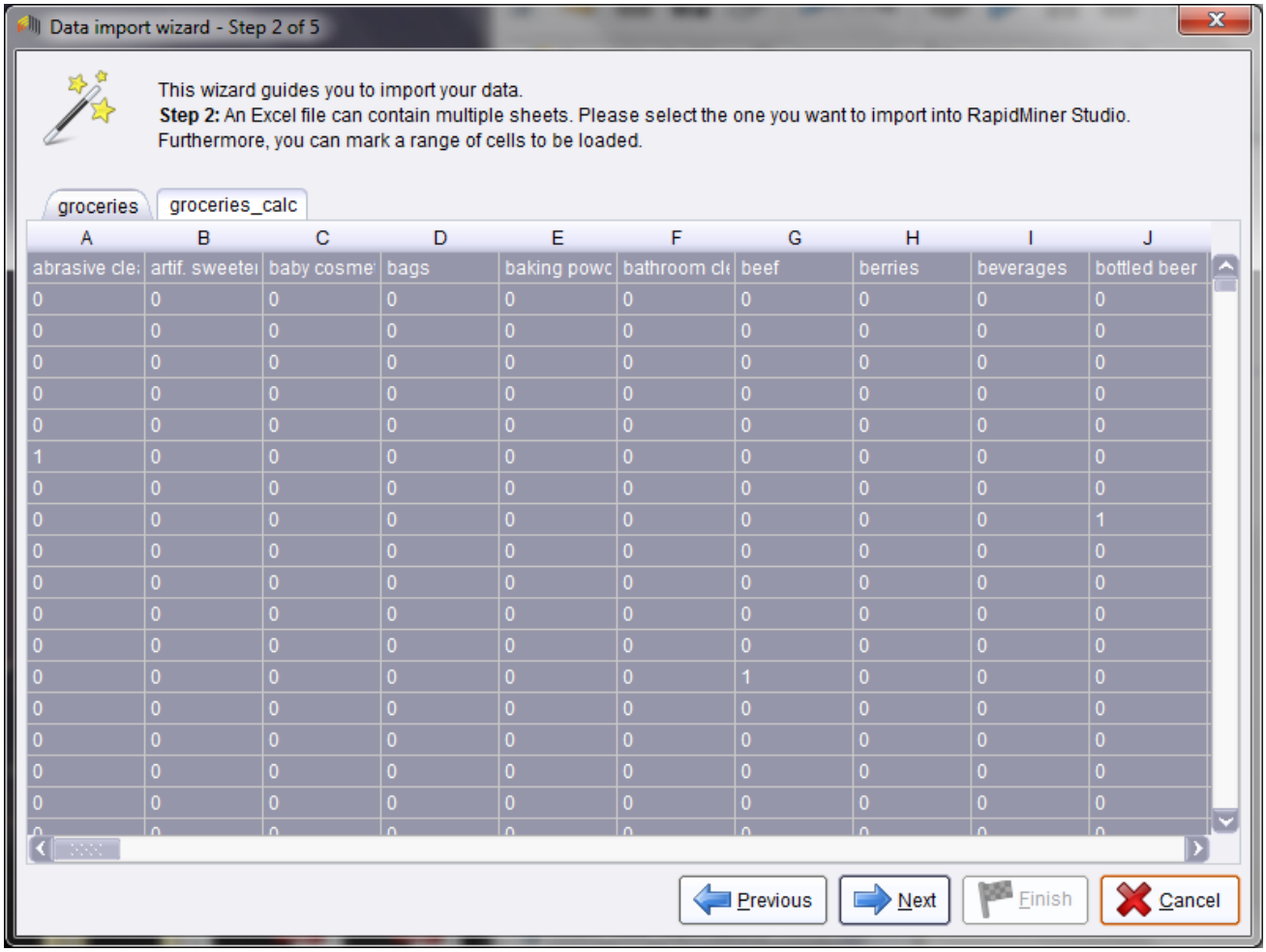
7. Εύρεση συχνών συνόλων, κάτι που μπορεί να γίνει απλά μετρώντας σε πόσα καλάθια εμφανίζεται ο κάθε συνδυασμός ειδών, και επιλέγοντας τους συνδυασμούς που εμφανίζονται συχνότερα από κάποιο όριο.
8. Δημιουργία κανόνων, όπου, για κάθε συχνό σύνολο, κάποια είδη των συνόλου χρησιμοποιούνται ως υπόθεση και τα υπόλοιπα ως πρόβλεψη.

Σημειώνεται ότι η κύρια πρόκληση στο πρόβλημα αυτό είναι οι μεγάλες ανάγκες σε υπολογιστική ισχύ και μνήμη, ιδιαίτερα όσο αυξάνει το μέγεθος του συνόλου δεδομένων. Ο γνωστότερος και απλούστερος, ίσως, αλγόριθμος εύρεσης συχνών συνόλων είναι ο Apriori, ενώ πιο αποτελεσματικός είναι ο FP-Growth. Το RapidMiner διαθέτει τελεστή για την εφαρμογή του FP-Growth, που είναι και η συνιστώμενη λύση, ιδιαίτερα για χρήστες που εργάζονται σε προσωπικό υπολογιστή, αφού έχει τις μικρότερες απαιτήσεις σε μνήμη και ταχύτητα επεξεργασίας. Επίσης προσφέρεται ο τελεστής **Create Association Rules** για την εξαγωγή κανόνων από συχνά σύνολα και ο **Apply Association Rules** για την εφαρμογή και επανεκτίμηση της εμπιστοσύνης των εξαχθέντων κανόνων σε νέο σύνολο δεδομένων.

7.3.2.3 Εισαγωγή και προσαρμογή των δεδομένων

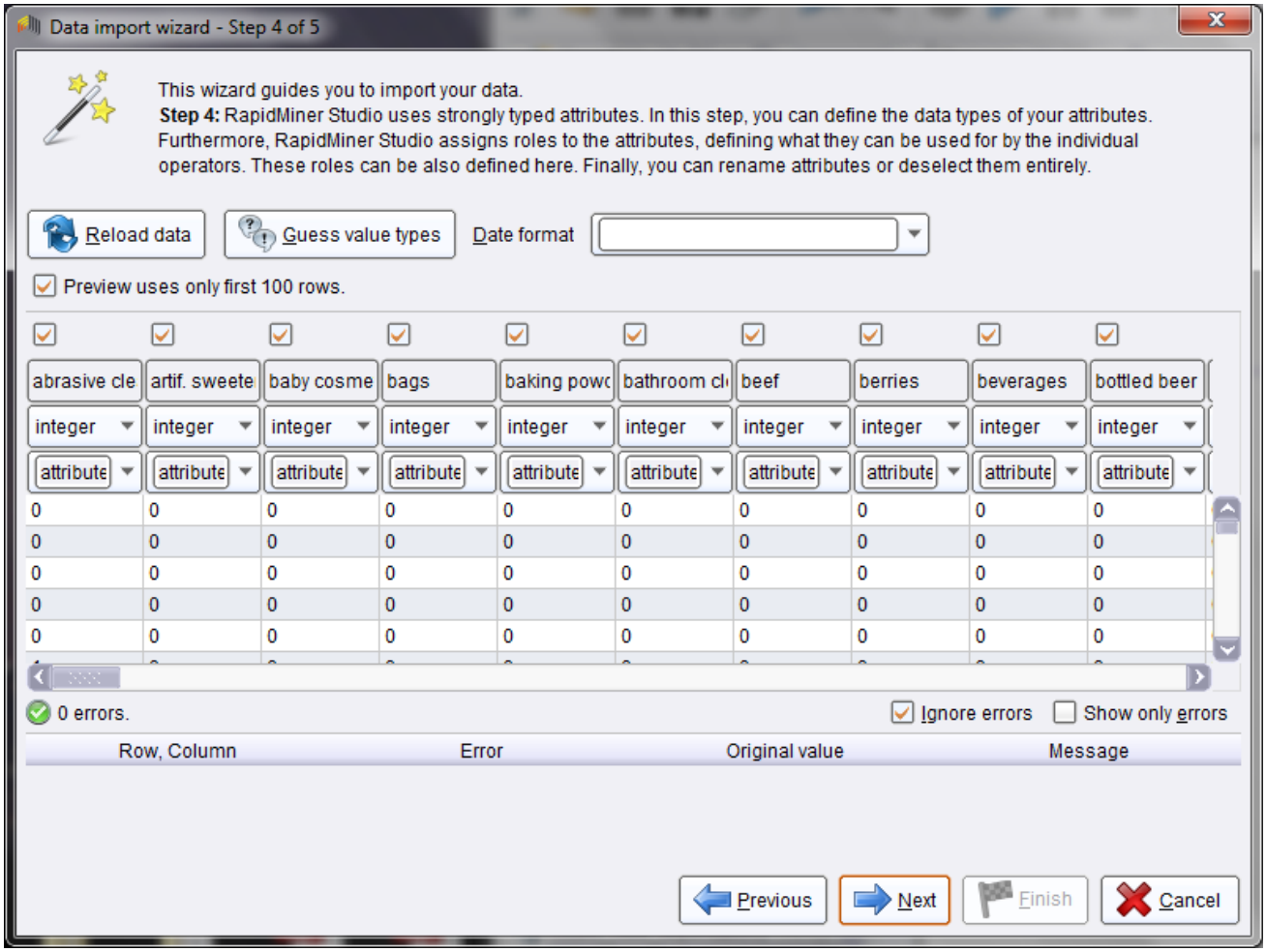
Τα δεδομένα εισόδου είναι διαθέσιμα σε μορφή φύλλου εργασίας Excel, όπου κάθε γραμμή αντιστοιχεί σε μια δοσοληψία και κάθε χαρακτηριστικό σε ένα προϊόν. Το περιεχόμενο κάθε κελιού είναι «1» αν στη δοσοληψία της γραμμής περιλαμβάνεται το προϊόν της στήλης και «0» αν δεν περιλαμβάνεται. Επειδή τα προϊόντα που διαθέτει το κατάστημα αναμένεται να είναι πολλά, αντίστοιχα μεγάλος θα είναι και ο αριθμός των στηλών του αρχείου εισόδου και, επομένως, και ο αριθμός των χαρακτηριστικών. Επίσης, επειδή ο αριθμός των προϊόντων που περιέχονται σε κάθε καλάθι αναμένεται να είναι πολύ μικρότερος του συνολικού αριθμού των προϊόντων του καταστήματος, κάθε γραμμή των δεδομένων θα έχει πολύ λίγα «1» και πάρα πολλά «0».

Για την εισαγωγή των δεδομένων από το Excel μπορούμε να χρησιμοποιήσουμε τον οδηγό εισαγωγής δεδομένων του RapidMiner. Στο πρώτο βήμα επιλέγουμε το αρχείο εισόδου, που στο παράδειγμα αυτό έχει το όνομα **Shopping_Groceries.xlsx**. Στο 2^{ο} βήμα, εμφανίζεται μια προεπισκόπηση των δεδομένων και, αν στο αρχείο Excel περιλαμβάνονται περισσότερα από ένα φύλλα εργασίας, μπορούμε να επιλέξουμε το σωστό (Σχήμα 7.24). Παρατηρούμε ότι κάθε στήλη-χαρακτηριστικό έχει το όνομα ενός προϊόντος και κάθε κελί περιέχει την τιμή *0* ή *1*, ανάλογα με την παρουσία ή όχι του προϊόντος της στήλης στη δοσοληψία της γραμμής.



Σχήμα 7.24. *Προεπισκόπηση δεδομένων εισόδου και επιλογή φύλλου εργασίας.*

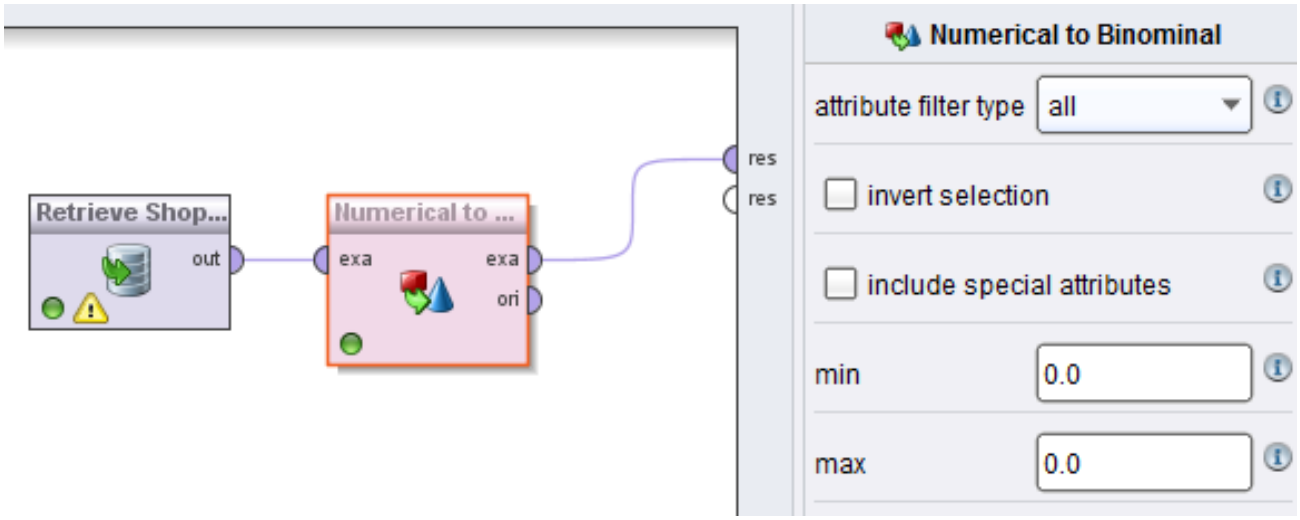
Στο επόμενο βήμα μπορούμε να καθορίσουμε τα ονόματα των χαρακτηριστικών. Στο παράδειγμα αυτό δε χρειάζεται να κάνουμε κάποια ενέργεια, αφού η πρώτη γραμμή ερμηνεύτηκε σωστά ως τα ονόματα των χαρακτηριστικών. Στο 4^{ο} βήμα (Σχήμα 7.25) καθορίζουμε τον τύπο δεδομένων και το ρόλο του κάθε χαρακτηριστικού. Όλα τα χαρακτηριστικά αναγνωρίστηκαν από το RapidMiner ως ακέραιοι και θεωρήθηκαν απλά χαρακτηριστικά (attribute). Γνωρίζουμε ότι το περιεχόμενο όλων των χαρακτηριστικών είναι λογικού τύπου (binominal), αφού το *0* ή *1* δεν εκφράζει κάποια ποσότητα, αλλά την παρουσία ή όχι κάποιου προϊόντος. Επειδή όμως η αλλαγή της επιλογής από integer σε binominal για όλα τα χαρακτηριστικά απαιτεί κόπο που μπορούμε να αποφύγουμε, είναι προτιμότερο να αφήσουμε την επιλογή integer και να πραγματοποιήσουμε αργότερα τη μετατροπή σε binominal. Στο τελευταίο βήμα, επιλέγουμε το όνομα και τη θέση στο αποθετήριο όπου θα αποθηκευτούν τα δεδομένα. Επιλέξαμε το όνομα “Shopping_basket” και το φάκελο data εντός του Local Repository.



Σχήμα 7.25*. Ο ορισμός του τύπου δεδομένων και του ρόλου κάθε χαρακτηριστικού.*

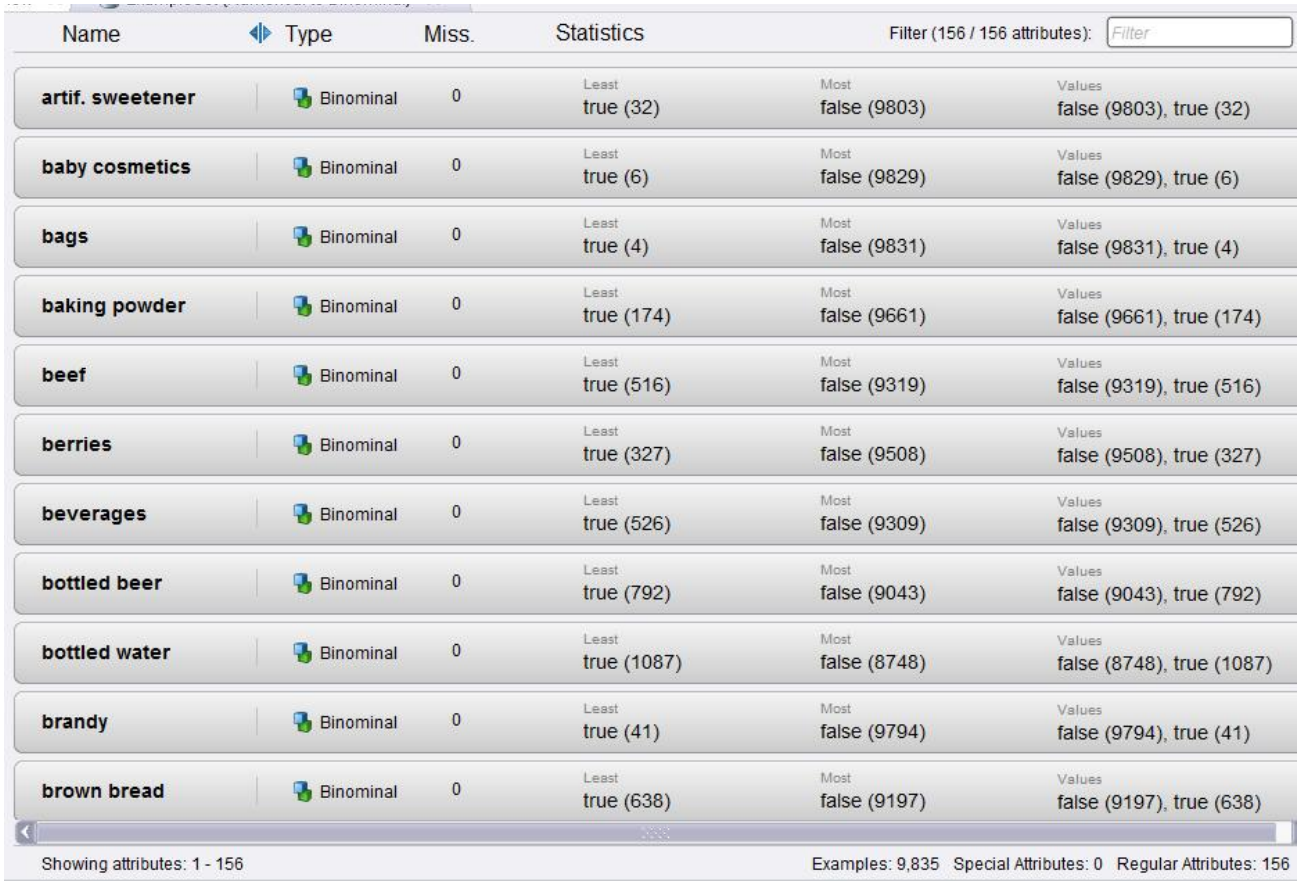
7.3.2.4 Επισκόπηση των δεδομένων

Για να παρατηρήσουμε τα δεδομένα, είναι σκόπιμο να τα εισάγουμε σε μια νέα διαδικασία και να τα μετατρέψουμε από αριθμητικά σε λογικά, ώστε το RapidMiner να εμφανίσει τα σωστά στατιστικά (συχνότητες και όχι μέσες τιμές). Η διαδικασία για την εισαγωγή και προετοιμασία των δεδομένων φαίνεται στο Σχήμα 7.26. Ο τελεστής **Retrieve** διαβάζει το σετ δεδομένων Shopping_basket από το Αποθετήριο και το τελεστής **Numerical to Binominal** μετατρέπει τα δεδομένα από integer σε binominal.



Σχήμα 7.26. *Εισαγωγή και τροποποίηση του τύπου των δεδομένων.*

Ο **Numerical to Binominal** βρίσκεται στην κατηγορία **Data Transformation**, στο φάκελο **Type Conversion**. Δεν απαιτείται καμία τροποποίηση στις παραμέτρους του, επειδή είναι προκαθορισμένο να εκλαμβάνει ως άρνηση (*false*) το *0* και ως κατάφαση (*true*) οποιαδήποτε άλλη αριθμητική τιμή, που στην περίπτωσή μας είναι το *1*. Η έξοδος του **Numerical to Binominal** οδηγείται προσωρινά στην έξοδο της διαδικασίας, ώστε να μπορέσουμε να παρατηρήσουμε τα δεδομένα στο στάδιο αυτό.



Σχήμα 7.27. *Προβολή στατιστικών στοιχείων για τα χαρακτηριστικά.*

Στο Σχήμα 7.27 παρατηρούμε τα στατιστικά στοιχεία για τα δεδομένα. Στο σετ δεδομένων περιλαμβάνονται 156 χαρακτηριστικά (δηλαδή προϊόντα) και 9835 παραδείγματα (δηλαδή δοσοληψίες). Μπορούμε επίσης να παρατηρήσουμε τη συχνότητα αγοράς του κάθε προϊόντος, π.χ. το artificial sweetener (τεχνητό γλυκαντικό) εμφανίζεται σε 32 καλάθια από τα 9835, το baby cosmetics (μωρουδιακά καλλυντικά) μόλις σε 6, ενώ το bottled water (εμφιαλωμένο νερό) σε 1087.

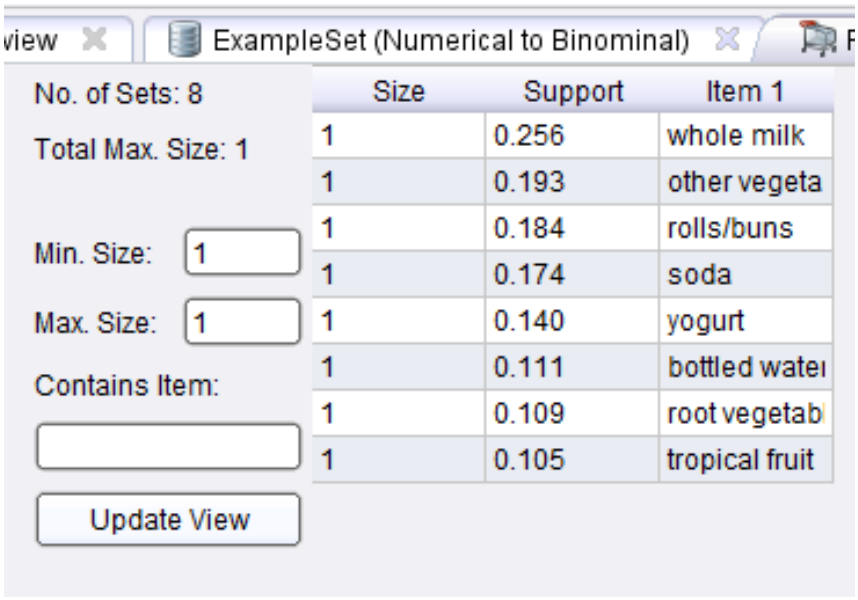
7.3.2.5 Μοντελοποίηση

Ο κατάλληλος τελεστής για την εύρεση των συχνών συνόλων (Frequent itemsets) είναι ο **FP-Growth**, που βρίσκεται στην κατηγορία **Modeling**, στο φάκελο **Association and Item Set Mining**. Ο τελεστής αυτός υπολογίζει όλα τα συχνά σύνολα σε ένα σύνολο παραδειγμάτων, χρησιμοποιώντας για το σκοπό αυτό μια δενδροειδή δομή. Τα γράμματα FP στην ονομασία του αντιστοιχούν στις λέξεις Frequency Pattern (πρότυπα συχνότητας) . Όλα τα συχνά σύνολα κωδικοποιούνται σε ένα δέντρο, το οποίο είναι ιδιαίτερα συμπυκνωμένη μορφή παράστασής τους. Πλεονέκτημα του **FP-Growth**, συγκριτικά με μη δενδροειδείς αλγορίθμους όπως ο γνωστός **Apriori** (βλέπε Κεφάλαιο 6), είναι οι μικρότερες ανάγκες σε μνήμη και επεξεργαστική ισχύ, ώστε να είναι δυνατή η επεξεργασία ακόμα και πολύ μεγάλων συνόλων δεδομένων. Όλα τα χαρακτηριστικά των δεδομένων εισόδου του **FP-Growth** είναι υποχρεωτικό να είναι τύπου binominal.

Οι σημαντικότερες από τις παραμέτρους του FP-Growth είναι:

- **Positive value**. Προσδιορίζει την τιμή που εκλαμβάνεται ως θετική, δηλαδή ως παρουσία του προϊόντος. Αν μείνει κενό, ισχύει αυτό που καθορίζεται από τα ίδια τα δεδομένα. Στην περίπτωσή μας, τα δεδομένα μετατράπηκαν σε binominal προσέχοντας να ισχύει η τιμή 1 ως θετική, επομένως δε χρειάζεται να επανακαθοριστεί το 1 ως positive value. (**Προσοχή:** αν για κάποιο λόγο αντιστραφούν το θετικό με το αρνητικό, τα αποτελέσματα θα είναι λανθασμένα).
- **Min support**. Καθορίζει την ελάχιστη υποστήριξη που πρέπει να διαθέτει κάθε συχνό σύνολο για να συμπεριληφθεί στα αποτελέσματα. Ως υποστήριξη ορίζεται το ποσοστό των παραδειγμάτων στα οποία εμφανίζεται (π.χ. αν ένας συνδυασμός προϊόντων εμφανίζεται μόνο σε 10 καλάθια, η υποστήριξη για αυτό το εύρημα είναι 10/9835 = 0,001 ή 0,1%.).
- Η επιλογή **find min mumber of itemsets.** Όταν η επιλογή αυτή είναι ενεργοποιημένη, το τελεστής μειώνει διαδοχικά το όριο υποστήριξης, ώστε να εξάγει τουλάχιστον έναν αριθμό συχνών συνόλων, που ορίζεται στην παράμετρο **Min number of itemsets**. Όταν η επιλογή δεν είναι ενεργοποιημένη, εξάγονται μόνο τα συχνά σύνολα που ικανοποιούν το όριο ελάχιστης υποστήριξης, όσο λίγα και αν είναι αυτά. Η λειτουργία με ενεργοποιημένη την επιλογή εύρεσης ελάχιστου αριθμού συχνών συνόλων είναι ίσως πρακτική, γιατί μας δίνει αποτέλεσμα χωρίς να πειραματιζόμαστε οι ίδιοι με το όριο υποστήριξης.

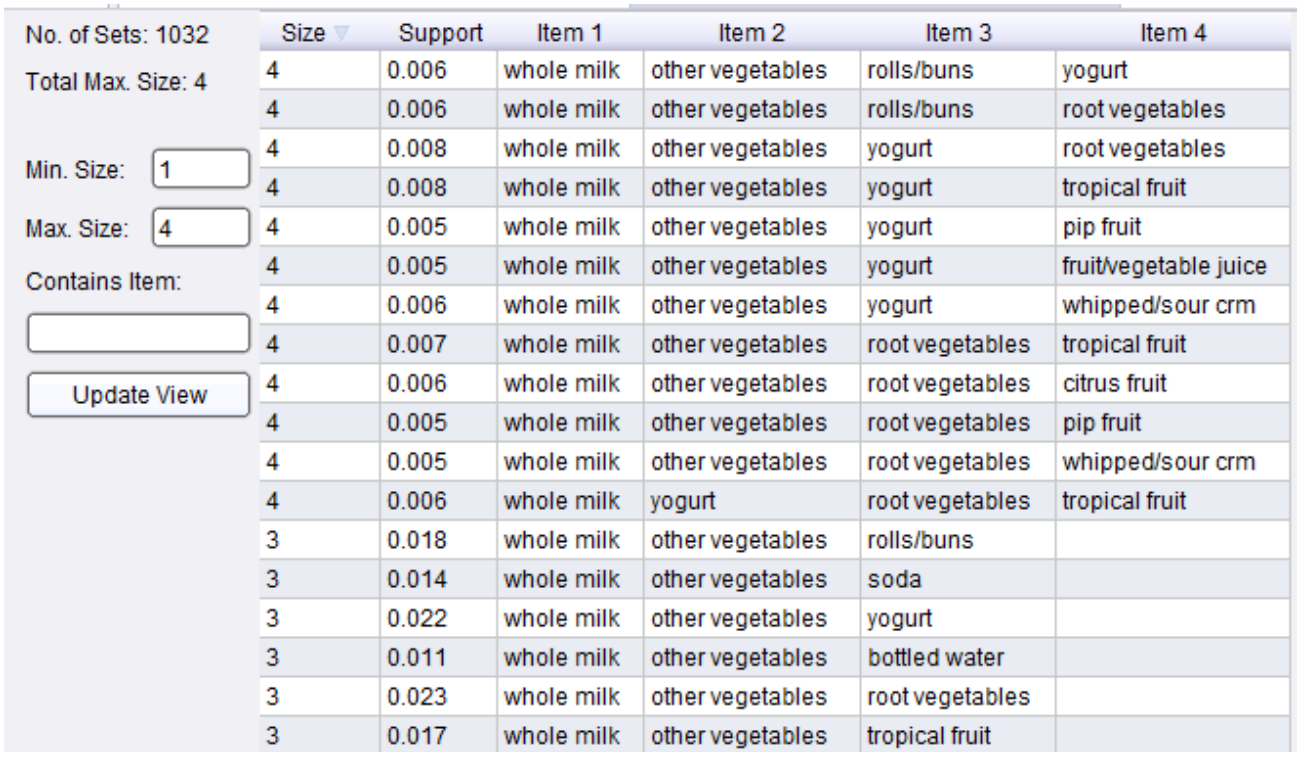
Η ρύθμιση των παραμέτρων του **FP-Growth** γίνεται με πειραματισμό και επαναληπτικές διορθωτικές ενέργειες. Για το σκοπό αυτό, μπορούμε να οδηγήσουμε την έξοδο του **FP-Growth** στην έξοδο της διαδικασίας, ώστε να παρατηρήσουμε τα συχνά σύνολα που θα βρεθούν στο σετ δεδομένων. Η προκαθορισμένη τιμή για την ελάχιστη υποστήριξη είναι 0,95. Δοκιμάζοντας να εκτελέσουμε τη διαδικασία με την τιμή αυτή και χωρίς να ενεργοποιήσουμε το **find min number of itemsets**, παίρνουμε το αποτέλεσμα **no itemsets found** (δε βρέθηκε κανένα συχνό σύνολο). Προφανώς το όριο είναι υπερβολικά υψηλό, κάτι που θα έπρεπε να περιμένουμε, αφού είναι πολύ απίθανο να εμφανίζονται κάποια σύνολα ειδών στο 95% των καλαθιών! Δοκιμάζοντας ως όριο για το min support την τιμή 0.1, παίρνουμε το αποτέλεσμα του σχήματος 7.28. Παρατηρούμε ότι προκύπτουν ελάχιστα σύνολα ενός μόνο είδους, που είναι φυσικά άχρηστα.



Size	Support	Item 1
1	0.256	whole milk
1	0.193	other vegeta
1	0.184	rolls/buns
1	0.174	soda
1	0.140	yogurt
1	0.111	bottled wate
1	0.109	root vegetab
1	0.105	tropical fruit

Σχήμα 7.28. *Τα συχνά σύνολα για ελάχιστη υποστήριξη 0.1 είναι ελάχιστα και αποτελούνται μόνο από 1 στοιχείο.*

Κατεβάζοντας το όριο στο 0.005, εξάγεται ικανοποιητικός αριθμός από συχνά σύνολα, με μέγιστο μέγεθος 4 στοιχεία. Το όριο υποστήριξης 0.005 σημαίνει ότι για να χαρακτηριστεί ένα σύνολο συχνό, αρκεί να βρεθεί σε περίπου 50 περιπτώσεις από τα 9.835 παραδείγματα. Αν μειωθεί και άλλο το όριο, παίρνουμε περισσότερα σύνολα με ακόμα μεγαλύτερο μέγεθος, των οποίων όμως η υποστήριξη είναι υπερβολικά μικρή.



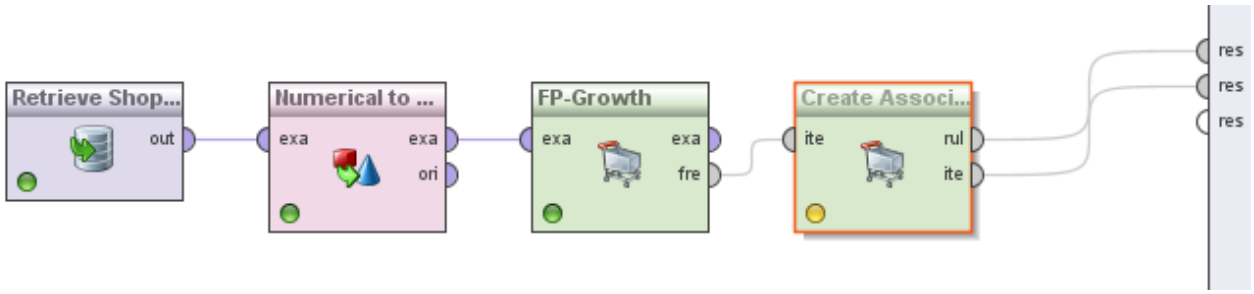
Size	Support	Item 1	Item 2	Item 3	Item 4
4	0.006	whole milk	other vegetables	rolls/buns	yogurt
4	0.006	whole milk	other vegetables	rolls/buns	root vegetables
4	0.008	whole milk	other vegetables	yogurt	root vegetables
4	0.008	whole milk	other vegetables	yogurt	tropical fruit
4	0.005	whole milk	other vegetables	yogurt	pip fruit
4	0.005	whole milk	other vegetables	yogurt	fruit/vegetable juice
4	0.006	whole milk	other vegetables	yogurt	whipped/sour crm
4	0.007	whole milk	other vegetables	root vegetables	tropical fruit
4	0.006	whole milk	other vegetables	root vegetables	citrus fruit
4	0.005	whole milk	other vegetables	root vegetables	pip fruit
4	0.005	whole milk	other vegetables	root vegetables	whipped/sour crm
4	0.006	whole milk	yogurt	root vegetables	tropical fruit
3	0.018	whole milk	other vegetables	rolls/buns	
3	0.014	whole milk	other vegetables	soda	
3	0.022	whole milk	other vegetables	yogurt	
3	0.011	whole milk	other vegetables	bottled water	
3	0.023	whole milk	other vegetables	root vegetables	
3	0.017	whole milk	other vegetables	tropical fruit	

Σχήμα 7.29. *Τα συχνά σύνολα με ελάχιστη υποστήριξη 0,005 (παρουσιάζονται ταξινομημένα κατά φθίνον μέγεθος)*

Η εύρεση των συχνών συνόλων μπορεί στη συνέχεια να οδηγήσει στην εύρεση κανόνων συσχέτισης (association rules). Οι κανόνες συσχέτισης βασίζονται στα συχνά σύνολα, αλλά η δομή τους μας δίνει περισσότερη πληροφορία και τους καθιστά χρησιμότερους σε άμεση εφαρμογή. Ενώ ένα συχνό σύνολο

αποτελείται από αντικείμενα που είναι συσχετισμένα μεταξύ τους επειδή εμφανίζονται συχνά μαζί, π.χ. γάλα και δημητριακά, οι κανόνες συσχέτισης έχουν ένα προηγούμενο (antecedent) και ένα επόμενο (consequent) ή, αλλιώς, μια υπόθεση και μια πρόβλεψη. Υπάρχουν περισσότεροι από έναν πιθανοί κανόνες που μπορεί να συνδέουν μεταξύ τους τα αντικείμενα ενός συχνού συνόλου, από τους οποίους κάποιος μπορεί να έχει μεγαλύτερο ποσοστό εμπιστοσύνης (confidence) και, επομένως, μεγαλύτερη αξία. Π.χ. το συχνό σύνολο **{γάλα, δημητριακά}** μπορεί να έχει ποσοστό εμφάνισης 10% , το **γάλα** μόνο του 50% και τα **δημητριακά** 15%, αν από τους 100 πελάτες, 50 αγοράζουν γάλα (ανεξάρτητα του αν αγοράζουν ή όχι δημητριακά), 15 αγοράζουν δημητριακά (ανεξάρτητα του αν αγοράζουν ή όχι γάλα) και επίσης, για 10 από τους προηγούμενους πελάτες ισχύει ότι αγοράζουν ταυτόχρονα γάλα και δημητριακά. Με βάση αυτές τις υποθετικές συχνότητες, ο κανόνας **γάλα → δημητριακά** θα έχει εμπιστοσύνη 20%, αφού από τους 50 που αγοράζουν γάλα, μόνο οι 10 αγοράζουν ταυτόχρονα και δημητριακά (10/50=20%), όμως ο κανόνας **δημητριακά → γάλα** έχει εμπιστοσύνη 67%, αφού από τους 15 που αγοράζουν δημητριακά, οι 10 αγοράζουν και γάλα. Επομένως, για κάποιον που αγοράζει δημητριακά μπορούμε να προβλέψουμε με μεγαλύτερη βεβαιότητα ότι θα αγοράσει και γάλα και, επομένως, ο αντίστοιχος κανόνας έχει μεγαλύτερη αξία, τόσο από τον αντίστροφο κανόνα, όσο και από τη γνώση του συχνού συνόλου.

Στο Σχήμα 7.30 φαίνεται η συνολική διαδικασία μοντελοποίησης. Τα παραδείγματα, μετά τη μετατροπή τους σε binominal, οδηγούνται στον **FP-Growth** για την εύρεση συχνών συνόλων. Η θύρα *fre* (frequent itemsets) οδηγείται στη θύρα εισόδου *ite* (items) του τελεστή **Create Association Rules**, ώστε, στη συνέχεια, τα συχνά σύνολα να μετατραπούν σε κανόνες συσχέτισης. Στην έξοδο της διαδικασίας *res* μπορούμε να οδηγήσουμε τόσο τους κανόνες, όσο και τα συχνά σύνολα.



Σχήμα 7.30. *Η συνολική διαδικασία εύρεσης κανόνων συσχέτισης*

Στις παραμέτρους του Create Association Rules μπορεί να καθοριστεί το κριτήριο επιλογής κανόνων και το αντίστοιχο όριο. Με τις παραμέτρους αυτούς, επιλέγονται οι σημαντικότεροι κανόνες και φιλτράρονται αυτοί που δε θα είχαν αξία. Το προκαθορισμένο και πιο συχνά χρησιμοποιούμενο κριτήριο είναι το ποσοστό εμπιστοσύνης των κανόνων (Confidence). Εισάγοντας ως Min confidence την τιμή 0,6, οι κανόνες που επιλέγονται ως έγκυροι είναι αυτοί που το συμπέρασμά τους είναι αληθινό τουλάχιστον στο 60% των περιπτώσεων εφαρμογής του κανόνα. Σημειώνεται ότι το Confidence αγνοεί την υποστήριξη, δηλαδή το πόσο συχνά είναι εφαρμόσιμος ο κανόνας. Τα υπόλοιπα κριτήρια που προσφέρονται (lift, conviction, ps, laplace) είναι πολύ χρήσιμα για την αξιολόγηση της αξίας των κανόνων, προτείνεται όμως να επιλεγεί ως κριτήριο το Confidence, ώστε να αποκλειστούν όλοι οι κανόνες που θα ήταν σίγουρα άχρηστοι λόγω της χαμηλής αξιοπιστίας τους. Στη συνέχεια, οι επιλεχθέντες κανόνες μπορούν να αξιολογηθούν από τον αναλυτή με βάση τις ανάγκες του, συνεκτιμώντας τις τιμές των υπολοίπων κριτηρίων. Η χρησιμότητα των κριτηρίων αυτών αναφέρεται στην Ενότητα 6.1.2 του Κεφαλαίου 6 και επιγραμματικά είναι:

- **Lift.** Εκφράζει το πόσο απέχουν τα συσχετιζόμενα αντικείμενα από την ανεξαρτησία και επομένως το πόσο πραγματικά επηρεάζει η υπόθεση το αποτέλεσμα. Τιμές κοντά στο 1 δείχνουν ανεξαρτησία των αντικειμένων, ενώ υψηλότερες τιμές δείχνουν μεγαλύτερο ενδιαφέρον του κανόνα.
- **Conviction.** Δείχνει το κατά πόσο περιέχεται χρήσιμη πληροφορία στη φορά του κανόνα, ώστε να εκτιμήσουμε ποιο είδος μέσα σε ένα σύνολο είναι αυτό που οδηγεί στην αγορά του άλλου.

- **Gain, Laplace και ps**. Ειδικότερα κριτήρια που σχετίζονται με εκτίμηση της ποσότητας πληροφορίας που περιέχεται στον κανόνα. Μεγαλύτερες τιμές δείχνουν ότι ο κανόνας περιέχει περισσότερη χρήσιμη πληροφορία.

No.	Premises	Conclusion	Support	Co...	LaPlace	Gain	p-s	Lift	Convic...
25	domestic eggs, curd	whole milk	0.005	0.725	0.998	-0.009	0.003	2.836	2.704
24	curd, butter	whole milk	0.005	0.714	0.998	-0.009	0.003	2.795	2.606
23	yogurt, root vegetables, tropical fruit	whole milk	0.006	0.700	0.998	-0.011	0.004	2.740	2.482
22	other vegetables, root vegetables, pip fruit	whole milk	0.005	0.675	0.997	-0.011	0.003	2.642	2.291
21	whipped/sour crm, butter	whole milk	0.007	0.660	0.997	-0.014	0.004	2.583	2.190
20	pip fruit, whipped/sour crm	whole milk	0.006	0.648	0.997	-0.013	0.004	2.537	2.117
19	yogurt, butter	whole milk	0.009	0.639	0.995	-0.020	0.006	2.500	2.062
18	root vegetables, butter	whole milk	0.008	0.638	0.995	-0.018	0.005	2.496	2.055
17	whole milk, root vegetables, citrus fruit	other vegetables	0.006	0.633	0.997	-0.013	0.004	3.273	2.200
16	other vegetables, yogurt, pip fruit	whole milk	0.005	0.625	0.997	-0.011	0.003	2.446	1.985
15	pip fruit, domestic eggs	whole milk	0.005	0.624	0.997	-0.012	0.003	2.440	1.978
14	tropical fruit, curd	whole milk	0.007	0.623	0.996	-0.016	0.004	2.437	1.974
13	tropical fruit, butter	whole milk	0.006	0.622	0.996	-0.014	0.004	2.436	1.972
12	domestic eggs, margarine	whole milk	0.005	0.622	0.997	-0.011	0.003	2.434	1.969
11	domestic eggs, butter	whole milk	0.006	0.621	0.996	-0.013	0.004	2.431	1.965
10	other vegetables, yogurt, tropical fruit	whole milk	0.008	0.620	0.995	-0.017	0.004	2.426	1.958
8	pip fruit, curd	whole milk	0.005	0.617	0.997	-0.011	0.003	2.416	1.945
9	other vegetables, yogurt, fruit/vegetable juice	whole milk	0.005	0.617	0.997	-0.011	0.003	2.416	1.945
7	whole milk, root vegetables, pip fruit	other vegetables	0.005	0.614	0.997	-0.012	0.004	3.171	2.087
5	tropical fruit, domestic eggs	whole milk	0.007	0.607	0.996	-0.016	0.004	2.376	1.895
6	other vegetables, root vegetables, whipped/sour	whole milk	0.005	0.607	0.997	-0.012	0.003	2.376	1.895
4	other vegetables, yogurt, root vegetables	whole milk	0.008	0.606	0.995	-0.018	0.005	2.373	1.891
3	pip fruit, whipped/sour crm	other vegetables	0.006	0.604	0.996	-0.013	0.004	3.124	2.039
2	bottled water, butter	whole milk	0.005	0.602	0.996	-0.013	0.003	2.357	1.872
1	root vegetables, onions	other vegetables	0.006	0.602	0.996	-0.013	0.004	3.112	2.027

Σχήμα 7.31. *Οι κανόνες που εξήχθησαν, ταξινομημένοι κατά φθίνουσα εμπιστοσύνη.*

Στο Σχήμα 7.31 παρουσιάζονται οι κανόνες που προέκυψαν επιλέγοντας ως κριτήριο το confidence με όριο το 0.6, ταξινομημένοι κατά φθίνουσα τιμή του confidence. Δημιουργήθηκαν 25 κανόνες από τους οποίους ο πρώτος λέει ότι:

domestic eggs (αυγά), curd (τυρόπηγμα) → whole milk (πλήρες γάλα)

Ο κανόνας έχει υποστήριξη 0.005, εμπιστοσύνη 0.725, lift 2.836 και conviction 2.704. Οι τιμές των κριτηρίων μας λένε ότι ο συνδυασμός αυγά με τυρόπηγμα έχει πιθανότητα να βρεθεί σε κάποιο καλάθι 0.5% (δηλαδή δεν είναι και πολύ συχνός), όταν όμως ισχύει, με πιθανότητα 72.5% ο πελάτης θα έχει αγοράσει και γάλα. Επίσης το ότι ο πελάτης αγόρασε αυγά και τυρόπηγμα κάνει την πιθανότητα να αγοράσει και γάλα 2.8 φορές μεγαλύτερη από την πιθανότητα να αγοράσει ένας οποιοσδήποτε πελάτης γάλα, ενώ η τιμή 2.7 του conviction δείχνει ότι η φορά του κανόνα έχει αξία, δηλαδή αν ξέραμε ότι κάποιος αγόρασε γάλα, δε θα γνωρίζαμε με την ίδια βεβαιότητα ότι θα αγοράσει και αυγά και τυρόπηγμα.

Ταξινομώντας τους κανόνες κατά φθίνουσα τιμή του lift (Σχήμα 7.32), παρατηρούμε ότι μεγαλύτερη προβλεπτική αξία έχει ο κανόνας 17:

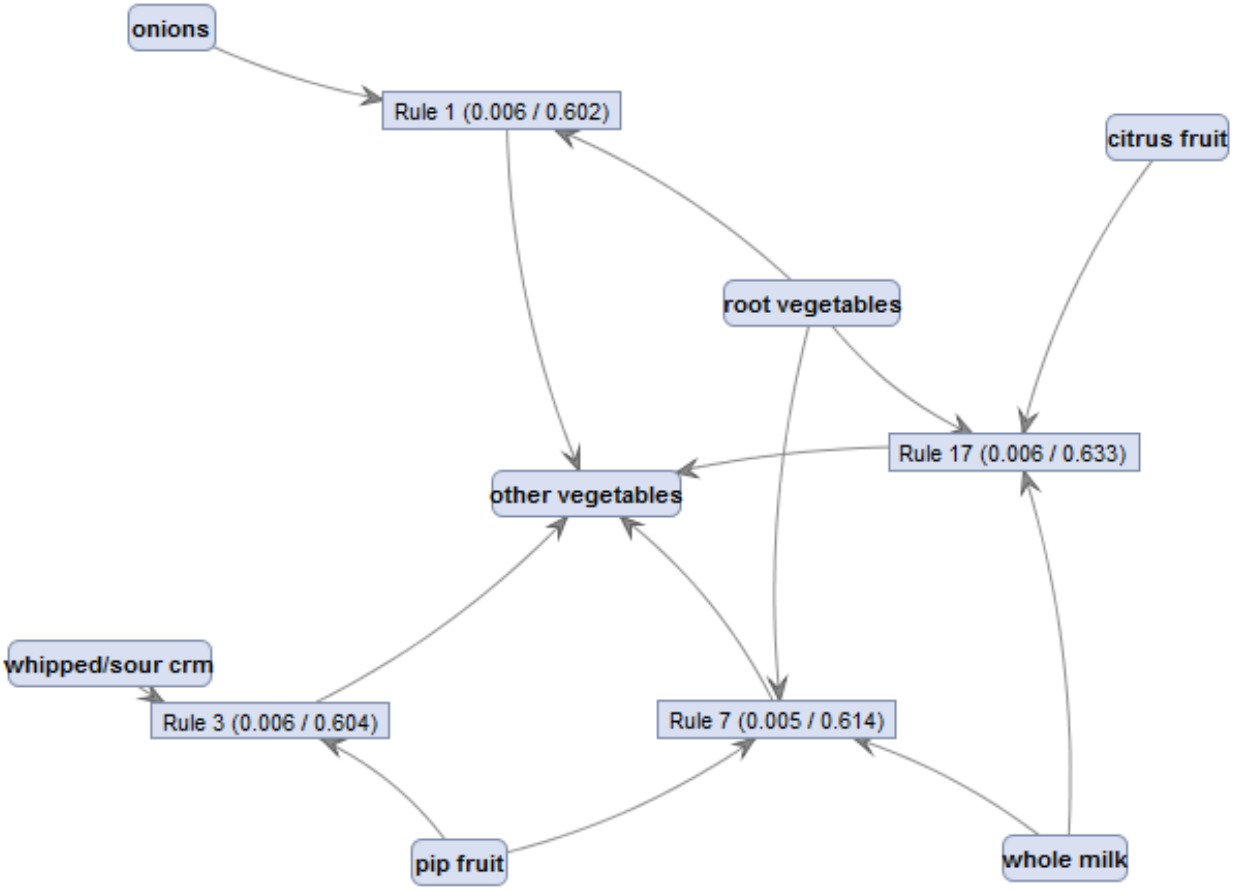
Whole milk (πλήρες γάλα), root vegetables (λαχανικά), citrus fruit (εσπεριδοειδή) → other vegetables (άλλα λαχανικά)

Ο κανόνας αυτός έχει υψηλότερο lift από τον προηγούμενο, αλλά με χαμηλότερο ποσοστό εμπιστοσύνης, δηλαδή προβλέπει κάτι πιο χρήσιμο, αλλά με μικρότερη βεβαιότητα.

No.	Premises	Conclusion	Support	Confid...	LaPlace	Gain	p-s	Lift	Convic...
17	whole milk, root vegetables, citrus fruit	other vegetables	0.006	0.633	0.997	-0.013	0.004	3.273	2.200
7	whole milk, root vegetables, pip fruit	other vegetables	0.005	0.614	0.997	-0.012	0.004	3.171	2.087
3	pip fruit, whipped/sour crm	other vegetables	0.006	0.604	0.996	-0.013	0.004	3.124	2.039
1	root vegetables, onions	other vegetables	0.006	0.602	0.996	-0.013	0.004	3.112	2.027
25	domestic eggs, curd	whole milk	0.005	0.725	0.998	-0.009	0.003	2.836	2.704
24	curd, butter	whole milk	0.005	0.714	0.998	-0.009	0.003	2.795	2.606
23	yogurt, root vegetables, tropical fruit	whole milk	0.006	0.700	0.998	-0.011	0.004	2.740	2.482
22	other vegetables, root vegetables, pip fruit	whole milk	0.005	0.675	0.997	-0.011	0.003	2.642	2.291
21	whipped/sour crm, butter	whole milk	0.007	0.660	0.997	-0.014	0.004	2.583	2.190
20	pip fruit, whipped/sour crm	whole milk	0.006	0.648	0.997	-0.013	0.004	2.537	2.117
19	yogurt, butter	whole milk	0.009	0.639	0.995	-0.020	0.006	2.500	2.062
18	root vegetables, butter	whole milk	0.008	0.638	0.995	-0.018	0.005	2.496	2.055
16	other vegetables, yogurt, pip fruit	whole milk	0.005	0.625	0.997	-0.011	0.003	2.446	1.985
15	pip fruit, domestic eggs	whole milk	0.005	0.624	0.997	-0.012	0.003	2.440	1.978
14	tropical fruit, curd	whole milk	0.007	0.623	0.996	-0.016	0.004	2.437	1.974
13	tropical fruit, butter	whole milk	0.006	0.622	0.996	-0.014	0.004	2.436	1.972
12	domestic eggs, margarine	whole milk	0.005	0.622	0.997	-0.011	0.003	2.434	1.969
11	domestic eggs, butter	whole milk	0.006	0.621	0.996	-0.013	0.004	2.431	1.965
10	other vegetables, yogurt, tropical fruit	whole milk	0.008	0.620	0.995	-0.017	0.004	2.426	1.958
8	pip fruit, curd	whole milk	0.005	0.617	0.997	-0.011	0.003	2.416	1.945
9	other vegetables, yogurt, fruit/vegetable juice	whole milk	0.005	0.617	0.997	-0.011	0.003	2.416	1.945
5	tropical fruit, domestic eggs	whole milk	0.007	0.607	0.996	-0.016	0.004	2.376	1.895
6	other vegetables, root vegetables, whipped/sour	whole milk	0.005	0.607	0.997	-0.012	0.003	2.376	1.895
4	other vegetables, yogurt, root vegetables	whole milk	0.008	0.606	0.995	-0.018	0.005	2.373	1.891
2	bottled water, butter	whole milk	0.005	0.602	0.996	-0.013	0.003	2.357	1.872

Σχήμα 7.32. *Οι κανόνες συσχέτισης ταξινομημένοι κατά φθίνουσα τιμή του lift.*

Το RapidMiner διαθέτει και εργαλείο γραφικής παράστασης των κανόνων, με δυνατότητα επιλογής του στοιχείου στο οποίο θέλουμε να εστιάσουμε και διαδραστική επιλογή του κριτηρίου και ορίου με βάση τα οποία θέλουμε να επιλεγούν οι κανόνες που θα προβληθούν. Στο Σχήμα 7.33 παρουσιάζονται σε μορφή γραφήματος οι κανόνες που οδηγούν στο συμπέρασμα other vegetables (άλλα λαχανικά).



Σχήμα 7.33. *Οι κανόνες που οδηγούν στην πρόβλεψη αγοράς λαχανικών, σε μορφή γραφήματος.*

7.3.2.6 Εφαρμογή και αξιολόγηση του μοντέλου

Η διαδικασία που παρουσιάστηκε καταλήγει στην παραγωγή κανόνων, που εκφράζουν τη γνώση που αποκομίστηκε από την ανάλυση. Το ενδιαφέρον ενός στελέχους της επιχείρησης επικεντρώνεται στη μελέτη των ίδιων των κανόνων, που αποτελούν ένα μοντέλο των αγοραστικών συνηθειών των πελατών της επιχείρησης. Αντίθετα, η αυτοματοποιημένη εκτέλεση των κανόνων για την παραγωγή προβλέψεων δεν έχει κάποιο πρακτικό ενδιαφέρον στο πρόβλημα της παρούσας εφαρμογής, αφού δε μας ενδιαφέρει να εισάγουμε τα στοιχεία κάποιου καλαθιού και να προβλέψουμε τι ακόμα θα ήθελε ο συγκεκριμένος πελάτης (κάτι τέτοιο μπορεί να είχε ενδιαφέρον σε κάποια άλλη εφαρμογή π.χ. αν θέλαμε να συστήσουμε κάποια επιπλέον αγορά σε έναν πελάτη που παρήγγειλε ένα σύνολο ειδών). Η αξιολόγηση και αξιοποίηση του μοντέλου μπορεί να γίνει λοιπόν από ένα ειδικευμένο στέλεχος, μελετώντας τους κανόνες και κρίνοντας την αξία και εφαρμοσιμότητά τους. Μπορεί επίσης να πραγματοποιηθούν επιπλέον εκτελέσεις της διαδικασίας με διαφορετικές παραμέτρους, ώστε να «εξερευνηθούν» ευρύτερα οι συσχετίσεις που κρύβονται στα δεδομένα.

Μια εκτίμηση του αποτελέσματος του Σχήματος 7.30 θα ήταν ότι οι κανόνες που προέκυψαν είναι μάλλον λίγοι (25) και έχουν ως συμπέρασμα 2 μόνο προϊόντα, other vegetables και whole milk, ενώ ο συνολικός αριθμός προϊόντων του καταστήματος είναι μεγαλύτερος από 150. Οι περισσότεροι μάλιστα προβλέπουν την αγορά γάλακτος, που έτσι και αλλιώς βρίσκεται στο 25% των καλαθιών. Επίσης παρατηρούμε ότι η υποστήριξη των κανόνων είναι μεταξύ 0.5% και 0.9%, που σημαίνει ότι αφορούν λιγότερες από 1 στις 100 περιπτώσεις. Τι θα μπορούσαμε να κάνουμε για να εξάγουμε περισσότερα και χρησιμότερα ευρήματα; Μεταβάλλοντας κάποιες παραμέτρους, θα μπορούσαμε να προσαρμόσουμε καλύτερα τη διαδικασία στα δεδομένα. Μειώνοντας το όριο min support στον FP-Growth από 0.005 σε 0.002, εξάγονται πολύ περισσότερα συχνά σύνολα (για την ακρίβεια 4254, χωρίς τα μονομελή). Στη συνέχεια, ο Create Association Rules ανακαλύπτει 395 κανόνες που περνούν το όριο ελάχιστης εμπιστοσύνης του 0.6 (ίδιο με προηγουμένως). Διατάσσοντας τους κανόνες κατά φθίνον ποσοστό εμπιστοσύνης, λαμβάνουμε τα

αποτελέσματα του Σχήματος 7.34. Παρατηρούμε ότι παράγονται κανόνες με μεγαλύτερα ποσοστά εμπιστοσύνης και lift, όπως ο 395:

Whole milk (πλήρες γάλα), root vegetables (λαχανικά), tropical fruit (τροπικά φρούτα), citrus fruit (εσπεριδοειδή) → other vegetables (άλλα λαχανικά)

No.	Premises	Conclusion	Support	Co...	LaPla...	Gain	p-s	Lift	Convic...
395	whole milk, root vegetables, tropical fruit, citrus frui	other vegetables	0.003	0.886	1.000	-0.004	0.002	4.578	7.057
394	other vegetables, pork, butter	whole milk	0.002	0.846	1.000	-0.003	0.002	3.312	4.839
393	other vegetables, yogurt, root vegetables, fruit/vege	whole milk	0.002	0.833	1.000	-0.003	0.001	3.261	4.467
391	tropical fruit, herbs	whole milk	0.002	0.821	0.999	-0.003	0.002	3.215	4.169
392	other vegetables, yogurt, root vegetables, citrus frui	whole milk	0.002	0.821	0.999	-0.003	0.002	3.215	4.169
390	rolls/buns, yogurt, root vegetables, tropical fruit	whole milk	0.002	0.815	0.999	-0.003	0.002	3.189	4.020
386	rolls/buns, herbs	whole milk	0.002	0.800	0.999	-0.004	0.002	3.131	3.722
387	whole milk, tropical fruit, grapes	other vegetables	0.002	0.800	0.999	-0.003	0.002	4.135	4.033
388	yogurt, curd, butter	whole milk	0.002	0.800	0.999	-0.004	0.002	3.131	3.722
389	whole milk, yogurt, root vegetables, fruit/vegetable j	other vegetables	0.002	0.800	0.999	-0.003	0.002	4.135	4.033
385	other vegetables, yogurt, root vegetables, pip fruit	whole milk	0.002	0.793	0.999	-0.004	0.002	3.104	3.598
384	yogurt, root vegetables, butter	whole milk	0.003	0.789	0.999	-0.005	0.002	3.090	3.536
383	curd, hmburger mt	whole milk	0.003	0.788	0.999	-0.004	0.002	3.083	3.510
382	root vegetables, tropical fruit, citrus fruit	other vegetables	0.004	0.786	0.999	-0.007	0.003	4.061	3.764
381	other vegetables, domestic eggs, curd	whole milk	0.003	0.784	0.999	-0.005	0.002	3.067	3.443
380	root vegetables, tropical fruit, fruit/vegetable juice	other vegetables	0.003	0.781	0.999	-0.004	0.002	4.038	3.687
376	curd, herbs	whole milk	0.002	0.778	0.999	-0.003	0.001	3.044	3.350
377	whole milk, whipped/sour crm, onions	other vegetables	0.002	0.778	0.999	-0.003	0.002	4.020	3.629
378	rolls/buns, tropical fruit, beef	whole milk	0.002	0.778	0.999	-0.003	0.001	3.044	3.350
379	root vegetables, tropical fruit, sausage	whole milk	0.003	0.778	0.999	-0.004	0.002	3.044	3.350
375	other vegetables, root vegetables, brown bread	whole milk	0.003	0.775	0.999	-0.005	0.002	3.033	3.309
371	root vegetables, rice	whole milk	0.002	0.774	0.999	-0.004	0.002	3.030	3.297
372	whole milk, root vegetables, sliced cheese	other vegetables	0.002	0.774	0.999	-0.004	0.002	4.001	3.572
373	other vegetables, yogurt, tropical fruit, citrus fruit	whole milk	0.002	0.774	0.999	-0.004	0.002	3.030	3.297
374	whole milk, root vegetables, tropical fruit, pip fruit	other vegetables	0.002	0.774	0.999	-0.004	0.002	4.001	3.572
370	root vegetables, tropical fruit, domestic eggs	whole milk	0.003	0.771	0.999	-0.004	0.002	3.019	3.257
368	other vegetables, bottled beer, domestic eggs	whole milk	0.002	0.769	0.999	-0.003	0.001	3.010	3.226
369	root vegetables, citrus fruit, frozen vegetables	other vegetables	0.002	0.769	0.999	-0.003	0.002	3.976	3.495
364	root vegetables, tropical fruit, frozen vegetables	whole milk	0.002	0.767	0.999	-0.004	0.002	3.000	3.191
365	root vegetables, pip fruit, whipped/sour crm	whole milk	0.002	0.767	0.999	-0.004	0.002	3.000	3.191

Σχήμα 7.34. *Οι κανόνες που προκύπτουν μειώνοντας το όριο υποστήριξης κατά την εύρεση των συχνών συνόλων.*

Ένας παρατηρητικός αναγνώστης θα πρόσεξε ότι ο παραπάνω κανόνας 395 είναι παρόμοιος με τον κανόνα 17 του Σχήματος 7.30, με μόνη διαφορά ότι στο μέρος της υπόθεσης του πρώτου υπάρχει επιπλέον το είδος των τροπικών φρούτων. Ο νέος αυτός κανόνας έχει υψηλότερη εμπιστοσύνη (0.886 αντί 0.633), υψηλότερο lift (4.578 αντί 3.273), αλλά μικρότερη υποστήριξη (0.003 αντί 0.006). Συγκρίνοντας τους δύο κανόνες, από τη μία έχουμε έναν πιο γενικό κανόνα του οποίου η συνθήκη είναι πιο πιθανό να εμφανιστεί και από την άλλη έναν πιο ειδικό, που αντιστοιχεί σε σπανιότερη περίπτωση αλλά προβλέπει με μεγαλύτερη βεβαιότητα κάτι που έχει μεγαλύτερη αξία. Ο κανόνας 395 ουσιαστικά εμπεριέχεται στον 17, αφού αποτελεί ειδική περίπτωσή του. Σχολιάζοντας, λοιπόν, τα αποτελέσματα που προέκυψαν μετά την τροποποίηση του ορίου υποστήριξης των συχνών συνόλων, θα λέγαμε ότι διευρύναμε την εξαχθείσα γνώση σε πιο ειδικές περιπτώσεις. Το κατά πόσο αυτό είναι χρήσιμο θα κριθεί από αυτόν που προτίθεται να αξιοποιήσει τα αποτελέσματα στην πράξη. Π.χ. αν γνωρίζουμε ότι αυτός που αγοράζει γάλα, καρότα και λεμόνια έχει καλή

πιθανότητα να αγοράσει και άλλα λαχανικά, προσθέτει τίποτα το ότι αν αγοράσει και τροπικά φρούτα, τότε έχει ακόμα μεγαλύτερη πιθανότητα να αγοράσει άλλα λαχανικά; Σε μια πρώτη ματιά, η επιπλέον γνώση φαίνεται άχρηστη, δεν αποκλείεται όμως να έχουμε στα σχέδιά μας μια εκστρατεία προώθησης τροπικών φρούτων!

Παρατηρήσεις:

- Η διαδικασία εξαγωγής γνώσης, όπως αυτή του παραδείγματος, πρέπει να θεωρείται επαναληπτική. Μεταβάλλοντας κάποιες παραμέτρους ή κάποια στοιχεία της σχεδίασης της διαδικασίας, μπορεί να επιτύχουμε την ανάδυση επιπλέον χρήσιμης γνώσης.
- Στην περίπτωση που τα ευρήματα δε μας ικανοποιούν, ακόμα και μετά από κάποια προσπάθεια βελτίωσης της διαδικασίας, θα πρέπει ίσως να αναθεωρήσουμε το πρόβλημα και να επανασχεδιάσουμε εκ βάθρων τη σύλληψη. Τα αποτελέσματα του παραπάνω παραδείγματος μπορεί να αλλοιώνονται από γεγονότα που δεν καταγράφονται στα δεδομένα, όπως π.χ. αν στις προσφορές της ημέρας γίνεται τακτικά έκπτωση στα καρότα και στο γιαούρτι, μπορεί τα δύο αυτά προϊόντα να βρεθούν συσχετισμένα μεταξύ τους, όχι επειδή σχετίζονται πραγματικά, αλλά επειδή τυχαίνει να προωθούνται τις ίδιες ημέρες. Σε μια τέτοια περίπτωση θα χρειαζόταν επιπλέον πληροφορία και ενδεχομένως διαφορετική μέθοδος ανάλυσης.
- Θα πρέπει να έχουμε πάντα υπόψη ότι αυτό που αντιλαμβανόμαστε ως αποτυχία να εξάγουμε χρήσιμη γνώση μπορεί να οφείλεται στο ότι απλούστατα αυτό που ψάχνουμε δεν υπάρχει ή τουλάχιστον δεν προκύπτει από τα δεδομένα που διαθέτουμε. Αν π.χ. δεν παράγεται κανένας κανόνας που να συσχετίζει τον καφέ με κάποια άλλα προϊόντα, το πιθανότερο είναι ότι οι πελάτες μας στην πραγματικότητα αγοράζουν καφέ όταν τους τελειώνει και όχι σε συνδυασμό με κάτι άλλο.

7.3.3 Μελέτη των προσδοκιών των πελατών από το ξενοδοχείο τους

7.3.3.1 Ορισμός του προβλήματος

Η εφαρμογή αυτή έχει ως σκοπό τη μελέτη των προσδοκιών ή αναγκών των πελατών από το ξενοδοχείο τους. Το ενδιαφέρον εστιάζεται στην εύρεση συσχετίσεων ανάμεσα στα χαρακτηριστικά ή υπηρεσίες που μπορεί να προσφέρει ένα ξενοδοχείο, ώστε, γνωρίζοντας τις ανάγκες που εκφράζουν κάποιοι πελάτες, να προβλέπουμε τι άλλο είναι πιθανόν να βρίσκεται μέσα στις επιθυμίες τους. Ένας επιπλέον στόχος είναι η συσχέτιση των προσδοκιών των πελατών με άλλα χαρακτηριστικά τους, όπως χώρα ή ηλικία, και με χαρακτηριστικά του ταξιδιού όπως η διάρκεια, ο σκοπός και η κατηγορία κόστους.

Η εφαρμογή βασίζεται σε πραγματικά δεδομένα έρευνας ικανοποίησης πελατών από τις υπηρεσίες του ξενοδοχείου τους. Η έρευνα πραγματοποιήθηκε το καλοκαίρι του 2010 και απευθύνθηκε σε τουρίστες που επισκέφθηκαν παραλιακές περιοχές της Βορείου Ελλάδας. Το ερωτηματολόγιο σχεδιάστηκε για τη συλλογή πληροφορίας σχετικά με τις προσδοκίες και την ικανοποίηση των πελατών από το ξενοδοχείο τους και περιελάμβανε επίσης στοιχεία για το σκοπό και τη διάρκεια του ταξιδιού, καθώς και δημογραφικά στοιχεία των ερωτώμενων. Το σετ δεδομένων διατίθεται σε μορφή αρχείου MS-Excel μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/hotels_Northern_Greece.xlsx .

7.3.3.2 Σχεδιασμός

Τα δεδομένα που αφορούν τις προσδοκίες των πελατών από το ξενοδοχείο τους προέρχονται από αντίστοιχη ενότητα του ερωτηματολογίου, όπου ο ερωτώμενος καλείται να βαθμολογήσει σε μια 5-βάθμια κλίμακα το βαθμό συμφωνίας του με την πρόταση: «Οι προσδοκίες μου σχετικά με το ξενοδοχείο ήταν …», ακολουθούμενη από 33 διαφορετικά πιθανά στοιχεία του ξενοδοχείου. Τα στοιχεία για τα οποία καλείται ο ερωτώμενος να προσδιορίσει το κατά πόσο συμπεριλαμβάνονται στις προσδοκίες του είναι παροχές ή ποιοτικά χαρακτηριστικά, όπως εγκαταστάσεις για άτομα με ειδικές ανάγκες, παροχή internet, θυρίδα ασφαλείας, πισίνα για παιδιά, τοπική κουζίνα, ποιότητα ύπνου, εξυπηρετικό προσωπικό, καθαριότητα, σεσουάρ μαλλιών κλπ. Για καθένα από τα 33 στοιχεία, ο ερωτώμενος σημειώνει τη διαβάθμιση που τον

εκφράζει, ανάμεσα στην απόλυτη συμφωνία και την απόλυτη διαφωνία. Στον πίνακα 7.1 παρουσιάζεται ο πλήρης κατάλογος με τις προτεινόμενες προσδοκίες που περιλαμβάνονται στο ερωτηματολόγιο.

Οι προσδοκίες μου σχετικά με το ξενοδοχείο ήταν …						
		Συμφωνώ απόλυτα [5]	**Συμφωνώ [4]**	**Ούτε συμφωνώ/ ούτε διαφωνώ [3]**	**Διαφωνώ [2]**	**Διαφωνώ απόλυτα [1]**
1	Εγκαταστάσεις για άτομα με ειδικές ανάγκες					
2	Υπηρεσίες περιποίησης και χαλάρωσης					
3	Νυχτερινός φωτισμός					
4	Πάρκινγκ					
5	Παροχή internet					
6	Θυρίδα ασφαλείας στο δωμάτιο					
7	Πισίνα για παιδιά					
8	Τοποθεσία ξενοδοχείου/ προσβασιμότητα					
9	Ποιότητα και ποικιλία του φαγητού					
10	Να ανήκει σε αλυσίδα					
11	Μικρή απόσταση από παραλία					
12	Τοπική κουζίνα					
13	Σεσουάρ μαλλιών					
14	Κατάλληλο και εξυπηρετικό προσωπικό					
15	Ποιότητα ύπνου					
16	Τοπικό παραδοσιακό στυλ					
17	Υπηρεσία δωματίου					
18	Bar στο ξενοδοχείο					
19	Μενού ειδικής διατροφής					
20	Αύρα δωματίου					
21	Μέγεθος δωματίου					
22	Δραστηριότητες αναψυχής/ απασχόλησης					
23	Αθλητικές εγκαταστάσεις (γυμναστήριο, πισίνες, γήπεδα, κλπ.)					
24	Χώρους για κατοικίδια					
25	Οικογενειακού τύπου ξενοδοχείο					
26	Ευρύχωρο σαλόνι					
27	Τηλεόραση και ποικιλία καναλιών					
28	Ρυθμιστής θερμοκρασίας δωματίου					
29	Ασφάλεια					

30	Υπηρεσίες πλυσίματος-σιδερώματος ρούχων					
31	Καθαριότητα					
32	Αναλώσιμα μπάνιου (σαπούνια, σαμπουάν, πετσέτες, κ.ά.)					
33	Εστιατόριο στο ξενοδοχείο					

Πίνακας 7.1. *Οι προσδοκίες των πελατών από το ξενοδοχείο τους.*

Το πρόβλημα της εύρεσης συσχετίσεων ανάμεσα σε υπηρεσίες ή χαρακτηριστικά ενός ξενοδοχείου που μπορεί να επιθυμεί ένας πελάτης είναι παρόμοιο με αυτό της ανάλυσης καλαθιού αγορών, αφού και σε αυτήν την περίπτωση, αναζητούμε στοιχεία που εμφανίζονται με μεγάλη συχνότητα μαζί. Μια διαφορά του παρόντος προβλήματος από αυτό του καλαθιού αγορών είναι ότι, αντί της παρουσίας ή όχι ενός είδους σε ένα καλάθι, τα δεδομένα που χρησιμοποιούνται είναι ο βαθμός συμφωνίας ενός ερωτώμενου με το αν ισχύει ή όχι η προσδοκία για ένα στοιχείο, ο οποίος εκφράζεται σε 5-βάθμια κλίμακα. Για να διαμορφωθεί το πρόβλημα στη γνωστή μορφή της εύρεσης κανόνων συσχέτισης μέσα από συχνά σύνολα, απαιτείται η τροποποίηση των τιμών κλίμακας σε δυαδικές τιμές αληθές/ψευδές. Για το σκοπό αυτό, μπορούμε να θεωρήσουμε ως αληθές (δηλ την παρουσία της προσδοκίας) την απάντηση συμφωνώ (4) ή συμφωνώ απόλυτα (5), και ως ψευδές (δηλ απουσία της προσδοκίας) τις απαντήσεις ούτε συμφωνώ ούτε διαφωνώ (3), διαφωνώ (2), διαφωνώ απόλυτα (1). Για την τροποποίηση των δεδομένων μπορεί να χρησιμοποιηθεί το τελεστής **Map**, με τον οποίο μπορούμε να αντιστοιχίσουμε τις τιμές 1 ως 5 στις τιμές 0,1 ή true/false.

Στη συνέχεια, αναγόμαστε στο γνωστό πρόβλημα της εξόρυξης συνόλων αντικειμένων, που επιλύεται με τη χρήση των τελεστών **FP-Growth** και **Create Association Rules**. Στην εφαρμογή αυτή, ο αριθμός των προς συσχέτιση στοιχείων (33) είναι πολύ μικρότερος από τους χιλιάδες κωδικούς προϊόντων ενός καταστήματος λιανικής, που θα είχαμε να αντιμετωπίσουμε στο κλασικό πρόβλημα ανάλυσης καλαθιού αγορών. Ωστόσο, είναι αρκετά μεγάλος ώστε, σε συνδυασμό και με τη φύση του προβλήματος εύρεσης συσχετισμών, να καθιστά καταλληλότερη τη μέθοδο της εξόρυξης συνόλων αντικειμένων.

Σημείωση: Η εύρεση συσχετίσεων ανάμεσα στις απαντήσεις ερωτώμενων, και γενικότερα η ποσοτική ανάλυση δεδομένων πρωτογενών ερευνών με ερωτηματολόγιο, είναι μια τυπική κατηγορία προβλημάτων στατιστικής ανάλυσης, που μπορούν να επιλυθούν με διάφορες μεθόδους της κλασικής στατιστικής ή της πολυδιάστατης στατιστικής ανάλυσης. Οι μέθοδοι αυτές δεν παρουσιάζονται εδώ, γιατί είναι εκτός των πλαισίων του βιβλίου, σημειώνεται όμως ότι σε κάποιες περιπτώσεις αποτελούν προτιμότερη επιλογή. Γενικά ισχύει ότι οι στατιστικές μέθοδοι πλεονεκτούν σε αξιοπιστία και είναι προτιμότερες σε απλά και τυποποιημένα προβλήματα, όπου τα σετ δεδομένων δεν είναι ιδιαίτερα μεγάλου μεγέθους. Από την άλλη μεριά, μέθοδοι εξόρυξης, όπως η εύρεση συχνών συνόλων, μπορούν να αντιμετωπίσουν μεγάλα σύνολα δεδομένων και να βρουν συσχετίσεις μεταξύ ιδιαίτερα μεγάλου αριθμού στοιχείων, κάτι που είναι πρακτικά αδύνατο για τις γνωστές στατιστικές μεθόδους. Ένα ακόμα πλεονέκτημα των μεθόδων εξόρυξης είναι ότι προσφέρουν πολυποίκιλους τρόπους εφαρμογής και δυνατότητες πειραματισμού σε δεδομένα με περιορισμένη δομή και άγνωστα χαρακτηριστικά, ενώ το κύριο μειονέκτημά τους είναι ότι δεν εγγυώνται την αξιοπιστία και τη χρησιμότητα του αποτελέσματος.

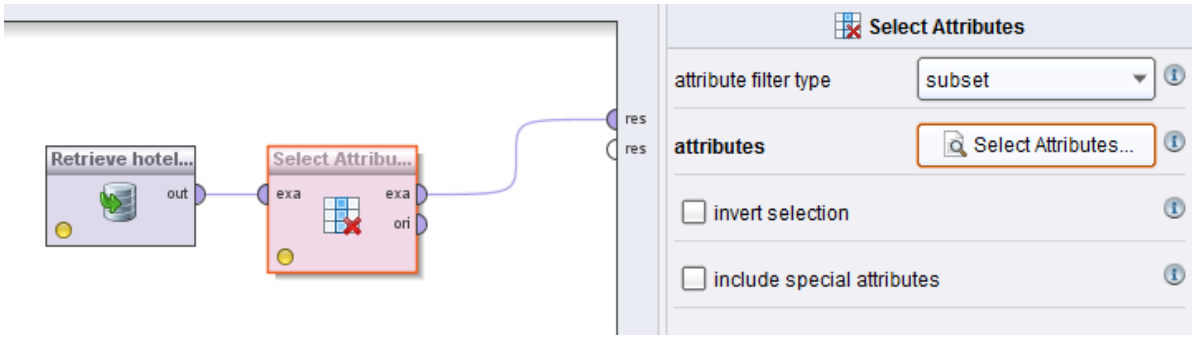
7.3.3.3 Εισαγωγή και προσαρμογή των δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν ως είσοδος προέρχονται από έρευνα με ερωτηματολόγια και έχουν κωδικοποιηθεί κατάλληλα σε πρόγραμμα στατιστικής ανάλυσης (βλέπε Σχήμα 2.7 της ενότητας 3.4.2 του Κεφαλαίου 2). Από το πρόγραμμα αυτό μπορούν να εξαχθούν πολύ απλά σε μορφή φύλλου δεδομένων Excel, που είναι ισοδύναμη μορφή και απολύτως τυποποιημένη, ώστε να μπορούν να εισαχθούν άμεσα στο RapidMiner. Η δομή των δεδομένων είναι η γνωστή μορφή πίνακα (Σχήμα 7.35), όπου κάθε στήλη αντιστοιχεί σε ένα χαρακτηριστικό (στοιχείο προσδοκίας, ικανοποίηση, δημογραφικά στοιχεία) και κάθε γραμμή ένας ερωτώμενος (στην περίπτωσή μας παράδειγμα). Τα ονόματα των στηλών είναι σύντομα ονόματα που βασίζονται στα στοιχεία του ερωτηματολογίου, ενώ οι τιμές των κελιών που αφορούν τις προσδοκίες των πελατών αντιστοιχούν στις διαβαθμίσεις 1 ως 5 (πλήρης διαφωνία ως πλήρης συμφωνία) ή στη μη απάντηση (τιμή 99).

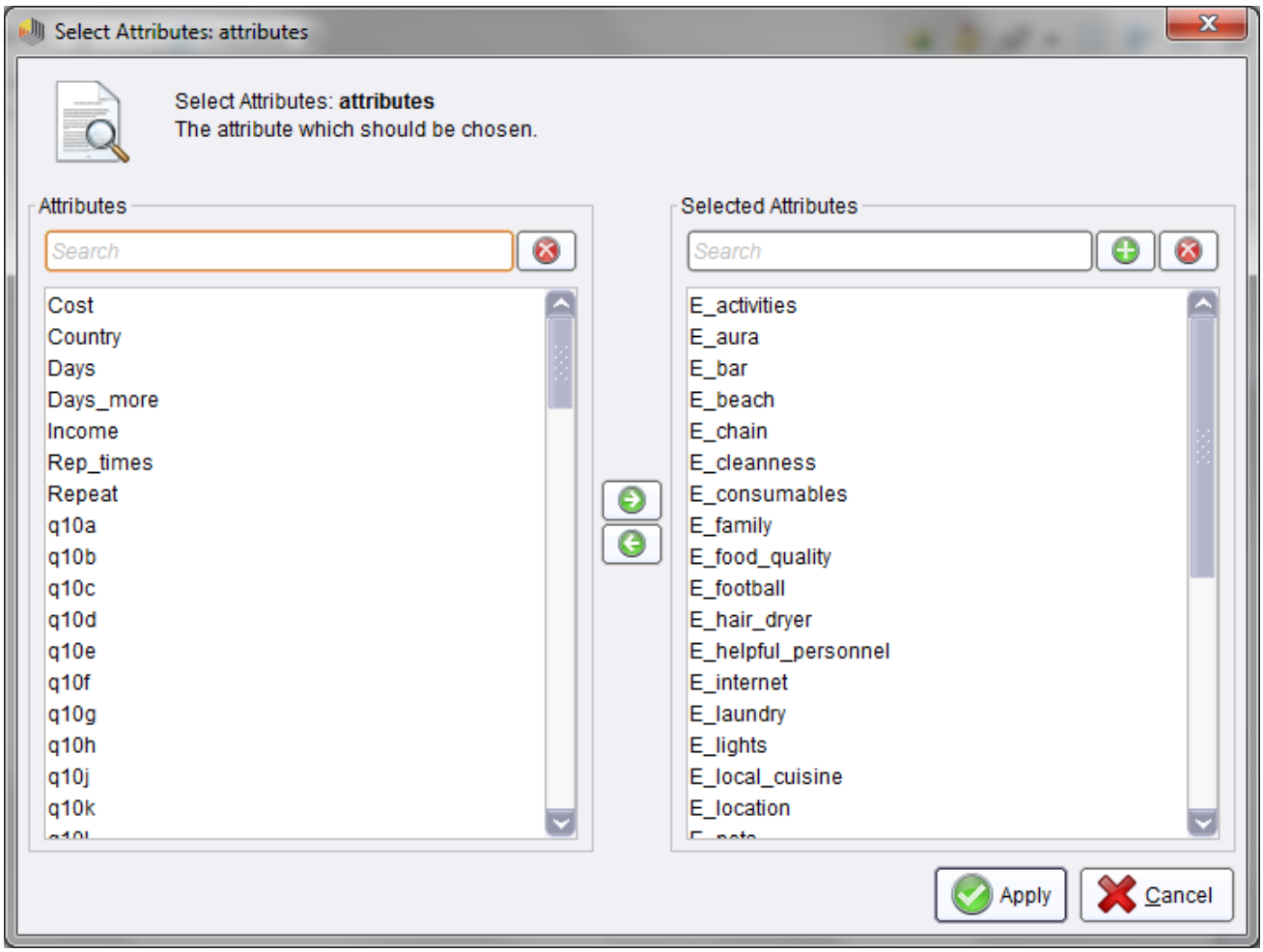
	A	B	C	D	E	F	G
1	IND	E_special_needs	E_spa	E_lights	E_parking	E_internet	E_safebox
2	1	4	1	1	4	4	5
3	2	99	99	99	99	5	5
4	3	99	99	99	99	5	5
5	4	99	99	99	99	5	5
6	5	99	99	99	99	5	5
7	6	99	99	99	99	99	99
8	7	99	99	99	99	5	4
9	8	99	99	99	99	5	5
10	9	3	99	5	99	5	5
11	10	99	99	99	99	5	5
12	11	99	99	4	99	5	5
13	12	99	99	4	99	5	5
14	13	4	3	2	1	5	5
15	14	4	4	4	5	5	5
16	15	4	5	5	4	5	4
17	16	3	3	3	3	4	5
18	17	3	3	3	3	5	5
19	18	4	3	2	3	5	5
20	19	2	2	1	2	5	5

Σχήμα 7.35. *Τα αρχικά δεδομένα της έρευνας σε μορφή φύλλου δεδομένων*

Η εισαγωγή των δεδομένων γίνεται κατά τα γνωστά, χρησιμοποιώντας τον οδηγό που προσφέρεται στα εργαλεία του αποθετηρίου **Import data into an existing repository** και αποθηκεύοντάς τα στο τοπικό αποθετήριο με το όνομα που επιθυμούμε (εδώ ονομάστηκε hotels_expectations). Τα δεδομένα της έρευνας είναι διαθέσιμα στο αρχείο Excel **Hotels_Nothern_Greece.xlsx**. Οι τιμές προς εισαγωγή παριστάνονται ως ακέραιοι αριθμοί και αντιστοιχούν σε κάποια κλίμακα. Ωστόσο, τα αντίστοιχα χαρακτηριστικά δε θα πρέπει να χαρακτηριστούν ως αριθμητικά, αλλά ως ονομαστικά (polynominal), επειδή οι τιμές δεν έχουν την έννοια της ποσότητας, αλλά εκλαμβάνονται ως κωδικοί κάποιας άλλης απάντησης (π.χ. το 2 δεν είναι διπλάσιο του 1, αλλά αντιστοιχεί στην επόμενη διαθέσιμη απάντηση). Στο Σχήμα 7.36 φαίνονται οι τελεστές για την ανάγνωση των δεδομένων από το αποθετήριο και την επιλογή των επιθυμητών χαρακτηριστικών. Σημειώνεται ότι τα δεδομένα αφορούν τη συνολική έρευνα που πραγματοποιήθηκε για τις προσδοκίες και την ικανοποίηση των πελατών ξενοδοχείων και, επομένως, θα πρέπει να επιλεγούν τα χαρακτηριστικά που αφορούν τη συγκεκριμένη εφαρμογή. Πατώντας στο κουμπί της παραμέτρου attributes του τελεστή Select Attributes (Σχήμα 7.36), επιλέγουμε από όλα τα διαθέσιμα χαρακτηριστικά, αυτά που αφορούν στοιχεία προσδοκίας (Σχήμα 7.37).



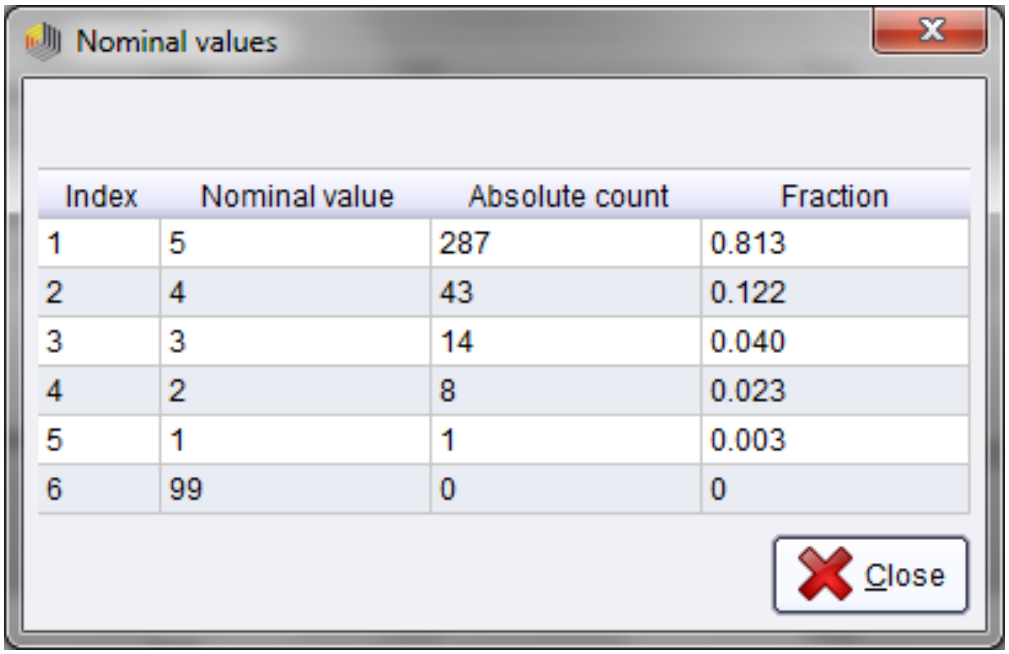
Σχήμα 7.36. *Ανάγνωση των δεδομένων και επιλογή χαρακτηριστικών.*



Σχήμα 7.37. *Η οθόνη επιλογής χαρακτηριστικών*

Στο σημείο αυτό είναι καλή τακτική να δούμε προσεκτικά τα δεδομένα, οδηγώντας τα στην έξοδο της διαδικασίας και πηγαίνοντας στην καρτέλα Statistics της προβολής Results (Σχήμα 7.38). Έχει ενδιαφέρον να δούμε τις κατανομές των τιμών ώστε να έχουμε μια πρώτη εικόνα για τις προτιμήσεις των ερωτώμενων. Στο Σχήμα 7.39 φαίνεται η κατανομή των απαντήσεων σχετικά με την ποιότητα φαγητού, όπου παρατηρούμε ότι η πολύ μεγάλη πλειονότητα των ερωτώμενων τη θέτει σε υψηλό βαθμό προσδοκίας. Αντίθετα, αν παρατηρήσει κανείς την κατανομή των απαντήσεων σχετικά με το νυχτερινό φωτισμό στους εξωτερικούς χώρους, διαπιστώνει ότι το μεγαλύτερο ποσοστό (47%) είναι ουδέτεροι και μόνο το 20% εκφράζει υψηλή προσδοκία.

Σχήμα 7.38. *Η καρτέλα προβολής στατιστικών στοιχείων των δεδομένων εισόδου.*

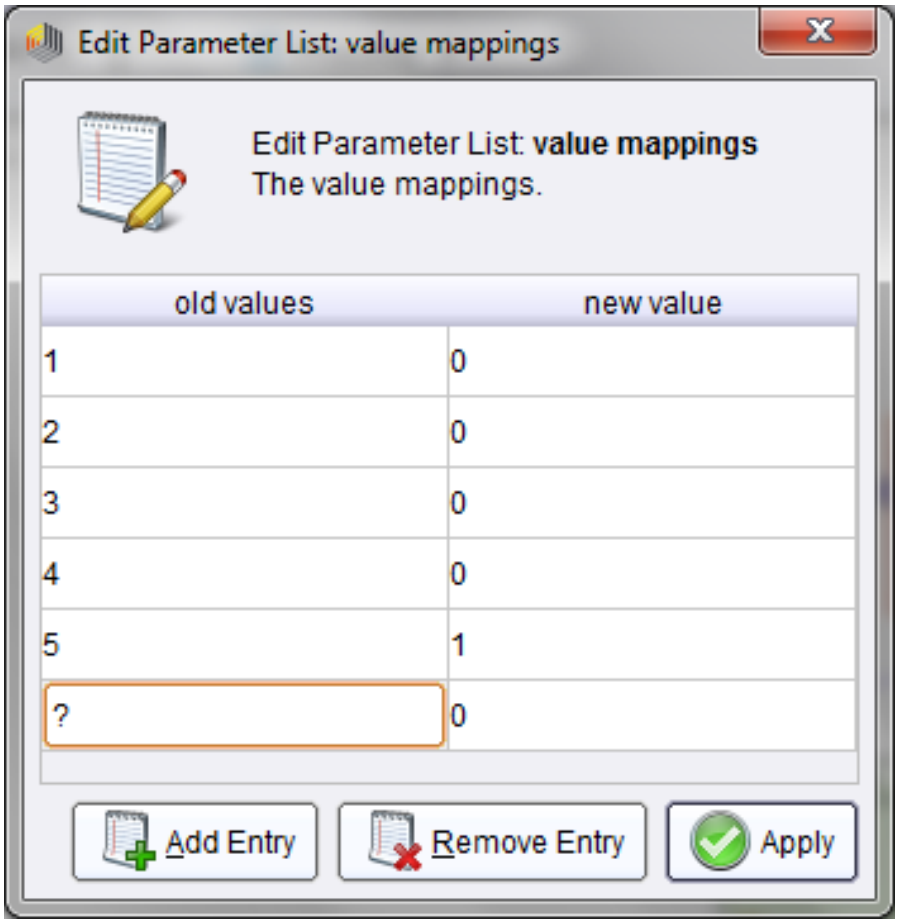


Σχήμα 7.39*. Οι συχνότητες των απαντήσεων σχετικά με την ποιότητα φαγητού.*

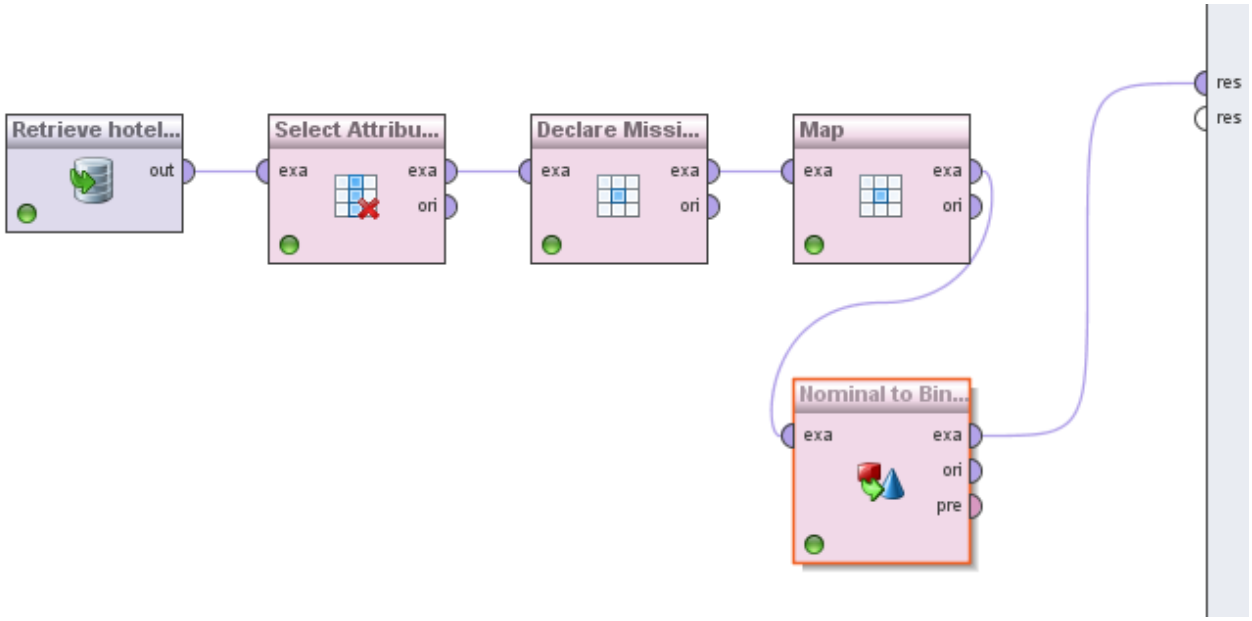
Στα αρχικά δεδομένα, η τιμή 99 έχει χρησιμοποιηθεί για να παραστήσει το αναπάντητο στοιχείο (missing value), αλλά το RapidMiner δεν το γνωρίζει αυτό και τη θεωρεί ως μια έγκυρη ονομαστική τιμή. Καλό είναι να δηλωθεί η τιμή 99 ως ανύπαρκτη τιμή, με χρήση του τελεστή **Declare Missing Value**.

Το επόμενο βήμα προετοιμασίας των δεδομένων είναι η μετατροπή τους από 5-βάθμια κλίμακα σε δυαδικά. Η απεικόνιση των τιμών πραγματοποιείται με τον τελεστή **Map**, που βρίσκεται στην κατηγορία **Data Transformation**, στο φάκελο **Value Modification**. Με τη βοήθεια των παραμέτρων του **Map**,

επιλέγουμε τα χαρακτηριστικά που επιθυμούμε να επηρεαστούν (στην περίπτωσή μας είναι όλα τα χαρακτηριστικά που έχουν ήδη επιλεγεί στο προηγούμενο βήμα από τον **Select Attributes**) και καθορίζουμε τα ζευγάρια παλαιών και νέων τιμών. Παρατηρώντας τα δεδομένα και λαμβάνοντας υπόψη τη φύση του προβλήματος, μπορούμε να αντιληφθούμε ότι μεγάλο μέρος των ερωτώμενων απαντάει θετικά για πολλά στοιχεία προσδοκίας, με την έννοια ότι τα θεωρεί θετικά και τα επιθυμεί, χωρίς όμως στην πραγματικότητα να τα θέτει πραγματικά σε υψηλή προτεραιότητα. Αν θέλουμε να επικεντρωθούμε στην μελέτη της ισχυρής προσδοκίας και όχι της χαλαρής θετικής προδιάθεσης, θα πρέπει να θεωρήσουμε ως παρουσία της προσδοκίας την απάντηση «Συμφωνώ απόλυτα», ενώ, αντίθετα, την απάντηση «Συμφωνώ» να τη συγχωνεύσουμε με την ουδετερότητα και τη μη συμφωνία. Με βάση το σκεπτικό αυτό, οι τιμές 1 ως και 4 απεικονίζονται στο 0 (false – απουσία στοιχείου) και μόνο η τιμή 5 στο 1 (true – παρουσία στοιχείου) (Σχήμα 7.40). Η επιλογή αυτή δεν είναι απόλυτη, αλλά απόφαση του αναλυτή και φυσικά μπορεί κάποιος να καταλήξει σε αυτή μετά από πειραματισμό και εκτίμηση του τελικού αποτελέσματος. Επίσης, αναφέρεται ότι η τιμή ? αντιστοιχεί στη μη απάντηση (missing), η οποία αντιστοιχίζεται και αυτή στην απουσία προσδοκίας. Τέλος, ο Map πρέπει να ακολουθηθεί από τον τελεστή Nominal to Binary, που μετατρέπει τις τιμές 0 και 1 σε λογικές τιμές που να τις αντιλαμβάνεται το RapidMiner ως True/False. Στο Σχήμα 7.41 απεικονίζεται η συνολική διαδικασία ανάγνωσης και προετοιμασίας των δεδομένων.



Σχήμα 7.40. *Η απεικόνιση των τιμών 5-βάθμιας κλίμακας σε δυαδικές 0-1.*



Σχήμα 7.41. *Η διαδικασία ανάγνωσης και προετοιμασίας των δεδομένων.*

7.3.3.4 Επισκόπηση των δεδομένων

Τα δεδομένα έχουν έρθει στη μορφή 0-1, όπου το 1 (true) δηλώνει ισχυρή προσδοκία για το αντίστοιχο στοιχεί και το 0 (false) την απουσία προσδοκίας. Παρατηρώντας τα στατιστικά των χαρακτηριστικών, μπορούμε να διακρίνουμε ότι υπάρχουν χαρακτηριστικά με ελάχιστη υποστήριξη, όπως τα μενού ειδικής δίαιτας, και άλλα με υψηλά ποσοστά εμφάνισης, όπως η καθαριότητα και τα αναλώσιμα μπάνιου (Σχήμα 7.42).

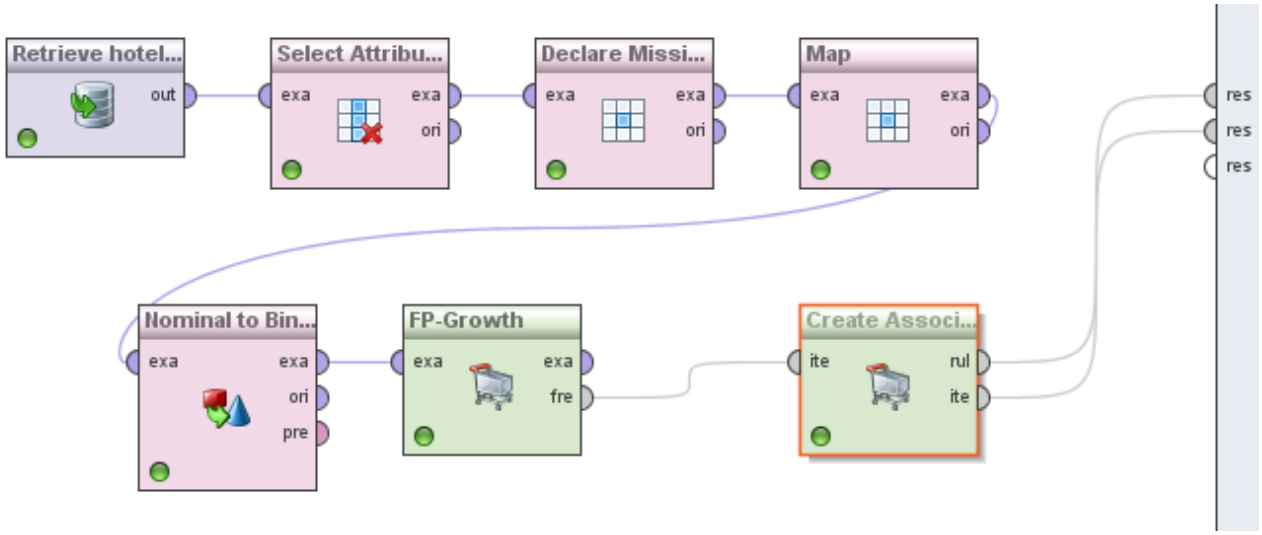
Name	Type	Miss.	Statistics			Filter (33 / 33 attr
E_family	Binominal	0	Least 1 (125)	Most 0 (294)	Values 0 (294), 1 (125)	
E_spacious_foyer	Binominal	0	Least 1 (91)	Most 0 (328)	Values 0 (328), 1 (91)	
E_tv	Binominal	0	Least 0 (164)	Most 1 (255)	Values 1 (255), 0 (164)	
E_temperature_adj	Binominal	0	Least 1 (105)	Most 0 (314)	Values 0 (314), 1 (105)	
E_safety	Binominal	0	Least 0 (139)	Most 1 (280)	Values 1 (280), 0 (139)	
E_laundry	Binominal	0	Least 0 (105)	Most 1 (314)	Values 1 (314), 0 (105)	
E_cleanness	Binominal	0	Open chart	Least 0 (35)	Most 1 (384)	
E_consumables	Binominal	0	Least 0 (66)	Most 1 (353)	Values 1 (353), 0 (66)	
E_restaurant	Binominal	0	Least 0 (132)	Most 1 (287)	Values 1 (287), 0 (132)	

Σχήμα 7.42. *Τα στατιστικά των προσδοκιών των πελατών.*

7.3.3.5 Μοντελοποίηση

Η μοντελοποίηση πραγματοποιείται με χρήση των FP-Growth και Create Association Rules, όπως στην εφαρμογή της ανάλυσης καλαθιού αγορών. Η ρύθμιση των παραμέτρων γίνεται και πάλι μετά από δοκιμές. Η τιμή 0,8 για την ελάχιστη υποστήριξη των συχνών συνόλων αποδεικνύεται υψηλή και επιλέγεται η τιμή 0,5, η οποία οδηγεί στην εξαγωγή 108 συχνών συνόλων με μέγιστο μέγεθος τα 5 στοιχεία. Στον Create Association Rules επιλέχθηκε ως κριτήριο η εμπιστοσύνη (Confidence) με ελάχιστο όριο το 0,6, αποδίδοντας 573 κανόνες. Στο Σχήμα 7.43 φαίνεται η τελική συνολική διαδικασία και στο Σχήμα 7.44 οι κανόνες που εξήχθησαν, ταξινομημένοι κατά φθίνουσα τιμή του lift. Οι κανόνες με το μεγαλύτερο ενδιαφέρον είναι ενδεικτικά:

E_bar → E_room_service, Confidence=0,972, Lift=1,471
E_consumables, E_food_quality → E_laundry, E_restaurant, Confidence=0,86, Lift=1,452
E_internet → E_safebox, Confidence=0,89, Lift=1,402



Σχήμα 7.43*. Η συνολική διαδικασία εξαγωγής κανόνων συσχέτισης.*

No.	Premises	Conclusion	Support	Confid...	LaPla...	Gain	p-s	Lift	Convi...
165	E_room_service	E_bar	0.504	0.762	0.905	-0.819	0.161	1.471	2.023
520	E_bar	E_room_service	0.504	0.972	0.991	-0.532	0.161	1.471	12.25
348	E_consumables, E_food_quality	E_laundry, E_restaurant	0.527	0.860	0.947	-0.699	0.164	1.453	2.913
349	E_consumables, E_food_quality	E_cleanness, E_laundry, E_restaurant	0.527	0.860	0.947	-0.699	0.164	1.453	2.913
350	E_cleanness, E_consumables, E_food_quality	E_laundry, E_restaurant	0.527	0.860	0.947	-0.699	0.164	1.453	2.913
401	E_laundry, E_restaurant	E_consumables, E_food_quality	0.527	0.891	0.960	-0.656	0.164	1.453	3.551
402	E_laundry, E_restaurant	E_cleanness, E_consumables, E_food_quality	0.527	0.891	0.960	-0.656	0.164	1.453	3.551
403	E_cleanness, E_laundry, E_restaurant	E_consumables, E_food_quality	0.527	0.891	0.960	-0.656	0.164	1.453	3.551
270	E_cleanness, E_room_service	E_consumables, E_laundry, E_food_quality	0.520	0.816	0.929	-0.754	0.157	1.431	2.341
427	E_consumables, E_laundry, E_food_quality	E_cleanness, E_room_service	0.520	0.912	0.968	-0.621	0.157	1.431	4.129
215	E_cleanness, E_food_quality	E_consumables, E_laundry, E_restaurant	0.527	0.786	0.914	-0.814	0.158	1.427	2.101
497	E_consumables, E_laundry, E_restaurant	E_cleanness, E_food_quality	0.527	0.957	0.985	-0.575	0.158	1.427	7.608
317	E_consumables, E_food_quality	E_cleanness, E_laundry, E_room_service	0.520	0.848	0.942	-0.706	0.151	1.410	2.626
359	E_cleanness, E_laundry, E_room_service	E_consumables, E_food_quality	0.520	0.865	0.949	-0.683	0.151	1.410	2.866
154	E_cleanness, E_food_quality	E_restaurant, E_room_service	0.506	0.754	0.901	-0.835	0.146	1.405	1.886
477	E_restaurant, E_room_service	E_cleanness, E_food_quality	0.506	0.942	0.980	-0.568	0.146	1.405	5.700
277	E_safebox	E_internet	0.520	0.820	0.930	-0.749	0.149	1.402	2.301
396	E_internet	E_safebox	0.520	0.890	0.959	-0.649	0.149	1.402	3.313
287	E_cleanness, E_food_quality	E_laundry, E_restaurant	0.556	0.829	0.931	-0.785	0.159	1.401	2.389
469	E_laundry, E_restaurant	E_cleanness, E_food_quality	0.556	0.940	0.978	-0.628	0.159	1.401	5.445

Σχήμα 7.44. *Οι κανόνες με υψηλή εμπιστοσύνη και τον υψηλότερο δείκτη lift.*

7.3.3.6 Εφαρμογή και αξιολόγηση του μοντέλου

Οι κανόνες που προκύπτουν είναι πολλοί και παρέχουν ικανοποιητικά επίπεδα υποστήριξης και εμπιστοσύνης. Μια αρνητική παρατήρηση είναι ότι δεν υπάρχουν κανόνες με ιδιαίτερα υψηλό δείκτη lift, που σημαίνει ότι τα στοιχεία στων οποίων την πρόβλεψη οδηγούν οι κανόνες δεν εξαρτώνται με ιδιαίτερα ισχυρό τρόπο από τα στοιχεία της υπόθεσης. Ωστόσο, οι κανόνες δεν μπορούν χαρακτηριστούν ως προφανείς, αλλά φαίνεται ότι διαθέτουν κάποια αξία, εκμεταλλεύσιμη από ένα στέλεχος διοίκησης. Μια ακόμα παρατήρηση (που ισχύει γενικότερα σε πολλές αναλύσεις αυτού του είδους) είναι η σημαντική αλληλοεπικάλυψη ανάμεσα στους κανόνες.

7.3.4 Μελέτη της επίδρασης επιμέρους στοιχείων ικανοποίησης στη συνολική ικανοποίηση των πελατών ξενοδοχείων

7.3.4.1 Ορισμός του προβλήματος

Σκοπός της εφαρμογής αυτής είναι να μελετηθεί το ποια στοιχεία ενός ξενοδοχείου επηρεάζουν την ικανοποίηση των πελατών τους, έτσι ώστε η διοίκηση του ξενοδοχείου να δώσει την ανάλογη έμφαση στις προσφερόμενες υπηρεσίες και την ποιότητά τους. Η εφαρμογή είναι συνέχεια του προηγούμενου παραδείγματος και το σετ δεδομένων που θα χρησιμοποιηθεί αποτελεί μέρος της ίδιας πραγματικής έρευνας αξιολόγησης ποιότητας υπηρεσιών ξενοδοχείων. Το ερωτηματολόγιο, μέσω του οποίου συγκεντρώθηκαν τα πρωτογενή δεδομένα, περιλαμβάνει (μεταξύ άλλων) στοιχεία του επισκέπτη (π.χ. ηλικία, φύλο), στοιχεία για το ταξίδι (π.χ. αν έρχεται για πρώτη φορά), μία ερώτηση για τη συνολική ικανοποίηση από το ξενοδοχείο, καθώς και μια ερώτηση πολλαπλών στοιχείων όπου ο ερωτώμενος καλείται να προσδιορίσει την ικανοποίησή του από ένα μεγάλο αριθμό παροχών και χαρακτηριστικών του ξενοδοχείου. Το σετ δεδομένων είναι το ίδιο με αυτό της προηγούμενης εφαρμογής (Ενότητα 3.3) και διατίθεται σε μορφή αρχείου MS-Excel μέσω του συνδέσμου: www.ba.teithe.gr/eBook_Data_and_Business_Intelligence/hotels_Northern_Greece.xlsx .

Χρησιμοποιώντας ως στόχο τη συνολική ικανοποίηση από το ξενοδοχείο και ως χαρακτηριστικά (α) την ικανοποίηση των πελατών από τα επιμέρους στοιχεία του ξενοδοχείου και (β) τα στοιχεία του πελάτη, επιθυμούμε να κατασκευάσουμε ένα μοντέλο που να «μάθει» να προβλέπει τη συνολική ικανοποίηση.

7.3.4.2. Σχεδιασμός

Το πρόβλημα που περιγράφηκε είναι πρόβλημα κατάταξης. Τα δεδομένα που διαθέτουμε από την έρευνα μπορούν να χρησιμοποιηθούν ως παραδείγματα, για τα οποία γνωρίζουμε τα στοιχεία ικανοποίησης, τα στοιχεία του πελάτη και το τελικό αποτέλεσμα συνολικής ικανοποίησης. Το μοντέλο που πρέπει να αναπτυχθεί μπορεί να είναι ένα δέντρο απόφασης ή μια μη-δενδροειδής μέθοδος εξαγωγής κανόνων που να μαθαίνει από παραδείγματα. Επομένως, η διαδικασία επικεντρώνεται γύρω από έναν τελεστή της κατηγορίας **Classification and Regression**.

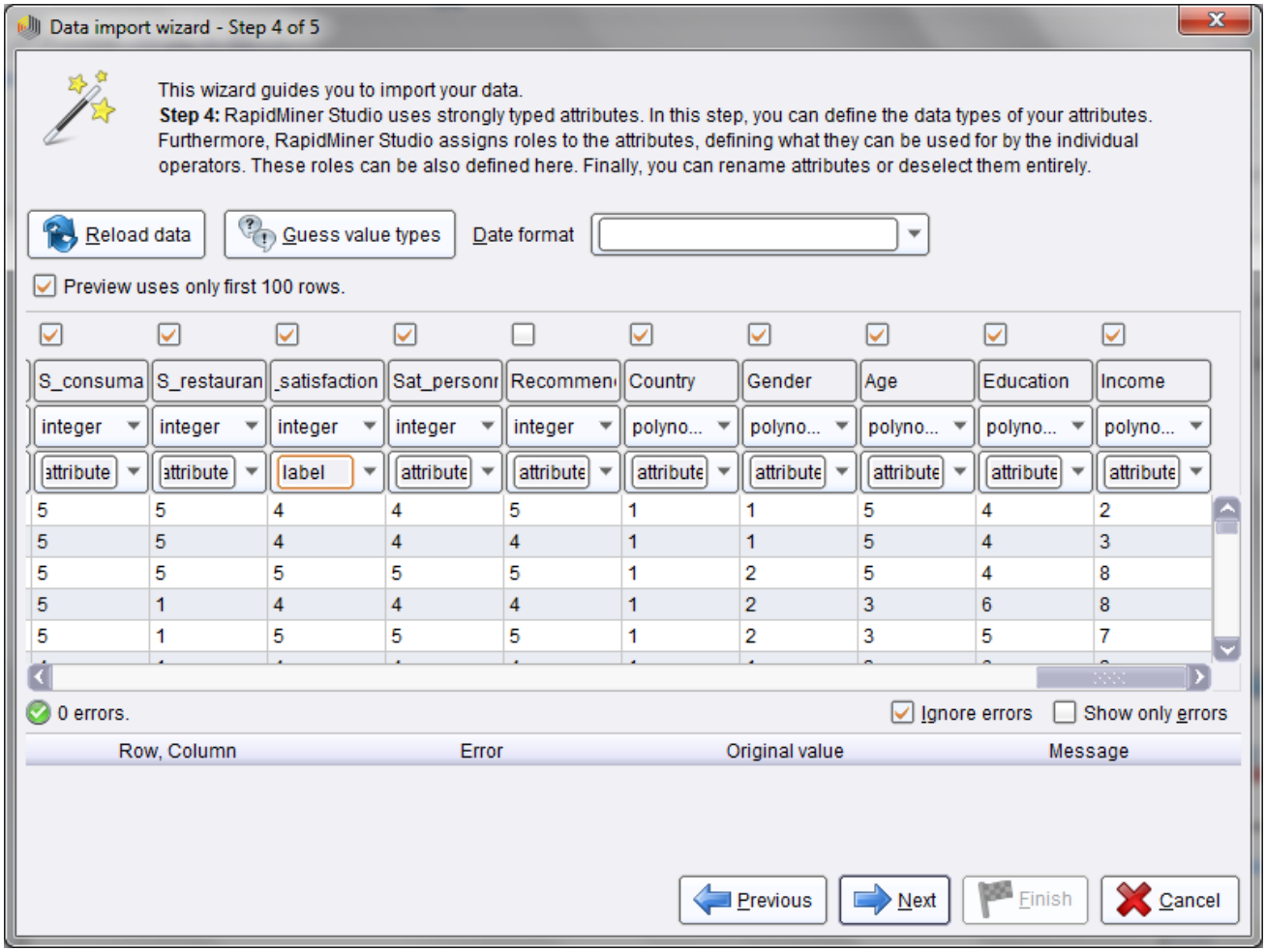
Τα δεδομένα περιλαμβάνουν καθαρά ονομαστικά χαρακτηριστικά (polynominal), όπως το φύλο, το επίπεδο εκπαίδευσης και το αν ο επισκέπτης έρχεται για πρώτη φορά στο συγκεκριμένο προορισμό, αλλά και χαρακτηριστικά που, ενώ είναι ποσοτικής φύσης, έχουν ήδη μετατραπεί κατά την έρευνα σε ονομαστικά, μέσω του χωρισμού σε τάξεις, όπως η ηλικία και το εισόδημα (π.χ. αντί της ακριβούς ηλικίας, έχει καταγραφεί η ηλικιακή κατηγορία του επισκέπτη: (1) 18-25, (2) 26-35, (3) 36-45, (4) 46-55, (5) 56-65, (6) >65).

Μια άλλη κατηγορία χαρακτηριστικών, που είναι και η καθοριστικότερη, είναι οι απαντήσεις σχετικά με την ικανοποίηση, που έχουν καταγραφεί σε κλίμακα 1-5 (από απόλυτη δυσαρέσκεια μέχρι την απόλυτη ικανοποίηση). Οι τιμές αυτές μπορούν να θεωρηθούν ως ποσοτικές, κάτι που συνηθίζεται στις κοινωνικές έρευνες, προκειμένου να πραγματοποιηθεί ανάλυση με ποσοτικές μεθόδους. Η προσέγγιση αυτή συχνά αποδίδει ικανοποιητικά αποτελέσματα, αλλά βασίζεται σε αμφιλεγόμενες παραδοχές και έχει κατακριθεί από μεγάλη μερίδα ερευνητών. Στη συγκεκριμένη εφαρμογή, εφόσον δεν κατευθυνόμαστε προς κάποια καθαρά ποσοτική μέθοδο (π.χ. παλινδρόμηση), προτείνεται η χρήση των τιμών ικανοποίησης ως μια ποιοτική κλίμακα. Επιπλέον, προτείνεται η τροποποίηση της 5-βάθμιας κλίμακας σε 3-βάθμια (Χαμηλή, Μέση, Υψηλή), αφού η λεπτή διάκριση ανάμεσα στο π.χ. συμφωνώ απόλυτα και το συμφωνώ δε φαίνεται να εμπεριέχει αξιοποιήσιμη πληροφορία, αλλά μάλλον άχρηστη λεπτομέρεια που θα δυσκολέψει την εκμάθηση.

Ως κύρια μέθοδος μοντελοποίησης επιλέγεται η ανάπτυξη ενός δέντρου αποφάσεων, που έχει το πλεονέκτημα ότι παρέχει παραστατικά και εύκολα ερμηνεύσιμα αποτελέσματα. Επιπλέον, το ανεπτυγμένο δέντρο μπορεί εύκολα να μετατραπεί σε μορφή κανόνων, ώστε να διαθέτουμε το αποτέλεσμα και σε αυτήν την εναλλακτική μορφή. Η επιλογή του κατάλληλου μοντέλου εκμάθησης είναι σημαντική απόφαση, που μπορεί να ληφθεί με βάση τη φύση των δεδομένων, την εμπειρία του ερευνητή και μετά από πειραματισμό. Μια καλή επιλογή για το συγκεκριμένο πρόβλημα είναι τα δέντρα **CHAID**, τα οποία βασίζονται στην απόσταση χ^2 (chi-squared) για τον υπολογισμό της εγγύτητας των αντικειμένων (δηλ για τον υπολογισμό του μέτρου που κρίνει αν ένα παράδειγμα είναι πιο κοντά στη μια ή στην άλλη κατηγορία). Η απόσταση χ2 είναι ένα μέτρο που θεωρείται καταλληλότερο για ονομαστικά δεδομένα, όπως συμβαίνει στην περίπτωσή μας. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο τελεστής **Decision Tree** (όπως στην εφαρμογή της Ενότητας 3.1), ο οποίος μπορεί να χειριστεί ταυτόχρονα ονομαστικά και ποσοτικά δεδομένα. Αναφέρεται επίσης ως επιλογή ο τελεστής **Rule induction**, που έχει τη δυνατότητα να εξάγει ένα σύνολο κανόνων χωρίς τη χρήση δέντρου, βασιζόμενος στη μεγιστοποίηση της πληροφορίας (information gain).

7.3.4.3. Εισαγωγή, προετοιμασία και επισκόπηση δεδομένων

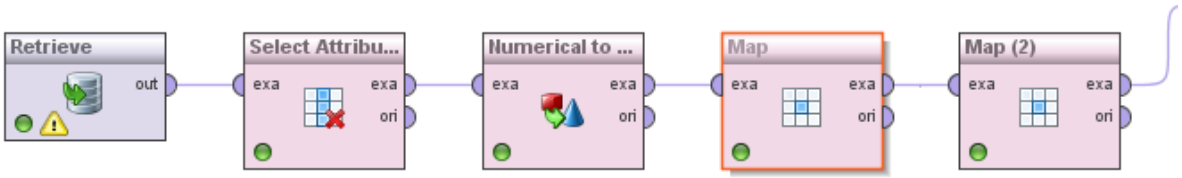
Τα δεδομένα εισόδου είναι διαθέσιμα σε μορφή φύλλου εργασίας Excel στο αρχείο **hotels_Northern_Greece.xlsx** (είναι το ίδιο με αυτό που χρησιμοποιήθηκε στην Ενότητα 3.3). Η εισαγωγή τους μπορεί να γίνει στο RapidMiner μέσω του οδηγού **Import Excel Sheet**, καθορίζοντας παράλληλα τον τύπο δεδομένων των χαρακτηριστικών και ορίζοντας το πεδίο Total_satisfaction ως το χαρακτηριστικό-στόχο (label). Κατά το 4^{ο} βήμα του οδηγού (Σχήμα 7.45), θα πρέπει να από-επιλεγούν τα χαρακτηριστικά που δεν αφορούν το πρόβλημα (όπως αυτά που αφορούν τις προσδοκίες) και να παραμείνουν επιλεγμένα τα στοιχεία ικανοποίησης (η ονομασία τους ξεκινάει από S_), η συνολική ικανοποίηση (Total_satisfaction), η ηλικία (Age), το κόστος του δωματίου (Cost), η χώρα προέλευσης (Country), το φύλο (Gender), το εισόδημα (Income) και το αν έχουν ξαναεπισκεφθεί τον προορισμό (Repeat). Όλα τα χαρακτηριστικά αναγνωρίζονται ως integer, αλλά επειδή τα αριθμητικά ψηφία έχουν την έννοια κωδικών και όχι ποσότητας, πρέπει να τροποποιηθούν σε Polynominal. Ειδικά τα χαρακτηριστικά που αφορούν κλίμακα (1-5), μπορούν να παραμείνουν ως integer κατά τη φάση εισαγωγής (αφού οι τιμές εμπεριέχουν κάποιου είδους ποσοτική πληροφορία) και η μετατροπή τους να πραγματοποιηθεί εντός της διαδικασίας ανάλυσης με τη βοήθεια του κατάλληλου τελεστή, ανάλογα με τη μέθοδο που θα χρησιμοποιηθεί.



Σχήμα 7.45. *Το βήμα επιλογής χαρακτηριστικών και προσδιορισμού του τύπου και του ρόλου τους.*

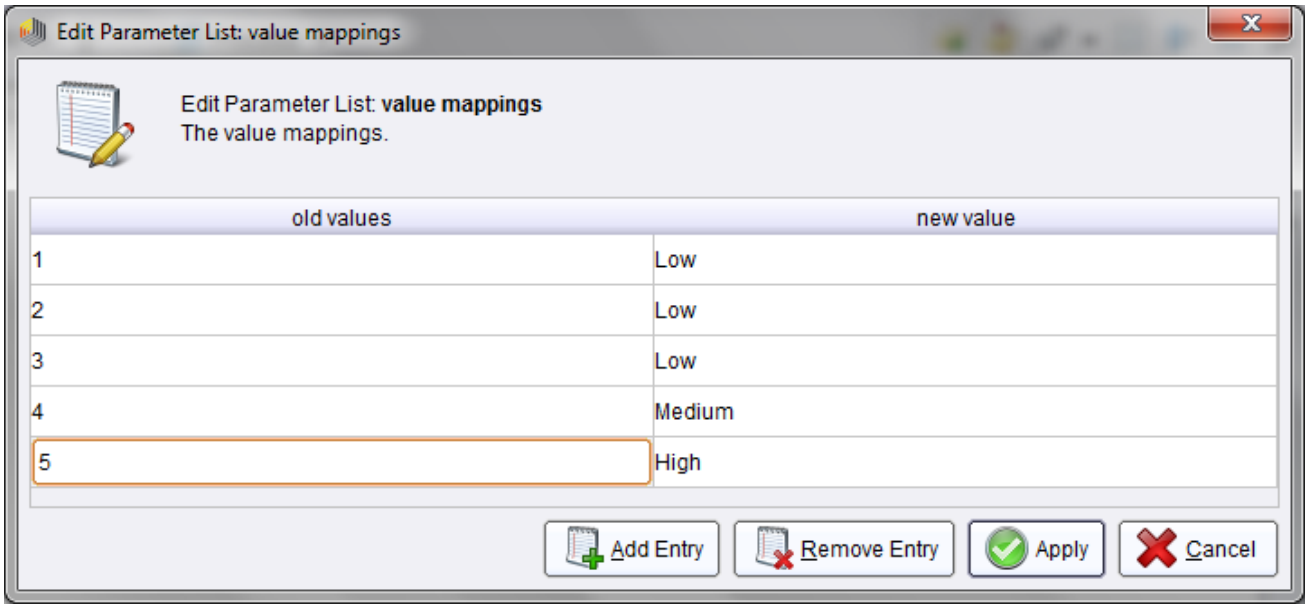
Η προετοιμασία των δεδομένων περιλαμβάνει την επιλογή των χαρακτηριστικών, τη μετατροπή των αριθμητικών χαρακτηριστικών σε ονομαστικά και την τροποποίηση της κλίμακας. Παρατηρώντας τα στατιστικά των ακατέργαστων δεδομένων, και συγκεκριμένα το γράφημα της συνολικής ικανοποίησης (Total_satisfaction), φαίνεται ότι οι απαντήσεις επικεντρώνονται γύρω από το 4 και είναι ελάχιστοι οι ερωτώμενοι που δήλωσαν απόλυτα δυσαρεστημένοι (1) ή δυσαρεστημένοι (2). Μετά από αυτήν την παρατήρηση, κρίνεται σκόπιμο να τροποποιηθεί η κλίμακα με τρόπο που να ενισχύει τις αντιθέσεις (αφού θα ήταν άσκοπο να μελετήσουμε την ικανοποίηση και τη δυσαρέσκεια αν όλοι είναι ικανοποιημένοι). Με τη χρήση του τελεστή **Map**, οι τιμές 1,2 και 3 (δηλ απόλυτα δυσαρεστημένος ως ουδέτερος) μπορούν να συγχωνευθούν στην τιμή Low (χαμηλή ικανοποίηση), η τιμή 4 μπορεί να απεικονιστεί στην τιμή Medium (μέτρια ικανοποίηση) και η 5 στην High (υψηλή ικανοποίηση). Επίσης, με τη χρήση ενός ακόμα τελεστή **Map**, η 5-βάθμια κλίμακα μέτρησης των στοιχείων ικανοποίησης μπορεί να μετατραπεί σε 3-βάθμια, αντιστοιχίζοντας τις τιμές 1 και 2 στο βαθμό ικανοποίησης Low, την 3 στο Medium και τις τιμές 4 και 5 στο βαθμό High.

Στο Σχήμα 7.46 φαίνεται το πρώτο μέρος της διαδικασίας, που αντιστοιχεί στην προετοιμασία των δεδομένων. Ο **Select Attributes** χρησιμοποιείται για την επιλογή των χαρακτηριστικών που συμμετέχουν στο πρόβλημα και ο **Numerical to Nominal** στη μετατροπή των αριθμητικών χαρακτηριστικών σε ονομαστικά. Στις παραμέτρους του τελευταίου, πρέπει να επιλεγούν όλα τα χαρακτηριστικά και επιπρόσθετα να ενεργοποιηθεί η επιλογή **include special attributes**, ώστε να ισχύσει η μετατροπή και στο χαρακτηριστικό Total_satisfaction (Επειδή το Total_satisfaction έχει δηλωθεί ως τύπου label και όχι ως απλό χαρακτηριστικό, ο **Numerical to Nominal** δε θα το επηρέαζε χωρίς την ενεργοποίηση της παραπάνω επιλογής).



Σχήμα 7.46. *Η διαδικασία ανάγνωσης και προετοιμασίας των δεδομένων.*

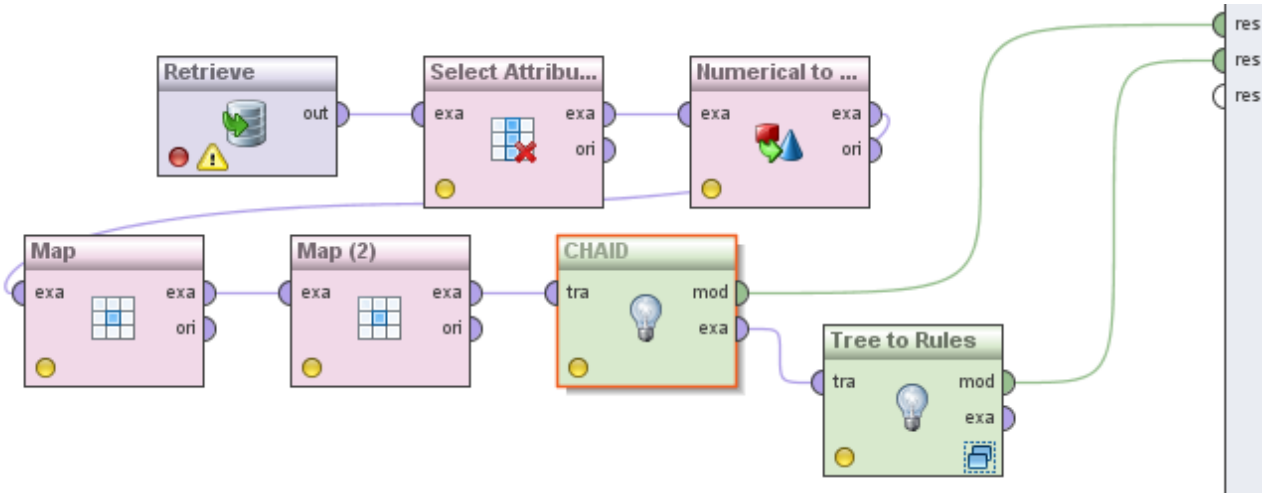
Ο τελεστής **Map** χρησιμοποιείται δύο φορές, μία για την τροποποίηση της κλίμακας του χαρακτηριστικού Total_satisfaction (Σχήμα 7.47) και μία για την κλίμακα όλων των στοιχείων ικανοποίησης.



Σχήμα 7.47. *Ο επανακαθορισμός της κλίμακας της συνολικής ικανοποίησης.*

7.3.4.4. Μοντελοποίηση

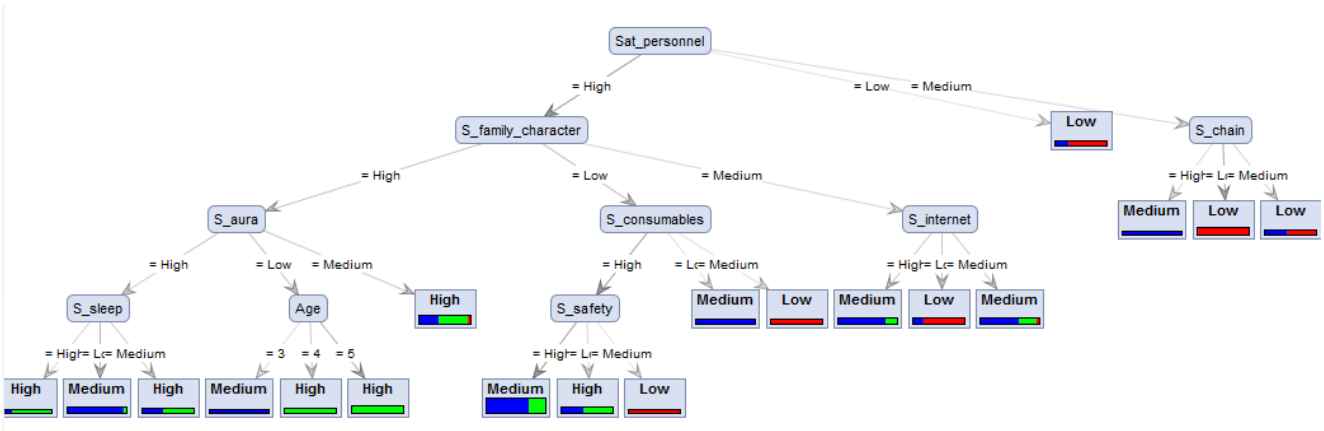
Σε συνέχεια του σχεδιασμού της Ενότητας 3.4.2, επιλέχθηκε η μοντελοποίηση με χρήση δέντρου CHAID και στη συνέχεια η εξαγωγή κανόνων. Στο Σχήμα 7.48 παρουσιάζεται η συνολική διαδικασία της εφαρμογής, όπου διακρίνεται ο τελεστής **CHAID** (που βρίσκεται στην κατηγορία **Tree Induction**) και ο **Tree to Rules** (που βρίσκεται στην κατηγορία **Rule Induction**).



Σχήμα 7.48. *Η συνολική διαδικασία μοντελοποίησης με χρήση δέντρων CHAID και εξαγωγή κανόνων.*

Οι σημαντικότερες παράμετροι και οι τιμές που επιλέχθηκαν για τον CHAID είναι:

- **Minimal size for split** = 4. Κάτω από αυτό το όριο, οι κόμβοι θεωρούνται τελικοί και δε διασπώνται περεταίρω.
- **Minimal leaf size** = 5. Το ελάχιστο μέγεθος που πρέπει να έχει ένα φύλλο για να κρατηθεί ως έγκυρο.
- **Minimal gain** = 0.1. Το όριο του κέρδους σε πληροφορία που θα προκύψει από τη διάσπαση ενός κόμβου με βάση κάποιο χαρακτηριστικό. Το όριο καθορίζει το αν θα διασπαστεί ή όχι ένας κόμβος.
- **Maximal depth** = 20. Το όριο για το μέγιστο βάθος του δέντρου (ο μέγιστος αριθμός κόμβων μέχρι τον οποίο μπορεί να αναπτυχθεί ένα κλαδί).
- **No pre-pruning**: απενεργοποιημένο, έτσι ώστε να επιτρέπεται το προ-κλάδεμα, δηλ η μη ανάπτυξη του δέντρου προς μια κατεύθυνση που κρίθηκε ότι δε θα συνεισφέρει στην προβλεπτική ικανότητα του δέντρου.
- **No pruning**: απενεργοποιημένο, έτσι ώστε να επιτρέπεται το κλάδεμα, δηλ η κατάργηση κλάδων που δε συνεισφέρουν ικανοποιητικά στην προβλεπτική ικανότητα το δέντρου.



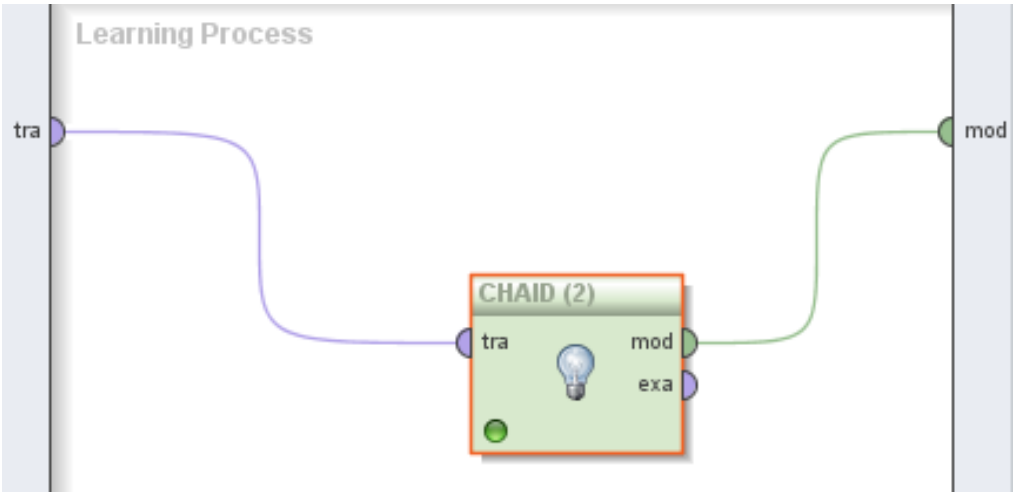
Σχήμα 7.49. *Το δέντρο πρόβλεψης της συνολικής ικανοποίησης από τα επιμέρους στοιχεία ικανοποίησης και τα χαρακτηριστικά του επισκέπτη.*

Για τη μετατροπή του δέντρου σε μορφή κανόνων χρησιμοποιήθηκε ο **Tree to Rules**, ο οποίος είναι σύνθετος, δηλαδή εμπεριέχει κάποιον άλλον τελεστή που καθορίζει την ανάπτυξη του δέντρου εκμάθησης, ο οποίος είναι συγκεκριμένα ο **CHAID**. Μετά την εισαγωγή του Tree to Rules στη διαδικασία, κάνουμε διπλό κλικ στο μπλε εικονιδιάκι του τελεστή αυτού, που βρίσκεται στο κάτω δεξιά μέρος του. Στο παράθυρο **Learning Process** που θα ανοίξει, εισάγουμε ένα νέο αντίγραφο του **CHAID**, ρυθμίζουμε τις παραμέτρους και επιστρέφουμε στην κύρια διαδικασία (Σχήμα 7.50).

Σημείωση: Στη διαδικασία που δημιουργήθηκε υπάρχουν τώρα 2 αντίγραφα του **CHAID**. Το ένα δημιουργεί το δέντρο και εξάγει το παραγόμενο μοντέλο στη θύρα εξόδου της διαδικασίας, κάτι που μας επιτρέπει να λάβουμε ως αποτέλεσμα το δέντρο. Το άλλο αντίγραφο βρίσκεται στο εσωτερικό του τελεστή εξαγωγής κανόνων και παράγει το δέντρο από όπου θα αντληθούν οι κανόνες, οι οποίοι, στη συνέχεια, οδηγούνται σε δεύτερη θύρα εξόδου και λαμβάνονται ως παράλληλο αποτέλεσμα. Η διάταξη αυτή ίσως δεν είναι η καλύτερη, γιατί το σύστημα αναγκάζεται να εκτελέσει δύο φορές την ίδια (αρκετά απαιτητική) διαδικασία. Ο λόγος για τον οποίο πραγματοποιήθηκε η σχεδίαση αυτή είναι καθαρά εκπαιδευτικός. Σε έναν πιο έμπειρο χρήστη συνιστάται να συμπεριληφθεί μόνο ο **CHAID** που βρίσκεται εντός του **Tree to Rules** και η επισκόπηση του δέντρου να γίνει εισάγοντας στον πρώτο ένα **Breakpoint** (βρίσκεται εύκολα στο πάνω δεξιά μέρος της προβολής σχεδίασης - περισσότερες πληροφορίες είναι διαθέσιμες στο εγχειρίδιο του RapidMiner).

Παρατηρώντας το δέντρο αποφάσεων που προκύπτει (Σχήμα 7.49), διαπιστώνουμε τα εξής ευρήματα:

- Το χαρακτηριστικό που τοποθετείται στη ρίζα του δέντρου είναι το **Sat_Personnel**, δηλαδή το χαρακτηριστικό με το οποίο επιτυγχάνεται ο πρώτος καλύτερος διαχωρισμός των πελατών στις κατηγορίες χαμηλής, μέτριας ή υψηλής συνολικής ικανοποίησης είναι η ικανοποίηση από το προσωπικό του ξενοδοχείου. Οι πελάτες που δηλώνουν χαμηλή ικανοποίηση από το προσωπικό (Low) δηλώνουν και χαμηλή συνολική ικανοποίηση (παρατηρείστε το φύλλο Low στα δεξιά του σχήματος). Αυτοί που δηλώνουν μέτρια ικανοποίηση από το προσωπικό, διασπώνται σε 3 υποομάδες, ανάλογα με το αν το ξενοδοχείο είναι μέρος αλυσίδας ή όχι (S_chain). Όσοι από αυτούς απαντούν High (δηλ το ξενοδοχείο τους είναι μέρος αλυσίδας), κατατάσσονται στη μέτρια συνολική ικανοποίηση, ενώ όσοι απαντούν Low ή Medium (δηλ το ξενοδοχείο τους δεν είναι μέρος αλυσίδας ή δεν τους ενδιαφέρει) κατατάσσονται στη χαμηλή συνολική ικανοποίηση.
- Όσοι είναι ικανοποιημένοι από το προσωπικό, διαχωρίζονται ανάλογα με το αν διαθέτει το ξενοδοχείο οικογενειακό χαρακτήρα. Ακολουθώντας μια διαδρομή που οδηγεί σε φύλλο χαρακτηρισμένο ως Low (αν π.χ. θέλουμε να διαπιστώσουμε τις συνθήκες που οδηγούν σε χαμηλή ικανοποίηση), παρατηρούμε π.χ. ότι αυτοί που είναι δυσαρεστημένοι από τον οικογενειακό χαρακτήρα του ξενοδοχείου και μετρίως ικανοποιημένοι από τα αναλώσιμα, έχουν χαμηλή συνολική ικανοποίηση.
- Από τα χαρακτηριστικά του πελάτη που εισήχθησαν στην ανάλυση (ηλικία, φύλο, μορφωτικό επίπεδο, χώρα, εισόδημα), παρατηρούμε ότι εμφανίζεται μόνο η ηλικία (Age), και μάλιστα σε χαμηλό σημείο του δέντρου. Αυτό δείχνει ότι τα δημογραφικά χαρακτηριστικά των πελατών δεν έχουν καλή προβλεπτική αξία της ικανοποίησης, σε σύγκριση με τις παροχές και τα χαρακτηριστικά του ξενοδοχείου.



Σχήμα 7.50. *Ο τελεστής εκμάθησης στο εσωτερικό του Tree to Rules είναι ο CHAID.*

Το αποτέλεσμα της διαδικασίας σε μορφή κανόνων είναι αυτό που φαίνεται στον Πίνακα 7.2. Οι κανόνες αυτοί αντιστοιχούν σε διαδρομές που μπορεί να ακολουθήσει κάποιος ξεκινώντας από τη ρίζα του δέντρου και καταλήγοντας, μέσω διακλαδώσεων, σε κάποιο φύλλο. Σημαντικό στοιχείο για την αξιολόγηση των κανόνων είναι ο αριθμός των παραδειγμάτων που κατατάσσονται με βάση αυτόν στον τελικό κόμβο (φύλλο) στον οποίο οδηγεί. Π.χ. ο πρώτος κανόνας του Πίνακα 7.2 οδηγεί στο συμπέρασμα High και τα στοιχεία (2/6/0) δείχνουν ότι ο κανόνας ισχύει για συνολικά 8 παραδείγματα, από τα οποία τα 2 είναι στην πραγματικότητα Medium και τα 6 High. Τα συγκεκριμένα νούμερα δείχνουν ότι ο κανόνας ισχύει για σχετικά λίγες περιπτώσεις και ότι δεν είναι απόλυτα ακριβής (αφού 2 παραδείγματα κατηγορίας Medium κατατάσσονται λανθασμένα ως High). Πολύ ισχυρότερος και ακριβέστερος κανόνας είναι ο

if Sat_personnel = Medium and S_chain = Low then Low (1 / 0 / 35)

Ο παραπάνω κανόνας ενεργοποιείται για 36 παραδείγματα, από τα οποία κατατάσσει ορθά στην κατηγορία Low τα 35.

RuleModel
if Sat_personnel = High and S_family_character = High and S_aura = High and S_sleep = High then High (2 / 6 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = High and S_sleep = Low then Medium (21 / 1 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = High and S_sleep = Medium then High (5 / 7 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = Low and Age = 3 then Medium (5 / 0 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = Low and Age = 4 then High (0 / 17 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = Low and Age = 5 then High (1 / 34 / 0)
if Sat_personnel = High and S_family_character = High and S_aura = Medium then High (17 / 24 / 2)
if Sat_personnel = High and S_family_character = Low and S_consumables = High and S_safety = High then Medium (72 / 27 / 0)
if Sat_personnel = High and S_family_character = Low and S_consumables = High and S_safety = Low then High (9 / 11 / 0)
if Sat_personnel = High and S_family_character = Low and S_consumables = High and S_safety = Medium then Low (0 / 0 / 5)
if Sat_personnel = High and S_family_character = Low and S_consumables = Low then Medium (13 / 0 / 0)
if Sat_personnel = High and S_family_character = Low and S_consumables = Medium then Low (0 / 0 / 10)
if Sat_personnel = High and S_family_character = Medium and S_internet = High then Medium (17 / 4 / 0)
if Sat_personnel = High and S_family_character = Medium and S_internet = Low then Low (5 / 0 / 18)
if Sat_personnel = High and S_family_character = Medium and S_internet = Medium then Medium (14 / 6 / 1)
if Sat_personnel = Low then Low (3 / 0 / 8)
if Sat_personnel = Medium and S_chain = High then Medium (7 / 0 / 0)
if Sat_personnel = Medium and S_chain = Low then Low (1 / 0 / 35)
if Sat_personnel = Medium and S_chain = Medium then Low (5 / 0 / 6)
correct: 330 out of 419 training examples.

Πίνακας 7.2. *Το μοντέλο πρόβλεψης της συνολικής ικανοποίησης από τα επιμέρους στοιχεία ικανοποίησης και τα χαρακτηριστικά του επισκέπτη σε μορφή κανόνων.*

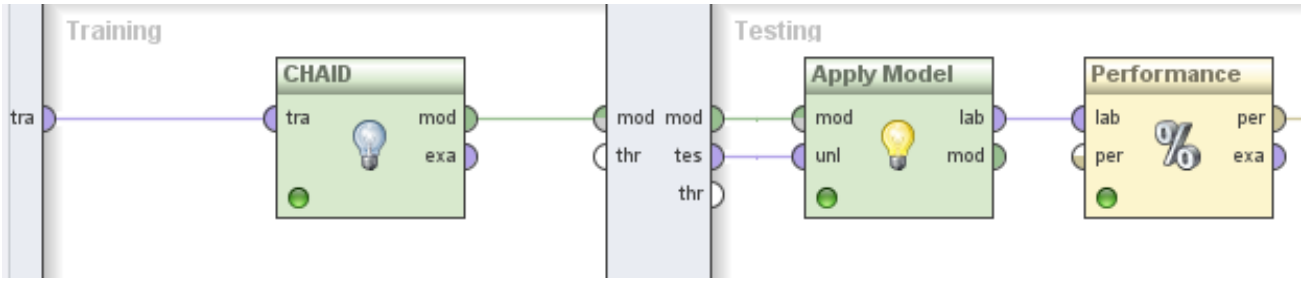
7.3.4.5. Εφαρμογή και αξιολόγηση του μοντέλου

Οι δύο μορφές του μοντέλου - σε μορφή δέντρου και σε μορφή κανόνων – είναι ουσιαστικά ισοδύναμες. Οι διαφορές είναι ότι ο δέντρο αποτελεί πιο συμπαγή και εύκολα παρατηρήσιμη μορφή, στην οποία είναι εμφανής ο ρόλος του κάθε χαρακτηριστικού, και επομένως προσφέρεται για την ποιοτική μελέτη και επεξήγηση του υπό διερεύνηση φαινομένου. Το μοντέλο σε μορφή κανόνων είναι προτιμότερο για την ευθεία εφαρμογή του στην πρόβλεψη άγνωστων παραδειγμάτων.

Ένα μέτρο της ακρίβειας του μοντέλου είναι η συνολική ακρίβεια που υπολογίζεται κατά τη δημιουργία των κανόνων. Η ακρίβεια αυτή φαίνεται στο κάτω μέρος του Πίνακα 7.2 και είναι μετρημένη ως 330 σωστά από τα 419 παραδείγματα, δηλαδή 78,7%. Κατά τη ρύθμιση του μοντέλου, μπορεί να γίνει προσπάθεια να βελτιωθεί η τιμή αυτή, σημειώνεται όμως ότι δεν αποτελεί απόλυτο κριτήριο για την επιτυχία ή όχι το μοντέλου. Για την ακριβέστερη αξιολόγηση, απαιτείται έλεγχος τύπου cross-validation, όπως παρουσιάστηκε στην ενότητα 3.1.7 του κεφαλαίου αυτού. Για το σκοπό αυτό, θα πρέπει να εισαχθεί στη διαδικασία ο τελεστής **X-Validation,** στο εσωτερικό του οποίου θα πρέπει να εισαχθούν ο **CHAID**, στο σκέλος της εκπαίδευσης, και οι **Apply Model** και **Performance**, για την εφαρμογή και μέτρηση της

ακρίβειας του μοντέλου (Σχήμα 7.51). Το αποτέλεσμα της αξιολόγησης φαίνεται στον Πίνακα 7.3. Εκεί φαίνεται ότι π.χ. από τα παραδείγματα που αντιστοιχούν στη χαμηλή ικανοποίηση (συνολικά 85), τα 73 κατατάσσονται ορθώς ως Low, τα 10 κατατάσσονται λανθασμένα ως Medium και τα 2 λανθασμένα ως High.

Σημείωση: η αξιοπιστία του μοντέλου, αλλά και ο πλούτος των ευρημάτων που θα αναδείξει, εξαρτώνται σε μεγάλο βαθμό από την ποσότητα των δεδομένων (και φυσικά και από την ποιότητα). Τα δέντρα αποφάσεων και οι μέθοδοι εξαγωγής κανόνων ταιριάζουν καλύτερα σε μεγάλα σετ δεδομένων. Ο λόγος είναι ότι, με τις διαδοχικές διασπάσεις κόμβων, μπορεί να δημιουργηθούν μικρές ομάδες που δεν εκπροσωπούνται αξιόπιστα. Στη συγκεκριμένη εφαρμογή, το μέγεθος του διαθέσιμου δείγματος θεωρείται μικρό, επομένως θα πρέπει να είμαστε επιφυλακτικοί ως προς την αξιοπιστία ορισμένων ευρημάτων.



Σχήμα 7.51. *Η υποδιαδικασία στο εσωτερικό του τελεστή Validation.*

	true Medium	**true High**	**true Low**	**class precision**
pred. Medium	129	32	10	75.44%
pred. High	53	104	2	65.41%
pred. Low	15	1	73	82.02%
class recall	65.48%	75.91%	85.88%	

Πίνακας 7.3. *Ο πίνακας σύγχυσης και η ακρίβεια πρόβλεψης κάθε κατηγορίας.*

Τα αποτελέσματα του cross-validation δείχνουν ότι το μοντέλο έχει αποκρυσταλλώσει με ικανοποιητική (αλλά όχι άριστη) ακρίβεια το φαινόμενο. Αν αποδεικνυόταν μη εφικτή η επίτευξη ικανοποιητικής ακρίβειας, το μοντέλο θα ήταν άχρηστο. Επόμενο βήμα είναι η ουσιαστική ποιοτική αξιολόγηση της ρεαλιστικότητας και της χρησιμότητας του μοντέλου από ένα στέλεχος διοίκησης/μάρκετινγκ και η αξιοποίησή του. Ενδεικτικά είναι τα ευρήματα και το σκεπτικό που αναφέρονται στην παράγραφο 3.4.4. Τονίζεται ότι η φιλοσοφία των μεθόδων που παρουσιάζονται είναι η ανάδειξη κρυμμένων προτύπων μέσα από τα δεδομένα, χωρίς εγγυημένο αποτέλεσμα ή αξιοπιστία. Επομένως είναι σημαντική η ανάπτυξη κρίσης και εμπειρίας από την πλευρά του ερευνητή. Επίσης, συνήθως απαιτείται πειραματισμός με διαφορετικές εναλλακτικές μεθόδους, αφού σε πολλές περιπτώσεις δεν υπάρχει μία μόνο σωστή διαδικασία, αλλά διαφορετικές προσεγγίσεις, με διαφορετικά πλεονεκτήματα η καθεμιά, μπορούν να προσφέρουν συμπληρωματική γνώση. Προτείνεται στον αναγνώστη να εφαρμόσει στα ίδια δεδομένα τις εναλλακτικές μεθόδους μοντελοποίησης που αναφέρθηκαν στην παράγραφο του σχεδιασμού (**Decision Tree** και **Rule Induction**) και να συγκρίνει τα αποτελέσματα.

Βιβλιογραφία/Αναφορές

Νανόπουλος Α., & Μανωλόπουλος Ι. (2008). *Εισαγωγή στην εξόρυξη και τις αποθήκες δεδομένων*, Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Laudon K., C., & Laudon J., P. (2009). *Πληροφοριακά Συστήματα Διοίκησης*, Ελληνική έκδοση, Αθήνα: Εκδόσεις Κλειδάριθμος.

Matlab, Mathworks (n.d.). Retrieved 30 October 2015 from http://www.mathworks.com/

North M. (2012). *Data Mining for the Masses*, licensed under a Creative Commons Attribution 3.0 License, Retrieved from Amazon.com, ISBN-13: 978-0615684376

Python (n.d.). Retrieved 30 October 2015 from https://www.python.org/

R The project for statistical computing (n.d.). Retrieved 30 October 2015 from https://www.r-project.org/

*Rapidminer, Predictive Analytics Reimagined (*n.d.*)*, Retrieved 30 October 2015 from https://rapidminer.com/

WEKA, The University of Waikato (n.d.). Retrieved 30 October 2015 from https://weka.wikispaces.com/

Κεφάλαιο 8. Μοντελοποίηση Γνώσης και Βάσεις Γνώσης

Σύνοψη
Η πρόκληση με την οποία ασχολούμαστε στο κεφάλαιο αυτό, είναι η αποτύπωση σε ηλεκτρονική μορφή της γνώσης που εξάγεται από δεδομένα, ώστε να μπορεί να εισαχθεί σε μια Βάση Γνώσης και να αξιοποιηθεί από ένα ευφυές σύστημα πληροφορικής. Σκοπός του κεφαλαίου είναι η εισαγωγή στη μοντελοποίηση γνώσης και ο διαχωρισμός των τεχνικών επιχειρηματικής ευφυΐας που βασίζονται σε γνώση από αυτές που βασίζονται στα δεδομένα και την πληροφορία. Παρουσιάζονται τα κυριότερα μοντέλα γνώσης, με ιδιαίτερη αναφορά στις οντολογίες και τα συστήματα κανόνων. Στο κεφάλαιο αυτό περιέχονται βασικά στοιχεία για την κατανόηση των εννοιών και της χρησιμότητας των αντίστοιχων τεχνολογιών. Για την καλύτερη κατανόηση της σχετικής φιλοσοφίας και την παρουσίαση των τεχνολογιών, παρατίθεται ένα ολοκληρωμένο παράδειγμα εφαρμογής από πραγματική έρευνα, ενός συστήματος στήριξης αποφάσεων βασισμένο σε γνώση και παρουσιάζεται η δομή του, το μοντέλο γνώσης που αναπτύχθηκε με βάση οντολογίες και μηχανή κανόνων, καθώς και η διαδικασία εξαγωγής συμπερασμάτων.

Προαπαιτούμενη γνώση
Κεφάλαιο 2. Δεδομένα και Πληροφορίες, Κεφάλαιο 6. Μέθοδοι εξόρυξης γνώσης από δεδομένα, Κεφάλαιο 7.Εφαρμογές επιχειρηματικής Ευφυΐας

8.1 Ορισμός και σημασία της Γνώσης

Η διαχείριση γνώσης (Knowledge Management) είναι η διαδικασία της αναγνώρισης, επιλογής, οργάνωσης και χρήσης των σημαντικών πληροφοριών και δεξιοτήτων, αποκαλούμενων «Γνώση», που παράγονται και χρησιμοποιούνται σε έναν οργανισμό. Η γνώση είναι κεφάλαιο «τεχνογνωσίας», «εμπειρίας» και «μνήμης» του οργανισμού. Αποτελεί κεφάλαιο που συσσωρεύεται με το χρόνο και έχει αξία. Χωρίς σύστημα διαχείρισης γνώσης, η γνώση είναι άτυπη, αδόμητη και δύσχρηστη, ενώ εξαρτάται αποκλειστικά από τους ανθρώπους, είναι εύκολο να χαθεί και δύσκολο να διακινηθεί και να αξιοποιηθεί πλήρως. Αντίθετα, η δόμηση και συστηματική διαχείριση της γνώσης επιτρέπει την αποδοτική και αποτελεσματική χρήση της στη λύση προβλημάτων, το σχεδιασμό και τη λήψη αποφάσεων.

Η γνώση ορίζεται ως επιλεγμένη πληροφορία που είναι εφαρμόσιμη στη λύση ενός προβλήματος και προσαρμόσιμη στο περιβάλλον και τις παραμέτρους του πραγματικού κόσμου. Διακρίνεται σε δύο βασικές κατηγορίες:

- **Υπονοουμένη γνώση (tacit knowledge)**, που είναι υποκειμενική και εμπειρική. Είναι με απλά λόγια, αυτό που «ξέρει» κάποιος να κάνει, χωρίς να μπορεί να περιγράψει ακριβώς το πώς και το γιατί. Αυτό το είδος γνώσης δύσκολα εκφράζεται, μεταφέρεται και εξάγεται από δεδομένα.
- **Κατηγορηματική γνώση (explicit knowledge)**, που είναι αντικειμενική, ορθολογιστική και τεχνική, και είναι δυνατόν να περιγραφεί σε διάφορες μορφές, όπως κανόνες και διαδικασίες.

Η αξία της γνώσης σε σχέση με την πληροφορία, εκτός του ότι είναι άμεσα εφαρμόσιμη στην επίλυση ενός προβλήματος, έγκειται και στο ότι είναι δυναμική στη φύση της και εξελίσσεται με το πέρασμα του χρόνου. Ενώ η πληροφορία χάνει την αξία της με τον καιρό και δύσκολα επαναχρησιμοποιείται, η γνώση συντηρείται, επεκτείνεται και γενικεύεται, ώστε να αποτελεί ένα διαχρονικό κεφάλαιο. Γενίκευση είναι όταν από έναν αριθμό περιπτώσεων (π.χ. τις απαντήσεις σε μια έρευνα) εξάγεται ένας γενικός κανόνας που ισχύει σε όλες τις αντίστοιχες περιπτώσεις (π.χ. η αποδοχή ενός προϊόντος από μια μερίδα καταναλωτών).

Η ψηφιακή τεχνολογία συμβάλλει στην ανάπτυξη και αξιοποίηση της γνώσης. Τα συστήματα πληροφορικής, μέσω ειδικών μεθόδων και τεχνολογιών, που θα παρουσιαστούν στο κεφάλαιο αυτό, επιτρέπουν τη μεταφορά γνώσης από τον πραγματικό κόσμο στον Η/Υ, ώστε να μπορούμε να την επεξεργαστούμε, να την αποθηκεύσουμε, να τη διαμοιράσουμε και να τη χρησιμοποιήσουμε αποτελεσματικά.

Σύμφωνα με την προσέγγιση αυτή, η γνώση αντιμετωπίζεται ως κάτι που μπορεί να συλλεχθεί, να παρασταθεί ως αντικείμενο, να συσκευαστεί και να διαμοιραστεί ως προϊόν ή να διαφυλαχθεί ως κεφάλαιο.

Ενώ η πληροφορία παράγεται από τα δεδομένα με κατάλληλη οργάνωση και επεξεργασία, όπως παρουσιάστηκε στο κεφάλαιο 5, η γνώση παράγεται από την πληροφορία με ειδικές μεθόδους ανάλυσης και «εξόρυξης γνώσης», όπως αυτές που παρουσιάστηκαν στα κεφάλαια 6 και 7. Η πρόκληση με την οποία ασχολούμαστε στο παρόν κεφάλαιο, είναι η αποτύπωση της εξαχθείσας γνώσης σε ηλεκτρονική μορφή, ώστε να μπορεί να εισαχθεί σε μια Βάση Γνώσης και να αξιοποιηθεί από ένα «ευφυές» σύστημα πληροφορικής. Η απλούστερη μορφή ηλεκτρονικής γνώσης είναι η ταξινομημένη πληροφορία, σε συνδυασμό με κάποιο προηγμένο μηχανισμό αναζήτησης, που είναι ερμηνεύσιμη μόνο από τον άνθρωπο. Στην περίπτωση αυτή το σύστημα δεν καταλαβαίνει το γνωστικό περιεχόμενο, αλλά λειτουργεί σαν μια αποθήκη πληροφοριών. Αντίθετα, αυτό στο οποίο αναφερόμαστε στο βιβλίο αυτό με τον όρο ηλεκτρονική διαχείριση γνώσης, είναι η παράστασή της σε μορφή που να μπορεί να καταλάβει και να επεξεργαστεί ένα σύστημα πληροφορικής, ώστε να μπορεί να προτείνει το ίδιο το σύστημα λύση σε συγκεκριμένα προβλήματα.

8.2 Μοντελοποίηση γνώσης

8.2.1 Σκοπός και διαδικασία μοντελοποίησης

Ένας από τους κύριους σκοπούς της Επιχειρηματικής Ευφυΐας, είναι η εξαγωγή από ένα σύνολο δεδομένων, είτε από το εσωτερικό είτε από το εξωτερικό περιβάλλον μιας επιχείρησης, της γνώσης που είναι χρήσιμη για επίλυση προβλημάτων στήριξης αποφάσεων και σχεδιασμού. Αντιπροσωπευτικές μέθοδοι και τεχνικές για το σκοπό αυτό παρουσιάστηκαν στο κεφάλαιο 6 και η εφαρμογή τους επιδείχθηκε στο κεφάλαιο 7. Το ευρύτερο πεδίο της παραγωγής, διαχείρισης και χρήσης γνώσης σε ηλεκτρονική μορφή είναι γνωστό ως Knowledge Engineering (Feigenbaum & McCorduck, 1983). Ένας μεγάλος αριθμός μεθόδων έχουν αναφερθεί σε αυτόν το χώρο τις τελευταίες δεκαετίες, που ωθήθηκαν από τη μεγάλη ανάπτυξη των τεχνολογιών πληροφορικής και οδήγησαν σε δημοφιλείς εφαρμογές στο χώρο του μάρκετινγκ και της διοίκησης επιχειρήσεων (Shadbolt & Milton, 1999).

Στο κεφάλαιο αυτό, εστιάζουμε στη μοντελοποίηση και διαχείριση της εξαχθείσας ηλεκτρονικής γνώσης, με σκοπό την αυτόματη συλλογιστική, σε αντιδιαστολή με τα συστήματα διαχείρισης πληροφορίας και τα συστήματα διαχείρισης γνώσης που περιορίζονται στη διαχείριση γνώσης κατανοητής από τον άνθρωπο. Αυτό που διαφοροποιεί τις εξεταζόμενες μεθόδους μοντελοποίησης γνώσης σε ηλεκτρονική μορφή από τα συστήματα που βασίζονται στην πληροφορία είναι ότι δεν περιορίζονται στην οργάνωση και διακίνηση πληροφορίας/γνώσης ώστε αυτή να είναι διαθέσιμη στον ειδικό για την επίλυση ενός προβλήματος, αλλά στοχεύουν στην αποτύπωση της γνώσης με τρόπο που να είναι κατανοητή από τον Η/Υ και να συνοδεύεται από εργαλεία επίλυσης προβλημάτων, ώστε το πρόβλημα να επιλύεται από τον ίδιο τον Η/Υ. Έτσι, ο χρήστης θα μπορεί να θέτει ερωτήματα δίνοντας παραμέτρους και να παίρνει απαντήσεις από μια συλλογιστική μηχανή, χωρίς να χρειάζεται να επιλύσει ο ίδιος όλα τα σκέλη του προβλήματος, αλλά να επικεντρώνεται στην αξιολόγηση των λύσεων και τη λήψη των τελικών αποφάσεων (Schreiber, 2008).

Ανάλογα με τη φύση του προβλήματος και των διαθέσιμων δεδομένων, η εξαγωγή γνώσης μπορεί να πραγματοποιηθεί από πολλές διαφορετικές μεθόδους, που μπορεί να βασίζονται σε στατιστική ανάλυση (κυρίως για ποσοτικά δεδομένα) ή σε αλγοριθμικές μεθόδους που επεξεργάζονται λογικές σχέσεις ανάμεσα σε ιδιότητες και ταιριάζουν σε ποιοτικά δεδομένα. Κάθε τέτοια μέθοδος, είτε ανήκει στην κατηγορία του Data Mining (ανακάλυψη προτύπων σε μεγάλους όγκους δεδομένων, π.χ. πωλήσεων σε συστήματα δοσοληψιών ή τα ημερολόγια μεγάλων ιστοτόπων), είτε στην κατηγορία της στατιστικής (π.χ. ανάλυση δεδομένων πρωτογενούς έρευνας), παράγει γνώση, η οποία, ανάλογα με τη φύση της, μοντελοποιείται και με διαφορετικό τρόπο. Υπάρχουν διάφορα θέματα προς επίλυση για την επιτυχημένη παράσταση της γνώσης ενός τομέα και την κατασκευή μιας βάσης γνώσης, καθώς και συγκεκριμένες προτεινόμενες μεθοδολογίες-πλαίσια. Τα κύρια ζητήματα που πρέπει να αντιμετωπιστούν κατά το σχεδιασμό και την ανάπτυξη ενός συστήματος βασισμένου σε γνώση (KBS) είναι τα ακόλουθα:

- Αναπαράσταση της γνώσης, που αφορά την επιλογή της κατάλληλης γλώσσας για την κωδικοποίηση της αποκτηθείσας γνώσης.
- Απόκτηση γνώσης, που είναι η διαδικασία εξαγωγής της γνώσης πεδίου, που μπορεί να κατέχει ένας ειδικός ή που κρύβεται στα δεδομένα.

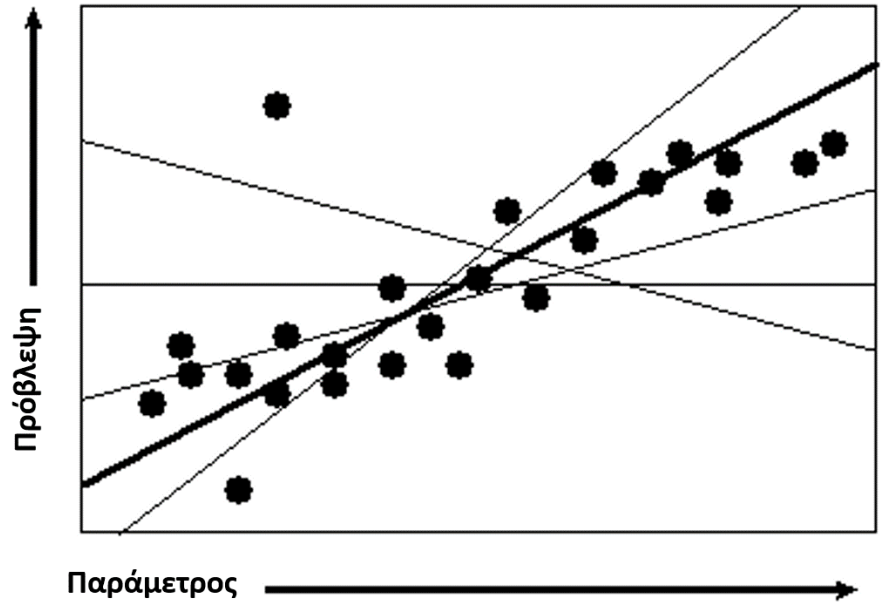
- Ανάπτυξη μηχανισμού εξαγωγής συμπερασμάτων, που αφορά το σχεδιασμό και υλοποίηση ενός μεταφραστή, ικανού να ερμηνεύει και να επεξεργάζεται τα στοιχεία της κωδικοποιημένης γνώσης.
- Ανάπτυξη μηχανισμού ελέγχου του συλλογισμού, που περιέχει γνώση σε υψηλότερο επίπεδο (μετα-γνώση) σχετικά με τον τρόπο οργάνωσης και εφαρμογής της κύριας γνώσης κατά τη λειτουργία των συλλογισμών.
- Επαλήθευση της γνώσης, δηλαδή έλεγχος για την ορθότητα της Βάσης Γνώσης.
- Εύρεση και εξήγηση λύσεων, που σχετίζεται με την αλληλεπίδραση ανθρώπου-συστήματος και αφορά την παρουσίαση των λύσεων στον χρήστη, καθώς και του πού βασίστηκαν και με ποιον τρόπο προέκυψαν οι λύσεις αυτές.
- Ανάπτυξη διεπαφών χρήστη που να καλύπτουν όλες τις λειτουργίες εισαγωγής και πρόσβασης γνώσης, εξαγωγής συμπερασμάτων και επαλήθευσης.

Μοντελοποίηση Γνώσης είναι η διαδικασία αναπαράστασης και κωδικοποίησης της γνώσης και της λογικής λειτουργίας της σε μορφή που να μπορούμε να τη διαχειριστούμε σε ένα σύστημα πληροφορικής. Ως μοντέλο γνώσης εννοούμε μια παράσταση ή έκφραση του τμήματος του πραγματικού κόσμου που μας ενδιαφέρει για την επίλυση ενός προβλήματος. Ο σχεδιασμός του κατάλληλου Μοντέλου Γνώσης (Knowledge Model - KM) είναι η σημαντικότερη πρόκληση στην ανάπτυξη ενός Συστήματος Βασισμένου σε Γνώση (KBS). Το KM πρέπει να διαθέτει τις απαιτούμενες δυνατότητες εκφραστικότητας (Expressiveness) που καλύπτουν τις ανάγκες του συγκεκριμένου προβλήματος, όχι μόνο σε ορθότητα και αποτελεσματικότητα, αλλά και σε δυνατότητες επαναχρησιμοποίησης / διαμοιρασμού του περιεχομένου, επεκτασιμότητας και συντηρησιμότητας.

Έχουν προταθεί αρκετές μεθοδολογίες σχεδιασμού και υλοποίησης ενός KM (Ligeza, 2006, Schreiber et al, 1999), όπως συστήματα βασισμένα σε κανόνες (Rule-based systems), μοντέλα βασισμένα σε εκμάθηση περιπτώσεων (case-based reasoning), οντολογίες, σημασιολογικά δίκτυα, νευρωνικά δίκτυα, κλπ. Η διαδικασία μοντελοποίησης θεωρείται γενικά κυκλική και υποκειμενική, με την έννοια ότι μπορεί να υπάρχουν περισσότερες από μία λύσεις, που όλες αποτελούν ατελείς αναπαραστάσεις του πραγματικού κόσμου και οφείλουν να προσαρμόζονται συνεχώς σε ένα κόσμο που επίσης μεταβάλλεται. Είναι επίσης αναγνωρισμένο ότι υπάρχουν διαφορετικοί τρόποι αναπαράστασης γνώσης και ότι ο τρόπος με τον οποίο γίνεται η σύλληψη και αναπαράσταση κάποιου προβλήματος επηρεάζει την ποιότητα της λύσης που θα επιτευχθεί. Επομένως η επιλογή του κατάλληλου σχήματος αναπαράστασης, δηλαδή του κατάλληλου πλαισίου μοντελοποίησης γνώσης και του κατάλληλου μοντέλου, είναι κρίσιμα για την επιτυχία του εγχειρήματος. Στη συνέχεια, παρουσιάζονται τα σημαντικότερα Μοντέλα Γνώσης, μερικά από τα οποία έχουν ήδη αναφερθεί στα Κεφάλαια 6 και 7 ως μέθοδοι εξαγωγής γνώσης. Στο κεφάλαιο αυτό, εστιάζουμε περισσότερο σε δύο από τους βασικότερους τύπους, που είναι οι Οντολογίες και τα Μοντέλα Κανόνων.

8.2.2 Στατιστικά μοντέλα

Τα στατιστικά μοντέλα βασίζονται στην εκτίμηση των παραμέτρων ενός αριθμητικού μοντέλου (π.χ. των συντελεστών μιας εξίσωσης) από έναν αριθμό δειγμάτων. Στη συνέχεια, το μοντέλο αυτό χρησιμοποιείται για την πρόβλεψη μελλοντικών τιμών, θεωρώντας ότι προσεγγίζει το σχετικό φαινόμενο. Αντιπροσωπευτικές μέθοδοι είναι οι διάφοροι τύποι παλινδρόμησης και γραμμικής ή μη γραμμικής προσέγγισης και η μέθοδος κατάταξης του Bayes. Το πεδίο εφαρμογής των στατιστικών μοντέλων περιορίζεται σε ποσοτικά δεδομένα, ενώ το αποτέλεσμά τους μπορεί να είναι μια ποσοτική εκτίμηση ή η πρόβλεψη ενός στοιχείου, όπως η κατάταξη ενός ατόμου σε μια συγκεκριμένη κατηγορία. Πλεονέκτημα των στατιστικών μοντέλων αποτελεί το ότι, μαζί με το αποτέλεσμα, παρέχουν και θεωρητικά θεμελιωμένη εκτίμηση της αξιοπιστίας του αποτελέσματος.

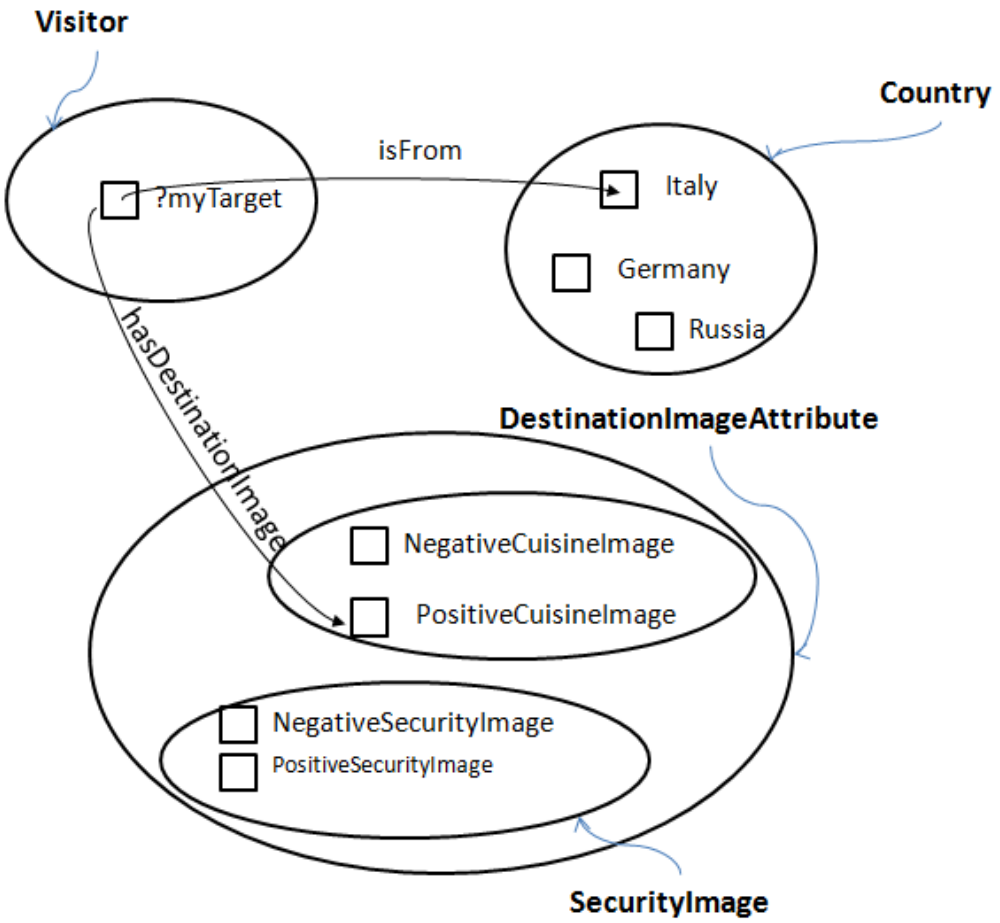


Σχήμα 8.1. *Η γραμμική προσέγγιση ως απλό στατιστικό μοντέλο πρόβλεψης*

8.2.3 Οντολογίες

Οι οντολογίες αποτελούνται από ορισμούς εννοιών και αντικειμένων ώστε να μπορεί να εκφραστεί ο πραγματικός κόσμος στη γλώσσα του Η/Υ (Gruber, 1993). Μια οντολογία αποτελεί μια περιγραφή ενός πεδίου ενδιαφέροντος, που περιλαμβάνει ορισμούς εννοιών, αντικειμένων, τύπων αντικειμένων (κλάσεων) και σχέσεων μεταξύ τους, παρέχοντας μια εννοιολογική βάση για την επίλυση προβλημάτων. Περικλείουν γνώση για το πώς είναι οργανωμένος ο χώρος που μας ενδιαφέρει και είναι χρήσιμες ως ορολογία ή λεξιλόγιο, που θα αποτελέσει τη βάση για να εκφραστούν οι κανόνες σε μοντέλα κανόνων. Επίσης, μπορεί να διαμοιραστεί μεταξύ συστημάτων και έτσι επιτρέπει την ενοποίηση ετερογενών πηγών πληροφορίας. Κυρίως όμως, προσφέρει μια προτυποποιημένη βασική ορολογία που είναι προαπαιτούμενο για τη διατύπωση της γνώσης με τυπικό τρόπο και στη συνέχεια την εφαρμογή λογικών αναλύσεων.

Στον ορισμό των εννοιών, τύπων και αντικειμένων μέσα σε μια οντολογία, υπάρχει ιεραρχική δομή (π.χ. το παντελόνι και η μπλούζα ανήκουν στα ενδύματα, τα ενδύματα ανήκουν στα προϊόντα) και κληρονομικότητα στα χαρακτηριστικά (π.χ. η τιμή είναι χαρακτηριστικό του τύπου προϊόντα και επομένως ισχύει κληρονομικά για τα ενδύματα, τα τρόφιμα, κλπ.).



Σχήμα 8.2. *Απόσπασμα οντολογίας από το χώρο του τουριστικού μάρκετινγκ*

Το παράδειγμα του σχήματος 8.2 είναι η σχηματική παράσταση ενός αποσπάσματος οντολογίας που αναφέρεται στον τουρισμό και περιλαμβάνει τις έννοιες του επισκέπτη, της χώρας και της εικόνας ενός τουριστικού προορισμού. Φαίνονται ως κύκλοι οι κλάσεις (Classes), που μπορεί να περιλαμβάνουν υποκλάσεις (subclasses) και ως τετραγωνάκια τα αντικείμενα (instances ή individuals) που περιλαμβάνονται σε μια κλάση, π.χ. τα αντικείμενα *Italy*, *Germany*, *Russia* ανήκουν στην κλάση *Country*. Οι ιδιότητες, που παριστάνονται ως βέλη, μπορεί να συνδέουν μεταξύ τους αντικείμενα ή κλάσεις και έχουν συγκεκριμένο πεδίο ορισμού και πεδίο τιμών. Π.χ. η ιδιότητα **isFrom** έχει πεδίο ορισμού την κλάση *Visitor* και πεδίο τιμών την κλάση *Country*, δηλαδή συνδέει έναν επισκέπτη με μια χώρα, προσδίδοντάς του την ιδιότητα ότι «προέρχεται από αυτήν τη χώρα». Η κλάση *DestinationImageAttribute* περιλαμβάνει τα στοιχεία της εικόνας ενός τουριστικού προορισμού και περιλαμβάνει τις υποκλάσεις *SecurityImage* και *CuisineImage*, που αναφέρονται στην εικόνα για την ασφάλεια και την εικόνα για την κουζίνα, αντίστοιχα. Κάθε υποκλάση περιλαμβάνει ως αντικείμενα την αρνητική και τη θετική εικόνα για το συγκεκριμένο στοιχείο. Έτσι, στο παράδειγμα του σχήματος, με την ιδιότητα **hasDestinationImage** προσδίδουμε σε κάποιον επισκέπτη την ιδιότητα *PositiveCuisineImage*, δηλ θετική εικόνα για την κουζίνα.

Οι οντολογίες καλύπτουν διάφορα γνωστικά πεδία, είτε γενικά είτε ειδικότερα, και είναι κατά κανόνα δημοσιευμένες στο σημασιολογικό ιστό ώστε να επιτρέπουν την ανταλλαγή γνώσης μεταξύ συστημάτων. Σε μια πρόσφατη ανασκόπηση των διαθέσιμων οντολογιών για το χώρο του τουρισμού (Prantner et al, 2007), βρέθηκε ότι υπάρχει ένας σημαντικός αριθμός από δημοσιευμένες οντολογίες, όπως οι QUALL-ME (Ou et al, 2008) και DERI e-tourism (DERI, n.d.). Ωστόσο διαπιστώθηκε ότι όλες οι προσπάθειες, εκτός από ένα γενικό μέρος, είναι εξειδικευμένες σε συγκεκριμένα προβλήματα και είναι απίθανο κάποια από αυτές να καλύψει πλήρως τις ανάγκες του προβλήματος ενός συγκεκριμένου έργου. Επομένως, για την επίλυση ενός προβλήματος, η συνηθέστερη προσέγγιση είναι η υιοθέτηση μιας πολύ-τμηματικής οντολογίας, που να αποτελείται από την εισαγωγή κατάλληλων υπαρχόντων οντολογιών ως τμήματα (modules) και η

συμπλήρωσή της με τμήματα εξειδικευμένα στο δικό μας πρόβλημα. Το εξειδικευμένο μέρος της οντολογίας που αφορά τη συγκεκριμένη εφαρμογή, πρέπει να αναπτυχθεί με επιμέλεια, ώστε να αντιστοιχεί στις πηγές γνώσης και να προβλεφθεί δυνατότητα επέκτασης, έτσι ώστε το μοντέλο να μπορεί να προσαρμόζεται δυναμικά στις μελλοντικές απαιτήσεις του προβλήματος.

8.2.4 Μηχανές κανόνων (Rule-based systems)

Τα μοντέλα γνώσης που βασίζονται σε κανόνες, περιλαμβάνουν κανόνες της μορφής:

Αν C_1 και C_2 και ... και C_n → Ε

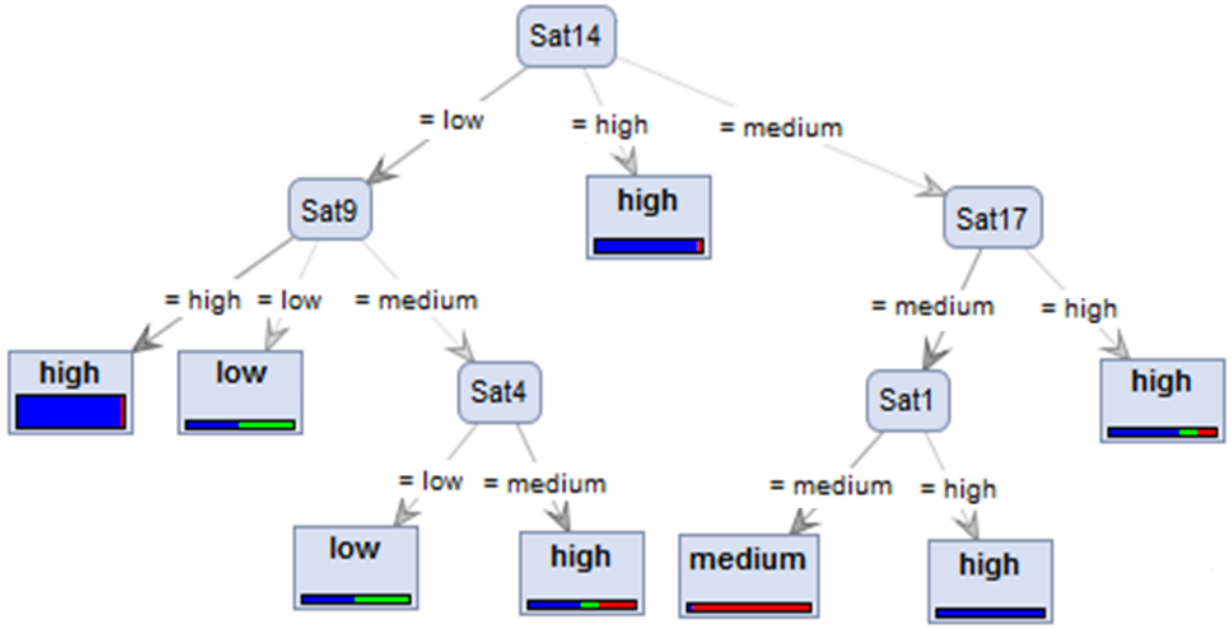
Όπου C_1, ..., C_n είναι συνθήκες που συζευκτικά μεταξύ τους απαρτίζουν το υποθετικό μέρος του κανόνα και Ε είναι το αποτέλεσμα. Κάθε συνθήκη αποτελεί έναν όρο της λογικής παράστασης του κανόνα και μπορεί να περιλαμβάνει το αν ένα αντικείμενο έχει μια ιδιότητα ή αν ανήκει σε μια κατηγορία ή το αν ισχύει μια παράσταση που συνδέει μεταβλητές και τιμές. Το αποτέλεσμα Ε παριστάνει ένα λογικό συμπέρασμα που μπορεί να αντιστοιχεί σε μια πρόβλεψη ή μια σύσταση ή ένα ενδιάμεσο αποτέλεσμα όπως π.χ. η πρόσδωση μιας ιδιότητας σε κάποιο αντικείμενο (Ligêza, 2006). Κανόνες αυτού του τύπου μπορούν να αποθηκευτούν σε Βάση Γνώσης και να χρησιμοποιηθούν από μηχανές εξαγωγής συμπερασμάτων.

Στα προηγούμενα κεφάλαια, έγινε αναφορά σε κανόνες αυτής της βασικής μορφής και παρουσιάστηκαν εφαρμογές εξαγωγής τους. Οι κανόνες αυτοί δημιουργήθηκαν από μεθόδους εξαγωγής γνώσης, ωστόσο προορίζονταν για ερμηνεία από τον άνθρωπο-αναλυτή. Για να είναι ο κανόνας κατανοητός από μια συλλογιστική μηχανή, είναι απαραίτητο η κάθε συνθήκη αλλά και το συνεπαγόμενο αποτέλεσμα να είναι διατυπωμένα με αυστηρά προσδιορισμένη ορολογία. Σε απλές περιπτώσεις, για αυτόν το σκοπό χρησιμοποιούνται ονόματα μεταβλητών και μαθηματικές εκφράσεις (π.χ. Αν Ηλικία > 35 → Καλός_πελάτης = 1). Σε πιο σύνθετες περιπτώσεις, είναι απαραίτητος ο ορισμός ειδικού δομημένου λεξιλογίου και η διατύπωση συνθηκών με χρήση ιδιοτήτων που έχουν οριστεί για τη συγκεκριμένη εφαρμογή. Αυτό είναι δυνατό όταν το μοντέλο κανόνων βασίζεται σε μια οντολογία. Ως τελεστές για τη διατύπωση παραστάσεων μπορούν να χρησιμοποιηθούν βασικοί τελεστές που ορίζονται στη γλώσσα της οντολογίας (π.χ. OWL) και οι ιδιότητες (properties) που περιέχει η οντολογία. Είναι λοιπόν σαφές ότι το μοντέλο κανόνων εξαρτάται άμεσα από την οντολογία, εφόσον για να είναι δυνατή η διατύπωση ενός κανόνα, πρέπει να υπάρχει η πρόβλεψη για τις απαραίτητες κλάσεις και ιδιότητες στον ορισμό της οντολογίας.

Σημαντικό πλεονέκτημα των μοντέλων κανόνων είναι ότι οι κανόνες έχουν φυσική ερμηνεία, δηλαδή η κάθε συνθήκη και το αποτέλεσμα του κανόνα έχουν κάποιο ξεκάθαρο νόημα. Επίσης, για κάθε προβλεπόμενο αποτέλεσμα, το σύστημα δίνει εξήγηση για το ποιες συνθήκες και ποιοι κανόνες οδήγησαν σε αυτό.

8.2.5 Δέντρα αποφάσεων

Το αποτέλεσμα των μεθόδων εξόρυξης που βασίζονται στην κατασκευή δέντρων, είναι τα δέντρα αποφάσεων, που αποτελούν και μοντέλα της εξαχθείσας γνώσης. Τα μοντέλα αυτά αποτελούνται από ένα σύνολο διακλαδώσεων που έχουν τη μορφή ενός δέντρου, που ξεκινάει από μια ρίζα και περιλαμβάνει κλαδιά που καταλήγουν σε φύλλα. Κάθε μονοπάτι που ξεκινάει από τη ρίζα και καταλήγει σε ένα φύλλο, ακολουθώντας μια σειρά διακλαδώσεων, αποτελεί και μια πιθανή απόφαση, αφού η κάθε διακλάδωση αποτελεί μια συνθήκη που καθορίζει με βάση ένα κριτήριο το αποτέλεσμα προς το οποίο θα οδηγηθεί η διαδικασία απόφασης. Τα δέντρα αποφάσεων είναι κατάλληλα για λήψη αποφάσεων και για αυτόματη κατάταξη, έχοντας ως πλεονέκτημα το ότι παρέχουν ποσοτική εκτίμηση της αξιοπιστίας του συμπεράσματος.



Σχήμα 8.3. *Παράδειγμα δέντρου αποφάσεων για την πρόβλεψη της συνολικής ικανοποίησης το πελάτη από τα επιμέρους στοιχεία ικανοποίησης.*

8.2.6 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα προσομοιώνουν τη λειτουργία του ανθρώπινου εγκεφάλου και έχουν την ικανότητα να εκπαιδεύονται με τη βοήθεια παραδειγμάτων. Ένα νευρωνικό δίκτυο μπορεί να μάθει από έναν αριθμό αντιπροσωπευτικών ερωτημάτων (παραδείγματα), αποκρυσταλλώνοντας γνώση σχετικά με ένα φαινόμενο και γενικεύοντας τη γνώση αυτή ώστε ένα εκπαιδευμένο νευρωνικό δίκτυο να μπορεί να δίνει απαντήσεις με αντίστοιχο τρόπο σε άγνωστα ερωτήματα.

Τα νευρωνικά δίκτυα αποτελούν μοντέλα γνώσης που δεν προφέρουν ερμηνεία του φαινομένου, αφού δε μπορούμε να ξέρουμε το λόγο για τον οποίο έδωσαν μια συγκεκριμένη απάντηση. Έχουν περίπλοκη δομή και η γνώση που περιέχουν είναι αποθηκευμένη σε ένα μεγάλο αριθμό συντελεστών που δεν έχουν νόημα για τον άνθρωπο. Λειτουργούν ποσοτικά, εκτελώντας μεγάλο πλήθος αριθμητικών υπολογισμών, αλλά το αποτέλεσμά τους είναι ποιοτικό (π.χ. απόφαση, κατάταξη).

8.3 Βάσεις Γνώσης και Συστήματα Διαχείρισης Γνώσης

Το Σύστημα Διαχείρισης Γνώσης (Knowledge Mangement System – KMS) έχει ως σκοπό να συσσωρεύσει τα αποτελέσματα των μεθόδων εξόρυξης γνώσης και να τα διαθέσει προς χρήση σε μορφή ηλεκτρονικής γνώσης προς χρήστες μη-ειδικούς στην ανάλυση δεδομένων και που δε χρειάζεται να έχουν πρόσβαση στα αρχικά δεδομένα. Μπορεί να χρησιμοποιηθεί για στήριξη αποφάσεων ή στην υποβοήθηση σχεδιασμού.

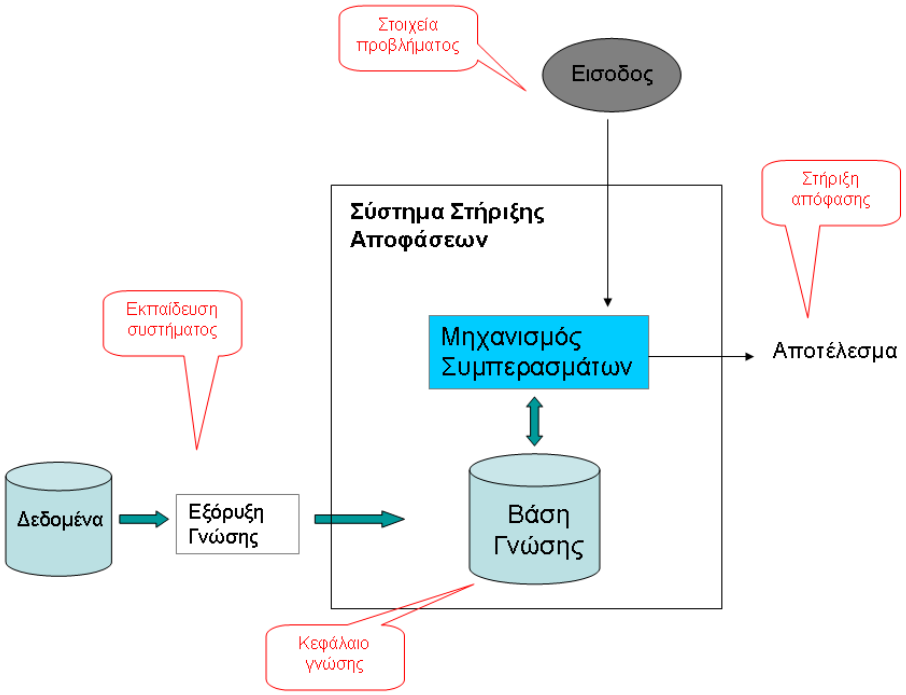
Το KMS είναι σχεδιασμένο γύρω από ένα Μοντέλο Γνώσης (ΚΜ), που μπορεί να είναι οποιοδήποτε από αυτά της προηγούμενης ενότητας ή ένα σύνθετο μοντέλο που αποτελείται από πολλά επιμέρους τμήματα διαφορετικού τύπου. Το ΚΜ επιλέγεται και διαμορφώνεται ανάλογα με τη φύση του προβλήματος, έτσι ώστε να μπορεί να εκφράσει όλη την απαραίτητη γνώση που απαιτείται και αποτελεί μια έκφραση του μέρους του πραγματικού κόσμου που μας ενδιαφέρει.

Ένα KMS περιλαμβάνει:

1. Μια Βάση Γνώσης, όπου είναι αποθηκευμένη όλη η γνώση που έχει συλλεγεί. Η Βάση Γνώσης είναι σχεδιασμένη σύμφωνα με το ΚΜ.

2. Ένα μηχανισμό εξαγωγής συμπερασμάτων, ο οποίος δέχεται σαν είσοδο τις παραμέτρους ενός προβλήματος, χρησιμοποιεί τη γνώση που βρίσκεται στη Βάση Γνώσης και παρέχει ως έξοδο το συμπέρασμα.

Για να μπορέσει να λειτουργήσει ένα KMS, πρέπει να έχει προηγηθεί η φάση εισαγωγής περιεχομένου ή εκπαίδευσής του, δηλαδή η εξαγωγή γνώσης από τα δεδομένα και η εισαγωγή της στη Βάση Γνώσης. Κατά τη φάση της λειτουργίας του συστήματος ως σύστημα Στήριξης Αποφάσεων, ο χρήστης αρχικά εισάγει τα στοιχεία του προβλήματος. Στη συνέχεια, ένας ευφυής μηχανισμός χρησιμοποιεί την αποθηκευμένη γνώση και τα δεδομένα εισόδου για να παράγει τη λύση στο πρόβλημα (Σχήμα 8.4). Ο ευφυής μηχανισμός που θα χρησιμοποιηθεί εξαρτάται από το ΚΜ, π.χ. στην περίπτωση μοντέλου κανόνων είναι ένας μηχανισμός συμπερασμάτων (Inference Engine), στην περίπτωση του νευρωνικού δικτύου είναι ο μηχανισμός ανάκλησης (Recall) του νευρωνικού δικτύου, στην περίπτωση ενός στατιστικού μοντέλου είναι το κατάλληλο σύστημα εξισώσεων, ενώ στην περίπτωση της οντολογίας, η απάντηση δίνεται από ένα μηχανισμό λογικών ερωτημάτων (Description Logic Query).



Σχήμα 8.4. *Σύστημα Διαχείρισης Γνώσης που χρησιμοποιείται για υποστήριξη απόφασης*

8.4 Παράδειγμα εφαρμογής στη στήριξη αποφάσεων μάρκετινγκ τουριστικών προορισμών

8.4.1 Σκοπός και πεδίο εφαρμογής

Η εφαρμογή Επιχειρηματικής Ευφυΐας που παρουσιάζεται ως παράδειγμα στην ενότητα αυτή, προέρχεται από ερευνητικό έργο και περιλαμβάνει αποτελέσματα πραγματικής έρευνας σχετικά με τον τουρισμό της Βορείου Ελλάδος. Αφορά την ανάπτυξη ενός Συστήματος Στήριξης Απόφασης για το μάρκετινγκ τουριστικών προορισμών. Το σύστημα βασίζεται σε Σύστημα Διαχείρισης Γνώσης (KMS), που είναι ικανό να χειριστεί τα αποτελέσματα ανάλυσης δεδομένων ως επαναχρησιμοποιούμενη «Γνώση», σε μορφή κανόνων. Τα δεδομένα που αποτελούν είσοδο για την εφαρμογή προέρχονται από πρωτογενή έρευνα με ερωτηματολόγιο. Η

εξαγωγή γνώσης πραγματοποιήθηκε με μεθόδους πολυδιάστατης στατιστικής ανάλυσης, η οποία κρίθηκε ως καταλληλότερη ώστε να αναδυθούν πρότυπα και τάσεις, να διαπιστωθούν οι παράγοντες που επηρεάζουν τις αποφάσεις των τουριστών και να εντοπιστούν συσχετίσεις ανάμεσα σε παραμέτρους της τουριστικής αγοράς, που προσδίδουν βέλτιστες προοπτικές επιτυχίας.

Στόχος της εφαρμογής ήταν, αντί της οπτικής επισκόπησης από τον ερευνητή, τα αποτελέσματα της ανάλυσης να συσσωρευτούν σε Βάση Γνώσης, ώστε να μπορούν να αξιοποιηθούν από ένα στέλεχος διοίκησης και προβολής τουριστικών προορισμών, μέσω ενός ευφυούς συστήματος Στήριξης Αποφάσεων. Με τον τρόπο αυτό, τα συμπεράσματα περισσότερων επιμέρους ερευνών θα μπορούν να συγκεραστούν, θα κεφαλαιοποιούνται και η γνώση θα γενικεύεται σε μορφή κανόνων και μοντέλων πρόβλεψης, ανεξάρτητα από τα αρχικά δεδομένα. Η γνώση αυτή θα μπορεί να εξελίσσεται προσθέτοντας νέα αποτελέσματα ερευνών και θα είναι άμεσα εφαρμόσιμη από μη-αναλυτές στο σχεδιασμό πιο αποτελεσματικών προωθητικών ενεργειών, στην επιλογή αγοράς στόχου, στη βελτίωση τουριστικών πακέτων και σε αποφάσεις σχετικά με βελτιωτικές παρεμβάσεις στον τουριστικό προορισμό.

8.4.2 Πηγές γνώσης

Οι πηγές από τις οποίες προήλθε το γνωστικό περιεχόμενο του παραπάνω συστήματος ήταν δύο ανεξάρτητες πρωτογενείς έρευνες, που πραγματοποιήθηκαν κατά το διάστημα 2012-13 στη Βόρεια Ελλάδα. Οι δύο έρευνες πραγματοποιήθηκαν με τη βοήθεια ερωτηματολογίων σε δείγμα περίπου 2000 και 400 ξένων επισκεπτών, αντίστοιχα. Η πρώτη αφορούσε την αντιλαμβανόμενη εικόνα της πόλης της Θεσσαλονίκης ως τουριστικού προορισμού, όπως εκφράζεται από διαφόρων τύπων επισκέπτες και η ανάλυση των παραγόντων που επηρεάζουν τη διαμόρφωση της εικόνας αυτής, την επιλογή του προορισμού και την ικανοποίηση. Το ερωτηματολόγιο περιελάμβανε στοιχεία για την εμπειρία των επισκεπτών και τα δημογραφικά χαρακτηριστικά τους, την εικόνα που έχουν για ένα σύνολο επιμέρους στοιχείων της πόλης και της χώρας, τις προτεραιότητές τους στην επιλογή του προορισμού και τις προσδοκίες τους, καθώς και την ικανοποίησή τους.

Η δεύτερη έρευνα αφορούσε τις απαιτήσεις και την ικανοποίηση επισκεπτών από το ξενοδοχείο τους. Εκτός από δημογραφικά χαρακτηριστικά του επισκέπτη και το σκοπό του ταξιδιού του, το ερωτηματολόγιο ζητούσε από τους ερωτώμενους να προσδιορίσουν σε σχέση με ένα σύνολο πιθανών χαρακτηριστικών του ξενοδοχείου, τις προσδοκίες τους και το βαθμό ικανοποίησής του για το καθένα, καθώς και τη συνολική ικανοποίησή τους.

Η ανάλυση των πρωτογενών δεδομένων για την εξαγωγή γνώσης πραγματοποιήθηκε με μεθόδους πολυδιάστατης παραγοντικής ανάλυσης, και συγκεκριμένα με εφαρμογή συνδυασμού πολλαπλής ανάλυσης αντιστοιχιών και ανιούσας ιεραρχικής ταξινόμησης. Περισσότερες πληροφορίες για τις έρευνες αυτές, την ανάλυση και τα αποτελέσματα είναι διαθέσιμα στα άρθρα που αναφέρονται στη βιβλιογραφία (Stalidis & Karapistolis, 2014a), (Stalidis & Karapistolis, 2014b). Το στοιχείο που είναι σημαντικό να αναφερθεί στο σημείο αυτό είναι οι ιδιαιτερότητες των συγκεκριμένων μεθόδων ανάλυσης σε σχέση με κλασικές ποσοτικές μεθόδους, που είναι η ικανότητά τους (α) να αναδείξουν σύνθετες σχέσεις ανάμεσα σε μεγάλο αριθμό μεταβλητών, χωρίς να υπάρχει ανάγκη προσδιορισμού εκ των προτέρων κάποιου μοντέλου, (β) να συμπεριλάβουν μεγάλο αριθμό ποιοτικών μεταβλητών (π.χ. φύλο, ύπαρξη ή όχι υπηρεσίας, θετική/αρνητική εικόνα, κλπ.), χωρίς την ανάγκη τεχνητής ποσοτικοποίησης με χρήση κλιμάκων και (γ) να συνθέσουν προφίλ συμπεριφοράς και να προσδιορίσουν αντιπροσωπευτικές ομάδες ατόμων. Τα αποτελέσματα των αναλύσεων είναι σε μορφή συσχετίσεων ανάμεσα σε χαρακτηριστικά, ομαδώσεων που ορίζουν κλάσεις και την τυπολογία τους, καθώς και κατάταξης ατόμων σε ομοιογενείς ομάδες. Αυτό σημαίνει ότι τα αποτελέσματα αυτά είναι ποιοτικού και όχι ποσοτικού χαρακτήρα και το μοντέλο που θα ταίριαζε να τα εκφράσει πρέπει να μπορεί να παραστήσει τύπους, συσχετίσεις και λογικές σχέσεις.

Μεταξύ των αποτελεσμάτων της έρευνας σχετικά με την τουριστική εικόνα της Θεσσαλονίκης, αναφέρεται ενδεικτικά ότι από την ανάλυση των παραγόντων που οδήγησαν στην επιλογή της πόλης προέκυψαν 4 κλάσεις επισκεπτών: (1) αυτοί που ελκύονται κυρίως από τη φήμη και την ιστορία της πόλης, ενδιαφέρονται για επισκέψεις σε μουσεία και θεωρούν τις υποδομές ως σημαντικό παράγοντα (στην κατηγορία αυτή κατατάχθηκε το 27,4%), (2) αυτοί που ελκύονται από τη νυχτερινή ζωή και το lifestyle, την ελληνική κουζίνα και επηρεάστηκαν από φίλους (8,7%), (3) αυτοί που ενδιαφέρονται για τις φυσικές ομορφιές, το κλίμα, το φυσικό περιβάλλον και τις ευκαιρίες για εκδρομές (54%) και (4) οι τουρίστες που ενδιαφέρονται αποκλειστικά για τις παραλίες (9,9%) .

8.4.3 Το Μοντέλο Γνώσης

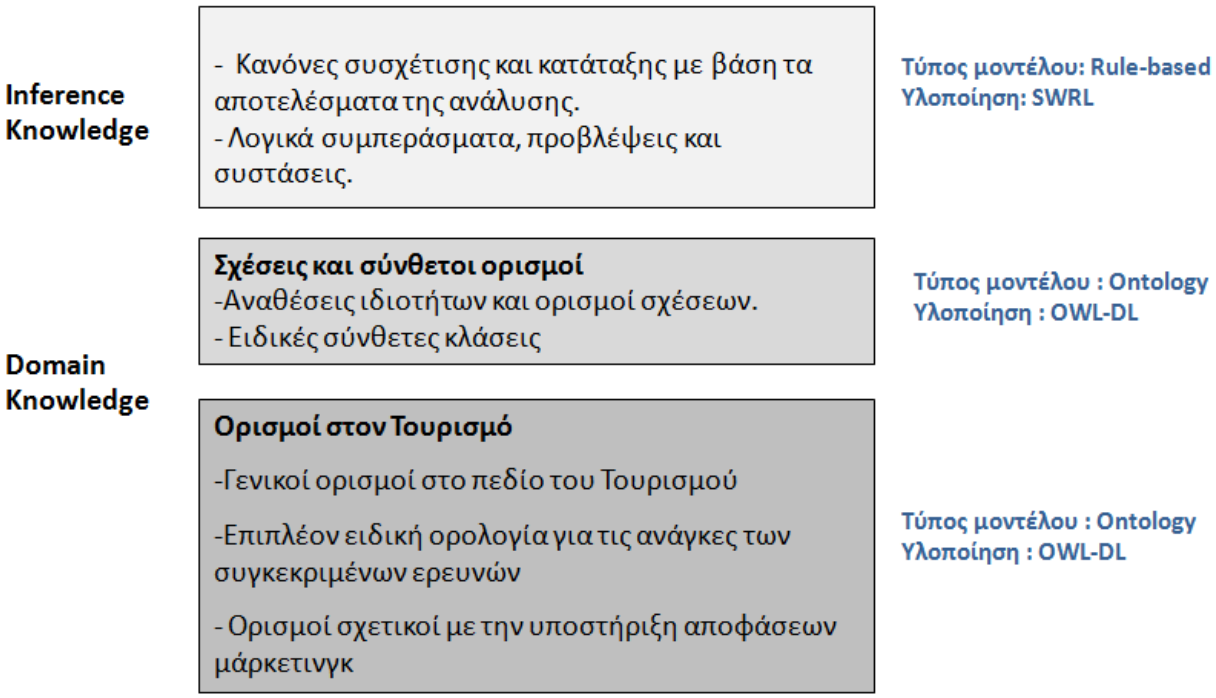
8.4.3.1 Τεχνολογίες μοντελοποίησης

Έχουν προταθεί διάφορα τεχνολογικά πλαίσια για τη μοντελοποίηση γνώσης, όπως το Common KADS Schreiber et al, 1999) and Model-based and Incremental Knowledge Engineering (MIKE). Στην παρούσα εφαρμογή, υιοθετήθηκε το πλαίσιο που υποστηρίζεται από την πλατφόρμα Protégé (Protégé, n.d.), η οποία βασίζεται σε οντολογίες και αποτελεί ένα ευρέως διαδεδομένο εργαλείο, όχι μόνο για την ανάπτυξη μοντέλων, αλλά και για το διαμοιρασμό δομημένης γνώσης μέσω του σημασιολογικού ιστού. Ως βασική πλατφόρμα για την ανάπτυξη του ΚΜ, την υλοποίηση της Βάσης Γνώσης και του μηχανισμού στήριξης απόφασης, χρησιμοποιήθηκε η έκδοση Protégé OWL 4.2, που βασίζεται στη γλώσσα Web Ontology Language (OWL) (OWL, 2013).

Η OWL υποστηρίζει σημασιολογικούς ορισμούς και μηχανισμούς λογικών συνεπαγωγών που επιτρέπουν όχι μόνο την έκφραση γνωστών γεγονότων αλλά και την εξαγωγή των λογικών τους συνεπαγωγών, έτσι ώστε να παράγεται γνώση που δεν έχει ρητά δηλωθεί. Η OWL έχει επίσης τη δυνατότητα ενσωμάτωσης λογικής, επιπλέον των ορισμών της οντολογίας, καθώς και ανταλλαγής γνώσης μέσω του σημασιολογικού ιστού. Αναφέρεται ακόμα ότι η OWL διατίθεται σε 3 διαφορετικούς τύπους, με διαφορετικό επίπεδο εκφραστικότητας, ενώ ήδη είναι διαθέσιμη η OWL 2, με ακόμα περισσότερους τύπους, βελτιστοποιημένους ο καθένας και για διαφορετικές κατηγορίες εφαρμογών. Έχει επιλεγεί ο τύπος OWL Description Logic, ως ο καταλληλότερος για εφαρμογές με έμφαση στις λογικές συνεπαγωγές, με υποστήριξη ειδικής γλώσσας για υποβολή λογικών ερωτημάτων (DL-queries). Τέλος, αναφέρεται ότι χρησιμοποιήθηκε η γλώσσα SWRL (Semantic Web Rule Language) (SWRL, 2014) για τη σύνταξη κανόνων, η οποία υποστηρίζεται από το Protégé και συνεργάζεται άμεσα με τις οντολογίες σε OWL.

8.4.3.2 Δομή Μοντέλου Γνώσης

Ο σχεδιασμός του Μοντέλου Γνώσης πραγματοποιήθηκε έτσι ώστε ταιριάζει στις διαθέσιμες πηγές γνώσης, δηλαδή στην περίπτωσή μας, έτσι ώστε να μπορεί να εκφράσει τα αποτελέσματα της πολυδιάστατης ανάλυσης δεδομένων, κατά την εφαρμογή της σε δεδομένα πρωτογενών ερευνών με ερωτηματολόγια. Σχεδιάστηκε έτσι ώστε να μπορεί να δεχτεί αποτελέσματα περισσότερων συμπληρωματικών ή διαδοχικών ερευνών και επίσης να εκφράσει διαθέσιμη πληροφορία από δευτερογενείς έρευνες. Κατά δεύτερο λόγο, το μοντέλο αναπτύχθηκε σύμφωνα με τη χρήση για την οποία προορίζεται η Βάση Γνώσης, δηλαδή για την παροχή προτάσεων προς τους μαρκετίστες για την υποστήριξη αποφάσεων μάρκετινγκ τουριστικών προορισμών. Τέλος, το μοντέλο θεωρήθηκε δυναμικό, δηλαδή ανοιχτό σε μελλοντικές επεκτάσεις/τροποποιήσεις, ώστε να προσαρμόζεται στη νέα γνώση και επεκτάσιμο σε νέες πιθανές πηγές γνώσης, όπως π.χ. γνώση εξαχθείσα από χρήστες του διαδικτύου.



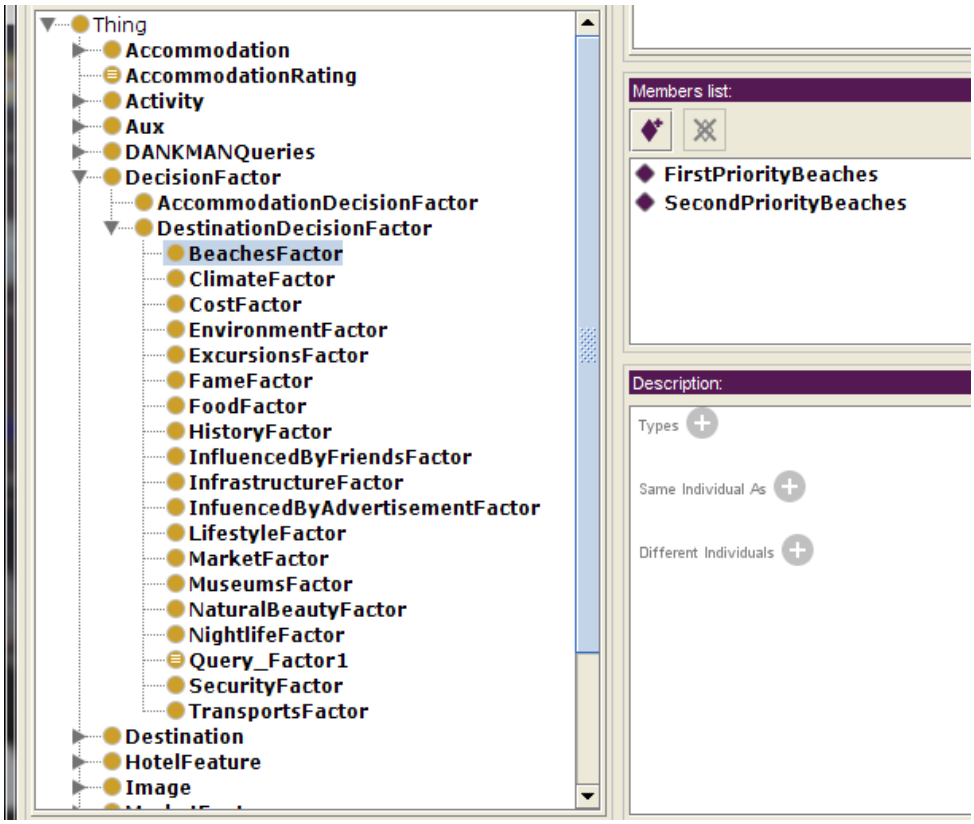
Σχήμα 8.5. *Η δομή του Μοντέλου Γνώσης περιλαμβάνει 3 επίπεδα.*

Η δομή του μοντέλου φαίνεται στο Σχήμα 8.5 και περιλαμβάνει τα τμήματα: (α) ορολογία στο πεδίο του τουρισμού, (β) σχέσεις και ειδική ορολογία προσαρμοσμένη στο πρόβλημα και (γ) συμπερασματική γνώση.

Η ανάπτυξη του μοντέλου ώστε να μπορεί να εκφράσει αποτελεσματικά την απαιτούμενη γνώση για το συγκεκριμένο πρόβλημα και η εισαγωγή των αποτελεσμάτων της ανάλυσης στη Βάση Γνώσης ήταν μια διαδικασία πολλαπλών βημάτων που παραμένει δυναμική σε όλη τη λειτουργία του συστήματος. Η διαδικασία αυτή περιλαμβάνει την ερμηνεία των αποτελεσμάτων της ανάλυσης δεδομένων, την αξιολόγηση, επαλήθευση και επιλογή τους, καθώς και την οργάνωση/ομαδοποίησή τους ώστε να μπορούν να κωδικοποιηθούν σύμφωνα με το παραπάνω μοντέλο 3 επιπέδων. Στη συνέχεια παρουσιάζεται αναλυτικότερα το περιεχόμενο του κάθε επιπέδου.

8.4.3.3 Ορολογία στο πεδίο του τουρισμού

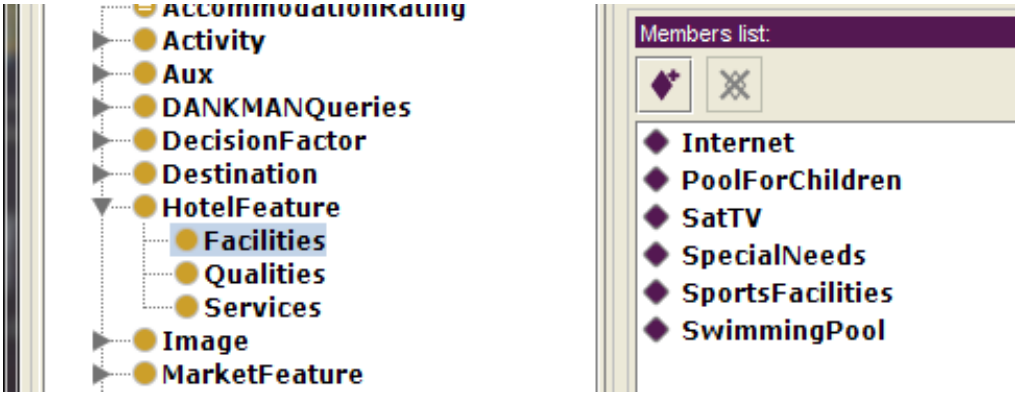
Το τμήμα αυτό περιλαμβάνει τη βασική ορολογία σχετικά με τις έννοιες που περιλαμβάνονται στο συγκεκριμένο πρόβλημα (π.χ. επισκέπτης, προορισμός, ταξίδι, ξενοδοχείο, κλπ.) και τα χαρακτηριστικά τους (π.χ. ο επισκέπτης έχει ηλικία, χώρα προέλευσης, επίπεδο εκπαίδευσης, κλπ.). Η ορολογία αυτή είναι οργανωμένη ιεραρχικά σε κλάσεις (δηλ τύποι αντικειμένων), υποκλάσεις (δηλ πιο ειδικοί τύποι που κληρονομούν τα χαρακτηριστικά των ανώτερων κλάσεων στις οποίες ανήκουν) και άτομα (individuals) (που μπορεί να ανήκουν σε κάποιες κλάσεις και να συνδέονται με κάποιες ιδιότητες). Για παράδειγμα, οι κλάσεις Hotel και Camping είναι ειδικοί τύποι (υποκλάσεις) της κλάσης Accommodation, ενώ τα άτομα RoomService, Internet και SwimmingPool ανήκουν στην κλάση HotelFeature. Μέρος της οντολογίας περιέχει βασικούς ορισμούς του χώρου του τουρισμού, που μπορεί να θεωρηθούν σταθεροί, και ένα μέρος εξαρτάται από το συγκεκριμένο πρόβλημα και επεκτείνει τους βασικούς ορισμούς, ώστε να υποστηρίζονται οι ειδικές ανάγκες των πηγών γνώσης του προβλήματος. Στο τελευταίο περιλαμβάνονται όροι για να εκφράσουν μεταβλητές που χρησιμοποιήθηκαν στο ερωτηματολόγιο (π.χ. η εικόνα που έχει ο επισκέπτης για τον προορισμό, οι παράγοντες επιλογής του, κλπ.). Ενδεικτικά: Η κλάση *DestinationDecisionFactor* και οι υποκλάσεις *BeachesFactor, ClimateFactor*, κλπ., που περιλαμβάνουν αντικείμενα του τύπου *FirstPriorityBeaches, SecondPriorityBeaches*, κλπ. (Σχήμα 8.6) καλύπτουν τις ανάγκες να εκφραστούν οι απαντήσεις στην ενότητα του ερωτηματολογίου της έρευνας σχετικά με τους παράγοντες επιλογής του προορισμού. Η κλάση *VisitorFeature* περιλαμβάνει ως υποκλάσεις τα χαρακτηριστικά του επισκέπτη που προκύπτουν από την ενότητα του ίδιου ερωτηματολογίου (π.χ. ηλικία, χώρα, κλπ.).



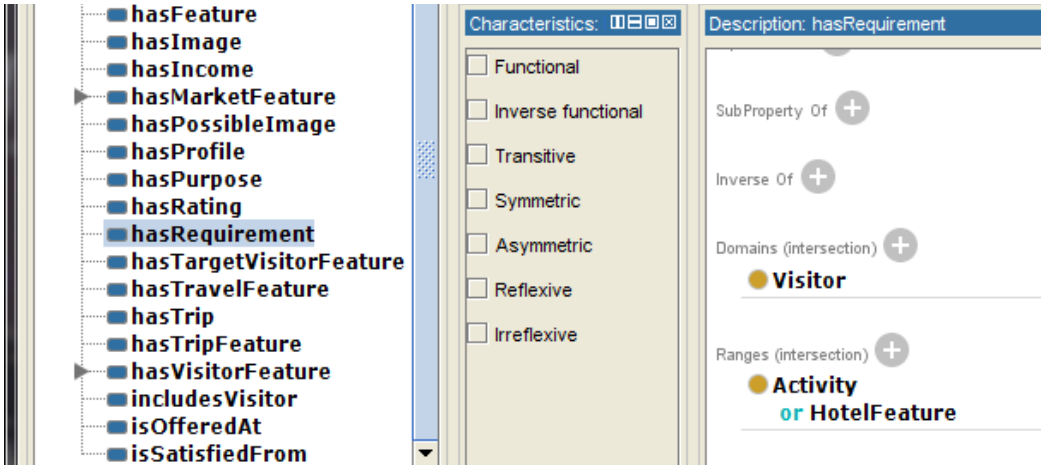
Σχήμα 8.6. *Το μέρος της οντολογίας που αφορά τους παράγοντες επιλογής προορισμού που μπορεί να εκφράσει ένας επισκέπτης(σε περιβάλλον Protégé-OWL 4.2).*

Από την έρευνα σχετικά με την ικανοποίηση των επισκεπτών από το ξενοδοχείο τους προέρχεται η κλάση HotelFeature, που περιλαμβάνει τις υποκλάσεις Facilities, Qualities και Services, οι οποίες περιλαμβάνουν ως αντικείμενα τις εγκαταστάσεις, χαρακτηριστικά και υπηρεσίες που μπορεί να διαθέτει ένα ξενοδοχείο και για τις οποίες ερωτάται ο επισκέπτης (Σχήμα 8.7). Επίσης έχει οριστεί η ιδιότητα hasRequirement, που συνδέει έναν επισκέπτη με κάποια από τα παραπάνω στοιχεία ως απαιτήσεις, και η isSatisfiedFrom, που συνδέει έναν επισκέπτη με τα χαρακτηριστικά από τα οποία είναι ικανοποιημένος/δυσαρεστημένος. Στο Σχήμα 8.8 φαίνεται ότι σύμφωνα με τον ορισμό της ιδιότητας hasRequirement, το πεδίο ορισμού της είναι η κλάση *Visitor* και το πεδίο τιμών της οι κλάσεις *Activity* και *HotelFeature*, δηλ η ιδιότητα αυτή μπορεί να συνδέει έναν επισκέπτη με μια δραστηριότητα ή ένα χαρακτηριστικό ξενοδοχείου.

Επίσης, έχουν συμπεριληφθεί ορισμοί που απαιτούνται για τη λήψη αποφάσεων. Έχουν οριστεί έννοιες όπως το τμήμα αγοράς (κλάση *MarketSegment*) και τα χαρακτηριστικά της αγοράς όπως η κλάση *MarketFeature* με υποκλάσεις τις *MarketRepeatVisit* (τάση να επαναλάβουν την επίσκεψη στην ίδια πόλη)και *MarketValue*, που η τελευταία περιλαμβάνει τις *MarketAcceptance*, *MarketExpenses*, *MarketShare* και *MarketSize*.



Σχήμα 8.7. *Οι ορισμοί των χαρακτηριστικών/υπηρεσιών των ξενοδοχείων που μπορεί να επιθυμεί ή από τα οποία μπορεί να είναι ευχαριστημένος/δυσαρεστημένος ένας πελάτης.*



Σχήμα 8.8. *Ο ορισμός της ιδιότητας hasRequirement, που μπορεί να συνδέει έναν επισκέπτη με μια δραστηριότητα ή ένα χαρακτηριστικό ξενοδοχείου.*

8.4.3.4 Σχέσεις και ειδική ορολογία προσαρμοσμένη στο πρόβλημα

Επιπλέον των βασικών ορισμών, ένα σημαντικό στοιχείο της προσέγγισης που ακολουθήθηκε στη μοντελοποίηση ήταν ο ορισμός ομάδων ατόμων ή ιδιοτήτων ή με άλλα λόγια συστάδες ή κλάσεις που προέκυψαν από το στάδιο Ανάλυσης Δεδομένων, έτσι ώστε να είναι δυνατή η αναφορά σε αυτές στη Βάση Γνώσης, η περιγραφή τους και η ανάθεση ιδιοτήτων σε αυτές. Για παράδειγμα, από την ανάλυση της ενότητας του ερωτηματολογίου σχετικά με την εικόνα της Θεσσαλονίκης, προέκυψαν 5 κλάσεις επισκεπτών με διαφορετικά κριτήρια ο καθένας στην επιλογή προορισμού του. Έτσι, ορίστηκε π.χ. η κλάση *VisitorForNature*, που αντιστοιχεί στον επισκέπτη που επέλεξε τον προορισμό του με προτεραιότητα τις φυσικές ομορφιές, το κλίμα και τις εκδρομές.

Αντίστοιχα, από την ανάλυση προέκυψαν συγκεκριμένες κλάσεις επισκεπτών με βάση την εικόνα τους για τον προορισμό (π.χ. NegativeWithStyle, PositiveWithAll, κλπ.) ή τις απαιτήσεις από το ξενοδοχείο τους, αλλά και πιο σύνθετες δομές από συνδυασμό μεταβλητών, όπως 5 κλάσεις στις οποίες διαχωρίστηκαν οι επισκέπτες με βάση τη χώρα προέλευσής τους και το πόσες φορές έχουν έρθει στο παρελθόν στο ίδιο προορισμό. Οι κλάσεις που ορίστηκαν για την έκφραση των τύπων αυτών, ονομάστηκαν σύνθετες (composite classes) και δεν ήταν γνωστές εκ των προτέρων αλλά η εισαγωγή τους στην οντολογία ήταν δυνατή μόνο μετά την εύρεσή τους κατά την ανάλυση δεδομένων της πρωτογενούς έρευνας. Τέτοιου είδους κλάσεις αναμένεται να προκύψουν και στο μέλλον, όσο θα προστίθενται στη Βάση Γνώσης αποτελέσματα νέων ερευνών και το μοντέλο θα εξελίσσεται.

8.4.3.5 Συμπερασματική γνώση

Το 3ο και υψηλότερο επίπεδο του μοντέλου εκφράζει τις σύνθετες συσχετίσεις ανάμεσα στις κλάσεις, τα άτομα και τις ιδιότητές τους, σε μορφή κανόνων. Έτσι ενσωματώνεται λογική και υπολογισμοί, ώστε να είναι δυνατό να εκφραστούν προβλέψεις ή συστάσεις που προκύπτουν από ένα σύνολο συνθηκών και παραμέτρων εισόδου. Εφόσον η συνθετότητα της γνώσης που πρέπει να διαχειριστούμε είναι μεγαλύτερη από αυτήν της ιεραρχικής οργάνωσης σε κλάσεις και των σχέσεων μεταξύ ατόμων, είναι απαραίτητη η προσθήκη ενός μοντέλου κανόνων για την επίτευξη της επιπλέον εκφραστικότητας που απαιτείται. Η μορφή των κανόνων είναι αυτή της παραγράφου 2.3, δηλαδή αποτελούνται από μια σειρά συνθηκών, που αν ισχύουν όλες μαζί, οδηγούν σε ένα αποτέλεσμα. Οι συνθήκες και το αποτέλεσμα εκφράζονται χρησιμοποιώντας ως λεξικό την οντολογία OWL. Για τη σύνταξη των κανόνων χρησιμοποιήθηκε η γλώσσα SWRL (SWRL, 2015), οποία βασίζεται στην OWL και υποστηρίζεται από το περιβάλλον ανάπτυξης Protégé.

Παράδειγμα κανόνων: Η ανάλυση των χαρακτηριστικών των ταξιδιών με μέθοδο συσταδοποίησης ανέδειξε 5 προφίλ ταξιδιών, μεταξύ των οποίων το τυπικό ταξίδι για διακοπές με βασικές απαιτήσεις, που συνήθως έχει διάρκεια 2 εβδομάδων και το κόστος του δωματίου που επιλέγεται είναι μεταξύ 50 και 100€. Ονομάστηκε StandardQualityVacation. Συσχετίζοντας τα χαρακτηριστικά των ταξιδιών με το δημογραφικό προφίλ των επισκεπτών και τις απαιτήσεις τους από το ξενοδοχείο, προέκυψε ότι οι επισκέπτες που έρχονται για τυπικό ταξίδι διακοπών βασικών απαιτήσεων, όταν είναι της ηλικιακής κατηγορίας 56-65 ετών, έχουν ως βασική απαίτηση την ασφάλεια και την καθαριότητα (αυτό είναι μόνο ένα από τα πολλά ευρήματα που αναφέρεται ως παράδειγμα). Τα ευρήματα αυτά διατυπώνονται ως κανόνες ως εξής:

- **Αν** το ταξίδι είναι τύπου «διακοπές» **και** το ταξίδι έχει ως ιδιότητα τη διάρκεια 2 εβδομάδων **και** το ταξίδι έχει ως ιδιότητα το κόστος 50-100€ **τότε** το ταξίδι είναι τύπου StandardQualityVacation
  ή σε γλώσσα SWRL:
- *Vacation(?mytrip)*, hasTravelFeature *(?mytrip, 2WeekVisit*), hasTravelFeature*(?mytrip, Cost50to100*) → *StandardQualityVacationTrip(?mytrip)*
  Και
- **Αν** το ταξίδι είναι τύπου StandardQualityVacation **και** ο επισκέπτης έχει ηλικία 56-65 **τότε** ο επισκέπτης έχει ως ιδιότητα την απαίτηση Ασφάλεια **και** ο επισκέπτης έχει ως ιδιότητα την απαίτηση Καθαριότητα
  ή σε γλώσσα SWRL:
- *StandardQualityVacationTrip(?mytrip)*, hasAge*(?myvisitor, Age56-65*) → hasRequirement*(?myvisitor, Security*), hasRequirement*(?myvisitor, Cleanness*)

Η εφαρμογή σχετικά με τον τουρισμό της Θεσσαλονίκης είχε συνολικά περισσότερους από 150 κανόνες, που προέκυψαν από την ανάλυση των δύο πρωτογενών ερευνών.

8.4.4 Εξαγωγή συμπερασμάτων και στήριξη απόφασης

Η αξιοποίηση της γνώσης του συστήματος από ένα στέλεχος διοίκησης/μάρκετινγκ γίνεται με τη βοήθεια ερωτημάτων. Ο μηχανισμός ερωτημάτων υποστηρίζεται από μια συμπερασματική μηχανή (Inference Engine) που είναι ενσωματωμένη στο περιβάλλον διαχείρισης γνώσης Protégé. Η μηχανή αυτή, εφαρμόζει στη δηλωμένη γνώση τα αξιώματα λογικής και τους κανόνες που έχουμε εισάγει και εξάγει ως αποτέλεσμα τη συνεπαγόμενη γνώση και την απάντηση στο ερώτημα.

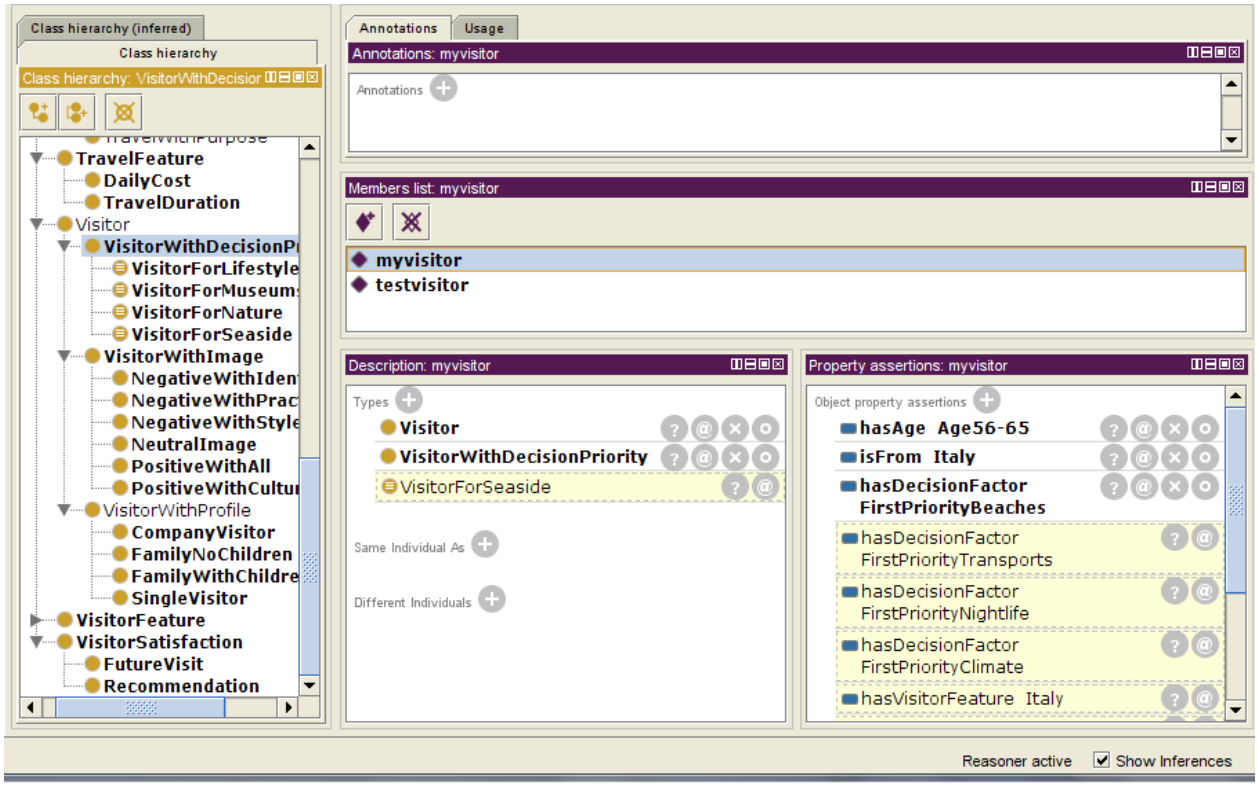
Παράδειγμα: Ας υποθέσουμε ότι επιθυμούμε να μελετήσουμε τους επισκέπτες που έρχονται από την Ιταλία για διακοπές στη θάλασσα και ότι ο στόχος μας είναι η ηλικιακή κατηγορία 56-65. Δημιουργούμε ως μεταβλητή ένα άτομο με όνομα π.χ. *myvisitor*, που να εκφράζει τον άγνωστο επισκέπτη, και το εντάσσουμε στην κλάση Visitor. Μετά, προσδίδουμε στο myvisitor τις εξής ιδιότητες, που εκφράζουν τις παραμέτρους εισόδου:

- hasAge *Age56-65*
- isFrom *Italy*
- hasDecisionFactor *FirstPriorityBeaches*

Μόλις ενεργοποιήσουμε τη συμπερασματική μηχανή, η εκτέλεση των κανόνων έχει ως αποτέλεσμα την πρόσδωση επιπλέον ιδιοτήτων στον myvisitor, που αποτελούν τη συνεπαγόμενη γνώση:

- Ο myvisitor εντάσσεται στην κλάση *VisitorForSeaside* (ειδικός τύπος επισκέπτη που ενδιαφέρεται κυρίως για τις παραλίες)
- hasDecisionFactor *FirstPriorityTransports*
- hasDecisionFactor *FirstPriorityNightlife*
- hasDecisionFactor *FirstPriorityClimate*

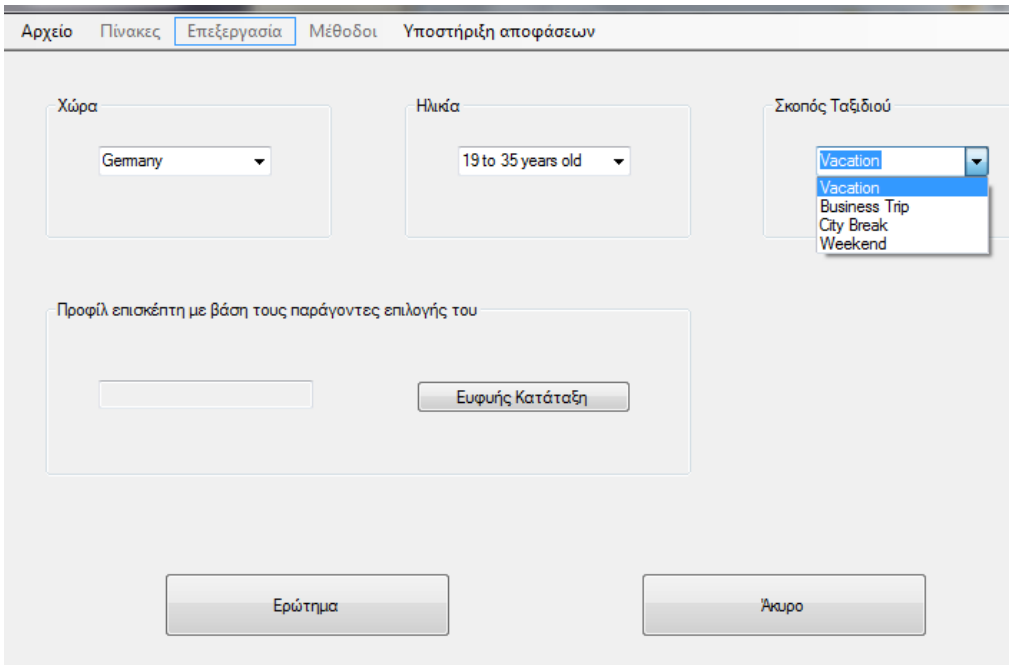
Στο Σχήμα 8.9 φαίνονται στο περιβάλλον Protégé οι ιδιότητες του υποθετικού επισκέπτη myvisitor (δεξιά). Οι ιδιότητες που εμφανίζονται σε λευκό φόντο είναι η δηλωμένη γνώση (αυτές που δώσαμε εμείς για να ορίσουμε το πρόβλημα) και με κίτρινο φόντο αυτές που συμπέρανε ο μηχανισμός με βάση τους κανόνες. Σε κατάλληλη γλώσσα υποβολής ερωτημάτων, που λέγεται Description Logic Query (DL-query) και αποτελεί μέρος της OWL (δεν είναι σκόπιμη λεπτομερέστερη αναφορά στο σημείο αυτό), μπορούμε να υποβάλλουμε ερωτήματα προς τη συνεπαγόμενη οντολογία όπως π.χ. ποια είναι η εικόνα του επισκέπτη για τη Θεσσαλονίκη, ποιες είναι οι προτεραιότητές του στην επιλογή προορισμού και στην επιλογή ξενοδοχείου, ποια η αξία του μεριδίου αγοράς που του αντιστοιχεί, ποια η δεκτικότητα αυτής της αγοράς, κ.ά.



Σχήμα 8.9. *Οι δηλωμένες και οι συνεπαγόμενες ιδιότητες του υπό διερεύνηση υποθετικού επισκέπτη*

Στο παραπάνω παράδειγμα, οι ενέργειες που έγιναν από το χρήστη για να πάρει απάντηση στο ερώτημά του πραγματοποιήθηκαν μέσα στο περιβάλλον της πλατφόρμας διαχείρισης γνώσης, κάτι που απαιτεί σχετική εξοικείωση. Για να γίνει το ευφυές σύστημα πιο φιλικό προς ένα χρήστη χωρίς γνώσεις πληροφορικής, απαιτείται η δημιουργία μιας εφαρμογής που μεσολαβεί ανάμεσα στον χρήστη/στέλεχος διοίκησης και το κυρίως KBS. Η εφαρμογή αυτή παρέχει μια σειρά από φόρμες που καθοδηγούν το χρήστη στην εισαγωγή παραμέτρων και την εκτέλεση ολοκληρωμένων σεναρίων διερεύνησης, και φροντίζει για την κωδικοποίηση των παραμέτρων αυτών με τον κατάλληλο τρόπο και την αποστολή τους στο KBS. Στη συνέχεια, θα επιδειχθεί η λειτουργία μιας τέτοιας εφαρμογής στην εκτέλεση μιας ολοκληρωμένης διαδικασίας στήριξης απόφασης.

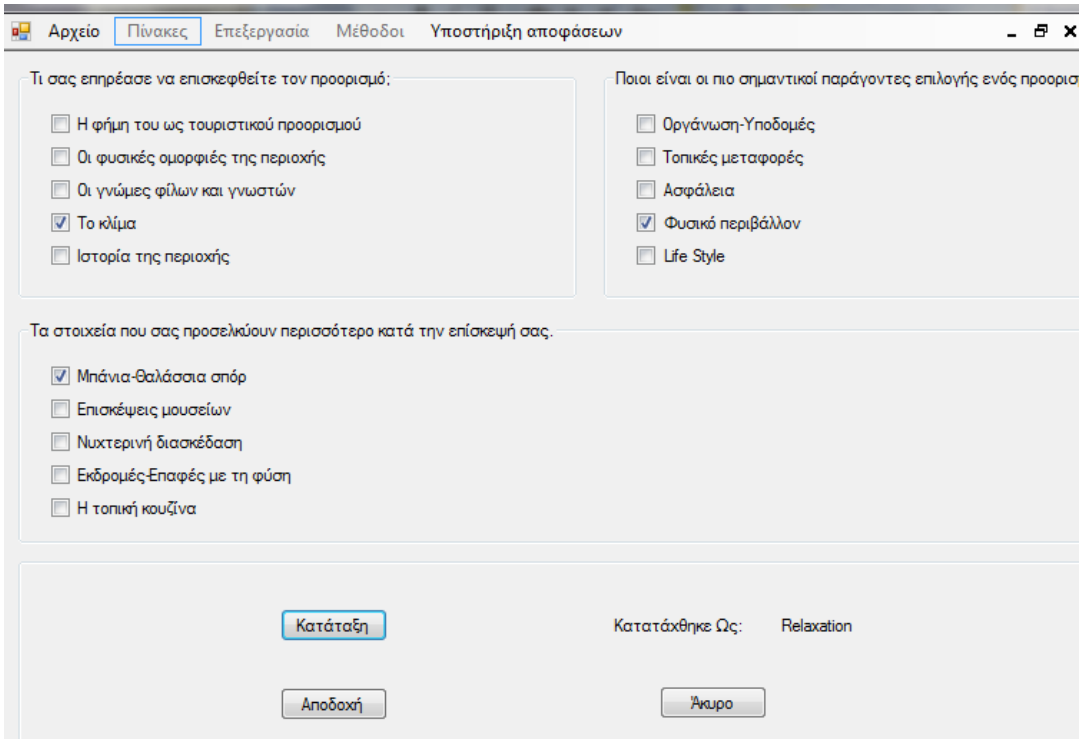
<u>Παράδειγμα διαδικασίας στήριξης απόφασης:</u> Επιλογή αγοράς στόχου – εκτίμηση μεγέθους και αξίας τμημάτων αγοράς. Η διαδικασία εκτελείται μέσω κατάλληλων φορμών με τις οποίες μπορούμε να καθορίσουμε χώρα, ηλικία, σκοπό ταξιδιού και επίσης το στοχευόμενο προφίλ του επισκέπτη με βάση τους παράγοντες επιλογής του. Το σύστημα απαντάει σχετικά με το μέγεθος και την αξία της αγοράς που αντιστοιχεί στις επιλογές αυτές. Οποιαδήποτε από τις παραμέτρους μπορεί να μείνει απροσδιόριστη, οπότε το σύστημα προτείνει για αυτές τυχόν καλύτερες επιλογές που γνωρίζει.



Σχήμα 8.10. *Η αρχική φόρμα της διαδικασίας Επιλογής Αγοράς-στόχου*

Ο χρήστης ξεκινάει από το μενού «Υποστήριξη Αποφάσεων» επιλέγοντας τη διαδικασία «Επιλογή αγοράς στόχου». Στην κύρια οθόνη που εμφανίζεται (Σχήμα 8.10), ο χρήστης μπορεί να επιλέξει τις παραμέτρους εισόδου. Στο παραπάνω παράδειγμα επιλέχθηκε σαν χώρα προέλευσης η Γερμανία, οι ηλικίες 19-35 και σαν σκοπός ταξιδιού οι διακοπές.

Για την επιλογή του προφίλ επισκέπτη με βάση τους παράγοντες επιλογής του δεν υπάρχει μενού αλλά κουμπί «Ευφυής κατάταξη», που ανοίγει νέα οθόνη (Σχήμα 8.11). Στην οθόνη επιλογής προφίλ επισκέπτη, μπορούμε να συμπληρώσουμε 1 ή 2 επιλογές στις 3 ερωτήσεις που εμφανίζονται σχετικά με τους παράγοντες επιλογής του επισκέπτη. Σημειώνεται ότι οι 3 ερωτήσεις και οι 15 συνολικά επιλογές προέρχονται από συγκεκριμένη ενότητα του ερωτηματολογίου της έρευνας για την εικόνα της Θεσσαλονίκης και από την ανάλυση των αντίστοιχων δεδομένων της έρευνας προέκυψαν 5 κλάσεις επισκεπτών: Lifestyle, Beaches and watersports, Nature and Excurions, Relaxation και Sights and museums.



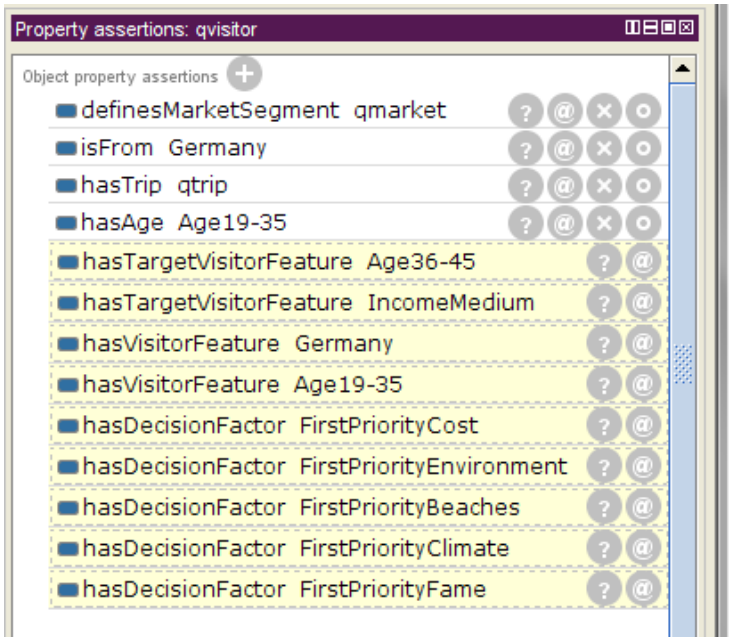
Σχήμα 8.11. *Η φόρμα ευφυούς κατάταξης του επισκέπτη σε τύπο με βάση τους παράγοντες επιλογής του.*

Το σύστημα είναι εκπαιδευμένο να κατατάσσει έναν επισκέπτη σε μια από τις 5 γνωστές κλάσεις με βάση τις επιλογές του, ακόμα και αν τα δεδομένα εισόδου είναι ελλιπή π.χ. κάποια ερώτηση είναι κενή. Στο παράδειγμα του Σχήματος 8.10, οι επιλογές που συμπληρώθηκαν οδήγησαν στην κατάταξη του στοχευόμενου επισκέπτη στο προφίλ "Relaxation". Αν ο χρήστης αποδεχθεί την προτεινόμενη κατάταξη, αυτή μεταφέρεται στην προηγούμενη οθόνη ως παράμετρος του ερωτήματος.

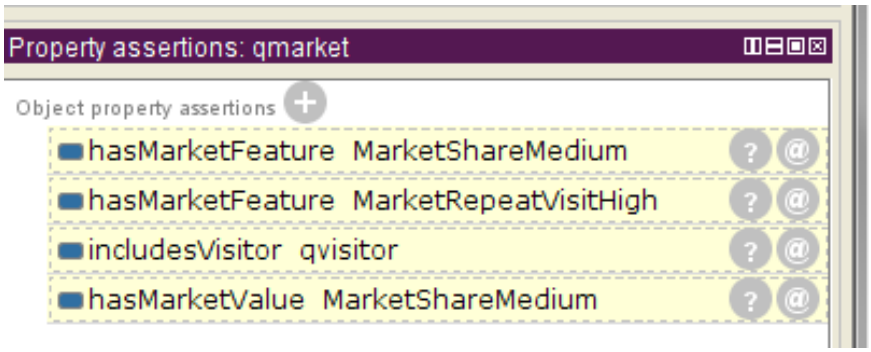
Πατώντας «Ερώτημα» , οι επιλογές κωδικοποιούνται σε μια έκφραση OWL που θα οδηγηθεί στη συνέχεια προς εισαγωγή στο Protégé. Το Protégé αντιλαμβάνεται τις νέες παραμέτρους και, αφού συγχρονιστεί η συμπερασματική μηχανή με τα νέα δεδομένα, υπολογίζει την συνεπαγόμενη γνώση. Για τη διευκόλυνση του χρήστη, έχει οριστεί και αποθηκευτεί στην οντολογία κατάλληλο προκαθορισμένο ερώτημα τύπου DL-Query, το οποίο ο χρήστης μπορεί να επιλέξει και να λάβει τις ιδιότητες που αφορούν την αξία του επιλεχθέντος τμήματος αγοράς και τα τυχόν χαρακτηριστικά που προτείνεται να στοχευθούν, ώστε να βελτιστοποιηθεί η αξία του τμήματος αγοράς.

Το αποτέλεσμα που λαμβάνουμε προκύπτει από τις ιδιότητες με τις οποίες συνδέονται ο επισκέπτης και το αντίστοιχο τμήμα αγοράς, όπως φαίνονται στο Σχήμα 8.12. Σύμφωνα με αυτές:

- Έχει οριστεί ότι ο επισκέπτης είναι από τη Γερμανία (isFrom Germany), έχει ηλικία 19-35 (hasAge19-35) και ανήκει στο προφίλ Relaxation (ανήκει στην κλάση VisitoForRelaxation).
- Έχουν εξαχθεί οι ιδιότητες ότι στοχευόμενο χαρακτηριστικό είναι η ηλικία 36-45 (hasTargetVisitoFeature Age36-45), στοχευόμενο χαρακτηριστικό είναι το μέσο εισόδημα (hasTargetVisitoFeature IncomeMedium), έχει ως παράγοντες επιλογής πρώτη προτεραιότητα το κόστος (hasDecisionFactor FirstPriorityCost), το περιβάλλον (hasDecisionFactor FirstPriorityEnvironment), τις παραλίες (hasDecisionFactor FirstPriorityBeaches), το κλίμα (hasDecisionFactor FirstPriorityClimate) και τη φήμη του προορισμού (hasDecisionFactor FirstPriorityFame).
- Το μερίδιο αγοράς έχει τις ιδιότητες ότι είναι μέτριο (hasMarketFeature MarketShareMedium) και ότι έχει υψηλό βαθμό επαναλαμβανόμενων επισκέψεων (hasMarketFeature MarketRepeatVisitHigh).



(α)



(β)

Σχήμα 8.12. *Οι ιδιότητες που προκύπτουν για (α) τον υπό διερεύνηση επισκέπτη και (β) το τμήμα αγοράς που του αντιστοιχεί.*

Στη συνέχεια του σεναρίου, ο χρήστης μπορεί να επιστρέψει στην αρχική οθόνη της εφαρμογής και να επαναδιαμορφώσει τις παραμέτρους του ερωτήματος. Εφόσον προτάθηκε η ηλικία 36-45, ας υποθέσουμε ότι ο χρήστης θέλει να εξετάσει καλύτερα αυτήν την εκδοχή. Στο μενού Ηλικία, μπορεί ο χρήστης να αλλάξει την επιλογή σε "36 to 45 years old" και να πατήσει Ερώτημα. Αυτή τη φορά (δηλ για ηλικία 36-45), το σύστημα υποδεικνύει ότι η στοχευόμενη επίσκεψη είναι διάρκειας 2 εβδομάδων (2WeekVisit), έχει υψηλό βαθμό επαναληψιμότητας (MarketRepeatVisitHigh), έχει μεγάλο μερίδιο αγοράς (MarketShareHigh) και χαρακτηρίζεται ως ταξίδι διακοπών. Παρατηρείται ότι όταν είχε επιλεγεί ηλικία 19-35, το σύστημα εκτίμησε μέτριο μερίδιο αγοράς και συνέστησε την ηλικία 36-45, ενώ όταν επιλέχθηκε η ηλικία 36-45, εκτίμησε μεγάλο μερίδιο αγοράς.

Βιβλιογραφία/Αναφορές

OWL (2013, May 30). *W3C Recommendation* . Retrieved from http://www.w3.org/TR/owl-features/

Protégé (n.d.). Retrieved 30 May 2014 from http://protege.stanford.edu/

Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R., Shadbolt, N., Van de Velde W. and Wielinga, B. (1999). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press.

Stalidis, G. & Karapistolis, D. (2014a), Knowledge discovery and computerized reasoning to assist tourist destination marketing. *International Journal on Strategic Innovative Marketing*, Vol.01, pp 103-119, DOI: 10.15556/IJSIM.01.02.004

Stalidis, G. & Karapistolis, D. (2014b). Tourist Destination Marketing Supported by Electronic Capitalization of Knowledge. *Procedia - Social and Behavioral Sciences*, pp. 110-118, doi: 10.1016/j.sbspro.2014.07.024

SWRL (2004, May 30). *A Semantic Web Rule Language Combining OWL and RuleML W3C Member Submission 21 May 2004*, National Research Council of Canada, Network Inference, and Stanford University. Retrieved from http://www.w3.org/Submission/SWRL/

Feigenbaum, E. A. & McCorduck, P. (1983). *The fifth generation* (1st ed.), Reading, MA: Addison-Wesley, ISBN 978-0-201-11519-2, OCLC 9324691

Shadbolt N. & Milton, N. (1999). *From knowledge engineering to knowledge management*. British Journal of Management, vol. 10, 1999, 309–322.

Schreiber, G. (2008). Knowledge Engineering. In: F. van Harmelen, V. Lifschitz, B. Porter (Eds.), *Handbook of Knowledge Representation* (pp. 929–946), Elsevier.

Ligêza, A., (2006). Logical foundations for rule-based systems. *Studies in Computational Intelligence*, vol. 11, 2nd Ed., Springer-Verlag Berlin Heidelberg.

Gruber,T.,(1993). A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, 5(2), pp. 199-220.

Prantner, K., Ding, Y., Luger, M., Yan, Z. (2007). *Tourism Ontology and Semantic Management System: State-of-the-Arts Analysis*. Proceedings of IADIS International Conference WWW/Internet 2007, pp.111-115.

Ou, S., Pekar, V., Orasan, C., Spurk, C., Negri M., (2008). *Development and Alignment of a Domain-Specific Ontology for Question Answering, in European Language Resources Association (ELRA)* (ed.): Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

DERI OnTour Ontology (n.d.). Retrieved 15 December 2011 from http://etourism.deri.at/ont/index.html

Κεφάλαιο 9. Συστήματα συστάσεων (Recommender systems)

Σύνοψη
Σκοπός του κεφαλαίου είναι να εισάγει τον αναγνώστη στα συστήματα συστάσεων (recommender systems) και στην εξατομίκευση των υπηρεσιών και των συστημάτων web (web personalization). Τα συστήματα συστάσεων είναι συστήματα που φιλτράρουν πληροφορίες και προσπαθούν να προβλέψουν την αξιολόγηση ή προτίμηση που θα έδινε ο χρήστης σε ένα αντικείμενο που δεν έχει αξιολογήσει ακόμα και να του προτείνουν λύσεις. Διάφορες επιχειρήσεις όπως Mystrands και Stumbleupon,το Yahoo και η Sun έχουν χρησιμοποιήσει recommender systems. Στον εμπορικό τομέα το Amazon συνειδητοποίησε νωρίς τις δυνατότητες των recommender systems και μέχρι σήμερα παραμένει ένα από τα πιο αντιπροσωπευτικά παραδείγματα των εταιριών που τα εφαρμόζουν. Στα πλαίσια ανάπτυξης ενός συστήματος συστάσεων παρουσιάζονται στο κεφάλαιο αυτό διάφορες μέθοδοι ανάλυσης δεδομένων, αρχιτεκτονικές συστημάτων καθώς και παραδείγματα συστημάτων που οι υπηρεσίες τους είναι διαθέσιμες στο διαδίκτυο.

Προαπαιτούμενη γνώση
Κεφάλαιο 1.Εισαγωγή στη βασισμένη σε δεδομένα επιχειρηματική ευφυΐα, Κεφάλαιο 2.Δεδομένα και Πληροφορίες, Κεφάλαιο 5.Μετατροπή των δεδομένων σε πληροφορία, Κεφάλαιο 6.Μέθοδοι εξόρυξης γνώσης από δεδομένα, Κεφάλαιο 8.Μοντελοποίηση Γνώσης και Βάσεις Γνώσης

9.1 Συστήματα Συστάσεων: Εισαγωγή

Στη σημερινή εποχή όπου η πληθώρα το πληροφοριών αλλά και των υπηρεσιών και προϊόντων ενισχύεται με την περαιτέρω εξάπλωση του διαδικτύου, ένα μεγάλο πρόβλημα ανακύπτει. Πώς ως χρήστες-πελάτες να επιλέξουμε αυτό που μας ενδιαφέρει, αυτό μας ταιριάζει; Οι πληροφορίες είναι χαοτικά πολλές για να μπορέσουμε να φιλτράρουμε αυτό που θέλουμε. Από την μεριά της επιχείρησης, το ίδιο πρόβλημα επαναδιατυπώνεται. Πώς ως επιχείρηση μπορούμε να γνωρίζουμε ποιος πελάτης θέλει τι και πώς θα γνωρίζουμε πόσο ευχαριστημένος είναι από τις επιλογές του. Αλλά και κάτι σημαντικό ακόμα. Πώς θα μπορούμε να προβλέπουμε τις ανάγκες ενός πελάτη εάν δεν γνωρίζουμε κάτι γι' αυτόν; Η τεχνολογία και οι εφαρμογές της μας οδηγούν στο χώρο της εξατομίκευσης της πληροφόρησης και των υπηρεσιών. Αναπτύσσεται η εξατομίκευση των υπηρεσιών (Kardaras & Karakostas, 2012), η εξατομίκευση του περιεχομένου και του τρόπου επικοινωνίας στον παγκόσμιο ιστό (Miao et al, 2007; Bunt et al, 2007; Kardaras et al, 2011; Kardaras et al, 2013). Εξέλιξη αυτής της τάσης είναι τα συστήματα συστάσεων.

Τα συστήματα συστάσεων ή προτάσεων (recommender systems ή recommendation systems), είναι πληροφοριακά συστήματα που βασίζονται σε αλγόριθμους και στόχο έχουν να προτείνουν τα πιο κατάλληλα προϊόντα ή υπηρεσίες σε άτομα, ομάδες ή και επιχειρήσεις. Η Amazon.com επιτυγχάνει πωλήσεις ύψους μεταξύ 30%-70% του συνόλου των πωλήσεων της, μέσω προτάσεων σε πελάτες της. Παρομοίως και η Netflix.com. Οι προτάσεις διαμορφώνονται μετά από ανάλυση των χαρακτηριστικών και την πρόβλεψη των προτεραιοτήτων και των αναγκών των ατόμων που πρόκειται να πληροφορηθούν την πρόταση του συστήματος (Ansari, et al, 2000); Adomavicius & Tuzhilin, 2005; Xiao, & Benbasat, 2007). Είναι γεγονός, και όλοι έχουμε σχετική πια εμπειρία, ότι οι ποσότητες των διαθέσιμων πληροφοριών στο διαδίκτυο είναι τεράστιες. Πληροφορίες που σχετίζονται με κάθε σχεδόν δραστηριότητα σύγχρονης κοινωνίας. Το πρόβλημα που παρουσιάζεται για ένα/μία χρήστη του διαδικτύου είναι πως από αυτήν την πληθώρα των πληροφοριών θα επιλέξει αυτήν ή αυτές τις πληροφορίες που πραγματικά θα είναι χρήσιμες. Ας σκεφτούμε το παρακάτω σενάριο. Στο διαδίκτυο υπάρχουν πολλοί ιστοχώροι που παρουσιάζουν ξενοδοχεία με παρόμοια αλλά πολλά διαφορετικά χαρακτηριστικά ανάλογα με τον τύπο του ξενοδοχείου, τη θέση του, τις υπηρεσίες που προσφέρει, κλπ. Χιλιάδες ξενοδοχεία, χιλιάδες χαρακτηριστικά. Εάν μια οικογένεια θέλει να πάει διακοπές, ποιό ξενοδοχείο να επιλέξει ή σε ποια να εστιάσει την προσοχή της; Σε ποιο τόπο διακοπών; Με ποιές προσφερόμενες υπηρεσίες; Με ποια κριτήρια να επιλέξει η οικογένεια και ποια κριτήρια εκφράζουν καλύτερα τις ανάγκες και τις επιθυμίες της; Η ανάπτυξη των συστημάτων συστάσεων σκοπεύουν να απαντήσουν σε τέτοιου είδους ερωτήσεις σε μεγάλο εύρος προϊόντων και υπηρεσιών. Με την ανάλυση των χαρακτηριστικών συμπεριφοράς αλλά και άλλων πληροφοριών σχετικά με τους χρήστες ένα σύστημα συστάσεων προσπαθεί να προβλέψει τις ανάγκες και τις επιθυμίες του χρήστη και να προτείνει την κατάλληλη πρόταση.

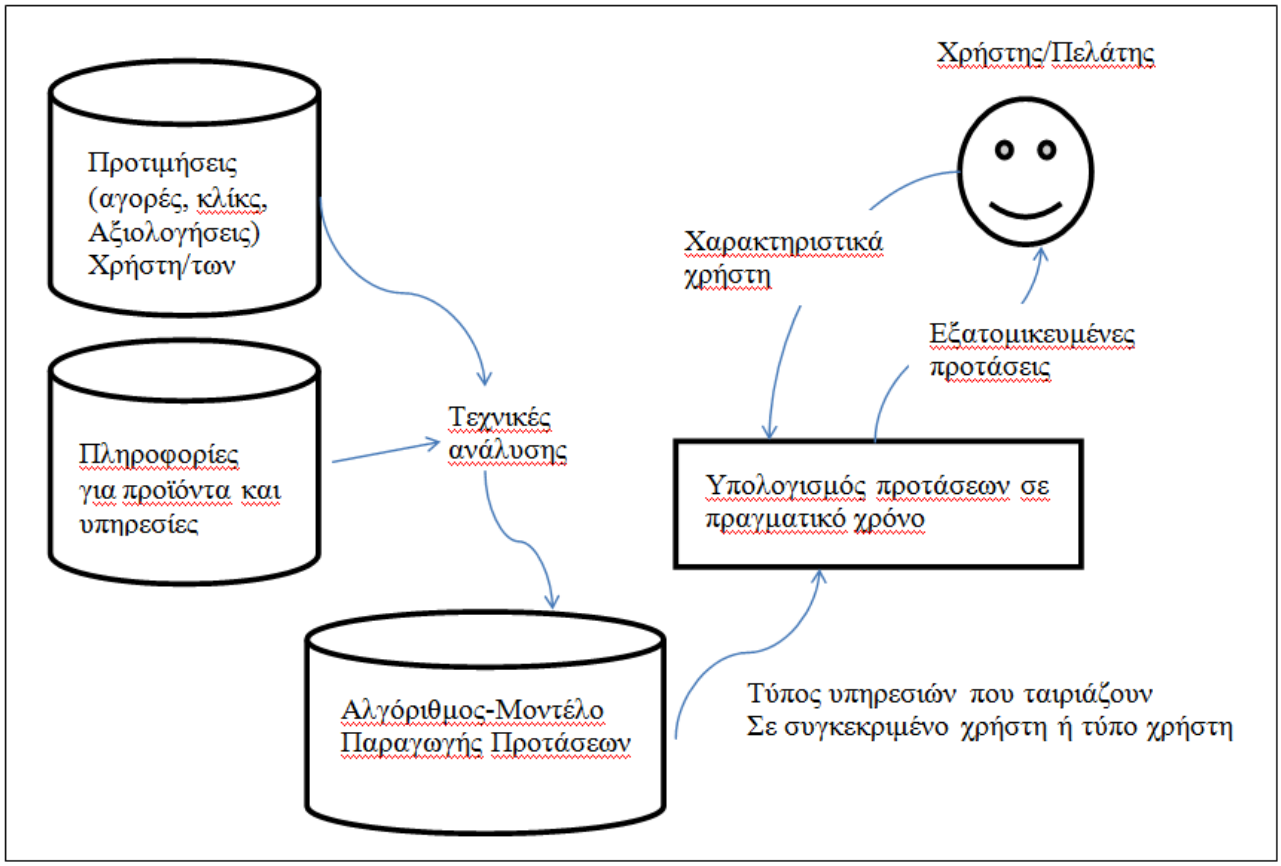
Οι πληροφορίες χρηστών που συνήθως αναλύονται αφορούν:

- Τις συνήθειες ενός χρήστη στο διαδίκτυο, δηλαδή τι κατά προτεραιότητα ψάχνει και το υπηρεσίες χρησιμοποιεί,
- Δημογραφικά στοιχεία,
- Το χρόνο στο οποίο ο χρήστης ψάχνει για κάτι, πόσο συχνά το ψάχνει, μετά από ποια σειρά γεγονότων τις ψάχνει, κλπ.
- Τις δραστηριότητές του όπως αυτές δημοσιοποιούνται στα κοινωνικά δίκτυα που αφορούν το χρήστη. Οι πληροφορίες αυτές είναι προσβάσιμες είτε απ' ευθείας από τη σελίδα του χρήστη είτε εμμέσως από τις σελίδες άλλων χρηστών συνδεδεμένων με τον χρήστη για τον οποίο διαμορφώνεται η πρόταση.

Επίσης συλλέγονται και αναλύονται πληροφορίες σχετικά με τα προϊόντα ή τις υπηρεσίες για τις οποίες αναπτύσσεται ένα σύστημα συστάσεων. Για παράδειγμα, εάν ένα σύστημα σχεδιάζεται για να προτείνει σε ένα χρήστη ποιά ταινία να παρακολουθήσει, τότε το σύστημα αυτό θα πρέπει να έχει στη βάση δεδομένων του αξιολογήσεις ταινιών. Οι αξιολογήσεις αυτές μπορεί ενδεικτικά και ανάλογα με τη σχεδίαση και τη υλοποίηση του συστήματος, να βασίζονται σε πληροφορίες όπως:

- Το συγγραφέα του σεναρίου της ταινίας,
- Το σκηνοθέτη,
- Τον/τους πρωταγωνιστές,
- Του ηθοποιούς,
- Το είδος της ταινίας,
- Το έτος ή δεκαετία παραγωγής,
- Κάποιες λέξεις κλειδιά, κλπ.

Διαγραμματικά η αρχιτεκτονική ενός συστήματος συστάσεων φαίνεται στο επόμενο Σχήμα 9.1.



Σχήμα 9.1.*Αρχιτεκτονική Συστήματος Συστάσεων*

Για την ανάπτυξη ενός συστήματος συστάσεων συλλέγονται δεδομένα για τις προτιμήσεις των χρηστών. Τα δεδομένα συλλέγονται από το διαδίκτυο με βάση την πλοήγηση κάθε χρήστη, τις αγορές του, συλλέγονται επίσης και από τα κοινωνικά δίκτυα αλλά και από τα POS (point of sale), στα σημεία πωλήσεων από όπου ένας χρήστης-πελάτης αγοράζει αγαθά και υπηρεσίες. Τα δεδομένα αυτά αποθηκεύονται σε βάση δεδομένων. Ομοίως συλλέγονται δεδομένα σχετικά με τις διαθέσιμες υπηρεσίες και προϊόντα καθώς και τα ιδιαίτερα χαρακτηριστικά τους. Οι πληροφορίες αυτές αναλύονται έτσι ώστε για κάθε χρήστη (ή τύπο χρήστη) να είναι γνωστό τι υπηρεσίες προτιμά και σε ποιο βαθμό. Εάν υποθέσουμε ότι είναι διαθέσιμα τα δεδομένα για *m* χρήστες ($u_i, i = 1,...,m$) και για *n* χαρακτηριστικά υπηρεσιών (s_j , $j = 1,...,n$), τότε για κάθε χρήστη ή ομάδα χρηστών υπολογίζεται ο βαθμός προτίμησης κάθε χαρακτηριστικού υπηρεσίας ($r_{i,j}$) όπου $i = 1,...m$ και $j = 1,...,n$. Η ανάλυση των δεδομένων δημιουργεί ένα μοντέλο το οποίο χρησιμοποιείται ώστε να μπορεί να προσδιορίσει προτάσεις σε χρήστες που σε πραγματικό χρόνο ζητούν εξατομικευμένες λύσεις, έστω και εάν οι προτιμήσεις των συγκεκριμένων αυτών χρηστών δεν είναι γνωστές. Δηλαδή το σύστημα συστάσεων θα προβλέψει ποιές θα πρέπει να είναι προτιμήσεις του χρήστη. Ο βαθμός επιτυχίας των προβλέψεων καθορίζει τελικά και το βαθμό αξιοπιστίας του συστήματος.

9.2 Τεχνικές και Στρατηγικές Ανάλυσης

Οι τεχνικές που χρησιμοποιούνται για την διαμόρφωση των συστάσεων αποτελούνται από τις πιο παραδοσιακές τεχνικές και από τις πιο πρόσφατες. Οι τεχνικές και οι αλγόριθμοι που χρησιμοποιούνται για την ανάλυση των προτιμήσεων των πελατών, ή την ανάλυση των χαρακτηριστικών των προϊόντων αλλά και για διαμόρφωση των συστάσεων προέρχονται από την στατιστική, την τεχνητή νοημοσύνη, όπως οι αλγόριθμοι συσταδοποίησης (clustering), η ασαφής λογική (fuzzy logic, fuzzy clustering, κλπ.), οι αλυσίδες Μαρκόφ (Markov chains), η παλινδρόμηση (regression), κλπ. Η ανάλυση των σχετικών πληροφοριών για τους χρήστες και τις προτιμήσεις τους καθώς και πληροφοριών σχετικών με τις διαθέσιμες υπηρεσίες και προϊόντα βασίζεται σε τεχνικές που χωρίζονται στις παρακάτω παραδοσιακές κατηγορίες τεχνικών και συστημάτων (Ricci, et al.,2011; Lu., et al, 2015).

- Τεχνικές με βάση το περιεχόμενο (content-based), προσδιορίζουν τις προτάσεις τους σύμφωνα με τις προτιμήσεις που έχει κάνει ο χρήστης στο παρελθόν.
- Τεχνικές συνεργατικού φιλτραρίσματος (collaborative filtering), οι οποίες προτείνουν σε ένα χρήστη προϊόντα και υπηρεσίες που άλλοι χρήστες με παρόμοιες με αυτόν προτιμήσεις έχουν επιλέξει στο παρελθόν.
- Τεχνικές με βάση τη γνώση (knowledge-based), οι οποίες χρησιμοποιούν κανόνες της μορφής "if <> then <> else", οι οποίοι κανόνες εκφράζουν στο πως μια υπηρεσία ή ένα χαρακτηριστικό της καλύπτει μια ανάγκη του χρήστη.
- Τεχνικές Υβριδικές (Hybrid), αποτελούν συνδυασμό των παραπάνω.

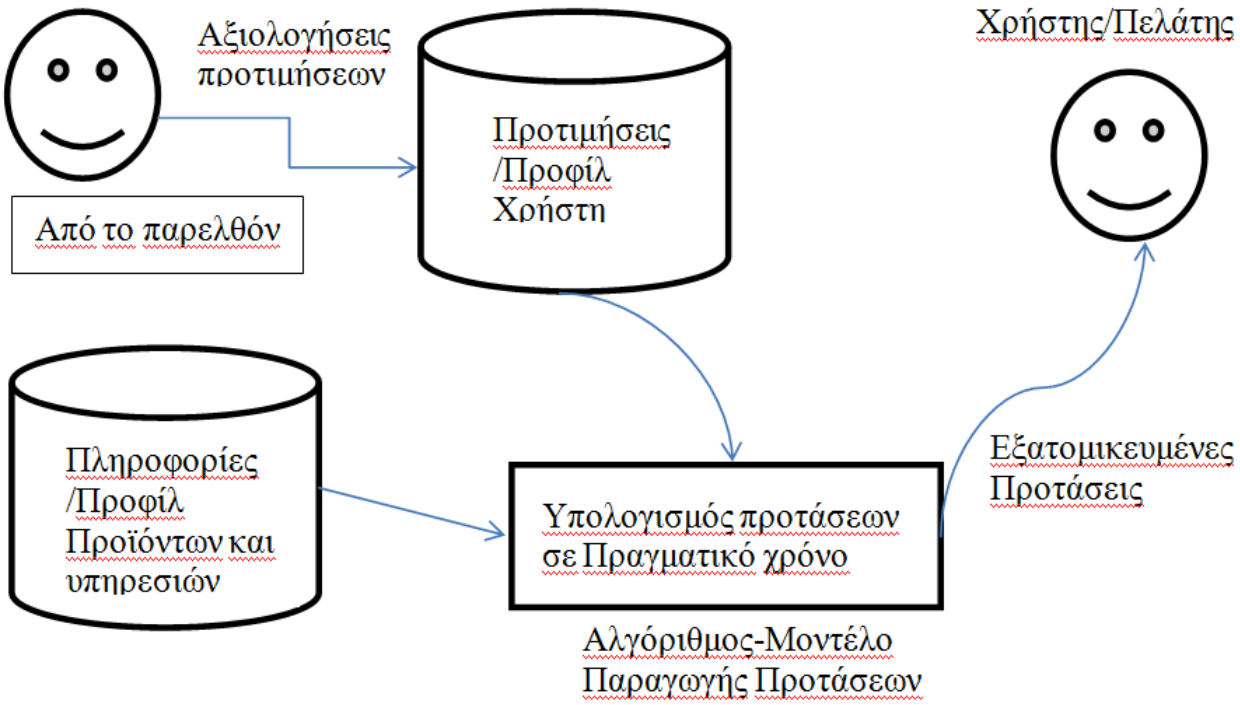
Η κάθε κατηγορία τεχνικών βεβαίως έχει τα ιδιαίτερα πλεονεκτήματα αλλά και μειονεκτήματά της. Η ανάπτυξη νέων τεχνικών αποτελεί ένα από τους πιο έντονα δραστήριους τομείς έρευνας στα συστήματα συστάσεων. Νέες μέθοδοι ανάλυσης των διαθέσιμών δεδομένων, νέες μέθοδοι αξιολόγησης των συστημάτων, κλπ. Πιο πρόσφατα έχουν αναπτυχθεί και αναπτύσσονται τεχνικές που βασίζονται στην ασαφή λογική (fuzzy logic), στα κοινωνικά δίκτυα (social networks based), στην μοντελοποίηση της εμπιστοσύνης (trust based), στην ανάλυση του περιβάλλοντος και των συνθηκών διαμόρφωσης μιας πρότασης (context-aware based), κλπ.

9.2.1 Τεχνικές με βάση το περιεχόμενο (content-based)

Οι τεχνικές αυτές εστιάζουν στην σύγκριση αντικειμένων για την εύρεση κοινών χαρακτηριστικών. Στην περίπτωση εγγράφων αναζητούνται ομοιότητες στα θέματα που πραγματεύονται, ενώ στην περίπτωση αντικειμένων ή άλλου είδους προϊόντων και υπηρεσιών τα κοινά χαρακτηριστικά αναζητούνται μέσα από τις περιγραφές και τις προδιαγραφές τους.

Οι τεχνικές με βάση το περιεχόμενο προτείνουν προϊόντα και υπηρεσίες με βάση το προφίλ του χρήστη, το οποίο είναι μοναδικό για κάθε έναν. Το προφίλ ενός χρήστη διαμορφώνεται με βάση τα

ενδιαφέροντα, τις προτιμήσεις και τις επιλογές που έχει κάνει στο παρελθόν ο συγκεκριμένος χρήστης. Για να είναι δυνατή η διαμόρφωση των προτάσεων πρέπει να είναι γνωστές οι προτιμήσεις των χρηστών αλλά και η αξιολόγηση των εναλλακτικών υπηρεσιών/προϊόντων με βάση τα κριτήρια και τις προτεραιότητες των χρηστών. Το ενδιαφέρον ενός χρήστη μπορεί να είναι η γνώμη του χρήστη για ένα προϊόν μέσω αξιολόγησης, οι προηγούμενες αγορές που έχει κάνει ή η περιήγησή του σε ιστοσελίδες σχετικές με αυτό. Έτσι το σύστημα συστάσεων θα μπορεί να προτείνει προϊόντα, υπηρεσίες με χαρακτηριστικά πιο κοντά σε αυτά που ζητάει ο χρήστης και με παρόμοια χαρακτηριστικά με αυτά που εμπεριέχονται στο προφίλ του. Τα συστήματα συστάσεων που χρησιμοποιούν τεχνικές με βάση το περιεχόμενο έχουν τη δομή και τις αρχές λειτουργίας που απεικονίζονται στο Σχήμα 9.2.



Σχήμα 9.2.*Δομή και Αρχές λειτουργίας ενός συστήματος συστάσεων με βάση το περιεχόμενο*

Οι επιλογές ενός χρήστη και οι αξιολογήσεις (ratings) αυτών που έχει κάνει στο παρελθόν καταγράφονται και αποτελούν μέρος του προφίλ του χρήστη. Στην περίπτωση που τα χαρακτηριστικά των αντικειμένων (προϊόντων, υπηρεσιών, ενδιαφερόντων, κ.ά.) είναι αποθηκευμένα σε δομημένες μορφές κειμένου (π.χ. Βάσεις δεδομένων, xml documents) η δημιουργία των προφίλ και η σύγκριση των αντικειμένων είναι απλή διαδικασία. Στις περισσότερες όμως περιπτώσεις, τα χαρακτηριστικά αυτά πρέπει να εξαχθούν από μη δομημένο κείμενο, όπως περιγραφές, το συνοδευτικό κείμενο σε μια ιστοσελίδα ή γνώμες άλλων χρηστών σε φυσική γλώσσα. Στην περίπτωση αυτή τα συστήματα χρησιμοποιούν τεχνικές μοντελοποίησης κειμένου για τη μετατροπή τους σε μια δομημένη αναπαράσταση (βάση δεδομένων).

Αντλώντας στοιχεία για τις προτιμήσεις που έχει κάνει ένας χρήστης στο παρελθόν, δημιουργείται το προφίλ του συγκεκριμένου χρήστη και αποθηκεύεται σε βάση δεδομένων. Τα προφίλ χρηστών δηλαδή, οργανώνουν πληροφορίες για τα ενδιαφέροντα των χρηστών, σχετικά με προϊόντα που έχουν τραβήξει την προσοχή τους. Μια μαθηματική συνάρτηση (η μορφή της, οι μεταβλητές, κλπ. διαφοροποιούνται από σύστημα σε σύστημα) προσδιορίζει το βαθμό στον οποίο ένας χρήστης ενδιαφέρεται για μια υπηρεσία ή ένα χαρακτηριστικό της. Δεδομένα συλλέγονται δυνητικά για όλους τους χρήστες. Τα χαρακτηριστικά των προϊόντων, που ουσιαστικά αποτελούν το σύνολο των δυνατών προτάσεων προς τους χρήστες επίσης αποθηκεύονται σε βάση δεδομένων. Οι προτάσεις δημιουργούνται προσδιορίζοντας «σε πιο βαθμό παρομοιάζει μια υπηρεσία ή χαρακτηριστικό της με προηγούμενη επιλογή του χρήστη». Εάν για παράδειγμα,

ένας πελάτης είχε παρακολουθήσει ταινία κωμωδίας και είχε δώσει καλή αξιολόγηση, τότε το σύστημα συστάσεων θα προτείνει πάλι ταινία κωμωδίας. Βεβαίως η ανάλυση των προτιμήσεων αλλά και η ανάλυση των εναλλακτικών συστάσεων δεν γίνεται με βάση ενός μόνο κριτηρίου αλλά με βάση την όσο το δυνατόν πληρέστερη αντίληψη του προφίλ του πελάτη αλλά και των χαρακτηριστικών των εναλλακτικών προτάσεων. Δηλαδή, ο υπολογισμός των προτάσεων καθορίζεται από το βαθμό ομοιότητας (degree of similarity) ενός ή περισσοτέρων χαρακτηριστικών μιας υπηρεσίας ή προϊόντος με τις προτιμήσεις του χρήστη, όπως αυτές έχουν διαμορφωθεί από προηγούμενες επιλογές του. Έχουν διατυπωθεί αρκετοί τρόποι υπολογισμού της ομοιότητας μεταξύ εναλλακτικών προτάσεων. Ένας δημοφιλής είναι ο παρακάτω τύπος:

$$sim(s_i, s_j) = \frac{\sum_k w_{ki} w_{kj}}{\sqrt{\sum_k w_{ki}^2} * \sqrt{\sum_k w_{kj}^2}},$$

όπου $sim(s_i, s_j)$ δηλώνει το βαθμό ομοιότητας μεταξύ της υπηρεσίας s_i και s_j αντίστοιχα, ενώ τα w_{ki} και w_{kj} είναι οι βαθμοί προτίμησης για το χαρακτηριστικό (k) της υπηρεσίας/προϊόντος (i) και (j) αντίστοιχα.

Τα συστήματα παραγωγής συστάσεων που βασίζονται στο περιεχόμενο, παρουσιάζουν κάποια σημαντικά μειονεκτήματα. Εξαρτώνται αποκλειστικά από τις παρεχόμενες πληροφορίες για κάθε αντικείμενο, με αποτέλεσμα όταν οι πληροφορίες είναι ανεπαρκείς οι παραγόμενες προτάσεις να μην είναι ικανοποιητικές. Η χρήση τους μπορεί να είναι επιτυχής σε περιπτώσεις που τα δεδομένα είναι δομημένα, ωστόσο, η χρήση αποκλειστικά ενός συστήματος βασισμένου στο περιεχόμενο δεν είναι επαρκής στις περισσότερες περιπτώσεις. Για το λόγο αυτό είναι επιθυμητό να ακολουθείται ένας συνδυασμός προσεγγίσεων.

9.2.1.1 Μοντελοποίηση χρηστών

Στα συστήματα συστάσεων βασισμένα στο περιεχόμενο, η διαδικασία δημιουργίας των προφίλ των χρηστών με βάση τις προτιμήσεις τους και η δημιουργία προτάσεων με βάση αυτές τις προτιμήσεις είναι μια μορφή ταξινόμησης της γνώσης που το σύστημα μαθαίνει σχετικά με το προφίλ των χρηστών (classification learning). Το σύστημα εκπαιδεύεται ώστε να αναγνωρίζει τα χαρακτηριστικά ενός αντικειμένου (προϊόντος, υπηρεσίας, ενδιαφέροντος, κ.ά.) και να το κατηγοριοποιήσει με βάση τα υπάρχοντα προφίλ (μοντέλα) των χρηστών. Η μοντελοποίηση των χρηστών περιλαμβάνει:

- τη διαδικασία συλλογής των πληροφοριών - δεδομένων που σχετίζονται με τον κάθε χρήστη ειδικά,
- τη μετατροπή των δεδομένων που συλλέγονται σε κατάλληλες μορφές, τέτοιες ώστε να είναι δυνατή η περαιτέρω επεξεργασία τους,
- την οργάνωσή τους σε κατάλληλες δομές, που θα επιτρέψουν τη χρήση τους για την παραγωγή των προτάσεων και, τέλος,
- τη συντήρηση και ανανέωσή τους, αντικατοπτρίζοντας τα τρέχοντα ενδιαφέροντα των χρηστών (μη στατικά αλλά μεταβλητά προφίλ χρηστών).

Οι βασικές μέθοδοι μοντελοποίησης του προφίλ των χρηστών χρησιμοποιούν διαφορετικές προσεγγίσεις όπως οι παρακάτω:

- σταθμισμένες λέξεις κλειδιά (weighted keywords)
- σημασιολογικά δίκτυα (semantic networks)
- σταθμισμένες έννοιες (weighted concepts)

9.2.1.2 Σταθμισμένες λέξεις κλειδιά (weighted keywords)

Τα προφίλ με λέξεις-κλειδιά (όρος που προέρχεται από την τεχνολογία ανάκτησης πληροφοριών) αποτελούνται από διανύσματα λέξεων όπου κάθε μια λέξη συνοδεύεται από μία τιμή, το βάρος της. Οι λέξεις

υποδηλώνουν τα ενδιαφέροντα του χρήστη και το βάρος κάθε μίας το βαθμό ενδιαφέροντός του χρήστη για τη λέξη αυτή, όπως φαίνονται στον Πίνακα 1.

Τέχνη		
	Πορτραίτα	0.60
	Γλυπτική	0.72
Αθλητισμός		
	Καλαθοσφαίριση	0.88
	Τένις	0.27

Πίνακας 9.1. *Λέξεις κλειδιά και το βάρος που υποδηλώνει τη σημασία των λέξεων για το χρήστη*

Οι λέξεις μπορούν να εξάγονται είτε έμμεσα, αυτόματα από τις ιστοσελίδες - "έγγραφα" που επισκέπτεται ο χρήστης (δεδομένα σε δομημένη μορφή-κείμενο που με κατάλληλες τεχνικές μοντελοποίησης εγγράφων εξάγονται οι λέξεις κλειδιά και το βάρος αυτών), είτε άμεσα μέσω ανατροφοδότησης από τον ίδιο τον χρήστη.

Η σημαντικότητα (βάρος) w_{kj} του όρου (λέξη-κλειδί) t_k στο έγγραφο d_j είναι μια συνάρτηση της συχνότητας του t_k στο d_j, του αριθμού των εγγράφων που περιλαμβάνουν τον όρο t_k και του συνολικού αριθμού εγγράφων μιας ομάδας εγγράφων $D = \{d_1, d_2, ..., d_N\}$. Η πιο συνηθισμένη μέθοδος υπολογισμού των βαρών είναι η «Συχνότητα Όρου-Αντίστροφη Συχνότητα Εγγράφου» (Term Frequency – Inverse Document Frequency ή TF-IDF). Σύμφωνα με τη μέθοδο αυτή, το βάρος ενός όρου ισούται με τη συχνότητα με την οποία εμφανίζεται ο όρος στο κείμενο-έγγραφο επί την αντίστροφη συχνότητα εμφάνισης του όρου σε όλα τα κείμενα. Λέξεις που εμφανίζονται συχνά σε ένα κείμενο, αλλά είναι σπάνιοι στα υπόλοιπα κείμενα, έχουν μεγαλύτερο βάρος σε σχέση με λέξεις που εμφανίζονται με μεγάλη συχνότητα σε όλα τα έγγραφα (π.χ. λέξεις όπως το "και", τα άρθρα, κλπ.). Οι όροι με μικρό βάρος δεν επηρεάζουν το προφίλ του χρήστη.

Η συνάρτηση TF-IDF δίνεται παρακάτω:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) * \log \frac{N}{n_k}$$

N: ο συνολικός αριθμός των εγγράφων της ομάδας D

n_k: ο αριθμός των εγγραφών στα οποία εμφανίζεται τουλάχιστον μια φορά ο όρος t_k.

Ο υπολογισμός των βαρών γίνεται ως εξής:

$$w_{k,j} = \frac{TF - IDF(t_k, d_j)}{\sqrt{\sum_s TF - IDF(t_s, d_j)^2}}$$

και στη συνέχεια υπολογίζονται οι βαθμοί ομοιότητας, των χαρακτηριστικών των υπηρεσιών και των προτεραιοτήτων των χρηστών.

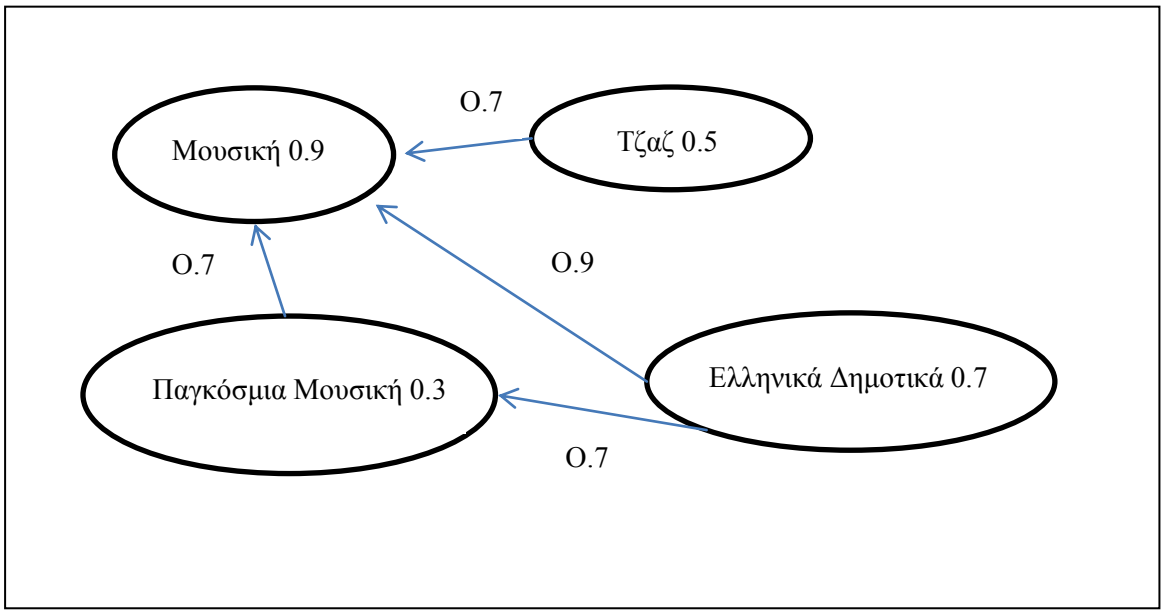
Στην περίπτωση που ο χρήστης έχει περισσότερα του ενός ενδιαφέροντα, η μέθοδος αυτή δεν είναι αποτελεσματική καθώς το προφίλ του χρήστη θα οδηγήσει σε προτάσεις που βρίσκονται στο μέσο όρο των διαφορετικών ενδιαφερόντων του. Η λύση στο πρόβλημα αυτό δόθηκε με τη δημιουργία πολλαπλών διανυσμάτων λέξεων κλειδιών, ένα για κάθε πεδίο ενδιαφέροντος του χρήστη. Στο σύστημα WebMate (Chen & Sycara, 1998), κάθε φορά που ένας χρήστης δηλώνει άμεσα ότι μια ιστοσελίδα "του αρέσει", εκτελείται ο αλγόριθμος με τις Σταθμισμένες λέξεις κλειδιά. Το σύστημα πραγματοποιεί τα παρακάτω:

- αφαιρεί από τις ιστοσελίδες τις κοινές λέξεις

- κάνει αποκοπή των καταλήξεων για την αποφυγή ύπαρξης πολλαπλών εγγραφών για την ίδια λέξη, λόγω της χρήσης της σε πληθυντικό αριθμό ή σε άλλο πρόσωπο.
- εξάγει τις λέξεις-κλειδιά, και
- πραγματοποιεί τον υπολογισμό των βαρών με την χρήση του αλγόριθμου TF-IDF
- αποδίδει στις λέξεις που περιέχονται σε τίτλους και επικεφαλίδες πρόσθετο βάρος.
- δημιουργεί Ν διανύσματα για τις πρώτες Ν λέξεις (όπου Ν ο προκαθορισμένος αριθμός διανυσμάτων που θα αποθηκευτούν), εάν προκύψουν περισσότερα από Ν διανύσματα, τότε γίνεται σύγκρισή τους με βάση την ομοιότητα συνημιτόνου (cosine similarity) και τα διανύσματα που παρουσιάζουν την μεγαλύτερη ομοιότητα συνενώνονται.
- ταξινομεί τα διανύσματα σε φθίνουσα σειρά με βάση τα βάρη των λέξεων.

9.2.1.3 Σημασιολογικά δίκτυα (semantic networks)

Η βασική δομή του δικτύου αποτελείται από κόμβους και συνδέσεις μεταξύ αυτών. Οι κόμβοι αναπαριστούν τις λέξεις-κλειδιά και οι συνδέσεις, την εννοιολογική συσχέτιση των λέξεων. Οι λέξεις-κλειδιά συνοδεύονται από τη σημαντικότητά τους. Το βάρος κάθε σύνδεσης αντιπροσωπεύει τη συχνότητα της εμφάνισης των συνδεδεμένων λέξεων (Σχήμα 9.3). Η συλλογή των δεδομένων γίνεται τις περισσότερες φορές με την άμεση συμμετοχή του χρήστη.



Σχήμα 9.3. *Σημασιολογικό Δίκτυο*

Η μοντελοποίηση της συσχέτισης δύο λέξεων μέσω των κοινών τους εμφανίσεων καθιστούν τη μέθοδο αυτή πλεονεκτικότερη σε σχέση με αυτή των σταθμισμένων λέξεων-κλειδιά. Παραλλαγές των σημασιολογικών δικτύων περιλαμβάνουν την αντικατάσταση των λέξεων είτε με μια ομάδα συνωνύμων είτε με μια έννοια που εμπεριέχει αρκετές διαφορετικές λέξεις (επιλύοντας με αυτό τον τρόπο το θέμα της ύπαρξης ίδιων λέξεων με διαφορετική εννοιολογική σημασία). Η απεικόνιση των σχέσεων λέξεων και εννοιών μπορεί να προέρχεται είτε από κάποιο προϋπάρχον λεξικό, π.χ. Wordnet (Το WordNet περιέχει 100.000 λέξεις ταξινομημένες σε 80.000 σετ συνωνύμων) ή να δημιουργηθεί εξαρχής. Διαφορετική προσέγγιση αποτελεί η χρήση στερεότυπων και πλαισίων (frames), όπου το προφίλ του χρήστη αναπαρίσταται με ένα πλαίσιο που εμπεριέχει τα προσωπικά του δεδομένα και μια ομάδα ενδιαφερόντων, τα στερεότυπα, τα οποία μοντελοποιούνται σαν σημασιολογικά δίκτυα. Κάθε προφίλ περιέχει πολλαπλά σημασιολογικά δίκτυα, ένα για κάθε διαφορετικό πεδίο ενδιαφέροντος του χρήστη.

Το σύστημα "ifweb" (Aniscar & Tasso, 1997) βασίζεται στη δημιουργία προφίλ χρήστη με σημασιολογικά δίκτυα. Στην περίπτωση αυτή, ο χρήστης αξιολογεί ένα έγγραφο θετικά, αρνητικά ή ουδέτερα. Το σύστημα

- εξάγει τις λέξεις-κλειδιά από το έγγραφο,
- εξετάζει αν υπάρχουν ή όχι στο δίκτυο,
- εάν υπάρχουν αναπροσαρμόζει τα βάρη των λέξεων - κλειδιών ανάλογα με την αξιολόγηση του
- χρήστη, εάν δεν υπάρχουν τις προσθέτει στο δίκτυο.

Το σύστημα InfoWeb (Gentili et al., 2003) χρησιμοποιεί στερεότυπα και περιλαμβάνει:

- επιλογή προκαθορισμένου αντιπροσωπευτικού αριθμού πρότυπων εγγράφων, για διάφορες θεματικές ενότητες,
- εφαρμογή του αλγορίθμου centorid-based k-means clustering στα πρότυπα έγγραφα
- ομαδοποίηση των εγγράφων,
- εντοπισμός των εγγράφων που βρίσκονται κοντά στο κέντρο κάθε κατηγορίας (centroid)
- προσδιορισμός των στερεοτύπων,
- αξιολόγηση των στερεοτύπων (αντιπροσωπευτικών έγγραφων κάθε κατηγορίας) από τον χρήστη ως θετικά, αρνητικά ή ουδέτερα,
- δημιουργία σημασιολογικού δικτύου - αναπαράσταση μέσω μοντέλου vector-space. Οι λέξεις κλειδιά που προκύπτουν από την αξιολόγηση των στερεότυπων εισάγονται στο δίκτυο ως ασύνδετα σημεία μεταξύ τους. Στη συνέχεια πραγματοποιούνται οι συνδέσεις μεταξύ τους καθιστώντας τα σημεία κόμβους. Μέσω της ανατροφοδότησης από τον χρήστη επεκτείνεται το δίκτυο και τροποποιούνται τα βάρη και οι συνδέσεις μεταξύ των κόμβων.

Το σύστημα WIFS (Micarelli & Sciarronne, 2004) αποτελεί προέκταση του προηγούμενου συστήματος. Κάθε θεματικό πεδίο διαθέτει το δικό του σημασιολογικό δίκτυο, το οποίο δημιουργείται με βάση στο σύνολο των εγγράφων που διατίθενται στο δίκτυο δημιουργώντας στερεότυπα μέσω της ομαδοποίησής τους. Το σύστημα περιλαμβάνει:

- την αρχική δημιουργία σημασιολογικών δικτύων για διάφορα θεματικά πεδία από ειδικούς
- την απάντηση από το κάθε χρήστη ενός ερωτηματολογίου προκειμένου να προσδιοριστούν τα στερεότυπα που θα ενεργοποιηθούν
- τη δημιουργία μοντέλου πλαισίου (frame) για κάθε χρήστη, το οποίο περιέχει την κεφαλή (header) με τα στοιχεία του χρήστη και τα ενεργοποιημένα στερεότυπα για αυτόν και το σώμα (body) με τα ενδιαφέροντά του. Το σώμα αποτελείται από θέσεις (slots). Κάθε θέση (slot) περιέχει (α) το πεδίο ενδιαφέροντος (domain), (β) τα επιμέρους θεματικά πεδία (topics), (γ) τον βαθμό ενδιαφέροντος του χρήστη για κάθε πεδίο (παίρνοντας άλλοτε θετικές και άλλοτε αρνητικές τιμές), (δ) ένα σημασιολογικό δίκτυο με σχετικούς όρους που συνυπάρχουν με τον όρο του θεματικού πεδίου και τέλος (ε) συνδέσμους τεκμηρίωσης (justification links), που δικαιολογούν την προσθήκη του συγκεκριμένου θεματικού πεδίου(αναφορά στο αρχικό ερωτηματολόγιο, ανατροφοδότηση από τον χρήστη ή άμεση επεξεργασία του προφίλ του χρήστη)
- τη συντήρηση του μοντέλου πλαισίου για κάθε χρήστη μέσω εξαγωγής λέξεων κλειδιών από μια ιστοσελίδα, (α) εάν αυτές υπάρχουν ήδη στο προφίλ του χρήστη, ανανεώνεται το βάρος του πεδίου ενδιαφέροντος καθώς και τα βάρη στο αντίστοιχο σημασιολογικό δίκτυο, (β) εάν αυτές δεν υπάρχουν ήδη στο προφίλ του χρήστη προστίθενται με τα κατάλληλα βάρη είτε σε επίπεδο slots είτε σε επίπεδο κόμβων.

Το προφίλ του χρήστη μπορεί να αλλάξει είτε με απευθείας επέμβαση του χρήστη είτε με αυτόματο τρόπο όταν αλλάξουν τα ενδιαφέροντα του. Στην περίπτωση αλλαγής των ενεργών στερεότυπων του χρήστη τότε τα slots που συνδέονται με συνδέσμους τεκμηρίωσης με τα παλαιά στερεότυπα θα σβηστούν.

9.2.1.4 Σταθμισμένες έννοιες (weighted concepts)

Η δομή σε αυτή τη μέθοδο είναι παραπλήσια με τα σημασιολογικά δίκτυα. Η διαφορά έγκειται στο ότι στη θέση των λέξεων ή των συνωνύμων χρησιμοποιούνται γενικότερες προϋπάρχουσες έννοιες, οι οποίες προέρχονται από ιεράρχηση, ταξινομίες (reference taxonomy), λεξικά ή και από οντολογίες. Στην περίπτωση χρήσης έτοιμων λεξικών (WordNet) ή οντολογιών (των οποίων το μέγεθος είναι μεγάλο) η επεξεργασία καθίσταται δύσκολη και χρονοβόρα λόγω της υπερ-εξειδίκευσης. Προτιμούνται λεξικά μικρότερου μεγέθους που περιέχουν μόνο τις απαραίτητες έννοιες και προκύπτουν μετά από διαδικασία εκπαίδευσης για την αντιστοίχηση των λέξεων στις κατάλληλες έννοιες. Το σύστημα PERSONA (Tanudjaja &Mui, 2002) χρησιμοποιεί προφίλ χρηστών βασισμένο στην ιεραρχία εννοιών του Open Directory Project (ODP). Η αναπαράσταση των προφίλ γίνεται με το χρωματισμό δέντρων.

9.2.1.5 Ο Αλγόριθμος του Rocchio

Οι μέθοδοι που βασίζονται στην υπόθεση ότι οι χρήστες μπορούν να χαρακτηρίσουν μόνοι τους την σημαντικότητα ή τη χρησιμότητα ενός εγγράφου, συχνά αναφέρονται ως, μέθοδοι σχετικής γνώμης. Έτσι η επιτυχία των αλγορίθμων που κάνουν χρήση των διανυσμάτων χώρου (Vector Space) οφείλεται στην ικανότητα του χρήστη να δημιουργεί ερωτήματα επιλέγοντας από μία ομάδα αντιπροσωπευτικών λέξεων. Η δημιουργία των ερωτημάτων, γίνεται με βάση τις προηγούμενες αναζητήσεις αλλά και προτιμήσεις του χρήστη. Ο σχολιασμός των χρηστών σχετικά με τη χρησιμότητα των αποτελεσμάτων μιας αναζήτησης (στην περίπτωση των συστημάτων συστάσεων μιας διατυπωμένης πρότασης), λαμβάνεται υπόψη για τον επανασχεδιασμό και βελτίωση του αρχικού ερωτήματος . Ο αλγόριθμος του Rocchio είναι ευρέως χρησιμοποιούμενος και βασίζεται στη διαμόρφωση του αρχικού ερωτήματος, τα αποτελέσματα του οποίου αξιολογούν οι χρήστες αναλόγως τα σχετικά και τα μη σχετικά αντικείμενα-αποτελέσματα. Η γενική αρχή είναι να επιτρέψει στους χρήστες να αξιολογήσουν τα έγγραφα που προτείνονται από το σύστημα σε σχέση με τις ανάγκες πληροφόρησης τους (relevance feedback). Αυτή η μορφή της ανατροφοδότησης χρησιμοποιείται στη συνέχεια για να βελτιώσει σταδιακά το προφίλ του χρήστη ή να εκπαιδεύσει τον αλγόριθμο. Τα βήματα του αλγορίθμου είναι τα παρακάτω:

1. Χρήστης: Υποβολή Ερωτήματος.
2. Σύστημα: Επιστροφή συνόλου ανακτημένων εγγράφων.
3. Χρήστης: Κατάδειξη των Σχετικών και των Μη Σχετικών εγγράφων.
4. Σύστημα: Υπολογισμός καλύτερης κατάταξης εγγράφων σύμφωνα με την
5. σχετική ανατροφοδότηση του χρήστη.
6. Σύστημα: Παρουσίαση αναθεωρημένου συνόλου εγγράφων.

Η ανατροφοδότηση μπορεί να επιτευχθεί σε μία ή περισσότερες επαναλήψεις. Είναι δύσκολο να διαμορφωθεί ένα καλό ερώτημα, όταν δεν είναι γνωστό το σύνολο των εγγράφων. Με την ανατροφοδότηση μπορεί να γίνει βελτίωση του. Ο αλγόριθμος του Rocchio χρησιμοποιεί την παρακάτω σχέση (1):

$$Q_1 = \alpha\ Q_0 + \frac{\beta}{n_1}\sum_{i=1}^{n_1} R_i - \frac{\gamma}{n_2}\sum_{i=1}^{n_2} S_i$$

Όπου

Q_1 είναι το τροποποιημένο διάνυσμα του οποίου τα στοιχεία το πλησιάζουν ή το απομακρύνουν από την αρχική επιλογή του χρήστη.

Q_0 είναι το διάνυσμα του αρχικού ερωτήματος του χρήστη.

R_i είναι το διάνυσμα των συναφών (σχετικών) κειμένων.

S_i είναι το διάνυσμα των μη συναφών (σχετικών) κειμένων.

n_1 είναι ο αριθμός των συναφών (σχετικών) κειμένων που επιλέγηκαν, δηλαδή τα κείμενα στα οποία εμφανίστηκε τουλάχιστον ένας όρος που ενδιαφέρει το χρήστη.

n_2 είναι ο αριθμός των μη συναφών (σχετικών) κειμένων που επιλέγηκαν.

α,β, και γ είναι οι συντελεστές που ρυθμίζουν τη σημαντικότητα των συναφών και των μη συναφών κειμένων. Σε αρκετές μελέτες α=1, β=0,75 και το γ=0,25.

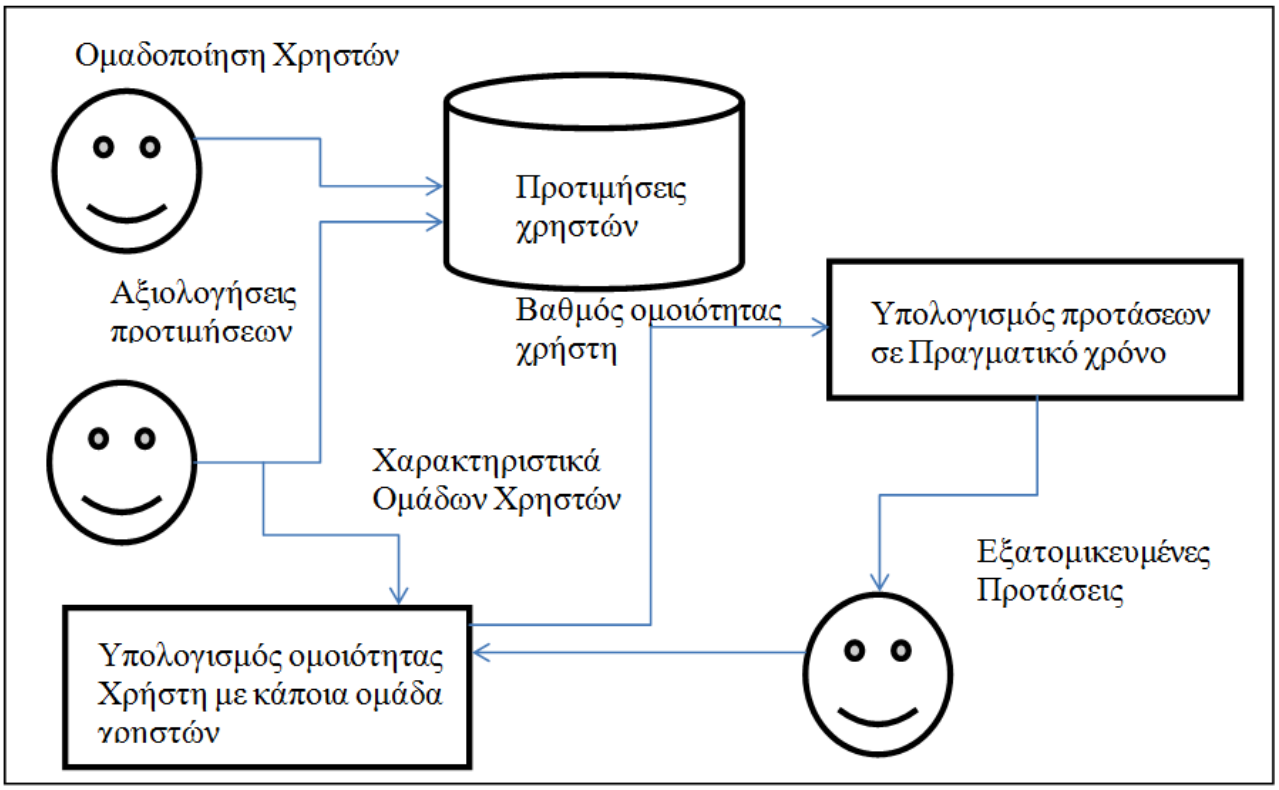
Εάν, για παράδειγμα, θεωρήσουμε το παρακάτω σύνολο εννοιών ή κειμένων ή επιλογών (στα συστήματα συστάσεων):

Σύνολο Επιλογών (ΣΕ)= {θέατρα, κινηματογράφοι, ζωντανή μουσική, εστιατόρια, καφέ), το διάνυσμα του αρχικού ερωτήματος ενός χρήστη που αναφέρεται σε *θέατρα* και *ζωντανή μουσική* έχει την παρακάτω μορφή:

Q_0=[1,0,1,0,0]. Τα στοιχεία των διανυσμάτων R_i και S_i δηλώνουν για κάθε κείμενο που ανακτάται τον αριθμό των εμφανίσεων του κάθε όρου από το ΣΕ. Έστω ότι το R_i=[2,0,5,2,3], τότε γίνεται σαφές ότι στο κείμενο (i) έχουμε 2 εμφανίσεις του όρου «θέατρα», καμία του όρου «κινηματογράφοι», 5 φορές εμφανίστηκε ο όρος «ζωντανή μουσική», κλπ. Αντίστοιχα ορίζεται και το διάνυσμα S_i. Μετά τις αριθμητικές πράξεις κατά την εφαρμογή του τύπου (1), υπολογίζεται το τροποποιημένο διάνυσμα Q_1. Τα διαδοχικά διανύσματα-αποτελέσματα που προκύπτουν από την διαδοχική εφαρμογή του αλγόριθμου αξιολογούνται με βάση το κέντρο βάρους (centroid) των συντεταγμένων (στοιχείων) των διανυσμάτων.

9.2.2 Τεχνικές συνεργατικού φιλτραρίσματος (collaborative filtering)

Οι προτάσεις σύμφωνα με το συνεργατικό φιλτράρισμα βασίζονται στη δημιουργία μοντέλου χρηστών, συλλέγοντας πληροφορίες και προτιμήσεις από πολλούς χρήστες. Είναι ίσως η πιο διαδεδομένη προσέγγιση για την ανάπτυξη συστήματος συστάσεων. Όλες οι πληροφορίες αναλύονται με στόχο να προσδιοριστούν οι χρήστες που οι προτιμήσεις τους είναι πιο κοντά, δηλαδή έχουν μεγαλύτερο βαθμό ομοιότητας. Για παράδειγμα ας υποθέσουμε ότι στόχος μας είναι η δημιουργία συστήματος προτάσεων που θα προτείνει ξενοδοχεία. Συλλέγονται λοιπόν στοιχεία από τουρίστες, τα οποία στη συνέχεια αναλύονται ώστε να διαμορφωθούν οι ομάδες τουριστών με παρόμοιες προτιμήσεις. Τότε για κάποιο πελάτη ξενοδοχείου, που ανήκει σε μια ομάδα με παρόμοιες προτιμήσεις, μπορεί να διατυπωθεί πρόταση για ξενοδοχείο-α, με βάση τις προτιμήσεις των άλλων τουριστών της ομάδας του. Ο συγκεκριμένος πελάτης δεν έχει επισκεφθεί ποτέ το προτεινόμενο-α ξενοδοχεία, αλλά επειδή άλλοι πελάτες με όμοιες προτιμήσεις με αυτόν, έχουν χρησιμοποιήσει τις υπηρεσίες των προτεινόμενων ξενοδοχείων, τότε πιθανότατα οι προτάσεις να έχουν ενδιαφέρον και για τον ίδιο. Οι αρχές δομής και λειτουργίας των συστημάτων συστάσεων που χρησιμοποιούν το συνεργατικό φιλτράρισμα απεικονίζονται στο Σχήμα 9.4.



Σχήμα 9.4. *Δομή και Αρχές λειτουργίας ενός συστήματος συστάσεων με βάση το συνεργατικό φιλτράρισμα*

Για κάθε χρήστη λοιπόν υπολογίζεται η ομοιότητα του με άλλους χρήστες ή ομάδες χρηστών. Για τον υπολογισμό αυτό δημιουργείται ένας πίνακας R, που συνδέει χρήστες και χαρακτηριστικά και του οποίου τα στοιχεία $r_{a,i}$ δηλώνουν το βαθμό προτίμησης του χρήστη (a) για το χαρακτηριστικό (i) και r_a είναι η μέση αξιολόγηση του χρήστη (a) για όλα τα χαρακτηριστικά μιας υπηρεσίας ή προϊόντος. Τότε χρησιμοποιώντας το δημοφιλή τύπο συσχέτισης του Pearson, υπολογίζεται η ομοιότητα δύο χρηστών.

$$sim(a,b) = \frac{\sum_i (r_{ai} - r_a)(r_{bi} - r_b)}{\sqrt{\sum_i (r_{ai} - r_a)^2 \sum_i (r_{bi} - r_b)^2}},$$

όπου $sim(a,b)$ δηλώνει το βαθμό ομοιότητας μεταξύ των χρηστών a και b αντίστοιχα. Βεβαίως, η ομοιότητα υπολογίζεται για κάθε ζεύγος χρηστών.

Ο πίνακας R μπορεί διαγραμματικά να αποδοθεί όπως παρακάτω:

	Χαρακτηριστικό-1	**Χαρακτηριστικό-2**	**Χαρακτηριστικό-3**
Χρήστης-1	7	2	**?**
Χρήστης-2	8	3	5
Χρήστης-3	9	5	2
Χρήστης-4	**?**	2	3
Χρήστης-5	9	5	3

Πίνακας 9.2.*.Πίνακας προτιμήσεων χαρακτηριστικών από χρήστες*

Όπως φαίνεται από τον πίνακα, είναι γνωστές οι προτιμήσεις των χρηστών αλλά όχι όλες. Για το χρήστη «Χρήστης-1» δεν είναι γνωστός ο βαθμός προτίμησης του/της για το «Χαρακτηριστικό-3». Παρομοίως για το «Χρήστης-4» κα το «Χαρακτηριστικό-1». Μπορεί να εκτιμηθεί ώστε να μπορεί αποφασιστεί εάν το «Χαρακτηριστικό-3» θα πρέπει να προταθεί ή όχι; Μπορεί να εκτιμηθεί εμμέσως. Με βάση τις εκτιμήσεις των άλλων χρηστών. Ποιών από όλους; Θα ληφθούν υπόψη οι προτιμήσεις μόνο των χρηστών με τους οποίους ο «Χρήστης-1» έχει βαθμό ομοιότητας, και πιο συγκεκριμένα τους περισσότερο όμοιους του, με βάση το βαθμό ομοιότητας $sim(a,b)$. Οι προτάσεις λοιπόν προς το «Χρήστης-1», θα επιλεγούν από τις προτιμήσεις των «όμοιων» χρηστών. Οι προτάσεις θα είναι οι N πρώτες σε ταξινόμηση, ταξινομημένες με βάση τους σταθμισμένους (λαμβάνοντας δηλαδή υπόψη τους βαθμούς προτίμησης όλων των όμοιων χρηστών) βαθμούς προτίμησης $r_{a,i}$, για κάθε χαρακτηριστικό. Ένας τρόπος προσδιορισμού των προτάσεων δίνεται πιο κάτω:

$$p_{a,i} = \bar{r}_a + \frac{\sum_k r_{u,i} - \bar{r}_u * sim(a,u)}{\sum_k sim(a,u)},$$

όπου $p_{a,i}$ είναι η πρόβλεψη του συστήματος συστάσεων ότι στο χρήστη (a) ταιριάζει και προτείνεται το χαρακτηριστικό της υπηρεσίας/προϊόντος (i), και το $sim(a,u)$ δηλώνει το βαθμό ομοιότητας μεταξύ των χρηστών (a) και , (u), τα $r_{a,i}$ δηλώνουν το βαθμό προτίμησης του χρήστη (a) για το χαρακτηριστικό (i) και το (K), δηλώνει το σύνολο των όμοιων χρηστών με το χρήστη (α), για τον οποίο γίνεται η πρόταση.

Οι τεχνικές συνεργατικού φιλτραρίσματος προτιμούνται σε περιπτώσεις όπου για λόγους ιδιωτικότητας (privacy) δεν είναι διαθέσιμες αξιολογήσεις συγκεκριμένων χρηστών. Επίσης σε ορισμένες περιπτώσεις όπως ανάλυση μουσικής ή βίντεο, είναι πιο δύσκολο για λογισμικό να αναλύσει τα χαρακτηριστικά τους, οπότε η ανάλυση των χαρακτηριστικών τους βασίζεται σε γνώμες πολλών χρηστών (Jin & Si, 2004).

9.2.3 Τεχνικές με βάση τη γνώση (knowledge-based)

Τα συστήματα συστάσεων χρησιμοποιούν γνώση σχετικά με τα χαρακτηριστικά των υπηρεσιών και των προϊόντων που αποτελούν δυνητικές προτάσεις. Η γνώση αναφέρεται στο πως και κατά, πόσο κάθε χαρακτηριστικό και υπηρεσία θα μπορούσε να εξυπηρετήσει μία ή και περισσότερες ανάγκες κάποιου χρήστη. Επίσης χρησιμοποιούν και γνώση σχετικά με τις ανάγκες και προτεραιότητες των χρηστών. Οι προτάσεις διαμορφώνονται από ένα μηχανισμό εξαγωγής συμπερασμάτων (inference engine), ο οποίος προσπαθεί να προσδιορίσει ποιες υπηρεσίες και χαρακτηριστικά μπορούν να καλύψουν μια ή περισσότερες ανάγκες ενός χρήστη. Ο μηχανισμός εξαγωγής συμπερασμάτων αποτελείται συνήθως από κανόνες της μορφής “if <> then <> else”. Όπως και στις προηγούμενες τεχνικές έτσι και εδώ οι προτάσεις διαμορφώνονται με βάση κάποια συνάρτηση ομοιότητας η οποία λαμβάνει υπόψη το βαθμό κατά τον οποίο τα χαρακτηριστικά μιας υπηρεσίας ανταποκρίνονται στις ανάγκες και τα ενδιαφέροντα του χρήστη. Ο βαθμός ικανοποίησης της ανάγκης του χρήστη προσδιορίζει και τη χρησιμότητα της πρότασης για τον συγκεκριμένο χρήστη. Η τεχνική με βάση τη γνώση προτιμάται σε περιπτώσεις όπου οι δυνητικές προτάσεις αφορούν ακριβά προϊόντα, ή προϊόντα που δεν αγοράζονται συχνά και επομένως δεν υπάρχουν αρκετές αξιολογήσεις. Επίσης είναι χρήσιμη όταν πρόκειται να προταθούν προϊόντα τεχνολογίας όπου η διάσταση του χρόνου (απαξίωση προϊόντος, ταχύτητα αλλαγών μεγάλη) αλλά και όπου έχουμε περιπτώσεις σαφούς διατύπωσης συγκεκριμένων και ιδιαίτερων προτεραιοτήτων και αναγκών χρηστών, όπου δύσκολα θα μπορούσαν να ενταχθούν σε ευρύτητα σύνολα αναγκών. Τέλος, η τεχνική αυτή θα προτιμηθεί όταν οι συναρτήσεις ομοιότητας (similarity) στα πλαίσια της τεχνικής με βάση το περιεχόμενο δεν είναι αρκετή για να εξαχθούν επιτυχείς προτάσεις. Η γνώση σχετικά με τις υπηρεσίες και τα προϊόντα συνήθως αναπαριστάται χρησιμοποιώντας οντολογίες (ontologies), όπως περιγράφηκε στο Κεφάλαιο 8.

9.2.4 Τεχνικές Υβριδικές (Hybrid)

Αποτελούν συνδυασμό των παραπάνω, αλλά με πιο συχνό συνδυασμό αυτό που συνδυάζει το συνεργατικό φιλτράρισμα με άλλες τεχνικές ώστε να αποφευχθούν τα προβλήματα της ψυχρής εκκίνησης και των αραιών δεδομένων τα οποία περιγράφονται πιο κάτω.

Το σύστημα OBIWAN (Gauch, et. al., 2003) είναι υβριδικό καθώς χρησιμοποιεί έγγραφα τα οποία είναι οργανωμένα σε εννοιολογικές ομάδες (προσέγγιση συμβατή με τα σημασιολογικά δίκτυα) αλλά χρησιμοποιεί και λεξικά όπως τα Magellan, Lycos, καθώς και το Open Directory Project (προσέγγιση συμβατή την στάθμιση σε έννοιες). Το σύστημα συλλέγει δεδομένα για το χρήστη κατά τη διάρκεια πλοήγησής του τα οποία αποθηκεύονται στη λανθάνουσα μνήμη (cache memory) από το ιστορικό των ερευνών του αλλά και από τον proxy server.

Το Bibster (Haase et al., 2004) είναι ακόμη ένα υβριδικό σύστημα που συνδυάζει μοντέλα ενδιαφερόντων των χρηστών που αποτυπώνουν τις έννοιες που αποτελούν τις προτεραιότητες στα ενδιαφέροντα των χρηστών και αλγόριθμους collaborative filtering. Τα προφίλφ των χρηστών διαμορφώνονται είτε αυτόματα από το σύστημα είτε μετά από παρεμβάσεις των χρηστών που μπορούν να μεταβάλλουν τα ενδιαφέροντά τους. Η προσέγγιση collaborative filtering ενεργοποιείται κάθε φορά που πραγματοποιείται μία αλλαγή στα ενδιαφέροντα ενός χρήστη κοινοποιώντας τις αλλαγές αυτές και στους άλλους χρήστες. Ο συνδυασμός του προφίλ του χρήστη και του collaborative filtering, παρέχει το πλεονέκτημα ότι ανατροφοδοτούνται τα προφίλ των χρηστών (και με παρέμβαση του χρήστη και με collaborative filtering)

9.3 Προβλήματα στην Ανάπτυξη των Συστημάτων Συστάσεων

Για την ανάπτυξη των συστημάτων συστάσεων θα πρέπει πρώτα να αξιολογηθούν και να αντιμετωπιστούν διάφορα προβλήματα τα σημαντικότερα των οποίων είναι:

- **Αραιά και Διάσπαρτα Δεδομένα (Data Sparsity).** Οι προτάσεις βασίζονται σε γνώση των προτεραιοτήτων των χρηστών και των αξιολογήσεών τους σε υπηρεσίες και προϊόντα. Δεν είναι σπάνιο όμως το φαινόμενο όπου αρκετοί χρήστες δεν αξιολογούν τα περισσότερα από τα διαθέσιμα χαρακτηριστικά των υπηρεσιών και των προϊόντων. Το αποτέλεσμα είναι να είναι διαθέσιμα δεδομένα με αρκετές ελλείψεις, δηλαδή να είναι αραιά και διάσπαρτα και αυτό συνεπώς επηρεάζει την ευστοχία και την αποδοχή των προτάσεων δηλαδή, την αποτελεσματικότητα του συστήματος συστάσεων. Το πρόβλημα αυτό παρουσιάζεται κυρίως σε περιπτώσεις όπου έχουμε μεγάλο αριθμό χαρακτηριστικών που θα πρέπει να έχουμε αξιολογήσεις σε σχέση με τον αριθμό των χρηστών-πελατών που έχουν χρησιμοποιήσει τις υπηρεσίες και κάνουν αξιολογήσεις. Για να αρθεί το πρόβλημα θα πρέπει να συλλεχθούν περισσότερα δεδομένα.
- **Το πρόβλημα της ψυχρής εκκίνησης (Cold-Start Problem)**. Νέα χαρακτηριστικά και υπηρεσίες, καθώς και νέοι χρήστες αποτελούν θέματα προς αντιμετώπιση, γιατί εφόσον είναι νέα στοιχεία στο σύστημα, πιθανόν να μην υπάρχει η ανάλογη και κατάλληλη πληροφόρηση. Επομένως η οποιαδήποτε διαμόρφωση πρότασης είναι αμφίβολη ως προς την επιτυχία της. Ένα χαρακτηριστικό υπηρεσίας (ειδικά στα συστήματα συνεργατικού φιλτραρίσματος), δεν μπορεί να αποτελέσει δυνητική πρόταση, εφόσον δεν έχει τουλάχιστον μία αξιολόγηση (Schein, et al. 2002). Αυτό το πρόβλημα απασχολεί επίσης στις περιπτώσεις όπου ένας χρήστης δεν έχει ιδιαίτερα συνηθισμένες προτεραιότητες, αλλά οι ανάγκες του διαφοροποιούνται γενικότερα. Για την αντιμετώπιση του προβλήματος, όταν λείπουν δεδομένα αξιόλογης χαρακτηριστικών, μια συνηθισμένη στρατηγική είναι να πραγματοποιείται πρώτα συλλογή και ανάλυση αναγκών των χρηστών με τεχνικές βασισμένες στο περιεχόμενο. Με αυτό τον τρόπο γίνονται διαθέσιμες κάποιες πρώτες αξιολογήσεις οι οποίες θα μπορούσαν να υποστηρίξουν προτάσεις έως ότου περισσότερα δεδομένα γίνουν διαθέσιμα μέσω άλλων τεχνικών όπως για παράδειγμα του συνεργατικού φιλτραρίσματος (Mooney & Roy 2000). Η περίπτωση στην οποία το σύστημα δέχεται ένα νέο χρήστη είναι σαφώς πιο δύσκολη, εφόσον χωρίς προηγούμενες αξιολογήσεις δεν είναι δυνατό να προσδιοριστούν «παρόμοιοι» χρήστες ή να διαμορφωθεί το προφίλ του χρήστη με

βάση το περιεχόμενο. Ερευνητικές προσπάθειες, όπως στο *active learning*, εστιάζουν στο πρόβλημα έτσι ώστε να μπορούν να διατυπωθούν προτάσεις με την ελάχιστη δυνατή πληροφόρηση είτε σε προφίλ χρηστών είτε σε αξιολογήσεις υπηρεσιών και προϊόντων (Harpale, & Yang, 2008; Jin & Si, 2004). Ο όρος *active learning*, αναφέρεται στην ανάπτυξη συναρτήσεων που μπορούν να «μάθουν» τις προτιμήσεις με τα ελάχιστα δυνατά δεδομένα.

- **Το πρόβλημα της Απάτης (Fraud).** Η αναπτυσσόμενη χρήση των συστημάτων συστάσεων για επιχειρηματικούς λόγους, καθιστά τα συστήματα ολοένα και πιο σημαντικά για την κερδοφορία των ηλεκτρονικών κυρίως επιχειρήσεων. Έτσι εμφανίζεται και το πρόβλημα της απάτης (Burke, et al., 2005) όπου κάποιος προσπαθεί να αυξήσει πλασματικά το βαθμό επιθυμίας ενός προϊόντος (push attacks), δείχνοντας έτσι πόσο πολύ το θέλουν άλλοι πελάτες. Με παρόμοιο τρόπο, άλλο είδος απάτης προσπαθεί να μειώσει το βαθμό επιθυμίας ενός προϊόντος (nuke attacks). Η πραγματοποίηση αυτών των ειδών απάτης πραγματοποιείται με τη δημιουργία ψευδών προφίλ, την ανάπτυξη πολλών λογαριασμών χρηστών αλλά και προϋποθέτουν την γνώση σχετικά με κάποια στοιχεία όπως για παράδειγμα την μέση τιμή αξιολόγησης (rating) ενός χαρακτηριστικού (Lam & Riedl, 2004). Τότε αυτός που πραγματοποιεί την επίθεση με στόχο την απάτη, διαμορφώνει αρκετές κριτικές και αξιολογήσεις με τιμές γύρω από τη μέση αξιολόγηση. Μετά διατυπώνει και μια αξιολόγηση σημαντικά πάνω από τη μέση τιμή με στόχο να την παρασύρει προς τα πάνω, διαμορφώνοντας μια πιο «επιθυμητή» εικόνα για το προϊόν. Μελέτες δείχνουν ότι οι πρακτικές αυτές είναι αρκετά αποτελεσματικές (Lam & Riedl, 2004), αν και δεν μπορούν να εφαρμοστούν σε αξιολογήσεις χρηστών που διαμορφώνονται με τεχνικές βασισμένες στο περιεχόμενο.

Άλλα προβλήματα είναι τα παρακάτω:

- **Το πρόβλημα των συνώνυμων (The synonyms problem).** Πολλές φορές τα συστήματα αδυνατούν να αντιληφθούν μια έννοια εάν είναι διατυπωμένη με συνώνυμα. Για παράδειγμα, οι όροι «παιδική κινηματογραφική ταινία» και «παιδικό κινηματογραφικό έργο», μπορεί να προσδιορίζουν διαφορετικές προτάσεις.
- **Το πρόβλημα του Γκρι και του Μαύρου πρόβατου (Grey sheep problem, Black sheep problem).** Ο όρος *grey sheep* αναφέρεται στο χρήστη/τες που οι προτιμήσεις τους δεν διαφοροποιούνται αρκετά, έτσι ώστε να μπορούν να βοηθηθούν από την όποια ομοιότητα τους με άλλες ομάδες χρηστών. Αντιθέτως ο όρος *black sheep* αναφέρεται σε χρήστες των οποίων οι προτιμήσεις είναι ριζικά διαφορετικές, κάνοντας έτσι την πρόβλεψη για προτάσεις εξαιρετικά δύσκολη.

9.4 Πλεονεκτήματα και Μειονεκτήματα Τεχνικών.

Κάθε μια διαφορετική στρατηγική στην διαμόρφωση των συστάσεων, κάθε μια διαφορετική τεχνική δεν αποτελεί πανάκεια αλλά έχει πλεονεκτήματα αλλά και μειονεκτήματα. Ο πιο κάτω πίνακας συνοψίζει τα υπέρ και τα κατά για κάθε μια κατηγορία τεχνικών.

9.4.1 Τεχνική-Μέθοδος: Με βάση το περιεχόμενο (content-based)

Πλεονεκτήματα:

- Οι τεχνικές αυτές για να εφαρμοστούν δεν χρειάζονται αξιολογήσεις από πολλούς χρήστες. Για τη διατύπωση των συστάσεων τους βασίζονται μόνο στις προτιμήσεις του συγκεκριμένου χρήστη για τον οποίο προορίζονται και οι προτάσεις. Κατά συνέπεια δεν χρειάζεται να υπολογιστεί και ο βαθμός ομοιότητας κάθε χρήστη με άλλους.
- Βασιζόμενες οι τεχνικές σε προτιμήσεις του ίδιου του χρήστη, και όχι άλλων «όμοιων» του, οι τεχνικές αυτές είναι σε θέση να προτείνουν προϊόντα και υπηρεσίες σε χρήστες-πελάτες με ξεχωριστά γούστα. Αυτό είναι ένα σημαντικό χαρακτηριστικό εφόσον τα συστήματα δεν προτείνουν μόνο «προφανείς» λύσεις. Επίσης δεν αντιμετωπίζουν πρόβλημα όταν δεν

υπάρχουν προηγούμενες αξιολογήσεις κάποιου προϊόντος καθώς τα δεδομένα του κάθε προϊόντος είναι έτσι και αλλιώς διαθέσιμα. Αυτό που χρειάζεται είναι οι προτιμήσεις του χρήστη οι οποίες είναι και αυτές δεδομένες από το χρήστη. Α αξιολόγηση του προϊόντος γίνεται με βάση τις προτεραιότητες του χρήστη και τα χαρακτηριστικά του προϊόντος.
- Τα συστήματα με βάση το περιεχόμενο μπορούν να εξηγήσουν τη λογική πίσω από τη διαμόρφωση των προτάσεών τους, αυξάνοντας έτσι την εμπιστοσύνη των χρηστών προς αυτά και συνεπώς την αποδοχή των προτάσεων και τελικά την επιτυχία των συστημάτων.

Μειονεκτήματα

- Η επιτυχία ενός συστήματος συστάσεων βασισμένο στη γνώση εξαρτάται από την ποιότητα και τη ποσότητα των δεδομένων σχετικά με τα χαρακτηριστικά των προϊόντων και των υπηρεσιών. Εάν δεν είναι διαθέσιμα τότε το σύστημα αντιμετωπίζει προβλήματα, ώστε να διαχωρίσει ποια προϊόντα θα αποτελούν καλές προτάσεις. Συνήθως οι περιγραφές των προϊόντων δεν είναι πλήρης, ή δεν έχουν την απαραίτητη για ένα σύστημα προτάσεων πληροφόρηση, καθιστώντας τις προτάσεις επισφαλείς.
- Για να είναι το σύστημα αξιόπιστο , πρέπει να έχει αρκετά δεδομένα από προηγούμενες επιλογές του χρήστη αλλά το βαθμό ικανοποίησης του από αυτές τι επιλογές.
- Τα συστήματα αυτά δεν έχουν τρόπους που να προσδιορίζουν προτάσεις πέρα από τις προφανείς προτάσεις που ένας χρήστης θέλει. Μπορούν να προτείνουν σε πελάτες με ξεχωριστά γούστα, αλλά δεν μπορούν για αυτό τον πελάτη να προτείνουν κάτι ξεχωριστό. Προτείνουν περισσότερα από τα «ίδιου» ύφους και χαρακτηριστικών προϊόντα που συνήθως επιλέγει ο πελάτης. Δηλαδή, το σύστημα δε θα προσδιορίσει νέα προϊόντα και υπηρεσίες που δεν έχουν σχέση με όσα έχει κοιτάξει ο χρήστης στο παρελθόν αλλά που πιθανώς να τον ενδιαφέρουν.

9.4.2 Τεχνική-Μέθοδος: Με συνεργατικό φιλτράρισμα (collaborative filtering)

Πλεονεκτήματα

- Έχουν καλύτερη απόδοση. Τα συστήματα με συνεργατικό φιλτράρισμα προσδιορίζουν επιτυχείς προτάσεις στις περισσότερες περιπτώσεις.
- Είναι σχετικά πιο απλά και εύκολα στη σχεδίαση και κατασκευή τους.
- Σταθερότητα προτάσεων. Από τη στιγμή που έχουν διαμορφωθεί οι ομοιότητες μεταξύ χαρακτηριστικών υπηρεσιών και προϊόντων, προτάσεις για ένα νεό προϊόν μπορούν να γίνουν αμέσως.

Μειονεκτήματα

- Απαιτούν μεγάλο όγκο δεδομένων για τους χρήστες και τα προϊόντα, για να μπορούν να εξάγουν ασφαλείς προτάσεις. Δηλαδή αντιμετωπίζουν το πρόβλημα με τα αραιά δεδομένα.
- Οι πληροφορίες για τα προϊόντα πρέπει να είναι άμεσα συγκρίσιμες και τυποποιημένες, έτσι ώστε όταν ένας χρήστης διατυπώνει μια προτίμηση να μπορεί το σύστημα να γνωρίζει τα ιδιαίτερα χαρακτηριστικά αυτής και να μπορεί να την συγκρίνει με άλλη.
- Κάνουν την υπόθεση, όχι πάντα σωστή, ότι η προηγούμενη συμπεριφορά ενός χρήστη θα τον χαρακτηρίζει για πάντα, χωρίς να λαμβάνουν υπόψη τις συνθήκες που διαμορφώνουν μια συγκεκριμένη συμπεριφορά. (contextual information).
- Αντιμετωπίζουν τα προβλήματα με το *Grey sheep, Black sheep*.
- Χρειάζεται να πάρουν μέτρα προστασίας για την αποφυγή απάτης.

9.4.3 Τεχνική-Μέθοδος: Με βάση τη γνώση (knowledge-based)

Πλεονεκτήματα

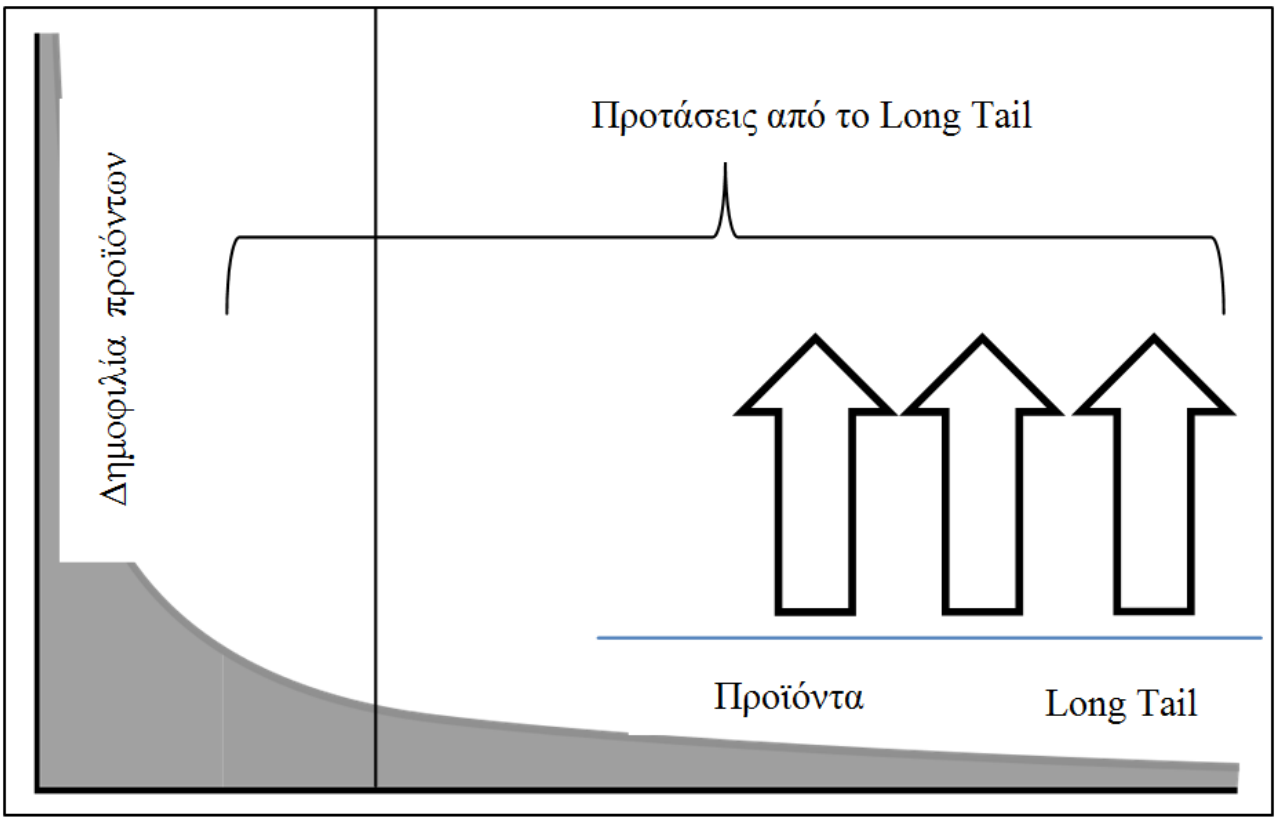
- Δεν αντιμετωπίζουν πρόβλημα με την «ψυχρή εκκίνηση», εφόσον με τους κανόνες που έχουν μπορούν να προτείνουν λύσεις συνδυάζοντας χαρακτηριστικά και όχι κριτικές. Οι κανόνες περιγράφουν το «γιατί» ένα χαρακτηριστικό καλύπτει μια ανάγκη, χωρίς να εξαρτάται από κριτικές.
- Επιδεικνύουν ψηλή ποιότητα στις προτάσεις που διατυπώνουν.
- Μπορούν να προσομοιώσουν διαλόγους μεταξύ πωλητή και πελάτη κατά τη διάρκεια της πώλησης.

Μειονεκτήματα

- Απαιτείται μεγάλη προσπάθεια για τη διατύπωση των κανόνων.
- Είναι βασικά στατικά συστήματα, που εάν δεν αλλάξουμε τους κανόνες δεν αλλάζουν τη συμπεριφορά τους και συνεπώς τις προτάσεις τους.
- Δεν προσαρμόζονται σε αλλαγές και τάσεις που διαμορφώνονται σε μικρά χρονικά διαστήματα.

9.5 Αξιολόγηση Συστημάτων Συστάσεων

Η αξιολόγηση των συστημάτων μπορεί να γίνει με διάφορα κριτήρια όπως για παράδειγμα τη συμβολή τους στην, αύξηση των πωλήσεων, ή στην προώθηση προϊόντων και υπηρεσιών (one to one marketing), κλπ. Επίσης μπορούν να αξιολογηθούν με βάση των click-through-rates, του βαθμού ικανοποίησης των πελατών, του βαθμού επιστροφής των πελατών, κλπ. Πέραν όμως των επιχειρηματικών δεικτών απόδοσης έχουν διαμορφωθεί και δείκτες απόδοσης που μετρούν την ποιότητα των προτάσεων που μπορεί ένα σύστημα να προτείνει. Για παράδειγμα, ένα επιτυχημένο σύστημα συστάσεων θα πρέπει να μπορεί να προτείνει προϊόντα και υπηρεσίες από το Long Tail, δηλαδή να μην περιορίζει τις προτάσεις σε προφανή προϊόντα, αλλά να μπορεί να συστήνει προϊόντα σχετικά άγνωστα για τα οποία όμως ο πελάτης θα είχε ενδιαφέρον. Το σημαντικό, αλλά ταυτόχρονα και ασυνήθιστο με το long tail είναι ότι το 20% των προϊόντων, που βρίσκονται στο long tail, δηλαδή που έχουν την μικρότερη συχνότητα, έχουν το 80% της σημαντικότητας. Συνεπώς ένα σύστημα που μπορεί να προτείνει από το long tail αυξάνει την επίδρασή του στην επιχειρηματική αποτελεσματικότητα.



Σχήμα 9.5. *Long Tail και επιτυχία των συστημάτων συστάσεων*

Η επιτυχημένη λειτουργία ενός συστήματος συστάσεων αξιολογείται είτε με on-line παρακολουθώντας τις προτάσεις και τις αντιδράσεις των πελατών για κάποιο χρονικό διάστημα, είτε off-line με ερωτηματολόγια και πειράματα σε εργαστηριακό χώρο. Έχουν αναπτυχθεί συστήματα μέτρησης της απόδοσης όπως το παρακάτω:

$$\Pr ecision = \frac{gp}{sum(p)},$$

όπου το (*gp*) αναφέρεται στις καλές και επιτυχημένες προτάσεις, προς το σύνολο των διατυπωμένων συστάσεων $sum(p)$. Άλλος δείκτης είναι ο δείκτης *recall*:

$$recall = \frac{gp}{sum(gp)},$$

όπου το (*gp*) αναφέρεται πάλι στις καλές και επιτυχημένες προτάσεις, προς το σύνολο των επιτυχημένων συστάσεων $sum(gp)$.

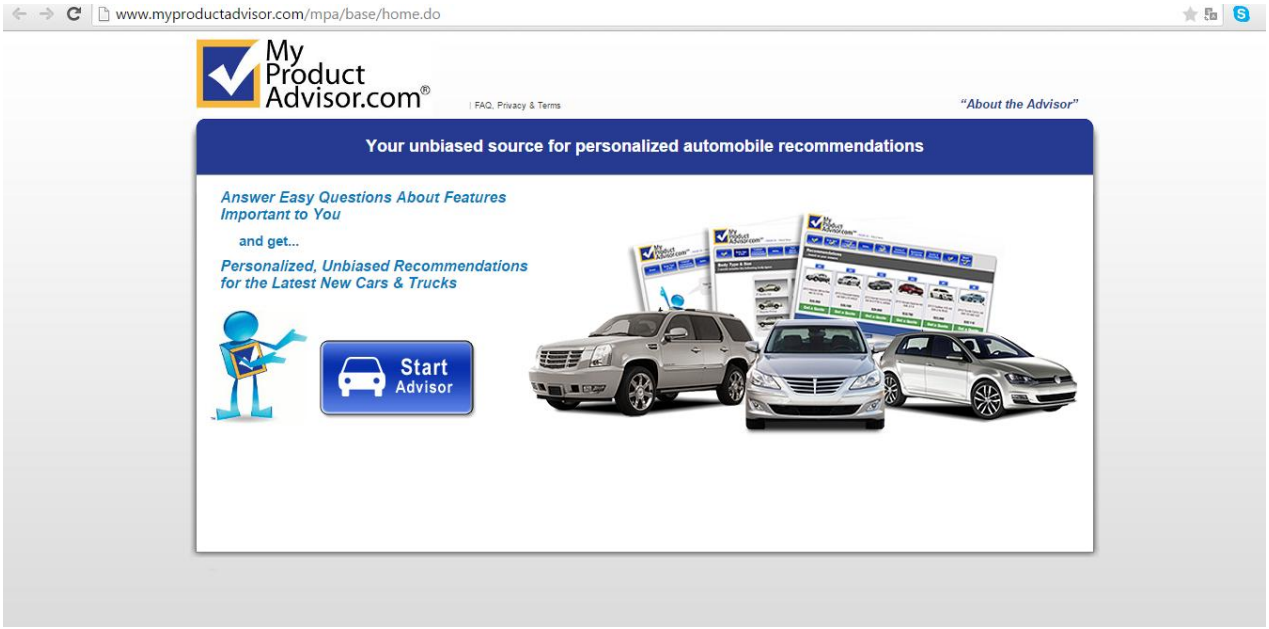
Από τους πιο δημοφιλείς τρόπους αξιολόγησης είναι ο υπολογισμός του RMSE (Root Mean Squared Error), ως εξής:

$$RMSE = \sqrt{\frac{1}{n}\sum_{n}(\hat{r}_{u,i} - r_{u,i})^2}.$$

Η αξιολόγηση αυτή προϋποθέτει τη διενέργεια (*n*) πειραμάτων με το σύστημα συστάσεων και τη καταγραφή των αξιολογήσεων των προτάσεων που προβλέπει το σύστημα για κάθε χρήστη. Η παράμετρος $\hat{r}_{u,i}$, αναφέρεται στην πρόβλεψη τους συστήματος σχετικά με την αξιολόγηση που θα έδινε ο χρήστης (*u*) για το χαρακτηριστικό (*i*). Οι προβλέψεις συγκρίνονται με τις πραγματικές αξιολογήσεις $r_{u,i}$ που δίνει ο κάθε χρήστης (*u*) για κάθε χαρακτηριστικό(*i*). Με τον τύπο του RMSE, συγκρίνονται (*n*) προβλέψεις του συστήματος με τις αντίστοιχες (*n*) πραγματικές αξιολογήσεις χρηστών.

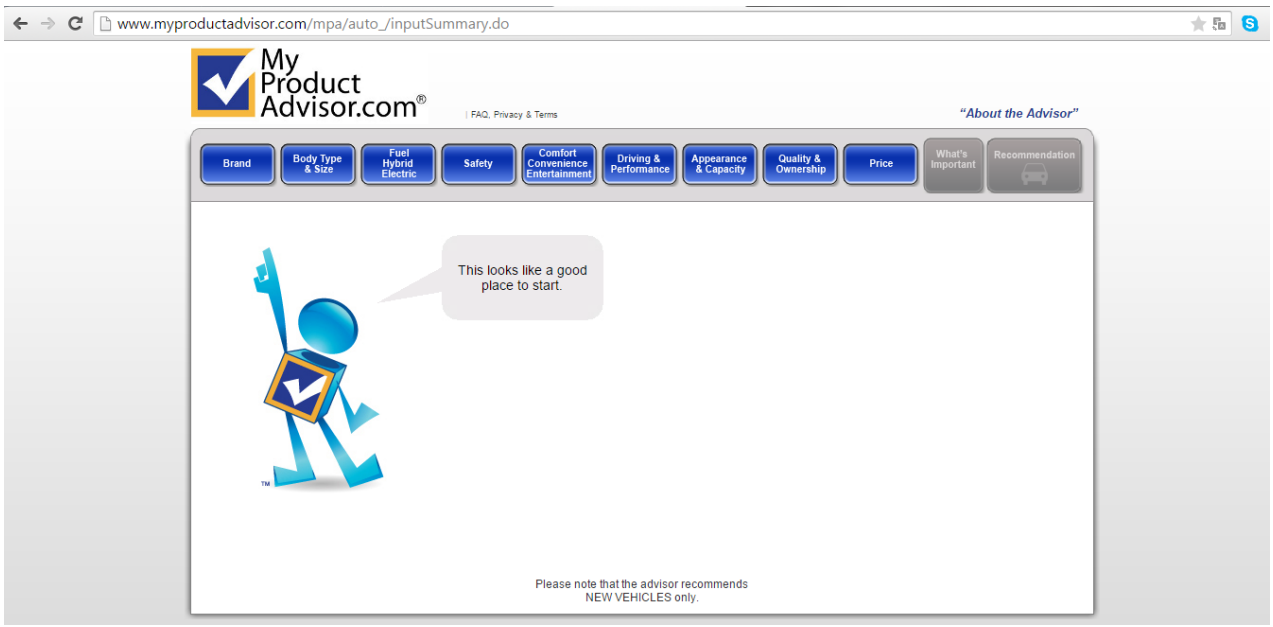
9.6 Εφαρμογές και Παραδείγματα

Εμπειρία χρήσης από συστήματα συστάσεων μπορούμε να έχουμε από ιστοχώρους που χρησιμοποιούμε αρκετά συχνά. Για παράδειγμα το *www.amazon.com*, το οποίο προτείνει προϊόντα χρησιμοποιώντας τεχνικές με βάση το περιεχόμενο. Όταν κάποιος πελάτης-χρήστης επιλέξει να πληροφορηθεί ή να αγοράσει ένα προϊόν, τότε η amazon προτείνει αντίστοιχα προϊόντα που άλλοι πελάτες είχαν αγοράσει. Αντίστοιχα ένας άλλος δημοφιλής ιστοχώρος ο www.linkedin.com, προτείνει στα μέλη του να συνδεθούν με άλλα μέλη που ίσως γνωρίζουν, θέσεις εργασίας που ίσως ενδιαφέρουν, εταιρείες που ίσως ενδιαφέρουν αλλά και ομάδες (groups) που ενδεχομένως ενδιαφέρουν το χρήστη. Το Linkedin.com, έχει αναπτυχθεί χρησιμοποιώντας την πλατφόρμα Apache Hadoop (2015). Υπάρχουν πολλά παραδείγματα συστημάτων συστάσεων που έχουν αναπτυχθεί σε μεγάλο εύρος επιχειρηματικών δραστηριοτήτων, όπως στην *ηλεκτρονική διακυβέρνηση, στον τουρισμό, στην εκπαίδευση, στο ηλεκτρονικό εμπόριο, στις ηλεκτρονικές βιβλιοθήκες, για τα τηλεοπτικά προγράμματα, για μουσική, για ταινίες*, κλπ. (Lu et al, 2015). Ενδεικτικά αναφέρονται τα πιο κάτω. Στο χώρο της αγοράς αυτοκινήτου έχει αναπτυχθεί το «my product advisor» στη διεύθυνση www.myproductadvisor.com. Η κεντρική σελίδα του συστήματος παρουσιάζεται πιο κάτω (Σχήμα 9.6).



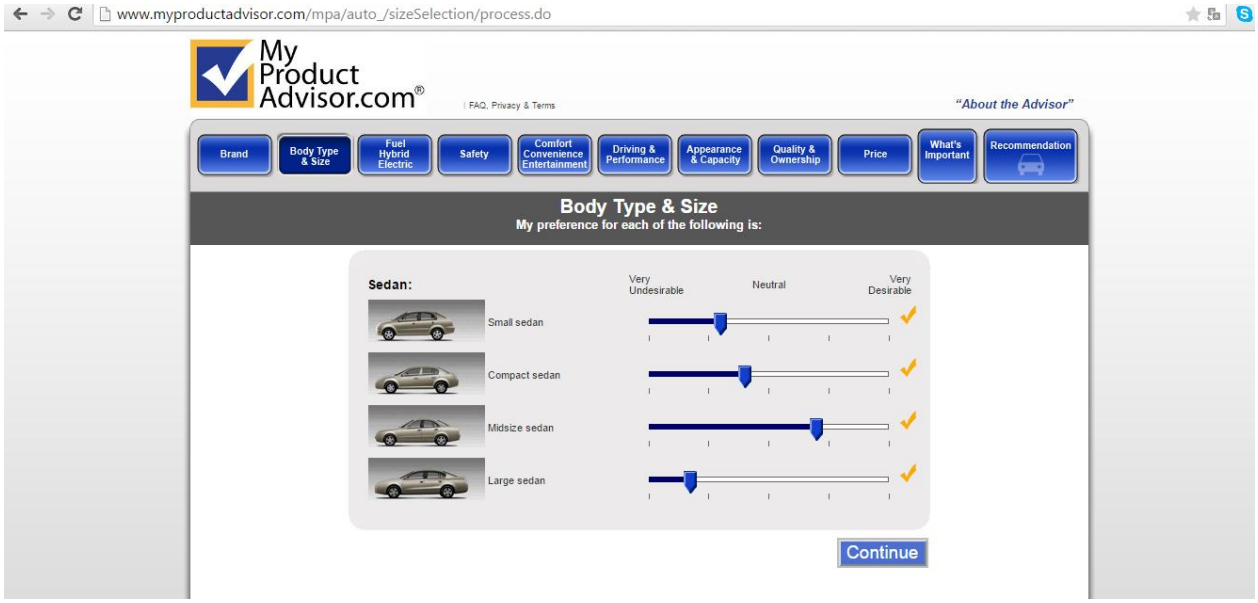
Εικόνα 9.6. *Σύστημα συστάσεων "myproduct.com"*

Το σύστημα δεν διατηρεί το προφίλ του πελάτη αλλά μέσα από μια σειρά ερωτήσεων καταγράφει τις προτεραιότητες του πελάτη ως αναφορά την αγορά ενός αυτοκινήτου και στη συνέχεια του προτείνει ποια ταιριάζουν στις ανάγκες του. Η τεχνική με την οποία το σύστημα ζητά από το χρήστη να συμπληρώσει τις προτεραιότητές του μέσα από μια σειρά ερωτήσεων ονομάζεται "check-box-personalisation" (Kardaras et al, 2013). Η έναρξη καταγραφής των απαιτήσεων του πελάτη παρουσιάζεται στην επόμενη εικόνα (Σχήμα 9.7).



Εικόνα 9.7. *Παράγοντες που λαμβάνονται υπόψη για την παραγωγή συστάσεων.*

Το σύστημα επικοινωνεί με το χρήστη και καταγράφει τις προτεραιότητές του για μια σειρά χαρακτηριστικών, όπως τη μάρκα, τον τύπο αμαξώματος, τις επιδόσεις κλπ., αλλά και το πόσο επιθυμητή είναι η κάθε επιλογή χαρακτηριστικού. Η επόμενη εικόνα δείχνει τη βαρύτητα που θεωρεί ο πελάτης-χρήστης για κάθε εναλλακτική.

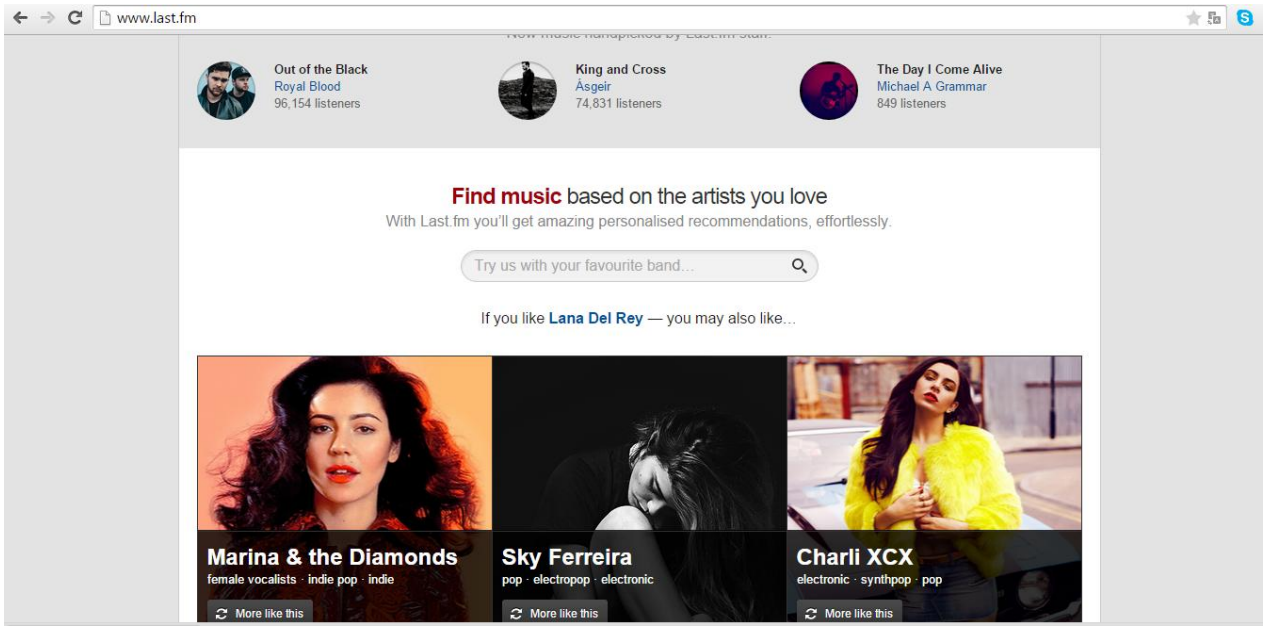


Εικόνα 9.8. *Ο χρήστης δηλώνει τις προτεραιότητές του.*

Το σύστημα αφού συλλέξει τις απαραίτητες πληροφορίες από το χρήστη διαμορφώνει τις προτάσεις του. Οι προτάσεις αφορούν και εναλλακτικές που ενδεχομένως ο χρήστης δεν ανέμενε. Το σύστημα ελέγχει και προτείνει ένα (ή περισσότερα) αυτοκίνητα, έστω και εάν για παράδειγμα είναι διαφορετικού τύπου αμαξώματος ή μάρκας από αυτήν που έχει επιλέξει ο πελάτης, αρκεί να πληροί τις ανάγκες του πελάτη. Το σύστημα επιτρέπει στο χρήστη να διακόψει και να συνεχίσει την καταγραφή των πληροφοριών σε μεταγενέστερη χρονική στιγμή, διευκολύνοντας έτσι τη χρήση του.

Στο χώρο της μουσικής συναντάμε τα www.last.fm, το http://www.apple.com/itunes/?cid=OAS-US-DOMAINS-itunes.com και το http://www.Pandora.comγνωστό και ως Pandora radio, το οποίο όμως είναι διαθέσιμο μόνο στις ΗΠΑ, Αυστραλία και Νέα Ζηλανδία.

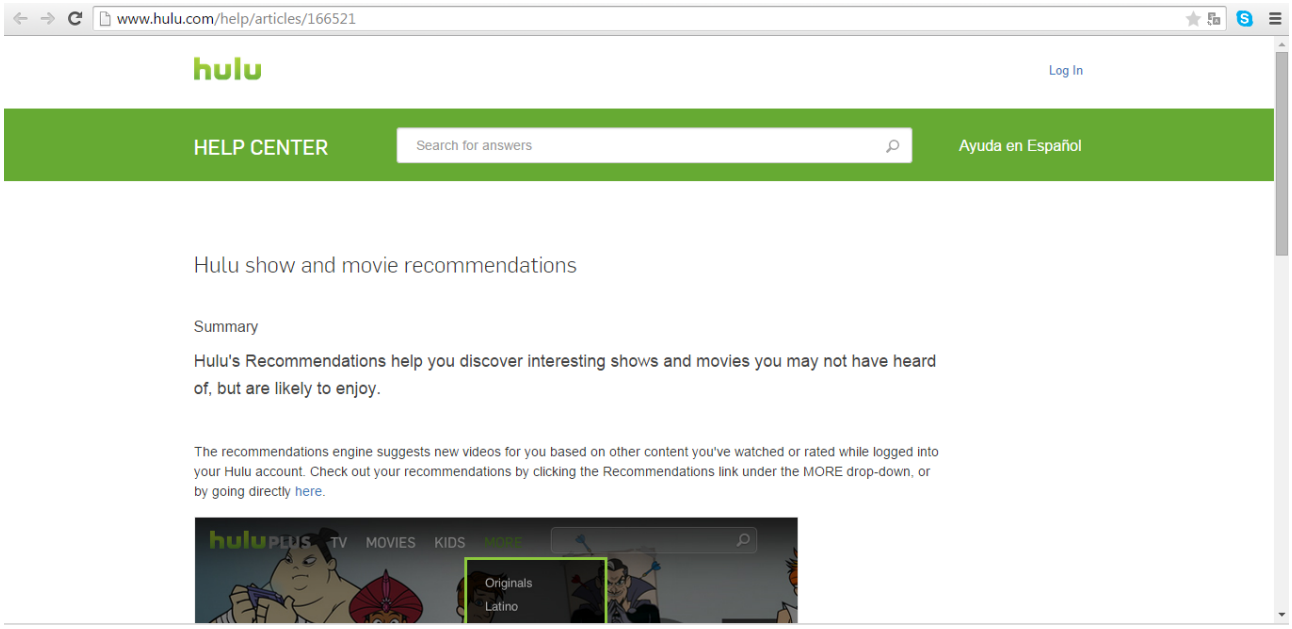
Το σύστημα "last.fm", δημιουργεί το προφίλ του χρήστη, συγκεντρώνοντας και αναλύοντας πληροφορίες σχετικά με τραγούδια που προτιμά να ακούει ο κάθε χρήστης, είτε στο διαδίκτυο ή σε φορητές συσκευές στις οποίες έχει εγκατασταθεί λογισμικό για τη καταγραφή των προτιμήσεων του χρήστη. Η διαδικασία αυτή ονομάζεται "scrobbling", και στη last.fm έχουν καταγραφεί πάνω από 50 εκατομμύρια "scrobbles", δηλαδή καταγεγραμμένα τραγούδια στα προφίλ των χρηστών. Το σύστημα εφαρμόζει την τεχνική με βάση το συνεργατικό φιλτράρισμα. Το σύστημα "last.fm" ανήκει από το 2007 στην εταιρεία CBS Interactive, που δραστηριοποιείται στη παραγωγή και διαχείριση περιεχομένου διασκέδασης (entertainment) στο διαδίκτυο.



Εικόνα 9.9. *Σύστημα συστάσεων στη μουσική "Last.fm".*

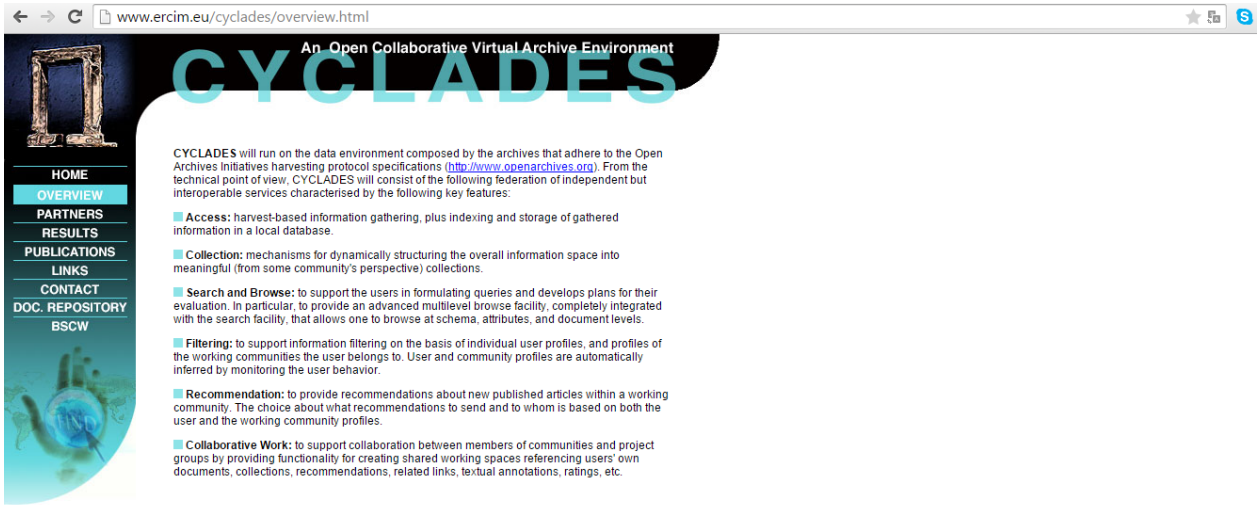
Το *Pandora radio*, εφαρμόζει την τεχνική με βάση το περιεχόμενο, και χρησιμοποιεί πληροφορίες από 400 περίπου χαρακτηριστικά τραγουδιών. Όταν τα τραγούδια παίζουν, οι χρήστες μπορούν να αφήσουν αξιολογήσεις διαμορφώνοντας έτσι το προφίλ του χρήστη.

Στο χώρο του κινηματογραφικών ταινιών βρίσκουμε το www.imdb.com το οποίο είναι συντόμευση του "internet movie database". Είναι ένα παράδειγμα συστήματος με βάση το περιεχόμενο. Έχει οργανώσει πληροφορίες για πάνω από 3 εκατομμύρια ταινιών, έχει πάνω από 60 εκατομμύρια μέλη τα οποία είναι διαμορφωμένα σε 6.5 εκατομμύρια προφίλ. Άλλο παράδειγμα είναι το www.netflix.com, το www.seethisnext.com και το www.hulu.com, τα οποία παρουσιάζει επίσης εξατομικευμένες προτάσεις ταινιών.



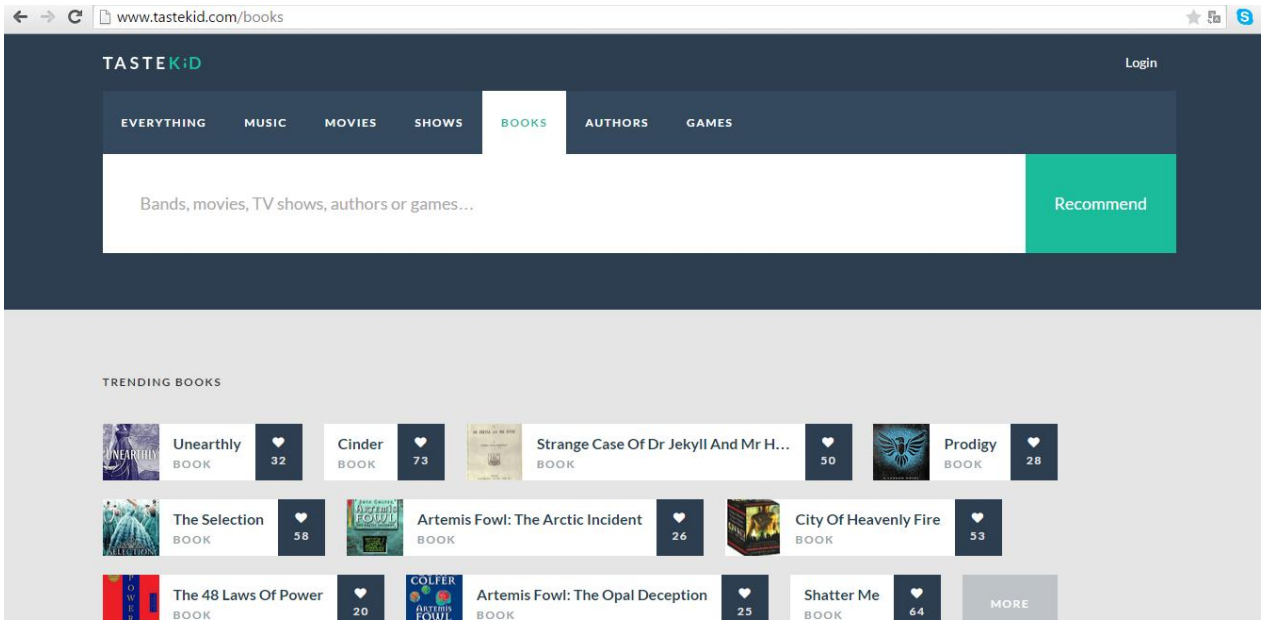
Εικόνα 9.10*. Σύστημα συστάσεων στις κινηματογραφικές ταινίες "hulu.com".*

Το hulu.com το οποίο βοηθά χρήστες να παρακολουθήσουν ταινίες και βίντεο που είναι κοντά στα ενδιαφέροντά τους, αλλά επίσης βοηθά και παραγωγούς βίντεο, κλπ. να «ανακαλυφθούν» από τους χρήστες. Το hulu.com βασίζεται στις κριτικές των χρηστών του για να διαμορφώνει και να εξελίσσει το προφίλ τους. Συστήματα συστάσεων στις ηλεκτρονικές βιβλιοθήκες συναντάμε με το CYCLADES, σύστημα του πανεπιστημίου Stanford, των ΗΠΑ (Lu, et al, 2015).



Εικόνα 9.11. *Σύστημα συστάσεων CYCLADES στις υπηρεσίες ηλεκτρονικών βιβλιοθηκών.*

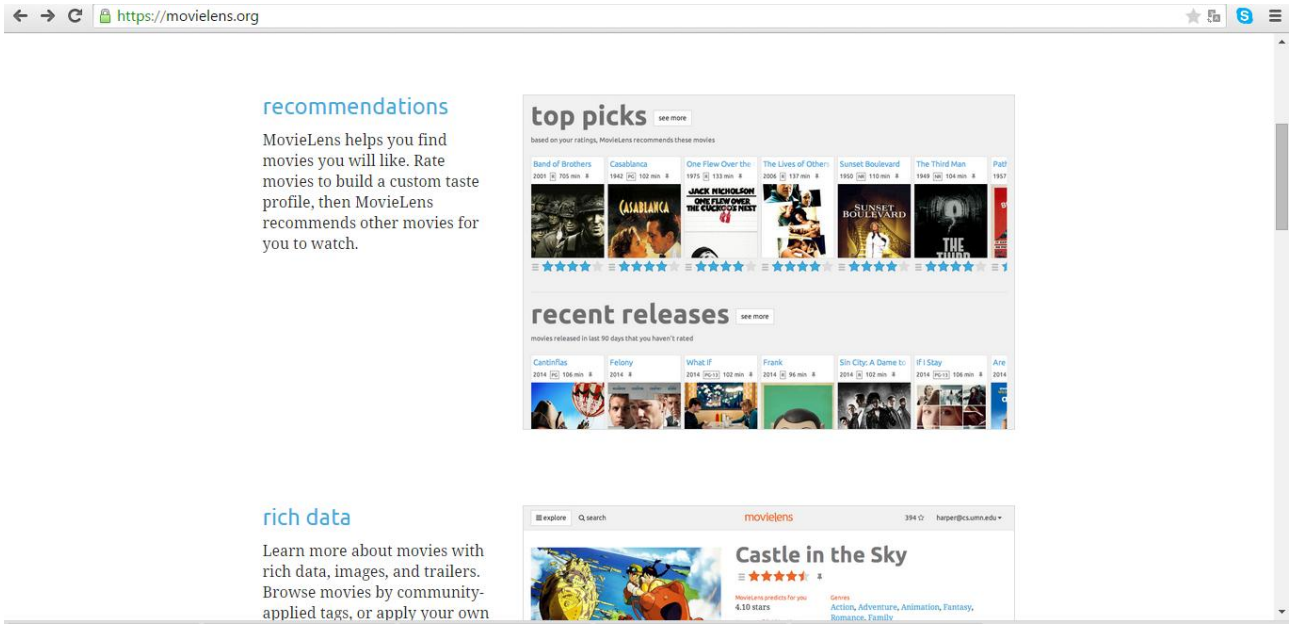
Το σύστημα, χρησιμοποιώντας κριτικές από χρήστες, είναι παράδειγμα ενός υβριδικού συστήματος συστάσεων, εφόσον συνδυάζει και μεθόδους με βάση το περιεχόμενο αλλά και με βάση το συνεργατικό φιλτράρισμα. Ένα σύστημα επίσης υβριδικό, (με βάση το περιεχόμενο αλλά το συνεργατικό φιλτράρισμα) με προτάσεις από ταινίες, μουσική, βιβλία, συγγραφείς, τηλεοπτικά προγράμματα είναι το http://www.tastekid.com/books.



Εικόνα 9.12. *Υβριδικό Σύστημα συστάσεων για ταινίες, βιβλία, κλπ.*

Οι χρήστες αξιολογούν ταινίες, μουσική, βιβλία, κλπ. και με βάση αυτές τις κριτικές το σύστημα δημιουργεί το προφίλ των χρηστών αλλά και των χαρακτηριστικών των υπηρεσιών και προϊόντων που προσφέρει.

Ιδιαίτερο ενδιαφέρον έχει το σύστημα www.movielens.org, το οποίο παρέχει εξατομικευμένες προτάσεις για κινηματογραφικές ταινίες.



Εικόνα 9.13. *Το Σύστημα συστάσεων για ταινίες movielens.*

Πρόκειται για ένα ερευνητικό έργο του πανεπιστημίου της Μινεσότα και το οποίο διαθέτει χιλιάδες δεδομένα με αξιολογήσεις χρηστών που χρησιμοποιούνται όχι μόνο για τη λειτουργία του συστήματος αλλά είναι διαθέσιμα και σε ερευνητές που ασχολούνται με τα συστήματα συστάσεων διεθνώς.

9.7 Συμπεράσματα

Η τάση για εξατομίκευση των υπηρεσιών και των προϊόντων, είναι αποτέλεσμα πολλών παραγόντων όπως του αυξημένου ανταγωνισμού, των πιο απαιτητικών αλλά και καλύτερα ενημερωμένων καταναλωτών, αλλά βεβαίως και της πληθώρας των διαθέσιμων πληροφοριών. Τα συστήματα συστάσεων αναπτύσσονται και απασχολούν την ερευνητική κοινότητα με αυξανόμενο ενδιαφέρον και ένταση.

Σύμφωνα με το τους (Lu, et al, 2015), οι επικρατούσες τεχνικές για την ανάπτυξη συστημάτων συστάσεων είναι οι τεχνικές με βάση το περιεχόμενο, με βάση το συνεργατικό φιλτράρισμα και με βάση τη γνώση. Υπάρχουν αρκετοί τομείς εφαρμογής, οι οποίοι παρουσιάζονται στον πιο κάτω πίνακα, μαζί με τις τεχνικές που χρησιμοποιούν (Lu, et al, 2015).

Χώρος Εφαρμογής	**Με Βάση το Περιεχόμενο**	**Με βάση το Συνεργατικό Φιλτράρισμα**	**Με βάση τη Γνώση**
Ηλεκτρονική Διακυβέρνηση	1	5	1
Ηλεκτρονικό Επιχειρείν		1	3
Ηλεκτρονικές Αγορές	3	1	
Ηλεκτρονικός Τουρισμός	5	9	9
Ηλεκτρονικές Βιβλιοθήκες	2	2	
Ηλεκτρονική Μάθηση	2		11

Πίνακας 9.3. *Περιοχές εφαρμογής και τεχνικές στα συστήματα συστάσεων*

Με την προβλεπόμενη μεγάλη ανάπτυξη στο χώρο των Big Data, η διαθεσιμότητα των δεδομένων θα αυξάνει διαρκώς, προσφέροντας μεγάλες ευκαιρίες δημιουργίας καινοτομικών εφαρμογών μέσα από την τεχνολογία και τη ανάλυση δεδομένων. Η ταυτόχρονη αύξηση των χρηστών του διαδικτύου, η ανάπτυξη του ηλεκτρονικού επιχειρείν και των ηλεκτρονικών υπηρεσιών, δημιουργούν τις κατάλληλες συνθήκες για την δημιουργία συστημάτων συστάσεων. Νέες τεχνικές και τεχνολογίες πρέπει να αναπτυχθούν και να εφαρμοστούν ώστε να αντιμετωπιστούν τα προβλήματα με τα συστήματα συστάσεων και να τα καταστήσουν αξιόπιστα παρέχοντας πολύτιμες υπηρεσίες στους χρήστες τους.

Βιβλιογραφία/Αναφορές

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp. 734–749.

Asnicar, F., & Tasso, C. (1997). *ifWeb: A prototype of user model-based intelligent agent for document filtering and navigation in the world wide web*. Proceedings of Workshop Adaptive Systems and User Modeling on the World Wide Web, June 2-5, Chia Laguna, Sardinia, Italy, pp: 3-12.

Ansari, A., S. Essegaier, S., Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, 37(3) pp. 363–375.

Apache Hadoop (n.d.). Retrieved 10 April 2015 from https://hadoop.apache.org/

Bunt, A., Carenini, G., Conati, C. (2007). Adaptive content presentation for the web, in: *Lecture notes in computer science*, Vol. 4321, Springer, Berlin/Heidelberg, pp. 409-432.

Burke, R. (2000). Knowledge-based Recommender Systems. In: A. Kent (ed.): *Encyclopedia of Library and Information Systems*, 69, Sup. 32. Publ. Marcel Dekker.

Burke, R., Mobasher, B., Bhaumik, R., Williams, C. (2005). *Segment-based injection attacks against collaborative filtering recommender systems*. In ICDM '05: Proceedings of the fifth IEEE international conference on data mining, pp. 577–580. Washington, DC: IEEE Computer Society. Houston, Texas.

Chen, L. & Sycara, K. (1998). *WebMate: a personal agent for browsing and searching*. In Proceedings of the second international conference on Autonomous agents (AGENTS '98), Katia P. Sycara and Michael Wooldridge (Eds.). ACM, New York, NY, USA, pp. 132-139.

Gauch, S., Chaffeeb, J., Pretschnerc, A. (2003). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems: An international journal*, 1, pp. 219–234.

Gentili, G., Micarelli, A., Sciarrone, F. (2003). InfoWeb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17, pp. 715-744.

Haase, P., Stojanovic, N., Sure, Y., Volker, J. (2004). *Bibster - a semantics-based bibliographic peer-to-peer system*. In Proc. of the Third Int. Semantic Web Conference, Hiroshima, Japan, NOV.

Harpale, A. S., & Yang, Y. (2008). *Personalized active learning for collaborative filtering*. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, Singapore, pp. 91–98. New York: ACM.

Jin, R., & Si, L. (2004). *A Bayesian approach toward active learning for collaborative filtering*. In UAI '04: Proceedings of the 20th conference on uncertainty in artificial intelligence, Banff, Canada, pp. 278–285. Arlington: AUAI Press.

Kardaras D., Mamakou X., and V. Karakostas (2011). *Adaptive web site design based on fuzzy user profiles, usability rules and design patterns*. Proceedings of the 1st European Workshop on HCI Design and Evaluation: The influence of domains, Cyprus University of Technology, the European University Cyprus in collaboration with SIGCHI Cyprus, 8 April, Limassol Cyprus, Printed by: IRIT Press, Toulouse, France, ISBN: 978-2-917490-13-6

Kardaras, D.K., & Karakostas, B. (2012). *Service customization using web technologies*, IGI Global, USA, ISBN 13: 978-1466616042.

Kardaras D.K., Karakostas B., Mamakou X. (2013). Content presentation personalisation and media adaptation in tourism web sites using Fuzzy Delphi Method and Fuzzy Cognitive Maps, *Expert Systems with Applications*, 40(6), pp. 2331-2342.

Lam, S. K., & Riedl, J. (2004). *Shilling recommender systems for fun and profit*. In WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, pp. 393–402, New York: ACM.

Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G. (2015). Recommender systems application developments: A survey. *Decision Support Systems*, 74, pp. 12-32.

Melville, P., Mooney, R. J., & Nagarajan, R. (2002). *Content-boosted collaborative filtering for improved recommendations*. In Proceedings of the eighteenth national conference on artificial intelligence (AAAI-02), Edmonton, Alberta, pp. 187–192.

Miao, C., Yangb, Q., Fangc, H., Goha, A. (2007). A cognitive approach for agent-based personalized recommendation, *Knowledge-Based Systems*, 20, pp. 397-405.

Micarelli, A., & Sciarrone, F. (2004). *Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System*. User Modeling and User-Adapted Interaction 14, June 2-3, pp. 159-200.

Mooney, R. J., & Roy, L. (June 2000). *Content-based book recommending using learning for text categorization*. In Proceedings of the fifth ACM conference on digital libraries, San Antonio, Texas, pp. 195–204.

Ricci, F., Rokach, L., Shapira, B., Kantor, P. (2011). *Recommender Systems Handbook*. Springer Publ. ISBN 978-0-387-85819-7.

Schein, A. I., Popescul, A., Ungar, L. H., Pennock, D. M. (2002). *Methods and metrics for cold-start recommendations.* In SIGIR ’02: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pp. 253–260, New York: ACM. Tampere, Finland.

Tanudjaja, F., & Mui, L. (2002). *Persona: A Contextualized and Personalized Web Search*. Hawaii International Conference on System Sciences, p. 67, 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 3.

Xiao, B., & Benbasat, I. (2007). E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact, *MIS Quarterly*, 31(1), pp. 137-209.