ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηαμτοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

Εισαγωγή

Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός μοντέλου πρόβλεψης περιλαμβάνουν **κατηγοριοποίηση, παλινδρόμηση, ανάλυση χρονολογικών σειρών και πρόβλεψη**.

Εισαγωγή

Οι βασικότερες από τις μεθόδους της εξόρυξης δεδομένων

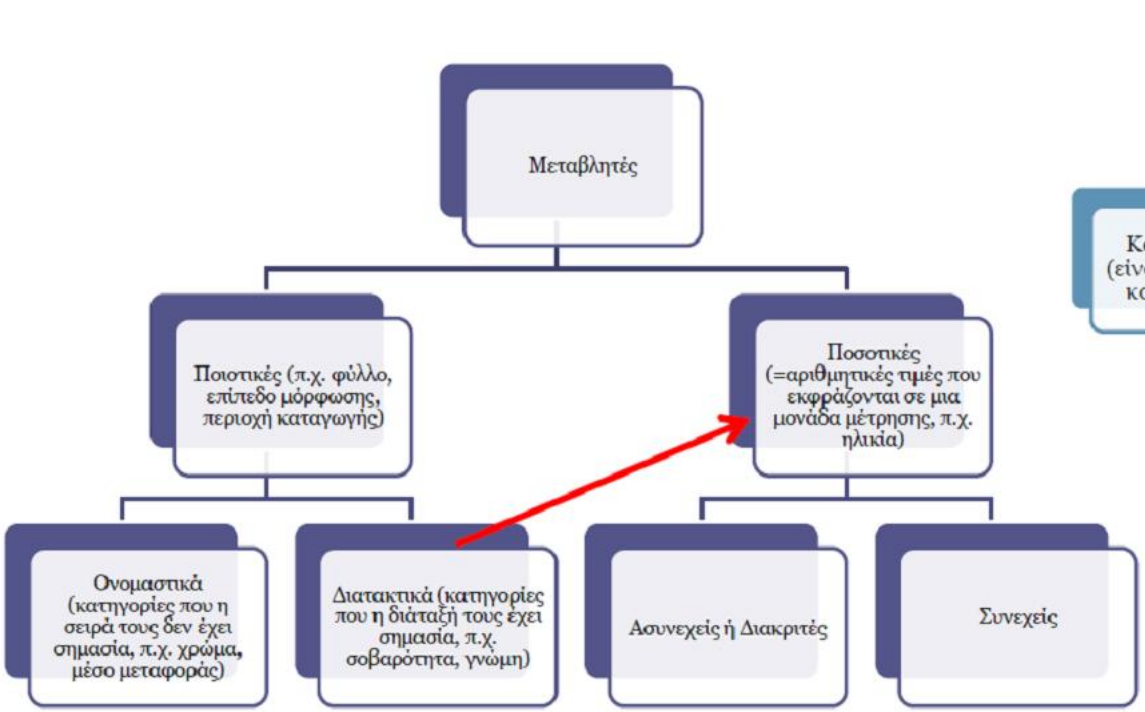
Εισαγωγή

Προεπεξεργασία

Χαμένες Τιμές
Θορυβώδη Δεδομένα
Κανονικοποίηση
Κατασκευή νέων πεδίων
Μείωση Διαστάσεων (δηλ. στηλών) καιΕπιλογή Χαρακτηριστικών

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Κατηγορίες Μεταβλητών



ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Α. Καθαρισμός δεδομένων (Data cleaning)
Συμπλήρωση των χαμένων τιμών, απαλοιφή θορύβου, απομάκρυνση των outliers, διόρθωση ασυνεπειών, απαλοιφή πλεονασμού

Β. Ενοποίηση δεδομένων (Data integration)
Ενοποίηση πολλαπλών βάσεων δεδομένων, κύβων δεδομένων ή αρχείων, απαλοιφή πλεονασμού

Γ. Μετασχηματισμός δεδομένων (Data transformation) και Διακριτοποίηση δεδομένων (Data discretization)
Κανονικοποίηση, Μετατροπή των numerical τιμών σε nominal

Δ. Μείωση δεδομένων (Data reduction)
Μείωση διαστατικότητας, μείωση πληθυκότητας, συμπίεση δεδομένων

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Κατηγοριοποίηση (Classification)

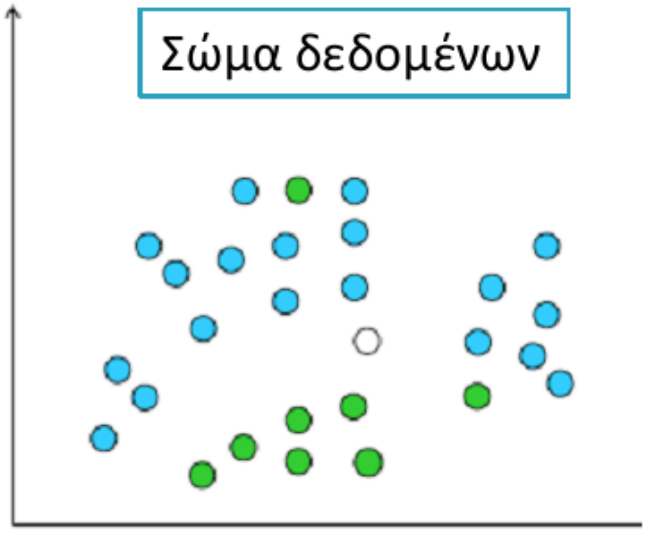
Η κατηγοριοποίηση απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες ή κατηγορίες-κλάσεις. Αναφέρεται συχνά σαν εποπτευμένη μάθηση, επειδή οι κατηγορίες-κλάσεις καθορίζονται πριν ακόμη εξεταστούν τα δεδομένα. Οι αλγόριθμοι κατηγοριοποίησης απαιτούν οι κατηγορίες να ορίζονται με βάση τις τιμές τω ν γνωρισμάτων των δεδομένων. Συχνά περιγράφουν αυτές τις κατηγορίες κοιτάζοντας τα χαρακτηριστικά δεδομένων που είναι ήδη γνωστό ότι ανήκουν στις κατηγορίες.
Η αναγνώριση προτύπου (Pattern recognition) αποτελεί ένα είδος κατηγοριοποίησης, όπου ένα πρότυπο εισόδου κατηγοριοποιείται σε μία από διάφορες κατηγορίες, με βάση την εγγύτητα του ως προς αυτές τις προκαθορισμένες κατηγορίες

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

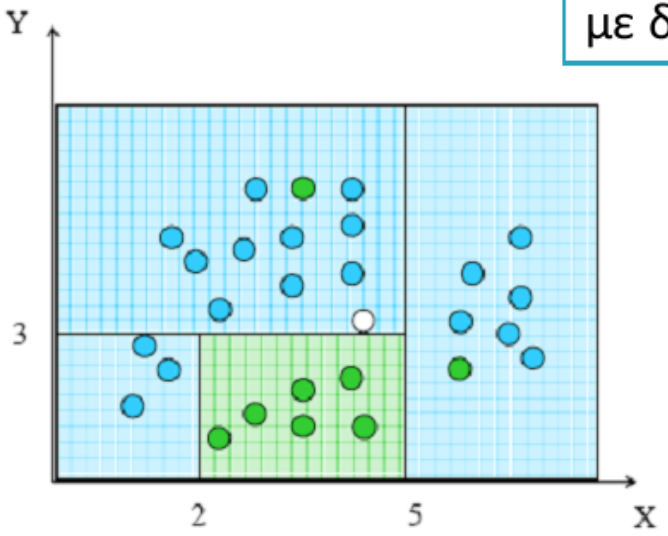
Στόχος είναι η δημιουργία ενός μοντέλου – κατηγοριοποιητή (classifier) με βάση τα υπάρχοντα δεδομένα Ουσιαστικά, είναι η μάθηση μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο (συνήθως αναπαρίσταται ως ένα διάνυσμα τιμών για τις χαρακτηριστικές του ιδιότητες) σε μία τιμή μιας κατηγορικής μεταβλητής, η οποία η οποία είναι γνωστή και ως κλάση (ή κατηγορία). Στην κατηγοριοποίηση, το αποτέλεσμα που θέλουμε να προβλέψουμε είναι η κλάση των δειγμάτων. Ή κλάση μπορεί να πάρει διακριτές τιμές από ένα πεπερασμένο σύνολο. Αντίθετα, κατά την πρόβλεψη με χρήση τεχνικών όπως η παλινδρόμηση, η μεταβλητή-στόχος μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός

Για το σκοπό αυτό, εφαρμόζονται τεχνικές, τις οποίες κατατάσσουμε, σε δυο βασικές κατηγορίες που είναι τα δέντρα απόφασης και τα Νευρωνικά δίκτυα. Επιπλέον, υπάρχουν άλλες δυο κατηγορίες, που είναι οι στατιστικοί αλγόριθμοι και οι αλγόριθμοι απόστασης

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ



Μοντέλο κατηγοριοποίησης με δέντρα αποφάσεων



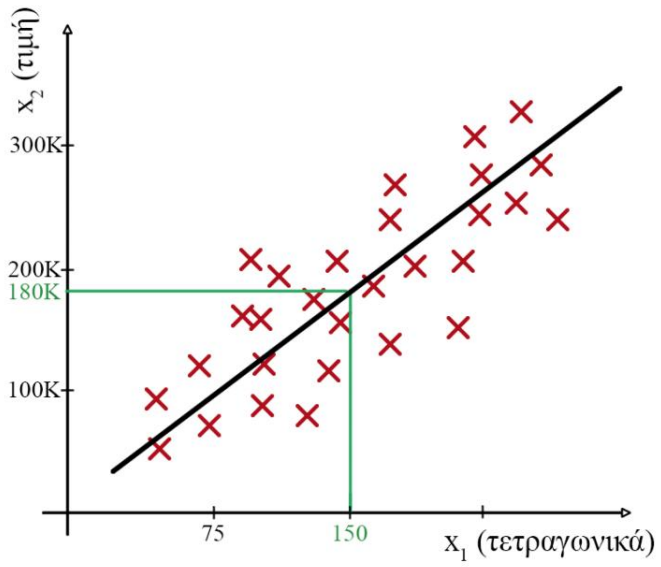
if X > 5 then **blue**
else if Y > 3 then **blue**
else if X > 2 then **green**
else **blue**

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Παλινδρόμηση (Regression)

Μια σχετική διαδικασία με την κατηγοριοποίηση είναι η παλινδρόμηση, στόχος της οποίας είναι η μάθηση ή αλλιώς η εκπαίδευση (training) μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή. Πρόκειται για μια, επίσης, προγνωστική μέθοδο. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable).
Η παλινδρόμηση (Regression), εφαρμόζεται κυρίως στην στατιστική και στα Νευρωνικά δίκτυα, όπου προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο δυνατό λάθος.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ



Οι μεταβλητές είναι τα τετραγωνικά ενός σπιτιού και η τιμή πώλησης του σε χιλιάδες Ευρώ. Ή γραμμική παλινδρόμηση προσαρμόζει μια ευθεία στα δείγματα του συνόλου δεδομένων, τα οποία σηματοδοτούνται με κόκκινο Χ. Ή προσαρμογή γίνεται με βάση μια συνάρτηση απόστασης ή συνάρτηση κόστους, την τιμή της οποία θέλουμε να ελαχιστοποιήσουμε.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η παλινδρόμηση, χρησιμοποιώντας μια βάση αριθμητικών δεδομένων, αναπτύσσει μια μαθηματική σχέση, η οποία ταιριάζει στα δεδομένα αυτά. Στην συνέχεια, η μαθηματική αυτή σχέση χρησιμοποιείται για την πρόβλεψη κάποιας μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτήν νέα αριθμητικά δεδομένα.

Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται πιο αποτελεσματικά, όταν έχουμε συνεχή ποσοτικά δεδομένα (βάρος, ταχύτητα ή ηλικία). Ενώ,δεν είναι τόσο αποτελεσματική με κατηγορικά (δεδομένα που προέρχονται από μμεταβλητές οι τιμές των οποίων εκφράζουν τάξεις ή κατηγορίες) δεδομένα

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Πρόβλεψη (Prediction)

Η πρόβλεψη μπορεί να θεωρηθεί ως ένα είδος κατηγοριοποίησης. Η διαφορά είναι ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μια μελλοντική κατάσταση παρά σε μια τρέχουσα.

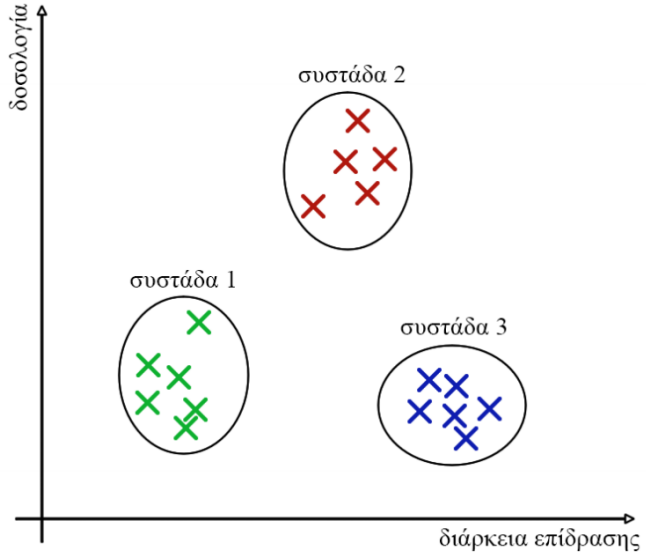
ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι η εργασία του μερισμού ενός συνόλου δεδομένων σε ομάδες ομοίων στοιχείων,
clusters. Τα δεδομένα ομαδοποιούνται σε σύνολα με βάση κάποιο κριτήριο ομοιότητας. Η συσταδοποίηση δεν βασίζεται σε προκαθορισμένες κλάσεις.

Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα. Ουσιαστικά αναζητείται ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, για να περιγράψει τα δεδομένα.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ



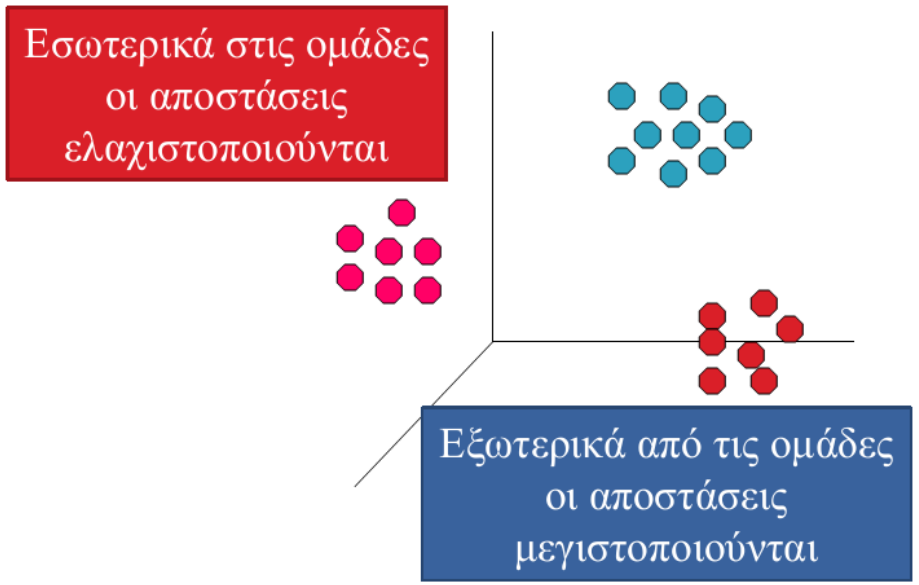
Στο παράδειγμα βλέπουμε το αποτέλεσμα συσταδοποίησης φαρμακευτικών δεδομένων. Έχουν δημιουργηθεί 3 συστάδες με βάση τα χαρακτηριστικά «δοσολογία» και «διάρκεια επίδρασης».

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

•Τα δεδομένα μιας ομάδας να είναι πιο όμοια μεταξύ τους
•Τα δεδομένα ξεχωριστών ομάδων να είναι λιγότερο όμοια μεταξύ τους

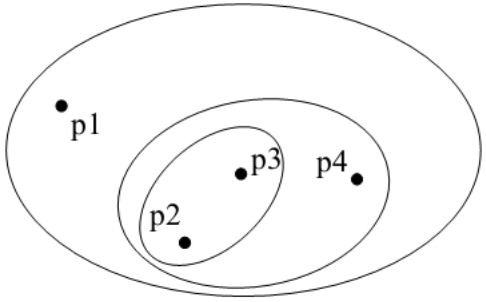
Μέτρα ομοιότητας

• Ευκλείδεια απόσταση, αν οι ιδιότητες είναι συνεχείς
•Άλλα μέτρα, εξαρτώμενα της εφαρμογής

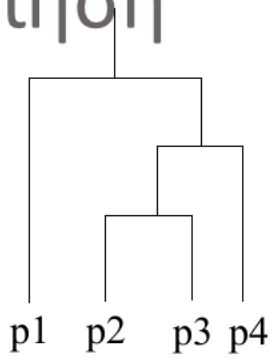


ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

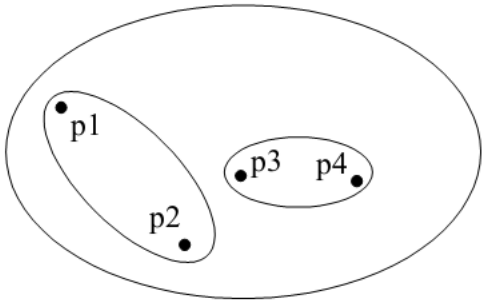
Ιεραρχική Συσταδοποίηση



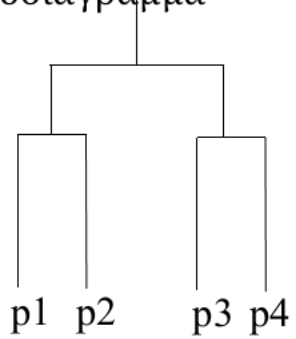
Παραδοσιακή Ιεραρχική Συσταδοποίηση



Παραδοσιακό Δενδροδιάγραμμα



Μη παραδοσιακή Ιεραρχική Συσταδοποίηση



Μη παραδοσιακό Δενδροδιάγραμμα

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

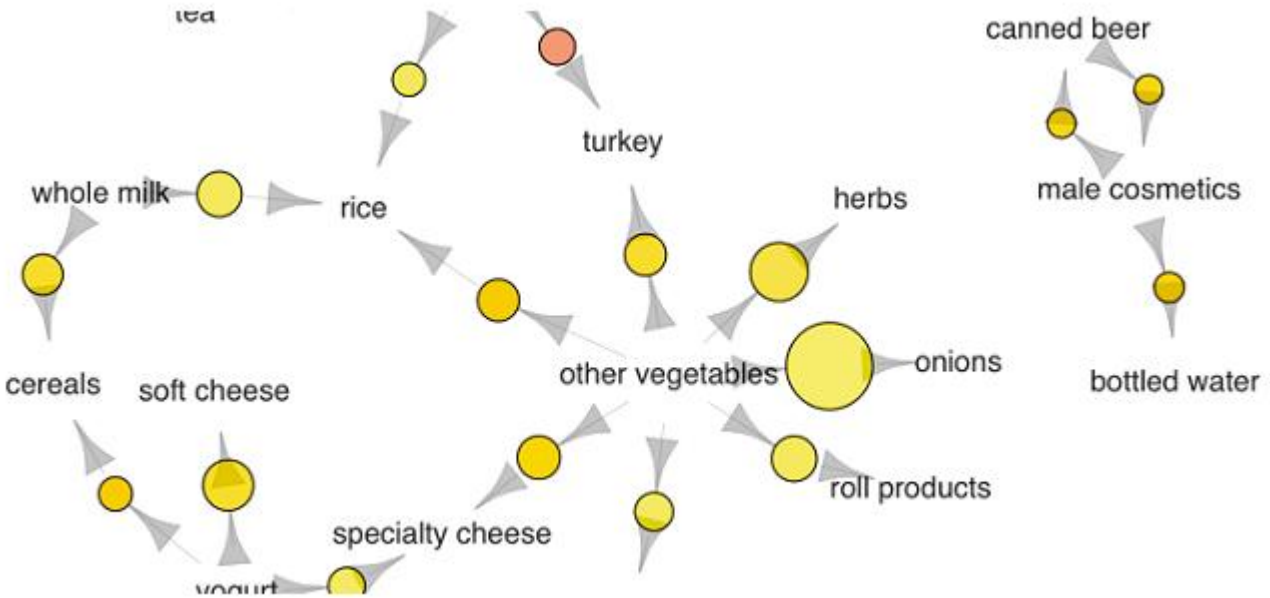
Κανόνες συσχέτισης (Association)

Η ανάλυση συσχέτισης αναφέρεται στη διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Έχει προσελκύσει ιδιαίτερο ενδιαφέρον, καθώς οι κανόνες συσχέτισης παρέχουν έναν
συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές
από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνω-ρισμάτων ενός συνόλου των δεδομένων.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Αυτοί οι συσχετισμοί παρουσιάζονται στη μορφή A →B, όπου τα A και B αποτελούν σύνολα που αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων που αναλύουμε. Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης A →B προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ



ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

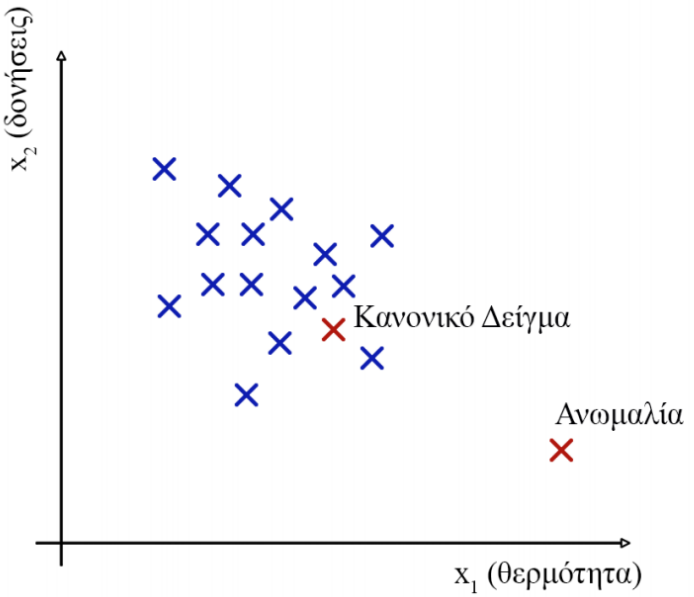
Οπτικοποίηση

Ή οπτικοποίηση των δεδομένων συχνά βοηθάει στην καλύτερη κατανόηση όχι μόνο των ίδιων των δεδομένων, αλλά και των συσχετίσεων που μπορεί να υπάρχουν μεταξύ τους. Ωστόσο, οπτικοποίηση μπορεί να γίνει μόνο για συγκεκριμένο αριθμό διαστάσεων. Αυτό σημαίνει ότι για σύνολα δεδομένων με πολλά χαρακτηριστικά, η οπτικοποίηση τους είναι ανέφικτη ή εναλλακτικά αρκούμαστε στην οπτικοποίηση ενός μικρού μέρους αυτών. Σε κάθε περίπτω-ση, οι οπτικοποιήσεις θα πρέπει να συνοδεύονται και από τους αντίστοιχους στατιστικούς ελέγχους, προκει-μένου να βεβαιωθούμε για την εγκυρότητα των συσχετίσεων που απεικονίζονται.

ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Ανίχνευση Ανωμαλιών

Ή ανίχνευση ανωμαλιών εστιάζει στην ανακάλυψη αποκλίσεων στα δεδομένα σε σχέση με αντίστοιχα δεδομένα, τα οποία έχουν συλλεχθεί στο παρελθόν ή με τυπικές τιμές των δεδομένων αυτών. Παρουσιάζεται ένα τέτοιο παράδειγμα, στο οποίο με κόκκινο φαίνονται ένα κανονικό δείγμα, κοντά στα υπόλοιπα με φυσιολογικές τιμές δείγματα, και ένα ανώμαλο δείγμα, του οποίου η τιμή απέχει αρκετά από τα υπόλοιπα.



ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η Ανίχνευση ανωμαλίας είναι μια τεχνική, η οποία χρησιμοποιείται, κυρίως, για τον εντοπισμό ασυνήθιστων μοτίβων, τα οποία δεν συμμορφώνονται με την αναμενόμενη συμπεριφορά και ονομάζονται "outliers". Οι βασικές κατηγορίες των ανωμαλιών, είναι παρακάτω τρεις:

Ανωμαλίες Σημείων (Point Anomalies): Μια μοναδική εμφάνιση δεδομένων είναι ανώμαλη εάν είναι πολύ μακριά από τα υπόλοιπα.

Συγκεντρωτικές Ανωμαλίες (Collective Anomalies): Η ανωμαλία είναι συγκεκριμένη για το περιβάλλον. Αυτός ο τύπος ανωμαλίας είναι κοινός στα δεδομένα χρονοσειρών

Συλλογικές Ανωμαλίες (Contextual Anomalies): Ένα σύνολο υποθέσεων δεδομένων συλλογικά βοηθά στην ανίχνευση ανωμαλιών.