ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

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

2022

Εισαγωγή

Σκοπός

Εισαγωγή στα μεγάλα δεδομένα

Εισαγωγή στην εξόρυξη δεδομένων

Χαρακτηριστικά μεγάλων δεδομένων

Ανάλυση μεθόδων επεξεργασίας μεγάλων δεδομένων

- Κατηγοριοποίηση
- Συσταδοποίηση
- Παλινδρόμηση
- Κανόνες συσχέτισης

Εφαρμογές στο πακέτα WEKA

Εισαγωγή

Ταξινόμηση

- δέντρα αποφάσεων.
- Ταξινομητές βασισμένοι σε κανόνα.
- Ταξινομητές κλασσικοί,
- Bayesian και νευρωνικοί.
- Ταξινομητές πλησιέστερων γειτόνων.
- Ανάλυση ομοιότητα, απόσταση,
- Χαρακτηριστικά αλγορίθμων ομαδοποίησης.
- Ιεραρχική ομαδοποίηση

Εισαγωγή

Διαδικασία Εξετάσεων:

3 ΕΡΓΑΣΙΕΣ

1. Από ομάδα βιβλίων σχετικών διαλέγετε 1 βιβλίο και από αυτό αναλύετε 1 κεφάλαιο. Πλήρη ανάλυση του κεφαλαίου (όχι πιστή μετάφραση) – 3 Μ
2. Πλήρης ανάλυση ένας θέματος από τις παρουσιάσεις του μαθήματος σε μορφή διπλωματικής – 4 Μ
3. Πλήρη ανάλυση εφαρμογής στο πακέτο WEKA – 3Μ

Παράδοση εργασιών μέχρι την τελευταία μέρα εξέτασης του μαθήματος

Εισαγωγή

- Η εξαγωγή χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων
- **Ουσιαστικά, η εξόρυξη δεδομένων είναι η διαδικασία της αυτόματης ανακάλυψης χρήσιμων πληροφοριών μέσω μεγάλων όγκων δεδομένων και η εξαγωγή παλαιότερων μη αναγνωρισμένων και δυνητικά χρήσιμων πληροφοριών**
- Η εξόρυξη δεδομένων προβλέπει επίσης συμπεριφορές και μελλοντικές τάσεις που βοηθούν τα άτομα να γίνουν πιο προληπτικά και να κάνουν ακριβέστερες αποφάσεις που βασίζονται στις πληροφορίες.

Εισαγωγή

- Η εξόρυξη δεδομένων καθιστά όλη τη διαδικασία διαχείρισης πληροφοριών πιο γρήγορη, ευκολότερη και πιο αποτελεσματική.

- **Ορισμός 1**: Η ανακάλυψη γνώσης σε βάσεις δεδομένων (ΑΓΒΔ) είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα.
- **Ορισμός 2**: Η εξόρυξη γνώσης από δεδομένα είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με τη διαδικασία ΑΓΒΔ

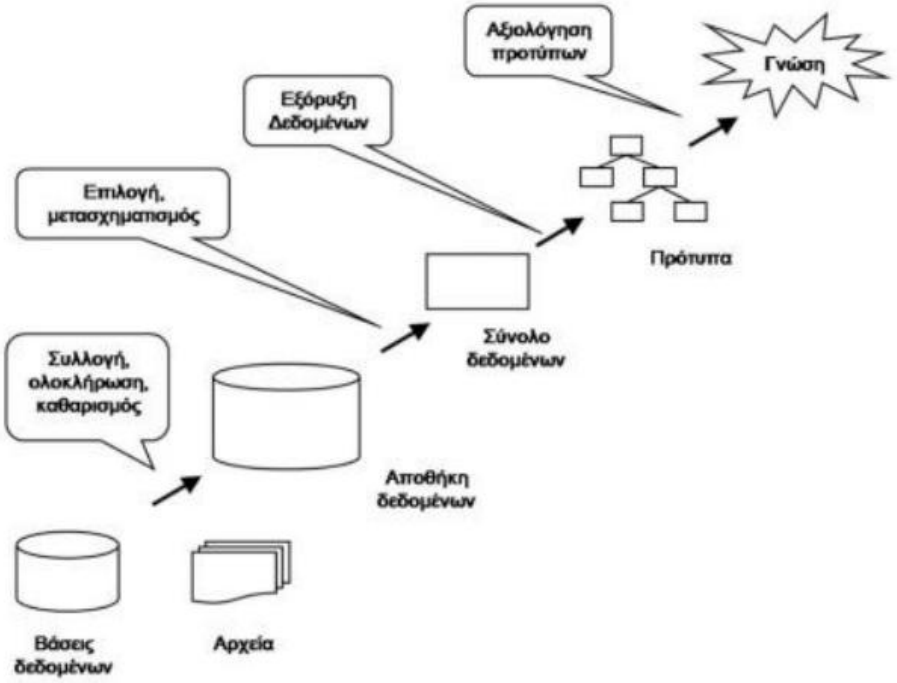
Εισαγωγή

- Η ΑΓΒΔ είναι μια διαδικασία που περιλαμβάνει πολλά διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα, και οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες είναι η έξοδος.
- Η διαδικασία από μόνη της είναι διαδραστική και συνήθως απαιτείται πολύς χρόνος για την ολοκλήρωση της. Για να διασφαλιστεί η χρησιμότητα και η ακρίβεια των αποτελεσμάτων της διαδικασίας, συνήθως χρειάζεται η συνεργασία ειδικών του πεδίου εφαρμογής με ειδικούς της διαδικασίας ΑΓΒΔ καθ'όλη τη διάρκεια της διαδικασίας αυτής.

Εισαγωγή

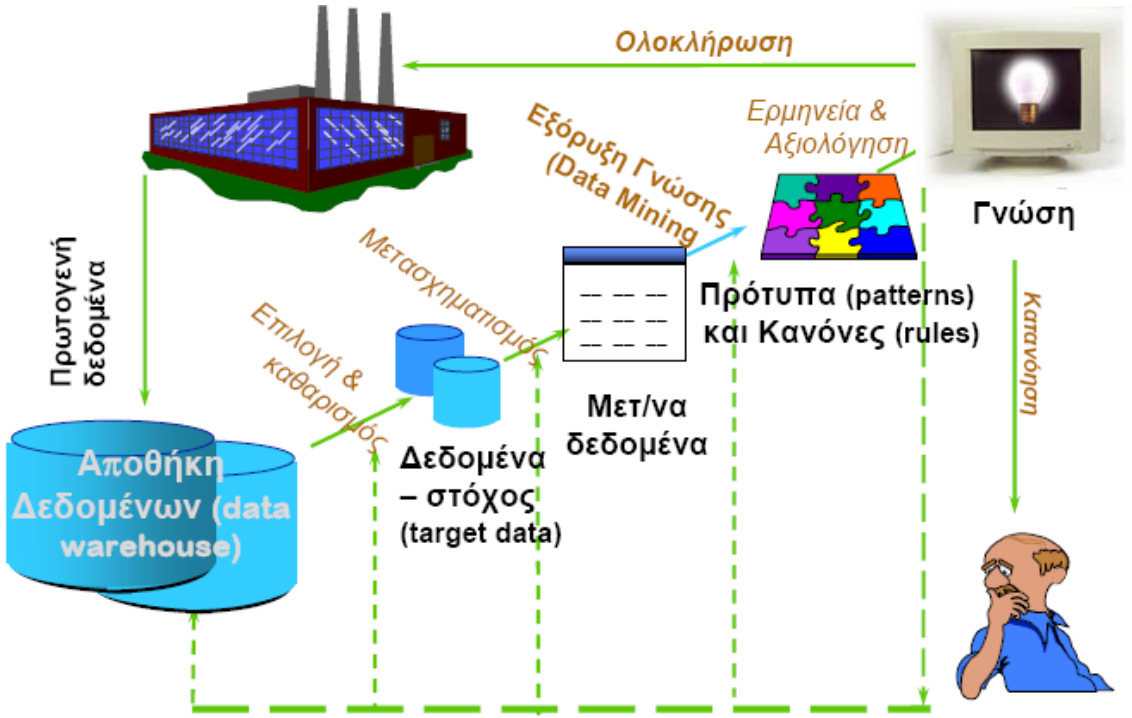
Τα βασικά στάδια της ΑΓΒΔ είναι:

1. Συλλογή Δεδομένων (Data Collection)
2. Προεπεξεργασία Δεδομένων (Preprocessing)
3. Μετασχηματισμός Δεδομένων (Transformation)
4. Εξόρυξη Δεδομένων (Data Mining)
5. Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)



Εισαγωγή

- Πληθώρα ορισμών
  - Μη τετριμμένη εξαγωγή υποκρυπτόμενης, άγνωστης και εν δυνάμει χρήσιμης πληροφορίας από τα δεδομένα
  - Εξερεύνηση και ανάλυση, με αυτόματο ή ημι-αυτόματο τρόπο, μεγάλων ποσοτήτων δεδομένων για την ανακάλυψη χρήσιμων προτύπων



Εισαγωγή

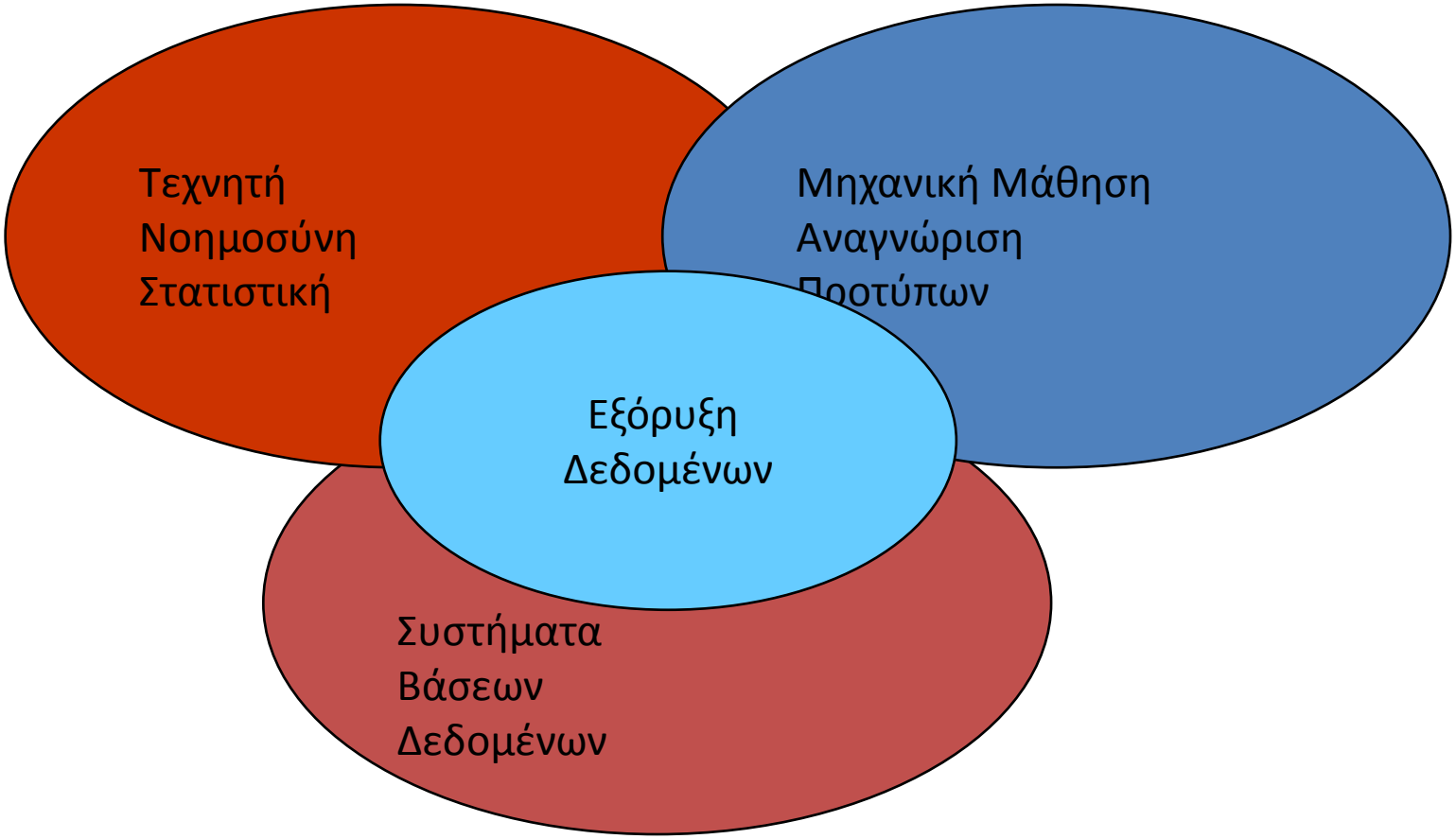
- Εφαρμογές Εξόρυξης Γνώσης

Οι τράπεζες και άλλοι πάροχοι χρηματοπιστωτικών υπηρεσιών μπορούν να εξάγουν δεδομένα σχετικά με τους λογαριασμούς, τις συναλλαγές και τις προτιμήσεις των πελατών τους για να ανταποκριθούν καλύτερα στις ανάγκες τους. Μπορούν επίσης να συλλέξουν δεδομένα από ιστοσελίδες και τις αλληλεπιδράσεις στα κοινωνικά μέσα ενημέρωσης για να αυξήσουν την αφοσίωση των υπαρχόντων πελατών και να προσελκύσουν νέους πελάτες.

Εισαγωγή

- Τα εκπαιδευτικά ιδρύματα μπορούν να επωφεληθούν από την εξόρυξη δεδομένων με την ανάλυση της συλλογής δεδομένων για την πρόβλεψη των μελλοντικών μαθησιακών συμπεριφορών και των επιδόσεων των φοιτητών και στη συνέχεια με τη χρήση αυτών των γνώσεων για τη βελτίωση των μεθόδων διδασκαλίας ή των προγραμμάτων διδασκαλίας.
- Οι πάροχοι υγειονομικής περίθαλψης μπορούν να εξορύσσουν και να αναλύουν δεδομένα για τον προσδιορισμό καλύτερων τρόπων παροχής φροντίδας στους ασθενείς και μείωσης του κόστους. Με τη βοήθεια της εξόρυξης δεδομένων, μπορούν να προβλέψουν πόσους ασθενείς θα χρειαστεί να φροντίσουν και τι τύπο υπηρεσιών θα χρειαστούν οι ασθενείς αυτοί.

Εισαγωγή



Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από τη συλλογή των δεδομένων στην ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά.

Εισαγωγή

Τα βήματα από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

1. **Καθαρισμός δεδομένων (Data cleaning):** Στο βήμα αυτό, αφαιρούνται από τη βάση δεδομένων αυτά που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα.
2. **Ενσωμάτωση δεδομένων (Data integration):** Σε αυτό το βήμα τα δεδομένα που έχουν συλλεχθεί, πολλές φορές ανομοιογενή και από πολλές διαφορετικές πηγές, ενσωματώνονται σε μια κοινή βάση δεδομένων.
3. **Επιλογή δεδομένων (Data selection):** Από όλα εκείνα τα δεδομένα που έχουμε στη διάθεση μας, επιλέγονται προσεκτικά εκείνα που είναι σχετικά και χρήσιμα για την ανάλυση που θα ακολουθήσει.
4. **Τροποποίηση δεδομένων (Data transformation):** Τα δεδομένα που έχουμε επιλέξει δέχονται τις απαραίτητες τροποποιήσεις έτσι ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.
5. **Εξόρυξη δεδομένων (Data mining):** Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.
6. **Αξιολόγηση προτύπων (Pattern evaluation):** Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

Εισαγωγή

Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων, καθώς επίσης την ταξινόμηση και την βελτιστοποίηση των πόρων της.

- **Πρόβλεψη:** Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις. Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

Εισαγωγή

- **Αναγνώριση:** Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.

- **Περιγραφή:** Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

- **Ταξινόμηση:** Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.

Εισαγωγή

- **Βελτιστοποίηση:** Μεταξύ των άλλων σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα.

Εισαγωγή

- Τα "μεγάλα δεδομένα" είναι ένα πεδίο που διαχειρίζεται τους τρόπους ανάλυσης και τη συστηματική απόσπαση πληροφοριών ή διαφορετικά διαχειρίζεται τα πακέτα δεδομένων που είναι πολύ μεγάλα ή περίπλοκα για να τα διαχειριστεί ένα παραδοσιακό λογισμικό εφαρμογών επεξεργασίας δεδομένων.
- Η τρέχουσα χρήση του όρου μεγάλα δεδομένα τείνει να αναφέρεται στη χρήση αναλυτικών στοιχείων πρόβλεψης, αναλύσεων συμπεριφοράς χρηστών ή ορισμένων άλλων προηγμένων μεθόδων ανάλυσης δεδομένων που εξάγουν αξία από δεδομένα και σπανίως σε ένα συγκεκριμένο μέγεθος συνόλου δεδομένων.

Εισαγωγή

- Λόγω του πλήθους των χαρακτηριστικών που διέπουν τα μεγάλα δεδομένα είναι δύσκολη η απεικόνιση του συνόλου των χαρακτηριστικών και εργασία πάνω σε αυτά. Έτσι δημιουργήθηκε η ανάγκη ανάπτυξης αλγορίθμων οι οποίοι βοηθούν την επεξεργασία και την οπτικοποίηση πολυδιάστατων δεδομένων

Εισαγωγή

- Τα μεγάλα δεδομένα συνδέθηκαν αρχικά με τρεις βασικές έννοιες, τον όγκο, την ποικιλία και την ταχύτητα

- Άλλες έννοιες που συνδέθηκαν αργότερα με την έννοια των μεγάλων δεδομένων είναι η εγκυρότητα (δηλαδή το ποσοστό του θορύβου ή της χαοτικής πληροφορίας που υπάρχει στα δεδομένα) και η αξία τους

- Ο όρος χρησιμοποιείται από τη δεκαετία του '90

Εισαγωγή

- Τα μεγάλα δεδομένα περιλαμβάνουν συνήθως σύνολα δεδομένων με μεγέθη πέρα από την ικανότητα των εργαλείων λογισμικού που χρησιμοποιούνται συνήθως για τη συλλογή, επεξεργασία, διαχείριση και επεξεργασία δεδομένων εντός ενός αποδεκτού χρόνου
- Η φιλοσοφία των μεγάλων δεδομένων περιλαμβάνει μη δομημένα, ημιδομημένα και δομημένα δεδομένα, ωστόσο η κύρια εστίαση είναι στα μη δομημένα δεδομένα
- Το μέγεθος των μεγάλων δεδομένων είναι μια ρευστή έννοια, καθώς από το 2012 κυμαίνεται από μερικές δεκάδες terabytes έως πολλά exabytes δεδομένων

Εισαγωγή

- Τα μεγάλα δεδομένα απαιτούν ένα σύνολο τεχνικών και τεχνολογιών με νέες ολοκληρωμένες μορφές που να είναι ικανές να αποκαλύψουν πληροφορίες από σύνολα δεδομένων που είναι ποικίλα, πολύπλοκα και μαζικής κλίμακας

- Το 2016 δόθηκε ένας ορισμός που δήλωνε ότι "**τα μεγάλα δεδομένα αντιπροσωπεύουν τα στοιχεία της πληροφορίας που χαρακτηρίζονται από τόσο μεγάλο όγκο, ταχύτητα και ποικιλία που απαιτούν συγκεκριμένη τεχνολογία και αναλυτικές μεθόδους για τη μετατροπή τους σε αξία**"

Εισαγωγή

- Η Gartner (η μεγαλύτερη επιχείρηση στον κόσμο που ασχολείται με την τεχνολογική έρευνα και συμβουλευτική), το 2012 έδωσε τον εξής ορισμό:
- «Τα big data είναι υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών».
- Στα «μεγάλα δεδομένα» συγκαταλέγονται όλες οι πληροφορίες των social media που είναι προσβάσιμες σε όλους μας και βρίσκονται στο Διαδίκτυο, δηλαδή φωτογραφίες, video και κείμενα, καθώς και όλα τα «κλειστά δεδομένα» των διαφόρων εταιριών αλλά και των κυβερνήσεων
- **η Gartner πρότεινε έναν ορισμό που περιλάμβανε τα "τρία Vs (Volume, Velocity, Variety)": τον όγκο, την ταχύτητα και την ποικιλία**

Εισαγωγή

- Ο ορισμός αυτός έχει επαναληφθεί από τη NIST (Nist Big Dataprogram, 2013) και διευρυνθεί από την IBM (IBM, 2013) για να συμπεριλάβει και ένα τέταρτο V: την πιστότητα (Veracity)
- Η Oracle αποφεύγει την χρήση των Vs για να καταλήξει σε έναν ορισμό. Αντ 'αυτού η Oracle (J.P Dijcks ORACLE, 2013) υποστηρίζει ότι τα μεγάλα στοιχεία είναι η δημιουργία αξίας από παραδοσιακές σχεσιακές βάσεις δεδομένων με στόχο τη λήψη επιχειρηματικών αποφάσεων, η οποία είναι εμπλουτισμένη με νέες πηγές μη-δομημένων δεδομένων.

Εισαγωγή

Χαρακτηριστικά

- “ακαταστασία” των Μεγάλων Δεδομένων
- συσχέτιση προβάλει τη στατιστική σχέση μεταξύ των δεδομένων
- μέσω των συσχετίσεων των Μεγάλων Δεδομένων μπορούμε να κάνουμε προβλέψεις

Εισαγωγή

- Τα τρία Vs (Ο όγκος, η ποικιλία και η ταχύτητα αποτελούν τα βασικά χαρακτηριστικά των μεγάλων δεδομένων, γνωστά και ως τα τρία Vs (Volume, Variety, Velocity), έχουν επεκταθεί περαιτέρω σε άλλα συμπληρωματικά χαρακτηριστικά των μεγάλων δεδομένων

Εισαγωγή


Μεγάλα Δεδομένα
Αξία
Ακρίβεια
Ποικιλομορφία
Ταχύτητα
Όγκος

Εισαγωγή

- Τα μεγάλα δεδομένα μπορούν να περιγραφούν από τα ακόλουθα χαρακτηριστικά

Όγκος

Η ποσότητα των παραγόμενων και αποθηκευμένων δεδομένων. Το μέγεθος των δεδομένων καθορίζει την αξία και τη δυνητική γνώση και αν μπορεί βάσει αυτού να θεωρηθούν μεγάλα δεδομένα ή όχι

Ποικιλομορφία

Ο τύπος και η φύση των δεδομένων. Αυτό βοηθά τα άτομα που τα αναλύουν να χρησιμοποιήσουν αποτελεσματικά την προκύπτουσα γνώση. Τα μεγάλα δεδομένα αντλούνται από κείμενο, εικόνες, ήχο, βίντεο και επιπλέον, τα κομμάτια που λείπουν ολοκληρώνονται μέσω της σύντηξης δεδομένων

Εισαγωγή

- Η **ποικιλομορφία** των μεγάλων δεδομένων, αναφέρεται στο είδος της δομής των δεδομένων. Το είδος της δομής περιλαμβάνει δεδομένα τα οποία είναι δομημένα, μη δομημένα και ήμι-δομημένα. Τα δεδομένα τα οποία είναι οργανωμένα και τα οποία επιπλέον είναι εύκολα προσβάσιμα και διαθέσιμα από έναν υπολογιστή, χαρακτηρίζονται ως δομημένα. Συνήθως αποθηκεύονται σε σχεσιακές βάσεις δεδομένων ή υπολογιστικά φύλλα (όπως αυτά του Excel), δηλαδή είναι περιορισμένα σε συγκεκριμένα πρότυπα, με τιμές σε συγκεκριμένα πεδία.

Εισαγωγή

- **Ταχύτητα**

Η ταχύτητα με την οποία παράγονται και επεξεργάζονται τα δεδομένα για να ανταποκριθούν στις απαιτήσεις και τις προκλήσεις που βρίσκονται στην πορεία της αύξησης και της ανάπτυξής τους. Μεγάλα δεδομένα είναι συχνά διαθέσιμα σε πραγματικό χρόνο. Σε σύγκριση με τα μικρά δεδομένα, τα μεγάλα δεδομένα παράγονται συνεχώς. Τα δύο είδη ταχύτητας που σχετίζονται με τα μεγάλα δεδομένα είναι η συχνότητα παραγωγής και η συχνότητα χειρισμού, καταγραφής και δημοσίευσης

Εισαγωγή

- **Εγκυρότητα**

Είναι ο εκτεταμένος ορισμός για μεγάλα δεδομένα, ο οποίος αναφέρεται στην ποιότητα των δεδομένων και την τιμή των δεδομένων. Η ποιότητα των συλλεγόμενων δεδομένων μπορεί να ποικίλει σημαντικά, επηρεάζοντας την ακριβή ανάλυση τους.Τα δεδομένα πρέπει να υποβάλλονται σε επεξεργασία με προηγμένα εργαλεία (αναλυτικά στοιχεία και αλγόριθμους) για την αποκάλυψη σημαντικών πληροφοριών.

Εισαγωγή

- **Αξιοπιστία (Validity)**

Η αξιοπιστία των μεγάλων δεδομένων αφορά «την ορθότητα και την ακρίβεια των δεδομένων ως προς την προοριζόμενη χρήση τους». Εξαρτάται, δηλαδή, από την προοριζόμενη χρήση, με αποτέλεσμα να είναι αξιόπιστο για κάποια χρήση αλλά πιθανώς να μην είναι αξιόπιστο για μία άλλη.

- **Αστάθεια (Volatility)**

Η αστάθεια των δεδομένων υπεισέρχεται όταν τα δεδομένα ενδέχεται να αλλάξουν μελλοντικά, ή ακόμη και να καταστραφούν. Η αστάθεια αυξάνεται μαζί με τον όγκο, την ποικιλομορφία και την ταχύτητα των δεδομένων. Δεδομένα με αστάθεια μπορεί να είναι για παράδειγμα οι συναλλαγές 10 ετών ενός πελάτη, καθώς αυξάνεται το κόστος αποθήκευσης, ασφάλειας και ανάκτησης, το οποίο οδηγεί στην ανάγκη αντιμετώπισης ή καταστροφής τους μετά από ένα συγκεκριμένο χρονικό περιθώριο.

Εισαγωγή

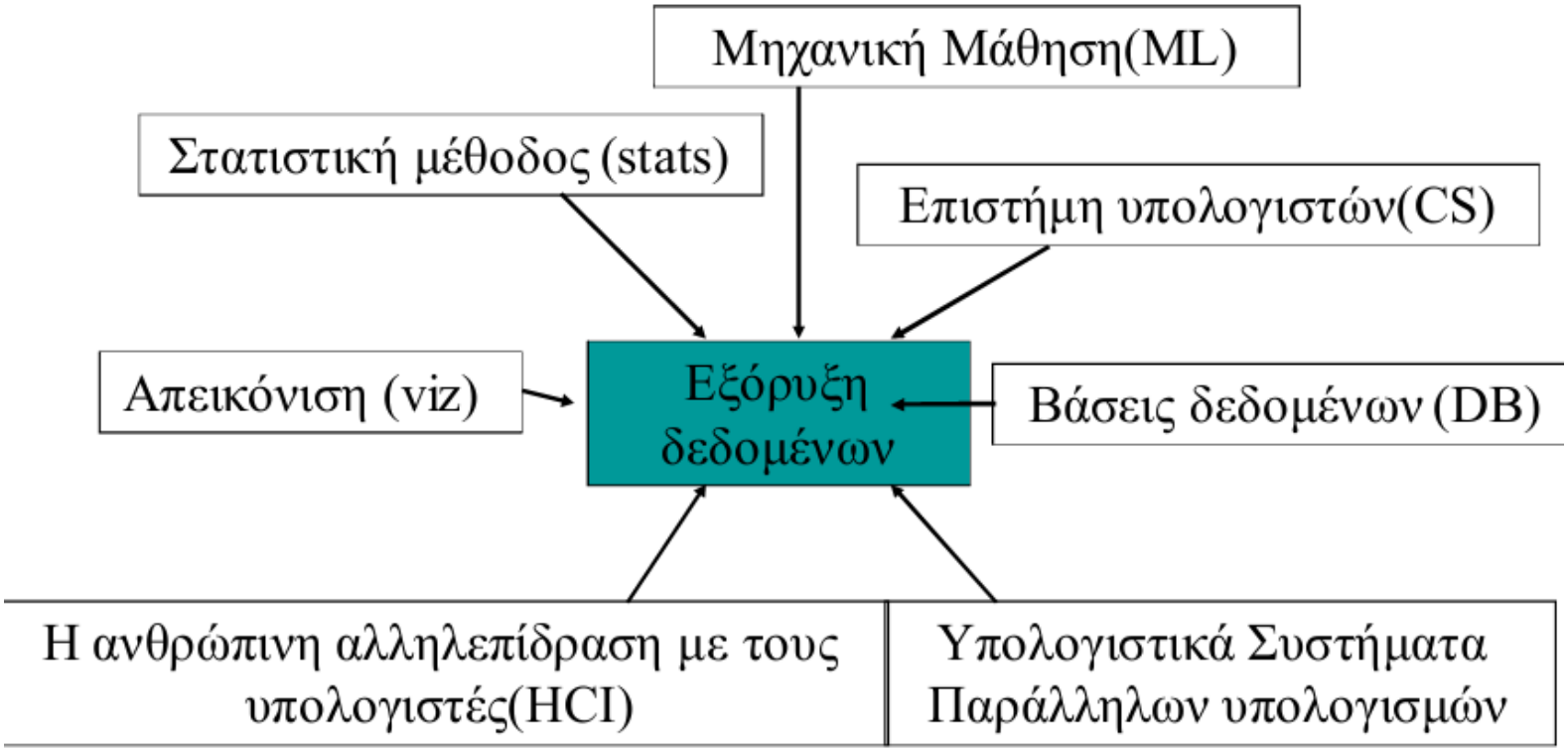
- **Μεταβλητότητα (Variability)**

Η μεταβλητότητα ως χαρακτηριστικό των μεγάλων δεδομένων αναφέρεται στην απροσδόκητη αλλαγή που μπορεί να υποστούν τα χαρακτηριστικά των μεγάλων δεδομένων, όπως για παράδειγμα η δομή ή η ποιότητα. Αυτό συνεπάγεται την ανάγκη για την ευελιξία των επιχειρήσεων σε πιθανές αλλαγές στην επεξεργασία και την αξιοποίηση των μεγάλων δεδομένων λόγω της μεταβλητότητας (Vijendra, 2016).

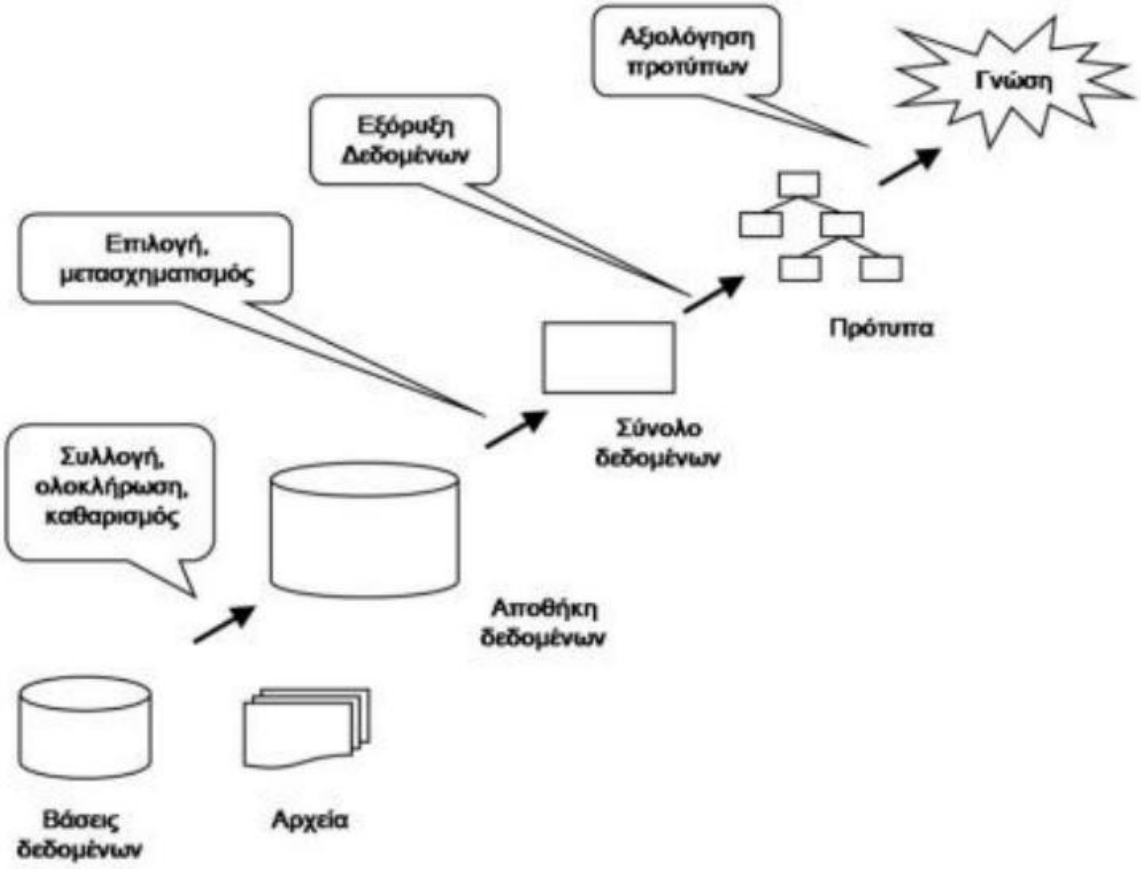
- **Οπτικοποίηση (Visualization)**

Η οπτικοποίηση των δεδομένων αφορά την ικανότητα της αφηρημένης απεικόνισης δεδομένων για την απόφαση ορθών τεχνικών αποθήκευσης, εξαγωγής και μετάδοσης. Αυτά τα δεδομένα μπορεί να προέρχονται από ετερογενείς πηγές και τύπους δεδομένων (Vijendra, 2016).

Εισαγωγή


Μηχανική Μάθηση(ML)
Στατιστική μέθοδος (stats)
Επιστήμη υπολογιστών(CS)
Απεικόνιση (viz)
Εξόρυξη δεδομένων
Βάσεις δεδομένων (DB)
Η ανθρώπινη αλληλεπίδραση με τους υπολογιστές(HCI)
Υπολογιστικά Συστήματα Παράλληλων υπολογισμών

Εισαγωγή


Αξιολόγηση προτύπων
Γνώση
Εξόρυξη Δεδομένων
Επιλογή, μετασχηματισμός
Πρότυπα
Σύνολο δεδομένων
Συλλογή, ολοκλήρωση, καθαρισμός
Αποθήκη δεδομένων
Βάσεις δεδομένων
Αρχεία

Εισαγωγή

Συλλογή Δεδομένων

Το πρώτο βήμα της ΑΓΒΔ είναι η συλλογή και η αποθήκευση των δεδομένων. Η συλλογή των δεδομένων συνήθως γίνεται είτε αυτόματα, π.χ. με χρήση αισθητήρων, είτε μη αυτόματα, π.χ. με χρήση ερωτηματολογίων. Δυσλειτουργία στους αισθητήρες ή αδυναμία απάντησης κάποιας ερώτησης στα ερωτηματολόγια μπορεί να οδηγήσει σε θορυβώδη ή ελλιπή δεδομένα. Τα συγκεκριμένα προβλήματα, που ενδεχομένως να προκύψουν κατά τη συλλογή δεδομένων, αναλαμβάνει να τα αντιμετωπίσει το επόμενο στάδιο.

Εισαγωγή

Προεπεξεργασία Δεδομένων

Τα δεδομένα που πρόκειται να χρησιμοποιηθούν κατά την διαδικασία, ίσως να είναι λανθασμένα ή ελλιπή. Ίσως υπάρχουν ανώμαλα δεδομένα από πολλαπλές πηγές που περιλαμβάνουν διαφορετικούς τύπους δεδομένων και διαφορετικές μονάδες μέτρησης. Σε αυτό το βήμα μπορούν να χρησιμοποιηθούν πολλές και διαφορετικές δραστηριότητες. Τα λανθασμένα δεδομένα μπορεί να διορθωθούν ή να αφαιρεθούν, ενώ τα ελλιπή δεδομένα πρέπει να συλλεχθούν ή να εκτιμηθούν (συχνά χρησιμοποιώντας εργαλεία εξόρυξης γνώσης από δεδομένα).

Εισαγωγή

Μετασχηματισμός Δεδομένων

Ο μετασχηματισμός των δεδομένων αποτελεί το τρίτο στάδιο της ΑΓΒΔ. Ουσιαστικά, πρόκειται για τη μετατροπή των δεδομένων κάτω από ένα κοινό πλαίσιο, για επεξεργασία. Χρησιμοποιείται κυρίως για την εξομάλυνση των δεδομένων και απομάκρυνση θορύβου, για τη συνάθροιση των δεδομένων, δηλαδή για την παραγωγή σύνοψης τους, για την κανονικοποίηση τους, δηλαδή την κλιμάκωση των χαρακτηριστικών του συνόλου δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών, ή τέλος για τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα. Ειδικές μορφές μετασχηματισμού αποτελούν η διακριτοποίηση και η συμπίεση

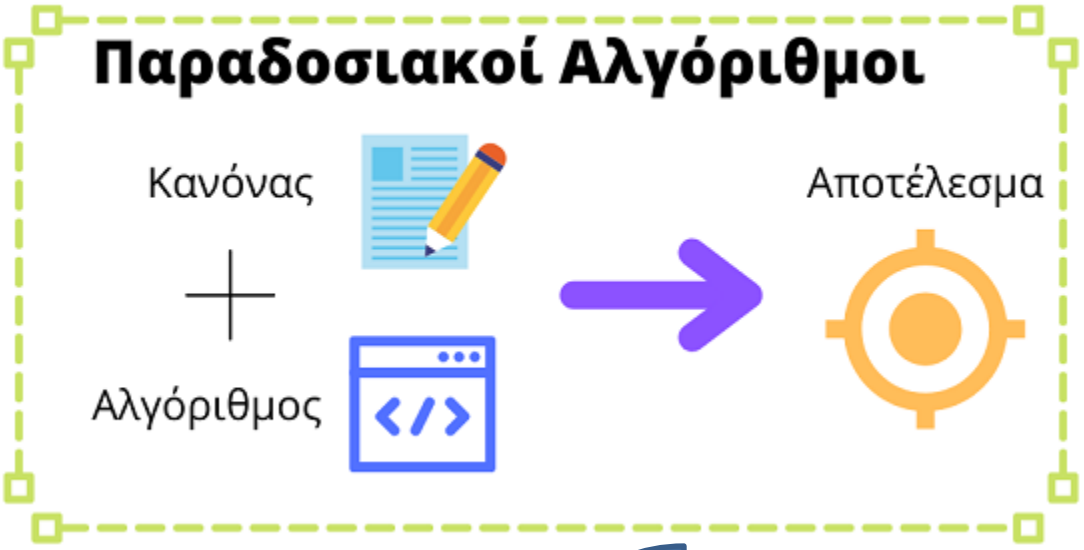
Εισαγωγή

Μηχανική μάθηση

Ο στόχος της μηχανικής μάθησης είναι η εύρεση χρήσιμων αναπαραστάσεων των δεδομένων και η αξιοποίηση αυτών των αναπαραστάσεων για την εξαγωγή συμπερασμάτων σε μελλοντικά δεδομένα, με παρόμοια χαρακτηριστικά

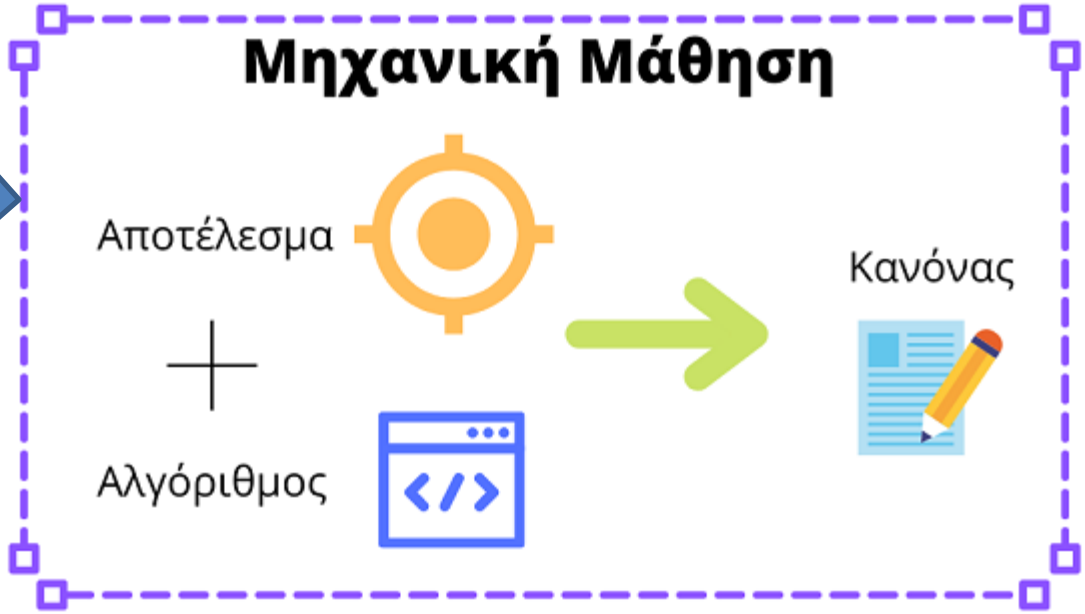
Χρησιμοποιεί στοιχεία από τομείς της στατιστικής, της γραμμικής άλγεβρας, της τεχνητής νοημοσύνης και της επιστήμης των υπολογιστών

Εισαγωγή

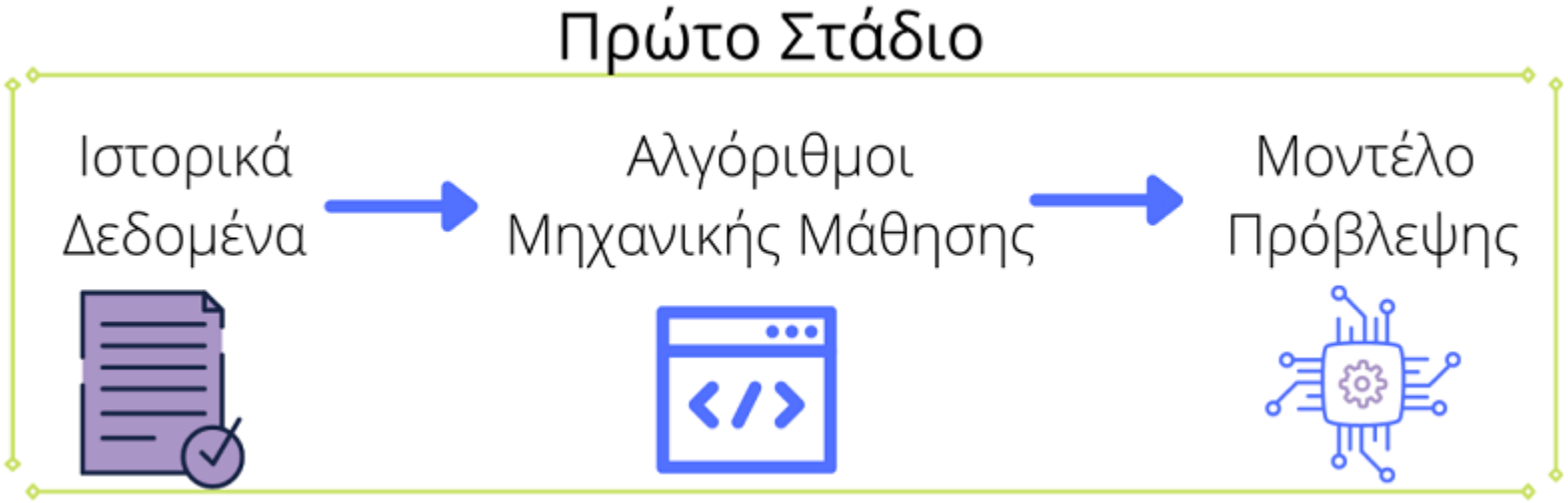


Οι παραδοσιακοί αλγόριθμοι της τεχνητής νοημοσύνης είναι ρητά προγραμματισμένοι. Χρησιμοποιούν κανόνες όπως το εάν και το ίσως (if – else) για να βγάλουν συμπεράσματα.

Οι αλγόριθμοι της μηχανικής μάθησης χρησιμοποιούν πολλά «λυμένα παραδείγματα», ώστε να «μάθουν» τους υποκείμενους κανόνες και σχέσεις που προϋπάρχουν στα δεδομένα, δίχως να τους έχουν προμηθευτεί



Εισαγωγή



Στο πρώτο στάδιο συλλέγονται τα ιστορικά δεδομένα, τα οποία αξιοποιούνται στον αλγόριθμο της μηχανικής μάθησης, ο οποίος στο πέρας της μάθησης, μας επιστρέφει ένα μοντέλο πρόβλεψης.

Εισαγωγή

Στο δεύτερο στάδιο μπορούμε να πραγματοποιήσουμε προβλέψεις χρησιμοποιώντας το μοντέλο πρόβλεψης, απλά με την χρήση δεδομένων τα οποία δεν έχει γνωρίσει ποτέ ο αλγόριθμος.

Εισαγωγή

Εξόρυξη γνώσης από δεδομένα

Σε αυτό το στάδιο της ΑΓΒΔ εφαρμόζεται κάποιος αλγόριθμος για την παραγωγή ενός μοντέλου.
Έχοντας καθαρίσει και μετασχηματίσει τα δεδομένα, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο, ώστε να δημιουργηθεί κάποιο μοντέλο, συνήθως κατηγοριοποίησης ή πρόβλεψης. Θέλουμε να χρησιμοποιήσουμε το μοντέλο αυτό, το οποίο δημιουργήθηκε με βάση κάποια γνωστά δεδομένα, έτσι ώστε να μπορεί να μας δώσει απάντηση για την τιμή ενός χαρακτηριστικού-μεταβλητής στόχου για νέα, άγνωστα δεδομένα.

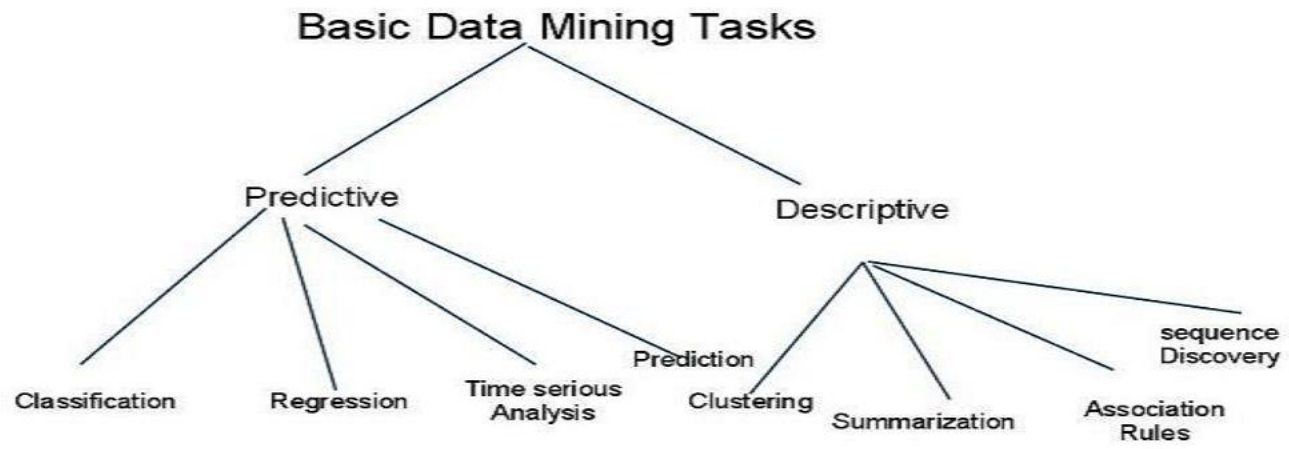
Εισαγωγή

Ερμηνεία και Αξιολόγηση

Είναι πολύ σημαντικό το πώς θα παρουσιαστούν στους χρήστες τα αποτελέσματα της εξόρυξης γνώσης, επειδή η χρησιμότητα ή μη των αποτελεσμάτων μπορεί να εξαρτάται ακριβώς από αυτή την παρουσίαση.

Εισαγωγή

Τα μοντέλα που παράγονται από το στάδιο της Εξόρυξης Δεδομένων διακρίνονται σε δυο βασικούς τύπους: τα μοντέλα πρόβλεψης (predictive) και τα περιγραφικά μοντέλα (descriptive).



Εισαγωγή

Στόχος ενός μοντέλου πρόβλεψης (predictive model) είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών. Η μοντελοποίηση πρόβλεψης μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων .

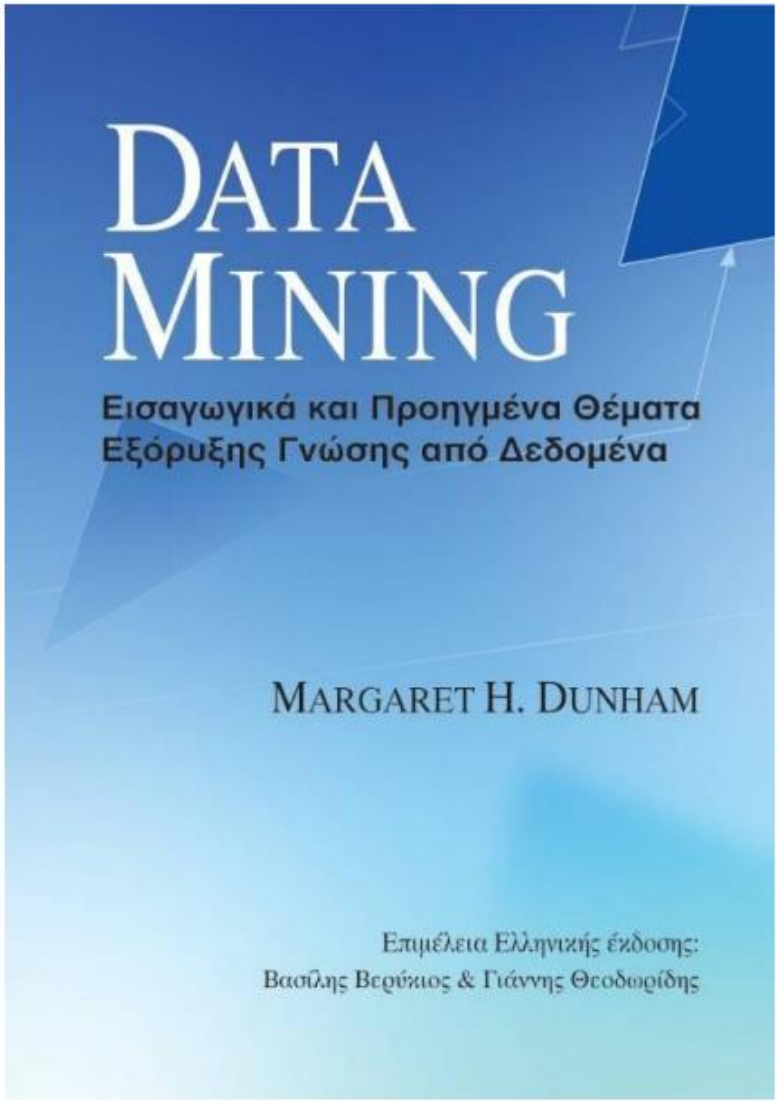
Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός μοντέλου πρόβλεψης περιλαμβάνουν **κατηγοριοποίηση, παλινδρόμηση, ανάλυση χρονολογικών σειρών και πρόβλεψη**.

Εισαγωγή

Ένα περιγραφικό μοντέλο (descriptive model) αναγνωρίζει πρότυπα (patterns) ή σχέσεις (relations) που υπάρχουν στα δεδομένα. Αντίθετα από το προβλεπτικό, το περιγραφικό μοντέλο λειτουργεί σαν ένα μέσο που διερευνά τις ιδιότητες των δεδομένων που εξετάζονται, όχι να προβλέπει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών συνήθως θεωρούνται σαν περιγραφικές εργασίες από τη φύση τους.

ΒΙΒΛΙΑ

- Data Mining (στα Ελληνικά)
  - Margaret H.,Dunham
    - Έτος Έκδοσης: 2004
    - Εκδότης: ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ
    - Αριθμός σελίδων: 315
    - Κωδικός ISBN: 960-8105-72-2



ΒΙΒΛΙΑ

- Εισαγωγή στην Εξόρυξη Δεδομένων και τις Αποθήκες Δεδομένων
- Αλ. Νανόπουλος - Γ. Μανωλόπουλος
  - Έτος Έκδοσης: 2008
  - Εκδότης: ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ
  - Αριθμός σελίδων: 384
  - Κωδικός ISBN: 978-960-6759-17-8