ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΛΟΠΟΙΗΣΗ ΠΡΟΗΓΜΕΝΩΝ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΠΡΟΒΛΗΜΑΤΑ

Παχούλας Γεώργιος

Επιβλέπων: Στύλιος Χρυσόστομος

Καθηγητής

Άρτα, Σεπτέμβριος 2019

IMPLEMENTATION OF DATA MINE'S ADVANCED TECHNIQUES IN REAL PROBLEMS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 16/9/2019

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

   Στύλιος Χρυσόστομος,

2. Μέλος επιτροπής

   Καρβέλης Πέτρος,

3. Μέλος επιτροπής

   Όνομα Επίθετο,

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Παχούλας Γεώργιος

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ τον κύριο Στύλιο Χρυσόστομο και τον κύριο Καρβέλη Πέτρο για την πολύτιμη υποστήριξη και καθοδήγηση στα πλαίσια της εκπόνησης της πτυχιακής εργασίας. Τέλος, ευχαριστώ όλους τους καθηγητές του τμήματος που με τις γνώσεις τους διδάσκουν και εμπνέουν τους φοιτητές.

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια οι εξελίξεις στη δημιουργία και συλλογή δεδομένων έχουν συντελέσει στην παραγωγή συνόλων δεδομένων μεγάλου μεγέθους. Η ευκολία με την οποία τα δεδομένα μπορούν να συλλεχθούν και να αποθηκευτούν έχει δημιουργήσει ιδιαίτερη δυσκολία στην περεταίρω ανάλυση τους. Για αυτό τον λόγο «δημιουργήθηκε» η εξόρυξη δεδομένων για να καλύψει τις ανάγκες των υπαρχόντων μεθόδων ανάλυσης. Ο σκοπός της πτυχιακής είναι να εξοικειωθεί ο αναγνώστης με τους όρους δεδομένα, πληροφορία. Έπειτα να μάθει τι είναι εξόρυξη δεδομένων καθώς και το πότε δημιουργήθηκε καθώς και της κατηγορίες στις οποίες χωρίζεται. Επίσης θα γνωρίσει κάποιους αλγορίθμους εξόρυξης δεδομένων καθώς και τομείς στους οποίους η εξόρυξη αποτελεί σημαντικό κομμάτι των λειτουργιών τους.

Λέξεις-κλειδιά: δεδομένα, εξόρυξη δεδομών, μηχανική μάθηση, αλγόριθμοι, κατηγορίες εξόρυξης δεδομένων

ABSTRACT

In recent years, developments in the creation and collection of data have contributed to the production of large-scale data sets. The ease with which data can be collected and stored has made it particularly difficult to analyze them further. For this reason, data mining was "created" to meet the needs of existing analytical methods. The purpose of the thesis is to familiarize the reader with the terms data, information. Then find out what data mining is and when it was created as well as its categories. He will also be familiar with some data mining algorithms as well as areas where mining is an important part of their operations.

Keywords: data, data mining, algorithms, machine learning, categories of data mining

Περιεχόμενα

ΠΕΡΙΛΗΨΗ 7

ABSTRACT 8

Πίνακας εικόνων 11

1. Δεδομένα 12

1.1 Τι είναι τα δεδομένα **12**

2. Επιστήμη Δεδομένων (Data Science) 14

2.1 Τι είναι η επιστήμη δεδομένων **14**

2.2 Ιστορική εξέλιξη **14**

2.3 Εργαλεία για την επιστήμη δεδομένων **15**

2.4 Τα βήματα της επιστήμης δεδομένων **16**

2.5 Η σχέση της επιστήμης δεδομένων και εξόρυξης δεδομένων **16**

3. Μηχανική μάθηση 17

3.1 Τι είναι η μηχανική μάθηση **17**

3.2 Μέθοδοι μηχανικής μάθησης **17**

3.3 Μοντέλα μηχανικής μάθησης **19**

3.4 Η σχέση μεταξύ εξόρυξης δεδομένων - μηχανικής μάθησης **22**

4. Εξόρυξη Δεδομένων (Data Mining) 24

4.1 Τι είναι η εξόρυξη δεδομένων **24**

4.2 Ιστορική εξέλιξη **25**

4.3 Η διαδικασία ανακάλυψης γνώσης **26**

4.4 Κατηγορίες της εξόρυξης δεδομένων **27**

4.6 Γλώσσες προγραμματισμού και εργαλεία για εξόρυξη δεδομένων **32**

3. Τεχνικές της εξόρυξης δεδομένων 35

3.1 Κατηγοριοποίηση **35**

3.1.1 Αλγόριθμος κοντινότερου γείτονα 37

3.2 Ομαδοποίηση **39**

3.2.1 Αλγόριθμος K-Means 39

6. Εφαρμογές της εξόρυξης δεδομένων 41

5. Υλοποίηση εφαρμογής 44

5.1 Δεδομένα της εφαρμογής **45**

5.2 Αλγόριθμος της εφαρμογής **48**

5.2.1 Αλγόριθμοι ενίσχυσης (Boosting) 48

5.2.2 Αλγόριθμος Gradient Descent 49

5.2.3 Αποτελέσματα 51

Βιβλιογραφία 52

Παράρτημα 56

Πίνακας εικόνων

Εικόνα 1 Διάφοροι τύποι δεδομένων 13
Εικόνα 2 Οι μέθοδοι μηχανικής μάθησης και οι τομείς εφαρμογής τους..... 19
Εικόνα 3 Παράδειγμα δέντρου απόφασης 20
Εικόνα 4 Παράδειγμα-αντιστοιχία ανθρώπινου νευρώνα με τεχνητό..... 21
Εικόνα 5 Παράδειγμα αναπαράστασης των μηχανών υποστήριξης σε 2D χώρο 22
Εικόνα 6 Η σχέση επιστήμης δεδομένων με μηχανική μάθηση..... 23
Εικόνα 7. Βήματα για την ανακάλυψη γνώσης..... 26
Εικόνα 8. Κατηγορίες της εξόρυξης δεδομένων. 27
Εικόνα 9.Σύγκριση προγνωστικής με περιγραφικής εξόρυξης. 28
Εικόνα 10. Παράδειγμα ανάλυσης συσχετίσεων..... 30
Εικόνα 11. Παράδειγμα ανάλυσης συστάδων. 30
Εικόνα 12 Παράδειγμα ανίχνευσης ανωμαλιών..... 31
Εικόνα 13 "Εργαλεία" για την εξόρυξη δεδομένων..... 34
Εικόνα 14 Συγκριτικός πίνακας χρήσης των διαφόρων εργαλείων για την εξόρυξη δεδομένων 34
Εικόνα 15 Παράδειγμα πριν την εκτέλεση του αλγορίθμου KNN..... 37
Εικόνα 16. Παράδειγμα μετά το τέλος του αλγορίθμου KNN. 38
Εικόνα 17 Παράδειγμα ομαδοποίησης K-Means με 2 centroid 40
Εικόνα 18 Αναπαράσταση του SSE και εύρεση του σημείου "αγκώνα"..... 40
Εικόνα 19. Η εξόρυξη δεδομένων στον τομέα της υγείας. 41
Εικόνα 20. Η εξόρυξη δεδομένων στον τομέα των αγορών..... 42
Εικόνα 21. Ανάλυση καλαθιού αγοράς. 43
Εικόνα 22 Επαναληπτική διαδικασία ενδυνάμωσης του μοντέλου..... 49
Εικόνα 23 Ο ρυθμός μάθησης σε σχέση με το κόστος 50
Εικόνα 24 Σύγκριση ρυθμού μάθησης..... 50

1. Δεδομένα

1.1 Τι είναι τα δεδομένα

Δεδομένα είναι κάθε σύνολο από χαρακτήρες τα οποία συλλέγονται και «μεταφράζονται» συνήθως για ανάλυση. Μπορούν να περιέχουν κείμενο ,αριθμούς, εικόνες, ήχο, ή βίντεο. Γενικά δεδομένα είναι μία εναλλακτική λέξη για την πληροφορία. (Anon., 2019) Τα δεδομένα τα οποία δεν έχουν αναλυθεί ακόμη ονομάζονται ακατέργαστα δεδομένα (raw data). Τα ακατέργαστα δεδομένα ("μη επεξεργασμένα δεδομένα") είναι μια συλλογή αριθμών ή χαρακτήρων τα οποία δεν έχουν υποστεί εκκαθάριση και διόρθωση από τους ερευνητές. (Joris Toonders, 2014) Τα ακατέργαστα δεδομένα πρέπει να διορθωθούν για να απομακρυνθούν οι υπερβολικές τιμές ή τα προφανή λάθη εγγραφής οργάνων ή δεδομένων (π.χ., ανάγνωση θερμόμετρου από μια υπαίθρια περιοχή της Αρκτικής που καταγράφει τροπική θερμοκρασία). Τα δεδομένα έπειτα από αύτη την επεξεργασία γίνονται κατάλληλα για χρήση.

Η λέξη δεδομένα χρησιμοποιήθηκε για πρώτη φορά το 1640. Αρχικά σήμαινε “transmissible and storable computer information” το 1946.

Στον κλάδο της πληροφορικής και των επιχειρήσεων ο όρος δεδομένα αναφέρεται στην πληροφορία η οποία είναι πιο εύκολα κατανοητή από έναν υπολογιστή παρά από τον άνθρωπο. (Anon., 2018)

Υπάρχουν διάφοροι τύποι δεδομένων αλλά οι πιο κοινοί είναι:

Προσωπικά Δεδομένα (Personal Data):

Προσωπικά δεδομένα είναι οτιδήποτε προσδιορίζει κάποιον. Μπορεί να είναι το ονοματεπώνυμο, το email ή η τοποθεσία. Πολλές εταιρίες συλλέγουν προσωπικά δεδομένα (κυρίως ιστοσελίδες κοινωνικής δικτύωσης) με σκοπό την προβολή περιεχομένου το οποίο απευθύνεται ξεχωριστά για τον καθένα (targeted ads).

Δεδομένα Συναλλαγής (Transactional Data):

Τα δεδομένα συναλλαγής είναι ιδιαίτερα σημαντικά για τις επιχειρήσεις καθώς συμβάλουν στο να αυξήσουν το ανταγωνιστικό τους πλεονέκτημα μέσω της βελτίωσης των λειτουργιών τους (π.χ. πιο αποτελεσματικό μάρκετινγκ).

Δεδομένα από το διαδίκτυο (Web Data):

Τα δεδομένα αυτά αναφέρονται σε οποιονδήποτε τύπο δεδομένων που συλλέγονται από το διαδίκτυο π.χ. για ερευνητικούς λόγους. Είναι πολύ σημαντικά καθώς δίνουν την δυνατότητα σε κάποιον να τα χρησιμοποιήσει χωρίς να τα έχει δημιουργήσει ο ίδιος.

Εικόνα 1 Διάφοροι τύποι δεδομένων

2. Επιστήμη Δεδομένων (Data Science)

2.1 Τι είναι η επιστήμη δεδομένων

Η εμφάνιση προηγμένων τεχνολογιών στον τομέα της επιστήμης των υπολογιστών είχε ως αποτέλεσμα στη μαζική αύξηση των δεδομένων. Πολλές εταιρείες πρέπει να αναλύουν και να αντλούν σημαντικές πληροφορίες κάτι που είναι αρκετά δύσκολο με τόσο μεγάλο όγκο δεδομένων. Την λύση στο πρόβλημα αυτό έρχεται να δώσει η επιστήμη των δεδομένων (Data Science). Η επιστήμη των δεδομένων είναι ένα διεπιστημονικό πεδίο που χρησιμοποιεί επιστημονικές μεθόδους, διαδικασίες, αλγορίθμους και συστήματα για την εξαγωγή γνώσεων από δομημένα και αδόμητα δεδομένα. Αποτελεί ένα συνδυασμό της στατιστικής (statistics), της ανάλυσης δεδομένων (data analysis), της μηχανικής μάθησης (machine learning) και άλλων σχετικών μεθόδων οι οποίες συμβάλλουν στην «κατανόηση» και ανάλυση δεδομένων. Χρησιμοποιεί τεχνικές και θεωρίες που αντλούνται από πολλούς τομείς όπως των μαθηματικών, των στατιστικών, της επιστήμης των υπολογιστών και της επιστήμης των πληροφοριών. Η επιστήμη των δεδομένων (Data Science) αποτελεί μια από τις πιο σύγχρονες εργασίες του 21ου αιώνα. Έχει χαρακτηριστεί ως η πιο «δημοφιλής» δουλεία του 21ου αιώνα από το Harvard Business Review. Οι βιομηχανίες χρειάζονται «επιστήμονες δεδομένων» (data scientists) που μπορούν να τους βοηθήσουν να λάβουν ισχυρές αποφάσεις με βάση δεδομένα. Υπάρχουν άφθονες θέσεις στον τομέα της επιστήμης των δεδομένων. Αυτό οφείλεται στο γεγονός ότι τα δεδομένα είναι πανταχού παρόντα. (Dhar, 2013) (Leek, 2013)

2.2 Ιστορική εξέλιξη

Αρχικά ο όρος χρησιμοποιήθηκε σαν υποκατάστατο για την επιστήμη των υπολογιστών από τον Pete Naur το 1960. Αργότερα ο Naur εισήγαγε τον όρο datalogy. Το 1974 δημοσίευσε μια συνοπτική ερεύνα των μεθόδων των υπολογιστών στην οποία χρησιμοποίησε τον όρο επιστήμη δεδομένων με τον οποίο αναφερόταν στις σύγχρονες μεθόδους επεξεργασίας δεδομένων που χρησιμοποιούνται σε ένα ευρύ φάσμα εφαρμογών. Ο σύγχρονος ορισμός της επιστήμης δεδομένων «σχεδιάστηκε» για πρώτη φορά κατά τη διάρκεια του δεύτερου ιαπωνικού-γαλλικού στατιστικού «συνεδρίου» που διοργανώθηκε στο πανεπιστήμιο Montpellier II το 1992 στην Γαλλία. Οι συμμετέχοντες κατανόησαν την ανάγκη για την δημιουργία

μιας νέας επιστήμης με ιδιαίτερη έμφαση στα δεδομένα, την οποία πλαισίωσαν με καθιερωμένες έννοιες και αρχές στατιστικής και ανάλυσης δεδομένων σε συνδυασμό με την εκτεταμένη χρήση της αυξανόμενης ισχύος των ηλεκτρονικών υπολογιστών. Αργότερα το 1996 ο όρος αυτός χρησιμοποιήθηκε για πρώτη φορά στον τίτλο της διάσκεψης της International Federation of Classification Societies (IFCS) «Επιστήμη των δεδομένων, ταξινόμηση και συναφείς μέθοδοι» (Yves Escoufier, 1995)

2.3 Εργαλεία για την επιστήμη δεδομένων

Python: Η πιο δημοφιλής γλώσσα για την επιστήμη δεδομένων. Προσφέρει μια μεγάλη ποικιλία βιβλιοθηκών που υποστηρίζουν τη λειτουργία της επιστήμης δεδομένων.

SAS (Statistical Analysis System): Σουίτα λογισμικού για την διευκόλυνση διαφόρων στατιστικών πράξεων.

R: Γλώσσα προγραμματισμού που προσφέρει διάφορα πακέτα για τη την οπτικοποίηση και ανάλυση δεδομένων.

TensorFlow: Βιβλιοθήκη μηχανικής μάθησης που χρησιμοποιείται για την εφαρμογή αλγορίθμων βαθιάς μάθησης. Υποστηρίζεται από γραφικές μονάδες επεξεργασίας (GPU).

Apache Spark: Παρέχει δυνατότητες επεξεργασίας και ανάλυσης δεδομένων σε μεγάλα δεδομένα (big data).

Tableau: Λογισμικό απεικόνισης που χρησιμοποιείται για την δημιουργία διαδραστικών γραφημάτων.

D3.js: Βιβλιοθήκη σε γλώσσα προγραμματισμού javascript για την δημιουργία διαδραστικών απεικονίσεων.

2.4 Τα βήματα της επιστήμης δεδομένων

Τα 5 βήματα στην επιστήμη δεδομένων

1. Εξαγωγή δεδομένων (Data Extraction): Το πρώτο βήμα είναι η «παραλαβή» των δεδομένων. Αυτό μπορεί να γίνει με την μορφή «ερωτημάτων» (queries) σε βάσεις δεδομένων (SQL & No SQL).
2. Προ επεξεργασία των δεδομένων (Data Preprocessing): Το βήμα αυτό περιλαμβάνει τον καθαρισμό των δεδομένων και την αντικατάσταση των τιμών που λείπουν. Αποτελεί το πιο σημαντικό βήμα καθώς διαμορφώνει - οργανώνει τα τελικά αποτελέσματα τα οποία θα χρησιμοποιηθούν για περεταίρω ανάλυση.
3. Ανάλυση δεδομένων (Data analysis): Περιλαμβάνει την χρήση διαφόρων στατιστικών μεθόδων όπως στατιστικά και περιγραφικά στοιχεία για την εύρεση προτύπων και τάσεων μέσα στα δεδομένα.
4. Δημιουργία προβλέψεων (Generating Predictions): Στο βήμα αυτό δημιουργούνται προβλέψεις χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης. Οι προβλέψεις εκτελούνται στα ιστορικά δεδομένα για την πρόβλεψη μελλοντικών γεγονότων καθώς και στην δημιουργία μοντέλων για την σύλληψη μοτίβων μέσα στα δεδομένα.
5. Βελτιστοποίηση του μοντέλου (Optimizing models): Το τελικό βήμα είναι η βελτιστοποίηση του μοντέλου μηχανικής μάθησης με σκοπό να βελτιώσει την επίδοση και την ακρίβεια του μοντέλου.

2.5 Η σχέση της επιστήμης δεδομένων και εξόρυξης δεδομένων

Η εξόρυξη δεδομένων και η επιστήμη των δεδομένων είναι δύο από τα σημαντικότερα θέματα της τεχνολογίας. Και τα δύο αυτά πεδία περιστρέφονται γύρω από τα δεδομένα. Ωστόσο, ο τρόπος που χρησιμοποιούν τα δεδομένα καθώς και οι γνώσεις που απαιτούνται για την διεξαγωγή των αποτελεσμάτων σε αυτούς τους τομείς είναι διαφορετικός. (Team, 2019)

- Η επιστήμη δεδομένων αποτελεί ένα σύνολο λειτουργιών που περιλαμβάνει και την εξόρυξη δεδομένων.

- Η επιστήμη δεδομένων ασχολείται τόσο με δομημένα όσο και αδόμητα δεδομένα ενώ η εξόρυξη δεδομένων ασχολείται μόνο με δομημένες πληροφορίες.

3. Μηχανική μάθηση

3.1 Τι είναι η μηχανική μάθηση

Η μηχανική μάθηση ή αλλιώς μάθηση μέσω μηχανής (**M**achine **L**earning) είναι μια εφαρμογή της τεχνητής νοημοσύνης (AI) και βασίζεται στην επιστημονική μελέτη αλγορίθμων και στατιστικών μοντέλων που χρησιμοποιούν τα ηλεκτρονικά συστήματα για την εκτέλεση συγκεκριμένων εργασιών χωρίς τη χρήση ρητών οδηγιών. Η μηχανική μάθηση σχετίζεται στενά με τις στατιστικές υπολογιστικής, η οποία επικεντρώνεται στην πραγματοποίηση προβλέψεων με τη χρήση ηλεκτρονικών υπολογιστών. Η διαδικασία μάθησης ξεκινά με παρατηρήσεις ή δεδομένα γνωστά ως δεδομένα εκπαίδευσης (training set) τα οποία χρησιμοποιούνται για την δημιουργία ενός μαθηματικού μοντέλου το οποίο αναζητεί μοτίβα στα δεδομένα με σκοπό να λάβει καλύτερες αποφάσεις στο μέλλον. Η εξόρυξη δεδομένων είναι ένα πεδίο μελέτης στο πλαίσιο της μηχανικής μάθησης και επικεντρώνεται στην διερευνητική ανάλυση δεδομένων μέσω της μάθησης χωρίς επίβλεψη. Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. (Bishop, 2006)

3.2 Μέθοδοι μηχανικής μάθησης

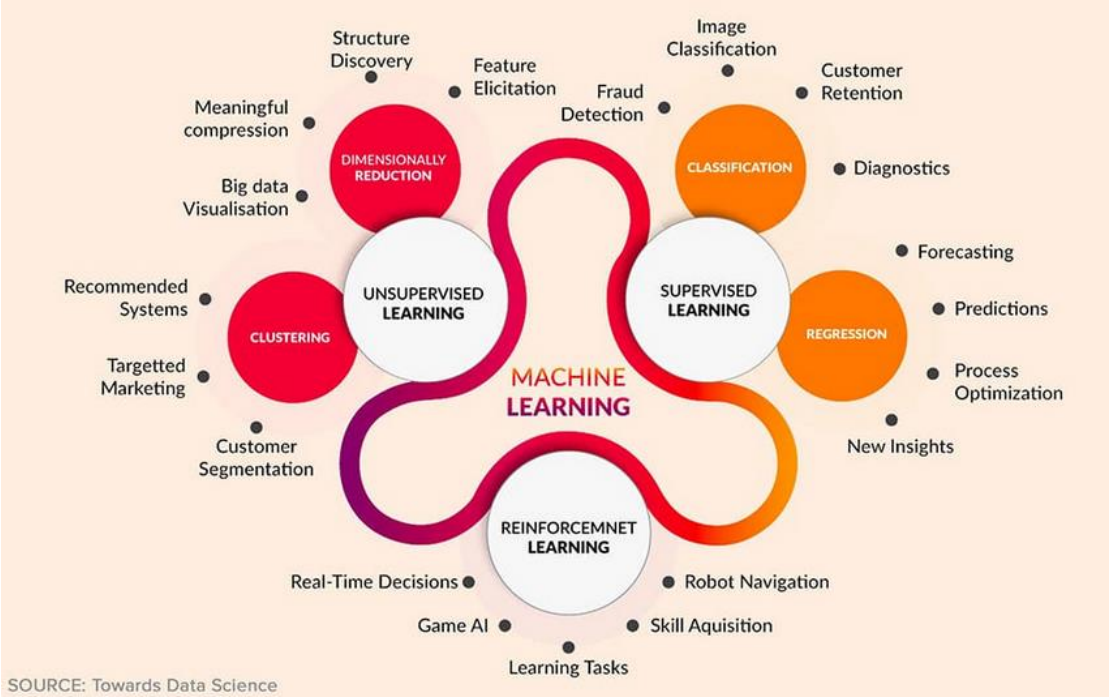
Αλγόριθμοι μάθησης με εποπτεία (Supervised machine learning algorithms): Οι αλγόριθμοι αυτοί χρησιμοποιούνται σε παραδείγματα με «ετικέτα» δηλαδή σε παραδείγματα όπου για μία συγκεκριμένη είσοδο γνωρίζουμε την επιθυμητή έξοδο. Ο αλγόριθμος μάθησης λαμβάνει ένα σύνολο εισόδων μαζί με τις αντίστοιχες σωστές εξόδους και ο αλγόριθμος μαθαίνει συγκρίνοντας την πραγματική του έξοδο με τις σωστές εξόδους για να εντοπίσει σφάλματα. Στη συνέχεια τροποποιεί το μοντέλο ανάλογα. Μέσω μεθόδων όπως η ταξινόμηση (classification), η παλινδρόμηση

(regression), η πρόβλεψη (prediction) και η αύξηση της κλίσης (Gradient Boosting), η εποπτευόμενη μάθηση χρησιμοποιεί πρότυπα για την πρόβλεψη των τιμών της ετικέτας σε πρόσθετα μη επισημασμένα δεδομένα. Η εποπτευόμενη μάθηση χρησιμοποιείται συνήθως σε εφαρμογές όπου ιστορικά δεδομένα προβλέπουν πιθανά μελλοντικά γεγονότα. Για παράδειγμα, μπορεί να προβλέψει πότε οι συναλλαγές με πιστωτικές κάρτες είναι πιθανό να είναι δόλιες ή ποιος ασφαλισμένος πελάτης είναι πιθανό να υποβάλει αξίωση.

Αλγόριθμοι μάθησης χωρίς εποπτεία (Unsupervised learning algorithms): Η μάθηση χωρίς επιτήρηση χρησιμοποιείται σε δεδομένα που δεν έχουν ιστορικές ετικέτες. Ο στόχος είναι να διερευνηθούν τα δεδομένα και να βρεθεί κάποια δομή εντός. Η μη επιτηρούμενη μάθηση λειτουργεί καλά σε δεδομένα συναλλαγών. Για παράδειγμα, μπορεί να εντοπίσει τμήματα πελατών με παρόμοια χαρακτηριστικά, τα οποία μπορούν να αντιμετωπιστούν με τον ίδιο τρόπο σε καμπάνιες μάρκετινγκ. Ή μπορεί να βρει τα κύρια χαρακτηριστικά που χωρίζουν τα τμήματα πελατών από το άλλο. Οι δημοφιλείς τεχνικές περιλαμβάνουν αυτο-οργανωμένους χάρτες, χαρτογράφηση πλησιέστερου γείτονα, συγκέντρωση k-mean και αποσύνθεση μοναδικής τιμής. Αυτοί οι αλγόριθμοι χρησιμοποιούνται επίσης για την ταξινόμηση θεμάτων κειμένου, την υποβολή προτάσεων και την εξακρίβωση των αποδόσεων δεδομένων.

Αλγόριθμοι μάθησης με ενδυνάμωση (Reinforcement learning algorithms): Η μέθοδος αυτή χρησιμοποιείται συχνά στον τομέα της ρομποτικής και στην πλοήγηση. Με την ενίσχυση της μάθησης ο αλγόριθμος ανακαλύπτει μέσω δοκιμών και σφαλμάτων ποιες ενέργειες αποφέρουν τα μεγαλύτερα οφέλη. Η μάθηση αυτή έχει τρία βασικά χαρακτηριστικά. Τον «πράκτορα»-παράγοντα ο οποίος είναι υπεύθυνος για την λήψη αποφάσεων, το περιβάλλον δηλαδή οτιδήποτε αλληλεπιδρά με τον παράγοντα και τις ενέργειες δηλαδή τις επιλογές που μπορεί να διαλέξει ο παράγοντας. Σκοπός του είναι ο παράγοντας να επιλέξει της κατάλληλες επιλογές που θα μεγιστοποιήσουν την αναμενόμενη ανταμοιβή.

Αλγόριθμοι μάθησης με ημι-εποπτεία (**Semi-supervised algorithms**): Οι αλγόριθμοι ημι-εποπτευόμενης εκμάθησης συνδυάζουν της εποπτευόμενη και την μη εποπτευόμενη μάθηση καθώς χρησιμοποιούν τόσο δεδομένα με ετικέτα όσο και χωρίς. Συνήθως μια μικρή ποσότητα δεδομένων με ετικέτα και ένα μεγάλο ποσό χωρίς ετικέτες. Αυτός ο τύπος μάθησης μπορεί να χρησιμοποιηθεί με μεθόδους όπως η ταξινόμηση, η παλινδρόμηση και η πρόβλεψη. Η μάθηση μέσω αλγορίθμων ημι-εποπτείας είναι χρήσιμη όταν το κόστος που συνδέεται με την επισήμανση είναι πολύ υψηλό για να επιτρέψει μια πλήρως εκπαιδευμένη διαδικασία κατάρτισης. Τα συστήματα που χρησιμοποιούν αυτή τη μέθοδο είναι σε θέση να βελτιώσουν σημαντικά την ακρίβεια της μάθησης. α πρώτα παραδείγματα αυτού του είδους περιλαμβάνουν την αναγνώριση του προσώπου ενός προσώπου σε μια web cam. (Team, 2019)



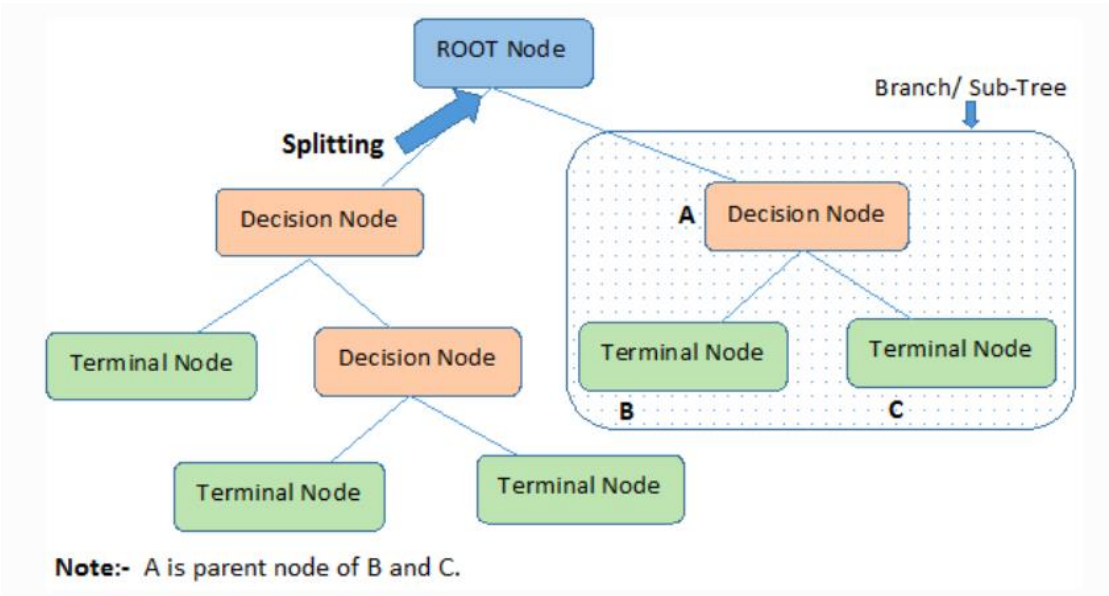
Εικόνα 2 Οι μέθοδοι μηχανικής μάθησης και οι τομείς εφαρμογής τους

3.3 Μοντέλα μηχανικής μάθησης

Δέντρα απόφασης

Τα δέντρα απόφασης ανήκουν στην κατηγορία αλγορίθμων μάθησης με εποπτεία και σε αντίθεση με άλλους αλγορίθμους της κατηγορίας αυτής μπορούν να λύσουν προβλήματα κατηγοριοποίησης (classification) και οπισθοδρόμηση (regression). Ο

στόχος του αλγορίθμου είναι η δημιουργία ενός μοντέλου εκπαίδευσης που μπορεί να χρησιμοποιηθεί για να προβλέψει την τάξη ή την τιμή της επιθυμητής μεταβλητής κάνοντας χρήση κανόνων απόφασης που προέρχονται από προηγούμενα δεδομένα εκπαίδευσης. Για να προβλέψουμε την κατηγορία μια εγγραφής ξεκινάμε από την «ρίζα» του δέντρου, συγκρίνουμε τις τιμές που έχει η «ρίζα» του δέντρου με τις τιμές της εγγραφής και ανάλογα την τιμή προχωράμε στο επόμενο «κλαδί» του δέντρου. Τα δέντρα απόφασης χωρίζονται σε 2 κατηγορίες. Στα κατηγορηματικά (categorical) δέντρα τα οποία βασίζονται σε κατηγορίες και στα συνεχή (continuous) τα οποία βασίζονται σε συνεχή τιμές. Μερική αλγόριθμοι που χρησιμοποιούνται στα δέντρα απόφασης είναι ID3, CART, C4.5. (Chauhan, 2019)

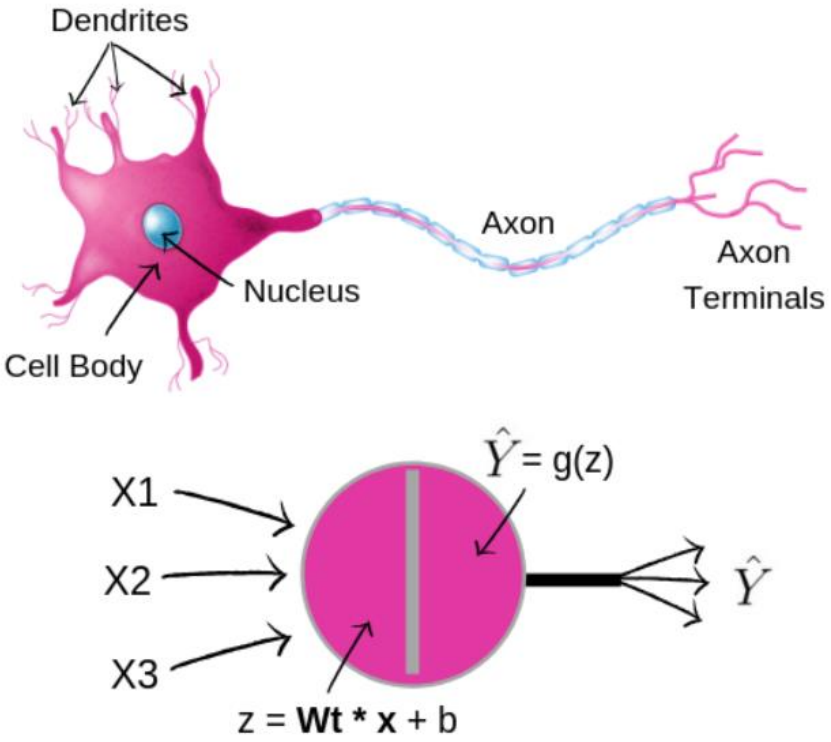


Εικόνα 3 Παράδειγμα δέντρου απόφασης

Τεχνητά νευρωνικά δίκτυα

Ο αλγόριθμος αυτός βασίζεται σε επίπεδα εισόδου-εξόδου καθώς και σε κρυφά επίπεδα και λειτουργούν με παρόμοιο τρόπο όπως τα νευρωνικά δίκτυα του ανθρώπινου εγκεφάλου. Τα νευρωνικά δίκτυα ανήκουν στην κατηγορία αλγορίθμων μάθησης με επιτήρηση καθώς οι ερευνητές θα πρέπει να κατηγοριοποιήσουν τα δεδομένα προκειμένου να εκπαιδεύσουν το μοντέλο και να εξάγουν αποτελέσματα. Δέχονται σαν είσοδο ένα σύνολο δεδομένων που

ονομάζεται επίπεδο εισαγωγής (input layer) και τα αποτελέσματα του αλγορίθμου ονομάζονται επίπεδο εξόδου (output layer). Στην περίπτωση που τα δεδομένα είναι πιο πολύπλοκα μπορεί να υπάρξουν επιπλέον κρυφά επίπεδα μεταξύ της εισόδου και της εξόδου τα οποία εκτελούν πολύπλοκους υπολογισμούς. Όταν ένα τεχνητό νευρωνικό δίκτυο περιέχει παραπάνω από ένα κρυφό επίπεδο θεωρείτε ας δίκτυο βαθιάς μάθησης (deep learning) και το βάθος εξαρτάται από τον αριθμό των κρυφών επιπέδων. (Schott, 2019)

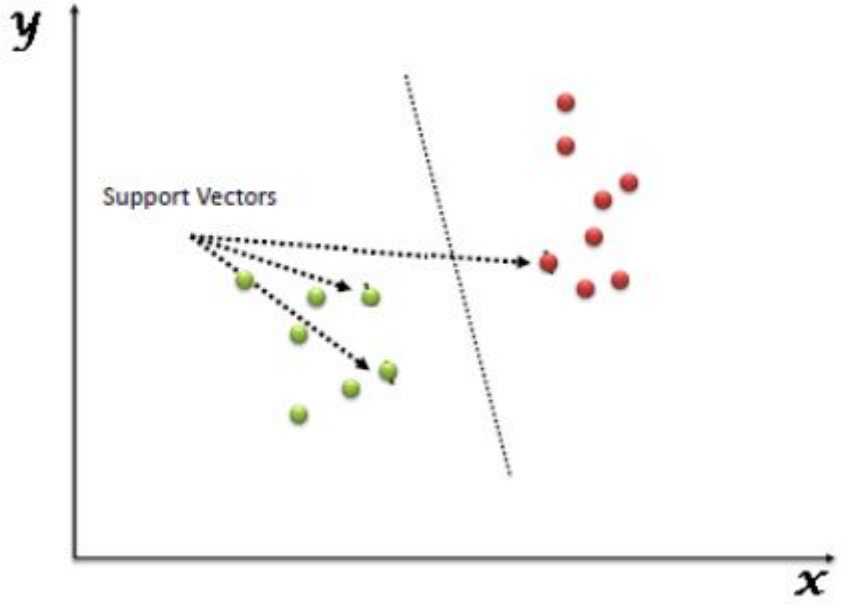


Εικόνα 4 Παράδειγμα-αντιστοιχία ανθρώπινου νευρώνα με τεχνητό

Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης ανήκουν στις κατηγορίες αλγορίθμων μάθησης με εποπτεία. Στον αλγόριθμο αυτό αναπαριστούμε κάθε δεδομένο σαν ένα σημείο σε ένα n-διάστατο χώρο όπου n είναι ο αριθμός των χαρακτηριστικών που έχει το πρόβλημά μας και ο τιμή του κάθε χαρακτηριστικού αντιπροσωπεύει τις συντεταγμένες του στοιχείου στον χώρο. Έπειτα κατηγοριοποιούμε τα δεδομένα βρίσκοντας τρόπο να χωρίσουμε τα δεδομένα σε υποχώρους (hyperplane) οι οποίοι θα διαφοροποιούν αρκετά τις κατηγορίες μεταξύ τους. Τα διανύσματα υποστήριξης

είναι απλά οι συντεταγμένες των παρατηρήσεων που προβάλλονται στον χώρο. (Ray, 2017)



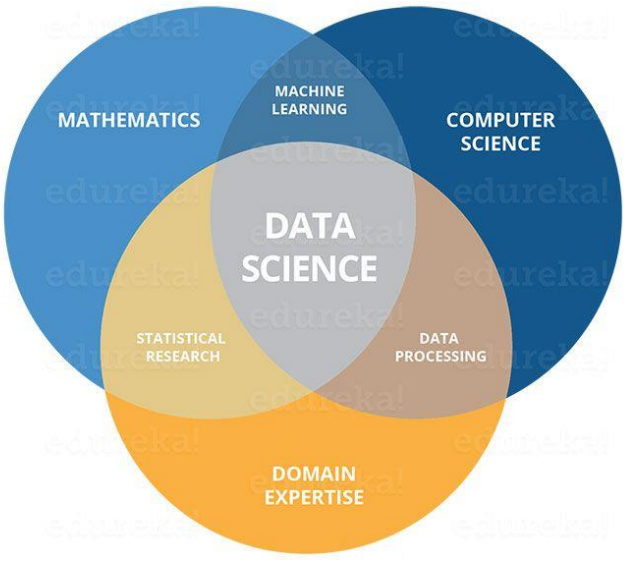
Εικόνα 5 Παράδειγμα αναπαράστασης των μηχανών υποστήριξης σε 2D χώρο

3.4 Η σχέση μεταξύ εξόρυξης δεδομένων - μηχανικής μάθησης

Τόσο η εξόρυξη δεδομένων όσο και η μηχανική μάθηση εμπίπτουν στον τομέα της Επιστήμης των Δεδομένων (Data Science). Χρησιμοποιούν συχνά τις ίδιες μεθόδους αλλά η μηχανική μάθηση επικεντρώνεται στην πρόβλεψη, βασισμένη σε γνωστές ιδιότητες που αντλήθηκαν από τα δεδομένα εκπαίδευσης, ενώ η εξόρυξη δεδομένων επικεντρώνεται στην ανακάλυψη προηγούμενων άγνωστων ιδιοτήτων στα δεδομένα. Και οι δύο μέθοδοι χρησιμοποιούνται για την επίλυση σύνθετων προβλημάτων, συνεπώς, πολλοί άνθρωποι (λανθασμένα) χρησιμοποιούν τους δύο όρους εναλλακτικά. Αυτό δεν είναι τόσο περίεργο, δεδομένου ότι η μηχανική μάθηση χρησιμοποιείται μερικές φορές ως μέσο για τη διεξαγωγή χρήσιμης εξόρυξης δεδομένων. Η εξόρυξη δεδομένων χρησιμοποιεί πολλές μεθόδους μηχανικής αλλά με διαφορετικούς στόχους. Από την άλλη πλευρά η μηχανική μάθηση χρησιμοποιεί επίσης μεθόδους εξόρυξης δεδομένων ως μάθηση χωρίς επίβλεψη για την βελτίωση της ακρίβειας. (Li, 2017)

Σύγκριση εξόρυξης δεδομένων και μηχανικής μάθησης

Basic for comparison	Data mining	Machine Learning
Meaning	Extracting knowledge from a large amount of data	Introduce new algorithm from data as well as past experience
History	Introduce in 1930, initially referred as knowledge discovery in databases	Introduce in near 1950, the first program was Samuel's checker-playing program
Responsibility	Data mining is used to get the rules from existing data	Machine Learning teaches the computer to learn and understand the given rules
Techniques involve	Data mining is more of a research using methods like machine learning	Self-learned and trains system to do the intelligent task



Εικόνα 6 Η σχέση επιστήμης δεδομένων με μηχανική μάθηση

4. Εξόρυξη Δεδομένων (Data Mining)

4.1 Τι είναι η εξόρυξη δεδομένων

Τα τελευταία χρόνια το μέγεθος των βάσεων δεδομένων έχει αυξηθεί ραγδαία. Αυτή η ταχύτατη ανάπτυξη του διαθέσιμου όγκου δεδομένων είναι αποτέλεσμα της μηχανοργάνωσης (οργάνωση και υποστήριξη, με την χρήση ηλεκτρονικών υπολογιστών) της κοινωνίας μας και της ταχείας ανάπτυξης εργαλείων συλλογής και αποθήκευσης δεδομένων. (Christopher, 2010) Οι επιχειρήσεις παράγουν γιγαντιαία σύνολα δεδομένων π.χ. δεδομένα συναλλαγών πωλήσεων, περιγραφές προϊόντων, κ.τ.λ. Στο τομέα των επιστημών δημιουργούνται συνεχώς τεράστιος όγκος δεδομένων (petabyte) από διαδικασίες όπως μετρήσεις και παρατηρήσεις πειραμάτων. Η ιατρική και η βιομηχανία υγείας παράγουν τεράστια ποσά δεδομένων από ιατρικά αρχεία, παρακολούθηση ασθενών και ιατρική απεικόνιση. Αυτό το εκρηκτικά αυξανόμενο, ευρέως διαθέσιμο και γιγαντιαίο σύνολο δεδομένων καθιστά τον καιρό μας πραγματικά την ηλικία των δεδομένων. (R Ragavi, 2018) Αυτό έχει οδηγήσει σε ένα αυξανόμενο ενδιαφέρον για την ανάπτυξη ευέλικτων εργαλείων, ικανών να εξάγουν αυτόματα γνώσεις από δεδομένα και να τις μετατρέπουν σε «οργανωμένη» γνώση. Αυτή η αναγκαιότητα έχει οδηγήσει στην «γέννηση» της εξόρυξης δεδομένων.

Εξόρυξη δεδομένων (DM) είναι η διαδικασία με την οποία εξάγεται πληροφορία από ένα τεράστιο σύνολο δεδομένων, χρησιμοποιώντας μεθόδους μηχανικής μάθησης, στατιστικής και συστημάτων βάσης δεδομένων. (Christopher, 2010) Η εξόρυξη δεδομένων αποκαλύπτει συσχετίσεις, μοτίβα, αλλαγές, ανωμαλίες και σημαντικές δομές από ένα μεγάλο όγκο ακατέργαστων (raw) δεδομένων, που είναι αποθηκευμένα σε βάσεις δεδομένων, αποθήκες δεδομένων η άλλα αποθετήρια πληροφορίων. Η εξόρυξη δεδομένων αποκαλείται επίσης αποκάλυψη γνώσεων σε βάσεις δεδομένων ή ανακάλυψη γνώσης και εξόρυξη δεδομένων (KDD). (R Ragavi, 2018)

Η διαφορά μεταξύ της ανάλυσης δεδομένων (data analysis) και της εξόρυξης δεδομένων (data mining) είναι ότι η ανάλυση δεδομένων χρησιμοποιείται για τη δοκιμή μοντέλων και υποθέσεων σχετικά με το σύνολο δεδομένων π.χ. αποτελεσματικότητα μιας καμπάνιας μάρκετινγκ. Αντίθετα η εξόρυξη δεδομένων χρησιμοποίει μοντέλα μηχανικής μάθησης και στατιστικά μοντέλα για την

ανακάλυψη «κρυφών» μοτίβων σε ένα μεγάλο όγκο δεδομένων. (Soumen Chakrabarti, 2006)

4.2 Ιστορική εξέλιξη

Κατά την δεκαετία του ’80 οι αποθηκευτικές δυνατότητες των ηλεκτρονικών υπολογιστών αυξήθηκαν αρκετά έχοντας ως αποτέλεσμα πολλές εταιρείες να αποθηκεύουν περισσότερα δεδομένα συναλλαγών. Οι συλλογές δεδομένων που αποθηκεύονταν ήταν πολύ μεγάλες για να αναλυθούν με παραδοσιακές στατιστικές προσεγγίσεις. Πραγματοποιήθηκαν διάφορα συνέδρια για να εξεταστεί πως οι εξελίξεις εκείνης της περιόδου στον τομέα της τεχνητής νοημοσύνης μπορούν να συμβάλουν στην ανακάλυψη πληροφοριών από μεγάλα σύνολα δεδομένων. (Christopher, 2010) Το 1989 ο Gregory Piatetsky-Shapiro Ρώσος αναλυτής δεδομένων επινόησε τον όρο αποκάλυψη γνώσεις σε βάσεις δεδομένων Knowledge Discovery in Databases (KDD). Η διαδικασία αυτή οδήγησε το 1995 στην πρώτη διεθνή σύσκεψη για την «ανακάλυψη γνώσης και την εξόρυξη δεδομένων» στο Μόντρεαλ και την κυκλοφορία του περιοδικού Data Mining and Knowledge Discovery το 1997. Αυτή ήταν και η περίοδος κατά την οποία δημιουργήθηκαν πολλές πρώιμες εταιρείες εξόρυξης δεδομένων και εισήχθησαν προϊόντα στην αγορά. Αυτή η αύξηση της δημοτικότητας μπορεί να αποδοθεί στην πρόοδο της τεχνολογίας, η ικανότητα επεξεργασίας των υπολογιστών και οι διαθέσιμες δυνατότητες αποθήκευσης δεδομένων σήμαιναν ότι η επεξεργασία μεγάλων όγκων δεδομένων χρησιμοποιώντας υπολογιστές «γραφείου» ήταν μια ρεαλιστική δυνατότητα. (COENEN, 2004)

Μία από τις πρώτες εφαρμογές της εξόρυξης δεδομένων ήταν η ανίχνευση απάτης με χρήση πιστωτικής κάρτας. Μελετώντας την αγοραστική συμπεριφορά του καταναλωτή μπορούσαν να προκύψουν κάποια τυπικά μοτίβα αγορών-συναλλαγών. Οι αγορές που πραγματοποιούνταν εκτός αυτού του μοτίβου μπορούσαν να είναι είτε υλικό για να εμπλουτίσουν το υπάρχον μοτίβο είτε κάποια «παραβίαση» στην οποία έπρεπε να παρέμβει ο ιδιοκτήτης της κάρτας και να ακυρώσει-αρνηθεί την συναλλαγή.

4.3 Η διαδικασία ανακάλυψης γνώσης

Πολλοί άνθρωποι θεωρούν ότι η εξόρυξη δεδομένων (data mine) είναι συνώνυμο με την ανακάλυψη γνώσης από δεδομένα (knowledge discover) ενώ άλλοι θεωρούν την εξόρυξη ως ένα σημαντικό βήμα στην διαδικασία ανακάλυψης της γνώσης. Η διαδικασία ανακάλυψης γνώσης αποτελεί μία επαναληπτική των ακόλουθων βημάτων:

Data Cleaning: Στο βήμα αυτό αφαιρούνται από την συλλογή δεδομένα θορύβου (παραμορφωμένα) ή άσχετα δεδομένα. Επίσης γίνεται συμπλήρωση των ελλειπουσών τιμών και διόρθωση στις ασυνέπειες των δεδομένων.

Data Integration: Στο βήμα αυτό δεδομένα από πολλές διαφορετικές πηγές συνδυάζονται σε ένα σύνολο δεδομένων. Οι πηγές αυτές μπορεί να περιλαμβάνουν βάσεις δεδομένων, κύβους δεδομένων.

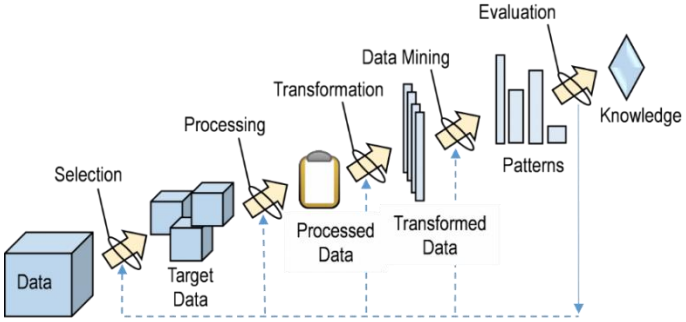
Data Selection: Στο βήμα αυτό ανακτώνται από την βάση δεδομένων τα δεδομένα που θα χρησιμοποιηθούν για ανάλυση.

Data Transformation: Στο βήμα αυτό τα δεδομένα μετατρέπονται σε φόρμες κατάλληλες για εξόρυξη.

Data Mining: Στο βήμα αυτό εφαρμόζονται διάφοροι μέθοδοι εξόρυξης με σκοπό την εξαγωγή «κρυφών» μοτίβων μέσα στα δεδομένα.

Pattern Evaluation: Στο βήμα αυτό γίνεται αξιολόγηση των μοτίβων που έχουν προκύψει από το προηγούμενο βήμα.

Knowledge Presentation: Στο βήμα αυτό γίνεται παρουσίαση των αποτελεσμάτων (εξορυγμένης γνώσης) μέσω τεχνικών απεικόνισης.



Εικόνα 7. Βήματα για την ανακάλυψη γνώσης

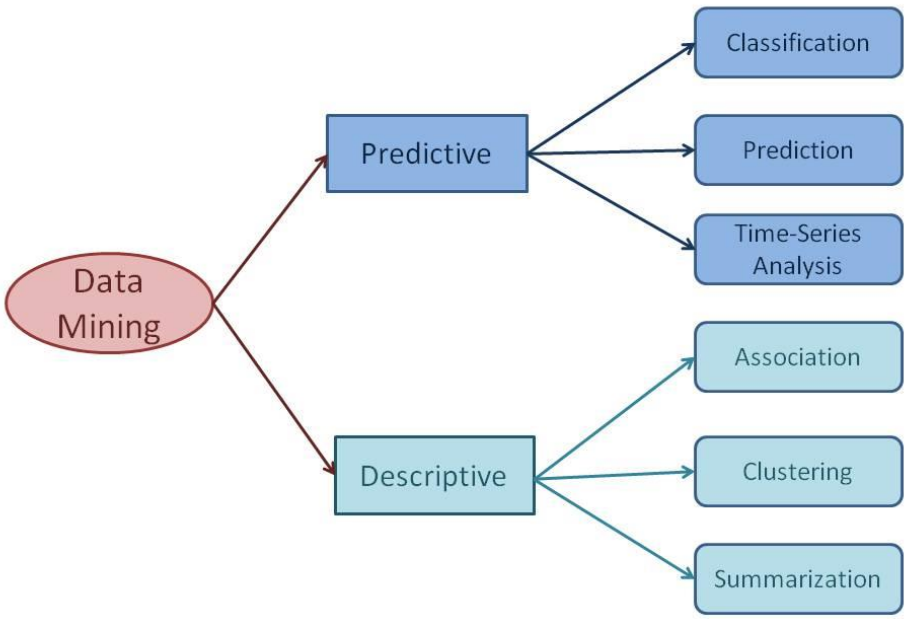
4.4 Κατηγορίες της εξόρυξης δεδομένων

Προγνωστικές εργασίες (Predictive tasks)

Ο σκοπός αυτής της μεθόδου είναι η πρόβλεψη της μελλοντικής τιμής ενός χαρακτηριστικού βάσει των τιμών άλλων χαρακτηριστικών. Τα χαρακτηριστικά που πρέπει να προβλεφθούν ονομάζονται στόχοι ή εξαρτημένες μεταβλητές ενώ τα χαρακτηριστικά που συμβάλλουν στην πρόβλεψη ονομάζονται ανεξάρτητες μεταβλητές. (Pang-Ning Tan, 2017,2010) Παράδειγμα μεθόδου πρόβλεψης είναι ο γιατρός που προσπαθεί να διαγνώσει μια ασθένεια με βάση τα αποτελέσματα των ιατρικών εξετάσεων ενός ασθενούς.

Περιγραφικές εργασίες (Descriptive tasks)

Στόχος αυτών των εργασιών είναι να βρεθούν πρότυπα-μοτίβα (τάσεις, συσχετισμοί, ανωμαλίες) που περιγράφουν τις βασικές σχέσεις που υπάρχουν στα δεδομένα. Πολλές φορές απαιτούνται τεχνικές μετεπεξεργασίας ώστε να επικυρωθούν τα δεδομένα. (Pang-Ning Tan, 2017,2010)



Εικόνα 8. Κατηγορίες της εξόρυξης δεδομένων.

Οι περιγραφικοί μέθοδοι χρησιμοποιούν λειτουργίες μάθησης χωρίς επιτήρηση (unsupervised) ενώ προγνωστικοί μέθοδοι χρησιμοποιούν τεχνικές μάθησης με εποπτεία (supervised). Αυτός είναι και ο λόγος για τον οποίο η περιγραφική ανάλυση δεν είναι σε θέση να προβλέψει άγνωστες-μελλοντικές τιμές αλλά επικεντρώνεται περισσότερο στην εγγενή διάταξη και τις σχέσεις. Αντίθετα η προγνωστικοί μέθοδοι καθορίζει ένα σύνολο δεδομένων για μελλοντική πρόβλεψη.

BASIS FOR COMPARISON	DESCRIPTIVE MINING	PREDICTIVE MINING
Basic	It identifies, what happened in the past by analyzing stored data	It describes, what can happen in the future with the help past data analysis.
Require	Data aggregation and data mining	Statistics and forecasting methods
Preciseness	Provides accurate data	Produces results does not ensure accuracy.
Type of approach	Reactive	Proactive
Practical analysis methods	Standard reporting, query/drill down and ad-hoc reporting.	Predictive modelling, forecasting, simulation and alerts.

Εικόνα 9.Σύγκριση προγνωστικής με περιγραφικής εξόρυξης.

4.5 Βασικές προσεγγίσεις της εξόρυξης δεδομένων

Κάποιες από τις βασικές εργασίες εξόρυξης δεδομένων είναι προγνωστική μοντελοποίηση, ανάλυση συστάδων, ανάλυση ανωμαλιών, ανάλυση συσχέτισης.

Προγνωστικό μοντέλο (predicting model)

Είναι μία διαδικασία που χρησιμοποιείται για την δημιουργία ενός μοντέλου με σκοπό την περιγραφή του «άγνωστου» χαρακτηριστικού (στόχου) ως συνάρτηση των «γνωστών» χαρακτηριστικών (επεξηγηματικών). (Pang-Ning Tan, 2017,2010) Ανάλογα με τον τύπο των μεταβλητών-χαρακτηριστικών χωρίζονται σε **κατηγοριοποίηση** (classification) για διακριτές μεταβλητές και σε **παλινδρόμηση** (regression) για συνεχείς μεταβλητές. Διακριτές τιμές αποτελούν ¨μεμονωμένες¨ αριθμητικές τιμές είναι δηλαδή στοιχεία ενός συνόλου τα οποία μπορούν να αντιστοιχηθούν ένα προς ένα με στοιχεία του συνόλου των θετικών ακέραιων αριθμών για παράδειγμα ο αριθμός των παιδιών μιας οικογένειας ή αριθμός των δωματίων μιας κατοικίας. Οι συνεχείς τιμές μπορούν να πάρουν αριθμητικές τιμές που καλύπτουν ολόκληρο διάστημα τιμών των πραγματικών αριθμών από το -∞ έως το + ∞ για παράδειγμα η θερμοκρασία ή η ηλικία.

Ανάλυση συσχέτισης (association analysis)

Χρησιμοποιείται για την ανακάλυψη συσχετίσεων μεταξύ αντικειμένων σε μεγάλο όγκο δεδομένων. Οι συσχετίσεις αυτές αναπαρίστανται με μορφή κανόνων συνεπαγωγής ή υποσυνόλων χαρακτηριστικών.

Οι κανόνες συσχέτισης προσφέρουν μία απεικόνιση συσχετίσεων της μορφής:

$$X \rightarrow Y \text{ (ύπαρξη συσχέτισης)}$$

Η ισχύ του κανόνα συσχέτισης μπορεί να μετρηθεί με βάση την υποστήριξη (support) και την εμπιστοσύνη (confidence). Η υποστήριξη καθορίζει το πόσο συχνά είναι εφαρμόσιμος ο κανόνας στο σύνολο δεδομένων, ενώ η εμπιστοσύνη καθορίζει πόσο συχνά τα αντικείμενα του Y εμφανίζονται σε περιπτώσεις που περιέχουν X. Με την εμπιστοσύνη μετράμε την αξιοπιστία του συμπεράσματος

που προκύπτει από τον κάθε κανόνα ενώ με την υποστήριξη διακρίνουμε τους κανόνες που προκύπτουν τυχαία. (Piatetsky-Shapiro, 1991)

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

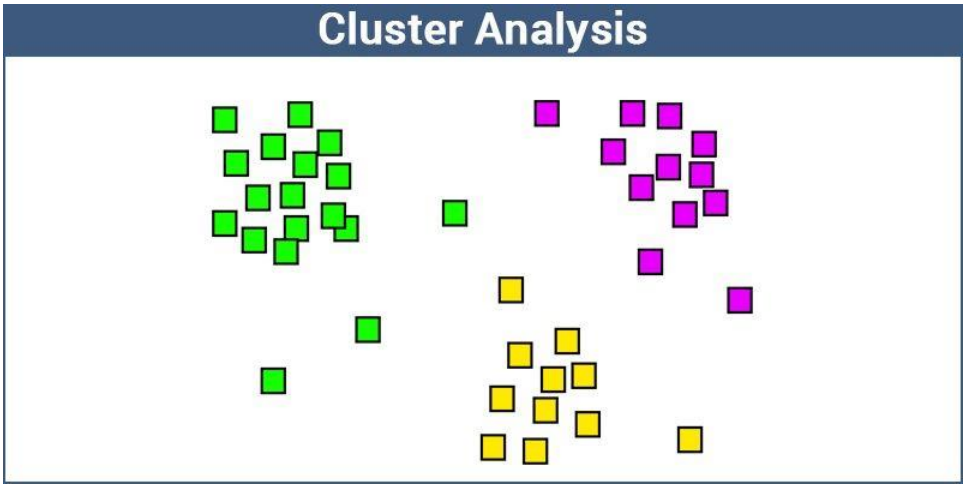
Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Εικόνα 10. Παράδειγμα ανάλυσης συσχετίσεων.

Ανάλυση συστάδων (clyster analysis)

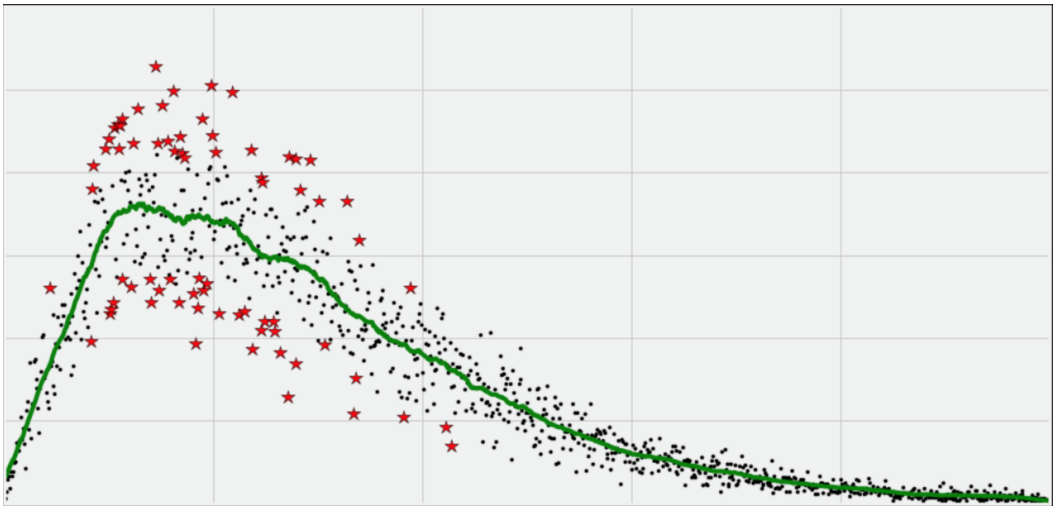
Είναι η διαδικασία με την οποία ομαδοποιούνται αντικείμενα της ίδιας ομάδας (cluster) με τέτοιο τρόπο ώστε να είναι περισσότερο παρόμοια μεταξύ τους παρά με εκείνα σε άλλες ομάδες (clusters). (Michael Hahsler, 2005) Όσο πιο μεγάλη είναι η ομοιογένεια μεταξύ των αντικειμένων μιας ομάδας τόσο πιο εμφανής είναι η συσταδοποίηση.



Εικόνα 11. Παράδειγμα ανάλυσης συστάδων.

Ανίχνευση ανωμαλιών (anomaly detection)

Στην ανίχνευση ανωμαλιών γνωστή και ως ανίχνευση αποκλίσεων (devation detection) στόχος είναι η εύρεση αντικειμένων-παρατηρήσεων που διαφέρουν σημαντικά από την πλειοψηφία των δεδομένων. (Victoria J. Hodge, n.d.) Τα διαφορετικά αυτά αντικείμενα είναι γνωστά ως ακραίες τιμές (outliers), θόρυβος (noise), αποκλίσεις (exceptions) και συνήθως στα διαγράμματα διασποράς των δεδομένων θα βρίσκονται πολύ μακριά από άλλα σημεία δεδομένων. (Choudhary, 2017) Χρησιμοποιώντας τα αποτελέσματα που θα προκύψουν (ανωμαλίες) μπορεί έπειτα να διακρίνει τις πραγματικές ανωμαλίες και να αποφύγει τον χαρακτηρισμό των ορθών αντικειμένων ως «ανωμαλίες». Ιδιαίτερα σημαντικό είναι ένας ανιχνευτής ανωμαλιών να διαθέτει υψηλό ρυθμό ανίχνευσης και χαμηλό ρυθμό εσφαλμένων «ειδοποιήσεων».



Εικόνα 12 Παράδειγμα ανίχνευσης ανωμαλιών

4.6 Γλώσσες προγραμματισμού και εργαλεία για εξόρυξη δεδομένων

Κάποιες από τις γλώσσες προγραμματισμού που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι:

Python: Η Python είναι μια ευρέως χρησιμοποιούμενη γλώσσα υψηλού επιπέδου στην κοινότητα των επιστημών δεδομένων που βασίζεται σε διερμηνέα. Η Python είναι μια ευέλικτη γλώσσα που διαθέτει μια μεγάλη ποικιλία βιβλιοθηκών για πολλαπλούς ρόλους. Έχει αναδειχθεί ως μία από τις πιο δημοφιλείς επιλογές για την εξόρυξη δεδομένων λόγω της ευκολίας της στην μάθηση και των χρήσιμων βιβλιοθηκών (Pandas, Numpy, scikit-learn). Η αναγνωσιμότητα του κώδικα που παρατηρείται από την Python καθιστά επίσης μια δημοφιλή επιλογή για την εξόρυξη δεδομένων.

R: Η γλώσσα R αποτελεί ένα από τα πιο συχνά χρησιμοποιούμενα εργαλεία στην εξόρυξη δεδομένων και στην χρηματοοικονομική μοντελοποίηση. Αποκαλύφθηκε το 1997 ως υποκατάστατο του Matlab. Παρέχει εργαλεία για στατιστική ανάλυση. Επίσης επιτρέπει την δημιουργία γραφικών υψηλού επιπέδου για την οπτικοποίηση των δεδομένων καθώς και διασύνδεση προς άλλες γλώσσες. Το μόνο μειονέκτημα του R είναι ότι δεν είναι μια γλώσσα προγραμματισμού γενικής χρήσης που σημαίνει ότι δεν χρησιμοποιείται για εργασίες εκτός του στατιστικού προγραμματισμού. (Bachheriya, 2019)

Julia: Η Julia αποτελεί μια εκφραστική γλώσσα υψηλού επιπέδου που καλύπτει τα κενά της Python και της R. Έχει ως στόχο να προσφέρει την ευκολία και την παραγωγικότητα της Python με τη μαθηματική αντοχή της R και την απόδοση της C. Λόγω της ταχύτερης εκτέλεσης, η Julia έχει γίνει μια τέλεια επιλογή για την αντιμετώπιση πολύπλοκων έργων που περιέχουν μεγάλο όγκο δεδομένων. Παρ ‘όλα αυτά βρίσκεται ακόμα σε πρώιμο στάδιο και χρειάζεται ακόμα αρκετές προσθήκες για να φτάσει στο ίδιο επίπεδο με την Python και την R. (Rao, 2018)

Scala: Αποτελεί μια επέκταση της γλώσσας προγραμματισμού Java. Πρόκειται για μια γενική γλώσσα προγραμματισμού που έχει χαρακτηριστικά μιας αντικειμενοστραφούς γλώσσας. Χρησιμοποιείτε συχνά με το Apache Spark (πλατφόρμα δεδομένων) πράγμα που την καθιστά ιδανική γλώσσα για μεγάλους όγκους δεδομένων. Ένα σημαντικό χαρακτηριστικό της Scala είναι η ικανότητά της να διευκολύνει την παράλληλη επεξεργασία σε μεγάλη κλίμακα.

RapidMiner: Αποτελεί ένα από τα πιο δημοφιλή εργαλεία για την εξόρυξη δεδομένων. Είναι γραμμένο σε Java αλλά δεν απαιτεί από τον χρήστη να γνωρίζει κώδικα για να το χρησιμοποιήσει. Επιπλέον παρέχει διάφορες λειτουργίες εξόρυξης δεδομένων, όπως προεπεξεργασία δεδομένων, εκπροσώπηση δεδομένων, φιλτράρισμα, ομαδοποίηση κλπ.

Apache Mahout: Το Mahout αποτελεί επέκταση της Hadoop Big Data Platform της Apache και αναπτύχθηκε προκειμένου να αντιμετωπίσουν την αυξανόμενη ανάγκη για εξόρυξη δεδομένων. Ως αποτέλεσμα, περιλαμβάνει διάφορες λειτουργίες μάθησης μηχανών όπως ταξινόμηση, παλινδρόμηση, ομαδοποίηση κλπ.

Weka: Το Weka είναι ένα λογισμικό εξόρυξης δεδομένων ανοικτού κώδικα. Όπως και το RapidManer δεν απαιτεί από τον χρήστη να γνωρίζει κάποια γλώσσα προγραμματισμού για την λειτουργία του καθώς διαθέτει ένα απλό στη χρήση GUI. Μέσω του Weka μπορεί κάποιος να χρησιμοποιεί αλγορίθμους εκμάθησης μηχανών απευθείας ή να τις εισάγει μέσω κώδικα Java. Παρέχει μια ποικιλία εργαλείων όπως οπτικοποίηση, προεπεξεργασία, ταξινόμηση, ομαδοποίηση κλπ.

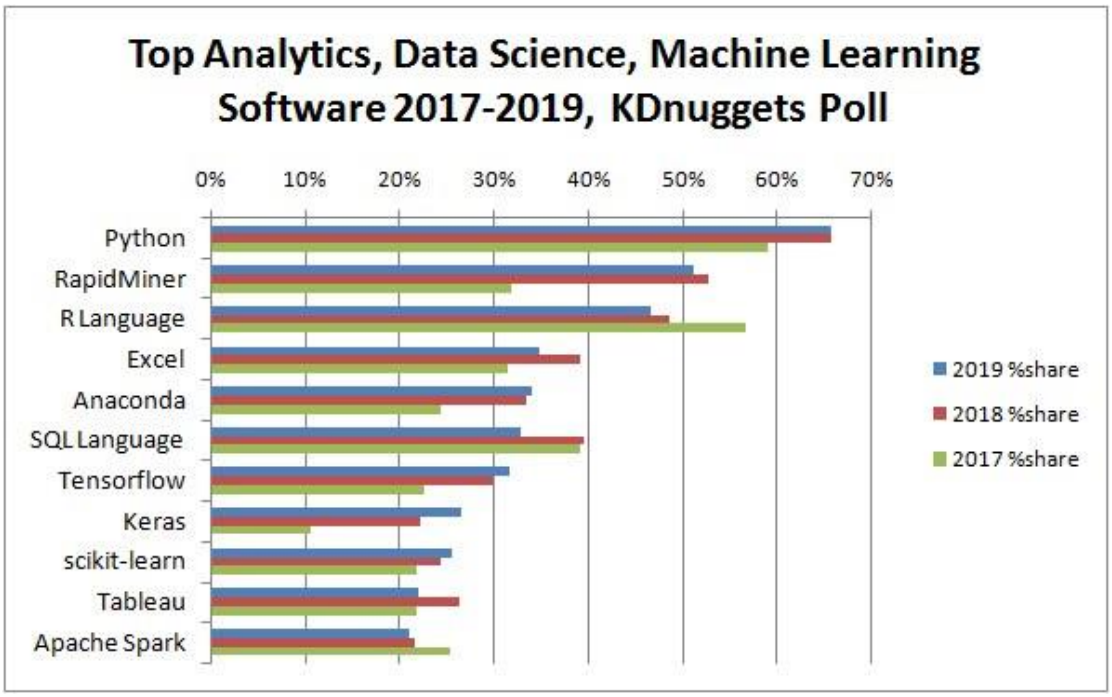
TeraData: Το TeraData γνωστό και ως βάση δεδομένων TeraData παρέχει υπηρεσίες αποθήκευσης που αποτελούνται από εργαλεία εξόρυξης δεδομένων. Μπορεί να αποθηκεύει δεδομένα με βάση τη χρήση τους, δηλαδή να αποθηκεύει δεδομένα που χρησιμοποιούνται λιγότερο συχνά στην 'αργή' του ενότητα και παρέχει γρήγορη πρόσβαση σε δεδομένα που χρησιμοποιούνται συχνά.

Oracle DataMining: Το Oracle Datamining είναι ένα εξαιρετικό εργαλείο για την ταξινόμηση, ανάλυση και πρόβλεψη δεδομένων. Επιτρέπει στους χρήστες της να

εκτελούν εκροές δεδομένων στις βάσεις δεδομένων SQL για να εξαγάγουν προβολές και σχήματα.

Orange: Αποτελεί διάσημο λογισμικό γραμμενό σε Python για την ενσωμάτωση εργαλείων μηχανικής μάθησης και εξόρυξης δεδομένων και προσφέρει διαδραστικές και αισθητικές απεικονίσεις στους χρήστες του. (Team, 2019)

Εικόνα 13 "Εργαλεία" για την εξόρυξη δεδομένων



Εικόνα 14 Συγκριτικός πίνακας χρήσης των διαφόρων εργαλείων για την εξόρυξη δεδομένων

3. Τεχνικές της εξόρυξης δεδομένων

3.1 Κατηγοριοποίηση

Κατηγοριοποίηση είναι η διαδικασία εκμάθησης μιας συνάρτησης-στόχου (target function) η οποία αντιστοιχίζει κάθε σύνολο χαρακτηριστικών x σε μια από τα προκαθορισμένες ετικέτες κατηγορίας y. Η κατηγοριοποίηση ξεκινά με μια συλλογή από εγγραφές (δεδομένα κατάρτισης-training data). Κάθε εγγραφή η οποία ονομάζεται στιγμιότυπο ή δείγμα χαρακτηρίζεται από ένα ζεύγος (x,y) όπου x είναι το σύνολο των χαρακτηριστικών και y ένα ειδικό χαρακτηριστικό στόχος το οποίο ορίζεται ως ετικέτα κλάσης. Το σύνολο των χαρακτηριστικών x μπορεί να είναι είτε διακριτά είτε συνεχή αλλά η ετικέτα κλάσης είναι πρέπει να είναι διακριτό χαρακτηριστικό, ενώ στην παλινδρόμηση είναι συνεχές χαρακτηριστικό. Αυτό αποτελεί και μια βασική διαφορά ανάμεσα στην κατηγοριοποίηση και στην παλινδρόμηση. Η συνάρτηση-στόχος ονομάζεται και μοντέλο κατηγοριοποίησης (classification model). (Pang-Ning Tan, 2017,2010)

Η κατηγοριοποίηση δεδομένων αποτελείται από δύο μέρη, από το βήμα μάθησης (learning step) και από βήμα κατηγοριοποίησης (classification step). Παραδείγματα τεχνικών που χρησιμοποιούν μοντέλα κατηγοριοποίησης είναι οι κατηγοριοποιητές δένδρων απόφασης, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης και οι κατηγοριοποιητές Bayes. (Pang-Ning Tan, 2017,2010) Στο βήμα μάθησης χρησιμοποιείται ένας αλγόριθμος μάθησης (learning algorithm) για την δημιουργία ενός μοντέλου που ταιριάζει καλύτερα στη σχέση μεταξύ του συνόλου χαρακτηριστικών και της ετικέτας κλάσης των δεδομένων εισόδου. Το μοντέλο αυτό κατασκευάζεται από ένα σύνολο εγγραφών των οποίων οι ετικέτες κλάσης είναι γνωστές και ονομάζεται σύνολο εκπαίδευσης (training set). Το μοντέλο αυτό ονομάζεται και ταξινομητής. Έπειτα ο ταξινομητής εφαρμόζεται σε ένα σύνολο ελέγχου, το οποίο χρησιμοποιείται για να υπολογιστεί η ακρίβεια των κανόνων κατηγοριοποιήσεις. Αν η ακρίβεια των κανόνων είναι υψηλή μπορούν να χρησιμοποιηθούνε και σε καινούργια σύνολα δεδομένων. Η εκτίμηση για την ακρίβεια-απόδοση του μοντέλου κατηγοριοποίησης, βασίζεται στο πλήθος εγγραφών ελέγχου που έχουν προβλεφθεί σωστά και λανθασμένα από το μοντέλο.

Σαν μέτρο απόδοσης ορίζεται η ακρίβεια :

$$Accuracy = \frac{NC}{T}$$

Όπου NC είναι το πλήθος των σωστών προβλέψεων και T το συνολικό πλήθος των προβλέψεων.

Ισοδύναμα η απόδοση του μοντέλου μπορεί να εκφραστεί με βάση τον ρυθμό σφάλματος (error rate) :

$$Error\ rate = \frac{NW}{T}$$

Όπου NW είναι το πλήθος των λανθασμένων προβλέψεων και T το συνολικό πλήθος των προβλέψεων.

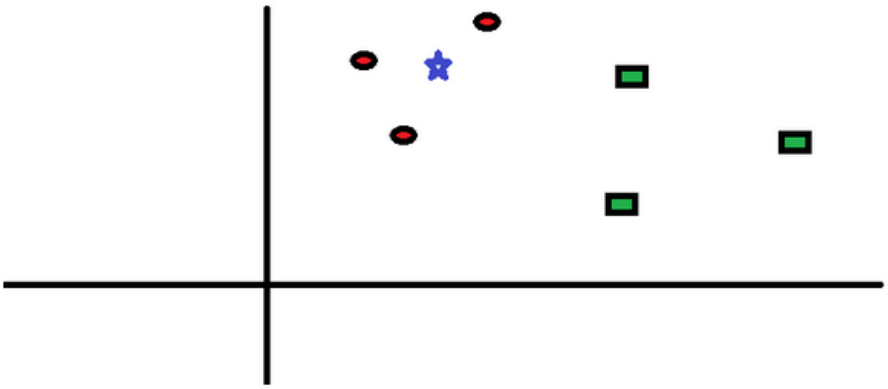
Ανάλογα με το μέτρο που χρησιμοποιείται για την απόδοση τα μοντέλα επιτυγχάνουν τη μεγαλύτερη ακρίβεια ή ισοδύναμα το μικρότερο ρυθμό σφάλματος.

3.1.1 Αλγόριθμος κοντινότερου γείτονα

Ο αλγόριθμος του κοντινότερου γείτονα (K Nearest Neighbors-KNN) αναπαριστά κάθε εγγραφή ενός συνόλου δεδομένων ως σημεία δεδομένων σε ένα χώρο d διαστάσεων, όπου d είναι το πλήθος των χαρακτηριστικών. Ο KNN ανήκει στους αλγορίθμους που ονομάζονται απρόθυμοι μαθητές (lazy learners) δηλαδή καθυστερούν την διαδικασία δημιουργίας μοντέλου μέχρι να γίνει απαραίτητο για την κατηγοριοποίηση των δειγμάτων ελέγχου. Ο αλγόριθμος υπολογίζει την απόσταση ανάμεσα σε κάθε δείγμα για να καθορίσει τη λίστα των πλησιέστερων γειτόνων. (Pang-Ning Tan, 2017,2010) Έπειτα ανάλογα με την τιμή του K διαλέγονται οι K κοντινότεροι γείτονες. Αν το πλήθος των κοντινότερων γειτόνων είναι ένα τότε το σημείο δεδομένου αποδίδεται στην ίδια κατηγορία με του γείτονα. Αν το πλήθος των κοντινότερων γειτόνων ξεπερνάει τους 2 τότε στο σημείο αποδίδεται η κατηγόρια στην οποία ανήκουν οι περισσότεροι κοντινότεροι γείτονες (στρατηγική της ψήφου πλειοψηφίας).

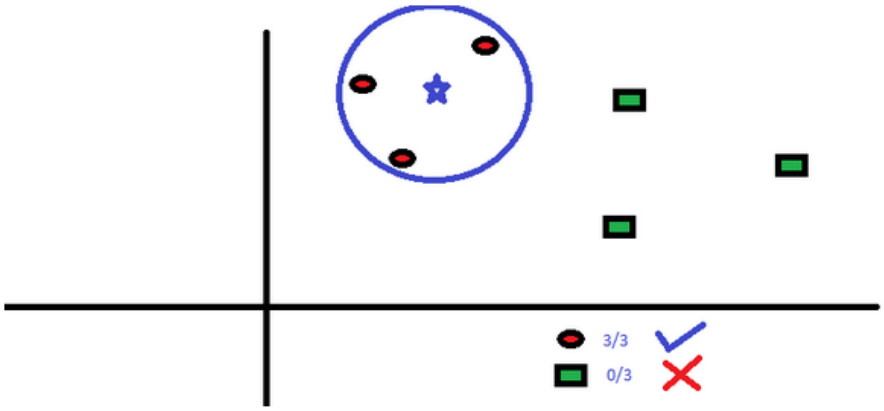
Παράδειγμα λειτουργίας του KNN:

Στην παρακάτω εικόνα έχουμε ένα σύνολο δεδομένων. Γνωρίζουμε την κατηγορία των κόκκινων κύκλων (κκ) και των πράσινων τετραγώνων και θέλουμε να βρούμε την κατηγορία του μπλέ αστεριού (μα). Το αστέρι μπορεί να είναι είτε κκ είτε μα.



Εικόνα 15 Παράδειγμα πριν την εκτέλεση του αλγορίθμου KNN.

Έστω ότι διαλέγουμε το K=3 οπότε μέσα στον κύκλο θα βρίσκονται τα 3 κοντινότερα σημεία στο σημείο ελέγχου (αστέρι). Τα 3 κοντινότερα σημεία είναι όλα κόκκινοι κύκλοι οπότε μπορούμε ότι και το νέο σημείο (μπλε αστέρι) ανήκει στην κατηγορία των κόκκινων κύκλων.



Εικόνα 16. Παράδειγμα μετά το τέλος του αλγορίθμου KNN.

Πλεονεκτήματα:

- Ο αλγόριθμος είναι εύκολος και απλός για να εφαρμοστεί.
- Δεν χρειάζεται η δημιουργία μοντέλου κατηγοριοποιήσεις
- Ο αλγόριθμος είναι ευέλικτος. Μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση και αναζήτηση

Μειονεκτήματα:

- Ο αλγόριθμος καθυστερεί σημαντικά καθώς το πλήθος των δεδομένων αυξάνεται.

3.2 Ομαδοποίηση

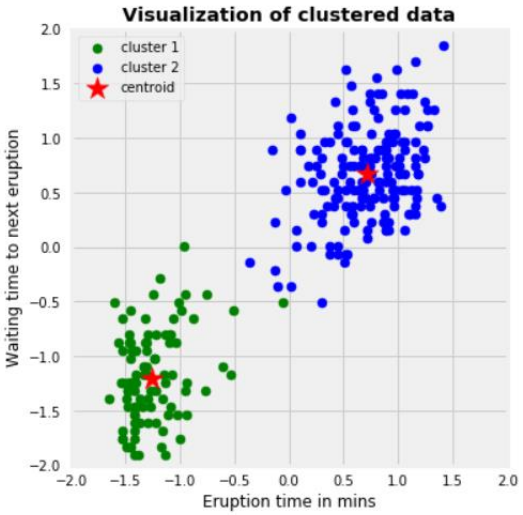
Αποτελεί μια από τις πιο κοινές τεχνικές ανάλυσης δεδομένων. Βασίζεται στον διαχωρισμό των στοιχείων σε ομάδες έτσι ώστε τα στοιχεία που ανήκουν στην ίδια υποομάδα (σύμπλεγμα) να είναι όσο το δυνατόν πιο παρόμοια, ενώ τα στοιχεία άλλων ομάδων να είναι πολύ διαφορετικά σε σχέση με τις άλλες ομάδες σύμφωνα με ένα μέτρο ομοιότητας. Το μέτρο ομοιότητας μπορεί να είναι η ευκλείδεια απόσταση ή η απόσταση που βασίζεται στην συσχέτιση. Η τεχνική αυτή ανήκει στους αλγορίθμους μάθησης χωρίς εποπτεία καθώς δεν γνωρίζουμε την πραγματική κατάσταση προκειμένου να την συγκρίνουμε με την έξοδο του αλγορίθμου για να αξιολογήσουμε την απόδοσή του, θέλουμε απλώς να διερευνήσουμε τη δομή των δεδομένων ομαδοποιώντας τα σημεία δεδομένων σε ξεχωριστές υποομάδες.

3.2.1 Αλγόριθμος K-Means

Ο αλγόριθμος K-Means είναι ένας επαναληπτικός αλγόριθμος ο οποίος προσπαθεί να χωρίσει το σύνολο δεδομένων σε ένα καθορισμένο αριθμό K μη επικαλυπτόμενων υποομάδων όπου κάθε σημείο θα ανήκει σε μία μόνο ομάδα. Τα σημεία δεδομένων εντός του συμπλέγματος θα είναι όσο το δυνατόν πιο παρόμοια μεταξύ τους διατηρώντας ταυτόχρονα τις υποομαδες-συστάδες όσο το δυνατόν πιο μακρία (διαφορετικές). Αναθέτει ένα σημείο-δεδομένο σε μία συστάδα με τέτοιο τρόπο ώστε το άθροισμα της τετραγωνικής απόστασης μεταξύ των σημείων δεδομένων και του αριθμητικού μέσου όλων των σημείων δεδομένων που ανήκουν σε αυτή την συστάδα (centroid) να είναι το ελάχιστο.

Τα βήματα του αλγορίθμου K-Means είναι:

1. Ομαδοποιεί τα δεδομένα σε K ομάδες.
2. Επιλέγει τυχαία σημεία K ως κέντρα των ομάδων-συμπλεγμάτων.
3. Αντιστοιχίζει τα αντικείμενα στο πλησιέστερο κέντρο συμπλέγματος σύμφωνα με τη συνάρτηση ευκλείδειων αποστάσεων.
4. Υπολογίζει το centroid σε κάθε σύμπλεγμα (μέσο όλων των αντικειμένων σε κάθε σύμπλεγμα).
5. Επανάληψη των βημάτων 2, 3 και 4 έως ότου αντιστοιχιστούν τα ίδια σημεία σε κάθε ομάδα και δεν θα υπάρχει καμία αλλαγή στα centroids.

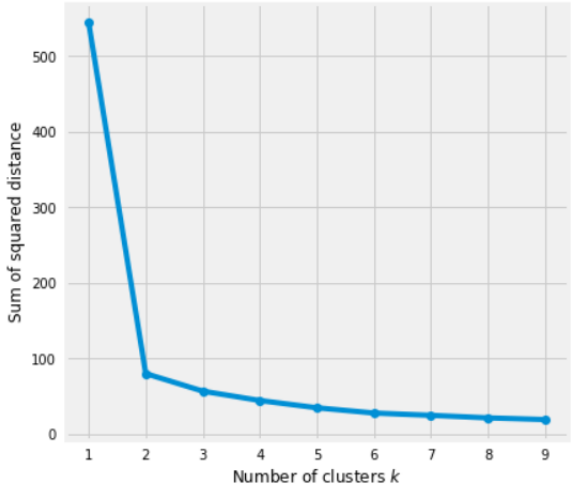


Εικόνα 17 Παράδειγμα ομαδοποίησης K-Means με 2 centroid

Στον αλγόριθμο K-Means δεν υπάρχει μια σταθερή μέτρηση αξιολόγησης για να αξιολογήσουμε τα αποτελέσματα του αλγορίθμου. Στον αλγόριθμο αυτό αξιολογούμε πόσο καλά αποδίδει το μοντέλο με βάση τον αριθμό K των διαφορετικών συστάδων. Ένας τρόπος για να αξιολογήσουμε αν έχουμε επιλέξει τον σωστό αριθμό K συστάδων είναι η μέθοδος αγκώνα (Elbow method). (Dabbura, 2018)

Μέθοδος αγκώνα (Elbow method)

Με την μέθοδο αγκώνα μπορούμε να βρούμε την καλύτερο δυνατό αριθμό των K συστάδων. Βασίζεται στο άθροισμα της τετραγωνικής απόστασης μεταξύ των σημείων δεδομένων και των κεντροειδών (centroids) τους που έχουν εκχωρηθεί. Διαλέγουμε τον αριθμό K στο σημείο όπου το SSE αρχίζει να ισιώνει και σχηματίζει έναν αγκώνα. Μερικές φορές είναι ακόμα δύσκολο να καταλάβουμε έναν καλό αριθμό συστάδων που πρέπει να χρησιμοποιηθούν, επειδή η καμπύλη μειώνεται μονοτονικά και μπορεί να μην δείχνει κανένα αγκώνα

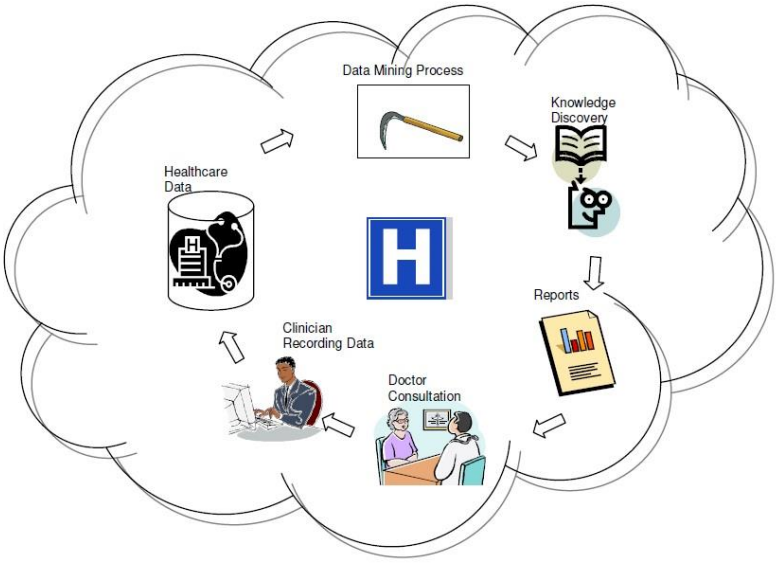


Εικόνα 18 Αναπαράσταση του SSE και εύρεση του σημείου "αγκώνα"

6. Εφαρμογές της εξόρυξης δεδομένων

Εξόρυξη δεδομένων στον τομέα της υγειονομικής περίθαλψης

Η εξόρυξη δεδομένων χρησιμοποιείται σε μεγάλο βαθμό και στον τομέα της σύγχρονης ιατρικής καθώς έχει μεγάλες δυνατότητες βελτίωσης των συστημάτων υγείας. Οι ερευνητές έχοντας όλες τις πληροφορίες του ασθενούς, όπως τα ιατρικά αρχεία, οι φυσικές εξετάσεις και τα μοτίβα θεραπείας, χρησιμοποιούν προσεγγίσεις εξόρυξης δεδομένων όπως πολυδιάστατες βάσεις δεδομένων, μηχανική μάθηση, στατιστική και οπτικοποίηση δεδομένων για την σύγκριση και την αντίθεση των συμπτωμάτων, των αιτιών καθώς και της θεραπείας με σκοπό να βρεθεί η πιο αποτελεσματική αντιμετώπιση για μια συγκεκριμένη ασθένεια ή κατάσταση. (Health, χ.χ.) Επιτρέπει την αποδοτικότερη και αποτελεσματικότερη διαχείριση των πόρων υγείας, εντοπίζοντας κινδύνους, προβλέποντας ασθένειες που είναι πιο πιθανές σε ορισμένα τμήματα του πληθυσμού (π.χ. σε ηλικιωμένους) ακόμα και τη διάρκεια νοσηλείας στο νοσοκομείο. (Anon., χ.χ.) Επιπλέον συμβάλει στην ανίχνευση της απάτης και της κατάχρησης προς τις ασφαλιστικές υγειονομικής περίθαλψης με τον εντοπισμό ασυνήθιστων μορφών ιατρικών απαιτήσεων από κλινικές, γιατρούς, εργαστήρια ή άλλους ακόμη και παράνομων συνταγών. Η Google το 2009 κατάφερε να προβλέψει με επιτυχία την εξάπλωση της χειμερινής γρίπης συγκρίνοντας τις αναζητήσεις που έγιναν με την λέξη γρίπη την περίοδο 2003-2008 δημιουργώντας ένα στατιστικό μοντέλο.



Εικόνα 19. Η εξόρυξη δεδομένων στον τομέα της υγείας.

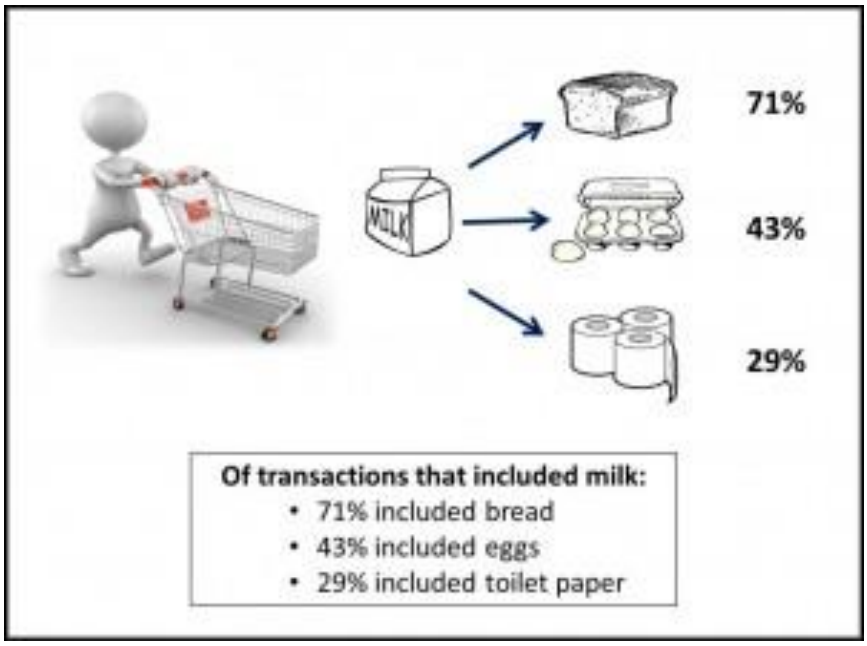
Εξόρυξη δεδομένων στον τομέα των αγορών

Οι τεχνικές εξόρυξης δεδομένων επιτρέπουν στις επιχειρήσεις να κατανοήσουν κρυμμένα μοτίβα μέσα στα δεδομένα συναλλαγών συμβάλλοντας έτσι στον σχεδιασμό νέων εκστρατειών μάρκετινγκ με άμεσο και οικονομικό τρόπο. Με την χρήση αυτών των τεχνικών μπορεί να προσδιοριστεί η αποτελεσματικότητα μιας συγκεκριμένης προώθησης σε διαφορετικά μέσα ενημέρωσης ή ακόμη σε διαφορετικές γεωγραφικές τοποθεσίες. (Choudhary, 2017) Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να εφαρμοστεί για να ελέγξει ποιο τμήμα πελατών ανταποκρίνεται θετικά σε μια προσφορά, πόσο αποτελεσματική είναι η προώθηση σε σχέση με το κόστος και τα οφέλη, ποια κανάλια μέσων έχουν πετύχει για διαφορετικές καμπάνιες στο παρελθόν κ.ο.κ.. Με την ανάλυση των πληροφοριών οι εταιρείες μπορούν να προσδιορίσουν τα πρότυπα αγοραστικής συμπεριφοράς των πελατών και στη συνέχεια να καταλήξουν σε πιο αποτελεσματικές προωθήσεις και διαφημίσεις.

Εικόνα 20. Η εξόρυξη δεδομένων στον τομέα των αγορών.

Ανάλυση καλαθιού αγοράς (Market Basket Analysis)

Η ανάλυση του καλαθιού αγοράς αποτελεί μια τεχνική μοντελοποίησης που βασίζεται στην θεωρία οτι αν κάποιος αγοράσει μια συγκεκριμένη ομάδα προιόντων είναι πιθανό να αγοράσει και μια άλλη ομάδα προιόντων. Αυτή η τεχνική μπορεί να επιτρέψει στον ιδιοκτήτη του καταστήματος να κατανοήσει την συμπεριφορά ενός αγοραστή και τις ανάγκες του. Χρησιμοποιόντας αυτές τις πληροφορίες ο ιδιοκτήτης του καταστήματος μπορεί να προβεί σε «στρατηγικές» αλλαγές προς οφελός του π.χ. αλλαγή στην διάταξη του καταστήματος. (Bell, n.d.)



Εικόνα 21. Ανάλυση καλαθιού αγοράς.

Στην ανάλυση του καλαθιού αγοράς συνήθως χρησιμοποιούνται οι κανόνες σύνδεσης (association rules) και αποσκοπούν στον εντοπισμό ισχυρών κανόνων μέσα από τα δεδομένα συναλλαγών.

5. Υλοποίηση εφαρμογής

Σκοπός της υλοποίησης είναι η εφαρμογή αλγορίθμων εξόρυξης δεδομένων με την χρήση τεχνικών μηχανικής μάθησης για την πρόβλεψη ορισμένων τιμών-χαρακτηριστικών. Η υλοποίηση αναπτύχθηκε σε γλώσσα python και έγινε χρήση των βιβλιοθηκών matplotlib, scikit-learn.

Η γλώσσα python

Η γλώσσα python είναι μια γλώσσα υψηλού επιπέδου υποστηρίζει τόσο το διαδικαστικό όσο και το αντικειμενοστραφές προγραμματισμό. Δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ στο ερευνητικό κέντρο Centrum Wiskunde & Informatica το 1989 και κυκλοφόρησε για πρώτη φορά το 1991. Το μεγαλύτερο πλεονέκτημα της γλώσσα αυτής είναι η αναγνωσιμότητα του κώδικα και η ευκολία χρήσης της. Το συντακτικό της και το πλήθος των βιβλιοθηκών που διαθέτει επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ότι θα ήταν δυνατόν σε γλώσσες όπως η Java και η C++.

Η βιβλιοθήκη scikit-learn

Η βιβλιοθήκη scikit-learn αποτελεί μια βιβλιοθήκη μηχανικής μάθησης και βασίζεται στην βιβλιοθήκη SciPy. Αναπτύχθηκε αρχικά από τον David Cournapeau ως project στο Google Summer of Code το 2007. Διαθέτει διάφορους αλγορίθμους ταξινόμησης, παλινδρόμησης και συστοιχίας, K μέσων, DBSCAN. Επίσης συνεργάζεται με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy.

Η βιβλιοθήκη matplotlib

Η βιβλιοθήκη matplotlib είναι μια βιβλιοθήκη για την σχεδίαση και την ενσωμάτωση γραφημάτων σε εφαρμογές που χρησιμοποιούν GUI μέσω αντικειμενοστραφούς API. Η βιβλιοθήκη γράφτηκε αρχικά από τον John D.Hunter. Μπορεί να δημιουργήσει ιστογράμματα, φάσματα ισχύος, πίνακες σφαλμάτων και πολλά ακόμη με ελάχιστες γραμμές κώδικα.

5.1 Δεδομένα της εφαρμογής

Τα δεδομένα που θα χρησιμοποιηθούν στην εφαρμογή προέρχονται από ειδικούς αισθητήρες που έχουν τοποθετηθεί σε πλοία και ονομάζονται VDR (Voyage Data Recorder). Το VDR αποτελεί ένα σύστημα καταγραφής δεδομένων το οποίο πρέπει να φέρουν όλα σχεδόν τα σκάφη που κατασκευάστηκαν από την 1η Ιουλίου 2002 και μετά προκειμένου να συμμορφώνονται με τις απαιτήσεις της σύμβασης για την ασφάλεια της ζωής στη θάλασσα SOLAS (Safety of Life at Sea). Το VDR συνδέεται με διάφορους αισθητήρες του σκάφους και μπορεί να καταγράψει δεδομένα όπως την τοποθεσία GPS του σκάφους, την ταχύτητα και την κατεύθυνση του ανέμου, τη θερμοκρασία, το βάθος, τη στάθμη της μπαταριάς, των δεξαμενών νερού, καυσίμων και πολλών άλλων δεδομένων. Οι πληροφορίες αποθηκεύονται σε ασφαλή και ανακτήσιμη μορφή σε μία προστατευτική μονάδα αποθήκευσης που παρέχει προστασία από παραβίαση και από άλλα περιστατικά όπως φυσικά που μπορούν να συμβούν στην θάλασσα περιστατικά (πυρκαγιά, έκρηξη, σύγκρουση, βύθιση κλπ.). Αν και ο πρωταρχικός σκοπός των δεδομένων του VDR είναι για την διερεύνηση των παραγόντων που συνέβαλλαν σε κάποιο ατύχημα οι πληροφορίες αυτές μπορεί επίσης να χρησιμοποιηθούν για προληπτική συντήρηση, παρακολούθηση της αποτελεσματικότητας τη απόδοσης, ανάλυση βλαβών από καιρικές συνθήκες, αποφυγή ατυχημάτων και εκπαιδευτικούς σκοπούς για την βελτίωση της ασφάλειας καθώς κα για την μείωση του κόστος λειτουργίας.

Συγκεκριμένα τα δεδομένα που έχουμε από το VDR είναι:

Time: Timestamp.

Latitude: Global Position System Latitude coordinates.

Longitude: Global Position System Longitude coordinates.

TWS: True Wind Speed current value.

TWS (med): True Wind Speed median value.

TWS (avg): True Wind Speed average value.

TWS (min): True Wind Speed minimum value.

TWS (max): True Wind Speed maximum value.

TWD: True Wind Direction current value.

TWD (med): True Wind Direction median value.

TWD (avg): True Wind Direction average value.

TWD (min): True Wind Direction minimum value.

TWD (max): True Wind Direction maximum value.

TWA: True Wind Angle current value.

TWA (med): True Wind Angle median value.

TWA (avg): True Wind Angle average value.

TWA (min): True Wind Angle minimum value.

TWA (max): True Wind Angle maximum value.

AWS (med): Apparent Wind Speed current value.

AWS (avg): Apparent Wind Speed average value.

AWS (min): Apparent Wind Speed minimum value.

AWS (max): Apparent Wind Speed maximum value.

AWA (med): Apparent Wind Angle current value.

AWA (avg): Apparent Wind Angle average value.

AWA (min): Apparent Wind Angle minimum value.

AWA (max): Apparent Wind Angle maximum value.

Heading: Heading of the ship current (compass direction) current value.

Heading (min): Heading of the ship (compass direction) minimum value.

Heading (max): Heading of the ship (compass direction) maximum value.

Pressure: Pressure current value.

Pressure (min): Pressure minimum value.

Pressure (max): Pressure maximum value.

Air: Air temperature current value.

Air (min): Air temperature minimum value.

Air (max): Air temperature maximum value.

Humidity Outside: Humidity Outside current value.

Humidity Outside (min): Humidity Outside minimum value.

Humidity Outside (max): Humidity Outside maximum value.

Από όλα αυτά τα δεδομένα που περιέχει το dataset της εφαρμογής μας στην εφαρμογής μας θα χρησιμοποιήσουμε μόνο κάποια συγκεκριμένα χαρακτηριστικά τα οποία είναι TWD (avg), Humidity Outside, Air, Pressure, TWS (med).

Επεξεργασία των δεδομένων

Αρχικά θα πρέπει να γίνει εκκαθάριση του θορύβου που μπορεί να υπάρχει στα δεδομένα όπως μηδενικές μετρήσεις ή ακόμη και ελλιπείς μετρήσεις διότι αυτές θα επηρεάσουνε αρνητικά την διαμόρφωση του μοντέλου και γενικότερα τα αποτελέσματα της εφαρμογής. Το τελικό αρχείο με τα διορθωμένα δεδομένα ονομάζεται all_data(fixed).csv. Να σημειωθεί ότι μετρήσεις είναι με διάστημα 15 λεπτών.

Στον κώδικα διαβάζουμε το αρχείο αυτό μέσω της συνάρτησης read_csv της βιβλιοθήκης pandas και το μετατρέπει σε μία δομή dataframe. Στην συνάρτηση read_csv δηλώνουμε κάποια ορίσματα όπως το ποιο θα είναι το διαχωριστικό στο αρχείο δεδομένων και ποιες στήλες θα χρησιμοποιήσουμε. Στην συνέχεια μετατρέπουμε το dataframe αυτό σε πίνακα numpy για πιο εύκολη διαχείριση των δεδομένων.

```
###################Load_the_data#############################
step=48 ##step=4 for 1 hours interval ##step=24 for 6 hours interval
##step=48 for 12 hours interval
datasheet
=pd.read_csv("all_data(fixed).csv",sep=",",header=0,usecols=["TWD(avg)","Hu
midityOutside","Air","Pressure","TWS(med)"]) #Anagnwsi twn dedomenwn apo to
csv
np_datasheet=datasheet[["TWD(avg)","HumidityOutside","Air","Pressure","TWS(
med)"]].to_numpy(dtype="float64") #allagi tis seiras twn column kai
metatropi se numpy array
```

Έπειτα θα πρέπει να χωρίσουμε το dataset δηλαδή τον numpy πίνακα σε train set και test set. Τα δεδομένα που χρησιμοποιούμε χωρίζονται συνήθως σε δεδομένα εκπαίδευσης και δεδομένα δοκιμών. Το σετ κατάρτισης περιέχει μια γνωστή έξοδο και το μοντέλο χτίζεται πάνω σε αυτά τα δεδομένα προκείμενου να γίνει

γενικευμένο σε άλλα δεδομένα αργότερα. Το σετ δοκιμών (test set) χρησιμοποιείται για να ελέγξουμε επαληθεύσουμε την πρόβλεψη του μοντέλου μας.

```
X=np_datasheet[:,:-1]
X=X.astype(np.float64)

y=np_datasheet[:,4]
y=y.astype(np.float64)
X_train,X_test,y_train,y_test = train_test_split(X, y,
train_size=0.70,shuffle=False)
```

Οι πρώτες 4 στήλες (TWD(avg), Humidity Outside, Air, Pressure) αποτελούν το train set ενώ η τελευταία στήλη αποτελεί το test set.

5.2 Αλγόριθμος της εφαρμογής

Στην υλοποίηση θα εφαρμόσουμε τον αλγόριθμο Gradient Boosting Regressor που παρέχεται μέσα από την βιβλιοθήκη scikit-learn.

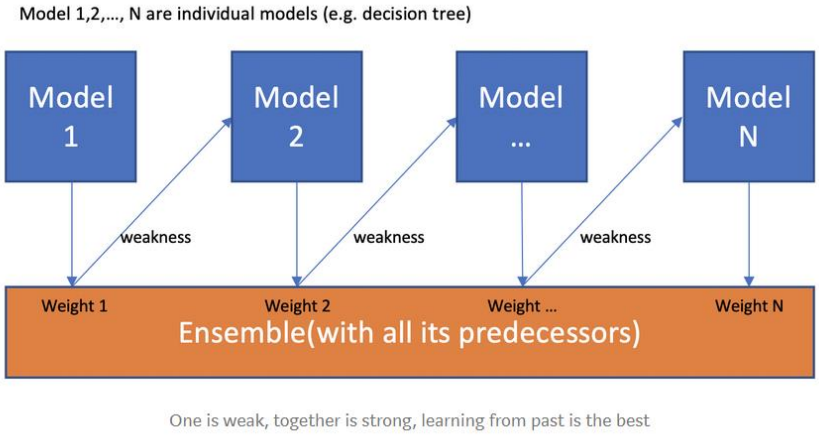
Αλγόριθμος Gradient Boosting

Ο αλγόριθμος Gradient Boosting (ενίσχυση της κλίσης-διαφοράς) είναι μια τεχνική μηχανικής μάθησης για προβλήματα παλινδρόμησης και ταξινόμησης, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης, συνήθως δέντρων αποφάσεων. Αποτελεί ένα συνδυασμό των αλγορίθμων ενίσχυσης **Boosting** και του αλγορίθμου **Gradient Descent** (διαβάθμιση κλίσης). (Singh, 2018)

5.2.1 Αλγόριθμοι ενίσχυσης (Boosting)

Ο όρος ενίσχυση απευθύνεται σε ένα σύνολο αλγορίθμων οι οποίοι μετατρέπουν έναν αδύναμο «μαθητή» σε ισχυρό. Αποτελεί μια μέθοδο για την βελτίωση των προβλέψεων οποιοδήποτε μοντέλου μάθησης. Η ιδέα της ενίσχυσης είναι να εκπαιδεύσει αδύναμους μαθητές διαδοχικά, προσπαθώντας ο καθένας να διορθώσει τον προκάτοχό του. Για να βρούμε έναν αδύναμο κανόνα (μαθητή) εφαρμόζουμε βασικούς κανόνες μηχανικής μάθησης με διαφορετική κατανομή ο καθένας

(Κανονική, Bernoulli, Poisson, κ.τ.λ.). Κάθε φορά που εκτελείτε ένας αλγόριθμος δημιουργεί και έναν αδύναμο κανόνα (μαθητή). Αυτή η διαδικασία γίνεται αρκετές φορές καθώς είναι επαναληπτική συνδυάζοντας στο τέλος όλους αυτούς τους κανόνες σε ένα ισχυρό κανόνα πρόβλεψης. (Ray, 2015)



Εικόνα 22 Επαναληπτική διαδικασία ενδυνάμωσης του μοντέλου

Για παράδειγμα αν θέλουμε να διαχωρίσουμε τα email σε spam και όχι spam πρέπει να θέσουμε κάποια κριτήρια για να μπορέσουμε να κάνουμε τον διαχωρισμό όπως :

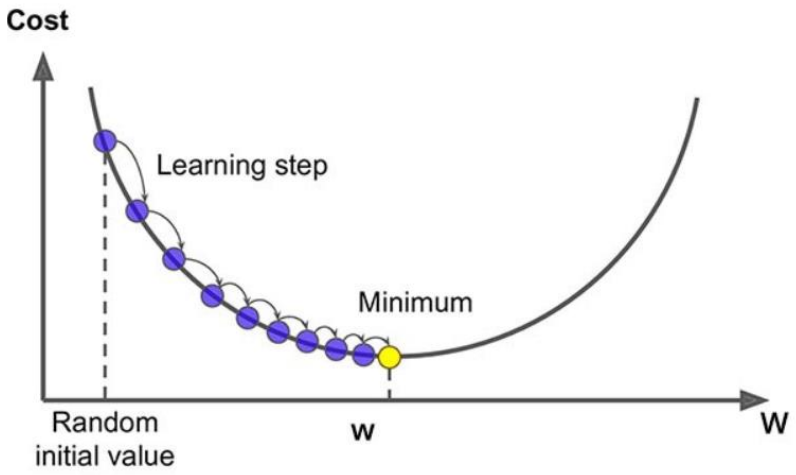
- Email με μία μόνο φωτογραφία αποτελεί spam.
- Email μόνο με υπερσυνδέσμους αποτελεί spam.
- Email από επίσημη και γνωστή πηγή δεν αποτελεί spam.

Μπορεί να ορίσαμε τα κριτήρια αλλά το καθένα από αυτά ξεχωριστά δεν είναι αρκετά ισχυρά ώστε να διαχωρίσει τα email με επιτυχία. Τα κριτήρια αυτά αποτελούν τους αδύναμους «μαθητές». Για να γίνει ισχυρό το μοντέλο πρέπει να συνδυαστούν όλα τα κριτήρια.

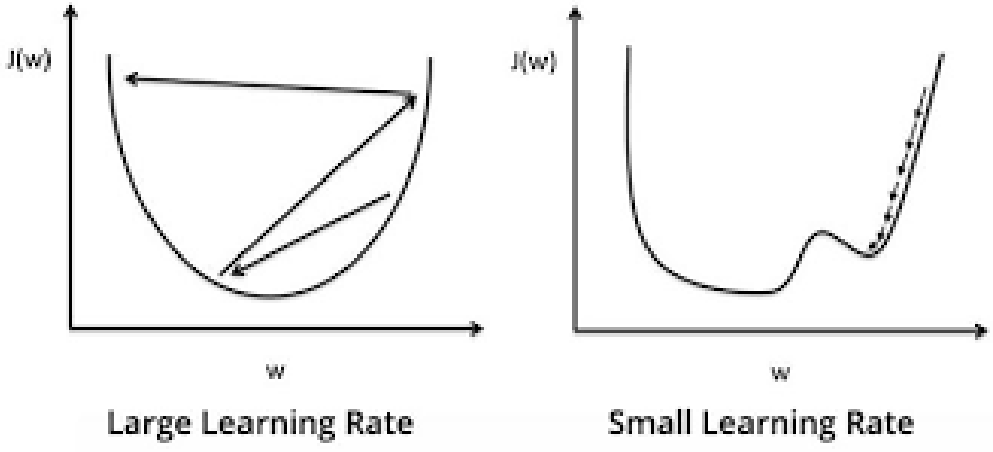
5.2.2 Αλγόριθμος Gradient Descent

Ο αλγόριθμος Gradient Descent αποτελεί μια τεχνική βελτιστοποίησης που χρησιμοποιείται σε πολλά προβλήματα μηχανικής μάθησης. Βασίζεται στην μείωση της συνάρτησης κόστους. Η συνάρτηση κόστους είναι η σχέση μεταξύ της «προβλεπόμενης» τιμής και της πραγματικής τιμής. Ο ρυθμός εκμάθησης έχει μεγάλη σημασία για τον αλγόριθμο. Με μικρό ρυθμό εκμάθησης γίνονται πολλές

επαναλήψεις μέχρι να βρεθεί (συγκλίνει) στο τοπικό ελάχιστο ενώ ο μεγάλος ρυθμός εκμάθησης θα οδηγούσε σε υπέρβαση της βέλτιστης τιμής. (Pandey, 2019) Η συνάρτηση απώλειας (lost function) υπολογίζει το σφάλμα για ένα μόνο παράδειγμα εκπαίδευσης, ενώ η συνάρτηση κόστους (cost function) είναι ο μέσος όρος των συναρτήσεων απώλειας για όλα τα παραδείγματα εκπαίδευσης. (Ruder, 2016)



Εικόνα 23 Ο ρυθμός μάθησης σε σχέση με το κόστος

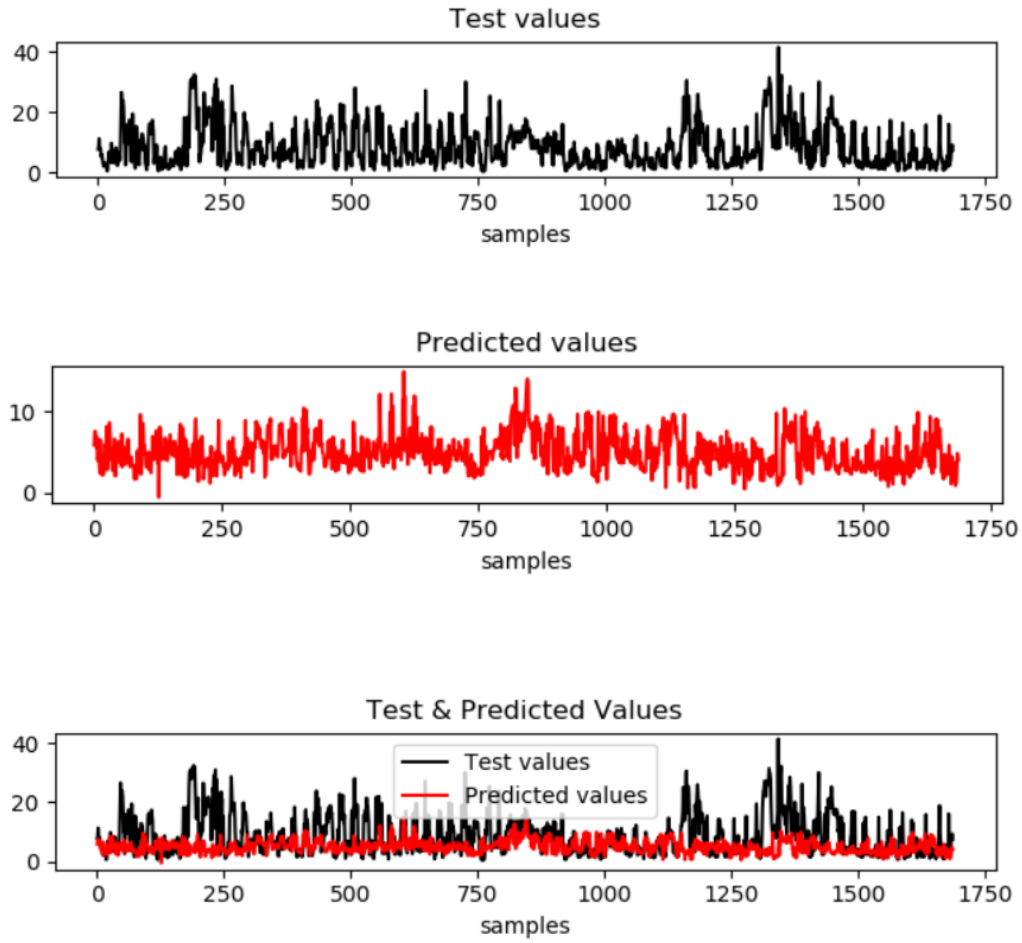


Εικόνα 24 Σύγκριση ρυθμού μάθησης

5.2.3 Αποτελέσματα

Σαν μέτρο εγκυρότητας του αλγορίθμου θα χρησιμοποιήσουμε το μέσο τετραγωνικό σφάλμα (mean squared error). Το MSE μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων δηλαδή τη μέση τετραγωνική διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών.

Το MSE (**Mean squared error**) της εφαρμογής μας είναι **56.36** τιμή αρκετά μεγάλη που σημαίνει ότι το μοντέλο πρόβλεψης μας έχει αρκετά μεγάλο αριθμό σφάλματος. Ο λόγος που προκύπτει αυτό είναι γιατί τα δεδομένα της εφαρμογής μας είναι εποχιακά δηλαδή έχουν μετρήσεις τόσο από χειμερινή περίοδο όσο και από καλοκαιρινή με αποτέλεσμα να υπάρχουν διακυμάνσεις στις τιμές της θερμοκρασίας, της υγρασίας, της ατμοσφαιρικής πίεσης κ.τ.λ. επηρεάζοντας έτσι αρνητικά την διαμόρφωση του μοντέλου μας. Στο παρακάτω γράφημα επιβεβαιώνεται το σφάλμα του μοντέλου από την διαφορά ανάμεσα στις προβλεπόμενες τιμές (κόκκινη γραμμή) και στις πραγματικές τιμές (μαύρες γραμμές).



Βιβλιογραφία

Anon., 2018. *import.io.* [Ηλεκτρονικό]
Available at: https://www.import.io/post/what-is-data-and-why-is-it-important/
[Πρόσβαση 15 Ιούλιος 2019].

Anon., 2019. *Computer Hope.* [Ηλεκτρονικό]
Available at: https://www.computerhope.com/jargon/d/data.htm
[Πρόσβαση 10 Αύγουστος 2019].

Anon., χ.χ. *Archer Software.* [Ηλεκτρονικό]
Available at: https://archer-soft.com/en/blog/data-mining-healthcare
[Πρόσβαση 25 Αύγουστος 2019].

Bachheriya, A. K., 2019. *Medium.* [Ηλεκτρονικό]
Available at: https://medium.com/datadriveninvestor/top-6-data-science-programming-languages-for-2019-39ba1b6819a8
[Πρόσβαση Ιούνιος 2019].

Bell, M., χ.χ. *Albion Research Ltd.* [Ηλεκτρονικό]
Available at: https://www.albionresearch.com/data_mining/market_basket.php

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* s.l.:Springer.

Chauhan, N. S., 2019. *Towards Data Science.* [Ηλεκτρονικό]
Available at: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4
[Πρόσβαση 2019].

Choudhary, P., 2017. *Datascience.* [Ηλεκτρονικό]
Available at: https://www.datascience.com/blog/python-anomaly-detection
[Πρόσβαση 25 Αύγουστος 2019].

Christopher, C., 2010. *ENCYCLOPÆDIA BRITANNICA.* [Ηλεκτρονικό]
Available at: https://www.britannica.com/technology/data-mining
[Πρόσβαση 24 Ιούνιος 2019].

COENEN, F., 2004. *Data Mining: Past, Present and Future,* Cambridge: s.n.

Dabbura, I., 2018. *Towards Data Science.* [Ηλεκτρονικό]
Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
[Πρόσβαση 24 Ιούνιος 2019].

Dhar, V., 2013. Data Science and Prediction. *Communications of the ACM*, Δεκέμβριος, pp. 64-73.

Dina Fawzy, S. M. N. B., 2016. The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics. *Asian Journal of Applied Sciences* , Ιούνιος, pp. 756-759.

Health, U., χ.χ. *Data Mining In Healthcare,* s.l.: s.n.

Joris Toonders, 2014. *WIRED.* [Ηλεκτρονικό]
Available at: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/
[Πρόσβαση 17 Ιούνιος 2019].

Leek, J., 2013. *Simply Statistics.* [Ηλεκτρονικό]
Available at: https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/
[Πρόσβαση 20 Δεκέμβριος 2019].

Li, H., 2017. *SAS.* [Ηλεκτρονικό]
Available at: https://www.sas.com/en_us/insights/analytics/machine-learning.html
[Πρόσβαση 25 Σεπτέμβριος 2019].

Michael Hahsler, B. G.̈. C. B. K. H., 2005. Introduction toarules– A computationalenvironment for mining association rules andfrequent item sets. *Journal of Statistical Software*.

Pandey, P., 2019. *Towards Data Science.* [Ηλεκτρονικό]
Available at: https://towardsdatascience.com/understanding-the-mathematics-behind-gradient-descent-dde5dc9be06e
[Πρόσβαση 25 Αύγουστος 2019].

Pang-Ning Tan, M. S. V. K., 2017,2010. *Εισαγωγή στην εξόρυξη δεδομένων.* Αθήνα: Εκδόσεις Τζιόλα.

Piatetsky-Shapiro, G., 1991. *Discovery, Analysis, and Presentation of Strong Rules,* s.l.: Knowledge Discovery in Databases.

P, R., 2014. *BigData-MadeSimple.* [Ηλεκτρονικό]
Available at: https://bigdata-madesimple.com/14-useful-applications-of-data-mining/
[Πρόσβαση 1 Σεπτεμβρίος 2019].

R Ragavi, B. S. V. S. A. S., 2018. Data Mining Issues and Challenges: A Review. *International Journal of Advanced Research in Computer and Communication Engineering*, Νοέμβριος, pp. 118-119.

Rao, V. S., 2018. *Technotification.* [Ηλεκτρονικό]
Available at: https://www.technotification.com/2018/06/data-mining-programming.html
[Πρόσβαση Ιούνιος 2019].

Ray, S., 2015. *Analytics Vidhya.* [Ηλεκτρονικό]
Available at: https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/
[Πρόσβαση 25 Αύγουστος 2019].

Ray, S., 2017. *Analytics Vidhya.* [Ηλεκτρονικό]
Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
[Πρόσβαση 28 Αύγουστος 2019].

Ruder, S., 2016. *ruder.io.* [Ηλεκτρονικό]
Available at: https://ruder.io/optimizing-gradient-descent/
[Πρόσβαση 25 Αύγουστος 2019].

Schott, M., 2019. *Medium.* [Ηλεκτρονικό]
Available at: https://medium.com/capital-one-tech/artificial-neural-networks-for-machine-learning-79c67d0681e9
[Πρόσβαση 21 Αύγουστος 2019].

Singh, H., 2018. *Towards Data Science.* [Ηλεκτρονικό]
Available at: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
[Πρόσβαση 19 Αύγουστος 2019].

Soumen Chakrabarti, M. E. U. F. J. G. H. S. M. G. P.-S. W. W., 2006. *https://www.kdd.org/.* [Ηλεκτρονικό]
Available at: https://www.kdd.org/curriculum/index.html
[Πρόσβαση 27 Ιανουάριος 2014].

Team, D., 2019. *DataFlair.* [Ηλεκτρονικό]
Available at: https://data-flair.training/blogs/data-mining-and-data-science/
[Πρόσβαση 22 Δεκέμβριος 2019].

Team, E. S., 2019. *Expert Systems.* [Ηλεκτρονικό]
Available at: https://expertsystem.com/machine-learning-definition/
[Πρόσβαση 19 Σεπτέμβριος 2019].

Victoria J. Hodge, J. A., χ.χ. *A Survey of Outlier Detection Methodologies,* York: s.n.

Yves Escoufier, B. F. D. L. H. N. O. B., 1995. Data Science and Its Applications. Στο: s.l.:Academic Press.

Παράρτημα

Παραθέτω ολόκληρο τον κώδικα της υλοποίησης.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_absolute_error

###################Load_the_data#########################
step=4 ##step=4 for 1 hours interval ##step=24 for 6 hours interval
##step=48 for 12 hours interval
datasheet
=pd.read_csv("all_data(fixed).csv",sep=",",header=0,usecols=["TWD(avg)","Hu
midityOutside","Air","Pressure","TWS(med)"]) #Anagnwsi twn dedomenwn apo to
csv
np_datasheet=datasheet[["TWD(avg)","HumidityOutside","Air","Pressure","TWS(
med)"]].to_numpy(dtype="float64") #allagi tis seiras twn column kai
metatropi se numpy array


np_datasheet=np_datasheet[0:np_datasheet.size:step]
X=np_datasheet[:,:-1]
X=X.astype(np.float64)

y=np_datasheet[:,4]
y=y.astype(np.float64)
X_train,X_test,y_train,y_test = train_test_split(X, y,
train_size=0.70,shuffle=False)

alpha = 0.95
model = GradientBoostingRegressor(loss='quantile', alpha=alpha,
                                  n_estimators=250, max_depth=3,
                                  learning_rate=.1, min_samples_leaf=9,
                                  min_samples_split=9)


model.fit(X_train, y_train)

y_upper = model.predict(X_test)

model.set_params(alpha=1.0-alpha)

model.fit(X_train,y_train)

y_lower = model.predict(X_test)

model.set_params(loss='ls')
model.fit(X_train,y_train)

y_pred=model.predict(X_test)
N=len(X_test)
xx = np.atleast_2d(np.linspace(0, N-1, N)).T
xx = xx.astype(np.float32)
fig=plt.figure()
```

```
plt.plot(xx, y_test, 'g-', label=r'$f(x) = x\,\sin(x)$')
#plt.plot(Xtest,y,'b.',markersize=5,label=u'Observations')
plt.plot(xx,y_pred,'r-',label=u'Prediction')
plt.plot(xx, y_upper, 'k-')
plt.plot(xx, y_lower, 'k-')
plt.fill(np.concatenate([xx, xx[::-1]]),
         np.concatenate([y_upper, y_lower[::-1]]),
         alpha=.5, fc='b', ec='None', label='90% prediction interval')
plt.xlabel('$x$')
plt.ylabel('$f(x)$')
#plt.ylim(-5, 100)
plt.legend(loc='upper left')
plt.show()

from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_absolute_error
model_score = model.score(X_train,y_train)
# Have a look at R sq to give an idea of the fit ,
# Explained variance score: 1 is perfect prediction
print('R2 sq: ',model_score)
y_predicted = model.predict(X_test)


# The mean squared error
print("Mean squared error: %.2f"% mean_squared_error(y_test, y_predicted))
print("Mean absolute error: %.2f"% mean_absolute_error(y_test,
y_predicted))
# Explained variance score: 1 is perfect prediction
print('Test Variance score: %.2f' % r2_score(y_test, y_predicted))



fig1, axs= plt.subplots(3,1,constrained_layout=True)
#ax.scatter(y_test, y_predicted, edgecolors=(0, 0, 0))
axs[0].plot(y_test,'-k')

axs[0].set_xlabel('samples')

axs[0].set_title('Test values')

axs[1].plot(y_predicted,'-r')

axs[1].set_xlabel('samples')

axs[1].set_title('Predicted values')

axs[2].plot(y_test,'-k',y_predicted,'-r')
axs[2].set_xlabel('samples')
axs[2].legend(['Test values','Predicted values'])
axs[2].set_title('Test & Predicted Values')

plt.show()
```