ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ **ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ**

ΠΕΡΙΦΕΡΕΙΑ ΗΠΕΙΡΟΥ

ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ

ΔΙΑΧΕΙΡΙΣΗΣ

Ε.Π. ΠΕΡΙΦΕΡΕΙΑΣ ΗΠΕΙΡΟΥ

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

“Ηπειρος” 2014-2020

ΚΩΔΙΚΟΣ ΠΡΑΞΗΣ (ΕΡΓΟΥ): ΗΠ1ΑΒ-00169

ΕΠΩΝΥΜΙΑ ΦΟΡΕΑ: Όμιλος Κολιού ΑΒΕΕ.

Εξαγωγές αγροτικών προϊόντων

Ακτινίδια - Πορτοκάλια - Μανταρίνια - Φράουλες

ΣΥΝΤΟΜΟΓΡΑΦΙΑ ΦΟΡΕΑ: Όμιλος Κολιού ΑΒΕΕ

Τίτλος Πρότασης

Σύγχρονο σύστημα αξιολόγησης της ποιότητας των ακτινιδίων, ιχνηλασιμότητα των παραγόμενων προϊόντων ακτινιδίου και ευφυής διαχείριση της αλυσίδας εφοδιασμού με βάση προηγμένες εφαρμογές Πληροφορικής ICT - FoodAware

ΙΑΝΟΥΑΡΙΟΣ, 2022

ΕΡΕΥΝΗΤΙΚΗ ΕΝΟΤΗΤΑ 6

ΠΑΡΑΔΟΤΕΟ: ΠΕ 6.1

Έκθεση σχετικά με τις τεχνικές και τους αλγόριθμους εξόρυξης δεδομένων οι οποίοι χρησιμοποιούνται για την εξαγωγή γνώσης και κανόνων από καλλιέργειες

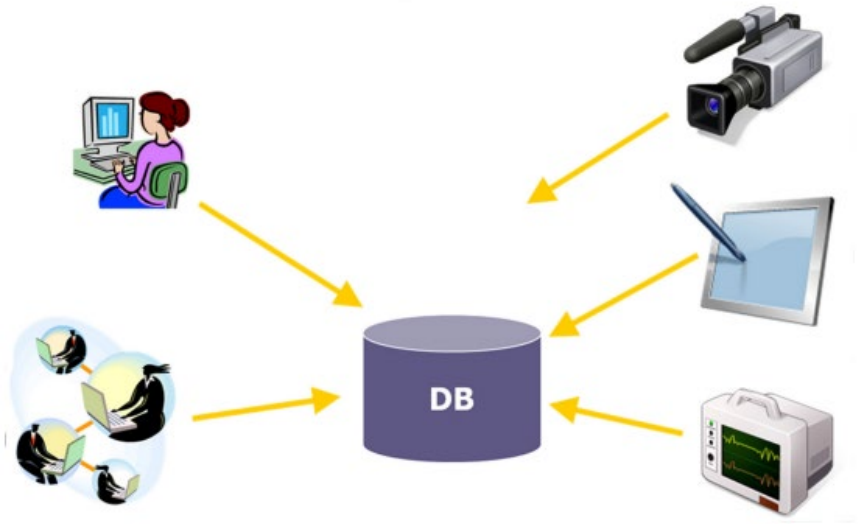
ΠΕΡΙΛΗΨΗ

Η εξόρυξη δεδομένων είναι ένα επιστημονικό πεδίο που ασχολείται με την ανάλυση, επεξεργασία και εξαγωγή γνώσης από δεδομένα, κυρίως μεγάλα δεδομένα, τα οποία βρίσκονται σε ακατέργαστη μορφή. Το παρόν παραδοτέο ασχολείται με την ανάλυση εννοιών και αλγορίθμους που εφαρμόζονται στην εξόρυξη δεδομένων και την μηχανική μάθηση. Αρχικά, αναλύεται η έννοια της εξόρυξης δεδομένων και η συνεισφορά άλλων κλάδων όπως η στατιστική και η μηχανική μάθηση στην εξόρυξη δεδομένων. Στη συνέχεια, περιγράφεται η έννοια των δεδομένων, τι είδους δεδομένα υπάρχουν, και η διαδικασία επεξεργασίας των δεδομένων, ώστε να είναι αποτελεσματική η εφαρμογή αλγορίθμων εξόρυξης δεδομένων και εξαγωγή γνώσης από αυτά. Τέλος, αναλύονται σύγχρονοι αλγόριθμοι εξόρυξης δεδομένων.

ΠΕΡΙΕΧΟΜΕΝΑ

1 Εισαγωγή

Στην σημερινή εποχή, ένας τεράστιος όγκος δεδομένων παράγεται από διάφορες πηγές, φυσικές και τεχνολογικές, καθώς και από ανθρώπινο παράγοντα. Το μεγαλύτερο μέρος αυτών των δεδομένων βρίσκονται σε ακατέργαστη μορφή (raw data). Η καταγραφή των δεδομένων αυτών μπορεί να προέρχεται από συλλογές μετρήσεων, κινητές συσκευές, τα κοινωνικά δίκτυα, δεδομένα αισθητήρων, δεδομένα παρακολούθησης, υποβολή ερωτημάτων, διαδικτυακές περιηγήσεις κλπ. Το χαμηλό κόστος αποθήκευσης επέτρεψε την διατήρηση των δεδομένων αυτών σε πρωτογενή μορφή, με σκοπό την διεξαγωγή γνώσης μέσω της ανάλυσης τους.



Εικόνα 1.1 Καταγραφή Δεδομένων (Κύρκος, 2015)

Η Εξόρυξη Δεδομένων (Data Mining - ΕΔ) χρησιμοποιείται για την αντιμετώπιση των προκλήσεων διαχείρισης αυτών των δεδομένων και την αποκάλυψη κρυφών μοτίβων (προτύπων – pattern recognition) και την κωδικοποίηση των πληροφοριών που συλλέγονται. Τα δεδομένα είναι εξαιρετικά χρήσιμα, με προϋπόθεση την αποτελεσματική αξιοποίηση της συλλογικής τους νοημοσύνης. Ουσιαστικά τα ακατέργαστα δεδομένα βρίσκονται σε μια μορφή που ο ανθρώπινος νους, με την περιορισμένη αναλυτική του δυνατότητα, δεν μπορεί να εξάγει χρήσιμες πληροφορίες. Η ΕΔ αποτελεί λύση σε αυτό το πρόβλημα. Αυτό επιτυγχάνεται με την βοήθεια

τεχνικών και αλγορίθμων που εφαρμόζει η ΕΔ, σε συνδυασμό με άλλες επιστήμες (στατιστική, μηχανική μάθηση, αναγνώριση προτύπων κλπ.), για την ανάλυση των δεδομένων. Η ΕΔ έχει εφαρμογή σε διάφορους τομείς στην σημερινή εποχή, όπως: στο εμπόριο, στην επιστημονική έρευνα, στην κλίμακα (σε μέγεθος δεδομένων και διάσταση χαρακτηριστικών).

Η διαχείριση γνώσης είναι η διαδικασία απόκτησης, δημιουργίας, σύνθεσης, διασποράς και χρήσης πληροφοριών, με εφαρμογή σε πολλούς τομείς. Για παράδειγμα, η Εξόρυξη Δεδομένων μπορεί να είναι πολύ χρήσιμη στις επιχειρήσεις, τόσο για την διαχείριση πληροφορίων όσο και για την απόκτηση γνώσεων μέσα από τις αλληλεπιδράσεις και τις δραστηριότητες με τους πελάτες. Πιθανά οφέλη περιλαμβάνουν αυξημένη κερδοφορία, μειωμένο κόστος, αυξημένη αποτελεσματικότητα, ταχεία ανταπόκριση και προσαρμογή και, τέλος, αυξημένη κερδοφορία.

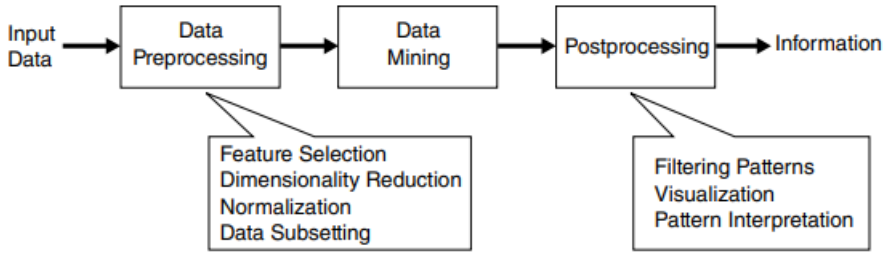
Η ΕΔ έχει εφαρμογή και σε γεωργικά προβλήματα. Μερικά παραδείγματα είναι η πρόβλεψη προβλημάτων κατά την διαδικασία ζύμωσης κρασιού, με σκοπό την διόρθωση και την εξασφάλιση μιας καλής ζύμωσης, η βελτίωση πρόβλεψης του καιρικών συνθηκών, η εκτίμηση υγρασίας του εδάφους, κ.ά. (Mucherino, Papajorgji and Pardalos, 2009). Τεχνικές και αλγόριθμοι Εξόρυξης Δεδομένων που χρησιμοποιούνται για την υλοποίηση των αναφερόμενων παραδειγμάτων θα αναλυθούν σε παρακάτω κεφάλαιο.

2 Βασικές Έννοιες

Η Εξόρυξη Δεδομένων ορίζεται ως "η χρήση αποτελεσματικών τεχνικών για την ανάλυση τεράστιου όγκου συνόλου δεδομένων και την εξαγωγή χρήσιμων και πιθανώς απροσδόκητων προτύπων από τα δεδομένα". (Tan, Steinbach, Karpatne and Kumar, 2006).

2.1 Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης

Η ΕΔ αποτελεί αναπόσπαστο μέρος της ανακάλυψης γνώσης από δεδομένα (KDD), η οποία είναι η διαδικασία μετατροπής των ακατέργαστων δεδομένων σε χρήσιμη πληροφορία. Αυτή η διαδικασία αποτελείται από μια σειρά μετασχηματισμών, από την προεπεξεργασία των δεδομένων έως τη μετα-επεξεργασία των αποτελεσμάτων της ΕΔ.



Εικόνα 2.1 Διαδικασία ανακάλυψης γνώσης από Δεδομένα (KDD - Tan, Steinbach and Kumar, 2006)

Αρχικά, θα πρέπει να σημειωθεί πως τα δεδομένα υπάρχουν διαθέσιμα σε ποικίλες μορφές και είναι δυνατόν να προέρχονται από κατηγορίες, διαφορετικές μεταξύ τους. Επίσης ανάλογα με την μορφή της ανάλυσης και τον τρόπο που αυτή σχεδιάζεται, είναι δυνατόν να πρέπει να χρησιμοποιηθούν σε περισσότερο από μια κατηγορίες δεδομένων. Για παράδειγμα, αν θεωρήσει κανείς το παράδειγμα της εξόρυξης δεδομένων από μια σχεσιακή βάση, τότε χρησιμοποιούνται μόνο τα δεδομένα που υπάρχουν καταγεγραμμένα στην βάση, τα οποία στην συντριπτική πλειοψηφία προέρχονται από κοινές κατηγορίες. Ενώ στην περίπτωση που πρόκειται να γίνει εξόρυξη από συστήματα που σχεδιάζονται για λειτουργία στο

διαδίκτυο, τα πράγματα είναι εντελώς διαφορετικά, καθώς εκεί υπάρχει τεράστιος όγκος δεδομένων με διαφορετική κατηγορία το καθένα.

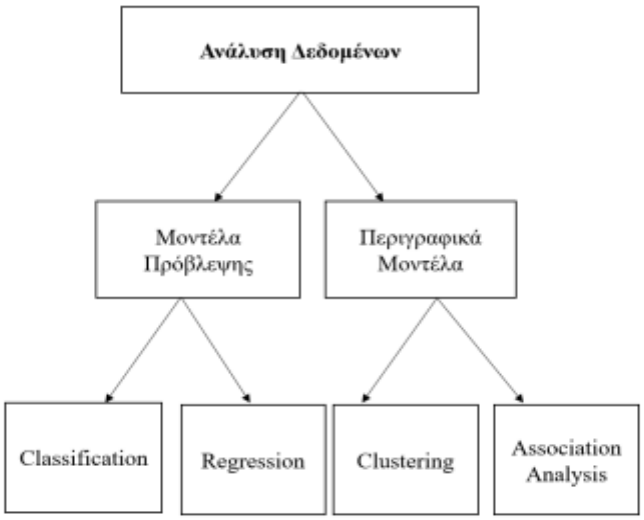
Επομένως, ο σκοπός της προ – επεξεργασίας (ή προετοιμασίας) είναι να μετασχηματιστούν όλα τα υφιστάμενα δεδομένα σε κατάλληλη μορφή για την υποβοήθηση της μετέπειτα λειτουργίας του αλγορίθμου εξόρυξης. Ο στόχος αυτός, σε ορισμένες περιπτώσεις είναι εύκολα υλοποιήσιμος ενώ σε κάποιες, πιο σύγχρονες, προσεγγίσεις είναι ιδιαίτερα σύνθετος, λόγω της ποικιλίας και της ανομοιομορφίας των δεδομένων. Ορισμένες βασικές επιδιώξεις είναι η αποβολή του θορύβου που περιέχουν τα δεδομένα, να προσαρμόσουν με συγκεκριμένο τρόπο όλα τα στοιχεία σε κοινό σύστημα και άλλες.

Στη συνέχεια, μετά την λειτουργία του αλγορίθμου εξόρυξης, υπάρχει περεταίρω ανάλυση. Η ανάλυση αυτή κυρίως αποσκοπεί στο να γίνουν τα δεδομένα όσο το δυνατόν πιο φιλικά και κατανοητά στον χρήστη, που θα τα χρησιμοποιήσει και εν τέλει θα λάβει την πληροφορία. Πιο συγκεκριμένα αναφέρεται ότι, η πληροφορία που δημιουργείται έχει σκοπό να τροφοδοτήσει κάποιο άλλο σύστημα, όπως για παράδειγμα ένα σύστημα συστάσεων ή ένα σύστημα υποστήριξης αποφάσεων, και γενικά να βοηθήσει τους αναλυτές να μπορούν εύκολα να ερμηνεύσουν τα μοτίβα και τις συσχετίσεις μεταξύ των δεδομένων. Για να είναι εφικτό κάτι τέτοιο θα πρέπει να υπάρξει μια παρουσίαση των συσχετίσεων με απλό και κατανοητό, για τον άνθρωπο τρόπο. Επομένως στο στάδιο αυτό, γίνεται χρήση στατιστικών τεχνικών, διαγραμμάτων απεικόνισης και άλλων εργαλείων, τα οποία γενικώς έχουν αποδεδειγμένα καλύτερη επίδραση στην κατανόηση των χρηστών.

2.2 Κατηγορίες συστημάτων Εξόρυξης Δεδομένων

Ο κλάδος της εξόρυξης δεδομένων έχει επεκταθεί πάρα πολύ τα τελευταία χρόνια, αφού όλο ένα και περισσότερες επιστήμες κάνουν χρήση των τεχνικών του, με σκοπό να βελτιώσουν τις αποφάσεις που λαμβάνονται και γενικώς να εκμεταλλευτούν τις δυνατότητες που παρέχει η πληθώρα των υπαρχόντων δεδομένων. Επομένως είναι φυσικό επόμενο ότι η μεγάλη απήχηση, έχει οδηγήσει στην δημιουργία διαφορετικών απαιτήσεων, οι οποίες με την σειρά τους οδηγούν

σε διαφορετικό σχεδιασμό και λειτουργία νέων αλγορίθμων. Έτσι έχουμε φτάσει στο σημείο όπου υπάρχει πληθώρα διαθέσιμων αλγορίθμων. Οι αλγόριθμοι αυτοί είναι αρκετά διαφορετικοί μεταξύ τους και ο καθένας από αυτούς ταιριάζει σε διαφορετικές κατηγορίες δεδομένων. Ωστόσο, είναι εφικτός ένας πρωταρχικός διαχωρισμός σε δύο βασικούς τύπους μοντέλων, των οποίων οι επιμέρους στόχοι ικανοποιούνται από διαφορετικές κατηγορίες αλγορίθμων.



Εικόνα 2.2 Τεχνικές Εξόρυξης και Ανάλυσης Δεδομένων

Όσον αφορά τον διαχωρισμό ως προς των τύπο των μοντέλων, εκεί διακρίνονται δύο βασικοί τύποι, ο πρώτος απαρτίζεται από μοντέλα τα οποία προβλέπουν (predictive models) και από μοντέλα τα οποία περιγράφουν (descriptive models - Tan, Steinbach, Karpatne and Kumar, 2006). Οι δύο αυτοί τύποι διαφέρουν κυρίως, ως προς τον στόχο λειτουργίας τους. Όπως προδίδεται και από τις αντίστοιχες ονομασίες, τα μεν πρώτα μοντέλα προσπαθούν να προβλέψουν τιμές και απαιτήσεις ορισμένων βασικών χαρακτηριστικών που εξετάζονται, ενώ τα δε δεύτερα λειτουργούν με στόχο τον συνδυασμό δεδομένων έτσι ώστε να δημιουργήσουν πρότυπα, μοτίβα και σχέσεις μεταξύ των δεδομένων και των ιδιοτήτων τους.

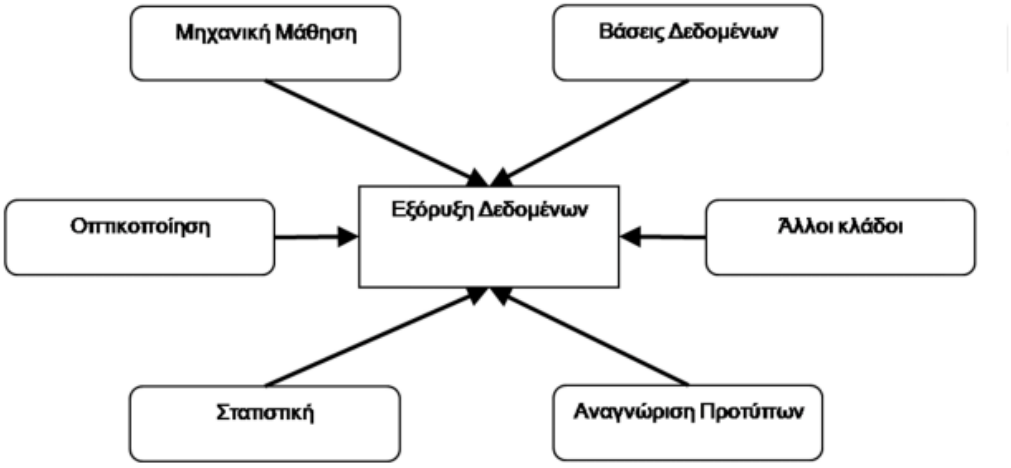
Ο παραπάνω διαχωρισμός εμπίπτει σε ένα πολύ αρχικό επίπεδο ανάλυσης. Ενώ σε πιο τεχνικό επίπεδο μπορούμε να σημειώσουμε πως υπάρχουν ορισμένες

κατηγορίες αλγορίθμων, οι οποίες επιτελούν συγκεκριμένους διαχωρισμούς και λειτουργίες μεταξύ των δεδομένων και των ιδιοτήτων αυτών. Οι κατηγορίες αυτές ικανοποιούν τόσο τους στόχους των περιγραφικών όσο και τους στόχους των μοντέλων πρόβλεψης. Επιπρόσθετα, στο παρόν σημείο αναφέρεται ότι, υπάρχουν πάρα πολλές κατηγοριοποιήσεις και αντίστοιχα πάρα πολλοί αλγόριθμοι που εμπίπτουν σε κάθε μία από αυτές.

Επιγραμματικά αναφέρεται ότι οι κατηγορίες αυτές είναι η ταξινόμηση (classification), η παλινδρόμηση (regression), η συσταδοποίηση (clustering) και η ανάλυση συσχέτισης (association analysis).

2.3 Εξόρυξη Δεδομένων και συνδυασμός με άλλους κλάδους

Η ΕΔ αντλεί μεθοδολογίες από πολλούς επιστημονικούς κλάδους, με μερικά παραδείγματα να είναι η Στατιστική, η Τεχνητή Νοημοσύνη, η Μηχανική Μάθηση, οι Βάσεις Δεδομένων, οι μηχανές αναζήτησης κλπ. Αυτοί οι κλάδοι από μόνοι τους καθιστούν αδύνατη την εξαγωγή γνώσης από δεδομένα μεγάλου όγκου, ή η διαδικασία είναι πολύ αργή και σε μερικές περιπτώσεις ανακριβής. Για παράδειγμα, η Στατιστική αν και προσφέρει λύσεις ανάλυσης δεδομένων, δεν λαμβάνει υπόψιν των μεγάλο όγκο των δεδομένων, παρομοίως και η Μηχανική Μάθηση και η Αναγνώριση Προτύπων (Κύρκος, 2015). Από την άλλη, ο κλάδος των Βάσεων Δεδομένων μπορεί να αποθηκεύσει μεγάλου όγκου δεδομένα, δεν προσφέρει τεχνικές ανάλυσης των δεδομένων. Η ΕΔ συνδυάζει αυτούς τους κλάδους, παρέχοντας μια λύση στα προβλήματα του κάθε κλάδου που από μόνοι τους δεν μπορούν να αντιμετωπίσουν.



Εικόνα 2.3. Η Εξόρυξη Δεδομένων ως αποτέλεσμα συνδυασμών άλλων κλάδων. (Κύρκος, 2015)

2.3.1 Στατιστική

Η επιστήμη της Στατιστικής βοηθά στην ανάλυση των δεδομένων, καθώς μπορεί να αναγνωρίσει μοντέλα ή πρότυπα που κρύβονται στις σχέσεις μεταξύ των μεταβλητών των δεδομένων. Η Στατιστική μπορεί να δώσει μια μαθηματική ερμηνεία των δεδομένων, το οποίο όμως δεν μπορεί να γενικευτεί για πολύ μεγάλα σύνολα δεδομένων (datasets). Η Στατιστική χρησιμοποιείται στην ΕΔ για την εις βάθος διερεύνηση των δεδομένων και την απόκτηση γνώσεων από αυτά. Οι στατιστικές τεχνικές εφαρμόζονται στους αλγορίθμους που χρησιμοποιούνται στην ΕΔ, που επιτρέπει την απόκτηση νέας γνώσης από τα δεδομένα.

2.3.2 Συστήματα Βάσεων Δεδομένων και Αποθήκες Δεδομένων

Η έρευνα στα συστήματα Βάσεων Δεδομένων (ΒΔ – Databases) επικεντρώνεται στην δημιουργία, συντήρηση και χρήση βάσεων για οργανισμούς και τελικούς χρήστες. Στα συστήματα ΒΔ έχουν καθιερωθεί ιδιαίτερα αναγνωρισμένες αρχές στην μοντελοποίηση δεδομένων, στις γλώσσες ερωτήσεων (query languages – π.χ. SQL), στην επεξεργασία ερωτημάτων και μεθόδους βελτιστοποίησης, στην αποθήκευση δεδομένων και, τέλος, στις μεθόδους ευρετηρίασης και πρόσβασης των δεδομένων, με αποτέλεσμα οι ΒΔ να

είναι γνωστές για την υψηλή επεκτασιμότητά τους στην επεξεργασία πολύ μεγάλου όγκου δεδομένων.

Πολλές εργασίες εξόρυξης δεδομένων πρέπει να χειρίζονται μεγάλα σύνολα δεδομένων ή ακόμη και ροές δεδομένων (streaming data). Επομένως, η ΕΔ μπορεί να κάνει καλή χρήση των συστημάτων ΒΔ, ώστε να έχει αποτελέσματα υψηλής απόδοσης και επεκτασιμότητας σε μεγάλα σύνολα δεδομένων. Επιπλέον, η ΕΔ μπορεί να χρησιμοποιηθεί για την επέκταση δυνατοτήτων των ΒΔ και να ικανοποιήσει τις ανάγκες απαιτητικών χρηστών, με εξαιρετικές απαιτήσεις ανάλυσης δεδομένων.

Σύγχρονα συστήματα ΒΔ ενσωματώνουν δυνατότητες ανάλυσης δεδομένων με την βοήθεια των Αποθηκών Δεδομένων και της ΕΔ. Μια Αποθήκη Δεδομένων αποτελείται από δεδομένα, που προέρχονται από πολλαπλές πηγές και διάφορα χρονικά πλαίσια. Ενοποιεί τα δεδομένα σε πολυδιάστατο χώρο, για να σχηματιστούν κύβοι δεδομένων (data cubes). Τα μοντέλα κύβων δεδομένων, όχι μόνο διευκολύνουν την διαδικτυακή αναλυτική επεξεργασία (OLAP), αλλά προωθεί και την ΕΔ σε πολυδιάστατα δεδομένα (Han, Kamber and Pei, 2012).

2.3.3 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι πεδίο της επιστήμης υπολογιστών που ανήκει στον τομέα της Τεχνητής Νοημοσύνης και ασχολείται με τη δημιουργία αλγορίθμων που έχουν ως στόχο να «μαθαίνουν» από δεδομένα, δηλαδή να αποκτούν την ικανότητα στην επιπλέον πρόσκτηση γνώσης μέσω της αλληλεπίδρασης με το περιβάλλον στο οποίο δραστηριοποιούνται και την ικανότητα να βελτιώνουν μέσω της επανάληψης τον τρόπο τον οποίο εκτελούν την ενέργεια αυτή. Σύμφωνα με τον Άρθουρ Σάμουελ (Samuel, 1959), η μηχανική μάθηση «*δίνει τη δυνατότητα στους υπολογιστές να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί*».

Ο πιο επίσημος ορισμός δόθηκε από τον Τομ Μ. Μίτσελ (Mitchell, 1997), σύμφωνα με τον οποίο: «*Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από μία εμπειρία Ε, σε σχέση με μία σειρά από έργα Τ και απόδοση μετρημένη με Ρ, αν η απόδοση του στα έργα Τ, μετρημένη με Ρ, βελτιώνεται με την εμπειρία Ε*».

Εφαρμόζεται σε μία πληθώρα εφαρμογών, στις οποίες ο ρητός σχεδιασμός και προγραμματισμός δεν είναι εφικτός, όπως για παράδειγμα στις μηχανές αναζήτησης, στην οπτική αναγνώριση χαρακτήρων, στα φίλτρα ανεπιθύμητων μηνυμάτων, κ.λπ. Η ΕΔ χρησιμοποιεί τεχνικές πρόβλεψης ή κατηγοριοποίησης της μηχανικής μάθησης. Με τη μηχανική μάθηση σε ένα σύστημα, είναι εφικτή η πρόβλεψη, που μέσω της ανατροφοδότησης (feedback), μπορεί να οδηγήσει στη μάθηση. Η μάθηση βασίζεται στα παραδείγματα, την αποθηκευμένη γνώση, και την ανατροφοδότηση. Όταν μελλοντικά συμβεί ανάλογη περίπτωση, η ανατροφοδότηση χρησιμοποιείται για να κάνει την ίδια πρόβλεψη ή για να κάνει μια εντελώς διαφορετική πρόβλεψη. Η στατιστική είναι πολύ σημαντική σε προγράμματα μηχανικής μάθησης γιατί τα αποτελέσματα των προβλέψεων πρέπει να είναι στατιστικά σημαντικά.

Είδη Μηχανικής Μάθησης

Τα είδη της μηχανικής μάθησης χωρίζονται βάσει του αποτελέσματος και του τρόπου με τον οποίο λειτουργεί το σύστημα εκμάθησης και ανατροφοδοτείται. Υπάρχουν τέσσερις κύριες κατηγορίες:

- Επιβλεπόμενη μάθηση
- Μη-επιβλεπόμενη μάθηση
- Ημι-επιβλεπόμενη μάθηση
- Ενισχυτική μάθηση

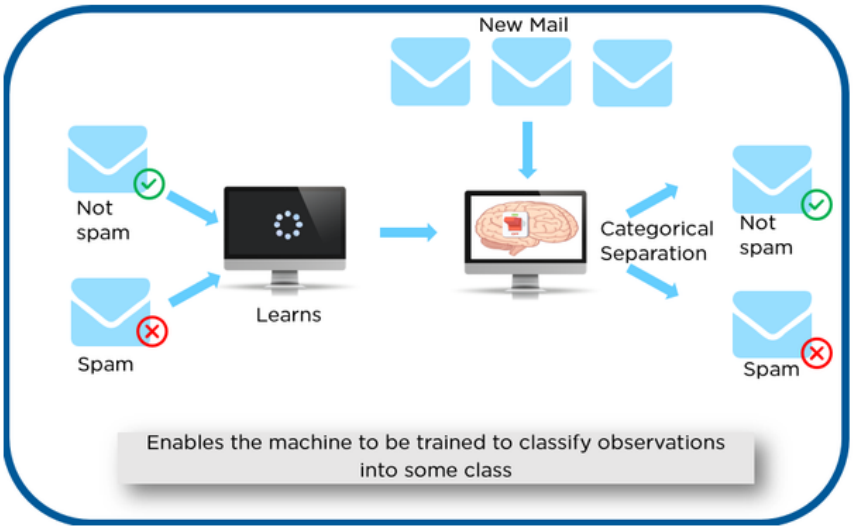
<u>Επιβλεπόμενη μάθηση (Supervised Learning)</u>

Ο όρος supervised learning αναφέρεται στη διαδικασία του να τροφοδοτήσεις έναν αλγόριθμο με εγγραφές στις οποίες μια μεταβλητή απόκρισης (output variable) που μας ενδιαφέρει είναι γνωστή και ο αλγόριθμος προσπαθεί να «μάθει» πώς να προβλέψει την τιμή με νέες εγγραφές όπου το αποτέλεσμα είναι άγνωστο. Δηλαδή, μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μία ή περισσότερες επεξηγηματικές μεταβλητές (input variable). Το σύνολο δεδομένων είναι μια συλλογή από ζευγάρια εισόδων και εξόδου της μορφής (x_i , y_i), $i = 1,2,\ldots,n$. Το x_i ονομάζεται διάνυσμα χαρακτηριστικών, όπου κάθε τιμή του διανύσματος

περιγράφει το i-οστό στοιχείο του συνόλου δεδομένων (εγγραφή). Το y_i είναι η "ετικέτα" και είναι συνήθως ένα στοιχείο από ένα σύνολο κατηγοριών ή ένας πραγματικός αριθμός, αλλά μπορεί να είναι και μια πιο σύνθετη δομή (διάνυσμα, γράφος, κ.ο.κ). Το x_i είναι το σύνολο των ανεξάρτητων μεταβλητών και το y_i είναι η εξαρτημένη μεταβλητή. Πρακτικά, το μοντέλο με ένα σύνολο δεδομένων εκπαίδευσης βρίσκει σχέσεις ανάμεσα στις τιμές του διανύσματος χαρακτηριστικών, αφού γνωρίζει ποια είναι η επιθυμητή έξοδος. Έπειτα από την διαδικασία εκπαίδευσης και αφού θα έχει ικανοποιητικά αποτελέσματα, το μοντέλο θα πρέπει να είναι σε θέση να προβλέπει την κατηγορία (ή ετικέτα) μιας άγνωστης εγγραφής, με μεγάλη πιθανότητα, σωστά. Αλγόριθμοι που βασίζονται σε αυτήν την μάθηση, είναι αλγόριθμοι κατηγοριοποίησης και παλινδρόμησης. (Burkov, 2019; Heidenreich, 2018)

Γνωστοί αλγόριθμοι/τεχνικές επιβλεπόμενης μάθησης είναι:

- Δέντρα Απόφασης
- Support Vector Machines (SVM)
- Naïve Bayes
- K-nearest neighbors (kNN) κ.ά.



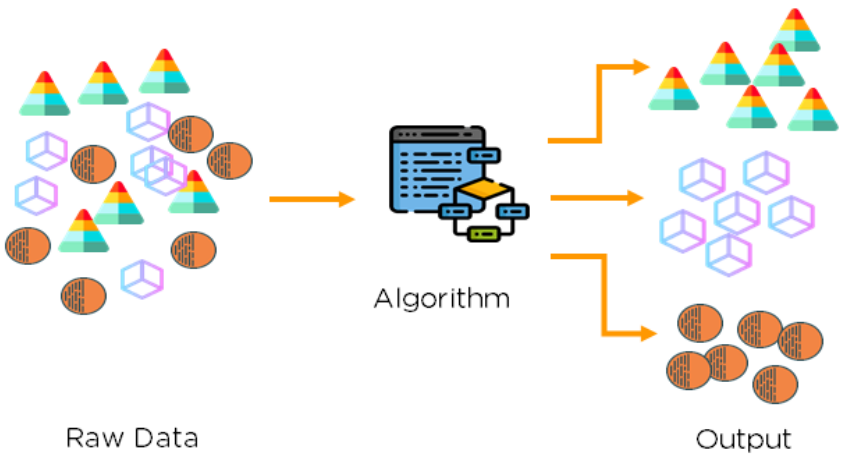
Εικόνα 2.4 Παράδειγμα Επιβλεπόμενης Μάθησης. Κατηγοριοποίηση email. (Heidenreich, 2018)

Μη-επιβλεπόμενη μάθηση (Unsupervised learning)

Η μη-επιβλεπόμενη μάθηση χρησιμοποιείται όταν δεν υπάρχει μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Πιο συγκεκριμένα, ο όρος unsupervised learning αναφέρεται στην ανάλυση που κάποιος επιχειρεί, για να μάθει κάτι άλλο για τα δεδομένα πέρα από την πρόβλεψη της τιμής μίας μεταβλητής που τον ενδιαφέρει, όπως για παράδειγμα, το αν ανήκει σε κάποιο cluster. Οι τεχνικές μάθησης χωρίς επίβλεψη χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά, διερευνώνται οι σχέσεις μεταξύ των δεδομένων ώστε να ανακαλυφθεί η γενική δομή τους.

Με άλλα λόγια, πρόκειται για αυτοματοποιημένη παραγωγή νέας γνώσης, όπου δεν είναι γνωστές οι επιθυμητές έξοδοι για το σύνολο εκπαίδευσης και το μοντέλο κατασκευάζεται γενικά για κάποιο σύνολο εισόδων με σκοπό να βρει τη δομή του συνόλου αυτού. Ενδεικτικοί αλγόριθμοι αυτής της κατηγορίας μάθησης είναι:

- Ο αλγόριθμος Apriori,
- k-means,
- Density-based spatial clustering of applications with noise (DBSCAN),
- Autoencoders
- Local Outlier Factor κ.ά.



Εικόνα 2.5. Παράδειγμα συσταδοποίησης. (Heidenreich, 2018)

Ημι-επιβλεπόμενη μάθηση (semi-supervised learning)

Στην ημι-επιβλεπόμενη μάθηση, τα δεδομένα περιέχουν και δεδομένα με ετικέτα και χωρίς ετικέτα. Ο αριθμός των δεδομένων με ετικέτα είναι συνήθως πολύ υψηλότερος. Σκοπός της ημι-επιβλεπόμενης μάθησης είναι ίδιος με την επιβλεπόμενη μάθηση, με την διαφορά ότι η χρήση πολλών δεδομένων χωρίς ετικέτα οδηγούν στην δημιουργία ενός καλύτερου μοντέλου, για κάποιες περιπτώσεις (Burkov, 2019; Heidenreich, 2018).

Ενισχυτική μάθηση (Reinforcement learning)

Ενισχυτική μάθηση είναι μια τεχνική, όπου το μοντέλο επηρεάζεται από το περιβάλλον. Το μοντέλο αντιλαμβάνεται την κατάσταση του περιβάλλοντος, ως ένα διάνυσμα χαρακτηριστικών. Το μοντέλο είναι σε θέση να δρα. Η κάθε δράση του μοντέλου ανταμείβεται ή να τιμωρείται, και μπορεί να αλλάξει την κατάσταση του περιβάλλοντος. Σκοπός του μοντέλου είναι να εκτελεί την καλύτερη δράση, σε κάθε κατάσταση του περιβάλλοντος. Η ενισχυτική μάθηση χρησιμοποιείται σε ηλεκτρονικά παιχνίδια, στην ρομποτική, σε προσομοιώσεις και σε συστήματα διαχείρισης πόρων (Burkov, 2019; Heidenreich, 2018). Ενδεικτικοί αλγόριθμοι ενισχυτικής μάθησης είναι οι:

- Monte Carlo
- Q-Learning
- SARSA
- DQN

3 Δεδομένα

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, προαπαιτούμενο της εξόρυξης δεδομένων είναι η «προετοιμασία» των δεδομένων. Τα δεδομένα του πραγματικού κόσμου συνήθως είναι θορυβώδη, τεράστια σε όγκο και μπορεί να προέρχονται από διαφορετικές πηγές. Η γνώση σχετικά με τα δεδομένα είναι χρήσιμη για την προεπεξεργασία των δεδομένων, που είναι το πρώτο σημαντικό βήμα της διαδικασίας ΕΔ. Ερωτήματα όπως από τι χαρακτηριστικά (attributes) αποτελούνται τα δεδομένα, τι τιμές παίρνουν τα χαρακτηριστικά, ποια χαρακτηριστικά είναι διακριτά και ποια παίρνουν πραγματικές τιμές, πως κατανέμονται οι τιμές, αν υπάρχουν τρόποι απεικόνισης των δεδομένων, αν υπάρχουν ακραίες τιμές (outliers), αν υπάρχει τρόπος μέτρησης ομοιότητας μεταξύ κάποιων εγγραφών κ.ά., βοηθούν στην επακόλουθη ανάλυση.

3.1 Εγγραφές και τύποι χαρακτηριστικών

Τα σύνολα δεδομένων αποτελούνται από εγγραφές. Μια εγγραφή είναι μια οντότητα, η οποία περιγράφεται από τα χαρακτηριστικά της. Οι εγγραφές μπορεί να αναφερθούν και ως δείγματα, παραδείγματα, στιγμιότυπα, σημεία δεδομένων (data points) ή αντικείμενα. Συνήθως, σε μια ΒΔ ή και κάποιων ειδών αρχείων (π.χ. csv αρχεία) οι εγγραφές αντιστοιχούνται με τις γραμμές, ενώ τα χαρακτηριστικά με τις στήλες.

Η λέξη «χαρακτηριστικό» είναι ταυτόσημη με τα ουσιαστικά διάσταση και μεταβλητή, αναλόγως τον κλάδο της επιστήμης και της βιβλιογραφίας. Ο όρος χαρακτηριστικό προτιμάται στην Μηχανική Μάθηση, ο όρος μεταβλητή στην στατιστική και ο όρος διάσταση σε Αποθήκες Δεδομένων. Στην ΕΔ ο όρος χαρακτηριστικό είναι ο επικρατέστερος. Ένα σύνολο χαρακτηριστικών χρησιμοποιείται για την περιγραφή μιας εγγραφής και ονομάζεται διάνυσμα χαρακτηριστικών. Για παράδειγμα, ένα διάνυσμα χαρακτηριστικών που περιγράφει έναν πελάτη μπορεί να περιλαμβάνει το αναγνωριστικό του πελάτη, όνομα και διεύθυνση.

Ο τύπος του χαρακτηριστικού εξαρτάται από τις τιμές που παίρνει το χαρακτηριστικό και μπορεί να είναι:

- Κατηγορικό (nominal - categorical)
- Δυαδικό (binary)
- Ιεραρχικό (ordinal)
- Αριθμητικό (numeric)

Κατηγορικό χαρακτηριστικό

Οι τιμές ενός κατηγορικού χαρακτηριστικού είναι σύμβολα ή ονόματα. Κάθε τιμή αντιπροσωπεύει κάποιο είδος κατηγορίας, κώδικα ή κατάστασης. Οι τιμές αυτές δεν ταξινομούνται. Στην επιστήμη των υπολογιστών, οι τιμές των κατηγορηματικών χαρακτηριστικών είναι γνωστές και ως απαριθμήσεις (enumerations - Han, Kamber and Pei, 2012).

Παρόλο που έχει αναφερθεί ότι τα κατηγορικά χαρακτηριστικά παίρνουν τιμές σε μορφή συμβόλων ή ονομάτων, μπορεί να αναπαρασταθούν και με αριθμούς. Όμως, σε αυτές τις περιπτώσεις οι αριθμοί δεν προορίζονται να χρησιμοποιηθούν ποσοτικά, δηλαδή η χρήση μαθηματικών πράξεων δεν έχει νόημα στις τιμές αυτές.

Επειδή στις κατηγορικές τιμές δεν έχει νόημα η ταξινόμηση, δεν έχει ουσία να βρεθεί και η μέση τιμή ή η διάμεσος για ένα τέτοιο χαρακτηριστικό, δεδομένου ενός συνόλου δεδομένων. Ωστόσο, ένα πράγμα που είναι ενδιαφέρον να βρεθεί, είναι η πιο συνηθισμένη τιμή του χαρακτηριστικού. Η εύρεση αυτής της τιμής, είναι σημαντική για τα μέτρα κεντρικής τάσης (Han, Kamber and Pei, 2012).

Δυαδικό χαρακτηριστικό

Ένα δυαδικό χαρακτηριστικό είναι ένα κατηγορικό χαρακτηριστικό με μόνο δύο καταστάσεις: 0 ή 1, όπου 0 σημαίνει συνήθως η απουσία του χαρακτηριστικού και 1 σημαίνει η ύπαρξη. Τα δυαδικά χαρακτηριστικά αναφέρονται και ως boolean εάν οι δύο καταστάσεις αντιστοιχούν σε αληθινές και ψευδείς.

Ένα δυαδικό χαρακτηριστικό είναι συμμετρικό εάν και οι δύο καταστάσεις είναι εξίσου πολύτιμες, δηλαδή δεν υπάρχει καμία προτίμηση στην οποία το

αποτέλεσμα πρέπει να κωδικοποιηθεί το αποτέλεσμα ως 0 ή 1. Ένα τέτοιο παράδειγμα είναι το φύλο. Αντίθετα, ένα δυαδικό χαρακτηριστικό θεωρείται ασύμμετρο όταν οι κατηγορίες δεν είναι εξίσου σημαντικά, όπως τα θετικά και τα αρνητικά αποτελέσματα μιας ιατρικής εξέτασης στον κορονοϊό. Συμβατικά, κωδικοποιούμε το πιο σημαντικό αποτέλεσμα, το οποίο είναι συνήθως το πιο σπάνιο, με 1 και το άλλο με 0 (Han, Kamber and Pei, 2012).

Ιεραρχικό χαρακτηριστικό

Ένα ιεραρχικό χαρακτηριστικό είναι ένα χαρακτηριστικό με τιμές που έχουν μια κατάταξη μεταξύ τους, αλλά το μέγεθος μεταξύ των διαδοχικών τιμών δεν είναι γνωστό (π.χ. μικρός – μεσαίος – μεγάλος). Τα ιεραρχικά χαρακτηριστικά είναι χρήσιμα για την καταχώριση αξιολογήσεων που δεν μπορούν να μετρηθούν αντικειμενικά. Έτσι, αυτού του είδους χαρακτηριστικών χρησιμοποιούνται κυρίως σε ερωτηματολόγια για αξιολογήσεις.

Οι κατηγορικές, δυαδικές και ιεραρχικές τιμές είναι ποιοτικές μετρήσεις, δηλαδή περιγράφουν ένα χαρακτηριστικό ενός αντικείμενου χωρίς να δίνει ένα πραγματικό μέγεθος ή μια ποσότητα. Παρακάτω θα αναλυθούν ποσοτικά χαρακτηριστικά, τα οποία παρέχουν ποσοτικές μετρήσεις ενός αντικειμένου (Han, Kamber and Pei, 2012).

Αριθμητικό χαρακτηριστικό

Ένα αριθμητικό χαρακτηριστικό είναι ποσοτικό, δηλαδή είναι μια μετρήσιμη ποσότητα που μπορεί να αναπαρασταθεί με ακέραιες ή πραγματικές τιμές και επιτρέπονται αριθμητικές πράξεις ανάμεσα σε δύο τιμές. Μπορούν να εκφραστούν με ίσα διαστήματα (interval-scaled) ή με αναλογική (ratio-scaled).

Τα χαρακτηριστικά ίσων διαστημάτων εκφράζουν μετρήσιμη ποσότητα, όπου μετριούνται με διαστήματα ίσου μεγέθους. Οι τιμές αυτού του είδους κλίμακα μέτρησης μπορεί να είναι αρνητικές, 0 ή θετικές. Έτσι, πέρα από την δυνατότητα κατάταξης, επιτρέπεται και η ποσοτικοποίηση της διαφοράς μεταξύ δύο χαρακτηριστικών. Παραδείγματα τιμών διαστημάτων είναι η θερμοκρασία και το IQ.

Η αναλογική κλίμακα μέτρησης είναι ένα αριθμητικό χαρακτηριστικό με ένα εγγενές μηδενικό σημείο, δηλαδή μπορεί να πάρει μόνο μη αρνητικές τιμές. Σε αυτήν την μέτρηση μπορούμε να χαρακτηρίσουμε μια τιμή ως πολλαπλάσιο (αναλογία) μιας άλλης τιμής. Επιπρόσθετα, οι τιμές είναι ταξινομήσιμες και μας επιτρέπεται ο υπολογισμός διαφοράς μεταξύ τιμών, ο μέσος όρος, η μέση τιμή και η επικρατούσα τιμή μεταξύ άλλων. Χαρακτηριστικά παραδείγματα είναι το βάρος, ηλικία, χρόνια εμπειρίας, ύψος, συντεταγμένες κλπ.

3.1.1 Ποιότητα δεδομένων

Οι εφαρμογές εξόρυξης δεδομένων συχνά εφαρμόζονται σε δεδομένα που συλλέχθηκαν για άλλο σκοπό ή για μελλοντικές, αλλά απροσδιόριστες εφαρμογές. Για αυτόν τον λόγο, η εξόρυξη δεδομένων δεν μπορεί συνήθως να επωφεληθεί από τα σημαντικά οφέλη της «αντιμετώπισης ζητημάτων ποιότητας στην πηγή». Αντιθέτως, μεγάλο μέρος των στατιστικών ασχολείται με το σχεδιασμό πειραμάτων ή ερευνών που επιτυγχάνουν ένα προκαθορισμένο επίπεδο ποιότητας δεδομένων. Επειδή η αποτροπή προβλημάτων ποιότητας δεδομένων συνήθως δεν αποτελεί επιλογή, η εξόρυξη δεδομένων επικεντρώνεται στον εντοπισμό και τη διόρθωση προβλημάτων της ποιότητας των δεδομένων και στη χρήση αλγορίθμων που μπορούν να ανεχθούν κακή ποιότητα δεδομένων. Το πρώτο βήμα, η ανίχνευση και η διόρθωση, ονομάζεται συχνά καθαρισμός δεδομένων.

3.1.1.1 Θέματα Μέτρησης και Συλλογής Δεδομένων

Η προσδοκία ότι τα δεδομένα θα είναι πάντα σε άψογη κατάσταση δεν είναι ρεαλιστική. Μπορεί να υπάρχουν προβλήματα λόγω ανθρώπινου λάθους, περιορισμών των συσκευών μέτρησης ή ελαττωμάτων στη διαδικασία συλλογής δεδομένων. Μπορεί να λείπουν τιμές ή ακόμα και ολόκληρα αντικείμενα δεδομένων. Σε άλλες περιπτώσεις, μπορεί να υπάρχουν πλαστά ή διπλότυπα αντικείμενα. δηλ. πολλαπλά αντικείμενα δεδομένων που αντιστοιχούν όλα σε ένα μόνο «πραγματικό» αντικείμενο. Για παράδειγμα, μπορεί να υπάρχουν δύο διαφορετικά αρχεία για ένα άτομο που έχει ζήσει πρόσφατα σε δύο διαφορετικές

διευθύνσεις. Ακόμα κι αν όλα τα δεδομένα είναι παρόντα και "φαίνονται καλά", μπορεί να υπάρχουν ασυνέπειες.

3.1.1.2 Σφάλματα Μέτρησης και Συλλογής Δεδομένων

Ο όρος σφάλμα μέτρησης αναφέρεται σε οποιοδήποτε πρόβλημα προκύπτει από τη διαδικασία μέτρησης. Ένα κοινό πρόβλημα είναι ότι η τιμή που καταγράφεται διαφέρει από την πραγματική τιμή σε κάποιο βαθμό. Για συνεχείς ιδιότητες, η αριθμητική διαφορά της μετρούμενης και της αληθινής τιμής ονομάζεται σφάλμα. Ο όρος σφάλμα συλλογής δεδομένων αναφέρεται σε σφάλματα, όπως η παράλειψη αντικειμένων δεδομένων ή τιμών χαρακτηριστικών ή η ακατάλληλη συμπερίληψη ενός αντικειμένου δεδομένων. Τόσο τα σφάλματα μέτρησης όσο και τα σφάλματα συλλογής δεδομένων μπορεί να είναι είτε συστηματικά είτε τυχαία.

3.1.1.3 Θόρυβος και Τεχνουργήματα

Ο θόρυβος είναι η τυχαία συνιστώσα ενός σφάλματος μέτρησης. Μπορεί να περιλαμβάνει την παραμόρφωση μιας τιμής ή την προσθήκη ψευδών αντικειμένων. Ο όρος θόρυβος χρησιμοποιείται συχνά σε σχέση με δεδομένα που έχουν χωρική ή χρονική συνιστώσα. Σε τέτοιες περιπτώσεις, τεχνικές από την επεξεργασία σήματος ή εικόνας μπορούν συχνά να χρησιμοποιηθούν για τη μείωση του θορύβου και, ως εκ τούτου, να βοηθήσουν στην ανακάλυψη μοτίβων (σημάτων) που μπορεί να «χαθούν στο θόρυβο». Ωστόσο, η εξάλειψη του θορύβου είναι συχνά δύσκολη και πολλή δουλειά στην εξόρυξη δεδομένων επικεντρώνεται στην επινόηση ισχυρών αλγορίθμων που παράγουν αποδεκτά αποτελέσματα, ακόμη και όταν υπάρχει θόρυβος. Τα σφάλματα δεδομένων μπορεί να είναι αποτέλεσμα ενός πιο ντετερμινιστικού φαινομένου, όπως μια ράβδωση στο ίδιο σημείο σε ένα σύνολο φωτογραφιών. Τέτοιες ντετερμινιστικές παραμορφώσεις των δεδομένων αναφέρονται συχνά ως τεχνουργήματα.

3.1.1.4 Εγκυρότητα, Μεροληψία και Ακρίβεια

Στη στατιστική και την πειραματική επιστήμη, η ποιότητα της διαδικασίας μέτρησης και τα δεδομένα που προκύπτουν μετρούνται με εγκυρότητα και μεροληψία. Εγκυρότητα είναι η εγγύτητα των επαναλαμβανόμενων μετρήσεων (της ίδιας ποσότητας) μεταξύ τους. Μεροληψία είναι μια συστηματική ποσότητα

που μετριέται. Η εγκυρότητα μετριέται συχνά με την τυπική απόκλιση ενός συνόλου τιμών, ενώ η μεροληψία μετριέται λαμβάνοντας τη διαφορά μεταξύ του μέσου όρου του συνόλου τιμών και της γνωστής τιμής της ποσότητας που μετριέται. Η μεροληψία μπορεί να προσδιοριστεί μόνο για αντικείμενα των οποίων η μετρούμενη ποσότητα είναι γνωστή με μέσα, εκτός της τρέχουσας κατάστασης. Είναι σύνηθες να χρησιμοποιείται ο γενικότερος όρος, ακρίβεια, για να αναφέρεται στο βαθμό του σφάλματος μέτρησης στα δεδομένα.

Η ακρίβεια αναφέρεται στην εγγύτητα των μετρήσεων στην πραγματική τιμή της μετρούμενης ποσότητας. Η ακρίβεια εξαρτάται από την εγκυρότητα και την μεροληψία, αλλά δεδομένου ότι είναι μια γενική έννοια, δεν υπάρχει συγκεκριμένος τύπος για την ακρίβεια όσον αφορά αυτές τις δύο ποσότητες. Μια σημαντική πτυχή της ακρίβειας είναι η χρήση σημαντικών ψηφίων. Ο στόχος είναι να χρησιμοποιηθούν μόνο τόσα ψηφία για να αναπαραστήσουν το αποτέλεσμα μιας μέτρησης ή υπολογισμού, όσα δικαιολογούνται από την εγκυρότητα των δεδομένων.

Ζητήματα όπως τα σημαντικά ψηφία, η εγκυρότητα, η μεροληψία και η ακρίβεια παραλείπονται μερικές φορές, αλλά είναι σημαντικά για την εξόρυξη δεδομένων καθώς και για τις στατιστικές και την επιστήμη. Πολλές φορές, τα σύνολα δεδομένων δεν συνοδεύονται από πληροφορίες για την ακρίβεια των δεδομένων, και επιπλέον, τα προγράμματα που χρησιμοποιούνται για την ανάλυση επιστρέφουν αποτελέσματα χωρίς καμία τέτοια πληροφορία. Ωστόσο, χωρίς κάποια κατανόηση της ακρίβειας των δεδομένων και των αποτελεσμάτων, ένας αναλυτής διατρέχει τον κίνδυνο να διαπράξει σοβαρά λάθη στην ανάλυση δεδομένων.

3.1.1.5 Ακραίες Τιμές

Οι ακραίες τιμές είναι είτε αντικείμενα δεδομένων που, κατά κάποια έννοια, έχουν χαρακτηριστικά που είναι διαφορετικά από τα περισσότερα από τα άλλα αντικείμενα δεδομένων στο σύνολο δεδομένων, είτε τιμές ενός χαρακτηριστικού που είναι ασυνήθιστες σε σχέση με τις τυπικές τιμές για αυτό χαρακτηριστικό. Υπάρχει σημαντικό περιθώριο στον ορισμό των ακραίων τιμών και πολλοί διαφορετικοί ορισμοί έχουν προταθεί από τις κοινότητες στατιστικών και εξόρυξης δεδομένων. Επιπλέον, είναι σημαντικό να γίνει διάκριση μεταξύ των

εννοιών του θορύβου και των ακραίων τιμών. Οι ακραίες τιμές μπορεί να είναι νόμιμα αντικείμενα ή τιμές δεδομένων. Έτσι, σε αντίθεση με τον θόρυβο, οι ακραίες τιμές μπορεί μερικές φορές να παρουσιάζουν ενδιαφέρον. Στην ανίχνευση απάτης και εισβολής στο δίκτυο, για παράδειγμα, ο στόχος είναι η εύρεση ασυνήθιστων αντικειμένων ή γεγονότων μεταξύ ενός μεγάλου αριθμού κανονικών.

3.1.1.6 Απούσες Τιμές

Δεν είναι ασυνήθιστο για ένα αντικείμενο να λείπουν μία ή περισσότερες τιμές χαρακτηριστικών. Σε ορισμένες περιπτώσεις, οι πληροφορίες δεν συλλέχθηκαν, για παράδειγμα στην περίπτωση που μερικοί άνθρωποι αρνούνται να δώσουν την ηλικία ή το βάρος τους. Σε άλλες περιπτώσεις, ορισμένα χαρακτηριστικά δεν ισχύουν για όλα τα αντικείμενα. Για παράδειγμα, συχνά, οι φόρμες έχουν μέρη υπό όρους που συμπληρώνονται μόνο όταν ένα άτομο απαντά σε μια προηγούμενη ερώτηση με συγκεκριμένο τρόπο, αλλά για λόγους απλότητας, όλα τα πεδία αποθηκεύονται. Ανεξαρτήτως, οι τιμές που λείπουν θα πρέπει να λαμβάνονται υπόψη κατά την ανάλυση δεδομένων.

Υπάρχουν διάφορες στρατηγικές (και παραλλαγές αυτών των στρατηγικών) για την αντιμετώπιση δεδομένων που λείπουν, καθεμία από τις οποίες μπορεί να είναι κατάλληλη σε ορισμένες περιπτώσεις. Ορισμένες από αυτές παρουσιάζονται παρακάτω.

3.1.1.6.1.1 Εξάλειψη των αντικειμένων δεδομένων ή των χαρακτηριστικών

Μια απλή και αποτελεσματική στρατηγική είναι η εξάλειψη αντικειμένων με τιμές που λείπουν. Ωστόσο, ακόμη και ένα μερικώς καθορισμένο αντικείμενο δεδομένων περιέχει ορισμένες πληροφορίες και εάν πολλά αντικείμενα έχουν τιμές που λείπουν, τότε μια αξιόπιστη ανάλυση μπορεί να είναι δύσκολη ή αδύνατη. Ωστόσο, εάν ένα σύνολο δεδομένων έχει μόνο μερικά αντικείμενα στα οποία λείπουν τιμές, τότε ίσως είναι σκόπιμο να παραλειφθούν. Μια σχετική στρατηγική είναι η εξάλειψη των χαρακτηριστικών που λείπουν τιμές. Αυτό πρέπει να γίνεται με προσοχή, ωστόσο, καθώς τα χαρακτηριστικά που έχουν εξαλειφθεί μπορεί να είναι αυτά που είναι κρίσιμα για την ανάλυση.

3.1.1.6.1.2 Εκτίμηση τιμών που λείπουν

Μερικές φορές τα δεδομένα που λείπουν μπορούν να εκτιμηθούν αξιόπιστα. Για παράδειγμα, έστω ένα σύνολο δεδομένων που έχει πολλά παρόμοια σημεία δεδομένων. Σε αυτήν την περίπτωση, οι τιμές χαρακτηριστικών των σημείων που βρίσκονται πιο κοντά στο σημείο με την τιμή που λείπει χρησιμοποιούνται συχνά για την εκτίμηση της τιμής που λείπει. Εάν το χαρακτηριστικό είναι συνεχές, τότε χρησιμοποιείται η μέση τιμή χαρακτηριστικού των πλησιέστερων γειτόνων. Εάν το χαρακτηριστικό είναι κατηγορηματικό, τότε μπορεί να ληφθεί η πιο συχνά εμφανιζόμενη τιμή χαρακτηριστικού.

3.1.1.6.1.3 Παράβλεψη των τιμών που λείπουν κατά την ανάλυση

Πολλές προσεγγίσεις εξόρυξης δεδομένων μπορούν να τροποποιηθούν ώστε να αγνοηθούν οι τιμές που λείπουν. Για παράδειγμα, έστω ότι τα αντικείμενα ομαδοποιούνται και πρέπει να υπολογιστεί η ομοιότητα μεταξύ ζευγών αντικειμένων δεδομένων. Εάν ένα ή και τα δύο αντικείμενα ενός ζεύγους έχουν τιμές που λείπουν για ορισμένα χαρακτηριστικά, τότε η ομοιότητα μπορεί να υπολογιστεί χρησιμοποιώντας μόνο τα χαρακτηριστικά που δεν έχουν τιμές που λείπουν. Είναι αλήθεια ότι η ομοιότητα θα είναι μόνο κατά προσέγγιση, αλλά εκτός εάν ο συνολικός αριθμός των χαρακτηριστικών είναι μικρός ή ο αριθμός των τιμών που λείπουν είναι μεγάλος, αυτός ο βαθμός ανακρίβειας μπορεί να μην έχει μεγάλη σημασία. Ομοίως, πολλά σχήματα ταξινόμησης μπορούν να τροποποιηθούν ώστε να λειτουργούν με τιμές που λείπουν.

3.1.1.7 Ασυνεπείς Τιμές

Τα δεδομένα μπορεί να περιέχουν ασυνεπείς τιμές. Έστω ένα πεδίο διεύθυνσης, όπου αναφέρονται τόσο ο ταχυδρομικός κώδικας όσο και η πόλη, αλλά η καθορισμένη περιοχή ταχυδρομικού κώδικα δεν περιέχεται σε αυτήν την πόλη. Μπορεί το άτομο που εισήγαγε αυτές τις πληροφορίες να μετέφερε δύο ψηφία ή ίσως ένα ψηφίο να αναγνώστηκε εσφαλμένα όταν οι πληροφορίες σαρώθηκαν από μια χειρόγραφη φόρμα. Ανεξάρτητα από την αιτία των ασυνεπών τιμών, είναι σημαντικό να εντοπιστούν και, εάν είναι δυνατόν, να διορθωθούν τέτοια προβλήματα.

Ορισμένοι τύποι ασυνεπειών είναι εύκολο να εντοπιστούν. Για παράδειγμα, το ύψος ενός ατόμου δεν πρέπει να είναι αρνητικό. Σε άλλες περιπτώσεις, μπορεί να χρειαστεί να επέμβει μια εξωτερική πηγή πληροφοριών. Για παράδειγμα, όταν μια ασφαλιστική εταιρεία επεξεργάζεται αξιώσεις για αποζημίωση, ελέγχει τα ονόματα και τις διευθύνσεις στα έντυπα αποζημίωσης σε μια βάση δεδομένων των πελατών της.

Μόλις εντοπιστεί μια ασυνέπεια, μερικές φορές είναι δυνατή η διόρθωση των δεδομένων. Ένας κωδικός προϊόντος μπορεί να έχει ψηφία "ελέγχου" ή μπορεί να είναι δυνατός ο διπλός έλεγχος ενός κωδικού προϊόντος σε σχέση με μια λίστα γνωστών κωδικών προϊόντων και, στη συνέχεια, να διορθωθεί ο κωδικός εάν είναι λανθασμένος, αλλά κοντά σε έναν γνωστό κωδικό. Η διόρθωση μιας ασυνέπειας απαιτεί πρόσθετες ή περιττές πληροφορίες.

3.1.1.8 Διπλότυπα Δεδομένα

Ένα σύνολο δεδομένων μπορεί να περιλαμβάνει αντικείμενα δεδομένων που είναι διπλότυπα ή σχεδόν διπλότυπα το ένα του άλλου. Πολλοί άνθρωποι λαμβάνουν διπλές αποστολές επειδή εμφανίζονται σε μια βάση δεδομένων πολλές φορές με ελαφρώς διαφορετικά ονόματα. Για τον εντοπισμό και την εξάλειψη τέτοιων διπλότυπων, δύο βασικά ζητήματα πρέπει να αντιμετωπιστούν. Πρώτον, εάν υπάρχουν δύο αντικείμενα που αντιπροσωπεύουν πραγματικά ένα μεμονωμένο αντικείμενο, τότε οι τιμές των αντίστοιχων χαρακτηριστικών μπορεί να διαφέρουν και αυτές οι ασυνεπείς τιμές πρέπει να επιλυθούν. Δεύτερον, χρειάζεται προσοχή ώστε να αποφευχθεί ο τυχαίος συνδυασμός αντικειμένων δεδομένων που είναι παρόμοια, αλλά όχι διπλά, όπως δύο διαφορετικά άτομα με πανομοιότυπα ονόματα. Ο όρος deduplication χρησιμοποιείται συχνά για να αναφερθεί στη διαδικασία αντιμετώπισης αυτών των θεμάτων.

Σε ορισμένες περιπτώσεις, δύο ή περισσότερα αντικείμενα είναι πανομοιότυπα σε σχέση με τα χαρακτηριστικά που μετρήθηκαν από τη βάση δεδομένων, αλλά εξακολουθούν να αντιπροσωπεύουν διαφορετικά αντικείμενα. Εδώ, τα διπλότυπα είναι νόμιμα, αλλά ενδέχεται να προκαλέσουν προβλήματα σε ορισμένους αλγόριθμους, εάν η πιθανότητα πανομοιότυπων αντικειμένων δεν λαμβάνεται ειδικά υπόψη στη σχεδίασή τους.

3.1.1.9 Ζητήματα που σχετίζονται με τις εφαρμογές

Τα ζητήματα ποιότητας δεδομένων μπορούν επίσης να εξεταστούν από την άποψη της εφαρμογής, όπως εκφράζεται από τη δήλωση "τα δεδομένα είναι υψηλής ποιότητας εάν είναι κατάλληλα για τη χρήση για την οποία προορίζονται". Αυτή η προσέγγιση στην ποιότητα των δεδομένων έχει αποδειχθεί αρκετά χρήσιμη, ιδιαίτερα στις επιχειρήσεις και τη βιομηχανία. Παρόμοια άποψη υπάρχει επίσης στις στατιστικές και στις πειραματικές επιστήμες, με έμφαση στον προσεκτικό σχεδιασμό των πειραμάτων για τη συλλογή των δεδομένων που σχετίζονται με μια συγκεκριμένη υπόθεση. Όπως και με τα ζητήματα ποιότητας σε επίπεδο μέτρησης και συλλογής δεδομένων, υπάρχουν πολλά ζητήματα που είναι ειδικά για συγκεκριμένες εφαρμογές και πεδία.

3.1.1.10 Επικαιρότητα

Ορισμένα δεδομένα αρχίζουν να «γερνούν» μόλις συλλεχθούν. Ειδικότερα, εάν τα δεδομένα παρέχουν ένα στιγμιότυπο κάποιου συνεχιζόμενου φαινομένου ή διαδικασίας, όπως η αγοραστική συμπεριφορά των πελατών ή τα μοτίβα περιήγησης στο Web, τότε αυτό το στιγμιότυπο αντιπροσωπεύει την πραγματικότητα μόνο για περιορισμένο χρονικό διάστημα. Εάν τα δεδομένα είναι παλιά, τότε είναι και τα μοντέλα και τα μοτίβα που βασίζονται σε αυτά.

3.1.1.11 Συνάφεια

Τα διαθέσιμα δεδομένα πρέπει να περιέχουν τις απαραίτητες πληροφορίες για την εφαρμογή. Για παράδειγμα, έστω το έργο της κατασκευής ενός μοντέλου που προβλέπει το ποσοστό ατυχημάτων για τους οδηγούς. Εάν παραληφθούν πληροφορίες σχετικά με την ηλικία και το φύλο του οδηγού, τότε είναι πιθανό το μοντέλο να έχει περιορισμένη ακρίβεια, εκτός εάν αυτές οι πληροφορίες είναι έμμεσα διαθέσιμες μέσω άλλων χαρακτηριστικών.

Είναι επίσης δύσκολο να βεβαιωθεί ότι τα αντικείμενα σε ένα σύνολο δεδομένων είναι σχετικά. Ένα κοινό πρόβλημα είναι η μεροληψία δειγματοληψίας, η οποία εμφανίζεται όταν ένα δείγμα δεν περιέχει διαφορετικούς τύπους αντικειμένων σε αναλογία με την πραγματική εμφάνισή τους στον πληθυσμό. Για παράδειγμα, τα δεδομένα της έρευνας περιγράφουν μόνο εκείνους που απαντούν στην έρευνα. Επειδή τα αποτελέσματα μιας

ανάλυσης δεδομένων μπορούν να αντικατοπτρίζουν μόνο τα δεδομένα που υπάρχουν, η μεροληψία δειγματοληψίας συνήθως οδηγεί σε εσφαλμένη ανάλυση.

3.1.1.12 Γνώση για τα Δεδομένα

Στην ιδανική περίπτωση, τα σύνολα δεδομένων συνοδεύονται από τεκμηρίωση που περιγράφει διαφορετικές πτυχές των δεδομένων. Η ποιότητα αυτής της τεκμηρίωσης μπορεί, είτε να βοηθήσει, είτε να εμποδίσει την επακόλουθη ανάλυση. Για παράδειγμα, εάν η τεκμηρίωση προσδιορίζει πολλά χαρακτηριστικά ως στενά συνδεδεμένα, αυτά τα χαρακτηριστικά είναι πιθανό να παρέχουν εξαιρετικά περιττές πληροφορίες και μπορεί να αποφασιστεί να διατηρηθεί μόνο ένα. Εάν η τεκμηρίωση είναι κακή, ωστόσο, τότε η ανάλυση των δεδομένων μπορεί να είναι εσφαλμένη. Άλλα σημαντικά χαρακτηριστικά είναι η ακρίβεια των δεδομένων, ο τύπος των χαρακτηριστικών (ονομαστική, τακτική, διάστημα, λόγος), η κλίμακα μέτρησης (π.χ. μέτρα ή πόδια για το μήκος) και η προέλευση των δεδομένων.

3.1.2 Μέτρα ομοιότητας και ανομοιότητας

Σε εφαρμογές εξόρυξης δεδομένων, όπως την συσταδοποίηση ή την ανάλυση ακραίων τιμών (outlier analysis), χρειαζόμαστε έναν τρόπο αξιολόγησης στο πόσο μοιάζουν δύο αντικείμενα μεταξύ τους. Για παράδειγμα, ένα κατάστημα ενδιαφέρεται για την ομαδοποίηση-συσταδοποίηση των πελατών με όμοια χαρακτηριστικά (π.χ. εισόδημα, ηλικία, κατοικία) για λόγους μάρκετιγνκ. Στην ανάλυση ακραίων τιμών, μας ενδιαφέρει η εύρεση αντικειμένων που δεν μοιάζουν με άλλα αντικείμενα, η οποία έχει εφαρμογή στην αναγνώριση απάτης.

Υπάρχουν διάφορες προσεγγίσεις για την μέτρηση ομοιότητας μεταξύ δύο αντικειμένων. Μια προσέγγιση είναι να θεωρούνται τα αντικείμενα ως διανύσματα n-διάστατου χώρου και να υπολογίζεται η ομοιότητα τους ως το συνημίτονο της γωνίας που σχηματίζουν:

$$\cos(x,y)= x^*yx^*y$$

όπου $x.y$ το εσωτερικό γινόμενο των διανυσμάτων και $\|x\|$ η Ευκλείδεια νόρμα του διανύσματος x. Αυτή η ομοιότητα είναι γνωστή ως η ομοιότητα συνημίτονου.

Η ομοιότητα μεταξύ των αντικειμένων μπορεί επίσης να δίδεται από την συσχέτισή τους. Μια τέτοια μέθοδος συσχέτισης είναι η συσχέτιση Pearson. Δεδομένης της συνδιακύμανσης Σ των σημείων-δεδομένων x και y και την τυπική απόκλισή τους σ, υπολογίζεται η συσχέτιση Pearson ως:

$$Pearson(x,y) = (x,y)x*y$$

Ενδεικτικά, άλλες προσεγγίσεις είναι η απόσταση Minkowski, η ευκλείδεια, η απόσταση Manhattan.

3.1.3 Μετρικές εξόρυξης δεδομένων

Οι μετρικές εξόρυξης δεδομένων μπορούν να οριστούν ως ένα σύνολο μετρήσεων που βοηθούν στον προσδιορισμό της αποτελεσματικότητας ενός τεχνικής/αλγορίθμου εξόρυξης δεδομένων. Η χρήση τους αποσκοπεί στην εύρεση της καλύτερης τεχνικής. Υπάρχουν διαφορετικοί τύποι εξόρυξης δεδομένων, και ο κάθε τύπος έχει δικές του μετρικές. Σε πολλές περιπτώσεις, μια μοναδική μετρική μπορεί να μην επαρκεί για την αξιολόγηση του αποτελέσματος. Σε αυτές τις περιπτώσεις, η ακρίβεια της αξιολόγησης επιτυγχάνεται με την χρήση πολλαπλών μετρικών. Η επιλογή των κατάλληλων μετρικών για κάθε εφαρμογή εξόρυξης δεδομένων είναι πολύ σημαντική.

Οι μετρικές εξόρυξης δεδομένων γενικά χωρίζονται στις εξής κατηγορίες:

- Ακρίβεια (accuracy)
- Αξιοπιστία (reliability)
- Χρησιμότητα (usefulness)

Ακρίβεια

Η ακρίβεια είναι ένα μέτρο για το πόσο καλά το μοντέλο συσχετίζει ένα αποτέλεσμα με τα χαρακτηριστικά των δεδομένων που έχουν παρασχεθεί.

Υπάρχουν διαφορετικές μετρικές ακρίβειας, οι οποίες όλες εξαρτώνται από τα δεδομένα που χρησιμοποιούνται. Σε πραγματικά σενάρια όμως ενδέχεται να λείπουν τιμές (missing values), ή να έχουν αλλάξει μετά από πολλαπλή επεξεργασία. Στην φάση της εξερεύνησης και ανάπτυξης του μοντέλου, συνηθίζεται να δεχόμαστε ένα ποσοστό σφάλματος, ειδικά αν πρόκειται για δεδομένα με ομοιόμορφα χαρακτηριστικά. Οι μετρικές ακρίβειας πρέπει να συνδυάζονται με μετρικές αξιοπιστίας.

Αξιοπιστία

Η αξιοπιστία είναι μια μετρική που αξιολογεί ένα μοντέλο εξόρυξης δεδομένων σε διαφορετικά σύνολα δεδομένων. Ένα μοντέλο εξόρυξης δεδομένων είναι αξιόπιστο όταν παράγει τον ίδιο τύπο προβλέψεων ή βρίσκει γενικά τα ίδια είδη μοτίβων, ανεξάρτητα από τα δεδομένα δοκιμών (test data) που παρέχονται.

Χρησιμότητα

Η χρησιμότητα περιλαμβάνει διάφορες μετρικές που μας ενημερώνουν εάν το μοντέλο παρέχει χρήσιμες πληροφορίες. Για παράδειγμα, μοντέλα που μπορεί να είναι ακριβή και αξιόπιστα, εάν δεν εξηγούν το φαινόμενο ή δεν μπορούν να γενικευτούν για άλλα σύνολα δεδομένων, δεν είναι χρήσιμα.

Η μέτρηση της αποτελεσματικότητας ή της χρησιμότητας δεν είναι πάντα εύκολη. Στην πραγματικότητα, διαφορετικές μετρικές θα μπορούσαν να χρησιμοποιηθούν για διαφορετικές τεχνικές και επίσης με βάση το τι και πόσο μας ενδιαφέρει το αποτέλεσμα του μοντέλου. Για παράδειγμα, από επιχειρηματική προοπτική, μας ενδιαφέρει η απόδοση της επένδυσης (ROI). Η μετρική ROI εξετάζει τη διαφορά μεταξύ του κόστους του μοντέλου και την εξοικονόμηση ή οφέλη της χρήσης του μοντέλου. Φυσικά, η ποσοτικοποίηση της επιστροφής είναι δύσκολη να πραγματοποιηθεί. Θα μπορούσε να μετρηθεί λαμβάνοντας υπόψιν αυξημένες πωλήσεις, ή μειωμένες διαφημιστικές δαπάνες ή και τα δύο.

3.2 Διαδικασία Ανακάλυψης Γνώσης

Όσο αφορά τον τρόπο λειτουργίας των συστημάτων εξόρυξης δεδομένων, αυτός είναι πολύ συγκεκριμένος και σαφώς ορισμένος. Πιο συγκεκριμένα, υπάρχουν καταγεγραμμένοι αλγόριθμοι που εφαρμόζονται ανάλογα με το επιθυμητό τελικό αποτέλεσμα, την μορφή και την φύση των δεδομένων και άλλους παράγοντες, ενώ πριν και μετά την λειτουργία κάθε αλγορίθμου, υπάρχουν οι διαδικασίες προ – επεξεργασίας και μετέπειτα επεξεργασίας των δεδομένων, μέχρις ότου αυτά να φτάσουν σε τέτοια μορφή που να είναι ικανά να προσφέρουν γνώσεις είτε στο σύστημα είτε σε κάποιον χρήστη.

Άλλοι βλέπουν την Εξόρυξη Δεδομένων ως ένα βήμα στη διαδικασία της ανακάλυψης γνώσης, η οποία αποτελείται από τα παρακάτω βήματα:

1. **Καθαρισμός Δεδομένων**, για την απομάκρυνση του θορύβου, εσφαλμένων δεδομένων και πληροφοριών που είναι περιττές.
2. **Ενσωμάτωση Δεδομένων**, από όπου μπορεί να γίνει ο συνδυασμός πολλαπλών πηγών δεδομένων.
3. **Συλλογή και Επιλογή Δεδομένων**, όπου τα δεδομένα που σχετίζονται με την διαδικασία της ανάλυσης ανακτώνται από τη βάση δεδομένων.
4. **Μετασχηματισμός Δεδομένων**, όπου τα δεδομένα μετασχηματίζονται ή ενοποιούνται σε μορφές κατάλληλες για εξόρυξη, εκτελώντας λειτουργίες σύνοψης ή συνάθροισης για παράδειγμα.
5. **Εξόρυξη Δεδομένων**, η διαδικασία και την οποία χρησιμοποιούνται έξυπνες μέθοδοι για να ανακαλυφθούν πρότυπα).
6. **Ερμηνεία / Αξιολόγηση Προτύπων**, να αναγνωριστούν τα πραγματικά ενδιαφέροντα πρότυπα που οδηγούν στην δημιουργία γνώσης, Η αξιολόγηση γίνεται με βάση κάποιους δείκτες που δείχνουν την αποτελεσματικότητα, και αυτοί είναι έμπιστοι και υποστηρικτικοί.
7. **Παρουσίαση της Γνώσης**, χρησιμοποιούνται τεχνικές παρουσίασης και τα δεδομένα οπτικοποιούνται, τα αποτελέσματα τις περισσότερες φορές παρουσιάζονται με τη μορφή διαγραμμάτων.

Κάποια από τα στάδια της διαδικασίας ανακάλυψης γνώσης μπορεί να επαναληφθούν αρκετές φορές. Τα στάδια της επιλογής, της προεπεξεργασίας, και του μετασχηματισμού συχνά αναφέρονται και ως στάδια προετοιμασίας των δεδομένων. Η προσπάθεια που απαιτείται σε κάθε στάδιο της διαδικασίας δεν είναι η ίδια.

Καθαρισμός των Δεδομένων

Αφετηριακό σημείο είναι τα πηγαία δεδομένα. Κατά κανόνα, τα δεδομένα αυτά είναι αποθηκευμένα σε διάφορες πηγές, όπως συστήματα παρακολούθησης συναλλαγών, ανεξάρτητες βάσεις δεδομένων, ανεξάρτητα αρχεία, εξωτερικές πηγές κλπ. Επίσης, τα δεδομένα αυτά πάσχουν από διάφορα προβλήματα όπως σφάλματα, αντιφάσεις, χαμένες τιμές κλπ. Τα αρχικά δεδομένα πρέπει να συλλεχθούν από τις διάφορες πηγές, να ομογενοποιηθούν και να καθαριστούν. Προβληματικά δεδομένα, τα οποία περιέχουν εσφαλμένες, ακραίες ή χαμένες τιμές μπορεί να αποπροσανατολίσουν τους αλγόριθμους εξόρυξης και να οδηγήσουν στην εξαγωγή άκυρων και εσφαλμένων προτύπων. Ορισμένοι αλγόριθμοι εξόρυξης δεδομένων έχουν τη δυνατότητα να αντιμετωπίζουν ενδογενώς τα προβλήματα των δεδομένων, ωστόσο αυτό δεν ισχύει γενικώς και επίσης, ο τρόπος αντιμετώπισης δεν είναι πάντα ο καλύτερος. Για τον λόγο αυτό, είναι προτιμότερο ο καθαρισμός των δεδομένων να γίνει από τον αναλυτή ως ανεξάρτητη εργασία και με τρόπο ελεγχόμενο από αυτόν. Συνήθως, τα δεδομένα, αφού απαλλαγούν από τα προβλήματα τους, αποθηκεύονται σε μια Αποθήκη Δεδομένων.

Συλλογή

Στο στάδιο της επιλογής των δεδομένων ορίζεται το σύνολο των δεδομένων στο οποίο θα αναζητηθούν πρότυπα. Η επιλογή των δεδομένων είναι μία διαδικασία πολύ σημαντική καθώς έμμεσα προσδιορίζει και τα αποτελέσματα που θα λάβουμε. Πολλές φορές, ο τρόπος αποθήκευσης τους δεν είναι συμβατός με τους αλγορίθμους ανακάλυψης γεγονότων που απαιτείται για την εξαγωγή των δεδομένων και την οργάνωση τους σε δομές που θα είναι προσβάσιμες από αυτούς.

Μετασχηματισμός

Η προεπεξεργασία των δεδομένων αποτελεί σημαντικό και απαραίτητο στάδιο της εξόρυξης δεδομένων, πριν την εφαρμογή των αλγορίθμων. Τα δεδομένα συνήθως συλλέγονται με διαδικασίες που μπορεί να μην διευκολύνουν την εξόρυξη γνώσης από αυτά, με αποτέλεσμα να περιέχουν θόρυβο και να επηρεάζεται αρνητικά η ποιότητά τους. Οι διαδικασίες περιλαμβάνουν συλλογή δεδομένων από ερωτηματολόγια, συσκευές μέτρησης όπως αισθητήρες ή ιατρικά όργανα, συνεντεύξεις, παρατηρήσεις ή το διαδίκτυο. Ο θόρυβος και οι ασυνεπείς τιμές μπορεί να προέρχονται από ανθρώπινο λάθος κατά την εισαγωγή τιμών ή από προβλήματα συσκευών.

Οι μέθοδοι της Εξόρυξης Δεδομένων είναι ισχυρώς καθοδηγούμενες από τα δεδομένα (data driven). Αυτό σημαίνει ότι το όποιο αποτέλεσμα θα αντληθεί απευθείας από τα δεδομένα. Για τον λόγο αυτό, η επιλογή των κατάλληλων δεδομένων είναι κομβικής σημασίας. Επιλογή δεδομένων σημαίνει καταρχάς επιλογή των κατάλληλων γνωρισμάτων ή χαρακτηριστικών. Η επιλογή των γνωρισμάτων είναι άμεσα συναρτώμενη με την εργασία που εκτελεί ο αναλυτής. Ορισμένα γνωρίσματα μπορεί να είναι χρήσιμα για μια εργασία, ενώ ορισμένα άλλα για κάποια άλλη. Ο αναλυτής αρχικά επιλέγει τα γνωρίσματα τα οποία θεωρεί ότι περιέχουν ουσιαστική πληροφορία που σχετίζεται με την ανάλυση του.

Η αρχική υποκειμενική επιλογή χαρακτηριστικών δεν είναι αρκετή. Στο αρχικό στάδιο, ο αναλυτής αποκλείει τα γνωρίσματα που εμφανώς δεν σχετίζονται με την ανάλυση του. Στη συνέχεια όμως, η επιλογή μπορεί να μην είναι προφανής, καθώς το ίδιο μέγεθος μπορεί να καταγράφεται με διαφορετικούς τρόπους.

Στο στάδιο της επιλογής χαρακτηριστικών γίνεται και ο μετασχηματισμός των δεδομένων. Για παράδειγμα, μπορεί να γίνει αναγωγή αριθμητικών τιμών σε άλλες αριθμητικές τιμές ή μετατροπή αριθμητικών τιμών σε ονομαστικές τιμές. Τέτοιες εργασίες γίνονται συνήθως για να προσαρμοστούν τα δεδομένα σε απαιτήσεις των μεθόδων ανάλυσης. Για παράδειγμα, ορισμένες μέθοδοι κατηγοριοποίησης επηρεάζονται ισχυρά από πεδία με μεγάλες τιμές και ασθενώς από πεδία με μικρές

τιμές. Σε τέτοιες περιπτώσεις, οι τάξεις μεγέθους των τιμών στα διάφορα πεδία πρέπει να είναι συγκρίσιμες. Το τελικό αποτέλεσμα αυτού του σταδίου είναι ένα σύνολο δεδομένων που θα χρησιμοποιηθεί για την εξαγωγή προτύπων.

Εκπαίδευση

Στο στάδιο αυτό επιλέγεται ποια τεχνική θα ακολουθηθεί, δηλαδή επιλέγεται ποιος αλγόριθμος θα εφαρμοστεί. Αυτό προσδιορίζεται από το είδος της γνώσης που θα αναζητηθεί. Οι δύο κατηγορίες γνώσης που μπορούν να προκύψουν είναι τα πρότυπα πληροφόρησης και τα πρότυπα πρόβλεψης. Η ουσιαστική ανακάλυψη γνώσης και εξόρυξη δεδομένων, συχνά υποκαθιστούν ο ένας τον άλλον. Σύμφωνα με την παραπάνω διαδικασία, το στάδιο της εξόρυξης δεδομένων είναι ένα μόνο βήμα στην όλη διαδικασία εξόρυξης γνώσης.

Ερμηνεία αποτελεσμάτων

Μόλις δημιουργηθεί το μοντέλο πρέπει να γίνει εκτίμηση των αποτελεσμάτων και η ερμηνεία της σημαντικότητας τους. Η ακρίβεια για παράδειγμα μπορεί να αποτελέσει ένα μέγεθος το οποίο προσδιορίζει το ποιο μοντέλο θα επιλεχθεί. Στα προβλήματα κατηγοριοποίησης η “confusion matrix” αποτελεί ένα καλό εργαλείο για την κατανόηση των αποτελεσμάτων. Για την καλύτερη ερμηνεία των αποτελεσμάτων απαιτείται η συνδρομή διάφορων γραφικών απεικονίσεων. Επομένως στο στάδιο αυτό, γίνεται χρήση στατιστικών τεχνικών, διαγραμμάτων απεικόνισης και άλλων εργαλείων, τα οποία γενικώς έχουν αποδεδειγμένα καλύτερη επίδραση στην κατανόηση των χρηστών.

Παρουσίαση γνώσης

Η οπτικοποίηση δεδομένων είναι η εμφάνιση πληροφοριών σε μορφή γραφημάτων ή πίνακα. Η επιτυχής οπτικοποίηση απαιτεί τη μετατροπή των δεδομένων (πληροφοριών) σε οπτική μορφή, έτσι ώστε τα χαρακτηριστικά των δεδομένων και οι σχέσεις μεταξύ στοιχείων ή ιδιοτήτων δεδομένων να μπορούν να αναλυθούν. Ο στόχος της οπτικοποίησης είναι η ερμηνεία των

οπτικοποιημένων πληροφοριών από ένα άτομο και ο σχηματισμός ενός νοητικού μοντέλου της πληροφορίας.

Στην καθημερινή ζωή, οπτικές τεχνικές όπως γραφήματα και πίνακες είναι συχνά η προτιμώμενη προσέγγιση που χρησιμοποιείται για να εξηγήσει τον καιρό, την οικονομία και τα αποτελέσματα των πολιτικών εκλογών. Ομοίως, ενώ οι αλγοριθμικές ή οι μαθηματικές προσεγγίσεις είναι αυτές στις οποίες δίνεται συχνά έμφαση στους περισσότερους τεχνικούς κλάδους, οι οπτικές τεχνικές που περιλαμβάνονται στην εξόρυξη δεδομένων μπορούν να διαδραματίσουν βασικό ρόλο στην ανάλυση δεδομένων. Μερικές φορές η χρήση τεχνικών οπτικοποίησης στην εξόρυξη δεδομένων αναφέρεται ως εξόρυξη οπτικών δεδομένων.

4 Τεχνικές και Αλγόριθμοι Εξόρυξης Δεδομένων

Αν και σήμερα τα πακέτα λογισμικού εξόρυξης δεδομένων θεωρούνται ότι είναι αυτοματοποιημένα, απαιτείται ακόμα να δίνονται μερικές κατευθυντήριες γραμμές από τους χρήστες. Η αναμενόμενη μέθοδος αλγορίθμου εξόρυξης δεδομένων είναι μια από αυτές τις απαιτήσεις. Επομένως, για τη χρησιμοποίηση των εργαλείων εξόρυξης δεδομένων, οι χρήστες πρέπει να έχουν μια βασική γνώση των μεθόδων αυτών. Οι διαφορετικοί τύποι των μεθόδων εξόρυξης δεδομένων μπορούν να ταξινομηθούν εναλλακτικά. Εντούτοις, γενικά εμπίπτουν σε έξι ευρείες κατηγορίες που είναι: Περιγραφή στοιχείων, ανάλυση εξάρτησης - συσχέτισης, ταξινόμηση και πρόβλεψη, συσταδοποίηση - ομαδοποίηση, ανάλυση ακραίων τιμών (outliers) και ανάλυση εξέλιξης.

Μια κατηγοριοποίηση των διαφορετικών τύπων εξόρυξης δεδομένων είναι:

- Πρόβλεψη
    - Κατηγοριοποίηση
    - Πρόβλεψη Τιμής
- Διαμερισμός
    - Συσταδοποίηση - Ομαδοποίηση
- Ανάλυση Σύνδεσης
    - Ανακάλυψη συσχετισμών
    - Ανακάλυψη ακολουθιακών προτύπων
    - Ανακάλυψη όμοιων χρονοσειρών
- Ανακάλυψη Αποκλίσεων
    - Στατιστική
    - Απεικόνιση

4.1 Κύριες κατηγορίες αλγορίθμων

Ταξινόμηση

Ξεκινώντας θα αναφερθούμε στην **ταξινόμηση (classification)** η οποία είναι μία από τις πλέον διαδεδομένες κατηγορίες αλγορίθμων στην ανάλυση δεδομένων. Η συγκεκριμένη κατηγορία ικανοποιεί ως επί το πλείστο τους στόχους των μοντέλων πρόβλεψης. Πιο συγκεκριμένα οι αλγόριθμοι αυτοί, προσπαθούν να δημιουργήσουν τις προϋποθέσεις έτσι ώστε κάθε είδους

αντικείμενο που πρόκειται να αξιολογηθεί, να είναι εφικτό να ταξινομηθεί σε ήδη υπάρχουσες κατηγορίες δεδομένων. Η ταξινόμηση αυτή γίνεται με βάση τα χαρακτηριστικά του συγκεκριμένου αντικειμένου και με γνώμονα το πόσο αυτά ταιριάζουν με άλλες κατηγορίες. Ένα σημαντικό θέμα, είναι το πόση ομοιότητα θεωρείται ικανοποιητική ούτος ώστε να γίνει η κατάταξη ενός αντικειμένου σε μια ευρύτερη ομάδα. Επίσης σημαντικό είναι το ζήτημα που ανακύπτει σε περιπτώσεις όπου τα διαθέσιμα δεδομένα είναι τέτοια ώστε να μην μπορούν να δημιουργηθούν ευδιάκριτες ομάδες.

Τα παραπάνω ζητήματα επιλύονται, στον βαθμό του εφικτού, από τον εκάστοτε αλγόριθμο με διαφορετικές προσεγγίσεις, οι οποίες θα παρουσιαστούν κάτωθι κατά την ανάλυση των αλγορίθμων. Όπως φανερώνεται από τα προηγούμενα, για την εκτέλεση των αλγορίθμων της συγκεκριμένης κατηγορίας είναι απολύτως απαραίτητες δύο ξεχωριστές βάσεις δεδομένων. Η πρώτη χρησιμοποιείται ως βοηθητική, με στόχο την τροφοδοσία του μοντέλου με δεδομένα και κατ' επέκταση την δημιουργία των ομάδων (training data set). Ενώ η δεύτερη, είναι κύρια βάση που περιέχει τα δεδομένα, των οποίων επιζητείται η ταξινόμηση.

Επιπρόσθετα, αξίζει να σημειωθεί ότι τέτοιου είδους αλγόριθμοι μπορούν να χρησιμοποιηθούν είτε σε βάσεις που περιέχουν διακριτά ή πεπερασμένα δεδομένα (classification) είτε σε βάσεις με συνεχή τύπο δεδομένων (regression) με μικρές παραλλαγές. Ωστόσο αμφότεροι οι αλγόριθμοι είναι ίδιοι, με ελάχιστες παραλλαγές. Ολοκληρώνοντας αναφέρεται ότι οι κυρίαρχες κατηγορίες αλγορίθμων, που χρησιμοποιούνται στο πλαίσιο της ταξινόμησης, είναι τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα. Οι τεχνικές αυτές έχουν συγκεκριμένα πλεονεκτήματα και αδυναμίες, ενώ ταιριάζουν σε διαφορετικούς τύπους δεδομένων και σχολιάζονται παρακάτω.

<u>Παλινδρόμηση</u>

Η επόμενη βασική κατηγορία, που σημειώνεται, είναι η **παλινδρόμηση (Regression)**. Είναι κλάδος της στατιστικής που εξετάζει την σχέση ανάμεσα σε μια τιμή εξόδου (εξαρτημένη τιμή) και μιας τιμής εισόδου (απλή παλινδρόμηση)

ή περισσότερων τιμών (πολλαπλή παλινδρόμηση) . (Gorunescu, 2011**)** Μαθηματικά, ένα μοντέλο παλινδρόμησης είναι μια συνάρτηση με την οποία μπορούμε να προβλέψουμε την τιμή της ανεξάρτητης μεταβλητής μέσω των ανεξάρτητων.

Η πρώτη μορφή παλινδρόμησης που χρησιμοποιήθηκε για την εύρεση εξίσωσης μιας ευθείας είναι η μέθοδος ελαχίστων τετραγώνων, που χρησιμοποίησαν οι Legendre (Legendre, 1805) και Gauss (Gauss, 1809) για τον προσδιορισμό της τροχιάς αντικειμένων γύρω από τον ήλιο. Όμως, ο όρος της παλινδρόμησης επινοήθηκε από τον Galton το 1886 (Galton, 1886), όπου περιέγραψε το φαινόμενο ότι το ύψος απογόνων, οι οποίοι είχαν ψηλούς προγόνους τείνει να πέφτει προς τον μέσο όρο. Για τον Galton, ο όρος παλινδρόμηση ήταν μόνο βιολογικής σημασίας. Αργότερα, οι Yule (Yule, 1897) και Pearson (Pearson, 1903) γενίκευσαν τον όρο Παλινδρόμηση και τον συνδύασαν με την στατιστική. Στις δημοσιεύσεις τους, η κοινή κατανομή (joint distribution) των εξαρτημένων και ανεξάρτητων μεταβλητών θεωρείται να είναι κανονική (γκαουσιανή). Ο Fisher (Fisher, 1922-1925), αντιθέτως, θεώρησε ότι η υποθετική κατανομή είναι κανονική, αλλά η κοινή κατανομή δεν χρειάζεται να είναι απαραίτητα κανονική.

Συσταδοποίηση

Η επόμενη βασική κατηγορία, που σημειώνεται, είναι η **συσταδοποίηση (clustering)**. Η συγκεκριμένη κατηγορία περιέχει αλγορίθμους, το οποίων η λειτουργία προσπαθεί να ικανοποιήσει τους στόχους που επιτάσσει το περιγραφικό μοντέλο ανάλυσης δεδομένων. Πιο συγκεκριμένα, ο βασικός στόχος των τεχνικών αυτής της κατηγορίας είναι ο καταμερισμός ενός συνόλου δεδομένων, με ανομοιογένειες μεταξύ τους, σε συγκεκριμένες συστάδες με συμπαγές και συναφές περιεχόμενο δεδομένων. Με πιο απλά λόγια, από ένα γενικό σύνολο επιδιώκεται η δημιουργία ειδικών ομάδων δεδομένων. Η συσταδοποίηση είναι εντελώς διαφορετική διεργασία από την ταξινόμηση, καθώς δεν βασίζει την λειτουργία της στην κατάταξη νέων δεδομένων σε ήδη κατασκευασμένες ομάδες δεδομένων, αλλά προσπαθεί να δημιουργήσει τις ομάδες αυτές των δεδομένων ή συστάδων, όπως αναγράφονται στο πλαίσιο της

ανάλυσης δεδομένων. Κάτω από αυτό το πρίσμα θα μπορούσαμε να θεωρήσουμε ότι η συσταδοποίηση είναι μια απαραίτητη διαδικασία για την ορθή διεξαγωγή της ταξινόμησης, στην συνέχεια. Εν τέλει, πρόκειται για δύο διαφορετικές διεργασίες μεταξύ τους, οι οποίες ωστόσο είναι απαραίτητο να συνυπάρχουν.

Ανάλυση Συσχέτισης

Η **ανάλυση συσχέτισης (association analysis)** είναι η επόμενη κατηγορία αλγορίθμων που θα παρουσιαστεί. Η παρούσα κατηγορία, επίσης, χρησιμοποιείται στο πλαίσιο της επίτευξης των στόχων των μοντέλων περιγραφής. Επιπρόσθετα, από πληθώρα εργασιών, θεωρείται αρκετά σύγχρονη και χρήσιμη κατηγορία και λόγω αυτού έχει συγκεντρώσει πολυάριθμες εφαρμογές.

Οι τεχνικές αυτής της κατηγορίας, επικεντρώνονται στην εύρεση μοτίβων και συσχετίσεων που θεωρείτε πως υπάρχουν μεταξύ διαφορετικών κατηγοριών δεδομένων, ωστόσο δεν είναι εύκολο να εντοπιστούν εξ' αρχής. Επίσης είναι αρκετά σύνηθες το φαινόμενο της ύπαρξης ισχυρών δεσμών μεταξύ δεδομένων που είναι εντελώς ανόμοια μεταξύ τους και επί της ουσίας αυτοί οι δεσμοί επιδιώκεται να εντοπιστούν. Ενώ, στην αντίπερα όχθη, πολλές φορές συναντώνται και οι σχέσεις μεταξύ συναφών, αλλά όχι όμοιων, ομάδων δεδομένων. Αντίθετα, με τις μη προφανείς περιπτώσεις, τα μοτίβα μεταξύ των συναφών κατηγοριών δεν ενδιαφέρουν τόσο καθώς δεν παρέχουν κάποια ιδιαίτερη γνώση για τα μοτίβα συσχέτισης, παρά μόνο την απολύτως βασική.

Μια από τις πιο γνωστές εφαρμογές, της συγκεκριμένης κατηγορίας, η οποία εν πολλοίς αποτέλεσε και το εναρκτήριο λάκτισμα για την εδραίωσή της, είναι αυτή της ανάλυσης του καλαθιού του super market. Σύμφωνα με την συγκεκριμένη ανάλυση επιδιώκεται η εκλογίκευση των επιλογών που γίνονται κατά την αγορά προϊόντων, από διάφορους αγοραστές. Ενώ σε δεύτερο στάδιο γίνεται προσπάθεια για την δημιουργία μοτίβων και για γενίκευση αυτών. Η τακτική αυτή θεωρήθηκε ότι αποδίδει και για τον λόγο αυτό, αρκετά super market, κυρίως στην Αμερική, την έχουν υιοθετήσει και προσπαθούν να

δημιουργήσουν την διαρρύθμιση τόσο του χώρου όσο και των ραφιών τους, ακολουθώντας μοντέλα ανάλυσης συσχέτισης.

4.2 Αλγόριθμοι Εξόρυξης Δεδομένων

4.2.1 Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση είναι μια από τις πιο διαδεδομένες τεχνικές πρόβλεψης στην στατιστική και την μηχανική μάθηση. Χρησιμοποιείται για την εύρεση σχέσης μεταξύ μιας ή περισσότερων ανεξάρτητων τιμών και μιας εξαρτημένης. Υπάρχουν δύο ειδών γραμμικής παλινδρόμησης:

- Η απλή γραμμική παλινδρόμηση
- Η πολλαπλή γραμμική παλινδρόμηση

Έστω, ένα διάνυσμα $x^t = (x_1, x_2, \dots, x_\mathrm{p})$, όπου p ο αριθμός των μεταβλητών και x^t το ανάστροφο διάνυσμα. Η πρόβλεψη της εξαρτημένης τιμής y μπορεί να υπολογιστεί με τον εξής τύπο:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} x_j \hat{\beta}_j,$$

Στην περίπτωση που στο διάνυσμα χ προστεθεί και ένας άσσος και ο συντελεστής β_0 στο διάνυσμα συντελεστών, ο τύπος μπορεί να γραφεί σαν γινόμενο 2 διανυσμάτων ως εξής:

$$\hat{y} = x^t \hat{\beta},$$

Η πιο διαδεδομένη τεχνική για την κατασκευή του γραμμικού μοντέλου πάνω σε ένα σύνολο εκπαίδευσης, που περιγράφει ένα φαινόμενο, είναι η μέθοδος των ελάχιστων τετραγώνων (MET-RSS). Σε αυτήν την μέθοδο επιλέγεται κάθε συντελεστής και σκοπός είναι η ελαχιστοποίηση της σχέσης. Ο τύπος είναι:

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x^t \beta)^2,$$

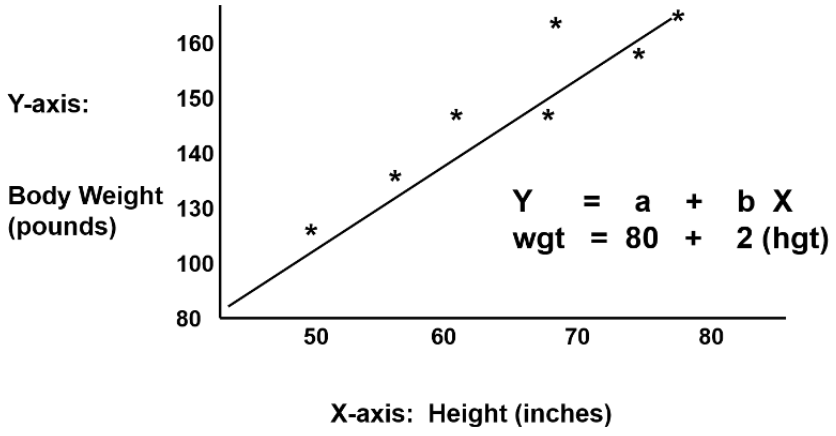
όπου n ο αριθμός παρατηρήσεων του συνόλου δεδομένων. Η συνάρτηση 3 είναι τετραγωνική, επομένως υπάρχει πάντα ελάχιστη τιμή. Η εξίσωση μπορεί να γραφεί:

$$RSS(\beta) = (y - X\beta)^t(y - X\beta),$$

όπου X ένας πίνακας n × p, με κάθε γραμμή του πίνακα να είναι ένα διάνυσμα του συνόλου εκπαίδευσης και y να είναι ένα διάνυσμα n μεγέθους που περιέχει τις τιμές εξόδου. Τέλος, αν το γινόμενο X^tX είναι αντιστρέψιμος πίνακας, η εξίσωση μπορεί να γραφεί:

$$\hat{\beta} = (X^tX)^{-1}X^ty.$$

Άλλα γνωστά γραμμικά μοντέλα είναι οι αλγόριθμοι παλινδρόμησης Ridge και Lasso. Αυτοί οι αλγόριθμοι κανονικοποιούν τις παραμέτρους, και, στην περίπτωση του Lasso μπορεί να εξαλείψει και εντελώς κάποιους συντελεστές. Η κανονικοποίηση που κάνει ο Ridge, είναι και γνωστή σαν L2, ενώ η κανονικοποίηση του Lasso ως L1. (Hastie, Tibshirani and Friedman, 2009)



Εικόνα 3.6 Παράδειγμα απλής γραμμικής παλινδρόμησης (Simple Linear Regression, 2022)

4.2.2 Πλησιέστεροι γείτονες

Ο αλγόριθμος των k-πλησιέστερων γειτόνων (KNN) είναι ένας απλός αλγόριθμος επιβλεπόμενης μάθησης, που μπορεί να χρησιμοποιηθεί τόσο για κατηγοριοποίηση, όσο και για παλινδρόμηση. Ο KNN υποθέτει ότι όμοιες εγγραφές έχουν μικρή απόσταση μεταξύ τους. Η χρήση του κατάλληλου μέτρου απόστασης από το πρόβλημα, ωστόσο μια δημοφιλής και συχνά χρησιμοποιούμενη προσέγγιση είναι η ευκλείδεια απόσταση.

Ο KNN βρίσκει τις αποστάσεις μεταξύ της εγγραφής, για το οποίο είναι να βρεθεί η ετικέτα και όλων των υπόλοιπων εγγραφών του συνόλου δεδομένων. Χαρακτηριστικό αυτού του αλγορίθμου είναι η υπερπαράμετρος κ, η οποία δηλώνει πόσους γείτονες να λαμβάνει υπόψιν για την εύρεση της ετικέτας. Για την εύρεση κατάλληλης τιμής της υπερπαραμέτρου k, πρέπει να τρέξουμε τον αλγόριθμο πολλές φορές με διαφορετικές τιμές k και επιλέγουμε εκείνη την τιμή, που ελαχιστοποιεί των αριθμό των σφαλμάτων και ταυτόχρονα συντηρεί την ικανότητα του αλγορίθμου να προβλέπει με ακρίβεια.

Αν πρόκειται για πρόβλημα κατηγοριοποίησης παίρνει την κατηγορία της πλειοψηφίας των κ κοντινότερων γειτόνων, ενώ στην περίπτωση της παλινδρόμησης η τιμή της ετικέτας ισούται με τον μέσο όρο των κ πλησιέστερων γειτόνων.

Ο αλγόριθμος είναι ο εξής:

Βήμα 1. Αρχικοποίηση της υπερπαραμέτρου **k**.

Βήμα 2. Για κάθε εγγραφή στο σύνολο δεδομένων:

1. Υπολόγισε την απόσταση μεταξύ της νέας εγγραφής και την τρέχων εγγραφή.
2. Πρόσθεση την απόσταση και τον δείκτη της εγγραφής σε μια συλλογή.

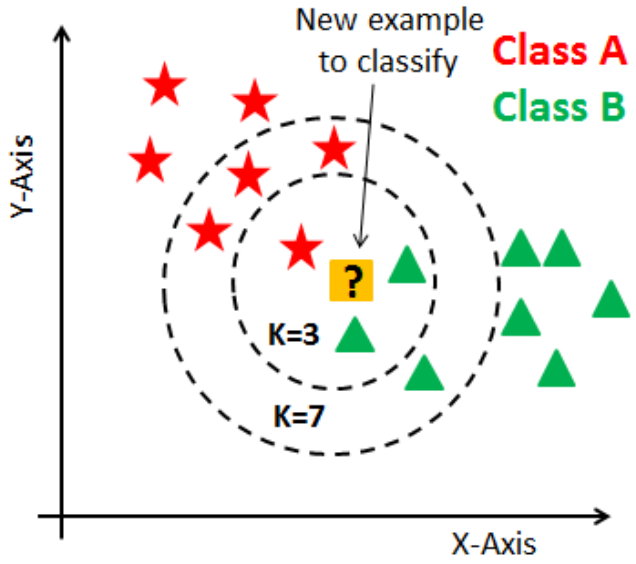
Βήμα 3. Ταξινόμησε σε αύξουσα σειρά την συλλογή που περιέχει τις αποστάσεις.

Βήμα 4. Επέλεξε τις **k** πρώτες αποστάσεις, μαζί με τους δείκτες των εγγραφών.

Βήμα 5. Βρες τις ετικέτες από τις πλησιέστερες εγγραφές.

Βήμα 6. Αν πρόκειται για πρόβλημα παλινδρόμησης, επέστρεψε τον μέσο όρο των πλησιέστερων εγγραφών.

Βήμα 7. Αν πρόκειται για πρόβλημα κατηγοριοποίησης, επέστρεψε την πλειοψηφία της κατηγορίας από τις πλησιέστερες εγγραφές.

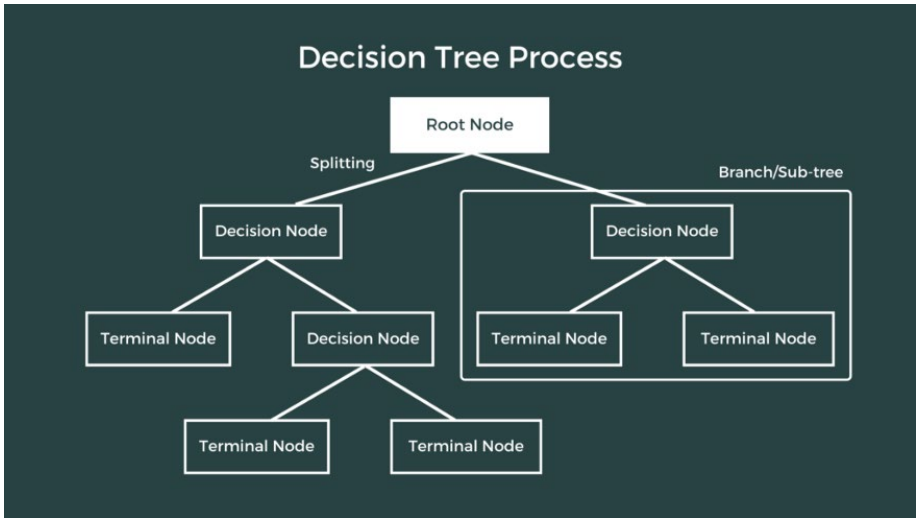


Εικόνα 3.11 Παράδειγμα του KNN αλγορίθμου. (K-Nearest Neighbor (KNN) Algorithm for Machine Learning, 2022)

4.2.3 Δέντρα απόφασης και Random Forests

Τα δέντρα απόφασης τόσο στην παλινδρόμηση, όσο και στην κατηγοριοποίηση χρησιμοποιούν μοντέλα μορφής δέντρου. Ένα Δέντρο απόφασης αποτελείται από:

- Ρίζα: Είναι η αρχή του δέντρου που αντιπροσωπεύει όλα τα δεδομένα.
- Κόμβους απόφασης: Είναι οι κόμβοι που διαχωρίζονται σε δύο ή περισσότερους κόμβους.
- Φύλλο/ Τερματικός κόμβος: Κόμβοι, οι οποίοι δεν διαχωρίζονται άλλο. Συνήθως, αυτοί οι κόμβοι είναι το τελικό αποτέλεσμα.

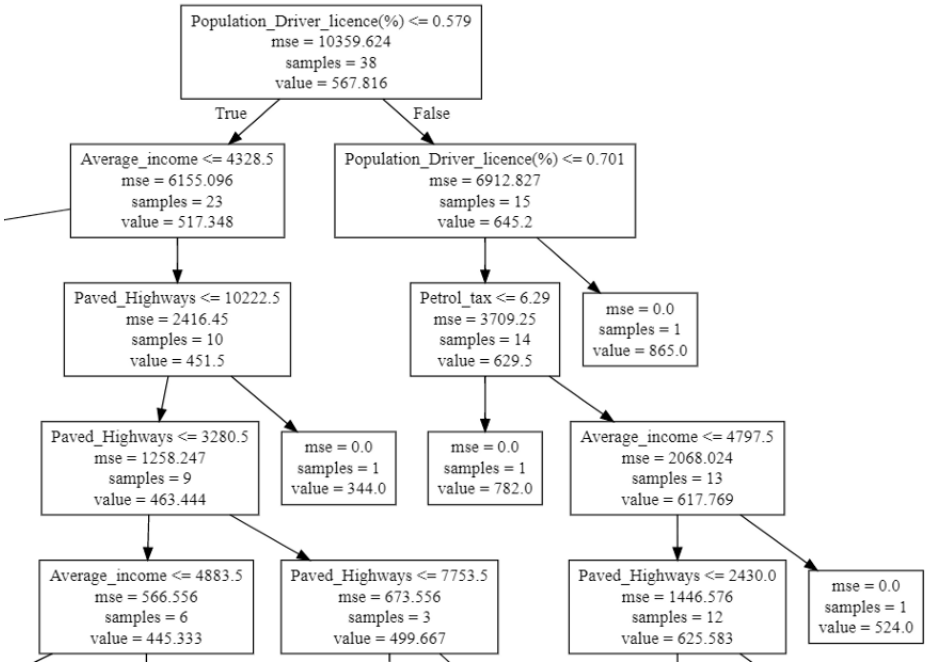


Εικόνα 4.5 Δομή απόφασης δέντρου (The Click Reader, 2017)

Η διαδικασία διαχωρισμού των κόμβων ξεκινά από την ρίζα και ακολουθείται από ένα υποδέντρο που οδηγεί σε τερματικό κόμβο που περιέχει την τελική πρόβλεψη. Η κατασκευή του δέντρου απόφασης ξεκινά από πάνω προς τα κάτω, διαλέγοντας σε κάθε κόμβο μια μεταβλητή που διαχωρίζει το σύνολο δεδομένων κατά τον βέλτιστο τρόπο.

Σε κάθε κόμβο υπολογίζεται με μια συνάρτηση κόστους η ακρίβεια του μοντέλου. Συνήθως, χρησιμοποιείται ο τύπος RMSE (ROOT MEAN SQUARE ERROR) και είναι:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$



Εικόνα 4.6. Παράδειγμα απόφασης δέντρου. (The Click Reader, 2017)

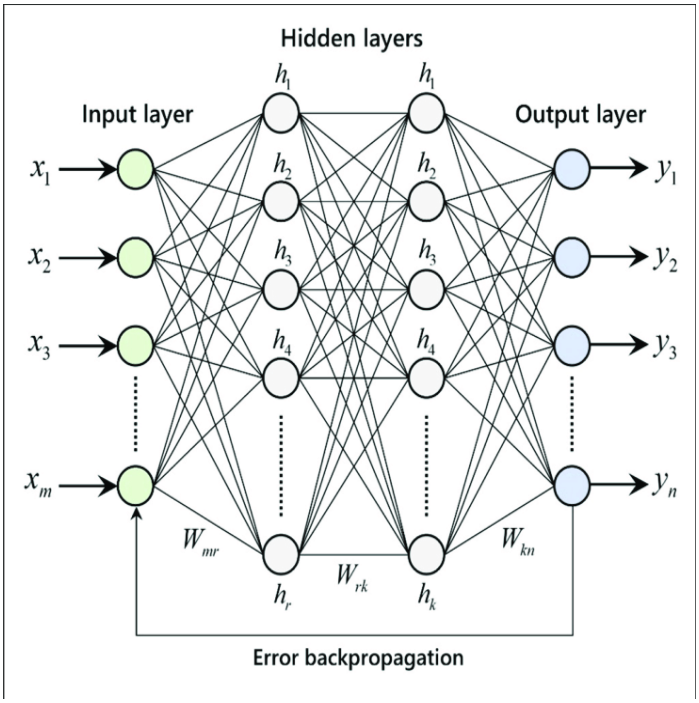
Τα Random Forests είναι ένα σύνολο από δέντρα απόφασης, όπου κάθε δέντρο απόφασης διαλέγει τυχαία κάποια χαρακτηριστικά. Αυτή η προσέγγιση βοηθά στον περιορισμό σφάλματος, λόγω μεροληψίας και διακύμανσης, κάνοντας τα Random Forests μια πιο εύρωστη τεχνική, σε σχέση με τα δέντρα απόφασης. (Liberman, 2017)

4.2.4 Τεχνητά Νευρωνικά Δίκτυα

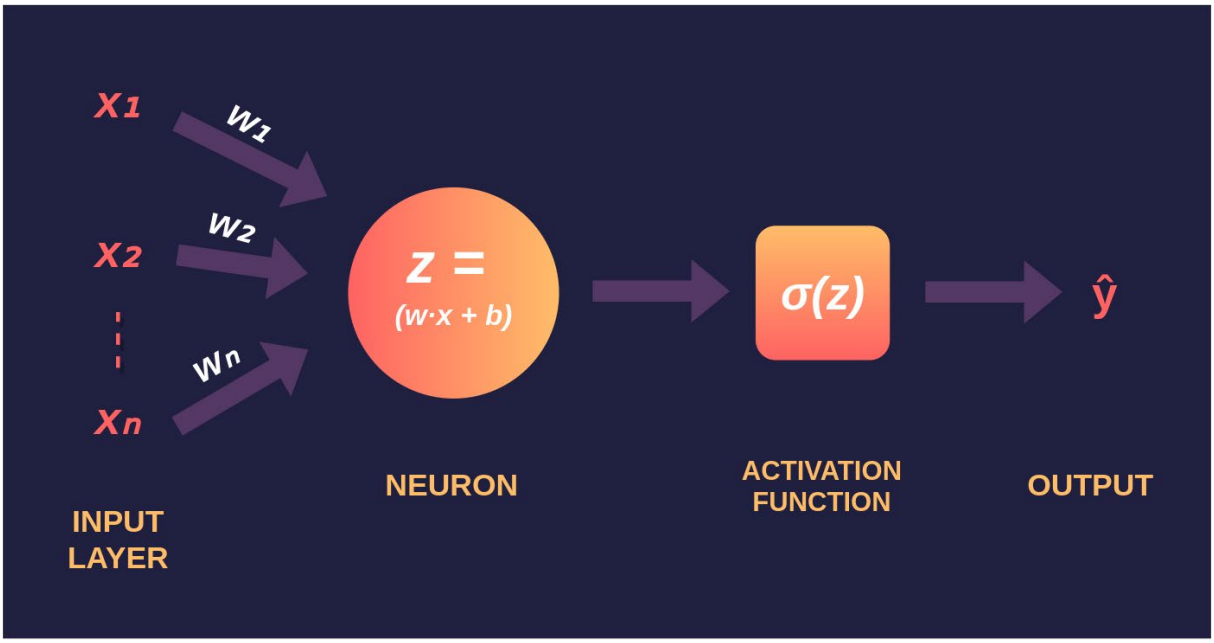
Στα μοντέλα τεχνητών νευρωνικών δικτύων (ΤΝΔ) η συνάρτηση που προσομοιώνει την σχέση ανάμεσα σε δεδομένα είναι η $y = f^*(x)$, όπου η $f(x)$ αποτελεί την σύνθεση περισσότερων συναρτήσεων και είναι της μορφής $f(x) = (f^{(n)} \circ f^{(n-1)} \circ \dots \circ f^{(1)})(x)$, όπου $f^{(i)}$ το i – οστό επίπεδο του δικτύου, n το βάθος του δικτύου και $f^{(n)}$ η έξοδος του δικτύου. Τα υπόλοιπα επίπεδα ονομάζονται κρυφά επίπεδα. Ένα ΤΝΔ μπορεί να αναπαρασταθεί με ένα κατευθυνόμενο ακυκλικό γράφο, που περιγράφει την σύνθεση μεταξύ των συναρτήσεων. Τα κρυφά επίπεδα έχουν την ευθύνη να αναγνωρίσουν τα διάφορα μοτίβα που μπορεί να υπάρχουν στα δεδομένα. Τα δεδομένα εκπαίδευσης κατευθύνουν το επίπεδο εξόδου, για να βγάζει όσο το δυνατόν πιο κοντινά αποτελέσματα στην

πραγματική έξοδο. Ο αλγόριθμος εκπαίδευσης είναι υπεύθυνος να αξιοποιήσει τα κρυφά επίπεδα με τέτοιον τρόπο ώστε να βγάζουν αποδεκτές λύσεις.

Το Perceptron είναι η πιο απλή μορφή ενός νευρωνικού δικτύου και εφευρέθηκε από τον Frank Rosenblatt το 1958. Αποτελούνται από n εισόδους, έναν νευρώνα και μια έξοδο, όπου n ο αριθμός των χαρακτηριστικών ενός συνόλου δεδομένων. Ένα ΤΝΔ έχει δύο φάσεις: την πρόσθια τροφοδότηση (forward propagation) και την ανατροφοδότηση (backpropagation). Η διαδικασία περάσματος δεδομένων μέσα στο ΤΝΔ είναι η πρόσθια τροφοδότηση, ενώ ανατροφοδότηση είναι η διαδικασία ενημέρωσης των βαρών.



Εικόνα 4.7. Ένα πολυεπίπεδο Τεχνητό Νευρωνικό Δίκτυο (Education, 2022)

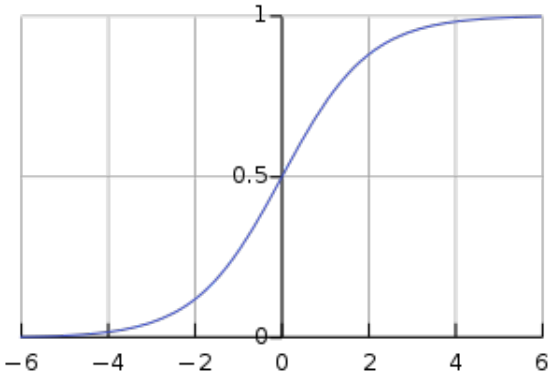


Εικόνα 4.8. Ένα απλό τεχνητό νευρωνικό δίκτυο. (A Gentle Introduction To Math Behind Neural Networks, 2022)

Κάθε είσοδος x_i πολλαπλασιάζεται με το αντίστοιχο βάρος του w_i και τα γινόμενα αθροίζονται. Επομένως, η σχέση είναι $\Sigma = (x_1 w_1) + (x_2 w_2) + \cdots + (x_n w_n)$. Αν x και w δύο διανύσματα, όπου $x = (x_1, x_2, \dots, x_n)$ και $w = (w_1, w_2, \dots, w_n)$, τότε το άθροισμα των γινομένων μπορεί να γραφεί $\Sigma = x \cdot w$. Τα βάρη αντιπροσωπούν την ισχύς μεταξύ των νευρώνων και αποφασίζουν κατά πόσο επηρεάζει κάθε χαρακτηριστικού την έξοδο του νευρώνα. Δηλαδή, ένα χαρακτηριστικό με μεγαλύτερη τιμή βάρους από ένα άλλο, επηρεάζει περισσότερο την έξοδο του νευρώνα από το χαρακτηριστικό με την μικρότερη τιμή βάρους. Επίσης, στο τελικό άθροισμα προστίθεται μια τιμή πόλωσης (bias), που αντισταθμίζει την συνάρτηση ενεργοποίησης και την μετακινεί προς τα δεξιά ή τα αριστερά για να παραχθούν επιθυμητές τιμές εξόδου. Η σχέση που δημιουργείται είναι η εξής $z = x \cdot w + b$, όπου b η τιμή πόλωσης. Η τιμή του z περνιέται από μια συνάρτηση ενεργοποίησης στην περίπτωση του Perceptron, ή περισσότερες όταν έχει περισσότερα κρυφά επίπεδα (μια συνάρτηση ενεργοποίησης για κάθε κρυφό επίπεδο). Οι συναρτήσεις ενεργοποίησης συμβάλλουν στον χρόνο εκπαίδευσης του ΤΝΔ. Η πιο συνηθισμένη συνάρτηση που χρησιμοποιείται σαν συνάρτηση ενεργοποίησης είναι η λογιστική (Εικόνα

2.4), γνωστή και ως σιγμοειδής. Χρησιμοποιώντας σαν συνάρτηση ενεργοποίησης την λογιστική, η πρόβλεψη του ΤΝΔ υπολογίζεται από τον τύπο:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{z}}$$



Εικόνα 4.9. Λογιστική συνάρτηση (Logistic function - Wikipedia, 2022)

Εφόσον, έχει υπολογιστεί η τιμή πρόβλεψης από το ΤΝΔ, το νευρωνικό δίκτυο ξεκινά την διαδικασία της ανατροφοδότησης, ξεκινώντας από την εύρεση του σφάλματος. Η ΜΕΤ είναι πιο διαδεδομένη μέθοδος εύρεσης σφαλμάτων και η διαδικασία είναι ίδια με αυτήν που αναφέρθηκε στην γραμμική παλινδρόμηση. Για την βελτιστοποίηση της εκπαίδευσης των τεχνητών νευρωνικών δικτύων χρησιμοποιείται συνήθως ο αλγόριθμος καθόδου με παραγώγους (gradient descent). Η διαδικασία της ανατροφοδότησης και ο αλγόριθμος καθόδου με παραγώγους επαναλαμβάνεται μέχρι να συγκλίνει και ενημέρωση των βαρών γίνεται ως εξής:

$$w_i = w_i - (a \cdot \frac{\partial C}{\partial w_i})$$

$$b = b - (a \cdot \frac{\partial C}{\partial b})$$

όπου C = MSE = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

και

$$\frac{\partial C}{\partial w_i} = \frac{2}{n} \cdot sum(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot x_i$$

$$\frac{\partial C}{\partial b} = \frac{2}{n} \cdot sum(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z))$$

4.2.5 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) είναι μια τεχνική η οποία ανήκει στην ομάδα των μηχανών εκμάθησης (learning machines) και ως στόχο έχει την επεξεργασία δεδομένων. Χρησιμοποιείται σε προβλήματα ταξινόμησης και στην προσέγγιση της μορφής της συνάρτησης σε προβλήματα παλινδρόμησης. Η λογική μια μηχανής εκμάθησης είναι να δίνει την τιμή yi μιας συνάρτησης (άγνωστη προς εμάς) που αντιστοιχεί σε δοσμένο σημείο $\vec{xi}$. Αυτό γίνεται ως εξής. Για Δεδομένο σύνολο l σημείων $\vec{xi} \in R^n$ και έχοντας τις αντίστοιχες τιμές yi∈R που παίρνει η άγνωστη συνάρτηση, εκπαιδεύουμε τη μηχανή εκμάθησης να μάθει τη σχέση που συνδέει τα $\vec{xi}$ με τα yi . Δηλαδή, η μηχανή μαθαίνει την αντιστοίχιση τα $\vec{xi} \rightarrow$yi και έτσι για ένα σημείο $\vec{xm}$, διαφορετικό από αυτά του συνόλου l της εκμάθησης, θα μας δώσει την τιμή ym που θα έπαιρνε η άγνωστη συνάρτηση.

Στην περίπτωση ταξινόμησης με τα SVM, το σύνολο των σημείων l αποτελείται από δύο υποσύνολα τα k και n. Έτσι, το αποτέλεσμα της συνάρτησης θα είναι +1 ή -1 (yi=+1 ή yi=-1) ανάλογα σε ποιο υποσύνολο ανήκει το δοθέν σημείο $\vec{xi}$. Τα δύο αυτά υποσύνολα ονομάζονται κλάσεις και η τιμή +1 (-1) είναι η "ετικέτα" της κλάσης. Δηλαδή, σε αυτή τη περίπτωση τα SVM μαθαίνουν να κατατάσσουν σωστά τα σημεία $\vec{xi}$ στις δύο κλάσεις. Τα σημεία $\vec{xi}$ και οι αντίστοιχες τιμές τους, yi, αποτελούν την πληροφορία εκπαίδευσης (training set). Τα σημεία $\vec{xi}$ ονομάζονται πρότυπα εκπαίδευσης (training patterns) ενώ οι τιμές yi που αντιστοιχούν σε αυτά, στόχοι εκπαίδευσης (training targets). Για να

αποφασιστεί σε ποια κλάση ανήκει ένα νέο σημείο, θα πρέπει να βρεθεί που είναι το όριο της κάθε κλάσης, δηλαδή να βρεθεί μια γραμμή (δισδιάστατος χώρος) που να διαχωρίζει τις δύο κλάσεις. Τα SVMs, για να πετυχαίνουν το διαχωρισμό των κλάσεων χρησιμοποιούν ευθείες. Έτσι, από τη σχετική θέση του προς κατάταξη σημείου και της διαχωριστικής ευθείας θα μπορεί να βγει το συμπέρασμα σε πιά κλάση ανήκει το σημείο.

Το σκεπτικό πίσω από τις μηχανές διανυσμάτων υποστήριξης είναι ότι αν έχουμε επιλέξει αυτό που μεγιστοποιεί το περιθώριο είναι λιγότερο πιθανό να κατηγοριοποιηθεί λάθος ένα άγνωστο αντικείμενο στο μέλλον.

4.2.6 Bayesian ταξινομητές

Η Naive Bayes κατηγοριοποίηση χρησιμοποιεί τις πιθανότητες για την επίλυση του προβλήματος κατάταξης σε κατηγορίες. Είναι μοντέλα τα οποία αναθέτουν «ταμπέλες» κλάσεων στα στοιχεία ενός προβλήματος, όπου οι κλάσεις προέρχονται από ένα πεπερασμένο σετ. Βασίζονται στην αρχή πως η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη της τιμής οποιουδήποτε άλλου χαρακτηριστικού δοσμένης της κλάσης. Τα κύρια οφέλη της Naive Bayes κατηγοριοποίησης είναι ότι δεν επηρεάζεται από απομονωμένα σημεία θορύβου και μη σχετικά χαρακτηριστικά, και μπορεί να χειρίζεται τις τιμές που λείπουν αγνοώντας τες κατά την εκτίμηση υπολογισμού πιθανοτήτων. Ωστόσο, η υπόθεση για την ανεξαρτησία των δεδομένων δεν μπορεί να ισχύει για ορισμένες ιδιότητες, που θα μπορούσαν να συσχετιστούν.

Βασίζεται τον ορισμό της δεσμευμένης πιθανότητας και του θεωρήματος Bayes. Το θεώρημα του Bayes, υπολογίζει την υπό συνθήκη πιθανότητα P($H|X$), δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση Η με δεδομένο ότι ισχύει το γεγονός Χ. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα P($H|X$) δίνεται από την Εξίσωση

$$\mathrm{P\ (H/X)} = \frac{P(H) * P(\frac{H}{X})}{P(X)}$$

όπου P(H) είναι η εκ των προτέρων πιθανότητα να ισχύει η υπόθεση H, P(X) είναι η εκ των προτέρων πιθανότητα να συμβεί το γεγονός X και P(X|H) είναι η πιθανότητα να συμβεί το γεγονός X με δεδομένο ότι ισχύει η υπόθεση H.

4.2.7 K-Means

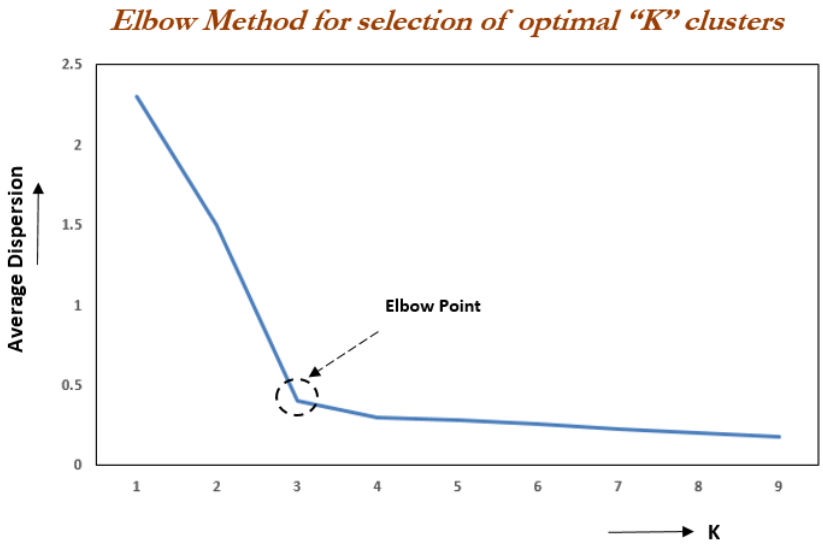
Ο αλγόριθμος K-means είναι ένας αλγόριθμος μη-επιβλεπόμενης μάθησης, ο οποίος χωρίζει το σύνολο δεδομένων σε ένα καθορισμένο αριθμό K μη επικαλυπτόμενων υποομάδων, όπου κάθε εγγραφή θα ανήκει σε μια μόνο ομάδα. Στόχος του αλγορίθμου είναι, οι εγγραφές που βρίσκονται σε μια ομάδα να είναι όσο πιο όμοια γίνεται με το κέντρο της συστάδας (centroid), με βάση τα χαρακτηριστικά τους. Η ομοιότητα υπολογίζεται με βάση μέτρα απόστασης, με κλασικό παράδειγμα την ευκλείδεια απόσταση.

Τα βήματα του αλγορίθμου K-Means είναι τα εξής:

1. Ομαδοποιεί τα δεδομένα σε K ομάδες.
2. Επιλέγει τυχαία σημεία K ως κέντρα των ομάδων.
3. Αντιστοιχίζει τα αντικείμενα στο πλησιέστερο κέντρο συμπλέγματος σύμφωνα με τη συνάρτηση ευκλείδειων αποστάσεων.
4. Υπολογίζει το centroid σε κάθε ομάδα (μέσο όλων των αντικειμένων σε κάθε ομάδα).
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.

Η αξιολόγηση της απόδοσης του αλγόριθμου γίνεται με βάση τον αριθμό των ομάδων. Μια διάσημη προσέγγιση για την αξιολόγηση αυτήν είναι η μέθοδος του αγκώνα (Elbow method). Με την μέθοδο του αγκώνα μπορούμε να βρούμε τον καλύτερο αριθμό για την παράμετρο K (αριθμό συστάδων). Βασίζεται στο άθροισμα της τετραγωνικής απόστασης (Sum Squared Error – SSE) μεταξύ των δεδομένων και των κεντροειδών. 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. (Dabbura, 2018)



Εικόνα 3.13. Αναπαράσταση του SSE με βάση τον αριθμό των K (Statistics for Machine Learning, 2022)

4.2.8 Συσταδοποίηση που βασίζεται στην πυκνότητα

Οι αλγόριθμοι αυτοί βασίζονται στην πυκνότητα γύρω από κάποιο σημείο που θεωρείται συστάδα. Αντιπροσωπευτικός αλγόριθμος είναι ο DBScan. Όμοια με την ομαδοποίηση με βάση τη σύνδεση, βασίζεται σε σημεία σύνδεσης εντός ορισμένων ορίων απόστασης. Όμως, συνδέει μόνο σημεία που, ικανοποιούν ένα κριτήριο πυκνότητας και στην αρχική παραλλαγή ορίζεται ως ένας ελάχιστος αριθμός αντικειμένων στο εσωτερικό αυτής της ακτίνας. Ένα σύμπλεγμα αποτελείται από όλα τα συνδεδεμένα αντικείμενα (τα οποία μπορούν να σχηματίσουν ένα σύμπλεγμα ενός αυθαίρετου σχήματος, σε αντίθεση με πολλές άλλες μεθόδους) συν όλα τα αντικείμενα που βρίσκονται εντός εμβέλειας αυτών των αντικειμένων. Ωστόσο η μέθοδος αναμένει κάποια μείωση της πυκνότητας για την ανίχνευση των συνόρων της συστάδας και δεν μπορεί να ανιχνεύσει τις εγγενείς δομές ομάδων που είναι διαδεδομένες στην πλειονότητα των δεδομένων πραγματικής ζωής.

4.2.9 Ιεραρχική Συσταδοποίηση

Η ομαδοποίηση που βασίζεται στη σύνδεση, επίσης γνωστή ως ιεραρχική ομαδοποίηση, βασίζεται στην κεντρική ιδέα πως τα αντικείμενα σχετίζονται πιο πολύ με τα αντικείμενα που είναι κοντά τους παρά με αυτά που είναι πιο μακριά. Αυτοί οι αλγόριθμοι συνδέουν "αντικείμενα" για να σχηματίσουν "συστάδες", με βάση την απόστασή τους. Ένα σύμπλεγμα μπορεί να περιγραφεί σε μεγάλο βαθμό από τη μέγιστη απόσταση που απαιτείται για τη σύνδεση τμημάτων του συμπλέγματος. Σε διαφορετικές αποστάσεις, θα σχηματιστούν διαφορετικές ομάδες, γεγονός που μπορεί να αναπαρασταθεί χρησιμοποιώντας ένα δενδρόγραμμα, που εξηγεί και την κοινή ονομασία «ιεραρχική ομαδοποίηση». Αυτοί οι αλγόριθμοι δεν παρέχουν μια ενιαία τμηματοποίηση του συνόλου των δεδομένων, αλλά αντίθετα προσφέρουν μια εκτεταμένη ιεραρχία συστάδων που συγχωνεύονται μεταξύ τους σε ορισμένες αποστάσεις. Ωστόσο δεν παράγουν μια μοναδική κατάτμηση του συνόλου των δεδομένων, αλλά μια ιεραρχία από την οποία ο χρήστης εξακολουθεί να πρέπει να επιλέξει τα κατάλληλα συμπλέγματα και επιπλέον δεν είναι πολύ ισχυροί έναντι των ακραίων τιμών, οι οποίες είτε θα εμφανίζονται ως πρόσθετοι πόλοι ή ακόμη και να προκαλέσουν συγχώνευση άλλων συστάδων.

4.2.10 TF-IDF

Κάθε κείμενο αναπαρίσταται υπό τη μορφή διανύσματος σε ένα πολυδιάστατο χώρο, όπου κάθε διάστασή του αντιπροσωπεύει ένα μοναδικό όρο μιας συλλογής κειμένων. Επίσης, σε κάθε

διάσταση αντιστοιχεί ένας πραγματικός αριθμός ο οποίος εξαρτάται από τη συχνότητα εμφάνισης του εκάστοτε όρου κάθε φορά στο κείμενο. Η μέθοδος TF-IDF στοχεύει στο να σταθμίσει όλους τους όρους μιας συλλογής κειμένων. Με λίγα λόγια δηλαδή, στόχος της είναι να αποδώσει το αντίστοιχο βάρος σε κάθε όρο και κατ' επέκταση σε κάθε διάσταση του πολυδιάστατου αυτού χώρου. Αυτό συμβαίνει γιατί η απλή αρίθμηση ενός όρου σε ένα κείμενο δεν αρκεί για να μας πληροφορήσει για τη σημαντικότητα του όρου αυτού και τη βαρύτητα της πληροφορίας που περιέχει. Η μέθοδος αυτή αποτελείται από τις ποσότητες TF και

IDF. Η ποσότητα TF (συχνότητα όρου) υποδηλώνει το πόσες φορές εμφανίζεται ένας όρος σε ένα κείμενο. Από την άλλη η ποσότητα IDF υποδηλώνει το πόσο ένας όρος είναι διαδεδομένος σε ένα κείμενο αλλά και σε ολόκληρη τη συλλογή κειμένων. Στόχος της μεθόδου αυτής μέσω του βάρους TF–IDF είναι η επιλογή εκείνων των όρων που αποτυπώνουν καλύτερα το περιεχόμενο ενός κειμένου. Για τον προσδιορισμό του βάρους ενός όρου είναι εξίσου σημαντικές και οι δύο ποσότητες TF και IDF. Υποθέτουμε πως N είναι ο συνολικός αριθμός στοιχείων που μπορούν να προταθούν στους χρήστες και πως η λέξη κλειδί k_i εμφανίζεται σε n_i από αυτά. Επιπλέον υποθέτουμε ότι $f_{i,j}$ είναι ο αριθμός των φορών που η λέξη κλειδί k_i εμφανίζεται στο έγγραφο di. Τότε η $TF_{I,j}$ είναι η term συχνότητα τη λέξης κλειδί ki στο έγγραφο di και προσδιορίζεται ως:

$$TF_{i,j}=\frac{f_{i,j}}{max_z f_{z,j}}$$

όπου το μέγιστο υπολογίζεται πάνω στις συχνότητες $f_{z,j}$ όλων των λέξεων κλειδιά k_z που εμφανίζονται στο έγγραφο d_j . Ωστόσο, υπάρχουν λέξεις κλειδιά που εμφανίζονται σε έγγραφα αλλά δεν είναι χρήσιμες στο διαχωρισμό σχετικού και μη σχετικού εγγράφου. Για το λόγο αυτό χρησιμοποιείται η inverse document frequency της λέξης κλειδί k_i ως

$$IDF_I = \log\frac{N}{ni}$$

Τότε το βάρος TF-IDF της λέξης κλειδί k_i στο έγγραφο dj προσδιορίζεται ως $w_{I,j} = TF_{i.j}*IDF_i$.

4.2.11 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης είναι μία μέθοδος βασισμένη σε κανόνες που χρησιμοποιεί η Μηχανική Μάθηση με στόχο την εύρεση σχέσεων ανάμεσα στις μεταβλητές μεγάλων βάσεων δεδομένων. Για την εύρεση των σχέσεων αυτών

χρησιμοποιούνται μετρικές ενδιαφέροντος. Χρησιμοποιούνται ευρέως στην ανάλυση της αγοράς, αλλά και στη Βιοπληροφορική, στην ανίχνευση επιθέσεων, κ.ά.

Το πρόβλημα των κανόνων συσχέτισης ορίζεται ως εξής (Agrawal, Imielinski and Swami, 1993):

Έστω πως το $I = \{i_1, i_2, \dots, i_n\}$ είναι ένα σύνολο από n δυαδικά χαρακτηριστικά που ονομάζονται *αντικείμενα* και έστω ότι το $D = \{t_1, t_2, \dots, t_m\}$ ένα σύνολο συναλλαγών που ονομάζεται *βάση δεδομένων*. Κάθε συναλλαγή D έχει ένα μοναδικό αριθμό συναλλαγής και περιέχει ένα υποσύνολο αντικειμένων από το I. Σαν κανόνα ορίζουμε μία σχέση της μορφής $X \Rightarrow Y, X, Y \subseteq I$. Κάθε κανόνας αποτελείται από δύο σύνολα αντικειμένων, τα X, Y, με το X να ονομάζεται *προηγούμενο* ή αριστερή πλευρά (left hand side) και το Yονομάζεται *προκύπτων* ή δεξιά πλευρά (right hand side). Ένα πολύ απλό παράδειγμα κανόνα συσχέτισης είναι το παρακάτω: $\{αυγά, αλεύρι\} \Rightarrow \{γάλα\}$, που σημαίνει ότι αν έχουν αγοραστεί αυγά και αλεύρι, θα αγοραστεί και γάλα.

Για την σωστή επιλογή κανόνων από όλους τους πιθανούς συνδυασμούς κανόνων που μπορούν να υπάρχουν χρησιμοποιούνται διάφορες μετρικές. Οι πιο δημοφιλείς είναι η *υποστήριξη* (support) και η *εμπιστοσύνη* (confidence), οι τιμές των οποίων πρέπει να περνούν κάποια ελάχιστα κατώφλια (thresholds).

Έστω τα X, Y σύνολα αντικειμένων, $X \Rightarrow Y$ ο κανόνας συσχέτισης και T ένα σύνολο συναλλαγών μίας δεδομένης βάσης δεδομένων.

Η υποστήριξη μας δείχνει πόσο συχνά εμφανίζεται το σύνολο αντικειμένων στη βάση δεδομένων και σαν υποστήριξη του X ως προς το T, ορίζεται το ποσοστό των συναλλαγών t στη βάση δεδομένων οι οποίες περιέχουν το σύνολο αντικειμένων X:

$$supp(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Η εμπιστοσύνη είναι ενδεικτικό του πόσο συχνά έχει αποδειχθεί ο κανόνας σωστός. Πιο συγκεκριμένα, η εμπιστοσύνη ενός κανόνα $X \Rightarrow Y$, ως προς ένα σετ συναλλαγών T, είναι το ποσοστό των συναλλαγών που περιέχουν τόσο το X όσο και το Y:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Η εμπιστοσύνη μπορεί να θεωρηθεί σαν μία εκτίμηση της υπό συνθήκη πιθανότητας $P(E_X)$, όπου E_X, E_Y, τα γεγονότα ότι μία συναλλαγή περιλαμβάνει τα σύνολα αντικειμένων X, Y αντίστοιχα. Είναι ουσιαστικά η πιθανότητα να βρεθεί η δεξιά πλευρά του κανόνα σε συναλλαγές, υπό την προϋπόθεση ότι αυτές οι συναλλαγές περιέχουν και την αριστερή πλευρά του κανόνα.

Βιβλιογραφία

Agrawal, R., Imieliński, T. and Swami, A., 1993, June. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).

Analytics Vidhya. 2022. K Means Clustering Simplified in Python | K Means Algorithm. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/> [Accessed 11 July 2022].

Burkov, A., 2019. The Hundred-Page Machine Learning Book by Andriy Burkov.

Education, I., 2022. *What are Neural Networks?*. [online] Ibm.com. Available at: <https://www.ibm.com/cloud/learn/neural-networks> [Accessed 17 April 2022].

En.wikipedia.org. 2022. *Logistic function - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Logistic_function> [Accessed 17 April 2022].

Fisher RA. 1925. Theory of statistical estimation. ***Proc. Cambridge Philos. Soc.*** , pp. 700–725.

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". The Journal of the Anthropological Institute of Great Britain and Ireland, pp. 246–263.

Gorunescu, F., 2011. *Data mining*. Berlin: Springer.

Han, J., Kamber, M. and Pei, J., 2012. *DATA MINING*. 3rd: MORGAN KAUFMANN,p.26, 41-65.

Liberman, N., 2017. Decision Trees and Random Forests.

Medium. 2022. *A Gentle Introduction To Math Behind Neural Networks*. [online] Available at: <https://towardsdatascience.com/introduction-to-math-behind-neural-networks-e8b60dbbdeba> [Accessed 17 April 2022].

Medium. 2022. K-Nearest Neighbor(KNN) Algorithm for Machine Learning. [online] Available at: <https://medium.com/@sudhanshugupta_66164/k-nearest-neighbor-knn-algorithm-for-machine-learning-1b506eb2c4a4> [Accessed 11 July 2022].

Medium. 2018. *What are the types of machine learning?*. [online] Available at: <https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f> [Accessed 15 April 2022].

Mitchell, T., 1997. Machine Learning. MacGraw-Hill Companies. *Inc., Boston*. Mucherino, A., Papajorgji, P. and Pardalos, P., 2009. *Data mining in agriculture*. Dordrecht: Springer, pp.19-20.

Pearson, K. and Lee, A. (1903) On the Laws of Inheritance in Man: Inheritance of Physical Characters. Biometrika, pp. 357-462

Samuel, A., 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*

Tan, P., Steinbach, M., Karpatne, A. and Kumar, V., 2006. *Introduction to data mining*. pp.2,3.

Taunk, K., De, S., Verma, S. and Swetapadma, A., 2019. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*.

The Click Reader, 2017. *Decision Tree Regression Explained with Implementation in Python*. [online] Medium. Available at: <https://medium.com/@theclickreader/decision-tree-regression-explained-with-implementation-in-python-1e6e48aa7a47> [Accessed 17 April 2022].

Yule G U, 1897a,``On the theory of correlation'' Journal of the Royal Statistical Society , pp. 812 -854

Κύρκος Ε., 2015. Εξόρυξη Γνώσης από Δεδομένα. [online] <https://repository.kallipos.gr/bitstream/11419/1233/2/Kef._6.pdf> [Accessed 13 April 2022].