# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

## Πρόγραμμα Μεταπτυχιακών Σπουδών
## «Προηγμένα Συστήματα Πληροφορικής»

## Μεταπτυχιακή Διατριβή

| | |
|---|---|
| Τίτλος Διατριβής | **Εξόρυξη Δεδομένων και Μηχανική Μάθηση**<br>**Data Mining and Machine Learning** |
| Ονοματεπώνυμο Φοιτητή | **Γριβοκωστόπουλος Κωνσταντίνος** |
| Πατρώνυμο | **Ιωάννης** |
| Αριθμός Μητρώου | **ΜΠΣΠ/16041** |
| Επιβλέπων | **Βέργαδος Δημήτριος, Καθηγητής** |

Ημερομηνία Παράδοσης

**Ιούλιος 2024**

**Τριμελής Εξεταστική Επιτροπή**

| | | |
|---|---|---|
| Δημήτριος Βέργαδος | Ευθύμιος Αλέπης | Διονύσιος Σωτηρόπουλος |
| Καθηγητής | Καθηγητής | Επ. Καθηγητής |

# Περιεχόμενα

## Περίληψη

Η παρουσία της αβεβαιότητας στις επιχειρηματικές δραστηριότητες είναι σταθερός παράγοντας από την ίδρυση των εταιρειών, συνεχίζοντας να διαδραματίζει σημαντικό ρόλο στο σημερινό επιχειρηματικό τοπίο. Καθώς οι επιχειρήσεις αντιμετωπίζουν αυτή την αβεβαιότητα, έχουν στραφεί σε τεχνικές συλλογής και ανάλυσης δεδομένων για να εξαγάγουν πολύτιμες γνώσεις. Με την πρόοδο της τεχνολογίας και τον

παγκόσμιο χαρακτήρα των επιχειρήσεων, ο τεράστιος όγκος δεδομένων που πρέπει να προηγηθούν οι εταιρείες έχει κάνει τις παραδοσιακές μεθόδους έρευνας απαρχαιωμένες. Κατά συνέπεια, οι επιχειρήσεις έπρεπε να αγκαλιάσουν νέους επιστημονικούς κλάδους και να ενισχύσουν τους υπάρχοντες για να διαχειριστούν αποτελεσματικά και να αξιοποιήσουν αυτόν τον τεράστιο όγκο δεδομένων για ανταγωνιστικό πλεονέκτημα. Η παρούσα διπλωματική εργασία στοχεύει να διερευνήσει και να αναλύσει σε βάθος τις έννοιες της μηχανικής μάθησης και της εξόρυξης δεδομένων.

## Abstract

The presence of uncertainty in business activities has been a constant factor since the inception of companies, continuing to play an important role in today's business landscape. As businesses face this uncertainty, they have turned to data collection and analysis techniques to extract valuable insights. With the advancement of technology and the global nature of business, the sheer volume of data that companies must navigate has rendered traditional research methods obsolete. Consequently, businesses have had to embrace new disciplines and strengthen existing ones to effectively manage and leverage this vast amount of data for competitive advantage. This thesis aims to explore and analyze in depth the concepts of machine learning and data mining.

## Σύνοψη

Αυτή η διατριβή εμβαθύνει στη θεωρητική εξερεύνηση και ανάλυση της μηχανικής μάθησης και της εξόρυξης δεδομένων, εστιάζοντας στην ανάπτυξη και εφαρμογή αλγορίθμων για την αντιμετώπιση των προκλήσεων του πραγματικού κόσμου. Αποτελούμενη από έξι κεφάλαια, η παρούσα διπλωματική στοχεύει να παρέχει μια διεξοδική εξέταση του αντικειμένου.

1. Εισαγωγή στο Data mining

2. Εισαγωγή στο Machine Learning

3. Ανασκόπηση αλγορίθμων μηχανικής μάθησης

4. Αξιολόγηση της μηχανικής μάθησης

5. Ταξινόμηση (classification) και εφαρμογές

6. Machine learning στο Azure

7. Παρουσίαση των Γλωσσών Προγραμματισμού R και Python

# 1. Εισαγωγή στη θεωρία της Εξόρυξης Δεδομένων

Η διαδικασία της εξόρυξης δεδομένων μας δίνει την δυνατότητα να ανακαλύψουμε πολύτιμες και απλές συσχετίσεις και μοτίβα εντός των υπαρχόντων δεδομένων. Διάφοροι τομείς αναφέρονται σε αυτή τη διαδικασία με διαφορετικούς όρους, όπως εξόρυξη γνώσης, ανακάλυψη πληροφοριών, συλλογή πληροφοριών και επεξεργασία δεδομένων. Αρχικά, οι στατιστικολόγοι, οι ερευνητές βάσεων δεδομένων και η επιχειρηματική κοινότητα εισήγαγαν τον όρο "εξόρυξη δεδομένων" για να αντλήσουν χρήσιμες πληροφορίες από τα δεδομένα. Ένα σημαντικό στοιχείο της εξόρυξης δεδομένων είναι η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (KDD), η οποία στοχεύει κυρίως στην εντοπισμό πολύτιμων πληροφοριών εντός ενός συνόλου δεδομένων. Η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων είναι μια ολοκληρωμένη διαδικασία που περιλαμβάνει διάφορα στάδια, συμπεριλαμβανομένης της προετοιμασίας των δεδομένων, της επιλογής, του καθαρισμού και της ακριβούς ερμηνείας. Ο βασικός στόχος αυτής της διαδικασίας είναι η εξαγωγή πολύτιμων και ουσιαστικών πληροφοριών από μεγάλα σύνολα δεδομένων. Σε αντίθεση με τις παραδοσιακές μεθόδους ανάλυσης δεδομένων και στατιστικών προσεγγίσεων, η εξόρυξη δεδομένων ενσωματώνει ένα ευρύ φάσμα τεχνικών που προέρχονται από πολλά πεδία. [51]Αυτά περιλαμβάνουν αριθμητική ανάλυση, αντιστοίχιση προτύπων και κλάδους εντός της τεχνητής νοημοσύνης, όπως η μηχανική μάθηση και τα νευρωνικά δίκτυα. Αξιοποιώντας αυτές τις διεπιστημονικές προσεγγίσεις, η εξόρυξη δεδομένων επιτρέπει τον εντοπισμό κρυφών προτύπων και τάσεων που μπορεί να μην είναι άμεσα εμφανείς μέσω συμβατικών μέσων ανάλυσης. Κατά συνέπεια, το KDD διαδραματίζει κεντρικό ρόλο στην αποκάλυψη πολύτιμων γνώσεων που μπορούν να οδηγήσουν τις διαδικασίες λήψης αποφάσεων με πληροφόρηση και να διευκολύνουν τις εξελίξεις σε διάφορους κλάδους και τομείς. Ορισμένες διαδικασίες εξόρυξης δεδομένων ακολουθούν την παραδοσιακή προσέγγιση που βασίζεται σε υποθέσεις στην ανάλυση δεδομένων. Επιπλέον, είναι σύνηθες να υιοθετείται μια προσέγγιση που εστιάζει στα ίδια τα δεδομένα, βοηθώντας τον αλγόριθμο εξόρυξης στον εντοπισμό τάσεων, προτύπων και σχέσεων που μπορούν να συμβάλλουν στη λήψη αποφάσεων. Αυτές οι δύο προσεγγίσεις αποδίδουν διαφορετικά αποτελέσματα, είτε ένα μοντέλο είτε ένα πρότυπο. Η προσέγγιση του μοντέλου μοιάζει με παραδοσιακές στατιστικές μεθόδους, με στόχο να συνοψίσει ένα σύνολο δεδομένων και να εντοπίσει και να περιγράψει τα κυρίαρχα χαρακτηριστικά του. Παραδείγματα τέτοιων μοντέλων περιλαμβάνουν ανάλυση συστάδων, μοντέλα παλινδρόμησης για πρόβλεψη και ταξινόμηση βάσει δέντρων. Στη δημιουργία μοντέλων, μερικές φορές χρησιμοποιούνται εμπειρικά ή μηχανιστικά μοντέλα.

Τα εμπειρικά μοντέλα, γνωστά και ως business μοντέλα, στοχεύουν στη δημιουργία σχέσεων χωρίς να βασίζονται σε καμία υποκείμενη θεωρία. Από την άλλη πλευρά, τα μηχανιστικά μοντέλα, γνωστά επίσης ως φαινομενολογικά μοντέλα, βασίζονται σε θεωρίες ή μηχανισμούς παραγωγής δεδομένων.[43] Κατά συνέπεια, η εστίαση της εξόρυξης δεδομένων, εξ ορισμού, είναι πρωτίστως στην εκπλήρωση των επιχειρηματικών αναγκών. Μια άλλη πτυχή της εξόρυξης δεδομένων είναι η ανίχνευση προτύπων, η οποία προσπαθεί να εντοπίσει μικρές αποκλίσεις από τον κανόνα που μπορεί να υποδηλώνουν ασυνήθιστα

πρότυπα συμπεριφοράς. Ένα παράδειγμα αυτού περιλαμβάνει την παρακολούθηση ασυνήθιστων δαπανών για τον εντοπισμό απάτης με πιστωτικές κάρτες. Η βασική ιδέα της εξόρυξης δεδομένων είναι η εξαγωγή πολύτιμων πληροφοριών από τεράστιες ποσότητες δεδομένων χρησιμοποιώντας αυτές τις μεθόδους. Ωστόσο, όταν πρόκειται για λειτουργικές βάσεις δεδομένων, η εξαγωγή μοντέλων μπορεί να είναι δύσκολη λόγω της πολυπλοκότητας των δεδομένων. Επιπλέον, η προγνωστική ισχύς των αλγορίθμων εξόρυξης μπορεί να μειωθεί καθώς αυξάνεται ο αριθμός των ανωμαλιών στα δεδομένα. Ένα κρίσιμο βήμα προ επεξεργασίας στην εξόρυξη δεδομένων είναι η κατασκευή μιας αποθήκης δεδομένων, η οποία περιλαμβάνει τον καθαρισμό και την ενοποίηση των δεδομένων. Ωστόσο, αυτό το βήμα δεν είναι υποχρεωτικό και μπορεί να παραλειφθεί σε περιπτώσεις όπου μια μεγαλύτερη αποθήκη δεδομένων περιέχει δεδομένα από πολλές πηγές. Σε τέτοιες περιπτώσεις, το έργο γίνεται μνημειώδες, χρειάζεται χρόνια για να ολοκληρωθεί και κοστίζει εκατομμύρια ευρώ. Οι αποθήκες δεδομένων μπορεί να είναι λειτουργικές ή σχεσιακές βάσεις δεδομένων, καθώς και λογικές ή φυσικές βάσεις δεδομένων. Η συγχώνευση τεχνητής νοημοσύνης και στατιστικών τεχνικών οδηγεί σε συστήματα Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων (KDD), τα οποία βοηθούν στην εύρεση συσχετίσεων, αλληλουχιών, ταξινομήσεων, ομάδων και προβλέψεων. Η σχεσιακή βάση χρησιμεύει ως το σημείο εισόδου για τα περισσότερα δεδομένα, τα οποία στη συνέχεια καθαρίζονται και μεταφέρονται στην αποθήκη δεδομένων για επεξεργασία και παρουσίαση. Μετά από μια ορισμένη περίοδο, τα δεδομένα είτε καθαρίζονται, συνοψίζονται μαζί με άλλες πληροφορίες ή αρχειοθετούνται.

Οι τρεις τυπικές συνιστώσες στην αρχιτεκτονική που αφορά την αποθήκευση δεδομένων είναι οι εξής:

- Το backend, γνωστό και ως λογισμικό απόκτησης δεδομένων, είναι ένα βασικό στοιχείο που είναι απαραίτητο για την εξαγωγή δεδομένων από παλαιότερα συστήματα και εξωτερικές πηγές. Μόλις συλλεχθούν τα δεδομένα, στη συνέχεια υποβάλλονται σε επεξεργασία, ενοποιούνται και συνοψίζονται πριν φορτωθούν στην αποθήκη δεδομένων.
- Το λογισμικό front-end, σχεδιασμένο τόσο για χρήστες όσο και για εφαρμογές, διευκολύνει την ανάκτηση και την εξέταση δεδομένων εντός του Datawarehouse.

## 1.1 Γενικά

Ο όρος «εξόρυξη δεδομένων» χρησιμοποιείται για να περιγράψει ένα συγκεκριμένο σύνολο δραστηριοτήτων που εξάγουν πολύτιμες νέες πληροφορίες από δεδομένα. Αυτή η έννοια δεν είναι νέα για τους στατιστικολόγους και αναφέρεται επίσης ως διαγραφή δεδομένων ή εξόρυξη δεδομένων, με σκοπό την ανακάλυψη προτύπων.Όταν τα δεδομένα χρησιμοποιούνται επανειλημμένα για την αναγνώριση ή την επιλογή μοτίβων, μπορεί να προκύψει μια διαδικασία που ονομάζεται "σκάρωση". Αυτό συμβαίνει όταν τα αλγόριθμα εξόρυξης δεδομένων εφαρμόζονται επανειλημμένα σε ένα σύνολο δεδομένων για την ανίχνευση μοτίβων. Ωστόσο, αυτή η διαδικασία μπορεί να είναι εξαντλητική και να παραβλέπει ορισμένα μοτίβα, καθώς πολλά από αυτά είναι αποτέλεσμα τυχαίων παραλλαγών και μπορεί να μην είναι πραγματικά σημαντικά.Ο στόχος της ανάλυσης δεδομένων είναι να μοντελοποιήσει την υποκείμενη δομή των δεδομένων, αναδεικνύοντας συνεπή και ουσιαστικά μοτίβα. Επομένως, είναι σημαντικό να αναγνωρίζουμε τα πραγματικά ζητήματα και τις σημαντικές πληροφορίες που προκύπτουν από τα δεδομένα μας.Η εξόρυξη δεδομένων προσφέρει ένα χρήσιμο εργαλείο για τις επιχειρήσεις, καθώς τους επιτρέπει να επικεντρωθούν στις πιο σχετικές πληροφορίες στις βάσεις δεδομένων τους. Ωστόσο, είναι ουσιώδες να κατανοούμε πλήρως την επιχειρηματική λογική, τα δεδομένα που έχουμε στη διάθεσή μας, καθώς και τις μεθόδους ανάλυσης που χρησιμοποιούνται, προκειμένου να εξάγουμε σημαντικές και χρήσιμες πληροφορίες από αυτά.Είναι σημαντικό να αναφέρουμε ότι οι μέθοδοι πρόβλεψης που προέρχονται από την εξόρυξη δεδομένων μπορεί να μην είναι απαραίτητα η αιτία μιας συγκεκριμένης συμπεριφοράς. Υπάρχουν πολλοί παράγοντες που μπορούν να επηρεάσουν την ακρίβεια των συμπερασμάτων, συμπεριλαμβανομένης της μεταβλητότητας, της επιλογής δειγμάτων, του πληθυσμού δεδομένων και της ισοδυναμίας του μοντέλου. Είναι σημαντικό να αναγνωρίσουμε ότι οι σχέσεις που προσδιορίζονται μέσω της εξόρυξης δεδομένων μπορεί να μην

αντικατοπτρίζουν απαραίτητα την πραγματική αξία του μοντέλου για τον οργανισμό. Ως εκ τούτου, είναι σημαντικό να επικυρωθούν αυτά τα μοντέλα στο κατάλληλο πλαίσιο. Πολλές βιομηχανίες έχουν βιώσει θετικά αποτελέσματα από τη χρήση της εξόρυξης δεδομένων, οδηγώντας σε εξοικονόμηση κόστους και αύξηση εσόδων. [38]Ακολουθούν ορισμένες μέθοδοι που μπορούν να χρησιμοποιηθούν για να επιτευχθούν αυτά τα οφέλη:

- Βοηθά στη μείωση των εξόδων στα αρχικά στάδια του κύκλου ζωής ενός προϊόντος κατά τη φάση έρευνας και ανάπτυξης, εξοικονομώντας τελικά χρήματα για την εταιρεία.
- Χρήση περιορισμών στις στατιστικές διαδικασίες και την εξόρυξη σε αυτοματοποιημένες διαδικασίες παραγωγής.
- Ραγδαία μείωση στο κόστος αλληλογραφίας με την αποφυγή αποστολής αλληλογραφίας σε πελάτες που δεν είναι δεκτικοί σε προσφορές.
- Η ενίσχυση των στρατηγικών μάρκετινγκ και των εξατομικευμένων προσφορών μέσω της διαχείρισης πελατειακών σχέσεων είναι μια κοινή πρακτική μεταξύ των επιχειρήσεων. Πολλές εταιρείες χρησιμοποιούν την εξόρυξη δεδομένων για να προσελκύσουν και να διατηρήσουν πελάτες σε κάθε στάδιο της διαδρομής των πελατών.
- Το προφίλ πελατών είναι ένα πολύτιμο εργαλείο για τις εταιρείες να εντοπίζουν πιθανούς πελάτες που μοιράζονται παρόμοια χαρακτηριστικά και να τους στοχεύουν για προσφορές προϊόντων. Αυτό επιτρέπει στις εταιρείες να αναλύουν τη συμπεριφορά των καταναλωτών, να προσελκύουν νέους πελάτες και να εμποδίζουν τους τρέχοντες πελάτες να φύγουν (μείωση των ποσοστών εκτροπής). Είναι πιο οικονομικό για τις εταιρείες να διατηρούν υπάρχοντες πελάτες παρά να αποκτούν συνεχώς νέους.

Η εξόρυξη δεδομένων είναι ένα πολύ επίκαιρο θέμα και επομένως μπορεί να συμβάλει στις περισσότερες επιχειρηματικές ροές. Μερικά παραδείγματα των τομέων όπου μπορεί να συμβάλει είναι:

- Εταιρείες τηλεπικοινωνιών και πιστωτικών καρτών -οι δύο αυτές εταιρείες είναι πρωτοπόρες στην εφαρμογή της εξόρυξης δεδομένων με πρωταρχικό τους στόχο την ανίχνευση απάτης.
- Οι εταιρείες λιανικής χρησιμοποιούν δεδομένα πωλήσεων και στοιχεία πελατών για να κατανοήσουν τις ανάγκες των καταναλωτών και να προωθήσουν προϊόντα.
- Οι χρηματοπιστωτικές εταιρείες αναλύουν τα δεδομένα συναλλαγών και τις χρηματοοικονομικές συνήθειες για να ανιχνεύσουν ανωμαλίες ή τάσεις στις αγορές.
- Οι εταιρείες τεχνολογίας χρησιμοποιούν δεδομένα χρήστη για να βελτιστοποιήσουν τις υπηρεσίες τους και να προσφέρουν εξατομικευμένες εμπειρίες στους χρήστες.
- Οι εταιρείες ασφαλείας χρησιμοποιούν δεδομένα καταγραφής συμβάντων και ανάλυσης κινδύνου για την ανίχνευση και την αποτροπή κυβερνοεπιθέσεων.
- Οι επιστημονικοί ερευνητές χρησιμοποιούν δεδομένα από πειράματα και μετρήσεις για να ανακαλύψουν νέες γνώσεις και να κατανοήσουν φαινόμενα σε διάφορους τομείς, όπως η φυσική, η ιατρική και η κλιματολογία.

## 1.2 Σχέση με την στατιστική

Οι στατιστικές μέθοδοι και η εξόρυξη δεδομένων στοχεύουν στην αποκάλυψη συσχετισμών ή μοτίβων στα δεδομένα. Ενώ ορισμένοι μπορεί να θεωρούν την εξόρυξη δεδομένων ως ένα υποσύνολο στατιστικών λόγω των παρόμοιων στόχων τους, είναι σημαντικό να αναγνωρίσουμε ότι η εξόρυξη δεδομένων ενσωματώνει ιδέες, εργαλεία και μεθόδους από διάφορους κλάδους, όπως η τεχνολογία βάσεων δεδομένων και η μηχανική μάθηση. Επιπλέον, η εξόρυξη δεδομένων εστιάζει σε διαφορετικά θέματα σε σύγκριση με τους στατιστικολόγους. [50]Οι στατιστικές, ωστόσο, διαδραματίζουν κρίσιμο ρόλο στην εξόρυξη δεδομένων,

ιδιαίτερα στην ανάπτυξη και αξιολόγηση μοντέλων. [50]Οι στατιστικές δοκιμές αποτελούν σημαντικό εργαλείο για τους αλγορίθμους μηχανικής μάθησης σε πολλά επίπεδα. Καταρχάς, χρησιμοποιούνται για την κατασκευή κανόνων και δέντρων απόφασης. Αυτό συμβαίνει καθώς οι αλγόριθμοι αυτοί προσπαθούν να εκπαιδεύσουν μοντέλα που μπορούν να προβλέπουν την τάση ή την κατηγορία στην οποία ανήκουν νέα δεδομένα, με βάση τα ήδη υπάρχοντα. Επιπλέον, οι στατιστικές δοκιμές είναι χρήσιμες για την αντιμετώπιση του φαινομένου της υπερπροσαρμογής (overfitting). Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης, με αποτέλεσμα να χάνει την ικανότητα γενίκευσης σε νέα δεδομένα. [41]Οι στατιστικές δοκιμές μπορούν να βοηθήσουν στην ανίχνευση και την αποτροπή αυτού του φαινομένου, επιτρέποντας τη βελτίωση της απόδοσης των μοντέλων. Συνήθως, αυτές οι δοκιμές χρησιμοποιούνται όχι μόνο για την επικύρωση μοντέλων μηχανικής μάθησης και την αξιολόγηση τους, αλλά και για την αξιολόγηση της απόδοσής τους. Κατά τη διάρκεια αυτών των δοκιμών, τα μοντέλα εκπαιδεύονται σε ένα σύνολο δεδομένων εκπαίδευσης και στη συνέχεια αξιολογούνται σε ένα ανεξάρτητο σύνολο δεδομένων ελέγχου. Το σύνολο δεδομένων ελέγχου συνήθως δεν χρησιμοποιείται κατά την εκπαίδευση του μοντέλου, και χρησιμοποιείται για να δοκιμαστεί η ικανότητα γενίκευσης του. Αυτή η διαδικασία μάς επιτρέπει να αξιολογήσουμε την απόδοση του μοντέλου σε νέα, πραγματικά δεδομένα που δεν έχει δει κατά τη διάρκεια της εκπαίδευσής του. Στην επόμενη ενότητα, θα εμβαθύνουμε σε μια περιεκτική επισκόπηση των τεχνικών στατιστικής ανάλυσης που χρησιμοποιούνται συχνά. Μεταξύ αυτών των τεχνικών είναι οι περιγραφικές και οπτικές μέθοδοι, οι οποίες περιλαμβάνουν απλές περιγραφικές στατιστικές όπως:

- Μέσοι όροι και μέτρο μεταβολής: Οι μέσοι όροι, όπως ο μέσος όρος και ο μέσος όρος, χρησιμοποιούνται για να κατανοήσουμε την τάση ή την τιμή ενός συνόλου δεδομένων. Το μέτρο μεταβολής, όπως η τυπική απόκλιση, μας δίνει μια ιδέα για το πόσο τα δεδομένα διαφέρουν ή διασπείρονται γύρω από τον μέσο όρο.
- Μετρήσεις και ποσοστά: Οι μετρήσεις χρησιμοποιούνται για να περιγράψουν την απόλυτη τιμή ενός φαινομένου, ενώ τα ποσοστά χρησιμοποιούνται για να δείξουν το μέγεθος ενός φαινομένου σε σχέση με το σύνολο.
- Διασταυρούμενες καρτέλες και απλές συσχετίσεις: Οι διασταυρούμενες καρτέλες χρησιμοποιούνται για να εξετάσουν τη σχέση μεταξύ δύο μεταβλητών, ενώ οι απλές συσχετίσεις εστιάζουν στη σχέση μεταξύ δύο μεταβλητών σε ένα σημείο χρόνου ή σε ένα συγκεκριμένο γεγονός. Αυτές οι δοκιμές μπορούν να δείξουν εάν υπάρχει κάποια στατιστικά σημαντική σχέση μεταξύ των μεταβλητών.

Η χρήση τεχνικών περιγραφής και οπτικοποίησης είναι εξαιρετικά πλεονεκτική για την κατανόηση της δομής δεδομένων. Αυτή η ανακάλυψη στην τεχνολογία βοηθά σημαντικά στην ερμηνεία τεράστιων ποσοτήτων δεδομένων και στην κατανόηση της υποκείμενης δομής. Μια σειρά εργαλείων, όπως ιστογράμματα, διαγράμματα κουτιών, διαγράμματα διασποράς και πολυδιάστατα γραφήματα επιφανειών, χρησιμοποιούνται σε αυτή την προσέγγιση. Η ανάλυση συστάδων αποτελεί μια μέθοδο οργάνωσης πληροφοριών σχετικά με μεταβλητές, με στόχο τον εντοπισμό σχετικά ομοιογενών συστάδων. Για παράδειγμα, αν αναλύουμε δεδομένα σχετικά με τα χαρακτηριστικά των πελατών ενός καταστήματος, οι συστάδες μπορεί να αναπαριστούν ομάδες πελατών με παρόμοια χαρακτηριστικά, όπως ηλικία, φύλο ή αγοραστικά συνήθεια. Αυτό μας επιτρέπει να κατανοήσουμε καλύτερα τις διαφορετικές ομάδες πελατών και να προσαρμόσουμε τις στρατηγικές μας ανάλογα.Από την άλλη πλευρά, η ανάλυση συσχέτισης παρέχει πληροφορίες για το πώς συσχετίζονται δύο μεταβλητές μεταξύ τους. Για παράδειγμα, μια ανάλυση συσχέτισης μεταξύ θερμοκρασίας και πωλήσεων μπορεί να δείξει εάν υπάρχει κάποια συσχέτιση μεταξύ της αύξησης της θερμοκρασίας και των αυξημένων πωλήσεων. Αυτό μας βοηθά να κατανοήσουμε τις σχέσεις μεταξύ διαφορετικών μεταβλητών και να καταλάβουμε τις πιθανές επιδράσεις μεταξύ τους.Τέλος, η ανάλυση διάκρισης χρησιμοποιείται για τον προσδιορισμό του πώς διάφοροι παράγοντες μπορούν να επηρεάσουν τη συμμετοχή σε συγκεκριμένες ομάδες. Για παράδειγμα, μπορεί να αναλύσουμε ποιοι παράγοντες επηρεάζουν την απόφαση ενός ατόμου να αγοράσει ένα προϊόν ή να χρησιμοποιήσει ένα

συγκεκριμένο υπηρεσία. Αυτό μας βοηθά να κατανοήσουμε ποιοι παράγοντες επηρεάζουν τις αποφάσεις των ανθρώπων και πώς μπορούμε να προβλέψουμε τη συμπεριφορά τους σε συγκεκριμένες καταστάσεις. Μπορεί να θεωρηθεί ως το αντίστροφο μιας μονόδρομης πολυμεταβλητής ανάλυσης διακύμανσης (MANOVA), με τα επίπεδα της ανεξάρτητης μεταβλητής στο MANOVA να γίνονται οι κατηγορίες της εξαρτημένης μεταβλητής στη διακριτική ανάλυση και οι εξαρτημένες μεταβλητές στο MANOVA να γίνονται προγνωστικοί σύνδεσμοι στη διακριτική ανάλυση. Η Παραγοντική Ανάλυση είναι μια σημαντική τεχνική που χρησιμοποιείται στην ανάλυση δεδομένων. [55]Βοηθά στην κατανόηση των υποκείμενων λόγων για τη συσχέτιση μεταξύ μιας ομάδας μεταβλητών. Η Παραγοντική Ανάλυση συχνά χρησιμοποιείται για τη μείωση του αριθμού των μεταβλητών και την ανίχνευση της δομής στις σχέσεις μεταξύ αυτών. Η διερευνητική Παραγοντική Ανάλυση αποσκοπεί στη διερεύνηση και την αναζήτηση της παραγοντικής δομής των μεταβλητών, ενώ η επιβεβαιωτική Παραγοντική Ανάλυση στοχεύει στην επιβεβαίωση της υποτιθέμενης παραγοντικής δομής, με βάση την προϋπόθεση ότι είναι γνωστή εκ των προτέρων. Η Ανάλυση Παλινδρόμησης, από την άλλη πλευρά, είναι μια στατιστική τεχνική που χρησιμοποιείται για την εκτίμηση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών, με τη χρήση της μίας ως εξαρτημένης και της άλλης ως ανεξάρτητης. Αν και η Ανάλυση Παλινδρόμησης δεν συνεπάγεται αιτιοκρατική σχέση, μπορεί να βοηθήσει στην πρόβλεψη της εξαρτημένης μεταβλητής με βάση την ανεξάρτητη. Η λογιστική παλινδρόμηση είναι μια ειδική περίπτωση της Ανάλυσης Παλινδρόμησης που χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι δυαδική ή ποιοτική. Χρησιμοποιεί μια μέθοδο μέγιστης πιθανότητας για να προσαρμόσει τα δεδομένα σε μια λογιστική καμπύλη, με σκοπό τη μέγιστη πιθανότητα πρόβλεψης της εξαρτημένης μεταβλητής.

# 2. Εισαγωγή στη θεωρία της Μηχανικής Μάθησης

Η μηχανική μάθηση είναι ένας τεράστιος και πολύπλοκος κλάδος που επικεντρώνεται στην ανάπτυξη αλγορίθμων που έχουν σχεδιαστεί για να κάνουν προβλέψεις με την ανάλυση και την επεξεργασία δεδομένων. Ο πρωταρχικός στόχος των εργασιών μηχανικής μάθησης είναι ο εντοπισμός και η εκμάθηση μιας συνάρτησης που αντιστοιχίζει αποτελεσματικά τον τομέα εισόδου (που αντιπροσωπεύεται από τα δεδομένα) στον τομέα εξόδου (που αντιπροσωπεύεται από πιθανές προβλέψεις).

$$f:X \rightarrow Y$$

Αυτή η διαδικασία περιλαμβάνει την επιλογή κατάλληλων συναρτήσεων, που δηλώνονται ως f, από διάφορες κατηγορίες συναρτήσεων με βάση τον συγκεκριμένο αλγόριθμο εκμάθησης που χρησιμοποιείται. Τελικά, η μηχανική μάθηση επιτρέπει την εξαγωγή πολύτιμων γνώσεων και προτύπων από δεδομένα, οδηγώντας σε βελτιωμένες δυνατότητες πρόβλεψης και ικανότητες λήψης αποφάσεων. [62]Ο Mitchell (1997) ορίζει την "μάθηση" ως εξής: '"Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία (experience) E σε σχέση με κάποια τάξη εργασιών(tasks) T και απόδοσης(performance) P, εάν η απόδοσή του σε εργασίες T, όπως μετράται από το P, βελτιώνεται με την εμπειρία E[2]. Το μέτρο απόδοσης P μας δίνει ποσοτικά τον βαθμό με τον οποίο ένας αλγόριθμος μηχανικής μάθησης βελτιώνεται καθώς εκτελείται. Η ακρίβεια είναι ένα συνηθέστερο μέτρο απόδοσης στις εργασίες ταξινόμησης, καθώς παρέχει μια απλή και ευανάγνωστη μετρική για το πόσο σωστά κατατάσσονται τα παραδείγματα στις διάφορες κατηγορίες.Η εμπειρία E, στην οποία υποβάλλονται οι αλγόριθμοι μηχανικής μάθησης, αντιπροσωπεύει τα σύνολα δεδομένων που χρησιμοποιούνται για την εκπαίδευση και τη δοκιμή αυτών των αλγορίθμων. Αυτά τα σύνολα δεδομένων περιέχουν παραδείγματα, καθένα από τα οποία ανήκει σε μια συγκεκριμένη κατηγορία ή κλάση. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου, ενώ τα δεδομένα δοκιμής χρησιμοποιούνται για να αξιολογηθεί η απόδοσή του στο να κατατάσσει σωστά τα παραδείγματα σε κατηγορίες που δεν έχει δει κατά τη διάρκεια της εκπαίδευσης..

## 2.1 Σχέση της Μηχανικής Μάθησης με την στατιστική

Τα πεδία της μηχανικής μάθησης και της στατιστικής είναι στενά συνδεδεμένα, με τους στατιστικολόγους να χρησιμοποιούν όλο και περισσότερο τεχνικές μηχανικής μάθησης, δημιουργώντας έναν νέο τομέα που είναι γνωστός ως στατιστική μάθηση. Έχουν κοινοποιηθεί διάφορες απόψεις σχετικά με τη σχέση μεταξύ μηχανικής μάθησης και στατιστικών. Μια αξιοσημείωτη άποψη προέρχεται από τον Larry Wasserman, έναν Καναδό στατιστικολόγο, ο οποίος πιστεύει ότι και τα δύο πεδία επιδιώκουν τελικά να απαντήσουν στο ίδιο ερώτημα: πώς μπορούμε να εξαγάγουμε γνώση από δεδομένα; Από την άλλη πλευρά, ο Michael I. Jordan, ένας Αμερικανός επιστήμονας, υποστηρίζει ότι η μηχανική μάθηση βασίζεται ουσιαστικά στις θεμελιώδεις αρχές και τα εργαλεία της στατιστικής, υποδηλώνοντας μια συγχώνευση των δύο πεδίων κάτω από την ομπρέλα της Επιστήμης Δεδομένων. Επεκτείνοντας αυτές τις σκέψεις, ο Leo Breiman, ένας Αμερικανός στατιστικολόγος, σημείωσε τη σύγκλιση των αλγοριθμικών μοντέλων στη στατιστική μοντελοποίηση με αλγόριθμους μηχανικής μάθησης όπως τα Random Forests. [54]

Για να επεξηγήσουμε καλύτερα αυτή τη διάκριση, εξετάζουμε το ακόλουθο σενάριο: ένας στατιστικολόγος και ένας μηχανικός μηχανικής μάθησης περιγράφουν το αποτέλεσμα του ίδιου μοντέλου. Ο στατιστικολόγος πιθανότατα θα δώσει έμφαση στις γνώσεις που αποκτήθηκαν από το μοντέλο και τις συνέπειες για την κατανόηση του υποκείμενου φαινομένου. Αντίθετα, ο Μηχανικός Μηχανικής Μάθησης πιθανότατα θα επικεντρωθεί στην απόδοση του μοντέλου και στην ικανότητά του να παράγει ακριβείς προβλέψεις. Μια κρίσιμη διάκριση μεταξύ των δύο έγκειται στον τρόπο με τον οποίο επεξεργάζονται δεδομένα. Οι στατιστικές απαιτούν την κατανόηση της συλλογής δεδομένων και την επιλογή των κατάλληλων παραμέτρων για ακριβείς προβλέψεις. Αντίθετα, η μηχανική μάθηση αγνοεί τυχόν προκαθορισμένες σχέσεις μεταξύ μεταβλητών και χρησιμοποιεί όλα τα διαθέσιμα δεδομένα για να καθορίσει τις παραμέτρους που θα οδηγήσουν σε επιτυχημένες προβλέψεις. Στην πραγματικότητα, η ακρίβεια των προβλέψεων μηχανικής μάθησης τείνει να αυξάνεται καθώς αυξάνεται ο όγκος των δεδομένων. Αυτό εξηγεί γιατί οι στατιστικές βασίζονται συχνά σε μικρότερα σύνολα δεδομένων, ενώ η μηχανική εκμάθηση βρίσκει χρησιμότητα σε σενάρια όπου είναι προσβάσιμες εκτενείς ομάδες δεδομένων. Συνοπτικά, ενώ η μηχανική μάθηση και οι στατιστικές μοιράζονται κοινό έδαφος και κοινό στόχο, τα θεμέλια, οι προσεγγίσεις επεξεργασίας δεδομένων και οι προοπτικές επίλυσης προβλημάτων τα ξεχωρίζουν ως ξεχωριστά πεδία μελέτης. Ωστόσο, η θεμελιώδης ασυμφωνία μεταξύ αυτών των πεδίων έγκειται στην προσέγγισή τους στην επίλυση προβλημάτων. Η μηχανική μάθηση στοχεύει στη βελτιστοποίηση και τη βελτίωση της αποτελεσματικότητας, προσπαθώντας να επιτύχει το καλύτερο δυνατό αποτέλεσμα. Από την άλλη πλευρά, οι στατιστικές επικεντρώνονται στην εξαγωγή ουσιαστικών συμπερασμάτων από τα δεδομένα, δίνοντας προτεραιότητα στην ερμηνεία των αποτελεσμάτων. Παρά τις σημαντικές ομοιότητες και τον αμοιβαίο στόχο τους, η μηχανική μάθηση και οι στατιστικές αποκλίνουν σημαντικά ως προς τα θεμέλια και την προέλευσή τους. Η στατιστική, ως κλάδος των μαθηματικών, προϋπήρχε της εμφάνισης των υπολογιστών και έχει μακρά ιστορία. Από την άλλη πλευρά, η μηχανική μάθηση αποτελεί ένα σχετικά νέο επιστημονικό πεδίο που αναδύθηκε από τη σφαίρα της τεχνητής νοημοσύνης. Η άνθησή του οφείλεται στις προόδους της επιστήμης των υπολογιστών, καθώς και στη διαθεσιμότητα όλο και περισσότερων δεδομένων και υπολογιστικών πόρων. Αυτό το πεδίο επικεντρώνεται στην ανάπτυξη αλγορίθμων και τεχνικών που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και εμπειρίες, χωρίς να χρειάζεται να προγραμματιστούν εξαρχής συγκεκριμένες ενέργειες. Αυτό το νέο πεδίο έχει επικεντρωθεί στην ανάπτυξη αλγορίθμων που μπορούν να αναγνωρίζουν πρότυπα στα δεδομένα, να κάνουν προβλέψεις και να λαμβάνουν αποφάσεις βασισμένες σε αυτά.[51]

## 2.2 Είδη Μηχανικής Μάθησης

Αντίθετα με την κλασική διαδικασία εκτίμησης ή πρόβλεψης (π.χ., Προσομοίωση) όπου το πρόγραμμα γνωρίζει ήδη το μοντέλο (γνώση) και το χρησιμοποιεί, η διαδικασία της Μηχανικής Μάθησης εφαρμόζεται

ένα βήμα πριν, όταν το σύστημα προσπαθεί να βρει το μοντέλο (γνώση) για να προχωρήσει μετά στην εκτίμηση και την πρόβλεψη.[47]

Στη συνεχώς εξελισσόμενη σφαίρα της μηχανικής μάθησης, καταλαβαίνουμε ότι νέες μεθοδολογίες αναδύονται συνεχώς, εμπίπτοντας σε διακριτές κατηγορίες γνωστές ως εποπτευόμενη μάθηση, χωρίς επίβλεψη και ενισχυτική μάθηση.

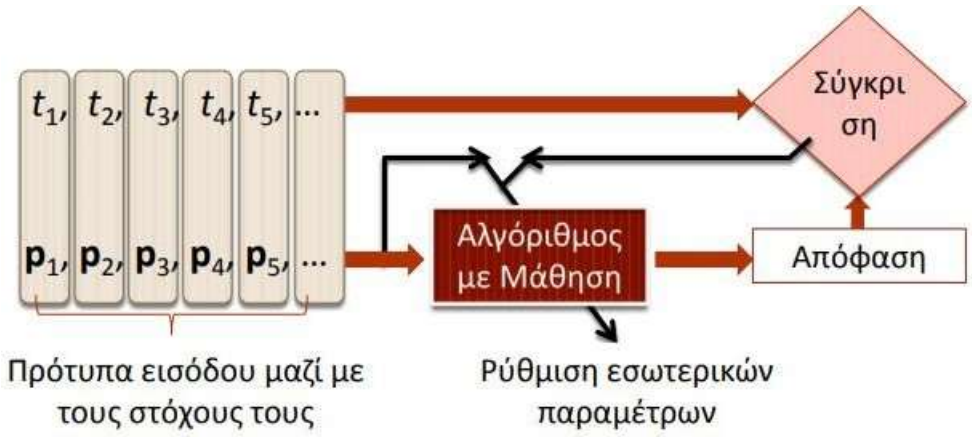
### 2.2.1 Επιβλεπόμενη μάθηση (supervised learning)

Στην εποπτευόμενη μάθηση, ο πρωταρχικός στόχος είναι να αναπτυχθεί ένας κανόνας ή μια συνάρτηση χρησιμοποιώντας ένα σύνολο γνωστών δεδομένων εισόδου και αντίστοιχων επιθυμητών δεδομένων εξόδου, με απώτερο στόχο τη γενίκευση αυτής της συνάρτησης για να γίνουν προβλέψεις σε περιπτώσεις όπου τα δεδομένα εξόδου είναι άγνωστα. Αυτή η συνάρτηση, γνωστή ως συνάρτηση στόχος, χρησιμεύει ως εργαλείο για την πρόβλεψη της τιμής της μεταβλητής εξόδου (εξαρτημένη μεταβλητή) με βάση τις μεταβλητές εισόδου (ανεξάρτητες μεταβλητές). Η έννοια της εποπτευόμενης μάθησης περιστρέφεται γύρω από την υπόθεση της επαγωγικής μάθησης, η οποία υποστηρίζει ότι τα μοτίβα που παρατηρούνται στα δεδομένα εκπαίδευσης μπορούν να προεκταθούν για να γίνουν ακριβείς προβλέψεις για αόρατα δεδομένα, σύμφωνα με την οποία:

*«Κάθε υπόθεση h που προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει».*[35]

Η επιβλεπόμενη μάθηση χρησιμοποιείται σε δύο κύριες περιπτώσεις προβλημάτων:

- Προβλήματα παλινδρόμησης (regression) - Αφορούν τη δημιουργία μοντέλων που σκοπό έχουν την πρόβλεψη αριθμητικών τιμών.
- Προβλήματα ταξινόμησης (classification) - Στοχεύουν στη δημιουργία μοντέλων πρόβλεψης διακριτών κατηγοριών ή τάξεων.




**Εικόνα 1 Επιβλεπόμενη Μάθηση**

**Πηγή: Κ. Διαμαντάρας, Μηχανική Μάθηση – Μάθημα 1, Βασικές Έννοιες, Τμήμα Πληροφορικής, ΤΕΙ Θεσσαλονίκης, 2011**

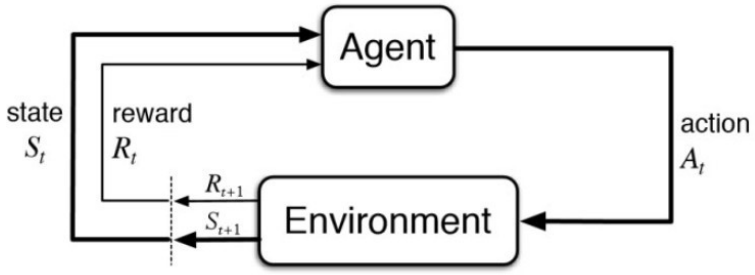
### 2.2.2 Μη επιβλεπόμενη μάθηση (unsupervised learning)

Στον τομέα της μάθησης χωρίς επίβλεψη, το πρωταρχικό καθήκον του αλγορίθμου είναι να αποκαλύψει την υποκείμενη δομή μέσα στα δεδομένα εισόδου. Αυτό το κάνει κατασκευάζοντας μοντέλα συσχέτισης χρησιμοποιώντας δεδομένα παρατήρησης, όλα χωρίς προηγούμενη γνώση των επιθυμητών εξόδων. Οι μέθοδοι μάθησης χωρίς επίβλεψη βρίσκουν εκτεταμένη εφαρμογή στην ανάλυση συσχετισμού, η οποία στοχεύει στην αποκάλυψη σχέσεων μέσα σε τεράστια σύνολα δεδομένων. [46]Επιπλέον, η ομαδοποίηση είναι ένας άλλος τομέας όπου χρησιμοποιείται η μάθηση χωρίς επίβλεψη, που περιλαμβάνει τη δημιουργία ομάδων ή συμπλεγμάτων από ένα δεδομένο σύνολο δεδομένων. Η ομαδοποίηση, γνωστή και ως clustering, αποτελεί σημαντικό μέρος της διαδικασίας εξόρυξης δεδομένων. Ο βασικός στόχος είναι να ομαδοποιήσει τα αντικείμενα με βάση τις ομοιότητές τους, έτσι ώστε τα αντικείμενα εντός του ίδιου ομαδικού συμπλέγματος να έχουν περισσότερες ομοιότητες μεταξύ τους από ό,τι με αντικείμενα σε άλλα συμπλέγματα. Η διαδικασία αυτή μπορεί να είναι χρήσιμη για την ανάλυση δεδομένων όπου δεν υπάρχουν ήδη γνωστές κατηγορίες ή ετικέτες. Αντίθετα, με την ομαδοποίηση, τα δεδομένα μπορούν να οργανωθούν σε φυσικές ή στατιστικά ομάδες, αναδεικνύοντας μοτίβα και δομές που ενδέχεται να μην είναι εμφανή από προηγούμενες αναλύσεις. Συνολικά, η ομαδοποίηση αποτελεί ένα σημαντικό εργαλείο για την εξαγωγή πληροφοριών από δεδομένα χωρίς προκαθορισμένες κατηγορίες, επιτρέποντας την ανακάλυψη νέων μοτίβων και συσχετίσεων..


**Εικόνα 2 Μη Επιβλεπόμενη Μάθηση**

**Πηγή: Κ. Διαμαντάρας, Μηχανική Μάθηση – Μάθημα 1, Βασικές Έννοιες, Τμήμα Πληροφορικής, ΤΕΙ Θεσσαλονίκης, 2011**

### 2.2.3 Ενισχυτική μάθηση (reinforcement learning)




**Εικόνα 3 Ενισχυτική Μάθηση**

**Πηγή: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html**

Η ενισχυτική μάθηση είναι μια διαδικασία όπου ένας αλγόριθμος διδάσκεται να λαμβάνει αποφάσεις μέσω δοκιμής και λάθους σε ένα περιβάλλον. Αυτή η μέθοδος διαφέρει από την εποπτευόμενη μάθηση καθώς δεν βασίζεται σε δεδομένα με ετικέτα για να καθοδηγήσει τις ενέργειες του αλγορίθμου. Αντίθετα, η ενισχυτική μάθηση χρησιμοποιεί ένα σύστημα ανταμοιβών και κυρώσεων για να ενθαρρύνει τον αλγόριθμο να κάνει επιλογές που οδηγούν στην υψηλότερη δυνατή ανταμοιβή. Ο απώτερος στόχος είναι ο αλγόριθμος να μάθει πώς να πλοηγείται στο περιβάλλον και να επιτυγχάνει έναν συγκεκριμένο στόχο χωρίς άμεση καθοδήγηση από εξωτερική πηγή. Αυτή η προσέγγιση είναι χρήσιμη για εργασίες όπως ο σχεδιασμός και η βελτιστοποίηση της κίνησης του ρομπότ σε εργοστασιακούς χώρους. [46]

## 2.3 Εργασίες Μηχανών Μάθησης

Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για μια ποικιλία εργασιών, με δύο κοινές εφαρμογές που είναι η ανάλυση παλινδρόμησης και η ταξινόμηση. Η ανάλυση παλινδρόμησης περιλαμβάνει την προσέγγιση της σχέσης μεταξύ των μεταβλητών προκειμένου να προβλεφθεί μια συγκεκριμένη τιμή με βάση τα δεδομένα εισόδου. Η ανάλυση παλινδρόμησης μπορεί να είναι χρήσιμη για την πρόβλεψη των μελλοντικών τιμών των μετοχών στη χρηματοπιστωτική αγορά. Από την άλλη πλευρά, η ταξινόμηση περιλαμβάνει τον προσδιορισμό της κλάσης στην οποία ανήκει μια συγκεκριμένη είσοδος. Αυτό επιτυγχάνεται με τη δημιουργία μιας συνάρτησης f: Rn → {1, ..., n}. Ένα παράδειγμα ενός ευρέως χρησιμοποιούμενου προβλήματος ταξινόμησης είναι η αναγνώριση αντικειμένων σε συστήματα τεχνητής νοημοσύνης, η οποία μπορεί να εφαρμοστεί για την ταξινόμηση αντικειμένων σε μια αποθήκη και την ανάθεσή τους στον σωστό προορισμό τους. Η τρέχουσα τεχνολογία αναγνώρισης αντικειμένων συχνά χρησιμοποιεί αλγόριθμους βαθιάς μάθησης για βελτιωμένη ακρίβεια.

## 2.4 Επιλογή Βέλτιστου Ελεγχόμενου Αλγορίθμου Εκμάθησης

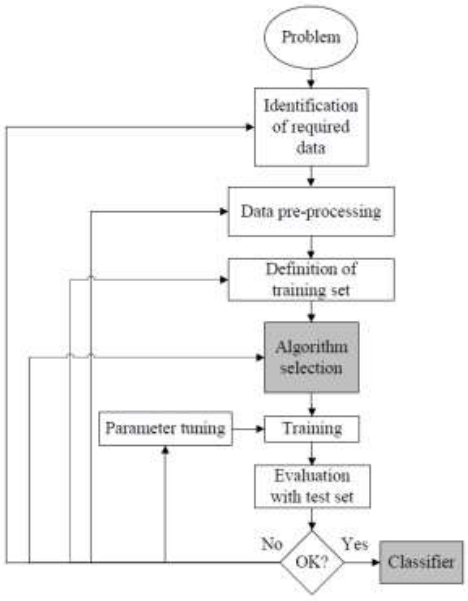
Προκειμένου να προσδιοριστεί η καταλληλότερη προσέγγιση μηχανικής εκμάθησης για μια συγκεκριμένη εργασία, είναι απαραίτητο να αξιολογηθούν τα στοιχεία που απαιτούνται για μια αποτελεσματική εποπτευόμενη διαδικασία μηχανικής μάθησης. Ο Κοτσιμάντης [6] περιγράφει μια μεθοδολογία για την ανάπτυξη ενός επιτυχημένου ταξινομητή που μπορεί να γενικευτεί αποτελεσματικά σε νέες περιπτώσεις δεδομένων. Αυτή η διαδικασία απεικονίζεται στην εικόνα 4.

Οι αρχικές φάσεις αυτού του αγωγού έχουν σημαντική σημασία καθώς παίζουν καθοριστικό ρόλο στη διαμόρφωση της αποτελεσματικότητας του ταξινομητή. Ο προσδιορισμός των βασικών δεδομένων περιλαμβάνει τον προσδιορισμό και την επιλογή των πιο συναφών χαρακτηριστικών. Με την εξάλειψη των περιττών ή περιττών χαρακτηριστικών, το μέγεθος των δεδομένων μπορεί να ελαχιστοποιηθεί. Μια υπερχείλιση περιττών δεδομένων μπορεί να εμποδίσει την ικανότητα ενός αλγορίθμου μάθησης να ανιχνεύει μοτίβα μέσα στα δεδομένα ή να οδηγήσει σε λανθασμένα αποτελέσματα. Το στάδιο προεπεξεργασίας δεδομένων είναι κρίσιμο στη διαδικασία της εξόρυξης δεδομένων, καθώς συμβάλλει στη

βελτίωση της ποιότητας και της εγκυρότητας των δεδομένων που χρησιμοποιούνται για την εκπαίδευση των μοντέλων. Οι κύριοι στόχοι του σταδίου προεπεξεργασίας περιλαμβάνουν:

- Καθαρισμός δεδομένων: Αφορά την αφαίρεση ή τη διόρθωση ανεπιθύμητων ή ανακριβών δεδομένων, όπως απουσιάζουσες τιμές, ανωμαλίες ή ακραίες τιμές (outliers).
- Κανονικοποίηση: Στόχος είναι η αναγκαία προσαρμογή των κλίμακων των δεδομένων ώστε να είναι συγκρίσιμα. Αυτό μπορεί να γίνει μέσω τεχνικών όπως η τυποποίηση ή η κανονικοποίηση Min-Max.
- Επιλογή χαρακτηριστικών: Αφορά την επιλογή των πιο σημαντικών χαρακτηριστικών για το μοντέλο, απορρίπτοντας εκείνα που δεν προσφέρουν πολλή πληροφορία ή προκαλούν θόρυβο.
- Αντιμετώπιση απουσιάζουσων τιμών: Συμπλήρωση ή αφαίρεση των απουσιάζουσων τιμών ανάλογα με την περίπτωση.
- Μείωση διαστάσεων: Μειώνει τον αριθμό των χαρακτηριστικών, βελτιώνοντας έτσι την απόδοση και την αποδοτικότητα του μοντέλου.

Το αποτέλεσμα του σταδίου προεπεξεργασίας είναι ένα επεξεργασμένο σύνολο δεδομένων που μπορεί να χρησιμοποιηθεί από τους αλγορίθμους μηχανικής μάθησης για εκπαίδευση και αξιολόγηση των μοντέλων.




**Εικόνα 4 Supervised machine learning pipeline used to create a successful classifier**
**Πηγή: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html**

## 2.5 Περιγραφή και Επεξεργασία των Δεδομένων

Ένας τρόπος για να περιγραφούν τα σύνολα δεδομένων που χρησιμοποιούνται στην εκπαίδευση και τη δοκιμή αλγορίθμων μηχανικής μάθησης είναι μέσω της χρήσης ενός πίνακα σχεδίασης [3]. Αυτός ο πίνακας σχεδίασης είναι ουσιαστικά ένας πίνακας που περιέχει όλα τα σημεία δεδομένων, όπου κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο χαρακτηριστικό. Για να το δείξουμε, ας εξετάσουμε ένα σενάριο όπου έχουμε μια συλλογή από 10 φωτογραφίες που καταγράφουν ένα αντικείμενο σε ανάλυση 1600x1200 και κάθε φωτογραφία συνδέεται με τρία διαφορετικά χαρακτηριστικά. Σε αυτήν την περίπτωση, το σύνολο

δεδομένων μπορεί να αναπαρασταθεί από έναν πίνακα σχεδίασης που συμβολίζεται ως X ∈ R10 × 1600 × 1200 × 3. Αξίζει να σημειωθεί ότι τα σύνολα δεδομένων που χρησιμοποιούνται στην εκπαίδευση και τη δοκιμή αλγορίθμων μηχανικής μάθησης μπορεί να κυμαίνονται από σχετικά απλά έως μεγάλα και πολύπλοκος. Παράδειγμα ενός σχετικά απλού συνόλου δεδομένων είναι το σύνολο δεδομένων Iris [7].

Το συγκεκριμένο dataset περιλαμβάνει 150 δείγματα, καθένα από τα οποία αποτελείται από τέσσερα στοιχεία δεδομένων και μπορεί να αναπαρασταθεί με έναν πίνακα X ∈ R150 × 4. Ωστόσο, όταν ασχολούμαστε με φωτογραφίες, τα δεδομένα μπορούν να είναι αρκετά μεγαλύτερα λόγω της υψηλής ανάλυσης της εικόνας. Για παράδειγμα, μια εικόνα με ανάλυση 1900 × 1080 pixels θα είχε ως αποτέλεσμα 6.156.000 σημεία δεδομένων. Κάθε pixel στην εικόνα αντιστοιχεί σε ένα σημείο δεδομένων με τιμές x, y και z. Η επεξεργασία τόσο μεγάλου όγκου δεδομένων ανά εικόνα απαιτεί σημαντική υπολογιστική ισχύ, ιδιαίτερα κατά την εκπαίδευση και δοκιμή αλγορίθμων μηχανικής μάθησης. Με τη συνεχή αύξηση του όγκου δεδομένων που παράγονται καθημερινά, οι αλγόριθμοι μηχανικής μάθησης μπορούν τώρα να εκπαιδεύονται σε πιο περίπλοκα datasets.

## 2.6 Λόγοι Κλιμάκωσης της Μηχανικής Μάθησης

Ο όγκος των δεδομένων και των συμβάντων που συλλέγονται και αποθηκεύονται σε καθημερινή βάση αυξάνεται σταθερά, ιδιαίτερα στους τομείς του Διαδικτύου και των οικονομικών. Οι εταιρείες τηλεπικοινωνιών συγκεντρώνουν τεράστιες ποσότητες δεδομένων από τους πελάτες τους, μετρώντας σε Petabytes ανά ώρα, ενώ δισεκατομμύρια οικονομικές συναλλαγές πραγματοποιούνται σε όλο τον κόσμο κάθε μέρα. Με τον πολλαπλασιασμό των αισθητήρων σε αυτούς τους τομείς, η δεξαμενή των πιθανών δεδομένων εκπαίδευσης επεκτείνεται εκθετικά. Αυτή η πρόκληση επιδεινώνεται από την πολυπλοκότητα κάθε σημείου δεδομένων, το οποίο συχνά περιέχει πολλές μεταβλητές. Ωστόσο, όταν πρόκειται για την επεξεργασία αυτού του τεράστιου όγκου δεδομένων για εργασίες μηχανικής εκμάθησης, μια ενιαία μονάδα επεξεργασίας μπορεί να μην είναι αρκετή. Σε τέτοιες περιπτώσεις, τεχνικές παραλληλοποίησης μπορούν να χρησιμοποιηθούν για τον διαχωρισμό του φόρτου εργασίας μεταξύ πολλαπλών μονάδων επεξεργασίας, επιταχύνοντας έτσι σημαντικά τη συνολική διαδικασία. Στον σημερινό κόσμο, δημιουργείται ένας τεράστιος όγκος δεδομένων κάθε μέρα, που αποτελείται από δισεκατομμύρια μεμονωμένα σημεία δεδομένων. Κάθε ένα από αυτά τα σημεία δεδομένων συνδέεται με χιλιάδες διαφορετικά χαρακτηριστικά, με αποτέλεσμα έναν αστρονομικό αριθμό ζευγών σημείου δεδομένων-χαρακτηριστικών, που φτάνει την κλίμακα των 10^12 ανά ημέρα. Η αποθήκευση τέτοιων τεράστιων συνόλων δεδομένων απαιτεί εκατοντάδες Petabyte αποθήκευσης.

Όταν η κλίμακα μιας εργασίας στον τομέα της μηχανικής μάθησης είναι υπερβολικά μεγάλη για έναν μόνο επεξεργαστή μηχανής, μπορούν να χρησιμοποιηθούν τεχνικές παραλληλοποίησης για την επιτάχυνση της διαδικασίας. Σύμφωνα με τον Beckerman [1], υπάρχουν τέσσερις κύριες παράμετροι που χαρακτηρίζουν μια εργασία μηχανικής μάθησης με μεγάλη κλίμακα[67]:

- Η εργασία βασίζεται σε ένα μεγάλο σύνολο δεδομένων.
- Τα δεδομένα εισόδου έχουν υψηλή διαστασιολόγηση.
- Η πολυπλοκότητα του μοντέλου και του αλγορίθμου είναι υψηλή.
- Υπάρχουν περιορισμοί στο χρόνο για την εκτέλεση των αναλύσεων.

Για παράδειγμα, η συλλογή δεδομένων από εταιρείες όπως η Vodafone μπορεί να περιλαμβάνει τεράστιο αριθμό δεδομένων. Επιπλέον, η χρήση πολύπλοκων αλγορίθμων που αναλύουν δεδομένα βίντεο ή εικόνας απαιτεί την αντιμετώπιση δεδομένων με υψηλή διαστασιολόγηση, με τον αριθμό των διαστάσεων εισόδου να μπορεί να φτάσει το μέγεθος του 10^6. Επιπλέον, η χρήση πιο περίπλοκων μη γραμμικών μοντέλων ή αλγορίθμων μπορεί να απαιτεί σημαντικούς υπολογιστικούς πόρους για την επίτευξη υψηλών επιπέδων

ακρίβειας. Τέλος, σε εφαρμογές που χρησιμοποιούν αισθητήρες και απαιτούν προβλέψεις σε πραγματικό χρόνο, υπάρχει η ανάγκη για γρήγορη επεξεργασία και ανάλυση δεδομένων, όπως στην περίπτωση των αυτόνομων οχημάτων..

Σε ορισμένες περιπτώσεις, ο χρόνος που απαιτείται για την εξαγωγή συμπερασμάτων περιορίζεται από συγκεκριμένες συνθήκες προκειμένου να ολοκληρωθεί αποτελεσματικά μια εργασία, όπως το περπάτημα ή η οδήγηση. Ο Bekkerman υποστήριξε ότι σε αυτά τα τέσσερα σενάρια, είναι ζωτικής σημασίας να χρησιμοποιηθούν εξαιρετικά παραλληλισμένες αρχιτεκτονικές όπως οι GPU ή να υιοθετηθούν αποτελεσματικά υπολογιστικά πλαίσια όπως το DryadLINQ [1]. Πολυάριθμα ερευνητικά άρθρα έχουν εμβαθύνει στη σύγκριση διαφορετικών αλγορίθμων μηχανικής μάθησης, ωστόσο μια οριστική απάντηση στο ερώτημα ποιος αλγόριθμος κυριαρχεί για μια δεδομένη εφαρμογή παραμένει αδιευκρίνιστη. Ο Mooney υποστηρίζει ότι δεν υπάρχει ένας αλγόριθμος που να ταιριάζει σε όλους που να έχει καλύτερη απόδοση από όλους τους άλλους παγκοσμίως. Βασιζόμενος σε αυτήν την έννοια, ο Wolpert αποδεικνύει μέσω της θεωρητικής ανάλυσης ότι κανένας ταξινομητής δεν μπορεί να έχει σταθερά καλύτερη απόδοση από όλους τους άλλους. Το "No Free Lunch Theorem of Optimization" όπως διατυπώνεται από τους Wolpert και Macready τονίζει περαιτέρω ότι ενώ δεν υπάρχει γενική τεχνική βελτιστοποίησης, ένας προσαρμοσμένος αλγόριθμος μπορεί να υπερέχει όταν έχει σχεδιαστεί ειδικά για μια συγκεκριμένη δομή προβλήματος. Παρά την απουσία σαφούς νικητή, οι συγκριτικές μελέτες προσφέρουν πολύτιμες γνώσεις σχετικά με την απόδοση του αλγορίθμου και τη δυνατότητα εφαρμογής για συγκεκριμένες εργασίες. Επιπλέον, η διεξαγωγή μελετών κλιμάκωσης μπορεί να ρίξει φως στο πώς αυξάνεται ο υπολογιστικός χρόνος καθώς μεγαλώνει η κλίμακα υλοποίησης.

## 2.7 Εφαρμογή Μηχανικής Μάθησης Μεγάλης Κλίμακας

Τα τελευταία δέκα χρόνια, υπήρξε μια αξιοσημείωτη και ουσιαστική επέκταση της μηχανικής εκμάθησης μεγάλης κλίμακας σε ένα ευρύ φάσμα εφαρμογών, κυρίως λόγω δύο βασικών παραγόντων. Πρώτον, υπήρξε μια αύξηση στη διαθεσιμότητα εκτεταμένων συνόλων δεδομένων σε πολλούς σύγχρονους τομείς. Αυτή η εισροή δεδομένων έχει προσφέρει άφθονες ευκαιρίες στους αλγόριθμους μηχανικής μάθησης να αξιοποιήσουν και να εξάγουν πολύτιμες πληροφορίες. Επιπλέον, η ανάπτυξη και η συνεχής εξέλιξη των πλαισίων προγραμματισμού και των αρχιτεκτονικών υλικού έχουν συμβάλει σημαντικά στον πολλαπλασιασμό της μηχανικής μάθησης. Αυτές οι εξελίξεις έχουν ενδυναμώσει πολλούς αλγόριθμους μάθησης παρέχοντάς τους αποτελεσματικά εργαλεία και πλατφόρμες για την επεξεργασία και ανάλυση τεράστιων ποσοτήτων δεδομένων. Αυτή η σύγκλιση μεγαλύτερων συνόλων δεδομένων και βελτιωμένης τεχνολογικής υποδομής έχει παίξει καθοριστικό ρόλο στην ταχεία ανάπτυξη και υιοθέτηση της μεγάλης κλίμακας μηχανικής μάθησης τα τελευταία χρόνια [42].

Υπάρχουν δύο συγκεκριμένοι τομείς όπου αυτή η ανάγκη για αποτελεσματικό χειρισμό δεδομένων είναι ιδιαίτερα εμφανής - η οπτική ανίχνευση αντικειμένων για αυτόνομα συστήματα και η αναγνώριση ομιλίας. Σε αυτές τις περιπτώσεις, η ικανότητα ακριβούς αναγνώρισης αντικειμένων ή κατανόησης της προφορικής γλώσσας είναι υψίστης σημασίας. Το Machine Learning Algorithms In-Depth στοχεύει στην αντιμετώπιση αυτής της ανάγκης, προσφέροντας μια ολοκληρωμένη κατανόηση των διαφόρων αλγορίθμων, που κυμαίνονται από τους πιο αποτελεσματικούς έως την επιτυχημένη εφαρμογή τους σε ένα ευρύ φάσμα εφαρμογών. Τα τελευταία χρόνια, υπάρχει μια αυξανόμενη ζήτηση για αλγόριθμους μηχανικής μάθησης που είναι ικανοί να χειρίζονται τεράστιες ποσότητες δεδομένων που αποθηκεύονται σε κατανεμημένες πλατφόρμες. Αυτή η αυξημένη ανάγκη μπορεί να αποδοθεί στις εξελίξεις στα πλαίσια προγραμματισμού και στα σχέδια υλικού. Επιπλέον, η εμφάνιση αισθητήρων που είναι ικανοί να λαμβάνουν γρήγορες αποφάσεις βάσει πολύπλοκων χαρακτηριστικών έχει τονίσει περαιτέρω τη σημασία των αρχιτεκτονικών υλικού που επιτρέπουν τον αποτελεσματικό παράλληλο υπολογισμό σε εφαρμογές μηχανικής μάθησης.Για να εξασφαλιστεί μια ολοκληρωμένη και αποτελεσματική ερευνητική μελέτη, πραγματοποιήθηκε η ακόλουθη βιβλιογραφικήανασκόπηση:

Για τον προσδιορισμό των πιο επιτυχημένων εποπτευόμενων αλγορίθμων μηχανικής μάθησης, πραγματοποιείται εκτενής ανάλυση δημοσιευμένων άρθρων που συγκρίνει την απόδοσή τους σε διάφορα σύνολα δεδομένων. Για να αξιολογηθεί η αποτελεσματικότητά τους, η απόδοση μετράται και αξιολογείται βάσει τριών βασικών κριτηρίων:

(α) Ταχύτητα ταξινόμησης: Αναφέρεται στον χρόνο που απαιτείται για την ταξινόμηση ενός ερωτήματος.

(β) Ακρίβεια: Αποτελεί τον αριθμό των σωστών προβλέψεων διαιρεμένος με τον συνολικό αριθμό των προβλέψεων.

(γ) Ανοχή στο θόρυβο: Αναφέρεται στον βαθμό με τον οποίο ένας αλγόριθμος μηχανικής μάθησης μπορεί να αντιμετωπίσει τον θόρυβο, όπως ασήμαντα ή περιττά χαρακτηριστικά.

# 3. Ανασκόπηση των Αλγορίθμων

Σε αυτό το κεφάλαιο παρουσιάζονται τα ευρήματα από τη βιβλιογραφική έρευνα. Αναλύονται και εξετάζονται οι πιο δημοφιλείς και ισχυροί αλγόριθμοι μηχανικής μάθησης στον τομέα της εξόρυξης δεδομένων, συγκεκριμένα:

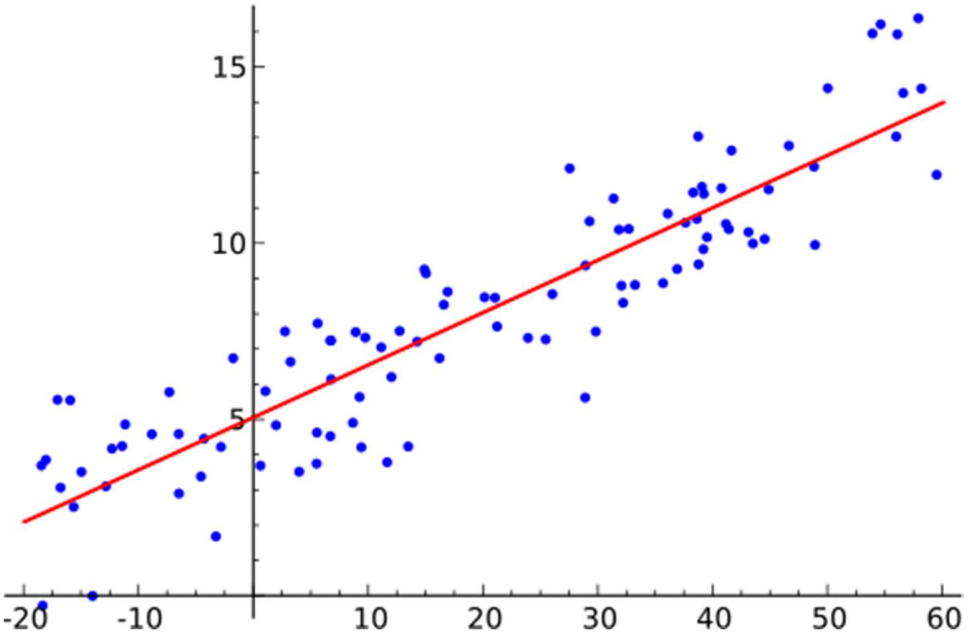
- Γραμμική Παλινδρόμηση (Linear Regression)
- Πολυμεταβλητή Ανάλυση Δεδομένων (MultiVariate Data Analysis, MVDA)
- Τυχαία Δάση (Random Forests)
- Ταξινομητές Naive-Bayes (Naive-Bayes classifiers)
- Δέντρα Αποφάσεων (Decision Trees)
- Νευρωνικά Δίκτυα (Neural Networks)
- Βοηθητικά Διανύσματα (Support Vector Machines)
- k-Nearest-Neighbor

Τα νευρωνικά δίκτυα είναι ιδιαίτερα ενδιαφέροντα λόγω της ευρείας χρήσης τους στην επεξεργασία πολυδιάστατων δεδομένων, ειδικά στον τομέα της ταξινόμησης αντικειμένων 3D. Στην εργασία γίνεται λεπτομερής ανάλυση των βασικών αρχών κάθε αλγορίθμου, εξετάζονται προσεκτικά τα δυνατά και τα αδύνατα σημεία τους, και πραγματοποιείται ολοκληρωμένη αξιολόγηση των δυνατοτήτων τους σε σχέση με την κλιμάκωση..

## 3.1 Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι ένας τρόπος για να καταλάβουμε πόσο διαφορετικά πράγματα σχετίζονται μεταξύ τους. Εξετάζει ένα πράγμα που μπορεί να αλλάξει (που ονομάζεται εξαρτημένη μεταβλητή) και πώς επηρεάζεται από άλλα πράγματα που δεν αλλάζουν (που ονομάζονται ανεξάρτητες μεταβλητές). [3]Εάν υπάρχει μόνο μία ανεξάρτητη μεταβλητή, ονομάζεται απλή γραμμική παλινδρόμηση, και εάν υπάρχουν περισσότερες από μία, ονομάζεται πολλαπλή γραμμική παλινδρόμηση. Ο στόχος είναι να κατανοήσουμε πώς οι ανεξάρτητες μεταβλητές μπορούν να αλλάξουν την τιμή της εξαρτημένης μεταβλητής.

Η διαδικασία της γραμμικής παλινδρόμησης αποτελείται από δύο διακριτές φάσεις. Στην αρχική φάση, γνωστή ως Φάση Α, οι ερευνητές διερευνούν εάν υπάρχει συσχέτιση μεταξύ των ανεξάρτητων και των εξαρτημένων μεταβλητών. Εάν ανακαλυφθεί μια σχέση, ξεκινά η Φάση Β, όπου λαμβάνει χώρα η ανάλυση παλινδρόμησης. Αυτό περιλαμβάνει τη δημιουργία ενός γραμμικού μοντέλου που αναπαριστά με ακρίβεια τα δεδομένα χρησιμοποιώντας γραμμικές συναρτήσεις. Σε αυτό το μοντέλο, ο όρος "ei" υποδηλώνει το τυχαίο σφάλμα ή θόρυβο, το οποίο έχει μέσο όρο μηδέν και διακύμανση σ2, είναι η ακόλουθη:




**Εικόνα 5 Παράδειγμα απλής γραμμικής παλινδρόμησης. Διάγραμμα διασποράς τιμών {x, y} με ανεξάρτητη μεταβλητή την x. Η κόκκινη ευθεία είναι η βέλτιστη εξίσωση y=α+βx που μοντελοποιεί τα σημεία**

**Πηγή: https://el.wikipedia.org/**

Με το μοντέλο που είναι άμεσα διαθέσιμο σε εμάς, μπορούμε να προβλέψουμε με ακρίβεια την τιμή της εξαρτημένης μεταβλητής y για κάθε δεδομένο x, καθώς και να εκτιμήσουμε τον βαθμό συσχέτισης μεταξύ y και x. Αυτό μας επιτρέπει να αναλύσουμε τη σχέση μεταξύ των δύο μεταβλητών και να εντοπίσουμε ποια υποσύνολα του x μπορεί να είναι περιττά στην πρόβλεψη του y. Ο απώτερος στόχος της γραμμικής παλινδρόμησης είναι να προσδιοριστεί η βέλτιστη συνάρτηση που αντιπροσωπεύει με ακρίβεια τη σύνδεση μεταξύ των μεταβλητών {xi, yi}, υπογραμμίζοντας τη σημασία της ακριβούς εκτίμησης των παραμέτρων α και β.Η μέθοδος των ελαχίστων τετραγώνων, η οποία έχει υιοθετηθεί ευρέως για την εκτίμηση των παραμέτρων α και β και για τον προσδιορισμό της εξίσωσης της καλύτερης ευθείας γραμμής για τα δεδομένα,

$$\hat{\beta} = \frac{v\sum_{i=1}^{v} x_i y_i - \left(\sum_{i=1}^{v} x_i\right)\left(\sum_{i=1}^{v} y_i\right)}{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2} \qquad \text{όπου } \bar{y} = \frac{1}{v}\sum_{i=1}^{v} y_i, \quad \bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i.$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

**Εξίσωση 1**

εισήχθη αρχικά από τον διάσημο Γάλλο μαθηματικό Legendre το 1805. Αυτή η μέθοδος επικεντρώνεται στην ελαχιστοποιώντας το άθροισμα των τετραγωνικών κατακόρυφων αποστάσεων μεταξύ των σημείων (xi, yi) και της ευθείας yi=α+βxi. Οι τιμές των α και β που αποδίδουν το ελάχιστο άθροισμα αναφέρονται ως εκτιμητές ελαχίστων τετραγώνων (least square estimators) καιυπολογίζονται από τις παρακάτω σχέσεις: [42]

Ενώ η ευθεία $\hat{y} = \hat{\alpha} + \hat{\beta}x$ καλείται ευθεία ελαχίστων τετραγώνων.

Στην μηχανική μάθηση, εκτός από τη μέθοδο ελαχίστων τετραγώνων, συχνά χρησιμοποιείται ο αλγόριθμος απότομης καθόδου (Gradient Descent) για την εκτίμηση των παραμέτρων. Ο αλγόριθμος αυτός επιδιώκει την ελαχιστοποίηση μιας συνάρτησης κόστους, προσαρμόζοντας επανειλημμένα τις παραμέτρους μέχρι να βρει το τοπικό ή και το γενικό ελάχιστο. Η ουσία του αλγορίθμου είναι να ακολουθεί την κατεύθυνση της κλίσης της συνάρτησης κόστους και να πραγματοποιεί μικρές προσαρμογές στις παραμέτρους σε κάθε βήμα, με σκοπό τη σύγκλιση προς το ελάχιστο.

Για να ελαχιστοποιήσουμε τη συνάρτηση κόστους των ελαχίστων τετραγώνων, χρησιμοποιούμε μια συνάρτηση γραμμικής παλινδρόμησης γνωστή ως συνάρτηση υπόθεσης. Στόχος μας είναι να βρούμε τις τιμές των θ0 και θ1 που θα έχουν ως αποτέλεσμα την ελάχιστη τιμή της εξίσωσης J(θ0, θ1), όπου το m αντιπροσωπεύει τον αριθμό των δειγμάτων {x, y}. Ο αλγόριθμος απότομης κατάβασης ξεκινά με την εκχώρηση αρχικών τιμών στα θ0 και θ1, και στη συνέχεια ενημερώνει επαναληπτικά αυτές τις τιμές μέχρι να συγκλίνουν στη βέλτιστη λύση.

Αυτή η αρχή ισχύει όταν η εξαρτημένη μεταβλητή μπορεί να εκφραστεί ως γραμμική συνάρτηση μιας ή περισσότερων ανεξάρτητων μεταβλητών. Ωστόσο, σε πολλές περιπτώσεις, η εξαρτημένη μεταβλητή είναι ένας συνδυασμός πολλαπλών ανεξάρτητων μεταβλητών, όπου η σχέση που προκύπτει είναι:

y_i: η τιμή της εξαρτημένης μεταβλητής

x_1, x_2, x_3, ....x_m: οι τιμές των ανεξάρτητων μεταβλητών

b_1, b_2, b_3, ..., b_m: οι συντελεστές παλινδρόμησης που υποδηλώνουν την επίδραση των μεταβλητών x_i

ε: το σφάλμα

Η πολλαπλή γραμμική παλινδρόμηση είναι μια προσέγγιση που χρησιμοποιείται όταν η εξαρτημένη μεταβλητή επηρεάζεται από πολλές ανεξάρτητες μεταβλητές. Σε αντίθεση με την απλή γραμμική παλινδρόμηση, όπου έχουμε μία εξαρτημένη μεταβλητή που εξαρτάται από μία ανεξάρτητη μεταβλητή, στην πολλαπλή γραμμική παλινδρόμηση έχουμε πολλές ανεξάρτητες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή.Η τιμή της εξαρτημένης μεταβλητής y σε μια πολλαπλή γραμμική παλινδρόμηση καθορίζεται από τις συστηματικές επιδράσεις των ανεξάρτητων μεταβλητών x1,x2,x3,…,xmx1,x2,x3,…,xm και έναν παράγοντα

τυχαίου σφάλματος ee, ο οποίος αντιπροσωπεύει όλους τους άλλους παράγοντες που επηρεάζουν το yy, εκτός από τις ανεξάρτητες μεταβλητές x1,x2,x3,…,xmx1,x2,x3,…,xm.Στην πολλαπλή γραμμική παλινδρόμηση, είναι σημαντικό οι ανεξάρτητες μεταβλητές να είναι ανεξάρτητες μεταξύ τους. Εάν υπάρχει συσχέτιση μεταξύ των μεταβλητών, θα πρέπει να επιλεγεί μόνο μία για την αποφυγή προβλημάτων πολυσυγκραμμικότητας. Επιπλέον, ο πίνακας που περιέχει τις ανεξάρτητες μεταβλητές XX θα πρέπει να έχει λιγότερες στήλες (ανεξάρτητες μεταβλητές) και περισσότερες σειρές (δείγματα), δηλαδή πολλά αντικείμενα και λίγες μεταβλητές.Ένα βασικό βήμα στην ανάπτυξη μοντέλων πολλαπλής γραμμικής παλινδρόμησης είναι η λήψη ορισμένων υποθέσεων. Δύο από τις πιο κοινές υποθέσεις είναι η γραμμικότητα και η ανεξαρτησία των σφαλμάτων. Η υπόθεση της γραμμικότητας υποδηλώνει ότι η μέση τιμή της εξαρτημένης μεταβλητής εξαρτάται γραμμικά από τις ανεξάρτητες μεταβλητές. Η υπόθεση της ανεξαρτησίας των σφαλμάτων υποθέτει ότι τα σφάλματα είναι ασυσχέτιστα μεταξύ τους και ακολουθούν κανονική κατανομή. Επίσης, υποθέτουμε ότι οι αναμενόμενες τιμές των σφαλμάτων είναι μηδέν και ότι τα σφάλματα έχουν σταθερή διακύμανση σε όλες τις τιμές των προβλεπόμενων μεταβλητών.Η γραμμική παλινδρόμηση ήταν η αρχική μορφή ανάλυσης παλινδρόμησης που προσέλκυσε μεγάλο ενδιαφέρον και χρησιμοποιήθηκε ευρέως σε διάφορα πεδία, λόγω της απλότητας των γραμμικών μοντέλων και της ευκολίας στην ερμηνεία των στατιστικών χαρακτηριστικών των εκτιμήσεων που προέκυψαν.

## 3.2 Πολυμεταβλητή Ανάλυση Δεδομένων (MVDA)

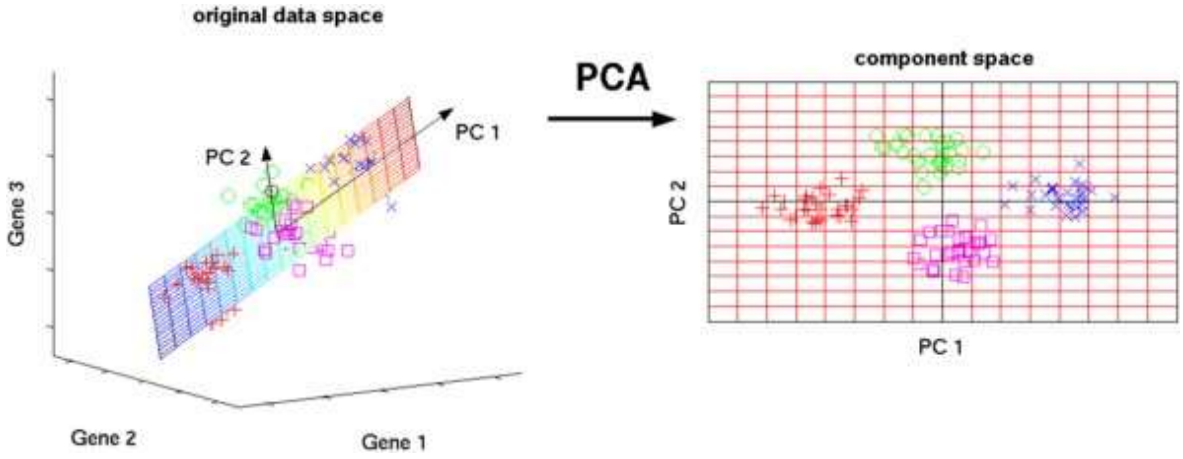
Η ανάλυση πολυμεταβλητών δεδομένων είναι μια εξελιγμένη τεχνική που επιτρέπει την εξέταση πολλαπλών μεταβλητών απόκρισης και τη διαχείριση πολλών συσχετισμένων ανεξάρτητων μεταβλητών ταυτόχρονα. Αυτή η μέθοδος περιλαμβάνει τη μείωση των διαστάσεων ενός πολυδιάστατου χώρου χωρίς να τροποποιούνται μεμονωμένες παραμέτρους ξεχωριστά. Η ανάλυση κύριας συνιστώσας (PCA) και η ανάλυση μερικών ελαχίστων τετραγώνων (PLS) είναι δύο κοινές μέθοδοι που χρησιμοποιούνται στην πολυμεταβλητή ανάλυση δεδομένων. Αυτές οι προσεγγίσεις είναι απαραίτητα εργαλεία για την κατανόηση πολύπλοκων συνόλων δεδομένων και τον εντοπισμό προτύπων και σχέσεων μέσα στα δεδομένα.

### 3.2.1 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis)

Η ανάλυση κύριας συνιστώσας είναι μια στατιστική τεχνική που περιλαμβάνει τη χρήση ενός μαθηματικού μετασχηματισμού για τη μετατροπή ενός συνόλου δεδομένων που περιέχει πιθανώς συσχετισμένες μεταβλητές σε ένα νέο σύνολο μεταβλητών που ονομάζονται κύρια συστατικά, τα οποία δεν σχετίζονται γραμμικά μεταξύ τους. Αυτή η μέθοδος χρησιμοποιείται συνήθως στην πολυμεταβλητή ανάλυση για τον εντοπισμό ενός μικρού αριθμού βασικών μεταβλητών από ένα μεγαλύτερο σύνολο. Για παράδειγμα, εάν υπάρχουν n παρατηρήσεις και p μεταβλητές, ο αριθμός των κύριων συνιστωσών θα περιοριστεί στο ελάχιστο n-1 ή p. Ο στόχος αυτού του μετασχηματισμού είναι να μεγιστοποιήσει τη διακύμανση του πρώτου κύριου στοιχείου, καταγράφοντας τη μεγαλύτερη μεταβλητότητα στα δεδομένα, με κάθε επόμενο στοιχείο να συλλαμβάνει την επόμενη υψηλότερη διακύμανση. Ένα από τα κύρια πλεονεκτήματα της ανάλυσης κύριων συνιστωσών είναι η ικανότητά της να εξηγεί ένα σημαντικό μέρος της συνολικής μεταβλητότητας μεταξύ των αρχικών μεταβλητών χρησιμοποιώντας μόνο μερικές νέες μεταβλητές.

Η ανάλυση του κύριου συστατικού βασίζεται σε δύο βασικούς παράγοντες. Πρώτον, δημιουργεί ένα σύνολο μη συσχετισμένων μεταβλητών, γνωστών ως κύριες συνιστώσες, από τις αρχικές μεταβλητές. Αυτό επιτρέπει τη μέτρηση διαφορετικών διαστάσεων των στοιχείων. Δεύτερον, η μεταβλητότητα μεταξύ αυτών των κύριων συνιστωσών είναι διατεταγμένη με φθίνουσα σειρά, με την πρώτη συνιστώσα να εξηγεί τη μεγαλύτερη μεταβλητότητα, ακολουθούμενη από τη δεύτερη συνιστώσα και ούτω καθεξής. Η τεχνική βασίζεται στον πίνακα συσχέτισης των μεταβλητών, με υψηλούς θετικούς ή αρνητικούς συντελεστές συσχέτισης να είναι απαραίτητοι για την επιτυχή ανάλυση. Είναι σημαντικό να επιτευχθεί μια ισορροπία, καθώς οι μεταβλητές με εξαιρετικά υψηλές τιμές συσχέτισης θεωρούνται περιττές και δεν πρέπει να περιλαμβάνονται στην ανάλυση. [46]

Το πρώτο βήμα για την ανάλυση κυρίων συνιστωσών είναι ο μετασχηματισμός των (X ,Y) μεταβλητών. Στη συνέχεια, το πρώτο κύριο συστατικό Z1 δημιουργείται συνδυάζοντας τις μεταβλητές p χρησιμοποιώντας βάρη α11X1+α12X2+...+α1pXp, ακολουθούμενη από τη δημιουργία πρόσθετων στοιχείων όπως Z2=α21X1+α22X2+...+α2pXp και Z3=α31X1+α32X2 ...+α3pXp. Αυτή η διαδικασία συνεχίζεται μέχρι να δημιουργηθούν ένα σύνολο p συνιστωσών Zi από τις αρχικές μεταβλητές. Είναι σημαντικό να τονιστεί ότι τα στοιχεία πρέπει να είναι ασυσχετισμένα μεταξύ τους, δηλαδή οι συντελεστές συσχέτισής τους πρέπει να είναι μηδενικοί. Οι συντελεστές aij στις συνιστώσες εξισώσεις αντιπροσωπεύουν την ειδική στάθμιση της μεταβλητής j στη συνιστώσα i, διασφαλίζοντας ότι επιτυγχάνεται η μέγιστη διακύμανση των μεταβλητών z. Ο υπολογισμός των συντελεστών στάθμισης aij βασίζεται στον πίνακα συνδιακύμανσης C των αρχικών μεταβλητών. Με τη διαδικασία τυποποίησης των αρχικών μεταβλητών, ο πίνακας των συνδιακυμάνσεων μετασχηματίζεται στον πίνακα των συσχετισμών. Αυτός ο πίνακας συσχετίσεων χρησιμεύει ως το θεμελιώδες στοιχείο στην ανάλυση των κύριων συστατικών. Οι διακυμάνσεις των κύριων συστατικών, που δηλώνονται ως λi, αναφέρονται ως ιδιοτιμές ή χαρακτηριστικές ρίζες. Είναι σημαντικό να σημειωθεί ότι αυτές οι ιδιοτιμές πρέπει να τηρούν τη συνθήκη λ1,λ2,...,λP > 0 και το άθροισμά τους πρέπει να ισούται με το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών, που συμβολίζονται ως c11+c22+...+cpp . Με άλλα λόγια, οι ιδιοτιμές αντιπροσωπεύουν συλλογικά τη συνολική διακύμανση που υπολογίζεται από τις αρχικές μεταβλητές. Επιπλέον, οι συσχετίσεις, που ονομάζονται rij, μεταξύ των αρχικών μεταβλητών και των κύριων συνιστωσών είναι γνωστές ως φορτίσεις. Αυτές οι φορτίσεις παρέχουν μια εικόνα για το επίπεδο επιρροής που ασκούν οι αρχικές μεταβλητές στη δημιουργία των στοιχείων, υποδεικνύοντας ουσιαστικά τον βαθμό στον οποίο συμβάλλουν στη συνολική κατασκευή των εξαρτήσεων. Συμπερασματικά, μπορεί να ειπωθεί ότι η διαδικασία της ανάλυσης των κύριων συστατικών περιλαμβάνει πολλά βασικά βήματα. Αυτά τα βήματα περιλαμβάνουν τον μετασχηματισμό των αρχικών μεταβλητών X1, X2,... Xp ώστε να έχουν μοναδιαία διακύμανση και μέσο όρο μηδέν, υπολογισμό του πίνακα συσχέτισης, προσδιορισμό των συντελεστών στάθμισης aij και εύρεση των ιδιοτιμών λ1, λ2... λp. Ο σκοπός αυτών των βημάτων είναι ο εντοπισμός και η εξάλειψη στοιχείων που συνεισφέρουν ελάχιστη μεταβλητότητα στο συνολικό σύνολο δεδομένων.




**Εικόνα 6 Ανάλυση Κυρίων Συνιστωσών**

**Πηγή: https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html/2**

### 3.2.2 Μέθοδος Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares)

Τα Μερικά Ελάχιστα Τετράγωνα (PLS) είναι μια εξελιγμένη στατιστική τεχνική που μοιράζεται ομοιότητες με την ανάλυση κύριου συστατικού (PCA). Παρόλα αυτά, το PLS έχει αποδειχθεί ότι ξεπερνά την τυπική παλινδρόμηση και το PCA σε περιπτώσεις όπου συμβαίνει υπερπροσαρμογή. Αυτό συμβαίνει όταν ο αριθμός

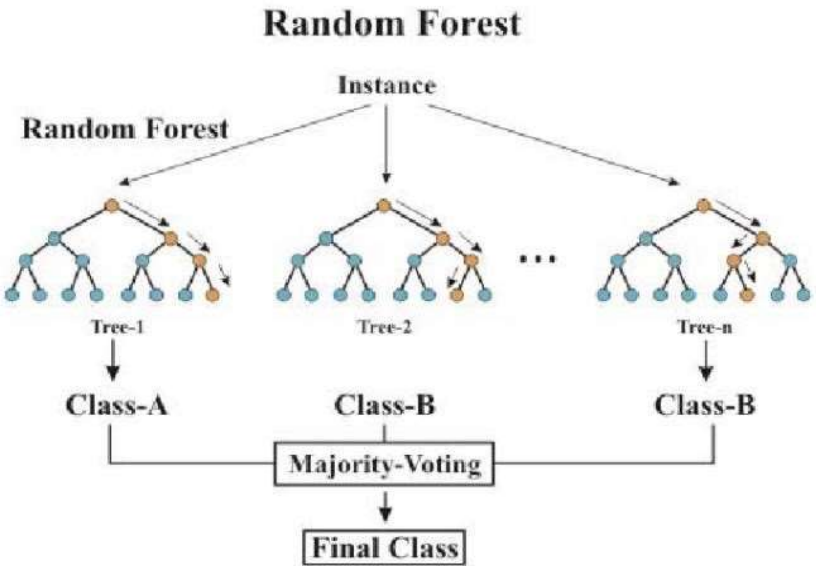
των μεταβλητών στον πίνακα πρόβλεψης υπερβαίνει τον αριθμό των παρατηρήσεων, με αποτέλεσμα ένα μοντέλο που ταιριάζει καλά στα υπάρχοντα δεδομένα, αλλά δυσκολεύεται να προβλέψει με ακρίβεια νέα σημεία δεδομένων. Ο στόχος των μερικών ελαχίστων τετραγώνων (PLS) είναι να προσδιορίσει βασικά στοιχεία του X που μπορούν να προβλέψουν αποτελεσματικά το Y συλλαμβάνοντας την υψηλότερη δυνατή συνδιακύμανση και εξάγοντας λανθάνουσες μεταβλητές που αντιπροσωπεύουν τη μεγαλύτερη διασπορά στη μεταβλητή απόκρισης, με αποτέλεσμα ένα ισχυρό μοντέλο πρόβλεψης. Οι λανθάνουσες μεταβλητές διαδραματίζουν κρίσιμο ρόλο στη μέθοδο PLS, γι' αυτό το PLS αναφέρεται συχνά ως προβολή σε λανθάνουσες δομές. Επιπλέον, η μερική ανάλυση ελάχιστων τετραγώνων (PLS-DA) είναι μια παραλλαγή του PLS που χρησιμοποιείται ειδικά για εργασίες ταξινόμησης. Το dataset περιλαμβάνει δύο πίνακες: τον X, που περιέχει τις ανεξάρτητες μεταβλητές για N παρατηρήσεις, και τον Y, που περιέχει τις εξαρτημένες μεταβλητές για τις ίδιες παρατηρήσεις. [81],[85]Η μέθοδος παλινδρόμησης μέριμνας (Partial Least Squares, PLS) αναλύει και τους δύο πίνακες X και Y σε έναν συνδυασμό κοινών ορθογωνικών παραγόντων και συγκεκριμένων φορτίων. Αυτό σημαίνει ότι ο X μπορεί να αναπαρασταθεί ως TPT, όπου ο T είναι ο πίνακας βαθμολογίας (και ορισμένες παραλλαγές της τεχνικής δεν απαιτούν τον T να έχει πρότυπα μονάδας). Ο πίνακας T, γνωστός ως πίνακας βαθμολογίας, και ο πίνακας P, που δεν είναι ορθογώνιος στο PLS, είναι βασικά συστατικά αυτής της διαδικασίας αποσύνθεσης. [81],[85]. Αντίστοιχα, η Y μπορεί να εκτιμηθεί ως $\hat{Y}$ = TBCT, με τον B να είναι ένας διαγώνιος πίνακας με βάρη παλινδρόμησης ως διαγώνια στοιχεία. Οι στήλες του T αντιπροσωπεύουν τις λανθάνουσες μεταβλητές και όταν ο αριθμός τους ταιριάζει με τις διαστάσεις του X, έχουμε αποσυνθέσει επιτυχώς τον X. Σημαντικό είναι να σημειωθεί ότι, ενώ η Y υπολογίζεται χρησιμοποιώντας αυτήν τη μέθοδο, το $\hat{Y}$ μπορεί να μην είναι ίσο με τον Y γενικά. Η επιλογή των λανθανουσών μεταβλητών στη μεθοδολογία PCA (Ανάλυση Κύριων Συνιστωσών) μπορεί να γίνει με διάφορους τρόπους. Ένας τρόπος είναι η χρήση όλων των ορθογωνικών φορέων που καλύπτουν τον χώρο των στηλών του πίνακα X για να αναπαραστήσουν τον χώρο των δεδομένων.Για να προσδιοριστεί το T (ο πίνακας των απεικονίσεων των δεδομένων στον νέο χώρο), απαιτούνται πρόσθετες συνθήκες. Συνήθως, η επιλογή γίνεται με βάση τη διακύμανση των δεδομένων ή τη συμβολή των κύριων συνιστωσών στην αναπαράσταση των αρχικών δεδομένων. Συγκεκριμένα, οι κύριες συνιστώσες που προσφέρουν τις μεγαλύτερες διακυμάνσεις επιλέγονται συνήθως για να αναπαραστήσουν τον χώρο των δεδομένων με λιγότερες διαστάσεις. Σε ορισμένες περιπτώσεις, μπορεί να χρησιμοποιηθούν κριτήρια όπως οι ελάχιστες απώλειες πληροφορίας ή η εκτίμηση της επίδοσης του μοντέλου με διαφορετικές συνιστώσες..

Η μερική παλινδρόμηση ελαχίστων τετραγώνων είναι μια μέθοδος που χρησιμοποιείται για την εκτίμηση του γραμμικού συνδυασμού μεταξύ δύο συνόλων δεδομένων, τα οποία εκφράζονται από τις μεταβλητές X και Y. Αυτή η μέθοδος επιδιώκει να βρει έναν γραμμικό συνδυασμό των στηλών των X και Y που μεγιστοποιεί τη συνδιακύμανσή τους. Για να επιτευχθεί αυτό, πρέπει να προσδιοριστούν δύο σύνολα βαρών, w και c. Όταν εντοπιστεί η πρώτη λανθανούσα μεταβλητή, αφαιρείται από τα X και Y και η διαδικασία επαναλαμβάνεται έως ότου ο πίνακας X γίνει μηδενικός. Η διαδικασία ξεκινά με τη δημιουργία δύο πινάκων, E=X και F=YF, οι οποίοι στη συνέχεια μετασχηματίζονται σε Z-scores. Πριν από την έναρξη της διαδικασίας επανάληψης, ο φορέας u αρχικοποιείται με τυχαίες τιμές. Τα βήματα που ακολουθούν περιλαμβάνουν τον υπολογισμό των βαρών w και c, την εκτίμηση των βαθμολογιών τους και την ενημέρωση των φορτίων για τον X και τον Y. Η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθούν οι επιθυμητές τιμές των φορτίων και ο πίνακας X γίνει μηδενικός..

## 3.3 Τυχαία Δάση (Random Forests)

Τα τυχαία δάση, επίσης γνωστά ως τυχαία δάση αποφάσεων, είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα παλινδρόμησης, κατηγοριοποίησης ή ταξινόμησης. Αυτό επιτυγχάνεται με τη δημιουργία ενός συνόλου δέντρων αποφάσεων, τα οποία συνεργάζονται για να παράγουν το τελικό αποτέλεσμα. Αν και μοιάζουν με τα δέντρα αποφάσεων, τα τυχαία δάση αποδεικνύονται πιο αποτελεσματικά σε περιπτώσεις υπερβολικής προσαρμογής, όπου τα απλά δέντρα αποφάσεων αποτυγχάνουν.

Το πρώτο σύστημα για τα τυχαία δάση αποφάσεων αναπτύχθηκε από τον Tin Kam Ho το 1995. Αυτό επιτεύχθηκε με τη χρήση της μεθόδου των τυχαίων υποσυνόλων ως μέσο για την υλοποίηση της προσέγγισης της "στοχαστικής διάκρισης" για την ταξινόμηση, η οποία είχε προταθεί λίγα χρόνια νωρίτερα από τον Kleinberg. Καθώς πέρασε ο χρόνος, εμφανίστηκαν βελτιωμένες εκδοχές του αλγορίθμου του Ho, οδηγώντας στον συνδυασμό της ιδέας του "bagging" με την τυχαία επιλογή χαρακτηριστικών. Αυτός ο συνδυασμός δημιούργησε μια συλλογή δέντρων αποφάσεων με ελεγχόμενη διακύμανση, γνωστή ως "τυχαία δάση". Η συνολική δομή του τυχαίου δάσους παρέχει αξιόπιστες προβλέψεις και αναγνωρίστηκε ως ένα ισχυρό εργαλείο μηχανικής μάθησης. Η σημαντική ανάπτυξη αυτής της μεθόδου οδήγησε στην αναγνώριση του εμπορικού σήματος "τυχαία δάση" από τους Breiman και Cutler..




**Εικόνα 7 Απεικόνιση ταξινομητή υλοποιημένου με αλγόριθμο τυχαίων δασών**
**Πηγή: https://www.semanticscholar.org**

Είναι εμφανές ότι τα τυχαία δάση περιλαμβάνουν πολλά δέντρα αποφάσεων και έχουν κοινά χαρακτηριστικά. Παρ' όλα αυτά, υπάρχουν σημαντικές διαφορές σε σχέση με τα κλασικά δέντρα αποφάσεων. Στα τυχαία δάση, αντί να επιλέγονται τα χαρακτηριστικά που μεταφέρουν την περισσότερη πληροφορία για την αρχική διαίρεση του δέντρου, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών από τα δεδομένα βάσης. Αυτό το υποσύνολο χαρακτηριστικών χρησιμοποιείται τυχαία σε κάθε κόμβο του δέντρου. Όταν ολοκληρώνεται η κατασκευή ενός τυχαίου δάσους, ένα τμήμα των δεδομένων χρησιμοποιείται για την εκπαίδευση, ενώ το υπόλοιπο χρησιμοποιείται για τον έλεγχο του σφάλματος. Η ομοιότητα μεταξύ των δέντρων ενός τυχαίου δάσους είναι σημαντική, καθώς εξαρτάται από την κοινή χρήση χαρακτηριστικών κατά την κατασκευή και επηρεάζει το σφάλμα του δέντρου.

Το bootstrapping είναι μια μέθοδος στα στατιστικά όπου ένα σύνολο δεδομένων δειγματοληπτείται τυχαία με αντικατάσταση, που χρησιμοποιείται συνήθως για τη μέτρηση της αβεβαιότητας ενός μοντέλου μηχανικής μάθησης. Είναι μια πολύτιμη τεχνική καθώς επιτρέπει τη δημιουργία νέων δειγμάτων από υπάρχοντα δεδομένα χωρίς την ανάγκη πρόσθετης συλλογής δεδομένων. Ουσιαστικά, η μέθοδος bootstrap περιλαμβάνει τη δημιουργία πολλαπλών συνόλων εκπαίδευσης με επαναδειγματοληψία των δεδομένων με αντικατάσταση από το αρχικό σύνολο. Αυτά τα νέα σετ εκπαίδευσης μπορούν στη συνέχεια να χρησιμοποιηθούν από μεθόδους "meta-learner" ή "ensemble" για τη μείωση της διακύμανσης πρόβλεψης και τη βελτίωση της συνολικής ακρίβειας πρόβλεψης.

Η τεχνική του Bootstrap Aggregation, ή αλλιώς bagging, είναι μια αποτελεσματική μέθοδος για τη βελτίωση της απόδοσης και της σταθερότητας αλγορίθμων μηχανικής μάθησης, είτε πρόκειται για αλγόριθμους ταξινόμησης είτε παλινδρόμησης. Αυτή η τεχνική έχει ως βασικό στόχο τη μείωση της διακύμανσης και την αντιμετώπιση του φαινομένου overfitting.

Η λειτουργία του bagging βασίζεται στη δημιουργία πολλών υποσυνόλων του συνόλου εκπαίδευσης, χρησιμοποιώντας τυχαία επιλογή με επανάληψη (Bootstrap). Έπειτα, για κάθε υποσύνολο, εκπαιδεύεται ένας αλγόριθμος μάθησης, όπως ένα δένδρο αποφάσεων. Τελικά, οι προβλέψεις από όλους τους αλγορίθμους συγκεντρώνονται και συνδυάζονται με κάποιον τρόπο (συνήθως μέσο όρο) για την τελική απόφαση.Μέσω της δημιουργίας πολλών υποσυνόλων και της εκπαίδευσης πολλαπλών αλγορίθμων, το bagging βοηθά στη μείωση της διακύμανσης του μοντέλου, καθώς και στην αποφυγή του overfitting. Επιπλέον, ενισχύει τη σταθερότητα του μοντέλου, καθώς η μέση τιμή των προβλέψεων από πολλαπλούς αλγορίθμους μπορεί να παραμείνει σταθερή, ανεξάρτητα από μικρές διακυμάνσεις στα δεδομένα εκπαίδευσης..

Η ιδέα είναι ότι δημιουργούμε πολλαπλά υποσύνολα του συνόλου εκπαίδευσης, κάθε ένα από τα οποία επιλέγεται τυχαία και με αντικατάσταση. Στη συνέχεια, εκπαιδεύουμε ξεχωριστά μοντέλα με τα διαφορετικά υποσύνολα αυτά. Το τελικό μοντέλο συνδυάζει τις προβλέψεις από όλα αυτά τα μοντέλα, συνήθως με μέσο όρο.Αυτή η προσέγγιση βοηθά στη μείωση της διακύμανσης του μοντέλου, καθώς οι προβλέψεις προέρχονται από διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης και επομένως είναι λιγότερο ευαίσθητες στις μικρές διακυμάνσεις των δεδομένων εκπαίδευσης. Αυτό μπορεί να βοηθήσει στην αντιμετώπιση του overfitting και να οδηγήσει σε πιο σταθερές προβλέψεις. Ωστόσο, όπως αναφέρει και ο Breiman, σε ορισμένες περιπτώσεις μπορεί να υπάρξει μια μικρή μείωση της απόδοσης, κυρίως σε περιπτώσεις όπου οι διαδικασίες που προσπαθούμε να μοντελοποιήσουμε είναι ήδη σταθερές. [3]

## 3.4 Ταξινομητές Naive-Bayes (Naive-Bayes classifiers)

Οι Μπεϋζιανοί ταξινομητές είναι σχεδιασμένοι για να προσεγγίσουν την πιθανοτική σχέση μεταξύ των χαρακτηριστικών ενός συνόλου δεδομένων και των κατηγοριών στις οποίες ανήκουν. Ο στόχος των ταξινομητών Naive-Bayes είναι να εκτιμήσουν την πιθανότητα ενός δείγματος να ανήκει σε μια συγκεκριμένη κατηγορία, βάσει των χαρακτηριστικών του δείγματος.[22] Αυτοί οι ταξινομητές βασίζονται στο θεώρημα του Bayes και υποθέτουν ανεξαρτησία μεταξύ των χαρακτηριστικών. Ένα παράδειγμα απλούστερου Μπεϋζιανού ταξινομητή είναι ο Naïve Bayesian, ο οποίος υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών και προσεγγίζει κάθε χαρακτηριστικό ως ανεξάρτητο προς τα άλλα, γεγονός που απλοποιεί τους υπολογισμούς και βοηθάει στην αποτελεσματική κατηγοριοποίηση. Το θεώρημα του Bayes είναι μια σημαντική αρχή στη στατιστική και την πιθανοτική, που περιγράφει τον τρόπο με τον οποίο ενημερωνόμαστε για την πιθανότητα της εμφάνισης μιας συγκεκριμένης συμβάντος με βάση τη διαθέσιμη πληροφορία. Η εξίσωση του θεωρήματος του Bayes είναι η εξής:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

**Εξίσωση 2**

όπου:

- Η P(A|B) είναι η πιθανότητα του γεγονότος Α να συμβεί δεδομένου ότι συνέβη το Β, ενώ η P(B|A) είναι η πιθανότητα του γεγονότος Β να συμβεί δεδομένου ότι συνέβη το Α.
- Η P(A) και P(B) είναι οι απλές πιθανότητες των γεγονότων Α και Β αντίστοιχα.
- Η P(A) ονομάζεται εκ των προτέρων πιθανότητα επειδή αναφέρεται στην πιθανότητα του γεγονότος Α χωρίς να ληφθεί υπόψη το Β, ενώ η P(B) είναι η πιθανότητα του δεδομένου Β ανεξάρτητα από την υπόθεση Α.

Η εκ των υστέρον πιθανότητα P(A|B) χρησιμοποιείται για να ενημερώσει την πιθανότητα του γεγονότος Α βάσει της παρατήρησης του Β. Αντίστοιχα, η P(B|A) προσφέρει πληροφορίες για την πιθανότητα του Β δεδομένου ότι η υπόθεση Α ισχύει.Ο Μπεϋζιανός ταξινομητής βασίζεται στο θεώρημα του Bayes, το οποίο περιγράφει τη σχέση μεταξύ της πιθανότητας ενός γεγονότος με βάση την εμπειρία ή την προηγούμενη πληροφορία. Στην ταξινόμηση, αυτό αναφέρεται στην υπολογιστική πρόβλεψη της κατηγορίας ενός δείγματος βάσει των πιθανοτήτων συσχέτισής του με κάθε κατηγορία.Όταν τα δεδομένα εκπαίδευσης δεν καλύπτουν πλήρως τα χαρακτηριστικά, ο Μπεϋζιανός ταξινομητής αντιμετωπίζει δυσκολίες στον υπολογισμό των πιθανοτήτων αυτών. Για να αντιμετωπίσει αυτό το πρόβλημα, μπορεί να χρησιμοποιήσει την υπόθεση της ανεξαρτησίας των χαρακτηριστικών, δηλαδή το θεώρημα του Bayes, υποθέτοντας ότι η παρουσία ενός χαρακτηριστικού δεν επηρεάζει την παρουσία ή την απουσία άλλων. Αυτή η προσέγγιση μπορεί να επιτρέψει στο μοντέλο να προβλέπει κατηγορίες ακόμη και όταν τα δεδομένα είναι ατελή.

### 3.4.1 Gaussian Naive Bayes

Η προσέγγιση Gaussian Naive Bayes είναι μια επέκταση των ταξινομητών Naive Bayes που μπορεί να αποφέρει εξαιρετικά αξιόπιστα αποτελέσματα όταν τα δεδομένα εκπαίδευσης υποτίθεται ότι ακολουθούν μια κατανομή Gauss. Αυτή η προσέγγιση θεωρείται η ευκολότερη μεταξύ των διαθέσιμων εναλλακτικών λόγω της απλότητάς της στους υπολογισμούς, οι οποίοι περιλαμβάνουν αποκλειστικά τον προσδιορισμό του μέσου όρου και της τυπικής απόκλισης των δεδομένων εκπαίδευσης. Για να το δείξουμε, ας εξετάσουμε ένα σενάριο όπου τα δεδομένα εκπαίδευσης περιλαμβάνουν ένα συνεχές χαρακτηριστικό, που συμβολίζεται ως x. Για να εφαρμόσουμε τον Gaussian Naive Bayes, ταξινομούμε πρώτα τα δεδομένα ανά κλάση και στη συνέχεια υπολογίζουμε τον μέσο όρο και τη διακύμανση του x σε κάθε κλάση. Για μια δεδομένη κλάση Ck, έστω το μk αντιπροσωπεύει τον μέσο όρο των τιμών x που σχετίζονται με το Ck και έστω σ2k τη διακύμανση. Τώρα, ας πούμε ότι έχουμε συγκεντρώσει μια τιμή παρατήρησης, που συμβολίζεται ως v. Μπορούμε να προσδιορίσουμε την κατανομή πιθανότητας της v δεδομένης κλάσης Ck, συνδέοντας το v στην εξίσωση μιας κανονικής κατανομής, χρησιμοποιώντας τις παραμέτρους μέσου όρου και διακύμανσης.

### 3.4.2 Multinomial Naive Bayes

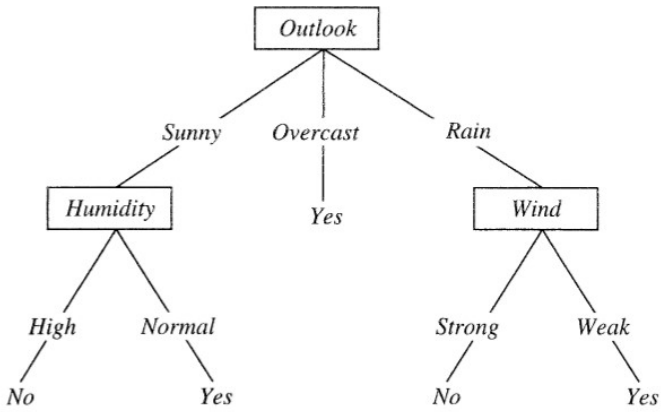
Οι ταξινομητές multinomial Bayes ανήκουν σε μια διαφορετική κατηγορία από τους Gaussian ταξινομητές. Δεν είναι κατάλληλοι για συνεχείς μεταβλητές, αλλά συχνά χρησιμοποιούνται για προβλήματα ταξινόμησης κειμένου. Με τη χρήση ενός πολυωνυμικού μοντέλου, τα δείγματα αντιστοιχούν στις συχνότητες εμφάνισης των γεγονότων από ένα πολυώνυμο. Κάθε χαρακτηριστικό του διανύσματος αντιπροσωπεύει τη συχνότητα εμφάνισης ενός ενδεχομένου.

### 3.4.3 Bernoulli Naive Bayes

Το μοντέλο Bernoulli Naive Bayes είναι μια εκδοχή των Μπεϋζιανών ταξινομητών που χρησιμοποιείται ευρέως στην ταξινόμηση κειμένων. Σε αντίθεση με το προηγούμενο μοντέλο multinomial, όπου λαμβάνεται υπόψη η συχνότητα εμφάνισης των χαρακτηριστικών, το μοντέλο Bernoulli λαμβάνει υπόψη μόνο την παρουσία ή την απουσία των δυαδικών χαρακτηριστικών σε ένα έγγραφο.Συγκεκριμένα, σε αυτό το μοντέλο, κάθε χαρακτηριστικό αντιπροσωπεύει την παρουσία ή την απουσία ενός συγκεκριμένου όρου σε ένα έγγραφο. Αυτό καθιστά τον ταξινομητή κατάλληλο για ταξινόμηση μικρών εγγράφων όπου η συχνότητα των όρων δεν είναι τόσο σημαντική όσο η παρουσία τους.Για παράδειγμα, σε ένα πρόβλημα κατηγοριοποίησης κειμένων, το μοντέλο Bernoulli Naive Bayes μπορεί να χρησιμοποιηθεί για να κατανοήσει εάν ένα κείμενο περιέχει ή όχι συγκεκριμένες λέξεις ή όρους, αγνοώντας τον αριθμό των φορών που εμφανίζονται αυτοί οι όροι..

## 3.5 Δέντρα αποφάσεων (Decision Trees)

Οι αλγόριθμοι δένδρων αποφάσεων είναι ένας τύπος μεθόδου μηχανικής μάθησης που περιλαμβάνει την κατασκευή μιας δενδρικής δομής για την κατηγοριοποίηση δεδομένων με βάση τις τιμές διαφορετικών χαρακτηριστικών. Κάθε κόμβος στο δέντρο αντιπροσωπεύει ένα συγκεκριμένο χαρακτηριστικό που ταξινομείται και οι κλάδοι του δέντρου αντιπροσωπεύουν τις πιθανές τιμές που μπορεί να λάβει αυτό το χαρακτηριστικό. Ο αλγόριθμος C4.5 είναι ένας δημοφιλής αλγόριθμος δέντρων απόφασης που αναπτύχθηκε από τον Quinlan ως μια βελτίωση του προηγούμενου αλγορίθμου του, του ID3. Στην ουσία, ένα δέντρο αποφάσεων είναι μια δομή που αποτελείται από μια σειρά κανόνων που καθοδηγούν τη διαδικασία λήψης αποφάσεων. Κάθε κανόνας οδηγεί είτε σε έναν άλλο κανόνα, είτε σε μια τελική απόφαση. Ο αλγόριθμος C4.5 βασίζεται σε μια σειρά από επιλογές χαρακτηριστικών και κατασκευάζει ένα δέντρο που μπορεί να χρησιμοποιηθεί για την ταξινόμηση δεδομένων ή τη λήψη αποφάσεων σε ένα πρόβλημα. Για παράδειγμα, μπορεί να έχουμε έναν κανόνα που λέει ότι θα παίξουμε τένις εκτός εάν ο καιρός είναι εξαιρετικά κακός, κάτι που θα οδηγήσει στη συνέχεια στην απόφαση να μην παίξουμε. Αυτή η διαδικασία μπορεί να απεικονιστεί ως δομή δέντρου, με κάθε κλάδο να αντιπροσωπεύει ένα διαφορετικό σημείο απόφασης. Συνολικά, οι αλγόριθμοι του δέντρου αποφάσεων είναι ένα ισχυρό εργαλείο για την οργάνωση και την ανάλυση δεδομένων, επιτρέποντας σε πολύπλοκες διαδικασίες λήψης αποφάσεων να αναλύονται σε απλά, λογικά βήματα.




**Εικόνα 8 : Ένα δέντρο απόφασης**

Χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης, ένας υπολογιστής έχει τη δυνατότητα να συνάγει δέντρα αποφάσεων μέσω της ανάλυσης διαφόρων παραδειγμάτων που παρουσιάζουν διαφορετικές συνθήκες (χαρακτηριστικά) και αντίστοιχα αποτελέσματα (ταξινομήσεις ή αποφάσεις). Αυτά τα δέντρα αποφάσεων μπορούν να αναπαρασταθούν με τη μορφή ενός πίνακα ή μιας συλλογής εκφράσεων Boolean, επιτρέποντας μια συνοπτική και οργανωμένη αναπαράσταση. Για παράδειγμα, το προαναφερθέν δέντρο μπορεί να μετατραπεί σε γραπτή μορφή που ενσωματώνει τη δομή και τις πληροφορίες του:

(Outlook = Sunny ∧ Humidity = Normal) ∨ (Outlook = Overcast) ∨ (Outlook = Rain ∧Wind=Weak)

Για να δημιουργηθεί ένα δέντρο απόφασης που ταξινομεί τιμές όπως (1, 2, 3, 4) ή {άσπρο, μαύρο}, μπορεί να χρησιμοποιηθεί ο επαναληπτικός αλγόριθμος ID3 (Iterative Dichotomiser 3). Αυτός ο αλγόριθμος εξετάζει τα συνεχή και κατηγορικά χαρακτηριστικά και βασίζεται στην έννοια της εντροπίας και της κέρδους πληροφορίας για τη διαίρεση των δεδομένων σε διακριτές κατηγορίες. Αυτός είναι ένας άπληστος αλγόριθμος, που σε κάθε κόμβο του δέντρου επιλέγει το χαρακτηριστικό που προσφέρει τη μέγιστη πληροφορία για την ταξινόμηση. Κατά την επεξεργασία ενός κόμβου, ο αλγόριθμος δοκιμάζει διαδοχικά τα διαφορετικά χαρακτηριστικά, λαμβάνοντας υπόψη τα προηγούμενα αποτελέσματα. Η διαδικασία αυτή συνεχίζεται μέχρι όλα τα παραδείγματα του εκπαιδευτικού συνόλου να ταξινομηθούν σωστά, πράγμα που

ενδέχεται να οδηγήσει σε υπερβολική υπολογιστική πολυπλοκότητα και μειωμένη απόδοση στα δεδομένα εκπαίδευσης.

| OUTLOOK | TEMPERATURE | HUMIDITY | WINDY | PLAY |
|---|---|---|---|---|
| sunny | 85 | 85 | false | Don't Play |
| sunny | 80 | 90 | true | Don't Play |
| overcast | 83 | 78 | false | Play |
| rain | 70 | 96 | false | Play |
| rain | 68 | 80 | false | Play |
| rain | 65 | 70 | true | Don't Play |
| overcast | 64 | 65 | true | Play |
| sunny | 72 | 95 | false | Don't Play |
| sunny | 69 | 70 | false | Play |
| rain | 75 | 80 | false | Play |
| sunny | 75 | 70 | true | Play |
| overcast | 72 | 90 | true | Play |
| overcast | 81 | 75 | false | Play |
| rain | 71 | 80 | true | Don't Play |

**Πίνακας 1: Ο πίνακας που αντιστοιχεί στο παραπάνω δέντρο**

Σε αυτό το συγκεκριμένο σενάριο, είναι σημαντικό να σημειωθεί ότι δύο από τα χαρακτηριστικά που αναλύονται, η θερμοκρασία και η υγρασία, δεν διαθέτουν συνεχή εύρη. Το ID3, ο αλγόριθμος που χρησιμοποιείται, δεν είναι εξοπλισμένος για να χειρίζεται απευθείας τέτοιες καταστάσεις. Ωστόσο, στην παρακάτω συζήτηση, θα διερευνήσουμε πιθανές επεκτάσεις που θα μπορούσαν να επιτρέψουν στο ID3 να χειρίζεται αποτελεσματικά αυτούς τους τύπους χαρακτηριστικών. Είναι σημαντικό να κατανοήσουμε ότι ο τελικός στόχος της δημιουργίας ενός δέντρου αποφάσεων δεν είναι απλώς η σύνοψη των υπαρχόντων δεδομένων, αλλά μάλλον η ακριβής ταξινόμηση νέων περιπτώσεων που μπορεί να προκύψουν στο μέλλον. Ως εκ τούτου, κατά την κατασκευή μοντέλων ταξινόμησης, είναι σημαντικό όχι μόνο να υπάρχει μια σταθερή βάση δεδομένων εκπαίδευσης για την κατασκευή του μοντέλου, αλλά και να υπάρχουν ξεχωριστά δεδομένα δοκιμής για την αξιολόγηση της απόδοσης και της ακρίβειάς του. Μια πιο απλή απεικόνιση από τη σφαίρα της χρηματιστηριακής ανάλυσης που εστιάζει αποκλειστικά σε διακριτά δεδομένα είναι η έννοια του Κέρδους ως μεταβλητής, η οποία μπορεί να ταξινομηθεί σε δύο διακριτές κατηγορίες: {up, down}.

Τα μη κατηγορηματικά χαρακτηριστικά του είναι: [39]

| ATTRIBUTE | POSSIBLE VALUES |
|---|---|
| age | old, midlife, new |
| competition | no, yes |
| type | software, hardware |

**Πίνακας 2 : Attributes**

Και το εκπαιδευτικό σετ είναι:

| AGE | COMPETITION | TYPE | PROFIT |
|---|---|---|---|
| old | yes | swr | down |
| old | no | swr | down |
| old | no | hwr | down |
| mid | yes | swr | down |
| mid | yes | hwr | down |
| mid | no | hwr | up |
| mid | no | swr | up |
| new | yes | swr | up |
| new | no | hwr | up |
| new | no | swr | up |

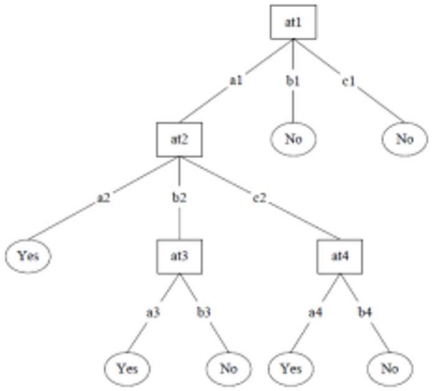
**Πίνακας 3: Εκπαιδευτικό σετ**

Οι βασικές ιδέες πίσω από το ID3 είναι ότι:

Ένα δέντρο αποφάσεων είναι μια δομή δεδομένων όπου κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό που δεν έχει ακόμη κατηγοριοποιηθεί και κάθε κλάδος αντιπροσωπεύει μια πιθανή τιμή που θα μπορούσε να λάβει το χαρακτηριστικό. Το τέλος ενός κλάδου, ή φύλλου, υποδεικνύει το προβλεπόμενο αποτέλεσμα ή την τιμή του χαρακτηριστικού που έχει κατηγοριοποιηθεί με βάση τη διαδρομή από τη ρίζα

του δέντρου προς αυτό το φύλλο. Αυτή η δομή βοηθά στην οργάνωση και οπτικοποίηση των διαδικασιών λήψης αποφάσεων με σαφή και συστηματικό τρόπο.

Στο δέντρο αποφάσεων, κατά την προσέγγιση ενός κόμβου, επιλέγουμε κάθε φορά το μη κατηγορημένο χαρακτηριστικό που παρέχει τη μέγιστη πληροφορία μεταξύ των χαρακτηριστικών που δεν έχουν ακόμη ληφθεί υπόψη στη διαδρομή από τη ρίζα. Αυτή η προσέγγιση καθορίζει την έννοια ενός "καλού" δέντρου αποφάσεων. [34]Η εντροπία χρησιμοποιείται για να μετρήσει τον τρόπο με τον οποίο ένας κόμβος περιέχει πληροφορία. Αυτό ορίζει το τι εννοούμε με το "καλό". Η έννοια αυτή εισήχθη από τον Claude. Shannon στη Θεωρία της Πληροφορίας.




| at1 | at2 | at3 | at4 | Class |
|---|---|---|---|---|
| a1 | a2 | a3 | a4 | Yes |
| a1 | a2 | a3 | b4 | Yes |
| a1 | b2 | a3 | a4 | Yes |
| a1 | b2 | b3 | b4 | No |
| a1 | c2 | a3 | a4 | Yes |
| a1 | c2 | a3 | b4 | No |
| b1 | b2 | b3 | b4 | No |
| c1 | b2 | b3 | b4 | No |

**Εικόνα 9 Decision tree with yes no classes**

Ένα μέτρο της αξίας των στοιχείων pi στο set S είναι η εντροπία. η εντροπία χρησιμοποιείται ως μέτρο της ανασφάλειας ή του χάους σε ένα σύστημα. Στο πλαίσιο των δέντρων απόφασης, η εντροπία χρησιμοποιείται για να εκτιμήσει την αμφιβολία ή την αβεβαιότητα που σχετίζεται με την κατηγορία μιας δεδομένης παρατήρησης.

Η εντροπία ορίζεται ως:

$$Entropy(S) = -p_{pos} log_2(p_{pos}) - p_{neg} log_2(p_{neg})$$

**Εξίσωση 3**

όπου pi είναι η πιθανότητα εμφάνισης της κατηγορίας i. Σε ένα δέντρο απόφασης, η εντροπία χρησιμοποιείται για να αξιολογήσει το πόσο καλά ένα χαρακτηριστικό διαχωρίζει τα δεδομένα σε διαφορετικές κατηγορίες. Σκοπός είναι να επιλεγούν τα χαρακτηριστικά που μειώνουν την εντροπία και κατά συνέπεια διαχωρίζουν αποτελεσματικά τα δεδομένα.Ο αλγόριθμος δέντρων απόφασης συνήθως χρησιμοποιεί μέτρα όπως η εντροπία ή η καθαρότητα Gini για να επιλέξει το καλύτερο χαρακτηριστικό για διαχωρισμό σε κάθε βήμα του δέντρου.

Το κέρδος πληροφορίας υπολογίζεται ως η διαφορά μεταξύ της εντροπίας του γονικού κόμβου και του κλασμένου συνόλου των εντροπιών των παιδικών κόμβων που προκύπτουν από το διαχωρισμό. Η τύπος για το κέρδος πληροφορίας IG είναι ο εξής:

$$IG(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

**Εξίσωση 4**

Όπου:

- S είναι το γονικό dataset.
- A είναι το χαρακτηριστικό που εξετάζεται για τον διαχωρισμό.
- Values(A) είναι οι πιθανές τιμές του χαρακτηριστικού

- Sv είναι το υποσύνολο του S για το οποίο το χαρακτηριστικό A έχει την τιμή v.
- |S| και |Sv | υποδηλώνουν τον αριθμό των παραδειγμάτων στα σύνολα S και Sv , αντίστοιχα.

Το χαρακτηριστικό με το υψηλότερο κέρδος πληροφορίας επιλέγεται για το διαχωρισμό σε κάθε κόμβο του δέντρου απόφασης. Αυτή η διαδικασία επαναλαμβάνεται αναδρομικά μέχρι όλα τα παραδείγματα σε έναν κόμβο να ανήκουν στην ίδια κατηγορία ή να μην υπάρχουν άλλα χαρακτηριστικά για διαχωρισμό.

## 3.6 Ο Αλγόριθμος ID3 (Iterative Dichotomiser 3)

Ο αλγόριθμος ID3 εφαρμόζεται για τη δημιουργία ενός δέντρου αποφάσεων από ένα σύνολο μη κατηγοριοποιημένων χαρακτηριστικών C1, C2, .., Cn, ένα κατηγορηοποιημένο χαρακτηριστικό C και ένα σύνολο εκπαίδευσης T των εγγραφών.Ο αλγόριθμος ID3 κατασκευάζει δέντρα αποφάσεων με μια προσέγγιση από την κορυφή προς τα κάτω και με επιλογή κατά την εκτέλεση. Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

- Ξεκινάμε με ένα σύνολο εκπαίδευσης S που περιλαμβάνει χαρακτηριστικά και ταξινομήσεις.
- Καθορίζουμε το καλύτερο χαρακτηριστικό στο S βάσει ενός κριτηρίου (που θα προσδιοριστεί σύντομα).
- Διαιρούμε το S σε υποσύνολα που αντιστοιχούν στις πιθανές τιμές του καλύτερου χαρακτηριστικού.
- Δημιουργούμε έναν κόμβο δέντρου απόφασης που περιέχει το καλύτερο χαρακτηριστικό, τοποθετημένο στη ρίζα του δέντρου.
- Επαναληπτικά δημιουργούμε νέους κόμβους δέντρου απόφασης χρησιμοποιώντας τα υποσύνολα δεδομένων που δημιουργήθηκαν στο προηγούμενο βήμα. Τα χαρακτηριστικά δεν μπορούν να επαναχρησιμοποιηθούν στον ίδιο κόμβο. Αν ένα υποσύνολο δεδομένων συμφωνεί συνεχώς στην ταξινόμηση, επιλέγουμε αυτήν την ταξινόμηση. Αν δεν υπάρχουν περισσότερα χαρακτηριστικά διαθέσιμα για διαίρεση, επιλέγουμε την πιο δημοφιλή ταξινόμηση.

Ο ψευδοκώδικας που παρέχεται υποθέτει διακριτά χαρακτηριστικά και δυαδικές ταξινομήσεις (ναι ή όχι). Αντιμετωπίζει αντιφάσεις στα δεδομένα εκπαίδευσης επιλέγοντας την πιο δημοφιλή ετικέτα ταξινόμησης σε περίπτωση σύγκρουσης.

***def id3(examples, classification_attribute, attributes):***
***create a root node for the tree***
***if all examples are positive/yes:***
***return root node with positive/yes label***
***else if all examples are negative/no:***
***return root node with negative/no label***
***else if there are no attributes left:***
***return root node with most popular***
***classification_attribute label***
***else:***
***best_attribute = attribute from attributes that best***
***classifies examples***
***assign best_attribute to root node***
***for each value in best_attribute:***
***add branch below root node for the value***

***branch_examples = [examples that have that value***
***for best_attribute]***
***if branch_examples is empty:***
***add leaf node with most popular***
***classification_attribute label***
***else:***
***add subtree id3(branch_examples,***
***classification_attribute,***
***attributes - best_attribute)*** [90]

Ο αλγόριθμος λειτουργεί αναδρομικά, διαιρώντας τα δεδομένα βάσει του καλύτερου χαρακτηριστικού και αφαιρώντας το από τα διαθέσιμα χαρακτηριστικά για περαιτέρω διαίρεση σε κάθε αναδρομική κλήση. Οι βασικές περιπτώσεις ελέγχονται πρώτα, συμπεριλαμβανομένου του ότι όλα τα παραδείγματα έχουν την ίδια ταξινόμηση, δεν υπάρχουν περισσότερα χαρακτηριστικά για διαίρεση ή δεν υπάρχουν παραδείγματα που παραμένουν.

**Χρησιμοποιώντας συντελεστές κέρδους**

Το Gain (κέρδος) είναι ένα μέτρο που χρησιμοποιείται για να αξιολογήσει τη σημασία των χαρακτηριστικών σε ένα δέντρο απόφασης. Η ιδέα είναι να προτιμά τα χαρακτηριστικά που προσφέρουν περισσότερη πληροφορία για την ταξινόμηση των δεδομένων. [44]

Ορισμένα χαρακτηριστικά με μεγάλο αριθμό διαφορετικών τιμών μπορεί να έχουν μέγιστο Gain. Αυτό συμβαίνει επειδή, όπως σωστά παρατηρήσατε, το Info (D, T) είναι 0, που σημαίνει ότι το χαρακτηριστικό παρέχει απόλυτη πληροφορία για την ταξινόμηση των δεδομένων. Ωστόσο, αυτό μπορεί να οδηγήσει σε προβλήματα υπερ-ειδίκευσης (overfitting), καθώς ένα χαρακτηριστικό με πολλές διαφορετικές τιμές μπορεί να προσαρμοστεί υπερβολικά στα δεδομένα εκπαίδευσης και να μην γενικεύει καλά σε νέα δεδομένα.Επομένως, είναι σημαντικό να λαμβάνουμε υπόψη μας την ισορροπία μεταξύ του μεγέθους του Gain και του αριθμού των διαφορετικών τιμών ενός χαρακτηριστικού κατά την επιλογή των χαρακτηριστικών για την κατασκευή του δέντρου απόφασης. [12]Για να αντισταθμίσει αυτό ο Quinlan προτείνει να χρησιμοποιούμε τον ακόλουθο αντί για κέρδος:

$$\text{GainRatio}(D,T)=\frac{Gain(D,T)}{Splitinfo(D,T)}$$

**Εξίσωση 5**

όπου SplitInfo (D, T) είναι οι πληροφορίες που οφείλονται στη διάσπαση του T με βάση της τιμής του κατηγοριοποιημένου χαρακτηριστικού D.

## 3.8 C4.5 Επεκτάσεις του Αλγορίθμου ID3

Ο αλγόριθμος C4.5, που αναπτύχθηκε από την Quinlan το 1993, εισάγει δύο βασικές βελτιώσεις στη μέθοδο ID3: επιλογή διακύμανσης και μειωμένο σφάλμα. Μια σημαντική βελτίωση είναι η ικανότητα χειρισμού αριθμητικών χαρακτηριστικών, επιπλέον των κατηγορικών χαρακτηριστικών που υποστηρίζονται από το ID3. Τα σημεία διακοπής επιλέγονται στρατηγικά για τη βελτιστοποίηση του κέρδους πληροφοριών, επιτρέποντας πιο ακριβείς διαχωρισμούς στο δέντρο αποφάσεων. Σε αντίθεση με τα κατηγορικά χαρακτηριστικά, τα αριθμητικά χαρακτηριστικά διατηρούν την πληροφοριακή τους αξία ακόμη και μετά τη χρήση τους σε προηγούμενες διαιρέσεις, καθώς κάθε νέο διαχωρισμό μπορεί να αποκαλύψει πρόσθετες πληροφορίες με βάση διαφορετικά κριτήρια. [39]

Το κλάδεμα με μειωμένο σφάλμα είναι μια τεχνική στο κλάδεμα δέντρων αποφάσεων όπου ένα ολόκληρο υποδέντρο αντικαθίσταται με έναν κόμβο ενός φύλλου εάν ταξινομεί με ακρίβεια ένα υψηλό ποσοστό παραδειγμάτων, συνήθως γύρω στο 95%. Αυτό βοηθά στον εξορθολογισμό των διαδικασιών λήψης αποφάσεων και εμποδίζει το μοντέλο να γίνει υπερβολικά πολύπλοκο. Επιπλέον, εάν ένα υποδέντρο μέσα στο δέντρο βρεθεί ότι έχει σημαντικό σφάλμα προσέγγισης, μπορεί να μετακινηθεί σε υψηλότερο επίπεδο στη δομή του δέντρου.

Ο αλγόριθμος C4.5 χρησιμοποιεί τα δεδομένα εκπαίδευσης και μια ευρετική μέθοδο για να εκτιμήσει τα σφάλματα προσέγγισης[10]. Κατά τη διάρκεια της δημιουργίας του δέντρου, ο αλγόριθμος αξιολογεί τις διαφορετικές διαιρέσεις των δεδομένων και επιλέγει αυτή που θα οδηγήσει στη μεγαλύτερη μείωση του σφάλματος πρόσθεσης (ή του σφάλματος εκτίμησης) στους κόμβους του δέντρου. Στη συνέχεια, η παράμετρος που εκτιμάται είναι ο περιορισμός της επικινδυνότητας, ο οποίος καθορίζει πότε το δέντρο πρέπει να σταματήσει την ανάπτυξή του για να αποφευχθεί η υπερεκπαίδευση.

Ορίζουμε τους παρακάτω όρους:

- N είναι ο συνολικός αριθμός δειγμάτων,
- f αναπαριστά τον παρατηρούμενο ρυθμό σφάλματος, ο οποίος υπολογίζεται ως ο λόγος του αριθμού των λανθασμένων προβλέψεων προς τον συνολικό αριθμό των δειγμάτων (f = E / N),
- q είναι ο πραγματικός ρυθμός σφάλματος,
- c αντιπροσωπεύει τη βεβαιότητα, η οποία στο C4.5 ορίζεται ως 0.25.

Στο C4.5, μια συντηρητική εκτίμηση του σφάλματος, συμβολίζεται με το γράμμα e, δίνεται από την εξίσωση:

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}.$$

**Εξίσωση 6**

Η τιμή του z αντιπροσωπεύει τον αριθμό των τυπικών αποκλίσεων που αντιστοιχούν στο ανώτατο όριο παραβάσεων. Η αρχική πτυχή απορρέει από το προβλεπόμενο βάθος καταγραφής του δέντρου, το οποίο απαιτεί την ανάλυση όλων των περιπτώσεων και χαρακτηριστικών σε κάθε επίπεδο, με αποτέλεσμα τον παράγοντα mn. Η επόμενη πτυχή προκύπτει από την αφαίρεση των υποδέντρων, καθώς η αντικατάσταση μιας υποσυστοιχίας συνεπάγεται λειτουργία O(n). Για κάθε κόμβο στο δέντρο, απαιτείται αξιολόγηση σφάλματος, ακολουθούμενη από πιθανή αντικατάσταση οποιουδήποτε κόμβου. Η διαδικασία επαναταξινόμησης κάθε επιπέδου του δέντρου δεν είναι ομοιόμορφη, καθώς απαιτεί χρόνο O(logn) για την επαναταξινόμηση μιας υποτάξης. Ως αποτέλεσμα, το συνολικό υπολογιστικό κόστος για την ανύψωση των υποδέντρων είναι O(n(logn)^2).

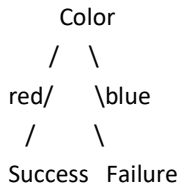
Το C4.5 είναι ένας δημοφιλής αλγόριθμος μηχανικής μάθησης για την κατασκευή δέντρων αποφάσεων. Παρόλο που θεωρείται ως ένας γρήγορος αλγόριθμος και έχει χρησιμοποιηθεί ευρέως σε πολλές εφαρμογές, υπάρχουν εναλλακτικοί αλγόριθμοι μηχανικής μάθησης που έχουν υπερβεί την απόδοσή του σε ορισμένα πραγματικά δεδομένα. Ωστόσο, το C4.5 και το C5.0 (η βελτιωμένη έκδοση του C4.5) παραμένουν αξιόπιστοι και ευρέως αναγνωρισμένοι ως αναφορές στον τομέα της μηχανικής μάθησης λόγω της απόδοσής τους και της ακρίβειάς τους.Οι εναλλακτικοί αλγόριθμοι μπορεί να περιλαμβάνουν το Random Forests, το Gradient Boosting Machines (GBM), το XGBoost, και άλλους. Αυτοί οι αλγόριθμοι έχουν συχνά επιδόσεις και χαρακτηριστικά που τους καθιστούν κατάλληλους για συγκεκριμένα προβλήματα μηχανικής μάθησης, όπως η εξερεύνηση μεγάλων συνόλων δεδομένων ή η αντιμετώπιση του φαινομένου της υπερεκπαίδευσης (overfitting). Ωστόσο, η επιλογή του αλγορίθμου εξαρτάται συχνά από τα χαρακτηριστικά του συγκεκριμένου προβλήματος και την πολυπλοκότητά του.

## 3.9 Δημιουργία Δέντρων Αποφάσεων

Το δέντρο αποφάσεων που δημιουργείται χρησιμοποιώντας το εκπαιδευτικό σύνολο δεδομένων, συνήθως, εστιάζει σωστά στην ταξινόμηση των περισσότερων δειγμάτων από αυτό το σύνολο. Ωστόσο, η διαδικασία αυτή μπορεί να είναι αρκετά πολύπλοκη, με μη ομοιόμορφες διαδρομές.

Το κλάδεμα του δέντρου αποφάσεων περιλαμβάνει την αφαίρεση ολόκληρων κόμβων μικρών φύλλων και την αντικατάστασή τους με μεγαλύτερα φύλλα κόμβων. Αυτή η αντικατάσταση είναι απαραίτητη όταν ένας κανόνας λήψης απόφασης οδηγεί στο συμπέρασμα ότι το αναμενόμενο ποσοστό σφάλματος στο μικρότερο φύλλο είναι υψηλότερο από αυτό του μεγαλύτερου. [44]

Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα απλό δέντρο απόφασης όπως το παρακάτω:




Στο εκπαιδευτικό σύνολο, έχουμε ένα επιτυχημένο δείγμα με χρώμα κόκκινο και δύο αποτυχημένα με χρώμα μπλε. Στο σετ δοκιμών, όμως, βρίσκουμε τρία αποτυχημένα δείγματα με χρώμα κόκκινο και ένα επιτυχημένο με χρώμα μπλε. Σε αυτήν την περίπτωση, μπορεί να αξιολογήσουμε την αντικατάσταση αυτού του δευτερεύοντος φύλλου με έναν μόνο κόμβο αποτυχίας. Μετά την αντικατάσταση, θα έχουμε μόνο δύο σφάλματα, αντί για πέντε αποτυχημένα δείγματα. Σε αυτό το παράδειγμα, ο Winston[23] επιδεικνύει την εφαρμογή της ακριβούς δοκιμής του Fisher ως μέσο προσδιορισμού της εξάρτησης μεταξύ ενός χαρακτηριστικού κατηγορίας και ενός μη κατηγοριοποιημένου χαρακτηριστικού. Εάν δεν υπάρχει εξάρτηση, υποδηλώνει ότι δεν χρειάζεται να συμπεριληφθεί το μη κατηγοριοποιημένο χαρακτηριστικό στη διαδρομή δέντρου αποφάσεων. [34]Οι Quinlan και Breiman προτείνουν πιο εξελιγμένα heuristics κλαδέματος.

Η εξαγωγή ενός συνόλου κανόνων από ένα δέντρο αποφάσεων είναι μια απλή διαδικασία. Κάθε διαδρομή από τη ρίζα σε ένα φύλλο στο δέντρο αποφάσεων αντιπροσωπεύει έναν κανόνα, ο οποίος μπορεί εύκολα να κατασκευαστεί συνδυάζοντας τις ετικέτες των κόμβων και των τόξων.

**Οι κανόνες που προκύπτουν μπορούν να απλουστευθούν:**

Στο πλαίσιο των συστημάτων που βασίζονται σε κανόνες, ο όρος "LHS" αναφέρεται στην αριστερή πλευρά ενός κανόνα. Αυτή η αριστερή πλευρά μπορεί να τροποποιηθεί αφαιρώντας ορισμένες συνθήκες, με αποτέλεσμα ένα τροποποιημένο LHS. Εάν τα υποσύνολα του συνόλου εκπαίδευσης που ικανοποιούν το αρχικό LHS και το τροποποιημένο LHS είναι πανομοιότυπα, τότε μπορούμε με σιγουριά να αντικαταστήσουμε το τροποποιημένο LHS στη θέση του αρχικού LHS εντός του κανόνα.

Ένας κανόνας μπορεί να εξαλειφθεί με τη χρήση άλλων κανόνων όπως "αν δεν ισχύει κανένας άλλος κανόνας".[39]

Η διαδικασία κατασκευής ενός δέντρου αποφάσεων σε ψευδοκώδικα συνήθως περιλαμβάνει τα ακόλουθα βήματα:

**python**

```
def construct_decision_tree(data):
    # Έλεγχος εάν έχει επιτευχθεί το κριτήριο τερματισμού
    if stopping_criteria(data):
        # Αν ναι, ανατίθεται η πιο συχνή ετικέτα κλάσης σε όλο το σύνολο εκπαίδευσης
        return assign_majority_class(data)
```

```
else:
    # Υπολογισμός της τιμής διαίρεσης για όλα τα χαρακτηριστικά και επιλογή του καλύτερου χαρακτηριστικού
    best_feature = find_best_split(data)
    # Διαίρεση του κόμβου σε πολλαπλούς κόμβους με βάση το επιλεγμένο χαρακτηριστικό
    subsets = split_node(data, best_feature)
    decision_tree = {}
    for value, subset in subsets.items():
      # Αναδρομική κλήση για κάθε υποσύνολο εκπαίδευσης
      decision_tree[value] = construct_decision_tree(subset)
    return decision_tree
```

Σε αυτό το παράδειγμα, η συνάρτηση stopping_criteria() ελέγχει εάν έχει επιτευχθεί το κριτήριο τερματισμού, όπως για παράδειγμα η επίτευξη ενός ελάχιστου αριθμού εγγραφών σε έναν κόμβο ή η επίτευξη μέγιστου βάθους του δέντρου. Η συνάρτηση find_best_split() υπολογίζει το χαρακτηριστικό που παρουσιάζει την καλύτερη τιμή διαίρεσης για την τρέχουσα κατάσταση. Η συνάρτηση split_node() διαιρεί τον κόμβο σε πολλαπλούς κόμβους, έναν για κάθε διαφορετική τιμή του επιλεγμένου χαρακτηριστικού. Η ανάλυση των υποσυνόλων συνεχίζεται αναδρομικά μέχρι να επιτευχθεί το κριτήριο τερματισμού για κάθε υποσύνολο.

**Ανάλυση scalability**

Η κατασκευή μονομεταβλητών δέντρων απόφασης παρουσιάζει ορισμένα πλεονεκτήματα και εφαρμογές στην εξόρυξη δεδομένων. Καθώς λαμβάνουν υπόψη μόνο ένα χαρακτηριστικό, είναι πιο ευαίσθητα στην ανάλυση της σχέσης μεταξύ του χαρακτηριστικού αυτού και της κλάσης που προσπαθούμε να προβλέψουμε. Η χρονική πολυπλοκότητα της κατασκευής ενός μονομεταβλητού δέντρου απόφασης είναι συνήθως λιγότερο απαιτητική σε σύγκριση με πολυμεταβλητούς ταξινομητές, επιτρέποντας την ανάπτυξη γρήγορα και αποδοτικά, ειδικά σε μεγάλα σύνολα δεδομένων.Η αποτελεσματική ανάπτυξη μονομεταβλητών δέντρων απόφασης μπορεί να επιτευχθεί μέσω διαφόρων τεχνικών, συμπεριλαμβανομένης της παραλληλοποίησης αλγορίθμων. Η παραλληλοποίηση επιτρέπει την αποδοτική εκτέλεση των αλγορίθμων σε πολυπύρηνους ή κατανεμημένους υπολογιστικούς πόρους, μειώνοντας σημαντικά τον χρόνο που απαιτείται για την εκπαίδευση του μοντέλου. Αυτό είναι ιδιαίτερα χρήσιμο σε περιβάλλοντα εξόρυξης δεδομένων με μεγάλα σύνολα δεδομένων, όπου η ταχύτητα εκπαίδευσης μπορεί να είναι κρίσιμη για την αποτελεσματική ανάλυση και τη λήψη αποφάσεων..

Οι Olcay και Onur[32] προτείνουν τρεις διαφορετικές μεθόδους για δίπλα-δίπλα εφαρμογές του αλγορίθμου C4.5:

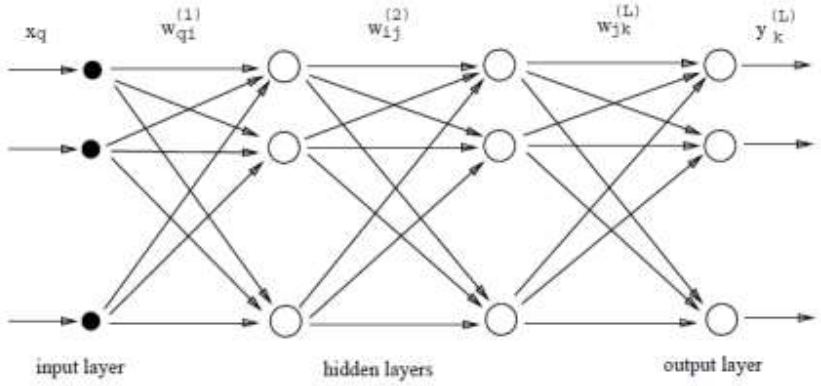
- Παραλληλοποίηση με βάση τα χαρακτηριστικά.
- Παραλληλοποίηση με κόμβους.
- Παράλληλη παραμετροποίηση δεδομένων.

Οι ερευνητές έχουν συμπεράνει ότι η παραλληλοποίηση βασισμένη στα χαρακτηριστικά μπορεί να επιταχύνει σημαντικά την εκπαίδευση σε σύνολα δεδομένων με υψηλή παραμετροποίηση ή μεγάλο μέγεθος, καθώς και σε σύνολα δεδομένων που περιέχουν δέντρα με πολλούς κόμβους. Ωστόσο, είναι ζωτικής σημασίας να υπάρχει ισορροπημένη κατανομή του φορτίου εργασίας μεταξύ των διαθέσιμων επεξεργαστών. Οι προσομοιώσεις έχουν δείξει ότι η αποτελεσματικότητα και η επιτάχυνση που επιτυγχάνεται εξαρτώνται σημαντικά από την αποτελεσματική κατανομή του φόρτου εργασίας μεταξύ των διαθέσιμων επεξεργαστών.

## 3.10 Artificial Neural Networks

Τα τελευταία χρόνια, τα νευρωνικά δίκτυα έχουν αποκτήσει σημαντική δημοτικότητα και πλέον θεωρούνται ένας από τους πιο ισχυρούς αλγόριθμους μηχανικής μάθησης που υπάρχουν. Έχουν αποδειχθεί ότι ξεπερνούν άλλους αλγόριθμους, όπως το Support Vector Machines σε διάφορες εφαρμογές, ιδιαίτερα σε εργασίες που σχετίζονται με την αναγνώριση προτύπων. Ένα νευρωνικό δίκτυο είναι ένα πολύπλοκο σύστημα που αποτελείται από μεμονωμένες μονάδες που ονομάζονται νευρώνες. Αυτά τα συστήματα οργανώνονται συνήθως σε τρία κύρια επίπεδα: το επίπεδο εισόδου, το οποίο λαμβάνει τα δεδομένα εισόδου. το επίπεδο εξόδου, το οποίο παράγει την τελική έξοδο του δικτύου. και το κρυφό στρώμα, το οποίο συνδέει νευρώνες από τα στρώματα εισόδου και εξόδου, επιτρέποντας τη διεξαγωγή πολύπλοκων υπολογισμών. [43]

Ένα παράδειγμα ενός νευρικού δικτύου απεικονίζεται στην εικόνα 10.




**Εικόνα 10 Τεχνητό νευρωνικό δίκτυο με τροφοδοσία προς τα εμπρός, επιτρέπει μόνο τα σήματα να ταξιδεύουν προς τα εμπρός από την είσοδο στην έξοδο**

**Πηγή: https://eclass-cybele.cce.uoa.gr/courses/CCEMECH181/**

Αυτό το τεχνητό νευρωνικό δίκτυο τροφοδοσίας επιτρέπει μόνο τη ροή πληροφοριών από την είσοδο στην έξοδο. Αποτελείται από τρία κύρια στοιχεία: τη δομή του δικτύου, τις λειτουργίες εισόδου και ενεργοποίησης και τα βάρη που έχουν εκχωρηθεί στις συνδέσεις εισόδου. Αυτά τα στοιχεία επιλέγονται στην αρχή και παραμένουν σταθερά κατά τη φάση της προπόνησης. Η αποτελεσματικότητα του νευρωνικού δικτύου επηρεάζεται σε μεγάλο βαθμό από τις τιμές των βαρών, τα οποία προσαρμόζονται κατά τη διάρκεια της προπόνησης για τη βελτίωση της απόδοσης. Διάφορες μέθοδοι εκπαίδευσης, όπως η Back Propagation, μπορούν να χρησιμοποιηθούν για την εκπαίδευση του νευρωνικού δικτύου. Άλλες τεχνικές, όπως ο αλγόριθμος εξάλειψης βάρους, προσαρμόζουν αυτόματα την τοπολογία του δικτύου, ενώ οι αλγόριθμοι χρησιμοποιούνται για τον προσδιορισμό της δομής του δικτύου και την εκπαίδευση των βαρών.

**Τοποθέτηση Νευρωνικών Δικτύων**

Το μοντέλο του νευρωνικού δικτύου περιλαμβάνει παραμέτρους που είναι άγνωστες αρχικά, γνωστές ως βάρη, και ο στόχος είναι να βρεθούν τιμές γι' αυτές που επιτρέπουν στο μοντέλο να προσαρμοστεί καλά στα δεδομένα εκπαίδευσης. [39]

Δηλώνουμε το σύνολο των βαρών με θ, που αποτελείται από

$$\{\alpha_{0m}, \alpha_m;\ m = 1, 2, \ldots, M\}\ \ M(p+1) \text{ weights},$$
$$\{\beta_{0k}, \beta_k;\ k = 1, 2, \ldots, K\}\ \ K(M+1) \text{ weights}.$$

**Εξίσωση 7**

$$\{\alpha_{0m}, \alpha_m;\ m = 1, 2, \ldots, M\}\ \ M(p+1) \text{ weights},$$
$$\{\beta_{0k}, \beta_k;\ k = 1, 2, \ldots, K\}\ \ K(M+1) \text{ weights}.$$

**Εξίσωση 8**

Για την ταξινόμηση χρησιμοποιούμε είτε το τετραγωνικό λάθος είτε τη διασταυρούμενη εντροπία (απόκλιση):

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2.$$

**Εξίσωση 9**

Για λόγους παλινδρόμησης,για την ταξινόμηση χρησιμοποιούμε είτε τετραγωνικό σφάλμα είτε διασταυρούμενη εντροπία (απόκλιση):

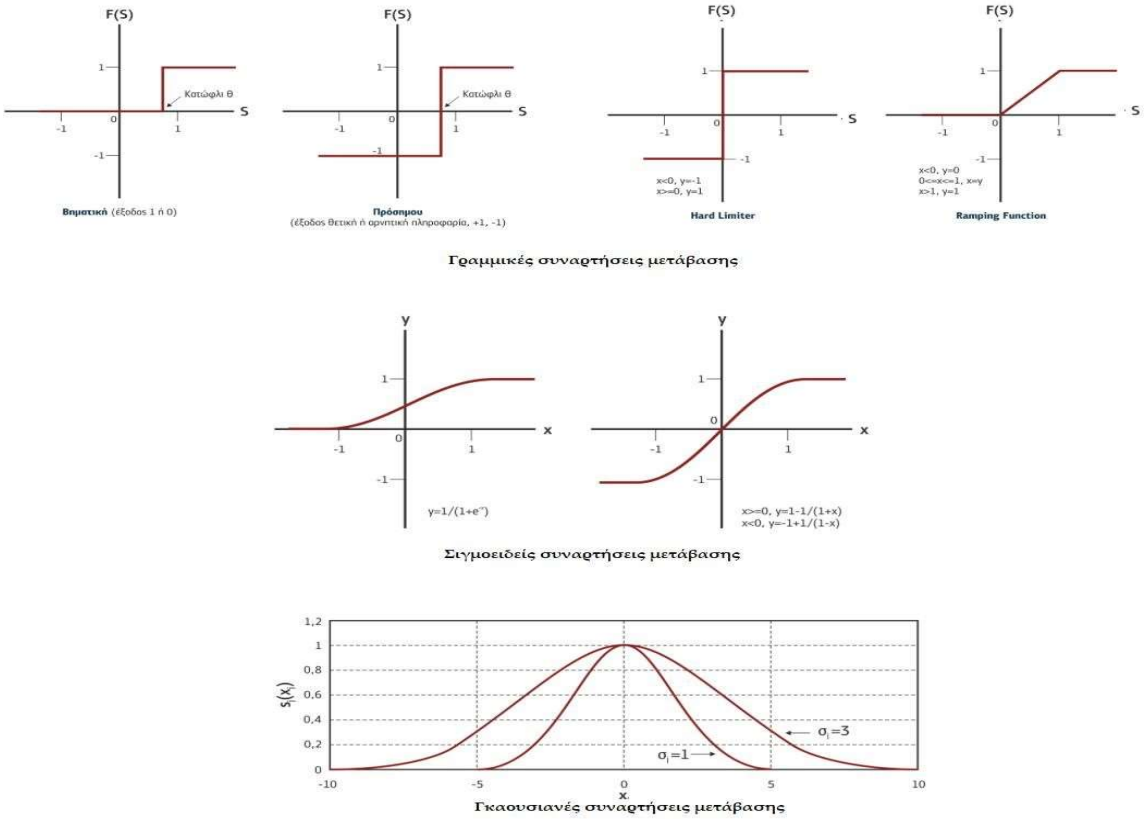
$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log f_k(x_i),$$

**Εξίσωση 10**

και ο αντίστοιχος ταξινομητής είναιG (x) = rgmaxkfk(x).

Χρησιμοποιώντας την ενεργοποίηση softmax και τη συνάρτηση καταλληλότητας cross-entropy, το νευρωνικό δίκτυο γίνεται ουσιαστικά ένα γραμμικό μοντέλο παλινδρόμησης στα κρυφά επίπεδα, με όλες τις παραμέτρους να εκτιμώνται μέσω της μεγιστοποίησης της πιθανότητας.[42] Γενικά προτιμάται να μην στοχεύουμε στον καθολικό ελαχιστοποιητή R (θ) καθώς μπορεί να οδηγήσει σε υπερπροσαρμογή. Αντίθετα, είναι απαραίτητο να εισαχθεί κάποια μορφή τακτοποίησης, είτε μέσω ποινής είτε μέσω πρόωρου τερματισμού, για να εξασφαλιστεί μια πιο αξιόπιστη και γενικεύσιμη λύση. Η επόμενη ενότητα παρέχει περισσότερες πληροφορίες σχετικά με τις λεπτομέρειες σχετικά με τον τρόπο ελαχιστοποίησης του R (θ). Μια κοινή μέθοδος για να επιτευχθεί αυτό είναι μέσω της παραμετροποίησης της κλίσης, η οποία αναφέρεται ως ανάδραση-διάδοση σε αυτό το πλαίσιο. Επειδή το μοντέλο είναι αρκετά περίπλοκο, ο προσδιορισμός της κλίσης μπορεί να γίνει με σχετική ευκολία εφαρμόζοντας τον κανόνα της αλυσίδας για διαφοροποίηση. Αυτή η διαδικασία περιλαμβάνει τη διεξαγωγή μιας σάρωσης προς τα εμπρός και προς τα πίσω στο δίκτυο, εστιάζοντας στην παρακολούθηση μόνο των συγκεκριμένων ποσοτήτων που σχετίζονται με κάθε μεμονωμένη μονάδα.

Οι γραμμικές συναρτήσεις μετάβασης, όπως οι βηματικές, οι συναρτήσεις κατωφλίου και οι συναρτήσεις προσήμου, είναι απλές συναρτήσεις που μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση γραμμικών σχέσεων μεταξύ εισόδου και εξόδου σε νευρωνικά δίκτυα. Παρόλα αυτά, αυτές οι απλές συναρτήσεις μετάβασης έχουν περιορισμένη ικανότητα προσαρμογής σε πιο πολύπλοκα προβλήματα, καθώς δεν μπορούν να αναπαραστήσουν πλήρως τη μη γραμμική φύση των περισσότερων πραγματικών δεδομένων.Οι μη γραμμικές συναρτήσεις, όπως οι σιγμοειδείς και οι γκαουσιανές συναρτήσεις, είναι πιο πολύπλοκες και ευέλικτες συναρτήσεις μετάβασης που επιτρέπουν στα νευρωνικά δίκτυα να μάθουν πιο πολύπλοκες σχέσεις μεταξύ εισόδου και εξόδου. Οι σιγμοειδείς συναρτήσεις, για παράδειγμα, προσφέρουν μια ομαλή μετάβαση από την είσοδο στην έξοδο και είναι πολύ δημοφιλείς σε νευρωνικά δίκτυα, ειδικά στα προβλήματα ταξινόμησης όπου χρειάζεται να εκτελείται πιθανοτική ταξινόμηση. Οι γκαουσιανές συναρτήσεις είναι ιδιαίτερα χρήσιμες σε προβλήματα παλινδρόμησης και μπορούν να προσαρμοστούν σε μη γραμμικές κατανομές των δεδομένων με περισσότερη ευελιξία από τις γραμμικές συναρτήσεις.

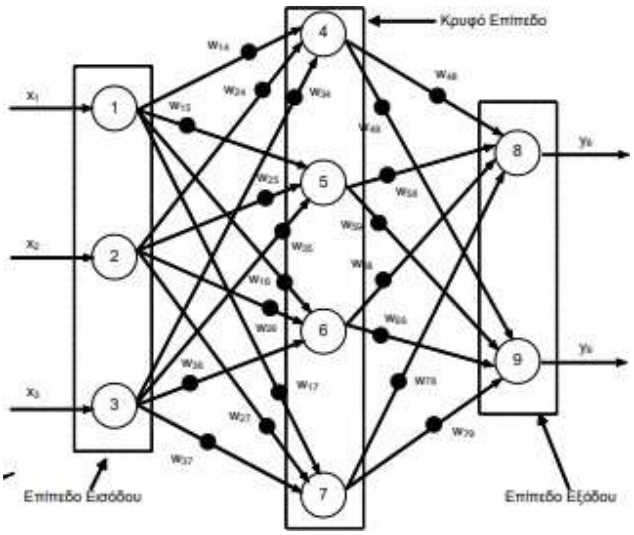



**Εικόνα 11 Γραφική αναπαράσταση των συχνότερα χρησιμοποιούμενων συναρτήσεων μεταφοράς**

**Πηγή: http://repfiles.kallipos.gr/html_books/93/index.html**

Ένα νευρωνικό δίκτυο δημιουργείται συνδέοντας πολλούς τεχνητούς νευρώνες μέσω συναπτικών συνδέσεων και συναρτήσεων ενεργοποίησης. Είναι δομημένο σε τρία επίπεδα: το στρώμα εισόδου, το στρώμα εξόδου και το κρυφό στρώμα, το οποίο δεν είναι άμεσα ορατό από τα άλλα επίπεδα. Κάθε στρώμα περιέχει διαφορετικούς τύπους νευρώνων - νευρώνες εισόδου, νευρώνες εξόδου και κρυφούς νευρώνες - με τα σήματα εξόδου ενός στρώματος να χρησιμεύουν ως είσοδος για το επόμενο. [45]Είναι σημαντικό να σημειωθεί ότι τα νευρωνικά δίκτυα μπορούν να έχουν περισσότερα από ένα κρυφά επίπεδα, καθώς είναι δυνατά πολλαπλά κρυφά επίπεδα.

Τα νευρωνικά δίκτυα διαχωρίζονται με βάση δύο βασικά κριτήρια: τη συνδεσιμότητα των νευρώνων και την κατεύθυνση της ροής πληροφοριών. [47]Όσον αφορά τη συνδεσιμότητα, τα νευρωνικά δίκτυα μπορούν να κατηγοριοποιηθούν ως πλήρως συνδεδεμένα, όπου όλοι οι νευρώνες σε ένα επίπεδο συνδέονται με όλους τους νευρώνες στο επόμενο επίπεδο ή μερικώς συνδεδεμένοι, όπου οι συνδέσεις είναι πιο επιλεκτικές.




***Εικόνα 12 Τεχνητό Νευρωνικό Δίκτυο απλής τροφοδότησης 3-4-2 Πλήρως συνδεδεμένο***

***Πηγή: Ιώαννης Βλαχάβας, Τεχνητή Νοημοσύνη, Τεχνητά Νευρωνικά Δίκτυα, Τμήμα Πληροφορικής ΑΠΘ, Θεσσαλονίκη 2013***

Οι νευρωνικές αρχιτεκτονικές χωρίζονται συνήθως σε δύο βασικές κατηγορίες: τα δίκτυα προώθησης και τα δίκτυα με ανατροφοδότηση.

- Δίκτυα Προώθησης (Feedforward): Σε αυτά τα δίκτυα, η πληροφορία κινείται μόνο προς τα εμπρός, από τα επίπεδα εισόδου προς τα επίπεδα εξόδου, χωρίς επαναλήψεις. Δεν υπάρχουν συνδέσεις που να επιστρέφουν προς προηγούμενα επίπεδα. Αυτό καθιστά την εκπαίδευση και τη χρήση τους πιο απλές και κατανοητές.
- Δίκτυα με Ανατροφοδότηση (Feedback ή Recurrent): Εδώ, οι νευρώνες ενός επιπέδου μπορούν να συνδέονται όχι μόνο με νευρώνες προηγούμενου επιπέδου, αλλά και με άλλους νευρώνες του ίδιου επιπέδου ή ακόμη και με προηγούμενα επίπεδα. Αυτό επιτρέπει στην πληροφορία να ρέει προς τα πίσω και για τις συνδέσεις να υπάρχουν κύκλοι. Τα δίκτυα με ανατροφοδότηση είναι πιο κοντά στην αναπαράσταση της πολυπλοκότητας των βιολογικών νευρωνικών δικτύων, αλλά η εκπαίδευσή τους είναι πιο πολύπλοκη και δύσκολη.

Το στάδιο της εκπαίδευσης σε ένα νευρωνικό δίκτυο συνήθως περιλαμβάνει την προσαρμογή των βαρών του δικτύου, έτσι ώστε για ένα συγκεκριμένο είσοδο να παράγεται μια επιθυμητή έξοδο. Αυτό μπορεί να γίνει με διάφορους αλγορίθμους μάθησης, όπως οι αλγόριθμοι επιβλεπόμενης μάθησης (όπου παρέχονται τόσο εισόδους όσο και επιθυμητές εξόδους), αλλά και οι αλγόριθμοι μη επιβλεπόμενης μάθησης (όπου το δίκτυο πρέπει να αντιληφθεί τη δομή των δεδομένων χωρίς επιθυμητές εξόδους).

Αφού ολοκληρωθεί η εκπαίδευση, ακολουθεί η φάση της ανάκλησης, όπου το δίκτυο δοκιμάζεται με νέα δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Κατά την ανάκληση, η απόδοση του δικτύου εκτιμάται μέσω της σύγκρισης των προβλεπόμενων εξόδων με τις πραγματικές τιμές εξόδου.

### 3.10.1 Θέματα στην Εκπαίδευση Νευρωνικών Δικτύων

**Σωστό μέγεθος κρυμμένου στρώματος**

Από την άλλη πλευρά, ένα σημαντικό πλεονέκτημα της χρήσης τεχνητών νευρωνικών δικτύων είναι η ικανότητά τους να χειρίζονται δεδομένα με πολύπλοκα και πολυδιάστατα χαρακτηριστικά, όπως εικόνες. Μία από τις κύριες προκλήσεις στην υλοποίηση των νευρωνικών δικτύων είναι ο καθορισμός του κατάλληλου μεγέθους για το κρυφό στρώμα. Εάν ο αριθμός των νευρώνων δεν προσδιορίζεται με ακρίβεια, το προκύπτον σύστημα μπορεί να μην λειτουργεί αποτελεσματικά σε καταστάσεις που δεν αναμενόταν. Αντίθετα, εάν χρησιμοποιούνται πάρα πολλοί κόμβοι, υπάρχει κίνδυνος υπερφόρτωσης και μπορεί να καταστεί αδύνατο να βρεθεί η επιθυμητή βέλτιστη τιμή. Τα πιο σοβαρά μειονεκτήματα των τεχνητών νευρωνικών δικτύων περιλαμβάνουν το υψηλό κόστος υπολογιστής ισχύος που απαιτούν, τη μεγάλη απαίτηση σε επεξεργαστική ισχύ και φυσική μνήμη, καθώς και τη δυσκολία κατανόησής τους για τον μέσο χρήστη της μηχανικής μάθησης. [6] [9].

**Αρχικές τιμές**

Συγκεκριμένα, όταν τα βάρη είναι κοντά στο μηδέν, η λειτουργική όψη του σιγμοειδούς (Εικόνα 13) συμπεριφέρεται με περίπου γραμμικό τρόπο, με αποτέλεσμα το νευρωνικό δίκτυο να γίνει ουσιαστικά ένα γραμμικό μοντέλο. Συνήθως, τυχαίες τιμές κοντά στο μηδέν επιλέγονται ως αρχικά βάρη. Ως αποτέλεσμα, η προετοιμασία του μοντέλου είναι κυρίως γραμμική αλλά γίνεται μη γραμμική καθώς αυξάνεται το βάρος. Προκειμένου να εισαχθούν μη γραμμικότητες όπου χρειάζεται, τα μηδενικά πλέγματα τοποθετούνται σε συγκεκριμένες κατευθύνσεις. Η χρήση βαρών που είναι ακριβώς μηδενικά έχει ως αποτέλεσμα μηδενικές παραγώγους και τέλεια συμμετρία, αποτρέποντας τελικά οποιαδήποτε πρόοδο στον αλγόριθμο. Από την άλλη πλευρά, η έναρξη με μεγάλα βάρη συχνά αποδίδει μη ικανοποιητικές λύσεις. [47]

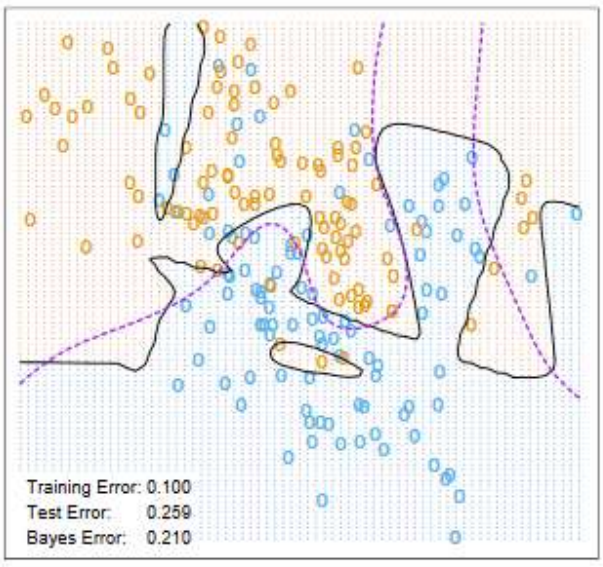
**Overfitting**

Τα νευρωνικά δίκτυα είναι συχνά επιρρεπή σε υπερπροσαρμογή λόγω των πολυάριθμων βαρών που εμπλέκονται, καθώς μπορεί να προσπαθήσουν να προσαρμόσουν τα δεδομένα πολύ κοντά στο συνολικό ελάχιστο σφάλμα. Στα αρχικά στάδια της ανάπτυξης νευρωνικών δικτύων, μια μέθοδος γνωστή ως πρόωρη διακοπή εφαρμόστηκε είτε σκόπιμα είτε ακούσια για να αποφευχθεί η υπερπροσαρμογή. Αυτή η τεχνική

περιλαμβάνει εκπαίδευση του μοντέλου για περιορισμένο χρονικό διάστημα και παύση πριν φτάσει στο απόλυτο ελάχιστο σφάλμα. Ξεκινώντας με μια βασική γραμμική λύση και χρησιμοποιώντας πρόωρη διακοπή, το τελικό μοντέλο ενθαρρύνεται να είναι πιο γραμμικό. Ο προσδιορισμός του βέλτιστου σημείου στάσης υποβοηθάται από τη χρήση δεδομένων επικύρωσης, τα οποία βοηθούν στον εντοπισμό του πότε αρχίζει να αυξάνεται το σφάλμα επικύρωσης.

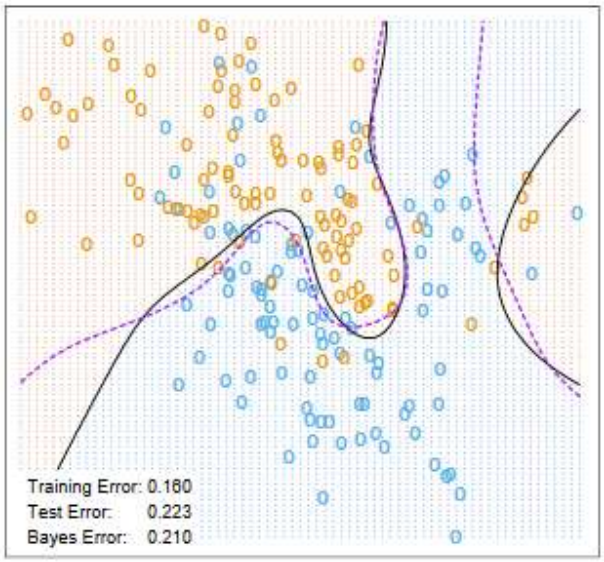
**Κλιμάκωση των εισροών**

Η ορθή κλιμάκωση των εισόδων επηρεάζει σημαντικά την αποτελεσματική κλιμάκωση των βαρών στο κάτω επίπεδο, με αποτέλεσμα να μπορεί να έχει σημαντική επίδραση στην ποιότητα του τελικού αποτελέσματος.



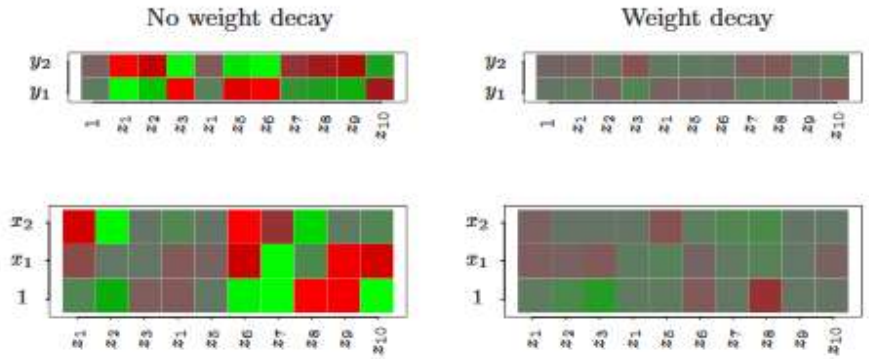

**Εικόνα 13 Νευρωνικό Δίκτυο - 10 Μονάδες, χωρίς Βάρος αποσύνθεσης**

**Πηγή : https://ikee.lib.auth.gr/record/292355/files/thesis_Fitsios_7547.pdf**




**Εικόνα 14** ***Νευρωνικό δίκτυο - 10 μονάδες, αποσύνθεση βάρους = 0,02***
**Πηγή : https://ikee.lib.auth.gr/record/292355/files/thesis_Fitsios_7547.pdf**

Το πρώτο πάνελ δεν αποσυνθέτει το βάρος και ξεπερνά τον όγκο των δεδομένων εκπαίδευσης, ενώ το δεύτερο πάνελ μειώνει το βάρος και φτάνει σε παρόμοιο επίπεδο σφάλματος με το βέλτιστο βάρος Bayes. Και τα δύο χρησιμοποιούν τη λειτουργία ενεργοποίησης softmax και τη μέθοδο σφάλματος cross-entropy. Η μείωση του βάρους είναι κατά 0,02.




**Εικόνα 15 Χάρτες θερμότητας των εκτιμώμενων βαρών από την εκπαίδευση των δικτύων νευρώνων. Η οθόνη κυμαίνεται από φωτεινό πράσινο (αρνητικό) κόκκινο (θετικό).**

Αρχικά, είναι προτιμότερο να τυποποιηθούν όλες οι εισόδοι, έτσι ώστε να έχουν μέση τιμή μηδέν και τυπική απόκλιση. Η τυποποίηση των εισόδων εξασφαλίζει την ισότιμη μεταχείριση όλων των εισόδων κατά τη διαδικασία της κανονικοποίησης. Αυτό επιτρέπει στον χρήστη να επιλέξει ένα ευρύ φάσμα για τα τυχαία αρχικά βάρη.

**Αριθμός κρυφών μονάδων και επιπέδων**

Συνήθως, είναι προτιμότερο να υπάρχει ένας επιπλέον αριθμός κρυφών μονάδων παρά να υπάρχουν πολύ λίγες. Όταν ο αριθμός των κρυφών μονάδων είναι πολύ λίγος, το μοντέλο ενδέχεται να μην έχει την

ευελιξία που απαιτείται για να ανιχνεύσει μη γραμμικά μοτίβα στα δεδομένα. Αντίθετα, με μια πληθώρα κρυφών μονάδων, τα βάρη μπορούν να ενημερωθούν αποτελεσματικά μέσω κατάλληλων τεχνικών τακτοποίησης. Συνήθως, ο αριθμός των κρυφών μονάδων κυμαίνεται μεταξύ 5 έως 100, αυξάνοντάς τον ανάλογα με τον αριθμό των εισροών και των περιπτώσεων εκπαίδευσης. Είναι συνηθισμένο να δίνουμε προτεραιότητα σε έναν σχετικά μεγαλύτερο αριθμό μονάδων και να τις εκπαιδεύουμε χρησιμοποιώντας κανονικοποίηση.. Η αναζήτηση του κατάλληλου αριθμού κρυφών επιπέδων καθοδηγείται από την εμπειρία και το πείραμα. Κάθε επίπεδο εξυπηρετεί το σκοπό της εξαγωγής χαρακτηριστικών από τα δεδομένα εισόδου για σκοπούς καταγραφής ή ταξινόμησης.

**Πολλαπλά ελάχιστα**

Η συνάρτηση σφάλματος R (θ) δεν είναι ομοιόμορφη και έχει πολλά τοπικά ελάχιστα. Κατά συνέπεια, η τελική λύση που λαμβάνεται επηρεάζεται σε μεγάλο βαθμό από τα αρχικά επιλεγμένα βάρη. Για να μετριαστεί αυτό, συνιστάται να δοκιμάσετε πολλές τυχαίες αρχικές ρυθμίσεις και να επιλέξετε τη λύση με το χαμηλότερο (τιμωρημένο) σφάλμα. Μια υποκειμενική προσέγγιση θα μπορούσε να περιλαμβάνει τη λήψη των μέσων προβλέψεων σε όλα τα δίκτυα της συλλογής ως τελική πρόβλεψη (Ripley, 1996). Αυτό είναι προτιμότερο από τον απλό υπολογισμό του μέσου όρου των βαρών, καθώς η μη γραμμική φύση του μοντέλου υποδηλώνει ότι αυτός ο μέσος όρος μπορεί να αποφέρει κακά αποτελέσματα. Μια άλλη προσέγγιση είναι το bagging, το οποίο περιλαμβάνει τον περιορισμό των προβλέψεων των δικτύων εκπαίδευσης χρησιμοποιώντας μια τυχαία κατανομή των δεδομένων εκπαίδευσης. [47]

## 3.11 Bayesian Neural Nets και το NIPS 2003 Challenge

Ο διαγωνισμός που περιγράφεται εστιάζει στην εξαγωγή χαρακτηριστικών και την κατάταξη των δεδομένων σε διάφορες κατηγορίες προβλημάτων ταξινόμησης. Η επιτυχία των Neal και Zhang φαίνεται ότι οφείλεται σε έναν συνδυασμό πολλών τεχνικών και μοντέλων μηχανικής μάθησης, συμπεριλαμβανομένων νευρωνικών δικτύων Bayes.Τα νευρωνικά δίκτυα Bayes είναι μια κατηγορία μοντέλων μηχανικής μάθησης που συνδυάζουν την πληροφορία από το Bayes' theorem με την αρχιτεκτονική των νευρωνικών δικτύων. Αυτό επιτρέπει στα δίκτυα να αντιμετωπίσουν αποδοτικά προβλήματα αναγνώρισης προτύπων και ταξινόμησης δεδομένων. Η επιτυχία των Neal και Zhang μπορεί να οφείλεται στην καλή προσαρμογή των νευρωνικών δικτύων Bayes στο συγκεκριμένο πρόβλημα, καθώς και στη συνδυαστική χρήση με άλλες τεχνικές επεξεργασίας δεδομένων.Συνοψίζοντας, η επιτυχία των Neal και Zhang στον διαγωνισμό φαίνεται να οφείλεται σε μια συνεκτική προσέγγιση που συνδυάζει νευρωνικά δίκτυα Bayes με άλλες τεχνικές επεξεργασίας δεδομένων, όπως η εξαγωγή χαρακτηριστικών και η ανίχνευση ανιχνευτών θορύβου, για να επιτύχουν καλύτερη απόδοση στον διαγωνισμό.

| Dataset | Domain | Feature Type | p | Percent Probes | $N_{tr}$ | $N_{val}$ | $N_{te}$ |
|---|---|---|---|---|---|---|---|
| Arcene | Mass spectrometry | Dense | 10,000 | 30 | 100 | 100 | 700 |
| Dexter | Text classification | Sparse | 20,000 | 50 | 300 | 300 | 2000 |
| Dorothea | Drug discovery | Sparse | 100,000 | 50 | 800 | 350 | 800 |
| Gisette | Digit recognition | Dense | 5000 | 30 | 6000 | 1000 | 6500 |
| Madelon | Artificial | Dense | 500 | 96 | 2000 | 600 | 1800 |

**Πίνακας 4**

Τα σύνολα δεδομένων της πρόκλησης NIPS 2003 περιέχουν διάφορα σύνολα δεδομένων, με κάθε σύνολο δεδομένων να έχει διαφορετικό αριθμό χαρακτηριστικών που αντιπροσωπεύονται στη στήλη p. Στο σύνολο δεδομένων Dorothea, τα χαρακτηριστικά είναι δυαδικά. Επιπλέον, το σύνολο δεδομένων περιλαμβάνει πληροφορίες σχετικά με τον αριθμό των δεδομένων εκπαίδευσης (Ntr), τα δεδομένα επικύρωσης (Nval) και τα δεδομένα δοκιμής (Nte) που είναι διαθέσιμα για ανάλυση.

Οι Neal και Zhang επιχείρησαν επίσης διάφορες μορφές προ-επεξεργασίας των χαρακτηριστικών[51]:

1. μονομεταβλητή διαλογή χρησιμοποιώντας δοκιμές t, και
2. αυτόματο προσδιορισμό συνάφειας.

Σαν τελευταία μέθοδο χρησιμοποίησαν τη μέθοδο ARD.Η μέθοδος ARD (Automatic Relevance Determination) προσφέρει μια ενδιαφέρουσα προσέγγιση στο πρόβλημα της εκμάθησης και αξιολόγησης χαρακτηριστικών σε μοντέλα μηχανικής μάθησης. Υπάρχουν τρία βασικά χαρακτηριστικά αυτής της προσέγγισης που είναι σημαντικά για την επιτυχία της:

- Κοινή Προηγούμενη Μεταβλητότητα (Common Prior Variance): Η ιδέα ότι τα βάρη των χαρακτηριστικών μοιράζονται μια κοινή προηγούμενη μεταβλητότητα συμβάλλει στην απλοποίηση του μοντέλου και τη μείωση της πολυπλοκότητας. Αυτό βοηθά στην αποφυγή της υπερπροσαρμογής και της εξασφάλισης ενός πιο γενικού μοντέλου που μπορεί να γενικεύει καλύτερα σε νέα δεδομένα.
- Εκ των Υστέρων Κατανομές και Απόρριψη Χαρακτηριστικών: Οι εκ των υστέρων κατανομές για τις παραμέτρους μπορούν να βοηθήσουν στον προσδιορισμό της σημαντικότητας των χαρακτηριστικών. Χαρακτηριστικά με χαμηλή οπίσθια διακύμανση είναι πιθανό να απορρίπτονται, καθιστώντας το μοντέλο πιο αποδοτικό και εστιάζοντας στα πιο σημαντικά χαρακτηριστικά.
- Ενσωματωμένη Εξαγωγή Χαρακτηριστικών: Η δυνατότητα του μοντέλου να αποφασίζει αυτόματα ποια χαρακτηριστικά είναι σημαντικά και ποια μπορούν να παραλειφθούν ενισχύει την ευελιξία του. Αυτό επιτρέπει στο μοντέλο να προσαρμοστεί καλύτερα στην πολυπλοκότητα των δεδομένων και να εξάγει πιο ευαίσθητες και αξιόπιστες προβλέψεις.

Η μέθοδος ARD παρέχει ένα πλαίσιο εργασίας που διευκολύνει την αποτελεσματική διαχείριση της πολυπλοκότητας των μοντέλων μηχανικής μάθησης και την αξιολόγηση της σημασίας των χαρακτηριστικών. Αυτό προωθεί την ανάπτυξη πιο αποτελεσματικών και γενικευμένων μοντέλων. Σύμφωνα με τους Neal και Zhang, η κύρια πρόκληση που αντιμετωπίζουν οι ερευνητές στον έλεγχο των χαρακτηριστικών είναι η βελτίωση της υπολογιστικής αποτελεσματικότητας. Όταν έχουν να κάνουν με μεγάλο αριθμό λειτουργιών, η διαδικασία MCMC (Markov Chain Monte Carlo) τείνει να επιβραδύνεται. Ωστόσο, αυτό δεν σημαίνει ότι πρέπει να εγκαταλείψουν την επιλογή των χαρακτηριστικών. Ο βασικός τους στόχος είναι να κατανοήσουν γιατί η μέθοδος Bayes ήταν επιτυχής. Οι Neal και Zhang υποστηρίζουν ότι η δύναμη των σύγχρονων μεθόδων Bayes δεν βρίσκεται στην αντιμετώπισή τους ως εργαλεία επίσημης διαδικασίας εξαγωγής συμπερασμάτων.[51] Είναι απίθανο να περιμένει κανείς ακρίβεια ασφαλίστρων σε ένα πολυδιάστατο, πολύπλοκο μοντέλο νευρωνικών δικτύων. Η προσέγγιση Bayesian/MCMC παρέχει μια αποτελεσματική τεχνική δειγματοληψίας τμημάτων του μοντέλου και την εκτίμηση του μέσου όρου των προβλέψεων για μοντέλα υψηλής πιθανότητας. Για παράδειγμα, στον τομέα της ρομποτικής, μπορούμε να χρησιμοποιήσουμε την προσέγγιση Bayesian/MCMC για να εκτιμήσουμε την πορεία ενός ρομπότ σε ένα περιβάλλον βάσει της τρέχουσας κατανομής θέσης του.

- Οι καμπύλες και οι προσαυξήσεις είναι μη-Bayesian διαδικασίες που παρουσιάζουν παρόμοια επίπεδα αξιοπιστίας με το MCMC σε ένα μπεϋζιανό μοντέλο. Για παράδειγμα, στην ανάλυση δεδομένων, μπορούμε να χρησιμοποιήσουμε καμπύλες και προσαυξήσεις για να εκτιμήσουμε την πρόβλεψη των τιμών ακινήτων με βάση την προηγούμενη κίνηση της αγοράς.
- Η Bayesian προσέγγιση προσαρμόζει τα δεδομένα και διαταράσσει τις παραμέτρους με βάση τις τρέχουσες εκτιμήσεις της κατανομής θέσης. Για παράδειγμα, στον τομέα της αναγνώρισης φωνής, μπορούμε να χρησιμοποιήσουμε Bayesian προσέγγιση για να βελτιώσουμε την ακρίβεια της αναγνώρισης λόγου με βάση τις προηγούμενες προβλέψεις.
- Το bagging διαταράσσει τα δεδομένα ακολουθώντας ένα ανεξάρτητο και ομοιόμορφα κατανεμημένο (i.d.) μοτίβο και στη συνέχεια εκτιμά το μοντέλο για να αποκτήσει ένα νέο σύνολο παραμέτρων μοντέλου. Για παράδειγμα, στον τομέα της αναγνώρισης εικόνας, μπορούμε να χρησιμοποιήσουμε το bagging για να βελτιώσουμε την ακρίβεια της αναγνώρισης αντικειμένων μέσω της συλλογής πολλών μοντέλων και της εκπαίδευσής τους σε διαφορετικά δείγματα.
- Το Boosting εφαρμόζεται σε ένα μοντέλο που είναι προσθετικό στα μοντέλα κάθε μεμονωμένου εκπαιδευτικού σταδίου, τα οποία αποκτώνται χρησιμοποιώντας μη-i.i.d. δείγματα. Για παράδειγμα, στον τομέα της ανάλυσης χρηματιστηριακών δεδομένων, μπορούμε να χρησιμοποιήσουμε το

Boosting για να αναπτύξουμε ένα μοντέλο πρόβλεψης των τιμών των μετοχών, λαμβάνοντας υπόψη την επίδραση παραγόντων που δεν ακολουθούν ένα συγκεκριμένο μοτίβο.

**Ανάλυση κλιμάκωσης**

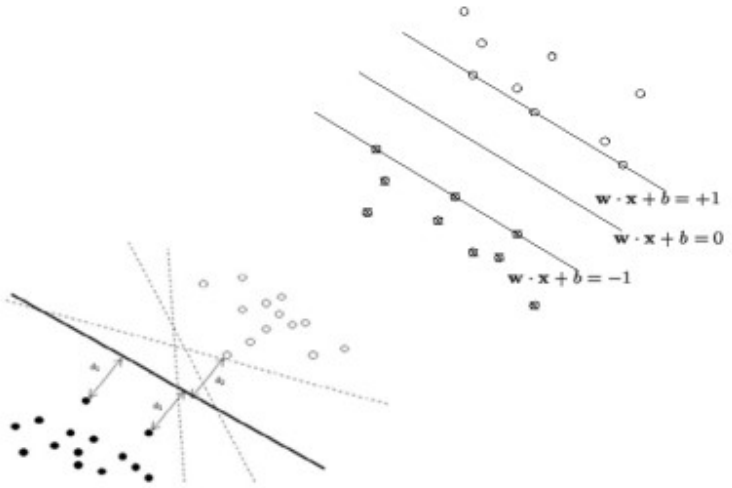
Για να συζητήσουμε την επεκτασιμότητα των νευρωνικών δικτύων χρησιμοποιείται ο αλγόριθμος backpropagation. Σύμφωνα με την περιγραφή του Kotsiantis [6], ο αλγόριθμος backpropagation περιλαμβάνει συνήθως τα ακόλουθα έξι βήματα:

- Δίνεται ένα παράδειγμα ως είσοδος στο δίκτυο.
- Η έξοδος του δικτύου συγκρίνεται με το επιθυμητό αποτέλεσμα για κάθε παράδειγμα, και υπολογίζεται το σφάλμα για κάθε νευρώνα.
- Υπολογίζεται το τοπικό σφάλμα για κάθε νευρώνα.
- Τροποποιούνται τα βάρη για τη μείωση του τοπικού σφάλματος.
- Το σφάλμα προωθείται προς τα πίσω, δίνοντας έμφαση στους νευρώνες που είναι πιο ευαίσθητοι.
- Η διαδικασία επαναλαμβάνεται μέχρι να χρησιμοποιηθεί το σφάλμα ως λανθασμένη πρόβλεψη.

Η διαδικασία επίτευξης της βέλτιστης διάταξης βαρών σε ένα νευρωνικό δίκτυο μπορεί να είναι χρονοβόρα και απαιτητική υπολογιστικά, ανάλογα με το μέγεθος του συνόλου εκπαίδευσης και την πολυπλοκότητα του δικτύου. Η χρονική πολυπλοκότητα είναι συνήθως ανάλογη του πλήθους των περιπτώσεων εκπαίδευσης και των βαρών που πρέπει να προσαρμοστούν.Για την εκπαίδευση ενός νευρωνικού δικτύου, συνήθως χρειάζεται μεγάλος όγκος δεδομένων. Η ποσότητα αυτή εξαρτάται από το ποσοστό σφαλμάτων που είναι αποδεκτό (ε) και εκφράζεται ως T = O(N/ε), όπου T είναι το μέγεθος του συνόλου εκπαίδευσης και N ο αριθμός των δειγμάτων εκπαίδευσης. Επίσης, η χρήση μεγάλων παράλληλων υπολογιστών μπορεί να επιταχύνει την εκπαίδευση, καθώς η feedforward λειτουργία είναι O(N), επιτρέποντας την ταυτόχρονη επεξεργασία δεδομένων σε πολλαπλούς πυρήνες. Εν συνεχεία, τα βάρη μπορούν να αντιγραφούν σε σειριακούς υπολογιστές για την εφαρμογή του μοντέλου. Ωστόσο, η ακριβής ποσότητα εκπαίδευσης που απαιτείται εξακολουθεί να εξαρτάται από τα χαρακτηριστικά του προβλήματος και την πολυπλοκότητα του μοντέλου [33].

## 3.12 Support Vector Machines

Τα μηχανήματα διανυσμάτων υποστήριξης (SVM) είναι ένας είδος αλγορίθμου ταξινόμησης που βασίζεται στην αρχή της ελαχιστοποίησης του κινδύνου σφάλματος. Αυτή η τεχνική πηγάζει από τη θεωρία της υπολογιστικής μάθησης και στοχεύει στο να εντοπίσει τον βέλτιστο τρόπο διαχωρισμού των διαφορετικών κατηγοριών εντός ενός συνόλου δεδομένων εκπαίδευσης. Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, το SVM λειτουργεί εντοπίζοντας ένα υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κατηγοριών, δημιουργώντας τη μεγαλύτερη δυνατή απόσταση μεταξύ τους. Αυτή η προσέγγιση απεικονίζεται οπτικά στο Σχήμα 16.




**Εικόνα 16 Δημιουργία του βέλτιστου διαχωριστικού υπερπλάνου με τη χρήση φορέων στήριξης [9]**
**Πηγή: https://eclass-cybele.cce.uoa.gr/courses/CCEMECH181/**

Το SVM έχει σαν θεμέλιο του την αναζήτηση του μέγιστου περιθωρίου και του βέλτιστου υπερεπιπέδου, το οποίο οδηγεί σε βελτιωμένη γενίκευση. Γι' αυτό, τόσο η ταξινόμηση των δεδομένων εκπαίδευσης όσο και οι προβλέψεις για τα μελλοντικά δεδομένα βελτιώνονται. Για να βρεθεί το μέγιστο περιθώριο υπερφόρτωσης, το SVM αποσκοπεί στη μεγιστοποίηση της συνάρτησης υποστήριξης σχετικά με τα βάρη (w) και τη σταθερά (b). Στη συνέχεια, τα δεδομένα που βρίσκονται κοντά στα όρια της απόφασης ορίζονται ως διανύσματα υποστήριξης, ενώ η λύση που παρέχει το SVM είναι ένας γραμμικός συνδυασμός αυτών των διανυσμάτων. Μια πιθανή δυσκολία κατά την εφαρμογή των SVM είναι η αδυναμία εύρεσης διαχωριστικού υπερεπιπέδου σε περιπτώσεις που τα δεδομένα περιέχουν σφάλματα. Γι' αυτό, σε πραγματικά προβλήματα που περιλαμβάνουν μη γραμμικά διαχωρίσιμα δεδομένα, μια λύση είναι η περιγραφή των δεδομένων σε έναν τροποποιημένο χώρο χαρακτηριστικών. Η χρήση των SVM προσφέρει πλεονεκτήματα, όπως η βάση τους σε καθιερωμένη θεωρία και η απαίτηση λίγων εκπαιδευτικών δειγμάτων. Από την πλευρά της υπολογιστικής απόδοσης, η εκπαίδευση και η ταξινόμηση μεγάλων SVM αλγορίθμων μπορεί να απαιτεί σημαντικούς υπολογιστικούς πόρους και χρόνο.[50]

**SVM και kernels**

Η διαδικασία που περιγράφηκε προηγουμένως για τον ταξινομητή Support Vector Machine (SVM) βασίζεται σε γραμμικά όρια στον χώρο των χαρακτηριστικών εισόδου. Ωστόσο, μπορούμε να επεκτείνουμε αυτήν τη διαδικασία κάνοντας την προ-επεξεργασία πιο ευέλικτη με τη διεύρυνση του χώρου των χαρακτηριστικών χρησιμοποιώντας επεκτάσεις βάσης, όπως πολυώνυμα ή σφήνες.

Η εφαρμογή γραμμικών ορίων στο διευρυμένο χώρο συνήθως οδηγεί σε καλύτερο διαχωρισμό των κλάσεων εκπαίδευσης, καθώς μεταφέρει μη γραμμικά όρια στον αρχικό χώρο. Μόλις επιλεγούν οι κατάλληλες βάσεις hm(x)hm(x), όπου m=1,...,Mm=1,...,M, η διαδικασία παραμένει παρόμοια με την προηγούμενη.

Ο αλγόριθμος SVM βασίζεται σε αυτήν την ιδέα, επιτρέποντας στον εκτεταμένο χώρο να έχει δυνητικά υψηλή διάσταση σε ορισμένα σενάρια. Μπορεί να φαίνεται ότι οι υπολογισμοί θα μπορούσαν να γίνουν συντριπτικοί και ότι ένας υπερβολικός αριθμός βασικών συναρτήσεων θα μπορούσε να οδηγήσει σε προβλήματα διαχωρισμού δεδομένων και υπερφόρτωση. Ο ταξινομητής SVM αποδίδει αποτελεσματικά στην αντιμετώπιση προβλημάτων συναρμολόγησης μέσω της χρήσης ενός συγκεκριμένου κριτηρίου και μορφής εγκατάστασης. Αυτός αποτελεί ένα κομμάτι της ευρύτερης κατηγορίας προβλημάτων που περιλαμβάνουν την εξομάλυνση γραμμών.

**Ανάλυση κλιμάκωσης**

Οι τεχνικές Support Vector Machine (SVM) περιλαμβάνουν την επίλυση ενός προβλήματος τετραγωνικού προγραμματισμού (QP), όπου ο αριθμός των περιπτώσεων εκπαίδευσης, που αντιπροσωπεύονται από το N, καθορίζει την πολυπλοκότητα. Ωστόσο, οι παραδοσιακές μέθοδοι QP που χρησιμοποιούνται για την επίλυση αυτού του προβλήματος βασίζονται σε εκτεταμένες λειτουργίες μήτρας και χρονοβόρους αριθμητικούς υπολογισμούς. Αυτή η προσέγγιση αποδεικνύεται εξαιρετικά αναποτελεσματική και μη πρακτική, ιδιαίτερα για εφαρμογές μεγάλης κλίμακας, κάτι που είναι γνωστός περιορισμός της μεθόδου SVM. Ευτυχώς, υπάρχουν εναλλακτικές μέθοδοι, όπως η Διαδοχική Ελάχιστη Βελτιστοποίηση (SMO) [35], που μπορούν να λύσουν αποτελεσματικά το πρόβλημα QP αναλύοντάς το σε υποπροβλήματα χωρίς να βασίζονται σε αριθμητικά βήματα βελτιστοποίησης QP ή πρόσθετη αποθήκευση μήτρας. Μια άλλη νέα προσέγγιση στην εκτίμηση SVM περιλαμβάνει τον προσδιορισμό του ελάχιστου περιβάλλοντος μιας ομάδας αντικειμένων [10]. Είναι σημαντικό να σημειωθεί ότι οι μέθοδοι SVM ασχολούνται κυρίως με προβλήματα δυαδικής ταξινόμησης. Στην περίπτωση προβλημάτων πολλαπλών τάξεων, η πρόκληση χωρίζεται σε πολλαπλές εργασίες ταξινόμησης. [45]

## 3.13 k-Nearest-Neighbor

Η μέθοδος ταξινόμησης k-Nearest-Neighbor, γνωστή και ως kNN, αναζητά τα k πλησιέστερα αντικείμενα σε ένα σύνολο εκπαίδευσης από ένα δεδομένο αντικείμενο που δοκιμάζουμε, και στη συνέχεια αναθέτει μια ετικέτα βάσει της κυριότητας της κλάσης μεταξύ των k πλησιέστερων γειτόνων. Για την εκτέλεση του αλγορίθμου απαιτούνται τρία βασικά στοιχεία: ένα σύνολο αντικειμένων με ετικέτες, ένα μέτρο απόστασης ή ομοιότητας, και ο αριθμός k των πλησιέστερων γειτόνων.

Ένα δημοφιλές μέτρο εγγύτητας που χρησιμοποιείται για την ταξινόμηση kNN είναι η "ευκλείδια απόσταση", η οποία ορίζεται από τον τύπο:

$$D(x,y) = (\sum_{i=1}^{m} |x_i - y_i|^2)^{1/2}$$

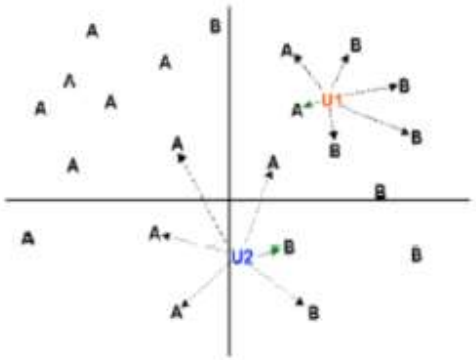
**Εξίσωση 11**

όπου xi και yi είναι τα διανύσματα χαρακτηριστικών δύο αντικειμένων, και m είναι ο αριθμός των χαρακτηριστικών.

Πέραν της ευκλείδιας απόστασης, υπάρχουν και άλλα μέτρα απόστασης που χρησιμοποιούνται για τον ορισμό της απόστασης μεταξύ των δειγμάτων σε ένα σύνολο δεδομένων. Μερικά παραδείγματα περιλαμβάνουν:

- Η μετρική Minkowski, η οποία γενικεύει την ευκλείδια απόσταση και περιλαμβάνει ως ειδικές περιπτώσεις την Ευκλείδια απόσταση (όταν το p=2p=2) και την απόλυτη απόσταση (όταν το p=1p=1).
- Η μετρική Camberra, που συνήθως χρησιμοποιείται για δεδομένα που περιλαμβάνουν μικρό αριθμό μη μηδενικών χαρακτηριστικών.
- Η μετρική Chebyshev, που ορίζει την απόσταση ως το μέγιστο απόλυτο διαφορών μεταξύ των αντίστοιχων χαρακτηριστικών των δύο δειγμάτων.

Συχνά, χρησιμοποιούνται και στρατηγικές ζύγισης για να αλλάξουν την επίδραση των γειτόνων, με σκοπό την επίτευξη πιο ακριβών αποτελεσμάτων.




**Εικόνα 17 Αναπαράσταση ενός γραφήματος k-Nearest-Neighbor.**

[10]Όταν αποκτηθεί η λίστα k-πλησιέστερων γειτόνων, τα αντικείμενα δοκιμής κατηγοριοποιούνται σύμφωνα με την τάξη πλειοψηφίας, που δίνεται στην εξίσωση 12:

$$Majority\ Voting: Y^{'} = argmax \sum_{(x_i, y_i) \in D_z} I(v = y_i)$$

**Εξίσωση 12**

**Πλεονεκτήματα και προκλήσεις**

Πράγματι, οι αλγόριθμοι k-Nearest-Neighbor έχουν αρκετά πλεονεκτήματα, όπως η ευκολία κατανόησης και η εφαρμοσιμότητά τους. Ωστόσο, όπως και με κάθε αλγόριθμο, υπάρχουν κάποια προβλήματα που μπορεί να επηρεάσουν τη συμπεριφορά τους. Συγκεκριμένα, οι Kotsiantis και συνεργάτες περιγράφουν τους ακόλουθους τρεις κύριους λόγους:

- Μεγάλες απαιτήσεις αποθήκευσης: Καθώς ο αλγόριθμος πρέπει να αποθηκεύει όλα τα επισημασμένα δείγματα, οι απαιτήσεις αποθήκευσης μπορεί να είναι σημαντικές, ειδικά για μεγάλα σύνολα δεδομένων.
- Ευαισθησία στην επιλογή της λειτουργίας ομοιότητας: Η απόδοση του αλγορίθμου εξαρτάται σε μεγάλο βαθμό από την κατάλληλη επιλογή της συνάρτησης ομοιότητας ή απόστασης μεταξύ των δεδομένων. Η επιλογή αυτή μπορεί να είναι υποκειμενική και να επηρεάζει τα αποτελέσματα.
- Λείπει μια βασική μέθοδος για την επιλογή k: Η επιλογή του κατάλληλου αριθμού k, δηλαδή των πλησιέστερων γειτόνων που θα χρησιμοποιηθούν για την ταξινόμηση, μπορεί να είναι μια δύσκολη διαδικασία και να επηρεάσει την απόδοση του αλγορίθμου.

Παρά τα παραπάνω, οι αλγόριθμοι k-Nearest-Neighbor παραμένουν αποτελεσματικοί σε πολλές περιπτώσεις, ειδικά όταν αντιμετωπίζουμε πολυτροπικές κατηγορίες..

**Υπολογιστικές εκτιμήσεις**

Οι αλγόριθμοι k-Nearest-Neighbor έχουν ως μειονέκτημα τον υπολογιστικό φόρτο που απαιτείται τόσο για την εύρεση των γειτόνων όσο και για την αποθήκευση ολόκληρου του συνόλου εκπαίδευσης. Ο υπολογιστικός φόρτος μπορεί να μειωθεί χρησιμοποιώντας γρήγορους αλγορίθμους για την εξεύρεση πλησιέστερων γειτόνων, καθώς και μειώνοντας τις απαιτήσεις αποθήκευσης.

Οι αλγόριθμοι που προτάθηκαν για τη μείωση του υπολογιστικού φόρτου περιλαμβάνουν:

- Μείωση των υπολογισμών για την εφαπτόμενη απόσταση: Μια προσέγγιση που μειώνει τους υπολογισμούς για την απόσταση μεταξύ των σημείων είναι η ανάπτυξη αναλογιών του αλγορίθμου K-means μέσα σε αυτήν τη μετρική απόσταση.
- Μείωση των απαιτήσεων αποθήκευσης: Οι προσεγγίσεις που μειώνουν τις απαιτήσεις αποθήκευσης περιλαμβάνουν διάφορες τεχνικές συμπύκνωσης του συνόλου εκπαίδευσης. Αυτό μπορεί να γίνει κρατώντας μόνο τα σημεία εκπαίδευσης που είναι κοντά στα όρια των αποφάσεων και αφαιρώντας τα υπόλοιπα.

Οι παραπάνω προσεγγίσεις ενδέχεται να μειώσουν τον υπολογιστικό φόρτο και τις απαιτήσεις αποθήκευσης του αλγορίθμου k-Nearest-Neighbor, καθιστώντας τον πιο αποτελεσματικό και ευέλικτο σε εφαρμογές μηχανικής μάθησης.

**Ανάλυση κλιμάκωσης**

Το μειονέκτημα του ταξινομητή kNN είναι ότι δεν έχει ειδική φάση εκπαίδευσης, αλλά απαιτεί να αποθηκεύει όλες τις τιμές χαρακτηριστικών στη μνήμη. Η διαδικασία κατασκευής του μοντέλου είναι υπολογιστικά ανέξοδη, αλλά η φάση ταξινόμησης μπορεί να είναι υπολογιστικά εντατική λόγω της ανάγκης να εκχωρηθούν ετικέτες σε κάθε έναν από τους k-πλησιέστερους γείτονες. Αυτό περιλαμβάνει τον υπολογισμό της μέτρησης της απόστασης για όλες τις περιπτώσεις στο σετ με ετικέτα, το οποίο μπορεί να είναι ιδιαίτερα δαπανηρό για μεγάλα σύνολα δεδομένων.Για να αντιμετωπιστεί αυτό το ζήτημα, έχουν αναπτυχθεί διάφορες μέθοδοι, όπως ο αλγόριθμος ανεστραμμένης ένωσης δείκτη και ο αλγόριθμος άπληστου φιλτραρίσματος. Ο αλγόριθμος ανεστραμμένης ένωσης δείκτη είναι ένας από τους ταχύτερους αλγορίθμους για τη δημιουργία λεπτομερών γραφημάτων kNN, αλλά έχει ασυμπτωτική χρονική πολυπλοκότητα O(n^2), ενώ ο αλγόριθμος άπληστου φιλτραρίσματος προσφέρει χρονική πολυπλοκότητα O(n) και βελτιωμένη απόδοση.

# 4. Αξιολόγηση της Μηχανικής Μάθησης

Μια άλλη κρίσιμη πτυχή της μηχανικής μάθησης περιλαμβάνει την πρόκληση του προσδιορισμού ποια αποτελέσματα που παράγονται από ένα πρόγραμμα υπολογιστή είναι ακριβή και ποια είναι ελαττωματικά. Ένα σενάριο πρόβλεψης αγοραστικής συμπεριφοράς σε μια πλατφόρμα ηλεκτρονικού εμπορίου μπορεί να περιλαμβάνει τα ακόλουθα βήματα. Το πρόγραμμα καταγράφει δεδομένα σχετικά με το εάν ο πελάτης αγόρασε πράγματι αντικείμενα ή όχι, τα οποία μπορούν να χρησιμοποιηθούν για την αξιολόγηση της αποτελεσματικότητας του αλγορίθμου. Ωστόσο, το πρόβλημα γίνεται πιο περίπλοκο σε ερευνητικά πεδία με σπάνια ή ανύπαρκτα δεδομένα από τον πραγματικό κόσμο, όπως η αξιολόγηση της ποιότητας των μεταφράσεων εγγράφων. Αυτή η εργασία απαιτεί πρόσθετη ανθρώπινη προσπάθεια για την κατηγοριοποίηση των μεταφράσεων σε συγκεκριμένες κατηγορίες προκειμένου να συγκριθούν τα αποτελέσματα που παράγονται από το ηλεκτρονικό πρόγραμμα. Όπως εξηγείται στην υποενότητα, η αξιολόγηση των εργασιών ταξινόμησης συνήθως περιλαμβάνει τη διαίρεση του συνόλου δεδομένων σε ένα σύνολο εκπαίδευσης και ένα σύνολο δοκιμής. Στη συνέχεια, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται χρησιμοποιώντας το σετ εκπαίδευσης, ενώ το σύνολο δοκιμής χρησιμοποιείται για τον υπολογισμό μετρήσεων απόδοσης που αξιολογούν την ποιότητα του αλγορίθμου. Η χρήση διασταυρούμενης επικύρωσης είναι μια σημαντική πρακτική για την αντιμετώπιση της υπερπροσαρμογής σε αλγόριθμους

μηχανικής μάθησης. Αυτή η διαδικασία επιτρέπει την αξιολόγηση της γενικότητας του μοντέλου, αντί να είναι υπερβολικά προσαρμοσμένο στα δεδομένα εκπαίδευσης.Με τη διασταυρούμενη επικύρωση, το σύνολο δεδομένων διαιρείται σε πολλαπλά υποσύνολα και εκτελούνται επανειλημμένα εκπαιδεύσεις και αξιολογήσεις χρησιμοποιώντας διαφορετικά υποσύνολα ως σετ δοκιμής και εκπαίδευσης. Αυτό επιτρέπει στο μοντέλο να εκπαιδευτεί σε διαφορετικά σετ δεδομένων και να αξιολογηθεί σε πολλές δοκιμές, προσφέροντας έτσι μια πιο αξιόπιστη εκτίμηση της απόδοσης.Όσον αφορά τις μετρήσεις απόδοσης, είναι σημαντικό να επιλεχθούν κατάλληλες μετρικές αξιολόγησης ανάλογα με το είδος του προβλήματος και τους στόχους της εφαρμογής. Για παράδειγμα, σε προβλήματα ταξινόμησης μπορούν να χρησιμοποιηθούν μετρικές όπως η ακρίβεια, η ευαισθησία (recall), η ειδικότητα (specificity), ο F1-βαθμός κ.λπ. Σε προβλήματα παλινδρόμησης, μπορούν να χρησιμοποιηθούν μετρικές όπως η μέση απόκλιση των τετραγώνων (mean squared error - MSE), ο απόλυτος μέσος όρος των αποκλίσεων (mean absolute error - MAE) κ.λπ.Συνολικά, η διασταυρούμενη επικύρωση και η σωστή επιλογή μετρικών αξιολόγησης συνιστούν σημαντικά εργαλεία για την αξιολόγηση και τη βελτίωση της απόδοσης των αλγορίθμων μηχανικής μάθησης.. Οι κύριοι παράγοντες για την αξιολόγηση της απόδοσης ενός αλγορίθμου μηχανικής μάθησης είναι οι ακόλουθοι:

- Η ακρίβεια (precision) μετρά τον αριθμό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των προβλέψεων που έγιναν σε μια κατηγορία. Για παράδειγμα, αν ένας αλγόριθμος αναγνώρισης φωνής αναγνωρίζει 90 από τις 100 φορές τη φωνή σωστά, η ακρίβεια του θα είναι 90%.
- Η ανάκληση (recall) μετρά τον αριθμό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών παραδειγμάτων σε μια κατηγορία. Αν συνολικά υπάρχουν 100 φορές που θα έπρεπε να αναγνωριστεί η φωνή, και ο αλγόριθμος αναγνωρίζει σωστά 90 από αυτές, τότε η ανάκλησή του θα είναι 90%.
- Το F1-score είναι ένας συνδυασμός της ακρίβειας και της ανάκλησης και υπολογίζεται ως ο αντίστροφος του αρμονικού μέσου των δύο. Αυτό το μέτρο λαμβάνει υπόψη τόσο τις ψευδείς θετικές όσο και τις ψευδείς αρνητικές προβλέψεις.
- Η καμπύλη ROC (Receiver Operating Characteristic) αναπαριστά τη σχέση μεταξύ της ανάκλησης και του επιπέδου των ψευδών θετικών προβλέψεων. Μια καλή καμπύλη ROC εμφανίζει μια κλίση προς τα πάνω προς την αριστερή πλευρά του γραφήματος.

Αυτοί είναι μερικοί από τους βασικούς παράγοντες που πρέπει να ληφθούν υπόψη κατά την αξιολόγηση ενός αλγορίθμου μηχανικής μάθησης.

Ένα σημαντικό ζήτημα με την εσφαλμένη ταξινόμηση είναι ότι τα αποτελέσματά της μπορεί να διαφέρουν σημαντικά ανάλογα με την κατανομή των δεδομένων μεταξύ διαφορετικών ετικετών κλάσεων. Για παράδειγμα, η επίτευξη ποσοστού εσφαλμένης ταξινόμησης 0,03 μπορεί να φαίνεται εντυπωσιακή με την πρώτη ματιά, αλλά θα μπορούσε να είναι σχετικά εύκολο να επιτευχθεί εάν το 97% του συνόλου δεδομένων χαρακτηρίζεται ως κατηγορία Α και μόνο το 3% ως κατηγορία Β. Ομοίως, ο αριθμός των κατηγοριών διαθέσιμο μπορεί επίσης να επηρεάσει το ποσοστό εσφαλμένης ταξινόμησης. Ένα σύστημα μηχανικής εκμάθησης με ποσοστό εσφαλμένης ταξινόμησης 20% ή χαμηλότερο θα θεωρηθεί πιο αποτελεσματικό για ένα σύνολο δεδομένων τριών κλάσεων σε σύγκριση με ένα σύνολο δεδομένων δύο κλάσεων.Το συγκριτικό benchmarking αποτελεί μια διαδικασία στην οποία συγκρίνουμε την απόδοση ενός δείκτη με μια τιμή αναφοράς, με στόχο να βελτιώσουμε την εκτίμησή του. Ένα παράδειγμα είναι η χρήση του για να αξιολογήσουμε την απόδοση ενός ταξινομητή σε ένα πρόβλημα δυαδικής ταξινόμησης. Μπορεί να χρησιμοποιηθεί για να εξετάσουμε την απόδοση του ταξινομητή σε σύγκριση με μια τυχαία ή βασική στρατηγική. Για παράδειγμα, μπορεί να μετρήσει το ποσοστό των λανθασμένων προβλέψεων και να συγκριθεί με ένα προκαθορισμένο ποσοστό.Η τιμή ακρίβειας (precision), που είναι επίσης γνωστή ως θετική πρόβλεψη, ορίζεται ως το ποσοστό των σωστών προβλέψεων σε σχέση με τις όλες τις προβλεπόμενες περιπτώσεις. Για παράδειγμα, μπορεί να μετρήσει το ποσοστό των πελατών που πραγματικά πραγματοποίησαν αγορά από τους πελάτες που προβλέφθηκε ότι θα πραγματοποιήσουν αγορά. Η τιμή ευαισθησίας (recall), επίσης γνωστή ως ευαισθησία ή αληθής θετικός λόγος, ορίζεται ως το ποσοστό των πραγματικών θετικών περιπτώσεων που ταξινομήθηκαν σωστά από τον ταξινομητή, δηλαδή το ποσοστό των

παρατηρήσεων που ανήκουν στη θετική κλάση και προβλέφθηκαν σωστά ως θετικές σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων. Για παράδειγμα, μπορεί να μετρήσει το ποσοστό των πελατών που πραγματικά πραγματοποίησαν αγορά ανάμεσα σε όλους τους πελάτες που έπρεπε να πραγματοποιήσουν αγορά.

Το F-Measure προσπαθεί να συνδυάσει τις τιμές ευαισθησίας και ακρίβειας, χρησιμοποιώντας το μέσο όρο ανάμεσά τους.:

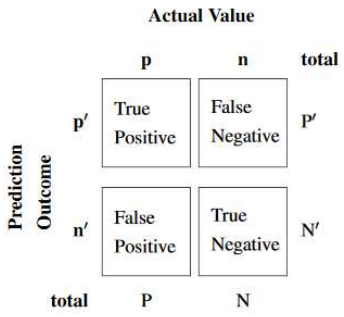
$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

**Εξίσωση 13**

Ο πίνακας σύγχυσης είναι ένα αποτελεσματικό εργαλείο για την αξιολόγηση της απόδοσης ενός αλγορίθμου μηχανικής μάθησης. Αναπαριστά τα αποτελέσματα των προβλέψεων σε ένα συνοπτικό πίνακα, όπου οι γραμμές αντιστοιχούν στις πραγματικές κλάσεις των δεδομένων και οι στήλες στις προβλέψεις του μοντέλου. Ο πίνακας αυτός διακρίνει τέσσερα διαφορετικά είδη προβλέψεων [15]:

- Αληθινά θετικά (True Positives - TP): Δείχνει τον αριθμό των δειγμάτων που προβλέφθηκαν σωστά ως θετικά.
- Ψευδή θετικά (False Positives - FP): Αντιπροσωπεύει τον αριθμό των δειγμάτων που προβλέφθηκαν λανθασμένα ως θετικά (ενώ στην πραγματικότητα είναι αρνητικά).
- Αληθινά αρνητικά (True Negatives - TN): Δείχνει τον αριθμό των δειγμάτων που προβλέφθηκαν σωστά ως αρνητικά.
- Ψευδή αρνητικά (False Negatives - FN): Αντιπροσωπεύει τον αριθμό των δειγμάτων που προβλέφθηκαν λανθασμένα ως αρνητικά (ενώ στην πραγματικότητα είναι θετικά).

Με τη χρήση του πίνακα σύγχυσης, μπορούμε να υπολογίσουμε διάφορα μέτρα αξιολόγησης της απόδοσης, όπως η ακρίβεια, η ευαισθησία, η ειδικότητα και άλλα. Είναι ένα ισχυρό εργαλείο που μας επιτρέπει να κατανοήσουμε πώς αποδίδει ένα μοντέλο σε διαφορετικά σενάρια και να προσδιορίσουμε πιθανές προσαρμογές ή βελτιώσεις. [45]




**Εικόνα 18**

Το κύριο μειονέκτημα ενός πίνακα σύγχυσης είναι ότι απαιτεί ανθρώπινη ερμηνεία.

## 4.1 Προβλήματα χαμένων δεδομένων

Στο εν λόγω κεφάλαιο μελετάται η θεωρία της έλλειψης δεδομένων, η οποία παρουσιάζεται από τους Little και Rubin [31]. Εξετάζονται οι διάφορες κατηγορίες των χαμένων δεδομένων και παρουσιάζεται μια ανάλυση των τυχαίων ποσοστών σε περιπτώσεις πολλαπλής έλλειψης δεδομένων.

Οι κύριες κατηγορίες των χαμένων δεδομένων είναι οι εξής:

- Missing Completely at Random (MCAR): Σε αυτήν την περίπτωση, τα δεδομένα λείπουν τυχαία, ανεξάρτητα από τις τιμές των άλλων μεταβλητών ή δεδομένων.
- Missing at Random (MAR): Εδώ, η πιθανότητα έλλειψης δεδομένων εξαρτάται από τις τιμές άλλων παρατηρούμενων μεταβλητών, αλλά όχι από τις λείπουσες τιμές ή τις μη παρατηρούμενες μεταβλητές.
- Not Missing at Random (NMAR) ή Non-Ignorable: Σε αυτήν την περίπτωση, η πιθανότητα έλλειψης δεδομένων εξαρτάται από τις μη παρατηρούμενες τιμές ή τις λανθάνουσες μεταβλητές.

Κάθε κατηγορία έλλειψης δεδομένων έχει διαφορετικές επιπτώσεις στην ανάλυση δεδομένων και στις εκτιμήσεις που μπορούν να γίνουν με βάση αυτά τα δεδομένα.

## 4.2 Η απουσία σε τυχαία θεώρηση

Η κατανόηση και οι επιπτώσεις της έλλειψης δεδομένων σε έναν τυχαίο καταλογισμό μπορεί να είναι πολύπλοκες, ειδικά όταν αυτή εμφανίζεται σε πολλαπλές μεταβλητές με διαφορετικά μοτίβα ελλείψεων. Οι Little και Rubin έχουν εξετάσει πολλά παραδείγματα, υποθέτοντας ότι το σύνολο δεδομένων μπορεί να διαχωριστεί σε δύο μέρη: ένα που επηρεάζεται από την έλλειψη και ένα που παρατηρείται πάντα.

Αν ορίσουμε το υπο-διάνυσμα που παρατηρείται πάντα ως xa τότε ένα μοντέλο δεδομένων που λείπει μπορεί να περιγραφεί από τη μορφή P(R=rjX=x;). Αυτή η προσέγγιση ικανοποιεί την έλλειψη σε τυχαία κατάσταση.

Είναι σημαντικό να σημειωθεί ότι ο περιορισμός της λείανσης των απουσιάζουσων τιμών (MAR) εφαρμόζεται σε κάθε μεταβλητή απόκρισης r και τις σχετικές μεταβλητές x. Σε ορισμένες περιπτώσεις, τα λείποντα δεδομένα σε ένα περιβάλλον δεδομένων που παράγεται από έναν συγκεκριμένο μηχανισμό μπορεί να είναι τυχαία, ενώ σε άλλες περιπτώσεις δεν είναι τυχαία. Για να διασφαλιστεί ότι το μοντέλο απουσιάζοντων δεδομένων εισάγει τυχαία λείανση δεδομένων, πρέπει να τηρούνται οι περιορισμοί του MAR για κάθε διάνυσμα r και την αντίστοιχη αναπαράσταση του x.

| $X \backslash R$ | 0 0 | 0 1 | 1 0 | 1 1 |
|---|---|---|---|---|
| 0 0 | $\alpha$ | $\beta$ | $\gamma$ | $1-\alpha-\beta-\gamma$ |
| 0 1 | $\alpha$ | $\delta$ | $\gamma$ | $1-\alpha-\delta-\gamma$ |
| 1 0 | $\alpha$ | $\beta$ | $\lambda$ | $1-\alpha-\beta-\lambda$ |
| 1 1 | $\alpha$ | $\delta$ | $\lambda$ | $1-\alpha-\delta-\lambda$ |

**Πίνακας 5 Περιορισμοί που έχουν τεθεί για την παραμετροποίηση του P (R = rjX = x) από τον όρο MAR.**

## 4.3 Επίδραση ατελών δεδομένων στα συμπεράσματα

Στη μελέτη του, ο Ρούμπιν[45] ερεύνησε πώς η παρουσία δεδομένων που λείπουν επηρεάζει το συμπέρασμα του Μπεϋζιανού. Αυτά τα ευρήματα τεκμηριώθηκαν επίσης στην πιο πρόσφατη δημοσίευση του Schafer. Σε αυτήν την ενότητα, θα αναλύσουμε την επίδραση των τυχαίων δεδομένων που λείπουν στο συμπέρασμα Bayes και θα τη συγκρίνουμε με την επίδραση των μη τυχαίων δεδομένων που λείπουν. Επιπλέον, τα αποτελέσματα για τη μέγιστη πιθανότητα και το μέγιστο οπίσθιο συμπέρασμα παρουσιάζουν παρόμοια μοτίβα και αναλύονται εκτενώς από τους Little και Rubin.

Μελετάται ένα παραμετρικό μοντέλο πάνω στις μεταβλητές δεδομένων, τις κρυφές μεταβλητές και τους δείκτες απόκρισης, με την υπόθεση ότι η προηγούμενη κατανομή είναι παραγοντοποιημένη. Χωρίς να απλοποιούμε την κατάσταση, η σχέση μεταξύ των απουσιών δεδομένων και του πλήρους μοντέλου δεδομένων εξαρτάται τόσο από τις τιμές των απουσιών όσο και από τις κρυφές μεταβλητές τιμές. Κάτω από τις τυχαίες απουσίες ή εάν η απουσία είναι εντελώς τυχαία, η πιθανότητα P(r_n|xn, xm, z; μ) παραμένει σταθερή για όλες τις τιμές των xm και z όταν δίνονται τα r_n και xo^n. Η μεταβλητή θ είναι πλήρως

ανεξάρτητη από τις μεταβλητές r_n, μ και το μοντέλο δεδομένων που λείπει. Ως εκ τούτου, μπορούμε να αγνοήσουμε τον όρο που αντιστοιχεί στο μοντέλο δεδομένων που λείπει, καθώς δεν επηρεάζει τη μεταβλητή θ. Επιπλέον, η ενσωμάτωση δεδομένων πέρα από τις απουσιάζουσες τιμές περιορίζεται στο πλήρες μοντέλο δεδομένων.

Το αποτέλεσμα αυτών των απλουστεύσεων είναι τα παρατηρούμενα δεδομένα που φαίνονται στην παρακάτω Εξίσωση 12.

$$P^{obs}(\theta|\{\mathbf{x}_n,\mathbf{r}_n\}_{1:N},\omega) \propto P(\theta|\omega)\prod_{n=1}^{N}\int P(\mathbf{x}_n^o,\mathbf{z}|\theta)d\mathbf{z}$$

**Εξίσωση 14**

| $\mathbf{x}$ | $P(\mathbf{x})$ | $P(\mathbf{R}=[0,0]\|\mathbf{x})$ | $P(\mathbf{R}=[0,1]\|\mathbf{x})$ | $P(\mathbf{R}=[1,0]\|\mathbf{x})$ | $P(\mathbf{R}=[1,1]\|\mathbf{x})$ |
|---|---|---|---|---|---|
| 0 0 | $a$ | $\alpha$ | $\beta$ | $\gamma$ | $1-\alpha-\beta-\gamma$ |
| 0 1 | $b$ | $\alpha$ | $\delta$ | $\gamma$ | $1-\alpha-\delta-\gamma$ |
| 1 0 | $c$ | $\alpha$ | $\beta$ | $\lambda$ | $1-\alpha-\beta-\lambda$ |
| 1 1 | $d$ | $\alpha$ | $\delta$ | $\lambda$ | $1-\alpha-\delta-\lambda$ |

**Πίνακας 6 Προδιαγραφές πραγματικών P (X = x) και P (R = rjX = x) ικανοποιώντας MAR.**

Όταν οι συνθήκες MCAR και MAR ισχύουν, μπορούμε πράγματι να αγνοήσουμε το μοντέλο δεδομένων που λείπει χωρίς να αμφισβητηθεί η εγκυρότητα των συμπερασμάτων. Αυτό σημαίνει ότι μπορούμε να καταλήξουμε σε σωστά συμπεράσματα βασιζόμενοι μόνο στα διαθέσιμα δεδομένα, χωρίς να χρειάζεται να λάβουμε υπόψη τα λείποντα δεδομένα. Ωστόσο, όταν οι συνθήκες MCAR και MAR αποτύχουν και τα λείποντα δεδομένα δεν είναι τυχαία, τότε η αγνόηση του μοντέλου δεδομένων που λείπει μπορεί να οδηγήσει σε ανεπαρκείς ή λανθασμένες εκτιμήσεις.Αν βασιστούμε αποκλειστικά στα παρατηρούμενα δεδομένα για να κάνουμε συμπεράσματα για το μοντέλο δεδομένων, χωρίς να λάβουμε υπόψη τα λείποντα δεδομένα, τότε οι εκτιμήσεις μας μπορεί να μην είναι αξιόπιστες. Αυτό επηρεάζει επίσης την ικανότητά μας να προβλέψουμε με ακρίβεια τα μελλοντικά δεδομένα. Επομένως, είναι σημαντικό να λαμβάνουμε υπόψη τις συνθήκες υπό τις οποίες απουσιάζουν τα δεδομένα και να χρησιμοποιούμε κατάλληλες μεθόδους για την αντιμετώπισή τους όταν κάνουμε ανάλυση δεδομένων.

## 4.4 Συμπεράσματα και μη σωστές προδιαγραφές μοντέλου

Είναι σημαντικό να κατανοήσουμε ότι η απλή εφαρμογή ενός τυχαίου μοντέλου δεδομένων που λείπουν δεν είναι αρκετή όταν η συνθήκη MAR δεν επαληθεύεται. Επιπλέον, αν χρησιμοποιηθεί ένα λανθασμένο μοντέλο δεδομένων, η εξαίρεση θα θεωρηθεί επίσης μη έγκυρη. Είναι αξιοσημείωτο ότι, ακόμα κι αν η υπόθεση MAR ισχύει για την υπόθεση σε ένα αληθινό μοντέλο, η υπόθεση σχετικά με τις παραμέτρους ενός βασικού μοντέλου δεδομένων μπορεί να παραμείνει μη έγκυρη, κάτι που είναι αρκετά απροσδόκητο. Αυτό το ζήτημα μπορεί να προκύψει σε βιομηχανίες όπως η μηχανική μάθηση, όπου οργανωμένα μοντέλα χρησιμοποιούνται συχνά για την ανάπτυξη τεχνικών μάθησης που μπορούν να οπτικοποιηθούν και να αποτρέψουν την υπερβολική καταπόνηση. Για να εξηγήσουμε περαιτέρω αυτό το πρόβλημα, ας εξετάσουμε ξανά ένα βασικό σύνολο δεδομένων διανυσμάτων σε δύο διαστάσεις που μπορεί να περιέχει τιμές που λείπουν. Οι παράμετροι που καθορίζουν τον τρόπο λειτουργίας του μηχανισμού δεδομένων, συμβολίζονται ως P(R=rj|X=x), βρίσκονται στον Πίνακα 7 μαζί με την αντίστοιχη κατανομή δεδομένων, P(X=x). Είναι σημαντικό να σημειωθεί ότι το μοντέλο δεδομένων που λείπει ικανοποιεί τη συνθήκη MAR. Τα απουσιάζοντα δεδομένα περιέχονται σε κάθε περίπτωση δεδομένων που προέρχεται από το αρχικό μοντέλο. Οι μεταβλητές δεδομένων X1 και X2 συνήθως δεν είναι ανεξάρτητες με αυτό το αρχικό υλικό επεξεργασίας. Ωστόσο, στην πρακτική της μηχανικής μάθησης, υποθέτουμε ότι η διαδικασία εκμάθησης είναι απλούστερη και χρησιμοποιεί ένα πιο απλό μοντέλο.

Ας εξετάσουμε την παρατηρούμενη συνάρτηση πιθανότητας δεδομένων. Όπου X είναι το διάνυσμα δεδομένων, x είναι ένα συγκεκριμένο σύνολο τιμών για τα δεδομένα, θ είναι το σύνολο των παραμέτρων του μοντέλου, και P(Xi=xi|θ) είναι η πιθανότητα να παρατηρηθεί η τιμή xi για τη μεταβλητή Xi δεδομένου των παραμέτρων θ. Για τις μέγιστες εκτιμήσεις των παραμέτρων θ, χρησιμοποιούμε τους πολλαπλασιαστές Lagrange για να επιβάλουμε την κανονικοποίηση. Έστω L(θ,λ) η συνάρτηση Lagrange, όπου λ είναι οι πολλαπλασιαστές Lagrange.Στόχος είναι να μεγιστοποιήσουμε τη συνάρτηση L(θ,λ)L(θ,λ) ως προς τις παραμέτρους θ και λ. Αυτό μας δίνει τις εκτιμήσεις θ για τις παραμέτρους μας. Η λύση αυτής της εξίσωσης μας δίνει τις μέγιστες πιθανές εκτιμήσεις των παραμέτρων μας.

$$
\begin{aligned}
\mathcal{L}^{obs} &= \sum_n \sum_d \sum_v r_{dn}[x_{dn} = v] \log \theta_{vd} \\
\frac{\partial \mathcal{L}^{obs}}{\partial \theta_{vd}} &= \sum_n \frac{r_{dn}[x_{dn} = v]}{\theta_{vd}} - \lambda = 0 \\
\lambda \theta_{vd} &= \sum_n r_{dn}[x_{dn} = v] \\
\sum_v \lambda \theta_{vd} &= \sum_v \sum_n r_{dn}[x_{dn} = v] \\
\lambda &= \sum_n r_{dn} \\
\theta_{vd} &= \frac{\sum_n r_{dn}[x_{dn} = v]}{\sum_n r_{dn}}
\end{aligned}
$$

**Εξίσωση 15**

Αυτή η διαδικασία εκτίμησης μέγιστης πιθανότητας είναι ασυμπτωτικά ισοδύναμη με τον υπολογισμό P (Xd = vjRd = 1) χρησιμοποιώντας τις πραγματικές αρχικές παραμέτρους.

Ο αλγόριθμος μέγιστης πιθανοφάνειας (Expectation Maximization algorithm - EM) είναι ένα πολύτιμο εργαλείο για την εκτίμηση των παραμέτρων της πραγματικής κατανομής όταν υπάρχουν λείποντα δεδομένα.Με τη χρήση του αλγορίθμου EM, επαναπροσδιορίζουμε τις παραμέτρους της πραγματικής κατανομής με βάση τα παρατηρούμενα δεδομένα. Κάθε επανάληψη του αλγορίθμου βελτιώνει τις εκτιμήσεις των παραμέτρων, μέχρις ότου να συγκλίνει σε μια σταθερή τιμή. Με αυτόν τον τρόπο μπορούμε να ανακτήσουμε τις παραμέτρους της κατανομής και να χρησιμοποιήσουμε αυτές τις εκτιμήσεις για την πρόβλεψη των παραμέτρων των οριακών κατανομών. Έτσι, μπορούμε να αντιμετωπίσουμε αποτελεσματικά τα λείποντα δεδομένα και να εξάγουμε αξιόπιστες εκτιμήσεις.

Ορίζουμε τις μεταβλητές μέτρησης:

| $\beta-\delta$ | True $P(X_1=1)$ | Est. $P(X_1=1)$ | True $P(X_1=1\|R_1=1)$ | Est. $P^M(X_1=1)$ |
|---|---|---|---|---|
| 0.5 | 0.8000 | $0.7999 \pm 0.0007$ | 0.7961 | $0.7961 \pm 0.0007$ |
| 1.0 | 0.8000 | $0.8004 \pm 0.0006$ | 0.7917 | $0.7923 \pm 0.0006$ |
| 1.5 | 0.8000 | $0.7996 \pm 0.0006$ | 0.7868 | $0.7860 \pm 0.0007$ |
| 2.0 | 0.8000 | $0.8011 \pm 0.0007$ | 0.7812 | $0.7826 \pm 0.0008$ |
| 2.5 | 0.8000 | $0.7990 \pm 0.0007$ | 0.7750 | $0.7737 \pm 0.0008$ |
| 3.0 | 0.8000 | $0.8000 \pm 0.0007$ | 0.7679 | $0.7679 \pm 0.0007$ |
| 3.5 | 0.8000 | $0.7994 \pm 0.0008$ | 0.7596 | $0.7582 \pm 0.0009$ |
| 4.0 | 0.8000 | $0.7999 \pm 0.0009$ | 0.7500 | $0.7501 \pm 0.0010$ |
| 4.5 | 0.8000 | $0.7992 \pm 0.0010$ | 0.7386 | $0.7379 \pm 0.0010$ |
| 5.0 | 0.8000 | $0.7986 \pm 0.0010$ | 0.7250 | $0.7241 \pm 0.0010$ |

**Πίνακας 7 Αποτέλεσμα μιας προσομοίωσης σχετικά με τις παραμέτρους εκπαίδευσης**

$$
\begin{aligned}
C_{ij}^1 &= \sum_n [r_1=1][r_2=1][x_1=i][x_2=j] \\
C_i^2 &= \sum_n [r_1=1][r_2=0][x_1=i] \\
C_j^3 &= \sum_n [r_1=0][r_2=1][x_2=j]
\end{aligned}
$$

**Εξίσωση 16**

Μια επανάληψη του αλγορίθμου EM μπορεί να οριστεί χρησιμοποιώντας την τρέχουσα εκτίμηση για τις παραμέτρους και τις παρατηρούμενες μεταβλητές μετρήσεων. Η διαδικασία του αλγορίθμου ξεκινά με μια αρχική εκτίμηση για τις παραμέτρους και στη συνέχεια επαναλαμβάνει τις δύο φάσεις, τη φάση του "Expectation" (E-step) και τη φάση του "Maximization" (M-step), μέχρις ότου η σύγκλιση επιτευχθεί.

Στην φάση του E-step, υπολογίζονται οι "αναμενόμενες" τιμές για τις κρυφές μεταβλητές με βάση τις τρέχουσες εκτιμήσεις των παραμέτρων. Στη φάση του M-step, εκτιμώνται οι νέες τιμές των παραμέτρων βασιζόμενες στις αναμενόμενες τιμές των κρυφών μεταβλητών και στις παρατηρούμενες μεταβλητές μετρήσεων.Αυτή η επανάληψη συνεχίζεται μέχρις ότου η μέγιστη αύξηση στην πιθανοφάνεια είναι κάτω από ένα προκαθορισμένο κατώφλι ή μέχρι να επιτευχθεί μια άλλη συνθήκη σύγκλισης. Κατά τη διάρκεια αυτών των επαναλήψεων, οι εκτιμήσεις των παραμέτρων συγκλίνουν στις πραγματικές τους τιμές.. Δίνουμε την επανάληψη EM στην παρακάτω Εξίσωση

$$
\phi_{ij} \leftarrow \frac{1}{N}\left(C_{ij}^1 + C_i^2 \frac{\phi_{ij}}{\sum_j \phi_{ij}} + C_j^3 \frac{\phi_{ij}}{\sum_i \phi_{ij}}\right)
$$

**Εξίσωση 17**

Ας ξεκινήσουμε με την παράθεση των πραγματικών παραμέτρων όπως έχουν οριστεί:

- Ο παράγοντας δ έχει τιμή 0.1.
- Ο παράγοντας β μεταβάλλεται από 0.15 σε 0.6.
- Η μεταβλητή t λαμβάνει τιμές από 1 έως 10.

Αυτά είναι τα πραγματικά χαρακτηριστικά που χρησιμοποιούμε για την προσομοίωση. Ακολούθως, θα πρέπει να δημιουργήσουμε τις αντίστοιχες μη ολοκληρωμένες περιπτώσεις δεδομένων. Αυτό σημαίνει ότι θα δημιουργήσουμε σύνολα δεδομένων όπου λείπουν κάποιες παρατηρήσεις. Θα χρησιμοποιήσουμε το μοντέλο που ορίζεται στον Πίνακα 7 για τη δημιουργία αυτών των περιπτώσεων. Στη συνέχεια, μπορούμε να

χρησιμοποιήσουμε τον αλγόριθμο ΕΜ για να εκτιμήσουμε τις παραμέτρους σε κάθε περίπτωση δεδομένων και να εξετάσουμε την επίδραση της μεταβολής του β στην εκτίμησή του. Αυτό μας επιτρέπει να αξιολογήσουμε εάν υπάρχει μεροληψία στις εκτιμήσεις μας όταν λείπουν τα δεδομένα. Επίσης, μπορούμε να εξετάσουμε πώς αλλάζει η ακρίβεια των εκτιμήσεων καθώς μεταβάλλουμε την τιμή του β. Το επόμενο βήμα είναι να εφαρμόσουμε τον αλγόριθμο ΕΜ σε κάθε περίπτωση δεδομένων και να αξιολογήσουμε τα αποτελέσματα.

$$a = \phi_{00} = 0.1 \qquad b = \phi_{01} = 0.1 \qquad c = \phi_{10} = 0.7 \qquad d = \phi_{11} = 0.1$$
$$\alpha = 0.1 \qquad \beta = 0.1 + t0.05 \qquad \delta = 0.1 \qquad \gamma = 0.2$$

Το πείραμα που περιγράφεται παραπάνω προσομοιώνει τη διαδικασία της εκτίμησης παραμέτρων σε ένα παραγοντοποιημένο μοντέλο δεδομένων χρησιμοποιώντας τον αλγόριθμο ΕΜ, καθώς και τον τρόπο με τον οποίο αυτή η εκτίμηση επηρεάζεται από την ύπαρξη μεροληψίας στα δεδομένα λόγω της λείψης τυχαίων παρατηρήσεων. Τα αποτελέσματα υποδεικνύουν επίσης ότι η εκτίμηση όλων των παραμέτρων του πραγματικού μοντέλου δεδομένων χρησιμοποιώντας τον αλγόριθμο ΕΜ δεν υπόκειται σε προκατάληψη.

Τα συμπεράσματα από αυτήν τη μελέτη είναι ότι η αποτελεσματική εκτίμηση των παραμέτρων σε ένα παραγοντοποιημένο μοντέλο δεδομένων εξαρτάται από την ποιότητα των δεδομένων και την τήρηση των υποθέσεων. Ακόμη και η ύπαρξη τυχαίων δεδομένων δεν εξασφαλίζει αμεροληψία στις εκτιμήσεις, ειδικά όταν τα πραγματικά δεδομένα δεν πληρούν τις υποθέσεις ανεξαρτησίας που καθορίζονται από το μοντέλο.

## 4.5 Ταξινόμηση με Χαμένα Δεδομένα

Σε αυτό το κεφάλαιο, θα διερευνήσουμε διαφορετικές τεχνικές για τον χειρισμό χαρακτηριστικών που λείπουν στις εργασίες ταξινόμησης. Η παρουσία χαρακτηριστικών που λείπουν επηρεάζει σε μεγάλο βαθμό τη διαδικασία ταξινόμησης, καθώς οι κοινώς χρησιμοποιούμενοι αλγόριθμοι μάθησης όπως η λογιστική παλινδρόμηση, οι μηχανές διανυσμάτων υποστήριξης (SVM) και τα νευρωνικά δίκτυα δεν είναι εγγενώς εξοπλισμένα για να χειρίζονται χαρακτηριστικά εισόδου που λείπουν. Είναι σημαντικό να εξερευνήσουμε μεθόδους ταξινόμησης που είναι ανθεκτικές σε λείποντα χαρακτηριστικά και μπορούν να παρέχουν ακριβείς προβλέψεις ακόμα και όταν τα δεδομένα περιέχουν απουσιάζουσες πληροφορίες. Αυτό είναι ιδιαίτερα σημαντικό στην πράξη, όπου στον πραγματικό κόσμο συχνά υπάρχουν απουσιάζουσες τιμές ή απουσιάζουσες παρατηρήσεις.Θα ξεκινήσουμε παρέχοντας μια ολοκληρωμένη επισκόπηση των στρατηγικών που μπορούν να χρησιμοποιηθούν για την αντιμετώπιση του ζητήματος των ελλιπών δεδομένων στις εργασίες ταξινόμησης. Γενικά, οι γενικοί ταξινομητές είναι σε θέση να μάθουν ένα μοντέλο που περιλαμβάνει τόσο τις ετικέτες όσο και τα χαρακτηριστικά. Αυτό το εγγενές χαρακτηριστικό τους επιτρέπει να μαθαίνουν αποτελεσματικά από ατελείς παρουσίες δεδομένων και να κάνουν ακριβείς προβλέψεις ακόμη και με την παρουσία χαρακτηριστικών που λείπουν. Προχωρώντας προς τα εμπρός, θα εξερευνήσουμε μια σειρά από στρατηγικές που μπορούν να χρησιμοποιηθούν με οποιονδήποτε διακριτικό ταξινομητή, όπως η διαγραφή περιπτώσεων, ο υπολογισμός και η κατηγοριοποίηση υποχώρων. Αυτές οι στρατηγικές προσφέρουν πολύτιμες προσεγγίσεις για τον χειρισμό δεδομένων που λείπουν και τη βελτίωση των αποτελεσμάτων ταξινόμησης. Τέλος, θα εμβαθύνουμε σε ένα πλαίσιο ταξινόμησης που έχει σχεδιαστεί ειδικά για ελλιπή δεδομένα. Αυτό το πλαίσιο περιλαμβάνει τη βελτίωση της αναπαράστασης εισόδου με την ενσωμάτωση ενός διανύσματος δεικτών απόκρισης, επιτρέποντας έτσι τη χρήση πλήρων ταξινομητών δεδομένων.

Η Γραμμική Διακριτική Ανάλυση είναι μόνο ένας τύπος ταξινομητή που εμπίπτει σε μια ευρύτερη κατηγορία. Σε αυτή τη μελέτη, διερευνούμε τη χρήση τεχνικών μέγιστης πιθανότητας και μέγιστης υπό όρους μάθησης στο πλαίσιο ενός μοντέλου κανονικοποιημένης γραμμικής διακριτικής ανάλυσης με δεδομένα που λείπουν. Επιπλέον, εξετάζουμε διάφορες προσεγγίσεις διακριτικής ταξινόμησης, όπως η λογιστική παλινδρόμηση, τα πολυεπίπεδα νευρωνικά δίκτυα και η λογιστική παλινδρόμηση πυρήνα.

### 4.5.1 Frameworks για Ταξινόμηση

Η διαχείριση των λειπόντων δεδομένων είναι ένα σημαντικό θέμα στη μηχανική μάθηση και οι ταξινομητές μπορούν να αντιμετωπίσουν αποτελεσματικά αυτό το πρόβλημα μέσω διαδικασιών όπως η περιθωριοποίηση. Η περιθωριοποίηση επιτρέπει στους ταξινομητές να αντιμετωπίσουν τα λείποντα δεδομένα με τρόπο που διατηρεί την απόδοσή τους και μειώνει τον αντίκτυπο των απουσιών στην ακρίβεια του μοντέλου.Οι διακριτικοί ταξινομητές, οι οποίοι χρησιμοποιούνται συνήθως σε διάφορες εφαρμογές, χρησιμοποιούν διάφορες δημοφιλείς τεχνικές για την αντιμετώπιση του ζητήματος των ελλειπόντων δεδομένων. Αυτές οι τεχνικές περιλαμβάνουν τη διαγραφή κεφαλαίων, τον καταλογισμό και την εκμάθηση υποχώρου. Μια αξιοσημείωτη πτυχή αυτών των μεθόδων είναι ότι μπορούν να συνδυαστούν με οποιονδήποτε ταξινομητή που λειτουργεί σε πλήρη δεδομένα. Στις επόμενες παραγράφους, θα εμβαθύνουμε σε αυτές τις στρατηγικές για αποτελεσματικό χειρισμό δεδομένων που λείπουν. Επιπλέον, θα διερευνήσουμε μια μοναδική προσέγγιση που περιλαμβάνει τη μετατροπή ενός ταξινομητή πλήρους δεδομένων σε έναν που μπορεί να χειριστεί δεδομένα που λείπουν. Αυτή η προσέγγιση περιλαμβάνει τη βελτίωση της εισόδου με την ενσωμάτωση δεικτών απόκρισης, αυξάνοντας έτσι την ικανότητα του ταξινομητή να χειρίζεται δεδομένα που λείπουν.

#### 4.5.1.1 Γενικοί Ταξινομητές

Οι γενικοί ταξινομητές έχουν σχεδιαστεί για να καταγράφουν τη σχέση μεταξύ ετικετών και χαρακτηριστικών σε ένα σύνολο δεδομένων. Σε περιπτώσεις όπου λείπουν τιμές για ορισμένα χαρακτηριστικά, αυτές μπορούν να ληφθούν υπόψη με την περιθωριοποίηση των δεδομένων. Τα μοντέλα υπό όρους, όπως το Naive Bayes και η Linear Discriminant Analysis, είναι σε θέση να χειριστούν αποτελεσματικά αυτήν τη διαδικασία περιθωριοποίησης. Είναι σημαντικό να αντιμετωπιστούν τα δεδομένα που λείπουν κατά τη διάρκεια της φάσης εκμάθησης προκειμένου να διασφαλιστεί η ακριβής εκπαίδευση του μοντέλου.Η εφαρμογή του αλγόριθμου Expectation Maximization είναι συνήθως απαραίτητη για να επιτευχθεί αυτό. Ωστόσο, οι γενετικοί ταξινομητές απαιτούν τη διατύπωση ρητών υποθέσεων σχετικά με την κατανομή του χώρου χαρακτηριστικών, ενώ οι διακριτικοί ταξινομητές δεν απαιτούν τέτοιες υποθέσεις.

#### 4.5.1.2 Διαγραφή Υποθέσεων

Η διαγραφή περιπτώσεων λόγω έλλειψης τιμών χαρακτηριστικών είναι μια απλή προσέγγιση για τη διαχείριση των λειπόντων δεδομένων, αλλά έχει το μειονέκτημα της απώλειας πληροφοριών. Η αφαίρεση περιπτώσεων μπορεί να επηρεάσει την ακρίβεια και την απόδοση του μοντέλου, ειδικά εάν οι περιπτώσεις που διαγράφονται αντιπροσωπεύουν ένα σημαντικό ποσοστό του συνόλου δεδομένων.Για να αντιμετωπιστεί αυτό το πρόβλημα, υπάρχουν εναλλακτικές προσεγγίσεις. Μία από αυτές είναι η αναπλήρωση των λειπόντων τιμών με τη χρήση μεθόδων ανακατασκευής, όπως η απόδοση της μέσης τιμής ή η ανάθεση τιμών βάσει των γειτονικών παρατηρήσεων. Αυτές οι μέθοδοι μπορούν να βοηθήσουν στη διατήρηση του συνόλου δεδομένων και την αποφυγή απώλειας πληροφοριών.

#### 4.5.1.3 Ταξινόμηση και Καταλογισμός

Ο καταλογισμός είναι μια προσέγγιση που χρησιμοποιείται συνήθως στο στατιστικό πεδίο για την αντιμετώπιση του ζητήματος των στοιχείων που λείπουν. Όταν στοχεύουμε στον προσδιορισμό μιας μέσης απόδοσης, ο μέσος όρος ενός συγκεκριμένου στοιχείου, που συμβολίζεται ως d, υπολογίζεται αποκλειστικά με βάση τις περιπτώσεις όπου παρατηρείται το αντίστοιχο χαρακτηριστικό, που αναφέρεται επίσης ως d. Κατά συνέπεια, αυτή η μέση τιμή χρησιμοποιείται στη συνέχεια ως αντικατάσταση για το χαρακτηριστικό d σε περιπτώσεις όπου το χαρακτηριστικό d δεν παρατηρείται. Στην περίπτωση της απόδοσης παλινδρόμησης, δημιουργείται αρχικά μια συλλογή μοντέλων παλινδρόμησης, ειδικά σχεδιασμένα για να χειρίζονται χαρακτηριστικά που λείπουν, αξιοποιώντας αυτά που παρατηρούνται. Αυτά τα μοντέλα στη συνέχεια χρησιμοποιούνται κατά τη διάρκεια της μαθησιακής διαδικασίας για να συμπληρώσουν τα χαρακτηριστικά που λείπουν χρησιμοποιώντας προβλεπόμενες τιμές που λαμβάνονται μέσω μοντελοποίησης μαθηματικής παλινδρόμησης.Η παλινδρόμηση παραμέτρων και ο μέσος όρος θεωρούνται μέθοδοι μονής καταλογισμού, όπου μια μεμονωμένη τιμή εκχωρείται σε μεταβλητές που λείπουν σε ένα σύνολο δεδομένων. Ωστόσο, ο πολλαπλός καταλογισμός προσφέρει μια διαφορετική προσέγγιση αντικαθιστώντας πολλαπλές τιμές για

κάθε μεταβλητή που λείπει. Αυτό περιλαμβάνει τη δειγματοληψία τεκμαρτών τιμών από ένα μοντέλο τεκμαρτών και την εφαρμογή τυποποιημένων τεχνικών σε κάθε ολοκληρωμένο σύνολο δεδομένων.. Το βασικό πλεονέκτημα του πολλαπλού καταλογισμού σε σχέση με τον απλό καταλογισμό είναι η ικανότητά του να συλλαμβάνει την αβεβαιότητα που προκαλείται από τα δεδομένα που λείπουν πιο αποτελεσματικά. Υπάρχουν εξειδικευμένες παραλλαγές πολλαπλών καταλογισμών που ευθυγραμμίζονται στενά με τις τεχνικές Bayes, όπως οι μέθοδοι Monte Carlo της αλυσίδας Markov. Αυτές οι προσεγγίσεις μπορούν να θεωρηθούν ως ένα μέσο απόδοσης δεδομένων που λείπουν εξετάζοντας μια κατανομή χρησιμότητας σε όλο τον χώρο χαρακτηριστικών.

#### 4.5.1.4 Ταξινόμηση σε υπο-διαστήματα: Μειωμένα Μοντέλα

Η προσέγγιση του μειωμένου μοντέλου που περιγράφεται αποτελεί μια ενδιαφέρουσα λύση για το χειρισμό δεδομένων που λείπουν, καθώς επιτρέπει τη χρήση παραδοσιακών τεχνικών μηχανικής μάθησης για την εκπαίδευση ξεχωριστών μοντέλων για κάθε μοναδικό σύνολο παρατηρούμενων τιμών. Αυτό μπορεί να οδηγήσει σε βελτιωμένη ακρίβεια ταξινόμησης σε σύγκριση με τη χρήση ενός μόνο ταξινομητή που λαμβάνει υπόψη όλα τα χαρακτηριστικά.Ωστόσο, όπως αναφέρει ο Tresp et al., ένα από τα βασικά μειονεκτήματα αυτής της προσέγγισης είναι ότι ο αριθμός των διαφορετικών μοτίβων ελλειπούσων λειτουργιών είναι εκθετικός ως προς τον αριθμό των χαρακτηριστικών. Σε περιπτώσεις όπου ο αριθμός των χαρακτηριστικών αυξάνεται, η προσέγγιση αυτή μπορεί να γίνει δυσκολότερη ή ακόμη και αδύνατη λόγω του μεγάλου αριθμού διαφορετικών μοτίβων. Στην περίπτωση των Sharpe και Solly, το σύνολο δεδομένων περιείχε μόνο τέσσερα χαρακτηριστικά και μόνο τέσσερα διαφορετικά μοτίβα ελλειπούσων χαρακτηριστικών. Αυτό καθιστά εφικτή την εφαρμογή της προσέγγισης μειωμένου μοντέλου χωρίς να προκύψουν σοβαρά προβλήματα λόγω του μεγάλου αριθμού διαφορετικών μοτίβων. Ωστόσο, σε πιο πολύπλοκα σύνολα δεδομένων με περισσότερα χαρακτηριστικά, η προσέγγιση αυτή μπορεί να αποτύχει ή να απαιτήσει περισσότερη προσοχή και προσαρμογή.

### 4.6 Γραμμική Ανάλυση Διακρίσεων

Η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis - LDA) είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για την ταξινόμηση δεδομένων. Η αρχική ιδέα του LDA προέρχεται από τον Ronald A. Fisher, ο οποίος πρότεινε έναν τρόπο να βρει μια γραμμική συνάρτηση η οποία μπορεί να χωρίσει δύο ή περισσότερες κατηγορίες των δεδομένων. Η LDA αναζητά έναν γραμμικό μετασχηματισμό των χαρακτηριστικών που μεγιστοποιεί τη διακριτική ικανότητα μεταξύ των κατηγοριών.Μια επέκταση της ιδέας του Fisher προέρχεται από την προοπτική της ταξινόμησης μέγιστης πιθανότητας. Αυτή η προσέγγιση βασίζεται σε ένα μοντέλο Gaussian που εξαρτάται από την τάξη, δηλαδή υποθέτει ότι οι κατηγορίες ακολουθούν διαφορετικές κανονικές κατανομές. Στόχος είναι η ελαχιστοποίηση του σφάλματος ταξινόμησης.Η Τετραγωνική Διακριτική Ανάλυση (Quadratic Discriminant Analysis - QDA) είναι μια επέκταση της LDA που διαφοροποιείται στο ότι επιτρέπει διαφορετικές κανονικές κατανομές για κάθε κατηγορία, αντί να υποθέτει μια κοινή κανονική κατανομή.Η Τακτοποιημένη Διακριτική Ανάλυση (Regularized Discriminant Analysis - RDA) είναι μια επέκταση της LDA που χρησιμοποιεί έναν παραμετροποιημένο τρόπο για την εκτίμηση των κατανομών των χαρακτηριστικών.Για να αντιμετωπιστούν τα λείποντα χαρακτηριστικά, ορισμένοι ερευνητές προτείνουν τη χρήση επιπλέον ταξινομητών που εκπαιδεύονται για διαφορετικά μοτίβα λείψανων δεδομένων, όπως κάνανε οι Sharpe και Solly. Αυτή η προσέγγιση μπορεί να βελτιώσει την ακρίβεια της ταξινόμησης όταν αντιμετωπίζουμε δεδομένα με λείποντα χαρακτηριστικά..

### 4.6.1 Γραμμική Ανάλυση Διακρίσεων Fisher

Η αναφορά σε ένα κριτήριο που προτείνεται από τον Fisher για τη δυαδική ρύθμιση ταξινόμησης είναι η χρήση του κριτηρίου του μέσου όρου των δύο κλάσεων (μ1, μ-1) που πρέπει να διαχωρίζονται μέγιστα από την επιλογή της διακύμανσης κατεύθυνσης w.Ουσιαστικά, αυτό το κριτήριο επιδιώκει να μεγιστοποιήσει την απόσταση μεταξύ των μέσων των δύο κλάσεων ενώ λαμβάνει υπόψη τη διακύμανση των κλάσεων. [44]

Το δεύτερο κριτήριο που προτείνει ο Fisher είναι ότι το επιλεγμένο w πρέπει να ελαχιστοποιήσει την εσωτερική διακύμανση των δύο κατηγοριών μέσα στην κλάση S, όπως εκφράζεται στην παρακάτω εξίσωση. Ο ορισμός της μήτρας διακύμανσης S δίνεται στην Εξίσωση 16.

Η εσωτερική διακύμανση S μπορεί να οριστεί ως:.

$$\mathbf{S} = \sum_{n=1}^{N} \sum_{c \in \{-1,1\}} [y_n = c](\mathbf{x}_n - \boldsymbol{\mu}_c)(\mathbf{x}_n - \boldsymbol{\mu}_c)^T$$

**Εξίσωση 18**

### 4.6.2 Ανάλυση της Γραμμικής Διάκρισης ως την Μέγιστη Ταξινόμηση

Ο Welch ανέπτυξε την έννοια της ταξινόμησης μέγιστης πιθανότητας, η οποία συνίσταται στο να επιλεγεί η κατηγορία που έχει τη μεγαλύτερη πιθανότητα να προκύψει δεδομένων των παρατηρούμενων χαρακτηριστικών. Αυτή η προσέγγιση συνδέθηκε με το μοντέλο τάξης Gauss, όπου η συνδιακύμανση είναι κοινή και οι μέσοι όροι είναι διακριτοί. Με βάση αυτήν την προσέγγιση, ο Welch κατέληξε στο συμπέρασμα ότι ο γραμμικός διαχωρισμός του Fisher μπορεί να προκύψει από τη διάταξη της οπίσθιας κατανομής. Αυτή η προσέγγιση κατέληξε στη συμβατική υπόθεση για γραμμική διακριτή ανάλυση, η οποία αποτελεί μια διαδεδομένη τεχνική στη μηχανική μάθηση.

### 4.6.3 Τετραγωνική Ανάλυση Διακρίσεων

Η τετραγωνική διάκριση επεκτείνει τη Γραμμική διάκριση του Fisher προσαρμόζοντας διαφορετικούς πίνακες συσχέτισης για κάθε τάξη. Αυτό συνεπάγεται μια οπίσθια κατανομή κατάταξης που δεν μπορεί να απλοποιηθεί σε μια γραμμική συνάρτηση, αλλά σχηματίζει μια τετράγωνη επιφάνεια. Όπως στην περίπτωση της γραμμικής διάκρισης, ο κανόνας ταξινόμησης είναι να ανατεθεί ένα διάνυσμα xnxn στην κλάση με την υψηλότερη οπίσθια πιθανότητα. Για παράδειγμα, αν έχουμε ένα σύνολο δεδομένων με εικόνες που ανήκουν σε δύο κλάσεις (π.χ. γάτες και σκύλους), η τετραγωνική διάκριση θα προσαρμόσει διαφορετικούς πίνακες συσχέτισης για κάθε κλάση, λαμβάνοντας υπόψη την εσωτερική δομή και τις σχέσεις μεταξύ των εικόνων σε κάθε κλάση. Κατά την ταξινόμηση μιας νέας εικόνας, η τετραγωνική διάκριση θα επιλέξει την κλάση με την υψηλότερη οπίσθια πιθανότητα, λαμβάνοντας υπόψη τους πίνακες συσχέτισης που έχουν εκπαιδευτεί για κάθε κλάση. [77].

### 4.6.4 Ρυθμιζόμενη ανάλυση διακρίσεων

Για να επεξηγήσει περαιτέρω αυτό το σημείο, ο Friedman παρέχει μια συναρπαστική επίδειξη της συνάρτησης διάκρισης, η οποία βασίζεται στη φασματική ανάλυση του πίνακα συνδιακύμανσης. Σε αυτή την ενότητα, θα εμβαθύνουμε στο θέμα των στρατηγικών τακτοποίησης για διαφορική γραμμική και τετραγωνική ανάλυση διάκρισης. Μία από τις μεγαλύτερες προκλήσεις σε αυτόν τον τομέα έγκειται στην ακριβή εκτίμηση των πινάκων συνδιακύμανσης, ειδικά όταν πρόκειται για δεδομένα υψηλών διαστάσεων και αραιά. Είναι ευρέως αναγνωρισμένο ότι καθώς μειώνεται ο αριθμός των περιπτώσεων δεδομένων, η δειγματοληπτική διακύμανση της τυπικής εκτίμησης συνδιακύμανσης τείνει να αυξάνεται. Επιπλέον, όταν ο αριθμός των περιπτώσεων δεδομένων (Nc) είναι μικρότερος από τον αριθμό των διαστάσεων (D), δεν μπορούν να αναγνωριστούν όλες οι παράμετροι συνδιακύμανσης για μια συγκεκριμένη κλάση (c).

### 4.6.5 Μη Επιτηρούμενη Συρρίκνωση

Ο Friedman προτείνει ότι οι ιδιοτιμές της συνάρτησης συνδιακύμανσης δείγματος παρουσιάζουν ένα συστηματικό μοτίβο, με μικρότερες ιδιοτιμές να τείνουν να είναι χαμηλού μεγέθους και μεγαλύτερες ιδιοτιμές υψηλότερες σε ένταση. Στο παρελθόν, οι μέθοδοι για την τακτοποίηση των εκτιμήσεων συσχέτισης αποσκοπούσαν στην αντιμετώπιση τυχόν προκατάληψης που υπήρχε στην κατανομή των ιδιοτιμών.

### 4.6.6 Επιτηρούμενη Συρρίκνωση

Ο Friedman[65] έφερε επανάσταση στον τομέα εισάγοντας μια πρωτοποριακή τεχνική γνωστή ως κανονικοποίηση συνδιακύμανσης για την περίπτωση της τετραγωνικής βάσης. Αυτή η μέθοδος περιλαμβάνει την ελαχιστοποίηση της εκτίμησης του σφάλματος πρόβλεψης που λαμβάνεται μέσω διασταυρούμενης επικύρωσης. Η υποκείμενη ιδέα περιστρέφεται γύρω από τη μείωση του μεγέθους των πινάκων

συνδιακύμανσης κατά κατηγορία ώστε να ταιριάζει με τον πίνακα συνδιακύμανσης ομαδοποιημένης ενσωμάτωσης μιας παραμέτρου στάθμισης α. Επιπλέον, αυτή η εκτίμηση μειώνεται περαιτέρω σε μια κλιμακούμενη έκδοση του πίνακα ταυτότητας με την εισαγωγή μιας δεύτερης παραμέτρου στάθμισης c.

### 4.6.6 LDA και Χαμένα Δεδομένα

Η Γραμμική Ανάλυση Διακρίσεων (LDA) είναι ένα γραμμικό μοντέλο που χρησιμοποιείται για την ταξινόμηση δεδομένων. Μια από τις εφαρμογές του LDA είναι η αντιμετώπιση των λειπόντων χαρακτηριστικών εισόδου στα δεδομένα.Η τάξη πιθανότητας ενός διανύσματος δεδομένων με ελλείψεις λειτουργίες εισόδου μπορεί να υπολογιστεί χρησιμοποιώντας το LDA. Αυτό συνήθως γίνεται με τη χρήση των παραμέτρων που εκτιμώνται κατά την εκπαίδευση του μοντέλου.Οι οπίσθιες πιθανότητες κάθε κατηγορίας, δηλαδή η πιθανότητα να ανήκει ένα δεδομένο σε μια συγκεκριμένη κατηγορία, μπορούν να υπολογιστούν χρησιμοποιώντας την κατανομή που προκύπτει από το LDA μοντέλο.

$$P(\mathbf{X}_n^o = \mathbf{x}_n^o | Y_n = c) = |2\pi\Sigma^{oo}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_n^o - \mu_c^o)^T(\Sigma^{oo})^{-1}(\mathbf{x}_n^o - \mu_c^o)\right)$$

$$P(Y = c | \mathbf{X}_n^o = \mathbf{x}_n^o) = \frac{\theta_c |2\pi\Sigma^{oo}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_n^o - \mu_c^o)^T(\Sigma^{oo})^{-1}(\mathbf{x}_n^o - \mu_c^o)\right)}{\sum_c \theta_c |2\pi\Sigma^{oo}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_n^o - \mu_c^o)^T(\Sigma^{oo})^{-1}(\mathbf{x}_n^o - \mu_c^o)\right)}$$

**Εξίσωση 19**

**Μέγιστη εκτίμηση πιθανότητας**

Η μέγιστη εκτίμηση πιθανότητας των μέσων παραμέτρων υπολογίζεται από ελλιπή δεδομένα όπως φαίνεται στην παρακάτω Εξίσωση.[38]

$$\mu_{dc} = \frac{\sum_{n=1}^{N}[y_n = c][r_{dn} = 1]x_{dn}}{\sum_{n=1}^{N}[y_n = c][r_{dn} = 1]}$$

**Εξίσωση 20**

### 4.6.7 Λογιστική Παλινδρόμηση(Logistic Regression)

Η λογιστική παλινδρόμηση αναφέρεται σε μια στατιστική τεχνική που χρησιμοποιείται για τη μοντελοποίηση της σχέσης μεταξύ μιας δυαδικής ή κατηγορικής μεταβλητής κλάσης και ενός συνόλου χαρακτηριστικών. Το όνομά της προέρχεται από τη χρήση της λογιστικής συνάρτησης στην αντιστοίχιση των πραγματικών τιμών στο διάστημα [0, 1].Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί σε διάφορες περιπτώσεις, συμπεριλαμβανομένων των πολυκατηγορικών προβλημάτων κατηγοριοποίησης και της δυαδικής ταξινόμησης. Συνήθως, το μοντέλο προσαρμόζεται με τη χρήση μεθόδων όπως η μέθοδος μέγιστης πιθανοφάνειας.Σε περιπτώσεις όπου τα δεδομένα περιέχουν απουσιάζουσες τιμές, η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί με διάφορους τρόπους. Οι στρατηγικές περιλαμβάνουν τον καταλογισμό, όπου οι απουσιάζουσες τιμές αντικαθίστανται με εκτιμήσεις, τα μειωμένα μοντέλα, όπου χρησιμοποιούνται μόνο τα χαρακτηριστικά με πλήρη δεδομένα, και οι δείκτες απόκρισης, που λαμβάνουν υπόψη την πιθανότητα απουσιάζουσας τιμής κατά την εκπαίδευση του μοντέλου.Η λογιστική παλινδρόμηση είναι μια ισχυρή και ευέλικτη τεχνική που μπορεί να χρησιμοποιηθεί σε ποικίλες στατιστικές και μηχανικής μάθησης εφαρμογές, και μπορεί να προσαρμοστεί για να αντιμετωπίσει διάφορα προβλήματα που αφορούν την απουσία δεδομένων.

### 4.6.7.2 Μέγιστη Εκτίμηση Πιθανότητας

Η Μέγιστη Εκτίμηση Πιθανότητας (Maximum Likelihood Estimation - MLE) είναι μια μέθοδος για την εύρεση των βέλτιστων παραμέτρων ενός μοντέλου, με βάση την πιθανότητα των παρατηρήσεων από τα δεδομένα. Για παράδειγμα, στην παλινδρόμηση, η MLE χρησιμοποιείται για να βρει τις παραμέτρους που μεγιστοποιούν την πιθανότητα να παρατηρηθούν τα πραγματικά δεδομένα, δεδομένου ενός μοντέλου. Για

παράδειγμα, ας υποθέσουμε ένα μοντέλο παλινδρόμησης που προβλέπει τις τιμές του καταναλωτή με βάση το εισόδημα. Η MLE θα χρησιμοποιηθεί για να εκτιμήσει τις παραμέτρους του μοντέλου, δηλαδή το καλύτερο προσαρμοσμένο γραμμικό μοντέλο που ταιριάζει με τα δεδομένα, με βάση την πιθανότητα να παρατηρηθούν οι πραγματικές τιμές κατανάλωσης.

Η μέγιστοποίηση της συνάρτησης πιθανοφάνειας σε προβλήματα πιθανοτικής ταξινόμησης όπως οι λογιστικές παλινδρομήσεις απαιτεί συνήθως τη χρήση ενός επαναληπτικού αλγορίθμου βελτιστοποίησης, όπως ο αλγόριθμος gradient descent. Αυτό συμβαίνει επειδή συνήθως δεν υπάρχει αναλυτική λύση κλειστού τύπου για τη βέλτιστη λύση, οπότε πρέπει να χρησιμοποιηθούν αριθμητικές μέθοδοι για τη βελτιστοποίηση της συνάρτησης κόστους. Ο αλγόριθμος gradient descent είναι ένα παράδειγμα επαναληπτικού αλγορίθμου που χρησιμοποιείται συχνά για αυτόν το σκοπό.

#### 4.6.7.3 Ρυθμίσεις για Λογιστική Παλινδρόμηση

Στη λογιστική παλινδρόμηση, η τακτοποίηση (regularization) είναι ένα σημαντικό εργαλείο για την αποφυγή του overfitting και τη βελτίωση της γενικευτικής ικανότητας του μοντέλου. Δύο κύριοι τύποι τακτοποίησης είναι η ρύθμιση L2 και η ρύθμιση L1.Η ρύθμιση L2, γνωστή και ως λαπλασιανή ρύθμιση, προσθέτει έναν όρο κανονικοποίησης στη συνάρτηση κόστους, βασιζόμενος στο τετράγωνο των βαρών του μοντέλου. Αυτός ο όρος βοηθά στην αποφυγή υπερεκπαίδευσης και στην εξασφάλιση μιας πιο γενικευμένης μοντελοποίησης.Από την άλλη πλευρά, η ρύθμιση L1, γνωστή και ως λαπλασιανή ρύθμιση, προσθέτει έναν όρο κανονικοποίησης βασιζόμενο στην απόλυτη τιμή των βαρών του μοντέλου. Αυτός ο όρος έχει την ιδιότητα ότι τείνει να μειώσει ορισμένα βάρη στο μηδέν, εκτελώντας έτσι επιλογή χαρακτηριστικών και τακτοποίηση ταυτόχρονα. Αυτό είναι χρήσιμο σε περιπτώσεις όπου υπάρχουν πολλά χαρακτηριστικά και θέλουμε να επιλέξουμε τα σημαντικότερα.Η επιλογή μεταξύ L1 και L2 ρύθμισης εξαρτάται συχνά από το πρόβλημα που αντιμετωπίζουμε. Ενώ η L2 ρύθμιση είναι πιο κοινή, η L1 ρύθμιση μπορεί να είναι προτιμητέα όταν ενδιαφερόμαστε για επιλογή χαρακτηριστικών.Υπάρχουν διάφορες μέθοδοι επίλυσης για τη λογιστική παλινδρόμηση με L1 ρύθμιση. Αυτές οι μέθοδοι συνήθως επιδιώκουν να επιλύσουν αποτελεσματικά το πρόβλημα επιλογής χαρακτηριστικών και ταυτόχρονα να ελαχιστοποιήσουν το σφάλμα του μοντέλου. Οι προτεινόμενες μέθοδοι περιλαμβάνουν προσαρμοσμένα εσωτερικά σημεία, τροποποιημένες με ελάχιστα τετράγωνα και τροποποιημένες μεθόδους quasi-Newton περιορισμένης μνήμης.Συνολικά, η επιλογή μεταξύ L1 και L2 ρύθμισης και η χρήση κατάλληλων μεθόδων τακτοποίησης εξαρτάται από τη φύση του προβλήματος και τις απαιτήσεις της εφαρμογής.

## 4.7 Νευρώνας Perceptron,Support Vector Machines

Η γραμμική διάκριση και η λογιστική παλινδρόμηση θεωρούνται συχνά προηγμένοι ταξινομητές. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η αντικειμενική τους συνάρτηση δεν επιδιώκει απευθείας την ελαχιστοποίηση του σφάλματος ταξινόμησης. Η γραμμική διάκριση εστιάζει στη μεγιστοποίηση της πιθανότητας κοινών χαρακτηριστικών και ετικετών, ενώ η λογιστική παλινδρόμηση προσπαθεί να βελτιστοποιήσει την πιθανότητα των ετικετών με βάση τα χαρακτηριστικά. Προχωρώντας προς την επόμενη ενότητα, θα εμβαθύνουμε στο θέμα των perceptrons και των μηχανών διανυσμάτων υποστήριξης (SVMs). Το perceptron είναι ένας ταξινομητής με μονή στρώση που στοχεύει να ελαχιστοποιήσει άμεσα το σφάλμα ταξινόμησης. Από την άλλη πλευρά, τα SVM επιλέγουν ένα βέλτιστο υπερεπίπεδο εξετάζοντας προσεκτικά την ισορροπία μεταξύ του σφάλματος ταξινόμησης και της ποινής πολυπλοκότητας.

### 4.7.1 Perceptrons

Το perceptron είναι ένας τύπος δυαδικού γραμμικού ταξινομητή που χρησιμοποιείται για την επιλογή ενός υπερεπίπεδου προκειμένου να βελτιστοποιηθεί το σφάλμα ταξινόμησης. Αυτό γίνεται μέσω του κανόνα εκμάθησης του perceptron, ο οποίος αναπτύχθηκε αρχικά από τον Rosenblatt ως αλγόριθμος στοχαστικής κλίσης. Σε αυτόν τον αλγόριθμο, τα στιγμιότυπα δεδομένων παρουσιάζονται διαδοχικά και εάν μια τρέχουσα παρουσία δεδομένων ταξινομηθεί εσφαλμένα, γίνονται προσαρμογές στις παραμέτρους. Η παράμετρος α είναι ο ρυθμός εκμάθησης.

$$f_{Pct}(\mathbf{w}, b) = \sum_{n=1}^{N} \frac{1}{2} |y_n - \text{sgn}(\mathbf{w}^T \mathbf{x}_n + b)|$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha y_n \mathbf{x}_n$$

$$b \leftarrow b + \alpha y_n$$

**Εξίσωση 21**

Ο αλγόριθμος εκτελείται μέχρι να διασχίσει το πλήρες σετ εκπαίδευσης, χωρίς να υπάρξουν ενημερώσεις στις παραμέτρους. Η ερμηνεία του perceptron σε ένα επίπεδο ως διαχωριστικό επίπεδο, μαζί με τους περιορισμούς που συνεπάγονται αυτό, κατανοήθηκε αργότερα από τους Minsky και Papert.[56]. Ο αλγόριθμος perceptron δεν θα συγκλίνει εάν τα δεδομένα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα. Αντίθετα, θα παρουσιάσει αστάθεια και θα έχει ως αποτέλεσμα την ανάπτυξη κύκλων στις τιμές των παραμέτρων.Για την αντιμετώπιση του προβλήματος μη διαχωρίσιμων συνόλων εκπαίδευσης, η μέθοδος perceptron έχει βελτιωθεί με τη χρήση της soft margin support vector machine (SVM). Η SVM επιτρέπει την ύπαρξη μικρών παραβάσεων στον αρχικό διαχωρισμό, επιτρέποντας την καλύτερη διαχωρισιμότητα των κλάσεων ενώ παράλληλα αποφεύγει την υπερεκπαίδευση (overfitting).

### 4.7.2 Hard Margin Support Vector Machines

Όπως αναφέρθηκε προηγουμένως, υπάρχει μια τεράστια ποικιλία υπερεπιπέδων που μπορούν να ταξινομήσουν αποτελεσματικά ένα σύνολο δεδομένων με διακριτές κατηγορίες. Ο κανόνας εκμάθησης perceptron οδηγεί σε ένα υπερεπίπεδο που επιλέγεται ουσιαστικά τυχαία από αυτήν την εκτεταμένη συλλογή. Αντίθετα, ο Vapnik[45] προσδιορίζει το βέλτιστο διαχωριστικό υπερεπίπεδο ως αυτό που μεγιστοποιεί την ελάχιστη απόσταση μεταξύ κάθε σημείου δεδομένων στο σύνολο εκπαίδευσης, διαιρώντας το υπερεπίπεδο σύμφωνα με μια συγκεκριμένη εξίσωση.Δεδομένου ότι ο ίδιος κανόνας απόφασης προκύπτει όταν οι παράμετροι w και b πολλαπλασιάζονται με μια σταθερά, μπορεί να χρησιμοποιηθεί οποιαδήποτε κλίμακα του βέλτιστου w.

**Μέθοδοι επέκτασης βάσης και μέθοδοι Kernel**

Οι μέθοδοι ταξινόμησης που αναφέρθηκαν παραπάνω παρουσιάζουν μια περιορισμένη δυνατότητα αναπαράστασης αποφάσεων, καθώς οι επιφάνειες απόφασης περιγράφονται από γραμμικά υπερπλάνα. Μόνο η τετραγωνική διακριτική ανάλυση και ορισμένες μορφές κανονικοποιημένων διακρίσεων ανάλυσης παρουσιάζουν μη γραμμικές επιφάνειες απόφασης. Συνήθως, οι περιορισμοί που επιβάλλονται από τα γραμμικά μοντέλα μπορούν να ξεπεραστούν με τη χρήση όρων αλληλεπίδρασης υψηλότερης τάξης. Για παράδειγμα, εάν υπάρχει ισχυρή αλληλεπίδραση μεταξύ δύο χαρακτηριστικών x1 και x2, μπορεί να συμπεριληφθεί ένας όρος αλληλεπίδρασης x1x2 στο μοντέλο. Αυτό επεκτείνει το υπερπλάνο που περιγράφει την απόφαση σε μη γραμμική μορφή.Οι βασικές επεκτάσεις και οι μέθοδοι πυρήνα γενικεύουν αυτήν τη βασική στρατηγική, επιτρέποντας στον γραμμικό ταξινομητή να γίνει μη γραμμικός. Οι αλγόριθμοι που προκύπτουν από αυτήν την προσέγγιση είναι συχνά κυρτοί για μια σταθερή επέκταση της βάσης ή για λειτουργία πυρήνα..

**Επέκταση βάσης**

Η ιδέα των όρων αλληλεπίδρασης μπορεί να γενικευθεί με τον υπολογισμό παραγώγων χαρακτηριστικών, όπου κάθε παράγοντας του διανύσματος χαρακτηριστικών μετασχηματίζεται σε μια κλιμακούμενη τιμή. Γενικότερα, μπορούμε να υποθέσουμε την ύπαρξη μιας συνάρτησης Φ(x)Φ(x) που μετατρέπει διανυσματικά διανύσματα DD διαστάσεων σε διανυσματικούς MM φορείς. Συνήθως το MM θα είναι μεγαλύτερο από το DD, επομένως ο φορέας zz αναφέρεται ως επέκταση του xx. Υποθέτοντας ότι η συνάρτηση Φ(x)Φ(x) είναι σταθερή, ο φορέας znzn μπορεί να υπολογιστεί για κάθε xnxn. Είναι ένα ευέλικτο πλαίσιο για την εφαρμογή μη γραμμικών μοντέλων σε δεδομένα που παρουσιάζουν μη γραμμική συσχέτιση μεταξύ των χαρακτηριστικών. Η ιδέα είναι να μετασχηματίσετε τα αρχικά χαρακτηριστικά σε ένα νέο χώρο

χαρακτηριστικών χρησιμοποιώντας μη γραμμικές συναρτήσεις, όπως η συνάρτηση χαρακτηριστικών βάσης (basis function), και στη συνέχεια να εφαρμόσετε ένα γραμμικό μοντέλο στο νέο χώρο χαρακτηριστικών.Η μετασχηματισμένη μορφή των χαρακτηριστικών μπορεί να καθοριστεί από την εφαρμογή της βάσης σε κάθε χαρακτηριστικό. Οι συναρτήσεις βάσης μπορεί να είναι, για παράδειγμα, πολυωνυμικές, RBF (Radial Basis Function), στοχαστικοί νευρώνες κ.λπ. Η σημασία επιλογής των συναρτήσεων βάσης είναι κρίσιμη για την απόδοση του μοντέλου.Ο μετασχηματισμός με μη γραμμικές συναρτήσεις βάσης επιτρέπει την προσαρμογή στην πολυπλοκότητα των δεδομένων, αλλά η επιτυχής εφαρμογή αυτής της τεχνικής απαιτεί την κατανόηση και την επιλογή κατάλληλων συναρτήσεων βάσης.

**Μέθοδοι kernel**

Σε πολλά μοντέλα γραμμικής ταξινόμησης, οι βέλτιστες παράμετροι μπορούν να εκφραστούν ως γραμμικός συνδυασμός των δεδομένων. Αυτό σημαίνει ότι ο κανόνας ταξινόμησης εξαρτάται μόνο από τα εσωτερικά γινόμενα των διανυσμάτων χαρακτηριστικών. Για παράδειγμα, στη λογιστική παλινδρόμηση και στους perceptrons, οι βέλτιστες παράμετροι μπορούν να εκφραστούν ως γραμμικός συνδυασμός των φορέων δεδομένων. Επίσης, στην περίπτωση του SVM, η ανάλυση του διπλού προβλήματος βελτιστοποίησης αποκαλύπτει ότι οι βέλτιστες παράμετροι μπορούν να εκφραστούν ως γραμμικός συνδυασμός των φορέων δεδομένων.Αυτή η ιδιότητα των μοντέλων επιτρέπει μια αποτελεσματική και γρήγορη εκπαίδευση, καθώς ο υπολογισμός των βέλτιστων παραμέτρων απαιτεί μόνο γραμμικές πράξεις στα δεδομένα εκπαίδευσης.

**Kernels για την κατάταξη δεδομένων που λείπουν**

Η επιλογή μιας κατάλληλης συνάρτησης πυρήνα είναι ένα κύριο πρόβλημα στην εφαρμογή των μεθόδων πυρήνα. Αυτό είναι ιδιαίτερα σημαντικό στα πλαίσια όπου χαρακτηριστικά είναι απαραίτητα να αντιμετωπίσουν τιμές που λείπουν. Για να αντιμετωπιστεί αυτό το πρόβλημα, μπορούμε να εξετάσουμε τη συνάρτηση πυρήνα K(xi,ri,xj,rj)K(xi,ri,xj,rj), όπου οι δείκτες απόκρισης riri και rjrj συμπεριλαμβάνονται ρητά στη συνάρτηση πυρήνα.Η χρήση τέτοιου τύπου πυρήνα επιτρέπει στο μοντέλο να λαμβάνει υπόψη τη σχέση μεταξύ των χαρακτηριστικών και των αποκρίσεων τους. Αυτό είναι χρήσιμο όταν οι αποκρίσεις συνδέονται στενά με τα χαρακτηριστικά, όπως συχνά συμβαίνει σε πραγματικά σενάρια. Με αυτόν τον τρόπο, μπορούμε να αντιμετωπίσουμε την απώλεια δεδομένων και να εκτιμήσουμε πιο ακριβώς τα μοντέλα πρόβλεψης ή ταξινόμησης.Η επιλογή αυτού του είδους πυρήνα εξαρτάται σημαντικά από τα συγκεκριμένα δεδομένα και τη φύση του προβλήματος. Είναι σημαντικό να εξετάσουμε τη σχέση μεταξύ των χαρακτηριστικών και των αποκρίσεων προκειμένου να επιλέξουμε τον κατάλληλο πυρήνα που θα αντικατοπτρίζει αυτή τη σχέση με τον καλύτερο δυνατό τρόπο.

## 4.8 Ταξινόμηση Νευρωνικών Δικτύων και Ελλιπή Δεδομένα

Η διαδικασία συνδυασμού των στρατηγικών καταλογισμού και μειωμένων μοντέλων με πολυεπίπεδα νευρωνικά δίκτυα είναι παρόμοια με το πώς γίνεται με την λογιστική παλινδρόμηση. Συγκεκριμένα, εκπαιδεύουμε ξεχωριστά νευρωνικά δίκτυα για κάθε μοντέλο καταλογισμού και επιλέγουμε παραμέτρους συντονισμού ξεχωριστά για κάθε μοντέλο όταν συνδυάζεται με διασταυρούμενη επικύρωση.Αντίστοιχα, όταν συνδυάζουμε με μειωμένα μοντέλα, ορίζουμε ξεχωριστές παραμέτρους επικύρωσης για κάθε μοντέλο. Χρησιμοποιούμε όλα τα στιγμιότυπα δεδομένων με τις απαραίτητες παρατηρούμενες διαστάσεις κατά την εκπαίδευση κάθε μειωμένου μοντέλου, όπως και στην λογιστική παλινδρόμηση. Αυτή η προσέγγιση επιτρέπει την αποτελεσματική εκπαίδευση και συνδυασμό διαφορετικών μοντέλων με τη χρήση διαφορετικών στρατηγικών.

Όταν τα νευρωνικά δίκτυα συνδυάζονται με το πλαίσιο υποδείγματος απόκρισης, είναι παρόμοιο με την λογιστική παλινδρόμηση. Η προσθήκη του διανύσματος του δείκτη απόκρισης ως μέρος της αναπαράστασης εισόδου επηρεάζει μόνο το πρώτο κρυφό επίπεδο.

# 5. Ταξινόμηση (classification) και Εφαρμογές

Η ταξινόμηση περιλαμβάνει τη διαδικασία κατηγοριοποίησης ενός αντικειμένου σε μια συγκεκριμένη κλάση από ένα προκαθορισμένο σύνολο κλάσεων. Σε αντίθεση με την παλινδρόμηση, η ταξινόμηση ασχολείται με διακριτές τιμές-στόχους. Ο κύριος στόχος της ταξινόμησης είναι η ανάθεση δεδομένων σε ξεχωριστές κλάσεις με βάση συγκεκριμένα κριτήρια. Συνήθως, κάθε μοτίβο εκχωρείται μόνο σε μία κλάση, καθιστώντας τις τάξεις αμοιβαία αποκλειστικές. Σε πολλές εργασίες αναγνώρισης προτύπων, αντιμετωπίζεται ένα πρόβλημα ταξινόμησης δύο τάξεων όπου μια κλάση αντιπροσωπεύει μοτίβα που ικανοποιούν μια συγκεκριμένη συνθήκη (κλάση 1) ενώ η άλλη κλάση αντιπροσωπεύει μοτίβα που δεν ικανοποιούν αυτήν την συνθήκη (κλάση 0). Ουσιαστικά, η ετικέτα κλάσης είναι μια δυαδική μεταβλητή που υποδεικνύει την παρουσία ή την απουσία μιας συγκεκριμένης συνθήκης. Υπάρχουν διάφορες μέθοδοι για την κωδικοποίηση κλάσεων σε αριθμητικές τιμές.

Για την μέθοδο της ταξινόμησης θα παρουσιάσουμε δύο διαφορετικές προσεγγίσεις σχεδίασηςταξινομητών.

**Περιπτώσεις δύο κλάσεων**

Η απόσταση από ένα σημείο xx στο επίπεδο απόφασης y(x)=0y(x)=0 στο διάνυσμα βαρών ww είναι η κατάλληλη νόρμα του σημείου x ως εξής:

Απόσταση=|wTx|||w||

όπου ||w|| είναι η νόρμα του διανύσματος βαρών w. Αυτό είναι εφικτό επειδή το y(x)=wTx+w0 ορίζει το όριο απόφασης και το wTx είναι η εξίσωση ενός επιπέδου. Έτσι, η απόσταση είναι η κάθετη απόσταση από το σημείο x στο επίπεδο που ορίζεται από το w

Αν έχουμε ένα σημείο x και το x⊥ είναι η ορθογώνια προβολή του σημείου x στην επιφάνεια απόφασης, τότε ισχύει:

y(x)=wTx+w0

y(x⊥)=wTx⊥+w0=0

Πολλαπλασιάζοντας και τα δύο μέλη με wT, προσθέτοντας w0w0 και χρησιμοποιώντας τις παραπάνω σχέσεις, παίρνουμε:

wTx=−w0

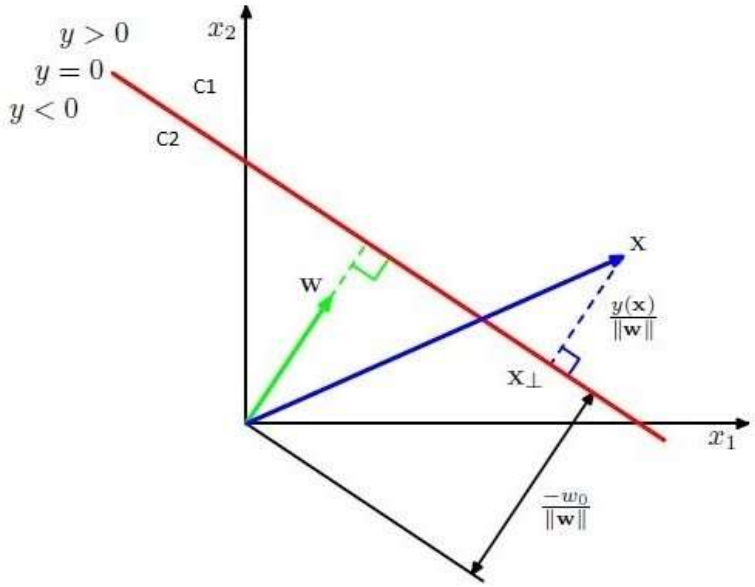
wTx⊥+w0=0

Παρατηρούμε ότι το wTx είναι η κατεύθυνση του xx ως προς το w, ενώ το wTx⊥ είναι η κατεύθυνση του x⊥ ως προς το w. Έτσι, το wTx και το wTx⊥ είναι κάθετα μεταξύ τους. Από τη γεωμετρία του εσωτερικού γινομένου δύο διανυσμάτων, γνωρίζουμε ότι:

wTx·wTx⊥=0

Άρα η σχέση wTx·wTx⊥=0 επιβεβαιώνει ότι τα δύο διανύσματα είναι κάθετα μεταξύ τους

μια γεωμετρική απεικόνιση των οποίων φαίνεται παρακάτω.




**Εικόνα 19 Απεικόνιση της γεωμετρίας δυο διαχωρίσιμων κλάσεων. Η επιφάνεια απόφασης φαίνεται με κόκκινο χρώμα , είναι κάθετη ως προς το w.**

**Πηγή:**
**https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/13644/Koursos_mwe1803.pdf?sequence=1&isAllowed=y**

**Περιπτώσεις πολλαπλών κλάσεων**

Στη μέθοδο που συζητήσαμε προηγουμένως, τα δεδομένα μας κατηγοριοποιήθηκαν μόνο σε δύο κατηγορίες. Ωστόσο, σε πολλές πραγματικές καταστάσεις, συνήθως υπάρχει ένας πολύ μεγαλύτερος αριθμός τάξεων. Είναι σημαντικό να κατανοήσουμε ότι ανεξάρτητα από τον αριθμό των κλάσεων, κάθε χαρακτηριστικό εκχωρείται μόνο σε μία κλάση. Αυτό δεν πρέπει να συγχέεται με την ταξινόμηση Η πολλαπλή ετικέτα είναι μια δυνατότητα που επιτρέπει την αντιστοίχιση πολλαπλών ετικετών σε ένα μόνο στοιχείο. Όταν αντιμετωπίζουμε προβλήματα ταξινόμησης πολλών κλάσεων, μια προσέγγιση είναι να αναλύσουμε το σύνολο δεδομένων σε πολλαπλά σύνολα δεδομένων δυαδικής ταξινόμησης και να εκπαιδεύσουμε ένα μοντέλο δυαδικής ταξινόμησης για το καθένα. Οι δύο πιο συνηθισμένες μέθοδοι για αυτό είναι το One-vs-Rest και το One-vs-One. Με το One-vs-Rest, το σύνολο δεδομένων πολλαπλών κλάσεων χωρίζεται σε πολλαπλά προβλήματα δυαδικής ταξινόμησης, με εκπαιδευμένο δυαδικό ταξινομητή για κάθε ένα και προβλέψεις που γίνονται χρησιμοποιώντας το καλύτερο μοντέλο. Από την άλλη πλευρά, το One-vs-One περιλαμβάνει τη δημιουργία ενός συνόλου δεδομένων για κάθε τάξη έναντι κάθε άλλης κλάσης.

Στο αρχικό σενάριο χρειάζονται ταξινομητές K-1, ενώ στην τελευταία περίπτωση είναι απαραίτητοι ταξινομητές K(k-1)/2 [6]. Ωστόσο, η πολυπλοκότητα της τεχνικής κλιμακώνεται όταν έχουμε να κάνουμε με εκτεταμένα σύνολα δεδομένων.

Στην πιο απλή περίπτωση μπορούμε να θεωρήσουμε K κλάσεις που περιλαμβάνουν γραμμικές συναρτήσεις της μορφής :

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

**Εξίσωση 22**

ένα χαρακτηριστικό ταξινομείται στην κλάση $C_k$ αν $y_k(x)$ $y_j(x)$ για κάθε $j =/ k$.

## 5.1 Μέθοδοι εξόρυξης δεδομένων και μηχανικής μάθησης για βιώσιμες έξυπνες πόλεις και ταξινόμηση δικτύου

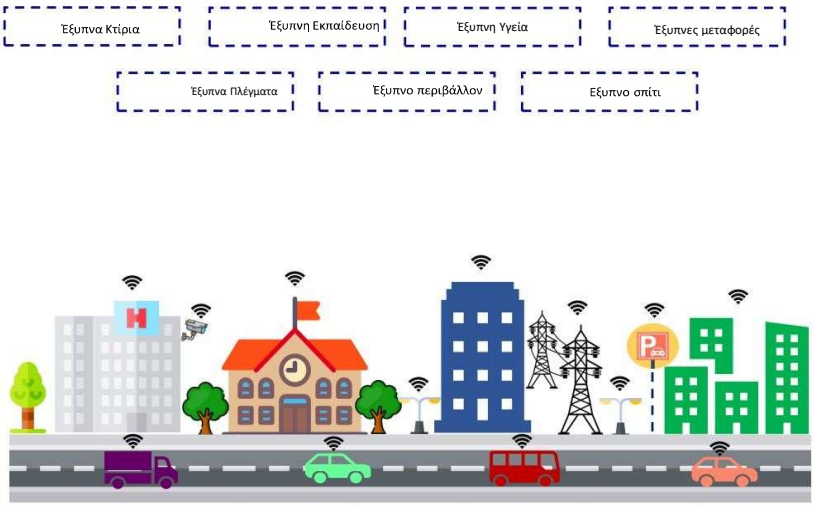
Οι εφαρμογές IoT είναι απαραίτητες για τη σύνδεση διαφόρων αντικειμένων, όπως αισθητήρες, ενεργοποιητές, έξυπνες συσκευές και συστήματα οικιακού αυτοματισμού, στο Διαδίκτυο με σκοπό τη μετάδοση και την επεξεργασία δεδομένων. Αυτή η τεχνολογία είναι ζωτικής σημασίας για την εφαρμογή πρωτοβουλιών έξυπνων πόλεων, οι οποίες στοχεύουν στη βελτίωση της ποιότητας ζωής των κατοίκων αξιοποιώντας τις προόδους στην τεχνολογία των επικοινωνιών. Οι έξυπνες πόλεις περιλαμβάνουν ένα ευρύ φάσμα καινοτόμων λύσεων σε διάφορους τομείς, συμπεριλαμβανομένων των έξυπνων κτιρίων, της υγειονομικής περίθαλψης, της εκπαίδευσης, των μεταφορών, των δικτύων και των κατοικιών, όλα προσανατολισμένα στην παροχή βολικών υπηρεσιών για τους πολίτες. Ωστόσο, οι διαφορετικές ανάγκες αυτών των εφαρμογών παρουσιάζουν προκλήσεις και πολυπλοκότητες στη διαχείριση δικτύων. Το Διαδίκτυο των Πραγμάτων (IoT) έχει αναδειχθεί ως μια σημαντική τεχνολογική πρόοδος που επηρεάζει πολλές πτυχές της καθημερινής μας ζωής. Η ευρεία επιρροή του μπορεί να φανεί στην ικανότητά του να οδηγεί την οικονομική ανάπτυξη, οδηγώντας σε σημαντικές επενδύσεις από εταιρείες τεχνολογίας και ερευνητικά ιδρύματα για την περαιτέρω ανάπτυξη και διερεύνηση λύσεων IoT.

Η ποικιλία των εφαρμογών σε ένα δίκτυο προσφέρει μια πληθώρα προκλήσεων όσον αφορά την ποιότητα της υπηρεσίας (QoS). Οι απαιτήσεις αυτές πρέπει να ληφθούν υπόψη κατά τον σχεδιασμό και τη λειτουργία του δικτύου, προκειμένου να εξασφαλιστεί η επίτευξη βέλτιστης απόδοσης και λειτουργικότητας. Ας εξετάσουμε μερικές από τις προκλήσεις που αντιμετωπίζονται σε διάφορες εφαρμογές:

- Βιντεοπαρακολούθηση και επείγουσες ανάγκες: Σε περιπτώσεις όπως η βιντεοπαρακολούθηση για την αντίδραση σε τροχαία ατυχήματα ή την ανίχνευση συμφορημένων δρόμων, οι ανάγκες εύρους ζώνης και χαμηλού jitter είναι κρίσιμες για τη μετάδοση αποτελεσμάτων σε πραγματικό χρόνο.
- Διαδικτυακά παιχνίδια και τηλεφωνία: Εδώ, η απρόσκοπτη αλληλεπίδραση απαιτεί χαμηλή καθυστέρηση και απώλεια πακέτων, καθώς οι καθυστερήσεις μπορούν να επηρεάσουν σοβαρά την εμπειρία του χρήστη.
- Έξυπνες πόλεις και ποικιλία στόχων: Οι εφαρμογές έξυπνων πόλεων, όπως ο έλεγχος της κυκλοφορίας, έχουν μοναδικές απαιτήσεις που πρέπει να ληφθούν υπόψη. Για παράδειγμα, ένα σύστημα διαχείρισης κυκλοφορίας θα χρειαστεί υψηλή αξιοπιστία και γρήγορη ανταπόκριση για να αποτρέψει ατυχήματα.

Για να αντιμετωπιστούν αυτές οι προκλήσεις, η σωστή σχεδίαση του δικτύου, η χρήση κατάλληλων πρωτοκόλλων επικοινωνίας και η διασφάλιση αξιοπιστίας και επίδοσης μέσω των μηχανισμών QoS είναι ζωτικής σημασίας. Επίσης, η σωστή διαχείριση του δικτύου και η παρακολούθηση της απόδοσης είναι απαραίτητες για τη διασφάλιση της ικανοποιητικής λειτουργίας των εφαρμογών.

Επομένως, αυτές οι προκλήσεις θα πρέπει να αντιμετωπιστούν σωστά.




**Εικόνα 20 Τυπική αρχιτεκτονική ενός δικτύου έξυπνων πόλεων.**

**Πηγή: https://www.mdpi.com/2624-6511/4/2/24**

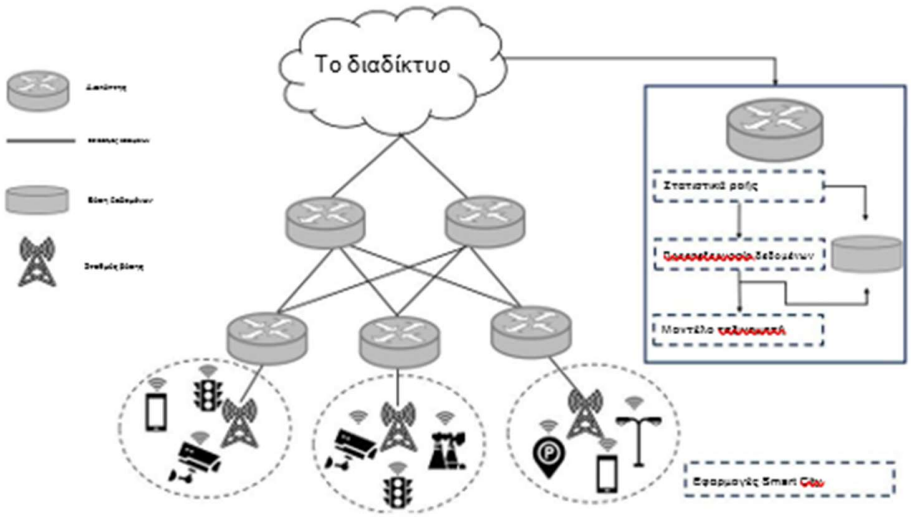
Στην αρχή της, η αρχιτεκτονική του Διαδικτύου δεν έδινε προτεραιότητα στις ανάγκες Ποιότητας Υπηρεσίας (QoS). Ο πρωταρχικός στόχος ήταν απλώς η αποτελεσματική παράδοση δεδομένων. Ωστόσο, έχουν γίνει προσπάθειες για την αντιμετώπιση των απαιτήσεων QoS μέσω της ανάπτυξης Ολοκληρωμένων Υπηρεσιών και Διαφοροποιημένων Υπηρεσιών. Το Integrated Services στοχεύει στην εγγύηση QoS τόσο για εφαρμογές πολλαπλής διανομής όσο και για εφαρμογές unicast, δεσμεύοντας πόρους δικτύου και διατηρώντας την κατάσταση ανά ροή σε κάθε δρομολογητή. Ενώ αυτή η προσέγγιση προσθέτει πολυπλοκότητα και μπορεί να επηρεάσει την επεκτασιμότητα του δικτύου, οι Διαφοροποιημένες Υπηρεσίες αμβλύνουν αυτές τις ανησυχίες ομαδοποιώντας τις ροές κυκλοφορίας σε κλάσεις QoS χρησιμοποιώντας το πεδίο Κώδικας διαφοροποιημένου σημείου υπηρεσίας. Παρά τις προσπάθειες αυτές, καμία προσέγγιση δεν έχει υιοθετηθεί ευρέως σε δίκτυα μεγάλης κλίμακας. Η ταξινόμηση της κυκλοφορίας διαδραματίζει κρίσιμο ρόλο σε διάφορες εφαρμογές δικτύου, συμπεριλαμβανομένης της παρακολούθησης, της διαχείρισης QoS και της ασφάλειας. Επιτρέπει τη διαφοροποίηση των ροών κίνησης με βάση τον τύπο εφαρμογής τους, όπως η διάκριση μεταξύ VoIP και κανονικής ροής δεδομένων. Αυτή η ταξινόμηση διευκολύνει την κατανομή των κατάλληλων πόρων, όπως το εύρος ζώνης και η καθυστέρηση, για την κάλυψη των ειδικών απαιτήσεων κάθε εφαρμογής, διασφαλίζοντας αξιόπιστη υποστήριξη για QoS. Υπάρχουν πολλές τεχνικές για την ταξινόμηση της κυκλοφορίας που δεν απαιτούν τροποποίηση της κεφαλίδας TCP/IP. Η βαθιά επιθεώρηση πακέτων χρησιμεύει ως εναλλακτική μέθοδος για την ταξινόμηση της κυκλοφορίας, που περιλαμβάνει την ανάλυση των ωφέλιμων φορτίων πακέτων για τον εντοπισμό προκαθορισμένων μοτίβων ή υπογραφών. Αν και αυτή η τεχνική είναι αποτελεσματική σε πολλές περιπτώσεις, παρουσιάζει προκλήσεις που

σχετίζονται με το απόρρητο, καθώς η επιθεώρηση μπορεί να αποκαλύψει ευαίσθητες πληροφορίες. Επιπλέον, η βαθιά επιθεώρηση πακέτων δυσκολεύεται να χειριστεί κρυπτογραφημένα δεδομένα και μπορεί να είναι υπολογιστικά απαιτητική, καθιστώντας την λιγότερο κατάλληλη για ορισμένες εφαρμογές.

Η ταξινόμηση της κυκλοφορίας με χρήση αλγορίθμων μηχανικής μάθησης έχει γίνει αντικείμενο αυξανόμενου ερευνητικού ενδιαφέροντος λόγω των προοπτικών για αυξημένη ακρίβεια και αποτελεσματικότητα στη διαχείριση της κυκλοφορίας δεδομένων. Η διαδικασία αυτή συνήθως περιλαμβάνει τα εξής βασικά βήματα:

- Ορισμός Χαρακτηριστικών Κίνησης: Αρχικά, τα χαρακτηριστικά κίνησης προσδιορίζονται και επιλέγονται για να αντιπροσωπεύουν διάφορα χαρακτηριστικά της ροής δεδομένων, όπως το μέγεθος του πακέτου, ο ρυθμός μετάδοσης και άλλα σχετικά.
- Εκπαίδευση Μοντέλου Μηχανικής Μάθησης: Στη συνέχεια, αναπτύσσεται και εκπαιδεύεται ένα μοντέλο μηχανικής μάθησης, όπως ένα νευρωνικό δίκτυο ή ένας ταξινομητής SVM, χρησιμοποιώντας τα προαναφερθέντα χαρακτηριστικά για να αναγνωρίσει πρότυπα και σχέσεις στα δεδομένα εκπαίδευσης.
- Ταξινόμηση Κυκλοφορίας Δεδομένων: Το εκπαιδευμένο μοντέλο χρησιμοποιείται στη συνέχεια για την ταξινόμηση της κυκλοφορίας δεδομένων, προβλέποντας τις κατηγορίες εντός της ροής κυκλοφορίας, όπως οι τύποι εφαρμογών ή οι κατηγορίες κίνησης.

Η χρήση αυτών των τεχνικών στη βελτίωση της Ποιότητας Υπηρεσίας (QoS) σε έξυπνα δίκτυα είναι ένα παράδειγμα εφαρμογής που έχει σημαντική σημασία για τη διαχείριση της κυκλοφορίας και τη βελτίωση της απόδοσης των δικτύων σε ποικίλες συνθήκες όπως φαίνεται στην επόμενη εικόνα.




**Εικόνα 21 Τυπικό σενάριο που απεικονίζει την ταξινόμηση της κυκλοφορίας με βάση τη μηχανική εκμάθηση σε ένα δίκτυο έξυπνων πόλεων**

Πολυάριθμες ερευνητικές μελέτες έχουν διερευνήσει την αποτελεσματικότητα της μηχανικής μάθησης στην ταξινόμηση της κυκλοφορίας σε διαφορετικά σύνολα δεδομένων και υποδομές δικτύου. Ορισμένοι έχουν διερευνήσει τη χρήση της μηχανικής εκμάθησης για την ταξινόμηση της κυκλοφορίας και τη βελτιστοποίηση της ποιότητας των υπηρεσιών σε εφαρμογές έξυπνων πόλεων σε διάφορα επίπεδα δικτύου. Για παράδειγμα, οι Aureli et al. ανέπτυξε μια τεχνική δυναμικής

ταξινόμησης που ονομάζεται διαφοροποιημένες υπηρεσίες βασισμένες στη μάθηση, η οποία εκχωρεί δυναμικά κλάσεις υπηρεσιών σε πακέτα IP με βάση τα χαρακτηριστικά κίνησης που προσδιορίζονται μέσω μεθόδων μηχανικής μάθησης, όπως η ανάλυση γραμμικής διάκρισης και η ομαδοποίηση k-means. Επικεντρώθηκαν στην αντιμετώπιση ζητημάτων όπως η άνιση κατανομή της κυκλοφορίας μεταξύ των τάξεων. Ενώ η προσέγγισή μας και των Aureli et al. μοιραζόμαστε έναν κοινό στόχο ταξινόμησης της κυκλοφορίας, διαφέρουμε ως προς τη χρήση των εποπτευόμενων αλγορίθμων μηχανικής εκμάθησης για την ταξινόμηση της κυκλοφορίας δικτύου σε ένα ευρύτερο φάσμα κλάσεων.

Στη μελέτη τους, ο Zhongsheng και οι συνεργάτες του εισήγαγαν μια Μηχανή Διανυσμάτων Υποστήριξης (SVM) ως μέθοδο ταξινόμησης της κυκλοφορίας δικτύου μέσα στα δίκτυα της πανεπιστημιούπολης. Συλλέγοντας δεδομένα και δημιουργώντας χαρακτηριστικά, χρησιμοποίησαν επιτυχώς το SVM για να επιτύχουν υψηλά επίπεδα ακρίβειας, με τα αποτελέσματα να δείχνουν ποσοστά ακρίβειας 99,31% και 96,12% όταν ελέγχθηκαν τόσο με προκατειλημμένα όσο και με αμερόληπτα δείγματα. Ωστόσο, οι ερευνητές εστίασαν αποκλειστικά στο SVM και δεν εξερεύνησαν άλλους αλγόριθμους μηχανικής μάθησης, παρά το γεγονός ότι η επίτευξη υψηλής ακρίβειας δεν είναι πάντα ο πρωταρχικός στόχος. Στην πραγματικότητα, για εφαρμογές σε πραγματικό χρόνο, η ελαχιστοποίηση της καθυστέρησης είναι συχνά πιο κρίσιμη από τη μεγιστοποίηση της ακρίβειας. Ως εκ τούτου, είναι σημαντικό να ληφθούν υπόψη και οι χρόνοι εκτέλεσης διάφορων αλγορίθμων μηχανικής μάθησης προκειμένου να καθοριστεί η καταλληλότερη προσέγγιση για ένα δεδομένο σενάριο.

Ο Al-Turjman επικεντρώθηκε στην αντιμετώπιση του ζητήματος της ασύρματης συνδεσιμότητας σε περιβάλλοντα που κινούνται γρήγορα σε έξυπνες πόλεις. Το προκύπτον πλαίσιο χρησιμοποιεί την τεχνολογία LTE για τη βελτίωση της ποιότητας των υπηρεσιών για κινητές εφαρμογές, ενώ παράλληλα μειώνει την εμφάνιση καθυστερήσεων και σφαλμάτων κατά τη μεταφορά δεδομένων σε πραγματικό χρόνο. Για την αξιολόγηση διαφορετικών πτυχών της ποιότητας της υπηρεσίας, το πλαίσιο ενσωματώνει μια Markovian διαδικασία που βασίζεται στο πρότυπο IEEE 802.16, εξετάζοντας συγκεκριμένα παράγοντες όπως η μέση καθυστέρηση πακέτων. Επιπρόσθετα, γίνεται πρόταση για τον σχεδιασμό κινητών οχημάτων cloud, λαμβάνοντας υπόψη διάφορους παράγοντες όπως τα κυκλοφοριακά μοτίβα και τις καιρικές συνθήκες. Αυτά τα οχήματα θα χρησιμοποιούν την υπάρχουσα υποδομή κινητής τηλεφωνίας για ροή δεδομένων και βίντεο. Ωστόσο, η πρόταση δεν εξετάζει τα πιθανά οφέλη από την ενσωμάτωση τεχνικών μηχανικής μάθησης, οι οποίες θα μπορούσαν ενδεχομένως να οδηγήσουν σε βέλτιστες και αποτελεσματικότερες διαδικασίες λήψης αποφάσεων.

Στη μελέτη τους, οι Yao et al. πρότειναν μια νέα μέθοδο ταξινόμησης της κυκλοφορίας ειδικά σχεδιασμένη για έξυπνα δίκτυα πόλεων. Χρησιμοποίησαν τεχνικές βαθιάς μάθησης, ιδιαίτερα ένα μοντέλο δικτύου κάψουλας, για να επιτύχουν αποτελεσματική ταξινόμηση. Ο κύριος στόχος της προτεινόμενης μεθόδου τους ήταν να εξαλείψει την ανάγκη για χειροκίνητη επιλογή των χαρακτηριστικών κυκλοφορίας του δικτύου. Από την άλλη πλευρά, η προσέγγισή αυτή εστιάζει στη βελτίωση της επιλογής χαρακτηριστικών με τη χρήση τεσσάρων εποπτευόμενων αλγορίθμων μηχανικής εκμάθησης. Ο απώτερος στόχος είναι η βελτίωση της ποιότητας των υπηρεσιών (QoS) στα έξυπνα δίκτυα πόλεων μέσω της επιτυχούς ταξινόμησης της κίνησης του δικτύου. Σε παρόμοιο πνεύμα, οι Miao et al. διεξήγαγαν μια συγκριτική ανάλυση έξι αλγορίθμων μηχανικής μάθησης για την ταξινόμηση της κυκλοφορίας. Αυτοί οι αλγόριθμοι περιελάμβαναν Naive Bayes, Random Forest (RF), Support Vector Machine (SVM), H2O, K-Nearest Neighbors (KNN) και Decision Tree (DT). Για την εξαγωγή χαρακτηριστικών, χρησιμοποίησαν ανάλυση κύριων συστατικών και αξιολόγησαν τον αντίκτυπό της στα αποτελέσματα ταξινόμησης. Τα πειραματικά ευρήματα αποκάλυψαν ότι οι RF και KNN ξεπέρασαν τους άλλους αλγόριθμους όσον αφορά τη συνολική απόδοση. Χωρίς τη συμπερίληψη της ανάλυσης του κύριου στοιχείου, το RF πέτυχε ακρίβεια 92,92% ενώ το KNN πέτυχε 84,56%. Ωστόσο, όταν χρησιμοποιήθηκαν αλγόριθμοι ταξινόμησης της κυκλοφορίας, η ακρίβεια βελτιώθηκε σημαντικά, φτάνοντας το 99,08% για το RF και το 97,16% για το KNN. Αξίζει να σημειωθεί ότι παρόλο που το σύνολο δεδομένων μας αποτελούνταν από κίνηση δεδομένων πανεπιστημιούπολης, ενώ το σύνολο δεδομένων τους επικεντρώθηκε στην κίνηση δεδομένων ISP,

ελήφθησαν υπόψη και οι δύο τύποι κίνησης δικτύου κορμού. Ως εκ τούτου, υπάρχουν ομοιότητες στην κίνηση δεδομένων που αναλύθηκαν και στις δύο μελέτες.

Στη μελέτη τους, οι Perera et al. διεξήγαγαν μια σύγκριση έξι διαφορετικών εποπτευόμενων αλγορίθμων μάθησης με σκοπό την ταξινόμηση στον τομέα της ανάλυσης κυκλοφορίας. Αυτοί οι αλγόριθμοι περιελάμβαναν Naive Bayes, Bayesian network, Random Forest, Decision Tree, Naive Bayes tree και multilayer perceptron. Οι ερευνητές πραγματοποίησαν μια σειρά πειραμάτων χρησιμοποιώντας δύο διαφορετικές τεχνικές επιλογής χαρακτηριστικών και κατηγοριοποίησαν τα δεδομένα σε πέντε διαφορετικές κατηγορίες κυκλοφορίας.

Τα ευρήματα έδειξαν ότι οι αλγόριθμοι RF και DT επέδειξαν ανώτερη ακρίβεια ταξινόμησης, επιτυγχάνοντας μέσο ποσοστό ακρίβειας 96% και 95% αντίστοιχα. Ο Rahman και οι συνεργάτες του παρουσίασαν ένα πλαίσιο ρομποτικής cloud σχεδιασμένο για χρήσεις έξυπνων πόλεων. Σε αυτό το πλαίσιο, ένα ρομπότ χρησιμοποιεί υπηρεσίες cloud αναθέτοντας εργασίες για τη βελτίωση της ποιότητας της υπηρεσίας (QoS) και της συνολικής απόδοσης του συστήματος. Καθιέρωσαν μια πρόκληση βελτιστοποίησης για ένα κατευθυνόμενο κυκλικό γράφημα, το οποίο στη συνέχεια επιλύθηκε χρησιμοποιώντας έναν γενετικό αλγόριθμο για τον προσδιορισμό των πιο αποτελεσματικών αποφάσεων φόρτωσης εργασιών.

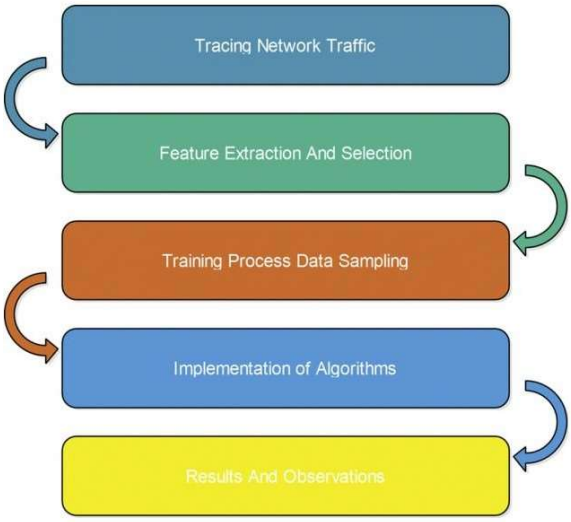
Συνοψίζοντας, διάφοροι αλγόριθμοι μηχανικής μάθησης έχουν χρησιμοποιηθεί για την αξιολόγηση της αποτελεσματικότητας του ταξινομητή σε σχέση με εποπτευόμενους αλγόριθμους. Επιπλέον, έχει διεξαχθεί εκτενής έρευνα για τεχνικές βαθιάς μάθησης, με αποτέλεσμα την πρόταση πολλαπλών προσεγγίσεων για τη βελτίωση της ποιότητας των υπηρεσιών (QoS) σε δίκτυα έξυπνων πόλεων. Ξεχωρίζοντας από προηγούμενες μελέτες, η έρευνά αυτή περιλαμβάνει μια ολοκληρωμένη διερεύνηση της απόδοσης των εποπτευόμενων αλγορίθμων ταξινόμησης, δηλαδή Υποστήριξη Διανυσματικών Μηχανών (SVM), Τυχαίου Δάσους (RF), Κ-Πλησιότερους Γείτονες (KNN) και Δέντρα Αποφάσεων (DT), με ο στόχος της ενίσχυσης της QoS σε δίκτυα έξυπνων πόλεων και της ακριβούς ταξινόμησης της κυκλοφορίας του δικτύου με βάση στατιστικά χαρακτηριστικά.

## 5.2 Τεχνικές ταξινόμησης

### 5.2.1 Τεχνική ταξινόμησης κυκλοφορίας βάσει θυρών

Το Πρωτόκολλο Ελέγχου Μετάδοσης (TCP) και το Πρωτόκολλο Δεδομένων Χρήστη (UDP) είναι δύο ευρέως χρησιμοποιούμενα πρωτόκολλα που επιτρέπουν πολλαπλές ροές επικοινωνίας χρησιμοποιώντας συγκεκριμένους αριθμούς θυρών μεταξύ κοινών τελικών σημείων IP. Σε μια ολοκληρωμένη μελέτη που διεξήχθη από τους Nguyen και Armitage το 2008, ανακαλύφθηκε ότι πολλές εφαρμογές επιλέγουν μια «γνωστή» θύρα για να διευκολύνουν τις τοπικές επικοινωνίες κεντρικού υπολογιστή τους. Αυτό σημαίνει ότι αυτές οι εφαρμογές χρησιμοποιούν με συνέπεια έναν συγκεκριμένο αριθμό θύρας για τις ανάγκες επικοινωνίας τους. Με απλούστερους όρους, όταν μια εφαρμογή δικτύου επιθυμεί να δημιουργήσει μια σύνδεση, δηλώνει πρώτα τον συγκεκριμένο αριθμό θύρας της όπως αναφέρεται στο μητρώο του Internet Assigned Numbers Authority (IANA). Στη συνέχεια, η υποδομή δικτύου κατηγοριοποιεί και κατευθύνει αποτελεσματικά την εισερχόμενη κίνηση με βάση τον καταχωρημένο αριθμό θύρας. Αυτό διασφαλίζει ότι δημιουργούνται και διατηρούνται οι κατάλληλες ροές επικοινωνίας για ομαλή και αποτελεσματική μετάδοση δεδομένων. Για την αποτελεσματική παρακολούθηση και ανάλυση της κυκλοφορίας του δικτύου, ένας ταξινομητής δικτύου διαθέτει τη δυνατότητα να εξετάζει πακέτα TCP SYN. Με αυτόν τον τρόπο, μπορεί να αναγνωρίσει τη δημιουργία μιας νέας σύνδεσης TCP διακομιστή-πελάτη στην πλευρά της υπηρεσίας. Το πακέτο TCP SYN παίζει κρίσιμο ρόλο στην έναρξη μιας συνεδρίας μέσω μιας τριμερής χειραψίας, όπως εξηγήθηκε διεξοδικά από τους Abbasi, Ezzati-Jivan, Bellaiche, Talhi και Dagenais το 2019. Αυτή η διαδικασία χειραψίας περιλαμβάνει την ανταλλαγή συγκεκριμένων μηνυμάτων ελέγχου μεταξύ των πελάτη και διακομιστή, οδηγώντας τελικά στη δημιουργία μιας σταθερής και ασφαλούς σύνδεσης. [44]

Ένα παράδειγμα αυτού είναι η εφαρμογή email που χρησιμοποιεί το Simple Mail Transfer Protocol (SMTP) για την αποστολή email. Λειτουργεί στον κοινά αναγνωρισμένο αριθμό θύρας 25 ενώ λαμβάνει email από το Ταχυδρομείο.




**Εικόνα 22 Μοντέλο ταξινόμησης κίνησης δικτύου**

Το πρωτόκολλο POP3 χρησιμοποιεί τον αριθμό θύρας 110, ενώ οι εφαρμογές Ιστού χρησιμοποιούν συνήθως τον αριθμό θύρας 80. Οι τεχνικές που βασίζονται σε θύρες είναι εξαιρετικά πολύτιμες και αποτελεσματικές για την ταξινόμηση και τον εντοπισμό εφαρμογών δικτύου μέσα σε ένα δίκτυο μεγάλης κυκλοφορίας. Ωστόσο, αυτές οι τεχνικές έχουν ορισμένους περιορισμούς, ιδιαίτερα όταν οι εφαρμογές χρησιμοποιούν δυναμικούς αριθμούς θυρών για τις επικοινωνίες τους, παρόμοια με τις εφαρμογές P2P. Πολλές εφαρμογές δεν βασίζονται σε σταθερούς αριθμούς θυρών, αλλά χρησιμοποιούν δυναμικούς αριθμούς θυρών, όπως P2P, Napster και Kazaa. Οι αριθμοί δυναμικής θύρας είναι ουσιαστικά μη καταχωρημένοι στην Αρχή Εκχωρημένων Αριθμών Διαδικτύου (IANA). Στην έρευνά τους, οι Moore και Papagiannaki βρήκαν ότι η προσέγγιση με βάση τα λιμάνια για την αναγνώριση της κυκλοφορίας δυσκολεύεται να επιτύχει ποσοστά ακρίβειας που ξεπερνούν το 70% και στερείται αποτελεσματικής ακρίβειας ταξινόμησης (Li et al., 2018; Moore & Papagiannaki, 2005). Madhukar and Sen et al. (2004) εξέτασε επίσης τους περιορισμούς αυτής της μεθόδου που βασίζεται σε θύρες. Ένας από τους κύριους λόγους για τις ελλείψεις της τεχνικής που βασίζεται σε θύρες είναι η εξάρτησή της από δυναμικούς αριθμούς θυρών, κάτι που τελικά εμποδίζει την αποτελεσματικότητά της.

### 5.2.2 Τεχνική ταξινόμησης κυκλοφορίας βάσει ωφέλιμου φορτίου

Οι ερευνητές Καραγιάννης κ.α. (2004) και Sen et al. (2004) διεξήγαγαν μελέτες για την ανάλυση των υπογραφών κυκλοφορίας σε επίπεδο εφαρμογής στην ταξινόμηση κυκλοφορίας P2P ωφέλιμου φορτίου. Τα ευρήματά τους αποκάλυψαν ότι αυτή η προσέγγιση μπορεί να μειώσει τα ψευδώς θετικά και τα ψευδώς αρνητικά στην αναγνώριση της κυκλοφορίας P2P κατά 5%. Ομοίως, οι Moore και Papagiannaki (2005) χρησιμοποίησαν τεχνικές που βασίζονται σε λιμάνι και ωφέλιμο φορτίο για να προσδιορίσουν με ακρίβεια διάφορες εφαρμογές δικτύου. Προκειμένου να αντιμετωπιστεί το πρόβλημα με την τεχνική που βασίζεται στη θύρα, έχει εισαχθεί μια νέα προσέγγιση γνωστή ως τεχνική Deep Packet Inspection (DPI). Αυτή η τεχνική περιλαμβάνει την ανάλυση του περιεχομένου των πακέτων για τον εντοπισμό των μοναδικών χαρακτηριστικών της κίνησης εφαρμογών δικτύου. Το DPI θεωρείται ως εναλλακτική λύση στην τεχνική που βασίζεται στο λιμάνι και είναι η δεύτερη μέθοδος που χρησιμοποιείται για το σκοπό αυτό. Ωστόσο,

αναπτύχθηκε ειδικά για την ταξινόμηση της κυκλοφορίας εφαρμογών P2P που χρησιμοποιεί δυναμικούς αριθμούς θύρας αντί για σταθερό αριθμό θύρας.

Η διαδικασία διαλογής τους ξεκινά με την ανάλυση της ροής με βάση τον αριθμό θύρας. Εάν δεν αναγνωριστεί ο γνωστός αριθμός θύρας, η κίνηση μεταφέρεται στο επόμενο βήμα. Το δεύτερο βήμα περιλαμβάνει τον έλεγχο για την παρουσία μιας υπογραφής εντός της ροής. Εάν δεν υπάρχει άλλη τεχνική ανίχνευσης, το πακέτο εξετάζεται για τυχόν πρωτόκολλα που ήδη υπάρχουν. Εάν και οι δύο από αυτές τις δοκιμές αποτύχουν, ελέγχεται εξονυχιστικά το πρώτο Kbyte υπογραφών πρωτοκόλλου. Τα ευρήματά τους δείχνουν ότι οι πληροφορίες θύρας μπορούν να ταξινομήσουν με επιτυχία περίπου το 69% της κίνησης στο Διαδίκτυο με βάση τα συνολικά byte. Επιπλέον, η χρήση πληροφοριών από την αρχική ροή μπορεί να αυξήσει την ακρίβεια ταξινόμησης σε περίπου 79%. Ωστόσο, αυτή η μέθοδος δεν είναι ιδανική για την ταξινόμηση της κίνησης στο Διαδίκτυο, καθώς απαιτεί ακριβό εξοπλισμό για την παρακολούθηση των μοτίβων ωφέλιμου φορτίου. Επιπλέον, δεν ταξινομεί αποτελεσματικά την κίνηση κρυπτογραφημένων εφαρμογών και απαιτεί συνεχείς ενημερώσεις στα μοτίβα υπογραφής για νέες εφαρμογές.

### 5.2.3Ταξινόμηση βάσει στατιστικών ιδιοτήτων κυκλοφορίας

Προκειμένου να αντιμετωπιστούν οι προκλήσεις που δημιουργούνται από ζητήματα που βασίζονται σε θύρες και ωφέλιμο φορτίο, έχει εισαχθεί μια νέα μέθοδος ταξινόμησης που βασίζεται στις στατιστικές ιδιότητες της κίνησης δικτύου για την κατηγοριοποίηση και τον εντοπισμό διαφορετικών εφαρμογών. Αυτά τα στατιστικά χαρακτηριστικά, όπως η διάρκεια ροής, το μήκος του πακέτου, ο χρόνος μεταξύ άφιξης πακέτων και η αδράνεια ροής, παίζουν καθοριστικό ρόλο στη διάκριση μεταξύ των διαφόρων τύπων εφαρμογών σε ένα δίκτυο. Και οι δύο μελέτες είχαν στόχο να παρέχουν μια ολοκληρωμένη κατανόηση των προτύπων κυκλοφορίας και των επιπτώσεών τους στην απόδοση του δικτύου. Εξετάζοντας διαφορετικές πτυχές των χαρακτηριστικών της κυκλοφορίας, αυτοί οι ερευνητές προσπάθησαν να συμβάλουν στην ανάπτυξη μοντέλων και συστημάτων που μπορούν να χειριστούν καλύτερα και να ανταποκριθούν στις συγκεκριμένες απαιτήσεις διαφόρων εφαρμογών και παιχνιδιών. Ομοίως, οι Lang, Armitage, Branch και Choo (2003) διεξήγαγαν μια μελέτη με έμφαση στην ανάπτυξη ενός συστήματος μοντέλου κυκλοφορίας ειδικά για το διαδικτυακό παιχνίδι Half-Life. Στόχος τους ήταν να διερευνήσουν τον πιθανό μελλοντικό αντίκτυπο αυτού του διαδικτυακού παιχνιδιού στα δίκτυα IP. Ανέλυσαν διάφορες πτυχές της κίνησης του παιχνιδιού, όπως το μήκος των πακέτων, τους χρόνους άφιξης μεταξύ των πακέτων και τους μέσους ρυθμούς πακέτων και byte που μεταδίδονται ανά δευτερόλεπτο. Σε μια πρόσφατη μελέτη που διεξήχθη από τον Paxson (2020), πραγματοποιήθηκε μια εξέταση σχετικά με τη συσχέτιση μεταξύ των προτύπων κυκλοφορίας και των στατιστικών τους χαρακτηριστικών. Οι ερευνητές στόχευσαν να δημιουργήσουν ένα εμπειρικό μοντέλο που θα μπορούσε να περιγράφει αποτελεσματικά τα χαρακτηριστικά των συνδέσεων, συμπεριλαμβανομένων παραγόντων όπως τα byte, οι διάρκειες και η περιοδικότητα άφιξης για συγκεκριμένα πακέτα TCP. Ωστόσο, είναι σημαντικό να σημειωθεί ότι το εύρος του μοντέλου τους περιοριζόταν σε ορισμένες συνδέσεις εφαρμογών όπως TELNET, FTP, SMTP και NNTP.

### 5.2.4 Σύνοψη

Οι κρίσιμες διαδικασίες που εμπλέκονται στην κατηγοριοποίηση της κίνησης στο Διαδίκτυο περιγράφονται παρακάτω.

•**Network Traffic Capturing** :

Ένα από τα κρίσιμα και ουσιαστικά βήματα στη διαδικασία είναι η παρακολούθηση της ροής της κυκλοφορίας του δικτύου προκειμένου να δημιουργηθούν ολοκληρωμένα σύνολα δεδομένων για ακριβή κατηγοριοποίηση της κυκλοφορίας. Κατά τη διάρκεια αυτού του σταδίου, χρησιμοποιούνται διάφορες εφαρμογές sniffer όπως Wireshark, tcpdump, Snort και Nmap για τον εντοπισμό και την ανάλυση της πραγματικής κίνησης στο Διαδίκτυο (Orebaugh, Ramirez, & Beale, 2019; Jacobson, Leres, & McCanne, 2020; Lyon, 2009).

•**Εξαγωγή και επιλογή χαρακτηριστικών**:

Αυτό το στάδιο περιλαμβάνει την κρίσιμη διαδικασία εξαγωγής και αναγνώρισης χαρακτηριστικών που θα χρησιμοποιηθούν στο μοντέλο μηχανικής μάθησης. Χωρίς την κατάλληλη εξαγωγή χαρακτηριστικών, η

μέθοδος μηχανικής εκμάθησης δεν μπορεί να εφαρμοστεί αποτελεσματικά για την αναγνώριση μοτίβων κίνησης δικτύου. Αυτή η μέθοδος περιλαμβάνει τη δημιουργία χαρακτηριστικών από τα δεδομένα που συλλέγονται κατά τον εντοπισμό κυκλοφορίας, όπως το μέγεθος, το μήκος και η διάρκεια του πακέτου. Αυτά τα εξαγόμενα χαρακτηριστικά χρησιμοποιούνται στη συνέχεια για την εκπαίδευση των ταξινομητών μηχανικής μάθησης. Για την εξαγωγή χαρακτηριστικών από τα δεδομένα κίνησης, διάφορες εφαρμογές λογισμικού είναι διαθέσιμες στο διαδίκτυο για δημόσια χρήση. Για παράδειγμα, ένα σενάριο Perl μπορεί να χρησιμοποιηθεί για τη δημιουργία ή εξαγωγή χαρακτηριστικών, ενώ ένα άλλο εργαλείο γνωστό ως NetMate μπορεί να εξαγάγει ένα ευρύ φάσμα στατιστικών χαρακτηριστικών, έως και 44 συνολικά, από τη ροή δεδομένων κίνησης (Dupuy, Sengupta, Wolfson και Yemini , 1991).

•**Διαδικασία εκπαίδευσης ή δειγματοληψία**:

Σε αυτό το στάδιο, το σύνολο δεδομένων υφίσταται δειγματοληψία μέσω μιας μεθόδου εποπτευόμενης μάθησης. Τα δείγματα αναλύονται πρώτα για να εντοπιστεί οποιαδήποτε άγνωστη επισκεψιμότητα κατηγορίας. Ουσιαστικά, αυτό το βήμα περιλαμβάνει τη διαίρεση του συνόλου δεδομένων σε σύνολα εκπαίδευσης και δοκιμών προκειμένου να εφαρμοστούν ταξινομητές μηχανικής μάθησης (ML). Ενώ κάθε βήμα στη διαδικασία είναι κρίσιμο, αυτό το συγκεκριμένο βήμα έχει μεγάλη σημασία, καθώς είναι απαραίτητο για την ακριβή ταξινόμηση ή αναγνώριση της κυκλοφορίας στο Διαδίκτυο χωρίς τη διαδικασία δειγματοληψίας των δεδομένων εκπαίδευσης και δοκιμής.

•**Υλοποίηση αλγορίθμων ML**:

Σε αυτό το βήμα, το σύνολο δεδομένων για εκπαίδευση και τα σύνολα δοκιμών διεξάγονται χρησιμοποιώντας επιλεγμένους ταξινομητές ML.

•**Αποτελέσματα και Παρατήρηση**:

Μόλις εκτελεστούν οι ταξινομητές ML, το εργαλείο εφαρμογής προσομοίωσης προσφέρει ολοκληρωμένα αποτελέσματα και δεδομένα, όπως στατιστικά ακριβείας, διάρκεια δοκιμών και εκπαίδευσης και ποσοστά ανάκλησης (Shafiq et al., 2016), μεταξύ άλλων.

## 5.3 Σύνολα δεδομένων για ταξινόμηση κυκλοφορίας

Τα δεδομένα διαδραματίζουν κρίσιμο ρόλο στην ταξινόμηση της κίνησης στο Διαδίκτυο. Χωρίς δεδομένα, καθίσταται αδύνατη η αποτελεσματική ταξινόμηση και κατανόηση της αναγνώρισης κίνησης στο Διαδίκτυο. Για να αποκτήσουμε μια πλήρη κατανόηση αυτής της διαδικασίας, είναι απαραίτητο να εμβαθύνουμε στα δεδομένα που χρησιμοποιούνται στο δίκτυο μέσω μεθόδων μηχανικής μάθησης (ML). Επιπλέον, πριν ξεκινήσουμε την τεχνική ταξινόμησης ή αναγνώρισης που χρησιμοποιεί τεχνικές ML ή DM, είναι επιτακτική ανάγκη να αναγνωρίσουμε ότι διάφοροι ερευνητές χρησιμοποιούν διαφορετικούς τύπους συνόλων δεδομένων για τις αντίστοιχες μελέτες τους. Ωστόσο, αυτή η ενότητα θα παρέχει μια λεπτομερή περιγραφή των συνόλων δεδομένων που χρησιμοποιούνται, εστιάζοντας ειδικά στην τεχνική ταξινόμησης της κυκλοφορίας δικτύου χρησιμοποιώντας ταξινομητές μηχανικής μάθησης.

### 5.3.1 Εργαλεία ανίχνευσης κυκλοφορίας για την καταγραφή της κυκλοφορίας

Ο πρωταρχικός στόχος στη σφαίρα της ταξινόμησης της κυκλοφορίας δικτύου είναι η ακριβής παρακολούθηση και τεκμηρίωση των μοτίβων κυκλοφορίας προκειμένου να διαφυλαχθεί το απόρρητο των χρηστών και οι ευαίσθητες πληροφορίες. Μια θεμελιώδης εργασία από τους Allman και Paxson (2007) υπογράμμισε τη σημασία της προσοχής κατά την κοινή χρήση δεδομένων μετρήσεων, προσφέροντας πολύτιμες γνώσεις για την αποτελεσματική διαχείριση των μετρήσεων δεδομένων.[34] Είναι αξιοσημείωτο ότι διάφορα εργαλεία και βιβλιοθήκες όπως tcpdump (Jacobson, Leres, & McCanne, 1987), WinDump (Frankel, Greely, & Sawyer, 1999), WinPcap (Huang, Wang, Aluru, Yang, & Hillier, 2003) και To Wireshark (Orebaugh et al., 2019) μπορεί να χρησιμοποιηθεί για την ανίχνευση κίνησης δικτύου. Ωστόσο, είναι σημαντικό να αναγνωρίσουμε ότι η χρήση αυτών των εφαρμογών απαιτεί άφθονο χώρο αποθήκευσης για την υποδοχή των δεδομένων που συλλέγονται.

### 5.3.2Δεδομένα επιπέδου πακέτου

Στη μελέτη που διεξήχθη από τους Buczak και Guven (2016), αποδείχθηκε ότι τα πακέτα κίνησης δικτύου μπορούν να συλληφθούν και να εντοπιστούν μέσω διαφόρων εφαρμογών που χρησιμοποιούνται από τους χρήστες. Αυτές οι εφαρμογές είναι σε θέση να ανιχνεύσουν τα πακέτα που λαμβάνονται και μεταδίδονται μέσω της φυσικής διεπαφής χρησιμοποιώντας εργαλεία όπως το WinPCap και το Libpcap, ειδικά στο λειτουργικό σύστημα Windows.

### 5.3.3Δεδομένα NetFlow

Τα δεδομένα NetFlow προσφέρουν σημαντικές πληροφορίες για την παρακολούθηση και τη διαχείριση της κυκλοφορίας σε ένα δίκτυο υπολογιστών. Το NetFlow έκδοσης 5, που περιγράφεται εδώ, ορίζει μια ροή δικτύου ως ένα μονόκλινο σύνολο πακέτων που έχουν κοινά χαρακτηριστικά. Αυτά τα χαρακτηριστικά περιλαμβάνουν:

- Διεύθυνση IP προέλευσης: Η διεύθυνση IP του αποστολέα του πακέτου.
- Διεύθυνση IP προορισμού: Η διεύθυνση IP του παραλήπτη του πακέτου.
- Θύρα προέλευσης: Ο αριθμός θύρας του αποστολέα.
- Θύρα προορισμού: Ο αριθμός θύρας του παραλήπτη.
- Πρωτόκολλο IP: Το πρωτόκολλο που χρησιμοποιείται για τη μεταφορά των δεδομένων, όπως TCP, UDP κλπ.
- Τύπος υπηρεσίας IP: Ο τύπος υπηρεσίας που χρησιμοποιείται, όπως VoIP, FTP κλπ.
- Χρόνος: Ο χρόνος καταγραφής του πακέτου.

Αυτά τα χαρακτηριστικά παρέχουν σημαντικές πληροφορίες για την ανάλυση της κυκλοφορίας δεδομένων και τη λήψη αποφάσεων σχετικά με τη διαχείριση του δικτύου..

### 5.3.4Σύνολο δεδομένων ιχνών κυκλοφορίας KDD99 και NSL KDD

Το KDD99, που αναπτύχθηκε από την DARPA, αποτελεί ένα πρωτοποριακό σύνολο δεδομένων αναφοράς συστήματος ανίχνευσης εισβολής. Μέσα σε αυτό το σύνολο δεδομένων, ένα ευρύ φάσμα τύπων επιθέσεων ταξινομείται σε διακριτές ομάδες, συγκεκριμένα Probe, DOS, R2L και U2R, όπως προσδιορίστηκε στην έρευνα των Chapaneri και Shah το 2019. Περιλαμβάνοντας τόσο κανονικές όσο και επιθέσεις κίνησης, αυτό το σύνολο δεδομένων περιλαμβάνει συνολικά σαράντα -ένα χαρακτηριστικά. Αξίζει να σημειωθεί ότι το σύνολο δεδομένων ταξινόμησης ανωμαλίας κυκλοφορίας IoT (Bibri, 2018) έχει αναδειχθεί ως δημοφιλής επιλογή μεταξύ των ερευνητών για την αποτελεσματική κατηγοριοποίηση των ανωμαλιών κυκλοφορίας IoT και των επιθέσεων εισβολής. Μια ολοκληρωμένη εξερεύνηση από τους Buczak και Guven το 2016 αποκάλυψε την παρουσία σημαντικού χάους στους τομείς των ML, DM και KDD. Η μελέτη τους ανέδειξε το KDD ως μια συστηματική προσέγγιση για την εξαγωγή πολύτιμων πληροφοριών από δεδομένα εισόδου, που χρησιμεύει ως μια ζωτικής σημασίας διαδικασία βήμα προς βήμα στο πεδίο.

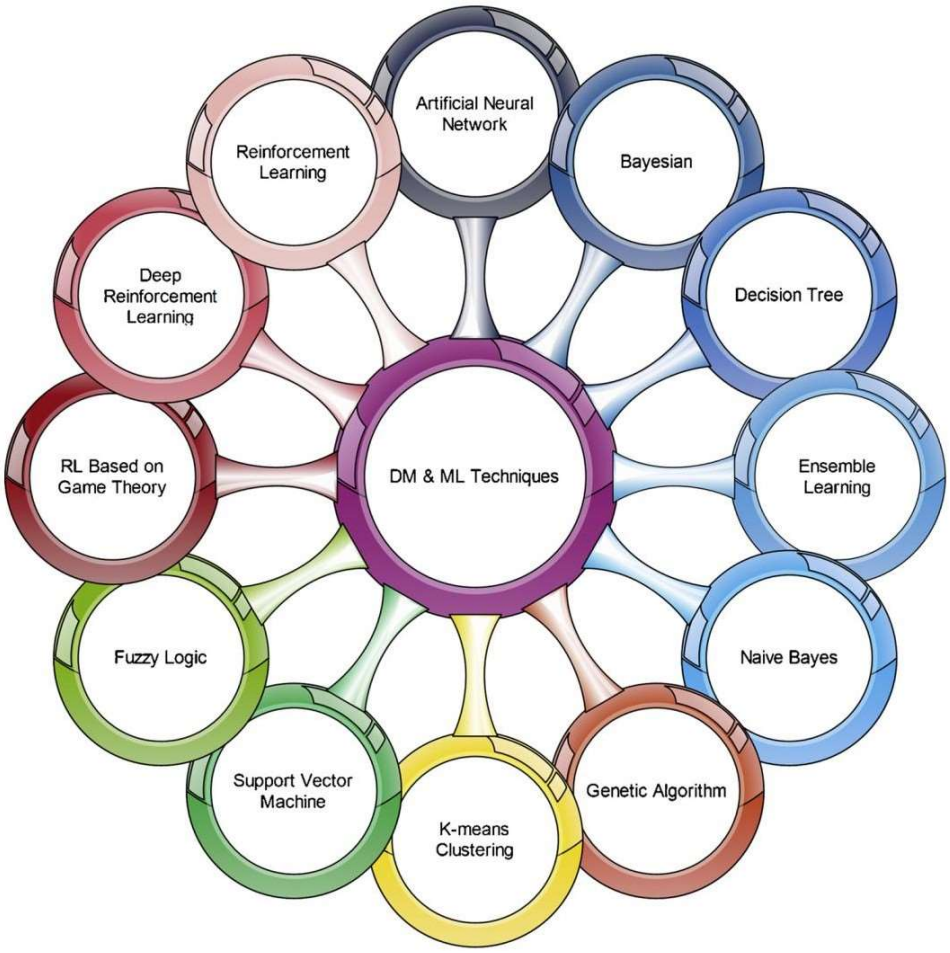
### 5.3.5Ίχνη κυκλοφορίας Auckland II

Το έργο Waikato Internet Traffic Storage είναι υπεύθυνο για τη συλλογή και την τεκμηρίωση όλης της κίνησης στο Διαδίκτυο. Ως μέρος αυτού του έργου, ορισμένα δεδομένα ιχνών καθίστανται ελεύθερα προσβάσιμα στους ερευνητές. Αν και υπάρχουν πολλά διαθέσιμα σύνολα δεδομένων για τον εντοπισμό μη κανονικής κίνησης δικτύου, συμπεριλαμβανομένων των Auckland I, IX, IV και VI, επιλέγουμε συγκεκριμένα το Auckland II. Ο λόγος πίσω από αυτήν την επιλογή είναι ότι το σύνολο δεδομένων Auckland II έχει αποδειχθεί εξαιρετικά αξιόπιστο στην ακριβή ταξινόμηση και χρήση της κίνησης στο Διαδίκτυο, καθιστώντας το έναν προτιμώμενο πόρο για πολλούς ερευνητές που διεξάγουν τις ερευνητικές τους μελέτες.

### 5.3.6. Ίχνη κυκλοφορίας UNIBS

Το UNIBS είναι ένα σημαντικό σύνολο δεδομένων κίνησης που χρησιμοποιείται για την ταξινόμηση της κίνησης στο Διαδίκτυο. Δημιουργήθηκε από τον καθηγητή F. Gringoli και την ερευνητική του ομάδα. Το UNIBS είναι ελεύθερα προσβάσιμο και έχει χρησιμοποιηθεί σε πολλές ερευνητικές εργασίες στον τομέα της αναγνώρισης και ταξινόμησης της κίνησης δικτύου.Η έρευνα που, διεξήχθη από τους Bhuyan, Bhattacharyya και Kalita το 2015, επικεντρώνεται στην ανάπτυξη μιας αποτελεσματικής υλοποίησης συστήματος γνωστής ως GT. Το GT πιθανόν να αναφέρεται στην εφαρμογή ή το εργαλείο που χρησιμοποιήθηκε για την ανάλυση και την επεξεργασία των δεδομένων από το σύνολο UNIBS. Τέτοιου είδους εφαρμογές μπορούν να χρησιμοποιηθούν για να εξάγουν γνώση από τα δεδομένα κίνησης του Διαδικτύου, όπως η αναγνώριση και η κατηγοριοποίηση των ειδών της κίνησης ή η ανίχνευση ανωμαλιών και κακόβουλων δραστηριοτήτων.

## 5.4 Μέθοδοι ML και DM για ταξινόμηση κυκλοφορίας

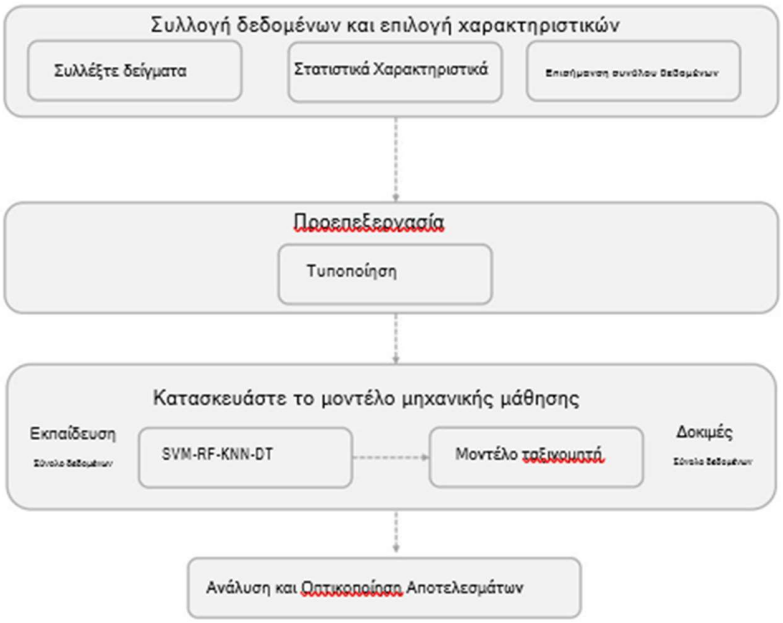



**Εικόνα 23 Τεχνικές DM και ML.**

**Μελέτη ταξινόμησης της κυκλοφορίας με βάση τη μηχανική μάθηση**

Η προσέγγισή στην ταξινόμηση της κυκλοφορίας βάσει μηχανικής μάθησης αποτελείται από τέσσερα βήματα, όπως απεικονίζεται στην Εικόνα 25. Το πρώτο βήμα περιλαμβάνει τη συλλογή δεδομένων και την επιλογή χαρακτηριστικών. Σε αυτό το βήμα, συλλέγουμε δείγματα ροής κυκλοφορίας και

επιλέγουμε προσεκτικά τα σχετικά χαρακτηριστικά για αξιολόγηση. Για να προετοιμάσουμε τα σύνολα δεδομένων για εκπαίδευση και δοκιμή, εξαλείφουμε τυχόν μη στατιστικά χαρακτηριστικά και εκχωρούμε σε κάθε δείγμα την κατάλληλη ετικέτα κλάσης με βάση τη μη αυτόματη επισήμανση. Προχωρώντας στο βήμα προεπεξεργασίας, κανονικοποιούμε τα χαρακτηριστικά κλιμακώνοντάς τα χρησιμοποιώντας μια τεχνική κανονικοποίησης. Τέλος, προχωράμε στην κατασκευή ενός μοντέλου μηχανικής μάθησης και στην ανάλυση και οπτικοποίηση των αποτελεσμάτων που προέκυψαν από τη διαδικασία ταξινόμησης. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν το σύνολο δεδομένων εκπαίδευσης για να δημιουργήσουν ένα κατάλληλο μοντέλο για την ταξινόμηση της κυκλοφορίας. Η μελέτη μας περιλαμβάνει μια σύγκριση τεσσάρων ευρέως χρησιμοποιούμενων εποπτευόμενων αλγορίθμων μηχανικής μάθησης, δηλαδή SVM, RF, KNN και DT. Για να αξιολογήσουμε την αποτελεσματικότητα κάθε αλγορίθμου, χρησιμοποιούμε το σύνολο δεδομένων δοκιμής για σκοπούς αξιολόγησης. Αυτή η αξιολόγηση συνεπάγεται τη χρήση τεσσάρων διακριτών μετρήσεων, δηλαδή την ακρίβεια, την ακρίβεια, την ανάκληση και τη βαθμολογία F1, για τη μέτρηση της απόδοσης του αλγορίθμου. Επιπλέον, ενσωματώνουμε την τεχνική διασταυρούμενης επικύρωσης k-fold για την ολοκληρωμένη αξιολόγηση της απόδοσης της ταξινόμησης.



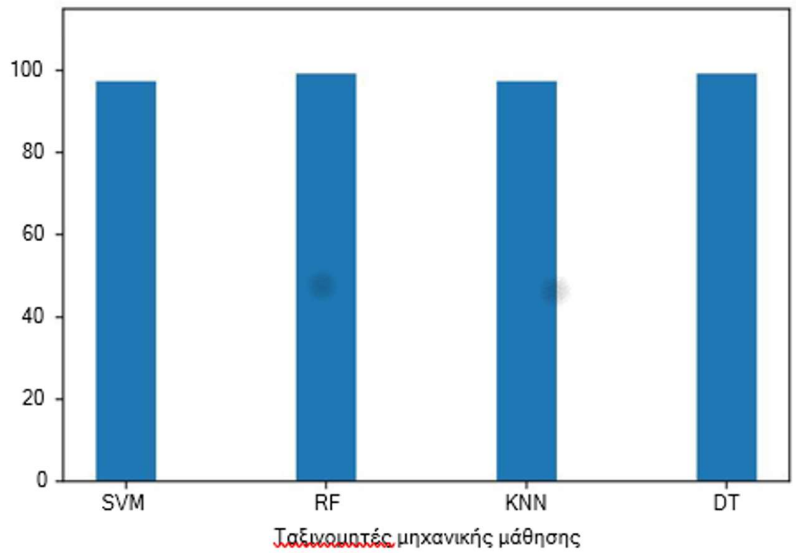

**Εικόνα 24 Βήματα για τη δημιουργία και την αξιολόγηση προτεινόμενων αλγορίθμων μηχανικής μάθησης.**

**Πηγή:** https://eclass-cybele.cce.uoa.gr/courses/CCEMECH181/

Για να βελτιωθεί η ακρίβεια κάθε αλγορίθμου, έγιναν πειράματα με διάφορες παραμέτρους μοντέλου. Για παράδειγμα, στην περίπτωση του SVM, χρησιμοποιήθηκε ενας γραμμικός πυρήνας. Αυτός ο πυρήνας επιτρέπει στο SVM να συμμετέχει σε εποπτευόμενη μάθηση. Με αυτόν τον τρόπο, το SVM είναι σε θέση να προσδιορίσει ένα υπερεπίπεδο που διαχωρίζει αποτελεσματικά τα δείγματα, μεγιστοποιώντας έτσι το περιθώριο μεταξύ διαφορετικών κλάσεων. Αυτή η διαδικασία διευκολύνεται από τη χρήση φορέων υποστήριξης [33]. Στη μελέτη, πραγματοποιήθηκε μια ανάλυση τεσσάρων αλγορίθμων μηχανικής μάθησης,

δηλαδή Υποστήριξη Διανυσματικών Μηχανών (SVM), Τυχαίου Δάσους (RF), Κ-Πλησιότερους Γείτονες (KNN) και Δέντρα Απόφασης (DT). Στόχος ήταν να συγκριθεί η απόδοση αυτών των αλγορίθμων εφαρμόζοντάς τους και εξετάζοντας τα αποτελέσματα. Συνολικά, η μελέτη περιελάβανε τη διεξοδική εφαρμογή και αξιολόγηση των αλγορίθμων SVM, RF, KNN και DT. Δόθηκε ιδιαίτερη προσοχή στη βελτιστοποίηση των ρυθμίσεων παραμέτρων προκειμένου να επιτευχθεί η μέγιστη δυνατή ακρίβεια. Επιπλέον, έγιεν εμβάθυνση στις ιδιαιτερότητες του SVM, τονίζοντας τη χρήση ενός γραμμικού πυρήνα και την ικανότητά του να εκτελεί εποπτευόμενη μάθηση σε χώρο υψηλών διαστάσεων.

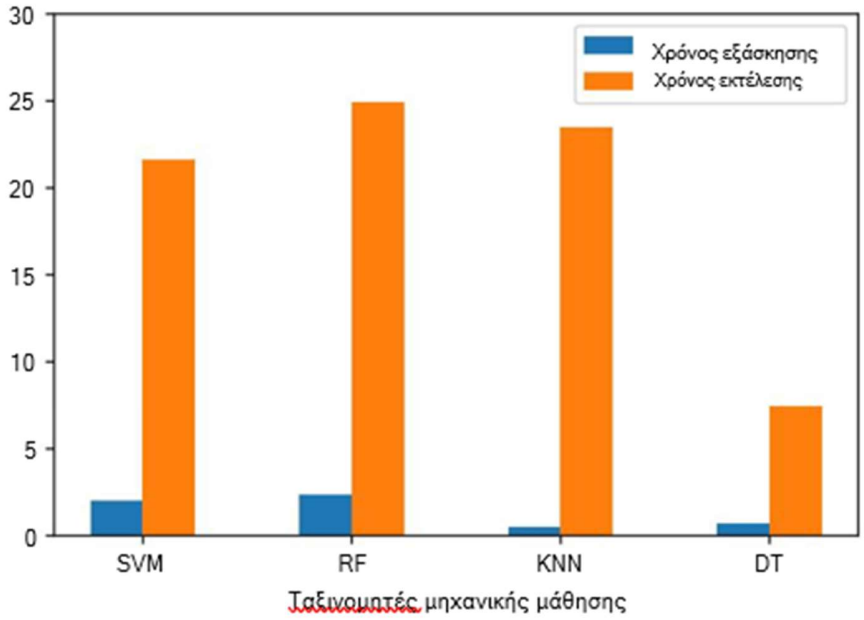
Το μοντέλο SVM πέτυχε μια εντυπωσιακή μέση ακρίβεια 97,41%, επιδεικνύοντας την ανώτερη απόδοσή του (όπως φαίνεται στο Σχήμα 4). Μια λεπτομερής ανάλυση της βαθμολογίας ακρίβειας, ανάκλησης και F1 για κάθε κατηγορία κυκλοφορίας κατά τη χρήση SVM μπορεί να βρεθεί στην εικόνα 26. Τα αποτελέσματα υπογραμμίζουν ότι οι κατηγορίες Interactive και Multimedia παρουσιάζουν χαμηλότερες βαθμολογίες σε σύγκριση με άλλες ετικέτες ταξινόμησης. Αυτή η διαφορά στην απόδοση μπορεί να αποδοθεί στον μικρότερο αριθμό δειγμάτων που είναι διαθέσιμα για αυτές τις κατηγορίες, επηρεάζοντας τελικά την ακρίβεια ταξινόμησης.




**Εικόνα 25 Μέση ακρίβεια αλγορίθμων μηχανικής μάθησης**
**Πηγή:** https://eclass-cybele.cce.uoa.gr/courses/CCEMECH181/

Η αξιολόγηση της απόδοσης των αλγορίθμων μηχανικής μάθησης είναι κρίσιμη πέρα από την απλή μέτρηση της ακρίβειας. Παράγοντες όπως ο χρόνος εκπαίδευσης και ο χρόνος εκτέλεσης παίζουν σημαντικό ρόλο στον χαρακτηρισμό της αποτελεσματικότητας αυτών των αλγορίθμων. Ο χρόνος εκπαίδευσης αναφέρεται στη διάρκεια που χρειάζεται για να μάθει ένα μοντέλο από ένα δεδομένο σύνολο δεδομένων, ενώ ο χρόνος εκτέλεσης περιλαμβάνει ολόκληρη τη διαδικασία χειρισμού δεδομένων, προεπεξεργασίας και αξιολόγησης μοντέλου. Αυτές οι μετρήσεις είναι απαραίτητες για τον προσδιορισμό της αποδοτικότητας και της αποτελεσματικότητας των μοντέλων μηχανικής εκμάθησης. Το Σχήμα 27 δείχνει τους χρόνους εκπαίδευσης και εκτέλεσης για διάφορους αλγόριθμους μηχανικής εκμάθησης. Ο αλγόριθμος SVM χρειάζεται 2,01 δευτερόλεπτα για να εκπαιδευτεί και 21,59 δευτερόλεπτα για να εκτελεστεί. Από την άλλη πλευρά, ο αλγόριθμος RF απαιτεί 2,36 δευτερόλεπτα για εκπαίδευση και 24,90 δευτερόλεπτα για εκτέλεση. Κατά συνέπεια, ο RF θεωρείται ο πιο αργός αλγόριθμος καθώς δημιουργεί και υπολογίζει πολλά δέντρα απόφασης. Αντίθετα, ο αλγόριθμος KNN έχει χρόνο εκπαίδευσης μόνο 0,49 δευτερολέπτων αφού δεν κατασκευάζει μοντέλα κατά τη διάρκεια της εκπαίδευσης. Αντίθετα, αποθηκεύει τα δεδομένα εκπαίδευσης

για μελλοντική ταξινόμηση. Το KNN χρειάζεται περισσότερο χρόνο για να μετρήσει την απόσταση από το σημείο δεδομένων k-πλησιέστερου γείτονα. Ο αλγόριθμος DT, από την άλλη πλευρά, έχει χρόνο εκπαίδευσης 0,72 δευτερόλεπτα και χρόνο εκτέλεσης 7,47 δευτερόλεπτα. Επομένως, ο DT είναι ο ταχύτερος αλγόριθμος από άποψη χρόνου εκπαίδευσης και εκτέλεσης.




**Εικόνα 26 Χρόνοι εκπαίδευσης και εκτέλεσης αλγορίθμων μηχανικής μάθησης.**
**Πηγή:** https://eclass-cybele.cce.uoa.gr/courses/CCEMECH181/

Με το πέρασμα του χρόνου και τις προόδους της τεχνολογίας, η έννοια των έξυπνων πόλεων κερδίζει δημοτικότητα. Ο πρωταρχικός στόχος της εφαρμογής έξυπνων λύσεων είναι να βελτιώσουμε την ποιότητα της καθημερινότητάς μας καθιστώντας τις πιο βολικές, παραγωγικές και αποτελεσματικές. Οι έξυπνες πόλεις περιλαμβάνουν ένα ευρύ φάσμα απαιτήσεων εφαρμογών, δεδομένων και ποιότητας υπηρεσιών (QoS), οι οποίες θέτουν προκλήσεις για την αποτελεσματική διαχείριση της κυκλοφορίας. Μια προσέγγιση για την αντιμετώπιση αυτών των προκλήσεων είναι μέσω της ταξινόμησης της κυκλοφορίας, η οποία μπορεί να βοηθήσει στη διαχείριση διαφόρων πτυχών του δικτύου, συμπεριλαμβανομένης της εξασφάλισης υποστήριξης QoS.

Οι παραδοσιακές μέθοδοι για την κατηγοριοποίηση της κυκλοφορίας, όπως οι μέθοδοι που βασίζονται σε θύρα και η βαθιά επιθεώρηση πακέτων, αντιμετωπίζουν προβλήματα με κρυπτογραφημένα δεδομένα και αλλαγή αριθμών θυρών. Αντίθετα, οι αλγόριθμοι μηχανικής μάθησης προσφέρουν μια λύση για τη διαχείριση της ποιότητας των υπηρεσιών και τον χειρισμό πολύπλοκων καταστάσεων. Τα ευρήματά της παραπάνω μελέτης έδειξαν ότι η ενσωμάτωση στατιστικών χαρακτηριστικών βελτίωσε την ακρίβεια της ταξινόμησης της κυκλοφορίας χρησιμοποιώντας μηχανική εκμάθηση. Ο αλγόριθμος DT έδειξε το υψηλότερο μέσο ποσοστό ακρίβειας στο 99,18%, ενώ ο αλγόριθμος KNN είχε τη χαμηλότερη μέση ακρίβεια στο 97,16%. Επιπλέον, επισημάναμε τους περιορισμούς της βασιζόμενης αποκλειστικά σε μεθόδους που

βασίζονται σε θύρα για την ταξινόμηση της κυκλοφορίας, καθώς εξαρτώνται από συγκεκριμένους αριθμούς θυρών για τη διαφοροποίηση των ροών δικτύου.

# 6. Μηχανική Μάθηση στο Azure

## 6.1 Τι είναι το Azure Machine Learning;

[51]Η διαδικασία εκπαίδευσης και βελτίωσης ενός επιτυχημένου μοντέλου μηχανικής μάθησης είναι μια σύνθετη και απαιτητική προσπάθεια, που απαιτεί σημαντική επένδυση χρόνου και πόρων. Το Azure Machine Learning προσφέρει μια βολική λύση που βασίζεται σε σύννεφο για τον εξορθολογισμό των απαραίτητων βημάτων που σχετίζονται με την προετοιμασία δεδομένων, την εκπαίδευση ενός μοντέλου και τη δημιουργία μιας υπηρεσίας πρόβλεψης. Χρησιμοποιώντας το Azure Machine Learning, οι χρήστες μπορούν να κατασκευάσουν μοντέλα μηχανικής μάθησης ταξινόμησης με ευκολία.

Ένα από τα πιο σημαντικά πλεονεκτήματα του Azure Machine Learning είναι η ικανότητά του να ενισχύει την παραγωγικότητα των επιστημόνων δεδομένων. Με την αυτοματοποίηση διαφόρων επίπονων εργασιών που εμπλέκονται στην εκπαίδευση μοντέλων, δίνει τη δυνατότητα τους επαγγελματίες να εξοικονομήσουν χρόνο και να επικεντρωθούν σε πιο πολύτιμες πτυχές της δουλειάς τους. Επιπλέον, το Azure Machine Learning επιτρέπει στους επιστήμονες δεδομένων να αξιοποιήσουν τη δύναμη των υπολογιστικών πόρων που βασίζονται σε σύννεφο, οι οποίοι φιλοξενούν αποτελεσματικά τεράστιες ποσότητες δεδομένων.

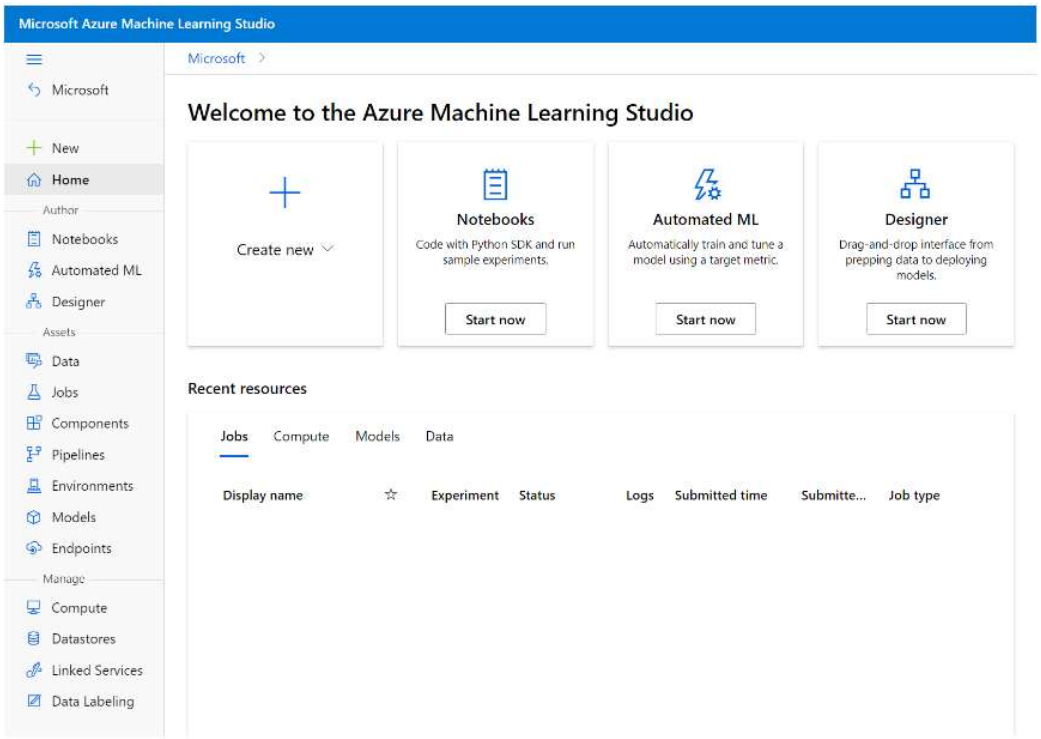
**Χώρος εργασίας Azure Machine Learning**

Για να χρησιμοποιήσουμε τη Μηχανική Εκμάθηση Azure, το αρχικό βήμα περιλαμβάνει τη δημιουργία ενός πόρου χώρου εργασίας στη συνδρομή μας στο Azure. Στη συνέχεια, αυτός ο χώρος εργασίας μπορεί να χρησιμοποιηθεί για την αποτελεσματική επίβλεψη διαφόρων πτυχών του φόρτου εργασίας μηχανικής εκμάθησης, συμπεριλαμβανομένης της διαχείρισης δεδομένων, των υπολογιστικών πόρων, της κωδικοποίησης, της ανάπτυξης μοντέλων και άλλων σχετικών τεχνουργημάτων.

Αφού δημιουργήσουμε έναν χώρο εργασίας Azure Machine Learning, μπορούμε να αναπτύξουμε λύσεις με την υπηρεσία μηχανικής εκμάθησης Azure είτε με εργαλεία προγραμματιστή είτε με την πύλη web στούντιο Azure Machine Learning.

**Azure Machine Learning Studio**

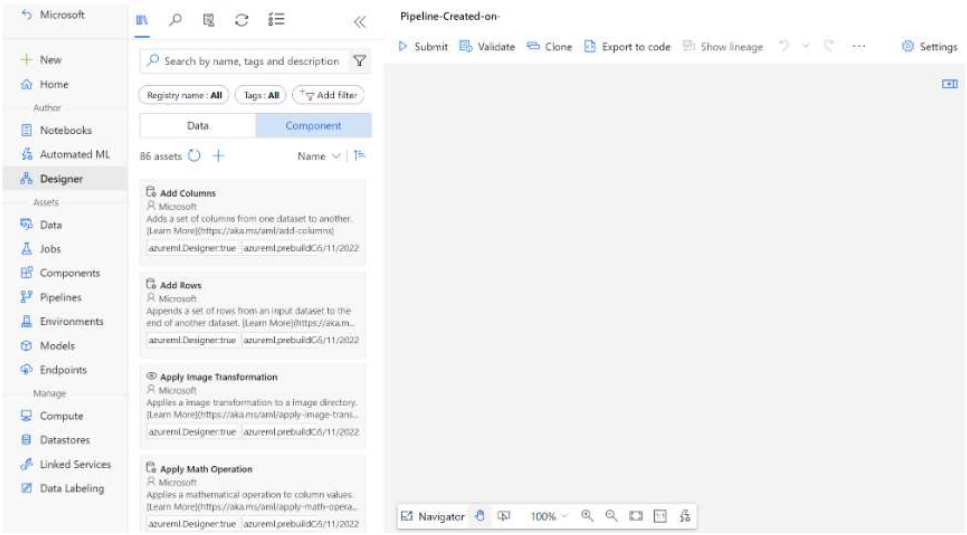
Το Azure Machine Learning Studio είναι μια ολοκληρωμένη διαδικτυακή πλατφόρμα που έχει σχεδιαστεί ειδικά για τη δημιουργία και την εφαρμογή λύσεων μηχανικής εκμάθησης εντός του περιβάλλοντος Azure. Αυτό το προηγμένο εργαλείο προσφέρει μια πληθώρα λειτουργιών και χαρακτηριστικών που βοηθούν σημαντικά τους επιστήμονες δεδομένων σε διάφορες πτυχές της εργασίας τους. Από την προετοιμασία δεδομένων έως την εκπαίδευση μοντέλων, από τη δημοσίευση υπηρεσιών πρόβλεψης έως την παρακολούθηση της χρήσης, το στούντιο μηχανικής εκμάθησης Azure παρέχει μια ολοκληρωμένη λύση για τον εξορθολογισμό ολόκληρης της διαδικασίας μηχανικής εκμάθησης.




**Azure Machine Learning studio**

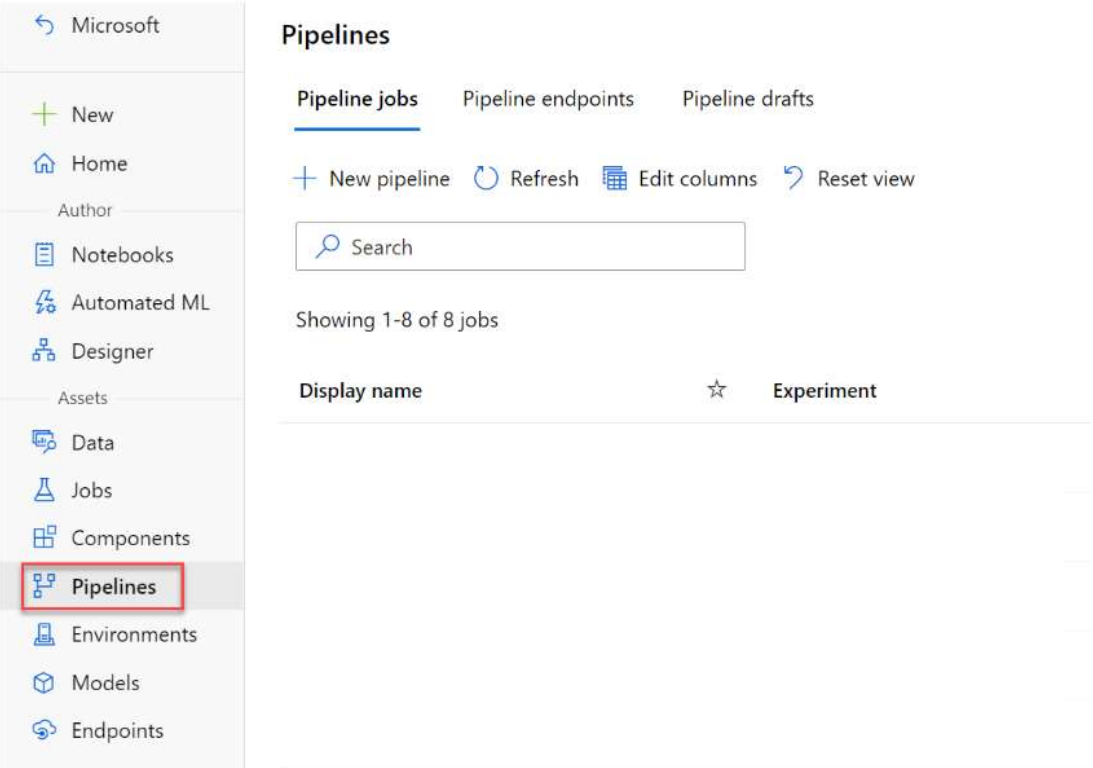
Στο στούντιο μηχανικής εκμάθησης Azure, υπάρχουν διάφοροι τρόποι για τη σύνταξη ταξινόμησης μοντέλων μηχανικής εκμάθησης. Ένας τρόπος είναι να χρησιμοποιήσουμε μια οπτική διεπαφή που ονομάζεται σχεδιαστής που μπορούμε να χρησιμοποιήσουμε για να εκπαιδεύσουμε, να δοκιμάσουμε και να αναπτύξουμε μοντέλα μηχανικής εκμάθησης. Η διεπαφή μεταφοράς και απόθεσης χρησιμοποιεί σαφώς καθορισμένες εισόδους και εξόδους που μπορούν να κοινοποιηθούν, να επαναχρησιμοποιηθούν και να ελέγχονται από την έκδοση.

Κάθε έργο σχεδιαστή, γνωστό ως αγωγός, έχει ένα αριστερό πλαίσιο για πλοήγηση και έναν καμβά στο δεξί σας χέρι. Για να χρησιμοποιήσουμε τον σχεδιαστή, προσδιορίζουμε τα δομικά στοιχεία ή εξαρτήματα που χρειάζονται για το μοντέλο μας, τα τοποθετούμε και τα συνδέουμε στον καμβά μας και εκτελούμε μια εργασία μηχανικής εκμάθησης.



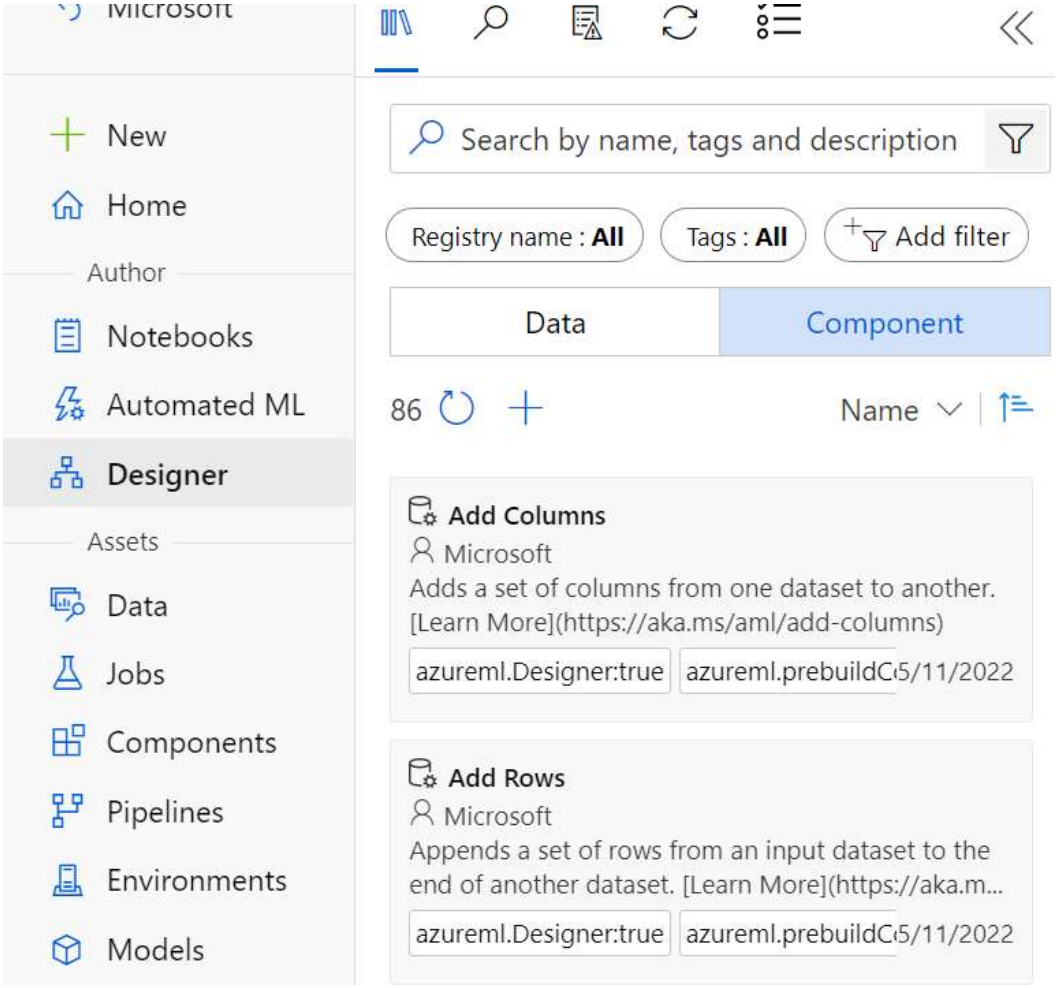

**Pipelines**

Οι Pipelines μας επιτρέπουν να οργανώνουμε, να διαχειριζόμαστε και να επαναχρησιμοποιούμε πολύπλοκες ροές εργασιών μηχανικής εκμάθησης σε έργα και χρήστες. Μια διοχέτευση ξεκινά με το σύνολο δεδομένων από το οποίο θέλουμε να εκπαιδεύσουμε το μοντέλο. Κάθε φορά που εκτελούμε μια διοχέτευση, η διαμόρφωση της διοχέτευσης και τα αποτελέσματά της αποθηκεύονται στον χώρο εργασίας μας ως εργασία διοχέτευσης.



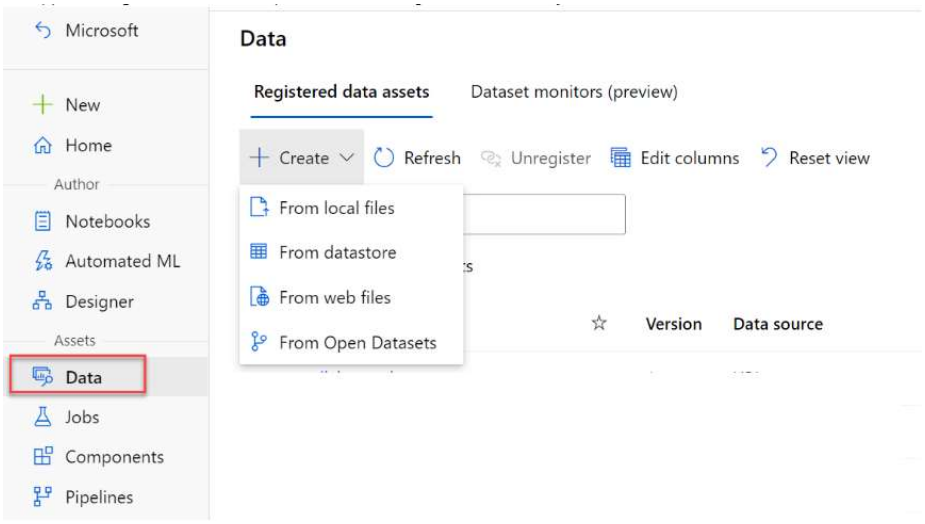
**Components**

Ένα στοιχείο Azure Machine Learning ενσωματώνει ένα βήμα σε έναν αγωγό μηχανικής μάθησης. Μπορούμε να σκεφτούμε ένα στοιχείο ως συνάρτηση προγραμματισμού και ως δομικό στοιχείο για αγωγούς Azure Machine Learning. Σε ένα έργο διοχέτευσης, μπορούμε να αποκτήσουμε πρόσβαση σε στοιχεία και στοιχεία δεδομένων από την καρτέλα Βιβλιοθήκη στοιχείων του αριστερού πίνακα.




**Datasets**

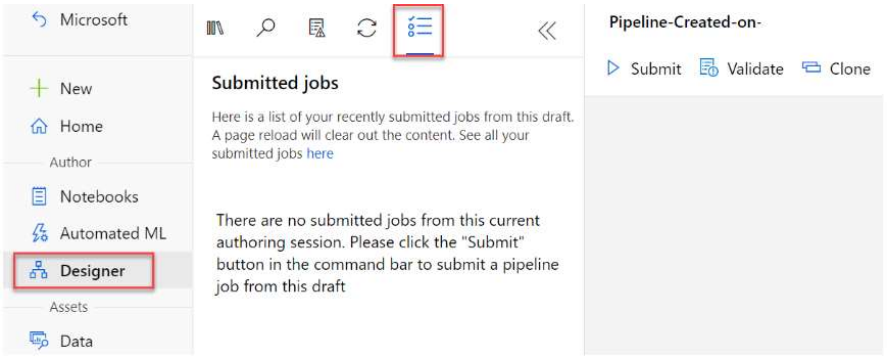
Μπορούμε να δημιουργήσουμε στοιχεία δεδομένων στη σελίδα Δεδομένα από τοπικά αρχεία, χώρο αποθήκευσης δεδομένων, αρχεία web και Open Datasets. Αυτά τα στοιχεία δεδομένων θα εμφανίζονται μαζί με τυπικά δείγματα συνόλων δεδομένων στη βιβλιοθήκη στοιχείων του σχεδιαστή.

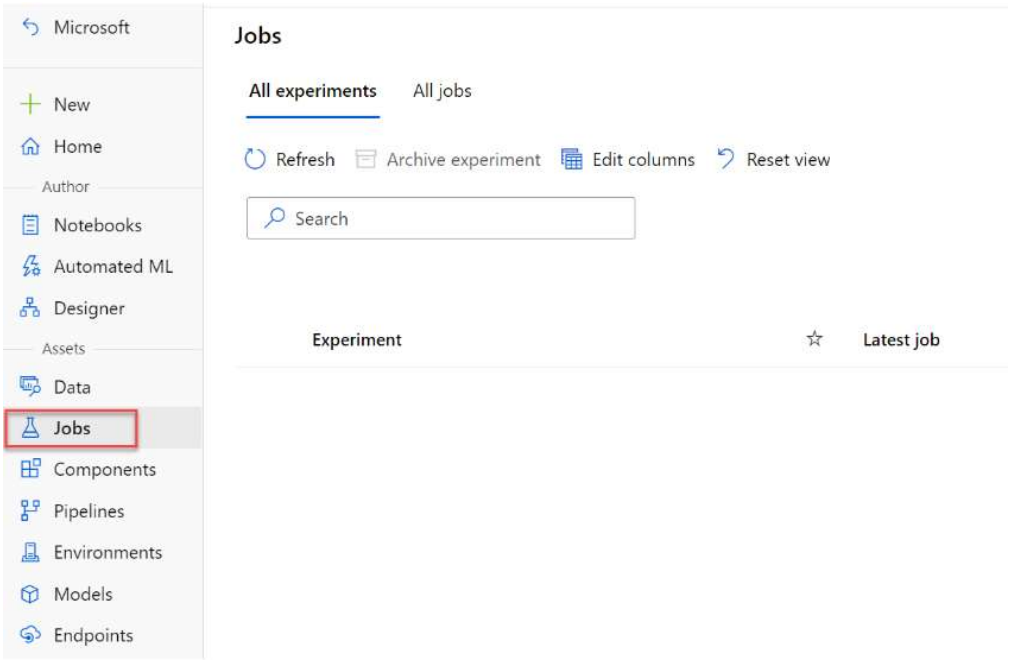



**Azure Machine Learning Jobs**

Μια εργασία Azure Machine Learning (ML) εκτελεί μια εργασία έναντι ενός καθορισμένου στόχου υπολογισμού. Οι εργασίες επιτρέπουν τη συστηματική παρακολούθηση για τον πειραματισμό και τις ροές εργασιών μηχανικής εκμάθησης. Μόλις δημιουργηθεί μια εργασία, το Azure ML διατηρεί ένα αρχείο εκτέλεσης για την εργασία. Μπορούμε να δούμε όλες τις εγγραφές των εργασιών μας στο στούντιο Azure ML.

Στο σχεδιαστή μας, μπορούμε να αποκτήσουμε πρόσβαση στην κατάσταση μιας εργασίας διοχέτευσης χρησιμοποιώντας την καρτέλα Υποβλήθηκαν εργασίες στο αριστερό παράθυρο.






**Βήματα για την ταξινόμηση στο Azure**

Μπορούμε να σκεφτούμε τα βήματα για την εκπαίδευση και την αξιολόγηση ενός μοντέλου μηχανικής μάθησης ταξινόμησης ως εξής:

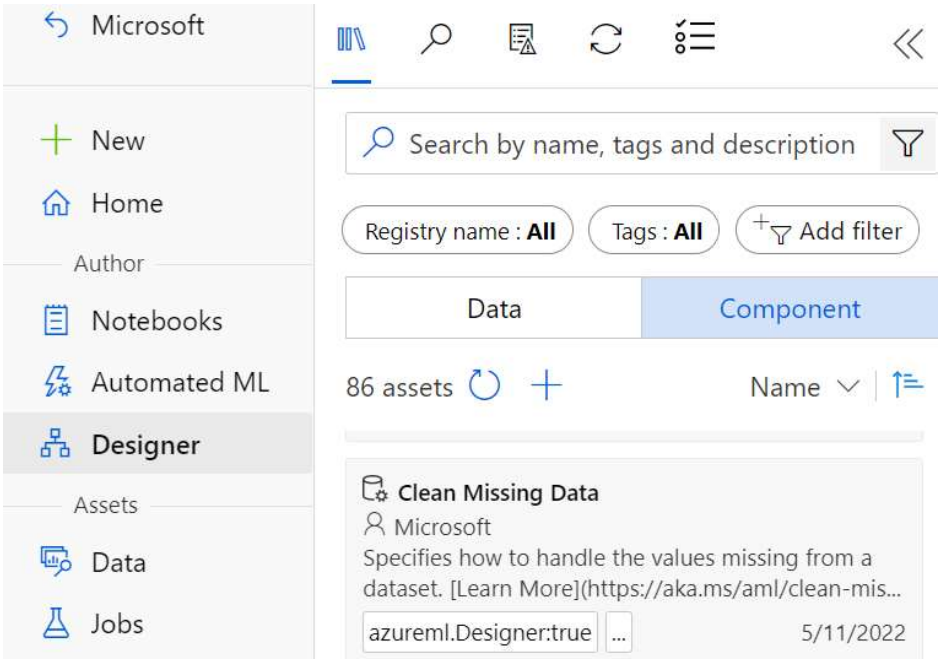
- Προετοιμασία δεδομένων: Προσδιορίστε τα χαρακτηριστικά και την ετικέτα σε ένα σύνολο δεδομένων. Προ επεξεργαστείτε ή καθαρίστε και μετασχηματίστε τα δεδομένα όπως απαιτείται.
- Εκπαίδευση του Μοντέλου: Χωρίστε τα δεδομένα σε δύο ομάδες, μια εκπαίδευσης και ένα σύνολο επικύρωσης. Εκπαιδεύστε ένα μοντέλο μηχανικής μάθησης χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης. Δοκιμάστε το μοντέλο μηχανικής εκμάθησης για απόδοση χρησιμοποιώντας το σύνολο δεδομένων επικύρωσης.
- Αξιολόγηση της απόδοσης: Συγκρίνουμε πόσο κοντά είναι οι προβλέψεις του μοντέλου με τις γνωστές ετικέτες. Αφού εκπαιδεύσουμε ένα μοντέλο μηχανικής εκμάθησης, πρέπει να μετατρέψουμε τη γραμμή εκπαίδευσης σε μια διοχέτευση συμπερασμάτων σε πραγματικό χρόνο.

Στη συνέχεια, μπορούμε να αναπτύξουμε το μοντέλο ως εφαρμογή σε διακομιστή ή συσκευή, ώστε να μπορούν να το χρησιμοποιήσουν άλλοι.

Ας ακολουθήσουμε αυτά τα τέσσερα βήματα όπως εμφανίζονται στο Azure Designer.

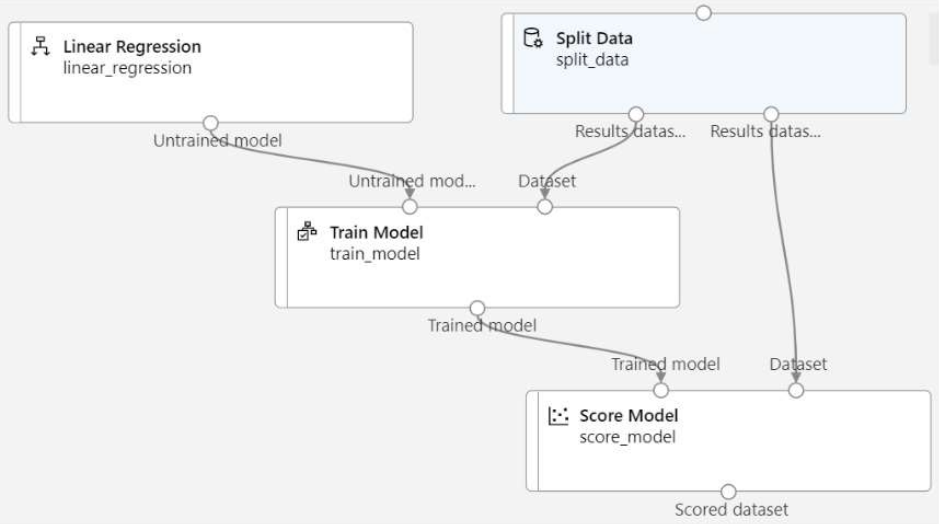
**Προετοιμασία δεδομένων**

Ο σχεδιαστής μηχανικής εκμάθησης Azure έχει πολλά προκατασκευασμένα στοιχεία που μπορούν να χρησιμοποιηθούν για την προετοιμασία δεδομένων για εκπαίδευση. Αυτά τα στοιχεία μάς επιτρέπουν να καθαρίζουμε δεδομένα, να ομαλοποιούμε τις λειτουργίες, να ενώνουμε πίνακες και πολλά άλλα.



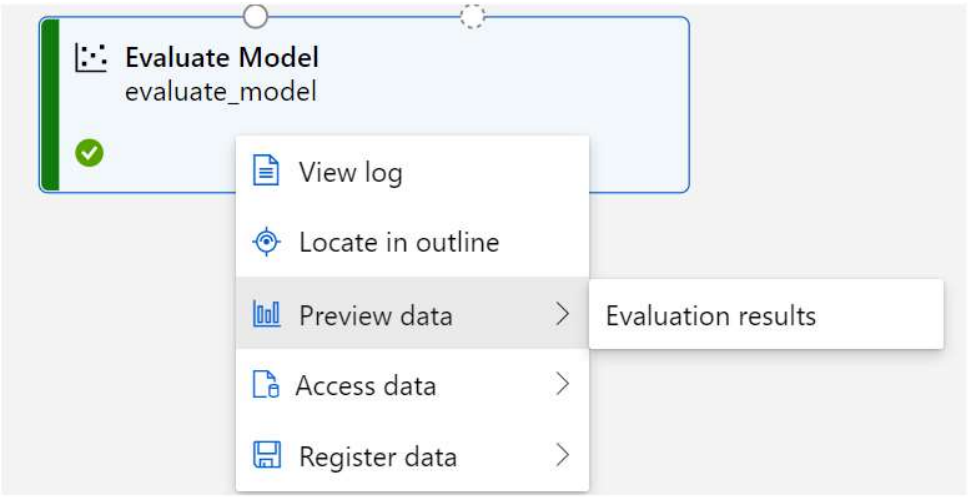

**Εκπαίδευση μοντέλου**

Για να εκπαιδεύσουμε ένα μοντέλο ταξινόμησης, χρειαζόμαστε ένα σύνολο δεδομένων που περιλαμβάνει ιστορικά χαρακτηριστικά, χαρακτηριστικά της οντότητας για την οποία θέλουμε να κάνουμε μια πρόβλεψη και γνωστές τιμές ετικέτας. Η ετικέτα είναι ο δείκτης κλάσης που θέλουμε να εκπαιδεύσουμε ένα μοντέλο να προβλέπει. Είναι κοινή πρακτική να εκπαιδεύουμε το μοντέλο χρησιμοποιώντας ένα υποσύνολο δεδομένων, ενώ συγκρατούμε ορισμένα δεδομένα με τα οποία μπορούμε να δοκιμάσουμε το εκπαιδευμένο μοντέλο. Αυτό μας δίνει τη δυνατότητα να συγκρίνουμε τις ετικέτες που προβλέπει το μοντέλο με τις πραγματικές γνωστές ετικέτες στο αρχικό σύνολο δεδομένων.Θα χρησιμοποιήσουμε το στοιχείο Score Model του σχεδιαστή για να δημιουργήσουμε την προβλεπόμενη τιμή ετικέτας κλάσης. Μόλις συνδέσουμε όλα τα στοιχεία, θα θέλουμε να εκτελέσουμε ένα πείραμα, το οποίο θα χρησιμοποιήσει το στοιχείο δεδομένων στον καμβά για να εκπαιδεύσει και να βαθμολογήσει ένα μοντέλο.




**Αξιολογούμε την απόδοση**

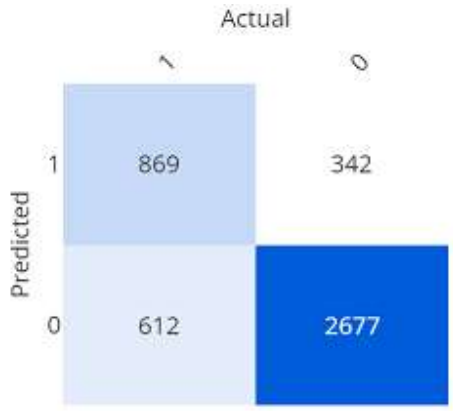
Μετά την εκπαίδευση ενός μοντέλου, είναι σημαντικό να αξιολογήσουμε την απόδοσή του. Υπάρχουν πολλές μετρήσεις απόδοσης και μεθοδολογίες για την αξιολόγηση του πόσο καλά κάνει προβλέψεις ένα μοντέλο. Μπορούμε να ελέγξουμε τις μετρήσεις αξιολόγησης στη σελίδα ολοκληρωμένης εργασίας κάνοντας δεξί κλικ στο στοιχείο Αξιολόγηση μοντέλου.




**Πίνακας σύγχυσης**

Ο πίνακας σύγχυσης εμφανίζει περιπτώσεις όπου τόσο οι προβλεπόμενες όσο και οι πραγματικές τιμές ήταν 1 (γνωστές ως αληθινά θετικά) επάνω αριστερά και περιπτώσεις όπου τόσο οι προβλεπόμενες όσο και οι πραγματικές τιμές ήταν 0 (αληθινά αρνητικά) κάτω δεξιά. Τα άλλα κελιά εμφανίζουν περιπτώσεις όπου οι προβλεπόμενες και οι πραγματικές τιμές διαφέρουν (ψευδώς θετικά και ψευδώς αρνητικά).

Για ένα μοντέλο δυαδικής ταξινόμησης όπου προβλέπουμε μία από τις δύο πιθανές τιμές, ο πίνακας σύγχυσης είναι ένα πλέγμα 2x2 που δείχνει τις προβλεπόμενες και πραγματικές μετρήσεις τιμών για τις κλάσεις 0 και 1, παρόμοιο με αυτό:




Στιγμιότυπο οθόνης μήτρας σύγχυσης που εμφανίζει πραγματικές και προβλεπόμενες μετρήσεις τιμών.

Για ένα μοντέλο ταξινόμησης πολλών κατηγοριών (όπου υπάρχουν περισσότερες από δύο πιθανές κλάσεις), η ίδια προσέγγιση χρησιμοποιείται για τον πίνακα κάθε πιθανού συνδυασμού πραγματικών και προβλεπόμενων μετρήσεων τιμών - έτσι ένα μοντέλο με τρεις πιθανές κατηγορίες θα είχε ως αποτέλεσμα έναν πίνακα 3x3 με διαγώνια γραμμή κελιών όπου ταιριάζουν οι προβλεπόμενες και οι πραγματικές ετικέτες.

Οι μετρήσεις μπορούν να προκύψουν από τον πίνακα σύγχυσης και περιλαμβάνουν:

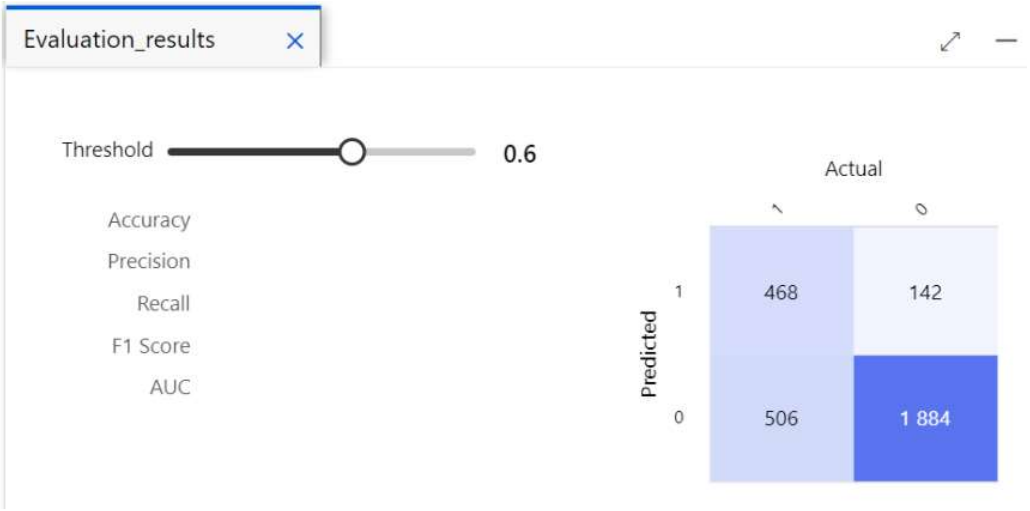
- Ακρίβεια: Ο λόγος των σωστών προβλέψεων (αληθινά θετικά + αληθινά αρνητικά) προς τον συνολικό αριθμό προβλέψεων.
- Ακρίβεια: Το κλάσμα των περιπτώσεων που ταξινομούνται ως θετικές και είναι πραγματικά θετικές (ο αριθμός των αληθινών θετικών διαιρεμένος με τον αριθμό των αληθινών θετικών συν ψευδώς θετικών).
- Ανάκληση: Το κλάσμα των θετικών περιπτώσεων που προσδιορίστηκαν σωστά (ο αριθμός των αληθινών θετικών διαιρεμένος με τον αριθμό των αληθινών θετικών συν ψευδώς αρνητικών).
- Βαθμολογία F1: Μια συνολική μέτρηση που ουσιαστικά συνδυάζει ακρίβεια και ανάκληση.

Από αυτές τις μετρήσεις, η ακρίβεια είναι η πιο διαισθητική. Ωστόσο, πρέπει να είστε προσεκτικοί σχετικά με τη χρήση της ακρίβειας ως μέτρησης του πόσο καλά λειτουργεί ένα μοντέλο. Ας υποθέσουμε ότι μόνο το 3% του πληθυσμού είναι διαβητικός. Θα μπορούσαμε να δημιουργήσουμε ένα μοντέλο που προβλέπει πάντα το 0 και θα ήταν 97% ακριβές, αλλά δεν θα βοηθούσε στην σωστή πρόβλεψη περιπτώσεων διαβήτη. Για αυτόν τον λόγο, οι περισσότεροι επιστήμονες δεδομένων χρησιμοποιούν άλλες μετρήσεις όπως η ακρίβεια και η ανάκληση για να αξιολογήσουν την απόδοση του μοντέλου ταξινόμησης.

**Επιλογή κατωφλίου**

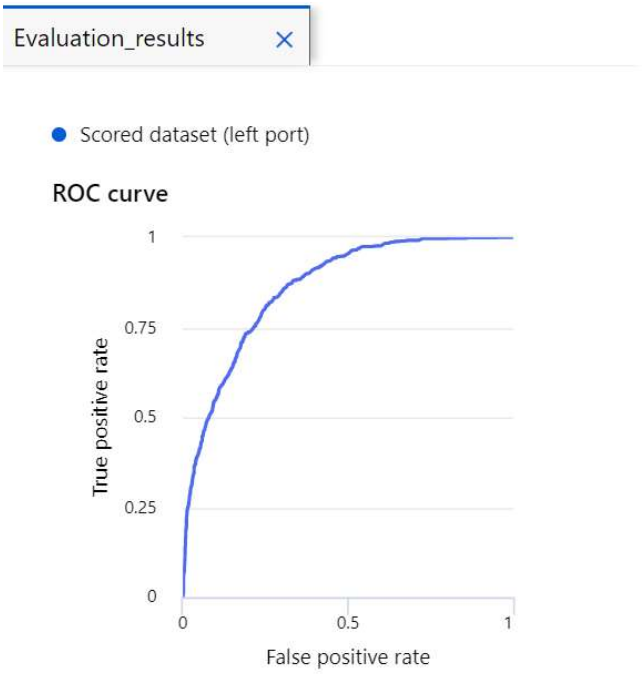
Ένα μοντέλο ταξινόμησης προβλέπει την πιθανότητα για κάθε πιθανή κλάση. Με άλλα λόγια, το μοντέλο υπολογίζει μια πιθανότητα για κάθε προβλεπόμενη ετικέτα. Στην περίπτωση ενός μοντέλου δυαδικής ταξινόμησης, η προβλεπόμενη πιθανότητα είναι μια τιμή μεταξύ 0 και 1. Από προεπιλογή, μια προβλεπόμενη πιθανότητα που περιλαμβάνει ή πάνω από 0,5 οδηγεί σε πρόβλεψη κλάσης 1, ενώ μια πρόβλεψη κάτω από αυτό το όριο σημαίνει ότι υπάρχει μεγαλύτερη πιθανότητα μιας αρνητικής πρόβλεψης (να θυμηθούμε ότι οι πιθανότητες για όλες τις κλάσεις αθροίζονται σε 1), οπότε η προβλεπόμενη κλάση θα είναι 0.Ο σχεδιαστής διαθέτει ένα χρήσιμο

ρυθμιστικό κατωφλίου για να ελέγχει πώς θα άλλαζε η απόδοση του μοντέλου ανάλογα με το καθορισμένο όριο.




**ROC curve and AUC metric**

Ένας άλλος όρος για ανάκληση είναι True positive rate και έχει μια αντίστοιχη μέτρηση που ονομάζεται False positive rate, η οποία μετρά τον αριθμό των αρνητικών περιπτώσεων που εσφαλμένα προσδιορίστηκαν ως θετικές σε σύγκριση με τον αριθμό των πραγματικών αρνητικών περιπτώσεων. Η γραφική παράσταση αυτών των μετρήσεων μεταξύ τους για κάθε πιθανή τιμή κατωφλίου μεταξύ 0 και 1 οδηγεί σε μια καμπύλη, γνωστή ως καμπύλη ROC. Σε ένα ιδανικό μοντέλο, η καμπύλη θα ανέβαινε μέχρι την αριστερή πλευρά και στην κορυφή, έτσι ώστε να καλύπτει ολόκληρη την περιοχή του γραφήματος. Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη, της μετρικής AUC, (η οποία μπορεί να είναι οποιαδήποτε τιμή από 0 έως 1), τόσο καλύτερη είναι η απόδοση του μοντέλου. Μπορούμε να δούμε την καμπύλη ROC στα Αποτελέσματα Αξιολόγησης.




**Binary Classification**

Σε αυτό το σημείο, θα επικεντρωθούμε σε ένα παράδειγμα δυαδικής ταξινόμησης, όπου το μοντέλο πρέπει να προβλέψει μια ετικέτα που ανήκει σε μία από τις δύο κατηγορίες. Σε αυτήν την άσκηση, θα εκπαιδεύσουμε έναν δυαδικό ταξινομητή για να προβλέψουμε εάν ένας ασθενής θα πρέπει να υποβληθεί σε εξέταση για διαβήτη με βάση ορισμένα ιατρικά δεδομένα.

**Εξερεύνηση των δεδομένων**

Εκτελούμε το ακόλουθο κελί για να φορτώσουμε ένα αρχείο CSV δεδομένων σε ένα πλαίσιο δεδομένων Pandas:[51]

```
import pandas as pd


# load the training dataset
!wget https://raw.githubusercontent.com/MicrosoftDocs/mslearn-introduction-to-machine-learning/main/Data/ml-basics/diabetes.csv
diabetes = pd.read_csv('diabetes.csv')
diabetes.head()
```

Αυτά τα δεδομένα αποτελούνται από διαγνωστικές πληροφορίες για ορισμένους ασθενείς που έχουν ελεγχθεί για διαβήτη.Ας διαχωρίσουμε τα χαρακτηριστικά από τις ετικέτες - θα ονομάσουμε τα χαρακτηριστικά X και την ετικέτα y:

```
# Separate features and labels
features = ['Pregnancies','PlasmaGlucose','DiastolicBloodPressure','TricepsThickness','SerumInsulin','BMI','DiabetesPedigree','Age']
label = 'Diabetic'
X, y = diabetes[features].values, diabetes[label].values
```

```
for n in range(0,4):
    print("Patient", str(n+1), "\n  Features:",list(X[n]), "\n  Label:", y[n])
```

Τώρα ας συγκρίνουμε τις κατανομές χαρακτηριστικών για κάθε τιμή ετικέτας.

```
from matplotlib import pyplot as plt
%matplotlib inline
features = ['Pregnancies','PlasmaGlucose','DiastolicBloodPressure','TricepsThickness','SerumInsulin','BMI','DiabetesPedigree','Age']
for col in features:
    diabetes.boxplot(column=col, by='Diabetic', figsize=(6,6))
    plt.title(col)
plt.show()
```

Για ορισμένα από τα χαρακτηριστικά, υπάρχει μια αξιοσημείωτη διαφορά στην κατανομή για κάθε τιμή ετικέτας. Ειδικότερα, οι εγκυμοσύνες και η ηλικία εμφανίζουν αξιοσημείωτα διαφορετικές κατανομές για διαβητικούς ασθενείς από ό,τι για μη διαβητικούς ασθενείς. Αυτά τα χαρακτηριστικά μπορεί να βοηθήσουν στην πρόβλεψη εάν ένας ασθενής είναι διαβητικός ή όχι.

**Διαχωρισμός των δεδομένων**

Το σύνολο δεδομένων μας περιλαμβάνει γνωστές τιμές για την ετικέτα, επομένως μπορούμε να τις χρησιμοποιήσουμε για να εκπαιδεύσουμε έναν ταξινομητή έτσι ώστε να βρίσκει μια στατιστική σχέση μεταξύ των χαρακτηριστικών και της τιμής της ετικέτας. αλλά πώς θα ξέρουμε αν το μοντέλο μας είναι καλό; Πώς ξέρουμε ότι θα προβλέψει σωστά όταν το χρησιμοποιούμε με νέα δεδομένα με τα οποία δεν έχει εκπαιδευτεί; Λοιπόν, μπορούμε να εκμεταλλευτούμε το γεγονός ότι έχουμε ένα μεγάλο σύνολο δεδομένων με γνωστές τιμές ετικετών, να χρησιμοποιήσουμε μόνο μερικές από αυτές για να εκπαιδεύσουμε το μοντέλο και να κρατήσουμε πίσω για να δοκιμάσουμε το εκπαιδευμένο μοντέλο - επιτρέποντάς μας να συγκρίνουμε τις προβλεπόμενες ετικέτες με τις ήδη γνωστές ετικέτες στο σετ δοκιμής.Στην Python, το πακέτο scikit-learn περιέχει μεγάλο αριθμό συναρτήσεων που μπορούμε να χρησιμοποιήσουμε για να δημιουργήσουμε ένα μοντέλο μηχανικής εκμάθησης - συμπεριλαμβανομένης μιας συνάρτησης train_test_split που διασφαλίζει ότι λαμβάνουμε μια στατιστικά τυχαία κατανομή δεδομένων εκπαίδευσης και δοκιμής. Θα το χρησιμοποιήσουμε για να χωρίσουμε τα δεδομένα σε 70% για εκπαίδευση και να συγκρατήσουμε το 30% για δοκιμές.

```
from sklearn.model_selection import train_test_split
# Split data 70%-30% into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=0)
print ('Training cases: %d\nTest cases: %d' % (X_train.shape[0], X_test.shape[0]))
```

## 6.2 Εκπαίδευση μοντέλου δυαδικής ταξινόμησης

Υπάρχουν διάφοροι αλγόριθμοι που μπορούμε να χρησιμοποιήσουμε για να εκπαιδεύσουμε το μοντέλο. Σε αυτό το παράδειγμα, θα χρησιμοποιήσουμε την Logistic Regression, η οποία (παρά το όνομά της) είναι ένας καλά καθιερωμένος αλγόριθμος ταξινόμησης. Εκτός από τις δυνατότητες και τις ετικέτες εκπαίδευσης, θα χρειαστεί να ορίσουμε μια παράμετρο τακτοποίησης. Αυτό χρησιμοποιείται για να εξουδετερώσει οποιαδήποτε προκατάληψη στο δείγμα και να βοηθήσει το μοντέλο να γενικευτεί καλά αποφεύγοντας την υπερβολική προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. [51]

```
# Train the model
from sklearn.linear_model import LogisticRegression
# Set regularization rate
reg = 0.01
# train a logistic regression model on the training set
model = LogisticRegression(C=1/reg, solver="liblinear").fit(X_train, y_train)
print (model)
```

Τώρα έχουμε εκπαιδεύσει το μοντέλο χρησιμοποιώντας τα δεδομένα εκπαίδευσης, μπορούμε να χρησιμοποιήσουμε τα δεδομένα δοκιμής που κρατήσαμε πίσω για να αξιολογήσουμε πόσο καλά προβλέπει. Και πάλι, το scikit-learn μπορεί να μας βοηθήσει να το κάνουμε αυτό. Ας ξεκινήσουμε χρησιμοποιώντας το μοντέλο για να προβλέψουμε ετικέτες για το δοκιμαστικό μας σύνολο και να συγκρίνουμε τις προβλεπόμενες ετικέτες με τις γνωστές ετικέτες:

```
predictions = model.predict(X_test)
print('Predicted labels: ', predictions)
print('Actual labels:    ' ,y_test)
```

Οι πίνακες ετικετών είναι πολύ μεγάλοι για να εμφανίζονται στην έξοδο του notebook, επομένως μπορούμε να συγκρίνουμε μόνο μερικές τιμές. Ακόμα κι αν εκτυπώσαμε όλες τις προβλεπόμενες και πραγματικές ετικέτες, υπάρχουν πάρα πολλές από αυτές για να γίνει ένας λογικός τρόπος αξιολόγησης του μοντέλου. Ευτυχώς, το scikit-learn έχει μερικά ακόμη κόλπα στο μανίκι του και παρέχει ορισμένες μετρήσεις που μπορούμε να χρησιμοποιήσουμε για να αξιολογήσουμε το μοντέλο.

Το πιο προφανές πράγμα που μπορεί να θέλουμε να κάνουμε είναι να ελέγξουμε την ακρίβεια των προβλέψεων - με απλά λόγια, ποιο ποσοστό των ετικετών προέβλεψε σωστά το μοντέλο;

```
from sklearn.metrics import accuracy_score
print('Accuracy: ', accuracy_score(y_test, predictions))
```

Η ακρίβεια επιστρέφεται ως δεκαδική τιμή - μια τιμή 1,0 θα σήμαινε ότι το μοντέλο είχε το 100% των προβλέψεων σωστές,ενώ μια ακρίβεια 0,0 είναι, λοιπόν, αρκετά άχρηστη.

**Περίληψη**

Εδώ ετοιμάσαμε τα δεδομένα μας χωρίζοντάς τα σε σύνολα δεδομένων δοκιμής και εκπαίδευσης και εφαρμόσαμε λογιστική παλινδρόμηση. Το μοντέλο ήταν σε θέση να προβλέψει εάν οι ασθενείς είχαν διαβήτη με εύλογη ακρίβεια. Στο επόμενο σημείο θα εξετάσουμε εναλλακτικές λύσεις για την ακρίβεια που μπορεί να είναι πολύ πιο χρήσιμες στη μηχανική εκμάθηση.

Η αναφορά ταξινόμησης περιλαμβάνει τις ακόλουθες μετρήσεις για κάθε τάξη (0 και 1)

Ακρίβεια: Από τις προβλέψεις που έκανε το μοντέλο για αυτήν την κατηγορία, ποια αναλογία ήταν σωστή;

Ανάκληση: Από όλες τις περιπτώσεις αυτής της κλάσης στο σύνολο δεδομένων δοκιμής, πόσες εντόπισε το μοντέλο.

F1-Score: Μια μέση μέτρηση που λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση.

Υποστήριξη: Πόσες περιπτώσεις αυτής της κλάσης υπάρχουν στο σύνολο δεδομένων δοκιμής.

Η αναφορά ταξινόμησης περιλαμβάνει επίσης μέσους όρους για αυτές τις μετρήσεις, συμπεριλαμβανομένου ενός σταθμισμένου μέσου όρου που επιτρέπει την ανισορροπία στον αριθμό των περιπτώσεων κάθε κατηγορίας.Επειδή πρόκειται για πρόβλημα δυαδικής ταξινόμησης, η κατηγορία 1 θεωρείται θετική και η ακρίβεια και η ανάκλησή της είναι ιδιαίτερα

Μπορούμε να ανακτήσουμε αυτές τις τιμές από μόνες τους χρησιμοποιώντας τις μετρήσεις precision_score και recall_score στο scikit-learn (που από προεπιλογή προϋποθέτουν ένα δυαδικό μοντέλο ταξινόμησης).

```
from sklearn.metrics import precision_score, recall_score
print("Overall Precision:",precision_score(y_test, predictions))
print("Overall Recall:",recall_score(y_test, predictions))
```

Οι μετρήσεις ακρίβειας και ανάκλησης προέρχονται από τέσσερα πιθανά αποτελέσματα πρόβλεψης:

- Αληθινά θετικά: Η προβλεπόμενη ετικέτα και η πραγματική ετικέτα είναι και οι δύο 1.
- Λανθασμένα θετικά: Η προβλεπόμενη ετικέτα είναι 1, αλλά η πραγματική ετικέτα είναι 0.
- Αρνητικά Λανθασμένα: Η προβλεπόμενη ετικέτα είναι 0, αλλά η πραγματική ετικέτα είναι 1.
- Αληθινά αρνητικά: Η προβλεπόμενη ετικέτα και η πραγματική ετικέτα είναι και οι δύο 0.

Αυτές οι μετρήσεις γενικά παρουσιάζονται σε πίνακα για το σύνολο δοκιμής και εμφανίζονται μαζί ως πίνακας σύγχυσης, ο οποίος έχει την ακόλουθη μορφή:

| | |
|---|---|
| TN | FP |
| FN | TP |

Σημειώστε ότι οι σωστές (αληθινές) προβλέψεις σχηματίζουν μια διαγώνια γραμμή από πάνω αριστερά προς τα κάτω δεξιά - αυτοί οι αριθμοί θα πρέπει να είναι σημαντικά υψηλότεροι από τις ψευδείς προβλέψεις εάν το μοντέλο είναι καλό.

Στην Python, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση sklearn.metrics.confusion_matrix για να βρούμε αυτές τις τιμές για έναν εκπαιδευμένο ταξινομητή:

```
[51]from sklearn.metrics import confusion_matrix

# Print the confusion matrix
cm = confusion_matrix(y_test, predictions)
print (cm)
```

Μέχρι τώρα, θεωρούσαμε τις προβλέψεις από το μοντέλο ως ετικέτες κατηγορίας είτε 1 είτε 0. Στην πραγματικότητα, τα πράγματα είναι λίγο πιο σύνθετα από αυτό. Οι στατιστικοί αλγόριθμοι μηχανικής μάθησης, όπως η λογιστική παλινδρόμηση, βασίζονται στην πιθανότητα. Έτσι, αυτό που στην πραγματικότητα προβλέπεται από έναν δυαδικό ταξινομητή είναι η πιθανότητα ότι η ετικέτα είναι αληθής (P(y)) και η πιθανότητα ότι η ετικέτα είναι ψευδής (1 - P(y)). Μια τιμή κατωφλίου 0,5 χρησιμοποιείται για να αποφασιστεί εάν η προβλεπόμενη ετικέτα είναι 1 (P(y) > 0,5) ή 0 (P(y) <= 0,5). Μπορούμε να χρησιμοποιήσουμε τη μέθοδο predict_proba για να δούμε τα ζεύγη πιθανοτήτων για κάθε περίπτωση:

```
y_scores = model.predict_proba(X_test)
```

```
print(y_scores)
```

Η απόφαση να βαθμολογηθεί μια πρόβλεψη ως 1 ή 0 εξαρτάται από το όριο με το οποίο συγκρίνονται οι προβλεπόμενες πιθανότητες. Εάν αλλάζαμε το όριο, θα επηρέαζε τις προβλέψεις. και επομένως αλλάξτε τις μετρήσεις στον πίνακα σύγχυσης. Ένας συνηθισμένος τρόπος αξιολόγησης ενός ταξινομητή είναι η εξέταση του αληθινού θετικού ποσοστού (που είναι ένα άλλο όνομα για ανάκληση) και του ψευδώς θετικού ποσοστού για μια σειρά πιθανών ορίων. Στη συνέχεια, αυτοί οι ρυθμοί σχεδιάζονται έναντι όλων των πιθανών ορίων για να σχηματιστεί ένα γράφημα γνωστό ως διάγραμμα λαμβανόμενου χαρακτηριστικού χειριστή (ROC), όπως αυτό:

```
from sklearn.metrics import roc_curve
from sklearn.metrics import confusion_matrix
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

# calculate ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_scores[:,1])

# plot ROC curve
fig = plt.figure(figsize=(6, 6))
# Plot the diagonal 50% line
plt.plot([0, 1], [0, 1], 'k--')
# Plot the FPR and TPR achieved by our model
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
```

Το διάγραμμα ROC δείχνει την καμπύλη των αληθών και ψευδών θετικών ποσοστών για διαφορετικές τιμές κατωφλίου μεταξύ 0 και 1. Ένας τέλειος ταξινομητής θα είχε μια καμπύλη που πηγαίνει ευθεία προς τα πάνω στην αριστερή πλευρά και ευθεία στην κορυφή. Η διαγώνια γραμμή κατά μήκος του γραφήματος αντιπροσωπεύει την πιθανότητα σωστής πρόβλεψης με τυχαία πρόβλεψη 50/50. οπότε προφανώς θέλουμε η καμπύλη να είναι υψηλότερη από αυτό (ή το μοντέλο μας δεν είναι καλύτερο από το να μαντεύουμε!).Η περιοχή κάτω από την καμπύλη (AUC) είναι μια τιμή μεταξύ 0 και 1 που ποσοτικοποιεί τη συνολική απόδοση του μοντέλου. Όσο πιο κοντά στο 1 είναι αυτή η τιμή, τόσο καλύτερο είναι το μοντέλο. Για άλλη μια φορά, το scikit-Learn περιλαμβάνει μια συνάρτηση για τον υπολογισμό αυτής της μέτρησης. [51]

```
from sklearn.metrics import roc_auc_score

auc = roc_auc_score(y_test,y_scores[:,1])
```

```
print('AUC: ' + str(auc))
```

**Εκτελούμε προεπεξεργασία σε αγωγό**

Σε αυτήν την περίπτωση, η καμπύλη ROC και η AUC της υποδεικνύουν ότι το μοντέλο αποδίδει καλύτερα από μια τυχαία εικασία, κάτι που δεν είναι κακό, δεδομένου ότι πραγματοποιήσαμε πολύ μικρή προεπεξεργασία των δεδομένων.Στην πράξη, είναι σύνηθες να εκτελούμε κάποια προεπεξεργασία των δεδομένων για να διευκολύνουμε τον αλγόριθμο να προσαρμόσει ένα μοντέλο σε αυτό. Υπάρχει μια τεράστια γκάμα μετασχηματισμών προεπεξεργασίας που μπορούμε να εκτελέσουμε για να ετοιμάσουμε τα δεδομένα μας για μοντελοποίηση, αλλά θα περιοριστούμε σε μερικές κοινές τεχνικές:

- Κλιμάκωση αριθμητικών χαρακτηριστικών ώστε να είναι στην ίδια κλίμακα. Αυτό αποτρέπει τα χαρακτηριστικά με μεγάλες τιμές από το να παράγουν συντελεστές που επηρεάζουν δυσανάλογα τις προβλέψεις.
- Κωδικοποίηση κατηγορικών μεταβλητών. Για παράδειγμα, χρησιμοποιώντας μια τεχνική κωδικοποίησης one hot, μπορούμε να δημιουργήσουμε μεμονωμένα δυαδικά χαρακτηριστικά (true/false) για κάθε πιθανή τιμή κατηγορίας.

Για να εφαρμόσουμε αυτούς τους μετασχηματισμούς προεπεξεργασίας, θα χρησιμοποιήσουμε μια δυνατότητα Scikit-Learn που ονομάζεται pipelines. Αυτά μας επιτρέπουν να ορίσουμε ένα σύνολο βημάτων προεπεξεργασίας που τελειώνουν με έναν αλγόριθμο. Στη συνέχεια, μπορούμε να προσαρμόσουμε ολόκληρη τη διοχέτευση στα δεδομένα, έτσι ώστε το μοντέλο να ενσωματώνει όλα τα βήματα προεπεξεργασίας καθώς και τον αλγόριθμο παλινδρόμησης. Αυτό είναι χρήσιμο, γιατί όταν θέλουμε να χρησιμοποιήσουμε το μοντέλο για να προβλέψουμε τιμές από νέα δεδομένα, πρέπει να εφαρμόσουμε τους ίδιους μετασχηματισμούς (με βάση τις ίδιες στατιστικές κατανομές και κωδικοποιήσεις κατηγορίας που χρησιμοποιούνται με τα δεδομένα εκπαίδευσης).

```
# Train the model
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LogisticRegression
import numpy as np
# Define preprocessing for numeric columns (normalize them so they're on the same scale)
numeric_features = [0,1,2,3,4,5,6]
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())])
# Define preprocessing for categorical features (encode the Age column)
categorical_features = [7]
categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])
# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])
# Create preprocessing and training pipeline
```

```
pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                           ('logregressor', LogisticRegression(C=1/reg, solver="liblinear"))])
# fit the pipeline to train a logistic regression model on the training set
model = pipeline.fit(X_train, (y_train))
print (model)
```

Ο αγωγός περιλαμβάνει τα βήματα προεπεξεργασίας καθώς και την εκπαίδευση μοντέλων.Ας χρησιμοποιήσουμε το μοντέλο που εκπαιδεύτηκε από αυτόν τον αγωγό για να προβλέψουμε ετικέτες για το δοκιμαστικό μας σύνολο και ας συγκρίνουμε τις μετρήσεις απόδοσης με το βασικό μοντέλο που δημιουργήσαμε προηγουμένως.

```
# Get predictions from test data
predictions = model.predict(X_test)
y_scores = model.predict_proba(X_test)
# Get evaluation metrics
cm = confusion_matrix(y_test, predictions)
print ('Confusion Matrix:\n',cm, '\n')
print('Accuracy:', accuracy_score(y_test, predictions))
print("Overall Precision:",precision_score(y_test, predictions))
print("Overall Recall:",recall_score(y_test, predictions))
auc = roc_auc_score(y_test,y_scores[:,1])
print('AUC: ' + str(auc))
# calculate ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_scores[:,1])
# plot ROC curve
fig = plt.figure(figsize=(6, 6))
# Plot the diagonal 50% line
plt.plot([0, 1], [0, 1], 'k--')
# Plot the FPR and TPR achieved by our model
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
```

Τα αποτελέσματα φαίνονται λίγο καλύτερα, επομένως η προεπεξεργασία των δεδομένων έχει κάνει τη διαφορά.

**Δοκιμάζουμε έναν διαφορετικό αλγόριθμο**

Τώρα ας δοκιμάσουμε έναν διαφορετικό αλγόριθμο. Προηγουμένως χρησιμοποιούσαμε έναν αλγόριθμο λογιστικής παλινδρόμησης, ο οποίος είναι ένας γραμμικός αλγόριθμος. Υπάρχουν πολλά είδη αλγορίθμων ταξινόμησης που θα μπορούσαμε να δοκιμάσουμε, όπως:

- Υποστήριξη διανυσματικών μηχανών αλγόριθμοι: Αλγόριθμοι που ορίζουν ένα υπερεπίπεδο που διαχωρίζει τις κλάσεις.
- Αλγόριθμοι που βασίζονται σε δέντρα: Αλγόριθμοι που δημιουργούν ένα δέντρο αποφάσεων για την επίτευξη μιας πρόβλεψης
- Αλγόριθμοι συνόλου: Αλγόριθμοι που συνδυάζουν τις εξόδους πολλαπλών βασικών αλγορίθμων για τη βελτίωση της γενίκευσης.

Αυτή τη φορά, θα χρησιμοποιήσουμε τα ίδια βήματα προεπεξεργασίας όπως πριν, αλλά θα εκπαιδεύσουμε το μοντέλο χρησιμοποιώντας έναν αλγόριθμο συνόλου που ονομάζεται Random Forest που συνδυάζει τις εξόδους πολλαπλών δέντρων τυχαίων αποφάσεων

```
from sklearn.ensemble import RandomForestClassifier
# Create preprocessing and training pipeline
pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                           ('logregressor', RandomForestClassifier(n_estimators=100))])
# fit the pipeline to train a random forest model on the training set
model = pipeline.fit(X_train, (y_train))
print (model)
```

Ας δούμε τις μετρήσεις απόδοσης για το νέο μοντέλο.

```
predictions = model.predict(X_test)
y_scores = model.predict_proba(X_test)
cm = confusion_matrix(y_test, predictions)
print ('Confusion Matrix:\n',cm, '\n')
print('Accuracy:', accuracy_score(y_test, predictions))
print("Overall Precision:",precision_score(y_test, predictions))
print("Overall Recall:",recall_score(y_test, predictions))
auc = roc_auc_score(y_test,y_scores[:,1])
print('\nAUC: ' + str(auc))
# calculate ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_scores[:,1])
# plot ROC curve
fig = plt.figure(figsize=(6, 6))
# Plot the diagonal 50% line
plt.plot([0, 1], [0, 1], 'k--')
# Plot the FPR and TPR achieved by our model
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
```

Χρησιμοποιούμε το μοντέλο για εξαγωγή συμπερασμάτων

Τώρα που έχουμε ένα αρκετά χρήσιμο εκπαιδευμένο μοντέλο, μπορούμε να το αποθηκεύσουμε για χρήση αργότερα για να προβλέψουμε ετικέτες για νέα δεδομένα: [51]

```
import joblib
# Save the model as a pickle file
filename = './diabetes_model.pkl'
joblib.dump(model, filename)
# Load the model from the file
model = joblib.load(filename)
# predict on a new sample
# The model accepts an array of feature arrays (so you can predict the classes of multiple patients in a single call)
# We'll create an array with a single array of features, representing one patient
X_new = np.array([[2,180,74,24,21,23.9091702,1.488172308,22]])
print ('New sample: {}'.format(list(X_new[0])))
# Get a prediction
pred = model.predict(X_new)
# The model returns an array of predictions - one for each set of features submitted
# In our case, we only submitted one patient, so our prediction is the first one in the resulting array.
print('Predicted class is {}'.format(pred[0]))
```

**Αξιολόγηση των μοντέλων ταξινόμησης**

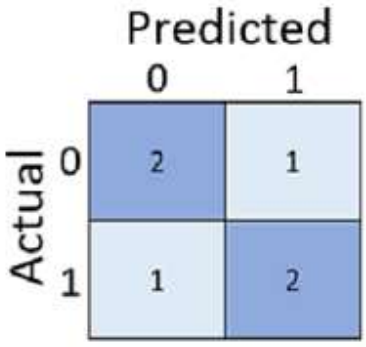
Η ακρίβεια της εκπαίδευσης ενός μοντέλου ταξινόμησης εξαρτάται σημαντικά από το πόσο καλά θα λειτουργήσει αυτό το μοντέλο όταν του παρέχονται νέα δεδομένα. Εξάλλου, εκπαιδεύουμε μοντέλα ώστε να μπορούν να χρησιμοποιηθούν σε νέα δεδομένα που βρίσκουμε στον πραγματικό κόσμο. Έτσι, αφού έχουμε εκπαιδεύσει ένα μοντέλο ταξινόμησης, θα πρέπει να αξιολογήσουμε την απόδοση του σε ένα σύνολο νέων, αόρατων δεδομένων.

Στις προηγούμενες ενότητες, δημιουργήσαμε ένα μοντέλο που θα πρόβλεψε εάν ένας ασθενής είχε διαβήτη ή όχι με βάση το επίπεδο γλυκόζης στο αίμα του. Τώρα, όταν εφαρμόζονται σε ορισμένα δεδομένα που δεν ήταν μέρος του σετ εκπαίδευσης, λαμβάνουμε τις ακόλουθες προβλέψεις:

| x | y | ŷ |
|---|---|---|
| 83 | 0 | 0 |
| 119 | 1 | 1 |
| 104 | 1 | 0 |
| 105 | 0 | 1 |
| 86 | 0 | 0 |
| 109 | 1 | 1 |

**Πίνακας *8 προβλέψεων***

Το x αναφέρεται στο επίπεδο γλυκόζης στο αίμα, το y αναφέρεται στο εάν είναι πραγματικά διαβητικό και το ŷ αναφέρεται στην πρόβλεψη του μοντέλου ως προς το εάν είναι διαβητικός ή όχι. Ο απλός υπολογισμός του πόσες προβλέψεις ήταν σωστές είναι μερικές φορές παραπλανητικός ή πολύ απλοϊκός για να κατανοήσουμε τα είδη των σφαλμάτων που θα κάνει στον πραγματικό κόσμο. Για να λάβουμε πιο λεπτομερείς πληροφορίες, μπορούμε να καταγράψουμε τα αποτελέσματα σε μια δομή που ονομάζεται πίνακας σύγχυσης, ως εξής:




**Πίνακας 9** πίνακας σύγχυσης

Ο πίνακας σύγχυσης δείχνει τον συνολικό αριθμό των περιπτώσεων όπου:

- Το μοντέλο προέβλεψε 0 και η πραγματική ετικέτα είναι 0 (αληθινά αρνητικά, επάνω αριστερά)
- Το μοντέλο προέβλεψε το 1 και η πραγματική ετικέτα είναι 1 (αληθινά θετικά, κάτω δεξιά)
- Το μοντέλο προέβλεψε το 0 και η πραγματική ετικέτα είναι 1 (ψευδώς αρνητικά, κάτω αριστερά)
- Το μοντέλο προέβλεψε το 1 και η πραγματική ετικέτα είναι 0 (ψευδώς θετικά, πάνω δεξιά)

Τα κελιά στη μήτρα σύγχυσης είναι συχνά σκιασμένα, έτσι ώστε οι υψηλότερες τιμές να έχουν βαθύτερη απόχρωση. Αυτό διευκολύνει την προβολή μιας ισχυρής διαγώνιας τάσης από πάνω αριστερά προς τα κάτω δεξιά, επισημαίνοντας τα κελιά όπου η προβλεπόμενη τιμή και η πραγματική τιμή είναι ίδια.

Από αυτές τις βασικές τιμές, μπορούμε να υπολογίσουμε μια σειρά από άλλες μετρήσεις που μπορούν να μας βοηθήσουν να αξιολογήσουμε την απόδοση του μοντέλου. Για παράδειγμα:

- Accuracy: (TP+TN)/(TP+TN+FP+FN) - από όλες τις προβλέψεις, πόσες ήταν σωστές;
- Recall: TP/(TP+FN) - από όλες τις περιπτώσεις που είναι θετικές, πόσες εντόπισε το μοντέλο;
- Precision: TP/(TP+FP) - από όλες τις περιπτώσεις που το μοντέλο προέβλεψε να είναι θετικές, πόσες είναι πραγματικά θετικές;

**Δημιουργία πολυκλασικών μοντέλων ταξινόμησης**

Είναι επίσης δυνατό να δημιουργηθούν μοντέλα ταξινόμησης πολλαπλών τάξεων, στα οποία υπάρχουν περισσότερες από δύο πιθανές κατηγορίες. Για παράδειγμα, η κλινική υγείας μπορεί να επεκτείνει το μοντέλο διαβήτη για να ταξινομήσει τους ασθενείς ως:

Μη διαβητικός
Διαβητικός τύπου 1
Διαβητικός τύπου 2

Οι μεμονωμένες τιμές πιθανότητας κλάσης θα εξακολουθούσαν να αθροίζονται στο σύνολο του 1, καθώς ο ασθενής είναι σίγουρα μόνο σε μία από τις τρεις κατηγορίες και η πιο πιθανή κατηγορία θα προβλεπόταν από το μοντέλο.

## 6.3 Χρήση μοντέλων ταξινόμησης πολλαπλών τάξεων

Η ταξινόμηση πολλαπλών κλάσεων μπορεί να θεωρηθεί ως ένας συνδυασμός πολλαπλών δυαδικών ταξινομητών. Υπάρχουν δύο τρόποι με τους οποίους προσεγγίζουμε το πρόβλημα:

[51]One vs Rest (OVR), όπου δημιουργείται ένας ταξινομητής για κάθε πιθανή τιμή κλάσης, με θετικό αποτέλεσμα για περιπτώσεις όπου η πρόβλεψη είναι αυτή η κατηγορία και αρνητικές προβλέψεις για περιπτώσεις όπου η πρόβλεψη είναι οποιαδήποτε άλλη κατηγορία. Για παράδειγμα, ένα πρόβλημα ταξινόμησης με τέσσερις πιθανές κατηγορίες σχήματος (τετράγωνο, κύκλος, τρίγωνο, εξάγωνο) θα απαιτούσε τέσσερις ταξινομητές που προβλέπουν:

τετράγωνο ή όχι
κύκλος ή όχι
τρίγωνο ή όχι
εξάγωνο ή όχι

One vs One (OVO), στο οποίο δημιουργείται ένας ταξινομητής για κάθε πιθανό ζεύγος κλάσεων. Το πρόβλημα ταξινόμησης με τέσσερις κατηγορίες σχήματος θα απαιτούσε τους ακόλουθους δυαδικούς ταξινομητές:

τετράγωνο ή κύκλο
τετράγωνο ή τρίγωνο
τετράγωνο ή εξάγωνο
κύκλο ή τρίγωνο
κύκλος ή εξάγωνο
τρίγωνο ή εξάγωνο

Και στις δύο προσεγγίσεις, το συνολικό μοντέλο πρέπει να λάβει υπόψη όλες αυτές τις προβλέψεις για να καθορίσει σε ποια κατηγορία ανήκει το αντικείμενο.

Ευτυχώς, στα περισσότερα πλαίσια μηχανικής μάθησης, συμπεριλαμβανομένου του scikit-learn, η εφαρμογή ενός μοντέλου ταξινόμησης πολλαπλών κλάσεων δεν είναι σημαντικά πιο περίπλοκη από τη δυαδική ταξινόμηση - και στις περισσότερες περιπτώσεις, οι εκτιμητές που χρησιμοποιούνται για τη δυαδική ταξινόμηση υποστηρίζουν έμμεσα την ταξινόμηση πολλαπλών κλάσεων αφαιρώντας έναν αλγόριθμο OVR, ένα OVO αλγόριθμο ή επιτρέποντας την επιλογή ενός από τα δύο.

**Εκπαίδευση και αξιολόγηση μοντέλων ταξινόμησης πολλαπλών τάξεων**

**Ταξινόμηση πολλαπλών τάξεων**

Στο τελευταίο σημείο, εξετάσαμε τη δυαδική ταξινόμηση. Αυτό λειτουργεί καλά όταν οι παρατηρήσεις δεδομένων μπορούν να ταξινομηθούν σε μία από δύο κατηγορίες ή κλάσεις, όπως "αληθές" ή "ψευδές". Όταν τα δεδομένα μπορούν να ταξινομηθούν σε περισσότερες από δύο κατηγορίες, θα πρέπει να χρησιμοποιείται ένας αλγόριθμος ταξινόμησης πολλαπλών κατηγοριών.Η ταξινόμηση πολλαπλών κλάσεων μπορεί να θεωρηθεί ως συνδυασμός πολλών δυαδικών ταξινομητών .

**Εξερευνούμε τα δεδομένα**

Θα χρησιμοποιήσουμε ένα σύνολο δεδομένων που περιέχει παρατηρήσεις τριών διαφορετικών ειδών πιγκουίνων.

```
import pandas as pd
# load the training dataset
!wget https://raw.githubusercontent.com/MicrosoftDocs/mslearn-introduction-to-machine-learning/main/Data/ml-basics/penguins.csv
penguins = pd.read_csv('penguins.csv')
# Display a random sample of 10 observations
sample = penguins.sample(10)
sample
```

Το σύνολο δεδομένων περιέχει τις ακόλουθες στήλες:

- CulmenLength: Το μήκος σε mm του culmen του πιγκουίνου.
- CulmenDepth: Το βάθος σε mm του culmen του πιγκουίνου.
- FlipperLength: Το μήκος σε mm του πτερυγίου του πιγκουίνου.
- BodyMass: Η μάζα σώματος του πιγκουίνου σε γραμμάρια.
- Species: Μια ακέραια τιμή που αντιπροσωπεύει το είδος του πιγκουίνου.

Η στήλη Είδη είναι η ετικέτα που θέλουμε να εκπαιδεύσουμε ένα μοντέλο να προβλέπει. Το σύνολο δεδομένων περιλαμβάνει τρία πιθανά είδη, τα οποία κωδικοποιούνται ως 0, 1 και 2. Τα πραγματικά ονόματα ειδών αποκαλύπτονται από τον παρακάτω κώδικα:

```
penguin_classes = ['Adelie', 'Gentoo', 'Chinstrap']
print(sample.columns[0:5].values, 'SpeciesName')
for index, row in penguins.sample(10).iterrows():
```

```
    print('[',row[0], row[1], row[2], row[3], int(row[4]),']',penguin_classes[int(row[4])])
# Count the number of null values for each column
penguins.isnull().sum()
```

Φαίνεται ότι λείπουν ορισμένες τιμές χαρακτηριστικών, αλλά δεν λείπουν ετικέτες. Ας σκάψουμε λίγο πιο βαθιά και ας δούμε τις σειρές που περιέχουν μηδενικά.

```
# Show rows containing nulls
penguins[penguins.isnull().any(axis=1)]
```

Υπάρχουν δύο σειρές που δεν περιέχουν καθόλου τιμές χαρακτηριστικών (NaN σημαίνει "not a number"), επομένως αυτές δεν θα είναι χρήσιμες για την εκπαίδευση ενός μοντέλου. Ας τα απορρίψουμε από το σύνολο δεδομένων.

```
# Drop rows containing NaN values
penguins=penguins.dropna()
#Confirm there are now no nulls
penguins.isnull().sum()
```

Ας διερευνήσουμε πώς σχετίζονται τα χαρακτηριστικά με την ετικέτα δημιουργώντας μερικά γραφήματα πλαισίου. [51]

```
from matplotlib import pyplot as plt
%matplotlib inline
penguin_features = ['CulmenLength','CulmenDepth','FlipperLength','BodyMass']
penguin_label = 'Species'
for col in penguin_features:
    penguins.boxplot(column=col, by=penguin_label, figsize=(6,6))
    plt.title(col)
plt.show()
```

Από τις γραφικές παραστάσεις του πλαισίου, φαίνεται ότι τα είδη 0 και 2 (Adelie και Chinstrap) έχουν παρόμοια προφίλ δεδομένων για το βάθος culmen, το μήκος του πτερυγίου και τη μάζα σώματος, αλλά τα Chinstraps τείνουν να έχουν μακρύτερα culmen. Το είδος 1 (Gentoo) τείνει να έχει αρκετά σαφώς διαφοροποιημένα χαρακτηριστικά από τα άλλα. που θα μας βοηθήσει να εκπαιδεύσουμε ένα καλό μοντέλο ταξινόμησης.

**Προετοιμάζουμε τα δεδομένα**

Όπως και στην περίπτωση της δυαδικής ταξινόμησης, τα χαρακτηριστικά και οι ετικέτες πρέπει να διαχωριστούν και τα δεδομένα να χωριστούν σε υποσύνολα εκπαίδευσης και επικύρωσης πριν από την εκπαίδευση του μοντέλου. Θα εφαρμόσουμε επίσης μια τεχνική στρωματοποίησης κατά τον διαχωρισμό των δεδομένων για τη διατήρηση της αναλογίας κάθε τιμής ετικέτας στα σύνολα δεδομένων εκπαίδευσης και επικύρωσης.

```
from sklearn.model_selection import train_test_split
# Separate features and labels
penguins_X, penguins_y = penguins[penguin_features].values, penguins[penguin_label].values
# Split data 70%-30% into training set and test set
x_penguin_train, x_penguin_test, y_penguin_train, y_penguin_test = train_test_split(penguins_X, penguins_
y,
                                                    test_size=0.30,
                                                    random_state=0,
                                                    stratify=penguins_y)
print ('Training Set: %d, Test Set: %d \n' % (x_penguin_train.shape[0], x_penguin_test.shape[0]))
```

**Εκπαιδεύουμε και αξιολογούμε έναν ταξινομητή πολλαπλών κλάσεων**

Τώρα που έχουμε ένα σύνολο χαρακτηριστικών εκπαίδευσης και αντίστοιχες ετικέτες εκπαίδευσης, μπορούμε να προσαρμόσουμε έναν αλγόριθμο ταξινόμησης πολλαπλών κλάσεων στα δεδομένα για να δημιουργήσουμε ένα μοντέλο. Οι περισσότεροι αλγόριθμοι ταξινόμησης scikit-learn υποστηρίζουν εγγενώς την ταξινόμηση πολλαπλών κλάσεων. Θα δοκιμάσουμε έναν αλγόριθμο λογιστικής παλινδρόμησης.

```
from sklearn.linear_model import LogisticRegression
# Set regularization rate
reg = 0.1
# train a logistic regression model on the training set
multi_model = LogisticRegression(C=1/reg, solver='lbfgs', multi_class='auto', max_iter=10000).fit(x_penguin
_train, y_penguin_train)
print (multi_model)
```

Τώρα μπορούμε να χρησιμοποιήσουμε το εκπαιδευμένο μοντέλο για να προβλέψουμε τις ετικέτες για τα χαρακτηριστικά δοκιμής και να συγκρίνουμε τις προβλεπόμενες ετικέτες με τις πραγματικές ετικέτες:

```
penguin_predictions = multi_model.predict(x_penguin_test)
print('Predicted labels: ', penguin_predictions[:15])
print('Actual labels   : ' ,y_penguin_test[:15])
```

Ας δούμε μια αναφορά ταξινόμησης.

```
from sklearn. metrics import classification_report
print(classification_report(y_penguin_test, penguin_predictions))
```

Όπως και με τη δυαδική ταξινόμηση, η αναφορά περιλαμβάνει μετρήσεις ακριβείας και ανάκλησης για κάθε κατηγορία. Ωστόσο, ενώ με τη δυαδική ταξινόμηση θα μπορούσαμε να

επικεντρωθούμε στις βαθμολογίες για τη θετική τάξη. Σε αυτήν την περίπτωση, υπάρχουν πολλές κλάσεις, επομένως πρέπει να εξετάσουμε μια συνολική μέτρηση (είτε τη μακροεντολή είτε τον σταθμισμένο μέσο όρο) για να καταλάβουμε πόσο καλά αποδίδει το μοντέλο και στις τρεις κατηγορίες.Μπορούμε να λάβουμε τις συνολικές μετρήσεις ξεχωριστά από την αναφορά χρησιμοποιώντας τις τάξεις βαθμολογίας μετρήσεων scikit-learn, αλλά με αποτελέσματα πολλαπλών κλάσεων πρέπει να καθορίσουμε ποια μέση μέτρηση θέλουμε να χρησιμοποιήσουμε για ακρίβεια και ανάκληση.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score
print("Overall Accuracy:",accuracy_score(y_penguin_test, penguin_predictions))
print("Overall Precision:",precision_score(y_penguin_test, penguin_predictions, average='macro'))
print("Overall Recall:",recall_score(y_penguin_test, penguin_predictions, average='macro'))
```

Τώρα ας δούμε τον πίνακα σύγχυσης για το μοντέλο μας:

```
from sklearn.metrics import confusion_matrix
# Print the confusion matrix
mcm = confusion_matrix(y_penguin_test, penguin_predictions)
print(mcm)
```

Ο πίνακας σύγχυσης δείχνει την τομή των προβλεπόμενων και των πραγματικών τιμών ετικέτας για κάθε τάξη - με απλά λόγια, οι διαγώνιες τομές από πάνω αριστερά προς κάτω δεξιά υποδεικνύουν τον αριθμό των σωστών προβλέψεων. Όταν ασχολούμαστε με πολλές κατηγορίες, είναι γενικά πιο διαισθητικό να το απεικονίσουμε ως χάρτη θερμότητας, όπως αυτό:

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
plt.imshow(mcm, interpolation="nearest", cmap=plt.cm.Blues)
plt.colorbar()
tick_marks = np.arange(len(penguin_classes))
plt.xticks(tick_marks, penguin_classes, rotation=45)
plt.yticks(tick_marks, penguin_classes)
plt.xlabel("Predicted Species")
plt.ylabel("Actual Species")
plt.show()
```

Τα πιο σκούρα τετράγωνα στην γραφική παράσταση του πίνακα σύγχυσης υποδεικνύουν μεγάλους αριθμούς περιπτώσεων και ελπίζουμε ότι μπορούμε να δούμε μια διαγώνια γραμμή πιο σκούρων τετραγώνων που υποδεικνύει περιπτώσεις όπου η προβλεπόμενη και η πραγματική ετικέτα είναι ίδιες.Για μοντέλα ταξινόμησης πολλαπλών κατηγοριών, δεν είναι δυνατή μια ενιαία καμπύλη ROC που να δείχνει τα

πραγματικά θετικά έναντι των ψευδώς θετικών ποσοστών. Ωστόσο, μπορούμε να χρησιμοποιήσουμε τις τιμές για κάθε τάξη σε σύγκριση One vs Rest (OVR) για να δημιουργήσουμε ένα γράφημα ROC για κάθε τάξη.

```
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
# Get class probability scores
penguin_prob = multi_model.predict_proba(x_penguin_test)
# Get ROC metrics for each class
fpr = {}
tpr = {}
thresh ={}
for i in range(len(penguin_classes)):
    fpr[i], tpr[i], thresh[i] = roc_curve(y_penguin_test, penguin_prob[:,i], pos_label=i)
# Plot the ROC chart
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label=penguin_classes[0] + ' vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label=penguin_classes[1] + ' vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label=penguin_classes[2] + ' vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.show()
```

Για να ποσοτικοποιήσουμε την απόδοση ROC, μπορούμε να υπολογίσουμε μια συνολική περιοχή κάτω από τη βαθμολογία της καμπύλης που υπολογίζεται κατά μέσο όρο σε όλες τις καμπύλες OVR.

```
auc = roc_auc_score(y_penguin_test,penguin_prob, multi_class='ovr')
print('Average AUC:', auc)
```

**Προεπεξεργασία δεδομένων σε αγωγό**

Και πάλι, όπως και με τη δυαδική ταξινόμηση, μπορούμε να χρησιμοποιήσουμε μια διοχέτευση για να εφαρμόσουμε βήματα προ-επεξεργασίας στα δεδομένα πριν τα τοποθετήσουμε σε έναν αλγόριθμο για την εκπαίδευση ενός μοντέλου. Ας δούμε αν μπορούμε να βελτιώσουμε τον προγνωστικό των πιγκουίνων, κλιμακώνοντας τα αριθμητικά χαρακτηριστικά σε βήματα μετασχηματισμού πριν από την προπόνηση. Θα δοκιμάσουμε επίσης έναν διαφορετικό αλγόριθμο (μια μηχανή υποστήριξης διανυσμάτων) [51]

```
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.svm import SVC
# Define preprocessing for numeric columns (scale them)
```

```
feature_columns = [0,1,2,3]
feature_transformer = Pipeline(steps=[
   ('scaler', StandardScaler())
   ])
# Create preprocessing steps
preprocessor = ColumnTransformer(
   transformers=[
      ('preprocess', feature_transformer, feature_columns)])
# Create training pipeline
pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                           ('regressor', SVC(probability=True))])
# fit the pipeline to train a linear regression model on the training set
multi_model = pipeline.fit(x_penguin_train, y_penguin_train)
print (multi_model)
```

Τώρα μπορούμε να αξιολογήσουμε το νέο μοντέλο.

```
# Get predictions from test data
penguin_predictions = multi_model.predict(x_penguin_test)
penguin_prob = multi_model.predict_proba(x_penguin_test)
# Overall metrics
print("Overall Accuracy:",accuracy_score(y_penguin_test, penguin_predictions))
print("Overall Precision:",precision_score(y_penguin_test, penguin_predictions, average='macro'))
print("Overall Recall:",recall_score(y_penguin_test, penguin_predictions, average='macro'))
print('Average AUC:', roc_auc_score(y_penguin_test,penguin_prob, multi_class='ovr'))
# Confusion matrix
plt.imshow(mcm, interpolation="nearest", cmap=plt.cm.Blues)
plt.colorbar()
tick_marks = np.arange(len(penguin_classes))
plt.xticks(tick_marks, penguin_classes, rotation=45)
plt.yticks(tick_marks, penguin_classes)
plt.xlabel("Predicted Species")
plt.ylabel("Actual Species")
plt.show()
```

Χρησιμοποιούμε το μοντέλο με νέες παρατηρήσεις δεδομένων.Τώρα ας αποθηκεύσουμε το εκπαιδευμένο μοντέλο μας για να μπορέσουμε να το χρησιμοποιήσουμε ξανά αργότερα.

```
import joblib
# Save the model as a pickle file
```

```
filename = './penguin_model.pkl'
joblib.dump(multi_model, filename)
```

Τώρα έχουμε ένα εκπαιδευμένο μοντέλο. Ας το χρησιμοποιήσουμε για να προβλέψουμε την τάξη μιας νέας παρατήρησης πιγκουίνου: [51]

```
# Load the model from the file
multi_model = joblib.load(filename)
# The model accepts an array of feature arrays (so you can predict the classes of multiple penguin observatio
ns in a single call)
# We'll create an array with a single array of features, representing one penguin
x_new = np.array([[50.4,15.3,224,5550]])
print ('New sample: {}'.format(x_new[0]))
# The model returns an array of predictions - one for each set of features submitted
# In our case, we only submitted one penguin, so our prediction is the first one in the resulting array.
penguin_pred = multi_model.predict(x_new)[0]
print('Predicted class is', penguin_classes[penguin_pred])
```

Μπορούμε επίσης να υποβάλουμε μια παρτίδα παρατηρήσεων πιγκουίνου στο μοντέλο και να λάβουμε πίσω μια πρόβλεψη για κάθε ένα.

```
# This time our input is an array of two feature arrays
x_new = np.array([[49.5,18.4,195, 3600],
      [38.2,20.1,190,3900]])
print ('New samples:\n{}'.format(x_new))
# Call the web service, passing the input data
predictions = multi_model.predict(x_new)
# Get the predicted classes.
for prediction in predictions:
  print(prediction, '(' + penguin_classes[prediction] +')')
```

# 7. Γλώσσες προγραμματισμού στη Μηχανική Μάθηση

Στο κεφάλαιο αυτό θα παρουσιάσουμε δύο απαραίτητες γλώσσες προγραμματισμού στο κόσμο της μηχανικής μάθησης. Οι δύο αυτές γλώσσες είναι η R και η Python. Και η Python και η R αποτελούν εξαιρετικές επιλογές για την επιστήμη δεδομένων, με κάθε μία από αυτές να προσφέρει μοναδικά πλεονεκτήματα και εφαρμογές που μπορούν να καλύψουν διαφορετικές ανάγκες και προτιμήσεις.

## 7.1 Παρουσίαση της Γλώσσας Προγραμματισμού R

Ένα από τα κύρια πλεονεκτήματα της R είναι η εκτεταμένη συλλογή στατιστικών μεθόδων του, που κυμαίνονται από γραμμικό και μη γραμμικό προγραμματισμό έως ανάλυση χρονοσειρών και ομαδοποίηση. [91] Επιπλέον, το R προσφέρει ένα ευρύ φάσμα τεχνικών για τη δημιουργία διαφόρων τύπων γραφημάτων και οι βιβλιοθήκες του επεκτείνονται συνεχώς για να ικανοποιήσουν τις εξελισσόμενες ανάγκες των χρηστών. Ενώ η γλώσσα προγραμματισμού S προτιμάται συχνά για ερευνητικές δραστηριότητες που περιλαμβάνουν στατιστικές μεθόδους, η R συμμετέχει επίσης ενεργά σε αυτές τις δραστηριότητες παρέχοντας μια λογική ανοιχτού κώδικα. Η γλώσσα προγραμματισμού R παίζει σημαντικό ρόλο στην επίλυση προβλημάτων που σχετίζονται με τεχνικές αριθμητικής ανάλυσης και μηχανικής μάθησης. Αναγνωρίζεται ευρέως ως το ιδανικό περιβάλλον για στατιστικούς υπολογισμούς και γραφήματα, καθιστώντας το ένα ισχυρό εργαλείο στον τομέα της Επιχειρηματικής Ανάλυσης. Το R αναπτύχθηκε από τους Ross Ihaka και Robert Gentleman στο Πανεπιστήμιο του Auckland στη Νέα Ζηλανδία. Επί του παρόντος, συντηρείται και αναπτύσσεται περαιτέρω από την R Development Core Team, η οποία περιλαμβάνει μέλη όπως τα Chambers. Το όνομα "R" προέρχεται από τα αρχικά των δημιουργών του, καθώς και μια παιχνιδιάρικη αναφορά στη γλώσσα προγραμματισμού S. Παρά το γεγονός ότι είναι σχετικά νέο σε σύγκριση με άλλες γλώσσες, η R δημιουργήθηκε για πρώτη φορά το 1992, με την αρχική έκδοση να κυκλοφόρησε το 1995 και τη σταθερή έκδοση beta το 2000. Στη σημερινή εποχή που βασίζεται στα δεδομένα, όπου οι μηχανές και οι συσκευές παράγουν τεράστιες ποσότητες δεδομένων, η δημοτικότητα του R αναμένεται να αυξηθεί εκθετικά. Ωστόσο, είναι σημαντικό για τους προγραμματιστές και τους χρήστες να γνωρίζουν τόσο τα πλεονεκτήματα όσο και τα μειονεκτήματα που συνεπάγεται η χρήση του R.

**Βασικά Πλεονεκτήματα:**

- **Εξειδικευμένη για Στατιστικό Υπολογισμό**:

Η R έχει σχεδιαστεί ειδικά για στατιστική ανάλυση και υπολογισμούς, καθιστώντας την ένα εξαιρετικά ισχυρό εργαλείο για επιστήμονες δεδομένων που εστιάζουν στην ανάλυση στατιστικών δεδομένων.

- **Ισχυρή Οπτικοποίηση Δεδομένων:**

Η R είναι ιδιαίτερα ισχυρή στην οπτικοποίηση δεδομένων, με βιβλιοθήκες όπως η ggplot2 που προσφέρουν ευρύ φάσμα δυνατοτήτων για τη δημιουργία εκτενών και προσαρμόσιμων γραφημάτων και πινάκων.

- **Πλήθος Πακέτων**:

Η R διαθέτει τεράστιο αριθμό πακέτων για διάφορες στατιστικές και αναλυτικές εργασίες, που φιλοξενούνται στο Comprehensive R Archive Network (CRAN), προσφέροντας μια πληθώρα επιλογών και εργαλεία για τον χρήστη.

- **Προηγμένα Στατιστικά Μοντέλα**:

Η R είναι η προτιμώμενη επιλογή για την κατασκευή και εφαρμογή προηγμένων στατιστικών μοντέλων και αναλύσεων, καθιστώντας την πολύ δημοφιλή σε ακαδημαϊκούς και ερευνητικούς κύκλους.

Η φύση ανοιχτού κώδικα της R είναι ένα σημαντικό πλεονέκτημα που το έχει ξεχωρίσει από την έναρξή του. Αυτό σημαίνει ότι όχι μόνο η R είναι ελεύθερα προσβάσιμη στο Διαδίκτυο, αλλά ο πηγαίος κώδικας της είναι επίσης ανοιχτός για τροποποίηση και επέκταση από οποιονδήποτε χωρίς να χρειάζεται άδεια. Αυτή η ανοιχτή προσέγγιση ενθαρρύνει τη συνεργασία και την καινοτομία εντός της κοινότητας R, οδηγώντας σε συνεχή βελτίωση και εξέλιξη της γλώσσας. Ένα από τα βασικά πλεονεκτήματα της χρήσης της R είναι το εκτεταμένο οικοσύστημα πρόσθετων «επεκτάσεων» που ενισχύουν και εμπλουτίζουν την εμπειρία του χρήστη. Εάν υπάρχει μια νέα στατιστική τεχνική ή εργαλείο, το πιθανότερο είναι ότι υπάρχει ήδη διαθέσιμη επέκταση για αυτήν στην R. Αυτό δείχνει ότι η R είναι μια εξαιρετικά προσαρμόσιμη γλώσσα, που δίνει τη δυνατότητα στους προγραμματιστές να δημιουργούν τα δικά τους εργαλεία και να αναλύουν τα δεδομένα αποτελεσματικά και αποτελεσματικά.

Η R προσφέρει μια πληθώρα πλεονεκτημάτων, συμπεριλαμβανομένης της ευελιξίας και της μεγάλης γκάμα χαρακτηριστικών του. Ένα αξιοσημείωτο πλεονέκτημα είναι η ικανότητά του να ενσωματώνει απρόσκοπτα τον κώδικα C και C++ στο περιβάλλον R. Αυτό σημαίνει ότι οι χρήστες δεν περιορίζονται στη

χρήση μόνο της R για όλες τις εργασίες, αλλά μπορούν να επιλέξουν τα καταλληλότερα εργαλεία για κάθε συγκεκριμένη εργασία. Επιπλέον, ο κώδικας που είναι γραμμένος σε R είναι εύκολα προσβάσιμος στον χρήστη, επιτρέποντας γρήγορες και αποτελεσματικές τροποποιήσεις. Ακόμη και μια μικρή προσαρμογή σε μια εργασία απαιτεί μόνο μια μικρή αλλαγή στον κώδικα, με αποτέλεσμα σημαντική εξοικονόμηση χρόνου.

Ένα από τα αξιοσημείωτα επιτεύγματα της R είναι η ικανότητά της να αντικατοπτρίζει τη διαδικασία σκέψης των χρηστών της. Επιπλέον, η R λειτουργεί κυρίως με διανύσματα, αντιμετωπίζοντας τα δεδομένα ως συνεκτικές οντότητες και όχι ως ξεχωριστούς αριθμούς, ευθυγραμμίζοντας πιο στενά με τα ανθρώπινα γνωστικά πρότυπα. Κατά ειρωνικό τρόπο, η επάρκεια σε γλώσσες όπως η C ή η Fortran μπορεί να εμποδίσει τη διαδικασία εκμάθησης και να μειώσει την αποτελεσματικότητα της χρήσης της R. Αυτές οι γλώσσες τείνουν να προσεγγίζουν τα προβλήματα ως προκλήσεις προγραμματισμού, ενώ η R ενθαρρύνει μια πιο φυσική προσέγγιση επίλυσης προβλημάτων.

Η R χρησιμοποιείται ευρέως από πολλές κορυφαίες εταιρείες που απασχολούν επιστήμονες δεδομένων. Συγκεκριμένα, εταιρείες όπως η Google και το Facebook, δεδομένης της φύσης των υπηρεσιών τους, βασίζονται συχνά στην R για ανάλυση δεδομένων. Επιπλέον, οι επιστήμονες δεδομένων της Microsoft προτιμούν το R ως το κύριο εργαλείο τους όταν εφαρμόζουν τεχνικές μηχανικής εκμάθησης σε δεδομένα από διάφορες πηγές όπως το Bing, το Azure, το Office, τα τμήματα πωλήσεων, μάρκετινγκ και οικονομικών. [44]Εκτός από αυτούς τους τεχνολογικούς γίγαντες, πολλές άλλες εταιρείες, όπως η Bank of America, η Ford, η TechCrunch, η Uber και η Trulia υιοθετούν επίσης το R σε μεγάλη κλίμακα.

Η R δεν είναι μόνο ένα εργαλείο που χρησιμοποιείται στις βιομηχανίες, αλλά είναι επίσης ιδιαίτερα δημοφιλής στον ακαδημαϊκό χώρο και στους ερευνητές. Η δημοτικότητά της στα πανεπιστήμια είναι ζωτικής σημασίας, καθώς παρέχει ένα γόνιμο έδαφος για την ανάπτυξη ταλέντων νέων επιστημόνων που μπορούν στη συνέχεια να συμβάλουν στην ανάπτυξη των βιομηχανιών.Η εκπαίδευση στη γλώσσα προγραμματισμού R κατά τη διάρκεια των ακαδημαϊκών χρόνων βοηθάει στην ανάπτυξη των κορυφαίων μυαλών σε αυτόν τον τομέα. Αυτό μπορεί να οδηγήσει σε μεγαλύτερη σημασία της R και στη βιομηχανία, καθώς οι επαγγελματίες που έχουν εκπαιδευτεί σε αυτήν τη γλώσσα καταφέρνουν να φέρουν τις γνώσεις και τις δεξιότητές τους στην αγορά εργασίας. Επιπλέον, καθώς η ανάλυση δεδομένων εξελίσσεται, οι επαγγελματίες πρέπει να διατηρούν στενή επαφή με τους συναδέλφους τους στον ακαδημαϊκό χώρο. Μέσω αυτής της συνεργασίας και του διαλόγου, μπορούν να ενημερώνονται για νέες τεχνικές και μεθοδολογίες, βελτιώνοντας έτσι την αποδοτικότητα και την αποτελεσματικότητά τους στην εργασία τους..

Το πακέτο γραφικής παράστασης και η επέκταση ggplot2 θεωρούνται ευρέως ως μερικά από τα κορυφαία εργαλεία για την οπτικοποίηση δεδομένων στην R. Ένα από τα κύρια πλεονεκτήματα του ggplot2 είναι ότι μόλις οι χρήστες εξοικειωθούν με τη σύνταξή του, αποκτούν επίσης την ικανότητα να παρουσιάζουν αποτελεσματικά και καθαρά τα στοιχεία τους. Επιπλέον, όταν το ggplot2 χρησιμοποιείται σε συνδυασμό με το dplyr, μια άλλη επέκταση R που επικεντρώνεται στην ανάλυση δεδομένων, μπορεί να παράγει τα επιθυμητά αποτελέσματα με ελάχιστη προσπάθεια και χρόνο. Είναι επίσης σημαντικό να σημειωθεί ότι υπάρχει μια υποκείμενη τυπική δομή για την οπτικοποίηση δεδομένων, η οποία παρέχει ένα εξαιρετικά δομημένο πλαίσιο για τη δημιουργία ακριβών και ενημερωτικών απεικονίσεων.

**Δυσκολίες και Μειονεκτήματα της R**

- **Χαμηλή Ταχύτητα Εκτέλεσης:**

Η R είναι γνωστή για την σχετικά χαμηλή απόδοση εκτέλεσης, ειδικά όταν διαχειρίζεται μεγάλα σύνολα δεδομένων ή πολύπλοκους υπολογισμούς. Σε αυτές τις περιπτώσεις, η ταχύτητα της μπορεί να είναι σημαντικά χαμηλότερη σε σύγκριση με άλλες γλώσσες όπως η C++ ή η Java.

- **Υψηλή Κατανάλωση Μνήμης:**

Η R καταναλώνει πολύ μνήμη, καθώς φορτώνει ολόκληρα σύνολα δεδομένων στη μνήμη RAM. Αυτό μπορεί να περιορίσει τη δυνατότητα επεξεργασίας πολύ μεγάλων δεδομένων και να απαιτεί μηχανήματα με υψηλή μνήμη RAM.

- **Απότομη Καμπύλη Εκμάθησης:**

Η R έχει μια απότομη καμπύλη εκμάθησης για νέους χρήστες, κυρίως λόγω της μοναδικής σύνταξης και των στατιστικών μεθόδων που πρέπει να μάθουν. Αυτό μπορεί να αποθαρρύνει τους αρχάριους.

- **Διαχείριση Σφαλμάτων:**

Ο χειρισμός σφαλμάτων στην R μπορεί να είναι δύσκολος και ορισμένα μηνύματα σφάλματος μπορεί να είναι δύσκολα στην κατανόηση και την επίλυση.

- **Μη Ιδανική για Ανάπτυξη Ιστού και Εφαρμογών:**

Ενώ η R είναι εξαιρετική για στατιστική ανάλυση και επιστήμη δεδομένων, δεν είναι η καλύτερη επιλογή για την ανάπτυξη διαδικτυακών εφαρμογών ή γενικών εφαρμογών λογισμικού. Οι δυνατότητες της σε αυτούς τους τομείς είναι περιορισμένες.

- **Περιορισμένο Οικοσύστημα σε Σύγκριση με άλλες Γλώσσες:**

Αν και η R έχει μια ενεργή κοινότητα και πολλές πακέτα, το οικοσύστημά της δεν είναι τόσο εκτεταμένο όσο αυτό της Python, ιδίως όσον αφορά τις βιβλιοθήκες και τα εργαλεία για μη στατιστικές εφαρμογές.

- **Υποστήριξη:**

Η υποστήριξη για την R δεν είναι τόσο ευρεία όσο για άλλες γλώσσες, και μπορεί να είναι δύσκολο να βρεθούν πόροι και απαντήσεις σε συγκεκριμένα προβλήματα.

- **Περιορισμένη Υποστήριξη Πολυνηματικότητας:**

Η R δεν έχει ισχυρή υποστήριξη για πολυνηματικές και παραλληλισμένες εργασίες, γεγονός που μπορεί να περιορίσει την απόδοσή της σε περιπτώσεις που απαιτείται εκτέλεση πολλαπλών διεργασιών ταυτόχρονα.

- **Προβλήματα Συμβατότητας Πακέτων:**

Καθώς η R εξελίσσεται, μπορεί να προκύψουν προβλήματα συμβατότητας μεταξύ παλαιών και νέων εκδόσεων πακέτων. Αυτό μπορεί να προκαλέσει αστάθεια σε έργα που εξαρτώνται από πολλά εξωτερικά πακέτα.

- **Ενσωμάτωση με Άλλες Γλώσσες και Συστήματα:**

Η ενσωμάτωση της R με άλλες γλώσσες προγραμματισμού ή πλατφόρμες μπορεί να είναι δύσκολη, καθιστώντας την λιγότερο ευέλικτη για έργα που απαιτούν πολυγλωσσικές λύσεις.

## 7.2 Παρουσίαση της Γλώσσας Προγραμματισμού Python

Η Python είναι μια υψηλού επιπέδου, διερμηνευμένη γλώσσα προγραμματισμού, ευρέως αναγνωρισμένη για την απλότητα και την ευαναγνωσιμότητά της. Δημιουργήθηκε από τον Guido van Rossum και κυκλοφόρησε για πρώτη φορά το 1991. Η Python δίνει μεγάλη έμφαση στην αναγνωσιμότητα του κώδικα, επιτρέποντας στους προγραμματιστές να εκφράζουν ιδέες και έννοιες με λιγότερες γραμμές κώδικα σε σύγκριση με γλώσσες όπως η C++ ή η Java. Η φιλοσοφία σχεδιασμού της Python βασίζεται στην καθαρότητα και την απλότητα, καθιστώντας την εξαιρετική επιλογή τόσο για αρχάριους όσο και για έμπειρους προγραμματιστές.

**Κύρια Χαρακτηριστικά της Python:**

- **Εύκολη στην Ανάγνωση και τη Γραφή:**

Η σύνταξη της Python είναι απλή και μοιάζει με τη φυσική γλώσσα, γεγονός που την καθιστά ευκολότερη στην εκμάθηση και τη χρήση. Οι καθαρές και κατανοητές γραμμές κώδικα βοηθούν στην ταχύτερη ανάπτυξη και συντήρηση του κώδικα.

- **Διερμηνευμένη Γλώσσα:**

Ο κώδικας Python εκτελείται γραμμή προς γραμμή, γεγονός που απλοποιεί τον εντοπισμό σφαλμάτων και επιτρέπει διαδραστικό προγραμματισμό. Η δυνατότητα άμεσης εκτέλεσης κώδικα διευκολύνει την ανάπτυξη και τον πειραματισμό με νέες ιδέες.

- **Γενικού Σκοπού:**

Η Python είναι εξαιρετικά ευέλικτη και μπορεί να χρησιμοποιηθεί για μια μεγάλη ποικιλία εφαρμογών, όπως ανάπτυξη ιστού, ανάλυση δεδομένων, τεχνητή νοημοσύνη, επιστημονικούς υπολογισμούς, αυτοματοποίηση και πολλά άλλα. Αυτή η πολυχρηστικότητα την καθιστά χρήσιμη σε πολλούς διαφορετικούς τομείς.

- **Εκτενής Βιβλιοθήκη:**

Η Python διαθέτει μια μεγάλη βιβλιοθήκη που παρέχει ενότητες και λειτουργίες για κοινές εργασίες, όπως είσοδος/έξοδος αρχείων, συστημικές κλήσεις και διαχείριση δεδομένων. Αυτές οι βιβλιοθήκες επιταχύνουν την ανάπτυξη και μειώνουν τον αριθμό των γραμμών κώδικα που πρέπει να γράψει ο προγραμματιστής.

- **Διαλειτουργικότητα:**

Η Python είναι διαθέσιμη σε πολλά λειτουργικά συστήματα, όπως Windows, macOS και Linux, καθιστώντας την εξαιρετικά φορητή. Ο ίδιος κώδικας μπορεί να εκτελεστεί σε διαφορετικές πλατφόρμες χωρίς αλλαγές.

- **Δυναμική Τυποποίηση:**

Οι μεταβλητές στην Python δεν χρειάζονται ρητές δηλώσεις, και ο τύπος τους μπορεί να αλλάξει δυναμικά κατά την εκτέλεση του προγράμματος. Αυτό προσφέρει ευελιξία στον προγραμματιστή και επιταχύνει την ανάπτυξη.

- **Κοινότητα και Οικοσύστημα:**

Η Python έχει μια μεγάλη, ενεργή κοινότητα που συνεισφέρει σε ένα πλούσιο οικοσύστημα βιβλιοθηκών και πλαισίων, όπως το Django για ανάπτυξη ιστού, το NumPy και το pandas για ανάλυση δεδομένων, και το TensorFlow και το PyTorch για μηχανική μάθηση. Αυτή η κοινότητα παρέχει επίσης υποστήριξη και πόρους για τους προγραμματιστές, διευκολύνοντας την επίλυση προβλημάτων και την εκμάθηση νέων τεχνολογιών.

**Δημοφιλείς Χρήσεις της Python:**

- **Ανάπτυξη Ιστού:**

Τα πλαίσια όπως το Django και το Flask κάνουν την ανάπτυξη ιστοσελίδων εύκολη και αποδοτική, προσφέροντας εργαλεία και δομές για την κατασκευή ισχυρών και επεκτάσιμων εφαρμογών ιστού.

- **Επιστήμη Δεδομένων και Μηχανική Μάθηση:**

Βιβλιοθήκες όπως οι NumPy, pandas, Matplotlib, Scikit-learn, TensorFlow και PyTorch υποστηρίζουν τη διαχείριση, την ανάλυση και τη μηχανική μάθηση δεδομένων, κάνοντας την Python μια από τις κυριότερες γλώσσες σε αυτούς τους τομείς.

- **Αυτοματοποίηση και Σενάρια:**

Η απλότητα και η ισχυρή βιβλιοθήκη της Python την καθιστούν δημοφιλή επιλογή για τη συγγραφή σεναρίων για αυτοματοποίηση επαναλαμβανόμενων εργασιών, αυξάνοντας την αποδοτικότητα και μειώνοντας τα ανθρώπινα λάθη.

- **Επιστημονικοί Υπολογισμοί:**

Εργαλεία όπως το SciPy και το SymPy χρησιμοποιούνται για επιστημονικούς και μαθηματικούς υπολογισμούς, παρέχοντας προηγμένες δυνατότητες για ερευνητές και επιστήμονες.

- **Ανάπτυξη Λογισμικού:**

Η Python συχνά χρησιμοποιείται για την ανάπτυξη επιτραπέζιων εφαρμογών και εργαλείων λογισμικού, χάρη στην ευκολία χρήσης και την υποστήριξη από πολλά γραφικά περιβάλλοντα.

**Δυσκολίες και Μειονεκτήματα της Python**

Παρόλο που η Python είναι μια εξαιρετικά δημοφιλής και ισχυρή γλώσσα προγραμματισμού, έχει και ορισμένα μειονεκτήματα και περιορισμούς που μπορεί να αποτελέσουν πρόκληση για τους προγραμματιστές. Εδώ παρατίθενται μερικές από τις βασικές δυσκολίες και τα μειονεκτήματα της Python:

- **Χαμηλή Ταχύτητα Εκτέλεσης:**

Η Python είναι μια διερμηνευμένη γλώσσα, γεγονός που σημαίνει ότι ο κώδικας εκτελείται γραμμή προς γραμμή. Αυτό μπορεί να οδηγήσει σε πιο αργή απόδοση σε σύγκριση με γλώσσες όπως η C ή η C++.

- **Βελτιστοποίηση Επεξεργασίας:**

Για εφαρμογές που απαιτούν υψηλή απόδοση και βελτιστοποίηση, η Python μπορεί να μην είναι η καλύτερη επιλογή, καθώς οι απαιτήσεις σε πόρους και η καθυστέρηση μπορεί να είναι υψηλές.

- **Υψηλή Κατανάλωση Μνήμης:**

Η Python χρησιμοποιεί αρκετή μνήμη, κάτι που μπορεί να αποτελέσει πρόβλημα σε εφαρμογές που απαιτούν αποτελεσματική διαχείριση μνήμης ή σε περιβάλλοντα με περιορισμένους πόρους.

- **Λιγότερες Επιλογές για Κινητές Εφαρμογές:**

Αν και υπάρχουν βιβλιοθήκες όπως το Kivy και το BeeWare, η Python δεν είναι τόσο διαδεδομένη για την ανάπτυξη κινητών εφαρμογών όσο άλλες γλώσσες, όπως το Swift για iOS ή το Kotlin για Android.

- **Global Interpreter Lock (GIL):**

Η ύπαρξη του Global Interpreter Lock (GIL) σημαίνει ότι σε περιβάλλοντα πολλών νημάτων (multithreading), μόνο ένα νήμα εκτελείται κάθε φορά. Αυτό μπορεί να περιορίσει την απόδοση των εφαρμογών που απαιτούν πολυνηματική εκτέλεση.

- **Προβλήματα Λόγω Δυναμικής Τυποποίησης:**

Η δυναμική τυποποίηση της Python μπορεί να οδηγήσει σε σφάλματα που εμφανίζονται κατά την εκτέλεση του προγράμματος, αντί να ανιχνεύονται κατά την κατασκευή του (compile-time). Αυτό μπορεί να δυσκολέψει τον εντοπισμό και την επίλυση των σφαλμάτων.

- **Λιγότερες Βιβλιοθήκες και Πλαίσια:**

Σε σύγκριση με άλλες γλώσσες προγραμματισμού, η Python έχει λιγότερες βιβλιοθήκες και πλαίσια για την ανάπτυξη εφαρμογών σε κινητές συσκευές.

- **Προβλήματα Κλιμάκωσης:**

Η ευελιξία και η απλότητα της Python μπορούν να οδηγήσουν σε προγράμματα που είναι δύσκολα στη συντήρηση και την κλιμάκωση, ειδικά σε μεγάλα έργα με πολλούς προγραμματιστές.

- **Μετάβαση από Python 2 σε Python 3:**

Η μετάβαση από Python 2 σε Python 3 δημιούργησε ασυμβατότητες και προκάλεσε προβλήματα σε προγραμματιστές που είχαν μεγάλα έργα γραμμένα στην παλαιότερη έκδοση.

Συμεπρασματικά η συνδυαστική απλότητα, η ευαναγνωσιμότητα και η ευελιξία της Python την καθιστούν μια ισχυρή γλώσσα για μια ευρεία γκάμα εφαρμογών. Οι εκτενείς βιβλιοθήκες και η υποστήριξη από την κοινότητα ενισχύουν περαιτέρω τις δυνατότητές της, καθιστώντας την μια προτιμώμενη επιλογή για πολλούς προγραμματιστές και οργανισμούς. Είτε ενδιαφέρόμαστε για ανάπτυξη ιστού, επιστήμη δεδομένων, αυτοματοποίηση ή ανάπτυξη λογισμικού, η Python προσφέρει τα εργαλεία και τους πόρους που χρειαζόμαστε για να πετύχουμε τους στόχους μας.

Και η Python και η R έχουν σημαντική θέση στην εργαλειοθήκη της επιστήμης δεδομένων. Πολλοί επιστήμονες δεδομένων επιλέγουν να μάθουν και τις δύο γλώσσες για να αξιοποιήσουν τα μοναδικά πλεονεκτήματα της καθεμίας. Αν κάποιος είναι αρχάριος και θέλει να μάθει μια γλώσσα που προσφέρει ευρύ φάσμα εφαρμογών, η Python μπορεί να είναι η καλύτερη επιλογή. Αντίθετα, αν η εργασία επικεντρώνεται στη στατιστική ανάλυση και την οπτικοποίηση δεδομένων, η R μπορεί να είναι πιο κατάλληλη για τις ανάγκες.

Τελικά, η καλύτερη επιλογή εξαρτάται από τις συγκεκριμένες ανάγκες, τους επαγγελματικούς στόχους και τον τύπο των έργων που σκοπεύουμε να ασχοληθούμε.

## Συμπεράσματα – Περίληψη

Στα πλαίσια της παρούσας διπλωματικής μελετήθηκαν διάφοροι αλγόριθμοι μηχανικής μάθησης και εξόρυξης δεδομένων. Αναλύθηκαν τα πλεονεκτήματα και τα μειονεκτήματα των αλγορίθμων και καταλήξαμε ότι αναλόγως του προβλήματος και του στόχου που έχουμε θέσει δεν υπάρχει πιο αποδοτικός αλγόριθμος που θα επιλύει όλα τα προβλήματα. Επίσης αναλύθηκε τι συμβαίνει σε περιπτώσεις που λείπουν δεδομένα και είδαμε εφαρμογές της μηχανικής μάθησης στο cloud (Azure computing) αλλά σε περιπτώσεις IoT.

Πιο αναλυτικά στα κεφάλαια 1 και 2 έγινε παρουσίαση της μηχανικής μάθησης και της εξόρυξης δεδομένων. Παρουσιάστηκε η σχέση της εξόρυξης δεδομένων με την επιστήμη της στατιστικής και μελετήσαμε τα είδη της μηχανική μάθησης. Τέλος έγινε μια σύντομη περιγραφή της εφαρμογής της μηχανικής μάθησης μεγάλης κλίμακας.

Στο κεφάλαιο 3 έγινε ανασκόπηση των βασικότερων αλγορίθμων μηχανικής μάθησης οι οποίοι είναι:

- Γραμμική Παλινδρόμηση (Linear Regression)
- Πολυμεταβλητή Ανάλυση Δεδομένων (MultiVariate Data Analysis, MVDA)
- Τυχαία Δάση (Random Forests)
- Ταξινομητές Naive-Bayes (Naive-Bayes classifiers)
- Δέντρα αποφάσεων (Decision Trees)
- Ο Αλγόριθμος ID3 (Iterative Dichotomiser 3)
- Δημιουργία Δέντρων Αποφάσεων
- Artificial Neural Networks
- Bayesian Neural Nets και το NIPS 2003 Challenge
- Support Vector Machines
- k-Nearest-Neighbor

Έγινε παρουσίαση των αλγορίθμων, πως χρησιμοποιούνται στη μηχανική μάθηση καθώς και παραδείγματα αυτών.

Στο κεφάλαιο 4 έγινε αξιολόγηση της μηχανικής μάθησης δλδ πως αξιολογείται η απόδοση ενός αλγορίθμου μηχανικής μάθησης. Επίσης έγινε παρουσίαση ενός βασικού προβλήματος της μηχανικής μάθησης , αυτό των χαμένων δεδομένων (missing data). Εξετάστηκαν οι διάφορες κατηγορίες των χαμένων δεδομένων και παρουσιάστηκε μια ανάλυση των αλγορίθμων που βοηθούν στην επίλυση του προβλήματος. Τέλος είδαμε την ταξινόμηση στην περίπτωση του προβλήματος των χαμένων δεδομένων.

Στο κεφάλαιο 5 μελετήσαμε τις εφαρμογές και τις τεχνικές της ταξινόμησης. Πιο συγκεκριμένα παρουσιάστηκε μια μελέτη ταξινόμησης στην περίπτωση των έξυπνων πόλεων και της ταξινόμησης του δικτύου. Επίσης έγινε παρουσίαση διαφόρων μεθόδων ταξινόμησης του δικτύου και παρουσίαση της απόδοσης των διαφόρων αλγορίθμων που χρησιμοποίησε η μελέτη.

Στο κεφάλαιο 6 έγινε παρουσίαση της χρήσης του Azure της Microsoft και πως αυτό μπορεί να χρησιμοποιηθεί για machine learning. Είναι ένα εργαλείο που βοηθά στην εύκολη κατασκευή μοντέλων μηχανικής μάθησης και ταξινόμησης αυτών.

Τέλος στο κεφάλαιο 7 παρουσιάστηκαν οι 2 βασικές γλώσσες προγραμματισμού που χρησιμοποιούνται στη μηχανική μάθηση, τα πλεονεκτήματα και μειονεκτήματά τους καθώς και εφαρμογές αυτών.

Μια μελλοντική συνέχεια της παρούσας διπλωματικής θα μπορούσε να είναι η εις βάθος ανάλυσης των εφαρμογών της μηχανικής μάθησης στην καθημερινότητα και πως ή εξόρυξη δεδομένων μπορεί να βελτιώσει την ίδια την καθημερινότητα όπως στο παράδειγμα των έξυπνων πόλεων.

## Βιβλιογραφία


1. Andrew Ng. «Linear Regression - LMS algorithm». CS229 Lecture notes. Standford University. Pg. 4–7.
2. Breiman, Leo (1996). "Bagging predictors". Machine Learning. 24 (2): 123–140. doi:10.1007/BF00058655.
3. Buzea, Pacheco, & Robbie, Nanomaterials and nanoparticles: Sources and toxicity, 2007, Department of Physics, Queen's University, Kingston, Ontario K7L 3N6, Canada doi: 10.1116/1.2815690
4. Chomenidis et al, Jaqpot Quattro: A novel computational web platform for modelling and analysis in nanoinformatics, 2017, doi: 10.1021/acs.jcim.7b00223
5. Cros A.F.A. (1863) Action de l'alcool amylique sur l'organisme, Thesis, University of Strasbourg, Strasbourg, France.
6. Dubitzky, Werner; Granzow, Martin; Berrar, Daniel (2007). Fundamentals of data mining in genomics and proteomics. Springer Science & Business Media. p. 178.
7. Efron, B. (1979) "Bootstrap methods: Another look at the jackknife", The Annals of Statistics7 (1): 1-26
8. Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials, Final report of the Nanocomput project, 2017
9. Ghosh Pallab, Introduction to Nanomaterials & Nanotechnology, Lecture 1, Department of Chemical Engineering IIT Guwahati, Guwahati–781039,India
10. Hervε Abdi, Partial Least Squares (PLS) Regression, The University of Texas at Dallas
11. Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pg. 278–282.
12. Hugo Kubinyi, Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis, 1988
13. John Dearden, The History and Development of Quantitative Structure-Activity Relationships (QSARs), 2016
14. Le Roux, Nicolas; Bengio, Yoshua; Fitzgibbon, Andrew (2012). «Improving First and Second-Order Methods by Modeling Uncertainty». Optimization for Machine Learning. MIT Press, pg. 404.
15. Marcu LG, Harriss-Phillips WM. In silico modelling of treatment-induced tumour cell kill: developments and advances. Comput Math Methods Med. 2012, doi: 10.1155/2012/960256
16. Meyer H. Zur Theorie der Alkoholnarkose Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung? Arch. Exp. Pathol. Pharmakol. 42, 109-18, 1899
17. MI Jordan, "Statistics and machine learning", 2014
18. Miramontes P. Un modelo de autómata celular para la evolución de los ácidos nucleicos [A cellular automaton model for the evolution of nucleic acids]. Tesis de doctorado en matemáticas. UNAM. 1992.

19. MoS2/TiO2 Heterostructures as Nonmetal Plasmonic Photocatalysts for Highly Efficient Hydrogen Evolution. Energy & Environmental Science
20. Nano, The Magazine for Small Science, What is Nanotechnology. A guide
21. Novel natural nanomaterial spins off from spider-mite genome sequencing. Phys.Org (May 23, 2013)
22. Overton E. (1901) Studien über die Narkose, Fischer, Jena, Germany,
23. Phil Simon (March 18, 2013). Too Big to Ignore: The Business Case for Big Data. Wiley, σελ. 89. ISBN 978-1-118-63817-0.
24. QSAR & QSPR, Alexandre Varnek, Faculté de Chimie, ULP, Strasbourg, FRANCE
25. Richet C. On the relationship between the toxicity and the physical properties of substances. Comptes Rendus Societe de Biologie 9, 775, 1893
26. S. Haykin. Νευρωνικά Δίκτυα και Μηχανική Μάθηση (3η Έκδοση)(Ε. Γκαγκάτσιου Μετάφραση), Παπασωτηρίου, 2010
27. Sarfaraz K. Niazi, Handbook of Preformulation, Chemical, Biological and Botanical Drugs, Second Edition, 2019, pg 105
28. Synthetic Nanomaterials Risk Assessment and Risk Management Basic report for the Swiss Action Plan, Environmental studies, Summary of the publication «Synthetische Nanomaterialien», Federal Office for the Environment FOEN and by the Federal Office of Public Health FOPH Bern, 2007, www.bafu.admin.ch/uw-0721-d
29. Todeschini, R., & Consonni Viviana. (2009). Molecular Descriptors for Chemoinformatics. (R.Mannhold, H. Kubinyi, & G. Folkers, Eds.) (2nd ed.). Wiley.
30. Δημόπουλος, Β., Τσαντίλη-Κακουλίδου, Ά. 2015. Βασικές αρχές σχεδιασμού και ανάπτυξης φαρμάκων. Κεφ. 3. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: https://repository.kallipos.gr/bitstream/11419/5884
31. Ελένη Βροντάκη et al, Ποσοτικές σχέσεις Δομής – Δράσης Τριών Διαστάσεων (3d -QSAr): Σύντομη Ανασκόπηση
32. Κ. Γεωργούλη, ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ, Μια εισαγωγική Προσέγγιση ΕλληνικάΑκαδημαϊκά Συγγράματα, 2015 http://repfiles.kallipos.gr/html_books/93/index.html
33. Καλαθάκης Χρήστος, Διπλωματική Εργασία ΔΙΑΓΝΩΣΤΙΚΗ ΑΕΡΙΟΣΤΡΟΒΙΛΩΝ ΜΕ ΧΡΗΣΗ ΜΗΧΑΝΩΝ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ (SUPPORT VECTOR MACHINES SVM), Σχολή Μηχανολόγων Μηχανικών ΕΜΠ, 2010
34. Κυπαρισσίδης, Καμμώνα, Χαϊτίδου, 2008, Εφαρμογές Νανοτεχνολογίας στην Ιατρική Μια Υπόσχεση για το Μέλλον, Intellectum | Τεύχος 04 / Μάιος 2008
35. Λούρου Σταυρούλα, Νανοτεχνολογία και εφαρμογές, Πτυχιακή Εργασία, Τμήμα Ηλεκτρονικής, ΤΕΙ Λαμίας, 2012
36. Μαγκανάρη Ειρήνη, ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΔΙΑΔΙΚΑΣΙΑ ΠΑΡΑΓΩΓΗΣ ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΒΙΟΜΗΧΑΝΙΑΣ ΠΑΡΑΓΩΓΗΣ ΤΣΙΜΕΝΤΟΥ, Διπλωματική Εργασία, Πανεπιστήμιο Πειραιά, 2006
37. Μακρίδης Αντώνιος, In-vitro αξιολόγηση μαγνητικών νανοσωματιδίων ως φορέων μαγνητικής υπερθερμίας», Διπλωματική Εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2013
38. Μπούτσικας Μιχαήλ. «Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)». Σημειώσεις μαθήματος "Στατιστικά Προγράμματα". Πανεπιστήμιο Πειραιώς
39. Παπαδόπουλος Ν. Αθανάσιος, Πρόβλεψη Τροχιών σε Δεδομένα Κίνησης με Βαθιά Νευρωνικά Δίκτυα, Διπλωματική Εργασία, Πανεπιστήμιο Πειραιώς, Δεκέμβριος 2018
40. Περρέα Δέσποινα, Εναλλακτικές Μέθοδοι, Ιατρική Σχολή Πανεπιστημίου Αθηνών.
41. Πετρίδης, Δ., 2015. Ανάλυση πολυμεταβλητών τεχνικών. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/2126

42. Τ. Σελλής, Τεχνητή Νοημοσύνη, Διάλεξη 10η, Νευρωνικά Δίκτυα - Μηχανική Μάθηση,2007, Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ Ε.Μ.Π.
43. Τσιάρα Αγγελική, Ταξινόμηση Εικόνων με Τυχαία Δάση, Μεταπτυχιακή Εργασία, Πανεπιστήμιο Ιωαννίνων, 2012
44. http://kelifos.physics.auth.gr/COURSES/neural/K8.pdf
45. http://www.business-analytics.gr/news/1211-machine-learning-vs-statistics
46. https://www.wikipedia.org/
47. https://dspace.lib.uom.gr/handle/2159/20262
48. https://core.ac.uk/download/pdf/323472768.pdf
49. https://dione.lib.unipi.gr/xmlui/handle/unipi/9467
50. https://nemertes.library.upatras.gr/server/api/core/bitstreams/cf58af52-bfc7-4fd8-bf31-3004dbc6b881/content
51. https://azure.microsoft.com/en-us/products/machine-learning
52. http://dmst.aueb.gr/gr/Courses/5sem/23_dioik_epix_texn/PPTS/Paper1A4.pdf
53. Imran? Ghaffar, Z.; Alshahrani, A.; Fayaz, M.; Alghamdi, AM; Gwak, J. A Topical Review on Machine Learning, Software Defined Networking, Internet of Things Applications: Research Limitations and Challenges. Electronics 2021, 10, 880.
54. Gyrard, A.; Zimmermann, A.; Sheth, A. Δημιουργία εφαρμογών που βασίζονται στο IoT για έξυπνες πόλεις: Πώς μπορούν να βοηθήσουν οι κατάλογοι οντολογίας; IEEE Internet Things J. 2018, 5, 3978–3990.
55. Kirimtat, A.; Krejcar, O.; Kertesz, A.; Tasgetiren, MF Future Trends and Current State of Smart City Concepts: A Survey. IEEE Πρόσβαση 2020, 8, 86448–86467.
56. Roblek, V.; Meš ko, M. Smart City Knowledge Management: Holistic Review and the Analysis of the Urban Knowledge Management. Στα Πρακτικά του 21ου Ετήσιου Διεθνούς Συνεδρίου για την Έρευνα Ψηφιακής Κυβέρνησης, Σεούλ, Κορέα, 15–19 Ιουνίου 2020. σελ. 52–60.
57. Tcholtchev, N.; Schieferdecker, I. Βιώσιμη και αξιόπιστη τεχνολογία πληροφοριών και επικοινωνιών για ανθεκτικές έξυπνες πόλεις. Έξυπνες πόλεις 2021, 4, 156–176.
58. Mohanty, SP; Choppali, U.; Κουγιανός, Ε. Όλα όσα θέλατε να μάθούμε για τις έξυπνες πόλεις: Το Διαδίκτυο των πραγμάτων είναι το σπονδυλική στήλη. Κατανάλωση IEEE. Ηλεκτρόνιο. Mag. 2016, 5, 60–70.
59. Alharbi, F.; Fei, Z. Βελτίωση της ποιότητας των υπηρεσιών για κρίσιμες ροές στο Έξυπνο Δίκτυο με χρήση δικτύωσης που καθορίζεται από λογισμικό.
60. Στα Πρακτικά του 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Σίδνεϊ, Αυστραλία, 6–9 Νοεμβρίου 2016. σελ. 237–242.
61. Naphade, M.; Banavar, G.; Harrison, C.; Paraszczak, J.; Morris, R. Οι έξυπνες πόλεις και οι προκλήσεις καινοτομίας τους. Υπολογιστής 2011, 44, 32–39.
62. Huang, N.; Liao, I.; Liu, H.; Wu, S.; Chou, C. Ένα δυναμικό σύστημα διαχείρισης QoS με πλατφόρμα ταξινόμησης ροής για δίκτυα καθορισμένα από λογισμικό. In Proceedings of the 2015 8th International Conference on Ubi-Media Computing (UMEDIA), Κολόμπο, Σρι Λάνκα, 24–26 Αυγούστου 2015; σελ. 72–77.
63. Binsahaq, A.; Sheltami, TR; Salah, K. A Survey on Autonomic Provisioning and Management of QoS in SDN Networks. IEEE Access 2019, 7, 73384–73435.
64. Braden, RT; Clark, DDD; Shenker, S. Integrated Services in the Internet Architecture: An Overview. Στο RFC 1633; IETF: Fremont, CA, USA, 1994.
65. Baker, F.; Μαύρο, DL; Nichols, K.; Blake, SL Ορισμός του πεδίου διαφοροποιημένων υπηρεσιών (Πεδίο DS) στα IPv4 και IPv6 Κεφαλίδες. Στο RFC 2474; IETF: Fremont, CA, USA, 1998.

66. AlZoman, R.; Alenazi, MJF που εκμεταλλεύεται το SDN για τη βελτίωση της QoS των δικτύων έξυπνων πόλεων ενάντια σε αποτυχίες συνδέσμων. Στα Πρακτικά του Έβδομου Διεθνούς Συνεδρίου 2020 για τα Καθορισμένα Συστήματα Λογισμικού (SDS), Παρίσι, Γαλλία, 20–23 Απριλίου 2020. σελ. 100–106.
67. MSc_Thesis_Nordin_Sahla.pdf
68. Tahaei, H.; Afifi, F.; Asemi, A.; Zaki, F.; Anuar, NB Η άνοδος της ταξινόμησης της κυκλοφορίας στα δίκτυα IoT: Μια έρευνα. J. Netw. Υπολογιστής. Appl. 2020, 154, 102538
69. Dainotti, A.; Pescape, A.; Claffy, KC Θέματα και μελλοντικές κατευθύνσεις στην ταξινόμηση της κυκλοφορίας. IEEE Netw. 2012, 26, 35–40.
70. Nguyen, TTT; Armitage, G. Μια έρευνα τεχνικών για την ταξινόμηση της κίνησης στο Διαδίκτυο με χρήση μηχανικής μάθησης. IEEE Commun. Surv. Παιδαγωγός. 2008, 10, 56–76. Pacheco, F.; Exposito, E.; Gineste, M.; Baudoin, C.; Aguilar, J. Towards the Deployment of Machine Learning Solutions in Network Ταξινόμηση Κυκλοφορίας: Συστηματική Έρευνα. IEEE Commun. Surv. Παιδαγωγός. 2019, 21, 1988–2014.
71. Park, B.; Win, Y.; Chung, J.; Kim, κα. Χονγκ, JWK Λεπτομερής ταξινόμηση κυκλοφορίας με βάση τον λειτουργικό διαχωρισμό. Int. J. Netw. Manag. 2013, 23, 350–381.
72. Aceto, G.; Dainotti, A.; de Donato, W.; Pescape, A. PortLoad: Λαμβάνοντας το καλύτερο από δύο κόσμους στην ταξινόμηση της κυκλοφορίας. In Proceedings of the 2010 INFOCOM IEEE Conference on Computer Communications Workshops, San Diego, CA, USA, 15–19 March 2010; σελ. 1–5.
73. Salman, O.; Elhajj, I.; Kayssi, A.; Chehab, A. A Review on Machine Learning Approaches for Internet Traffic Classification. Αννα. Τηλεπικοινωνία. 2020, 673–710.
74. Alqudah, N.; Yaseen, Q. Machine Learning for Traffic Analysis: A Review. Procedia Comput. Sci. 2020, 170, 911–916.
75. Xie, J.; Yu, FR; Huang, T.; Xie, R.; Liu, J.; Wang, C.; Liu, Y. A Survey of Machine Learning Techniques Applied to Software Καθορισμένη Δικτύωση (SDN): Ερευνητικά ζητήματα και προκλήσεις. IEEE Commun. Surv. Παιδαγωγός. 2019, 21, 393–430.
76. Zhongsheng, W.; Jianguo, W.; Sen, Y.; Jiaqiong, G. Αναγνώριση κυκλοφορίας και ανάλυση κυκλοφορίας με βάση τη μηχανή διανυσμάτων υποστήριξης. Συμφωνία. Υπολογιστής. Πρακτική. Exp. 2020, 32, e5292.
77. Al-Turjman, F. Πρόσβαση μέσου έξυπνης πόλης για εφαρμογές έξυπνης κινητικότητας στο Διαδίκτυο των πραγμάτων. Μεταφρ. Emerg. Τηλεπικοινωνία. Τεχνολ.2020, e3723.
78. Yao, H.; Gao, P.; Wang, J.; Zhang, P.; Jiang, C.; Han, Z. Capsule Network Assisted Traffic Classification Mechanism for Smart Cities. IEEE Internet Things J. 2019, 6,7515–7525.
79. Miao, Y.; Ruan, Z.; Pan, L.; Zhang, J.; Xiang, Y. Ολοκληρωμένη ανάλυση δεδομένων κίνησης δικτύου. Συμφωνία. Υπολογιστής. Πρακτική. Exp. 2018,30, e4181.
80. Perera, P.; Tian, YC; Fidge, C.; Kelly, W. A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. Στην Επεξεργασία Νευρωνικών Πληροφοριών; Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, ESM, Eds.; Springer International Publishing: Cham, Switzerland, 2017; σελ.445–454.
81. Rahman, A.; Jin, J.; Cricenti, A.; Rahman, A.; Yuan, D. A Cloud Robotics Framework of Optimal Task Offloading for Smart City Applications. Στα Πρακτικά του 2016IEEE Global Communications Conference (GLOBECOM), Ουάσιγκτον, DC, ΗΠΑ, 4–8 Δεκεμβρίου 2016. σελ. 1–7.
82. Moore, AW; Zuev, D. Ταξινόμηση κίνησης στο Διαδίκτυο με χρήση τεχνικών ανάλυσης Bayesian. SIGMETRICS Εκτελέστε. Eval. Αναθ. 2005,33, 50–60.
83. Zhang, C.; Wang, X.; Li, F.; He, Q.; Huang, M. Ταξινόμηση εφαρμογών δικτύου βασισμένη σε βαθιά μάθηση για SDN. Μεταφρ. Emerg.Τηλεπικοινωνία. Τεχνολ. 2018, 29, e3302.

84. Cao, J.; Fang, Z.; Qu, G.; Sun, H.; Zhang, D. Ένα ακριβές μοντέλο ταξινόμησης κυκλοφορίας που βασίζεται σε μηχανές διανυσμάτων υποστήριξης. Int. J.Netw. Manag. 2017, 27, e1962.
85. Yuan, R.; Li, Z.; Guan, X.; Xu, L. Μια μέθοδος μηχανικής εκμάθησης που βασίζεται σε SVM για ακριβή ταξινόμηση της κίνησης στο Διαδίκτυο. Inf. Συστ.Εμπρός. 2010, 12, 149–156.
86. Cotton, M.; Eggert, L.; Αγγίξτε, DJD; Westerlund, M.; Cheshire, S. Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Number Registry.
87. Abbasi, H., Ezzati-Jivan, N., Bellaiche, M., Talhi, C., & Dagenais, MR (2019). Τεχνική ανίχνευσης επιθέσεων edos που βασίζεται σε μηχανική μάθηση με χρήση ανάλυσης ίχνους εκτέλεσης.
88. Akande, A., Cabral, P., Gomes, P., & Casteleyn, S. (2019). Η κατάταξη της Λισαβόνας για τις έξυπνες βιώσιμες πόλεις στην Ευρώπη. Sustainable Cities and Society, 44, 475–487.
89. Aljawarneh, S., Aldwairi, M., & Yassein, MB (2018). Σύστημα ανίχνευσης εισβολής που βασίζεται σε ανωμαλίες μέσω ανάλυσης επιλογής χαρακτηριστικών και δημιουργίας υβριδικού αποδοτικού μοντέλου. Journal of Computational Science, 25, 152–160.
90. https://s3-us-west-2.amazonaws.com/gae-supplemental-media/id3-algorithm-for-decision-treespdf/ID3-Algorithm-for-Decision-Trees.pdf
91. http://ikee.lib.auth.gr/record/292940/files/%CE%94%CE%B9%CF%80%CE%BB%CF%89%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AE%20R%20-%20%CE%96.%CE%9B.pdf