ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Διπλωματική Εργασία στην Κατηγοριοποίηση Δεδομένων.
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ – ΠΜΣ 'ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ' 'ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ'

ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΦΟΙΤΗΤΗΣ : ΚΑΤΣΑΝΤΩΝΗΣ ΑΓΓΕΛΟΣ ΜΕ1608

ΟΚΤΩΒΡΙΟΣ 2020

ΥΠΕΥΘΥΝΟΣ ΚΑΘΗΓΗΤΗΣ : ΜΑΓΚΛΟΓΙΑΝΝΗΣ ΗΛΙΑΣ - ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ – ΠΜΣ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ –
ΠΕΙΡΑΙΑΣ - ΑΤΤΙΚΗ 2020

[1]

Ευχαριστώ τους γονείς μου και τα αδέρφια μου.

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία έχει ώς στόχο την παρουσίαση και την εφαρμογή αρχικών στατιστικών τεχνικών και αλγορίθμων μηχανικής μάθησης για την πρόβλεψη και κατηγοριοποίηση δεδομένων (Machine Learning and Classification).

Σε πρώτη φάση γίνεται εισαγωγή στην κατηγοριοποίηση και αναλυτικής δεδομένων και δή στην ανάλυση δεδομένων αθλητικού χαρακτήρα έτσι ώστε να γίνουν κατανοητά στον αναγνώστη ορολογίες και τεχνικές ανάλυσης που αφορούν δεδομένα αθλητικών διοργανώσεων.

Έπειτα, αναλύεται η σπουδαιότητα της Διερευνητικής Ανάλυσης Δεδομένων (Exploratory Data Analysis) διαμέσου της παρουσίασης δεδομένων διαμέσου της έρευνας προτύπων και τάσεων σε αυτά , με αποτέλεσμα τον καθορισμό και την παρουσίαση κατανοητής και χρήσιμης πληροφορίας προς τον τελικό χρήστη. Σε αυτό το κομμάτι θα χρησιμοποιηθεί η γλώσσα προγραμματισμού python και η βιβλιοθήκη pandas ώς εργαλείο ανάλυσης και η πρακτική δοκιμή θα γίνει σε ένα σύνολο δεδομένων που περιέχει συμβάντα τα οποία προέρχονται από το SFPD Crime Incident Reporting System και είναι διαθέσιμα στην ηλεκτρονική ιστοσελίδα Kaggle.com

Στην συνέχεια , παρουσιάζονται δεδομένα από το NBA (National Basketball Association) . Αφού τα δεδομένα εξεταστούν υπόκεινται σε μια αρχική ανάλυση έτσι ώστε να καθοριστεί εάν είανι αξιόπιστα για δοκιμή στους αλγορίθμους μηχανικής μάθησης.Τα δεδομένα αυτά υπόκεινται σε δοκιμή από τους εξής αλγορίθμους : Deep Learning, Logistic Regression, Generalized Linear Model, Gradient Boosted Trees, Random Forest, Naive Bayes, Decision Tree, Fast Large Margin, Support Vector Machine.

Ώς απώτερος στόχος της ανάλυσης είναι αν μια ομάδα θα περάσει στα play offs (εξαρτημένη μεταβλητή)βάση των χαρακτηριστικών κάθε ομάδας (ανεξάρτητες μεταβλητές)

Έπειτα αφού παρουσιαστούν όλοι οι δοκιμαζόμενοι αλγόριθμοι μηχανικής μάθησης , παρουσιάζονται τα μέτρα αποδοτικότητάς τους και αξιολογούνται βάση των εξής μετρικών : ακρίβειας , ευαισθησίας , περιοχή κάτω από την καμπύλη ROC κ.α. αφού έχει προηγηθεί τυχαία επιλογή ανεξάρτητων μεταβλητών , με επικύρωση αλγορίθμων το train-test set με αναλογία 70%-30%.

Έπειτα παρουσιάζονται αποτελέσματα με χρηση της μεθόδου Forward Feature Selection η οποία είναι μέθοδος επιλογής χαρακτηριστικών (ανεξάρτητων μεταβλητών) διαμέσου αλγορίθμων μηχανικής μάθησης έτσι ώστε να επιλέγονται τα χαρακτηριστικά τα οποία προσφέρουν το μέγιστο στην πρόβλεψη και στην κατηγοριοποίηση. Το σημαντικό πλεονέκτημα αυτής της μεθόδου είναι ότι από ένα αρχικό σύνολο ανεξάρτητων μεταβλητών διαμέσου αλγορίθμων μηχανικής μάθησης χρησιμοποιούνται μόνο τα χαρακτηριστικά αυτά που προσφέρουν ουσιαστικά στην πρόβλεψη και με αυτή την διαδιαδικασία μειώνονται η πολυπλοκότητα και ο τελικός επεξεργαστικός χρόνος και αυξάνεται η ακρίβεια και η ευαισθησία των αλγορίθμων.

Εισαγωγή : Ανάλυση δεδομένων

Στη σημερινή εποχή τα δεδομένα μπορεί να προέρχονται από πολλές πηγές όπως ψηφιακές συσκευές, κάμερες, κοινωνικά δίκτυα, προσομοιώσεις και άλλα. Τα δεδομένα μπορεί να έχουν διάφορες μορφές όπως δομημένοι πίνακες, γραφήματα, εικόνες, ήχος, βίντεο, χρονοσειρές, κείμενα και άλλα. Ο όρος Αναλυτική (Analytics) είναι ο απόγονος των Συστημάτων Υποστήριξης Αποφάσεων (Decision Support Systems, DSS) και της Επιχειρηματικής Ευφυίας (Business Intelligence, BI). Ένα σύστημα υποστήριξης αποφάσεων είναι μια εφαρμογή ηλεκτρονικού προγράμματος που αναλύει επιχειρηματικά δεδομένα και τα παρουσιάζει, έτσι ώστε οι χρήστες να μπορούν να λαμβάνουν επιχειρηματικές αποφάσεις ευκολότερα. Πρόκειται για μια «ενημερωτική εφαρμογή» (για να την διακρίνει από μια «επιχειρησιακή εφαρμογή» που συλλέγει τα δεδομένα κατά τη διάρκεια της κανονικής επιχειρηματικής λειτουργίας). Τα πρώτα DSS εμφανίστηκαν στη δεκαετία του 1970. Με την πάροδο του χρόνου, πρόσθετες αιτήσεις υποστήριξης αποφάσεων όπως εκτελεστικά συστήματα πληροφοριών, online αναλυτική επεξεργασία (OLAP) και άλλα έγιναν δημοφιλή. Στη δεκαετία του 1990, ο Howard Dresner, πρότεινε τον όρο επιχειρηματική ευφυΐα. Ένας τυπικός ορισμός είναι ότι "η BI είναι μια ευρεία κατηγορία εφαρμογών, τεχνολογιών και διαδικασιών για τη συλλογή, την αποθήκευση, την πρόσβαση και την ανάλυση δεδομένων ώστε να βοηθήσουν τους επιχειρηματικούς χρήστες να λαμβάνουν καλύτερες αποφάσεις. Η επιχειρηματική ευφυΐα είναι μια τεχνολογικά καθοδηγούμενη διαδικασία για την ανάλυση δεδομένων και την παρουσίαση ενεργητικών πληροφοριών που βοηθούν στελέχη, διαχειριστές και άλλους εταιρικούς τελικούς χρήστες να λαμβάνουν ενημερωμένες επιχειρηματικές αποφάσεις. Το BI περιλαμβάνει μια ευρεία ποικιλία εργαλείων, εφαρμογών και μεθοδολογιών που επιτρέπουν στους οργανισμούς να συλλέγουν δεδομένα από εσωτερικά συστήματα και εξωτερικές πηγές, να τα προετοιμάζει για ανάλυση, να αναπτύσσει και να εκτελεί ερωτήματα ενάντια σε αυτά τα δεδομένα και να δημιουργεί αναφορές, πίνακες ελέγχου και οπτικοποιήσεις δεδομένων για να κάνει τα αναλυτικά αποτελέσματα διαθέσιμο στους υπεύθυνους για τη λήψη αποφάσεων, καθώς και στους επιχειρησιακούς εργαζόμενους. Ο ορισμός του BI μπορεί να θεωρηθεί ως ομπρέλα για όλες τις εφαρμογές που υποστηρίζουν τη λήψη αποφάσεων. Σε επίπεδο επιχειρήσεων, τα δυνητικά οφέλη των συστημάτων BI περιλαμβάνουν την επιτάχυνση και τη βελτίωση της διαδικασίας λήψης αποφάσεων, τη βελτιστοποίηση των εσωτερικών επιχειρηματικών διαδικασιών, την αύξηση της λειτουργικής αποτελεσματικότητας, την προώθηση νέων εσόδων και την επίτευξη ανταγωνιστικού πλεονεκτήματος έναντι των επιχειρηματικών ανταγωνιστών. Τα συστήματα BI μπορούν επίσης να βοηθήσουν τις εταιρείες να εντοπίσουν τις τάσεις της αγοράς και να εντοπίσουν επιχειρηματικά προβλήματα που πρέπει να αντιμετωπιστούν.

Το BI εξελίχθηκε από τα DSS και η αναλυτική από τα DSS. Το BI μπορεί επίσης να θεωρηθεί ως «λήψη δεδομένων μέσα» (σε μια περιοχή δεδομένων ή αποθήκη) και «λήψη δεδομένων έξω» (ανάλυση των δεδομένων που είναι αποθηκευμένο). Μια δεύτερη ερμηνεία της αναλυτικής είναι ότι είναι το τμήμα «get data out» του BI. Η τρίτη ερμηνεία είναι ότι η αναλυτική είναι η χρήση αλγορίθμων π.χ., μάθησης μηχανών, νευρωνικών δικτύων, κ.α. για την ανάλυση δεδομένων [1].

Η Ανάλυση Δεδομένων (Data Analysis) είναι η διαδικασία εξέτασης συνόλων δεδομένων προκειμένου να εξαχθούν συμπεράσματα σχετικά με τις πληροφορίες που περιέχουν, όλο

και περισσότερο με τη βοήθεια εξειδικευμένων συστημάτων και λογισμικού. Οι τεχνολογίες και οι τεχνικές ανάλυσης δεδομένων χρησιμοποιούνται ευρέως σε οργανισμούς για να τις επιτρέψουν να λαμβάνουν πιο ενημερωμένες επιχειρηματικές αποφάσεις και από επιστήμονες και ερευνητές για να επαληθεύσουν ή να διαψεύσουν επιστημονικά μοντέλα, θεωρίες και υποθέσεις. Ως όρος, η ανάλυση δεδομένων αναφέρεται κατά κύριο λόγο σε μια ποικιλία εφαρμογών, από τη βασική επιχειρησιακή ευφυΐα, την online αναλυτική επεξεργασία, κ.α. Η ανάλυση δεδομένων σε μερικές περιπτώσεις ταυτίζεται με τα Data Mining και Knowledge Discovery in Databases (KDD). Αναφέρεται δε ως τη διαδικασία ανακάλυψης αξιοποιήσιμης πληροφορίας σε μεγάλες βάσεις δεδομένων χρησιμοποιώντας αλγορίθμους που ανακαλύπτουν κρυμμένα πρότυπα σε δεδομένα.

Η ανάλυση δεδομένων είναι ένα σχετικά νέο και γρήγορα εξελισσόμενο σύνολο εργαλείων τον επιχειρηματικό κόσμο και τα εργαλεία αυτά προσαρμόζονται όλο και περισσότερο στον κόσμο του αθλητισμού. Η ανάλυση δεδομένων περιλαμβάνει προηγμένα στατιστικά στοιχεία, διαχείριση δεδομένων, οπτικοποίηση δεδομένων και διάφορα άλλα. Οι τεχνικές ανάλυσης δεδομένων μπορούν να βοηθήσουν τις επιχειρήσεις να αυξήσουν τα έσοδα, να βελτιώσουν την επιχειρησιακή αποτελεσματικότητα, να βελτιστοποιήσουν τις εκστρατείες μάρκετινγκ και τις προσπάθειες εξυπηρέτησης πελατών, να ανταποκριθούν ταχύτερα στις τάσεις των αναδυόμενων αγορών και να αποκτήσουν ανταγωνιστικό πλεονέκτημα έναντι των ανταγωνιστών τους. Ανάλογα με τη συγκεκριμένη εφαρμογή, τα δεδομένα που αναλύονται μπορούν να αποτελούνται είτε από ιστορικές εγγραφές είτε από νέες πληροφορίες που έχουν υποβληθεί σε επεξεργασία για χρήσεις αναλυτικών δεδομένων σε πραγματικό χρόνο. Οι αναλύσεις δεδομένων μπορούν επίσης να διαχωριστούν σε ανάλυση ποσοτικών δεδομένων και ανάλυση ποιοτικών δεδομένων. Ο πρώτος περιλαμβάνει ανάλυση αριθμητικών δεδομένων με ποσοτικά προσδιορίσιμες μεταβλητές που μπορούν να συγκριθούν ή να μετρηθούν στατιστικά. Η ποιοτική προσέγγιση είναι πιο ερμηνευτική, εστιάζει στην κατανόηση του περιεχομένου μη-αριθμητικών δεδομένων όπως το κείμενο, οι εικόνες, ο ήχος και το βίντεο, συμπεριλαμβανομένων κοινών φράσεων, θεμάτων και απόψεων [2].

Η ανάλυση δεδομενών διακρίνεται σε προγνωστική (predictive), διερευνητική (exploratory) και κανονιστική (prescriptive). Οι προγνωστικές αναλύσεις υποδεικνύουν τι θα συμβεί στο μέλλον. Οι μέθοδοι που χρησιμοποιούνται κυρίως είναι η ανάλυση παλινδρόμησης (regression analysis), η μηχανική μάθηση (machine learning) και τα τεχνητά νευρωνικά δίκτυα (artificial neural networks). Η βασική εφαρμογή της προγνωστικής ανάλυσης δεδομένων είναι το μάρκετινγκ όπου ο στόχος είναι να κατανοηθούν κσλύτερα τα προφίλ των καταναλωτών, οι ανάγκες και οι προτιμήσεις τους. Η διερευνητική ανάλυση δεδομένων αποσκοπεί στην εύρεση των σχέσεων μεταξύ των δεδομένων. Η κανονιστική ανάλυση προτείνει τι ενέργεια πρέπει να γίνει για να ληφθεί μια απόφαση, όπως για παράδειγμα οι οδηγίες που δίνει ένα σύστημα Global Positioning System (GPS) σε ένα όχημα. Οι κανονιστικές αναλύσεις μπορούν να εντοπίσουν τις βέλτιστες λύσεις, βάσει ενός συνόλου διαθέσιμων πόρων [3].

Οι πιο προηγμένοι μέθοδοι ανάλυσης δεδομένων περιλαμβάνουν την εξόρυξη δεδομένων, η οποία περιλαμβάνει τη διαλογή σε μεγάλα σύνολα δεδομένων για τον εντοπισμό τάσεων, μοτίβων και σχέσεων, την προγνωστική ανάλυση, η οποία επιδιώκει να προβλέψει τη συμπεριφορά των πελατών, τις αποτυχίες του εξοπλισμού και άλλα μελλοντικά γεγονότα και τη μηχανική μάθηση, μια τεχνική τεχνητής νοημοσύνης που

χρησιμοποιεί αυτοματοποιημένους αλγόριθμους για να μετατρέψει τα σύνολα δεδομένων πιο γρήγορα από ό, τι οι επιστήμονες των δεδομένων μπορούν να κάνουν μέσω της συμβατικής αναλυτικής μοντελοποίησης.

Σε προηγμένες εφαρμογές ανάλυσης δεδομένων, μεγάλο μέρος της απαιτούμενης εργασίας λαμβάνει χώρα εκ των προτέρων, συλλέγοντας, ενσωματώνοντας και προετοιμάζοντας δεδομένα και στη συνέχεια αναπτύσσοντας, δοκιμάζοντας και αναθεωρώντας τα αναλυτικά μοντέλα για να εξασφαλιστεί ότι παράγουν ακριβή αποτελέσματα.

Η διαδικασία ανάλυσης ξεκινά με τη συλλογή δεδομένων, στην οποία εντοπίζονται οι πληροφορίες που χρειάζονται για μια συγκεκριμένη εφαρμογή ανάλυσης δεδομένων. Μόλις ολοκληρωθεί η συλλογή των απαιτούμενων δεδομένων, το επόμενο βήμα είναι ο εντοπισμός και η διόρθωση των προβλημάτων ποιότητας δεδομένων που θα μπορούσαν να επηρεάσουν την ακρίβεια των εφαρμογών της ανάλυσης όπως για παράδειγμα η εξάλειψη των σφαλμάτων και των διπλών εγγραφών.

Οι εφαρμογές ανάλυσης δεδομένων υποστηρίζουν μια μεγάλη ποικιλία επιχειρηματικών χρήσεων. Για παράδειγμα, οι τράπεζες και οι εταιρείες πιστωτικών καρτών αναλύουν τα πρότυπα δαπανών για την πρόληψη της απάτης και της κλοπής ταυτότητας. Οι εταιρείες ηλεκτρονικού εμπορίου και οι πάροχοι υπηρεσιών μάρκετινγκ κάνουν ανάλυση clickstream για να εντοπίσουν τους επισκέπτες του ιστότοπου που είναι πιο πιθανό να αγοράσουν ένα συγκεκριμένο προϊόν ή υπηρεσία με βάση τα πρότυπα πλοήγησης και προβολής σελίδας. Οι οργανισμοί υγειονομικής περίθαλψης δίνουν δεδομένα ασθενών για αξιολόγηση της αποτελεσματικότητας των θεραπειών για διάφορες ασθένειες [1].

Ανάλυση δεδομένων στον αθλητισμό

Η εφαρμογή της στατιστικής στον αθλητισμό είναι ένα πεδίο που παρέχει εξειδικευμένη μεθοδολογία για τη συλλογή και την ανάλυση δεδομένων αθλητισμού προκειμένου να ληφθούν αποφάσεις για επιτυχή σχεδιασμό και την εφαρμογή νέων στρατηγικών. Πριν από τον 21^{ο} αιώνα, η λήψη αποφάσεων στον αθλητισμό βασίστηκε κυρίως στις πληροφορίες που αποκτήθηκαν από την παρατήρηση. Αυτό έχει αλλάξει με τεχνολογικές εξελίξεις, που σχετίζονται κυρίως με την απόκτηση δεδομένων και τη διαθεσιμότητα προσωπικών υπολογιστών. Η πρακτική των αθλητικών αναλύσεων είναι εδώ και δεκαετίες, αλλά οι πρόσφατες εξελίξεις στη συλλογή δεδομένων και την τεχνολογία διαχείρισης έχουν διευρύνει σημαντικά το πεδίο εφαρμογής της. Η χρήση δεδομένων και στατιστικών έχει καταστεί παραγωγική σε όλα τα μεγάλα αθλήματα. Στην πραγματικότητα, μεγάλο μέρος των επαγγελματικών ομάδων στις Ηνωμένες Πολιτείες τώρα συνηθίζουν να χρησιμοποιούν τις υπηρεσίες των επαγγελματικών στατιστικολόγων για να υποστηρίξουν τις δραστηριότητές τους. Η παρακολούθηση του μέσου όρου του παίκτη του μπέιζμπολ ως βάση για τη μέτρηση του δυναμικού ή της ικανότητας είναι μόνο ένα από τα πολλά παραδείγματα εφαρμογής αναλυτικών στοιχείων στον αθλητισμό [4].

Οι αθλητικές αναλύσεις είναι μια συλλογή σχετικών ιστορικών στατιστικών στοιχείων τα οποία, όταν εφαρμόζονται σωστά, μπορούν να προσφέρουν ένα ανταγωνιστικό πλεονέκτημα σε μια ομάδα ή ένα άτομο. Μέσω της συλλογής και της ανάλυσης αυτών των στοιχείων, οι αθλητικές αναλύσεις ενημερώνουν τους παίκτες, τους προπονητές και το λοιπό προσωπικό για να διευκολύνουν τη λήψη αποφάσεων τόσο κατά τη διάρκεια όσο και πριν από αθλητικά γεγονότα. Ουσιαστικά, οι αθλητικές αναλύσεις είναι η πρακτική της

εφαρμογής μαθηματικών και στατιστικών αρχών στον αθλητισμό και τις σχετικές περιφερειακές δραστηριότητες. Παρόλο που υπάρχουν πολλοί παράγοντες και προτεραιότητες ειδικά για τη βιομηχανία, οι αθλητικοί αναλυτές χρησιμοποιούν τις ίδιες βασικές μεθόδους και προσεγγίσεις όπως κάθε άλλου είδους αναλυτής δεδομένων. Η καθιέρωση παραμέτρων για τη μέτρηση και η συνεχής συλλογή δεδομένων από ένα ευρύ δείγμα είναι η βάση της διαδικασίας ανάλυσης. Αυτά τα δεδομένα επιλύονται και βελτιστοποιούνται για να βελτιωθεί η ακρίβεια και η χρηστικότητα των αποτελεσμάτων.

Ο όρος Αθλητική Αναλυτική (Sport Analytics) περιλαμβάνει έννοιες από διάφορους τομείς όπως η Στατιστική, η Πληροφορική, η Διοίκηση και οι Επιστήμες Υγείας. Ως εκ τούτου, η αθλητική αναλυτική περιγράφεται ευρέως ως η διαδικασία διαχείρισης δεδομένων, η εφαρμογή μοντέλου πρόβλεψης και η χρήση συστημάτων πληροφοριών για τη λήψη αποφάσεων για να αποκτήσει ανταγωνιστικό πλεονέκτημα ένα αθλητικό σωματείο ή έναν αθλητή στο πεδίο του παιχνιδιού. Για παράδειγμα, οι αθλητικές ομάδες χρησιμοποιούν στατιστική ανάλυση για να αξιολογήσουν παίκτες για να καθορίσουν την καλύτερη στρατηγική παιχνιδιού. Οι αθλητικές ενώσεις αναπτύσσουν βαθμολογίες των παικτών και των ομάδων, για να αξιολογήσουν τους υπάρχοντες κανόνες και να μελετήσουν τη σκοπιμότητα της εισαγωγής νέων κανόνων. Μέσω της συλλογής και της ανάλυσης σχετικών ιστορικών στατιστικών στοιχείων, οι αθλητικές αναλύσεις ενημερώνουν τους παίκτες, τους προπονητές και το λοιπό προσωπικό για να διευκολύνουν τη λήψη αποφάσεων τόσο κατά τη διάρκεια όσο και πριν από αθλητικά γεγονότα [5].

Ο όρος Αθλητική Αναλυτική αναφέρεται στη διαχείριση δομημένων ιστορικών δεδομένων, στην εφαρμογή προγνωστικών αναλυτικών μοντέλων που χρησιμοποιούν αυτά τα δεδομένα και στη χρήση του Πληροφοριακού Συστήματος για την ενημέρωση των υπευθύνων λήψης αποφάσεων ώστε να βοηθήσουν τα σωματεία τους να αποκτήσουν ένα ανταγωνιστικό πλεονέκτημα. Οι τρεις βασικές συνιστώσες ενός προγράμματος αθλητικών αναλύσεων ειναι η διαχείριση δεδομένων, τα προγνωστικά μοντέλα και τα πληροφοριακά συσρτήματα. Η Αθλητική Αναλυτική αναφέρεται στους υπεύθυνους για τη λήψη αποφάσεων του σωματείου, π.χ. εκπρόσωποι του προσωπικού, προπονητές, εκπαιδευτές κτλ. Τα απαραίτητα δεδομένα συλλέγονται και στη συνέχεια μετασχηματίζονται σε αξιοποιήσιμη πληροφορία. Εισάγονται στο πληροφοριακό σύστημα και προκύπτουν πληροφορίες που υποστηρίζουν τη λήψη αποφάσεων.

Η χρήση της ανάλυσης δεδομένων στον αθλητισμό, προέρχεται από την εξέλιξη της υπολογιστικής ισχύος και τη διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων τόσο στις ομάδες όσο και στο κοινό. Η πρόσβαση στις πληροφορίες που οι ανταγωνιστές δεν έχουν οδηγεί στην παροχή πλεονεκτημάτων στις ομάδες, γεγονός που είναι ορατό και στις επιχειρήσεις. Ομάδες όπως οι Oakland A's, Tampa Bay Rays και San Antonio Spurs αποτελούν παραδείγματα της χρήσης ανάλυσης δεδομένων. Για παράδειγμα, οι Rays ήταν από τις πρώτες ομάδες που χρησιμοποίησαν δεδομένα από το Pitch F/X, που ανιχνεύουν την πορεία της μπάλας.

Πολλοί οργανισμοί κατανοούν το ρίσκο που σχετίζεται με τη μη επένδυση σε συστήματα ανάλυσης δεδομένων στην διαδικασία λήψης αποφάσεων. Τα ρίσκο στο να μην κατανοηθεί τόσο το πρόγραμμα αναλύσεων όσο και η ενσωμάτωσή του σε έναν οργανισμό, διασαφηνίστηκαν μέσω των αποτελεσμάτων της πρόσφατης έρευνας Sports Analytics Use Survey (SAUS). Είκοσι επτά άτομα που εκπροσωπούν ομάδες εμπλέκονται στα National Football League, Major League Baseball, the National Basketball Association και the English

Premier League απάντησαν σε ερωτήσεις σχετικά με τη χρήση των αθλητικών αναλύσεων στις ομάδες τους. Δύο ερωτηθέντες στην ίδια ομάδα (ένας που ανήκει στο προσωπικό και ο άλλος στον τομέα της πληροφορικής) κατέδειξαν δύο τελείως διαφορετικές προοπτικές όσον αφορά τη διαθεσιμότητα και την εφαρμογή αναλυτικών στοιχείων στο σωματείο τους. Ο εργαζόμενος στον τομέα της πληροφορικής γνώριζε για την προοπτική της ανάλυσης δεδομένων ενώ ο εργαζόμενος στο προσωπικό του σωματείου είχε σχετική άγνοια για τις δυνατότητες της ανάλυσης δεδομένων. Κάθε σωματείο χρησιμοποιεί την ανάλυση δεδομένων με διαφορετικό τρόπο. Αυτό εξαρτάται από το επίπεδο της χρηματικής επένδυσης και τις στρατηγικές του σωματείου σε μακροπρόθεσμο χρονικό ορίζοντα. Η ανάλυση δεδομένων μπορεί να εφαρμοστεί σε οποιοδήποτε άθλημα. Εντός του ιδίου αθλήματος, δύναται να υπάρχουν διαφορές. Για παράδειγμα, στην καλαθοσφαίριση μια σχολική ομάδα δεν έχει πρόσβαση στην ίδια ποσότητα δεδομέμων με μια ομάδα που κάνει πρωταθλητισμό.

Ένα μοντέλο ανάλυσης αθλητικών δεδομένων έχει δύο κύριους στόχους. Πρώτον, μπορεί να οδηγήσει σε μείωση του χρόνου λήψης αποφάσεων από την επιλογή με το αξιολογηθούν όλοι οι παίκτες και οι επιλογές μιας ομάδας. Το μοντέλο αποβλέπει στην εξαγωγή προτάσεων για λήψη αποφάσεων σε σύντομο χρονικό διάστημα από την απλή συγκέντρωση και αποθήκευση των δεδομένων. Ο δεύτερος στόχος είναι να παρέχει στους υπεύθυνους λήψης αποφάσεων μια πρόταση στο πως να διαχειριστούν την ομάδα και τους παίκτες. Η συνεχής ροή δεδομένων και η στατιστική επεξεργασία τους μπορεί να οδηγήσει σε ακριβέστερες εκτιμήσεις των προοπτικών ενός παίκτη σε επαγγελματικό επίπεδο.

Για παράδειγμα, στην περίπτωση μιας ομάδας καλαθοσφαίρησης, μπορεί να παρακολουθείται η απόδοσης ενός παίκτη κατά τη διάρκεια της αγωνιστικής περιόδου και να παρέχεται μια πρόβλεψη για ένα μελλοντικό παιχνίδι. Επίσης, ένας προπονητής που προετυοιμάζει την ομάδα του, μπορεί να λάβει πληροφόρηση για τις δυνατότητες και τις αδυναμίες τόσο της ομάδας του όσο και της αντίπαλης. Ο προπονητής μπορεί να αποφασίσει για τη σύνθεση της πεντάδας που θα χρησιμοποιήσει και να την εισάγει το μοντέλο λήψης αποφάσεων και αυτό με τη σειρά του να το αξιολογήσει. Υπό την απουσία του μοντέλου ανάλυσης και δεδομένων των στενών χρονικών περιθωρίων, ο προπονητής καλείται να αξιολογήσει από μόνος τους διαφορετικά σενάρια περί της σύνθεσης της ομάδας, γεγονός που κάτι τέτοιο θα είναι χρονοβόρο με κίνδυνο να μην ολοκληρωθεί.

Βαδικές κατηγορίες δεδομένων για ένα μοντέλο ανάλυσης, είναι τα οικονομικά στοιχεία της ομάδας, τα ιατρικά στοιχεία των παικτών και στοιχεία που σχετίζονται με την απόδοση των παικτών. Επίσης, τα δεδομένα μπορεί να περιλαμβάνουν βίντεο της ομάδας κατά τη διάρκεια ενός παιχνιδιού. Οι αθλητικοί ιατροί χρησιμοποιούν στατιστικές μεθόδους για να κατανοήσουν τη φυσική και τη ψυχολογική κατάσταση των αθλητών. Κάθε αθλητικό σωματείο αντιμετωπίζει τις δικές της προκλήσεις. Εισάγοντας και αναπτύσσοντας αναλυτικά στοιχεία ως μέρος της διαδικασίας λήψης αποφάσεων, αλλά η κατανόηση των συνιστωσών ενός προγράμματος ανάλυσης θα βοηθήσει τους διαχειριστές να μεγιστοποιήσουν το ανταγωνιστικό πλεονέκτημα τους.

Η ανάλυση αθλητικών δεδομένων διαθέτει πολλές εφαρμογές σε ένα αθλητικό περιβάλλον, συμπεριλαμβανομένης της διαχείρισης των ατομικών και ομαδικών επιδόσεων. Οι προπονητές μπορούν να χρησιμοποιήσουν τα δεδομένα για να βελτιστοποιήσουν τα προγράμματα άσκησης για τους παίκτες τους και να αναπτύξουν σχέδια διατροφής για να μεγιστοποιήσουν τη φυσική τους κατάσταση. Η ανάλυση

αθλητικών δεδομένων χρησιμοποιείται επίσης συχνά στην ανάπτυξη τακτικών και ομαδικών στρατηγικών. Με χιλιάδες παιχνίδια που αξίζουν τα δεδομένα για μελέτη, οι αναλυτές μπορούν να αναζητήσουν μοτίβα σε ένα ευρύ δείγμα μεγέθους όσον αφορά το σχηματισμό, τις στρατηγικές αντιμετώπισης και άλλες βασικές μεταβλητές.

Η πρακτική ανάλυση δεδομένων έχει πολλές εφαρμογές για την επιχειρηματική πλευρά του αθλητισμού. Δεδομένου ότι οι περισσότερες επαγγελματικές αθλητικές ομάδες λειτουργούν ως επιχειρήσεις, αναζητούν πάντα τρόπους βελτίωσης των πωλήσεων και μείωσης των εξόδων σε ολόκληρο τον οργανισμό τους. Ορισμένοι αθλητικοί αναλυτές επικεντρώνονται ειδικά σε θέματα σχετικά με την εμπορία και την πώληση αθλητικών εισιτηρίων και προϊόντων ομάδας.

Υπάρχουν δύο βασικές πτυχές της αθλητικής ανάλυσης - επί τόπου και εκτός πεδίου αναλύσεις. Οι αναλύσεις σε πεδίο ασχολούνται με τη βελτίωση της επιτόπιας απόδοσης των ομάδων και των παικτών. Πετάει βαθιά σε πτυχές όπως η τακτική του παιχνιδιού και η γυμναστική των παικτών. Οι αναλύσεις εκτός πεδίου ασχολούνται με την επιχειρηματική πλευρά του αθλητισμού. Τα αναλυτικά στοιχεία εκτός πεδίου επικεντρώνονται στη βοήθεια ενός αθλητικού οργανισμού ή μοντέλων επιφάνειας σώματος και πληροφοριών μέσω δεδομένων που θα βοηθήσουν στην αύξηση των πωλήσεων των εισιτηρίων και των εμπορευμάτων, στη βελτίωση της εμπλοκής ανεμιστήρων κλπ. Οι αναλύσεις εκτός πεδίου βασικά χρησιμοποιούν δεδομένα για να βοηθήσουν τους κατόχους δικαιωμάτων να λαμβάνουν αποφάσεις που θα οδηγούσαν σε υψηλότερες ανάπτυξη και αυξημένη κερδοφορία.

Καθώς η τεχνολογία έχει προχωρήσει κατά τη διάρκεια του τελευταίου αριθμού ετών, η συλλογή δεδομένων έχει γίνει πιο εμπεριστατωμένη και μπορεί να διεξαχθεί με σχετική ευκολία. Οι προόδους στη συλλογή δεδομένων επέτρεψαν την ανάπτυξη των αθλητικών αναλύσεων, γεγονός που οδήγησε στην ανάπτυξη προηγμένων στατιστικών καθώς και σε συγκεκριμένες αθλητικές τεχνολογίες που επιτρέπουν να διεξάγονται προσομοιώσεις παιχνιδιών από ομάδες πριν από το παιχνίδι, να βελτιώνονται οι στρατηγικές απόκτησης και εμπορίας ανεμιστήρων ακόμη και να κατανοήσουν τον αντίκτυπο της χορηγίας σε κάθε ομάδα καθώς και στους οπαδούς της.

Ένας άλλος σημαντικός αντίκτυπος των αθλητικών αναλύσεων που είχαν στον επαγγελματικό αθλητισμό είναι σε σχέση με τον αθλητικό τζόγο. Σε βάθος τα αθλητικά αναλύματα έχουν πάρει αθλητικά τυχερά παιχνίδια σε νέα επίπεδα, είτε πρόκειται για πρωταθλήματα αθλητικών φαντασμάτων είτε για νυχτερινά στοιχήματα, οι καλύτεροι έχουν τώρα περισσότερες πληροφορίες στη διάθεσή τους για να βοηθήσουν στη λήψη αποφάσεων. Ορισμένες εταιρείες και ιστοσελίδες έχουν αναπτυχθεί για να βοηθήσουν τους οπαδούς να ενημερωθούν για τις ανάγκες των στοιχημάτων τους

Ξεκινώντας από το 1977, ο Bill James αυτο-δημοσίευσε ένα ετήσιο βιβλίο με τίτλο «Baseball Abstract» που θεωρείται από πολλούς ως την αρχή των αθλητικών αναλύσεων. Αν και δεν χρησιμοποίησε ακόμη τα βασικά εργαλεία της Στατιστικής, όπως η ποσαρμογή μοντέλων ή οι γραφικές απεικονίσεις για το μεγαλύτερο μέρος της γραφής του, η εργασία αυτή είχε επιρροή baseball. Το κύριο χαρακτηριστικό του έργου του ήταν ότι οι απόψεις του βασίζονταν σε στοιχεία που περιέχονταν στα δεδομένα. Η ταινία του 20111 με τίτλο «Moneyball» βασίζεται στο βιβλίο του Michael Lewis με τίτλο «Moneyball: The Art of Winning an Unfair Game». Στο βιβλίο, ο Lewis διερευνά μερικές από τις καινοτόμες έννοιες του James και παρουσιάζει, όπως ο Billy Beane, ο γενικός διευθυντής μιας ομάδας Major

League Baseball (MLB), που καλείται Oakland Athletics, προσάρμοσε αυτές τις έννοιες σε πρακτική εφαρμογή. Ένα από τα πολλά τα πρόσφατα ακαδημαϊκά βιβλία που έχουν γραφτεί για τις αθλητικές αναλύσεις είναι το «Handbook of Statistical Methods and Analyses in Sports» που συν-εκδόθηκε από τον Jim Albert, τον Mark E. Glickman, τον Tim B. Swartz και τον Ruud H. Koning. Παρέχει επισκόπηση των στατιστικών μεθόδων στα μεγάλα αθλήματα και περιγράφει τις προκλήσεις και τα προβλήματα που αντιμετωπίζει η στατιστική έρευνα στον αθλητισμό. Ο κλάδος της Στατιστικής έχει αναγνωρίσει ότι ο αθλητισμός είναι πλούσιος σε δεδομένα και ότι ενδιαφέροντα στατιστικά προβλήματα προκύπτουν στον αθλητισμό. Δύο από τα πρώιμες εργασίες στην στατιστική είναι αυτά των Elderton (1945) και Wood (1945). Και οι δυο εργασίες αφορούν την κατανομή της τρέχουσας βαθμολογίας στο Test cricket. Ως άλλο πρώιμο παράδειγμα, ο Mosteller (1952) εξέτασε το πρόβλημα της εκτίμησης της πιθανότητας ότι η καλύτερη ομάδα κερδίζει το World Series στο MLB. Με την κατανόηση της ανάγκης να προωθηθεί η ανάπτυξη της στατιστικής στον αθλητισμό, ξεκίνησε η Αμερικανική Στατιστική Ένωση (American Statistical Association) που ξεκίνησε ένα ένα ξεχωριστό τμήμα για το «Statistics in Sports (SIS)» (SIS) το 1992. Προωθεί τις δημοσιεύσεις που επικεντρώνεται στην εφαρμογή της στατιστικής στον αθλητισμό. Ο τομέας SIS επίσης προωθεί συναντήσεις συναντήσεις που αφιερώνονται σε αθλητικές αναλύσεις, παρέχει καθοδήγηση σταδιοδρομίας, δεδομένα και online φόρουμ αθλητικών στατιστικών. Με την ταχεία εξέλιξη στον τομέα των αθλητικών αναλύσεων, έχουν δημιουργηθεί μερικά ερευνητικά περιοδικά που είναι αφιερωμένα στις στατιστικές στον αθλητισμό. Δύο γνωστά περιοδικά είναι το The Journal of Quantitative Analysis in Sports που αποτελεί έκδοση του American Statistical Association και το Journal of Sports Analytics. Το τελευταίο είναι ένα πιο πρόσφατο περιοδικό που επικεντρώνεται στις πρακτικές εφαρμογές που εξυπηρετούν τους ιδιοκτήτες ομάδων, γενικούς διευθυντές, προπονητές, οπαδούς και ακαδημαϊκούς. Υπάρχουν πολλά άλλα επιστημονικά περιοδικά τα οποία δεν επικεντρώνονται στην στατιστική όπως τα Journal of Sports Economics, Journal of Sports Sciences, American Journal of Sports Medicine, International Journal of Computer Science in Sport, Journal of Sports Science και Medicine and International Journal of Sports Science and Engineering. Εκτός από τα περιοδικά και τα βιβλία, οι αθλητικοί αναλυτές επίσης δημοσιεύουν το έργο τους σε ιστολόγια blog. Υπάρχουν καθημερινές αναλύσεις παιχνιδιών σε ιστότοπους (π.χ. www.soccermetrics.net, www.hockeyanalysis.com). Τα συνέδρια αθλητικών αναλύσεων αποτελούν επίσης μια πλατφόρμα για επαγγελματίες, ερευνητές και φοιτητές για να συζητήσουν σχετικά θέματα στον αθλητισμό. Το Συνέδριο MIT Sloan Sports Analytics είναι ένα από τα τα ιδιαίτερα αναγνωρισμένα ετήσια συνέδρια. Άλλα αθλητικά συνέδρια περιλαμβάνουν τα the New England Symposium on Statistics in Sport, MathSport International και the Australasian Conference on Mathematics and Computers in Sport. Ο συνδυασμός κοινοτήτων, φόρουμ στο διαδίκτυο και άλλων ειναι ευεργετικός τόσο στους εμπλεκόμενους με τα αθλήματα όσο και στους ερευνητές. Η Εταιρεία για την Έρευνα Αμερικανικού Μπέιζμπολ (Society for American Baseball Research, SABR) δημιουργήθηκε το 1971 με το στόχο να δοθεί στα μέλη της η δυνατότητα να δημοσιεύσουν τα ευρήματά τους στο περιοδικό «Baseball Research Journal» το οποίο περιλαμβάνει ιστορίες, βιογραφίες, στατιστικές, κριτικές βιβλίων και άλλες πτυχές του Μπέιζμπολ. Η SABR έχει περίπου 6000 μέλη, συμπεριλαμβανομένων ραδιοτηλεοπτικών φορέων, συγγραφέων, καθώς και πολλών πρώην παικτών. Στο Πανεπιστήμιο Simon Fraser (SFU), μια ομάδα από προπονητές και φοιτητές δημιούργησαν το Sports Analytic Group

(SAG) το 2015 με στόχο τη δημιουργία βασικών αναλυτικών στοιχείων ώστε να βοηθήσουν τους υπεύθυνους λήψης αποφάσεων στην αθλητική οργάνωση. Διοργανώνουν τακτικές συνομιλίες που δίδονται από τους προπονητές και τους επαγγελματίες για να κατανοήσουν τις πραγματικές καταστάσεις στον αθλητισμό και να διεξάγουν αποτελεσματική έρευνα (www.sfu.ca/sportsanalytics.html). Μέσα από την οργάνωση δύο συνεδρίων, του Vancouver Hockey Analytics Conference και του Ascadia Symposium on Statistics in Sports, η SAG κατάφερε να παρουσιάσει εμπειρογνώμονες στον τομέα των αθλητικών αναλύσεων αναδεικνύοντας τη δυνατότητα των αναλύσεων στον αθλητισμό. Η σύγχρονη τεχνολογία έχει διευκολύνει την οπτικοποίηση των στατιστικών πληροφοριών στο κοινό. Το σύστημα κάμερας SportVU χρησιμοποιείται στο National Basketball Association (NBA) και παρακολουθεί τις θέσεις σε πραγματικό χρόνο των παικτών και την μπάλα 25 φορές το δευτερόλεπτο. Στο Παγκόσμιο Κύπελλο της FIFA του 2014, ένα σύστημα υπό την ονομασία «Matrics», συγκέντρωνε δεδομένα για την παράδοση στατιστικών στοιχείων σε πραγματικό χρόνο.Κατά τη διάρκεια των αγώνων, οι σχολιαστές ήταν σε θέση να εξηγήσουν τις απόψεις τους με τη βοήθεια αυτών των οπτικών δεδομένων. Η ανάλυση αθλητικών δεδομένων βρίσκει επίσης εφαρμογή στην ανάλυση των τιμών των εισητηρίων και των αθλητικών στοιχημάτων.

Αναφορικά με μερικά παραδείγματα χρήσης ανάλυσης δεδομένων στον αθλητισμό, στις ΗΠΑ η εταιρεία Prozone sports data είναι από τους πρωτοπόρους στη χρήση δεδομένων παρακολούθησης παικτών για τεχνική και τακτική στατιστική ανάλυση. Μέχρι πρόσφατα, η απόδοση των παικτών αποτιμήθηκε κυρίως σε περιστατικά όπως τα περάσματα και οι ρίψεις για γκολ. Με την παρακολούθηση περαιτέρω δεδομένων, αναπτύσσονται νέα μέτρα για την αξιολόγηση των παικτών όπως η δημιουργία χώρου για τους συμπαίκτες του, η εφαρμογή αμυντικής πίεση και η μείωση των περασμάτων του αντιπάλου. Κάποιες από τις χρήσιμες μετρήσεις σε ανάλυση δεδομένων στην καλαθοσφαίρηση είναι απόπειρες ρίψεων, οι εύστοχες ρίψεις, οι ελεύθερες βολές, ο ρυθμός αμυντικής ανάκαμψης και άλλα. Σχετικά με το cricket, στην ιστοσελίδα www.espncricinfo.com υπάρχουν πληροφορίες από το 1770 έως σήμερα αναφορικά με παιχνίδια. Το ice hockey είναι ένα διάσημο άθλημα στις ΗΠΑ, Καναδά και Σκανδιναβικές χώρες. Ειδικα στις ΗΠΑ, στο National Hockey League (NHL). Ομοίως με την καλαθοσφαίριση, έχουν αναπτυχθεί κατά την πάροδο των ετών τυπικοί στατιστικοί δείκτες για παιχνίδια. Σημαντική βάση δεδομένων είναι το NHL Real Time Scoring System. Επίσης, έχουν εγκατασταθεί κάμερες σε πολλές αρένες του NHL με σκοπό την παρακολούθηση των παικτών και την περαιτέρω συγκέντρωση δεδομένων. Το baseball είναι από τα πιο λαοφιλή αθλήματα στις ΗΠΑ. Η διάθεση δεδομένων προς το κοινό έγινε τη δεκαετία του 1980 με την έκδοση του Baseball Abstract του Bill James, ο οποίος είναι εκ των ιδρυτών του Sabermetrics. Στη σημερινή εποχή, κάθε ομάδα του baseball έχει το δικό του προσωπικό για ανάλυση δεδομένων. Με την ενσωμάτωση καμερών στα γήπεδα και με τις τεχνολογίες PITCHf/x και FIELDf/x, το baseball έχει περάσει σε μια νέα περίοδο εξέλιξης της ανάλυσης δεδομένων. Για παράδειγμα, για κάθε βολή το σύστημα PITCHf/x παρέχει πληροφορίες για άνω των 70 μεταβλητών. Σημαντικό άθλημα στις ΗΠΑ αποτελεί το football. Η ιστοσελίδα του National Football League (NFL), www.nfl.com, είναι η βασική πηγή για δεδομένα. Επίσης, δεδομένα παρέχει και η www.advancedfootballanalytics.com. Κατηγοριοποιεί τα δεδομένα σε 4 κλάσεις: Ανάλυση ομάδων, ανάλυση παικτών, ανάλυση παιχνιδιών και πιθανότητες παιχνιδιών [4].

Διερευνητική ανάλυση δεδομένων

Η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis, EDA) είναι μια προσέγγιση για ανάλυση δεδομένων που χρησιμοποιεί μια ποικιλία τεχνικών (ως επί το πλείστον γραφικές) ώστε:

1. Να αποκαλυφτεί η υποκείμενη δομή ενός συνόλου δεδομένων.
2. Να εξαχθούν σημαντικές μεταβλητές.
3. Να ανιχνεύει τις ακραίες τιμές και τις μη τυπικές τιμές.
4. Να δοκιμασθούν υποκειμενικές παραδοχές.
5. Να αναπτυχθούν μοντέλα.

Η EDA δεν είναι όμοια με τα στατιστικά γραφικά, αν και οι δύο όροι χρησιμοποιούνται σχεδόν αδιακρίτως. Τα στατιστικά γραφικά είναι μια συλλογή τεχνικών που βασίζονται σε γραφικά και εστιάζοντας σε ένα χαρακτηριστικό δεδομένων. Η EDA είναι μια προσέγγιση στην ανάλυση δεδομένων που αναβάλλει τις συνηθισμένες υποθέσεις για το τι είδους μοντέλου ταιριάζει στα δεδομένα με την πιο άμεση προσέγγιση του επιτρέποντας στα ίδια τα δεδομένα να αποκαλύπτουν την υποκείμενη δομή και μοντέλο. Η EDA δεν είναι απλή συλλογή τεχνικών. Η EDA είναι μια προσέγγιση ως προς το πώς αναλύουμε ένα σύνολο δεδομένων. Οι περισσότερες τεχνικές της EDA είναι τύπου γραφικών. Συγκεκριμένα:

1. Απεικόνιση ακατέργαστων δεδομένων (π.χ. Data traces, histograms, bihistograms, probability plots, lag plots και block plots).
2. Απεικόνιση απλών στατιστικών όπως γραφήματα μέσης τιμής, τυπικής απόκλισης, box plots, διακύμανσης και άλλων.

Υπάρχουν και μη γραφικές τεχνικές που εστιάζουν κυρίως στον υπολογισμό στατιστικών τιμών. Επίσης, οι τεχνικές μπορούν να διακριθούν σε μονομεταβλητές (univariate) και πολυμεταβλητές (multivariate). Στις μονομεταβλητές τεχνικές, η έξοδος ενός μοντέλου εξαρτάται από μια είσοδο ενώ στις πολυμεταβλητές, η έξοδος εξαρτάται από πολλές εισόδους.

Ο κύριος στόχος της EDA είναι να μεγιστοποίησει την εποπτεία του αναλυτή σε ένα σύνολο δεδομένων με την εξέταση της δομής των δεδομένων και παρέχοντας τα εξής:

1. Μοντέλο που προκύπτει από τα δεδομένα
2. Εκτίμηση των παραμέτρων του μοντέλου και των αβεβαιοτήτων των παραμέτρων
3. Λίστα με ακραίες τιμές και άλλων μη τυπικών τιμών που ενυπάρχουν στα δεδομένα

Σε γενικά πλαίσια, η ανάλυση δεδομένων μπορεί να διακριθεί σε κλασσική (classical), διερευνητική (exploratory) και Bayesian. Όλες οι προσεγγίσεις αποβλέπουν στην επίλυση ενός επιστημονικού προβλήματος και εξάγουν κάποια συμπεράσματα. Η διαφορά έγκειται στα ενδιάμεσα στάδια. Για την κλασσική ανάλυση, τα στάδια είναι:

Πρόβλημα => Δεδομένα => Μοντέλο => Ανάλυση => Συμπεράσματα

Για τη EDA, τα στάδια είναι:

Πρόβλημα => Δεδομένα => Ανάλυση => Μοντέλο => Συμπεράσματα

Για την ανάλυση Bayesian, τα στάδια είναι:

Πρόβλημα => Δεδομένα => Μοντέλο => Εκ των Προτέρων Κατανομή > Ανάλυση => Συμπεράσματα

Έτσι για την κλασική ανάλυση μετά τη συλλογή δεδομένων ακολουθεί η διαμόρφωση ενός μοντέλου (ομαλότητα, γραμμικότητα, κ.λπ.) και η ανάλυση, η εκτίμηση και οι δοκιμές που ακολουθούν επικεντρώνονται τις παραμέτρους αυτού του μοντέλου. Για την EDA προηγείται η ανάλυση με στόχο να συμπεράνει ποιο μοντέλο θα ήταν κατάλληλο. Τέλος, για μια Bayesian ανάλυση, ο αναλυτής επιχειρεί να ενσωματώσει επιστημονική εμπειρογνωμοσύνη στην ανάλυση, επιβάλλοντας μια ανεξάρτητη κατανομή δεδομένων στις παραμέτρους των επιλεγμένων μοντέλο. Η ανάλυση αποτελείται συνεπώς από το συνδυασμό της εκ των προτέρων κατανομών των παραμέτρων και των δεδομένων ώστε να γίνουν υποθέσεις για τις παρταμέτρους του μοντέλου.

Η κλασσική ανάλυση αντιστοιχεί μοντέλα (ντερτεμινιστικά και πιθανοκρατικά) στα δεδομένα. Για παράδειγμα, ντερτεμινιστικά μοντέλα είναι τα παλινδρομικά μοντέλα (regression models) και τα Analysis of Variance (ANOVA) μοντέλα. Τα πιο κοινά πιθανοκρατικά μοντέλα υποθέτουν ότι τα σφάλματα ενός ντερτεμινιστικού μοντέλου ακολουθούν κανονική κατανομή. Η EDA δεν αντιστοιχεί μοντέλα στα δεδομένα. Η έμφαση είναι στα ίδια τα δεδομέμα. Μετά την εξέταση της δομής τους, επιλέγεται μοντέλο. Συνεπώς, στην κλασσική ανάλυση το επίκεντρο του ενδιαφέροντος είναι ο τύπος του μοντέλου και οι παράμετροι του ενώ στην EDA, είναι τα δεδομένα (δομή, ακραίες τιμές, ασυνέχειες, κτλ.). Μια άλλη σημαντική διαφορά μεταξύ της κλασσικής ανάλυσης και της EDA, είναι ότι στην κλασσική ανάλυση δεν λαμβάνονται υπόψη όλα τα δεδομένα παρα μόνο κάποιοι εκτιμητές. Τα δεδομένα χαρτογραφούνται σε ένα νέο χώρο και αντικαθίστονται με κάποιους δείκτες. Συνεπώς, προκύπτει μια μείωση των εν χρήση δεδομένων. Το πλεονέκτημα αυτής της προσέγγισης είναι ότι χρησιμοποιούνται μόνο ορισμένα χαρακτηριστικά όπως για παράδειγμα η τοποθεσία ενός δεδομένου στο χώρο διαστάσεων. Μειονέκτημα αποτελεί ότι δεν λαμβάνονται υπόψη άλλοι στατιστικοί δείκτες όπως η αυτοσυσχέστιση, η κύρτωση και άλλοι. Στην EDA όλα τα διαθέσιμα δεδομένα λαμβάνονται υπόψη κατά τη διάρκεια της ανάλυσης [6]-[7].

Η EDA έχει βρει εφαρμογή σε πλήθος προβλημάτων. Στην αναφορά [8], η EDA χρησιμοποιείται για την ανάλυση της κατανάλωσης της ηλεκτρικής ενέργειας του Ελληνικού ενεργειακού συστήματος για την περίοδο 2002-2006. Εξετάζεται με τη χρήση στατιστικών δεικτών και γραφικών η σχέση της κατανάλωσης με άλλες μεταβλητές όπως το εθνικό ακαθάριστο προϊόν και η τιμή της χονδρεμπορικής αγοράς ενέργειας. Τα αποτελέσματα της ανάλυσης μπορούν να βρουν εφαρμογή σε περιπτώσεις μελέτης της χρονολογικής εξέλιξης της κατανάλωσης και σε μελέτες πρόβλεψης ηλεκτρικού φορτίου. Στην αναφορά [9], εξετάζεται η εφαρμογή της EDA στην ακύρωση κράτησης θυρίδας σε διηπειρωτικές τακτικές θαλάσσιες μεταφορές και πιο συγκεκριμένα, για τις μεταφορές μεταξύ της Ασίας και της δυτικής ακτής των ΗΠΑ. Σε μια διηπειρωτική υπηρεσία μεταφοράς

εμπορευματοκιβωτίων, οι διαχειριστές των εμπορευματοκιβωτίων παραλαμβάνουν τις κρατήσεις των πελατών τους για το φορτίο τους λίγες εβδομάδες πριν το πλοίο αναχωρήσει από ένα συγκεκριμένο λιμάνι. Στην πράξη, ορισμένες από αυτές τις κρατήσεις ακυρώνονται τελικά χωρίς φόρτωση εμπορευματοκιβωτίων στα πλοία, γεγονός που οδηγεί σε χαμηλό συντελεστή φόρτωσης και απώλεια εσόδων. Στην εργασία εξετάζονται δείκτες όπως τα ποσοστά ακύρωσης του ταξιδιού, τα χαρακτηριστικά των κρατήσεων και οι παράγοντες που μπορεί να επηρεάσουν τις συμπεριφορές ακύρωσης. Στην αναφορά [10], η EDA εφαρμόζεται ως μέρος της διαδικασίας συσταδοποίησης κειμένων. Οι συστάδες με όμοια κείμενα συνδυάζονται για να δημιουργήσουν ένα σύνολο συστάδων που θα αναφέρεται σε ένα συγκεκριμένο θέμα. Η ανάλυση της εργασίας περιλαμβάνει 330000 άρθρα εφημερίδων που σχετίζονται με την πολιτική των New York Times από την 1η Ιανουαρίου του 1900 έως την 31η Δεκεμβρίου του 2015.

Η πρακτική δοκιμή θα γίνει σε ένα σύνολο δεδομένων που περιέχει συμβάντα τα οποία προέρχονται από το SFPD Crime Incident Reporting System , δηλαδή από το σύστημα αναφοράς εγκληματικών περιστατικών του αστυνομικού τμήματος του Σαν Φρανσίσκο στις Ηνωμένες Πολιτείες Αμερικής. Αυτό το σύνολο δεδομένων βρίσκεται ελεύθερα προσβάσιμο στο site **Kaggle .com.**

Περιγραφή πεδίων δεδομένων

Dates – Χρονική στιγμή που έγινε το συμβάν

Category – κατηγορία του εγκληματικού συμβάντος

Descript – λεπτομερείς περιγραφή του συμβάντος

DayOfWeek – ημέρα της εβδομάδας

PdDistrict – όνομα του αστυνομικού τμήματος

Resolution – πως επιλύθηκε το συμβάν

Address – η κατά προσέγγιση διεύθυνση που σημειώθηκε το συμβάν

X – Γεωγραφικό μήκος

Y – Γεωγραφικό πλάτος

Έξοδος Python.Pandas.DataFrame

RangeIndex: 2500 entries, 0 to 2499

Data columns (total 9 columns):

Dates 2500 non-null datetime64[ns]

Category 2500 non-null object

Descript 2500 non-null object

DayOfWeek 2500 non-null object

PdDistrict 2500 non-null object

Resolution 2500 non-null object

Address 2500 non-null object

X 2500 non-null float64

Y 2500 non-null float64

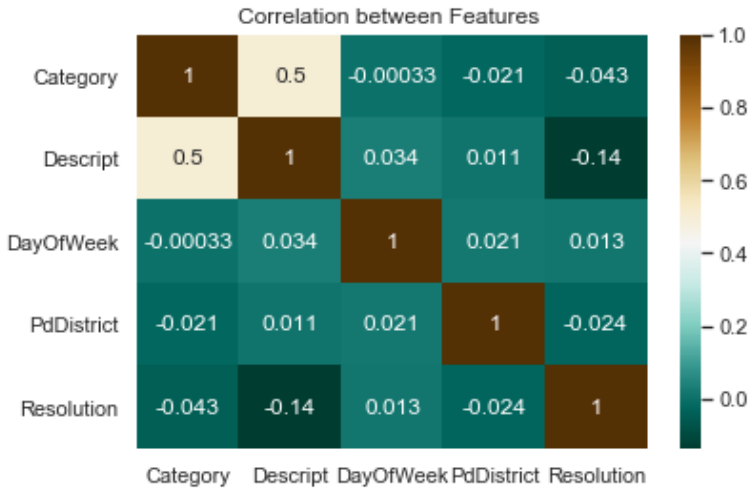
Dtypes: datetime64 [ns](1), float64(2), object(6)

Για την διερευνητική αναλυτική δεδομένων θα χρειαστεί ένα υποσύνολο των δεδομένων οπότε για την ανάλυση θα αφαιρεθούν τα εξής χαρακτηριστικά : X , Y .

Έπειτα αφού ομαδοποιηθούν τα δεδομένα , μέσω της γλώσσας προγραμματισμού Python και των βιβλιοθηκών αυτής Pandas και Numpy μπορούν ,χωρίς την ανάπτυξη μοντέλου ,να διεξαχθούν χρήσιμές πληροφορίες για τον τύπο εγκλημάτων.

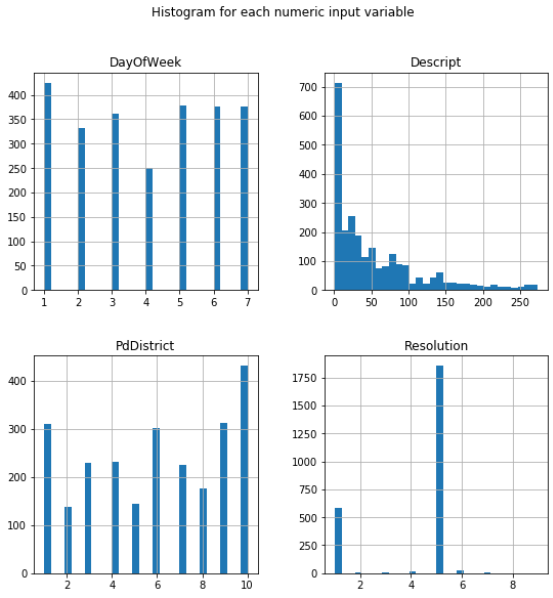
Διαδικασία επεξεργασίας δεδομένων – Αρχική ανάλυση και προ-επεξεργασία

Το σύνολο δεδομένων δεν έχει ελλείπουσες τιμές .Επίσης , με την βοήθεια της βιβλιοθήκης Pandas παρέχεται η δυνατότητα να παρθεί η ημέρα από την χρονοσφραγίδα των παρατηρήσεων . Επιπλέον , για κάθε χαρακτηριστικό από το σύνολο δεδομένων μπορεί να παρουσιαστεί ένα heatmap για την μεταξύ τους συνδιακύμανση , χρησιμοποιώντας την βιβλιοθήκη Seaborn για γραφήματα.



Γράφημα 1.Heatmap των συνδιακυμάνσεων των χαρακτηριστικών του συνόλου δεδομένων.

Για κάθε κατηγορία PdDistrict , DayofWeek και Resolution κατασκευάζεται ένα σύνολο ζευγαριού key-value (dictionary) , το οποίο αντιστοιχεί τις κατηγορίες με έναν αντίστοιχο αριθμό. Αυτός ο μετασχηματισμός βοηθάει στην περεταίρω κατανόηση των δεδομένων και βοηθάει στην καλύτερη εισαγωγή των δεδομένων σε αλγορίθμους μηχανικής μάθησης. Ένα ιστόγραμμα (Γράφημα 2) για κάθε κατηγορία χαρακτηριστικών με βάση τα συνολικά συμβάντα θα βοηθήσει για την περαιτέρω κατανόηση δεδομένων .

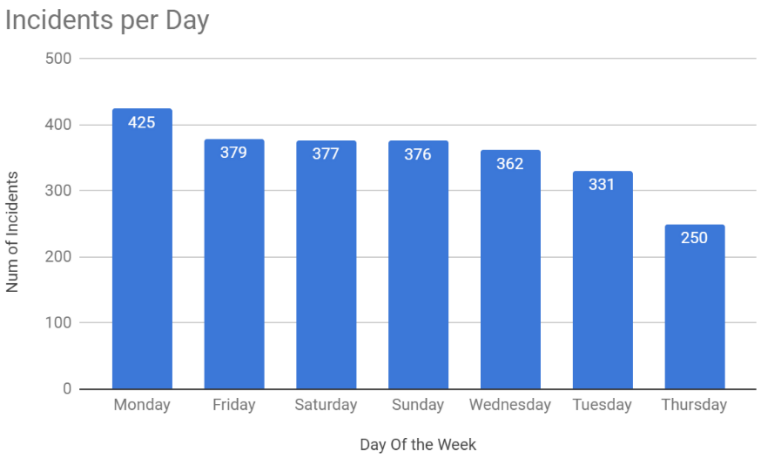


Γράφημα 2 - Ιστογράμματα και κατανομές δεδομένων

Συμπεράσματα Exploratory Analysis στα δεδομένα εγκλήματος.

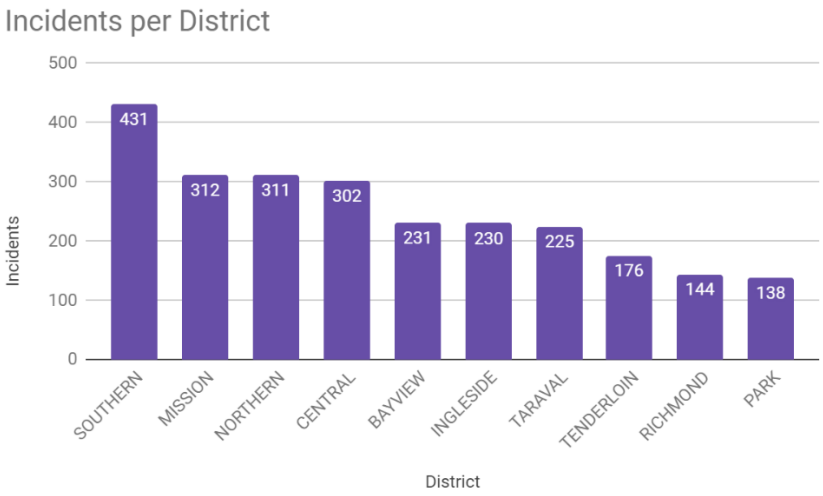
Η διερευνητική ανάλυση παρουσιάζει χρήσιμες πληροφορίες , οι οποίες μπορούν να αξιοποιηθούν κατάλληλα από τους αρμόδιους φορείς και να βελτιωθούν με αυτό τον τρόπο οι υφιστάμενες διαδικασίες αντιμετώπισης εγκλημάτων.

Στο Γράφημα 3 παρατηρείται ότι τα περισσότερα συμβάντα γίνονται την ημέρα Δευτέρα και έπειτα Παρασκευή με Κυριακή.



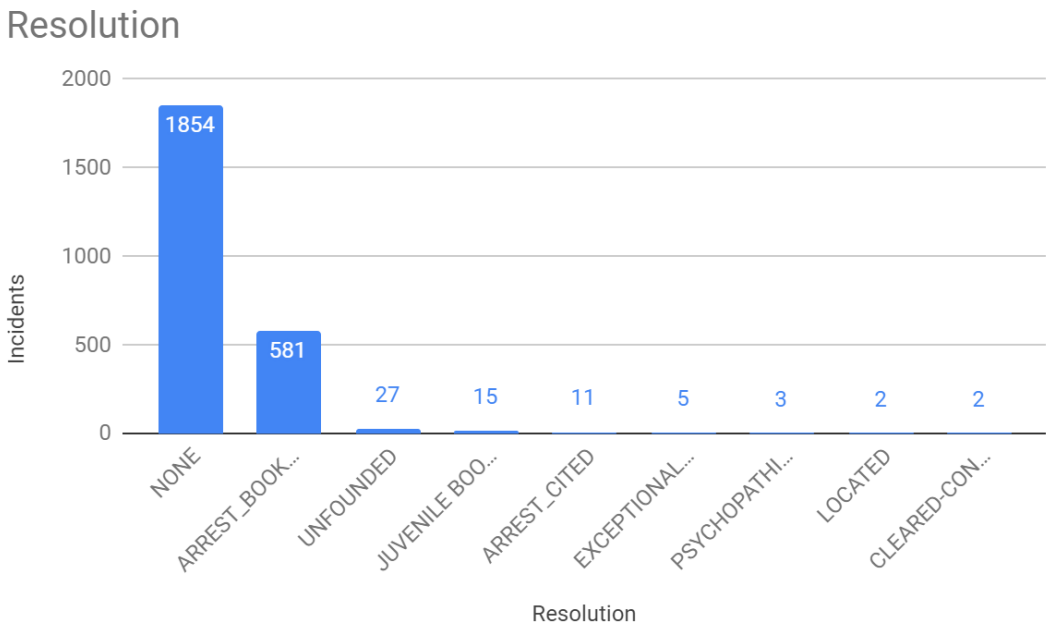
Γράφημα 3. Συνολικά συμβάντα ανά ημέρα

Στο Γράφημα 4 παρουσιάζονται οι περιοχές ευθύνης του αστυνομικού τμήματος με τα συνολικά συμβάντα ανά περιοχή. Βλέπουμε ότι το νότιο τμήμα έχει την περισσότερη εγκληματικότητα από όλες τις περιοχές.



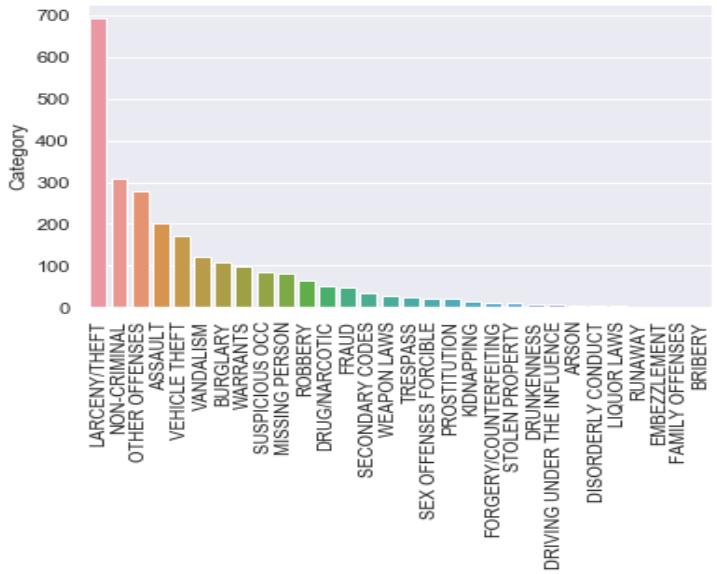
Γράφημα 4 - Συνολικά Συμβάντα ανα περιοχή ευθύνης του αστυνομικού τμήματος

Στο Γράφημα 5 παρουσιάζονται συνολικά οι καταλήξεις που έχουν τα συμβάντα. Παρατηρείται ότι τα περισσότερα συμβάντα μένουν άλυτα χωρίς να διαλευκανθούν ενώ αμέσως επόμενη κατάληξη είναι η προσαγωγή ή η φυλάκιση του δράστη. Οι επόμενες κατηγορίες κατάληξης είναι πολύ χαμηλές.



Γράφημα 5 - Σύνολα κατάληξης συμβάντων

Έπειτα , στο Γράφημα 6 βλέπουμε ένα γράφημα για τα σύνολα των συμβάντων ανά τύπο εγκλήματος. Η κατηγορία με τα περισσότερα συμβάντα είναι η κλοπή ενώ η κατηγορία με τις λιγότερες παρατηρήσεις είναι η δωροδοκία. Έτσι βγαίνει το συμπέρασμα ότι οι πιο κοινές κατηγορίες συμβάντων είναι οι: ['LARCENY/THEFT' (Κλοπή), 'NON-CRIMINAL' (όχι εγκληματική κατάληξη), 'OTHER OFFENSES' (άλλες παραβάσεις), 'ASSAULT' (επιθέσεις), 'VEHICLE THEFT' (κλοπή οχήματος)] οι οποίες αποτελούν το 82.56% των συμβάντων.

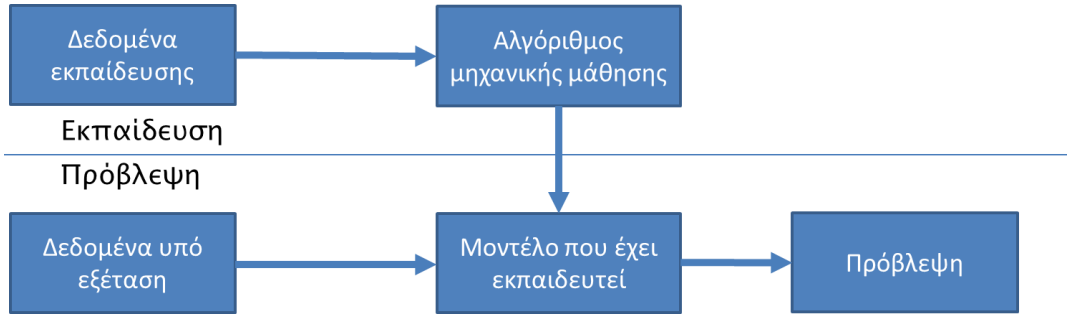


Γράφημα 6 - Σύνολα συμβάντων ανα τύπο συμβάντων

Αναγνώριση προτύπων

Η Αναγνώριση Προτύπων (Pattern Recognition) είναι η επιστημονική περιοχή της οποίας στόχος είναι η ταξινόμηση του αντικείμενα σε διάφορες κατηγορίες ή κλάσεις. Ανάλογα με την εφαρμογή, αυτά τα αντικείμενα μπορεί να είναι εικόνες ή κυματομορφές σήματος ή οποιοδήποτε είδος μετρήσεων που χρειάζονται να ταξινομηθούν. Η αναγνώριση προτύπων έχει μακρά ιστορία, αλλά πριν από τη δεκαετία του 1960 ήταν κυρίως η παραγωγή της θεωρητικής έρευνας στον τομέα των στατιστικών. Όπως και σε άλλες επιστημονικές περιοχές, η εξέλιξη των υπολογιστών αύξησε τη ζήτηση για πρακτικές εφαρμογές αναγνώρισης προτύπων, οι οποίες με τη σειρά τους δημιούργησαν νέες απαιτήσεις για περαιτέρω θεωρητικές εξελίξεις. Βασικές εφαρμογές της αναγνώρισης προτύπων είναι η όραση μηχανής, η αναγνώριση χαρακτήρων, η αναγνώριση ομιλίας, η ιατρική απεικόνιση και η εξόρυξη δεδομένων. Ένα σύστημα όρασης μηχανής καταγράφει τις εικόνες μέσω κάμερας και τις αναλύει για να παράγει περιγραφές του τι απεικονίζεται. Μια τυπική εφαρμογή ενός συστήματος μηχανικής όρασης είναι στον τομέα της μεταποιητικής βιομηχανίας, είτε για αυτοματοποιημένη οπτική επιθεώρηση είτε για αυτοματοποίηση στη γραμμή συναρμολόγησης. Με την ανάλυση των εικόνων μπορεί να γίνει ανίχνευση ελλατωματικών προϊόντων. Η αναγνώριση χαρακτήρων είναι μια άλλη σημαντική περιοχή για την αναγνώριση προτύπων. Η οπτική αναγνώριση χαρακτήρων βρίσκει εφαρμογή σε συστήματα σάρωσης τόσο σε εφαρμογές γραφείου όσο και σε μεγάλες εταιρίες [11].

Μέρος της αναγνώρισης προτύπων είναι η Μηχανική Μάθηση (Machine Learning). Η ΜΜ αναφέρεται σε ένα κλάδο της Τεχνητής Νοημοσύνης που σχετίζεται με το σχεδιασμό και ανάπτυξη αλγορίθμων που επιτρέπουν τους υπολογιστές να συμπεριφέρονται βάσει εμπειρικών δεδομένων. Σχετίζεται με την ικανότητα ορισμένων συστημάτων να μαθαίνουν. Ένα σύστημα ΜΜ χρησιμοποιεί διάφορες τεχνικές ώστε να λαμβάνει μια απόφαση βάσει των υπαρχόντων δεδομένων. Ένας αλγόριθμος ΜΜ είναι ένα πρόγραμμα που δείχνει το σύστημα ή τη μηχανή πώς να μαθαίνει ή εκπαιδεύεται. Στο Σχήμα 1 παρουσιάζεται το γενικό διάγραμμα ενός συστήματος ΜΜ.



Σχήμα 1 - Αναπαράσταση λειτουργίας ενός συστήματος ΜΜ

Στον προγραμματισμό των υπολογιστών, η πολυπλοκότητα συνήθως υπάρχει στον κώδικα. Δηλαδή για την επίλυση ενός σχετικά πολύπλοκου προβλήματος απαιτείται αντίστοιχα πολύπλοκος κώδικας σε κάποια γλώσσα προγραμματισμού που να δίνει εντολές στον υπολογιστή για την υλοποίηση του αλγορίθμου επίλυσης του προβλήματος. Στη ΜΜ, οι αλγόριθμοι (προγράμματα) είναι σχετικά απλοί. Η πολυπλοκότητα υπάρχει στα δεδομένα. Ο πυρήνας της ΜΜ είναι η αυτοματοποιημένη εκμάθηση της δομής των δεδομένων. Συνήθως τα δεδομένα έχουν μικρό οικονομικό κόστος και βρίσκονται σε αφθονία. Αντιθέτως, η εξειδικευμένη γνώση που απαιτείται για την επίλυση ενός προβλήματος έχει μεγάλο οικονομικό κόστος και συνήθως είναι σπάνια.

Η εκπαίδευση (μάθηση) ενός συστήματος ΜΜ γίνεται είτε βάσει βελτιστοποίησης ενός κριτηρίου κόστους (αντικειμενική συνάρτηση) είτε βάσει προηγούμενης εμπειρίας. Η μάθηση μπορεί να οριστεί επίσης ως η διαδικασία σύμφωνα με την οποία ένα σύστημα βελτιώνει την απόδοση του μέσω της εμπειρίας. Η ΜΜ χρησιμοποιείται όταν:

i) Δεν υπάρχει προηγούμενη ανθρώπινη γνώση (π.χ. πλοήγηση στον πλανήτη Άρη)
ii) Οι άνθρωποι δεν μπορούν να περιγράψουν την εξειδικευμένη γνώση τους (expertise) (π.χ. αναγνώριση ομιλίας)
iii) Η λύση ενός προβλήματος μεταβάλλεται στο χρόνο (π.χ. δρομολόγηση σε υπολογιστές)
iv) Οι λύσεις δεν είναι γενικές αλλά στοχεύουν αποκλειστικά σε μια περίπτωση (π.χ. βιομετρική)

Μερικές εφαρμογές της ΜΜ είναι οι εξής:

i) Αναγνώριση προτύπων (π.χ. φωνή, πρόσωπα, δεδομένων από αισθητήρας, ιατρικών εικόνων, κτλ.)
ii) Προβλήματα βελτιστοποίησης (π.χ. μη γραμμικά, κτλ.)
iii) Ανίχνευση μη ομαλότητας (π.χ. στη χρήση πιστωτικής κάρτας, σε δεδομένα λειτουργίας πυρηνικών σταθμών, κτλ.)
iv) Πρόβλεψη (π.χ. τιμών σε χρηματιστήριο, αιολικής ταχύτητας, κτλ.)

Η ΜΜ διακρίνεται σε τρεις κατηγορίες:

1) Επιβλεπόμενη Μάθηση (Supervised Learning): Μάθηση μέσω ενός συνόλου εκπαίδευσης που έχει ετικέτες (labels). Παράδειγμα: Ανίχνευση spam emails μέσω

εκπαίδευσης από ενός συνόλου emails που έχουν εκ των προτέρων χαρακτηριστεί με ετικέτες/επιγραφές.

2) Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Ανακάλυψη προτύπων σε δεδομένα χωρίς ετικέτες. Παράδειγμα: Συσταδοποίηση όμοιων εγγράφων βάσει του περιεχομένου του κειμένου.
3) Ενισχυτική Μάθηση (Reinforcement Learning): Μάθηση μέσω ανατροφοδότησης (ποινή ή βράβευση). Παράδειγμα: Μάθηση για παίξιμο σκάκι (ήττες και νίκες).

Στην επιβλεπόμενη μάθηση υπάρχει ένας δάσκαλος/κριτής που παρέχει τις επιθυμητές εξόδους ενός προβλήματος ή εκβάσεις ενός πειράματος. Δηλαδή, είναι εκ των προτέρων γνωστή η επιθυμητή έξοδος σε ένα πρόβλημα. Η επιβλεπόμενη μάθηση βρίσκει εφαρμογή σε προβλήματα κατηγοριοποίησης (classification) ή παλινδρόμησης (regression). Μερικοί αλγόριθμοι επιβλεπόμενης μάθησης είναι οι εξής: Naïve Bayes, Gaussian Discriminant Analysis (GDA), Hidden Markov models (HMM), Probabilistic graphical models, K-nearest neighbors, Kernel regression, Kernel density estimation, Local regression, Classification and regression tree (CART), decision tree, κτλ. Στην επιβλεπόμενη μάθηση το σύστημα πρέπει να «μάθει» επαγωγικά μια συνάρτηση που ονομάζεται συνάρτηση στόχος (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου ή χαρακτηριστικά. Στην επιβλεπόμενη μάθηση διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής. Η ταξινόμηση ή κατηγοριοποίηση αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) (π.χ. ομάδα αίματος). Η παλινδρόμηση ή παρεμβολή αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (π.χ. πρόβλεψη ισοτιμίας νομισμάτων ή τιμής μετοχής).

Στην ενισχυτική μάθηση παρέχονται σε ένα μοντέλο κάποια κριτήρια επιτυχίας, όπως η νίκη σε ένα παιχνίδι ή η πλοήγηση σε μια επιφάνεια με εμπόδια. Ένας πράκτορας (agent), δηλ. η οντότητα που μαθαίνει και λαμβάνει αποφάσεις, αποβλέπει στο να βελτιστοποιήσει την αλληλεπίδραση του με ένα δυναμικό περιβάλλον μέσω διαδικασίας trial and error. Οτιδήποτε άλλο εκτός του πράκτορα ονομάζεται περιβάλλον. Ένα πρόβλημα ενισχυτικής μάθησης περιλαμβάνει τρία μέρη: Το περιβάλλον, τη συνάρτηση τιμής της κατάστασης και τη συνάρτηση ενίσχυσης. Ο πράκτορας και το περιβάλλον αλληλεπιδρούν συνεχώς, με τον πρώτο να επιλέγει ενέργειες και το δεύτερο να αποκρίνεται σε αυτές και να του παρουσιάζει καινούριες καταστάσεις. Το περιβάλλον δίνει στον πράκτορα ανταμοιβές (rewards). Πρόκειται για ειδικές αριθμητικές τιμές τις οποίες ο πράκτορας προσπαθεί μακροπρόθεσμα να μεγιστοποιήσει.

Στη μη επιβλεπόμενη μάθηση, δεν υπάρχει η παρουσία δασκάλου/κριτή. Η μη επιβλεπόμενη μάθηση μπορεί να βρει εφαρμογή σε: Εξαγωγή χαρακτηριστικών (features extraction), συσταδοποίηση (clustering) και ανίχνευση μη ομαλών δεδομένων (anomaly detection) [12].

Εισαγωγή στην κατηγοριοποίηση δεδομένων

Ένα τυπικό πρόβλημα αναγνώρισης προτύπων είναι ο διαχωρισμός ενός συνόλου προτύπων $O=\{o_1,o_2,..\}$ σε C υποσύνολα $O_1,O_2,...,O_C$ που περιλαμβάνουν πρότυπα που ανήκουν στο ίδιο υποσύνολο και τα υποσύνολα είναι αμοιβαίως αποκλειόμενα:

$$O=\bigcup_{i=1}^{C}O_i \text{ και } (\forall i,j\in\{1,2,...,C\}, i\neq j) O_i\cap O_j=\varnothing \quad (1)$$

Αυτός ο διαχωρισμός ορίζεται από μια χαρτογράφηση (mapping) που καλείται κατηγοριοποιητής (classifier) $\Psi: O\to\Theta$ όπου $\Theta=\{O_1,O_2,...,O_C\}$ είναι ένα σύνολο κλάσεων. Για λόγους απλότητας θεωρούμε ότι η χαρτογράφηση Ψ λαμβάνει τιμές από το σύνολο των δεικτών των κλάσεων $\Theta=\{1,2,...,C\}$ δηλαδή των ετικετών των κλάσεων αντί των ίδιων των κλάσεων. Προφανώς οι ετικέτες των συστάδων μπορεί να είναι διαφορετικές από τους αριθμούς $1,2,...,C.$ Για παράδειγμα, μπορεί να λαμβάνουν τιμές -1 και 1 σε ένα πρόβλημα 2 κλάσεων. Η αναγνώριση προτύπων συνήθως λαμβάνει χώρα σε χαρακτηριστικά (features) που χαρακτηρίζουν τα πρότυπα παρά απ` ευθείας στα ίδια τα πρότυπα. Ο εξαγωγέας χαρακτηριστικών (feature extractor) είναι η χαρτογράφηση από το χώρο των προτύπων O στο χώρο των χαρακτηριστικών $X, \varphi: O\to X.$ Στη συνέχεια θεωρούμε μια χαρτογράφηση από το χώρο των χαρακτηριστικών στο χώρο των κλάσεων $\psi: X\to\Theta$ που καλείται επίσης κατηγοριοποιητής. Συνεπώς η έννοια του κατηγοριοποιητή έχει δύο σημασίες: κατηγοριοποίηση των προτύπων και κατηγοριοποίηση των χαρακτηριστικών. Η σύνθεση των δύο αυτών χαρτογραφήσεων συνιστά τον κατηγοριοποιητή $\Psi=\psi\circ\varphi.$ Με άλλα λόγια, η χαρτογράφηση

$$O\xrightarrow{\Psi}\Theta \quad (2)$$

αναλύεται σε

$$O\xrightarrow{\varphi}X\xrightarrow{\Psi}\Theta \quad (3)$$

Η χαρτογράφηση Ψ λαμβάνει τιμές από το σύνολο των ετικετών των κλάσεων $\Theta=\{1,2,...,C\}$ παρά των ίδιων των κλάσεων. Ο κατηγοριοποιητής Ψ θεωρείται γνωστός για ένα υποσύνολο όλων των προτύπων, που καλείται σύνολο εκπαίδευσης (learning set). Το σύνολο εκπαίδευσης είναι ένα υποσύνολο $L\subset O$ στο οποίο οι κλάσεις είναι γνωστές, δηλαδή για κάθε πρόττυπο αυτού του υποσυνόλου $o\subset L$ η τιμή $\Psi(o)$ είναι γνωστή. Ο στόχος της αναγνώρισης προτύπων είναι να κατασκευαστεί ο κατηγοριοποιητής Ψ δοθέντος ενός συνόλου εκπαίδευσης,

$$\Psi: O\to\Theta \quad (4)$$

και θεωρώντας ότι αυτή η χαρτογράφηση είναι γνωστή για $L\subset O.$ Ο κατηγοριοποιητής αναλύεται σε εξαγωγέα χαρακτηριστικών

$$\varphi: O\to X \quad (5)$$

και σε κατηγοριοποιητή χαρακτηριστικών ή αλγόριθμο κατηγοριοποίησης

$$\psi : X \to \Theta \quad (6)$$

Δοθέντος ένος συνόλου προτύπτων εκπαίδευσης $O \supset L = \{l_1, l_2, ..., l_M\}$ και του διαχωρισμού του σε κλάσεις

$$L = \bigcup_{i=1}^{C} L_i \text{ και } (\forall i \in \{1,2,...,C\})\ L_i \subset O_i \quad (7)$$

Οι κλάσεις εκπαίδευσης είναι επίσης αμοιβαίως αποκλειόμενες:

$$L = \bigcup_{i=1}^{C} L_i \text{ και } (\forall i, j \in \{1,2,...,C\}, i \neq j)\ L_i \cap \subset L_j = \varnothing \quad (8)$$

Το σύνολο εκπαίδευσης θα πρέπει να περιέχει πρότυποα από όλες τις κλάσεις στο O. Επίσης, κάθε κλάση εκπαίδευσης διαχωρίζεται στο υποσύνολο εκπαίδευσης Tr και στο υποσύνολο ελέγχου (test) Ts, ώστε [13]

$$L_i = Tr_i \cup Ts_i, Tr_i \cap Ts_i = \varnothing, i = 1,2,..,C \quad (9)$$

Παρουσίαση συνόλου δεδομένων και αρχική επεξεργασία τους.

Τα δεδομένα αφορούν το πιο δημοφιλές πρωτάθλημα μπάσκετ της Αμερικής , National Basketball Association . Κάθε γραμμή από το τελικό σύνολο δεδομένων περιέχει στατιστικές κάθε ομάδας για κάθε ολοκληρωμένη σεζόν και μια στήλη που υποδεικνύει για το εάν η ομάδα προκ9ρίθηκε στα playoffs ή όχι.

Λόγω της ανάγκης να συγκεντρωθούν όσο το δυνατόν περισσότερες δεδομένα σε ένα σταθερό σύστημα τουρνουά, συγκεντρώθηκαν δεδομένα από σεζόν 2005-2006 έως 2015-2016. Αυτό αποφασίστηκε λόγω ότι το NBA ακολουθεί το ίδιο διαχωριστικό σύστημα από την εποχή 2006-2006. Από το σύνολο δεδομένων του NBA, η σεζόν 2011-2012 αποκλείστηκε, καθώς μειώθηκαν από τα κανονικά 82 παιχνίδια ανά ομάδα σε 66, λόγω αδράνειας των ομάδων για περίπου δύο μήνες.

Το σύνολο δεδομένων του NBA έχει 300 (30 ομάδες ανά σεζόν) παρατηρήσεις των 30 χαρακτηριστικών.

Η συλλογή δεδομένων αποτελεί το πρώτο δομικό στοιχείο της διαδικασίας της μηχανικής μάθησης . Στη συνέχεια έρχεται ή αρχική ανάλυση δεδομένων (Initial Data Analysis -IDA). Η IDA βεβαιώνει ότι τα δεδομένα είναι καθαρά, σωστά και πλήρη για περαιτέρω διερευνητική ανάλυση. Η διαδικασία της IDA περιλαμβάνει την προετοιμασία των δεδομένων με τις σωστές ονομασίες και τύπους δεδομένων για τις μεταβλητές, ελέγχοντας για τις τιμές που λείπουν και τις ακραίες τιμές. Είναι ένα σημαντικό βήμα στη δημιουργία την κατανόηση και τις πληροφορίες από τα δεδομένα. Αυτή η πρώτη ματιά δημιουργεί τη διαίσθηση και την κατανόηση των προτύπων και των τάσεων. Σε αυτό ειδικά το σύνολο δεδομένων όμως και οι ακραίες τιμές έχουν πολύ μεγάλο ρόλο στην διεξαγωγή των αποτελεσμάτων.

Καλώντας τη λειτουργία str () του R, βλέπουμε για το σύνολο δεδομένων το Πίνακα 1 για το NBA. Οι έξοδοι δείχνουν τέσσερις χρήσιμες πληροφορίες:

• Ο αριθμός των γραμμών και των στηλών στα δεδομένα

• Όνομα μεταβλητής ή κεφαλίδα στήλης στα δεδομένα

• Τύπος δεδομένων κάθε μεταβλητής

• Τιμές δείγματος για κάθε μεταβλητή

```
'data.frame':   300 obs. of  30 variables:
 $ Season      : Factor w/ 10 levels "2005-2006","2006-2007",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Qualified   : int  1 0 1 0 1 1 1 0 1 1 ...
 $ Team        : Factor w/ 35 levels "Atlanta Hawks",..: 28 32 35 33 9 17 15 27 8 30 ...
 $ G           : int  82 82 82 82 82 82 82 82 82 82 ...
 $ MP          : int  19955 19830 19830 19955 19930 19755 19855 19905 19880 19830 ...
 $ FG          : int  3430 3077 2975 3013 3078 3039 2992 3001 2948 2954 ...
 $ FGA         : int  7167 6711 6656 6639 6672 6355 6607 6546 6375 6500 ...
 $ FG.         : num  0.479 0.459 0.447 0.454 0.461 0.478 0.453 0.458 0.462 0.454 ...
 $ X3P         : int  837 605 497 608 350 497 552 375 416 494 ...
 $ X3PA        : int  2097 1631 1394 1620 1076 1441 1583 1031 1113 1408 ...
 $ X3P.        : num  0.399 0.371 0.357 0.375 0.325 0.345 0.349 0.364 0.374 0.351 ...
 $ X2P         : int  2593 2472 2478 2405 2728 2542 2440 2626 2532 2460 ...
 $ X2PA        : int  5070 5080 5262 5019 5596 4914 5024 5515 5262 5092 ...
 $ X2P.        : num  0.511 0.487 0.471 0.479 0.487 0.517 0.486 0.476 0.481 0.483 ...
 $ FT          : int  1189 1652 1889 1653 1721 1616 1618 1770 1818 1704 ...
 $ FTA         : int  1475 2104 2496 2089 2312 2310 2172 2330 2322 2173 ...
 $ FT.         : num  0.806 0.785 0.757 0.791 0.744 0.7 0.745 0.76 0.783 0.784 ...
 $ ORB         : int  778 1013 1035 864 903 858 970 873 1030 852 ...
 $ DRB         : int  2650 2233 2344 2291 2486 2675 2488 2425 2431 2473 ...
 $ TRB         : int  3428 3246 3379 3155 3389 3533 3458 3298 3461 3325 ...
 $ AST         : int  2179 1696 1523 1593 1921 1692 1734 1653 1473 1825 ...
 $ STL         : int  549 622 658 529 698 522 628 651 593 608 ...
 $ BLK         : int  412 306 339 272 463 442 350 404 488 299 ...
 $ TOV         : int  1088 1208 1143 1071 1216 1186 1143 1159 1112 1199 ...
 $ PF          : int  1683 1933 1855 1965 1863 1871 1894 1712 1834 1672 ...
 $ PTS         : int  8886 8411 8336 8287 8227 8191 8154 8147 8130 8106 ...
 $ PLAYER.COUNT: int  15 14 14 15 17 17 17 16 16 14 ...
 $ AVG.HT      : num  2.04 2.04 2.07 2.04 2.01 ...
 $ AVG.AGE     : num  26.3 24.8 25.6 26.4 27 26.9 24.6 25.4 26.6 26.9 ...
 $ SALARY....  : int  72475222 47251655 48960907 54633268 54211523 59858488 59143352 60084965 69186797 58751977 ...
```

Εικόνα 1 - Οθόνη εξόδου συνάρτησης str() από το R-Studio

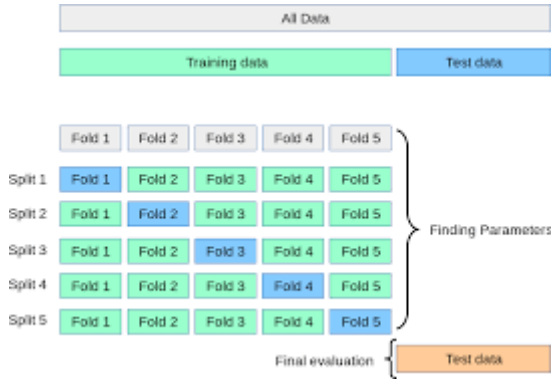
Η προεπεξεργασία των δεδομένων έγινε στο περιβάλλον εργασίας του RapidMiner. Το RapidMiner είναι ένα λογισμικό ανάλυσης δεδομένων και μηχανικής μάθησης το οποίο βασίζεται στην λογική του workflow oriented software δηλαδή με ένα πλήρως παραμετροποιήσιμο περιβάλλον διαδικασιών και ροών εργασίας να γίνεται η υλοποίηση και η δοκιμή τεχνικών μηχανικής μάθησης και προ-επεξεργασίας δεδομένων.

Όλα τα χαρακτηριστικά του συνόλου δεδομένων μετατράπηκαν στο διάστημα [0-1] , είτε σε συνεχές είτε σε διακριτό διάστημα (0 ή 1). Αυτός ο μετασχηματισμός έγινε διότι οι αλγόριθμοι της μηχανικής μάθησης δέχονται και επεξεργάζονται καλύτερα τα δεδομένα σε αριθμητικό διάστημα (0-1) και επιπλέον υπάρχει το πλεονέκτημα ότι τα χαρακτηριστικά (features) δεν χάνουν την εσωτερική διακύμανσή τους . Τέλος σημαντικό είναι το πλεονέκτημα του ότι σε αλγορίθμους όπως παραδείγματος χάριν τα Νευρωνικά δίκτυα (NN) και Μηχανές Διανυσμάτων Υποστήριξης (SVM) όπου αναθέτονται βάρη σε χαρακτηριστικά η κανονικοποίηση προσφέρει θετικά αποτελέσματα στην απόδοση των αλγορίθμων [12]

Η εξακρίβωση της απόδοσης κάθε αλγορίθμου έγινε με διαδικασία δισταυρούμενης επικύρωσης (Cross-Validation)

Διασταυρούμενη επικύρωση (Cross-Validation)

Η διασταυρούμενη επικύρωση, μερικές φορές ονομάζεται εκτίμηση εκ - περιστροφής, δοκιμή είναι οποιαδήποτε από τις διάφορες παρόμοιες τεχνικές επικύρωσης μοντέλου για την αξιολόγηση του τρόπου γενίκευσης των αποτελεσμάτων μιας στατιστικής ανάλυσης σε ένα ανεξάρτητο σύνολο δεδομένων [20]. Χρησιμοποιείται κυρίως σε περιπτώσεις όπου ο στόχος είναι η πρόβλεψη και κάποιος θέλει να υπολογίσει με ακρίβεια πόσο ακριβές ένα μοντέλο πρόβλεψης θα εκτελεστεί στην πράξη. Σε ένα πρόβλημα πρόβλεψης, ένα μοντέλο συνήθως λαμβάνει ένα σύνολο δεδομένων με γνωστά δεδομένα για την οποία εκτελείται η κατάρτιση (σύνολα δεδομένων κατάρτισης) και ένα σύνολο δεδομένων άγνωστων δεδομένων (ή δεδομένων πρώτης προβολής) με τα οποία δοκιμάζεται το μοντέλο (που ονομάζεται σύνολο δεδομένων επικύρωσης ή σύνολο δοκιμής). Ο στόχος της διασταυρούμενης επικύρωσης είναι να δοκιμάσει την ικανότητα του μοντέλου να προβλέψει νέα δεδομένα που δεν χρησιμοποιήθηκαν για την εκτίμησή του, προκειμένου να επισημάνει προβλήματα όπως η υπερφόρτωση (overfitting) ή η μεροληψία επιλογής (biased selection) και να δώσει μια εικόνα για το πώς το μοντέλο θα γενικευθεί σε ένα ανεξάρτητο σύνολο δεδομένων (δηλαδή ένα άγνωστο σύνολο δεδομένων, για παράδειγμα από ένα πραγματικό πρόβλημα).



Εικόνα 2. Διαδικασία Cross-Validation

Ένας κύκλος διασταυρούμενης επικύρωσης περιλαμβάνει τη διαίρεση δείγματος δεδομένων σε συμπληρωματικά υποσύνολα, την εκτέλεση της ανάλυσης σε ένα υποσύνολο (που ονομάζεται σετ κατάρτισης) και την επικύρωση της ανάλυσης στο άλλο υποσύνολο (που ονομάζεται σύνολο επικύρωσης ή σύνολο δοκιμών). Για να μειωθεί η μεταβλητότητα, στις περισσότερες μεθόδους πραγματοποιούνται πολλαπλοί γύροι διασταυρούμενης επικύρωσης χρησιμοποιώντας διαφορετικά διαμερίσματα και τα αποτελέσματα επικύρωσης συνδυάζονται (π.χ. κατά μέσο όρο) στους γύρους για να δώσουν μια εκτίμηση της προβλεπτικής απόδοσης του μοντέλου

Συνοπτικά, η εγκάρσια επικύρωση συνδυάζει (μέσες τιμές) τα μέτρα καταλληλόλητας στην πρόβλεψη για την εξαγωγή ακριβέστερης εκτίμησης της απόδοσης πρόβλεψης του μοντέλου.[21]

Ακρίβεια, απόδοση και μετρικές μήτρας σύγχυσης (confusion matrix)

Ένας πίνακας σύγχυσης είναι ένας πίνακας που συχνά χρησιμοποιείται για να περιγράψει την απόδοση ενός μοντέλου ταξινόμησης (ή "ταξινομητή") σε ένα σύνολο δεδομένων δοκιμής (test data) για τα οποία είναι γνωστές οι πραγματικές τιμές.

Ας ξεκινήσουμε με ένα πρότυπο μήτρας σύγχυσης για έναν δυαδικό ταξινομητή (αν και μπορεί εύκολα να επεκταθεί στην περίπτωση περισσότερων από δύο τάξεων):

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Πίνακας 1 - Παράδειγμα Μήτρας Σύγχυσης (Confusion Matrix)

Υπάρχουν δύο πιθανές προβλεπόμενες κλάσεις: "ναι" και "όχι". Εάν προβλέψαμε την ύπαρξη μιας ασθένειας, για παράδειγμα, «ναι» σημαίνει ότι έχουν την ασθένεια και το «όχι» σημαίνει ότι δεν έχουν την ασθένεια.

Ο ταξινομητής πραγματοποίησε συνολικά 165 προβλέψεις (π.χ., δοκιμάστηκαν 165 ασθενείς για την παρουσία της νόσου αυτής).

Από τις 165 περιπτώσεις, ο ταξινομητής προέβλεψε "ναι" 110 φορές και "όχι" 55 φορές.

Στην πραγματικότητα, 105 ασθενείς στο δείγμα έχουν την ασθένεια και 60 ασθενείς δεν το κάνουν.

Ας οριστούν τώρα οι πιο βασικοί όροι :

- **Αληθώς θετικά (TP – True Positive):** Αυτές είναι περιπτώσεις στις οποίες προβλέφθηκε ναι (έχουν την ασθένεια) και έχουν την ασθένεια.
- **Αληθώς αρνητικά (TN – True Negative):** Δεν προβλέφθηκε κανένα και δεν έχουν την ασθένεια.
- **Ψευδώς θετικά (FP – False Positive):** Προβλέφθηκε ναι, αλλά στην πραγματικότητα δεν έχουν την ασθένεια. (Επίσης γνωστό ως "σφάλμα τύπου I".)
- **Ψευδώς αρνητικά (FN – False Negative):** Προβλέψαμε όχι, αλλά στην πραγματικότητα έχουν την ασθένεια. (Επίσης γνωστό ως "σφάλμα τύπου II").

- **Ακρίβεια:** Συνολικά, πόσο συχνά είναι ο ταξινομητής σωστός;

- (TP + TN) / σύνολο = (100 + 50) / 165 = 0,91

- **Ποσοστό εσφαλμένης ταξινόμησης:** Συνολικά, πόσο συχνά είναι λάθος;
  - (FP + FN) / σύνολο = (10 + 5) / 165 = 0,09
  - ισοδύναμη με 1 μείον Ακρίβεια
  - επίσης γνωστή ως "Βαθμός σφάλματος"
- **Αληθινό θετικό ποσοστό:** Όταν είναι ναι, πόσο συχνά προδίδει ναι;
  - TP / πραγματικό ναι = 100/105 = 0,95
  - επίσης γνωστή ως "ευαισθησία" ή "ανάκληση"
- **False Positive Rate:** Όταν δεν υπάρχει πραγματικά, πόσο συχνά προβλέπει ότι ναι;
  - FP / actual no = 10/60 = 0,17
- **Ο πραγματικός αρνητικός ρυθμός:** Όταν πραγματικά δεν υπάρχει, πόσο συχνά δεν προβλέπει;
  - TN / πραγματικό όχι = 50/60 = 0,83
  - ισοδύναμη με 1 μείον ψευδώς θετικό ποσοστό
  - επίσης γνωστή ως "ειδικότητα"
- **Ακρίβεια:** Όταν προβλέπει ναι, πόσο συχνά είναι σωστό;
  - TP / προβλεπόμενη ναι = 100/110 = 0,91
- **Επικράτηση:** Πόσο συχνά συμβαίνει η κατάσταση ναι στο δείγμα μας;
  - πραγματική ναι / συνολική = 105/165 = 0,64

Μερικοί όροι που αξίζει να αναφερθούν [19]:

- **Μηδενικός ρυθμός σφάλματος:** Αυτό είναι πόσο συχνά θα ήταν λάθος αν προέβλεπε πάντα την τάξη πλειοψηφίας. (Στο παράδειγμα μας, το μηδενικό ποσοστό σφάλματος θα είναι 60/165 = 0,36 γιατί αν προβλέπετε πάντα ναι, θα κάνατε λάθος μόνο για τις 60 "όχι" περιπτώσεις.) Αυτό μπορεί να είναι μια χρήσιμη βασική μέτρηση για να συγκρίνετε τον ταξινομητή σας. Ωστόσο, ο καλύτερος ταξινομητής για μια συγκεκριμένη εφαρμογή θα έχει μερικές φορές υψηλότερο ρυθμό σφάλματος από τον μηδενικό ρυθμό σφάλματος.
- **Cohen's Kappa:** Αυτό είναι ουσιαστικά ένα μέτρο του πόσο καλά ο ταξινομητής εκτελεί σε σύγκριση με το πόσο καλά θα είχε εκτελέσει απλά τυχαία. Με άλλα λόγια, ένα μοντέλο θα έχει υψηλή βαθμολογία Kappa αν υπάρχει μεγάλη διαφορά μεταξύ της ακρίβειας και του μηδενικού ποσοστού σφάλματος.
- **F Μετρική (F-Measurement):** Αυτός είναι ο σταθμισμένος μέσος όρος του πραγματικού θετικού ποσοστού (ανάκληση) και της ακρίβειας.
- **Καμπύλη ROC:** Πρόκειται για ένα συνηθισμένο γράφημα που συνοψίζει την απόδοση ενός ταξινομητή σε όλα τα πιθανά όρια. Δημιουργείται με την

αντιστοίχιση του πραγματικού θετικού ρυθμού (άξονας y) έναντι του ψευδώς θετικού ρυθμού (άξονας x) καθώς μεταβάλλετε το κατώφλι για την αντιστοίχιση παρατηρήσεων σε μια δεδομένη κλάση.

Επιλογή των ανεξάρτητων μεταβλητών μέσω της μεθόδου Forward Selection

Ο τρόπος επιλογής χαρακτηριστικών Forward-Selection ξεκινά με μια κενή επιλογή χαρακτηριστικών και σε κάθε επανάληψη, προσθέτει κάθε μη χρησιμοποιημένο χαρακτηριστικό από τα δεδομένα . Για κάθε πρόσθετο χαρακτηριστικό, η απόδοση εκτιμάται χρησιμοποιώντας τους εσωτερικούς χειριστές, π.χ. μια διασταυρούμενη επικύρωση (Cross-Validation). Μόνο το χαρακτηριστικό που δίνει την υψηλότερη αύξηση της απόδοσης προστίθεται στην επιλογή. Στη συνέχεια ξεκινάει μια καινούρια επανάληψη με την τροποποιημένη επιλογή. Αυτή η εφαρμογή αποφεύγει οποιαδήποτε πρόσθετη κατανάλωση μνήμης εκτός από τη μνήμη που χρησιμοποιήθηκε αρχικά για την αποθήκευση των δεδομένων και της μνήμης που μπορεί να χρειαστεί για την εφαρμογή των εσωτερικών χειριστών. Υπάρχουν τρεις διαφορετικές επιλογές για το πότε θα σταματήσει η αυτή η επαναληπτική διαδικασία:

1. Χωρίς αύξηση: Η επανάληψη διαρκεί όσο υπάρχει αύξηση της απόδοσης.
2. Χωρίς ιδιαίτερη αύξηση: Η επανάληψη διαρκεί όσο η αύξηση είναι τουλάχιστον τόσο υψηλή όσο καθορίζεται, είτε σχετική είτε απόλυτη. Η παράμετρος ελάχιστης σχετικής αύξησης χρησιμοποιείται για τον προσδιορισμό της ελάχιστης σχετικής αύξησης εάν η παράμετρος σχετικής αύξησης χρήσης έχει οριστεί ως αληθής. Διαφορετικά, η ελάχιστη παράμετρος απόλυτης αύξησης χρησιμοποιείται για τον προσδιορισμό της ελάχιστης απόλυτης αύξησης.
3. χωρίς σημαντική αύξηση: Η επανάληψη σταματά μόλις η αύξηση δεν είναι σημαντική για το επίπεδο που καθορίζεται από την παράμετρο alpha.

Η παράμετρος των κερδοσκοπικών γύρων καθορίζει τον αριθμό των γύρων που θα εκτελεστούν σε μια σειρά, μετά την πρώτη φορά που πληρούται το κριτήριο στάσης. Αν η απόδοση αυξηθεί ξανά κατά τη διάρκεια των κερδοσκοπικών γύρων, η επιλογή θα συνεχιστεί. Διαφορετικά, όλα τα επιπρόσθετα επιλεγμένα χαρακτηριστικά θα αφαιρεθούν, σαν να μην είχαν εκτελεστεί κερδοσκοπικοί γύροι. Αυτό μπορεί να βοηθήσει να αποφευχθούν αυτοί οι γύροι και να βρεθεί το τοπικό βέλτιστό σημείο.

Η επιλογή χαρακτηριστικών, δηλαδή η ερώτηση για τις πιο σχετικές λειτουργίες για προβλήματα κατηγοριοποίησης ή παλινδρόμησης, είναι μία από τις κύριες εργασίες εξόρυξης δεδομένων.

Ένα ευρύ φάσμα μεθόδων αναζήτησης έχουν ενσωματωθεί στο RapidMiner συμπεριλαμβανομένων εξελικτικών αλγορίθμων. Για όλες τις μεθόδους αναζήτησης, χρειαζόμαστε μια μέτρηση απόδοσης που υποδεικνύει πόσο καλά θα εκτελεστεί ένα

σημείο αναζήτησης (ένα υποσύνολο χαρακτηριστικών) στο συγκεκριμένο σύνολο δεδομένων.

Η επιλογή του υποσυνόλου γίνεται με βάση αλγορίθμους μηχανικής μάθησης , δηλαδή με την συμβολή των αλγορίθμων Naïve Bayes, Svm,K-nn

Αλγόριθμοι εποπτευόμενης μάθησης (Supervised Machnine Learning Algorithms)

Support Vector Machines

Τα Support Vector Machines (SVMs) είναι αλγόριθμος που προτάθηκε από τους C. Cortes και V. Vapnik το 1995. Στη βασική μορφή τους πρόκειται για μη πιθανοκρατικό δυαδικό γραμμικό κατηγοριοποιητή που χρησιμοποιείται για να διαχωρίσει ένα σύνολο προτύπων $O=\{o_1,o_2,...,o_N\}$ σε δύο κλάσεις με ετικέτες -1 και 1, δηλαδή, $\Theta=\{O_{-1},O_1\}$. Θεωρούμε ότι τα χαρακτηριστικά που χαρακτηρίζουν τα πρότυπα είναι $\varphi: O\to X; X\subset R^M$. Αναζητείται η χαρτογράφηση $\psi: X\to\{-1,1\}$. Θεωρούμε ότι οι κλάσεις O_{-1} και O_1 είναι γραμμικώς διαχωρίσιμες στον Ευκλείδιο χώρο R^M των χαρακτηριστικών και ότι υπάρχει ένα υπερπλάνο (hyperplane) που διαχωρίζει τις κλάσεις. Θεωρούμε επίσης ότι η ακόλουθη σχέση ορίζει ένα τέτοιο υπερπλάνο H':

$$w'^T x+b'\equiv w'x+b'=0 \tag{11}$$

όπου το x δηλώνει το υπερπλάνο στον Ευκλείδιο χώρο R^M, w' είναι ένα τυπικό διάνυσμα στο υπερπλάνο, $b'\in R$ είναι βαθμωτό μέγεθος και $w'^T x$ δηλώνει το γινόμενο των πινάκων και όπου για διανύσματα είναι ισοδύναμο με το βαθμωτό γινόμενο $w'x$. Τα πρότυπα κατηγοριοποιούνται ανάλογα με την πλευρά του υπερεπιπέδου στην οποία βρίσκονται. Εάν το υπερπλάνο διαχωρίζει τις κλάσεις O_{-1} και O_1 τότε ισχύουν οι παρακάτω ανισότητες:

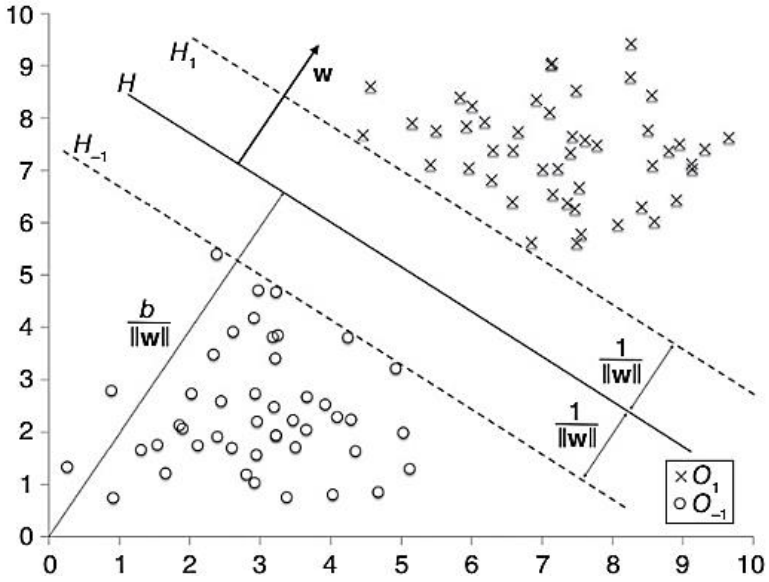
$w'x_i+b'<0$ για όλα $x_i=\varphi(o_i)$ έτσι ώστε $o_i\in O_{-1}$

$w'x_i+b'>0$ για όλα $x_i=\varphi(o_i)$ έτσι ώστε $o_i\in O_1$

Συνεπώς, προκύπτει ο παρακάτω κανόνας κατηγοριοποίησης

$$\psi(x_i)=\mathrm{sgn}(w'x_i+b') \tag{12}$$

Στο Σχήμα 2 παρουσιάζεται μια περίπτωση κατηγοριοποίησης με το SVM.



Σχήμα 2: Περίπτωση κατηγοριοποίησης με SVM [13].

Όπως φαίνεται από το Σχήμα 2, μπορεί να υπάρχουν πολλά διαχωριστικά υπερπλάνα. Το πρόβλημα της κατηγοριοποίησης μέσω SVM ανάγεται σε πρόβλημα καθορισμού του υπερπλάνου. Εάν υπάρχουν περισσότερα του ενός, θα πρέπει να ευρεθεί το καταλληλότερο. Θεωρούμε αποστάσεις μεταξύ του υπερπλάνου και πρότυπα και στις δύο πλευρές του. Για μια δεδομένη κλάση λαμβάνουμε ένα πρότυπο του οποίου η απόσταση από το υπερπλάνο είναι η ελάχιστη μέσα στην κλάση. Τα πρότυπα που αντιστοιχούν στις ελάχιστες αποστάσεις καλούνται Support Vectors (SVs), δηλαδή διανύσματα που ξεκινούν από τις αρχές του συστήματος συντεταγμένων και τερματίζουν στο πρότυπο. Αυτά μπορεί να χρησιμοποιηθούν για να κατασκευαστούν τα συνοριακά υπερπλάνα H'_{-1} και H'_{1}, ένα για κάθε κλάση. Η περιοχή μεταξύ αυτών καλείται περιθώριο (margin) δεν περιέχει κάποιο πρότυπο και αντιστοιχεί στη μικρότερη απόσταση από το συνοριακό υπερπλάνο. Η μεγιστοποίηση της απόστασης μεταξύ των συνοριακών υπερπλάνων, δηλαδή με την αύξηση του πλάτους του περιθωρίου, διαμορφώνται ισομερώς τα συνοριακά υπερπλάνα. Με την τοποθέτηση ενός διαχωριστικού υπερπλάνου στο μέσο του συνοριακού περιθωρίου μπορεί να γίνει η εν λόγω διαμόρφωση. Θεωρώντας ότι το διαχωριστικό υπερπλάνο ορίζεται ως $w'x + b' = 0$ προκύπτουν οι εξής ανισότητες

$w'x_i + b' \le -c$ για τα πρότυπα x_i που προέρχονται από την κλάση O_{-1}

$w'x_i + b' > c$ για τα πρότυπα x_i που προέρχονται από την κλάση O_1

Στις παραπάνω ανισότητες η μεταβλητή c είναι μια σταθερά. Μετά από κλιμάκωση, προκύπτουν οι εξής εξισώσεις για το διαχωριστικό υπερπλάνο H και τα διαχωριστικά υπερπλάνα H_{-1} και H_1, αντίστοιχα:

$$wx + b = 0 \quad (13)$$

$$wx + b = -1 \quad (14)$$

$$wx + b = 1 \quad (15)$$

Οι ανισότητες για τα πρότυπα των δύο κλάσεων είναι

$wx + b \le -1$ (16)

$wx + b \ge 1$ (17)

με $w = w'/c$ και $b = b'/c$.

Μπορούμε να συνδυάσουμε τις δύο ανισότητες ως

$y_i(wx_i + b) \ge 1$ (18)

για τα πρότυπα των δύο κλάσεων, όπου $y_i = \psi(x_i) = \psi(\varphi(o_i))$ είναι η ετικέτα της κλάσης του o_i. Η απόσταση μεταξύ του διαχωριστικού υπερπλάνου H και των συνοριακών υπερπλάνων H_{-1} και H_1 είναι ίση με $1/\|w\|$ ώστε η απόσταση μεταξύ των διαχωριστικών είναι ίση με $2/\|w\|$. Συνεπώς, η βέλτιστη τοποθέτηση του διαχωριστικού υπερπλάνου, δηλαδή η τοποθέτηση που μεγιστοποιεί το περιθώριο απαιτεί τη μεγιστοποίηση του όρου $1/\|w\|$ υπό τον περιορισμό

$$\min_{i=1,2,..,n} |wx_i + b| = 1 \quad (19)$$

Η μεγιστοποίηση του $1/\|w\|$ ισοδυναμεί με την ελαχιστοποίηση του $\|w\|^2/2$. Το πρόβλημα μπορεί να διαμορφωθεί ως πρόβλημα βελτιστοποίησης με τετραγωνική αντικειμενική συνάρτηση $\|w\|^2 = w \cdot w = w_1^2 + w_2^2 + w_3^2 + ... + w_M^2$ με γραμμικούς περιορισμούς. Τέτοιου είδους πρόβλημα λύνεται με τη συνάρτηση Lagrange

$$L(w,b,a) = \frac{1}{2}\|w\| - \sum_{i=1}^{N} a_i[y_i(wx_i + b) - 1] \quad (20)$$

όπου $a = (\alpha_1, \alpha_2, ..., \alpha_N)$ είναι ένα διάνυσμα με μη αρνητικούς πολλαπλασιαστές Lagrange. Η ποσότητα $a_i[y_i(wx_i + b) - 1]$ είναι μη αρνητική έτσι ώστε το $L(w,b,a)$ θα μεγιστοποιηθεί θεωρώντας την παρακάτω συνθήκη

$a_i[y_i(wx_i + b) - 1] = 0, i = 1, 2, ..., N$ (21)

που υπονοεί ότι $a_i = 0$ εάν το x_i είναι support vector. Λαμβάνοντας τη μερική παράγωγο ως προς το w

$$\frac{\partial}{\partial w} L(w,b,a) = w - \sum_{i=1}^{N} a_i y_i x_i = 0$$

προκύπτει ότι

$$w = \sum_{i=1}^{N} a_i y_i x_i = 0 \quad (22)$$

και σχετικά με την παράμετρο b έχουμε

$$\frac{\partial}{\partial w} L(w,b,a) = \sum_{i=1}^{N} a_i y_i x_i = 0 \quad (23)$$

Η (20) μπορεί να γραφτεί ως

$$L(w,b,a)=\frac{1}{2}w^T w-\sum_{i=1}^{N}a_i y_i w x_i-b\sum_{i=1}^{N}a_i y_i+\sum_{i=1}^{N}a_i$$

Η τελική έκφραση για τη βελτιστοποίηση είναι

$$L(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j x_i^T x_i \tag{24}$$

υπό τους περιορισμούς

$$\alpha_i\geq 0, i=1,2,...,N, \sum_{i=1}^{N}a_i y_i=0 \tag{25}$$

Όλη η προηγούμενη ανάλυση αναφέρεται σε προβλήματα που είναι γραμμικώς διαχωρίσιμα. Αυτή η συνθήκη δεν ισχύει πάντοτε. Για παράδειγμα, σε μια περίπτωση που τα πρότυπα δύο κλάσεων είναι τόσο κοντά μεταξύ τους όπου δεν μπορούν να διαχωριστούν με ένα υπερπλάνο. Σε αυτή την περίπτωση η (18) γράφεται ως

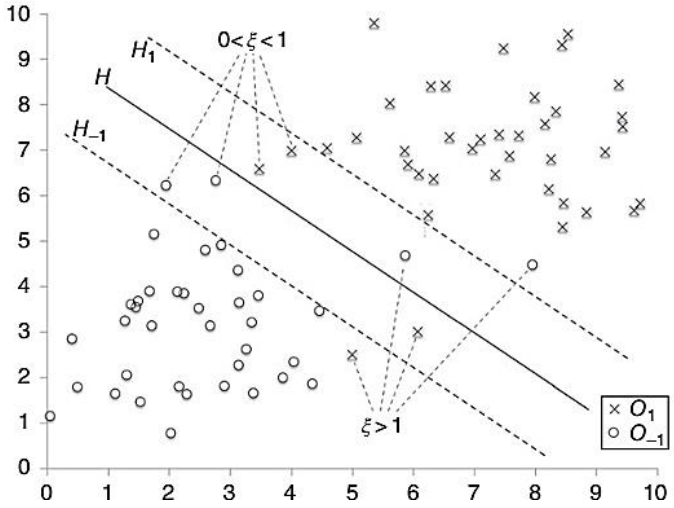
$$y_i(wx_i+b)\geq 1-\xi_i \tag{26}$$

όπου $\xi_i\geq 0, i=1,2,...,N$

Για $0<\xi_i<1$ το πρότυπο x_i είναι μέσα στο περιθώριο και πιο συγκεκριμένα μεταξύ τπυ διαχωριστικού υπερπλάνου και του συνοριακού υπερπλάνου στο οποίο ανήκει. Για $1<\xi_i$ το πρότυπο x_i βρίσκεται προς την πλευρά του άλλου συνοριακού υπερπλάνου. Το πρόβλημα της ελαχιστοποίησης στην περίπτωση μη γραμμικώς διαχωρίσιμων προτύπων γράφεται ως

$$\frac{\|w\|^2}{2}+C\sum_{i=1}^{N}\xi_i \tag{27}$$

και θεωρώντας τους περιοσμούς της (26). Η μεταβλητή C είναι θετικός πραγματικός αριθμός και καλείται ρυθμιστική παράμετρος (regularization parameter) και ελέγχει τη μορφή του περιθωρίου. Στο Σχήμα 3 παρουσιάζεται ένα παράδειγμα με μη γραμμικώς διαχωρίσιμα πρότυπα.



Σχήμα 3: Περίπτωση κατηγοριοποίησης με SVM με μη γραμμικώς διαχωρίσιμα πρότυπα [13].

Μια ιδανική κατηγοριοποίηση με μηδενικό σφάλμα είναι εφικτή στην περίπτωση δύο γραμμικώς διαχωρίσιμων κλάσεων. Για πιο σύνθετα προβλήματα, θα πρέπει να γίνει ένας μετασχηματισμός των προτύπων από ένα $M-$διάστατο χώρο σε ένα χώρο μεγαλύτερης διάστασης. Πιο συγκεκριμένα, ένα διάνυσμα χαρακτηριστικών $x \in I^M$ μπορεί να μετασχηματιστεί σε ένα διάνυσμα $g(x) \in I^K$ χρησιμοποιώντας μια συνάρτηση $g: I^M \to I^K$, όπου $I=[a,b] \subset R$ και $M < K$. Το βαθμωτό γινόμενο $x \cdot x_i$ μπορεί να αντικατασταθεί με το $g(x) \cdot g(x_i)$ και σε αυτή την περίπτωση το διαχωριστικό υπερπλάνο στο χώρο I^K αντιστοιχελι σε μια πιο σύνθετη υπερεπιφάνεια (hypersurface) στον αρχικό χώρο I^M των χαρακτηριστικών. Μια τέτοια υπερεπιφάνεια μπορεί να διαχωρίσει τα πρότυπα σε κλάσεις που δεν είναι γραμμικώς διαχώρισιμα. Η (12) μπορεί να γραφτεί ως

$$\psi(x_i) = \operatorname{sgn}(g(w) \cdot g(x)) = \operatorname{sgn}\left(\sum_{i \in SV} a_i^0 y_i g(x) \cdot g(x_i) + b^0 \right) \tag{28}$$

Αντί για τον υπολογισμό του μετασχηματισμού και στη συνέχεια του βαθμωτού γινομένου, μπορεί να χρησιμοποιηθεί μια ειδική συνάρτηση στο χώρο I^M. Μπορεί να αντικατασταθεί το βαθμωτό γινόμενο $g(x) \cdot g(x_i)$ με μια βαθμωτή συνάρτηση $K(x, x_i)$, που καλείται συνάρτηση πυρήνα (kernel function):

$$g(x) \cdot g(x_i) = K(x, x_i) \tag{29}$$

Έχουμε

$$\psi(x_i) = \operatorname{sgn}\left(\sum_{i \in SV} a_i^0 y_i K(x, x_i) + b^0 \right) \tag{30}$$

Υπάρχουν διάφορες συναρτήσεις που μπορούν να χρησιμοποιηθούν:

- Γραμμική συνάρτηση πυρήνα με μία παράμετρο c
  $K(x,y) = \langle x, y \rangle + c$
- Πολυωνυμική συνάρτηση πυρήνα βαθμού d
  $K(x,y) = (\langle x, y \rangle + c)^d$
- Gaussian συνάρτηση πυρήνα με παράμετρο γ που καλείται επίσης συνάρτηση ακτινωτής βάσης (radial basis function)
  $K(x,y) = \exp(-\gamma \|x-y\|^2)$
- Υπερβολική εφαπτομενική συνάρτηση πυρήνα με παράμετρο γ
  $K(x,y) = \tanh(\gamma \cdot \langle x, y \rangle + c)$
- Laplace συνάρτηση πυρήνα
  $K(x,y) = \exp(-\gamma |x-y|)$
- Sinc συνάρτηση πυρήνα
  $K(x,y) = \operatorname{sin} c(|x-y|) = \dfrac{\sin(|x-y|)}{|x-y|}$
- Sinc2 συνάρτηση πυρήνα

$$K(x,y) = \sin c2(|x-y|) = \frac{\sin(\|x-y\|^2)}{\|x-y\|^2}$$

- Quadtratic συνάρτηση πυρήνα

$$K(x,y) = 1 - \frac{\|x-y\|^2}{\|x-y\|^2 + c}$$

- Minimum συνάρτηση πυρήνα

$$K(x,y) = \sum_i \min(x_i - y_i)$$

όπου $\langle x, y \rangle$ είναι το βαθμωτό γινόμενο, $\|x-y\| = \sqrt{\sum_i (x_i - y_i)^2}$ είναι η Ευκλείδια νόρμα και $|x-y| = \sum_i |x_i - y_i|$ είναι η Manhattan νόρμα [13].

Generalized Linear Model

Η γραμμική παλινδρόμηση (linear regression) αναφέρεται στη γραμμική μοντελοποίηση της σχέσης μεταξύ μιας ή ή περισσότερων μεταβλητών σε μια άλλη. Διερευνάται δηλαδή η επίδραση των ανεξάρτητων μεταβλητών (independent variables) στην εξαρτημένη μεταβλητή (dependent variable). Οι ανεξάρτητες μεταβλητές καλούνται και επεξηγηματικές (explanatory). Η γραμμική παλινδρόμηση αποτελεί την πιο απλή εκδοχή της ανάλυσης παλινδρόμησης όπου η εξαρτημένη μεταβλητή εξαρτάται από μια μόνο εξαρτημένη. Η σχέση μεταξύ των δύο μοντελοποιείται θεωρώντας γραμμική συνάρτηση η οποία είναι άγνωστη και εκτιμάται από τα δεδομένα. Έστω $y_i, i = 1,2,...,n$ είναι η εξαρτημένη μεταβλητή και $x_j, j = 1,2,..., p$ οι ανεξάρτητες μεταβλητές. Η εξαρτημένη μεταβλητή εκφράζεται ως γραμμικός συνδυασμός των εξαρτημένων με την προσθήκη ενός όρου σφάλματος ε_i :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \quad (31)$$

όπου $\beta_0, \beta_1, \beta_2, ..., \beta_p$ είναι οι συντελεστές παλινδρόμησης και υπολογίζονται με τη μέθοδο των ελαχίστων τετραγώνων. Το σφάλμα ε_i καταγράφει όλους τους άλλους παράγοντες που επηρεάζουν τη εξαρτημένη μεταβλητή y_i εκτός από τις ανεξάρτητες μεταβλητές x_j. Θεωρούμε ότι τα σφάλματα ε_i είναι ανεξάρτητα μεταξύ τους και ακολουθούν κανονική κατανομή (normal distribution), $\varepsilon_i \square N(0, \sigma^2)$. Τα δύο βασικά μειονεκτήματα της γραμμικής παλινδρόμησης είναι ότι το εύρος τιμών του y_i είναι μειωμένο π.χ. λαμβάνει δυαδικές τιμές και ότι η διακύμανση (variance) του y_i εξαρτάται από το μέσο όρο (mean).

Το γενικευμένο γραμμικό μοντέλο (Generalized Linear Model, GLM) είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης που επιτρέπει μεταβλητές απόκρισης που έχουν

μοντέλα κατανομής σφάλματος διαφορετικά από μια κανονική κατανομή. Το GLM γενικεύει τη γραμμική παλινδρόμηση επιτρέποντας στο γραμμικό μοντέλο να σχετίζεται με τη μεταβλητή απόκρισης μέσω μιας συνάρτησης ζεύξης και επιτρέποντας στο μέγεθος της διακύμανσης κάθε μέτρησης να είναι συνάρτηση της προβλεπόμενης τιμής. Το GLM εφαρμόζεται σε περιπτώσεις εξαρτημένων μεταβλητών με αυθαίρετες κατανομές ή άλλες κατανομές όπως Poisson ή Bernoulli. Το GLM αποτελείται από ένα γραμμικό προγνώστη (linear predictor) η_i

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} \quad (32)$$

και δύο συναρτήσεις, μια συνάρτηση σύνδεσης (link function) που περιγράφει πως ο μέσος όρος $E(y_i) = \mu_i$ εξαρτάται από το γραμμικό προγνώστη, δηλαδή $g(\mu_i) = \eta_i$ και μια συνάρτηση διακύμανσης (variance function) που περιγράφει πως η διακύμανσης $\mathrm{var}(Y_i)$ εξαρτάται από το μέσο, δηλαδή $\mathrm{var}(Y_i) = \varphi V(\mu)$ όπου ο παράγοντας διασποράς (dispersion factor) φ είναι μια σταθερά. Δηλαδή το GLM αποτελείται από τρία στοιχεία, α) μια τυχαία συνιστώσα που αναφέρεται αναφέρεται στην κατανομή πιθανότητας της εξαρτημένης μεταβλητής π.χ. κανονική κατανομή στη γραμμική παλινδρόμηση ή διωνυμική κατανομή στη δυαδική λογιστική παλινδρόμηση (ονομάζεται επίσης μοντέλο θορύβου ή μοντέλο σφάλματος), β) μια συστηματική συνιστώσα που είναι ο γραμμικός προγνώστης και γ) μια συνάρτηση σύνδεσης που καθορίζει τη σύνδεση μεταξύ τυχαίων και συστηματικών στοιχείων (αναφέρει πώς η αναμενόμενη τιμή της εξαρτημένης μεταβλητής σχετίζεται με το γραμμικό πρόβλεψη των επεξηγηματικών μεταβλητών) [14]-[15].

Logistic Regression

Η λογιστική παλινδρόμηση (logistic regression) χρησιμοποιείται για να μοντελοποιήσει την πιθανότητα μιας συγκεκριμένης κατηγορίας. Στη γραμμική παλινδρόμηση, η εξαρτημένη μεταβλητή λαμβάνει αριθμητικές τιμές. Η δυαδική λογική παλινδρόμηση είναι ένας ειδικός τύπος παλινδρόμησης όπου η δυαδική εξαρτημένη μεταβλητή σχετίζεται με ένα σύνολο επεξηγηματικών μεταβλητών, οι οποίες μπορούν να είναι διακριτές ή/και συνεχείς. Το σημαντικό σημείο που πρέπει να σημειωθεί εδώ είναι ότι στην γραμμική παλινδρόμηση, οι αναμενόμενες τιμές της μεταβλητής απόκρισης διαμορφώνονται με βάση τον συνδυασμό των τιμών που λαμβάνονται από τις ανεξάρτητες μεταβλητές. Η λογιστική παλινδρόμηση εφαρμόζεται, για παράδειγμα, όταν θέλουμε να μοντελοποιήσουμε τις πιθανότητες μιας μεταβλητής απόκρισης ως συνάρτηση ορισμένων επεξηγηματικών μεταβλητών, π.χ. «επιτυχία» σε ένα διαγωνισμό ή όταν θέλουμε να εκτελέσουμε περιγραφικές αναλύσεις διακρίσεων. Επίσης στην περίπτωση που θέλουμε να προβλέψουμε τις πιθανότητες ότι τα άτομα εμπίπτουν σε δύο κατηγορίες της δυαδικής απόκρισης ως συνάρτηση ορισμένων επεξηγηματικών μεταβλητών. Στη δυαδική λογιστική παλινδρόμηση, το αποτέλεσμα συνήθως κωδικοποιείται ως «0» ή «1». Εάν ένα συγκεκριμένο παρατηρούμενο αποτέλεσμα για τη εξαρτημένη μεταβλητή είναι το αξιοσημείωτο δυνατό αποτέλεσμα (αναφέρεται ως «επιτυχία» ή «περίπτωση»), συνήθως κωδικοποιείται ως «1» και το αντίθετο αποτέλεσμα (που αναφέρεται ως «αποτυχία» ή «μη περίπτωση») ως «0». Στη λογιστική παλινδρόμηση, οι πιθανότητες της εξαρτημένης

μεταβλητής που λαμβάνουν μια συγκεκριμένη τιμή διαμορφώνονται με βάση το συνδυασμό των τιμών που λαμβάνονται από τις ανεξάρτητες μεταβλητές. Έστω p η πιθανότητα για μια εξαρτημένη δυαδική μεταβλητή Y να λάβει την τιμή 1. Η πιθανότητα να λάβει την τιμή 0 είναι $1-p$. Θεωρούμε μια γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών x_j, $j=1,2,...,p$ και του συμβάντος $Y=1$, δηλαδή

$$l=\log_b \frac{p}{1-p}=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi} \quad (33)$$

όπου l είναι η λογαριθμική απόδοση (log-odds) και b είναι η βάση του λογαρίθμου. Έχουμε

$$\frac{p}{1-p}=b^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi}} \quad (34)$$

Η πιθανότητα του συμβάντος $Y=1$, είναι

$$p=\frac{1}{1+b^{-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi})}} \quad (35)$$

Σύμφωνα με την παραπάνω σχέση, εάν είναι γνωστά οι συντελεστές $\beta_0,\beta_1,\beta_2,...,\beta_p$ μπορεί να υπολογιστεί η πιθανότητα $Y=1$. [14]-[15]

Decision Tree

Ένα δένδρο είναι μια γενική δομή δεδομένων ως ακυκλικός γράφος. Σε ένα τέτοιο δένδρο δεν υπάρχουν κυκλικές διαδρομες και κάθε δύο κόμβοι συνδέονται. Δηλαδή, για κάθε δύο κόμβους υπάρχει μια μοναδική διαδρομή. Ένα δένδρο είναι ένας γράφος $T=(V,E)$ όπου V είναι ένα πεπερασμένο σύνολο κόμβων και $E\subset\{\{u,v\}:u,v\in V\}$ είναι οι ακμές του δένδρου. Ισχύουν τα εξής:

- Υπάρχει ένας διακεκριμένος κόμβος $v\in V$, ονομάζεται η ρίζα του δέντρου.
- Όλοι οι άλλοι κόμβοι (εκτός από τη ρίζα) είναι χωρισμένοι σε k ή λιγότερα αμοιβαίως αποκλειόμενα ζευγάρια υποσυνόλων.
- Κάθε τέτοιο υποσύνολο είναι ένα δέντρο με τη δική του ρίζα και μπορούμε να το ονομάσουμε υποδιαίρεση.
- Για κάθε δευτερεύουσα διάταξη υπάρχει μια (μη κατευθυνόμενη) ακμή η οποία συνδέει τη ρίζα v και τη ρίζα του υποδένδρου.
- Ο κόμβος v ονομάζεται γονέας της ρίζας του υποδένδρου. Οι ρίζες των υποδένδρων ονομάζονται παιδιά του κόμβου v.

Ένα υποσύνολο με ακριβώς έναν κόμβο δημιουργεί ένα (εκφυλισμένο) δέντρο χωρίς υποκείμενα. Αυτό το δέντρο ονομάζεται φύλλο. Ένα δέντρο ονομάζεται *k*-tree υποθέτοντας ότι τουλάχιστον ένας από τους κόμβους του έχει *k* παιδιά και, φυσικά, κάθε κόμβος του δεν έχει περισσότερα από *k* παιδιά. Προφανώς, ο προαναφερόμενος ορισμός υποδηλώνει ότι ένα δέντρο είναι μη κατευθυνόμενο ακυκλικό-συνδεδεμένο γράφημα. Για κάθε κόμβο, υπάρχει μια μοναδική διαδρομή από τη ρίζα σε αυτό τον κόμβο. Ομοίως, μπορούμε να πούμε ότι κάθε κόμβος, φυσικά εκτός από τη ρίζα του δέντρου, έχει ακριβώς έναν γονέα. Ο

αριθμός των άκρων σε μια διαδρομή ονομάζεται μήκος της διαδρομής. Το ύψος ενός δέντρου είναι το μήκος της μεγαλύτερης διαδρομής από τη ρίζα σε ένα φύλλο.

Ένα δένδρο απόφασης (decision tree) είναι μια δενδροειδή αναπαράσταση όπου:

- Κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων ή των δεδομένων.
- Κάθε κλαδί που εξέρχεται από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο.
- Τα κλαδιά φύλλα αναφέρονται σε μια απόφαση/δράση.

Τα δένδρα απόφασης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών).

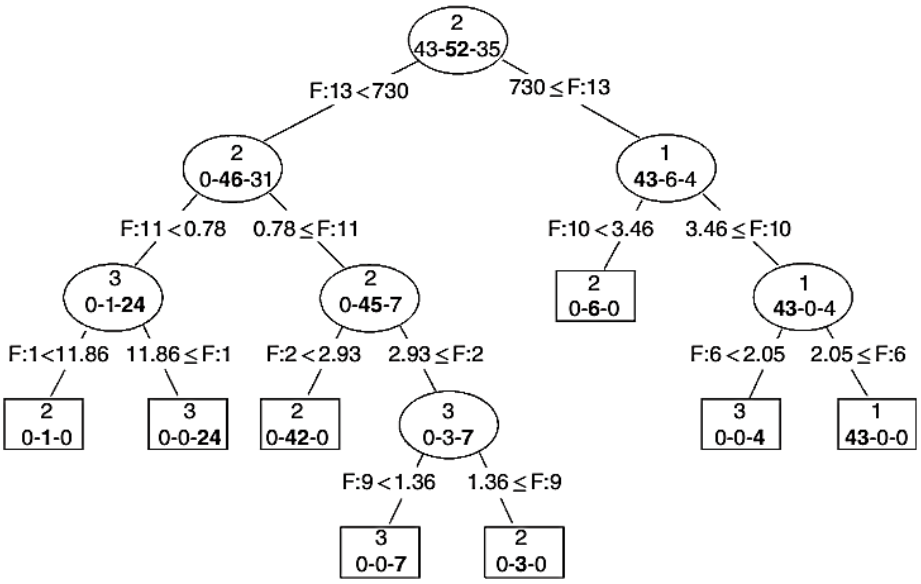
Ένα δέντρο απόφασης είναι ένα δέντρο, το οποίο θα μπορούσε να χρησιμοποιηθεί ως κατηγοριοποιητής. Σε ένα δέντρο αποφάσεων, κάθε ο κόμβος έχει ένα υποσύνολο προτύπων εκπαίδευσης που του έχουν ανατεθεί. Η διαδικασία αντιστοίχισης των κόμβων με τα πρότυπα πραγματοποιείται στο στάδιο της κατασκευής του δένδρου. Η κατασκευή είναι, στην πραγματικότητα, η εκπαίδευση του κατηγοριοποιητη. Κατά τη διάρκεια της κατασκευής, μπορούμε να χρησιμοποιήσουμε γνώσεις σχετικά με τα πρότυπα που έχει οριστεί σε έναν δεδομένο κόμβο για να σχηματίσει μια ετικέτα για τον προσδιορισμό αυτού του κόμβου. Στην ρίζα έχει κατανεμηθεί όλο το σύνολο εκπαίδευσης. Αν ένας κόμβος v δεν είναι φύλλο, τότε το σύνολο των προτύπων εκπαίδευσης χωρίζεται σε υποσύνολα, τα οποία εκχωρούνται σε παιδιά αυτού του κόμβου (ταυτόχρονα διατηρείται ένας γονέας διατηρεί όλα τα πρότυπα που έχουν ανατεθεί στα παιδιά του). Το σύνολο χωρίζεται σε υποσύνολα με τη χρήση του ενός επιλεγμένου χαρακτηριστικού. Η διαδικασία διαίρεσης ολοκληρώνεται όταν υπάρχει κάποια κατάσταση τερματισμού που έχει ικανοποιηθεί. Μία από αυτές τις συνθήκες είναι όταν όλα τα πρότυπα που έχουν εκχωρηθεί σε ένα κόμβο ανήκουν στην ίδια κλάση. Σε αυτήν την περίπτωση, όταν ανήκουν όλα τα πρότυπα στον κόμβο μια μόνη κλάση, ο κόμβος αυτός γίνεται ένα φύλλο και φέρει την ετικέτα κατηγορίας αυτών των προτύπων. Έχοντας κατασκευάσει ένα τέτοιο δέντρο, μπορούμε να το χρησιμοποιήσουμε για την κατηγοριοποίησης νέων προτύπων. Ένα νέο πρότυπο μετακινείται από τη ρίζα σε ένα φύλλο σύμφωνα με τις τιμές των χαρακτηριστικών του, δηλαδή, για κάθε κόμβο το πρότυπο θα μετακινηθεί σε ένα παιδί σύμφωνα με την τιμή ενός χαρακτηριστικού που χρησιμοποιείται για τον διαχωρισμό αυτού του κόμβου.

Για την κατασκευή ενός δένδρου απόφασης, τίθενται τα παρακάτω ερωτήματα προς διερεύνηση:

- Πώς επιλέγουμε ένα χαρακτηριστικό σε έναν συγκεκριμένο κόμβο για να πραγματοποιήσουμε μια διαίρεση;
- Πώς μπορούμε να βρούμε τον αριθμό των υποσυνόλων για το δεδομένο χαρακτηριστικό σε έναν κόμβο, δηλαδή τον αριθμός παιδιών του συγκεκριμένου κόμβου;
- Πότε πρέπει να σταματήσουμε τη διαδικασία διαίρεσης το σύνολο των προτύπων εκπαίδευσης;
- Ποια ετικέτα κλάσης πρέπει να αντιστοιχιστεί σε ένα δεδομένο φύλλο, ειδικά σε ποια ετικέτα κατηγορίας πρέπει να εκχωρείται όταν ένα φύλλο περιέχει πρότυπα που προέρχονται από περισσότερες της μίας κλάσης;

Η απάντηση στην πρώτη ερώτηση δεν είναι προφανής. Διαισθητικά, θα ήταν λογικό να επιλεχτεί ένα τέτοιο χαρακτηριστικό, το οποίο προσφέρει έναν καλό διαχωρισμό μεταξύ των κλάσεων. Η επιλογή ενός χαρακτηριστικού σχετίζεται με το δεύτερο ερώτημα, δηλαδή με τον τρόπο εύρεσης του αριθμού των υποσυνόλων για το δεδομένο χαρακτηριστικό, δηλαδή με το πόσα παιδιά θα πρέπει να αντιστοιχιστούν στο γονέα. Η ιδανική λύση θα ήταν να επιλεχτεί
ένα χαρακτηριστικό, το οποίο επιτρέπει τον διαχωρισμό σε υποσύνολα με τρόπο που το καθένα υποσύνολο περιέχει πρότυπα από μία μόνο κατηγορία. Σε μια τέτοια περίπτωση, μια απαραίτητη (αλλά σε ορισμένες περιπτώσεις ανεπαρκής) κατάσταση, που απαιτείται για να χωρίσει ένα χαρακτηριστικό έτσι ώστε το αποτέλεσμα να παρέχει πλήρη κατηγοριοποίηση, είναι να έχουμε αριθμό των υποσυνόλων ίσο με τον αριθμός των κλάσεων, έστω *C*. Σε μια τέτοια περίπτωση, λαμβάνουμε ένα υποδένδρο ύψους ίσο με 1, όπου κάθε ένα από τα *C* παιδιά του ανήκουν σε μία κλάση. Να σημειωθεί ότι για ορισμένα χαρακτηριστικά μπορεί να συμβεί το γεγονός ότι τα πρότυπα που ανήκουν σε διαφορετικές κλάσεις θα έχουν την ίδια τιμή του χαρακτηριστικού. Δεδομένου ότι συνήθως κατασκευάζουμε ένα μοντέλο κατηγοριοποίησης με περισσότερα από ένα χαρακτηριστικά, συνήθως λαμβάνεται ο αριθμός των υποσυνόλων μικρότερος από τον αριθμό των κλάσεων. Εάν γίνει διάσπαση για λιγότερα υποσύνολα από τον αριθμό των κλάσεων που αντιπροσωπεύονται στο σύνολο των προτύπων, τότε κάποια υποσύνολα θα περιλαμβάνουν πρότυπα που ανήκουν σε περισσότερες από μία κλάσεις. Στη συνέχεια, αυτά τα υποσύνολα θα υποβάλλονται αναδρομικά σε περαιτέρω διαχωρισμό. Η επανάληψη αυτής της διαδικασίας τελικά οδηγεί στο σχηματισμό υποσυνόλων, συμπεριλαμβανομένων προτύπων μόνο από μία κατηγορία. Παρόλα αυτά, δεν είναι εφικτή η λύση που αναφέρεται στο ότι κάθε υποσύνολο που έχει εκχωρηθεί σε ένα φύλλο θα περιέχει πρότυπα από μία μόνο κλάση. Συνεπώς, αναζητούνται άλλες σχεδόν βέλτιστες λύσεις.

Θεωρούμε ως παράδειγμα το σύνολο δεδομένων wine που είναι διαθέσιμο από το UCI Machine Learning Repository. Το σύνολο δεδομένων $O = \{o_1, o_2, ..., o_{178}\} = O_1 \cup O_2 \cup O_3$ περιλαμβάνει 178 πρότυπα από 3 κλάσεις που χαρακτηρίζονται από 13 χαρακτηριστικά. Μια αρχική επισκόπηση των δεδομένων αποκάλυψε ότι ένα χαρακτηριστικό σχετιζόνταν με άλλα δύο και συνεπώς δεν λήφθηκε υπόψη (το 7^{ο} χαρακτηριστικό). Το σύνολο διαχωρίστηκε σε σύνολο εκπαίδευσης και ελέγχου με αναλογία 7:3. Το σύνολο εκπαίδευσης περιελάμβανε τα πρότυπα 43, 52 και 35 στις κλάσεις O_1, O_2 και O_3, αντίστοιχα. Το δένδρο που κατασκευάστηκε φαίνεται στο Σχήμα 4.



Σχήμα 4: Το δέντρο αποφάσεων που δημιουργήθηκε από το σύνολο wine.

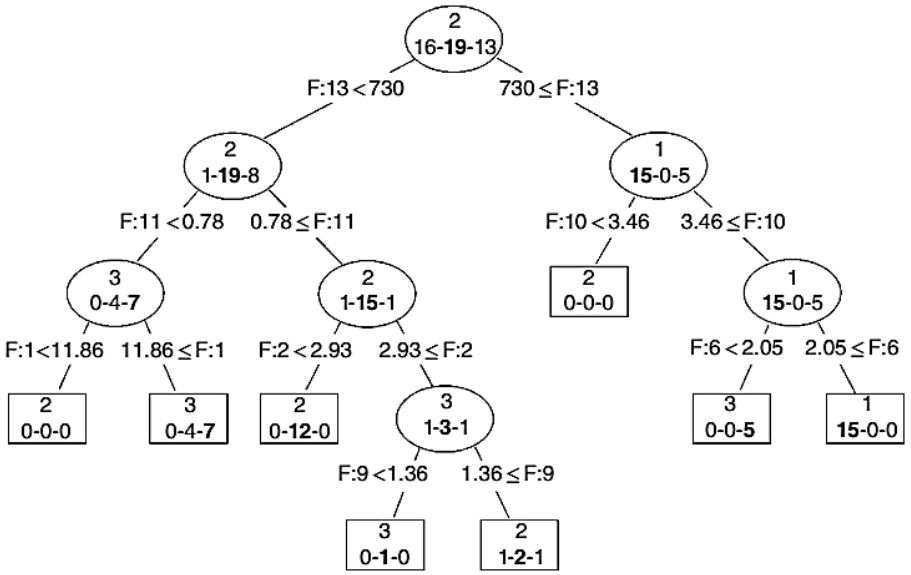
Οι κόμβοι απεικονίζονται με ελλείψεις (μη φύλλων) και ορθογώνια (κόμβοι φύλλων). Ο συμβολισμός x-y-z (π.χ., 43-52-35 στη ρίζα) ενημερώνει για τον αριθμό των προτύπων από τις κατηγορίες O_1, O_2 και O_3, αντίστοιχα, που έχουν εκχωρηθεί σε έναν δεδομένο κόμβο. Ένας μοναδικός αριθμός εντός κόμβων δέντρου, τοποθετημένος πάνω από το x-y-z, ενημερώνει για την ετικέτα της κλάσης πλειοψηφίας σε έναν κόμβο. Η κλάση πλειοψηφίας είναι επίσης που υποδεικνύεται με έναν αριθμό προτύπων με έντονη γραφή, για παράδειγμα, στη ρίζα έχουμε 43-**52**-35 που αναφέρει ότι η δεύτερη τάξη είναι η τάξη πλειοψηφίας σε αυτόν τον κόμβο. Ανισότητες στις άκρες δείχνουν τη συνθήκη διάσπασης του συνόλου των προτύπων του γοένα και το χαρακτηριστικό που χρησιμοποιήθηκε για να γίνει η διάσπαση [13].

Το δένδρο που κατασκευάστηκε έχει τις εξής ιδιότητες:

- Το σύνολο εκπαίδευσης 43-52-35 καταχωρύθηκε στη ρίζα.
- Το 13ο χαρακτηριστικό και η τιμή 730 επιλέχτηκαν για να διαχωρίσουν το σύνολο των προτύπων σε δύο υποσύνολα.
- Αμφότερα τα υποσύνολα, που περιέχουν τα πρότυπα 0-46-31 και 43-6-4 καταχορύθηκαν στα παιδιά της ρίζας.
- Το 11ο χαρακτηριστικό και η τιμή 0.78 θεωρήθηκαν για να διαχωρίσουν το σύνολο προτύπων 0-46-31 σε δύο υποσύνολα.
- Το 10ο χαρακτηριστικό και η τιμή 3.46 θεωρήθηκαν για να διαχωρίσουν το σύνολο προτύπων 43-6-4 σε δύο υποσύνολα.

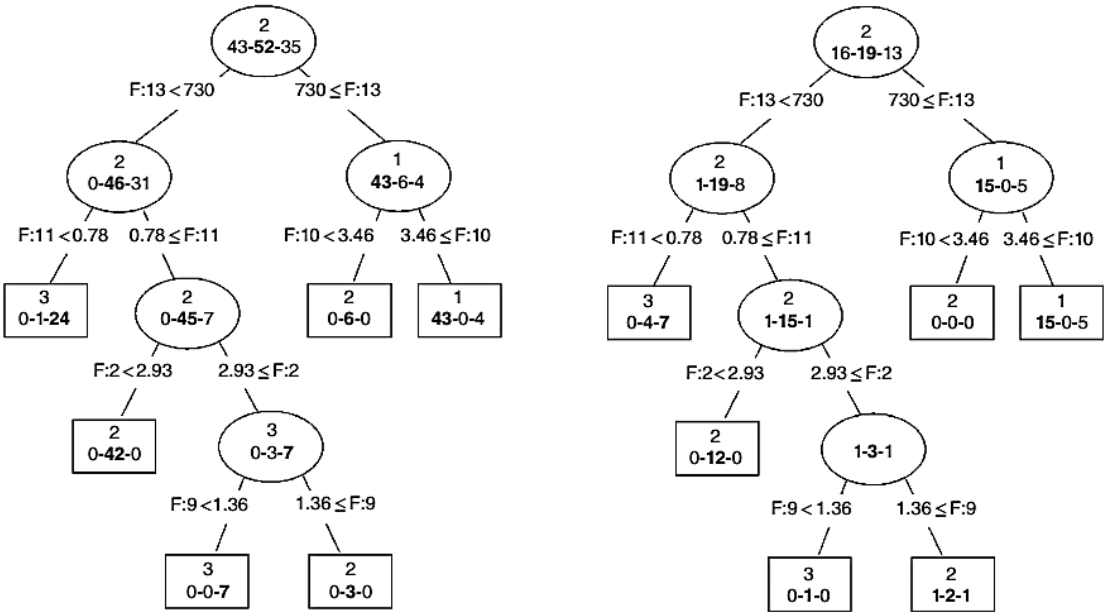
Η διαδικασία διαχωρισμού σταμάτησε όταν ένα σύνολο προτύπων σε έναν κόμβο περιελάμβανε πρότυπα μόνο από μία κατηγορία. Επομένως, κάθε φύλλο αντιπροσωπεύει μία κλάση. Στο δένδρο του Σχήματος 2.4 οι εσωτερικοί κόμβοι (ελλείψεις) και τα φύλλα (ορθογώνια) περιέχουν δύο σειρές αριθμών. Η δεύτερη σειρά περιέχει τρεις ακέραιους γραμμένους υπό τη μορφή x-y-z, όπου το x δηλώνει πόσα πρότυπα ανήκουν στην κλάση O_1 και έχουν εκχωρυθεί στο συγκεκριμένο κόμβο, το y αναφέρεται την κλάση O_2 και το z αναφέρεται στην κλάση O_3. Επιπλέον, σε κάθε κόμβο, ο μέγιστος ακέραιος μφανίζεται με

έντονους χαρακτήρες και αναφέρεται στην κλάση με τα περισσότερα πρότυπα. Ένας κόμβος που περιέχει πρότυπα που ανήκουν στην ίδια τάξη έγινε ένα φύλλο του δέντρου. Μόλις το δέντρο έχει κατασκευαστεί, κάθε πρότυπο του συνόλου ελέγχου μπορεί να κατηγοριοποιηθεί ως εξής: Αρχικά περνάει μέσα από τη ρίζα και φτάνει σε ένα φύλλο. Στο σύλλο παρατηρούμε ποια τάξη περιέχει τα περισσότερα πρότυπα και θεωρείται ότι το συγκεκριμένο πρότυπο ανήκει σε αυτή. Για το σύνολο ελέγχου προκύπτει το δένδρο του Σχήματος 5.



Σχήμα 5: Το δέντρο αποφάσεων για το σύνολο ελέγχου [13].

Η δομή του δένδρου διαμορφώθηκε κατά την εκπαίδευση. Για το σύνολο ελέγχου, η δομή δεν αλλάζει παρά μόνο η σύνθεση των κλάσεων. Παρατηρούμε ότι τα πρότυπα ελέγχου δεν κατηγοριοποίηθηκαν με βέλτιστο τρόπο. Συγκεκριμένα, 3 από τα 48 πρότυπα έχουν ταξινομηθεί εσφαλμένα. Επίσης, δεν χρησιμοποιήθηκαν όλα τα μέρη του δένδρου. Τέσσερα φύλλα και ένας εσωτερικός κόμβος δεν συμμετείχαν στη διαδικασία κατηγοριοποίησης του συνόλου ελέγχου. Επιπλέον, υπάρχει ένας εσωτερικός κόμβος και ένα φύλλο, όπου η πλειονότητα των προτύπων ελέγχου δεν ταιριάζει με τις αναμενόμενες ταξινομήσεις που ανακαλύφθηκαν από την εκπαίδευση. Ως εκ τούτου, μπορούμε να συμπεράνουμε ότι η δομή του δέντρου θα μπορούσε να απλοποιηθεί με τη διαγραφή ορισμένων κόμβων. Στο Σχήμα 6 παρουσιάζονται δύο δένδρα, όπου το αριστερό αναφέρεται στην εκπαίδευση και το δεξιό στον έλεγχο.



Σχήμα 6: Απλοποιημένα δένδρα αποφάσεων [13].

Τα απλοποιημένα δένδρα ελήφθησαν με τη διαδικασία του κλαδέματος (pruning) κάποιων κόμβων με αποτέλεσματα κάποιοι εσωτερικοί κόμβοι γίνονται φύλλα. Ο κανόνας που ακολουθήθηκε για το κλάδεμα είναι ότι ο εσωτερικός κόμβος που γίνεται φύλλο περιέχει λιγότερο από 10% λαθασμένων ταξινομημένων προτύπων.

Θεωρούμε ένα σύνολο προτύπων $O = \{o_1, o_2, ..., o_N\}$ που έχει διαχωριστεί σε C κλάσεις $O_1, O_2, ..., O_C$. Η κατανομή πιθανότητας p_k της κλάσης O_k στο σύνολο των κλάσεων υπολογίζεται ως

$$p_k = \frac{|O_k|}{|O|}, k = 1, 2, ..., C \tag{36}$$

Ο δείκτης $\hat{k}$ εκφράζει την πιο συχνή κλάση στο σύνολο O:

$$\hat{k} = \arg\max_k p_k \tag{37}$$

Η διαφορετικότητα (diversity) τψν κλάσεψων σε ένα σύνολο προτύπων είναι ένα μέτρο που λαμβάνει χαμηλές τιμές όταν τα περισσότερα πρότυπα προέρχονται από μία κλάση. Το πεδίο τιμών του είναι [0,1]. Τιμές κοντά στο 0 αντιστοιχούν στην περίπτωση που κάθε πρότυπο ανήκει στην ίδια κλάση, δηλαδή $|O_{\hat{k}}| = |O|$ και $|O_1| = ... = |O_{\hat{k}-1}| = |O_{\hat{k}+1}| = ... = |O_C| = 0$ ή ισοδύναμα $p_{\hat{k}} = 1$ και $p_1 = ... = p_{\hat{k}-1} = p_{\hat{k}+1} = ... = p_C = 0.$ Εάν λάβει τιμές κοντά στο 1, αναφέρεται στην περίπτωση που όλες οι κλάσεις αντιπροσωπεύονται ίσα, δηλαδή $|O_1| = |O_2| = ... = |O_C|$ ή ισοδύναμα $p_1 = p_2 = ... = p_C$.

Ο Index of Incorrect Classification $R(O)$ αποτιμάει τη διαφορετικότητα και εκφράζει το λόγο του αριθμού των προτύπων που δεν ανήκουν στην κλάση με τα περισσότερα μέλη προς τον αριθμό όλων των προτύπων:

$$R(O) = p_1 + ... + p_{\hat{k}-1} + p_{\hat{k}+1} + ... + p_C = 1 - p_{\hat{k}} \quad (38)$$

Ο δείκτης είναι ίσος με 0 όταν όλα τα πρότυπα ανήκουν στην ίδια κλάση και είναι ίσος με 1-1/*C* όταν όλα τα πρότυπα κατανέμονται ίσα σε όλες τις κλάσεις.

Αναφορικά με τη διαφορετικότητα (διασπορά) της πληροφορίας που υπάρχει σε ένα διαχωρισμό είναι η εντροπία (entropy). Για παράδειγμα, στο σύνολο δεδομένων wine όπου κάθε χαρακτηριστικό είναι πραγματική τιμή, υπολογίζουμε όλους τους πιθανούς διαχωρισμούς ενός χαρακτηριστικού και υπολογίζουμε την εντροπία. Επιλέγεται ο διαχωρισμός με την μικρότερη εντροπία δηλαδή αυτός με τη μικρότερη διασπορά της πληροφορίας. Η εντροπία της πληροφορίας σε ένα σύνολο προτύπων $O = \{o_1, o_2, ..., o_N\}$ είναιν ισοδύναμη με την εντροπία μιας τυχαίας μεταβλητής όπως ορίζεται από την (36), δηλαδή από την κατανομή των πιθανοτήτων σε ένα σύνολο κλάσεων. Η εντροπία $E(O)$ ορίζεται ως

$$E(O) = -\sum_{i=1}^{C} \frac{|O_i|}{|O|} \log_2 \frac{|O_i|}{|O|} = -\sum_{i=1}^{C} p_i \log_2 p_i \quad (39)$$

Από τη στιγμή που ο όρος $\log_2 0$ δεν ορίζεται, θεωρούμε ότι $\log_2 0 = \lim_{p \to 0^+} \log_2 p = 0$. Όσο μεγαλύτερη είναι η εντροπία τόσο, τόσο μεγαλύτερη είναι διαφορετικά των δεδομένων. Τιμές κοντά στο 0 σημαίνει οτι όλα τα πρότυπα ανήκουν στην ίδια κλάση. Το σύνολο τιμών είναι [0,1]. [13]

Random Forest

Μια κατηγορία κατηγοριοποιητών είναι οι σύνθετοι κατηγοριοποιητές (ensemble classifiers) όπου γίνεται συνδυασμός διάφορων κατηγοριοποιητών. Δημιουργείται ένας μεγάλος αριθμός απλών κατηγοριοποιητών που όλοι εφαρμόζονται στο ίδιο πρόβλημα και παρέχει ο κάθε ένας από αυτούς μία λύση. Η τελική κατηγοριοποίηση προκύπτει από το συνδυασμό των αποτελεσμάτων κάθε επιμέρους κατηγοριοποιητή. Ο κατηγοριοποιητής random forest θεωρεί τα δένδρα αποφάσεων ως τους επιμέρους κατηγοριοποιητές. Συνεπώς η απόδοση του random forest εξαρτάται από την απόδοση κάθε επιμέρους κατηγοριοποιητή. Ο αλγόριθμος random forest είναι ο εξής:

Δεδομένα: Σύνολο προτύπων εκπαίδευσης $O = \{o_1, o_2, ..., o_N\}$

Σύνολο χαρακτηριστικών $F = \{f_1, f_2, ..., f_M\}$

B - ο αριθμός των επιμέρους κατηγοριοποιητών

p - ο αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν για να κατασκευάσουν τους επιμέρους κατηγοριοποιητές

Αλγόριθμος: **θεώρησε** τα βάρη $w_i = 1/N, i = 1, 2, ..., N$

Για $k = 1$ έως *B* **τότε**

ξεκίνα

κατασκεύασε ένα σύνολο εκπαίδευσης $O^{(k)}$ με τυχαία δειγματοληψία με αντικατάσταση του συνόλου εκπαίδευσης $O = \{o_1, o_2, ..., o_N\}$, χρησιμοποίησε τα βάρη w_i για δειγματοληψία

κατασκεύασε ένα δένδρο απόφασης για $\psi^{(k)}$ για το σύνολο εκπαίδευσης $O^{(k)}$ **ως ακολούθως:**

ξεκίνα

για κάθε κόμβο του κατασκευασμένου δένδρου **κάνε**

εάν ο κόμβος δεν είναι φύλλο **κάνε**

ξεκίνα

επέλεξε p χαρακτηριστικά από το σύνολο F όλων των χαρακτηριστικών με τυχαία δειγματοληψία χωρίς αντικατάσταση, θεώρησε $p << M$

βρες το καλύτερο χαρακτηριστικό από τα επιλεχθέντα p

βρες το καλύτερο διαχωρισμό από το σύνολο εκπαίδευσης

χρησιμοποίησε κάθε υποσύνολο του διαχωρισμό για να κατασκευάσεις τον αντίστοιχο κόμβο του δένδρου

τέλος

τέλος

τέλος

Αποτέλεσμα: το σύνολο των δένδρων απόφασης $\psi^{(1)}, \psi^{(2)}, ..., \psi^{(B)}$

Για ένα δεδομένο πρότυπο x η ανάθεση σε κλάση γίνεται ως

$$\psi(x) = \arg\max_{1 \le k \le C} \sum_{i=1}^{B} \delta_{k, \psi^{(i)}(x)} \tag{40}$$

όπου $\delta_{i,j}$ είναι η Kronecker συνάρτηση δέλτα.

Η διαδικασία κατασκευής του δένδρου τοπικά δεν χρησιμοποιεί όλα τα πρότυπα εκπαίδευσης παρά μόνο ένα μικρό υποσύνολο τους. Το υποσύνολο των χαρακτηριστικών που χρησιμοποιούνται για την κατασκευή του δένδρου λαμβάνεται με τυχαία δειγματοληψία *p* χωρίς αντικατάσταση από το σύνολο των χαρακτηριστικών. Η ίδια παράμετρος *p* είναι που χρησιμοποιείται για την κατασκευή όλων των δέντρων αποφάσεων. Για την επιλογή της παραμέτρου έχει προταθεί στη βιβλιογραφία ο εμπειρικός κανόνας $p \approx \sqrt{M}$. Το random forest έχει βρει εφαρμογή σε προβλήματα κατηγοριοποίησης με χιλιάδες χαρακτηριστικά [13].

Gradient Boosted Trees

Για το συνδυασμό επιμέρους κατηγοριοποιητών έχουν προταθεί οι τεχνικές boosting και bagging. Ο όρος bagging είναι το ακρωνύμιο των όρων **b**ootstrap **agg**regat**ing**. Η τεχνική bagging δημιουργεί σύνολα εκπαίδευσης με δειγματοληψία με επανατοποθέτηση. Για την κατηγοριοποίηση μιας νέας παρατήρησης γίνεται ψηφοφορία (συνδυασμός) μεταξύ των

κατηγοριοποιητών, και το πρότυπο ελέγχου τοποθετείται στην κλάση που συγκέντρωσε τις περισσότερες ψήφους. Κατασκευάζονται *B* σύνολα προτύπων εκπαίδευσης $O^{(1)}, O^{(2)}, \ldots, O^{(B)}$ από ένα δεδομένο σύνολο εκπαίδευσης $O = \{o_1, o_2, \ldots, o_N\}$. Για κάθε σύνολο εκπαίδευσης κατασκευάζεται ένας κατηγοριοποιητής. Κάθε σύνολο εκπαίδευσης $O^{(i)}$ δημιουργείται με τυχαία δειγματοληψία με επανατοποθέτηση από το σύνολο O. Η δειγματοληψία βασίζεται σε μια ομοιόμορφη κατανομή πιθανοτητών στο σύνολο O, δηλαδή κάθε πρότυπο μπορεί να ληφθεί με ίση πιθανότητα 1/*N*. Το boosting μπορεί να θεωρηθεί γενίκευση του bagging και υιοθετεί κατανομή πιθανότητας για συνεχή σύνολα εκπαίδευσης. Η πιθανότητα να επιλεχτούν πρότυπα που έχουν κατηγοριοποιηθεί με εσφαλμένο τρόπο από τον τρέχοντα κατηγοριοποιητή αυξάνει εφόσον τέτοια πρότυπα είναι δύσκολο να διδαχτούν. Συνεπώς δίνεται έμφαση στο ρόλο που έχουν κατά την κατασκευή του κατηγοριοποιητή. Αυξάνει λοιπόν η πιθανότητα να επιλεχτούν πρότυπα που είναι δύσκολο να κατηγοριοποιηθούν από τον επόμενο στη σειρά κατηγοριοποιηθεί. Αυτό γίνεται με την εκχώρηση τιμής βάρους στο πρότυπο που είναι δύσκολο να κατηγοριοποιηθεί. Κάθε κατηγοριοποιητής έχει ένα συντελεστή βάρους ο οποίος εκφράζει την ακρίβεια του. Η τελική απόφαση για την κατηγοριοποίηση γίνεται με ψηφοφορία και λαμβάνοντας υπόψη τα βάρη.

Η τεχνική ενίσχυση της κλίσης (gradient boosting) κατασκευάζει ένα κατηγοριοποιητή συνδυάζοντας επιμέρους μέσω ενός αριθμού επαναλήψεων. Έστω $m = 1, 2, \ldots, M$ ο αριθμός των επαναλήψεων της ενίσχυσης της κλίσης και J_m ο αριθμός των φύλλων του δένδρου κατά την $m-$ιοστή επανάληψη. Το δέντρο διαχωρίζει το χώρο εισόδου J_m σε διαχωρισμένες περιοχές $R_{1m}, \ldots, R_{J_m m}$ και προβλέπει μια σταθερή τιμή σε κάθε περιοχή. Η έξοδος $h_m(x)$ του δένδρου για είσοδο x είναι

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} 1_{R_{jm}}(x) \tag{41}$$

όπου b_{jm} είναι η προβλεπόμενη τιμή στην περιοχή R_{jm}. Έστω F_m το μοντέλο του κατηγοριοποιητή. Στη συνέχεια οι συντελεστές b_{jm} πολλαπλασιάζονται με μια τιμή γ_m έτσι ώστε να ελαχιστοποιηθεί η συνάρτηση σφάλματος (δηλ. η τετραγωνική διαφορά μεταξύ της πραγματικής εξόδου του κατηγοριοποιητή και της ιδανικής) και το μοντέλο ανανεώνεται ως

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{41}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x) + \gamma_m h_m(x)) \tag{42}$$

όπου L είναι η συνάρτηση σφάλματος μεταξύ της επιθυμητής (ιδανικής) εξόδου y και της πραγματικής F. [13]

Naive Bayes

Η κατανομή πιθανότητας στο σύνολο των κλάσεων ονομάζεται a priori πιθανότητα ενώ αυτή στο σύνολο των χαρακτηριστικών class conditional πιθανότητα. Στη συνέχεια, για ένα

δεδομένο πρότυπο που χαρακτηρίζεται από τιμές χαρακτηριστικών για κάθε κλάση, υπολογίζουμε την πιθανότητα ότι αυτό το πρότυπο ανήκει σε μια δεδομένη κλάση. Αυτές οι πιθανότητες ονομάζονται a posteriori πιθανότητες. Το πρότυπο αποδίδεται στην κλάση με την υψηλότερη a posteriori πιθανότητα. Αυτός ο συγκεκριμένος ταξινομητής ελαχιστοποιεί την πιθανότητα λανθασμένης ταξινόμησης και που ονομάζεται κατηγοριοποιητής Bayes. Θεωρούμε ότι για ένα πρόβλημα κατηγοριοποίησης, υπάρχει διαθέσιμη πληροφορία για μερικά πρότυπα που ανήκουν σε κάθε κλάση και τα χαρακτηριστικά σε κάθε κλάση. Αυτή η πληροφορία μπορεί να προέρχεται από το σύνολο εκπαίδευσης. Θεωρούμε επίσης ότι είναι γνωστή η πιθανότητα $P(O_i)$ που αναφέρεται ότι ένα δεδομένο πρότυπο ανήκει στην κλάση $O_i, O_i \in \Theta$. Δηλαδή είναι γνωστή η κατανομή πιθανότητας της κλάσης

$$p:\Theta \to [0,1], \sum_{O_k \in \Theta} p(O_k) = 1 \tag{43}$$

όπου p είναι η πυκνότητα πιθανότητας και P η πιθανότητα του συμβάντος. Η κατανομή πιθανότητας της κλάσης ορίζεται για ένα σύνολο κλάσεων. Τιμές της κατανομής είτε είναι γνωστές είτε καθορίζονται από το σύνολο εκπαίδευσης. Σε αυτή την περίπτωση η κατανομή πιθανότητας $p(O_i)$ είναι ίση με την πιθανότητα ότι το δεδομένο πρότυπο ανήκει σε αυτή την κλάση $P(O_i)$. Στη συνέχεια θεωρούμε ότι ένα πρότυπο από την κλάση O_k χαρακτηρίζεται από ένα διάνυσμα X από το χώρο των χαρακτηριστικών R^M, δηλαδή $X \in R^M$. Επιπλέον θεωρούμε ότι είτε γνωρίζουμε τη επί συνθήκη πιθανότητα της κλάσης $P(X|O_k)$ ή μπορούμε να την υπολογίσουμε. Θεωρούμε ότι η πυκνότητα πιθανότητας $p(x|O_k)$ είναι συνεχής συνάρτηση. Αυτή ορίζει την κατανομή πιθανότητας στο χώρο των χαρακτηριστικών για την κλάση O_k. Η a posteriori πιθανότητα δίνεται από την παρακάτω σχέση

$$p(O_k|x) = \frac{p(x|O_k) \cdot P(O_k)}{p(x)} \tag{44}$$

όπου $x \in R^M$, $p(x)$ είναι ο παράγοντας κανονικοποίησης και αναφέρεται στην πυκνότητα πιθανότητας

$$p(x) = \sum_{k=1}^{C} p(x|O_k) \cdot P(O_k) \tag{45}$$

Η κανονικοποίηση εξασφαλίζει ότι όλες το άθροισμα όλων των posteriori πιθανότητων $p(x|O_\kappa)$ είναι ίσο με 1,

$$\sum_{k=1}^{C} p(x|O_k) = 1 \tag{46}$$

Θεωρούμε ότι οι πιθανότητες $P(X|O_k)$ για τις τιμές των χαρακτηριστικών $X \in R^M$ για τις κλάσεις $O_k, k = 1,2,..,C$ είναι γνωστές. Η conditional πιθανότητα $P(O_k|x)$ δίνεται ως

$$P(O_k|x) = \frac{P(x|O_k) \cdot P(O_k)}{P(x)} \quad (47)$$

Ο παράγοντας κανονικοποίησης είναι

$$P(x) = \sum_{k=1}^{C} P(x|O_k) \cdot P(O_k) \quad (48)$$

Η φόρμουλα για τον κατηγοριοποιητή Bayes είναι

$$\Psi(x) = \arg\max_k P(O_k|X) \quad (49)$$

Ο κατηγοριοποιητής Bayes ταξινομεί ένα άγνωστο πρότυπο που χαρακτηρίζεται από το διάνυσμα χαρακτηριστικών X στην κλάση O_k εάν

$$P(O_k|X) > P(O_j|X), j = 1,2,...,C, j \neq k \quad (50)$$

Ο παράγοντας κανοκικοποίησης είναι σταθερός και ανεξάρτητος των κλάσεων. Έχουμε

$$P(O_k|X) \cdot P(O_k) > P(O_j|X) \cdot P(O_j), j = 1,2,...,C, j \neq k \quad (51)$$

Ο κατηγοριοποιητής Bayes χωρίζει το χώρο των χαρακτηριστικών σε περιοχές απόφασης (decision regions) $D^1, D^2, ..., D^C$ που ορίζονται από την (50). Τα όρια μεταξύ των κλάσεων καλούνται περιοχές απόφασης ή όρια απόφασης (decision boundaries).

Εάν ο αριθμός των προτύπων εκπαίδευσης είναι πολύ μεγάλος, τότε ο υπολογισμός της πυκνότητας πιθανοτήτας καθίσταται μη πρακτικός. Αυτό το πρόβλημα είναι γνωστό ως curse of dimensionality. Για να υπερνικηθεί το πρόβλημα curse of dimensionality, θεωρείται ότι οι συντεταγμένες σε μια $M-$διάστατη πυκνότητα πιθανότητας είναι στατιστικά ανεξάρτητες. Αυτή η υπόθεση επιτρέπει την αποσύνθεση της $M-$διάστατης κατανομή πιθανότητας $p: R^M \to [0,1]$ σε ένα γινόμενο μονοδιάστατων κατανομών $p_l: R \to [0,1], l = 1,2,...,M$

$$p(x) = \prod_{l=1}^{M} p_l(x_l) \quad (52)$$

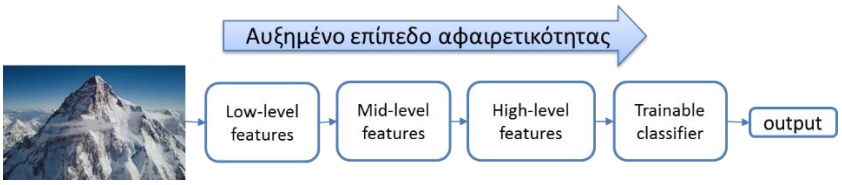
όπου $x = (x_1, x_2, ..., x_M)^T$ και ο δείκτης T δηλώνει αναστροφή του διανύσματος. Μπορούμε να εκτιμήσουμε τη συνάρτηση πυκνότητας σε κάθε διάσταση ξεχωριστά και με αυτό το τρόπο να αποφύγουμε το curse of dimensionality. Ο κατηγοριοποιητής Bayes που χρησιμοποιεί την αποσύνθεση της (52) καλείται naive Bayes [13].

Deep learning

Η Τεχνητή Νοημοσύνη (Artificial Intelligence) αναφέρεται σε οποιαδήποτε τεχνική που επιτρέπει τους υπολογιστές να μιμούνται την ανθρώπινη συμπεριφορά. Η Μηχανική Μάθηση είναι υποσύνολο της τεχνητής νοημοσύνης και αναφέρεται στην ικανότητα για μάθηση διάφορων συστημάτων χωρίς τον εκ των προτέρων αποκλειστικό προγραμματισμό για μάθηση. Η Βαθία Μάθηση (Deep Learning, DL) είναι υποσύνολο της μηχανικής μάθησης και αναφέρεται σε μάθηση υποκειμενικών χαρακτηριστικών των δεδομένων με τη χρήση Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ). Η DL βρίσκει κυρίως εφαρμογή σε προβλήματα

αναγνώρισης προτύπων και πιο συγκεκριμένα εικόνας, ήχου και κειμένου. Σε γενικά πλαίσια, μάθηση είναι είναι η διαδικασία της τροποποίησης της τιμής των βαρών του ΤΝΔ, ώστε δοθέντος συγκεκριμένου διανύσματος εισόδου να παραχθεί συγκεκριμένο διάνυσμά εξόδου. Κατά τη μάθηση βαθμιαία τροποποιούνται οι τιμές των βαρών.

Η DL αποβλέπει στη μάθηση λεπτομερών ιεραρχικών χαρακτηριστικών μέσω μιας διαδικασίας πολλών βημάτων κατά την εκπαίδευσης ΤΝΔ. Στην εικόνα 3 παρουσιάζεται ένα παράδειγμα για αναγνώριση εικόνας.



Εικόνα 3 - Παράδειγμα αναγνώρισης Εικόνας με Deep Learning

Τα χαρακτηριστικά ιεραρχικοποιούνται με αυξημένο επίπεδο αφαιρετικότητας. Κάθε επίπεδο είναι ένας εκπαιδεύσιμος μη γραμμικός μετασχηματισμός των χαρακτηριστικών. Για την αναγνώρισης εικόνας έχουμε Pixel → edge → texton → motif → part → object. Στην περίπτωση αναγνώρισης κειμένου έχουμε Character → word → word group → clause → sentence → story.

Η DL αναφέρεται σε ένα σύνολο αλγορίθμων μηχανικής μάθησης τέτοιων ώστε:

- Αναπτύσσουν ιεραρχική δομή και αναπαράστασης των πρωτογενών και δευτερογενών χαρακτηριστικών, που αντιπροσωπεύουν διαφορετικά επίπεδα αφαίρεσης.
- Χρησιμοποιούν ένα σύνολο πολλών στρωμάτων νευρώνων (ή άλλων μονάδων επεξεργασίας) διαφόρων ειδών για βαθμιαία εξαγωγή χαρακτηριστικών και του μετασχηματισμού τους προκειμένου να επιτευχθεί μια ιεράρχηση δευτερογενών, χαρακτηριστικών που μπορούν να οδηγήσουν σε καλύτερα τελικά αποτελέσματα. Με αυτόν τον τρόπο, προσπαθούν να προσδιορίσουν χαρακτηριστικά υψηλότερου επιπέδου που προέρχονται από χαρακτηριστικά χαμηλότερου επιπέδου.

Οι στρατηγικές DL έχουν την ικανότητα:

- Να ανανεώνουν μόνο ένα επιλεγμένο τμήμα των νευρώνων που ανταποκρίνονται καλύτερα στα δεδομένα εισόδου, ενώ οι άλλοι νευρώνες και οι παράμετροί τους (π.χ. βάρη, κατώφλια) δεν ανανεώνονται.
- Αποφεύγεται η σύνδεση όλων των νευρώνων μεταξύ των διαδοχικών στρωμάτων, οπότε δεν χρησιμοποιείται η στρατηγική συνολικής σύνδεσης που χρησιμοποιείται συνήθως σε πολυστρωματικά perceptron και σε άλλα δίκτυα, αλλά επιτρέπονται οι νευρώνες να ειδικεύονται στην αναγνώριση υπο-προτύπων που μπορούν να εξαχθούν από τα περιορισμένα υποσύνολα των εισόδων.
- Να δημιουργούν συνδέσεις μεταξύ διαφόρων επιπέδων και υπο-δικτύων, όχι μόνο μεταξύ διαδοχικών επιπέδων.
- Να χρησιμοποιούν πολλά υπο-δίκτυα που μπορούν να συνδεθούν με διάφορους τρόπους προκειμένου να επιτρέψουν σε αυτούς τους νευρώνες των υπο-δικτύων να ειδικεύονται στον ορισμό ή την αναγνώριση περιορισμένων υποσυνόλων χαρακτηριστικών ή υποκατηγοριών.

Σε αρχιτεκτονικές DL οι νευρώνες μπορούν να έχουν συνδέσεις εισόδου που να προέρχονται από διαφορετικά στρώματα συνδυάζοντας την έξοδο των νευρώνων που είναι στην είσοδο τους με τη δική τους έξοδο. Μια αρχιτεκτονική μπορεί να περιλαμβάνει αρκετά υποδίκτυα και στρώματα διαφορετικού τύπου. Τα στρώματα υποδειγματολοψίας (subsampling) μπορούν να συνδεθούν με συνελικτικά (convolutional) στρώματα. Σε κάθε στρώματα μπορεί να υπάρχουν πολλά υποδίκτυα του ιδίου τύπου. Οι αρχιτεκτονικές DL διαιρούν την επεξεργασία των νευρώνων σε νέες φάσεις, μερικές είναι βασικές και μερικές δευτερεύουσες. Μπορούν να περιλαμβάνουν συνελικτικά ή συγκεντρωτικά (pooling) στρώματα. Συνήθως το τελευταίο στρώμα είναι επιβλεπόμενης μάθησης και εκπαιδεύεται με αλγόριθμους όπως backpropagation, κτλ. ώστε να συλλέξει και να εξάγει τα δεδομένα.

Τα βασικά ΤΝΔ που βασίζονται σε DL είναι τα Συνεκτικά ΤΝΔ (Convolutional Neural Networks), οι Αυτό-κωδικοποιητές (Autoencoders), τα Restricted Boltzmann Machines και τα Αναδρομικά ΤΝΔ (Recurrent Neural Networks). Τα συνεκτικά ΤΝΔ είναι πολύ-στρωματικά perceptron με τη διαφορά ότι δεν υπάρχουν όλες οι συνδέσεις των προηγούμενων νευρώνων με τους επόμενους. Οι νευρώνες σε κάθε στρώμα συνδέονται μόνο με μια μικρή περιοχή του προηγούμενου στρώματος σε αντίθεση με ΤΝΔ πλήρους σύνδεσης που υπάρχουν συνδέσεις από τους όλους τους νευρώνες του προηγούμενου στρώματος σε όλους τους νευρώνες του επόμενου στρώματος. Οι Αυτοκωδικοποιητές είναι ΤΝΔ που εκπαιδεύονται με μη επιβλεπόμενο τρόπο ώστε να παράγουν έξοδο ίδια με την είσοδο. Αποτελούν δομικά μέρη μεγαλύτερων αρχιτεκτονικών DL. Ο κύριος στόχος μιας τέτοιας εκπαίδευσης είναι να βρεθεί ένα μειωμένος αριθμός νευρώνων σε σχέση με τη διάσταση των δεδομένων εισόδου που να είναι σε θέση να αντιπροσωπεύει τα δεδομένα εισόδου χωρίς παραμόρφωση. Στόχος η κωδικοποίηση των δεδομένων εισόδου και η παρουσίαση τους προς τα επόμενα στρώματα. Δηλαδή εξάγουν τα πιο σημαντικά χαρακτηριστικά των δεδομένων εισόδου. Τα Restricted Boltzmann Machines είναι ΤΝΔ που μαθαίνουν μια κατανομή πιθανότητας πάνω από το σύνολο των εισόδων. Τα Αναδρομικά ΤΝΔ έχουν τη δυνατότητα επιλεκτικά να μεταφέρουν πληροφορία μέσω ακολουθιακών βημάτων, ενώ να γίνεται η επεξεργασία ενός κάθε χρονική ακολουθιακού δεδομένου. Η απλούστερη μορφή ενός πλήρους συνδεδεμένου αναδρομικού δικτύου είναι ένα πολυστρωματικό perceptron όπου σε κάθε κρυμμένο στρώμα εισάγονται ξανά τα δεδομένα εισόδου με καθυστέρηση. Σχηματίζεται μια «μνήμη» προηγούμενων εισόδων που καθορίζουν την εσωτερική κατάσταση του δικτύου και κατ` επέκταση την έξοδο του δικτύου. Μπορούν να χρησιμοποιούν την εσωτερική τους κατάσταση (μνήμη) για να επεξεργάζονται ακολουθίες εισόδων [16].

Fast Large Margin

Ο αλγόριθμος Fast Large Margin βρίσκει μια συνάρτηση απόφασης για τα πρότυπα x διάστασης n που ανήκουν είτε στην κλάση *A* είτε στην κλάση *B*. Η είσοδος στον αλγόριθμο είναι p ζεύγη προτύπων εκπαίδευσης εισόδου-εξόδου (x_i, y_i):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$$

όπου $\begin{cases} y_k = 1, & \varepsilon\acute{\alpha}\nu\ x_k \in A \\ y_k = -1, & \varepsilon\acute{\alpha}\nu\ x_k \in B \end{cases}$ (53)

Από αυτά τα πρότυπα εκπαίδευσης ο αλγόριθμος βρίσκει τις παραμέτρους της συνάρτησης απόφασης $D(x)$ κατά τη διάρκεια της εκπαίδευσης. Μετά την εκπαίδευση η κατηγοριοποίηση ενός νέου προτύπου γίνεται σύμφωνα με τον παρακάτω κανόνα

$$\begin{cases} x \in A, & \varepsilon\acute{\alpha}\nu\ D(x) > 0 \\ x \in B, & \varepsilon\acute{\alpha}\nu\ D(x) \le 0 \end{cases} \tag{54}$$

Η συνάρτηση απόφασης είναι

$$D(x) = \sum_{i=1}^{N} w_i \varphi_i(x) + b \tag{55}$$

Στην προηγούμενη εξίσωση τα $\varphi_i(x)$ είναι οι προκαθορισμένες συναρτήσεις του x και τα w_i και b είναι οι προσαρμόσιμες παράμετροι της συνάρτησης απόφασης. Στο διπλό (dual) χώρο, η συνάρτηση απόφασης είναι

$$D(x) = \sum_{k=1}^{p} a_k K(x_k, x) + b \tag{56}$$

όπου a_k είναι οι παράμετροι που πρέπει να προσαρμοστούν και x_k είναι τα πρότυπα εκπαίδευσης. Η συνάρτηση K είναι ένας προκαθορισμένος πυρήνας. Για παράδειγμα, οι συμμετρικοί πυρήνες είναι της μορφής

$$K(x, x') = \sum_i \varphi_i(x) \cdot \varphi_i(x') \tag{57}$$

Οι παράμετροι w_i καλούνται ευθύς (direct) παράμετροι και δίνονται από την παρακάτω σχέση

$$w_i = \sum_{k=1}^{p} a_k \varphi_i(x_k) \tag{58}$$

Ο αλγόριθμος Fast Large Margin εκπαιδεύει γραμμικούς κατηγοριοποιητές. Αρχικά διαμορφώνεται το περιθώριο (margin) μεταξύ του συνόρου (boundary) της κλάσης και των προτύπων εκπαίδευσης. Η συνάρτηση απόφασης είναι

$$D(x) = w \cdot \varphi(x) + b \tag{59}$$

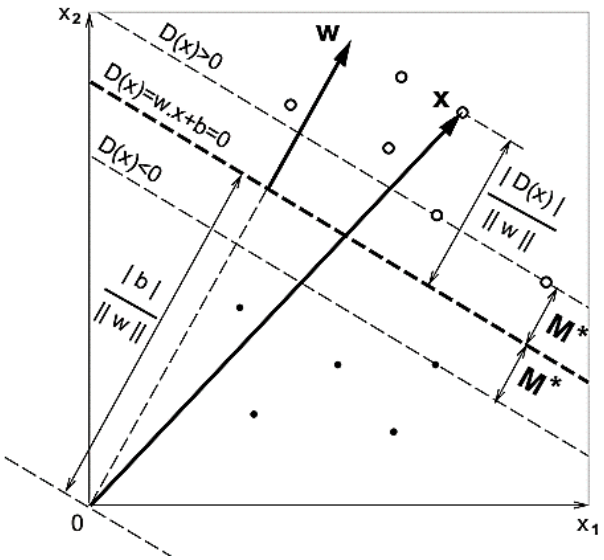
όπου w και $\varphi(x)$ είναι $N-$ διάσταση διανύσματα και b είναι μια παράμετρος πόλωσης. Η συνάρτηση απόφασης ορίζει ένα διαχωριστικό υπερπλάνο στο χώρο φ. Η απόσταση μεταξύ του υπερπλάνου και του προτύπου είναι $D(x)/\|w\|$, όπως φαίνεται στο Σχήμα 8. Θεωρώντας ότι υπάρχει ένας διαχωρισμός μεταξύ του συνόλου εκπαίδευσης με ένα περιθώριο M και του συνόρου της κλάσης, όλα τα πρότυπα εκπληρώνουν την παρακάτω ανισότητα

$$y_k \frac{D(x)}{\|w\|} \ge M \tag{60}$$

Ο αντικειμενικός στόχος του αλγορίθμου είναι να βρει το διάνυσμα παραμέτρων w που μεγιστοποιεί το M

$$M^* = \max_{w,\|w\|=1} M \tag{61}$$

υπό τον περιορισμό $y_k D(x) \geq M$, $k = 1, 2, ..., p$.



Σχήμα 7: Maximum margin γραμμική συνάρτηση απόφασης [17].

Το όριο M^* προσεγγίζεται για όλα εκείνα τα πρότυπα που ικανοποιούν την παρακάτω συνθήκη

$$\min_k y_k D(x_k) = M^* \tag{62}$$

Αυτά τα πρότυπα ονομάζονται υποστηρικτικά (supporting) πρότυπα του ορίου απόφασης. Στο Σχήμα 7 απεικονίζεται η συνάρτηση απόφασης με το μέγιστο περιθώριο. Το πρόβλημα της εύρεσης ενός υπερπλάνο στο χώρο φ χώρο με το μέγιστο περιθώριο ανάγεται σε ένα minimax πρόβλημα

$$\max_{w,\|w\|=1} \min y_k D(x_k) \tag{63}$$

Το γινόμενο μεταξύ των όρων $\|w\|$ και M διαμορφώνεται ως εξής

$$\|w\| \cdot M = 1 \tag{64}$$

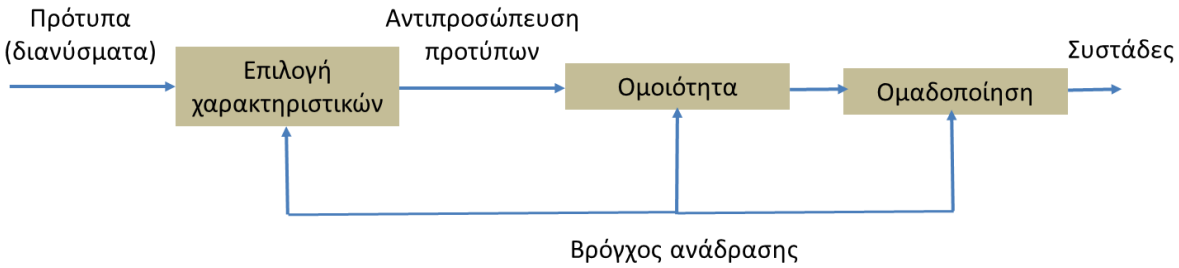
Συνεπώς η μεγιστοποίηση του περιθωρίου M είναι ισοδύναμη με την ελαχιστοποίησης της νόρμας $\|w\|$. Συνεπώς το πρόβλημα της εύρεσης ένα διαχωριστικό υπερπλάνο μεγίστου περιθωρίου w^* αντιστοιχεί στην επίλυση του παρακάτω τετραγωνικού προβλήματος

$$\min_w \|w\|^2 \tag{65}$$

υπό τον περιορισμό $y_k D(x) \geq M$, $k = 1, 2, ..., p$. Το μέγιστο περιθώριο είναι $M^* = 1/\|w^*\|$. [17]

Αλγόριθμος μη εποπτευόμενης μάθησης (Unsupervised Machine Learning Algorithm)

Η συσταδοποίηση είναι μια διαδικασία που χωρίζει τα δεδομένα σε συστάδες (ομάδες, υποσύνολα) σύμφωνα με το βαθμό ομοιότητα τους. Ο βασικός στόχος είναι να προσδιοριστούν οι συνθήκες που δείχνουν πως ορίζονται οι συστάδες. Η συσταδοποίηση είναι ιδανική σε προβλήματα που δεν υπάρχει εκ των προτέρων πληροφορία για τη δομή των δεδομένων. Στο Σχήμα 8 παρουσιάζεται η γενική δομή ενός συστήματος συσταδοποίησης.



Σχήμα 8: Αναπαράσταση λειτουργίας ενός συστήματος συσταδοποίησης.

Ως πρότυπα (patterns) νοούνται τα δεδομένα εισόδου. Συνήθως αναπαρίστανται ως διανύσματα πεπερασμένης διάστασης. Η επιλογή χαρακτηριστικών (feature selection) αναφέρεται στην επιλογή εκείνων των τιμών από τα πρότυπα που θα εισαχθούν σε ένα αλγόριθμο συσταδοποίησης. Η αντιπροσώπευση προτύπων (patterns representation) αναφέρεται στην τεχνική που χρησιμοποιείται για να ετοιμαστούν τα πρότυπα προς συσταδοποίηση. Μπορεί να περιλαμβάνει μετασχηματισμό των δεδομένων (π.χ. από το πεδίο του χρόνου στο πεδίο της συχνότητας) ή/και κανονικοποίηση (normalization). Η ομοιότητα (similarity) περιλαμβάνει την επιλογή ενός μαθηματικού κριτηρίου ή δείκτη αξιολόγησης (clustering validity indicator) που αποτιμά το σφάλμα συσταδοποίησης. Η ομαδοποίηση αναφέρεται στην ίδια την διαδικασία της συσταδοποίησης. Ο βρόγχος ανάδρασης (feedback loop) αναφέρεται στην τελική αξιολόγηση της συσταδοποίησης και στην αναγκαιότητα η όχι να αλλάξουν η αντιπροσώπευση των προτύπων, ο δείκτης αξιολόγησης ή ο ίδιος ο αλγόριθμος συσταδοποίησης. Μερικές βασικές έννοιες της συσταδοποίησης είναι οι παρακάτω:

- Συστάδα (cluster): ομάδα με δεδομένα μεγάλης ομοιότητας
- Κεντροειδές (centroid): ο μέσος όρος των δεδομένων που ανήκουν στην ίδια συστάδα
- Ιδιότητα μέλους συστάδας (cluster membership): αναφέρεται στο ποια συστάδα ανήκει ένα διάνυσμα μετά την ολοκλήρωση της συσταδοποίησης
- Αποστάσεις μεταξύ συστάδων (cluster distances): μπορούν να οριστούν με διαφορετικούς τρόπους, συνήθως αντιστοιχούν στις αποστάσεις μεταξύ των κεντροειδών

Η συσταδοποίηση μπορεί να χρησιμοποιηθεί ως ένα εργαλείο για την εύρεση δομών μέσα σε δεδομένα ή σαν ένα εργαλείο προ-επεξεργασίας των δεδομενών για άλλους αλγορίθμους. Μερικές βασικές εφαρμογές της συσταδοποίησης είναι οι εξής:

- Βιολογία: ταξινόμηση ζωικών οργανισμών (ειδών)
- Ανάκτηση πληροφορίας: συσταδοποίηση εγγράφων
- Χρήση γης: Αναγνώριση περιοχών με όμοια χρήση από βάσεις δεδομένων δορυφορικών εικόνων
- Μάρκετινγ: Δημιουργία προφίλ χρηστών και καταναλωτικών προτύπων
- Σχεδιασμός πόλεων: Αναγνώριση τύπων οικιών, γεωγραφικών περιοχών, κτλ.
- Κλίμα: Εύρεση προτύπων της ατμόσφαιρας, των ωκεανών, κτλ.

Αναφορικά με τα δεδομένα, αυτά μπορούν να διακριθούν στους εξής τύπους:
- Πραγματικές τιμές/συνεχείς μεταβλητές (π.χ. οικονομικά μεγέθη, κατανάλωση φυσικού αερίου, κτλ.)
- Δυαδικές τιμές (π.χ. φύλο)
- Κατηγορίες (π.χ. χρώμα, μήνες, κτλ.)
- Ιεραρχίες (π.χ. διευθυντής, υπο-διευθυντής, κτλ.)
- Μεικτά δεδομένα (π.χ. αναφέρονται σε 2 ή περισσότερες κατηγορίες)

Ο πυρήνας της συσταδοποίησης είναι ο αλγόριθμος συσταδοποίησης (clustering algorithm). Αυτοί μπορούν να διακριθούν στις ακόλουθες κατηγορίες:

1) Διαχωριστικοί αλγόριθμοι (Partitional algorithms). Η λειτουργία τους αποσκοπεί στην εύρεση του βέλτιστου διαχωρισμού των δεδομένων σε ένα προκαθορισμένο αριθμό συστάδων. Αυτό γίνεται μέσω της βελτιστοποίησης αντικειμενικής συνάρτησης μέσω ενός αριθμού πεπερασμένων επαναλήψεων. Βασικοί εκπρόσωποι αυτής της κατηγορίας είναι οι K-Means και K-Medoids.
2) Ιεραρχικοί αλγόριθμοι (Hierarchical algorithms): Η ιεραρχική συσταδοποίηση ομαδοποιεί τα δεδομένα όχι σε ένα βήμα όπως στη διαχωριστική συσταδοποίηση, αλλά μέσω ενός αριθμού βημάτων. Αναφέρεται στη δημιουργία υπο-συστάδων μέσα σε συστάδες οργανωμένες με τη μορφή δένδρου (δενδρόγραμμα). Το δενδρόγραμμα αποτελείται από κόμβους στον κορμό του (που αντιστοιχούν σε ομάδες), από κόμβους στα φύλλα (που αντιστοιχούν στις υπο-συστάδες) και από τη ρίζα (που αντιστοιχεί στο σύνολο των συστάδων). Χαρακτηριστικό γνώρισμα της ιεραρχικής συσταδοποίησης είναι το γεγονός ότι αναθέτει τα πρότυπα στις διάφορες συστάδες οριστικά. Βασικοί εκπρόσωποι αυτής της κατηγορίας είναι οι Single Linkage, Complete Linkage και Ward`s Method.
3) Ασαφείς αλγόριθμοι (Fuzzy algorithms): Έχουν παρόμοια λειτουργία με τους διαχωριστικούς. Κάθε πρότυπο κατανέμεται σε όλες τις συστάδες με βαθμό συμμετοχής (membership degree). Βασικός εκπρόσωπος αυτής της κατηγορίας είναι οι Fuzzy C-Means.
4) Λοιποί αλγόριθμοι που δεν ανήκουν στις άνω κατηγορίες όπως το Self-Organizing Map, DBSCAN, κτλ. [12]

Περιγραφή του αλγορίθμου K-Means

Έστω $X=\{x_i\}, i=1,2,...,n$ είναι ένα σύνολο n-διάστατων αντικειμένων. Σκοπός της λειτουργίας του αλγορίθμου είναι η ομαδοποίηση τους σε *K* συστάδες. Το σύνολο των συστάδων αναφέρεται ως $C=\{c_k\}, k=1,2,...,K$. Κάθε κάθε συστάδα αντιπροσωπεύεται από το κεντροειδές της c_k το οποίο αποτελεί τον αριθμητικό μέσο των αντικειμένων που ανήκουν στη k -ιοστή συστάδα:

$$c_k = \frac{1}{M}\sum_{m=1}^{M} x_m \tag{66}$$

όπου M είναι ο αριθμός των αντικείμενων x_m του αρχικού συνόλου X που ανήκουν στη k -ιοστή συστάδα. Ως τετραγωνικό σφάλμα της *k*-ιοστής συστάδας $J(c_k)$ ορίζεται το άθροισμα όλων των Ευκλείδειων αποστάσεων μεταξύ κάθε αντικειμένου x_i από το κεντροειδές c_k :

$$J(c_k) = \sum_{x_i \in c_k} \left\| x_i - c_k \right\|^2 \tag{67}$$

όπου η νόρμα $\| \ \|$ αναφέρεται στην Ευκλείδεια απόσταση. Ο K-Means μέσω συνεχών επαναλήψεων βρίσκει εκείνη την ομαδοποίηση του έτσι ώστε το συνολικό άθροισμα των τετραγωνικών σφαλμάτων όλων των συστάδων J είναι ελάχιστο:

$$J = \sum_{k=1}^{K} \sum_{x_i \in c_k} \left\| x_i - c_k \right\|^2 \tag{68}$$

Σύμφωνα με την αναφορά [18], η ελαχιστοποίηση του J είναι πρόβλημα τύπου NP-hard ακόμα και για αριθμό συστάδων ίσο με $k=2$. Συνεπώς ο K-Means μπορεί να συγκλίνει μόνο σε τοπικά ελάχιστα. Η λειτουργία του αλγορίθμου είναι η εξής: Ο αλγόριθμος ξεκινάει με τυχαία επιλογή *k* προτύπων που θα τοποθετηθούν ως τα αρχικά κεντροειδή. Στη συνέχεια, τα υπόλοιπα πρότυπα τοποθετούνται στις *k* συστάδες έτσι ώστε το σφάλμα J να ελαχιστοποιηθεί μεταξύ δύο διαδοχικών επαναλήψεων. Ως επανάληψη νοείται μία πλήρης τροφοδότηση όλων των προτύπων στον αλγόριθμο. Σε κάθε επανάληψη, γίνεται εκ νέου υπολογισμός (ανανέωση) των κεντροειδών σύμφωνα με την (66). Εφόσον το σφάλμα J μειώνεται με την αύξηση του αριθμού των συστάδων, θα πρέπει να ελαχιστοποιηθεί μόνο για συγκεκριμένο και σταθερό αριθμό συστάδων. Το διάγραμμα ροής της λειτουργίας του K-Means απεικονίζεται στο Σχήμα 9. Τα βήματα λειτουργίας είναι τα εξής:

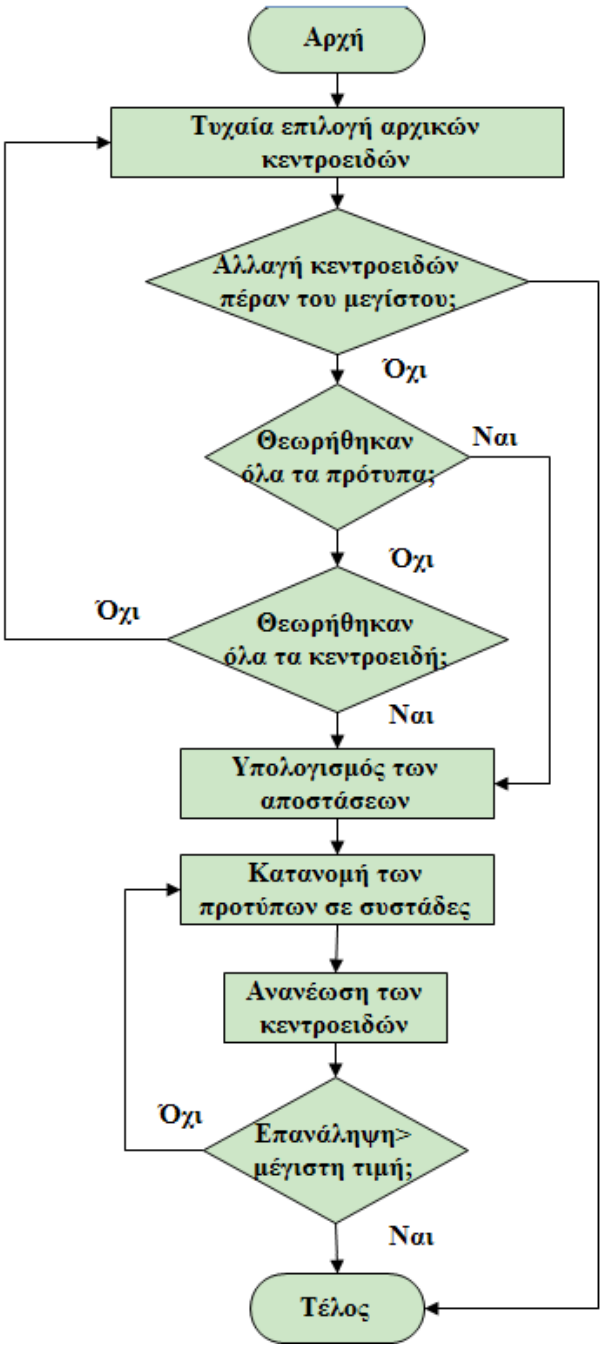
Βήμα 1: Έστω ένα σύνολο δεδομένων περιέχει *N* πρότυπα. Ο χρήστης αποφασίζει για τον τελικό αριθμό συστάδων, έστω *k*.

Βήμα 2: Λαμβάνει χώρα τυχαία επιλογή *k* προτύπων από τα *N* πρότυπα. Αυτά θέτονται ως τα αρχικά κεντροειδή.

Βήμα 3: Για κάθε ένα από τα εναπομείναντα *N-k* πρότυπα υπολογίζονται οι Ευκλείδειες αποστάσεις από τα *k* κεντροειδή. Τα πρότυπα κατανέμονται σε εκείνες τις συστάδες που αντιστοιχούν τα κεντροειδή με τις ελάχιστες τιμές των αποστάσεων.

Βήμα 4: Μόλις υπολογιστούν όλες οι αποστάσεις των *N-k* πρότυπων, δηλ. ολοκληρωθεί μία επανάληψη πραγματοποιείται η ανανέωση των αρχικών κεντροειδών.

Βήμα 5: Τα Βήματα 3 και 4 επαναλαμβάνονται μέχρι να παρέλθει ο μέγιστος αριθμός επαναλήψεων ή ελαχιστοποιηθεί το σφάλμα μεταξύ δύο διαδοχικών επαναλήψεων.



Σχήμα 9: Διάγραμμα λειτουργίας του αλγορίθμου K-Means.

Οι παράμετροι του αλγορίθμου που θα πρέπει να τεθούν από το χρήστη είναι οι εξής:

i) Ο επιθυμητός αριθμός των συστάδων *k*. Θα πρέπει να ισχύει η συνθήκη $2 \le k < N$, δηλ. να είναι μεγαλύτερος ίσος του 2 και μικρότερος του αριθμού των προτύπων.
ii) Ο μέγιστος αριθμός των επαναλήψεων.
iii) Το κατώφλι βελτίωσης του σφάλματος J μεταξύ δύο διαδοχικών επαναλήψεων ε. Έστω t η τρέχουσα επανάληψη. Είναι $J(t+1) - J(t) \le \varepsilon$.

Για την εύρεση των παραπάνω τιμών, οι γενικοί κανόνες είναι οι εξής:

i) Ο αλγόριθμος εκτελείται ξεχωριστά για διαφορετικό αριθμό συστάδων και για κάθε αριθμό σημειώνεται η τιμή του εκάστοτε δείκτη αξιολόγησης. Ο επιθυμητός αριθμός συστάδων είναι εκείνος που αντιστοιχεί στο «γόνατο» της καμπύλης του δείκτη αξιολόγησης.
ii) Συνήθως τίθεται μεγάλος αριθμός επαναλήψεων, π.χ. άνω του 500.
iii) Συνήθως τίθεται πολύ μικρό κατώφλι, π.χ. $\varepsilon = 10^{-6}$.

Ο αλγόριθμος K-Means τερματίζει όταν ισχύει μία από τις εξής συνθήκες: α) Έχει παρέλθει ο μέγιστος αριθμός επαναλήψεων ή β) η τιμή μείωσης του σφάλματος είναι μικρότερη ή ίση με το κατώφλι. Συνεπώς θα πρέπει να επιλέγεται μεγάλος αριθμός επαναλήψεων ώστε να υπάρχει η δυνατότητα για τον αλγόριθμο να εκτελεστεί πολλές φορές και να οδηγήσει σε πολύ μικρή τιμή σφάλματος J.Αξίζει να σημειωθεί πως λόγω της απλότητας, της ταχύτητας εκτέλεσης και της κατανοητής λειτουργίας του, ο αλγόριθμος K-Means έχει βρει εφαρμογή σε πλήθος προβλημάτων και είναι διαθέσιμος σε πακέτα λογισμικού όπως Matlab, Mathematica, WEKA, R, κτλ. [12], [18]

Αποτελέσματα Αλγορίθμων

Σε αυτή την ενότητα θα παρουσιαστούν τα αποτελέσματα των προαναφερθέντων αλγορίθμων . Επιπρόσθετα θα παρουσιαστούν γνωρίσματα απόδοσης αλγορίθμων και θα δοθεί μεγάλη έμφαση στην ακρίβεια των αλγορίθμων η οποία θα αποτελέσει καθοριστικό μέτρο σύγκρισης.

Αλγόριθμοι χωρίς επιλογή χαρακτηριστικών (Χωρίς Forward Selection)

Area Under the Curve

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα για area under the curve για κάθε αλγόριθμο ξεχωριστά

Model	AUC
Deep Learning	92.07%
Generalized Linear Model	91.56%
Gradient Boosted Trees	85.61%
Random Forest	80.46%
Naive Bayes	79.57%
Logistic Regression	73.02%
Decision Tree	65.45%
Fast Large Margin	64.99%
Support Vector Machine	50.00%

Πίνακας 2 - Ποσοστά Περιοχής κάτω από την Καμπύλη ROC

Ακρίβεια (Accuracy)

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα για την ακρίβεια κάθε αλγορίθμου στα δεδομένα *χαρακτηριστική* είναι και η τυπική απόκλιση κάθε αλγορίθομου που εξετάστηκε.

Model	Accuracy	Standard Deviation
Deep Learning	85.38%	0.037
Logistic Regression	80.62%	0.031
Generalized Linear Model	79.81%	0.039
Gradient Boosted Trees	77.52%	0.072
Random Forest	71.67%	0.058
Naive Bayes	70.90%	0.056
Decision Tree	64.71%	0.039
Fast Large Margin	53.90%	0.028
Support Vector Machine	52.95%	0.021

Πίνακας 3 - Ακρίβεια Αλγορίθμων και τυπική απόκλιση

Ευαισθησία (Sensitivity)

Στον παρακάτω πίνακα παρουσιάζεται η ευαισθησία (Sensitivity) για κάθε αλγόριθμο.

Model	Sensitivity
Deep Learning	93.56%
Generalized Linear Model	88.28%
Naive Bayes	86.06%
Decision Tree	84.56%
Gradient Boosted Trees	84.56%
Logistic Regression	83.50%
Random Forest	81.06%
Fast Large Margin	2.00%
Support Vector Machine	0.00%

Πίνακας 4 - Ευαισθησία Αλγορίθμων σε ποσοστιαίες τιμές

Πρόβλεψη με χρήση Forward Selection τεχνικής για τα χαρακτηριστικά εισόδου (ανεξάρτητων μεταβλητών)

Με την χρήση της τεχνικής Forward Selection παρατειρείται μικρή αύξηση στην ακρίβεια αλγορίθμων αλλά μικρότερη διακύμανση ανάμεσα στα χαρακτηριστικά εισόδου (ανεξάρτητες μεταβητές) με αποτέλεσμα να υπάρχει ένα πιο αξιόπιστο αποτέλεσμα.

Η τεχνική αυτή χρησιμοποιεί αλγορίθμους μηχανικής μάθησης έτσι ώστε να αναγνωριστούν ποιά από τα χαρακτηριστικά εισόδου (ανεξάρτητες μεταβλητές) συμβάλουν περισσότερο στην ακριβή και αμερόληπτη πρόβλεψη των εξεταζόμενων αλγορίθμων.

Αυτό το στάδιο ανήκει στην προεπεξεργασία των δεδομένων δηλαδή στο στάδιο αφου έχουν καθαριστεί τα δεδομένα από θόρυβο όπως ακραίες τιμές , μη αναγνώσιμες τιμές εισόδους και άλλα σφάλματα που ενδεχομένως να υπάρχουν στα αρχικά δεδομένα.

Στην συνέχεια ακολουθούν αναλυτικοί ,ανα εξεταζόμενο αλγόριθμο ,πίνακες που αφορούν την ακρίβειά τους και άλλα χαρακτηριστικά υπό την χρήση της forward selection τεχνικής .

Στην πρώτη φάση δοκιμάζεται ο αλγόριθμος **K-Nearest Neighbor** με τα εξής αποτελέσματα.

Model	Accuracy	Standard Deviation
Support Vector Machine	74.71%	6.60%
Generalized Linear Model	74.43%	4.62%
Logistic Regression	74.43%	4.62%
Deep Learning	73.76%	5.76%
Decision Tree	73.52%	9.36%
Naive Bayes	70.62%	3.01%
Random Forest	69.95%	8.31%
Gradient Boosted Trees	69.62%	7.96%
Fast Large Margin	56.33%	7.31%

Πίνακας 5 - Ακρίβεια Αλγορίθμων με χρήση Foward Selection K-nn

Παρατηρείται χαμηλή μείωση απόδοσης της ακρίβειας των αλγορίθμων αλλά και σημαντική ποσοστιαία αύξηση της τυπικής απόκλισης με αποτέλεσμα

Έπειτα γίνεται χρήση του αλγορίθμου **Logistic Regression** που αποδίδει τα εξής :

Model	Accuracy	Standard Deviation
Generalized Linear Model	80.76%	3.39%
Logistic Regression	80.76%	3.39%
Deep Learning	80.57%	0.52%
Naive Bayes	76.71%	3.87%
Gradient Boosted Trees	73.52%	2.66%
Decision Tree	71.57%	7.04%
Random Forest	71.52%	5.46%
Fast Large Margin	71.52%	4.44%
Support Vector Machine	59.14%	6.95%

Πίνακας 6 - Ακρίβεια Αλγορίθμων με χρήση Foward Selection LogReg

Παρατηρείται μικρή αύξηση της ακρίβειας των αλγορίθμων και παρόλο που ο αλγόριθμος Deep Learning έχει μικρότερη απόδοση κατά μερικές δεκαδικές μονάδες , η τυπική του απόκλιση είναι η χαμηλότερη με αποτέλεσμα να έχουμε έναν πιο αποδοτικό αλγόριθμο που αποδίδει καλύτερα συγκριτικά με τους προηγούμενους.

Στην συνέχεια για την τεχνική forward selection χρησιμοποιείται ο αλγόριθμος **Naive bayes** με τα εξής αποτελέσματα:

Model	Accuracy	Standard Deviation
Gradient Boosted Trees	76.52%	5.98%
Naive Bayes	75.67%	3.91%
Deep Learning	74.57%	3.52%
Generalized Linear Model	73.76%	2.87%
Logistic Regression	73.76%	2.87%
Random Forest	73.52%	5.45%
Decision Tree	72.52%	8.74%
Support Vector Machine	61.81%	7.83%
Fast Large Margin	54.90%	3.92%

Πίνακας 7 - Ακρίβεια Αλγορίθμων με χρήση Foward Selection Naive Bayes

Παρατηρείται χειροτέρευση της ακρίβειας των αλγορίθμων και σημαντική ποσοστιαία αύξηση της τυπικής απόκλισης των αλγορίθμων με αποτέλεσμα να μην υπάρχει ουσιαστική βελτίωση στα εξεταζόμενα μοντέλα.

Τέλος, εξετάζεται ο αλγόριθμος **SVM(Support Vector Machine)** και στον παρακάτω πίνακα καταγράφεται η απόδοση του ώς κατηγοριοποιήτης της τεχνικής Forward Selection.

Model	Accuracy	Standard Deviation
Deep Learning	84.48%	4.06%
Fast Large Margin	84.33%	4.02%
Generalized Linear Model	83.57%	5.23%
Logistic Regression	83.57%	5.23%
Gradient Boosted Trees	77.52%	6.08%
Support Vector Machine	72.81%	2.64%
Naive Bayes	72.57%	2.29%
Decision Tree	72.52%	8.74%
Random Forest	71.52%	5.46%

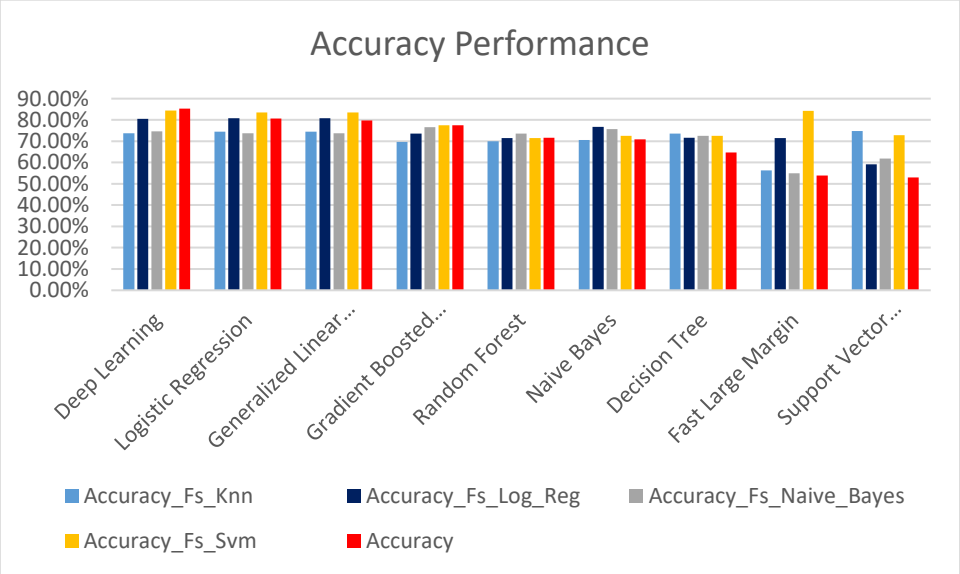
Πίνακας 8 - Ακρίβεια Αλγορίθμων με χρήση Foward Selection SVM

Παρατηρείται σημαντική βελτίωση στην ακρίβεια των αλγορίθμων και παράλληλα χαμηλή αύξηση, έως και μηδενική , της τυπικής απόκλισης , με αποτέλεσμα ο αλγόριθμος Support Vector Machine να αποτελεί σημαντικός αλγόριθμος καθορισμού των ανεξάρτητων μεταβλητών όταν χρησιμοποιείται σε συνδυασμό με την τεχνική Forward Selection.

Σύγκριση Αποτελεσμάτων

Σε αυτή την ενότητα θα παρουσιαστούν τα αποτελέσματα της ακρίβειας των αλγορίθμων καθώς και η τυπική απόκλισή τους.

Σκοπός αυτής της σύγκρισης είναι να βρεθεί ποιος αλγόριθμος έχει μεγαλύτερη ακρίβεια με τα λιγότερα χαρακτηριστικά εισόδου και με την εν δυνάμει λιγότερη τυπική απόκλιση ώστε να αποδίδει καλύτερα κατά την πρόβλεψη.



Γράφημα 7. Συγκεντρωτικός Πίνακας Αποδόσεων Ακρίβειας Αλγορίθμων

Αναφορικά με τον παραπάνω Γράφημα , διακρίνεται ότι ο αλγόριθμος Deep Learning αποδίδει τα μέγιστα χωρίς την χρήση της μεθόδου προεπεξεργασίας Forward Selection με ποσοστό ακρίβειας **85.38%**

Έπειτα ακολουθεί με την χρήση της μεθόδου Forward Selection και με τον αλγόριθμο SVM ως κατηγοριοποιητή των χαρακτηριστικών εισόδου, ο αλγόριθμος Deep Learning αποδίδει 84.48% , Logistic Regression 83.57% , στο ίδιο επίπεδο βρίσκεται και ο Generalized Linear Model με 83.57% και ,με *μεγάλη βελτίωση της ακρίβειάς* του ο Fast Large Margin με 84.33%

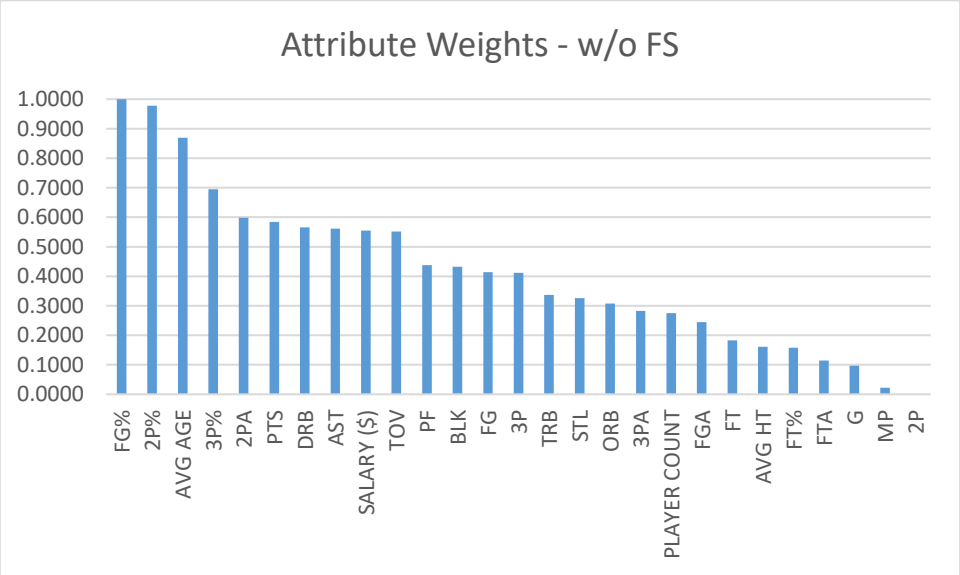
Πίνακας 9 . Αναλυτικός Πίνακας Αποδώσεων Αλγορίθμων (Με κόκκινο οι μέγιστες αποδώσεις)

Model	Accuracy_Fs_Knn	Accuracy_Fs_Log_Reg	Accuracy_Fs_Naive_Bayes	Accuracy_Fs_Svm	Accuracy
Deep Learning	73.76%	80.57%	74.57%	84.48%	85.38%
Logistic Regression	74.43%	80.76%	73.76%	83.57%	80.62%
Generalized Linear Model	74.43%	80.76%	73.76%	83.57%	79.81%
Gradient Boosted Trees	69.62%	73.52%	76.52%	77.52%	77.52%
Random Forest	69.95%	71.52%	73.52%	71.52%	71.67%
Naive Bayes	70.62%	76.71%	75.67%	72.57%	70.90%
Decision Tree	73.52%	71.57%	72.52%	72.52%	64.71%
Fast Large Margin	56.33%	71.52%	54.90%	84.33%	53.90%
Support Vector Machine	74.71%	59.14%	61.81%	72.81%	52.95%

Αναλυτικά στον Πίνακα Αποδόσεων Αλγορίθμων περιγράφονται όλες οι αποδόσεις των αλγορίθμων με και χωρίς την χρήση της Forward Selection τεχνικής.

Συμπεράσματα

Στην εφαρμογή προγνωστικών αλγορίθμων χωρίς την εφαρμογή της τεχνικής forward selection τα βάρη στις μεταβλητές εισόδου κατανέμονται όπως περιγράφονται στο παρακάτω γράφημα.



Γράφημα 8. Κατανομή βαρών σε χαρακτηριστικά εισόδου.(χωρίς Foward Selection)

Τη μεγαλύτερη απόδοση ακρίβειας αλγορίθμου χωρίς την τεχνική forward selection την είχε ο αλγόριθμος Deep Learning με ποσοστό ακρίβειας 85.38% .

Ο Αλγόριθμος deep learning έδωσε μεγαλύτερη βαρύτητα στα εξής χαρακτηριστικά:

Πίνακας 10.Βάρη στα χαρακτηριστικά εισόδου.Deep Learning Algorithm

Attribute	**Weight**
FG%	1.0000
2P%	0.9784
AVG AGE	0.8693
3P%	0.6951
2PA	0.5983
PTS	0.5842
DRB	0.5658
AST	0.5610
SALARY ($)	0.5546

TOV	0.5513

Βιβλιογραφία

[1] M. Rouse, “Data analytics”, διαθέσιμο στο https://searchdatamanagement.techtarget.com/definition/data-analytics

[2] B. S. Xia and P. Gong, “Review of business intelligence through data analysis”, Benchmarking, vol. 21, pp. 300-311, April 2014

[3] R. K. Pearson, “Exploratory Data Analysis Using R”, Chapman & Hall/CRC. USA, 2018

[4] R. Minusha Silva, “Sports Analytics”, PhD Thesis, Simon Fraser University. Canada, 2016

[5] T. W. Miller, “Sports Analytics and Data Science: Winning the Game with Methods and Models”, Pearson Education, Inc. USA, 2016

[6] C. Croarkin and P. Tobias, “Handbook of Statistical Methods”, National Institute of Standards and Technology, USA, 2012

[7] A. T. Jebb, S. Parrigon and S. E. Woo, “Exploratory data analysis as a foundation of inductive research”, Human Resource Management Review, vol. 27, pp. 265-276, June 2017

[8] H. Tyralis, G. Karakatsanis, K. Tzouka and N. Mamassis, “Exploratory data analysis of the electrical energy demand in the time domain in Greece”, Energy, vol. 134, pp. 902-918, September 2017

[9] H. Zhao, Q. Meng and Y. Wang, “Exploratory data analysis for the cancellation of slot booking in intercontinental container liner shipping: A case study of Asia to US West Coast Service”, Transportation Research Part C, vol. 106, pp. 243-263, September 2019

[10] S. Castano, A. Ferrara and S. Montanelli, “Exploratory analysis of textual data streams”, Future Generation Computer Systems, vol. 68, pp. 391-406, March 2017

[11] S. Theodoridis and K. Koutroumbas, “Pattern Recognition”, Academic Press, USA, 2008

(12) Howard Demuth & Mark Beale Neural Network Toolbox, http://www.image.ece.ntua.gr/courses_static/nn/matlab/nnet.pdf)

[12] R. Xu and D. Wunsch, “Clustering”, Wiley-IEEE Press, USA, 2008

[13] W. Homenda and W. Pedrycz, “Pattern Recognition: A Quality of Data Perspective”, John Wiley & Sons, Inc. USA, 2018

[14] S. Mishra and A. Datta-Gupta, “Chapter 4 - Regression Modeling and Analysis”, Applied Statistical Modeling and Data Analytics, Elsevier Inc. USA, 2018

[15] C. M. Bishop, “Pattern Recognition and Machine Learning”, Springer Science+Business Media, LLC. Singapore, 2006

[16] I. Goodfellow, Y. Bengio and A. Courville, “Deep Learning”, The MIT Press, USA, 2016

[17] B. E. Boser, I. M. Guyon and V. N. Vapnik, “A training algorithm for optimal margin classifiers”, Proceeding of the COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152

[18] A. K. Jain, “Data clustering: 50 years beyond K-means”, Pattern Recognition Letters, vol. 31, pp. 651-666, June 2010

[19] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). *Journal of Machine Learning Technologies*. **2**(1): 37–63.

[20] [Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on*

Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. **2** (12): 1137–1143. CiteSeerX 10.1.1.48.529.]

[21] [Grossman, Robert; Seni, Giovanni; Elder, John; Agarwal, Nitin; Liu, Huan (2010). "Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions". Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool]

Υπόμνημα Πινάκων, Σχημάτων, Εικονών και Γραφημάτων

Γραφήματα:

Σχήματα:

Εικόνες:

Πίνακες: