ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ

ΜΕΛΕΤΗ ΝΑΥΤΙΚΩΝ ΑΤΥΧΗΜΑΤΩΝ ΚΑΙ ΠΡΟΒΛΕΨΗ ΔΑΠΑΝΩΝ ΜΕΣΩ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΠΑΠΑΘΑΝΑΣΙΟΥ ΔΗΜΗΤΡΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

Περιεχόμενα

Ευρετήριο Διαγραμμάτων

Ευρετήριο Πινάκων

Πρόλογος

Όπως όλες οι εταιρείες με επιχειρηματική δραστηριότητα έτσι και οι ναυτιλιακές καλούνται να δράσουν με σκοπό την κερδοφορία η οποία μετρά την αποτελεσματικότητα της επιχείρησης να χρησιμοποιεί την περιουσία της ώστε να παράγει κέρδος και αξία. Το ανθρώπινο δυναμικό που βρίσκεται εν πλω, κατά κύριο λόγο αποτελεί τον παράγοντα που διασφαλίζει την ομαλή και προγραμματισμένη πορεία ενός εμπορικού ταξιδιού το οποίο ταξίδι αποτελεί την βασικότερη δραστηριότητα μιας ναυτιλιακής εταιρείας. Λόγω του σύνθετου περιβάλλοντος και των απρόβλεπτων συνθηκών που χαρακτηρίζουν ένα ταξίδι, η ναυτιλιακή εταιρεία βρίσκεται συνεχώς σε εγρήγορση και πρέπει να είναι έτοιμη να ανταπεξέλθει, σκεπτόμενη πάντα με εμπορική σκοπιά, σε νέα δεδομένα. Τα ξαφνικά ατυχήματα που γίνονται πάνω σε πλοία του εμπορικού ναυτικού αποτελούν μερίδιο των αστάθμητων και απρόβλεπτων συμβάντων και μπορεί να έχουν άμεσο αντίκτυπο στην πορεία του πλοίου καθώς και στην συνολική απόδοση του ταξιδιού.

Ο σκοπός της διπλωματικής εργασίας είναι η θεωρητική και πειραματική μελέτη αλγορίθμων εποπτευόμενης μηχανικής μάθησης με τη χρήση του προγραμματιστικού περιβάλλοντος R. Απώτερος στόχος είναι η δημιουργία και η ανάπτυξη αξιόλογων μοντέλων πρόγνωσης τα οποία θα προβλέπουν τη σημαντικότητα των ατυχημάτων που συμβαίνουν κατά τη διάρκεια του ταξιδιού ενός εμπορικού πλοίου, παράμετρος η οποία αποτελεί βασικό χαρακτηριστικό για την πρόβλεψη της δαπάνης που απαιτείται για την αντιμετώπισή του.

Γενικά το πρόβλημα της πρόβλεψης αποτελεί ένα από τα μεγαλύτερα σημεία ενδιαφέροντος στον κλάδο της μηχανικής μάθησης με ευρύ ερευνητικό φάσμα. Η εφαρμογή της πρόβλεψης βρίσκεται σε ποικίλα επιστημονικά πεδία όπως είναι η ιατρική, μετεωρολογία, βιολογία καθώς επίσης και στη βιομηχανία-οικονομία όπως για παράδειγμα στον κλάδο της αυτοκινητοβιομηχανίας, τράπεζες ασφαλιστικές κλπ. έχοντας διάφορες πρακτικές εφαρμογές. Στο χώρο της ναυτιλίας η χρήση μοντέλων πρόβλεψης είναι επίσης διαδεδομένη ωστόσο δεν υπάρχει κάποια προηγούμενη εφαρμογή για το θέμα της παρούσας εργασίας.

Τα δεδομένα που θα χρησιμοποιήσουμε προέρχονται από ανώνυμη βάση δεδομένων. Αρχικά θα μελετήσουμε τη συμπεριφορά και την απόδοση των αλγορίθμων μηχανικής μάθησης και στη συνέχεια θα ενισχύσουμε την απόδοση της μέσα από τεχνικές βελτιστοποίησης. Τέλος θα συγκρίνουμε τα αποτελέσματα και θα εξάγουμε τα ανάλογα αποτελέσματα.

Abstract

Like all business entities, shipping companies are called upon to act for profitability, which is how well a company utilizes its resources in the purpose to generate profit and shareholder value. The onboard human resources are the most essential factor that ensures a smooth and planned course of a business voyage, which is the main activity of a shipping company. Due to the complex environment and unforeseen circumstances that characterize a voyage, the shipping company is constantly on the alert and must be ready to manage, the new data and the new situation. Sudden accidents that happen on merchant ships are part of the unforeseen and unexpected events and can have a direct impact on the ship's course as well as the overall performance of the voyage.

The main goal of this master thesis is to study the theoretical and experimental algorithms of supervised machine learning using the R programmatic environment. The final goal is to develop and deploy valuable forecasting models that predict the importance of accidents occurring during a voyage of a merchant ship, which is a key feature of predicting the required total treatment cost.

In general, prediction is one of the core points of interest in machine learning field with a wide range of research. The application of the forecast is carried out in various scientific majors such as medicine, meteorology, biology as well as the industry, such as the automobile industry, insurance companies, banks, etc. with various practical applications. In the shipping industry, the use of forecasting models is also widespread, but there is no prior application to the subject of the present master thesis.

The data we will use comes from an anonymous database. Primarily we will study the behavior and the performance of the applied machine learning algorithms and then enhance its performance through optimization techniques. Finally, we will compare the results and extract the analogous results.

Δομή διπλωματικής εργασίας

Στο πρώτο κεφάλαιο της διπλωματικής εργασίας δίνεται εισαγωγή στη μηχανική μάθηση καθώς και οι βασικές έννοιες που την διέπουν. Αναλυτικότερα αναφέρονται οι μεθοδολογίες της μηχανικής μάθησης οι διαδικασίες των δεδομένων που χρησιμοποιούνται καθώς επίσης η εποπτευόμενη μηχανική μάθηση. Στο δεύτερο κεφάλαιο παρουσιάζονται οι διαφορετικοί αλγόριθμοι της εποπτευόμενης μηχανικής μάθησης. Επιπλέον, περιγράφονται ο τρόπος με τον οποίο χρησιμοποιούνται καθώς και τα κριτήρια επιλογής τους ανάλογα την περίπτωση. Στο τρίτο κεφάλαιο στα πλαίσια της εκτίμησης του υπό κατασκευή μοντέλου, αναλύεται η

εγκυρότητα των κατηγοριοποιητών. Περιγράφονται οι τεχνικές ανάλυσης και εκτίμησης της απόδοσης των αλγορίθμων (accuracy, RT, ROC curve κλπ.) καθώς και διάφορες τεχνικές βελτιστοποίησης της απόδοσης των αλγορίθμων. Επίσης περιγράφεται η έννοια της υπεργενίκευσης (overfitting). Στο τέταρτο κεφάλαιο δίνονται οι βασικές ναυτιλιακές γνώσεις ώστε να γίνει σωστά η γεφύρωση των μοντέλων πρόβλεψης με τα ναυτικά ατυχήματα. Αναλύεται σε βάθος η αξία που προκύπτει απο την πρόβλεψη του κόστους των ναυτικών ατυχημάτων. Στο πέμπτο κεφάλαιο θα παρουσιαστούν οι προτεινόμενες τεχνικές πρόβλεψης καθώς και τα αποτελέσματα που δίνει ο κάθε αλγόριθμος. Επιπλέον, περιγράφονται τα μέτρα απόδοσης του κάθε αλγορίθμου για το συγκεκριμένο πρόβλημα, όπως αυτά αναλύονται από τη θεωρητική σκοπιά της εκτίμησης των αλγορίθμων (accuracy, LogLoss, ROC curve κλπ.). Τέλος δίνονται τα τελικά αποτελέσματα των υπολογισμών και γίνεται μια σύγκριση ως προς την καταλληλόλητα και την ακρίβεια των αλγορίθμων καθώς και οι προτάσεις για μελλοντική μελέτη. Τέλος στο έκτο κεφάλαιο δίνεται η βιβλιογραφία.

1.Εξόρυξη Δεδομένων & Μηχανική Μάθηση

1.1 Βασικές έννοιες Μηχανικής Μάθησης

Η μάθηση είναι ένα σύνθετο βιολογικό και πνευματικό φαινόμενο που έχει μελετηθεί από διάφορους κλάδους της επιστήμης όπως ψυχολογία, ιατρική, βιολογία και άλλοι. Το εύρος των διαδικασιών της μάθησης είναι αρκετά μεγάλο και η ένταξη τους σε μία και μοναδική κατηγορία δεν μπορεί να είναι βάσιμη. Επίσης τα όσα γράφονται και λέγονται για τη μάθηση αποτελούν επιστημονικές υποθέσεις που εξάγονται από την παρατήρηση και τη μελέτη των αποτελεσμάτων της. Έχοντας υπόψη τα παραπάνω και γνωρίζοντας ότι κανένας ορισμός της μάθησης δεν μπορεί να είναι ικανοποιητικός, παρατίθεται παρακάτω ένας ορισμός που προτάθηκε από τον Kimble (Kimble, 1980) ο οποίος μπορεί να θεωρηθεί αρκετά αντιπροσωπευτικός: «Μάθηση είναι μια σχετικά σταθερή αλλαγή σε μια δυνατότητα της συμπεριφοράς, η οποία συμβαίνει ως αποτέλεσμα ενισχυμένης πρακτικής». Ένας ακόμη ορισμός είναι του Gagné σύμφωνα με τον οποίο (Gagné, 1975) η μάθηση είναι η διαδικασία που υποβοηθά τους οργανισμούς να τροποποιήσουν τη συμπεριφορά τους σε ένα σχετικά σύντομο χρονικό διάστημα και με ένα μόνιμο τρόπο, έτσι ώστε η ίδια η τροποποίηση ή αλλαγή

να μη χρειαστεί να συμβεί κατ’ επανάληψη σε κάθε νέα περίπτωση. Η αλλαγή ή τροποποίηση αυτή γίνεται αντιληπτή από το ίδιο το πρόσωπο που μαθαίνει, αφού από τη στιγμή που θα έχει ολοκληρωθεί η μάθηση, θα είναι σε θέση να εκτελεί ορισμένες πράξεις που δεν θα μπορούσε να κάνει προηγουμένως.

Κάθε άνθρωπος αναπτύσσει ένα μοντέλο για το περιβάλλον στο οποίο ζει ύστερα από την παρατήρηση και τη μελέτη των χαρακτηριστικών του περιβάλλοντος. Η δημιουργία ενός τέτοιου μοντέλου, ονομάζεται επαγωγική μάθηση (inductive learning) ενώ η διαδικασία γενικότερα ονομάζεται επαγωγή (induction). Κάθε άνθρωπος συλλέγει ομαδοποιεί και συσχετίζει τις εμπειρίες και τις παραστάσεις του και με τον τρόπο αυτό δημιουργεί πρότυπα (patterns). Οι παράγοντες μάθησης αναφέρονται σε όλα εκείνα τα στοιχεία που είναι δυνατό να επηρεάσουν τη διαδικασία και το αποτέλεσμα της μάθησης. Οι παράγοντες αυτοί και οι κατηγοριοποιήσεις τους δεν είναι μοναδικοί και καθολικά αποδεκτοί. Έτσι μηχανική μάθηση (Machine Learning) ορίζεται η ανάπτυξη μοντέλων και μοτίβων από τα δεδομένα που έχει επεξεργαστεί ο υπολογιστής. Πολλοί και διάφοροι ορισμοί έχουν προταθεί για την μηχανική μάθηση, γνωστότεροι ωστόσο είναι οι παρακάτω: Mitchell (1997), "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E". Witten & Frank (2000), "Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον". Η διαδικασία της μηχανικής μάθησης αποτελείται από την εκπαίδευση και τον έλεγχο. Πιο λεπτομερώς αυτό σημαίνει ότι το σύνολο που εκπαιδεύεται προετοιμάζει κατάλληλα τον αλγόριθμο του συστήματος ώστε να νέα δεδομένα που θα εξυπηρετήσουν τη διαδικασία του ελέγχου να δώσουν αποτελέσματα εφάμιλλα με αυτά της εκπαίδευσης. Το πρώτο βήμα είναι να αποφασιστούν οι κατηγορίες ταξινόμησης στις οποίες θα καταλήξει ο διαχωρισμός των κλάσεων. Οι κατηγορίες ταξινόμηση είναι δύο, π.χ. εάν το δείγμα που αφορά η παρατήρηση έχει καεί από τον ήλιο ή όχι.

1.2 Μέθοδοι Μηχανικής Μάθησης

Στο γενικό πλαίσιο ο τομέας της Μηχανικής Μάθησης αναπτύσσεται μέσα από τρεις μεθόδους μάθησης. Αυτές είναι η επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση και η ενισχυτική μάθηση. Πιο αναλυτικά:

Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα: Ταξινόμησης (Classification) ο Πρόγνωσης (Prediction) ο Διερμηνείας (Interpretation)

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα: ο Ανάλυσης Συσχετισμών (Association Analysis) ο Ομαδοποίησης (Clustering)

Ενισχυτική Μάθηση (Reinforcement Learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

Στο χώρο της Μηχανικής Μάθησης υπάρχει ένας κατάλληλος τρόπος μάθησης για να λύσει ένα πιθανό πρόβλημα και για κάθε μέθοδο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί. Στις παρακάτω παραγράφους θα περιγράφει αναλυτικότερα η μέθοδος επιβλεπόμενης μηχανικής μάθησης καθώς είναι και το αντικείμενο της παρούσας εργασίας.

1.3 Εποπτευόμενη μηχανική μάθηση

Η εποπτευόμενη μηχανική μάθηση είναι ένα σύστημα κατά το οποίο κατάλληλοι αλγόριθμοι βασιζόμενοι σε ένα σύνολο δεδομένων εισόδου αλλά και εξόδου δημιουργούν μια συνάρτηση η οποία μπορεί να γενικευθεί και να παράγει προβλέψεις στις περιπτώσεις όπου τα δεδομένα εξόδου είναι άγνωστα. Το σύστημα αυτό θα πρέπει να «μάθει» και να λάβει τη γνώση από τα δεδομένα εισόδου και από τα δεδομένα εξόδου ώστε να μπορέσει να δημιουργήσει τη συνάρτηση η οποία θα μπορεί να κάνει την πρόβλεψη βάση της γενίκευσης αυτής. Η

συνάρτηση πρόγνωσης έχει κάποια βασικά χαρακτηριστικά τα οποία ισχύουν σε κάθε περίπτωση ανάλυσης. Μερικά από αυτά περιγράφονται στις επόμενες προτάσεις. Τα δεδομένα εισόδου περιγράφονται βάση των γνωρισμάτων που έχουν και έχουν χαρακτηριστεί ως σημαντικά από την αρχή της μελέτης του προβλήματος που καλείται να αναλύσει το σύστημα. Κάθε είσοδος, δεδομένη ή μη, που δέχεται η συνάρτηση πρόβλεψης χαρακτηρίζεται ως στιγμιότυπο και με τον τρόπο αυτό δημιουργείται το σύνολο στιγμιότυπων. Αυτά τα δεδομένα εισόδου αποτελούν το σύνολο εκπαίδευσης. Επίσης από τα δεδομένα εισόδου λαμβάνουμε και το σύνολο ελέγχου του αλγορίθμου. Έτσι τα δεδομένα εισόδου χωρίζονται σε σύνολο εκπαίδευσης και σε σύνολο ελέγχου. Η συνάρτηση που απεικονίζει μια είσοδο από το σύνολο εκπαίδευσης στη γνωστή της έξοδο καλείται συνάρτηση στόχου. Η συνάρτηση που αναπαριστά μια είσοδο από το σύνολο εκπαίδευσης στη γνωστή έξοδο λέγεται συνάρτηση στόχου και η τιμή που επιστρέφει για ένα στιγμιότυπο από το σύνολο των δεδομένων δίνεται σε μια μεταβλητή που λέγεται μεταβλητή στόχου. Μέσω της διαδικασίας εκπαίδευσης και με τη βοήθεια της συνάρτησης λάθους που εντοπίζει τη διαφορά της μεταβλητής στόχου από την επιθυμητή έξοδο επιτυγχάνεται η βελτίωση της συνάρτησης στόχου. Η μεταβλητή εξόδου μπορεί να είναι οτιδήποτε ωστόσο κυρίως οι μέθοδοι που χρησιμοποιούνται έχουν σαν υπόθεση ότι η μεταβλητή εξόδου είναι είτε κατηγορική (ναι ή όχι) είτε συνεχής (η τιμή μιας μετοχής). Στην πρώτη περίπτωση όπου έχουμε κατηγορική μεταβλητή το πρόβλημα είναι κατηγοριοποίησης (classification) ενώ στη δεύτερη περίπτωση όπου η μεταβλητή είναι συνεχής αριθμός το πρόβλημα είναι γνωστό ως παλινδρόμηση (regression). Η ειδοποιός διαφορά με την μη εποπτευόμενη μάθηση είναι ότι στην εποπτευόμενη μέθοδο το σύνολο δεδομένων δίνεται και για τις τιμές εισόδου αλλά και για τις τιμές εξόδου ενώ στη μη εποπτευόμενη δεν είναι γνωστές οι τιμές εξόδου. Για το λόγο αυτό το μοντέλο πιθανότητας θα περιγράφεται ως $p(y|\mathbf{x})$

Γνωστότεροι αλγόριθμοι Επιβλεπόμενης Επαγωγικής Μάθησης είναι:

Δέντρα Απόφασης (Decision Trees), Μάθηση βασισμένη σε Επεξηγήσεις (Explanation-Based Learning), Μάθηση βασισμένη σε Περιπτώσεις (Case-Based Learning), Μάθηση Νευρωνικών δικτύων (π.χ. για Backpropagation Neural Networks), Μάθηση Μέσω Στατιστικών Μεθόδων (π.χ. μάθηση κατά Bayes), Συλλογική Μάθηση από Ενδυνάμωση (Boosting) κ.ά.

Η ταξινόμηση, η πρόγνωση και η διερμηνεία είναι από τις πιο συχνές και διαδεδομένες λειτουργίες πολλών συστημάτων που βασίζονται στη μηχανική μάθηση και χρησιμοποιούν υπολογιστική δύναμη. Η επιλογή του κατάλληλου αλγορίθμου σχετίζεται με την φύση των

δεδομένων, των αριθμό των χαρακτηριστικών, τον αριθμό των περιπτώσεων. Με άλλα λόγια δεν υποστηρίζεται η έννοια του «καλύτερου» αλγόριθμου αλλά αυτή του καταλληλότερου αλγόριθμου βάση του προβλήματος που πρέπει να διερευνηθεί κάθε φορά. Οι αλγόριθμοι μηχανική μάθησης απαιτούν ακρίβεια (accuracy) , (precision) καθώς και το ελάχιστο σφάλμα και βάση αυτών βαθμολογούνται για την καταλληλόλητα τους.

Γενικά το πεδίο της εποπτευόμενης μηχανικής μάθησης είναι αυτό που χρησιμοποιείται περισσότερο σε σχέση με τις υπόλοιπες επιλογές.

1.3.1 Classification

Ο στόχος στην μέθοδο της κατηγοριοποίησης είναι η δημιουργία ενός συστήματος ικανού να ταξινομήσει ορθά μια νέα παρατήρηση η οποία δεν έχει χρησιμοποιηθεί στο παρελθόν ούτε είχε χρησιμοποιηθεί κατά την εκπαίδευση του συστήματος. Έχοντας ένα σύνολο δεδομένων εκπαίδευσης που περιέχουν παρατηρήσεις των οποίων οι ιδιότητες είναι γνωστές δημιουργούνται κατηγορίες (κλάσεις) στις οποίες κατηγοριοποιούνται οι νέες παρατηρήσεις. Σε πολλά προβλήματα και περιπτώσεις εκπαίδευσης ανάλογων συστημάτων κατηγοριοποίησης, τα δεδομένα εξόδου παίρνουν τιμές από μια έως δύο κλάσεις. Ένα παράδειγμα είναι η ανάλυση και κατηγοριοποίηση εξετάσεων από ασθενείς που πάσχουν από ηπατίτιδα. Σε αυτή την περίπτωση από τις γνωστές εκ των προτέρων τιμές εξόδου οι κλάσεις που δημιουργούνται είναι δύο και αφορούν το αν πρόκειται να πεθάνει ο ασθενής ή όχι. Με άλλα λόγια οι κλάσεις της κατηγοριοποίησης είναι ναι ή όχι. Αυτή η περίπτωση καλείται δυαδική κατηγοριοποίηση καθώς οι τιμές μπορούν να είναι από δύο κλάσεις. Εάν οι κλάσεις είναι περισσότερες από δύο η μέθοδος θα καλείται multiclass classification.

2. Αλγόριθμοι Μηχανικής Μάθησης

2.1Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι εμπνευσμένα από τον τρόπο με τον οποίο λειτουργεί ο ανθρώπινος εγκέφαλος. Τα τεχνητά νευρωνικά δίκτυα είναι μαζικά παράλληλα υπολογιστικά συστήματα που αποτελούνται από ένα εξαιρετικά μεγάλο αριθμό απλών επεξεργαστών με πολλές διασυνδέσεις.

Ο ανθρώπινος εγκέφαλος μπορεί να επεξεργάζεται τεράστιες ποσότητες πληροφοριών χρησιμοποιώντας δεδομένα που αποστέλλονται από τις ανθρώπινες αισθήσεις (όραση, ακοή, αφή κλπ.) . Η επεξεργασία αυτή γίνεται από νευρώνες, οι οποίοι ενεργοποιούνται από ηλεκτρικά σήματα που διέρχονται μέσω αυτών έχοντας ως βασική παράμετρο την μετάδοση ή μη των προαναφερθέντων ηλεκτρικών σημάτων. Οι σύγχρονοι υπολογιστές, σε πολλούς τομείς όπως είναι οι αριθμητικοί υπολογισμοί, ο χειρισμός συμβόλων κ.α., έχουν ξεπεράσει τον άνθρωπο σε μεγάλο βαθμό. Ωστόσο δεν είναι δυνατόν από έναν υπολογιστή να εκτελεστούν ίδιες διεργασίες με έναν άνθρωπο όπως για παράδειγμα η αναγνώριση των συγγενικών προσώπων ενός ανθρώπου σε ένα μεγάλο πλήθος ατόμων. Έτσι γίνεται αντιληπτό ότι το βιολογικό νευρωνικό δίκτυο είναι ανώτερο του τεχνητού και ως εκ τούτου η έμπνευση για την ανάπτυξη μιας εναλλακτικής χαλαρής κράτησης, αποκεντρωμένη αρχιτεκτονική που μιμείται τον εγκέφαλο. Η έμφαση δίνεται, όπως και σε πολλούς κλάδους της πληροφορικής, στην υπολογιστική δύναμη που διατίθεται και είναι ικανή να δώσει πολλαπλά οφέλη στο σύστημα των τεχνιτών νευρωνικών δικτύων. Τα κύρια χαρακτηριστικά του κάθε νευρώνα είναι:

Δενδρίτες: Σημεία εισόδου κάθε νευρώνα που λαμβάνουν εισροή από άλλους νευρώνες του δικτύου υπό μορφή ηλεκτρικών παλμών

Cell Body: Δημιουργεί συμπεράσματα από τις εισόδους των δενδριτών και αποφασίζει τι δράση που πρέπει να αναληφθεί

Ακροδέκτες Axon: Μεταδίδουν τις εξόδους με τη μορφή ηλεκτρικών παλμών στον επόμενο νευρώνα

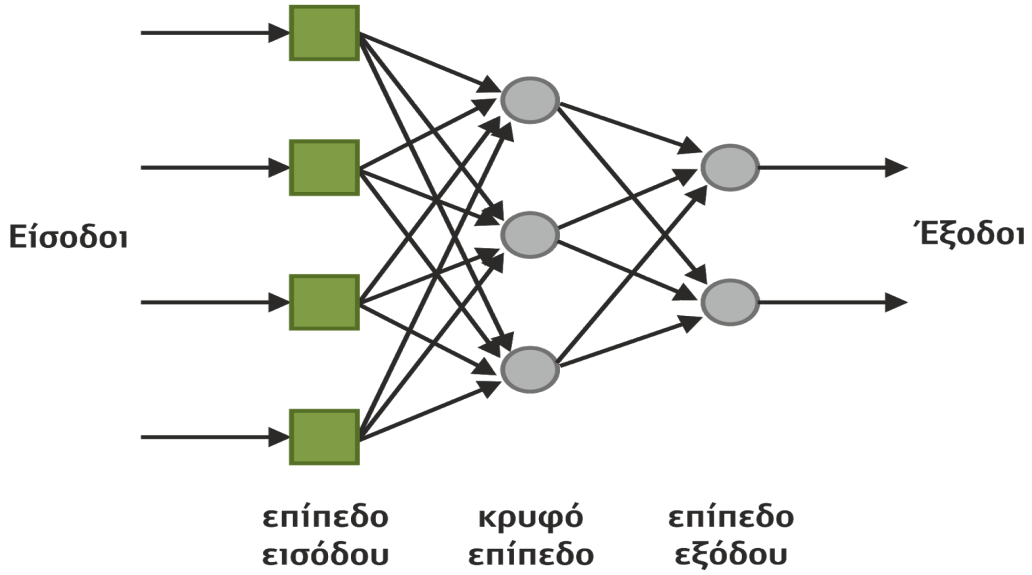
Μερικοί από τους πιο δημοφιλής λόγους χρήσης είναι η αναγνώριση προτύπου κειμένου για να μετατρέψει τις εικόνες σε κείμενο και στη συνέχεια σε οποιαδήποτε μορφή ανάλογα με τις ανάγκες των τελικών χρηστών, αναγνώριση εικόνων, απαλοιφή του θορύβου από ένα σύνολο δεδομένων, κλπ.

Η λειτουργία των νευρωνικών δικτύων είναι παρόμοια με τη δομή του βιολογικού νευρώνα, όπου ο νευρώνας ορίζεται ως κεντρική μονάδα επεξεργασίας, η οποία εκτελεί μια μαθηματική λειτουργία για να παράγει μία έξοδο από ένα σύνολο εισόδων. Η έξοδος ενός νευρώνα είναι συνάρτηση του σταθμισμένου αθροίσματος των εισροών συν την «τιμή ενεργοποίησης». Κάθε νευρώνας εκτελεί μια πολύ απλή λειτουργία που περιλαμβάνει την ενεργοποίηση του στην

περίπτωση όπου η συνολική ποσότητα του σήματος που λαμβάνεται υπερβαίνει ένα όριο ενεργοποίησης, όπως φαίνεται στο παρακάτω σχήμα:

Έχοντας ως δεδομένο ότι η ΑΝΝ είναι ένα σύνολο μαθηματικών προσεγγίσεων, η λειτουργία ενός νευρωνικού δικτύου μπορεί να διατυπωθεί ως ο υπολογισμός των εξόδων από όλους τους νευρώνες, ο οποίος είναι ένας εξ ολοκλήρου ντετερμινιστικός υπολογισμός. Κάθε σύστημα ΑΝΝ έχει τα παρακάτω χαρακτηριστικά:

- Στρώμα εισόδου
- Κρυφό στρώμα
- Στρώμα εξόδου
- Συναπτικά βάρη
- Λειτουργία ενεργοποίησης



Εικόνα 1: Επίπεδα Νευρωνικού Δικτύου

Επίπεδα νευρώνων

Κάθε νευρωνικό δίκτυο που επεξεργάζεται ένα πρόβλημα έχει την ακόλουθη αρχιτεκτονική:

Υπάρχει ένα σύνολο εισόδων, ένας επεξεργαστής και ένα σύνολο εξόδων. Οι είσοδοι σχηματίζουν το στρώμα εισόδου, το ενδιάμεσο επίπεδο το οποίο εκτελεί και τη διεργασία και καλείται κρυφό επίπεδο και το οι έξοδοι σχηματίζουν το στρώμα εξόδου. Το κρυμμένο επίπεδο έχει την ικανότητα να μετατρέψει την τιμή εισόδου στην επιθυμητή έξοδο. Η κατανόηση του κρυφού επιπέδου απαιτεί γνώση των βαρών, της ΄προκατάληψης’ και των λειτουργιών ενεργοποίησης.

Βάρη και ‘προκατάληψη’ (bias)

Τα βάρη σε ένα ANN είναι ο σημαντικότερος παράγοντας για τη μετατροπή μιας εισόδου στην αντίστοιχη έξοδο. Αυτό είναι παρόμοιο με την κλίση στην γραμμική παλινδρόμηση, όπου πολλαπλασιάζεται το βάρος στην είσοδο για να προστεθεί και να διαμορφώσει την έξοδο. Τα βάρη είναι αριθμητικές παράμετροι οι οποίες καθορίζουν πόσο ισχυρά επηρεάζουν ο ένας τον

άλλο νευρώνα. Για έναν τυπικό νευρώνα, εάν οι είσοδοι είναι x1, x2 και x3, τότε τα συναπτικά βάρη που πρόκειται να του εφαρμοστούν, υποδεικνύονται ως w1, w2 και w3. Η έξοδος θα είναι:

$$y = f(x) = \sum xi\ wi$$

όπου i είναι 1 στον αριθμό εισόδων. Με τον παραπάνω υπολογισμό μπορούμε να φτάσουμε στο σταθμισμένο άθροισμα. Η μεροληψία (bias) είναι μια πρόσθετη παράμετρος που χρησιμοποιείται για να ρυθμίσει την έξοδο μαζί με το σταθμισμένο άθροισμα των εισροών στο νευρώνα. Η επεξεργασία του γίνεται από έναν νευρώνα και δηλώνεται ως εξής:

Output = sum (weights*inputs) + bias

Σε αυτή την έξοδο (output) εφαρμόζεται μια συνάρτηση η οποία ονομάζεται λειτουργία ενεργοποίησης. Η είσοδος του επόμενου επιπέδου είναι η έξοδος των νευρώνων του προηγούμενου επιπέδου.

2.2 Naïve Bayes

Οι Μπεϋζιανοί ταξινομητές είναι στατιστικοί ταξινομητές και μπορούν να προβλέψουν πιθανότητες κλάσης, όπως την πιθανότητα ότι μια πλειάδα ανήκει σε μια συγκεκριμένη κλάση. Το μοντέλο πιθανοτήτων είναι γνωστό από την στατιστική σύμφωνα με το οποίο ισχύει ότι:

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

P (A|B) όπου είναι η δεσμευμένη πιθανότητα του ενδεχομενου A, δοθέντος του ενδεχομένου B.

P (A) όπου είναι η πιθανότητα του ενδεχομενου A

P (B) όπου είναι η πιθανότητα του ενδεχομένου B

P (B|A) όπου είναι η πιθανότητα του ενδεχομένου B δοθέντος του ενδεχομένου B

Στην περίπτωση της μηχανικής μάθησης η παραπάνω υπόθεση που δίνεται από τη στατιστική μεταφράζεται ως ακολούθως: αρχικά υπάρχει η πιθανότητα η υπόθεση A να ταξινομεί σωστά τα στιγμιότυπα P (A) και η πιθανότητα η υπόθεση να ταξινομεί σωστά τα στιγμιότυπα ενός συγκεκριμένου χώρου εκπαίδευσης (έστω για το παράδειγμά μας B) συμβολίζεται με P (A|B) . Η τελευταία καλείται δεσμευμένη πιθανότητα. Η λειτουργία των ταξινομητών κατά Bayes στηρίζονται στην υπόθεση ότι το μοντέλο κατασκευής σχετίζεται άμεσα με την κατανομή των πιθανοτήτων που παρουσιάζουν τα στιγμιότυπα του προβλήματος αναφορικά με την κλάση στην οποία ανήκουν. Άρα έχοντας μια τελείως διαφορετική κατανόηση του χώρου υποθέσεων σε σχέση με άλλους ταξινομητές κατασκευάζεται ένας ταξινομητής κατά τον οποίο αναζητείται η υπόθεση με την μεγαλύτερη πιθανότητα να ταξινομεί σωστά τα στιγμιότυπα του συνόλου εκπαίδευσης.

Κατά τη δημιουργία ενός μοντέλου πρόβλεψης χρησιμοποιείται για την εκτίμηση της πιθανότητας ενός στιγμιότυπου να ανήκει σε μια συγκεκριμένη κλάση έχοντας ως προϋπόθεση ότι τα χαρακτηριστικά είναι μεταξύ τους ανεξάρτητα. Ωστόσο πολλές φορές στην πράξη αυτό δεν συμβαίνει καθώς στη φύση τα περισσότερα χαρακτηριστικά που διέπουν ένα υπό μελέτη αντικείμενο είναι εξαρτημένα το ένα με το άλλο. Μια πρακτική δυσκολία στην εφαρμογή της μάθησης κατά Bayes είναι η απαίτηση για τη γνώση πολλών τιμών πιθανοτήτων. Όταν αυτές οι τιμές δεν είναι δυνατό να υπολογιστούν επακριβώς, υπολογίζονται κατ' εκτίμηση από παλαιότερες υποθέσεις, δηλαδή με εμπειρική γνώση. Η παραπάνω δυσκολία εφαρμογής έχει δώσει μεγάλη πρακτική αξία σε μια απλουστευμένη εκδοχή της μάθησης κατά Bayes, τον απλό ταξινομητή Bayes, στον οποίο γίνεται η παραδοχή ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους και η μάθηση στηρίζεται σε στατιστικά στοιχεία .Με άλλα λόγια οι απλοϊκοί Μπεϋζιανοί ταξινομητές υποθέτουν ότι η επίδραση της τιμής ενός χαρακτηριστικού σε μια δοσμένη κλάση είναι ανεξάρτητη από τις τιμές των άλλων χαρακτηριστικών. Αυτή η υπόθεση καλείται υπό συνθήκη ανεξαρτησία κλάσης, δηλαδή δοσμένης της κλάσης μιας πλειάδας, οι τιμές των χαρακτηριστικών θεωρούνται ότι είναι υπό συνθήκη ανεξάρτητες η μία με την άλλη. Αυτό απλοποιεί τους υπολογισμούς και για το λόγο αυτό καλείται “naïve”. Όταν ισχύει αύτη η

υπόθεση, τότε ο Naïve Bayes ταξινομητής είναι ο πιο ακριβής σε σύγκριση με όλους τους άλλους ταξινομητές.

Βέλτιστος ταξινομητής Bayes

Κατά την ταξινόμηση ενός στιγμιότυπου, προτεραιότητα έχει η ορθή ταξινόμηση του συγκεκριμένου στιγμιότυπου στην ανάλογη κλάση, χωρίς να έχει κάποια σημασία η επίδοση του ταξινομητή στο σύνολο των στιγμιοτύπων. Για το λόγο αυτό είναι πολύ χρήσιμο να κατασκευάζεται μια υπόθεση που θα ταξινομεί τα περισσότερα στιγμιότυπα σωστά και όχι απαραίτητα όλα.

Αφελής ταξινομητής - naïve Bayes

Ο κατηγοριοποιητής naïve Bayes πρόκειται για μια απλοποιημένη μορφή των αρχικών εκδοχών και στην ουσία η ειδοποιός διαφορά έχει να κάνει με την υπόθεση ανεξαρτησίας που εφαρμόζεται στα δεδομένα. Αν υποθέσουμε λοιπόν ότι οι τιμές των δεδομένων είναι ανεξάρτητες μεταξύ τους δημιουργείται ένα απλούστερο μοντέλο που δεν απαιτεί μεγάλη υπολογιστική δύναμη για να εκτελεστεί και ούτε μεγάλο χρόνο εκτέλεσης. Στην φύση αυτό δεν είναι ρεαλιστικό καθώς τις περισσότερες φορές η τιμή ενός χαρακτηριστικού επηρεάζει την τιμή που θα πάρει το άλλο. Ωστόσο τα αποτελέσματα που δίνει αυτή η μέθοδος είναι πολύ καλά. Domingos & Pazzani, 1997). Μια ακόμη μέθοδος για την αντιμετώπισης των συνεχών τιμών στα χαρακτηριστικά είναι η διακριτοποίηση πριν από την εκπαίδευση (Yang & Webb, 2003). Ο Ευέλικτος Ταξινομητής Bayes (Flexible Bayes ή FB Classifier) είναι ο ταξινομητής που προσδιορίζει με μεγαλύτερη ακρίβεια την άγνωστη κατανομή τιμών των συνεχών χαρακτηριστικών ενός προβλήματος, επιχειρώντας να προσεγγίσει την πυκνότητα πιθανότητάς τους μέσω κανονικών πυρήνων (Gaussian Kernels). Με αυτό τον τρόπο, ουσιαστικά επιτυγχάνεται να υπολογιστούν με μεγαλύτερη ακρίβεια η μέση τιμή και η διασπορά της κατανομής των τιμών κάθε χαρακτηριστικού για την εκάστοτε κλάση. Όπως γίνεται αντιληπτό, η χρήση του FB πλεονεκτεί κατά πολύ εκείνης του NB, στην περίπτωση που η κατανομή των συνεχών τιμών των χαρακτηριστικών του χώρου ενός προβλήματος δεν προσεγγίζεται από την κανονική κατανομή. Το αρνητικό αυτής της διαδικασίας είναι ότι αυξάνεται η πολυπλοκότητα του σταδίου της εκπαίδευσης έχοντας άμεσο αντίκτυπο στον αποθηκευτικό χώρο καθώς και στον χρόνο εκτέλεσης που απαιτείται. Πιο συγκεκριμένα απαιτείται μεγάλος χώρος

αποθήκευσης για να αποθηκεύσει όλες τις διαφορετικές τιμές που συναντά στα στιγμιότυπα εκπαίδευσης μιας συγκεκριμένης κλάσης που αντιστοιχούν σε κάθε χαρακτηριστικό, καθώς αποτελούν τη μέση τιμή της εκάστοτε επιμέρους κατανομής και καταναλώνεται περισσότερος χρόνος για τον υπολογισμό των τιμών των επιμέρους συναρτήσεων πυκνότητας πιθανότητας του κάθε χαρακτηριστικού, για κάθε στιγμιότυπο εκπαίδευσης της δεδομένης κλάσης .

2.3 Μηχανές Διανυσμάτων Υποστήριξης - SVM

Οι Μηχανές Διανυσμάτων Υποστήριξης SVM (Support Vector Machines) βασίζονται στη Θεωρία στατιστικής μάθησης (Statistical Learning Theory) και στα νευρωνικά δίκτυα τύπου Perceptron. Έχουν εδραιωθεί ως μια από τις πιο διαδεδομένες μεθόδους (γραμμικής και μη) παρεμβολής και ταξινόμησης, με πλήθος εφαρμογών όπως αναγνώριση γραφής (handwriting recognition), ταξινόμηση κειμένων (text categorization), στον κλάδο της ιατρικής της οικονομίας και άλλα. Γενικά, στην περίπτωση της ταξινόμησης, οι SVMs προσπαθούν να βρουν μια υπερ-επιφάνεια (hypersurface) που να διαχωρίζει στο χώρο των παραδειγμάτων τα αρνητικά από τα θετικά παραδείγματα. Η υπερεπιφάνεια αυτή επιλέγεται έτσι, ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα (maximum margin hypersurface). Πιο αναλυτικά, οι Μηχανές Διανυσμάτων Υποστήριξης είναι μια μέθοδος μηχανικής μάθησης για δυαδικά προβλήματα ταξινόμησης. Προβάλλουν τα σημεία του συνόλου εκπαίδευσης σε έναν χώρο περισσοτέρων διαστάσεων και βρίσκουν το υπερεπίπεδο το οποίο διαχωρίζει βέλτιστα τα σημεία των δύο τάξεων. Τα άγνωστα σημεία ταξινομούνται σύμφωνα με την πλευρά του υπερεπιπεδου στην οποία βρίσκονται. Τα διανύσματα τα οποία ορίζουν το υπερεπίπεδο το οποίο χωρίζει τις δύο τάξεις ονομάζονται διανύσματα υποστήριξης (support vectors).

Χρησιμοποιούν μια συνάρτηση πυρήνα π.χ. συνάρτηση πυρήνα ακτινωτής βάσης

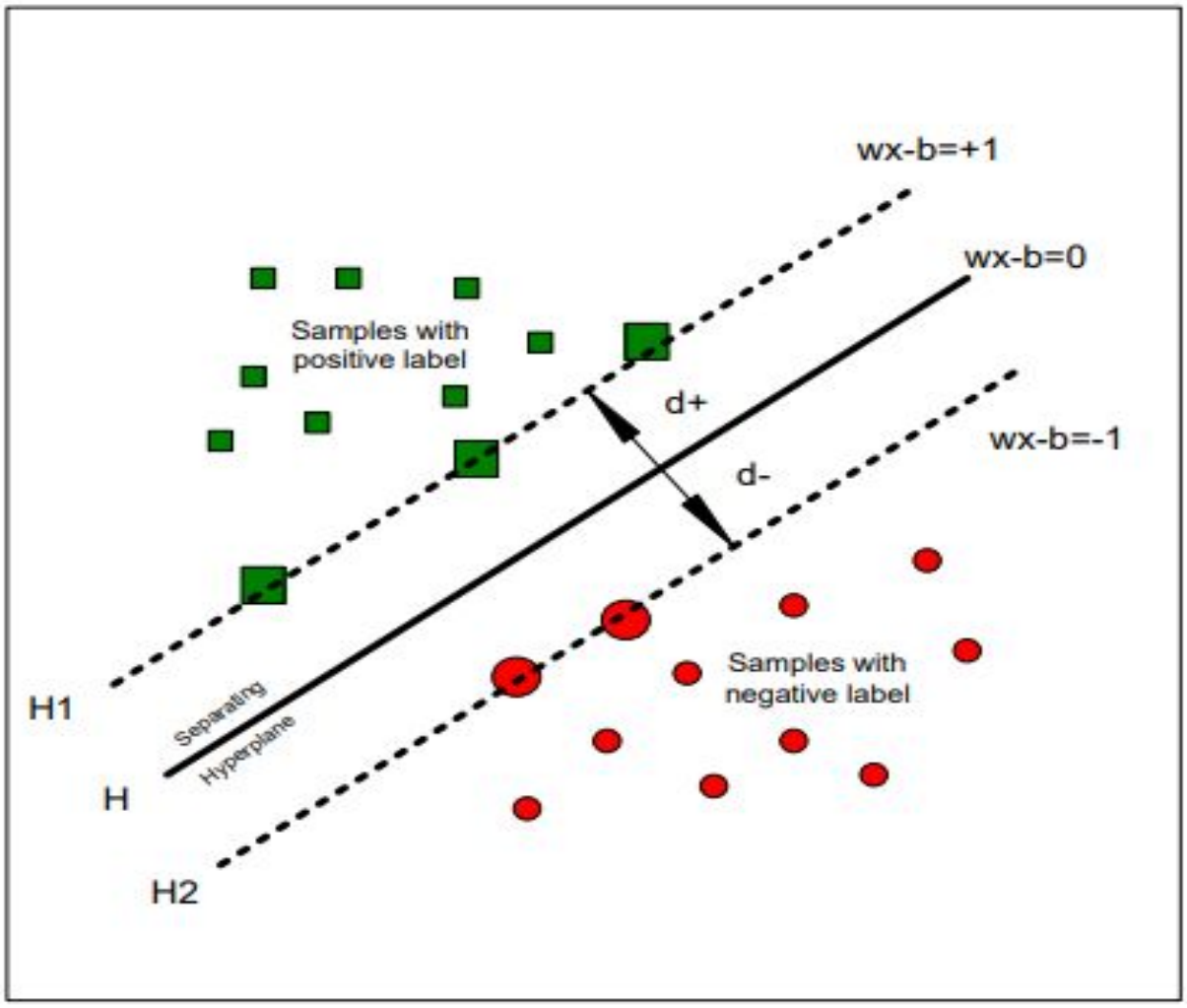
$$K(s,z) = exp(-\gamma||s-z||^2)$$

όπου, s είναι τα διανύσματα υποστήριξης, z είναι τα διανύσματα γνωρισμάτων των αντικειμένων ελέγχου και γ παράμετρος που καθορίζει το μέγεθος του πυρήναΧρησιμοποιούν το όριο απόφασης της κατηγοριοποίησης

$$\sum_{i=1}^{ns} l_i \, K(s_i \, z) + b = \Delta$$

όπου ns είναι το πλήθος των διανυσμάτων υποστήριξης s_i, l_i είναι οι ετικέτες των αντίστοιχων διανυσμάτων υποστήριξης, a, b παράμετροι που υπολογίζονται κατά τη διαδικασία εκμάθησης και Δ η τιμή κατωφλίου για την εξισορρόπηση των ρυθμών των ψευδώς θετικών και ψευδών αρνητικών. Η εξίσωση αυτή ουσιαστικά δείχνει ότι οι δύο κατηγορίες δεδομένων μπορούν να διαχωριστούν από ένα υπερεπίπεδο.

Στη θεωρία της στατιστικής μάθησης το πρόβλημα της εποπτευόμενης μάθησης διαμορφώνεται ως εξής. Δίνεται ένα σύνολο δεδομένων εκπαίδευσης έστω $\{(x_1, y_1) \ldots (x_l, y_l)\}$ σε $R^n \times R$ το οποίο ακολουθεί άγνωστη κατανομή πιθανότητας P(x, y), έχει συνάρτηση απώλειας V(y, f(x)) που μετρά το σφάλμα, για ένα δεδομένο x, όπου f(x) είναι η "προβλεπόμενη" τιμή της πραγματικής τιμής y. Το πρόβλημα συνίσταται στην εύρεση μιας συνάρτησης f που ελαχιστοποιεί το σφάλμα στα νέα δεδομένα η οποία είναι, . Στη στατιστική μοντελοποίηση θα επιλέγαμε ένα μοντέλο από το χώρο της αρχικής υπόθεσης, που θα είναι πιο κοντά στη συνάρτηση σφάλματος που αναφέραμε παραπάνω. Επίσης στην παρακάτω εικόνα (....), βλέπουμε το βέλτιστο όριο απόφασης το οποίο περιγράφεται από τη συνεχόμενη γραμμή μαζί με τις δύο διακεκομμένες γραμμές οι οποίες συνήθως καλούνται υποστηρικτικά όρια απόφασης. Τα σημεία δεδομένων που είναι οριακά κατά μήκος αυτών των δύο υποστηρικτικών ορίων είναι τα διανύσματα υποστήριξης. Αυτά τα δεδομένα είναι τα μόνα σημεία δεδομένων που επηρεάζουν τη θέση του ορίου απόφασης. Αυτό σημαίνει ότι όσο λιγότεροι είναι οι φορείς υποστήριξης, τόσο πιο γενικεύσιμο θα είναι το μοντέλο.



Εικόνα 2: Μηχανές Διανυσμάτων Υποστήριξης

Αναλυτικότερα, σε μια δίτομη ταξινόμηση η μοντελοποίηση ξεκινά με το δείγμα εκπαίδευσης T (xi, yi), (όπως είναι και το σύνολο δεδομένων εκπαίδευσης που αναφέραμε παραπάνω) όπου I = 1..., N είναι το πλήθος των προτύπων, xi (διάνυσμα M-χαρακτηριστικών) η εισαγόμενη πληροφορία που χρησιμοποιείται για την εκπαίδευση σχετικά με το πρότυπο i που στην ουσία αποτελούν τις ανεξάρτητες μεταβλητές του προβλήματος και yi ∈ {+1, -1} η αντίστοιχη έξοδος (εξαρτημένη μεταβλητή). Στόχος της ανάλυσης είναι η κατασκευή μιας συνάρτησης f(x) η οποία να διαχωρίζει τις θετικές από τις αρνητικές περιπτώσεις. Στην απλούστερη περίπτωση η f ορίζεται από ένα υπερεπίπεδο b = wx ως εξής:

$$f(x) = sgn(wx - b)$$

όπου w είναι ένα διάνυσμα των συντελεστών των x χαρακτηριστικών και b μια σταθερά. Αποδεικνύεται ότι το βέλτιστο υπερεπίπεδο -και συνεπώς η βέλτιστη συνάρτηση- είναι εκείνο με το μέγιστο περιθώριο διαχωρισμού ανάμεσα στις δύο κλάσεις. Για τον υπολογισμό του περιθωρίου αρκεί να εφαρμοστεί ο τύπος υπολογισμού της απόστασης ενός σημείου από μια ευθεία. Πιο συγκεκριμένα όπως προκύπτει και από το Σχήμα 1.4 η απόσταση του Η από το Η1 είναι:

$$d^+ = | -1 - b |/||w||$$

Ομοίως η απόσταση του Η από το Η2 είναι:

$$d^- = | 1 - b |/||w||$$

οπότε η απόσταση ανάμεσα στην Η1 και την Η2 δηλαδή το περιθώριο έστω π θα είναι:

$$\pi = d^+ d^- = 2 / ||w||$$

Συνεπώς, προκειμένου να μεγιστοποιηθεί το περιθώριο διαχωρισμού θα πρέπει να ελαχιστοποιηθεί το ||w|| με την προϋπόθεση ότι δεν υπάρχουν άλλα σημεία ενδιάμεσα στις Η1 και Η2 , δηλαδή να ισχύει

$$x_i \, w - b \geq 1, \quad y = +1$$

$$x_i \, w - b \leq 1, \quad y = +1$$

Οι δυο παραπάνω σχέσεις μπορούν να συνδυαστούν με την παρακάτω σχέση:

$$y_i (x_i \, w - b) \geq 1$$

Βασικό πλεονεκτήματα της μεθόδου είναι ότι δεν είναι επιρρεπής στην υπερπροσαρμογή του αλγορίθμου σε συγκεκριμένο σύνολο δεδομένων σε σχέση με άλλες μεθόδους (overfitting). Επίσης οι Μηχανές Διανυσμάτων Υποστήριξης αποτελούν μία ανθεκτική μέθοδο έναντι της ύπαρξης θορύβου στα δεδομένα. Στηρίζεται στην επίλυση ενός προβλήματος κυρτού

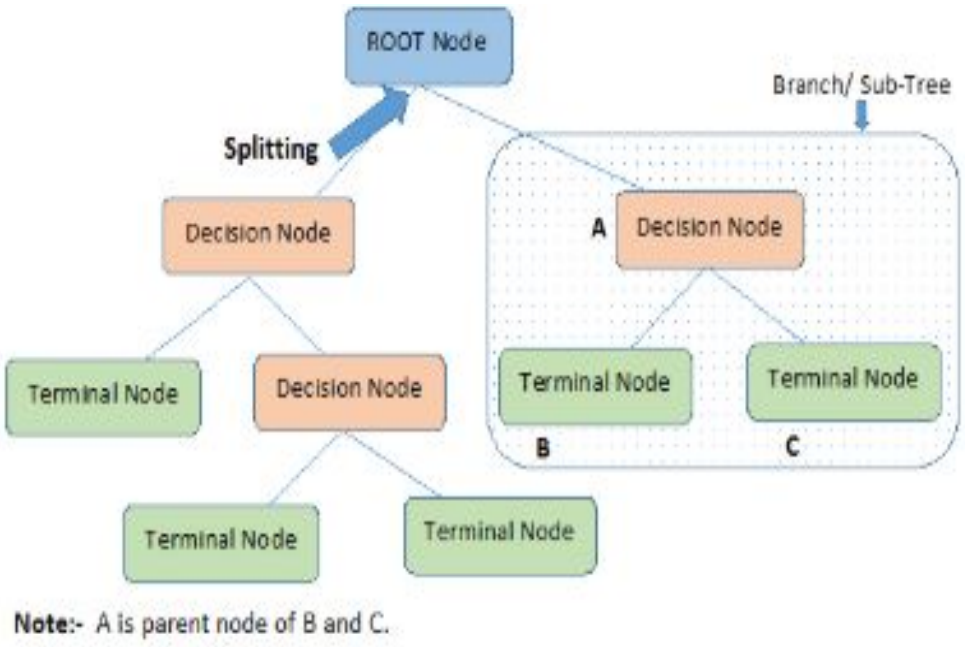
τετραγωνικού προγραμματισμού. Το σημαντικό πλεονέκτημα ως προς αυτή την ιδιότητα των SVMs είναι ότι το πρόβλημα βελτιστοποίησης ,λόγω κυρτότητας, παρουσιάζει ολικό ελάχιστο δίνοντας μοναδική βέλτιστη επιλογή, κάτι που δεν συμβαίνει σε άλλες μεθόδους όπως τα Νευρωνικά Δίκτυα που μπορούν να παγιδευτουν σε τοπικά ελάχιστα. Ακόμη, παρουσιάζουν σημαντική ικανότητα γενίκευσης σε μη γραμμικά διαχωρίσιμα δεδομένα ενσωματώνοντας το τέχνασμα του πυρήνα (kernel trick). Με την εφαρμογή συναρτήσεων πυρήνα είναι δυνατή η παραγωγή μη γραμμικών μοντέλων που οδηγούν σε γραμμικότητα σε χώρους μεγαλύτερων διαστάσεων.

Ωστόσο η ενσωμάτωση νέας γνώσης δεν είναι δυνατή. Πιο συγκεκριμένα, όταν ένα σύστημα εκπαιδεύεται σε κάποιο training set, η προσθήκη κάποιων παρατηρήσεων στο σύνολο εκπαίδευσης δεν δίνει τη δυνατότητα αναπροσαρμογής του αλγορίθμου που έχει ήδη κατασκευαστεί και έτσι η εκπαίδευση του βασίζεται στο νέο σύνολο δεδομένων εκπαίδευσης. Επίσης η εκπαίδευση ενός SVM σε κάποια προβλήματα είναι χρονοβόρα.

2.4 Δέντρα Απόφασης

Τα Δέντρα Απόφασης-ΔΑ (Decision Trees) είναι ο γνωστότερος αλγόριθμος επιβλεπόμενης Μάθησης και έχει εφαρμοστεί με επιτυχία σε πολλούς τομείς όπως είναι η ιατρική, η τεχνολογία, ο στρατηγικός σχεδιασμός του marketing και άλλα. Τα ΔΑ οδηγούν στη δημιουργία μιας δενδροειδούς μορφής που τα φύλλα αποτελούν τις κατηγορίες ταξινόμησης (classes). Η δενδροειδής αυτή μορφή μπορεί να αναγνωριστεί και ως ένα σύνολο κανόνων που καλούνται κανόνες ταξινόμησης (classification rules) και έχει ως βασικές προϋποθέσεις για την ορθή λειτουργία του αλγορίθμου την ύπαρξη ενός συνόλου χαρακτηριστικών, την ύπαρξη προκαθορισμένων κατηγοριών ταξινόμησης για τον διαχωρισμού τον οποίο θα επιδιώξει ο αλγόριθμος και την ύπαρξη επαρκούς αριθμού παρατηρήσεων που θα τροφοδοτήσουν το σύστημα με τιμές εκπαίδευσης. Ως πρώτο βήμα ορίζονται οι κατηγορίες στις οποίες θα καταλήξει ο διαχωρισμός. Έπειτα ορίζονται τα υπόλοιπα χαρακτηριστικά που θα πρέπει να αναγνωριστούν κατά την επεξεργασία των τιμών εισόδου ώστε να καταλήξει σε ένα σύνολο δειγμάτων με κοινά χαρακτηριστικά. Στα δέντρα απόφασης δημιουργούνται κόμβοι και κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού. Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται

με τον κόμβο και στα κλαδιά - φύλλα έχουμε το τι συνέβη. Τα δέντρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων χαρακτηριστικών. Συγκεκριμένα ο J48 δημιουργεί δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης χρησιμοποιώντας την έννοια της εντροπίας πληροφοριών. Ο J48 έχει ως συνθήκη ότι κάθε χαρακτηριστικό των δεδομένων μπορεί να χρησιμοποιηθεί για να λάβει μια απόφαση, η οποία στη συνέχεια χωρίζει τα δεδομένα σε μικρότερα υποσύνολα. Ύστερα χρησιμοποιεί το ομαλοποιημένο κέρδος πληροφοριών (Information gain - διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Όποιο χαρακτηριστικό έχει το υψηλότερο ομαλοποιημένο κέρδος πληροφοριών χρησιμοποιείται για να ληφθεί μια απόφαση. Παρακάτω δίνεται ένα σχηματικό παράδειγμα.



Εικόνα 3: Δέντρα απόφασης

Ένας επίσης πολύ γνωστός αλγόριθμος, που είναι η πρώτη μορφή του J48, είναι ο ID3 ο οποίος σε κάθε κόμβο του δέντρου αναζητά μεταξύ των χαρακτηριστικών του συνόλου δειγμάτων εκπαίδευσης το χαρακτηριστικό το οποίο διαχωρίζει καλύτερα τα δεδομένα. Εάν το

χαρακτηριστικό διαχωρίζει πλήρως το σύνολο εκπαίδευσης, τότε ο αλγόριθμος σταματά. Αλλιώς, λειτουργεί αναδρομικά για όλες τις πιθανές τιμές των διαχωρισμένων υποσυνόλων, για να εντοπίσει το καλύτερό τους χαρακτηριστικό. Ο αλγόριθμος εφαρμόζει αναδρομικά μια συνεχή αναζήτηση, ώστε να επιλέξει το καλύτερο χαρακτηριστικό και δεν ανατρεχει σε προηγουμενα χαρακτηριστικά που έχει χρησιμοποιήσει στη δεδομένη πορεία του για να τα επανεξετάσει. Ο ID3, όπως και κάθε άλλος επαγωγικός αλγόριθμος, μπορεί να κατηγοριοποιήσει λανθασμένα ένα νέο στιγμιότυπο για πολλούς διαφορετικούς λόγους. Ο πιο απλός είναι να μην έχει γίνει καλός σχεδιασμός και η συγκέντρωση των δειγμάτων να μην έχει βασιστεί σε σωστά χαρακτηριστικά. Το πρόβλημα δεν αφορά τον αλγόριθμο, αλλά οδηγεί σε κακής ποιότητας συμπεράσματα στα φύλλα του δέντρου που παράγει. Άλλος λόγος είναι ότι μη επαρκές πλήθος των δειγμάτων θα οδηγήσει σε πρόωρες συγκλίσεις και υπεραπλουστευμένα συμπεράσματα. Τέλος, βασικότερο πρόβλημα για το οποίο ευθύνεται ο αλγόριθμος είναι η κακή επιλογή μεθόδου για τον ευρετικό ή συνήθως στατιστικό εντοπισμό του “καλού” χαρακτηριστικού.

Το καλό χαρακτηριστικό λοιπόν επιλέγεται με τη μέθοδο της εντροπίας και του κέρδους πληροφορίας. Η εντροπία είναι η ομοιογένεια των τιμών του αρχικού συνόλου ως προς κάποιο χαρακτηριστικό. Είναι με άλλα λόγια ο βαθμός βεβαιότητας ή αβεβαιότητας των ενδεχομένων που περιλαμβάνονται σε ένα σύνολο δεδομένων. Εάν για παράδειγμα τα δείγματα του συνόλου εκπαίδευσης είναι ομοιογενή ως προς μια κατηγορία, τότε η εντροπία του ισούται με μηδέν. Εάν οι κατηγορίες είναι διαφορετικές και τα δείγματα που ανήκουν σε καθεμία έχουν το ίδιο πλήθος, τότε η εντροπία είναι 1 και έχουμε τη μέγιστη τιμή που μπορεί να πάρει η εντροπία και έχουμε έναν τέλειο διαχωρισμό των κατηγοριών. Η συνολική εντροπία λοιπόν ενός συνόλου δεδομένων περιγράφεται από τον παρακάτω τύπο.

$$\text{Entropy (S)} = \sum_{i=1}^{n} -p_i \log p_i$$

Όπου p_1, $p2 \ldots p_i$ οι πιθανότητες του κάθε ενδεχομένου που περιλαμβάνεται στο σύνολο.

Από την παραπάνω εξίσωση γίνεται αντιληπτό ότι για να υπολογιστεί η συνολική εντροπία υπολογίζεται αρχικά η εντροπία του κάθε χαρακτηριστικού η οποία είναι απαραίτητη για να υπολογιστεί το κέρδος πληροφορίας όπου περιγράφεται στην παρακάτω σχέση.

$$Gain\ (S,A)=Entropy\ (S)-Entropy\ (S,A)$$

Η εντροπία που προκύπτει ως αποτέλεσμα αφαιρείται από την εντροπία πριν από το διαχωρισμό. Το αποτέλεσμα είναι το κέρδος πληροφορίας (Gain) και πρακτικά περιγράφει το πόση πληροφορία περιέχει ένα χαρακτηριστικό. Άρα σε πρακτικό επίπεδο το σύνολο εκπαίδευσης χωρίζεται ως προς τα διαφορετικά χαρακτηριστικά και για κάθε διαφορετικό χαρακτηριστικό γίνεται ο υπολογισμός της εντροπίας για κάθε διαφορετική τιμή του χαρακτηριστικού και στη συνέχεια αφού πολλαπλασιαστεί με το ποσοστό των δειγμάτων που διαθέτουν την αντίστοιχη τιμή προστίθεται για να υπολογιστεί η συνολική εντροπία του διαχωρισμένου τμήματος. Το κέρδος πληροφορίας πρέπει να υπολογιστεί για κάθε χαρακτηριστικό που είναι ικανό να εφαρμοστεί για το διαχωρισμό ενός υποσυνόλου που είναι μη πλήρως διαχωρισμένο όταν το υποσύνολο είναι σε ένα κόμβο του δέντρου. Ο αλγόριθμος ID3 τρέχει αναδρομικά διαχωρίζοντας σε κάθε κύκλο έναν κόμβο του δέντρου που δεν είναι φύλλο, έως ότου όλα τα δείγματα κατηγοριοποιηθούν.

2.5 K-nn

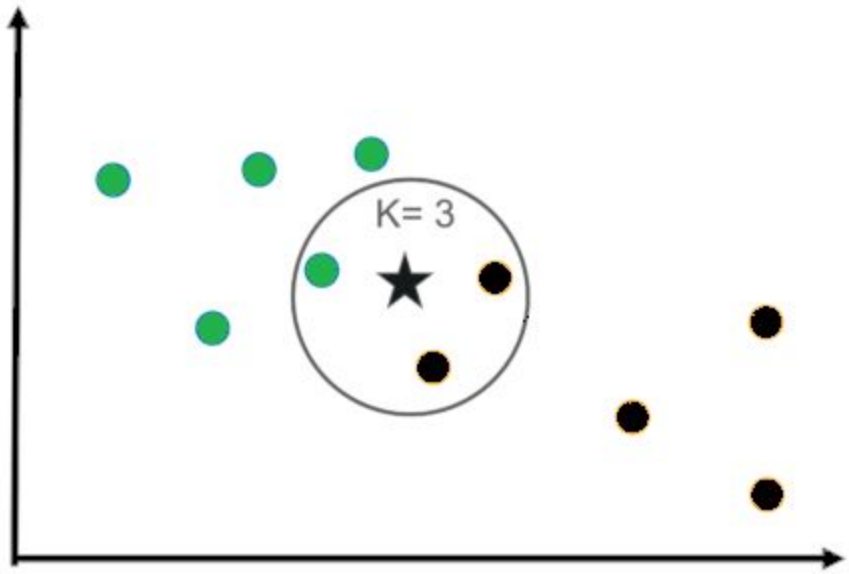
Οι Κατηγοριοποιητές Βασισμένοι σε Παραδείγματα είναι από τους απλούστερους αλγορίθμους μηχανικής μάθησης και αποτελεί μια χρήσιμη τεχνική όπου έχει εφαρμογή στην κατηγοριοποίση καθώς και σε προβλήματα παλινδρόμησης. σε αντίθεση με τις παραπάνω μεθόδους οι IBC αλγόριθμοι δεν δημιουργούν κάποιο μοντέλο γενίκευσης που προκύπτει από τα δεδομένα εκπαίδευσης αλλά όταν χρειαστεί να κατηγοριοποιηθεί μια νέα παρατήρηση, τη συγκρίνουν με γνωστές παρατηρήσεις του συνόλου εκπαίδευσης. γενικά το σύστημα αυτό λαμβάνει υπ όψιν τη συνεισφορά των γειτονικών παρατηρήσεων, έτσι ώστε οι πλησιέστεροι γείτονες να

συνεισφέρουν περισσότερο στον μέσο όρο σε σχέση με τους πιο μακρινούς γείτονες. Ετσι, κάθε πλειάδα εκπαίδευσης αναπαρίσταται από ένα σημείο σε n-διάστατο χώρο και αποθηκεύονται σε ένα ν-διάστατο χώρο και όταν είναι να μελετήσουμε μια άγνωστη πλειάδα, ένας ταξινομητής k – κοντινότερων γειτόνων αναζητά το πρότυπο χώρου για τις k πλειάδες εκπαίδευσης που βρίσκονται πλησιέστερα στην άγνωστη πλειάδα. Αυτές οι k πλειάδες εκπαίδευσης είναι οι k «κοντινότεροι γείτονες» της άγνωστης πλειάδας. Έστω ότι υπάρχει ένα πρόβλημα κατηγοριοποίησης, όπου οι παρατηρήσεις αποτελούνται από δύο αριθμητικά πεδία και το γνώρισμα της κλάσης και η κάθε παρατήρηση μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Η παρατήρηση Χ απέχει από την παρατήρηση Υ, απόσταση d(X,Y) μέσα στον δισδιάστατο χώρο. Η απόσταση d(X,Y) μπορεί να υπολογιστεί με τη βοήθεια της Ευκλείδειας απόστασης που περιγράφεται παρακάτω.

$$d(X, Y) = \sqrt{((x1 - y1)2 + (x2 - y2)2)}$$

όπου χ1, y1 οι τιμές των Χ, Υ για την πρώτη διάσταση και x2, y2 οι τιμές των Χ, Υ για τη δεύτερη διάσταση. Ο αλγόριθμος αναζητά μέσα στον δισδιάστατο χώρο τα k σημεία-παρατηρήσεις που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Ο κατηγοριοποιητής εκχωρεί τη νέα παρατήρηση στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων γειτόνων. Εάν οριστεί ότι k=1, τότε η νέα παρατήρηση εκχωρείται στην κλάση που υπάρχει η πιο όμοια παρατήρησης εκπαίδευσης.

Στο παρακάτω σχήμα υπάρχουν δύο κλάσεις στις οποίες είναι δυνατόν να κατηγοριοποιηθεί η νέα παρατήρηση που δίνεται ώς αστερίσκος. έτσι έχουμε την πρώτη κλάση που περιγράφεται με πράσινο χρώμα και τη δεύτερη με μαύρο. η τιμή του k ειναι 3 και εντοπίζονται τα πλησιέστερα σημεία τα οποία είναι τις δεύτερης κλάσης που χαρακτηρίζονται ώς μαύρα. Έστι η νέα παρατήρηση ανήκει πλέον στη δεύτερη κλάση.



Εικόνα 4: Κατηγοριοποιητής K-nn

Ο παραπανω αλγόριθμος βελτιώνεται όταν δεν λαμβάνεται η απόφαση κατηγοριοποίησης με ισότιμη ψηφοφορία μεταξύ των επιλεγμένων γειτόνων, αλλά να συνεισφέρουν ανάλογα με το πόσο κοντά βρίσκονται στην νέα παρατήρηση. Ένας απλός τρόπος για να γίνει αυτό είναι να εκχωρηθούν συντελεστές βαρύτητας ψήφου στα επιλεγμένα σημεία. Οι συντελεστές θα μπορούσαν να είναι ίσοι με 1/d, όπου d η απόσταση του εκάστοτε σημείου από τη νέα παρατήρηση. Μία αδυναμία στον υπολογισμό της ομοιότητας με βάση την Ευκλείδεια απόσταση είναι το γεγονός ότι οι μεταβλητές με μεγάλο εύρος τιμών επηρεάζουν περισσότερο το αποτέλεσμα από τις μεταβλητές με μικρό εύρος τιμών. Εάν πχ οι παρατηρήσεις έχουν δύο γνωρίσματα Α και Β και το Α παίρνει τιμές από 1 έως 1000, ενώ το Β παίρνει τιμές από 1 έως 10, τότε το γνώρισμα Α επηρεάζει δυσανάλογα την απόσταση σε σχέση με το γνώρισμα Β. Το

πρόβλημα αυτό αντιμετωπίζεται με κανονικοποίηση των αριθμητικών τιμών. Αυτό μπορεί να επιτευχθεί διαιρώντας τις τιμές των γνωρισμάτων με την περιοχή τιμών των γνωρισμάτων.

Η παρακάτω συνάρτηση παρουσιάζει των καθορισμό των βαρών με βάση την Ευκλείδεια απόσταση.

$$d(X, Y) = \sqrt{\sum wii * (xi - yi)2}$$

Η ακρίβεια του αλγορίθμου μπορεί να υποβαθμιστεί σοβαρά από την ύπαρξη θορυβώδους ή άσχετου χαρακτηριστικού ή αν οι κλίμακες χαρακτηριστικών δεν συμφωνούν με τη σημασία τους. Επίσης να σημειώσουμε ότι η μέτρηση της απόστασης γίνεται συνήθως με την ευκλείδεια απόσταση. Η καλύτερη επιλογή του k εξαρτάται από τα δεδομένα και δεν υπάρχει κάτι καλύτερο εκ των προτέρων. Αυτό που όμως γενικά ισχύει είναι ότι οι μεγαλύτερες τιμές του k μειώνουν την επίδραση του θορύβου στην ταξινόμηση αλλά καθιστούν τα όρια μεταξύ κατηγοριών λιγότερο διακριτά. Για το λόγο αυτό η διαδικασία του πειράματος είναι καθοριστική για να ορίσουμε το k με κατάλληλο τρόπο.

Η Ευκλείδεια απόσταση προϋποθέτει την ύπαρξη αριθμητικών τιμών των γνωρισμάτων ώστε να μπορέσει να υπολογίσει την απόσταση των παρατηρήσεων. ‘Ετσι όταν υπάρχουν παρατηρήσεις που δεν έχουν αριθμητικές τιμές αλλα ονομαστικές χρησιμοποιούνται παραλλαγές της Ευκλείδειας απόστασης. Στην απλούστερη εκδοχή τους οι συναρτήσεις αυτές επιστρέφουν την τιμή 0 εάν οι τιμές του ίδιου ονομαστικού γνωρίσματος δύο διαφορετικών παρατηρήσεων είναι ίδιες, ενώ αν είναι διαφορετικές επιστρέφουν την τιμή 1. Επίσης, έχουν προταθεί αλγόριθμοι που υπολογίζουν την απόσταση αντικειμένων που αποτελούνται και από αριθμητικές και από ονομαστικές τιμές. Συνοπτικά, οι κατηγοριοποιητές k-NN διαθέτουν απλό αλγόριθμο, είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών. και σε πολλές περιπτώσεις πετυχαίνουν υψηλές επιδόσεις κατηγοριοποίησης. Ορισμένα από τα βασικά μειονεκτήματα τους είναι τα ακόλουθα. Είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων και Τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k. Επίσης όταν ο αριθμός των γειτόνων είναι πολύ μεγάλος γίνονται

πολλές συγκρίσεις μεταξύ παρατηρήσεων και απαιτείται πολύς χρόνος για να γίνει η κατηγοριοποίηση των νέων παρατηρήσεων.

3. Εγκυρότητα κατηγοριοποιητών

Η τελική απόδοση ενός μοντέλου είναι συνάρτηση πολλών παραμέτρων και δε θα πρέπει βεβιασμένα ο χρήστης να βγάζει συμπεράσματα. Οι παράμετροι αυτοί πολλές φορές είναι ανεξάρτητοι μεταξύ τους γεγονός που βοηθά πολύ στη βελτίωση των τιμών των παραμέτρων, έχοντας ως στόχο την ολική βελτίωση του συστήματος. Συγκεκριμένα μελετάται κάθε παράμετρος χωριστά και υπολογίζεται η καλύτερη τιμή της και στη συνέχεια η καλύτερη τιμή που έχει βρεθεί από τις υπό εξέταση παραμέτρους, ορίζεται ως σημείο αναφοράς για τις υπόλοιπες παραμέτρους. Έτσι μπορούμε να συγκρίνουμε και μεταξύ τους τις παραμέτρους και να έχουμε χρήσιμα συμπεράσματα. Το στάδιο λοιπόν της αξιολόγησης είναι πολύ σημαντικό καθώς λαμβάνει χώρα μετά την εκπαίδευση του μοντέλου και πριν την εφαρμογή του από τον τελικό χρήστη. Με την επαλήθευση και τον έλεγχο των αποτελεσμάτων το πειραματικό στάδιο ολοκληρώνεται και το σύστημα μπορεί να κριθεί ως λειτουργικό. Το πιο σημαντικό κριτήριο για την επιλογή ενός συστήματος κατηγοριοποίησης (classification) είναι η αποτελεσματικότητα της κατηγοριοποίησης των δεδομένων που έχει απόλυτη σχέση με την απόδοση των προβλέψεων. Επίσης είναι πολύ σημαντικό αυτό το στάδιο γιατί δίνεται η δυνατότητα να βελτιωθεί η απόδοση του συστήματος ακολουθώντας βήματα που υποδεικνύουν τα μέτρα απόδοσης.

3.1 Ανισοκατανομή κλάσεων σε δυαδικό μοντέλο κατηγοριοποίησης

Τα διάφορα προβλήματα που επιλύει η μέθοδος της μηχανικής μάθησης διαφέρουν σε πολλά σημεία και δεν είναι έχουν πάντα τα ίδια χαρακτηριστικά. Ένα από τα σημεία που διαφέρουν τα προβλήματα μεταξύ τους σχετίζεται με τον αριθμό των παρατηρήσεων που περιλαμβάνει η κάθε κλάση. Υπάρχουν προβλήματα που οι παρατηρήσεις κατά την περίοδο της εκπαίδευση είναι ισομερώς κατανεμημένες και υπάρχουν και άλλα στα οποία δεν είναι. Για παράδειγμα, μια εταιρεία η οποία στην επιχειρηματική της δραστηριότητα περιλαμβάνει μια γραμμή παραγωγής στην οποία θέλει να κάνει αναγνώριση εικόνων των προϊόντων που περνούν από αυτή ώστε να μπορεί να προβλέπει τον αριθμό και τον τύπο συσκευασιών που θα χρειαστεί, ενδέχεται να έχει

ισομερώς κατανεμημένες τις παρατηρήσεις, καθώς υποθέτουμε ότι έχει ισομερώς μοιράσει την γκάμα των προϊόντος της για να κάνει διασπορά στην επένδυση. Στην αντίθετη περίπτωση έχουμε μια τράπεζα που θέλει να κάνει ανίχνευση απάτης (fraud detection) στη χρήση των πιστωτικών καρτών των πελατών της. Σε αυτή την περίπτωση προφανώς και οι κλοπές των καρτών είναι η μειονότητα των περιπτώσεων χρήσης και σαν αποτέλεσμα τα δεδομένα δεν θα είναι ισομερώς κατανεμημένα στις κλάσεις. Σε τέτοια σύνολα δεδομένων όπου οι παρατηρήσεις δεν είναι ισομερώς κατανεμημένες στις κλάσεις, το γενικό ποσοστό επιτυχών προβλέψεων δεν επαρκεί για την εκτίμηση των κατηγοριοποιητών. Επίσης, η ανισοκατανομή των κλάσεων έχει επιπτώσεις στην εκπαίδευση των κατηγοριοποιητών καθώς μελέτες έχουν δείξει ότι τα παραγόμενα μοντέλα τείνουν να προβλέπουν καλύτερα την πλειοψηφούσα κλάση και πολύ χειρότερα τη μειοψηφούσα κλάση. Η φυσική κατανομή των παρατηρήσεων σε κλάσεις συχνα δεν είναι η καλύτερη κατανομή για την εκπαίδευση ενός κατηγοριοποιητή. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες τεχνικές επαναδειγματοληψίας (resampling). Η τυχαία υποδειγματοληψία (random undersampling) είναι μια τεχνική, η οποία απομακρύνει με τυχαίο τρόπο παρατηρήσεις της πλειοψηφούσας κλάσης, μέχρι να επιτευχθεί ίσο πλήθος παρατηρήσεων για την κάθε κλάση. Ένα μειονέκτημα της τεχνικής αυτής είναι ότι διαγράφονται παρατηρήσεις οι οποίες μπορεί να περιέχουν ουσιαστική πληροφορία και ένα ακόμη είναι ότι μπορεί να οδηγήσει σε υπερπροσαρμογή καθώς η τυχαία υπερδειγματοληψία (random oversampling) αναπαράγει με τυχαίο τρόπο τις παρατηρήσεις της μειοψηφούσας κλάσης, μέχρι να ισοσταθμιστεί το πλήθος των κλάσεων. Για την αντιμετώπιση αυτών των προβλημάτων έχουν προταθεί τεχνικές, οι οποίες επιλεκτικά αναπαράγουν παρατηρήσεις της μειοψηφούσας κλάσης ή/και διαγράφουν παρατηρήσεις της πλειοψηφούσας κλάσης.

3.2 Υπεργενίκευση

Σ αυτή την παράγραφο γίνεται αναφορά στην υπεργενίκευση και στο πλαίσιο στο οποίο επηρεάζει ένα μοντέλο πρόβλεψης. Η υπεργενίκευση (overfitting) είναι ένα σφάλμα της μοντελοποίησης που συμβαίνει όταν μια συνάρτηση είναι πολύ κατάλληλη για ένα συγκεκριμένο σύνολο δεδομένων με αποτέλεσμα να μη μπορεί να δώσει ικανοποιητικά αποτελέσματα σε ένα νέο σύνολο δεδομένων το οποίο θα επεξεργαστεί ο αλγόριθμος για πρώτη φορά. Συμβαίνει με άλλα λόγια όταν το μοντέλο μας δε γενικεύει καλά από το σύνολο δεδομένων εκπαίδευσης του αλγορίθμου, προς νέα δεδομένα. Στην στατιστική το επίπεδο της καλής προσαρμογής αναφέρεται στο πόσο κοντά οι προβλεπόμενες τιμές του μοντέλου ταιριάζουν με τις

παρατηρούμενες (αληθείς) τιμές. Στην αντίθετη περίπτωση της υπεργενίκευσης (overfitting) το μοντέλο έχει ‘μάθει’ το θόρυβο αντί για την τάση (pattern / trend) των δεδομένων και εν τέλει ταιριάζει με το σύνολο δεδομένων εκπαίδευσης αλλά έχει κακή εφαρμογή με νέα σύνολα δεδομένων. Ένα μοντέλο που έχει μάθει τον θόρυβο αντί για το σήμα θεωρείται "overfit" επειδή ταιριάζει με το σύνολο δεδομένων κατάρτισης αλλά έχει κακή εφαρμογή με νέα σύνολα δεδομένων. Η υπερφόρτωση του μοντέλου γενικά συμβαίνει όταν το μοντέλο γίνεται πολύ περίπλοκο με αποτέλεσμα να μην είναι σε θέση να εξηγήσει τα χαρακτηριστικά των υπό μελέτη δεδομένων. Στην πραγματικότητα, τα δεδομένα που μελετώνται συχνά έχουν κάποιο βαθμό σφάλματος ή τυχαίο θόρυβο μέσα σε αυτά. Έτσι, ο αλγόριθμος προσπαθώντας να καταστήσει ένα καλό μοντέλο σε ελαφρώς ανακριβή δεδομένα μπορεί να μολύνει το μοντέλο με σημαντικά λάθη και να μειώσει την προβλεπτική ισχύ του. Ωστόσο πέραν του προβλήματος της υπεργενίκευσης υπάρχει και το αντίθετο πρόβλημα, το underfitting οπού είναι αναγκαίο να αποφευχθεί κατά τη δημιουργία ενός μοντέλου. Το underfitting συμβαίνει όταν ένα μοντέλο είναι πολύ απλό και έχει δημιουργηθεί από πολύ λίγα χαρακτηριστικά ή υπερβολικά μορφοποιημένο σύνολο δεδομένων - γεγονός που το καθιστά άκαμπτο στη διαδικασία της μάθησης. Οι απλοί αλγόριθμοι εκμάθησης τείνουν να έχουν λιγότερες διακυμάνσεις στις προβλέψεις τους, αλλά περισσότερη μεροληψία έναντι εσφαλμένων αποτελεσμάτων (The Bias-Variance Tradeoff). Από την άλλη μεριά, οι σύνθετοι αλγόριθμοι εκμάθησης τείνουν να έχουν μεγαλύτερη διακύμανση στις προβλεψεις τους. Τόσο η απόκλιση όσο και η διακύμανση είναι μορφές σφάλματος της πρόβλεψης στη μηχανική μάθηση. Συνήθως, μπορούμε να μειώσουμε το σφάλμα από τη μεροληψία (bias), αλλά μπορεί να αυξήσουμε το σφάλμα από τη διακύμανση (variance) ή και αντίστροφα. Αυτός ο συνδυασμός μεταξύ μεροληψίας και διακύμανσης είναι μια βασική αρχή στην στατιστική και στη μηχανική μάθηση και επηρεάζει όλους τους αλγόριθμους μάθησης υπό επίβλεψη.

Η συνήθης πρακτική αντιμετώπισης αποτελείται από τα παρακάτω στάδια: αρχικά χωρίζεται το αρχικό μας σύνολο δεδομένων σε ξεχωριστά υποσύνολα εκπαίδευσης (train) και δοκιμών (test). Αυτή η μέθοδος ελέγχει το πόσο καλά το μοντέλο μας θα εφαρμοστεί σε νέα δεδομένα. Εάν το μοντέλο αποδώσει πολύ καλύτερα στο σύνολο εκπαίδευσης από ό, τι στο σύνολο δοκιμών, τότε πιθανόν να έχει υπεργενίκευση. Για παράδειγμα, θα είναι προβληματικό έναν το μοντέλο δώσει 99% ακρίβεια στο σύνολο εκπαίδευσης, αλλά μόνο 55% ακρίβεια στο σύνολο δοκιμών. Δεύτερον, χρησιμοποιείται ως σημείο αναφοράς και σύγκρισης ένα πολύ απλό μοντέλο. Στη συνέχεια, δοκιμάζονται πιο σύνθετοι αλγόριθμοι, και έχοντας ένα σημείο αναφοράς μπορεί να

διαπιστωθεί εάν η πρόσθετη πολυπλοκότητα αξίζει να χρησιμοποιηθεί και δίνει ποιοτικότερα αποτελέσματα. Εάν δύο μοντέλα έχουν συγκρίσιμες επιδόσεις, τότε γενικά επιλέγεται το πιο απλό. Η ανίχνευση της υπεργενίκευση είναι χρήσιμη, αλλά δεν επιλύει το πρόβλημα. Όπως αναφέρθηκε και παραπάνω, υπάρχουν ποικίλες μεθοδολογίες για την αποφυγή της και παρακάτω μελετώνται κάποιες από αυτές στις επόμενες παραγράφους.

3.3 Μέτρα απόδοσης

Σε ρεαλιστικά προβλήματα αυτό που συχνά έχει σημασία είναι η μείωση του κόστους εσφαλμένων κατηγοριοποιήσεων, και όχι η αύξηση του ρυθμού ακρίβειας. Για μια επιχείρηση, σημασία έχει η λήψη επικερδών αποφάσεων και όχι η διατύπωση πολλών επιτυχών προβλέψεων. Για την υποθετική τράπεζα του προηγούμενου παραδείγματος αυτό που έχει αξία είναι ο σωστός εντοπισμός των ύποπτων συναλλαγών των πιστωτικών καρτών των πελατών της.

Η αξία ενός μοντέλου συνίσταται στην ικανότητά του να προβλέπει την κλάση των άγνωστων παρατηρήσεων. Όταν ένα μοντέλο αφομοιώνει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, και δεν καταγράφει τις γενικότερες σχέσεις που περιγράφουν τα δεδομένα τότε παρουσιάζει υπερπροσαρμογή. Συχνά αυτό οφείλεται στην πολυπλοκότητα του συστήματος και σαν αποτέλεσμα έχουμε πολύ υψηλές αποδόσεις στο γνωστό σύνολο δεδομένων εκπαίδευσης και πολύ χαμηλές αποδόσεις σε νέα δεδομένα.

Όπως γίνεται σε πολλές ερευνητικές εφαρμογές, έτσι και σε αυτή, θα εξεταστούν πληθώρα κατηγοριοποιητών ώστε να συγκρίνουμε διαφορετικά μοντέλα για να γίνει η επιλογή του καλύτερου. Για την εκτίμηση της ικανότητας ενός συστήματος να κατατάσσει σωστά άγνωστες παρατηρήσεις υπάρχουν πολλές τεχνικές και στα πλαίσια της εργασίας παρατίθενται οι ακόλουθοι.

3.3.1 Μέθοδος Holdout

Το σύνολο δεδομένων διασπάται σε δύο υποσύνολα όπου το πρώτη θα χρησιμοποιηθεί ως σύνολο εκπαίδευσης και το δεύτερο ως σύνολο επικύρωσης. Κατά την επαλήθευση καταγράφονται οι προβλέψεις του μοντέλου με την πραγματική κλάση της παρατήρησης. Μια κλασσική πρακτική είναι να χρησιμοποιούνται τα δύο τρίτα του αρχικού συνόλου ως σύνολο εκπαίδευσης και το ένα τρίτο ως σύνολο επικύρωσης. Η επίδοση του μοντέλου είναι το ποσοστό

των σωστών προβλέψεων. Επίσης, μια συνήθης παραλλαγή της μεθόδου είναι η τυχαία υποδειγματοληψία (random subsampling) όπου πρακτικά γίνεται πολλές φορές επανάληψη της μεθόδου holdout όπου σε κάθε επανάληψη δημιουργούνται νέα σύνολα εκπαίδευσης και επικύρωσης, κάνοντας πάντα τυχαία δειγματοληψία.

3.3.2 Μέθοδος Leave-one-out

Η τεχνική αυτή ακολουθεί το σκεπτικό της μεθόδου cross validation και ουσιαστικά το k=n όπου n= πλήθος παρατηρήσεων οι οποίες απαρτίζουν το σύνολο των δεδομένων. Για κάθε μια παρατήρηση το μοντέλο εκπαιδεύεται από τις υπόλοιπες n-1 παρατηρήσεις και επικυρώνει τη μέθοδο κάνοντας χρήση την επιλεγμένη παρατήρηση. Η διαδικασία επαναλαμβάνεται n φορές και τελικά υπολογίζεται το ποσοστό των σωστών παρατηρήσεων.

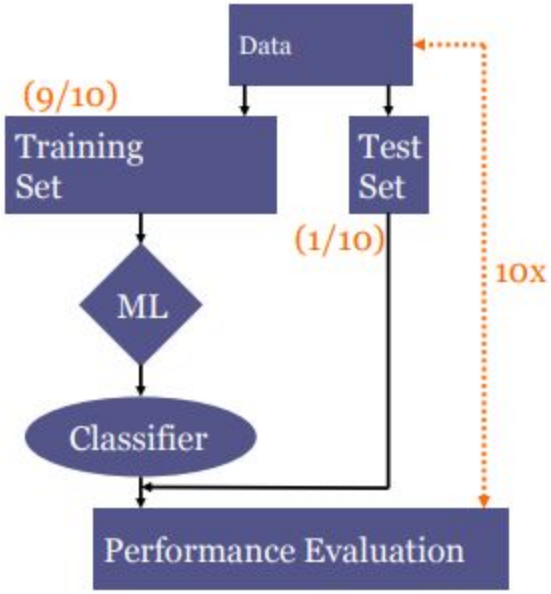
3.3.3 Μέθοδος Bootstrap

Στη μέθοδο bootstrap δημιουργούνται πάλι πολλαπλά σύνολα επικύρωσης με δειγματοληψία αλλα η διαφορά έγκειται στο γεγονός ότι η δειγματοληψία γίνεται με επανατοποθέτηση. Κάθε παρατήρηση που επιλέγεται να συμμετάσχει στο σύνολο επικύρωσης δεν αφαιρείται από το αρχικό δείγμα. Με αυτόν τον τρόπο, μια παρατήρηση μπορεί να επιλεγεί περισσότερες από μία φορές για να συμμετάσχει στο ίδιο σύνολο επικύρωσης. Σε γενικές γραμμές οι έρευνες έχουν δείξει ότι μεγαλύτερη επιτυχία και αποτελεσματικότητα έχει η μέθοδος cross validation σε σχέση με τις υπολοιπες η οποία περιγράφεται αναλυτικότερα παρακάτω.

3.3.4 Cross validation

Η μέθοδος cross validation είναι ένα ισχυρό προληπτικό μέτρο κατά της υπεργενίκευσης. Από το αρχικό σύνολο εκπαίδευσης δημιουργούνται υπό σύνολα δεδομένων εκπαίδευσης. Ο τρόπος της δημιουργίας αυτών των υποσυνόλων προϋποθέτουν μια εναλλαγή στο διαχωρισμό του αρχικού συνόλου. Στην τυπική k-fold cross validation, τα δεδομένα χωρίζονται σε k υποσύνολα. Στη συνέχεια, εκτελείται με προσοχή ο αλγόριθμος στα k-1 σύνολα δεδομένων ενώ χρησιμοποιούμε το εναπομείναν υποσύνολο για τη διαδικασία της δοκιμής (testing). Έτσι επιτυγχάνεται και διατηρείται το σύνολο δοκιμών ως ένα πραγματικά αόρατο, ως προς τη

διαδικασία εκπαίδευσης, σύνολο δεδομένων για την επιλογή του τελικού μοντέλου. Η παρακάτω εικόνα εξηγεί σχηματικά τη διαδικασία.



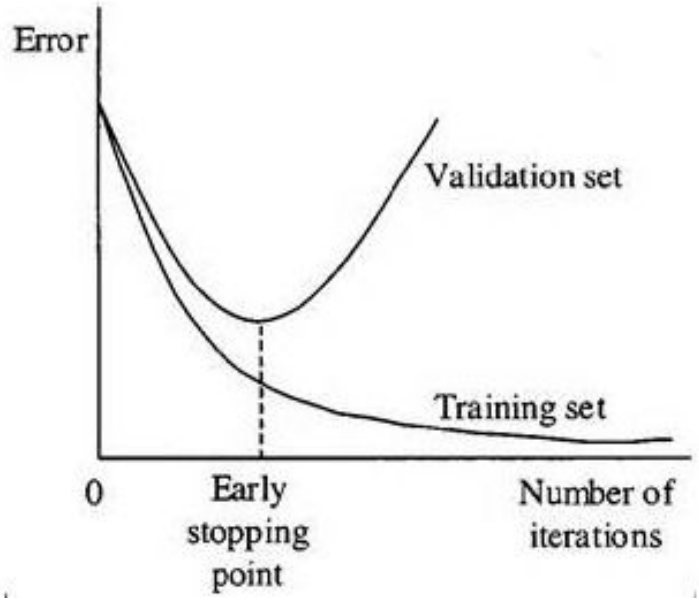
Εικόνα 5: Cross Validation

Αφαίρεση χαρακτηριστικών.

Κάποιοι αλγόριθμοι έχουν την δυνατότητα να αφαιρούν αυτόματα άσχετα δεδομένα εισόδου. Για όσους δεν το κάνουν, ο χρήστης μπορεί να τα αφαιρέσει χειροκίνητα ώστε να βελτιώσει τον τρόπο με τον οποίο θα γενικεύσει το μοντέλο. Ένας ενδιαφέρον τρόπος για να γίνει η προκείμενη αφαίρεση των θορυβωδών δεδομένων εισόδου είναι να εξεταστεί το κατά πόσο τα υπό επεξεργασία δεδομένα εισόδου είναι σχετικά και χρήσιμα στο μοντέλο που δημιουργείται. Με άλλα λόγια ελέγχεται εάν δίνουν ικανοποιητική απάντηση σε ερωτήματα που αφορούν το πρόβλημα. Ένας ακόμη τρόπος εντοπισμού τέτοιων δεδομένων συμβαίνει όταν κάτι δεν έχει νόημα ή αν είναι δύσκολο να δικαιολογήσει την ύπαρξή του.

3.3.5 Early stopping

Όταν εκπαιδεύεται ένας αλγόριθμος εκμάθησης επαναληπτικά, μπορεί να μετρηθεί πόσο καλά εκτελείται κάθε επανάληψη του μοντέλου. Αυτό είναι χρήσιμο διότι μέχρι έναν ορισμένο αριθμό επαναλήψεων, οι νέες επαναλήψεις βελτιώνουν το μοντέλο. Μετά από αυτό το σημείο, ωστόσο, η ικανότητα του μοντέλου να γενικεύει μπορεί να αποδυναμώσει καθώς ξεκινά να υπερισχύει των δεδομένων εκπαίδευσης. Για το λόγο αυτό η έγκαιρη διακοπή αφορά τη διακοπή της διαδικασίας κατάρτισης πριν ο αλγόριθμος περάσει εκείνο το σημείο. Στην εικόνα 6 δίνεται οπτικά η διαδικασία.



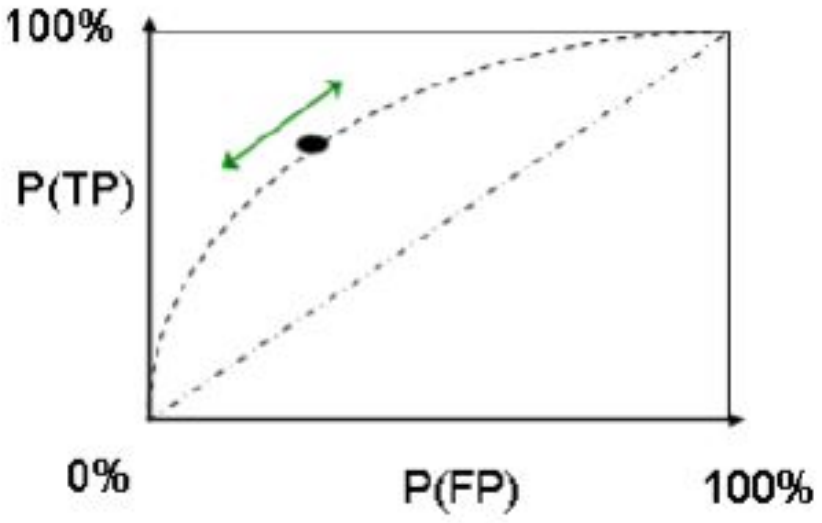
Εικόνα 6: Early Stopping

Γενικότερα υπάρχει ένα ευρύ φάσμα τεχνικών που αναγκάζουν τεχνητά ένα μοντέλο να είναι απλούστερο και η μέθοδος που εφαρμόζεται εξαρτάται από τον τύπο του αλγορίθμου που χρησιμοποιείτε. Για παράδειγμα, μπορεί να κλαδευτεί ένα δέντρο αποφάσεων, να γίνει dropout σε ένα νευρωνικό δίκτυο ή να προστεθεί μια παράμετρος ποινής στη συνάρτηση κόστους στην παλινδρόμηση.

3.3.6 Καμπύλη ROC

Η καμπύλη ROC (Receiver Operating Characteristics) είναι μια γραφική παράσταση που δείχνει την απόδοση ενός μοντέλου κατηγοριοποίησης συναρτήσει της αλλαγής του κατωφλίου διαχωρισμού. Η επίδοση των ταξινομητών στον χώρο ROC συμβολίζεται με καμπύλη. Είναι ιδιαίτερα χρήσιμη επειδή παρέχει μία ορατή αναπαράσταση των σχετικών διαφορών μεταξύ των πλεονεκτημάτων (από τις αληθείς θετικές κατηγοριοποιήσεις) και μειονεκτημάτων/κόστη (εσφαλμένες θετικές κατηγοριοποιήσεις) της ταξινόμησης σε σχέση με τις κατανομές δεδομένων. Η καμπύλη αυτή δείχνει τα TPR (TruePositiveRate) στον άξονα των y προς τα FPR (False Positive Rate) στον άξονα των x και η απόδοση κάθε ταξινομητή αναπαρίσταται σαν ένα σημείο της καμπύλης. Στην εικόνα 7 δίνεται ένα υπόδειγμα.

Όπου TPR = TP/(TP +FN) (θετικές τιμές) και FPR = FP / (TN +FP) (αρνητικές τιμές)



Εικόνα 7: Καμπύλη ROC

Με άλλα λόγια η τεχνική αξιολόγησης με καμπύλες ROC χρησιμοποιεί την αναλογία δύο απλών στηλών που βασίζονται σε μετρικές αξιολογήσεις, δηλαδή την αληθή-θετική τιμή (TP rate) και την εσφαλμένη θετική τιμή (FP rate), οι οποίες ορίζονται με τους παρακάτω τύπους:

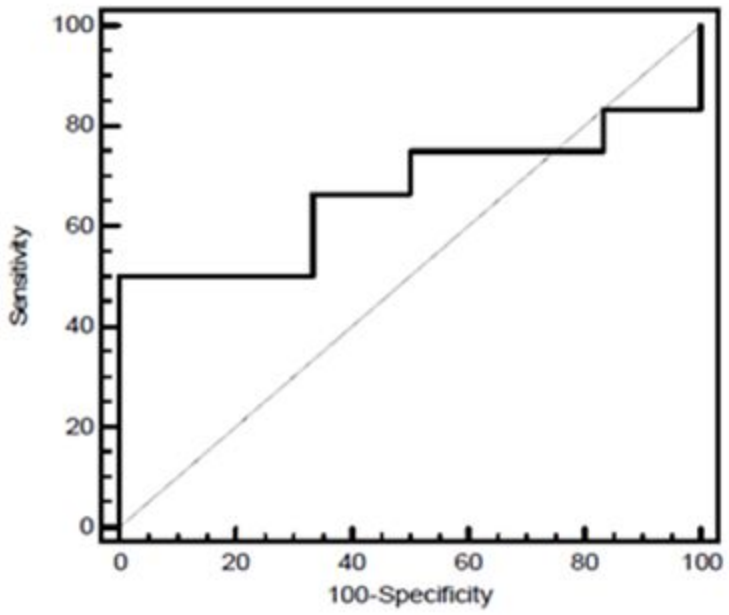
$$TP\ RATE\ =\ \frac{TP}{PC}$$

$$FP\ RATE\ =\ \frac{FP}{NC}$$

Έτσι για να συγκρίνουμε τους ταξινομητές μας θα χρησιμοποιήσουμε το μέτρο σύγκρισης που είναι η περιοχή κάτω από την καμπύλη ROC που είναι ευρέως γνωστή ως Area Under ROC Curve (AUC). Η AUC εκφράζει το ποσοστό του χώρου που βρίσκεται κάτω από την καμπύλη,

και παίρνει τιμές από 0 έως 1. Όσο πιο κοντά στο 1 είναι οι τιμές τόσο καλύτερο το αποτέλεσμα μας.

Η διαγώνια γραμμή τυχαίας πρόβλεψης έχει AUC = 0,5. Συνεπώς, κάθε ταξινομητής που είναι καλύτερος της τυχαίας πρόβλεψης θα έχει AUC > 0,5. Όσο μεγαλύτερο το AUC για έναν ταξινομητή τόσο καλύτερος είναι. Το παρακάτω γράφημα δίνει σχηματικά την ιδέα της καμπύλης. Στην εικόνα 8 δίνεται ένα επίσης μια γραφική απεικόνιση της καμπύλης.



Εικόνα 8: Καμπύλη ROC

Η ακρίβεια και η τιμή σφάλματος δίνονται αντίστοιχα από τους τύπους:

Accuracy = (TP + TN) / (PC + NC)

Error Rate = 1 – Accuracy

Όπου, TP (True Positive), TN (True Negative) και PC, NC (Column Counts).

Αλλες μετρικές αξιολόγησης αποτελούν η ακρίβεια προσέγγισης (precision), η ανάκληση (recall), το μέτρο -F (F-measure) και ο μέσος όρος -G (G-mean). Οι παράγοντες αυτοί ορίζονται παρακάτω αντίστοιχα:

Precision =TP / (TR+FP)

Recall =TP / (TP +FN)

F- Measure = $[(1+ b)^2 *Recall * Precision] /[b^2*Recall + Precision]$

G- Mean = $[TP /(TP+ FN) *TN /(TN + FP)]^{1/2}$,

Όπου FN (False Negatives), FP (False Positives) και b ένας συντελεστής για τον καθορισμό της σχετικής σημασίας της ακρίβειας προσέγγισης έναντι της ανάκλησης (Guoand Viktor, 2004; Weiss, 2004, Provostetal. 1998; Sunetal. 2007).

Συνοπτικά αναφέρεται παρακάτω η περιγραφή της κάθε μετρικής.

Accuracy = αριθμός σωστών προβλέψεων / σύνολο προβλέψεων . μετρά πόσο συχνά ο κατηγοριοποιητής κάνει σωστή πρόβλεψη.

Precision = σωστές θετικές / σωστές θετικές + λάθος θετικές. Μετράει πόσες είναι οι σωστές απο τις θετικές προβλέψεις. Επικαιροποιεί την ικανότητα του κατηγοριοποιητή να να μην κατηγοριοποιεί ένα αρνητικό παράδειγμα ως θετικό.

Recall = σωστές θετικές / σωστές θετικές + λάθος αρνητικές

4. Ανάλυση ναυτικών ατυχημάτων

4.1 Ορισμός προβλήματος - πεδίο έρευνας

Ο βασικός στόχος που αλγορίθμου είναι να προβλέψει με ακρίβεια τη σημαντικότητα ενός περιστατικού που θέτει σε κίνδυνο την υγεία και τη σωματική ακεραιότητα ενός ναυτικού κατά τη διάρκεια ενός ταξιδιού ποντοπόρου πλοίου.

Κατά τη διάρκεια ενός ταξιδιού παντός είδους εμπορικού πλοίου, υπάρχουν πολλά περιστατικά που μπορούν να αποτελέσουν πρόβλημα στην ομαλή διεξαγωγή και ολοκλήρωση της αποστολής. Ακράδαντα παραδείγματα αποτελούν τα μηχανικά προβλήματα, οι κακές καιρικές συνθήκες που καθυστερούν το χρονοδιάγραμμά του ταξιδιού, προβλήματα ή ατυχήματα στο κύτος του πλοίου, καταστροφή ή φθορά του φορτίου, καθυστερήσεις λόγω ανεφοδιασμού καθώς και ατυχήματα και ασθένειες που πιθανόν να προσβάλουν το πλήρωμα του πλοίου. Η παρούσα εργασία ασχολείται με τη πρόβλεψη και την ανάλυση των περιστατικών υγείας που μπορεί να προκύψουν. Πιο συγκεκριμένα μέριμνα της μελέτης είναι να δημιουργηθεί ένα σύστημα ικανό να αξιολογεί τη σημαντικότητα των περιστατικών που θα κλονίσουν την υγεία ενός ναυτικού, τον οικονομικό αντίκτυπο που μπορεί να έχει η αντιμετώπιση του περιστατικού, τις νομικές αξιώσεις που μπορεί να προκύψουν, καθώς επίσης και την επίπτωση που θα έχει στην ομαλή συνέχιση του ταξιδιού. Κάθε σύγχρονη ναυτιλιακή εταιρεία διαθέτει πληροφοριακό σύστημα ικανό να αποθηκευσει και να επεξεργαστεί πληθώρα δεδομένων που αφορούν όλες τις πτυχές της.

Για να γίνει περισσότερο αντιληπτή η αναγκαιότητα ενός μοντέλου πρόβλεψης θα πρέπει να σημειωθεί και να γίνει η αντιπαραβολή συγκρίνοντας τον τρόπο αντιμετώπισης ενός ατυχήματος ή μιας ασθένειας από ιατρικής άποψης, όταν αυτό συμβαίνει στη στεριά. Αρχικά όπως είναι αντιληπτό, και βάση κοινής λογικής, όταν σε έναν άνθρωπο συμβαίνει κάτι ιατρικής φύσης, η πρώτη αναφορά γίνεται σε έναν γενικό ιατρό εάν δε μπορεί ο ασθενής να καθορίσει τον πόνο και το σημείο εντοπισμού της όχλησης. Ωστόσο εάν αυτό είναι αντιληπτό, λόγω παραδείγματος χάριν πρότερης εμπειρίας με παρόμοιο σύμπτωμα, ο ασθενής μπορεί να απευθυνθεί σε ένα γιατρό συγκεκριμένης ειδικότητας. Η εξέταση γίνεται δια ζώσης και το πόρισμα μπορεί να δοθεί

σχετικά άμεσα. Η έναρξη της θεραπείας μπορεί επίσης να ξεκινήσει σε σύντομο χρονικό διάστημα βάση της άμεσης πρόσβασης σε διαγνωστικό κέντρο, νοσοκομείο ή άλλο ίδρυμα αντιμετώπισης ασθενειών (φυσιοθεραπευτήρια, κέντρα αποκατάστασης, κλπ.) Ωστόσο, όπως μπορεί να γίνει και πάλι εύκολα αντιληπτό, οι προαναφερθείσες συνθηκες και πρακτικές δεν είναι εφαρμόσιμες στη θάλασσα. Ένα μεγάλο μέρος της ναυτιλιακής βιομηχανίας (εταιρίες, τμήματα ναυτιλιακών επιχειρήσεων) ασχολείται αποκλειστικά με θέματα ασφάλειας εν πλω με συνεχή έρευνα, εξέλιξη και αναδιαμόρφωση των κανονισμών ασφαλείας, ανάλογα με τα περιστατικά και τα αποτελέσματα που συμβαίνουν κατά τη διάρκεια των ταξιδιών. Πέραν των ίδιων των εταιριών που μεριμνούν για αυτά τα θέματα, υπάρχουν και οργανισμοί που είναι υπεύθυνοι για την εποπτεία και τη συμμόρφωση των ναυτιλιακών εταιρειών αλλά και όλων των μερών που εμπλέκονται με το διεθνές εμπόριο. Μέρος λοιπόν των λειτουργιών τους είναι η επίβλεψη των συνθηκών εργασίας πάνω σε ένα εμπορικό πλοίο με σκοπό τη βελτίωση του τρόπου ζωής και εργασίας πάνω στο πλοίο και την αποφυγή ατυχημάτων. Επίσης για τα ζητήματα αυτά υπάρχει πληθώρα βιβλιογραφία ανατρέχοντας σε ακαδημαϊκά συγγράμματα ναυτιλιακών σπουδών. Τα δεδομένα της παρούσας εργασίας έχουν προκύψει από ανώνυμη πηγή διαθέσιμα σε διαδικτυακό τόπο. Έχοντας λοιπόν ορίσει την κεντρική ιδέα και το πεδίο της έρευνας, στις επόμενες παραγράφους θα παρουσιαστούν τα βασικά χαρακτηριστικά σχετικά με τη δομή, τις διαδικασίες και το περιβάλλον εφαρμογής της ναυτιλιακής δραστηριότητας με σκοπό να συνδεθεί η ανάλυση των ναυτικών ατυχημάτων με τη χρήση μηχανικής μάθησης.

4.2 Ορισμός ναυτιλιακής εταιρείας και εμπορικού πλοίου

Οι ναυτιλιακές επιχειρήσεις είναι πολύπλοκοι οργανισμοί που καλούνται να λειτουργήσουν σε ένα ιδιαίτερα σύνθετο περιβάλλον το οποίο δεν έχει γεωγραφικά όρια και ούτε τις ίδιες συνθήκες λειτουργίας με τις εταιρείες που βρίσκονται στη στεριά. Οι ιδιαιτερότητες λοιπον αυτού του περιβάλλοντος αναγκάζουν τις επιχειρήσεις να επιλέγουν μεθόδους και επιλογές που ενισχύουν την ανταγωνιστικότητα τους όπως η διαφοροποίηση του θεσμικού φορολογικού και νομικού πλαισίου στο οποίο λειτουργούν, η ποικιλία του ανθρώπινου δυναμικού από χώρες πέραν του τόπου λειτουργίας, η τοποθέτηση των κεντρικών γραφείων και αποθηκών και άλλα πολλά. Όλοι αυτοί οι παράγοντες οδηγούν αναπόφευκτα σε προσαρμογές που αφορούν την οργάνωση και τη διοίκηση των επιχειρήσεων προκειμένου να αναζητηθούν λύσεις για την

βελτίωση της αποτελεσματικότητας και της αποδοτικότητάς τους. Το περιβάλλον λειτουργίας των εταιρειών αυτών μεταβάλλεται διαρκώς για πολλούς και διάφορους λόγους. Ένας από αυτούς είναι η παγκοσμιοποίηση η οποία τα τελευταία χρόνια παρουσιάζεται με μεγαλύτερη ένταση και έχει αναδιαμορφώσει τις πτυχές του εμπορίου και σε συνέχεια την οργάνωση και διοίκηση της επιχείρησης τόσο σε διοικητικό όσο και σε εμπορικό επίπεδο. Σε κάθε περίπτωση όμως, η ποιότητα της παραγόμενης υπηρεσίας και η επικέντρωση στον ανθρώπινο παράγοντα είναι οι παράμετροι που συνδέεται άμεσα με την αποδοτικότητα και την επιτυχία. Στο νέο πλαίσιο λειτουργίας, οι ναυτιλιακές εταιρείες καλούνται να δώσουν λύσεις για την καλύτερη προσαρμογή τους στα νέα δεδομένα και την αύξηση της αποτελεσματικότητας σε νέες συνθήκες.

Κατά το άρθρο 1 παρ. 1 του Κώδικα Ιδιωτικού Δικαίου (Κ.Ι.Ν.Δ) πλοίο είναι κάθε σκάφος, χωρητικότητας καθαράς τουλάχιστον δέκα κόρων προορισμένο να κινείται αυτοδυνάμως στη θάλασσα. Όσον αφορά την αυτοδύναμη κίνηση, ο νόμος αναφέρει ότι το πλοίο θα πρέπει να διαθέτει τα δικά του μέσα πλεύσεως και προώσεως ανεξάρτητα με το πώς παράγεται η ενέργεια που τροφοδοτεί την πλεύση.

4.3 Αξιωματικοί και πλήρωμα εμπορικού ναυτικού

Ο πλοίαρχος είναι ο επικεφαλής του πληρώματος και λόγω της θέσης και των υποχρεώσεών που ορίζονται από το νόμο, συγκεντρώνει πολλές και διάφορες αρμοδιότητες κατα την διαρκεια της πλοιαρχίας του. Ο πλοίαρχος είναι υπεύθυνος για την ολοκλήρωση ενός ταξιδιού και λαμβάνει όλα τα αναγκαία μέτρα για την επίλυση προβλημάτων και για την τήρηση της απαιτούμενης ασφάλειας του πληρώματος. Είναι σε θεση να εκπροσωπήσει τα πρόσωπα που συνδέονται άμεσα σε νομικό επίπεδο με τη ναυτιλιακή. Στην περίπτωση λοιπόν ενός ναυτικού ατυχήματος η αναζήτηση των ευθυνών ξεκινά από τον πλοίαρχο και είναι εκείνος που θα λάβει την πρωτοβουλία για την επίλυση του ζητήματος σε συνεννόηση με το κέντρο επιχειρήσεων της ναυτιλιακής που βρίσκεται στη στεριά.
Το πλήρωμα ενός πλοίου είναι οι εργαζόμενοι του πλοίο, ανήκουν στο Προσωπικό καταστρώματος, και κατά την διάρκεια του πλου εκτελούν τις ακόλουθες εργασίες:

Αρχικά υπάρχουν οι εργασίες γεφύρας κατά φυλακές, δηλαδή βάρδιες (συνήθως τετραωρίες) σε κάθε μια των οποίων συμμετέχει ένας Αξιωματικός (Υποπλοίαρχος ή Ανθυποπλοίαρχος), ένας ναύτης πηδαλιούχος και ένας ναύτης οπτήρας. Ο Αξιωματικός Γεφύρας, με την είσοδό του στη Γέφυρα, ενημερώνεται αμέσως από τον αποχωρούντα Αξιωματικό για οτιδήποτε κρίνεται απαραίτητο, για την θέση του πλοίου, και στη συνέχεια προσέχει για την σωστή λειτουργία των φανών ναυσιπλοΐας, την τήρηση της πορείας του πλοίου, Ο Πηδαλιούχος με την είσοδό του στη Γέφυρα ενημερώνεται για την ακολουθητέα πορεία του πλοίου, επαναλαμβάνοντας και επιβεβαιώνοντας αυτή, την οποία και συνεχίζει να ακολουθεί εκτελώντας τις εντολές του Αξιωματικού Γεφύρας στους ακριβείς χειρισμούς του πηδαλίου. Ο οπτήρας ναύτης με την είσοδό του στη Γέφυρα ενημερώνεται και αυτός σχετικά από τον προηγούμενο για τυχόν στόχους, ή άλλου συμβάντος, στη συνέχεια παρακολουθεί την πορεία του πλοίου, τον γύρω θαλάσσιο χώρο, τους πλοϊκούς φανούς, αναφέροντας στον Αξιωματικό Γεφύρας παν ότι αντιληφθεί. Φυσικά υπάρχουν και οι γενικές εργασίες του σκάφους. Εκτός των παραπάνω απασχολουμένων οι υπόλοιποι εκτός βαρδιών ασχολούνται με καθαρισμούς, λιπάνσεις, σφυροκρούσεις, βαφές, προετοιμασία κυτών για επικείμενη φόρτωση κ.λπ.
Κατά την παραμονή του πλοίου εντός λιμένα ή σε όρμο ή σε αγκυροβόλιο, το προσωπικό καταστρώματος ασχολείται, επίσης κατά βάρδιες, σε γενικές εργασίες, καθώς και με την φύλαξη του σκάφους και τον έλεγχο του περί την αγκυροβολία θαλάσσιου χώρου.

Απο τα παραπάνω συμπεραίνουμε ότι κάθε μια θέση αποτελεί μια ξεχωριστή λειτουργία του πλοίου και είναι απαραίτητη για την ομαλή πορεία του ταξιδιού. Επίσης αξίζει να αναφέρουμε ότι κάθε θέση είναι αναντικατάστατη (ειδικά όσο οι βαθμοί των αξιωματικών μεγαλώνουν) και για το λόγο αυτό εάν ένας αξιωματικός ή ναύτης του πλοίου τεθεί εκτός των καθηκόντων του κάποιο άλλο μέλος του πληρώματος θα πρέπει να αντικαταστήσει το πόστο του και τη βάρδιά του έως ότου η διαχειρίστρια εταιρεία στείλει τον αντικαταστάτη. Η εταιρεία έχει ως μεγαλύτερη παράμετρο απόφασης για τον αριθμό των αξιωματικών / ναυτικών που θα ναυτολογηθούν, το νόμο και τα διεθνή πρότυπα στελέχωσης πλοίων που ορίζονται από τους σχετικούς οργανισμούς (IMO - International Marine Organization) και ο αριθμός των αξιωματικών και πληρώματος δε μπορεί να είναι μικρότερος αυτών. Η εταιρία επίσης μπορεί να αποφασίσει να στελεχώσει και με επιπλέον άτομα ένα πλοίο (π.χ. Να έχει δύο καπετάνιους) αλλά αυτό σπάνια γίνεται στην πράξη και όταν συμβαίνει είναι κυρίως για εκπαιδευτικούς σκοπούς. Βέβαια αξίζει να σημειωθεί ότι υπάρχουν και συγκεκριμένα πλοίο που απαιτούν μεγαλύτερο αριθμό

πληρώματος σε σχέση με άλλους τύπους πλοίων όπως είναι για παράδειγμα τα βαπόρια που μεταφέρουν υγροποιημένο φυσικό αέριο και απαιτούν σχεδόν τον διπλάσιο αριθμό πληρώματος σε σχέση με τα βαπόρια μεταφοράς χύδην φορτίου. Ωστόσο η δεύτερη παράμετρος για την απόφαση αυτή είναι το οικονομικό κόστος και όπως είναι φυσιολογικό τίθεται στον οικονομικό προγραμματισμό. Ωστόσο οι περισσότερες ναυτιλιακές δεν αντιμετωπίζουν τους μισθούς ως ένα απλό διαχειριστικό κόστος που πρέπει διαρκώς να μειώνεται και να ελαχιστοποιηθεί, καθώς αναγνωρίζουν τη σημαντικότητα των ανθρώπων και έτσι οι μισθοί αντιμετωπίζονται ώς επένδυση από την οποία περιμένεις κέρδος. Φυσικά προς αποφυγή παρερμηνείας αναφέρεται πως η προηγούμενη πρόταση βρίσκει ως αποδέκτες πολλές επιχειρήσεις και πολλούς κλάδους της οικονομίας οι οποίοι αναγνωρίζουν την υπεραξία του ανθρώπινου παράγοντα. Ο λόγος λοιπόν που γίνεται ειδική μνεία είναι λόγω της ιδιαιτερότητας και των αντικειμενικών δυσκολιών του ναυτικού επαγγέλματος όπως είναι η απόσταση από την οικογένεια και την πατρίδα, τα μακρινά ταξίδια έως το επόμενο λιμάνι, εργασία τις κυριακές, αργίες και άλλα. Από τα παραπάνω γίνεται αντιληπτό ότι η χρήση του μοντέλου πρόβλεψης έρχεται να βοηθήσει στο πρόβλημα της αντικατάστασης του ναυτικού που νοσεί και μπορεί να δώσει άμεσα απάντηση στο αν είναι συνετό απο οικονομικής και ιατρικής άποψης να αντικαταστήσει η ναυτιλιακή τον συγκεκριμένο ναυτικό. Αν δηλαδή το αποτέλεσμα του μοντέλου δείξει μεγάλη επικινδυνότητα για την εξέλιξη του συμβάντος η διαχειρίστρια εταιρεία από τη στεριά μπορεί να συμβουλεύσει τον καπετάνιο να πράξει τα δέοντα σχετικά με την αντικατάσταση του ναυτικού που νοσεί παρόλου που φαινομενικά η κατάσταση του μπορεί να δείχνει επιλύσιμη εντός του πλοίου.

4.4 Θαλάσσιοι κίνδυνοι και πρωτόκολλο ασφαλείας

Τα ναυτικά ατυχήματα πολλές φορές προξενούνται έπειτα από ανθρώπινο σφάλμα. Λίγες είναι οι περιπτώσεις όπου ελαττωματικά εργαλεία ή μηχανήματα έχουν δημιουργήσει το πρόβλημα και ακόμα πιο λίγες είναι εκείνες όπου το αίτιο βρίσκεται στον καιρό ή σε κάποια άλλη εξωγενή δύναμη. Ωστόσο έχουν συμβεί και πολύ μεγάλα εμπορικά ατυχήματα τα οποία προκλήθηκαν από ελαττωματική κατασκευή, αλλά αυτά κυρίως συνέβαιναν στο παρελθόν. Όπως και σε πολλούς τομείς της οικονομίας τα τελευταία χρόνια με τη ραγδαία τεχνολογική εξέλιξη οι μηχανές και οι μέθοδοι συντηρησης έχουν διαμορφώσει ένα πολύ πιο ασφαλες περιβάλλον λειτουργίας σε σχέση με το παρελθόν. Για το λόγο αυτό η σύγχρονη ναυτιλιακή βιομηχανία έχει επιστήσει την προσοχή στην προστασία του ανθρώπου μέσα από διάφορες δράσεις. Η

διαδικασία που ακολουθείται σε περίπτωση εμφάνισης περιστατικών υγείας είναι η ακόλουθη. Αρχικά, ο καπετάνιος ενημερώνει το γιατρό της εταιρείας για το πρόβλημα που αντιμετωπίζει και κοινοποιεί την πληροφορία αυτή στο τμήμα ανθρώπινου δυναμικού πληρωμάτων (crew department) στο τμήμα επιχειρήσεων (operations department) και συνήθως στο τμήμα HSQE (Health Security Quality Environment). Ύστερα ανάλογα με τη σημαντικότητα του περιστατικού ενημερώνονται και τμήματα όπως είναι το τμήμα ασφαλειών (insurance department) και αν είναι απαιτητό ακόμα και η νομική διεύθυνση (legal department). Ο γιατρός λοιπόν μετά την πρώτη αναγγελία αναλαμβάνει τη θεραπεία και δίνει οδηγίες στον καπετάνιο σχετικά με τη περίθαλψη τα φάρμακα και τις εργασίες που μπορεί να εκτελέσει ο εν λόγω ναυτικός.

Ο ρόλος του γιατρού σε αυτή τη φάση είναι πολύ σημαντικός καθώς το πλοία διαθέτουν ευρύ φάσμα φαρμάκων που μπορουν να καταστείλουν ή να καθυστερήσουν την επιδείνωση του περιστατικού. Έτσι λοιπόν, ανάλογα με την πορεία του ασθενή κρίνεται εάν θα επισκεφθεί κάποιον γιατρό στο επόμενο λιμάνι. Η διάγνωση από το γιατρό του λιμανιού που βρίσκεται ενδιάμεσα του τελικού προορισμού είναι πολύ σημαντική για δύο κυρίως λόγους. Πρώτον γιατί η αρχική διάγνωση δε γίνεται δια ζώσης αλλά από τηλέφωνο, εικόνες και e-mail και ο δεύτερος είναι διότι δίνεται η δυνατότητα να εξεταστεί ο ναυτικός από ιατρό ειδικότητας (υπο την έννοια ότι ο γιατρός της εταιρείας που δίνει την πρώτη εκτίμηση είναι ως επί το πλείστον παθολόγος). Έτσι για παράδειγμα εάν υπάρχει ένα ορθοπεδικό ζήτημα ή ένα ατύχημα από πτώση, ο πλέον κατάλληλος γιατρός είναι ένας χειρουργός ορθοπεδικός. Η διάγνωση του γιατρού που βρίσκεται στο λιμάνι σταθμού είναι επίσης πολύ σημαντική διότι αποφασίζει αν ο ναυτικός είναι σε θέση να εκτελέσει τα καθηκοντα του (fit for duty / unfit for duty) ή αν πρεπει να τεθεί εκτός εργασίας. Επίσης αποφασίζει αν θα μπορεί να επαναπατριστεί με αεροπλάνο σε περίπτωση που δεν είναι ικανός προς εργασία (unfit for duty), πόσες μέρες ή ώρες θα πρέπει να παραμείνει εκτός εργασίας, ενώ παράλληλα να είναι εντός του πλοίου μέχρι να αναρρώσει πλήρως και άλλες τέτοιες λεπτομέρειες. Όλες αυτές οι λεπτομέρειες είναι σημαντικές διότι δε γνωρίζει κανείς με βεβαιότητα αν θα υποτροπιάσει ένα φαινομενικά απλό συμβάν με αποτέλεσμα όλα τα παραπανω να αποτελούν πλέον απαραίτητα πειστήρια που θα κοινοποιηθούν στους διάφορους φορείς ασφάλισης που χρησιμοποιεί η ναυτιλιακή για να αποζημιώσει τα περιστατικά της και φυσικά για τις αντίστοιχες κρατικές αρχές που πιθανόν να ελέγξουν το συμβάν (αν είναι εργατικό ατύχημα κλπ). Απο τη στιγμή που θα συμβει η θα αναγγελθεί ένα συμβάν ώς σοβαρό θα ειδοποιηθεί και το αντίστοιχο P&I Club.

4.5 Κόστος περίθαλψης -όφελος πρόβλεψης δαπανών

Το κόστος περίθαλψης ποικίλει ανάλογα με το γεωγραφικό μέρος, την ηλικία του ναυτικού, την ασθένεια, τα διαθέσιμα ιατρικά μέσα θεραπείας και της απόφασης που θα παρθεί από το κέντρο επιχειρήσεων της εταιρείας σχετικά με το πλάνο θεραπείας. Βλεποντας τις παραμέτρους με αντικειμενικότητα η μεταβλητή που επηρεάζει περισσότερο το κόστος περίθαλψης είναι το μέρος στο οποίο συμβαίνει το γεγονός. Κάθε τόπος έχει τη δική του οικονομία και οι τιμές της χώρας κυμαίνονται σε αντίστοιχα επίπεδα. Αυτό φυσικά επηρεάζει τις δαπάνες και τις τιμές που σχετίζονται με την υγεία. Το κόστος αντιμετώπισης μιας κήλης στην Αμερική είναι τελείως διαφορετικό από το κόστος που θα είχε το ίδιο περιστατικό στην Αφρική. Ως δεύτερη σημαντικότερη παράμετρος είναι η απόφαση της διοίκησης της εταιρείας για το πλάνο θεραπείας. Το που και πότε θα συμβεί ένα ατυχές περιστατικό ατυχήματος ή ασθένειας δεν είναι σε θέση να το γνωρίζει κανείς καθώς βρίσκεται αποκλειστικά στη σφαίρα της τυχαιότητας. Αυτό που μπορεί λοιπον να κάνει ο δέκτης του περιστατικού αυτού, που στην περίπτωσή μας ειναι μια ναυτιλιακή εταιρεία, είναι να οργανώσει ένα αποτελεσματικό σύστημα αντιμετώπισης των κινδύνων που θα βασίζεται στην κάθε περίπτωση ξεχωριστά. Στα πλαίσια λοιπόν της οργάνωσης και του οικονομικού σχεδιασμού η εταιρεία που γνωρίζει εκ των προτέρων απο προηγούμενη εμπειρία πόσο θα δαπανήσει για μια ανάλογη περίπτωση είναι σε θέση να διαμορφώσει το πρόγραμμα της περίθαλψης κάνοντας κινήσεις από την αρχή. Αυτό είναι σημαντικό διότι η εμπειρία έχει δείξει ότι πολλές φορές η ναυτιλιακή εταιρεία αποφασίζει να αναδιαμορφώσει το πρόγραμμα περίθαλψης όταν αυτό είναι ήδη σε προχωρημένο στάδιο. Ωστόσο εκείνη τη στιγμή δεν είναι εφικτό να συμβεί από πρακτική και ιατρικής άποψης καμία αλλαγή. Επίσης, έχει παρατηρηθεί ότι οργανώνονται διορθωτικές κινήσεις με σκοπό τη μείωση των εξόδων αφού όμως έχει λάβει έκταση ένα περιστατικό και έχει τραβήξει την προσοχή της διεύθυνσης. Κάτι τέτοιο όμως δεν είναι δυνατόν να συμβεί ειδικά σε περιπτώσεις όπου εμπλέκεται ο παράγοντας της ανθρώπινης υγείας υπό την έννοια ότι αν δε μπορεί για παράδειγμα να μετακινηθεί ο ασθενείς λόγω ιατρικών περιορισμών, ότι και να αποφασίσει η ανώτατη διοίκηση της εταιρείας ο ασθενής θα παραμείνει στο μέρος που του χορηγείται η θεραπεία. Ένα χαρακτηριστικό παράδειγμα είναι τα ξαφνικά σοβαρά ατυχήματα που συμβαίνουν. Τις περισσότερες φορές, τη στιγμή που συμβαίνει ένα περιστατικό, η πρώτη κίνηση είναι η άμεση διακομιδή του ασθενούς στο πλησιέστερο νοσοκομείο. Ωστόσο το πλησιέστερο νοσοκομείο ενδέχεται να μην είναι και το καταλληλότερο για τη συγκεκριμένη περίπτωση και

μάλιστα μπορεί να λείπουν ιατρικές ειδικότητες και βασικός ιατρικός εξοπλισμός. Αυτό μπορεί να έχει σαν αποτέλεσμα να επιμηκυνθει ο χρονος περιθαλψης, η εταιρεία να χρειαστεί να μεταφέρει τον απαραίτητο γιατρό στο μέρος της νοσηλείας καθώς ακόμα και να ξανα μεταφέρει τον ασθενή σε διαφορετικό μέρος όπου θα του παρέχονται όλα τα απαραίτητα μέσα για τη συνέχιση της θεραπείας. Όλα τα παραπάνω λοιπόν αποτελούν αστάθμητό παράγοντα για την εταιρεία και σαν αποτέλεσμα έχουν να αυξήσουν τα έξοδα θεραπείας αλλα και τις εργατοώρες των ανθρώπων που εργάζονται στη στεριά και οργανώνουν το ταξίδι του πλοίου.

5. Αποτελέσματα Ανάλυσης

5.1 Δεδομένα

Προτού αναφέρουμε τα ιδιαίτερα χαρακτηριστικά των τιμών εισόδου που περιγράφονται στον πίνακα 1.1. αξίζει να αναφέρουμε ότι υπάρχει αμφίδρομη και συνεχείς επικοινωνία μεταξύ των πλοίων που ταξιδεύουν και του κέντρου επιχειρήσεων στη στεριά. Αυτό συμβαίνει για λόγους που αφορούν τη βασική δραστηριότητα των ναυτιλιακών επιχειρήσεων και δε θα αναλυθεί περισσότερο στα πλαίσια αυτής της εργασίας. Αυτή λοιπόν η άμεση επαφή που υπάρχει μας επιτρέπει να ανανεώνουμε άμεσα τα δεδομένα εισόδου και να διευρύνεται το φάσμα των περιπτώσεων που μπορούν να προβλεφθούν καθώς επίσης και η ακρίβεια αυτών. Όπως μπορούμε να παρατηρήσουμε και από τον πίνακα, υπάρχουν δεδομένα που αφορούν τους ναυτικούς, το πλοίο, καθώς και τις συνθήκες. Διαπιστώνουμε ότι το πρόβλημα είναι πολύ παραγοντικό, με την έννοια ότι οι παράγοντες είναι ανεξάρτητοι μεταξύ τους και αποτελούν διαφορετικές μεταβλητές. Έτσι τα πρώτα πεδία του πίνακα αφορούν τη σειρά που αναφέρθηκε στην παραπάνω πρόταση. Μέρος της μελέτης μας θα είναι επίσης και η προσπάθεια απλοποίησης του μοντέλου μειώνοντας τα χαρακτηριστικά εισόδου. Σκοπός αυτής της κίνησης είναι να μειωθεί ο χρόνος πρόβλεψης που συστήματος και κυρίως να μειωθεί ο χρόνος που απαιτείται πρακτικά να αποσταλούν από το πλοίο οι παραπάνω πληροφορίες και να συλλεχθούν να επεξεργαστούν από αρμόδια τμήματα της ναυτιλιακής στη στεριά. Αυτό θα

μειώσει ουσιαστικά το χρόνο ανταπόκρισης που θα έχει η εταιρεία για το συμβάν και θα λάβει ταχύτερα αποφάσεις. Ωστόσο, για να συμβεί αυτό θα πρέπει να ελέγξουμε κατά πόσο βελτιώνεται το σύστημα με τη μείωση χαρακτηριστικών χωρίς να υπάρχει παράλληλα μείωση της ακρίβειας και των άλλων μετρικών αξιολόγησης. Στο προηγούμενο κεφάλαιο αναφέρθηκε ότι εμπειρικά έχει αποδειχθεί ότι τα ατυχήματα προκαλούνται κυρίως από τον ανθρώπινο παράγοντα. Ένας δεύτερος στόχος είναι να διερευνηθεί και η συσχέτιση των εξόδων ενός ατυχήματος / ασθένειας με τον ανθρώπινο παράγοντα και συγκεκριμένα εάν μπορεί να γίνει πρόβλεψη των εξόδων μόνο με τα δεδομένα που περιγράφουν τις λεπτομέρειες του ανθρώπινου παράγοντα και όχι αυτά των επιμέρους συνθηκών (μέρος, λιμάνι, πλοίο κλπ.). Επιστρέφοντας στα δεδομένα, σημειώνεται ότι τα χαρακτηριστικά χωρίζονται σχεδόν ίσα τόσο σε κατηγορικές όσο και σε συνεχείς τιμές. Ως επί το πλείστων όλες οι μεταβλητές είναι αρκετά ευκολονοητες με μια απλή ανάγνωση. Διευκρινίζεται ωστόσο η έβδομη μεταβλητή (sickness code) η οποία είναι η αναγγελία του τύπου ασθένειας ή ατυχήματος που στέλνεται από το πλοίο. Ο παρακάτω πίνακας περιγράφει τα δεδομένα εισόδου.

No	Μεταβλητές	Ερμηνεία	Τιμές
1	Class / Severity	χαρακτηρισμος ατυχηματος	1=Σοβαρο 2=Μη Σοβαρό
2	P&I Club	αναγγελία στα Club	1=Ναι 2=Οχι
3	Vessel Rank	μεγεθος πλοίων	συνεχείς
4	season	εποχη του έτους	διακριτές 1-4
5	rank description	αξιωμα πληρωματος	διακριτές 1-36
6	number of certificates	αριθμων διπλωματων/εκπαιδευσεων	συνεχείς
7	sickness code	κωδικοποιηση ατυχηματος	διακριτές 1-150
8	call port ranking	λιμανι ελλιμενισμου	διακριτές 1-106
9	age	ηλικία ναυτικού	συνεχείς
10	gender code	φύλλο ναυτικού	Ανδρας=1

			Γυναίκα=2
11	months of service	μήνες πλευσιμης υπηρεσίας	συνεχείς

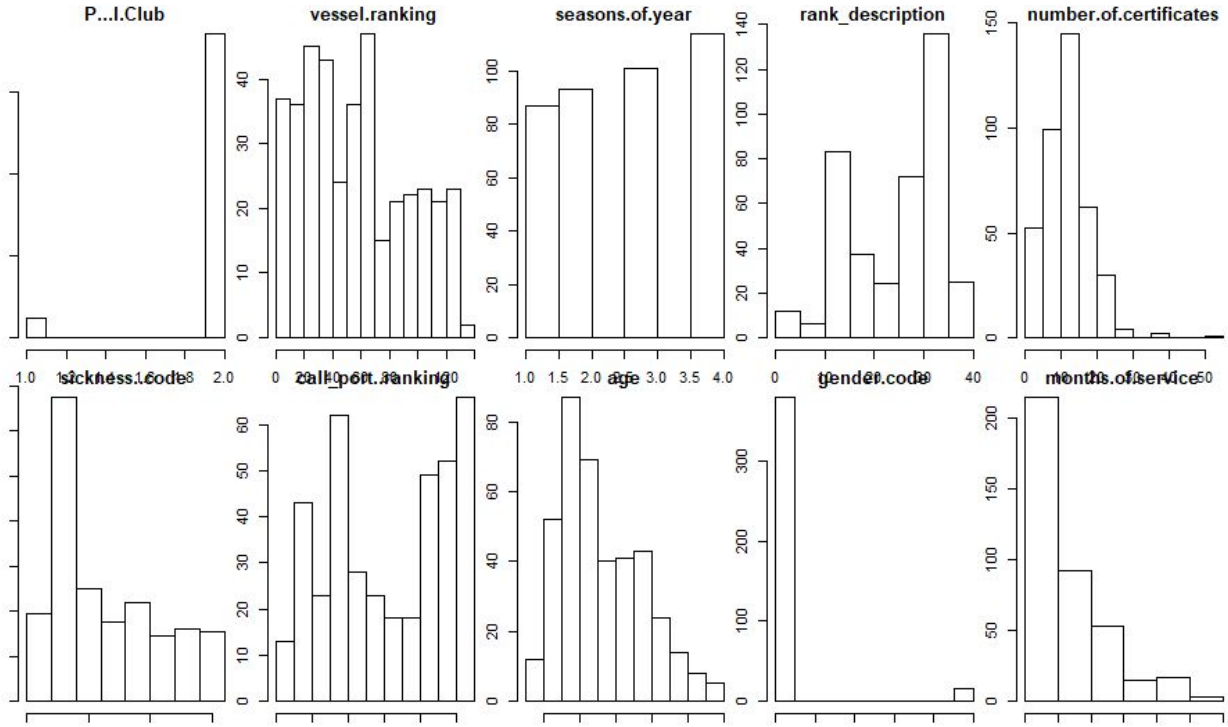
Πίνακας 1: Περιγραφή Δεδομένων

5.2 Περιγραφική Στατιστική

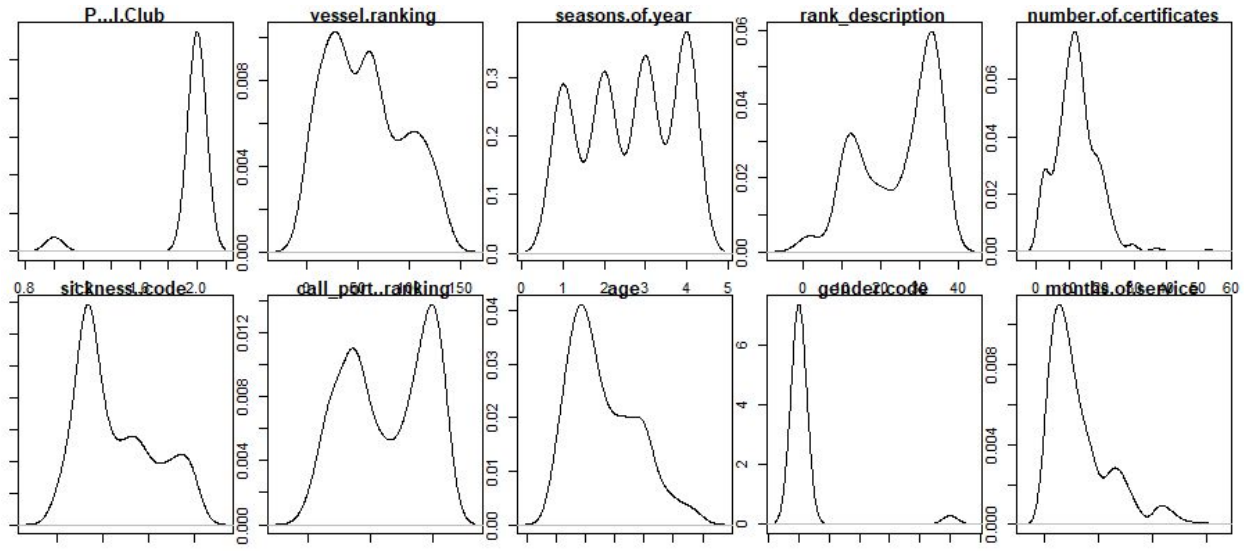
Για την κατανόηση και την περιγραφή των δεδομένων παρουσιάζεται παρακάτω η αναλυτική περιγραφή των δεδομένων μέσω descriptive statistics της R. Όπως σε κάθε πρόβλημα έτσι και στην μηχανική μάθηση είναι πολλές φορές αναγκαιο και ωφελιμο να προηγηθεί η επεξεργασία των δεδομένων ώστε να γίνουν αντιληπτα τα βασικά χαρακτηριστικά που τα διεπουν.

Ιστογραμμα και Διάγραμμα πυκνότητας

Αρχικά τα ιστογράμματα παρέχουν ένα διάγραμμα ράβδων με το διαχωρισμό των αριθμητικων χαρακτηριστικών τα οποία χωρίζονται σε κλάδους με το ύψος κάθε ράβδου να δείχνει τον αριθμό των περιπτώσεων που εμπίπτουν σε κάθε κλάδο. Είναι χρήσιμα καθώς δίνουν μια ένδειξη της κατανομής ενός χαρακτηριστικού. Στα συγκεκριμένα δεδομένα δεν φαίνεται ξεκάθαρα η κανονική κατανομή. Επίσης με το διάγραμμα πυκνότητας εξομαλύνουμε τα ιστογράμματα σε γραμμές χρησιμοποιώντας μια γραφική παράσταση πυκνότητας. Αυτό είναι χρήσιμο για μια πιο αφηρημένη απεικόνιση της κατανομής κάθε μεταβλητής .Η μεταβλητή call port ranking ακολουθεί double Gaussian κατανομή.



Διάγραμμα 1: Ιστόγραμμα

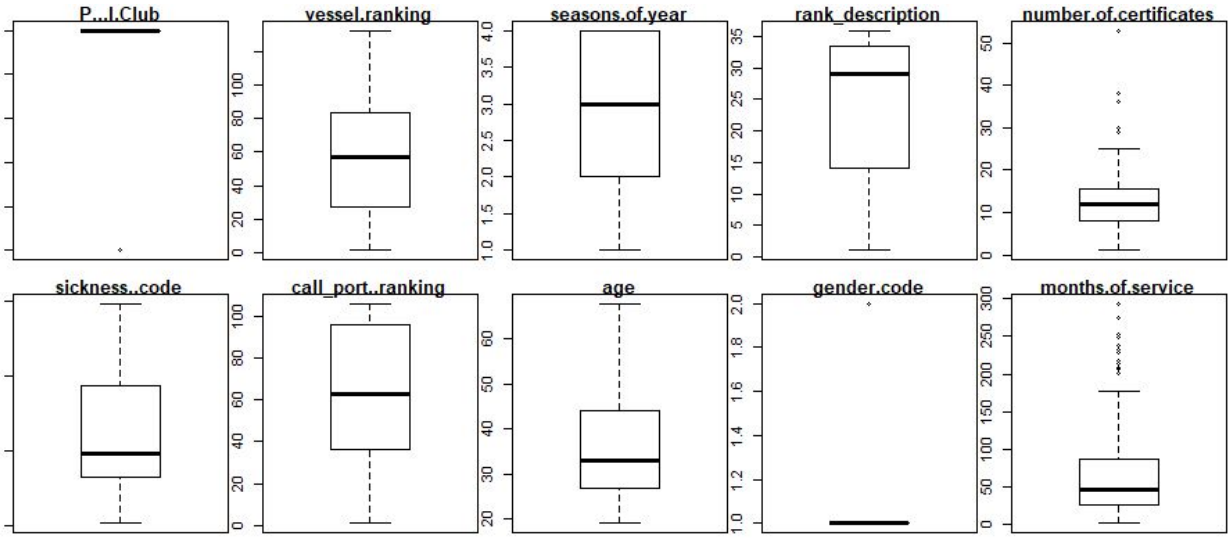


Διάγραμμα 2: Πυκνότητα Πιθανότητας

Box plot Whisker plot

Η κατανομή των δεδομένων δίνεται και με διαφορετικό τρόπο χρησιμοποιώντας τα box plot και whisker plot. Το κουτί καταγράφει το μέσον 50% των δεδομένων, η γραμμή δείχνει το διάμεσο και τοι γραμμές που εκτείνονται πάνω και κάτω από το κουτί δείχνουν την έκταση των δεδομένων. Οποιεσδήποτε κουκίδες έξω από τις γραμμές είναι πιθανόν ακραίες τιμές. Τα δεδομένα έχουν παρόμοιο εύρος καθώς επίσης το πλάτος του χαρακτηριστικου months of service & number of certificates φαίνεται να παρουσιάζει ακραίες τιμές. Ακόμα, διαπιστώνουμε ότι όσο περισσότερη εμπειρία και εκπαίδευση έχει ο ναυτικός τόσο λιγότερα τα ατυχήματα.

Όλα τα δεδομένα έχουν παρόμοιο εύρος (και τις ίδιες μονάδες των εκατοστών). Επίσης το πλάτος του χαρακτηριστικου months of service & number of certificates έχει μερικές ακραίες τιμές για αυτό το δείγμα δεδομένων. Ωστόσο διαπιστώνεται ότι όσο περισσότερη εμπειρία και εκπαίδευση έχει ο ναυτικός τόσο λιγότερα τα ατυχήματα.

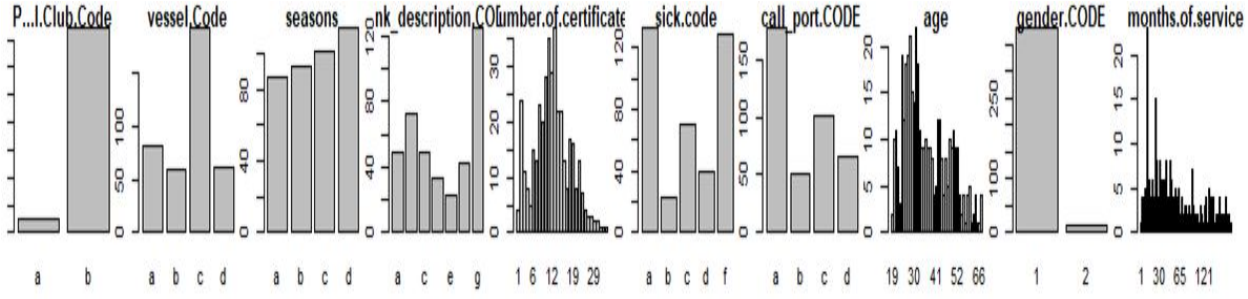


Διάγραμμα 3: Box Plot

Bar plots

Σε σύνολα δεδομένων που έχουν κατηγορηματικά και όχι αριθμητικά χαρακτηριστικά, μπορούν επίσης να δημιουργηθούν γραφήματα που δίνουν μια ιδέα για το ποσοστό των περιπτώσεων που ανήκουν σε κάθε κατηγορία και έτσι παρατηρούνται γραφήματα που έχουν μια καλή μικτή κατανομή και άλλα οπου η κατανομή ρέπει προς μια κλάση. Έτσι παρατηρούνται ορισμένα χαρακτηριστικά έχουν όμοια διαμοιρασμένη κατανομή, ενώ άλλα όχι. Εμφανές παράδειγμα στα

δεδομένα είναι η κλάση P&I Club. Με σκοπό την ανάλυση των δεδομένων υπό το πρίσμα των κατηγορικών δεδομένων τροποποιήθηκε το αρχικό σύνολο δεδομένων ώστε να μελετηθούν ευρύτερα πεδία συσχετισης. Συγκεκριμένα τα χαρακτηριστικά που περιλαμβανουν διακριτες τιμές τροποποιήθηκαν ώστε να προβάλλουν μεγαλύτερες ομάδες μελέτης στις οποίες εμπίπτουν πλέον περισσότερα περιστατικά. Για παράδειγμα το χαρακτηριστικο sickness code, στο αρχικο συνολο δεδομένων λαμβάνει 145 κλάσσεις διότι τόσες είναι οι κλάσεις τις οποίες έχει το σύνολο δεδομένων με σκοπό τη λεπτομερή καταγραφή των περιστατικών. Ωστόσο υπάρχει αξία στην μελέτη συλλογικών ομάδων ατυχημάτων. Για παράδειγμα όλες οι τιμές του αρχικού συνόλου δεδομένων που αναπαριστούν όλες τις αναγγελίες για τα ατυχήματα στην περιοχή του κεφαλιού, πλέον στο σύνολο δεδομένων accidents2 θα εμπιπτουν σε μια κλάση d. Ομοίως τροποποιούνται και τα υπόλοιπα χαρακτηριστικά. Ακόμη ένας λόγος για τη δημιουργία του επιπλέον συνόλου είναι η απουσία προηγούμενης μελετης για το συγκεκριμένο θέμα.



Διάγραμμα 4: Bar plots

Πίνακας συσχετίσεων

Παρακάτω υπολογίζεται η συσχέτιση μεταξύ κάθε ζευγαριού χαρακτηριστικών. Αυτά τα ζεύγη συσχετίσεων μπορούν να αναπαρασταθούν σε ένα πίνακα συσχετίσεων για να αναλυθεί οπτικά η σύγκριση των χαρακτηριστικών των οποίων οι τιμές αλλάζουν από κοινού. Χρησιμοποιήθηκε η τεκμηρίωση κατά την οποία το έντονο μπλε αντιπροσωπεύει θετική συσχέτιση και λευκό την αρνητική. Ο πίνακας είναι συμμετρικός και ότι οι διαγώνιες ιδιότητες συσχετίζονται απόλυτα θετικά (και αυτό το λόγο παραλείπονται οι διπλές τιμές). Ορισμένα από τα χαρακτηριστικά είναι αρκετά συσχετισμένα όπως είναι η ηλικία του ναυτικού σε σχέση με την εμπειρία, και σε μικρότερο βαθμό είναι ο βαθμός της εκπαίδευσης με την ηλικία και με την εμπειρία στη θάλασσα.

	P...I.Club	vessel.ranki	seasons.of.y	rank_descrip	number.of.c	sickness..co	call_port..ranking	age	gender.cod	months.of.service
P...I.Club	1	-0.0675887	0.1012554	0.0019405	-0.1192138	0.0171357	0.01843756	-0.002837703	0.0505327	0.04132228
vessel.ranking		1	0.0919263	0.0137117	-0.1522653	0.0676872	0.03870957	-0.043871213	0.037345	-0.00624776
seasons.of.year			1	-0.0378214	-0.1020365	-0.0129246	0.05502032	-0.048213433	0.104164	-0.06031203
rank_description				1	0.0823442	0.0220941	0.08989759	0.247265476	0.0473965	0.29270924
number.of.certificates					1	0.0188602	-0.0223292	0.212506032	-0.1537135	0.22432199
sickness..code						1	-0.04531692	0.112194565	0.0367768	0.01744594
call_port..ranking							1	0.029507656	-0.0639507	-0.04288116
age								1	-0.199438	0.68974955
gender.code									1	-0.15009749
months.of.service										1

Πίνακας 2: Συσχετίσεις Δεδομένων

Μέσος όρος και τυπική απόκλιση skewness

Παρατηρείται ότι δεν υπάρχουν ακραίες τιμές καθώς η τυπική απόκλιση δεν παρουσιάζει μεγάλη απόκλιση από το μέσο όρο του κάθε χαρακτηριστικού. Ώς εκ τουτου δε θεωρήθηκε απαραίτητο να γίνει κάποια σχετική τροποποίηση. Εάν μια κατανομή φαίνεται να ακολουθεί την Γκαουσιανή κατανομή αλλά ωθείται πολύ αριστερά ή δεξιά, είναι χρήσιμο να γνωρίζουμε την απόκλιση. Γενικά είναι πολύ πιο εύκολο ο χρηστης να καταλάβει απο τα διαγράμματα την αίσθησης της “παράκαμψης” -όπως ένα ιστόγραμμα ή γραφική παράσταση πυκνότητας- και είναι πιο δύσκολο να το καταλάβει από την εξέταση των μέσων, των τυπικών αποκλίσεων και των τεταρτημορίων. Παρόλα αυτά, ο υπολογισμός της κλίσης προς τα εμπρός δίνει μια αναφορά που μπορεί να χρησιμοποιηθεί αν αποφασίσει να διορθωθεί / τροποποιηθεί το σύνολο των τιμών για ένα ή παραπάνω χαρακτηριστικά. Τα δεδομένα της παρούσας εργασίας δεν ακολουθούν ακριβώς την κανονική κατανομή κάτι που είναι που είναι αντιληπτό τόσο απο τα διαγράμματα όσο και το μέτρο skewness. Συγκεκριμένα πολλές μεταβλητές έχουν δεξιά την ουρά της κατανομής και άλλες αριστερα.

	P&I Club	vessel	season	rank	certificates	sickness	port	age	gender	service
min	1	1	1	1	1	1	1	19	1	1
median	2	57	3	29	12	48	63	33	1	47
mean	1.9	57.21	24.66	2.61	12.26	64.68	62.81	36.33	1.03	65.14
max	2	133	4	36	53	149	106	68	2	294

SD	0.23	36.9	1.12	9.82	6.38	41.74	32.99	11.37	0.19	57.35
skewness	-3.66	0.35	-0.13	-0.57	1.05	0.6	-0.11	0.67	4.81	1.4

Πίνακας 3: Στατιστική Ανάλυση Δεδομένων

5.3 Rough Set Theory

Μέρος του συστήματος πρόβλεψης είναι η ανάδειξη των πιο αξιόλογων χαρακτηριστικών του συνόλου δεδομένων ώστε να γίνει βελτιστοποίηση της διαδικασίας. Παρόλο που αυτή η διαδικασία δεν αποσκοπεί αποκλειστικά στη βελτίωση του αλγορίθμου, είναι εν γένει βοηθητική ακόμη και για τον αλγόριθμο. Συγκεκριμένα κύριο μέλημα για τη σωστή λειτουργία του μοντέλου είναι η συλλογή όλων των απαραίτητων δικαιολογητικών που χρειάζεται το σύστημα ώστε να δοθεί και μια σωστή εκτίμηση. Όσο λιγότερα λοιπόν είναι τα απαραίτητα δικαιολογητικά τόσο πιο σύντομα θα τα συγκεντρώσει ο καπετάνιος ώστε να τα στείλει με το διαβιβαστικό μήνυμα στη στεριά. Με τον τρόπο αυτό φυσικά δε βελτιώνεται ο χρόνος απόδοσης των αλγορίθμων αλλα ο χρόνος που απαιτείται για την αποτελεσματική διεκπεραίωση του προβλήματος. Η μέθοδος Rough Set Theory είναι μια από τις ευρύτατα διαδεδομένες και αποδεκτές διαδικασίες για την επίλυση τέτοιων ζητημάτων.

Οι βασικοί της στόχοι της μεθόδου ‘’Rough Set Theory” είναι η επιλογή χαρακτηριστικών ουσιαστικής σημασίας, η μείωση των δεδομένων, η εξαγωγή μοτίβων (πρότυπων, κανόνων σύνδεσης). Το κύριο χαρακτηριστικό που κάνει τη αυτή τη μέθοδο να ξεχωρίζει είναι ότι επιτρέπει τη μείωση των αρχικών δεδομένων, δηλαδή την εύρεση ελάχιστων συνόλων δεδομένων με την ίδια γνώση όπως και στα αρχικά δεδομένα μέσω της αξιολόγησης της σημασίας των δεδομένων. Επίσης προσφέρει μαθηματικά εργαλεία για την εύρεση κρυμμένων μοτίβων στα δεδομένα προσδιορίζει τις επιμέρους ή συνολικές εξαρτήσεις των δεδομένα, εξαλείφει τα περιττά δεδομένα, δίνει προσέγγιση στις μηδενικές τιμές (αν υπάρχουν, καθώς και σε ελλείπουσες τιμές). Δεν χρειάζεται καμία ειδική επεξεργασία ή πρόσθετη πληροφορία σχετικά με τα δεδομένα. Γενικά είναι ευκολονόητη και προσφέρει απλή ερμηνεία των αποτελεσμάτων που έχουν αποκτηθεί. Είναι επίσης κατάλληλη για ταυτόχρονη (παράλληλη/κατανεμημένη) επεξεργασία.

Κάνοντας την ανάλυση μέσω από το πληροφοριακό περιβάλλον του weka το χαρακτηριστικό με τη μεγαλύτερη γνώση είναι το “P & I Club” και δεύτερο είναι το ‘’sickness code”. Και τα δυο χαρακτηριστικά έχουν άμεση σχέση με την μετέπειτα εξέλιξη ενός ατυχήματος και όχι με τις συνθήκες εκείνης της στιγμής. Δηλαδή φαίνεται ότι μεγαλύτερη γνώση έχει η τελική έκβαση και όχι τα αρχικά δεδομένα του συμβάντος. Αυτό είναι λογικό και ανταποκρίνεται στην ανάγκη ενός συστήματος που αντλεί αξία από το ιστορικό των περιστατικών και μπορεί να δώσει πρόβλεψη.

Rough Set Theory
1. P&I Club
2. Sickness code

Πίνακας 4: Αποτέλεσμα Rough Set Theory

5.4 Συγκρίσεις και αξιολόγηση Κατηγοριοποιητών

Στις περισσότερες περιπτώσεις η καλύτερη μετρική είναι η χρήση ενός συγκεκριμένου ορίου (threshold) όπως είναι Precision, Accuracy, Recall. Ο λόγος για τον οποίο είναι πιο χρήσιμες από τη καμπύλη ROC (η π.χ το μέτρο απόδοσης Gin) είναι το γεγονός ότι δίνουν ένα καθορισμένο όριο ενώ οι άλλες παρουσιάζουν ένα άθροισμα όλων των υποψήφιων ορίων που θα μπορούσε να χρησιμοποιήσει το μοντέλο. Αυτό σημαίνει ότι το όριο πρέπει να καθοριστεί από το χρήστη, κάτι που είναι απαραίτητο όταν η μελέτη έχει και εμπορικό σκοπό. Στην ουσία τη καμπύλη ROC (ή το μέτρο απόδοσης Gin)i βοηθούν στην επιλογή του ορίου και όχι στο να δηλώσουν με ακρίβεια αν ο αλγόριθμός είναι αποτελεσματικός ή όχι. Επίσης μπορούν να απαντήσουν στην ερώτηση αν είναι αρκετά καλός ο αλγόριθμός δεδομένου ότι ο χρήστης επιλέγει ένα όριο στην τύχη. Για να εντοπίσουμε ποιο μέτρο απόδοσης μας ενδιαφέρει θα πρέπει να μελετήσουμε και να ερμηνεύσουμε τι σημαίνει και τι ακριβώς αναλύει το κάθε μέτρο ακρίβειας. Για παράδειγμα το accuracy επιβραβεύει τους αλγορίθμους ανάλογα με το πόσο καλά προβλέπουν τις αρνητικές αλλα και τις θετικές τιμές. Αν θέλουμε να επιλέξουμε μια μετρική που θα δίνει βάση στους αλγορίθμους που θα προβλέπουν ορθά κυρίως τις θετικές τιμές τότε

θα χρησιμοποιηθεί το μέτρο ακρίβειας precision. Σε δεδομένα που παρουσιάζουν μη ισορροπημένες κατηγορίες, μπορεί να αναδείξει περισσότερη γνώση η μετρική Kappa, η οποία παρουσιάζει τις ίδιες πληροφορίες λαμβάνοντας υπόψη την ισορροπία των κλάσεων.

Accuracy - Kappa - LogLoss

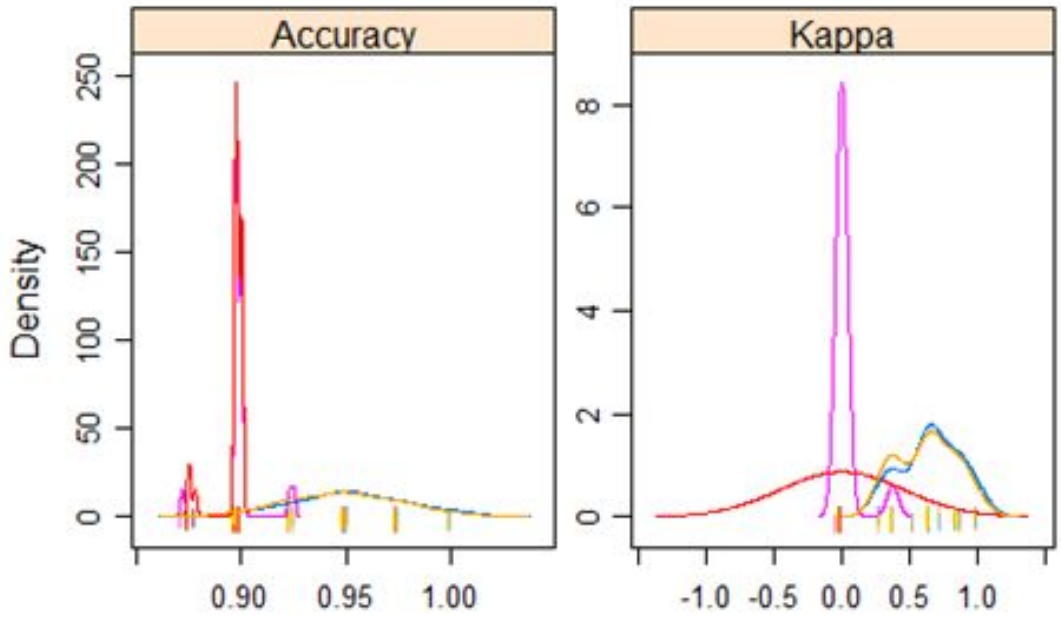
Η ακρίβεια και Kappa είναι οι μετρήσεις που χρησιμοποιούνται ως επι το πλείστον για την αξιολόγηση αλγορίθμων σε δυαδικά προβλήματα ταξινόμησης. Η ακρίβεια είναι το ποσοστό των σωστά ταξινομημένων στιγμιότυπων από όλες τις περιπτώσεις. Είναι πιο χρήσιμο σε μια δυαδική ταξινόμηση από ένα πολυκατηγοριακό πρόβλημα κατηγοριοποίησης επειδή μπορεί να είναι λιγότερο σαφές ακριβώς πώς η ακρίβεια διασπάται σε αυτές τις κατηγορίες. Η Kappa - Cohen's Kappa- είναι σαν την ακρίβεια , εκτός από το ότι είναι κανονικοποιημένη στη βάση της τυχαίας πιθανότητας για το σύνολο των δεδομένων. Είναι ένα πιο χρήσιμο μέτρο για χρήση σε προβλήματα που παρουσιάζουν ανισορροπία στις τάξεις (π.χ. διαίρεση 70% έως 30% για τις κλάσεις 1 και 2). Η λογαριθμική απώλεια (ή LogLoss) χρησιμοποιείται για την αξιολόγηση της δυαδικής ταξινόμησης, αλλά είναι πιο συνηθισμένη στους αλγορίθμους πολυ-ταξικής ταξινόμησης. Συγκεκριμένα, αξιολογεί τις πιθανότητες που εκτιμούν οι αλγόριθμοι. Σε πολλές περιπτώσεις ισχύει ότι όσο μικρότερη είναι τιμή του logloss ως προς το 0 τόσο καλύτερο για την αξιοπιστία του κατηγοριοποιητή. Ωστόσο αυτό δεν θα πρέπει να εκτιμηθεί ως ο απολυτος γενικός κανόνας για η ερμηνεία έχει να κάνει με την κατανομή των δεδομένα, με το αν είναι δυαδική κατηγοριοποίηση, με την ισορροπία των κλάσεων και άλλα.

Ο πίνακας 5 μας δείχνει τα αποτελέσματα των μεθόδων βάση του cross validation. Κατά την εφαρμογή του cross validation έγιναν 10 επαναλήψεις με αλλαγή του random seed σε κάθε επανάληψη (1 έως 10). Προτού σχολιάσουμε τα αποτελέσματα της έρευνας θα πρέπει να αναφέρουμε ότι ως κατώτατο όριο αποτέλεσμα ταξινομητή (baseline) παίρνουμε αυτό που προέρχεται από τον αλγόριθμο Zero R. Ο συγκεκριμένος αλγόριθμος έχει ως κριτήριο ταξινόμησης την πιο δημοφιλή κλάση που εμφανίζεται στο σύνολο δεδομένων εκπαίδευσης. Αυτό είναι αλήθεια καθώς το αποτέλεσμα που δίνει είναι 89,62% ακρίβεια. Με άλλα λόγια εάν έπρεπε να ‘μαντέψει’ ο χρήστης το ενδεχόμενο να μην είναι αρκετά σοβαρό το ατύχημα θα είχε πιθανότητα ίση με (354: οι περιπτώσεις όπου το ατύχημα δεν ήταν αρκετά σοβαρό) / 395: σύνολο δεδομένων) = 89,62 %.

	Zero - R	DT	NN	SVM	KNN	NB
Accuracy	**0.896**	**0.952**	**0.90**	**0.948**	**0.897**	**0.889**
Kappa		0.66	0	0.62	0.023	0
ROC	0.486	0.776	0.551	0.696	0.549	0.793
Log Loss		0.272	0.327	0.210	1.60	0.532
RMSE		0.21	0.22	0.23	0.29	0.25

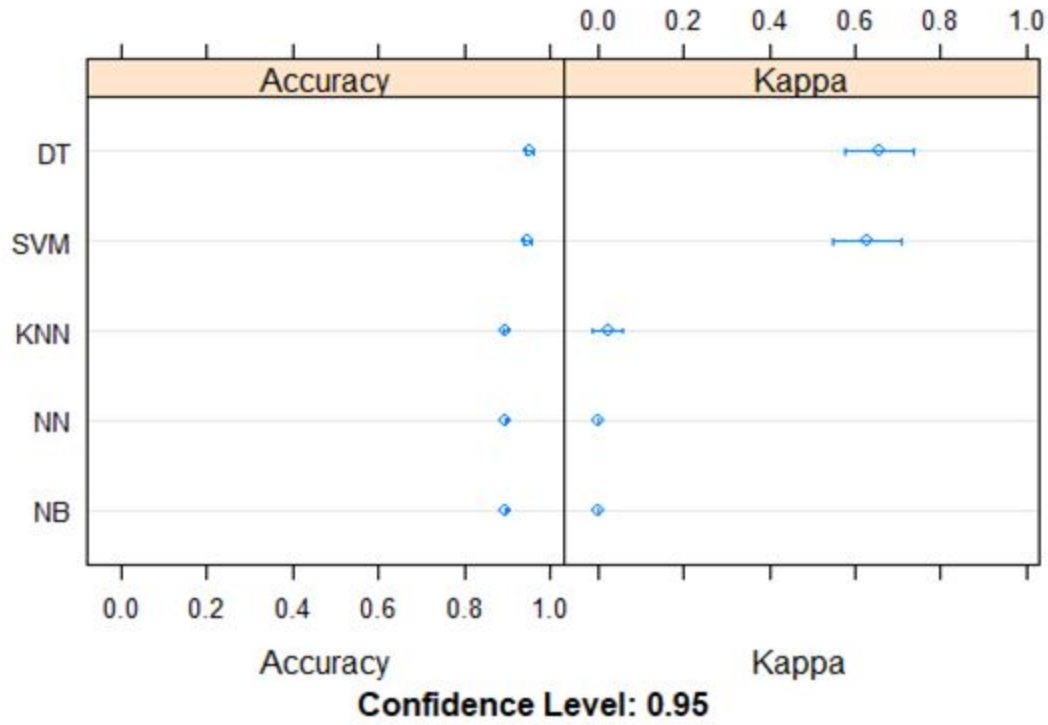
Πίνακας 5: Αποτελέσματα Ταξινομητών

Επίσης δίνεται η κατανομή της ακρίβειας του μοντέλου ως διάγραμμα πυκνότητας. Αυτός είναι ένας χρήσιμος τρόπος για να αξιολογηθεί το overlap της εκτιμώμενης συμπεριφοράς των αλγορίθμων.

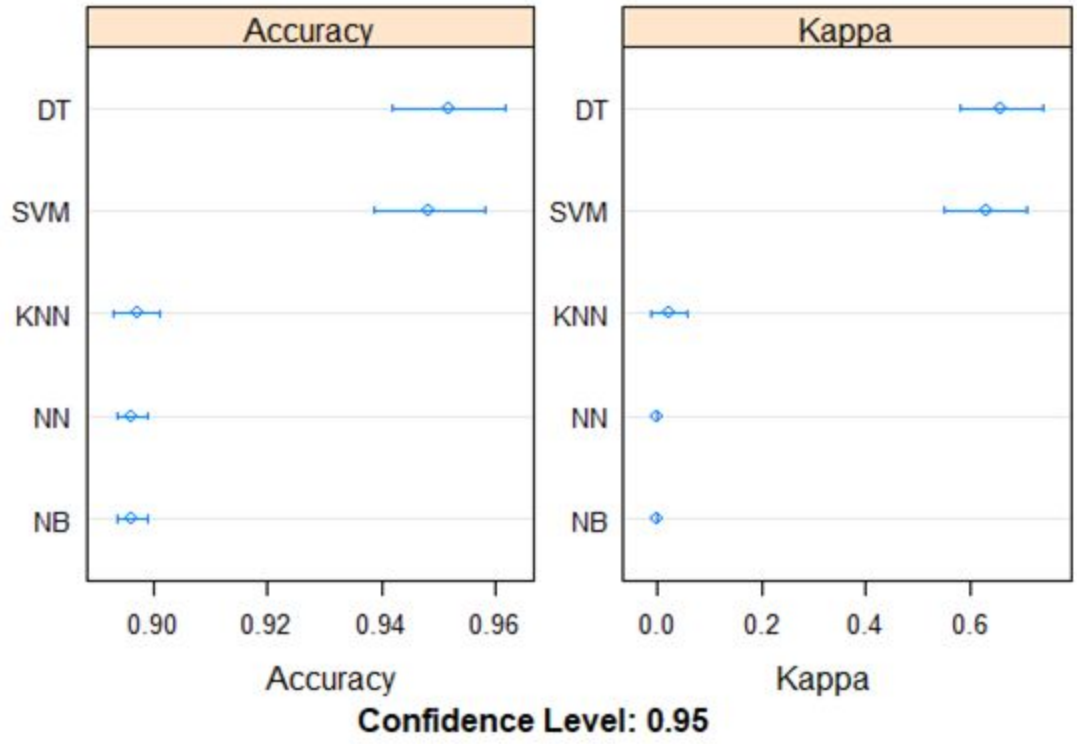


Διάγραμμα 5: Density plot by Class

Τα διαγράμματα 6 και 7 δίνουν την ακρίβεια των ταξινομητών.



Διάγραμμα 6: Σύγκριση ακρίβειας ταξινομητών



Διάγραμμα 7: Σύγκριση ακρίβειας ταξινομητών

Μια γενική παρατήρηση είναι ότι κανένας εκτιμητής δεν είναι χειρότερος σε αποτελεσματικότητα από τον Zero R που είναι το σημείο αναφοράς. Παρατηρούμε επίσης ότι σε όλους τους ταξινομητές η ακρίβεια είναι μεγαλύτερη από 90% εκτός από τον KNN & NB. Τη μεγαλύτερη τιμή έχουν οι αλγόριθμοι DT και SVM οι οποίοι είχαν και ελάχιστη διαφορά μεταξύ τους. Την επόμενη θέση παίρνουν τα νευρωνικά δίκτυα τα οποία έδωσαν ικανοποιητικά αποτελέσματα.

Η καμπύλη ROC ακολουθεί τα αποτελέσματα που έδωσε το μέτρο accuracy δίνοντας βαρύτητα στο αποτέλεσμα του βέλτιστου κατηγοριοποιητή. Ακολούθως και οι υπόλοιπες μετρικές Kappa, LogLoss και RMSE έδωσαν τιμές στους ταξινομητές ανάλογες με τις τιμές της ακρίβειας. Αυτό και πάλι συνηγορεί υπέρ των δέντρων απόφασης ώς τον καταλληλότερο αλγοριθμο για το πρόβλημα της εργασίας.

Τα δέντρα απόφασης γενικά λειτουργούν με κανόνες απόφασης και καταληκτικά με φύλλα και κόμβους. Η ρίζα του δέντρου είναι ένα από τα χαρακτηριστικά που ο αλγόριθμος κρίνει

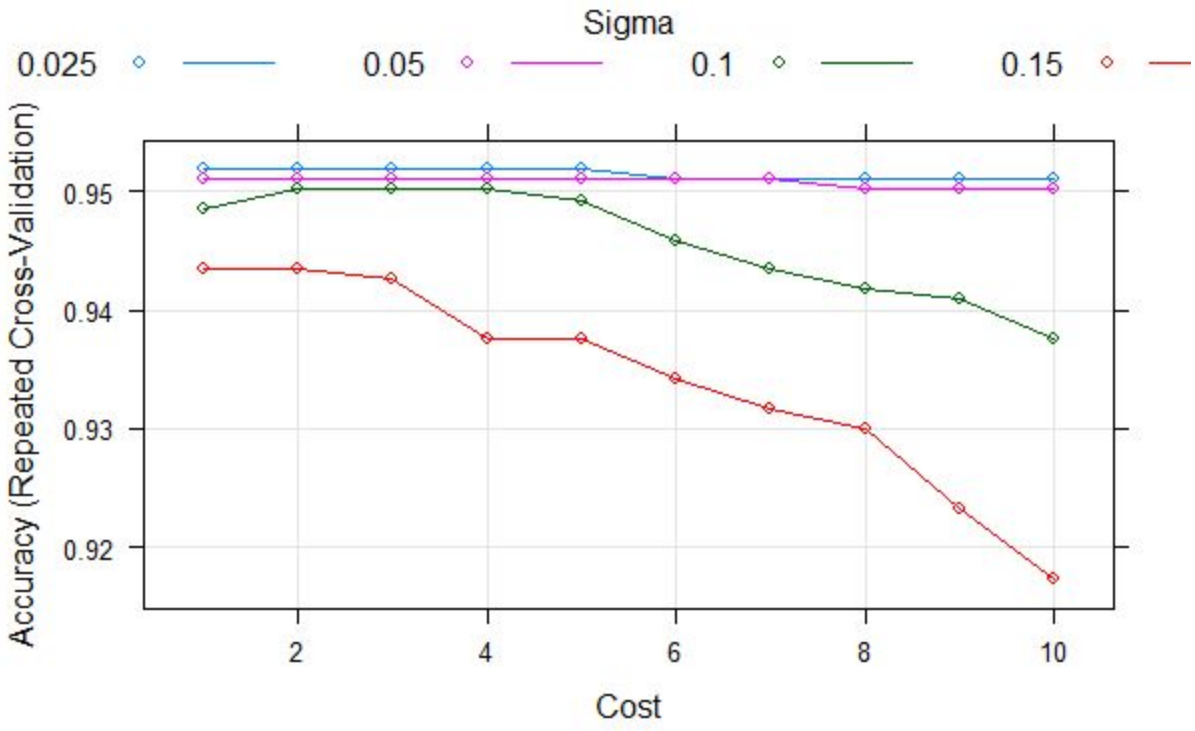
προσφορότερο να επιλεγεί πρώτο. Το χαρακτηριστικό της ρίζας είναι που παρουσιάζει και το μεγαλύτερο κέρδος πληροφορίας. Το γεγονός ότι τα δεδομένα προέρχονται από μια καλά οργανωμένη βάση δεδομένων σίγουρα ευνοεί τη δημιουργία ανάλογων κανόνων που δίνουν σαφή και ασφαλή συμπεράσματα. Τα δέντρα απόφασης έχουν επίσης καλή εφαρμογή (σε σχέση με άλλους ταξινομητές) σε προβλήματα που αφορούν έρευνες των καταναλωτών, διαφημιστικές καμπάνιες για την προώθηση προϊόντων κλπ. Οπου και εκεί υπάρχει συνήθως σωστή καταγραφή δεδομένων χωρίς ελλείπουσες τιμές, θόρυβο κλπ. Ακόμα τα δέντρα απόφασης έχουν πρακτική εφαρμογή και σε αρκετούς κλάδους της οικονομίας. Είναι αποδεδειγμένο ότι λειτουργούν στον τραπεζικό τομέα, στις online υπηρεσίες, (subscription), στον τομέα της ενέργειας και άλλες. Ένα κλασικό παράδειγμα είναι στον τραπεζικό τομέα και συγκεκριμένα στη διαδικασία δανειοδότησης ιδιωτών και επιχειρήσεων. Ένα μοντέλο δέντρου απόφασης μπορεί με ακρίβεια να δώσει πρόβλεψη για την εγκυρότητα και την καταλληλόλητα του υποψήφιου δανειολήπτη (αν θα μπορεί να εξυπηρετήσει τις μελλοντικές υποχρεώσεις). Εδώ και πάλι αυτό που κάνει σημαντική διαφορά είναι η πλήρης καταγραφή των δεδομένων.

Ένα βασικό χαρακτηριστικό που οδηγεί σε ακριβή αποτελέσματα για τα δέντρα απόφασης είναι οι δηλώσεις που έχουν αποτέλεσμα να διαχωρίζεται το σύνολο δεδομένων καλά. Με τον όρο καλά εννοούμε ότι όσο πιο ομογενοποιημένα είναι τα δεδομένα μετά από κάθε διαχωρισμό τόσο πιο επιτυχημένος θεωρείται ο διαχωρισμός. Το γεγονός ότι τα δεδομένα του συνόλου δεδομένων απαντούν εύκολα στις ερωτήσεις του αλγορίθμου που δημιουργούν τα φύλλα και τους κόμβους του δέντρου απόφασης είναι λογικό, καθώς για κάθε εγγραφή ναυτικού που παθαίνει ατύχημα υπάρχει πλήρες ιστορικό για τις παραμέτρους που υπήρχαν κατά το συμβάν. Τα δέντρα απόφασης πέραν της χρησιμότητας των προβλέψεων, βοηθούν στην αποτύπωση των χαρακτηριστικών που οδήγησαν στο τελικό αποτέλεσμα. Για παράδειγμα, στις ηλεκτρονικές εγγραφές (subscription) που αναφέρθηκαν παραπάνω το σημαντικότερο στοιχείο αναμφίβολα είναι η πρόβλεψη των ανθρώπων που τελικά θα γίνουν συνδρομητές. Ωστόσο, για το χρήστη είναι επίσης σημαντικό να γνωρίζει το κανάλι διαφήμισης που οδήγησε το πελάτη στην αγορά, αν δηλαδή ο πελάτης οδηγήθηκε μέσω της διαφήμισης των μέσων κοινωνικής δικτύωσης, ή μέσω από τις μηχανές αναζήτησης κλπ. Έτσι θα μπορεί να αποφασιστεί η πλέον αποδοτική στρατηγική διαφήμισης. Έτσι και στην περίπτωσή των ναυτικών ατυχημάτων μπορεί να φανεί χρήσιμο το κανάλι των χαρακτηριστικών που οδηγούν στο αποτέλεσμα ώστε να μπορεί να γίνει παραπάνω μελέτη και κινήσεις βελτίωσης.

Καταληκτικά, στην παρούσα εργασία, και με την προκειμένη επεξεργασία των δεδομένων την καλύτερη απόδοση έχουν ο αλγόριθμος των δέντρων αποφάσεων και ο SVM με μικρές ποιοτικές διαφορές μεταξύ τους. Κατάφεραν να εκπαιδεύσουν το training set σε πολύ ικανοποιητικό επίπεδο και χωρίς να κάνουν overfitting. Τη χειρότερη επίδοση παρουσίασε ο αλγόριθμος Naive Bayes τόσο σε επίπεδο ακρίβειας όσο και σε άλλα μέτρα απόδοσης. Η αυξημένη ακρίβεια των καλύτερων ταξινομητών έχει πρακτικο και εμπορικό ενδιαφέρον για την εταιρία καθώς κάθε συμβουλευτικό εργαλείο που βοηθά στις αποφάσεις είναι κρίσιμο να έχει υψηλή απόδοση ώστε να υπάρχει εμπιστοσύνη και βεβαιότητα κατά τη χρήση του.

5.5 Παραμετροποίηση αλγορίθμων - Tuning

Τελευταίο βήμα της ανάλυσης είναι η παραμετροποίηση των καλύτερων αλγορίθμων με σκοπό τη βελτίωση της αποτελεσματικότητάς τους. Καθώς αυτό γίνεται στα πλαίσια της αναζήτησης μεγαλύτερης ακρίβειας, ωστόσο είναι χρήσιμη μέθοδος για να εξετάσει ο χρήστης τα ήδη υπάρχοντα αποτελέσματα των αλγορίθμων. Στη συγκεκριμένη μελέτη θα αναλυθεί εκτενέστερα ο SVM ταξινομητής που είναι ο ένας από τους δύο καλύτερους. Πιο αναλυτικά μεταβλήθηκαν οι τιμές για τις παραμέτρους sigma (η οποία είναι smoothing parameter) και C (η οποία cost constraint). Η sigma πήρε τιμές 0.025, 0.05, 0.1, 0.15 που είναι κοντά στην default τιμή 0,1 και η C έλαβε τιμες απο 1 εώς 10. Παρακάτω, δίνονται τα αποτελέσματα του τροποποιημένου αλγορίθμου καθώς και το διάγραμμα 8 στο οποίο περιγράφονται τα αποτελέσματα που δινουν οι παραμετροποιήσεις που έγιναν. Το συμπέρασμα είναι ότι ο SVM αλγόριθμος αποδίδει καλυτερα για sigma = 0.025 και C = 1 (όπως φαίνεται και από το διάγραμμα σύγκρισης 8). Αρα γίνεται αντιληπτό ότι η αλλαγη του sigma απο 0,15 σε 0,025 έχει νόημα καθώς βελτίωσε την ακρίβεια.



Διάγραμμα 8: Σύγκριση παραμετροποίησης

Support Vector Machines with Radial Basis Function Kernel

395 samples

10 predictor

2 classes: 'serious', 'various'

Pre-processing: Box-Cox transformation (8)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 356, 356, 355, 355, 355, 355, ...

Resampling results across tuning parameters:

sigma	C	Accuracy	Kappa
0.025	1	0.9518569	0.6577513
0.025	2	0.9518569	0.6577513
0.025	3	0.9518569	0.6577513
0.025	4	0.9518569	0.6577513
0.025	5	0.9518569	0.6577513
0.025	6	0.9510022	0.6510456
0.025	7	0.9510022	0.6510456
0.025	8	0.9510022	0.6510456
0.025	9	0.9510022	0.6510456
0.025	10	0.9510022	0.6510456
0.050	1	0.9510022	0.6510456
0.050	2	0.9510022	0.6510456
0.050	3	0.9510022	0.6510456
0.050	4	0.9510022	0.6510456
0.050	5	0.9510022	0.6510456
0.050	6	0.9510022	0.6510456
0.050	7	0.9510022	0.6510456
0.050	8	0.9501475	0.6421166
0.050	9	0.9501475	0.6421166
0.050	10	0.9501475	0.6421166
0.100	1	0.9484595	0.6279672
0.100	2	0.9501475	0.6421166

0.100 3 0.9501475 0.6421166

0.100 4 0.9501475 0.6421166

0.100 5 0.9492928 0.6331876

0.100 6 0.9459167 0.6026808

0.100 7 0.9434370 0.5750687

0.100 8 0.9417490 0.5641963

0.100 9 0.9408943 0.5517186

0.100 10 0.9375182 0.5243158

0.150 1 0.9434370 0.5774683

0.150 2 0.9434370 0.5774683

0.150 3 0.9425823 0.5707626

0.150 4 0.9375599 0.5163763

0.150 5 0.9375386 0.5295185

0.150 6 0.9341625 0.5100750

0.150 7 0.9315984 0.4856706

0.150 8 0.9299104 0.4885826

0.150 9 0.9231369 0.4424865

0.150 10 0.9172191 0.4311095

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were sigma = 0.025 and C = 1.

5.6 Συμπεράσματα και Μελλοντική μελέτη

Απο τα αποτελέσματα γίνεται κατανοητό ότι υπάρχει όφελος στην ακριβή καταγραφή των δεδομένων που αποτελούν μέρος της επιχειρησιακής λειτουργίας. Συγκεκριμένα παρατηρήθηκε

ωφέλιμη γνώση στα χαρακτηριστικά P&I Club και sickness code τα οποία είχαν και την ανάλογη συμμετοχή στην εκπαίδευση των αλγορίθμων. Για το λόγο αυτό μια πρόταση είναι να αναλυθεί περισσότερο η φύση των ατυχημάτων και ο τρόπος με τον οποίο προκαλούνται. Κάτι τέτοιο είναι δυνατόν να πραγματοποιηθεί στα πλαίσια των ναυτιλιακών εταιρειών με τη συμβολή των επαγγελματιών υγείας της εταιρίας (όπου συνήθως είναι ο medical fleet advisor), του καπετάνιου και του ημερολογίου της γεφυρας στο οποία καταγράφονται όλα τα γεγονότα που συμβαίνουν. Ακόμα θα είναι ωφέλιμο να μελετηθούν σε μεγαλύτερη έκταση όλα τα ατυχήματα που οδηγούν σε αναγγελία του P&I Club με απώτερο σκοπό την αποφυγή τους. Με άλλα λόγια μπορεί να μελετηθεί το πως θα γίνει πρόληψη των προβλημάτων που προκαλούνται από αυτά τα ατυχήματα. Οι τρόποι ποικίλλουν και πιθανόν να έχουν να κάνουν με την κατάλληλη εκπαίδευση του προσωπικού, των ναυτικών, καθώς και με τα μέρη των πλοίων που ελλοχεύουν κινδύνους. Ακόμη, μέσα από τη συνεργασία των κατάλληλων τμημάτων θα μπορούσαν να αλλάξουν διαδικασίες στην καθημερινότητα του πλοίου οι οποίες φαίνεται να αυξάνουν τον κίνδυνο εμφάνισης σχετικών ατυχημάτων.

Επίσης στα πλαίσια μελλοντικής μελέτης αξίζει να διερευνηθούν και να αναλυθούν ακόμη περισσότερα χαρακτηριστικά που να βελτιώνουν το μοντέλο πρόβλεψης και να ορίζουν με ακόμη μεγαλύτερη ακρίβεια τις δαπάνες για ένα ατύχημα. Καθώς τα χαρακτηριστικά και οι συνθήκες που συνδέονται με ένα ατύχημα είναι πολλές, υπάρχει ευρύ πεδίο αναζήτησης χαρακτηριστικών. Επιπλέον, σε μια μελλοντική μελέτη θα μπορούσε να διερευνηθεί ακόμη περισσότερο το επίπεδο στο οποίο επηρεάζουν τα πιο σημαντικά χαρακτηριστικά του συνόλου των δεδομένων, στο στάδιο της εκπαίδευσης. Με διάφορες διαθέσιμες τεχνικές (rough set theory) βρέθηκαν εκείνα τα χαρακτηριστικά που απλοποιούν το πακέτο δεδομένων. Με αυτή την έννοιά θα μπορούσε να προχωρήσει περισσότερο η προτεινόμενη μελέτη.

6. Βιβλιογραφία

1. Ομάδες ταξινομητών για την αύξηση της ακρίβειας των μεθόδων μηχανικής μάθησης και εξόρυξης γνώσης . Σωτήρης Β. Κωτσιαντής - Διδακτορική Διατριβή
2. Analysing port and shipping operations using big dat - Christopher Bonham, Alex Noyvirt, Ioannis Tsalamanis and Sonia Williams
3. Neural Networks with R Authors Giuseppe Ciaburro Balaji Venkateswaran
4. McGrawHill - Machine Learning -Tom Mitchell
5. ML Machine Learning-A Probabilistic Perspective Author: Kevin P Murphy
6. Κεφαλαιο 4 Μηχανική Μάθηση - Γεωργούλη Κατερίνα ΤΕΙ Αθηνας
7. Decision Trees MIT 15.097 Course Notes Cynthia Rudin
8. Analysis of Classification Algorithms Applied to Hepatitis Patients - T.Karthikeyan, PhD. / P.Thangaraju
9. Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach Fadl Mutaher Ba-Alwi, Houzifa M. Hintaya
10. AN ASSESSMENT OF VESSEL-SOURCE OIL POLLUTION INCIDENTS IN THE MEDITERRANEAN SEA USING INDUCTIVE MACHINE LEARNING METHODOLOGIES A. B. Alexopoulos G. Dounias
11. Using Machine Learning Methods for Evaluating the Quality of Technical Documents - Michael LUCKERT Moritz SCHAEFER-KEHNER
12. Predicting Shipping Freight Rate Movements Using Recurrent Neural Networks and AIS Data Stian Røyset Salen Gisle Hoel Århus
13. Μηχανική Μάθηση και εφαρμογή σε Οικονομικά δεδομένα - Φαζάκης Νικόλαος
14. A Machine-Learning Approach to Predict Main Energy Consumption under Realistic Operational Conditions
15. Joan P. Petersen, Ole Winther & Daniel J. Jacobse
16. Predicting arrival times for tankers ships using recurrent neural networks - Raymond Hardij
17. Fuzzy-roughnearestneighbourclassificationandprediction RichardJensena,∗,ChrisCorneli

18. Neural Network Model with Monte Carlo Algorithm for Electricity Demand Forecasting in Queensland - Binbin Yong
19. Business Administration Strategy for a small/medium sized Nordic shipping line - Ágúst Þór Ragnarsson June 2013
20. An Examination into the Structure of Freight Rates in the Shipping Freight Markets - Stefan van Dellen
21. Ανάπτυξη Προσθετικών Μηχανών Διανυσμάτων Υποστήριξης: Μεθοδολογία και εφαρμογή στη πρόβλεψη ενδονοσοκομειακού θανάτου - Γκολφινοπούλου Βασιλική