ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες"

Ειδίκευση : "Μεγάλα Δεδομένα και Αναλυτική"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναλυτική Αθλητικών Δεδομένων : Αξιολόγηση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη νικητήριας ομάδας για το Αγγλικό πρωτάθλημα ποδοσφαίρου (EPL)

Σεβαστή Γιάχου

Α.Μ.: ΜΕ2007

Επιβλέπων Καθηγητής:

Ηλίας Μαγκλογιάννης, Καθηγητής

ΠΕΙΡΑΙΑΣ

ΦΕΒΡΟΥΑΡΙΟΣ 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναλυτική Αθλητικών Δεδομένων : Αξιολόγηση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη νικητήριας ομάδας για το Αγγλικό πρωτάθλημα ποδοσφαίρου (EPL)

Σεβαστή Γιάχου

Α.Μ.: ΜΕ2007

ΠΕΡΙΛΗΨΗ

Οι τεχνικές εξόρυξης δεδομένων έχουν εφαρμοστεί με επιτυχία σε πολλούς επιστημονικούς, βιομηχανικούς και επιχειρηματικούς τομείς. Στον τομέα του επαγγελματικού αθλητισμού είναι γνωστό πως συλλέγονται τεράστιες ποσότητες δεδομένων για κάθε παίκτη, προπόνηση, ομάδα, παιχνίδι και σεζόν, ωστόσο η αποτελεσματική χρήση αυτών των δεδομένων εξακολουθεί να είναι περιορισμένη. Πολλοί αθλητικοί οργανισμοί έχουν αρχίσει να συνειδητοποιούν ότι υπάρχει πληθώρα αναξιοποίητων γνώσεων που περιέχονται στα δεδομένα τους και υπάρχει αυξανόμενο ενδιαφέρον για τεχνικές για τη χρήση αυτών. Ο στόχος αυτής της μελέτης είναι η ανάπτυξη ισχυρών μοντέλων με σκοπό την πρόβλεψη της νικητήριας ομάδας του αγγλικού πρωταθλήματος με την υψηλότερη δυνατή ακρίβεια, χρησιμοποιώντας και αξιολογώντας την απόδοση των σχετικών με το πρόβλημα, αλγορίθμων επιβλεπόμενης μηχανικής μάθησης. Χρησιμοποιήθηκαν στατιστικά στοιχεία για είκοσι δύο σεζόν του Αγγλικού Πρωταθλήματος τα οποία αποκτήθηκαν με την μέθοδο web-scrapping από την ιστοσελίδα transfermarkt. Τόσο η απόκτηση των δεδομένων, όσο και η υλοποίηση έγινε εξολοκλήρου σε γλώσσα προγραμματισμού Python και το επίπεδο ακρίβειας που επιτεύχθηκε κατά την υλοποίηση του προβλεπτικού μοντέλου, ανέρχεται σε 90 %.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ:

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ανάλυση αθλητικών δεδομένων, τεχνητά νευρωνικά δίκτυα, μηχανική μάθηση, feed forward MLP

ABSTRACT

Data mining techniques have been successfully applied in many scientific, industrial and business fields. In the field of professional sports it is known that huge amounts of data are collected for every player, practice, team, game and season, but the effective use of this data is still limited. Many sports organizations are beginning to realize that there is a wealth of untapped knowledge contained in their data and there is a growing interest in techniques to use it. The objective of this study is the development of robust models to predict the winning team of the English league with the highest possible precision using and evaluating the performance of problem-specific supervised machine learning algorithms. Statistics for the twenty two seasons of the English Championship were used which were obtained by using web-scraping method from the transfermarkt website. Both the acquisition of the data and the implementation was done entirely in Python programming language and the level of accuracy achieved during the implementation of the predictive model is 90 %.

SUBJECT AREA:

KEYWORDS: Sports data analysis, artificial neural networks, machine learning, feed forward MLP

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΟΡΙΣΜΟΙ

Training: Η εκπαίδευση ενός νευρωνικού δικτύου είναι η διαδικασία εύρεσης τιμών για τα βάρη και τα bias, έτσι ώστε για ένα δεδομένο σύνολο τιμών εισόδου, οι υπολογισμένες τιμές εξόδου να ταιριάζουν πολύ με τις γνωστές, σωστές τιμές-στόχους.

Validation: Το σύνολο δεδομένων επικύρωσης παρέχει μια αμερόληπτη αξιολόγηση της προσαρμογής ενός μοντέλου στο σύνολο δεδομένων εκπαίδευσης, ενώ συντονίζει τις υπερπαραμέτρους του μοντέλου (π.χ. τον αριθμό των κρυφών στρωμάτων) σε ένα νευρωνικό δίκτυο.

Testing: Το πρώτο βήμα είναι να αλλάξει το νευρωνικό δίκτυο από μια λειτουργία εκμάθησης σε μια λειτουργία «τρεξίματος». Στη συνέχεια, εκτελεί τα ίδια δεδομένα εκπαίδευσης που μόλις χρησιμοποιήθηκαν μέσω του συστήματος για να παρατηρηθεί το ποσοστό σφάλματος που προκύπτει από τη σύγκριση της εξόδου του νευρωνικού δικτύου με το αναμενόμενο αποτέλεσμα από τα δεδομένα.

Learning Rate: Είναι μια ρυθμιζόμενη υπερπαράμετρος που χρησιμοποιείται στην εκπαίδευση νευρωνικών δικτύων, η οποία έχει μικρή θετική τιμή, συχνά στο εύρος μεταξύ 0 και 1. Το ποσοστό εκμάθησης ελέγχει πόσο γρήγορα το μοντέλο προσαρμόζεται στο πρόβλημα. Πιθανώς αποτελεί την πιο σημαντική υπερπαράμετρο για το μοντέλο.

Epoch: Όσον αφορά τα τεχνητά νευρωνικά δίκτυα, μια επανάληψη αναφέρεται σε έναν κύκλο του πλήρους συνόλου δεδομένων εκπαίδευσης. Συνήθως, η εκπαίδευση ενός νευρωνικού δικτύου διαρκεί περισσότερες από μερικές επαναλήψεις.

Momentum: Το momentum του νευρωνικού δικτύου είναι μια απλή τεχνική που βελτιώνει συχνά τόσο την ταχύτητα όσο και την ακρίβεια της εκπαίδευσης του.

Cross Validation: Το σύνολο εκπαίδευσης χωρίζεται σε k υποσύνολα (k fold cross validation). Κάθε φορά αφαιρείται από το σύνολο εκπαίδευσης ένα υποσύνολο που θα χρησιμοποιηθεί για την επαλήθευση και η εκπαίδευση πραγματοποιείται με τα εναπομείναντα (k-1) υποσύνολα. Η διαδικασία επαναλαμβάνεται k φορές και τα k εκπαιδευμένα δίκτυα χρησιμοποιούνται για την τελική πρόβλεψη (π.χ. πλειοψηφία ή μέσος όρος).

1 Εισαγωγή

Το πεδίο της Αναλυτικής Αθλητικών Δεδομένων (Sports analytics) είναι ένα πολλά υποσχόμενο ερευνητικό πεδίο που περιλαμβάνει την απόκτηση πολύτιμων πληροφοριών σχετικά με τον αγώνα, βάσει παλαιότερων παιχνιδιών, ή ακόμη και παιχνιδιών που βρίσκονται σε εξέλιξη. Η πρόβλεψη του τελικού αποτελέσματος του αγώνα αποδεικνύεται πολύ ωφέλιμη για τα μέλη της ομάδας, τους προπονητές και για τους στοιχηματίες. Για παράδειγμα, οι προπονητές μπορούν να αναπτύξουν την τακτική των επικείμενων παιχνιδιών βασιζόμενοι στο αποτέλεσμα προηγούμενων αγώνων ή στατιστικών που σχετίζονται με συγκεκριμένους παίκτες [1].

Το ποδόσφαιρο είναι παγκοσμίως ένα πολύ δημοφιλές άθλημα και κατέχει το μεγαλύτερο μερίδιο στην βιομηχανία των αθλητικών στοιχημάτων, η οποία αναπτύσσεται με ταχύ ρυθμό.

Η εφαρμογή προγνωστικών αναλυτικών τεχνικών έχει χρησιμοποιηθεί με επιτυχία σε πολλά διαφορετικά αθλήματα όπως ποδόσφαιρο, μπάσκετ, κρίκετ, ράγκμπι και χόκεϊ. Η πρόβλεψη αποτελέσματος ενός παιχνιδιού είναι φυσικά ένας από τους πιο προφανείς στόχους στο πεδίο του Sports Analytics.

Ακόμα κι αν η πρόβλεψη της έκβασης του αγώνα στο ποδόσφαιρο, στην βασική της μορφή έχει τρία αποτελέσματα (win-loss-draw), υπάρχουν και άλλα δεδομένα που η πρόβλεψη τους παρουσιάζει ενδιαφέρον. Τα πιο δημοφιλή στοιχήματα ποδοσφαίρου αφορούν όχι μόνο το αποτέλεσμα, αλλά και την πρόβλεψη του αριθμού του σκορ, αριθμός κόρνερ, τα ελεύθερα σουτ ή ακόμα και τις κάρτες.

Αρκετές μελέτες στη βιβλιογραφία της στατιστικής και της έρευνας έχουν προηγουμένως εξετάσει τα αποτελέσματα πρόβλεψης στο πεδίο των Sports Analytics με κλασσικές τεχνικές μηχανικής μάθησης για ταξινόμηση , αλλά η χρήση νευρωνικών δικτύων για το σκοπό αυτό είναι ο πιο πρόσφατος τομέας μελέτης. Η ισχυρή τεχνική των νευρωνικών δικτύων έχει αποδειχθεί να είναι αποτελεσματική στην παραγωγή μοντέλων ταξινόμησης μεγάλης ακρίβειας και σε άλλους τομείς. Η προσέγγιση που περιγράφεται στην παρούσα εργασία,

αξιοποιεί το γεγονός ότι τα νευρωνικά δίκτυα είναι αποδοτικά στην αναγνώριση μοτίβων και την χαρτογράφηση αυτών σε εξόδους.

Για την εργασία αυτή, κατασκευάστηκαν τα μοντέλα επιβλεπόμενης μηχανικής μάθησης Naïve Bayes (MultinomialNB), Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest όπως και ένα μοντέλο τεχνητού νευρωνικού δικτύου (ANN) για την πρόβλεψη της νικητήριας ομάδας στην Αγγλική Premiere League. Η ανάπτυξη έγινε εξολοκλήρου σε γλώσσα προγραμματισμού Python.

2 Σχετικές έρευνες και επιστημονικό κενό

Πληθώρα ακαδημαϊκών μελετών έχουν πραγματοποιηθεί πρόσφατα που αφορούν στην πρόβλεψη έκβασης ποδοσφαιρικών αγώνων. Οι μεθοδολογίες μπορούν να κατηγοριοποιηθούν με βάση το είδος δεδομένων που χρησιμοποιήθηκαν, το στάδιο πρόβλεψης (δηλαδή κατά τη διάρκεια ή πριν από το παιχνίδι, ή ακόμα και τη σεζόν), τον τύπο του αποτελέσματος που πρέπει να προβλεφθεί και τις τεχνικές πρόβλεψης που χρησιμοποιούνται.

Ο C. Reep θεωρείται πως είναι ο πρώτος αναλυτής δεδομένων. Το 1968 μαζί με τον B. Benjamin, δημοσίευσε μια στατιστική ανάλυση των προτύπων παιχνιδιού στο ποδόσφαιρο, χρησιμοποιώντας ως σύνολο δεδομένων 578 αγώνες στο χρονικό διάστημα μεταξύ 1953 και 1967. Τα τελευταία 20 χρόνια, εξελιγμένες τεχνικές, αλγόριθμοι και εργαλεία αναπτύχθηκαν για την ανάλυση αθλητικών δεδομένων, ενώ άρθρα και εργασίες που σχετίζονται με αθλητικές αναλύσεις δημοσιεύονται συνεχώς [2].

Στις μελέτες που αφορούν στην χρήση τεχνητών νευρωνικών δικτύων για την πρόβλεψη αποτελέσματος αγώνων ποδοσφαιρικών ομάδων, δύο προσεγγίσεις έχουν χρησιμοποιηθεί για τη μοντελοποίηση των αποτελεσμάτων. Η πρώτη μοντελοποιώντας τα γκολ που σημειώθηκαν και παραχωρήθηκαν από κάθε ομάδα [3] και η δεύτερη, μοντελοποιώντας άμεσα το αποτέλεσμα win-draw-lost είτε ως ποσοστό όπως στην περίπτωση των [4].

Η μελέτη των Arabzad et al. , ενσωματώνει ως χαρακτηριστικά την ομάδα, τη μορφή της ομάδας σε ολόκληρο το πρωτάθλημα μέχρι την ημερομηνία διεξαγωγής του αγώνα, τη μορφή της ομάδας στους 4 τελευταίους αγώνες, την ποιότητα των τελευταίων αντιπάλων και την εβδομάδα του αγώνα. Η προβλεπόμενη απόδοση είναι ο αριθμός των γκολ για την γηπεδούχο ομάδα και την φιλοξενούμενη ομάδα και λόγω τουλάχιστον δύο περιορισμών που μπορεί να αποδίδουν σε ένα πιο ακριβές μοντέλο, όπως η επένδυση των συλλόγων και ο καιρός [3].

Στην δεύτερη προσέγγιση ανήκει, η μελέτη των McCabe και Trevathan , η οποία προσπαθεί να προβλέψει το αποτέλεσμα των αγώνων ποδοσφαίρου ως νίκη

εντός έδρας, εκτός έδρας ή ισοπαλία. Η ετικέτα προβλέπεται βάσει ενός πλούσιου συνόλου χαρακτηριστικών και ενσωματώνει τις βολές στον στόχο , τα γκολ εναντίον, τρέχον ρεκόρ νίκης-απώλειας, τη γενική επίδοση εντός και εκτός έδρας, την απόδοση της ομάδας στους προηγούμενους 4 αγώνες, τη τρέχουσα θέση στην κατάταξη, τοποθεσία και διαθεσιμότητα παίκτη. Με αυτό το σύνολο χαρακτηριστικών οδηγήθηκαν σε ακρίβεια 54% και υποδεικνύουν ότι ένα πλουσιότερο διάνυσμα χαρακτηριστικών μπορεί να ενισχύσει την ακρίβεια του μοντέλου [5]. Αντίστοιχη ακρίβεια επιτυγχάνουν και οι Aslan και Inceoglu με δύο πιο απλά μοντέλα τεσσάρων και δύο χαρακτηριστικών αντιστοίχως, βαθμολογώντας διαφορετικά τα παιχνίδια εντός και εκτός έδρας για την κάθε ομάδα ανάλογα με το αν είναι η γηπεδούχος ή όχι [6].

Η μελέτη των Huang και Chang , οποίοι δημιούργησαν ένα MLP για να προβλέψουν το αποτέλεσμα των ποδοσφαιρικών αγώνων των σταδίων μετά το στάδιο της δημιουργίας ομίλων του Παγκόσμιου Κυπέλλου του 2006, έχοντας ως είσοδο τους αγώνες των ομίλων. Στη περίπτωση αυτή δεν υπάρχει δυνατότητα ισοπαλίας, καθώς μία από τις δύο ομάδες πρέπει να προχωρήσει στον επόμενο γύρο. Ενσωμάτωσαν ένα σύνολο χαρακτηριστικών που περιλαμβάνει λεπτομέρειες του αγώνα όπως : το σκορ, τα σουτ εκτός στόχου και εντός στόχου, τα κόρνερ, το ελεύθερα λάκτισμα, κατοχή μπάλας και φάουλ. Παρά το μικρό σύνολο δεδομένων βρήκαν ακρίβεια 77% για τους αγώνες μετά τα στάδια του ομίλου [7].

Τέλος, οι Tax and Joustra είναι η μόνοι που ενσωμάτωσαν τις αποδόσεις στοιχημάτων στο μοντέλο τους. Εκτός από αυτές τις πιθανότητες, επίσης ενσωμάτωσαν τα γκολ υπέρ, γκολ εναντίον, αποτελέσματα σε προηγούμενους αγώνες, κορυφαίους σκόρερ, ημέρες από τον προηγούμενο αγώνα και το αποτέλεσμα (νίκη, ισοπαλία, ήττα) της ομάδας σε ποσοστά. Με αυτό το διάνυσμα χαρακτηριστικών, κατάφεραν να προβλέψουν τους ποδοσφαιρικούς αγώνες της Ολλανδικής Eredivisie με 55%. Για να βελτιωθεί το μοντέλο τους, το σετ χαρακτηριστικών τους μπορεί να εμπλουτιστεί και να δοκιμαστεί για άλλα πρωταθλήματα [8] .

Η πρόβλεψη της έκβασης ενός αγώνα είναι σημαντική όμως η πρόβλεψη της απόδοσης της ομάδας για ολόκληρη σεζόν είναι σημαντικότερη. Είναι προφανές ότι αποτελεί μεγάλη πρόκληση η πρόβλεψη της μακροπρόθεσμη απόδοση μιας ομάδας και είναι ακόμα μεγαλύτερη πρόκληση η πρόβλεψη της απόδοσή της σε σύγκριση με την απόδοση άλλων ομάδων.

Περιορισμένη εργασία έχει πραγματοποιηθεί σε αυτό το δύσκολο έργο μέχρι στιγμής. Ένα από τα πιο ενδιαφέροντα αλλά και σχεδόν ανεξερεύνητα πεδία είναι η πρόβλεψη του τελικού πίνακα κατάταξης ενός πρωταθλήματος.

Οι Van Haaren και Davis τόνισαν τη δυσκολία πρόβλεψης ακριβής θέση μιας ομάδας στον τελικό πίνακα καθώς εξαρτάται από την τελική θέση κάθε άλλης ομάδας του πρωταθλήματος [9]. Άλλο ένα εμπόδιο για τους μέθοδος ήταν ο αριθμός των αγώνων που έληξαν με ισοπαλία. Τα συστήματα κατάταξης που χρησιμοποιούνται για την προσομοίωση αποτελεσμάτων αγώνων δυσκολεύονται στην πρόβλεψη ισοπαλίας. Με αποτέλεσμα να οδηγηθούν σε μεγάλη απόκλιση των προβλεπόμενων πόντων για κάθε ομάδα. Παρόλα αυτά, υπέδειξαν δύο ουσιαστικές μετρικές για την αξιολόγηση της ποιότητας των προβλεπόμενων βαθμολογικών πινάκων, το ποσοστό των σωστά προβλεπόμενων σχετικών θέσεων και το Μέσο Τετράγωνο Σφάλμα (MSE) σε σχέση με τις θέσεις αυτές.

Ο Oberstone ανέπτυξε ένα μοντέλο πολλαπλής παλινδρόμησης, καταλήγοντας σε έξι ανεξάρτητες μεταβλητές τις οποίες εκτίμησε ως επαρκείς για την πρόβλεψη του τελικού πίνακα της English Premier league προβλέποντας τους συνολικούς βαθμούς των ομάδων αντί των ακριβών θέσεων [10]. Χρησιμοποίησε επίσης την κατανομή F ώστε να διερευνήσει ποιες είναι εκείνες οι ενέργειες στο γήπεδο που διαφοροποιούν τις τέσσερις καλύτερες ομάδες από τις υπόλοιπες στο πρωτάθλημα επιτυγχάνοντας εξαιρετικά αποτελέσματα.

Στην παρούσα μελέτη, γίνεται απόπειρα της κατασκευής προβλεπτικών μοντέλων με αλγορίθμους μηχανικής μάθησης Naïve Bayes (MultinomialNB), Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest με στόχο την πρόβλεψη του τελικού πίνακα κατάταξης του Αγγλικού

πρωταθλήματος, προβλέποντας τις θέσεις των ομάδων με την καλύτερη δυνατή ακρίβεια.

3 Αλγόριθμοι επιβλεπόμενης ταξινόμησης

3.1 Λογιστική Παλινδρόμηση (Logistic Regression)

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

3.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM)

Η κατηγοριοποίηση των δεδομένων στηρίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου που διαχωρίζει τα δεδομένα δημιουργώντας το μέγιστο περιθώριο. Στην περίπτωση που ο γραμμικός διαχωρισμός είναι αδύνατος,

γίνεται χρήση κατάλληλων απεικονίσεων που μεταφέρουν το σύνολο των δεδομένων σε μεγαλύτερη διάσταση ώστε να επιτευχθεί τελικά ο διαχωρισμός τους. Η ικανότητα γενίκευσης της χρήσης των SVM σε μη γραμμικά δεδομένα στηρίζεται στο τέχνασμα του πυρήνα (kernel trick). Κάθε μηχανή διανυσμάτων υποστήριξης είναι ένας δυαδικός ταξινομητής, έχει δηλαδή τη δυνατότητα κατηγοριοποίησης σε δύο κλάσεις. Η συνάρτηση που υλοποιεί ο αλγόριθμος είναι : $g(X)= w^3 \varphi(X)+b$.

Το «X» είναι το διάνυσμα χαρακτηριστικών, το «w» είναι το βάρος του διανύσματος και το «b» είναι το διάνυσμα μεροληψίας. Η φ(X) είναι η μη γραμμική χαρτογράφηση από το χώρο εισόδου σε μεγάλο χώρο διαστάσεων [11].

3.3 Κ- Εγγύτεροι Γείτονες (K-Nearest Neighbors-KNN)

Ο K Nearest Neighbors (KNN) είναι μια τυπική μέθοδος μηχανικής μάθησης που έχει χρησιμοποιηθεί στην εξόρυξη δεδομένων. Η ιδέα είναι ότι χρησιμοποιείται ένα μεγάλο μέρος των δεδομένων εκπαίδευσης, όπου κάθε σημείο δεδομένων χαρακτηρίζεται από ένα σύνολο μεταβλητών. Εννοιολογικά, κάθε σημείο σχεδιάζεται σε ένα χώρο μεγάλης διαστάσεως, όπου κάθε άξονας στον χώρο αντιστοιχεί σε μια μεμονωμένη μεταβλητή. Όταν έχουμε ένα νέο δοκιμαστικό σημείο δεδομένων, θέλουμε να μάθουμε τους πλησιέστερους γείτονες του Κ δηλαδή, περισσότερο όμοιους. Ο αριθμός Κ επιλέγεται τυπικά ως η τετραγωνική ρίζα του Ν, δηλαδή ο συνολικός αριθμός των σημείων του συνόλου εκπαίδευσης δεδομένων. Η KNN έχει το πλεονέκτημα ότι είναι μη παραμετρική, δηλαδή μπορεί να χρησιμοποιηθεί ακόμα και όταν οι μεταβλητές είναι κατηγορηματικές. Συνήθως, χρησιμοποιείται κάποια μορφή προ παραγωγής, για παράδειγμα, η αναζήτηση ευρετηρίου (indexing). Αντί να χρησιμοποιηθούν όλα τα σημεία δεδομένων, μπορούν να χρησιμοποιηθούν επιλεγμένα σημεία δεδομένων που αντιπροσωπεύουν μεμονωμένα τμήματα σημείων για να διευκολύνουν την αναζήτηση ενάντια σε ένα νέο στοιχείο, και στη συνέχεια εμφανίζονται τα γειτονικά σημεία με το πιο όμοιο πρωτότυπο [12].

3.4 Δέντρα απόφασης (Decision Trees)

Τα δέντρα απόφασης είναι οι πιο εξελιγμένες μέθοδοι για τη διαίρεση ομάδων αντικειμένων σε κατηγορίες (class). Ένα δέντρο απόφασης ταξινομεί τα στοιχεία δεδομένων σε ένα πεπερασμένο αριθμό προκαθορισμένων κλάσεων. Οι κόμβοι δέντρων φέρουν ετικέτες με τα ονόματα των χαρακτηριστικών, τα τόξα φέρουν ετικέτες με τις πιθανές τιμές του χαρακτηριστικού και τα φύλλα επισημαίνονται με τις διαφορετικές κλάσεις.

Τα δέντρα αποφάσεων εισήχθησαν στο σύστημα ID3 [13], ως ένας από τους πρώτους αλγορίθμους εξόρυξης δεδομένων. Ένα αντικείμενο ταξινομείται ακολουθώντας μια διαδρομή κατά μήκος του δέντρου που σχηματίζεται από τα τόξα που αντιστοιχούν στις τιμές των χαρακτηριστικών του. Ένας απόγονος του ID3 που χρησιμοποιείται συχνά σήμερα για την οικοδόμηση των δέντρων αποφάσεων είναι ο αλγόριθμος C4.5 [13]. Λαμβάνοντας υπόψη ένα σύνολο C στοιχείων, ο αλγόριθμος C4.5 αναπτύσσει πρώτα ένα δέντρο απόφασης χρησιμοποιώντας τον αλγόριθμο «διαίρει και βασίλευε». Τα χαρακτηριστικά μπορούν να είναι είτε αριθμητικά (numeric) είτε ονομαστικά (nominal) και αυτό καθορίζει τη μορφή των αποτελεσμάτων των δοκιμών. Για ένα αριθμητικό χαρακτηριστικό A όπου $\{A \leq t,\ A > t\}$ όπου το όριο t βρίσκεται με κατώφλι (threshold) C στις τιμές του A και επιλέγοντας τη διαίρεση μεταξύ διαδοχικών τιμών που μεγιστοποιεί το κριτήριο αυτό. Ένα χαρακτηριστικό A με διακριτές τιμές έχει προεπιλογή ένα αποτέλεσμα για κάθε τιμή, αλλά μια επιλογή επιτρέπει την ομαδοποίηση των τιμών σε δύο ή περισσότερα υποσύνολα με ένα αποτέλεσμα για κάθε υποσύνολο. Το αρχικό δέντρο στη συνέχεια κλαδεύεται (pruning) για να αποφευχθεί η υπερφόρτωση (overfitting).

Ένα μειονέκτημα των δέντρων αποφάσεων είναι ότι αυξάνονται υπερβολικά όσον αφορά τις διαστάσεις τους σε πραγματικές εφαρμογές εξόρυξης δεδομένων, με αποτέλεσμα την δυσκολία κατανόησή τους.

3.5 Τυχαία Δάση (Random Forest)

Τα random forests αποτελούν [12] την γενίκευση των decision trees, όπου η εκτίμηση για κάθε κόμβο προκύπτει ως το μέσο από τις εκτιμήσεις που δίνουν για αυτό τον κόμβο, ένα μεγάλο σύνολο από random trees (και τα οποία δημιουργούν το forest) [13]. Με αυτό τον τρόπο μειώνεται η διασπορά των εκτιμήσεων και επομένως και τα σφάλματα της μεθόδου [16]. Για να επιτευχθεί η μείωση της διασποράς στις εκτιμήσεις και ο υψηλός βαθμός ακρίβειας της εκτίμησης, θα πρέπει τα ξεχωριστά δέντρα που αποτελούν το forest να είναι όσο πιο ασυσχέτιστα γίνεται. Για αυτό τον σκοπό, στους αλγόριθμους των random forests:

- Χρησιμοποιείται συνήθως η μέθοδος bootstrap aggregating (ή bagging), δηλαδή η κατασκευή των επιμέρους trees γίνεται βάσει τυχαία επιλεγμένων υποσυνόλων του πληθυσμού.
- Χρησιμοποιείται συνήθως η μέθοδος random subspace method (ή attribute bagging), ώστε σε κάθε σπάσιμο να εξετάζεται ένα τυχαίο υποσύνολο των χαρακτηριστικών.

Σαν αποτέλεσμα, για την πρακτική χρήση των random forests, πρέπει συνήθως να ορισθούν [16] οι παρακάτω τρεις παράμετροι (hyperparameters):

- Ο αριθμός των decision trees για το forest.
- Ο αριθμός των τυχαία επιλεγμένων μεταβλητών που θα εξετάζονται σε κάθε σπάσιμο.
- Ο ελάχιστος αριθμός παρατηρήσεων που θα πρέπει να έχει κάθε τερματικός κόμβος, που καθορίζει και την πολυπλοκότητα των δέντρων.

Για τον ορισμό των δύο τελευταίων παραμέτρων και για να βελτιστοποιηθεί η ακρίβεια του μοντέλου, μπορεί να χρησιμοποιηθεί και η μέθοδος του cross-validation.

Τα random forests είναι πολύ διαδεδομένα και στα πλεονεκτήματά τους αναφέρονται:

- Μείωση του overfitting σε σχέση με τα απλά decision trees.

- Μοντελοποίηση γραμμικών αλλά και μη γραμμικών σχέσεων.
- Αρκετά καλές και ακριβείς εκτιμήσεις.
- Εξέταση πολλών πιθανών μεταβλητών (high dimensionality).

Η κριτική που υπάρχει για την χρήση τους είναι:

- Δεν υπάρχει διαφάνεια και έλεγχος στο πως λειτουργεί το μοντέλο, εκτός από τον ορισμό των παραμέτρων. Λειτουργεί σαν ‘black box’.
- Δουλεύει καλύτερα σε προβλήματα classification ενώ σε regression προβλήματα μπορεί να εμφανισθούν θέματα overfitting

3.6 Naive Bayes

Ο ταξινομητής Naive Bayes είναι ένας από τους πιο τυπικούς και δημοφιλής αλγορίθμους κατηγοριοποίησης /ταξινόμησης οι οποίοι βασίζονται στο θεώρημα του Thomas Bayes. Οι αλγόριθμοι αυτοί μοιράζονται μια κοινή αρχή, δηλαδή κάθε ζεύγος χαρακτηριστικών που ταξινομείται είναι ανεξάρτητο το ένα από το άλλο. Η βασική ιδέα του συγκεκριμένου κατηγοριοποιητή είναι ότι τα στοιχεία /χαρακτηριστικά του διανύσματος είναι στατιστικά ανεξάρτητα. Ο αλγόριθμος αυτός είναι γρήγορος σε συνθήκες πραγματικού χρόνου. Μπορεί να κωδικοποιηθεί εύκολα και οι προβλέψεις του γίνονται σε τάχιστο χρόνο.

3.7 Τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks)

Τα Νευρωνικά Δίκτυα, επίσης γνωστά ως Τεχνητά Νευρωνικά Δίκτυα (ANN), είναι συστήματα που βασίζονται σε μια συλλογή κόμβων (νευρώνες) που μοντελοποιούν σε αλγοριθμικό επίπεδο το συνδέσεις μεταξύ νευρώνων στον ανθρώπινο εγκέφαλο. Κάθε νευρώνας μπορεί να λάβει ένα σήμα από νευρώνες και να το μεταδώσει σε άλλους νευρώνες. Δύο νευρώνες συνδέονται από μια άκρη που έχει ένα βάρος που της αποδίδεται, το οποίο διαμορφώνει τη σημασία της εισόδου αυτού του νευρώνα στην έξοδο του άλλου νευρώνα. Ένα νευρωνικό δίκτυο αποτελείται συνήθως από ένα στρώμα εισόδου, με έναν νευρώνα ανά μεταβλητή εισόδου για το μοντέλο, ένα στρώμα εξόδου, αποτελούμενο από έναν

μόνο νευρώνα που θα δώσει το αποτέλεσμα ταξινόμησης ή παλινδρόμησης, και ένας αριθμός κρυφών επιπέδων μεταξύ των δύο, που περιέχει έναν μεταβλητό αριθμό νευρώνων σε κάθε στρώμα.

4 Ορισμός του προβλήματος

Στόχος της παρούσας εργασίας είναι η ανάπτυξη ισχυρών μοντέλων μηχανικής μάθησης Naïve Bayes (MultinomialNB), Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest και νευρωνικού δικτύου, με σκοπό την πρόβλεψη της νικητήριας ομάδας του αγγλικού πρωταθλήματος με την υψηλότερη δυνατή ακρίβεια, χρησιμοποιώντας και αξιολογώντας την απόδοση των σχετικών με το πρόβλημα, αλγορίθμων.

5 Σύνολο δεδομένων και επεξεργασία

5.1 Το Αγγλικό Πρωτάθλημα (English Premier League)

- **Γενικά Στοιχεία**

Το πρωτάθλημα πραγματοποιείται στο διάστημα μεταξύ Αυγούστου και Μαΐου και περιλαμβάνει όλες τις ομάδες, για συνολικά 38 αγώνες κάθε σεζόν. Συμμετέχουν συνολικά 20 ομάδες σε κάθε σεζόν από την σεζόν 1995-1996 όπου και καθιερώθηκε η τρέχουσα μορφή. Μέχρι το 2002-03 η μόνη προθεσμία μεταγραφών κάθε σεζόν έληγε στις 31 Μαρτίου. Οι ομάδες μπορούσαν να αποκτήσουν παίκτες καθ' όλη τη διάρκεια της σεζόν, εκτός τους τελευταίους δύο μήνες σε αντίθεση με την λογική των δύο μεταγραφικών περιόδων- η πρώτη πριν την έναρξη της εκάστοτε περιόδου με διάρκεια από την 10η Ιουνίου έως και την 1η Σεπτεμβρίου και η ενδιάμεση κατά την διάρκεια της σεζόν με διάρκεια από την 1η έως την 31η Ιανουαρίου.

- **Κανονισμοί και Βαθμολογικό Σύστημα**

Τρεις πόντοι απονέμονται για τη νίκη, ένας βαθμός για την ισοπαλία και κανένας βαθμός για την ήττα, με την ομάδα με τους περισσότερους πόντους στο τέλος της σεζόν να κερδίζει τον τίτλο της Premier League και να αναδεικνύεται πρωταθλήτρια.

5.2 Σύνολο Δεδομένων (Dataset)

Το σύνολο δεδομένων αφορά σε χρονικό διάστημα 22 ετών και καλύπτει τις σεζόν 2000-2022 αποτελούμενο από συνολικά 440 εγγραφές (20 ομάδες επί 22 έτη). Η επιλογή του χρονικού διαστήματος αυτού έγινε με γνώμονα την μέγιστη δυνατή πληρότητα των δεδομένων που απαιτήθηκαν για την μελέτη καθώς και την συνοχή των κανονισμών που αφορούν στον αριθμό των ομάδων που συμμετέχουν στο πρωτάθλημα, τις μεταγραφικές περιόδους κ.α. Κατασκευάστηκε από το μηδέν με την μέθοδο web-scrapping με τα αρχικά δεδομένα να ανακτώνται από την ιστοσελίδα transfermarkt και κάποια συμπληρωματικά από την wikipedia καθώς δεν υπήρχε κάποιο διαθέσιμο

dataset σε γνωστά αποθετήρια που να καλύπτει όλο το χρονικό διάστημα μελέτης. Χρησιμοποιήθηκε η γλώσσα προγραμματισμού python και συγκεκριμένα οι βιβλιοθήκες selenium και bs4 (BeautifulSoup).

Από την παραπάνω μέθοδο δημιουργήθηκαν πρωτογενή αρχεία με δεδομένα που αφορούν στις ομάδες(οικονομικά στοιχεία, αποδόσεις, αριθμός παικτών κτλ), στους παίκτες(ηλικία, εθνικότητα κτλ) καθώς και στις μεταγραφές τους μεταξύ των ομάδων ανά σεζόν όπως και στους αγώνες(αριθμός οπαδών που παρακολούθησαν στα γήπεδα).

5.3 Κατασκευή Χαρακτηριστικών (Feature Engineering)

Ορισμένα χαρακτηριστικά δημιουργήθηκαν από τα πρωτογενή αρχεία και άλλα εντάχθηκαν αυτούσια όπως προέκυψαν από την διαδικασία του web scrapping (αφορά στα χαρακτηριστικά *pos, team, gf, ga, gd, pts, avg_att, total_market_value*) τα οποία περιγράφονται στον πίνακα που ακολουθεί :

Χαρακτηριστικό(Feature)	Περιγραφή / Τύπος
pos	Η θέση της ομάδας στο πρωτάθλημα για την κάθε σεζόν / αριθμητικό
team	Το όνομα της ομάδας / κατηγορικό
gf	Τα γκολ που έχει επιτύχει συνολικά σε μια σεζόν / αριθμητικό
ga	Τα γκολ που έχει δεχτεί συνολικά σε μια σεζόν / αριθμητικό
gd	Η διαφορά μεταξύ των γκολ που έχει επιτύχει και δεχτεί / αριθμητικό
pts	Οι συνολικοί πόντοι βαθμολογίας σε μια σεζόν / αριθμητικό
season	Σε ποια σεζόν αντιστοιχούμε
avg_att	Ο μέσος όρος των θεατών που παρακολούθησαν τους αγώνες δια ζώσης σε μία σεζόν / αριθμητικό
avg_time	Αφορά την συνοχή της ομάδας δλδ τον χρόνο που κάθε παίκτης της βρίσκεται εκεί από την ημέρα που πήγε μέχρι και την αποτιμώμενη στιγμή. / αριθμητικό
years_in_epl	Ο αριθμός των ετών που συμμετέχει η ομάδα στο πρωτάθλημα / αριθμητικό
big6	Εάν βρίσκεται βαθμολογικά μεταξύ των 6 πρώτων ομάδων συνολικά για όλο το διάστημα μελέτης/ αριθμητικό
num_trans_in	Ο αριθμός των παικτών που εισήχθησαν στην ομάδα για την σεζόν / αριθμητικό
num_trans_out	Ο αριθμός των παικτών που έφυγαν από την ομάδα για την σεζόν / αριθμητικό

squad	Ο συνολικός αριθμός των παικτών ανά σεζόν / αριθμητικό
age	Ο μέσος όρος ηλικίας των παικτών μιας ομάδας ανά σεζόν / αριθμητικό
foreigners	Ο αριθμός των παικτών άλλων εθνικοτήτων ανά σεζόν / αριθμητικό
total_market_value	Η αποτιμώμενη αξία της ομάδας / αριθμητικό

Πίνακας 5.3-1 : Τα χαρακτηριστικά του dataset

Το χαρακτηριστικό *avg_time* αφορά αποτιμώμενη την συνοχή της ομάδας δηλαδή τον χρόνο που κάθε παίκτης της βρίσκεται εκεί από την ημέρα που ενεργοποιήθηκε το συμβόλαιο του μέχρι και την αποτιμώμενη στιγμή.

Το χαρακτηριστικό *years_in_epl* είναι ο αριθμός των ετών που συμμετέχει η ομάδα στο πρωτάθλημα για όλες τις σεζόν που εξετάζονται και στόχος του είναι να αναδείξει ομάδες που έχουν ανέβει κατηγορία πρόσφατα από εκείνες που διαχρονικά διαγωνίζονται και έχουν μεγαλύτερη εμπειρία.

Το χαρακτηριστικό *big6* αφορά στις ομάδες (Arsenal, Chelsea, Liverpool, Manchester United, Manchester City, Tottenham Hotspur) οι οποίες διαθέτουν τα μεγαλύτερα στάδια, έχουν τις μεγαλύτερες ομάδες οπαδών οι οποίοι παρακολουθούν τους αγώνες τους τόσο εντός όσο και εκτός έδρας και άρα και τα μεγαλύτερα έσοδα από τις υπόλοιπες ομάδες του πρωταθλήματος. Χρησιμοποιήθηκε η δυαδική μορφή με 1 για τις ομάδες που πληρούν τις προϋποθέσεις και 0 για τις υπόλοιπες.

Τα επιμέρους αρχεία που παρήχθησαν συνδυάστηκαν σε ένα ενιαίο αρχείο που αποτελεί και το τελικό σύνολο δεδομένων με χρήση της γλώσσας προγραμματισμού Python

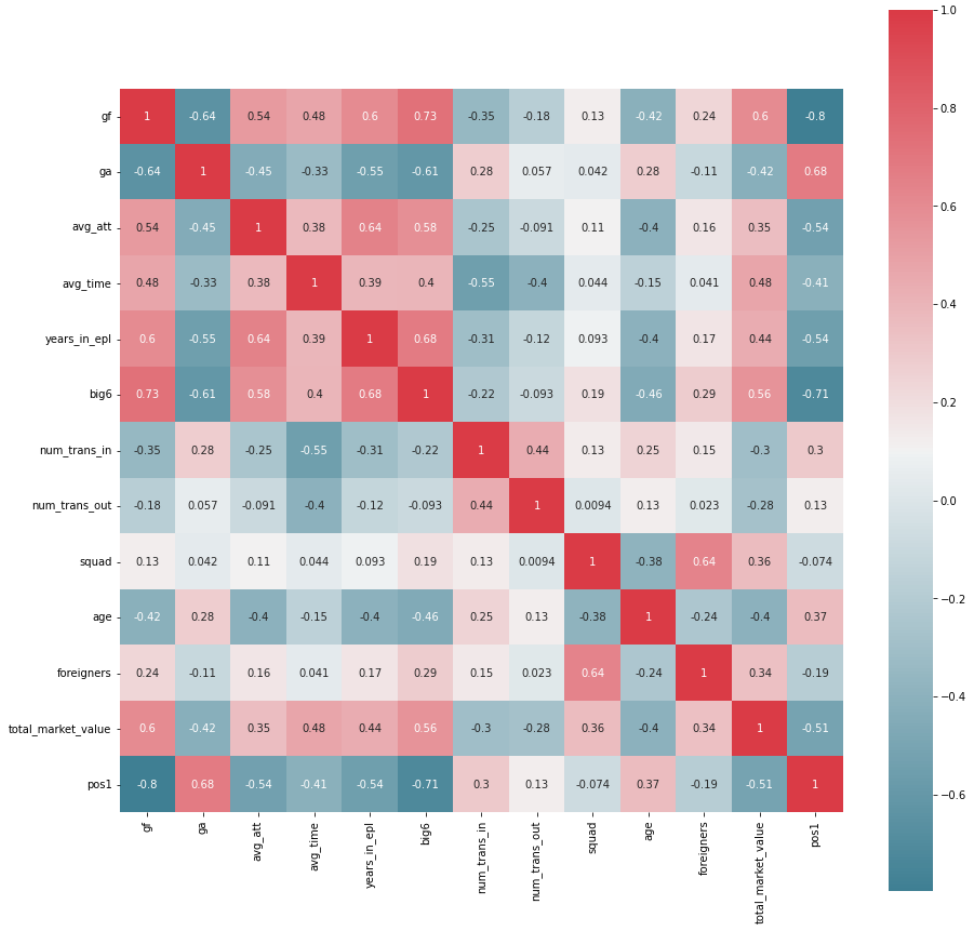
5.4 Επιλογή Χαρακτηριστικών (Feature Selection)

Η επιλογή χαρακτηριστικών είναι η διαδικασία που χρησιμοποιείται για την επιλογή των μεταβλητών εισόδου που είναι πιο σημαντικές για τα μοντέλα επιβλεπόμενης μηχανικής μάθησης. Με αυτόν τον τρόπο δύναται να βελτιωθεί η ακρίβεια του μοντέλου, εξασφαλίζεται μικρότερο υπολογιστικό κόστος και καθίσταται πιο εύκολο να κατανοηθεί και να εξηγηθεί.

- **Αρχική επιλογή χαρακτηριστικών με χρήση Correlation Matrix**

Ένας πίνακας συσχέτισης είναι ένας πίνακας που εμφανίζει τους συντελεστές συσχέτισης για διαφορετικές μεταβλητές. Ο πίνακας απεικονίζει τη συσχέτιση μεταξύ όλων των πιθανών ζευγών τιμών σε έναν πίνακα. Είναι ένα ισχυρό εργαλείο για τη σύνοψη ενός μεγάλου συνόλου δεδομένων και τον εντοπισμό και την οπτικοποίηση μοτίβων στα δεδομένα.

Αποτελείται από γραμμές και στήλες που δείχνουν τις μεταβλητές. Κάθε κελί σε έναν πίνακα περιέχει τον συντελεστή συσχέτισης η οποία κυμαίνεται από [-1 ,1]. Τιμές κοντά στο μηδέν υποδηλώνουν ότι δεν υπάρχει γραμμική τάση μεταξύ των δύο μεταβλητών. Όσο πιο κοντά στο 1 είναι η συσχέτιση, τόσο πιο θετικά συσχετίζονται και τόσο ισχυρότερη είναι η σχέση τους, δηλαδή αν αυξηθεί το ένα θα αυξηθεί και το άλλο. Μια συσχέτιση πιο κοντά στο -1 είναι αντίστροφη δηλαδή εάν η μία μεταβλητή αυξηθεί, η άλλη θα μειωθεί.



Γράφημα 5.4-1: Correlation Matrix

Από το παραπάνω correlation matrix προκύπτει ότι οι μεταβλητές *squad*, *num_trans_out* και *foreigners* εμφανίζουν χαμηλή συσχέτιση.

- **Sequential Feature Selection-SFS**

Πρόκειται για μια τεχνική επιλογής χαρακτηριστικών που χρησιμοποιείται στη μηχανική εκμάθηση για την επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών από ένα μεγαλύτερο σύνολο χαρακτηριστικών. Αυτή η τεχνική περιλαμβάνει την επαναληπτική επιλογή χαρακτηριστικών, είτε προς τα εμπρός είτε προς τα πίσω.

Forward

Η διαδικασία επιλογής ξεκινά με ένα κενό σύνολο χαρακτηριστικών και προσθέτει επαναληπτικά το πιο πολλά υποσχόμενο χαρακτηριστικό στο σύνολο, με βάση κάποια μέτρηση αξιολόγησης, όπως η ακρίβεια ή το ποσοστό σφάλματος. Σε κάθε βήμα, ο αλγόριθμος αξιολογεί την απόδοση του μοντέλου με το πρόσθετο χαρακτηριστικό και επιλέγει το χαρακτηριστικό που βελτιώνει περισσότερο την απόδοση. Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής, όπως ένας προκαθορισμένος αριθμός χαρακτηριστικών ή ένα όριο απόδοσης.

Backward

Πρόκειται για την αντίστροφη διαδικασία ξεκινώντας με ολόκληρο το σύνολο χαρακτηριστικών και αφαιρεί επαναληπτικά το λιγότερο σημαντικό χαρακτηριστικό από το σύνολο, με βάση κάποια μέτρηση αξιολόγησης. Σε κάθε βήμα, ο αλγόριθμος αξιολογεί την απόδοση του μοντέλου μετά την αφαίρεση ενός χαρακτηριστικού και επιλέγει το χαρακτηριστικό προς κατάργηση που προκαλεί τη μικρότερη μείωση της απόδοσης. Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής, όπως ένας προκαθορισμένος αριθμός χαρακτηριστικών ή ένα όριο απόδοσης.

Τόσο οι τεχνικές διαδοχικής επιλογής χαρακτηριστικών προς τα εμπρός όσο και προς τα πίσω μπορούν να είναι χρήσιμες για τη μείωση του αριθμού των χαρακτηριστικών σε ένα σύνολο δεδομένων και τη βελτίωση της απόδοσης και της ερμηνείας των μοντέλων μηχανικής εκμάθησης. Ωστόσο, μπορεί να είναι

υπολογιστικά ακριβή για μεγάλα σύνολα δεδομένων και μπορεί να μην αποδίδουν πάντα το βέλτιστο υποσύνολο χαρακτηριστικών.

Εφαρμόστηκαν οι παραπάνω μέθοδοι με την σειρά που περιγράφονται και τα αποτελέσματα από την επιλογή χαρακτηριστικών είναι :

Ο SFSF προτείνει :

```
Best accuracy score: 0.87
Best subset (indices): (0, 1, 3, 5, 7, 9, 10, 11)
Best subset (corresponding names): ('0', '1', '3', '5', '7', '9', '10', '11')
```

Εικόνα 5.4-1: Αποτελέσματα του αλγορίθμου Sequential Feauture Selection- Forward

Τα χαρακτηριστικά που θεωρεί σημαντικά είναι 8 και είναι τα *gf, ga, avg_time, big6, num_trans_out, age, foreigners,total_market_value.*

Και ο SFSB

```
Best accuracy score: 0.88
Best subset (indices): (0, 1, 3, 6, 8)
Best subset (corresponding names): ('0', '1', '3', '6', '8')
```

Εικόνα 5.4-2: Αποτελέσματα του αλγορίθμου Sequential Feauture Selection- Backward

Τα χαρακτηριστικά που θεωρεί σημαντικά είναι 5 και είναι τα *gf, ga, avg_time, big6, num_trans_in, squad.*

- **Feature Selection with ANOVA F-measure technique**

Η επιλογή χαρακτηριστικών με χρήση της μετρικής ANOVA F μέσω της συνάρτησης f_classif() είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για να επιλέξει σημαντικά χαρακτηριστικά από ένα σύνολο δεδομένων σε ένα πρόβλημα ταξινόμησης. Η μετρική ANOVA F είναι ένα στατιστικό τεστ που χρησιμοποιείται για να προσδιορίσει εάν οι μέσες τιμές δύο ή περισσότερων ομάδων είναι ίσες και μπορεί να χρησιμοποιηθεί για να αξιολογήσει τη σημαντικότητα της σχέσης μεταξύ κάθε χαρακτηριστικού και της μεταβλητής στόχου.

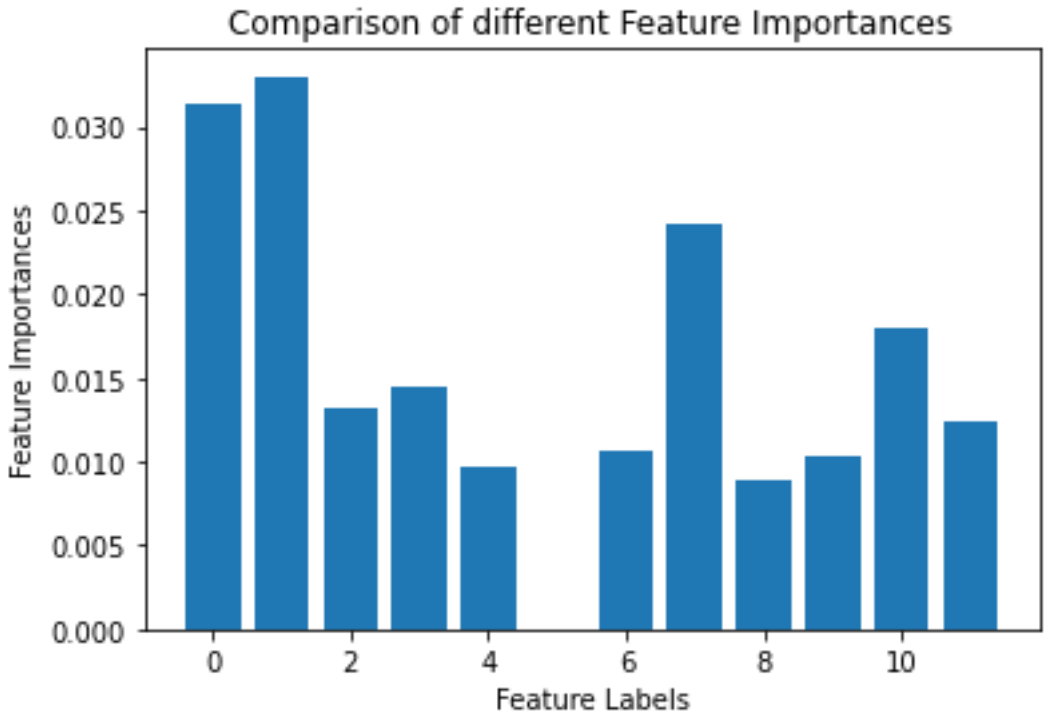
Η συνάρτηση f_classif() στη βιβλιοθήκη scikit-learn της Python είναι μια ενσωματωμένη μέθοδος επιλογής χαρακτηριστικών που χρησιμοποιεί τη

μετρική ANOVA F για να βαθμολογήσει κάθε χαρακτηριστικό στο σύνολο δεδομένων. Η συνάρτηση παίρνει ως παραμέτρους το σύνολο δεδομένων εισόδου και την αντίστοιχη μεταβλητή στόχο και επιστρέφει έναν πίνακα F-score και p-values για κάθε χαρακτηριστικό.

Η βαθμολογία F- measure μετρά τη διαφορά μεταξύ των μέσων τιμών κάθε κλάσης για ένα δεδομένο χαρακτηριστικό, ενώ η τιμή p υποδεικνύει τη στατιστική σημασία της διαφοράς. Μια υψηλότερη βαθμολογία F- measure και μια χαμηλότερη τιμή p υποδηλώνουν ότι το χαρακτηριστικό είναι πιο σχετικό με τη μεταβλητή στόχο και θα πρέπει να διατηρηθεί για περαιτέρω ανάλυση.

Η χρήση επιλογής χαρακτηριστικών με το ANOVA F-measure μέσω της συνάρτησης f_classif() μπορεί να βοηθήσει στη μείωση της διάστασης του συνόλου δεδομένων, στη βελτίωση της ακρίβειας και της απόδοσης του μοντέλου μηχανικής εκμάθησης και στην παροχή πληροφοριών για τα πιο σημαντικά χαρακτηριστικά για το πρόβλημα ταξινόμησης.

Τα αποτελέσματα που παρήχθησαν με την χρήση της μεθόδου φαίνονται στο διάγραμμα που ακολουθεί :



Γράφημα 5.4-2 : Αποτελέσματα αλγορίθμου ANOVA F-measure

Πρόκειται για τα χαρακτηριστικά : *gf, ga, avg_att, avg_time , num_trans_out, foreigners.*

Την τελική επιλογή των χαρακτηριστικών που επιλέχθηκαν ως δεδομένα εισόδου για τα μοντέλα μας απεικονίζει ο πίνακας που ακολουθεί :

Feature Selection	
gf	***ga***
avg_time	***avg_att***
big6	***num_trans_out***
age	***foreigners***
total_market_value	

Πίνακας 5.4-1 : Τα τελικά δεδομένα εισόδου

Τα χαρακτηριστικά αυτά επιλέχθηκαν λαμβάνοντας υπόψιν τα αποτελέσματα των παραπάνω μεθόδων αλλά και κάποια από αυτά εμπειρικά καθώς είναι γνωστό πως θεωρούνται σημαντικά για τους συλλόγους και τους μάνατζερ. Για παράδειγμα ο αριθμός των οπαδών που ακολουθεί την ομάδα στα εντός και εκτός έδρας παιχνίδια παρά τις καιρικές συνθήκες τονώνει το ηθικό της ή ο μέσος όρος ηλικίας θεωρείται σημαντικός για την φόρμα της επειδή οι μικρότεροι παίκτες έχουν περισσότερες αντοχές και επιθετικότητα και τέλος η «άξια» της ομάδας , το πόσο αποτιμάται συνολικά, μπορεί να καθορίζει τις μεταγραφές της τόσο σε οικονομικούς όρους, όσο και σε επίπεδο ελκυστικότητας για την καριέρα των παικτών.

6 Μεθοδολογία

6.1 Γενικά

Για την υλοποίηση των αλγορίθμων επιβλεπόμενης μηχανικής μάθησης και του νευρωνικού δικτύου ακολουθήθηκαν τα εξής βήματα:

1. Με την μεθοδολογία που περιγράφηκε στο προηγούμενο κεφάλαιο, πραγματοποιήθηκε ο προσδιορισμός των χαρακτηριστικών (features) που χρησιμοποιήθηκαν.
2. Χρησιμοποιήθηκαν οι κατάλληλες βιβλιοθήκες sklearn, TensorFlow και Keras σε γλώσσα προγραμματισμού python
3. Ελέγχθηκε το σύνολο των δεδομένων για κενές τιμές και αποκαταστάθηκαν με προβλέψεις που παρήχθησαν χρησιμοποιώντας την μέθοδο Linear Regression**.**
4. Επιλέχθηκαν τα κατάλληλα δεδομένα εισόδου με τις τεχνικές που περιγράφονται στο κεφάλαιο 5.4
5. Στην συνέχεια έγινε ο σχετικός μετασχηματισμός των κατηγορικών μεταβλητών σε δυαδική αριθμητική μορφή, μέσω της διαδικασίας One-Hot-Encoding και της ετικέτας με την χρήση του LaberEnconder () της βιβλιοθήκης sklearn και για το νευρωνικό δίκτυο, μετατράπηκε η ετικέτα σε one-hot encoded μεταβλητή με χρήση της μεθόδου keras to_categorically
6. Με χρήση της βιβλιοθήκης sklearn, χωρίσαμε τα δεδομένα σε train και test σετ.
7. Εξαιτίας της ανισορροπίας των κλάσεων , εφαρμόστηκε η τεχνική SMOTE επαναχωρίστηκαν τα δεδομένα σε train και test σετ.
8. Σε επόμενο βήμα, προκειμένου να καθοριστεί η αρχιτεκτονική του νευρωνικού δικτύου που κατασκευάσαμε, με την βοήθεια της βιβλιοθήκης sklearn και συγκεκριμένα με την μέθοδο GridSearchCV, αποπειραθήκαμε αρχικά να κατασκευάσουμε το μοντέλο μας, συντονίζοντας παράλληλα τις υπερπαράμετρους του δικτύου. Εξαιτίας του πολύ μεγάλου χρόνου που απαιτείται για την ολοκλήρωση της

διαδικασίας, χρησιμοποιήθηκε η μέθοδος HalvingGridSearchCV από την πειραματική βιβλιοθήκη της sklearn, η οποία έχει την ίδια ακριβώς ακρίβεια στο ένα δέκατο του χρόνου κατά προσέγγιση όπως υποστηρίζεται στην επίσημη σελίδα της, γεγονός που διαπιστώσαμε και εμείς στην πρώτη παράλληλη δοκιμή.

9. Κατασκευάστηκαν τα μοντέλα επιβλεπόμενης μηχανικής μάθησης με χρήση της βιβλιοθήκης sklearn
10. Επιλέχθηκαν οι κατάλληλες μετρικές για το πρόβλημα (multiclass ταξινόμηση) και με την μέθοδο GridSearchCV ρυθμίστηκαν οι υπερπαράμετροι και
11. Τέλος, έγινε η αξιολόγηση της πειραματικής διαδικασίας και η ανάλυση των αποτελεσμάτων αναδεικνύοντας τα πιο ακριβή μοντέλα

6.2 Προ-επεξεργασία

Στην συνέχεια ελέγχθηκε η πληρότητα των δεδομένων για κενές τιμές. Οι κενές τιμές που διαπιστώθηκε ότι υπάρχουν στο σύνολο των δεδομένων αφορούν στην στήλη που περιείχε τις τιμές της μεταβλητής *total_market_value* και ήταν 80 σε αριθμό. Αυτό οφείλεται στο ότι για τις πρώτες τέσσερις χρονιές δεν διατίθονταν τα στοιχεία αυτά στην βάση δεδομένων της σελίδας από όπου ανακτήθηκαν. Προκειμένου να επιλυθεί αυτό το ζήτημα οι κενές τιμές αντικαταστάθηκαν με προβλέψεις που παρήχθησαν χρησιμοποιώντας την μέθοδο Linear Regression.

Στην συνέχεια επειδή ο στόχος είναι η πρόβλεψη της θέσης της ομάδας στο πρωτάθλημα για την συγκεκριμένη σεζόν, για την μεταβλητή-στόχο υπήρχαν 20 διακριτές τιμές και άρα 20 κλάσεις, γεγονός που καθιστά το πρόβλημα υψηλής πολυπλοκότητας. Έτσι δημιουργήθηκαν τρεις νέες κλάσεις [1,2,3] οι οποίες ουσιαστικά αποτελούν bins. Τα bins αυτά περιέχουν :

Bins	
1	Όλες τις τιμές **pos**=1
2	Όλες τις τιμές **pos** [2-6]
3	Όλες τις τιμές **pos** [7-20]

Πίνακας 6.2-1: Οι νέες κλάσεις που δημιουργήθηκαν

Δημιουργήθηκε έτσι μια νέα μεταβλητή **pos1** με τιμές [1,2,3] η οποία και αποτελεί και την μεταβλητή-στόχο. Παρόλο που η μορφή των δεδομένων παρουσιάζεται ως αριθμητική παραμένει κατηγορική μεταβλητή και έτσι στην συνέχεια με χρήση της μεθόδου LabelEncoder(), μετατράπηκε σε αριθμητική μορφή [1 2 3]= [0 1 2].

Ακόμη, οι στήλες που αφορούν στα ονόματα των ομάδων είναι και αυτές της μορφής object/string και τις χρησιμοποιούμε ως εισόδους στο μοντέλο. Με τη βοήθεια της εντολής get_dummies() θα καταλήξουμε σε στήλες της μορφής: HomeTeam_Arsenal, AwayTeam_Newcastle και ούτω καθεξής. Οι στήλες αυτές έχουν δυαδικές τιμές: 0 για όλες τις ομάδες εκτός από αυτή που αφορούν τα δεδομένα της εκάστοτε γραμμής. Έτσι, μετά από αυτήν την διαδικασία ο αριθμός των χαρακτηριστικών που χρησιμοποιήθηκαν ως inputs ανέρχεται στα 52.

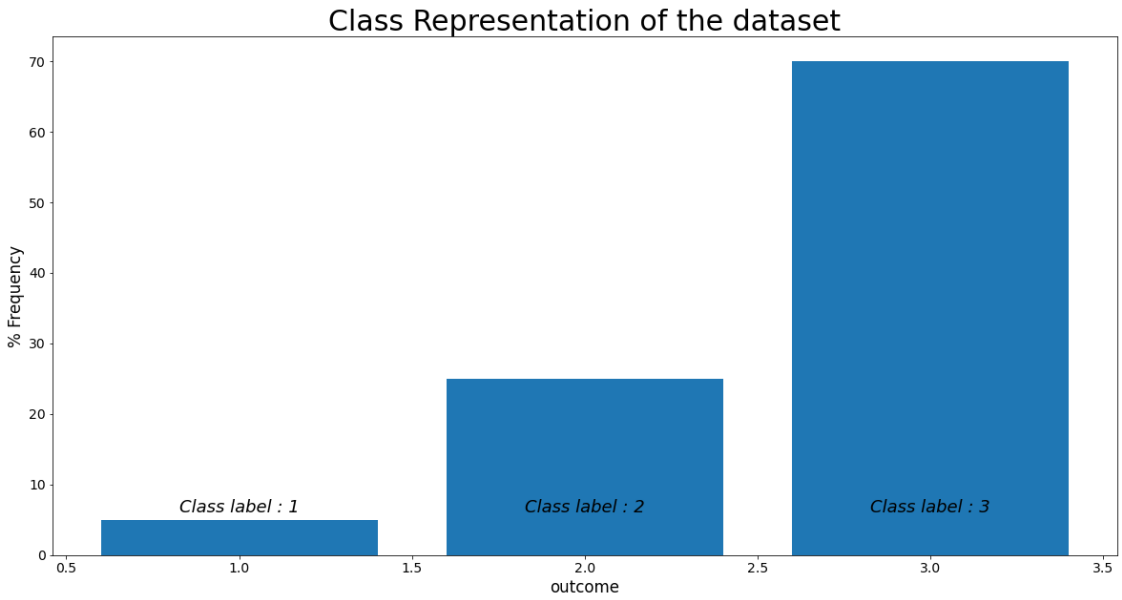
Επιπρόσθετα όλα τα αριθμητικά δεδομένα κανονικοποιήθηκαν σε εύρος [0 1] χρησιμοποιώντας τη μέθοδο StandardScaler().

Επιπλέον για την τελική εκτέλεση του νευρωνικού δικτύου, μετατράπηκε η ετικέτα σε one-hot encoded μεταβλητή με χρήση της μεθόδου keras to_categorically καθώς για την εισαγωγή της στους αλγορίθμους που εξετάστηκαν απαιτείται να βρίσκονται σε διανυσματική μορφή.

Τέλος με χρήση της βιβλιοθήκης sklearn, χωρίσαμε τα δεδομένα σε train και test σετ με αναλογία 80-20 χωρίς «τυχαίο ανακάτεμα».

6.3 Κατανομή των κλάσεων

Προκύπτει λοιπόν μετά την διαδικασία που περιγράφηκε άνωθεν, η εξής κατανομή των «νέων» κλάσεων όπως φαίνεται και στο σχήμα που ακολουθεί :



Γράφημα 6.3-1: Η κατανομή των κλάσεων

Παρατηρούμε όπως είναι αναμενόμενο ότι πλέον οι κλάσεις είναι σαφώς μη ισορροπημένες. Η ανακατανομή των δεδομένων είναι αυτή που φαίνεται στο παρακάτω διάγραμμα :

Γράφημα 6.3-2: Η κατανομή των δεδομένων πριν την εφαρμογή του SMOTE

- **Synthetic Minority Over-sampling Technique (SMOTE)**

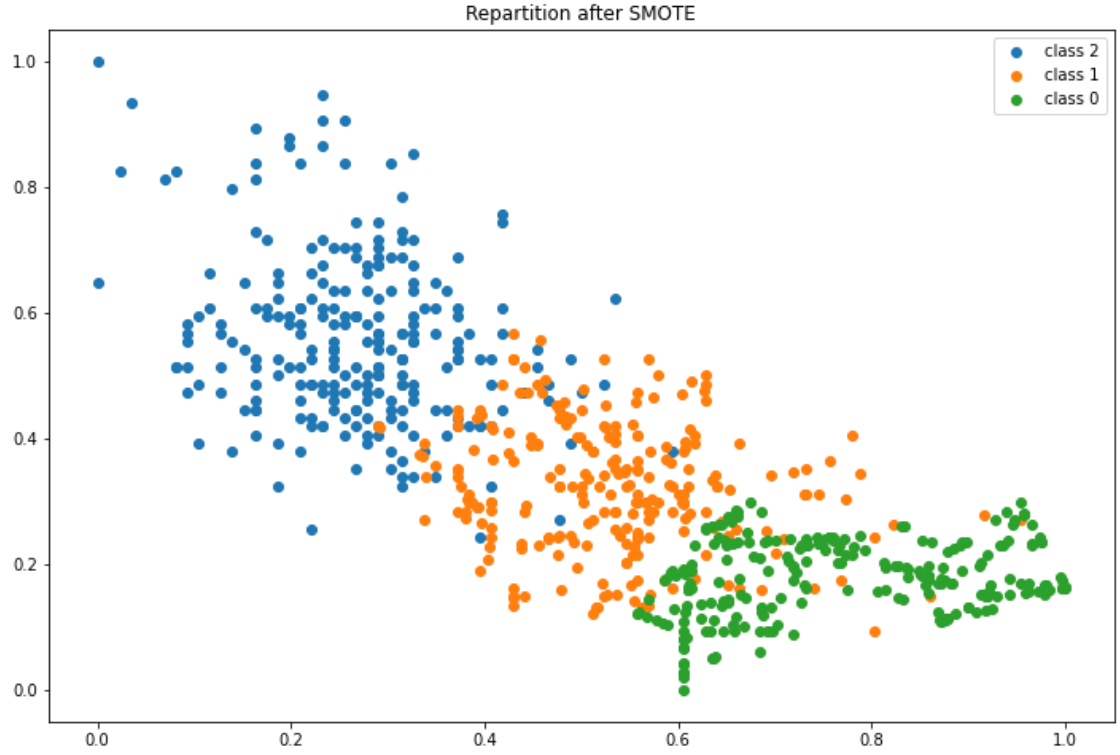
Είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την αντιμετώπιση του προβλήματος των μη ισορροπημένων συνόλων δεδομένων. Σε ένα σύνολο δεδομένων, όταν ο αριθμός των στιγμιότυπων που ανήκουν σε μια κλάση είναι σημαντικά χαμηλότερος από τον αριθμό των στιγμιότυπων που ανήκουν σε μια άλλη κλάση, το σύνολο δεδομένων λέγεται ότι είναι μη ισορροπημένο.

Το SMOTE δημιουργεί νέα συνθετικά δείγματα της κατηγορίας που μειοψηφεί παρεμβάλλοντας μεταξύ υπαρχόντων instances. Ο αλγόριθμος επιλέγει τυχαία μια περίπτωση της μειοψηφούσας κλάσης και βρίσκει τους k πλησιέστερους γείτονες της. Στη συνέχεια επιλέγει έναν από αυτούς τους γείτονες και δημιουργεί ένα νέο στιγμιότυπο επιλέγοντας ένα σημείο στο τμήμα της γραμμής που συνδέει τις δύο περιπτώσειπερις. Η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί το επιθυμητό επίπεδο υπερδειγματοληψίας της μειοψηφίας.

Το σύνολο δεδομένων που θα προκύψει θα έχει μια ισορροπημένη κατανομή στιγμιότυπων μεταξύ των κλάσεων, η οποία μπορεί να βελτιώσει την απόδοση

των αλγορίθμων μηχανικής μάθησης που είναι ευαίσθητοι στην ανισορροπία κλάσεων. Το SMOTE χρησιμοποιείται ευρέως σε διάφορους τομείς της μηχανικής μάθησης, συμπεριλαμβανομένου του εντοπισμού της απάτης, της ιατρικής διάγνωσης και της ανάλυσης πιστωτικού κινδύνου

Η επίλυση αυτού του προβλήματος, έγινε με την τεχνική **SMOTE** της βιβλιοθήκης imblearn. Στο γράφημα που ακολουθεί παρουσιάζεται η ανακατανομή των δεδομένων μετά την εφαρμογή της τεχνικής υπερδειγματοληψίας :



Γράφημα 6.3-3: Η κατανομή των δεδομένων μετά την εφαρμογή του SMOTE

6.4 Μετρικές

- **Accuracy**

Η accuracy εκτιμά την ορθότητα των αποτελεσμάτων μιας εργασίας. Αν υποθέσουμε ένα σύστημα, στο οποίο όλες οι σωστές απαντήσεις που υπάρχουν είναι X, και λαμβάνουμε απαντήσεις Y, εκ των οποίων οι N είναι σωστές, τότε η ακρίβεια ορίζεται σαν: ακρίβεια= N/Y. Η ακρίβεια είναι μια απλή μετρική που μας

επιτρέπει να κατανοήσουμε την απόδοση του μοντέλου ταξινόμησης, μέσω του ποσοστού παραδειγμάτων που έχει προβλεφθεί σωστά. Η ακρίβεια είναι πάντα μεταξύ 0 και 1.

- **Precision**

Precision είναι η αναλογία μεταξύ του αριθμού των θετικών δειγμάτων που ταξινομούνται σωστά στον συνολικό αριθμό των δειγμάτων που ταξινομήθηκαν ως θετικά.

- **F-mesure**

Η μετρική F-measure είναι ο αρμονικός μέσος των μετρικών της ακρίβειας (precision) και της ανάκλησης, λαμβάνοντας υπόψιν και τις δύο μετρήσεις. Χρησιμοποιούμε τον αρμονικό μέσο αντί για έναν απλό μέσο όρο, λόγω του ότι “τιμωρεί” τις ακραίες τιμές. Ένας ταξινομητής με ακρίβεια 1.0 και ανάκληση 0.0 έχει έναν απλό μέσο όρο 0.5, αλλά βαθμολογία F1 0. Το μέτρο F έχει μια διαισθητική έννοια. Εκφράζει πόσο ακριβής είναι ο ταξινομητής (πόσες περιπτώσεις ταξινομεί σωστά), καθώς και πόσο ισχυρός είναι (δεν χάνει σημαντικό αριθμό instances).

- **Cohen's kappa coefficient**

Ο συντελεστής k του Cohen (κ) είναι ένα στατιστικό μέτρο που χρησιμοποιείται για τη μέτρηση της αξιοπιστίας μεταξύ των τιμών για ποιοτικά (κατηγορικά) δεδομένα. Γενικά θεωρείται ότι είναι πιο ισχυρό μέτρο από τον απλό υπολογισμό του ποσοστού συμφωνίας, καθώς λαμβάνει υπόψιν τη πιθανότητα να συμβεί η συμφωνία τυχαία. Οι τιμές που παίρνει είναι στο εύρος [0,1] με τιμές πάνω από 0.6 να υποδεικνύουν καλή συμφωνία.

- **ROC (Area Under the Curve-AUC)**

Ένα άλλο δημοφιλές στατιστικό εργαλείο είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curves – ROC curves), οι οποίες εξ ορισμού χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός συστήματος με διχοτομικά εξαγόμενα αποτελέσματα. Παραδοσιακά, το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve-

AUC) χρησιμοποιείται ως συνοπτικός δείκτης ακρίβειας ενός τεστ, και είναι χρήσιμο ως περιγραφικό μέτρο της συνολικής απόδοσης ενός τεστ.

6.5 Αλγόριθμοι Μηχανικής Μάθησης- Μοντελοποίηση

Οι αλγόριθμοι μηχανικής μάθησης που υλοποιήθηκαν και αξιολογήθηκαν είναι οι *Naïve Bayes (MultinomialNB), Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest* οι οποίοι είναι κατάλληλοι για το multiclass classification πρόβλημα της παρούσας μελέτης και παρέχονται από την βιβλιοθήκη sklearn.

- **Grid Search CV**

Είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για τον συντονισμό των υπερπαραμέτρων ενός μοντέλου. Οι υπερπαράμετροι είναι παράμετροι ενός μοντέλου που δεν μαθαίνονται από τα δεδομένα αλλά ορίζονται πριν από την εκπαίδευση, όπως π.χ. ο ρυθμός εκμάθησης (learning rate). Η επιλογή των υπερπαραμέτρων μπορεί να έχει σημαντικό αντίκτυπο στην απόδοση ενός μοντέλου. Η διαδικασία περιλαμβάνει τον καθορισμό ενός εύρους τιμών για κάθε υπερπαράμετρο και στη συνέχεια την εκπαίδευση του μοντέλου σε όλους τους πιθανούς συνδυασμούς αυτών των τιμών, την αξιολόγηση της απόδοσης κάθε μοντέλου χρησιμοποιώντας ένα validation set και την επιλογή του συνδυασμού υπερπαραμέτρων που αποδίδει την καλύτερη απόδοση σύμφωνα με την επιλεχθείσα μετρική.

Στο πλαίσιο των αλγορίθμων ταξινόμησης, η αναζήτηση πλέγματος μπορεί να χρησιμοποιηθεί για τον συντονισμό υπερπαραμέτρων όπως ο αριθμός των κρυφών επιπέδων σε ένα νευρωνικό δίκτυο ή η παράμετρος λειτουργίας πυρήνα σε μια μηχανή διανύσματος υποστήριξης. Μπορεί να είναι ιδιαίτερα χρήσιμη τεχνική σε προβλήματα ταξινόμησης πολλαπλών κλάσεων όπου η επιλογή των υπερπαραμέτρων μπορεί να έχει σημαντικό αντίκτυπο στην ακρίβεια του μοντέλου.

Η εκτέλεση της τεχνικής αυτής, μπορεί να υποδείξει τον βέλτιστο συνδυασμό υπερπαραμέτρων που μεγιστοποιούν την ακρίβεια του μοντέλου σε ένα

validation set, βελτιώνοντας την ικανότητα του μοντέλου να γενικεύει σε νέα δεδομένα.

Χρησιμοποιώντας την τεχνική Grid Search CV για κάθε έναν από τους αλγορίθμους επιλέχθηκαν οι ακόλουθοι υπερπαράμετροι, όπως φαίνεται στον ακόλουθο πίνακα :

Classifier	Hyperparameters
Naïve Bayes	MultinomialNB(), alpha= 1e-05
Logistic Regression	C=100, penalty=I2, solver=newton-cg, multi_class='multinomial'
K-Neighbors	metric='manhattan', n_n= 10, weights= 'uniform',algorithm='ball_tree'
SVC	C=50, gamma=scale, kernel= rbf
Decision Tree	class_weight={0: 10, 1: 1}, max_depth= 5
Random Forest	criterion= 'entropy', max_depth= 9, n_estimators= 200

Πίνακας 6.5-1: Οι προτεινόμενες παράμετροι όπως προέκυψαν από την μέθοδο Grid Search CV

6.6 Αρχιτεκτονική ΝΔ– Μοντελοποίηση

Για να επιλέξουμε το βέλτιστο μοντέλο, δεν χρησιμοποιήσαμε τη κλασική τεχνική των πολλών δοκιμών για το προσδιορισμό των καταλληλότερων τιμών των παραμέτρων. Χρησιμοποιήθηκε η βιβλιοθήκη sklearn και συγκεκριμένα η μέθοδος HalvingGridSearchCV, η οποία μας πρότεινε ως καταλληλότερο μοντέλο αυτό που περιγράφεται αναλυτικά στην συνέχεια.

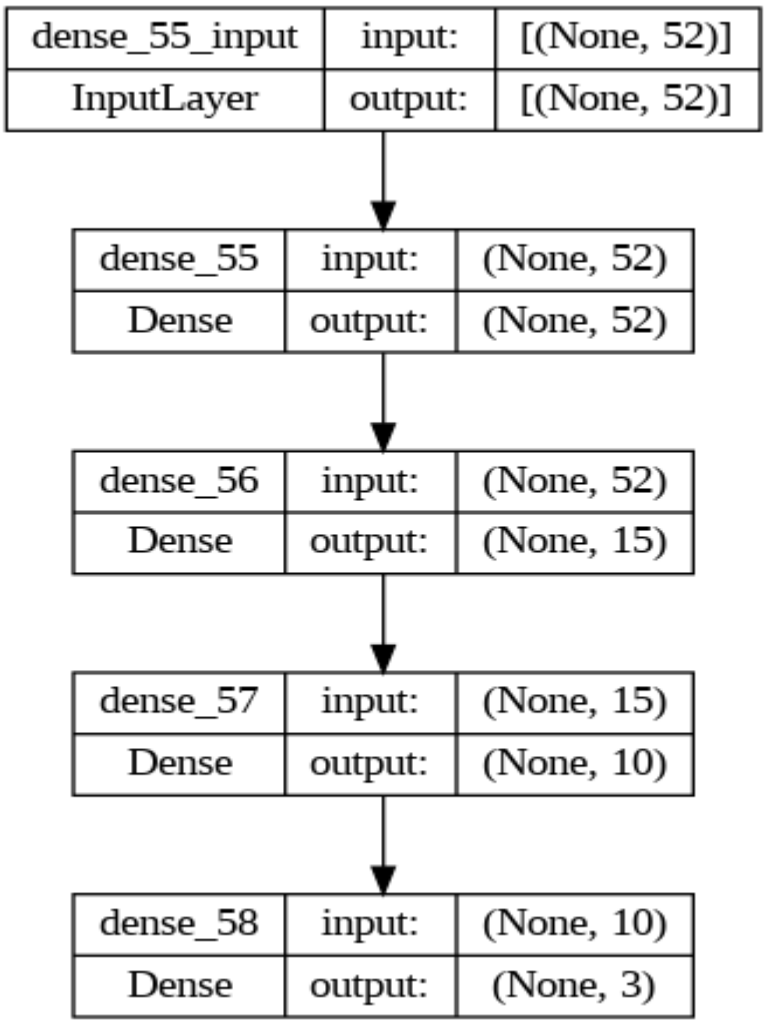
Το μοντέλο του νευρωνικού δικτύου που κατασκευάστηκε για την υλοποίηση της παρούσας εργασίας, αποτελείται από ένα input layer με 52 εισόδους (όσα και τα χαρακτηριστικά εισόδου) με 52 νευρώνες και συνάρτηση ενεργοποίησης relu, 2 hidden layers με 15 και 10 νευρώνες αντίστοιχα με την ίδια συνάρτηση ενεργοποίησης relu. Τέλος, το output layer με 3 εξόδους με συνάρτηση ενεργοποίησης softmax, καθώς είναι καταλληλότερη για προβλήματα multi-class classifications.

Αρχικά δοκιμάστηκε ο Adam optimizer, ο οποίος μαθαίνει με ταχύτερο ρυθμό από τους υπόλοιπους βελτιστοποιητές, είναι πιο σταθερός, δεν επιφέρει καμία

σημαντική μείωση στην ακρίβεια και έχει ευρύτερο φάσμα επιτυχημένων learning rates. Όμως εξαιτίας της κατανομής των κλάσεων, όπως συζητήθηκε παραπάνω, περισσότερο κατάλληλος φάνηκε ο SGDClassifier, ο οποίος υποστηρίζει τη ταξινόμηση πολλαπλών κλάσεων συνδυάζοντας πολλαπλούς ταξινομητές σε ένα one versus all σχήμα. Κατά τη φάση της αξιολόγησης, υπολογίζεται το confidence score για κάθε ταξινομητή και επιλέγεται η κλάση με το μεγαλύτερο confidence score.

Για το συγκεκριμένο πρόβλημα ταξινόμησης, χρησιμοποιήθηκε ως loss function η ενδεδειγμένη μέθοδος CategoricalCrossentropy, η οποία υπολογίζει ένα σκορ που συνοψίζει τη μέση διαφορά μεταξύ της πραγματικής και της προβλεπόμενης κατανομής πιθανότητας για όλες τις τάξεις του προβλήματος. Το σκορ ελαχιστοποιείται με βέλτιστη τιμή το 0.

Στην συνέχεια προχωρήσαμε στον τελικό συντονισμό του μοντέλου μας για τις υπερπαραμέτρους learning rate και momentum με την μέθοδο των δοκιμών.



Εικόνα 6.6-1: Αρχιτεκτονική νευρωνικού δικτύου

Τέλος, εκπαιδεύτηκε το μοντέλο μας με το training σετ και αξιολογήθηκε με βάση τα δεδομένα του test σετ και τις μετρικές που περιεγράφησαν παραπάνω δίνοντας περισσότερη βαρύτητα στην macro F-measure και του συντελεστή Cohen's kappa, εξαιτίας του ότι πρόκειται για multiclass classification πρόβλημα.

7 Αποτελέσματα πειραματικής μελέτης

7.1 Αποτελέσματα Logistic Regression

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκαν οι υπερπαράμετροι : C=100, penalty= I2, solver = newton-cg multi_class= multinomial.

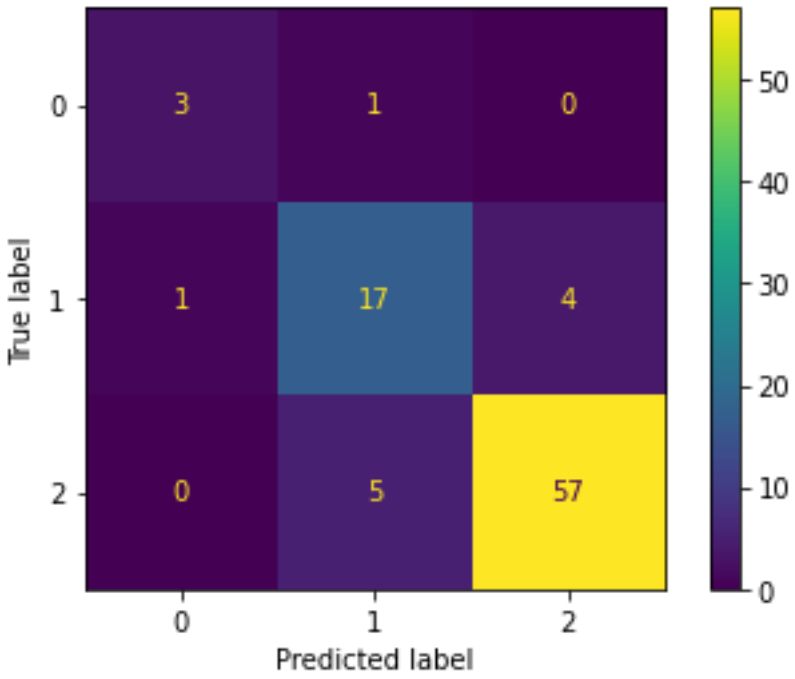
```
accuracy: 0.875
roc_auc_score: 0.9759698661253685
cohen_kappa_score: 0.7186046511627907
Classification Report :
              precision    recall  f1-score   support

           0       0.75      0.75      0.75         4
           1       0.74      0.77      0.76        22
           2       0.93      0.92      0.93        62

    accuracy                           0.88        88
   macro avg       0.81      0.81      0.81        88
weighted avg       0.88      0.88      0.88        88
```

Εικόνα 7.1-1: Classification Report – Logistic Regression

Το μοντέλο είναι αξιόπιστο λόγω του μεγάλου Kappa = 0.72 και το Confusion Matrix φαίνεται ότι πετυχαίνει μια σχετικά υψηλή ακρίβεια με το macro-F-Measure να είναι 81%. Το F-Measure είναι ο αρμονικός μέσος όρος του Precision και του Recall. Για την τελευταία κλάση των δεδομένων βλέπουμε ότι προσεγγίζουμε ακρίβεια πάνω από 90% ενώ για τις 2 πρώτες από 75% για την πρώτη και στο 74% για την δεύτερη.



Εικόνα 7.1-2: Confusion Matrix- Logistic Regression

Από τα 4 παραδείγματα στην πρώτη κλάση ταξινομήθηκαν σωστά τα 3 , 17 από τα 22 στην δεύτερη και 57 από τα 62 στην τρίτη κλάση.

7.2 Αποτελέσματα SVM

Για την ακρίβεια ο ταξινομητής είναι ο SVC, η built-in εκδοχή για multiclass προβλήματα της βιβλιοθήκης sklearn.

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκαν οι υπερπαράμετροι : C=50, gamma=scale, kernel=rbf

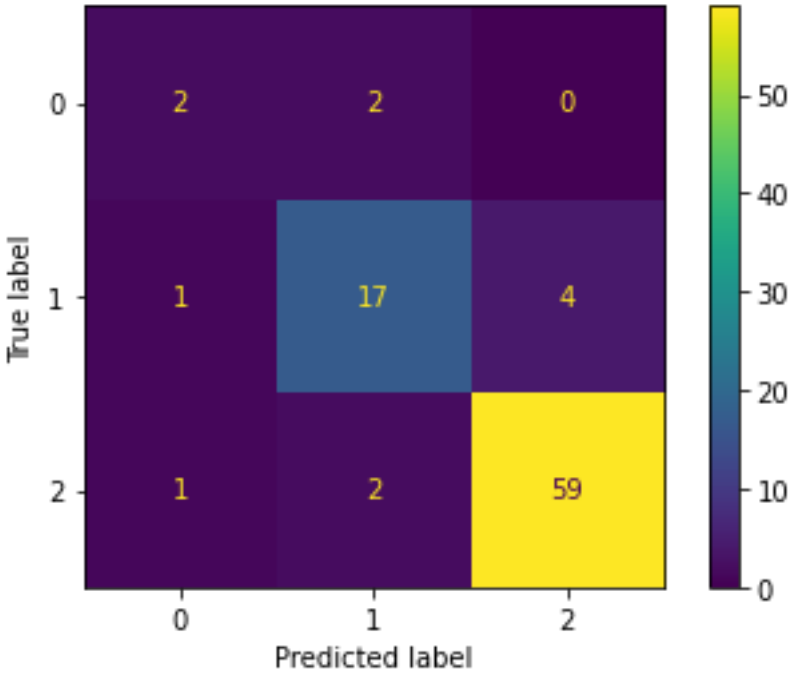
```
accuracy: 0.8863636363636364
roc_auc_score: 0.9621283839764702
cohen_kappa_score: 0.7380952380952381
Classification Report :
              precision    recall  f1-score   support

           0       0.50      0.50      0.50         4
           1       0.81      0.77      0.79        22
           2       0.94      0.95      0.94        62

    accuracy                           0.89        88
   macro avg       0.75      0.74      0.74        88
weighted avg       0.88      0.89      0.89        88
```

Εικόνα 7.2-1: Classification Report -SVC

Το μοντέλο είναι αξιόπιστο λόγω του μεγάλου Kappa = 0.74 και το Confusion Matrix φαίνεται ότι πετυχαίνει μια σχετικά καλή ακρίβεια με το macro-F-Measure να είναι 74%. Το F-Measure είναι ο αρμονικός μέσος όρος του Precision και του Recall. Για την τελευταία κλάση των δεδομένων βλέπουμε ότι προσεγγίζουμε ακρίβεια πάνω από 90% ενώ για τις 2 πρώτες από 50% για την πρώτη και στο 81% για την δεύτερη.



Εικόνα 7.2-2: Confusion Matrix-SVC

Από τα 4 παραδείγματα του συνόλου επαλήθευσης στην πρώτη κλάση ταξινομήθηκαν σωστά τα 2, 17 από τα 22 στην δεύτερη και 59 από τα 62 στην τρίτη κλάση.

7.3 Αποτελέσματα KNN

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκαν οι υπερπαράμετροι : metric=manhattan, n_neighbors= 10, weights= uniform, algorithm=ball_tree.

Όπως βλέπουμε και στο classification report που ακολουθεί η μέθοδος αυτή παρουσιάζει χαμηλή τιμή τόσο της μετρικής Kappa = 0.54 όσο και της macro-F-Measure που ισούται με 65%. Αυτό υποδεικνύει ότι δεν έχει αξιοπιστία και καλή απόδοση αντίστοιχα. Ειδικά για την πρώτη κλάση που είναι και αυτή που μας ενδιαφέρει περισσότερο η ακρίβεια είναι στο 40%, στην δεύτερη στο 64% και στην τρίτη πολύ καλύτερη γύρω στο 90%.

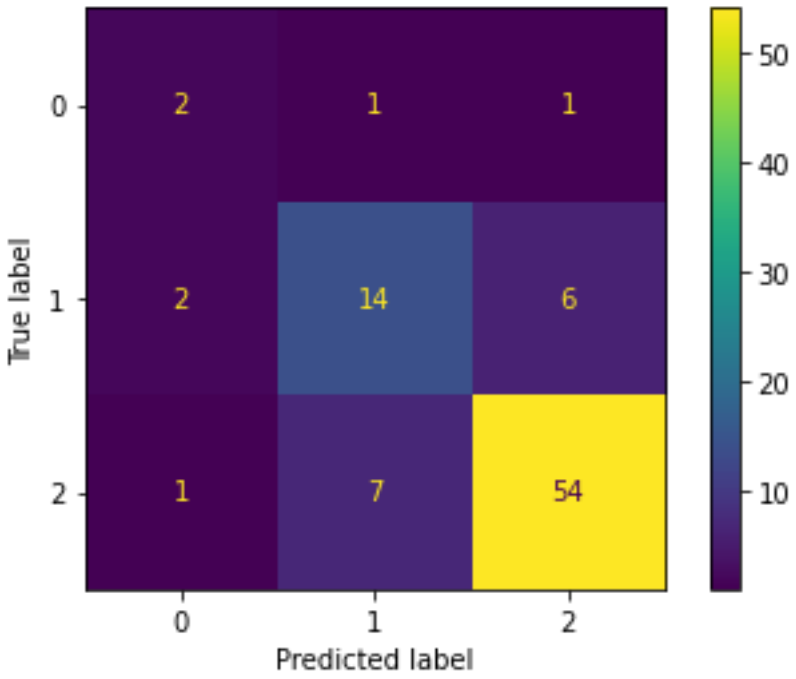
```
accuracy: 0.7954545454545454
roc_auc_score: 0.9621283839764702
cohen_kappa_score: 0.5419317524580682
Classification Report :
              precision    recall  f1-score   support

           0       0.40      0.50      0.44         4
           1       0.64      0.64      0.64        22
           2       0.89      0.87      0.88        62

    accuracy                           0.80        88
   macro avg       0.64      0.67      0.65        88
weighted avg       0.80      0.80      0.80        88
```

Εικόνα 7.3-1: Classification Report KNN

Από τα 4 παραδείγματα του συνόλου επαλήθευσης στην πρώτη κλάση ταξινομήθηκαν σωστά τα 2, τα 14 από τα 22 στην δεύτερη και 54 από τα 62 στην τρίτη κλάση.



Εικόνα 7.3-2: Confusion Matrix – KNN

7.4 Αποτελέσματα Δέντρων απόφασης

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκαν οι υπερπαράμετροι : max_depth=5, class_weight={0: 10, 1: 1}.

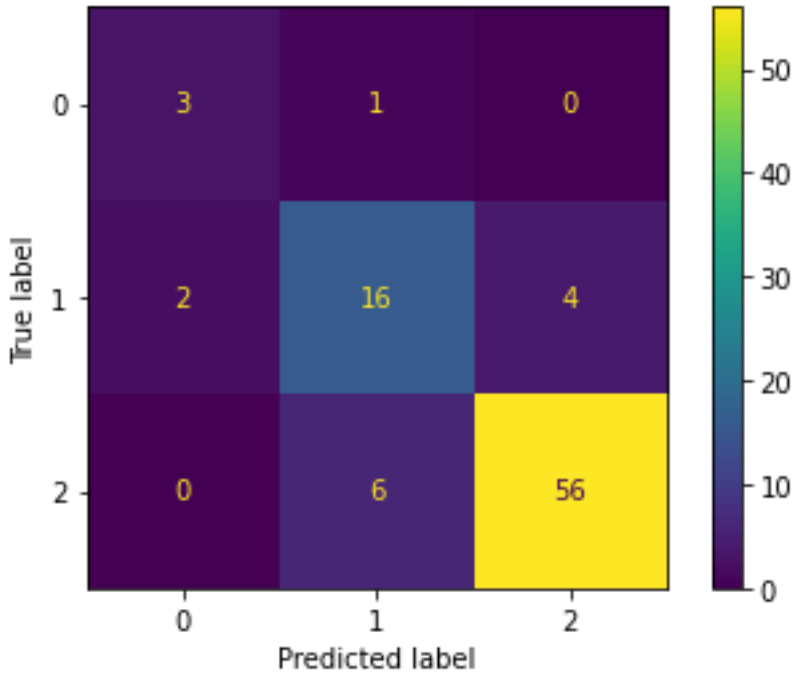
```
accuracy: 0.8522727272727273
roc_auc_score: 0.9615320929734805
cohen_kappa_score: 0.6729559748427674
Classification Report :
              precision    recall  f1-score   support

           0       0.60      0.75      0.67         4
           1       0.70      0.73      0.71        22
           2       0.93      0.90      0.92        62

    accuracy                           0.85        88
   macro avg       0.74      0.79      0.77        88
weighted avg       0.86      0.85      0.85        88
```

Εικόνα 7.4-1: Classification Report –Decision Trees

Το μοντέλο είναι σχετικά αξιόπιστο λόγω του Kappa = 0.67 και από το Confusion Matrix φαίνεται ότι πετυχαίνει μια συνολικά σχετικά καλή ακρίβεια με το macro-F-Measure να είναι 77%. Το F-Measure είναι ο αρμονικός μέσος όρος του Precision και του Recall. Για την τελευταία κλάση των δεδομένων βλέπουμε ότι προσεγγίζουμε ακρίβεια πάνω από 90% ενώ για τις 2 πρώτες από 60% για την πρώτη και στο 70% για την δεύτερη.



Εικόνα 7.4-2: Confusion Matrix – Decision Trees

Από τα 4 παραδείγματα του συνόλου επαλήθευσης στην πρώτη κλάση ταξινομήθηκαν σωστά τα 3, τα 16 από τα 22 στην δεύτερη και 56 από τα 62 στην τρίτη κλάση.

7.5 Αποτελέσματα Naïve Bayes

Για την ακρίβεια ο ταξινομητής είναι ο MultinomialNB, η built-in εκδοχή για multiclass προβλήματα της βιβλιοθήκης sklearn.

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκε η υπερπαράμετρος : alpha= 1e-05.

Όπως βλέπουμε και στο classification report που ακολουθεί η μέθοδος αυτή παρουσιάζει χαμηλή τιμή τόσο της μετρικής Kappa = 0.42 όσο και της macro-F-Measure που ισούται με 54%.

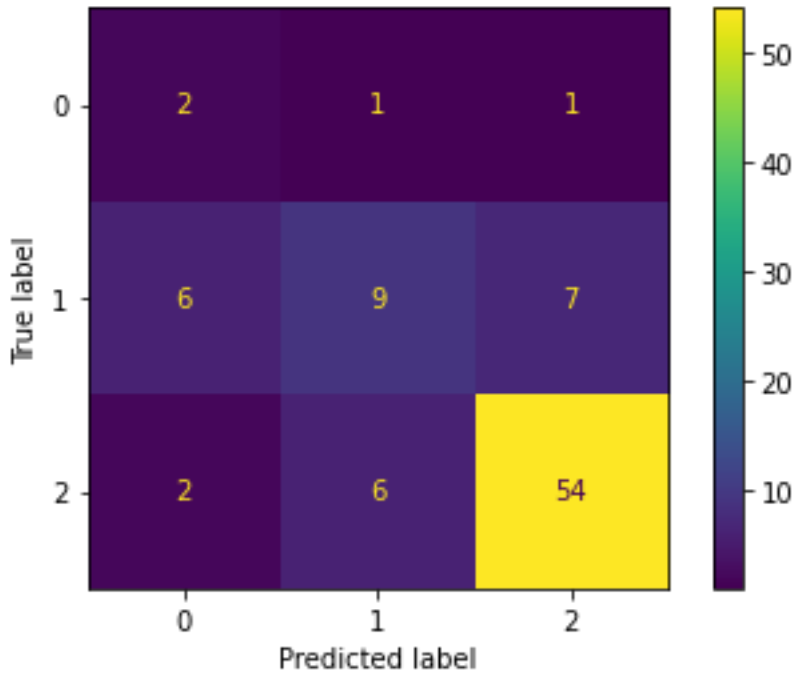
```
accuracy: 0.7386363636363636
roc_auc_score: 0.9209050518859131
cohen_kappa_score: 0.4230330672748005
Classification Report :
              precision    recall  f1-score   support

           0       0.20      0.50      0.29         4
           1       0.56      0.41      0.47        22
           2       0.87      0.87      0.87        62

    accuracy                           0.74        88
   macro avg       0.54      0.59      0.54        88
weighted avg       0.76      0.74      0.75        88
```

Εικόνα 7.5-1: Classification Report- Naïve Bayes

Αυτό υποδεικνύει ότι έχει πολύ χαμηλή αξιοπιστία και ακρίβεια αντίστοιχα, χάνοντας σημαντικό αριθμό instances. Ειδικά για την πρώτη κλάση που είναι και αυτή που μας ενδιαφέρει περισσότερο η ακρίβεια είναι στο 20% , στην δεύτερη στο 56% και στην τρίτη πολύ καλύτερη γύρω στο 90%.



Εικόνα 7.5-2: Confusion Matrix - Naïve Bayes

Από τα 4 παραδείγματα του συνόλου επαλήθευσης στην πρώτη κλάση ταξινομήθηκαν σωστά τα 2, τα 9 από τα 22 στην δεύτερη και 54 από τα 62 στην τρίτη κλάση.

7.6 Αποτελέσματα Random Forest

Από την εφαρμογή του αλγορίθμου Grid Search CV επιλέχθηκαν οι υπερπαράμετροι : criterion= entropy, max_depth= 10, n_estimators= 400.

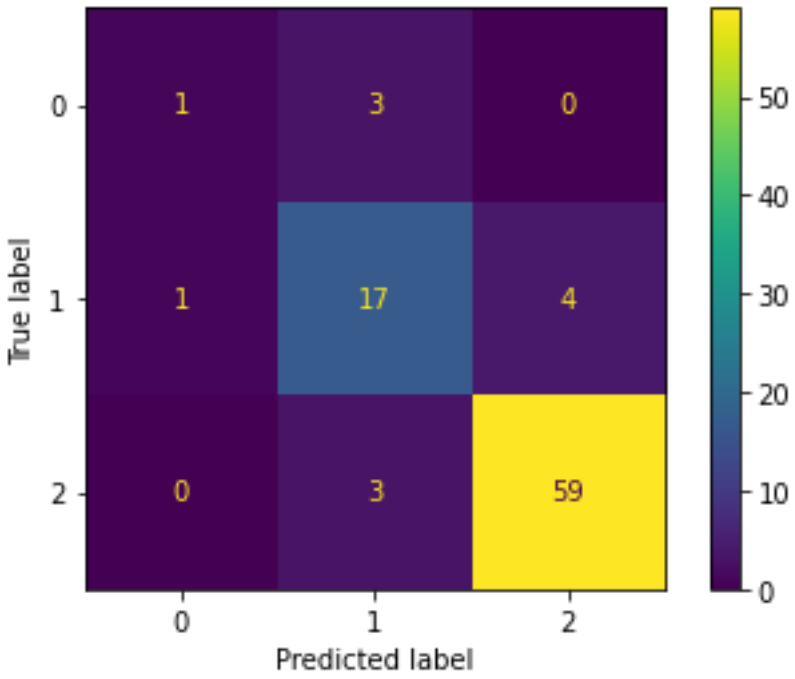
```
accuracy: 0.875
roc_auc_score: 0.9925431001746791
cohen_kappa_score: 0.7087845968712394
Classification Report :
              precision    recall  f1-score   support

           0       0.50      0.25      0.33         4
           1       0.74      0.77      0.76        22
           2       0.94      0.95      0.94        62

    accuracy                           0.88        88
   macro avg       0.73      0.66      0.68        88
weighted avg       0.87      0.88      0.87        88
```

Εικόνα 7.6-1: Classification Report – Random Forest

Το μοντέλο είναι σχετικά αξιόπιστο λόγω του Kappa = 0.7 και από το Confusion Matrix φαίνεται ότι πετυχαίνει συνολικά ακρίβεια με το macro-F-Measure να είναι ίσο 68%. Το F-Measure είναι ο αρμονικός μέσος όρος του Precision και του Recall. Για την τελευταία κλάση των δεδομένων βλέπουμε ότι προσεγγίζουμε ακρίβεια πάνω από 90% ενώ για τις 2 πρώτες από 50% για την πρώτη και πάνω από 70% για την δεύτερη. Εξαιτίας όμως του χαμηλού F-Measure για την πρώτη κλάση, από τα 4 παραδείγματα του συνόλου επαλήθευσης ταξινομήθηκε σωστά μόνο το 1, ενώ ταξινομήθηκαν τα 17 από τα 22 στην δεύτερη και 59 από τα 62 στην τρίτη κλάση.



Εικόνα 7.6-2 : Confusion Matrix – Random Forest

7.7 Αποτελέσματα Τεχνητών Νευρωνικών Δικτύων

Προκειμένου να εκπαιδευτεί το νευρωνικό δίκτυο με την μεγαλύτερη δυνατή ακρίβεια, αποφεύγοντας το overfitting, πραγματοποιήθηκαν αρκετές δοκιμές, όπως αυτές δίνονται στο παρακάτω πίνακα σχετικά με τις κατάλληλες τιμές για τις παραμέτρους learning rate και momentum ενώ από το συντονισμό με την μέθοδο HalvingGridSearchCV, όπως αναφέρθηκε παραπάνω επιλέχθηκε αριθμός epoch=100 και batch size=738 ίσο δηλαδή με το μήκος του συνόλου δεδομένων μετά τη εφαρμογή των μεθόδων train-test-split και SMOTE.

momemtum	Learning rate	Accuracy (%)	F-Measure	loss	Kappa
0.6	0.1	**88**	0.88	0.34	0.72
0.6	0.09	**88**	0.79	0.38	0.72
0.6	0.06	**86**	0.79	0.35	0.77
0.5	0.09	**83**	0.69	0.44	0.62
0.7	0.09	**90**	0.83	0.29	0.77
0.9	0.09	**90**	0.83	0.29	0.76
0.9	0.1	**90**	0.81	0.28	0.76

0.9	0.01	**77**	0.61	0.67	0.52

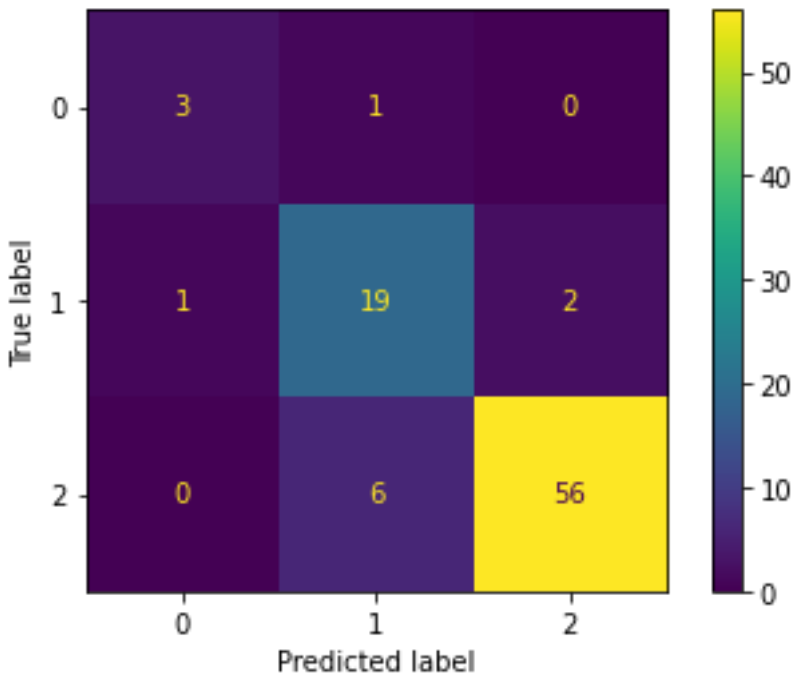
Πίνακας 7.7-1: Δοκιμές μοντέλου Neural Network

Ο συνδυασμός παραμέτρων που εκπαιδεύει καλύτερα το μοντέλο μας και παρουσιάζει την μικρότερη τιμή σε loss μαζί με την μεγαλύτερη τιμή της ακρίβειας είναι αυτός που επισημαίνεται (με κόκκινο χρώμα) στο παραπάνω πίνακα.

	precision	recall	f1-score	support
0	0.75	0.75	0.75	4
1	0.76	0.86	0.81	22
2	0.97	0.92	0.94	62
accuracy			0.90	88
macro avg	0.83	0.84	0.83	88
weighted avg	0.90	0.90	0.90	88

Εικόνα 7.7-1: Classification Report – Neural Network

Το μοντέλο είναι αξιόπιστο λόγω του μεγάλου Kappa = 0.78 και από το Confusion Matrix φαίνεται ότι πετυχαίνει μια σχετικά υψηλή ακρίβεια με το macro-F-Measure να είναι 83%. Για την τελευταία κλάση των δεδομένων βλέπουμε ότι προσεγγίζουμε ακρίβεια πάνω από 95% ενώ για τις 2 πρώτες από 75% για την πρώτη και στο 76% για την δεύτερη.



Εικόνα 7.7-2: Confusion Matrix – Neural Network

Από τα 4 παραδείγματα του συνόλου επαλήθευσης στην πρώτη κλάση ταξινομήθηκαν σωστά τα 3, τα 19 από τα 22 στην δεύτερη και 56 από τα 62 στην τρίτη κλάση.

7.8 Συνοπτικά αποτελέσματα

Στο παρακάτω πίνακα παρουσιάζονται τα συνοπτικά αποτελέσματα από τις εκτελέσεις των αλγορίθμων μετά την ρύθμιση των υπερπαραμέτρων τους

Classifier	Accuracy (%)	F-Measure	Kappa
Logistic Regression	**88**	0.81	0.72
SVC	**89**	0.74	0.74
KNN	**80**	0.65	0.54
Decision Trees	**85**	0.77	0.67
Naïve Bayes	**74**	0.54	0.42
Random Forest	**88**	0.68	0.70
ANN	**90**	0.83	0.78

Πίνακας 7.8-1 : Συγκριτικός πίνακας αποτελεσμάτων προβλεπτικών μοντέλων

Παρατηρούμε ότι από τις κλασσικές τεχνικές η λογιστική παλινδρόμηση είναι εκείνη η οποία παρουσιάζει αποτελέσματα εφάμιλλα με το νευρωνικό δίκτυο. Έχει μικρή διαφορά στο σκορ της μετρικής F-measure, της μετρικής Kappa και της ακρίβειας. Το μοντέλο με την χειρότερη απόδοση είναι Naïve Bayes με F-measure=0.54 και Kappa = 0.42.

8 Συμπεράσματα και διερεύνηση

Για την εργασία αυτή, κατασκευάστηκαν τα μοντέλα επιβλεπόμενης μηχανικής μάθησης Naïve Bayes (MultinomialNB), Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest όπως και ένα μοντέλο τεχνητού νευρωνικού δικτύου (ANN) με σκοπό την πρόβλεψη της νικητήριας ομάδας στην Αγγλική Premiere League. Η ανάπτυξη έγινε εξολοκλήρου σε γλώσσα προγραμματισμού Python. Χρησιμοποιήθηκαν δεδομένα από 22 σεζόν της English Premier League τα οποία αποκτήθηκαν με την μέθοδο web scrapping από την ιστοσελίδα transermarkt.

Από τα προβλεπτικά μοντέλα που αναπτύχθηκαν τόσο αυτό του νευρωνικού δικτύου όσο και του Logistic Regression αποδείχθηκαν αξιόπιστα και με καλή ακρίβεια και ικανότητα διαχωρισμού των κλάσεων. Εξαιτίας της τόσο κοντινής απόδοσης τους θα ήταν θεμιτό να αξιολογηθούν τα μοντέλα και σε διαφορετικά πρωταθλήματα και σε όσο το δυνατόν μεγαλύτερα σύνολα δεδομένων ώστε να αξιολογηθεί συγκριτικά η απόδοση τους. Γενικά τα μοντέλα της λογιστικής παλινδρόμησης είναι λιγότερα επιρρεπή σε overfitting λόγω του ότι περιέχουν απλούστερες σχέσεις μεταξύ της ετικέτας και των μεταβλητών σε αντίθεση με το νευρωνικό που έχει περισσότερο περίπλοκη δομή και στηρίζεται πολύ στα δεδομένα του συνόλου εκπαίδευσης . Σε μικρά σύνολα δεδομένων όμως, μπορεί να μην αποδίδει το ίδιο καλά με το νευρωνικό δίκτυο.

Η παρούσα μελέτη διεξήχθη στηριζόμενη σε στοιχεία τα οποία αφορούν μόνο στα στατιστικά των ομάδων. Μια ενδιαφέρουσα επέκταση θα ήταν να συμπεριληφθούν και στατιστικά στοιχεία παικτών, συμπεριλαμβανομένου και εκείνων που αφορούν σε τραυματισμούς, ή επαναλαμβανόμενη κούραση, ακόμα και τις αποδόσεις από στοιχηματικούς παράγοντες, αναρτήσεις σε social media που αφορούν σε όλα τα παραπάνω (ομάδες, παίκτες κτλ) ακόμα και στις καιρικές συνθήκες τις ημέρες των αγώνων, δημιουργώντας έτσι ένα υψηλότερης πολυπλοκότητας τεχνητό νευρωνικό δίκτυο, με περισσότερη πληροφορία.

9 ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] K. M. Langaroudi και R. M. Yamaghani, «Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey,» Lahijan Branch,Iran, 2019.

[2] V. C. P. a. C. Tjortjis, «Sports Analytics for Football League Table and Player Performance Prediction,» σε *2020 11th International Conference on Information*, Piraeus, Greece, 2020.

[3] M. A. S. Arabzad, T. . M. Araghi, S. S. Nezhad και N. Ghofrani, «Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League,» *Applied Research on Industrial Engineering,* τόμ. 1, αρ. 3, pp. 159-179, 3 10 2014.

[4] K. Sujatha, T. Godhavari και N. P. Bhavani , «Football Match Statistics Prediction using Artificial Neural Networks,» *International Journal of Mathematical and Computational Methods ,* τόμ. 3, pp. 1-8, 2018.

[5] A. Mccabe και J. Trevathan, «Artificial Intelligence in Sports Prediction,» σε *Fifth International Conference on Information Technology: New Generations*, 2008.

[6] B. Aslan και M. Inceoglu, «A Comparative Study on Neural Network Based Soccer Result Prediction,» σε *Seventh International Conference on Intelligent Systems Design and Applications*, 2007.

[7] K. Huang και W. Chang, «A neural network method for prediction of 2006 World Cup Football Game,» σε *The 2010 International Joint Conference on Neural Networks*, 2010.

[8] N. Tax και Y. P. Joustra, «Predicting the Dutch football competition using public data: A machine learning approach,» τόμ. 10, αρ. 10, p. 1– 13, 2015.

[9] Van Haaren, J. and Davis, J, «Predicting the Final League Tables of,» σε *5th int'l conf. mathematics in sport*, 2015.

[10] J. Oberstone, «Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success,» *Journal of Quantitative Analysis in Sports,* τόμ. 5, 2009.

[11] M. S. &. R. R. Neethu, «Sentiment analysis in twitter using machine learning techniques,» σε *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).*, 2013.

[12] P. Nadkarni, Clinical Research Computing-A Practitioner's Handbook, Academic Press, 2016.

[13] J. Quinlan, «Induction of Decision Trees,» *Machine Learning,* τόμ. 1, pp. 81-106, 1986.

[14] L. Breiman, «Bagging predictors,» *Machine Learning 24,* p. 123–140, 1996.

[15] L. Breiman, «Random Forests,» *Machine Learning 45,* p. 5–32.

[16] J. &. L. S. &. V. S. G. Beutel, «Does Machine Learning Help us Predict Banking Crises?,» *Journal of Financial Stability,* τόμ. 45, p. 100693, 2019.