ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΕΚΒΑΣΗΣ ΑΓΩΝΩΝ ΠΟΔΟΣΦΑΙΡΟΥ ΜΕ ΑΛΓΟΡΙΘΜΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΕΝΡΙΚ ΜΑΖΑΙ
Α.Μ. 18390080

Επιβλέπουσα Καθηγήτρια
Παναγιώτα Τσελέντη, ΕΔΙΠ

Αθήνα - Αιγάλεω, Σεπτέμβριος 2024

Διπλωματική Εργασία

Πρόβλεψη Έκβασης Αγώνων Ποδοσφαίρου με Αλγορίθμους Μηχανικής Μάθησης

Prediction of Football Matches Outcomes with Machine Learning Algorithms

Ενρίκ Μαζάι

A.M. 18390080

Επιβλέπουσα Καθηγήτρια: Παναγιώτα Τσελέντη, ΕΔΙΠ

Εγκρίθηκε από την κάτωθι τριμελή εξεταστική επιτροπή:

ΤΣΕΛΕΝΤΗ ΠΑΝΑΓΙΩΤΑ	ΕΔΙΠ	Panagiota Tselenti Digitally signed by Panagiota Tselenti Date: 2024.10.08 12:46:59 +03'00'
ΤΡΟΥΣΣΑΣ ΧΡΗΣΤΟΣ	ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ	Christos Troussas Digitally signed by Christos Troussas Date: 2024.10.08 13:00:30 +03'00'
ΑΚΡΙΒΗ ΚΡΟΥΣΚΑ	ΕΔΙΠ	Akrivi Krouska Digitally signed by Akrivi Krouska Date: 2024.10.08 13:15:58 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Ενρίκ Μαζάι του Αγκρόν, με αριθμό μητρώου 18390080 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι: είμαι συγγραφέας αυτής της Διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου.

Ο Δηλών

Ευχαριστίες

Με την εκπλήρωση αυτής της διπλωματικής εργασίας θα ήθελα να ευχαριστήσω ιδιαίτερα την καθηγήτρια μου και επιβλέπουσα κυρία Παναγιώτα Τσελέντη που με βοήθησε με την καθοδήγηση και τις εποικοδομητικές παρατηρήσεις της.

Ακόμα θα ήθελα να ευχαριστήσω τους συναδέλφους του τμήματος μου για την συμπαράσταση και την βοήθεια τους κατά την διάρκεια όλων των χρόνων της σχολής.

Τέλος, ένα ιδιαίτερο ευχαριστώ οφείλω στην οικογένειά μου και στους φίλους μου, που με την αμέριστη στήριξη και την πίστη τους σε μένα, μου έδωσαν τη δύναμη να φέρω εις πέρας αυτή την προσπάθεια.

Περίληψη

Η παρούσα διπλωματική εξετάζει τη χρήση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων, με στόχο την αξιολόγηση διάφορων μοντέλων που μπορούν να βελτιώσουν την ακρίβεια των προβλέψεων. Στην αρχή, διεξήχθη αναλυτική βιβλιογραφική ανασκόπηση, κατά την οποία αναλύθηκαν οι προσεγγίσεις και οι τεχνικές από προηγούμενες έρευνες, εντοπίζοντας αδυναμίες και περιορισμούς στις μεθόδους πρόβλεψης. Στη συνέχεια, παρουσιάστηκε το θεωρητικό υπόβαθρο των αλγορίθμων μηχανικής μάθησης, όπως Logistic Regression, Random Forest, Gradient Boosting, k-Nearest Neighbors και Multi-Layer Perceptron, εξετάζοντας πώς αυτές οι τεχνικές μπορούν να βελτιστοποιηθούν για πιο ακριβείς προβλέψεις. Η εργασία εστιάζει κυρίως στην προσέγγιση με ταξινόμηση δύο κλάσεων (binary classification) για την πρόβλεψη των αποτελεσμάτων. Ακολούθησε μια ανάλυση των βημάτων δημιουργίας μοντέλων μηχανικής μάθησης, από τη συλλογή των δεδομένων μέχρι τη βελτιστοποίησή τους. Στο πειραματικό μέρος, χρησιμοποιήθηκαν πραγματικά δεδομένα από ποδοσφαιρικούς αγώνες και εφαρμόστηκαν τεχνικές feature engineering για τη βελτίωση της απόδοσης των αλγορίθμων. Μετά από δύο φάσεις εκτέλεσης και βελτιστοποίησης των αλγορίθμων, τα αποτελέσματα έδειξαν βελτιώσεις στην ακρίβεια των προβλέψεων, ενώ πραγματοποιήθηκε επίσης ανάλυση για την ανίχνευση φαινομένων overfitting και underfitting. Η εργασία εξετάζει και την ταξινόμηση πολλαπλών κλάσεων (Multiclass Classification), υπογραμμίζοντας τις προκλήσεις και τις δυνατότητές της. Τέλος, η μελέτη συγκρίνει τα αποτελέσματα των μοντέλων σε όρους ακρίβειας και απόδοσης, αναδεικνύοντας τις διαφορές ανάμεσα στις διάφορες προσεγγίσεις και αναγνωρίζει τους περιορισμούς που υπήρξαν . Συνολικά, η εργασία συμβάλλει στην κατανόηση και βελτίωση των υπαρχουσών μεθόδων πρόβλεψης ποδοσφαιρικών αγώνων, προσφέροντας πιο αξιόπιστα μοντέλα που μπορούν να εφαρμοστούν για τη λήψη αποφάσεων σε επαγγελματικό επίπεδο ή την ανάπτυξη στρατηγικών σε στοιχηματικές και αθλητικές αναλύσεις.

Λέξεις-κλειδιά

Αλγόριθμοι Μηχανική Μάθηση, Ταξινόμηση, Δυαδική Ταξινόμηση, Λογιστική Παλινδρόμηση, Random Forest, Gradient Boosting, K-Nearest Neighbors, Multi-Layer Perceptron, Feature Engineering

Abstract

This thesis examines the use of machine learning algorithms for predicting the outcomes of football matches, aiming to evaluate various models that can improve the accuracy of predictions. Initially, a comprehensive literature review was conducted, analyzing the approaches and techniques from previous research, identifying weaknesses and limitations in the prediction methods. Subsequently, the theoretical background of machine learning algorithms was presented, such as Logistic Regression, Random Forest, Gradient Boosting, k-Nearest Neighbors, and Multi-Layer Perceptron, exploring how these techniques can be optimized for more accurate predictions. The work primarily focuses on the binary classification approach for predicting outcomes. An analysis of the steps involved in creating machine learning models was then performed, from data collection to optimization. In the experimental section, real data from football matches was used, and feature engineering techniques were applied to improve the performance of the algorithms. After two phases of execution and optimization of the algorithms, the results showed improvements in prediction accuracy, and an analysis was also conducted for detecting overfitting and underfitting phenomena. The study examines also multiclass classification, highlighting its challenges and potential. Finally, the research compares the results of the models in terms of accuracy and performance, highlighting the differences between various approaches and recognizing the limitations that existed. Overall, the work contributes to understanding and improving existing methods for predicting football match outcomes, providing more reliable models that can be applied for decision-making at a professional level or for developing strategies in betting and sports analytics.

Keywords

Machine Learning Algorithms, Classification, Binary Classification, Logistic Regression, Random Forest, Gradient Boosting, K-Nearest Neighbors, Multi-Layer Perceptron, Feature Engineering

ΠΙΝΑΚΕΣ

ΕΙΚΟΝΕΣ

Περιεχόμενα

1. Εισαγωγή

Στη σύγχρονη εποχή, η χρήση της τεχνητής νοημοσύνης (ΤΝ) επεκτείνεται συνεχώς σε ολοένα και περισσότερους τομείς της ζωής μας, και ο χώρος του αθλητισμού δεν αποτελεί εξαίρεση. Το ποδόσφαιρο, αποτελεί το πιο δημοφιλές άθλημα στον κόσμο, συγκεντρώνει περίπου 5 δισεκατομμύρια φιλάθλους παγκοσμίως με βάση στατιστικά της FIFA (International Federation of Association Football) και έχει τεράστια οικονομική επίδραση στις σημερινές κοινωνίες. Στην Αγγλία το ποδόσφαιρο προσφέρει 94.000 χιλιάδες θέσεις εργασίας ενώ αποφέρει 4.6 δις εκατομμύρια ευρώ στο κρατικό προϋπολογισμό. (Premier League, 2024). Ενώ το γεγονός ότι περίπου 1,5 δισεκατομμύρια άνθρωποι παρακολούθησαν τον τελικό του Παγκοσμίου Κυπέλλου το 2022 κάνει σαφές την έκταση και τη σημασία του αθλήματος παγκοσμίως. (FIFA, 2023)

Τα τελευταία χρόνια, η ενσωμάτωση της τεχνολογίας στον χώρο του ποδοσφαίρου είναι εντυπωσιακή, με τη συλλογή μεγάλου όγκου στατιστικών δεδομένων και την οπτικοποίησή τους. Αυτή η αύξηση των δεδομένων έχει παρακινήσει τις ομάδες να προσλάβουν αναλυτές διαφόρων ειδικοτήτων, οι οποίοι χρησιμοποιούν αυτά τα δεδομένα για να παρέχουν χρήσιμες πληροφορίες, όπως προτεινόμενα συστήματα αντιμετώπισης αντιπάλων, αυτοματοποιημένο σύστημα σκάουτινγκ παικτών, συλλογή στατιστικών στοιχείων κ.α.

Λόγω του τεράστιου όγκου των διαθέσιμων δεδομένων και με την πρόοδο της τεχνητής νοημοσύνης οι δυνατότητες χρήση της προς όφελος των ομάδων αυξάνεται. Μια νέα χρήση των δεδομένων είναι η αξιοποιήση τους για δημιουργία μοντέλων πρόβλεψης των ποδοσφαιρικών αγώνων. Υπάρχει μια τάση τα τελευταία χρόνια για την ανάπτυξη τέτοιων μοντέλων, με πολλούς ερευνητές να προσπαθούν να αποκρυπτογραφήσουν τη σύνθετη φύση του αθλήματος για να επιτύχουν ακριβείς προβλέψεις. Αυτές οι προβλέψεις μπορούν να χρησιμοποιηθούν για πολλούς σκοπούς από τα τεχνικά μέλη των ομάδων για την ανάπτυξη στρατηγικών, από τους απλούς χρήστες για οικονομικό κέρδος με το ποντάρισμα επίλογων μέσω των στοιχηματικών εταιριών, αλλά και για προσωπική χρήση από τους φιλάθλους που απολαμβάνουν το άθλημα.

Καθώς η συλλογή και ανάλυση δεδομένων στο ποδόσφαιρο συνεχίζει να εξελίσσεται, η τεχνητή νοημοσύνη αναδεικνύεται ως ένας σημαντικός σύμμαχος στη διαδικασία λήψης αποφάσεων και βελτιστοποίησης στρατηγικών. Η ΤΝ, με την ικανότητά της να επεξεργάζεται μεγάλα δεδομένα και να αναγνωρίζει μοτίβα που θα ήταν αδύνατον να εντοπιστούν από τον άνθρωπο, προσφέρει νέες δυνατότητες σε πολλούς τομείς του αθλήματος.

Η επόμενη ενότητα θα εξετάσει τη συμβολή της τεχνητής νοημοσύνης και θα θέσει τις βάσεις για την κατανόηση των τεχνικών και μεθοδολογιών που εφαρμόζονται στην πρόβλεψη ποδοσφαιρικών αγώνων.

1.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (ΤΝ) αποτελεί έναν από τους πιο συναρπαστικούς και καινοτόμους τομείς της σύγχρονης τεχνολογίας. Αναφέρεται στην ανάπτυξη υπολογιστικών συστημάτων που μπορούν να εκτελούν εργασίες οι οποίες συνήθως απαιτούν ανθρώπινη νοημοσύνη. Αυτές οι εργασίες περιλαμβάνουν την αναγνώριση προτύπων, την κατανόηση φυσικής γλώσσας, τη λήψη αποφάσεων και την επίλυση προβλημάτων. Η ΤΝ συνδυάζει διάφορους τομείς της επιστήμης των υπολογιστών, όπως η μηχανική μάθηση, η ανάλυση δεδομένων και η ρομποτική, για να δημιουργήσει έξυπνα συστήματα.

Η πρόοδος στην ΤΝ οφείλεται κυρίως στην αύξηση της υπολογιστικής ισχύος και στην προσβασιμότητα μεγάλων ποσοτήτων δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης, ειδικότερα, έχουν επιτρέψει στους ερευνητές να αναπτύξουν μοντέλα που μπορούν να μάθουν από τα δεδομένα και να βελτιώνονται με την πάροδο του χρόνου. Αυτά τα μοντέλα χρησιμοποιούνται ευρέως σε πολλούς τομείς, όπως η υγειονομική περίθαλψη, η χρηματοοικονομική ανάλυση, η ανίχνευση απάτης και ο αθλητισμός.

Η εφαρμογή της ΤΝ στον αθλητισμό, και ειδικά στην πρόβλεψη αποτελεσμάτων αγώνων, έχει προσελκύσει το ενδιαφέρον τόσο των ερευνητών όσο και των επαγγελματιών του χώρου. Τα μοντέλα πρόβλεψης που χρησιμοποιούν την ΤΝ μπορούν να αναλύουν μεγάλους όγκους δεδομένων, όπως τα στατιστικά των παικτών, τις αγωνιστικές τους επιδόσεις, εξωτερικούς παράγοντες όπως οι καιρικές συνθήκες, αλλά και άλλους παράγοντες που μπορεί να επηρεάσουν το αποτέλεσμα ενός αγώνα. Μέσω της ανάλυσης αυτών των δεδομένων, τα μοντέλα μπορούν να προβλέπουν με μεγαλύτερη ακρίβεια τα πιθανά αποτελέσματα, βοηθώντας έτσι τις ομάδες και τους αναλυτές να λαμβάνουν πιο ενημερωμένες αποφάσεις.

Η εξέλιξη της ΤΝ συνεχίζεται με ταχείς ρυθμούς, και οι δυνατότητές της για καινοτομία και βελτίωση είναι απεριόριστες. Ως εκ τούτου, η κατανόηση και η αξιοποίηση της ΤΝ και των τεχνικών μηχανικής μάθησης είναι κρίσιμης σημασίας για την αντιμετώπιση των προκλήσεων του μέλλοντος.

1.2 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας υποτομέας της τεχνητής νοημοσύνης που επικεντρώνεται στην ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα. Αντί να ακολουθούν προκαθορισμένες εντολές, τα συστήματα μηχανικής μάθησης αναγνωρίζουν μοτίβα και τάσεις στα δεδομένα και χρησιμοποιούν αυτή τη γνώση για να λαμβάνουν αποφάσεις ή να κάνουν προβλέψεις.

Υπάρχουν διάφοροι τύποι μηχανικής μάθησης, όπως η επιβλεπόμενη, η μη επιβλεπόμενη και η ενισχυτική μάθηση. Στην επιβλεπόμενη μάθηση, τα μοντέλα εκπαιδεύονται με δεδομένα που περιέχουν ετικέτες, δηλαδή γνωστές απαντήσεις. Αντίθετα, στη μη επιβλεπόμενη μάθηση, τα δεδομένα δεν έχουν ετικέτες, και το μοντέλο πρέπει να βρει μόνο του τα μοτίβα. Η ενισχυτική μάθηση επικεντρώνεται στη λήψη αποφάσεων και στη βελτιστοποίηση της συμπεριφοράς μέσω αλληλεπίδρασης με ένα δυναμικό περιβάλλον.

Η μηχανική μάθηση έχει πολλές πρακτικές εφαρμογές, από τη βελτίωση των αποτελεσμάτων αναζητήσεων στο διαδίκτυο και την εξατομίκευση προτάσεων προϊόντων, μέχρι την πρόγνωση καιρού και την ανίχνευση απάτης σε χρηματοοικονομικές συναλλαγές. Στον αθλητισμό, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την ανάλυση της απόδοσης των αθλητών, την πρόβλεψη των αποτελεσμάτων αγώνων και τη βελτίωση των στρατηγικών παιχνιδιού.

Η συνεχής ανάπτυξη και βελτίωση των αλγορίθμων μηχανικής μάθησης καθιστά αυτή την τεχνολογία κρίσιμη για την επίλυση σύνθετων προβλημάτων και την προώθηση της καινοτομίας σε διάφορους τομείς της καθημερινής ζωής.

1.3 Κατηγορίες Αλγορίθμων Μηχανικής Μάθησης

Επιβλεπόμενη Μάθηση (Supervised Learning):

Στην επιβλεπόμενη μάθηση, οι αλγόριθμοι κατηγοριοποιούνται κυρίως σε δύο κατηγορίες: ταξινόμηση (classification) και παλινδρόμηση (regression). Η ταξινόμηση χρησιμοποιείται όταν οι ετικέτες είναι κατηγορίες, ενώ η παλινδρόμηση χρησιμοποιείται όταν οι ετικέτες είναι συνεχείς τιμές. Στο πειραματικό μέρος της διπλωματικής μας εργασίας, θα χρησιμοποιηθούν διάφοροι αλγόριθμοι ταξινόμησης, όπως οι Random Forest Classifier, Gradient Boosting, Logistic Regression και MLP κ.α.

Η επιβλεπόμενη μάθηση θα εφαρμοστεί για την πρόβλεψη των αποτελεσμάτων αθλητικών αγώνων, χρησιμοποιώντας ένα σύνολο δεδομένων που περιλαμβάνει διάφορες μεταβλητές κυρίως αριθμητικές που έχουν να κάνουν με στατιστικές επιδόσεις των ομάδων σε προηγούμενους αγώνες καθώς και τις αντίστοιχες ετικέτες που αντιπροσωπεύουν το αποτέλεσμα του αγώνα (νίκη γηπεδούχου ή φιλοξενούμενου). Οι αλγόριθμοι θα εκπαιδευτούν με αυτό το σύνολο δεδομένων και θα αξιολογηθούν με βάση την ικανότητά τους να προβλέπουν σωστά τα αποτελέσματα σε νέα δεδομένα.

Η συγκριτική ανάλυση των αποτελεσμάτων των διαφόρων αλγορίθμων θα μας επιτρέψει να κατανοήσουμε ποιος αλγόριθμος αποδίδει καλύτερα στις συγκεκριμένες συνθήκες και δεδομένα του προβλήματός μας. Αυτό θα συμβάλει στην επιλογή του βέλτιστου μοντέλου για την πρόβλεψη των αποτελεσμάτων αγώνων και θα ενισχύσει τη δυνατότητα λήψης ακριβών και αξιόπιστων προβλέψεων.

Η επιβλεπόμενη μάθηση είναι η κύρια κατηγόρια αλγορίθμων που θα χρησιμοποιηθούν στην έρευνα ,όμως υπάρχουν και άλλες σημαντικές κατηγορίες.

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning):

Σε αυτή την κατηγορία, τα δεδομένα δεν συνοδεύονται από ετικέτες εξόδου. Οι αλγόριθμοι προσπαθούν να εντοπίσουν κρυφά μοτίβα ή δομές στα δεδομένα. Τυπικές μέθοδοι περιλαμβάνουν τον διαχωρισμό σε ομάδες (clustering) και την ανάλυση κύριων συνιστωσών (PCA).

Ημιαυτόματη Μάθηση (Semi-Supervised Learning):

Αυτή η κατηγορία βρίσκεται μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης. Οι αλγόριθμοι εκπαιδεύονται χρησιμοποιώντας ένα μικρό σύνολο επισημασμένων δεδομένων και ένα μεγαλύτερο σύνολο μη επισημασμένων δεδομένων.

1.4 Εφαρμογές της Τεχνητής Νοημοσύνης στο Ποδόσφαιρο

Η ΤΝ και η μηχανική μάθηση έχουν βρει εφαρμογή σε διάφορες πτυχές του ποδοσφαίρου, προσφέροντας λύσεις που στοχεύουν στην βελτίωση των επιδόσεων των επαγγελματιών και συνολικότερα στην ποιότητα του αθλήματος (Ahmed, 2023). Συγκεκριμένα κάποιες εφαρμογές της ΤΝ στο ποδόσφαιρό είναι:

1. **Ανάλυση Απόδοσης Παικτών και Ομάδων:** Μέσω της ανάλυσης δεδομένων από αγώνες και προπονήσεις, η ΤΝ επιτρέπει στους προπονητές να εξετάσουν λεπτομερώς την απόδοση των παικτών, εντοπίζοντας τόσο σημεία βελτίωσης όσο και αδυναμίες των αντιπάλων. (García-Aliaga, 2021)
2. **Πρόβλεψη Αποτελεσμάτων:** Τα μοντέλα μηχανικής μάθησης χρησιμοποιούν ιστορικά δεδομένα, πληροφορίες για τραυματισμούς και καιρικές συνθήκες για να προβλέψουν την έκβαση των αγώνων, παρέχοντας κρίσιμες πληροφορίες σε προπονητές και στοιχηματικές εταιρείες. Με αυτή την κατηγορία θα ασχοληθεί και η παρούσα διπλωματική.
3. **Ανάλυση Βίντεο και Τεχνικών Δεδομένων:** Η ΤΝ αναλύει βίντεο αγώνων και κατηγοριοποιεί αυτόματα φάσεις του παιχνιδιού, επιτρέποντας στους αναλυτές να επικεντρωθούν στις κρίσιμες στιγμές που μπορούν να καθορίσουν το αποτέλεσμα ενός αγώνα. (J. Xing, 2011)
4. **Διαχείριση Κούρασης και Πρόληψη Τραυματισμών:** Αξιοποιώντας δεδομένα από αισθητήρες και συστήματα ανάλυσης, οι ομάδες μπορούν να προσαρμόζουν τις προπονήσεις και να προβλέπουν πότε οι παίκτες είναι πιο ευάλωτοι σε τραυματισμούς. (Van Eetvelde, 2021)
5. **Αναγνώριση Ταλέντων:** Η ΤΝ βοηθάει στον εντοπισμό νέων ταλέντων μέσα από την ανάλυση μεγάλου όγκου δεδομένων, επιτρέποντας στις ομάδες να ανακαλύψουν και να αναπτύξουν νεαρούς παίκτες με υψηλό δυναμικό. (Jauhiainen S., 2019)
6. **Βελτιστοποίηση Στρατηγικής Παιχνιδιού:** Μέσω ανάλυσης σε πραγματικό χρόνο, η ΤΝ παρέχει στρατηγικές προτάσεις, βοηθώντας τους προπονητές να λαμβάνουν αποφάσεις με βάση την εξέλιξη του αγώνα και τα δεδομένα που συλλέγονται κατά τη διάρκεια του παιχνιδιού. (Wang, 2024)

Αυτές οι εφαρμογές καταδεικνύουν τη δύναμη της ΤΝ στη μετατροπή του ποδοσφαίρου σε ένα πιο επιστημονικό άθλημα, όπου οι αποφάσεις λαμβάνονται βάσει στατιστικών αναλύσεων.

1.5 Σκοπός και Στόχοι της Εργασίας

Σκοπός

Σκοπός της διπλωματικής εργασίας είναι η δημιουργία ενός μοντέλου πρόβλεψης ποδοσφαιρικών αγώνων χρησιμοποιώντας διάφορους αλγορίθμους μηχανικής μάθησης, βρίσκοντας την βέλτιστή επιλογή έπειτα από σύγκριση των αποτελεσμάτων. Η εργασία στοχεύει στην αξιοποίηση ενός εκτεταμένου και ποικιλόμορφου συνόλου δεδομένων, ώστε να παραχθούν πιο αξιόπιστα και γενικεύσιμα αποτελέσματα, ξεπερνώντας τα όρια προηγούμενων ερευνών .

Στόχοι

Ο στόχος της διπλωματικής μας εργασίας είναι να αξιοποιήσουμε ένα επαρκές και αντιπροσωπευτικό σύνολο δεδομένων, καλύπτοντας ποικιλία χρονικών περιόδων και πρωταθλημάτων. Στοχεύουμε να εξερευνήσουμε τις δυνατότητες που προσφέρουν οι αλγόριθμοι μηχανικής μάθησης στην πρόβλεψη ποδοσφαιρικών αποτελεσμάτων, αναδεικνύοντας τη σημασία της ανάλυσης δεδομένων σε αυτό το πεδίο.

Θα εφαρμόσουμε τόσο τους πιο διαδεδομένους αλγορίθμους όσο και εκείνους που θεωρούνται πιο ιδιαίτερα υποσχόμενοι, ενώ θα ενσωματώσουμε και νέα χαρακτηριστικά που δεν έχουν χρησιμοποιηθεί σε προηγούμενες έρευνες. Με αυτόν τον τρόπο, φιλοδοξούμε να δημιουργήσουμε ένα πιο αξιόπιστο και γενικεύσιμο μοντέλο πρόβλεψης για ποδοσφαιρικούς αγώνες, το οποίο θα μπορεί να εφαρμοστεί ευρέως και με ασφάλεια στο μέλλον.

Διάρθρωση της αναφοράς

Η εργασία που ακολουθεί επικεντρώνεται στην δημιουργία ενός μοντέλου πρόβλεψης της έκβασης ποδοσφαιρικών αγώνων.

- **Κεφάλαιο 1:** Εισαγωγή

  Σε αυτό το κεφάλαιο, θα παρουσιάσαμε το πλαίσιο και τη σημασία της έρευνας. Αναλύσαμε την εφαρμογή της μηχανικής μάθησης στον τομέα των αθλητικών προβλέψεων, θέτοντας τις βάσεις για την περαιτέρω μελέτη και την ανάπτυξη του μοντέλου μας.
- **Κεφάλαιο 2:** Βιβλιογραφική Έρευνα

  Στο κεφάλαιο αυτό θα εξετάσουμε την υπάρχουσα βιβλιογραφία σχετικά με την πρόβλεψη ποδοσφαιρικών αγώνων, εντοπίζοντας παλαιότερες μελέτες, μεθόδους και τεχνικές που έχουν χρησιμοποιηθεί. Η έρευνα αυτή θα μας βοηθήσει να κατανοήσουμε τα πλεονεκτήματα και τα μειονεκτήματα των προηγούμενων προσεγγίσεων.
- **Κεφάλαιο 3:** Μεθοδολογία

  Στην ενότητα αυτή, θα περιγράψουμε αναλυτικά τα βήματα που θα ακολουθήσουμε για την ανάπτυξη του μοντέλου πρόβλεψης. Θα παρουσιαστούν οι τεχνικές προ-επεξεργασίας δεδομένων, οι αλγόριθμοι που θα χρησιμοποιηθούν, καθώς και οι μέθοδοι αξιολόγησης του μοντέλου μας.
- **Κεφάλαιο 4:** Πειραματικό Μέρος

  Σε αυτό το κεφάλαιο, θα υλοποιήσουμε το μοντέλο που σχεδιάσαμε, εφαρμόζοντας τα δεδομένα και τις μεθόδους που έχουμε επιλέξει. Θα αναλύσουμε τα αποτελέσματα των πειραμάτων και θα συγκρίνουμε τις επιδόσεις των διαφορετικών αλγορίθμων.
- **Κεφάλαιο 5:** Αποτελέσματα

  Σε αυτή την ενότητα, θα παρουσιάσουμε και θα αναλύσουμε τα αποτελέσματα που προκύπτουν από την εφαρμογή του μοντέλου. Θα συγκρίνουμε την ακρίβεια, την αποτελεσματικότητα και την αξιοπιστία των ευρημάτων για κάθε αλγόριθμο αλλά και για τις διάφορες προσεγγίσεις που ακολουθήσαμε.
- **Κεφάλαιο 6:** Συμπεράσματα

  Στο τελευταίο κεφάλαιο, θα συνοψίσουμε τα ευρήματα της έρευνάς μας, αξιολογώντας την αποτελεσματικότητα του μοντέλου πρόβλεψης και τονίζοντας τους περιορισμούς μας. Θα συζητήσουμε τις προοπτικές για μελλοντική έρευνα και βελτιώσεις, καθώς και τη σημασία των αποτελεσμάτων μας στον τομέα των αθλητικών προβλέψεων.

2. Βιβλιογραφική Έρευνα

2.1 Έρευνες Προηγούμενων ετών

Η συνεχής άνοδος της τεχνητής νοημοσύνης και των αλγορίθμων μηχανικής μάθησης έχουν οδηγήσει σε αρκετές έρευνες πάνω στο θέμα της πρόβλεψης αγώνων ποδοσφαίρου. Οι έρευνες αυτές ποικίλουν λόγω των διαφορετικών χαρακτηριστικών που χρησιμοποιούνται στα μοντέλα, τους αλγορίθμους αλλά και της προσέγγισης που ακολουθούν.

Αρκετές έρευνες προσεγγίζουν το πρόβλημα χωρίζοντας την κατηγοριοποίηση σε 3 κλάσεις που είναι μία για κάθε πιθανό αποτέλεσμα: νίκη γηπεδούχου, ισοπαλία και νίκη φιλοξενούμενου. Αυτή την προσέγγιση παρατηρούμε στην έρευνα (Berrar, 2019) όπου με την χρήση του αλγορίθμου K-nearest neighbour (KNN) κατάφερε μια ακρίβεια της τάξης του 53.88%. Σε αυτή την έρευνα χρησιμοποιήθηκαν 216.743 αγώνες από το 2000 έως το 2017 και η ακρίβεια επιτεύχθηκε με χρήση 8 features. Ο μεγάλος όγκος των δεδομένων που χρησιμοποιήθηκε μας δείχνει ότι αποτελεί μια έγκυρη πηγή για το τι μπορούμε να περιμένουμε σε παρόμοιες προσεγγίσεις με 3 κλάσεις.

Επιπλέον μια ακόμα έρευνα με 3 κλάσεις (Kaur, 2019) κατάφερε με την χρήση του Random Forest να πετύχει ακρίβεια 57%. Μια αδυναμία αυτής της έρευνας είναι πως επιλέχθηκαν δεδομένα μόνο μιας ποδοσφαιρικής σεζόν και από ένα πρωτάθλημα της Αγγλίας καθιστώντας το αποτέλεσμα μη ικανοποιητικό για γενική χρήση.

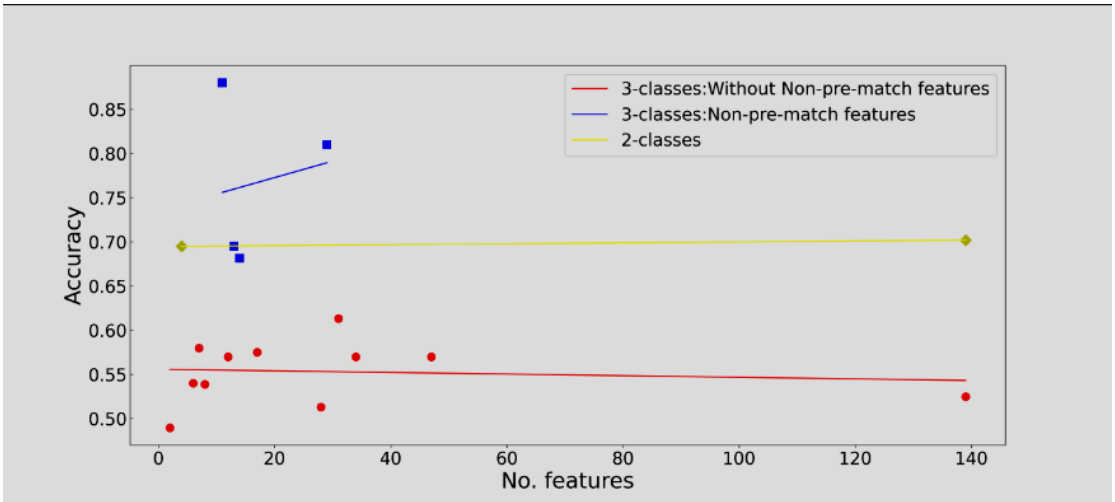
Παρόμοια αποτελέσματα παρατηρήθηκαν και στην έρευνα (Eşme, 2018) όπου επιτεύχθηκε επίσης 57% ακρίβεια αυτή την φορά με την χρήση του KNN αλγορίθμου. Η αδυναμία αυτής της έρευνας ήταν πως τα δεδομένα ήταν μιας ποδοσφαιρική σεζόν και αποκλειστικά ενός πρωταθλήματος της Τουρκίας.

Η μεγαλύτερη ακρίβεια που παρατηρήθηκε για το πρόβλημα μας με την προσέγγιση με κατηγοριοποίηση 3 κλάσεων ήταν του (Sani, 2019) , όπου επιτεύχθηκε ακρίβεια 68% με τη χρήση του Random Forest. Η σημαντική αύξηση στην ακρίβεια οφείλεται στη χρήση χαρακτηριστικών του αγώνα που δεν είναι διαθέσιμα κατά την έναρξη του, όπως το σκορ στο ημίχρονο, γεγονός που παρέχει στο μοντέλο πληροφορίες που δεν είναι συνήθως διαθέσιμες κατά τη φάση της πρόβλεψης. Επομένως τα αποτελέσματα της έρευνας μας προσφέρουν μια νέα προσέγγιση στο πρόβλημα, όμως δεν μπορούν να αξιοποιηθούν ουσιαστικά σε κάποιο σημαντικό τομέα.

Πολλές προσπάθειες για καλύτερη επίλυση του προβλήματος οδήγησε και σε προσεγγίσεις με κατηγοριοποίηση 2 κλάσεων χωρίς τον συνυπολογισμό της ισοπαλίας ως τελικό αποτέλεσμα. Μια τέτοια έρευνα είχαμε (N. Danisik, 2019) όπου η ακρίβεια έφτασε το

70% και επιτεύχθηκε με την χρήση του LSTM Regression με δεδομένα 5 χρόνων και 5 διαφορετικών πρωταθλημάτων.

Μια αναλυτική βιογραφική έρευνα πάνω στο θέμα εκτέλεσαν οι Yiming Ren & Teo Sunak (Susnjak, 2022) όπου συγκέντρωσε διάφορες έρευνες, συγκρίνοντας την τελική τους ακρίβεια δημιούργησαν το παρακάτω διάγραμμα (**Εικόνα 1**).



Εικόνα 1 - Διάγραμμα με ακρίβεια διάφορων ερευνών προς τον αριθμό των χαρακτηριστικών τους (Susnjak,2022)

Όπως παρατηρήσαμε και εμείς, οι περισσότερες έρευνες που προσεγγίζουν το θέμα με κατηγοριοποίηση 3 κλάσεων έχουν μέση ακρίβεια περίπου 55%, ενώ όσες πλησιάζουν το 60% έχουν διεξαχθεί με περιορισμένα δεδομένα που δεν επιφέρουν αξιόπιστα συμπεράσματα. Αυτό συμβαίνει για διάφορους λόγους, όπως η περιορισμένη ποικιλία των χαρακτηριστικών που χρησιμοποιούνται στα μοντέλα, η επιλογή δεδομένων μόνο από μία ποδοσφαιρική σεζόν ή από ένα μεμονωμένο πρωτάθλημα, καθώς και η χρήση χαρακτηριστικών που δεν είναι διαθέσιμα πριν την έναρξη του αγώνα.

Οι έρευνες που έχουν 2 κλάσεις στο πειραματικό τους μέρος έχουν μέση ακρίβεια 70% ενώ έρευνες που χρησιμοποιούν χαρακτηριστικά που δεν είναι διαθέσιμα κατά τη φάση της πρόβλεψης παρουσιάζουν υψηλές τιμές στις προβλέψεις τους από 70% εώς 90%, χωρίς όμως να μπορούν να αξιοποιηθούν κάπου ουσιαστικά όπως προείπαμε αυτά τα μοντέλα πρόβλεψης.

Αλλάζοντας οπτική, πέρα από τις τεχνικές προσεγγίσεις του προβλήματος, έχει υπάρξει σημαντική εστίαση στην ταυτοποίηση των πιο ισχυρών παραγόντων που επηρεάζουν την απόδοση των παικτών σε ποδοσφαιρικούς αγώνες. Πολλές μελέτες έχουν εξετάσει παραμέτρους που ξεφεύγουν από τις αμιγώς αθλητικές επιδόσεις, όπως οι ημέρες ξεκούρασης και η διάρκεια ταξιδιού πριν από τον αγώνα (Maxime Settembrea, 2024). Παρά το γεγονός ότι αυτά τα στοιχεία μπορούν να ενσωματωθούν σε ένα μεγάλο σύνολο χαρακτηριστικών, δεν είναι βέβαιο ότι θα αποτελέσουν τους πιο ισχυρούς ή κρίσιμους παράγοντες για την ακρίβεια του μοντέλου πρόβλεψης. Επομένως, ενώ τέτοιοι παράγοντες μπορεί να βελτιώσουν την

απόδοση του μοντέλου, η επίδρασή τους πρέπει να αξιολογηθεί προσεκτικά στο πλαίσιο πιο ουσιαστικών χαρακτηριστικών που επηρεάζουν άμεσα το αποτέλεσμα του αγώνα,

Στον τομέα των προβλέψεων αθλητικών γεγονότων είναι αρκετά διαδεδομένη επίσης η χρήση του ELO Ranking, μια μονάδα μέτρησης που κατατάσσει τις ομάδες με βάση την δυναμική τους χρησιμοποιώντας διάφορα στατιστικά. Η FIFA έχει εκδώσει την επίσημη φόρμουλα υπολογισμού και αρκετές έρευνες έχουν χρησιμοποιήσει το ELO Ranking (FIFA, 2018). Ορισμένες από αυτές έχουν παρουσιάσει καλά αποτελέσματα (Lars Magnus Hvattum, 2010), όμως η χρήση του παρουσιάζει κάποιες ελλείψεις σχετικά με την φόρμα των ομάδων καθώς δεν υπολογίζεται σωστά το βάρος μια νίκης. Συγκεκριμένα, το ELO Ranking σύστημα δεν λαμβάνει υπόψη τη διαφορά στο σκορ μεταξύ των ομάδων. Αυτό σημαίνει ότι είτε μια ομάδα κερδίσει οριακά με ένα γκολ διαφορά, είτε επικρατήσει με μια συντριπτική νίκη, η κατάταξή της θα αλλάξει το ίδιο. (Wheatcroft, 2021)**.**

Στις μέρες μας προκειμένου να αποδωθεί καλύτερα η δυναμική των ομάδων για τα μοντέλα πρόβλεψης, δημιουργούνται συνεχώς νέες φόρμουλες ή εξετάζονται καινούργιες στατιστικές έρευνες και παραλλαγές των συνδιασμών τους.. Η αποκρυπτογράφηση αυτού του περίπλοκου ζητήματος είναι ο λόγος που μπορούμε να εντοπίσουμε αρκετές διαφορετικές προσεγγίσεις στο χώρο των προβλεψέων.

2.2 Συμπεράσματα Βιβλιογραφικής Ερευνας

Όπως έχουμε δει, η πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων με τη χρήση μηχανικής μάθησης έχει μελετηθεί σε πολλές ερευνητικές εργασίες τα τελευταία χρόνια. Ωστόσο, πολλές από αυτές δεν έχουν συγκρίνει εκτενώς διαφορετικές μεθόδους μεταξύ τους, αλλά επικεντρώθηκαν σε μια συγκεκριμένη μέθοδο ή έναν αλγόριθμο. Ένα επιπλέον ζήτημα είναι η έλλειψη σαφών, κοινών παραγόντων που επηρεάζουν την απόδοση των μοντέλων. Κάθε έρευνα χρησιμοποιεί διαφορετικά χαρακτηριστικά (features) για να εκπαιδεύσει τα μοντέλα της, γεγονός που δυσχεραίνει τη σύγκριση μεταξύ τους και την ανεύρεση καθολικά αποδεκτών και αποτελεσματικών παραμέτρων. Ακόμα, πολλές φορές τα δεδομένα που χρησιμοποιήθηκαν για τις έρευνες δεν είναι αρκετά για να δώσουν ασφαλή αποτελέσματα. Αυτό οφείλεται στον περιορισμένο αριθμό των δεδομένων, όπως η χρήση δεδομένων από ένα μόνο πρωτάθλημα ή από μία μόνο χρονιά. Αυτή η προσέγγιση δεν μπορεί να παράγει αξιόπιστα μοντέλα, καθώς τα μοντέλα αυτά δεν είναι δυνατόν να γενικευτούν και να εφαρμοστούν σε μεγαλύτερη κλίμακα στο μέλλον, περιορίζοντας έτσι την αξιοπιστία και την εφαρμοσιμότητά τους σε διαφορετικά πλαίσια και περιόδους.

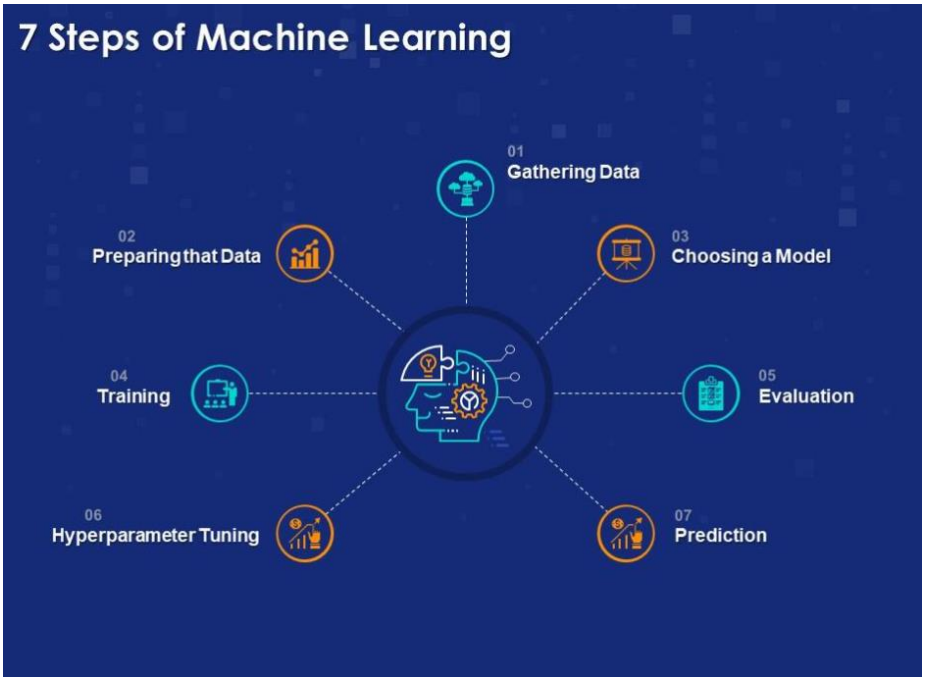
Συνεπώς, καθίσταται δύσκολη η γενίκευση των αποτελεσμάτων σε ευρύτερα πλαίσια, γεγονός που υπογραμμίζει την ανάγκη για μια νέα, πιο ολοκληρωμένη μελέτη. Αυτή θα πρέπει να εστιάσει σε μεγαλύτερο και πιο αντιπροσωπευτικό δείγμα δεδομένων, εφαρμόζοντας ταυτόχρονα πιο εξελιγμένες τεχνικές πρόβλεψης. Εξίσου σημαντική θα είναι η αναζήτηση νέων, ισχυρών χαρακτηριστικών που μπορούν να βελτιώσουν την ακρίβεια των μοντέλων πρόβλεψης. Ακόμα, ιδιαίτερη σημασία πρέπει να δοθεί στην ταξινόμηση 2 κλάσεων, καθώς οι υπάρχουσες έρευνες έχουν δείξει ότι αυτή η προσέγγιση μπορεί να προσφέρει υψηλότερη ακρίβεια, με μέσο όρο 70% (**Εικόνα 1**), αλλά δεν έχει εξεταστεί επαρκώς με ευρύτερο και πιο ποικιλόμορφο σύνολο δεδομένων. Επομένως, η περαιτέρω μελέτη σε αυτή την προσέγγιση μπορεί να οδηγήσει σε πιο αξιόπιστα και εφαρμόσιμα αποτελέσματα, που θα προσφέρουν πολύτιμες πληροφορίες για μελλοντική έρευνα.

3. Μεθοδολογία

Η δημιουργία ενός μοντέλου μηχανικής μάθησης περιλαμβάνει διάφορα στάδια, τα οποία είναι κρίσιμα για την επιτυχή ανάπτυξη και αξιολόγηση του μοντέλου. Τα βασικά στάδια είναι τα εξής:

1. **Συλλογή Δεδομένων** Η διαδικασία ξεκινά με τη συλλογή δεδομένων που είναι απαραίτητα για την εκπαίδευση του μοντέλου. Τα δεδομένα πρέπει να είναι αντιπροσωπευτικά και να περιέχουν τις πληροφορίες που είναι σημαντικές για το πρόβλημα που θέλουμε να λύσουμε.
2. **Προ-επεξεργασία Δεδομένων** Μετά τη συλλογή, τα δεδομένα πρέπει να καθαριστούν και να προετοιμαστούν. Αυτό περιλαμβάνει τη διαχείριση των ελλιπών ή θορυβωδών δεδομένων, την κανονικοποίηση των τιμών και την επιλογή των σχετικών χαρακτηριστικών.
3. **Επιλογή Μοντέλου** Επιλέγουμε τον κατάλληλο αλγόριθμο ή τα κατάλληλα αλγοριθμικά μοντέλα που θα χρησιμοποιηθούν για την εκπαίδευση. Η επιλογή αυτή βασίζεται στη φύση του προβλήματος (π.χ. ταξινόμηση, παλινδρόμηση) και στα χαρακτηριστικά των δεδομένων.
4. **Εκπαίδευση του Μοντέλου** Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για να εκπαιδεύσουμε το μοντέλο. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει να συσχετίζει τις εισόδους με τις επιθυμητές εξόδους.
5. **Αξιολόγηση του Μοντέλου** Μετά την εκπαίδευση, το μοντέλο αξιολογείται χρησιμοποιώντας ένα σύνολο δεδομένων δοκιμής. Χρησιμοποιούμε μετρικές όπως η ακρίβεια, η precision, η recall και το F1-score για να αξιολογήσουμε την απόδοση του μοντέλου.
6. **Βελτιστοποίηση του Μοντέλου** Με βάση τα αποτελέσματα της αξιολόγησης, κάνουμε βελτιστοποιήσεις στο μοντέλο, όπως τη ρύθμιση υπερπαραμέτρων, την προσθήκη περισσότερων δεδομένων ή την αλλαγή του αλγορίθμου.
7. **Ανάπτυξη και Εφαρμογή**: Το μοντέλο δοκιμάζεται σε πραγματικά δεδομένα και αναπτύσσεται στο περιβάλλον παραγωγής. Η παρακολούθηση της απόδοσης του μοντέλου στον πραγματικό κόσμο είναι απαραίτητη για να διασφαλιστεί η συνέπεια και η αξιοπιστία του.

Η προσεκτική εκτέλεση αυτών των σταδίων είναι απαραίτητη για την ανάπτυξη ενός αποτελεσματικού και αξιόπιστου μοντέλου μηχανικής μάθησης ε.



Εικόνα 2 - Βήματα Δημιουργίας ενός Μοντέλου Μηχανικής Μάθησης

3.1 Συλλογή Δεδομένων

Η συλλογή δεδομένων αποτελεί ένα από τα πλέον κρίσιμα στάδια στην ανάπτυξη μοντέλων μηχανικής μάθησης, καθώς η ποιότητα και η ακρίβεια των δεδομένων επηρεάζουν άμεσα την απόδοση του μοντέλου. Για την επίτευξη της βέλτιστης ακρίβειας, είναι απαραίτητο να βρεθούν δεδομένα από αξιόπιστες πηγές που περιέχουν χαρακτηριστικά κατάλληλα για την έρευνά μας. Η αναζήτηση αυτών των πηγών μπορεί να πραγματοποιηθεί μέσω του διαδικτύου, αξιοποιώντας κυρίως ελεύθερες βάσεις δεδομένων που παρέχουν σχετικές πληροφορίες.

Ωστόσο, πολλές φορές, η ενσωμάτωση δεδομένων από πολλαπλές πηγές μπορεί να ενισχύσει τη συνολική ακρίβεια και πληρότητα του συνόλου δεδομένων. Σε ορισμένες περιπτώσεις, κρίνεται απαραίτητο να συλλέξουμε οι ίδιοι τα δεδομένα απευθείας από τις ιστοσελίδες που περιέχουν τις πληροφορίες που χρειαζόμαστε. Αυτή η διαδικασία, γνωστή ως **web scraping**, περιλαμβάνει την αυτοματοποιημένη εξαγωγή δεδομένων από ιστοσελίδες, προκειμένου να δημιουργηθεί ένα προσαρμοσμένο σύνολο δεδομένων που να ανταποκρίνεται στις ανάγκες της συγκεκριμένης ανάλυσης. Το web scraping απαιτεί προσεκτικό σχεδιασμό και συμμόρφωση με τους όρους χρήσης των ιστοσελίδων, ώστε να διασφαλίζεται η νόμιμη και ηθική χρήση των συλλεγμένων δεδομένων.

Η χρήση του **feature importance** εντάσσεται στο στάδιο της **επιλογής χαρακτηριστικών (feature selection)** κατά την προ-επεξεργασία των δεδομένων.

Feature Importance: Αυτή η διαδικασία χρησιμοποιείται για την αξιολόγηση της σχετικής σημασίας των χαρακτηριστικών που συνεισφέρουν στο μοντέλο. Μέσω αυτής της μεθόδου, μπορούν να εντοπιστούν τα πιο σημαντικά χαρακτηριστικά που έχουν τη μεγαλύτερη επιρροή στην απόδοση του μοντέλου. Τα λιγότερο σημαντικά χαρακτηριστικά ενδέχεται να αφαιρεθούν από το μοντέλο για να μειωθεί η πολυπλοκότητά του χωρίς να μειωθεί σημαντικά η ακρίβεια.

3.2 Προ-επεξεργασία Δεδομένων

Η **προ-επεξεργασία των δεδομένων** είναι ένα από τα πιο σημαντικά βήματα στη διαδικασία ανάπτυξης ενός μοντέλου μηχανικής μάθησης. Σε αυτό το στάδιο, τα συλλεγμένα δεδομένα καθαρίζονται, μετασχηματίζονται και οργανώνονται κατάλληλα, ώστε να μπορούν να χρησιμοποιηθούν αποτελεσματικά από τον αλγόριθμο. Η διαδικασία αυτή περιλαμβάνει διάφορες ενέργειες, όπως:

- **Καθαρισμός δεδομένων**: Αφαίρεση ή διόρθωση τυχόν λανθασμένων ή ελλιπών δεδομένων που μπορεί να επηρεάσουν την ακρίβεια του μοντέλου. Αυτό μπορεί να περιλαμβάνει τη διαχείριση των ελλειπουσών τιμών (π.χ. συμπλήρωση με μέσες τιμές ή αφαίρεση των σχετικών παρατηρήσεων) και τον εντοπισμό και την αντιμετώπιση των ανωμαλιών ή των ακραίων τιμών.
- **Μετασχηματισμός δεδομένων**: Τροποποίηση των δεδομένων ώστε να είναι συμβατά με τον αλγόριθμο που θα χρησιμοποιηθεί. Αυτό μπορεί να περιλαμβάνει την κανονικοποίηση των αριθμητικών χαρακτηριστικών, τη μετατροπή κατηγορικών δεδομένων σε αριθμητικά μέσω τεχνικών όπως η κωδικοποίηση "one-hot", και τον χειρισμό της ημερομηνίας και της ώρας.
- **Κανονικοποίηση δεδομένων**: Η κανονικοποίηση (normalization) είναι μια τεχνική μετασχηματισμού των αριθμητικών δεδομένων, ώστε οι τιμές τους να περιορίζονται σε ένα συγκεκριμένο εύρος, όπως το [0,1]. Αυτό είναι ιδιαίτερα σημαντικό για αλγορίθμους που βασίζονται σε αποστάσεις (π.χ. K-Nearest Neighbors) ή εκείνους που επηρεάζονται από το μέγεθος των χαρακτηριστικών (π.χ. νευρωνικά δίκτυα). Ένας από τους κοινούς τρόπους κανονικοποίησης είναι η εφαρμογή της ελάχιστης-μέγιστης κλιμάκωσης (min-max scaling), όπου το εύρος των δεδομένων περιορίζεται μεταξύ του 0 και του 1. Ένας άλλος τρόπος είναι η τυποποίηση (standardization), κατά την οποία οι τιμές κατανέμονται σύμφωνα με μια κανονική κατανομή με μέσο όρο 0 και τυπική απόκλιση 1.

- **Δημιουργία νέων χαρακτηριστικών (feature engineering)**: Η διαδικασία εξαγωγής ή δημιουργίας νέων χαρακτηριστικών από τα υπάρχοντα δεδομένα, τα οποία μπορεί να προσφέρουν καλύτερη κατανόηση των υποκείμενων σχέσεων και να βελτιώσουν την απόδοση του μοντέλου.

Η σωστή προ-επεξεργασία των δεδομένων εξασφαλίζει ότι το μοντέλο μηχανικής μάθησης θα μπορέσει να μάθει με τον καλύτερο δυνατό τρόπο από τα δεδομένα, μεγιστοποιώντας την ακρίβεια και την αποτελεσματικότητά του

3.3 Επιλογή Αλγορίθμου

Εισαγωγή στους Αλγορίθμους Μηχανικής Μάθησης

Μετά από ενδελεχή βιβλιογραφική ανασκόπηση και ανάλυση προηγούμενων ερευνών στον τομέα της πρόβλεψης ποδοσφαιρικών αγώνων, επιλέχθηκε μια σειρά αλγορίθμων μηχανικής μάθησης που χαρακτηρίζονται από διαφορετικές υποκείμενες μεθοδολογίες (A. Singh, 2016). Η επιλογή αυτή αποσκοπεί στη διερεύνηση ποικίλων προσεγγίσεων και στην αξιολόγηση της απόδοσης κάθε αλγορίθμου στο συγκεκριμένο πρόβλημα.

Ο κύριος στόχος της έρευνας είναι ο εντοπισμός των αλγορίθμων που επιδεικνύουν τη βέλτιστη απόδοση, καθώς και η ανάδειξη τυχόν αδυναμιών και πλεονεκτημάτων στις διαφορετικές μεθοδολογίες. Οι αλγόριθμοι που επιλέχθηκαν κατατάσσονται σε διάφορες κατηγορίες, ανάλογα με τη μεθοδολογία που χρησιμοποιούν:

- **Λογιστική Παλινδρόμηση (Logistic Regression)**: Κατατάσσεται στους γραμμικούς αλγορίθμους και ανήκει στην κατηγορία των αλγορίθμων παλινδρόμησης, παρά το γεγονός ότι εφαρμόζεται σε προβλήματα ταξινόμησης.
- **Random Forest**: Πρόκειται για έναν αλγόριθμο ενισχυτικών μεθόδων (ensemble learning), και συγκεκριμένα για μια τεχνική που συνδυάζει πολλαπλά δέντρα απόφασης (decision trees) για την ενίσχυση της ακρίβειας της πρόβλεψης.
- **Gradient Boosting**: Επίσης ανήκει στην κατηγορία των ενισχυτικών αλγορίθμων (ensemble methods) και συγκεκριμένα στους αλγορίθμους ενίσχυσης (boosting), οι οποίοι εκπαιδεύουν διαδοχικά μοντέλα για να διορθώσουν τα λάθη των προηγούμενων.
- **K-Nearest Neighbors - k-NN**: Ανήκει στην κατηγορία των αλγορίθμων βασισμένων σε παραδείγματα (instance-based learning), όπου οι προβλέψεις γίνονται με βάση την εγγύτητα των νέων δεδομένων σε σημεία που έχουν ήδη κατηγοριοποιηθεί.
- **Multi-Layer Perceptron (MLP)**: Ανήκει στην κατηγορία των αλγορίθμων βασισμένων σε νευρωνικά δίκτυα (neural network-based algorithms). Πρόκειται για ένα τεχνητό νευρωνικό δίκτυο που περιλαμβάνει τουλάχιστον τρία στρώματα

(εισαγωγής, κρυφά και εξαγωγής) και χρησιμοποιεί την κατάβαση της κλίσης (gradient descent) για την εκπαίδευση του μοντέλου.

Κάθε ένας από αυτούς τους αλγορίθμους εφαρμόστηκε με συγκεκριμένες παραμέτρους, προκειμένου να εξεταστεί η απόδοσή τους στην πρόβλεψη των αποτελεσμάτων ποδοσφαιρικών αγώνων. Η διαφοροποίηση στις μεθοδολογίες των αλγορίθμων επιτρέπει μια ολοκληρωμένη ανάλυση και σύγκριση, προσφέροντας πολύτιμα συμπεράσματα για την απόδοση κάθε τεχνικής στο συγκεκριμένο πλαίσιο.

Στις επόμενες ενότητες θα ακολουθήσει αναλυτική παρουσίαση κάθε αλγορίθμου, περιλαμβάνοντας τη θεωρητική του βάση, τις παραμέτρους που χρησιμοποιήθηκαν, καθώς και το τρόπο λειτουργίας τους στο θέμα μας.

3.3.1 Logistic Regression

Θεωρητικό Υπόβαθρο

Η **Logistic Regression** (Λογιστική Παλινδρόμηση) είναι ένας στατιστικός αλγόριθμος που χρησιμοποιείται κυρίως για δυαδική ταξινόμηση, δηλαδή για την πρόβλεψη της πιθανότητας ότι ένα δεδομένο δείγμα ανήκει σε μία από δύο κατηγορίες. Σε αντίθεση με τη γραμμική παλινδρόμηση, η οποία μοντελοποιεί μια συνεχή εξαρτημένη μεταβλητή, η Logistic Regression επικεντρώνεται σε μια δυαδική εξαρτημένη μεταβλητή.

Ο πυρήνας του αλγορίθμου στηρίζεται στην **λογιστική συνάρτηση (sigmoid function)**, η οποία λαμβάνει ως είσοδο μια γραμμική συνάρτηση των χαρακτηριστικών εισόδου και επιστρέφει μια τιμή μεταξύ 0 και 1, η οποία ερμηνεύεται ως πιθανότητα. Η συνάρτηση sigmoid δίνεται από τη σχέση:

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

όπου z είναι η γραμμική συνάρτηση των εισόδων.

Η εκπαίδευση του μοντέλου πραγματοποιείται με την **μέγιστη εκτίμηση πιθανότητας (maximum likelihood estimation - MLE)**, μια μέθοδο που επιλέγει τις τιμές των συντελεστών έτσι ώστε να μεγιστοποιηθεί η πιθανότητα να παρατηρηθούν τα δεδομένα που δόθηκαν.

Πλεονεκτήματα

1. **Ευκολία Ερμηνείας**: Η Logistic Regression παρέχει συντελεστές που μπορούν εύκολα να ερμηνευτούν, καθώς κάθε συντελεστής δείχνει την αλλαγή στις πιθανότητες για την κατηγορία 1 με μια μονάδα αλλαγής στη μεταβλητή.
2. **Υπολογιστική Αποδοτικότητα**: Είναι γρήγορη και υπολογιστικά ελαφριά, γεγονός που την καθιστά κατάλληλη για μεγάλα σύνολα δεδομένων.

3. **Καλή Απόδοση σε Γραμμικά Διαχωρίσιμα Δεδομένα**: Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, η Logistic Regression μπορεί να αποδώσει πολύ καλά, παρέχοντας σαφείς και αξιόπιστες προβλέψεις.
4. **Ευκολία Εφαρμογής και Σταθερότητα**: Η Logistic Regression είναι αρκετά εύκολη στην υλοποίηση και σταθερή σε εφαρμογές με μικρότερες ή μετριοπαθείς ποσότητες δεδομένων.

Μειονεκτήματα

1. **Περιορισμένη σε Γραμμικά Χωρισμένα Δεδομένα**: Η Logistic Regression υποθέτει ότι υπάρχει μια γραμμική σχέση μεταξύ των χαρακτηριστικών και του λογαριθμού των πιθανοτήτων. Σε περιπτώσεις όπου τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, η απόδοση του μοντέλου μπορεί να είναι περιορισμένη.
2. **Ευαισθησία στα Ακραία Σημεία**: Η Logistic Regression μπορεί να επηρεαστεί από τα ακραία σημεία (outliers), τα οποία μπορεί να αλλοιώσουν τους συντελεστές του μοντέλου και να οδηγήσουν σε αναξιόπιστες προβλέψεις.
3. **Περιορισμένη Ικανότητα Μάθησης Σύνθετων Σχέσεων**: Σε σύγκριση με άλλους αλγορίθμους, όπως τα νευρωνικά δίκτυα ή οι μέθοδοι δέντρων, η Logistic Regression δεν μπορεί να μάθει σύνθετες σχέσεις και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.

Παρά τα μειονεκτήματα, η Logistic Regression παραμένει ένας δημοφιλής αλγόριθμος λόγω της απλότητας και της αξιοπιστίας του, ειδικά σε περιπτώσεις όπου η γραμμικότητα είναι ικανοποιητική προσέγγιση. Εφαρμόζεται ευρέως σε τομείς όπως η ιατρική για την πρόβλεψη δυαδικών εκβάσεων (π.χ., η πιθανότητα εμφάνισης μιας ασθένειας) και στη χρηματοοικονομική για την ανάλυση πιστωτικού κινδύνου.

3.3.2 Random Forest

Θεωρητικό Υπόβαθρο

Το **Random Forest** είναι ένας αλγόριθμος ταξινόμησης και παλινδρόμησης που ανήκει στην κατηγορία των μεθόδων ενίσχυσης (ensemble methods). Βασίζεται στη δημιουργία ενός συνόλου από δέντρα απόφασης (decision trees) για την επίλυση προβλημάτων ταξινόμησης ή παλινδρόμησης. Το Random Forest συνδυάζει τις προβλέψεις πολλαπλών δέντρων απόφασης για να παράγει μια πιο ακριβή και σταθερή πρόβλεψη από ό,τι μπορεί να επιτύχει ένα μεμονωμένο δέντρο απόφασης.

Ο αλγόριθμος λειτουργεί ως εξής:

1. **Δημιουργία Δειγμάτων από τα Δεδομένα (Bootstrap Sampling)**: Το Random Forest δημιουργεί πολλαπλά υποσύνολα των δεδομένων εκπαίδευσης μέσω δειγματοληψίας

με επανάθεση (bootstrapping). Κάθε υποσύνολο χρησιμοποιείται για να εκπαιδεύσει ένα ξεχωριστό δέντρο απόφασης.

2. **Κατασκευή Δέντρων Απόφασης (Decision Trees)**: Για κάθε υποσύνολο δεδομένων, κατασκευάζεται ένα δέντρο απόφασης. Κατά τη δημιουργία κάθε δέντρου, σε κάθε διακλάδωση (split) επιλέγεται τυχαία ένα υποσύνολο χαρακτηριστικών (features) από τα οποία θα γίνει η διαίρεση, γεγονός που προσθέτει ποικιλία και αποτρέπει την υπερπροσαρμογή (overfitting).
3. **Συνδυασμός των Προβλέψεων (Ensemble of Trees)**: Για να ληφθεί η τελική πρόβλεψη, οι προβλέψεις όλων των δέντρων συνδυάζονται. Στην ταξινόμηση, η τελική απόφαση λαμβάνεται με πλειοψηφική ψήφο (majority voting), ενώ στην παλινδρόμηση, υπολογίζεται ο μέσος όρος των προβλέψεων.

Πλεονεκτήματα

1. **Ανθεκτικότητα στο Overfitting**: Σε αντίθεση με τα μεμονωμένα δέντρα απόφασης, τα οποία είναι επιρρεπή στο overfitting, το Random Forest γενικεύει καλύτερα σε μη ορατά δεδομένα λόγω του συνδυασμού πολλών δέντρων.
2. **Υψηλή Ακρίβεια**: Το Random Forest συχνά επιτυγχάνει υψηλή ακρίβεια στις προβλέψεις, ιδιαίτερα σε σύνθετα προβλήματα ταξινόμησης και παλινδρόμησης.
3. **Αντιμετώπιση Χαμένων Δεδομένων**: Ο αλγόριθμος έχει τη δυνατότητα να χειρίζεται με επιτυχία τις περιπτώσεις όπου υπάρχουν ελλιπή δεδομένα, λόγω της φύσης του να χρησιμοποιεί τυχαία υποσύνολα των δεδομένων και των χαρακτηριστικών.
4. **Υπολογιστική Αποδοτικότητα**: Αν και η εκπαίδευση πολλών δέντρων μπορεί να είναι χρονοβόρα, η παράλληλη επεξεργασία καθιστά τον αλγόριθμο αποδοτικό για εφαρμογές σε μεγάλα σύνολα δεδομένων.
5. **Σημασία Χαρακτηριστικών (Feature Importance)**: Το Random Forest παρέχει ενδείξεις για τη σημασία των χαρακτηριστικών που χρησιμοποιούνται στο μοντέλο, βοηθώντας στην κατανόηση του τι οδηγεί τις προβλέψεις.

Μειονεκτήματα

1. **Υπολογιστική Πολυπλοκότητα**: Αν και το Random Forest είναι αποδοτικό για μεγάλα σύνολα δεδομένων, η διαδικασία εκπαίδευσης μπορεί να είναι αργή και να απαιτεί πολλούς υπολογιστικούς πόρους, ειδικά αν το σύνολο δεδομένων και ο αριθμός των δέντρων είναι μεγάλα.
2. **Δυσκολία Ερμηνείας**: Παρά το γεγονός ότι τα δέντρα απόφασης είναι εύκολα κατανοητά και ερμηνεύσιμα, το σύνολο δέντρων που προκύπτει από το Random Forest καθιστά δύσκολη την ερμηνεία του τελικού μοντέλου και των αποφάσεών του.
3. **Δυνατότητα για Overfitting σε Εξαιρετικά Θορυβώδη Δεδομένα**: Αν και το Random Forest είναι ανθεκτικό στο overfitting σε πολλά σενάρια, μπορεί ακόμα να

υπερπροσαρμοστεί σε πολύ θορυβώδη δεδομένα, ειδικά αν δεν χρησιμοποιηθεί σωστά η παραμετροποίηση.

Το Random Forest έχει ευρεία εφαρμογή σε ποικίλους τομείς, όπως στην ιατρική, τη χρηματοοικονομική, τη βιοπληροφορική και την ανάλυση κειμένου, λόγω της ισχυρής απόδοσης και της ευελιξίας του σε προβλήματα ταξινόμησης και παλινδρόμησης. Συχνά χρησιμοποιείται σε περιπτώσεις όπου είναι κρίσιμη η ακρίβεια των προβλέψεων και απαιτείται ανθεκτικότητα σε θορυβώδη δεδομένα.

3.3.3 Gradient Boosting

Θεωρητικό Υπόβαθρο

Το **Gradient Boosting** είναι μια τεχνική ενίσχυσης (ensemble technique) που δημιουργήθηκε από τον Jerome Friedman και χρησιμοποιείται ευρέως για προβλήματα ταξινόμησης και παλινδρόμησης. Ο αλγόριθμος αυτός βασίζεται στην ιδέα της ενίσχυσης, κατά την οποία οι προβλέψεις ενός συνόλου ασθενών μοντέλων (weak learners), όπως τα δέντρα απόφασης, συνδυάζονται για να δημιουργήσουν ένα ισχυρότερο μοντέλο.

Ο τρόπος λειτουργίας του Gradient Boosting περιλαμβάνει τα εξής βήματα:

1. **Εκπαίδευση Αρχικού Μοντέλου**: Αρχικά, εκπαιδεύεται ένα απλό μοντέλο, όπως ένα μικρό δέντρο απόφασης, με σκοπό να προβλέψει το στόχο (target variable).
2. **Υπολογισμός των Υπολοίπων (Residuals)**: Τα υπόλοιπα (residuals), δηλαδή οι διαφορές μεταξύ των πραγματικών τιμών και των τιμών που προέβλεψε το μοντέλο, υπολογίζονται. Τα υπόλοιπα αυτά αντιπροσωπεύουν τα σφάλματα του μοντέλου.
3. **Εκπαίδευση Νέου Μοντέλου στα Υπόλοιπα**: Ένα νέο μοντέλο εκπαιδεύεται για να προβλέψει τα υπόλοιπα, δηλαδή να διορθώσει τα λάθη του προηγούμενου μοντέλου.
4. **Συνδυασμός των Μοντέλων**: Το νέο μοντέλο προστίθεται στο σύνολο, και η διαδικασία επαναλαμβάνεται μέχρις ότου επιτευχθεί ορισμένος αριθμός μοντέλων ή δεν μπορούν πλέον να μειωθούν τα σφάλματα. Το τελικό αποτέλεσμα είναι ένας συνδυασμός όλων των μοντέλων που έχουν εκπαιδευτεί, όπου κάθε μοντέλο έχει εκπαιδευτεί για να διορθώσει τα λάθη των προηγούμενων.
5. **Βελτιστοποίηση Μέσω της Κατάβασης του Γραμμικού Κλίσης (Gradient Descent)**: Η βελτιστοποίηση γίνεται με τη χρήση της κατάβασης της γραμμικής κλίσης, για να ελαχιστοποιηθεί η συνάρτηση κόστους (loss function). Κάθε νέο μοντέλο προστίθεται με τέτοιο τρόπο, ώστε να κινείται προς την κατεύθυνση της μέγιστης μείωσης του σφάλματος.

Πλεονεκτήματα

1. **Υψηλή Ακρίβεια**: Το Gradient Boosting συχνά επιτυγχάνει εξαιρετική ακρίβεια, ειδικά σε προβλήματα ταξινόμησης και παλινδρόμησης με μεγάλη πολυπλοκότητα.
2. **Διαχείριση Σύνθετων Δεδομένων**: Μπορεί να χειριστεί πολύπλοκα σύνολα δεδομένων και να αποδώσει καλά ακόμη και όταν τα δεδομένα είναι ανομοιογενή ή μη γραμμικά.
3. **Ευελιξία στη Χρήση Διαφορετικών Συναρτήσεων Κόστους**: Ο αλγόριθμος μπορεί να προσαρμοστεί για να χρησιμοποιήσει διάφορες συναρτήσεις κόστους, επιτρέποντας τη βελτιστοποίηση για διαφορετικούς τύπους προβλημάτων.
4. **Εξισορρόπηση του Σφάλματος (Bias-Variance Trade-off)**: Το Gradient Boosting έχει την ικανότητα να μειώνει το σφάλμα λόγω προκατάληψης (bias) και το σφάλμα λόγω διακύμανσης (variance), κάτι που καθιστά το μοντέλο περισσότερο ισορροπημένο.

Μειονεκτήματα

1. **Ανθεκτικότητα στο Overfitting**: Παρά την ισχυρή απόδοσή του, το Gradient Boosting είναι επιρρεπές στο overfitting αν δεν ρυθμιστεί σωστά, ιδιαίτερα όταν χρησιμοποιείται σε θορυβώδη δεδομένα ή με πολύ μεγάλο αριθμό μοντέλων.
2. **Απαιτήσεις Υπολογιστικής Ισχύος**: Η εκπαίδευση ενός Gradient Boosting μοντέλου είναι συχνά αργή και απαιτεί σημαντική υπολογιστική ισχύ, ειδικά όταν χρησιμοποιείται σε μεγάλα σύνολα δεδομένων.
3. **Δυσκολία στην Παράλληλη Επεξεργασία**: Σε αντίθεση με το Random Forest, όπου τα δέντρα απόφασης μπορούν να εκπαιδευτούν παράλληλα, το Gradient Boosting απαιτεί σειριακή εκπαίδευση, καθώς κάθε μοντέλο εξαρτάται από τα προηγούμενα.
4. **Δυσκολία στην Ερμηνεία**: Η πολυπλοκότητα του μοντέλου καθιστά δύσκολη την κατανόηση και ερμηνεία των προβλέψεων, παρόλο που υπάρχουν εργαλεία για την ανάλυση της σημασίας των χαρακτηριστικών.

Το Gradient Boosting χρησιμοποιείται ευρέως σε εφαρμογές όπου η ακρίβεια είναι κρίσιμη, όπως σε προβλέψεις πωλήσεων, αναγνώριση εικόνας, ανίχνευση απάτης, και ανάλυση κινδύνου. Η ικανότητά του να προσαρμόζεται σε διαφορετικά προβλήματα το καθιστά έναν από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης.

3.3.4 K-Nearest Neighbors – kNN

Θεωρητικό Υπόβαθρο

Ο **K-Nearest Neighbors (KNN)** είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται τόσο για ταξινόμηση όσο και για παλινδρόμηση. Ο αλγόριθμος KNN βασίζεται στην ιδέα ότι η ταξινόμηση ενός δείγματος μπορεί να προκύψει από τη συνδυασμένη πληροφορία των πιο κοντινών γειτόνων του στο χώρο χαρακτηριστικών.

Στην ταξινόμηση, η πρόβλεψη γίνεται με βάση την κατηγορία των K πλησιέστερων γειτόνων. Ο αριθμός των γειτόνων (K) είναι μια υπερπαράμετρος που καθορίζει πόσοι γείτονες θα ληφθούν υπόψη για την απόφαση της κατηγορίας. Ο αλγόριθμος χρησιμοποιεί συνήθως μια μετρική απόστασης, όπως η Ευκλείδεια απόσταση, για να βρει τους πλησιέστερους γείτονες.

Για να κατατάξει ένα νέο δείγμα, ο αλγόριθμος ακολουθεί τα εξής βήματα:

1. Υπολογίζει την απόσταση μεταξύ του νέου δείγματος και όλων των παραδειγμάτων εκπαίδευσης.
2. Επιλέγει τους K γείτονες που είναι οι πιο κοντινοί.
3. Καθορίζει την κατηγορία του νέου δείγματος με βάση την πλειοψηφία των κατηγοριών των K γειτόνων.

Πλεονεκτήματα

1. **Απλότητα**: Ο αλγόριθμος KNN είναι πολύ απλός στην κατανόηση και την υλοποίηση. Δεν απαιτεί καμία μαθηματική εκπαίδευση ή προϋποθέσεις σχετικά με τη μορφή των δεδομένων.
2. **Μη Παραμετρικός**: Δεν υποθέτει καμία συγκεκριμένη μορφή κατανομής για τα δεδομένα, δηλαδή δεν απαιτεί την εκτίμηση παραμέτρων ή την προσαρμογή σε μοντέλα.
3. **Ευκολία Αντιμετώπισης Ελλιπών Δεδομένων**: Αν υπάρχουν ελλείψεις σε ορισμένα χαρακτηριστικά, ο KNN μπορεί να προσαρμοστεί για να χειριστεί τέτοιες καταστάσεις, καθώς εξετάζει μόνο τα χαρακτηριστικά που είναι διαθέσιμα για την εκτίμηση της απόστασης.
4. **Ευέλικτο**: Μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση, ανάλογα με τον τύπο της εξόδου που απαιτείται.

Μειονεκτήματα

1. **Επιβάρυνση Υπολογισμών**: Ο KNN μπορεί να γίνει υπολογιστικά ακριβός, ειδικά σε μεγάλα σύνολα δεδομένων, καθώς απαιτεί την υπολογισμό της απόστασης του νέου δείγματος με όλα τα παραδείγματα εκπαίδευσης.
2. **Ευαισθησία στο Υποκείμενο Χώρο**: Η απόδοση του αλγορίθμου μπορεί να μειωθεί αν τα χαρακτηριστικά δεν είναι κανονικοποιημένα ή αν υπάρχουν χαρακτηριστικά με διαφορετική κλίμακα.

3. **Επιβάρυνση Μνήμης**: Απαιτεί την αποθήκευση όλων των παραδειγμάτων εκπαίδευσης στη μνήμη, γεγονός που μπορεί να περιορίσει την κλίμακα των δεδομένων που μπορεί να επεξεργαστεί.
4. **Αναγνώριση Πολύπλοκων Δομών**: Ο KNN μπορεί να δυσκολευτεί να αναγνωρίσει πολύπλοκες δομές ή σχέσεις μεταξύ χαρακτηριστικών, καθώς βασίζεται αποκλειστικά στην τοπική γειτνίαση των παραδειγμάτων.

Ο αλγόριθμος KNN χρησιμοποιείται ευρέως σε διάφορες εφαρμογές μηχανικής μάθησης και εξόρυξης δεδομένων. Ορισμένα παραδείγματα εφαρμογών περιλαμβάνουν:

- **Κατηγοριοποίηση Εγγράφων**: Εντοπισμός κατηγοριών ή θεμάτων κειμένων με βάση την ομοιότητα με άλλα έγγραφα.
- **Σύσταση Προϊόντων**: Προτάσεις προϊόντων σε αγοραστές με βάση την ομοιότητα με άλλους χρήστες που έχουν παρόμοια χαρακτηριστικά.
- **Ανάλυση Εικόνας**: Αναγνώριση αντικειμένων ή προτύπων σε εικόνες με βάση την ομοιότητα με εκπαιδευμένα δείγματα.

Ο KNN είναι ιδιαίτερα χρήσιμος σε προβλήματα όπου η ομοιότητα μεταξύ των παραδειγμάτων μπορεί να καθορίσει την κατηγορία ή την τιμή των δεδομένων. Ενώ η απλότητά του το καθιστά ελκυστικό για διάφορες εφαρμογές, οι περιορισμοί του όσον αφορά την επιβάρυνση υπολογισμών και την αποθήκευση δεδομένων πρέπει να ληφθούν υπόψη κατά την εφαρμογή του.

3.3.5 Multi Layer Perceptron - MLP

Θεωρητικό Υπόβαθρο

Το Multi-Layer Perceptron (MLP) είναι ένας τύπος νευρωνικού δικτύου που χρησιμοποιείται ευρέως σε προβλήματα ταξινόμησης και παλινδρόμησης. Ανήκει στην κατηγορία των επιβλεπόμενων αλγορίθμων μάθησης και είναι γνωστός για την ικανότητά του να μαθαίνει μη γραμμικές σχέσεις μεταξύ των εισόδων και των εξόδων μέσω της διαδικασίας της Οπισθοδιάδοσης (backpropagation).

Ο τρόπος λειτουργίας του MLP περιλαμβάνει τα εξής βήματα:

1. **Δομή Νευρωνικού Δικτύου:** Το MLP αποτελείται από έναν ή περισσότερους κρυφούς νευρωνικούς στρώματα (hidden layers) που τοποθετούνται μεταξύ της εισόδου και της εξόδου. Κάθε στρώμα περιλαμβάνει νευρώνες που είναι συνδεδεμένοι με τους νευρώνες του επόμενου στρώματος. Οι συνδέσεις αυτές έχουν συντελεστές βαρών (weights) που προσαρμόζονται κατά την εκπαίδευση.
2. **Εκπαίδευση του Δικτύου:** Κατά την εκπαίδευση, τα δεδομένα εισόδου περνούν μέσα από τα κρυφά στρώματα και κάθε νευρώνας υπολογίζει μια σταθμισμένη άθροιση των

εισόδων του, την οποία στη συνέχεια περνάει από μια συνάρτηση ενεργοποίησης (activation function), όπως η ReLU ή η sigmoid. Το αποτέλεσμα αυτό χρησιμοποιείται για να προβλεφθεί η έξοδος του δικτύου.

3. **Οπισθοδιάδοση** (Backpropagation): Το σφάλμα (error) μεταξύ της προβλεπόμενης και της πραγματικής τιμής υπολογίζεται μέσω μιας συνάρτησης κόστους (loss function). Στη συνέχεια, το σφάλμα αυτό διαδίδεται πίσω μέσα στο δίκτυο, ενημερώνοντας τα βάρη των συνδέσεων μέσω του αλγορίθμου της κατάβασης του γραμμικού κλίσης (gradient descent), ώστε το σφάλμα να μειώνεται σε κάθε επανάληψη.
4. **Βελτιστοποίηση μέσω Επαναλήψεων:** Το δίκτυο εκπαιδεύεται σε πολλαπλές επαναλήψεις (epochs) μέχρι να επιτευχθεί ένα αποδεκτό επίπεδο ακρίβειας. Οι παράμετροι όπως το μέγεθος του κρυφού στρώματος, ο ρυθμός μάθησης (learning rate), και ο αριθμός των επαναλήψεων καθορίζουν την απόδοση του μοντέλου.

Πλεονεκτήματα

1. **Ικανότητα Μάθησης Μη Γραμμικών Σχέσεων:** Το MLP είναι εξαιρετικά ικανό να μαθαίνει περίπλοκες μη γραμμικές σχέσεις στα δεδομένα, κάτι που το καθιστά κατάλληλο για προβλήματα με υψηλή πολυπλοκότητα.
2. **Προσαρμοστικότητα:** Μπορεί να προσαρμοστεί σε διαφορετικά προβλήματα μεταβάλλοντας τη δομή του δικτύου, όπως τον αριθμό των στρωμάτων και των νευρώνων ανά στρώμα.
3. **Υψηλή Απόδοση σε Ποικίλα Δεδομένα:** Το MLP μπορεί να αποδώσει καλά σε διάφορα είδη δεδομένων, από εικόνες και ήχο μέχρι κατηγορίες και τιμές.
4. **Εκμάθηση Πολύπλοκων Μοτίβων:** Χάρη στην πολυεπίπεδη δομή του, το MLP μπορεί να μάθει και να αναγνωρίσει περίπλοκα μοτίβα και συσχετίσεις στα δεδομένα.

Μειονεκτήματα

1. **Απαιτήσεις Υπολογιστικής Ισχύος:** Η εκπαίδευση ενός MLP μπορεί να είναι υπολογιστικά απαιτητική, ιδιαίτερα σε μεγάλα δίκτυα με πολλά στρώματα και νευρώνες.
2. **Επιρροή από τα Υπερπαραμετρικά:** Η απόδοση του εξαρτάται σημαντικά από την επιλογή των υπερπαραμέτρων, όπως ο αριθμός των κρυφών στρωμάτων, ο αριθμός των νευρώνων, και ο ρυθμός μάθησης. Η ρύθμισή τους μπορεί να είναι δύσκολη και απαιτεί πειραματισμό.
3. **Κίνδυνος Υπερεκπαίδευσης (Overfitting):** Εάν το MLP δεν ρυθμιστεί σωστά, μπορεί να υπερεκπαιδευτεί, αποδίδοντας καλά στα δεδομένα εκπαίδευσης αλλά όχι στα νέα, μη γνωστά δεδομένα.

3.4 Μεθοδολογία Εκτέλεσης Αλγορίθμου

Η εκτέλεση ενός μοντέλου μηχανικής μάθησης είναι ένα σημαντικό βήμα στη διαδικασία κατασκευής και αξιολόγησης του μοντέλου. Αφορά την προετοιμασία των δεδομένων, τη ρύθμιση των παραμέτρων του μοντέλου και τη διαίρεση των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής. Η σωστή διαχείριση αυτών των βημάτων εξασφαλίζει ότι το μοντέλο θα έχει καλή απόδοση και ικανότητα γενίκευσης σε άγνωστα δεδομένα.

1. Χωρισμός των Δεδομένων

Πριν εκπαιδεύσουμε το μοντέλο, είναι απαραίτητο να διαχωρίσουμε τα δεδομένα μας σε διαφορετικά σύνολα. Συνήθως, αυτό γίνεται με τον εξής τρόπο:

- **Εκπαιδευτικό Σύνολο (Training Set)**: Το μεγαλύτερο μέρος των δεδομένων χρησιμοποιείται για την εκπαίδευση του μοντέλου, ώστε να μάθει τις σχέσεις μεταξύ των χαρακτηριστικών και της εξαρτημένης μεταβλητής.
- **Σύνολο Δοκιμής (Test Set)**: Ένα μικρότερο μέρος των δεδομένων χρησιμοποιείται για να αξιολογηθεί η απόδοση του μοντέλου μετά την εκπαίδευση, δίνοντας μια εκτίμηση της ικανότητας γενίκευσής του σε νέα δεδομένα.

Συνήθως, το εκπαιδευτικό σύνολο αντιστοιχεί στο 70-80% των δεδομένων, ενώ το σύνολο δοκιμής είναι το υπόλοιπο 20-30%. Αυτό μπορεί να γίνει χρησιμοποιώντας την εντολή train_test_split() από τη βιβλιοθήκη scikit-learn.

2. Ρύθμιση Παραμέτρων του Μοντέλου

Η επιλογή των παραμέτρων είναι κρίσιμη για τη σωστή εκπαίδευση του μοντέλου. Υπάρχουν δύο τύποι παραμέτρων που πρέπει να καθοριστούν:

- **Υπερπαράμετροι (Hyperparameters)**: Αυτές είναι παράμετροι που ορίζονται πριν την εκπαίδευση του μοντέλου και δεν προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης. Για παράδειγμα, στον αλγόριθμο Random Forest, η υπερπαράμετρος "αριθμός δέντρων" (n_estimators) είναι σημαντική για την απόδοση του μοντέλου. Άλλες υπερπαράμετροι περιλαμβάνουν το ποσοστό μάθησης (learning rate), το μέγεθος των φύλλων στα δέντρα απόφασης, και η regularization για την αποφυγή overfitting.

3. Διαδικασία Εκπαίδευσης

Η εκπαίδευση του μοντέλου πραγματοποιείται χρησιμοποιώντας το εκπαιδευτικό σύνολο. Στη διαδικασία αυτή, το μοντέλο προσπαθεί να μάθει τις σχέσεις μεταξύ των χαρακτηριστικών (features) και της εξαρτημένης μεταβλητής (target). Η εκπαίδευση περιλαμβάνει την εκτέλεση του αλγορίθμου που επιλέχθηκε.

Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο βελτιστοποιεί μια συνάρτηση κόστους (loss function) με στόχο να ελαχιστοποιήσει το σφάλμα πρόβλεψης. Η επιλογή της σωστής

συνάρτησης κόστους εξαρτάται από το είδος του προβλήματος που αντιμετωπίζουμε (π.χ. κατάταξη, παλινδρόμηση).

3.5 Αξιολόγηση Μοντέλου

Για την αξιολόγηση των μοντέλων, θα χρησιμοποιηθούν οι εξής μετρικές:

- **Ακρίβεια (Accuracy)**: Η ακρίβεια υπολογίζει το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των προβλέψεων. Υπολογίζεται ως:

$$Accuracy = TP + \frac{TN}{TP + TN + FP + FN}$$

όπου:

  - TP = True Positives (Αληθινά Θετικά)
  - TN = True Negatives (Αληθινά Αρνητικά)
  - FP = False Positives (Ψευδή Θετικά)
  - FN = False Negatives (Ψευδή Αρνητικά)

Η ακρίβεια παρέχει μια γενική εικόνα της συνολικής απόδοσης του μοντέλου, αλλά μπορεί να είναι παραπλανητική σε ανισόρροπα σύνολα δεδομένων.

- **Precision (Ευστοχία)**: Το Precision μετρά την ακρίβεια των προβλέψεων θετικής κατηγορίας και υπολογίζεται ως:

$$Precision = \frac{TP}{TP + FP}$$

Αυτή η μέτρηση είναι ιδιαίτερα χρήσιμη όταν το κόστος των ψευδών θετικών είναι υψηλό.

- **Recall (Ανάκληση)**: Το Recall μετρά την ικανότητα του μοντέλου να αναγνωρίζει σωστά τα πραγματικά θετικά και υπολογίζεται ως:

$$Recall = \frac{TP}{TP + FN}$$

Αυτή η μέτρηση είναι σημαντική όταν το κόστος των ψευδών αρνητικών είναι υψηλό.

- **F1-Score**: Το F1-Score είναι ο αρμονικός μέσος του Precision και του Recall, παρέχοντας μια ισορροπημένη μέτρηση της απόδοσης. Υπολογίζεται ως:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Το F1-Score είναι χρήσιμο όταν θέλουμε να συνδυάσουμε τις πληροφορίες από το Precision και το Recall σε μία μόνο μέτρηση.

• **Macro Avg:** Ο μέσος όρος των Precision, Recall και F1-Score υπολογίζεται ισότιμα για κάθε κατηγορία, χωρίς να λαμβάνει υπόψη το μέγεθος των κατηγοριών. Είναι χρήσιμο για τη μέτρηση της απόδοσης σε ισόρροπα σύνολα δεδομένων.

• **Weighted Avg:** Ο σταθμισμένος μέσος όρος των Precision, Recall και F1-Score, λαμβάνοντας υπόψη το πλήθος των δειγμάτων σε κάθε κατηγορία. Είναι χρήσιμο για την αξιολόγηση σε ανισόρροπα σύνολα δεδομένων, καθώς αντικατοπτρίζει τη συνολική απόδοση του μοντέλου.

Χρήση των Μετρικών

Οι μετρικές απόδοσης θα χρησιμοποιηθούν για να αξιολογηθεί η ικανότητα των μοντέλων να προβλέπουν τα αποτελέσματα των ποδοσφαιρικών αγώνων. Η ακρίβεια θα δώσει μια γενική εικόνα της συνολικής απόδοσης, ενώ οι Precision, Recall και F1-Score θα προσφέρουν μια πιο λεπτομερή αξιολόγηση για κάθε κατηγορία (π.χ., νίκη γηπεδούχου, νίκη φιλοξενούμενου). Αυτή η λεπτομερής ανάλυση είναι απαραίτητη για την κατανόηση της ικανότητας του μοντέλου να διαχειρίζεται τις διαφορετικές κατηγορίες και να εντοπίσει πιθανές αδυναμίες που χρειάζονται βελτίωση.

3.6 Βελτιστοποίηση

Η **βελτιστοποίηση υπερπαραμέτρων** (hyperparameter tuning) είναι μια κρίσιμη διαδικασία για την επίτευξη της βέλτιστης απόδοσης ενός μοντέλου μηχανικής μάθησης. Οι υπερπαράμετροι είναι παράμετροι που καθορίζονται πριν από την εκπαίδευση του μοντέλου και δεν ενημερώνονται κατά τη διάρκεια της εκπαίδευσης. Αυτές οι παράμετροι επηρεάζουν τη δομή του μοντέλου και τη διαδικασία εκπαίδευσής του, και η σωστή ρύθμισή τους είναι απαραίτητη για την επίτευξη της καλύτερης απόδοσης.

Η διαδικασία της βελτιστοποίησης υπερπαραμέτρων περιλαμβάνει την αναζήτηση για τις βέλτιστες ρυθμίσεις αυτών των παραμέτρων μέσω διαφόρων τεχνικών, όπως:

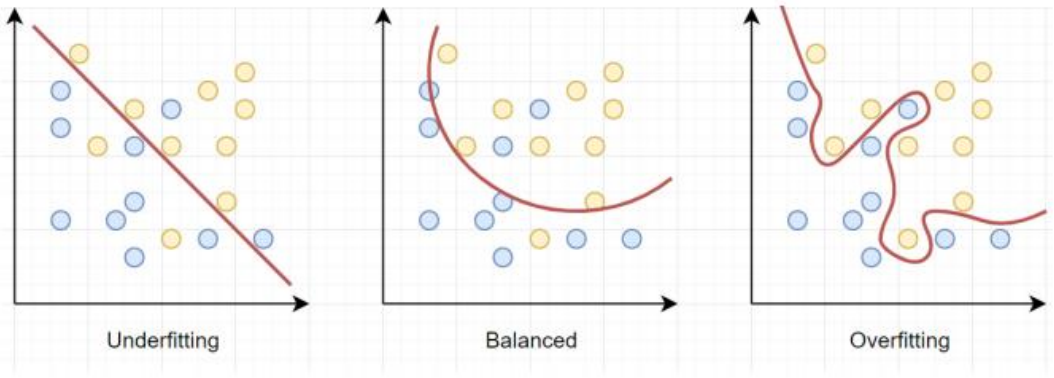
- **Αναζήτηση πλέγματος (Grid Search)**: Αυτή η μέθοδος περιλαμβάνει την εξέταση όλων των πιθανών συνδυασμών υπερπαραμέτρων μέσα σε ένα καθορισμένο εύρος τιμών. Αν και η μέθοδος αυτή μπορεί να είναι χρονοβόρα, είναι συχνά αποτελεσματική για την εύρεση των βέλτιστων ρυθμίσεων.
- **Αναζήτηση τυχαίας δειγματοληψίας (Random Search)**: Εδώ, αντί να εξετάσουμε όλους τους συνδυασμούς, επιλέγουμε τυχαία συνδυασμούς υπερπαραμέτρων από ένα προκαθορισμένο διάστημα. Αυτή η μέθοδος μπορεί να είναι λιγότερο εξαντλητική και πιο γρήγορη, ειδικά όταν ο αριθμός των υπερπαραμέτρων είναι μεγάλος.

3.7 Ανάλυση Απόδοσης

Μετά την ολοκλήρωση της βελτιστοποίησης των παραμέτρων ενός μοντέλου μέσω του **hypertuning**, είναι σημαντικό να αξιολογηθεί η απόδοσή του για να διασφαλίσουμε ότι δεν πάσχει από overfitting ή underfitting **(Εικόνα 3)** .

Αυτές οι δύο καταστάσεις μπορούν να επηρεάσουν σημαντικά την απόδοση του μοντέλου και απαιτούν προσεκτική διαχείριση για να εξασφαλίσουμε ότι το μοντέλο μας γενικεύει καλά σε νέα δεδομένα.

1. **Underfitting**: Εάν τόσο η ακρίβεια της εκπαίδευσης όσο και της δοκιμής είναι χαμηλή και δεν βελτιώνονται με την αύξηση των δεδομένων εκπαίδευσης, αυτό δείχνει ότι το μοντέλο μας είναι πολύ απλό για τα δεδομένα και παρουσιάζει underfitting.
2. **Overfitting**: Εάν η ακρίβεια της εκπαίδευσης είναι υψηλή αλλά η ακρίβεια της δοκιμής είναι σημαντικά χαμηλότερη, το μοντέλο παρουσιάζει overfitting. Αυτό σημαίνει ότι το μοντέλο έχει μάθει πολύ καλά τα δεδομένα εκπαίδευσης, αλλά δεν μπορεί να γενικεύσει σε νέα, άγνωστα δεδομένα.



Εικόνα 3 - Σύγκριση Μοντέλων με Underfitting - Balanced - Overfitting

Καμπύλες Μάθησης (Learning Curves)

Οι καμπύλες μάθησης είναι ένα από τα κύρια εργαλεία που χρησιμοποιούνται για την ανίχνευση του overfitting και του underfitting. Αυτές οι καμπύλες απεικονίζουν την απόδοση του μοντέλου σε σχέση με το μέγεθος του συνόλου δεδομένων εκπαίδευσης.

Confusion Matrix

Το Confusion Matrix είναι ένα εργαλείο που βοηθά στην κατανόηση της απόδοσης ενός ταξινομητή σε προβλήματα κατηγοριοποίησης. Αποτελείται από τέσσερις κατηγορίες αποτελεσμάτων:

True Positives (TP): Τα σωστά ταξινομημένα θετικά δείγματα.

True Negatives (TN): Τα σωστά ταξινομημένα αρνητικά δείγματα.

False Positives (FP): Τα αρνητικά δείγματα που ταξινομήθηκαν λανθασμένα ως θετικά (επίσης γνωστά ως σφάλματα τύπου Ι).
False Negatives (FN): Τα θετικά δείγματα που ταξινομήθηκαν λανθασμένα ως αρνητικά (σφάλματα τύπου ΙΙ).

Το **Confusion Matrix** προσφέρει πολύτιμες πληροφορίες για τις επιδόσεις του μοντέλου σε συγκεκριμένες κατηγορίες. Για παράδειγμα, αν οι **False Positives** (FP) είναι υψηλοί, αυτό σημαίνει ότι το μοντέλο έχει χαμηλή **Precision**, δηλαδή κάνει πολλά λάθη στην πρόβλεψη της θετικής κατηγορίας. Αν οι **False Negatives** (FN) είναι υψηλοί, αυτό υποδεικνύει χαμηλή **Recall**, δηλαδή το μοντέλο αποτυγχάνει να ανιχνεύσει αρκετές από τις πραγματικές θετικές περιπτώσεις

ROC Curve (Receiver Operating Characteristic)

Η **ROC καμπύλη** χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός δυαδικού ταξινομητή, απεικονίζοντας την ευαισθησία (True Positive Rate) έναντι της 1 - ειδικότητας (False Positive Rate) για διάφορα όρια πρόβλεψης.

- **True Positive Rate (TPR)**: Η αναλογία των σωστών θετικών ταξινομήσεων (επίσης γνωστό ως recall).
- **False Positive Rate (FPR)**: Η αναλογία των λανθασμένων θετικών ταξινομήσεων επί των συνολικών αρνητικών.

Η ROC καμπύλη μας δίνει μια συνολική εικόνα για την ικανότητα ενός ταξινομητή να διακρίνει μεταξύ των δύο κατηγοριών. Ο πιο συχνά χρησιμοποιούμενος δείκτης είναι το **AUC (Area Under the Curve)**. Μια καλή τιμή AUC πλησιάζει το 1, κάτι που σημαίνει ότι το μοντέλο έχει πολύ καλή απόδοση, ενώ μια τιμή AUC κοντά στο 0.5 δείχνει ένα μοντέλο που δεν διακρίνει σωστά τις κατηγορίες (τυχαία πρόβλεψη).

4. Πειραματικό Μέρος

Στο παρακάτω κεφάλαιο, θα αναλύσουμε το πείραμα που διεξήγαμε για την επίτευξη της μεγαλύτερης πιθανής ακρίβειας στο πρόβλημά μας. Επικεντρωνόμαστε κυρίως στην κατηγοριοποίηση δύο κλάσεων (νίκη γηπεδούχου, νίκη φιλοξενούμενου), αποκλείοντας την ισοπαλία από την ανάλυση. Επιπλέον, θα παρουσιάσουμε όλα τα βήματα που απαιτούνται για τη δημιουργία μοντέλων μηχανικής μάθησης, όπως εξηγήσαμε με βάση την μεθοδολογία του Κεφαλαίου 3, και θα αναλύσουμε τα αποτελέσματα, τονίζοντας τις βελτιώσεις που παρατηρούμε.

Στο τέλος της διαδικασίας, αφού εντοπίσουμε το βέλτιστο μοντέλο, θα υλοποιήσουμε το πρόβλημα με την προσέγγιση των τριών κλάσεων, ώστε να αποκτήσουμε μια ολοκληρωμένη εικόνα της λύσης μας και να πραγματοποιήσουμε ενδελεχή σύγκριση με παρόμοιες έρευνες που εξετάσαμε στο Κεφάλαιο 2.

Οι αλγόριθμοι που θα εξετάσουμε είναι:

- Logistic Regression
- Random Forest
- Gradient Boosting
- K-Nearest Neighbors (KNN)
- Multi-Layer Perceptron (MLP)

Η επιλογή αυτών των αλγορίθμων βασίστηκε σε προηγούμενες έρευνες που παρουσίασαν αισιόδοξα αποτελέσματα για την επίλυση παρόμοιων προβλημάτων, καθώς και σε προσωπική έρευνα που εντόπισε αυτούς τους αλγόριθμους ως ιδιαίτερα υποσχόμενους για την συγκεκριμένη εφαρμογή μας. Το γεγονός ότι αυτοί οι αλγόριθμοι έχουν θεωρηθεί αποτελεσματικοί για την αποκωδικοποίηση σύνθετων προβλημάτων μας οδήγησε στην απόφαση να τους εξετάσουμε εκτενώς στο πλαίσιο του πειράματός μας.

4.1 Εργαλεία

Για την δημιουργία και ανάπτυξη των μοντέλων χρησιμοποιήθηκε εξ ολοκλήρου η γλώσσα προγραμματισμού Python. Ο κώδικας αναπτύχθηκε στο προγραμματιστικό περιβάλλον του google colab και στο pycharm. Η επιλογή της γλώσσας έγινε διότι διαθέτει αρκετές βιβλιοθήκες που βοηθάνε στην απλοποίηση του θέματος μας.

Συγκεκριμένα έγινε χρήση των παρακάτω βιβλιοθηκών:

- **Pandas**: Για την διαχείριση των δεδομένων από τα csv αρχεία και την προεπεξεργασία τους (συγχώνευση, διαγραφή, φιλτράρισμα, μετατροπές κτλ)
- **Beautiful Soup**: Για την εξαγωγή δεδομένων από ιστοσελίδες μέσω της διαδικασίας που εξηγήσαμε προηγουμένως του web scrapping. Η βιβλιοθήκη περιέχει εντολές που αυτοματοποιούν την συλλογή επιμέρους στατιστικών με την εξαγωγή τους από τον html κώδικα των ιστοσελίδων.
- **Scikit-Learn :** Η δημοφιλής βιβλιοθήκη χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων και την αξιολόγηση τους με τις μετρικές που προσφέρει (Precision,F1-score,Accuracy κ.α.). Σε αυτή την βιβλιοθήκη υπάρχουν διάφοροι αλγόριθμοι μηχανικής μάθησης (Logistic Regression,Random Forest κτλ.), ενώ προσφέρει και τις εντολές για την ανάπτυξη των μοντέλων όπως ο διαχωρισμός των δεδομένων.

4.2 Συλλογή Δεδομένων

Βάση δεδομένων – web scrapping

Η βάση δεδομένων που χρησιμοποιήθηκε για το πειραματικό στάδιο αρχικά ήταν από την ιστοσελίδα Kaggle που περιέχει ελεύθερη πρόσβαση σε σύνολα δεδομένων από διάφορους τομείς όπως υγείας, αθλητισμού, οικονομίας κτλ. Η βάση μας περιείχε δεδομένα και στατιστικά από αγώνες από τα κυριότερα πρωταθλήματα της Ευρώπης (Cariboo, 2023) . Σε σχέση με άλλες έρευνες και πειράματα πάνω στο θέμα επιλέχθηκε να μην γίνει εστίαση σε μια ομάδα, μια σεζόν η ακόμα και ένα μοναδικό πρωτάθλημα ώστε να μπορεί το μοντέλο μας να μπορεί να χρησιμοποιηθεί σε μεγάλη κλίμακα.

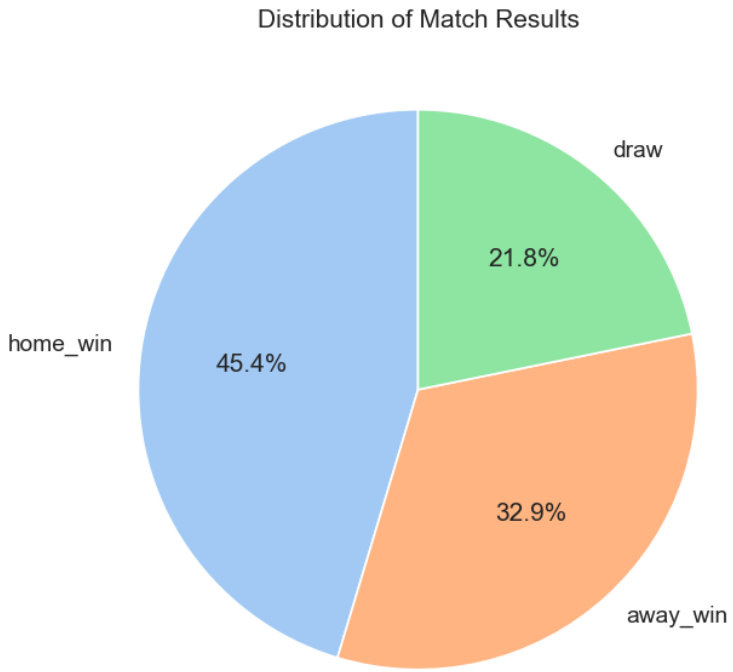
Η κύρια πηγή της βάσης δεδομένων ήταν η σελίδα transfermarkt (Transfermarkt, 2000) μια γερμανική εταιρία που είναι πολύ γνωστή στο χώρο του ποδοσφαίρου για τα έγκυρα στατιστικά που προσφέρει, ειδικότερα για τις χρηματιστηριακές αξίες των παιχτών και των ομάδων. Προκειμένου να εμπλουτίσουμε το αρχικό σύνολο των δεδομένων μας έγινε με την διαδικασία του web scrapping μια προσθήκη επιμέρους χαρακτηριστικών για τους αγώνες που προϋπήρχαν ήδη , αλλά και για συμπλήρωση τιμών που μπορεί να παραλείπονταν από το αρχικό αρχείο.

Έπειτα από την διαδικασία του web scrapping έγινε στόχευση αυτών των έγκυρων στατιστικών που προσφέρει η ιστοσελίδα για τις χρηματικές αξίες των ομάδων. Οι αξίες αυτές υποδηλώνουν πόσο κοστίζουν οι παίχτες που απαρτίζουν την ομάδα και μπορούν να θεωρηθούν έμμεσα ως τιμές ένδειξης της δυναμικότητας τους, επομένως όσο μεγαλύτερη η χρηματιστηριακή διαφορά τόσο μεγαλύτερες πιθανότητες έχει η ομάδα με την μεγαλύτερη αξία να κερδίσει τον αγώνα. Σε παρόμοιες έρευνες του χώρου όπως είδαμε τέτοιου είδους στατιστικά δεν έχουν χρησιμοποιηθεί αρκετά, αλλά προτιμούνται πιο περίπλοκες φόρμουλες όπως το ELO Ranking που αναλύσαμε στο Κεφάλαιο 2.

Το αρχείο περιείχε 59.538 αγώνες στην συνέχεια έπειτα από την προ επεξεργασία ο αριθμός τους μειώθηκε σε 30.691.

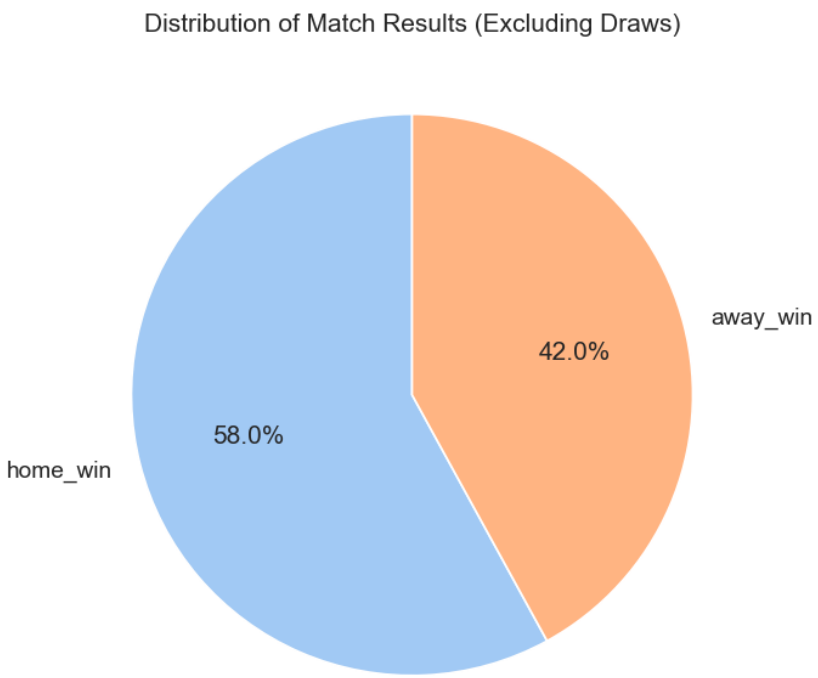
4.3 Ανάλυση Δεδομένων

Προκειμένου να κατανοήσουμε καλύτερα τα δεδομένα μας και το πρόβλημα που θα αντιμετωπίσουμε εξετάσαμε τα στατιστικά των αποτελεσμάτων όλων των αγώνων του αρχείου. Σύμφωνα με το pie chart στην **Εικόνα 4** βλέπουμε πως η νίκη της γηπεδούχου ομάδας υπερισχύει από τα υπόλοιπα αποτελέσματα με ποσοστό 45.4% σε σύγκριση με την νίκη του φιλοξενούμενου με 32.9%, ενώ η ισοπαλία δεν έχει μεγάλη συχνότητα εμφάνισης με μόλις 21.8%.



Εικόνα 4 - Κατανομή Αποτελεσμάτων Ποδοσφαιρικών Αγώνων

Για την προσέγγιση που θα ακολουθήσουμε παρακάτω μας ενδιαφέρει και η ανάλυση των αποτελεσμάτων χωρίς την ισοπαλία ως πιθανό αποτέλεσμα. **(Εικόνα 5)**



Εικόνα 5 - Κατανομή Αποτελεσμάτων Ποδοσφαιρικών Αγώνων - χωρίς την ισοπαλία

Αυτά τα διαγράμματα μας δείχνουν που περιμένουμε να κινηθούν και οι μετρικές των αλγορίθμων που θα δοκιμάσουμε αργότερα. Δηλαδή στην κατηγοριοποίηση με 2 κλάσης περιμένουμε οι αλγόριθμοι μας να εμφανίσουν μεγαλύτερη ακρίβεια στην πρόβλεψη νίκης της γηπεδούχου ομάδας λόγω της μεγαλύτερης συχνότητας εμφάνισης αυτού του αποτελέσματος.

4.4 Επιλογή Χαρακτηριστικών

Αφού έχουμε ετοιμάσει τα δεδομένα μας, η επόμενη φάση περιλαμβάνει την προσεκτική ανάλυση των χαρακτηριστικών για την αναγνώριση αυτών που έχουν τη μεγαλύτερη επίδραση στην πρόβλεψη του αποτελέσματος του αγώνα. Η διαδικασία αυτή περιλαμβάνει τη χρήση αλγορίθμων για τη μέτρηση της σημασίας των χαρακτηριστικών, με στόχο την κατανόηση του ποια χαρακτηριστικά συνεισφέρουν περισσότερο στη λήψη των αποφάσεων του μοντέλου.

Επιλογή Στόχου

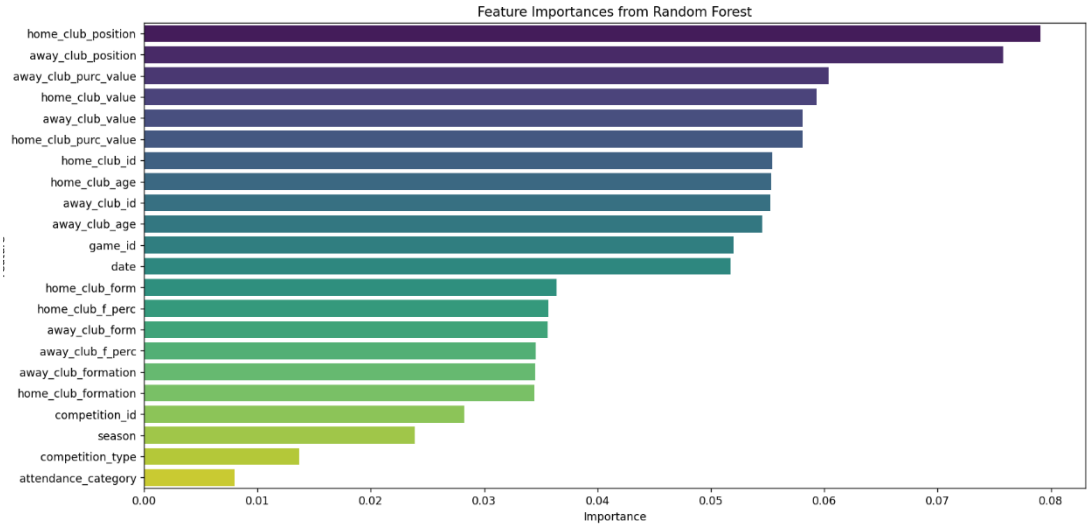
Η μεταβλητή στόχος στο αρχικό σύνολο μας αποτελεί τη στήλη result, η οποία περιέχει το τελικό αποτέλεσμα κάθε αγώνα. Η αρχική στήλη περιλάμβανε τρία πιθανά αποτελέσματα: home_win, draw, και away_win. Ωστόσο, για την ανάλυση της σημασίας των

χαρακτηριστικών, αποφασίσαμε να εξαιρέσουμε τα ισόπαλα αποτελέσματα και να επικεντρωθούμε μόνο σε δυαδικές κλάσεις (0 για νίκη της εντός έδρας ομάδας και 1 για νίκη της εκτός έδρας ομάδας). Αυτή η επιλογή διευκόλυνε την εφαρμογή αλγορίθμων που υποστηρίζουν δυαδική ταξινόμηση και απλοποίησε την ανάλυση των χαρακτηριστικών.

Διαδικασία Επιλογής Χαρακτηριστικών

Μετά την προετοιμασία των δεδομένων, εφαρμόσαμε τη μέθοδο Random Forest για την ανάλυση της σημασίας των χαρακτηριστικών **(Εικόνα 6).**

Αυτή η διαδικασία μας δείχνει τα ισχυρότερα χαρακτηριστικά και δεν αποτελεί την τελική μας επιλογή καθώς ο συνδυασμός του μπορεί να μην επιφέρει την βέλτιστη απόδοση για την πρόβλεψη μας. Μέσα από τη διαδικασία πειραματισμού, παρατηρήσαμε ότι τα χαρακτηριστικά season, home_club_formation, και away_club_formation, σε συνδυασμό με τα βασικά χαρακτηριστικά όπως οι θέσεις και οι αξίες των ομάδων, εμφάνισαν ιδιαίτερα υψηλή ακρίβεια στα μοντέλα μας.



Εικόνα 6 - Αποτελέσματα ανάλυσης σημασίας χαρακτηριστικών σε σχέση με το τελικό αποτέλεσμα

- **Home_club_position, Away_club_position, Home_club_value, Away_club_value**: Τα χαρακτηριστικά αυτά εμφανίζονται ως τα πιο σημαντικά χαρακτηριστικά. Αυτό επιβεβαιώνει ότι οι θέσεις και οι αξίες των ομάδων έχουν μεγάλη επίδραση στο αποτέλεσμα του αγώνα.
- **Season**: Παρόλο που η επίδραση της χρονιάς μπορεί να φαίνεται μικρή, ενδεχομένως παρέχει πολύτιμες πληροφορίες για την τάση της ομάδας κατά τη διάρκεια μιας συγκεκριμένης περιόδου.

- **Home and Away Club Formation**: Οι σχηματισμοί μπορεί να παίζουν κρίσιμο ρόλο στην στρατηγική των ομάδων και την προσαρμοστικότητα τους σε διαφορετικούς αντιπάλους, κάτι που μπορεί να μην αποτυπώνεται πλήρως στην ανάλυση των μεμονωμένων χαρακτηριστικών.

Αν και αυτά τα χαρακτηριστικά μπορεί να μην εμφανίζονται αρχικά ως τα πιο σημαντικά σύμφωνα με τις θεωρητικές μας προσδοκίες, η πρακτική ανάλυση και η δοκιμή των μοντέλων υπέδειξαν ότι η συνδυασμένη χρήση τους είχε εντυπωσιακά αποτελέσματα.

Στον **Πίνακα 1** έχουμε τα τελικά χαρακτηριστικά που επιλέχθηκαν:

Πίνακας 1 - Περιγραφή Επιλεγμένων Χαρακτηριστικών 1ης φάσης

Χαρακτηριστικά (Features)	Περιγραφή
Home_club_value	Η χρηματιστική αξία των παιχτών της αρχικής εντεκάδας του γηπεδούχου
Away_club_value	Η χρηματιστική αξία των παιχτών της αρχικής εντεκάδας του φιλοξενούμενου
Home_club_position	Η θέση κατάταξης της γηπεδούχου ομάδας στο πρωτάθλημα
Away_club_position	Η θέση κατάταξης της φιλοξενούμενης ομάδας στο πρωτάθλημα
Home_club_formation	Ο σχηματισμός της γηπεδούχου ομάδας
Away_club_formation	Ο σχηματισμός της φιλοξενούμενης ομάδας
Season	Την χρονιά που έχει διεξαχθεί ο συγκεκριμένος αγώνας

4.5 Προ-επεξεργασία

Φιλτράρισμα Δεδομένων:

1. Αρχικά για την ομαλή διεξαγωγή του πειράματος χρησιμοποιήσαμε φίλτρο στην ημερομηνία διεξαγωγής των αγώνων ώστε να έχουμε έγγραφες που έχουν διεξαχθεί από τις 1-1-2012 έως και την σεζόν 2022-2023.
2. Ακόμα χρησιμοποιήθηκαν μόνο αγώνες που αναφέρονταν σε εγχώρια πρωταθλήματα και αποκλείστηκαν εγγραφές αγώνων που αφορούσαν άλλες διοργανώσεις όπως εγχώρια και ευρωπαϊκά κύπελλα. Ο λόγος αυτής της ενέργειας ήταν για να μην υπάρξει σύγχυση στο μοντέλο μας με χαρακτηστικά που χρησιμοποιήθηκαν όπως η θέση των ομάδων στα πρωταθλήματα που δεν μπορεί να προσφέρει ξεκάθαρη πληροφορία όταν προέρχονται από διαφορετικές κατηγορίες.

Καθαρισμός Δεδομένων: Στην συνέχεια έγινε μετατροπή των τιμών '-' με NA με την βοήθεια του pandas και στην συνέχεια διαγράφτηκαν όσες εγγραφές δεν είχαν τιμές.

Μετασχηματισμός: Για την στήλη του αρχείου που περιέχει τους σχηματισμούς των ομάδων έγινε μετασχηματισμός της πληροφορίας με την χρήση regex expression ώστε να έχουν πιο απλοποιημένη μορφή (4-4-2 ,3-5-2 κτλ.) σε σχέση με την αρχική μορφή που περιείχε επιπλέον λέξεις 4-4-2 attacking, 4-4-2 diamond κτλ.

Encoding: Τέλος προκειμένου να έχουμε μόνο αριθμητικές τιμές για τους αλγορίθμους μας γίνεται μετατροπή των δεδομένων και για τις στήλες των σχηματισμών των ομάδων έγινε χρήση του one-hot encoding για μετατροπή των κατηγορηματικών τιμών σε αριθμητικές επίσης.

Scaling: Στην επόμενη φάση, εφαρμόσαμε κανονικοποίηση (scaling) των χαρακτηριστικών, χρησιμοποιώντας τον αλγόριθμο StandardScaler. Αυτή η διαδικασία εξασφαλίζει ότι όλα τα χαρακτηριστικά έχουν την ίδια κλίμακα και κατανομή, κάτι που είναι κρίσιμο για τη σωστή απόδοση των αλγορίθμων μηχανικής μάθησης, ειδικά για αλγορίθμους που είναι ευαίσθητοι στις κλίμακες των δεδομένων. Η κανονικοποίηση συνεισφέρει στη βελτίωση της ταχύτητας εκπαίδευσης και της ακρίβειας των μοντέλων.

Έπειτα την προ επεξεργασία των δεδομένων έχουμε σύνολο 30.691 εγγραφές για το πείραμα μας

4.6 1η φάση εκτέλεσης των αλγορίθμων

Αφού ολοκληρώσαμε την προετοιμασία των δεδομένων, προχωρήσαμε στην δημιουργία και εκτέλεση των επιλεγμένων μοντέλων μηχανικής μάθησης. Για την ανάλυση της απόδοσης των αλγορίθμων μας, ακολουθήσαμε τα εξής βήματα:

1. **Διαχωρισμός Δεδομένων**: Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης και δοκιμής με αναλογία 80%-20%. Ο διαχωρισμός αυτός διασφαλίζει ότι το μοντέλο θα εκπαιδευτεί σε ένα υποσύνολο των δεδομένων και θα αξιολογηθεί σε ένα άλλο, εξασφαλίζοντας έτσι την αξιοπιστία των αποτελεσμάτων.
2. **Κανονικοποίηση Δεδομένων**: Τα χαρακτηριστικά κανονικοποιήθηκαν χρησιμοποιώντας τον StandardScaler για να εξασφαλιστεί ότι όλα τα χαρακτηριστικά έχουν την ίδια κλίμακα. Η κανονικοποίηση βοηθά στη βελτίωση της απόδοσης των αλγορίθμων και στη σύγκριση των αποτελεσμάτων.
3. **Παράμετροι**: Δεν καθορίστηκαν συγκεκριμένοι παράμετροι στην συγκεκριμένη φάση του πειράματος επομένως έχουν χρησιμοποιηθεί οι προεπιλεγμένοι σε κάθε αλγόριθμο.

Μετά την εκτέλεση των αλγορίθμων έχουμε τα εξής αποτελέσματα του **Πίνακα 2**:

Πίνακας 2 - Αποτελέσματα Εκτέλεσης Αλγορίθμων 1ης φάσης

Αλγόριθμός	Accuracy
Logistic Regression	78%
Random Forest	78%
Gradient Boosting	78%
KNN	74%
MLP	77%

Οι αλγόριθμοι Logistic Regression, Random Forest και Gradient Boosting είχαν την μεγαλύτερη ακρίβεια με 78% ενώ ο MLP είχε εξίσου καλή απόδοση με 77%. Ο αλγόριθμος KNN είχε μικρότερη ακρίβεια της τάξης του 74% σε σύγκριση με τους άλλους αλγορίθμους γεγονός που δείχνει κάποιες αρχικές αδυναμίες. Αυτά τα αποτελέσματα αποτελούν τα πρώτα δείγματα του πειράματος χωρίς επεμβάσεις στο αρχικό dataset , στη συνέχεια θα ακολουθήσουμε κάποια βήματα για περαιτέρω βελτίωση.

4.7 Feature engineering

Μετά τα αποτελέσματα που είχαμε από την πρώτη εκτέλεση μας για να βελτιώσουμε το μοντέλο μας δημιουργήσαμε νέα χαρακτηριστικα (feature engineering). Τα νέα χαρακτηριστικά ήταν 4 και προέκυψαν από την προηγούμενη έρευνα με το διάγραμμα σημαντικότητας (feature importance) των χαρακτηριστικών που είχαμε ήδη διαθέσιμα στο αρχείο μας. Για να ενισχύσουμε την ακρίβεια του μοντέλου με τα πιο σημαντικά χαρακτηριστικά τα νέα αποτελούν την διαφορά των δύο ομάδων σε αυτές τις κατηγορίες. Συγκεκριμένα οι νέες προσθήκες περιέχονται παρακάτω στον **Πίνακα 3**:

Πίνακας 3 - Περιγραφή Επιπρόσθετων Χαρακτηριστικών 2ης φάσης

Χαρακτηριστικά (Features)	Περιγραφή
Position_difference	Η διαφορά των ομάδων στις θέσεις κατάταξης του πρωταθλήματος
Values_difference	Η χρηματιστική διαφορά των ομάδων
Form_difference	Η διαφορά της φόρμας των τελευταίων 5 αγώνων των ομάδων
Age_difference	Η διαφορά της μέσης ηλικίας των παιχτών των ομάδων

4.8 2η φάση εκτέλεσης των αλγορίθμων

Στην συνέχει αφού προσθέσουμε τα επιπλέον χαρακτηριστικά στον πρόγραμμα μας εκτελούμε ξανά τους αλγορίθμους για την 2η φάση.

Πίνακας 4 - Αποτελέσματα Εκτέλεσης Αλγορίθμων 2ης φάσης

Αλγόριθμός	Accuracy
Logistic Regression	79%
Random Forest	79%
Gradient Boosting	79%
KNN	74%
MLP	78%

Παρατηρούμε στον **Πίνακα 4** μετά την εκτέλεση των αλγορίθμων με τα επιπλέον χαρακτηριστικά μια γενική αύξηση της ακρίβειας των μοντέλων. Όλοι οι αλγόριθμοι αυξήθηκαν 1-2% στις προβλέψεις τους εκτός από το KNN, που λόγω της μεθοδολογίας του δεν μπορεί να βρεί εύκολα μοτίβα στο πρόβλημα μας στον ίδιο βαθμό με τους υπόλοιπους αλγορίθμους. Η ακρίβεια αν και δεν βελτιώθηκε σημαντικά μας δείχνει πως ενδεχομένως το αρχικό dataset με τα πρώτα χαρακτηριστικά που επιλέχθηκαν ήταν πολύ ισχυρά καθώς η ακρίβεια βρίσκεται ήδη σε υψηλά επίπεδα για το πείραμα μας.

Η περαιτέρω προσπάθεια για ανάπτυξη νέων χαρακτηριστικών θα προσφέρει σημαντική ενίσχυση στα μοντέλα , επομένως μπορούμε να προχωρήσουμε στο επόμενο βήμα για την εύρεση των βέλτιστων παραμέτρων για τους αλγόριθμούς.

4.9 Βελτιστοποίηση

Σε αυτό το βήμα έχοντας το τελικό dataset μπορούμε να αναζητήσουμε με ισχυρές μεθόδους τους κατάλληλους παραμέτρους για τον κάθε αλγόριθμο ξεχωριστά. Οι τελικοί παράμετροι θα διαμορφώσουν το τρόπο εκτέλεσης των αλγορίθμων και θα προσφέρουν το ιδανικό περιβάλλον για την ανάπτυξη του βέλτιστου μοντέλου. Οι μέθοδοι που θα χρησιμοποιήσουμε είναι ο Grid Search και ο Random Search για τον MLP, ο τρόπος λειτουργίας τους εξετάστηκε στο κεφάλαιο 3.6.

Παρακάτω ακολουθούν οι παράμετροι που επιλέχθηκαν και τα αποτελέσματα για τον κάθε αλγόριθμο.

Logistic Regression

Παράμετροι που εξετάστηκαν:

- Penalty: Το είδος της κανονικοποίησης (L1, L2) που εφαρμόζεται στο μοντέλο.
- C: Ο αντίστροφος της ισχύος της κανονικοποίησης, με τιμές [0.1, 1, 10].
- Solver: Οι αλγόριθμοι βελτιστοποίησης liblinear και saga.
- Max Iter: Ο μέγιστος αριθμός επαναλήψεων κατά την εκπαίδευση του μοντέλου, με τιμές [100, 200, 300].

Βέλτιστες Παράμετροι:

- Penalty: L2
- C: 1
- Solver: liblinear
- Max Iter: 100

Στην **Εικόνα 7** παρατηρούμε τις μετρικές αξιολόγησης του Logistic Regression με τελική ακρίβεια 79%

```
Evaluating Best Model: Logistic Regression with Grid Search
Accuracy: 0.7967095618178857

Classification Report:
              precision    recall  f1-score   support

    home_win       0.82      0.84      0.83      3648
    away_win       0.76      0.73      0.74      2491

    accuracy                           0.80      6139
   macro avg       0.79      0.79      0.79      6139
weighted avg       0.80      0.80      0.80      6139
```

Εικόνα 7 - Αποτελέσματα Βελτιστοποίησης του Logistic Regression

Random Forest

Παράμετροι που εξετάστηκαν:

- n_estimators: Ο αριθμός των δέντρων (50, 100, 200).
- max_depth: Το μέγιστο βάθος των δέντρων (None, 10, 20, 30).
- min_samples_split: Το ελάχιστο πλήθος δειγμάτων για διαχωρισμό κόμβων (2, 5, 10).
- min_samples_leaf: Το ελάχιστο πλήθος δειγμάτων ανά φύλλο (1, 2, 4).
- bootstrap: Αν θα χρησιμοποιηθούν bootstrap δείγματα για την κατασκευή των δέντρων (True, False).

Βέλτιστες Παράμετροι:

- n_estimators: 200
- max_depth: 20
- min_samples_split: 2
- min_samples_leaf: 4
- bootstrap: True

Στην **Εικόνα 8** παρατηρούμε τις μετρικές αξιολόγησης του Random Forest με τελική ακρίβεια 80%

```
Evaluating Best Model: Random Forest with Grid Search
Accuracy: 0.8001303143834501

Classification Report:
              precision    recall  f1-score   support

    home_win       0.83      0.83      0.83      3648
    away_win       0.75      0.75      0.75      2491

    accuracy                           0.80      6139
   macro avg       0.79      0.79      0.79      6139
weighted avg       0.80      0.80      0.80      6139
```

Εικόνα 8 - Αποτελέσματα Βελτιστοποίησης του Random Forest

Gradient Boosting

Παράμετροι που εξετάστηκαν:

- n_estimators: Ο αριθμός των αθροιστικών μοντέλων (100, 200).
- learning_rate: Ο ρυθμός μάθησης (0.1, 0.2).
- max_depth: Το μέγιστο βάθος κάθε δέντρου (3, 5).
- min_samples_split: Ο ελάχιστος αριθμός δειγμάτων για διαχωρισμό κόμβων (2).

Βέλτιστες Παράμετροι:

- n_estimators: 200
- learning_rate: 0.1
- max_depth: 3
- min_samples_split: 2

Στην **Εικόνα 9** παρατηρούμε τις μετρικές αξιολόγησης του Gradient Boosting με τελική ακρίβεια 80%

```
Evaluating Gradient Boosting with Best Parameters
Accuracy: 0.8001303143834501

Classification Report:
              precision    recall  f1-score   support

    home_win       0.83      0.84      0.83      3648
    away_win       0.76      0.75      0.75      2491

    accuracy                           0.80      6139
   macro avg       0.79      0.79      0.79      6139
weighted avg       0.80      0.80      0.80      6139
```

Εικόνα 9 - Αποτελέσματα Βελτιστοποίησης του Gradient Boosting

KNN

Παράμετροι που εξετάστηκαν:

- n_neighbors: 3, 5, 7, 9 (Αριθμός γειτόνων)
- weights: 'uniform', 'distance' (Ομοιόμορφα βάρη ή βάρη με βάση την απόσταση)
- p: 1, 2 (p=1 για Manhattan απόσταση, p=2 για Ευκλείδεια απόσταση)

Καλύτερες Παράμετροι (Best Parameters):

- n_neighbors: 9
- weights: 'uniform'
- p: 1 (Ευκλείδεια απόσταση)

Στην **Εικόνα 10** παρατηρούμε τις μετρικές αξιολόγησης του KNN με τελική ακρίβεια 76%

```
Evaluating Best Model: K-Nearest Neighbors with Grid Search
Accuracy: 0.7638051799967421

Classification Report:
              precision    recall  f1-score   support

    home_win       0.79      0.82      0.81      3648
    away_win       0.72      0.68      0.70      2491

    accuracy                           0.76      6139
   macro avg       0.76      0.75      0.75      6139
weighted avg       0.76      0.76      0.76      6139
```

Εικόνα 10 - Αποτελέσματα Βελτιστοποιήσης του KNN

MLP

Παράμετροι που εξετάστηκαν:

- hidden_layer_sizes: [(50,), (100,), (100, 50), (100, 100)] (Μέγεθος κρυφών επιπέδων)
- activation: ['identity', 'logistic', 'tanh', 'relu'] (Συνάρτηση ενεργοποίησης)
- solver: ['lbfgs', 'sgd', 'adam'] (Αλγόριθμος βελτιστοποίησης)
- alpha: [0.0001, 0.001, 0.01, 0.1] (Συντελεστής regularization)
- learning_rate: ['constant', 'invscaling', 'adaptive'] (Ρυθμός εκμάθησης)

Καλύτερες Παράμετροι (Best Parameters):

- hidden_layer_sizes: (100,50)
- activation: 'identity'
- solver: 'lbfgs'
- alpha: 0.01
- learning_rate: ‘constant’

Στην **Εικόνα 11** παρατηρούμε τις μετρικές αξιολόγησης του MLP με τελική ακρίβεια 79%.

```
Evaluating Best Model: Multi-Layer Perceptron with Random Search
Accuracy: 0.7965466688385731

Classification Report:
              precision    recall  f1-score   support

    home_win       0.82      0.84      0.83      3648
    away_win       0.76      0.73      0.74      2491

    accuracy                           0.80      6139
   macro avg       0.79      0.79      0.79      6139
weighted avg       0.80      0.80      0.80      6139
```

Εικόνα 11 - Αποτελέσματα Βελτιστοποιήσης του MLP

Έχοντας ολοκληρώσει και την διαδικασία της βελτιστοποίησης των αλγορίθμων με την βοήθεια των μεθόδων hypertuning παρατηρούμε στον **Πίνακα 5** πως έχουμε βελτίωση στην τελική ακρίβεια ξανά. Οι αλγόριθμοι Random Forest και Gradient Boosting ξεπέρασαν το 80% ενώ Logistic Regression και MLP πλησίασαν εξίσου με 79% . Ο αλγόριθμος KNN ξανά δεν μπόρεσε να φτάσει τις επιδόσεις των υπόλοιπων αλγορίθμων όμως μετά την βελτιστοποίηση ενισχύθηκε με 2% και έφτασε στο ικανοποιητικό 76%.

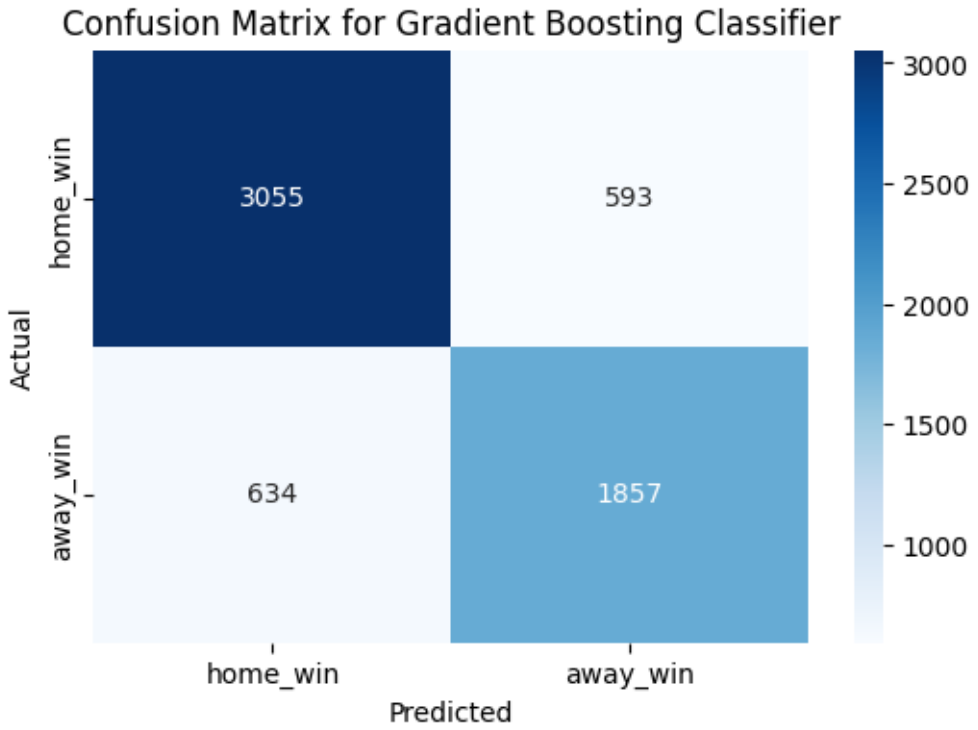
Πίνακας 5 - Σύγκριση Τελικών Αποτελεσμάτων

Αλγόριθμός	1η φάση Αρχικό Dataset	2η φάση Τελικό Dataset	Βελτιστοποίηση Παραμέτρων
Logistic Regression	78%	79%	79%
Random Forest	78%	79%	80%
Gradient Boosting	78%	79%	80%
KNN	74%	74%	76%
MLP	77%	78%	79%

Ολοκληρώνοντας την δημιουργία των μοντέλων προβλέψεις και έχοντας λάβει τα αποτελέσματα των μετρήσεων τους για να εμβαθύνουμε στην αξιοπιστία της έρευνας έγιναν επιπλέον έλεγχοι για την ποιότητα τους. Επειδή οι αλγόριθμοι παρουσίασαν παραπλήσιες τιμές στις μετρήσεις τους θα εξετάσουμε την βέλτιστη επίδοση που ήταν ο Gradient Boosting για το επόμενο βήμα ώστε να ελέγξουμε την ποιότητα του μοντέλου.

4.10 Ανάλυση Αξιοπιστίας Αποτελεσμάτων

Confusion Matrix



Εικόνα 12 - Confusion Matrix του Gradient Boosting

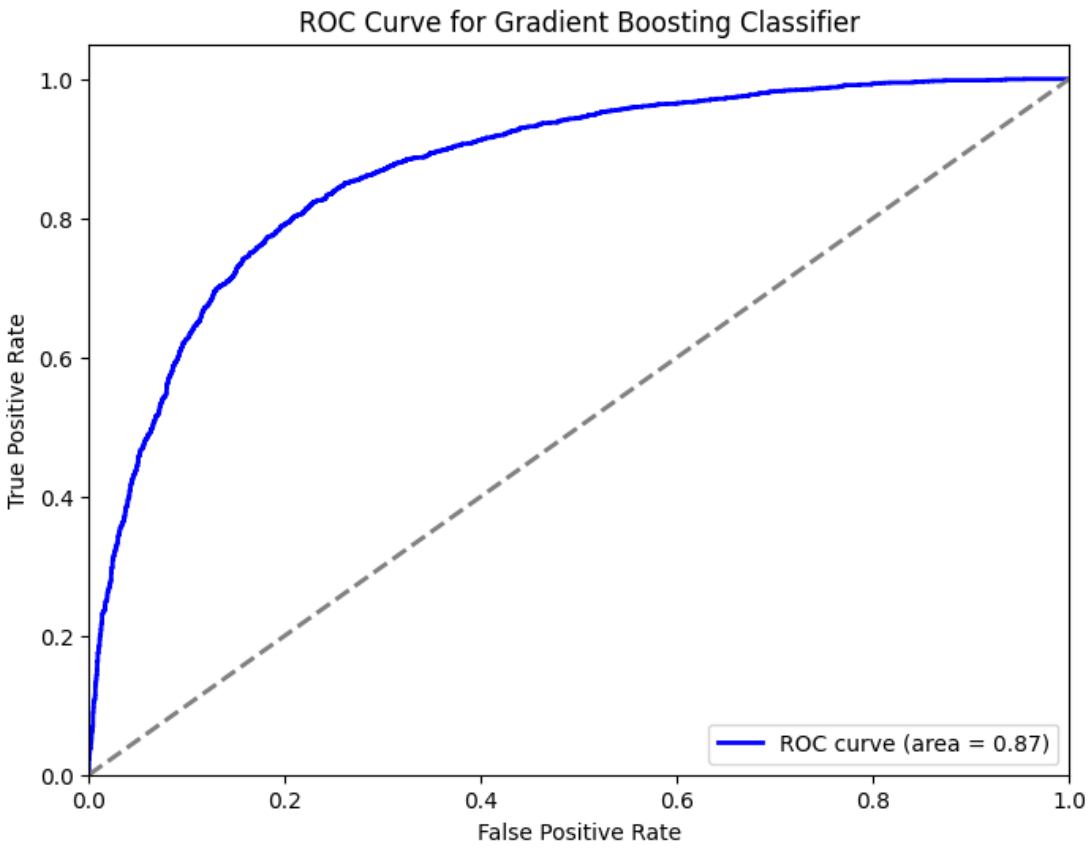
Το confusion matrix απεικονίζεται στην **Εικόνα 12** την απόδοση του **Gradient Boosting** για (νίκη γηπεδούχου "home_win" ή νίκη φιλοξενούμενου "away_win"). Κάθε στοιχείο του confusion matrix αναπαριστά τις προβλέψεις του μοντέλου σε σχέση με τις πραγματικές τιμές.

Ερμηνεία:

- Το μοντέλο έχει υψηλή ακρίβεια στις προβλέψεις για τη νίκη της γηπεδούχου ομάδας (3055 σωστές, 593 λάθος).
- Επίσης, έχει σχετικά καλή απόδοση στις προβλέψεις για τη νίκη της φιλοξενούμενης (1857 σωστές, 634 λάθος).

Το μοντέλο δείχνει μια καλή γενική απόδοση, με περισσότερες σωστές προβλέψεις παρά λάθη, αλλά υπάρχει περιθώριο βελτίωσης. Οι προβλέψεις για τη γηπεδούχο ομάδα φαίνονται ελαφρώς πιο ακριβείς σε σύγκριση με αυτές για τη φιλοξενούμενη ομάδα.

Roc Curve



Εικόνα 13 - ROC Curve του Gradient Boosting

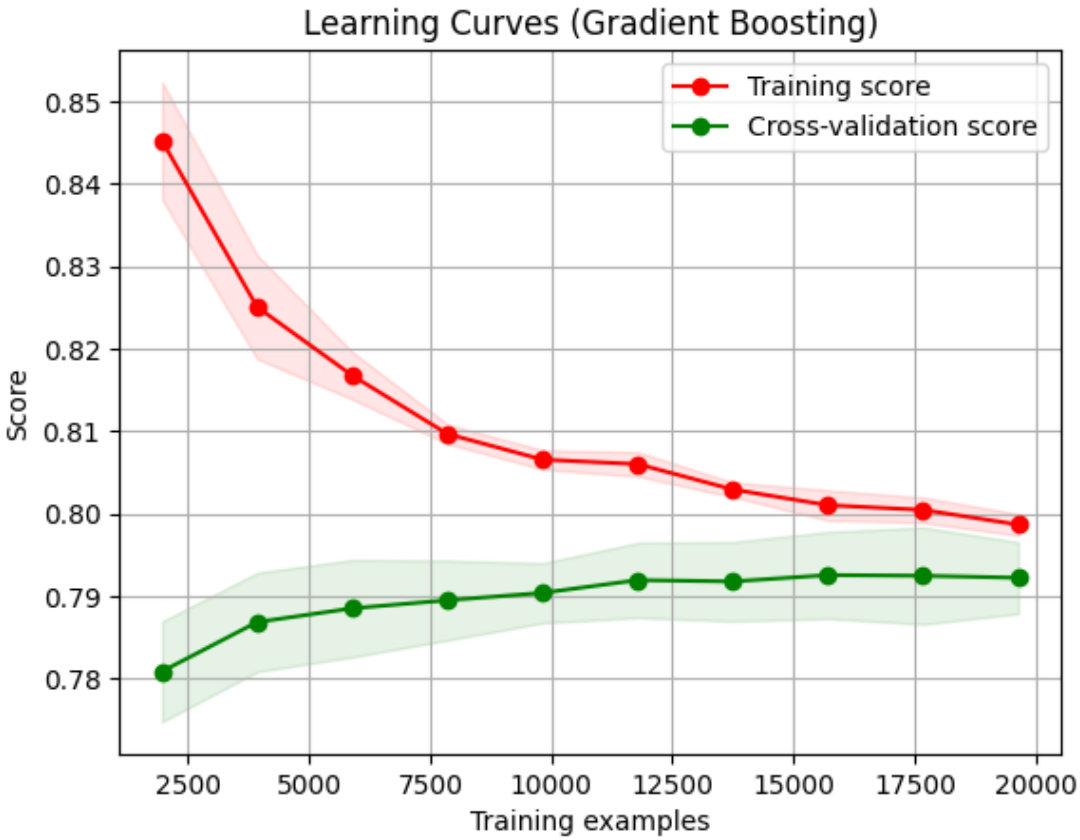
Ερμηνεία της καμπύλης ROC:

Η καμπύλη ROC της **Εικόνας 13** απεικονίζει τη σχέση μεταξύ του **True Positive Rate (TPR)**, ή αλλιώς **Recall**, και του **False Positive Rate (FPR)** για διάφορα κατώφλια απόφασης του μοντέλου. Η γραμμή διαγώνιας (γκρι γραμμή) αντιπροσωπεύει την απόδοση ενός τυχαίου ταξινομητή, όπου το μοντέλο κάνει τυχαίες προβλέψεις.

1. **Καμπύλη πάνω από τη διαγώνιο (τυχαίος ταξινομητής):**
    - Η μπλε καμπύλη βρίσκεται σημαντικά πάνω από τη γκρι διαγώνια γραμμή, η οποία αντιπροσωπεύει τον τυχαίο ταξινομητή. Αυτό σημαίνει ότι το μοντέλο είναι σαφώς καλύτερο από έναν τυχαίο ταξινομητή.
2. **Area Under the Curve (AUC = 0.87):**
    - Το **AUC (Area Under the Curve)** είναι **0.87**, το οποίο είναι ένα αρκετά καλό αποτέλεσμα.
    - Η τιμή του AUC κυμαίνεται από 0.5 (τυχαίος ταξινομητής) έως 1 (τέλειος ταξινομητής). Με το AUC να είναι 0.87, αυτό δείχνει ότι το μοντέλο έχει **ισχυρή διαχωριστική ικανότητα**, δηλαδή είναι σε θέση να διαχωρίσει με αρκετή ακρίβεια τις δύο κατηγορίες (home_win και away_win).

3. **Ισορροπία μεταξύ False Positives και True Positives:**
    - Στην περιοχή κοντά στην αρχή της καμπύλης (αριστερά), βλέπουμε ότι το μοντέλο έχει **υψηλό TPR** με χαμηλό **FPR**. Αυτό σημαίνει ότι το μοντέλο καταφέρνει να εντοπίσει πολλές σωστές προβλέψεις χωρίς να κάνει πολλά λάθη (false positives).
    - Καθώς προχωράμε προς τα δεξιά της καμπύλης, το FPR αυξάνεται, αλλά ταυτόχρονα αυξάνεται και το TPR. Αυτό δείχνει ότι το μοντέλο διατηρεί καλή ισορροπία μεταξύ του **εντοπισμού των σωστών προβλέψεων** και της **μείωσης των ψευδών θετικών προβλέψεων**.

Learning Curve



Εικόνα 14 - Learning Curves του Gradient Boosting

Το διάγραμμα στην **Εικόνα 14** δείχνει τις **καμπύλες μάθησης** για το **Gradient Boosting** μοντέλο, με σκορ για το **training set** (κόκκινη γραμμή) και το **cross-validation set** (πράσινη γραμμή) καθώς αυξάνεται ο αριθμός των παραδειγμάτων εκπαίδευσης. Οι καμπύλες μάθησης είναι χρήσιμες για να κατανοήσουμε πώς η απόδοση του μοντέλου βελτιώνεται καθώς αυξάνουμε το μέγεθος του συνόλου εκπαίδευσης.

Συμπεράσματα από τις καμπύλες μάθησης:

1. **Overfitting στην αρχή:**
    - Στην αρχή (λίγα παραδείγματα εκπαίδευσης), το μοντέλο αποδίδει πολύ καλά στο training set (**Training score ~ 0.85**), αλλά το σκορ για το cross-validation set είναι αρκετά χαμηλότερο (**Cross-validation score ~ 0.78**).
    - Αυτό υποδεικνύει **overfitting**, όπου το μοντέλο απομνημονεύει καλά τα δεδομένα εκπαίδευσης αλλά δεν γενικεύει καλά σε νέα δεδομένα.
2. **Μείωση του Overfitting:**
    - Καθώς αυξάνεται ο αριθμός των παραδειγμάτων εκπαίδευσης (περίπου από 5000 και μετά), παρατηρούμε ότι η απόδοση στο training set μειώνεται σταδιακά και πλησιάζει την απόδοση στο cross-validation set.
    - Το κλείσιμο της διαφοράς ανάμεσα στις δύο καμπύλες υποδεικνύει ότι το overfitting **μειώνεται** με την προσθήκη περισσότερων δεδομένων.
3. **Σταθεροποίηση των σκορ:**
    - Όσο αυξάνονται τα παραδείγματα εκπαίδευσης (προς τα δεξιά, στα 20.000 παραδείγματα), οι καμπύλες φαίνεται να **σταθεροποιούνται**. Το **training score** και το **cross-validation score** πλησιάζουν πολύ (training ~0.80, validation ~0.79).
    - Αυτό σημαίνει ότι το μοντέλο **γενικεύει καλύτερα** στα νέα δεδομένα καθώς μαθαίνει από περισσότερα παραδείγματα.
4. **Επάρκεια δεδομένων:**
    - Φαίνεται ότι καθώς προσθέτουμε περισσότερα δεδομένα, το μοντέλο συνεχίζει να βελτιώνεται, αν και σταδιακά η βελτίωση γίνεται μικρότερη.
    - Αυτό σημαίνει ότι το μοντέλο μπορεί να συνεχίσει να βελτιώνεται ελαφρώς με ακόμη περισσότερα δεδομένα, αλλά η βελτίωση είναι πιθανό να είναι μικρή.

4.11 Multiclass Classification

Στο τελικό στάδιο της παρούσας ερευνητικής διαδικασίας, με στόχο την περαιτέρω επέκταση της μελέτης και την ενίσχυση της χρησιμότητάς της για μελλοντικές ερευνητικές προσπάθειες, προχωρούμε στην εφαρμογή των ίδιων αλγορίθμων για την προσέγγιση του προβλήματος με ταξινόμηση 3 κλάσεων. Επομένως θα συμπεριλάβουμε την ισοπαλία στα πιθανά αποτελέσματα ενώ θα διατηρήσουμε το ίδιο πειραματικό περιβάλλον, με ίδια χαρακτηριστικά και παραμέτρους. Το νέο σύνολο δεδομένων αποτελείται από 40954 εγγραφές λόγω της προσθήκης των ισόπαλων αγώνων. Τα αποτελέσματα από την πολλαπλή ταξινόμηση φαίνονται στον **Πίνακα 6**.

Πίνακας 6 - Αποτελέσματα Αλγορίθμων σε Πολλαπλή Ταξινόμηση

Αλγόριθμός	Accuracy
Logistic Regression	59%
Random Forest	59%
Gradient Boosting	60%
KNN	55%
MLP	59%

Παρατηρούμε πως και στην πολλαπλή ταξινόμηση έχουμε τα αποτελέσματα που περιμέναμε με το gradient boosting να αποδίδει λίγο καλύτερα σε σχέση με τους υπόλοιπους αλγορίθμους ενώ ο KNN ξανά δεν φτάνει τις επιδόσεις των υπόλοιπων μεθόδων.

5. Ανάλυση Αποτελεσμάτων

5.1 Σύνοψη Αποτελεσμάτων

Συνοψίζοντας τα αποτελέσματα της έρευνας, οι μετρήσεις από τους διάφορους αλγορίθμους για το πρόβλημα ταξινόμησης δύο και τριών κατηγοριών υποδεικνύουν διαφορετικά επίπεδα απόδοσης όπως παρατηρούμε στον **Πίνακα 7**. Συγκεκριμένα, στο πρόβλημα του binary classification, οι αλγόριθμοι Random Forest και Gradient Boosting επέδειξαν την υψηλότερη ακρίβεια με **80%**, ενώ οι αλγόριθμοι Logistic Regression και MLP ακολούθησαν με 79%. Ο αλγόριθμος K-Nearest Neighbors (KNN) κατέγραψε την χαμηλότερη απόδοση, φτάνοντας στο 76%.

Πίνακας 7 - Σύγκριση Αποτελεσμάτων Δυαδικής και Πολλαπλής Ταξινόμησης

Αλγόριθμός	Binary Classification	Multiple Classification
Logistic Regression	79%	59%
Random Forest	80%	59%
Gradient Boosting	80%	60%
K-Nearest Neighbors (KNN)	76%	55%
Multi-Layer Perceptron (MLP)	79%	59%

Στο πρόβλημα της πολλαπλής ταξινόμησης, όλοι οι αλγόριθμοι παρουσίασαν σημαντική πτώση στην ακρίβειά τους, με τις τιμές να κυμαίνονται από 55% έως 60%. Πιο συγκεκριμένα, ο Gradient Boosting κατάφερε την υψηλότερη επίδοση με **60%**, ενώ οι αλγόριθμοι Logistic Regression, Random Forest, και MLP κατέγραψαν παρόμοιες αποδόσεις, φτάνοντας το 59%. Τέλος, ο KNN σημείωσε την χαμηλότερη ακρίβεια στο πρόβλημα της ταξινόμησης πολλαπλών κατηγοριών με 55%.

Αυτά τα αποτελέσματα υποδεικνύουν ότι οι περισσότεροι αλγόριθμοι μπορούν να χειριστούν σχετικά καλά την δυαδική ταξινόμηση, αλλά η απόδοσή τους μειώνεται αισθητά στην ταξινόμηση πολλαπλών κατηγοριών, με τον Gradient Boosting να αποδεικνύεται ο πιο ανθεκτικός στις διαφορετικές απαιτήσεις των δύο προβλημάτων

5.2 Συζήτηση Αποτελεσμάτων

Όπως είδαμε στα αποτελέσματα οι αλγόριθμοι που ανήκουν στην κατηγορία του ensemble learning παρουσιάζουν παρόμοια αποτελέσματα και ταιριάζουν περισσότερο στο πρόβλημα μας. Οι αλγόριθμοι αυτοί είναι ο **Random Forest** και ο **Gradient Boosting** ενώ όπως περιμέναμε ο δεύτερος παρουσιάζει ελάχιστα μεγαλύτερη ακρίβεια αφού αποτελεί μια βελτιωμένη εκδοχή του πρώτου όπου μόνο μετά την βελτιστοποίηση των παραμέτρων του κατάφερε να φτάσει τα υψηλά επίπεδα του.

Επιπλέον, οι αλγόριθμοι **Logistic Regression** και **Multi-Layer Perceptron (MLP)** παρουσιάζουν επίσης υψηλές τιμές ακρίβειας. Και οι δύο αλγόριθμοι χρησιμοποιούν αναδρομικές μεθόδους για να εκπαιδευτούν και να προσαρμοστούν στα δεδομένα. Παρόλο που ο Logistic Regression αποτελεί τον απλούστερο αλγόριθμο η ακρίβεια του βελτιώνεται αρκετά στο μοντέλο μας λόγω του scaling που έχει προηγηθεί στα χαρακτηριστικά, διαφορετικά θα εμφάνιζε αδυναμίες.

Ο αλγόριθμος του **K-Nearest Neighbors (KNN)** χρησιμοποιεί μια διαφορετική προσέγγιση από τους υπόλοιπους, καθώς δεν βασίζεται σε μοντέλα, αλλά σε αποστάσεις μεταξύ των δεδομένων σημείων. Κάθε νέα πρόβλεψη βασίζεται στη σύγκριση της με τους πιο κοντινούς γείτονές της στα δεδομένα εκπαίδευσης. Αυτή η απλή προσέγγιση, ωστόσο, δεν απέδωσε το ίδιο καλά με τους υπόλοιπους αλγόριθμους, καθώς επηρεάζεται από θόρυβο και απο περίπλοκα σύνολα δεδομένων, καθιστώντας τον λιγότερο αποδοτικό. Συμπερασματικά, από τους αλγόριθμους που εξετάσαμε, προτείνονται για περαιτέρω έρευνα όλοι εκτός από τον KNN, ο οποίος κατέγραψε επαρκή, αλλά όχι τόσο υψηλή απόδοση όσο οι υπόλοιποι.

Οι **Random Forest**, **Gradient Boosting**, **Logistic Regression** και **MLP** αποδείχθηκαν ιδιαίτερα αποτελεσματικοί, με τον Gradient Boosting να υπερέχει ελαφρώς λόγω των υψηλών του επιδόσεων και στα δυο προβλήματα με 2 και 3 κλάσεις. Επιπλέον τα διαγράμματα που δημιουργήθηκαν, όπως η **καμπύλη ROC**, το **confusion matrix** και τα **learning curves**, προσφέρουν σημαντικές ενδείξεις για την αποτελεσματικότητα του Gradient Boosting. Η **ROC καμπύλη** με περιοχή κάτω από την καμπύλη (AUC) ίση με **0.87** υποδεικνύει υψηλή διακριτική ικανότητα του αλγορίθμου, γεγονός που επιβεβαιώνει ότι μπορεί να ξεχωρίσει με ακρίβεια τις κατηγορίες σε δυαδικά προβλήματα. Το **confusion matrix** έδειξε ότι ο Gradient Boosting παρουσιάζει χαμηλά ποσοστά λανθασμένων ταξινομήσεων, ειδικά στη δυαδική ταξινόμηση, όπου οι τιμές true positives και true negatives είναι αρκετά ικανοποιητικές. Συγκεκριμένα, έχουμε καλύτερη πρόβλεψη για την νίκη του γηπεδούχου όπως περιμέναμε και από την έρευνα που προηγήθηκε λόγω μεγαλύτερης συχνότητας αυτού του αποτελέσματος.

Επιπλέον, η ανάλυση του **learning curve** αποκάλυψε ότι ο Gradient Boosting συνεχίζει να βελτιώνεται όσο αυξάνεται το πλήθος των δεδομένων εκπαίδευσης, υποδηλώνοντας ότι το μοντέλο είναι σε θέση να μάθει και να γενικεύσει καλύτερα με

περισσότερα δεδομένα. Η σύγκλιση του **training score** και του **cross-validation score** δείχνει ότι το μοντέλο βρίσκεται σε ένα σημείο ισορροπίας, όπου η εκπαίδευση και οι προβλέψεις του είναι αρκετά σταθερές χωρίς σημαντικά σημάδια overfitting με διαφορά μικρότερης του 1% ανάμεσα στις καμπύλες.

Συνολικά, ο Gradient Boosting επιδεικνύει ελαφρώς μεγαλύτερη απόδοση και προσαρμοστικότητα, γεγονός που τον καθιστά την πιο κατάλληλη επιλογή για τη συγκεκριμένη έρευνα. Η ικανότητά του να χειρίζεται αποτελεσματικά δεδομένα με σύνθετες σχέσεις και να βελτιώνει συνεχώς την απόδοσή του καθιστά τη χρήση του ιδιαίτερα ευνοϊκή για το περίπλοκο σύνολο δεδομένων μας.

6. Συμπεράσματα

Η παρούσα διπλωματική εργασία καταδεικνύει σημαντική πρόοδο σε σχέση με τις προηγούμενες έρευνες στον τομέα των μοντέλων πρόβλεψης ποδοσφαιρικών αγώνων. Όπως αποδεικνύεται από την ανάλυση των αποτελεσμάτων, η εργασία μας υπερβαίνει σημαντικά τα προηγούμενα ευρήματα. Ειδικότερα, στον τομέα της δυαδικής ταξινόμησης (binary classification), το καλύτερο προηγούμενο μοντέλο παρουσίασε μέγιστη ακρίβεια γύρω στο 70%, ενώ ο βέλτιστος αλγόριθμός μας πέτυχε ακρίβεια άνω του 80%.

Στην πολλαπλή ταξινόμηση (multiple classification), το βέλτιστο μοντέλο με τον αλγόριθμο του Gradient Boosting έφτασε ποσοστό ακρίβειας 60%, επιδεικνύοντας αισιόδοξη απόδοση ξεπερνώντας τον μέσο όρο αυτής της κατηγορίας κατά 5%. Αυτό δείχνει τη δυνατότητα του μοντέλου μας να παρέχει αξιόπιστα αποτελέσματα, ακόμη και σε πιο περίπλοκα σενάρια.

Η βελτίωση της ακρίβειας σε σύγκριση με παλαιότερες έρευνες μπορεί να αποδοθεί σε διάφορους παράγοντες. Πρώτον, η διαδικασία του **Web Scraping** αποτέλεσε κρίσιμο βήμα, επιτρέποντάς μας να συλλέξουμε δεδομένα που δεν είχαν επαρκώς χρησιμοποιηθεί σε προηγούμενες έρευνες, όπως οι χρηματιστηριακές αξίες των παικτών. Αυτές οι αξίες, οι οποίες συνήθως θεωρούνται δείκτες της ποιότητας των ομάδων, αποδείχθηκαν ιδιαίτερα ισχυρά χαρακτηριστικά για τα μοντέλα μας.

Επιπλέον, η ποικιλία του αρχικού dataset, το οποίο περιλάμβανε δεδομένα από πολλαπλές χρονιές και διοργανώσεις, προσέφερε μια πιο ολοκληρωμένη και αντιπροσωπευτική βάση για την ανάλυση όπως είχαμε θέσει ως αρχικό μας στόχο. Η δημιουργία νέων χαρακτηριστικών μέσω της διαδικασίας feature engineering, συνέβαλαν στη βελτίωση της απόδοσης των μοντέλων μας έστω και σε μικρό βαθμό. Η διαδικασία αυτή αποδεικνύει πόσο σημαντική είναι η στοχευμένη ανακάλυψη και ενσωμάτωσή τους, καθώς παρέχει κρίσιμες πληροφορίες που επιτρέπουν στα μοντέλα να αναγνωρίζουν κρυφές σχέσεις στα δεδομένα.

Ακόμα, η διαδικασία hyperparameter tuning, αν και αύξησε την ακρίβεια μόνο ελαφρώς, αναδεικνύει την αξία της βελτιστοποίησης των παραμέτρων των αλγορίθμων. Αυτά τα δύο βήματα, παρόλο που μπορεί να φαντάζουν ήπια, είναι καθοριστικά για την αναβάθμιση των μοντέλων μας και ανοίγουν το δρόμο για μελλοντική έρευνα και βελτίωση.

Είναι επίσης σημαντικό να αναφερθεί ότι, δεδομένου ότι το μοντέλο μας επικεντρώνεται στην πρόβλεψη ποδοσφαιρικών αγώνων, η απόλυτη ακρίβεια δεν μπορεί ποτέ να επιτευχθεί. Οι απρόβλεπτοι παράγοντες και το στοιχείο της έκπληξης που χαρακτηρίζει τον αθλητισμό καθιστούν τις προβλέψεις δύσκολες και υπόκεινται σε συνεχείς μεταβολές. Ωστόσο, η συστηματική προσέγγιση που υιοθετήσαμε, συνδυάζοντας ποιοτικά δεδομένα με προηγμένες

τεχνικές ανάλυσης, δείχνει τη δυνατότητα για πιο αναβαθμισμένες προβλέψεις στον τομέα της ποδοσφαιρικής ανάλυσης.

6.1 Περιορισμοί της Μελέτης

Φαίνεται ότι ένας από τους βασικούς περιορισμούς ήταν η αρχική επιλογή χαρακτηριστικών (features) που χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων. Ορισμένα από τα διαθέσιμα χαρακτηριστικά δεν παρείχαν επαρκείς πληροφορίες για να υποστηρίξουν αποτελεσματικά την εκπαίδευση, γεγονός που μας ώθησε στην ανάγκη δημιουργίας νέων χαρακτηριστικών. Παράλληλα, προχωρήσαμε και στην εξαγωγή επιπλέον στατιστικών δεδομένων από την ιστοσελίδα προέλευσης των δεδομένων, προκειμένου να εξασφαλίσουμε την πληρότητα και τη στατιστική ισχύ της έρευνας.

Μια επιπλέον αδυναμία που παρατηρήθηκε αφορά τον αριθμό των διαθέσιμων δεδομένων. Όπως έδειξε το γράφημα της καμπύλης μάθησης του μοντέλου Gradient Boosting (Εικόνα 14), υπάρχει μια μικρή ένδειξη υπερεκπαίδευσης (overfitting). Ωστόσο, το γράφημα αποκαλύπτει επίσης ότι με την προσθήκη περισσότερων δεδομένων οι καμπύλες συγκλίνουν, μειώνοντας τις διαφορές τους. Αυτό σημαίνει ότι η αύξηση του αριθμού των διαθέσιμων αγώνων στη βάση δεδομένων θα μπορούσε να συμβάλει στην εξάλειψη αυτού του φαινομένου.

Τέλος, μια αδυναμία της έρευνας είναι ότι δεν δόθηκε η ίδια προσοχή στην προσέγγιση του προβλήματος με την ταξινόμηση τριών κλάσεων, καθώς η έμφαση δόθηκε στον αρχικό στόχο, που ήταν η επίτευξη βέλτιστης απόδοσης στη δυαδική ταξινόμηση. Ωστόσο, η προσαρμογή του μοντέλου σε πολλαπλή ταξινόμηση απέδωσε πολύ καλά αποτελέσματα, παρόλο που δεν έγινε εκτενής έρευνα και δοκιμή σε αυτό το πεδίο, καθώς χρησιμοποιήθηκαν οι ίδιες μεταβλητές και παράμετροι που είχαν επιλεγεί για το προηγούμενο περιβάλλον.

6.2 Μελλοντική Έρευνα και Προτάσεις

Η χρήση οικονομικών στατιστικών, όπως οι χρηματιστηριακές αξίες των παικτών, αποδείχθηκε ιδιαίτερα ισχυρός δείκτης πρόβλεψης, ενισχύοντας την απόδοση των αλγορίθμων. Για το λόγο αυτό, προτείνεται η ενσωμάτωση τέτοιων στατιστικών στα μελλοντικά μοντέλα πρόβλεψης. Η οικονομική αξία των παικτών μπορεί να λειτουργήσει ως ένα σταθερό και αξιόπιστο μέτρο για την εκτίμηση της δυναμικότητας των ομάδων και την πιθανότητα επιτυχίας τους σε αγώνες.

Η μελλοντική έρευνα θα μπορούσε να επεκταθεί μέσω της εξερεύνησης συνδυασμών παραπάνω μεταβλητών και παραγόντων που ενδέχεται να επηρεάσουν την έκβαση ενός ποδοσφαιρικού αγώνα. Για παράδειγμα, η ανάλυση παραμέτρων όπως η ψυχολογία των παικτών, οι συνθήκες αγώνα (όπως ο καιρός και η κατάσταση του γηπέδου) και γενικότερα οι

στατιστικές που έχουν χρησιμοποιηθεί με επιτυχία από παλιότερες έρευνες μπορούν να συνδυαστούν με τα ευρήματα της παρούσας εργασίας, ώστε να δημιουργηθούν πιο αποτελεσματικά μοντέλα πρόβλεψης. Επιπλέον, η εφαρμογή νέων αλγορίθμων μηχανικής μάθησης, μπορεί να διευκολύνει τον εντοπισμό μοτίβων και σχέσεων που παραμένουν κρυφά σε παραδοσιακά μοντέλα.

Τέλος, η μελλοντική έρευνα θα μπορούσε να εξετάσει τη δυνατότητα εφαρμογής των μοντέλων σε άλλα αθλήματα, δημιουργώντας έτσι μια πλατφόρμα για τη διαρκή ανάπτυξη και εξέλιξη των προγνωστικών εργαλείων. Οι προοπτικές είναι πολλές, και η εμβάθυνση στην ανάλυση θα μπορούσε να οδηγήσει σε νέες και καινοτόμες προσεγγίσεις στον τομέα των αθλητικών προβλέψεων.

6.3 Συμβολή διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία συμβάλλει σημαντικά στην ανάπτυξη και κατανόηση των μοντέλων πρόβλεψης ποδοσφαιρικών αγώνων. Εισάγουμε νέες μεθόδους ανάλυσης δεδομένων και αλγορίθμων πρόβλεψης που προσφέρουν βελτιωμένη ακρίβεια και αξιοπιστία στις προβλέψεις των αγώνων. Χρησιμοποιώντας καινοτόμα στατιστικά στοιχεία, καταφέραμε να αναπτύξουμε μοντέλα που ξεπερνούν τα υπάρχοντα όρια των προηγούμενων ερευνών.

Τα ευρήματα της έρευνας παρέχουν πολύτιμα εργαλεία για επαγγελματίες του ποδοσφαίρου, όπως παίχτες, προπονητές και αναλυτές, αλλά και για τη βιομηχανία του στοιχηματισμού, δίνοντας τη δυνατότητα στους παίκτες να βελτιώσουν τις στρατηγικές τους με βάση των προβλέψεων, όπως έχει εφαρμοστεί και στο παρελθόν (Fátima Rodrigues, 2022) . Η εφαρμογή των μοντέλων που αναπτύξαμε μπορεί να προσφέρει ανταγωνιστικό πλεονέκτημα και να συνεισφέρει στη λήψη αποφάσεων κατά τη διάρκεια των αγώνων και στην στρατηγική προετοιμασίας.

Ακαδημαϊκά, η εργασία μας διευρύνει τη βιβλιογραφία στον τομέα των ποδοσφαιρικών προβλέψεων, προσφέροντας μεγαλύτερη βάθος σε προσεγγίσεις που δεν έχουν εξεταστεί εκτενώς, όπως η ταξινόμηση με δύο κλάσεις. Οι μελλοντικές έρευνες μπορούν να βασιστούν στα έγκυρα ευρήματά μας για να εξετάσουν περαιτέρω τη βελτίωση και εφαρμογή των μοντέλων ακόμα και σε άλλα αθλήματα.

Βιβλιογραφία

1. A. Singh, N. T. (2016). A review of supervised machine learning algorithms. *3rd International Conference on Computing for Sustainable Global Development* .
2. Ahmed, S. J. (2023). Machine Learning in Sports Analytics and Performance Prediction. *Medium*.
3. Berrar, D. L. (2019). special issue on machine learning for soccer. *Mach Learn 108, 1–7* .
4. Cariboo, D. (2023). *Football Data from Transfermarkt.* Ανάκτηση από Kaggle: https://www.kaggle.com/datasets/davidcariboo/player-scores
5. Eşme, E. &. (2018). Prediction of Football Match Outcomes . *International Journal of Machine Learning and Computing*. Ανάκτηση από https://www.researchgate.net/publication/323588333_Prediction_of_Football_Match_Outcomes_Based_On_Bookmaker_Odds_by_Using_k-Nearest_Neighbor_Algorithm/citations.
6. Fátima Rodrigues, A. P. (2022). Prediction of football match results with Machine Learning. *Procedia Computer Science*.
7. FIFA. (2018). Ανάκτηση από https://digitalhub.fifa.com/m/f99da4f73212220/original/edbm045h0udbwkqew35a-pdf.pdf.
8. FIFA. (2023). *The football landscape.*
9. García-Aliaga, A. M.-G.-S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching, 16(1), 148-157.*
10. J. Xing, H. A. (2011). Multiple Player Tracking in Sports Video: A Dual-Mode Two-Way Bayesian Inference Approach With Progressive Observation Modeling. *IEEE Transactions on Image Processing, vol. 20, no. 6, pp. 1652-1667, June 2011,.*
11. Jauhiainen S., Ä. S.-P. (2019). Talent identification in soccer using a one-class. *International Journal of Computer Science in Sport* .
12. Kaur, R. B. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*. Ανάκτηση από

https://www.researchgate.net/publication/324072605_Predictive_analysis_and_modelling_football_results_using_machine_learning_approach_for_English_Premier_League/citations.

13. Lars Magnus Hvattum, H. A. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*.
14. Maxime Settembrea, M. B. (2024). Factors associated with match outcomes in. *Journal of Sports Analytics* .
15. N. Danisik, P. L. (2019). Football Match Prediction using Players Attributes. *World Symposium on Digital Intelligence for Systems and Machines* . Ανάκτηση από https://sci-hub.se/https:/ieeexplore.ieee.org/document/8490613.
16. Premier League. (2024). *Economic and social impact of Premier League highlighted by report.*
17. Sani, Y. &. (2019). Ανάκτηση από https://www.mecs-press.org/ijisa/ijisa-v11-n7/IJISA-V11-N7-3.pdf.
18. Susnjak, Y. R. (2022). Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index. Ανάκτηση από https://ar5iv.labs.arxiv.org/html/2211.15734.
19. Transfermarkt. (2000). Ανάκτηση από Transfermarkt.com: https://www.transfermarkt.com/
20. Van Eetvelde, H. M. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. . *J EXP ORTOP 8, 27* .
21. Wang, Z. V. (2024). An AI assistant for football tactics. *Nat Commun 15, 1906* .
22. Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*.