



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**Πρόβλεψη Αποτελεσμάτων Αγώνων Ποδοσφαίρου
με Χρήση Μηχανικής Μάθησης και Νευρωνικών Δικτύων**

Διπλωματική Εργασία

Γεώργιος Δημήτριος Σκριάπας

Επιβλέπουσα: Ελένη Τουσίδου

Ιούνιος 2024



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

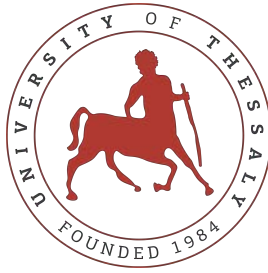
**Πρόβλεψη Αποτελεσμάτων Αγώνων Ποδοσφαίρου
με Χρήση Μηχανικής Μάθησης και Νευρωνικών Δικτύων**

Διπλωματική Εργασία

Γεώργιος Δημήτριος Σκριάπας

Επιβλέπουσα: Ελένη Τουσίδου

Ιούνιος 2024



UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Predicting Football Match Results
Using Machine Learning and Neural Networks

Diploma Thesis

Georgios Dimitrios Skriapas

Supervisor: Eleni Tousidou

June 2024

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **Ελένη Τουσίδου**

Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Βασιλακόπουλος Μιχάλης**

Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Φεύγας Αθανάσιος**

Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Γεώργιος Δημήτριος Σκριάπας

Διπλωματική Εργασία
Πρόβλεψη Αποτελεσμάτων Αγώνων Ποδοσφαίρου
με Χρήση Μηχανικής Μάθησης και Νευρωνικών Δικτύων

Γεώργιος Δημήτριος Σκριάπας

Περίληψη

Στη σημερινή εποχή, ο όγκος των δεδομένων αυξάνεται ραγδαία, δημιουργώντας την ανάγκη, αλλά και ευκαιρία, διαχείρισης και εξαγωγής χρήσιμων συμπερασμάτων. Η πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων αποτελεί μία από αυτές τις προκλήσεις. Στόχος της παρούσας εργασίας είναι η πρόβλεψη αποτελεσμάτων των αγώνων του αγγλικού πρωταθλήματος ποδοσφαίρου, χρησιμοποιώντας τεχνικές μηχανικής μάθησης και τεχνητά νευρωνικά δίκτυα. Για την υλοποίηση της εργασίας εξετάστηκαν αρκετές μέθοδοι. Χρησιμοποιήθηκαν δεδομένα αγώνων από έξι αγωνιστικές περιόδους, συλλεγμένα από διάφορους ιστότοπους. Οι μέθοδοι αξιολογήθηκαν με βάση μετρικές όπως η ακρίβεια πρόβλεψης και άλλες σχετικές. Τα αποτελέσματα δείχνουν ότι ορισμένες μέθοδοι αποδίδουν σταθερά καλύτερα, ενώ άλλες παρουσιάζουν χαμηλότερες επιδόσεις. Η μελέτη αναδεικνύει την ύπαρξη βελτιώσεων στην ακρίβεια πρόβλεψης μέσω της στρατηγικής επιλογής χαρακτηριστικών και της βελτιστοποίησης υπερπαραμέτρων. Τα αποτελέσματα υπογραμμίζουν την αποτελεσματικότητα των χρησιμοποιούμενων μεθόδων. Τέλος, παρουσιάζονται οι επιπτώσεις των αποτελεσμάτων στην αθλητική ανάλυση και προτείνονται κατευθύνσεις για μελλοντική έρευνα.

Λέξεις-κλειδιά:

Ποδόσφαιρο, Προβλέψεις, Μηχανική Μάθηση, Νευρωνικά Δίκτυα

Diploma Thesis
Predicting Football Match Results
Using Machine Learning and Neural Networks
Georgios Dimitrios Skriapas

Abstract

In today's world, the volume of data is growing rapidly, creating the need and opportunity to manage and draw useful conclusions. Predicting football match results is one of these challenges. The aim of this paper is to predict English Premier League match results using machine learning techniques and artificial neural networks. Several methods were considered to implement this project. Match data from six seasons, collected from various websites, were used. The methods were evaluated based on metrics such as prediction accuracy and other relevant metrics. The results show that some methods consistently perform better, while others underperform. The study highlights the existence of improvements in prediction accuracy through feature selection strategy and hyperparameter optimization. The results highlight the effectiveness of the methods used. Finally, the implications of the results for sports analysis are presented and directions for future research are suggested.

Keywords:

Football, Predictions, Machine Learning, Neural Networks

Πίνακας περιεχομένων

Περίληψη	x
Abstract	xi
Πίνακας περιεχομένων	xiii
Κατάλογος σχημάτων	xv
Κατάλογος πινάκων	xvii
Συντομογραφίες	xix
1 Εισαγωγή	1
1.1 Αντικείμενο της διπλωματικής	1
1.2 Οργάνωση του τόμου	2
2 Συναφείς Εργασίες	3
2.1 Εισαγωγή	3
2.2 Εργασίες αλγορίθμων μηχανικής μάθησης	3
2.3 Εργασίες νευρωνικών δικτύων	5
3 Θεωρητικό Υπόβαθρο	7
3.1 Μηχανική Μάθηση	7
3.2 Τεχνητά Νευρωνικά Δίκτυα	8
3.3 Αλγόριθμοι και Δίκτυα	9
3.3.1 Εισαγωγή στις Μέθόδους Μηχανικής Μάθησης	9
3.3.2 Gaussian Naive Bayes	9
3.3.3 Support Vector Machine	10

3.3.4	Random Forest	11
3.3.5	Logistic Regression	11
3.3.6	Gradient Boosting Decision Tree	12
3.3.7	Multilayer Perceptron	13
3.4	Προεπεξεργασία & Μηχανική Χαρακτηριστικών	14
3.5	Θεωρητική Ανάλυση Συντονισμού Υπερπαραμέτρων	16
3.6	Μετρικές Απόδοσης	17
4	Εργαλεία, Δεδομένα και Υπερπαραμέτροι	21
4.1	Python	21
4.2	Jupyter Notebook	21
4.3	Βιβλιοθήκες	22
4.4	Σύνολα Δεδομένων	23
4.5	Χαρακτηριστικά	23
4.6	Μοντέλα	25
4.7	Υπερπαραμέτροι κατά τον Συντονισμό	26
5	Αποτελέσματα και Συγκρίσεις	31
5.1	Εισαγωγή	31
5.2	Προσθήκη Κατηγοριών Χαρακτηριστικών	32
5.3	Α Φάση: Ανάλυση Αποτελεσμάτων 7 Πειραμάτων	33
5.4	Αφαίρεση Χαρακτηριστικών από το Νέο Σύνολο	37
5.5	Β Φάση: Ανάλυση Αποτελεσμάτων μετά την Επίλογή Χαρακτηριστικών	41
5.6	Συντονισμός Υπερπαραμέτρων	45
6	Συμπεράσματα	49
6.1	Εισαγωγή	49
6.2	Σύνοψη και συμπεράσματα	49
6.3	Μελλοντικές επεκτάσεις	51
	Βιβλιογραφία	53

Κατάλογος σχημάτων

3.1	Κατηγορίες Μηχανικής Μάθησης	8
3.2	Τεχνητό Νευρωνικό Δίκτυο	8
3.3	Μέθοδος Gaussian Naive Bayes	10
3.4	SVM Υπερεπίπεδο για ταξινόμηση	10
3.5	Δομή Random Forest	11
3.6	Logistic Regression	12
3.7	Διάγραμμα Ροής GBDT	13
3.8	Αρχιτεκτονική MLP	14
3.9	Γενική Μορφή Πίνακα Σύγκρισης 3 Κλάσεων	19
5.1	Πίνακες Μετρικών για τα μοντέλα GBDT και SVM	36
5.2	Πίνακας Συσχέτισης	37
5.3	Πίνακες Μετρικών για το μοντέλο SVM	45

Κατάλογος πινάκων

3.1	Πίνακας Μετρικών Απόδοσης	20
5.1	Πίνακας Αρχικών Μοντέλων	32
5.2	Συγκεντρωτικός Πίνακας Ακρίβειας 7 Πειραμάτων	33
5.3	Συγκεντρωτικός Πίνακας F1-score 7 Πειραμάτων	34
5.4	Τα 30 πιο σημαντικά χαρακτηριστικά σύμφωνα με την Mutual Information	40
5.5	Τα 15 επιλεγμένα χαρακτηριστικά σύμφωνα με την Forward Selection . . .	41
5.6	Συγκεντρωτικός Πίνακας Ακρίβειας μετά την Επιλογή Χαρακτηριστικών .	42
5.7	Συγκεντρωτικός Πίνακας F1-score μετά την Επιλογή Χαρακτηριστικών . .	42
5.8	Πίνακας Ακρίβειας Καλύτερων Μοντέλων πριν-μετά την Επιλογή Χαρακτη- ριστικών	44
5.9	Εύρος Τιμών Υπερπαραμέτρων Μοντέλων 88 και 86 Χαρακτηριστικών . .	46
5.10	Πίνακας Μετρικών Μοντέλων 88 Χαρακτηριστικών πριν-μετά τον Συντονισμό	46
5.11	Πίνακας Μετρικών Μοντέλων 86 Χαρακτηριστικών πριν-μετά τον Συντονισμό	47
5.12	Βέλτιστες Τιμές Υπερπαραμέτρων Μοντέλων 88 Χαρακτηριστικών	47
5.13	Βέλτιστες Τιμές Υπερπαραμέτρων Μοντέλων 86 Χαρακτηριστικών	48

Συντομογραφίες

MM	μηχανική μάθηση
κ.α	και άλλα
π.χ	παραδείγματος χάρη
HPs	Hyperparameters
GaussianNB	Gaussian Naive Bayes
GBDT	Gradient Boosting Decision Tree
RF	Random Forest
SVM	Support Vector Machine
MNB	Multinomial Naive Bayes
LR	Logistic Regression
LRG	Linear Regression
KNN	K-Nearest Neighbour
NB	Naive Bayes
MLP	Multi-layer Perceptron
GDS	Grad Student Descent
GS	Grid Search
RS	Randomized Search

Κεφάλαιο 1

Εισαγωγή

Το ποδόσφαιρο, ως το πιο δημοφιλές άθλημα παγκοσμίως, ελκύει τα βλέμματα από οπαδούς, αναλυτές αλλά και άλλων ειδών ενδιαφερόμενους. Στο πεδίο της αθλητικής ανάλυσης, η πρόβλεψη των αποτελεσμάτων των ποδοσφαιρικών αγώνων αποτελεί σημαντικό πεδίο μελέτης. Αξιοποιώντας παρελθοντικά αποτελέσματα αγώνων, στατιστικά στοιχεία παικτών και πληροφορίες σχετικά με τις επιδόσεις, η εφαρμογή μεθόδων Μηχανικής Μάθησης μπορεί να βοηθήσει στην πρόβλεψη των αποτελεσμάτων νέων αγώνων, καθώς και άλλων στοιχείων που αφορούν τα στατιστικά των αγώνων. Η σημασία μιας τέτοιας πρόβλεψης αποτελέσματος εκτείνεται πέρα από την απλή περιέργεια, επηρεάζοντας τις διαδικασίες λήψης αποφάσεων για τις ομάδες, τους προπονητές και τους αναλυτές. Είτε πρόκειται για τη διαμόρφωση τακτικών παιχνιδιού, τη βελτιστοποίηση της επιλογής παικτών είτε για την αξιολόγηση των δεικτών απόδοσης, η εφαρμογή των μεθόδων αυτών παρέχει πολύτιμη πρόβλεψη στην έντονη και απρόβλεπτη φύση των ποδοσφαιρικών αγώνων. Σε αυτό το πλαίσιο λοιπόν, η κατανόηση της πολυπλοκότητας της πρόβλεψης του αποτελέσματος των αγώνων καθίσταται απαραίτητη όχι μόνο για τη βελτίωση της απόδοσης των ομάδων αλλά και για την ευρύτερη προώθηση του τομέα του αθλητισμού.

1.1 Αντικείμενο της διπλωματικής

Η παρούσα έρευνα έχει ως στόχο να συμβάλει, έστω και ελάχιστα, στην πρόοδο της πρόβλεψης των αποτελεσμάτων ποδοσφαιρικών αγώνων. Αρχικά, η έρευνα προσπαθεί να αναπτύξει ισχυρά μοντέλα πρόβλεψης αξιοποιώντας αλγορίθμους μηχανικής μάθησης και τεχνητά νευρωνικά δίκτυα. Μέσω της ανάλυσης των δεδομένων αγώνων από το παρελθόν,

ο στόχος είναι να εντοπιστούν πρότυπα και τάσεις που επηρεάζουν σημαντικά τα αποτελέσματα των αγώνων. Επιπλέον, επιδιώκεται να διερευνηθεί ο αντίκτυπος των χαρακτηριστικών στην απόδοση των μοντέλων, με στόχο την ενίσχυση της ακρίβειας πρόβλεψης μέσω της προσεκτικής επιλογής και μετατροπής τους. Ακόμη, εξετάζεται η σημασία του συντονισμού των υπερπαραμέτρων ως προς την βελτιστοποίηση της απόδοσης ενός μοντέλου, μεγιστοποιώντας έτσι την επίδοση της πρόβλεψης του κάθε μοντέλου. Τέλος, πραγματοποιείται αξιολόγηση και σύγκριση διαφόρων τεχνικών μηχανικής μάθησης και τεχνικών νευρωνικών δικτύων για τον εντοπισμό των καταλληλότερων μεθόδων.

1.2 Οργάνωση του τόμου

Η εργασία περιλαμβάνει έξι κεφάλαια. Το πρώτο κεφάλαιο, θέτει τις βάσεις με την εισαγωγή του πεδίου του προβλήματος και την έμφαση στη σημασία της πρόβλεψης γεγονότων στον κόσμο του ποδοσφαίρου, ενώ παράλληλα περιγράφει τους ερευνητικούς στόχους. Στο δεύτερο κεφάλαιο, παρατίθεται η εξέταση της υπάρχουσας βιβλιογραφίας και των μελετών που σχετίζονται με τον τομέα αυτό, προσφέροντας πολύτιμες πληροφορίες σχετικά με τις μεθοδολογίες, τις τεχνικές και τους αλγορίθμους που χρησιμοποιήθηκαν σε προηγούμενες έρευνες. Στο τρίτο κεφάλαιο, περιγράφεται η θεωρητική ανάλυση της εργασίας αυτής, αναλύεται δηλαδή το θεωρητικό υπόβαθρο για κάθε αλγόριθμο που χρησιμοποιείται, γίνεται σχετική αναφορά στην περιγραφή της προεπεξεργασίας δεδομένων και της επιλογής χαρακτηριστικών, αναλύονται τεχνικές συντονισμού υπερπαραμέτρων και οι μετρικές αξιολόγησης που χρησιμοποιούνται. Στο τέταρτο κεφάλαιο, περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε, περιλαμβάνοντας τη διαδικασία συλλογής δεδομένων, τις κατηγορίες χαρακτηριστικών και τα κριτήρια επιλογής μοντέλων, δίνοντας ιδιαίτερη έμφαση και σε άλλες προσεγγίσεις όπως ο συντονισμός υπερπαραμέτρων. Στο πέμπτο κεφάλαιο, παρουσιάζονται τα ευρήματα της έρευνας, αναλύοντας την απόδοση των διαφόρων μοντέλων πρόβλεψης μέσα από διάφορες μετρικές και εξετάζοντας τις τιμές που προκύπτουν στις υπερπαραμέτρους των μεθόδων. Τέλος, στο έκτο κεφάλαιο ερμηνεύονται αυτά τα αποτελέσματα συζητώντας τις επιπτώσεις τους, τη σημασία τους, ενώ παράλληλα προτείνονται δρόμοι για μελλοντική έρευνα.

Κεφάλαιο 2

Συναφείς Εργασίες

2.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται ήδη υπάρχουσες μελέτες πάνω στην πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου τόσο με τη βοήθεια αλγορίθμων μηχανικής μάθησης όσο και με τη βοήθεια νευρωνικών δικτύων.

2.2 Εργασίες αλγορίθμων μηχανικής μάθησης

Στο επιστημονικό άρθρο [1], σκοπός των S. Hu και M. Fu, είναι να προβλέψουν τα αποτελέσματα ποδοσφαιρικών αγώνων της Premier League και της La Liga. Το μοντέλο χρειάζεται να είναι σε θέση να προβλέπει τους παρόντες και τους μελλοντικούς αγώνες. Κατά την επιλογή των δεδομένων, χρησιμοποίησαν δεδομένα αγώνων της Premier League και της La Liga για το χρονικό διάστημα 2018-2022. Προκειμένου το μοντέλο να πετύχει καλύτερα αποτελέσματα, πέρα των κοινότυπων χαρακτηριστικών ενός αγώνα δημιούργησαν δύο νέα χαρακτηριστικά, τη διαφορά στο σκορ ημιχρόνου και στο τελικό σκορ. Εξετάστηκαν τρεις αλγόριθμοι, συμπεριλαμβανομένου των Linear Regression (LRG), Gradient Boosting Decision Tree (GBDT), Random Forest (RF). Οι μέθοδοι αξιολόγησης που επέλεξαν ως μετρικές του συστήματος είναι το $R^2 - score$ και η ακρίβεια. Ως πιο αποτελεσματικό μοντέλο προέκυψε το RF.

Συνεχίζοντας με την βιβλιογραφική έρευνα, ενδιαφέρονσα είναι αυτή των Ishan Jawade, Rushikesh Jadhav, Mark Joseph Vaz και Vaishnavi Yamgekar, [2] στην οποία χρησιμοποιούνται διάφοροι αλγόριθμοι μηχανικής μάθησης για να προβλεφθούν τα αποτελέσματα της Pre-

mier League για νίκη εντός/εκτός έδρας ή ισοπαλία. Το σύνολο δεδομένων που επιλέχθηκε αφορά την σεζόν 2017-2018. Στην έρευνα χρησιμοποιήθηκαν χαρακτηριστικά που προσδιορίζουν την δυναμική των ομάδων σε άμυνα και επίθεση, η φόρμα κάθε ομάδας και το πλεονέκτημα που παρέχει η έδρα σε κάθε ομάδα. Υλοποιήθηκαν πέντε αλγόριθμοι, οι Support Vector Machine (SVM), Random Forest (RF), Multinomial Naive Bayes (MNB), Logistic Regression (LR) , K-Nearest Neighbours (KNN) και τα αποτελέσματά ταξινομήθηκαν σε 3 κλάσεις (Νίκη, Ισοπαλία και Ήττα). Από τα πειράματα διαπιστώθηκε ότι οι SVM, RF και MNB δεν δίνουν καλά αποτελέσματα ενώ μεταξύ των LR και KNN αποδείχθηκε ότι ο KNN είναι καλύτερος για πρόβλεψη, δίνοντας μάλιστα ακρίβεια 73,33% .

Πολύ καλή και αξιόλογη προσέγγιση έγινε και από τους Rahual Baboota, Harleen Kaur το 2019 [3] πάλι με δεδομένα από την Premier League. Τα δεδομένα που χρησιμοποίησαν καλύπτουν 11 αγωνιστικές περιόδους (2005-2016). Μέσω feature engineering, δημιουργήθηκε ένα σύνολο χαρακτηριστικών το οποίο χωρίστηκε σε δύο ευρείες κατηγορίες, την κατηγορία A που περιέχει τα ατομικά χαρακτηριστικά τόσο για τις γηπεδούχες όσο και για τις φιλοξενούμενες ομάδες και την κατηγορία B που περιέχει τα διαφορικά χαρακτηριστικά. Δοκιμάστηκαν τα μοντέλα Gaussian Naive Bayes (GaussianNB), SVM, RF και GBDT ρυθμίζοντας κάθε μοντέλο για βέλτιστη απόδοση και συντονισμό υπερπαραμέτρων. Μετά την υλοποίηση του πειράματος διαπιστώθηκε ότι το μοντέλο με τις καλύτερες επιδόσεις ήταν το μοντέλο GBDT, ακολουθούμενο από το RF, στη συνέχεια τα δύο μοντέλα SVM με πυρήνα RBF και linear kernel και τελευταίο το μοντέλο GaussianNB.

Το 2022 ο Wenbo YU [4] χρησιμοποίησε τα μοντέλα Logistic Regression (LR), SVM, XGBoost, βελτιωμένο XGBoost με δεδομένα συνολικά 3 σεζόν (2016-2019) από το πρωτάθλημα της Premier League για να επιτευχθεί η πρόβλεψη των αποτελεσμάτων των αγώνων σε δύο κλάσεις (Νίκη γηπεδούχου, Μη νίκη γηπεδούχου). Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων ήταν το F1-score και η ακρίβεια. Ύστερα από την εφαρμογή κάθε μοντέλου με χαρακτηριστικά το Goal Difference (Goal Difference = Goals Scored – Goals Conceded) κάθε ομάδας, τα ratings κάθε ομάδας, τους βαθμούς κάθε ομάδας και την φόρμα κάθε ομάδας στα τελευταία 3 παιχνίδια το μοντέλο LR εμφάνισε τη υψηλότερη επίδοση και στις δύο μετρικές. Αντιθέτως, την χαμηλότερη επίδοση εμφάνισε το βελτιωμένο XGBoost.

Ένα χρόνο νωρίτερα, το 2021 οι Yash Ajgaonkar, Prof. Anagha Patil, Kunal Bhoyar, Jenil Shah [5] απέδειξαν πως στο δικό τους σύστημα το SVM ήταν το μοντέλο που έδειξε τις καλύ-

τερες προβλέψεις. Συγκεκριμένα, το πρωτάθλημα που μελετήθηκε ήταν η Premier League. Τα δεδομένα που χρησιμοποιήσαν αφορούσαν τις σεζόν 2011-2019. Από τα 65 αρχικά χαρακτηριστικά ανά σεζόν επιλέχθηκαν για το πείραμα: η γηπεδούχος ομάδα, η φιλοξενούμενη ομάδα, τα σουτ στο στόχο της γηπεδούχου ομάδας, τα σουτ στο στόχο της φιλοξενούμενης ομάδας, τα κόρνερ της γηπεδούχου ομάδας, τα κόρνερ της φιλοξενούμενης ομάδας και τέσσερα χαρακτηριστικά που φανερώνουν την δυναμική των ομάδων. Εφαρμόστηκαν τρεις αλγόριθμοι, οι Naive Bayes (NB), RF και SVM. Τόσο χωρίς συντονισμό των υπερπαραμέτρων όσο και με συντονισμό των υπερπαραμέτρων το μοντέλο SVM κατέγραψε τις καλύτερες επιδόσεις.

2.3 Εργασίες νευρωνικών δικτύων

Στο άρθρο [6], οι S. Anfilets, S. Bezobrazov, V. Golovko, A. Sachenko, M. Komar ερεύνησαν την προσέγγιση τους για την πρόβλεψη των αποτελεσμάτων σε αγώνες ποδοσφαίρου, η οποία βασίζεται σε ένα Deep Multilayer Νευρωνικό Δίκτυο με elastic net κανονικοποίησης. Χρησιμοποιήθηκαν οι τελευταίοι 5, 10, 15, 35 αγώνες και υπολογίστηκαν για κάθε μία από αυτές τις περιπτώσεις οι μέσοι όροι των: σουτ, σουτ στο στόχο, κόρνερ, φάουλ, καρτών, γκολ (υπέρ-κατά) ημιχρόνου και ολόκληρης διάρκειας τόσο για την γηπεδούχο όσο και για την φιλοξενούμενη ομάδα. Επιπλέον υπολογίστηκε η τυπική απόκλιση των παραπάνω χαρακτηριστικών για τους τελευταίους 35 αγώνες. Η ακρίβεια της πρόβλεψης του συγκεκριμένου συστήματος κρίθηκε ικανοποιητική.

Κλείνοντας στην σχετική έρευνα της βιβλιογραφίας ενδιαφέρουσα ήταν και η έρευνα των S. Mohammad Arabzad, M.E. Tayebi Araghi, S.Sadi-Nezhad, Nooshin Ghofrani [7] που χρησιμοποίησαν Τεχνητά Νευρωνικά Δίκτυα για την πρόβλεψη του ακριβούς σκορ στους αγώνες της τελευταίας αγωνιστικής εβδομάδας για την επαγγελματική κατηγορία του Ιράν. Στο άρθρο, χρησιμοποιείται ένα MLP νευρωνικό δίκτυο για το οποίο χρησιμοποιήθηκαν ως δεδομένα εισόδου: η γηπεδούχος ομάδα, η φιλοξενούμενη ομάδα, η κατάσταση της γηπεδούχου και της φιλοξενούμενης στα τελευταία 4 παιχνίδια της, η βαθμολογία τους στο πρωτάθλημα, η ποιότητα των αντιπάλων που αντιμετώπισαν στα τελευταία 4 παιχνίδια, η αγωνιστική εβδομάδα και ο αριθμός του αγώνα. Τα αποτελέσματα που εξήχθησαν έδειξαν ότι το μοντέλο είχε αδυναμία στην πρόβλεψη ακριβούς σκορ σε περιπτώσεις ισοπαλιών αλλά ικανοποιητικό ποσοστό σε περιπτώσεις αγώνων με νικητή.

Κεφάλαιο 3

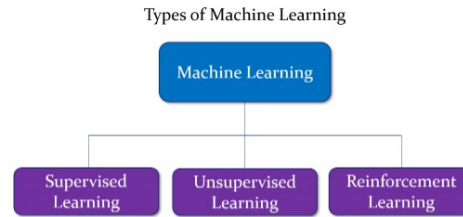
Θεωρητικό Υπόβαθρο

3.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένας κλαδος της Τεχνητής Νοημοσύνης που διερευνά την κατασκευή ικανών για προβλέψεις αλγορίθμων μέσα από αναγνώριση μοτίβων στο σύνολο δεδομένων. Στη σφαίρα αυτή αναδύονται τρεις θεμελιώδεις κατηγορίες (σχήμα 3.1), καθεμία από τις οποίες προσφέρει δικές της μεθοδολογίες και στόχους.

- **Μάθηση υπό επίβλεψη (Supervised Learning):** Περιλαμβάνει την εκπαίδευση μοντέλων σε σύνολα δεδομένων με ετικέτες, όπου τα δεδομένα εισόδου αντιστοιχίζονται με αντίστοιχες ετικέτες εξόδου. Αυτή η προσέγγιση, που περιλαμβάνει εργασίες όπως η ταξινόμηση και η παλινδρόμηση, αποσκοπεί στην ανάπτυξη σχέσεων μεταξύ των μεταβλητών εισόδου και των προβλέψεων της εξόδου.
- **Μάθηση χωρίς επίβλεψη (Unsupervised Learning):** Λειτουργεί χωρίς δεδομένα με ετικέτες, εστιάζοντας στην αναγνώριση μοτίβων ή δομών σε μη επισημειωμένα σύνολα δεδομένων. Η ομαδοποίηση και η μείωση της διαστατικότητας είναι κάποιες από τις τεχνικές που διευκολύνουν την εξαγωγή σημαντικών πληροφοριών από τα δεδομένα.
- **Ενισχυτική μάθηση (Reinforcement Learning):** Εισάγει ένα διαδραστικό στοιχείο, όπου οι πράκτορες μαθαίνουν να καθοδηγούνται σε περιβάλλοντα μέσω δοκιμής και σφάλματος, μεγιστοποιώντας την εύρεση της καλύτερης προσέγγισης.

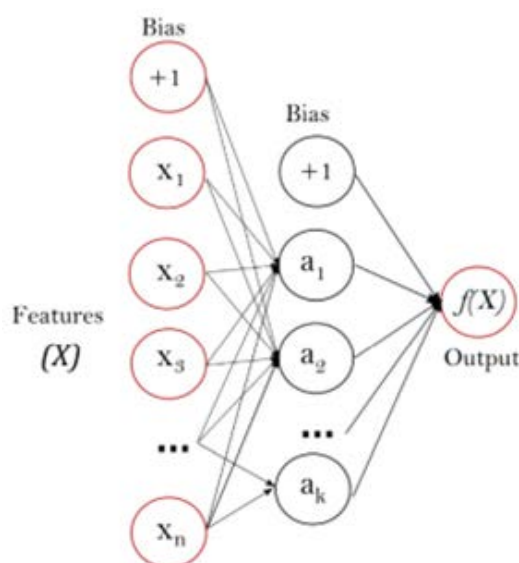
Στην προσπάθειά μας να προβλέψουμε τα αποτελέσματα ποδοσφαιρικών αγώνων, στην παρούσα έρευνα χρησιμοποιείται η επιβλεπόμενη μάθηση.



Σχήμα 3.1: Κατηγορίες Μηχανικής Μάθησης

3.2 Τεχνητά Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα, γνωστά και ως Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ), είναι συστήματα που βασίζονται σε μια συλλογή από κόμβους (νευρώνες) που μοντελοποιούν σε αλγοριθμικό επίπεδο τις συνδέσεις μεταξύ των νευρώνων στον ανθρώπινο εγκέφαλο. Κάθε νευρώνας μπορεί να λάβει ένα σήμα από νευρώνες και να το μεταβιβάσει σε άλλους νευρώνες. Δύο νευρώνες συνδέονται με μια ακμή στην οποία έχει εκχωρηθεί ένα βάρος, το οποίο μοντελοποιεί τη σημασία της εισόδου αυτού του νευρώνα για την έξοδο του άλλου νευρώνα. Ένα νευρωνικό δίκτυο αποτελείται συνήθως από ένα επίπεδο εισόδου, με έναν νευρώνα ανά μεταβλητή εισόδου για το μοντέλο, ένα στρώμα εξόδου, αποτελούμενο από έναν μόνο νευρώνα που θα δώσει το αποτέλεσμα ταξινόμησης ή παλινδρόμησης, και έναν αριθμό κρυφών στρωμάτων μεταξύ των δύο, που περιέχουν μεταβλητό αριθμό νευρώνων σε κάθε στρώμα. Στο σχήμα 3.2, απεικονίζεται η αρχιτεκτονική ενός νευρωνικού δικτύου με ένα κρυφό στρώμα.



Σχήμα 3.2: Τεχνητό Νευρωνικό Δίκτυο

3.3 Αλγόριθμοι και Δίκτυα

3.3.1 Εισαγωγή στις Μέθόδους Μηχανικής Μάθησης

Ο πυρήνας του κάθε μοντέλου είναι η μέθοδος MM που επιλέγει να χρησιμοποιήσει. Υπάρχουν πολλαπλές και διαφορετικές μέθοδοι για κάθε τύπο και κάθε προσέγγιση. Ωστόσο, η επιλογή της καταλληλότερης δεν είναι πάντα τόσο εύκολη. Παρόλο που ένα πρόβλημα μπορεί να φαίνεται πως ταιριάζει σε μία συγκεκριμένη μέθοδο ενδέχεται να υπάρχουν και άλλοι τύποι προσεγγίσεων που λειτουργούν εξίσου καλά. Οι μέθοδοι MM μπορεί να αποτελούνται από ένα σύνολο κανόνων, μια δενδρική δομή ή έναν τύπο δικτύου [8].

3.3.2 Gaussian Naive Bayes

Ένας από τους πιο δημοφιλείς πιθανοτικούς ταξινομητές είναι ο ταξινομητής Naive Bayes. Πρόκειται για μια πιθανοτική μέθοδο ταξινόμησης που βασίζεται στην "αυθαίρετη" υπόθεση ότι κάθε δύο διαφορετικά χαρακτηριστικά είναι ανεξάρτητα, επιτρέποντας έτσι τη χρήση του θεωρήματος Bayes:

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

Όπου:

- y είναι η κλάση-στόχος
- x_i είναι οι μεταβλητές πρόβλεψης (εισόδου)

Χρησιμοποιώντας την υπόθεση ανεξαρτησίας, αυτό οδηγεί στο εξής:

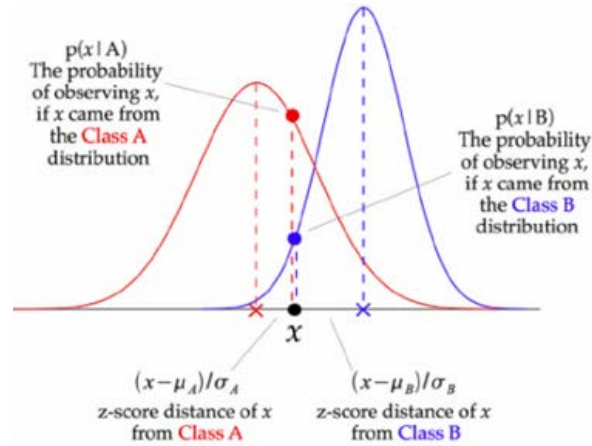
$$P(y | x_1, \dots, x_n) \propto P(y, x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y)$$

Για το Gaussian Naive Bayes, η πιθανότητα $P(x_i|y)$ των χαρακτηριστικών ακολουθεί κατανομή Gauss (σχήμα 3.3) που περιγράφεται από την παρακάτω εξίσωση:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}}$$

Όπου:

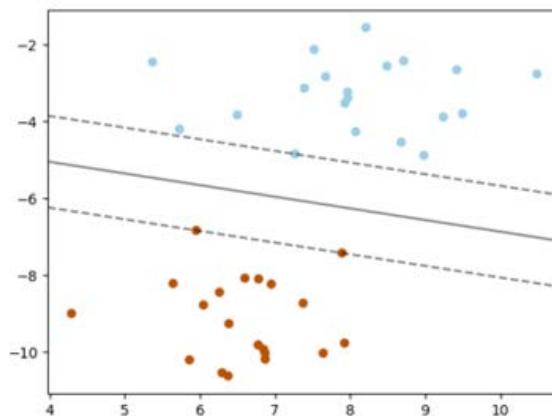
- μ_y είναι η μέση τιμή της μεταβλητής εξόδου
- σ_y είναι η τυπική απόκλιση της μεταβλητής εξόδου



Σχήμα 3.3: Μέθοδος Gaussian Naive Bayes

3.3.3 Support Vector Machine

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι μοντέλα μηχανικής μάθησης τόσο για ταξινόμηση όσο και για παλινδρόμηση. Ένα μοντέλο SVM αναπαριστά τα δεδομένα εκπαίδευσης ως σημεία στο χώρο, έτσι ώστε τα παραδείγματα που εμπίπτουν σε διαφορετικές κατηγορίες να χωρίζονται από ένα υπερεπίπεδο (σχήμα 3.4) που απέχει όσο το δυνατόν περισσότερο από το πλησιέστερο σημείο δεδομένων.



Σχήμα 3.4: SVM Υπερεπίπεδο για ταξινόμηση

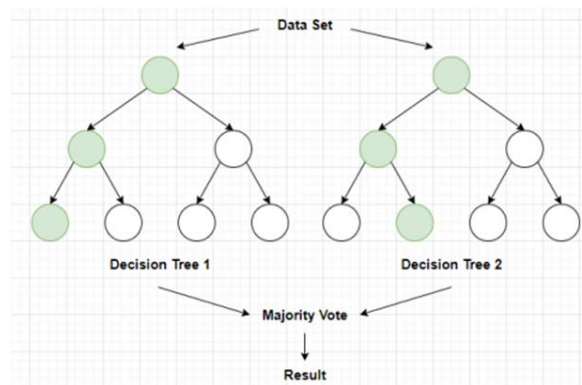
Οι νέες εισοδοι απεικονίζονται με τον ίδιο τρόπο όπως τα δεδομένα εκπαίδευσης και ταξινομούνται ως προς την κατηγορία στην οποία ανήκουν (πλευρά του υπερεπιπέδου). Όταν

τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, μπορεί να χρησιμοποιηθεί το τέχνασμα του πυρήνα, με τη χρήση διαφορετικών πιθανών συναρτήσεων πυρήνα, όπως συναρτήσεις ακτινικής βάσης (Radial Basis Functions - RBF) ή πολυωνυμικές συναρτήσεις (Polynomial Functions - Poly), προκειμένου να απεικονιστούν τα δεδομένα σε χώρους χαρακτηριστικών υψηλής διάστασης και να βρεθεί ένα κατάλληλο υπερεπίπεδο υψηλής διάστασης.

Το παραπάνω πρόβλημα ταξινόμησης μπορεί να επεκταθεί στην επίλυση προβλημάτων παλινδρόμησης με παρόμοιο τρόπο, εξαρτώμενο μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης για τη δημιουργία μιας πρόβλεψης παλινδρόμησης.

3.3.4 Random Forest

Ο Random Forest είναι ένας αλγόριθμος ταξινόμησης που αποτελείται από δέντρα αποφάσεων. Χρησιμοποιεί πολλαπλά δέντρα αντί για μόνο ένα ως αλγόριθμος δέντρων απόφασης. Ένα τυχαίο δάσος εκπαιδεύει τα δέντρα με τυχαία υποσύνολα των αρχικού συνόλου δεδομένων και το αποτέλεσμα καθορίζεται με ψηφοφορία πλειοψηφίας μεταξύ των δέντρων. Αυτό καθιστά τα τυχαία δάση πιο γενικευμένα από τα μεμονωμένα δέντρα απόφασης [9]. Τα τυχαία δάση μπορούν επίσης να χρησιμοποιηθούν για παλινδρόμηση. Τότε η αριθμητική πρόβλεψη γίνεται με τη λήψη του μέσου όρου της εξόδου των δέντρων. Ένα παράδειγμα ενός τυχαίου δάσους και των δέντρων απόφασης που περιέχει απεικονίζεται στο σχήμα 3.5.



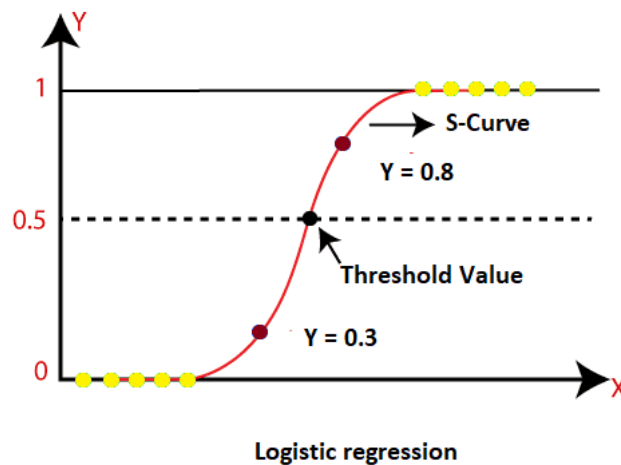
Σχήμα 3.5: Δομή Random Forest

3.3.5 Logistic Regression

Η λογιστική παλινδρόμηση είναι ένας αλγόριθμος με επίβλεψη που σχετίζεται με τη γραμμική παλινδρόμηση. Η πιο σημαντική διαφορά μεταξύ τους, είναι ότι η γραμμική παλινδρόμηση είναι ένας αλγόριθμος παλινδρόμησης, ενώ η λογιστική παλινδρόμηση παρά το

όνομά της είναι αλγόριθμος ταξινόμησης. Η λογιστική παλινδρόμηση προβλέπει διακριτές εξαρτημένες μεταβλητές ενώ η γραμμική παλινδρόμηση προβλέπει συνεχείς εξαρτημένες μεταβλητές. Η λογιστική παλινδρόμηση χρησιμοποιείται συνήθως ως δυαδικός αλγόριθμος ταξινόμησης και επομένως η έξοδος μπορεί να είναι μία από δύο κλάσεις, όπως αληθής ή ψευδής και ναι ή όχι. Η τιμή εξόδου βρίσκεται με τη χρήση πιθανοτήτων και καθορίζει σε ποια από τις δύο κλάσεις ανήκει το σημείο δεδομένων. Οι πιθανότητες υπολογίζονται χρησιμοποιώντας μια σιγμοειδή συνάρτηση, η οποία είναι μια συνάρτηση ενεργοποίησης. Πολλαπλασιάζει κάθε χαρακτηριστικό με ένα βάρος και στη συνέχεια τα προσθέτει. Τα σημεία δεδομένων που είναι χαμηλότερα από μια τιμή κατωφλίου, ταξινομούνται σε μία από τις κλάσεις και τα σημεία δεδομένων που είναι υψηλότερα από την τιμή κατωφλίου, ταξινομούνται στην άλλη κλάση (σχήμα 3.6). Η εξίσωση της λογιστικής παλινδρόμησης μπορεί να εκφραστεί όπως στην παρακάτω εξίσωση, όπου το y ισούται με την πιθανότητα εξόδου, και z είναι μια παράμετρος του μοντέλου που περιλαμβάνει την τιμή x , το μέσο της καμπύλης και την κλίμακα της καμπύλης. Η προκύπτουσα πιθανότητα θα είναι πάντα στο εύρος μεταξύ 0 και 1, το οποίο απαιτείται σε μια σιγμοειδή συνάρτηση.

$$y = \frac{1}{1 + e^{-z}}$$

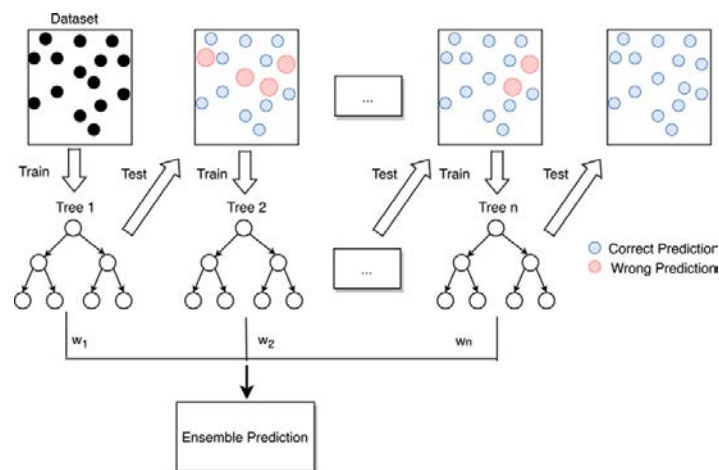


Σχήμα 3.6: Logistic Regression

3.3.6 Gradient Boosting Decision Tree

Ο Gradient Boosting Decision Tree (GBDT) είναι μια ισχυρή τεχνική που συνδυάζει την προγνωστική δύναμη των δέντρων απόφασης με τη μέθοδο συνόλου ενίσχυσης. Σε αντίθεση

με τα παραδοσιακά δέντρα αποφάσεων που κάνουν προβλέψεις ανεξάρτητα, το GBDT δημιουργεί διαδοχικά μια σειρά από δέντρα αποφάσεων, καθένα από τα οποία διορθώνει τα σφάλματα που έχουν γίνει από τα προηγούμενα (σχήμα 3.7). Αυτή η επαναληπτική διαδικασία επιτρέπει στο GBDT να βελτιώνει σταδιακά την ακρίβεια της πρόβλεψης εστιάζοντας στις περιπτώσεις όπου τα προηγούμενα δέντρα είχαν κακή απόδοση. Στον πυρήνα του, το GBDT λειτουργεί ελαχιστοποιώντας μια συνάρτηση απωλειών, όπως το μέσο τετραγωνικό σφάλμα για παλινδρόμηση ή η εντροπία για ταξινόμηση, μέσω μιας διαδικασίας βελτιστοποίησης βαθμωτής καθόδου. Κατά τη διάρκεια της εκπαίδευσης, κάθε επόμενο δέντρο απόφασης κατασκευάζεται για την ελαχιστοποίηση των υπολειπόμενων σφαλμάτων των προηγούμενων δέντρων. Αυτό σημαίνει ότι κάθε νέο δέντρο εκπαιδεύεται για να προβλέπει τα υπολείμματα που απομένουν από τις συνδυασμένες προβλέψεις του υπάρχοντος συνόλου. Προσθέτοντας επαναληπτικά δέντρα και προσαρμόζοντας τις προβλέψεις τους, το GBDT μαθαίνει αποτελεσματικά πολύπλοκα μοτίβα και σχέσεις μέσα στα δεδομένα, παράγοντας τελικά ένα ισχυρό μοντέλο πρόβλεψης.

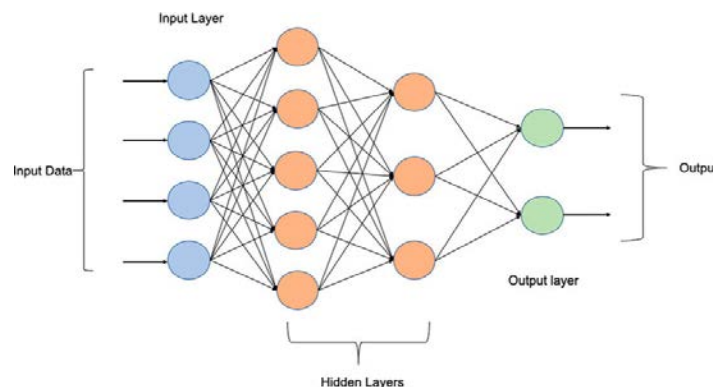


Σχήμα 3.7: Διάγραμμα Ροής GBDT

3.3.7 Multilayer Perceptron

Το MultiLayer Perceptron (MLP) είναι ένας τύπος τεχνητού νευρωνικού δικτύου, που αποτελείται και αυτό από διασυνδεδεμένα στρώματα κόμβων (νευρώνων). Αυτά τα στρώματα περιλαμβάνουν το στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου, και διευκολύνουν την ικανότητα του δικτύου να επεξεργάζεται πληροφορίες. Κάθε κόμβος εντός ενός στρώματος συνδέεται πλήρως με κάθε κόμβο στο επόμενο στρώμα, με

συνδέσεις που έχουν βάρος το οποίο αντικατοπτρίζει τη σημασία τους. Επίσης, αξιοποιούν συναρτήσεις ενεργοποίησης εντός των κρυφών στρωμάτων για να εισάγουν το στοιχείο της μη γραμμικότητας, επιτρέποντας στο δίκτυο να μαθαίνει περίπλοκα μοτίβα μέσα στα δεδομένα. Έτσι, μπορούν και λειτουργούν ως ισχυρά υπολογιστικά μοντέλα βασισμένα στην ιδέα της λειτουργίας των βιολογικών νευρώνων. Κατά τη λειτουργία τους, τα MLP λαμβάνουν ακατέργαστα δεδομένα εισόδου μέσω του στρώματος εισόδου, τα οποία στη συνέχεια επεξεργάζονται μέσω των κρυφών στρωμάτων. Εδώ, οι συνδέσεις με τα βάρη και οι συναρτήσεις ενεργοποίησης χειρίζονται τα δεδομένα εισόδου, επιτρέποντας στο δίκτυο να εξάγει σύνθετα χαρακτηριστικά και σχέσεις μεταξύ τους. Το στρώμα εξόδου συνθέτει αυτές τις επεξεργασμένες πληροφορίες για να παράγει προβλέψεις ή ταξινομήσεις με βάση τα πρότυπα που έχει μάθει. Τα MLP είναι ιδιαίτερα ικανά στο χειρισμό συνόλων που απαιτούν τη μοντελοποίηση μη γραμμικών σχέσεων, όπως η παλινδρόμηση και η ταξινόμηση, και υπερέχουν κυρίως σε εργασίες αναγνώρισης προτύπων. Μία εικόνα αρχιτεκτονικής MLP φαίνεται στο σχήμα 3.8.



Σχήμα 3.8: Αρχιτεκτονική MLP

3.4 Προεπεξεργασία & Μηχανική Χαρακτηριστικών

Η προεπεξεργασία δεδομένων και η μηχανική χαρακτηριστικών είναι κρίσιμα βήματα στην υλοποίηση της μηχανικής μάθησης, που επηρεάζουν σημαντικά την απόδοση και την αποτελεσματικότητα των μοντέλων πρόβλεψης. Οι διαδικασίες αυτές είναι απαραίτητες για τη μετατροπή των ακατέργαστων δεδομένων σε μορφή κατάλληλη έτσι ώστε να εκπαιδευτούν οι αλγόριθμοι μηχανικής μάθησης και να μπορέσουν να εξαχθούν ουσιαστικά συμπεράσματα.

Ένας από τους πρωταρχικούς λόγους για την προεπεξεργασία δεδομένων είναι η αντιμε-

τώπιση προβλημάτων, όπως τιμές που λείπουν και ασυνέπειες στα δεδομένα. Οι τιμές που λείπουν μπορούν να εισάγουν υποκειμενικότητα και να οδηγήσουν σε ανακριβείς προβλέψεις εάν δεν αντιμετωπιστούν σωστά, ενώ οι ασυνέπειες μπορούν να διαστρεβλώσουν τη διαδικασία μάθησης του μοντέλου. Οι τεχνικές προεπεξεργασίας δεδομένων, όπως ο υπολογισμός, η ανίχνευση ακραίων τιμών και η κανονικοποίηση των δεδομένων, βοηθούν στον περιορισμό αυτών των προβλημάτων και διασφαλίζουν ότι τα δεδομένα είναι καθαρά και αξιόπιστα για ανάλυση.

Η μηχανική χαρακτηριστικών, από την άλλη πλευρά, περιλαμβάνει τη δημιουργία νέων χαρακτηριστικών ή τον μετασχηματισμό των υπάρχοντων για την ενίσχυση της απόδοσης του μοντέλου. Η διαδικασία αυτή αποσκοπεί στην εξαγωγή σχετικών πληροφοριών, που μπορεί να μην είναι άμεσα εμφανείς. Για παράδειγμα, στο πλαίσιο της πρόβλεψης ποδοσφαιρικών αγώνων, η μηχανική χαρακτηριστικών μπορεί να περιλαμβάνει τη δημιουργία νέων μεταβλητών, όπως δείκτες απόδοσης παικτών και δείκτες φόρμας ομάδων ή δείκτες σημασίας αγώνων με βάση ιστορικά δεδομένα. Αυτά τα νέα χαρακτηριστικά παρέχουν στο μοντέλο πρόσθετες πληροφορίες για την καλύτερη δημιουργία σχέσεων στα δεδομένα. Επιπλέον, παίζει καθοριστικό ρόλο στη μείωση της διαστατικότητας, κάτι που βοηθάει σημαντικά την βελτίωση της απόδοσης του μοντέλου. Με την επιλογή ή τη δημιουργία κάποιων πιο σχετικών χαρακτηριστικών, επιτρέπεται στο μοντέλο να εστιάζει στις πτυχές των δεδομένων που δίνουν περισσότερη πληροφορία, απορρίπτοντας παράλληλα τις πληροφορίες που δεν προσφέρουν στο μοντέλο.

Οι μέθοδοι επιλογής χαρακτηριστικών κατηγοριοποιούνται σε διάφορες κατηγορίες. Οι βασικές κατηγορίες είναι οι μέθοδοι φίλτρου, wrapper και ενσωμάτωσης. Οι μέθοδοι φίλτρου βασίζονται σε στατιστικά τεστ για την επιλογή των σημαντικών χαρακτηριστικών ανεξάρτητα από το μοντέλο πρόβλεψης. Οι μέθοδοι wrapper χρησιμοποιούν το ίδιο το μοντέλο για να αξιολογήσουν την απόδοση του σε διαφορετικά υποσύνολα χαρακτηριστικών. Οι μέθοδοι ενσωμάτωσης συνδυάζουν τη διαδικασία επιλογής χαρακτηριστικών με την εκπαίδευση του μοντέλου και χρησιμοποιούνται συχνά με αλγόριθμους που έχουν ενσωματωμένους μηχανισμούς επιλογής χαρακτηριστικών όπως τα δέντρα απόφασης.

Μία από τις μεθόδους επιλογής χαρακτηριστικών είναι η μέθοδος του αμοιβαίου πληροφοριακού φίλτρου (Mutual Information). Αυτή η μέθοδος βασίζεται στη θεωρία της πληροφορίας και μετρά την εξάρτηση μεταξύ δύο μεταβλητών. Στην περίπτωση της επιλογής χαρακτηριστικών, η μέθοδος αξιολογεί πόση πληροφορία για την πρόβλεψη της εξαρτημένης

μεταβλητής (output) παρέχει κάθε χαρακτηριστικό (input). Χρησιμοποιείται ευρέως λόγω της απλότητας και της αποτελεσματικότητάς της στην ανίχνευση μη γραμμικών σχέσεων μεταξύ των μεταβλητών.

Μία άλλη σημαντική μέθοδος επιλογής χαρακτηριστικών είναι η εξαντλητική μέθοδος περιτύλιξης (Exhaustive Feature Selection). Σε αυτή τη μέθοδο, δοκιμάζονται όλα τα πιθανά υποσύνολα χαρακτηριστικών για να βρεθεί ο συνδυασμός που προσφέρει την καλύτερη απόδοση στο μοντέλο πρόβλεψης. Παρά το ότι είναι πολύ χρονοβόρα και υπολογιστικά απαιτητική, η μέθοδος αυτή διασφαλίζει ότι θα εντοπιστεί ο βέλτιστος συνδυασμός χαρακτηριστικών, καθιστώντας την κατάλληλη για περιπτώσεις όπου η ακρίβεια του μοντέλου είναι κρίσιμη.

Τέλος, υπάρχει και η μέθοδος Forward Feature Selection. Αυτή η μέθοδος περιλαμβάνει την εκκίνηση με μηδέν χαρακτηριστικά και την προσθήκη του ενός χαρακτηριστικού τη φορά, που βελτιώνει περισσότερο την απόδοση του μοντέλου. Η διαδικασία συνεχίζεται μέχρι να μην υπάρχει περαιτέρω βελτίωση στην απόδοση του μοντέλου. Η μέθοδος αυτή επιτρέπει την αναγνώριση των πιο σημαντικών χαρακτηριστικών και μπορεί να οδηγήσει σε πιο απλά και πιο κατανοητά μοντέλα.

3.5 Θεωρητική Ανάλυση Συντονισμού Υπερπαραμέτρων

Κάθε μοντέλο MM έχει κάποιες υπερπαραμέτρους (HPs) ο συντονισμός των οποίων είναι απαραίτητος για να λειτουργήσει ένα μοντέλο με τον καλύτερο δυνατό τρόπο. Η έννοια των HPs είναι διαφορετική από τις παραμέτρους. Οι HPs είναι οι παράμετροι που χρειάζονται αρχικοποίηση πριν από την εκπαίδευση του μοντέλου, καθώς αντιπροσωπεύουν την αρχιτεκτονική του μοντέλου. Αντίθετα, οι παράμετροι μπορούν να αρχικοποιηθούν και να ενημερωθούν κατά τη διάρκεια της εκπαίδευσης του μοντέλου [10]. Η αυτόματη ρύθμιση των HPs είναι μία από τις κύριες αρμοδιότητες στην αυτοματοποιημένη MM (AutoML) για να μπορέσει να μειώσει το βάρος των χειροκίνητων εργασιών και να βελτιώσει την απόδοση, την επεκτασιμότητα και την εγκυρότητα των διαφόρων ερευνών. Υπάρχουν διάφορες τεχνικές για τα προβλήματα που αφορούν τον συντονισμό υπερπαραμέτρων. Οι πιο κοινές και συμβατικές μέθοδοι είναι η χειροκίνητη αναζήτηση ή Grad Student Descent (GSD), η GridSearch (GS) και η RandomizedSearch (RS). Κάθε μία από αυτές ορίζεται ως εξής:

Grad Student Descent: Η GSD είναι η πιο βασική μέθοδος ρύθμισης των υπερπαρα-

μέτρων και μια τυπική προσέγγιση μεταξύ ερευνητών [11]. Σε αυτή τη μέθοδο, οι χρήστες δοκιμάζουν διαφορετικές τιμές υπερπαραμέτρων με βάση την εκτίμηση ή τη γνώση του πεδίου και επαναλαμβάνουν αυτή τη διαδικασία μέχρι να επιτύχουν ένα ικανοποιητικά βελτιωμένο αποτέλεσμα ή μέχρι να εξαντλήσουν το χρόνο τους. Η GSD χρειάζεται βαθιά κατανόηση των μοντέλων MM ή αρκετό χρόνο για να επιτευχθούν καλά αποτελέσματα. Ωστόσο, η πολύπλοκη φύση των μοντέλων MM και οι μεγάλοι χώροι αναζήτησης την καθιστούν μη πρακτική προσέγγιση.

Grid Search: Η GS είναι ο τρόπος «ωμής» αναζήτησης HP [12], με καθορισμένο το κατώτατο και το ανώτατο όριο, μαζί με συγκεκριμένα βήματα [13]. Η εργασία GS βασίζεται στο καρτεσιανό γινόμενο του διαφορετικού συνόλου τιμών, αξιολογεί κάθε διαμόρφωση και επιστρέφει το συνδυασμό με την καλύτερη απόδοση. Η GS έχει απλή υλοποίηση, ωστόσο, μπορεί να είναι πολύ αναποτελεσματική για μεγάλους χώρους αναζήτησης λόγω της εξαντλητικής φύσης του. Το πρόβλημα αυτό επιδεινώνεται όσο αυξάνονται οι διαστάσεις των δεδομένων.

Randomized Search: Η RS είναι μια άλλη κοινή και τυπική μέθοδος αναζήτησης υπερπαραμέτρων [14]. Η RS επιλέγει τις διαμορφώσεις των υπερπαραμέτρων σε τυχαία βάση (αντί να αξιολογεί κάθε διαμόρφωση) και επαναλαμβάνει αυτή τη διαδικασία έως τη στιγμή που οι προκαθορισμένοι πόροι τελειώσουν. Η RS είναι πολύ ταχύτερη από την GS, αλλά δεν ακολουθεί μια διαδρομή για την εύρεση της βέλτιστης διαμόρφωσης. Επιπλέον, η RS μπορεί να επιταχύνει την εξερεύνηση του χώρου των υπερπαραμέτρων μέσω της εκμετάλλευσης τυχαίων δειγμάτων, επιτρέποντας την εντοπισμό της βέλτιστης υπερπαραμετροποίησης με λιγότερους πόρους.

3.6 Μετρικές Απόδοσης

Η ανάλυση μετρικών επιδόσεων είναι το πιο σημαντικό στην έρευνα μεταξύ των μοντέλων MM, παρέχοντας πληροφορίες σχετικά με το πόσο καλά ένα μοντέλο γενικεύει σε δεδομένα που δεν έχει ξαναδεί. Μεταξύ διάφορων μετρικών απόδοσης οι πιο σημαντικές είναι το accuracy, το precision, το recall και το F1-score που η καθεμία προσφέρει με τον τρόπο της στην απόδοση του μοντέλου (πίνακας 3.1).

Accuracy (ακρίβεια): Ίσως η πιο σημαντική μετρική, μετρά το ποσοστό των σωστά προβλέψιμων περιπτώσεων επί του συνολικού αριθμού των περιπτώσεων. Παρέχει μια γενική

επισκόπηση του πόσο καλά αποδίδει το μοντέλο σε όλες τις κλάσεις. Ωστόσο, η ακρίβεια από μόνη της μπορεί να μην είναι επαρκής, ιδίως σε μη ισορροπημένα σύνολα δεδομένων όπου οι κλάσεις δεν αντιπροσωπεύονται εξίσου. Για παράδειγμα, σε ένα σύνολο δεδομένων όπου μια κλάση είναι κυρίαρχη, ένας ταξινομητής που προβλέπει την κυρίαρχη κλάση για όλες τις περιπτώσεις, θα μπορούσε να επιτύχει υψηλή ακρίβεια χωρίς να αποτυπώνει αποτελεσματικά τα πρότυπα των υπολοίπων κλάσεων.

Precision: Αποτελεί μια πολύ σημαντική μετρική που μετράει το ποσοστό των πραγματικών θετικών προβλέψεων σε σχέση με τις συνολικές θετικές προβλέψεις του μοντέλου. Αναπαριστά, συνήθως, το ποσοστό των σωστών προβλέψεων για μια συγκεκριμένη κλάση σε σχέση με το συνολικό πλήθος των περιπτώσεων που ο αλγόριθμος έχει ταξινομήσει σε αυτήν την κλάση.

Recall (ανάκληση): Μετρά το ποσοστό των πραγματικών θετικών προβλέψεων σε σχέση με το συνολικό πλήθος των πραγματικών θετικών περιπτώσεων σε ένα σύνολο δεδομένων. Είναι ένας σημαντικός δείκτης ευαισθησίας του μοντέλου στην αναγνώριση θετικών περιπτώσεων και θα μπορούσε να είναι ιδιαίτερα σημαντικός σε περιπτώσεις όπου οι ψευδείς αρνητικές προβλέψεις έχουν σημαντικές επιπτώσεις.

F1-score: Λαμβάνει υπόψη τόσο το precision όσο και το recall, παρέχοντας μια ισορροπία στην αξιολόγηση της απόδοσης για κάθε ταξινομητή, ιδίως σε σενάρια που οι κατανομές κλάσεων είναι άνισες. Το F1-score είναι ο αρμονικός μέσος όρος του precision και του recall, εξισορροπώντας έτσι, τη μεταβίβαση μεταξύ ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων. Κυμαίνεται συνήθως από 0 έως 1, με το 1 να υποδηλώνει τέλει precision και recall και το 0 να υποδηλώνει πολύ κακή απόδοση.

Κατά την ανάλυση αυτών των μετρικών, είναι σημαντικό να λαμβάνονται υπόψη τα ειδικά χαρακτηριστικά του συνόλου δεδομένων και οι στόχοι της εκάστοτε έρευνας. Για παράδειγμα, στην ιατρική διάγνωση, όπου τα ψευδώς αρνητικά αποτελέσματα μπορεί να έχουν σοβαρές συνέπειες, η μεγιστοποίηση του recall μπορεί να είναι πιο κρίσιμη από το precision. Αντίθετα, στην ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, το precision μπορεί να έχει προτεραιότητα για να αποφευχθεί η εσφαλμένη ταξινόμηση νόμιμων μηνυμάτων ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητα.

Επιπλέον, οι οπτικοποιήσεις, όπως τα confusion matrices (πίνακες σύγχυσης), μπορούν να παρέχουν μια λεπτομερή ανάλυση της απόδοσης του μοντέλου σε διάφορες κλάσεις, επισημαίνοντας περιοχές για βελτίωση. Στην περίπτωση ταξινόμησης αποτελεσμάτων αγώνων

σε 3 κλάσεις (νίκη γηπεδούχου, ισοπαλία, νίκη φιλοξενούμενης) η γενική μορφή του πίνακα σύγκρισης απεικονίζεται στο σχήμα 3.9, με κάθε τετράγωνο να υποδηλώνει:

Predicted \ Actual	Predicted		
	Predicted Away Wins	Predicted Draws	Predicted Home Wins
Actual Away Wins	T1	T2	T3
Actual Draws	T4	T5	T6
Actual Home Wins	T7	T8	T9

Σχήμα 3.9: Γενική Μορφή Πίνακα Σύγκρισης 3 Κλάσεων

- **T1:** Νίκη φιλοξενούμενης ομάδας, προβλέφθηκε ως νίκη φιλοξενούμενης ομάδας.
- **T2:** Νίκη φιλοξενούμενης ομάδας, προβλέφθηκε ως ισοπαλία.
- **T3:** Νίκη φιλοξενούμενης ομάδας, προβλέφθηκε ως νίκη γηπεδούχου ομάδας.
- **T4:** Ισοπαλία, προβλέφθηκε ως νίκη φιλοξενούμενης ομάδας.
- **T5:** Ισοπαλία, προβλέφθηκε ως ισοπαλία.
- **T6:** Ισοπαλία, προβλέφθηκε ως νίκη γηπεδούχου ομάδας.
- **T7:** Νίκη γηπεδούχου ομάδας, προβλέφθηκε ως νίκη φιλοξενούμενης ομάδας.
- **T8:** Νίκη γηπεδούχου ομάδας, προβλέφθηκε ως ισοπαλία.
- **T9:** Νίκη γηπεδούχου ομάδας, προβλέφθηκε ως νίκη γηπεδούχου ομάδας.

Μετρική	Ορισμός	Τύπος(π.χ Ισοπαλίες)
Accuracy	ποσοστό των σωστά προβλέψιμων περιπτώσεων επί του συνολικού αριθμού των περιπτώσεων	$\frac{T1+T5+T9}{\sum_{i=1}^9 T_i}$
Precision	ποσοστό των πραγματικών θετικών προβλέψεων σε σχέση με τις συνολικές θετικές προβλέψεις	$\frac{T5}{T2+T5+T8}$
Recall	ποσοστό των πραγματικών θετικών προβλέψεων σε σχέση με το συνολικό πλήθος των πραγματικών θετικών περιπτώσεων	$\frac{T5}{T4+T5+T6}$
F1-score	αρμονικός μέσος όρος του precision και του recall	$\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Πίνακας 3.1: Πίνακας Μετρικών Απόδοσης

Κλείνοντας, η αποτελεσματική ερμηνεία αυτών των μετρικών επιτρέπει τη λήψη τεκμηριωμένων αποφάσεων για την αξιολόγηση και τη βελτίωση του μοντέλου, οδηγώντας τελικά σε βελτιωμένη απόδοση και αξιοπιστία σε πραγματικές εφαρμογές.

Κεφάλαιο 4

Εργαλεία, Δεδομένα και Υπερπαράμετροι

4.1 Python

Η Python χρησιμεύει ως το εργαλείο για τον χειρισμό δεδομένων, την κατασκευή μοντέλων και την ανάπτυξη εφαρμογών στο πλαίσιο αυτής της έρευνας. Ξεκινώντας το 1989 υπό την καθοδήγηση του Ολλανδού προγραμματιστή Guido van Rossum και κυκλοφορώντας επίσημα το 1991, η Python ξεχωρίζει ως μια ευέλικτη, υψηλού επιπέδου, αντικειμενοστραφής γλώσσα προγραμματισμού. Η ευρεία υιοθέτησή της μπορεί να αποδοθεί στο φιλικό προς το χρήστη συντακτικό της, στην εκτεταμένη τυποποιημένη βιβλιοθήκη της και στο εύρος πακέτων τρίτων κατασκευαστών. Το αντικειμενοστραφές πλαίσιο της Python προωθεί την επεκτασιμότητα του κώδικα, ενώ το πλούσιο εύρος βιβλιοθηκών δίνει τη δυνατότητα στους προγραμματιστές να χρησιμοποιούν εργαλεία για διάφορους τομείς, από την ανάλυση δεδομένων έως την ανάπτυξη ιστοσελίδων. Επιπλέον, η υποστήριξη της Python για αυτοματοποίηση, μέσω χαρακτηριστικών όπως η κατανόηση λίστας και οι ενσωματωμένες συναρτήσεις, διευκολύνει την ταχεία δημιουργία πρωτοτύπων και τον πειραματισμό. Καθ' όλη τη διάρκεια αυτής της μελέτης, η χρήση διαφόρων βιβλιοθηκών και πακέτων Python υπογραμμίζει τη σημασία της Python ως θεμελιώδες εργαλείο στη σύγχρονη επιστήμη δεδομένων, τη μηχανική μάθηση και τις προσπάθειες ανάπτυξης λογισμικού.

4.2 Jupyter Notebook

Το Jupyter Notebook χρησιμεύει ως το κύριο εργαλείο για τη διεξαγωγή της ανάλυσης δεδομένων, τον πειραματισμό μοντέλων και την οπτικοποίηση των αποτελεσμάτων στην πα-

ρούσα μελέτη. Το Jupyter Notebook παρέχει μια διαδραστική πλατφόρμα που ενσωματώνει απρόσκοπτα τον κώδικα, τις οπτικοποιήσεις και το επεξηγηματικό κείμενο σε ένα ενιαίο αρχείο, προωθώντας τις συνεργατικές ροές εργασίας της έρευνας. Η προέλευσή του μπορεί να εντοπιστεί στο έργο IPython που ξεκίνησε από τον Fernando Pérez το 2001, ενώ το Jupyter παρουσιάστηκε επίσημα το 2014 ως εξέλιξη του IPython. Η πλατφόρμα υποστηρίζει πολλές γλώσσες προγραμματισμού, συμπεριλαμβανομένων των Python και R, καθιστώντας την ευέλικτη για διάφορες εργασίες. Η διεπαφή του Jupyter επιτρέπει στους ερευνητές να εξερευνούν επαναληπτικά δεδομένα, να δημιουργούν πρωτότυπα μοντέλα και να τεκμηριώνουν τα ευρήματα, ενισχύοντας τη διαφάνεια και την κατανόηση. Καθ' όλη τη διάρκεια αυτής της μελέτης, το Jupyter Notebook χρησιμεύει ως ένα εργαλείο για την απρόσκοπτη ενσωμάτωση της ανάλυσης δεδομένων και της ανάπτυξης μοντέλων, δίνοντας στους ερευνητές τη δυνατότητα να διερευνούν και να επικοινωνούν αποτελεσματικά τα συμπεράσματα.

4.3 Βιβλιοθήκες

Σε αυτό το πείραμα χρησιμοποιήθηκαν διάφορες βιβλιοθήκες για χειρισμό, οπτικοποίηση και ανάλυση των δεδομένων:

- **pandas**: Η pandas παρέχει ισχυρές δομές δεδομένων και συναρτήσεις για εργασία με δομημένα δεδομένα, όπως τα σύνολα δεδομένων.
- **datetime**: Χρησιμοποιείται για το χειρισμό λειτουργιών που σχετίζονται με την ημερομηνία και την ώρα, συμπεριλαμβανομένης της ανάλυσης συμβολοσειρών ημερομηνίας και της εκτέλεσης αριθμητικών πράξεων σε ημερομηνίες.
- **sklearn**: Παρέχει εργαλεία για ΜΜ στην Python και είναι μια από τις πιο δημοφιλείς βιβλιοθήκες για την ανάπτυξη μοντέλων ΜΜ. Περιλαμβάνει αλγορίθμους για ταξινόμηση, παλινδρόμηση, clustering, εκτίμηση, επιλογή χαρακτηριστικών κ.α.
- **seaborn**: Μια βιβλιοθήκη οπτικοποίησης στατιστικών δεδομένων που βασίζεται στην matplotlib, το seaborn διευκολύνει τη δημιουργία στατιστικών γραφικών.
- **matplotlib.pyplot**: Η matplotlib είναι μια ευρέως χρησιμοποιούμενη βιβλιοθήκη γραφικών παραστάσεων για τη δημιουργία στατικών, διαδραστικών και κινούμενων απεικονίσεων στην Python.

- **numpy**: Απαραίτητη για τους αριθμητικούς υπολογισμούς στην Python, η numpy παρέχει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και πίνακες, μαζί με μια συλλογή μαθηματικών συναρτήσεων για τη λειτουργία αυτών των πινάκων.
- **warnings**: Χρησιμοποιείται για την ένδειξη συγκεκριμένων προειδοποιητικών μηνυμάτων κατά τη διάρκεια της εκτέλεσης, οι προειδοποιήσεις βοηθούν στη διατήρηση της ομαλής λειτουργίας του κώδικα.

4.4 Σύνολα Δεδομένων

Ξεκινώντας από τα δεδομένα, αυτά αντλήθηκαν από διάφορες ιστοσελίδες, καθώς για πολλά από αυτά έγινε και χειροκίνητη προσθήκη. Χρησιμοποιήθηκαν συνολικά δεδομένα από έξι αγωνιστικές περιόδους από το 2017 έως το 2023 για αγώνες της Premier League. Για την δημιουργία του συνόλου δεδομένων, αρχικά προήλθαν δεδομένα από το <https://www.football-data.co.uk/> [15], όπου πάρθηκαν η ημερομηνία του αγώνα, η γηπεδούχος ομάδα, η φιλοξενούμενη ομάδα, τα γκολ της γηπεδούχου και τα γκολ της φιλοξενούμενης τόσο σε όλη την διάρκεια του αγώνα όσο και στο ημίχρονο, το τελικό αποτέλεσμα του αγώνα και το αποτέλεσμα ημιχρόνου, ο διαιτητής του αγώνα, τα σουτ, τα σουτ στον στόχο, τα φάουλ, τα κόρνερ, οι κίτρινες και οι κόκκινες κάρτες για κάθε ομάδα. Ύστερα, από τον ιστότοπο <https://fbref.com/en/> [16] πάρθηκαν τα xGoals της γηπεδούχου και της φιλοξενούμενης στον αγώνα καθώς επίσης η εβδομάδα της αγωνιστικής στην οποία ανήκει ο κάθε αγώνας. Αντίστοιχα από το https://github.com/footballdotpy/epl_scraper [17] προήλθαν και άλλα χαρακτηριστικά όπως η κατοχή μπάλας, οι συνολικές πάσες, τα τάκλιν και τα οφσάιντ. Τέλος, από το <https://www.fifaindex.com/> [18] πάρθηκαν τα FIFA Ratings για κάθε ομάδα και από το <https://www.transfermarkt.com/> [19] η χρηματιστηριακή αξία κάθε ομάδας. Η FIFA πρόκειται για την Διεθνή Ομοσπονδία Ποδοσφαίρου και είναι ο επίσημος φορέας από τον οποίο μπορούν να αντληθούν οι δυναμικές των ομάδων.

4.5 Χαρακτηριστικά

Ξεκινώντας, το βασικό και αρχικό μοντέλο περιέχει 4 χαρακτηριστικά, το αναγνωριστικό γηπεδούχου ομάδας, το αναγνωριστικό φιλοξενούμενης ομάδας, τον κωδικό αποτελέσματος

ημιχρόνου και το αναγνωριστικό διαιτητή. Για την υλοποίηση της πτυχιακής παράχθηκαν μέσω feature engineering συνολικά 88 νέα χαρακτηριστικά που κατηγοριοποιούνται ως εξής:

- **1η Κατηγορία:** Μέσοι όροι κατά τη διάρκεια της σεζόν μέχρι και πριν τον συγκεκριμένο αγώνα σε γκολ, αναμενόμενα γκολ, σουτ, σουτ στον στόχο, φάουλ, κόρνερ, κίτρινες και κόκκινες κάρτες, κατοχή, αριθμό πασών, τάκλιν και οφσάιντ. Στην κατηγορία αυτή ανήκουν τα χαρακτηριστικά που ξεκινάνε με Avg και είναι σύνολο 48 χαρακτηριστικά. Οι μέσοι όροι υπολογίζονται κάθε αγωνιστική περίοδο από την αρχή, που σημαίνει πως στην 1η αγωνιστική της κάθε αγωνιστικής περιόδου αυτά τα χαρακτηριστικά ξεκινάνε με τιμή 0.
- **2η Κατηγορία:** FIFA Ratings και χρηματιστηριακή αξία των ομάδων. Στην συγκεκριμένη κατηγορία ανήκουν 10 χαρακτηριστικά. Σε κάθε αγωνιστική περίοδο, κάθε ομάδα ανεξάρτητα του αν είναι γηπεδούχος ή φιλοξενούμενη έχει τις ίδιες τιμές σε rating και χρηματιστηριακή αξία καθ' όλη τη διάρκεια της αγωνιστικής περιόδου.
- **3η Κατηγορία:** Head to Head μέσοι όροι σε γκολ, αναμενόμενα γκολ, σουτ, σουτ στον στόχο, φάουλ, κόρνερ, κίτρινες και κόκκινες κάρτες, κατοχή, αριθμός πασών, τάκλιν και οφσάιντ για όλα τα πιθανά ζευγάρια ομάδων. Στην κατηγορία αυτή ανήκουν τα χαρακτηριστικά που ξεκινάνε με H2H και είναι σύνολο 24 χαρακτηριστικά. Με τον όρο H2H εννοούμε τους μέσους όρους στους μεταξύ τους αγώνες όλες τις προηγούμενες αγωνιστικές περιόδους.
- **4η Κατηγορία:** Φόρμα (κατάσταση) κάθε ομάδας (γηπεδούχου, φιλοξενούμενης) στους 5 τελευταίους αγώνες της. Η κατηγορία αυτή αποτελείται από 2 χαρακτηριστικά.

Ενδεικτικά αναφέρονται κάποια από τα χαρακτηριστικά:

- HomeTeamID -> Αναγνωριστικό γηπεδούχου ομάδας.
- RefereeID -> Αναγνωριστικό διαιτητή.
- HTR_code -> Ο κωδικός για το αποτέλεσμα ημιχρόνου.
- HATT -> Η βαθμολογία για την επίθεση της γηπεδούχου ομάδας.
- AMID -> Η βαθμολογία για το κέντρο της φιλοξενούμενης ομάδας.
- AMarketVal(€) -> Η χρηματιστηριακή αξία της φιλοξενούμενης ομάδας σε ευρώ.

- AvgASA -> Ο μέσος όρος σουτ που επιτρέπει η φιλοξενούμενη ομάδα στους αντιπάλους της.
- AvgHPOSSF(%) -> Ο μέσος όρος κατοχής μπάλας της γηπεδούχου ομάδας.
- AvgAxGF -> Ο μέσος όρος αναμενόμενων γκολ της φιλοξενούμενης ομάδας.
- HTFormPoints -> Οι πόντοι που κέρδισε η γηπεδούχος ομάδα στους τελευταίους 5 αγώνες της.
- H2HAvgHY -> Ο Head to Head μέσος όρος κίτρινων καρτών της γηπεδούχου ομάδας.

4.6 Μοντέλα

Τα μοντέλα που υλοποιήθηκαν επιλέχθηκαν με βάση την βιβλιογραφία που εξετάστηκε εκτενώς στο κεφάλαιο 2 αλλά και λόγω των χαρακτηριστικών που έχουν να προβλέπουν αποτελέσματα σε σύνολα δεδομένων όπως αυτά για τις προβλέψεις αποτελεσμάτων σε αγώνες ποδοσφαίρου. Παρακάτω εξηγούνται και οι λόγοι για τους οποίους οι συγκεκριμένοι αλγόριθμοι θα προσφέρουν στην έρευνα.

- Το **Support Vector Machine (SVM)** είναι γνωστό για την ικανότητά του να χειρίζεται δεδομένα υψηλών διαστάσεων και να ταξινομεί αποτελεσματικά τα δεδομένα σε διαφορετικές κλάσεις.
- Τα **Gradient Boosting Decision Trees (GBDT)** και **Random Forest (RF)** επιλέχθηκαν καθώς οι μέθοδοι ensemble, είναι ιδιαίτερα ικανές στο χειρισμό ετερογενών συνόλων δεδομένων, ενώ παράλληλα μετριάζουν την υπερπροσαρμογή. Στο πλαίσιο των ποδοσφαιρικών αγώνων, όπου το αποτέλεσμα μπορεί να επηρεαστεί από πολυάριθμους παράγοντες, επιλέγονται οι μέθοδοι ensemble για να αξιοποιήσουν τη γνώση πολλαπλών δέντρων απόφασης, με αποτέλεσμα αξιόπιστες προβλέψεις.
- Το **Logistic Regression (LR)** αποτελεί έναν ισχυρό αλγόριθμο για ταξινόμηση, καθώς μοντελοποιεί την πιθανότητα των διακριτών κατηγοριών βάσει γραμμικών συνδυασμών των χαρακτηριστικών. Μέσω αυτού, είναι δυνατή η πρόβλεψη ύπαρξης ενός παρατηρούμενου παράδειγματος σε μία από δύο ή περισσότερες κατηγορίες, καθιστώντας το ένα από τα βασικά εργαλεία στη MM.

- Ο **Gaussian Naive Bayes (GaussianNB)** βασίζεται στις πιθανολογικές αρχές του θεωρήματος Bayes και υποθέτει ότι τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα όπως είδαμε και στο κεφάλαιο 3. Κάτι τέτοιο δεν ισχύει στα σύνολα δεδομένων για πρόβλεψη αποτελεσμάτων σε αγώνες ποδοσφαίρου αλλά θα εξεταστεί στα αποτελέσματα τυχόν επιβεβαίωση/απόρριψη.
- Το **Multilayer Perceptron (MLP)** που είναι ένα τεχνητό νευρωνικό δίκτυο που φημίζεται για την ικανότητά του να συλλαμβάνει περίπλοκες μη γραμμικές σχέσεις στα δεδομένα. Αξιοποιώντας πολλαπλά στρώματα νευρώνων, μπορεί να αποκαλύψει πολύπλοκα μοτίβα και συσχετίσεις που μπορεί να διαφεύγουν από απλούστερα μοντέλα. Σε αυτό το πείραμα, συμπεριλήφθηκε ένα Multilayer Perceptron για να εξετασθούν τα πιθανά οφέλη των προσεγγίσεων βαθιάς μάθησης στην πρόβλεψη ποδοσφαιρικών αγώνων.

4.7 Υπερπαράμετροι κατά τον Συντονισμό

Όπως είδαμε στο προηγούμενο κεφάλαιο ο συντονισμός υπερπαραμέτρων είναι απαραίτητος καθώς είναι ένα κρίσιμο βήμα για τη βελτιστοποίηση της απόδοσης των μοντέλων μηχανικής μάθησης. Σε αυτό το πείραμα, χρησιμοποιήσαμε την αναζήτηση πλέγματος (GridSearch), λόγω της απλότητας και της αποτελεσματικότητάς της στην ολοκληρωμένη εξερεύνηση του χώρου υπερπαραμέτρων. Η αναζήτηση πλέγματος επιλέχθηκε σε σχέση με τις υπόλοιπες αναζητήσεις, επειδή αναζητά εξαντλητικά ένα προκαθορισμένο σύνολο τιμών υπερπαραμέτρων, διασφαλίζοντας ότι κανένας συνδυασμός δεν παραβλέπεται και παρέχοντας πιο ντετερμινιστικά αποτελέσματα, τα οποία είναι απαραίτητα για το πείραμα μας.

Για τον συντονισμό των υπερπαραμέτρων του **GaussianNB**, συντονίζονται οι ακόλουθες υπερπαράμετροι:

- `var_smoothing`, επηρεάζει την προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης.

Για τον **SVM**, οι βασικές υπερπαράμετροι που λήφθηκαν υπόψιν είναι οι ακόλουθες:

- `C`, ελέγχει το trade-off μεταξύ πολυπλοκότητας του μοντέλου και σφάλματος.
- `gamma`, καθορίζει την επιρροή των μεμονωμένων δειγμάτων εκπαίδευσης.
- `kernel`, καθορίζει το σχήμα του ορίου απόφασης.

- `random_state`, ορίζει τον τρόπο που γίνεται από τη γεννήτρια τυχαίων αριθμών η τυχοιοποίηση της διαδικασίας εκπαίδευσης.

Ο **Random Forest** συντονίστηκε για:

- `n_estimators`, καθορίζει τον αριθμό των δέντρων που συμπεριλαμβάνονται στο τελικό μοντέλο. Ένας μεγαλύτερος αριθμός εκτιμητών οδηγεί σε καλύτερη απόδοση, αλλά αυξάνει επίσης τον χρόνο εκπαίδευσης και την πολυπλοκότητα του μοντέλου.
- `criterion`, αναφέρεται στη μέθοδο που χρησιμοποιείται για τη μέτρηση της ποιότητας ενός διαχωρισμού σε κάθε κόμβο. Οι δύο συνήθεις επιλογές είναι το “gini” και η “έντροπία” .
- `max_depth`, ελέγχει το μέγιστο βάθος κάθε δέντρου στο δάσος. Ένα υψηλότερο βάθος μπορεί να οδηγήσει σε πιο σύνθετα μοντέλα που μπορούν να προσαρμοστούν καλύτερα στα δεδομένα εκπαίδευσης, αλλά μπορεί επίσης να οδηγήσει σε υπερπροσαρμογή.
- `min_samples_split`, καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για τη διάσπαση ενός εσωτερικού κόμβου. Μια υψηλότερη τιμή μπορεί να οδηγήσει σε λιγότερες διασπάσεις, προστατεύοντας έτσι από υπερπροσαρμογή.
- `random_state`, όπως αναφέρθηκε παραπάνω.
- `max_features`, καθορίζει τον αριθμό των χαρακτηριστικών που θα λαμβάνονται υπόψη κατά την αναζήτηση του καλύτερου διαχωρισμού σε κάθε κόμβο.

Ο **Logistic Regression** βελτιστοποιήθηκε ρυθμίζοντας τα:

- `penalty`, η ποινή ‘l1’ αναφέρεται στην κανονικοποίηση L1, η οποία προσθέτει την απόλυτη τιμή του μεγέθους των συντελεστών στη συνάρτηση κόστους. Η ποινή ‘l2’ αναφέρεται στην κανονικοποίηση L2, η οποία προσθέτει το τετράγωνο του μεγέθους των συντελεστών στη συνάρτηση κόστους.
- `C`, ελέγχει το αντίστροφο της έντασης της κανονικοποίησης.
- `solver`, καθορίζει τον αλγόριθμο που θα χρησιμοποιηθεί στο πρόβλημα βελτιστοποίησης. Ο ‘liblinear’ είναι κατάλληλος για μικρά σύνολα δεδομένων και υποστηρίζει τόσο

τις ποινές 'l1' όσο και τις ποινές 'l2'. Ο 'lbfgs' είναι ένας πιο ισχυρός επιλύτης που λειτουργεί καλά για πολυωνυμικές απώλειες και είναι καταλληλότερος για μεγαλύτερα σύνολα δεδομένων.

- `max_iter`, καθορίζει τον μέγιστο αριθμό επαναλήψεων που χρειάζεται ο επιλύτης για να συγκλίνει. Η αύξηση του μέγιστου αριθμού επαναλήψεων μπορεί να βελτιώσει τη σύγκλιση, αλλά αυξάνει επίσης τον υπολογιστικό χρόνο.
- `random_state`, όπως αναφέρθηκε παραπάνω.

Για τον **Gradient Boosting Decision Trees** ρυθμίστηκαν λεπτομερώς τα: `n_estimators`, `learning_rate`, `max_depth`, `min_samples_split`, `random_state`, `max_features`. Σχεδόν σε όλες τις υπερπαραμέτρους αναφερθήκαμε εκτενέστερα στο RF, συνεπώς γι' αυτό δεν αναλύθηκαν περαιτέρω. Η μοναδική υπερπαραμέτρος που χρήζει ανάλυσης είναι το `learning_rate`, που ελέγχει την συνεισφορά κάθε δέντρου.

Τέλος, το **Multilayer Perceptron** συντονίστηκε για τις παραμέτρους:

- `hidden_layer_sizes`, αναφέρεται στον αριθμό των νευρώνων σε κάθε κρυφό στρώμα. Παίρνει μια λίστα αντιπροσωπεύοντας τον αριθμό των νευρώνων σε κάθε στρώμα.
- `activation`, καθορίζει τη συνάρτηση ενεργοποίησης για τα κρυφά στρώματα. Το "tanh" αναφέρεται στη συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης, ενώ το "relu" αναφέρεται στη γραμμική μοναδιαία συνάρτηση ενεργοποίησης.
- `solver`, αναλαμβάνει τον αλγόριθμο βελτιστοποίησης που θα χρησιμοποιηθεί. Μπορεί να χρησιμοποιηθεί ο αλγόριθμος βελτιστοποίησης Adam, ο οποίος είναι ένας αλγόριθμος βελτιστοποίησης με προσαρμοστική μάθηση. Το 'sgd' αναφέρεται στη στοχαστική κάθοδο κλίσης.
- `alpha`, αντιπροσωπεύει τον όρο ποινής L2, ο οποίος «τιμωρεί» τα μεγάλα βάρη στο νευρωνικό δίκτυο. Βοηθά στην αποφυγή της υπερβολικής προσαρμογής.
- `learning_rate`, ο τρόπος με τον οποίο ενημερώνεται ο ρυθμός μάθησης κατά τη διάρκεια της εκπαίδευσης. Μπορεί να είναι 'adaptive' και να προσαρμόζεται με βάση το validation.

- `learning_rate_init`, ορίζει τον αρχικό ρυθμό μάθησης. Επίσης ορίζει το μέγεθος βήματος σε κάθε επανάληψη κατά τη διάρκεια της βελτιστοποίησης.
- `max_iter`, ο μέγιστος αριθμός επαναλήψεων για την εκπαίδευση του μοντέλου MLP.
- `random_state`, όπως αναφέρθηκε παραπάνω.
- `early_stopping`, επιτρέπει την πρόωρη διακοπή της εκπαίδευσης όταν η απόδοση στο σύνολο επικύρωσης δεν βελτιώνεται. Εάν οριστεί σε `True`, η εκπαίδευση σταματά όταν η βαθμολογία επικύρωσης δεν βελτιώνεται για διαδοχικές εποχές.

Κεφάλαιο 5

Αποτελέσματα και Συγκρίσεις

5.1 Εισαγωγή

Στην παρούσα αναφορά στο κεφάλαιο 3 εμβαθύνουμε στις διάφορες μεθόδους αλγορίθμων μηχανικής μάθησης και αρχιτεκτονικών νευρωνικών δικτύων, ενώ στο κεφάλαιο 4 έγινε ανάλυση ως προς το πώς δημιουργήθηκε το σύνολο δεδομένων που χρησιμοποιείται και πώς χειριζόμαστε αυτά τα χαρακτηριστικά στην διαδικασία των πειραμάτων. Τώρα, καθώς περνάμε στο κεφάλαιο αυτό, επιχειρείται ανάλυση των αποτελεσμάτων και η εστίασή μας μετατοπίζεται στα απτά αποτελέσματα που προκύπτουν από την εφαρμογή αυτών των μεθοδολογιών σε δεδομένα ποδοσφαίρου του πραγματικού κόσμου. Το κεφάλαιο χωρίζεται σε τρία μέρη. Στο πρώτο μέρος προστίθενται στα αρχικά μοντέλα κατηγορίες χαρακτηριστικών, με στόχο την εύρεση ενός νέου συνόλου χαρακτηριστικών που θα οδηγήσει σε βελτιωμένες επιδόσεις. Στο δεύτερο μέρος, στο νέο σύνολο χαρακτηριστικών γίνονται καινούργια πειράματα αφαιρώντας έναν αριθμό χαρακτηριστικών είτε ελεύθερα (κοιτάζοντας στον πίνακα συσχέτισης ποια παρουσιάζουν μεγάλη συσχέτιση μεταξύ τους) είτε χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών. Για τα καλύτερα μοντέλα που προκύπτουν εξετάζεται η συμπεριφορά τους σε κάθε μία από τις τρεις κλάσεις του προβλήματος. Τέλος, στο τρίτο μέρος επιχειρείται συντονισμός υπερπαραμέτρων για περαιτέρω βελτίωση της ακρίβειας των μοντέλων.

5.2 Προσθήκη Κατηγοριών Χαρακτηριστικών

Τα αρχικά μοντέλα της παρούσας εργασίας αποτελούνται από τέσσερα χαρακτηριστικά, το αναγνωριστικό γηπεδούχου ομάδας, το αναγνωριστικό φιλοξενούμενης ομάδας, τον κωδικό αποτελέσματος ημιχρόνου και το αναγνωριστικό διαιτητή. Αυτά τα τέσσερα χαρακτηριστικά αποτελούν το απλούστερο σύνολο χαρακτηριστικών που μπορεί να οδηγήσει σε πρόβλεψη αποτελέσματος ενός αγώνα. Στον Πίνακα 5.1 παρατίθενται τα αποτελέσματα των μοντέλων για τις μετρικές accuracy και F1-score. Το accuracy αποτελεί βιβλιογραφικά την σημαντικότερη μετρική ενώ το F1-score είναι ουσιαστικά ο μέσος όρος των μετρικών precision και recall. Γι' αυτούς τους λόγους προτιμήθηκαν όσον αφορά στην αξιολόγηση των αποτελεσμάτων.

	Accuracy	F1-score
GaussianNB	0.6162	0.4775
SVM	0.5810	0.5193
RF	0.5702	0.5220
LR	0.6135	0.4427
GBDT	0.6162	0.5480
MLP	0.6135	0.4442

Πίνακας 5.1: Πίνακας Αρχικών Μοντέλων

Καλύτερη επίδοση όπως φαίνεται έχουν τα μοντέλα GaussianNB, GBDT τα οποία έχουν μάλιστα την ίδια ακρίβεια πρόβλεψης. Ωστόσο το GBDT έχοντας μεγαλύτερο F1-score κρίνεται εν τέλει ως καλύτερο μοντέλο. Καλή επίδοση επίσης έχουν τα μοντέλα MLP και LR.

Συνεχίζοντας, στα επόμενα πειράματα προστίθενται στα αρχικά μοντέλα είτε μεμονωμένα είτε συνδυαστικά κατηγορίες χαρακτηριστικών που αναφέρθηκαν στο κεφάλαιο 4. Συνολικά στο πρώτο μέρος γίνονται τα εξής 7 πειράματα:

- **1Α Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 1η κατηγορία χαρακτηριστικών που αφορά τους μέσους όρους των διάφορων στατιστικών δεικτών ανά σεζόν και αποτελείται από 48 χαρακτηριστικά.
- **2Α Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 2η κατηγορία χαρακτηριστικών που αφορά τα FIFA ratings και την χρηματιστηριακή αξία των ομάδων

και αποτελείται από 10 χαρακτηριστικά.

- **3A Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 3η κατηγορία χαρακτηριστικών που αφορά τους Head to Head μέσους όρους και αποτελείται από 24 χαρακτηριστικά.
- **4A Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 4η κατηγορία χαρακτηριστικών που αφορά την φόρμα των ομάδων στους τελευταίους 5 αγώνες τους και αποτελείται από 2 χαρακτηριστικά.
- **5A Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 1η και 2η κατηγορία κατηγορία χαρακτηριστικών.
- **6A Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε την 1η, 2η και 3η κατηγορία χαρακτηριστικών.
- **7A Πείραμα:** Σε καθένα από τα αρχικά μοντέλα προσθέτουμε όλες τις κατηγορίες χαρακτηριστικών.

5.3 Α Φάση: Ανάλυση Αποτελεσμάτων 7 Πειραμάτων

Στους Πίνακες 5.2 και 5.3 παρατίθενται τα αποτελέσματα των μοντέλων για τα επτά πειράματα που περιγράφηκαν στην ενότητα 5.2 ως προς το accuracy και το F1-score αντίστοιχα:

	1A Πείραμα	2A Πείραμα	3A Πείραμα	4A Πείραμα	5A Πείραμα	6A Πείραμα	7A Πείραμα
GaussianNB	0.4342	0.5631	0.5578	0.6078	0.4605	0.4526	0.45
SVM	0.6105	0.6342	0.6184	0.5947	0.6131	0.6342	0.6368
RF	0.6052	0.6026	0.6131	0.6105	0.5921	0.6	0.6
LR	0.6078	0.6210	0.6236	0.6157	0.6263	0.6315	0.6289
GBDT	0.6157	0.6473	0.6052	0.6131	0.6210	0.6342	0.6368
MLP	0.6210	0.5763	0.6052	0.6026	0.5447	0.5868	0.5815

Πίνακας 5.2: Συγκεντρωτικός Πίνακας Ακρίβειας 7 Πειραμάτων

Στο **1A Πείραμα**, συγκρίνοντας τα αποτελέσματα με αυτά των αρχικών μοντέλων παρατηρείται καλύτερη επίδοση σε τρία μοντέλα: το MLP, το SVM και το RF. Ακόμη, βλέπουμε

	1A	2A	3A	4A	5A	6A	7A
	Πείραμα	Πείραμα	Πείραμα	Πείραμα	Πείραμα	Πείραμα	Πείραμα
GaussianNB	0.4318	0.5167	0.4943	0.4711	0.4581	0.4497	0.4469
SVM	0.4693	0.5694	0.5297	0.5152	0.4601	0.4933	0.4893
RF	0.5335	0.5460	0.5642	0.5584	0.5151	0.5233	0.5195
LR	0.4887	0.4570	0.5060	0.4519	0.4904	0.5222	0.5220
GBDT	0.5596	0.5922	0.5569	0.5533	0.5652	0.5636	0.5751
MLP	0.4665	0.4213	0.4415	0.4396	0.4104	0.4531	0.4502

Πίνακας 5.3: Συγκεντρωτικός Πίνακας F1-score 7 Πειραμάτων

ότι στο συγκεκριμένο πείραμα επιτυγχάνεται αύξηση στη μέγιστη τιμή ακρίβειας κατά 0,5% με το μοντέλο MLP, το οποίο καταφέρνει ακρίβεια 62,1%.

Στο **2A Πείραμα**, σε σχέση με τα αρχικά εμφανίζεται αύξηση στα μοντέλα GBDT, SVM, LR και RF. Παρατηρείται δηλαδή αύξηση σε τέσσερα από τα έξι μοντέλα. Επίσης υπάρχει αύξηση στη μέγιστη τιμή ακρίβειας κατά 3,1% την οποία σημειώνει το GBDT φτάνοντας σε νέα ακρίβεια 64,73%.

Στο **3A Πείραμα**, παρατηρήθηκε ως προς την ακρίβεια μικρή αύξηση σε τρία από τα έξι μοντέλα. Συγκεκριμένα, αυξήση παρουσίασαν τα μοντέλα LR, SVM και RF. Συνολικά η μέγιστη τιμή ακρίβειας αυξήθηκε κατά 0,8% την οποία σημείωσε το LR με νέα ακρίβεια 62,36%.

Στο **4A Πείραμα**, παρατηρήθηκε επίσης αύξηση ακρίβειας σε τρία από τα έξι μοντέλα. Εδώ αύξηση σημείωσαν τα μοντέλα LR, RF και SVM. Ωστόσο δεν υπήρξε αύξηση στη μέγιστη τιμή ακρίβειας και συνεπώς το συγκεκριμένο πείραμα κρίνεται ως το χειρότερο από αυτά που έχουν εξεταστεί μέχρι στιγμής.

Στο **5A Πείραμα**, υπήρξε αύξηση ακρίβειας στα μοντέλα LR, GBDT και SVM. Συνολικά η αύξηση στη μέγιστη τιμή ακρίβειας είναι της τάξης του 1% και σημειώθηκε από το μοντέλο LR με νέα τιμή ακρίβειας 62,63%.

Στο **6A Πείραμα**, σημειώθηκε αύξηση ακρίβειας σε τέσσερα από τα έξι μοντέλα. Βελτιωμένη επίδοση παρουσίασαν τα μοντέλα GBDT, SVM, LR, RF ενώ τα GBDT και SVM σημειώνοντας και τα δύο νέα ακρίβεια 63,42% (το μοντέλο GBDT επιτυγχάνει μεγαλύτερο F1-score) αύξησαν την μέγιστη τιμή ακρίβειας κατά 1,8% .

Στο **7A Πείραμα** σημειώθηκε αύξηση ακρίβειας και πάλι σε τέσσερα μοντέλα. Όπως και

στο προηγούμενο πείραμα βελτιωμένη επίδοση παρουσίασαν τα μοντέλα GBDT, SVM, LR και RF ενώ ξανά τα GBDT και SVM σημειώνοντας νέα ακρίβεια 63,68% (το GBDT επιτυγχάνει μεγαλύτερο F1-score) κατάφεραν να επιφέρουν αύξηση στην μέγιστη τιμή ακρίβειας κατά 2% .

Συγκεντρωτικά, συγκρίνοντας τα αρχικά μοντέλα με τα μοντέλα των πειραμάτων (7 περιπτώσεις), αυτό που αναδεικνύεται από την όλη διαδικασία είναι πως το βέλτιστο πείραμα είναι αυτό που προκύπτει με την πρόσθεση όλων των κατηγοριών χαρακτηριστικών (7Α πείραμα). Δηλαδή, το βέλτιστο πείραμα είναι αυτό που έχει συνολικά 88 χαρακτηριστικά, τα 4 που έχουν ήδη τα αρχικά μοντέλα και τα 84 νέα από τις 4 κατηγορίες που προστέθηκαν. Αν και το 2Α πείραμα ήταν αυτό που παρουσίασε τη μεγαλύτερη αύξηση στην μέγιστη τιμή ακρίβειας (της τάξης του 3%) επιλέχθηκε να ασχοληθούμε εκτενέστερα με το 7Α πείραμα καθώς πέτυχε αρκετά ικανοποιητική αύξηση στην μέγιστη τιμή ακρίβειας (της τάξης του 2%), παρουσίασε σε σχέση με το 2Α πείραμα αύξηση ακρίβειας στον ίδιο αριθμό μοντέλων (4 μοντέλα) και επιπρόσθετα προσέφερε σε γενικές γραμμές καλύτερες τιμές στα μοντέλα όπου υπήρξε η αύξηση. Για το 7Α πείραμα τα μοντέλα με υψηλότερη επίδοση είναι το GBDT και το SVM που προβλέπουν με ακρίβεια 63,68%.

Στο βέλτιστο πείραμα τα μοντέλα GBDT και SVM εμφάνισαν την μεγαλύτερη ακρίβεια και αυτό οφείλεται κυρίως στις ιδιότητες και την αποδοτικότητα αυτών των μεθόδων στην επεξεργασία και την ανάλυση πολύπλοκων δεδομένων. Η μέθοδος GBDT λειτουργεί εξαιρετικά καλά με μεγάλα σύνολα δεδομένων και μπορεί να ανιχνεύσει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και των αποτελεσμάτων. Επιπλέον, η μέθοδος SVM είναι ιδιαίτερα αποτελεσματική στη διαχωριστική ανάλυση των κλάσεων, αφού βρίσκει το βέλτιστο επίπεδο που διαχωρίζει τα δεδομένα στον χώρο των χαρακτηριστικών. Τέλος, και οι δύο μέθοδοι είναι σε θέση να αντιμετωπίσουν αποτελεσματικά την υπερπροσαρμογή και άρα γι' αυτό αναδείχθηκαν ως οι καταλληλότερες στο βέλτιστο πείραμα.

Εφόσον βρέθηκαν τα μοντέλα με την υψηλότερη ακρίβεια, στη συνέχεια επιχειρείται μια πιο λεπτομερής ανάλυση με σκοπό να βρεθούν οι επιδόσεις τους σε κάθε μία από τις τρεις κλάσεις του προβλήματος. Γι' αυτό το λόγο χρησιμοποιώντας τους τύπους μετρικών που αναφέρθηκαν στο 3ο κεφάλαιο (αναλογικά για όλες τις κλάσεις) υπολογίζονται για κάθε κλάση τα recall, precision και F1-score (σχήμα 5.1).

Κοιτάζοντας το σχήμα 5.1 από τα αποτελέσματα που προέκυψαν μπορούν να εξαχθούν για καθέ μοντέλο τα εξής συμπεράσματα:

	Recall	Precision	F1-score		Recall	Precision	F1-score
Home Wins	0.858	0.705	0.773	Home Wins	0.788	0.721	0.752
Draws	0.034	0.375	0.061	Draws	0.287	0.409	0.336
Away Wins	0.743	0.547	0.629	Away Wins	0.660	0.610	0.634

(α') Support Vector Machine

(β') Gradient Boosting Decision Trees

Σχήμα 5.1: Πίνακες Μετρικών για τα μοντέλα GBDT και SVM

- **SVM:** Εμφανίζει υψηλή επίδοση στην πρόβλεψη νικών εντός και εκτός έδρας, υποδεικνύοντας ότι όταν γίνεται μια πρόβλεψη για αυτά τα αποτελέσματα, συχνά είναι σωστή. Ωστόσο, η επίδοση του στις ισοπαλίες είναι πολύ χαμηλή (πολύ χαμηλή ανάκληση) που υποδηλώνει αδυναμία στην πρόβλεψη ισοπαλιών.
- **GBDT:** Παρουσιάζει πιο ισορροπημένη επίδοση. Επιτυγχάνει υψηλή επίδοση για την πρόβλεψη νικών εντός έδρας και εκτός έδρας, υποδεικνύοντας μια τάση για ακριβείς προβλέψεις στα αποτελέσματα αυτά. Η επίδοση του στις ισοπαλίες είναι μέτρια αλλά πολύ καλύτερη συγκριτικά με αυτή του SVM.

Συνολικά οι διαφορές που παρατηρούνται στο precision και το recall μεταξύ των δύο μοντέλων μπορούν να αποδοθούν στους αλγορίθμους τους και στον τρόπο που χειρίζονται την πολυπλοκότητα των δεδομένων. Μοντέλα όπως ο SVM συνήθως δίνουν προτεραιότητα στο precision, με στόχο να ελαχιστοποιήσουν τα ψευδή θετικά, ενώ μοντέλα όπως το GBDT δίνουν προτεραιότητα στο recall, με στόχο να ελαχιστοποιήσουν τα ψευδή αρνητικά. Αυτές οι διαφορές τονίζουν τις διαφορές μεταξύ precision και recall και τη σημασία της επιλογής της κατάλληλης μεθόδου. Επίσης, όπως παρατηρείται οι ισοπαλίες αποτελούν γενικώς πρόκληση και για τις δύο μεθόδους, με πολύ χαμηλές έως μέτριες τιμές σε όλες τις μετρικές.

Τελικώς, για το βέλτιστο πείραμα μεταξύ των μοντέλων GBDT και SVM επιλέγεται ως καλύτερο το GBDT λόγω μεγαλύτερου F1-score και καλύτερης επίδοσης στην πρόβλεψη ισοπαλιών.

- **2η Περίπτωση:** Συσχέτιση μεταξύ των μέσων όρων κατοχής μπάλας και των μέσων όρων συνολικών πασών.
- **3η Περίπτωση:** Συσχέτιση μεταξύ των μέσων όρων γκολ, αναμενόμενων γκολ, σουτ, σουτ στο στόχο.
- **4η Περίπτωση:** Συσχέτιση μεταξύ των head to head μέσων όρων γκολ, αναμενόμενων γκολ, σουτ, σουτ στο στόχο.
- **5η Περίπτωση:** Συσχέτιση μεταξύ των head to head μέσων όρων κατοχής μπάλας και των head to head μέσων όρων συνολικών πασών.

Επειδή για τις περισσότερες από τις παραπάνω περιπτώσεις τα χαρακτηριστικά που τις αποτελούν είναι πολλά θα έπρεπε σε κάθε περίπτωση να γίνουν πολλάπια πειράματα. Επιλέχθηκε λοιπόν ανά περίπτωση να γίνονται δύο πειράματα, κρατώντας κάθε φορά ως χαρακτηριστικό αυτό που κατά γενική ομολογία θεωρείται μεγαλύτερης σημασίας.

Στην **1η Περίπτωση**, την πρώτη φορά επιλέχθηκε να κρατηθούν ως χαρακτηριστικό τα rating που φανερώνουν την γενική δυναμική των ομάδων (Home Overall και Away Overall αντίστοιχα), ενώ την δεύτερη φορά κρατήθηκαν ως χαρακτηριστικό οι χρηματιστηριακές αξίες των ομάδων. Καλύτερα αποτελέσματα προέκυψαν κρατώντας τις χρηματιστηριακές αξίες των ομάδων, συνεπώς στους πίνακες 5.6 και 5.7 στην 1η στήλη, πείραμα 1B, παρουσιάζονται τα αποτελέσματα όλων των μοντέλων για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά χωρίς τα FIFA Ratings (σύνολο 80 χαρακτηριστικά).

Στην **2η Περίπτωση**, αρχικά επιλέχθηκε να κρατηθούν ως χαρακτηριστικό οι μέσοι όροι συνολικών πασών και στην συνέχεια οι μέσοι όροι κατοχής μπάλας. Πιο ικανοποιητικά αποτελέσματα προέκυψαν κρατώντας τους μέσους όρους κατοχής μπάλας, οπότε στους πίνακες 5.6 και 5.7 στην 2η στήλη, πείραμα 2B, παρουσιάζονται τα αποτελέσματα όλων των μοντέλων για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά χωρίς τους μέσους όρους συνολικών πασών (σύνολο 84 χαρακτηριστικά).

Στην **3η Περίπτωση**, την πρώτη φορά επιλέχθηκε να κρατηθούν ως χαρακτηριστικό οι μέσοι όροι γκολ και την δεύτερη φορά οι μέσοι όροι σουτ στο στόχο. Ιδανικότερα αποτελέσματα προέκυψαν χρησιμοποιώντας τους μέσους όρους σουτ στο στόχο, γι' αυτό στους πίνακες 5.6 και 5.7 στην 3η στήλη, πείραμα 3B, παρουσιάζονται τα αποτελέσματα όλων των μοντέλων για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά χωρίς τους μέσους όρους γκολ, αναμενόμενων γκολ, σουτ (σύνολο 76 χαρακτηριστικά).

Στην **4η Περίπτωση**, για αρχή επιλέχθηκε να κρατηθούν ως χαρακτηριστικό οι head to head μέσοι όροι γκολ και στην συνέχεια οι head to head μέσοι όροι σουτ στο στοχο. Βέλτιστα αποτελέσματα προέκυψαν με τους head to head μέσους όρους σουτ στο στόχο, συνεπώς στους πίνακες 5.6 και 5.7 στην 4η στήλη, πείραμα 4B, παρουσιάζονται τα αποτελέσματα όλων των μοντέλων για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά χωρίς τους head to head μέσους όρους γκολ, αναμενόμενων γκολ, σουτ (σύνολο 82 χαρακτηριστικά).

Στην **5η Περίπτωση**, την πρώτη φορά επιλέχθηκε να κρατηθούν ως χαρακτηριστικό οι head to head μέσοι όροι συνολικών πασών ενώ την δεύτερη φορά οι head to head μέσοι όροι κατοχής μπάλας. Καλύτερα αποτελέσματα προέκυψαν με τους head to head μέσους όρους κατοχής μπάλας, οπότε στους πίνακες 5.6 και 5.7 στην 5η στήλη, πείραμα 5B, παρουσιάζονται τα αποτελέσματα όλων των μοντέλων για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά χωρίς τους head to head μέσους όρους συνολικών πασών (σύνολο 86 χαρακτηριστικά).

Στη συνέχεια, για να γίνει μία πιο ολοκληρωμένη έρευνα, αποφασίστηκε να εξεταστούν επιπρόσθετα και δύο τεχνικές επιλογής χαρακτηριστικών. Η παρούσα εργασία έχει να κάνει με πρόβλημα ταξινόμησης όπου οι μεταβλητές εισόδου είναι συνδυασμός κατηγορικών και αριθμητικών μεταβλητών ενώ οι μεταβλητές εξόδου είναι κατηγορικές μεταβλητές γι' αυτό και η πρώτη τεχνική που επιλέχθηκε είναι η **Mutual Information**. Ως δεύτερη τεχνική επιλέχθηκε αρχικά η **Exhaustive Feature Selection** λόγω της λεπτομερούς αναζήτησης που πραγματοποιεί, καθώς πριν επιστρέψει το καλύτερο σύνολο χαρακτηριστικών αξιολογεί όλους τους πιθανούς συνδυασμούς. Ωστόσο δεν μπόρεσαν να ολοκληρωθούν τα πειράματα με αυτή την τεχνική επειδή είχε αυξημένες απαιτήσεις σε υπολογιστικούς πόρους οι οποίοι δεν ήταν διαθέσιμοι. Επόμενη τεχνική που επιλέχθηκε να δοκιμάσκει ως δεύτερη ήταν η **Forward Feature Selection**. Σε αυτή την τεχνική εξήχθησαν αποτελέσματα αλλά ο αριθμός επιλεγμένων χαρακτηριστικών περιορίστηκε σε χαμηλές τιμές λόγω των μεγάλων απαιτήσεων της τεχνικής ως προς χρόνο.

Όσον αφορά την **Mutual Information**, αρχικά υπολογίστηκε η τιμή συσχέτισης καθενός από τα 88 χαρακτηριστικά με την μεταβλητή στόχου (όσο πιο μεγάλη η τιμή συσχέτισης τόσο πιο σημαντικό το χαρακτηριστικό) και έπειτα εξετάστηκαν οι αποδόσεις όλων των μοντέλων για τέσσερις περιπτώσεις:

1. Για νεό σύνολο χαρακτηριστικών τα 10 πιο σημαντικά χαρακτηριστικά που προέκυψαν.

2. Για νέο σύνολο χαρακτηριστικών τα 30 πιο σημαντικά χαρακτηριστικά που προέκυψαν.
3. Για νέο σύνολο χαρακτηριστικών τα 50 πιο σημαντικά χαρακτηριστικά που προέκυψαν.
4. Για νέο σύνολο χαρακτηριστικών τα 70 πιο σημαντικά χαρακτηριστικά που προέκυψαν.

Καλύτερα αποτελέσματα προέκυψαν για νέο σύνολο χαρακτηριστικών τα 30 πιο σημαντικά χαρακτηριστικά (πίνακας 5.4) οπότε στους πίνακες 5.6 και 5.7 στην 6η στήλη, πείραμα 6B, παρουσιάζονται τα αποτελέσματα ως προς το accuracy και το F1-score γι' αυτό το σύνολο χαρακτηριστικών.

1. HTR_code	11. AvgAPOSSA(%)	21. AwayTeamID
2. HATT	12. H2HAvgAPOSS(%)	22. H2HAvgHS
3. AATT	13. AvgHSTF	23. HMarketVal(€)
4. AMarketVal(€)	14. AvgASTF	24. AOVR
5. H2HAvgHPOSS(%)	15. H2HAvgAPASS	25. AvgHSTA
6. HOVR	16. AvgASTA	26. AvgAGA
7. HDEF	17. HMID	27. H2HAvgHPASS
8. H2HAvgAG	18. HomeTeamID	28. H2HAvgHST
9. AvgAGF	19. AMID	29. AvgASA
10. H2HAvgAS	20. AvgAxGF	30. ADEF

Πίνακας 5.4: Τα 30 πιο σημαντικά χαρακτηριστικά σύμφωνα με την Mutual Information

Όσον αφορά την **Forward Feature Selection**, εξετάστηκαν τρεις περιπτώσεις:

1. Για σύνολο επιλεγμένων χαρακτηριστικών 5 χαρακτηριστικά.
2. Για σύνολο επιλεγμένων χαρακτηριστικών 10 χαρακτηριστικά.
3. Για σύνολο επιλεγμένων χαρακτηριστικών 15 χαρακτηριστικά.

Καλύτερα αποτελέσματα προέκυψαν για σύνολο επιλεγμένων χαρακτηριστικών 15 χαρακτηριστικά (πίνακας 5.5) οπότε στους πίνακες 5.6 και 5.7 στην 7η στήλη, πείραμα 7B, παρουσιάζονται τα αποτελέσματα ως προς το accuracy και το F1-score για αυτό το σύνολο επιλεγμένων χαρακτηριστικών.

GaussianNB: [HTR_code, ADEF, HDEF, AvgHCA, RefereeID, AvgHYF, AvgAPOSSA(%), AvgAYA, H2HAvgHOFF, HMID, H2HAvgAOFF, H2HAvgAF, H2HAvgHTAC, AvgHYA, AvgHOFFF]
SVM: [HTR_code, HOVR, AvgASTF, AvgAPASSF, AvgAPASSA, AvgAYA, AvgHTACA, AvgAGF, AvgHCA, AvgAFF, H2HAvgAOFF, AvgAPOSSF(%), AATT, HATT, AvgHxGF]
RF: [HTR_code, HOVR, H2HAvgHR, H2HAvgAR, HMID, H2HAvgAPASS, AvgAYF, AvgHFA, H2HAvgAG, HDEF, AMID, AvgAPASSA, HomeTeamID, AvgACF, AvgHRA]
LR: [HTR_code, ADEF, HDEF, AvgAxGF, AvgHYF, AvgHRA, AvgHSTF, HTFormPoints, AvgHGA, AvgHCF, H2HAvgAG, AvgHFA, H2HAvgAPASS, AATT, H2HAvgHY]
GBDT: [HTR_code, HOVR, AvgASTF, HDEF, AOVR, AMID, H2HAvgHG, H2HAvgHR, H2HAvgAPASS, AvgHRA, HMarketVal(€), HTFormPoints, AvgAYA, H2HAvgHS, AvgARA]
MLP: [HTR_code, ADEF, HDEF, AvgASF, AvgAPOSSA(%), AvgAYA, AvgHRA, AvgAFA, H2HAvgAY, AvgAOFFF, AvgHPOSSF(%), HomeTeamID, AvgAYF, AvgACA, AvgHPASSF]

Πίνακας 5.5: Τα 15 επιλεγμένα χαρακτηριστικά σύμφωνα με την Forward Selection

5.5 Β Φάση: Ανάλυση Αποτελεσμάτων μετά την Επίλογή Χαρακτηριστικών

Στους πίνακες 5.6 και 5.7 παρατίθενται τα αποτελέσματα των μοντέλων μετά την επιλογή χαρακτηριστικών (τόσο με βάση τον πίνακα συσχέτισης όσο και με τις τεχνικές επιλογής χαρακτηριστικών) ως προς το accuracy και το F1-score αντίστοιχα. Όπου:

- **1B Πείραμα:** Αναφέρεται στο σύνολο των 88 χαρακτηριστικών χωρίς τα FIFA Ratings.
- **2B Πείραμα:** Αναφέρεται στο σύνολο των 88 χαρακτηριστικών χωρίς τους μέσους όρους συνολικών πασών.
- **3B Πείραμα:** Αναφέρεται στο σύνολο των 88 χαρακτηριστικών χωρίς τους μέσους

	1B Πείραμα	2B Πείραμα	3B Πείραμα	4B Πείραμα	5B Πείραμα	6B Πείραμα	7B Πείραμα
GaussianNB	0.4552	0.4526	0.4447	0.4631	0.4526	0.5263	0.6078
SVM	0.6263	0.6368	0.6236	0.6315	0.6394	0.6289	0.6368
RF	0.6184	0.5947	0.6131	0.6105	0.6131	0.6210	0.6315
LR	0.6289	0.6131	0.6184	0.6236	0.6210	0.6131	0.6473
GBDT	0.6157	0.6473	0.6210	0.6368	0.6315	0.6342	0.6052
MLP	0.6	0.5868	0.5947	0.6078	0.5921	0.6315	0.6026

Πίνακας 5.6: Συγκεντρωτικός Πίνακας Ακρίβειας μετά την Επιλογή Χαρακτηριστικών

	1B Πείραμα	2B Πείραμα	3B Πείραμα	4B Πείραμα	5B Πείραμα	6B Πείραμα	7B Πείραμα
GaussianNB	0.4527	0.4485	0.4432	0.4606	0.4502	0.4975	0.5453
SVM	0.4714	0.4885	0.4804	0.4914	0.4921	0.5	0.5619
RF	0.5415	0.5072	0.5444	0.5379	0.5403	0.5482	0.5719
LR	0.5160	0.4976	0.5056	0.5049	0.5076	0.4655	0.5232
GBDT	0.5261	0.5880	0.5529	0.5720	0.5631	0.5810	0.5315
MLP	0.4558	0.4485	0.4603	0.4872	0.4555	0.5110	0.4418

Πίνακας 5.7: Συγκεντρωτικός Πίνακας F1-score μετά την Επιλογή Χαρακτηριστικών

όρους γκολ, αναμενόμενων γκολ, σουτ.

- **4B Πείραμα:** Αναφέρεται στο σύνολο των 88 χαρακτηριστικών χωρίς τους head to head μέσους όρους γκολ, αναμενόμενων γκολ, σουτ.
- **5B Πείραμα:** Αναφέρεται στο σύνολο των 88 χαρακτηριστικών χωρίς τους head to head μέσους όρους συνολικών πασών.
- **6B Πείραμα:** Αναφέρεται στα 30 πιο σημαντικά χαρακτηριστικά σύμφωνα με την Mutual Information.
- **7B Πείραμα:** Αναφέρεται στα 15 επιλεγμένα χαρακτηριστικά σύμφωνα με την Forward Selection.

Συγκρίνοντας τα αποτελέσματα του πίνακα 5.6 με αυτά του 7Α πειράματος (πίνακας 5.2) μπορούν να εξαχθούν τα εξής συμπεράσματα:

Στο **1B Πείραμα**, αύξηση ακρίβειας συναντάται σε τρία μοντέλα: το RF, το MLP και το GaussianNB. Ωστόσο δεν σημειώνεται αύξηση στην μέγιστη τιμή ακρίβειας οπότε και δεν εμφανίζονται συνολικά καλά αποτελέσματα.

Στο **2B Πείραμα**, παρατηρείται αύξηση ακρίβειας στα μοντέλα GBDT, MLP και GaussianNB. Επιπλέον, υπάρχει αύξηση στην μέγιστη τιμή ακρίβειας κατά 1% την οποία σημειώνει το μοντέλο GBDT φτάνοντας σε ακρίβεια 64,73%. Συνολικά, παράγονται ικανοποιητικά αποτελέσματα.

Στο **3B Πείραμα**, παρατηρείται μικρή αύξηση ως προς την ακρίβεια σε μόνο δύο μοντέλα. Συγκεκριμένα βελτίωση παρουσιάζουν τα μοντέλα RF και MLP. Ακόμη, δεν διαπιστώνεται αύξηση στην μέγιστη τιμή ακρίβειας οπότε το πείραμα θεωρείται εύκολα από τα μέχρι στιγμής χειρότερα.

Στο **4B Πείραμα**, αύξηση εμφανίζεται σε τρία από τα έξι μοντέλα. Εδώ, αύξηση παρουσιάζουν τα μοντέλα RF, MLP και GaussianNB. Ωστόσο, δεν παρατηρείται αύξηση στην μέγιστη τιμή ακρίβειας και γι' αυτό το πείραμα απορρίπτεται.

Στο **5B Πείραμα** σημειώνεται για πρώτη φορά αύξηση σε τέσσερα μοντέλα. Βελτιωμένη επίδοση σημειώνουν τα μοντέλα SVM, RF, MLP και GaussianNB. Στην περίπτωση αυτή διαπιστώνεται αύξηση στη μέγιστη τιμή ακρίβειας κατά 0,25% την οποία σημειώνει το μοντέλο SVM με νέα ακρίβεια 63,94%. Σε γενικό πλαίσιο παρουσιάζονται ικανοποιητικά αποτελέσματα και επομένως το πείραμα συμπεριλαμβάνεται στα υποψήφια βέλτιστα.

Στο **6B Πείραμα** παρατηρείται αύξηση ακρίβειας στα μισά μοντέλα. Συγκεκριμένα, αύξηση παρουσιάζουν τα μοντέλα MLP, RF και GaussianNB. Επίσης, δεν παρατηρείται καμία αύξηση στην μέγιστη τιμή ακριβείας. Συνολικά, το πείραμα δεν μπορεί να συγκαταλεγεί στα υποψήφια βέλτιστα.

Στο **7B Πείραμα** υπάρχει αύξηση ακρίβειας σε τέσσερα από τα έξι μοντέλα που δοκιμάστηκαν. Τα μοντέλα LR, RF, GaussianNB και MLP ήταν αυτά που σημείωσαν πρόοδο ενώ η μέγιστη τιμή ακρίβειας αυξήθηκε κατά 1% χάρη στο LR που πέτυχε ακρίβεια 64,73%. Λόγω των παραπάνω το συγκεκριμένο πείραμα ανήκει στα καλύτερα που εξετάστηκαν.

Συγκεντρωτικά όπως φαίνεται στον πίνακα 5.8, σε τρία από τα παραπάνω πειράματα (2B, 5B και 7B) προέκυψαν νέα σύνολα χαρακτηριστικών που θα μπορούσαν να επιφέρουν περαιτέρω βελτίωση στην μέγιστη τιμή ακρίβειας πετυχαίνοντας παράλληλα καλύτερη επί-

	Πριν	Μετά		
	7A Πείραμα	2B Πείραμα	5B Πείραμα	7B Πείραμα
GaussianNB	0.45	0.4526	0.4526	0.6078
SVM	0.6368	0.6368	0.6394	0.6368
RF	0.6	0.5947	0.6131	0.6315
LR	0.6289	0.6131	0.6210	0.6473
GBDT	0.6368	0.6473	0.6315	0.6052
MLP	0.5815	0.5868	0.5921	0.6026

Πίνακας 5.8: Πίνακας Ακρίβειας Καλύτερων Μοντέλων πριν-μετά την Επιλογή Χαρακτηριστικών

δοση τουλάχιστον στα μίσα μοντέλα. Συγκρίνοντας αυτά τα τρία πειράματα, αναδεικνύεται ως βέλτιστο το πείραμα 5B. Αρχικά, το πείραμα 2B εγκαταλείπεται πρώτο καθώς επιφέρει βελτιώσεις σε λιγότερα μοντέλα συγκριτικά με τα πειράματα 5B και 7B. Στη συνέχεια, καθώς μεταξύ των πειραμάτων 5B και 7B η διαφορά που παρατηρείται στην μέγιστη τιμή ακρίβειας είναι πολύ μικρή επιλέγεται τελικώς το πείραμα 5B. Συνεπώς, βέλτιστο πείραμα είναι το πείραμα που έχει για σύνολο χαρακτηριστικών το σύνολο των 88 χαρακτηριστικών χωρίς τους head to head μέσους όρους συνολικών πασών (σύνολο 86 χαρακτηριστικά). Για το 5B πείραμα, το μοντέλο που παρουσιάζει την καλύτερη επίδοση είναι το SVM με ακρίβεια πρόβλεψης 63,94%.

Στο βέλτιστο πείραμα, η μέθοδος SVM εμφάνισε την καλύτερη επίδοση. Αυτό οφείλεται στην ικανότητά που έχει να αντιμετωπίζει αποτελεσματικά προβλήματα ταξινόμησης κλάσεων. Όπως αναφέρθηκε και σε προηγούμενη υποενότητα η SVM είναι ιδιαίτερα αποτελεσματική σε διαχωρίσιμα δεδομένα και μπορεί να εξάγει ένα επίπεδο που διαχωρίζει αποτελεσματικά τις κλάσεις στον χώρο των χαρακτηριστικών. Επιπλέον, η SVM είναι ανθεκτική στην υπερπροσαρμογή και μπορεί να διαχειριστεί πολύπλοκα δεδομένα, γεγονός που την καθιστά πιο αποδοτική από τις υπόλοιπες.

Όπως στην ενότητα 5.3, εφόσον βρέθηκε το καλύτερο μοντέλο επιχειρείται μια πιο λεπτομερής ανάλυση με σκοπό να βρεθούν οι επιδόσεις του σε κάθε μία από τις τρεις κλάσεις του προβλήματος. Χρησιμοποιούνται και πάλι οι τύποι μετρικών που αναφέρθηκαν στο 3ο κεφάλαιο (αναλογικά για όλες τις κλάσεις) για να υπολογίστουν για κάθε κλάση τα recall, precision και F1-score (σχήμα 5.3).

	Recall	Precision	F1-score
Home Wins	0.858	0.702	0.771
Draws	0.034	0.375	0.061
Away Wins	0.752	0.557	0.639

(α') Support Vector Machine

Σχήμα 5.3: Πίνακες Μετρικών για το μοντέλο SVM

Κοιτάζοντας το σχήμα 5.3 για το μοντέλο SVM μπορούμε να αποφανθούμε ότι έχει πολύ καλή συμπεριφορά στην πρόβλεψη νικών εντός και εκτός έδρας και ότι είναι πολύ κακό στην πρόβλεψη ισοπαλιών με ιδιαίτερα χαμηλή τιμή ανάκλησης.

5.6 Συντονισμός Υπερπαραμέτρων

Σε αυτή τη φάση εφαρμόζεται συντονισμός υπερπαραμέτρων (hyperparameter tuning), μια επαναληπτική διαδικασία που αποσκοπεί στη βελτιστοποίηση της απόδοσης των μεθόδων που επιλέχθηκαν. Είναι πολύ σημαντικό αυτό το βήμα για να προσαρμοστούν τα δεδομένα ώστε να αποφευχθεί η υπερπροσαρμογή και η υποπροσαρμογή. Για κάθε μέθοδο, πραγματοποιήθηκε συντονισμός χρησιμοποιώντας την τεχνική GridSearch. Η τεχνική αυτή είναι μια τεχνική που διερευνά συστηματικά προκαθορισμένους συνδυασμούς τιμών υπερπαραμέτρων, δημιουργώντας ένα πλέγμα δυνατοτήτων. Αξιολογεί εξαντλητικά κάθε συνδυασμό χρησιμοποιώντας διασταυρούμενη επικύρωση, προσδιορίζοντας το σύνολο των υπερπαραμέτρων που βελτιστοποιούν τη μετρική ακρίβειας η οποία είναι αυτή που ενδιαφέρει σε ένα μοντέλο μηχανικής μάθησης. Στο κεφάλαιο 4 έγινε αναλυτική αναφορά στο ποιες είναι οι υπερπαραμέτροι που συντονίζονται για κάθε μέθοδο και τι σημαίνουν. Το εύρος των τιμών που χρησιμοποιήθηκαν προκειμένου να υπολογιστεί η βέλτιστη τιμή κάθε υπερπαραμέτρου στα μοντέλα των δύο συνόλων χαρακτηριστικών που βρέθηκαν στις ενότητες 5.3 και 5.5 απεικονίζονται στον πίνακα 5.9. Στη συνέχεια, για τα μοντέλα αυτών των συνόλων χαρακτηριστικών υπολογίστηκαν εκ νέου το accuracy και το F1-score (πίνακες 5.10 και 5.11).

Για **σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά**, μετά τον συντονισμό υπήρξε αύξηση σε πέντε από τα έξι μοντέλα που χρησιμοποιήθηκαν. Συγκεκριμένα, αύξηση ακρίβειας παρουσίασαν τα μοντέλα RF, LR, GBDT, MLP και GaussianNB. Μικρή μείωση σημείωσε το μοντέλο SVM. Παράλληλα, υπήρξε αύξηση στη μέγιστη τιμή ακρίβειας κατά 1,7% την

GaussianNB Tuning: Range = {var_smoothing = np.logspace (0,-9, num=100)}
SVM Tuning: Range = {C = [1,10,100], gamma = [0.0001,0.01,1, auto, scale], kernel = [linear, rbf, poly, sigmoid], random_state = 0}
RF Tuning: Range = {n_estimators = [60,80,100,120,140], criterion = [gini, entropy], max_depth = [1,2,3,4,5], min_samples_split = [2,3,4], random_state = 0, max_features = None}
LR Tuning: Range = {penalty = [L1, L2], C = [0.0001,0.01,1,100], solver = [liblinear, lbfgs], max_iter = [100,200,300], random_state = 0}
GBDT Tuning: Range = {n_estimators = [60,80,100,120,140], learning_rate = [0.01,0.05,0.1,0.2], max_depth = [1,2,3,4,5], min_samples_split = [2,3,4], random_state = 0, max_features = None}
MLP Tuning: Range = {h_layers = [(128,64,32), (200,100,50), (128,64), (128,64,32,16)], activation = [tanh, relu], solver = [adam, sgd], alpha = [0.0001,0.001,0.005], learning_rate = [adaptive], learning_rate_init = [0.0001,0.001]}

Πίνακας 5.9: Εύρος Τιμών Υπερπαραμέτρων Μοντέλων 88 και 86 Χαρακτηριστικών

	Πριν		Μετά	
	Accuracy	F1-score	Accuracy	F1-score
GaussianNB	0.45	0.4469	0.4847	0.4798
SVM	0.6368	0.4893	0.6347	0.4754
RF	0.6	0.5195	0.6542	0.6074
LR	0.6289	0.5220	0.6331	0.4750
GBDT	0.6368	0.5751	0.6384	0.5750
MLP	0.5815	0.4502	0.6184	0.4942

Πίνακας 5.10: Πίνακας Μετρικών Μοντέλων 88 Χαρακτηριστικών πριν-μετά τον Συντονισμό

οποία σημείωσε το μοντέλο RF φτάνοντας σε ακρίβεια 65,42%. Αξιοσημείωτο είναι ότι εδώ η ακρίβεια ξεπέρασε το 65% κάτι που δεν έχει παρατηρηθεί μέχρι τώρα στα πειράματα που έχουν γίνει. Οι βέλτιστες τιμές ως προς τον συντονισμό των παραμέτρων που υπολογίστηκαν για τα έξι μοντέλα, για αυτό το σύνολο χαρακτηριστικών, παρατίθενται στον πίνακα 5.12.

Για **σύνολο χαρακτηριστικών τα 86 χαρακτηριστικά**, μετά τον συντονισμό υπήρξε αύ-

	Πριν		Μετά	
	Accuracy	F1-score	Accuracy	F1-score
GaussianNB	0.4526	0.4502	0.4863	0.4816
SVM	0.6394	0.4921	0.6347	0.4752
RF	0.6131	0.5403	0.6505	0.5998
LR	0.6210	0.5076	0.6331	0.4750
GBDT	0.6315	0.5631	0.6410	0.5766
MLP	0.5921	0.4555	0.6199	0.5104

Πίνακας 5.11: Πίνακας Μετρικών Μοντέλων 86 Χαρακτηριστικών πριν-μετά τον Συντονισμό

GaussianNB Tuning: Optimal Value = {var_smoothing = 0.002848035868435802}
SVM Tuning: Optimal Value = {C = 10, gamma = 0.0001, kernel = rbf}
RF Tuning: Optimal Value = {n_estimators = 140, criterion = entropy, max_depth = 4, min_samples_split = 2, random_state = 0, max_features = None}
LR Tuning: Optimal Value = {C = 0.01, penalty = L1, solver = liblinear, max_iter = 100, random_state = 0}
GBDT Tuning: Optimal Value = {n_estimators = 100, learning_rate = 0.05, max_depth = 2, min_samples_split = 2, random_state = 0, max_features = None}
MLP Tuning: Optimal Value= {h_layers= (128,64,32,16), activation = tanh, solver = sgd, alpha = 0.0001, learning_rate = adaptive, learning_rate_init = 0.001, max_iter = 50, random_state=0, early_stopping =True}

Πίνακας 5.12: Βέλτιστες Τιμές Υπερπαραμέτρων Μοντέλων 88 Χαρακτηριστικών

ξηση ξανά στα ίδια πέντε μοντέλα. Τα RF, GBDT, LR, MLP, GaussianNB παρουσίασαν την βελτιωμένη επίδοση ενώ πάλι το SVM σημείωσε μικρή μείωση. Παράλληλα, η μέγιστη τιμή ακρίβειας αυξήθηκε κατά 1,1% την οποία πέτυχε το μοντέλο RF φτάνοντας ακρίβεια 65,05%. Εδώ η ακρίβεια δεν έφτασε τα επίπεδα της ακρίβειας που βρέθηκαν για σύνολο χαρακτηριστικών τα 88 χαρακτηριστικά, ωστόσο έφτασε αρκετά κοντά. Οι βέλτιστες τιμές ως προς τον συντονισμό των παραμέτρων που υπολογίστηκαν για τα έξι μοντέλα, για αυτό το σύνολο χαρακτηριστικών, παρατίθενται στον πίνακα 5.13.

Όπως παρατηρήθηκε, ο συντονισμός των υπερπαραμέτρων έφερε σημαντική βελτίωση

GaussianNB Tuning: Optimal Value = {var_smoothing = 0.0533669923120631}
SVM Tuning: Optimal Value = {C = 10, gamma = 0.0001, kernel = rbf}
RF Tuning: Optimal Value = {n_estimators = 80, criterion = entropy, max_depth = 4, min_samples_split = 2, random_state = 0, max_features = None}
LR Tuning: Optimal Value = {C = 0.01, penalty = L1, solver = liblinear, max_iter = 100, random_state = 0}
GBDT Tuning: Optimal Value = {n_estimators = 100, learning_rate = 0.05, max_depth = 2, min_samples_split = 3, random_state = 0, max_features = None}
MLP Tuning: Optimal Value = {h_layers = (128,64,32), activation = tanh, solver = adam, alpha = 0.0001, learning_rate = adaptive, learning_rate_init = 0.001, max_iter = 50, random_state = 0, early_stopping = True}

Πίνακας 5.13: Βέλτιστες Τιμές Υπερπαραμέτρων Μοντέλων 86 Χαρακτηριστικών

στην απόδοση των μοντέλων και στις δύο περιπτώσεις συνόλων χαρακτηριστικών. Η επιλογή βέλτιστων τιμών για τον αριθμό και το βάθος των δέντρων, σε συνδυασμό με τη χρήση του κριτηρίου entropy ήταν αυτά που οδήγησαν σε αύξηση ακρίβειας το μοντέλο RF. Το LR επίσης επωφελήθηκε από τον συντονισμό, με τη βέλτιστη τιμή του C να ορίζεται στο 0,01, τη χρήση της κανονικοποίησης στο L1 και τον επιλυτή σε liblinear υπήρξε μείωση της υπερεκπαίδευσης και συνεπώς βελτίωση της ακρίβειας. Ύστερα, στο μοντέλο GBDT, η επιλογή βέλτιστων τιμών για το βάθος των δέντρων, τον ελάχιστο αριθμό δειγμάτων και το ρυθμό μάθησης έπαιξαν καθοριστικό ρόλο στη βελτίωση που παρουσίασε. Ακόμη, για το MLP η προσαρμογή του στον καλύτερο αριθμό κρυφών επιπέδων και στον καλύτερο επιλυτή, η χρήση συνάρτησης ενεργοποίησης “tanh” και ρυθμού μάθησης ‘adaptive’ επέτρεψαν βελτίωση της απόδοσης του. Τέλος, στο μοντέλο GaussianNB η αλλαγή τιμής διακύμανσης διανομής επέφερε την αύξηση που παρατηρήθηκε. Έτσι, ο συντονισμός των παραμέτρων κατέστησε τα μοντέλα πιο αποτελεσματικά και ακριβή στην πρόβλεψη των αποτελεσμάτων ποδοσφαιρικών αγώνων.

Κεφάλαιο 6

Συμπεράσματα

6.1 Εισαγωγή

Στο τελευταίο κεφάλαιο, γίνεται εκτενής αναφορά στα συμπεράσματα που εξήχθησαν από όλη την διαδικασία που εφαρμόστηκε και επεξηγήθηκε παραπάνω, ενώ στη συνέχεια παρατίθενται προτάσεις για μελλοντικές επεκτάσεις της παρούσας δουλειάς.

6.2 Σύνοψη και συμπεράσματα

Κοιτάζοντας πίσω, προέκυψαν αρκετά και χρήσιμα συμπεράσματα από την έρευνα που έγινε πάνω στην πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων. Αρχικά, η προσεκτική επιλογή χαρακτηριστικών αποδείχθηκε καθοριστική στην μοντελοποίηση πρόβλεψης, με αποτέλεσμα τη σημαντική βελτίωση των επιδόσεων των μοντέλων κατά τη διάρκεια και των δύο φάσεων των πειραμάτων. Συγκεκριμένα, επιτεύχθηκε αύξηση στην ακρίβεια πρόβλεψης των μοντέλων με κατάλληλη επιλογή και ιδανικό μετασχηματισμό των χαρακτηριστικών.

Επιπλέον, η εξέταση σημαντικών μετρικών απόδοσης, όπως το accuracy, το precision, το recall και το F1-score, παρείχε σημαντικές πληροφορίες σχετικά με το πόσο καλά λειτούργησαν τα μοντέλα τόσο συνολικά όσο και σε κάθε μία από τις τρεις κλάσεις ξεχωριστά. Τα μοντέλα SVM και GBDT είχαν πάντα εξαιρετικά καλές επιδόσεις και βρίσκονταν στην κορυφή των αποτελεσμάτων.

Σημαντικά κέρδη στην ακρίβεια των μοντέλων επέφερε ο συντονισμός των υπερπαραμέτρων, ο οποίος αποδείχθηκε εξαιρετικά σημαντικός στη διαδικασία της προεπεξεργασίας. Ανέδειξε πόσο ζωτικής σημασίας είναι η βελτιστοποίηση των παραμέτρων των μοντέλων

προκειμένου να αξιοποιηθεί πλήρως η προβλεπτική τους ικανότητα. Μετά τον συντονισμό παρατηρήθηκε ότι τα μοντέλα RF και GBDT ξεχώρισαν για το εξαιρετικό F1-score και το accuracy τους και ανερίχθηκαν στην κορυφή των επιδόσεων λόγω της ικανότητάς τους να συλλαμβάνουν σύνθετα μοτίβα δεδομένων.

Όπως παρατηρήθηκε καθ'όλη την διάρκεια των πειραμάτων, υπήρχε συνεχώς ένα πρόβλημα με την ακρίβεια του μοντέλου GaussianNB. Με μια απλή ματιά σε όλα τα αποτελέσματα του GaussianNB μπορεί κάλλιστα να βγει το συμπέρασμα πως η επιλογή χρήσης του δεν ήταν ορθή. Κυμαίνεται συνεχώς σε χαμηλές τιμές μεταξύ 45-50%, και πολύ σπάνια σε τιμές ανάλογες του 62-63% που είναι οι τιμές που επιτεύχθηκαν στα περισσότερα μοντέλα. Υπάρχουν αρκετοί λόγοι για τους οποίους δεν υπάρχει καλή ακρίβεια πρόβλεψης. Αρχικά το GaussianNB υποθέτει ότι τα χαρακτηριστικά ακολουθούν γκαουσιανή (κανονική) κατανομή. Ωστόσο τα χαρακτηριστικά που υπάρχουν στο σύνολο δεδομένων της παρούσας εργασίας, δεν ακολουθούν πάντα την κανονική κατανομή κάτι που αποτελεί τον κύριο λόγο για τον οποίο δεν παρατηρείται η βέλτιστη απόδοση. Ένας άλλος λόγος είναι ότι το GaussianNB υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών. Κάτι τέτοιο έρχεται σε πλήρη αντίθεση με τα δεδομένα που υπάρχουν στο σύνολο δεδομένων που χρησιμοποιήθηκε, καθώς τα περισσότερα από αυτά είναι άμεσα συνδεδεμένα με διάφορα άλλα χαρακτηριστικά.

Τέλος, παρατηρείται πως τα αποτελέσματα της εργασίας είναι πολύ ικανοποιητικά, και στις περισσότερες περιπτώσεις ακόμα καλύτερα από αυτά που παρουσιάστηκαν και εξετάστηκαν στην βιβλιογραφία του 2ου κεφαλαίου. Για παράδειγμα όπως είδαμε στο [1], το μοντέλο RF είχε ακρίβεια 63,82% ενώ στην παρούσα εργασία χρησιμοποιώντας το εν λόγω μοντέλο επιτεύχθηκε ακρίβεια της τάξης του 65,42%. Παραλλήλα, βελτιωμένη κατά 5% εμφανίστηκε η ακρίβεια και για τα μοντέλα LR και GBDT. Μόνο στο [4] με την χρήση του μοντέλου LR αλλά και του μοντέλου SVM παρατηρήθηκαν ακρίβειες 67,3% και 66,67% αντίστοιχα, τιμές οι οποίες είναι οριακά καλύτερες από αυτές που υπολογίζονται στην παρούσα εργασία (63,31% και 63,47%, αντίστοιχα). Ωστόσο αυτό μπορεί να αιτιολογηθεί από το γεγονός ότι η πρόβλεψη στο [4] αφορούσε δυαδική ταξινόμηση και όχι ταξινόμηση πολλαπλών κλάσεων. Άλλος λόγος που θα μπορούσε να εξηγήσει την επίτευξη της μεγαλύτερης ακρίβειας είναι η διαφορετική προεπεξεργασία δεδομένων που γίνεται. Συνεπώς, η επιλογή των υπολοίπων πέντε μεθόδων (SVM, LR, RF, GBDT, MLP) ήταν ορθότατη μιας και κατέστη δυνατό να ξεπεραστούν τα αποτελέσματα της βιβλιογραφικής αναφοράς μέσω της διαφορετικής προετοιμασίας του συνόλου δεδομένων αλλά και μέσω του συντονισμού υπερπαραμέ-

τρων που ήταν μια διαδικασία με καρποφόρα αποτελέσματα.

6.3 Μελλοντικές επεκτάσεις

Όσον αφορά το μέλλον, η εργασία αυτή ανέδειξε ενδιαφέρουσες κατευθύνσεις για περαιτέρω έρευνα στον τομέα της πρόβλεψης ποδοσφαιρικών αγώνων. Επεκτείνοντας το βασικό σκέλος που βασίστηκε στη δημιουργία των μοντέλων και την επιλογή των χαρακτηριστικών υπάρχουν πολλές ενδιαφέρουσες κατευθύνσεις που θα μπορούσαν να ακολουθηθούν.

- Η διερεύνηση αυτοματοποιημένων αλγορίθμων επιλογής χαρακτηριστικών θα μπορούσε να δώσει περισσότερες πληροφορίες σχετικά με τη συνεισφορά των χαρακτηριστικών στα αποτελέσματα των αγώνων και να οδηγήσει σε βελτιωμένη απόδοση των μοντέλων.
- Η χρήση εξελιγμένων τρόπων διαχείρισης μη ισορροπημένων δεδομένων, όπως η υπερδειγματοληψία ή οι μέθοδοι συνόλου που έχουν σχεδιαστεί για την επίλυση ανισορροπημένων κλάσεων σε σύνολα πρόβλεψης ποδοσφαίρου, θα μπορούσαν να παραγάγουν μοντέλα που είναι πιο ανθεκτικά και αξιόπιστα.
- Μετρήσεις όπως η ROC-AUC μπορούν να παρέχουν περισσότερες πληροφορίες σχετικά με την αποτελεσματικότητα του μοντέλου, ιδίως όταν πρόκειται για την πρόβλεψη ποδοσφαιρικών αγώνων, όπου οι διαφορετικές μετρικές μπορεί να έχουν διαφορετική βαρύτητα.
- Η εξέταση εναλλακτικών μεθόδων προεπεξεργασίας δεδομένων και δημιουργίας μοντέλων θα μπορούσαν να παρέχουν διορατικές πληροφορίες σχετικά με την ενίσχυση της γενίκευσης και της αξιοπιστίας των μοντέλων.
- Αναλύσεις ευαισθησίας όπως Min-Max scaling θα μπορούσαν να προσφέρουν ιδιαίτερα χρήσιμες πληροφορίες σχετικά με τη σταθερότητα και την αξιοπιστία των μοντέλων.

Συμπεραίνοντας, υπάρχουν περιθώρια βελτίωσης της πρόβλεψης ποδοσφαιρικών αγώνων, τα οποία, ενδεχομένως, θα μπορούσαν να βοηθήσουν στην πρόβλεψη και άλλων συναφών πεδίων.

Βιβλιογραφία

- [1] Sicheng Hu and Min Fu. Football match results predicting by machine learning techniques. In *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, pages 72–76. IEEE, 2022”.
- [2] Ishan Jawade, Rushikesh Jadhav, M Vaz, and Vaishnavi Yamgekar. Predicting football match results using machine learning. *Maharashtra: International Research Journal of Engineering and Technology*, 2021.
- [3] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755, 2019.
- [4] Wenbo YU. Football result prediction based on machine learning. In Chi-Hua Chen, editor, *Applied Mathematics, Modeling and Computer Simulation*, volume 20 of *Advances in Transdisciplinary Engineering Series*, pages 787–792. IOS Press, 2021. doi:10.3233/ATDE220080.
- [5] Yash Ajgaonkar, Kunal Bhoyar, Anagha Patil, and Jenil Shah. Prediction of winning team using machine learning. *International Journal of Engineering Research Technology*, 2021.
- [6] Sergei Anfilets, Sergei Bezobrazov, Vladimir Golovko, Anatoliy Sachenko, Myroslav Komar, Raman Dolny, Valery Kasyanik, Pavlo Bykovyy, Egor Mikhno, and Oleksandr Osolinskyi. Deep multilayer neural network for predicting the winner of football matches. *International Journal of Computing*, 19(1):70–77, 2020.
- [7] S Mohammad Arabzad, Mohamad Ebrahim Tayebi Araghi, Soheil Sadi-Nezhad, and Nooshin Ghofrani. Football match results prediction using artificial neural networks;

- the case of iran pro league. *Journal of Applied Research on Industrial Engineering*, 1(3):159–179, 2014.
- [8] Tom Dietterich. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327, 1995.
- [9] Oswald Campesato. *Artificial intelligence, machine learning, and deep learning*. Mercury Learning and Information, Dulles, Virginia, 1 edition, 2020.
- [10] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Springer, New York, 2013th edition, 2013.
- [11] Steven Abreu. Automated architecture design for deep neural networks. *arXiv preprint arXiv:1908.10714*, 2019.
- [12] Marc Claesen, Jaak Simm, Dusan Popovic, Yves Moreau, and Bart De Moor. Easy hyperparameter search using optunity. *arXiv preprint arXiv:1412.1114*, 2014.
- [13] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4):1502–1509, 2016.
- [14] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.
- [15] Premier database. <https://www.football-data.co.uk/englandm.php>. Ημερομηνία πρόσβασης: 12-5-2024.
- [16] Premier league xgoals. <https://fbref.com/en/comps/9/schedule/Premier-League-Scores-and-Fixtures>. Ημερομηνία πρόσβασης: 12-5-2024.
- [17] Premier league advanced stats. https://github.com/footballdotpy/epl_scraper. Ημερομηνία πρόσβασης: 12-5-2024.
- [18] Fifa ratings. <https://www.fifaindex.com/>. Ημερομηνία πρόσβασης: 12-5-2024.

- [19] Team market value. <https://www.transfermarkt.com/>. Ημερομηνία πρόσβασης: 12-5-2024.