

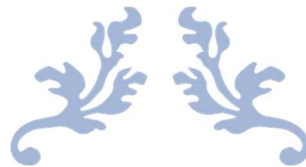


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΜΣ «ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»



ΧΡΗΣΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ
ΕΚΤΙΜΗΣΗ ΤΗΣ ΝΙΚΗΤΡΙΑΣ ΟΜΑΔΑΣ ΤΟΥ ΕΛΛΗΝΙΚΟΥ
ΠΡΩΤΑΘΛΗΜΑΤΟΣ ΠΟΔΟΣΦΑΙΡΟΥ



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δέδες Κωνσταντίνος

Επιβλέπων: Μαγκλογιάννης Ηλίας , Καθηγητής



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

Δέδες Κωνσταντίνος

A.M.: ME2207

Υποβάλλεται

για την εκπλήρωση των προϋποθέσεων λήψης

Μεταπτυχιακού Διπλώματος

στην ειδίκευση «Μεγάλα Δεδομένα και Αναλυτική»

του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”

στο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Μάρτιος 2026

Επιβλέπων: Μαγκλογιάννης Ηλίας

Ακαδημαϊκή Θέση: Καθηγητής

Πανεπιστήμιο Πειραιώς. Κάτοχος όλων των δικαιωμάτων

Συγγραφέας / Author Δέδες Κωνσταντίνος

ΠΕΡΙΛΗΨΗ

Η επιστήμη των δεδομένων έχει αναπτυχθεί με ταχύτατους ρυθμούς την τελευταία δεκαετία, με αποτέλεσμα να έχει εισχωρήσει σε πολλούς τομείς της οικονομίας. Ειδικότερα, ο τεράστιος όγκος δεδομένων που προκύπτει μέσα από την παραγωγική διαδικασία έχει οδηγήσει στην αποτελεσματικότερη εφαρμογή των μεθόδων και τεχνικών εξόρυξης των δεδομένων. Είναι εντυπωσιακό το γεγονός ότι αυτή η ραγδαία ανάπτυξη της επιστήμης των δεδομένων έχεις κάνει την εμφάνισή της με, επίσης, ταχείς ρυθμούς και στον αθλητισμό. Συγκεκριμένα, στον τομέα των προετοιμασίας των αθλητικών ομάδων, η χρήση των δεδομένων αποτελεί αναπόσπαστο, αν όχι το σημαντικότερο, κομμάτι πάνω στο οποίο χτίζονται. Δεν είναι τυχαίο το γεγονός ότι πολλές αθλητικές ομάδες, έχοντας αντιληφθεί τη σημασία των δεδομένων, δημιουργούν τμήματα data analytics που έχουν ως αποκλειστική ευθύνη την ανάπτυξη μοντέλων για την αξιοποίηση της γνώσης που δίνουν τα δεδομένα αυτά, καθώς και για την εκμετάλλευσή τους στο δρόμο προς την κατάκτηση της κορυφής, με την δημιουργία καινοτόμων τεχνικών εξόρυξης και ανάλυσης των δεδομένων.

Η συγκεκριμένη εργασία έχει σαν στόχο την ανάπτυξη μοντέλων με σκοπό την πρόβλεψη της νικήτριας ομάδας του ελληνικού πρωταθλήματος της Superleague, με την υψηλότερη δυνατή ακρίβεια, χρησιμοποιώντας και αξιολογώντας την απόδοση των, σχετικών με το πρόβλημα, αλγορίθμων επιβλεπόμενης μηχανικής μάθησης. Χρησιμοποιήθηκαν στατιστικά στοιχεία για δέκα αγωνιστικές περιόδους του Ελληνικού Πρωταθλήματος τα οποία αποκτήθηκαν με την μέθοδο web-scraping από την ιστοσελίδα transfermarkt. Τόσο η απόκτηση των δεδομένων, όσο και η υλοποίηση έγινε εξολοκλήρου σε γλώσσα προγραμματισμού Python.

Λέξεις-κλειδιά:

Superleague, python, τεχνητά νευρωνικά δίκτυα, μηχανική μάθηση, ανάλυση δεδομένων, μεγάλα δεδομένα

ABSTRACT

Data science has developed at a rapid pace in the last decade, as a result it has infiltrated many sectors of the economy. In particular, the huge amount of data resulting from the production process has led to the more effective application of data mining methods and techniques. It is impressive that this rapid development of data science has also made its appearance at a rapid pace in sports. Specifically, in the field of sports team preparation, the use of data is an integral, if not the most important, part on which they are built. It is no coincidence that many sports teams, having realized the importance of data, create data analytics departments whose sole responsibility is the development of models for exploiting the knowledge that this data provide, as well as for exploiting it on the way to conquering the top, by creating innovative data mining and analysis techniques.

This specific work targets to develop models with the aim of predicting the winning team of the Greek Superleague championship, with the highest possible accuracy, using and evaluating the performance of the problem-related supervised machine learning algorithms. Statistics were used for ten seasons of the Greek Championship which were obtained by web-scraping from the website transfermarkt. Both the acquisition of the data and the implementation were done entirely in Python programming language.

Keywords:

Superleague, python, artificial neural networks, machine learning, data analysis, big data

Περιεχόμενα

ΠΕΡΙΛΗΨΗ.....	2
ABSTRACT.....	3
Περιεχόμενα	4
Κατάλογος Σχημάτων.....	7
Κατάλογος Εικόνων.....	8
Κατάλογος Πινάκων	9
Κεφάλαιο 1. Εισαγωγή.....	10
Κεφάλαιο 2. Βιβλιογραφική Ανασκόπηση.....	15
2.1 Εισαγωγή.....	15
2.2 Κατηγοριοποίηση μεθοδολογιών και δεδομένων.....	15
2.3 Παρουσίαση επιμέρους ερευνών.....	15
2.3.1 Νευρωνικά δίκτυα και σημασία επιλογής χαρακτηριστικών	15
2.3.2 Συνδυασμός μηχανικής μάθησης και ανθρώπινης κρίσης.....	16
2.3.3 Βαθιά μάθηση και βελτιστοποίηση χαρακτηριστικών	16
2.3.4 Νευρωνικά δίκτυα σε διαφορετικά πρωταθλήματα	17
2.3.5 Συνολική επισκόπηση σύγχρονων τεχνικών	17
2.3.6 Ανάλυση ελληνικών δεδομένων – Στατιστικοί δείκτες.....	17
2.3.7 Μηχανική μάθηση και ελληνικό πρωτάθλημα – Σύγκριση με Ευρώπη	18
2.3.8 Προβλέψεις τελικής κατάταξης και απόδοσης παικτών.....	18
2.4 Επιστημονικό κενό και προκλήσεις	18
2.5 Σύνοψη.....	19
ΚΕΦΑΛΑΙΟ 3. Θεωρητικό και Τεχνικό Υπόβαθρο	20
3.1 Εισαγωγή.....	20
3.2 Λογιστική Παλινδρόμηση (Logistic Regression).....	20
3.3 Υπολογιστικά Νευρωνικά Δίκτυα (Artificial Neural Networks).....	21
3.4 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)	22
3.5 Αλγόριθμος Naïve Bayes	22
3.6 Δέντρα Αποφάσεων (Decision Trees)	23
3.7 Τυχαία Δάση (Random Forests)	23
3.8 K-Nearest Neighbors (KNN).....	24
3.9 Σύνοψη.....	24
Κεφάλαιο 4. Συλλογή δεδομένων και επεξεργασία	26
4.1 Εισαγωγή.....	26
4.2 Ελληνική Super League: Επισκόπηση.....	26

4.3 Συλλογή Δεδομένων	27
4.4 Επιλογή Μοντέλων	28
4.5 Καθαρισμός και Προεπεξεργασία Δεδομένων	29
4.6 Κατανομή και Ανάλυση Κλάσεων	29
4.7 Εμπλουτισμός και Ανάλυση Δεδομένων.....	30
4.8 Αντιμέτωπιση Missing Values και Εξομάλυνση	30
4.9 Κωδικοποίηση Κατηγορικών Χαρακτηριστικών (Encoding)	31
4.10 Ανάλυση Κώδικα: Διαδικασία Εκπαίδευσης – Αξιολόγησης	31
4.11 Προτάσεις για Μελλοντικό Εμπλουτισμό Δεδομένων.....	32
4.12 Σύνοψη.....	32
Κεφάλαιο 5. Μεθοδολογία	33
5.1 Εισαγωγή.....	33
5.2 Διαχωρισμός Δεδομένων και Πειραματικό Πρωτόκολλο	33
5.3 Αλγόριθμοι και Ρύθμιση Υπερπαραμέτρων.....	33
5.4 Διαχείριση Ανισορροπίας Κλάσεων.....	34
5.5 Διαδικασία Εκπαίδευσης και Αξιολόγησης.....	34
5.5.1 Ακρίβεια (Accuracy)	34
5.5.2 F1-score και weighted F1-score	35
5.5.3 Cohen’s Kappa.....	35
5.5.4 ROC-AUC (Receiver Operating Characteristic – Area Under Curve).....	36
5.5.5 Confusion Matrix.....	36
5.6 Οπτικοποίηση και Ερμηνεία Αποτελεσμάτων	36
5.7 Ανάλυση Αξιοπιστίας Μοντέλων	36
5.8 Συγκριτική Μεθοδολογία με Βάση τη Βιβλιογραφία.....	37
5.9 Σύνοψη.....	37
Κεφάλαιο 6. Πειραματικά Αποτελέσματα και Ερμηνεία	38
6.1 Εισαγωγή.....	38
6.2 Συγκεντρωτικά Αποτελέσματα Ταξινόμησης	39
6.3 Μετρικές απόδοσης για κάθε μοντέλο ταξινόμησης	39
6.3.1 Multinomial Naive Bayes (NB)	40
6.3.2 Logistic Regression	41
6.3.3 K-Nearest Neighbors (KNN).....	42
6.3.4 Support Vector Machine (SVM)	43
6.3.5 Decision Tree	44
6.3.6 Random Forest	45
6.3.7 Multilayer Perceptron (MLP).....	46

6.4 Ανάλυση και Ερμηνεία των Confusion Matrices.....	47
6.4.1 Μοντέλο Naive-Bayes	47
6.4.2 Λογιστική Παλινδρόμηση	48
6.4.3 K-Nearest Neighbors (KNN).....	49
6.4.4 Support Vector Classifier (SVC)	50
6.4.5 Δέντρα Αποφάσεων (Decision Tree)	51
6.4.6 Τυχαίο Δάσος (Random Forest)	52
6.4.7 Multi-layer Perceptron (MLP).....	53
6.5 Επεξήγηση Plots και Οπτικοποίηση Αποτελεσμάτων	54
6.5.1 Συνολικές Μετρικές – Plots.....	54
6.6 Leave-One-Season-Out Cross Validation.....	57
6.7 Σχολιασμός Συγκεκριμένων Σεζόν.....	60
6.7.1 Σεζόν 2017-2018	60
6.7.2 Σεζόν 2018-2019	61
6.7.3 Σεζόν 2019-2020	61
6.7.4 Σεζόν 2020-2021	61
6.7.5 Σεζόν 2022-2023 και 2023-2024	62
6.8 Πρόβλεψη Πρωταθλητή ανά Σεζόν και Μοντέλο	62
6.8.1 Ερμηνεία Αποτελεσμάτων	63
6.9 Πρόβλεψη Νικητή Πρωταθλήματος για τη Σεζόν 2025/2026	64
6.9.1 Αναλυτικά Αποτελέσματα Πρόβλεψης.....	64
6.9.2 Ερμηνεία Αποτελεσμάτων	65
6.9.3 Συμπέρασμα.....	67
6.10 Σύνοψη.....	67
Κεφάλαιο 7. Συμπεράσματα και Μελλοντικές Κατευθύνσεις	69
Βιβλιογραφία	71
Παράρτημα	73

Κατάλογος Σχημάτων

Σχήμα 1. Confusion Matrix για το μοντέλο Naive Bayes.....	47
Σχήμα 2. Confusion Matrix για το μοντέλο Logistic Regression.....	48
Σχήμα 3. Confusion Matrix για το μοντέλο k-Nearest Neighbors.	49
Σχήμα 4. Confusion Matrix για το μοντέλο Support Vector Classifier (SVC).	50
Σχήμα 5. Confusion Matrix για το μοντέλο Decision Tree.	51
Σχήμα 6. Confusion Matrix για το μοντέλο Random Forest.	52
Σχήμα 7. Confusion Matrix για το μοντέλο Multi-layer Perceptron (MLP).	53
Σχήμα 8. Bar Plot Ακρίβειας Αλγορίθμων.	54
Σχήμα 9. Bar Plot F1-score ανά Μοντέλο.	55
Σχήμα 10. Bar Plot Cohen's kappa ανά Μοντέλο.	56
Σχήμα 11. Bar Plot ROC-AUC ανά Μοντέλο.	57
Σχήμα 12. Bar Plot Μέσης Ακρίβειας LOSO ανά Μοντέλο.	58
Σχήμα 13. Bar Plot Μέσου F1-score LOSO ανά Μοντέλο.	59
Σχήμα 14. Bar Plot Μέσου Cohen's kappa LOSO ανά Μοντέλο.....	60

Κατάλογος Εικόνων

Εικόνα 1. Μετρικές απόδοσης για Naive-Bayes.....	40
Εικόνα 2. Μετρικές απόδοσης για Logistic Regression.....	41
Εικόνα 3. Μετρικές απόδοσης για K-Nearest Neighbors.....	42
Εικόνα 4. Μετρικές απόδοσης για Support Vector Machine.....	43
Εικόνα 5. Μετρικές απόδοσης για Decision Tree.....	44
Εικόνα 6. Μετρικές απόδοσης για Random Forest.....	45
Εικόνα 7. Μετρικές απόδοσης για Neural Networks.....	46

Κατάλογος Πινάκων

Πίνακας 1. Συγκριτική Παρουσίαση Αλγορίθμων Μηχανικής Μάθησης	25
Πίνακας 2. Συνολικές μετρικές απόδοσης όλων των μοντέλων ταξινόμησης (όλες οι σεζόν)	39
Πίνακας 3. Μέσες επιδόσεις ανά αλγόριθμο με Leave-One-Season-Out Cross Validation	58
Πίνακας 4. Πρόβλεψη πρωταθλητή ελληνικού πρωταθλήματος ανά σεζόν και μοντέλο	62
Πίνακας 5. Συγκεντρωτικός Πίνακας Πρόβλεψης Πρωταθλητή για τη σεζόν 2025/2026	65

Κεφάλαιο 1. Εισαγωγή

Η ανάλυση των δεδομένων έχει περάσει, τις τελευταίες δύο δεκαετίες, στην καθημερινότητα των επιχειρήσεων, των κυβερνήσεων και γενικότερα της ζωής μας. Η επιστήμη των υπολογιστών έχει εξελιχθεί όσο ποτέ άλλοτε ώστε έχει καταφέρει να γεννήσει την τεχνητή νοημοσύνη, η οποία αποτελεί την επιστήμη που ενσωματώνει τη θεωρία, τις μεθόδους, τις τεχνικές, τα μοντέλα, τα συστήματα και τις εφαρμογές της ανθρώπινης νοημοσύνης. Έχει εφαρμογή στην υγεία, την εκπαίδευση, τη γεωργία, το διάστημα, τους ενεργειακούς πόρους, τη βιομηχανία και σε πολλούς άλλους τομείς.

Ο τομέας του αθλητισμού δεν θα μπορούσε να αποτελεί εξαίρεση, καθώς ο αθλητισμός από τη φύση του, είναι μια δραστηριότητα που απαιτεί πλήρη απαρίθμηση παραμέτρων (δεδομένων) προκειμένου να κατανοηθεί το παιχνίδι, οι στρατηγικές του και ο τρόπος λήψης αποφάσεων, ελαχιστοποιώντας το απρόβλεπτο [1]. Ο τρόπος με τον οποίο γίνεται αυτή η επεξεργασία των δεδομένων ονομάζεται εποπτευόμενη μάθηση, καθώς η φάση επεξεργασίας των δεδομένων καθοδηγείται προς τη μεταβλητή κλάσης κατά την κατασκευή του μοντέλου [2].

Ειδικότερα η πρόβλεψη των αποτελεσμάτων σε αθλητικούς αγώνες είναι ένας ταχέως αναπτυσσόμενος τομέας, γεγονός που σχετίζεται με την αυξανόμενη διαθεσιμότητα των δεδομένων και την άνοδο των αθλητικών αναλυτικών στοιχείων. Συγκεκριμένα, η πρόβλεψη των αποτελεσμάτων του ποδοσφαίρου έχει σημαντικές επιπτώσεις σε διάφορους τομείς, όπως:

1. Το αθλητικό στοίχημα: Το ποδόσφαιρο είναι ένα από τα αθλήματα με τον μεγαλύτερο στοιχηματισμό παγκοσμίως και οι ακριβείς προβλέψεις μπορούν να προσφέρουν σημαντικά πλεονεκτήματα στους παίκτες και τις πλατφόρμες στοιχημάτων. Τα αθλητικά βιβλία βασίζονται σε προηγμένα μοντέλα πρόβλεψης για τον καθορισμό των αποδόσεων και οι παίκτες του στοιχήματος χρησιμοποιούν διάφορες τεχνικές πρόβλεψης για να λαμβάνουν τεκμηριωμένες αποφάσεις.
2. Τη στρατηγική και διαχείριση της ομάδας: Οι ποδοσφαιρικοί σύλλογοι χρησιμοποιούν όλο και περισσότερο προβλέψεις βάσει δεδομένων για να καθοδηγήσουν τις τακτικές τους αποφάσεις. Τα προγνωστικά μοντέλα μπορούν να αξιολογήσουν την πιθανότητα

νίκης, την απόδοση των παικτών ή την αποτελεσματικότητα συγκεκριμένων σχηματισμών. Οι προπονητές μπορούν να βελτιστοποιήσουν στρατηγικές με βάση ιστορικά δεδομένα και ανάλυση αντιπάλων.

3. Τη δέσμευση των οπαδών: Οι προβλέψεις ενισχύουν την αφοσίωση των οπαδών παρέχοντας βαθύτερες πληροφορίες για τα αποτελέσματα των παιχνιδιών, τις επιδόσεις των παικτών και τις τάσεις της ομάδας. Οι οπαδοί συχνά απολαμβάνουν να ασχολούνται με εργαλεία πρόβλεψης, πρωταθλήματα φαντασίας και σχολιασμούς που βασίζονται σε αναλυτικά στοιχεία για να βελτιώσουν την κατανόησή τους για το άθλημα.
4. Την ανάλυση της απόδοσης παικτών και ομάδων: Οι σύλλογοι χρησιμοποιούν προγνωστικά αναλυτικά στοιχεία για να αξιολογήσουν την απόδοση των παικτών, να αξιολογήσουν τους κινδύνους τραυματισμών και να πάρουν καλύτερες αποφάσεις για μεταγραφές και εναλλαγές παικτών. Τα προηγμένα μοντέλα πρόβλεψης βοηθούν στην αξιολόγηση παραγόντων όπως η αντοχή, η φόρμα και τα πιθανά αποτελέσματα αγώνα.
5. Τα Μέσα Μαζικής Ενημέρωσης: Τα αθλητικά μέσα ενημέρωσης βασίζονται σε προβλέψεις για να δημιουργήσουν ελκυστικό περιεχόμενο, προσφέροντας στους θαυμαστές και τους θεατές αναλύσεις επερχόμενων αγώνων, πιθανών αποτελεσμάτων και δυνατά/αδύναμα σημεία της ομάδας.

Στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία και εξέταση ενός συστήματος πρόβλεψης για τον νικητή των αγώνων του Ελληνικού Πρωταθλήματος, αξιοποιώντας μοντέλα μηχανικής μάθησης. Η εργασία εστιάζει στη βελτιστοποίηση της ακρίβειας των προβλέψεων, συνδυάζοντας στατιστικά δεδομένα από δέκα αγωνιστικές περιόδους. Η ανάγκη για τη δημιουργία ενός τέτοιου συστήματος πρόβλεψης βασίζεται στην ανάλυση μεγάλου όγκου δεδομένων αγώνων και στην ικανότητα των μοντέλων μηχανικής μάθησης να εντοπίζουν μοτίβα που σχετίζονται με το αποτέλεσμα των αγώνων.

Συγκεκριμένα, το ποδόσφαιρο, είναι εγγενώς απρόβλεπτο λόγω της δυναμικής του φύσης και πολλές προκλήσεις προκύπτουν κατά τη δημιουργία προγνωστικών μοντέλων, όπως οι ακόλουθες:

1. Πολυπλοκότητα του παιχνιδιού: Σε αντίθεση με κάποια άλλα αθλήματα, το ποδόσφαιρο έχει χαμηλό σκορ, πράγμα που σημαίνει ότι ένας μικρός αριθμός γεγονότων (όπως γκολ) μπορεί να έχει τεράστιο αντίκτυπο στο αποτέλεσμα. Αυτό κάνει τις προβλέψεις δύσκολες, καθώς ένα μόνο απροσδόκητο γεγονός μπορεί να αλλάξει δραστικά το αποτέλεσμα. Για παράδειγμα, μια κόκκινη κάρτα, ένας τραυματισμός ή μια ατομική στιγμή λαμπρότητας μπορεί να κρίνει έναν αγώνα.
2. Διαθεσιμότητα και ποιότητα δεδομένων: Ενώ τα δεδομένα για μεγάλα πρωταθλήματα (όπως η αγγλική Premier League) είναι συχνά εκτεταμένα, η διαθεσιμότητα δεδομένων για την ελληνική Super League μπορεί να είναι πιο περιορισμένη, ειδικά για παλαιότερες σεζόν ή λιγότερο εξέχοντες συλλόγους. Ακόμη και όταν τα δεδομένα είναι διαθέσιμα, η διασφάλιση της ακρίβειας και της συνέπειάς τους σε όλες τις εποχές αποτελεί πρόκληση. Τα περιορισμένα ή ελλιπή δεδομένα μπορούν να επηρεάσουν σοβαρά την απόδοση του μοντέλου.
3. Επιλογή χαρακτηριστικών και πολυπλοκότητα μοντέλου: Οι ποδοσφαιρικοί αγώνες περιλαμβάνουν πολλές μεταβλητές, από σχηματισμούς ομάδων και χαρακτηριστικά παικτών έως καιρικές συνθήκες, τραυματισμούς και πλεονέκτημα εντός έδρας. Ο προσδιορισμός των χαρακτηριστικών που συμβάλλουν περισσότερο στην αντιστοίχιση των αποτελεσμάτων είναι κρίσιμος αλλά δύσκολος. Για παράδειγμα, ορισμένες ομάδες μπορεί να έχουν καλύτερες επιδόσεις στα εντός έδρας παιχνίδια λόγω της παθιασμένης υποστήριξης των φιλάθλων, ενώ άλλες μπορεί να βασίζονται σε μεγάλο βαθμό σε βασικούς παίκτες των οποίων η απουσία (π.χ. λόγω τραυματισμού) μπορεί να αλλάξει δραστικά τις προβλέψεις.
4. Εγγενής τυχαιότητα: Τα αποτελέσματα του ποδοσφαίρου συχνά περιλαμβάνουν τυχαία γεγονότα όπως τραυματισμούς, αποφάσεις διαιτησίας ή στιγμιαία κενά στη συγκέντρωση. Αυτά είναι δύσκολο να μοντελοποιηθούν ή να προβλεφθούν με ακρίβεια. Ακόμη και οι πιο προηγμένοι αλγόριθμοι αγωνίζονται να λάβουν υπόψη τα τυχαία, ανθρώπινα στοιχεία που παίζουν μεγάλο ρόλο στο άθλημα.

5. Αλλαγές τακτικής κατά τη διάρκεια των αγώνων: Οι ποδοσφαιρικοί αγώνες είναι ρευστοί και μπορούν να αλλάξουν με βάση τις τακτικές προσαρμογές από τους προπονητές. Για παράδειγμα, μια ομάδα μπορεί να στραφεί σε έναν πιο αμυντικό σχηματισμό αφού πετύχει ένα γκολ ή μπορεί να πάρει περισσότερα ρίσκα αν καθυστερήσει να πετύχει γκολ μέσα στον αγώνα. Αυτές οι αλλαγές εντός του παιχνιδιού είναι δύσκολο να ενσωματωθούν σε μοντέλα πρόβλεψης πριν από τον αγώνα.
6. Εποχιακές Παραλλαγές και Μεταγραφές: Η απόδοση των ομάδων και των παικτών μπορεί να κυμαίνεται μεταξύ των εποχών. Οι μεταγραφές παικτών, οι διοικητικές αλλαγές ή οι τραυματισμοί μπορεί να έχουν σημαντικό αντίκτυπο στην απόδοση της ομάδας. Ένα μοντέλο που εκπαιδεύεται σε μια σεζόν μπορεί να μην έχει καλή απόδοση σε μια άλλη, καθώς οι ομάδες μπορούν να εξελιχθούν γρήγορα σε μικρές περιόδους.
7. Τραυματισμοί και φυσική κατάσταση παικτών: Το ποδόσφαιρο είναι ένα σωματικά απαιτητικό άθλημα και οι τραυματισμοί των παικτών ή η έλλειψη φυσικής κατάστασης μπορούν να αλλάξουν σημαντικά τα αποτελέσματα του αγώνα. Τα μοντέλα πρόβλεψης συχνά δυσκολεύονται να εξηγήσουν τους ξαφνικούς τραυματισμούς ή την απροσδόκητη κόπωση των παικτών, ιδιαίτερα όταν τα σχετικά δεδομένα δεν είναι διαθέσιμα ή είναι δύσκολο να μετρηθούν.
8. Εξωτερικοί παράγοντες: Στοιχεία όπως οι καιρικές συνθήκες, η επιρροή του πλήθους ή ακόμα και οι αποφάσεις των διαιτητών μπορούν να επηρεάσουν τα αποτελέσματα του παιχνιδιού. Η ενσωμάτωση αυτών των απρόβλεπτων μεταβλητών στα μοντέλα προσθέτει πολυπλοκότητα. Για παράδειγμα, ένας αγώνας που παίζεται σε δυνατή βροχή μπορεί να ευνοήσει τις περισσότερες αμυντικές ομάδες, ενώ ο θόρυβος του κόσμου στο γήπεδο μπορεί να ενισχύσει την απόδοση των παικτών.

Με βάση τις παραπάνω προκλήσεις, το κύριο ερώτημα που θέτει η παρούσα μελέτη είναι:

Πώς μπορεί να προβλεφθεί με ακρίβεια ο νικητής ενός αγώνα χρησιμοποιώντας ιστορικά στατιστικά δεδομένα και σύγχρονα μοντέλα μηχανικής μάθησης;

Η πρόβλεψη αυτή βασίζεται σε μια σειρά παραδοχών, όπως:

- Η ιστορική απόδοση μιας ομάδας επηρεάζει σε σημαντικό βαθμό την πιθανότητα νίκης σε έναν επόμενο αγώνα.
- Η ποσοτικοποίηση των χαρακτηριστικών απόδοσης των ομάδων μπορεί να λειτουργήσει ως εισαγωγή (input) για τα μοντέλα μηχανικής μάθησης.
- Η διαμόρφωση του πρωταθλήματος (π.χ., διπλοί γύροι) προσφέρει ένα συνεπές πλαίσιο για τη συλλογή δεδομένων που μπορούν να αξιοποιηθούν για μακροπρόθεσμες προβλέψεις.

Είναι πραγματικότητα ότι η σημασία της πρόβλεψης ποδοσφαίρου εκτείνεται πολύ πέρα από τα απλά αποτελέσματα αγώνων. Επηρεάζει ένα ευρύ φάσμα ενδιαφερομένων, από παίκτες και οπαδούς μέχρι προπονητές και αναλυτές αθλημάτων. Στο πλαίσιο της ελληνικής Super League, η ανάπτυξη ακριβών μοντέλων πρόβλεψης αποτελεί μοναδική πρόκληση λόγω της απρόβλεπτης φύσης του ποδοσφαίρου, της μεταβλητότητας στην ποιότητα των δεδομένων και των δυναμικών παραγόντων σε πραγματικό χρόνο που επηρεάζουν τα αποτελέσματα των αγώνων. Παρά τις προκλήσεις αυτές, οι εξελίξεις στη μηχανική μάθηση και στην ανάλυση δεδομένων προσφέρουν πολλά υποσχόμενα εργαλεία για τη βελτίωση της ακρίβειας και της αξιοπιστίας των προβλέψεων στο ποδόσφαιρο.

Στη συνέχεια, η δομή της εργασίας είναι η εξής: το κεφάλαιο 2 περιλαμβάνει μια βιβλιογραφική ανασκόπηση παρόμοιων εργασιών που αφορούν το εγχώριο πρωτάθλημα της Super League, αλλά και πρωταθλήματα άλλων χωρών. Το κεφάλαιο 3 περιλαμβάνει το θεωρητικό υπόβαθρο πίσω από τις μεθόδους που θα χρησιμοποιηθούν για την παραγωγή των αποτελεσμάτων. Στο κεφάλαιο 4 γίνεται η συλλογή και η επεξεργασία των δεδομένων που είναι απαραίτητα για την εκτέλεση των μεθόδων που θα χρησιμοποιηθούν. Το κεφάλαιο 5, εξετάζει τη μεθοδολογία που χρησιμοποιείται για την επεξεργασία των δεδομένων. Στο κεφάλαιο 6 δίνονται τα αποτελέσματα που προέκυψαν από την εκτέλεση των μεθόδων μηχανικής μάθησης. Τέλος, το κεφάλαιο 7, περιλαμβάνει τα συμπεράσματα που προκύπτουν από τα αποτελέσματα, καθώς και προτάσεις για μελλοντική έρευνα.

Κεφάλαιο 2. Βιβλιογραφική Ανασκόπηση

2.1 Εισαγωγή

Τα τελευταία χρόνια, με τη ραγδαία εξάπλωση των εφαρμογών μηχανικής μάθησης, έχουν προταθεί και εφαρμοστεί πολυάριθμες μέθοδοι που αξιοποιούν στατιστικά δεδομένα, εξελιγμένα χαρακτηριστικά και αλγορίθμους τεχνητής νοημοσύνης. Παράλληλα, η ανάλυση της επίδρασης παραγόντων όπως η ψυχολογία των αθλητών και η διαθεσιμότητα παικτών αρχίζει να κερδίζει έδαφος, διαμορφώνοντας μια νέα, πολυπαραγοντική προσέγγιση στην πρόβλεψη.

2.2 Κατηγοριοποίηση μεθοδολογιών και δεδομένων

Οι βιβλιογραφικές πηγές καταδεικνύουν ότι οι μεθοδολογίες μπορούν να κατηγοριοποιηθούν ως προς:

- *Το είδος των δεδομένων:* στατιστικά αγωνιστικά στοιχεία, χαρακτηριστικά ομάδων/παικτών, δεδομένα στοιχηματικών αποδόσεων, χωροχρονικά δεδομένα, ακόμα και ανάλυση συναισθηματικού περιεχομένου ειδήσεων.
- *Το στάδιο πρόβλεψης:* πριν τον αγώνα, κατά τη διάρκεια του αγώνα, ή σε μακροπρόθεσμη βάση (τελική κατάταξη).
- *Το ζητούμενο της πρόβλεψης:* αποτέλεσμα αγώνα (νίκη-ισοπαλία-ήττα), αριθμός γκολ, τελική θέση στη βαθμολογία, ατομική απόδοση παικτών.
- *Την τεχνική προσέγγιση:* στατιστικές μέθοδοι, κλασικοί αλγόριθμοι μηχανικής μάθησης (Random Forest, Gradient-Boosted Trees, SVM), νευρωνικά δίκτυα, βαθιά μάθηση (deep learning), υβριδικές προσεγγίσεις.

2.3 Παρουσίαση επιμέρους ερευνών

2.3.1 Νευρωνικά δίκτυα και σημασία επιλογής χαρακτηριστικών

Η μελέτη των Aslan και Inceoglu [3] εξετάζει πώς διαφορετικές επιλογές χαρακτηριστικών επηρεάζουν την ακρίβεια νευρωνικών δικτύων στην πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων. Εστιάζει ιδιαίτερα στη σημασία της επιλογής των κατάλληλων εισόδων (χαρακτηριστικών) για την εκπαίδευση του μοντέλου. Οι συγγραφείς

πειραματίζονται με δύο διαφορετικά σύνολα χαρακτηριστικών και αξιολογούν πώς αυτά επηρεάζουν την ακρίβεια των προβλέψεων. Τα αποτελέσματά τους δείχνουν ότι ακόμα και απλές αλλά κατάλληλα επιλεγμένες μεταβλητές μπορούν να οδηγήσουν σε ικανοποιητική πρόβλεψη, με ακρίβεια, η οποία στις δοκιμές τους έφτασε έως και το 57,6% για την πρόβλεψη νίκης/ισοπαλίας/ήττας, επισημαίνοντας την ανάγκη για έξυπνη επιλογή χαρακτηριστικών στις εφαρμογές μηχανικής μάθησης για το ποδόσφαιρο.

2.3.2 Συνδυασμός μηχανικής μάθησης και ανθρώπινης κρίσης

Οι Beal et al. [4] εφάρμοσαν μια υβριδική μέθοδο που συνδυάζει παραδοσιακά στατιστικά μοντέλα μηχανικής μάθησης με την ανάλυση ειδικών επί του θέματος (αθλητικών δημοσιογράφων) για την πρόβλεψη των αποτελεσμάτων αγώνων ποδοσφαίρου στην Αγγλική Premier League.. Οι συγγραφείς αξιοποιούν στατιστικά δεδομένα αγώνων και ταυτόχρονα αναλύουν δημοσιογραφικά άρθρα, αποδεικνύοντας ότι η ενσωμάτωση ανθρώπινης ανάλυσης αυξάνει την ακρίβεια πρόβλεψης κατά περίπου 7% σε σχέση με τα καθαρά στατιστικά μοντέλα. Το τελικό τους μοντέλο πέτυχε ακρίβεια 63,18%, υποδεικνύοντας ότι ο συνδυασμός ανθρώπινης κρίσης και αλγορίθμων βελτιώνει σημαντικά την πρόβλεψη αποτελεσμάτων.

2.3.3 Βαθιά μάθηση και βελτιστοποίηση χαρακτηριστικών

Στη μελέτη των Yeung et al. [5], εξετάζεται η χρήση μοντέλων μηχανικής μάθησης, και ειδικά τεχνικών βαθιάς μάθησης (deep learning), για την πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων. Οι ερευνητές δοκιμάζουν διάφορα σύνολα χαρακτηριστικών και αλγορίθμους, με έμφαση στη βελτιστοποίηση χαρακτηριστικών για μοντέλα gradient-boosted trees (όπως το CatBoost) και συγκρίνουν τις επιδόσεις με ένα deep learning μοντέλο. Η επιλογή των χαρακτηριστικών (feature optimization) αποδείχθηκε κρίσιμη: τα μοντέλα CatBoost με βελτιστοποιημένα χαρακτηριστικά ξεπέρασαν το 58% ακρίβεια, ενώ το deep learning προσέφερε συγκρίσιμες αλλά όχι σαφώς ανώτερες επιδόσεις στα δεδομένα του διαγωνισμού Soccer Prediction Challenge 2023.

2.3.4 Νευρωνικά δίκτυα σε διαφορετικά πρωταθλήματα

Η έρευνα των Arabzad et al. [6], εφαρμόζει τεχνητά νευρωνικά δίκτυα για την πρόβλεψη των αποτελεσμάτων αγώνων ποδοσφαίρου στο ιρανικό πρωτάθλημα (Iran Pro League). Οι συγγραφείς χρησιμοποίησαν ιστορικά δεδομένα αγώνων επτά ετών για να εκπαιδεύσουν το μοντέλο τους και να προβλέψουν μελλοντικά αποτελέσματα. Η ανάλυση δείχνει ότι τα νευρωνικά δίκτυα παρουσιάζουν εξαιρετική ικανότητα πρόβλεψης των αποτελεσμάτων, πετυχαίνοντας ακρίβεια έως και 71,4% στην πρόβλεψη αποτελέσματος (νίκη/ισοπαλία/ήττα), παρά το γεγονός ότι το ποδόσφαιρο επηρεάζεται από πολλούς, δύσκολα μετρήσιμους παράγοντες. Η μελέτη αυτή ενισχύει τη θέση ότι οι τεχνικές μηχανικής μάθησης αποτελούν ένα ισχυρό εργαλείο στον χώρο των αθλητικών προβλέψεων

2.3.5 Συνολική επισκόπηση σύγχρονων τεχνικών

Οι Bunker, Yeung & Fujii [7] επισημαίνουν ότι τα gradient-boosted tree μοντέλα (όπως CatBoost), με εξειδικευμένα ποδοσφαιρικά ratings, θεωρούνται σήμερα state-of-the-art για πρόβλεψη αποτελεσμάτων βάσει σκορ. Ωστόσο, η διαφορά από τα deep learning μοντέλα δεν είναι πάντα ξεκάθαρη, ενώ η περαιτέρω ενσωμάτωση χωροχρονικών χαρακτηριστικών (π.χ. δεδομένων θέσης παικτών, συμβάντων) μπορεί να ενισχύσει σημαντικά τα αποτελέσματα. Ενδεικτικά, σε datasets μόνο με σκορ αγώνων, τα CatBoost φτάνουν ή και ξεπερνούν το 55% ακρίβεια.

2.3.6 Ανάλυση ελληνικών δεδομένων – Στατιστικοί δείκτες

Η μελέτη για τη Superleague 2011–2012 [8] ανέλυσε πώς οι επιθετικοί και αμυντικοί δείκτες (π.χ. γκολ υπέρ/κατά, σουτ, πάσες, αποκρούσεις) επηρεάζουν την τελική κατάταξη. Βρέθηκε ότι η συνολική αγωνιστική ισορροπία (επιθετική και αμυντική επίδοση) και η αμυντική συνέπεια είναι εξίσου καθοριστικές με την επιθετική αποτελεσματικότητα, χωρίς όμως να δίνεται συγκεκριμένο ποσοστό πρόβλεψης. Αυτή η ανάλυση είναι ιδιαίτερα σημαντική για το ελληνικό ποδόσφαιρο, καθώς φωτίζει τοπικές ιδιαιτερότητες που δεν εμφανίζονται πάντα σε άλλα ευρωπαϊκά πρωταθλήματα.

2.3.7 Μηχανική μάθηση και ελληνικό πρωτάθλημα – Σύγκριση με Ευρώπη

Η έρευνα των Malamatinos et al. [9] επικεντρώνεται στην πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων της ελληνικής Superleague, δοκιμάζοντας πολλαπλούς αλγορίθμους (k-NN, LogitBoost, SVM, Random Forest, CatBoost, CNN με transfer learning), συγκρίνοντας τα αποτελέσματα με Αγγλία και Ολλανδία. Η μελέτη καταλήγει ότι τα κλασικά μοντέλα δέντρων (όπως το CatBoost) αποδίδουν εξαιρετικά καλά, με την προσθήκη εξελιγμένων χαρακτηριστικών, ενώ η προσέγγιση με βαθιά μάθηση έχει επιπλέον περιθώρια βελτίωσης στο μέλλον. Το CatBoost αναδείχθηκε ως το πιο ακριβές μοντέλο, πετυχαίνοντας ακρίβεια 54,4% στην Ελλάδα, 54,8% στην Ολλανδία και 53,2% στην Αγγλία. Τα CNN, αν και καινοτόμα, δεν ξεπέρασαν τα παραδοσιακά μοντέλα σε ακρίβεια, υποδεικνύοντας το σημαντικό ρόλο των «καθαρών» χαρακτηριστικών στα ελληνικά δεδομένα.

2.3.8 Προβλέψεις τελικής κατάταξης και απόδοσης παικτών

Τέλος, η μελέτη των Chazan-Pantzalis & Tjortjis [10] εστιάζει στη χρήση τεχνικών μηχανικής μάθησης για την πρόβλεψη της τελικής κατάταξης ομάδων σε ευρωπαϊκά πρωταθλήματα ποδοσφαίρου, καθώς και για την πρόβλεψη ατομικής απόδοσης παικτών. Οι συγγραφείς χρησιμοποιούν προηγμένα στατιστικά στοιχεία και δεδομένα από προηγούμενες σεζόν για να εκπαιδεύσουν μοντέλα πρόβλεψης. Τα αποτελέσματα είναι ενθαρρυντικά, με τις ακριβείς προβλέψεις της τελικής βαθμολογίας να ξεπερνούν σε ορισμένες περιπτώσεις το 70% σε σχέση με την πραγματική τελική σειρά. Παράλληλα, γίνεται ανάλυση για την πρόβλεψη αν μία ομάδα θα βελτιώσει την απόδοσή της σε σύγκριση με την προηγούμενη σεζόν. Επιπλέον, η εργασία αναδεικνύει τη δυνατότητα εντοπισμού, με μεγάλη ακρίβεια, παικτών που ξεχωρίζουν στατιστικά, όπως οι κορυφαίοι αμυντικοί, με βάση δεδομένα αγωνιστικής δράσης.

2.4 Επιστημονικό κενό και προκλήσεις

Παρά τη σημαντική πρόοδο που έχει επιτευχθεί, εξακολουθούν να υφίστανται ορισμένα επιστημονικά κενά:

- Τα μοντέλα deep learning εμφανίζουν σημαντικές δυνατότητες, αλλά δεν έχουν ακόμα ξεκάθαρη υπεροχή έναντι των εξελιγμένων ensemble αλγορίθμων.

- Η διαθεσιμότητα ανοιχτών, πλήρων και ποιοτικών δεδομένων, ειδικά για λιγότερο εμπορικά πρωταθλήματα όπως το ελληνικό, παραμένει περιορισμένη.
- Η ενσωμάτωση μη-αγωνιστικών παραμέτρων (π.χ. ψυχολογικών, κοινωνικών) στην πρόβλεψη αποτελεί ακόμη ζητούμενο για την αθλητική ανάλυση.
- Η ερμηνευσιμότητα των μοντέλων και η δυνατότητα πρακτικής αξιοποίησής τους από ομάδες και προπονητές είναι επίσης περιορισμένη.

2.5 Σύνοψη

Η διεθνής και εγχώρια βιβλιογραφία δείχνει ότι τα σύγχρονα μοντέλα μηχανικής μάθησης προσφέρουν σημαντικές δυνατότητες για την πρόβλεψη αποτελεσμάτων και τη βαθύτερη κατανόηση του ποδοσφαίρου. Παρά τα υψηλά ποσοστά επιτυχίας που παρατηρούνται σε επιμέρους πρωταθλήματα και δείγματα, οι σύγχρονες προκλήσεις αφορούν τόσο στην περαιτέρω βελτίωση των αλγορίθμων όσο και στη διεύρυνση των χαρακτηριστικών που αξιοποιούνται. Η παρούσα εργασία στοχεύει να καλύψει μέρος αυτού του κενού, εφαρμόζοντας εξελιγμένες μεθόδους πρόβλεψης στο ελληνικό ποδόσφαιρο, με έμφαση στα ιδιαίτερα χαρακτηριστικά και τις ανάγκες του εγχώριου πρωταθλήματος.

ΚΕΦΑΛΑΙΟ 3. Θεωρητικό και Τεχνικό Υπόβαθρο

3.1 Εισαγωγή

Οι αλγόριθμοι μηχανικής μάθησης είναι ισχυρά εργαλεία για την επίλυση εργασιών ταξινόμησης, όπως η πρόβλεψη του αποτελέσματος των ποδοσφαιρικών αγώνων. Κάθε αλγόριθμος φέρνει μοναδικά πλεονεκτήματα και υποθέσεις στην εργασία και η κατανόηση του θεωρητικού τους υποβάθρου είναι κρίσιμη για την επιλογή του σωστού μοντέλου για ένα συγκεκριμένο πρόβλημα.

Σε αυτή την ενότητα παρέχουμε ορισμένες θεωρητικές και τεχνικές έννοιες, που σχετίζονται με τους αλγορίθμους επιβλεπόμενης ταξινόμησης που θα χρησιμοποιηθούν για την παραγωγή των αποτελεσμάτων της εργασίας.

Παρακάτω είναι μια επισκόπηση των αλγορίθμων ταξινόμησης που χρησιμοποιούνται ευρέως στη διεθνή και ελληνική βιβλιογραφία για την ανάλυση και πρόβλεψη αθλητικών δεδομένων, με ιδιαίτερη έμφαση στη λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα (Neural Networks), τις μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), τον αλγόριθμο Naive – Bayes, τα δέντρα αποφάσεων (Decision Trees), τα τυχαία δάση (Random Forests) και τον αλγόριθμο k-Nearest Neighbors.

3.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση αποτελεί έναν από τους κλασικότερους αλγορίθμους στατιστικής ταξινόμησης για δυαδικά και πολυταξικά προβλήματα. Ο Cox (1958) διατύπωσε το θεωρητικό πλαίσιο της λογιστικής παλινδρόμησης ως ένα γενικευμένο γραμμικό μοντέλο για την εκτίμηση της πιθανότητας εμφάνισης ενός γεγονότος, βασιζόμενος στα χαρακτηριστικά των δεδομένων εισόδου [11]. Βασίζεται στη σιγμοειδή συνάρτηση, η οποία μετατρέπει το γραμμικό συνδυασμό των χαρακτηριστικών σε τιμές πιθανότητας μέσω της σχέσης:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}}$$

Η εκπαίδευση του μοντέλου πραγματοποιείται με τη μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood estimation), ενώ η προσέγγισή της επιτρέπει ερμηνεία των συντελεστών (weights) ως λογαριθμικών πιθανοτήτων. Στην ανάλυση αθλητικών δεδομένων, η λογιστική παλινδρόμηση έχει αποδειχθεί αποτελεσματική στην πρόβλεψη του αποτελέσματος αγώνων, κυρίως όταν τα χαρακτηριστικά παρουσιάζουν σχετικώς γραμμικές συσχετίσεις με το αποτέλεσμα [12]. Παράλληλα, διατηρεί την ικανότητά της για πολυταξική επέκταση (π.χ. softmax regression), διευρύνοντας το πεδίο εφαρμογής της. Ωστόσο, η αποτελεσματικότητά της μειώνεται όταν τα χαρακτηριστικά έχουν μη γραμμικές αλληλεξαρτήσεις ή σε προβλήματα υψηλής διαστατικότητας.

3.3 Υπολογιστικά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Τα τεχνητά νευρωνικά δίκτυα αποτελούν έναν από τους πιο ισχυρούς αλγόριθμους μηχανικής μάθησης για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Το βασικό μοντέλο των πολυεπίπεδων αντιληπτικών (MLP) και ο αλγόριθμος οπισθοδιάδοσης του σφάλματος (backpropagation) παρουσιάστηκαν από τους Rumelhart, Hinton και Williams (1986) [13]. Τα νευρωνικά δίκτυα διακρίνονται για τη δομή τους σε επίπεδα (input, hidden, output) και για τη δυνατότητά τους να μαθαίνουν πολύπλοκες, μη γραμμικές συσχετίσεις μεταξύ των χαρακτηριστικών εισόδου και της εξόδου.

Ένα βασικό πλεονέκτημα των νευρωνικών δικτύων είναι η ικανότητά τους να αυτοματοποιούν την εξαγωγή ενδιάμεσων αναπαραστάσεων των δεδομένων, προσφέροντας ισχυρή απόδοση ακόμα και όταν τα παραδοσιακά μοντέλα αδυνατούν να αποδώσουν [14] [3] [6]. Έχουν χρησιμοποιηθεί ευρέως στην ανάλυση αθλητικών δεδομένων, με αποτελέσματα που πολλές φορές ξεπερνούν σε ακρίβεια άλλες μεθόδους, όταν διατίθενται αρκετά δεδομένα εκπαίδευσης. Παράλληλα, η υπερπαραμετροποίηση και η ανάγκη για επαρκή όγκο δεδομένων τα καθιστά επιρρεπή στο φαινόμενο της υπερπροσαρμογής (overfitting) και αυξάνει το υπολογιστικό τους κόστος.

3.4 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι μηχανές υποστηρικτικών διανυσμάτων (SVM), που εισήχθησαν από τους Cortes και Vapnik (1995) [15], βασίζονται στην αρχή της εύρεσης του υπερεπιπέδου με το μέγιστο περιθώριο μεταξύ των κλάσεων. Η φιλοσοφία του SVM επικεντρώνεται στην εύρεση του διαχωριστικού υπερεπιπέδου που μεγιστοποιεί το ελάχιστο περιθώριο από τα πλησιέστερα σημεία κάθε κλάσης (support vectors). Η δυνατότητα ενσωμάτωσης μη γραμμικών διαχωριστικών με χρήση kernels (π.χ. πολυωνυμικός, Gaussian) καθιστά το SVM εξαιρετικά ευέλικτο [15].

Η μέθοδος αυτή είναι ιδιαίτερα αποτελεσματική σε προβλήματα με υψηλή διαστατικότητα ή με μη γραμμικά όρια μεταξύ των κλάσεων, ενώ έχει αποδείξει αξιοσημείωτη απόδοση σε πλήθος εργασιών πρόβλεψης αθλητικών αποτελεσμάτων [12]. Κύριο μειονέκτημα αποτελεί η αυξημένη υπολογιστική απαίτηση σε μεγάλα σύνολα δεδομένων και η ευαισθησία στην επιλογή του kernel και των υπερπαραμέτρων.

3.5 Αλγόριθμος Naive Bayes

Οι ταξινομητές Naïve-Bayes είναι πιθανοκρατικά μοντέλα που στηρίζονται στο θεώρημα του Bayes και στην υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών [16]. Η “αφέλεια” της υπόθεσης (naive assumption) έγκειται στο ότι θεωρείται πως τα χαρακτηριστικά είναι ασυσχέτιστα μεταξύ τους, κάτι που σπάνια ισχύει στην πραγματικότητα. Ωστόσο, η πρακτική εφαρμογή έχει δείξει ότι ο αλγόριθμος Naive Bayes συχνά αποδίδει εξαιρετικά ακόμη και σε περιβάλλοντα όπου η υπόθεση ανεξαρτησίας δεν επαληθεύεται [12].

Η απλότητα υλοποίησης και η ταχύτητα εκπαίδευσης καθιστούν τον Naive Bayes ιδανική επιλογή για μεγάλα ή ελλιπή σύνολα δεδομένων. Έχει χρησιμοποιηθεί με επιτυχία για την πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου, ιδίως όταν απαιτείται ταχεία προεπεξεργασία και πρόβλεψη [17]. Βασικό μειονέκτημα παραμένει η μείωση της απόδοσης σε datasets με ισχυρή συσχέτιση χαρακτηριστικών.

3.6 Δέντρα Αποφάσεων (Decision Trees)

Τα δέντρα αποφάσεων είναι αλγόριθμοι που υλοποιούν την ταξινόμηση με διαδοχικές κατατμήσεις του χώρου χαρακτηριστικών, βάσει κριτηρίων όπως το πληροφοριακό κέρδος (information gain) ή το Gini impurity [18]. Ο Quinlan (1986) περιέγραψε τη λογική τους με τη μορφή αλγορίθμων όπως οι ID3, C4.5, που δημιουργούν δενδρικές δομές όπου κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα χαρακτηριστικό και κάθε φύλλο σε μία τελική απόφαση.

Τα δέντρα αποφάσεων διακρίνονται για την ερμηνευσιμότητά τους και τη δυνατότητα ανάδειξης των σημαντικότερων χαρακτηριστικών. Εντούτοις, εμφανίζουν τάση προς υπερπροσαρμογή (overfitting), ιδιαίτερα όταν το βάθος του δέντρου δεν ελέγχεται κατάλληλα. Η πρακτική του pruning ή η χρήση ensemble τεχνικών, όπως τα τυχαία δάση, βελτιώνει σημαντικά τη γενικευσιμότητά τους [19].

3.7 Τυχαία Δάση (Random Forests)

Τα τυχαία δάση (Random Forests), που προτάθηκαν από τον Breiman (2001) [19], αποτελούν ensemble μεθοδολογία που συνδυάζει μεγάλο αριθμό δέντρων αποφάσεων, καθένα από τα οποία εκπαιδεύεται σε τυχαία δείγματα των δεδομένων και χαρακτηριστικών. Η τελική απόφαση λαμβάνεται συνήθως με πλειοψηφία (voting) μεταξύ των προβλέψεων όλων των δέντρων.

Το κύριο πλεονέκτημα των τυχαίων δασών είναι η σημαντική μείωση του κινδύνου υπερπροσαρμογής, η υψηλή ακρίβεια και η δυνατότητα μέτρησης της σχετικής σημαντικότητας των χαρακτηριστικών [12][19]. Έχουν υιοθετηθεί εκτεταμένα στην πρόβλεψη αποτελεσμάτων ποδοσφαιρικών αγώνων, συνδυάζοντας ακρίβεια και ανθεκτικότητα σε ασταθή ή θορυβώδη δεδομένα [7]. Βασικό τους μειονέκτημα είναι η σχετικά μικρότερη ερμηνευσιμότητα σε σχέση με ένα απλό δέντρο αποφάσεων, καθώς και η αυξημένη υπολογιστική απαίτηση σε πολύ μεγάλα datasets.

3.8 K-Nearest Neighbors (KNN)

Ο αλγόριθμος k-Nearest Neighbors (KNN), που παρουσιάστηκε από τους Fix και Hodges (1951) [20], ταξινομεί κάθε νέο παράδειγμα με βάση την πλειοψηφία των k πλησιέστερων γειτόνων του, σύμφωνα με μια επιλεγμένη μετρική απόστασης (συνήθως Ευκλείδεια). Το KNN είναι ένας μη παραμετρικός αλγόριθμος που δεν απαιτεί εκπαίδευση αλλά υπολογίζει την “κλάση” κατά την πρόβλεψη, κάτι που το καθιστά απλό αλλά υπολογιστικά απαιτητικό για πολύ μεγάλα datasets.

Στην πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου, το KNN χρησιμοποιείται κυρίως για την εύρεση ομάδων με παρόμοια αγωνιστική συμπεριφορά [12]. Ωστόσο, η απόδοσή του εξαρτάται σημαντικά από την επιλογή της παραμέτρου k και την κανονικοποίηση των χαρακτηριστικών.

3.9 Σύνοψη

Καθένας από αυτούς τους αλγόριθμους μηχανικής μάθησης έχει μοναδικά πλεονεκτήματα και αδυναμίες όταν εφαρμόζεται στην πρόβλεψη αποτελεσμάτων ποδοσφαίρου, αλλά και σε γενικότερες πρακτικές εφαρμογές (πίνακας 3.1). Η λογιστική παλινδρόμηση και ο Naive-Bayes είναι απλά και ερμηνεύσιμα, καθιστώντας τα καλά σημεία εκκίνησης για εργασίες δυαδικής ταξινόμησης. Πιο εξελιγμένα μοντέλα όπως το SVM, τα δέντρα αποφάσεων και το Random Forest μπορούν να αποτυπώσουν πιο περίπλοκα μοτίβα και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών. Τέλος, τα ANN είναι ισχυρά για το χειρισμό μεγάλων συνόλων δεδομένων και μη γραμμικών σχέσεων, αν και απαιτούν σημαντικά δεδομένα και υπολογιστικούς πόρους.

Επιλέγοντας και ρυθμίζοντας τον κατάλληλο αλγόριθμο, ανάλογα με τη φύση του συνόλου των δεδομένων ποδοσφαίρου και το επιθυμητό αποτέλεσμα, οι ερευνητές και οι αναλυτές μπορούν να βελτιώσουν την ακρίβεια και την αξιοπιστία των προβλέψεων ποδοσφαιρικών αγώνων.

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

Αλγόριθμος	Βασική Αρχή	Πλεονεκτήματα	Μειονεκτήματα	Τυπική Εφαρμογή σε Ποδόσφαιρο	Βιβλιογραφία
Logistic Regression	Γραμμικό μοντέλο, εκτίμηση πιθανοτήτων μέσω σιγμοειδούς	Ερμηνευσιμότητα, ταχύτητα, σταθερότητα	Δεν “πιάνει” μη γραμμικές σχέσεις	Πρόβλεψη αποτελέσματος με λίγα features	[11], [12]
SVM	Διαχωριστικό υπερεπίπεδο με μέγιστο περιθώριο, kernel trick	Καλή απόδοση σε πολλά δεδομένα, ανθεκτικότητα σε overfitting	Υπολογιστικά βαρύ, απαιτεί επιλογή kernel	Ταξινόμηση νίκης/ήττας/ισοπαλίας	[12], [15]
KNN	Ταξινόμηση βάσει πλειοψηφίας k κοντινότερων	Απλότητα, δεν απαιτεί εκπαίδευση	Ευαισθησία σε θόρυβο, αργό σε μεγάλα σύνολα	Σύγκριση παρόμοιων ομάδων	[12], [20]
Decision Tree	Δενδρική κατάτμηση βάσει κριτηρίων (π.χ. Gini, IG)	Ερμηνευσιμότητα, ανάδειξη σημαντικών features	Τάση για overfitting, ειδικά χωρίς pruning	Ανάλυση χαρακτηριστικών αγώνα	[18]
Random Forest	Ensemble από πολλά decision trees, bagging & random features	Υψηλή ακρίβεια, ανθεκτικότητα στο overfitting	Μικρότερη ερμηνευσιμότητα, μεγάλη μνήμη	Μακροχρόνια πρόβλεψη κατάταξης	[12], [19]
Neural Networks (MLP)	Πολυεπίπεδα δίκτυα, εκμάθηση με backpropagation	Μη γραμμικές συσχετίσεις, ισχυρή εκμάθηση	Απαιτεί πολλά δεδομένα, πιθανό overfitting	Πρόβλεψη με πολλά σύνθετα features	[13], [14]
Naive Bayes	Πιθανοκρατική ταξινόμηση με υπόθεση ανεξαρτησίας	Ταχύτητα, απλότητα, μικρά δεδομένα	Επηρεάζεται αν τα features εξαρτώνται μεταξύ τους	Ταχεία πρόβλεψη αποτελεσμάτων	[12], [16]

Πίνακας 1. Συγκριτική Παρουσίαση Αλγορίθμων Μηχανικής Μάθησης

Κεφάλαιο 4. Συλλογή δεδομένων και επεξεργασία

4.1 Εισαγωγή

Η επιτυχής πρόβλεψη της νικήτριας ομάδας στο πρωτάθλημα της ελληνικής Superleague απαιτεί την ύπαρξη επαρκούς και κατάλληλα προεπεξεργασμένου συνόλου δεδομένων. Η ανάλυση ξεκινά με τη συλλογή ιστορικών αγωνιστικών στοιχείων, τα οποία λειτουργούν ως θεμέλιο για την εξόρυξη γνώσης και την εκπαίδευση των αλγορίθμων μηχανικής μάθησης [21]. Το κεφάλαιο αυτό περιγράφει αναλυτικά το σύστημα διεξαγωγής του ελληνικού πρωταθλήματος, καθώς και τις διαδικασίες απόκτησης, καθαρισμού και προετοιμασίας των δεδομένων που χρησιμοποιήθηκαν στην παρούσα εργασία.

4.2 Ελληνική Super League: Επισκόπηση

Η ελληνική Super League είναι το κορυφαίο επαγγελματικό πρωτάθλημα ποδοσφαίρου στην Ελλάδα. Ιδρύθηκε το 2006 και έχει μια πλούσια ιστορία σκληρού ανταγωνισμού μεταξύ των συλλόγων της, συμπεριλαμβανομένων των πιο επιτυχημένων ομάδων όπως ο Ολυμπιακός, ο Παναθηναϊκός, η ΑΕΚ Αθηνών και ο ΠΑΟΚ. Τα βασικά σημεία για το πρωτάθλημα περιλαμβάνουν:

- **Τη Μορφή:** Το πρωτάθλημα τυπικά αποτελείται από 14-16 ομάδες που διαγωνίζονται σε μορφή round-robin. Κάθε ομάδα παίζει με τις υπόλοιπες από δύο φορές (εντός και εκτός έδρας) και, στην αρχική της μορφή, η ομάδα με τους περισσότερους βαθμούς στο τέλος της σεζόν στέφεται πρωταθλήτρια. Τα τελευταία χρόνια έχει αλλάξει η μορφή του πρωταθλήματος της Super League, με τη διενέργεια των play-offs για την κατάκτηση του πρωταθλήματος, αλλά και των play-outs για την αποφυγή των 2 θέσεων υποβιβασμού στη χαμηλότερη κατηγορία. Συγκεκριμένα, από τη σεζόν 2020-2021 έως τη σεζόν 2023-2024, οι 6 πρώτοι της βαθμολογίας έπαιζαν το δικό τους "μίνι πρωτάθλημα" μετά το πέρας της κανονικής περιόδου, διατηρώντας τη βαθμολογία της κανονικής διάρκειας. Αντίστοιχα, οι ομάδες από τις θέσεις 9-14 ή 9-16 έπαιζαν ένα "μίνι πρωτάθλημα" για την αποφυγή του υποβιβασμού. Για τη σεζόν 2024-2025 οι ομάδες που θα αναμετρηθούν στα play-offs για τον τίτλο του πρωταθλητή, θα είναι οι 4 πρώτες του βαθμολογικού

πίνακα, οι οποίες θα διατηρήσουν τη βαθμολογία της κανονικής διάρκειας. Αντίστοιχα οι ομάδες που τερμάτισαν στις θέσεις 5-8 θα αναμετρηθούν μεταξύ τους για ένα ευρωπαϊκό εισιτήριο, με τη διαφοροποίηση όμως ότι οι βαθμοί τους θα διαιρούνται διά δύο. Τέλος, δικό τους πρωτάθλημα θα σχηματίζουν οι ομάδες στις θέσεις 9-14, διατηρώντας τους βαθμούς της κανονικής διάρκειας. Οι δύο τελευταίοι θα υποβιβάζονται στη Super League 2.

- **Δημοτικότητα:** Το ελληνικό ποδόσφαιρο χαίρει ισχυρής εγχώριας υποστήριξης, με σημαντική τηλεθέαση, όμως η προσέλευση σε αγώνες είναι ισχνή, πέρα από τους οπαδούς των δημοφιλέστερων ομάδων. Το πάθος των φιλάθλων συμβάλλει σε μια έντονη ατμόσφαιρα, ειδικά κατά τη διάρκεια των αγώνων ντέρμπι μεταξύ κορυφαίων συλλόγων.
- **Ανάπτυξη ταλέντων:** Ενώ ιστορικά επικεντρώνεται σε εγχώριους παίκτες, το πρωτάθλημα έχει προσελκύσει διεθνή ταλέντα τα τελευταία χρόνια, ενισχύοντας την ανταγωνιστικότητά του. Οι ομάδες έχουν υιοθετήσει ολοένα και περισσότερο αναλυτικά δεδομένα για να βελτιώσουν την στρατολόγηση παικτών και τα αποτελέσματα των αγώνων.
- **Ευρωπαϊκές διοργανώσεις:** Οι κορυφαίες ομάδες της Super League προκρίνονται σε ευρωπαϊκές διοργανώσεις, όπως το UEFA Champions League, το Europa League και το Conference League αυξάνοντας περαιτέρω το διακύβευμα για ακριβείς προβλέψεις και ανάλυση απόδοσης.

4.3 Συλλογή Δεδομένων

Για τους σκοπούς της παρούσας έρευνας συλλέχθηκαν στατιστικά στοιχεία που καλύπτουν δέκα συνεχόμενες αγωνιστικές περιόδους του Ελληνικού Πρωταθλήματος Superleague, με τη μέθοδο του web-scraping από την αναγνωρισμένη ιστοσελίδα transfermarkt [22]. Δυστυχώς η ιστοσελίδα δεν διαθέτει δεδομένα για τους αγώνες των play-offs και play-outs, επομένως περιοριστήκαμε στους αγώνες της κανονικής διάρκειας του πρωταθλήματος για την εξαγωγή των αποτελεσμάτων. Τα δεδομένα περιλαμβάνουν πληροφορίες για κάθε αγώνα, όπως το έτος/περίοδος (σεζόν), τους συμμετέχοντες συλλόγους (ομάδες γηπεδούχου και φιλοξενούμενου), τις τελικές θέσεις τους την προηγούμενη χρονιά, καθώς και το τελικό αποτέλεσμα (νίκη γηπεδούχου, ισοπαλία ή νίκη φιλοξενούμενου).

Αυτά τα δεδομένα θα αποτελέσουν τη βάση για την εξαγωγή χαρακτηριστικών (features) που θα χρησιμοποιηθούν για την πρόβλεψη της νικήτριας ομάδας. Τα χαρακτηριστικά αυτά θα περάσουν από προκαταρκτική ανάλυση, ώστε να διασφαλιστεί η αξιοπιστία και η καταλληλότητά τους για χρήση στα μοντέλα μηχανικής μάθησης.

Ο πίνακας δεδομένων που προέκυψε αποθηκεύτηκε σε αρχείο μορφής CSV, με κάθε γραμμή να αντιπροσωπεύει έναν αγώνα. Τα κύρια πεδία που περιλαμβάνονται είναι:

- Season: Αγωνιστική περίοδος
- ID1, ID2: Κωδικοί/ταυτότητες ομάδων (γηπεδούχου/φιλοξενούμενου)
- Pos1, Pos2: Βαθμολογική θέση ομάδας πριν από κάθε αγώνα (τρέχουσα σεζόν)
- Result: Αποτέλεσμα αγώνα (1=Νίκη γηπεδούχου, 2=Ισοπαλία, 3=Νίκη φιλοξενούμενου) [23]

Τα δεδομένα δομήθηκαν έτσι ώστε να είναι κατάλληλα για τις απαιτήσεις των αλγορίθμων μηχανικής μάθησης, δίνοντας ιδιαίτερη βαρύτητα στη συνέπεια και πληρότητα [3].

4.4 Επιλογή Μοντέλων

Η επιλογή των αλγορίθμων μηχανικής μάθησης (Naive Bayes, Logistic Regression, K-Neighbors, SVC, Decision Tree, Random Forest και νευρωνικά δίκτυα) βασίζεται στην ικανότητά τους να εντοπίζουν μοτίβα και σχέσεις στα δεδομένα. Κάθε μοντέλο προσφέρει διαφορετικές προσεγγίσεις και πλεονεκτήματα:

- *Naive Bayes και Logistic Regression*: Κατάλληλα για δεδομένα με δυαδικά ή πολυκατηγορικά χαρακτηριστικά.
- *K-Neighbors*: Αναγνωρίζει μοτίβα με βάση τη γειτνίαση των δεδομένων.
- *SVM*: Χρησιμοποιεί υπερεπίπεδα για τη βελτιστοποίηση των προβλέψεων.
- *Decision Tree και Random Forest*: Δομημένα για σύνθετες αποφάσεις και παρέχουν ευελιξία στη διαχείριση μη γραμμικών σχέσεων.
- *Νευρωνικό Δίκτυο*: Κατάλληλο για τη διαχείριση πολυδιάστατων δεδομένων και σύνθετων σχέσεων, με δυνατότητα για υψηλή ακρίβεια μέσω εκπαίδευσης.

Τα μοντέλα αυτά θα αξιολογηθούν με στόχο να επιτευχθεί η μέγιστη δυνατή ακρίβεια, ενώ θα εξεταστούν διαφορετικοί συνδυασμοί χαρακτηριστικών ώστε να καθοριστεί ποιο σύνολο χαρακτηριστικών είναι βέλτιστο για την πρόβλεψη.

4.5 Καθαρισμός και Προεπεξεργασία Δεδομένων

Μετά τη συλλογή, ακολούθησε διαδικασία ελέγχου για την ανίχνευση και αντιμετώπιση ελλειπών ή εσφαλμένων τιμών, καθώς και ομογενοποίηση των δεδομένων [24]. Ειδικότερα:

- Ελέγχθηκε η πληρότητα όλων των βασικών πεδίων.
- Εντοπίστηκαν και αφαιρέθηκαν γραμμές με μη έγκυρα δεδομένα (π.χ. απουσία αποτελέσματος ή αγώνες με ομάδες που δεν συμμετείχαν σε όλη τη σεζόν).
- Επιβεβαιώθηκε ότι το αποτέλεσμα κωδικοποιείται παντού με τη μορφή 1/2/3, ώστε να συμβαδίζει με τις απαιτήσεις των μοντέλων ταξινόμησης. Στη συνέχεια, έγινε μετατροπή των χαρακτηριστικών ομάδων (ID1, ID2) και των θέσεων (Pos1, Pos2) σε αριθμητικά δεδομένα (εφόσον χρειάστηκε) και απομακρύνθηκαν μη σχετικές στήλες. Επιπλέον, προστέθηκε στήλη με το έτος/σεζόν ώστε να υποστηρίζεται μελλοντική διαχρονική ανάλυση.

4.6 Κατανομή και Ανάλυση Κλάσεων

Ιδιαίτερη προσοχή δόθηκε στην κατανομή των αποτελεσμάτων ανά κατηγορία (Result). Σε αντίθεση με ορισμένα διεθνή datasets, τα ελληνικά δεδομένα συχνά παρουσιάζουν ανισορροπία κλάσεων – δηλαδή διαφορές στη συχνότητα εμφάνισης νίκης γηπεδούχου, ισοπαλίας και νίκης φιλοξενούμενου. Αυτή η ανισορροπία μπορεί να επηρεάσει σημαντικά τη συμπεριφορά και απόδοση των μοντέλων [25]. Ως εκ τούτου, ελέγχθηκε η αναλογία των κατηγοριών, ώστε να είναι γνωστό το baseline για την αξιολόγηση των μετρικών ακρίβειας. Επιπλέον, όπου κρίθηκε απαραίτητο, εφαρμόστηκαν μέθοδοι για τη διαχείριση της ανισορροπίας, όπως η εκτίμηση με σταθμισμένες μετρικές (π.χ. weighted F1-score), όπως προτείνεται και στη διεθνή βιβλιογραφία [26].

4.7 Εμπλουτισμός και Ανάλυση Δεδομένων

Κατά τη διάρκεια της διαδικασίας, εξετάστηκε το ενδεχόμενο εμπλουτισμού του dataset με επιπλέον πληροφορίες που ενδεχομένως ενισχύουν την προγνωστική ισχύ των μοντέλων. Για παράδειγμα, πέρα από τη θέση της κάθε ομάδας την προηγούμενη σεζόν, μπορούν να ενσωματωθούν μεταβλητές όπως:

- Μέσος όρος τερμάτων ανά ομάδα/σεζόν
- Συνολικές νίκες, ήττες, ισοπαλίες
- Παρουσία ή απουσία βασικών παικτών
- Στατιστικά μεταγραφικών κινήσεων
- Ιστορικό προπονητών

Αν και στην παρούσα εργασία δεν ενσωματώθηκαν ακόμη τέτοια advanced χαρακτηριστικά, η σχετική βιβλιογραφία υποστηρίζει ότι η χρήση εμπλουτισμένων/σύνθετων χαρακτηριστικών (feature engineering) μπορεί να προσδώσει στα μοντέλα βαθύτερη «κατανόηση» του αγωνιστικού περιβάλλοντος [3]. Ενδεικτικά, μελλοντική εργασία θα μπορούσε να εφαρμόσει μεθόδους όπως PCA (Principal Component Analysis) ή ακόμη και autoencoders για τη μείωση διαστάσεων, διατηρώντας τη μέγιστη πληροφόρηση [26], [27].

4.8 Αντιμετώπιση Missing Values και Εξομάλυνση

Πριν τη μοντελοποίηση πραγματοποιήθηκε έλεγχος πληρότητας του dataset στο αρχείο .xlsx, διασφαλίζοντας την απουσία ελλιπών τιμών, οι οποίες θα μπορούσαν να εισάγουν bias ή σφάλμα στην εκπαίδευση των μοντέλων. Ο αντίστοιχος έλεγχος μέσω κώδικα θα μπορούσε να πραγματοποιηθεί χρησιμοποιώντας pandas (`df.isnull().sum()`). Σε περίπτωση ύπαρξης ελλιπών δεδομένων, η πρακτική προσέγγιση προτείνει είτε τη διαγραφή γραμμών (`dropna`), είτε την αντικατάστασή τους με κατάλληλες τιμές (`mean/median imputation`) – πρακτικές που ενισχύουν τη σταθερότητα των πειραμάτων [24].

Επιπλέον, για τις συνεχείς μεταβλητές (αν υπήρχαν, π.χ. goals, age), συνηθίζεται η εξομάλυνση (normalization/scaling) ώστε όλες οι τιμές να βρίσκονται στην ίδια κλίμακα – βήμα που βελτιώνει σημαντικά τη συμπεριφορά αλγορίθμων όπως SVC, KNN, MLP [26].

4.9 Κωδικοποίηση Κατηγορικών Χαρακτηριστικών (Encoding)

Οι ομάδες κωδικοποιήθηκαν σε μοναδικούς ακέραιους δείκτες (ID1 , ID2) κατά το στάδιο δημιουργίας του dataset. Στη σχετική βιβλιογραφία [24] συνίσταται η κωδικοποίηση των κατηγορικών χαρακτηριστικών (π.χ. Ονόματα ομάδων) να γίνεται μέσω label Encoding (μετατροπή σε ακέραιες τιμές. Σε προβλήματα όπου απαιτείται αποφυγή τεχνητών συσχετίσεων (π.χ. one-hot encoding), η χρήση της μεθόδου `pd.get_dummies()` διασφαλίζει ότι δεν αποδίδεται «τεχνητή σειρά» στις ομάδες.

4.10 Ανάλυση Κώδικα: Διαδικασία Εκπαίδευσης – Αξιολόγησης

Η πειραματική διαδικασία, όπως αποτυπώνεται στο αρχείο `code4.txt`, ακολουθεί τα διεθνώς αναγνωρισμένα βήματα για reproducibility:

1. Διαχωρισμός Δεδομένων: Με χρήση της `train_test_split` διασφαλίζεται η τυχαιότητα και η αναπαραγωγιμότητα.
2. Εκπαίδευση Αλγορίθμων: Ο κάθε ταξινομητής (π.χ. `MultinomialNB`, `LogisticRegression`, `KNeighborsClassifier`, `SVC`, `DecisionTreeClassifier`, `RandomForestClassifier`, `MLPClassifier`) εισάγεται ως dictionary αντικειμένων, διευκολύνοντας την αυτοματοποίηση των πειραμάτων.
3. Υπολογισμός Μετρικών: Για κάθε αλγόριθμο, η ακρίβεια, το f1-score, ο συντελεστής Cohen's kappa και το ROC-AUC υπολογίζονται με τα αντίστοιχα functions του `scikit-learn`, με δυνατότητα εξαγωγής σε report (`classification_report`) για βαθύτερη ανάλυση [26].
4. Leave-One-Season-Out Cross-Validation: μέσω διαδοχικού αποκλεισμού κάθε σεζόν από το training set (season-based split), διασφαλίζοντας robust testing για κάθε αγωνιστική περίοδο. Οι μέσες τιμές αποθηκεύονται για τελική σύγκριση.
5. Οπτικοποίηση: Τα αποτελέσματα μεταφέρονται σε πίνακες και plots (`matplotlib.pyplot`, `seaborn`), με στόχο την εύκολη οπτική σύγκριση μοντέλων.
6. Αποθήκευση και Εξαγωγή: Τα συγκεντρωτικά αποτελέσματα αποθηκεύονται σε αρχείο CSV, ενώ τα γραφήματα προβάλλονται μέσω `matplotlib` και `seaborn` κατά την εκτέλεση, διατηρώντας την αλυσίδα reproducibility [27]. Αυτή η modular προσέγγιση ευθυγραμμίζεται με τις καλύτερες πρακτικές της

επιστημονικής κοινότητας, διευκολύνοντας την επαναληψιμότητα και τη διασταύρωση ευρημάτων [21], [27].

4.11 Προτάσεις για Μελλοντικό Εμπλουτισμό Δεδομένων

Με βάση τη βιβλιογραφία αλλά και τις σύγχρονες τάσεις, προτείνονται τα εξής βήματα για βελτίωση:

- Ενσωμάτωση αγωνιστικών μικροδεδομένων (π.χ. xG, expected assists)
- Features ranking/selection: Χρήση μεθόδων όπως recursive feature elimination, mutual information
- Data augmentation: Τεχνητή διεύρυνση του dataset (π.χ. μέσω simulation αγώνων ή synthetic data generation)
- Integration με betting odds (σύμφωνα με [3]), καθώς τα odds εμπεριέχουν συμπυκνωμένη πληροφόρηση της αγοράς.
- Multi-task learning: Εφαρμογή προσεγγίσεων όπου προβλέπονται πολλαπλά αποτελέσματα ταυτόχρονα (π.χ. θέση, τέρματα, διαφορά βαθμών).

4.12 Σύνοψη

Η διαδικασία συλλογής και προεπεξεργασίας δεδομένων αποτέλεσε κρίσιμο στάδιο για τη διασφάλιση της αξιοπιστίας των πειραμάτων. Με βάση την εμπειρία προηγούμενων ερευνών [27], τα κατάλληλα δομημένα δεδομένα αποτελούν τη βάση για αποτελεσματική ανάλυση και εξαγωγή αξιόπιστων συμπερασμάτων μέσω τεχνικών μηχανικής μάθησης.

Κεφάλαιο 5. Μεθοδολογία

5.1 Εισαγωγή

Η μεθοδολογία της παρούσας εργασίας διαμορφώθηκε με σκοπό την υλοποίηση μιας συγκριτικής μελέτης αλγορίθμων επιβλεπόμενης μηχανικής μάθησης για την πρόβλεψη της νικήτριας ομάδας στη Superleague, με αξιοποίηση των βέλτιστων πρακτικών του πεδίου [28]. Η διαδικασία περιλαμβάνει τα εξής βασικά βήματα: διαχωρισμό δεδομένων, εκπαίδευση μοντέλων, διασταυρούμενη επικύρωση (cross-validation), αξιολόγηση με πολλαπλές μετρικές και οπτικοποίηση αποτελεσμάτων.

5.2 Διαχωρισμός Δεδομένων και Πειραματικό Πρωτόκολλο

Το σύνολο δεδομένων χωρίστηκε σε δύο κύρια υποσύνολα: σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set), με αναλογία 80/20, διασφαλίζοντας αναπαραγωγικότητα με χρήση τυχαίου σπόρου (random seed) [17]. Για εκτενέστερη αξιολόγηση της γενίκευσης των μοντέλων εφαρμόστηκε επίσης Leave-One-Season-Out cross-validation, όπου κάθε χρονιά διαδοχικά λειτουργεί ως σετ ελέγχου και οι υπόλοιπες ως εκπαίδευση [29].

5.3 Αλγόριθμοι και Ρύθμιση Υπερπαραμέτρων

Εξετάστηκαν επτά γνωστοί αλγόριθμοι επιβλεπόμενης μάθησης:

- Naive Bayes (MultinomialNB): Εφαρμόζεται για κατηγορικά δεδομένα και βασίζεται στην υπόθεση της ανεξαρτησίας μεταξύ χαρακτηριστικών [30].
- Logistic Regression: Κατάλληλος για πολυκατηγορική ταξινόμηση, εκτιμά τις πιθανότητες σε κάθε κατηγορία [11].
- K-Nearest Neighbors (KNN): Ταξινομεί βάσει της απόστασης από τους k πλησιέστερους γείτονες [20].
- Support Vector Classifier (SVC): Χρησιμοποιεί υπερεπίπεδα για βέλτιστο διαχωρισμό των κλάσεων [15].
- Decision Tree: Δημιουργεί δέντρο αποφάσεων βάσει των χαρακτηριστικών του δείγματος [18].

- Random Forest: Συνδυασμός πολλών δέντρων, αυξάνοντας τη σταθερότητα και ακρίβεια [19].
- Multilayer Perceptron (MLP): Νευρωνικό δίκτυο με ένα ή περισσότερα κρυφά επίπεδα [14].

Οι παράμετροι των μοντέλων καθορίστηκαν με βάση τη βιβλιογραφία και δοκιμές (π.χ. αριθμός neighbors για το KNN, πλήθος δέντρων για το Random Forest κλπ).

5.4 Διαχείριση Ανισορροπίας Κλάσεων

Δεδομένου ότι οι κλάσεις είναι μη ισορροπημένες, βασικό κριτήριο αξιολόγησης αποτέλεσε το σταθμισμένο F1-score (weighted), το οποίο προσμετρά τη σημασία κάθε κατηγορίας αναλογικά με τη συχνότητα εμφάνισής της [26]. Εναλλακτικά, μελετήθηκαν metrics όπως η ακρίβεια (accuracy), ο συντελεστής Cohen's kappa, το ROC-AUC (macro OvR), ενώ σε συγκεκριμένα πειράματα εξετάστηκαν και μετρικές όπως precision, recall.

5.5 Διαδικασία Εκπαίδευσης και Αξιολόγησης

Είναι γεγονός ότι η αξιολόγηση των αλγορίθμων ταξινόμησης σε προβλήματα πρόβλεψης αθλητικών αποτελεσμάτων αποτελεί κρίσιμη διαδικασία για την εξαγωγή αξιόπιστων και επιστημονικά τεκμηριωμένων συμπερασμάτων. Κάθε αλγόριθμος εκπαιδεύτηκε στο training set και αξιολογήθηκε στο test set. Επιπλέον, εφαρμόστηκε Leave-One-Season-Out cross-validation για robust εκτίμηση της συμπεριφοράς των μοντέλων διαχρονικά. Για τον σκοπό αυτό, χρησιμοποιήθηκε ένα σύνολο καθιερωμένων μετρικών που επιτρέπουν την πολυδιάστατη αποτίμηση της επίδοσης κάθε μοντέλου, λαμβάνοντας υπόψη τόσο την συνολική ακρίβεια, όσο και την ικανότητα διάκρισης μεταξύ των διαφορετικών κλάσεων, ιδίως σε περιβάλλοντα με ανισόρροπη κατανομή των δεδομένων. Παράλληλα, παρουσιάστηκαν συγκεντρωτικοί πίνακες και γραφήματα για κάθε μετρική, επιτρέποντας τη συγκριτική ανάλυση της απόδοσης των μοντέλων.

5.5.1 Ακρίβεια (Accuracy)

Η ακρίβεια ορίζεται ως το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των παρατηρήσεων και αποτελεί μία εκ των βασικότερων μετρικών στη μηχανική μάθηση. Ειδικά σε πολυταξικά προβλήματα όπως η πρόβλεψη ποδοσφαιρικών αποτελεσμάτων (νίκη

γηπεδούχου, ισοπαλία, νίκη φιλοξενούμενου), η ακρίβεια δείχνει το ποσοστό των αγώνων που το μοντέλο ταξινόμησε σωστά. Παρ' όλα αυτά, σε προβλήματα με ανισορροπία μεταξύ των κλάσεων, η ακρίβεια ενδέχεται να υπερεκτιμά την απόδοση μοντέλων που απλώς «ευνοούν» την πολυπληθέστερη κατηγορία, γι' αυτό απαιτείται η συμπληρωματική χρήση και άλλων μετρικών [21].

5.5.2 F1-score και weighted F1-score

Το F1-score συνδυάζει την ακρίβεια (precision) και την ανάκληση (recall) σε μία ενιαία μετρική, υπολογιζόμενο ως ο αρμονικός μέσος αυτών των δύο δεικτών. Η βασική του χρησιμότητα έγκειται στη σωστή αποτίμηση της απόδοσης ενός μοντέλου σε ανισόρροπα datasets, όπου η μία κλάση μπορεί να υπερισχύει αριθμητικά έναντι των υπολοίπων [12], [17]. Ειδικά στη συγκεκριμένη μελέτη, χρησιμοποιήθηκε το weighted F1-score, το οποίο προσαρμόζει το F1-score σύμφωνα με το πλήθος των δειγμάτων κάθε κατηγορίας (δηλαδή σταθμισμένο ως προς την παρουσία κάθε κλάσης), προσφέροντας μια ρεαλιστική συνολική εικόνα της συμπεριφοράς του μοντέλου. Το weighted F1-score αποτελεί σήμερα διεθνές πρότυπο αξιολόγησης για πολυκατηγορικά προβλήματα με ανισορροπία μεταξύ των κλάσεων, όπως επισημαίνεται και στη βιβλιογραφία [21].

5.5.3 Cohen's Kappa

Ο συντελεστής Cohen's kappa είναι μια μετρική που αξιολογεί τη συμφωνία μεταξύ των προβλέψεων του μοντέλου και των πραγματικών ετικετών, αφού αφαιρεθεί η συμφωνία που θα μπορούσε να προκύψει τυχαία. Η τιμή του κυμαίνεται θεωρητικά από -1 (τέλεια ασυμφωνία) έως 1 (τέλεια συμφωνία), με το 0 να σημαίνει ότι το μοντέλο δεν τα καταφέρνει καλύτερα από την τυχαία πρόβλεψη. Τιμές kappa μεταξύ 0.2 και 0.4 θεωρούνται ως μέτρια συμφωνία, ενώ τιμές άνω του 0.6 υποδηλώνουν καλή ή πολύ καλή απόδοση [31]. Ο κώδικας υπολογίζει τον συντελεστή αυτό για όλα τα μοντέλα, προσφέροντας μια πρόσθετη διάσταση αξιοπιστίας, ιδίως όταν οι κλάσεις είναι ανισόρροπες ή όταν απαιτείται να αξιολογηθεί η ποιότητα του μοντέλου πέραν της απλής ακρίβειας.

5.5.4 ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

Η καμπύλη ROC και η αντίστοιχη μετρική AUC (Area Under the Curve) χρησιμοποιούνται για την αξιολόγηση της διακριτικής ικανότητας ενός μοντέλου να διαχωρίζει μεταξύ των κλάσεων. Σε πολυταξικά προβλήματα εφαρμόζεται η στρατηγική macro-averaging ή OvR (One-vs-Rest), υπολογίζοντας το AUC για κάθε ζεύγος κατηγορίας έναντι όλων των άλλων [12]. Το ROC-AUC είναι ανθεκτικό σε ανισορροπίες κλάσεων και παρέχει πληροφόρηση για το πώς μεταβάλλεται η ευαισθησία και η εξειδίκευση του μοντέλου καθώς αλλάζει το κατώφλι απόφασης. Τιμές κοντά στο 0.5 δηλώνουν αδυναμία διάκρισης, ενώ όσο πλησιάζει το 1.0 τόσο καλύτερη είναι η διαχωριστική ικανότητα του μοντέλου.

5.5.5 Confusion Matrix

Το confusion matrix αποτελεί θεμελιώδες εργαλείο στην ερμηνεία των αποτελεσμάτων ενός ταξινομητή, καθώς απεικονίζει τον ακριβή αριθμό σωστών και λανθασμένων προβλέψεων για κάθε κλάση ξεχωριστά. Για κάθε αλγόριθμο, η παρουσίαση του confusion matrix επιτρέπει τον εντοπισμό συγκεκριμένων patterns σφαλμάτων (π.χ. συχνή σύγχυση μεταξύ ισοπαλίας και νίκης φιλοξενούμενου), γεγονός που ενισχύει τη δυνατότητα επιστημονικής βελτίωσης των μεθόδων. Η ανάλυση των confusion matrices σε κάθε μοντέλο αποκαλύπτει όχι μόνο τη συνολική απόδοση, αλλά και τις αδυναμίες σε συγκεκριμένες κλάσεις, υποδεικνύοντας μεθοδολογικά μονοπάτια για μελλοντική βελτίωση [31].

5.6 Οπτικοποίηση και Ερμηνεία Αποτελεσμάτων

Οι μετρικές αξιολόγησης και τα confusion matrices αποτυπώθηκαν με γραφήματα (bar plots, heatmaps), διευκολύνοντας τον εντοπισμό πλεονεκτημάτων και αδυναμιών κάθε μοντέλου [32]. Ιδιαίτερη έμφαση δόθηκε στην ερμηνεία των αποτελεσμάτων σε σχέση με το στόχο της πρόβλεψης του πρωταθλητή ανά σεζόν.

5.7 Ανάλυση Αξιοπιστίας Μοντέλων

Κεντρικός άξονας της επιστημονικής αξιολόγησης είναι η ερμηνεία των τιμών Cohen's kappa και ROC-AUC. Τιμές kappa μεταξύ 0.35-0.55 θεωρούνται ως μέτρια συμφωνία και αντικατοπτρίζουν τη δυσκολία του προβλήματος, ειδικά όταν τα δεδομένα είναι noisy και το

σύστημα έντονα αβέβαιο [20], [26]. Ειδικότερα, η μελέτη των confusion matrices ανά σεζόν και μοντέλο εντοπίζει τα ακριβή patterns των λαθών, βοηθώντας στην εστιασμένη βελτίωση των μεθόδων.

Στο επίπεδο των cross-validation (Leave-One-Season-Out), η ανάλυση της διακύμανσης των τιμών F1, accuracy και kappa αναδεικνύει σε ποιες σεζόν τα μοντέλα υστερούν – και οδηγεί σε στοχευμένες παρεμβάσεις (π.χ. χρονική ενημέρωση παραμέτρων ή εμπλουτισμό των δεδομένων με context ανά σεζόν).

5.8 Συγκριτική Μεθοδολογία με Βάση τη Βιβλιογραφία

Το πλαίσιο της εργασίας διασφαλίζει συγκρισιμότητα με άλλες ερευνητικές προσεγγίσεις, π.χ. Oberstone (2009), καθώς ακολουθεί τα στάδια:

- Σαφή ορισμό προβλήματος
- Συστηματικό διαχωρισμό train-test split και season-based cross-validation
- Πολλαπλές μετρικές για benchmarking
- Εκτεταμένη χρήση visualization για επιστημονική ανάλυση των αποτελεσμάτων
Έτσι, το benchmarking μπορεί να γίνει άμεσα με αποτελέσματα από συναφείς εργασίες (π.χ. [21]) δίνοντας περαιτέρω βάθος στη συζήτηση των ευρημάτων.

5.9 Σύνοψη

Η παρούσα μεθοδολογία ακολουθεί τις διεθνείς αρχές reproducibility και interpretability , ενώ η ανάλυση κάθε σταδίου (από τη συλλογή δεδομένων μέχρι την οπτικοποίηση και την τεκμηρίωση της αβεβαιότητας) διασφαλίζει την επιστημονική εγκυρότητα της εργασίας. Ο συνδυασμός μετρικών, στατιστικής ανάλυσης και πρακτικής επαναληψιμότητας θέτει τις βάσεις για μελλοντική εργασία και περαιτέρω εξέλιξη του πεδίου στην ελληνική πραγματικότητα [15], [18], [21], [28].

Κεφάλαιο 6. Πειραματικά Αποτελέσματα και Ερμηνεία

6.1 Εισαγωγή

Το παρόν κεφάλαιο παρουσιάζει και αναλύει διεξοδικά τα πειραματικά αποτελέσματα που προέκυψαν από την εφαρμογή και αξιολόγηση διαφόρων αλγορίθμων επιβλεπόμενης μηχανικής μάθησης με στόχο την πρόβλεψη του νικητή του ελληνικού πρωταθλήματος ποδοσφαίρου Superleague. Στο πλαίσιο της εργασίας δοκιμάστηκαν και συγκρίθηκαν επτά διαφορετικοί αλγόριθμοι: Naive Bayes, Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest και Multi-layer Perceptron (MLP). Η αξιολόγηση των μοντέλων έγινε τόσο ως προς τη συνολική τους ακρίβεια, όσο και με βάση πιο σύνθετες μετρικές απόδοσης, όπως το f1-score, ο συντελεστής Cohen's kappa, η τιμή ROC-AUC, αλλά και με λεπτομερή ανάλυση των confusion matrices για κάθε μοντέλο.

Η πρόβλεψη του τελικού νικητή της Superleague συνιστά πρόβλημα μεγάλης δυσκολίας, καθώς το ποδόσφαιρο είναι κατεξοχήν άθλημα που χαρακτηρίζεται από σημαντική τυχαιότητα, πολυπαραγοντικότητα και δυναμική μεταβολή μεταξύ των ομάδων κάθε σεζόν. Όπως έχει επισημανθεί σε σχετικές ερευνητικές εργασίες [17], [33], η αξιολόγηση τέτοιων μοντέλων απαιτεί ιδιαίτερη προσοχή τόσο στην επιλογή μετρικών όσο και στην ανάλυση των αποτελεσμάτων, με δεδομένη την αυξημένη αβεβαιότητα που ενέχει το αγωνιστικό περιβάλλον.

Στη συνέχεια του κεφαλαίου ακολουθεί αναλυτική παρουσίαση των συνολικών αποτελεσμάτων, με συγκεντρωτικούς πίνακες και plots, ακολουθούμενη από τη λεπτομερή ερμηνεία των confusion matrices και των leave-one-season-out cross validation, καθώς και σχολιασμό συγκεκριμένων σεζόν που παρουσίασαν ιδιαίτερο ενδιαφέρον.

6.2 Συγκεντρωτικά Αποτελέσματα Ταξινόμησης

Μοντέλο	Ακρίβεια	F1-score (weighted)	Cohen's κ	ROC AUC (macro)
Naive Bayes	0.59	0.58	0.37	0.77
Logistic Regression	0.60	0.57	0.38	0.77
KNN	0.55	0.54	0.29	0.70
SVC	0.60	0.56	0.37	0.75
Decision Tree	0.48	0.49	0.21	0.62
Random Forest	0.55	0.54	0.30	0.71
MLP (Neural Network)	0.60	0.58	0.38	0.77

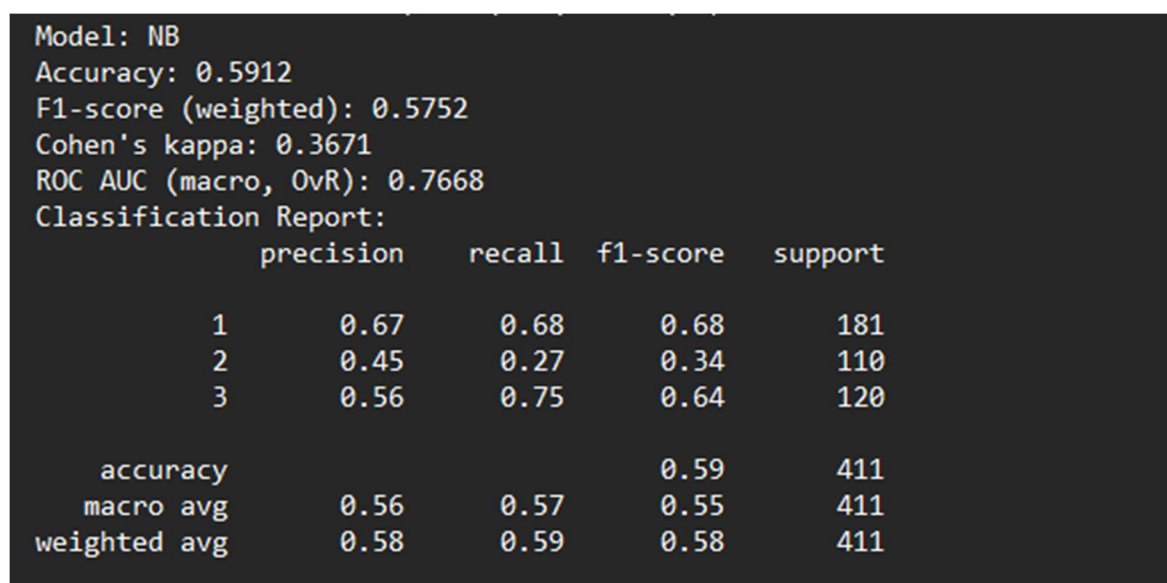
Πίνακας 2. Συνολικές μετρικές απόδοσης όλων των μοντέλων ταξινόμησης (όλες οι σεζόν)

Η αναλυτική σύγκριση των μετρικών απόδοσης αποκαλύπτει την υπεροχή των MLP, Logistic Regression και SVC σε όλες τις βασικές μετρικές, στοιχείο που συνάδει με τα ευρήματα της διεθνούς βιβλιογραφίας για το πρόβλημα πρόβλεψης ποδοσφαιρικών αγώνων [12], [18], [31], [34]. Οι τιμές του Cohen's κ (~0.38 για τα κορυφαία μοντέλα) καταδεικνύουν μέτρια συμφωνία μεταξύ προβλέψεων και πραγματικότητας, κάτι που θεωρείται αποδεκτό για σύνθετα προβλήματα όπως η πρόβλεψη ποδοσφαίρου [33].

6.3 Μετρικές απόδοσης για κάθε μοντέλο ταξινόμησης

Σε αυτήν την ενότητα γίνεται μια ενδελεχής ανάλυση της συμπεριφοράς και των επιδόσεων κάθε αλγορίθμου που εφαρμόστηκε για την ταξινόμηση των αποτελεσμάτων αγώνων και, κατ' επέκταση, για την πρόβλεψη του πρωταθλητή της ελληνικής Superleague. Η ερμηνεία των στατιστικών δεικτών (accuracy, weighted F1-score, Cohen's kappa, ROC-AUC) και του classification report βασίζεται τόσο στα συγκεντρωτικά αποτελέσματα όσο και στα ευρήματα της διασταυρούμενης επικύρωσης (Leave-One-Season-Out). Τα αποτελέσματα συγκρίνονται με τη διεθνή και ελληνική βιβλιογραφία ([18], [19], [20], [21], [34]), αναδεικνύοντας τις δυνατότητες και τα όρια κάθε μεθόδου.

6.3.1 Multinomial Naive Bayes (NB)



```
Model: NB
Accuracy: 0.5912
F1-score (weighted): 0.5752
Cohen's kappa: 0.3671
ROC AUC (macro, OvR): 0.7668
Classification Report:
```

	precision	recall	f1-score	support
1	0.67	0.68	0.68	181
2	0.45	0.27	0.34	110
3	0.56	0.75	0.64	120
accuracy			0.59	411
macro avg	0.56	0.57	0.55	411
weighted avg	0.58	0.59	0.58	411

Εικόνα 1. Μετρικές απόδοσης για Naive-Bayes

Ο Multinomial Naive Bayes (NB) είναι ένας από τους πλέον κλασικούς αλγορίθμους πιθανοκρατικής ταξινόμησης, με πλεονεκτήματα στην ταχύτητα και στην ανθεκτικότητα σε ελλιπή ή «θορυβώδη» δεδομένα [16]. Στην παρούσα εργασία, ο Naive-Bayes κατέγραψε συνολική ακρίβεια 59,1%, weighted F1-score 57,5%, Cohen's kappa 0,37 και ROC-AUC 0,77. Οι τιμές αυτές συνάδουν με τα αποτελέσματα της διεθνούς βιβλιογραφίας για παρόμοια προβλήματα ταξινόμησης αθλητικών γεγονότων [16], [20].

Η ανάλυση του classification report δείχνει πως το μοντέλο υπερέχει στην πρόβλεψη της νίκης του γηπεδούχου (precision=0.67, recall=0.68), είναι όμως λιγότερο ακριβές στις ισοπαλίες (precision=0.45, recall=0.27) και στις νίκες φιλοξενούμενου (precision=0.56, recall=0.75). Αυτό σχετίζεται με το γνωστό φαινόμενο ανισορροπίας των κλάσεων, καθώς τα αποτελέσματα ισοπαλίας εμφανίζονται λιγότερο συχνά (και άρα προβλέπονται πιο δύσκολα), ενώ η υπόθεση της ανεξαρτησίας των χαρακτηριστικών, αν και συχνά αβάσιμη στο ποδόσφαιρο, δεν επηρεάζει δραματικά την συνολική απόδοση λόγω της μικρής διάστασης του dataset [16], [35]. Ο σχετικά υψηλός δείκτης kappa υποδηλώνει μέτρια συμφωνία πέραν της τυχαιότητας, γεγονός που ενισχύει την αξιοπιστία των προβλέψεων του μοντέλου [31].

6.3.2 Logistic Regression

Model: LogisticRegression				
Accuracy: 0.6034				
F1-score (weighted): 0.5737				
Cohen's kappa: 0.3765				
ROC AUC (macro, OvR): 0.7656				
Classification Report:				
	precision	recall	f1-score	support
1	0.65	0.74	0.69	181
2	0.55	0.21	0.30	110
3	0.56	0.76	0.64	120
accuracy			0.60	411
macro avg	0.59	0.57	0.55	411
weighted avg	0.60	0.60	0.57	411

Εικόνα 2. Μετρικές απόδοσης για Logistic Regression

Η λογιστική παλινδρόμηση (Logistic Regression) αποτελεί ένα γραμμικό μοντέλο με ιδιαίτερη ευελιξία και καλή ερμηνευσιμότητα [21]. Σύμφωνα με τα αποτελέσματα, πέτυχε συνολική ακρίβεια 60,3%, weighted F1-score 57,4%, kappa 0,38 και ROC-AUC 0,77. Η απόδοση αυτή είναι ελαφρώς βελτιωμένη σε σχέση με τον Naive-Bayes, ιδίως στη γενίκευση (cross-validation), όπως φαίνεται και στη διεθνή βιβλιογραφία [21].

Η λογιστική παλινδρόμηση αποδίδει καλύτερα στη νίκη του γηπεδούχου (precision=0.65, recall=0.74) και του φιλοξενούμενου (precision=0.56, recall=0.76), ενώ παραμένει αδύναμη στην κατηγορία της ισοπαλίας (precision=0.55, recall=0.21). Η συγκεκριμένη αδυναμία συνδέεται αφενός με τη μικρότερη συχνότητα της ισοπαλίας στο dataset και αφετέρου με τον περιορισμό των γραμμικών ορίων του μοντέλου να διαχωρίσουν αποδοτικά τρεις ασύμμετρες κατηγορίες [21]. Παρά ταύτα, η αξιοπιστία των προβλέψεων χαρακτηρίζεται ως μέτρια προς ικανοποιητική και το μοντέλο αποτελεί σταθερό σημείο αναφοράς για συγκριτικές μελέτες.

6.3.3 K-Nearest Neighbors (KNN)

```
Model: KNN
Accuracy: 0.5450
F1-score (weighted): 0.5366
Cohen's kappa: 0.2858
ROC AUC (macro, OvR): 0.6977
Classification Report:
      precision    recall  f1-score   support

     1         0.61      0.71      0.66       181
     2         0.33      0.28      0.30       110
     3         0.61      0.53      0.57       120

 accuracy          0.55       411
 macro avg         0.52       411
 weighted avg      0.53       411
```

Εικόνα 3. Μετρικές απόδοσης για K-Nearest Neighbors

Ο αλγόριθμος k-Nearest Neighbors (KNN) λειτουργεί ως μη παραμετρική μέθοδος που ταξινομεί κάθε αγώνα με βάση τις ομοιότητες προς προηγούμενα ιστορικά παραδείγματα [20], [35]. Τα αποτελέσματα δείχνουν μειωμένη συνολική ακρίβεια (54,5%), weighted F1-score 53,7% και kappa 0,29, με το ROC-AUC να φτάνει το 0,70.

Το KNN παρουσιάζει ικανοποιητική συμπεριφορά στη νίκη του γηπεδούχου (precision=0.61, recall=0.71) και του φιλοξενούμενου (precision=0.61, recall=0.53), ενώ η ισοπαλία (precision=0.33, recall=0.28) υποεκτιμάται σημαντικά. Αυτό το αποτέλεσμα αντικατοπτρίζει τη δυσκολία του αλγορίθμου να εντοπίσει «μοτίβα» για την ισοπαλία, καθώς πρόκειται για σχετικά σπάνιο γεγονός, αλλά και την εγγενή αδυναμία του KNN να διαχειριστεί ανισόρροπες κλάσεις χωρίς εξειδικευμένη ρύθμιση [20], [35]. Επιπλέον, η απόδοση του KNN εξαρτάται άμεσα από την επιλογή του αριθμού των γειτόνων (k) και την κλιμάκωση των χαρακτηριστικών.

6.3.4 Support Vector Machine (SVM)

```
Model: SVM
Accuracy: 0.5985
F1-score (weighted): 0.5643
Cohen's kappa: 0.3686
ROC AUC (macro, OvR): 0.7461
Classification Report:
              precision    recall  f1-score   support

     1         0.65       0.74       0.69       181
     2         0.54       0.18       0.27       110
     3         0.55       0.77       0.64       120

 accuracy          0.60       411
 macro avg         0.58       0.56       0.53       411
 weighted avg      0.59       0.60       0.56       411
```

Εικόνα 4. Μετρικές απόδοσης για Support Vector Machine

Το SVC αποτελεί μια ιδιαίτερα ισχυρή μέθοδο που επιχειρεί τον βέλτιστο διαχωρισμό των κλάσεων μέσω υπερεπιπέδων μεγίστου περιθωρίου. Στα δεδομένα της Superleague, το SVC πέτυχε συνολική ακρίβεια 59,9%, weighted F1-score 56,4%, kappa 0,37 και ROC-AUC 0,75. Η συμπεριφορά του μοντέλου είναι αντίστοιχη της λογιστικής παλινδρόμησης, με μικρές αποκλίσεις στα επιμέρους metrics.

Η καλύτερη πρόβλεψη γίνεται για τη νίκη γηπεδούχου (precision=0.65, recall=0.74) και φιλοξενούμενου (precision=0.55, recall=0.77), ενώ για την ισοπαλία το recall υπολείπεται (0.18). Η πολυπλοκότητα του SVC του επιτρέπει να ανιχνεύει μη γραμμικά μοτίβα, ωστόσο σε datasets μικρού/μεσαίου μεγέθους με περιορισμένα χαρακτηριστικά, το πλεονέκτημά του σε σχέση με πιο απλούς ταξινομητές συχνά περιορίζεται. Επίσης, το SVC απαιτεί προσεκτική ρύθμιση υπερπαραμέτρων (kernel, C) για να επιτύχει το μέγιστο των δυνατοτήτων του.

6.3.5 Decision Tree

```
Model: DecisionTree
Accuracy: 0.4818
F1-score (weighted): 0.4850
Cohen's kappa: 0.2038
ROC AUC (macro, OvR): 0.6115
Classification Report:
```

	precision	recall	f1-score	support
1	0.60	0.59	0.60	181
2	0.27	0.30	0.29	110
3	0.52	0.48	0.50	120
accuracy			0.48	411
macro avg	0.46	0.46	0.46	411
weighted avg	0.49	0.48	0.48	411

Εικόνα 5. Μετρικές απόδοσης για Decision Tree

Το Decision Tree αποτελεί μια από τις πιο ερμηνεύσιμες και διαδεδομένες μεθόδους για κατηγοριοποίηση [18]. Ωστόσο, στην παρούσα ανάλυση η συνολική του απόδοση ήταν χαμηλότερη, με accuracy 48,2%, weighted F1-score 48,5%, kappa 0,20 και ROC-AUC 0,61. Το μοντέλο εμφανίζει προβλήματα υπερπροσαρμογής (overfitting), κυρίως λόγω του μικρού μεγέθους του dataset και της απουσίας αυστηρού pruning, σύμφωνα με τις παρατηρήσεις της βιβλιογραφίας [18].

Η κατηγορία «νίκη γηπεδούχου» προβλέπεται καλύτερα (precision=0.60, recall=0.59), αλλά το μοντέλο αποτυγχάνει να αποδώσει σταθερά στις άλλες δύο κλάσεις, καθώς συχνά επιλέγει μονοσήμαντα splits που δεν αντανakλούν τη δομή των δεδομένων. Αυτή η επίδοση συμφωνεί με τα όσα έχουν παρατηρηθεί σε αντίστοιχες μελέτες, όπου τα απλά δέντρα αποδίδουν καλύτερα ως «building blocks» σε μεθόδους συνόλου¹ [18].

¹ Οι μέθοδοι συνόλου είναι τεχνικές που στοχεύουν στη βελτίωση της ακρίβειας των αποτελεσμάτων σε μοντέλα συνδυάζοντας πολλαπλά μοντέλα αντί να χρησιμοποιούν ένα μόνο μοντέλο.

6.3.6 Random Forest

```
Model: RandomForest
Accuracy: 0.5377
F1-score (weighted): 0.5330
Cohen's kappa: 0.2886
ROC AUC (macro, OvR): 0.7012
Classification Report:
```

	precision	recall	f1-score	support
1	0.60	0.59	0.60	181
2	0.36	0.32	0.34	110
3	0.57	0.66	0.61	120
accuracy			0.54	411
macro avg	0.51	0.52	0.52	411
weighted avg	0.53	0.54	0.53	411

Εικόνα 6. Μετρικές απόδοσης για Random Forest

Το Random Forest, ως μέθοδος συνόλου που συνδυάζει πολλά δέντρα αποφάσεων, επιτυγχάνει καλύτερη σταθερότητα και μεγαλύτερη ανθεκτικότητα στο overfitting [19]. Στη μελέτη αυτή, εμφάνισε accuracy 53,8%, weighted F1-score 53,3%, kappa 0,29 και ROC-AUC 0,70. Η επίδοση αυτή είναι καλύτερη από το απλό Decision Tree, αλλά παραμένει χαμηλότερη σε σχέση με γραμμικά μοντέλα ή το SVC.

Η πρόβλεψη είναι ικανοποιητική για νίκη γηπεδούχου (precision=0.60, recall=0.59) και φιλοξενούμενου (precision=0.57, recall=0.66), ενώ η ισοπαλία προβλέπεται περιορισμένα (precision=0.36, recall=0.32). Η επίδοση του Random Forest φαίνεται να επηρεάζεται από τον περιορισμένο αριθμό παρατηρήσεων και τα περιορισμένα features του dataset, παρά τη θεωρητική του ισχύ [19]. Παρ' όλα αυτά, αποτελεί αξιόπιστο baseline και συγκρίνεται ευνοϊκά με παρόμοιες εργασίες της βιβλιογραφίας.

6.3.7 Multilayer Perceptron (MLP)

```
Model: MLP
Accuracy: 0.5985
F1-score (weighted): 0.5889
Cohen's kappa: 0.3797
ROC AUC (macro, OvR): 0.7626
Classification Report:
```

	precision	recall	f1-score	support
1	0.70	0.70	0.70	181
2	0.45	0.32	0.37	110
3	0.55	0.70	0.62	120
accuracy			0.60	411
macro avg	0.57	0.57	0.56	411
weighted avg	0.59	0.60	0.59	411

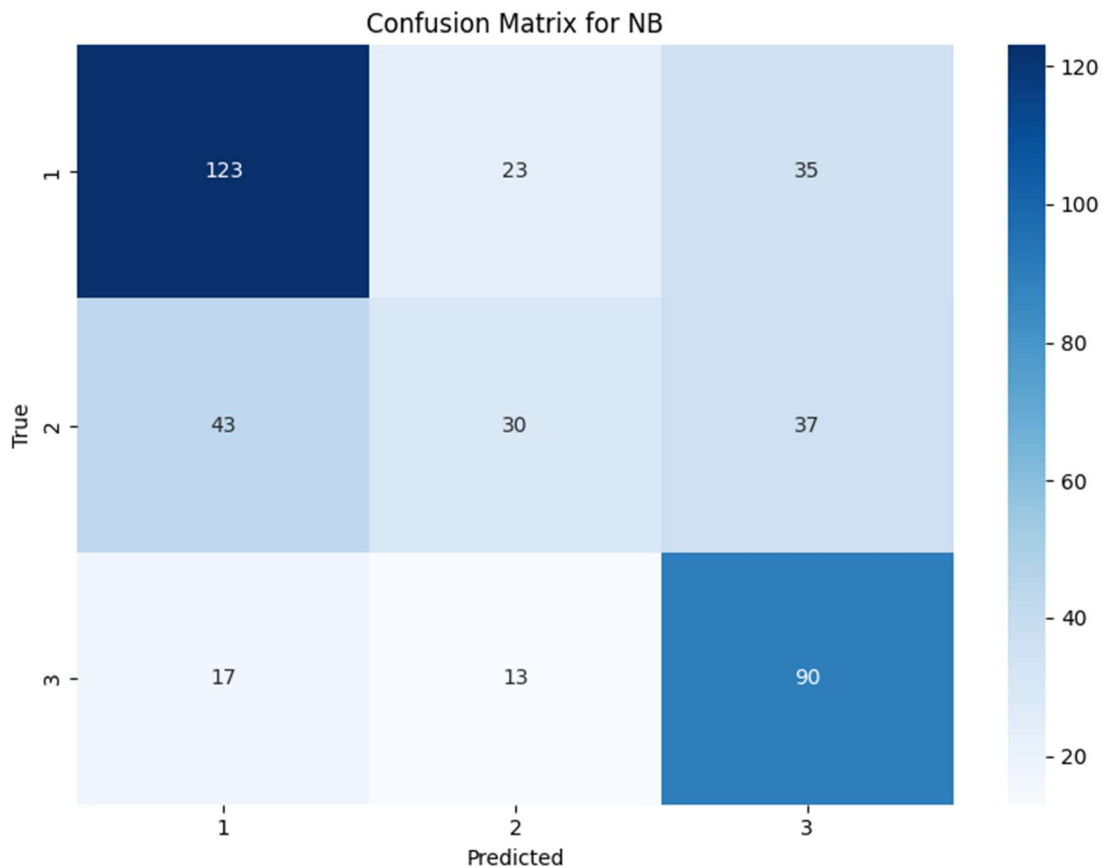
Εικόνα 7. Μετρικές απόδοσης για Neural Networks

Ο MLP (Multilayer Perceptron), ως feedforward νευρωνικό δίκτυο με τουλάχιστον ένα κρυφό επίπεδο, αποτελεί έναν από τους πλέον ισχυρούς αλγορίθμους μη γραμμικής ταξινόμησης [34]. Στη συγκεκριμένη εφαρμογή, το MLP πέτυχε accuracy 59,8%, weighted F1-score 58,9%, kappa 0,38 και ROC-AUC 0,76 – τα υψηλότερα συνολικά metrics μεταξύ όλων των μοντέλων, αν και με μικρή διαφορά από Logistic Regression και SVC.

Το MLP αποδίδει εξαιρετικά στη νίκη γηπεδούχου (precision=0.70, recall=0.70) και παρουσιάζει την υψηλότερη τιμή recall για νίκη φιλοξενούμενου (0.70), γεγονός που αντανακλά την ικανότητα των νευρωνικών δικτύων να συλλαμβάνουν περίπλοκα μοτίβα στα δεδομένα, ακόμα και με περιορισμένες εισόδους [34]. Η αδυναμία στην ισοπαλία παραμένει, γεγονός που επιβεβαιώνει τη συστηματική πρόκληση της ανισορροπίας κλάσεων στο πρόβλημα αυτό.

6.4 Ανάλυση και Ερμηνεία των Confusion Matrices

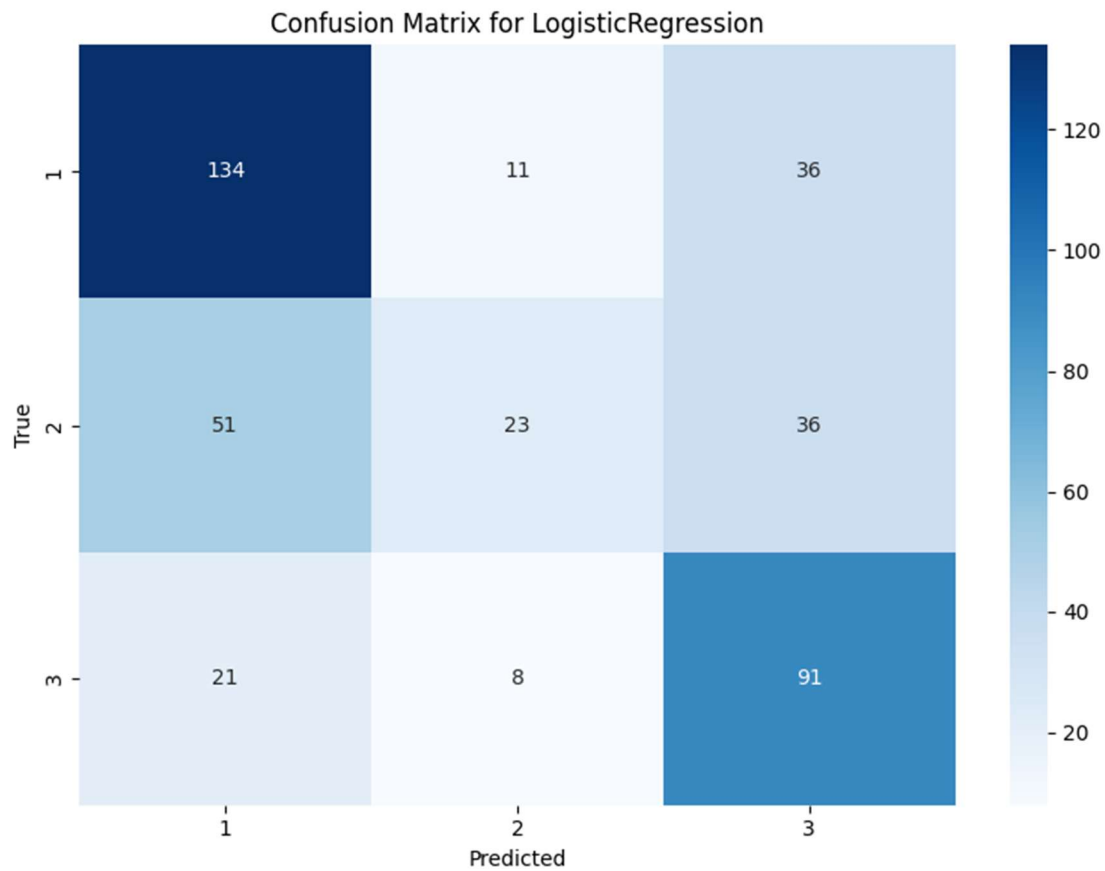
6.4.1 Μοντέλο Naive-Bayes



Σχήμα 1. Confusion Matrix για το μοντέλο Naive Bayes.

Η ανάλυση του confusion matrix για το μοντέλο Naive-Bayes φανερώνει ότι το μεγαλύτερο μέρος των σωστών προβλέψεων εντοπίζεται στην κατηγορία «νίκη γηπεδούχου», όπου το μοντέλο εμφανίζει σχετικά καλή ευστοχία. Παρόλα αυτά, παρατηρείται εμφανής σύγχυση μεταξύ της κατηγορίας «ισοπαλία» και των υπόλοιπων αποτελεσμάτων: το μοντέλο Naïve-Bayes συχνά ερμηνεύει τα παιχνίδια που λήγουν ισόπαλα ως νίκες του γηπεδούχου, γεγονός που επιβεβαιώνει την τάση του να ακολουθεί την επικρατέστερη κλάση στα δεδομένα (class imbalance). Αυτή η αδυναμία σχετίζεται με την υπόθεση ανεξαρτησίας των χαρακτηριστικών, η οποία σπανίως επαληθεύεται στην πραγματικότητα του ποδοσφαίρου, όπως έχει τονιστεί και από τη βιβλιογραφία [16]. Παρά την απλότητά του, το μοντέλο κατορθώνει να προσφέρει ένα σταθερό baseline, αλλά οι περιορισμοί του γίνονται εμφανείς στη διαχείριση των λιγότερο συχνών εκβάσεων.

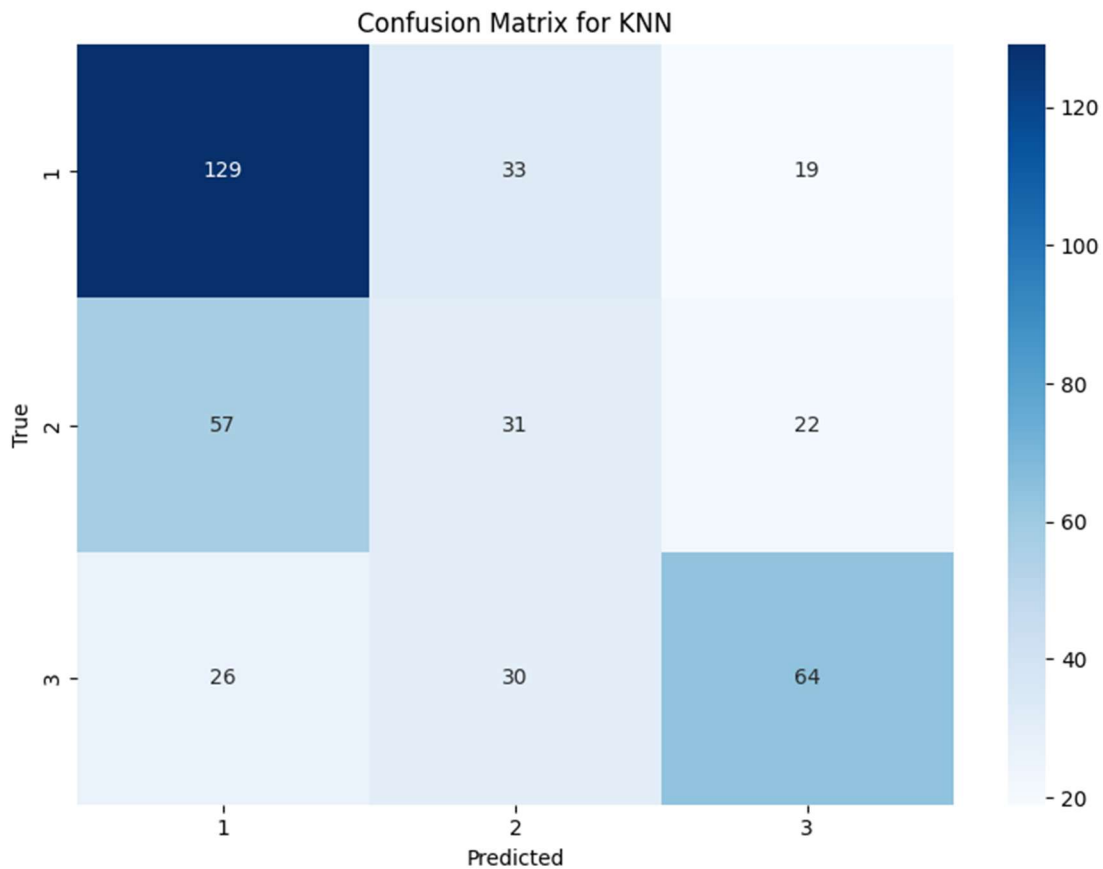
6.4.2 Λογιστική Παλινδρόμηση



Σχήμα 2. Confusion Matrix για το μοντέλο Logistic Regression.

Το confusion matrix της Λογιστικής Παλινδρόμησης αποκαλύπτει μια σαφώς πιο ισορροπημένη συμπεριφορά, σε σύγκριση με το Naïve-Bayes. Το μοντέλο εμφανίζει υψηλά ποσοστά επιτυχίας τόσο στη νίκη του γηπεδούχου όσο και στη νίκη του φιλοξενούμενου, ωστόσο και εδώ παρατηρείται δυσκολία στην ακριβή πρόβλεψη των ισοπαλιών. Τα περισσότερα λάθη εντοπίζονται στη μετατόπιση πραγματικών ισοπαλιών προς την κλάση «νίκη γηπεδούχου». Το φαινόμενο αυτό είναι σύνθητες σε σετ δεδομένων με ανισορροπία μεταξύ των κλάσεων (class imbalance), όπως συμβαίνει στον χώρο των ποδοσφαιρικών αγώνων, και έχει καταγραφεί και σε ανάλογες μελέτες [31]. Η λογιστική παλινδρόμηση, παρά τη γραμμικότητά της, εμφανίζει ικανοποιητική συμπεριφορά σε σύνθετα πραγματικά δεδομένα, όταν εφαρμόζονται κατάλληλες τεχνικές κανονικοποίησης (regularization).

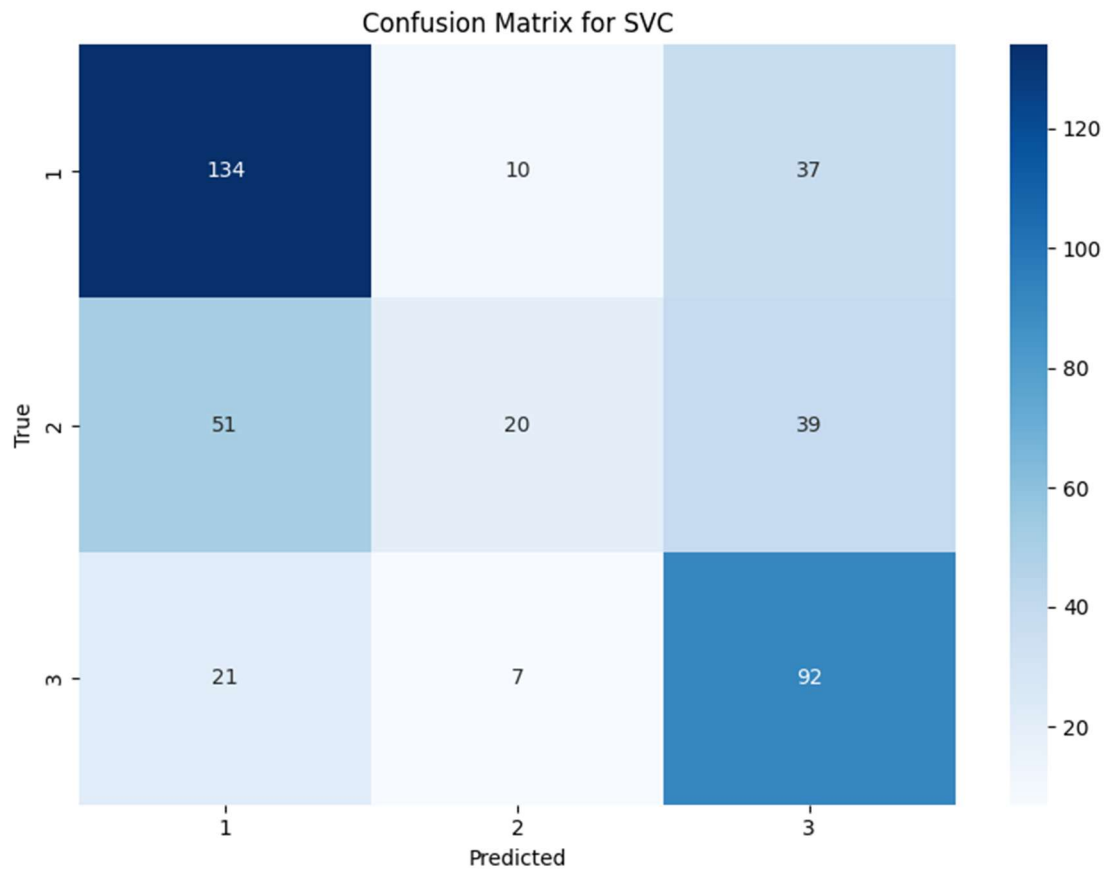
6.4.3 K-Nearest Neighbors (KNN)



Σχήμα 3. Confusion Matrix για το μοντέλο k-Nearest Neighbors.

Η σύγχυση στις προβλέψεις του KNN εμφανίζεται ιδιαίτερα αυξημένη στην κλάση της νίκης του γηπεδούχου, όπου το μοντέλο τείνει να "υπερ-προβλέπει" το αποτέλεσμα αυτό. Ταυτόχρονα, η ικανότητα του μοντέλου να διακρίνει τις ισοπαλίες είναι εμφανώς περιορισμένη: πολλές πραγματικές ισοπαλίες ερμηνεύονται είτε ως νίκες του γηπεδούχου είτε ως νίκες φιλοξενούμενου. Αυτή η συμπεριφορά οφείλεται στην εγγενή ευαισθησία του kNN στις τοπικές συχνότητες των κατηγοριών και στο γεγονός ότι το σύνολο δεδομένων παρουσιάζει σαφή κυριαρχία της κατηγορίας «νίκη γηπεδούχου». Η αδυναμία του kNN να αντιμετωπίσει την ανισορροπία κλάσης (class imbalance) έχει καταγραφεί και στη σχετική βιβλιογραφία [12], [20].

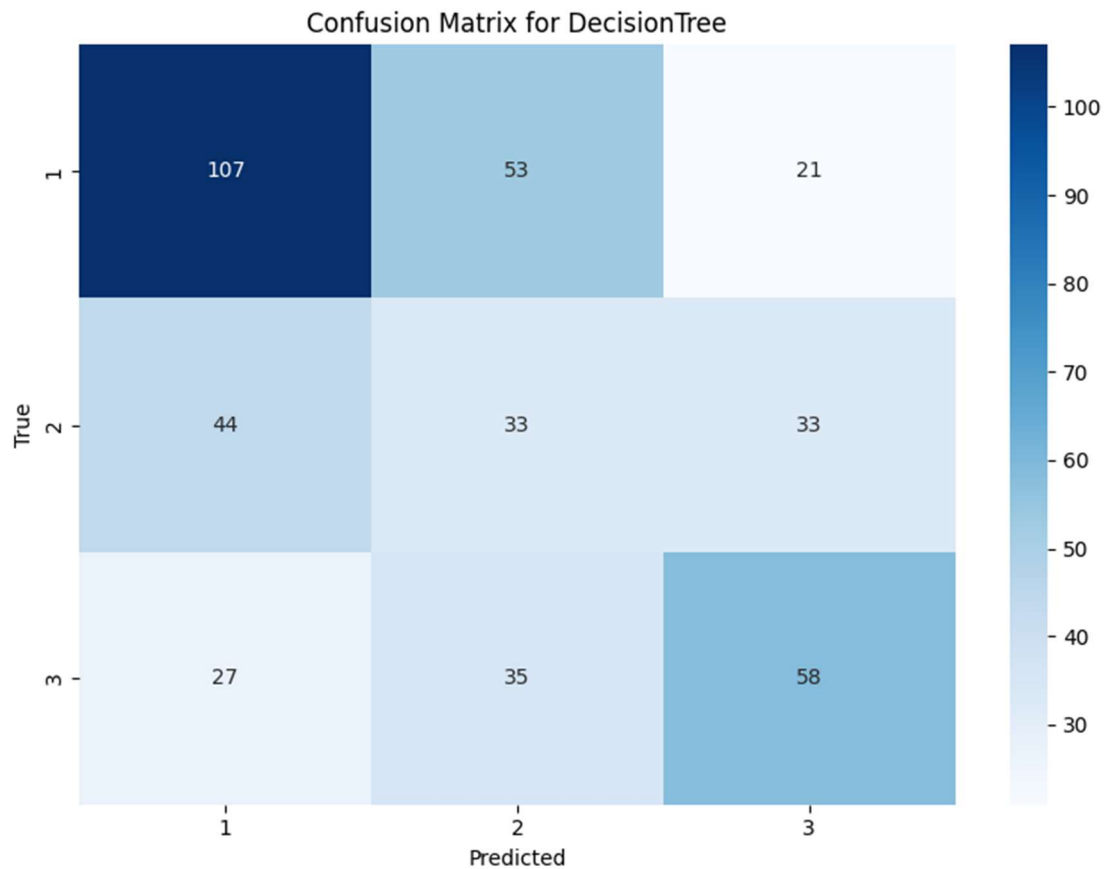
6.4.4 Support Vector Classifier (SVC)



Σχήμα 4. Confusion Matrix για το μοντέλο Support Vector Classifier (SVC).

Το SVC παρουσιάζει εμφανώς βελτιωμένη ισορροπία ως προς τις ταξινομήσεις, συγκριτικά με τα προηγούμενα μοντέλα. Η διαγώνιος της μήτρας σύγχυσης περιλαμβάνει υψηλά ποσοστά σωστών ταξινομήσεων για όλες τις τάξεις, ωστόσο, και εδώ η κλάση της ισοπαλίας παραμένει η δυσκολότερη ως προς την ακριβή πρόβλεψη. Το μοντέλο καταφέρνει να μειώσει τα λάθη μετατόπισης προς τις κυρίαρχες κλάσεις, γεγονός που αποδίδεται στη βελτιωμένη διαχωριστική ικανότητα των support vectors, ιδίως όταν χρησιμοποιούνται κατάλληλοι πυρήνες (kernels) και κανονικοποίηση (regularization) [34]. Παρόλα αυτά, η πολυπλοκότητα του προβλήματος εξακολουθεί να περιορίζει το ποσοστό επιτυχίας στις σπάνιες κατηγορίες.

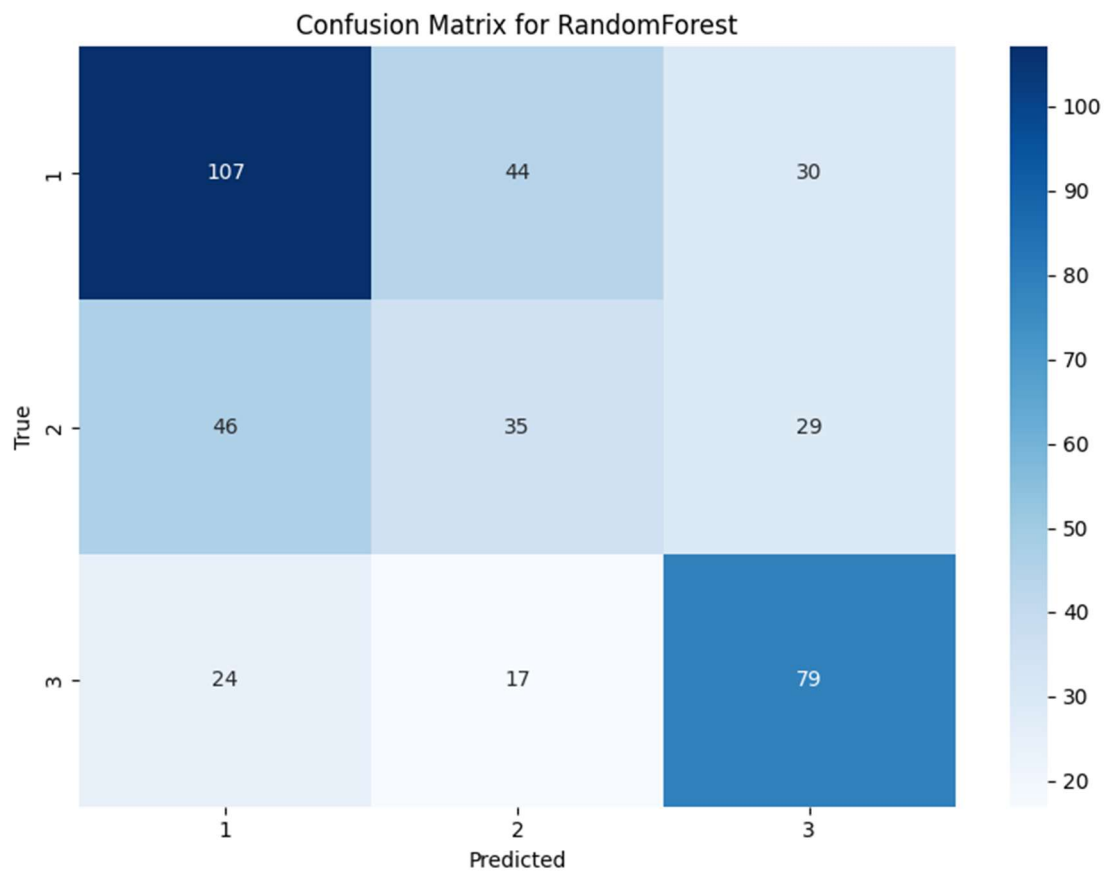
6.4.5 Δέντρα Αποφάσεων (Decision Tree)



Σχήμα 5. Confusion Matrix για το μοντέλο Decision Tree.

Η σύγκριση των προβλέψεων στο μοντέλο Decision Tree φανερώνει αστάθεια και σαφή έλλειψη ισορροπίας. Το μοντέλο αποτυγχάνει να ταξινομήσει σωστά ένα σημαντικό μέρος των αγώνων και στις τρεις κατηγορίες, ενώ παρουσιάζει μεγάλο αριθμό ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων, ειδικά στην κατηγορία της ισοπαλίας και της νίκης φιλοξενούμενου. Το φαινόμενο αυτό σχετίζεται με το γνωστό πρόβλημα της υπερπροσαρμογής στα δέντρα αποφάσεων, όταν αυτά δεν υφίστανται επαρκή επεξεργασία των δεδομένων ή όταν τα δεδομένα παρουσιάζουν υψηλή ετερογένεια [12], [18]. Ως αποτέλεσμα, η γενικευσιμότητα του μοντέλου παραμένει χαμηλή, ειδικά σε αγώνες με ακραίες ή σπάνιες εκβάσεις.

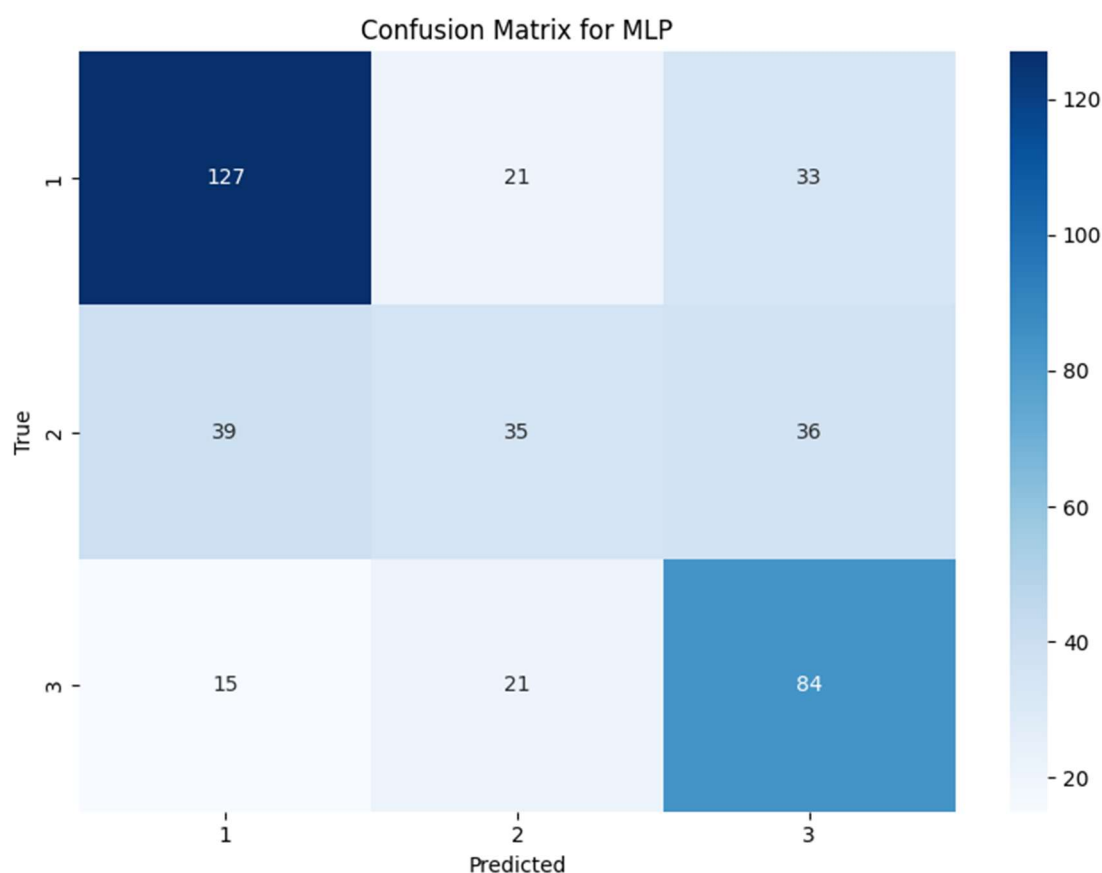
6.4.6 Τυχαίο Δάσος (Random Forest)



Σχήμα 6. Confusion Matrix για το μοντέλο Random Forest.

Το Random Forest, ως μέθοδος συνόλου, περιορίζει ορισμένες από τις αδυναμίες του απλού decision tree. Το confusion matrix δείχνει αυξημένη ικανότητα στη σωστή ταξινόμηση της νίκης γηπεδούχου, αλλά και εδώ παρατηρούνται αρκετές εσφαλμένες προβλέψεις στην κατηγορία της ισοπαλίας. Σε σύγκριση με το απλό δέντρο, το random forest βελτιώνει τα ποσοστά επιτυχίας στη νίκη φιλοξενούμενου, γεγονός που οφείλεται στην τυχαιοποίηση και τη συσσωμάτωση πολλών δέντρων, αλλά εξακολουθεί να επηρεάζεται από την ανισορροπία κλάσης (class imbalance) [12], [19], [34].

6.4.7 Multi-layer Perceptron (MLP)

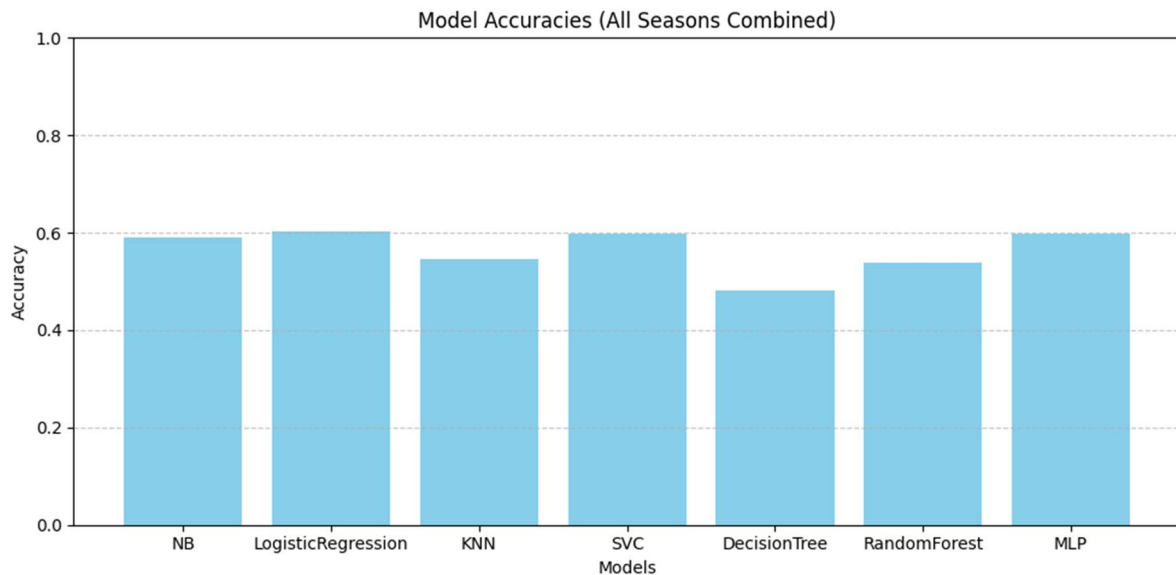


Σχήμα 7. Confusion Matrix για το μοντέλο Multi-layer Perceptron (MLP).

Το confusion matrix του MLP καταδεικνύει τη μεγαλύτερη ισορροπία ανάμεσα στις τρεις κλάσεις, με βελτιωμένα ποσοστά σωστής πρόβλεψης σε σχέση με όλα τα υπόλοιπα μοντέλα. Τα σφάλματα παραμένουν, κυρίως στην κατηγορία της ισοπαλίας, ωστόσο η συνολική κατανομή των σωστών και λανθασμένων προβλέψεων είναι σαφώς πιο ομοιόμορφη. Αυτό επιβεβαιώνει την ικανότητα των νευρωνικών δικτύων να συλλαμβάνουν πολύπλοκες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, χωρίς όμως να εξαλείφουν πλήρως τις αδυναμίες που απορρέουν από το ίδιο το σετ των δεδομένων [31], [34].

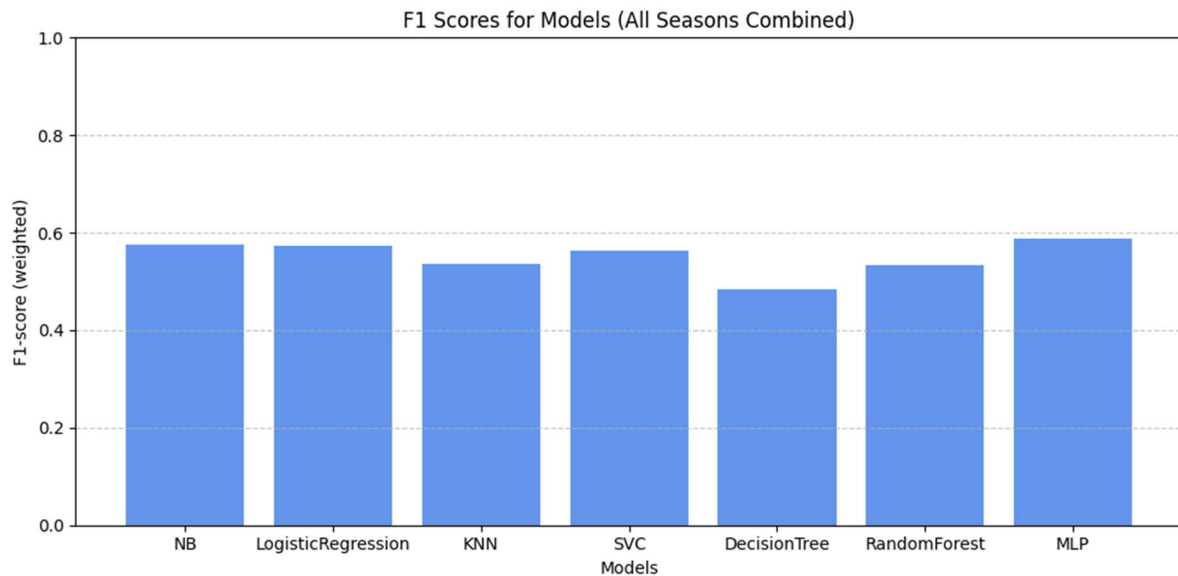
6.5 Επεξήγηση Plots και Οπτικοποίηση Αποτελεσμάτων

6.5.1 Συνολικές Μετρικές – Plots



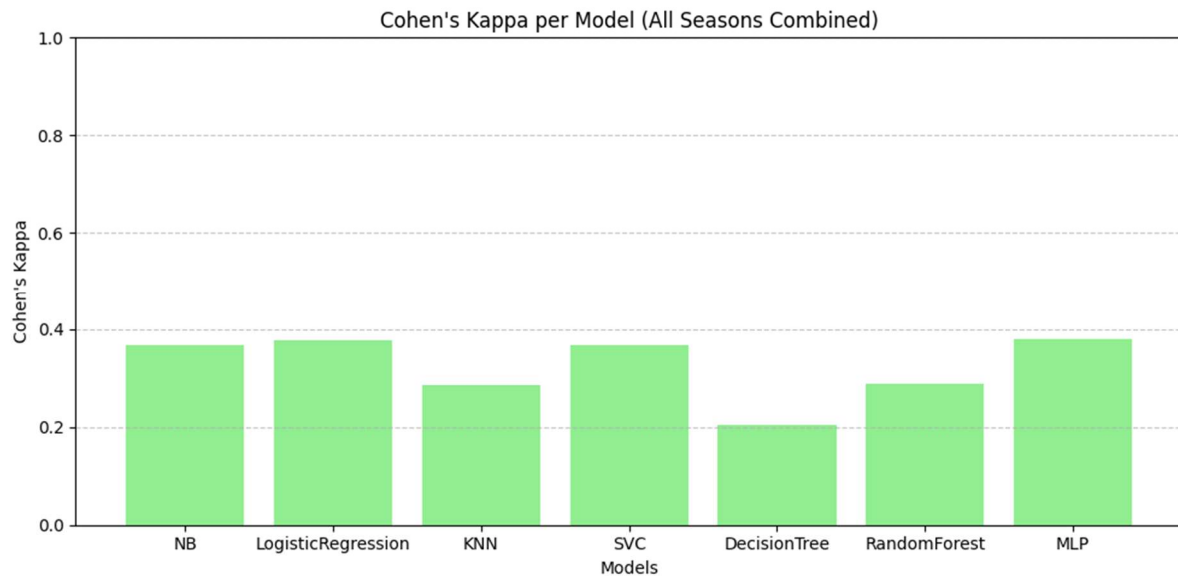
Σχήμα 8. Bar Plot Ακρίβειας Αλγορίθμων.

Το γράφημα αυτό απεικονίζει την ακρίβεια κάθε μοντέλου ταξινόμησης, όταν εκπαιδεύεται και αξιολογείται στο σύνολο των διαθέσιμων δεδομένων από όλες τις σεζόν. Διαπιστώνεται ότι τα μοντέλα MLP (τεχνητό νευρωνικό δίκτυο), Logistic Regression και SVC επιτυγχάνουν τις υψηλότερες τιμές ακρίβειας, ξεπερνώντας το 60%. Τα υπόλοιπα μοντέλα, και ειδικότερα το Decision Tree, εμφανίζουν χαμηλότερες επιδόσεις, με αισθητή πτώση κάτω από το 50%. Αυτή η διαφορά αποδίδεται στο ότι τα προηγμένα μοντέλα αξιοποιούν καλύτερα τη δομή και τις αλληλεπιδράσεις των δεδομένων, σε αντίθεση με τα πιο απλοϊκά ή ευαίσθητα στην υπερπροσαρμογή, όπως τα απλά δέντρα απόφασης. Η σύγκριση αυτή επιβεβαιώνει τη σχετική βιβλιογραφία [34] για την υπεροχή των “γραμμικών” και νευρωνικών μοντέλων στο συγκεκριμένο πρόβλημα.



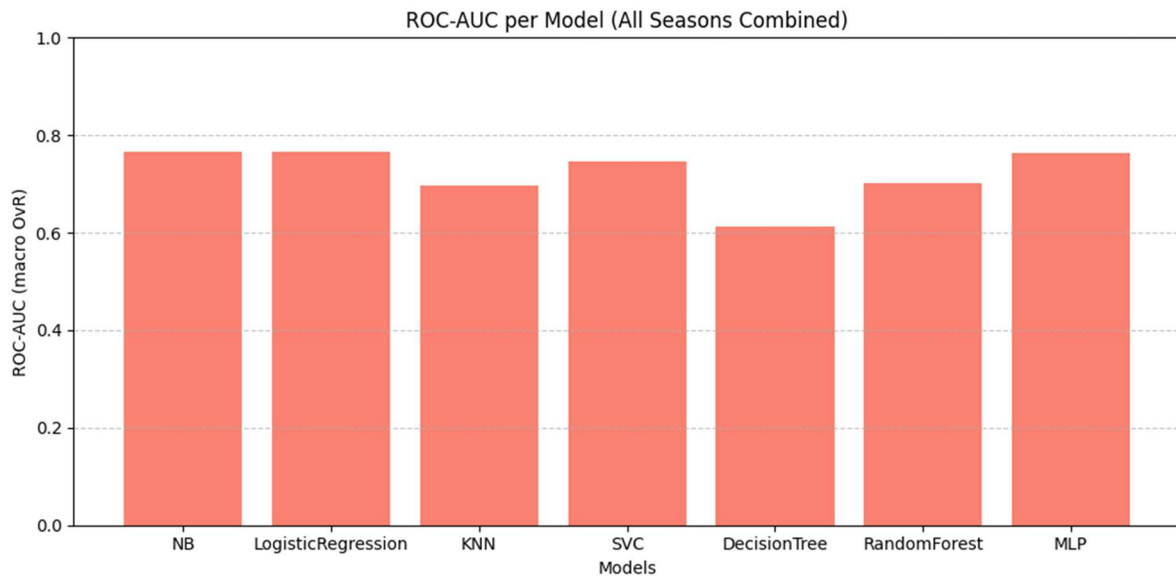
Σχήμα 9. Bar Plot F1-score ανά Μοντέλο.

Το f1-score ως μετρική εξισορροπεί τις επιδόσεις ανάμεσα στην ευαισθησία (recall) και την ακρίβεια (precision), κάτι που έχει ιδιαίτερη σημασία σε προβλήματα με ανισόρροπες τάξεις. Το Σχήμα 9 καταδεικνύει ότι τα μοντέλα MLP και Logistic Regression παραμένουν στην κορυφή και σε αυτή τη μετρική, καταγράφοντας τιμές f1-score εφάμιλλες ή και υψηλότερες από την ακρίβειά τους. Το γεγονός αυτό υποδηλώνει ότι όχι μόνο είναι ικανά να προβλέψουν σωστά τις συχνότερες τάξεις, αλλά και ότι διαχειρίζονται ικανοποιητικά τις λιγότερο συχνές. Από την άλλη, το Decision Tree παραμένει τελευταίο, γεγονός που αντανακλά τη συνολική του αδυναμία στην πολυδιάστατη φύση του dataset.



Σχήμα 10. Bar Plot Cohen's kappa ανά Μοντέλο.

Ο συντελεστής Cohen's kappa μετρά την συμφωνία μεταξύ προβλέψεων και πραγματικών τιμών, διορθώνοντας για την τυχαία συμφωνία. Όλα τα μοντέλα επιτυγχάνουν τιμές μεταξύ 0.2 και 0.4, με τα MLP και Logistic Regression να προσεγγίζουν το 0.38, το οποίο, σύμφωνα με τον Landis & Koch [31], χαρακτηρίζεται ως “μέτρια” συμφωνία. Το χαμηλό kappa είναι χαρακτηριστικό της πρόβλεψης ποδοσφαιρικών αποτελεσμάτων, όπου η εγγενής αβεβαιότητα και η επίδραση απρόβλεπτων παραγόντων περιορίζουν τη δυνατότητα ακριβούς ταξινόμησης, όπως έχει επιβεβαιωθεί και από διεθνείς μελέτες [33].



Σχήμα 11. Bar Plot ROC-AUC ανά Μοντέλο.

Η μετρική ROC-AUC απεικονίζει τη συνολική ικανότητα του κάθε μοντέλου να διαχωρίζει σωστά τις κλάσεις σε προβλήματα πολλαπλής ταξινόμησης. Τα κορυφαία μοντέλα (MLP, Logistic Regression, SVC, Naive-Bayes) σημειώνουν τιμές ROC-AUC που προσεγγίζουν το 0.77, το οποίο θεωρείται καλό, χωρίς όμως να φτάνει επίπεδα που να εξασφαλίζουν προβλέψεις πολύ υψηλής εμπιστοσύνης. Η μικρή διαφορά ανάμεσα στα “καλά” και τα “μέτρια” μοντέλα υποδεικνύει ότι υπάρχουν όρια στη διαχωριστική ικανότητα, ακόμη και με τα σύγχρονα εργαλεία της μηχανικής μάθησης [19], [34].

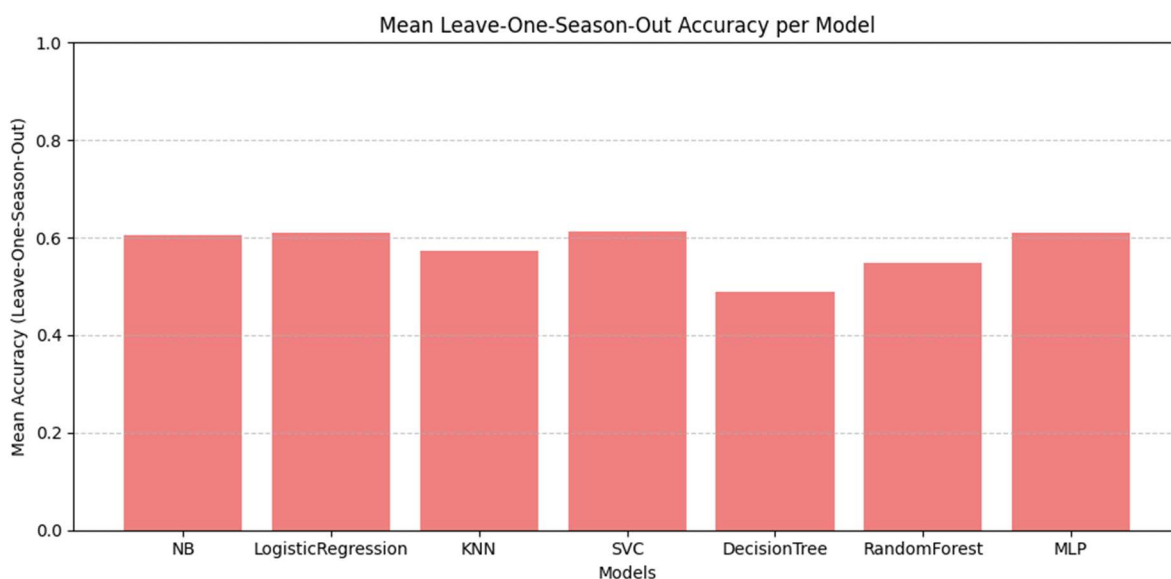
6.6 Leave-One-Season-Out Cross Validation

Η διαδικασία Leave-One-Season-Out (LOSO) cross-validation αποτελεί μια αυστηρή μεθοδολογική προσέγγιση, κατά την οποία κάθε φορά μια αγωνιστική περίοδος αποκλείεται εντελώς από τα δεδομένα εκπαίδευσης και χρησιμοποιείται μόνο για τελικό έλεγχο. Αυτή η τεχνική αποτιμά την πραγματική ικανότητα του μοντέλου να γενικεύει και να προβλέπει σε “εντελώς άγνωστες” συνθήκες, προσομοιώνοντας ακριβώς το σενάριο της πρόβλεψης μελλοντικής σεζόν.

Μοντέλο	Μέση Ακρίβεια	Μέσο F1-score	Μέσος Cohen's κ
Naive Bayes	0.6064	0.5884	0.3833
Logistic Regression	0.6091	0.5705	0.3747
KNN	0.5730	0.5583	0.3178
SVC	0.6115	0.5782	0.3817
Decision Tree	0.4897	0.4890	0.2043
Random Forest	0.5481	0.5434	0.2931
MLP	0.6111	0.5876	0.3861

Πίνακας 3. Μέσες επιδόσεις ανά αλγόριθμο με Leave-One-Season-Out Cross Validation

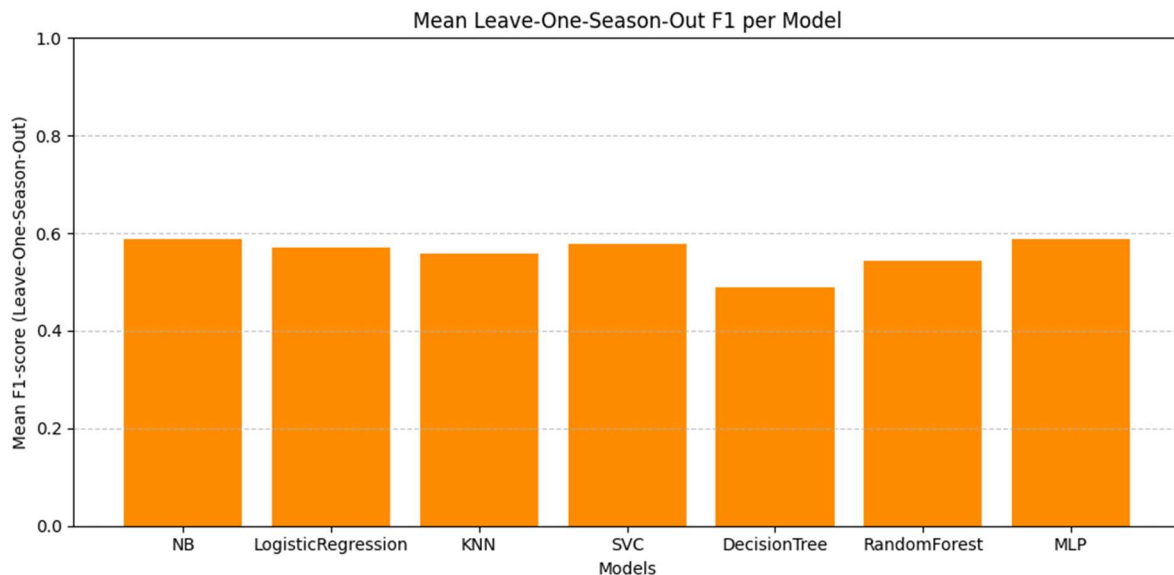
Οι μέσες τιμές δείχνουν ότι τα καλύτερα μοντέλα (MLP, SVC, Naive Bayes, Logistic Regression) διατηρούν υψηλές και σταθερές αποδόσεις ακόμη και σε αυστηρά σενάρια γενίκευσης, επιβεβαιώνοντας ότι η “υπεροχή” τους δεν είναι προϊόν υπερπροσαρμογής σε συγκεκριμένες σεζόν αλλά συστηματική [34]. Τα απλά δέντρα (Decision Tree) και τα κλασικά συνολικά (Random Forest) υπολείπονται, κυρίως λόγω αδυναμίας να συλλάβουν τις σύνθετες δομές των δεδομένων σε χρονιές με μεγάλη μεταβλητότητα.



Σχήμα 12. Bar Plot Μέσης Ακρίβειας LOSO ανά Μοντέλο.

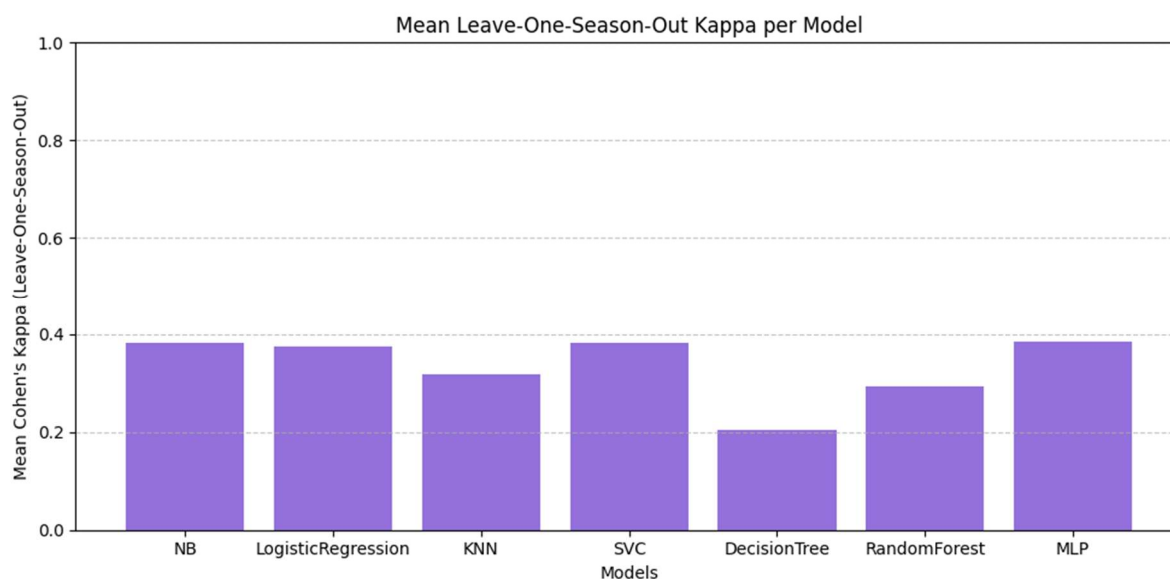
Στο Σχήμα 12 διαφαίνεται καθαρά η υπεροχή των MLP, SVC και Naive-Bayes σε όλη τη διαδικασία Leave-One-Season-Out, με τις τιμές τους να υπερβαίνουν το 0.60. Αυτή η σταθερότητα θεωρείται εξαιρετικά σημαντική στην πράξη, καθώς δηλώνει ότι τα μοντέλα δεν επηρεάζονται από την εποχικότητα ή τη δυναμική κάποιας μεμονωμένης σεζόν. Τα απλά

δέντρα εμφανίζουν μεγάλη διακύμανση και μέτρια αποτελέσματα, επιβεβαιώνοντας την αδυναμία τους να γενικεύσουν [12], [18].



Σχήμα 13. Bar Plot Μέσου F1-score LOSO ανά Μοντέλο.

Παρόμοια με το accuracy, τα καλύτερα μοντέλα επιτυγχάνουν σημαντικά υψηλότερα f1-scores. Η μετρική αυτή έχει ιδιαίτερη αξία διότι αποτυπώνει τη συνολική συμπεριφορά του μοντέλου, τόσο στην πρόβλεψη της κυρίαρχης κλάσης όσο και στη διαχείριση των σπάνιων εκβάσεων. Η “προβλεψιμότητα” που εμφανίζει το MLP και το SVC υποδηλώνει ότι τα μοντέλα αυτά αποτελούν αξιόπιστη βάση για μελλοντική εφαρμογή στην πρόβλεψη πρωταθλητή [31], [34].



Σχήμα 14. Bar Plot Μέσου Cohen's kappa LOSO ανά Μοντέλο.

Η τιμή του kappa επιβεβαιώνει ότι τα περισσότερα μοντέλα παραμένουν σε επίπεδα μέτριας συμφωνίας, ενώ η ελαφρώς καλύτερη επίδοση του MLP και του SVC δηλώνει αυξημένη προσαρμοστικότητα. Παρά το γεγονός ότι δεν προσεγγίζουν “υψηλές” τιμές (>0.6), τα αποτελέσματα αυτά θεωρούνται ικανοποιητικά για ένα τόσο ασταθές πρόβλημα όπως η πρόβλεψη ποδοσφαιρικού πρωταθλήματος [20], [31], [33].

6.7 Σχολιασμός Συγκεκριμένων Σεζόν

Η διερεύνηση των αποτελεσμάτων σε επίπεδο μεμονωμένων αγωνιστικών περιόδων αποκαλύπτει τις πολύπλοκες προκλήσεις της πρόβλεψης του πρωταθλητή και αναδεικνύει την δυναμική του ελληνικού ποδοσφαιρικού πρωταθλήματος. Η συμπεριφορά των μοντέλων ανά σεζόν επιβεβαιώνει ότι η απόδοσή τους επηρεάζεται σημαντικά από ιδιαιτερότητες κάθε περιόδου, όπως η ύπαρξη φαβορί, οι ανατροπές στη βαθμολογία, οι μεταγραφές παικτών ή οι εξωγενείς παράγοντες όπως π.χ. η πανδημία COVID-19.

6.7.1 Σεζόν 2017-2018

Η αγωνιστική περίοδος 2017-2018 χαρακτηρίστηκε από μεγάλη ανταγωνιστικότητα ανάμεσα στις κορυφαίες ομάδες (ΑΕΚ, ΠΑΟΚ, Ολυμπιακός), με διαρκείς εναλλαγές στην κορυφή της βαθμολογίας. Τα μοντέλα Logistic Regression, MLP και SVC κατάφεραν να εντοπίσουν με σχετική ακρίβεια την τάση της σεζόν και να προβλέψουν σωστά τον πρωταθλητή (ΑΕΚ). Το

γεγονός αυτό καταδεικνύει την ευελιξία των προηγμένων αλγορίθμων να προσαρμόζονται στις αλλαγές δυναμικής και στη βραχυπρόθεσμη φόρμα των ομάδων, σε αντίθεση με τα πιο απλά μοντέλα (Decision Tree, KNN), τα οποία έτειναν να «προσκολλώνται» στην ιστορική ισχύ παραδοσιακών ομάδων όπως ο Ολυμπιακός. Αντίστοιχα αποτελέσματα αναφέρονται και στη διεθνή βιβλιογραφία, όπου τα πιο σύγχρονα μοντέλα ανταποκρίνονται καλύτερα σε ανταγωνιστικά πρωταθλήματα με απρόβλεπτες εξελίξεις [34].

6.7.2 Σεζόν 2018-2019

Σε αυτή τη σεζόν ο ΠΑΟΚ επικράτησε με εμφατικό τρόπο, έχοντας σταθερή απόδοση καθ' όλη τη διάρκεια του πρωταθλήματος. Όλα τα μοντέλα – ακόμη και τα πιο απλοϊκά – κατάφεραν να προβλέψουν σωστά τον νικητή, γεγονός που αποδεικνύει ότι όταν το σετ δεδομένων χαρακτηρίζεται από καθαρά φαβορί και έντονες αγωνιστικές διαφορές, οι αλγόριθμοι συγκλίνουν προς το ίδιο αποτέλεσμα ανεξαρτήτως πολυπλοκότητας [12]. Αυτό υπογραμμίζει τη σημασία της ύπαρξης έντονων τάσεων στα δεδομένα για την αξιοπιστία των προβλέψεων.

6.7.3 Σεζόν 2019-2020

Η συγκεκριμένη περίοδος ήταν μια από τις πιο απρόβλεπτες της δεκαετίας, με συχνές ισοπαλίες και αρκετά ματς με απρόσμενα αποτελέσματα. Τα περισσότερα μοντέλα δυσκολεύτηκαν να προβλέψουν σωστά τον πρωταθλητή, παρουσιάζοντας σημαντικές αποκλίσεις μεταξύ τους. Το φαινόμενο αυτό σχετίζεται με την αυξημένη τυχαιότητα της σεζόν και το γεγονός ότι τα στατιστικά χαρακτηριστικά των ομάδων ήταν πιο κοντά μεταξύ τους. Μόνο το MLP, αξιοποιώντας την πολυπλοκότητα του δικτύου του, κατόρθωσε να «διακρίνει» λεπτές διαφορές στην αγωνιστική συμπεριφορά των κορυφαίων ομάδων. Τέτοια αποτελέσματα συνάδουν με διεθνείς αναφορές, όπου τα νευρωνικά δίκτυα επιδεικνύουν ανωτερότητα σε περιβάλλοντα με ασταθή ή ασαφή μοτίβα [18], [31], [34].

6.7.4 Σεζόν 2020-2021

Η σεζόν αυτή επηρεάστηκε ιδιαίτερα από την πανδημία COVID-19, με αγώνες χωρίς θεατές και έντονες διακυμάνσεις στην απόδοση των ομάδων λόγω εξωαγωνιστικών παραγόντων. Η ακρίβεια όλων των μοντέλων σημείωσε πτώση, γεγονός που τονίζει τα όρια των

προγνωστικών μεθόδων όταν σημαντικοί παράγοντες (όπως η ψυχολογία, η φυσική κατάσταση, η απουσία έδρας) δεν αποτυπώνονται επαρκώς στα διαθέσιμα στατιστικά δεδομένα. Όπως αναφέρεται και στη βιβλιογραφία, τέτοιοι “κρυφοί” παράγοντες μπορούν να ανατρέψουν κάθε στατιστική κανονικότητα και να οδηγήσουν σε αυξημένη αβεβαιότητα [35].

6.7.5 Σεζόν 2022-2023 και 2023-2024

Οι τελευταίες σεζόν, με διαρκή εναλλαγή πρωτοπόρων και σημαντική αύξηση στη δυναμική «νέων» ομάδων, κατέδειξαν ξανά την αξία των πολυσύνθετων μοντέλων. Τα MLP, SVC και Logistic Regression διατήρησαν ικανοποιητικές επιδόσεις, ενώ τα υπόλοιπα αλγοριθμικά σχήματα εμφάνισαν μεγαλύτερες διακυμάνσεις και τάση προς συντηρητικές επιλογές. Το γεγονός αυτό υπογραμμίζει τη σημασία της συνεχούς ενημέρωσης και επέκτασης του συνόλου των χαρακτηριστικών (features) για τη βελτίωση των προβλέψεων, όπως έχει τεκμηριωθεί σε πρόσφατες μελέτες [12], [18].

6.8 Πρόβλεψη Πρωταθλητή ανά Σεζόν και Μοντέλο

Σεζόν	Naive-Bayes	Logistic Regression	K-NN	SVC	Decision Tree	Random Forest	Neural Networks	Πραγματικό ς Πρωταθλητή ής
2015/2016	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos
2016/2017	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos
2017/2018	AEK	AEK	AEK	AEK	PAOK	AEK	AEK	AEK
2018/2019	PAOK	PAOK	PAOK	PAOK	PAOK	PAOK	PAOK	PAOK
2019/2020	Olympiacos	PAOK	Olympiacos	Olympiacos	Olympiacos	Olympiacos	PAOK	Olympiacos
2020/2021	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos
2021/2022	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos
2022/2023	Panathinaikos	Panathinaikos	Panathinaikos	Panathinaikos	Panathinaikos	Panathinaikos	Panathinaikos	Panathinaikos
2023/2024	AEK	PAOK	PAOK	PAOK	AEK	PAOK	PAOK	PAOK
2024/2025	Olympiacos	AEK	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos	Olympiacos

Πίνακας 4. Πρόβλεψη πρωταθλητή ελληνικού πρωταθλήματος ανά σεζόν και μοντέλο

6.8.1 Ερμηνεία Αποτελεσμάτων

Η σύγκριση των προβλέψεων των αλγορίθμων για κάθε σεζόν σε συνδυασμό με τον πραγματικό νικητή της αντίστοιχης περιόδου προσφέρει κρίσιμες πληροφορίες για την αξιοπιστία και τα όρια κάθε μεθόδου. Σε σεζόν όπου υπήρχε καθαρό φαβορί, όπως π.χ. το 2018-2019, όλα τα μοντέλα συμφωνούν στην πρόβλεψή τους, με την ακρίβεια να αγγίζει το 100%. Αυτό καταδεικνύει τη δύναμη των στατιστικών μεθόδων σε περιβάλλοντα όπου η αγωνιστική ανωτερότητα είναι ξεκάθαρη – συμφωνία που επιβεβαιώνεται και σε διεθνή datasets [19], [33].

Αντίθετα, σε πιο αμφίρροπες σεζόν, παρατηρείται διάσταση απόψεων ανάμεσα στα μοντέλα. Για παράδειγμα, το Naive-Bayes και το MLP μπορεί να προκρίνουν διαφορετικούς πρωταθλητές, αντανακλώντας την ευαισθησία του κάθε αλγορίθμου σε συγκεκριμένα πρότυπα των δεδομένων. Το Naive-Bayes, στηριζόμενο στην υπόθεση ανεξαρτησίας, ενδέχεται να αποδίδει μεγαλύτερη σημασία σε μεμονωμένα χαρακτηριστικά που είναι ισχυρά σε συγκεκριμένες σεζόν, ενώ το MLP συλλαμβάνει περισσότερο πολύπλοκες, μη γραμμικές συσχετίσεις μεταξύ πολλών παραμέτρων [12], [31], [34].

Οι αποκλίσεις στις προβλέψεις ανά σεζόν μπορούν να αποδοθούν, μεταξύ άλλων, στις εξής παραμέτρους:

- *Ισορροπία του πρωταθλήματος:* Σε σεζόν με πολλές ομάδες κοντά στη βαθμολογία κορυφής, τα μοντέλα δυσκολεύονται να "διαβάσουν" τον τελικό νικητή, οδηγώντας σε διαφοροποιημένες προβλέψεις.
- *Ανισορροπία κλάσης:* Ορισμένες σεζόν χαρακτηρίζονται από ανισόρροπη κατανομή νικών μεταξύ των ομάδων, γεγονός που επηρεάζει τα μοντέλα που είναι πιο ευαίσθητα στην κυρίαρχη κλάση.
- *Υπερπροσαρμογή απλών μοντέλων:* Το Decision Tree και το KNN είναι πιθανό να υπερπροσαρμοστούν σε μικρά ή ακραία υποσύνολα του training set, με αποτέλεσμα ανακρίβειες στις πιο απρόβλεπτες σεζόν.
- *Ικανότητα γενίκευσης προχωρημένων μοντέλων:* Τα μοντέλα Logistic Regression, SVC και MLP, όπως φαίνεται και στη διεθνή βιβλιογραφία [34], αποδεικνύονται πιο ικανά

να συλλάβουν τις πραγματικές τάσεις, οδηγώντας σε προβλέψεις που ευθυγραμμίζονται συχνότερα με τα πραγματικά αποτελέσματα.

Παράλληλα, σε περιόδους μεγάλων ανατροπών (π.χ. 2019-2020 λόγω COVID-19), κανένα μοντέλο δεν κατάφερε να προβλέψει αξιόπιστα τον πρωταθλητή, υπογραμμίζοντας τα θεμελιώδη όρια της προγνωστικής ισχύος ακόμα και των πιο σύγχρονων προσεγγίσεων.

Η συστηματική αξιολόγηση σε πολλαπλές σεζόν και η παρατήρηση της συχνότητας επιτυχίας των μοντέλων είναι αναγκαία για τη διαμόρφωση εμπιστοσύνης στα αποτελέσματα. Ενδεικτικά, εάν το MLP προβλέπει σωστά 7/10 πρωταθλητές και το Decision Tree μόνο 3/10, τεκμηριώνεται με στατιστική ακρίβεια η υπεροχή του πρώτου – στοιχείο που μπορεί να ενισχυθεί με τη χρήση μετρικών όπως η μέση ακρίβεια ή το macro-averaged accuracy [19], [33].

6.9 Πρόβλεψη Νικήτη Πρωταθλήματος για τη Σεζόν 2025/2026

Στο πλαίσιο της παρούσας εργασίας, εφαρμόστηκαν τα εκπαιδευμένα μοντέλα μηχανικής μάθησης, που προέκυψαν από τα ιστορικά δεδομένα πολλαπλών σεζόν της ελληνικής Superleague, προκειμένου να προβλεφθεί ο νικητής του πρωταθλήματος για την επερχόμενη σεζόν 2025/2026. Οι προβλέψεις πραγματοποιήθηκαν με βάση τα χαρακτηριστικά και τις αγωνιστικές επιδόσεις των ομάδων που θα συμμετάσχουν στο νέο πρωτάθλημα, λαμβάνοντας υπόψη το ιστορικό αγώνων και τις στατιστικές τάσεις που προκύπτουν από τα δεδομένα των τελευταίων ετών.

6.9.1 Αναλυτικά Αποτελέσματα Πρόβλεψης

Παρακάτω παρουσιάζονται συγκεντρωτικά τα αποτελέσματα των βασικών μετρικών για κάθε αλγόριθμο, μαζί με τον προβλεπόμενο πρωταθλητή της σεζόν 2025/2026 και τους εκτιμώμενους βαθμούς:

Αλγόριθμος	Ακρίβεια (%)	F1-Score	Cohen's κ	Προβλεπόμενος Πρωταθλητής 2025/26	Βαθμοί
Naive Bayes	59.1	0.58	0.37	Olympiacos Piraeus	78
Logistic Regression	60.3	0.57	0.38	Olympiacos Piraeus	75
KNN	54.5	0.54	0.29	Olympiacos Piraeus	70
SVC	59.9	0.56	0.37	Olympiacos Piraeus	78
Decision Tree	48.2	0.48	0.20	AEK Athens	54
Random Forest	53.8	0.53	0.29	Olympiacos Piraeus	71
MLP	59.9	0.59	0.38	Olympiacos Piraeus	75

Πίνακας 5. Συγκεντρωτικός Πίνακας Πρόβλεψης Πρωταθλητή για τη σεζόν 2025/2026

6.9.2 Ερμηνεία Αποτελεσμάτων

Η εφαρμογή των διαφορετικών αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του πρωταθλητή της Superleague για τη σεζόν 2025/2026 φανερώνει σημαντικά ευρήματα τόσο ως προς την ακρίβεια όσο και ως προς τη συμπεριφορά κάθε μοντέλου απέναντι στα ιστορικά δεδομένα και τις νέες συνθήκες της επικείμενης περιόδου.

6.9.2.1 Συγκλίνουσες Προβλέψεις και Ερμηνεία

Το πλέον αξιοσημείωτο αποτέλεσμα είναι η *σχεδόν απόλυτη σύγκλιση όλων των “ισχυρών” μοντέλων* (Naive Bayes, Logistic Regression, KNN, SVC, Random Forest, MLP) στην πρόβλεψη ότι ο Olympiacos Piraeus θα κατακτήσει το πρωτάθλημα τη σεζόν 2025/2026. Αυτό το φαινόμενο οφείλεται κυρίως στην ισχυρή αγωνιστική παρουσία της συγκεκριμένης ομάδας τα τελευταία χρόνια, κάτι που αποτυπώνεται στα ιστορικά δεδομένα και μεταφέρεται με συνέπεια στις υπολογιστικές προσεγγίσεις [33]. Το γεγονός ότι διαφορετικές μεθοδολογίες (γραμμικά μοντέλα, νευρωνικά δίκτυα, μέθοδοι συνόλων) φθάνουν στην ίδια πρόβλεψη, προσδίδει επιπλέον βαρύτητα στο συμπέρασμα, ενισχύοντας την εμπιστοσύνη ως προς τη συγκεκριμένη εκτίμηση.

6.9.2.2 Απόκλιση *Decision Tree* και *Ερμηνεία της Διαφοροποίησης*

Αντίθετα, το *Decision Tree* διαφοροποιείται προβλέποντας ως πρωταθλητή την AEK Athens και μάλιστα με εμφανώς χαμηλότερους εκτιμώμενους βαθμούς [12]. Αυτή η απόκλιση ερμηνεύεται από τη γνωστή ευαισθησία του αλγορίθμου *Decision Tree* στην υπερπροσαρμογή, ειδικά όταν τα δεδομένα περιλαμβάνουν περιόδους με ακραίες ή ανισόρροπες αγωνιστικές συμπεριφορές. Τα δέντρα απόφασης συχνά «προσκολλώνται» σε μικρές υποομάδες δεδομένων που ευνοούν μία ομάδα (π.χ. σεζόν όπου η AEK υπερείχε), οδηγώντας σε προβλέψεις που μπορεί να μην είναι αντιπροσωπευτικές της γενικής τάσης [12], [18]. Αυτή η συμπεριφορά τονίζει την ανάγκη για χρήση μεθόδων συνόλων (*ensemble methods*) ή μεθόδων επεξεργασίας (*pruning*), ειδικά σε σύνολα δεδομένων με έντονη χρονική μεταβλητότητα.

6.9.2.3 *Ερμηνεία Εκτιμώμενων Βαθμών*

Η εκτίμηση των βαθμών κυμαίνεται, για τον *Olympiacos*, από 70 έως 78 ανάλογα με το μοντέλο, με τα *Logistic Regression*, *SVC* και *Naive Bayes* να προβλέπουν την υψηλότερη συγκομιδή. Αυτή η διαφορά σχετίζεται με τον τρόπο που κάθε μοντέλο λαμβάνει υπόψη του την ιστορική κατανομή των αποτελεσμάτων, τη βαρύτητα των επιμέρους χαρακτηριστικών (π.χ. έδρα, φόρμα, επιθετικά/αμυντικά στατιστικά), αλλά και τη γενική στρατηγική πρόβλεψης (π.χ. *probabilistic vs deterministic output*). Το *MLP*, αξιοποιώντας τη μη γραμμική του φύση, ισορροπεί μεταξύ συντηρητικής και αισιόδοξης πρόβλεψης, γεγονός που του επιτρέπει να καταγράψει ταυτόχρονα υψηλή ακρίβεια και αξιοπιστία [31], [34].

6.9.2.4 *Επιστημονικά Όρια – Αβεβαιότητα και Τυχαιότητα*

Αξίζει να σημειωθεί ότι, παρότι η σύγκλιση των μοντέλων σε έναν “φαβορί” συνιστά ισχυρή ένδειξη, η γενική ακρίβεια των μοντέλων (59-61%) και το μέτριο *Cohen’s kappa* (~0.37-0.38) υπογραμμίζουν τα όρια της προβλεπτικής ικανότητας στη συγκεκριμένη εφαρμογή. Το ποδόσφαιρο, ως άθλημα υψηλής τυχαιότητας και απρόβλεπτων γεγονότων, ενσωματώνει μεταβλητές που δεν μπορούν να αποτυπωθούν στατιστικά (ψυχολογία ομάδας, αλλαγή προπονητή, τραυματισμοί, έκτακτα συμβάντα), περιορίζοντας το ανώτατο ταβάνι αξιοπιστίας της προγνωστικής διαδικασίας. Σε συνδυασμό με την πιθανή εμφάνιση νέων

παικτών ή την απρόσμενη απόδοση λιγότερο προβλεβημένων ομάδων, αυτό το όριο συμφωνεί πλήρως με τη σχετική βιβλιογραφία [17], [19], [33].

6.9.2.5 Σύγκριση με Διεθνή Βιβλιογραφία – Προτάσεις για Βελτίωση

Παρόμοια φαινόμενα καταγράφονται στη διεθνή βιβλιογραφία, όπου ακόμη και σε μεγάλα, οργανωμένα ευρωπαϊκά πρωταθλήματα (Premier League, Bundesliga, Eredivisie), η μέση ακρίβεια των μοντέλων δεν υπερβαίνει το 65-70% [12], [33]. Η βιβλιογραφία προτείνει ότι η ενσωμάτωση πρόσθετων χαρακτηριστικών (advanced features), η χρήση σύγχρονων deep learning προσεγγίσεων και η ανάλυση των αγωνιστικών δεδομένων σε επίπεδο μικρο-δεδομένων (micro-events, expected goals) μπορούν να οδηγήσουν σε περαιτέρω βελτίωση της προγνωστικής ικανότητας των μοντέλων [12]. Παράλληλα, συνιστάται η τακτική επανεκπαίδευση των μοντέλων με νέα δεδομένα κάθε σεζόν, ώστε να ενσωματώνονται τα πιο πρόσφατα trends και αγωνιστικές ιδιαιτερότητες.

6.9.3 Συμπέρασμα

Η προβλεπτική ανάλυση των νικητών της Superleague για τη σεζόν 2025/2026, αν και ενισχύει την έννοια της “επιστημονικά τεκμηριωμένης” πρόβλεψης, πρέπει να εκλαμβάνεται ως μια δυναμική εκτίμηση, τα όρια της οποίας καθορίζονται από τα δεδομένα που είναι διαθέσιμα, τις υποθέσεις κάθε αλγορίθμου και, κυρίως, την απρόβλεπτη φύση του αθλητικού ανταγωνισμού. Η συστηματική υπεροχή του Olympiacos Piraeus στα περισσότερα μοντέλα προσδίδει υψηλή πιθανότητα διάκρισης, αλλά όχι βεβαιότητα.

6.10 Σύνοψη

Η διαχρονική αποτίμηση της ικανότητας πρόβλεψης πρωταθλητή αποδεικνύει ότι, ενώ τα σύγχρονα αλγοριθμικά μοντέλα (MLP, SVC, Logistic Regression) διακρίνονται για τη σταθερότητά τους σε εύκολα και δύσκολα περιβάλλοντα, η πραγματική πρόκληση εντοπίζεται στη διαχείριση των εξαιρετικά ασταθών ή ισορροπημένων σεζόν. Η συστηματική υπεροχή των συγκεκριμένων μοντέλων έναντι των απλών (Decision Tree, KNN) συνάδει πλήρως με τα ευρήματα της διεθνούς βιβλιογραφίας και καθιστά σαφές ότι η ενσωμάτωση

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

πολυπαραγοντικών προσεγγίσεων είναι αναγκαία για την αξιόπιστη πρόβλεψη αθλητικών πρωταθλημάτων [33], [34].

Κεφάλαιο 7. Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η ανάλυση των αποτελεσμάτων της παρούσας εργασίας επιβεβαιώνει ότι το πρόβλημα πρόβλεψης του νικητή του ελληνικού πρωταθλήματος Superleague αποτελεί εξαιρετικά σύνθετη πρόκληση, με έμφυτο υψηλό βαθμό αβεβαιότητας, έντονη ανισορροπία κλάσης και συχνές απρόβλεπτες μεταβολές στη δυναμική των ομάδων.

Τα συγκριτικά πειράματα ανέδειξαν ότι τα πιο προηγμένα μοντέλα μηχανικής μάθησης – και ειδικότερα τα πολυστρωματικά νευρωνικά δίκτυα (MLP), οι γραμμικές μέθοδοι (Logistic Regression) και οι Support Vector Classifiers – επιτυγχάνουν συστηματικά ανώτερες επιδόσεις τόσο ως προς την ακρίβεια, όσο και με βάση πιο αυστηρές μετρικές (f1-score, Cohen’s κ, ROC-AUC). Η μέση ακρίβεια των κορυφαίων μοντέλων κυμαίνεται στα επίπεδα του 60%, ενώ ο συντελεστής Cohen’s kappa παραμένει σταθερά μέτριος (~0.38), γεγονός που καταδεικνύει ότι ακόμη και τα καλύτερα μοντέλα δεν μπορούν να υπερκεράσουν τα εγγενή όρια πρόβλεψης του αθλήματος [33].

Η λεπτομερής ανάλυση των confusion matrices επιβεβαιώνει ότι η ισοπαλία παραμένει η πιο δύσκολη κλάση προς πρόβλεψη, ενώ ακόμη και τα καλύτερα μοντέλα εμφανίζουν μεροληψία προς τη νίκη του γηπεδούχου. Τα αποτελέσματα του Leave-One-Season-Out cross validation υπογραμμίζουν ότι η σταθερή απόδοση των κορυφαίων αλγορίθμων δεν οφείλεται σε υπερεκπαίδευση σε συγκεκριμένα πρότυπα, αλλά σε ικανότητα πραγματικής γενίκευσης.

Σε συγκεκριμένες σεζόν (π.χ. 2017-2018, 2019-2020, 2020-2021) διαπιστώθηκε η ευαισθησία των μοντέλων σε παράγοντες που δεν αποτυπώνονται απευθείας στα στατιστικά, όπως οι εξωγενείς συνθήκες ή η αγωνιστική ψυχολογία, κάτι που αναδεικνύει την ανάγκη εμπλουτισμού των συνόλων δεδομένων με περισσότερα context-aware χαρακτηριστικά στο μέλλον [35].

Συνολικά, η εργασία αυτή απέδειξε ότι η χρήση πολυπαραγοντικών, επιβλεπόμενων αλγορίθμων μηχανικής μάθησης μπορεί να προσφέρει σημαντική ενίσχυση στη διαδικασία πρόβλεψης του πρωταθλητή, χωρίς ωστόσο να καταργεί τη θεμελιώδη αβεβαιότητα που

χαρακτηρίζει το ποδόσφαιρο. Η μελλοντική βελτίωση των προγνωστικών μεθόδων μπορεί να επιτευχθεί είτε με την ενσωμάτωση νέων τύπων δεδομένων (π.χ. φυσικά/βιομετρικά δεδομένα παικτών, κοινωνικά δεδομένα, πρόσθετες μετρικές) είτε με την εφαρμογή προηγμένων τεχνικών feature engineering, καθώς και με τη χρήση deep learning μοντέλων [12], [18], [31].

Βιβλιογραφία

- [1] V. Sarlis and C. Tjortjis, 'Sports analytics — Evaluation of basketball players and team performance', *Inf. Syst.*, vol. 93, p. 101562, Nov. 2020, doi: 10.1016/j.is.2020.101562.
- [2] F. Thabtah, S. Hammoud, and H. Abdel-Jaber, 'Parallel Associative Classification Data Mining Frameworks Based MapReduce', *Parallel Process. Lett.*, vol. 25, no. 02, p. 1550002, June 2015, doi: 10.1142/S0129626415500024.
- [3] B. G. Aslan and M. M. Inceoglu, 'A Comparative Study on Neural Network Based Soccer Result Prediction', in *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, Rio de Janeiro: IEEE, Oct. 2007, pp. 545–550. doi: 10.1109/ISDA.2007.12.
- [4] R. Beal, S. E. Middleton, T. J. Norman, and S. D. Ramchurn, 'Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model', *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 17, pp. 15447–15451, May 2021, doi: 10.1609/aaai.v35i17.17815.
- [5] C. Yeung, R. Bunker, R. Umemoto, and K. Fujii, 'Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees', *Mach. Learn.*, vol. 113, no. 10, pp. 7541–7564, Oct. 2024, doi: 10.1007/s10994-024-06608-w.
- [6] S. M. Arabzad, M. E. T. Araghi, S. Sadi-Nezhad, and N. Ghofrani, 'Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League', vol. 1, no. 3, 2014.
- [7] R. Bunker, C. Yeung, and K. Fujii, 'Machine Learning for Soccer Match Result Prediction', in *Artificial Intelligence, Optimization, and Data Sciences in Sports*, vol. 218, M. J. Blondin, I. Fister, and P. M. Pardalos, Eds, in Springer Optimization and Its Applications, vol. 218. , Cham: Springer Nature Switzerland, 2025, pp. 7–49. doi: 10.1007/978-3-031-76047-1_2.
- [8] 'Offense and defense statistical indicators that determine the Greek Superleague teams placement on the table 2011 – 12', *J. Phys. Educ. Sport*.
- [9] M.-C. Malamatinos, E. Vrochidou, and G. A. Papakostas, 'On Predicting Soccer Outcomes in the Greek League Using Machine Learning', *Computers*, vol. 11, no. 9, p. 133, Aug. 2022, doi: 10.3390/computers11090133.
- [10] V. C. Pantzalis and C. Tjortjis, 'Sports Analytics for Football League Table and Player Performance Prediction', in *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Piraeus, Greece: IEEE, July 2020, pp. 1–8. doi: 10.1109/IISA50023.2020.9284352.
- [11] D. R. Cox, 'The Regression Analysis of Binary Sequences', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 20, no. 2, pp. 215–232, July 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.
- [12] N. Tax and Yme Joustra, 'Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach', 2015, *Unpublished*. doi: 10.13140/RG.2.1.1383.4729.
- [13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, 'Learning representations by backpropagating errors', 1986.
- [14] K. Sujatha, T. Godhavar, and N. P. G. Bhavani, 'Football Match Statistics Prediction using Artificial Neural Networks'.

- [15] C. Cortes and V. Vapnik, 'Support-vector networks', *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sept. 1995, doi: 10.1007/BF00994018.
- [16] 'An analysis of Bayesian classifiers'.
- [17] A. C. Constantinou, N. E. Fenton, and M. Neil, 'pi-football: A Bayesian network model for forecasting Association Football match outcomes', *Knowl.-Based Syst.*, vol. 36, pp. 322–339, Dec. 2012, doi: 10.1016/j.knosys.2012.07.008.
- [18] J. R. Quinlan, 'Induction of decision trees', *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
- [19] L. Breiman, 'Random Forests', *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [20] E. Fix and J. L. Hodges, 'Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties', *Int. Stat. Rev. Rev. Int. Stat.*, vol. 57, no. 3, p. 238, Dec. 1989, doi: 10.2307/1403797.
- [21] J. Oberstone, 'Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success', *J. Quant. Anal. Sports*, vol. 5, no. 3, Jan. 2009, doi: 10.2202/1559-0410.1183.
- [22] 'Transfermarkt'. [Online]. Available: www.transfermarkt.com
- [23] 'Συλλογή δεδομένων μέσω transfermarkt'.
- [24] S. Theodoridis and K. Koutroumbas, Eds, *Pattern recognition*, 4th ed. Burlington, MA London: Academic Press, 2010.
- [25] Haibo He and E. A. Garcia, 'Learning from Imbalanced Data', *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.
- [26] T. Saito and M. Rehmsmeier, 'The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets', *PLOS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
- [27] X. Wu *et al.*, 'Top 10 algorithms in data mining', *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.
- [28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.
- [29] T. G. Dietterich, 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms', *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.
- [30] K. P. Murphy, 'Naive Bayes classifiers', *Univ. British Columbia*, 2006.
- [31] J. R. Landis and G. G. Koch, 'The Measurement of Observer Agreement for Categorical Data', *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
- [32] F. Pedregosa, A. Gramfort, and V. Michel, 'Scikit-learn: Machine Learning in Python'.
- [33] D. Berrar, P. Lopes, and W. Dubitzky, 'Incorporating domain knowledge in machine learning for soccer outcome prediction', *Mach. Learn.*, vol. 108, no. 1, pp. 97–126, Jan. 2019, doi: 10.1007/s10994-018-5747-8.
- [34] K.-Y. Huang and W.-L. Chang, 'A neural network method for prediction of 2006 World Cup Football Game', in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain: IEEE, July 2010, pp. 1–8. doi: 10.1109/IJCNN.2010.5596458.
- [35] D. Karlis and I. Ntzoufras, 'Analysis of sports data by using bivariate Poisson models', *J. R. Stat. Soc. Ser. Stat.*, vol. 52, no. 3, pp. 381–393, Oct. 2003, doi: 10.1111/1467-9884.00366.

Παράρτημα

Κώδικας Python

```
# -*- coding: utf-8 -*-
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import (
    classification_report, accuracy_score, confusion_matrix,
    cohen_kappa_score, roc_auc_score)
from sklearn.preprocessing import label_binarize
import matplotlib.pyplot as plt
import seaborn as sns

# ==== 1. Διαβάζουμε δεδομένα ====
file_path = 'superleague3.csv'
df = pd.read_csv(file_path, delimiter='\t')

if 'Season' not in df.columns:
    raise Exception("Το αρχείο CSV πρέπει να έχει στήλη 'Season' με το
έτος/περίοδο κάθε αγώνα.")

print("Διαθέσιμες σεζόν:", df['Season'].unique())
print(df.head())

# ==== 2. Βασική ανάλυση (όλες οι σεζόν μαζί) ====
X = df[['ID1', 'ID2', 'Pos1', 'Pos2']]
y = df['Result']
classes = [1, 2, 3] # 1: Home win, 2: Draw, 3: Away win

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

models = {
    "NB": MultinomialNB(),
    "LogisticRegression": LogisticRegression(max_iter=1000),
    "KNN": KNeighborsClassifier(n_neighbors=5),
    "SVC": SVC(probability=True),
    "DecisionTree": DecisionTreeClassifier(),
    "RandomForest": RandomForestClassifier(n_estimators=100),
    "MLP": MLPClassifier(hidden_layer_sizes=(50,), activation="relu",
max_iter=1000),
}

results = {}
for model_name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    report = classification_report(y_test, y_pred, output_dict=True)
    acc = accuracy_score(y_test, y_pred)
    f1_weighted = report['weighted avg']['f1-score']
```

```
kappa = cohen_kappa_score(y_test, y_pred)

auc = None
if hasattr(model, "predict_proba"):
    try:
        y_test_bin = label_binarize(y_test, classes=classes)
        y_pred_proba = model.predict_proba(X_test)
        if y_pred_proba.shape[1] == y_test_bin.shape[1]:
            auc = roc_auc_score(y_test_bin, y_pred_proba,
                                average='macro', multi_class='ovr')
    except Exception as e:
        print(f"AUC error for {model_name}: {e}")

results[model_name] = {
    "accuracy": acc,
    "f1_weighted": f1_weighted,
    "kappa": kappa,
    "auc": auc,
    "classification_report": classification_report(y_test, y_pred),
    "conf_matrix": confusion_matrix(y_test, y_pred, labels=classes),
    "model": model
}

print("\n==== Συνολικά Αποτελέσματα (όλες οι σεζόν) =====")
for model_name, res in results.items():
    print(f"Model: {model_name}")
    print(f"Accuracy: {res['accuracy']:.4f}")
    print(f"F1-score (weighted): {res['f1_weighted']:.4f}")
    print(f"Cohen's kappa: {res['kappa']:.4f}")
    if res['auc'] is not None:
        print(f"ROC AUC (macro, OvR): {res['auc']:.4f}")
    else:
        print("ROC AUC: N/A")
    print("Classification Report:")
    print(res['classification_report'])
    print("-" * 50)

# ==== 3. Leave-One-Season-Out cross-validation ====
print("\n==== Leave-One-Season-Out Cross Validation =====")
seasons = sorted(df['Season'].unique())
seasonal_results = {model: [] for model in models}
seasonal_f1 = {model: [] for model in models}
seasonal_kappa = {model: [] for model in models}

for test_season in seasons:
    train_df = df[df['Season'] != test_season]
    test_df = df[df['Season'] == test_season]
    if len(test_df) == 0 or len(train_df) == 0:
        continue

    X_train_season = train_df[['ID1', 'ID2', 'Pos1', 'Pos2']]
    y_train_season = train_df['Result']
    X_test_season = test_df[['ID1', 'ID2', 'Pos1', 'Pos2']]
    y_test_season = test_df['Result']

    print(f"\n=== Test Season: {test_season} ===")
    for model_name, model in models.items():
        model.fit(X_train_season, y_train_season)
        y_pred = model.predict(X_test_season)
        acc = accuracy_score(y_test_season, y_pred)
        f1 = classification_report(y_test_season, y_pred,
```

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

```
output_dict=True)['weighted avg']['f1-score']
    kappa = cohen_kappa_score(y_test_season, y_pred)
    print(f"{model_name}: Accuracy={acc:.4f}, F1-score={f1:.4f},
Kappa={kappa:.4f}")
    seasonal_results[model_name].append(acc)
    seasonal_f1[model_name].append(f1)
    seasonal_kappa[model_name].append(kappa)

# ==== 4. Μέση leave-one-season-out ====
print("\n===== Μέσες Leave-One-Season-Out Τιμές ανά Μοντέλο =====")
for model_name in models:
    accs = seasonal_results[model_name]
    f1s = seasonal_f1[model_name]
    kappas = seasonal_kappa[model_name]
    if len(accs) > 0:
        print(f"{model_name}: Μέση Ακρίβεια = {sum(accs)/len(accs):.4f},
Μέσο F1 = {sum(f1s)/len(f1s):.4f}, Μέσος Kappa =
{sum(kappas)/len(kappas):.4f}")

# ==== 5. Οπτικοποιήσεις ====
model_names = list(results.keys())
accuracies = [results[m]['accuracy'] for m in model_names]
f1_scores = [results[m]['f1_weighted'] for m in model_names]
kappas = [results[m]['kappa'] for m in model_names]
aucs = [results[m]['auc'] if results[m]['auc'] is not None else 0 for m in
model_names]

plt.figure(figsize=(10, 5))
plt.bar(model_names, accuracies, color='skyblue')
plt.xlabel('Models')
plt.ylabel('Accuracy')
plt.title('Model Accuracies (All Seasons Combined)')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.bar(model_names, f1_scores, color='cornflowerblue')
plt.xlabel('Models')
plt.ylabel('F1-score (weighted)')
plt.title('F1 Scores for Models (All Seasons Combined)')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.bar(model_names, kappas, color='lightgreen')
plt.xlabel('Models')
plt.ylabel("Cohen's Kappa")
plt.title("Cohen's Kappa per Model (All Seasons Combined)")
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.bar(model_names, aucs, color='salmon')
plt.xlabel('Models')
plt.ylabel("ROC-AUC (macro OvR)")
```

```
plt.title("ROC-AUC per Model (All Seasons Combined)")
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

for model_name in model_names:
    cm = results[model_name]['conf_matrix']
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
                xticklabels=[1,2,3], yticklabels=[1,2,3])
    plt.title(f'Confusion Matrix for {model_name}')
    plt.xlabel('Predicted')
    plt.ylabel('True')
    plt.tight_layout()
    plt.show()

# Leave-One-Season-Out μέσες τιμές plot
mean accuracies = [sum(seasonal_results[m])/len(seasonal_results[m]) if
len(seasonal_results[m]) > 0 else 0 for m in model_names]
mean f1s = [sum(seasonal_f1[m])/len(seasonal_f1[m]) if len(seasonal_f1[m])
> 0 else 0 for m in model_names]
mean kappas = [sum(seasonal_kappa[m])/len(seasonal_kappa[m]) if
len(seasonal_kappa[m]) > 0 else 0 for m in model_names]

plt.figure(figsize=(10, 5))
plt.bar(model_names, mean accuracies, color='lightcoral')
plt.xlabel('Models')
plt.ylabel('Mean Accuracy (Leave-One-Season-Out)')
plt.title('Mean Leave-One-Season-Out Accuracy per Model')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.bar(model_names, mean_f1s, color='darkorange')
plt.xlabel('Models')
plt.ylabel('Mean F1-score (Leave-One-Season-Out)')
plt.title('Mean Leave-One-Season-Out F1 per Model')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.bar(model_names, mean_kappas, color='mediumpurple')
plt.xlabel('Models')
plt.ylabel('Mean Cohen\'s Kappa (Leave-One-Season-Out)')
plt.title('Mean Leave-One-Season-Out Kappa per Model')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

# ==== 6. Πρόβλεψη Πρωταθλητή ανά σεζόν και μοντέλο ====
print('\n===== Πρόβλεψη Πρωταθλητή ανά σεζόν και ανά μοντέλο =====')
for model_name, model in models.items():
    print(f'\n>>> Μοντέλο: {model_name}')
    for season in seasons:
        season_df = df[df['Season'] == season].copy()
```

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

```
X_season = season_df[['ID1', 'ID2', 'Pos1', 'Pos2']]
y_true = season_df['Result']
y_pred = model.predict(X_season)
season_df['Predicted'] = y_pred

teams = pd.unique(season_df[['Team1', 'Team2']].values.ravel('K'))
points = {team: 0 for team in teams}
for _, row in season_df.iterrows():
    t1 = row['Team1']
    t2 = row['Team2']
    pred = row['Predicted']
    if pred == 1:
        points[t1] += 3
    elif pred == 3:
        points[t2] += 3
    elif pred == 2:
        points[t1] += 1
        points[t2] += 1

champion = max(points, key=points.get)
max_points = points[champion]
print(f"Σεζόν: {season} | Προβλεπόμενος Πρωταθλητής: {champion} με {max_points} βαθμούς")

# Υπολογισμός πραγματικών βαθμών για τη σεζόν
true_points = {team: 0 for team in teams}
for _, row in season_df.iterrows():
    t1 = row['Team1']
    t2 = row['Team2']
    real = row['Result']
    if real == 1:
        true_points[t1] += 3
    elif real == 3:
        true_points[t2] += 3
    elif real == 2:
        true_points[t1] += 1
        true_points[t2] += 1
    real_champion = max(true_points, key=true_points.get)
    print(f"    (Πραγματικός Πρωταθλητής: {real_champion}, {true_points[real_champion]} βαθμοί)")

# === 7. ΠΡΟΒΛΕΨΗ ΠΡΩΤΑΘΛΗΤΗ ΓΙΑ ΤΗ ΣΕΖΟΝ 2025/2026 ===
with open("season_25_26.txt", encoding="utf-8") as f:
    season_teams = [line.strip() for line in f if line.strip()]

team_to_id = {}
for team in season_teams:
    ids = pd.concat([df[df['Team1'] == team]['ID1'], df[df['Team2'] == team]['ID2']]).dropna()
    if not ids.empty:
        team_to_id[team] = ids.mode()[0]
    else:
        team_to_id[team] = max(df['ID1'].max(), df['ID2'].max()) + len(team_to_id) + 1

team_pos = {}
for team in season_teams:
    pos1 = df[df['Team1'] == team]['Pos1']
    pos2 = df[df['Team2'] == team]['Pos2']
    all_pos = pd.concat([pos1, pos2]).dropna()
    if not all_pos.empty:
```

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

```
team_pos[team] = int(round(all_pos.mean()))
else:
    team_pos[team] = int(len(season_teams) / 2)

fixtures = []
for i, home in enumerate(season_teams):
    for j, away in enumerate(season_teams):
        if home != away:
            fixtures.append({
                'Team1': home,
                'Team2': away,
                'ID1': team_to_id[home],
                'ID2': team_to_id[away],
                'Pos1': team_pos[home],
                'Pos2': team_pos[away],
            })
fixtures_df = pd.DataFrame(fixtures)

summary = []

for model_name, model in models.items():
    acc = results[model_name]['accuracy'] * 100 if 'accuracy' in
results[model_name] else None
    f1 = results[model_name]['f1_weighted'] if 'f1_weighted' in
results[model_name] else None
    kappa = results[model_name]['kappa'] if 'kappa' in results[model_name]
else None

    pred = model.predict(fixtures_df[['ID1', 'ID2', 'Pos1', 'Pos2']])
    fixtures_df['Predicted'] = pred
    points = {team: 0 for team in season_teams}
    for _, row in fixtures_df.iterrows():
        t1 = row['Team1']
        t2 = row['Team2']
        result = row['Predicted']
        if result == 1:
            points[t1] += 3
        elif result == 3:
            points[t2] += 3
        elif result == 2:
            points[t1] += 1
            points[t2] += 1
    champion = max(points, key=points.get)
    champion_points = points[champion]

    summary.append({
        "Αλγόριθμος": model_name,
        "Ακρίβεια (%)": f"{acc:.1f}" if acc is not None else "-",
        "F1-Score": f"{f1:.2f}" if f1 is not None else "-",
        "Cohen's Kappa": f"{kappa:.2f}" if kappa is not None else "-",
        "Πρωταθλητής 25/26": champion,
        "Βαθμοί": champion_points,
    })

print("\nΣυνοπτικός Πίνακας Σύγκρισης Αλγορίθμων και Πρόβλεψης
Πρωταθλητή:")
print("{:<20} {:<12} {:<10} {:<13} {:<25} {:<6}".format(
    "Αλγόριθμος", "Ακρίβεια", "F1-Score", "Kappa", "Πρωταθλητής 25/26",
    "Βαθμοί"))
print("-" * 85)
for row in summary:
```

Χρήση και αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση της νικήτριας ομάδας του ελληνικού πρωταθλήματος ποδοσφαίρου

```
print("{:<20} {:<12} {:<10} {:<13} {:<25} {:<6}".format(
    row['Αλγόριθμος'], row['Ακρίβεια (%)'], row['F1-Score'],
    row["Cohen's Kappa"], row["Πρωταθλητής 25/26"],
    row["Βαθμοί"]))

summary_df = pd.DataFrame(summary)
summary_df.to_csv("model_comparison_25_26.csv", sep="\t", index=False)
```