ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«ΠΛΗΡΟΦΟΡΙΚΗ»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**(Ελληνικά)** **ΠΡΟΓΝΩΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΑΓΩΝΩΝ ΒΟΛΕΥ ΜΕ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ** **(Αγγλικά)** **PREDICTION OF VOLLEYBALL MATCH RESULTS USING MACHINE LEARNING TECHNIQUES**
Ονοματεπώνυμο Φοιτητή	**ΓΕΩΡΓΙΑΔΗΣ ΜΗΤΡΟΦΑΝΗΣ**
Πατρώνυμο	**ΑΘΑΝΑΣΙΟΣ**
Αριθμός Μητρώου	**ΜΠΠΛ19009**
Επιβλέπων	**Δ. Σωτηρόπουλος, Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης **Δεκέμβριος 2024**

Τριμελής Εξεταστική Επιτροπή

Δ. Σωτηρόπουλος Επίκουρος Καθηγητής	Ε. Σακκόπουλος Αναπληρωτής Καθηγητής	Γ. Τσιχριντζής Καθηγητής

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Διπλωματική Εργασία

PREDICTION OF VOLLEYBALL MATCH RESULTS USING MACHINE LEARNING TECHNIQUES

Μητροφάνης Αθανασίου Γεωργιάδης

Τριμελής Εξεταστική Επιτροπή

Επίκουρος Καθηγητής Δ. Σωτηρόπουλος, Αναπληρωτής Καθηγητής Ε. Σακκόπουλος, Καθηγητής Γ. Τσιχριντζής

Πειραιάς, 2024

Περίληψη

Η συγκεκριμένη μεταπτυχιακή εργασία θα εστιάσει στην ανάλυση δεδομένων από το άθλημα της Πετοσφαίρισης (Volley Ball), με σκοπό την ανάπτυξη μιας καινοτόμου μεθόδου που θα επιτρέπει την ακριβή πρόβλεψη των αποτελεσμάτων των αγώνων. Για την υλοποίηση αυτού του στόχου, θα αξιοποιηθεί η μεθοδολογία της μηχανικής μάθησης, η οποία θα διασφαλίσει την αποτελεσματική και αποδοτική επεξεργασία των δεδομένων, επιτρέποντας την παραγωγή αξιόπιστων και χρήσιμων αποτελεσμάτων.

Στην παρούσα εργασία, θα πραγματοποιηθεί ανασκόπηση της σχετικής βιβλιογραφίας, εστιάζοντας σε προηγούμενες μελέτες και πηγές που έχουν διεξαχθεί σχετικά με την πρόβλεψη αποτελεσμάτων στον αθλητισμό. Αυτή η ανασκόπηση θα παρέχει ένα πλαίσιο κατανόησης των προκλήσεων και των επιτευγμάτων στον τομέα, αναδεικνύοντας τη σημασία της ανάλυσης δεδομένων και των σύγχρονων τεχνολογιών.

Για την επεξεργασία των δεδομένων θα χρησιμοποιηθούν διάφορα προγράμματα, μεταξύ των οποίων το Excel και η Python, τα οποία θα διευκολύνουν τη διαχείριση και ανάλυση των στοιχείων. Η Python, ειδικότερα, θα είναι κρίσιμη για την εφαρμογή αλγορίθμων μηχανικής μάθησης και για την παραγωγή των αναγκαίων στατιστικών αποτελεσμάτων.

Στη συνέχεια της εργασίας, θα παρουσιαστούν τα αποτελέσματα της εφαρμογής αυτών των μεθόδων σε ένα πλούσιο σύνολο δεδομένων. Τα δεδομένα που θα χρησιμοποιηθούν περιλαμβάνουν τα αποτελέσματα των αγώνων της Volley League αντρών στην Ελλάδα από το 2011 έως και σήμερα, προσφέροντας μια εκτενή βάση για την ανάλυση και τις συγκρίσεις.

Κλείνοντας, η εργασία αυτή θα αναφερθεί στις προκλήσεις που ενδέχεται να προκύψουν κατά την ανάπτυξη προγραμμάτων πρόβλεψης αποτελεσμάτων, καθώς και στις προοπτικές που διανοίγονται για τη χρήση παρόμοιων μεθόδων όχι μόνο στην Πετοσφαίριση, αλλά και σε άλλα αθλήματα. Η εμβάθυνση στην ανάλυση δεδομένων και η εφαρμογή προηγμένων τεχνικών μπορεί να προσφέρει σημαντικά οφέλη για προπονητές, αθλητές και αναλυτές, συμβάλλοντας στην βελτίωση των στρατηγικών και της απόδοσης στους αγώνες.

Λέξεις Κλειδιά: Volley Ball, Πετοσφαίριση, Βόλεϊ, Volley League, Random Forest, Python, Machine Learning.

Abstract

This specific master's thesis will focus on the analysis of data from the sport of volleyball, with the aim of developing an innovative method that will allow for the accurate prediction of match outcomes. To achieve this goal, the methodology of machine learning will be utilized, ensuring effective and efficient data processing, thereby enabling the production of reliable and useful results.

In this thesis, a review of the relevant literature will be conducted, focusing on previous studies and sources related to outcome prediction in sports. This review will provide a framework for understanding the challenges and achievements in the field, highlighting the importance of data analysis and modern technologies.

Various programs will be used for data processing, including Excel and Python, which will facilitate the management and analysis of the data. Python will be critical for implementing machine learning algorithms and producing the necessary statistical results.

Subsequently, the results of applying these methods to a rich dataset will be presented. The data used will include the results of men's matches in the Volley League in Greece from 2011 to the present, offering an extensive foundation for analysis and comparisons.

Finally, this work will address the challenges that may arise during the development of prediction programs, as well as the prospects for using similar methods not only in volleyball but also in other sports. Delving into data analysis and applying advanced techniques can provide significant benefits for coaches, athletes, and analysts, contributing to the improvement of strategies and performance in matches.

Keywords: Volleyball, Volley League, Random Forest, Python, Machine Learning.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου στην οικογένειά μου και στους φίλους μου που με στήριξαν και με άντεξαν κατά τη διάρκεια των σπουδών μου και της διπλωματικής μου εργασίας. Η αμέριστη στήριξη και η πίστη τους σε μένα αποτέλεσαν τον βασικό πυλώνα της επιτυχίας μου σε αυτή τη δύσκολη και απαιτητική περίοδο.

Ένα ιδιαίτερο ευχαριστώ ανήκει στον καθηγητή μου, Δρ. Διονύση Σωτηρόπουλο, από το Τμήμα Πληροφορικής του Πανεπιστημίου Πειραιά, που με καθοδήγησε καθ' όλη τη διάρκεια αυτής της εργασίας. Η βοήθειά του και οι εποικοδομητικές παρατηρήσεις του υπήρξαν καίρια για την ολοκλήρωση της εργασίας μου. Η καθοδήγηση και οι χρήσιμες συμβουλές του ήταν καθοριστικές για την επιτυχία της μελέτης μου.

Η ολοκλήρωση αυτής της διπλωματικής εργασίας σηματοδοτεί το τέλος ενός εξαιρετικά σημαντικού ταξιδιού στην ακαδημαϊκή μου πορεία. Υπήρξα τυχερός να έχω δίπλα μου εξαιρετικούς καθηγητές και συνεργάτες που συνέβαλαν καθοριστικά στην ανάπτυξή μου τόσο στον τομέα της γνώσης όσο και προσωπικά.

Ευχαριστώ θερμά όλους τους φίλους και συμφοιτητές μου που με στήριξαν όλα αυτά τα χρόνια και ιδιαίτερα τους συνεργάτες που είχα σε εργασίες και ασκήσεις. Η παρουσία σας έκανε την πορεία αυτή πιο ευχάριστη και υποστηρικτική.

Τέλος, δεν μπορώ παρά να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου, που πάντα πίστευε σε μένα και με ενθάρρυνε να προσπαθώ για τα όνειρά μου. Η υποστήριξή τους με γέμισε δύναμη και όραμα για το μέλλον.

Σας ευχαριστώ όλους από καρδιάς!

Πίνακας περιεχομένων

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Εισαγωγή

Πρόγνωση: η πρόβλεψη της εξέλιξης μιας κατάστασης ή κάποιων γεγονότων, βάσει της ορθής εκτίμησης δεδομένων. Ετυμολογικά, η λέξη προέρχεται από την αρχαία ελληνική «πρόγνωσις» (πρό + γνῶσις), όπου το «προ» δηλώνει το «πριν», «προς τα εμπρός» ή «εκ των προτέρων», ενώ το «γνώσις» σημαίνει «γνώση», «αντίληψη» ή «κατανόηση». Συνεπώς, «πρόγνωση» σημαίνει «η γνώση πριν» ή «η γνώση εκ των προτέρων», υποδηλώνοντας την προσπάθεια πρόβλεψης του μέλλοντος με βάση την υπάρχουσα γνώση. Αυτή η ετυμολογική ρίζα αντικατοπτρίζει την ουσία της έννοιας: η πρόβλεψη βασίζεται στην συστηματική ανάλυση ποσοτικών και ποιοτικών δεδομένων, συχνά συλλεγμένων από προηγούμενες παρατηρήσεις, όπως στατιστικά στοιχεία, ιστορικά αρχεία, αποτελέσματα πειραμάτων ή ακόμη και προσομοιώσεις. Η διαδικασία αυτή στοχεύει στην ακριβέστερη εκτίμηση της έκβασης μιας κατάστασης ή γεγονότων στο μέλλον, λαμβάνοντας υπόψη τις πιθανές αλληλεπιδράσεις μεταξύ των διαφόρων παραγόντων που επηρεάζουν το φαινόμενο. Η πρόγνωση δεν αποτελεί απλή μαντεία ή εικασία, αλλά μια προσπάθεια επιστημονικής προσέγγισης με εργαλεία τη λογική, τη στατιστική και την εμπειρία, με απώτερο σκοπό τη λήψη καλύτερων αποφάσεων και τον σχεδιασμό αποτελεσματικότερων δράσεων.

Στην αρχαία Ελλάδα είχαν αναπτύξει την τέχνη της μαντείας (πρόβλεψη των μελλούμενων γεγονότων,) σε πάρα πολύ μεγάλο βαθμό και την χρησιμοποιούσαν προκειμένου να μπορέσουν να προβλέψουν το μέλλον. Από την εποχή των 1500-2000 π.Χ., οι Έλληνες, επηρεασμένοι από τον σαμανισμό της Θράκης, πίστευαν στην ύπαρξη ψυχής, η οποία ήταν πιο ενεργητική κατά τον ύπνο. Η μαντική, τόσο στην άτεχνη όσο και στην έντεχνη μορφή της, χρησιμοποιήθηκε για την λήψη αποφάσεων σε όλους τους τομείς της ζωής, από στρατιωτικές εκστρατείες και ιδρύσεις αποικιών μέχρι καθημερινά ζητήματα όπως η υγεία και η οικογένεια. Η μαντική τέχνη, αρχικά συνδεδεμένη με τη θεραπευτική, άρχισε να απομπλέκεται από ιερουργίες με τον Ιπποκράτη, ανοίγοντας τον δρόμο για την επιστημονική ιατρική.

Ιστορικά τώρα καταγεγραμμένο είναι ότι οι νικητές των Ολυμπιακών Αγώνων αποκτούσαν μεγάλο κύρος και φήμη, όχι μόνο κατά την εποχή τους αλλά και για τις επόμενες γενιές. Οι τιμές που τους αποδίδονταν ήταν ιδιαίτερα σημαντικές και οι αθλητές απολάμβαναν ειδική μεταχείριση από την κοινωνία, με την αναγνώριση των επιτευγμάτων τους να εκδηλώνεται με πολλούς τρόπους, όπως μνημεία και αγάλματα προς τιμήν τους.

Από τους αρχαίους Ολυμπιακούς Αγώνες, όπου η πρόβλεψη του νικητή αποτελούσε αντικείμενο μαντείας, μέχρι τη σημερινή εποχή της πληροφορίας. Η πρόγνωση των αποτελεσμάτων σε αγώνες έχει τις ρίζες της σε εκείνα τα χρόνια, καθώς οι θεατές προσπαθούσαν να μαντέψουν ποιος θα είναι ο νικητής. Για το λόγο αυτό, επισκέπτονταν τα πιο γνωστά μαντεία της εποχής και ζητούσαν από τους μάντεις να τους παρέχουν προβλέψεις για τους νικητές. Σύμφωνα με ορισμένες αναφορές, οι

ιερείς χρησιμοποιούσαν τη μέθοδο της μαντείας βάζοντας τα δάχτυλά τους σε ένα υγρό, συνήθως από δαφνέλαιο, και στη συνέχεια προσπαθούσαν να διακρίνουν τον νικητή, μια πρακτική που έχει συνδεθεί με τη φράση «δε μύρισα τα νύχια μου» όταν η γνώση ήταν αβέβαιη.

Στον αθλητισμό, η πρόγνωση του νικητή αποτελεί ένα από τα πιο συναρπαστικά στοιχεία, τόσο για τους φιλάθλους όσο και για τους επαγγελματίες του χώρου. Παραδοσιακά, οι προβλέψεις βασίζονταν σε εμπειρικές εκτιμήσεις, την άποψη των ειδικών και την ιστορική εμπειρία. Ωστόσο, η ραγδαία ανάπτυξη της τεχνολογίας και η διαθεσιμότητα μεγάλων όγκων δεδομένων, που περιλαμβάνουν στατιστικά αγώνων, δεδομένα παικτών, ακόμη και πληροφορίες σχετικές με τις συνθήκες διεξαγωγής των αγώνων, έχουν ανοίξει νέους ορίζοντες στην ανάλυση και την πρόγνωση αθλητικών αποτελεσμάτων.

Στους πρώτους σύγχρονους Ολυμπιακούς Αγώνες του 1896, αλλά και στις μετέπειτα διοργανώσεις, η πρόγνωση των αποτελεσμάτων δεν ήταν οργανωμένη όπως τη γνωρίζουμε σήμερα. Η παράδοση της πρόγνωσης των αποτελεσμάτων ξεκίνησε με τις πρώτες αναφορές και εκτιμήσεις από αθλητικούς δημοσιογράφους της εποχής, οι οποίοι εξέφραζαν την άποψή τους για τους πιθανούς νικητές σε διάφορα αγωνίσματα, προσφέροντας μια αρχική μορφή ανάλυσης και προβλέψεων. Οι εκτιμήσεις αυτές βασίζονταν στις πληροφορίες που γνώριζαν οι δημοσιογράφοι που κύρια απασχόλησή τους ήταν ο αθλητισμός.

Από τότε, η επιθυμία για την πρόβλεψη των αποτελεσμάτων έχει αυξηθεί. Με τη βοήθεια της επιστήμης και της τεχνολογίας, έχουν δημιουργηθεί νέες τεχνικές και μέθοδοι πρόγνωσης που εφαρμόζονται στον αθλητισμό. Η ανάπτυξη μαθηματικών μοντέλων, νευρωνικών δικτύων και τεχνικών μηχανικής μάθησης έχει επιτρέψει την αποτελεσματική επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων. Αυτή η δυνατότητα ανάλυσης δεδομένων αποτελεί πλέον θεμελιώδη λίθο για την ακριβή πρόγνωση, τόσο στον αθλητισμό όσο και σε άλλους τομείς.

Τα τελευταία χρόνια, η τεχνολογική εξέλιξη έχει οδηγήσει στην άνθηση διαδικτυακών πλατφορμών και εφαρμογών που προσφέρουν υπηρεσίες πρόγνωσης αποτελεσμάτων για ένα ευρύ φάσμα αθλημάτων, καλύπτοντας τις ανάγκες των φιλάθλων, των στοιχηματικών εταιρειών, των αθλητικών οργανισμών και των επαγγελματιών του χώρου που προσφέρουν υπηρεσίες πρόγνωσης αποτελεσμάτων. Η ανάπτυξη προγνωστικών μοντέλων στον αθλητισμό βασίζεται στην ανάλυση δεδομένων που σχετίζονται με την φυσική κατάσταση των αθλητών, τις επιδόσεις τους σε προηγούμενους αγώνες, τη στρατηγική της ομάδας, καθώς και εξωτερικούς παράγοντες, όπως η έδρα του αγώνα ή οι καιρικές συνθήκες

Η πρόγνωση αποτελεί κρίσιμο στοιχείο στον αθλητισμό, καθώς επιτρέπει τη δημιουργία στρατηγικών, την αξιολόγηση της απόδοσης των αθλητών και τη διαμόρφωση προσδοκιών από το κοινό. Παράλληλα, τα τελευταία χρόνια, η μηχανική μάθηση, με τη δυνατότητα ανάλυσης μεγάλων δεδομένων, προσφέρει νέα εργαλεία και μεθόδους για την ακριβέστερη πρόγνωση των αποτελεσμάτων. Στον αθλητισμό, η πρόβλεψη αποτελεσμάτων βασιζόταν παραδοσιακά σε εμπειρικά δεδομένα, αλλά η εξέλιξη της τεχνολογίας και της τεχνητής νοημοσύνης επιτρέπει την επεξεργασία τεράστιων ποσοτήτων δεδομένων, προσφέροντας ακριβέστερες και πιο αξιόπιστες προβλέψεις.

Σύμφωνα με έρευνες στον τομέα της αθλητικής ανάλυσης, η μηχανική μάθηση έχει εφαρμοστεί με επιτυχία σε αθλήματα όπως το ποδόσφαιρο και το μπάσκετ, όπου αλγόριθμοι όπως οι Random Forests και Neural Networks έχουν χρησιμοποιηθεί για την πρόβλεψη των αποτελεσμάτων και την αξιολόγηση της απόδοσης των παικτών. Παρόλα αυτά, στον τομέα της Πετοσφαίρισης, η πρόγνωση αποτελεσμάτων δεν έχει μελετηθεί εκτενώς, οι μελέτες που εστιάζουν στην πρόγνωση αγώνων παραμένουν περιορισμένες, και αυτή η εργασία επιχειρεί να βοηθήσει ώστε να γεφυρώσει αυτό το κενό, χρησιμοποιώντας σύγχρονες μεθόδους μηχανικής μάθησης και δεδομένα που καλύπτουν αρκετά χρόνια. Η ανάλυση δεδομένων από αγώνες της Volley League αντρών της Ελλάδας, από το 2011 έως και σήμερα, και η εφαρμογή σύγχρονων τεχνικών μηχανικής μάθησης αναμένεται να προσφέρει πολύτιμα συμπεράσματα για την απόδοση των ομάδων και να συμβάλει στην ανάπτυξη ενός ακριβούς και αξιόπιστου προγνωστικού μοντέλου.

Το αποτέλεσμα κάθε αγώνα Πετοσφαίρισης εξαρτάται από ποικίλους παράγοντες, όπως η φυσική κατάσταση των παικτών, η στρατηγική, η δυναμική της ομάδας και οι συνθήκες του αγώνα (π.χ., έδρα, εξωτερικοί παράγοντες κ.λπ.). Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για να εντοπίσουν συσχετίσεις μεταξύ αυτών των παραμέτρων και να προβλέψουν με μεγαλύτερη ακρίβεια ποια ομάδα έχει περισσότερες πιθανότητες να κερδίσει.

Η συγκεκριμένη έρευνα στοχεύει στην ανάπτυξη ενός προγνωστικού μοντέλου που θα αξιοποιεί τα δεδομένα από τους αγώνες της Volley League αντρών στην Ελλάδα για να προσφέρει προβλέψεις με υψηλή ακρίβεια. Για την δημιουργία αυτού του μοντέλου, θα δοκιμαστούν διάφοροι αλγόριθμοι μηχανικής μάθησης, προκειμένου να βρεθεί ο καταλληλότερος που θα έχει την υψηλότερη ορθότητα στις προβλέψεις του. Η επιλογή του κατάλληλου αλγορίθμου είναι κρίσιμη για την ακρίβεια των προβλέψεων, καθώς κάθε αλγόριθμος διαθέτει διαφορετικά χαρακτηριστικά και δυνατότητες. Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης, όπως οι KNeighborsClassifier, GaussianNB, και Random Forest, που θα οδηγήσουν στην εξαγωγή πολύτιμων συμπερασμάτων για τους παράγοντες που επηρεάζουν το αποτέλεσμα των αγώνων και την ακριβέστερη εκτίμηση των πιθανών νικητών.

Βιβλιογραφική Επισκόπηση

Αξίζει να σημειωθεί ότι δεν υπάρχουν αρκετές εργασίες που ασχολούνται με τα Μοντέλα Μηχανικής Μάθησης και την πρόβλεψη των αποτελεσμάτων αγώνων Βόλεϊ. Αυτό έχει να κάνει με την απήχηση που έχει το συγκεκριμένο άθλημα. Από την αναζήτηση τόσο στο διαδίκτυο αλλά στο Scholar της Google οι περισσότερες εργασίες ασχολούντο με αθλήματα που έχουν μεγαλύτερη τηλεθέαση και είναι πιο γνωστά στο κοινό, παρά με εκείνα που δεν είναι ιδιαίτερα δημοφιλή. Τα αποτελέσματα της αναζήτησης δείχνουν ότι το μεγαλύτερο ποσοστό των εργασιών αφορούν αγώνες ποδοσφαίρου, μπάσκετ και τένις.

Στην εργασία του, ο Σ. Αρμενιάκου, χρησιμοποίησε δεδομένα αγώνων μπάσκετ. Στα συμπεράσματά του αναφέρει ότι τα Μοντέλα Συνόλου είναι πιο επιτυχημένα από τα βασικά Μοντέλα Μηχανικής Μάθησης. Αξίζει να σημειωθεί ότι τα δεδομένα που χρησιμοποίησε δεν είχαν τα ίδια χαρακτηριστικά με την παρούσα εργασία αλλά συμπεριελάμβαναν μεγαλύτερη πληθώρα χαρακτηριστικών, όπως ενέργειες των αθλητών. Ο μεγαλύτερος όγκος δεδομένων πιθανότατα να συνεπάγει τη μεγαλύτερη συσχέτιση μεταξύ των χαρακτηριστικών, το γεγονός αυτό εμποδίζει ένα μοντέλο να αξιοποιήσει την ουσιαστική πληροφορία που θα «κρύβεται» στο σύνολο εκπαίδευσης. (Αρμενιάκου, 2022)

Στην εργασία του ο Φ. Νικολόπουλος αναφέρει τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν, Μηχανικής Μάθησης Δάσους Αποφάσεων, Τυχαίου Δάσους και LightGBM, δεν παρουσίασαν μεγάλες διαφοροποιήσεις στις εκτιμήσεις τους. Ωστόσο ο ταξινομητής Τυχαίου Δάσους (Random Forest) ήταν αυτός που είχε τα καλύτερα αποτελέσματα σε σχέση με τις Banchmark προβλέψεις, στις οποίες δεν χρησιμοποιείται Μηχανική μάθηση. Θα πρέπει να σημειωθεί ότι η εργασία αφορά αποτελέσματα αγώνων τένις. (Νικολουλόπουλος, 2023)

Στην εργασία τους, οι Rodrigues και Pinto, που παρουσιάστηκε στο International Conference on Industry Sciences and Computer Science Innovation το 2022 παρουσιάζουν την ανάπτυξη μοντέλων τεχνικών μηχανικής μάθησης. Σύμφωνα με την εργασίας τους, χρησιμοποίησαν δεδομένα από αγώνες ποδοσφαίρου της σεζόν 2016-2017 της Αγγλικής Premier League. Αφού ολοκλήρωσαν την ανάλυση και την σύγκριση των διαφόρων αλγορίθμων, δημιούργησαν ένα μοντέλο με ακρίβεια 65,26%, υψηλότερο σε σχέση με προηγούμενες μελέτες. Ο αλγόριθμος που χρησιμοποίησαν για το μοντέλο τους ήταν ο Random Forest. (Rodrigues & Pinto, 2022)

Επίσης ο Herbinet σε εργασία που έκανε για το ποδόσφαιρο στο Imperial College στο Λονδίνο, παρουσίασε τα αποτελέσματα των μοντέλων ταξινόμησης και παλινδρόμησης για την πρόβλεψη αποτελεσμάτων αγώνων. Στην εργασία αυτή αναφέρει ότι εν τέλει το μοντέλο παλινδρόμησης που

χρησιμοποίησε δεν ήταν τόσο αποτελεσματικό με τα δεδομένα που χρησιμοποίησε. (HERBINET, 2018)

Το 2024, ο Μαζάι, στην εργασία του, χρησιμοποίησε ένα μοντέλο πρόβλεψης ποδοσφαιρικών αγώνων που κατάφερε να πετύχει ακρίβεια άνω του 80% στη δυαδική ταξινόμηση (νίκη/ήττα) και 60% στην πολλαπλή ταξινόμηση (νίκη/ισοπαλία/ήττα). Στην εργασία του αναφέρει ότι η βελτιωμένη ακρίβεια οφείλεται σε διάφορους παράγοντες, όπως οι χρηματιστηριακές αξίες των παικτών και η ποικιλία του dataset (πολλαπλές χρονιές και διοργανώσεις). Η δημιουργία νέων χαρακτηριστικών μέσω της διαδικασίας feature engineering και η βελτιστοποίηση των παραμέτρων των αλγορίθμων (hyperparameter tuning) συνέβαλαν επίσης στην αύξηση της απόδοσης του μοντέλου. Τέλος, αναφέρει ότι η συστηματική προσέγγιση που υιοθέτησε ίσως να έχει την δυνατότητα να ανοίξει το δρόμο για πιο αναβαθμισμένες προβλέψεις στον τομέα της ποδοσφαιρικής ανάλυσης. Επίσης αναφέρει ότι η απόλυτη ακρίβεια δεν μπορεί να επιτευχθεί λόγω των απρόβλεπτων παραγόντων και του στοιχείου της έκπληξης που χαρακτηρίζει τον αθλητισμό. (MAZAI, 2024)

Τον Ιούνιο του 2021 στην εργασία του ο Σ. Γερδέλης προκειμένου να μπορέσει να προβλέψει την πρόκριση των ομάδων στα playoffs σε Euroleague και NBA, χρησιμοποίησε στατιστικά δεδομένα από την κανονική περίοδο. Μελέτησε τρία σενάρια ανάλυσης: standardization, Pearson correlation και επιλογή avg χαρακτηριστικών, για κάθε σενάριο, εφαρμόζονται 5 κατηγοριοποιητές (SVM, LogReg, KNN, RFC, MLP) για την πρόβλεψη. Τα αποτελέσματα που βρήκε έδειξαν ότι το σενάριο standardization υπερέχει, ενώ η λογαριθμική παλινδρόμηση αποδεικνύεται ότι είναι ο πιο αποδοτικός αλγόριθμος, για την συγκεκριμένη εργασία. Επίσης πρότεινε σε μελλοντική έρευνα να συμπεριληφθούν τα στατιστικά playoffs και εφαρμογή του πιο αποδοτικού αλγορίθμου στο αντίθετο dataset. (Γαρδέλης, 2021)
Στην διπλωματική εργασία του, ο Μιχαλάκης εστίασε στην πρόβλεψη αποτελεσμάτων αγώνων Euroleague, αξιοποιώντας στατιστικά δεδομένα από το 2000 έως το 2020. Χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης, όπως SVM, Logistic Regression, K-NN, και Random Forest, προσπάθησε να ανακαλύψει την ικανότητα πρόβλεψης των αλγορίθμων μέσα από διάφορα σενάρια. Αρχικά εφάρμοσε τους αλγόριθμους χωρίς επιλογή χαρακτηριστικών, με την Logistic Regression να εμφανίζει καλύτερα αποτελέσματα. Στη συνέχεια, δοκίμασε τρία διαφορετικά σενάρια επιλογής χαρακτηριστικών: filtered, wrapper και embedded. Στο σενάριο filtered, ο SVM και η Logistic Regression υπερείχαν, ενώ στο wrapper, η Logistic Regression και τα Random Forests εμφάνισαν καλύτερη απόδοση. Τέλος, στο σενάριο embedded, η Logistic Regression L1 και τα Random Forests κατάφερε να επιτύχει την καλύτερη απόδοση. (Μιχαλάκης, 2024)

Επίσης η Αιγινίτη, στην διπλωματική της εργασία που αφορά όμως τον μηχανοκίνητο αθλητισμό, προσπάθησε να αναζητήσει τον ποιον κατάλληλο αλγόριθμο μηχανικής μάθησης που θα επιτύγχανε

την καλύτερη πρόβλεψη θέσεων τερματισμού στην Formula 1. Στην εργασίας της εφάρμοσε διάφορους αλγόριθμους (όπως παλινδρόμηση, συλλογικές μεθόδους βασισμένες σε δένδρα απόφασης και νευρωνικά δίκτυα) σε δεδομένα F1 που αφορούσαν οδηγούς, κατασκευαστές και αποτελέσματα αγώνων. Τα αποτελέσματα έδειξαν ότι το μοντέλο τμηματικών πολυωνύμων (piecewise polynomial splines) και το extreme gradient boosting (XGBoost) κατάφεραν να επιτύχουν την καλύτερη πρόβλεψη θέσεων τερματισμού, ενώ τα μοντέλα decision tree και AdaBoost είχαν χαμηλότερη απόδοση. (Αιγινίτη, 2024)

Από τα παραπάνω θα μπορούσαμε να βγάλουμε το συμπέρασμα ότι η έρευνα για την πρόβλεψη αποτελεσμάτων αγώνων βόλεϊ με τη χρήση μοντέλων μηχανικής μάθησης είναι αρκετά περιορισμένη, σε αντίθεση φυσικά με αθλήματα με μεγαλύτερη δημοτικότητα, όπως είναι το ποδόσφαιρο, το μπάσκετ και το τένις.

Σε αυτά τα αθλήματα, παρόλο που υπάρχουν μελέτες που αξιοποιούν μοντέλα μηχανικής μάθησης για την πρόβλεψη αποτελεσμάτων, η αποτελεσματικότητα των μοντέλων ποικίλλει ανάλογα με το άθλημα και το dataset. Τα μοντέλα συνόλου, όπως τα Random Forest, έχουν αποδειχθεί από τα πιο επιτυχημένα, ενώ η επιτυχία των μοντέλων εξαρτάται, κατά κύριο λόγο από τον όγκο και την ποιότητα των δεδομένων. Επίσης η ακρίβεια των προβλέψεων εξαρτάται και από την πολυπλοκότητα του αθλήματος. Γενικότερα πάντως, σύμφωνα με τις παραπάνω μελέτες για το ποδόσφαιρο, η ακρίβεια των προβλέψεων μπορεί να φτάσει πάνω από 80% εάν χρησιμοποιηθεί δυαδική ταξινόμηση (νίκη/ήττα). Αυτό είναι κάτι που θα μπορούσε να χρησιμοποιηθεί ως βάση για την πρόβλεψη των αποτελεσμάτων σε αγώνες βόλεϊ, καθώς το αποτέλεσμα των αγώνων μπορεί να είναι νίκη ή ήττα. Ωστόσο, η απόλυτη ακρίβεια δεν μπορεί να επιτευχθεί λόγω των απρόβλεπτων παραγόντων και του στοιχείου της έκπληξης που χαρακτηρίζει γενικότερα τον αθλητισμό, λόγω εξωγενών παραγόντων.

Μεθοδολογίες Μηχανικής Μάθησης

Λογιστική παλινδρόμηση

Η παλινδρόμηση είναι μία στατιστική τεχνική μοντελοποίησης που χρησιμοποιείται ευρέως για την συσχέτιση μιας εξαρτώμενης μεταβλητής και μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. (Παλινδρόμηση (στατιστική), 2023)

Η λογιστική παλινδρόμηση χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών για να μοντελοποιήσει την πιθανότητα κατηγοριοποίησης μίας παρατήρησης. Όταν χρειάζεται να προβλέψουμε την ύπαρξη ή μη ενός χαρακτηριστικού, τότε μπορούμε να χρησιμοποιήσουμε την

λογιστική παλινδρόμηση για να εκτιμήσουμε την πιθανότητα εμφάνισης του. Το παραπάνω μοντέλο δημιουργεί μία μη γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και της πιθανότητας της εξαρτημένης μεταβλητής, χρησιμοποιώντας μία λογιστική συνάρτηση. Έτσι δίνει την δυνατότητα της εκτίμησης της επίδρασης κάθε ανεξάρτητης μεταβλητής που θα μπορούσε να διαμορφώσει τις πιθανότητες των εξαρτημένων μεταβλητών. (ΣΚΟΥΦΑ, 2008)

Πιο συγκεκριμένα η λογιστική παλινδρόμηση χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών (x_1, x_2, ..., x_n) για να μοντελοποιήσει την πιθανότητα (P) μιας δυαδικής εξαρτημένης μεταβλητής (Y) να ανήκει σε μια συγκεκριμένη κατηγορία (συνήθως Y=1). Όταν χρειάζεται να προβλέψουμε την ύπαρξη ή μη ενός χαρακτηριστικού, τότε μπορούμε να χρησιμοποιήσουμε τη λογιστική παλινδρόμηση για να εκτιμήσουμε την πιθανότητα εμφάνισής του. Το μοντέλο αυτό δημιουργεί μια μη γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και της πιθανότητας της εξαρτημένης μεταβλητής, χρησιμοποιώντας τη λογιστική συνάρτηση (sigmoid function):

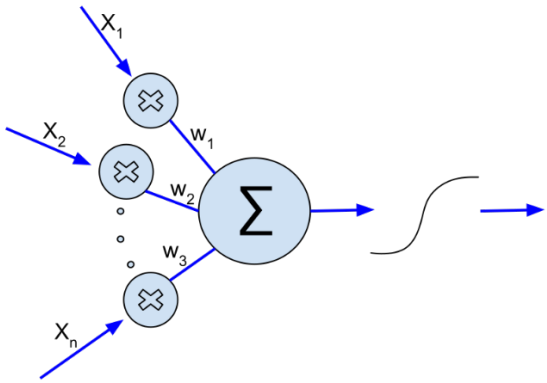
$P(Y=1 \mid x_1, x_2, ..., x_n) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)})$

όπου β_0 είναι η σταθερά (intercept) και β_1, β_2, ..., β_n είναι οι συντελεστές που αντιστοιχούν σε κάθε ανεξάρτητη μεταβλητή. Η λογιστική συνάρτηση μετατρέπει τη γραμμική συνάρτηση των ανεξάρτητων μεταβλητών σε μια πιθανότητα μεταξύ 0 και 1. Οι συντελεστές (β) υπολογίζονται συνήθως με τη μέθοδο της μέγιστης πιθανότητας (Maximum Likelihood Estimation - MLE) και η ερμηνεία τους γίνεται μέσω του λόγου αποδόσεων (Odds Ratio), όπου e^{β_i} αντιπροσωπεύει τον λόγο αποδόσεων και δείχνει την πολλαπλασιαστική αλλαγή στον λόγο αποδόσεων για μια αύξηση κατά μία μονάδα στην αντίστοιχη μεταβλητή x_i.

Νευρωνικά δίκτυα

Ιστορική αναδρομή

Το πρώτο μοντέλο νευρωνικού δικτύου παρουσιάστηκε το 1943 από τους McCulloch και Pitts. (W. S. McCulloch and W. Pitts, 1943)



Εικόνα 1 – Διάγραμμα ενός τεχνητού νευρώνα McCulloch-Pitts *(Alhama, 2017)*

Το μοντέλο McCulloch-Pitts αποτελεί μια απλοποιημένη μαθηματική αναπαράσταση ενός βιολογικού νευρώνα. Λειτουργεί ως μια λογική πύλη που δέχεται n εισόδους (x_1, x_2, ..., x_n), καθεμία με το αντίστοιχο βάρος (w_1, w_2, ..., w_n). Η έξοδος (y) του νευρώνα υπολογίζεται μέσω της συνάρτησης:

$$y = f(\Sigma(wi * xi) - \theta$$

όπου Σ(wi * xi) είναι το άθροισμα των σταθμισμένων εισόδων, θ είναι το κατώφλι και f είναι η συνάρτηση ενεργοποίησης, συνήθως η συνάρτηση βηματισμού (συνάρτηση Heaviside):

$$f(x) = H(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

Με κατάλληλη επιλογή βαρών και κατωφλιού, το μοντέλο μπορεί να αναπαραστήσει βασικές λογικές πύλες όπως AND, OR και NOT. Ωστόσο, έχει περιορισμούς, καθώς δεν μπορεί να αναπαραστήσει όλες τις λογικές συναρτήσεις με έναν μόνο νευρώνα.

Λίγα χρόνια πιο μετά, το 1949, ο Hebb δημοσίευσε το βιβλίο "Organization of behavior” μέσα στο οποίο αναφέρει:

«When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased. » (HEBB, 1949)

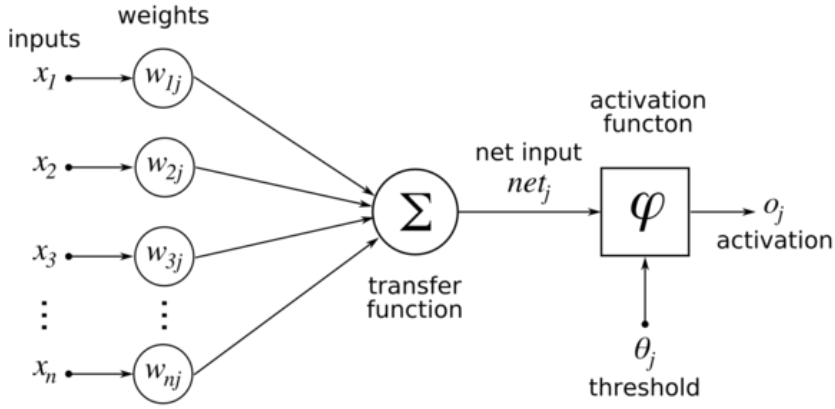
Ελεύθερη μετάφραση: *«Όταν ένας άξονας του κυττάρου Α είναι αρκετά κοντά για να διεγείρει ένα κύτταρο Β και επανειλημμένα ή επίμονα συμμετέχει στην πυροδότησή του, κάποια διαδικασία ανάπτυξης ή μεταβολική αλλαγή λαμβάνει χώρα σε ένα ή και στα δύο κύτταρα, με αποτέλεσμα η αποτελεσματικότητα του Α, ως ένα από τα κύτταρα που πυροδοτούν το Β, να αυξάνεται.»*

Η πρόταση αυτή του Hebb έγινε η βάση για την συσχετιστική μάθηση σε κυτταρικό επίπεδο. Μεταφέροντας την πρόταση του Hebb στο χώρο των νευρωνικών δικτύων, παρήχθησαν οι παρακάτω κανόνες:

1. Αν δύο νευρώνες εκατέρωθεν της σύνδεσης ενεργοποιούνται ταυτόχρονα, τότε η ισχύς της σύνδεσης αυξάνεται, δηλαδή αυξάνεται η τιμή του βάρους σύνδεσής τους.

2. Αν δύο νευρώνες εκατέρωθεν της σύνδεσης ενεργοποιούνται ασύγχρονα, η σύνδεσή τους εξασθενεί (μειώνεται η τιμή του βάρους σύνδεσης) ή και εξαφανίζεται. Με τον τρόπο αυτό συμπεριλαμβάνουμε και την "αρνητική μάθηση". Μια σύναψη με αυτά τα χαρακτηριστικά ονομάζεται hebbian synapse και επειδή υπεύθυνος για τις αλλαγές είναι ο συσχετισμός της προσυνδετικής και μετασυνδετικής λειτουργίας, αναφέρεται επίσης και σαν correlational synapse. (Πανεπιστήμιο Πατρών, 2024)

Αργότερα, το 1957, εφευρέθηκε ο νευρώνας Perceptron από τον Φρανκ Ρόζενμπλατ. Ο νευρώνας αυτός είναι ένα είδος τεχνητού νευρωνικού δικτύου και μπορεί να χαρακτηριστεί ως ένα απλό είδος εμπροσθοτροφοδοτούμενο (feed-forward) νευρωνικό δίκτυο, ένας γραμμικός ταξινομητής (linear classifier). (Wikipedia, 2024)



Εικόνα 2 – Το Perceptron του Φρανκ Ρόζενμπλατ (Frank Rosenblatt) *(Mitchell, 2012)*

Το Perceptron, ένας γραμμικός ταξινομητής, δέχεται n εισόδους (x_1, x_2, ..., x_n) με αντίστοιχα βάρη (w_1, w_2, ..., w_n) και μια πόλωση (b). Το άθροισμα των σταθμισμένων εισόδων υπολογίζεται ως

$$\text{net} = \Sigma(\text{wi} * \text{xi}) + \text{b}$$

Η έξοδος (y) προκύπτει εφαρμόζοντας μια συνάρτηση ενεργοποίησης φ στο net, συνήθως μια συνάρτηση βηματισμού (φ(x) = 1 αν x >= 0, και φ(x) = 0 αν x < 0) ή πρόσημο (φ(x) = +1 αν x > 0, και φ(x) = -1 αν x <= 0). Γεωμετρικά, το Perceptron ορίζει ένα υπερεπίπεδο

$$\Sigma(wi * xi) + b = 0$$

που χωρίζει τις εισόδους σε δύο κατηγορίες. Ο αλγόριθμος εκπαίδευσης του Perceptron προσαρμόζει τα βάρη και την πόλωση ώστε να ταξινομεί σωστά τα δεδομένα, αλλά έχει τον περιορισμό να λειτουργεί μόνο με γραμμικά διαχωρίσιμα δεδομένα.

Το 1959, ο Bernard Widrow μαζί με τον φοιτητή του Ted Hoff, προχώρησαν σε βελτιώσεις στο προσαρμοστικό φίλτρο ώστε να κάνει μία κλήση κατωφέρειας για κάθε σημεία δεδομένων, οδηγώντας τους στου κανόνες Delta και Adaline. Προς αποφυγήν της χειροκίνητης ρύθμισης βαρών, επινόησαν το menistor. Το menistor είναι ένα νανοηλεκτρονικό στοιχείο κυκλωμάτων που χρησιμοποιείται στην τεχνολογία μνήμης παράλληλης επεξεργασίας. Ουσιαστικά, μια αντίσταση με μνήμη ικανή να εκτελεί λογικές πράξεις και να αποθηκεύει πληροφορίες.

Παρά τις πολλές προσπάθειες, δεν κατάφεραν ποτέ να αναπτύξουν έναν αλγόριθμο εκπαίδευσης για ένα πολυεπίπεδο νευρωνικό δίκτυο. Το πιο μακριά που έφτασαν ήταν με τον Madaline Rule I (1962), ο οποίος είχε δύο στρώματα βαρών. Το πρώτο ήταν εκπαιδεύσιμο, αλλά το δεύτερο ήταν σταθερό. (Wikipedia, 2024)

Αρκετά αργότερα, το 1982, ο Hopfield παρουσίασε μία νέα κατηγορία δικτύων. Τα δίκτυα αυτά έχουν μεγάλες δυνατότητες σε υπολογισμούς και μπορούν να χρησιμοποιηθούν κυρίως σε δύο κατηγορίες προβλημάτων, προβλήματα μνήμης συνειρμού (associative memory) και προβλήματα βελτιστοποίησης (optimization). Αν και τα δίκτυα του Hopfield χρησιμοποιήθηκαν σε ήδη λυμένα προβλήματα, και μάλιστα με όχι με τα καλύτερα αποτελέσματα, έχουν ένα χαρακτηριστικό ότι οι νευρώνες τους συνέχεια αναπροσαρμόζονται με αποτέλεσμα τα δίκτυα να εκπαιδεύονται. Αυτό ήταν πολύ χρήσιμο για την κατανόηση της φύσης των προβλημάτων και αντιμετώπισης τους κάτω από μία νέα οπτική. (Αρετός, 2020)

Το 1985 οι Ackley, Hinton και Sejnowski ανέπτυξαν τον κανόνα εκμάθησης Boltzmann. Η μηχανή Boltzmann, που αποτελεί την πρώτη επιτυχή εφαρμογή ενός πολυεπίπεδου νευρωνικού δικτύου, βασίζεται σε μία ενέργεια κατάστασης που καθορίζεται από τα βάρη σύνδεσης στο στρώμα προτύπου. Η εκμάθηση ενός συνόλου προτύπως περιλαμβάνει την ελαχιστοποίηση αυτής της ενέργειας με τα βάρη προσαρμόζονται κατά την επανάληψη των δεδομένων. Όπως και το δίκτυο του Hopfield, η μηχανή Boltzmann μπορεί να ολοκληρώσει ελλείπουσες πληροφορίες όταν παρουσιάζεται ένα μερικό σχέδιο. (Κελεμενίδης, 2018)

Το 1986 παρουσιάστηκε ο αλγόριθμος οπίσθιας διάδοσης του λάθους (backpropagation algorithm) ο οποίος βοήθησε πολύ στην ανάπτυξη των μεθόδων εκπαίδευσης των πολυεπίπεδων νευρωνικών δικτύων. Αν και η ανακάλυψη αυτή έγινε περίπου την ίδια περίοδο από ανεξάρτητους ερευνητές, φαίνεται ότι η βασική ιδέα του αλγορίθμου ήταν αυτή που ανέφερε στην διπλωματική του διατριβή ο Paul Werbos το 1974.

Τα νευρωνικά δίκτυα σήμερα

Οι παραπάνω εξελίξεις οδήγησαν στην ανάπτυξη των τεχνητών νευρωνικών δικτύων. Οι έρευνες που ακολούθησαν βοήθησαν στο να ανοιχτούν νέοι δρόμοι για νέες εφαρμογές σε διάφορους τομείς.

Μερικοί τομείς που χρησιμοποιούν τα νευρωνικά δίκτυα είναι:

- Αναγνώριση εικόνων: Τα νευρωνικά δίκτυα μπορούν να αναγνωρίσουν αντικείμενα, πρόσωπα και άλλα χαρακτηριστικά σε εικόνες.
- Μετάφραση γλωσσών: Τα νευρωνικά δίκτυα μπορούν να μεταφράσουν κείμενα από μια γλώσσα σε μια άλλη.
- Ανάλυση δεδομένων: Τα νευρωνικά δίκτυα μπορούν να αναλύσουν δεδομένα για να εντοπίσουν τάσεις, μοτίβα και άλλες πληροφορίες.

- Ιατρική διάγνωση: Τα νευρωνικά δίκτυα μπορούν να βοηθήσουν στην διάγνωση ασθενειών, με βάση ιατρικά δεδομένα.
- Οικονομική πρόβλεψη: Τα νευρωνικά δίκτυα μπορούν να προβλέψουν την κίνηση των τιμών των μετοχών και άλλων οικονομικών δεδομένων.

Τα νευρωνικά δίκτυα, όπως τα πολυεπίπεδα perceptrons (MLP) ή τα convolutional neural networks (CNN), έχουν αποδειχθεί αποτελεσματικά σε πολλές εφαρμογές, όπως της πρόβλεψης αθλητικών αποτελεσμάτων. Αυτό οφείλεται στην μεγάλη ευελιξία που προσφέρουν και στην ικανότητά τους να μοντελοποιήσουν πολύπλοκες σχέσεις μεταξύ των δεδομένων.

Δέντρα απόφασης

Τα δέντρα απόφασης είναι ένας άλλος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης και παλινδρόμησης. Τα δέντρα απόφασης δημιουργούν μία ιεραρχική δομή που μοιάζει με διάγραμμα ροής, παρόμοια με ένα δέντρο, ώστε να οδηγηθούν σε αποφάσεις. Οι διακλαδώσεις στον κορμό του δέντρου (κόμβοι) αντιπροσωπεύουν τα χαρακτηριστικά, και κάθε κλαδί αντιπροσωπεύει μία απόφαση μέσα από την οποία θα οδηγηθεί σε μία πρόβλεψη (φύλο δέντρου). Στην ουσία η διαδρομή από την ρίζα του δέντρου έως τα φύλλα του απεικονίζει μία σειρά αποφάσεων που δείχνουν τον δρόμο από το πρόβλημα (ρίζα) μέχρι μία συγκεκριμένη πρόβλεψη (φύλλα).

Για την κατασκευή ενός δέντρου απόφασης, χρησιμοποιούνται διάφορες μετρικές, ανάλογα με το αν έχουμε πρόβλημα ταξινόμησης ή παλινδρόμησης. Στην ταξινόμηση, στόχος είναι η ελαχιστοποίηση της αβεβαιότητας. Αυτό επιτυγχάνεται με τη χρήση της εντροπίας (H):

$$H(Y) = -\Sigma\,(P(y_i) * log_2(P(y_i)))$$

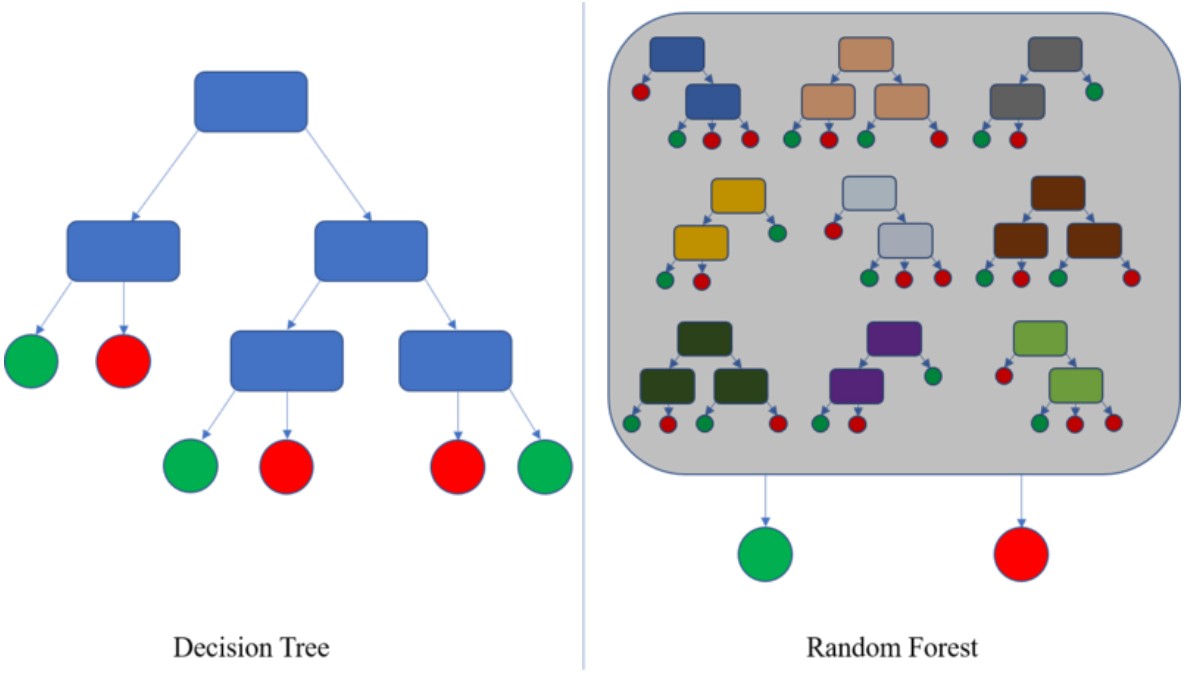
όπου $P(y_i)$ είναι η πιθανότητα η μεταβλητή Y να πάρει την τιμή y_i. Στη συνέχεια, υπολογίζουμε το κέρδος πληροφορίας (IG), το οποίο μετρά τη μείωση της εντροπίας μετά από μια διαίρεση:

$$\mathrm{IG}(Y, X_i) = \mathrm{H}(Y) - \Sigma\,(\mathrm{P}(x_{ij}) * \mathrm{H}(Y|X_i = x_{ij}))$$

όπου x_{ij} είναι η j-οστή τιμή του χαρακτηριστικού X_i. Στην παλινδρόμηση, χρησιμοποιούμε μετρικές όπως η διακύμανση ή το μέσο τετραγωνικό σφάλμα (MSE):

$$MSE = (1/N) * \Sigma\,(y_i - \hat{y}_i)^2$$

όπου N είναι ο αριθμός των δειγμάτων, y_i είναι η πραγματική τιμή και $\hat{y}_i$ είναι η προβλεπόμενη τιμή. Στα δέντρα παλινδρόμησης, ο αλγόριθμος επιλέγει το χαρακτηριστικό που ελαχιστοποιεί το MSE μετά τη διαίρεση. Οι αλγόριθμοι κατασκευής δέντρων, όπως οι ID3, C4.5 και CART, χρησιμοποιούν αυτές τις μετρικές. Ένα σημαντικό ζήτημα είναι η αποφυγή της υπερπροσαρμογής, όπου το δέντρο μαθαίνει πολύ καλά τα εκπαιδευτικά δεδομένα, αλλά αποδίδει άσχημα σε νέα. Τεχνικές όπως το κλάδεμα (αφαίρεση κλαδιών) χρησιμοποιούνται για την αντιμετώπιση αυτού του προβλήματος.



Εικόνα 3 – Decision Tree & Random Forest *(Jeremybeauchamp, 2020)*

Random Forests

Τα Τυχαία Δάση (Random Forest) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται και αυτός για εργασίες ταξινόμησης και παλινδρόμησης. Στην συγκεκριμένη εργασία θα χρησιμοποιηθεί ο αλγόριθμος αυτός για την πρόβλεψη των αποτελεσμάτων αγώνων βόλεϊ.

Τα Τυχαία Δάση δημιουργούν ένα σύνολο N δέντρων αποφάσεων, καθένα από τα οποία είναι εκπαιδευμένο σε ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης, που επιλέγεται με επαναδειγματοληψία (bootstrap sampling). Κατά την κατασκευή κάθε δέντρου, σε κάθε διακλάδωση (split) εξετάζεται μόνο ένας τυχαίος υποσύνολο

$$m$$

χαρακτηριστικών από το πλήρες σύνολο

$$p$$

, όπου

$$\mathrm{m} \ll \mathrm{p}$$

, για να καθοριστεί η καλύτερη διαίρεση.

Η τελική πρόβλεψη του αλγορίθμου είναι αποτέλεσμα της συνδυασμένης "ψήφου" των δέντρων, όπου

- Στην ταξινόμηση, η τελική κατηγορία προκύπτει από την πλειοψηφία ψήφων:
$$\hat{y} = \text{mode}\{h^1(x), h^2(x), \dots, h_n(x)\}$$
όπου $h_i(x)$ είναι η πρόβλεψη του i-οστού δέντρου για ένα δεδομένο x.

- Στην παλινδρόμηση, η τελική πρόβλεψη προκύπτει από τον μέσο όρο των προβλέψεων:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

Τα Random Forest υπερέχουν σε σχέση με τα Δέντρα Απόφασης διότι η μέθοδος τυχαίας δειγματοληψίας και επιλογής χαρακτηριστικών μειώνει την πιθανότητα υπερπροσαρμογής (overfitting) και οδηγεί σε μεγαλύτερη ακρίβεια. Επιπλέον, η διαδικασία bootstrap ενισχύει τη γενίκευση του μοντέλου.

Deep Learning (Βαθιά μάθηση)

Η εξέλιξη της μηχανικής μάθησης (machine learning) οδήγησε στην βαθιά μάθηση (deep learning). Αυτό είχε σαν αποτέλεσμα την ανάπτυξη του τομέα αυτού. Υπάρχουν όμως κάποιες διαφορές ανάμεσα στα δύο. Οι 3 πιο σημαντικές διαφορές είναι:

- Η βαθιά μάθηση μπορεί να μαθαίνει από τα δικά της λάθη (αυτομάθηση) σε αντίθεση με την μηχανική μάθηση η οποία χρειάζεται ανθρώπινη συμμετοχή ώστε να μπορούν να γίνουν οι διορθώσεις.
- Στην μηχανική μάθηση, από τη στιγμή που ένας αλγόριθμος μπει σε λειτουργία, υπάρχει επίβλεψη από τον χρήστη, ώστε να παρατηρεί τις διάφορες μεταβλητές. Αντίθετα στην βαθιά μάθηση, από τη στιγμή που ένας αλγόριθμος ξεκινήσει να λειτουργεί, έχει μεγαλύτερη ανεξαρτησία αναφορικά με την ανάλυση δεδομένων (ανεξαρτησία ανάλυσης δεδομένων).
- Τέλος η βαθιά μάθηση απαιτεί πολύ περισσότερα δεδομένα από τη μηχανική μάθηση. Αυτό έχει ως αποτέλεσμα να απαιτείται πολύ περισσότερη υπολογιστική ισχύ ώστε να υπάρχει η δυνατότητα να γίνει η ανάλυση των δεδομένων.

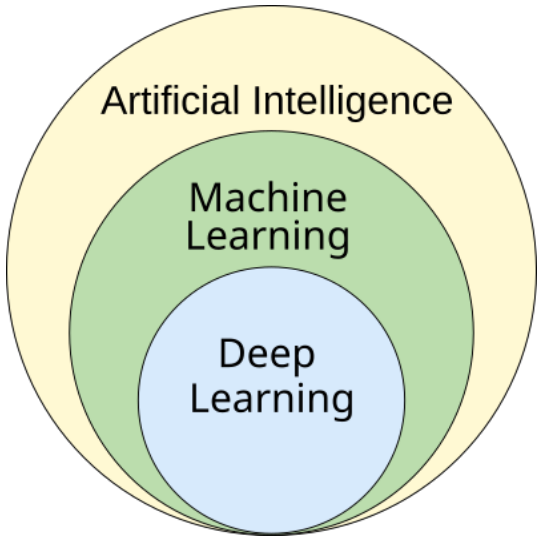
Γνωστά νευρωνικά δίκτυα της βαθιάς μάθησης είναι το RNN (Αναδρομικά Νευρωνικά Δίκτυα) και το LSTM (Long Short-Term Memory).

Το RNN είναι ιδανικό για την επεξεργασία δεδομένων που έχουν χρονική σειρά καθώς έχουν την δυνατότητα να «θυμούνται» προηγούμενα δεδομένα και αυτό τους δίνει την δυνατότητα να τα χρησιμοποιούν σε προβλέψεις μελλοντικών αποτελεσμάτων (ποσότητα δεδομένων).

Τα LSTM είναι η εξέλιξη των RNN, έχουν τα ίδια χαρακτηριστικά αλλά την δυνατότητα επεξεργασίας δεδομένων με μεγαλύτερη χρονική διάρκεια και εξαρτήσεις, δηλαδή μπορούν να υπολογίσουν περισσότερους παράγοντες οι οποίοι κάνουν πιο πολύπλοκη την πρόβλεψη του αποτελέσματος.

Και τα δύο έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά στην πρόβλεψη αθλητικών αποτελεσμάτων.

Για την αποπεράτωση της συγκεκριμένης εργασίας θα χρησιμοποιηθούν αλγόριθμοι μηχανικής μάθησης.



Εικόνα 4 – Artificial Intelligence *(Lollixzc, 2022)*

Σύνολο Δεδομένων

Συλλογή δεδομένων

Η ακριβής πρόβλεψη του αποτελέσματος αγώνων βόλεϊ είναι μια σύνθετη διαδικασία που δεν μπορεί να επιτευχθεί απλώς με τη δημιουργία ενός μαγικού μοντέλου που θα εξασφαλίσει αυτή τη δυνατότητα. Αντίθετα, απαιτείται η ανάπτυξη μιας ισχυρής και καλά δομημένης βάσης δεδομένων, η οποία θα είναι όσο το δυνατόν πιο πλήρης και ακριβής. Η αξιοποίηση αλγορίθμων μηχανικής μάθησης σε συνδυασμό με αυτή τη βάση δεδομένων θα επιτρέψει την εκπαίδευση των μοντέλων, ώστε να μπορούν να παράγουν τα καλύτερα δυνατά αποτελέσματα και να αναγνωρίζουν υποκείμενα πρότυπα.

Η συλλογή αυτών των δεδομένων είναι ζωτικής σημασίας και θα πρέπει να γίνεται από έγκυρες και αξιόπιστες πηγές, προκειμένου να διασφαλιστεί η πληρότητα και η ακρίβεια των πληροφοριών. Για τη συγκεκριμένη εργασία, η κύρια πηγή δεδομένων είναι η επίσημη ιστοσελίδα της Volley League στην Ελλάδα (www.volleyleague.gr). Αυτή η πλατφόρμα παρέχει όλα τα αποτελέσματα αγώνων από το 2011 μέχρι σήμερα, προσφέροντας μια εκτενή και αξιόπιστη βάση δεδομένων που είναι απαραίτητη για μια εμπεριστατωμένη ανάλυση.

Επιπλέον, προκειμένου να καλυφθούν τυχόν ελλείψεις και να επιβεβαιωθούν τα δεδομένα, πραγματοποιήθηκε και συλλογή αποτελεσμάτων για την ίδια χρονική περίοδο από ανθρώπους που κρατούσαν στατιστικά στις ομάδες ανδρών αυτής της κατηγορίας. Αυτή η διαδικασία όχι μόνο ενισχύει την αξιοπιστία των στοιχείων, αλλά και διασφαλίζει ότι η βάση δεδομένων που θα χρησιμοποιηθεί στην παρούσα εργασία είναι όσο το δυνατόν πιο πλήρης και ακριβής. Έτσι, παρέχεται μια σταθερή βάση για την ανάπτυξη των μοντέλων πρόβλεψης.

Σύμφωνα με τα παραπάνω, μέσα από την εργασία αυτή θα μπορέσει να δημιουργηθεί μια βάση δεδομένων με περίπου 1400 αποτελέσματα αγώνων. Αυτή η πλούσια συλλογή δεδομένων θα επιτρέψει τη δημιουργία ενός ισχυρού μοντέλου, το οποίο θα έχει τη δυνατότητα να προβλέπει με μεγαλύτερη ακρίβεια το αποτέλεσμα των αγώνων.

Η χρήση αξιόπιστων δεδομένων, σε συνδυασμό με την εφαρμογή προηγμένων αλγορίθμων μηχανικής μάθησης, θα δώσει τη δυνατότητα για την ανάπτυξη ενός συστήματος πρόβλεψης που θα μπορεί να χρησιμοποιηθεί για την ανάλυση και την πρόβλεψη των αποτελεσμάτων αγώνων βόλεϊ με υψηλό βαθμό ακρίβειας. Η προοπτική αυτή όχι μόνο θα ενισχύσει τις γνώσεις μας σχετικά με το άθλημα, αλλά θα προσφέρει και πολύτιμα εργαλεία για προπονητές, αθλητές και αναλυτές, διευρύνοντας έτσι τις δυνατότητες της αθλητικής ανάλυσης.

Επεξεργασία δεδομένων

Τα δεδομένα αυτά χρειάστηκε να συλλεχθούν από διάφορα σημεία, προκειμένου να διασφαλιστεί η πληρότητα και η ακριβής αναπαράσταση των στοιχείων. Η μεγάλη διάρκεια των αποτελεσμάτων, που εκτείνεται σε βάθος χρόνου, είχε ως αποτέλεσμα οι ομάδες είτε να αλλάζουν ονόματα είτε να μεταβιβάζονται σε διαφορετικές κατηγορίες. Αυτές οι αλλαγές καθιστούν την ανάλυση πιο περίπλοκη, καθώς πολλές ομάδες, παρά την αρχική τους συμμετοχή στην αγωνιστική περίοδο, ενδέχεται να έχουν διακόψει τη δραστηριότητά τους.

Προκειμένου να αντιμετωπιστούν όλα τα παραπάνω, ήταν αναγκαίος ένας λεπτομερής έλεγχος όλων των αγωνιστικών στοιχείων για κάθε χρονική περίοδο. Ο στόχος ήταν να εντοπιστούν και να αφαιρεθούν οι περιττές πληροφορίες που θα μπορούσαν να προκαλέσουν σύγχυση ή να επηρεάσουν αρνητικά την εκπαίδευση των μοντέλων αλλά και τα τελικά αποτελέσματα της ανάλυσης. Όπως είναι αναμενόμενο, δόθηκε ιδιαίτερη προσοχή στην επεξεργασία των δεδομένων καθώς οι αλλαγές στη δομή των πρωταθλημάτων και των ομάδων έχουν επηρεάσει σε μεγάλο βαθμό τα δεδομένα που χρησιμοποιούνται.

Η σωστή επεξεργασία και η κατηγοριοποίηση των πληροφοριών δεν είναι μόνο κρίσιμη για την ποιότητα των αναλύσεων αλλά και για την κατανόηση των τάσεων που έχουν προκύψει στον αθλητικό τομέα όλα αυτά τα χρόνια.

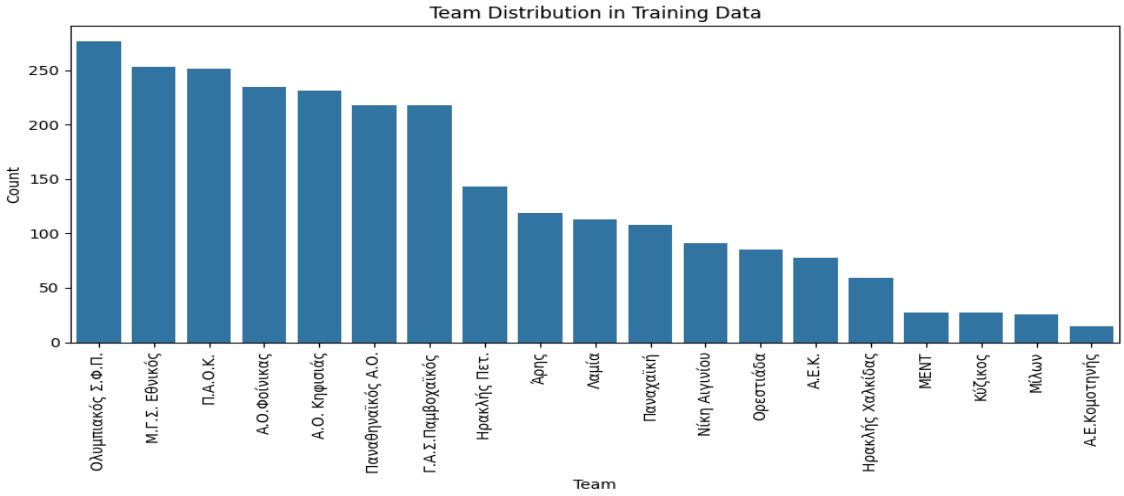
Πειραματικά Αποτελέσματα

Προκειμένου να καταλήξω στο τελικό μοντέλο που θα γίνει χρήση, προχώρησα στη δοκιμή διαφόρων διαφορετικών αλγορίθμων μηχανικής μάθησης. Για να μπορέσει να γίνει πιο σωστή αξιολόγηση των αποτελεσμάτων, χρησιμοποιήθηκαν 4 ομάδες που έχουν παρουσία σε όλες τις αγωνιστικές περιόδους από το 2011 μέχρι και σήμερα. Οι ομάδες αυτές είναι οι:

- Ολυμπιακός Σ.Φ.Π.
- Α.Ο. Φοίνικας
- Π.Α.Ο.Κ.
- Παναθηναϊκός Α.Ο..

KNeighborsClassifier

Στα πρώτα στάδια της εργασίας οι δοκιμές έγιναν με τον KNeighborsClassifier με την χρήση ενός προγράμματος Python.



Εικόνα 5 – Κατανομή των ομάδων στα train data (KNeighborsClassifier)

Ας δούμε όμως λίγο την λειτουργία του προγράμματος. Αρχικά φορτώνει το σύνολο των δεδομένων που είναι για εκπαίδευση και δοκιμή. Γίνεται ο καθαρισμός των δεδομένων και οι κωδικοποίηση των μεταβλητών, χρησιμοποιώντας το LabelEncoder. Τα χαρακτηριστικά που θα χρησιμοποιηθούν είναι τα:

- Hometeam (ομάδα που παίζει εντός έδρας)
- Hometeam_score (σκορ της ομάδας εντός έδρας)
- Awayteam (ομάδα που παίζει εκτός έδρας)

- Awayteam_score (σκορ της ομάδας εκτός έδρας

Στη συνέχεια, εκπαιδεύεται το μοντέλο με βάση τα δεδομένα που υπάρχουν στον αρχείο εκπαίδευσης (train_data.csv). Στο αρχείο αυτό υπάρχει μία πληθώρα δεδομένων ώστε ο αλγόριθμος να εκπαιδευτεί και στη συνέχεια, κάνοντας χρήση το αρχείο δοκιμής (test_data.csv), ελέγχεται η ακρίβεια των προβλέψεων του αλγορίθμου. Στα δύο csv αρχεία υπάρχουν το σύνολο των αποτελεσμάτων των αγώνων.

Από τα αποτελέσματα των δοκιμών που έγιναν η συνολική ακρίβεια του μοντέλου ήταν στο 0.71, που σημαίνει ότι το μοντέλο προβλέπει σωστά το αποτέλεσμα του 71% των αγώνων. Αυτό σημαίνει ότι το μοντέλο είναι αρκετά ακριβές στην πρόβλεψη των αποτελεσμάτων, με συνολική ακρίβεια 71%.

Αποτελέσματα στα test data για το σύνολο των ομάδων

Accuracy:	0.7136150234741784			
Ομάδα	Precision	Recall	F1-Score	Support
Α.Ε.Κ.	0.00	0.00	0.00	1
Α.Ο. Κηφισιάς	0.75	0.88	0.81	24
Α.Ο. Φοίνικας	**0.90**	**0.82**	**0.86**	**34**
Α.Σ. Ελπίς Αμπελοκήπων	0.00	0.00	0.00	2
Γ.Α.Σ. Παμβοχαϊκός	0.50	1.00	0.67	8
Ηρακλής Πετ.	1.00	0.40	0.57	10
Μ.Γ.Σ. Εθνικός	0.11	1.00	0.20	1
Μίλων	0.00	0.00	0.00	9
Νίκη Αιγινίου	0.00	0.00	0.00	0
Ο.Φ.Η.	0.00	0.00	0.00	7
Ολυμπιακός Σ.Φ.Π.	**0.85**	**1.00**	**0.92**	**39**
Ορεστιάδα	0.00	0.00	0.00	0
Π.Α.Ο.Κ.	**0.61**	**0.97**	**0.75**	**35**
Παναθηναϊκός Α.Ο.	**0.85**	**0.45**	**0.59**	**38**
Φίλιππος Βέροιας	0.00	0.00	0.00	5

Πίνακας 1 - Αποτελέσματα στα test data (KNeighborsClassifier)

Ας δούμε αναλυτικά τα αποτελέσματα για τις ομάδες:

- Ολυμπιακός Σ.Φ.Π.
- Α.Ο. Φοίνικας
- Π.Α.Ο.Κ.
- Παναθηναϊκός Α.Ο.

Ολυμπιακός Σ.Φ.Π.:

- **Precision**: 0.85, δηλαδή 85% των προβλέψεων για τον Ολυμπιακό ήταν σωστές.
- **Recall**: 1.00, δηλαδή 100% των αγώνων του Ολυμπιακού προβλέφθηκαν σωστά.
- **F1-Score**: 0.92, δηλαδή ο Ολυμπιακός είχε υψηλή συνολική ακρίβεια στις προβλέψεις.
- **Support**: 39, δηλαδή ο Ολυμπιακός έπαιξε 39 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Παναθηναϊκός Α.Ο.:

- **Precision**: 0.85, δηλαδή 85% των προβλέψεων για τον Παναθηναϊκό ήταν σωστές.
- **Recall**: 0.45, δηλαδή 45% των αγώνων του Παναθηναϊκού προβλέφθηκαν σωστά.
- **F1-Score**: 0.59, δηλαδή ο Παναθηναϊκός είχε μέτρια συνολική ακρίβεια στις προβλέψεις.
- **Support**: 38, δηλαδή ο Παναθηναϊκός έπαιξε 38 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Α.Ο. Φοίνικας:

- **Precision**: 0.90, δηλαδή 90% των προβλέψεων για τον Φοίνικα ήταν σωστές.
- **Recall**: 0.82, δηλαδή 82% των αγώνων του Φοίνικα προβλέφθηκαν σωστά.
- **F1-Score**: 0.86, δηλαδή ο Φοίνικας είχε υψηλή συνολική ακρίβεια στις προβλέψεις.
- **Support**: 34, δηλαδή ο Φοίνικας έπαιξε 34 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Π.Α.Ο.Κ.:

- **Precision**: 0.61, δηλαδή 61% των προβλέψεων για τον ΠΑΟΚ ήταν σωστές.
- **Recall**: 0.97, δηλαδή 97% των αγώνων του ΠΑΟΚ προβλέφθηκαν σωστά.
- **F1-Score**: 0.75, δηλαδή ο ΠΑΟΚ είχε μέτρια συνολική ακρίβεια στις προβλέψεις.
- **Support**: 35, δηλαδή ο ΠΑΟΚ έπαιξε 35 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

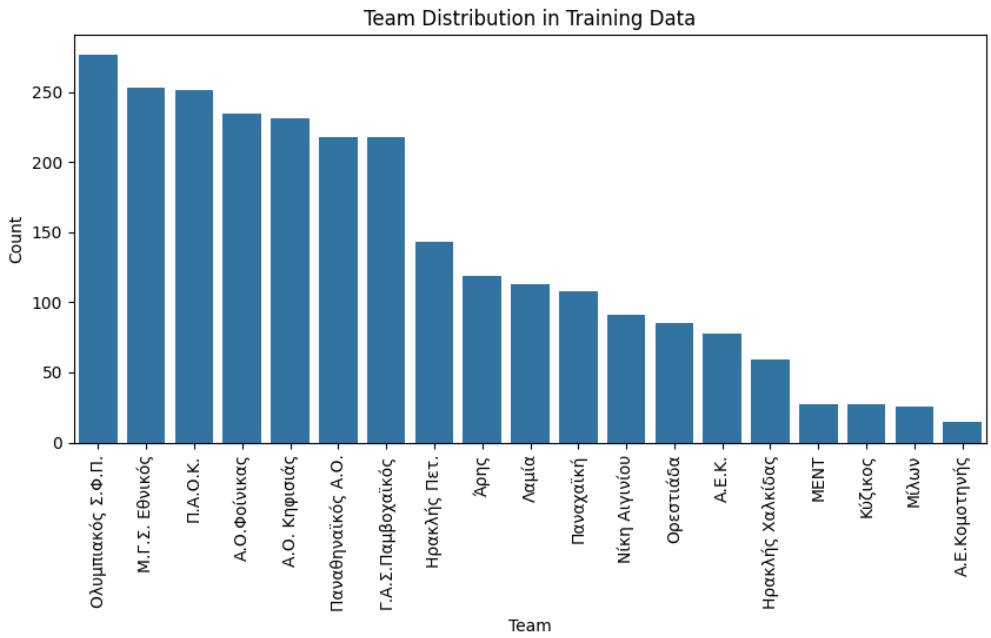
Συνολικά:

- Ο Ολυμπιακός και ο Φοίνικας είχαν υψηλή ακρίβεια στις προβλέψεις.
- Ο Παναθηναϊκός είχε μέτρια ακρίβεια.
- Ο ΠΑΟΚ είχε μέτρια ακρίβεια, με υψηλό recall και χαμηλό precision.

Αυτό υποδηλώνει ότι η ανάλυση είχε καλύτερη απόδοση στην πρόβλεψη των αποτελεσμάτων για τον Ολυμπιακό και τον Φοίνικα. Η πρόβλεψη για τον Παναθηναϊκό και τον ΠΑΟΚ ήταν πιο δύσκολη.

GaussianNB

Στο δεύτερο στάδιο της εργασίας οι δοκιμές έγιναν με έναν άλλο αλγόριθμο, τον GaussianNB.



Εικόνα 6 – Κατανομή των ομάδων στα train data (GaussianNB)

Όπως και στον KNeighborsClassifier, χρησιμοποιήθηκαν τα ίδια δεδομένα εκπαίδευσης (train_data.csv) και δοκιμής (test_data.csv). Έγινε η προεπεξεργασία των δεδομένων, όπως καθαρισμό και κωδικοποίηση των μεταβλητών. Σ' αυτήν την περίπτωση, χρησιμοποιούμε το LabelEncoder για να μετατρέψουμε τα ονόματα των ομάδων σε αριθμητικές τιμές.

Ο GaussianNB χρησιμοποίησε τα ίδια χαρακτηριστικά για να κάνει προβλέψεις:

- Hometeam (ομάδα που παίζει εντός έδρας)
- Hometeam_score (σκορ της ομάδας εντός έδρας)
- Awayteam (ομάδα που παίζει εκτός έδρας)
- Awayteam_score (σκορ της ομάδας εκτός έδρας)

Ο αλγόριθμος GaussianNB εκπαιδεύεται με τα δεδομένα εκπαίδευσης (train_data.csv).
Η διαδικασία εκπαίδευσης περιλαμβάνει τον υπολογισμό της κατανομής πιθανοτήτων για κάθε χαρακτηριστικό, υποθέτοντας ότι τα δεδομένα ακολουθούν μια κανονική κατανομή (Gaussian distribution).

Αφού εκπαιδευτεί το μοντέλο, μπορούμε να κάνουμε προβλέψεις για τα αποτελέσματα αγώνων χρησιμοποιώντας τα δεδομένα δοκιμής (test_data.csv). Ο αλγόριθμος GaussianNB χρησιμοποιεί την κατανομή πιθανοτήτων που υπολογίστηκε κατά την εκπαίδευση για να προβλέψει την πιθανότητα νίκης για κάθε ομάδα.

Από τα αποτελέσματα των δοκιμών που έγιναν η συνολική ακρίβεια του μοντέλου ήταν στο 0.27, που σημαίνει ότι το μοντέλο προβλέπει σωστά το αποτέλεσμα του 27% των αγώνων. Αυτό σημαίνει ότι το μοντέλο δεν είναι πολύ ακριβές στην πρόβλεψη των αποτελεσμάτων, με συνολική ακρίβεια 27%.

Πίνακας 2 - Αποτελέσματα από τα test data (GaussianNB)

Accuracy	0.27230046948356806			
Ομάδα	Precision	Recall	F1-Score	Support
Α.Ε.Κ.	0.00	0.00	0.00	1
Α.Ο. Κηφισιάς	0.33	0.58	0.42	24
Α.Ο.Φοίνικας	**0.25**	**0.12**	**0.16**	**34**
Α.Σ.Ελπίς Αμπελοκήπων	0.00	0.00	0.00	2
Γ.Α.Σ.Παμβοχαϊκός	0.00	0.00	0.00	8
Ηρακλής Πετ.	0.00	0.00	0.00	10
Λαμία	0.00	0.00	0.00	0
Μ.Γ.Σ. Εθνικός	0.00	0.00	0.00	1
Μίλων	0.00	0.00	0.00	9
Ο.Φ.Η.	0.00	0.00	0.00	7
Ολυμπιακός Σ.Φ.Π.	**0.28**	**0.90**	**0.43**	**39**
Π.Α.Ο.Κ.	**0.12**	**0.06**	**0.08**	**35**
Παναθηναϊκός Α.Ο.	**0.50**	**0.08**	**0.14**	**38**
Φίλιππος Βέροιας	0.00	0.00	0.00	5

Ας δούμε αναλυτικά τα αποτελέσματα για τις ίδιες ομάδες:

- Ολυμπιακός Σ.Φ.Π.
- Α.Ο. Φοίνικας
- Π.Α.Ο.Κ.
- Παναθηναϊκός Α.Ο.

Ολυμπιακός Σ.Φ.Π.:

- **Precision**: 0.28, δηλαδή 28% των προβλέψεων για τον Ολυμπιακό ήταν σωστές.
- **Recall**: 0.90, δηλαδή 90% των αγώνων του Ολυμπιακού προβλέφθηκαν σωστά.
- **F1-Score**: 0.43, που δείχνει μέτρια συνολική ακρίβεια.
- **Support**: 39, δηλαδή ο Ολυμπιακός έπαιξε 39 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Παναθηναϊκός Α.Ο.:

- **Precision**: 0.50, δηλαδή 50% των προβλέψεων για τον Παναθηναϊκό ήταν σωστές.
- **Recall**: 0.08, δηλαδή 8% των αγώνων του Παναθηναϊκού προβλέφθηκαν σωστά.
- **F1-Score**: 0.14, που δείχνει πολύ χαμηλή συνολική ακρίβεια.
- **Support**: 38, δηλαδή ο Παναθηναϊκός έπαιξε 38 αγώνες. στο dataset στο οποίο έγινε η πρόβλεψη

Α.Ο. Φοίνικας:

- **Precision**: 0.25, δηλαδή 25% των προβλέψεων για τον Φοίνικα ήταν σωστές.
- **Recall**: 0.12, δηλαδή 12% των αγώνων του Φοίνικα προβλέφθηκαν σωστά.
- **F1-Score**: 0.16, που δείχνει πολύ χαμηλή συνολική ακρίβεια.
- **Support**: 34, δηλαδή ο Φοίνικας έπαιξε 34 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

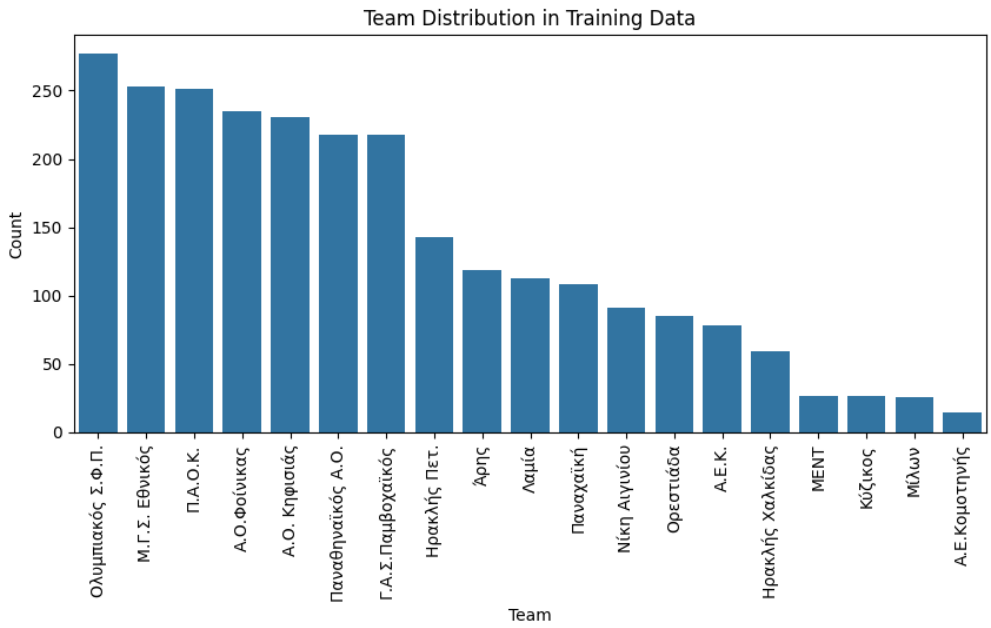
Π.Α.Ο.Κ.:

- **Precision**: 0.12, δηλαδή 12% των προβλέψεων για τον ΠΑΟΚ ήταν σωστές.
- **Recall**: 0.06, δηλαδή 6% των αγώνων του ΠΑΟΚ προβλέφθηκαν σωστά.
- **F1-Score**: 0.08, που δείχνει πολύ χαμηλή συνολική ακρίβεια.
- **Support**: 35, δηλαδή ο ΠΑΟΚ έπαιξε 35 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Συνολικά, φαίνεται ότι η πρόβλεψη για τις ομάδες του Ολυμπιακού, του ΠΑΟΚ, του Παναθηναϊκού και του Φοίνικα δεν ήταν πολύ ακριβής από το συγκεκριμένο μοντέλο.

Random Forest

Στην συνέχεια οι δοκιμές έγιναν με την Random Forest.



Εικόνα 7 – Κατανομή των ομάδων στα train data (Random Forest)

Παρόμοια με τους άλλους δύο αλγορίθμους, ο Random Forest φορτώνει τα δεδομένα και δημιουργεί μία λίστα με όλες τις ομάδες που εμφανίζονται στα δεδομένα και κάνει την κωδικοποίηση των μεταβλητών.

Στην συνέχεια δημιουργεί το σύνολο της εκπαίδευσης που βασίζεται στα ίδια τέσσερα χαρακτηριστικά:

- Hometeam: Ο αριθμός που αντιστοιχεί στην ομάδα που αγωνίζεται εντός έδρας.
- Hometeam_score: Η βαθμολογία της ομάδας που αγωνίζεται εντός έδρας.
- Awayteam: Ο αριθμός που αντιστοιχεί στην ομάδα που αγωνίζεται εκτός έδρας.
- Awayteam_score: Η βαθμολογία της ομάδας που αγωνίζεται εκτός έδρας.

Ο αλγόριθμος Random Forest μαθαίνει από τα δεδομένα εκπαίδευσης πώς συνδέονται τα τέσσερα χαρακτηριστικά με το αποτέλεσμα του αγώνα. Για παράδειγμα, μπορεί να μάθει ότι όταν η ομάδα που αγωνίζεται εντός έδρας έχει πολύ υψηλότερη βαθμολογία από την αντίπαλη ομάδα, τότε είναι πιο πιθανό να κερδίσει.

Κατά την πρόγνωση, ο αλγόριθμος λαμβάνει τα τέσσερα χαρακτηριστικά για τον αγώνα που θέλουμε να προβλέψει και εφαρμόζει τις γνώσεις που έχει αποκτήσει από τα δεδομένα εκπαίδευσης. Στη συνέχεια, ψηφίζει για το αποτέλεσμα του αγώνα, χρησιμοποιώντας όλα τα δέντρα αποφάσεων που έχει δημιουργήσει.

Από τα αποτελέσματα των δοκιμών που έγιναν η συνολική ακρίβεια του μοντέλου ήταν στο 0.87, που σημαίνει ότι το μοντέλο προβλέπει σωστά το αποτέλεσμα του 87% των αγώνων. Αυτό σημαίνει ότι το μοντέλο δεν είναι πολύ ακριβές στην πρόβλεψη των αποτελεσμάτων, με συνολική ακρίβεια 87%.

Πίνακας 3 - Αποτελέσματα από τα test data (Random Forest)

Accuracy	0.8732394366197183			
Ομάδα	Precision	Recall	F1-Score	Support
Α.Ε.Κ.	0.00	0.00	0.00	1
Α.Ο. Κηφισιάς	0.96	1.00	0.98	24
Α.Ο.Φοίνικας	**0.94**	**1.00**	**0.97**	**34**
Α.Σ.Ελπίς Αμπελοκήπων	0.00	0.00	0.00	2
Γ.Α.Σ.Παμβοχαϊκός	0.73	1.00	0.84	8
Ηρακλής Πετ.	1.00	0.70	0.82	10
Μ.Γ.Σ. Εθνικός	0.25	1.00	0.40	1
Μίλων	0.00	0.00	0.00	9
Νίκη Αιγινίου	0.00	0.00	0.00	0
Ο.Φ.Η.	0.00	0.00	0.00	7
Ολυμπιακός Σ.Φ.Π.	**1.00**	**1.00**	**1.00**	**39**
Π.Α.Ο.Κ.	**1.00**	**1.00**	**1.00**	**35**
Παναθηναϊκός Α.Ο.	**1.00**	**1.00**	**1.00**	**38**
Παναχαϊκή	0.00	0.00	0.00	0
Φίλιππος Βέροιας	0.00	0.00	0.00	5

Ας δούμε αναλυτικά τα αποτελέσματα για τις ίδιες ομάδες:

- Ολυμπιακός Σ.Φ.Π.
- Α.Ο.Φοίνικας
- Π.Α.Ο.Κ.
- Παναθηναϊκός Α.Ο.

Ολυμπιακός Σ.Φ.Π.:

- **Precision**: 1.0, δηλαδή 100% των προβλέψεων για τον Ολυμπιακό ήταν σωστές.
- **Recall**: 1.0, δηλαδή 100% των αγώνων του Ολυμπιακού προβλέφθηκαν σωστά.
- **F1-Score**: 0.97, που σημαίνει ότι το μοντέλο έχει μάθει να ταξινομεί σωστά όλες τις περιπτώσεις που σχετίζονται με τον Ολυμπιακό.
- **Support**: 39, δηλαδή ο Ολυμπιακός έπαιξε 39 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Παναθηναϊκός Α.Ο.:

- **Precision**: 1.0, δηλαδή 100% των προβλέψεων για τον Παναθηναϊκό ήταν σωστές.
- **Recall**: 1.0, δηλαδή 100% των αγώνων του Παναθηναϊκού προβλέφθηκαν σωστά.
- **F1-Score**: 1.0, που σημαίνει ότι το μοντέλο έχει μάθει να ταξινομεί σωστά όλες τις περιπτώσεις που σχετίζονται με τον Παναθηναϊκό.
- **Support**: 38, δηλαδή ο Παναθηναϊκός έπαιξε 38 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Α.Ο. Φοίνικας:

- **Precision**: 0.94, δηλαδή 94% των προβλέψεων για τον Φοίνικα ήταν σωστές.
- **Recall**: 1.0, δηλαδή 100% των αγώνων του Φοίνικα προβλέφθηκαν σωστά.
- **F1-Score**: 0.97 που σημαίνει ότι το μοντέλο έχει μάθει να ταξινομεί σωστά τις περισσότερες περιπτώσεις που σχετίζονται με τον Φοίνικα.
- **Support**: 34, δηλαδή ο Φοίνικας έπαιξε 34 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Π.Α.Ο.Κ.:

- **Precision**: 1.0, δηλαδή 100% των προβλέψεων για τον Π.Α.Ο.Κ. ήταν σωστές.
- **Recall**: 1.0, δηλαδή 100% των αγώνων του Π.Α.Ο.Κ. προβλέφθηκαν σωστά.
- **F1-Score**: 1.0, που σημαίνει ότι το μοντέλο έχει μάθει να ταξινομεί σωστά όλες τις περιπτώσεις που σχετίζονται με τον Π.Α.Ο.Κ..
- **Support**: 35, δηλαδή ο Π.Α.Ο.Κ. έπαιξε 35 αγώνες στο dataset στο οποίο έγινε η πρόβλεψη.

Συνολικά οι 4 ομάδες που χρησιμοποιήθηκαν στις δοκιμές φαίνονται να έχουν καλή αναγνωρισιμότητα από το συγκεκριμένο μοντέλο. Η μεγάλη υποστήριξη για αυτές τις ομάδες υποδηλώνει ότι το μοντέλο έχει εκπαιδευτεί σε ένα πλούσιο σύνολο δεδομένων, αυξάνοντας την αξιοπιστία των αποτελεσμάτων που δίνει.

Random Forest με K-Fold Cross Validation

Στην συνέχεια χρησιμοποιήθηκε ο αλγόριθμος Random Forest σε συνδυασμό με το K-Fold Cross Validation.

Από τα αποτελέσματα των δοκιμών που έγιναν βλέπουμε ότι οι τιμές της ακρίβειας (Accuracy) κυμαίνονται ανάμεσα στο 0.8893 και το 0.9146. Αυτό δείχνει ότι το μοντέλο είναι αρκετά σταθερό και δεν εξαρτάται από την συγκεκριμένη κατανομή των δεδομένων εκπαίδευσης και ελέγχου του κάθε fold.

Η μέση ακρίβεια του μοντέλου ήταν στο 0.9031 (Mean accuracy, δηλαδή τη μέση ακρίβεια παρέχει μια γενική εικόνα για την απόδοση του μοντέλου σε διαφορετικά υποσύνολα δεδομένων), είναι αρκετά υψηλή, κάτι που σημαίνει ότι το μοντέλο μπορεί να προβλέψει σωστά το αποτέλεσμα για πάνω από το 90% των αγώνων.

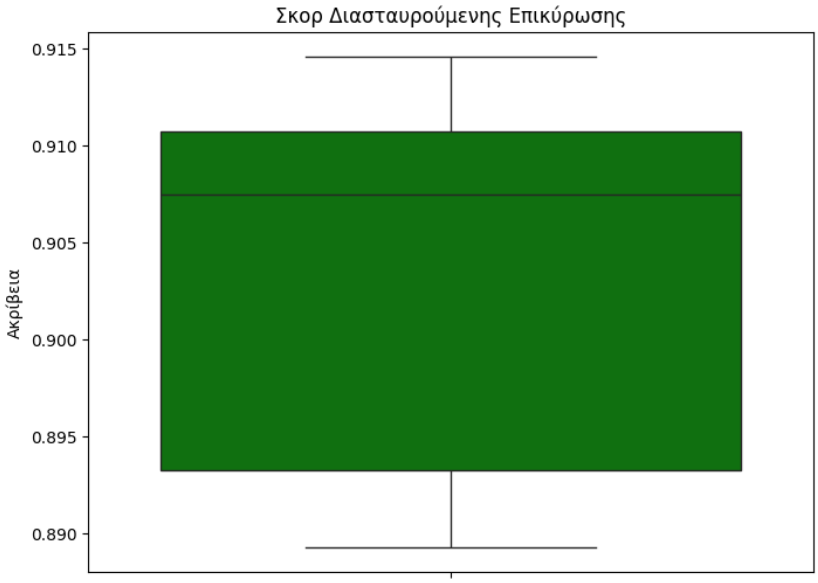
Η καλύτερη επίδοση είναι στο Fold 2, με ακρίβεια 0.9146 ενώ η χαμηλότερη είναι στο Fold 4 με ακρίβεια 0.8893. Η διαφορά των δύο δεν είναι μεγάλη και δείχνει ότι ίσως να υπάρχουν μικρές διαφορές στις επιδόσεις ανάλογα με την κατανομή των δεδομένων σε κάθε Fold.

Πιο αναλυτικά οι τιμές ακρίβειας που εμφανίστηκαν από το cross-validation ήταν:

Cross-validation accuracy scores	
Fold	Accuracy
Fold 1	0.8932384341637011
Fold 2	0.9145907473309609
Fold 3	0.9074733096085409
Fold 4	0.8892857142857142
Fold 5	0.9107142857142857
Mean	0.9030604982206405

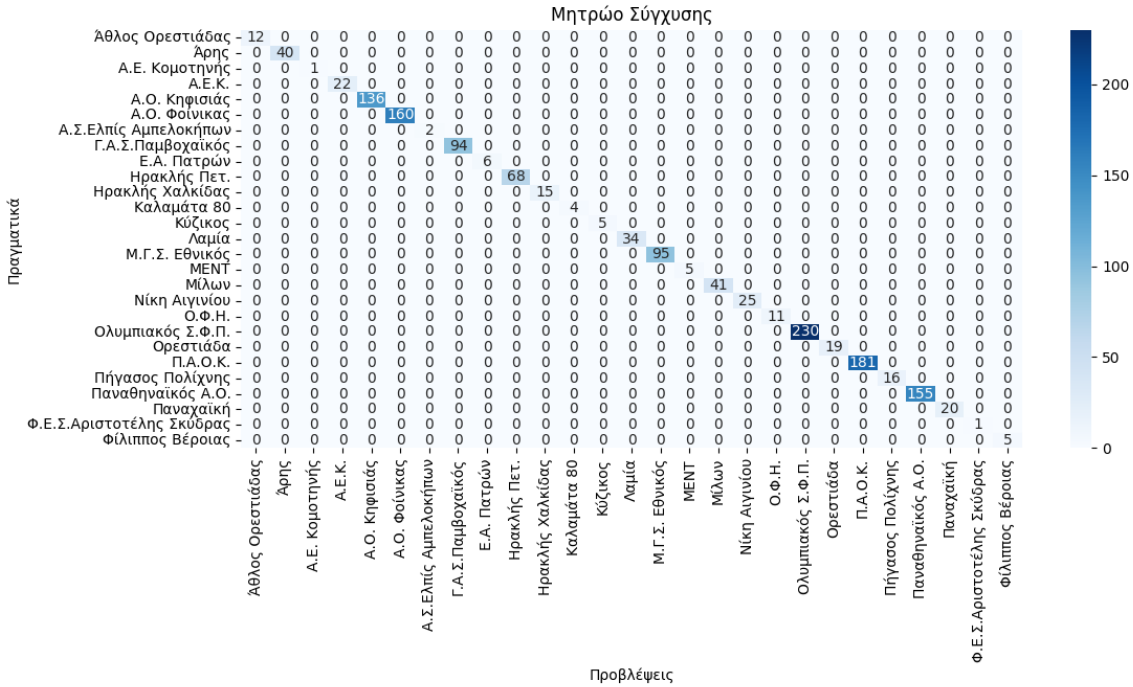
Εικόνα 8 – Cross-validation accuracy scores

Στο παρακάτω διάγραμμα μπορούμε να δούμε τα αποτελέσματα της ακρίβειας για κάθε ένα από τα K folds. Δηλαδή, δείχνουν πόσο καλά τα πήγε το μοντέλο σε κάθε γύρο επαλήθευσης.



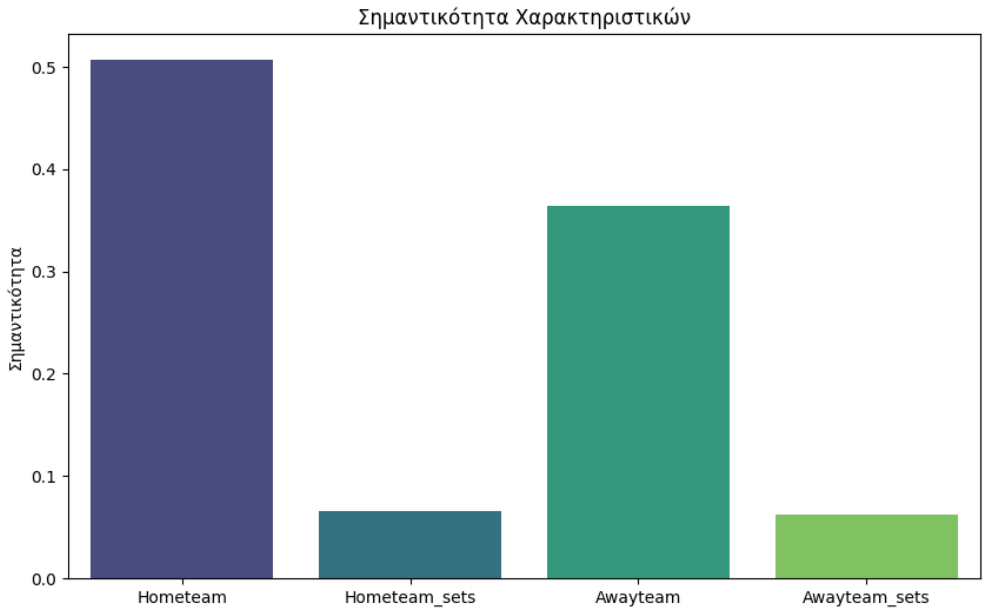
Εικόνα 9 – Cross-Validation Accuracy Scores

Από το παρακάτω διάγραμμα μπορούμε να δούμε τα πραγματικά αποτελέσματα σε σχέση με τις προγνώσεις. Οι τιμές που υπάρχουν στην διαγώνιο είναι οι σωστές προβλέψεις για την κάθε ομάδα. Οι τιμές που ίσως να υπήρχαν εκτός της διαγώνιου θα ήταν λανθασμένες προβλέψεις. Παρατηρούμε ότι, για τις ομάδες που χρησιμοποιήσαμε ως δείγμα σε όλα τα μοντέλα, δεν υπάρχουν λανθασμένες προβλέψεις (τιμές εκτός διαγώνιου). Τέλος παρατηρούμε ότι, στις ίδιες ομάδες, έχουμε και τις πιο υψηλές τιμές, κάτι το υποδεικνύει ότι το μοντέλο που χρησιμοποιήσαμε έχει καλή ακρίβεια.



Εικόνα 10 – Confusion Matrix

Στο επόμενο διάγραμμα μπορούμε να δούμε τα χαρακτηριστικά του μοντέλου μας και την σημαντικότητά τους.

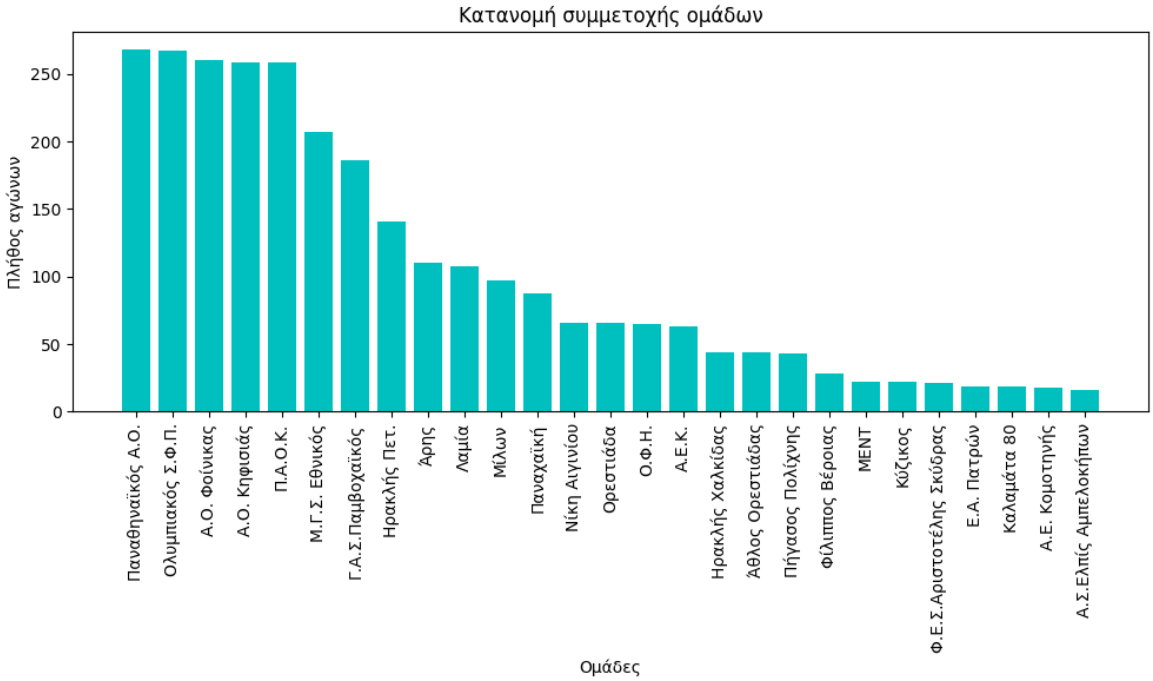


Εικόνα 11 – Σημαντικότητα Χαρακτηριστικών

Ανάλυση των χαρακτηριστικών:

- **Hometeam (Ομάδα εντός έδρας)**: Με την υψηλότερη σημαντικότητα, κοντά στο 0.5. Αυτό σημαίνει ότι η πληροφορία σχετικά με την εντός έδρας ομάδα ήταν η πιο καθοριστική για τις προβλέψεις του μοντέλου. Οι ομάδες που παίζουν εντός έδρας έχουν συχνά πλεονεκτήματα λόγω παραγόντων, όπως η εξοικείωση με το γήπεδο, τους τοπικούς φιλάθλους και την ευκολία μετακίνησης, καθώς δεν κουράζονται από την μετακίνηση του ταξιδιού.
- **Awayteam (Ομάδα εκτός έδρας)**: Αυτό είναι το δεύτερο πιο σημαντικό χαρακτηριστικό, με σημασία περίπου 0.3. Η πληροφορία σχετικά με την εκτός έδρας ομάδα είναι επίσης πολύ σημαντική για το μοντέλο. Οι φιλοξενούμενες ομάδες μπορεί να επηρεαστούν από το άγνωστο περιβάλλον του γηπέδου καθώς και από την τυχόν μετακίνηση (ή ίσως και ταξίδι) προς το γήπεδο, γεγονός που επηρεάζει την απόδοσή τους.
- **Hometeam_sets (Σετ της εντός έδρας ομάδας)**: Με πολύ μικρότερη σημαντικότητα, περίπου στο 0.1. Τα σετ που κέρδισε η εντός έδρας ομάδα δεν ήταν τόσο καθοριστικά στις αποφάσεις του μοντέλου. Αυτό δείχνει ότι η ταυτότητα της ομάδας είναι πιο σημαντική από την απόδοση της σε ένα συγκεκριμένο σετ.
- **Awayteam_sets (Σετ της εκτός έδρας ομάδας)**: Με τη χαμηλότερη σημαντικότητα, λίγο κάτω από 0.1. Τα σετ που κέρδισε η εκτός έδρας ομάδα είχαν ακόμα μικρότερη επιρροή στις προβλέψεις. Αυτό δείχνει ότι το μοντέλο δίνει μικρότερη σημασία στην απόδοση της φιλοξενούμενης ομάδας στα σετ, σε σχέση με την ταυτότητά της.

Η ανάλυση αυτή επιβεβαιώνει ότι για το μοντέλο, η ταυτότητα των ομάδων (εντός και εκτός έδρας) είναι οι πιο κρίσιμοι παράγοντες για την πρόβλεψη των αποτελεσμάτων των αγώνων βόλεϊ, με την απόδοση στα σετ να παίζει δευτερεύοντα ρόλο.



Εικόνα 12 – Κατανομή των ομάδων στα train data (Random Forest & KFold)

Από την εικόνα 12 βλέπουμε ότι ο Ολυμπιακός Σ.Φ.Π. και ο Παναθηναϊκός Α.Ο. είναι οι ομάδες με τις περισσότερες συμμετοχές, με πάνω από 250 αγώνες η κάθε μία. Επίσης ο Α.Ο. Φοίνικας και Π.Α.Ο.Κ . έχουν συμμετάσχει σε έναν αξιοσέβαστο αριθμό αγώνων, κυμαινόμενοι από περίπου 150 έως 200 αγώνες. Αυτός ήταν και ο λόγος που χρησιμοποιήθηκαν οι 4 αυτές ομάδες για την σύγκριση των αποτελεσμάτων.

Μαζί με τα παραπάνω, να συμπληρώσω ότι και οι 4 ομάδες είχαν

- **Precision**: 1.0, δηλαδή 100% των προβλέψεων για των ομάδων ήταν σωστές.
- **Recall**: 1.0, δηλαδή 100% των αγώνων των ομάδων προβλέφθηκαν σωστά.
- **F1-Score**: 1.0, που σημαίνει ότι το μοντέλο έχει μάθει να ταξινομεί σωστά όλες τις περιπτώσεις που σχετίζονται με τις συγκεκριμένες ομάδες.

Συνοψίζοντας τα παραπάνω καταλαβαίνουμε ότι το συγκεκριμένο μοντέλο έχει χαμηλή διακύμανση των αποτελεσμάτων, κάτι το οποίο σημαίνει ότι έχει ισχυρή σταθερότητα και δεν επηρεάζεται υπερβολικά από τυχαίες διαφορές στα δεδομένα εκπαίδευσης.

Με μέση ακρίβεια 0.9031, το μοντέλο παρουσιάζει εξαιρετική απόδοση. Επίσης η μικρή διακύμανση των αποτελεσμάτων ακρίβειας υποδεικνύει ότι στα περισσότερα σενάρια το μοντέλο μπορεί να κάνει μία σωστή πρόβλεψη. Σενάρια που θα μπορούσαν να «δυσκολέψουν» το συγκεκριμένο μοντέλο θα ήταν, για παράδειγμα, όταν εμπλέκαμε αγώνες με πιο ισοδύναμες ομάδες.

Συζήτηση αποτελεσμάτων

GaussianNB: Ο αλγόριθμος αυτός είχε τη χαμηλότερη απόδοση από όλους τους αλγορίθμους με ακρίβεια που φτάνει στο 27%. Οι πολύ χαμηλές τιμές των macro και weighted averages δείχνουν ότι το συγκεκριμένο μοντέλο δεν είναι κατάλληλο για την ανάλυση των εν λόγω δεδομένων που αφορούν την πρόβλεψη αποτελεσμάτων αγώνων βόλεϊ, πιθανώς λόγω της υποθέσεως της κανονικότητας των δεδομένων από τον GaussianNB..
Συνοπτικά τα αποτελέσματά του:

- **Accuracy**: 0.27
- **Macro avg**: Precision 0.11, Recall 0.12, F1-Score 0.09
- **Weighted avg**: Precision 0.24, Recall 0.27, F1-Score 0.19

KNeighborsClassifier: Ο αλγόριθμος αυτός είχε μεγαλύτερη ακρίβεια, ποσοστό 71%, σε σύγκριση με τον GaussianNB, όμως δεν ήταν τόσο υψηλό σε σύγκριση με το Random Forest. Από τα αποτελέσματά του βλέπουμε ότι το μοντέλο αυτό δεν έχει τη δυνατότητα να προβλέψει σωστά όλες τις κατηγορίες. Αυτό έχει ως αποτέλεσμα να μην έχει μεγάλη ακρίβεια.
Συνοπτικά τα αποτελέσματά του:

- **Accuracy**: 0.71
- **Macro avg**: Precision 0.37, Recall 0.43, F1-Score 0.36
- **Weighted avg**: Precision 0.70, Recall 0.71, F1-Score 0.68

Random Forest (Χωρίς K-Fold Cross-Validation): Ο αλγόριθμος Random Forest χωρίς τη χρήση K-Fold Cross-Validation έχει υψηλότερη ακρίβεια, με ποσοστό 87%, σε σχέση με τα δύο άλλα μοντέλα. Η ακρίβεια που έχει το κάνει πιο ισχυρό σε συγκεκριμένες κατηγορίες αλλά όχι συνολικά σε όλες. Σε σχέση με το επόμενο μοντέλο, δηλαδή του Random Forest με τη χρήση K-Fold Cross-Validation, είναι ελάχιστα πιο χαμηλά στην ακρίβεια αλλά λόγω του ότι δεν κάνει cross-validation, θα μπορούσαμε να ισχυριστούμε ότι τα αποτελέσματά του είναι πολύ αισιόδοξα. Επίσης διαπιστώνουμε ότι δεν συμπεριφέρεται το ίδιο σε όλες τις ομάδες και δίνει μεγαλύτερη βαρύτητα στις ομάδες με τους περισσότερους αγώνες.
Συνοπτικά τα αποτελέσματά του:

- **Accuracy**: 0.87
- **Macro avg**: Precision 0.50, Recall 0.55, F1-Score 0.51
- **Weighted avg**: Precision 0.87, Recall 0.88, F1-Score 0.87

Random Forest με K-Fold Cross-Validation (5-fold): Από τα αποτελέσματα διαπιστώθηκε ότι για το συγκεκριμένο dataset, τα καλύτερα αποτελέσματα τα έφερε ο αλγόριθμος Random Forest με χρήση K-Fold Cross-Validation (5-fold). Το μοντέλο αυτό εμφανίζει σταθερή μέση ακρίβεια, ελάχιστα

πάνω από το 90%. Σαν ποσοστό είναι πιο υψηλό από την ακρίβεια του προηγούμενου και τα scores που υπάρχουν σε κάθε fold, τα οποία ήταν αρκετά κοντά μεταξύ τους, δείχνουν ότι το συγκεκριμένο μοντέλο είναι σταθερό και δεν εξαρτάται τόσο από τα δεδομένα. Εξάλλου η χρήση του K-Fold Cross-Validation βοηθά στην αποφυγή overfitting, κάτι που το κάνει να παρέχει μία πιο αξιόπιστη εκτίμηση της απόδοσης.

- **Cross-validation accuracy scores:**
  - Fold 1: 0.8932
  - Fold 2: 0.9146
  - Fold 3: 0.9075
  - Fold 4: 0.8893
  - Fold 5: 0.9107
- **Mean accuracy:** 0.9031

Συμπεράσματα και Μελλοντική έρευνα

Συνοψίζοντας τα πειραματικά αποτελέσματα βλέπουμε τα παρακάτω.

Για μελλοντική έρευνα, προτείνεται η χρήση μεγαλύτερου dataset ώστε να μπορούν τα μοντέλα να εκπαιδευτούν ακόμα καλύτερα. Ένα μεγαλύτερο dataset θα βοηθήσει στη βελτίωση της γενίκευσης του μοντέλου και στην αποφυγή overfitting, καθώς περισσότερα δεδομένα παρέχουν ένα πιο πλούσιο σύνολο παραδειγμάτων για εκπαίδευση.

Με την προσθήκη περισσότερων χαρακτηριστικών, όπως στατιστικά των παικτών, παράγοντες όπως η ψυχολογία της ομάδας και άλλες εξωγενείς παράμετροι (π.χ., αλλαγές στην προπονητική ομάδα), θα μπορούσαμε να παραχθεί ένα πιο ολοκληρωμένο και ακριβές μοντέλο πρόβλεψης. Τα πρόσθετα χαρακτηριστικά μπορούν να προσφέρουν νέες διαστάσεις και πληροφορίες, αυξάνοντας την ακρίβεια των προβλέψεων.

Η ισορροπία των δεδομένων είναι κρίσιμη για την ακριβή πρόβλεψη των αποτελεσμάτων. Ένα ισορροπημένο dataset μπορεί να διασφαλίσει ότι το μοντέλο δεν θα ευνοεί συγκεκριμένες τάξεις (π.χ., νίκες συγκεκριμένων ομάδων), παρέχοντας πιο δίκαιες και ακριβείς προβλέψεις. Τεχνικές όπως η oversampling ή η undersampling μπορούν να χρησιμοποιηθούν για την επίτευξη αυτής της ισορροπίας.

Η χρήση πιο σύνθετων μοντέλων όπως τα νευρωνικά δίκτυα, τα οποία μπορούν να μάθουν πιο περίπλοκα μοτίβα από τα δεδομένα, θα μπορούσε να βελτιώσει περαιτέρω την ακρίβεια των προβλέψεων. Τα βαθιά νευρωνικά δίκτυα (deep learning) και τα αναδρομικά νευρωνικά δίκτυα (recurrent neural networks) είναι ειδικά κατάλληλα για προβλήματα όπου οι χρονικές ακολουθίες και τα μοτίβα έχουν σημασία.

Μια πιο λεπτομερής ανάλυση της σημασίας των διαφόρων χαρακτηριστικών μπορεί να οδηγήσει σε καλύτερη κατανόηση των παραγόντων που επηρεάζουν τα αποτελέσματα των αγώνων. Μέσω αυτής της ανάλυσης, μπορούμε να εντοπίσουμε ποια χαρακτηριστικά έχουν τη μεγαλύτερη επιρροή στις προβλέψεις και να επικεντρωθούμε στην ενίσχυση αυτών των παραγόντων.

Η ενσωμάτωση εξωτερικών δεδομένων, όπως ιστορικά δεδομένα για τη δυναμική των ομάδων, τις αποδόσεις των παικτών σε προηγούμενες σεζόν, και τα δεδομένα από παρόμοιες διοργανώσεις, μπορεί να προσφέρει πλούσια πληροφορίες που θα βελτιώσουν την ακρίβεια των προβλέψεων.

Εκτός από το 5-fold cross-validation, μπορούν να δοκιμαστούν και άλλες τεχνικές διασταυρούμενης επικύρωσης, όπως το 10-fold ή το stratified k-fold, για να διασφαλιστεί η σταθερότητα και η ακρίβεια των αποτελεσμάτων σε διαφορετικές κατανομές των δεδομένων.

Αυτές οι κατευθύνσεις μπορούν να συμβάλουν στη βελτίωση της απόδοσης των μοντέλων και στην ανάπτυξη πιο ακριβών και αξιόπιστων προβλέψεων για τα αποτελέσματα των αγώνων βόλεϊ.

Βιβλιογραφία

Κελεμενίδης, Ζ. (2018, Μάϊος). *Μεταπτυχιακή Διατριβή: Από το Perceptron στην Βαθειά Μάθηση. Εφαρμογή περιπτώσεων ταξινόμησης με 3 γενεές Νευρωνικών Δικτύων.* Ανάκτηση από Αποθετήριο Δημοκρίτειου Πανεπιστημίου: https://repo.lib.duth.gr/jspui/bitstream/123456789/13779/1/KelemenidisZ_2018.pdf

ΣΚΟΥΦΑ, Α. (2008, Νοέμβριος). *Μεταπτυχιακή Διπλωματική Εργασία.* Ανάκτηση από ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ: http://ikee.lib.auth.gr/record/112962/files/GRI-2009-2726.pdf

(2024). Ανάκτηση από https://eclass.upatras.gr/modules/document/file.php/CEID1041/Διδακτικές%20Σημείωσεις/Chapter2.pdf.

Alhama, R. G. (2017, Μάιος). *Wikimedia.* Ανάκτηση από https://commons.wikimedia.org/wiki/File:Artificial_Neuron.svg

Funcs, C. (2024, Μάϊος). *Wikimedia Commons.* Ανάκτηση από https://commons.wikimedia.org/wiki/File:Artificial_neuron_structure.svg

HEBB, D. 0. (1949). *The Organizationof Behavior.* Νέα Υόρκη.

HERBINET, C. (2018, Ιούνιος). *Predicting Football Results Using Machine Learning Techniques.* Ανάκτηση από Google Scolar: https://down.documentine.com/2e9ffc9e076525b0ab753a9cc74db806.pdf

Jeremybeauchamp. (2020, Δεκέμβριος). *Decision Tree vs. Random Forest.* Ανάκτηση από Wikimedia: https://commons.wikimedia.org/wiki/File:Decision_Tree_vs._Random_Forest.png

Lollixzc. (2022, Αύγουστος). *Machine learning as a subset of AI.* Ανάκτηση από Wikimedia: https://commons.wikimedia.org/wiki/File:AI_hierarchy.svg

Mitchell. (2012, Οκτώβριος). *Wikimedia.* Ανάκτηση από https://commons.wikimedia.org/wiki/File:Rosenblattperceptron.png

Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with Machine Learning. *Procedia Computer Science, 204*, 463-470.

W. S. McCulloch and W. Pitts. (1943, 12). A logical calculus of ideas immanent in nervous activity. *Bullettin of Mathematical Biophysics, 5*, 115–133.

Wikipedia. (2024). Ανάκτηση από https://en.wikipedia.org/wiki/Bernard_Widrow

Wikipedia. (2024). *wikipedia*. Ανάκτηση από https://el.wikipedia.org/wiki/Perceptron

Νικολουλόπουλος, Φ. (2023, Μάρτιος). *Πρόβλεψη Αποτελεσμάτων Αγώνων Τένις με Χρήση Μηχανικής Μάθησης και Ανάπτυξη Στοιχηματικής Στρατηγικής.* Ανάκτηση από Ψηφιακό Αποθετήριο Εθνικό Μετσόβιο Πολυτεχνίο: https://dspace.lib.ntua.gr/xmlui/bitstream/handle/123456789/57493/Τελικό-Αρχείο-Τόμος-11.pdf?sequence=1

Παλινδρόμηση (στατιστική). (2023, Οκτώβρης). Ανάκτηση από Wikipedia: https://el.wikipedia.org/wiki/Παλινδρόμηση_(στατιστική)#Γραμμικότητα

Πανεπιστήμιο Πατρών. (2024). *Πλατφόρμα Τηλεκπαίδευσης*. Ανάκτηση από Πλατφόρμα Τηλεκπαίδευσης: https://eclass.upatras.gr/modules/document/file.php/CEID1041/Διδακτικές%20Σημείωσεις/Chapter2.pdf

Αρετός, Ε. (2020, Ιούλιος). *Διπλωματική Εργασία: Τεχνητά νευρωνικά δίκτυα και εφαρμογές αυτών στην εκπαίδευση.* Ανάκτηση από Ελληνικό Ανοικτό Πανεπιστήμιο: https://apothesis.eap.gr/home

Αρμενιάκου, Σ. (2022, Οκτώβριος). *Πρόβλεψη Αποτελεσμάτων Αγώνων Μπάσκετ με Χρήση Τεχνικών Μηχανικής Μάθησης.* Ανάκτηση από Ψηφιακό Αποθετήριο Εθνικό Μετσόβιο Πολυτεχνίο: https://dspace.lib.ntua.gr/xmlui/bitstream/handle/123456789/56722/Thesis.pdf?sequence=1

Γαρδέλης, Σ. (2021, Ιούνιος). *Αναλυτική δεδομένων αγώνων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων και εξαγωγή γνώσης.* Ανάκτηση από ΔΙΩΝΗ: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/13576/Gardelis_ME1922.pdf?sequence=1&isAllowed=y

Μιχαλάκης, Ζ. (2024, Μάιος). *Ανάλυση δεδομένων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων με επιλογή χαρακτηριστικών.* Ανάκτηση από ΔΙΩΝΗ: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16461/Michalakis_ME2116.pdf?sequence=1&isAllowed=y

ΜΑΖΑΙ, Ε. (2024, Σεπτέμβριος). *Ιδρυματικό Αποθετήριο Πανεπιστημίου Δυτικής Αττικής .* Ανάκτηση από ΠΡΟΒΛΕΨΗ ΕΚΒΑΣΗΣ ΑΓΩΝΩΝ ΠΟΔΟΣΦΑΙΡΟΥ ΜΕ ΑΛΓΟΡΙΘΜΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ: https://polynoe.lib.uniwa.gr/xmlui/bitstream/handle/11400/7709/Mazai_18390080.pdf?sequence=1&isAllowed=y

Αιγινίτη, Χ. (2024, Σεπτέμβριος). *ΕΦΑΡΜΟΓΕΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΟ ΜΗΧΑΝΟΚΙΝΗΤΟ ΑΘΛΗΤΙΣΜΟ.* Ανάκτηση από ΔΙΩΝΗ: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16849/Aiginiti_MES22005.pdf?sequence=1&isAllowed=y

Παράρτημα 1 – Ο τελικός κώδικας της εργασίας

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, precision_score,
recall_score, f1_score
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import KFold, cross_val_score
from tabulate import tabulate
import os
import warnings
import sys

# Ορισμός καθαρισμού οθόνης
def clear_screen():
    os.system('cls' if os.name == 'nt' else 'clear')

# Καθαρισμός οθόνης πριν την έναρξη
clear_screen()

# Φόρτωση των δεδομένων εκπαίδευσης
train_data = pd.read_csv('volleyball_match_data.csv', delimiter=';')

# Συνδυασμός των στηλών Hometeam και Awayteam για προσαρμογή στο
LabelEncoder
all_teams = pd.concat([train_data['Hometeam'], train_data['Awayteam']],
axis=0)

# Κωδικοποίηση των κατηγορικών μεταβλητών (ομάδων και νικητών)
le_team = LabelEncoder()
le_team.fit(all_teams)
train_data['Hometeam'] = le_team.transform(train_data['Hometeam'])
train_data['Awayteam'] = le_team.transform(train_data['Awayteam'])

le_winner = LabelEncoder()
train_data['Winner'] = le_winner.fit_transform(train_data['Winner'])

# Προετοιμασία χαρακτηριστικών και στόχου
X = train_data[['Hometeam', 'Hometeam_sets', 'Awayteam',
'Awayteam_sets']]
y = train_data['Winner']

# Ορισμός του μοντέλου
model = RandomForestClassifier()

# Ρύθμιση της διασταυρούμενης επικύρωσης K-Fold
kf = KFold(n_splits=5, shuffle=True, random_state=42)

# Εκτέλεση διασταυρούμενης επικύρωσης
scores = cross_val_score(model, X, y, cv=kf, scoring='accuracy')

# Δημιουργία πίνακα με τα αποτελέσματα της διασταυρούμενης επικύρωσης
description = 'Cross-validation accuracy scores'
```

```
table = [[description, ""]]
table.append(["Fold", "Accuracy"])
for i, score in enumerate(scores):
    table.append([f"Fold {i+1}", score])
table.append(["Mean", scores.mean()])

# Εκτύπωση των αποτελεσμάτων σε πίνακα
print('\n\n' + tabulate(table, headers="firstrow", tablefmt="grid") +
'\n\n')

# Εκπαίδευση του μοντέλου σε όλο το σύνολο δεδομένων μετά τη
διασταυρούμενη επικύρωση
model.fit(X, y)

# Δημιουργία οπτικοποιήσεων

# Κατανομή συμμετοχής ομάδων (Ραβδόγραμμα)
team_counts = pd.concat([train_data['Hometeam'], train_data['Awayteam']],
axis=0).value_counts()
team_indices = team_counts.index.tolist()
team_names = le_team.inverse_transform(team_indices)

plt.figure(figsize=(10, 6))
plt.bar(team_names, team_counts, color='c')
plt.title("Κατανομή συμμετοχής ομάδων")
plt.xlabel("Ομάδες")
plt.ylabel("Πλήθος αγώνων")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()

# Σκορ Διασταυρούμενης Επικύρωσης (Δεξιά διάγραμμα)
plt.figure(figsize=(8, 6))
sns.boxplot(data=scores, color='g')
plt.title('Σκορ Διασταυρούμενης Επικύρωσης')
plt.ylabel('Ακρίβεια')
plt.show()

# Μητρώο Σύγχυσης (Θερμόγραμμα)
y_pred = model.predict(X)
cm = confusion_matrix(y, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
xticklabels=le_winner.classes_, yticklabels=le_winner.classes_)
plt.title('Μητρώο Σύγχυσης')
plt.xlabel('Προβλέψεις')
plt.ylabel('Πραγματικά')
plt.show()

# Σημαντικότητα Χαρακτηριστικών (Ραβδόγραμμα)
importances = model.feature_importances_
features = ['Hometeam', 'Hometeam_sets', 'Awayteam', 'Awayteam_sets']
plt.figure(figsize=(10, 6))
sns.barplot(x=features, y=importances, hue=features, palette='viridis',
dodge=False, legend=False)
plt.title("Σημαντικότητα Χαρακτηριστικών")
plt.ylabel("Σημαντικότητα")
```

```
plt.show()

# Υπολογισμός και εκτύπωση μετρικών για κάθε ομάδα
unique_teams = le_winner.classes_

precision_per_team = {}
recall_per_team = {}
f1_per_team = {}

for team in unique_teams:
    try:
        team_mask = (y == le_winner.transform([team])[0])
        team_y_true = team_mask.astype(int)
        team_y_pred = (y_pred ==
le_winner.transform([team])[0]).astype(int)

        precision_per_team[team] = precision_score(team_y_true,
team_y_pred)
        recall_per_team[team] = recall_score(team_y_true, team_y_pred)
        f1_per_team[team] = f1_score(team_y_true, team_y_pred)

    except Exception as e:
        print(f"Δεν ήταν δυνατός ο υπολογισμός μετρικών για την ομάδα
{team}: {e}")

# Δημιουργία πίνακα για τις μετρήσεις
table = []
for team in unique_teams:
    if team in precision_per_team:
        table.append([team, precision_per_team[team],
recall_per_team[team], f1_per_team[team]])

# Εκτύπωση του πίνακα
print(tabulate(table, headers=["Ομάδα", "Precision", "Recall", "F1-
Score"], tablefmt="grid"))

# Συνάρτηση για να προβλέψει τον νικητή του αγώνα
def predict_winner(hometeam, awayteam):
    # Προσθέστε οποιεσδήποτε νέες ομάδες συναντώνται στον κωδικοποιητή
    if hometeam not in le_team.classes_:
        le_team.classes_ = np.append(le_team.classes_, hometeam)
    if awayteam not in le_team.classes_:
        le_team.classes_ = np.append(le_team.classes_, awayteam)

    # Κωδικοποίηση των ονομάτων των ομάδων
    hometeam_encoded = le_team.transform([hometeam])[0]
    awayteam_encoded = le_team.transform([awayteam])[0]

    # Δημιουργία ενός DataFrame με 0-0 σετ για πρόβλεψη
    match_features = pd.DataFrame([[hometeam_encoded, 0,
awayteam_encoded, 0]],
                                  columns=['Hometeam', 'Hometeam_sets',
'Awayteam', 'Awayteam_sets'])

    # Προβλέψτε τον νικητή
    predicted_winner_encoded = model.predict(match_features)[0]
    predicted_winner =
```

```
le_winner.inverse_transform([predicted_winner_encoded])[0]

    return predicted_winner

# Εκτύπωση όλων των ομάδων (για να βοηθήσει στην αντιγραφή-επικόλληση)
message = '   Μοναδικές ομάδες - για χρήση στην αντιγραφή-επικόλληση:'
print('\n\n\n' + message)
print('-' * len(message))
for t in all_teams.sort_values().unique():
    print(t)

# Προτροπή χρήστη για εισαγωγή
def main_interaction():
    hometeam_input = input("Εισάγετε την γηπεδούχο ομάδα: ")
    awayteam_input = input("Εισάγετε την φιλοξενούμενη ομάδα: ")

    # Πρόβλεψη και εμφάνηση του νικητή
    predicted_winner = predict_winner(hometeam_input, awayteam_input)
    print(f'Ο προβλεπόμενος νικητής είναι: {predicted_winner}')

if __name__ == "__main__":
    main_interaction()
```

Παράρτημα 2 – Στιγμιότυπα από την εφαρμογή

Εικόνα 13 – Στιγμιότυπο από την εφαρμογή, πριν την πρόβλεψη του αγώνα.

Εικόνα 14 – Στιγμιότυπο από την εφαρμογή, αποτέλεσμα πρόβλεψης αγώνα.