ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη Εφαρμογής Ταξινόμησης Ποδοσφαιριστών

Του φοιτητή
Γεώργιου Τσακλίδη
Αρ. Μητρώου: 174953

Επιβλέπων
Χαράλαμπος Μπράτσας
Επίκουρος Καθηγητής

Ημερομηνία 7/9/2025

Ανάπτυξη Εφαρμογής Ταξινόμησης Ποδοσφαιριστών

24282

Γεώργιος Τσακλίδης

Χαράλαμπος Μπράτσας

27-10-2024

08-07-2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως πτυχιακή εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η παρούσα πτυχιακή εργασία εκπονήθηκε με σκοπό τη μελέτη και ανάπτυξη μίας εφαρμογής που συνδυάζει την ανάλυση δεδομένων με το ποδόσφαιρο, αξιοποιώντας τεχνικές μηχανικής μάθησης και ταξινόμησης. Το ενδιαφέρον για το συγκεκριμένο θέμα προέκυψε από τη διαρκώς αυξανόμενη συμβολή της τεχνολογίας στον αθλητισμό και την ανάγκη δημιουργίας εργαλείων που διευκολύνουν την κατανόηση και αξιοποίηση μεγάλων ποσοτήτων δεδομένων. Η εργασία επιχειρεί να γεφυρώσει το χάσμα μεταξύ θεωρητικών εννοιών ταξινόμησης δεδομένων και της πρακτικής εφαρμογής τους σε πραγματικά παραδείγματα από τον χώρο του ποδοσφαίρου. Στόχος είναι η υλοποίηση ενός συστήματος που επιτρέπει την εισαγωγή και επεξεργασία χαρακτηριστικών ποδοσφαιριστών, παρέχοντας στο τέλος προβλέψεις σχετικά με τη θέση που θα μπορούσαν να καταλάβουν στον αγωνιστικό χώρο. Η επιλογή του θέματος βασίστηκε τόσο σε προσωπικό ενδιαφέρον για την επιστήμη δεδομένων όσο και στην αγάπη για το ποδόσφαιρο, ενώ φιλοδοξία του συγγραφέα είναι το παρόν έργο να αποτελέσει ένα πρώτο βήμα προς πιο σύνθετες εφαρμογές στον τομέα της αθλητικής ανάλυσης.

Περίληψη

Στη σύγχρονη εποχή, η τεχνολογία και τα δεδομένα κατέχουν ολοένα και πιο κεντρικό ρόλο στον χώρο του ποδοσφαίρου. Από την καταγραφή κινήσεων των παικτών στις προπονήσεις έως τη μελέτη αποφάσεων κατά τη διάρκεια ενός αγώνα, η ανάλυση δεδομένων έχει μετατραπεί σε κρίσιμο εργαλείο αξιολόγησης και στρατηγικού σχεδιασμού.

Η παρούσα εργασία εστιάζει στην ανάπτυξη εφαρμογής ταξινόμησης ποδοσφαιριστών, αξιοποιώντας σύγχρονες μεθόδους μηχανικής μάθησης. Αρχικά συγκροτήθηκε βάση δεδομένων ενεργών παικτών, η οποία χρησιμοποιήθηκε ως σύνολο εκπαίδευσης, και αντίστοιχη βάση ανενεργών παικτών, που αποτέλεσε το σύνολο δοκιμών. Μέσω μιας φιλικής διεπαφής, ο χρήστης δύναται είτε να επιλέξει έναν ανενεργό ποδοσφαιριστή είτε να εισάγει χειροκίνητα τα χαρακτηριστικά ενός νέου παίκτη και να εφαρμόσει διαφορετικούς αλγορίθμους ταξινόμησης (k-NN, CNN, Random Forest).

Η εφαρμογή αποδίδει καταρχάς την καταλληλότερη περιοχή δράσης (άμυνα, κέντρο, επίθεση) και στη συνέχεια προτείνει τη βέλτιστη θέση για τον παίκτη, με βάση τα δεδομένα των ενεργών ποδοσφαιριστών. Η μεθοδολογία εμπλουτίζεται με γραφήματα και οπτικοποιήσεις που ενισχύουν την κατανόηση των αποτελεσμάτων. Τα συμπεράσματα της εργασίας καταδεικνύουν τις δυνατότητες και τα πλεονεκτήματα της αξιοποίησης αλγορίθμων μηχανικής μάθησης στον χώρο του αθλητισμού, θέτοντας παράλληλα προτάσεις για μελλοντική εξέλιξη και βελτίωση του συστήματος.

Application for Football Player Classification

George Tsaklidis

Abstract

In today's football industry, technology and data play an increasingly pivotal role. From tracking players' movements during training sessions to analyzing decision-making in competitive matches, data analysis has become an essential tool for performance evaluation and strategic planning. This thesis focuses on the development of a football player classification application using modern machine learning techniques. A database of active players was initially compiled and used as the training set, while a corresponding database of inactive players served as the testing set. Through a user-friendly interface, the user can either select an inactive football player or manually input the attributes of a new player and then apply various classification algorithms (k-NN, CNN, Random Forest).

The application first identifies the most suitable field area for the player (defense, midfield, attack) and subsequently recommends the optimal playing position, based on the data of active players. The methodology is further enhanced through graphs and visualizations, facilitating interpretation of the results. The conclusions highlight the potential and advantages of integrating machine learning algorithms into the field of sports, while also outlining suggestions for future improvements and system extensions.

Περιεχόμενα

Κατάλογος Σχημάτων

Συντομογραφίες

ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία
ML	Machine Learning
AI	Artificial Intelligence

Κεφάλαιο 1ο: Ταξινόμηση Δεδομένων

1.1. Εισαγωγή

1.1.1 Εισαγωγή στην έννοια ταξινόμησης δεδομένων

Η ταξινόμηση δεδομένων είναι μία θεμελιώδης διαδικασία στο πεδίο της μηχανικής μάθησης και της επιστήμης δεδομένων. Αφορά την αναγνώριση μοτίβων και τη δημιουργία μαθηματικών μοντέλων που μπορούν να αποδώσουν μια «ετικέτα» ή κατηγορία σε νέα δεδομένα, βασισμένα σε προηγούμενα παραδείγματα, των οποίων η κλάση είναι ήδη γνωστή. Με άλλα λόγια, πρόκειται για μια εποπτευόμενη μορφή μάθησης, στην οποία το μοντέλο εκπαιδεύεται χρησιμοποιώντας εισόδους και τις αντίστοιχες εξόδους τους. Η ταξινόμηση συναντάται καθημερινά σε εφαρμογές, όπως η αναγνώριση ανεπιθύμητης αλληλογραφίας (spam detection), η πρόβλεψη φθοράς εξαρτημάτων σε βιομηχανικά μηχανήματα, η ταξινόμηση ασθενειών βάσει ιατρικών δεδομένων, ή η αξιολόγηση αιτήσεων πίστωσης σε τραπεζικά συστήματα.

Υπάρχουν διάφοροι τύποι αλγορίθμων ταξινόμησης, καθένας με τα δικά του χαρακτηριστικά, πλεονεκτήματα και περιορισμούς, που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και των δεδομένων. Οι γραμμικοί αλγόριθμοι βασίζονται στην υπόθεση γραμμικής σχέσης μεταξύ χαρακτηριστικών και εξόδου και είναι κατάλληλοι για προβλήματα όπου οι κατηγορίες είναι διαχωρίσιμες με ευθεία γραμμή. Οι αλγόριθμοι βασισμένοι σε δέντρα απόφασης (Decision Trees, Random Forests) παρέχουν υψηλή ακρίβεια και ερμηνευσιμότητα, ενώ οι μέθοδοι κοντινότερου γείτονα, όπως ο k-NN, ταξινομούν με βάση την απόσταση σε έναν πολυχώρο χαρακτηριστικών. Επιπλέον, αλγόριθμοι όπως το Naive Bayes βασίζονται στη θεωρία πιθανοτήτων, ενώ τα νευρωνικά δίκτυα και ειδικά τα συνελικτικά (CNNs) είναι ισχυρά για πολύπλοκες δομές και υψηλής διάστασης δεδομένα. Η επιλογή του κατάλληλου μοντέλου είναι κρίσιμη για την απόδοση και την αξιοπιστία των προβλέψεων.

Ιστορικά, οι ρίζες της ταξινόμησης ως υπολογιστικό πρόβλημα ανάγονται στις αρχές της στατιστικής και της θεωρίας πιθανοτήτων. Στα τέλη του 20ού αιώνα, αλγόριθμοι όπως τα Decision Trees (δέντρα απόφασης)[1], οι μέθοδοι k-Nearest Neighbors (k-NN) και τα Support Vector Machines (SVM)[2] κυριάρχησαν στον χώρο, προσφέροντας καλά αποτελέσματα με σχετικά περιορισμένη υπολογιστική απαίτηση. Με την εξέλιξη της υπολογιστικής ισχύος και την άνοδο της εποχής των «Big Data», μοντέλα όπως τα Random Forests, αλλά και τα τεχνητά νευρωνικά δίκτυα, απέκτησαν την ικανότητα να επεξεργάζονται μεγάλα και πολύπλοκα σύνολα δεδομένων με εντυπωσιακή ακρίβεια.

Η ταξινόμηση διακρίνεται από άλλες μεθόδους μηχανικής μάθησης όπως η παλινδρόμηση ή η ομαδοποίηση (clustering). Σε αντίθεση με την παλινδρόμηση, η οποία προσπαθεί να προβλέψει μία συνεχή τιμή (π.χ. την τιμή ενός σπιτιού), η ταξινόμηση αποσκοπεί στην πρόβλεψη διακριτών κατηγοριών (π.χ. “υψηλό”, “μεσαίο”, “χαμηλό”). Από την άλλη, ενώ η ταξινόμηση είναι εποπτευόμενη, η ομαδοποίηση είναι μη εποπτευόμενη μέθοδος, χωρίς ετικέτες εκπαίδευσης, και χρησιμοποιείται κυρίως για την ανακάλυψη κρυμμένων δομών μέσα σε δεδομένα. Παρά ταύτα, πολλές φορές οι δύο μέθοδοι χρησιμοποιούνται συμπληρωματικά, ιδίως στα πρώτα στάδια εξερεύνησης δεδομένων.

Η εποπτευόμενη και η μη εποπτευόμενη μάθηση αποτελούν δύο βασικές κατηγορίες στη μηχανική μάθηση, με διαφορετικές προσεγγίσεις και στόχους. Η εποπτευόμενη μάθηση (supervised learning) χρησιμοποιεί δεδομένα που είναι επισημασμένα, δηλαδή κάθε παράδειγμα εκπαίδευσης συνοδεύεται

από την επιθυμητή έξοδο (ετικέτα). Ο στόχος είναι η δημιουργία ενός μοντέλου που μπορεί να προβλέψει τις ετικέτες για νέα, άγνωστα δεδομένα. Η ταξινόμηση (classification) και η παλινδρόμηση (regression) ανήκουν σε αυτήν την κατηγορία. Αντίθετα, η μη εποπτευόμενη μάθηση (unsupervised learning) εφαρμόζεται όταν δεν υπάρχουν ετικέτες στα δεδομένα. Ο στόχος της είναι η εύρεση μοτίβων ή δομών, όπως η ομαδοποίηση παρόμοιων σημείων (clustering) ή η μείωση διαστάσεων (π.χ. PCA). Παρότι πιο δύσκολη στην ερμηνεία, η μη εποπτευόμενη μάθηση είναι χρήσιμη στην εξερεύνηση δεδομένων.

Η επιλογή της κατάλληλης τεχνικής ταξινόμησης εξαρτάται από παράγοντες όπως το μέγεθος του συνόλου δεδομένων, η ύπαρξη ή όχι ανισορροπίας στις κατηγορίες, ο αριθμός και η ποιότητα των χαρακτηριστικών (features), καθώς και η ανάγκη ερμηνευσιμότητας του μοντέλου. Παράλληλα, η προεπεξεργασία των δεδομένων (data preprocessing), όπως η κανονικοποίηση τιμών, η διαχείριση ελλιπών τιμών ή η επιλογή των πιο σημαντικών χαρακτηριστικών, αποτελεί κρίσιμο στάδιο για την επιτυχία της ταξινόμησης.

Τέλος, η αναπαράσταση των αποτελεσμάτων μέσα από γραφήματα και τεχνικές οπτικοποίησης (όπως τα confusion matrix, τα SHAP plots ή η PCA απεικόνιση) βοηθά στην καλύτερη κατανόηση και αποδοχή του συστήματος από τους τελικούς χρήστες. Όλα τα παραπάνω καθιστούν την ταξινόμηση δεδομένων μία εξαιρετικά πολύτιμη τεχνολογία για προβλέψεις, υποστήριξη αποφάσεων και ανάλυση τάσεων σε ένα μεγάλο φάσμα εφαρμογών, από την επιστήμη έως τη βιομηχανία.

1.1.2 Η ταξινόμηση δεδομένων στο ποδόσφαιρο

Από τα μέσα του περασμένου αιώνα, το ποδόσφαιρο άρχισε να αποκτά στατιστική ταυτότητα. Ο Charles Reep, ήδη από τη δεκαετία του 1950, ανέπτυξε συστημικές αναλύσεις, δείχνοντας πόσο σημαντικές είναι οι σύντομες αλυσίδες πάσας (“chains of 3 or fewer”) για την επίτευξη γκολ. Τη δεκαετία του 1990, μικρές ομάδες όπως η Derby County και αργότερα μεγάλα κλαμπ άρχισαν να χρησιμοποιούν βίντεο και συστήματα παρακολούθησης, όπως Prozone και Opta, για τη συλλογή στοιχείων αγωνιστικής συμπεριφοράς. Τα δεδομένα αυτού του είδους αποτέλεσαν τη βάση για την πρώτη εποχή ταξινόμησης, όπου πρότυπα για ομαλές κινήσεις, τύπους επιθέσεων ή αμυντικών σχημάτων κωδικοποιήθηκαν σε ετικέτες.

Σήμερα, τόσο στην Premier League όσο και στο πανεπιστημιακό πεδίο υπάρχει εκτεταμένη χρήση AI, ML, δεδομένων εντοπισμού (tracking) και computer vision για ανάλυση τακτικών, πρόβλεψη τραυματισμών και ενίσχυση αποδόσεων . Ταυτόχρονα, κορυφαία παραδείγματα, όπως το συνεργατικό TacticAI της DeepMind με τη Liverpool, χρησιμοποιούν γεωμετρικά δίκτυα για στρατηγικές στις στημένες φάσεις[4].

Στο σύγχρονο ποδόσφαιρο, η ταξινόμηση και η ανάλυση δεδομένων δεν περιορίζεται πλέον μόνο σε στατικά στατιστικά (όπως γκολ ή ασίστ), αλλά διαχωρίζεται σε διαφορετικές κατηγορίες δεδομένων, που συνθέτουν ένα πλήρες αγωνιστικό προφίλ. Οι βασικές μορφές είναι: τα event data, που περιλαμβάνουν γεγονότα όπως πάσες, σουτ, τάκλιν και φάουλ, τα tracking data, που αναφέρονται σε συνεχείς συντεταγμένες X, Y, Z των παικτών ή της μπάλας και τα biometric data, δηλαδή μετρήσεις από wearable συσκευές (όπως GPS, παλμογράφοι), που δείχνουν επιβάρυνση, καρδιακούς παλμούς και κόπωση. Η ταξινόμηση αυτών των τύπων δεδομένων επιτρέπει στους αναλυτές να ξεχωρίσουν μοτίβα που δεν είναι ορατά στο γυμνό μάτι, δημιουργώντας βαθύτερες κατηγορίες, όπως το "tempo-based pressing" ή τα συστήματα μετάβασης. Αυτή η πολυεπίπεδη ταξινόμηση ενισχύει την ακριβή αξιολόγηση και παρέχει στις ομάδες ανταγωνιστικό πλεονέκτημα.

Βλέπουμε επίσης εφαρμογές στο πεδίο της πρόληψης τραυματισμών, όπου AI αναλύει βιομετρικά και φορετά δεδομένα, υπολογίζοντας επίπεδα κόπωσης ή κινδύνους τραυματισμού. Επιπλέον, μελέτες όπως των Moustakidis et al. (2023) συνδυάζουν XGBoost με SHAP για την εξηγήσιμη ταξινόμηση απόδοσης ομάδων, αναδεικνύοντας τους βασικούς παράγοντες που επηρεάζουν την τελική επίδοση[5].

Η ταξινόμηση δεν περιορίζεται στη συνολική απόδοση ενός ποδοσφαιριστή, αλλά εστιάζει και στον προσδιορισμό του ρόλου που διαδραματίζει μέσα στο γήπεδο. Αν και δύο μέσοι μπορεί να αγωνίζονται στην ίδια περιοχή, η λειτουργία τους διαφέρει: ο ένας είναι πιθανόν πιο δημιουργικός και ο άλλος περισσότερο ανασταλτικός. Μέσω αλγορίθμων μηχανικής μάθησης, μπορούμε να ταξινομήσουμε αυτούς τους ρόλους, συγκρίνοντας αγωνιστικά χαρακτηριστικά, όπως τύποι πάσας, θέση ανάκτησης μπάλας, ταχύτητα αντίδρασης και επιρροή στο transition. Έτσι, προκύπτουν προφίλ όπως "box-to-box", "holding midfielder", "inverted winger" ή "false nine". Αυτή η μορφή ταξινόμησης επιτρέπει σε προπονητές και αναλυτές να κατανοούν καλύτερα τη συνεισφορά κάθε παίκτη και να χτίζουν συστήματα βασισμένα όχι απλώς στη θέση, αλλά στον ρόλο και τη συμπεριφορά που επιδεικνύει ο παίκτης υπό πραγματικές συνθήκες αγώνα.

Η ταξινόμηση δεδομένων στο ποδόσφαιρο σήμερα βασίζεται σε σύνθετους δείκτες απόδοσης (KPI) και στατιστικά μοντέλα, που επιτρέπουν μια πιο αντικειμενική και επιστημονική αξιολόγηση. Ένα από τα πιο γνωστά παραδείγματα είναι το Expected Goals (xG), το οποίο αξιολογεί την ποιότητα μιας ευκαιρίας βάσει πολλών παραγόντων (θέση, τοποθέτηση τερματοφύλακα, απόσταση αμυντικών από την μπάλα κ.ά.). Άλλοι δείκτες όπως το Expected Assists (xA) ή τα progressive passes, βοηθούν στην κατηγοριοποίηση παικτών ανά ρόλο ή στυλ παιχνιδιού. Η εφαρμογή τέτοιων δεικτών στη μηχανική μάθηση επιτρέπει την αυτόματη κατηγοριοποίηση παικτών σε «δημιουργικούς μέσους», «box-to-box» ή «deep-lying playmakers». Η παρουσία αυτών των KPI ενισχύει την ερμηνευσιμότητα των μοντέλων ταξινόμησης και καθιστά την απόδοση συγκρίσιμη ανάμεσα σε διαφορετικές διοργανώσεις, χώρες ή επίπεδα, συμβάλλοντας στη βελτιστοποίηση του ταλέντου και της στρατηγικής.

Η ταξινόμηση δεδομένων παίζει κομβικό ρόλο και στο κομμάτι των μεταγραφών, ιδιαίτερα μέσω του λεγόμενου "data-driven scouting". Υπάρχουν ομάδες κολοσοί, όπως η Brentford, όπου έχουν ένα μοντέλο εύρεσης νέων ποδοσφαιριστών με προσδοκία (potential) να ανέβουν επίπεδα. Αντί οι ομάδες να βασίζονται αποκλειστικά στην προσωπική παρατήρηση ενός σκάουτερ, σήμερα χρησιμοποιούν αλγορίθμους για να εντοπίζουν παίκτες που ταιριάζουν σε συγκεκριμένα αγωνιστικά προφίλ. Για παράδειγμα, μια ομάδα που παίζει με high pressing μπορεί να αναζητά παίκτες με υψηλό pressing efficiency και καλό recovery rate. Μέσω της ταξινόμησης, μπορούν να εντοπιστούν ποδοσφαιριστές που ίσως είναι υποτιμημένοι, παρόλο που τα δεδομένα δείχνουν μεγάλη τακτική προσαρμοστικότητα. Επιπλέον, η χρήση clustering βοηθά στον ομαδοποιημένο εντοπισμό παικτών με παρόμοιο στυλ, γεγονός που διευκολύνει την αναζήτηση εναλλακτικών λύσεων όταν ένας παίκτης δεν είναι διαθέσιμος. Η ορθή ταξινόμηση εξοικονομεί χρόνο, μειώνει τα ρίσκα και προσφέρει ανταγωνιστικό πλεονέκτημα.

Παρότι η ταξινόμηση δεδομένων έχει αναδείξει νέα πρότυπα αξιολόγησης στο ποδόσφαιρο, εμφανίζονται σημαντικά ηθικά και πρακτικά ζητήματα. Πρώτον, η υπερβολική εξάρτηση από τα δεδομένα ενδέχεται να οδηγήσει στην «αορατότητα» παικτών των οποίων η προσφορά δεν μετριέται εύκολα με νούμερα. Παράδειγμα αποτελούν οι παίκτες με ηγετικό πνεύμα, επιρροή στα αποδυτήρια ή ψυχραιμία σε κρίσιμες φάσεις, στοιχεία που δύσκολα κωδικοποιούνται. Δεύτερον, υπάρχει ο κίνδυνος παγίωσης στερεοτύπων μέσα από τα δεδομένα, δηλαδή να ταξινομούνται παίκτες με βάση το φύλο, την ηλικία ή ακόμα και την εθνικότητα, όταν τα δεδομένα εκπαίδευσης είναι προκατειλημμένα. Επιπλέον, οι αλγόριθμοι μπορεί να οδηγήσουν σε «κλείδωμα» ρόλων, περιορίζοντας την ευελιξία ενός

παίκτη. Η σωστή ερμηνεία των δεδομένων, ο ανθρώπινος παράγοντας και η ηθική χρήση των εργαλείων είναι απαραίτητα για να παραμείνει η τεχνολογία εργαλείο και όχι εμπόδιο.

Στο μέλλον, αναμένουμε μεγαλύτερη χρήση deep learning, computer vision, real-time analytics και τεχνολογιών όπως οι ψηφιακοί δίδυμοι (digital twins), που θα συνδέουν με ακρίβεια τη φυσική κατάσταση και τη θέση των παικτών σε εικονικά μοντέλα, ολοκληρώνοντας έτσι τη σύνδεση αποτελεσματικότητας, ασφάλειας και στρατηγικής αξιοποίησης .

Αξίζει να σημειωθεί ότι στην σύγχρονη εποχή σχεδόν όλες οι επαγγελματικές ομάδες είτε έχουν εξελιχθεί είτε προσπαθούν να εξελιχθούν μέσω των δεδομένων. Οι πιο προχωρημένες ομάδες, παραδείγματος χάρη οι αγγλικές ομάδες, βασίζονται σε επαγγελματικές πλατφόρμες (όπως SciSports, Wyscout, Opta) για να αναλύουν δεδομένα και να βγάζουν συμπεράσματα από αυτές. Το ML στην σύγχρονη εποχή χρησιμοποιείται τόσο για την ανάλυση των αποδόσεων των ποδοσφαιριστών όσο και για τον τρόπο παιχνιδιού μιας ομάδας. Μελετώντας στατιστικά οι γνώστες του αθλήματος βγάζουν συμπεράσματα και οργανώνουν τον δικό τους τρόπο αντιμετώπισης μιας ομάδας μέσω αυτών των συμπερασμάτων. Επίσης, τα δεδομένα δεν παίζουν ρόλο μόνο στην αγωνιστική εικόνα αλλά και σε εξωγηπεδικό επίπεδο. Ανακαλύπτοντας συνεχώς νέους τρόπους να παρακολουθείς παίκτες βάση των στατιστικών τους, αρκετές ομάδες προσεγγίζουν πλέον πιο εύκολα ποδοσφαιριστές που ταιριάζουν στα θέλω και στην εικόνα μιας ομάδας. Στις μέρες μας, όλο και περισσότερες ομάδες έχουν πολλαπλά γκρουπ με άτομα που δουλεύουν πάνω στην ανάλυση των δεδομένων και χωρίζονται σε ακόμα πιο μικρές υποομάδες με βάση τον κλάδο τους. Για παράδειγμα μια ομάδα αναλυτών αγώνων χωρίζονται στον επικεφαλή και στα μέλη, τα οποία ο κάθε ένας τους ξεχωριστά ασχολείται με διαφορετικό κομμάτι όπως ανάλυση αντιπάλου συστήματος, ανάλυση αντιπάλων παικτών κ.ο.κ.

1.1.3 Η ταξινόμηση των δεδομένων στις βάσεις δεδομένων

Οι βάσεις δεδομένων αποτελούν θεμέλιο της ψηφιακής εποχής, καθώς επιτρέπουν την οργανωμένη αποθήκευση, διαχείριση και ανάκτηση μεγάλου όγκου πληροφοριών. Πρόκειται για δομές που επιτρέπουν την αποδοτική αναζήτηση, ταξινόμηση και επεξεργασία δεδομένων μέσω συστημάτων διαχείρισης βάσεων δεδομένων (DBMS), όπως το MySQL, PostgreSQL, SQLite και άλλα. Αντί τα δεδομένα να είναι αποθηκευμένα σε απλά αρχεία ή χειροκίνητες λίστες, η χρήση βάσεων δεδομένων εξασφαλίζει σταθερότητα, αξιοπιστία, ταχύτητα και δυνατότητα σύνθετων ερωτημάτων. Μέσω των κατάλληλων εντολών (SQL), ο χρήστης μπορεί να προσθέτει, να διαγράφει, να τροποποιεί ή να αναζητά πληροφορίες γρήγορα και με ακρίβεια. Οι βάσεις δεδομένων υποστηρίζουν επίσης σχέσεις μεταξύ διαφορετικών πινάκων και διασφαλίζουν την ακεραιότητα των πληροφοριών, κάτι που είναι κρίσιμο σε εφαρμογές όπου η δομημένη πληροφορία παίζει κομβικό ρόλο.

Η ταξινόμηση των δεδομένων στις βάσεις δεδομένων αποτελεί μία από τις βασικότερες λειτουργίες που διευκολύνουν τη διαχείριση και αξιοποίηση των πληροφοριών. Μέσω της ταξινόμησης (sorting), τα δεδομένα μπορούν να οργανωθούν με βάση ένα ή περισσότερα κριτήρια, όπως αριθμητικά πεδία, ημερομηνίες ή αλφαβητική σειρά, γεγονός που διευκολύνει την ανάγνωση, την εύρεση και την ανάλυση. Η SQL, η βασική γλώσσα διαχείρισης σχεσιακών βάσεων, υποστηρίζει ταξινόμηση μέσω της εντολής ORDER BY, επιτρέποντας αύξουσα (ASC) ή φθίνουσα (DESC) σειρά. Η σωστή ταξινόμηση καθίσταται κρίσιμη ειδικά όταν διαχειριζόμαστε μεγάλα σύνολα δεδομένων ή όταν εκτελούνται στατιστικές αναλύσεις. Επιπλέον, η ταξινόμηση συνδυάζεται συχνά με άλλες λειτουργίες όπως GROUP BY για ακόμα πιο εξειδικευμένη παρουσίαση των δεδομένων. Σε εφαρμογές όπως αυτή

της παρούσας πτυχιακής, η ταξινόμηση επιτρέπει την ευκολότερη αξιολόγηση ποδοσφαιριστών, την ιεράρχηση αποτελεσμάτων και την κατανόηση των πιο καθοριστικών χαρακτηριστικών.

Η πολύ-κριτηριακή ταξινόμηση αναφέρεται στη δυνατότητα ταξινόμησης δεδομένων με βάση περισσότερα από ένα κριτήρια ταυτόχρονα. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη όταν οι χρήστες επιθυμούν να εντοπίσουν εγγραφές που ικανοποιούν σύνθετα χαρακτηριστικά, π.χ. παίκτες που έχουν υψηλό συνολικό σκορ (overall) αλλά και συγκεκριμένη αντοχή (pace) ή ορατότητα (vision). Στις βάσεις δεδομένων, αυτό επιτυγχάνεται με SQL εντολές που χρησιμοποιούν διαδοχικά πεδία στο ORDER BY. Έτσι, οι παίκτες ταξινομούνται πρώτα με βάση το συνολικό σκορ και σε περίπτωση ισοβαθμίας, σύμφωνα με την ταχύτητά τους. Στην εφαρμογή της παρούσας πτυχιακής, η δυνατότητα αυτή επιτρέπει στους χρήστες να εντοπίζουν με μεγαλύτερη ακρίβεια ποδοσφαιριστές που ταιριάζουν στις προδιαγραφές τους. Παράλληλα, ενισχύει την ερμηνευσιμότητα των αποτελεσμάτων, καθώς προσφέρει πιο στοχευμένη παρουσίαση των δεδομένων, απαραίτητη για την κατανόηση των ρόλων και δυνατοτήτων κάθε παίκτη.

Η κανονικοποίηση και η σωστή μορφοποίηση των δεδομένων αποτελούν απαραίτητο στάδιο πριν από οποιαδήποτε διαδικασία ταξινόμησης, ειδικά όταν τα δεδομένα προέρχονται από πολλαπλές πηγές ή όταν συλλέγονται χειροκίνητα. Η κανονικοποίηση αφορά την ομογενοποίηση των τιμών, π.χ. με τη μετατροπή όλων των πεζών και κεφαλαίων χαρακτήρων σε κοινή μορφή ή την αφαίρεση ειδικών συμβόλων και κενών. Για παράδειγμα, ο όρος "Right" δεν πρέπει να αποθηκεύεται άλλοτε ως "right" και άλλοτε ως "RIGHT", καθώς αυτό θα προκαλέσει αναντιστοιχίες κατά την αναζήτηση ή ταξινόμηση. Επιπλέον, η τυποποίηση αριθμητικών πεδίων (όπως δεκαδικά ψηφία, null τιμές) διευκολύνει τη σωστή λειτουργία αλγορίθμων και ερωτημάτων. Στην εφαρμογή της πτυχιακής, το βήμα αυτό είναι κρίσιμο, ώστε η ταξινόμηση σε λίστες ή πίνακες παικτών να είναι αξιόπιστη και ο χρήστης να βλέπει ακριβείς και πλήρως αντιπροσωπευτικές πληροφορίες.

Η χρήση βάσης δεδομένων σε μια εφαρμογή, σε σύγκριση με την απλή τοπική αποθήκευση δεδομένων (π.χ. σε αρχεία .xls ή .csv), προσφέρει πολλαπλά πλεονεκτήματα. Πρώτα απ’ όλα, παρέχει ευκολότερη πρόσβαση σε πολύπλοκες και σχετιζόμενες πληροφορίες, επιτρέποντας δομημένες αναζητήσεις και διαχείριση μέσω SQL. Επιπλέον, υποστηρίζει καλύτερη επεκτασιμότητα, δηλαδή η εφαρμογή μπορεί να μεγαλώσει σε μέγεθος χωρίς σημαντική επιβάρυνση στη διαχείριση των δεδομένων. Ακόμη, η αξιοπιστία των βάσεων δεδομένων περιορίζει τα σφάλματα, ενώ εξασφαλίζεται μεγαλύτερη ασφάλεια μέσω δικαιωμάτων πρόσβασης. Τέλος, η χρήση βάσης διευκολύνει τη δημιουργία αντιγράφων ασφαλείας και την αποκατάσταση δεδομένων, κάτι που είναι πιο δύσκολο με αρχεία τοπικής αποθήκευσης. Η οργάνωση, η ταχύτητα και η σταθερότητα καθιστούν τις βάσεις δεδομένων ιδανική λύση για εφαρμογές που διαχειρίζονται δυναμικά δεδομένα.

Όσον αφορά τις διαφορές μεταξύ τοπικής και απομακρυσμένης βάσης δεδομένων, έγκειται στο πού αποθηκεύονται και πώς αποκτάται πρόσβαση στα δεδομένα. Μια τοπική βάση (όπως η SQLite) αποθηκεύεται απευθείας στον υπολογιστή όπου εκτελείται η εφαρμογή. Παρέχει ταχύτητα και ευκολία στην ανάπτυξη, χωρίς να απαιτείται σύνδεση στο διαδίκτυο. Αντίθετα, μια απομακρυσμένη βάση (π.χ. σε MySQL server) βρίσκεται σε εξυπηρετητή (server) και επιτρέπει ταυτόχρονη πρόσβαση από πολλαπλούς χρήστες και συσκευές μέσω δικτύου ή διαδικτύου. Αυτό δίνει πλεονέκτημα σε εφαρμογές με πολλούς χρήστες ή σε περιβάλλοντα cloud. Ωστόσο, η απομακρυσμένη πρόσβαση απαιτεί αυξημένα μέτρα ασφαλείας, έλεγχο συνδέσεων και μπορεί να επηρεαστεί από προβλήματα δικτύου. Συνολικά, η επιλογή μεταξύ των δύο εξαρτάται από το είδος και το μέγεθος της εφαρμογής, τις ανάγκες προσβασιμότητας και τη μελλοντική επεκτασιμότητα.

Τέλος, η ταξινόμηση στις βάσεις δεδομένων διαφέρει από την προβλεπτική ταξινόμηση που επιτυγχάνεται μέσω αλγορίθμων μηχανικής μάθησης (ML), όμως στην πράξη οι δύο διαδικασίες συνεργάζονται στενά. Στο πλαίσιο της παρούσας εφαρμογής, η ταξινόμηση της βάσης χρησιμοποιείται για την προετοιμασία των δεδομένων, την ιεράρχηση αποτελεσμάτων, καθώς και την εμφάνιση πληροφοριών με οργανωμένο τρόπο. Από την άλλη, οι αλγόριθμοι ML, όπως οι k-NN, CNN ή Random Forest, πραγματοποιούν προβλέψεις για την κατηγορία (θέση ή περιοχή) ανήκει ένας παίκτης με βάση τις τιμές των χαρακτηριστικών του. Το output αυτής της πρόβλεψης μπορεί με τη σειρά του να ενσωματωθεί και να ταξινομηθεί στην ίδια τη βάση, ώστε να καταχωρηθεί μόνιμα ή να συγκριθεί με άλλες προβλέψεις. Η σύνδεση αυτών των δύο επιπέδων – στατικής ταξινόμησης (βάση) και δυναμικής πρόβλεψης (ML) – προσφέρει ένα ισχυρό εργαλείο ανάλυσης και παρουσίασης αποτελεσμάτων.

1.1.4 Η ταξινόμηση δεδομένων στην παρούσα Π.Ε.

Ξεκινώντας την ανάλυση του θέματος της παρούσας πτυχιακής, το πρώτο βήμα ήταν ο εντοπισμός και η προετοιμασία των δεδομένων που θα χρησιμοποιούνταν για την εκπαίδευση και τη δοκιμή των αλγορίθμων ταξινόμησης. Από ιστοσελίδες με διαθέσιμα ανοιχτά datasets[8], εντοπίστηκε ένα αρχείο που περιλάμβανε τα στατιστικά των ενεργών ποδοσφαιριστών. Το dataset αυτό χρησιμοποιήθηκε ως το σύνολο εκπαίδευσης (training data), καθώς διέθετε πάνω από 40 χαρακτηριστικά ανά παίκτη, συμπεριλαμβανομένων στατιστικών απόδοσης (όπως short passing, dribbling, heading), φυσικών χαρακτηριστικών (ύψος, βάρος, ηλικία), και άλλων σημαντικών μεταβλητών που επηρεάζουν την αγωνιστική εικόνα ενός ποδοσφαιριστή. Φυσικά για την ανάλυση στην εφαρμογή κρατήθηκαν όσα χαρακτηριστικά ήταν χρήσιμα και καθαρίστηκαν δεδομένα (cleaning data) τα οποία δεν ενδιέφεραν τον διαχειριστή και πιθανώς και τον χρήστη, όπως μισθός, εμφανίσεις με την ομάδα, διάρκεια συμβολαίου κ.α..

Αντίθετα, η ανεύρεση ενός έτοιμου dataset με ποδοσφαιριστές, οι οποίοι είχαν αποσυρθεί από την ενεργό δράση, αποδείχθηκε ιδιαίτερα δύσκολη. Μετά από εκτενή αναζήτηση, δεν βρέθηκε πλήρες σύνολο δεδομένων, γεγονός που οδήγησε στη δημιουργία ενός dataset από την αρχή. Χρησιμοποιήθηκαν πληροφορίες από την πλατφόρμα Futwiz[9], η οποία φιλοξενεί στοιχεία για τους “icon players” του FIFA. Τα δεδομένα εισήχθησαν χειροκίνητα σε τοπικό αρχείο και στην συνέχεια φορτώθηκαν μέσω κώδικα στην βάση, διαμορφώνοντας ένα σύνολο δοκιμών (testing data) το οποίο εστιάζει σε ανενεργούς παίκτες, αξιοποιήσιμο για ταξινόμηση.

Η βάση δεδομένων που επιλέχθηκε ήταν η SQLite3, λόγω της αυτοτελούς λειτουργίας της, της ευκολίας εγκατάστασης και της ταχύτητας επεξεργασίας. Δεν απαιτεί ξεχωριστό server και συνεργάζεται άψογα με Python, καθιστώντας την ιδανική για μικρομεσαίες εφαρμογές με έμφαση στην τοπική ανάπτυξη. Παρότι μια απομακρυσμένη βάση θα επέτρεπε πρόσβαση από πολλαπλούς χρήστες, η τοπική επιλογή ενίσχυσε τη σταθερότητα, την ταχύτητα και την ευκολία υλοποίησης. Έτσι διαμορφώθηκε ένα σταθερό, αξιόπιστο και πλήρως διαχειρίσιμο περιβάλλον για την ανάλυση ταξινόμησης παικτών.

Σχήμα 1.1.4 Σχήμα Βάσης Δεδομένων

1.2. Πίνακας δεδομένων ενεργών ποδοσφαιριστών

Στον χώρο του ποδοσφαίρου, τα χαρακτηριστικά ενός ποδοσφαιριστή αποτελούν κρίσιμους δείκτες των ικανοτήτων του και των πιθανών ρόλων που μπορεί να αναλάβει εντός του αγωνιστικού χώρου. Αυτά τα χαρακτηριστικά χωρίζονται συνήθως σε κατηγορίες, όπως τεχνικά (τεχνική πάσα, ντρίμπλα, έλεγχος μπάλας), φυσικά (ταχύτητα, αντοχή, δύναμη), τακτικά (τοποθέτηση, αντίληψη παιχνιδιού), και ψυχολογικά (ψυχραιμία, επιθετικότητα, αποφασιστικότητα). Η ακριβής μέτρησή τους είναι αποτέλεσμα παρατήρησης, στατιστικής ανάλυσης και, πλέον, προηγμένων τεχνολογιών ανίχνευσης της απόδοσης. Τα χαρακτηριστικά αυτά δεν περιγράφουν απλώς τις δυνατότητες του παίκτη αλλά καθορίζουν και την προσαρμοστικότητά του σε συγκεκριμένες θέσεις ή τακτικές.

Με την πρόοδο της τεχνολογίας και την εδραίωση της ανάλυσης δεδομένων στο ποδόσφαιρο, τα χαρακτηριστικά των παικτών κωδικοποιούνται και ποσοτικοποιούνται με τρόπο που επιτρέπει την άμεση σύγκριση και επεξεργασία από υπολογιστικά συστήματα. Αυτό γίνεται είτε με βάση την απόδοση σε πραγματικές συνθήκες (π.χ. αριθμός επιτυχημένων μεταβιβάσεων ανά αγώνα) είτε μέσω αξιολογήσεων που παρέχονται από αναγνωρισμένες πλατφόρμες, όπως το video game FIFA (που χρησιμοποιείται στην Π.Ε.) ή η βάση δεδομένων της Opta. Έτσι, ένα χαρακτηριστικό όπως το "vision" (όραση παιχνιδιού) ή το "ball control" (έλεγχος μπάλας) μετατρέπεται σε αριθμητική τιμή από το 1 έως το 99, κάνοντας το προσιτό σε αλγορίθμους μηχανικής μάθησης.

Στην εφαρμογή μας, η πηγή των χαρακτηριστικών προέρχεται από το σύστημα αξιολόγησης του FIFA, το οποίο μετατρέπει κάθε ικανότητα του ποδοσφαιριστή σε μια κλίμακα 0–99. Για παράδειγμα, το χαρακτηριστικό attacking_short_passing, που αφορά την ικανότητα κοντινής πάσας, δεν προκύπτει απλά από στατιστικά αριθμητικά δεδομένα, αλλά βασίζεται σε ένα σύνθετο μείγμα στοιχείων, όπως η ακρίβεια σε μικρές αποστάσεις, η ταχύτητα αντίδρασης σε κοντινές συνεργασίες και η συνέπεια στην πάσα υπό πίεση. Οι τιμές που παρέχει το FIFA δημιουργούνται από χιλιάδες παρατηρήσεις, αναλύσεις βίντεο και απόψεις scouts, με στόχο να αποδώσουν όσο το δυνατόν πιο αντικειμενικά την πραγματική ποιότητα του παίκτη στο συγκεκριμένο τομέα. Αυτό επιτρέπει στο μοντέλο ταξινόμησης της εφαρμογής μας να έχει πρόσβαση σε έγκυρα και σταθερά δομημένα δεδομένα, τα οποία μπορούν να αξιοποιηθούν με συνέπεια στη μηχανική μάθηση.

Τα χαρακτηριστικά που χρησιμοποιήθηκαν ήταν τα εξής:

- Player_id
- Short name

- Player positions
- Overall
- Potential
- Age
- Height
- Weight
- Nationality Name
- Preferred Foot
- Weak Foot
- Pace
- Shooting
- Passing
- Dribbling
- Defending
- Physic
- Attacking_crossing
- Attacking_finishing
- Attacking_heading_accuracy
- Attacking_short_passing
- Attacking_volleys
- Skill_dribbling
- Skill_curve
- Skill_fk_accuracy
- Skill_long_passing
- Skill_ball_control
- Movement_acceleration
- Movement_sprint_speed
- Movement_agility
- Movement_reactions
- Movement_balance
- Power_shot_power
- Power_jumping
- Power_stamina
- Power_strength
- Power_long_shots
- Mentality_aggression
- Mentality_interceptions
- Mentality_positioning
- Mentality_vision
- Mentality_penalties
- Mentality_composure
- Defending_marking_awareness
- Defending_standing_tackle
- Defending_sliding_tackle

Αφού ολοκληρώθηκε αυτή η διαδικασία εκκαθάρισης των δεδομένων στην συνέχεια αφού αποθηκευτηκαν εκ νέου τα δεδομένα στο αρχείο, πραγματοποιήθηκε η μετατροπή, αυτήν την φορά από excel σε csv, έτσι ώστε να εύκολα διαχειρήσιμα με μορφή dataframe στην python και στην συνέχεια ως πίνακα στην βάση δεδομένων.

Αρχικά, για να αποθηκευτεί αυτός ο πίνακας στην βάση χρειαζόταν ένα αρχείο διαγραφής και εισαγωγής δεδομένων. Στην αρχή θα έπρεπε να υπάρχει ένα αρχείο που να διαγράφει και να

ξαναδημιουργεί τον πίνακα στην βάση. Με αυτήν την απόφαση δημιουργήθηκε ένα αρχείο που αρχικά ανοίγει την σύνδεση με την βάση, διαγράφει τον πίνακα players αν υπάρχει ήδη και στην συνέχεια εισάγει τα δεδομένα από το csv αρχείο, τα επεξεργάζεται και στο τέλος τα αποθηκεύει ξανά στον πίνακα στην βάση δεδομένων και εν τέλει κλείνει την σύνδεση με την βάση. Η επεξεργασία που γίνεται στα δεδομένα είναι ως εξής:

- Διαγραφή όλων των διπλότυπων εγγραφών
- Διαγραφή της στήλης potential (δεν ωφελεί κάπου στην ανάλυση)
- Διαγραφή όλων των παικτών που αγωνίζονται ως τερματοφύλακες
- Διαγραφή όλων των παικτών που έχουν overall κάτω από 69 (για την πιο κατατοπισμένη ανάλυση συγκριτικά με τους ανενεργούς ποδοσφαιριστές)
- Διαχωρισμός θέσεων κρατώντας μόνο την πρώτη και βασική θέση που αγωνίζεται ένας παίκτης
- Προσθήκη της στήλης area (που θα βοηθήσει στην εύρεση της περιοχής των μελλοντικών παικτών στην ανάλυση)
- Τοποθέτηση τιμών area στους παίκτες με βάση την θέση που αγωνίζονται
- Εμφάνιση του πίνακα στον τερματικό για επιβεβαίωση σωστής λειτουργίας

1.3. Πίνακας δεδομένων ανενεργών ποδοσφαιριστών

Οι ποδοσφαιριστές που έχουν αποσυρθεί από την ενεργό δράση εξακολουθούν να κατέχουν μια ιδιαίτερη θέση τόσο στην ιστορία του αθλήματος όσο και στην πολιτισμική συνείδηση των φιλάθλων. Συχνά προκαλούν έντονες συγκρίσεις με τους σύγχρονους παίκτες, είτε σε επίπεδο στατιστικών επιδόσεων είτε ως προς το στυλ παιχνιδιού, την επιρροή τους στην ομάδα ή την επίδρασή τους στην εποχή τους. Συζητήσεις όπως «ήταν καλύτερος ο Maradona ή ο Messi;» ή «πώς θα τα πήγαινε ο Pelé αν έπαιζε σήμερα;» αποτελούν ενδείξεις της διαχρονικής γοητείας αυτών των ποδοσφαιρικών θρύλων. Καθώς οι τεχνικές καταγραφής δεδομένων στο παρελθόν ήταν ελλιπείς ή λιγότερο αντικειμενικές, η προσπάθεια σύγκρισης με βάση αντικειμενικά κριτήρια καθίσταται ακόμη πιο απαιτητική, δημιουργώντας έτσι την ανάγκη για πιο συστηματική και αναλυτική αποτίμηση.

Το δημοφιλές βιντεοπαιχνίδι FIFA, κατανοώντας αυτήν τη διαχρονική αξία των παλαιών παικτών, εισήγαγε τους λεγόμενους Icon Players. Πρόκειται για ψηφιακές αναπαραστάσεις ιστορικών ποδοσφαιριστών όπως ο Johan Cruyff, ο Ronaldo Nazário, ο Zinedine Zidane, οι οποίοι περιλαμβάνονται στο παιχνίδι με στόχο να τιμηθεί η συμβολή τους στο ποδόσφαιρο και να δοθεί η δυνατότητα στους παίκτες να τους εντάξουν στις ομάδες τους, δίπλα σε σύγχρονους αστέρες. Τα στατιστικά αυτών των παικτών σχεδιάζονται προσεκτικά ώστε να αντανακλούν την απόδοσή τους στην ακμή της καριέρας τους, ενώ κάθε Icon μπορεί να εμφανίζεται με διαφορετικές κάρτες (π.χ. Baby, Mid, Prime), οι οποίες αντιστοιχούν σε διαφορετικές φάσεις της καριέρας τους. Η διατήρησή τους στο παιχνίδι, παρότι δεν είναι πλέον ενεργοί, επιτρέπει όχι μόνο τη δημιουργία ιστορικών ομάδων αλλά και τη σύγκριση του στυλ και των ικανοτήτων τους με τα δεδομένα των σύγχρονων παικτών μέσα από κοινές στατιστικές μεταβλητές. Αυτό καθιστά τους Icon Players ιδανική βάση για εφαρμογές που συνδυάζουν σύγχρονη ανάλυση δεδομένων με ιστορικές συγκρίσεις, όπως συμβαίνει και στην παρούσα πτυχιακή εργασία.

Για τον πίνακα δεδομένων με τους ανενεργούς ποδοσφαιριστές, χρειάστηκε να γραφτεί χειροκίνητα σε μορφή excel και στην συνέχεια να μετατραπεί σε csv για τις ανάγκες προσθήκης στην εφαρμογή. Αρχικά, με τον ίδιο τρόπο που έγινε επεξεργασία των δεδομένων στους ενεργούς παίκτες με το ίδιο

τρόπο έγινε και η εισαγωγή των στοιχείων των ανενεργών ποδοσφαιριστών. Έπρεπε τα χαρακτηριστικά των ενεργών ποδοσφαιριστών και τα χαρακτηριστικά των ανενεργών ποδοσφαιριστών να βρίσκονται στην ίδια σειρά, με τις ίδιες παραμέτρους και με τους ίδιους τύπους δεδομένων. Επίσης, έπρεπε να ληφθεί υπόψιν και ποιοι ποδοσφαιριστές θα προστεθούν έτσι ώστε η ανάλυση να έχει ουσιώδη αποτέλεσμα. Αρκετές φορές κατά την διάρκεια των δοκιμών της εφαρμογής προστέθηκαν και αφαιρέθηκαν ανενεργοί ποδοσφαιριστές για να υπάρξει μεγαλύτερο βάθος επιλογών και να μπορέσουμε να οδηγηθούμε σε ποιο ασφαλή συμπεράσματα.

Όπως και στους ενεργούς ποδοσφαιριστές, έτσι και στους ανενεργούς θα έπρεπε να υπάρχει ένα πρόγραμμα, το οποίο θα διαγράφει και θα ξανά αποθηκεύει τους παίκτες στον πίνακα της βάσης έπειτα από κάποιες δοκιμές. Οπότε το πρόγραμμα με τους ανενεργούς ποδοσφαιριστές κάνει τι εξής ενέργειες:

- Ανοίγει την σύνδεση με την βάση
- Φορτώνει το csv αρχείο σε ένα dataframe
- Διαγράφει τον προϋπάρχον πίνακα
- Δημιουργεί έναν πίνακα με σωστά ορισμένα πεδία (τύπους, κλειδία, κλπ)
- Εισάγει τα δεδομένα στο πίνακα
- Φιλτράρει το dataframe για να έχει μόνο τις προαπαιτούμενες στήλες
- Μετατρέπει τα δεδομένα ανά εγγραφή και τις εισάγει στον πίνακα
- Διαγράφει όλες τις τιμές του χαρακτηριστικού player positions (διότι θα βρεθεί από τους αλγορίθμους ταξινόμησης)
- Προσθέτει την στήλη area
- Κλείνει την σύνδεση στην βάση

1.4. Πίνακας δεδομένων περιληπτικών στοιχείων θέσεων

Η κάθε θέση στο ποδόσφαιρο παίζει ξεχωριστό ρόλο. Άλλες ευθύνες έχει ο κεντρικός αμυντικός, άλλες ένας μέσος και άλλες ένας επιθετικός. Ο στόπερ (κεντρικός αμυντικός), για παράδειγμα, έχει ως κύριο στόχο την αποτροπή επιθέσεων, την καλή τοποθέτηση και την εναέρια υπεροχή, ενώ ο επιθετικός εστιάζει στο σκοράρισμα, την εκμετάλλευση των ευκαιριών και την πίεση στην άμυνα του αντιπάλου. Ο μέσος, ιδίως ο κεντρικός ή «box-to-box», αποτελεί τον συνδετικό κρίκο ανάμεσα σε άμυνα και επίθεση, συμμετέχοντας σε μεταβάσεις, κυκλοφορία μπάλας και δημιουργία φάσεων. Οι απαιτήσεις κάθε θέσης είναι διαφορετικές: ένας παίκτης ο οποίος αγωνίζεται στον χώρο του κέντρου προφανώς θα χρειάζεται περισσότεροι αντοχή, διότι καλύπτει μεγάλη έκταση του αγωνιστικού χώρου κατά την διάρκεια ενός αγώνα

Στο παλαιότερο ποδόσφαιρο, ιδίως μέχρι και τη δεκαετία του 1990, οι θέσεις θεωρούνταν πολύ πιο στατικές. Ένας δεξιός μπακ, για παράδειγμα, σπάνια περνούσε τη μεσαία γραμμή, και οι παίκτες έμεναν «πιστοί» στην περιοχή δράσης τους. Οι σχηματισμοί ήταν αυστηροί, και οι ρόλοι εντός αυτών σχεδόν ανελαστικοί. Αντίθετα, στο σύγχρονο ποδόσφαιρο οι παίκτες οφείλουν να έχουν μεγαλύτερη ευελιξία, υψηλό επίπεδο φυσικής κατάστασης και δυνατότητα συμμετοχής τόσο στην επίθεση όσο και στην άμυνα. Ο ακραίος αμυντικός πλέον λειτουργεί συχνά ως εξτρά δημιουργός (inverted fullback ή overlapping fullback), ενώ οι επιθετικοί συχνά γυρίζουν πίσω για υποστήριξη στην κυκλοφορία της μπάλας (passing game), ή πιέζουν ψηλά σε φάση άμυνας. Παίκτες αλλάζουν ρόλους συνεχώς εντός παιχνιδιού, σε καταστάσεις build-up, γρήγορης μετάβασης (transitions) ή reorganizing μετά από κατοχή. Αυτή η πολυπλοκότητα οδηγεί και σε μεταβολή του τρόπου αξιολόγησης των

ποδοσφαιριστών, καθιστώντας σημαντικά όχι μόνο τα στατικά τους χαρακτηριστικά, αλλά και τα δυναμικά, όπως η λήψη απόφασης υπό πίεση ή η ικανότητα προσαρμογής σε διαφορετικές τακτικές.

Στην παρούσα Π.Ε. ο πίνακας δεδομένων με περιληπτική περιγραφή των θέσεων είναι ένας πίνακας όπου έχει για κάθε θέση λίγα λόγια για τον ρόλο του παίκτη που αγωνίζεται σε αυτήν την θέση. Αυτή η περιγραφή γράφτηκε σε excel και μετά έγινε μετατροπή σε csv για ευκολία αποθήκευσης στην βάση με πολύ λίγο κώδικα. Η εισαγωγή αυτών των στοιχείων γίνεται με τρεις γραμμές κώδικα όπου στην παρούσα Π.Ε. βρίσκονται σε σχόλιο στο αρχείο με τους ανενεργούς ποδοσφαιριστές.

Κεφάλαιο 2ο: Σχεδιασμός Εφαρμογής

2.1. Εισαγωγή

Η γλώσσα προγραμματισμού Python επιλέχθηκε ως κύριο εργαλείο ανάπτυξης της εφαρμογής λόγω της ευελιξίας, της απλότητας σύνταξης και της ισχυρής υποστήριξης βιβλιοθηκών για επιστημονικό υπολογισμό, μηχανική μάθηση και ανάλυση δεδομένων. Η Python προσφέρει υψηλού επιπέδου αφαιρετικότητα, γεγονός που επιτρέπει την ανάπτυξη σύνθετων εφαρμογών με λιγότερο κώδικα και μεγαλύτερη αναγνωσιμότητα. Παράλληλα, η ενσωμάτωση της βιβλιοθήκης Tkinter επιτρέπει την κατασκευή παραθύρων και διαδραστικών διεπαφών με σχετική ευκολία, καθιστώντας τη γλώσσα ιδιαίτερα χρήσιμη όχι μόνο για το backend αλλά και για το εμφανισιακό μέρος του προγράμματος. Η δυνατότητα δημιουργίας κουμπιών, μηνυμάτων, φορμών και διαλόγων με ελάχιστη προσπάθεια βοήθησε στον σχεδιασμό μιας φιλικής εμπειρίας χρήστη.

Επιπλέον, η Python ενδείκνυται για την αποτελεσματική οπτικοποίηση δεδομένων, μέσω βιβλιοθηκών όπως matplotlib, seaborn και plotly, ενώ τεχνικές όπως η Ανάλυση Κύριων Συνιστωσών (PCA) υποστηρίζονται πλήρως μέσω της βιβλιοθήκης scikit-learn. Αυτό επέτρεψε την απεικόνιση πολυδιάστατων χαρακτηριστικών παικτών σε δύο διαστάσεις με κατανοητό τρόπο, συμβάλλοντας καθοριστικά στην κατανόηση των ταξινομήσεων. Όσον αφορά την αποθήκευση και διαχείριση των δεδομένων, η ενσωμάτωση της SQLite μέσω του ενσωματωμένου πακέτου sqlite3 προσέφερε μια αξιόπιστη και ελαφριά λύση για τοπική βάση δεδομένων. Η σύνδεση με την κύρια εφαρμογή είναι απλή, χωρίς την ανάγκη εξωτερικού server ή περίπλοκης διαμόρφωσης, ενώ η συνεργασία της με τη βιβλιοθήκη pandas διευκολύνει σημαντικά την ανάγνωση και τροποποίηση δεδομένων.

Από την πρώτη στιγμή ανάληψης της πτυχιακής η πρώτη ιδέα ήταν να υπάρξει μια απλή διεπαφή προσιτή σε κάθε χρήστη και να δοθεί έμφαση στην σωστή παραγωγή αποτελεσμάτων μέσω των αλγορίθμων ταξινόμησης. Με αφορμή αυτήν την ιδέα διαμορφώθηκαν κουμπιά που κατευθύνουν τον χρήστη στην απλή κατανόηση της χρήσης της εφαρμογής διαλέγοντας στην αρχή αν θέλει να επιλέξει έναν από τους ήδη εισαγόμενους παίκτες της βάσης δεδομένων ή θέλει να εισάγει χειροκίνητα τα χαρακτηριστικά μόνος του. Στην συνέχεια, εμφανίζεται το μενού με τους αλγορίθμους που είναι πολύ απλό επιλέγοντας ποια μέθοδο θέλει να δοκιμάσει κάθε φορά. Τέλος, δόθηκε μεγάλη έμφαση στα παράθυρα με τα αποτελέσματα των αλγορίθμων καθώς θα έπρεπε ο κάθε χρήστης να αναγνωρίζει τα αποτελέσματα της ταξινόμησης, για αυτό χρησιμοποιήθηκαν διαφόρων ειδών plot.

2.2. Αρχικό πλάνο

Το αρχικό πλάνο της εφαρμογής προέβλεπε δύο δυνατότητες: είτε την επιλογή ενός ήδη καταχωρημένου παίκτη από τη βάση δεδομένων είτε την εισαγωγή ενός νέου παίκτη μέσω χειροκίνητης καταχώρησης. Η σχεδίαση του κεντρικού παραθύρου επικεντρώθηκε στην απλότητα και στη σαφή διάκριση των λειτουργιών, ώστε η χρήση της εφαρμογής να είναι κατανοητή από τον τελικό χρήστη. Για τον καθορισμό της ροής υλοποιήθηκε διάγραμμα λειτουργιών (diagram), το οποίο αποτύπωνε τις αλληλεπιδράσεις μεταξύ των παραθύρων. Στη συνέχεια, αναπτύχθηκε το αρχικό παράθυρο έναρξης και οι επιλογές που οδηγούν, μέσω κατάλληλων κουμπιών, σε διαφορετικά παράθυρα. Ανεξάρτητα από τη μέθοδο που θα επιλέξει ο χρήστης (προϋπάρχον παίκτη ή χειροκίνητη εισαγωγή), η διαδικασία κατέληγε στο μενού επιλογής του αλγορίθμου ταξινόμησης. Επιπλέον, σε

κάθε αλλαγή παραθύρου υλοποιήθηκε αναπροσαρμογή του μεγέθους του, ώστε να παρουσιάζεται μόνο το απαραίτητο περιεχόμενο και να αποφεύγεται η ύπαρξη μη ορατών στοιχείων.

Κατά την ανάπτυξη, παρουσιάστηκε το ζήτημα της διατήρησης του επιλεγμένου παίκτη καθ' όλη τη διάρκεια της διαδικασίας ταξινόμησης. Αρχικά επιχειρήθηκε η αποθήκευση των χαρακτηριστικών του παίκτη σε γραμμή ενός DataFrame, όμως η μετακίνηση της εγγραφής δημιουργούσε προβλήματα διάκρισης μεταξύ τιμών και κενών πεδίων (όπως η θέση ή η περιοχή). Για την επίλυση του προβλήματος επιλέχθηκε η χρήση μιας καθολικής μεταβλητής (global variable), η οποία αποθηκεύει τον μοναδικό αναγνωριστικό αριθμό (ID) του παίκτη και χρησιμοποιείται σε κάθε μέθοδο επιλογής (είτε προϋπάρχοντος είτε χειροκίνητης εισαγωγής).

2.3. Επιλογή ανενεργού ποδοσφαιριστή από την βάση

Αφού έγιναν όλες οι προαπαιτούμενες ενέργειες για να είναι οι ανενεργοί παίκτες στην βάση θα έπρεπε να υπάρχει και επιλογή εμφάνισης για να επιλεχθούν. Αρχικά, μέσα στην μέθοδο επιλογής ενός ανενεργού ποδοσφαιριστή πρέπει να υπάρχει εισαγωγή των δεδομένων από την βάση και φιλτράρισμα αυτών, έτσι ώστε ο χρήστης να βλέπει μόνο τα απολύτως απαραίτητα στοιχεία του παίκτη που θα ήθελε να επιλέξει. Οπότε, στο παράθυρο αυτό στο επάνω μέρος συναντάμε μια λίστα σε ένα drop down list όπου αναγράφονται τα ονόματα των παικτών. Στην συνέχεια, στο κέντρο του παραθύρου βλέπουμε έναν πίνακα με τα έξι βασικά χαρακτηριστικά του ποδοσφαιριστή, που είναι:

- Pace
- Shooting
- Passing
- Dribbling
- Defending
- Physic

Κατά την διάρκεια σχεδιασμού τόσο του παραθύρο όσο και της εφαρμογής διαπιστώθηκε η ανάγκη να εμφανίζεται η ηλικία του παίκτη και το overall του. Οι ανενεργοί παίκτες κάθε χρόνο που περνούσε στην καριέρα τους είχαν διαφορετικά χαρακτηριστικά, για παράδειγμα ο G.Bale έχει επιλεχθεί με βάση τις ικανότητες που είχε στην τελευταία χρονιά του συμβολαίου του στην Ρεάλ Μαδρίτης, ενώ ο A.Vieirinha επιλέχθηκε με τις ικανότητες που είχε το 2017 όταν και επέστρεψε στον ΠΑΟΚ και όχι με τις ικανότητες που αποσύρθηκε από το ποδόσφαιρο το 2025 στην ηλικία των 39 ετών. Για λόγους εμφάνισης αυτά τα δύο χαρακτηριστικά προστέθηκαν ακριβώς κάτω από τον προηγούμενο πίνακα σε έναν δικό τους ξεχωριστό πίνακα, διότι είναι πολύ πιο εύχρηστα εμφανισιακά, γιατί αν τα πρόσθετε κανείς παράλληλα με τα υπόλοιπα θα είχαμε ένα μακροσκελές παράθυρο το οποίο και ωραίο στο μάτι δεν θα ήταν και όλα τα στοιχεία δεν θα ήταν ευδιάκριτα στον χρήστη. Τέλος, στο κάτω μέρος του παραθύρου υπάρχει το κουμπί επιλογή που κρατάει το αναγνωριστικό (id) του παίκτη και μεταφέρει τον χρήστη στο παράθυρο επιλογής μεθόδου.

2.4. Χειροκίνητη εισαγωγή στοιχείων

Για την υλοποίηση της χειροκίνητης εισαγωγής χαρακτηριστικών παίκτη κρίθηκε απαραίτητη η πλήρης κατανόηση τόσο των μεταβλητών που περιγράφουν την αγωνιστική ταυτότητα ενός ποδοσφαιριστή όσο και των προϋποθέσεων που εξασφαλίζουν την ομαλή λειτουργία της εφαρμογής.

Τα δεδομένα που αξιοποιήθηκαν προέρχονται από τη βάση του παιχνιδιού **EA Sports FC 24**, το οποίο κατηγοριοποιεί τα επιμέρους χαρακτηριστικά σε έξι βασικές ομάδες. Επιπλέον, κάθε ποδοσφαιριστής διαθέτει δηλωμένη κύρια θέση, καθώς και δευτερεύουσες θέσεις στις οποίες μπορεί να αγωνιστεί με παρόμοια απόδοση.

Στο πλαίσιο της εφαρμογής, σχεδιάστηκαν πεδία εισαγωγής δεδομένων με αντίστοιχες ετικέτες, προκειμένου ο χρήστης να καταχωρεί τα στοιχεία και τα χαρακτηριστικά του παίκτη που επιθυμεί να προσθέσει. Η διαδικασία ξεκινά με την εισαγωγή των βασικών πληροφοριών ταυτότητας του ποδοσφαιριστή (π.χ. αναγνωριστικό παίκτη, όνομα, ηλικία, ύψος, βάρος, εθνικότητα, προτιμώμενο πόδι, βαθμός αδυναμίας αδύναμου ποδιού) και ακολουθεί η συμπλήρωση των αγωνιστικών χαρακτηριστικών.

Χαρακτηριστικά εισαγωγής από τον χρήστη:

- player_id (Το αναγνωριστικό του παίκτη όπου υπάρχει μέθοδος έλεγχος διπλότυπων εγγραφών σε περίπτωση που ο χρήστης εισάγει αναγνωριστικό, το οποίο υπάρχει ήδη στην βάση)
- short_name (Όπου ο χρήστης εισάγει το πρώτο γράμμα του ονόματος του παίκτη και το επίθετο του)
- age (Όπου υπάρχει έλεγχος για εύρος τιμών)
- height_cm (Όπου υπάρχει έλεγχος για εύρος τιμών)
- weight_kg (Όπου υπάρχει έλεγχος για εύρος τιμών)
- nationality_name
- preferred_foot (Όπου υπάρχει εμφανή λίστα επιλογής Left ή Right)
- weak_foot (Όπου υπάρχει και εδώ εμφανή λίστα για να επιλέξεις τιμή από 1 εως 5)

Μετά την συμπλήρωση των γενικών στοιχείων του παίκτη ο χρήστης εισάγει τιμές στο εύρος 0-99 για τα εξής χαρακτηριστικά:

- Attacking_crossing
- Attacking_finishing
- Attacking_heading_accuracy
- Attacking_shprt_passing
- Attacking_volleys
- Skill_dribbling
- Skill_curve
- Skill_fk_accuracy
- Skill_long_passing
- Skill_ball_control
- Movement_acceleration
- Movement_sprint_speed
- Movement_agility
- Movement_reactions
- Movement_balance
- Power_shot_power
- Power_jumping
- Power_stamina
- Power_strength
- Power_long_shots
- Mentality_aggression
- Mentality_interseptions

- Mentality_positioning
- Mentality_vision
- Mentality_penalties
- Mentality_composure
- Defending_marking_awarness
- Defending_standing_tackle
- Defending_sliding_tackle

Μετά την εισαγωγή όλων των ανωτέρω χαρακτηριστικών ο χρήστης έχει την επιλογή να πατήσει ένα εκ των τεσσάρων κάτω κουμπιών:

- Πίσω. Επιστρέφει ένα παράθυρο πίσω
- Καταχώρηση. Καταχωρεί τα δεδομένα στον πίνακα στην βάση, εμφανίζει τα κατάλληλα μηνύματα αν κάποιο πεδίο δεν είναι συμπληρωμένο και αν όλα είναι συμπληρωμένα σωστά συνεχίζει το πρόγραμμα
- Προεπισκόπηση. Δείχνει σε καινούργιο παράθυρο τα 6 βασικά χαρακτηριστικά του παίκτη, τα οποία τα υπολογίζει και τα προσθέτει και αυτά στον πίνακα της βάσης δεδομένων
- Καθαρισμός. Διαγράφει όλα τα περιεχόμενα των πεδίων (καθαρίζει ακόμα και τις επιλογές των δύο λιστών)

Αφού συμπληρωθούν όλα τα πεδία σωστά και περάσει ο χρήστης από τους ελέγχους τότε το πρόγραμμα τον αφήνει να συνεχίσει στην επιλογή μια μεθόδου.

2.5. Εμφάνιση βάσης

Μετά από την ολοκλήρωση της λειτουργικότητας της εφαρμογής και την προσθήκη του πίσω κουμπιού έλειπε μια επιβεβαίωση της λειτουργίας των αλγορίθμων ταξινόμησης. Στην αρχή, για την επιβεβαίωση ότι οι αλγόριθμοι λειτουργούν σωστά και η περιοχή αλλά και η θέση του παίκτη εισάγονται σωστά στην βάση, θα έπρεπε να κλείσει η εφαρμογή και να ελεγχθεί η βάση εκ νέου. Παρατηρώντας το πίσω κουμπί, το οποίο οδηγεί στην αρχή του προγράμματος, δηλαδή εκεί που ο χρήστης είτε πατάει το κουμπί και οδηγείται στο παράθυρο επιλογής ποδοσφαιριστή είτε την εισαγωγή ενός ποδοσφαιριστή χειροκίνητα. Ακριβώς από κάτω προστέθηκε ένα κουμπί όπου ο χρήστης και θα μπορεί να δει τα αποτελέσματα της καταχώρησης των αποτελεσμάτων των αλγορίθμων αλλά και θα μπορεί αν ανοίξει για πρώτη φορά την εφαρμογή να δει την προηγούμενη ανάλυση και τι εγγραφές υπάρχουν στην βάση.

Με αυτήν την λειτουργία μπορεί ο χρήστης να επιβεβαιώνει βλέποντας τα στοιχεία της βάσης και κάθε φορά που ανοίγει την εφαρμογή θα υπενθυμίζεται τι έχει αναλύσει.

2.6. Μενού επιλογής αλγορίθμου ταξινόμησης

Οι αλγόριθμοι ταξινόμησης που χρησιμοποιήθηκαν στην Π.Ε. ήταν τρεις, ο k-NN, ο CNN και ο Random Forest (αρχικά σε συνδιασμό με την SHAP και στην συνέχεια χωρίς).

Ο αλγόριθμος k-Nearest Neighbors (k-NN) είναι μία απλή αλλά αποτελεσματική εποπτευόμενη μέθοδος ταξινόμησης, η οποία βασίζεται στην υπόθεση ότι παρόμοια δεδομένα βρίσκονται κοντά στον χώρο χαρακτηριστικών. Ουσιαστικά, για κάθε νέο δείγμα, υπολογίζει την απόσταση του από τα υπάρχοντα δεδομένα και αναθέτει την κατηγορία που συναντάται πιο συχνά στους “k” κοντινότερους

γείτονές του. Στην εφαρμογή μας, η μέθοδος αυτή χρησιμοποιήθηκε για να προσφέρει ένα εύκολο και άμεσα ερμηνεύσιμο μοντέλο που ταξινομεί έναν ανενεργό ποδοσφαιριστή (icon player) με βάση την ομοιότητα των στατιστικών του σε ενεργούς παίκτες γνωστής θέσης ή περιοχής. Η χρήση του k-NN δικαιολογείται από τη σταθερότητά του σε μικρού και μεσαίου μεγέθους datasets και την ικανότητά του να εντοπίζει μοτίβα με φυσική εγγύτητα. Επιπλέον, η οπτικοποίηση της ταξινόμησης μέσω PCA και η χρήση γειτόνων ενισχύουν τη διαφάνεια και την κατανόηση των αποτελεσμάτων από τον τελικό χρήστη.

Αν και στη θεωρία τα Convolutional Neural Networks (CNN) είναι πολύπλοκα νευρωνικά δίκτυα που χρησιμοποιούνται κυρίως για επεξεργασία εικόνας, στην εφαρμογή μας ο όρος «CNN» χρησιμοποιείται ως απλουστευμένη αναπαράσταση μιας συνελικτικής λογικής ομαδοποίησης γύρω από κέντρα (centroids). Σε αυτό το πλαίσιο, ο αλγόριθμος λειτουργεί εντοπίζοντας τον μέσο όρο των χαρακτηριστικών κάθε κατηγορίας (περιοχής ή θέσης) και ταξινομεί τον παίκτη με βάση την εγγύτητά του σε αυτά τα κέντρα. Η μέθοδος επιλέχθηκε για την εφαρμογή επειδή προσφέρει έναν ενδιάμεσο δρόμο μεταξύ των στατιστικών (μηχανιστικά) μοντέλων και των πιο εξελιγμένων μαθησιακών μοντέλων. Επιτρέπει την απλή οπτική ερμηνεία (μέσω PCA plot) και βοηθά να αναδειχθούν οι «φυσικές ομάδες» μέσα στα δεδομένα. Λειτουργεί καλά για datasets με εμφανή ομαδοποίηση, όπως το ποδόσφαιρο, όπου παίκτες της ίδιας θέσης παρουσιάζουν κοινά προφίλ χαρακτηριστικών.

Η ίδια μέθοδος Random Forest ενισχύθηκε περαιτέρω στην εφαρμογή με τη χρήση του αλγορίθμου SHAP, ο οποίος παρέχει επεξηγήσεις για το πώς κάθε χαρακτηριστικό συμβάλλει στην πρόβλεψη. Ο SHAP βασίζεται στη θεωρία παιγνίων και κατασκευάζει τιμές Shapley που δείχνουν τη συνεισφορά κάθε χαρακτηριστικού θετικά ή αρνητικά στο τελικό αποτέλεσμα. Επιλέχθηκε για την εφαρμογή ώστε να καλύψει την ανάγκη εξηγησιμότητας του μοντέλου – ένα κρίσιμο ζητούμενο όταν οι προβλέψεις χρησιμοποιούνται από μη ειδικούς, όπως προπονητές ή αναλυτές. Με την παρουσίαση των SHAP values σε μορφή bar plot, ο χρήστης μπορεί να κατανοήσει με διαφάνεια ποια χαρακτηριστικά (π.χ. “vision” ή “acceleration”) οδήγησαν τον αλγόριθμο στην τελική ταξινόμηση. Αυτή η δυνατότητα καθιστά τον Random Forest πιο «κατανοητό», ενισχύοντας την αποδοχή του μοντέλου και προσδίδοντας επιστημονική τεκμηρίωση στις αποφάσεις του συστήματος.

Ο αλγόριθμος Random Forest αποτελεί έναν ensemble ταξινομητή που συνδυάζει πολλά δέντρα απόφασης για να βελτιώσει τη γενίκευση και να μειώσει την πιθανότητα υπερπροσαρμογής. Λειτουργεί δημιουργώντας διαφορετικά υποσύνολα των δεδομένων και εκπαιδεύοντας σε αυτά ξεχωριστά δέντρα, ενώ η τελική πρόβλεψη βασίζεται στη «δημοκρατία» όλων των δέντρων. Στην εφαρμογή, η Random Forest χρησιμοποιήθηκε για την εύρεση περιοχής ή θέσης ενός παίκτη, καθώς προσφέρει εξαιρετική ακρίβεια χωρίς να απαιτεί προσεκτικό tuning. Η επιλογή της έγινε λόγω της ανθεκτικότητάς της σε πολυδιάστατα δεδομένα και της καλής απόδοσης ακόμη και σε περιπτώσεις με επικαλύψεις μεταξύ των κατηγοριών. Επιπλέον, η δυνατότητα να εξάγει πιθανότητες για κάθε θέση/περιοχή επέτρεψε την κατασκευή matrix plot, δίνοντας στον χρήστη μια πιο αναλυτική εικόνα της προβλεπτικής αβεβαιότητας. Αυτό προσφέρει ευελιξία και ακρίβεια σε πραγματικές συνθήκες ανάλυσης παικτών.

Το παράθυρο της επιλογής μεθόδου είναι ένα απλό παράθυρο για κάθε χρήστη. Αποτελείται από τρεις μεθόδους, οι οποίες δίπλα από την ετικέτα τους έχουν ένα radio button, όπου έχει και την λειτουργία να μην μπορεί ο χρήστης να επιλέξει παραπάνω από μια μέθοδο την φορά. Η πρώτη μέθοδο, η οποία είναι η K-NN χρειάζεται και μια τιμή λόγω της λειτουργίας της (υπάρχει εξήγηση στο παρακάτω κεφάλαιο). Κάτω από τις μεθόδους υπάρχει το κουμπί της εκτέλεσης όπου εκτελείται η επιλεγμένη

μέθοδος από πάνω. Τέλος, κάτω αριστερά υπάρχει το κουμπί πίσω όπου μας επιστρέφει πίσω στο αρχικό μενού.

2.7. Αρχιτεκτονική της εφαρμογής

Η αρχιτεκτονική της εφαρμογής σχεδιάστηκε με στόχο τη διατήρηση σαφήνειας στη ροή, απλότητας στη χρήση και επεκτασιμότητας στη λειτουργικότητα. Το περιβάλλον βασίζεται σε αρχές modular προγραμματισμού, όπου κάθε ενότητα της διεπαφής έχει ανεξάρτητη ευθύνη και μπορεί να τροποποιηθεί ή να επεκταθεί χωρίς να επηρεάσει τον υπόλοιπο κώδικα. Το κυρίως παράθυρο της εφαρμογής αποτελεί τον κεντρικό κόμβο αλληλεπίδρασης του χρήστη με το σύστημα και ενοποιεί όλα τα επιμέρους στάδια λειτουργίας.

Η σχεδίαση βασίζεται στην χρήση της βιβλιοθήκης Tkinter, η οποία παρέχει βασικές δυνατότητες για τη δημιουργία γραφικού περιβάλλοντος χρήστη. Το βασικό παράθυρο (root) διαμορφώνεται έτσι ώστε να παρουσιάζει στον χρήστη δύο κύριες επιλογές: να επιλέξει έναν ήδη καταχωρημένο παίκτη από την υπάρχουσα βάση δεδομένων ή να εισάγει έναν νέο παίκτη με χειροκίνητο τρόπο. Οι επιλογές αυτές πραγματοποιούνται μέσω ξεκάθαρων κουμπιών και οδηγούν σε αντίστοιχα διαμορφωμένα παράθυρα με ειδικά widgets για καταχώρηση στοιχείων ή εμφάνιση δεδομένων.

Το παράθυρο με την επιλογή παικτών εμφανίζει πίνακες οι οποίοι έχουν οργανωθεί μέσω της βιβλιοθήκης ttk.Treeview. Παρουσιάζονται τα στατιστικά του κάθε παίκτη, ενώ ο χρήστης μπορεί να επιλέξει έναν παίκτη και να συνεχίσει στην επιλογή της μεθόδου ανάλυσης. Παράλληλα, έχει τη δυνατότητα να πατήσει κουμπί που του δείχνει τα ήδη καταχωρημένα δεδομένα στη βάση, κάτι ιδιαίτερα χρήσιμο για την επαλήθευση των αποτελεσμάτων.

Η χειροκίνητη εισαγωγή παικτών πραγματοποιείται σε ειδικό παράθυρο, το οποίο περιλαμβάνει δυναμικά πεδία εισόδου (Entry και Listbox), όπου ο χρήστης μπορεί να συμπληρώσει τα χαρακτηριστικά του παίκτη. Προστέθηκαν μηχανισμοί ελέγχου εγκυρότητας, όπως έλεγχοι αριθμητικής τιμής, εύρους τιμών, και έλεγχοι για την επιλογή λίστας, ώστε να αποφευχθούν εσφαλμένες εγγραφές. Αφού γίνει η καταχώρηση, ο παίκτης αποθηκεύεται στη βάση, υπολογίζονται αυτόματα τα βασικά συνολικά χαρακτηριστικά του (pace, shooting, passing, dribbling, defending, physic, overall) και εισάγονται κι αυτά μαζί με τα επιμέρους.

Ο κώδικας ακολουθεί πρακτικές όπως χρήση callback συναρτήσεων μέσω της παραμέτρου command, για κάθε κουμπί ή ενέργεια του χρήστη. Αυτό προσφέρει μεγάλη ευελιξία και διαχωρισμό ευθυνών μεταξύ του GUI και της λογικής του προγράμματος. Όταν ο χρήστης επιλέγει μια μέθοδο ανάλυσης, το σύστημα τον οδηγεί στο αντίστοιχο παράθυρο μέσω της κατάλληλης συνάρτησης, και ανοίγεται νέο Toplevel παράθυρο με το αποτέλεσμα.

Το μενού επιλογής αλγορίθμων είναι ιδιαίτερα απλό, όπως ήταν και ο αρχικός στόχος του σχεδιασμού. Ο χρήστης καλείται να επιλέξει έναν από τρεις αλγόριθμους (k-NN, Random Forest, CNN) μέσω RadioButtons, ώστε να μην μπορεί να επιλεγεί παραπάνω από ένας. Για την περίπτωση του k-NN ζητείται και η επιλογή της παραμέτρου k με έλεγχο τιμής. Κάθε επιλογή ενεργοποιεί το αντίστοιχο παράθυρο με ανάλυση της πρόβλεψης.

Κάθε επιμέρους αλγόριθμος καλεί παράθυρο με διαγράμματα, περιγραφές και δυνατότητα αποθήκευσης των αποτελεσμάτων στη βάση. Τα διαγράμματα χρησιμοποιούν matplotlib και έχουν ενσωματωθεί στο GUI με FigureCanvasTkAgg. Σε αυτά εμφανίζονται οι παίκτες, τα κέντρα των

περιοχών ή θέσεων, καθώς και ο παίκτης προς ανάλυση. Υπάρχουν χρωματικοί διαχωρισμοί για να κατανοεί ο χρήστης εύκολα τις προβλέψεις.

Η εφαρμογή περιλαμβάνει επίσης μηχανισμούς ανάδρασης για τον χρήστη. Σε κάθε κρίσιμο σημείο, όπως αποτυχία σύνδεσης με βάση, ελλιπή δεδομένα ή πρόβλημα στην εισαγωγή, εμφανίζεται παράθυρο ειδοποίησης μέσω messagebox με κατάλληλο μήνυμα. Αυτό διασφαλίζει την απρόσκοπτη εμπειρία χρήστη και ελαχιστοποιεί τα σφάλματα.

Εν τέλει, η εφαρμογή σχεδιάστηκε με τέτοιο τρόπο ώστε να είναι επεκτάσιμη. Νέα κουμπιά, νέες λειτουργίες και επιπλέον αλγόριθμοι μπορούν να προστεθούν εύκολα χωρίς να διαταραχθεί η υπάρχουσα λειτουργικότητα. Αυτό επιτυγχάνεται με την υιοθέτηση modular προγραμματισμού, όπου κάθε λειτουργία βρίσκεται σε δική της ενότητα, ξεχωριστή από τον κορμό της διεπαφής.

Η σύνδεση με την βάση δεδομένων γίνεται με SQLite3 και τα δεδομένα φορτώνονται με Pandas DataFrames, κάτι που διευκολύνει σημαντικά τη διαχείριση, φιλτράρισμα και εμφάνιση των δεδομένων. Η βάση δεδομένων παραμένει τοπική για λόγους απλότητας, ταχύτητας και ανεξαρτησίας από σύνδεση στο διαδίκτυο, ωστόσο η αρχιτεκτονική της εφαρμογής επιτρέπει και τη μελλοντική σύνδεση με απομακρυσμένες βάσεις.

Εν κατακλείδι, η αρχιτεκτονική της εφαρμογής στηρίζεται σε έναν ξεκάθαρο και ευέλικτο πυρήνα, όπου η απλότητα χρήσης συνδυάζεται με την ακριβή λειτουργία των αλγορίθμων. Κάθε τμήμα της εφαρμογής λειτουργεί συμπληρωματικά προς τον στόχο της σωστής ανάλυσης και παρουσίασης των αποτελεσμάτων στον χρήστη.

Το κύριο παράθυρο της εφαρμογής λειτουργεί ως κεντρικός κόμβος πλοήγησης και διαχείρισης της αλληλεπίδρασης του χρήστη με το σύστημα. Ο σχεδιασμός του ακολουθεί την αρχή του γραφικού περιβάλλοντος μονής ροής (single-window GUI controller), όπου όλες οι βασικές ενέργειες του χρήστη εκκινούνται ή διεκπεραιώνονται από μία ενοποιημένη διεπαφή. Η δομή και τα layout βασίζονται σε:

- Tkinter (root) παράθυρο, το οποίο αποτελεί την κύρια εφαρμογή
- Πολλαπλά Frame widgets για διαχωρισμό της διεπαφής σε περιοχές για την κατανομή των ετικετών και των κουπιών
- Χρήση pack & grid για την οργάνωση στοιχείων

Στο κεντρικό παράθυρο περιλαμβάνονται κουμπιά για τις βασικές λειτουργίες:

- Ανάλυση Θέσης. Εκκινεί τη διαδικασία πρόβλεψης περιοχής και θέσης, ανοίγοντας νέο παράθυρο με plot.
- Καθαρισμός ή Πίσω. Χρησιμοποιούνται για διαχείριση της ροής ή εξόδου από το σύστημα.

Κάθε κουμπί έχει ενσωματωμένο callback (command=...) που καλεί την αντίστοιχη συνάρτηση.

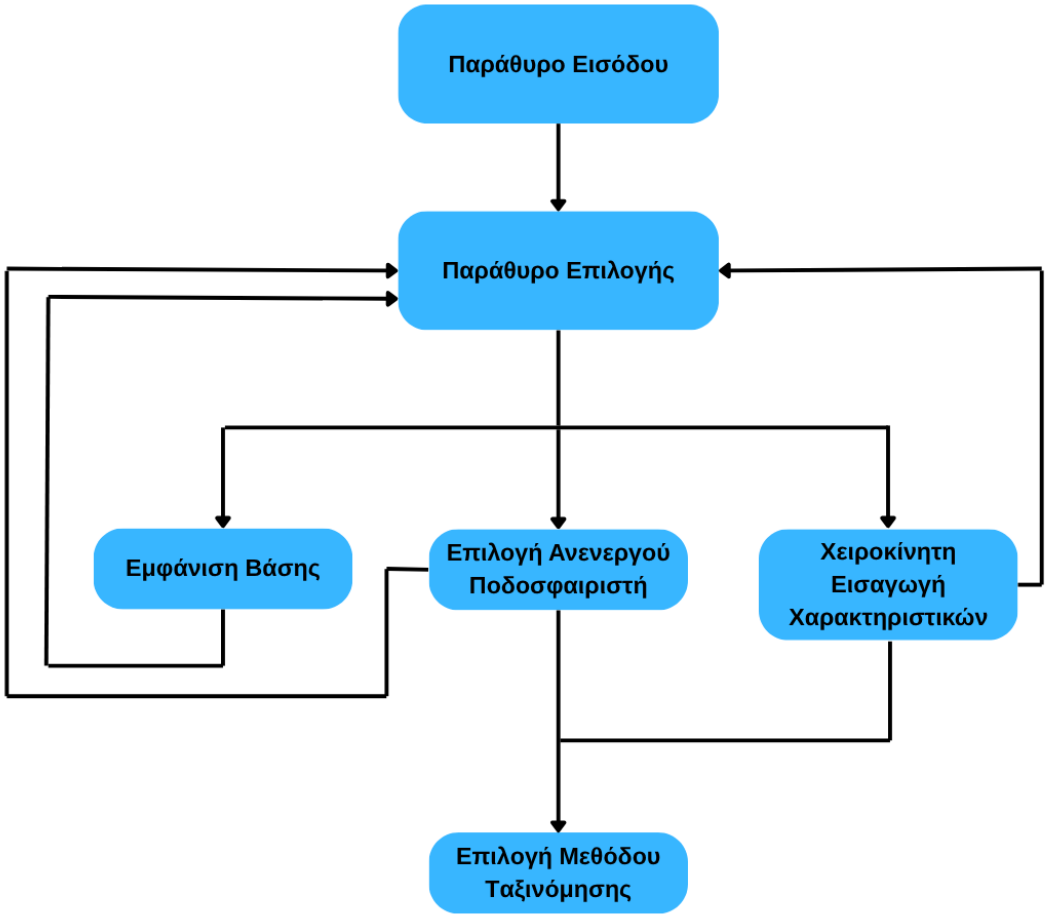
Όσον αφορά την σύνδεση με την βάση δεδομένων (Database Interface Layer), η μόνη ενέργεια που γίνεται είναι η ανάκτιση των απαραίτητων πινάκων στο πρόγραμμα και πολλαπλλα φιλτραρίσματα και επαληθεύσεις ύπαρξης στοιχείων. Η διασύνδεση είναι υλοποιημένη με κατάλληλη διαχείριση σφαλμάτων ώστε να ενημερώνεται ο χρήστης με messagebox.showerror() σε περίπτωση αστοχίας.

Για την λογική ανάλυση των παραθύρων χρησιμοποιείται ένα root, το οποίο συνεχώς αλλάζει σχήμα, τίτλο και σβήνετε το προηγούμενο περιεχόμενο, έτσι ώστε να εμφανιστεί το επόμενο. Για παράδειγμα όταν επιλέγουμε την χειροκίνητη εισαγωγή ενός παίκτη προφανώς το παράθυρο αλλάζει σχήμα και γίνεται πιο μεγάλο για να μπορέσουν όλα τα χαρακτηριστικά και τα κουμπιά να εμφανίζονται όλα σε

ένα παράθυρο. Η μοναδική εξαίρεση είναι όταν επιλέγουμε μέθοδο και εκτελούμε τους αλγορίθμους, καθώς εμφανίζονται καινούργια παράθυρα για να μπορούμε σε περίτωση που κάναμε κάποιο λάθος ή να επιλέξουμε έναν καινούργιο παίκτη να έχουμε αυτήν την δυνατότητα.

Με λίγα λόγια το κύριο παράθυρο ακολουθεί αρχές modular programming, καθώς:

- Διατηρεί καθαρό ρόλο. Επιλογή παικτών και στην συνέχεια κλήση μεθόδων ταξινόμησης/ανάλυσης
- Όλες οι υποστηρικτικές λειτουργίες υλοποιούνται σε ξεχωριστές συναρτήσεις εντός της απαραίτητης μεθόδου
- Υπάρχει δυνατότητα επέκτασης του παραθύροι με επιπλέον λειτουργίες χωίς να χρειάζεται να διαταχθεί η υπάρχουσα δομή
- Περιλαμβάνει μηχανισμούς επιβεβαίωσης/ενημέρωσης (μέσω extra εμφανιζόμενων παραθύρων τύπου messagebox), ώστε ο χρήστης να γνωρίζει τι ακριβώς συμβαίνει και να υπάρχουν σχεδόν μηδενικές πιθανότητες ο χρήστης να μην μπορέσει να κατανοήσει πλήρως την λειτουργία της εφαρμογής



Σχήμα 2.7 Διάγραμμα Ροής Βασικού Προγράμματος

Κεφάλαιο 3ο: Ανάλυση Αλγορίθμων Ταξινόμησης

3.1 Εισαγωγή

Η εποπτευόμενη (supervised) μάθηση αποτελεί σήμερα τον πυρήνα της ταξινόμησης δεδομένων, καθώς διευκολύνει τη μάθηση μιας συνάρτησης f:X→Y βάσει προσημασμένων δειγμάτων (xi,yi). Στην πράξη, ο στόχος είναι η γενίκευση, δηλαδή η ανάπτυξη μοντέλων με ικανότητα να προβλέπουν σωστά σε νέα, αόρατα δεδομένα. Για να επιτευχθεί αυτό, κεντρικές παράμετροι όπως ο συμβιβασμός bias–variance, η αποφυγή υπερπροσαρμογής (overfitting) και η εκτίμηση του σφάλματος με cross-validation αποτελούν βασικά εργαλεία.

Στο παρόν κεφάλαιο αναλύονται επιλεγμένοι αλγόριθμοι ταξινόμησης—k-NN, Random Forests, ελαφρά παραλλαγές CNN μαζί με μεθόδους μείωσης διαστάσεων (PCA) και τεχνικές ερμηνευσιμότητας (SHAP). Η επιλογή αυτών των τεχνικών βασίζεται σε θεωρητικές αρχές αλλά και εντοπισμένες αναλυτικές ανάγκες στον αθλητισμό συνδυάζοντας αξιοπιστία και ερμηνευσιμότητα, καθιστώντας το μοντέλο πρακτικά χρήσιμο.

Οι αλγόριθμοι αυτοί χρησιμοποιούνται σε:

- Ανάλυση απόδοσης παικτών και ομάδων.
- Πρόβλεψη θέσης ή ρόλου για έναν νέο ή υπό αξιολόγηση παίκτη.
- Σύγκριση με πρότυπα (π.χ. θρύλους του αθλήματος).
- Ανίχνευση ταλέντων μέσω clustering ή similarity scoring.

Στην παρούσα Π.Ε., υλοποιούνται αλγόριθμοι που συνδυάζουν κανονικοποίηση, PCA για μείωση διαστάσεων και μέτρηση απόστασης (NCC ή Euclidean distance) ώστε να εντοπιστεί η αγωνιστική περιοχή και θέση που ταιριάζει περισσότερο σε έναν παίκτη, βάσει των στατιστικών του.

3.2. Μέθοδος K-NN

Ο αλγόριθμος k-NN ανήκει στις παραδειγματοκεντρικές (instance-based) μεθόδους, καθώς η ταξινόμηση μιας νέας παρατήρησης βασίζεται στον υπολογισμό της απόστασης από δειγματικές περιπτώσεις και την επιλογή της πλειοψηφίας των k κοντινότερων ομοειδών. Χρήσιμες μετρικές απόστασης περιλαμβάνουν την Ευκλείδεια, την Manhattan, ή πιο εξειδικευμένες όπως το Minkowski βάσει της φύσης των χαρακτηριστικών. Οι κίνδυνοι του k-NN περιλαμβάνουν υψηλή ευαισθησία σε θόρυβο και διαφορετικές κλίμακες: η σωστή κλιμάκωση και επιλογή του k είναι κρίσιμη, όπως και η χρήση cross-validation για επιλογή του βέλτιστου k.

Από θεωρητική πλευρά, το κλασικό αποτέλεσμα των Cover & Hart (1967) δείχνει ότι ο k-NN δεν ξεπερνά σε σφάλμα διπλάσιο του Bayes-optimal για μεγάλες δειγματοληψίες, δίνοντας μια άμεση θεωρητική εγγύηση συνέπειας[3]. Στην εφαρμογή μας, όπου ταξινομούνται παίκτες βάσει πολυδιάστατων χαρακτηριστικών, το k-NN λειτουργεί ως baseline, με πλεονεκτήματα στην απλότητα, αλλά περιορισμούς στην κλίμακα και στα πολυδιάστατα δεδομένα. Η προσθήκη διαστάσεων μέσω PCA ή η χρήση SHAP για ερμηνευσιμότητα μπορεί να ενισχύσει το μοντέλο.

Επιπλέον, με την εκτέλεση k-NN πάνω σε PCA αναπαραστάσεις, αναδεικνύεται η δυνατότητα συνδυασμού απλών και προηγμένων τεχνικών διατηρώντας την ευελιξία και την επεξηγηματικότητα.

Βήματα λειτουργίας:

- Υπολογίζεται η απόσταση του νέου δείγματος από όλα τα υπάρχοντα σημεία στο dataset.
- Επιλέγονται οι k κοντινότεροι γείτονες.
- Το νέο δείγμα κατατάσσεται στην πιο συχνή κατηγορία (label) των γειτόνων αυτών.

Πλεονεκτήματα:

- Εύκολος στην κατανόηση και υλοποίηση.
- Δεν απαιτεί εκπαίδευση (είναι lazy learner).
- Ικανός να χειριστεί μη γραμμικά διαχωρίσιμα δεδομένα.

Μειονεκτήματα:

- Γίνεται αργός με μεγάλα σύνολα δεδομένων.
- Ευαίσθητος στο θόρυβο και στις ασυνεπείς ετικέτες.
- Η επιλογή του κατάλληλου k είναι κρίσιμη.

Στο ποδόσφαιρο αυτό μπορεί να βοηθήσει στην επίλυση προβλημάτων κατάταξης ενός νέου παίκτη σε ένα νέο σύνολο. Για παράδειγμα, σε αγωνιστικό επίπεδο μπορούμε να κατατάξουμε έναν παίκτη σε μια θέση χρησιμοποιώντας τα χαρακτηριστικά του συγκριτικά με τα χαρακτηριστικά άλλων παικτών που υπάρχουν στην βάση δεδομένων του προγράμματος και να βρούμε σε ποιον χώρο του γηπέδου θα δραστηριοποιείται καλύτερα. Επίσης, σε οικονομικό επίπεδο θα μπορούσαμε να συγκρίνουμε ως προς τον μισθό, την χρηματιστηριακή αξία και άλλων παραμέτρων έτσι ώστε να εξεταστεί οικονομικά μια πρόταση είτε ανανέωσης συμβολαίου είτε ακόμα και μεταγραφής.

Στην παρούσα Π.Ε. ο αλγόριθμος λειτουργεί σε δύο φάσεις. Στην πρώτη φάση αναλύει τον επιλεγμένο ανενεργό παίκτη, που είτε έχει διαλέξει είτε έχει εισάγει χειροκίνητα, και βρίσκει την περιοχή που θα αγωνιζόταν ο παίκτης. Οι περιοχές είναι χωρισμένες στην άμυνα, στο κέντρο και στην επίθεση. Για να πετύχουμε αυτήν την αναζήτηση το πρόγραμμα λειτουργεί ως εξής:

- Δηλώνονται από πριν σε ένα dataframe οι ανενεργοί παίκτες. Διότι η μέθοδος όπως έχουμε αναφέρει και πριν κρατάει ως παράμετρο μόνο το αναγνωριστικό του παίκτη προς ανάλυση. Με αυτόν τον τρόπο διασφαλίζουμε ότι θα βρούμε τα χαρακτηριστικά του.
- Δηλώνονται οι στήλες (χαρακτηριστικά) που θα χρησιμοποιηθούν για την ανάλυση.
- Η μέθοδος χρησιμοποιεί τρεις παραμέτρους. Μια είναι η τιμή του k για την σύγκριση με τον συγκεκριμένο αριθμό κοντινότερων γειτόνων. Μια είναι η τιμή του αναγνωριστικού του παίκτη. Τέλος, το dataframe με τους ανενεργούς παίκτες για την εύρεση των χαρακτηριστικών του παίκτη υπό ανάλυση.
- Στην αρχή της μεθόδου υπάρχουν κάποιο έλεγχοι ως προς την βάση για ομαλή λήψη στοιχείων ενεργών και ανενεργών παικτών.
- Μετά γίνεται έλεγχος αν τα στοιχεία του παίκτη προς ανάλυση είναι πλήρη.
- Γίνεται έλεγχος της τιμης k αν υπάρχει ο αριθμός τόσων κοντινότερων γειτόνων.
- Στην συνέχεια γίνεται ανάλυση των χαρακτηριστικών των ενεργών ποδοσφαιριστών.
- Μετά γίνεται η σύγκριση του ανανεργού παίκτη με τους ενεργούς.
- Βρίσκεται το αποτέλεσμα.
- Εμφανίζεται το αποτέλεσμα σε ένα παράθυρο με γήπεδο οπτικοποίησης της OPTA με χρωματισμένη την περιοχή δράσης του παίκτη.

Στην δεύτερη φάση μέσω εύρεσης της περιοχής δράσης του παίκτη, βρίσκουμε την θέση δράσης του. Αυτό γίνεται υιοθετώντας την περιοχή που έχει βρεθεί από την προηγούμενη μέθοδο και στην συνέχεια εκτελώντας ξανά την ίδια διαδικασία με την μέθοδο k-NN. Τα βήματα της επόμενης μεθόδου εύρεσης θέσης είναι ως εξής:

- Πάλι περνάμε ως παραμέτρους την περιοχή δράσης, τα χαρακτηριστικά του ανενεργού παίκτη (που τα βρήκαμε πριν), το αναγνωριστικό του παίκτη και την βάση με τους παίκτες
- Στην αρχή της μεθόδου υπάρχουν κάποιοι έλεγχοι ως προς την βάση για ομαλή λήψη στοιχείων ενεργών παικτών.
- Φιλτράρουμε τα δεδομένα, έτσι ώστε να έχουμε μόνο τους ενεργούς παίκτες που αγωνίζονται στην ίδια περιοχή με τον παίκτη προς ανάλυση.
- Χωρίζουμε τα χαρακτηριστικά και τις θέσεις των ενεργών παικτών
- Εκτελούμε την μέθοδο ανάλυσης των χαρακτηριστικών των ενεργών παικτών
- Λαμβάνουμε υπόψιν την τιμή k
- Βρίσκουμε σε ποια θέση αγωνίζονται οι περισσότεροι k κοντινότεροι γείτονες με βάση τον παίκτη που αναλύουμε
- Σε παράθυρο εμφανίζουμε πάλι σε γήπεδο ΟΡΤΑ την θέση που θα αγωνιζόταν ο παίκτης με σκιασμένο χρώμα εντός του γηπέδου
- Τέλος, υπάρχει ένα κουμπί που μπορεί ο χρήστης να εισάγει την περιοχή και την θέση της πρόβλεψης στον ανενεργό παίκτη και αυτές οι τιμές να αποθηκευτούν στην βάση.

Συμπερασματικά, η εφαρμογή της μεθόδου k-Nearest Neighbors (k-NN) στην παρούσα εργασία αποδείχθηκε ιδιαίτερα χρήσιμη για την ανάλυση ομοιότητας μεταξύ ποδοσφαιριστών με βάση τα στατιστικά τους χαρακτηριστικά. Η δυνατότητα του αλγορίθμου να εντοπίζει γειτονικές περιπτώσεις σε πολυδιάστατο χώρο επέτρεψε τη σύγκριση παικτών με βάση την αγωνιστική συμπεριφορά τους, οδηγώντας σε ρεαλιστικά συμπεράσματα σχετικά με πιθανές θέσεις ή και ρόλους που μπορούν να καλύψουν. Αν και ο αλγόριθμος παρουσιάζει ορισμένα μειονεκτήματα, όπως η ευαισθησία στο μέγεθος του συνόλου δεδομένων και στην επιλογή του k, στη συγκεκριμένη εφαρμογή, με κατάλληλη προεπεξεργασία και κανονικοποίηση, παρείχε αξιόπιστα αποτελέσματα που ενίσχυσαν τη χρηστικότητα του συστήματος υποστήριξης απόφασης για προπονητές ή αναλυτές ποδοσφαίρου. Το αποτέλεσμα της ανάλυσης έχει να κάνει με το σύνολο των δεδομένων που προϋπάρχουν στην βάση δεδομένων. Εκεί μπορεί να διαπιστώσει κανείς ότι κάποιες θέσεις έχουν πιο πολλούς παίκτες που αγωνίζονται στις συγκεκριμένες θέσεις. Ο αλγόριθμος είναι αρκετά κατατοπιστικός και παρουσιάζει σπουδαία αποτελέσματα, ειδικά αν ο χρήστης γνωρίζει τον ανενεργό παίκτη και κατανοεί πλήρη την ευαισθησία στην ξεχωριστή δήλωση κάθε τιμής του k σε κάθε επανάληψη.

Στην συγκεκριμένη Π.Ε. αξίζει να σημειωθεί ότι με βάση αυτόν τον αλγόριθμο είχαμε κάποια ενδιαφέρουσα αποτελέσματα. Υπήρξαν παίκτες που θα αγωνίζονταν σε περιοχές αναμενόμενες στο κοινό, διότι σε αυτές αγωνίζονταν κατά την εποχή τους. Υπήρχαν όμως και ποδοσφαιριστές που με βάση αυτόν τον αλγόρυθμο θα τους ταίριαζε μια άλλη θέση όχι πολύ μακριά από αυτήν που αγωνίστηκαν. Για παράδειγμα, ένας παίκτης που θα μπορούσε να παίζει στα πλάγια, και ο αλγόριθμος έδειχνε ότι θα ήταν πιο πρακτικός ως κεντρικός μέσος. Ή ένας αμυντικός λόγω της άριστης δημιουργικότητας του θα αγωνιζόταν σε θέση που έχει πιο ενεργό ρόλο στην συμβολή του παιχνιδιού.

3.3. Μέθοδος CNN

Οι Συνελικτικοί Νευρωνικοί Δικτύων (Convolutional Neural Networks - CNN) αποτελούν μια ισχυρή κατηγορία αλγορίθμων μηχανικής μάθησης. Χάρη στη δομή τους, που στηρίζεται σε φίλτρα (kernels) και πολλαπλά επίπεδα (convolutional, pooling, fully connected), τα CNN έχουν τη δυνατότητα να ανιχνεύουν πολύπλοκα πρότυπα μέσα από ακατέργαστα δεδομένα και να μαθαίνουν ιεραρχικές αναπαραστάσεις. Η χρήση τους επεκτείνεται πλέον και στην επεξεργασία αθλητικών δεδομένων, όπου μπορούν να εντοπίζουν στρατηγικά μοτίβα κίνησης, τακτικές ή στυλ παιχνιδιού παικτών, ιδίως όταν υπάρχουν χρονικές, οπτικές ή στατιστικές καταγραφές.

Αντίθετα, ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise), που είχε αρχικά επιλχθεί ως μέθοδος, ανήκει στην κατηγορία των αλγορίθμων ομαδοποίησης (clustering) και χρησιμοποιείται για την αναγνώριση ομάδων δεδομένων βάσει πυκνότητας. Είναι ιδιαίτερα αποτελεσματικός όταν τα δεδομένα δεν σχηματίζουν σφαιρικές συστάδες, καθώς δεν απαιτεί τον καθορισμό του αριθμού των clusters εκ των προτέρων και μπορεί να αγνοήσει θόρυβο. Σε αντίθεση με τα CNN, τα οποία μαθαίνουν παραμέτρους μέσα από επαναληπτική εκπαίδευση. Ο DBSCAN βασίζεται σε γεωμετρικά κριτήρια εγγύτητας και πυκνότητας, χωρίς να απαιτεί εποπτευόμενα δεδομένα ή μεγάλα σετ εκπαίδευσης.

Η βασική διαφορά μεταξύ των δύο μεθόδων έγκειται στο ότι ο CNN είναι αλγόριθμος μάθησης, ικανός να γενικεύσει και να προβλέψει, ενώ ο DBSCAN είναι μη εποπτευόμενη μέθοδος ομαδοποίησης, που χρησιμοποιείται για την εξερεύνηση της δομής ενός συνόλου δεδομένων χωρίς προκαθορισμένες ετικέτες. Η μέθοδος DBSCAN είναι για ταξινόμηση ενός ευρή συνόλο, ενώ στο πρόγραμμα μας θέλουμε να ταξινομήσουμε έναν ποδοσφαιριστή (ανενεργό) σε ένα σύνολο από ήδη ταξινομημένους παίκτες (ενεργοί).

Τα CNNs εκμεταλλεύονται τις τοπικές συσχετίσεις μέσω συνελιστικών φίλτρων και pooling, που περιορίζουν το πλήθος παραμέτρων και ενισχύουν τη γενίκευση, ιδιαίτερα σε δομημένα δεδομένα όπως εικόνες ή πίνακες. Η κλασική εργασία των LeCun et al. (1998) παρουσίασε το LeNet, μοντέλο-ορόσημο για αναγνώριση χαρακτήρων, με απόδοση και κοστολογικά βελτιστοποιημένη δομή[8].

Στο πλαίσιο της ταξινόμησης χαρακτηριστικών παικτών, μπορείς να χρησιμοποιήσεις πιο απλά "συνελικτικά σχήματα" δηλαδή, επίπεδα dense ή convolution-like λειτουργίες πάνω σε πίνακες χαρακτηριστικών, για να αντλήσεις θεματικές αλληλεπίδρασης μεταξύ αδυναμιών, θέσης και ικανότητας. Αυτό δημιουργεί πιο ιεραρχικές, σύνθετες αναπαραστάσεις, συμβάλλοντας στην πιο ακριβή και επεξηγηματική ταξινόμηση.

Στην παρούσα Π.Ε. ο αλγόριθμος CNN λειτουργεί σε δύο φάσεις. Στην πρώτη όπως και στον k-NN αναζητεί την περιοχή δράσης του παίκτη τον οποίο έχει επιλέξει ο χρήστης προς ανάλυση. Για να επιτευχθεί αυτή η ταξινόμηση χρειάστηκαν τα εξής βήματα:

- Αρχικά, πριν την αρχή της πρώτης μεθόδου δηλώθηκαν οι στήλες (χαρακτηριστικά) που θα συμμετέχουν στην ανάλυση.
- Στην αρχή του προγράμματος φορτώθηκαν ο πίνακας με τους ενεργούς και ο πίνακας με τους ανενεργούς παίκτες από την βάση.
- Στην συνέχεια, έγινε ανάλυση των χαρακτηριστικών των παικτών ξεχωρίζοντας φυσικά την περιοχή ως ξεχωριστό χαρακτηριστικό.
- Μετά βρέθηκαν τα κέντρα στον πολυχώρο των τριών περιοχών (άμυνας, κέντρου και επίθεσης).
- Στην συνέχεια, υπολογίστηκε η ευκλείδια απόσταση μεταξύ των κέντρων και του σημείου που βρισκόταν ο παίκτης προς ανάλυση.

- Ο παίκτης έπειτα ταξινομήθηκε με βάση το πιο κοντινό κέντρο σε αυτόν.
- Τέλος, χάρη στο διάγραμμα PCA μπορέσαμε να εμφανίσουμε τον χώρο της ανάλυσης με τους παίκτες να είναι με κύκλους (τρια διαφορετικά χρώματα για τις τρεις διαφορετικές περιοχές), τα κέντρα να είναι με X (ίδιο χρώμα με τους παίκτες της περιοχής τους) και τον παίκτη προς ανάλυση με μαύρο X για να ξεχωρίζει από το σύνολο των υπολοίπων.

Στην δεύτερη φάση μέσω εύρεσης της περιοχής δράσης του παίκτη, βρίσκουμε την θέση δράσης του. Αυτό γίνεται υιοθετώντας την περιοχή που έχει βρεθεί από την προηγούμενη μέθοδο και στην συνέχεια εκτελώντας ξανά την ίδια διαδικασία με την μέθοδο CNN. Τα βήματα της επόμενης μεθόδου εύρεσης θέσης είναι ως εξής:

- Αρχικά, πριν την εκκίνηση της μεθόδους εύρεσης παίκτη δηλώθηκαν οι θέσεις ανά περιοχή και τα χρώματα που θα εμφανίζονται στο διάγραμμα ανά θέση.
- Στην αρχή του προγράμματος φορτώθηκαν ο πίνακας με τους ενεργούς και ο πίνακας με τους ανενεργούς παίκτες από την βάση.
- Στην συνέχεια, έγινε ανάλυση των χαρακτηριστικών των παικτών ξεχωρίζοντας φυσικά την θέση ως ξεχωριστό χαρακτηριστικό.
- Μετά βρέθηκαν τα κέντρα των θέσεων με βάση την περιοχή που βρέθηκε στον προηγούμενο αλγόριθμο.
- Στην συνέχεια, υπολογίστηκε η ευκλείδια απόσταση μεταξύ των κέντρων και του σημείου που βρισκόταν ο παίκτης προς ανάλυση.
- Ο παίκτης έπειτα ταξινομήθηκε με βάση το πιο κοντινό κέντρο σε αυτόν.
- Χάρη στο διάγραμμα PCA μπορέσαμε να εμφανίσουμε τον χώρο της ανάλυσης με τους παίκτες να είναι με κύκλους (διαφορετικά χρώματα άνα θέση), τα κέντρα να είναι με X (ίδιο χρώμα με τους παίκτες της θέσης τους) και τον παίκτη προς ανάλυση με μαύρο X για να ξεχωρίζει από το σύνολο των υπολοίπων.
- Τέλος, υπάρχει ένα κουμπί που μπορεί ο παίκτης να εισάγει την περιοχή και την θέση της πρόβλεψης στον ανενεργό παίκτη και αυτές οι τιμές να αποθηκευτούν στην βάση.

Συμπερασματικά, η μέθοδος CNN, παρόλο που εφαρμόστηκε στην παρούσα Π.Ε. με έναν απλοποιημένο τρόπο χωρίς χρήση νευρωνικού δικτύου στην πλήρη του μορφή, αποδείχθηκε ιδιαίτερα αποτελεσματική για την ανάλυση και ταξινόμηση ανενεργών παικτών σε υπάρχουσες κατηγορίες ενεργών ποδοσφαιριστών. Μέσω της προσέγγισης με κέντρα περιοχών και θέσεων στον πολυδιάστατο χώρο των χαρακτηριστικών, η μέθοδος μπόρεσε να εντοπίσει με ακρίβεια την πλησιέστερη αγωνιστική ταυτότητα για κάθε παίκτη. Η απεικόνιση με PCA ενίσχυσε την κατανόηση και διαφάνεια των αποτελεσμάτων, προσφέροντας και οπτική επιβεβαίωση των ταξινομήσεων. Αν και η μεθοδολογία δεν περιλαμβάνει συνελικτικά επίπεδα όπως ένα πλήρες CNN, το όνομα χρησιμοποιείται συμβολικά για να περιγράψει την συστηματική προσέγγιση εντοπισμού συσχετίσεων, που θυμίζει την λειτουργία τέτοιων αλγορίθμων. Το τελικό αποτέλεσμα είναι ένα εργαλείο που βοηθά αξιόπιστα στην πρόβλεψη αγωνιστικού ρόλου ενός παίκτη, ακόμα κι αν λείπουν σύγχρονα δεδομένα απόδοσης.

Στην συγκεκριμένη Π.Ε. αξίζει να σημειωθεί ότι με βάση αυτόν τον αλγόριθμο είχαμε κάποια ενδιαφέρουσα αποτελέσματα. Είδαμε πόσο κοντά ένας παίκτης μπορεί να είναι σε δύο κέντρα περιοχών. Είδαμε επίσης, παίκτες να είναι πολύ μακρία από το σύνολο (θα μπορούσαμε να τους χαρακτηρίσουμε ως outliers αν ανήκαν στους ενεργούς παίκτες) και η κατάταξη τους σε κέντρο να είναι εμφανή. Όσον αφορά τις θέσεις, είχαμε πολύ ενδιαφέρον συμπεράσματα. Αρχικά, είδαμε κάποιοι παίκτες να είναι τόσο κοντά σε 2 ή και στα 3 κέντρα και η ευκλείδια απόσταση να τους κατατάξει πιθανόν σε μια θέση που στην πραγματικότητα να μην μπορούσαν να ανταποκριθούν. Επιπλέον, πολύ ενδιαφέρον ήταν και η απόσταση μεταξύ δύο κέντρων, η οποία πραγματικά ήταν μικρή, όπως για παράδειγμα οι ακραίοι αμυντικοί (στην ανάλυση των θέσεων της άμυνας) ή οι

ακραίοι επιθετικοί (στην ανάλυση των θέσεων της επίθεσης). Πριν μερικές δεκαετίες οι πλάγιες θέσεις είχαν πολύ διακριτά χαρακτηριστικά. Πλέον στις μέρες μας ένας παίκτης, ειδικά στις θέσεις των ακραίων επιθετικών δεν μπορεί να ξεχωρίσει. Μπορεί να βολεύει τον προπονητή, το σύστημα ή τον τρόπο παιχνιδιού να εναλλάσονται σε αυτές τις δύο θέσεις. Τέλος, είναι πάρα πολλά τα συμπεράσματα που μπορεί να βγάλει κανείς παρατηρόντας απλά μερικά αποτελέσματα ενός γραφήματος με την βοήθεια του συγκεκριμένου αλγορίθμου.

3.4. Μέθοδος Random Tree σε συνδιασμό με SHAP

Η μέθοδος Random Forest, γνωστή και ως Random Tree, αποτελεί έναν από τους πιο ισχυρούς και διαδεδομένους αλγορίθμους εποπτευόμενης μάθησης. Πρόκειται για ένα σύνολο από πολλαπλά δέντρα απόφασης (decision trees), τα οποία συνεργάζονται για να βελτιώσουν την ακρίβεια της πρόβλεψης. Η ιδέα πίσω από τη μέθοδο αυτή είναι ότι, αντί να βασιζόμαστε σε μία μόνο απόφαση από ένα μεμονωμένο δέντρο, λαμβάνουμε αποφάσεις από πολλά δέντρα και επιλέγουμε την πιο συχνή ή μέση τιμή (ανάλογα αν πρόκειται για ταξινόμηση ή παλινδρόμηση). Αυτή η τεχνική συναίνεσης μειώνει την υπερπροσαρμογή (overfitting) και αυξάνει τη γενίκευση του μοντέλου. Στην παρούσα εφαρμογή, ο Random Forest χρησιμοποιείται για την πρόβλεψη της αγωνιστικής περιοχής ή της θέσης ενός ποδοσφαιριστή, με βάση στατιστικά χαρακτηριστικά από την καριέρα του.

Τα Random Forests είναι υποσύνολο των ensemble learning, και λειτουργούν με τη λογική της συλλογικής ιδιοκτησίας: δημιουργούν πολλά δένδρα απόφασης μέσω bagging (bootstrap aggregation) και επιπλέον εισάγουν τυχαιότητα στην επιλογή χαρακτηριστικών σε κάθε κόμβο. Το αποτέλεσμα είναι σταθερή απόδοση και σημαντική μείωση της διακύμανσης (variance), ιδιαίτερα σε σύνολα δεδομένων με θόρυβο ή υψηλή ευαισθησία.

Στην παρούσα εφαρμογή, τα Random Forests προσφέρουν σαφή υπεροχή, καθώς μπορούν να αντιμετωπίσουν δεδομένα με θόρυβο, υψηλή διάσταση και να αξιολογήσουν τη σημαντικότητα κάθε χαρακτηριστικού (feature importance), δίνοντας πρακτικά εργαλεία στον χρήστη για να κατανοήσει ποια χαρακτηριστικά καθορίζουν τη θέση ή τον ρόλο του παίκτη.

Η SHAP χρησιμοποιήθηκε στην παρούσα Π.Ε. ως μέσο εξήγησης των προβλέψεων του αλγορίθμου Random Forest, προκειμένου να αναδειχθεί η σημασία και το βάρος κάθε χαρακτηριστικού στην τελική πρόβλεψη της θέσης ή περιοχής ενός ποδοσφαιριστή. Δεδομένου ότι τα μοντέλα τύπου Random Tree θεωρούνται «μαύρα κουτιά» (black boxes) λόγω της εσωτερικής πολυπλοκότητάς τους, η SHAP δίνει τη δυνατότητα στον χρήστη να κατανοήσει γιατί ένας παίκτης ταξινομήθηκε σε μια συγκεκριμένη θέση. Μέσω των SHAP plots, ο χρήστης μπορεί να δει ποια χαρακτηριστικά έσπρωξαν την πρόβλεψη προς μια συγκεκριμένη κατηγορία και ποια έδρασαν αντίθετα. Έτσι, η SHAP δεν προσφέρει μόνο αιτιολόγηση των αποτελεσμάτων, αλλά προσδίδει και μεγαλύτερη εμπιστοσύνη στη διαδικασία λήψης απόφασης του συστήματος. Στην πράξη, αυτό καθιστά την εφαρμογή πιο χρήσιμη και αποδεκτή από προπονητές, αναλυτές ή ακόμα και ερευνητές.

Στην παρούσα Π.Ε. ο αλγόριθμος Random Forest σε συνδιασμό με τον αλγόριθμο SHAP λειτουργεί σε δύο φάσεις. Στην πρώτη όπως και στον k-NN και το CNN αναζητεί την περιοχή δράσης του παίκτη τον οποίο έχει επιλέξει ο χρήστης προς ανάλυση. Για να επιτευχθεί αυτή η ταξινόμηση χρειάστηκαν τα εξής βήματα:

- Αρχικά, πριν την αρχή της πρώτη μεθόδου δηλώθηκαν οι στήλες (χαρακτηριστικά) που θα συμμετέχουν στην ανάλυση.
- Στην αρχή του προγράμματος φορτώθηκαν ο πίνακας με τους ενεργούς και ο πίνακας με τους ανενεργούς παίκτες από την βάση.
- Στην συνέχεια, έγινε ανάλυση των χαρακτηριστικών των παικτών ξεχωρίζοντας φυσικά την περιοχή ως ξεχωριστό χαρακτηριστικό.
- Έπειτα έγινε προετοιμασία των δεδομένων πριν την εκπαίδευση του αλγορίθμου Random Forest.
- Μετά βρέθηκε η πρόβλεψη της περιοχής με βάση το αποτέλεσμα της ανάλυσης του Random Forest.
- Στην συνέχεια, κάναμε επεξήγηση των τιμών SHAP που αναζητούμε για το σύνολο των δεδομένων που έχουμε στην κατοχή μας.
- Παίρνουμε τις τιμές SHAP για τον παίκτη προς ανάλυση.
- Φορτώνουμε τις τιμές SHAP σε ένα Dataframe.
- Κανουμε αναπαράσταση των αποτελεσμάτων χρησιμοποιώντας διάγραμμα με μπάρες και εμφανίζοντας όλα τα χαρακτηριστικά του παίκτη με την σειρά επίδρασης στην εύρεση της περιοχής του παίκτη.

Στην δεύτερη φάση μέσω εύρεσης της περιοχής δράσης του παίκτη, βρίσκουμε την θέση δράσης του. Αυτό γίνεται υιοθετώντας την περιοχή που έχει βρεθεί από την προηγούμενη μέθοδο και στην συνέχεια εκτελώντας ξανά την ίδια διαδικασία μόνο με την μέθοδο Random Forest αυτήν την φορά. Τα βήματα της επόμενης μεθόδου εύρεσης θέσης είναι ως εξής:

- Αρχικά, πριν την εκκίνηση της μεθόδου εύρεσης παίκτη δηλώθηκαν οι θέσεις ανά περιοχή.
- Στην αρχή του προγράμματος φορτώθηκαν ο πίνακας με τους ενεργούς και ο πίνακας με τους ανενεργούς παίκτες από την βάση.
- Στην συνέχεια, έγινε ανάλυση των χαρακτηριστικών των παικτών ξεχωρίζοντας φυσικά την θέση ως ξεχωριστό χαρακτηριστικό.
- Έπειτα έγινε προετοιμασία των δεδομένων πριν την εκπαίδευση του αλγορίθμου Random Forest.
- Μετά βρέθηκε η πρόβλεψη της περιοχής και η πιθανότητα με βάση το αποτέλεσμα της ανάλυσης του Random Forest.
- Σε Dataframe φορτώσαμε τα δεδομένα της πρόβλεψης και της πιθανότητας
- Με την χρήση matrix εμφανίσαμε τις θέσεις και την πιθανότητα του παίκτη να παίξει στην κάθε μια από αυτήν (όλες οι θέσεις είναι με βάση την περιοχή που βρέθηκε να ανήκει ο παίκτης)
- Τέλος, υπάρχει ένα κουμπί που μπορεί ο παίκτης να εισάγει την περιοχή και την θέση της πρόβλεψης στον ανενεργό παίκτη και αυτές οι τιμές να αποθηκευτούν στην βάση.

Ο συνδυασμός της μεθόδου Random Forest με τον αλγόριθμο SHAP αποτελεί ένα ιδιαίτερα ισχυρό εργαλείο για εφαρμογές που απαιτούν τόσο ακρίβεια στην πρόβλεψη, όσο και ερμηνευσιμότητα των αποτελεσμάτων. Η Random Forest εξασφαλίζει υψηλή αξιοπιστία και σταθερότητα στις προβλέψεις, εκμεταλλευόμενη τη δύναμη της συλλογικής απόφασης πολλαπλών δέντρων, ενώ η SHAP αποκαλύπτει τον ρόλο κάθε χαρακτηριστικού στην τελική πρόβλεψη, προσφέροντας διαφάνεια σε ένα κατά τα άλλα "μαύρο κουτί" μοντέλο. Οι δύο μέθοδοι σε συνδυασμό προσφέρουν έναν ισορροπημένο μηχανισμό, όπου η υψηλή απόδοση συνοδεύεται από αιτιολόγηση, στοιχείο απαραίτητο σε εφαρμογές που στοχεύουν στη λήψη αποφάσεων βασισμένων σε δεδομένα.

Στην παρούσα Π.Ε., οι μέθοδοι Random Forest και SHAP χρησιμοποιήθηκαν συνδυαστικά για την ταξινόμηση ανενεργών ποδοσφαιριστών σε αγωνιστική περιοχή και συγκεκριμένη θέση, βάσει στατιστικών χαρακτηριστικών. Ο αλγόριθμος Random Forest χρησιμοποιήθηκε για την πρόβλεψη της

κατηγορίας, με στόχο την εύρεση της βέλτιστης τακτικής ταξινόμησης ενός παίκτη, αξιοποιώντας πλήρως τα ιστορικά δεδομένα. Από την άλλη, ο SHAP ενσωματώθηκε για να παρέχει αιτιολόγηση της πρόβλεψης, εξηγώντας στον χρήστη ποια χαρακτηριστικά οδήγησαν σε ποιο αποτέλεσμα. Μέσα από διαγράμματα SHAP και πίνακες πιθανοτήτων, η εφαρμογή προσφέρει στον τελικό χρήστη όχι μόνο το αποτέλεσμα, αλλά και το σκεπτικό πίσω από αυτό, καθιστώντας την λύση πλήρως λειτουργική, επεξηγηματική και κατάλληλη για πρακτική αξιοποίηση από αναλυτές, προπονητές και ερευνητές του ποδοσφαίρου.

Αξίζει να σημειωθεί ότι αρχικά η πρώτη ιδέα ήταν αναπαράσταση και ταξινόμηση μέσω Random Forest. Στην συνέχεια αυτό φάνηκε να μην δουλεύει καθώς ο τεράστιος όγκος χαρακτηριστικών και η μεγάλη αναζήτηση τόσο της περιοχής όσο και των θέσεων καθιστούσε δύσκολη την περιγραφή του αποτελέσματος της ανάλυσης στον χρήστη. Έπειτα από αρκετή μελέτη και με την βοήθεια της μεθόδου SHAP, ήταν πλέον εύκολο να δώσουμε στον χρήστη να μια αναπαράσταση, η οποία θα τον βοηθήσει και στην κατανόηση του αποτελέσματος του αλγορίθμου αλλά και να οδηγηθεί εύκολα στο συμπέρασμα της σύγκρισης. Στην πρώτη αναπαράσταση μέσω γραφήματος δένδρων ήταν αρκετά δυσανάγνωστο το δένδρο καθώς κάθε κόμβος έκανε σύγκριση 5 χαρακτηριστικών και άνω μερικές φορές με αποτέλεσμα να μην μπορεί με το μάτι να διακρίνεις ποια διαδρομή ακολούθησε ο παίκτης που έχεις επιλέξει για ανάλυση. Με την χρήση της SHAP και την βοήθεια των Dataframe καταφέραμε να κρατήσουμε όλες τις καθοριστικές τιμές που έπαιξαν ρόλο στην ταξινόμηση του παίκτη σε περιοχή. Αναπαραστόντας αυτές τις τιμές σε διάγραμμα με μπάρες όπου στο πάνω μέρος είναι οι αρνητικές τιμές και στο κάτω οι θετικές τιμές, μπορεί ο κάθε χρήστης να διακρίνει τα χαρακτηριστικά που έπαιξαν τον καθοριστικό ρόλο στην ταξινόμηση του παίκτη σε περιοχή. Στην συνέχεια μετά από την εύρεση της περιοχής δράσης του παίκτη ήταν εύκολο σχετικά να βρεθεί η θέση του. Χρησιμοποιώντας αυτήν την φορά ένα MATRIX διάγραμμα για αναπαράσταση της πιθανής θέσης του παίκτη, χρησιμοποιήθηκε η ποσοστιαία αναπαράσταση της πιθανότητας του παίκτη να αγωνιστεί σε κάθε θέση που υπάρχει στην περιοχή δράσης του. Έτσι ο χρήστης μπορεί με βάση ποια χαρακτηριστικά έπαιξαν μεγαλύτερο ρόλο πριν και με τις πιθανότητες καταλληλότητας για κάθε θέση στην περιοχή δράσης να οδηγηθεί εύκολα σε κάποιο συμπέρασμα.

3.4 Ανάλυση Κυρίων Συνιστωσών (PCA)

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis – PCA) αποτελεί μία από τις σημαντικότερες τεχνικές μείωσης διαστασιμότητας στα δεδομένα. Στόχος της είναι η αναπαράσταση ενός πολυδιάστατου συνόλου δεδομένων σε λιγότερες διαστάσεις, διατηρώντας όσο το δυνατόν περισσότερη από τη συνολική διασπορά (variance). Η μέθοδος εισήχθη από τον Pearson και συστηματοποιήθηκε από τον Hotelling, ενώ έχει έκτοτε εδραιωθεί ως βασικό εργαλείο στατιστικής και μηχανικής μάθησης. Ουσιαστικά, το PCA μετασχηματίζει τα αρχικά χαρακτηριστικά σε νέες γραμμικές συνιστώσες (principal components), οι οποίες είναι ασυσχέτιστες μεταξύ τους και κατατάσσονται με βάση τη σπουδαιότητά τους.

Η χρησιμότητα του PCA εκτείνεται σε πολλούς τομείς της επιστήμης των δεδομένων. Χρησιμοποιείται για την αντιμετώπιση του προβλήματος της «κατάρας της διαστασιμότητας» (curse of dimensionality), που καθιστά δυσχερή την ανάλυση και οπτικοποίηση δεδομένων μεγάλου όγκου και πολυπλοκότητας. Επιπλέον, μειώνει τον θόρυβο, διευκολύνει την εξαγωγή κρυφών δομών και απλοποιεί τη διαδικασία εκπαίδευσης αλγορίθμων, καθώς τα μοντέλα χρειάζεται να επεξεργαστούν μικρότερο αριθμό μεταβλητών. Σύμφωνα με τη βιβλιογραφία, το PCA συχνά χρησιμοποιείται ως

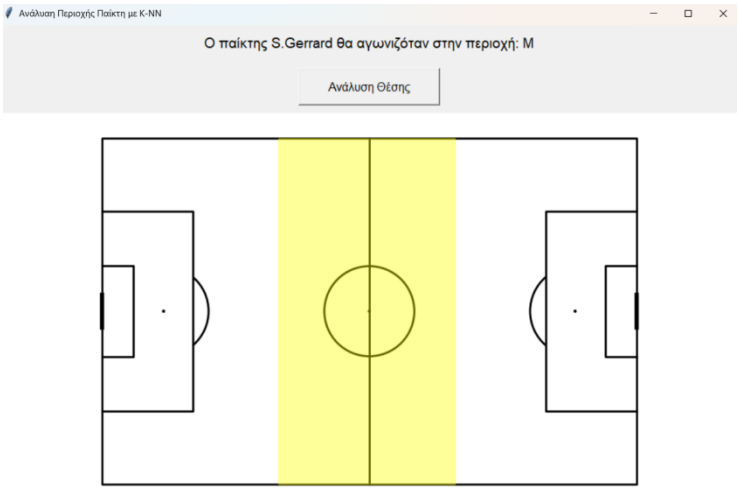
βήμα προεπεξεργασίας πριν από αλγορίθμους ταξινόμησης ή ομαδοποίησης, βελτιώνοντας τόσο την αποδοτικότητα όσο και την ακρίβεια[9].

Στην παρούσα εφαρμογή, το PCA αποδείχθηκε ιδιαίτερα χρήσιμο για την οπτικοποίηση των ποδοσφαιριστών σε έναν δισδιάστατο χώρο, επιτρέποντας την καλύτερη κατανόηση των σχέσεων μεταξύ των χαρακτηριστικών τους. Μέσα από τα διαγράμματα PCA, έγινε εφικτό να παρουσιαστεί στον χρήστη μια απλουστευμένη αλλά κατανοητή αναπαράσταση της κατανομής των παικτών, διευκολύνοντας τη διάκριση μεταξύ διαφορετικών θέσεων ή ρόλων. Παράλληλα, η μείωση διαστασιμότητας συνέβαλε στη βελτίωση της ταχύτητας εκτέλεσης των αλγορίθμων ταξινόμησης, χωρίς να χαθεί ουσιαστική πληροφορία. Έτσι, το PCA ενίσχυσε τόσο την αποδοτικότητα όσο και την ερμηνευσιμότητα της εφαρμογής.

Κεφάλαιο 4ο: Αποτέσματα

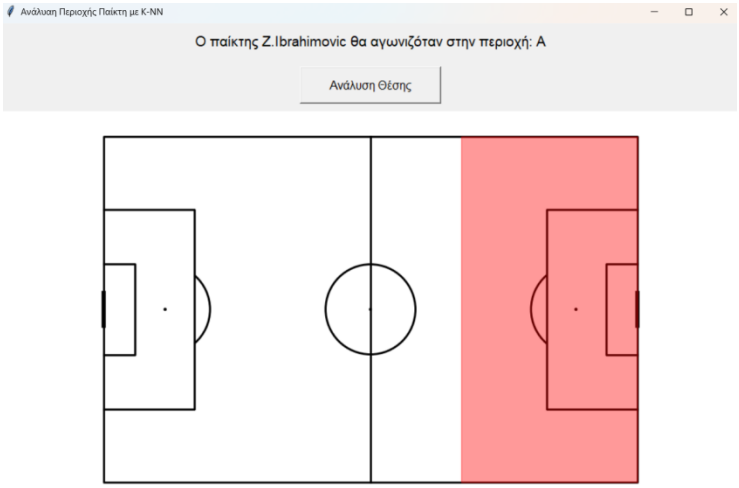
4.1 Αποτελέσματα αλγορίθμου K-NN

Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως αμυντικός (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό αμυντικό. Ο συγκεκριμένος ποδοσφαιριστής στην εποχή του ήταν ένας από τους καλύτερους κεντρικούς αμυντικούς της εποχής του. Τιμή για το k ήταν το 250. Όμως όσο και να αυξάναμε ή να μειώναμε την τιμή ο αλγόριθμος πάλι το ίδιο αποτέλεσμα μας βγάζει. Επίσης, με το κουμπί καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (A) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (CB).



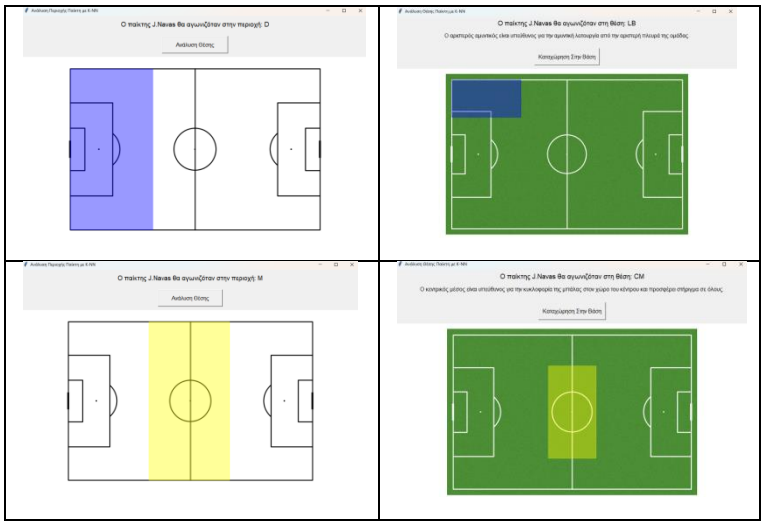
Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως μέσος (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό μέσο. Ο συγκεκριμένος ποδοσφαιριστής στην εποχή του ήταν ένας από τους καλύτερους κεντρικούς μέσους της εποχής του. Τιμή για το k ήταν το 500. Όμως όσο και να αυξάναμε ή να μειώναμε την τιμή ο αλγόριθμος πάλι το ίδιο αποτέλεσμα μας βγάζει. Επίσης, με το κουμπί

καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (M) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (CM).

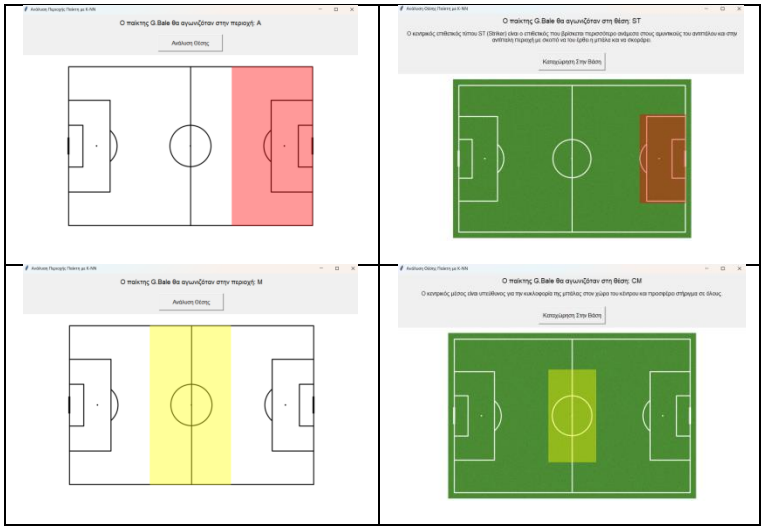


Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως επιθετικός (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό επιθετικό τύπου ST (Striker). Ο συγκεκριμένος ποδοσφαιριστής στην εποχή του ήταν ένας από τους καλύτερους κεντρικούς επιθετικούς της εποχής του. Τιμή για το k ήταν το 25. Όμως όσο και να αυξάναμε ή να μειώναμε την τιμή ο αλγόριθμος πάλι το ίδιο αποτέλεσμα μας βγάζει. Επίσης, με το κουμπί καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (A) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (ST).

Παραδείγματα άξια σχολιασμού του αλγορίθμου:

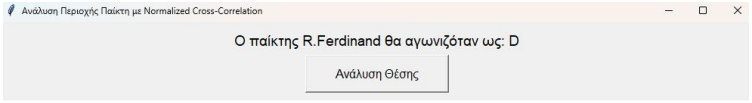


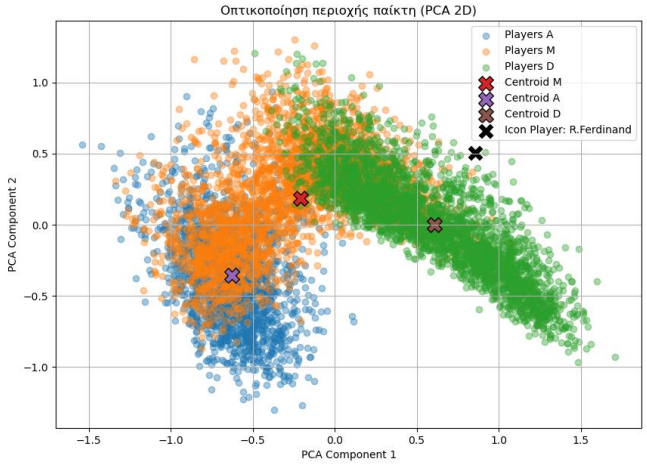
Ένα πολύ ενδιαφέρον συμπέρασμα σε αυτόν τον αλγόριθμο ήταν ο Jesus Navas. Όσοι θα τον θυμούνται είναι ο παίκτης που ξεκίνησε ως εξτρέμ και έχει αγωνιστεί σε ομάδες όπως η Μάντσεστερ Σίτυ και η Σεβίλλη. Στα τέλη της καριέρας του αγωνιζόταν ως δεξιός μπακ λόγω της έλλειψης τρεξιμάτων και αντοχής. Όπως έχω αναφέρει στην σύγχρονη εποχή δεν ξεχωρίζουν τόσο οι πλάγιες θέσεις στο γήπεδο. Ωστόσο, σε πρώτη ανάλυση με k<125 ο παίκτης αγωνίζεται ως αριστερός αμυντικός. Όμως με k>125 ο παίκτης τοποθετείται στην θέση του κεντρικού μέσου (CM).

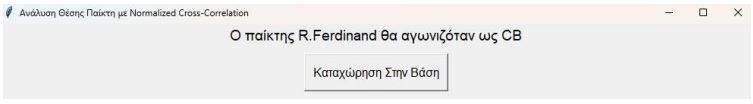


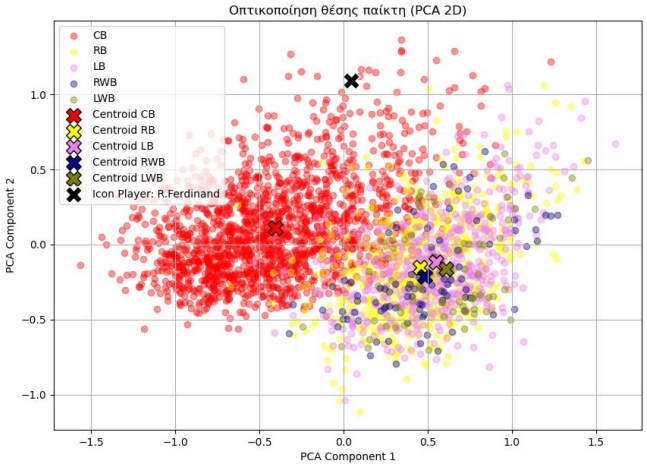
Ακόμα ένα θαυμάσιο σχολιασμού είναι η ανάλυση του Gareth Bale. Όλοι σχεδόν γνωρίζουμε την ιστορία του. Ο παίκτης που από αριστερός αμυντικός κατέλειξε στο τέλος της καριέρας του να αγωνίζεται κεντρικός επιθετικός. Ωστόσο με την θέση που άφησε εποχή είναι αυτή του εξτρέμ. Με βιογραφικό σε Τότεναμ και Ρεάλ Μαδρίτης ο Ουαλός ποδοσφαιριστής κατά την ανάλυση μας με k<277 αγωνίζεται ως κεντρικός επιθετικός, ωστόσο με τιμή μεγαλύτερη από αυτήν αγωνίζεται ως κεντρικός μέσος (CM). Όπως καταλαβαίνουμε όλοι είναι δύο τελείως διαφορετικές θέσεις στο γήπεδο με διαφορετικές απαιτήσεις και χρειάζονται διαφορετικά χαρακτηριστικά για να αγωνιστείς σε αυτές. Αφήνω έναν αστερίσκο στην ποιότητα των training data της εφαρμογής.

4.2 Αποτελέσματα αλγορίθμου CNN

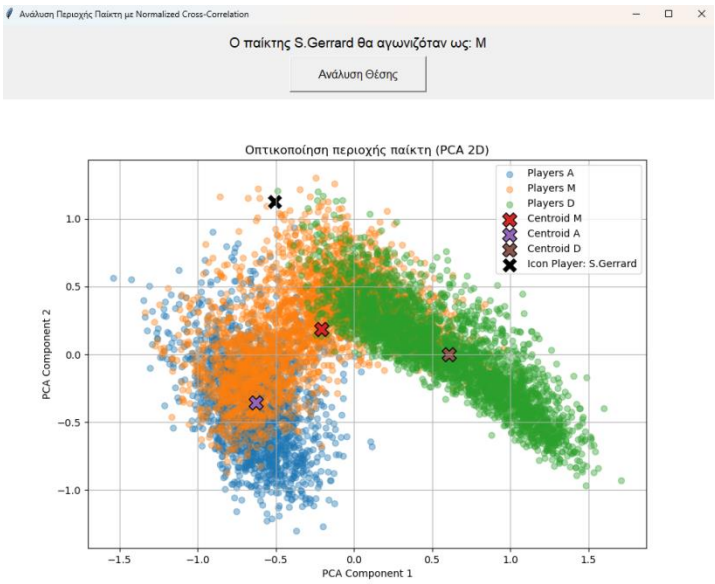


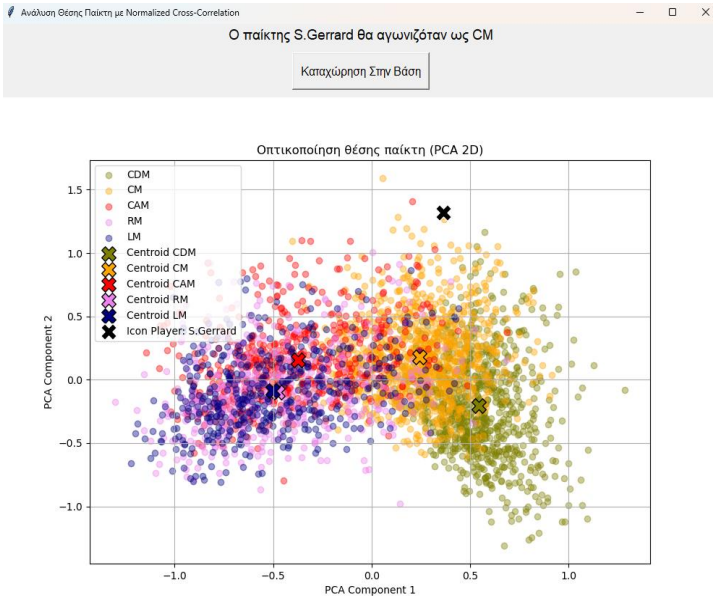




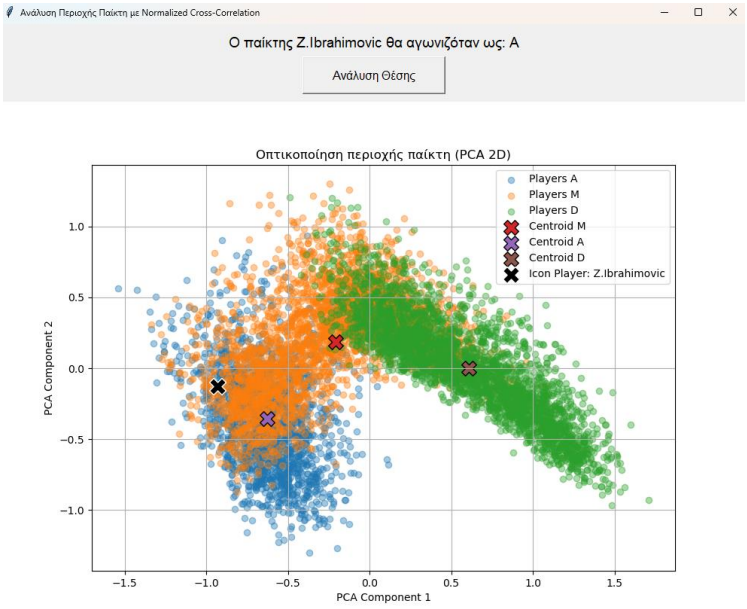


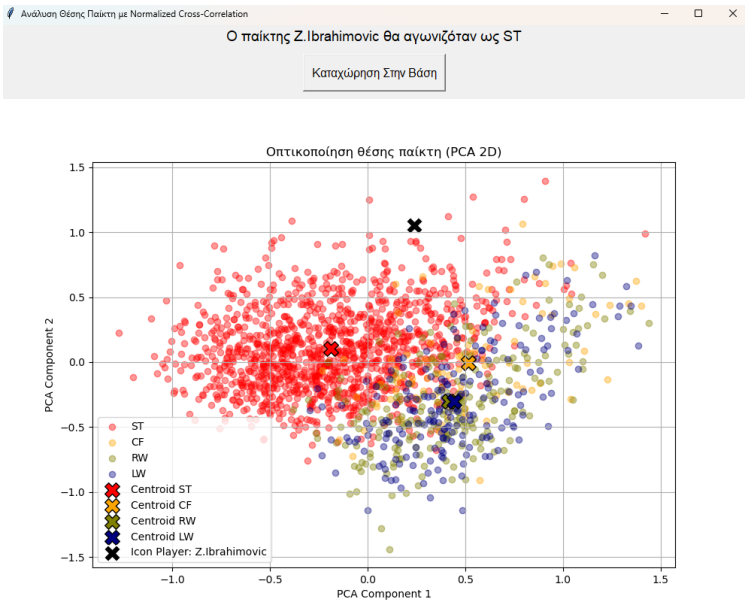
Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως αμυντικός (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό αμυντικό. Ο συγκεκριμένος ποδοσφαιριστής στην εποχή του ήταν ένας από τους καλύτερους κεντρικούς αμυντικούς της εποχής του. Όπως μπορούμε να παρατηρήσουμε από τα υπόλοιπα clusters (ομάδες) ο παίκτης δεν θα μπορούσα να αγωνίζεται σε άλλη περιοχή πέραν της άμυνας, ωστόσο αν θέλουμε να βρούμε μια δεύτερη θέση που θα μπορούσε να χρησιμοποιηθεί αυτή θα ήταν πιθανώς ως δεξιός αμυντικός (RB). Επίσης, με το κουμπί καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (A) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (CB).





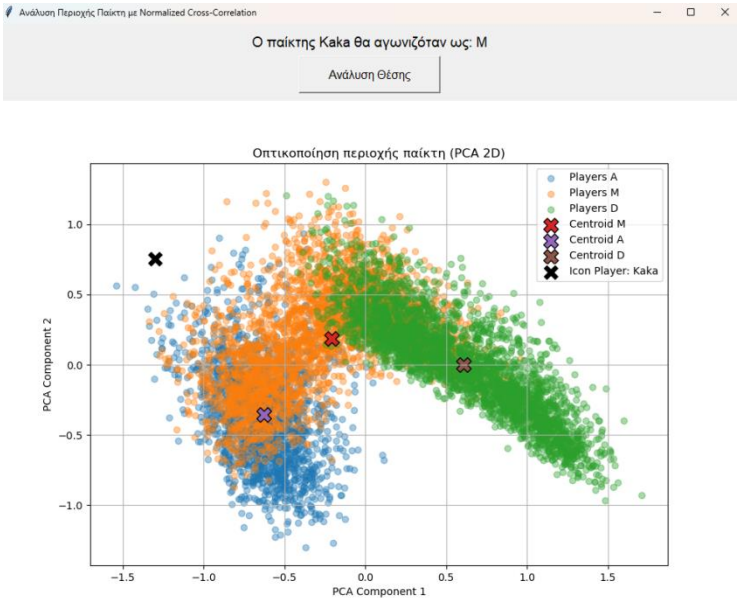
Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως μέσος (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό μέσο. Όπως μπορούμε να παρατηρήσουμε από τα υπόλοιπα clusters (ομάδες) ο παίκτης δεν θα μπορούσα να αγωνίζεται σε άλλη περιοχή πέραν του κέντρου, ωστόσο αν θέλουμε να βρούμε μια δεύτερη θέση που θα μπορούσε να χρησιμοποιηθεί αυτή θα ήταν πιθανώς ως αμυντικός μέσος (CDM). Επίσης, με το κουμπί καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (M) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (CM).

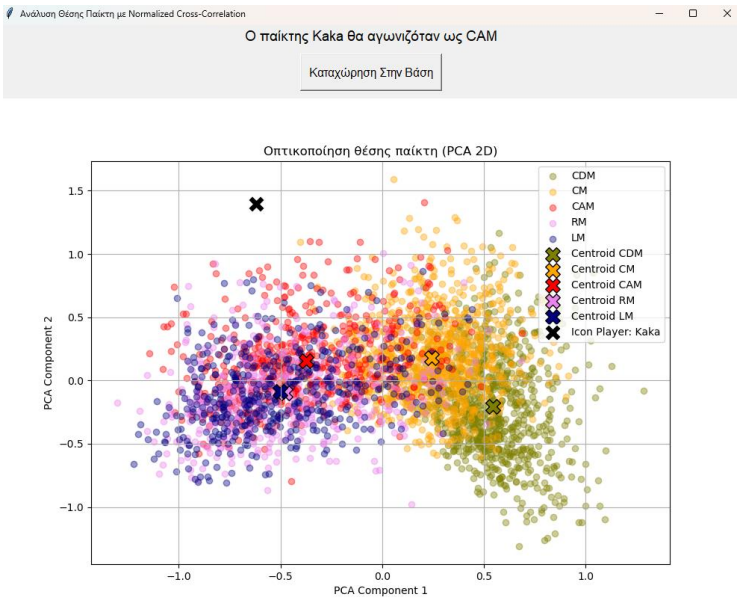




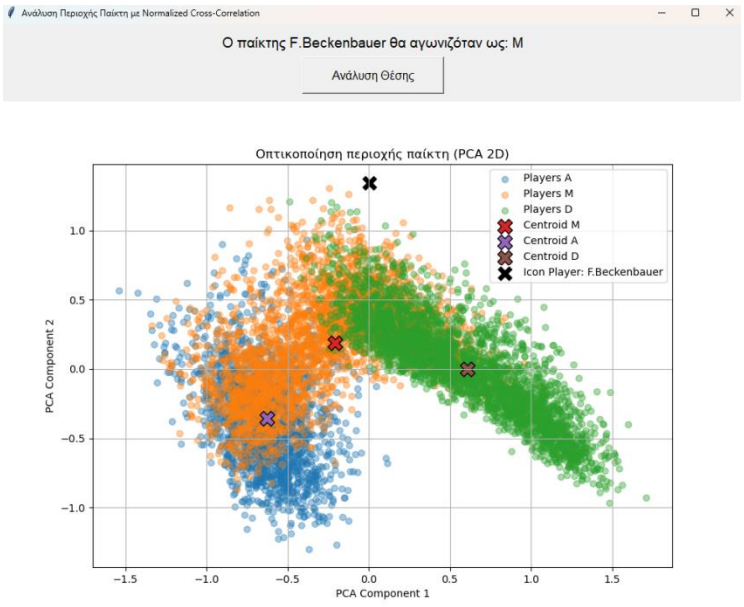
Σε αυτό το παράδειγμα, βλέπουμε την επιλογή ενός παίκτη ο οποίος θα αγωνιζόταν ως επιθετικός (κατά την πρώτη ανάλυση) και στην συνέχει πατώντας ανάλυση θέσης βλέπουμε ότι ο αλγόριθμος τον κατατάσσει ως κεντρικό επιθετικό τύπου ST (Striker). Όπως μπορούμε να παρατηρήσουμε από τα υπόλοιπα clusters (ομάδες) ο παίκτης δεν θα μπορούσα να αγωνίζεται σε άλλη περιοχή πέραν της επίθεσης, ωστόσο αν θέλουμε να βρούμε μια δεύτερη θέση που θα μπορούσε να χρησιμοποιηθεί αυτή θα ήταν πιθανώς ως επιθετικός τύπου CF. Επίσης, με το κουμπί καταχώρηση στην βάση θα καταχωρηθεί στην βάση το αποτέλεσμα της εύρεσης της περιοχής (A) και το αποτέλεσμα εύρεσης της θέσης του παίκτη (ST).

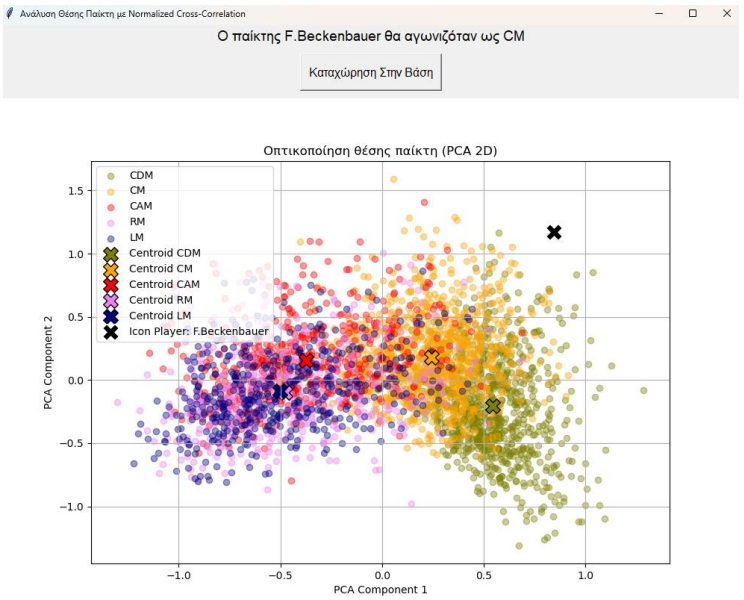
Παραδείγματα άξια σχολιασμού του αλγορίθμου:





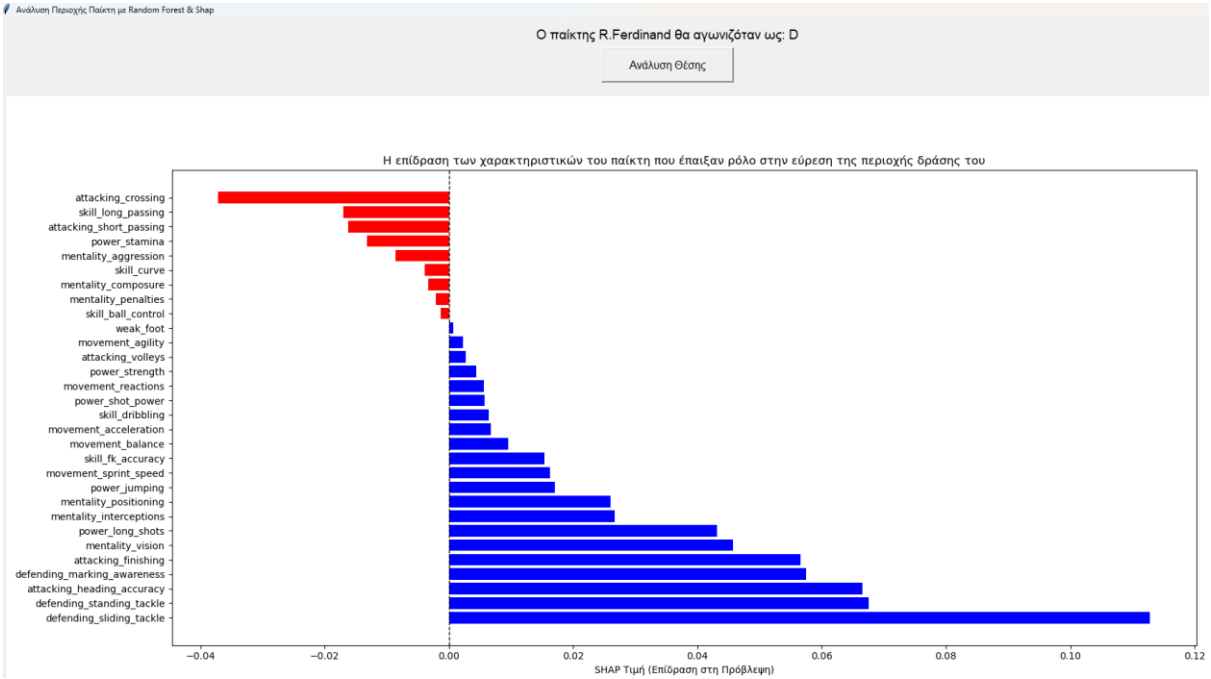
Με αφορμή τα clusters (ομάδες) που έχουν δημιουργηθεί αναζητήθηκαν κάποιοι παίκτες που πιθανόν να βρίσκονται ανάμεσα και να μην ξεχωρίζει η κατανομή του με γυμνό μάτι ή άμεσα. Ένας τέτοιος παίκτης ήταν ο Βραζιλιάνος Κακά με πέρασμα κυρίως από την Μίλαν και την Ρεάλ Μαδρίτης. Η αλήθεια είναι ότι αναμέσα σε δημοσιογράφους και αναλυτές χαρακτηρίζεται ως μεσοεπιθετικός. Ωστόσο όταν αναζητείς θέσεις, η θέση είναι μια και ανήκει είτε στο κέντρο είτε στην επίθεση. Στο συγκεκριμένο παράδειγμα διακρίνουμε ότι ο Κακά τοποθετείται στην περιοχή του κέντρου και στην συνέχεια στην θέση του επιθετικού μέσου (γνωστή και ως δεκαριού). Ωστόσο λόγω των υψηλών χαρακτηριστικών σε σχέση με το σύνολο των σημερινών ενεργών παικτών βλέπουμε ότι είναι σχετικά μακριά από τις υπόλοιπες κουκίδες (που αναπαριστούν η κάθε μια κάποιον παίκτη). Το συμπέρασμα ενός γνώστη του αθλήματος θα ήταν ότι ίσως ο Κακά να είναι αρκετά καλός οργανωτικά και για αυτό κατατάσσεται στην περιοχή του κέντρου ωστόσο έχει και πολλά επιθετικά χαρακτηριστικά και για αυτό του δίνεται μια επιθετική θέση στην περιοχή του κέντρου. Μια τυχόν δεύτερη θέση που θα μπορούσε να χρησιμοποιηθεί ο εν λόγω κύριος θα ήταν πιθανόν ως Αριστερός Μέσος (LM).

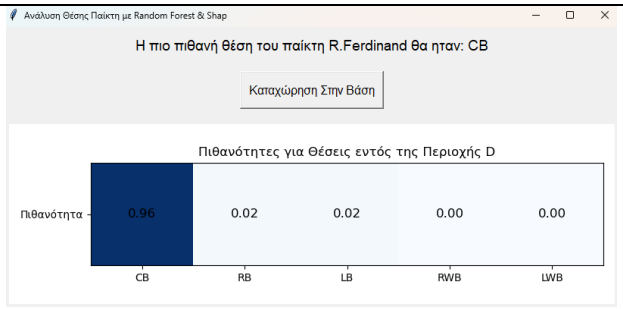




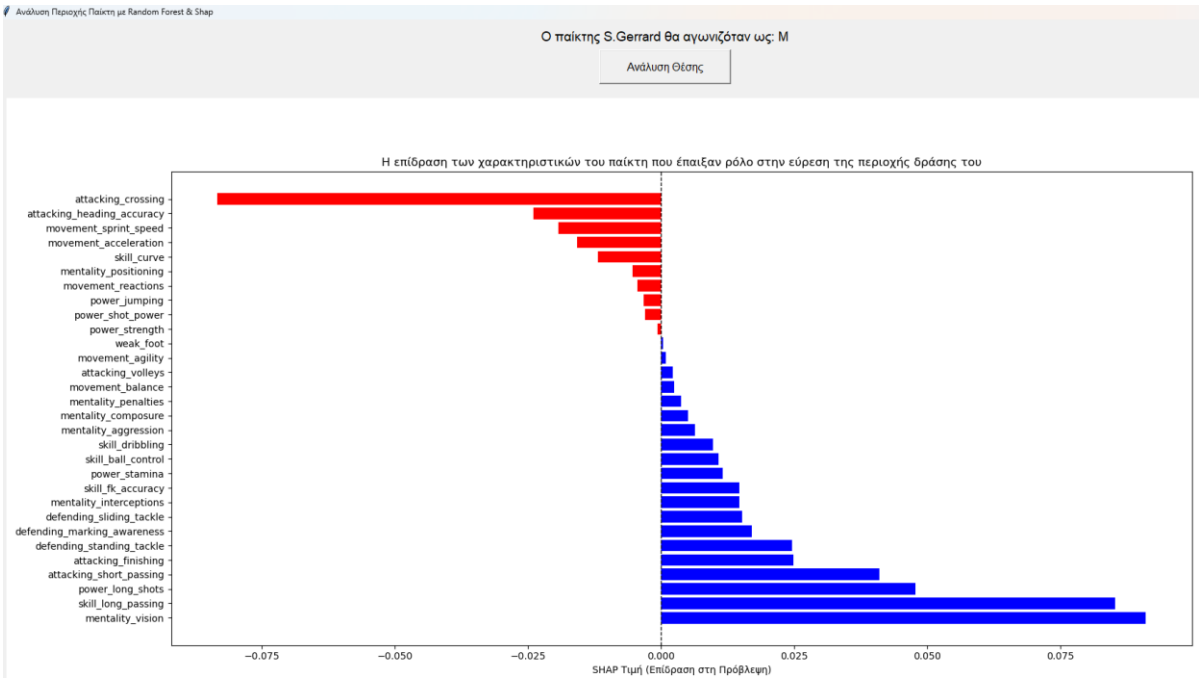
Για το επόμενο ενδιαφέρον συμπέρασμα της μεθόδου CNN θα πάρουμε έναν πολύ παλιό γνώριμο που η σημερινή κοινωνία τον γνωρίζει μέσω βίντεο στον υπολογιστή. Αυτός είναι ο Frank Beckenbauer. Ο Γερμανός που άφησε εποχή όντως ο πιο ολοκληρωμένος αμυντικός της εποχής του. Όλοι θα τον γνωρίζουν για την φανταστική σου τεχνική κατάρτιση και την δυνατότητα του να μπορεί να ντριμπλάρει όποιον αντίπαλι βρισκόταν μπροστά του. Όπως βλέπουμε στο παράθυρο της ανάλυσης ως προς την περιοχή ο παίκτης μας βρίσκεται μακριά από τα κέντρα λόγω των απίθανων χαρακτηριστικών του και πιο κοντινό κέντρο σε αυτόν για να μπορέσει να συμπεριληφθεί σε μια περιοχή είναι αυτή του μέσου (M). Ωστόσο, βλέποντας το διάγραμμα οπτικοποίησης της θέσης βλέπουμε ότι βρίσκεται μακριά από τους υπόλοιπους παίκτες και δεν είναι εύκολα να διακρίνεις σε ποια θέση θα αγωνιζόταν. Ο αλγόριθμος τον κατατάσσει ως κεντρικό μέσο (CM). Μια δεύτερη θέση που θα μπορούσε να αγωνιστεί πιθανώς να ήταν αυτή του αμυντικού μέσου (CDM) χάρη στις αμυντικές του ικανότητες.

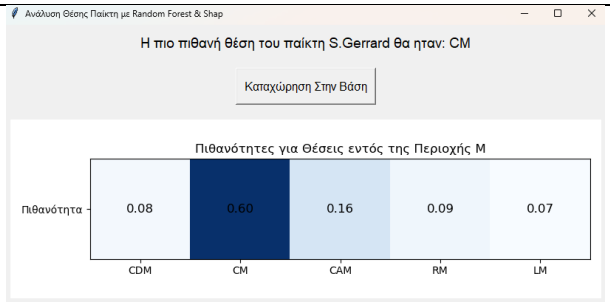
4.3 Αποτελέσματα αλγορίθμου Random Forest



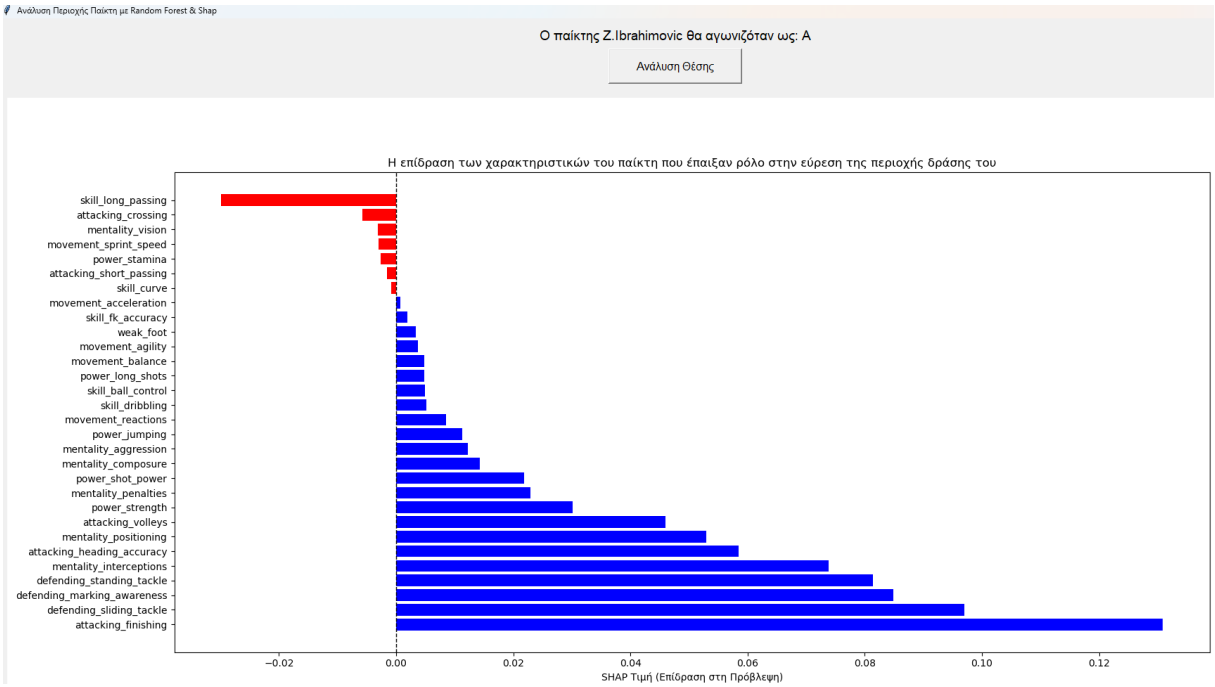


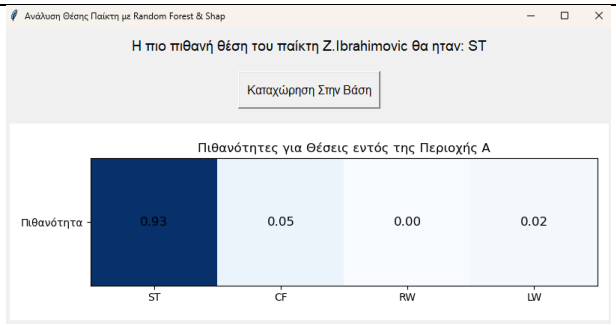
Σε αυτόν τον αλγόριθμο ταξινόμησης αρχικά χρειάστηκε και η βοήθεια της μεθόδου shap που μας βοήθησε να δούμε ποια χαρακτηριστικά έπαιξαν ρόλο στην ταξινόμηση του παίκτη ως προς την περιοχή. Σε αυτό το παράδειγμα βλέπουμε τον παίκτη να έχει κατηγοριοποιηθεί ως αμυντικός και βλέπουμε την επίδραση των χαρακτηριστικών που έπαιξαν ρόλο στην εύρεση της περιοχής δράσης του. Στην συνέχεια βρέθηκαν σε ποσοστιαία μορφή σε ποια θέση αγωνιζόταν ο παίκτης. Όπως μπορούμε να διακρίνουμε εμφανώς ο παίκτης κατά 96% θα αγωνιζόταν ως κεντρικός αμυντικός (CB).





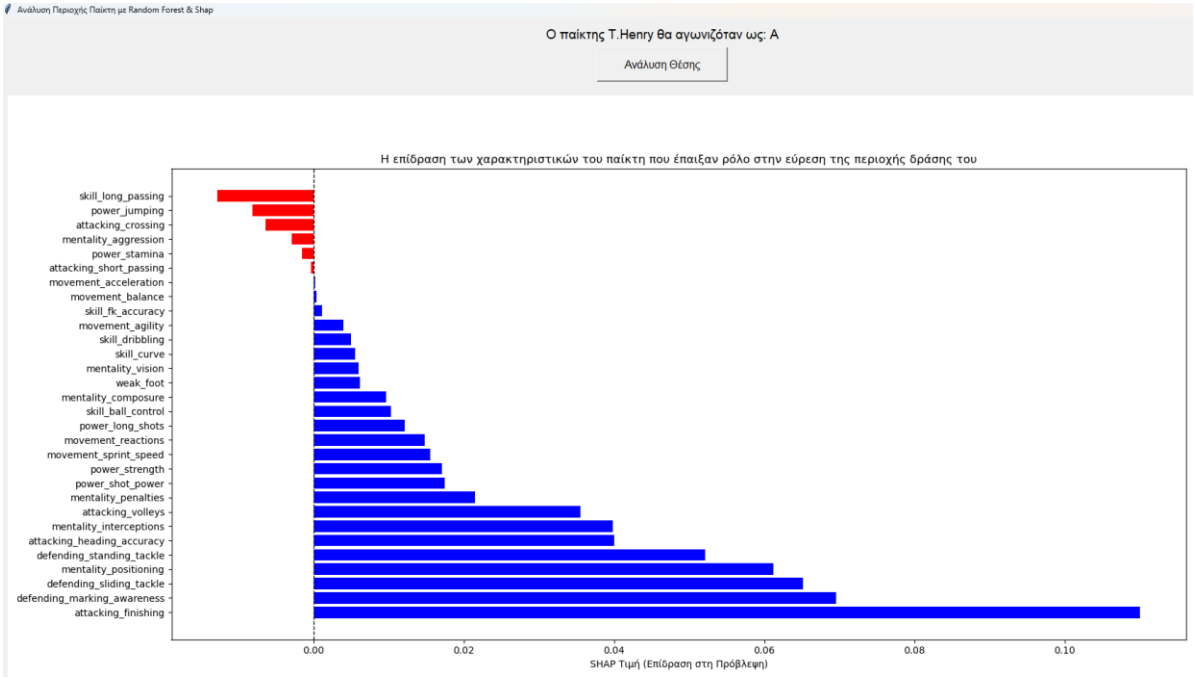
Σε αυτό το παράδειγμα βλέπουμε ότι ο παίκτης που αναλύουμε θα αγωνιζόταν ως κεντρικός με μέγιστο χαρακτηριστικό επίδρασης την ικανότητα του να βλέπει καλά το γήπεδο. Όσων αφορά την θέση είναι ξεκάθαρο ότι ο παίκτης μας θα αγωνιζόταν ως κεντρικός μέσος (CM). Και αν θέλουμε να δούμε τι επηρεάσαν αυτήν την επιλογή αρνητικά μπορούμε να δούμε τις κόκκινες μπάρες στο πρώτο γράφημα.

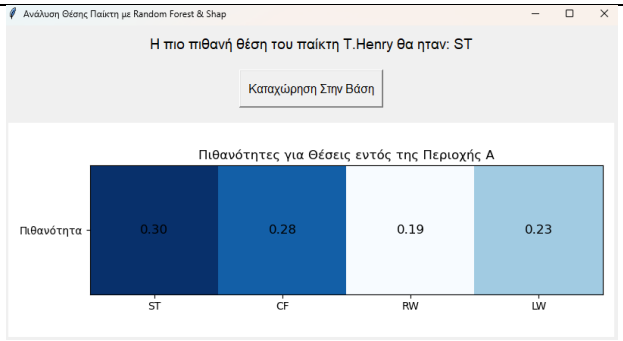




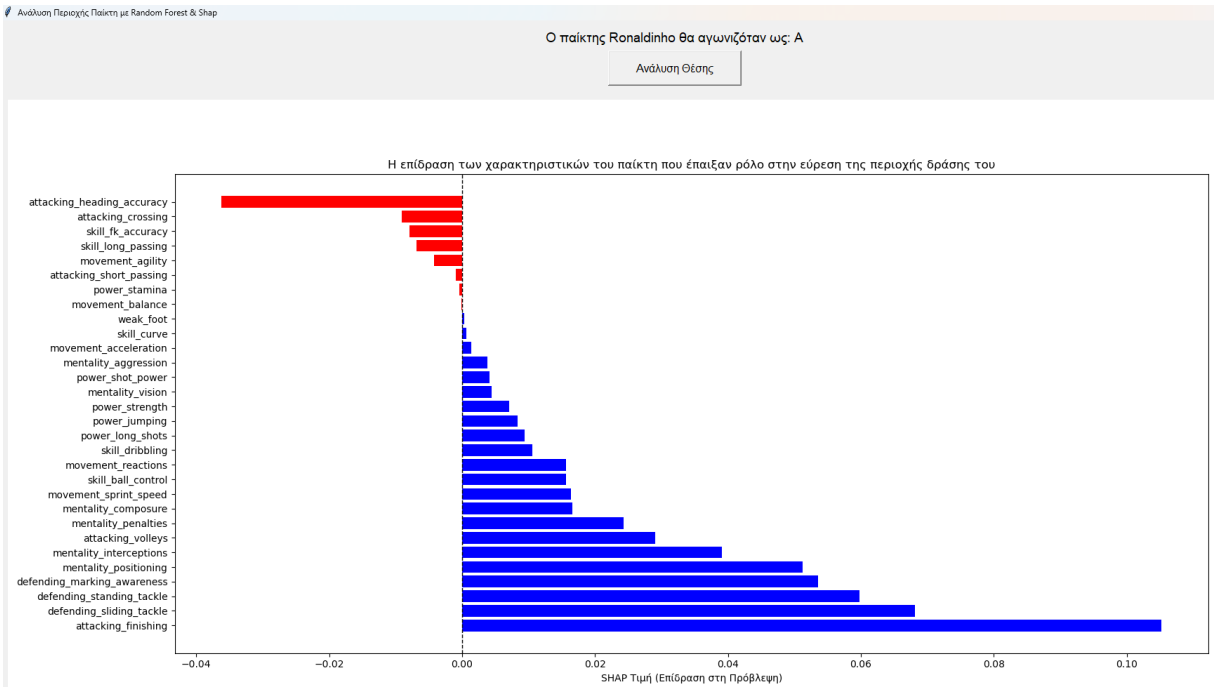
Σε αυτό το παράδειγμα βλέπουμε ότι ο παίκτης που αναλύουμε θα αγωνιζόταν ως επιθετικός με μέγιστο χαρακτηριστικό επίδρασης το τελείωμα του. Όσων αφορά την θέση είναι ξεκάθαρο ότι ο παίκτης μας θα αγωνιζόταν ως κεντρικός επιθετικός τύπου Striker (ST). Είναι εμφανές ότι ο παίκτης μας δεν θα αγωνιζόταν με μεγάλη επιτυχία σε κάποια άλλη θέση της επίθεσης.

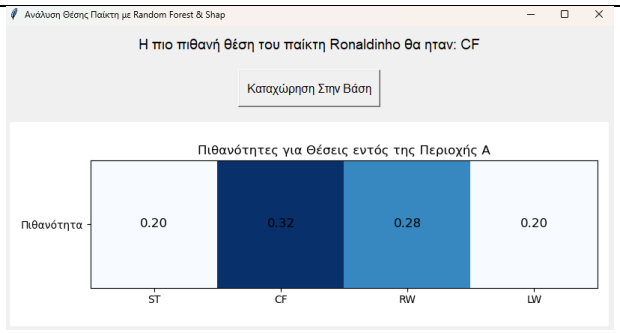
Παραδείγματα άξια σχολιασμού του αλγορίθμου:





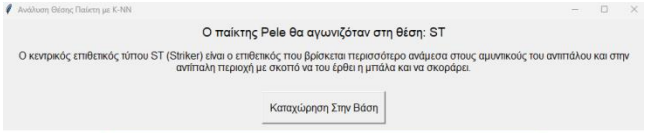
Με αφορμή την επίδραση των χαρακτηριστικών της συγκεκριμένης μεθόδου και κάποιους ήδη προϋπάρχον παίκτες που δεν είχαν κάποια σταθερή θέση προτίμησης επιλέχθηκε για δοκιμή ο Thiery Henry. Ο Γάλλος σούπερ σταρ της Άρσεναλ και της Μπαρτσελόνα διακρινόταν κυρίως για την τεχνική του ως προς την ταχύτητα, την ευελιξία αλλά και το τελείωμα του. Βλέποντας το επάνω γράφημα το χαρακτηριστικό που έπαιξε τον περισσότερο ρόλο ήταν η ικανότητα του στο τελείωμα των φάσεων. Στην συνέχεια όσον αφορά την εύρεση θέσης έχουμε ένα πολύ ενδιαφέρον αποτέλεσμα. Οι 4 θέσεις της επίθεσης ήταν σχεδόν μοιρασμένες ισόποσα. Φυσικά όλοι θα τον θυμούνται είτε ως αριστερό εξτρέμ (LW) είτε ως κεντρικό επιθετικό που έμενε κεντρικά για να πάρει την μπάλα (τύπου CF).

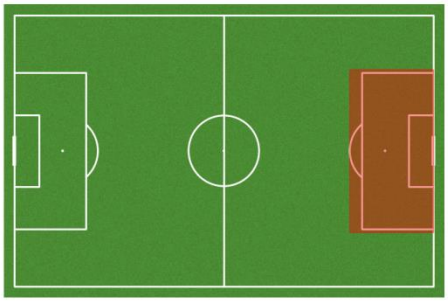


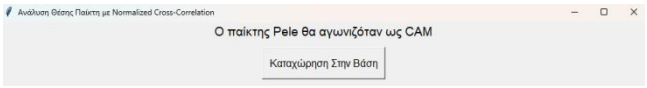


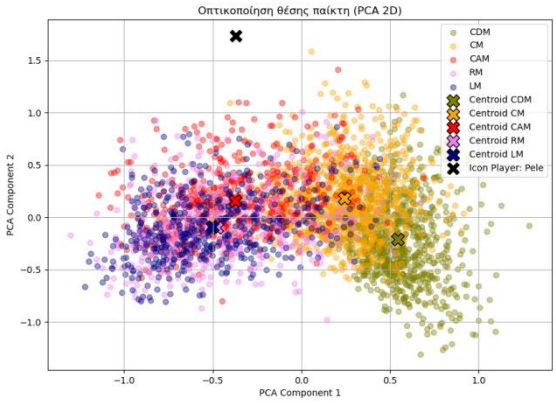
Ο επόμενος παίκτης που θεωρείται από πολλούς ως ο πιο ταλαντούχος παίκτης που έχει περάσει από το άθλημα του ποδοσφαίρου ήταν ο Ronaldinho. Ο Βραζιλιάνος σούπερ σταρ με πολλές συμμετοχές σε Μπαρτσελόνα και Μίλαν είναι σίγουρα ένας από τους πιο τεχνικούς παίκτες που πέρασαν από το ποδόσφαιρο. Φοβερός έλεγχος της μπάλας, σπουδαία τεχνική κατάρτιση και φυσικά το φοβερό του τελείωμα τον κατατάσσουν ως έναν σπουδαίο επιθετικό παίκτη. Αναμενόμενο ο αλγόριθμος ταξινόμησης να έχει μοιρασμένα τα ποσοστά όσον αφορά τις πιθανότητες του να αγωνιστεί στην κάθε θέση της επίθεσης.

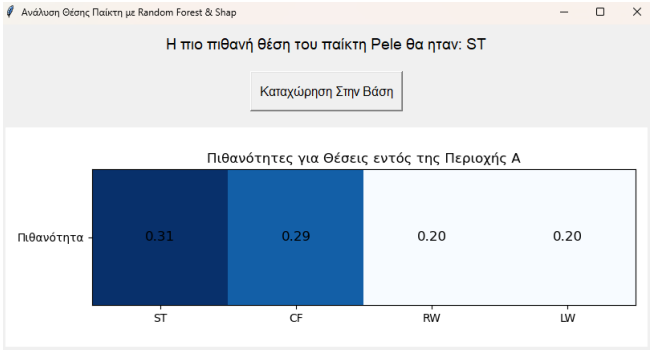
4.4 Ενδιαφέρον Συμπεράσματα Παικτών Μεταξύ Αλγορίθμων



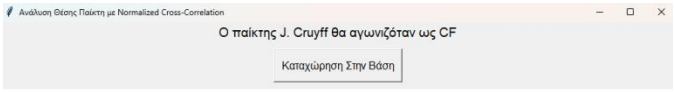


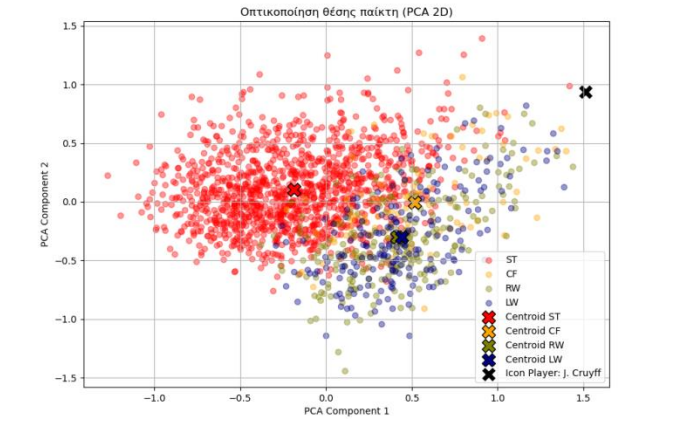


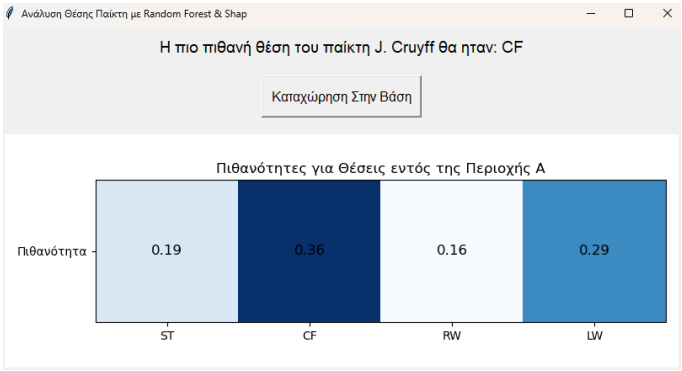




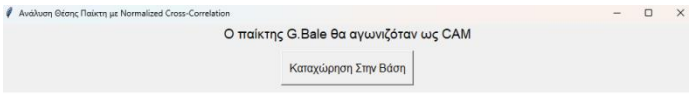
Φυσικά δεν θα γινόταν να μην γίνει πλήρη χρήση όλων των αλγορίθμων για κάποιους από τους κορυφαίους παίκτες που πέρασαν από την ιστορία του ποδοσφαίρου. Ένας εξ αυτών ήταν ο Pele. Ο Βραζιλιάνος είχε γίνει γνωστός σε κάθε άνθρωπο στον πλανήτη γη στην εποχή του. Ήταν ένας παίκτης που άφησε το σημάδι του από την πρώτη κιόλας διοργάνωση που αγωνίστηκε (Μουντιάλ 1958). Ο Pele αγωνιζόταν κυρίως στον χώρο πίσω από τον επιθετικό. Στον αλόρυθμο ταξινόμησης k-NN σε οποιαδήποτε τιμή σωστού εύρους αναζήτησης μας τον παρουσιάζει ως κεντρικό επιθετικό τύπου Striker (ST). Στον αλγόριθμο του CNN μας τον κατατάσσει οριακά ως κεντικό (ανάλυση περιοχής) με αποτέλεσμα όπως βλέπεται να είναι πολύ μακριά συγκριτικά με τους υπόλοιπους παίκτες και με πιο κοντινό του κέντρο να είναι αυτό του επιθετικού μέσου (CAM). Ενώ στον αλγόριθμο Random Forest αντικρίζουμε πόσο μοιρασμένες είναι οι πιθανότητες να αγωνιζόταν σε όλες τις θέσεις τις επίθεσης. Είναι φανταστικό πως μπορείς να βγάλεις μια ολοκληρωμένη εικόνα για έναν παίκτη ο οποίος αγωνιζόταν την επόχη του 60 και του 70.

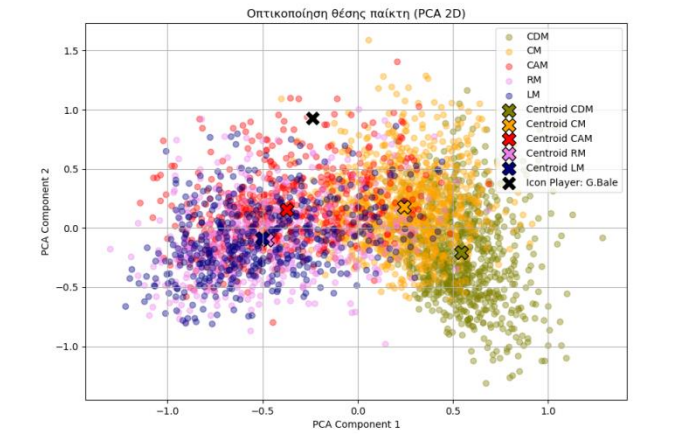


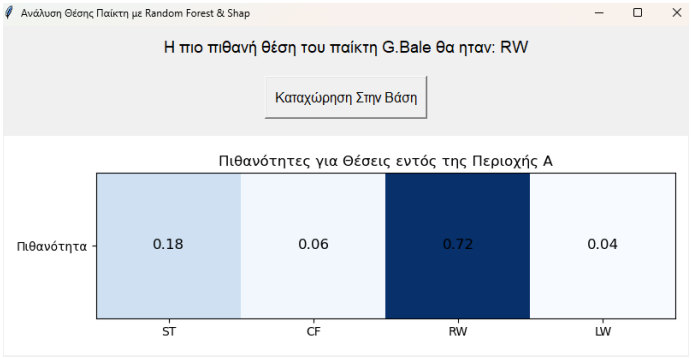




Άλλος ένας ποδοσφαιριστής που άφησε εποχή και το όνομά του ακούγεται συνεχώς, διότι ήταν το ίδιο καλός ποδοσφαιριστής όσο και προπονητής. Για κάποιους χαρακτηρίζεται ως ο πρώτος πρόγονος της προπονητικής του αθλήματος. Το σίγουρο είναι ότι άφησε εποχή. Εμείς θα δούμε αναλυτικά τι έκανε ως παίκτης και σε ποια θέση θα αγωνιζόταν. Ήταν γνωστός για την φοβερή του τεχνική κατάρτιση, την διείσδυσή του και το φοβερό έλεγχο της μπάλας. Ο αλγόριθμος k-NN τον κατατάσσει ως αριστερό εξτρέμ (LW). Ο αλγόριθμος του CNN φαίνεται από το μαύρο X που δείχνει τον παίκτη ότι τον κατατάσσει αρκετά μακριά από τους υπόλοιπους, ωστόσο το πιο κοντινό κέντρο σε αυτόν είναι αυτό του κεντρικούς επιθετικού τύπου (CF) και αμέσως κοντινότερο μετά εκείνο του αριστερού εξτρέμ. Τέλος, ο αλγόριθμος Random Forest μας επιβεβαιώνει σε αυτές τις δύο θέσεις δείχνοντας μας ότι με ποσοστό 36% για την θέση του κεντρικού επιθετικού τύπου CF και 29% στην θέση του αριστερού εξτρέμ.







Ένας παίκτης ο οποίος αναφέρθηκε σε προηγούμενο παράδειγμα αλλά αξίζει να αναλυθεί σε βάθος είναι ο Gareth Bale. Ο Ουαλός ξεκίνησε να γίνεται γνωστός στην ποδοσφαιρική κοινότητα ως αριστερός αμυντικός (LB) στην Τότεναμ. Με τον καιρό δείχνοντας ότι διαθέτει πολλά επιθετικά χαρακτηριστικά έγινε κατά σειρά αριστερός κεντρικός (LM) και στην συνέχεια μετά την μεταγραφή του στην Ρεάλ Μαδρίτης καθορίστηκε και επίσημα ως δεξιός εξτρέμ (RW) με μεγάλο ατού του να είναι να συγκλίνει με το αριστερό και να τελειώνει την φάση. Στον αλγόριθμο k-NN λόγω της σύγκρισης του στην τελευταία του γεμάτη χρονιά ποδοσφαιρικά (33 ετών) είχε ήδη χάσει ένα μεγάλο μέρος των ικανοτήτων του, με αποτέλεσμα ο αλγόριθμος σύμφωνα με τους ενεργούς ποδοσφαιριστές τον κατατάσσει ως κεντρικό επιθετικό τύπου Striker (ST). Ο αλγόριθμος CNN από την άλλη όμως τον κατατάσσει ως επιθετικό μέσο (CAM). Τέλος, ο αλγόριθμος Random Forest τον κατατάσσει σε τρίτη διαφορετική θέση, αξιολογώντας τον ως δεξιό εξτρέμ (RW) με ποσοστό 72%. Αυτό είναι το πιο ωραίο συμπέρασμα των αλγορίθμων ταξινόμησης, ότι μπορούν να σου βγάλουν έναν παίκτη σε τρεις διαφορετικές θέσεις αξιολογώντας τον διαφορετικά μεταξύ τους.

Κεφάλαιο 5ο: Αξιολόγηση αλγορίθμων

Η αξιολόγηση των αλγορίθμων ταξινόμησης αποτελεί ένα από τα σημαντικότερα στάδια σε οποιοδήποτε έργο μηχανικής μάθησης, καθώς επιτρέπει την αντικειμενική εκτίμηση της ποιότητας των μοντέλων. Ενώ η ποιοτική ανάλυση και η οπτικοποίηση των αποτελεσμάτων παρέχουν χρήσιμες πληροφορίες για συγκεκριμένα παραδείγματα, είναι απαραίτητο να παρουσιαστούν και ποσοτικά αποτελέσματα που να αποτυπώνουν την απόδοση κάθε αλγορίθμου στο σύνολο των δεδομένων. Για τον σκοπό αυτό χρησιμοποιούνται μετρικές όπως η Accuracy, η Precision, η Recall και η F1-Score, καθώς και το Confusion Matrix, το οποίο προσφέρει μια συνολική εικόνα των σωστών και λανθασμένων ταξινομήσεων.

Περιγραφή Μετρικών Αξιολόγησης:

- **Accuracy (Ακρίβεια)**
  Η ακρίβεια εκφράζει το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των παρατηρήσεων. Αντιπροσωπεύει μια γενική ένδειξη της επίδοσης του μοντέλου, ωστόσο σε περιπτώσεις μη ισορροπημένων κλάσεων (class imbalance) μπορεί να είναι παραπλανητική.
- **Precision (Ακρίβεια Θετικών Προβλέψεων)**
  Η precision δείχνει το ποσοστό των προβλέψεων μιας κλάσης που ήταν πραγματικά σωστές. Δηλαδή, μετράει πόσο «σίγουρος» είναι ο αλγόριθμος όταν προβλέπει μια συγκεκριμένη κλάση. Είναι ιδιαίτερα σημαντική όταν το κόστος ενός false positive είναι υψηλό.
- **Recall (Ευαισθησία / Ανάκληση)**
  Η recall μετράει το ποσοστό των πραγματικών παραδειγμάτων μιας κλάσης που το μοντέλο κατάφερε να αναγνωρίσει σωστά. Είναι σημαντική όταν το κόστος ενός false negative είναι υψηλό, δηλαδή όταν είναι κρίσιμο να μην παραλειφθεί καμία περίπτωση.
- **F1-Score**
  Η F1-Score αποτελεί τον αρμονικό μέσο της Precision και της Recall, παρέχοντας έναν πιο ισορροπημένο δείκτη, ειδικά σε προβλήματα με ανισορροπία δεδομένων. Όσο μεγαλύτερη είναι η τιμή της (κοντά στο 1), τόσο καλύτερη θεωρείται η συνολική απόδοση του μοντέλου.

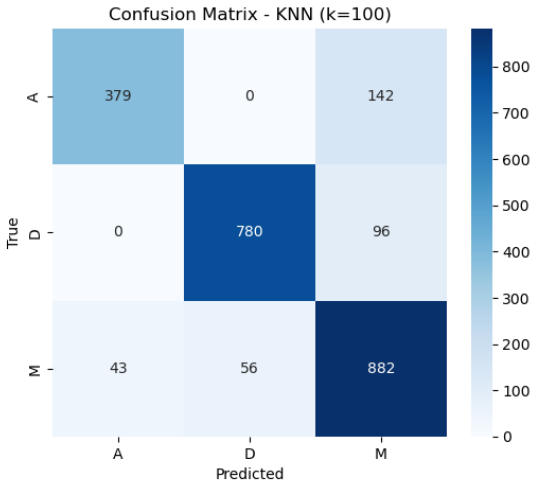
Με βάση το σύνολο δεδομένων των ενεργών ποδοσφαιριστών, πραγματοποιήθηκε **διαχωρισμός σε train και test set** και αξιολογήθηκαν τρεις αλγόριθμοι: **K-Nearest Neighbors (KNN)**, **Nearest Centroid Classifier (NCC)** και **Random Forest**.

5.1 K-Nearest Neighbors (KNN, k=100)

Ο αλγόριθμος KNN πέτυχε ικανοποιητικά αποτελέσματα, με υψηλή ακρίβεια στις κατηγορίες που διαθέτουν επαρκή αριθμό δειγμάτων. Παρόλο που είναι ευαίσθητος στον αριθμό των κοντινών γειτόνων (k) και στη διανομή των δεδομένων, κατάφερε να διακρίνει αποτελεσματικά τις κύριες περιοχές δράσης των παικτών.

- **Accuracy**: 0.86
- **Precision**: 0.87
- **Recall**: 0.84
- **F1-Score**: 0.85

Το confusion matrix έδειξε ότι οι περισσότερες λανθασμένες ταξινομήσεις προήλθαν από τη σύγχυση μεταξύ μέσων (M) και επιθετικών (A), γεγονός που δικαιολογείται από την εγγενή ομοιότητα των χαρακτηριστικών τους.



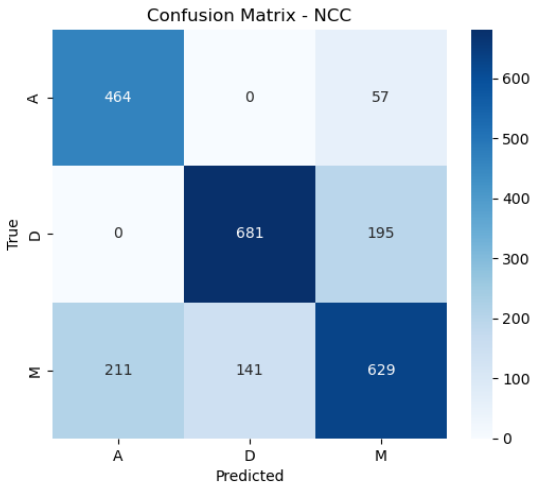
Σχήμα 5.1 Confusion Matrix - KNN

5.2 Nearest Centroid Classifier (NCC)

Ο NCC βασίζεται στον υπολογισμό του κεντροειδούς κάθε κλάσης και την απόδοση του δείγματος σε αυτόν που βρίσκεται πιο κοντά. Η μέθοδος αυτή είναι ιδιαίτερα απλή και αποδοτική, ωστόσο δεν μπορεί να μοντελοποιήσει σύνθετες σχέσεις. Αυτό αποτυπώθηκε και στα αποτελέσματα, όπου η ακρίβεια ήταν χαμηλότερη σε σχέση με τους άλλους αλγορίθμους.

- **Accuracy**: 0.75
- **Precision**: 0.74
- **Recall**: 0.77
- **F1-Score**: 0.75

Το confusion matrix κατέδειξε ότι ο αλγόριθμος έχει την τάση να συγχέει τους αμυντικούς (D) με τους μέσους (M), δείχνοντας την περιορισμένη ικανότητα του να διαχωρίζει κατηγορίες με κοντινές στατιστικές κατανομές.



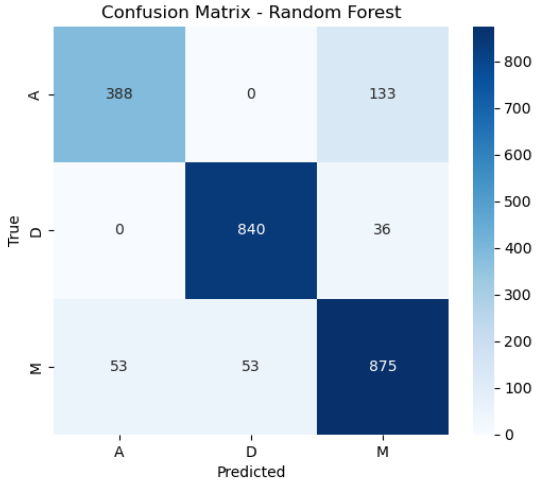
Σχήμα 5.2 Confusion Matrix - NCC

5.3 Random Forest (σε συνδιασμό με SHAP)

Ο Random Forest εμφάνισε την καλύτερη απόδοση, καθώς συνδυάζει πλήθος δέντρων απόφασης και πραγματοποιεί πλειοψηφική ψήφο, μειώνοντας έτσι τον κίνδυνο overfitting. Η απόδοση του ήταν σταθερά υψηλή σε όλες τις κλάσεις.

- **Accuracy**: 0.88
- **Precision**: 0.89
- **Recall**: 0.87
- **F1-Score**: 0.87

Το confusion matrix επιβεβαίωσε ότι ο αλγόριθμος μπορεί να διαχωρίσει με ακρίβεια όλες τις κατηγορίες, με μικρότερο ποσοστό λαθών σε σχέση με τους άλλους δύο αλγορίθμους. Η χρήση των μετρικών SHAP έδειξε επίσης ποια χαρακτηριστικά συνέβαλαν περισσότερο στην τελική πρόβλεψη.



Σχήμα 5.3 Confusion Matrix – Random Forest

Τα αποτελέσματα που προέκυψαν από την αξιολόγηση των αλγορίθμων ταξινόμησης επιβεβαιώνουν τη χρησιμότητα των μετρικών ως εργαλεία ποσοτικής ανάλυσης της απόδοσης. Η παρουσίαση του confusion matrix κατέδειξε τη διασπορά των σωστών και λανθασμένων ταξινομήσεων, ενώ οι μετρικές ακρίβειας (Accuracy), ανάκλησης (Recall), ακρίβειας πρόβλεψης (Precision) και ισορροπημένης επίδοσης (F1-score) παρείχαν μια πιο ολοκληρωμένη εικόνα της αποτελεσματικότητας των μοντέλων. Συνολικά, οι τιμές που καταγράφηκαν καταδεικνύουν ότι τα μοντέλα κατάφεραν να αποδώσουν με ικανοποιητική συνέπεια στις περισσότερες περιπτώσεις, παρά τις προκλήσεις που σχετίζονται με την πολυπλοκότητα και την ετερογένεια των δεδομένων ποδοσφαιριστών. Ως εκ τούτου, τα ευρήματα αυτά ενισχύουν την αξιοπιστία της εφαρμογής και καθιστούν εμφανή τη δυνατότητα περαιτέρω βελτιστοποίησης μέσω τροποποίησης υπερπαραμέτρων ή ενσωμάτωσης επιπλέον τεχνικών μηχανικής μάθησης.

Κεφάλαιο 6ο: Συμπεράσματα & Προτάσεις βελτίωσης

6.1. Συμπεράσματα

Η αξιοποίηση της Python ως βασικής τεχνολογίας ανάπτυξης της εφαρμογής αυτής αποδείχθηκε στρατηγικά σωστή επιλογή, τόσο από πλευράς λειτουργικότητας όσο και από άποψη ευκολίας ανάπτυξης. Η Python, ως γλώσσα υψηλού επιπέδου, προσφέρει πλήθος βιβλιοθηκών για μηχανική μάθηση (όπως scikit-learn), επεξεργασία δεδομένων (pandas, NumPy, dataframes) και οπτικοποίηση (matplotlib), επιτρέποντας την ταχεία ανάπτυξη και εύκολη ενσωμάτωση σύνθετων αλγορίθμων. Παράλληλα, η χρήση της βιβλιοθήκης Tkinter για την ανάπτυξη του γραφικού περιβάλλοντος χρήστη (GUI) διασφάλισε τη δημιουργία μιας εύχρηστης, απλής και κατανοητής διεπαφής, χωρίς περίπλοκες εξαρτήσεις. Επιπλέον, η ενσωμάτωση της SQLite3 επέτρεψε την απευθείας διαχείριση βάσεων δεδομένων με ελάχιστο προγραμματιστικό κόστος. Συνολικά, η Python λειτούργησε ως ενοποιητικό στοιχείο ανάμεσα στη διαχείριση δεδομένων, τη μηχανική μάθηση και τη διαδραστική παρουσίαση αποτελεσμάτων, επιβεβαιώνοντας τον ρόλο της ως εργαλείο αιχμής στην ανάπτυξη εκπαιδευτικών και ερευνητικών εφαρμογών.

Η εφαρμογή σχεδιάστηκε με βασικό γνώμονα τη φιλικότητα προς τον χρήστη, ώστε να μπορεί να χρησιμοποιηθεί εύκολα ακόμα και από άτομα χωρίς προηγούμενη εμπειρία σε συστήματα ανάλυσης δεδομένων ή μηχανική μάθηση. Από την αρχική οθόνη, ο χρήστης καθοδηγείται βήμα-βήμα με σαφή κουμπιά και απλές επιλογές, όπως η δυνατότητα εισαγωγής νέου παίκτη ή η επιλογή ήδη καταχωρημένου. Το περιβάλλον βασίζεται σε γραφικά Tkinter που διατηρούν καθαρό και ευανάγνωστο σχεδιασμό, ενώ τα μηνύματα σφαλμάτων ή οδηγιών προσφέρουν σαφή επεξήγηση κάθε ενέργειας. Επίσης, η απεικόνιση των αποτελεσμάτων με γραφήματα (όπως PCA plots, SHAP bars ή matrix προβλέψεων) καθιστά την ερμηνεία των δεδομένων προσιτή ακόμη και σε μη τεχνικούς χρήστες. Η προσέγγιση αυτή ενισχύει την κατανόηση της ταξινόμησης και συμβάλλει στη μεγαλύτερη αποδοχή και αξιοποίηση της εφαρμογής από προπονητές, φοιτητές ή αναλυτές.

Η εφαρμογή που αναπτύχθηκε στο πλαίσιο της παρούσας πτυχιακής εργασίας διαθέτει ένα πλήρως ολοκληρωμένο pipeline ανάλυσης, το οποίο καλύπτει όλα τα στάδια της επεξεργασίας δεδομένων και της εξαγωγής προβλέψεων. Από την εισαγωγή των χαρακτηριστικών των παικτών –είτε αυτόματα μέσω της βάσης δεδομένων είτε χειροκίνητα από τον χρήστη– έως την τελική ταξινόμηση περιοχής και θέσης, η ροή είναι απρόσκοπτη και εύχρηστη. Περιλαμβάνει κρίσιμα βήματα, όπως τον καθαρισμό και την κανονικοποίηση των δεδομένων, την εφαρμογή διαφορετικών αλγορίθμων μηχανικής μάθησης (k-NN, CNN, Random Forest), την οπτικοποίηση των αποτελεσμάτων μέσω PCA και heatmaps, καθώς και τη δυνατότητα αποθήκευσης των προβλέψεων στη βάση. Επιπλέον, η παρουσία εξηγήσιμων μεθόδων όπως η SHAP ενισχύει τη διαφάνεια του pipeline. Το σύστημα, ως εκ τούτου, δεν αποτελεί απλώς μια αποσπασματική ανάλυση, αλλά μια ολοκληρωμένη λύση για την ταξινόμηση και αξιολόγηση ποδοσφαιριστών.

Η εφαρμογή κατάφερε να ενσωματώσει αποτελεσματικά τεχνικές μηχανικής μάθησης, αποδεικνύοντας τη χρησιμότητα της τεχνολογίας στην ανάλυση αθλητικών δεδομένων και ειδικότερα στο ποδόσφαιρο. Μέσα από την αξιοποίηση διαφορετικών αλγορίθμων, όπως οι k-Nearest Neighbors (k-NN), Random Forest και Συνελικτικά Νευρωνικά Δίκτυα (CNN), το σύστημα ήταν σε θέση να αναλύει δεδομένα παικτών και να προσδιορίζει τόσο την αγωνιστική περιοχή όσο και τη θέση τους στο γήπεδο. Η ποικιλία μεθόδων επέτρεψε την ευελιξία στην επιλογή και τη δοκιμή διαφορετικών προσεγγίσεων, ενώ οι συγκρίσεις μεταξύ των αποτελεσμάτων ενίσχυσαν την αξιοπιστία των

προβλέψεων. Επιπλέον, η χρήση εργαλείων επεξήγησης όπως η SHAP βοήθησε στην κατανόηση των εσωτερικών διαδικασιών των μοντέλων. Συνολικά, η εφαρμογή πέτυχε έναν βασικό στόχο της σύγχρονης μηχανικής μάθησης: τη δημιουργία εργαλείων που είναι και αποδοτικά και κατανοητά από τον τελικό χρήστη.

Ένα από τα πιο εντυπωσιακά συμπεράσματα της εφαρμογής ήταν η ικανότητά της να διαχειρίζεται και να αναλύει έναν μεγάλο όγκο δεδομένων παικτών με ακρίβεια και συνέπεια. Η επεξεργασία χαρακτηριστικών από δεκάδες στήλες και εκατοντάδες εγγραφές ανέδειξε το βάθος που μπορεί να προσφέρει η μηχανική μάθηση σε συνδυασμό με αποτελεσματικές τεχνικές προεπεξεργασίας και οπτικοποίησης. Χάρη στη χρήση της βιβλιοθήκης pandas για διαχείριση δεδομένων, της scikit-learn για ταξινόμηση και της matplotlib για οπτικοποίηση, η εφαρμογή αποκάλυψε σχέσεις, πρότυπα και ταξινομήσεις που θα ήταν δύσκολο να αναγνωριστούν με μη αυτόματο τρόπο. Η προσέγγιση αυτή απέδειξε ότι, ακόμα και σε εφαρμογές όπως το ποδόσφαιρο όπου τα δεδομένα είναι πολυδιάστατα και πολυπαραγοντικά, είναι δυνατή η σε βάθος ανάλυση, ενισχύοντας τις δυνατότητες λήψης αποφάσεων τόσο σε τεχνικό όσο και σε στρατηγικό επίπεδο.

Η εφαρμογή διακρίνεται για την υψηλή ερμηνευσιμότητα των αποτελεσμάτων που παρέχει, γεγονός που ενισχύει τη χρηστικότητα και την αξιοπιστία της. Και οι τρεις μέθοδοι που χρησιμοποιούνται (k-NN, CNN και Random Forest με SHAP) προσφέρουν τη δυνατότητα οπτικοποίησης και ανάλυσης της πρόβλεψης με κατανοητό τρόπο για τον τελικό χρήστη. Ο αλγόριθμος k-NN αξιοποιεί την έννοια της εγγύτητας για να προσδιορίσει την κατηγορία ενός παίκτη, κάτι που αποτυπώνεται καθαρά σε PCA διαγράμματα. Η CNN μέθοδος, παρόλο που δεν είναι πραγματικό νευρωνικό δίκτυο στην υλοποίηση, χρησιμοποιεί μια παρόμοια λογική για να επιτύχει υψηλή ακρίβεια στις ταξινομήσεις. Τέλος, ο Random Forest συνδυάζεται με τη SHAP για την αιτιολόγηση των προβλέψεων, αποκαλύπτοντας ποια χαρακτηριστικά συνέβαλαν θετικά ή αρνητικά στην τελική απόφαση. Αυτό το στοιχείο προσδίδει διαφάνεια στη διαδικασία ταξινόμησης, καθιστώντας την εφαρμογή κατανοητή και αξιόπιστη.

Η εφαρμογή αποδεικνύεται ιδιαίτερα χρήσιμη ως εργαλείο συγκριτικής αξιολόγησης ποδοσφαιριστών, καθώς επιτρέπει την ανάλυση και αντιπαραβολή χαρακτηριστικών παικτών που έχουν αγωνιστεί σε διαφορετικές εποχές ή θέσεις. Μέσω των αλγορίθμων ταξινόμησης και των διαγραμμάτων που προβάλλονται, ο χρήστης μπορεί να τοποθετήσει έναν παίκτη (ενεργό ή αποσυρμένο) μέσα σε ένα συγκεκριμένο πλαίσιο αγωνιστικών απαιτήσεων, αναγνωρίζοντας τα δυνατά του σημεία σε σχέση με άλλους. Η δυνατότητα να συγκριθούν παίκτες ίδιου τύπου (π.χ. μέσοι ή επιθετικοί) ως προς τα ποιοτικά τους χαρακτηριστικά βοηθά στην εξαγωγή ποσοτικών και ποιοτικών συμπερασμάτων. Επιπλέον, οι visual μέθοδοι όπως PCA ή heatmaps ενισχύουν τη διαδικασία, παρέχοντας ένα ευκρινές πεδίο σύγκρισης που δεν βασίζεται απλώς στη φήμη ή στη μνήμη, αλλά σε δεδομένα. Έτσι, η εφαρμογή συμβάλλει στη δημιουργία μιας πιο αντικειμενικής βάσης αξιολόγησης.

Μια από τις πιο εντυπωσιακές πτυχές της εφαρμογής είναι η αποτελεσματικότητα των διαγραμμάτων στην ανάδειξη ουσιαστικών συμπερασμάτων μέσα από απλές οπτικές αναπαραστάσεις. Οι κουκίδες που αντιπροσωπεύουν τους παίκτες και τα X που συμβολίζουν τα κέντρα των κατηγοριών (περιοχών ή θέσεων) επιτρέπουν στον χρήστη να κατανοήσει με μια ματιά τη θέση και το αγωνιστικό προφίλ κάθε ποδοσφαιριστή. Μέσα από τεχνικές όπως το PCA, ο πολυδιάστατος χώρος μετατρέπεται σε ένα διαχειρίσιμο δισδιάστατο επίπεδο, στο οποίο οι αποστάσεις αποκτούν πραγματικό νόημα. Έτσι, γίνεται εύκολη η σύγκριση του υπό μελέτη παίκτη με τους ενεργούς συναδέλφους του, ενώ η απόστασή του από τα κέντρα κατηγοριών αποκαλύπτει με ακρίβεια σε ποια ομάδα ή θέση είναι πιο

πιθανό να ανήκει. Το γεγονός ότι ένα τόσο πολύπλοκο μοντέλο συνοψίζεται σε ένα διάγραμμα καθιστά την εφαρμογή εξαιρετικά χρηστική, εύληπτη και οπτικά καθοδηγούμενη.

Βέβαια αξίζει να σημειωθεί ότι για το συγκεκριμένο θέμα που είναι σχετικό με το ποδόσφαιρο και τις θέσεις είναι σχετικό το αποτέλεσμα της ανάλυσης. Και αυτό διότι μια σύγκριση που μπορεί μέσω ενός αλγορίθμου και μιας ταξινόμησης να σου βγάλει ένα συμπέρασμα να μην μπορεί να ισχύει στην πραγματικότητα. Στο ποδόσφαιρο δεν παίζουν μπάλα μόνο τα φυσικά χαρακτηρισικά και οι ικανότητες παιχνιδιού, αλλά η απόδοση του ποδοσφαιριστή την κάθε ημέρα. Οι αριθμητικοί δείκτες απόδοσης είναι ενδεικτικοί και δεν αποτυπώνουν πλήρως τη δυναμική ενός ποδοσφαιριστή, η οποία μπορεί να μεταβάλλεται σημαντικά ανάλογα με τις συνθήκες αγώνα. Αντίστοιχα δεν μπορούμε να θεωρήσουμε ως δεδομένο ότι το αποτέλεσμα του αλγορίθμου θα βγάλει ουσιώδη αποτέλεσμα. Δεν μπορούν όλοι οι παίκτες να παίξουν στον κεντρικό άξονα του γηπέδου ακόμα και αν διαθέτουν χαρακτηριστικά παρόμοια με αυτούς που ήδη αγωνίζονται εκεί. Συνεπώς, η τελική απόφαση για την αγωνιστική αξιοποίηση ενός ποδοσφαιριστή ανήκει στον προπονητή και το επιτελείο του, τα οποία συνδυάζουν τα στατιστικά δεδομένα με την προσωπική αγωνιστική παρατήρηση.

Στις μέρες μας στις προηγμένες χώρες οι περισσότεροι σύλλογοι έχουν βασιστεί στα δεδομένα και ότι με αυτά υπάρχουν περισσότερες πιθανότητες να αποφύγεις ένα λάθος, ενώ σε αυτές που είναι λιγότερο ανεπτυγμένες παραμένουν πεπισμένοι ότι η αντίληψη του ματιού δεν μπορεί να αντικατασταθεί από κανένα στατιστικό νούμερο και από κανέναν αλγόριθμο. Σε ορισμένες χώρες, κυρίως της Αφρικής και της Λατινικής Αμερικής, παρατηρείται μικρότερη διείσδυση της τεχνολογίας στην προπονητική διαδικασία, με έμφαση σε πιο παραδοσιακές μεθόδους. Πράγματι, ο τελευταίος παράγοντας που θα κρίνει αν κάτι είναι βάσιμο από την ανάλυση ενός αλγορίθμου είναι ο προπονητής και οι άνθρωποι που αποτελούν την ομάδα του.

Η εφαρμογή που αναπτύχθηκε μπορεί να αποτελέσει ένα ιδιαίτερα πολύτιμο εργαλείο για σκοπούς εκπαίδευσης και ανάλυσης, ειδικά στο πλαίσιο σχολών ποδοσφαιρικής ανάλυσης, προπονητικής ή ακόμα και σε ακαδημίες και ερασιτεχνικά σωματεία που επιθυμούν να εισαγάγουν τη λογική της data-driven προσέγγισης. Μέσα από ένα φιλικό περιβάλλον χρήστη και τη χρήση καθιερωμένων αλγορίθμων μηχανικής μάθησης, οι σπουδαστές και οι επαγγελματίες έχουν τη δυνατότητα να δουν στην πράξη πώς μπορεί να αξιοποιηθεί η στατιστική απεικόνιση για την εξαγωγή ουσιαστικών συμπερασμάτων. Επιπλέον, οι οπτικοποιήσεις (όπως το PCA και οι SHAP γραφικές αναλύσεις) προσφέρουν εξαιρετική διδακτική αξία, επιτρέποντας στους χρήστες να κατανοούν πώς διαφοροποιούνται οι παίκτες στον χώρο με βάση τα χαρακτηριστικά τους. Μελλοντικές κατευθύνσεις ανάπτυξης θα μπορούσαν να περιλαμβάνουν την ενσωμάτωση της εφαρμογής σε προγράμματα σπουδών επιστημόνων δεδομένων με κατεύθυνση το ποδόσφαιρο (Data Science in Football).

6.2 Προτάσεις Βελτίωσης

Μία σημαντική κατεύθυνση για τη βελτίωση της εφαρμογής αφορά το αισθητικό και λειτουργικό σκέλος της εμφάνισης. Αν και η τρέχουσα υλοποίηση παρέχει ένα απλό και κατανοητό περιβάλλον χρήστη, θα μπορούσε να ωφεληθεί από έναν εκσυγχρονισμό στη σχεδίαση του γραφικού περιβάλλοντος. Η χρήση μιας πιο μοντέρνας βιβλιοθήκης διεπαφής χρήστη θα μπορούσε να προσδώσει μεγαλύτερη ευελιξία στο design, πιο ομαλές μεταβάσεις και responsive συμπεριφορά σε διαφορετικά μεγέθη οθόνης. Επίσης, η ενσωμάτωση θεμάτων χρωματικής παραμετροποίησης, tooltips με επεξηγήσεις και καλύτερης διάταξης των στοιχείων θα βελτίωνε αισθητά την εμπειρία του τελικού χρήστη. Τέλος, η δυνατότητα αλλαγής γλώσσας ή η προσαρμογή του UI ανάλογα με το επίπεδο

εμπειρίας του χρήστη (π.χ. απλό/προχωρημένο mode) θα πρόσθεταν λειτουργικότητα και θα έκαναν την εφαρμογή ακόμα πιο φιλική και προσβάσιμη.

Μία σημαντική πρόταση βελτίωσης για την εφαρμογή αφορά την αναζήτηση και αξιοποίηση μίας πληρέστερης και ποιοτικά ανώτερης βάσης δεδομένων ποδοσφαιριστών. Η τρέχουσα βάση, αν και επαρκής για δοκιμές και βασικές αναλύσεις, ενδέχεται να περιλαμβάνει ελλιπή ή περιορισμένα χαρακτηριστικά, ιδίως όσον αφορά πιο λεπτομερείς στατιστικές αποδόσεις, χρονικές μεταβολές, ή τακτικά δεδομένα. Η ένταξη δεδομένων από πιο σύγχρονες ή επαγγελματικές πηγές όπως η Opta ή το StatsBomb, θα μπορούσε να εμπλουτίσει σημαντικά το περιεχόμενο, προσφέροντας ακριβέστερες πληροφορίες ανά αγώνα, σεζόν, ή ακόμη και ανά φάση παιχνιδιού. Αυτό θα επέτρεπε την εφαρμογή να πραγματοποιεί πιο εξειδικευμένες αναλύσεις, να βελτιώσει την ακρίβεια των μοντέλων της και να προσφέρει πιο ρεαλιστικά και αξιόπιστα αποτελέσματα. Η βελτίωση της βάσης δεδομένων αποτελεί επομένως θεμέλιο για την εξέλιξη της εφαρμογής σε επαγγελματικό επίπεδο.

Μια σημαντική προσθήκη όσων αφορά την βελτίωσης της εφαρμογής θα ήταν η ενίσχυση των δεδομένων με περισσότερους ανενεργούς ποδοσφαιριστές (icon players), προκειμένου να εμπλουτιστεί η βάση συγκρίσεων και να ενισχυθεί η ανάλυση ιστορικών στοιχείων. Στην παρούσα έκδοση, οι ανενεργοί παίκτες αποτελούν περιορισμένο δείγμα, γεγονός που περιορίζει τη δυνατότητα γενίκευσης των αποτελεσμάτων. Η εισαγωγή περισσότερων ιστορικών μορφών του ποδοσφαίρου από διαφορετικές χρονικές περιόδους, χώρες και στυλ παιχνιδιού, θα επιτρέψει την εξαγωγή πληρέστερων συμπερασμάτων για την εξέλιξη των χαρακτηριστικών των ποδοσφαιριστών, καθώς και για τη διαχρονική σύγκριση ταλέντου, φυσικών ικανοτήτων και τακτικής. Επιπλέον, θα εμπλουτίσει τη λειτουργικότητα της εφαρμογής σε επίπεδο εκπαιδευτικό και αναλυτικό, καθιστώντας την πιο ελκυστική για επαγγελματίες αναλυτές, ερευνητές και φιλάθλους που επιθυμούν να εξερευνήσουν τον "διάλογο" μεταξύ παλαιότερων και σύγχρονων ποδοσφαιριστών.

Ένα ακόμη κρίσιμο σημείο ενίσχυσης είναι η περαιτέρω εξειδίκευση της ανάλυσης ως προς τον ρόλο κάθε θέσης μέσα στο γήπεδο. Ενώ η τρέχουσα έκδοση κατατάσσει τους ποδοσφαιριστές σε βασικές θέσεις (όπως ST, CM, CB), δεν λαμβάνει υπόψη τις υποκατηγορίες ή τα αγωνιστικά προφίλ που υπάρχουν εντός της ίδιας θέσης. Για παράδειγμα, ένας επιθετικός μπορεί να λειτουργεί είτε ως target man, είτε ως false nine, με εντελώς διαφορετικά χαρακτηριστικά και ρόλους στο build-up της ομάδας. Αντίστοιχα, ένας μέσος μπορεί να είναι box-to-box, deep-lying playmaker ή αμυντικός χαφ. Η προσθήκη τέτοιων διαφοροποιήσεων θα επέτρεπε μια πιο πλούσια και ρεαλιστική απεικόνιση της αγωνιστικής ταυτότητας του κάθε παίκτη. Με τη βοήθεια επιπλέον χαρακτηριστικών ή προσαρμοσμένων συνδυασμών δεικτών, η εφαρμογή θα μπορούσε να ταξινομεί παίκτες όχι μόνο ως προς τη θέση τους, αλλά και ως προς τον αγωνιστικό ρόλο που ενσαρκώνουν, ενισχύοντας έτσι τη στρατηγική αξία της ανάλυσης.

Ένα πεδίο πιθανής βελτίωσης αφορά την ενσωμάτωση πιο εξελιγμένων και διαδραστικών γραφημάτων για την απεικόνιση των αποτελεσμάτων. Αν και η παρούσα έκδοση περιλαμβάνει ικανοποιητικά στατικά plots, όπως scatter plots με PCA και matrix προβολές πιθανοτήτων, η προσθήκη δυναμικών γραφημάτων (π.χ. με χρήση βιβλιοθηκών όπως Plotly ή Bokeh) θα ενίσχυε σημαντικά την εμπειρία του χρήστη. Οι χρήστες θα μπορούσαν να αλληλεπιδρούν με τα δεδομένα, να κάνουν zoom σε περιοχές ενδιαφέροντος, να τοποθετούν τον δείκτη του ποντικιού επάνω στα δεδομένα για να εμφανίζονται οι αντίστοιχες τιμές. Παράλληλα, εξελιγμένα heatmaps για θέσεις παικτών, player radar charts ή ακόμη και animated time-series διαγράμματα για συγκρίσεις μεταξύ εποχών θα προσέφεραν πολύπλευρη πληροφόρηση. Η αναβάθμιση της οπτικοποίησης σε αυτό το

επίπεδο θα έφερνε την εφαρμογή πιο κοντά στα πρότυπα επαγγελματικών εργαλείων scouting και ανάλυσης.

Ένας άξονας μελλοντικής ανάπτυξης είναι η ενσωμάτωση εναλλακτικών αλγορίθμων ταξινόμησης και πρόβλεψης, πέρα από αυτούς που ήδη χρησιμοποιούνται (k-NN, CNN, Random Forest). Για παράδειγμα, οι αλγόριθμοι Gradient Boosted Trees όπως το XGBoost και το LightGBM έχουν αποδειχθεί εξαιρετικά αποτελεσματικοί σε προβλήματα με μεγάλους πίνακες χαρακτηριστικών και ανισορροπία κατηγοριών, προσφέροντας συχνά καλύτερη ακρίβεια από τους Random Forest. Επιπλέον, η χρήση νευρωνικών δικτύων με dense layers (MLP) θα μπορούσε να ενισχύσει τις δυνατότητες γενίκευσης όταν υπάρχουν πολλά δεδομένα. Μια ακόμα ενδιαφέρουσα δυνατότητα είναι η ενσωμάτωση AutoML εργαλείων, τα οποία μπορούν να δοκιμάζουν αυτόματα διάφορα μοντέλα και υπερπαραμέτρους, επιλέγοντας το καλύτερο για κάθε υποσύνολο δεδομένων. Τέλος, η αξιοποίηση μη εποπτευόμενων αλγορίθμων clustering, όπως το HDBSCAN ή ακόμα και η χρήση deep clustering μοντέλων, θα μπορούσε να οδηγήσει σε νέα ευρήματα για τη δομή των ποδοσφαιρικών χαρακτηριστικών.

Η υποστήριξη πολλαπλών γλωσσών στην εφαρμογή αποτελεί μια ιδιαίτερα σημαντική πρόταση βελτίωσης, καθώς διευρύνει σημαντικά το κοινό στο οποίο μπορεί να απευθυνθεί. Επί του παρόντος, η εφαρμογή έχει αναπτυχθεί στα ελληνικά, κάτι που είναι ιδανικό για τοπική χρήση ή για εκπαιδευτικά ιδρύματα εντός Ελλάδας. Ωστόσο, με την ενσωμάτωση επιπλέον γλωσσών, όπως η αγγλική, η ισπανική ή και η γερμανική, η εφαρμογή θα μπορούσε να χρησιμοποιηθεί από ακαδημαϊκούς, προπονητές ή αναλυτές σε διεθνές επίπεδο. Η πολυγλωσσική υποστήριξη θα επέτρεπε επίσης τη χρήση της σε διεθνή συνέδρια, προγράμματα ανταλλαγής, ή ακόμη και σε επαγγελματικές ομάδες ποδοσφαίρου με διεθνές προσωπικό. Η τεχνική υλοποίηση μπορεί να γίνει με χρήση λεξικών (dictionaries) και αρθρωτής αρχιτεκτονικής στα μηνύματα διεπαφής, ώστε να είναι εύκολη η προσθήκη νέων γλωσσών στο μέλλον. Η διεθνοποίηση της εφαρμογής ενισχύει τον επαγγελματισμό και την εξωστρέφειά της.

Η ενίσχυση της ασφάλειας και η ενσωμάτωση μηχανισμών καταγραφής (logging) αποτελούν σημαντικές προτάσεις για τη βελτίωση της εφαρμογής, ειδικά αν προορίζεται για ευρύτερη ή επαγγελματική χρήση. Σε επίπεδο ασφάλειας, θα μπορούσαν να προστεθούν έλεγχοι ταυτοποίησης χρηστών και διαχωρισμός δικαιωμάτων πρόσβασης, διασφαλίζοντας ότι μόνο εξουσιοδοτημένοι χρήστες μπορούν να τροποποιούν δεδομένα στη βάση. Παράλληλα, η καταγραφή ενεργειών (logs) είναι απαραίτητη για την επίβλεψη της χρήσης της εφαρμογής, την αναγνώριση λαθών και την παρακολούθηση αλλαγών στα δεδομένα των παικτών. Ένα σύστημα μηχανισμού καταγραφής (logging) μπορεί να καταγράφει ενέργειες όπως προσθήκη ή επεξεργασία παικτών, επιλογή αλγορίθμων, αποθήκευση αποτελεσμάτων κ.ά. Αυτά τα δεδομένα συμβάλλουν τόσο στη διαφάνεια όσο και στην αποσφαλμάτωση (debugging) της εφαρμογής, καθιστώντας την πιο αξιόπιστη, ασφαλή και έτοιμη για πιθανή μελλοντική ανάπτυξη σε περιβάλλοντα πολλαπλών χρηστών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Internet Site

[1] Decision Tree, Wikipedia, 2025. [Online]. Available en.wikipedia.org

[6] Icons EA FC 24, Futwiz, 2025. [Online]. Available futwiz.com

Data Sheet

[5] Stefano Leone, “EA Sports FC 24 complete player dataset”, Sept 2023

Journal Articles

[2] Rishabh Singh, “Support Vector Machines (SVM), medium.com, Oct 2024

[3] Michael Peel, “Google DeepMind unveils AI football tactics coach honed with Liverpool”, ft.com, Mar 2024

[4] Dimitrios Tsaopoulos, “Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics”, mdpi.com, April 2023

[7] Joao Medeiros, “How data analytics killed the Premier League’s long ball game”, wired.com, Aug 2017

[9] Jolliffe, I., Cadima, J. “Principal Component Analysis: A Review and Recent Developments.” Philosophical Transactions of the Royal Society A, 2016

Lecture Notes and Educational Material

[8] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. “Gradient-Based Learning Applied to Document Recognition.” Proceedings of the IEEE, 1998

AI used for troubleshooting

OpenAI ChatGPT, GPT-4 (2025)