ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
Τμήμα Μηχανικών Βιομηχανικής Σχεδίασης & Παραγωγής

Διπλωματική Εργασία

Μηχανική Μάθηση στο Ποδόσφαιρο

Βασίλης Αποστολόπουλος
ΑΜ: 222017080

Επιβλέπων Καθηγητής: Γρηγόριος Νικολάου

Αθήνα, Ιούλιος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
Τμήμα Μηχανικών Βιομηχανικής Σχεδίασης & Παραγωγής

Diploma Thesis

Machine Learning in Football

Vassilis Apostolopoulos
Registration Number: 222017080

Supervisor: Grigorios Nikolaou

Athens, July 2023

Η διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Ονοματεπώνυμο	**Ψηφιακή Υπογραφή**
Γρηγόριος Νικολάου	GRIGORIOS NIKOLAOU Digitally signed by GRIGORIOS NIKOLAOU Date: 2023.07.24 11:14:58 +03'00'
Σουλτάνα Βασιλειάδου	Soultana Vasileiadou Digitally signed by Soultana Vasileiadou Date: 2023.07.25 08:53:08 +03'00'
Χρήστος Δρόσος	Christos Drosos Digitally signed by Christos Drosos Date: 2023.07.26 17:40:03 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Αποστολόπουλος Βασίλης με αριθμό μητρώου 222017080 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

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

Ο Δηλών,

Αποστολόπουλος Βασίλης

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ.Νικολάου Γρηγόρη για την καθοδήγηση και την υποστήριξη καθ'όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τα αξιότιμα μέλη της εξεταστικής επιτροπής, κ.Βασιλειάδου Σουλτάνα και κ.Δρόσο Χρήστο που με τίμησαν με τη συμμετοχή τους. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου για τη στήριξη όλα αυτά τα χρόνια.

Περίληψη
Η παρούσα διπλωματική εργασία πραγματεύεται τις εφαρμογές της μηχανικής μάθησης στο ποδόσφαιρο. Αρχικά, γίνεται μια εισαγωγή στο ποιες ανάγκες προέκυψαν στο ποδόσφαιρο και πώς μπορούν να ικανοποιηθούν από τη μηχανική μάθηση. Για την καλύτερη κατανόηση του θέματος, παρέχεται μια επισκόπηση του θεωρητικού υποβάθρου της μηχανικής μάθησης για προβλήματα ταξινόμησης. Στη συνέχεια, αναφέρονται οι τύποι επιχειρηματικής αναλυτικής, πώς αυτοί μπορούν να μεταφραστούν στο ποδόσφαιρο και τι δεδομένα συλλέγονται. Επομένως, αναλύεται η συλλογή τους και ο ρόλος του cloud computing στην επεξεργασία τους. Έπειτα, παρουσιάζονται οι εφαρμογές της μηχανικής μάθησης στο ποδόσφαιρο, όπως η ανάλυση απόδοσης, η πρόληψη τραυματισμών, η οικονομική βιωσιμότητα της ομάδας αλλά ακόμα και η διαιτησία. Επιπλέον, υλοποιούνται δύο εφαρμογές κώδικα. Η πρώτη αφορά το ποιο μοντέλο ταξινόμησης είναι πιο ικανό να προβλέψει αποτελέσματα αγώνων με βάση το στατιστικό xGoals, ενώ η δεύτερη εξετάζει ποιο μοντέλο ταξινόμησης μπορεί να προβλέψει τις θέσεις που μπορεί να αποδώσει καλύτερα ένας ποδοσφαιριστής βάσει των χαρακτηριστικών του. Τέλος, γίνεται μια ανασκόπηση για να προκύψει το συμπέρασμα αν όλα αυτά βελτιώνουν ή αλλοιώνουν το άθλημα και παραθέτονται μελλοντικές προτάσεις για περαιτέρω εξέλιξη.

Λέξεις κλειδιά: μηχανική μάθηση, ποδόσφαιρο, ταξινόμηση, δεδομένα, συλλογή, επεξεργασία, εφαρμογές.

Abstract

The present thesis deals with the applications of machine learning in football. Initially, an introduction is given to the needs that have emerged in football and how they can be satisfied by machine learning. For a better understanding of the subject, an overview of the theoretical background of machine learning for classification problems is provided. Next, the types of business analytics are mentioned, how they can be translated into football analytics and what data is collected. Consequently, their collection and the role of cloud computing in processing them are analysed. Thereafter, the applications of machine learning in football are presented, such as performance analysis, injury prevention, football clubs' financial sustainability and even refereeing. Additionally, two code applications are implemented. The first one concerns which classification model is more capable of predicting match results based on xGoals statistic, while the second one examines which classification model can better predict the positions that a player can perform based on their characteristics. Finally, a review is conducted to draw the conclusion of whether all these improve or alter the sport and future proposals are put forward for further development.

Keywords: machine learning, football, classification, data, collection, processing, applications.

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Εισαγωγή

Η μηχανική μάθηση χρησιμοποιείται όλο και περισσότερο σε αρκετούς τομείς, όπως η ιατρική και η αυτοβιομηχανία. Επιπλέον, έχει αρχίσει να εφαρμόζεται στον αθλητισμό, ειδικά στο ποδόσφαιρο.

Τα ποσά που διακυβεύονται στο κορύφαιο επίπεδο είναι υπέρογκα, της τάξης των δισεκατομμυρίων. Μια μικρή απόφαση μπορεί να αποφέρει ή να κοστίσει πολλά εκατομμύρια. Αυτό έχει αναγκάσει τους συλλόγους να λειτουργούν σαν επιχειρήσεις. Δηλαδή, να λαμβάνουν αποφάσεις με γνώμονα κάποια δεδομένα και σκοπό ένα βιώσιμο μοντέλο λειτουργίας τόσο αγωνιστικά όσο και οικονομικά. Η παραμικρή λεπτομέρεια μπορεί να λειτουργήσει ως πλεονέκτημα έναντι του αντιπάλου αλλά και φυσικά να βελτιώσει τον ίδιο το σύλλογο. Έτσι, έχει προκύψει η ανάγκη για μετρικές αξιολόγησης και ανάδειξη νέων στατιστικών.

Αυτές οι ανάγκες και προέκυψαν αλλά και μπορούν να ικανοποιηθούν από την εξέλιξη της τεχνολογίας η οποία έχει κομβικό ρόλο στη συλλογή και την ανάλυση μεγάλων δεδομένων. Πιο συγκεκριμένα, έχουν επιτευχθεί υψηλές ταχύτητες δικτύων, εύκολη πρόσβαση σε δίκτυα και χαμηλό κόστος αποθήκευσης δεδομένων. **[1]**

Τα βασικά του ποδοσφάιρου

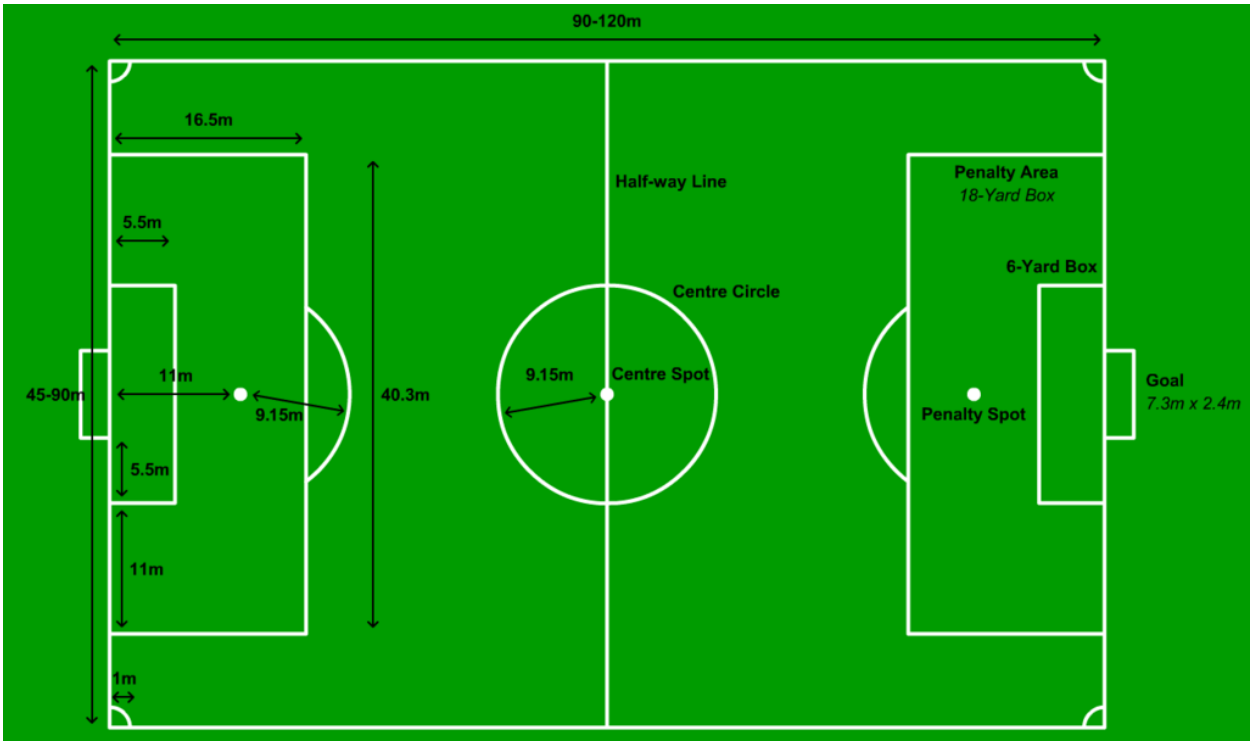
Το ποδόσφαιρο είναι ένα ομαδικό άθλημα ανάμεσα σε δύο ομάδες από έντεκα παίκτες η καθεμία και μία σφαιρική μπάλα. Ο αγώνας διεξάγεται σε ένα ορθογώνιο γήπεδο με φυσικό ή τεχνητό χλοοτάπητα και ένα μεταλλικό πλαίσιο στο μέσο κάθε μίας από τις μικρές πλευρές, το τέρμα. Σκοπός της κάθε ομάδας είναι να περάσει ολόκληρη την μπάλα στο αντίπαλο τέρμα, δηλαδή να σημειώσει τέρμα. Με απλά λόγια, να βάλει γκολ.

Οι παίκτες μπορούν να χειριστούν τη μπάλα με τα πόδια, τον κορμό και το κεφάλι. Η κάθε ομάδα έχει έναν τερματοφύλακα ο οποίος μπορεί να χειριστεί τη μπάλα με τα χέρια μόνο μέσα στις περιοχές του. Ο αγώνας αποτελείται από δύο ημίχρονα των 45 λεπτών. Η ομάδα που θα σημειώσει τα περισσότερα τέρματα κερδίζει. Αν οι δύο ομάδες σημειώσουν τον ίδιο αριθμό τερμάτων, ο αγώνας λήγει ισόπαλος.

Ο αγωνιστικός χώρος είναι σε σχήμα ορθογώνιου παραλληλόγραμμου και χαράσσεται με λευκές γραμμές. Η πλάγια γραμμή πρέπει να είναι από 90 μέχρι 120 μέτρα, ενώ η γραμμή τέρματος από 45 μέχρι 90 μέτρα. Το προτεινόμενο μήκος της πλάγιας γραμμής είναι 105 μέτρα, ενώ της γραμμής τέρματος είναι 68 μέτρα. Ο αγωνιστικός χώρος χωρίζεται σε δύο τμήματα με μια διχοτόμο γραμμή, τη γραμμή κέντρου, και περνά από το σημείο του κέντρου του γηπέδου και συναντά το μέσο κάθε πλάγιας γραμμής. Γύρω από το σημείο του κέντρου χαράσσεται κύκλος με ακτίνα 9,15 μέτρα. Επίσης, υπάρχουν δύο περιοχές κοντά στο τέρμα της κάθε ομάδας, η μικρή και η μεγάλη περιοχή. Κάθετα προς τη γραμμή τέρματος χαράσσονται δύο γραμμές που αρχίζουν από απόσταση 5,50 μέτρων από το εσωτερικό κάθετου δοκαριού. Αυτές οι γραμμές εκτείνονται μέσα στον αγωνιστικο χώρο σε απόσταση 5,50 μέτρων και ενώνονται με μια παράλληλη γραμμή με τη γραμμή τέρματος. Αυτή είναι η μικρή περιοχή. Ομοίως, χαράσσεται και η μεγάλη περιοχή αλλά η απόσταση αυτή τη φορά είναι 16,50 μέτρα. Μέσα σε κάθε περιοχή, αποτυπώνεται μια λευκή βούλα σε απόσταση 11 μέτρων από το μέσο του τέρματος. Έξω από την περιοχή, χαράσσεται ένα τόξο κύκλου με ακτίνα 9,15 μέτρα. Τέλος, σε κάθε γωνία του αγωνιστικού χώρου τοποθετείται

ένα κοντάρι με σημαία. Στο εσωτερικό κάθε γωνίας, χαράσσεται ένα τεταρτημόριο κύκλου με ακτίνα 1 μέτρο από κάθε κοντάρι. **[2]**

Το παιχνίδι έχει γενικά ελεύθερη ροή που σταματά μόνο όταν η μπάλα περάσει έξω από τον αγωνιστικό χώρο ή όταν διακοπεί από τον διαιτητή λόγω παράβασης των κανόνων. Μετά από διακοπή, το παιχνίδι ξαναρχίζει με συγκεκριμένο τρόπο επανεκκίνησης για κάθε περίπτωση. Τέλος, για την ομαλή διεξαγωγή του αγώνα υπάρχουν συγκεκριμένοι κανόνες και για την τήρησή τους είναι υπεύθυνος ο διαιτητής.



Εικόνα 1: Οι διαστάσεις ενός γηπέδου ποδοσφαίρου **[3]**

1. Θεωρητικό υπόβαθρο Μηχανικής Μάθησης

Ορισμός και ορολογία

Μηχανική μάθηση είναι η δημιουργία μοντέλων ή προτύπων με βάση ένα σύνολο δεδομένων από ένα υπολογιστικό σύστημα. Τα δεδομένα έχουν χαρακτηριστικά, τα οποία είναι αριθμητικά ή κατηγορίας, και είναι οι στήλες του συνόλου δεδομένων. Οι γραμμές ονομάζονται δείγματα ή στιγμιότυπα. Το σύνολο δεδομένων χωρίζεται σε εισόδους και εξόδους αλλά και σε δεδομένα εκπαίδευσης και ελέγχου. Οι έξοδοι ονομάζονται ετικέτες ή μέσα πρόβλεψης. **[1], [4]**

1.1 Είδη Μηχανικής Μάθησης

1. Μάθηση με Επίβλεψη (Supervised Learning)

Το σύστημα καλείται να μάθει μια έννοια ή μια συνάρτηση από ένα σύνολο δεδομένων η οποία αποτελεί την περιγραφή του μοντέλου. Κατά την εκπαίδευση, το σύστημα λαμβάνει τη σωστή απάντηση για κάθε παράδειγμα.

Το σύστημα εξετάζοντας μόνο ένα μέρος του συνόλου των περιπτώσεων καλείται να επάγει μια συνάρτηση στόχο που θα ισχύει για όλο το σύνολο. Αυτή η προσέγγιση ονομάζεται επαγωγική μάθηση. Η επαγωγική μάθηση στηρίζεται στην υπόθεση επαγωγικής μάθησης σύμφωνα με την οποία:

<<Κάθε υπόθεση h που προσεγγίζει καλά τη συνάρτηση στόχο για ένα μεγάλο σύνολο παραδειγμάτων εκπαίδευσης, θα την προσέγγιζει το ίδιο καλά και για άλλες περιπτώσεις του ίδιου προβλήματος που δεν έχει συναντήσει στην εκπαίδευση.>> ***[4]***

Υπάρχουν 2 είδη προβλημάτων μάθησης με επίβλεψη.
α) Ταξινόμηση (Classification)
Απαιτεί δημιουργία μοντέλων πρόβλεψης διακριτών κατηγοριών. (π.χ. Ναι/Όχι)

Ο στόχος είναι η αντιστοίχηση νέων δεδομένων που δεν χρησιμοποιήθηκαν στην εκπαίδευση του μοντέλου σε μια κατηγορία με τη μεγαλύτερη δυνατή ακρίβεια. **[1]**

β) Παρεμβολή ή Παλινδρόμηση (Regression)
Απαιτεί δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών. (π.χ. θερμοκρασία)

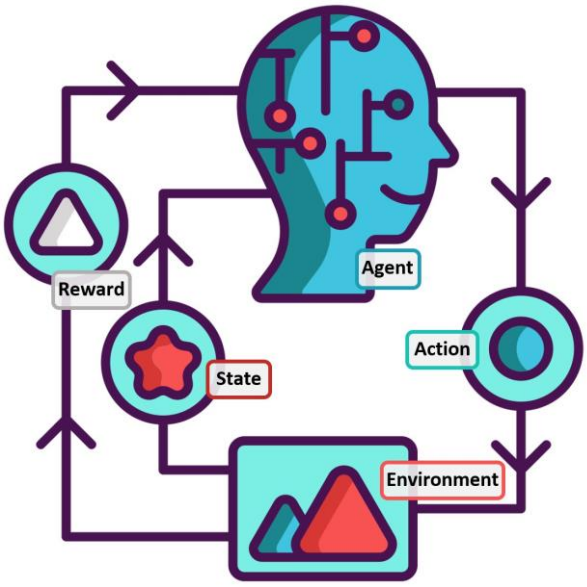
Είναι η διαδικασία προσδιορισμού της σχέσης μιας συνεχούς μεταβλητής y που ονομάζεται εξαρτημένη μεταβλητή ή έξοδος με μια ή περισσότερες μεταβλητές x_1, x_2, …, x_n που ονομάζονται ανεξάρτητες μεταβλητές. Πιο συγκεκριμένα, ο στόχος είναι να μοντελοποιηθεί η σχέση που έχουν τα ζευγάρια τιμών που έχουν δοθεί στο μοντέλο (δεδομένα εκπαίδευσης). Με τον υπολογισμό αυτής της σχέσης είναι δυνατή η πρόβλεψη της τιμής της εξόδου για νέες περιπτώσεις εισόδου. **[4]**

2. Μάθηση χωρίς Επίβλεψη (Unsupervised Learning)
Το σύστημα καλείται να ανακαλύψει μόνο του συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων μέσα από τη δημιουργία προτύπων χωρίς να γνωρίζει αν υπάρχουν ή πόσα και ποια είναι. Κατά την εκπαίδευση, το σύστημα δε λαμβάνει τη σωστή απάντηση για κάθε παράδειγμα. **[1], [4]**

Παράδειγμα
Εξατομικευμένες προτάσεις για προϊόντα με βάση τις προηγούμενες προτιμήσεις του χρήστη

3. Ενισχυτική Μάθηση (Reinforcement Learning)
Στόχος της ενισχυτικής μάθησης είναι η εύρεση της βέλτιστης συμπεριφοράς ενός λογισμικού <<πράκτορα>> σε ένα περιβάλλον, με βάση την ανταμοιβή που λαμβάνει σε μια τελική κατάσταση, ξεκινώντας από μια αρχική κατάσταση και επιλέγοντας μια σειρά ενεργειών. Το σύστημα δεν καθοδηγείται και πρέπει να ανακαλύψει μόνο του τις ενέργειες που θα του αποφέρουν το μεγαλύτερο κέρδος. **[1], [4]**

Μοντέλο Ενισχυτικής Μάθησης **[1]**
1. Παρατήρηση της κατάστασης (State)
2. Επιλογή δράσης (Action)
3. Εκτέλεση δράσης
4. Παρατήρηση νέας κατάστασης
5. Παρατήρηση για τυχόν επιβράβευση (Reward)
6. Μάθηση από την εμπειρία
7. Επανάληψη



Εικόνα 2: Μοντέλο ενισχυτικης μάθησης **[5]**

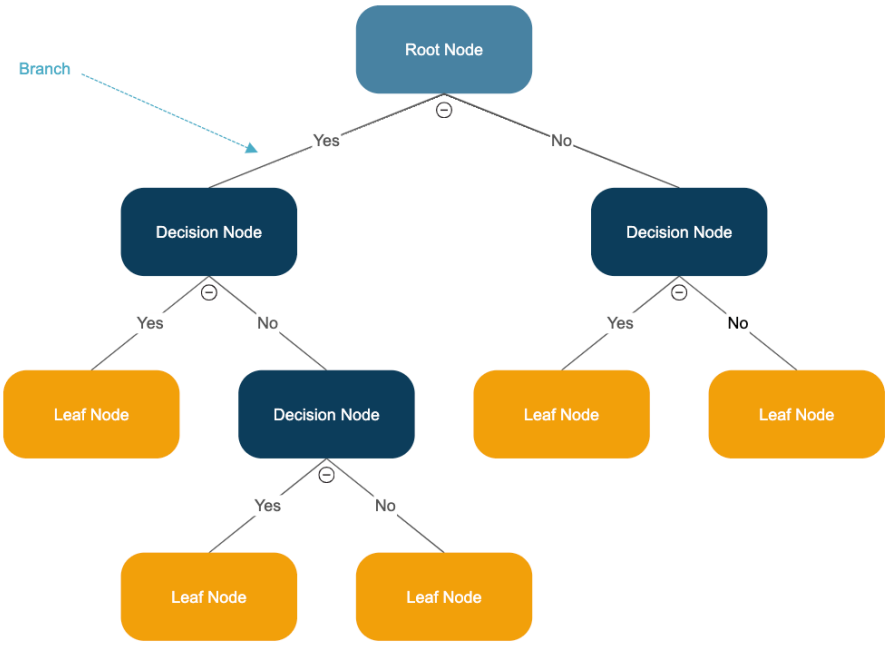
1.2 Αλγόριθμοι Μηχανικής Μάθησης για Ταξινόμηση

1. Δέντρα Αποφάσεων (Decision Trees)

Το αποτέλεσμά τους είναι μια γραφική αναπαράσταση των δεδομένων με ένα διάγραμμα δενδροειδούς δόμης. **[4]**

Δομή δέντρων απόφασης

Η βάση του δέντρου ονομάζεται ρίζα ή ριζικός κόμβος (Root Node). Από εκεί προκύπτουν οι κόμβοι απόφασης (Decision Node) που απεικονίζουν τις αποφάσεις που πρέπει να ληφθούν με βάση ένα κριτήριο διαχωρισμού. Οι κόμβοι φύλλων (Leaf Node) είναι τα αποτελέσματα αυτών των αποφάσεων. Εκτός από τον ριζικό κόμβο, όλοι οι κόμβοι έχουν μόνο έναν εισερχόμενο κλάδο (Branch) και δύο ή περισσότερους εξερχόμενους. **[1]**



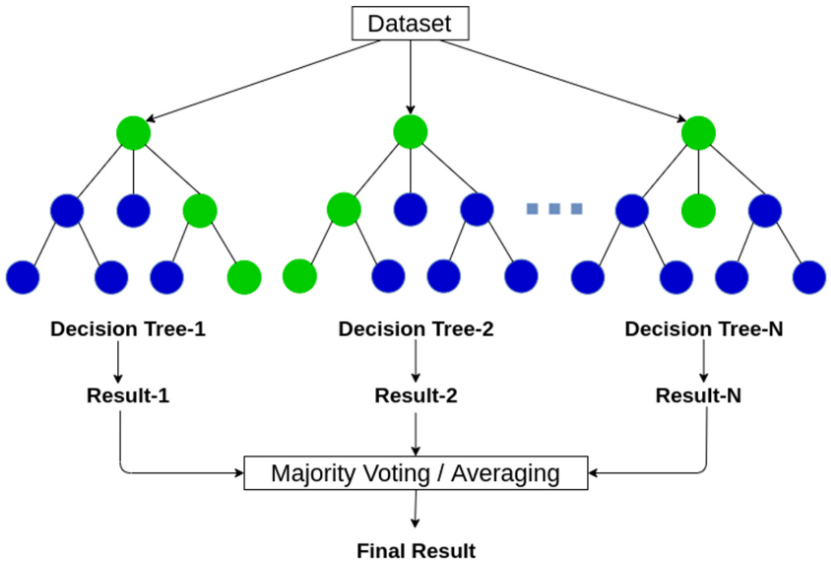
Εικόνα 3: Δομή Δέντρων Απόφασης **[6]**

Πλεονεκτήματα [1]
• Απλή δημιουργία και εύκολη χρήση
• Ευδιάκριτος τρόπος λήψης αποφάσεων
• Ανθεκτικά σε δεδομένα με θόρυβο
• Ανεπηρέαστα από απουσία δεδομένων
• Ικανότητα αναπαράστασης σαν σύνολο κανόνων

Μειονεκτήματα [1]
• Μεγάλος βαθμός εξάρτησης από το κριτήριο διαχωρισμού
• Ευαισθησία στην υπερπροσαρμογή (overfitting)
• Υπερβολικά μεγάλα δέντρα σε συγκεκριμένες περιπτώσεις

2. Αλγόριθμος Τυχαίου Δάσους (Random Forest)

Ο αλγόριθμος τυχαίου δάσους κατασκευάζει και συνδυάζει πολλά δέντρα αποφάσεων για προβλήματα ταξινόμησης και παρεμβολής. Η βασική ιδέα είναι ότι ένας μεγάλος αριθμός από δέντρα που δεν έχουν σχέση μεταξύ τους θα πάρουν καλύτερη απόφαση από κάθε δέντρο ξεχωριστά. Για να λειτουργήσει αυτό, απαιτείται χαμηλή συσχέτιση μεταξύ των επιμέρους δέντρων. Παρόλο που κάποια δέντρα θα προβλέψουν λάθος, η συνολική πρόβλεψη θα είναι προς τη σωστή κατεύθυνση. **[4]**



Εικόνα 4: Δομή Random Forest **[7]**

Πλεονεκτήματα [1]

- Επίτευξη υψηλών επιδόσεων
- Αποτελεσματικότητα σε μεγάλες βάσεις δεδομένων
- Χειρισμός χιλιάδων μεταβλητών εισόδου
- Εκτίμηση για τη σημαντικότητα των μεταβλητών στην ταξινόμηση
- Λειτουργία και με ελλειπή δεδομένα
- Εξισορρόπηση σφαλμάτων σε ανισόρροπα σύνολα δεδομένων

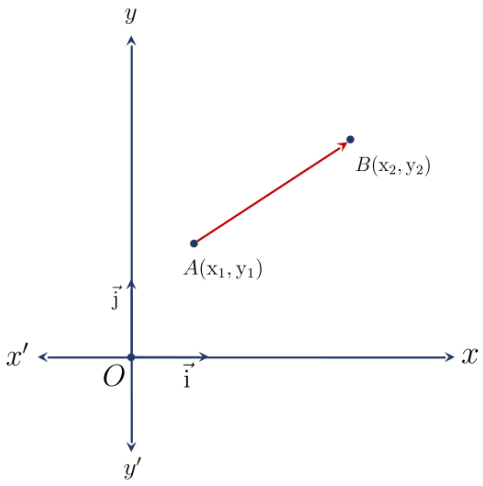
Μειονεκτήματα [1]
• Ασυνέχεια στις τιμές πρόβλεψης σε προβλήματα παρεμβολής
• Υπερπροσαρμογή σε δεδομένα με πολύ θόρυβο

Θεωρείται ένας από τους κορυφαίους αλγόριθμους ταξινόμησης και προσφέρει καλές επιδόσεις σε πολλά προβλήματα. **[4]**

3. Αλγόριθμος k-Πλησιέστερων Γειτόνων (k-Nearest Neighbors)

Βασίζεται στην υπόθεση ότι τα παραδείγματα μπορούν να αναπαρασταθούν ως σημεία σε έναν Ευκλείδειο χώρο R^n n διαστάσεων, όπου n ο αριθμός των χαρακτηριστικών. Κάθε νέα περίπτωση τοποθετείται στο χώρο ως νέο σημείο και η τιμή της κλάσης προσδιορίζεται με βάση την τιμή κλάσης των k πλησιέστερων γειτονικών του σημείων. Οι πλησιέστεροι γείτονες υπολογίζονται με βάση την Ευκλείδια απόστασή τους. **[4]**

Ευκλείδεια απόσταση 2 σημείων [8]
Έστω 2 σημεία $A(x_1,y_1)$ και $B(x_2,y_2)$. Η Ευκλείδια απόσταση d των 2 σημείων ισούται με το μέτρο του διανύσματος AB. Προφανώς, d≥0.
$d = \sqrt{[(x2 - x1)^2 + (y2 - y1)^2]}$

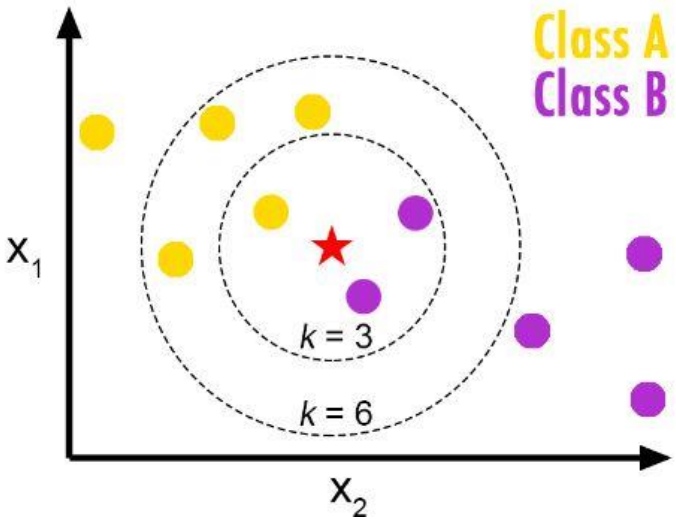


Εικόνα 5: Απόσταση 2 σημείων **[9]**

Ο καθορισμός της κλάσης γίνεται με βάση τη ψήφο πλειοψηφίας της κλάσης των k-πλησιέστερων γειτόνων ή με βαρύτητα στη ψήφο ανάλογα την απόσταση όπου συντελεστής βαρύτητας $w=\frac{1}{d^2}$ **[1]**

Επιλογή k

Το πολύ μικρό k έχει μεγαλύτερη επίδραση του θόρυβου στο αποτέλεσμα, ενώ με πολύ μεγάλο k υπάρχει πιθανότητα να υπάρχουν μες στη γειτονία σημεία και από άλλες κλάσεις με αποτέλεσμα να αυξάνεται το υπολογιστικό κόστος. Το προτεινόμενο k είναι $k=\sqrt{n}$, όπου n το πλήθος των περιπτώσεων. **[1], [4]**



Εικόνα 6: Παράδειγμα k-Nearest Neighbors για k=3 και k=6 **[10]**

Πλεονεκτήματα [1]
• Απλός στη χρήση
• Αξιόπιστος για μεγάλο εύρος προβλημάτων
• Μηδενικό υπολογιστικό κόστος εκπαίδευσης
• Ικανότητα προσέγγισης πολύπλοκων συναρτήσεων
• Γρήγορες απαντήσεις

Μειονεκτήματα [1]
• Υψηλό υπολογιστικό κόστος ταξινόμησης
• Υψηλές απαιτήσεις σε μνήμη
• Μη ευδιάκριτος τρόπος λήψης απόφασης

• Μεγάλος βαθμός εξάρτησης από τα δεδομένα
• Μεγάλος βαθμός επιρροής από το k

4. Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές (Naive Bayes Classifiers)
Βασίζεται στη στατιστική και συγκεκριμένα στο Θεώρημα του Bayes.

Θεώρημα Bayes
Υπολογίζει την υπό συνθήκη πιθανότητα P(A|B), δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση A με δεδομένο ότι ισχύει το γεγονός B.

Σύμφωνα με το Θεώρημα του Bayes:
$P(A|B) = P(B|A) * \frac{P(A)}{P(B)}$
όπου,
P(A): Η προϋπάρχουσα πιθανότητα να ισχύει η υπόθεση A.
P(B): Η πιθανότητα να συμβεί το γεγονός B χωρίς να γνωρίζουμε την υπόθεση που ισχύει.
P(B|A): Η πιθανότητα να συμβεί το γεγονός B αν ισχύει η υπόθεση A. **[4]**

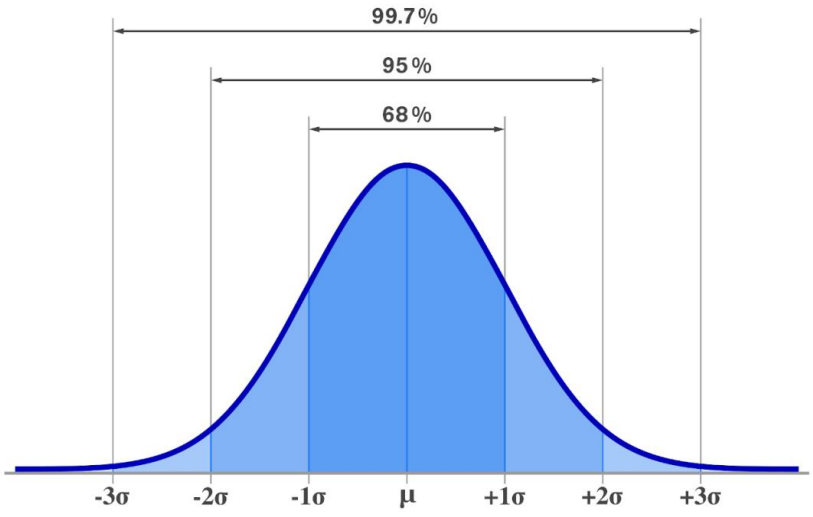
Gaussian Naive Bayes Classifier
Βασίζεται στο Θεώρημα του Bayes και στη Γκαουσιανή (Κανονική) κατανομή.

Κανονική κατανομή [11]
Είναι η καταλληλότερη να περιγράψει μια τυχαία συνεχή μεταβλητή. Συμβολίζεται ως $X \sim N(\mu, \sigma^2)$ και διαβάζεται ως η τυχαία μεταβλητή X ακολουθεί την κανονική κατανομή με συντελεστές μ και σ^2, όπου μ ο αριθμητικός μέσος και σ η τυπική απόκλιση.

<u>Χαρακτηριστικά της κανονικής κατανομής</u>
• Το υψηλότερο σημείο της κανονικής κατανομής είναι ο αριθμητικός μέσος μ.
• Η καμπύλη της κανονικής κατανομής είναι σε μορφή καμπάνας και έχει μόνο μια κορυφή.
• Η τυπική απόκλιση καθορίζει το πλάτος της καμπύλης.
• Το συνολικό εμβαδόν της περιοχής κάτω από την καμπύλη είναι 1 και κατανέμεται ισομερώς αριστέρα και δεξιά.
• Οι πιθανότητες για την κανονική τυχαία μεταβλητή δίνονται από το εμβαδόν της περιοχής κάτω από την καμπύλη. Δηλαδή,
→ 68,26% του συνολικού εμβαδού βρίσκεται σε απόσταση ± σ από τον αριθμητικό μέσο.
→ 95,44% του συνολικού εμβαδού βρίσκεται σε απόσταση ± 2σ από τον αριθμητικό μέσο.
→ 99,72% του συνολικού εμβαδού βρίσκεται σε απόσταση ± 3σ από τον αριθμητικό μέσο.

Συνάρτηση πυκνότητας πιθανότητας [11]

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$



Εικόνα 7: Γράφημα κανονικής κατανομής **[12]**

Ο αλγόριθμος Gaussian Naive Bayes είναι άμεση εφαρμογή του Θεωρήματος Bayes. Έστω X τα χαρακτηριστικά ενός παραδείγματος και y η κλάση που πρέπει να ταξινομηθεί. Τότε, η πιθανότητα να ανήκει το παράδειγμα στην κλάση y είναι: $P(y|X) = P(X|y) * \frac{P(y)}{P(X)}$

Για την πρόβλεψη ενός αγνώστου παραδείγματος, ο αλγόριθμος υπολογίζει τις πιθανότητες για κάθε κλάση και προβλέπει την κλάση με τη μεγαλύτερη πιθανότητα. Εφόσον το P(X) είναι ίδιο για όλες τις κλάσεις και το P(y) υπολογίζεται εύκολα ως το πλήθος των παρατηρήσεων που ανήκουν στην κλάση προς τις συνολικές παρατηρήσεις, το ζητούμενο είναι ο υπολογισμός του P(X|y). Αν θεωρηθεί ότι υπάρχει σχέση εξάρτησης μεταξύ των διαστάσεων του X, ο υπολογισμός γίνεται περίπλοκος. **[13]**

Πλεονεκτήματα του αλγόριθμου [14]
• Απλός, γρήγορος και αποδοτικός αλγόριθμος
• Αποτελεσματικός σε μεγάλα σύνολα δεδομένων
• Υψηλή απόδοση σε περίπτωση ανεξαρτησίας των δεδομένων
• Χαμηλό υπολογιστικό κόστος

Μειονεκτήματα [13], [14]
• Συχνό φαινόμενο η εμφάνιση μηδενικής πιθανότητας αν μια κατηγορία δεν υπάρχει στο training set.
• Η υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών εισόδου ισχύει σπάνια.

1.3 Μετρικές Απόδοσης (Performance Metrics) για Ταξινόμηση

Σε ένα μοντέλο κατηγοριοποίησης με 2 κλάσεις υπάρχουν 4 περιπτώσεις αποτελεσμάτων και μπορούν να περιγραφούν σε έναν 2x2 πίνακα, ο οποίος ονομάζεται πίνακας σύγχυσης (confusion matrix). **[4]**

• True Positive: Σωστή ταξινόμηση στη θετική κλάση
• True Negative: Σωστή ταξινόμηση στην αρνητική κλάση

• False Positive: Λάθος ταξινόμηση στη θετική κλάση
• False Negative: Λάθος ταξινόμηση στην αρνητική κλάση

Ιδανικά, FP=0 και FN=0. **[1]**

Αν οι κλάσεις είναι περισσότερες από 2, τότε ο πίνακας παίρνει τις αντίστοιχες διαστάσεις. Για παράδειγμα, για N κλάσεις ο πίνακας έχει N*N διαστάσεων. **[4]**

Πίνακας 1: Πίνακας σύγχυσης δυαδικής ταξινόμησης **[4]**

	Προβλεπόμενη Κλάση (Predicted Class)		
		Positive	**Negative**

Πραγματική Κλάση (Actual Class)	Positive	True Positive **(TP)**	False Negative **(FN)**
	Negative	False Positive **(FP)**	True Negative **(TN)**

1. Ακρίβεια (Accuracy)

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

Η ακρίβεια σε πολλές περιπτώσεις μπορεί να δώσει παραπλανητική πληροφόρηση και για αυτό είναι φρόνιμο να λαμβάνονται υπόψη και οι υπόλοιπες μετρικές. **[4]**

2. Ευστοχία (Precision)

Το ποσοστό των δειγμάτων που ταξινομούνται στη θετική κλάση και ανήκουν σε αυτήν.

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

3. Ανάκληση (Recall)

Το ποσοστό των δειγμάτων που ανήκουν στη θετική κλάση και ταξινομούνται σε αυτήν.

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

4. F-Measure ή F1-Score

Είναι ο αρμονικός μέσος της ευστοχίας και της ανάκλησης.

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

Η μέγιστη τιμή του είναι 1 όταν Precision=Recall=1.

Το Precision και το Recall εστιάζουν μόνο στη θετική κλάση και δεν λαμβάνουν υπόψη την αρνητική. **[4]**

5. Εξειδίκευση (Specificity)

Το ποσοστό των δειγμάτων που ταξινομούνται στην αρνητική κλάση και ανήκουν σε αυτήν. Ονομάζεται και True Negative Rate.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$$

6. Ευαισθησία (Sensitivity)

Το ποσοστό των δειγμάτων που ταξινομούνται στη θετική κλάση και ανήκουν σε αυτήν. Ονομάζεται και True Positive Rate. Ουσιαστικά, είναι ίδιο με το Recall.

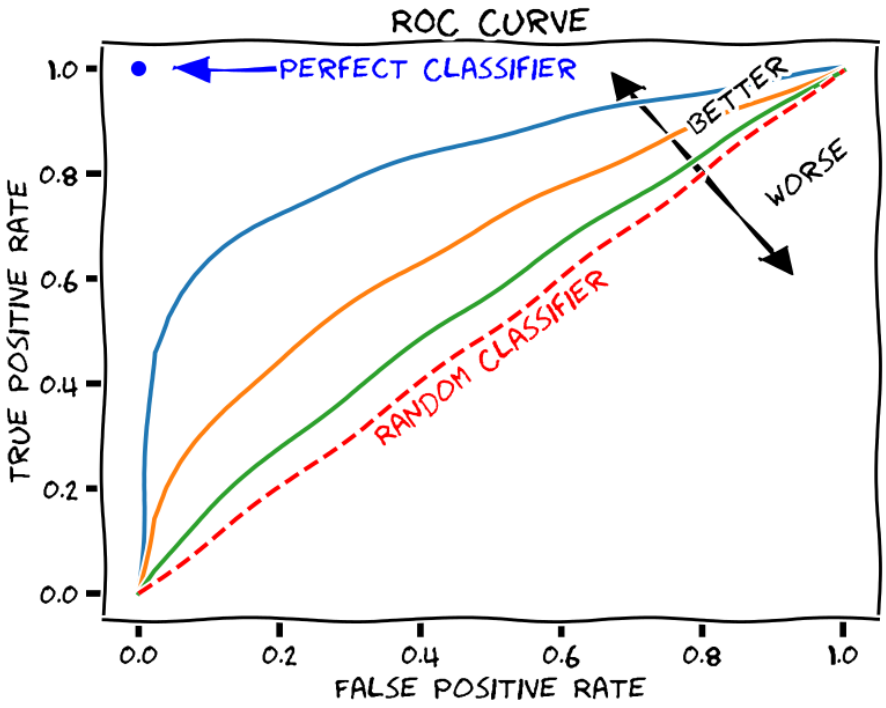
$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

Καμπύλη ROC
Γράφημα (1–Specificity), Sensitivity

1 – Specificity = 1 – $\frac{\text{TN}}{\text{TN+FP}}$ = $\frac{\text{FP}}{\text{TN+FP}}$ = False Positive Rate (FPR)

Ιδανικά, 1–Specificity=0 και Sensitivity=1.

Στον άξονα X απεικονίζεται το FPR, δηλαδή πόσα αρνητικά κατηγοριοποιούνται λάθος και στον άξονα Y το TPR, δηλαδή πόσα θετικά κατηγοριοποιούνται σωστά.



Εικόνα 8: Παραδείγματα καμπύλης ROC **[15]**

Όπως φαίνεται και στην παραπάνω εικόνα:
• Η διαγώνιος αναφέρεται σε μια τυχαία κατηγοριοποίηση.
Κάτω από τη διαγώνιο απεικονίζονται οι κατηγοριοποιήσεις σε αντίθετη κατηγορία από την πραγματική.

• Ο ιδανικός ταξινομητής βρίσκεται στο (0,1). **[1]**

Μετρική AUC (Area Under the Curve) [4]

Υπολογίζει την πιθανότητα ενός ταξινομητή να ταξινομήσει ένα τυχαίο θετικό παράδειγμα υψηλότερα από ένα τυχαίο αρνητικό. Επίσης, υπολογίζει το ποσοστό του χώρου κάτω από την καμπύλη ROC και απεικονίζει την καμπύλη με μια αριθμητική τιμή. Αναδεικνύει την ικανότητα του ταξινομητή στην αποφυγή λανθασμένων ταξινομήσεων, δηλαδή πόσο καλός είναι στο διαχωρισμό των θετικών από τα αρνητικά παραδείγματα.

$$\text{AUC} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right) = \frac{1}{2}\,(\text{TPR} + \text{TNR})$$

• Παίρνει τιμές από 0 έως 1. Ιδανικά, AUC = 1.
• Η διαγώνιος έχει AUC = 0.5
Αν ένας ταξινομητής έχει AUC>0.5 , είναι καλύτερος της τυχαίας πρόβλεψης.

Μικρο- και Μακρο- μέση τιμή [4], [16]
Σε περιπτώσεις προβλημάτων με πολλά διαφορετικά σύνολα δεδομένων ή με πολλές κλάσεις, υπάρχουν δύο μέθοδοι για τον υπολογισμό της συνολικής τους απόδοσης.

Ο μικρο-υπολογισμός (micro averaging) και ο μακρο-υπολογισμός (macro averaging) που υπολογίζουν τη μέση ευστοχία (average precision) και τη μέση ανάκληση (average recall) και στη συνέχεια τον αρμονικό τους μέσο (F1-Score).

Αρχικά, υπολογίζονται τα TP, TN, FP και FN για κάθε σύνολο δεδομένων ή για κάθε κλάση ξεχωριστά.

Μικρο-υπολογισμός
Υπολογίζονται τα επιμέρους αθροίσματα όλων των μετρικών και από αυτά υπολογίζονται το precision και το recall.

Μακρο-υπολογισμός
Υπολογίζονται οι μετρικές precision και recall για κάθε περίπτωση και στη συνέχεια υπολογίζεται η μέση τιμή τους.

Αν οι μικρο-μετρικές είναι σημαντικά μικρότερες από τις αντίστοιχες μακρο-μετρικές, υπάρχει σημαντικό πρόβλημα λανθασμένης ταξινόμησης των μεγαλύτερων σε μέγεθος κλάσεων, ενώ στις μικρές κλάσεις δεν υπάρχει κανένα πρόβλημα. Συμβαίνει ακριβώς το αντίθετο, αν οι μακρο-μετρικές είναι σημαντικά μικρότερες από τις αντίστοιχες μικρο-μετρικές.

Γενικά, ο μακρο-υπολογισμός αποτελεί απλά μια ένδειξη συνολικής απόδοσης του αλγόριθμου σε πολλές κλασεις ή σε πολλά σύνολα δεδομένων και δεν χρησιμοποιείται για τη λήψη κάποιας απόφασης για το πρόβλημα που μελετάται. Από την άλλη, ο μικρο-υπολογισμός αποτελεί μια χρήσιμη μετρική. **[4]**

Υπάρχει και μια τρίτη μέθοδος υπολογισμού της συνολικής απόδοσης που ονομάζεται σταθμισμένος υπολογισμός και αναφέρεται κυρίως στο F1-Score. Η σταθμισμένη τιμή υπολογίζεται πολλαπλασιάζοντας την τιμή F1-Score της κάθε κλάσης επί την αναλογία του αριθμού (support) των πραγματικών εμφανίσεων της κλάσης στο δοκιμαστικό σύνολο δεδομένων (test set). **[16]**

Πίνακας *2:* Παράδειγμα σταθμισμένου υπολογισμού για 3 κλάσεις **[16]**

Κλάση	**F1_Score**	**Support**	**Αναλογία Support**
Κλάση 1	0.67	3	0.3
Κλάση 2	0.40	1	0.1
Κλάση 3	0.67	6	0.6
Σύνολο	-	10	1

Weighted F1-Score = (0.67 * 0.3) + (0.4 * 0.1) + (0.67 * 0.6) = 0.64

Είναι μια πολύ χρήσιμη μέθοδος για μη ισορροπημένα σύνολα δεδομένων και χρειάζεται να δοθεί έμφαση στις κλάσεις με τα περισσότερα παραδείγματα.

Γενίκευση

Ένα πολύ σημαντικό ζήτημα είναι πώς θα αντιδράσει το μοντέλο σε νέες παρατηρήσεις που δεν ανήκουν στο σύνολο εκπαίδευσης. Αυτό ακριβώς καθορίζει την αξία του μοντέλου, δηλαδή η ικανότητά του να προβλέπει την κλάση άγνωστων παρατηρήσεων του πραγματικού κόσμου.

Η υπερπροσαρμογή (overfitting) παρουσιάζεται όταν το μοντέλο είναι υπερβολικά περίπλοκο. Δηλαδή, μπορεί να αφομοιώσει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, αλλά αδυνατεί να καταγράψει σχέσεις γενικότερης ισχύος. Το αποτέλεσμα της υπερπροσαρμογής είναι ιδιαίτερα υψηλές επιδόσεις στο σύνολο εκπαίδευσης, αλλά δυσανάλογα χαμηλές επιδόσεις σε νέες παρατηρήσεις. Για αυτόν τον λόγο, η ακρίβεια ενός μοντέλου πρέπει να εκτιμάται σε άγνωστες παρατηρήσεις. **[13]**

Η υποπροσαρμογή (underfitting) παρουσιάζεται όταν το μοντέλο είναι αρκετά απλό σε σχέση με την πολυπλοκότητα των δεδομένων. Αδυνατεί να μάθει τη σχέση μεταξύ εισόδων και εξόδων. Το μοντέλο και σε αυτήν την περίπτωση δεν γενικεύει καλά μακρία από τις τιμές παρατήρησης. **[1]**

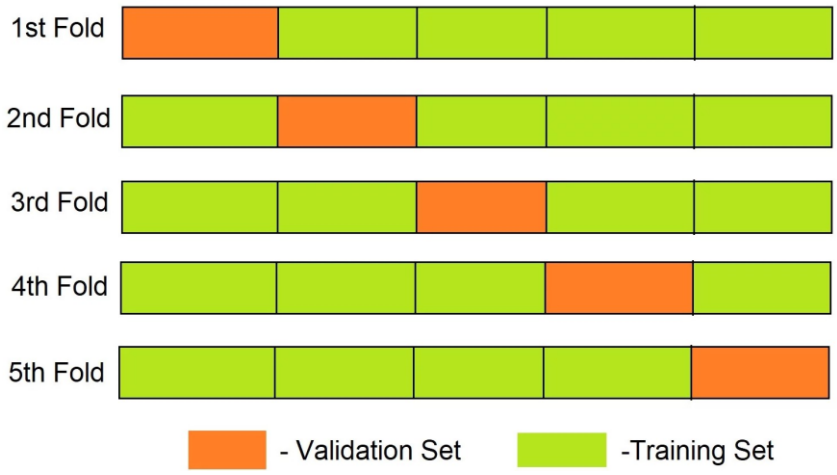
Διασταυρούμενη Επικύρωση (Cross-Validation)

Μια πολύ γνωστή μέθοδος για εκτίμηση ικανότητας γενίκευσης του μοντέλου είναι η διασταυρούμενη επικύρωση (Cross-Validation).

Μέθοδος k-fold Cross-Validation

Μια εκδοχή της μεθόδου είναι η k-fold cross-validation. Το σύνολο δεδομένων διαιρείται σε k υποσύνολα και η επιλογή τους είναι τυχαία. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης (validation set) και τα υπόλοιπα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης (training set). Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι

του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται k φορές, κάθε φορά με διαφορετικό σύνολο επικύρωσης και τα υπόλοιπα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου. **[13]**



Εικόνα 9: k-fold Cross-Validation για k=5 **[17]**

Πλεονεκτήματα [1]
• Καλή εκτίμηση της επίδοσης γενίκευσης του μοντέλου
• Αποτελεσματική μέθοδος και με λίγα δεδομένα

Μειονέκτημα [1]
• Μεγάλο υπολογιστικό κόστος λόγω της ανάγκης εκπαίδευσης του μοντέλου πολλές φορές

2. Αναλυτική

2.1 Επιχειρηματική Αναλυτική (Business Analytics) **[18]**

Η Επιχειρηματική Αναλυτική είναι η διαδικασία χρήσης στατιστικών μεθόδων, δεξιοτήτων, τεχνολογιών και πρακτικών για την ανάλυση ιστορικών δεδομένων και την απόκτηση νέων με σκοπό την υποστήριξη της λήψης αποφάσεων. Για τη διαχείριση των δεδομένων χρησιμοποιείται η επιχειρηματική ευφυΐα και αρκετές μεθοδολογίες, όπως η εξόρυξη δεδομένων και η στατιστική ανάλυση.

Διάφοροι τύποι επιχειρηματικής αναλυτικής βοηθούν στην ανάλυση και τη μετατροπή ακατέργαστων δεδομένων σε χρήσιμες πληροφορίες εντοπίζοντας και προβλέποντας τις τρέχουσες τάσεις για πιο έξυπνες επιχειρηματικές αποφάσεις.

Τύποι Επιχειρηματικής Αναλυτικής

Οι πιο δημοφιλείς τύποι επιχειρηματικής αναλυτικής είναι η περιγραφική, η διαγνωστική, η προβλεπτική και η προστακτική. Υπάρχει όμως και μια πέμπτη κατηγορία, η γνωστική ανάλυση. Είναι ένας τύπος επιχειρηματικής αναλυτικής που χρησιμοποιεί τη μηχανική μαθηση. Όλοι οι τύποι είναι αποτελεσματικοί ακόμα και αν χρησιμοποιηθούν ξεχωριστά, αλλά γίνονται ιδιαίτερα ισχυροί όταν χρησιμοποιούνται μαζί.

α) Περιγραφική αναλυτική (Descriptive analytics)

Αναλύει ιστορικά δεδομένα για να προσδιορίσει την αντίδραση μιας μονάδας σε ένα σύνολο μεταβλητών δεδομένων.

Περιλαμβάνει τα ακόλουθα βήματα:

- Επιλογή των μετρικών που θα χρησιμοποιηθούν για την αποτελεσματική αξιολόγηση της απόδοσης ανάλογα τον σκοπό
- Αναγνώριση των απαιτούμενων δεδομένων για την τρέχουσα κατάσταση
- Συλλογή και προεπεξεργασία δεδομένων
- Ανάλυση δεδομένων για την εύρεση μοτίβων και την αξιολόγηση της απόδοσης
- Παρουσίαση δεδομένων σε γραφήματα για να γίνουν κατανοητά από μη ειδικούς αναλυτές δεδομένων

Με λίγα λόγια, η περιγραφική αναλυτική λέει τι έγινε, αλλά όχι γιατί έγινε.

β) Διαγνωστική αναλυτική (Diagnostic αnalytics)

Η διαγνωστική αναλυτική δίνει την απάντηση στο ερώτημα γιατί έγινε. Χρησιμοποιώντας διαδοχικές αναλύσεις και εξόρυξη δεδομένων μπορεί να ανακαλύψει νέα δεδομένα και να βρει συσχετίσεις ώστε να γίνουν κατανοητοί οι λόγοι για κάποια συγκεκριμένα αποτελέσματα σχετικά με τα οικονομικά, το μάρκετινγκ, την κυβερνοασφάλεια και άλλους τομείς.

γ) Προγνωστική αναλυτική (Predictive analytics)

Λαμβάνει υπόψη τις τάσεις των ιστορικών δεδομένων για να προβλέψει την πιθανότητα συγκεκριμένων μελλοντικών αποτελεσμάτων. Χρησιμοποιεί τεχνικές όπως η εξόρυξη δεδομένων, οι αλγόριθμοι μηχανικής μάθησης και η στατιστική μοντελοποίηση.

δ) Προστακτική αναλυτική (Prescriptive analytics)

Δημιουργεί προτάσεις για την αντιμετώπιση παρόμοιων καταστάσεων στο μέλλον με βάση προηγούμενες. Χρησιμοποιεί εργαλεία όπως η στατιστική και οι αλγόριθμοι μηχανικής μάθησης για τα διαθέσιμα δεδομένα. Τέλος, παρέχει πληροφορίες σε αυτό που μπορεί να συμβεί, πότε και γιατί.

ε) Γνωστική αναλυτική (Cognitive analytics)

Συνδυάζει την τεχνητή νοημοσύνη και την ανάλυση δεδομένων. Εξετάζει τα διαθέσιμα δεδομένα και ανακαλύπτει τις βέλτιστες λύσεις για τα ερωτήματα που τίθενται. Με τη χρήση επεξεργασίας φυσικής γλώσσας και μηχανικής μάθησης μπορεί να κατανοτήσει μη δομημένα δεδομένα, όπως κείμενο, φωνή και εικόνες.

Οι τύποι analytics που θα χρησιμοποιηθούν στην κάθε περίπτωση εξαρτώνται από το πρόβλημα που πρέπει να λυθεί και τα διαθέσιμα δεδομένα.

Παραδείγματα των τύπων αναλυτικής στο ποδόσφαιρο

• **Περιγραφική Αναλυτική**
Ανάλυση παλαιότερων δεδομένων για την εύρεση μοτίβων όπως η απόδοση της ομάδας σε συγκεκριμένες αγωνιστικές και οικονομικές συνθήκες

• **Διαγνωστική αναλυτική**
Ανάλυση δεδομένων για την εύρεση αιτιών συγκεκριμένων προβλημάτων όπως τραυματισμούς και οικονομικές απώλειες

• **Προγνωστική αναλυτική**
Χρήση στατιστικών μοντέλων και μηχανικής μάθησης για πρόβλεψη μελλοντικών αποτελεσμάτων όπως η απόδοση ενός παίκτη ή τα έσοδα του συλλόγου

• **Προστακτική αναλυτική**
Παροχή προτάσεων όπως προσαρμογή της στρατηγικής του αγώνα ή βελτιστοποίησης των οικονομικών

• **Γνωστική αναλυτική**
Χρήση μηχανικής μάθησης για τη κατανόηση μη δομημένων δεδομένων όπως κείμενο, φωνή και εικόνες. Στο ποδόσφαιρο μη δομημένα δεδομένα μπορεί να είναι αναφορές scouting, δημοσιεύσεις στα μέσα κοινωνικής δικτύωσης και ειδήσεις για να κατανοηθεί η απόδοση ενός παίκτη ή η διάθεση των φιλάθλων.

2.2 Αναλυτική Ποδοσφαίρου (Football Analytics)

Όπου υπάρχουν δεδομένα, μπορούν να χρησιμοποιηθούν μοντέλα μηχανικής μάθησης και το ποδόσφαιρο παράγει τεράστιο όγκο δεδομένων. Οι ομάδες προσλαμβάνουν αναλυτές για την οργάνωση και την ερμηνεία αυτών των δεδομένων.

Η ανάλυση του ποδοσφαίρου έχει εξελιχθεί και δεν δίνεται σημασία μόνο από τα παραδοσιακά στατιστικά. Πλέον, χρησιμοποιούνται όλο και περισσότερο προηγμένες μετρικές με κυριότερη τα xGoals ή αλλίως Expected Goals.

Τι είναι τα xGoals; [19], [20], [21]

Είναι μονάδα μέτρησης της πιθανότητας του σουτ να μπει γκολ. Άρα, παίρνει τιμές από 0 έως 1. Αξιολογεί την ποιότητα του σουτ υπολογιζοντας την πιθανότητα αυτό το σουτ να γίνει γκολ. Ουσιαστικά, ποσοτικοποιεί την ποιότητα του σουτ.

Παρουσιάστηκε για πρώτη φορά το 2012 από μια εταιρεία ανάλυσης δεδομένων, την Opta. Πλέον, είναι ένα από τα βασικά στατιστικά ενός αγώνα ποδοσφαίρου. Είναι ένας σημαντικός δείκτης για την επιθετική λειτουργία μιας ομάδας, αλλά και την αποτελεσματικότητά της. Υπολογίζεται ότι έχουν αξιολογηθεί πάνω από 2.5 εκατομμύρια σουτ από 66.000 διαφορετικούς παίκτες. Ενδεικτικά, τα πέναλτι αξιολογούνται με 0.79 xGoals.

Εξαρτάται από πολλούς παράγοντες όπως:
• Η απόσταση από την εστία
• Η γωνία του σουτ σε σχέση με την εστία
• Τρόπος εκτέλεσης: κεφάλι ή πόδι
• Η πάσα πριν το σουτ: προωθημένη πάσα ή σέντρα
• Η ίδια η φάση: κανονική ροή παιχνιδιού, αντεπίθεση ή στατική φάση
• Απόσταση αμυντικών από τον εκτελεστή
• Τοποθέτηση τερματοφύλακα
• Επαφές εκτελεστή πριν το σουτ: μονοκόμματο σουτ ή τρίπλα πριν από αυτό

Δεν υπάρχει ένα συγκεκριμένο μοντέλο xGoals. Πολλές εταιρείες ανάλυσης δεδομένων έχουν τον δικό τους αλγόριθμο για την αξιολόγηση του σουτ δίνοντας διαφορετική σημασία στον κάθε παράγοντα. Βέβαια, οι αποκλίσεις δεν είναι σημαντικά μεγάλες.

xAssists [22], [23]
Μονάδα μέτρησης της πιθανότητας μιας πάσας να γίνει ασίστ, δηλαδή να γίνει η πάσα πριν το γκολ. Άρα, παίρνει τιμές από 0 έως 1. Έχει σκοπό να αξιολογήσει την ποιότητα της ευκαιρίας που δημιουργεί ένας παίκτης για τους συμπαίκτες του. Έχουν αναλυθεί εκατομμύρια πάσες για χιλιάδες παιχνίδια.

<u>Εξαρτάται από 5 παράγοντες:</u>
• Από πού έγινε η πάσα;
• Τύπος πάσας (προωθημένη, σέντρα, κεφαλιά)
• Φάση: κανονική ροή παιχνιδιού, αντεπίθεση ή στατική φάση
• Πού έλαβε την πάσα ο εκτελεστής;
• Απόσταση πάσας

Αξίζει να σημειωθεί ότι δεν εξαρτάται από το αν θα σουτάρει τελικά ο παίκτης που θα παραλάβει την πάσα.

Από τα xGoals προκύπτουν άλλες δύο μετρικές.
Η πρώτη είναι τα xGoalsAgainst, δηλαδή τα xGoals που δέχεται μια ομάδα από την αντίπαλο της.

Η δεύτερη είναι η xGoalsDifference, δηλαδή η διαφορά των xGoals από τα xGoalsAgainst. Αν μια ομάδα έχει xGoalsDifference > 0 έχει περισσότερες πιθανότητες να κερδίσει.

Πέρα από τις προηγμένες μετρικές, υπάρχουν και τα βασικά στατιστικά που προκύπτουν μετά από έναν αγώνα ποδοσφαίρου.

Στατιστικά ομάδων

• Γκολ
• Κατοχή μπάλας
• Σουτ
• Σουτ στην εστία
• Κόρνερ
• Επιτυχημένες πάσες
• Οφσαίντ
• Πλάγια άουτ
• Φάουλ κατά
• Κίτρινες κάρτες
• Κόκκινες κάρτες

Στατιστικά παικτών
• Γκολ
• Ασίστ
• Σουτ
• Σουτ στην εστία
• Επιτυχημένες πάσες
• Ανακτήσεις κατοχής μπάλας (κλεψίματα)
• Διανυθείσα απόσταση (σε χιλιόμετρα)

Στατιστικά τερματοφύλακα
• Σουτ κατά
• Σουτ στην εστία κατά
• Αποκρούσεις
• Γκολ κατά

Όπως αναφέρθηκε και στην εισαγωγή ένας ποδοσφαιρικός σύλλογος είναι μια επιχείρηση. Οπότε είναι πολύ σημαντικά και τα δεδομένα εκτός αγωνιστικού χώρου, δηλαδή τα οικονομικά δεδομένα.

Πίνακας 3: Οι πηγές εσόδων και εξόδων ενός ποδοσφαιρικού συλλόγου **[24]**

Έσοδα (+)	Έξοδα (-)
Επίδοση σε διοργανώσεις	Μισθοί ποδοσφαιριστών
Πώληση ποδοσφαιριστή	Αγορά ποδοσφαιριστή
Τηλεοπτικά δικαιώματα	Αμοιβές προσωπικού → Προπονητές, αναλυτές κλπ.

Εισιτήρια και παροχές γηπέδου	Εκτός έδρας αγώνες (οδοιπορικά/αεροπορικά)
Χορηγοί	Φόροι
Εισιτήρια διαρκείας και συνδρομές	Λειτουργικά κόστη τμημάτων υποδομής
Εμπορεύματα	Κόστος παραγωγής εμπορευμάτων
	Συντήρηση και οργάνωση γηπέδου
	Ιατρικά κόστη
	Έξοδα διαφήμισης και μάρκετινγκ
	Κυρώσεις

Βεβαίως, αυτά εξαρτώνται από τις διοργανώσεις που παιζει η κάθε ομάδα, την τοποθεσία και το μέγεθός της σαν σύλλογος. Κάποιες παίρνουν σημαντική ενίσχυση από τα τηλεοπτικά δικαίωματα, άλλες βασίζονται στην πώληση εισιτηρίων και άλλες βασίζονται στις ακαδημίες.

3. Συλλογή και επεξεργασία δεδομένων
3.1 Συλλογή δεδομένων

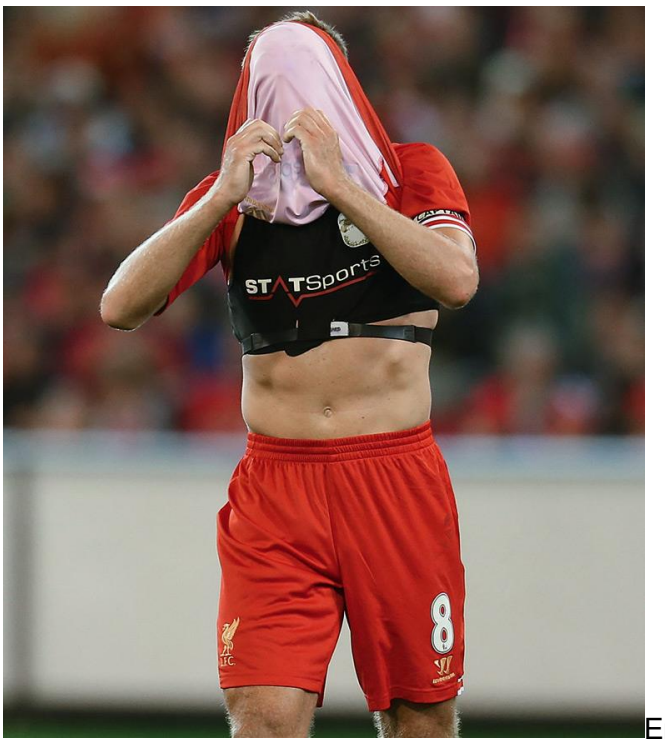
3.1.1 Αναλυτές

Η συλλογή δεδομένων πραγματοποιείται από μια ομάδα αναλυτών. Γίνεται χρήση ενός συστήματος συλλογής βασισμένο σε βίντεο για να αναλυθεί κάθε ενέργεια κατά τη διάρκεια του αγώνα. Οι αναλυτές έχουν πλήρη εικόνα του γηπέδου μέσω των υπολογιστών τους χάρη στις πολλές κάμερες που έχουν εγκατασταθεί στο γήπεδο. Αυτές οι κάμερες καλύπτουν κάθε γωνία του γηπέδου καθιστώντας δυνατή την παρακολούθηση κάθε κίνησης του παίκτη κατά τη διάρκεια του αγώνα. Με τη χρήση λογισμικού οι εικόνες από τις κάμερες μετατρέπονται σε data points. Πριν τον αγώνα υπάρχει χρόνος για τους αναλυτές να ελέγξουν τις αρχικές συνθέσεις για κάθε ομάδα και να βάλουν δείκτες για κάθε παίκτη. Μόλις αντιστοιχηθεί ο δείκτης, ο αναλυτής μπορεί να παρακολουθεί τις ενέργειες του παίκτη. Οι περισσότερες ενέργειες συλλέγονται χειροκίνητα, αλλά με τη βοήθεια της μηχανικής μάθησης και της μηχανικής όρασης, μπορεί να αυτομαποιηθεί η διαδικασία συλλογής δεδομένων από το βίντεο. Σε αυτό μπορεί να βοηθήσει και το επίπεδο τεχνολογίας που χρησιμοποιείται στο γήπεδο, γιατί η κίνηση των παικτών μπορεί να αυτοματοποιηθεί μέσω της χρήσης πολλαπλών HD καμερών. Οι διακοπές που συμβαίνουν σε έναν αγώνα είναι σωτήριες για τους αναλυτές καθώς τους δίνεται η δυνατότητα να γυρίσουν πίσω το βίντεο και να ελέγξουν τα δεδομένα τους. Τις περισσότερες φορές υπάρχει και ένας αναλυτής ελέγχου ποιότητας που είναι υπεύθυνος για να επιβεβαίωσει ότι τα δεδομένα που συλλέγουν οι αναλυτές είναι ακριβή. Επίσης, είναι υπεύθυνος να λάβει μια απόφαση σε μια δύσκολη φάση και οι αναλυτές έχουν διαφορετικά αποτελέσματα μεταξύ τους. Μόλις ολοκληρωθεί ο αγώνας, γίνεται περαιτέρω ανάλυση για να διασφαλιστεί ότι τα δεδομένα είναι όσο το δυνατόν ακριβή. **[25]**


Εικόνα 10: Οι αναλυτές εν δράσει **[26]**

3.1.2 Γιλέκα με GPS

Τα γιλέκα με GPS που φοράνε οι ποδοσφαιριστές κατά τη διάρκεια του αγώνα δίνουν χρήσιμες πληροφορίες σχετικά με ταχύτητα, απόσταση, τοποθέτηση και καρδιακό παλμό. Γνωρίζοντας αυτά τα δεδομένα ο προπονητής, μπορεί να διαχειριστεί καλύτερα τους ποδοσφαιριστές του και να μη διακινδυνεύσει τραυματισμούς. **[27], [28]**



Εικόνα 11: Γιλέκο με GPS που φοράνε οι ποδοσφαιριστές κατά τη διάρκεια του αγώνα **[29]**

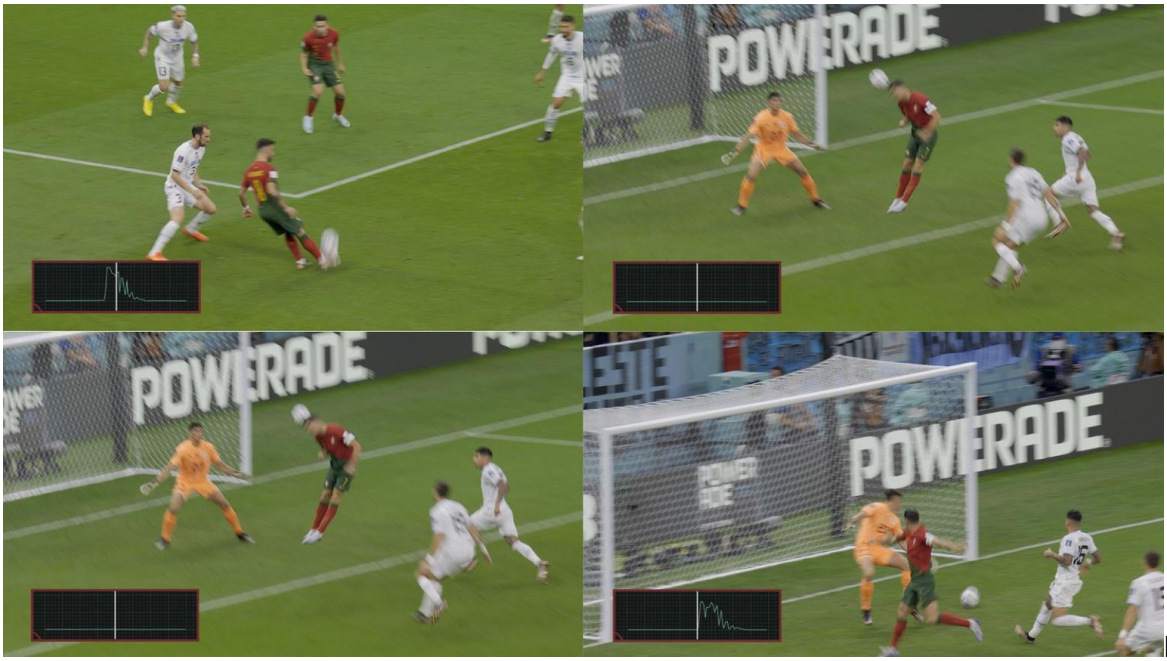
3.1.3 Αισθητήρες μπάλας

Για το Παγκόσμιο Κύπελλο 2022, η Adidas κατασκεύασε μια μπάλα που περιέχει έναν υψηλής ακρίβειας αισθητήρα αδρανειακής μέτρησης (IMU) 500Hz και την Connected Ball τεχνολογία της Adidas. Παρέχει πληροφορίες για τη θέση, την ταχύτητα, το ρυθμό περιστροφής (spin) και την επιτάχυνση της μπάλας. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν άμεσα από την τεχνολογία αυτόματου οφσάιντ και από το Video Assistant Referee (VAR), τα οποία θα αναλυθούν στο επόμενο κεφάλαιο. **[30], [31], [32]**

Εικόνα 12: Η μπάλα του Παγκοσμίου Κυπέλλου 2022 **[30]**

Η τεχνολογία της βοήθησε σε 2 πολύ οριακές φάσεις.

1. Για τη 2η αγωνιστική της φάσης των ομίλων Πορτογαλία και Ουρουγουάη βρίσκονται αντιμέτωπες και μέχρι το 54ο λεπτό το σκόρ είναι 0-0. Τότε, η Πορτογαλία κάνει το 1-0 έπειτα από σέντρα του Bruno Fernandes και ένα πιθανό άγγιγμα με το κεφάλι από τον Cristiano Ronaldo. Έγινε μέγαλη κουβέντα για το ποιος έβαλε το γκολ, παρόλο που η Πορτογαλία κέρδισε 2-0. Η FIFA ήρθε να δώσει την απάντηση λέγοντας πως από τις μετρήσεις του αισθητήρα της μπάλας δεν καταγράφηκε κάποια δύναμη από το κεφάλι του Cristiano Ronaldo. **[33]**



Εικόνα 13: Οι μετρήσεις του αισθητήρα στο γκολ του Bruno Fernandes **[33]**

2. Για την τελευταία αγωνιστική της φάσης των ομίλων, Ιαπωνία και Ισπανία είναι ισόπαλες με σκορ 1-1. Στο 51ο λεπτό, η Ιαπωνία πετυχαίνει 2ο γκολ μετά από υπερπροσπάθεια του Kaoru Mitoma να προλάβει την μπάλα πριν βγει έξω, να δίνει πάσα στον Ao Tanaka και αυτός να σκοράρει για το 2-1. Οι Ισπανοί διαμαρτύρονται έντονα ότι η μπάλα έχει περάσει εκτός αγωνιστικού χώρου πριν την πάσα του Mitoma και από τις επαναληπτικές προβολές φαίνεται να έχουν δίκιο. Ο διαιτητής συμβουλεύεται το Video Assistant Referee. Λίγα λεπτά αργότερα, ανακοινώνει ότι το γκολ μετράει κανονικά. Η μπάλα δεν πέρασε ολόκληρη εκτός αγωνιστικού χώρου για 1,88 χιλιοστά. Στις επαναληπτικές προβολές ξεγελούσε το σφαιρικό σχήμα της μπάλας σε συνδυασμό με τη γωνία της κάμερας. Χρειαζόταν η κάτοψη για να επιβεβαιώσει την απόφαση του διαιτητή. Φυσικά, ο διαιτητής συμβουλεύτηκε τις μετρήσεις του αισθητήρα της μπάλας. **[34]**


Εικόνα 14: Το οριακό γκολ της Ιαπωνίας **[34]**

3.2 Επεξεργασία δεδομένων

Για την επεξεργασία και την αποθήκευσή των δεδομένων τους οι ομάδες συνεργάζονται με εταιρείες που προσφέρουν cloud υπηρεσίες, όπως η Amazon Web Services ή εταιρείες παροχής στατιστικών και ανάλυσης δεδομένων όπως η Opta.

Τι είναι το Cloud Computing;

Είναι ένα ενιαίο σύνολο υπολογιστικών πόρων που προσφέρεται με διάφορες μορφές προκειμένου να υποστηρίξει διαφόρων ειδών υπηρεσίες και εφαρμογές. Το σύνολο των πόρων που χρησιμοποιούνται για την κάλυψη των αναγκών των εφαρμογών μπορεί να αυξομειώνεται, να ενεργοποιείται κατ' απαίτηση, να ελευθερώνεται ή να δεσμεύεται με ευέλικτο τρόπο. Χρησιμοποιείται ως βάση για την υποστήριξη νέων τεχνολογιών όπως η τεχνητή νοημοσύνη, οι πλατφόρμες IoT και εργαλεία Μηχανικής Μάθησης. **[35]**

Χαρακτηριστικά του Cloud Computing [35], [36]

1. Χρησιμοποίηση κατ' απαίτηση

• Ο χρήστης χρησιμοποιεί οποτεδήποτε θέλει τους πόρους που έχει δεσμεύσει στο νέφος για την υποστήριξη των εφαρμογών και των υπηρεσιών του

• Η χρήση των πόρων μπορεί να πραγματοποιηθεί και μέσω αυτοματοποιημένων διαδικασιών χωρίς να χρειάζεται ανθρώπινη παρέμβαση.

2. Υψηλή προσβασιμότητα

• Οι υπηρεσίες cloud είναι ευρέως προσπελάσιμες μέσω του διαδικτύου από το χρήστη καθώς επίσης και οι εφαρμογές και υπηρεσίες που αναπτύσσονται από αυτόν.

3. Συνεκμετάλλευση πόρων

• Δίνεται η δυνατότητα σε διαφορετικούς χρήστες να κάνουν χρήση των ίδιων πόρων του νέφους για την υποστήριξη διαφορετικών απομονωμένων εφαρμογών.

4. Ελαστικότητα
• Ελαστικότητα είναι η αυτοματοποιημένη δυνατότητα του νέφους να κλιμακώνει τους πόρους ανάλογα με τις ανάγκες.
• Καθορισμός του κόστους με βάση τη ζήτηση

5. Μετρούμενη χρησιμοποίηση πόρων
• Παρακολούθηση της χρησιμοποίησης των πόρων από τους χρήστες. Έτσι, μπορεί να γίνει η χρέωση με βάση την χρήση των πόρων ανάλογα το μέγεθός τους αλλά και την διάρκεια που χρησιμοποιήθηκαν.
• Εξέταση «περίεργης» χρήσης σε περίπτωση ζητήματος ασφάλειας

6. Ανθεκτικότητα
• Σε περίπτωση προβλήματος, γίνεται μετάβαση σε άλλους πόρους.

Τι προσφέρει η τεχνολογία cloud συγκριτικά με τις παλιές υποδομές; [35]
• Πιο εύκολη διαχείριση πόρων
• Μεγαλύτερη αποδοτικότητα της υποδομής
• Μεγαλύτερη διαθεσιμότητα των υπηρεσιών
• Πιο γρήγορη και ανώδυνη ανάκαμψη της υποδομής σε περίπτωση προβλήματος
• Μεγαλύτερη επεκτασιμότητα και ελαστικότητα στην χρήση πόρων
• Απομόνωση των υπηρεσιών και των εφαρμογών που υποστηρίζονται από την υποδομή.

Από τα παραπάνω προκύπτουν κάποια πλεονεκτήματα χρήσης και για τον παρόχο αλλά και για το χρήστη.

Πλεονεκτήματα χρήσης τεχνολογίας νέφους για τον πάροχο [35]
• Μεγαλύτερη αποδοτικότητα της υπάρχουσας υποδομής
• Μειωμένο ρίσκο αγοράς νέου εξοπλισμού που δεν θα είναι πλήρως αποδοτικός.
• Πιο εύκολη διαχείριση των πόρων
• Μεγαλύτερη διαθεσιμότητα των προσφερόμενων υπηρεσιών
• Εύκολος έλεγχος για προβλήματα ασφάλειας

Πλεονεκτήματα χρήσης τεχνολογίας νέφους για το χρήστη [35]
• Χρήση έτοιμων υποδομών και υπηρεσιών
• Γρήγορη ανάπτυξη λύσεων
• Εύκολη μεταφορά λογισμικού και εργαλείων στο cloud
• Μειωμένο λειτουργικό κόστος καθώς δεν χρειάζεται η δημιουργία και η συντήρηση υποδομής.
• Υψηλή διαθεσιμότητα των εφαρμογών
• Ομαλή λειτουργία των προϊόντων ακόμα και σε περιόδους υψηλής ζήτησης

Η ομαλή λειτουργία του cloud εξασφαλίζεται από τη σωστή λειτουργία και αλληλεπίδραση των υπηρεσιών που το αποτελούν. Η διαχείριση, η πρόσβαση, ο τρόπος διάθεσης των πόρων και άλλες λειτουργίες είναι αρμοδιότητα των cloud υπηρεσιών. Όπως το λογισμικό αποτελείται από μικρές υπηρεσίες έτσι και το νέφος αποτελείται από νεφοϋπολογιστικές υπηρεσίες με διαφορετικές αρμοδιότητες. **[35]**

Μοντέλα υπηρεσιών cloud
Τα μοντέλα υπηρεσιών καθορίζουν τον τρόπο χρήσης των πόρων του cloud. Επιπλέον, καθορίζουν τις ελευθερίες και τους περιορισμούς του χρήστη.

Τα βασικά τρία μοντέλα είναι: **[35], [36]**
1. Υποδομή ως Υπηρεσία (Infrastructure as a Service – IaaS)
2. Πλατφόρμα ως Υπηρεσία (Platform as a Service – PaaS)
3. Λογισμικό ως Υπηρεσία (Software as a Service – SaaS)

Αναλυτικότερα για το IaaS: **[35], [36]**
• Δέσμευση των πόρων σε επίπεδο vCPU, μνήμης, αποθήκευσης και δικτύου για την δημιουργία εικονικών μηχανών με προδιαγραφές που ορίζει ο χρήστης.
• Δυνατότητα επιλογής λειτουργικού συστήματος και έλεγχος σε επίπεδο εικονικής μηχανής
• Ο χρήστης έχει πλήρη ελευθερία μέσα στην εικονική μηχανή αλλά και απόλυτη ευθύνη για την εύρυθμη λειτουργία του συστήματος του.
• Είναι ιδανικό για συστήματα υποστήριξης εφαρμογών και έρευνα.

Δημοφιλείς πάροχοι και λογισμικά για παροχή IaaS είναι η Amazon Web Services (AWS), η Microsoft Azure και η Oracle Cloud Infrastructure.

Amazon Web Services (AWS)
Είναι ένα είδος public cloud που προσφέρει υπηρεσίες IaaS και PaaS. Χρησιμοποιείται για ανάπτυξη εφαρμογών από διάφορες εταιρείες και οργανισμούς καθώς διαθέτει τεράστια ποικιλία υπηρεσιών για τα περισσότερα σενάρια χρήσης. Προσφέρει ένα μοντέλο πληρωμής ανάλογα την χρήση, αλλά παρέχει και κάποιες δωρεάν υπηρεσίες. Τέλος, δεν απαιτεί εξειδίκευση για τον χρήστη και όλα τα θέματα διαχείρισης και συντήρησης σχετικά με την υποδομή τα αναλαμβάνει η Amazon. Προσφέρει μια εντυπωσιακή γκάμα εφαρμογών και

υπηρεσιών και έχει κερδίσει εταιρείες κολοσσούς σε διαφορετικούς τομείς, όπως Coca-Cola, Netflix, Heineken, NASA, Formula 1, McDonald's, Adidas και Nike. **[37], [38]**

Microsoft Azure
Είναι μια πλατφόρμα cloud που έχει δημιουργήσει η Microsoft. Προσφέρει πρόσβαση, διαχείριση και ανάπτυξη εφαρμογών και υπηρεσιών μέσω ενός παγκοσμίου δικτύου κέντρων δεδομένων. Υποστηρίζει πολλές γλώσσες προγραμματισμού και εργαλεία. Έχει πελάτες όπως ο ΟΠΑΠ, η Stoiximan, η Bridgestone, η Eurobank και η Mitsubishi. **[39], [40]**

Oracle Cloud Infrastructure
Είναι μια υπηρεσία cloud που προσφέρεται από την Oracle. Παρέχει διακομιστές, αποθήκευση, δίκτυο, εφαρμογές και υπηρεσίες μέσω ενός παγκοσμίου δικτύου κέντρων δεδομένων που το διαχειρίζεται η ίδια η Oracle. Η εταιρεία παρέχει όλα τα παραπάνω κατόπιν ζήτησης μέσω διαδικτύου. Υποστηρίζει γλώσσες προγραμματισμού, βάσεις δεδομένων, εργαλεία και εφαρμογές ανοικτού κώδικα. Έχει πάνω από 430.000 πελάτες. Κάποιοι από αυτούς είναι η Vodafone, η MasterCard, η Red Bull Racing αλλά ακόμα και ο ΕΟΠΠΥ. **[41], [42]**

Οι εφαρμογές του Cloud Computing στο ποδόσφαιρο

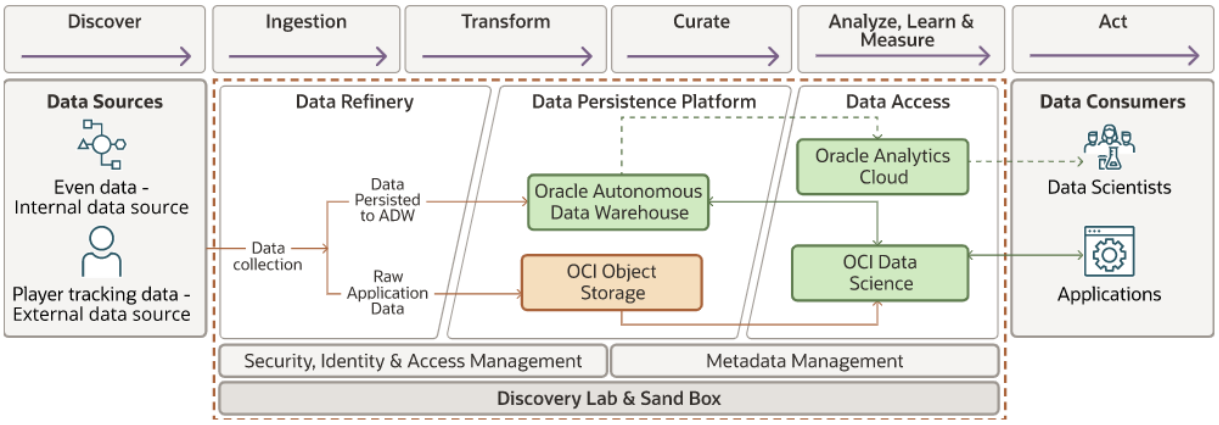
Φαίνεται ότι την αρχή σε αυτό το εγχείρημα την έχουν κάνει κυρίως ισπανικές ομάδες.
Η Barcelona έχει συνεργαστεί με την AWS για θέματα διαχείρισης ηλεκτρονικού εμπορίου, κοινωνικών μέσων και κινητών εφαρμογών. **[43]**

Από την άλλη, η Ρεάλ Μαδρίτης χρησιμοποιεί Microsoft Azure για να εμπλακεί άμεσα με τους εκατομμύρια φιλάθλους της σε όλο τον κόσμο. Έτσι, μπορεί να τους προσφέρει μια καλύτερη εμπειρία, αλλά και να αναλύει δεδομένα και πληροφορίες για να προσαρμόσει τις διαφημιστικές καμπάνιες της ώστε να αυξήσει τα έσοδα. Με τη λύση αυτή, ο σύλλογος μπορεί να ανακαλύψει προσωπικές προτιμήσεις και να το προωθήσει στον φίλαθλο μέσω εφαρμογής στο κινητό ή να δώσει μια προσφορά στο κατάστημα φιλάθλων αν ο φίλαθλος πραγματοποιήσει check-in στο γήπεδο μέσω κωδικού QR. Μελλοντικά, η Ρεάλ Μαδρίτης σκοπεύει να χρησιμοποιήσει μηχανική μάθηση με τη βοήθεια της Microsoft για να μπορεί να παρακολουθεί την κατάσταση των ποδοσφαιριστών αλλά και να βελτιστοποιήσει τις τιμές των εισιτηρίων, ώστε να προσελκύσει περισσότερους φιλάθλους, αλλά ταυτόχρονα να είναι και όσο το δυνατόν περισσότερο κερδοφόρο για το σύλλογο. **[44], [45]**

Άλλη μια ομάδα που χρησιμοποιεί Microsoft Azure είναι η Valencia. Κατάφερε να κυκλοφορήσει μια εφαρμογή που επιτρέπει στους φιλάθλους να παραγγέλνουν φαγητό και ποτά από τις θέσεις τους χωρίς να περιμένουν στην ουρά στο κυλικείο. Επίσης, κατά τη διάρκεια της πανδημίας, δημιούργησε έναν αλγόριθμο πρόβλεψης της πιθανότητας της προσέλευσης των φιλάθλων στο γήπεδο με βάση την προηγούμενη συμπεριφορά. Αν κάποιος φίλαθλος δεν ερχόταν για 2 συνεχόμενους αγώνες, η Βαλένθια έστελνε ένα email αν είναι όλα εντάξει. Έτσι, κατάφερε να αυξήσει την προσέλευση κατά 20%. **[46]**

Η Espanyol χρησιμοποιεί Oracle Financials Cloud επιτρέποντας στους επενδυτές να έχουν πρόσβαση σε οικονομικά δεδομένα σε πραγματικό χρόνο από οπουδήποτε. Μείωσε τους χρόνους διεκπεραίωσης πληρωμών κατά 80% και τα σφάλματα σε αναφορές κατά 25%

[47]. Αλλά και ομάδες εκτός Ευρώπης όπως η Seattle Sounders που αγωνίζεται στο πρωτάθλημα των ΗΠΑ, συνεργάζεται με την Oracle Cloud για την οργάνωση των πολυμέσων της που περιλαμβάνουν στιγμιότυπα πολλών σεζόν αλλά και στοιχεία μάρκετινγκ. Οι αναλυτές της Seattle και της Oracle δημιούργησαν ένα μοντέλο μηχανικής μάθησης χρησιμοποιώντας δεδομένα παρακολούθησης των παικτών. Πιο συγκεκριμένα, το μοντέλο αφορά τις πάσες που μπορούν να φέρουν σε μειονεκτική θέση την αμυντική γραμμή της αντίπαλης ομάδας. Το μοντέλο αυτό βοήθησε την Seattle Sounders να αξιολογήσει το στυλ παιχνιδιού κάθε παίκτη της ομάδας απέναντι σε κάθε αντίπαλο και να δημιουργήσει πιο αποτελεσματικές στρατηγικές παιχνιδιού. **[48]**



Εικόνα 15: Το μοντέλο της Seattle Sounders σε συνεργασία με την Oracle Cloud **[48]**

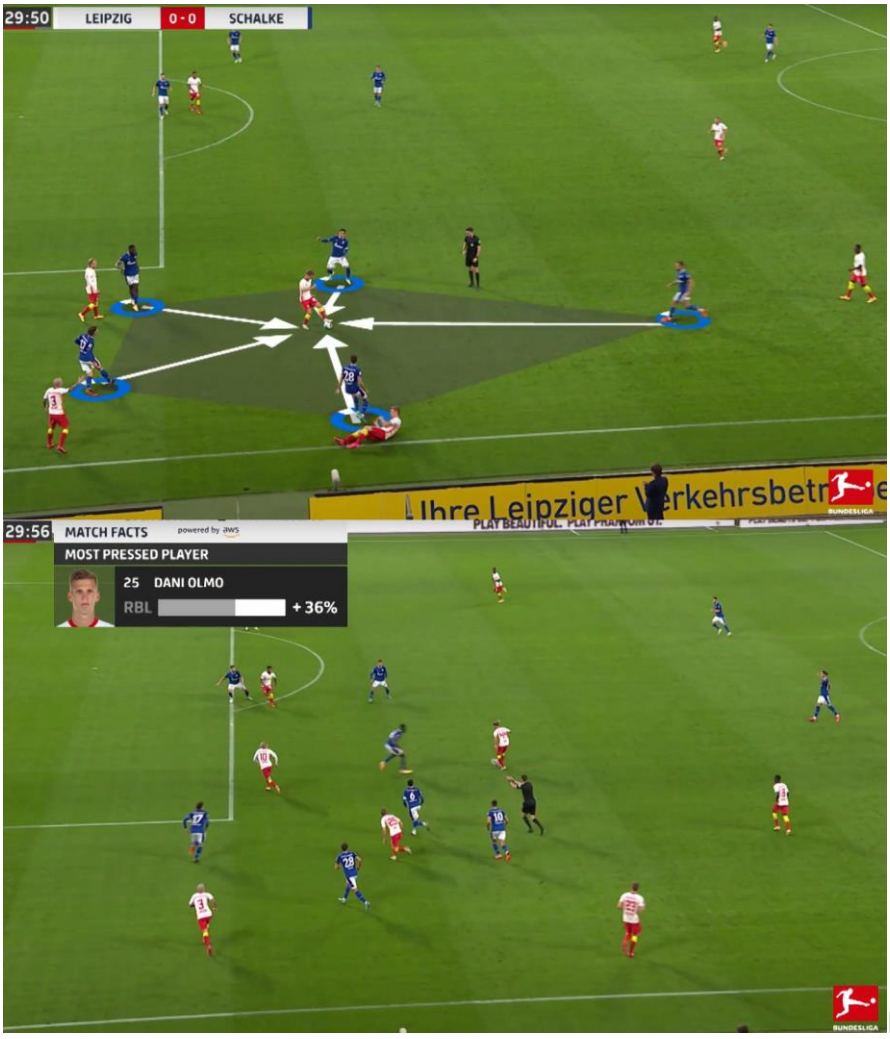
Εκτός από τις ομάδες, είναι εντυπωσιακό ότι ολόκληρα πρωταθλήματα χρησιμοποιούν cloud υπηρεσίες για να ενισχύσουν το προϊόν τους.

Bundesliga Insights

Η Bundesliga (Πρωτάθλημα Γερμανίας) σε συνεργασία με την AWS έχουν δημιουργήσει το Bundesliga Insights. Παρέχονται 14 πληροφορίες κατά την διάρκεια ενός αγώνα που κάνουν καλύτερη την εμπειρία του τηλεθεατή. **[49], [50]**

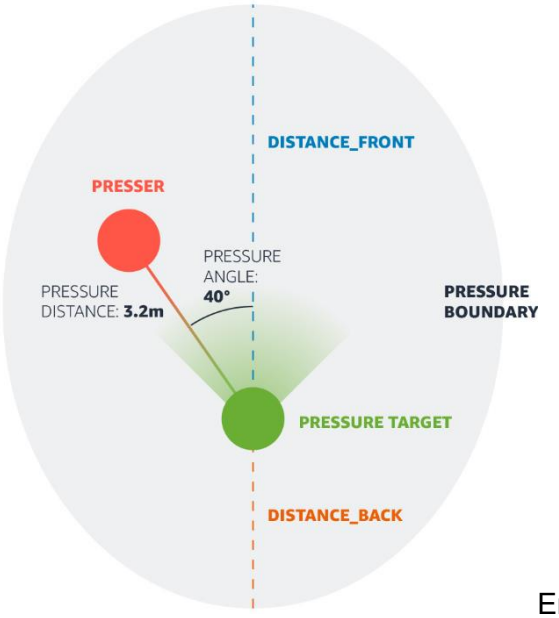
1. Most Pressed Player

Δείχνει πόσο συχνά και σε τι ποσοστό ένας ποδοσφαιριστής αντιμετωπίζει πίεση από τους αντιπάλους του μετρώντας τον αριθμό των αντιπάλων του, την απόστασή τους καθώς και την κατεύθυνση κίνησής τους. Αυτό το προηγμένο στατιστικό συγκρίνει τον αριθμό καταστάσεων πίεσης που δέχεται ένας ποδοσφαιριστής σε σχέση με τους συμπαίκτες του βοηθώντας να προσδιοριστεί ποιοι ποδοσφαιριστές δέχονται περισσότερη πίεση. **[51]**



Εικόνα 16: Η AWS υπολόγισε ότι το ποσοστό για την παραπάνω φάση ισούται με 36%. **[51]**

Πώς υπολογίζεται αυτή η πίεση; **[51], [52]**



Εικόνα 17: Ζώνη πίεσης **[51]**

Pressure = $(1 - \frac{d}{L})^q$ * 100%

όπου,

L = Distance_Back + (Distance_Front – Distance_Back) (z^3 + 0.3z) / 1.3

d: Η απόσταση του ποδοσφαιριστή που πιέζει από τον ποδοσφαιριστή που έχει τη μπάλα.

q: Ο εκθέτης που ρυθμίζει πόσο γρήγορα διασπάται η πίεση με την απόσταση.

Distance_Back: Η μέγιστη απόσταση που ο ποδοσφαιριστής μπορεί να πιεστεί στην κατεύθυνση που δεν βλέπει.

Distance_Front: Η μέγιστη απόσταση που ο ποδοσφαιριστής μπορεί να πιεστεί στην κατεύθυνση που βλέπει.

$z = \frac{1-\cos\Theta}{2}$

$-180^o \leq \Theta \leq 180^o$

Τα Θ και L είναι πολικές συντεταγμένες.

Ωστόσο ο ποδοσφαιριστής μπορεί να πιέζεται από 2 ή περισσότερους παίκτες. Σε αυτές τις περιπτώσεις, αθροίζεται η πίεση από τον καθένα ξεχωριστά για τη συνολική τιμή πίεσης σε μια δεδομένη χρονική στιγμή. Έτσι, οι τιμές πίεσης μπορεί να υπερβαίνουν το 1.

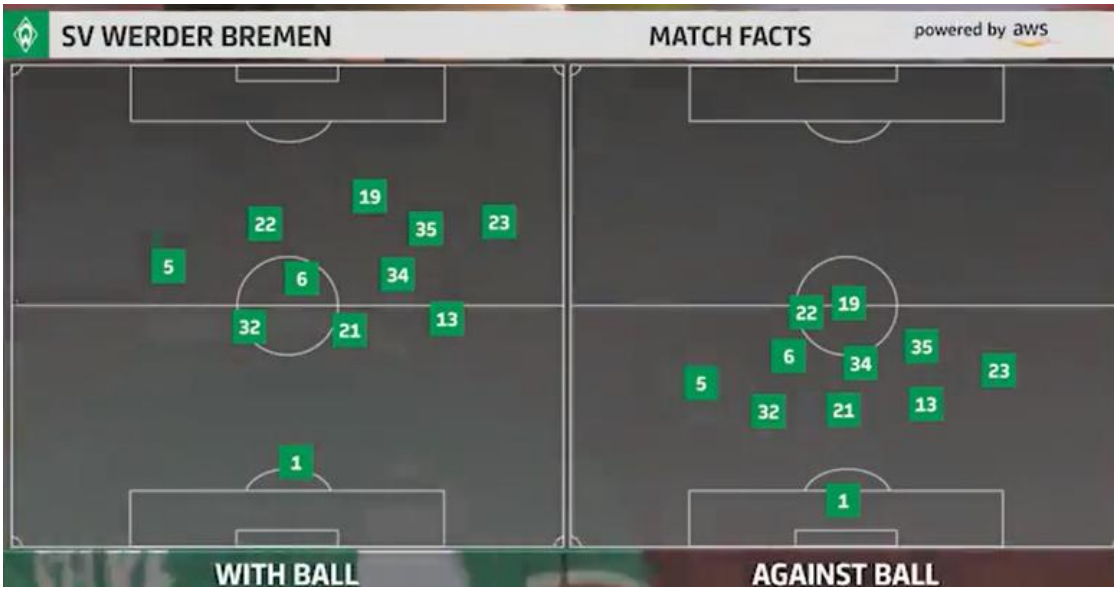
2. Attacking Zones

Οι ζώνες επίθεσης επιτρέπουν στους θεατές να δουν πού εστιάζουν οι ομάδες την επίθεσή τους για να δημιουργήσουν ευκαιρίες. Απεικονίζει τις περιοχές στις οποίες οι ομάδες εισέρχονται στο τελευταίο τρίτο του γηπέδου όταν επιτίθενται, χωρίζοντας αυτήν την περιοχή σε τέσσερις κάθετες ζώνες ίσου μεγέθους και δίνεται ένα ποσοστό για κάθε μία από αυτές. **[50], [53]**

Εικόνα 18: Παράδειγμα των Attacking Zones **[53]**

3. Average Positions
Παρέχει τις θέσεις των παικτών στο γήπεδο σε πραγματικό χρόνο. Τα δεδομένα υπολογίζονται χρησιμοποιώντας περίπου 3.6 εκατομμύρια data points ανά παιχνίδι.

4. Average Position: Trends
Μια βελτιωμένη εκδοχή του Average Positions. Επιτρέπει τη σύγκριση σχηματισμών με συγκεκριμένες τάσεις κατά τη διάρκεια του παιχνιδιού που μπορεί να προκύψουν μετά από κάποια γεγονότα, όπως μια κόκκινη κάρτα, ένα γκολ ή αν η ομάδα έχει την μπάλα ή όχι. **[49], [50]**

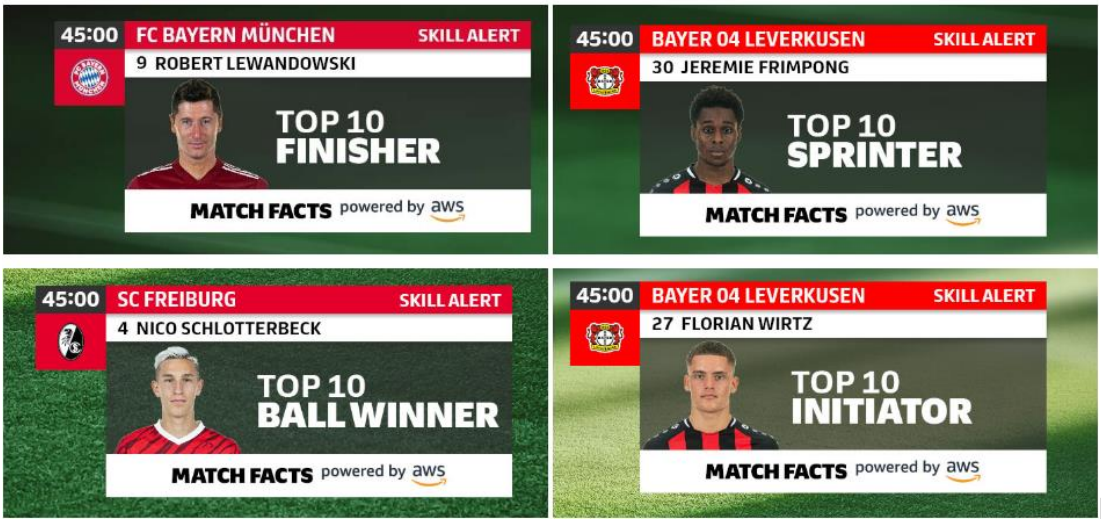


Εικόνα 19: Παράδειγμα του Average Position: Trends **[54]**

5. Skill

Κάθε ποδοσφαιριστής στην Bundesliga έχει την δική του ικανότητα που συμπληρώνει τους συμπαίκτες του και κάνει την ομάδα πιο δύνατη. Το Skill συνδυάζει και συγκρίνει αθροιστικά στατιστικά κάθε παίκτη για να αξιολογήσει τις ικανότητές του σε διαφορετικές κατηγορίες. Περιλαμβάνει τέσσερα προφίλ ποδοσφαιριστών.

• Finisher: Σκοράρει πολλά γκολ χωρίς να σπαταλάει ευκαιρίες.
• Sprinter: Οι πιο γρήγοροι ποδοσφαιριστές του πρωταθλήματος
• Initiator: Συγκεντρώνει πολλές ασιστ και περνάει δύσκολες πάσες.
• Ball Winner: Αναγκάζει τους αντιπάλους τους σε λάθη κερδίζοντας πολλές μονομαχίες είτε στο έδαφος είτε ψηλά αλλά και ανακόπτοντας αντίπαλες πάσες. **[50]**



Εικόνα 20: Τα 4 προφίλ ποδοσφαιριστών του insight Skill **[55]**

6. Goal Probability (xGoals)

Αξιολόγηση της πιθανότητας να πετύχει γκολ ένας ποδοσφαιριστής όταν σουτάρει από οποιαδήποτε θέση στο γήπεδο. Η πιθανότητα υπολογίζεται σε πραγματικό χρόνο για κάθε σουτ για να δώσει στους θεατές μια εικόνα για τη δυσκολία του σουτ και την πιθανότητα να γίνει γκολ. Η Bundesliga δίνει 77% πιθανότητες για το πέναλτι να γίνει γκολ. **[49], [50], [56]**

Εικόνα 21: Η παραπάνω φάση αξιολογείται με μόλις 4% πιθανότητα να γίνει γκολ. **[57]**

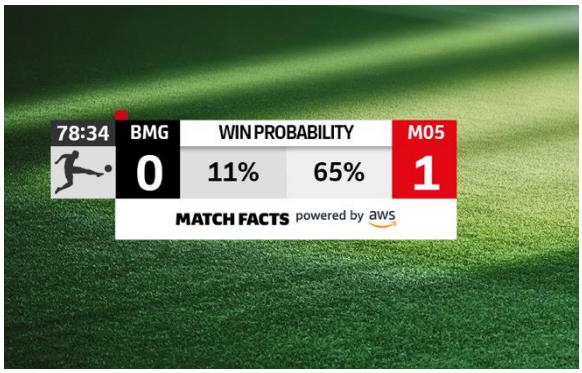
7. Speed Alert

Εμφανίζει πόσο γρήγορος είναι ένας ποδοσφαιριστής ανά πάσα στιγμή κατά τη διάρκεια ενός αγώνα και τους κατατάσσει συγκρίνοντας τη μέγιστη ταχύτητα μεταξύ ποδοσφαιριστών, ομάδων, και όλων των εποχών στη Bundesliga. **[49]**

Εικόνα 22: Παράδειγμα του Speed Alert **[50]**

8. Win Probability [49], [58]

Δείχνει πόσες πιθανότητες έχει η κάθε ομάδα να κερδίσει ανάλογα με την εξέλιξη του παιχνιδιού. Οι σημαντικές αλλαγές στην πιθανότητα νίκης γίνονται ορατές από ένα αντίστοιχο γραφικό.



Αναπτύχθηκε δημιουργώντας μοντέλα μηχανικής μάθησης που ανέλυσαν πάνω από 1000 προηγούμενα παιχνίδια. Το μοντέλο λαμβάνει ζωντανά τις εκτιμήσεις πριν από τον αγώνα και τις προσαρμόζει σύμφωνα με τα γεγονότα του αγώνα και με βάση τα χαρακτηριστικά που επηρεάζουν το αποτέλεσμα, συμπεριλαμβανομένων των εξής:

• Γκολ
• Πέναλτι
• Κόκκινες κάρτες
• Αλλαγές
• Χρόνος που πέρασε
• Ευκαιρίες κάθε ομάδας
• Στατικές φάσεις

Το μοντέλο εκπαιδευέται με νευρωνικό δίκτυο και χρησιμοποιεί μια κατανομή Poisson για να προβλέψει έναν ρυθμό r, όπου r ο ρυθμός των γκολ ανά λεπτό για κάθε ομάδα.

Η είσοδος στο μοντέλο είναι μια πλειάδα τριών χαρακτηριστικών.
α) τρέχουσα διαφορά τερμάτων
β) υπολοιπόμενος χρόνος παιχνιδιού
γ) ένα σύνολο χαρακτηριστικών

Το τρίτο χαρακτηριστικό είναι ένα σύνολο χαρακτηριστικών που περιγράφουν την τρέχουσα ροή του παιχνιδιού σε πραγματικό χρόνο και για τις δύο ομάδες, βασισμένα σε μετρικές απόδοσης. Αυτές περιλαμβάνουν διάφορες τιμές όπως τα xGoals με ιδιαίτερη προσοχή στα σουτ που πραγματοποιήθηκαν τα τελευταία 15 λεπτά. Επίσης, λαμβάνονται υπόψη κόκκινες κάρτες, πέναλτι, κόρνερ και ο αριθμός των επικίνδυνων στατικών φάσεων. Επικίνδυνη στατική φάση θεωρείται αυτή που εκτελείται από απόσταση μικρότερη των 25 μέτρων από την αντίπαλη εστία. Τέλος, το μοντέλο επηρεάζεται και από την αντικατάσταση ποδοσφαιριστών που έχουν κάποιο από τα Skill που αναφέρθηκαν προηγουμένως.

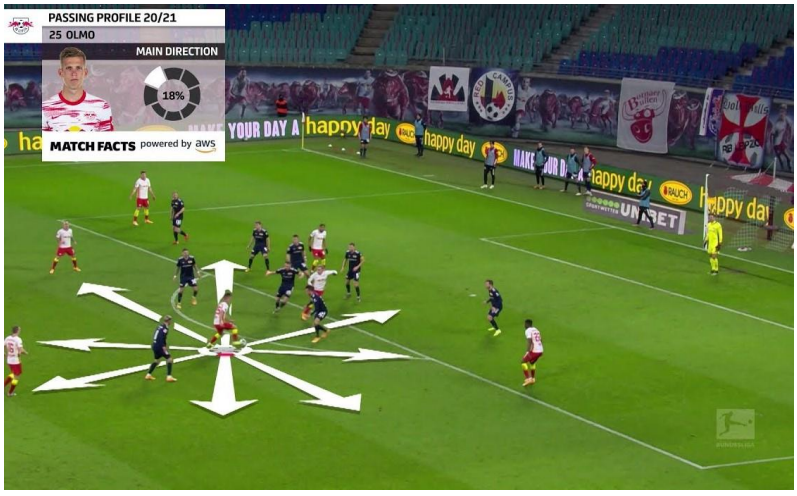
9. Shot Efficiency [49], [50]
Είναι η διαφορά των γκολ ενός ποδοσφαιριστή από τα xGoals του.
Δηλαδή, Shot Efficiency = Goals – xGoals

Αν το Shot Efficiency είναι μεγαλύτερο από 0, τότε ο ποδοσφαιριστής είναι αποτελεσματικός και εκμεταλλέυται και με το παραπάνω τις ευκαιρίες που του παρουσιάζονται.
Τη σεζόν 2020-2021 ο Robert Lewandowski σημείωσε ρεκόρ 41 τερμάτων, τα περισσότερα για έναν ποδοσφαιριστή μέσα σε μια σεζόν στο γερμανικό πρωτάθλημα. Τα xGoals του εκείνη τη σεζόν ήταν 28.7 . Άρα, είχε Shot Efficiency = +12.3
Με άλλα λόγια, έβαλε 12 γκολ περισσότερα από αυτά που έπρεπε να είχε βάλει σύμφωνα με το μοντέλο των xGoals.

10. Passing Profile

Το Passing Profile βοηθά τους φιλάθλους να κατανοήσουν πώς σκέφτονται οι ποδοσφαιριστές και πώς αποφασίζουν να δώσουν πάσα. Παρέχει βαθύτερες πληροφορίες για την ποιότητα και τη δύναμη της πάσας όπως και πληροφορίες για τις προτιμήσεις των ποδοσφαιριστών να πασάρουν, δηλαδή αν προτιμούν πάσες προς τα πίσω, κάθετες πάσες ή ψηλές διαγώνιες.

Έχουν χρησιμοποιηθεί σχεδόν 2 εκατομμύρια ιστορικά δεδομένα από πάσες για να φτιαχτεί ένα μοντέλο μηχανικής μάθησης που υπολογίζει τον βαθμό δυσκολίας της πάσας. Το μοντέλο είναι βασισμένο σε 26 παράγοντες, όπως η απόσταση της πάσας και η πίεση που δέχεται αυτός που κάνει την πάσα. **[49], [60]**



11. Set Piece Threat

Οι ομάδες πετυχαίνουν ένα σημαντικό μέρος των τερμάτων τους από στατικές φάσεις, δηλαδή φάουλ και κόρνερ. Το Set Piece Threat υπολογίζει την απειλή για αυτές τις καταστάσεις και μπορεί να αναδείξει πόσο απειλητική είναι μια ομάδα με τις στατικές φάσεις ή πόσο καλή είναι στην αντιμετώπισή τους.
Για την ποσοτικοποίησή αυτής της απειλής, λαμβάνεται υπόψη το αν προέκυψε τελική προσπάθεια ή όχι. Αν προέκυψε, λαμβάνεται υπόψη το xGoals που είχε αυτή η προσπάθεια. Τα παραπάνω δεν αθροίζονται και λαμβάνεται υπόψη η μέση τιμή ανά στατική φάση. Αυτό σημαίνει ότι μια ομάδα που κερδίζει πολλές στατικές φάσεις δεν θα έχει παραπάνω Set Piece Threat. Το αν προέκυψε τελική προσπάθεια συνεισφέρει 70% στο ενώ η τιμή του xGoals 30%. Τέλος, γίνεται σύγκριση όλων των ομάδων του πρωταθλήματος, για το πόσο πάνω ή κάτω είναι από τον μέσο όρο του πρωταθλήματος. **[49], [62]**

Για επιθετικές στατικές φάσεις: Ø League = score(team)/avg_score(league) – 1
Για αμυντικές στατικές φάσεις: Ø League = -score(team)/avg_score(league) + 1

12. Pressure Handling
Αναδεικνύει ποιοι ποδοσφαιριστές αποδίδουν εξαιρετικά υπό πίεση. Η βασική μετρική είναι το escape rate (ποσοστό διαφυγής) το οποίο δείχνει πόσο συχνά ξεφεύγει ένας ποδοσφαιριστής από καταστάσεις πίεσης και διατηρεί την κατοχή της μπάλας για την ομάδα του, δηλαδή μέσω πάσας, τρίπλας ή κερδίζοντας κάποια στατική φάση. Για κάθε ποδοσφαιριστή, υπολογίζουμε το ποσοστό διαφυγής διαιρώντας τον αριθμό διαφυγών που πέτυχε με τον αριθμό των καταστάσεων που βρέθηκε υπό πίεση. **[50], [63]**

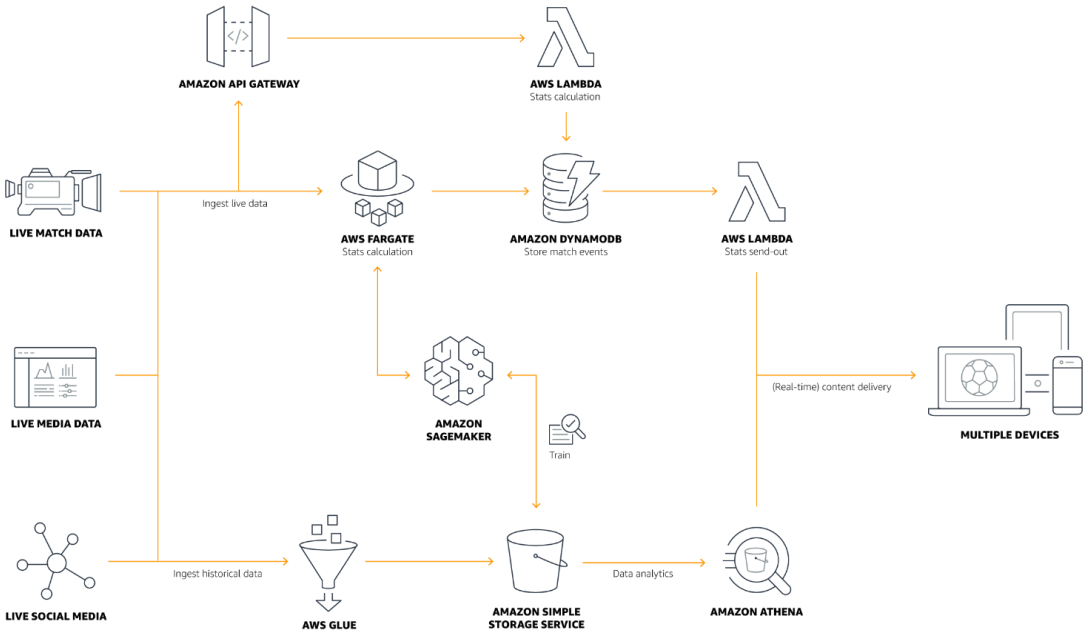
13. Keeper Efficiency [64]
Έχει παρόμοια μεθοδολογία με το Shot Efficiency. Αντί για xGoals έχουμε xSaves και αντί για σουτ έχουμε αποκρούσεις. Ουσιαστικά, συγκρίνει τις αποκρούσεις που έχει κάνει ένας τερματοφύλακας σε σχέση με αυτές που θα έπρεπε να είχε κάνει σύμφωνα με το μοντέλο των xSaves. Δηλαδή, Keeper Efficiency = Saves – xSaves.

14. Ball Recovery Time

Δείχνει πόσο χρόνο χρειάζεται μια ομάδα να ανακτήσει την κατοχή της μπάλας. Μέσω αλγορίθμων παρακολουθείται αυτόματα ποια ομάδα έχει τον έλεγχο της μπάλας και όταν η ομάδα χάνει την κατοχή, υπολογίζεται ο χρόνος που χρειάζεται για να ανακτήσει την μπάλα. **[49], [65]**

Εφαρμόζοντας μηχανική μάθηση στα δεδομένα

Χρησιμοποιώντας το ευρύ φάσμα δυνατοτήτων μηχανικής μάθησης της AWS που βασίζονται στο cloud, η Bundesliga παρέχει πληροφορίες για κάθε αγώνα. **[49]**



Premier League

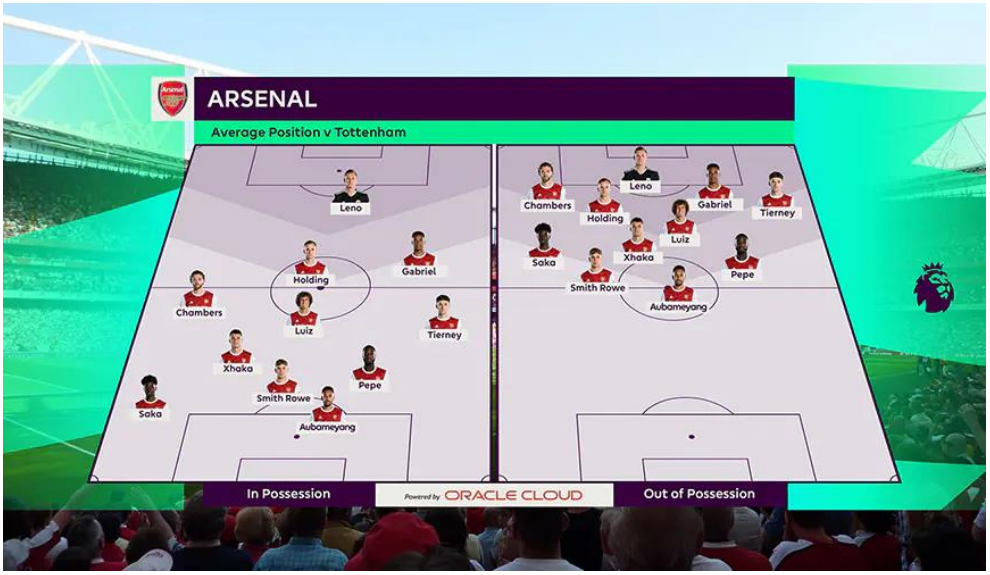
Η Premier League (Πρωτάθλημα Αγγλίας) συνεργάζεται με την Opta για τη συλλογή των δεδομένων και με την Oracle cloud παρέχουν πληροφορίες για τον αγώνα.
Μοντέλα μηχανικής μάθησης παράγουν άμεσα αποτελέσματα βασισμένα σε ζωντανές μεταδόσεις δεδομένων, δεδομένα παρακολούθησης σε πραγματικό χρόνο και συλλογής γεγονότων από τον κάθε ποδοσφαιριστή του πρωταθλήματος και από χιλιάδες προηγούμενα παιχνίδια. Οι φίλαθλοι έχουν άμεση πρόσβαση σε ένα εύρος πληροφοριών από τον κάθε αγώνα είτε βλέπουν από το σπίτι είτε ελέγχουν το αποτέλεσμα από τις φορητές τους συσκευές. **[66], [67]**

1. Attacking Threat

Πόσο πιθανόν να σκοράρει η ομάδα που έχει την κατοχή της μπάλας στα επόμενα 10 δευτερόλεπτα με βάση δεδομένα από χιλιάδες παιχνίδια.

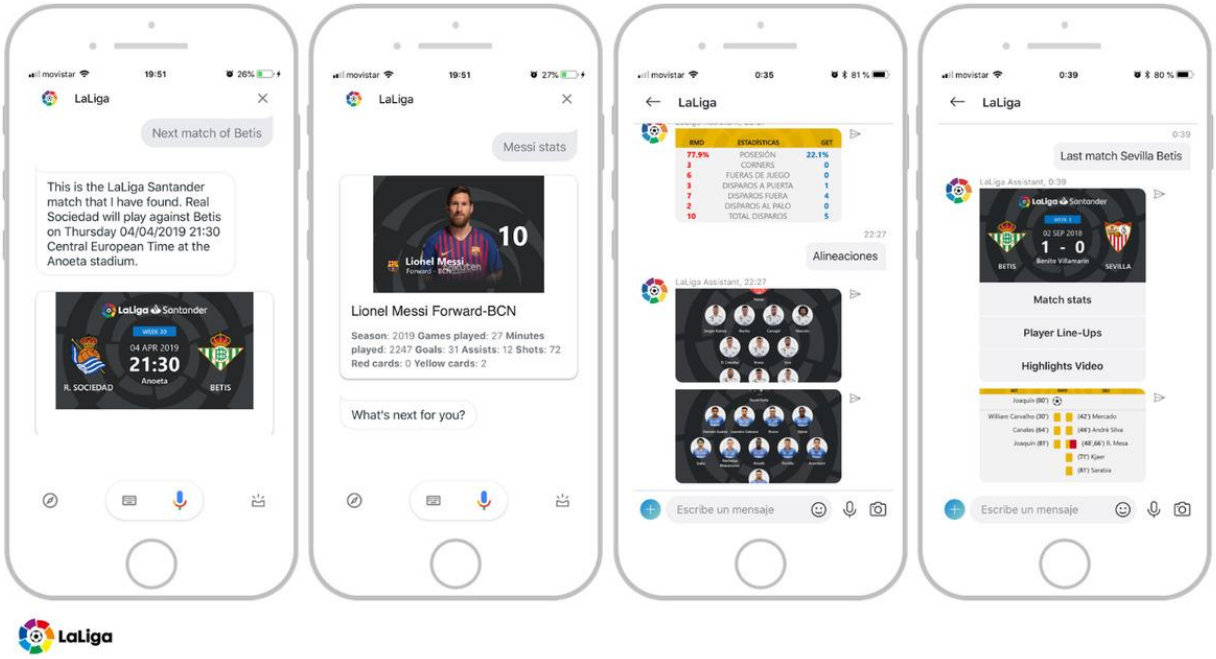
2. Win Probability
Χρησιμοποιώντας δεδομένα αγώνων από τις 4 τελευταίες σεζόν δείχνει την πιθανή έκβαση του αγώνα προσομοιώνοντας το υπόλοιπό του 100.000 φορές.

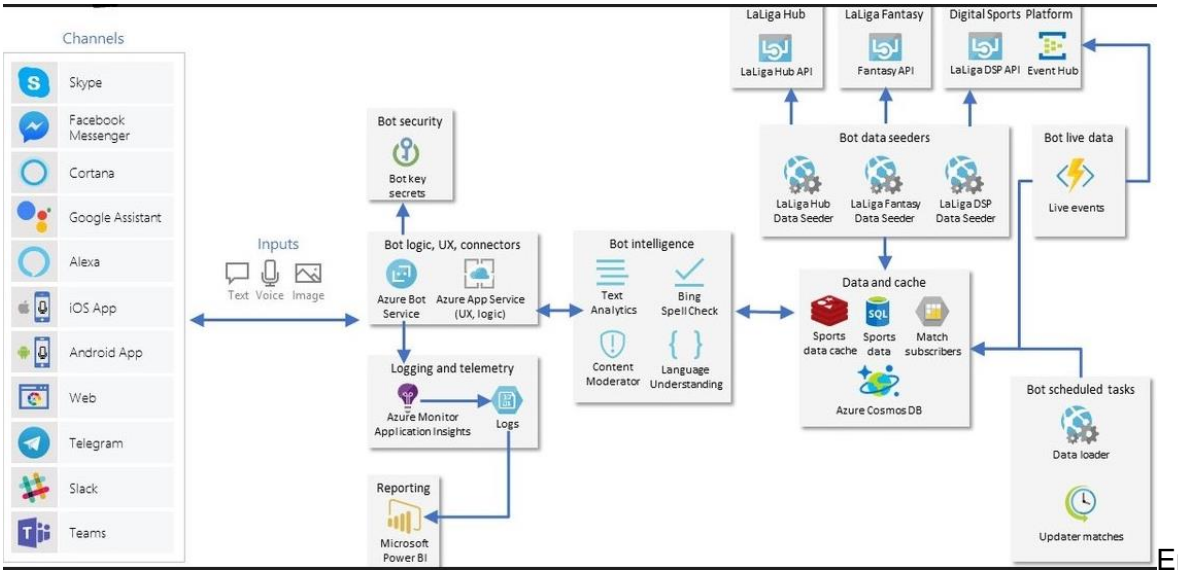
3. Average Position
Τι διάταξη χρησιμοποιεί μια ομάδα όταν έχει την μπάλα και όταν δεν την έχει.


ARSENAL
Average Position v Tottenham
Leno
Holding
Gabriel
Chambers
Luiz
Tierney
Xhaka
Pepe
Smith Rowe
Saka
Aubameyang
In Possession
Powered by ORACLE CLOUD
Out of Possession

La Liga

Η La Liga (Πρωτάθλημα Ισπανίας) ανέπτυξε μια πλατφόρμα ψηφιακής καινοτομίας βασισμένη στο Microsoft Azure. Αυτή η πλατφόρμα περιλαμβάνει έναν ψηφιακό βοηθό που βοηθά τους φιλάθλους να μάθουν περισσότερα για τις ομάδες ή τους παίκτες που θέλουν μέσω συνομιλητικής τεχνητής νοημοσύνης (conversational AI). Ο φίλαθλος έχει τη δυνατότητα να ρωτήσει ακόμα και μέσω της φωνής του. Απαντάει σε ερωτήματα όπως τον επόμενο αγώνα μιας ομάδας, τα στατιστικά ενός ποδοσφαιριστή αλλά και ποιος ήταν ο προηγούμενος αγώνας μεταξύ συγκεκριμένων ομάδων. Ο εικονικός βοηθός έχει κυκλοφορήσει για Google Assistant και Skype, αλλά αναμένεται να κυκλοφορήσει σε Messenger, Alexa, Cortana και άλλα κανάλια. Με αυτήν την ισχυρή ψηφιακή εμπειρία, η La Liga είναι σε θέση να αναπτύξει ακόμα περισσότερο την ταυτότητα της επωνυμίας της. **[68], [69]**





Εικόνα 36: Η αρχιτεκτονική του εικονικού βοηθού **[68]**

4. Εφαρμογές της Μηχανικής Μάθησης στο Ποδόσφαιρο

Στο προηγούμενο κεφάλαιο αναφέρθηκε ο ρόλος του cloud computing στην επεξεργασία των δεδομένων. Έτσι, οι ομάδες μπορούν να κάνουν καλύτερη την εμπειρία του φιλάθλου είτε είναι στο γήπεδο είτε παρακολουθεί τον αγώνα από την τηλεόραση. Επίσης, έγινε μια σύντομη αναφορά ότι οι ομάδες μπορούν να αναλύσουν θέματα όπως η αγωνιστική και η οικονομική τους απόδοση και πώς μπορούν να τα βελτιώσουν. Σε αυτό το κεφάλαιο θα αναλυθεί ο ρόλος και οι εφαρμογές της μηχανικής μάθησης στο ποδόσφαιρο.

4.1 Ανάλυση Απόδοσης (Performance Analysis)

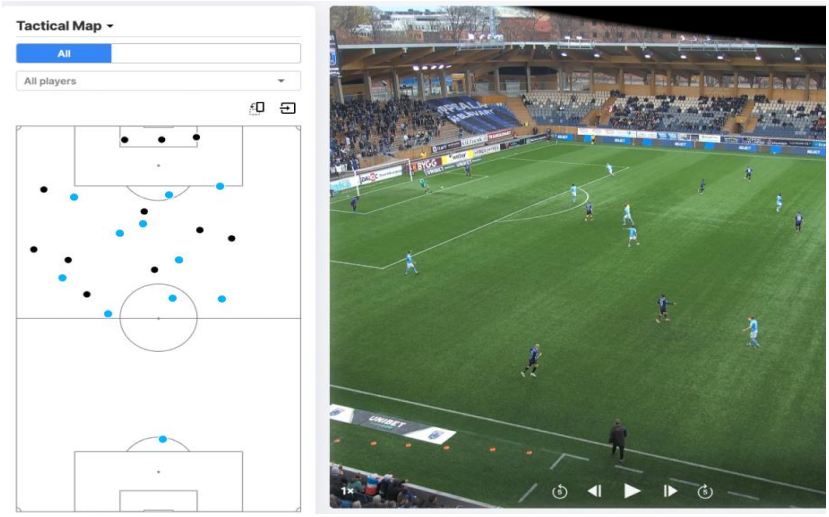
Αυτή η διαδικασία βασίζεται σε συστηματική παρατήρηση η οποία παρέχει έγκυρες, αξιόπιστες και λεπτομερείς πληροφορίες σχετικά με την απόδοση. Η ανάλυση απόδοσης μπορεί να βοηθήσει στην βελτίωση της διαδικασίας της προπόνησης παρέχοντας οπτική ανατροφοδότηση (ανάλυση βίντεο) και αντικειμενική στατιστική ανάλυση. Με αυτόν τον τρόπο, μια απόφαση μπορεί να ληφθεί πιο αντικειμενικά και όχι βάση υποθέσεων. Αυτό βοηθά τους αθλητές και τους προπονητές να κατανοήσουν ακριβώς τι έχουν κάνει σωστό ή λάθος και να παρουσίασουν συνεπείς επιδόσεις. Μπορεί να γίνει ανάλυση απόδοσης και των αντίπαλων ομάδων ώστε να εντοπιστούν αδύναμα σημεία με σκοπό να προετοιμαστεί μια στρατηγική για τον αγώνα. Πιο συγκεκριμένα, να αποφασιστεί ποιοι ποδοσφαιριστές θα χρησιμοποιηθούν, σε ποια θέση μπορεί να αποδώσει καλύτερα ο καθένας και σε ποια διάταξη. **[70]**

Ανάλυση βίντεο

Η ανάλυση βίντεο αποτελεί ένα πολύ χρήσιμο εργαλείο στην ανάλυση απόδοσης. Στο κορύφαιο επίπεδο ομάδων, εφαρμόζεται σε όλους τους αγώνες και στις περισσότερες προπονήσεις.

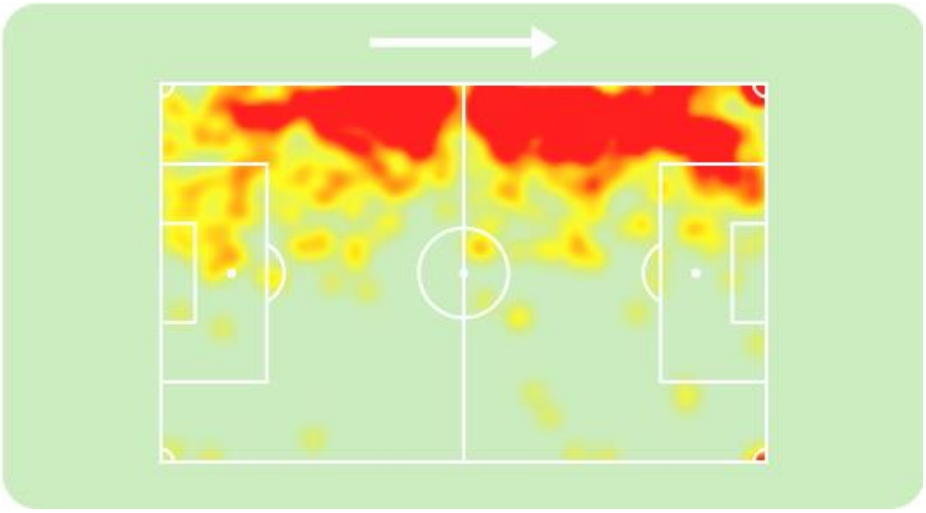
<u>Με χρήση εργαλείων μηχανικής μάθησης μπορούν να επιτευχθούν:</u> **[71]**

- Αναγνώριση ποδοσφαιριστών, ομάδων, μπάλας, διαιτητή και αγωνιστικού χώρου
- Αναγνώριση διάταξης που χρησιμοποιεί η κάθε ομάδα σε διάφορες συνθήκες του παιχνιδιού **[72]**
- Ιχνηλάτηση (Tracking) συγκεκριμένου ή συγκεκριμένων αντικειμένων
- Δημιουργία 2D όψης που διευκολύνει την όλη διαδικασία της ανάλυσης. **[73]**



Στις 2D όψεις υπάρχει η δυνατότητα να γίνει πολλές φορές προσομοίωση του αγώνα και με διαφορετικές συνθήκες ώστε να βγουν χρήσιμα συμπεράσματα. **[74]**

Από την ανάλυση βίντεο προκύπτει η ανάγκη οπτικοποίησης των κινήσεων ενός ποδοσφαιριστή. Αυτό γίνεται συχνά με χρήση θερμικού χάρτη (heat map). Ο θερμικός χάρτης απεικονίζει τις περιοχές μέσα στις οποίες κινείται ένας ποδοφαιριστής. Αυτό επιτυγχάνεται με την χρωματική παλέτα, δηλαδή όσο πιο σκούρα η απόχρωση τόσο πιο πολύ ώρα κινήθηκε σε αυτό το σημείο. Ένας θερμικός χάρτης μπορεί να αποκαλύψει ζητήματα τακτικής που δεν είναι ορατά με γυμνό μάτι. Παρόλα αυτά δεν παρουσιάζει αν οι ενέργειες του ποδοσφαιριστή είναι καλές ή κακές. **[75]**

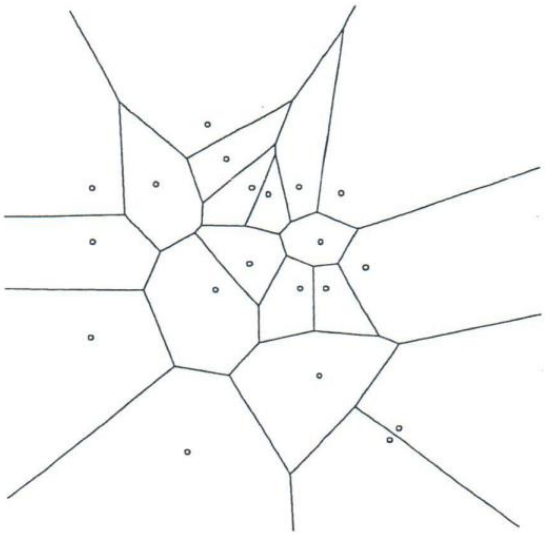


Από το παραπάνω heat map μπορεί να βγει το συμπέρασμα ότι ανήκει σε ποδοσφαιριστή που αγωνίζεται στη θέση του αριστερού μπακ και όπως φαίνεται έχει επιθετικές αρετές και δεν αρκείται μόνο στα αμυντικά του καθήκοντα. Πιο συγκεκριμένα, το παραπάνω heat map ανήκει στον ποδοσφαιριστή Κώστα Τσιμίκα και είναι μια αποτύπωση των κινήσεων του για τη σεζόν 2021-2022. Πράγματι, αγωνίζεται στην θέση του αριστερού μπακ και έχει αυτόν τον τρόπο παιχνιδιού.

Άλλο ένα εργαλείο που χρησιμοποιείται είναι το διάγραμμα Voronoi. Μπορεί να αναδείξει στοιχεία όπως οι τοποθετήσεις των αμυντικών αλλά και η γεωμετρία που εφαρμόζουν οι ομάδες για να έχουν τις ιδανικές αποστάσεις μέσα στον αγώνα. Αυτό το υπολογιστικό εργαλείο μπορεί να βοηθήσει τους προπονητές να βρουν κένα διαστήματα για να επιτεθεί η ομάδα τους, αλλά και να καλύψει χώρους στην άμυνα. Το διάγραμμα Voronoi ταιριάζει απόλυτα με την φιλοσοφία των προπονητών για το σύγχρονο ποδόσφαιρο.

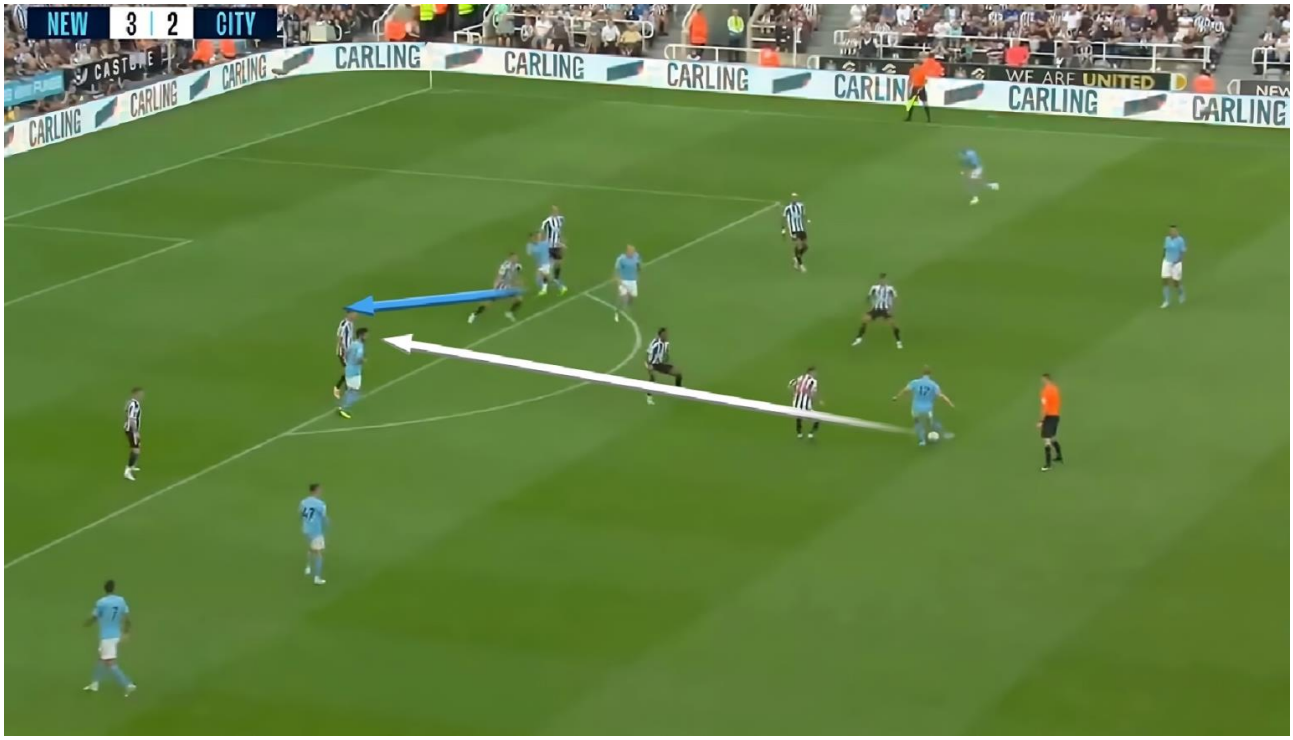
Τι είναι το διάγραμμα Voronoi;

Έστω P = {p1,p2,…,pn} ένα σύνολο σημείων στο επίπεδο τα οποία ονομάζονται κόμβοι. Σε κάθε έναν από τους κόμβους αναθέτουμε όλα τα σημεία του επίπεδου που είναι πιο κοντά σε αυτόν παρά σε οποιονδήποτε άλλον κόμβο, με βάση την Ευκλείδια απόσταση. Όλα αυτά τα σημεία διαμορφώνουν την περιοχή Voronoi του κόμβου. **[77]**



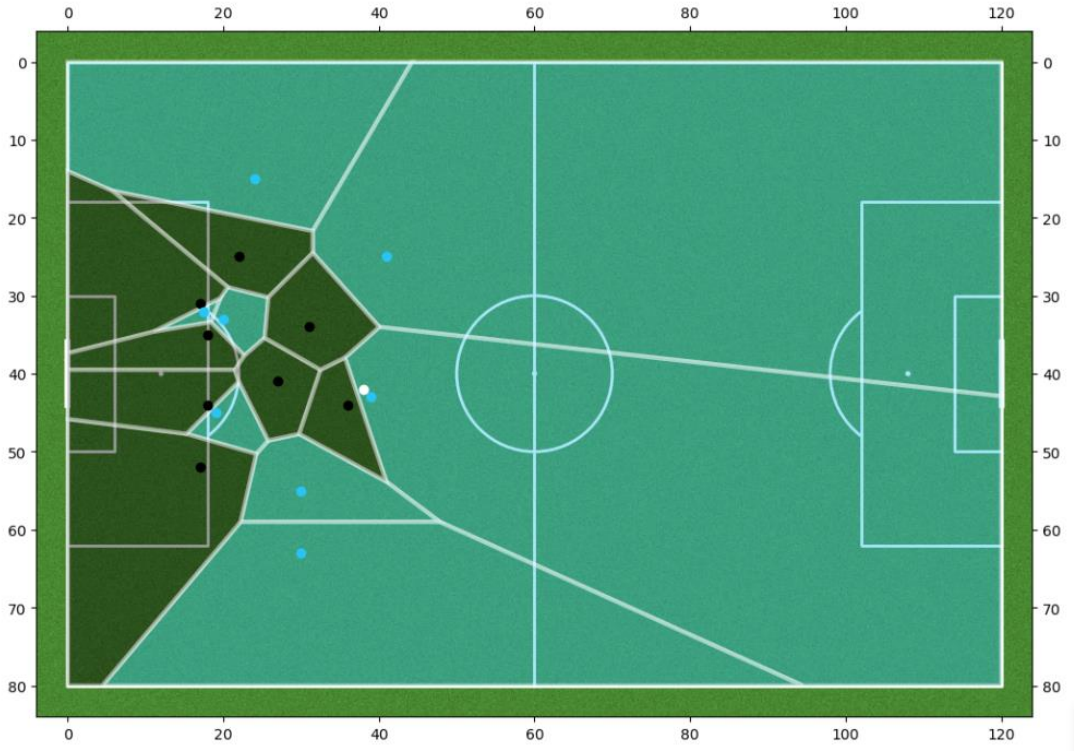
Η τακτική της Manchester City είναι σε μεγάλο βαθμό εμπνευσμένη από το διάγραμμα Voronoi. Είναι μία από τις καλύτερες ομάδες στην κυκλοφορία της μπάλας αλλά και στην πίεση των αντιπάλων της. Προσπαθεί να βρει χώρους στην επίθεση και να δημιουργήσει ευκαιρίες που δεν αντιμετωπίζονται εύκολα από την ομάδα που αμύνεται. Επίσης, όταν αμύνεται πιέζει ασφυκτικά και δεν αφήνει χώρους στην αντίπαλη ομάδα. Σε αυτό παίζει ρόλο και η ποιότητα των ποδοσφαιριστών της που έχουν αποδείξει ότι και σε κλειστούς χώρους έχουν την ικανότητα να βρουν τον τρόπο να είναι αποτελεσματικοί, αλλά και καταφέρνοντας να εξουδετερώσουν ακόμα και τις καλύτερες επιθέσεις. **[78]**

Για παράδειγμα, στην παρακάτω εικόνα φαίνεται μια επίθεση της Manchester City απέναντι σε μια αρκετά καλά οργανώμενη άμυνα που δεν θέλει να επιτρέψει μια κάθετη πάσα παρά μόνο παράλληλες. Κι όμως, ο ποδοσφαιριστής που έχει τη μπάλα περνάει μια κάθετη πάσα στον συμπαίκτη του που εκείνη την ώρα πατάει στην αντίπαλη περιοχή και εκείνος πετυχαίνει γκολ.

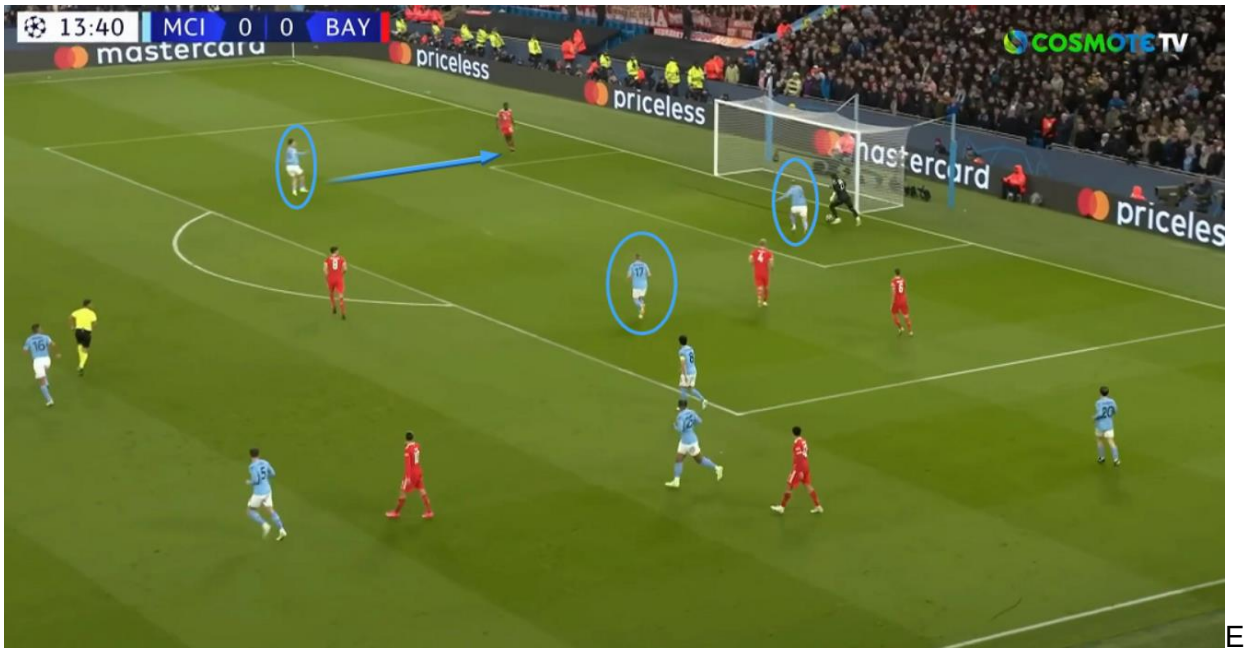


Εικόνα 40: Η απίστευτη ασίστ του ποδοσφαιριστή της Manchester City **[79], [80]**

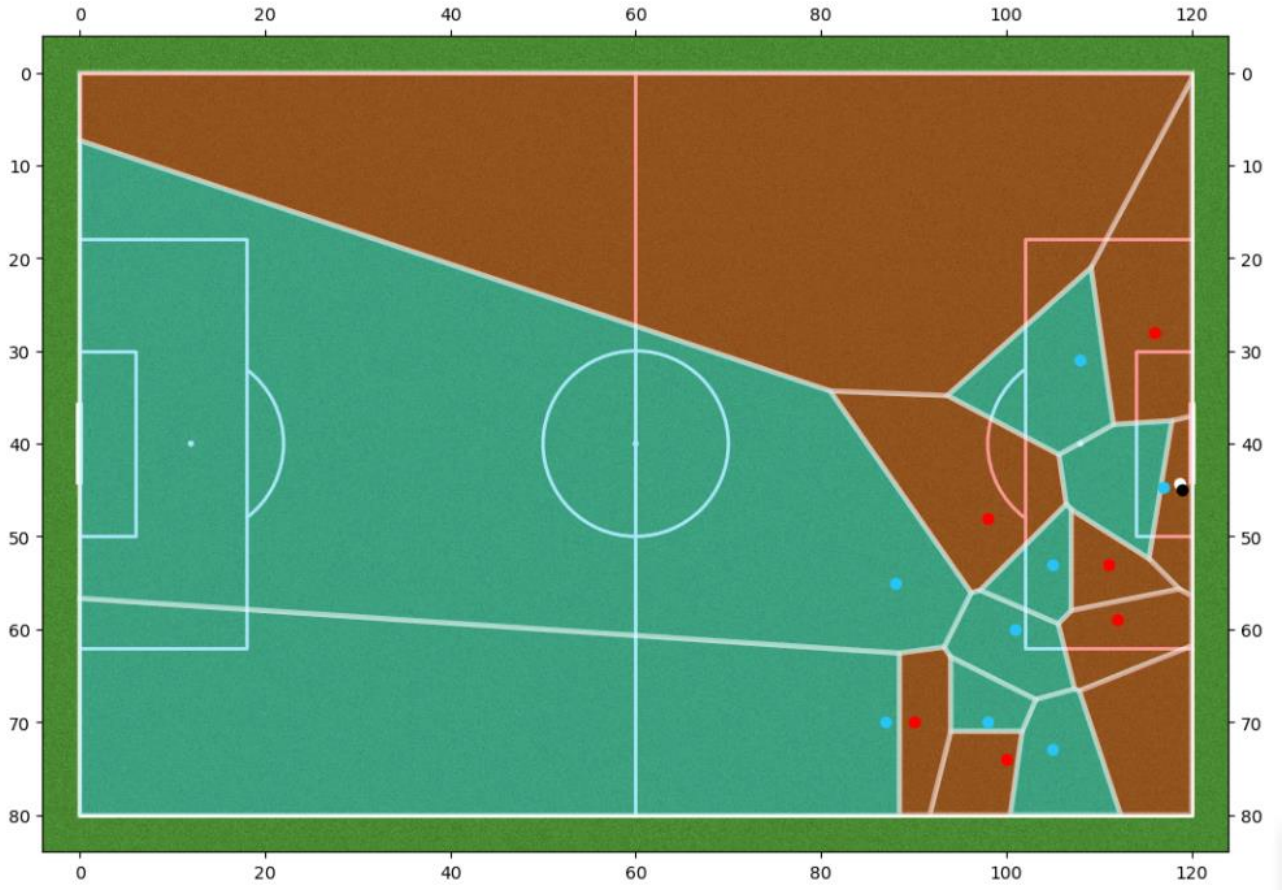
Η δυσκολία της πάσας μπορεί να αποτυπωθεί καλύτερα στο διάγραμμα Voronoi.



Όπως αναφέρθηκε και παραπάνω, η Manchester City είναι μία από τις καλύτερες ομάδες και στην πίεση των αντιπάλων της και δε διστάζει να πιέζει ακόμα και δυνατούς αντιπάλους, όπως η Μπάγερν Μονάχου. Στην παρακάτω φωτογραφία φαίνεται η ασφυκτική της πίεση να φτάνει μέχρι και τον αντίπαλο τερματοφύλακα όπου τελικά τον αναγκάζει σε λάθος πάσα και τη μπάλα να καταλήγει πλάγιο.



Εικόνα 42: Η ασφυκτική πίεση της Manchester City **[80], [81]**



Εικόνα 43: Το αντίστοιχο διάγραμμα Voronoi στο Jupyter Notebook **[Παράρτημα 1]**

4.2 Πρόληψη και αποκατάσταση τραυματισμών

Κάτι πολύ δυσάρεστο στον αθλητισμό είναι οι τραυματισμοί των αθλητών. Ευτυχώς, οι ομάδες έχουν πλέον τους τρόπους να μειώσουν την συχνότητα και τη σοβαρότητα των τραυματισμών. Ακόμα και αν δεν αποφευχθεί ο τραυματισμός, η αποκατάσταση είναι πιο ομαλή και ο αθλητής επανέρχεται στην αγωνιστική δράση σε πολύ καλή κατάσταση. Αυτό φυσικά οφείλεται ότι οι ομάδες έχουν δεδομένα να βελτιώσουν την προπόνηση και την αποθεραπεία. Οι αθλητές φοράνε γιλέκα με τεχνολογία GPS και βιντεοσκοπούνται σχεδόν σε όλες τις προπονήσεις.

Milan Lab

Είναι ένα ερευνητικό κέντρο υψηλής τεχνολογίας που ανήκει στην AC Milan. Παρέχει πληροφορίες για ιατρικά δεδομένα με σκοπό την πρόληψη και την αποκατάσταση τραυματισμών. Δημιουργήθηκε το 2002 και πέρα από την πρόληψη τραυματισμών, καθορίζει σε σημαντικό βαθμό το πρόγραμμα προπόνησης με σκοπό να διατηρηθούν σε καλή κατάσταση ακόμα και ποδοσφαιριστές μεγαλύτερης ηλικίας. Το Milan Lab πήρε γρήγορα αναγνωρισιμότητα καθώς τα επόμενα χρόνια η AC Milan ήταν μια ομάδα κορυφαίου επιπέδου που τρόμαζε κάθε αντίπαλο. Αυτό επισφραγίστηκε και με τίτλους με αποκορύφωμα να αναδειχθεί 2 φορές Πρωταθλήτρια Ευρώπης, το 2003 και το 2007. Η 2η κατάκτηση ήρθε με μέσο όρο ηλικίας των ποδοσφαιριστών της κοντά στα 30 έτη και τον αρχηγό της να πλησιάζει τα 40.
[82], [83]

4.3 Ο ρόλος του Scouting και των Ακαδημιών στη στελέχωση της πρώτης ομάδας

Αρκετές ομάδες δεν έχουν την πρόθεση και τη δυνατότητα να ξοδεύουν υπέρογκα ποσά για την απόκτηση ποδοσφαιριστών. Αντιθέτως, ψάχνουν να αγοράσουν φθηνά και να πουλήσουν ακριβά. Σε αυτό έχει συνεισφέρει δραματικά η ανάλυση δεδομένων. Οι ομάδες μπορούν να βρουν παίκτες και πρωταθλήματα που συχνά υποτιμώνται και να ανακαλύψουν μια πολύ καλή περίπτωση παίκτη για την ομάδα τους που θα τους αποφέρει και αγωνιστικές λύσεις αλλά και οικονομικό κέρδος στο μέλλον. Αυτό συνήθως εφαρμόζεται από τα τμήματα υποδομών ώστε να αναπτυχθούν ποδοσφαιριστές που τα επόμενα χρόνια θα βοηθήσουν το σύλλογο. Τέλος, δε φοβούνται να πουλήσουν ποδοσφαιριστές αν κριθεί ότι κοστολογούνται παραπάνω από αυτά που προσφέρουν. Έτσι, τους δίνεται η δυνατότητα επανεπένδυσης των χρημάτων τους είτε για scouting είτε για τις ακαδημίες. **[83]**

4.4 Εκτίμηση και διαχείριση οικονομικών δεδομένων

Ήδη πολλές ομάδες προσπαθούν να βελτιστοποιήσουν τις τιμές των εισιτηρίων και των εμπορευμάτων ώστε να προσελκύσουν φιλάθλους αλλά ταυτόχρονα με το μεγαλύτερο δυνατό κέρδος. Επίσης, πρέπει να διασφαλίσουν ότι όλες αυτές οι συναλλαγές θα γίνονται με ασφαλή και έγκυρο τρόπο. Όμως, οι ομάδες είναι υποχρεωμένες να κάνουν και προσεκτική διαχείριση των εξόδων τους για να συμμορφώνονται με τους κανονισμούς των ομοσπονδιών, με σημαντικότερο το Financial Fair Play. Σε όλα αυτά μπορεί να δώσει λύσεις η μηχανική μάθηση. **[24]**

Financial Fair Play

Το Financial Fair Play (FFP) είναι ένα σύστημα κανόνων που θεσπίστηκε από την Ευρωπαϊκή Ποδοσφαιρική Ομοσπονδία (UEFA) το 2009 με σκοπό την εξασφάλιση της οικονομικής βιωσιμότητας των συλλόγων. Πιο συγκεκριμένα, απαιτεί από τις ομάδες να λειτουργούν με οικονομική πειθαρχία και να μην ξοδεύουν περισσότερα χρήματα από τα έσοδα τους. Αυτό σημαίνει ότι δε μπορούν να επιχορηγούνται από πλούσιους ιδιοκτήτες, εταιρείες ή κράτη προκειμένου να αγοράζουν ακριβούς ποδοσφαιριστές ή να αυξάνουν τον μισθολογικό προϋπολογισμό τους. Οι σύλλογοι πρέπει να παρουσιάζουν ισορροπημένα οικονομικά αποτελέσματα και να μην οφείλουν υπερβολικά ποσά σε άλλους οργανισμούς. Η επιβολή του γίνεται μέσω της αξιολόγησης της οικονομικής κατάστασης των συλλόγων από την UEFA που ελέγχει τα βιβλία και τηρεί στοιχεία για τα έσοδα και τις δαπάνες τους. Οι σύλλογοι που δε συμμορφώνονται με το FFP έρχονται αντιμέτωπες με κυρώσεις όπως αφαίρεση βαθμών, αποκλεισμό από διοργανώσεις και υψηλά πρόστιμα. **[84], [85]**

4.5 Διαιτησία

Στην προσπάθεια που γίνεται να μειωθούν, ακόμα και να εξαλειφθούν τα διαιτητικά λάθη εξαιτίας του ανθρώπινου παράγοντα, η μηχανική μάθηση βοηθήσει ακόμα και σε οριακές φάσεις.

4.5.1. Τεχνολογία Γραμμής (Goal-line Technology)

Η τεχνολογία γραμμής είναι ένα τεχνικό μέσο που καθορίζει άμεσα αν η μπάλα έχει περάσει ολόκληρη τη γραμμή του τέρματος. Εισήχθη για πρώτη φορά το 2014. Η πληροφορία μεταδίδεται εντός ενός δευτερολέπτου εξασφαλίζοντας μια άμεση απόφαση του διαιτητή.

Οι διαιτητές λαμβάνουν την πληροφορία στα ρολόγια τους. Το σύστημα χρησιμοποιεί 14 υψηλής ταχύτητας κάμερες που τοποθετούνται περιμετρικά του γηπέδου.
Τα δεδομένα από τις κάμερες χρησιμοποιούνται για να δημιουργηθεί μια 3D κινούμενη εικόνα που εμφανίζεται στην τηλεόραση και στην τεράστια οθόνη μέσα στο γήπεδο. **[86]**

4.5.2. Video Assistant Referee (VAR)

Είναι ένα εργαλείο υποστήριξης για τους διαιτητές. Χρησιμοποιήθηκε επιτυχώς στο Παγκόσμιο Κύπελλο 2018 και έκτοτε έχει εφαρμοστεί σε πάνω από 100 διοργανώσεις παγκοσμίως.

Η ομάδα VAR υποστήριζει τη διαδικασία λήψης αποφάσεων του διαιτητή σε 4 περιπτώσεις.
1. Γκολ και παραβάσεις πριν το γκολ
2. Πέναλτι και παραβάσεις πριν το πέναλτι
3. Απευθείας κόκκινη κάρτα (όχι 2η κίτρινη)
4. Κάρτα σε λάθος ποδοσφαιριστή

Η ομάδα του VAR βρίσκεται σε μια αίθουσα μακριά από το γήπεδο και λαμβάνει το βίντεο από τις κάμερες του γηπέδου μέσω οπτικής ίνας. Αν κρίνει ότι πρέπει να παρέμβει, επικοινωνεί με τον διαιτητή μέσω συστήματος ραδιοφώνου με οπτικές ίνες. Η περιοχή όπου ο διαιτητής συμβουλεύεται το βίντεο βρίσκεται ανάμεσα στους πάγκους των ομάδων και παρακολουθεί το βίντεο από μια οθόνη.

Το VAR όπως είναι εύκολα αντιληπτό εξαρτάται σε μεγάλο βαθμό από τον ανθρώπινο παράγοντα. Όμως, έχει πρόσβαση στο Goal-line technology και στη νέα τεχνολογία ημιαυτόματου offside. **[88], [89]**

4.5.3. Τεχνολογία Ημιαυτόματου Οφσαίντ (Semi-automated Offside Technology)

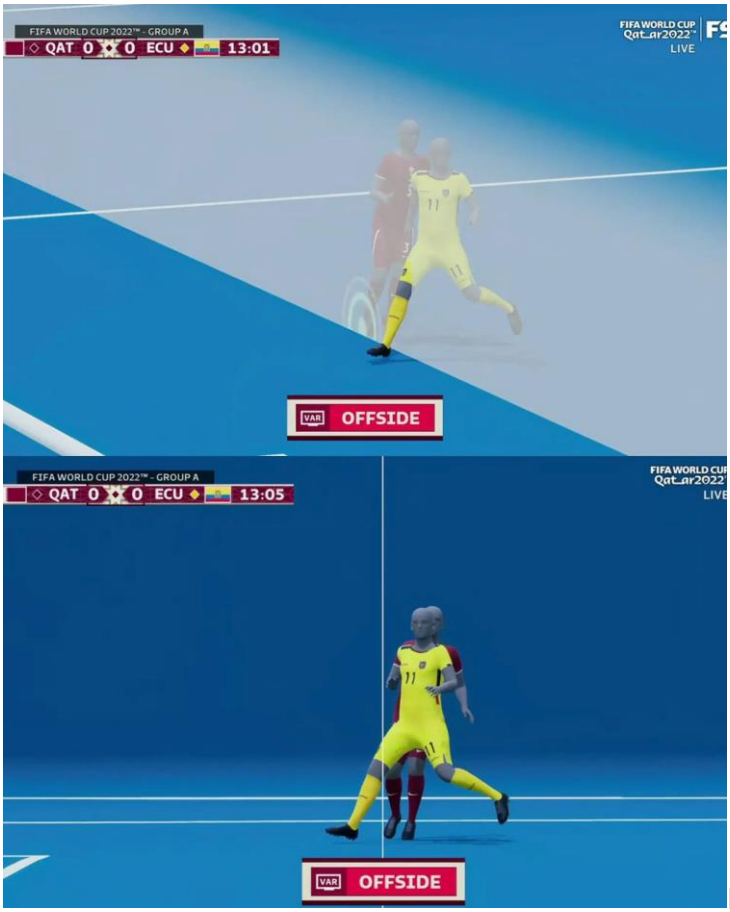
Είναι ένα εργαλείο υποστήριξης της λήψης αποφάσεων οφσαίντ. Έχει σκοπό να βελτιώσει την ταχύτητα και την ακρίβεια αυτών των αποφάσεων. Εφαρμόστηκε στο Παγκόσμιο Κύπελλο 2022.

Χρησιμοποιεί 12 κάμερες για την παρακολούθηση της μπάλας και έως και 29 σημείων δεδομένων για κάθε ποδοσφαιριστή, 50 φορές το δευτερόλεπτο, υπολογίζοντας την ακριβή θέση τους στον αγωνιστικό χώρο. Τα 29 σημεία δεδομένων περιλαμβάνουν όλα τα άκρα του σώματος που είναι σημαντικά για τη λήψη αποφάσεων οφσάιντ.

Συνδυάζοντας τα δεδομένα παρακολούθησης των άκρων και της μπάλας και εφαρμόζοντας τεχνητή νοημοσύνη, παρέχεται αυτόματη ειδοποίηση για οφσάιντ στην ομάδα VAR όταν η μπάλα ληφθεί από έναν επιθετικό σε θέση οφσαίντ τη στιγμή που η μπάλα έφυγε από τον συμπαίκτη του. Πριν ενημερωθεί ο διαιτητής στον αγωνιστικό χώρο, η ομάδα VAR επιβεβαιώνει την προτεινόμενη απόφαση ελέγχοντας χειροκίνητα το αυτόματα επιλεγμένο σημείο εκτέλεσης και την αυτόματα δημιουργημένη γραμμή οφσαίντ, η οποία βασίζεται στις υπολογισμένες θέσεις των άκρων των ποδοσφαιριστών. Αυτή η διαδικασία διαρκεί μερικά δευτερόλεπτα.

Μετά την επιβεβαίωση της απόφασης από τον διαιτητή στον αγωνιστικό χώρο, τα ίδια ακριβώς σημεία θέσης δεδομένων που χρησιμοποιήθηκαν για τη λήψη της απόφασης μετατρέπονται σε μια 3D κινούμενη εικόνα που απεικονίζει ακριβώς τη θέση των άκρων των

παικτών στη στιγμή που η μπάλα έπαιξε. Αυτή η εικόνα εμφανίζεται στην τηλεόραση και στις οθόνες του γηπέδου. **[90], [91]**



Εικόνα 46: Παράδειγμα χρήσης ημιαυτόματου οφσάιντ **[91]**

4.6 Η ανατρεπτική προσέγγιση της Brentford

Η Brentford αποφασίζει σε μεγάλο βαθμό με βάση τα δεδομένα σε κάθε τομέα του συλλόγου, ειδικά στην επιλογή των ποδοσφαιριστών.

Όλο αυτό ξεκίνησε το 2012 όταν ιδιοκτήτης του συλλόγου έγινε ο Matthew Benham, απόφοιτος φυσικής από το πανεπιστήμιο της Οξφόρδης. Κέρδισε πολλά χρήματα χρησιμοποιώντας μαθηματικά μοντέλα στο στοίχημα. Ίδρυσε την εταιρεία Smartodds που παρέχει εξειδικευμένα δεδομένα για αθλητικά γεγονότα και την στοιχηματική εταιρεία Matchbook. Έγινε ιδιοκτήτης της Brentford για να δοκιμάσει αν οι γνώσεις του μπορούν να βοηθήσουν την αγαπημένη του ομάδα. Τότε, η Brentford ήταν στη League One, στη 3η σε δυναμικότητα κατηγορία της Αγγλίας και δυσκολευόταν οικονομικά.

Ο Benham ήταν από τους πρώτους που κατάλαβε την αξία των xGoals και τα χρησιμοποίησε όχι μόνο για να βρει υποτιμημένους επιθετικούς, αλλά και για να κάνει μια εκτίμηση για την συνολική απόδοση της ομάδας. Κατάφερε να βρει παίκτες και πρωταθλήματα που υποτιμώνται από την αγορά. Ο στόχος της ήταν να φτιάξει ένα συμπαγές σύνολο που να ταιριάζουν οι ικανότητές τους στον αγωνιστικό χώρο, ώστε να μεγιστοποίησει την απόδοσή τους. Για παράδειγμα, αν αγόραζε έναν ψηλό επιθετικό, έπρεπε να έχει κάποιον να τον τροφοδοτεί με καλές σέντρες και κατά συνέπεια να έχει και κάποιον να κάνει διαγώνιες πάσες σε αυτούς που θα κάνουν σέντρες. Μέσα από αυτή τη διαδικασία βελτιώθηκαν πολλοί ποδοσφαιριστές και πουλήθηκαν σε αρκετά μεγαλύτερη τιμή από αυτήν που αγοράστηκαν. Τέλος, η Brentford δεν φοβάται να πουλήσει ποδοσφαιριστές που κρίνει ότι αποδίδουν παραπάνω σε σχέση με το αναμενόμενο και εκτιμά ότι στη συνέχεια θα πέσει η απόδοσή τους. **[83], [92], [93]**

Πίνακας 4: Οι 7 ακριβότερες πωλήσεις της Brentford σύμφωνα με το Transfermarkt **[94]**

Ποδοσφαιριστής	Ποσό αγοράς (€)	Ποσό πώλησης (€)	Κέρδος (€)
Ollie Watkins	7.220.000	34.000.000	26.780.000
Said Benrahma	1.700.000	23.100.000	21.400.000
Neal Maupay	2.000.000	15.560.000	13.560.000
Chris Mepham	0 (Ακαδημίες)	13.600.000	13.600.000
Ezri Konsa	2.850.000	13.300.000	10.450.000
Andre Gray	620.000	12.400.000	11.780.000
Scott Hogan	950.000	10.500.000	9.550.000
Σύνολο	**15.340.000**	**122.460.000**	**107.120.000**

Πίνακας 5: Οι 5 ακριβότεροι ποδοσφαιριστές από την τωρινή ομάδα της Brentford σύμφωνα με το Transfermarkt. (1/6/2023) **[94]**

Ποδοσφαιριστής	Ποσό αγοράς (€)	Τρέχουσα αξία (€)
Ivan Toney	5.600.000	50.000.000
Bryan Mbeumo	6.500.000	28.000.000
David Raya	3.350.000	25.000.000
Rico Henry	1.800.000	22.000.000
Kevin Schade	1.000.000 (δανεικός)	20.000.000

Όπως αναφέρθηκε και προηγουμένως, ο σύλλογος απέκτησε μια κουλτούρα analytics σε κάθε τομέα του. Αυτό φαίνεται από την πρόσληψη στατιστικολόγου για την ανάλυση δεδομένων κατά τη διάρκεια του αγώνα, αναλυτή στατικών φάσεων αλλά μέχρι και αναλυτή ύπνου. **[92], [93]**

Πλεόν, η Brentford είναι οικονομικά βιώσιμη και αγωνίζεται στην Premier League από τη σεζόν 2021-2022 όπου τερμάτισε 13η, ενώ σύμφωνα με το μοντέλο των Expected Points (xPTS) άξιζε να βρίσκεται στην 7η θέση. Τα 15 από τα 48 γκολ που πέτυχε, δηλαδή περίπου το 31%, προήλθαν από στατικές φάσεις και ήταν η 4η ομάδα με τα περισσότερα γκολ από στατικές φάσεις. **[95]**

Τη σεζόν 2022-2023 τερμάτισε στην 9η θέση ενώ πάλι άξιζε να βρίσκεται στην 7η θέση. Πέτυχε συνολικά 58 τέρματα, εκ των οποίων τα 16 πρόηλθαν από στατικές φάσεις (περίπου το 27%). Ήταν η 2η ομάδα με τα περισσότερα γκολ από στατικές φάσεις. **[96]**

Τι είναι το μοντέλο των xPTS; [97], [98]
Το μοντέλο των xPTS προσπαθεί να αποτυπώσει μια “δίκαιη” βαθμολογία του πρωταθλήματος. Οι ομάδες κανονικά παίρνουν 3 πόντους στη νική, 1 στην ισοπαλία και 0 στην ήττα.

Τα xPTS υπολογίζονται για κάθε αγώνα ως εξής:
xPTS = 3 * P(νίκης) + 1 * P(ισοπαλίας)
όπου P: πιθανότητα

Για τις πιθανότητες έκβασης του αγώνα λαμβάνονται υπόψη κυρίως τα xGoals κάθε ομάδας.

4.7 Οι εταιρείες ανάλυσης δεδομένων ως <<ατζέντηδες>> ποδοσφαιριστών

Ο Kevin De Bruyne απευθύνθηκε σε εταιρεία ανάλυσης δεδομένων για να τον βοηθήσει με τις διαπραγματεύσεις για την ανανέωση του συμβολαίου του με τη Manchester City, αλλά και για να προβλέψουν αν η ομάδα του θα είναι ικανή για επιτυχίες τα επόμενα χρόνια με γνώμονα την ηλικία και την ποιότητα των συμπαικτών του. Ανανέωσε τον Απρίλιο του 2021 μέχρι το καλοκαίρι του 2025 λαμβάνοντας μια αύξηση του μισθού του περίπου κατά 50.000 ευρώ ανά βδομάδα **[99]**. Από τη στιγμή που ανανέωσε, η Manchester City έχει κατακτήσει 2 πρωταθλήματα Αγγλίας, 1 κύπελλο Αγγλίας, αλλά στέφθηκε και Πρωταθλήτρια Ευρώπης το 2023. **[100], [101]**

Μια άλλη περίπτωση είναι αυτή του Memphis Depay. Όταν αποχώρησε από την Manchester United, προσέλαβε εταιρεία ανάλυσης δεδομένων ώστε να του υποδείξουν πιθανές ομάδες που ταιριάζει να πάει ώστε να αναγεννήσει την καρίερα του. Ο Depay πήρε τη συμβουλή να πάει στη γαλλική Λυών και σε 178 αγώνες έβαλε 76 γκολ και μοίρασε 55 ασίστ. Ανέβασε την χρηματιστηριακή του αξία από 25.000.000€ στα 45.000.000€ και πήρε μεταγραφή στην Barcelona. **[94], [101]**

5. Εφαρμογές κώδικα

5.1 Πρόβλεψη αποτελέσματος αγώνα με βάση τα xGoals

Η συγκεκριμένη εφαρμογή θα εξετάσει αρχικά αν τα xGoals παρέχουν μια αξιόπιστη εικόνα για το τελικό αποτέλεσμα ενός αγώνα και στη συνέχεια ποιο μοντέλο κατηγοριοποίησης μπορεί να προβλέψει καλύτερα αποτελέσματα αγώνων με βάση τα xGoals.

Τα δεδομένα αγώνων είναι από την ιστοσελίδα FBREF **[103]**, η οποία τα λαμβάνει από την Opta. Αφορούν τα 4 καλύτερα πρωταθλήματα (Αγγλίας, Ισπανίας, Γερμανίας, Ιταλίας) με βάση την κατάταξη της UEFA **[104]** και τις 6 τελευταίες σεζόν (2017-2023). Συνολικά είναι 8674 αγώνες. Μετά από την κατάλληλη επεξεργασία, τα δεδομένα αποθηκεύτηκαν σε μορφή csv. Θα γίνει χρήση του Jupyter Notebook.

• **Εισαγωγή βιβλιοθηκών** **[105], [106], [107], [108], [109]**

```
# Οργάνωση και ανάλυση δεδομένων
import pandas as pd

# Αριθμητικοί υπολογισμοί σε πίνακες
import numpy as np

# Ορισμός μέγιστου αριθμού εμφανιζόμενων γραμμών και στηλών
pd.options.display.max_rows = 10000
pd.options.display.max_columns = 15

# Οπτικοποίηση δεδομένων
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns

# LabelEncoder
from sklearn import preprocessing

# Κατανομή των δεδομένων σε εκπαίδευσης και ελέγχου.
from sklearn.model_selection import train_test_split

# Decision Trees
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier

# K-Nearest Neighbors
from sklearn.neighbors import KNeighborsClassifier

# Random Forest
from sklearn.ensemble import RandomForestClassifier

# Gaussian Naive Bayes
from sklearn.naive_bayes import GaussianNB

# Performance Metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report
from sklearn.preprocessing import label_binarize
from sklearn.metrics import roc_curve, auc

# Cross Validation
from sklearn.model_selection import cross_val_score
```

• **Ανάγνωση αρχείου δεδομένων τύπου csv**
data=pd.read_csv('xGoals FBREF.csv')

• **Μελέτη δεδομένων**
data.shape
→ 8674 δείγματα και 9 χαρακτηριστικά

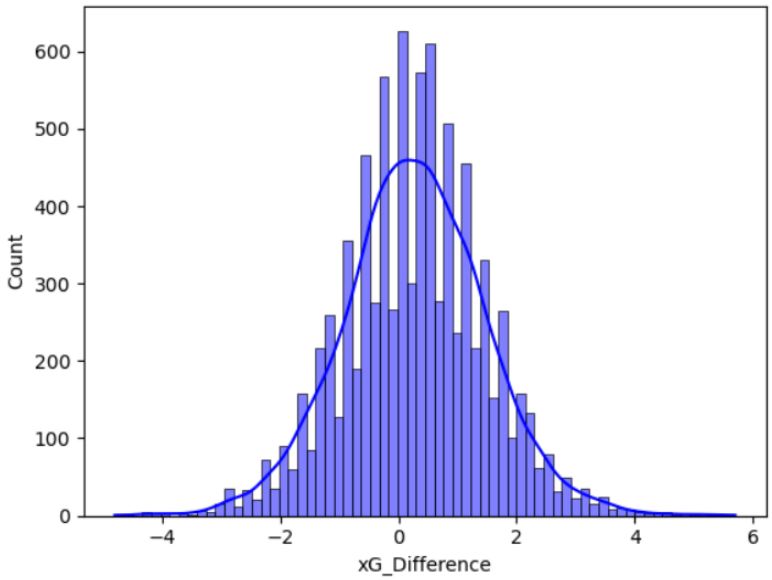
Εμφάνιση των 5 πρώτων γραμμών δεδομένων
data.head()

	Date	Home	Away	Home_Goals	Away_Goals	Result	xG_Home	xG_Away	xG_Difference
0	2017-08-11	Arsenal	Leicester City	4	3	1	2.5	1.5	1.0
1	2017-08-12	Watford	Liverpool	3	3	0	2.1	2.6	-0.5
2	2017-08-12	Crystal Palace	Huddersfield	0	3	2	1.1	1.5	-0.4
3	2017-08-12	West Brom	Bournemouth	1	0	1	1.3	0.5	0.8
4	2017-08-12	Chelsea	Burnley	2	3	2	1.5	0.6	0.9

Στήλες (Columns):
• Date: Ημερομηνία
• Home: Γηπεδούχος ομάδα
• Away: Φιλοξενούμενη ομάδα
• Home_Goals: Γκολ γηπεδούχου
• Away_Goals: Γκολ φιλοξενούμενης
• Result: Αποτέλεσμα (1=νίκη γηπεδούχου, 0=ισοπαλία, 2=νικη φιλοξενούμενης)
• xG_Home: xGoals γηπεδούχου
• xG_Away: xGoals φιλοξενούμενης
• xG_Difference: Διαφορά xGoals, δηλαδή xG_Difference = xG_Home – xG_Away

Καταμέτρηση μοναδικών τιμών για συγκεκριμένα χαρακτηριστικά
Κατανομή xG_Difference.

```
xG_Distribution = data.xG_Difference.value_counts()
sns.histplot(data=data, x='xG_Difference', kde=True, color='blue')
```

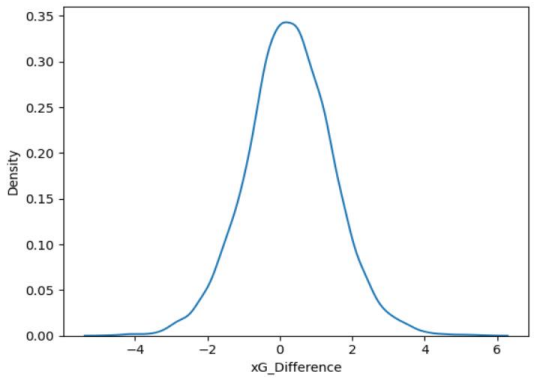


Από το παραπάνω ιστόγραμμα φαίνεται ότι πολλοί αγώνες έχουν την τάση προς την τιμή xG_Difference=0. Αυτό σημαίνει ότι υπάρχουν οριακοί αγώνες όπου δεν υπάρχει ξεκάθαρη υπεροχή κάποιας ομάδας. Σύμφωνα με την KDE που αναδεικνύει την κατανομή των δεδομένων σε μια συνεχή κλίμακα τιμών, η μέγιστη τιμή της καμπύλης KDE βρίσκεται στο x=0.19. Για αυτόν τον λόγο, θα ορίσουμε τους οριακούς αγώνες στο πεδίο [-0.2, 0.2], δηλαδή -0.2 ≤ xG_Difference ≤ 0.2.

Εύρεση τιμής x για τη μέγιστη τιμή της καμπύλης

```
x_values, y_values = sns.kdeplot(data['xG_Difference']).get_lines()[0].get_data()

# Εύρεση του αντίστοιχου x για τη μέγιστη τιμή του y
max_x = x_values[np.argmax(y_values)]
```

Η μέγιστη τιμή της καμπύλης KDE βρίσκεται στο x=0.19 .

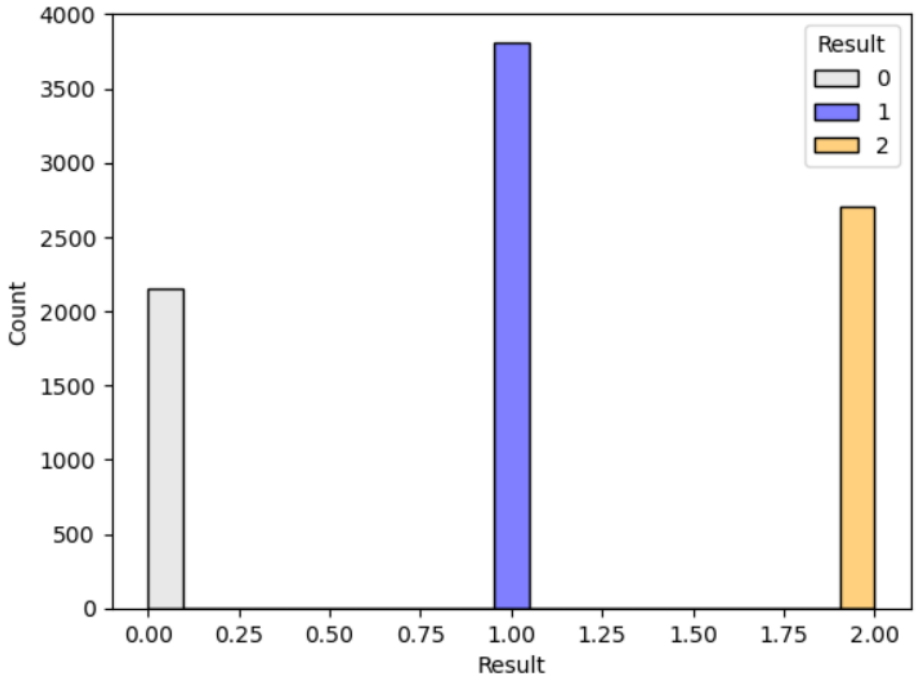


Μέσος όρος xG_Difference

```
mean_xG_Difference = data.xG_Difference.mean()
```

print("Ο μέσος όρος του xG Difference είναι:", round(mean_xG_Difference, 2))
Ο μέσος όρος του xG_Difference είναι 0.27.

Καταμέτρηση αποτελεσμάτων
data.Result.value_counts()
→ 3812 νίκες γηπεδούχου, 2153 ισοπαλίες και 2709 νίκες φιλοξενούμενης



α) Φιλτράρισμα πίνακα δεδομένων χρησιμοποιώντας συνθήκες

1. Υπεροχή γηπεδούχου (xG_Difference > 0.2)

Αριθμός αγώνων με υπεροχή γηπεδούχου: 4381

2. Οριακοί αγώνες (-0.2 ≤ xG_Difference ≤ 0.2)

Αριθμός οριακών αγώνων: 1480
Το ποσοστό των οριακών αγώνων ισούται περίπου με το 17% των συνολικών αγώνων.

3. Υπεροχή φιλοξενούμενης (xG_Difference < -0.2)

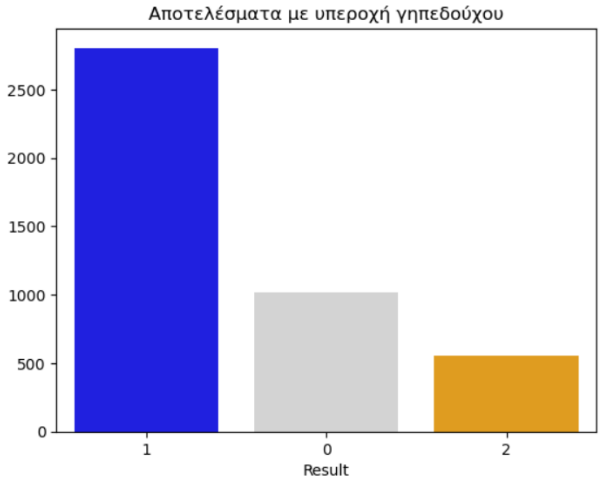
Αριθμός αγώνων με υπεροχή φιλοξενούμενης: 2813

Πίνακας 6: Σύνοψη αποτελεσμάτων ανάλογα το xG_Difference

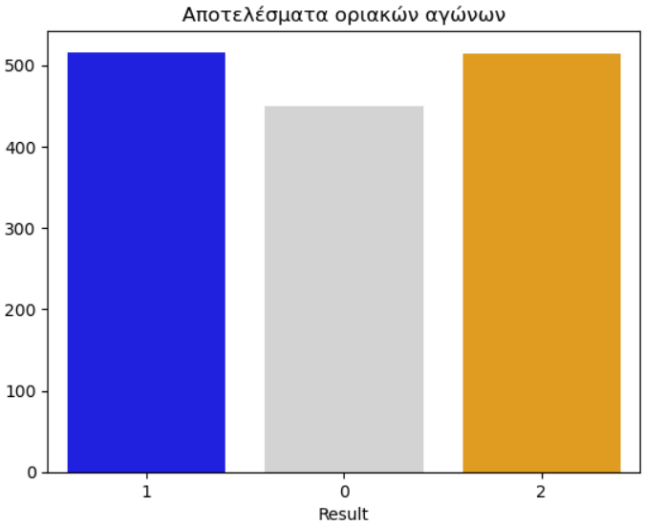
	1	0	2	Σύνολο
Υπεροχή γηπεδούχου	**2804**	1019	558	4381
Οριακοί αγώνες	516	**450**	514	1480
Υπεροχή φιλοξενούμενης	492	684	**1637**	2813
Σύνολο	3812	2153	2709	**8674**

1. Υπεροχή γηπεδούχου
Η υπεροχή του γηπεδούχου επιβεβαιώθηκε και στο αποτέλεσμα σε 2804 αγώνες, δηλαδή στο 64% των αγώνων που ο γηπεδούχος είχε περισσότερα xGoals, το οποίο είναι ένα αρκετά ικανοποιητικό ποσοστό.

Η υπεροχή του γηπεδούχου δεν επιβραβεύτηκε καθόλου σε 558 αγώνες από τους 4381, περίπου το 12.7%.

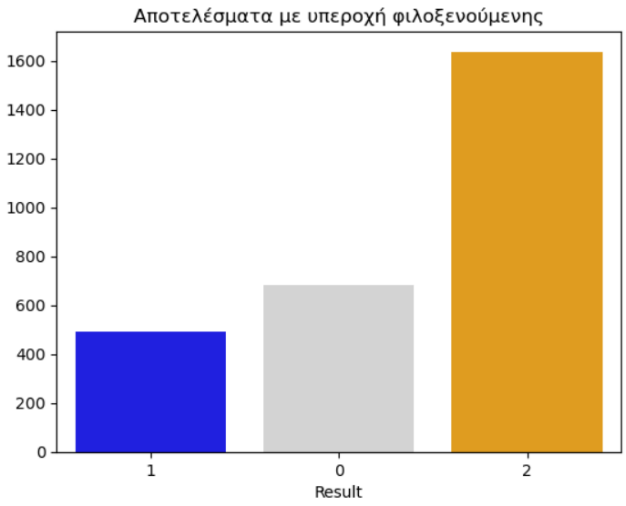


2. Οριακοί αγώνες
Παρατηρείται μια ισορροπημένη κατανομή των αποτελεσμάτων στους οριακούς αγώνες. Ουσιαστικά προκύπτει το συμπέρασμα ότι όταν ο αγώνας είναι οριακός, το αποτέλεσμα μπορεί να είναι οποιοδήποτε. Είναι αξιοσημείωτο οτί οι ισοπαλίες είναι λιγότερες από τα άλλα αποτελέσματα σε αυτήν τη συνθήκη.



3. Υπεροχή φιλοξενούμενης
Η υπεροχή της φιλοξενούμενης επιβεβαιώθηκε και στο αποτέλεσμα σε 1637 αγώνες, δηλαδή περίπου στο 58.2% των αγώνων που η φιλοξενούμενη ομάδα είχε περισσότερα xGoals, το οποίο είναι ένα αρκετά καλό ποσοστό.

Η υπεροχή της φιλοξενούμενης δεν επιβραβεύτηκε καθόλου σε 492 αγώνες από τους 2813, περίπου το 17.5%, ποσοστό που δεν θεωρείται μεγάλο αν ληφθεί υπόψη, ένας ανθρώπινος παράγοντας, ο παράγοντας της έδρας.

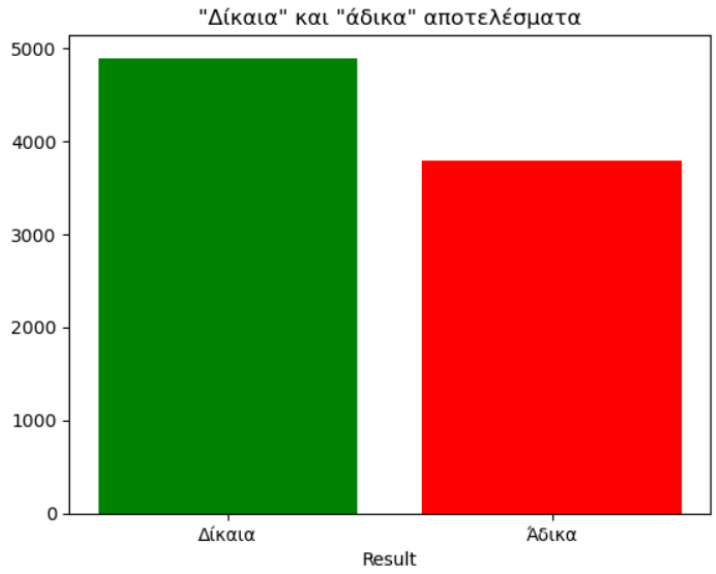


Δίκαια και άδικα αποτελέσματα
Θεωρούνται δίκαια τα αποτελέσματα όταν υπήρξε:

• Νίκη γηπεδούχου και xG_Difference > 0.2
• Ισοπαλία και -0.2 ≤ xG_Difference ≤ 0.2
• Νίκη φιλοξενούμενης και xG_Difference < -0.2

Άδικα θεωρούνται όλα τα άλλα αποτελέσματα.

Υπήρξε "δίκαιο" αποτέλεσμα σε 4891 από τους 8674 αγώνες, περίπου στο 56.4% των συνολικών αγώνων. Λαμβάνοντας υπόψη και τους οριακούς αγώνες, προκύπτει το συμπέρασμα ότι η υπεροχή μιας ομάδας στα xGoals της δίνει μια αρκετά μεγάλη πιθανότητα να κερδίσει και τον αγώνα ή τουλάχιστον να μη χάσει.



β) Ποιο μοντέλο κατηγοριοποίησης είναι πιο ικανό να προβλέψει τα αποτελέσματα ποδοσφαιρικών αγώνων με βάση τα xGoals;

Διαχωρισμός σε δεδομένα εκπαίδευσης και ελέγχου
X: Τα χαρακτηριστικά που θα χρησιμοποιηθούν ως είσοδοι στο μοντέλο.
y: Οι στήλες με τις ετικέτες για κάθε δείγμα ως έξοδοι του μοντέλου.

```
X = data['xG_Difference']
y = data['Result']
```

Χρήση του εργαλείου preprocessing της βιβλιοθήκης Scikit-learn για τη μετατροπή των κατηγορικών μεταβλητών 'xG_Difference' και 'Result' σε αριθμητικές μεταβλητές χρησιμοποιώντας τον LabelEncoder.

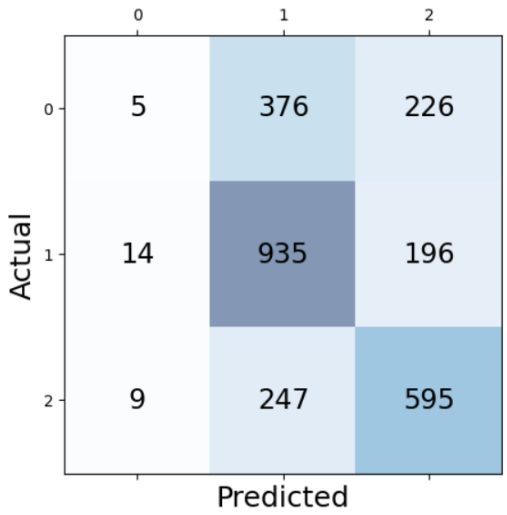
```
from sklearn import preprocessing

le = preprocessing.LabelEncoder()
X = le.fit_transform(data['xG_Difference'])
y = le.fit_transform(data['Result'])

X = np.array(X).reshape((-1, 1))
X,y
```

Χρήση του εργαλείου train_test_split της βιβλιοθήκης Scikit-learn για την κατανομή των δεδομένων σε εκπαίδευσης και ελέγχου

Επιλογή του 30% του συνόλου δεδομένων για έλεγχο και του 70% για την εκπαίδευση.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state=10)
```

1. Δέντρα Απόφασης (Decision Trees)



Πίνακας 7: Classification Report Decision Trees

	Precision	Recall	F1-Score	Support
0	0.18	0.01	0.02	607
1	0.60	0.82	0.69	1145
2	0.59	0.70	0.64	851
Accuracy			0.59	2603
Macro avg	0.45	0.51	0.45	2603
Weighted avg	0.50	0.59	0.52	2603

Result 0

α) Το μοντέλο προβλέπει αυτήν την κατηγορία ακόμα και αν δεν είναι απόλυτα σίγουρο. Αυτό μπορεί να έχει ως αποτέλεσμα την πρόβλεψη λάθος κατηγοριών. **(precision=0.18)**

β) Όταν αυτή η κατηγορία είναι δύσκολο να προβλεφθεί, το μοντέλο επιλέγει να μην πάρει το ρίσκο να προβλέψει λάθος. **(recall=0.01)**

Άρα, το μοντέλο έχει πολύ κακή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 0. **(F1-score=0.02)**

Αυτό κυρίως οφείλεται στον μικρό αριθμό των ισοπαλιών σε σχέση με τις άλλες κατηγορίες, που είναι μικρότερος ακόμα και στη συνθήκη των οριακών αγώνων.

Result 1

α) Όταν το μοντέλο προβλέπει την συγκεκριμένη κατηγορία την προβλέπει σωστά σε έναν μέτριο βαθμό. **(precision=0.60)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά. **(recall=0.82)**

Άρα, το μοντέλο έχει αρκετά καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 1. **(F1-score=0.69)**

Result 2

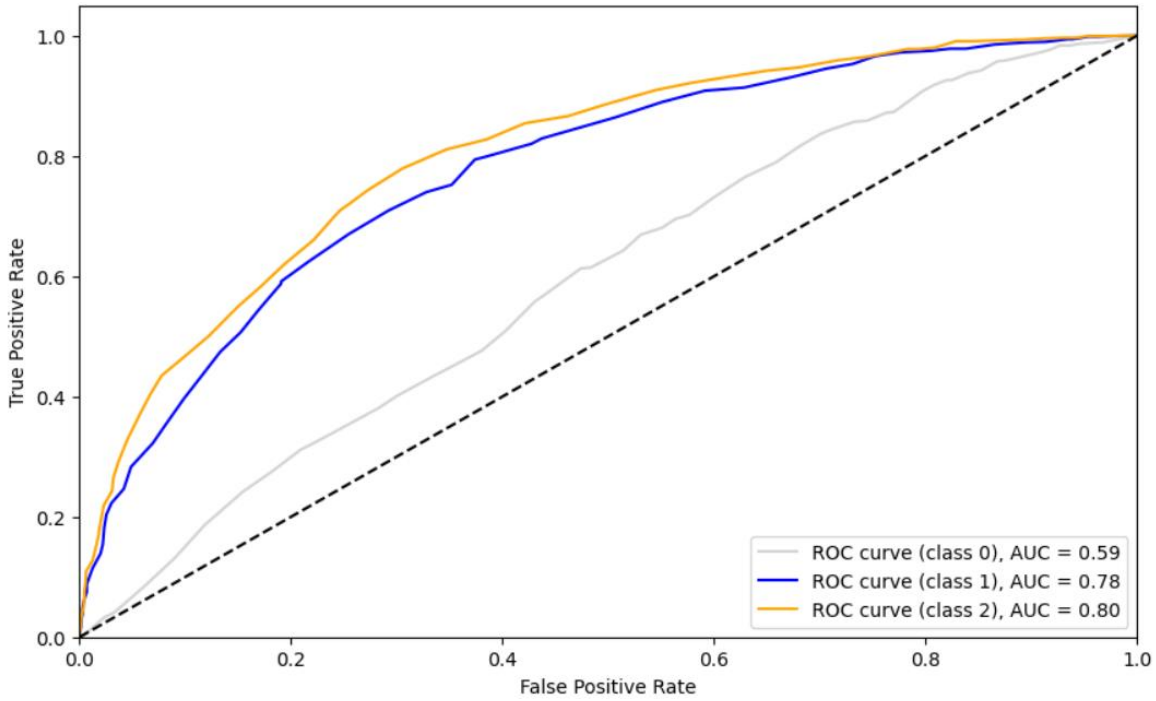
α) Το μοντέλο προβλέπει αυτήν την κατηγορία με μέτρια ακρίβεια. Είναι πολύ πιθανό να κάνει και λανθασμένες προβλέψεις. **(precision=0.59)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά με μια σχετικά υψηλή ακρίβεια. **(recall=0.70)**

Άρα, το μοντέλο έχει μια μέτρια προς καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 2. **(F1-score=0.64)**

Γενικά, το μοντέλο έχει μια μέτρια σταθμισμένη απόδοση **(weighted F1-score = 0.52)**.

Καμπύλη ROC



Result 0
AUC Score = 0.59 → Μέτρια ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 1
AUC Score = 0.78 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 2
AUC Score = 0.80 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 3 folds και η μέση επίδοση στο weighted F1-score είναι 0.491. Η απόδοση του μοντέλου είναι σχεδόν η ίδια είτε πρόκειται για καινούρια δεδομένα είτε για δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση. Άρα, το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα υπερπροσαρμογής (overfitting).

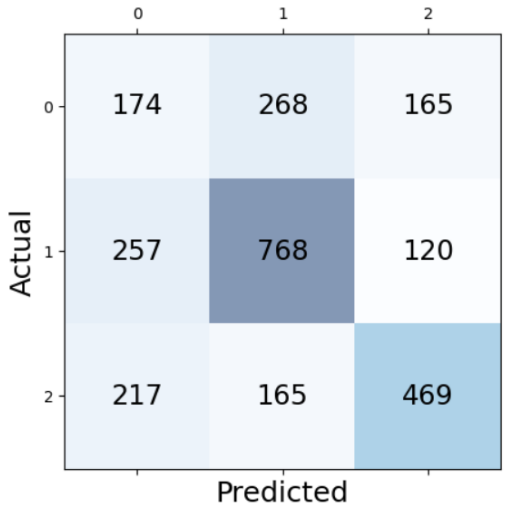
2. K-Nearest Neighbors

Κώδικας που αποτυπώνει τη σχέση του αριθμού του k με το weighted F1-Score του μοντέλου για τιμές k από 1 έως 15.

```
epidosiModelou = []
for k in range(1, 100):
    knnClass = KNeighborsClassifier(n_neighbors=k)
    knnClass.fit(X_train, y_train)
    pred = knnClass.predict(X_test)
    score = f1_score(y_test, pred, average='weighted')
    epidosiModelou.append(score)
print(f"Η καλύτερη επίδοση είναι {max(epidosiModelou):.3f} για K =
{epidosiModelou.index(max(epidosiModelou)) + 1}")
```

→ Η καλύτερη επίδοση είναι 0.544 για K=5

Άρα, επιλέγεται K=5 για την εκπαίδευση του μοντέλου.



Πίνακας 8: Classification Report k-Nearest Neighbors

	Precision	Recall	F1-Score	Support
0	0.27	0.29	0.28	607
1	0.64	0.67	0.65	1145
2	0.62	0.55	0.58	851
accuracy			0.54	2603
macro avg	0.51	0.50	0.51	2603
weighted avg	0.55	0.54	0.54	2603

Result 0

α) Το μοντέλο προβλέπει αυτήν την κατηγορία ακόμα και αν δεν είναι απόλυτα σίγουρο. Αυτό μπορεί να έχει ως αποτέλεσμα την πρόβλεψη λάθος κατηγοριών. **(precision=0.27)**

β) Όταν αυτή η κατηγορία είναι δύσκολο να προβλεφθεί, το μοντέλο επιλέγει να μην πάρει το ρίσκο να προβλέψει λάθος. **(recall=0.29)**

Άρα, το μοντέλο έχει χαμηλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 0. **(F1-score=0.28)**

Αυτό κυρίως οφείλεται στον μικρό αριθμό των ισοπαλιών σε σχέση με τις άλλες κατηγορίες, που είναι μικρότερος ακόμα και στη συνθήκη των οριακών αγώνων.

Result 1

α) Όταν το μοντέλο προβλέπει την συγκεκριμένη κατηγορία την προβλέπει σωστά σε έναν μέτριο βαθμό. **(precision=0.64)**

β) όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά με μια μέτρια ακρίβεια.**(recall=0.67)**

Άρα, το μοντέλο έχει μέτρια προς καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 1. **(F1-score=0.65)**

Result 2

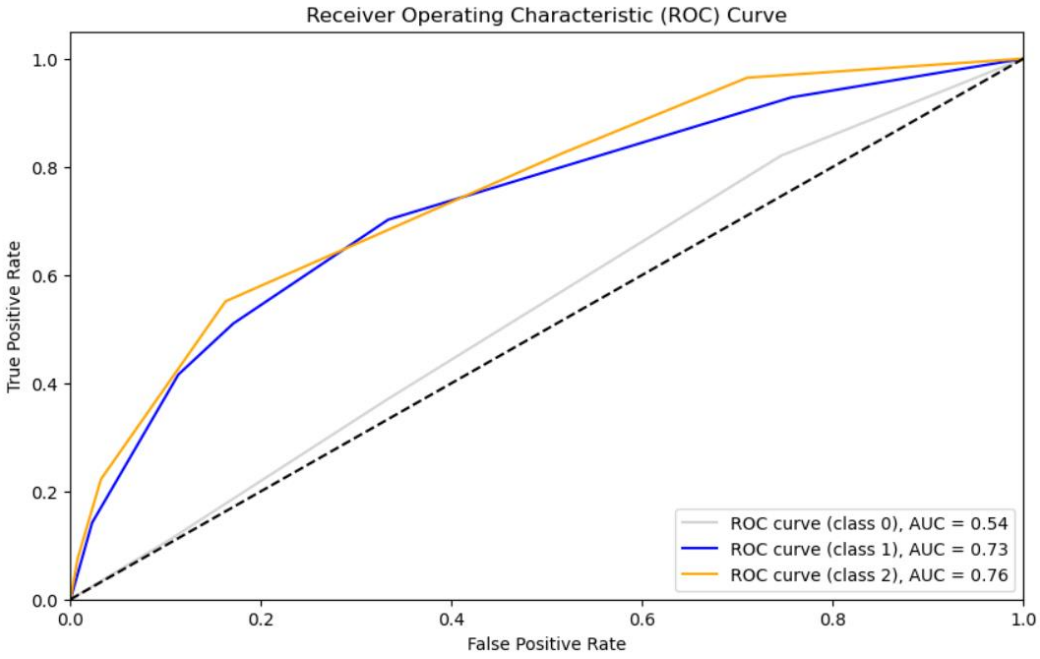
α) Το μοντέλο προβλέπει αυτήν την κατηγορία με μέτρια ακρίβεια. Είναι πολύ πιθανό να κάνει και λανθασμένες προβλέψεις. **(precision=0.62)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά με μια μέτρια ακρίβεια. **(recall=0.55)**

Άρα, το μοντέλο έχει μια μέτρια προς καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 2. **(F1-score=0.58)**

Γενικά, το μοντέλο έχει μια μέτρια σταθμισμένη απόδοση **(weighted F1-score=0.54)**.

Καμπύλη ROC



Result 0
AUC Score = 0.54 → Μέτρια ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

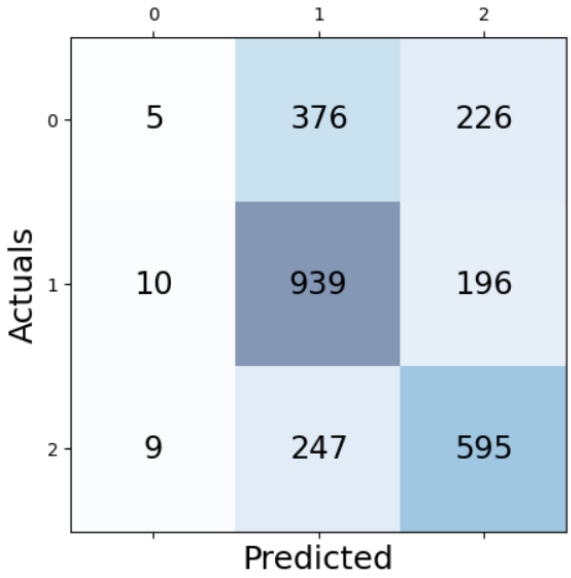
Result 1
AUC Score = 0.73 → Μέτρια προς υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 2
AUC Score = 0.80 → Μέτρια προς υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 11 folds και η μέση επίδοση στο weighted f1-score είναι 0.512. Η απόδοση του μοντέλου είναι σχεδόν η ίδια είτε πρόκειται για καινούρια δεδομένα είτε για δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση. Άρα, το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα υπερπροσαρμογής (overfitting).

3. Random Forest

Θα γίνει χρήση 500 εκτιμητών, δηλαδή 500 δέντρων απόφασης και 3 μέγιστων χαρακτηριστικών.



Πίνακας 9: Classification Report Random Forest

	Precision	Recall	F1-Score	Support
0	0.21	0.01	0.02	607
1	0.60	0.82	0.69	1145
2	0.59	0.70	0.64	851
accuracy			0.59	2603
macro avg	0.46	0.51	0.45	2603
weighted avg	0.50	0.59	0.52	2603

Result 0

α) Το μοντέλο προβλέπει αυτήν την κατηγορία ακόμα και αν δεν είναι απόλυτα σίγουρο. Αυτό μπορεί να έχει ως αποτέλεσμα την πρόβλεψη λάθος κατηγοριών. **(precision=0.21)**

β) Όταν αυτή η κατηγορία είναι δύσκολο να προβλεφθεί, το μοντέλο επιλέγει να μην πάρει το ρίσκο να προβλέψει λάθος. **(recall=0.01)**

Άρα, το μοντέλο έχει πολύ κακή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 0. **(F1-score=0.02)**

Result 1

α) Όταν το μοντέλο προβλέπει την συγκεκριμένη κατηγορία την προβλέπει σωστά σε έναν μέτριο βαθμό. **(precision=0.60)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά.**(recall=0.82)**

Άρα, το μοντέλο έχει αρκετά καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 1. **(F1-score=0.69)**

Result 2

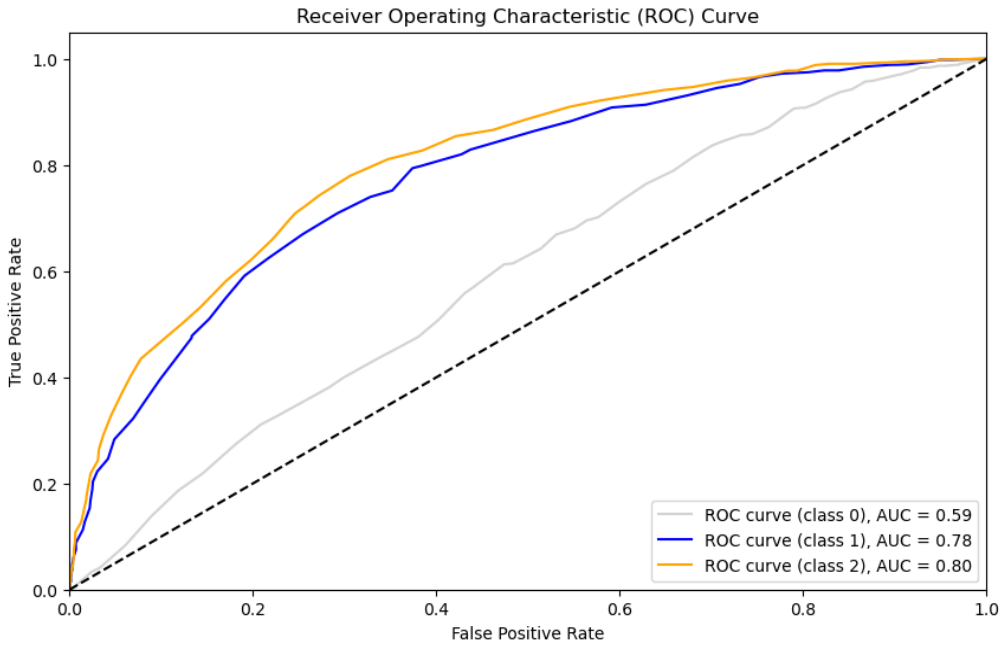
α) Το μοντέλο προβλέπει αυτήν την κατηγορία με μέτρια ακρίβεια. Είναι πολύ πιθανό να κάνει και λανθασμένες προβλέψεις. **(precision=0.59)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά με μια σχετικά υψηλή ακρίβεια. **(recall=0.70)**

Άρα, το μοντέλο έχει μια μέτρια προς καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 2. **(F1-score=0.64)**

Γενικά, το μοντέλο έχει μια μέτρια σταθμισμένη απόδοση **(weighted F1-score=0.52)**.

Καμπύλη ROC



Result 0

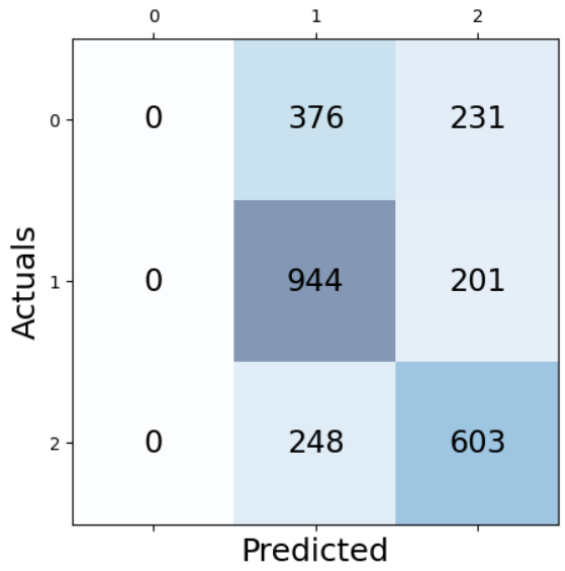
AUC Score = 0.59 → Μέτρια ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 1
AUC Score = 0.78 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 2
AUC Score = 0.80 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 3 folds και η μέση επίδοση στο weighted F1-score είναι 0.498. Η απόδοση του μοντέλου είναι σχεδόν η ίδια είτε πρόκειται για καινούρια δεδομένα είτε για δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση. Άρα, το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα υπερπροσαρμογής (overfitting).

4. Gaussian Naive Bayes



Πίνακας 10: Classification Report Gaussian Naive Bayes

	Precision	Recall	F1-Score	Support
0	0	0	0	607

1	0.60	0.82	0.70	1145
2	0.58	0.71	0.64	851
accuracy			0.59	2603
macro avg	0.39	0.51	0.45	2603
weighted avg	0.46	0.59	0.52	2603

Result 0
Το μοντέλο δεν ταξινομεί και δεν αναγνωρίζει καθόλου την κατηγορία 0. Αυτό αρχικά οφείλεται στον μικρό αριθμό των ισοπαλιών σε σχέση με τις άλλες κατηγορίες. Επίσης, ο αλγόριθμος Gaussian Naive Bayes είναι βασισμένος στην υπόθεση ανεξαρτησίας των χαρακτηριστικών. Άρα, το μοντέλο δεν έχει τη δυνατότητα να μάθει αποτελεσματικά τα χαρακτηριστικά που σχετίζονται με αυτήν την κατηγορία.

Result 1
α) Όταν το μοντέλο προβλέπει την συγκεκριμένη κατηγορία την προβλέπει σωστά σε έναν μέτριο βαθμό. **(precision=0.60)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά.**(recall=0.82)**

Άρα, το μοντέλο έχει αρκετά καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορία 1. **(F1-score=0.70)**

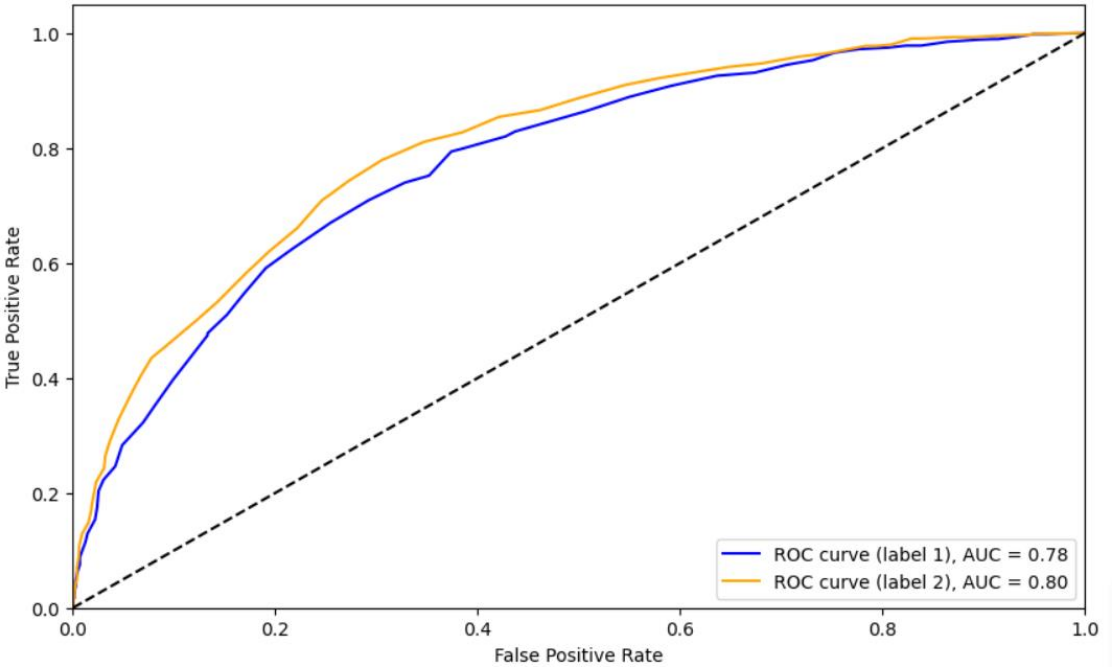
Result 2
α) Το μοντέλο προβλέπει αυτήν την κατηγορία με μέτρια ακρίβεια. Είναι πολύ πιθανό να κάνει και λανθασμένες προβλέψεις. **(precision=0.58)**

β) Όταν η συγκεκριμένη κατηγορία πρέπει να προβλεφθεί, το μοντέλο την προβλέπει σωστά με μια σχετικά υψηλή ακρίβεια. **(recall=0.71)**

Άρα, το μοντέλο έχει μια σχετικά καλή απόδοση στην ταξινόμηση και αναγνώριση της κατηγορίας 2. **(F1-score=0.64)**

Γενικά, το μοντέλο έχει μια μέτρια σταθμισμένη απόδοση **(weighted F1-score=0.52)**.

Καμπύλη ROC



Result 1
AUC Score = 0.78 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

Result 2
AUC Score = 0.80 → Υψηλή ικανότητα διαχωρισμού θετικών από αρνητικών παραδειγμάτων.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 3 folds και η μέση επίδοση στο weighted F1-score είναι 0.488. Η απόδοση του μοντέλου είναι σχεδόν η ίδια είτε πρόκειται για καινούρια δεδομένα είτε για δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση. Άρα, το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα υπερπροσαρμογής (overfitting).

Πίνακας 11: Σύγκριση μοντέλων κατηγοριοποίησης

	Weighted Precision	Weighted Recall	Weighted F1-Score	Cross-Validation Score
Decision Trees	0.50	0.59	0.52	0.491
K-Nearest Neighbors	**0.55**	**0.54**	**0.54**	**0.512**
Random Forest	0.50	0.59	0.52	0.498
Gaussian Naive Bayes	0.46	0.59	0.52	0.488

Το μοντέλο που θα επιλεχθεί είναι το k-Nearest Neighbors γιατί έχει καλύτερη σταθμισμένη απόδοση και καλύτερη μέση επίδοση γενίκευσης.

5.2 Πρόβλεψη των θέσεων που μπορεί να αποδώσει καλύτερα ένας ποδοσφαιριστής βάσει των χαρακτηριστικών του.

Η συγκεκριμένη εφαρμογή θα εξετάσει ποιο μοντέλο κατηγοριοποίησης μπορεί να προβλέψει καλύτερα τις πιο αποδοτικές θέσεις για έναν ποδοσφαιριστή βάσει των χαρακτηριστικών του.

Υπάρχουν πολλά παραδείγματα ποδοσφαιριστών που αγωνίζονται σε μια θέση και υπάρχει η αίσθηση ότι θα ήταν πιο αποδοτικοί σε μια άλλη θέση. Επίσης, δεν είναι λίγοι οι ποδοσφαιριστές που ξεκινούν την καριέρα τους από μια θέση και καταλήγουν να καθιερώνονται σε μία άλλη. Τέτοια παραδείγματα είναι και οι δύο καλύτεροι ποδοσφαιριστές στον κόσμο, ο Lionel Messi και ο Cristiano Ronaldo. Ο Lionel Messi ξεκίνησε την καριέρα του από τη θέση του δεξιού ακραίου επιθετικού (Right Winger) και πλέον αγωνίζεται πίσω από τον επιθετικό (Attacking Midfielder) σε έναν πιο ελεύθερο ρόλο για να αξιοποιηθούν όσο το δυνατόν περισσότερο τα επιθετικά του στοιχεία και να γίνει διαχείριση των τρεξιμάτων του. Ο Cristiano Ronaldo ξεκίνησε την καριέρα του από τη θέση του αριστερού ακραίου επιθετικού (Left Winger) και πλέον αγωνίζεται ως επιθετικός (Striker) λόγω της μείωσης της εκρηκτικότητας του, αλλά και της ικανότητας του να βάζει γκολ.

Για τους σκοπούς της εφαρμογής, θα χρησιμοποιηθούν δεδομένα από το βιντεοπαιχνίδι Football Manager 2023 **[110]** επειδή διαθέτει μια μεγάλη βάση δεδομένων και περιέχει πολλά χαρακτηριστικά των παικτών σε μια όσο γίνεται αντικειμενική αξιολόγηση. Τα χαρακτηριστικά βαθμολογούνται από το 0 έως το 20. Το ίδιο ισχύει και για τις θέσεις. Το κριτήριο κατηγοριοποίησης είναι ότι ο παίκτης θεωρείται ικανός να αγωνιστεί σε μια θέση αν η βαθμολογία του για αυτή τη θέση είναι μεγαλύτερη ή ίση του 15. Μετά από την κατάλληλη επεξεργασία, τα δεδομένα αποθηκεύτηκαν σε μορφή csv. και θα γίνει χρήση του Jupyter Notebook.

• **Εισαγωγή βιβλιοθηκών**

```
# Οργάνωση και ανάλυση δεδομένων
import pandas as pd

# Αριθμητικοί υπολογισμοί σε πίνακες
import numpy as np

# Ορισμός μέγιστου αριθμού εμφανιζόμενων γραμμών και στηλών
pd.options.display.max_rows = 10000
pd.options.display.max_columns = 100

# Οπτικοποίηση δεδομένων
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns

# Scaler για το χαρακτηριστικό Height
from sklearn.preprocessing import MinMaxScaler

# Κατανομή των δεδομένων σε εκπαίδευσης και ελέγχου.
from sklearn.model_selection import train_test_split

# K-Nearest Neighbors
from sklearn.neighbors import KNeighborsClassifier

# Random Forest
from sklearn.ensemble import RandomForestClassifier

# Performance Metrics
from sklearn.metrics import multilabel_confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report
from sklearn.preprocessing import label_binarize
from sklearn.metrics import roc_curve, auc

# Cross Validation
from sklearn.model_selection import cross_val_score
```

• **Ανάγνωση αρχείου δεδομένων τύπου csv**
data=pd.read_csv('FM Dataset.csv')

• **Αφαίρεση Goalkeepers**
Επειδή δε θα γίνει ταξινόμηση για τη θέση του τερματοφύλακα, αφαιρούνται και οι αντίστοιχοι ποδοσφαιριστές από το σύνολο δεδομένων.

data = data[data['Position'] != 'GK'].reset_index(drop=True)

• **Μελέτη δεδομένων**
data.shape
→ 7755 δείγματα και 59 χαρακτηριστικά

• **Scaling του ύψους στην κλίμακα 0-20**
scale_feature = ['Height']
scaler = MinMaxScaler(feature_range=(0, 20))
data[scale_feature] = scaler.fit_transform(data[scale_feature])

• **Εμφάνιση των 5 πρώτων γραμμών δεδομένων**

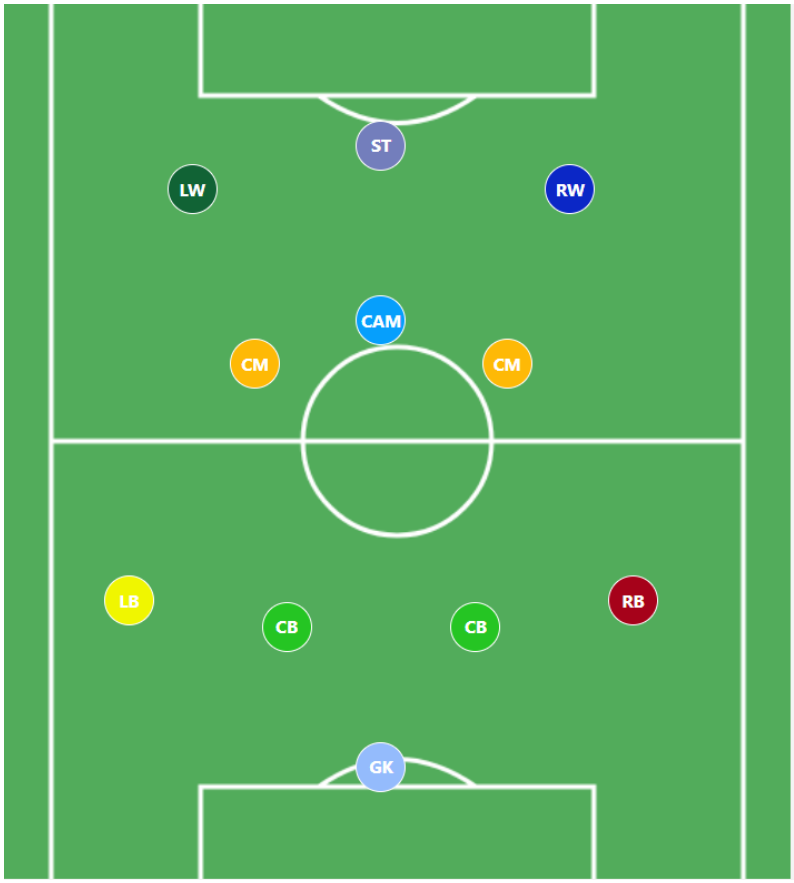
	Name	UniqueID	Position	CurrentAbility	Nationality	Club	Age	Height	LeftFoot	RightFoot	Corners	Crossing	Dribbling	Finishing
0	Kevin De Bruyne	18004457	M/AM RLC	189	Belgium	Manchester City	31	10.416667	16	20	14	19	15	16
1	Kylian Mbappé	85139014	AM/S RL	188	France	Paris Saint-Germain	23	9.166667	10	20	13	13	18	17
2	Robert Lewandowski	719601	S	186	Poland	Barcelona	33	12.083333	13	20	3	8	13	19
3	Erling Haaland	29179241	S	185	Norway,England	Manchester City	22	16.250000	20	11	7	10	14	18
4	Mohamed Salah	98028755	AM/S RL	185	Egypt	Liverpool	30	7.916667	20	8	12	14	17	17

• **Επιλογή χαρακτηριστικών (Feature Selection)**
Δε θα γίνει χρήση όλων των χαρακτηριστικών, γιατί κάποια αφορούν τον χαρακτήρα του ποδοσφαιριστή και όχι την ικανότητά του σε τεχνικά κομμάτια του αθλήματος.

Χαρακτηριστικά (Features)
1. Height → Ύψος
2. LeftFoot → Ικανότητα αριστερού ποδιού
3. RightFoot → Ικανότητα δεξιού ποδιού
4. Corners → Ικανότητα εκτέλεσης κόρνερ
5. Crossing → Σέντρες
6. Dribbling → Η ικανότητα να αποφεύγει τους αντιπάλους του με τη μπάλα.
7. Finishing → Ικανότητα σκοραρίσματος όταν παρουσιαστεί καλή ευκαιρία.
8. HeadingAccuracy → Ακρίβεια της κεφαλιάς
9. LongShots → Μακρινά σουτ
10. LongThrows → Μακρινά πλάγια
11. Marking → Ικανότητα μαρκαρίσματος των αντιπάλων
12. Passing → Ικανότητα πάσας
13. Tackling → Ικανότητα αναχαίτισης των αντιπάλων
14. Vision → Ικανότητα διαβάσματος του παιχνιδιού
15. Flair → Φυσικό ταλέντο για το δημιουργικό και το απρόβλεπτο
16. OffTheBallMovement → Κίνηση χωρίς τη μπάλα
17. Positioning → Τοποθέτηση στον αγωνιστικό χώρο
18. Acceleration → Επιτάχυνση
19. Jumping → Άλμα
20. Pace → Ταχύτητα
21. Strength → Δύναμη

Οι θέσεις που θα χρησιμοποιηθούν ως ετικέτες ταξινόμησης.
1. LeftBack → Αριστερός αμυντικός
2. CenterBack → Κεντρικός αμυντικός
3. RightBack → Δεξιός αμυντικός
4. CentralMidfielder → Κεντρικός μέσος
5. LeftWinger → Αριστερός εξτρέμ
6. AttackingMidfielder → Επιθετικός μέσος
7. RightWinger → Δεξιός εξτρέμ
8. Striker → Επιθετικός

Οι θέσεις που θα χρησιμοποιηθούν έχουν σημείο αναφοράς μια από τις πιο δημοφιλείς διατάξεις, το 4-3-3.



X: Τα χαρακτηριστικά που θα χρησιμοποιηθούν ως είσοδοι στο μοντέλο.
y: Οι στήλες με τις ετικέτες για κάθε δείγμα ως έξοδοι του μοντέλου.

X = data[features]
y = data[classification_labels]

Κριτήριο κατηγοριοποίησης
y[y<15] = 0
y[y>=15] = 1

Χρήση του εργαλείου train_test_split της βιβλιοθήκης Scikit-learn για την κατανομή των δεδομένων σε εκπαίδευσης και ελέγχου

Επιλογή του 25% του συνόλου δεδομένων για έλεγχο και του 75% για την εκπαίδευση.

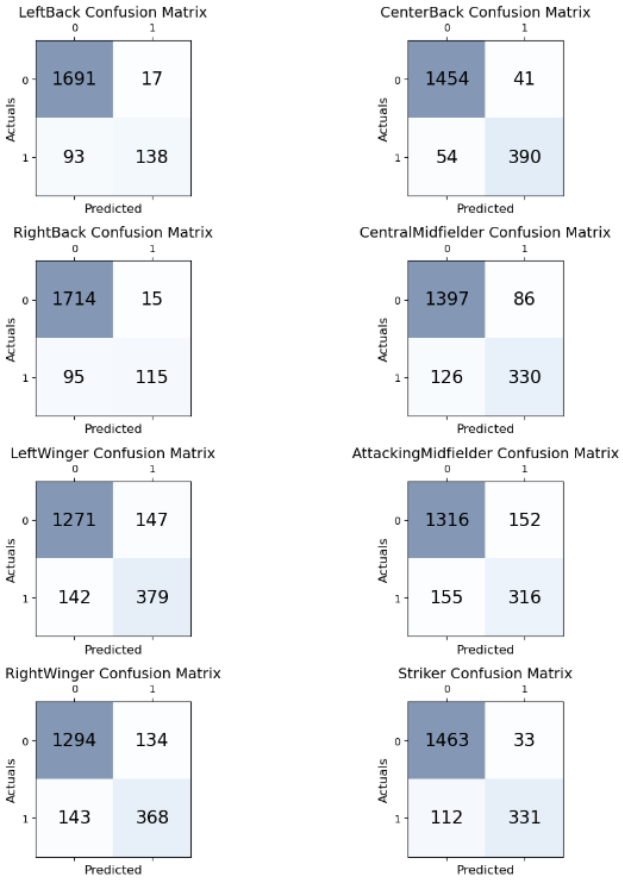
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state=10)
```

1. k-Nearest Neighbors

Κώδικας που αποτυπώνει τη σχέση του αριθμού του k με το weighted F1-Score του μοντέλου για τιμές k από 1 έως 15.

```
epidosiModelou = []
for k in range(1, 15):
    knnClass = KNeighborsClassifier(n_neighbors=k)
    knnClass.fit(X_train, y_train)
    pred = knnClass.predict(X_test)
    score = f1_score(y_test, pred, average='weighted')
    epidosiModelou.append(score)
print(f"Η καλύτερη επίδοση είναι {max(epidosiModelou):.3f} για K = {epidosiModelou.index(max(epidosiModelou)) + 1}")
```

→ Η καλύτερη επίδοση είναι 0.754 για K = 13
Άρα, επιλέγεται k=13 για την εκπαίδευση του μοντέλου.



Πίνακας 12: Classification Report k-Nearest Neighbors

	Precision	Recall	F1-Score	Support
LeftBack	0.89	0.60	0.72	231
CenterBack	0.90	0.88	0.89	444
RightBack	0.88	0.55	0.68	210
CentralMidfielder	0.79	0.72	0.76	456
LeftWinger	0.72	0.73	0.72	521
AttackingMidfielder	0.68	0.67	0.67	471
RightWinger	0.73	0.72	0.73	511
Striker	0.91	0.75	0.82	443
Micro avg	0.79	0.72	0.75	3287
Macro avg	0.81	0.70	0.75	3287
Weighted avg	0.80	0.72	0.75	3287
Samples avg	0.81	0.77	0.76	3287

Οι ετικέτες CenterBack και Striker έχουν υψηλό F1-Score, δηλαδή το μοντέλο μπορεί να ταξινομήσει και να αναγνωρίσει αυτές τις ετικέτες πάρα πολύ καλά.

Οι ετικέτες LeftBack, CentralMidfielder, LeftWinger και RightWinger έχουν μέτριο F1-Score, δηλαδή το μοντέλο μπορεί να ταξινομήσει και να αναγνωρίσει αυτές τις ετικέτες σχετικά καλά.

Οι ετικέτες RightBack και AttackingMidfielder έχουν κάτω από 0.70 F1-Score, δηλαδή το μοντέλο δυσκολεύεται λίγο παραπάνω να ταξινομήσει και να αναγνωρίσει αυτές τις κλάσεις σε σχέση με τις άλλες.

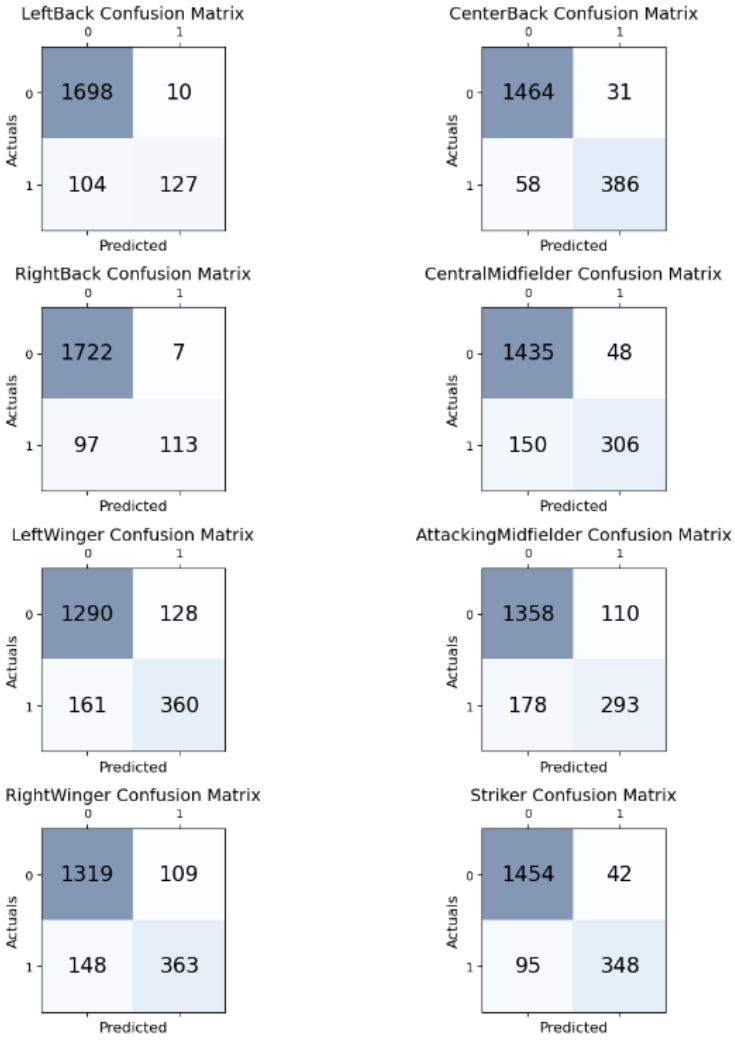
Το πολύ χαμηλότερο Recall στις ετικέτες LeftBack και RightBack σε σχέση με τα αντίστοιχα Precision οφείλεται στον μικρό αριθμό δειγμάτων που είχαν στο test set και το μοντέλο δυσκολεύεται να τις αναγνωρίσει, αλλά όταν είναι να τις προβλέψει έχει πολύ υψηλό Precision.

Το μοντέλο γενικά έχει μια μέτρια προς καλή σταθμισμένη απόδοση **(weighted F1-Score = 0.75)**.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 7 folds και η μέση επίδοση για το weighted F1-Score είναι 0.75.
Το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα overfitting. Τέλος, η απόδοση του μοντέλου είναι η ίδια είτε πρόκειται για καινούρια δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση είτε για δεδομένα που χρησιμοποιήθηκαν.

2. Random Forest

Θα γίνει χρήση 500 εκτιμητών, δηλαδή 500 δέντρων απόφασης και 5 μέγιστων χαρακτηριστικών.



Πίνακας 13: Classification Report Random Forest

	Precision	Recall	F1-Score	Support
LeftBack	0.93	0.55	0.69	231
CenterBack	0.93	0.87	0.90	444
RightBack	0.94	0.54	0.68	210
CentralMidfielder	0.86	0.67	0.76	456
LeftWinger	0.74	0.69	0.71	521
AttackingMidfielder	0.73	0.62	0.67	471
RightWinger	0.77	0.71	0.74	511
Striker	0.89	0.79	0.84	443
Micro avg	0.83	0.70	0.76	3287
Macro avg	0.85	0.68	0.75	3287
Weighted avg	0.83	0.70	0.75	3287
Samples avg	0.81	0.75	0.75	3287

Οι ετικέτες CenterBack και Striker έχουν υψηλό F1-Score, δηλαδή το μοντέλο μπορεί να ταξινομήσει και να αναγνωρίσει αυτές τις ετικέτες πάρα πολύ καλά.

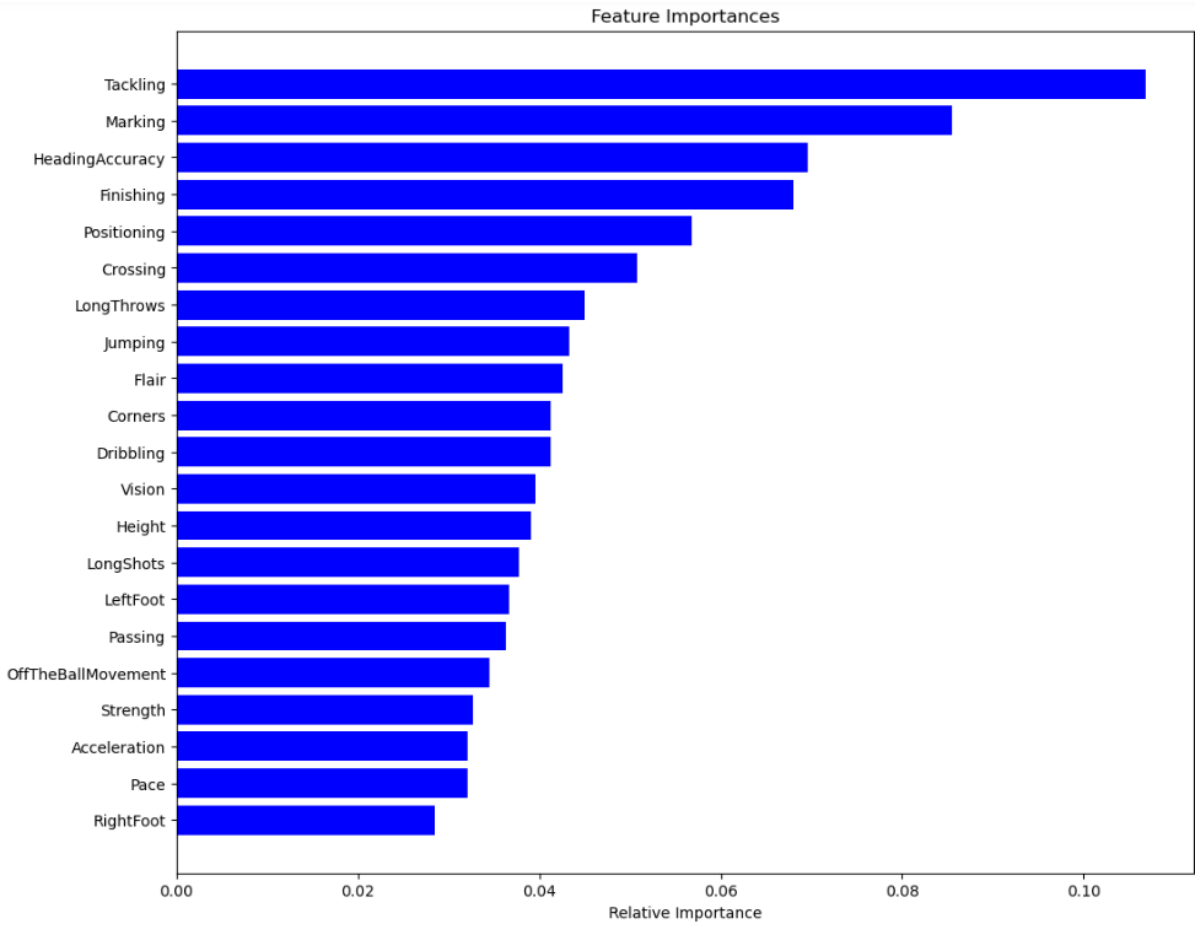
Οι ετικέτες CentralMidfielder, LeftWinger και RightWinger έχουν μέτριο F1-Score, δηλαδή το μοντέλο μπορεί να ταξινομήσει και να αναγνωρίσει αυτές τις ετικέτες σχετικά καλά.

Οι ετικέτες LeftBack, RightBack και AttackingMidfielder έχουν κάτω από 0.70 F1-Score, δηλαδή το μοντέλο δυσκολεύεται λίγο παραπάνω να ταξινομήσει και να αναγνωρίσει αυτές τις κλάσεις σε σχέση με τις άλλες.

Το πολύ χαμηλότερο Recall στις ετικέτες LeftBack και RightBack σε σχέση με τα αντίστοιχα Precision οφείλεται στον μικρό αριθμό δειγμάτων που είχαν στο test set και το μοντέλο δυσκολεύεται να τις αναγνωρίσει, αλλά όταν είναι να τις προβλέψει έχει πολύ υψηλό Precision.

Το μοντέλο γενικά έχει μια μέτρια προς καλή σταθμισμένη απόδοση **(weighted F1-Score = 0.75)**.

Σημασία χαρακτηριστικών (Feature Importances)



Weight	Feature
0.1068 ± 0.1351	Tackling
0.0855 ± 0.1249	Marking
0.0696 ± 0.0450	HeadingAccuracy
0.0680 ± 0.0693	Finishing
0.0567 ± 0.0961	Positioning
0.0508 ± 0.0211	Crossing
0.0450 ± 0.0133	LongThrows
0.0433 ± 0.0416	Jumping
0.0426 ± 0.0565	Flair
0.0413 ± 0.0286	Corners
0.0413 ± 0.0470	Dribbling
0.0396 ± 0.0150	Vision
0.0390 ± 0.0257	Height
0.0377 ± 0.0209	LongShots
0.0366 ± 0.0196	LeftFoot
0.0363 ± 0.0134	Passing
0.0345 ± 0.0292	OffTheBallMovement
0.0327 ± 0.0211	Strength
0.0321 ± 0.0126	Acceleration
0.0321 ± 0.0122	Pace
0.0285 ± 0.0194	RightFoot

Τα 5 πιο σημαντικά χαρακτηριστικά για την κατηγοριοποίηση:
• Tackling
• Marking
• HeadingAccuracy
• Finishing
• Positioning

• **Tackling, Marking, Positioning:** Είναι καθοριστικά στην κατηγοριοποίηση των αμυντικών και ειδικά των κεντρικών αμυντικών.

• **HeadingAccuracy:** Το υψηλότερο HeadingAccuracy το έχουν οι Strikers αλλά και οι κεντρικοί αμυντικοί επειδή είναι ψηλοί και πρωθούνται στα κόρνερ.

• **Finishing:** Το πιο σημαντικό χαρακτηριστικό για έναν Striker.

Για αυτό, οι κλάσεις CenterBack και Striker έχουν υψηλή απόδοση ταξινόμησης και αναγνώρισης.

Τα 5 λιγότερο σημαντικά χαρακτηριστικά για την κατηγοριοποίηση:
• OffTheBallMovement
• Strength
• Acceleration
• Pace
• RightFoot

Η κίνηση χωρίς τη μπάλα (OffTheBallMovement) δεν είναι από τα χαρακτηριστικά που καθορίζουν τη θέση ενός ποδοσφαιριστή. Συνήθως, την καλύτερη κίνηση χωρίς τη μπάλα την έχουν οι μεσοεπιθετικοί ποδοσφαιριστές (Attacking Midfielder, Winger, Striker). Αυτό επιβεβαιώνεται και στο σύνολο δεδομένων που χρησιμοποιήθηκε.

Το καλύτερο Strength στο σύνολο δεδομένων το έχουν κεντρικοί αμυντικοί, κεντρικοί μέσοι και επιθετικοί.

Το Acceleration και το Pace είναι πολύ σημαντικά για τους παίκτες που αγωνίζονται στα άκρα, είτε στην άμυνα είτε στην επίθεση. Βέβαια, στο σύγχρονο ποδόσφαιρο σχεδόν όλοι οι ποδοσφαιριστές είναι γρήγοροι και αθλητικοί, οπότε δικαιολογείται ότι αυτά τα δύο χαρακτηριστικά δεν είναι τοσο σημαντικά στην κατηγοριοποίηση. Στο σύνολο δεδομένων που χρησιμοποιήθηκε, οι πιο γρήγοροι παίκτες αγωνίζονται στις θέσεις των πλάγιων αμυντικών και του επιθετικού μέσου.

Οι ποδοσφαιριστές που έχουν ικανότητα αριστερού ποδιού μεγαλύτερη ή ίση του 15 είναι 2340, ενώ δεξιού ποδιού 5728. Οπότε, το μοντέλο δεν βρίσκει κάτι ιδιαίτερο στο χαρακτηριστικό RightFoot για να το ξεχωρίσει ως σημαντικό για την κατηγοριοποίηση. Μάλιστα, το θεωρεί ως το λιγότερο σημαντικό. Αξίζει να αναφερθεί ότι αυτό το χαρακτηριστικό παίζει ρόλο στις θέσεις των πλάγιων αμυντικών. Ανάλογα το πόδι, ο ποδοσφαιριστής αγωνίζεται και στην αντίστοιχη πλευρά. Αυτό δεν ισχύει στους Wingers, καθώς οι περισσότεροι αγωνίζονται με ευχέρεια και στις δύο πλευρές.

Τα Feature Importances δώσανε μια πολύ καλή εικόνα για τα αποτελέσματα στο classification report.

k-fold Cross-Validation για εκτίμηση της επίδοσης γενίκευσης του μοντέλου
Θα γίνει επιλογή 3 folds και η μέση επίδοση για το weighted F1-Score είναι 0.75.
Το μοντέλο είναι ικανό να γενικεύει σωστά σε καινούρια δεδομένα και δεν προκύπτει θέμα overfitting. Τέλος, η απόδοση του μοντέλου είναι η ίδια είτε πρόκειται για καινούρια δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση είτε για δεδομένα που χρησιμοποιήθηκαν.

Πίνακας 14: Σύγκριση μοντέλων κατηγοριοποίησης

	Weighted Precision	**Weighted Recall**	**Weighted F1-Score**	**Cross-Validation Score**
K-Nearest Neighbors	0.80	0.72	0.75	0.75
Random Forest	0.83	0.70	0.75	0.75

Τα δύο μοντέλα κατηγοριοποίησης που χρησιμοποιήθηκαν έχουν την ίδια σταθμισμένη απόδοση, αλλά και την ίδια μέση επίδοση γενίκευσης. Το k-Nearest Neighbors έχει ελαφρώς καλύτερη σταθμισμένη ανάκληση, ενώ το Random Forest ελαφρώς καλύτερη σταθμισμένη ευστοχία.

Για αυτό θα συγκρίνουμε και τις τιμές του μικρο-υπολογισμού που αποτελεί μια χρήσιμη μετρική για την επιλογή του μοντέλου. **[4]**

Πίνακας 15: Σύγκριση μοντέλων κατηγοριοποίησης με μικρο-υπολογισμό

	Micro Precision	**Micro Recall**	**Micro F1-Score**	**Cross-Validation Score**
K-Nearest Neighbors	0.79	0.72	0.75	0.751
Random Forest	**0.83**	**0.70**	**0.76**	**0.752**

Στον μικρο-υπολογισμό το μοντέλο Random Forest έχει ελαφρώς καλύτερη απόδοση από το k-Nearest Neighbors. Οπότε, το καλύτερο μοντέλο από τα δύο για την πρόβλεψη θέσης ενός ποδοσφαιριστή βάσει των χαρακτηριστικών του, είναι το Random Forest.

Βέβαια, η επιλογή του μοντέλου μπορεί να γίνει και ανάλογα τι παίκτες ψάχνει μια ομάδα είτε πρόκειται για μεταγραφή είτε για αυτούς που έχει ήδη. Για παράδειγμα, αν ψάχνει για αριστερό μπακ, μπορεί να διαλέξει το μοντέλο k-Nearest Neighbors επειδή στη συγκεκριμένη θέση έχει ελαφρώς καλύτερη απόδοση. Για τον ίδιο λόγο, αν ψάχνει για επιθετικό (Striker), μπορεί να διαλέξει το μοντέλο Random Forest.

6. Συμπεράσματα και μελλοντικές προτάσεις

Παρά τις λύσεις που προσφέρει η μηχανική μάθηση στο ποδόσφαιρο, είναι πολλόι αυτοί που διαφωνούν με τη χρήση της βασιζόμενοι στο ότι δεν είναι όλα αριθμοί και αλγόριθμοι και δε μπορούν να προβλεφθούν χαρακτηριστικά όπως είναι η ψυχολογία, το ένστικτο και η εμπειρία. Επίσης, αν τα δεδομένα είναι ανεπαρκή ή δεν έχουν αναλυθεί σωστά υπάρχει μεγάλη πιθανότητα να μη προκύψει η κατάλληλη απόφαση. Δημοφιλής άποψη είναι και ότι εργαλεία όπως το VAR που λαμβάνουν υπόψη την παραμικρή λεπτομέρεια, αλλοιώνουν τη μαγεία του αθλήματος λόγω της διστακτικότητας στους πανηγυρισμούς μήπως και το γκολ ακυρωθεί τελικά, αλλά και επειδή οι αποφάσεις μεταξύ παρόμοιων φάσεων, δεν είναι πάντα οι ίδιες. **[113], [114]**

Πράγματι, η βασισμένη σε δεδομένα προσέγγιση δε μπορεί να λειτουργήσει ως υποκατάστατο του επιδέξιου παιχνιδιού και της εξαιρετικής προπονητικής. Παρόλα αυτά, έχει καθιερωθεί ως μια σημαντική βελτίωση για αυτούς τους βασικούς παράγοντες επιτυχίας ενώ ακόμα όλο αυτό είναι σε πρώιμο στάδιο. Ακόμη και αν μειώνεται η μαγεία του αθλήματος κάποιες φορές, έχει εξασφαλιστεί σε μεγάλο βαθμό η διαφάνεια στις διαιτητικές αποφάσεις και αποφεύγονται εξόφθαλμα λάθη, όπως να μετρήσει ένα γκολ που μπήκε με το χέρι ή να μη μετρήσει γκολ που η μπάλα πέρασε ολόκληρη τη γραμμή τέρματος.

Μελλοντικά, υπάρχει η δυνατότητα να υλοποιηθούν και άλλες εφαρμογές που θα βελτιώσουν το επίπεδο του αθλήματος. Μία από αυτές περιλαμβάνει την τεχνολογία εικονικής πραγματικότητας (Virtual Reality Technology) η οποία ήδη έχει κάνει κάποια μικρά βήματα στον αθλητισμό προσφέροντας εργαλεία για ατομική προπόνηση όπως προσομοίωση συνθηκών αγώνα με σκοπό τη βελτίωση στη λήψη αποφάσεων. Από αυτήν τη διαδικασία, μπορούν να συλλεχθούν χρήσιμα δεδομένα για τους προπονητές **[117]**. Όμως, εκεί που με τη χρήση μηχανικής μάθησης μπορούν να γίνουν σημαντικές αλλαγές είναι η διαιτησία, με εφαρμογή τεχνολογιών που θα μειώσουν και άλλο τους χρόνους λήψης διαιτητικών αποφάσεων. Προτάσεις για το αμέσο μέλλον της διαιτησίας περιλαμβάνουν τη χρήση αισθητήρων στις μπάλες, όπως στο Παγκόσμιο Κύπελλο 2022, αλλά και τεχνολογία αυτόματου offside, παρόμοια με την τεχνολογία γραμμής τέρματος.

Βιβλιογραφία – Αναφορές – Διαδικτυακές Πηγές

[1] Γ. Νικολάου, Επιχειρηματική Ευφυΐα και Ανάλυση Μεγάλων Δεδομένων. **https://eclass.uniwa.gr/courses/IDPE184/**

[2] Federation Internationale de Football Association (FIFA), Laws of the Game https://digitalhub.fifa.com/m/3f3e15cc1ab8977b/original/datdz0pms85gbnqy4j3k-pdf

[3] https://www.sportzcraazy.com/football-field-dimensions/

[4] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου, Τεχνητή Νοημοσύνη, Δ΄ Έκδοση. : Εκδόσεις Πανεπιστημίου Μακεδονίας, 2020.

[5] Reinforcement Learning

[6] https://www.smartdraw.com/decision-tree/

[7] Random Forest

[8] Μαθηματικά Β΄ Γενικού Λυκείου, Θετικής και Τεχνολογικής Κατεύθυνσης, Ινστιτούτο Τεχνολογίας Υπολογιστών και Εκδόσεων Διόφαντος, 2013.

[9] https://study4maths.gr/2020/02/10/αποσταση-δυο-σημειων/

[10] https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

[11] Ε.Δημητριάδης, Στατιστική Επιχειρήσεων με Εφαρμογές σε SPSS και LISREL, 2η έκδοση, Εκδόσεις Κριτική, 2016.

[12] The Normal Distribution

[13] Ε.Κύρκος, Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων, Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις, 2015.

[14] https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

[15] https://medium.com/the-owl/evaluation-metrics-part-3-47c315e07222

[16] Micro, Macro & Weighted Averages of F1 Score

[17] https://www.askpython.com/python/examples/k-fold-cross-validation

[18] https://www.simplilearn.com/types-of-business-analytics-tools-examples-jobs-article

[19] https://www.youtube.com/watch?v=w7zPZsLGK18

[20] https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/

[21] https://theanalyst.com/eu/2022/04/evolving-expected-goals-xg/

[22] https://www.youtube.com/watch?v=H4kNa1cUvZM

[23] https://theanalyst.com/eu/2021/03/what-are-expected-assists-xa/

[24] Sports Analytics: How Different Sports Use Data Analytics

[25] https://the-elastico.com/how-is-football-data-collected/

[26] https://www.youtube.com/watch?v=ujakhyFWQ8E&list=WL&index=16

[27] https://www.sportperformanceanalysis.com/article/gps-in-professional-sports

[28] https://blog.isportsanalysis.com/the-advantages-of-using-gps-data-in-football/

[29] https://www.economist.com/1843/2018/04/24/how-gps-tracking-is-changing-football

[30] Official Match Ball of the FIFA World Cup™ 2022

[31] https://www.fifa.com/fifaplus/en/watch/weFbPIyOvE2WX4b-1hJzPw

[32] https://www.youtube.com/watch?v=TvgHs9FZzf8&list=WL&index=37

[33] Adidas technology proves Portugal captain did not score

[34] https://theathletic.com/3958521/2022/12/01/japan-goal-var-spain-world-cup-2022/

[35] Ε. Λελίγκου, Νεφοϋπολογιστική Μηχανική. https://eclass.uniwa.gr/courses/IDPE301/

[36] https://aws.amazon.com/what-is-cloud-computing/?nc1=f_cc

[37] https://aws.amazon.com/what-is-aws/?nc1=f_cc

[38] https://aws.amazon.com/solutions/case-studies/?hp=tile&tile=customerstories

[39] https://azure.microsoft.com/en-us/

[40] Microsoft Azure Customer Stories

[41] https://www.oracle.com/cloud/

[42] https://www.oracle.com/customers/

[43] Futbol Club Barcelona Case Study

[44] https://news.microsoft.com/europe/features/the-transformation-of-real-madrid/

[45] How Real Madrid scores fan engagement in the cloud

[46] Valencia CF creates hyper-personalized fan experiences with Dynamics 365

[47] https://www.oracle.com/customers/rcd-1-financials-cl/

[48] https://www.oracle.com/customers/seattle-sounders-fc/

[49] https://aws.amazon.com/sports/bundesliga/

[50] https://www.dfl.de/en/topics/match-data/bundesliga-match-facts/

[51] https://aws.amazon.com/sports/bundesliga/most-pressed-player/

[52] Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., Von Landesberger, T. and Weber, H. (2017). Visual Analysis of Pressure in Football. Data Mining and Knowledge Discovery, 31(6), pp. 1793-1839. doi: 10.1007/s10618-017-0513-2

[53] https://aws.amazon.com/sports/bundesliga/attacking-zones/

[54] https://www.youtube.com/watch?v=jSkvYOiAWKA

[55] Bundesliga Match Fact Skill: Quantifying football player qualities

[56] The tech behind the Bundesliga Match Facts xGoals

[57] https://www.youtube.com/watch?v=_vGhocyvKhA

[58] Bundesliga Match Fact Win Probability

[59] https://www.youtube.com/watch?v=419B95mTXEA

[60] The development of Bundesliga Match Fact Passing Profile

[61] https://www.youtube.com/watch?v=QERuLHVk9uA

[62] Bundesliga Match Fact Set Piece Threat

[63] Bundesliga Match Fact Pressure Handling

[64] Bundesliga Match Fact Keeper Efficiency

[65] Bundesliga Match Fact Ball Recovery Time

[66] https://www.oracle.com/uk/premier-league/

[67] https://www.premierleague.com/partners/oracle

[68] https://customers.microsoft.com/en-us/story/laliga-media-entertainment-azure

[69] https://www.youtube.com/watch?v=oZwtUXzXUi4

[70] Performance Analysis

[71] Football Games Analysis from video stream with Machine Learning

[72] A New Way Of Classifying Team Formations in Football

[73] https://www.spiideo.com/news/2d-tactical-maps/

[74] https://www.sportperformanceanalysis.com/article/artificial-intelligence-ai-in-sports

[75] https://blog.tryoliver.com/en/heat-maps-oliver

[76] https://www.sofascore.com/player/konstantinos-tsimikas/786259

[77] Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Υπολογιστική Γεωμετρία, Comp_Geometry_Chapter4

[78] https://footballatif.com/how-manchester-city-use-voronoi-diagram/

[79] https://www.youtube.com/watch?v=dP2cgT8698o

[80] https://www.nacsport.com/index.php?lc=en-gb

[81] https://www.youtube.com/watch?v=h_8U6-gUZFk&t=33s

[82] https://www.acmilan.com/en/club/venues/vismara/milan-lab

[83] https://soccerment.com/the-importance-of-football-analytics/

[84] https://www.bbc.com/sport/football/29361839

[85] UEFA Club Licensing and Financial Fair Play Regulations

[86] Goal-line Technology

[87] John Stones' goal-line clearance

[88] Video Assistant Referee (VAR)

[89] What is VAR in football and when is it used?

[90] Semi-automated offside technology

[91] https://dataconomy.com/2022/11/21/saot-semi-automated-offside-technology-2022/

[92] How Brentford's Moneyball Approach Works

[93] The Brentford FC Story: Running a football club through data

[94] https://www.transfermarkt.com/

[95] https://theanalyst.com/eu/2022/06/english-premier-league-2021-22-stats/

[96] https://theanalyst.com/eu/2022/10/premier-league-stats-2022-23/

[97] https://understat.com/league/EPL

[98] https://soccerment.com/soccerments-expected-points-model/

[99] Kevin De Bruyne Cash Earnings

[100] De Bruyne’s new contract agreed after data analysts convinced him

[101] How Analytics FC helped De Bruyne negotiate new Man City deal

[102] De Bruyne Champions League Trophy

[103] https://fbref.com/en/

[104] https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2023

[105] https://pandas.pydata.org/docs/

[106] https://numpy.org/

[107] https://matplotlib.org/

[108] https://seaborn.pydata.org/

[109] https://scikit-learn.org/stable/

[110] https://www.kaggle.com/datasets/platinum22/foot-ball-manager-2023-dataset

[111] https://createformation.com/create

[112] https://eli5.readthedocs.io/en/latest/overview.html

[113] Impart of VAR to Football and Controversy

[114] Soccer’s Video Assistant Referee has its pros and cons

[115] https://gr.pinterest.com/pin/533324780853003520/

[116] In Pictures: Frank Lampard's Disallowed Goal v Germany

[117] https://www.besoccer.com/new/how-is-vr-being-used-in-football-1248355

[118] https://mplsoccer.readthedocs.io/en/latest/#

[119] https://www.youtube.com/watch?v=BUKsBH9oQKg&t=1073s

Παράρτημα

Ο κώδικας των εφαρμογών στο Κεφάλαιο 5, αλλά και για τα διαγράμματα Voronoi **[118], [119]** στην ενότητα 4.1 βρίσκονται στον παρακάτω σύνδεσμο:
https://github.com/vassilis17080/ML-in-Football