Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΠΡΟΒΛΕΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΑΓΩΝΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΤΙΚΗΣ ΔΕΔΟΜΕΝΩΝ ΑΘΛΗΤΩΝ

Σταυρούλα Κωνσταντίνου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Οκτώβριος 2022

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής Ν. Πελέκης (Επιβλέπων)
- Αναπληρωτής Καθηγητής Ελ. Κοφίδης
- Αναπληρωτής Καθηγητής Σ. Μπερσίμης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

MATCH OUTCOME FORECASTING WITH SPORT ANALYTIC TECHNIQUES

Stavroula Konstantinou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
October 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Νίκο Πελέκη, τους γονείς μου και τον αδερφό μου, οι οποίοι με στήριξαν, ο καθένας με τον δικό του ξεχωριστό τρόπο σε αυτή τη προσπάθεια.

Περίληψη

Στην εποχή της ραγδαίας μετάδοσης της πληροφορίας και της συνεχούς εξέλιξης των μεθόδων συλλογής και ανάλυσης δεδομένων αθλητών, η παγκόσμια βιομηχανία του ποδοσφαίρου δε θα μπορούσε να μείνει αμέτοχη. Υπέρογκα χρηματικά ποσά διατίθενται σε αναλύσεις για κατάλληλο σχηματισμό ομάδων, και σχεδιασμό παιχνιδιών. Μεγάλα χρηματικά ποσά επενδύονται στην εξόρυξη γνώσης από αθλητικά δεδομένα, με σκοπό την αύξηση κερδών και θεαματικότητας. Οι μέθοδοι αξιολόγησης παικτών είναι ένας τομέας ανάλυσης που βοηθά στη χάραξη στρατηγικών και στη διαμόρφωση ισχυρών ποδοσφαιρικών ομάδων.

Στη μελέτη αυτή εφαρμόσαμε ένα καινοτόμο πλαίσιο αξιολόγησης και βαθμολογικής κατάταξης παικτών ποδοσφαίρου, το οποίο θεμελιώνεται σε μεθόδους μηχανικής μάθησης, και τροφοδοτείται με έναν συγκεκριμένο τύπο δεδομένων, τα λεγόμενα soccer-logs. Τα αποτελέσματα των αλγορίθμων έδειξαν ότι το πλαίσιο αυτό αποδίδει καλύτερα σε μεγάλου όγκου δεδομένα τέτοιας μορφής.

Λέξεις κλειδιά: ποδόσφαιρο, εξόρυξη γνώσης, μηχανική μάθηση, soccer-logs.

Abstract

In the era of rapid information transfer and the continuous development of sports data collection and analytics methods, the global soccer industry could not be uninvolved. Exorbitant amounts of money are invested in the proper team formation and game planning analysis. Considerable amounts of money are invested in sports data mining, with the intention of profit and ratings increase. Football players evaluation is an analysis field that helps in strategic planning and constructing powerful teams.

In this study we implement an innovative framework of rating and ranking football players which is based on machine learning algorithms and accepts a specific data type as input, the so-called soccer-logs. The results showed that this framework performs better on big data of this type.

Key words: soccer, data mining, machine learning, soccer-logs.

Περιεχόμενα

Κατάλογος Σχημάτων

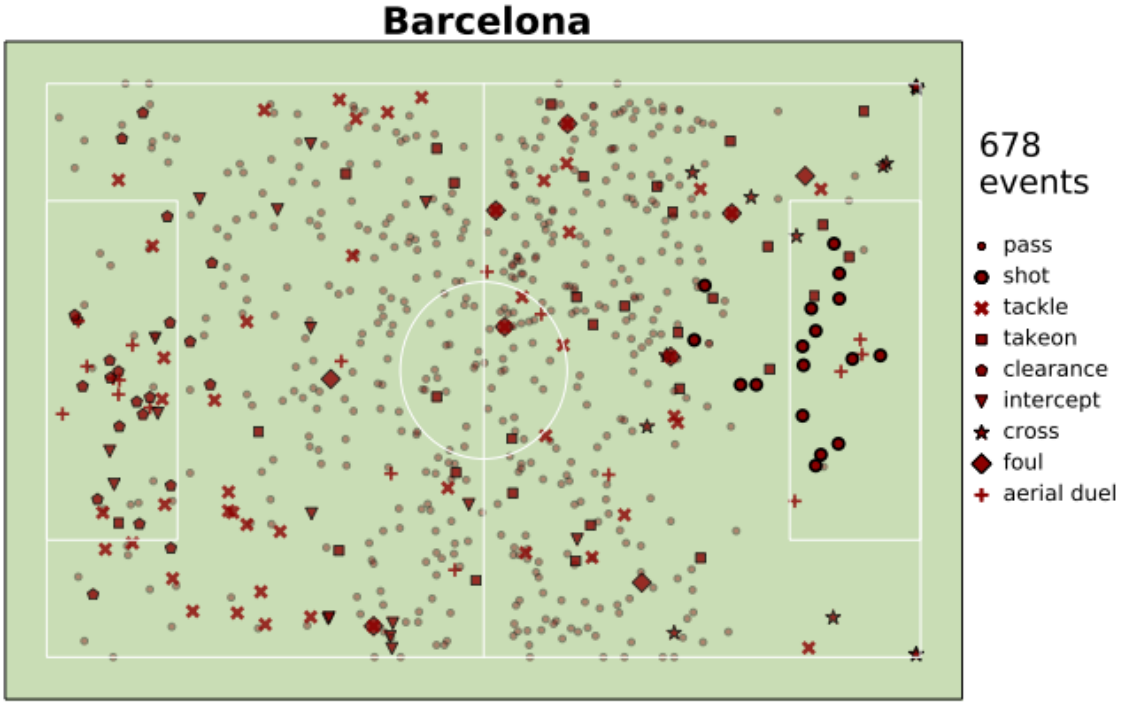
Κατάλογος Πινάκων

1. Βιβλιογραφική επισκόπηση

1.1. Εισαγωγή

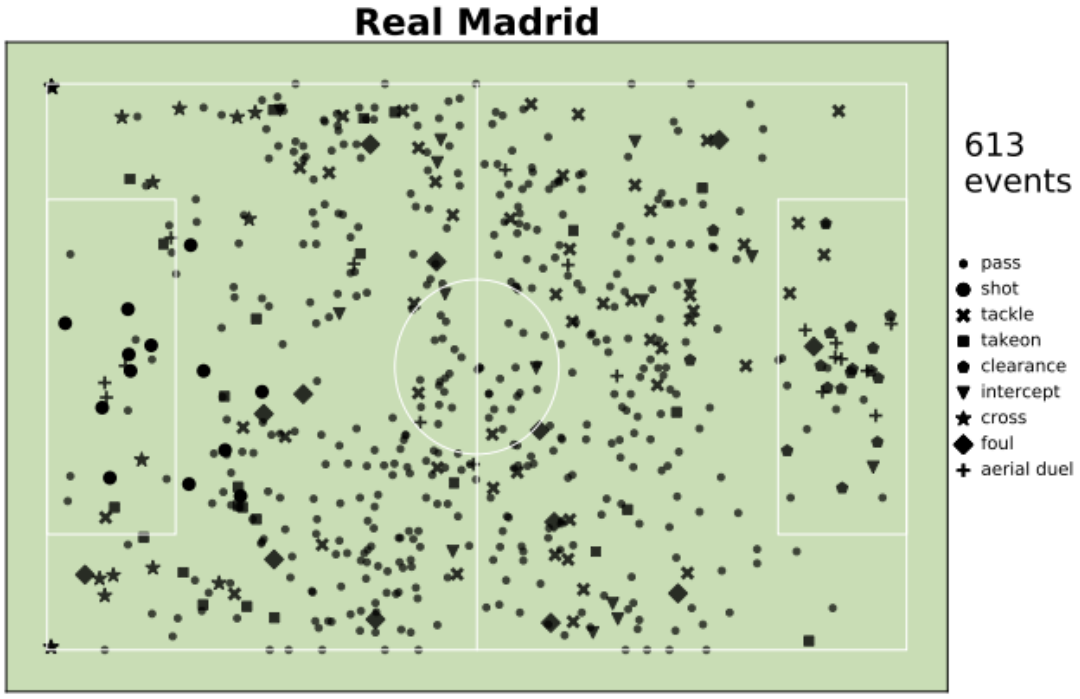
Η ποσοτική ανάλυση στο ποδόσφαιρο παρουσιάζει ιδιαίτερες δυσκολίες εξαιτίας της πολυπλοκότητάς του και της σχεδόν αδιάκοπης ροής της μπάλας κατά τη διάρκεια του παιχνιδιού. Σε αντίθεση με το μπασκετ ή το μπέιζμπολ, στο ποδόσφαιρο δεν είναι εύκολο να οριστούν κατάλληλα μέτρα που να περιγράφουν με σαφήνεια την αγωνιστική συμπεριφορά ενός παίκτη (Duch et al., 2010).

Τα Σχήματα 1 και 2 αποτελούν οπτικοποίηση των γεγονότων που σημειώθηκαν από κάθε ομάδα κατά τη διάρκεια του αγώνα της FC Barcelona ενάντια στην Real Madrid (Spanish League, 22nd March 2015). Κάθε σημείο δείχνει την τοποθεσία του γηπέδου όπου έλαβε χώρα κάποια χαρακτηριστική κίνηση όπως προσπέραση, βολή, τάκλιν, ντρίμπλα, φάουλ, κ.α. Από τις εικόνες αυτές μπορεί να γίνει αντιληπτή η πολυπλοκότητα του αθλήματος καθώς και η πλούσια πληροφορία η οποία μπορεί να εξορυχθεί απο ένα παιχνίδι. (Luca Pappalardo et al., 2017).



Σχήμα 1 : Κινήσεις της Barcelona στον αγώνα ενάντια της Real Madrid (Spanish League, 22nd March 2015).

Πηγή : Luca Pappalardo et al., 2017.



Σχήμα 2 : Κινήσεις της Real Madrid στον αγώνα ενάντια της Balcelona (Spanish League, 22nd March 2015).

.Πηγή : Luca Pappalardo et al., 2017.

Το πρόβλημα της αξιολόγησης επιδόσεων (performance) ποδοσφαιριστών προσελκύει το ενδιαφέρον πολλών εταιρειών, καθώς και της επιστημονικής κοινότητας, λόγω της διαθεσιμότητας τεράστιου όγκου δεδομένων που απαθανατίζουν τα γεγονότα (πάσες, σουτς κλπ.) που συμβαίνουν κατά τη διάρκεια ενός αγώνα. Δυστυχώς δεν υπάρχει κάποιο εδραιωμένο και ευρέως αποδεχούμενο μέτρο για να υπολογίζεται η ποιότητα της επίδοσης των παικτών σε όλες της τις όψεις. (Pappalardo et al., 2019)

Οι κατατάξεις (Rankings) των παικτών και η αξιολόγησή τους με μεθόδους ανάλυσης δεδομένων εξελίσσονται σε ολοένα και πιο κρίσιμα θέματα για την βιομηχανία του ποδοσφαίρου. Από τη μια πλευρά, πολλές αθλητικές εταιρείες, τηλεοπτικοί εκφωνητές και ιστοσελίδες όπως Opta, WhoScored.com και Sky, καθώς και πληθώρα από διαδικτυακές πλατφόρμες για fantasy football και ηλεκτρονικά παιχνίδια, χρησιμοποιούν εκτεταμένα στατιστικά ποδοσφαίρου για να συγκρίνουν τις επιδόσεις των παικτών. Σκοπός τους είναι να να δεσμεύσουν τους οπαδούς και να αυξήσουν το φανατισμό τους, κάνοντας κριτικές αναλύσεις, παρέχοντας γνώση (insights) και scoring μοτίβα (patterns). Από την άλλη πλευρά οι προπονητές και οι μάνατζερς των ομάδων ενδιαφέρονται για εργαλεία τα οποία να έχουν τη δυνατότητα να υποστηρίξουν την ανάλυση στρατηγικής (tactical analysis) και να ελέγχουν την ποιότητα των παικτών τους σε μενονωμένους

αγώνες ή και σε ολόκληρες σεζόνς. Επιπλέον, οι κυνηγοί ταλέντων αναζητούν εργαλεία που καθοδηγούνται από τα δεδομένα (data-driven) για να βελτιώσουν την ανάκτηση ταλαντούχων παικτών με επιθυμητά χαρακτηριστικά, βασιζόμενοι σε κριτήρια αξιολόγησης που λαμβάνουν υπόψιν την πολυπλοκότητα και την πολυδιάστατη φύση της ποδοσφαιρικής επίδοσης. Η διαλογή ταλέντων μέσα από ολόκληρο τον ποδοσφαιρικό κόσμο είναι ανέφικτη και χρονοβόρα για τον άνθρωπο. Οι βαθμολογίες (scores) των επιδόσεων όμως που προκύπτουν από μεθόδους αναλυτικής δεδομένων μπορούν να βοηθήσουν στην επιλογή μικρών υποσυνόλων των καλύτερων παικτών, οι οποίοι παρουσιάζουν κάποιους συγκεκριμένους περιορισμούς ή ένα συγκεκριμένο μοτίβο επιδόσεων. Με αυτόν τον τρόπο επιτρέπουν στους κυνηγούς ταλέντων και στους ομίλους να αναλύσουν ένα μεγαλύτερο σύνολο παικτών εξοικονομώντας χρόνο και χρήματα, και παράλληλα να διευρύνουν τις διαδικασίες διαλογής προσφέροντας ευκαιρίες καριέρας σε ταλαντούχους παίκτες. (Pappalardo et al., 2019)

Το πρόβλημα της αξιολόγησης και βαθμολογικής κατάταξης παικτών με μεθόδους ανάλυσης δεδομένων ενδιαφέρει όπως προαναφέρθηκε και την επιστημονική κοινότητα, εξαιτίας της τεράστιας ροής δεδομένων που γεννάται από (ημι-)αυτόματες τεχνολογίες αισθητήρων, όπως τα αποκαλούμενα soccer-logs, τα οποία περιγράφουν λεπτομερώς όλα τα χωροχρονικά γεγονότα που σημειώνονται από τους παίκτες κατά τη διάρκεια ενός αγώνα (όπως πάσες,ντρίμπλες, τάκλιν κλπ.). (Pappalardo et al., 2019)

Κατάταξη παικτών είναι ο ορισμός μιας σχέσης διάταξης μεταξύ των παικτών με βάση κάποιο μέτρο της επίδοσής τους σε μια σειρά από αγώνες. Η επίδοση με τη σειρά της μπορεί να μετρηθεί υπολογίζοντας τις αξιολογήσεις (ratings) κάθε παίκτη, βάσει δεδομένων, και ποσοτικοποιώντας έτσι την ποιότητα επίδοσής του σε ένα συγκεκριμένο παιχνίδι. Στη συνέχεια οι αξιολογήσεις συναθροίζονται για μια σειρά από παιχνίδια. Η διαδικασία αυτή είναι πολύπλοκη καθώς δεν υπάρχει ένας αντικειμενικός και κοινός ορισμός για την ποιότητα επίδοσης, καθώς πρόκειται για μια εγγενώς πολυδιάστατη έννοια. (Pappalardo et al., 2019)

Μέχρι σήμερα έχουν προταθεί αρκετοί αλγόριθμοι για την αξιολόγησης και την κατάταξη παικτών, αλλά οι περισσότεροι μειονεκτούν σε τρία βασικά σημεία. Πρώτον, οι προσεγγίσεις τους είναι μονοδιάστατες, και τα μέτρα αξιολόγησης εστιάζουν σε μία όψη της επίδοσης, για παράδειγμα κυρίως στα σουτς ή στις πάσες, με αποτέλεσμα να μην εκμεταλλεύονται σημαντικό μέρος της πληροφορίας που προσφέρουν τα soccer-logs. Αντιθέτως οι κυνηγοί ταλέντων χρειάζονται μέτρα αξιολόγησης που συνδυάζουν πολλές πτυχές της επίδοσης, από αμυντικές δεξιότητες μέχρι κατοχή της μπάλας και επιθετικές ικανότητες. Δεύτερον, οι υπάρχουσες προσεγγίσεις αξιολογούν αθλητικές επιδόσεις χωρίς να λαμβάνουν υπόψιν την ιδιαιτερότητα της θέσης που έχει κάθε παίκτης στο γήπεδο (πλάγιος αμυντικός, κεντρικός επιθετικός, κλπ.), και έτσι συγκρίνουν

παίκτες που εμπλέκονται σε διαφορετικά καθήκοντα. Τρίτον, ελλείψει ενός καλού συνόλου δεδομένων, οι περισσότερες προσεγγίσεις εξάγουν συμπεράσματα βασισμένα σε απλοϊκά μέτρα επίδοσης όπως για παράδειγμα στην εμπορική αξία (market value) ή στον αριθμό των γκολς. (Pappalardo et al., 2019)

Συνεπώς, υπάρχει ανάγκη από ένα συμπεριληπτικό πλαίσιο, ικανό να να παράγει μια πιο συνολική αξιολόγηση της αθλητικής επίδοσης, από σύνολα δεδομένων κατασκευασμένα από ειδικούς. (Pappalardo et al., 2019)

1.2. Συστήματα αξιολόγησης αθλητικών επιδόσεων. (Sports Rating systems)

Στην παράγραφο αυτή γίνεται μια περιήγηση σε συστήματα αξιολόγησης (Rating systems) παικτών/ομάδων στον ευρύτερο αθλητικό χώρο, με αύξουσα χρονολογική σειρά.

Ένα από τα πιο διάσημα συστήματα αξιολόγησης, το Elo Rating system, δημιουγήθηκε το 1959 από τον Arpad Emrick Elo, και στη συνέχεια υιοθετήθηκε το 1970 από την παγκόσμια σκακιστική ομοσπονδία (FIDE). Το σύστημα αυτό κατασκευάστηκε για τις ανάγκες αξιολόγησης επιδόσεων παικτών σκακιού, συνεπώς εφαρμόζεται ικανοποιητικά σε παιχνίδια 2 αντιπάλων. Πλέον, χρησιμοποιείται σαν βάση για μοντέλα πρόβλεψης και αξιολόγησης επιδόσεων και στον ευρύτερο αθλητικό χώρο. Σε ένα παιχνίδι δύο αντιπάλων, το σύστημα αξιολόγησης Elo υπολογίζει το βαθμό κατάταξης (ranking point) κάθε παίκτη μέσω ενός επαναληπτικού τύπου, κάθε φορά που ολοκληρώνεται ένας αγώνας. Ο επαναληπτικός τύπος ανανεώνει τους βαθμούς κατάταξης λαμβάνοντας υπόψιν 3 παράγοντες, οι οποίοι είναι ο προηγούμενος βαθμός, το πραγματικό αποτέλεσμα του αγώνα, και το αναμενόμενο αποτέλεσμα του αγώνα (Jan Lasek et al., 2013). Με βάση τη μέθοδο του Elo, όταν δύο παίκτες ανταγωνίζονται, αυτός με την υψηλότερη αξιολόγηση αναμένεται να έχει περισσότερες νίκες σε σχέση με αυτόν που έχει την χαμηλότερη. Επιπλέον, όσο μεγαλύτερη είναι η διαφορά μεταξύ των βαθμολογιών τους, τόσο πιθανότερο είναι να νικήσει ο παίκτης που έχει καταχθεί υψηλότερα (M. E. Glickman, 1995).

Το σύστημα που κατασκεύασε ο Arpad Emrick Elo έχει έκτοτε λάβει διάφορες προεκτάσεις και γενικεύσεις. Παρακάτω συνοψίζονται κάποιες από αυτές.

Ο Mark E. Glickman (1995) υποστηρίζει ότι το Elo Rating System έχει ένα σημαντικό μειονέκτημα, την έλλειψη αξιοπιστίας των παικτών. Ένας παίκτης που επιστρέφει στον αγωνιστικό χώρο έπειτα από κάποια χρόνια δεν πρέπει να αξιολογείται στην ίδια βάση που αξιολογείται ένας παίκτης ο οποίος αγωνίζεται κάθε εβδομάδα. Προτείνει μια μπεϋζιανή γενίκευση του συστήματος του Elo, το σύστημα αξιολόγησης Glicko, όπου λαμβάνεται υπόψιν ένα μέτρο ακρίβειας των

αξιολογήσεων, το rating deviation (RD) ή, σε στατιστική ορολογία, μια τυπική απόκλιση, η οποία μετρά την αβεβαιότητα που υπάρχει ως προς την αγωνιστική ικανότητα ενός παίκτη. Μεγάλες τιμές του μέτρου RD αντιστοιχούν σε αναξιόπιστες αξιολήσεις, καθώς υποδεικνύει ότι ο ο παίκτης δεν αγωνίζεται συχνά ή έχει λάβει μέρος σε πολύ λίγα παιχνίδια. Αντίθετα, χαμηλές τιμές του μέτρου RD υποδηλώνουν ότι ο παίκτης αγωνίζεται συχνά και οδηγούν σε πιο αξιόπιστες αξιολογήσεις. Το 2000 ο Mark E. Glickman αναπτύσσει μια βελτίωση του Glicko, το Glicko-2, όπου υπεισέρχεται ένα επιπλέον μέτρο, το rating volatility σ, το οποίο μετρά το βαθμό της αναμενόμενης διακύμανσης στην αξιολόγηση ενός παίκτη. Το μέτρο μεταβλητότητας σ είναι υψηλό όταν ο παίκτης έχει ακανόνιστες και απρόβλεπτες επιδόσεις, και χαμηλό όταν οι επιδόσεις του είναι σταθερές και συνεπείς.

Οι ερευνητές της Microsoft Ralf Herbrich et al. (2006) παρουσιάζουν το πλαίσιο TrueSkill ως μια γενίκευση του Elo Rating System, το οποίο δύναται να επεκταθεί σε παιχνίδια με περισσότερους από δύο παίκτες ,και κατασκευάστηκε κυρίως για online παιχνίδια. Το σύστημα έχει τη δυνατότητα να αξιολογεί τις ατομικές δεξιότητες των παικτών μιας ομάδας και να καταγράφει την αβεβαιότητα που ενδέχεται να υπάρχει για τις δυνατότητές τους μέσα από μια Μπεϋζιανή προσέγγιση.

Η φόρμουλα EloRatings.net, αποτελεί μια τροποποίηση της παραδοσιακής φόρμουλας του Elo Rating ώστε να εξυπηρετεί περισσότερο ως σύστημα αξιολόγησης στο ποδόσφαιρο, λαμβάνοντας υπόψιν κάποιες επιπλέον παραμέτρους. Αυτές είναι ένα ζύγισμα με βάση το είδος του αγώνα και τη σημαντικότητά του (φιλικός, Ευρωπαικός, Word Cup, κλπ), το πλεονέκτημα της έδρας και η διαφορά στον αριθμό των goals στο τελικό αποτέλεσμα

Μια ενδιαφέρουσα προσέγγιση που χτίστηκε πάνω στο Elo Rating System και χρησιμοποιεί τεχνικές μηχανικής μάθησης προτάθηκε απο τον Yannis Sismanis (2010). Το μοντέλο του, Elo++ , κέρδισε τον διαγωνισμό που διεξήγαγε το Kaggle “Chess Ratings: Elo vs the rest of the world” (Kaggle.com, 2010). Το Elo++ αποτελεί μια προέκταση της κλασσικής μεθόδου του Arpad Emrick Elo. Η βασική ιδέα του Yannis Sismanis ήταν ότι ένα σύστημα αξιολόγησης επιδόσεων πρέπει να βασίζεται περισσότερο στις αξιολογήσεις παικτών που έχουν λάβει μέρος σε πολλούς και πρόσφατους αγώνες, παρά σε αξιολογήσεις παικτών που συμμετείχαν σε λίγους και παλιούς αγώνες. Μια ακόμη ιδέα του ήταν ότι η δύναμη του παίκτη προς αξιολόγηση, δεν πρέπει να απέχει πολύ από τη δύναμη του αντιπάλου. Όμοια με άλλες προσεγγίσεις που έχουν γίνει στο παρελθόν, ο Elo++ προσθέτει στο βαθμό αξιολόγησης (rating) του παίκτη με τα λευκά πιόνια μια παράμετρο γ η οποία υποδηλώνει το πλεονέκτημα της πρώτης κίνησης. Η δεύτερη προσθήκη της μεθόδου, είναι στο κομμάτι του σφάλματος της λογιστικής παλινδρόμησης που εφαρμόζει για την πρόβλεψη του αποτελέσματος ενός αγώνα.

Η συνάρτηση του συνολικού σφάλματος λαμβάνει υπόψιν τη διαφορά μεταξύ των εκτιμημένων και των πραγματικών αποτελεσμάτων, τον χρονικό ορίζοντα στον οποίο συντελέστηκαν τα παιχνίδια, το πλήθος των αγώνων που συμμετείχε κάθε παίκτης, και τις αξιολογήσεις των αντιπάλων. Στο σημείο αυτό γίνεται μια κανονικοποίηση του σφάλματος με την τεχνική L2 regularization, για την αποφυγή του overfitting. Η L2 regularization προσθέτει στο συνολικό σφάλμα του μοντέλου (total loss) έναν επιπλέον όρο, ο οποιος στην περίπτωση του Elo++ είναι ο $\lambda^*\sum_{i=1}^{n}(r_i - a_i)^2$ και κατά την ελαχιστοποίηση του σφάλματος εμποδίζει τον μηδενισμό του. Επιπλέον ο όρος αυτός «τιμωρεί» τις αξιολογήσεις r_i του παίκτη i που είναι δυσανάλογα μακριά από τις μέσες αξιολογήσεις a_i των αντιπάλων του. Οι αξιολογήσεις ri υπολογίζονται ελαχιστοποιώντας τη συνάρτηση σφάλματος μέσω του αλγορίθμου Stochastic gradient descent (J. C. Spall, 2003) ο οποίος μετά από έναν αριθμό επαναλήψεων συγκλίνει στο τοπικό ελάχιστο της συνάρτησης σφάλματος.

Οι Anthony C. Constantinou et al. (2013) προτείνουν τα Pi ratings, ένα δυναμικό σύστημα αξιολόγησης ποδοσφαιρικών -και άλλων- ομάδων, που βασίζεται στη σχετική απόκλιση που υπάρχει στα σκορ των αγώνων μεταξύ δύο ομάδων. Τα Pi ratings υπερισχύνουν αρκετών από τις προτεινόμενες παραλλαγές του Elo ratings για το ποδόσφαιρο και αποδίδουν καλά σαν στοιχηματική στρατηγική. Το σύστημα αυτό λαμβάνει υπόψιν τρεις βασικές υποθέσεις: το πλεονέκτημα της έδρας, το γεγονός ότι τα πιο πρόσφατα αποτελέσματα έχουν μεγαλύτερη βαρύτητα από τα πιο παλιά για την εκτίμηση των ικανοτήτων μιας ομάδας, και το γεγονός ότι μια νίκη είναι πιο σημαντική από την απλή μείωση της διαφοράς στο σκορ. Τα Pi ratings υπολογίζουν την αναμενόμενη διαφορά των goals της ομάδας Υ έναντι σε έναν μέσο (average) αντίπαλο, στη βάση των τριών παραπάνω υποθέσεων. Ανανεώνονται αθροιστικά, και οι αποκλίσεις μεταξύ της εκτιμημένης και της παρατηρηθείσας διαφοράς στα goals καθορίζουν εαν το rating θα αυξηθεί ή θα μειωθεί. Το rating μιας ομάδας θα αυξηθεί εαν το σκορ υποδεικνύει υψηλότερη επίδοση από αυτήν που εκτίνησαν τα Pi ratings.

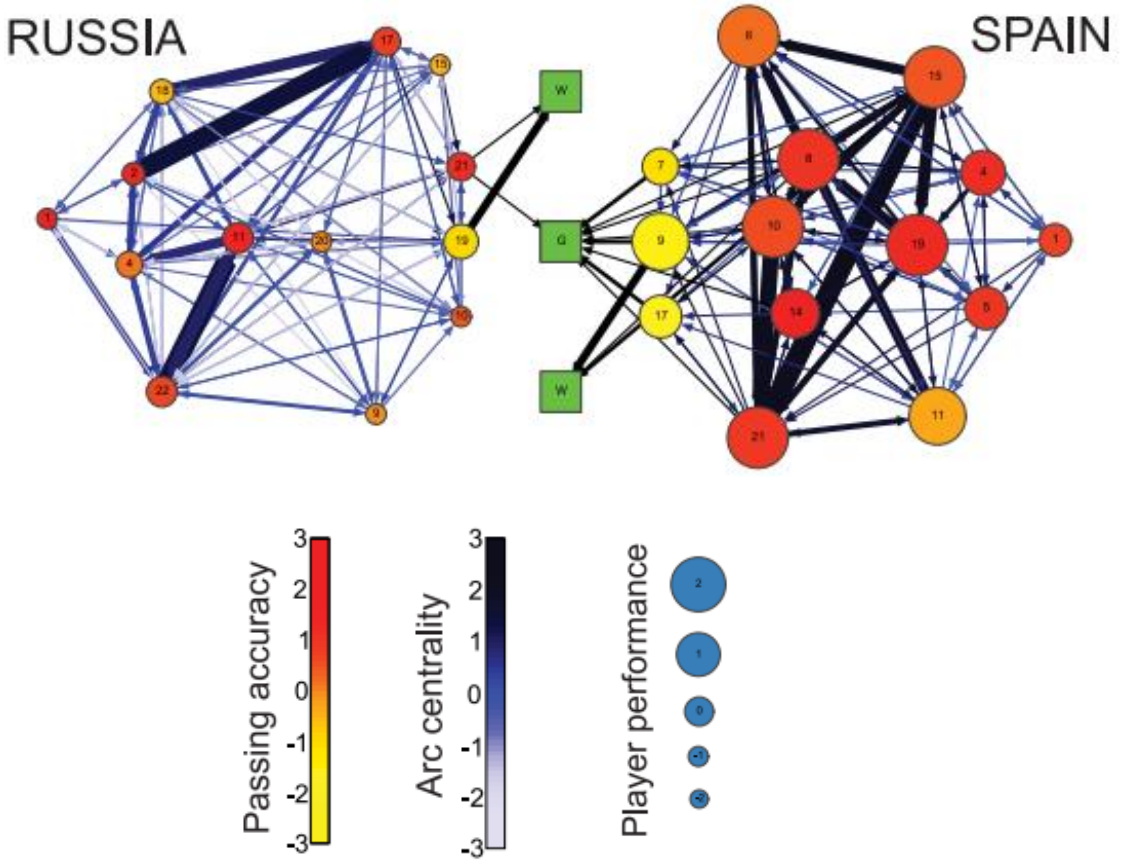
Πολλές από τις προσπάθειες που έχουν σημειωθεί για την βελτίωση των συστημάτων Elo και TrueSkill βαθμολογούν ομάδες ή παίκτες με βάση παρελθοντικές νίκες/ήττες και την εκτιμώμενη δύναμη του αντιπάλου. Ωστόσο, ο αγώνας είναι μια πολυδιάστατη έννοια, η οποία εμπεριέχει το στοιχείο του χρόνου, του χώρου και διάφορες ποιοτικές συνιστώσες της ατομικής και της συλλογικής επίδοσης των παικτών. Τα περισσότερα από τα προαναφερθέντα συστήματα αξιολόγησης δεν εστιάζουν σε αυτές τις πλευρές, και δεν εκμεταλλεύονται το εύρος της πληροφορίας που μπορεί να εξαχθεί από έναν αγώνα, με αποτέλεσμα να καθίστανται ανίκανα να παράξουν ένα σαφή χαρακτηρισμό και να αξιολογήσουν επαρκώς τη συνεισφορά των παικτών στον αγώνα (Pappalardo et al., 2019).

1.3. Data-driven συστήματα αξιολόγησης αθλητικών επιδόσεων στο ποδόσφαιρο.

Στην παράγραφο αυτή γίνεται μια περιήγηση σε ορισμένα συστήματα αξιολόγησης και κατάταξης επιδόσεων παικτών/ομάδων ποδοσφαίρου, τα οποία θεμελιώνονται σε μεθόδους ανάλυσης δεδομένων.

Μια από τις πρώτες προσπάθειες σε αυτό το πλαίσιο ήταν αυτή των Duch et al. (2010), οι οποίοι εμπνεύστηκαν από μεθόδους ανάλυσης κοινωνικών δικτύων ώστε να αποτυπώσουν τη συνεισφορά ενός παίκτη σε ένα match. Κατασκεύασαν μια δομή δικτύων η οποία απεικονίζει τη ροή της μπάλας μεταξύ των παικτών μιας ομάδας, το λεγόμενο δίκυο ροής (flow network). Οι κόμβοι συμβολίζουν τους παίκτες και τα τόξα είναι ζυγισμένα σύμφωνα με τον αριθμό των επιτυχών μεταβιβάσεων της μπάλας μεταξύ δύο παικτών. Προσθέτουν επίσης ένα κόμβο ο οποίος συμβολίζει τις επιτυχείς βολές (shots to goal), και έναν κόμβο για τις άστοχες βολές (shots wide). Οι παίκτες συνδέονται με αυτούς τους δύο κόμβους μέσω τόξων ζυγισμένων ανάλογα με το πλήθος των βολών τους. Για να περιγράψουν την ικανότητα ενός παίκτη να φτάσει τη μπάλα στο τέρμα του αντιπάλου χρησιμοποιούν την ακρίβεια της μεταβίβασης (passing accuracy), η οποία μετρά τις επιτυχείς πάσες του στους συμπαίκτες, και την ακρίβεια βολής (shooting accuracy), η οποία μετρά τις επιτυχημένες βολές προς το τέρμα. Συνδυάζοντας το δίκτυο ροής με την ακρίβεια μεταβίβασης και βολής, μετρούν την επίδοση ενός παίκτη μέσω του μέτρου Flow Centrality (FC). Το FC μετρά πόσες φορές ένας παίκτης ενεπλάκη σε μονοπάτια τα οποία κατέληξαν σε κόμβους-βολές. Μέσω της μετρικής τους οι Duch et al. κατατάσσουν τους παίκτες της UEFA European Championship 2008 και καταφέρνουν να κατατάξουν σωστά 8 από τους κορυφαίους 20 παίκτες της λίστας που ανακοινώθηκε μετά το πέρας του διαγωνισμού. Ωστόσο, οι ίδιοι κάνουν την παραδοχή ότι το σύστημά τους είναι πιο αποτελεσματικό για επιθετικούς και μέσους παίκτες.

Στο Σχήμα 3 απεικονίζεται ένα παράδειγμα δικτύου ροής του FC από έναν αγώνα Ρωσίας και Ισπανίας. Η τοποθεσία του κάθε κόμβου ορίζεται από την θέση του παίκτη το γήπεδο, και ο αριθμός του κόμβου αντιστοιχεί στον αριθμό που αναγράφεται στην μπλούζα του παίκτη. Οι κόμβοι είναι χρωματισμένοι σύφωνα με την ακρίβεια μεταβίβασης του παίκτη, και το μέγεθός τους διαφοροποιείται ανάλογα με την επίδοση του παίκτη. Το πλάτος των τόξων αυξάνεται εκθετικά σε σχέση με τον αριθμό των επιτυχών μεταβιβάσεων της μπάλας σε συμπαίκτες, ενώ το χρώμα τους υποδεικνύει το κανονικοποιημένο FC του τόξου.



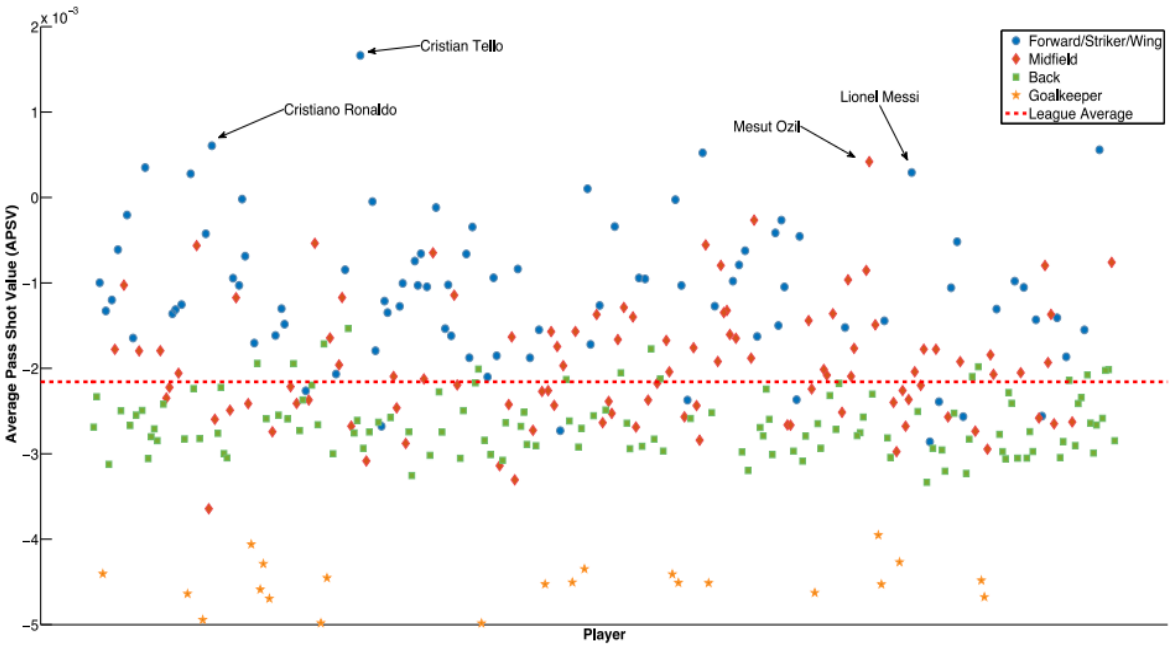
Σχήμα 3 : Παράδειγμα από εφαρμογή της Flow Centrality σε έναν αγώνα Ρωσίας και Ισπανίας.

Πηγή : Duch et al., 2010.

Στηριζόμενοι στην ιδέα ότι η πάσα είναι μια από τις πιο στρατηγικές κινήσεις στο ποδόσφαιρο οι Brooks et al. (2016) αναπτύσσουν το μέτρο Pass Shot Value (PSV), σε μια προσπάθεια να ποσοτικοποιήσουν και να μετρήσουν τη σημαντικότητα της πάσας, με βάση το αν αυτή καταλήγει σε γκολ ή όχι. Κατακερματίζουν το παιχνίδι στο επίπεδο των κατοχών της μπάλας (possessions), δηλαδή της χρονικής περιόδου που μια ομάδα κατέχει τη μπάλα χωρίς διακοπές ή χάνει τη μπάλα από την αντίπαλη ομάδα. Για το σκοπό αυτό χωρίζουν την περιοχή του γηπέδου σε 18 ζώνες και αναπαριστούν την πάσα ως ένα διάνυσμα 360 χαρακτηριστικών. Ο μέσος όρος των διανυσμάτων όλων των ολοκληρωμένων μεταβιβάσεων της μπάλας αποτελεί το διάνυσμα μιας κατοχής, και περιγράφει χωρικά την προέλευση και τον προορισμό της πάνω στις 18 ζώνες. Στη συνέχεια κανουν κατηγοριοποίηση των κατοχών της μπάλας ,μέσω του αλγορίθμου επιβλεπόμενης μάθησης Support Vector Machine (SVM), για να προβλέψουν εαν μια κατοχή καταλήγει σε σουτ ή όχι. Τα βάρη των χαρακτηριστικών που προκύπτουν από το τελικό μοντέλο είναι διανύσματα της

μορφής $w = [w^o, w^d, w^{od}]$, όπου w^o, w^d και w^{od} είναι βάρη χαρακτηριστικών που αντιστοιχούν σε διαφορετικές ζώνες για την προέλευση, τον προορισμό και τα ζεύγη προέλευσης-προορισμού της μπάλας. Τα βάρη αυτά υποδεικνύουν από ποιές υποπεριοχές του γηπέδου είναι πιο πιθανό να προκύψουν ευκαιρίες για σουτς. Έτσι, οι Brooks et al. μετρούν τη σημαντικότητα μιας πάσας με προέλευση τη ζώνη i και προορισμό τη j μέσω του αθροίσματος των βαρών, $PSV(i,j) = w_i^o + w_j^d + w_{ij}^{od}$. Τέλος, χρησιμοποιούν soccer-logs για να κατατάξουν τους παίκτες της La Liga 2012-13 που σημείωσαν περισσότερες από 200 πάσες, με βάση το μέσο PSV τους (APSV). Οπως σημειώνουν στην εργασία τους, το PSV μεροληπτεί υπέρ των επιθετικών παικτών. Οι Pappalardo et al. (2019) προσθέτουν ότι το PSV είναι ένα μέτρο το οποίο βασίζεται μόνο στις πάσες, παραλείποντας όλες τις υπόλοιπες κινήσεις που παρατηρούνται κατά τη διάρκεια ενός αγώνα, με αποτέλεσμα να κρίνεται αναξιόπιστο.

Στο Σχήμα 4 φαίνεται πως το μοντέλο κατατάσσει τους παίκτες με βάση την μέση τάση τους να εκτελούν πάσες που εξελίσσονται σε σουτ. Πιο συγκεκριμένα, τα σημεία στο γράφημα συμβολίζουν το APSV για κάθε παίκτη, ενώ τα χρώματα και τα σχήματα των σημείων διαφοροποιούνται ανάλογα τη θέση του παίκτη. Η κόκκινη διακεκομμένη γραμμή αντιπροσωπεύει το APSV του συνόλου των παικτών.

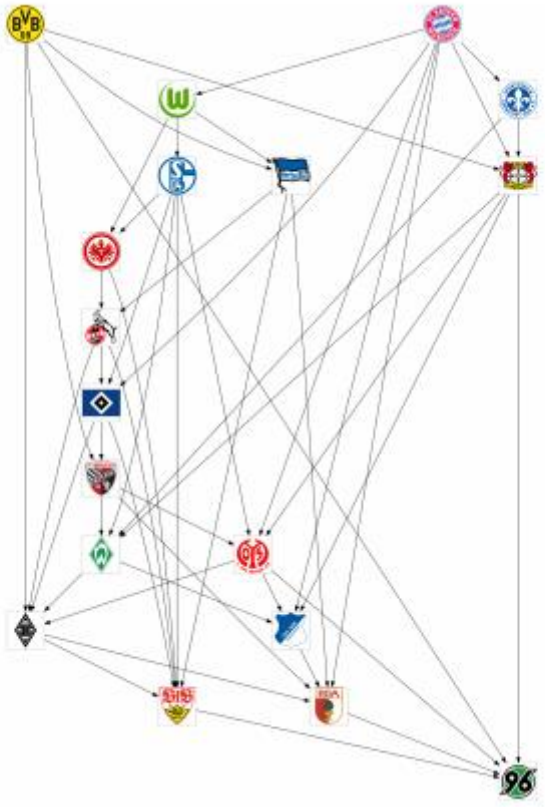


Σχήμα 4 : Παράδειγμα από εφαρμογή του PSV σε παίκτες της La Liga 2012-13.

Πηγή : Brooks et al., 2016.

Η κατάταξη (ranking) των ομάδων στο ποδόσφαιρο γίνεται συνήθως με βάση τους βαθμολογικούς πίνακες (league tables) που παρέχονται από τους διοργανωτές των πρωταθλημάτων. Αυτοί οι πίνακες είναι σχεδιασμένοι με τέτοιον τρόπο ώστε σε κάθε ομάδα να αντιστοιχίζεται ένας μοναδικός βαθμός. Ωστόσο στο χώρο της αναλυτικής αθλητικών δεδομένων το πρόβλημα της κατάταξης των ομάδων μπορεί να μετατραπεί σε πρόβλημα κατηγοριοποίησης των ομάδων σε μικρές κλάσεις, οι οποίες να έχουν μεταξύ τους ουσιαστικές διαφορές ως προς την ποιότητα της ποδοσφαιρικής συμπεριφοράς.

Οι Neumann et al. (2018) βασίστηκαν σε μια γενικευμένη εκδοχή του αλγορίθμου Agony για την ανίχνευση ιεραρχιών στα κοινωνικά δίκτυα των Gupte et al. (2011), προκειμένου να κατατάξουν με πιο «φυσικό» τρόπο τις ομάδες από τέσσερα μεγάλα Ευρωπαϊκά πρωταθλήματα, δημιουργώντας ουσιώδεις ιεραρχίες που αποτυπώνουν τα διαφορετικά επίπεδα ποδοσφαιρικής συμπεριφοράς. Παράλληλα διερευνούν απέναντι σε ποιούς αντιπάλους μια ομάδα μπορεί να αποδίδει καλύτερα ή όχι, μια πτυχή που συνήθως δεν λαμβάνεται υπόψιν στις κατατάξεις των ομάδων. Χρησιμοποιούν δύο ειδών γραφήματα δικτύων (networks), το κατευθυνόμενο (directed) και το μη κατευθυνόμενο (undirected). Στην πρώτη περίπτωση μια ακμή κατευθύνεται καθοδικά από μια κορυφή που συμβολίζει την ομάδα Α προς μια κορυφή που συμβολίζει την ομάδα Β εάν η Α νίκησε την Β. Το μήκος της ακμής είναι ζυγισμένο σύμφωνα με το σκορ του αγώνα της Α με την Β. Από αυτά τα γραφήματα ο αλγόριθμος κατασκευάζει ιεραρχίες από επίπεδα που συμβολίζουν την ποιότητα των ομάδων πηγαίνοντας από το υψηλότερο στο χαμηλότερο. Οι κορυφές-ομάδες οι οποίες δεν έχουν καθόλου εισερχόμενες ακμές, δηλαδή έχουν μόνο νίκες, κατατάσσονται στο επίπεδο 1. Συνεχίζοντας, οι κορυφές-ομάδες οι οποίες έχουν εισερχόμενα βέλη μόνο από ομάδες που έχουν ήδη καταταχθεί πηγαίνουν στο επίπεδο 2, κ.ο.κ. Στην δεύτερη περίπτωση μια ακμή ενώνει δύο κορυφές εάν οι αντίστοιχες ομάδες σημείωσαν ισοπαλία. Οι ισόπαλες ομάδες κατατάσσονται στο ίδιο επίπεδο. Ιδιαίτερη για την περίπτωση των κατευθυνόμενων δικτύων, οι Neumann et al. προτείνουν να ορίζεται ένας μέγιστος αριθμός αγωνιστικών ημερών (matchdays) για τον οποίο οι κατατάξεις δεν είναι αντικρουόμενες. Αντικρουόμενες κατατάξεις προκύπτουν στην περίπτωση που η ομάδα Α νίκησε την Β, η Β την C, και η C την Α. Δυστυχώς η συνθήκη αυτή δεν εξασφαλίζεται πάντα, με αποτέλεσμα να παράγονται κατατάξεις από πρωταθλήματα και τουρνουά οι οποίες είναι αυτοαναιρούμενες (Skinner et al., 2009). Στο Σχήμα 5 απεικονίζεται ένα κατευθυνόμενο ακυκλικό γράφημα δικτύου που έχει προκύψει από την εφαρμογή του Agony στην εργασία των Neumann et al. για τις 7 πρώτες αγωνιστικές ημέρες του Γερμανικού πρωταθλήματος της σεζόν 2015/16.



Σχήμα 5 : Οι πρώτες 7 αγωνιστικές ημέρες από την Γερμανική Bundesliga, σεζόν 2015/16.

Πηγή : Neumann et al., 2018.

Το πεδίο της αναλυτικής ποδοσφαιρικών δεδομένων (soccer analytics) έχει κερδίσει αναγνώριση σχετικά πρόσφατα, και πάσχει από την έλλειψη διαθέσιμων και οικονομικά προσιτών δεδομένων. Κατά συνέπεια, η συλλογή κατάλληλων γνωρισμάτων (features) παικτών/ομάδων για την κατασκευή διανυσμάτων (feature vectors) που τροφοδοτούν τους αλγορίθμους μηχανικής μάθησης για τους σκοπούς της εκάστοτε ανάλυσης είναι δύσκολη, ακριβή και χρονοβόρα. Οι Robert Müller et al. (2019) προτείνουν ως λύση στο παραπάνω πρόβλημα τον αλγόριθμο STEVE (Soccer Team Vectors), ο οποίος βρίσκει αυτόματα λανθάνοντα διανύσματα γνωρισμάτων ή αλλιώς αναπαραστάσεις (representations), χρησιμοποιώντας μονάχα την ελεύθερη πληροφορία που υπάρχει σχετικά με απολέσματα αγώνων. Ο STEVE είναι σχεδιασμένος έτσι ώστε να χρησιμοποιεί ελάχιστο μέρος από τη διαθέσιμη πληροφορία, δηλαδή μόνο τις νίκες/ήττες/ισοπαλίες μιας ομάδας ,τη σεζόν που αγωνίστηκε, και τους αντιπάλους που είχε. Παράλληλα, ανιχνεύει όμοιες και ανόμοιες ομάδες μέσω υπολογισμών ευκλείδειων αποστάσεων μεταξύ των αναπαραστάσεων, και κατατάσσει ομάδες με βάση τη δύναμη της καθεμίας. Η έννοια της ομοιότητας μεταξύ δύο ομάδων ορίζεται στην εργασία των Robert Müller et al. από τέσσερις βασικές υποθέσεις. Η πιο σημαντική είναι η υπόθεση ότι δύο ομάδες είναι όμοιες, εάν νικούν συχνά τους ίδιους αντιπάλους. Στην εργασία τους γίνεται επίσης αναφορά στην ανεπάρκεια που χαρακτηρίζει τις κατατάξεις που

προκύπτουν από τους βαθμολογικούς πίνακες, λόγω του γεγονότος ότι αφορούν μόνο μια σεζόν, καθώς και ομάδες ενός μόνο πρωταθλήματος. Κατά συνέπεια, μια κατάταξη ακανόνιστα επιλεγμένων ομάδων απο διαφορετικά πρωταθλήματα καθίσταται πολύπλοκη. Ο STEVE μετριάζει αυτό το πρόβλημα, και δοθείσης μια λίστας από ομάδες, δημιουργεί την προσομοίωση ενός τουρνουά όπου κάθε ομάδα αγωνίζεται ενάντια σε όλες τις υπόλοιπες. Η λίστα αυτή στη συνέχεια ανακατατάσσεται σύμφωνα με τον αριθμό από τις νίκες που είχε η κάθε ομάδα. Για να υπολογιστεί το αποτέλεσμα του αγώνα (νίκη/ήττα) μεταξύ δύο ομάδων α και β, υπολογίζεται το τετράγωνο της ευκλείδειας νόρμας:

$$\alpha = \| \Phi_\alpha - \Psi_\beta \|^2 \quad \text{και} \quad \beta = \| \Phi_\beta - \Psi_\alpha \|^2,$$

όπου Φi οι αναπαράστασεις για τις νίκες (winner representations) της ομάδας i , και Ψi οι αναπαραστάσεις για τις για τις ήττες της (loser representations). Εάν α < β τότε η αναπαράσταση με τις ήττες της β είναι πιο κοντά στην αναπαράσταση με τις νίκες της α, επομένως η ομάδα α είναι πιο δυνατή. Ανάλογα ερμηνεύεται και η περίπτωση όπου β < α. Στο Σχήμα 6 παρουσιάζονται δύο κατατάξεις από την εφαρμογή του STEVE. Κάθε λίστα αποτελείται από 12 διαφορετικές Ευρωπαϊκές ομάδες διαφόρων επιπέδων. Οι ομάδες σε κάθε διάταξη κατατάσσονται με φθίνουσα σειρά από αριστερά προς δεξιά. Πράγματι, οι ισχυρότερες ομάδες όπως οι Real Madrid, FC Bayern Munich, FC Barcelona και AS Roma είναι τοποθετημένες στην κορυφή, ενώ οι λιγότερο επιτυχημένες όπως οι FC Toulouse, Cardiff City, Fortuna Düsseldorf και Parma Calcio βρίσκονται στην ουρά κάθε λίστας.

[1] Real Madrid, FC Bayern Munich, Inter Milano, Liverpool FC, Borussia Dortmund, Ajax Amsterdam, FC Porto, Club Brugge KV, Werder Bremen, 1.FC Nuremberg, FC Toulouse, Cardiff City

[2] FC Barcelona, AS Roma, Atltico Madrid, Paris SG, Tottenham, PSV Eindhoven, Arsenal London, SL Benfica, Espanyol Barcelona, VFB Stuttgart, Fortuna Dusseldorf, Parma Calcio

Σχήμα 6 : Κατατάξεις που παρήχθησαν από τον STEVE.

Πηγή : Robert Müller et al., 2019.

Η επιτυχία μιας ποδοσφαιρικής ομάδας εξαρτάται σε μεγάλο βαθμό από τις ατομικές δεξιότητες των μονάδων που την απαρτίζουν. Ωστόσο, οι Nsolo et al. (2018) υπογραμμίζουν ότι πέραν της διαφορετικότητας που υπάρχει μεταξύ των παικτών, διαφοροποιούνται και οι θέσεις/ρόλοι τους εντός της ομάδας. Ακόμη, το στυλ του παιχνιδιού μπορεί να αλλάζει για μια ομάδα ανάλογα με το επίπεδο του αγώνα που έχει να αντιμετωπίσει κάθε φορά. Θεωρούν ότι τα στοιχεία αυτά πρέπει να συνυπολογίζονται σε μια διαδικασία αξιολόγησης επιδόσεων. Στην εργασία τους δεν προτείνουν κάποια δική τους μέθοδο αξιολόγησης αλλά χρησιμοποιούν αλγορίθμους μηχανικής μάθησης προκειμένου να βρουν ποιά χαρακτηριστικά/δεξιότητες δίνουν την καλύτερη πρόβλεψη της επιτυχίας των παικτών λαμβάνοντας υπόψιν τις θέσεις τους και έχοντας ως μέτρο σύγκρισης τις αξιολογήσεις του WhoScored.com. Αρχικά εξετάζουν σε 5 κορυφαία Ευρωπαϊκά πρωταθλήματα (Ιταλία, Ισπανία, Αγγλία, Γερμανία, Γαλλία), ποια γνωρίσματα είναι σημαντικά για τους αμυντικούς, τους μέσους, τους φόργουορντς και τους τερματοφύλακες. Στην ανάλυσή τους εφαρμόζουν και συγκρίνουν διαφορετικούς αλγορίθμους κατηγοριοποίησης προκειμένου να προβλέψουν εάν ένας παίκτης ανήκει στο 10% (κορυφαίοι παίκτες), 25% (καλοί παίκτες) ή 50% (μέτριοι παίκτες) των rankings του WhoScored.com. Το τελικό μοντέλο είναι πιο ακριβές για συγκεκριμένες θέσεις παικτών, ειδικότερα για τους φόργουοντς, και για συγκεκριμένα πρωταθλήματα, όπως το Αγγλικό Premier League.

1.4. Αξιολόγηση μέσω της εμπορικής αξίας (Market value).

Το ποδόσφαιρο, πέραν από ένα δημοφιλές άθλημα, αποτελεί ταυτόχρονα και μια μεγάλη επιχείρηση. Από μια διοικητική σκοπιά, οι πιο σημαντικές αποφάσεις που λαμβάνονται από τους μάνατζερς των ομάδων σχετίζονται με διαπραγματεύσεις για μεταγραφές παικτών. Κατά συνέπεια, θέματα όπως η αποτίμηση της εμπορικής αξίας (market value) ενός παίκτη, ο μισθός, τα έξοδα μεταγραφής, η δημοτικότητά του και η σχέση αυτών με την επίδοσή του προσελκύουν το ενδιαφέρον των ειδικών του ποδοσφαίρου και των ερευνητών. Η εμπορική αξία είναι μια εκτίμηση του κόστους μεταγραφής, ή πιο απλουστευμένα, η πιο πιθανή τιμή που εκτιμάται ότι κοστίζει ένας παίκτης στο χώρο της αγοράς. Αυτές οι τιμές παραδοσιακά εκτιμούνταν από τους ειδικούς του ποδοσφαίρου, όμως τα τελευταία χρόνια το έργο αυτό το έχει μετατεθεί και στον πληθοπορισμό. Παρόλο που οι ερευνητές έχουν βρει υψηλές συσχετίσεις μεταξύ της εμπορικής αξίας που εκτιμά ο πληθοπορισμός και του πραγματικού κόστους μεταγραφής, οι εκτιμησεις που παράγονται από το ετερογενές πλήθος δεν φημίζονται για την διαφάνειά τους, ανανεώνονται σπάνια επειδή απαιτούν τη συμμετοχή πολλών χρηστών, και δεν υπάρχουν αντίγραφα της διαδικασίας που ακολουθείται. Στο πρόβλημα αυτό, ο κλάδος της ανάλυσης δεδομένων δύναται να

προσφέρει μια συμπληρωματική του πληθοπορισμού προσέγγιση για την εκτίμηση του μεγέθους της εμπορικής αξίας (Oliver Müller et al., 2017).

Οι Stanojevic et al. (2016) στην εργασία τους χρησιμοποιούν soccer-logs για να εξετάσουν τη σχέση μεταξύ της αθλητικής επίδοσης ενός παίκτη και της εμπορικής του αξίας, όπως αυτή εκτιμάται από τα πλήθη. Εντοπίζουν μεγάλη απόκλιση μεταξύ των εκτιμημένων και των πραγματικών τιμών της εμπορικής αξίας, γεγονός που αποδίδουν στην έλλειψη σημαντικού μέρους πληροφορίας που σχετίζεται με την τάση του παίκτη προς τραυματισμούς, την ικανότητά του για διαφήμιση και αυτοπροβολή κ.α.

Οι Oliver Müller et al. (2017) ακολουθώντας μια παρόμοια με την παραπάνω προσέγγιση, χρησιμοποιούν soccer-logs και πολυμεταβλητή ανάλυση παλινδρόμησης για να εκτιμήσουν την εμπορική αξία παικτών από πέντε κορυφαία Ευρωπαϊκά πρωταθλήματα σε μια περίοδο που αποτελείται από έξι σεζόνς. Τα αποτελέσματά τους δείχνουν ότι οι εκτιμήσεις με μεθόδους ανάλυσης δεδομένων υπερνικούν αρκετούς από τους πρακτικούς περιορισμούς που έχει ο πληθοπορισμός, και είναι πιο ακριβή.

Ένας άλλο πεδίο έρευνας εστιάζει στην αναγνώριση μοτίβων επαγγελματικής ποδοσφαιρικής καριέρας. Το κομμάτι αυτό βοηθά τους προπονητές να επιτύχουν τις καλύτερες δυνατές μεταγραφές και συνθέσεις ομάδων. Επίσης, βοηθά τους παίκτες να ξεδιπλώσουν τις δεξιότητές τους, και να θέσουν στόχους για την βελτίωση της καριέρας τους. Οι Acs et al. (2021) στην εργασία τους αναζητούν τα χαρακτηριστικά των παικτών που επηρεάζουν περισσότερο την αξιολόγηση και την αποτίμησή τους. Χρησιμοποιούν τμήματα από χρονοσειρές διάρκειας τριών ετών για 4204 παίκτες, και για κάθε τμήμα δημιουργούν συστάδες (clusters) με βάση τις μεταβολές που είχε η εμπορική αξία μέσα στο εξεταζόμενο χρονικό διάστημα. Στη συνέχεια αναζητούν μοτίβα από καριέρες όπου η εμπορική αξία είχε σημειώσει εξαιρετική αύξηση, και εντοπίζοντας ομοιότητες καταφέρνουν να βρουν τα πέντε σημαντικότερα χαρακτηριστικά που πρέπει να εξασκήσει ένας παίκτης, και πώς, ώστε να επιτύχει μια τέτοια καριέρα.

2. Αναπαράσταση αθλημάτων με χρήση χωροχρονικών δεδομένων.

Τα χωροχρονικά δεδομένα (spatio-temporal data) είναι μια σειρά δειγμάτων που περιέχουν χρονοσήμανση και τοποθεσία κάποιων φαινομένων. Στον τομέα των αθλητικών ομάδων σκιαγραφούνται συνήθως δύο είδη τέτοιων δεδομένων: οι τροχιές αντικειμένων (object trajectories), που καταγράφουν την κίνηση παικτών ή της μπάλας, και τα αρχεία καγραφής γεγονότων (event logs) που καταγράφουν τη θέση και το χρόνο όπου συνέβησαν κάποια γεγονότα σε έναν αγώνα, όπως οι πάσες, τα φάουλς, τα σουτς κ.ο.κ. Αυτά τα δύο είδη δεδομένων διευκολύνουν την ανάλυση του παιχνιδιού, και παρόλο που είναι συμπληρωματικά μεταξύ τους με την έννοια ότι περιγράφουν διαφορετικές όψεις του παιχνιδιού, αν χρησιμοποιηθούν συνδυαστικά παρέχουν μια πιο πλούσια επεξήγηση. Για παράδειγμα, ο χωρικός σχηματισμός με τον οποίο διατάσσεται μια ομάδα στο γήπεδο θα είναι ξεκάθαρος σε ένα σετ από δεδομένα τροχιών παικτών. Όμως, αυτή η συγκεκριμένη διάταξη στο χώρο πιθανότατα εξαρτάται από το αν η ομάδα αυτή έχει την κατοχή της μπάλας, το οποίο μπορεί να εξακριβωθεί από τα event logs. (Gudmundsson et al., 2017)

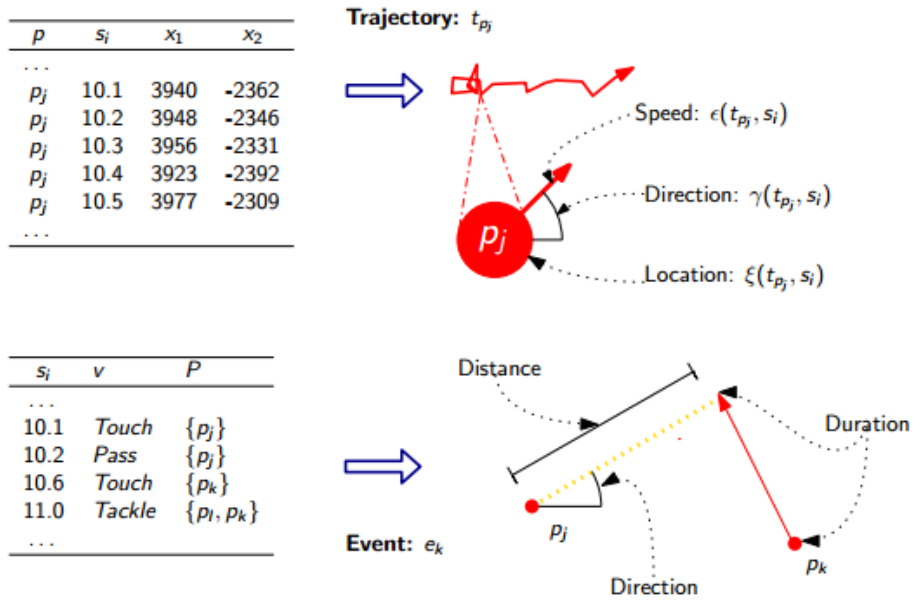
2.1. Τροχιές αντικειμένων

Οι κινήσεις των παικτών ή της μπάλας στην περιοχή του γηπέδου αποτυπώνονται σαν μια χρονοσφραγισμένη σειρά (timestamped sequence) από χωρικά σημεία στο επίπεδο, όπως το επάνω μέρος του Σχήματος 7. Στο σχήμα παρουσιάζονται κάποια δεδομένα κίνησης και ο τρόπος που αυτά αναπαρίστανται γεωμετρικά σαν τροχιές. Κάθε τροχιά είναι μια σειρά σημείων του χώρου, τα οποία μπορούν να χρησιμοποιηθούν για να προεκβληθεί η γεωμετρία ενός παίκτη με ένα συγκεκριμένο χρονικό βήμα (time-step). Παρομοίως, η γεωμετρία των γεγονότων, για παράδειγμα της προσπέρασης, μπορεί να υπολογιστεί από τις τροχιές των εμπλεκόμενων παικτών σε αυτήν. Οι τροχιές καταγράφονται με τη χρήση οπτικών- ή συσκευών-συστημάτων εντοπισμού/παρακολούθησης και επεξεργασίας. Τα οπτικά συστήματα εντοπισμού χρησιμοποιούν σταθερές κάμερες για την καταγραφή της κίνησης του παίκτη, και οι εικόνες στη συνέχεια υπόκεινται σε επεξεργασία για να υπολογιστούν οι τροχιές. Υπάρχουν αρκετοί εμπορικοί προμηθευτές που παρέχουν υπηρεσίες παρακολούθησης σε επαγγελματικές ομάδες και πρωταθλήματα. Οι συσκευές παρακολούθησης από την άλλη μεριά είναι τοποθετημένες στα ρούχα των παικτών ή ενσωματωμένες στη μπάλα, και εντοπίζουν τη θέση τους. Αυτά τα συστήματα μπορεί να βασίζονται σε GPS ή RF ID τεχνολογίες. Η διαθεσιμότητα των χωροχρονικών δεδομένων για την έρευνα ποικίλλει. Μερικά πρωταθλήματα καταγράφουν δεδομένα από όλα τα παιχνίδια, ενώ κάποιες ομάδες μπορεί να καταγράφουν γεγονότα μόνο στο στάδιό τους. Τα δεδομένα από πρωταθλήματα είναι πιο

ογκώδη και προσφέρονται για πειράματα που ελέγχουν εξωτερικούς παράγοντες όπως οι μεταβλητές καιρού, οι τραυματισμοί των παικτών κ.α. (Gudmundsson et al., 2017)

2.2. Event logs

Τα Event logs είναι μια ακολουθία από σημαντικά γεγονότα τα οποία συμβαίνουν στη διάρκεια ενός αγώνα, και μπορούν να διακριθούν σε δύο ευρείες κατηγορίες. Η πρώτη κατηγορία είναι τα γεγονότα τα οποία συνδέονται άμεσα με τον παίκτη (player events), όπως για παράδειγμα οι πάσες ή τα σουτς· η δεύτερη κατηγορία είναι τεχνικής φύσεως γεγονότα (technical events), όπως τα φάουλς, τα time-outs, η αρχή και το τέλος της περιόδου κ.α. Ένα παράδειγμα event logs δεδομένων με μια αναπαράστασή τους απεικονίζεται στην κάτω περίπτωση του Σχήματος 7. Αυτού του τύπου τα δεδομένα ενδέχεται εν μέρει να πηγάζουν από τις τροχιές των εμπλεκόμενων παικτών, αλλά μπορεί επίσης και να καταγράφονται απευθείας από την ανάλυση video, μια τακτική που ακολουθεί για παράδειγμα η Opta Sports Ltd. Αυτή η προσέγγιση εξυπηρετεί συνήθως σε περιπτώσεις αθλημάτων όπου υπάρχουν πρακτικές δυσκολίες στη σύλληψη των τροχιών των παικτών, όπως στο ράγκμπυ. Τα event logs διαφέρουν ποιοτικά από τις τροχιές των παικτών, λόγω του ότι δεν είναι πυκνά, αφού τα δείγματα καταγράφονται μόνον όταν συμβαίνει ένα γεγονός. Εν αντιθέσει με τις τροχιές όμως, θα μπορούσαν να χαρακτηριστούν πιο πλούσια σημασιολογικά, καθώς περιέχουν λεπτομέρειες για το είδος του γεγονότος, τους παίκτες που αναμείχθηκαν σε αυτό, κ.α. (Gudmundsson et al., 2017)

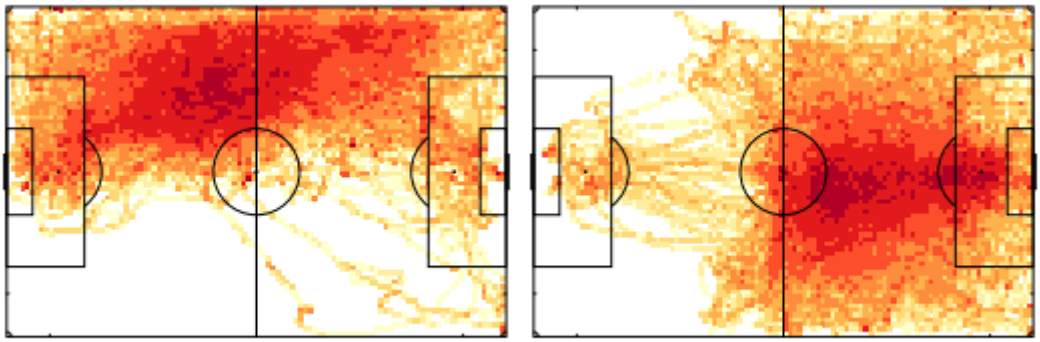


Σχήμα 7 : Παράδειγμα trajectory και event log δεδομένων και γεωμετρική τους αναπαράσταση.

Πηγή : Gudmundsson et al., 2017.

2.3. Συνήθης δυσκολία.

Οι τροχίες αντικειμένων και τα event logs είναι αναπαραστάσεις γεγονότων σε χαμηλό επίπεδο, και ενδέχεται να είναι απαιτητικές στην ανάλυση. Ένας τρόπος αντιμετώπισης αυτού του ζητήματος είναι ο διαχωρισμός του αγωνιστικού χώρου σε περιοχές, και η ανάθεση των τοποθεσιών των δεδομένων σε αυτές. Η συχνότητα ή η ένταση των γεγονότων που συμβαίνουν σε αυτούς τους υποχώρους μπορεί να αποτελέσει πηγή πληροφορίας για τους παίκτες και τους ρόλους τους, καθώς και για την πρόοδο ενός παιχνιδιού. Μια τεχνική διαχωρισμού του γηπέδου σε υποπεριοχές είναι για παράδειγμα οι χάρτες έντασης (intensity maps). Στο Σχήμα 8 απεικονίζεται ένα παράδειγμα τέτοιου χάρτη για το ποδόσφαιρο, που υποδεικνύει την περιοχή του γηπέδου την οποία απασχολεί ο ρόλος των Left-back παικτών στην περίπτωση (a), και Striker στην (b). Οι τροχιές των αθλητών έχουν προσανατολισμό επίθεσης από αριστερά προς τα δεξιά (Gudmundsson et al., 2017).



Σχήμα 8 : Χάρτες έντασης των θέσεων-ρόλων (a) Left-back, και (b) Striker.

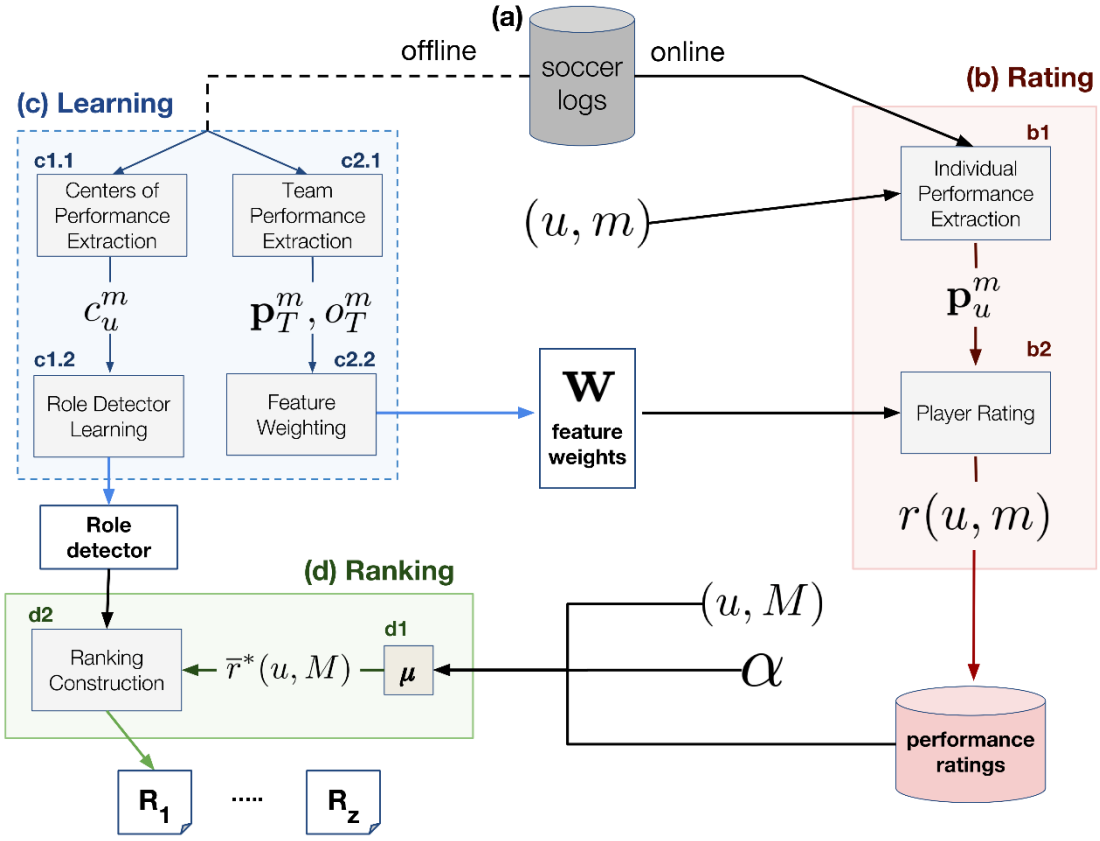
Πηγή : Gudmundsson et al., 2017.

3. Το πλαίσιο PlayeRank

Στο σημείο αυτό παρουσιάζονται αναλυτικά τα βήματα που ακολουθεί το πλαίσιο (framework) PlayRank ώστε να γίνει κατανοητή η ροή και οι ιδιαιτερότητές του.

Στο Σχήμα 9 απεικονίζεται το γενικό πλαίσιο PlayeRank, το οποίο δέχεται και λειτουργεί με soccer-logs δεδομένα. Σε αυτού του τύπου τα δεδομένα, που είναι ουσιαστικά event logs, ένα match μπορεί να κωδικοποιείται ώς ένα διάνυσμα < Id, type, position, timestamp >, όπου Id είναι ο κωδικός αναγνώρισης του παίκτη, το type αφορά το είδος του γεγονότος (προσπέραση, σούτ, τάκλιν, κλπ.) ,ενώ οι συνιστώσες position και timestamp δίνουν την χωροχρονική πληροφορία του συγκεκριμένου γεγονότος μέσα στο πεδίο του γηπέδου. Η μέθοδος κάνει την υπόθεση ότι τα soccer-logs δεδομένα είναι αποθηκευμένα σε μια βάση δεδομένων η οποία ανανεώνεται έπειτα από κάθε νέο αγώνα.

Είναι εμφανές από το Σχήμα 9 ότι το πλαίσιο απαρτίζεται από 3 βασικές φάσεις: αυτήν της μάθησης (Learning), της αξιολόγησης (Rating) και της κατάταξης (Ranking).



Σχήμα 9 : Σχεδιάγραμμα του πλαισίου PlayeRank.

Πηγή : https://github.com/mesosbrodleto/playerank/blob/master/README.md.

Η φάση της μάθησης είναι μια "offline" διαδικασία η οποία παράγει την απαραίτητη πληροφορία που θα χρησιμοποιηθεί στις ακόλουθες δύο φάσεις. Στο αρχικό αυτό στάδιο εκτελούνται δύο βήματα. Γίνεται ζύγισμα των χαρακτηριστικών (Feature weighting) και ανίχνευση θέσης/ρόλου του παίκτη (Role detector).

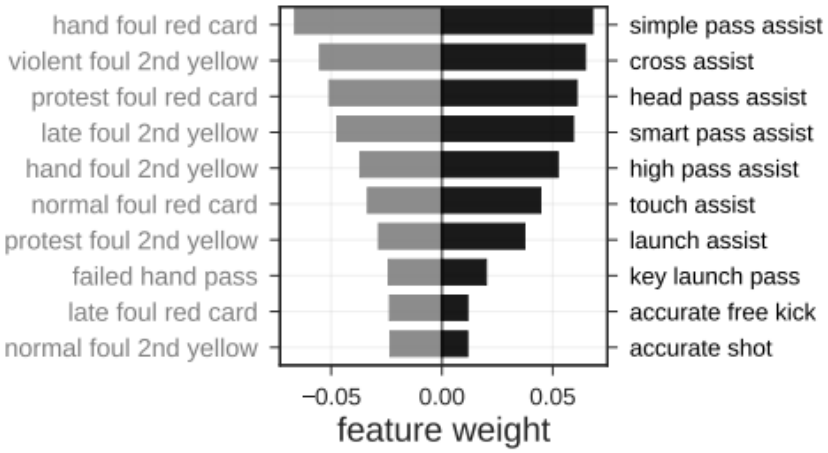
Για το ζύγισμα των χαρακτηριστικών η μέθοδος εξάγει το διάνυσμα της επίδοσης P_T^m της ομαδας T στο match m, και το αποτέλεσμα του αγώνα, O_T^m , όπου O_T^m=1 υποδηλώνει νίκη, και O_T^m=0 υποδηλώνει μή νίκη (ισοπολία, ήττα). Το διάνυσμα $P_T^m = [x_1^{(T)}, \ldots, x_n^{(T)}]$ της επιδοσης της ομάδας T στο παιχνίδι m υπολογίζεται αθροίζοντας τα επιμέρους διανύσματα επίδοσης όλων των παικτών της, U_T^m, σε αυτό το match.

$$P_T^m[i] = \sum_{u \in U_T^m} P_u^m[i] ,$$

όπου $P_u^m = [x_1, \ldots, x_n]$ είναι το διάνυσμα επίδοσης του παίκτη u στο m match. Η συνιστώσα xi περιγράφει ένα συγκεκριμένο χαρακτηριστικό της συμπεριφοράς του παίκτη u στον αγώνα. Συνήθως οι συνιστώσες x_i μετρούν πλήθος σημαντικών γεγονότων, όπως για παράδειγμα τον αριθμό των κόκκινων καρτών που δόθηκαν στον παίκτη u.

Στη συνέχεια γίνεται μια κατηγοριοποίηση μεταξύ του performance P_T^m της ομάδας T και του αποτελέσματος του αγώνα m, O_T^m , μέσω ενός γραμμικού κατηγοριοποιητή, όπως είναι για παράδειγμα ο Support Vector Machine (SVM). Όπως σημειώνουν οι Pappalardo et al. (2017), υπάρχει ισχυρή σχέση μεταξύ του performance μιας ομάδας, και του αποτελέσματος του αγώνα. Από αυτήν τη κατηγοριοποίηση παράγονται τα βάρη (weights) $w = [w_1, \ldots, w_n]$, τα οποία ποσοτικοποιούν την επίδραση των χαρακτηριστικών στο τελικό αποτέλεσμα του αγώνα, και τα οποία αργότερα θα χρησιμοποιηθούν στη φάση της αξιολόγησης.

Στο Σχήμα 10 παρουσιάζονται τα υψηλότερα (μαύρο) και τα χαμηλότερα (γκρι) βάρη χαρακτηριστικών που προέκυψαν από τον SVM. Τα χαρακτηριστικά που σχετίζονται με το assist φαίνεται να είναι τα πιο σημαντικά, ενώ οι κόκκινες/κίτρινες κάρτες έχουν τα χαμηλοτερα βάρη.



Σχήμα 10 : Παράδειγμα από βάρη w από τον αλγόριθμο SVM.

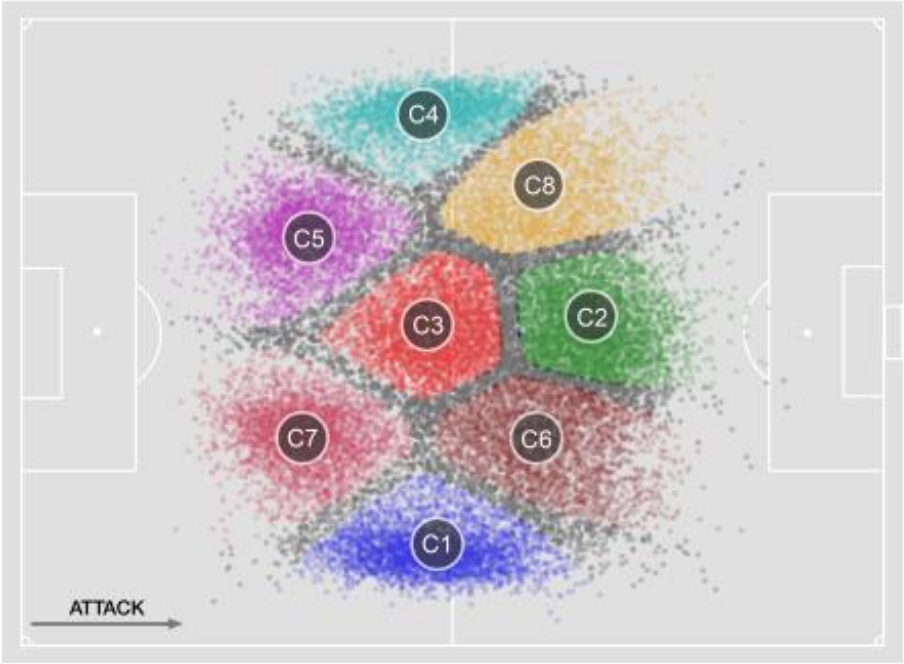
Πηγή : Pappalardo et al., 2019.

Στο κομμάτι της ανίχνευσης θέσης, αξίζει να σημειωθεί ότι οι αξιολογήσεις επιδόσεων έχουν νόημα υπό την προϋπόθεση ότι συγκρίνονται παίκτες που παίζουν σε παρόμοιες θέσεις. (S. Pettigrew, 2015), (Oliver Shulte and Zeyu Zhao, 2017). Επιπλέον, ένας παίκτης δεν διατηρείται στην ίδια θέση καθ'όλη τη διάρκεια ενός αγώνα, τα καθήκοντά του ενδέχεται να διαφοροποιούνται ανάλογα το σύστημα του παιχνιδιού που υιοθετείται. Για τους λόγους αυτούς σε αυτό το στάδιο εφαρμόζεται μια συσταδοποίηση με σκοπό την ανίχνευση της θέσης που κατέχει κάθε παίχτης σε ένα match ,με βάση τα soccer-logs. Πιο συγκεκριμένα, λαμβάνεται υπόψιν η μέση θέση κάθε παίκτη (average position). Αυτό απορρέει από το γεγονός ότι η θέση ενός παίκτη συνήθως ορίζεται σε σχέση με τις θέσεις των συμπαικτών του (A. Bialkowski et al., 2014). Έτσι, γίνεται χρήση του κέντρου της επίδοσης $c_u^m = (\bar{x}_u^m, \bar{y}_u^m)$ κάθε παίκτη u στο match m, όπου $\bar{x}_u^m$ *και* $\bar{y}_u^m$ είναι οι μέσες συντεταγμένες των γεγονότων που έλαβαν χώρα από τον u στο m, όπως αυτές εξήχθησαν από τα soccer-logs. Ακολουθεί μια συσταδοποίηση των κέντρων c_u^m μέσω του αλγορίθμου k-means (J. A. Hartigan and M. A. Wong, 1979), ώστε να ομαδοποιηθούν τα κέντρα c_u^m. Είναι σημαντικό να αναφερθεί ότι το PlayeRank προνοεί για την πιθανότητα ύπαρξης "υβριδικών" θέσεων, δηλαδή περιπτώσεων όπου ένα κέντρο επίδοσης μπορεί να εντοπίζεται ανάμεσα σε δύο ή περισσότερες συστάδες, και άρα να ταξινομείται σε περισσότερες από μια. Έτσι, δύναται να παραχθεί μια πιο "απαλή" συσταδοποίηση. Για κάθε κέντρο c_u^m στη συστάδα C_i, η μέθοδος υπολογίζει το k-silhouette Sk (c_u^m) αναφορικά με κάθε άλλη συστάδα C_k, (k≠i) ως εξής :

$$S_k\,(c_u^m) = \frac{\overline{d_k}\,(c_u^m) - \overline{d_\iota}\,(c_u^m)}{\max(\overline{d_\iota}\,(c_u^m)\,,\overline{d_k}\,(c_u^m))},$$

όπου $\overline{d_z}(c_u^m)$ η μέση απόσταση μεταξύ του c_u^m και όλων των υπόλοιπων σημείων μέσα στη συστάδα C_z. Ο αλγόριθμος αντιστοιχεί το κέντρο c_u^m σε κάθε συστάδα C_j για την οποία $Sj\ (c_u^m) \leq \delta_s$, όπου το δ_s είναι ένα κατώφλι που υποδεικνύει την ανοχή σε "υβριδικά" κέντρα. Εάν δεν υπάρχει τέτοιο j, το c_u^m ανατίθεται στη συστάδα C_i , η οποία έχει προκύψει από τον διαχωρισμό του k-means.

Στο Σχήμα 11 απεικονίζεται μια συσταδοποίηση των κέντρων επίδοσης σε 8 ομάδες. Κάθε χρώμα υποδηλώνει μια διαφορετική συστάδα, δηλαδή μία συγκεκριμένη θέση (role) παίκτη. Τα γκρι σημεία συμβολίζουν "υβριδικά" κέντρα επίδοσης.



Σχήμα 11 : Παράδειγμα συσταδοποίησης μέσω K-means.

Πηγή : Pappalardo et al., 2019.

Η φάση της αξιολόγησης είναι μια online διαδικασία η οποία τροφοδοτείται με την "offline" πληροφορία της προηγούμενης φάσης, κάθε φορά που ένα καινούριο match γίνεται διαθέσιμο στην βαση δεδομένων soccer-logs, και παράγει αξιολογήσεις για κάθε παίκτη στον συγκεκριμένο αγώνα. Αποτελείται από δύο βασικά βήματα: Εξαγωγή της ατομικής επίδοσης (individual performance extraction) και την αξιολόγηση του παίκτη (player rating).

Όσον αφορά την εξαγωγή της ατομικής επίδοσης κάθε παίκτη, αυτή μοντελοποιείται ως ένα διάνυσμα $P_u^m = [x_1,\ldots,x_n]$, του οποίου οι συνιστώσες, όπως έχει προαναφερθεί, περιγράφουν χαρακτηριστικά της αγωνιστικής

συμπεριφοράς του παίκτη u μέσα στο match m. Όπως σημειώνεται από τους Pappalardo et al. (2019), το PlayeRank είναι σχεδιασμένο με τέτοιον τρόπο ώστε να επιτρέπει στον χρήστη να δουλεύει με ποικίλα σύνολα χαρακτηριστικών.

Η αξιολόγηση ενός παίκτη u στο match m είναι ένα βαθμωτό μέγεθος που υπολογίζεται μεταξύ των τιμών των χαρακτηριστικών του παίκτη u στο match m, και των αντίστοιχων βαρών w των χαρακτηριστικών που έχουν παραχθεί από την φάση της μάθησης. Κάθε βάρος w μοντελοποιεί την σημαντικότητα των χαρακτηριστικών στην αξιολόγηση της ποιότητας της επίδοσης κάθε παίκτη. Πιο συγκεκριμένα, δοθέντος του πολυδιάστατου διανύσματος των χαρακτηριστικών P_u^m = [x_1,…,x_n] και των αντίστοιχων βαρών τους w = [w_1,…,w_n], η μέθοδος αξιολογεί την επίδοση ενός παίκτη u στο match m ως εξής:

$$r(u,m) = \frac{1}{R}\sum_{i=1}^{n} wi \times xi$$

Η ποσότητα r(u,m) καλείται αξιολόγηση επίδοσης (performance rating) του παίκτη u στο match m, όπου R είναι μια σταθερά κανονικοποίησης ,τέτοια ώστε r(u,m) ∈ [0,1].

Η παραπάνω σχέση προϋποθέτει την απουσία των goals σαν στοιχείο του συνόλου των χαρακτηριστικών. Ωστόσο, η μέθοδος μπορεί να προσαρμοστεί στην περίπτωση που ο χρήστης επιλέξει να θέσει τα goals σαν κύριο χαρακτηριστικό. Αυτό είναι δόκιμο κυρίως στις αξιολογήσεις επιθετικών παικτών (Pappalardo et al., 2019). Τότε, η αξιολόγηση της επίδοσης του παίκτη υπολογίζεται ως εξής:

$$r^*(u,m) = \alpha \times norm\_goals + (1 - \alpha) \times r(u,m) ,$$

Η ποσότητα r*(u,m) καλείται προσαρμοσμένη αξιολόγηση επίδοσης (adjusted performance rating) του παίκτη u στο match m, όπου norm_goals ο αριθμός των goals που επιτεύχθησαν από τον u στο m ,κανονικοποιημένο σε ένα εύρος [0,1], και α∈[0, 1] μια παράμετρος που δείχνει τη σημαντικότητα/βάρος των goals στη νέα αξιολόγηση. Προφανώς όταν α=0, r∗(u,m)=r(u,m) και όταν α=1, r*(u,m) = norm_goals.

Τελικά, η αξιολόγηση ενός παίκτη u σε μια σειρά από αγώνες M = (m_1,…,m_g) υπολογίζεται συναθροίζοντας τις επιμέρους αξιολογήσεις του u σε αυτούς τους αγώνες, μέσω μιας συνάρτησης συνάθροισης μ(r(u,m_1), . . . ,r(u,m_g)) που επιλέγει ο χρήστης. Οι Pappalardo et al. (2019) για το σκοπό αυτό χρησιμοποίησαν τη συνάρτηση του εκθετικού κινούμενου μέσου (EWMA). Συνεπώς, η ποιότητα της επίδοσης του u έπειτα από g αγώνες ισούται με:

$$\bar{r}(u, M) = \bar{r}(u,m_g) = \beta \times r(u,m_g) + (1 - \beta) \times \bar{r}(u,m_{g-1}) ,$$

όπου β είναι ένας παράγοντας εξομάλυνσης, ο οποίος καλείται συντελεστής βαρύτητας, και παίρνει τιμές στο διάστημα [0,1]. Το μέγεθος $\bar{r}$(u, M) καλείται

αξιολόγηση της επίδοσης του παίκτη u σε ένα σύνολο αγώνων M. Κατά τον ίδιο τρόπο υπολογίζεται η αντίστοιχη προσαρμοσμένη αξιολόγηση $\bar{r}*(u, M)$. Άρα, η ποιότητα της επίδοσης του u μετά από g παιχνίδια, υπολογίζεται σαν ο ζυγισμένος μέσος της αξιολόγησης $r(u,m_g)$ του τελευταίου αγώνα, και των προηγούμενων εξομαλυσμένων αξιολογήσεων $\bar{r}(u,m_{g-1})$ από τους προηγούμενους αγώνες. Η συνάρτηση EWMA επιτρέπει στον υπολογισμό της ποιότητας επίδοσης μεγαλύτερη βαρύτητα στους πιο πρόσφατους αγώνες και μικρότερη στους παλαιότερους.

Η φάση της κατάταξης είναι επίσης μια online διαδικασία κατά την οποία το PlayeRank κατασκευάζει ένα σετ $R_1,\ldots,R_z$ από κατατάξεις, με βάση τις θέσεις των παικτών (role-based rankings). Κάθε R_i αντιστοιχεί σε έναν από τους z ρόλους-θέσεις που έχουν ανιχνευθεί προηγουμένως στην φάση της μάθησης. Ένας παίκτης u κατατάσσεται στο R_i εάν τουλάχιστον στο x% των αγώνων του συνόλου M αυτός έχει εντοπιστεί στην θέση i. Το x είναι μια παράμετρος η οποία επιλέγεται από τον χρήστη και ανάλογα με την τιμή του, ένας παίκτης ενδέχεται να εμφανίζεται σε περισσότερες από μια κατατάξεις/σύνολα R_z σε διαφορετική σειρά, αφού αυτή εξαρτάται από το $\bar{r}(u,M)$. (Pappalardo et al., 2019.)

4. Πειραματικό μέρος

Έπειτα από τη θεωρητική περιήγηση στο θέμα της αξιολόγησης και βαθμολογικής κατάταξης ποδοσφαιριστών που έχει προηγηθεί, στο κεφάλαιο αυτό γίνεται παρουσίαση των δεδομένων που επιλέχθηκαν για την πραγμάτωση της εργασίας, περιγραφή της προετοιμασίας τους και της εφαρμογής του πλαισίου PlayeRank σε αυτά.

4.1. Περιγραφή των δεδομένων

Καθώς το πλαίσιο PlayeRank είναι προορισμένο να λειτουγεί με δεδομένα τύπου event logs ή αλλιώς soccer-logs, ήταν εύλογο η αναζήτηση των δεδομένων που επρόκειτο να χρησιμοποιηθούν να περιοριστεί σε τέτοιου είδους σύνολα. Τέτοιας φύσεως δεδομένα είναι αρκετά δυσεύρετα, αφού κατά κύριο λόγο συλλέγονται από εξειδικευμένες στον τομέα των sports analytics εταιρείες και αποτελούν ιδιοκτησία τους. Για την πραγματοποίηση της παρούσας εργασίας επιλέχθηκε μια ανοικτή συλλογή δεδομένων της εταιρείας Wyscout, την οποία παρουσιάζουν το 2019 οι Pappalardo et al. σε άρθρο τους που δημοσιεύθηκε στο επιστημονικό περιοδικό Nature. Η συλλογή αυτή είναι ένα μέρος του γιγαντιαίου συνόλου δεδομένων της Wyscout, το οποίο χρησιμοποιούν οι Pappalardo et al. (2019) στην σχετική εργασία τους για το πλαίσιο Playerank.

Το αρχικό σύνολο δεδομένων που χρησιμοποιούν οι Pappalardo et al. περιλαμβάνει 31,000,000 γεγονότα από προσεγγιστικά 20,000 αγώνες και 21,000 παίκτες, των 18 διακεκριμένων πρωταθλημάτων/διαγωνισμών: La Liga (Ισπανία), Premier League (Αγγλία), Serie A (Ιταλία), Bundesliga (Γερμανία), Ligue 1 (Γαλλία), Primeira Liga (Πορτογαλία), Super Lig (Τουρκία), Super League (Ελλάδα), Austrian Bundesliga (Αυστρία), Raiffeisen Super League (Ελβετία), Russian Football Championship (Ρωσία), Eredivisie (Ολλανδία), Superliga (Αργεντινή), Campeonato Brasileiro Série A (Βραζιλία), UEFA Champions League, UEFA Europa League, FIFA World Cup 2018 και UEFA Euro Cup 2016.

Το υποσύνολο που χρησιμοποιήθηκε στην παρούσα εργασία είναι αρκετά πιο περιορισμένο, και πιο συγκεκριμενα περιλαμβάνει όλα τα χωροχρονικά γεγονότα που καταγράφηκαν στους αγώνες της σεζόν 2017/2018, πέντε Ευρωπαϊκών πρωταθλημάτων πρώτης κατηγορίας: Ισπανικό, Ιταλικό, Αγγλικό, Γερμανικό και Γαλλικό, όπως επίσης και των World cup 2018 και European cup 2016. Η συλλογή αποτελείται από 1,941 αγώνες, 3,251,294 γεγονότα και 4,299 παίκτες. Υπάρχουν επτά επιμέρους υποσύνολα δεδομένων σε αρχεία της μορφής JSON (JavaScript Object Notation), τα αρχεία: coaches, competitions, events, matches, players, referees, και teams. Τα coaches, competitions και referees δεν

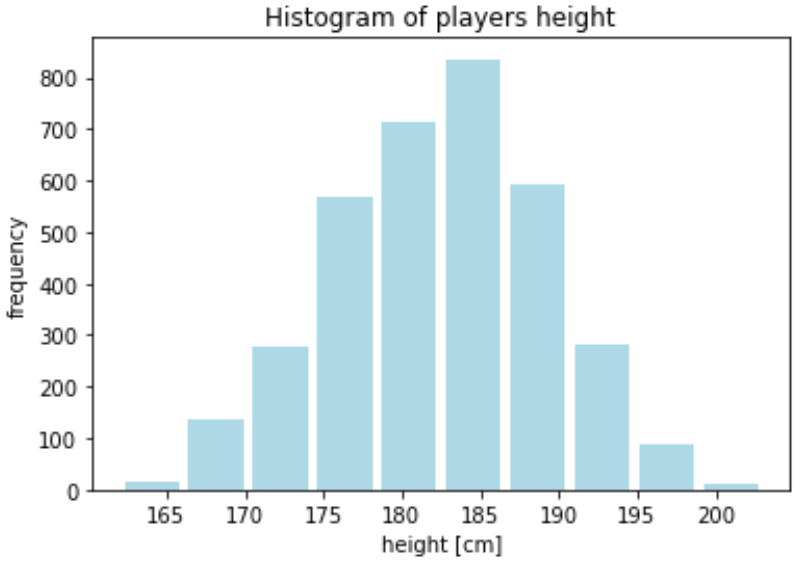
χρησιμοποιήθηκαν στην ανάλυση, επομένως είναι περιττό να επεκταθούμε σε αυτά.

Όσον αφορά τα σύνολα που χρησιμοποιήθηκαν, το players περιλαμβάνει περιγραφικά στοιχεία για όλους τους παίκτες που συμμετείχαν στους επτά διαγωνισμούς. Ένα παράδειγμα του περιεχομένου αυτού του αρχείου παρουσιάζεται στο Σχήμα 12. Περιγράφεται ο παίκτης Alfred John Momar N'Diaye ('shortName': "A. N'Diaye"), ο οποίος γεννήθηκε στην Γαλλία το 1990 ('birthArea', 'birthDate') και έχει σενεγαλέζικο διαβατήριο ('passportArea'). Ο αθλητής αυτός ανήκει προς το παρόν στην ομάδα 683 ('currentTeamId'), ζυγίζει 82 κιλά και έχει ύψος 187 εκατοστά ('weight', 'height'). Η βασική του θέση/ρόλος είναι μέσος ('role'), και το προτιμώμενο πόδι του είναι το δεξί ('foot').

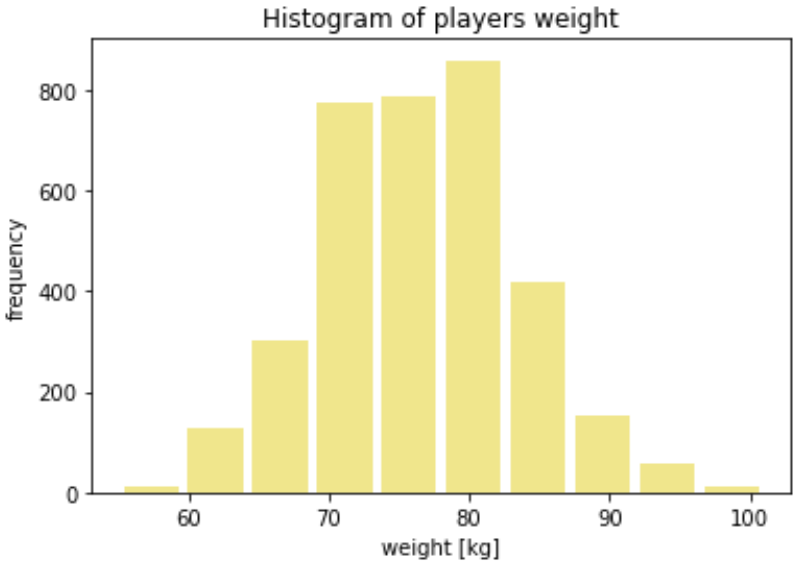

```
{'passportArea': {'name': 'Senegal',
  'id': '686',
  'alpha3code': 'SEN',
  'alpha2code': 'SN'},
 'weight': 82,
 'firstName': 'Alfred John Momar',
 'middleName': '',
 'lastName': "N'Diaye",
 'currentTeamId': 683,
 'birthDate': '1990-03-06',
 'height': 187,
 'role': {'code2': 'MD', 'code3': 'MID', 'name': 'Midfielder'},
 'birthArea': {'name': 'France',
  'id': '250',
  'alpha3code': 'FRA',
  'alpha2code': 'FR'},
 'wyId': 32793,
 'foot': 'right',
 'shortName': "A. N'Diaye",
 'currentNationalTeamId': 19314}
```


Σχήμα 12 : Απόσπασμα περιεχομένου από το αρχείο players.

Σύμφωνα με τα ιστογράμματα των Σχημάτων 13 και 14, η πλειοψηφία των παικτών έχει ύψος κυμαινόμενο κοντά στα 185 εκατοστά και βάρος κοντά στα 80 κιλά.



Σχήμα 13 : Ιστόγραμμα του ύψους των παικτών.



Σχήμα 14 : Ιστόγραμμα του βάρους των παικτών.

Τα σύνολο teams περιγράφει τις ομάδες τύπου clubs ή εθνικές ομάδες που συμμετέχουν στους επτά διαγωνισμούς. Στο Σχήμα 15 φαίνεται η δομή μιας ομάδας όπως αυτή σκιαγραφείται στο αρχείο teams. Πρόκειται για την Las Palmas, η οποία έχει έδρα την πόλη Las Palmas de Gran Canaria της Ισπανίας.

```
{'city': 'Las Palmas de Gran Canaria',
 'name': 'Las Palmas',
 'wyId': 714,
 'officialName': 'UD Las Palmas',
 'area': {'name': 'Spain',
  'id': '724',
  'alpha3code': 'ESP',
  'alpha2code': 'ES'},
 'type': 'club'}
```


Σχήμα 15 : Απόσπασμα περιεχομένου από το αρχείο teams.

Το σύνολο matches περιέχει πληροφορίες για όλους τους αγώνες των διαθέσιμων διαγωνισμών. Ένας αγώνας του αρχείου matches έχει τη μορφή που φαίνεται στο Σχήμα 16. Το συγκεκριμένο παράδειγμα είναι από το Ιταλικό πρωτάθλημα πρώτης κατηγορίας (‘competitionId’: 524) και πραγματοποιήθηκε στις 20 Μαϊου 2018 (‘date’ , ‘dateutc’), μεταξύ των ομάδων Napoli και Crotone, με τελικό σκορ 2-1 υπέρ της Napoli (‘label: Napoli - Crotone, 2 – 1’).


```
{'status': 'Played',
 'roundId': 4406278,
 'gameweek': 38,
 'teamsData': {'3197': {'scoreET': 0,
   'coachId': 287082,
   'side': 'away',
   'teamId': 3197,
   'score': 1,
   'scoreP': 0,
   'hasFormation': 1,
   'formation': {'bench': [{'playerId': 137048,
      'ownGoals': '0',
      'redCards': '0',
      'goals': 'null',
      'yellowCards': '0'},
     {'playerId': 21078,
      'ownGoals': '0',
      'redCards': '0',
      'goals': 'null',
```

```
      'goals': 'null',
      'yellowCards': '0'}],
    'substitutions': [{'playerIn': 122, 'playerOut': 99452, 'minute': 67},
     {'playerIn': 21171, 'playerOut': 21385, 'minute': 75},
     {'playerIn': 284315, 'playerOut': 40726, 'minute': 77}]},
   'scoreHT': 2}},
 'seasonId': 181248,
 'dateutc': '2018-05-20 16:00:00',
 'winner': 3187,
 'venue': '',
 'wyId': 2576332,
 'label': 'Napoli - Crotone, 2 - 1',
 'date': 'May 20, 2018 at 6:00:00 PM GMT+2',
 'referees': [{'refereeId': 377231, 'role': 'referee'},
  {'refereeId': 388407, 'role': 'firstAssistant'},
  {'refereeId': 393638, 'role': 'secondAssistant'},
  {'refereeId': 377271, 'role': 'fourthOfficial'}],
 'duration': 'Regular',
 'competitionId': 524}
```

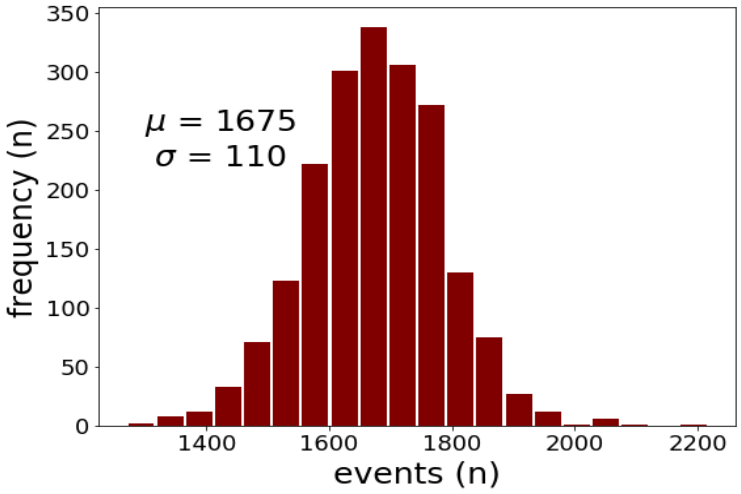

Σχήμα 16 (α), (β) : Αποσπάσματα αγώνα από το αρχείο matches.

Το τελευταίο και σημαντικότερο αρχείο, το events, περιέχει όλα τα γεγονότα που συνέβησαν ανά αγώνα των επτά διαγωνισμών. Ένα παράδειγμα γεγονότος παρουσιάζεται στο Σχήμα 17. Περιγράφει τις λεπτομέρειες της πάσας (‘eventName’: ‘pass’) με κωδικό 8 (‘eventId’: 8), και πιο συγκεκριμένα μιας απλής πάσας (‘subEventName’: ‘Simple pass’) η οποία παρήχθη από τον παίκτη 134383 (‘playerId’: 134383) της ομάδας 2444 (‘teamId’: 2444) στον αγώνα 2516739 (‘matchId’: 2516739), στο δευτερόλεπτο 875.49 και στο πρώτο μισό του αγώνα (‘eventSec’: 875.497474999999 , ‘matchPeriod’: ‘1H’). Επιπλέον, όπως μας πληροφορεί το στοιχείο ‘positions’, η εκκίνηση της πάσας έγινε στο σημείο του γηπέδου με συντεταγμένες (65,30) και ολοκληρώθηκε στο σημείο (86,37). Τέλος, σύμφωνα με τον κωδικό 1801 του ‘tags’, πρόκειται για μια έγκυρη (accurate) πάσα.

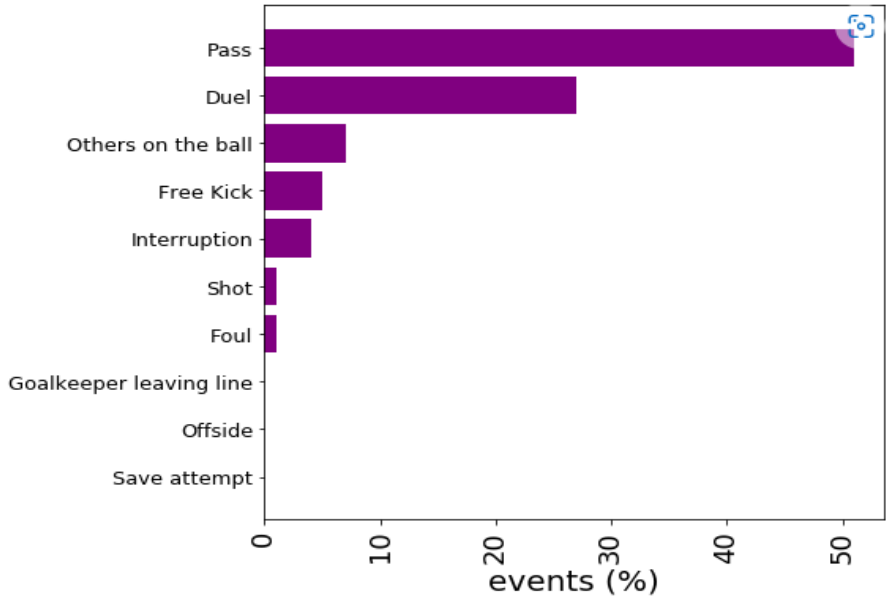

```
{'eventId': 8,
 'subEventName': 'Simple pass',
 'tags': [{'id': 1801}],
 'playerId': 134383,
 'positions': [{'y': 65, 'x': 30}, {'y': 86, 'x': 37}],
 'matchId': 2516739,
 'eventName': 'Pass',
 'teamId': 2444,
 'matchPeriod': '1H',
 'eventSec': 875.4974749999999,
 'subEventId': 85,
 'id': 179896753},
```


Σχήμα 17 : Παράδειγμα μιας απλής πάσας του αρχείου events.

Κατά μέσο όρο ένα match αποτελείται από 1,675 ± 110 γεγονότα, όπως φαίνεται και από την κατανομή των γεγονότων στο Σχήμα 18. Από το Σχήμα 19 πληροφορούμαστε ότι οι πάσες (passes) είναι το γεγονός που συμβαίνει συχνότερα, και καταλαμβάνει περίπου το 50% του συνόλου events. Τα duels, δηλαδή οι μονομαχίες/διεκδικήσεις της μπάλας (τάκλιν, ντρίμπλες, κλπ.) είναι τα επόμενα συχνότερα γεγονότα, και αποτελούν περίπου στο 28% του συνόλου. Αντίθετα, τα goals καταλαμβάνουν ποσοστό λιγότερο του 1% των συνολικών γεγονότων.

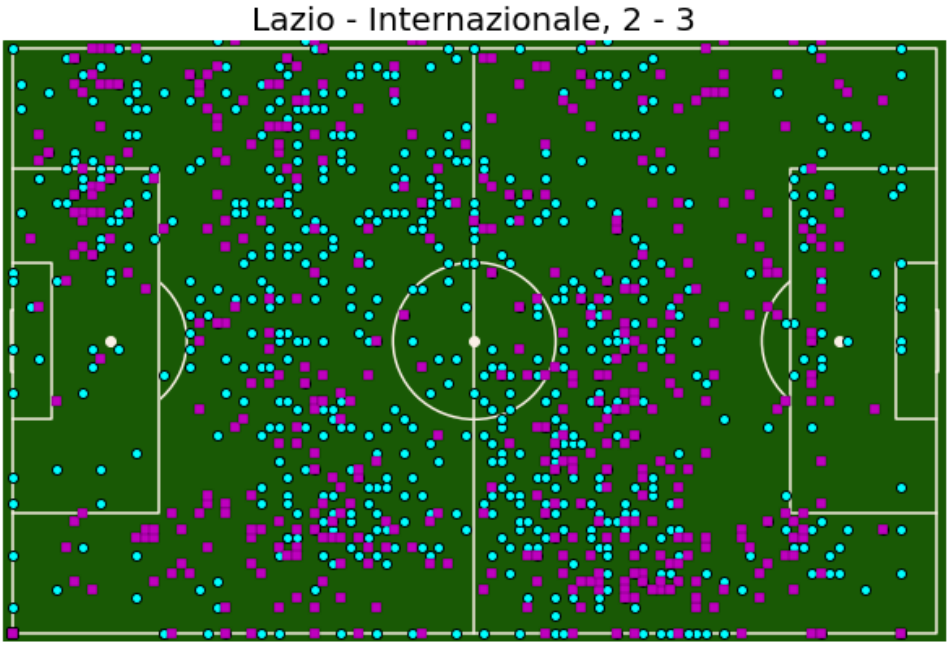


Σχήμα 18 : Η κατανομή του συνόλου των γεγονότων (events).



Σχήμα 19 : Συχνότητες ανά τύπο γεγονότος.

Στο Σχήμα 20 παρουσιάζεται μια οπτικοποίηση των γεγονότων του αγώνα της Lazio (γαλάζιο) ενάντια στην Internazionale (μωβ), στο Ιταλικό πρωτάθλημα ομάδων πρώτης κατηγορίας (20 Μαίου, 2018). Οι κουκίδες συμβολίζουν όλα τα γεγονότα (1,620) που καταγράφηκαν και είναι τοποθετημένες στα σημεία του γηπέδου όπου σημειώθηκαν.



Σχήμα 20 : Τα γεγονότα του αγώνα Lazio – Internazionale.

4.2. Προετοιμασία των δεδομένων

Η ποιότητα των δεδομένων και η πληροφορία που αντλείται από αυτά είναι καθοριστικά για την ικανότητα εκπαίδευσης ενός αλγορίθμου μηχανικής μάθησης αλλά και για την εξόρυξη ποιοτικής γνώσης. Ο καθαρισμός των δεδομένων και ο μετασχηματισμός τους στην κατάλληλη μορφή με βάση τις ιδιαιτερότητες του εκάστοτε αλγορίθμου, μπορεί να εξασφαλίσει πιο ακριβή αποτελέσματα. Συνήθεις πρακτικές που ακολουθούνται για το σκοπό αυτό είναι ο εντοπισμός και η διαχείριση ελλιπών και ακραίων τιμών (outliers), θορύβου και τυχόν ασυνεπών δεδομένων, η κανονικοποίηση, η μείωση των δεδομένων, η αντιμετώπιση πιθανών υψηλών συσχετίσεων μεταξύ μεταβλητών κ.α.

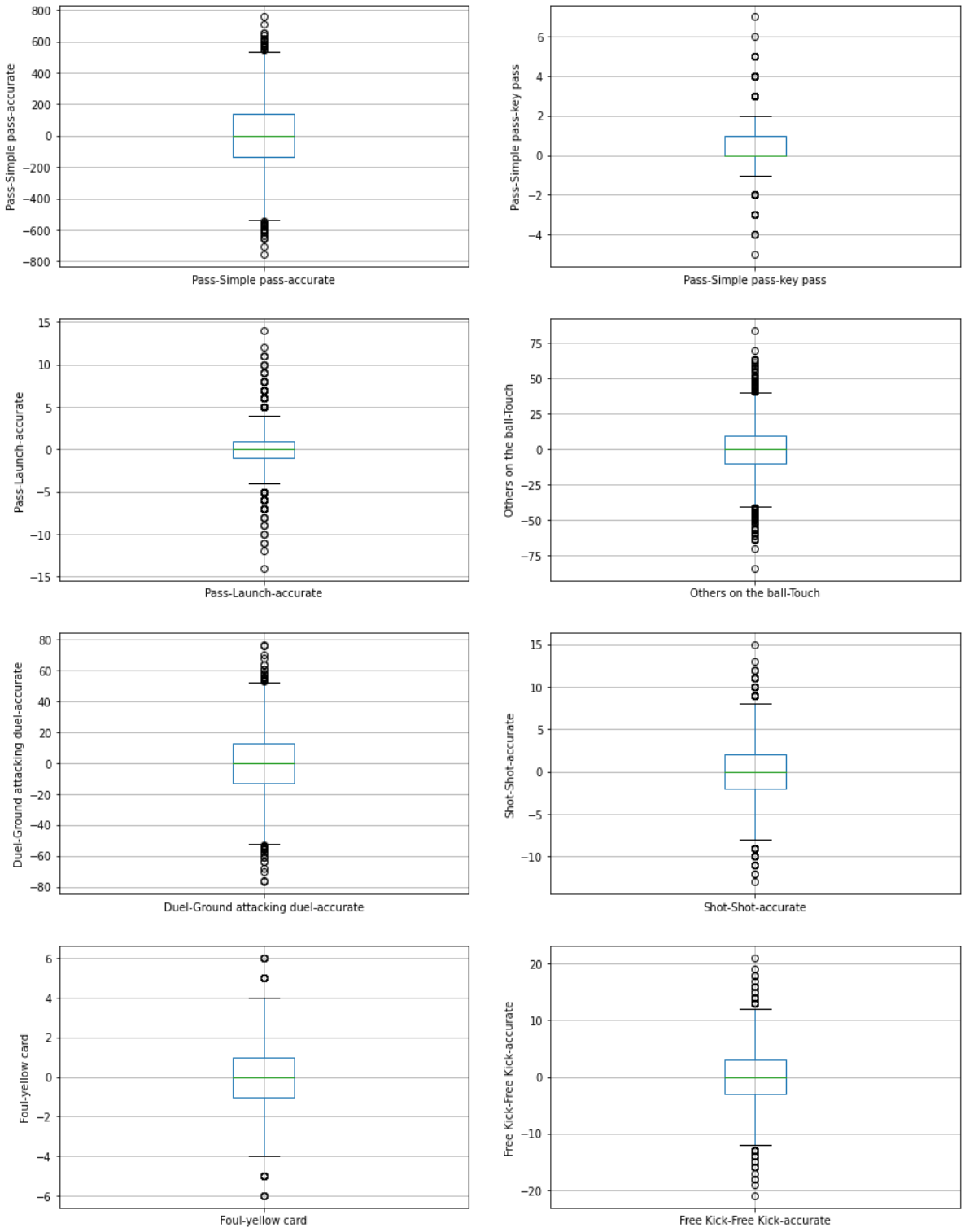
Στην παρούσα συλλογή δεδομένων η επιλογή των γνωρισμάτων έγινε με γνώμονα το είδος των γεγονότων και την έκβασή τους ή αλλιώς τον χαρακτηρισμό τους, για παράδειγμα ως έγκυρα ή μη. Το είδος ενός γεγονότος περιγράφεται από στα στοιχεία ‘type’ και ‘subtype’ του συνόλου events, ενώ η έκβασή του από το ‘tag’. Έτσι, ορίζεται ένα χαρακτηριστικό (performance

feature) για κάθε δυνατό συνδυασμό type, subtype και tag όπως φαίνεται στον Πίνακα 1. Τα χαρακτηριστικά μετρούν πόσες φορές εκτελέστηκε μια συγκεκριμένη κίνηση, παραδείγματος χάριν τον αριθμό των fouls με κιτρινη κάρτα. Επίσης, τα χαρακτηριστικά που σχετίζονταν με τον τερματοφύλακα εξαιρέθηκαν από την παρούσα ανάλυση καθώς ο ρόλος του και οι κανόνες που τον διέπουν δεν είναι συγκρίσιμοι με αυτούς των υπόλοιπων παικτών.

Πίνακας 1 : Χαρακτηριστικά.

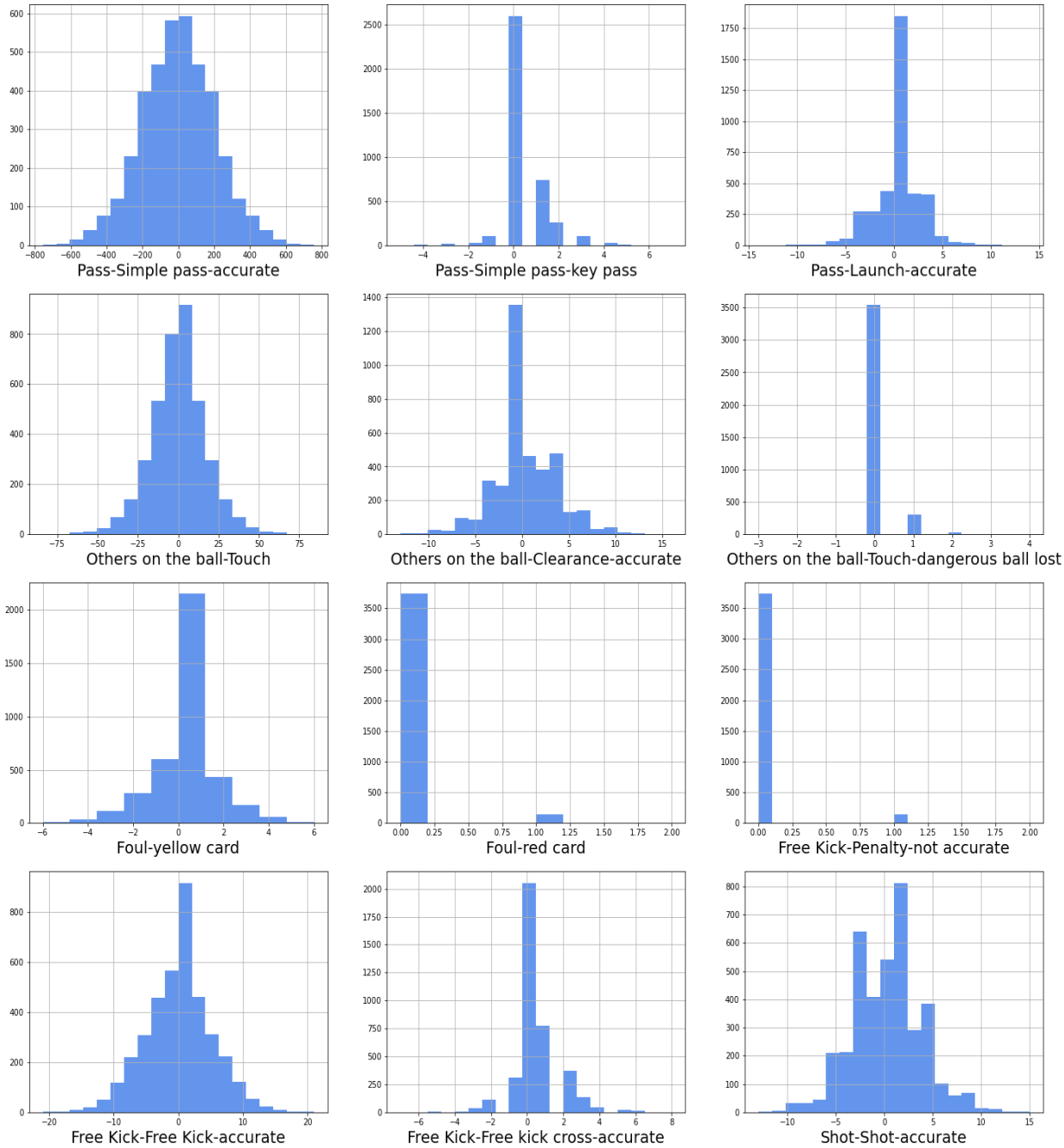
Type	**Subtypes**	**Tags**
pass	cross, hand, head, high, launch, smart, simple pass	accurate, not accurate, assist, key pass
foul		Yellow card, 2nd yellow card, red card
shot		accurate, not accurate
duel	air duel, ground attacking duel, ground defending duel, ground loose ball	accurate, not accurate
free kick	corner, normal kick, shot, cross, goal kick, throw in, penalty	accurate, not accurate, key pass, opportunity, assist, goal
offside	normal offside	
saves	Reflexes, normal save	accurate, not accurate
Others on the ball	acceleration, clearance, touch others	accurate, not accurate, goal, opportunity, interception, missed ball, feint, counterattack, assist, dangerous ball lost

Στην διερεύνηση των δεδομένων δεν εντοπίστηκαν ελλιπείς τιμές. Επίσης, λόγω του πλήθους των μεταβλητών που επιλέχθηκαν (71), η αναζήτηση για outliers έγινε δειγματοληπτικά με τη βοήθεια θηκογραμμάτων για μερικά από τα χαρακτηριστικά. Στα θηκογράμματα του Σχήματος 21 είναι εμφανές ότι υπάρχουν ακραίες τιμές στα δεδομένα, οι οποίες είναι δυνατόν να περιοριστούν με κατάλληλη ρύθμιση της παραμέτρου C του κατηγοριοποιητή SVM που θα εφαρμοστεί στη φάση της μάθησης. Παρόλα αυτά, τα χαρακτηριστικά που είχαν μικρή διασπορά τιμών (< 0.02) αφαιρέθηκαν.



Σχήμα 21: Θηκογράμματα χαρακτηριστικών.

Οι περισσότερες μεταβλητές έχουν μέση τιμή κοντά στο μηδέν, και το εύρος των τιμών τους διαφοροποιείται. Η κανονικοποίηση των χαρακτηριστικών προσδίδει μεγαλύτερη ακρίβεια στον αλγόριθμο κατηγοριοποίησης SVM. Στο Σχήμα 22 παρουσιάζονται οι κατανομές μερικών μεταβλητών.



Σχήμα 22 : Κατανομές χαρακτηριστικών.

Ο παρακάτω πίνακας περιέχει όλες τις μεταβλητές στις οποίες καταλήξαμε μετά την προπαρασκευή των δεδομένων.

Πίνακας 2 : Τελικά χαρακτηριστικά.

Duel-Air duel-not accurate	Pass-Smart pass-not accurate
Duel-Air duel-accurate	Pass-Launch-accurate
Duel-Ground loose ball duel-accurate	Pass-Head pass-accurate
Duel-Ground loose ball duel-not accurate	Pass-High pass-key pass
Duel-Ground attacking duel-not accurate	Pass-Head pass-key pass
Duel-Ground defending duel-not accurate	Pass-Simple pass-key pass
Duel-Ground defending duel-accurate	Pass-Smart pass-key pass
Duel-Ground attacking duel-accurate	Pass-Launch-not accurate
Free Kick-Free Kick-accurate	Free Kick-Throw in-not accurate
Free Kick-Free kick cross-accurate	Pass-Smart pass-accurate
Free Kick-Free kick cross-not accurate	Pass-Head pass-assist
Free Kick-Throw in-accurate	Pass-High pass-assist
Free Kick-Goal kick	Pass-Smart pass-assist
Free Kick-Free kick shot-not accurate	Pass-Simple pass-accurate
Free Kick-Penalty-not accurate	Pass-Cross-assist
Free Kick-Corner-not accurate	Pass-Simple pass-assist
Free Kick-Free Kick-not accurate	Others on the ball-Touch
Free Kick-Corner-accurate	Others on the ball-Clearance
Free Kick-Free kick shot-accurate	Others on the ball-Touch-dangerous ball lost
Free Kick-Penalty	Others on the ball-Touch-missed ball
Foul	Others on the ball-Touch-interception
Foul-red card	Others on the ball-Acceleration-not accurate
Foul-yellow card	Others on the ball-Clearance-accurate
Foul-second yellow card	Others on the ball-Touch-opportunity
Pass-Cross-not accurate	Others on the ball-Touch-feint
Pass-Cross-accurate	Others on the ball-Touch-counterattack
Pass-High pass-accurate	Others on the ball-Acceleration-accurate
Pass-High pass-not accurate	Others on the ball-Clearance-not accurate
Pass-Simple pass-not accurate	Shot-Shot-not accurate
Pass-Head pass-not accurate	Shot-Shot-accurate
Pass-Cross-key pass	

4.3. Κύριο πειραματικό μέρος

Έπειτα από την επιλογή των χαρακτηριστικών εκτελούμε τον κώδικα ώστε να παραχθούν οι αξιολογήσεις (performance ratings) και οι βαθμολογικές κατατάξεις για τους παίκτες (player rankings) του συνόλου δεδομένων μας. Τα αποτελέσματα παρουσιάζονται των αλγορίθμων παρουσιάζονται παρακάτω ανά φάση του πλαισίου Playerank.

4.3.1. Φάση μάθησης

Ζύγισμα των χαρακτηριστικών.

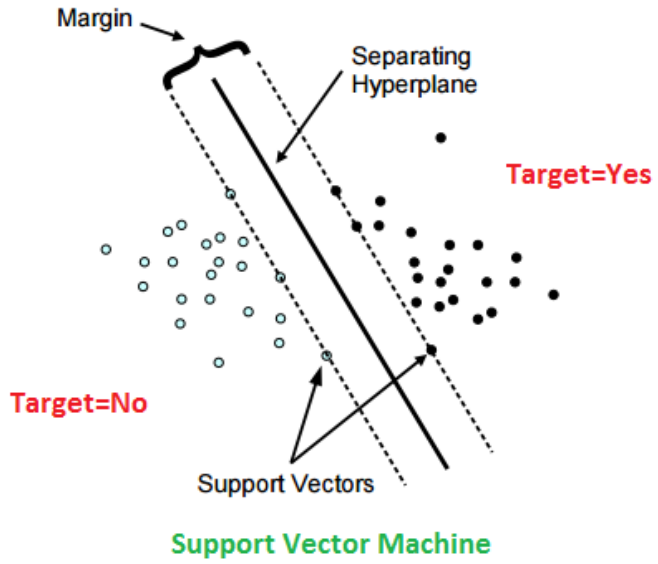
Στη διαδικασία αυτή, τα χαρακτηριστικά συναθροίζονται ανά ομάδα και ανά match ώστε να κατασκευαστούν τα διανύσματα επίδοσης σε επίπεδο ομάδας. Ακολουθεί μια κατηγοριοποίηση μεταξύ των διανυσμάτων αυτών και του αποτελέσματος του αγώνα. Το αποτέλεσμα του αγώνα ορίστηκε να λαμβάνει την τιμή 1 για τη νίκη, και 0 για την ήττα της ομάδας. Δοκιμάστηκαν και δύο διαφορετικοί ορισμοί, το αποτέλεσμα του match να παίρνει την τιμή 1 για την νίκη, και 0 για ήττα/ ισοπαλία, καθώς και η επιλογή της τιμής 1 για την νίκη, 0 για την ισοπαλία και 2 για την ήττα. Ωστόσο, οι δύο τελευταίοι ορισμοί οδηγούσαν σε χαμηλότερη ακρίβεια για τον γραμμικό SVM. (0.7696 και 0.6705 αντίστοιχα).

Το σύνολο δεδομένων το οποίο αποτελείται από τα τελικά μας χαρακτηριστικά και την μεταβλητή απόκρισης, που είναι η goal-scored, χωρίστηκε σε train (80% του συνόλου) και test set (20% του συνόλου) με τυχαίο τρόπο. Η πρακτική αυτή βοηθά στην εκπαίδευση μοντέλου μας, η οποία πραγματοποιείται στο train set. Το μοντέλο στη συνέχεια καλείται να ταξινομήσει σωστά τα άγνωστα σε αυτό δεδομένα του test set στις δύο κλάσεις της μεταβλητής απόκρισης.

Για την κατηγοριοποίηση εφαρμόστηκαν οι αλγόριθμοι κατηγοριοποίησης SVM με γραμμικό πυρήνα (LinearSVC) και η λογιστική παλινδρόμηση, οι οποίοι είχαν παρόμοια απόδοση.

Ο γραμμικός SVM είναι αλγόριθμος εποπτευόμενης μάθησης και εφαρμόζεται σε προβλήματα παλινδρόμισης, και κατά κύριο λόγο κατηγοριοποίησης. Κάθε μονάδα από τα δεδομένα μας (στη δική μας περίπτωση ένα διάνυσμα επίδοσης) θεωρείται σημείο ενός ν-διάστατου χώρου, όπου ν είναι το πλήθος των μεταβλητών μας. Η τιμή του κάθε χαρακτηριστικού είναι μια συγκεκριμένη συνιστώσα αυτής της μονάδας δεδομένου στο χώρο. Ο αλγόριθμος αυτός επιχειρεί να διαχωρίσει τα σημεία σε δύο κλάσεις μέσω ενός υπερεπιπέδου, με τον καλύτερο δυνατό τρόπο. Στόχος είναι η μεγιστοποίηση του περιθωρίου μεταξύ των δύο κατηγοριών. Ως περιθώριο (margin) ορίζεται ο κενός χώρος

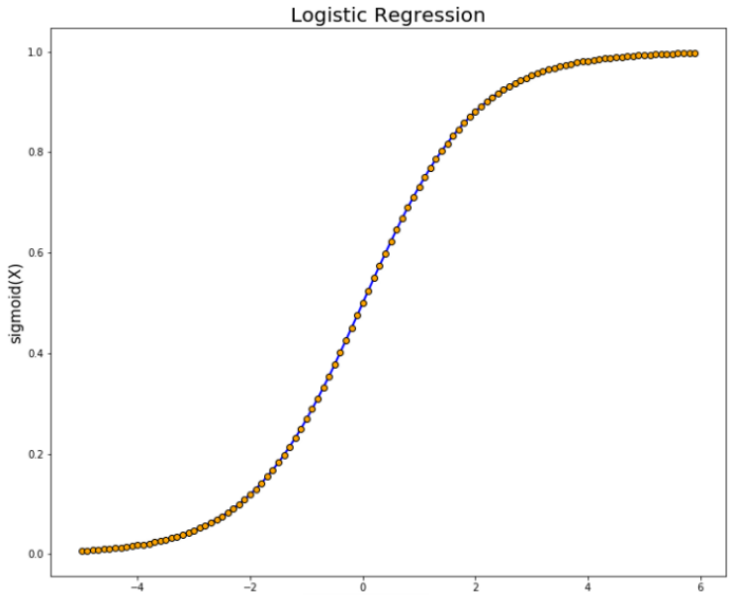
μεταξύ των δύο ευθειών που διέρχονται από τα κοντινότερα σημεία (support vectors). Τα νέα δεδομένα κατηγοριοποιούνται στις δύο κλάσεις ανάλογα με την πλευρά του περιθωρίου στην οποία εμπίπτουν. Στο Σχήμα 23 φαίνεται μια γραφική επεξήγηση του εν λόγω αλγορίθμου.



Σχήμα 23 : Γραφική αναπαράσταση του γραμμικού SVM.

Πηγή: https://www.kaggle.com/code/prashant111/svm-classifier-tutorial/notebook

Η λογιστική παλινδρόμηση είναι ένα μοντέλο ταξινόμησης των τιμών της μεταβλητής απόκρισης που στηρίζεται στη θεωρία πιθανοτήτων. Το μοντέλο ουσιαστικά προβλέπει την πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα σε μια καμπύλη σιγμοειδούς μορφής, την λογιστική καμπύλη (Σχήμα 24). Η δίτιμη λογιστική παλινδρόμηση η μεταβλητή απόκρισης αποτελεί το τυχαίο αποτέλεσμα εμφάνισης μιας από τις δύο δυνατές εκβάσεις του τύπου επιτυχία ή αποτυχία. Στην περίπτωσή μας επιτυχία είναι η νίκη της ομάδας, και αποτυχία η ήττα/ισοπαλία.



Σχήμα 24 : Καμπύλη λογιστικής παλινδρόμησης.

Πηγή: https://towardsdatascience.com/logistic-regression-explained-9ee73cede081

Στην μηχανική μάθηση είναι σημαντικό να αξιολογούμε την απόδοση (performance) ενός αλγορίθμου κατηγοριοποίησης. Η απόδοση μετράται συνήθως μέσω των μετρικών precision, recall, f1-score και accuracy. Η μετρική precision είναι το ποσοστό των περιπτώσεων που ο αλγόριθμος έχει ταξινομήσει σωστά σε μια κλάση ως προς τον αριθμό περιπτώσεων που έχει ταξινομήσει συνολικά, σωστά ή λανθασμένα, σε αυτή την κλάση. Η recall ή αλλιώς sensitivity είναι το ποσοστό των σωστά ταξινομημένων από τον αλγόριθμο περιπτώσεων σε μια κλάση ως προς τον αριθμό αυτών που ανήκουν πράγματι σε αυτή την κλάση. Πιο συγκεκριμένα, και με τη βοήθεια του Σχήματος 25, οι ορισμοί των δύο παραπάνω μέτρων είναι:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

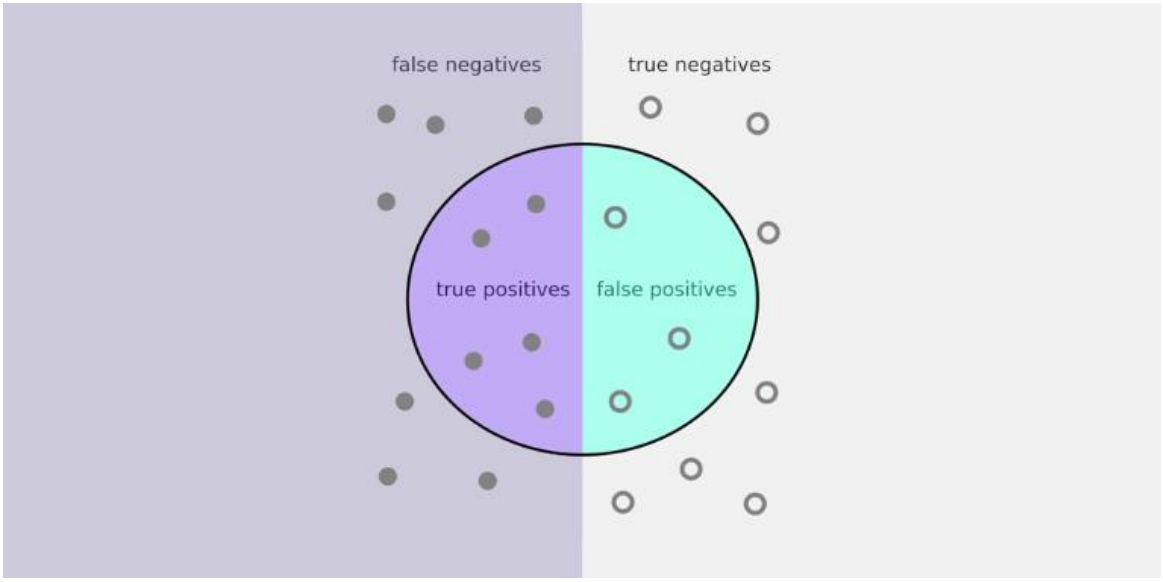
$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Η accuracy είναι το ποσοστό των σωστά ταξινομημένων περιπτώσεων προς τον συνολικό αριθμό των περιπτώσεων, δηλαδή:

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives}}$$

Η f1-score λαμβάνει υπόψιν τις precision και recall, περιγράφεται και ως ο αρμονικός μέσος των δύο, και ισούται με:

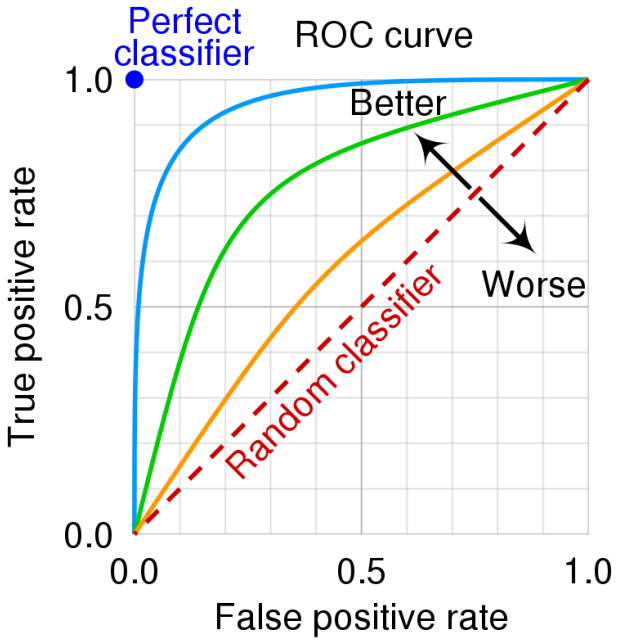
$$\text{f1-score} = 2 * \frac{precision * recall}{precision + recall}$$



Σχήμα 25 : Precision vs recall.

Πηγή: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

Συμπληρωματικά, σαν μέσο αξιολόγησης της διαγνωστικής ικανότητας ενός αλγορίθμου κατηγοριοποίησης σε δύο κλάσεις χρησιμοποιείται η καμπύλη AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) ή αλλιώς AUROC (Area Under the Receiver Operating Characteristics). Η καμπύλη ROC είναι η γραφική παράσταση της ευαισθησίας (sensitivity) σαν μια συνάρτηση του σφάλματος τύπου 1, και η περιοχή AUC είναι ένα μέτρο διαχωρισμού μεταξύ των δύο κλάσεων. Στο Σχήμα 26 παρουσιάζεται μια αναπαράσταση των μέτρων αυτών. Όσο πιο κοντά στη μονάδα είναι το εμβαδό της περιοχής AUC, τόσο καλύτερα μπορεί το μοντέλο να προβλέψει τις κλάσεις 0 και 1.



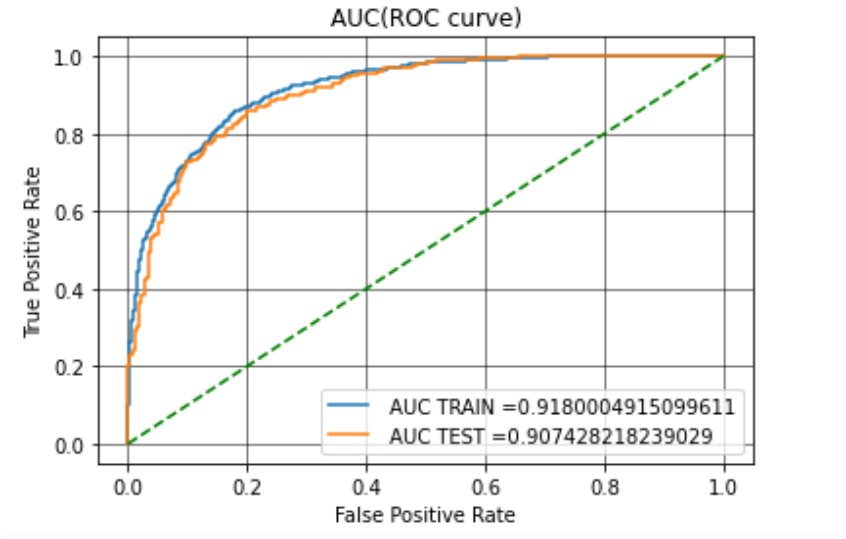
Σχήμα 26 : Area Under the Receiver Operating Characteristics.

Πηγή: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Τα αποτελέσματα του γραμμικού SVM φαίνονται στα Σχήματα 27 και 28. Η ακρίβεια (accuracy) του μοντέλου ισούται με 0.8211, ενώ το μέτρο AUC για το test set υπολογίστηκε να είναι 0.9074.

	precision	recall	f1-score	support
0	0.90	0.80	0.85	481
1	0.72	0.86	0.79	296
accuracy			0.82	777
macro avg	0.81	0.83	0.82	777
weighted avg	0.83	0.82	0.82	777

Σχήμα 27 : Αποτέλεσμα κατηγοριοποίησης με τον γραμμικό SVM.

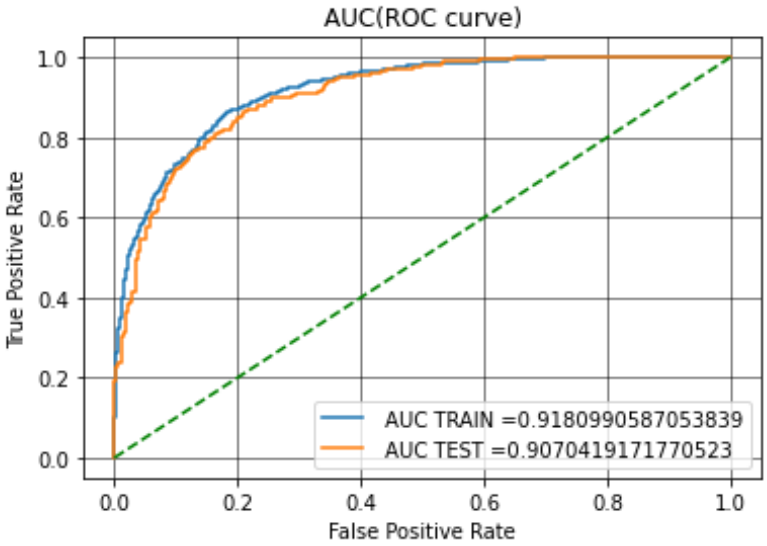


Σχήμα 28 : AUC με τον γραμμικό SVM.

Τα αποτελέσματα της λογιστικής παλινδρόμησης φαίνονται στα Σχήματα 29 και 30. Η ακρίβεια (accuracy) του μοντέλου ισούται με 0.8159, ενώ το μέτρο AUC για το test set υπολογίστηκε ίσο με 0.9070.

	precision	recall	f1-score	support
0	0.90	0.79	0.84	481
1	0.72	0.85	0.78	296
accuracy			0.82	777
macro avg	0.81	0.82	0.81	777
weighted avg	0.83	0.82	0.82	777

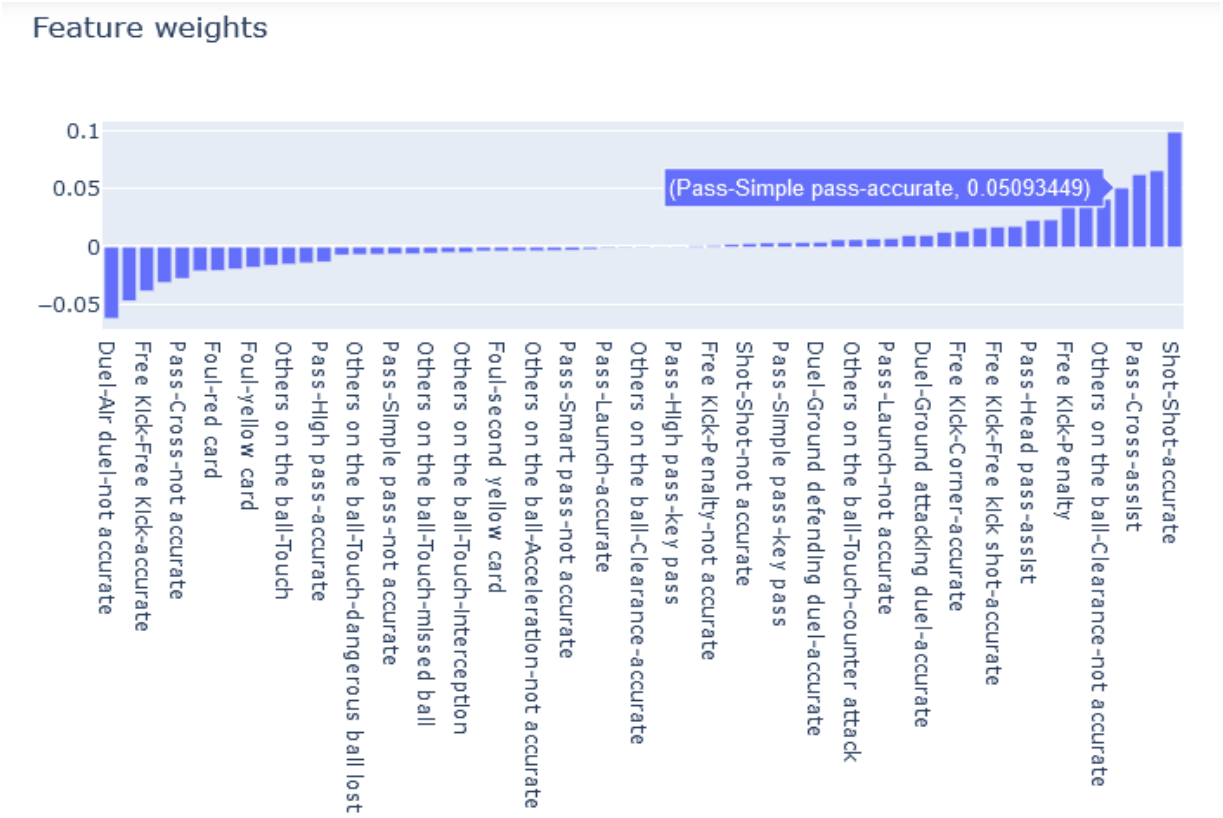
Σχήμα 29 : Αποτέλεσμα κατηγοριοποίησης με την λογιστική παλινδρόμηση.



Σχήμα 30 : AUC με την λογιστική παλινδρόμηση.

Για την πορεία της ανάλυσης χρησιμοποιούμε την κατηγοριοποίηση που παρήγαγε ο γραμμικός SVM.

Έπειτα από την κατηγοριοποίηση υπολογίστηκαν τα βάρη των χαρακτηριστικών, δηλαδή η συνεισφορά τους στο αποτέλεσμα του αγώνα. Μέσω του κώδικα παράγουμε ένα διαδραστικό διάγραμμα, το οποίο επιτρέπει την οπτική διερεύνηση των αποτελεσμάτων από το ζύγισμα των χαρακτηριστικών. Ένα στιγμιότυπο του διαγράμματος φαίνεται στο Σχήμα 31. Παρατηρούμε ότι το χαρακτηριστικό Shot-Shot-accurate, δηλαδή ένα έγκυρο σουτ, φαίνεται να έχει τη μεγαλύτερη θετική συνεισφορά στο αποτέλεσμα του αγώνα. Από την άλλη πλευρά ένα Duel-Air duel-not accurate, δηλαδή μια μη έγκυρη εναέρια διεκδίκηση της μπάλας που μπορεί να γίνει μεσω κεφαλιάς ή με το στήθος, φαίνεται να έχει την πιο αρνητική επιρροή στην έκβαση του αγώνα.



Σχήμα 31: Στιγμιότυπο διαδραστικού διαγράμματος για τα βάρη των χαρακτηριστικών.

Ανίχνευση θέσεων-ρόλων.

Βασική ιδιαιτερότητα του Playerank είναι, όπως έχει προαναφερθεί, το ότι λαμβάνει υπόψιν τις θέσεις των παικτών ώστε να παράξει τις αξιολογήσεις (ratings) τους, καθώς η σύγκριση παικτών με διαφορετικούς ρόλους δεν έχει ιδιαίτερο νόημα. Για το σκοπό αυτό υπολογίζεται το κέντρο επίδοσης ή αλλιώς οι μέσες συντεταγμένες θέσης για κάθε παίκτη ανά αγώνα. Στο σημείο αυτό δημιουργείται ένα dataframe το οποίο περιέχει τα κέντρα επίδοσης (avg_x , avg_y) για κάθε παίκτη (entity) ανά match, και τον αριθμό των γεγονότων που έχει σημειώσει ο κάθε ένας (n_events), όπως φαίνεται στο Σχήμα 32.

	match	entity	avg_x	avg_y	n_events
0	2499719	25413	69	44	53
1	2499719	370224	35	76	86
2	2499719	3319	63	43	108
3	2499719	120339	50	55	88
4	2499719	167145	54	67	93
...	...	...	...	...	...
46461	2058017	279545	38	12	57
46462	2058017	8200	41	56	20
46463	2058017	209091	52	21	11
46464	2058017	69411	71	36	14
46465	2058017	28115	62	90	15

Σχήμα 32 : Dataframe με τα κέντρα επίδοσης (avg_x,avg_y) των παικτών ανά match.

Στη συνέχεια εφαρμόζεται συσταδοποίηση μέσω του αλγορίθμου K-means, για να ομαδοποιηθούν τα κέντρα επίδοσης σε συστάδες-ρόλους.

Ο K-means αποτελεί έναν από τους δημοφιλέστερους και πιο απλούς αλγορίθμους μη εποπτευόμενης μάθησης, και η ιδέα πίσω από αυτόν είναι ουσιαστικά η ομαδοποίηση όμοιων δεδομένων, και στη συνέχεια η ανακάλυψη υποκείμενων μοτίβων ή πληφορορίας από τις ομάδες που προέκυψαν. Ο χρήστης δίνει στον αλγόριθμο το πλήθος k των συστάδων που επιθυμεί να παραχθούν. Ο K-means ξεκινά επιλέγοντας τυχαία k κεντροειδή (centroids), δηλαδή τα κέντρα των συστάδων, και τοποθετεί κάθε σημείο στη συστάδα με το πλησιέστερο κέντρο. Λειτουργεί επαναληπτικά, επαναπροσδιορίζοντας τη θέση των κεντροειδών, και εντάσσοντας ξανά κάθε σημείο στη συστάδα με το κοντινότερο κέντρο. Παράλληλα, μειώνει τη διασπορά εντός των συστάδων, ελαχιστοποιώντας τα τετράγωνα των Ευκλείδειων αποστάσεων μεταξύ των εσωτερικών τους σημείων. Ο αλγόριθμος σταματά όταν έχει εκπληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων, ή όταν τα κεντροειδή έχουν πια σταθεροποιηθεί και δεν υπάρχουν μετακινήσεις σημείων σε άλλες συστάδες.

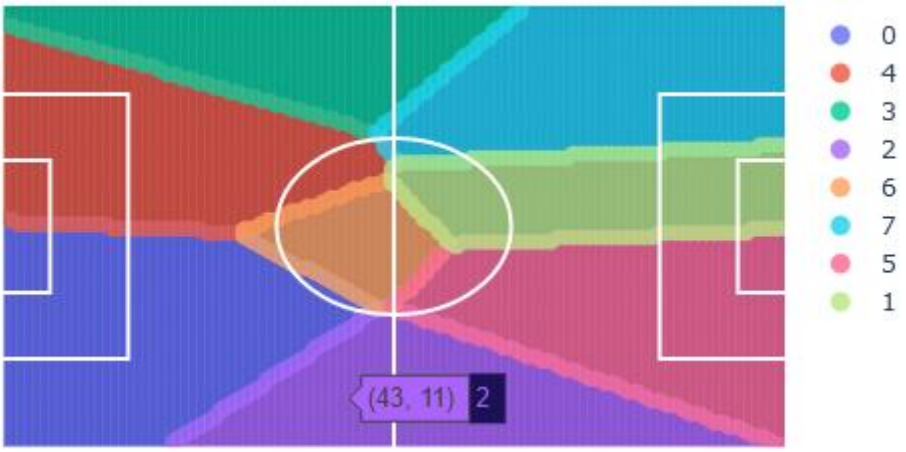
Για την αξιολόγηση της συσταδοποίησης βασιστήκαμε στο silhouette score, ένα μέτρο εκτίμησης της διαχωριστικής ικανότητας του αλγορίθμου, αλλά και της ομοιότητας που υπάρχει μεταξύ των σημείων εντός συστάδας. Η μετρική αυτή έχει εύρος τιμών [-1,1]. Όσο πιο κοντά στο +1 είναι η τιμή της, τόσο πιο ευδιάκριτες είναι οι συστάδες. Η τιμή 0 υποδεικνύει ότι οι ομάδες δεν διαφέρουν μεταξύ τους, ενώ η τιμή -1 ότι οι συστάδες έχουν δημιουργηθεί με λάθος τρόπο.

Καθώς ο ρόλος του τερματοφύλακα εξαιρείται από την ανάλυση, ο μέγιστος αριθμός συστάδων που μπορούμε να έχουμε είναι 10. Ο αλγόριθμος K-means εκτελέστηκε για k = 5,6,7,8,9,10 από δύο φορές. Την πρώτα φορά μέσω του ορίσματος kind = “single” υπαγορεύσαμε στον αλγόριθμο να ταξινομήσει κάθε κέντρο επίδοσης των παικτών σε μια μόνο συστάδα, αποφεύγοντας έτσι την ανίχνευση υβριδικών κέντρων. Τη δεύτερη φορά ορίσαμε kind = “multi” και επιτρέψαμε την αναζήτηση υβριδικών κέντρων, δηλαδή την ανάθεση κέντρων επίδοσης σε δύο ή περισσότερες συστάδες, με κατώφλι ανοχής σε υβριδικά κέντρα δ_s = 0.1. Ο καλύτερος αριθμός συστάδων όπως μπορούμε να συμπεράνουμε και από τα silhouette scores του Σχήματος 33 για τα διάφορα k, είναι k = 8 (ss = 0.385).

```
n_clust |silhouette
--------------------
5       |0.3648
6       |0.3644
7       |0.3533
8       |0.385
9       |0.3591
10      |0.368
Best: n_clust=8 (silhouette=0.385)
```

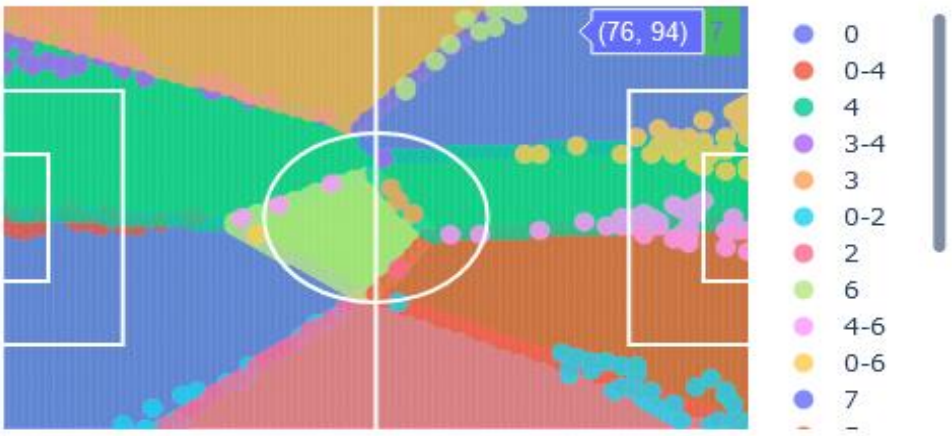
Σχήμα 33 : Silhouette scores για k=5,6,7,8,9.

Η συσταδοποίηση που έκανε ο 8-means χωρίς να υπολογίσει υβριδικά κέντρα φαίνεται στο Σχήμα 34, το οποίο είναι στιγμιότυπο από το διαδραστικό γράφημα που παράγει ο κώδικας. Στο συγκεκριμένο μπορεί να διαπιστώσει κανείς τις συντεταγμένες των κέντρων επίδοσης που περιέχονται σε κάθε συστάδα. Θεωρούμε ότι η επίθεση γίνεται από αριστερά προς τα δεξιά.



Σχήμα 34 : Συσταδοποίηση 8-means χωρίς υβριδικούς ρόλους.

Η συσταδοποίηση που έκανε ο 8-means συμπεριλαμβάνοντας και υβριδικούς ρόλους απεικονίζεται στο Σχήμα 35. Οι κουκίδες που υπάρχουν στα σύνορα των συστάδων είναι τα υβριδικά κέντρα επίδοσης, τα οποία μπορεί να αντιστοιχούν σε 2 ή παραπάνω ρόλους.



Σχήμα 35 : Συσταδοποίηση 8-means με υβριδικούς ρόλους.

Η ερμηνεία που δώσαμε στις συστάδες 0-7 που έχουν δημιουργηθεί παρουσιάζεται στον ακόλουθο πίνακα.

Πίνακας 3 : Ερμηνεία των 8 συστάδων που προέκυψαν από τον 8-Means.

συστάδα	**όνομα θέσης-ρόλου**	**περιγραφή θέσης-ρόλου**
0	δεξής κεντρικός αμυντικός	παίζει κοντά στην μεγάλη περιοχή της άμυνας, κυρίως στα δεξιά
1	κεντρικός επιθετικός	παίζει κοντά στο αντίπαλο τέρμα με σκοπό το σκοράρισμα
2	δεξής αμυντικός	παίζει στη δεξιά πλευρά του γηπέδου με κύριο σκοπό την άμυνα και δευτερεύοντα την επίθεση
3	αριστερός αμυντικός	παίζει στην αριστερή πλευρά του γηπέδου με κύριο σκοπό την άμυνα και δευτερεύοντα την επίθεση
4	αριστερός κεντρικός αμυντικός	παίζει κοντά στην μεγάλη περιοχή της άμυνας, κυρίως στα αριστερά
5	δεξής μέσος	πάιζει στη δεξιά περιοχή του κέντρου με σκοπό τη δημιουργία επιθέσεων
6	αμυντικός μέσος	παίζει άμυνα στο κέντρο του γηπέδου, μπροστά από τους κεντρικούς αμυντικούς. Είναι η σύνδεση της άμυνας με την επίθεση
7	αριστερός μέσος	πάιζει στην αριστερή περιοχή του κέντρου με σκοπό τη δημιουργία επιθέσεων

Τέλος, ένα χρήσιμο για τη συνέχεια αποτέλεσμα της φάσης αυτής είναι η παραγωγή του role matrix, ο οποίος έχει δομή dictionary, και περιέχει πληροφορία για τους ρόλους που αντιστοιχούν σε κάθε κέντρο επίδοσης.

4.3.2. Φάση αξιολόγησης

Στη φάση αυτή γίνεται εξόρυξη των διανυσμάτων επίδοσης των παικτών ανά αγώνα. Λαμβάνοντας υπόψιν τα βάρη των χαρακτηριστικών που υπολογίστηκαν στην προηγούμενη φάση, και τη σημαντικότητα των goals, την οποία θέσαμε ίση με το 10% του τελικού βαθμού αξιολόγησης, παράγονται οι αξιολογήσεις της επίδοσης (performance ratings) των παικτών ανά αγώνα, και κανονικοποιουνται στο διάστημα [0,1]. Για να υπολογίσουμε τον τελικό βαθμό αξιολόγησης κάθε παίκτη στο σύνολο των αγώνων που συμμετείχε χρησιμοποιούμε ως συνάρτηση συνάθροισης τον μέσο όρο.

4.3.3. Φάση βαθμολογικής κατάταξης - Αποτελέσματα

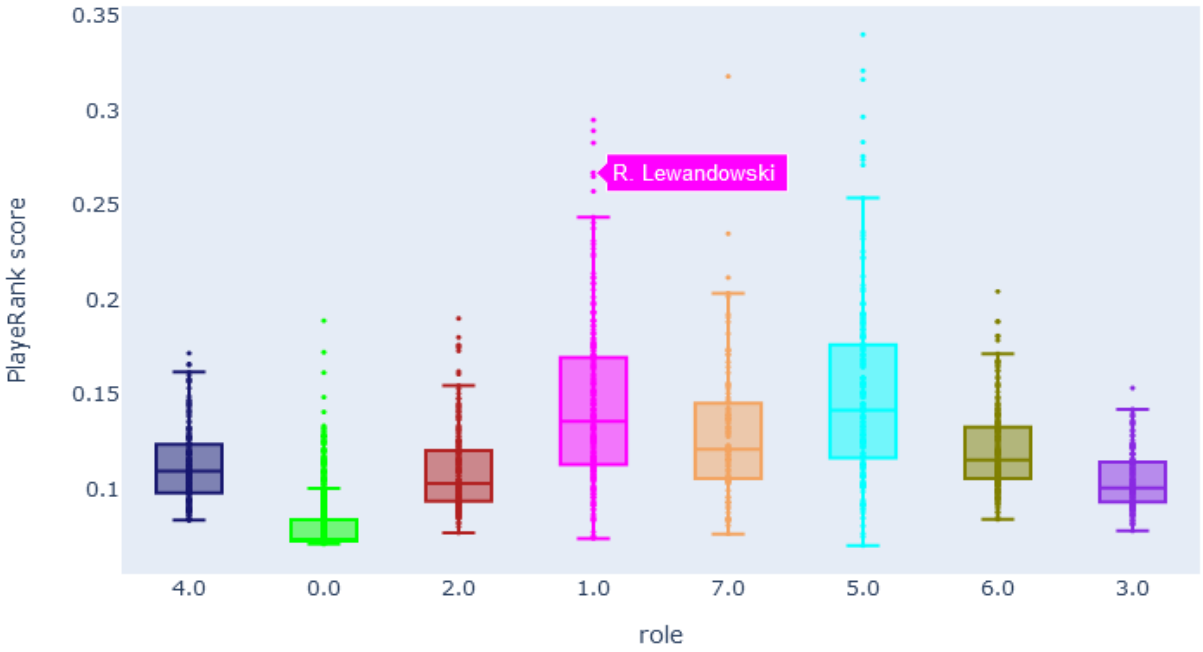
Τελικά, υπολογίζονται οι θέσεις-ρόλοι των παικτών με βάση τον role matrix που έχει προκύψει από την φάση της μάθησης, και κατασκευάζεται ένα dataframe το οποίο περιέχει τις βαθμολογίες των παικτών ανά ρόλο.

Στον Πίνακα 4 παρουσιάζονται οι 3 κορυφαίοι παίκτες κάθε θέσης με βάση τα rankings που παράχθηκαν για αυτούς. Οι βαθμοί περιέχονται στη στήλη Playerank.

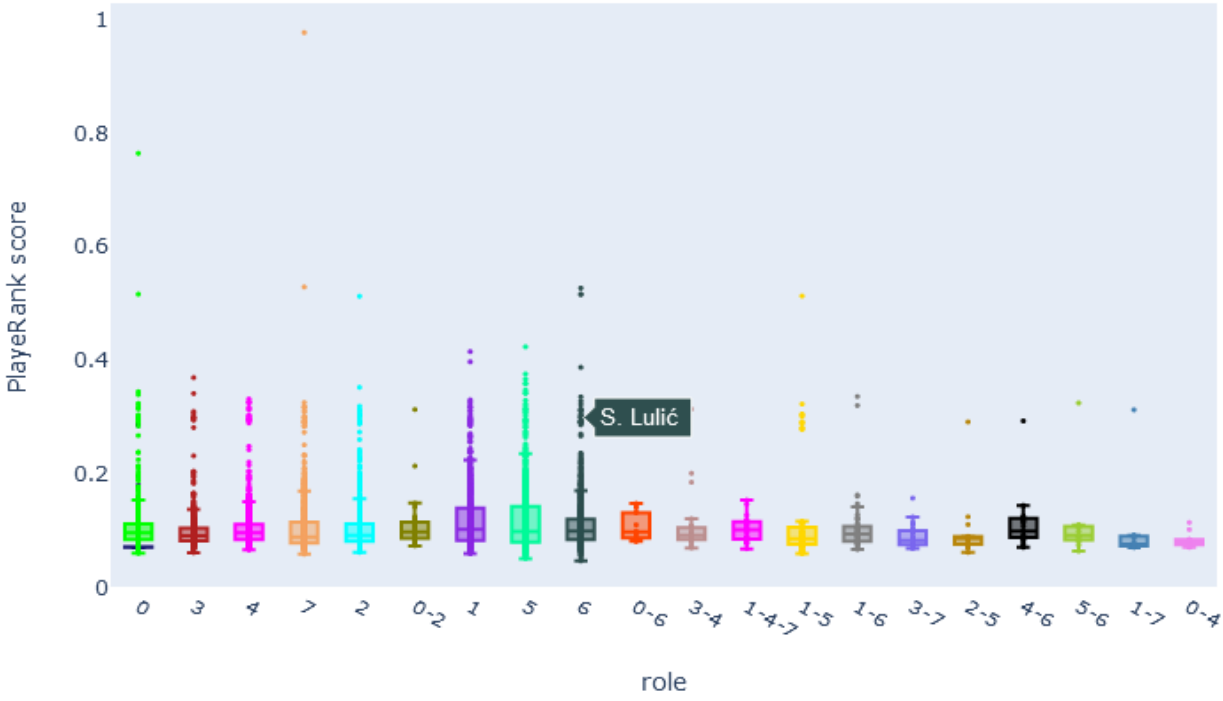
Στα Σχήματα 36 και 37 παρουσιάζονται οι κατανομές των βαθμολογιών των παικτών ανά ρόλο. Πιο συγκεκριμένα οι εικόνες αυτές είναι στιγμιότυπα από τα διαδραστικά διαγράμματα που παράγονται από τον κώδικα. Κάθε θηκόγραμμα αντιπροσωπεύει μια συστάδα (ρόλο), και περιέχει τις βαθμολογίες των παικτών που κατέχουν αυτή τη θέση. Θα μπορούσαμε να πούμε ότι οι κεντρικοί επιθετικοί (1) και οι δεξιοί μέσοι (5) έχουν τα υψηλότερα ratings, και αρκετά outlier ratings, το οποίο σημαίνει ότι στις θέσεις αυτές παρατηρούνται εξαιρετικά καλές επιδόσεις. Στο δεύτερο σχήμα φαίνεται το αντίστοιχο διάγραμμα συμπεριλαμβανομένων όμως και των υβριδικών κέντρων.

Πίνακας 4 : Βαθμολογίες των τριών κακλύτερων παικτών ανά ρόλο.

θέση-ρόλος	**αριθμός συστάδας**	**Παίκτης**	**Playerank**
δεξής κεντρικός αμυντικός	0	K. Koulibaly	0.172599
		Sergio Ramos	0.161821
		Z. Fedda	0.148874
κεντρικός επιθετικός	1	L. Messi	0.295248
		E. Cavani	0.289547
		S. Agüero	0.283173
δεξής αμυντικός	2	F. Ribéry	0.190448
		L. Insigne	0.180474
		F. Ghoulam	0.176415
αριστερός αμυντικός	3	Azpilicueta	0.153704
		M. De Sciglio	0.142514
		A. Valencia	0.140419
αριστερός κεντρικός αμυντικός	4	Aritz Elustondo	0.172088
		Naldo	0.166437
		V. Kompany	0.162104
δεξής μέσος	5	M. Icardi	0.340381
		P. Aubameyang	0.321260
		Neymar	0.316593
αμυντικός μέσος	6	J. Veretout	0.204673
		Illarramendi	0.189004
		A. Vidal	0.188561
αριστερός μέσος	7	Mohamed Salah	0.318273
		F. Thauvin	0.235198
		B. Traoré	0.212009



Σχήμα 36 : Στιγμιότυπο από διαδραστική εικόνα των κατανομών των βαθμολογιών ανά ρόλο.



Σχήμα 37 : Στιγμιότυπο από διαδραστική εικόνα των κατανομών των βαθμολογιών ανά ρόλο, λαμβάνοντας υπόψιν και τα υβριδικά κέντρα.

5. Συμπερασματολογία

5.1. Συμπεράσματα

Η εξέλιξη της τεχνολογίας βοηθά στην ανάπτυξη νέων μηχανισμών συλλογής ογκωδών και περισσότερο λεπτομερών δεδομένων, τα οποία προσφέρονται για έρευνα και εξυπηρετούν στην κατασκευή ακριβέστερων συστημάτων αξιολόγησης και βαθμολογικής κατάταξης αθλητών ή/και ομάδων.

Ο έλεγχος της καταλληλότητας αυτών των συστημάτων έχει ουσία να πραγματοποιείται με έναν ποιοτικό τρόπο βασισμένο στην ανθρώπινη κρίση, από εμπειρογνώμονες του χώρο, κυνηγούς ταλέντων κ.α., και όχι να αρκείται μόνο σε ποσοτικά μέτρα αξιολόγησης, όπως για παράδειγμα η εμπορική αξία παικτών ή ο αριθμός των γκολς, τα οποία είναι απλοϊκά και μονόπλερα.

Για την εγκυρότητα των αποτελεσμάτων μας δεν υπήρχε η πολυτέλεια να ληφθεί βοήθεια από εμπειρογνώμονες του ποδοσφαίρου, κάτι που δεν ήταν και τόσο αναγκαίο στα πλαίσια αυτής της διπλωματικής εργασίας. Ωστόσο έγινε σύγκριση με τα αποτελέσματα της εργασίας των Pappalardo et al. (2019), εστιασμένη στη φάση της μάθησης. Ο περιορισμένος όγκος δεδομένων που είχαμε στη διάθεσή σε σχέση με το αρχικό σύνολο δεδομένων των Pappalardo et al., και η διαφορετική επιλογή χαρακτηριστικών έδωσε ικανοποιητικά αποτελέσματα κατηγοριοποίησης (accuracy=0.82, AUC=0.9074), συγκριτικά με αυτά των Pappalardo et al. (accuracy=0.82, AUC=0.89). Αντίθετα, το ζύγισμα των χαρακτηριστικών στην δική μας περίπτωση δεν είναι νοηματικά απόλυτα σωστό. Μια σημαντική διαφορά εντοπίζεται στα βάρη των χαρακτηριστικών που αναφέρονται στα fouls. Πιο συγκεκριμένα, στην εργασία των Pappalardo et al. οι μεταβλητές των fouls κατείχαν την πιο αρνητική συνεισφορά στην έκβαση του παιχνιδιού, ενώ στην δική μας περίπτωση δεν φαίνονται να είναι τα πιο «επιβλαβή» γεγονότα για έναν αγώνα, παρουσιάζοντας παρόλα αυτά σημαντικά αρνητικά βάρη (βλ. Σχήμα 10 και Σχήμα 31). Όσον αφορά στην ανίχνευση ρόλων, οι δύο εργασίες συμφωνούν στα αποτελέσματα της συσταδοποίησης και στην νοηματοδότηση των συστάδων, έχοντας ωστόσο διαφορετικά silhouette scores, σαφώς χαμηλότερο στην δική μας περίπτωση (ss=0.385, έναντι του ss=0.43 των Pappalardo et al.). Τέλος, οι βαθμολογικές κατατάξεις που προέκυψαν από την εφαρμογή του PlayeRank στις δύο εργασίες δεν ταυτίζονται, καθώς πρόκειται για σύνολα δεδομένων διαφορετικής τάξης μεγέθους.

5.2. Μελλοντικές επεκτάσεις

Το πλαίσιο PlayeRank προσφέρεται για συγκριτική ανάλυση μεταξύ παικτών, και εξόρυξη γνώσης μέσα από διαφορετικές οπτικοποιήσεις των αποτελεσμάτων

του. Μια επέκταση της εργασίας μπορεί να είναι η συγκριτική μελέτη της απόδοσης πολλών διαφορετικών αλγορίθμων κατηγοριοποίησης και συσταδοποίησης που διέπουν τη φάση της μάθησης του πλαισίου. Μια δεύτερη και ιδιαιτέρως ενδιαφέρουσα επέκταση της εργασίας δεδομένου ότι το πλαίσιο ανιχνεύει ρόλους παικτών, είναι η διερεύνηση μέσω πιθανοτήτων ή εντροπίας, της προσαρμοστικότητας των παικτών, δηλαδή της ικανότητάς τους να αλλάζουν θέση από παιχνίδι σε παιχνίδι και πως συμπεριφέρονται εκεί.

6. Παράρτημα - Εργαλεία που χρησιμοποιήθηκαν

Το πρακτικό κομμάτι της εργασίας πραγματοποιήθηκε σε λογισμικό Windows 10 Home, και ο κώδικας εκτελέστηκε σε Python 3.8. Η εκτέλεση του κώδικα στην νεότερη έκδοση της Python 3.10 παρουσίασε κάποια προβλήματα στη χρήση της βιβλιοθήκης pandas, για αυτό το λόγο στραφήκαμε σε παλαιότερη έκδοση.

Ο κώδικάς μας είναι βασισμένος στο tutorial που διατίθεται στο Google Colab: https://colab.research.google.com/drive/1K7PENj9UrvuMpTmLpaPgwZg09wHf-2GE#scrollTo=yqPvk5w197M_ , και λειτουργεί καλώντας Classes αντικείμενα, τα οποία έχουν χτιστεί επάνω σε αυτά του ανοικτού αποθετηρίου Github: https://github.com/mesosbrodleto/playerank.

Για την εκτέλεση του κώδικα απαιτείται η εγκατάσταση των βιβλιοθηκών pandas, numpy, seaborn, json, joblib, matplotlib, sklearn, plotly, collections, scipy, glob.

7. Βιβλιογραφία

Acs, B. & Toka, L. (2021). A career in football: what is behind an outstanding market value? *In Machine Learning and Data Mining for Sports Analytics workshop (MLSA 2021)*.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S. & Matthews, I. (2014). Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data. *IEEE International Conference on Data Mining*. pp. 725-730. doi: 10.1109/ICDM.2014.133.

Brooks, J., Kerr, M., & Guttag, J. (2016). Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. *In Proceedings of the 22nd ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*. pp. 49–55. doi: 10.1145/2939672.2939695.

Constantinou, A. C. & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *In Journal of Quantitative Analysis in Sports*. doi: 10.1515/jqas-2012-0036.

Duch, J., Waitzman, J. S. & Nunes Amaral, L. A. (2010). Quantifying the Performance of Individual Players in a Team Activity. *PLOS ONE* 5, 6 (2010), 1–7. doi: 10.1371/journal.pone.0010937.

Glickman, M. E. (1995). A Comprehensive Guide to Chess Ratings. A subsequent version of this paper appeared in the American Chess Journal, 3, pp: 59-102.

Glickman, M. E. (2013). (revised). Example of the Glicko-2 system. http://www.glicko.net/glicko.html.

Glickman, M. E. (2016) (revised). The Glicko system. http://www.glicko.net/glicko.html.

Gudmundsson, J. & Horton, M. (2017). Spatio-Temporal Analysis of Team Sports. *Computing Surveys* 50, 2 (2017), 22:1–22:34. doi: 10.1145/ 3054132.

Gupte, M., Shankar, P., Li, J. & Muthukrishnan, S. (2011). Finding hierarchy in directed online social networks. *In Proceedings of the 20th international conference on World wide web*. pp. 557–566. doi: 10.1145/1963405.1963484.

Hartigan, J. A. & Wong, M. A. (1979). A k-means clustering algorithm. *JSTOR: Applied Statistics* 28, 1 (1979), pp. 100–108.

Herbrich, R., Minka, T. & Graepel, T. (2006). TrueSkill™: A Bayesian Skill Rating System. *In Procs of the 19th Intl Conf on Neural Information Processing Systems*. pp. 569–576.

Lasek, J., Szlávik, Z. & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition* 1, 1 (2013), pp. 27-46. doi: 10.1504/IJAPR.2013.052339.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research* 263, 2 (2017), 611–624. doi: 10. 1016/j.ejor.2017.05.005.

Müller, R., Langer, S., Ritz, F., Roch, C., Illium, S. & Linnhoff-Popien, C. (2019). Soccer Team Vectors. *In Machine Learning and Data Mining for Sports Analytics workshop (MLSA 2019)*. doi: 10.1007/978-3-030-43887-6_19.

Neumann, S., Ritter, J. & Budhathoki, K. (2018). Ranking the Teams in European Football Leagues With Agony. *In Machine Learning and Data Mining for Sports Analytics workshop (MLSA 2018)*. pp. 55–66. doi: 10.1007/978-3-030-17274-9_5

Nsolo, E., Lambrix, P., & Carlsson, N. (2018). Player valuation in European football. *In Machine Learning and Data Mining for Sports Analytics workshop (MLSA 2018)*. doi: 10.1007/978-3-030-17274-9_4.

Pappalardo, L. & Cintia, P. (2017). Quantifying the relation between performance and success in soccer. *Advances in Complex Systems* 20, 4 (2017). doi: 10.1142/S021952591750014X.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology* 10(5). pp: 1-27. doi: 10.1145/3343172.

Pappalardo, L., Cintia, P., Rossi, A., Ferragina, P., Pedreschiet, D. & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Sci Data* 6, 236 (2019). doi:10.1038/s41597-019-0247-7.

Pettigrew, S. (2015). Assessing the offensive productivity of NHL players using in-game win probabilities. *In MIT Sloan Sports Analytics Conference*. Convention and Exhibition Center, Boston, MA, USA.

Shulte, O. & Zhao, Z. (2017). Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact. *In MIT Sloan Sports Analytics Conference*. Hynes Convention Center, Boston, MA, USA.

Sismanis, Y. (2010). How I won the "Chess Ratings - Elo vs the Rest of the World" Competition. *CoRR*. Doi: 10.48550/arXiv.1012.4571.

Skinner, G.K. & Freeman, G.H. (2009). Soccer matches as experiments: how often does the 'best' team win? *Journal of Applied Statistics* 36(10). pp. 1087–1095. doi: 10.1080/02664760802715922.

Spall, J. C. (2003). Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. ISBN 0-471-33052-3. doi: 10.1002/0471722138.

Stanojevic, R. & Gyarmati, L. (2016). Towards Data-Driven Football Player Assessment. *In Proceedings of the IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. pp. 167-172. doi: 10.1109/ICDMW.2016.0031.

Opta Sports Ltd. Opta Live Performance Data. (2015).

http://www.optasports.com/about/what-we-do/live-performance-data.aspx.

World Football Elo Ratings. eloratings.net. (2022). Retrieved 27 September 2022.

https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

https://github.com/mesosbrodleto/playerank/blob/master/README.md.

https://www.kaggle.com/code/prashant111/svm-classifier-tutorial/notebook

https://towardsdatascience.com/logistic-regression-explained-9ee73cede081