UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN
INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO FIN DE MASTER

Application of Machine Learning and Data Analysis Techniques in Professional Football

JAVIER MAZARÍO PICAZO

2024

TRABAJO DE FIN DE MASTER

Título:	Aplicación de Técnicas de Aprendizaje Automático y Análisis de Datos en el Fútbol Profesional
Título (inglés):	Application of Machine Learning and Data Analysis Techniques in Professional Football
Autor:	Javier Mazarío Picazo
Tutor:	Carlos Ángel Iglesias Fernandez
Ponente:	Javier Mazarío Picazo
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO DE FIN DE MASTER

Application of Machine Learning and Data Analysis Techniques in Professional Football

FEBRERO 2024

Resumen

La analítica de datos y el uso de aprendizaje automático se está consolidando en el mundo del deporte, en concreto en el fútbol profesional. Cada vez son más los clubes profesionales que tienen equipos dedicados a este ámbito, los cuales toman cada vez más importancia a la hora de tomar decisiones. Esto abarca desde daciones a nivel de dirección deportiva y gestión del club, como en el análisis de rivales para la preparación de los partidos. En este trabajo se realiza un estudio del estado del arte del uso de estas técnicas, desde la evaluación de rendimiento de los jugadores y la identificación de patrones en el juego hasta la optimización de estrategias y la gestión de recursos dentro de los clubes.

El núcleo de este proyecto se centra en la aplicación de un algoritmo de aprendizaje automático no supervisado para clusterizar equipos y jugaodres con características similares. El estudio tiene todas las fases características de un proyecto de este tipo. En primer lugar, se realizó un estudio de los diferentes proveedores de datos existentes. Después, se llevó a cabo una etapa de preprocesameinto de los datos con el objetivo de la obtención y normalización de las diferentes estadísticas y métricas que se iban a evaluar en el estudio. Finalmente, se estudió la correlación entre las diferentes métricas y se aplicó el algoritmo K-Means. Una vez aplicado el algoritmo de clusterización, se realizó un análisis de los resultados con el objetivo de sacar conclusiones acerca del rendimiento de equipos y jugadores.

Adicionalmente, se desarrolló una aplicación interactiva en Streamlit, que ofrece una plataforma para la visualización de variedad de métricas y estadísticas, facilitando de esta forma su accesibilidad y comprensión. Una herramienta valiosa para entrenadores y analistas, que sirve como complemento para la toma de decisiones basadas en datos.

Palabras clave: Aprendizaje automático, Fútbol profesional, Análisis de datos, K-Means, Streamlit

Abstract

Data analytics and the use of machine learning is consolidating in the world of sports, specifically in professional football. More and more professional clubs have teams dedicated to this area, which are becoming increasingly important when it comes to making decisions. This ranges from donations at the level of sports management and club management, as in the analysis of rivals for the preparation of the matches. In this paper, a state-of-the-art study of the use of these techniques is carried out, from the evaluation of player performance and the identification of patterns in the game to the optimization of strategies and the management of resources within the clubs.

The core of this project focuses on the application of an unsupervised machine learning algorithm to cluster teams and players with similar characteristics. The study has all the characteristic phases of such a project. First, a study of the different existing data providers was carried out. Then, a data preprocessing stage was performed in order to obtain and normalize the different statistics and metrics to be evaluated in the study. Finally, the correlation between the different metrics was studied and the K-Means algorithm was applied. Once the clustering algorithm was applied, an analysis of the results was performed to draw conclusions about the performance of the teams and players.

Additionally, an interactive application was developed in Streamlit, which offers a platform for visualization of a variety of metrics and statistics, thus facilitating their accessibility and understanding. A valuable tool for coaches and analysts, it serves as a complement to data-driven decision making.

Keywords: Machine learning, Professional football, Data analysis, K-Means, Streamlit

Agradecimientos

A todas aquellas personas que me han acompañado a lo largo de este camino.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Context

The world of sports has always been a good place for data enthusiasts. In recent years, the development of technologies has allowed the integration of data science into sports, and football is no exception. Data are transforming how we understand the game, from analyzing player performance [5] to predicting the outcomes of matches [4].

Applying analytics to football opens up a world that transforms the industry at all levels, from football coaches or talent scouts to finance departments. Each role has different analytics needs, meaning that the data must also differ. In this document, we intend to review the various applications of data analytics in football.

The group formed by football coaches and staff needs to analyze the performance of the team as a whole, as well as the individual performance of each player, identifying team strengths and weaknesses. To this end, they should focus on the event and player analytics data. Also, it is necessary to understand the fitness of the team. For this purpose, they must collect and evaluate physical and mental health data to improve the player's performance. Other fields of application are football operations, financial, and marketing departments. What is necessary here is to increase the organization's efficiency, maximizing

the return on your investment at all levels of the club. This includes player level, attendance records, and marketing impact through forecasts and predictions. Furthermore, despite the controversy, gambling is a major player in football. The most relevant information in this field is historical data and trends, which are the fundamentals of betting. Lastly, there is an ever-growing need to supplement entertainment programs with data and their corresponding visualizations for easy understanding.

Keeping this in mind, the objective of the project is to make a systematic review of the application of data analysis and machine learning techniques applied to football. Furthermore, these techniques will be applied to a specific case study using free data provided by StatsBomb[23]. We decided to use the data of this provider because it is one of the market leaders and offers a large amount of free data with a high level of detail.

1.2 Project goals

The main objective of the project is to learn and apply the main data analytics and machine learning techniques used in professional football.

- Study of data analysis techniques and machine learning algorithms.
- Study of different data providers and its characteristics.
- Overall performance analysis of teams.
- Overall performance analysis of specific players.
- Comparison between teams from the same and different leagues.
- Comparison between players.
- Application of machine learning techniques.

1.3 Structure of this document

The remainder of this document is structured as follows:

Chapter 2 covers the research carried out in relation to other existing work in the field of data analytics and the application of machine learning techniques in the world of football.

Chapter 3 describes the different technologies and tools used for the development of this project, highlighting their main characteristics and the advantages of their use in this specific study.

Chapter 4 outlines the whole process carried out for the application of the machine learning algorithm. The process has four main phases, Problem Definition, Data Processing, Modeling, and Evaluation.

Chapter 5 shows a guide to use the interactive application developed in Streamlit for the visualization of metrics that serves as a complementary tool for coaches and analysts for data-driven decision making.

Chapter 6 concludes the project with the conclusions drawn from the work done, as well as an analysis of the results with the objectives of the project and discusses future possible work that can be done in this area.

CHAPTER 2

State of Art

In this section, it is proposed to review the available work in football analytics to obtain a global vision of the different applications in the fields mentioned previously.

2.1 Data

Good data is crucial to generate accurate and meaningful analytical insights. To ensure data quality, following best practices for data collection, storage, and analysis is important. In addition, it is crucial to have a clear understanding of the data being used, including its sources and its limitations.

There are several sources of data in football analytics, for example Statsbomb, Opta, Stats Perform, and WyScout, among others. According to the study conducted by Left Field [6], in which they reviewed how 27 professional football teams use data, football clubs acquire mainly two types of data: event data and tracking data.

Event data records every action during a game, such as passes, shots, interceptions, blocks, etc. By analyzing event data, analysts can identify patterns, trends, and correlations that can help coaches and players make more informed decisions. As a result of the study [6], it was concluded that all clubs, except one, surveyed purchased event data. Most clubs obtain event data for over 30 competitions, all purchasing data from Opta and/or

Statsbomb. The survey found that 22 out of 27 clubs purchase Statsbomb data, indicating its market leading position. As clubs mature in analytics, they develop customized and proprietary insights and value raw data from event providers more than pre-packaged insights delivered through platforms and front-ends.

Tracking data refers to the collection and analysis of player and ball movement data during a game and can provide a wealth of information about player performance, including distance covered, speed, acceleration, deceleration, and positioning. Consequently, it allows the creation of visualizations, such as heat maps and player trajectories, that can help coaches and players better understand their performance and identify areas for improvement. In this field, SkillCorner is one of the market leaders in delivering tracking data and derivatives of tracking data captured via computer vision technology. This technology uses advanced algorithms to track the movement of players and the ball during a game based on broadcast footage. Left Field Research [6] concluded that just over two-thirds of the respondents are SkillCorner customers. The number of clubs willing to pay for tracking data (instead of receiving them as part of league-wide deals) can indicate the proportion of clubs that possess the necessary data science resources to extract valuable insights from this type of data.

As mentioned above, Statsbomb is one of the most important companies in football data. It started as a football analysis blog in 2013 with the goal of establishing a hub for top-tier sport analysis on the Web. Over time, StatsBomb transformed into the premier destination for high-quality, data-driven content, ultimately evolving into a thriving global community of analysts. The business that now represents StatsBomb began as a consulting service to teams from leagues and competitions around the world. Initially, they bought the data from other data companies, but realized that it was necessary to get better data. For this reason, they partnered with Arqam FC, an analytics business that had developed a way to collect and analyze football matches in great detail, and StatsBomb data was born. They knew of the data limitations during that period. Consequently, they undertook the task of incorporating novel metrics to address these deficiencies. [23]

With the aim of promoting research and facilitating entry into the world of data analytics, StatsBomb offers free data from some competitions. The leagues for which it offers free data are:

- Messi Data Biography: all matches of Messi with Barcelona in LaLiga between 2004 and 2021.

- UEFA European Football Championship (male) 2020.
- UEFA European Football Championship (female) 2022.
- FIFA World Cup (male) 2018 and 2022.
- FIFA World Cupa (female) 2019
- Invencible season of Arsenal 2003-2004
- Some Champions League Finals between 2000 and 20019.
- FA Women's Super League (England) between 2018 and 2021.
- National Women's Soccer League (United States) 2018.
- Indian Super League 2021-2022.
- La Liga 2015-2016.
- Premier League 2015-2016.
- Bundesliga 2015-2016.
- Serie A 2015-2016.
- Ligue 1 2015-2016.

The complimentary data offering showcases the same comprehensive specifications that are offered in the paid packages. The data feed encompasses an array of essential aspects, including the positions of attacking and defending players in all shooting scenarios, the actions of the goalkeeper during the development of the play, and exhaustive data on all players applying pressure on the player with the ball during the defensive phase. Additionally, valuable information is offered on the particulars of each pass made during the game, such as the foot used to execute it, the height of the pass, and several other key variables that provide an added layer of depth to the data analysis.

StatsBomb offers an API that allows access to your data and a package (statsbombpy) that makes it easy to extract data from the API and use it in Python. For every competition, the API includes different types of data: competitions, matches, events, and lineups. Also, in some competitions they have more detailed data with StatsBomb 360.

The competitions.json file contains information about the competitions available in StatsBomb's database, such as the Premier League, La Liga, the Bundesliga, etc. Information includes the name of the competition, an identifier for each tournament, the country

where it is played, and the period of time that it covers. In the matches.json file, information about football matches is stored, such as the home team, the away team, the date and time of the match, the final score, and the competition it is played in. Each match has an identifier with it is possible to access the event data of that match. The events.json file contains more detailed data about the events in a match, such as goals, yellow and red cards, shots, passes, fouls, and interceptions. These data can be used to analyze player and team performance in different game situations. With the lineups.json file, it is possible to obtain information about the lineups of the teams in a match, including player names, position, and formation. These data can be used to analyze team strategy and player performance in different positions. Finally, StatsBomb also offers 360 data, which allows for a full view of the field and a better understanding of the game. Incorporates teammate and opposition locations into the events collected per match. Through the incorporation of this additional context into every event, along with exclusive new metrics, StatsBomb 360 facilitates more advanced performance and recruitment analysis.

In this project, the StatsBomb data are going to be used because it provides a large sample of free data from different teams and leagues, which allows an analysis of team performance and individual performance.

2.2 Team performance

One of the branches of data analytics in football is the measurement of team performance. To this end, Pratas, Volossovitch, and Carita [17] conducted a systematic review of goal scoring in elite male football leagues. Covers two main approaches used to recognize key performance indicators related to goal scoring, the static and dynamic approach. Although most studies were conducted using the static approach, the authors provide an argument for adopting a dynamic approach for a complete understanding of how goals are scored in football. The study emphasizes the importance of using data to formulate theories that explain how a team increases its chances of winning, and that the relevance of performance metrics varies depending on the game context. Different key performance indicators are identified such as pass accuracy, scoring efficiency, zones in which possessions started, playing style, and space-time coordination.

In the study “Prerequisites for Winning a League Using Data Analytics” [7] some metrics are introduced with the aim of measuring the performance of a team. The first is Passes Allowed Per Defensive Action (PPDA), which assists in ascertaining the extent to which a team applies pressure on the opposing team to regain possession of the ball in advanced

areas of the field, thereby increasing the likelihood of the team scoring a goal. This is the way to measure the pressing factor of a team; it is calculated by dividing the number of passes allowed by the defending team by the total number of defensive actions; both values will be calculated with reference to a specific area of the pitch. The higher the PPDA, the lower the pressing. The other metric used in the study is Expected Goals (xG), which measures the probability that a goal will be scored from a particular shot or scoring opportunity. The xG model takes into account various factors such as the location of the shot, the angle of the shot, the type of shot and other contextual information to assign a value between 0 and 1 that represents the probability that the shot will result in a goal. This metric provides a more objective and data-driven approach to evaluating the quality of scoring chances and the performance of players and teams in front of the goal. At the same time, Expected Goals Against (xGA) is a metric that measures the expected goals of the opponent. The combination of these metrics generate another metric called Expected Points (xPoints), which estimates the number of points a team is expected to earn from a particular match based on their performance. The xP model takes into account various factors such as the number and quality of chances created and conceded, the scoreline, and other contextual information to assign a value between 0 and 3 that represents the likelihood of a team winning, drawing, or losing the match.

The research carried out by Sarmento, Campanico and Marcelino [20] focuses on the relationship between physical activity patterns and the effectiveness of the game in football teams. According to the study, high intensity activity patterns contribute significantly to team performance. Compared to players on less successful teams, successful teams covered greater distances with the ball and performed more actions, such as passes, tackles, dribbling, and shots at the target. The study also found that at very high intensity running, players from successful teams had a higher average of goals and total shots on target. The study emphasizes the importance of physical activity patterns and the efficacy of the game to improve the performance of the football team.

2.3 Player individual analysis

It is possible to approach football analytics from another perspective, which deals with the players as a separate entity to analyze its performance.

Proposed by Caya and Bourdon [5], the technical skills of football players are closely related to data and performance. There are Business Intelligence and analytics techniques that can effectively capture individual athlete performance in their pursuit of positive out-

comes. In this field, Hughes M, Caudrelier T, James N, Redwood-Brown A, Donnelly I, Kirkbride A, Duschesne C [9], used a large number of performance analysts to discuss and define sets of performance indicators for each position in football. The positions used were:

- Goalkeepers.
- Full Backs.
- Centre Backs.
- Holding Midfield.
- Attacking Midfield.
- Wide Midfield.
- Strikers.

Each group of experts presented and discussed different key performance indicators. The authors of the study interpreted these results and generated the following table.

PERFORMANCE INDICATORS	GK	Full Backs	Centre Backs	HM	AM	WM	Strikers
Physiological	Height Strength Power Agility Coordination Reaction Time	Speed Power Stamina	Height Strength Speed Power Stamina	Stamina Speed Power Strength	Stamina Speed Power Strength	Speed Stamina Power Strength	Speed Agility Power Strength Stamina
Tactical	Vision Organisation Communication Distribution	Support play When to cross Passing Running off the ball Forcing offside	Vision Organisation Communication Passing	Vision Organisation Communication	Vision Organisation Communication	Vision Organisation Communication	Vision – awareness of space Anticipation Organisation Communication
Technical – Def	Shot stopping Coordination Recovery speed Save Punch	Tackle Pressing opposition Interception – anticipation Clearance Defensive header	Tackle Defensive header Pressing opposition Interception – anticipation Clearance	Tackle Pressing opposition Interception – anticipation Heading	Tackle Pressing opposition Interception – anticipation Heading	Tackle Pressing opposition Cover full-back Interception – anticipation Heading	Tackle Pressing opposition Interception – anticipation Heading
Technical – Att	Passing Throw Ball control with feet Kick Tackle	Tackle Interception – anticipation Dribbling Running with the ball Clearance Defensive header	Passing Heading Running with the ball Support play Dribbling Crossing Shooting	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Shooting Heading Reception Dribbling Passing Running with the ball Support play Crossing
Psychological	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language

GK – Goal Keepers; HM – Holding Midfield; AM – Attacking Midfield; WM – Wide Midfield.

Figure 2.1: Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position [9]

The authors extracted some conclusions from the study. Seven general classifications of positions in a football team, and their corresponding key performance indicators, were defined for each of these classifications within five categories. Most of the indicators are similar among the different positions; only the order of priority varies, with the exception of the goalkeeper. This enabled generic sets of skills required for outfield players in football. In addition, emphasis was also placed on the importance of finding a way to rate the execution of these skills in further research.

Another problem encountered when measuring a player's performance is the importance of each player in a build-up play. For this purpose, Karun Singh introduces a new concept called Expected Threat (xT) [21]. This framework is based on four main concepts:

- Reward individual player actions. The model evaluates passes and dribbles on the basis of how much it contributed to the build-up play.
- Event-level data. Framework does not use any player tracking data, only have a list of events with attributes such as player in possession, time elapsed in the match, start and end location, etc.
- Independence from and result of possession. Actions will be evaluated in isolation, without regard to what happened before and after possession.
- Recognize 'threatening' positions. When assessing pitch locations, we should consider more than just xG. The expected goal model assumes a direct shot, but some areas allow for easier ball movement into higher xG zones. To accurately gauge the threat, we must account for the potential of multiple actions, as xG only focuses on one (shooting) from the current position.

The article presents a simplified football buildup play model, categorizing player actions into shooting, passing, and dribbling until they lose possession or score. Using the 2017-2018 Premier League data, the author analyzes the behaviors of players in different pitch zones, identifying the probabilities of move and shoot, the transition matrix of move and the probabilities of goals. This model resembles a Markov model, treating pitch locations as states with transitions. While it currently focuses on successful moves, it could potentially include attempted moves, though this would add complexity. The 16x12 grid is used, resulting in 192 zones, assigning a value (V_x,y) to each zone. When the ball is in a zone (x,y), a player can shoot (with an expected payoff, g_x,y) or move the ball to one of the 192 possible zones. To calculate the expected return for each choice, a move transition matrix (T_x, y) is used based on historical data. The xT metric combines weighted pay-offs

for shooting and moving, offering a quantitative way to evaluate pitch locations' threat potential, beyond immediate shooting, for tactical analysis and decision-making.

Continuing with the topic, the investigation by Spearman and Basye [22] presents a model for ball control in football based on two key concepts: time-to-intercept and time-to-control. Using this physics-based approach to modeling pass probabilities at the per player level, athletes can gain a competitive advantage in understanding their performance and improving their field skills.

According to the study by Pappalardo, Cintia, Pedreschi, Giannotti, and Barabási [15], there is also another sample of events generated by a player which would be considered in the evaluation of technical skills. This study is based on the individual player rating by three journalists according to the personal interpretation of the player performance. This research outlines the limitations of human evaluation in considering different characteristics for the development of the evaluation process. The figure shows graphs for goalkeepers, defenders, midfielders, and forwards showing the significance of each attribute, normalized between 0 and 1, in the human rating process for typical football positions. The plots demonstrate that, with the aid of machine learning models, many of the features have minimal impact on the evaluation process carried out by human judges.

The research of Aquino, Alves, Fuini and Garganta [2] reveal several deficiencies in the current approach to measuring skill-related variables. First, there is a notable absence of clear definitions and classifications of these variables. Additionally, two additional limitations were identified: the lack of contextualization in the sample and the disregard for situational variables of the match, such as location, quality of opponent and status. Furthermore, a representative task design is lacking to accurately assess skill-related performance. According to the literature, a set of criteria are proposed to assess the ability of football players to understand the relationship between skill acquisition and player development for sporting excellence.

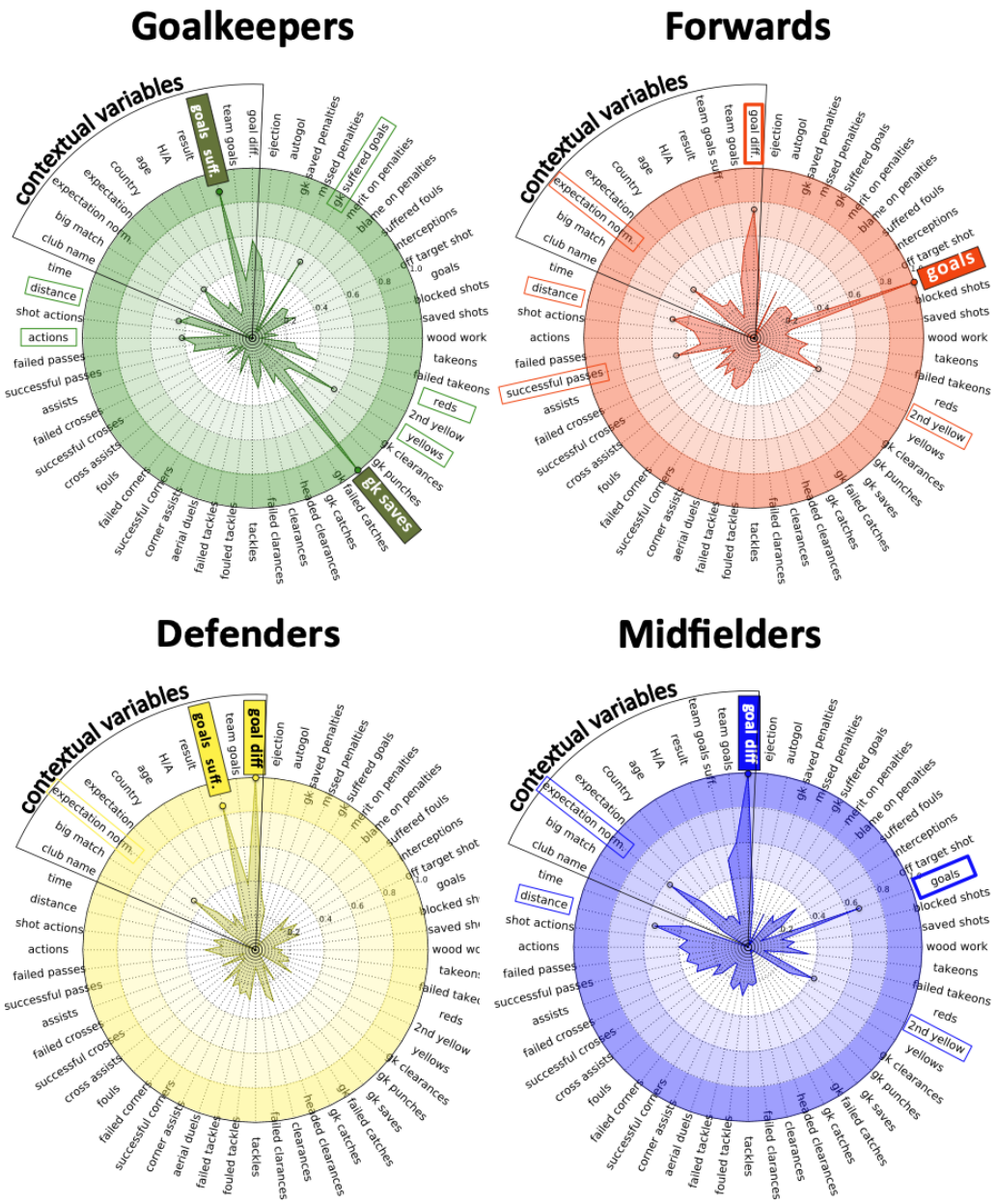


Figure 2.2: Human Perception of Performance [15]

2.4 Other fields

In addition, another relevant and controversial area in football is the sports betting industry. Predicting the outcomes of football matches and tournaments has become a significant area of interest, and researchers are continuously developing models that use different methodologies for forecasting. These models are constructed using advanced computing and machine learning techniques, which make the sports betting industry highly competitive.

An example of research in this field is [4], which illustrates a forecasting system to make profits in the sports betting market using Machine Learning (ML) techniques. They compare different ML techniques: Logistic regression, K-Nearest Neighbors, Support Vector Machine, Nave Bayes, and Random Forest, as well as a four-layer Artificial Neural Network (ANN).

The system works in two phases. In the first phase, they extract betting odds using web scraping techniques. The metrics used are 1x2, over-under 2.5, and goal-no goal. In the second phase, ML techniques are used to make predictions using the collected data as input. After a comparative analysis between techniques, it can be concluded that ANN is the most suitable option. They introduced the Return-on-Investment (ROI) metric to evaluate the probability that an investment will perform well. They evaluated the system by observing the ROI trend while varying the threshold, which was established based on the output of the neural network.

The system was tested in matches that are yet to be played. Of 89 matches, the system obtained 65 positive outcomes, 13 negative outcomes, and 9 considered it better not to bet. Additionally, the ROI was calculated; with 10 credits on each game, the system reached a gain of 999.7 credits out of 790 played with an overall ROI of 26 54%.

CHAPTER 3

Enabling Technologies

This chapter offers a brief review of the main technologies that have made possible this project, as well as some of the related published works.

3.1 Programming and Development Environment Technologies

3.1.1 Python

Python [18] is a versatile high-level programming language known for its simplicity, readability, and a wide ecosystem of libraries and frameworks. Its adoption in multiple fields has grown exponentially in recent years, such as data science and machine learning.

This language comes with a vast standard library that covers a wide range of functionalities, from file handling to web development and data manipulation. This extensive library reduces the need for developers to reinvent the wheel, as they can leverage pre-built modules and functions to streamline their projects. The application provides a thriving ecosystem of third-party packages and frameworks that cater to almost every domain and industry. For data analysis, libraries like Pandas and NumPy excel, while web development is made easier with frameworks like Django and Flask. Machine learning enthusiasts can explore TensorFlow, PyTorch, and Scikit-Learn. The availability of these packages accelerates development and empowers developers to tackle complex problems efficiently. In professional football, where large amounts of data are generated from matches, training sessions, and player performance, Python's data-handling capabilities have proven indispensable. Analysts can quickly load, clean, and analyze complex datasets, enabling them to make data-driven decisions effectively.

Python has a vibrant and supportive community of developers, analysts, and data scientists. This community-driven approach has led to the creation of extensive documentation, tutorials, and open source projects specifically tailored for sports analytics. As a result, professionals in the field can easily access resources to improve their skills and stay up-to-date with the latest developments.

In summary, Python's versatility, readability, and extensive ecosystem of libraries and frameworks have solidified its position as one of the most popular programming languages for a wide range of applications, from web development to data science and beyond. Its community-driven development, cross-platform compatibility, and simplicity continue to attract developers and organizations looking to harness the power of technology for various purposes. As Python continues to evolve, its influence on the world of technology is expected to expand even further.

3.1.2 Streamlit

Streamlit [26] is a powerful and user-friendly open source Python library that has gained significant traction in the development of data projects. It is designed to simplify the process of creating interactive and data-driven web applications.

The first of the points that can be highlighted is its simplicity and intuitive syntax. Streamlit excels at rapid prototyping, allowing developers and data scientists to quickly turn data scripts into interactive web applications, including those with minimal web development experience. Second, note that this library provides a wide range of widgets and components that enable developers to create interactive data visualizations, such as charts, sliders, buttons, and text input, among others. All this with a high degree of customization; users can apply themes, styles, and layouts to match their branding or design preferences. Another important point is that Streamlit seamlessly integrates with popular data analysis libraries such as Pandas, NumPy, and Matplotlib. This makes it an excellent choice to turn data analysis scripts into interactive dashboards and reports that can be shared. Finally, it is worth mentioning that this library makes web application deployment a breeze. With a single command, developers can share their applications with colleagues or make them accessible to a broader audience on the Web. Streamlit Sharing, Streamlit's own hosting platform, further simplifies the deployment process.

In summary, Streamlit's simplicity and interactive capabilities have made it an essential tool for data professionals, developers, and anyone looking to create web applications with data at their core. Its rapid prototyping capabilities, ease of deployment, and active community support have solidified its position as a versatile library for a wide range of applications, including data visualization, reporting, machine learning, and data storytelling.

3.2 Data Modeling Technologies

3.2.1 Pandas

Pandas [14] is an open-source Python library designed to facilitate data manipulation and analysis. It has become an indispensable tool for data scientists, analysts, and researchers worldwide.

This library introduces two primary data structures, Series and DataFrame. A series is an object similar to an array that can contain a variety of data types, such as numbers, strings, and so on. On the other hand, a DataFrame is a two-dimensional object with

columns and rows that provides powerful data manipulation capabilities, including filtering, merging, grouping, and pivot operations. In addition, Pandas offers a wide array of functions for data cleaning, transformation, and exploration. It can be used to manage missing values, facilitate data type conversions, and integrate datasets. These operations are crucial to ensuring the precision and trustworthiness of the data used in a project. All of this combined with the capacity to integrate Pandas with other libraries commonly used in data science, including matplotlib, seaborn, scikit-learn, etc.

Due to all of these factors, Pandas has become an indispensable tool for data analysis and manipulation in the Python ecosystem. Its intuitive data structures, versatile data handling capabilities, and seamless integration with other libraries make it a go-to choice for data scientists and analysts. Whether dealing with small datasets or big data, Pandas empowers users to efficiently explore, transform, and analyze data, making it an essential part of the data science toolkit.

3.2.2 Numpy

Numpy [13] is a foundational library in the Python programming ecosystem. It was created to address the limitations of Python for numerical and mathematical computations. NumPy provides a versatile and efficient framework to work with arrays and matrices, making it an indispensable tool for scientists, engineers, and data analysts.

At the core of NumPy is the ndarray (n-dimensional array), a flexible and efficient data structure. This array allows the representation of vectors, matrices, and higher-dimensional data with ease. In addition, it offers a rich set of mathematical functions and operations for array manipulation, including element-wise operations, allowing for efficient and concise execution of complex mathematical operations across entire arrays; and essential aggregation functions, including mean, median, and standard deviation, crucial tools for comprehensive data analysis. All of this combined with the capacity to integrate Pandas with other libraries commonly used in data science, including matplotlib, seaborn, scikit-learn, etc.

In conclusion, NumPy stands as the foundational cornerstone of numerical computation within the Python ecosystem, furnishing an indispensable arsenal for scientific investigation, meticulous data analysis, and rigorous engineering pursuits. Distinctive attributes, such as its versatile multidimensional arrays, extensive suite of mathematical functions, adept broadcasting capabilities, and seamless harmonization with auxiliary libraries, make

NumPy an elemental preference for individuals engaged in the manipulation and analysis of numerical datasets.

3.2.3 Scikit-Learn

Scikit-Learn [16] is an open-source Python machine learning library that has become a staple in the data science community. It provides a wide range of supervised and unsupervised learning algorithms through a consistent interface. Built on top of other prominent libraries like NumPy, SciPy, and matplotlib, scikit-learn is designed for efficient and straightforward implementation of complex machine learning tasks. From classification and regression to clustering and dimensionality reduction, scikit-learn's well-organized API allows both novices and experts to deploy models with ease.

The library's strengths lie in its extensive documentation, active community, and rich ecosystem of tools. Scikit-learn comes with numerous built-in datasets for experimentation, and it integrates seamlessly with the Python scientific stack, enhancing its functionality and scope. Its commitment to best practices and the constant addition of cutting-edge algorithms ensure that users are always at the forefront of machine learning technology. Whether for academic research or commercial applications, scikit-learn remains a premier choice for developing predictive models and analyzing data.

3.2.4 StatsBomb API

The StatsBomb API, a prominent data service in the domain of sports analytics, represents a key resource for professionals, researchers, and fans immersed in the realm of football. Statsbomb offers two types of packages to access its API, depending on the language you use, R or Python.

StatsBomb promotes and facilitates new research and analysis at all levels of football analytics. For this purpose, StatsBomb has chosen to provide access to select leagues of their data, free of charge, to the public. This offering is intended exclusively for research efforts and people who are genuinely passionate about football analytics [24]. The data are furnished in the form of JSON files sourced from the StatsBomb Data API, organized into the following structure:

- Information pertaining to competitions and seasons is contained within the "competitions.json" file.

- Matches associated with each competition and season are stored in the "matches" directory. Within this directory, individual folders are named according to competition IDs, while each file is denoted by a season ID within the respective competition.
- The events and lineups data for each match can be found in the "events" and "lineups" directories, respectively. In both cases, files are named on the basis of their corresponding match ID.
- For selected matches, the StatsBomb 360 data is available and is housed within the "three-sixty" directory. Each file in this directory is named after the match ID.

3.3 Data Visualization Technologies

3.3.1 Matplotlib

Matplotlib [11] is an open source data visualization library for the Python programming language. It stands as an indispensable tool in the world of scientific computing and data analysis. Its longevity and established reputation as a reliable graphing library have solidified its position as the de facto choice for creating static, animated, and interactive visualizations.

It is characterized by its versatility, offering a diverse range of plotting functions and customization options, which makes it adaptable to a wide range of data visualization requirements. From basic line and scatter plots to complex 3D visualizations, Matplotlib caters to a multitude of use cases. The library's extensive configurability enables users to tailor every aspect of a plot, from axis labels and tick marks to legend placement and annotations. This level of customization empowers researchers to convey complex data insights with precision. In addition, it is worth mentioning that Matplotlib seamlessly integrates with various data analysis libraries and frameworks, including NumPy and Pandas. This interoperability ensures a streamlined workflow for data scientists, allowing for easy data manipulation and visualization.

3.3.2 Mplsoccer

Mplsoccer is an open source Python library for plotting football charts in Matplotlib and loading open data from StatsBomb [12]. In contrast to generic data visualization libraries, Mplsoccer is specifically designed for the visualization of football data. This tool offers a specialized set of tools and functionalities customized to the specific requirements of football analytics, which will be of great assistance in the completion of this project.

A distinguishing feature of Mplsoccer is its ability to plot player positions, movements, and events on the football field. Analysts can visualize player heatmaps, pass networks, and shot maps with ease. Additionally, using Mplsoccer, statistical data can be integrated into visualizations that allow for a more complete report. Through the integration of player statistics into dynamic graphics of the display of match statistics on the pitch, the library enhances the depth of football analysis.

CHAPTER 4

Methodology

This chapter presents the methodology used in this work.

This chapter represents a comprehensive guide to the methodology used in the development of this project. In works similar to this one, there are different stages that go through problem definition, data processing, modelling, and evaluation.

Figure 4.1: Phases of a data project

4.1 Problem definition

The initial step in this project is to identify and articulate a problem that needs to be addressed. A well-defined problem serves as the cornerstone for effective problem-solving, providing a clear direction to tackle the challenge at hand.

As mentioned previously, the main objective of this project is to perform a comprehensive data analysis of events that occur during football matches. The aim is to perform an extensive performance assessment that includes both collective team dynamics and individual player contributions. This endeavour will involve the meticulous calculation of various statistical metrics, facilitating the acquisition of valuable insights into the intricate world of football. Through the use of data analysis techniques, patterns, trends, and concealed relationships within match events can be revealed. Furthermore, a diverse range of machine learning models is intended for deployment to further advance the analysis. These models will not only contribute to an enhanced understanding of the game but will also enable the generation of informed predictions, shedding light on potential future match outcomes.

4.1.1 Team performance

In order to simplify the analysis, the game will be broken down into three stages: defensive actions, building actions, and finishing actions. The goal is to perform both quantitative and qualitative analysis of the data in order to see trends and try to draw conclusions. Each phase encapsulates a unique aspect of the game, offering valuable insights into the intricacies of football strategy and team dynamics.

4.1.1.1 Finishing actions

This phase centers on the critical moments when teams approach the opponent's goal in an attempt to score. Finishing action analysis reveal the effectiveness of attacking strategies, the precision of goal-scoring opportunities. For this purpose, several statistics have been selected and will be emphasized in order to carry out the study from a quantitative standpoint:

- Total number of shots.
- Number of shots on target.
- Number of shots off target.
- Number of shots blocked by opponents.
- Number of goals.
- Shot rate per goal.

In order to measure the quality of goal-scoring opportunities, the Expected Goals (xG) metric is introduced. In this project, the xG offered by StatsBomb [25] is used. This metric is designed to assess the probability that a shot will end in a goal. The model leverages historical data from numerous shots that share similar attributes to predict the probability of a goal, typically quantified on a scale ranging from 0 to 1. For example, a shot with an xG value of 0.2 is expected to be converted twice in every 10 attempts. Although each xG model may exhibit unique features, several fundamental factors have traditionally formed the core inputs for the vast majority of Expected Goals models. These factors include the shot's distance from the goal, the angle at which the shot is taken, the specific body part used for the shot, and the nature of the assist or preceding action (such as throughball, cross, set-piece, dribble, etc.). The StatsBomb xG model incorporates additional crucial details such as the goalkeeper's positioning and status, the locations of all attackers and defenders within the frame, and the height at which the shot makes impact. This enhanced model provides a more precise and comprehensive assessment of the quality of the scoring opportunity.

Expected Goals models hold significance due to their exceptional accuracy in predicting future team and player performance. When evaluating team performance, the xG models surpass current goal difference and basic shot count metrics such as the total shot ratio (TSR) in terms of predictiveness. These models can help to gain a deeper understanding of the underlying quality of a team, transcending their current results. This allows for

the identification of teams that may be either overachieving or underperforming relative to their expected statistics, potentially signaling an impending shift in their results. Naturally, there exists a certain degree of residual variance between actual goals and Expected Goals over specific time frames. This variance arises from the binary outcome of shots, which can result in either a goal or no goal, in contrast to the xG values that span a probabilistic scale from 0 to 1. An independent study conducted by Lars Maurath indicates that, depending on the quality of the model, we can anticipate that approximately 79% to 93% of the team seasons would align the goals with xG within a 95% confidence interval.

4.1.1.2 Building actions

This phase refers to the actions and processes involved in constructing and advancing attacks, particularly from the team's defensive or midfield positions. These actions encompass a wide range of activities that contribute to the team's offense strategy and ball progression. The statistics selected to make a quantitative study are as follows.

- Total number of passes.
- Number of short passes.
- Number of long passes.
- Number of passes into final third of the pitch.
- Number of crosses.
- Mean distance of passes.

Furthermore, to improve the understanding of player interactions and relationships on the field, passing networks will be used. In particular, the type of passing networks used will be player passing networks, where nodes are players and links are weighted by the number of passes between them. The position of the players represented in these nets is calculated as the mean position from which the players have made passes. To this end, the passing networks will be tested using the code obtained from the *Friends of Tracking* repository [10].

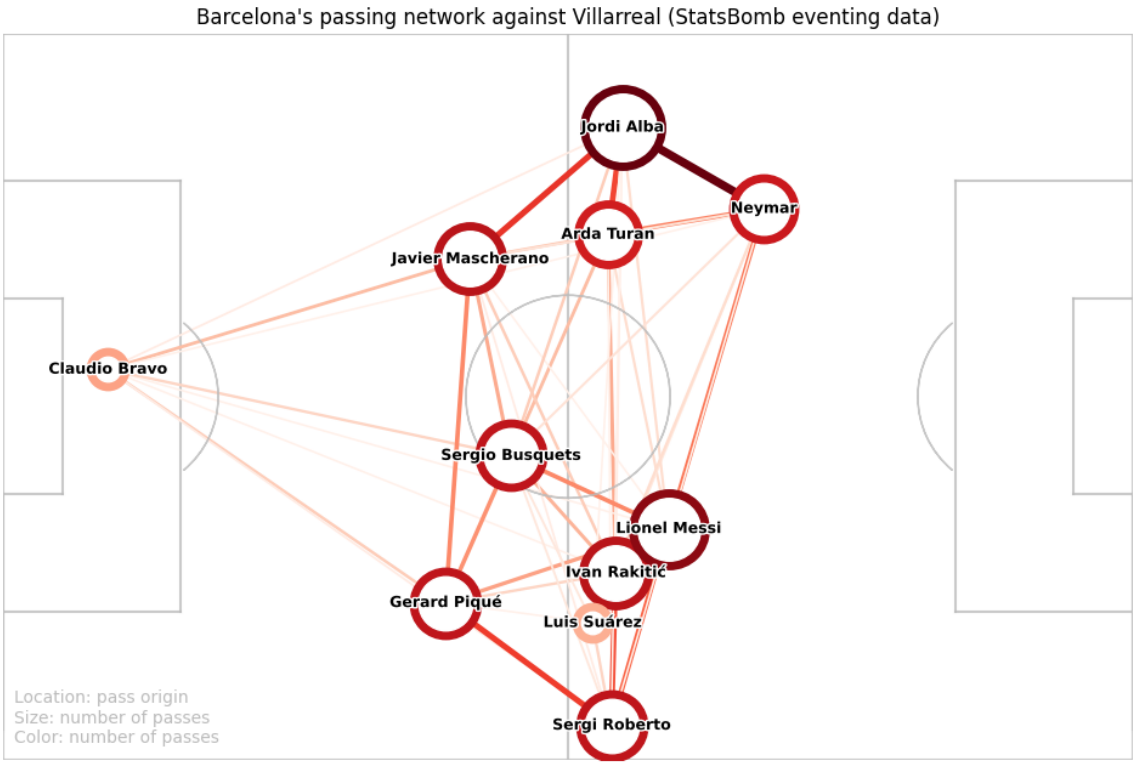


Figure 4.2: Example Player Passing Network

In addition, the centralization index will be calculated for each player's football passing network. This index is used to measure the centralization of the passing. A higher centralization index indicates that passing is more centralized around a few key players, while a lower index suggests a more distributed passing pattern where more players are involved in passing the ball. The method used to calculate this index is the one used by *David Sumpter* in his Soccermatics course [27].

1. Calculate the number of successful passes by each player.

2. Find the player who made the most passes.

3. Calculate the denominator: is calculated as ten times the total sum of passes made by all players. This represents the total possible passes in the system if passing was equally distributed among players.

4. Calculate the numerator: Calculated as the sum of the differences between the maximum number of successful passes made by one player and the number of successful passes made by each player. This represents the concentration of passes around a few key players.

5. Calculate the centralization index as the division of the numerator by the denominator previously calculated.

The centralization index must be a value between 0 and 1. A value of 0 indicates perfect decentralization, which means that the passes are distributed equally among all players in the network. On the other hand, a value of 1 indicates perfect centralization, which means that all passes are performed in a single player. This index provides insight into the passing patterns and strategies of a football team.

4.1.1.3 Defensive actions

The defensive phase of a football game refers to the period in which a team is primarily focused on preventing the opposing team from scoring a goal. During the defensive phase, the main objectives are to protect the own goal, regain possession of the ball, and disrupt the opponent's attacking plays. Some of the statistics that will be used to analyze this phase are as follows.

- Number of recoveries in different zones of the pitch.
- Number of committed fouls.
- Number of duels won.
- Number of shots from opponents.
- Number of clean sheets.
- Expected goals conceded (xGA).
- Goals conceded.

In addition, to analyze the pressure made by a team, the PPDA metric is introduced. PPDA (Passes Allowed per Defensive Action) [29] was defined by Colin Trainor on the StatsBomb blog with an article called "Defensive Metrics: Measurement of the intensity of a high press". It is a metric that can quantify the extent and aggression of high presses used by teams, both during a season and in any specific match. Calculated by dividing the number of passes allowed by the defending team by the total number of defensive actions; both values will be calculated with reference to a specific area of the pitch. The author underscores the importance of conducting analytical studies of this nature, highlighting the advantages they offer coaches, teams, and enthusiasts. With the establishment of a measurable metric, coaches can quickly gauge the potency of the high-striking tactics employed by

their adversaries in recent matches, allowing them to formulate suitable counterstrategies. Furthermore, both teams and fans can assess the level of pressure exerted by their own team on the opposition, especially in deep defensive positions. This evaluation improves the understanding of the team's defensive efficiency and serves as a foundation for analyzing defensive performance over time. In short, this metric mitigates the subjective nature of describing the pressing intensity by adding it to a single numerical value. This simplifies the process of comparing, analyzing and ranking the pressing strategies of different teams, thereby fostering a more comprehensive understanding of defensive dynamics. Coaches, analysts, and fans can use this quantifiable measure to gain insight into team performance, make well-founded decisions, and refine their strategic planning.

The defensive actions used by the author are tackles, interceptions, challenges (failed tackles), and fouls. He remarks the importance of measuring the pressure that the defending team puts on the opposition players when they are in possession of the ball and not about whether the opposition team made or completed passes. Another key point is the decision to select the specific area of the pitch to measure the intensity of pressing. The author selects the area of the pitch with an x-coordinate greater than 40. As can be seen in the image 4.3, the perspective is based on the team playing from left to right. Defensive actions, which include tackles, interceptions, challenges, and fouls, are focused specifically on events that occur to the right of the reference line, denoted by "x = 40." This encompasses defensive actions in both the attacking half of the team and a small portion of their own defensive half.

The author has chosen the value of "x = 40" as the boundary for the pressing metric, primarily due to the consideration that some teams may strategically concede possession in their opponents' half but then apply intense pressure once the ball crosses the halfway line. Setting "x = 40" ensures that such teams receive recognition for their pressure as soon as the ball enters their own half. Moreover, the choice strikes a balance by excluding actions in the defensive third of the pitch (e.g., "x = 33"), where tackles and challenges are commonly expected from all teams. To validate this boundary selection, a correlation analysis was conducted, comparing pressing metric values (PPDA) at different potential boundaries (x = 33, 40, and 50). The findings revealed a strong correlation, with "x = 40" explaining more than 90% of the variation in the metric at the other potential boundaries. Therefore, the decision to adopt "x = 40" as the boundary is in line with practical football considerations and effectively captures the intensity of defensive actions, regardless of the specific cutoff value, underscoring the importance of choosing a suitable boundary for the pressing metric to measure defensive pressure effectively.

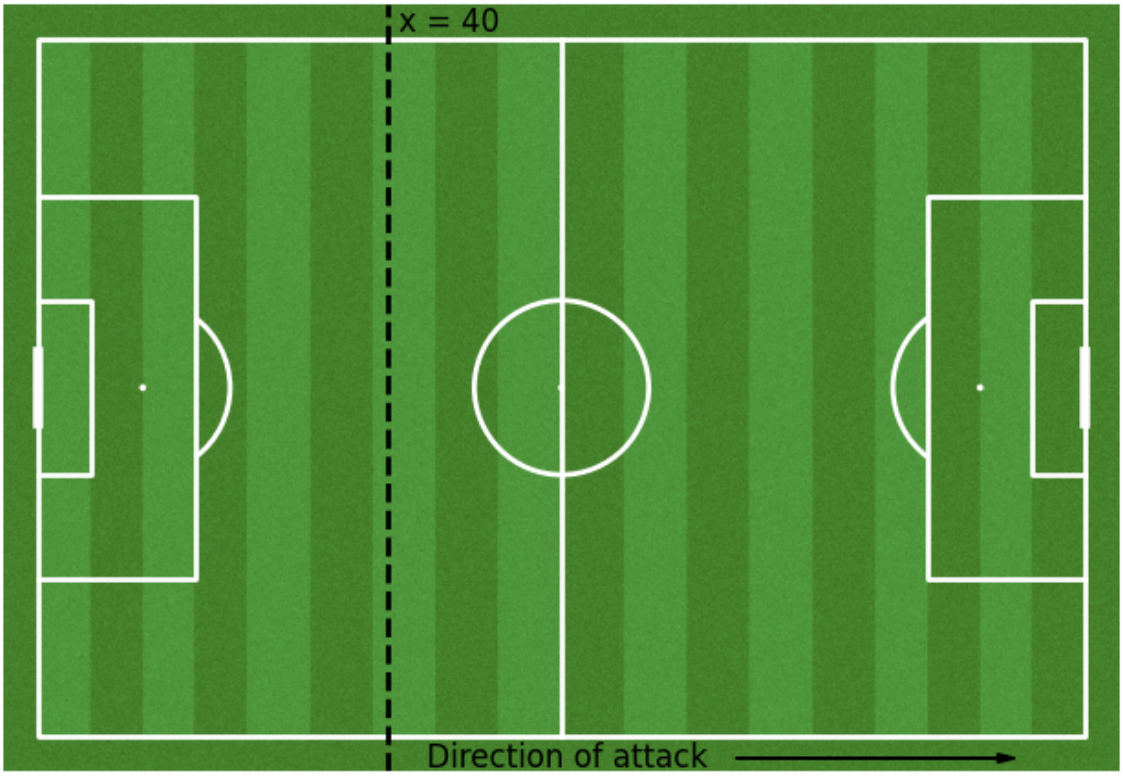


Figure 4.3: Zone of the pitch where it is calculated PPDA

In short, PPDA is a metric that allows us to quantify the pressure exerted by a team in a specific area of the field. A reduced PPDA value signifies that the defending team more actively interferes with the opposition's passing sequences, signifying a heightened defensive pressure level. On the contrary, a higher PPDA value signifies a decreased defensive intensity, as the defending team allows a greater share of uncontested passes.

4.1.2 Player performance

With the aim of analyzing player performance, it is necessary to divide players into groups according to their position. Four main groups will be distinguished: goalkeepers, defenders, midfielders, and forwards. It is true that within each of these groups there are different positions and roles, but this differentiation will be valid for this analysis. The main objective is to define some statistics and metrics that allow to measure the player performance and compare players of the same group.

As *David Sumpter* mentioned in the section of scouting of his Soccematics course [28], there are two major problems with the player statistics. On the one hand, the context of an action. For example, only with the number of completing passes is not known if the player was completing simple passes between defenders in the backline or more challenging

attacking passes. To address this, an effort will be made to provide context to the numbers under analysis. This may involve incorporating visualizations or examining statistics in specific scenarios to enhance understanding of the action. Furthermore, the comparison of different players within the same league will be explored. One proposed solution is the use of percentiles, where players are ranked according to their scores and each player is assigned a percentile score, ranging from the highest (100%) to the lowest (0%). When performing this ranking, it is essential to compare them with players in a similar position within the same league. In addition, the author proposes the normalization of the statistics by the minutes played by each player.

4.1.2.1 Goalkeepers

The statistics used to measure the goalkeeper's performance are as follows:

- Number of goals conceded.
- Number of shots on target received.
- Number of saves.
- Save percentage = shots on target - goal against/ shot on target.
- The distribution of saves in relation to the part of the body with which they are performed.
- Number of penalty kicks saved.
- Number of passes. Distinguish between short and long passes.

4.1.2.2 Defenders

To measure the performance of the defenders, the following statistics are used:

- Number of ball recoveries.
- Number of duels won.
- Number of blocks.
- Number of interceptions.
- Number of clearances.
- Number of committed fouls.

4.1.2.3 Midfielders

The following statistics are used to assess the midfielder's performance:

- Number of ball recoveries.
- Number of duels won.
- Number of interceptions.
- Number of committed fouls.
- Number of passes.
- Number of passes that end in the final third.
- Number of passes that precede a shot.
- Number of assists.
- Number of goals.
- Number of dribbles.

4.1.2.4 Forwards

The forward's performance is evaluated using the following statistics:

- Number of passes that end in final third.
- Number of passes that precede a shot.
- Number of assists.
- Number of shots.
- Number of shots on target.
- Shot rate per goal.
- Expected goals (xG) generated.
- Number of goals.
- Number of dribbles.

In order to evaluate the actions performed by a player, the expected threat metric (xT) will be introduced. The origin of this metric is Sarah Rudd's work 'framework for tactical analysis and assessment of individual offensive production in soccer using Markov chains' [21]. A Markov chain is a mathematical model used to describe a sequence of events where the outcome of each event depends only on the current state and not on the previous one. This assumption does not always hold in football, but it is a good start. The name expected threat (xT) was introduced by Karun Singh in a blog post in 2018 [21] where he tried to create a method to assign the credit of each player involved in a build-up play. His method is based on four main concepts:

- Reward individual player actions. The model evaluates passes and dribbles on the basis of how much it contributed to the build-up play.
- Event-level data. Framework does not use any player tracking data, only have a list of events with attributes such as player in possession, time elapsed in the match, start and end location, etc.
- Independence from and result of possession. Actions will be evaluated in isolation, without regard to what happened before and after possession.
- Recognize 'threatening' positions. When assessing pitch locations, we should consider more than just xG. The expected goal model assumes a direct shot, but some areas allow for easier ball movement into higher xG zones. To accurately gauge the threat, we must account for the potential of multiple actions, as xG only focuses on one (shooting) from the current position.

This metric assesses the potential threat to the opponent of the pitch locations. The author uses a 16x12 grid system on the field, dividing it into 192 zones, with each zone represented by the value $V_{x,y}$ derived from their algorithm. When a player is in zone (x,y), they face a choice between shooting or passing the ball. Opting for a shot yields an expected payoff of $g_{x,y}$, determined by the probability of scoring from that zone based on historical data. In contrast, choosing to pass provides 192 potential zones for ball movement, each with its associated expected payoff denoted as $V_{z,w}$. To calculate the expected payoffs for all 192 passing options, the move transition matrix $T_{x,y}$, informed by past data, is used to indicate the likely ball movement of the player. The reward for passing the ball to zone (z,w) becomes $T_{x,y}\rightarrow$(z,w) $\times$ $V_{z,w}$, signifying the probability of moving to that zone multiplied by the reward expected from that zone. To calculate the expected overall payoff for ball movement, the author aggregates this value across all 192 potential zones. The ultimate metric for zone (x,y) is derived by incorporating both the payoff for shooting and the payoff for ball movement, taking into account their respective probabilities: $s_{x,y}$ represents the

likelihood of shooting from zone (x,y) based on historical data, while $m_{x,y}$ represents the probability of passing the ball. Subsequently, xT is defined as the sum of the weighted payoffs for shooting and passing the ball.

$$\mathbf{xT}_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{z=1}^{16}\sum_{w=1}^{12} T_{(x,y)\to(z,w)}\mathbf{xT}_{z,w})$$

Figure 4.4: Expected Threat (xT) formula

The formula initially contains a cyclic dependency problem because it requires knowing the xT values for all zones before calculating the xT for any specific zone. The solution to this issue is an iterative approach. Starting with xT = 0 for all zones, the formula is evaluated iteratively until convergence, typically within 4-5 iterations. This process not only resolves the cyclic dependency but also enhances interpretability. In the first iteration, it essentially approximates an expected goal (xG) model, considering only shooting. The second iteration introduces the possibility of "move, then shoot". Extending this rationale to further iterations, in the third iteration, scenarios such as "move, move, shoot" are taken into consideration, looking up to two moves ahead of the checkmate. This concept significantly enhances the interpretability of xT. Instead of exist as a score on an arbitrary scale, it assumes a highly intuitive meaning, akin to expected goals. $xT_{x,y}$ at iteration n represents the probability of scoring within the next n actions.

In short, expected threat is a metric that allows one to quantify the performance of a player in the construction of a play and the impact that the creation of danger has on the creation of scoring chances.

4.1.3 Machine Learning Algorithms in Football

With all the data processed and analyzed in the previous sections, machine learning algorithms will be used with the objective of clustering players and teams.

In order to complement the performance study, both for players and teams, K-Means will be used with the objective of identifying patterns and making predictions. The article written by *Rio RizKi Aryanto*, *"Clustering NBA Player using K-Means"* [3], has served as inspiration. This article is a great example of using K-Means for clustering in the realm of professional sports, from which insights can be obtained into adpating the methodology for clustering both football teams and players. On the other hand, another article that has

served as a reference has been *"K-means for player clustering"* [1] by *Ricardo André*, in which he uses the algorithm to cluster football players.

K-Means is an unsupervised learning algorithm, which means it doesn't require a training dataset with known outcomes. Instead, it operates on the inherent structure of the data to categorize them into distinct groups based on specified features. In the context of this work, these characteristics could encompass a myriad of metrics such as total shots taken, accuracy, goal scoring ability, defensive skills, and so forth. By clustering players or teams with similar attributes, K-Means helps to understand the underlying patterns that define performance levels in professional football.

The algorithm works by specifying a predefined number of clusters (K), where it iteratively assigns each data point to the nearest cluster centroid and then recalculates the centroids based on the newly assigned points. This process continues until there is no change in cluster assignments or the change is below a specified threshold, thus ensuring convergence to a solution that minimizes the variance within the cluster.

Using K-Means to group players and teams in professional football can provide a granular understanding of different styles and strategies. It can help isolate groups of players who exhibit similar playing characteristics, or teams that adhere to comparable tactical doctrines. Additionally, the insights derived from this clustering can be crucial for opponent analysis, talent scouting, and performance optimization, providing a substantial analytical foundation for decision making both in and out of the field.

To carry out this study, this algorithm will be used in each of the phases of the game defined to measure the team's performance (finishing actions, building actions, and defensive actions); and in the case of the grouping of players, the classification previously made will be used (goalkeepers, defenders, midfielders, and forwards), with their respective statistics and metrics.

4.2 Data

4.2.1 Data Structure

The data provided by StatsBomb cover a wide range of events that occur during a match, from basic actions like passes and shots to more intricate details like player positions, tactical shifts, and specific event outcomes. To access these data, as mentioned above, StatsBomb

has its own API [24] that facilitates its use. In addition, all the data are correctly structured; this makes it much easier for the final user. In the StatsBomb repository, there are several documents explaining in detail each type of data set available for querying through the API, however, in this section we will provide a detailed description of those that have been most relevant for this study.

The aim of this project is to analyze different teams and players. For this purpose, open data from different leagues will be used corresponding to the 2015/16 season. The leagues used are: Premier League, Bundesliga, Serie A, and La Liga. First of all, it is necessary to check the competitions available in the StatsBomb API. With the function *competitions()* it is possible to consult the available leagues with open data. The competitions used and its identifiers are shown in the next table.

competition_id	season_id	country_name	competition_name
9	27	Germany	1. Bundesliga
11	27	Spain	La Liga
2	27	England	Premier League
12	27	Italy	Serie A

Table 4.1: Competitions and its respective identifiers

Another function is *matches()*, in which, specifying the season and the competition identifier, it returns a data set in which each row will have information related to the different matches played in that league in that specific season. This data set consists of different attributes which include:

Variable	Variable Type	Description
match_id	Integer	The unique identifier for the match
match_date	Date	The date of the match
kick_off	Time	The time of the match
home_team	String	The name of the home team in this match
away_team	String	The name of the away team in this match
home_score	Integer	The final score of the home team
away_score	Integer	The final score of the away team
match_week	Integer	Week number of the match in the competition

Table 4.2: Matches data set

To carry out the study proposed in this work, the events for each competition that will be analyzed are necessary. With this aim exist the function *competitions_events()*, which returns all events that have occurred during a season. This function will be used to obtain the events for each of the competitions, specifying the necessary parameters (country, division, season, and gender). In the returned data set, each of the rows corresponds to an event and the columns will show information about that event. Among the existing columns are.

Variable	**Variable Type**	**Description**
id	Integer	The unique identifier for the event
minute	Integer	The minutes on the clock at the time of this event.
type	Object	Name of the event type.
play_pattern	Object	Name of the play pattern relevant to this event
team	Object	Name of the team this event relates to
player	Object	Name of the player this event relates to
location	Array[x, y]	Array containing two integer values. X and Y coordinates of the event.

Table 4.3: Events data set

The *type* variable can take different values depending on the type of event. In addition, each of the events has its own associated properties, for which specific columns are reserved with the name *eventName_proppertyName*, for example, *shot_outcome*. The different types of events and some of the most important properties for this work are as follows.

Event type	**Description**
50-50	Two players challenging to recover a lose ball
Bad Behavior	Instances when a player receives a card due to an infringement
Ball Receipt	Instances when a player receives a ball
Ball Recovery	Recovery ball lost by a teanmate on bad touch or dribble
Block	Preventing it from reaching its intended target.
Carry	Represents a player controlling the ball at their feet while moving or standing still.
Clearance	Action where a player kicks or heads the ball away from his own goal without the intended recipient

Event type	**Description**
Dribble	Indicates when a player attempts to beat an opponent while retaining possession
Dribbled Past	Indicates that a player was beaten by an opponent's dribble
Duel	Represents a contest between two opposing players.
Foul Committed	Any infringement that is penalized as foul play by a referee. Offsides are not tagged as a foul committed
Foul Won	Defined as where a player wins a free kick or penalty for his team after being fouled by an opposing player
Goalkeeper	Actions that can be done by the goalkeeper
Half End	Signals the referee whistle to finish a part of the match
Half Start	Signals referee whistle to start a match period
Injury Stopage	A stop in play due to an injury
Interception	Preventing an opponent's pass from reaching their teammates by moving to the passing lane or reacting to intercept it
Misscontrol	represents when a player loses the ball due to a bad touch.
Pass	Indicates when the ball is passed between teammates
Player Off	Added if the player left the game permanently. For scenarios where no subs are left but the player cannot return to pitch due to injury,
Pressure	Represents applying pressure to an opposing player who is receiving, carrying, or releasing the ball
Shot	Represents an attempt to score
Substitution	Indicates a player substitution

Table 4.4: Events data set

4.2.2 Data Processing

Data processing is a vital step in data analysis, ensuring the extraction of meaningful information. The dataset treatment methodology provides a structured approach to managing datasets, ensuring their quality and relevance. This includes data cleansing, which corrects errors and inconsistencies; preprocessing, which standardizes and normalizes data; and integration, which combines data from different sources. The data must then be transformed for analysis and the relevant features must be selected. Fortunately, the data with which the study will be carried out are correctly structured and do not have errors or inconsistencies. The process to be carried out is to generate a data set for each of the phases to be analyzed with the data aggregated by team or player.

Team dataset

This dataset stores all statistics that describe each phase of the game aggregated by team. Regarding the final phase, the *Shot* type event has been used to filter the event data set. The idea is to obtain a column for each of the statistics described in the problem definition. To do this, a grouping will be made for each of the teams in each of the leagues with data from the entire season, and a count will be made of the shot-type events. In addition, to emphasize the quality of these shots, we will filter according to the outcome (on target, off target, post, blocked by a defense, goal). Another column will indicate the number of total goals expected for the season. In addition, some statistics will be calculated, such as the percentage of shots on target out of the total, the shot-to-goal ratio, the shot-on-target-to-goal ratio, or the ratio of expected goals per shot. On the other hand, for the building phase, the event data set has been filtered by *Pass* type events. First, a column has been created for each of the types of passes: long pass, with a length greater than 30 yards and height is *High Pass* type; short pass, with a length less than 30 yards and height is *Ground Pass* type; final third pass, this type measures those passes that generate greater danger taking into account those that are made from x<80 and end in the final third, x>80, knowing that the field measurements offered by StatsBomb are standardized at 120x80; cross pass, this type is used to create scoring opportunities, aiming to find attacking players making runs into the penalty area; to filter this type of pass, there is a field called *pass_cross*. Second, after defining the different types of passes to be considered, the number of complete and incomplete passes for each type was calculated to measure the accuracy of each of the teams. Lastly, it has information related to the defensive phase. The defensive actions that will be measured are recoveries, getting back a ball in possession of the opposing team, duels, a 50-50 contest between two players of opposing sides in the match, and fouls committed. These actions will be grouped according to the area of the

field where they are performed, giving an idea of how aggressive a team is in different areas of the field. Additionally, the actions of the opposing team will also be taken into account, such as shots taken, goals scored, and expected goals generated. Finally, the PPDA metric average for the season will also be included for each team. For its calculation, the defensive actions taken into account are as follows: Tackle, Interception, Foul Committed, Block, and 50-50, all of them performed in the field zone x>70.

Player dataset

This data set stores statistics describing each of the positions described above, aggregated by player. It includes detailed information about players, including their names, teams, and specific positions such as Goalkeeper, Defender, Midfielder, and Forward. In relation to the goalkeeper position, the data set has several statistics. First, the saved amounts will be taken into account and broken down according to the part of the body with which they were made (both hands, chest, head, left foot, right foot, left hand and right hand) and whether they were penalty kicks or not. Additionally, the percentage of saves will be calculated in relation to the total number of shots on goal received. Second, the total number of goals and its breakdown according to whether it was a penalty goal or not. On the other hand, for the defender position, the statistics to be quantified are mostly the same as in the team data set corresponding to the defensive phase, but instead of being aggregated by the team, they are aggregated by player: recoveries record the number of times a player regains possession of the ball; duels, counts the total number of confrontations a player engages in to gain possession of the ball; the total number of fouls committed; shot blocked, notes the instances where a player has successfully blocked an opponent's shot; clearances, the count of times a player successfully clears the ball away; pressure, measure the frequency of a player applying pressure on an opponent; and interceptions, tracks the number of times a player interrupts the opposition's play progression. The same applies to the statistics related to midfielders and forwards; the information is the same as that stored in the team data set but now aggregated by the player.

4.3 Modelling

As mentioned above, the machine learning algorithm used to cluster is K-Means. The main objective of the analysis is to group similar teams and players according to the various statistics described in the previous section.

4.3.1 Team dataset

A correlation matrix study was carried out. The result of this study is a table similar to that shown in Figure 4.5 that shows the correlation between the variables of a data set. The value is in the range of -1 to 1. A value close to 0 indicates that there is no correlation, a value close to 1 signifies a strong positive relationship, and a value close to -1 shows a strong negative relationship between the two.

It can be observed that “total_shots”, “onTarget”, “goals” and “xG_total” all have high positive correlations, as well as “shots_against”, “shots_onTarget_against”, “goals_against” and “goals_against”. This is expected because more shots generally lead to more goals and a higher expected goals (xG) value indicates a higher quality of chances created, which should correlate with actual goals scored. On the other hand, “average_pass_length” and “total_passes” show a significant negative correlation, indicating that as the number of passes increases, the length of each pass tends to be shorter, which may reflect a possession-based style of play.

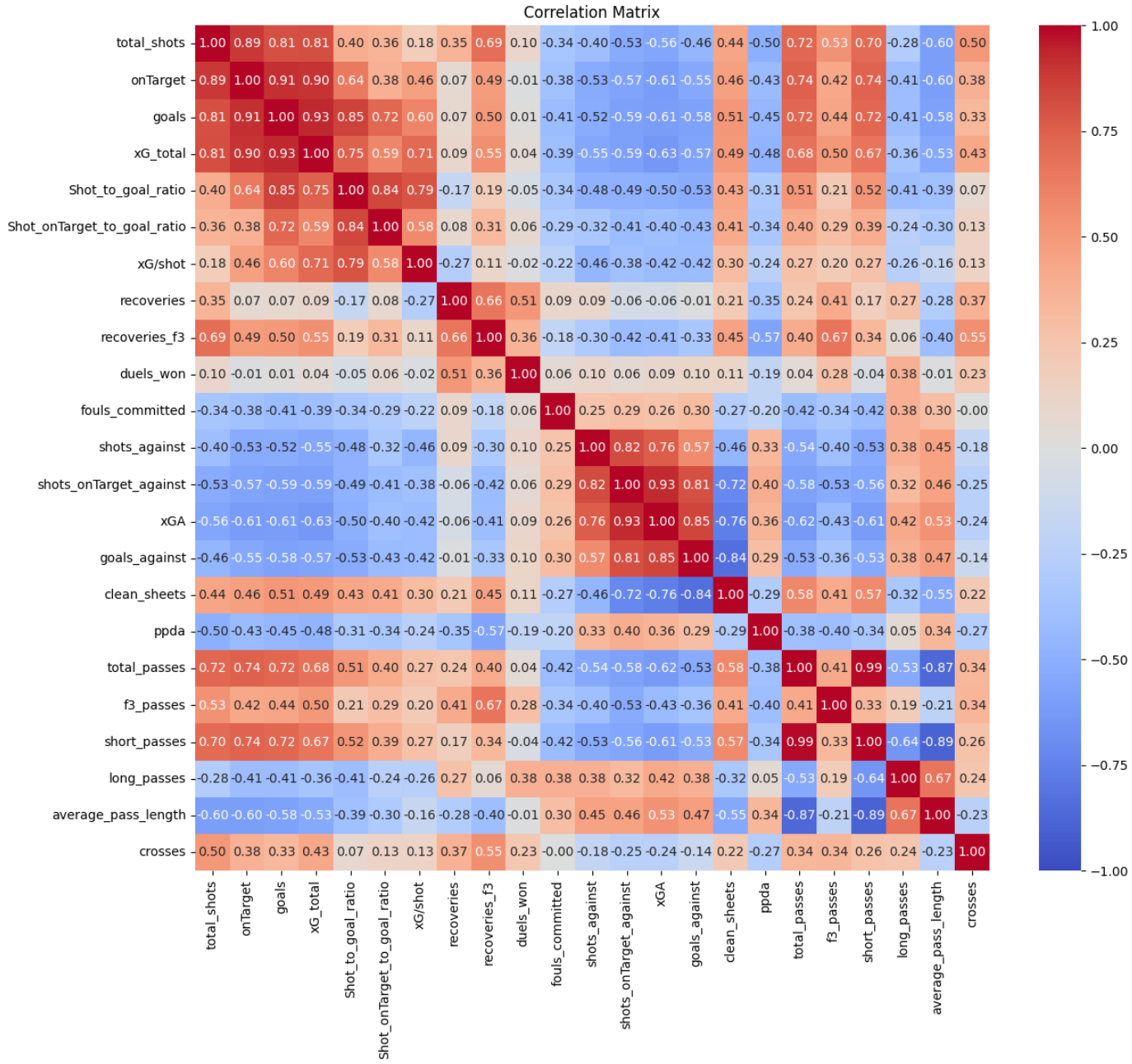


Figure 4.5: Correlation matrix, team dataset

K-Means is a distance-based algorithm; it is sensitive to the scale of the data. For this reason, *Standard Scaler* [8] technique is used. This technique ensures standardization of the features around the mean and unit variance. The scaler computes the mean of each feature in the dataset. Then, it subtracts this mean from each data point. This centers the feature at zero and removes any bias caused by features that naturally carry a higher average value than others. On the other hand, after centring the data, this technique calculates the standard deviation for each feature and divides the centered feature by it. This process transforms the data to have a variance of one, making the variation of the feature consistent across the dataset. In short, the use of *Standard Scaler* ensures that the clustering algorithm treats all features equally, leading to more meaningful and balanced clusters.

After standardizing the data, another key phase in the analysis is determining how many clusters are used. For this purpose, *Elbow method* is used. This method is a heuristic that is used to determine the optimal number of clusters for K-Means clustering. It involves running the clustering algorithm multiple times over a range of k values and calculating a score that represents the variance within the clusters for each k. The result of this analysis can be observed in Fig. 4.6, which shows that the optimum number of clusters to be used is 4.

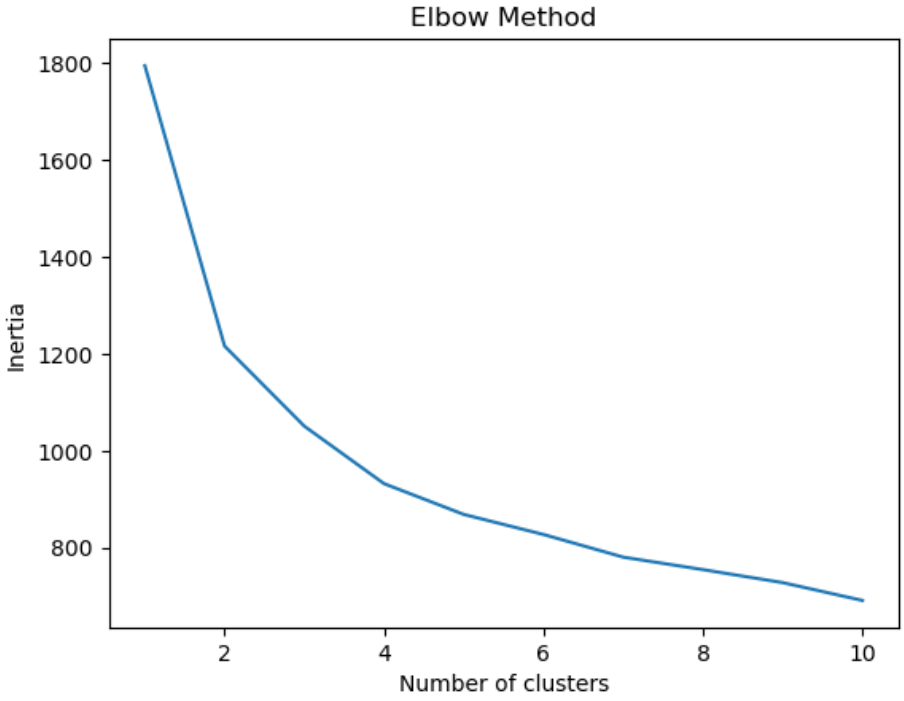


Figure 4.6: Elbow method, team dataset

4.3.2 Player dataset

In this section, the same study was carried out as was done with teams, but now with players. It was divided into groups according to the positions on the field: goalkeepers, defenders, midfielders, and forwards. For each group, the statistics and metrics mentioned in the problem definition section were used. Images of correlation analysis and elbow method can be found in the Appendix.

Goalkeepers

For this analysis, the player dataset was filtered for the goalkeeper position. The study tries to group different styles of goalkeeping, highlighting different strengths and weaknesses in terms of passing distribution, stopping shots, and handling penalty situations. These insights can be valuable for understanding different goalkeeping styles and for the formulation of team strategies. Furthermore, to avoid distortion of the analysis, the dataset has also been filtered for goalkeepers with a higher number of games played than the average (17).

First, a correlation matrix study was carried out, as can be seen in Figure C.1. It should be noted, as is logical, that there is a strong negative correlation between "average_pass_length" and "short_passes", in contrast to "long_passes", which has a strong positive correlation with it. On the other hand, there is a strong correlation between "non_penalty_goals_against" and "total_goals_against", because most goals come from open play. In addition, it is possible to highlight the strong positive correlation between "penalty_saves" and "penalty_saves_percentage%"; and a moderate negative correlation between "save%" and "total_goals_against" and "non_penalty_goals_against", respectively.

After performing this correlation study, as in the previous case, the data was standarized and the number of clusters used was determined using the *Elbow method*, reaching the conclusion that 4 is the optimal number.

Defenders

The dataset of players was filtered for the position of defenders and for those who had played more than 18 matches. The fields used in the study are those described in the problem definition.

Regarding the correlation study represented in Figure C.3, some points can be highlighted. First, "total_passes" has a very strong positive correlation with "short_passes", due to the fact that players who tend to do a high number of passes tend to make short passes,

which may reflect that those players play in teams that tend to have a possession-based style. Furthermore, continuing with the fields related to the passes, it can be observed that there is a strong positive correlation between "long_passes" and "f3_passes"; and between "total_passes" and "recoveries". On the other hand, "pressure" has a high positive correlation with "recoveries", "duels_won", "fouls_committed" and "shot_blocked"; this shows a strong relationship between actively contesting for the ball and applying pressure, which are key components of aggressive defensive play.

Finally, after standardization and the *Elbow method*, it was concluded that the optimal number of clusters is 4.

Midfielders

In this case, the player dataframe was filtered by midfielder position and those players that have played more than 19 matches.

Regarding the correlation analysis observed in Figure C.5, there are some interesting points beyond those commented on in the defenders section. On the one hand, a high positive correlation can be observed between "crosses" and "total_shots_assists", indicating that after a cross, there is usually a completion of the play in the form of a shot. "crosses" has also a moderate correlation with "dribbles_f3_completed", due to the fact that the cross is usually preceded by a completed dribble in the last third of the field. On the other hand, "dribbles_f3_completed" has a moderate positive correlation with "total_shots" and "xG_total", which could indicate that a completed dribble in this area of the field leads to more dangerous shooting actions. In addition, as logical, the fields "total_shots", "goals" and "xG_total" have a strong positive correlation between them.

The result of the analysis of the *Elbow method* after standardizing the data is that the optimal number of clusters to use is also 4.

Forwards

For this group, the position used to filter the player dataframe is forward, and the average number of matches played is 19.

This group shares fields with the previous ones; for this reason, the analysis of the correlation that can be observed in Figure C.7 is similar in most cases. The fields "total_passes", "f3_passes", and "total_shots_assists" have a high positive correlation between them. In addition to "total_shots", "goals" and"xG_total", which also have a strong positive cor-

relation. The new fields that can be discussed in this group are "shot_to_goal_ratio" and "shot_onTarget_to_goal_ratio", which, as is logical, also have a high positive correlation.

As in the previous cases, after standardizing the data and performing the study using the *Elbow method*, the optimal number of clusters was deduced to be 4.

4.4 Evaluation

In the preceding sections, all processes were explored to apply the K-Means clustering algorithm. At this point, the results obtained after applying the model to each of the cases described above will be examined and discussed. This section aims to assess the validity of the formed groups and interpret their significance in the context of professional football. The different statistics emerging from each group will be explored, ensuring that they fit both the empirical data and the conceptual understanding of the roles of football.

4.4.1 Team dataset

As observed in the analysis with *Elbow method*, the optimal number of clusters is 4. The main characteristics of each of the equipment that makes up the different clusters are described below. Then, it will be tried to draw some conclusions once the different clusters have been broken down.

Cluster 0: Relegation Contenders

This cluster has 27 teams: Rayo Vallecano, Frosinone, Crystal Palace, Atalanta, Hellas Verona, RC Deportivo La Coruña, Udinese, Sunderland, Watford, Levante UD, Sampdoria, Getafe, Carpi, Granada, Chievo, Ingolstadt, Stoke City, Eibar, Norwich City, Palermo, Real Betis, Newcastle United, Espanyol, West Bromwich Albion, Bologna, Aston Villa and Sporting Gijón. Based on the mean values of each of the statistics by which the study is conducted, the style of these teams can be defined as conservative.

From the point of view of the attack, this group has a low mean number of shots and the lowest mean number of shots on target. In terms of efficiency in front of goal, they have performed below expectations, scoring a mean of 38 goals out of 40.3 expected goals. In addition, they have a shot-to-goal ratio of 8.78 and a shot-on-target-to-goal ratio of 27.86, values lower than those of the rest of the clusters. Therefore, from a finishing point of view, it can be highlighted that this group generally generates fewer scoring opportunities and has lower efficiency than the other clusters.

In terms of the game-building phase, the teams in this group have the lowest mean number of passes and the highest mean pass length, 20.97 meters. This could suggest a style of play that is less focused on building attacks through controlled possession, if not through long balls, with a more direct and reactive style.

Finally, with an emphasis on defensive aspects, some points can be highlighted. This is the cluster with the second-highest value of PPDA, 14. This indicates that, in general, these teams tend to have a low block when they are not in possession of the ball. On the other hand, this is the cluster with the second-highest mean of recoveries and duels won, as well as the one that commits the most fouls, indicating that they are aggressive teams in terms of defence. -highest mean of recoveries and duels won, as well as the one that commits the most fouls, indicating that they are aggressive teams in terms of defence. However, this aggression is not reflected in defensive efficiency, as the group with the highest mean number of shots against, shots on target against, and the highest number of goals conceded, as well as the group with the lowest mean of clean sheets

Cluster 1: Title Contenders

This group contains 12 teams: Real Madrid, Napoli, Tottenham Hotspur, Bayern Munich, Manchester City, Juventus, Barcelona, AS Roma, Arsenal, Fiorentina, Borussia Dortmund and Atlético Madrid. In this case, there are the teams that usually stand out and dominate their respective leagues.

With regard to the attack aspect, the statistics show strong offensive capabilities. These teams have the highest mean of shots and shots on target. Their goal-scoring efficiency is remarkable, scoring 77 goals from 67.48 expected goals, which could indicate that their attacking players have above-mean qualities. Furthermore, 12.82% of their total shots end up beating the goalkeeper, and of their shots on target, 34.49% end up in the back of the net, which are the highest percentages among all clusters, demonstrating a clinical approach in their attacking plays.

From a build-up play, this cluster leads in the number of total passes made, highlighting a clear tendency to make short passes with a mean pass distance of 18.35 meters, the shortest of all the clusters. They also have the highest number of passes made in the last third of the opponent's field, as well as the highest number of crosses into the box. This indicates that they are teams that tend to set the pace of matches and are quite dominant in the game.

Defensively, these teams present the lowest value of PPDA (11.92); this is in addition to the number of recoveries in the last third of the opponent's field, in which they also stand out as the cluster with the highest mean (467), an indication of a more aggressive pressing style and a proactive effort to win back possession. Also noteworthy is their record of clean sheets (16), complemented by the lowest mean of shots against and goals against, conceding 30 goals of 32.6 expected goals, this suggests not only aggressive play but also defensive discipline and effectiveness.

Cluster 2: Mid-table teams

The teams that are part of this cluster are 19: Wolfsburg, VfB Stuttgart, Schalke 04, Torino, Borussia Mönchengladbach, Werder Bremen, Swansea City, Augsburg, Las Palmas, Valencia, FC Köln, Hoffenheim, FSV Mainz 05, Hannover 96, Hamburger SV, Eintracht Frankfurt, Darmstadt 98, Villarreal and Hertha Berlin. In this group, German teams predominate; this may be due to the fact that there are fewer teams in the Bundesliga and, therefore, fewer matches.

In terms of offensive phase, this cluster has the lowest mean of total shots and the second lowest mean of shots on target. They achieve goals (42.7) slightly above their expected goals (42.1), indicating a slight edge in finishing effectiveness. Furthermore, these teams have a decent 10.33% shot-to-goal ratio and a low value of shot-on-target-to-goal ratio, 29%. In short, they have moderate attacking metrics, with a fair balance between shots taken and goals scored.

Regarding the game creation phase, these teams have a low mean of passes with an mean pass length of 20.72 meters, the second with the highest value. It should also be noted that they are the cluster with the lowest mean of passes to the final third and the lowest mean of crosses. This indicates that they are the set of teams that least reach the final third of the opponent's field, so it can be deduced that they tend not to be in control of matches but rather adopt a more reactive approach.

Defensively, the cluster shows the highest value of PPDA (15.05); this, added to the fact that they are the cluster with the lowest number of recoveries in the last third of the opponent's field, confirms the above comments about the reactive approach of the teams of these groups. They are the group with the second-highest mean of shots against, conceding 49.4 goals out of 49.5 expected goals and keeping eight clean sheets.

Cluster 3: European Contenders

This cluster is made up of 20 teams: Liverpool, AC Milan, West Ham United, Chelsea, Inter Milan, Southampton, Leicester City, Lazio, Everton, Sassuolo, Empoli, Bayer Leverkusen, AFC Bournemouth, Genoa, Sevilla, Real Sociedad, Athletic Club, Málaga, Celta Vigo and Manchester United.

This cluster presents an interesting offensive profile, with a high mean of shots and a reasonably high conversion rate, scoring 51.05 goals from 50.03 expected goals, indicating a proactive attack. Their shot-to-goal and shot-on-target-to-goal rates stand at 10.16% and 30.89%, respectively.

The build-up play for these teams is characterized by a significant number of passes with a slightly lower mean pass length of 19.68 meters, suggesting a preference for a more dynamic and possibly wing-oriented attacking style, as evidenced by their substantial number of crosses.

On the defensive end, these teams show a PPDA value of 12.62, the second lowest value, implying a calculated pressing game. They are the group with the highest mean of recoveries and duels won; a large part of these recoveries take place in the last third of the opponent's field. They have a considerable number of shots against them, receiving 45.35 goals from 44.47 expected goals and leaving a clean sheet 11.9 times.

In the following, the previously made description of each of the clusters will be complemented with some graphs that allow a better evaluation of the results. The features of each of the game phases (attack, build-up, defense) will be compared in order to see how the teams behave in each of these phases. First, the mean values for the attack phase are plotted, as shown in Figure 4.7. This figure confirms what has been seen in the description of each of the clusters. Cluster 1 stands out due to its dominance of the attack, as it is the one with the highest mean values of total shots, shots on target and goals. It is notable not only for its total number of shots, but also for its efficiency in front of the goal, being the one with the best shot-to-goal and shot-on-target-to-goal rate, as well as the one with the best goal return based on his expected goals. On the other hand, cluster 0 and 2 have similar profiles across the mean values of shots, shots on target and goals. However, cluster 2 is slightly superior in terms of efficiency. Cluster 3 seems to be in the middle ground, with respect to the number of shots and goals, between Cluster 1 and Clusters 0 and 2. However, cluster 3 has similar values to cluster 2 in the features that measure efficiency.

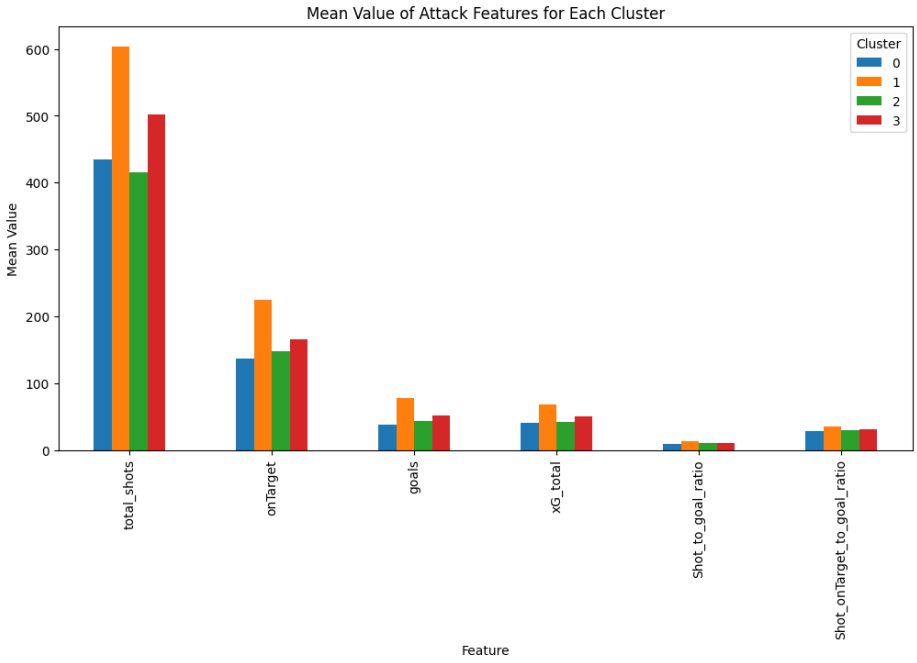


Figure 4.7: Mean values attack features

Figure 4.8 shows the distribution of the build-up features for each cluster. As was the case with the attack, it can be seen that cluster 1 also stands out in terms of the number of total passes, indicating that these teams tend to dominate matches through a possession-based style. In addition, if we look at the features related to the type of pass and the place where they are made, we can deduce that cluster 1 tends to make short passes in the last third of the opponent's field, a fact also related to the number of scoring chances seen above. Furthermore, it can be seen that the teams in this group are the ones that make the most crosses into the box, so these teams tend to carry out their attacks through the flanks. As for the other clusters, it is found that cluster 0 and cluster 2 have similar values in terms of passing, although the distribution of passes is different, cluster 0 has a greater tendency to play long balls, as well as being more prone to play on the flanks since it has a considerably higher mean value of crosses. On the other hand, cluster 3 again has values that are in the middle ground between clusters 0 and 2 and cluster 1. This group stands out for having a large number of passes in the last third of the opponent's field, as well as a large number of crosses into the box.

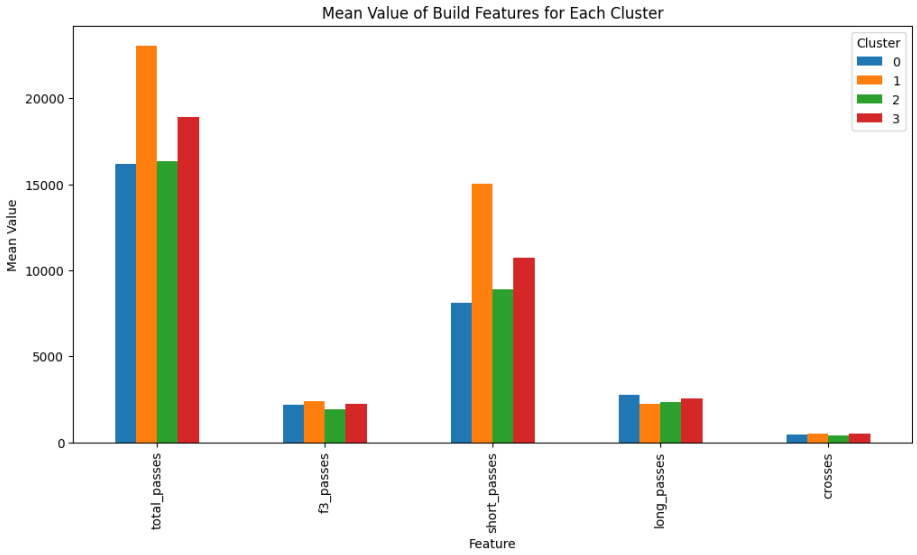


Figure 4.8: Mean values build features

In Figure 4.9 are represented the mean values of the defense characteristics are represented for each group. Unlike the attacking and construction phases, in this defensive phase there is no such clear predominance in all the statistics of cluster 1. It does stand out in the metrics related to shots conceded and goals conceded; in addition, it excels in terms of recoveries in the last third of the opponent's field, which indicates that they play a defensive game based on high pressure. On the other hand, cluster 0 stands out for having relatively high values of recoveries and duels won; it is also the cluster with the highest number of fouls committed, which indicates that they base their defense on a more physical style. They also stand out negatively in metrics related to shots conceded and goals conceded, being the cluster with the worst numbers. As for cluster 2, it can be seen that they are the set of teams that make the fewest recoveries and win the fewest duels, a fact that contrasts with a high number of fouls committed, which may indicate that they do not measure defensive actions well. They also stand out negatively in the statistics of shots conceded and goals conceded. At the same time, cluster 3 stands out for its great ability to recover the ball, as well as to win duels. As for the statistics of shots against and goals conceded, it has values between those of cluster 1 and those of clusters 0 and 2.

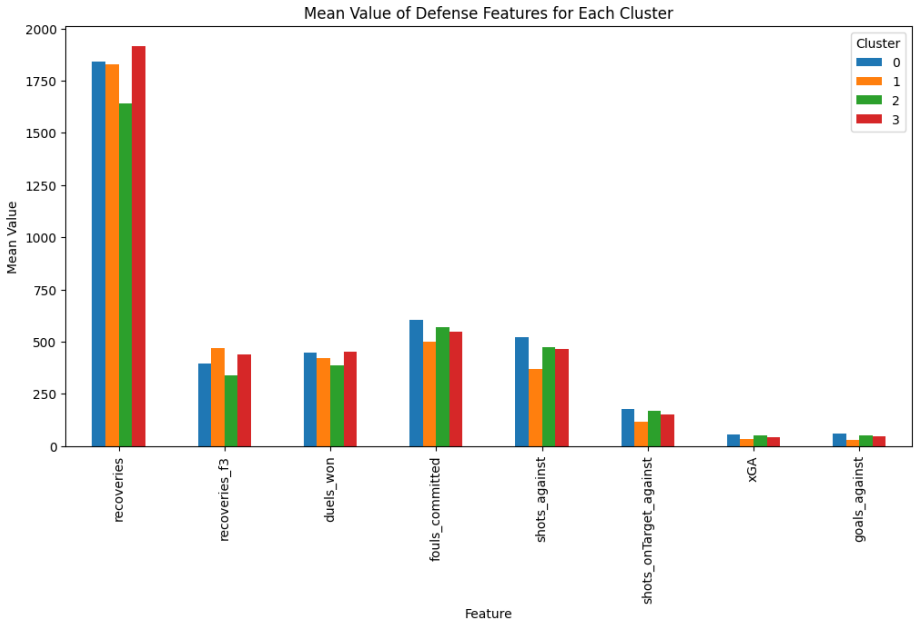


Figure 4.9: Mean values defense features

In order to check whether the different groups with their respective characteristics correctly reflect the position of the clubs in the season's ranking, the data used in the model have been crossed with the statistics corresponding to the ranking (position, wins, draws, losses, points, points per game). For a better understanding of this, use has been made of the visualization of the type box plot has been used [30], as shown in Figure 4.10, a standardized graphical representation showing the distribution of data based on a summary of five numbers: minimum, first quartile (Q1), median, third quartile (Q3), and maximum:

- Minimum: The smallest data point excluding any outliers, and is represented by the end of the lower whisker.
- First Quartile (Q1): The median of the lower half of the data, marking the 25th percentile. It is at the bottom of the box.
- Median: The middle value of the data when ordered from lowest to highest, marking the 50th percentile. It is the line inside the box.
- Third Quartile (Q3): The median of the upper half of the dataset, marking the 75th percentile. It is at the top of the box.
- Maximum: The largest data point, excluding any outliers, is represented by the end of the upper whisker.

The Interquartile Range (IQR) is the distance between Q1 and Q3 and represents the central 50% of the data. The whiskers extend from the extremes of the box (Q1 and Q3) to the highest and lowest values that are within 1.5 times the IQR of these quartiles. Data points outside this range are considered outliers and are plotted as individual points.

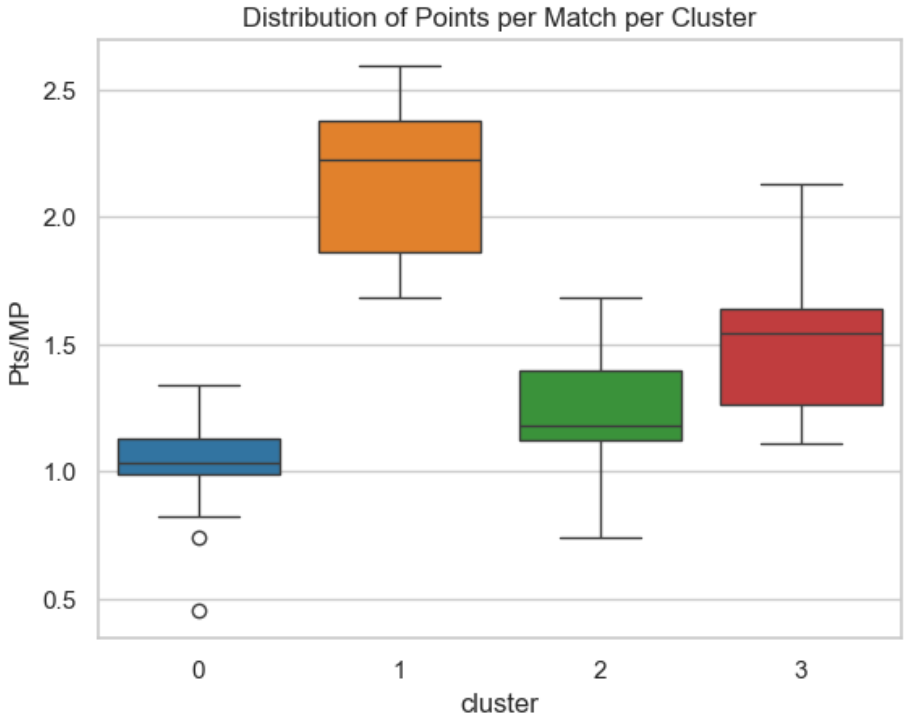


Figure 4.10: Points per match by cluster

The representation of points per match confirms some of the trends that had been observed when evaluating the mean values of each of the clusters. Teams in cluster 1, which excelled in almost all aspects previously studied, are those who obtain the most points per match, with an average of 2.14. On the other hand, the teams in cluster 0, which had the worst values in most of the metrics studied, obtained the fewest points per game, with an average of 1.03. This cluster also has a fairly narrow interquartile range, which implies less variability in points per game, highlighting the existence of two outliers that have lower performance than the rest. As for cluster 2, it is observed that it has values that are between clusters 0 and 1. Its interquartile range is wider than that of cluster 0 but narrower than that of cluster 1, indicating moderate variability in equipment performance. The average number of points per game for this cluster is 1.23. Finally, it can be seen that cluster 3 has slightly higher values than cluster 2. The existence of greater variability in the data stands out, with equipment that exceeds expectations compared to the rest of the group. The average value of points per game in this cluster is 1.49.

In conclusion, it can be said that the features selected for the study are relatively good in determining the ranking performance of the different teams since the values in the different metrics selected in the different phases of the game (attack, construction, and defence) reflect the final result of the season.

4.4.2 Player dataset

In this section, the results obtained from the clustering performed for each of the defined positions (goalkeeper, defender, midfielder, and forward) will be analyzed. The aim is to see the different roles that may exist in each of the positions. For this purpose, the average values of each of the defined metrics have been studied.

4.4.2.1 Goalkeepers

Cluster 0

The goalkeeper defined in this cluster is a player who makes a large number of passes, mostly short, with an average pass length of 34.19 meters. This indicates that they play in teams that have an associative character and come out playing the ball from the back. They are the cluster with the lowest average of goals conceded (33.06), and although they do not concede many goals from penalties, with an average of 2.8, they do not stand out for being specialists in saving penalties. In terms of the percentage of savings, they have the second-highest average with a value of 72.64%. They show to be balanced between the use of hands and feet to make saves, somewhat superior to the use of the left foot over the right and the right hand over the left.

The goalkeeper that describes this cluster excels in most of the analyzed metrics, so it can be deduced that this group includes some of the most valued goalkeepers, such as: Manuel Neuer (Bayern Munich), Claudio Bravo (F.C Barcelona) or Gianlugi Buffon (Juventus).

Cluster 1

This cluster has a similar average number of total passes as cluster 0, but the distribution is quite different, with the long pass being the predominant one, with an average pass distance of 47.31 meters, which indicates that they play in teams with a more direct style. The goalkeepers in this group are those with the highest number of average goals against (56.4), a value well above that of the other clusters. They do not stand out for having a good percentage of penalty kicks saved, with an average of 11.27%; nor for its overall stop percentage, being the cluster with the lowest average (67.8%). They show a balance of saves in terms of the use of hands and feet, with a slightly higher use of the right hand over the

left, as well as the right foot over the left.

This cluster includes some of the goalkeepers of the lower-ranked teams that have competed to avoid relegation, for example: Iván Cuéllar (Sporting Gijón), Bradley Guzan (Aston Villa), Vicente Guaita (Getafe) or Ron-Robert Zieler (Hannover 96).

Cluster 2

The type of goalkeeper described in this cluster stands out because it has a high average number of passes, the predominant being the long pass, with an average passing distance of 42.64 meters. Therefore, it could be said that the goalkeeper of this cluster actively participates in the team's game by means of a more direct style. They are the second cluster with the second highest number of goals against (41.2). However, they stand out for having the best average of total saves (74.53%) and the best average of saved penalties (28.39%), a percentage much higher than the rest of the clusters. They show a balanced distribution of saves between the left and right foot, as well as a certain preference for making saves with the right hand instead of the left.

Among the most prominent names of goalkeepers in this cluster are some household names such as Keylor Navas (Real Madrid), Bernd Leno (Bayer Leverkusen), Samir Handanovi (Inter Milan), or Adrián San Miguel (West Ham United).

Cluster 3

This is the cluster with the lowest average by difference in the number of passes made (853.7), a number much lower than the next cluster (1071.27). This indicates that the type of goalkeeper in this cluster participates to a lesser extent in the team's game. They have the longest average passing distance (46.41), so when they do intervene, they tend to do it longer. The distinguishing feature of these goalkeepers is the number of goals conceded, with the second-lowest average (33.39). They also have the second-best percentage of penalties saved (16%) and a considerably good percentage of total saves (71.19%). As for the distribution of saves with the different parts of the body, no significant differences are observed between the use of both hands and feet.

In short, the goalkeepers in this cluster stand out in some of the metrics studied, such as the number of goals conceded, which is why here we find outstanding names such as: Jan Oblak (Atlético Madrid), Petr Čech (Arsenal), David de Gea (Manchester United) or Kasper Schmeichel (Leicester City).

4.4.2.2 Defenders

Cluster 0

The defenders belonging to this group stand out for their high average number of passes (1369.69), mostly short passes (747.7), with a high average number of passes to the opponent's final third (163.84). They are the cluster that makes the fewest crosses (5), so this gives us an idea of where these players stand in the defensive line. They also stand out for having the highest average passing distance (22.46 meters). From a defensive point of view, they are the second cluster with the second highest average of recoveries (113.9), being also second in the number of duels won (29.9) and fouls committed (32.86). They also have a high number of blocked shots (58.8) and interceptions (48.78). The most outstanding metric is the number of clearances made, with an average of 173.4.

The metrics studied indicate that the players in this cluster tend to play in the centre back position, both left and right. To corroborate this, the position that each player occupies within the defence was added. Counting the distribution of positions in the cluster, it is observed that there are 45 left center backs, 44 right center backs, six center backs (who can play on both sides); and only four right full backs. Some names stand out, such as Gerard Piqué (Barcelona), Diego Godín (Atletico Madrid), Sergio Ramos (Real Madrid) or Andrea Barzagli (Juventus)

Cluster 1

This cluster shows a somewhat lower amount than that shown by cluster 0, with an average of 1161.59 total passes; most of them are short passes (611.16). In this case, there are a large number of centres (44.97), and the average passing distance is lower than in the previous case, with 17.9 meters. These defenders do not stand out, especially for their number of recoveries (90.45), although they maintain similar numbers in terms of duels won (28.06) and fouls committed (29.4). They also have lower numbers in terms of blocked shots (49.7), interceptions (30.16), and much lower clearances (58.32). One metric in which they improve on the previous cluster is the number of pressure actions performed (305.96).

In this case, the distribution of positions in the different players of the cluster is 46 right backs, 27 left backs, eight right wing-backs, five left-wing backs, five right centre-backs and four left centre-backs. This fact is demonstrated by the data, as they have a higher number of centre-backs, so they are players who tend to reach dangerous areas on the wing and put balls into the area, as well as slightly lower values in the set of purely defensive metrics such as clearances, interceptions, or blocked shots. Some names stand out, such as

Daniel Carvajal (Real Madrid), Joshua Kimmich (Bayern Munich), Philipp Lahm (Bayern Munich) or Patrice Evra (Juventus).

Cluster 2

The players in this cluster have the highest average of total passes (1757.38), most of them short passes (911.84), so these defenders are a very active part of the team's game. The large number of crosses made (65.2) stands out, indicating that they are players who reach dangerous areas from which they send crosses into the area; they are also the cluster with the highest number of passes made to the last third of the opponent's field (217.29). They are the cluster with the highest average of recoveries (145.18), duels won (42.4) and fouls committed (39.04), which shows great aggressiveness when recovering the ball through defensive actions. In addition, they are also the leaders in blocked shots (71.2), interceptions (50.38), and pressure actions (424.78).

This cluster stands out in most of the metrics studied, both in those related to passing and in purely defensive metrics. In addition, the different positions found within this cluster are 38 left backs, 32 right backs, seven left center backs, two left-wing backs, and two right center backs. This shows that they are players who actively participate in their team's play and are capable of reaching dangerous areas where they generate goal scoring opportunities through crosses into the box. Some of the most prominent names are Filipe Luis (Atletico Madrid), David Alaba (Bayern Munich), Alessandro Florenzi (AS Roma) and César Azpilicueta (Chelsea).

Cluster 3

This is the cluster with the lowest average of passes made (834.69), which denotes a lower participation in the team's game. The distribution is still logically mostly of short passes (451.13), but its low average of crosses (6.4), as well as passes to the last third of the opponent's field (91.9) stand out. Furthermore, they are the second cluster with the second highest average in passing distance (21.32). As for defensive metrics, they have values well below the rest of the clusters in most of them: turnovers (65.1), duels won (19.57), fouls committed (22.25), blocked shots (39.2), interceptions (27.96), and pressure actions (196.45).

This group of players has the worst averages in most of the metrics studied. The distribution of positions in the cluster is: 46 right center backs, 37 left center backs, 11 right backs, 9 left backs, and 1 center back. So, the predominant position is center back, both right and left. The names that can be highlighted in this group are Raphaël Varane

(Real Madrid), Gary Cahill (Chelsea), John Terry (Chelsea) or Jérôme Boateng (Bayern Munich).

4.4.2.3 Midfielders

Cluster 0

This cluster defines a type of midfielder who has a significant average of passes made (1162), among which stands out his large number of crosses (55.46), being the cluster with the lowest average pass placement (16.86). He does not particularly stand out for being the type of midfielder who takes the ball to the last third of the opponent's field from more backward positions; his average number of passes of this type is 132.13. However, what he does stand out is the creation of goal actions, with an average number of assist shots of 42.26, of which only 5.41 ended in a goal. In addition, this midfielder profile stands out for its large number of completed dribbles (50.6), having completed more than half of them in the last third of the opponent's field (26.2). This is related to the number of fouls that he receives, with an average of 51.88, a value very similar to that of the cluster with the highest value (53.06). His ability to generate danger by shooting also stands out, with a total shot average of 56.63, generating expected goals of 5.36 and scoring a total of 5.88 goals. As for defensive metrics, this profile does not stand out especially in any of them; the only ones in which it has reasonably good numbers are recoveries (132.49) and pressure actions (483.2).

In short, it is a midfielder profile that has a great capacity for overflow and the generation of danger, both in the form of assistance to teammates and in the form of shooting. Some outstanding names in this group are: Mesut Özil (Arsenal), Kevin De Bruyne (Manchester City), Francisco Román Alarcón (Real Madrid), who are profiles more of the classic playmaker type; although there are also other players who are more of the winger type such as Carrasco (Atletico Madrid), Kingsley Coman (Bayern Munich) or Riyad Mahrez (Leicester City).

Cluster 1

In this group are the midfielders with the highest average of total passes (1990.5), their average of short passes (1249.06) is higher than the average of total passes of the rest of the clusters. This indicates that they are the players through whom all the team's game construction takes place. They also have a relatively high number of crosses (32.9), as well as a large number of passes to the last third of the opponent's field (298.19). They are the cluster with the second-highest average of shot assists (39.2), which translates into 3.73

goal assists. This midfielder profile is the one that receives the most fouls (53.07); this may be due to the fact that they have the ball in their possession during many phases of the match; in addition, they are the cluster with the second-highest average of completed dribbles (29.08), although only a small part of them are in the last third of the opponent's field (9.19). However, they are a group that has a relatively high average of shots (36.9), although they score only 2.57 goals out of 2.41 expected goals. So it can be deduced that in general they take those shots from positions that are unlikely to result in a goal. As for defensive metrics, they are the cluster with the best values in all of them, recoveries (183.66), duels won (50.42), fouls committed (50.7), interceptions (44.36), and pressure actions (625.6).

In conclusion, it can be deduced that this cluster defines a type of total midfielder, who dominates both playmaking and defense-related tasks. Some names stand out: Toni Kroos (Real Madrid), Sergio Busquets (Barcelona), Xabi Alonso (Bayern Munich) and N'Golo Kanté (Leicester City).

Cluster 2

In this case, this group of players has a considerable passing average (1199.82), although lower than that of cluster 1 (1990.5), which indicates that these players are also highly participative in the game, although with a somewhat secondary role with respect to the previous cluster. Their low number of crosses (14.39) stands out, which could imply that they do not occupy dangerous positions on the field. In addition, they have very few shot assists (17.04), which translates into 1.35 goal assists. They also have a low number of completed dribbles (20.03), of which only 5.58 are in the opponent's final third of the field; they still have a considerable number of fouls conceded (39.96). In terms of scoring, they do not stand out for having a high average of shots taken (24.08), with an average of 1.62 expected goals, of which they have scored 1.48. In terms of defensive actions, they stand out in terms of turnovers (135.55), duels won (42.77), fouls committed (50.27), interceptions (37.02) and pressure actions (542.38). These data denote high participation in defensive aspects.

In short, this cluster defines a type of midfielder who actively participates with a secondary role in the construction of the team's play and whose main task is to provide balance through his defensive actions. Several names stand out: Carlos Henrique Casemiro (Real Madrid), Morgan Schneiderlin (Manchester United), Nemanja Matic (Chelsea) and Leon Goretzka (Schalke 04).

Cluster 3

Finally, this cluster includes players who do not have active participation in the team's game, as shown by the averages of the different metrics studied. They have the lowest average of passes made (652.87), the lowest average of passes made to the last third of the opponent's field (78.98), as well as low participation in the generation of chances, having the lowest average of shot assists (16.56) which translates into an average of 1.48 goal assists. They are the cluster with the lowest average of completed dribbles (17.75), although it is worth noting that almost half of these (8.05) take place in the last third of the opponent's field, which indicates that they tend to occupy dangerous areas. They also do not stand out in the generation of danger through shooting, as they have the lowest average of total shots (22.42), scoring 1.79 goals out of 1.83 expected goals. As for defensive metrics, they have the worst averages in all of them, recoveries (77.2), duels won (18.67), fouls committed (26.24), interceptions (16.6) and pressure actions (306.79).

The midfielder described in this cluster could be said to have no active participation in the construction phase of the play, while his defensive contribution is not remarkable either, which leads us to think that they are not indispensable players in their respective teams or that they play in teams that do not usually control the games. Some names stand out: Mateo Kovačić (Real Madrid), Julian Brandt (Bayer Leverkusen), Sami Khedira (Juventus), or Bojan Krkc Pérez (Stoke City).

4.4.2.4 Forwards

Cluster 0

This cluster is defined as a forward who likes to actively participate in his team's game, as reflected in his passing metrics, being the group with the highest average of passes made (1107.2), of which 110.68 are passes made to the last third of the opponent's field. They also stand out for their generation of key passes, with an average of 40.10 shot assists that translate into 5.07 goal assists. His generation of danger comes through dribbling, with an average of 53.95 completed dribbles, more than half of them (28.82) made in the last third of the opponent's field. This means that they are the cluster with the highest average of fouls received (57.5). In terms of offensive production through shooting, they are the cluster with the second-highest average of total shots (61.46), although far behind the first (125.84). They are not the most effective cluster in front of the goal; only 10.52% of their shots end in goal, and 29.72% of their shots on goal are goals. They have an average of 6.53 goals from 5.87 expected goals, so they score more goals than they should according to the odds of their chances. In terms of defensive actions, they are the most active cluster,

with the highest average in duels won (25.39), second in fouls committed (39.39), and first in pressing actions (474.63).

In short, this is a type of forward who is very active in all aspects of the game, both offensively and defensively. He stands out mainly for his generation of danger through assistance and dribbling, a statement that can be corroborated by the predominant position within the cluster, which is that of the winger. Some of the names in this group are Leroy Sané (Bayern Munich), Mohamed Salas (AS Roma), Alexis Sánchez (Arsenal), or Raheem Sterling (Manchester City).

Cluster 1

This type of forward, unlike the previous one, does not actively participate in the team's game. It is the cluster with the lowest average of total passes (476.52), being the worst cluster in terms of passes to the last third of the opponent's field (45.36) and generation of danger through assists (16.7), which translates into an average of 1.74 assists. Nor do they stand out for their dribbling, with an average of 20.43 completed dribbles, of which 11.15 are in the opponent's final third of the field. As for their shooting metrics, they do not have a high volume of shots taken (32.64), which translates into a 2.74 goals from 3.4 expected goals. Only 8.36% of their total shots and 23.4% of their shots on target end up being goals. In terms of defensive efforts, they are the cluster with the worst averages in all metrics, duels won (10.98), fouls committed (23.03), and pressure actions (245.01).

In short, it can be said that the type of player described in this cluster does not actively participate in the team's play, nor does he stand out for his generation of danger through dribbling or key passes. In terms of his performance in front of the goal, his numbers do not either stand out positively. So it can be assured that this is a type of player that, if he belongs to a team in the upper part of the table, is not the first choice in terms of the offensive plot of the team; or plays in teams that do not generate many scoring chances. Some notable names in this cluster are Fernando Llorente (Sevilla), Mario Balotelli (AC Milan), Álvaro Negredo (Valencia), or Jesse Lingard (Manchester United).

Cluster 2

This cluster defines a type of forward that actively participates in the team's game by passing, with an average of 1022.31 passes made, of which 94.1 are passes made to the last third of the opponent's field. The players in this cluster generate many dangerous opportunities by passing, being the group with the highest average of shot assists (40.84) and goal assists (6.47). In addition, they are players with a tendency to dribble, having

the best values in completed dribbles (55.42), of which more than half (35.21) are dribbles completed in the last third of the opponent's field; this translates into a high value of fouls received (57.26). They stand out widely in terms of shooting-related metrics, being the cluster with the highest average of total shots (125.84). They have a goals average of 24.31 from 20.18 expected goals, values well above the rest of the clusters. 19.65% of their total shots and 42. 8% of their shots at the target end in goals. In terms of defensive metrics, they do not stand out in any of them, being the third cluster in duels won (12.57), in fouls committed (33.73), and the second in pressure actions (370.32).

In short, this cluster defines a very complete type of striker, capable of generating danger from passing, dribbling, and shooting. In this group, we find the most top players, with some names like Cristiano Ronaldo (Real Madrid), Leo Messi (Barcelona), Robert Lewandowski (Bayern Munich) or Harry Kane (Tottenham Hotspur).

Cluster 3

The players in this cluster are not very participative with respect to the team's passing game, being the cluster with the second-lowest average of total passes (541.83), of which only 53.23 are passes to the last third of the opponent's field. They also have quite low numbers in terms of generating dangerous actions through passing, with an average of 18.38 shot assists, which translates into 2.68 goal assists. They are the cluster with the lowest average in completed dribbles (19.89), of which 10.94 are dribbles completed in the last third of the opponent's field. Regarding the goal scoring aspect, they do not generate many scoring chances, with an average of 52.54 total shots, but they prove to be highly effective, with 17.86% of their total shots and 43.45% of their shots on goal ending in a goal. This translates into an average of 8.97 goals from 8.16 expected goals, the second cluster with the best numbers in this aspect. In terms of defensive aspects, they do not stand out in any of the metrics studied: duels won (13.26), fouls committed (39.84), and pressure actions (359.73).

In short, the player described in this cluster is a type of forward who does not actively participate in the team's passing game but has a clear finishing profile, with great effectiveness in front of goal. Some prominent names in this group are Iago Aspas (Celta de Vigo), Edin Džeko (AS Roma), Aleksandar Mitrović (Newcastle United) or Olivier Giroud (Arsenal).

CHAPTER 5

Streamlit Application

In this chapter, the Streamlit application developed for the visualization of match reports is presented.

One of the tasks of the coaching staff and the analyst team is to prepare for matches by analyzing the opponents. From this need arises the development that will be explained in this chapter. As in the study, the data provided by Statsbomb [23] is used to create a tool that allows us to visualize different aspects of the matches. As mentioned above, the application is developed with Streamlit [26], an open-source Python library that allows the creation of web applications for data projects.

Data corresponding to the leagues used in the analysis in the previous chapter (La Liga, Bundesliga, Premier League, and Serie A) are available in the application, all with data corresponding to the 2015/16 season. The idea of development is to be able to have a report of each of the matches played in that season for each of the aforementioned leagues. The report contains information about the match; the data and visualizations that can be consulted are as follows:

- The final score of the match.
- The lineups of both teams.
- Pass map.
- Heat map.
- Shots map.
- Passing network.

Next, the structure of the application, its views, and different options will be described. First, there is a drop-down menu on the left side where there are different filters, as shown in Figure 5.1. The first is the filter to select the league to be displayed. As mentioned above, the leagues that can be selected are as follows. La Liga, Bundesliga, Premier League, and Serie A; all with data from the 2015/16 season. The second of the available filters corresponds to the selection of the day of the championship. And finally, the third filter allows you to select the match you want from the match day selected in filter 2. Once the match is selected, the report information is displayed, as shown on the right side of 5.1, where the result of the match and the lineups of both teams can be seen.

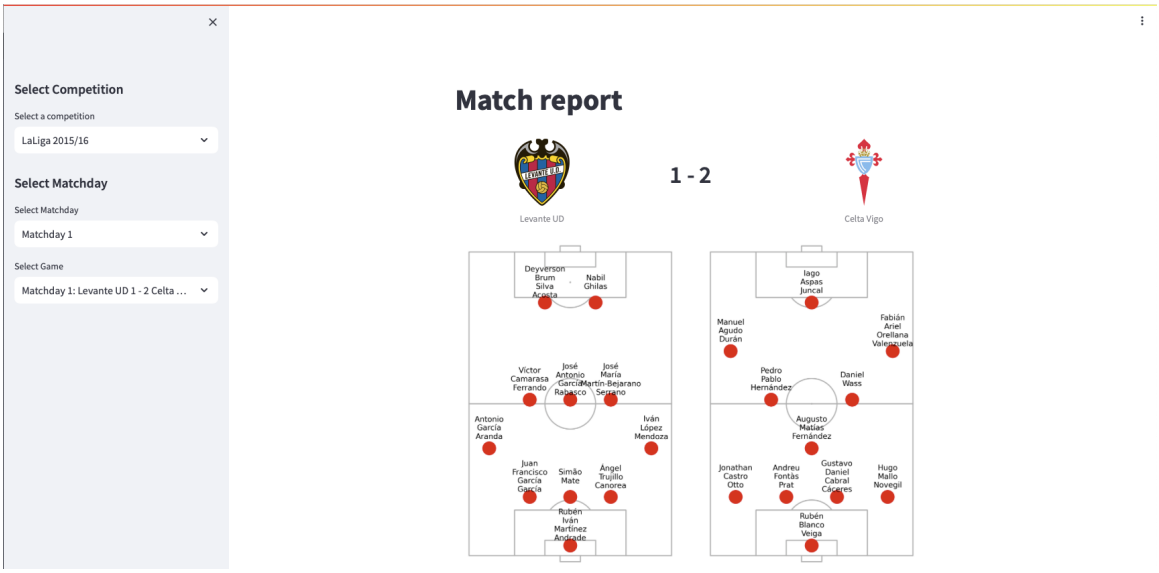


Figure 5.1: Drop-dpwn menu with competition, matchday and match selection.

Below the team lineups, another set of filters is displayed, allowing you to choose the team, player, and type of visualization that you want to show.

Figure 5.2: Filters to choose team, player and display type.

The first type of visualization is the one shown in Figure 5.3, the pass map. This display allows you to see the passes that a player has made, differentiating between completed (red arrow) and incomplete (grey arrow), as well as the different areas of the field from which each of them was made.

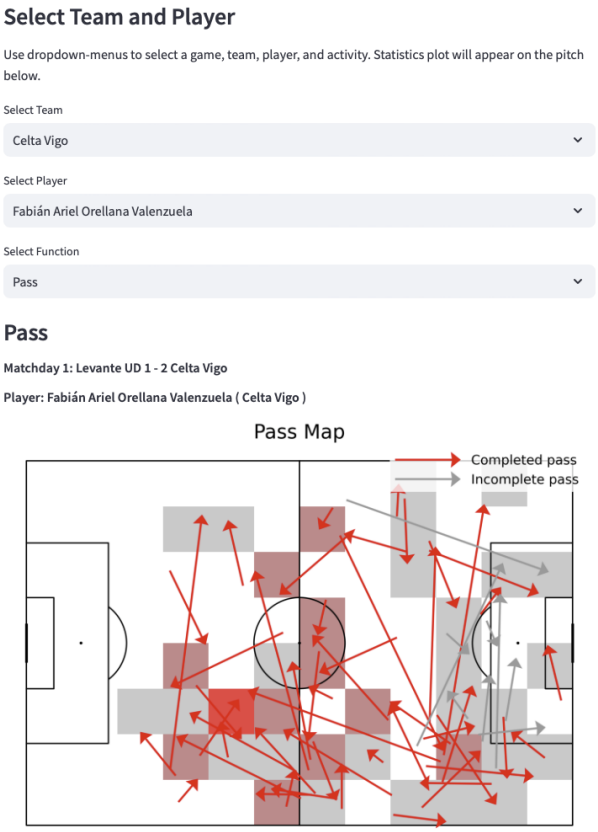


Figure 5.3: Passmap

The next of the available visualization options is the heat map, as shown in Fig. 5.4. This display allows you to know the areas of the field where the player usually moves, indicating the intensity of the heat by the percentage of time spent in each of the areas.

Figure 5.4: Heatmap

The third option is the shooting map, which can be seen in Figure 5.5. This display is similar to the pass map; it allows one to visualize the shots taken by the player, differentiating between those that go to the goal (red arrow) and those that don't (grey arrow).

Figure 5.5: Shotmap

Finally, there is the type of network passing visualization, as shown in Fig. 5.6. This type of visualization allows one to see how the players are related within the field. First, the average position from which the different players on the team have made their passes is represented by a circle; the larger the circle, the greater the number of passes they have made. Then, these circles are joined together according to the number of passes that have been made between players; the greater the number of passes, the thicker and browner the line will be. In addition, a table has been added below the display that allows you to see the number of passes made by each player in the match.

Select Team and Player

Use dropdown-menus to select a game, team, player, and activity. Statistics plot will appear on the pitch below.

Select Team

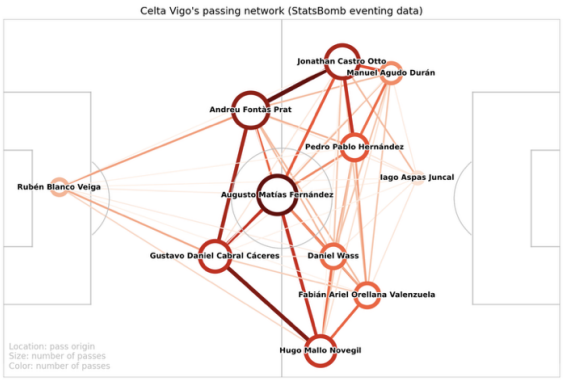
Celta Vigo

Select Player

Fabián Ariel Orellana Valenzuela

Select Function

Passing Network



	num_passes
Andreu Fontàs Prat	65
Augusto Matías Fernández	72
Daniel Wass	38
Fabián Ariel Orellana Valenzuela	39
Gustavo Daniel Cabral Cáceres	54
Hugo Mallo Novegil	52
Iago Aspas Juncal	9
Jonathan Castro Otto	60
Manuel Agudo Durán	30
Pedro Pablo Hernández	43
Rubén Blanco Veiga	20

Figure 5.6: Passing network

CHAPTER 6

Conclusions

This chapter describes the goals achieved by the master thesis following some of the key points developed in the project.

6.1 Achieved Goals

The goals achieved for this project are the following.

- A study of the different existing techniques in data analysis and machine learning applied to professional football has been carried out. This study covers the investigation of the different data providers that currently exist in the market and their main features to the review of some studies related to the analysis of player and team performance.
- To analyze the performance of players and teams, a machine learning algorithm has been applied to group them. This process has allowed a more detailed understanding of all phases of this process, problem definition, data processing, modeling, and evaluation. This clustering has allowed us to compare teams and players, finding similarities and differences between them.
- A tool has been developed that generates reports for soccer matches. This development allows us to obtain information on the different matches studied in order to facilitate the work that can develop the staff of a team when analyzing rivals.

6.2 Conclusion

To conclude this thesis, we shall make a recapitulation of the entire process. The realization of this process has allowed me to conduct a more detailed study of how data analysis and machine learning algorithms are applied in professional football.

First, a study of the different available data providers and their main characteristics was carried out, with the aim of studying which one was the best suited to the conditions necessary to carry out this project. After this, through a review of different available works, the characteristics to be studied in order to perform a performance analysis of both teams and players were proposed. These characteristics try to describe, in the case of teams, the three main phases of the game (defense, construction, and attack), and for players, each of the team's lines (goalkeeper, defence, midfielder, and forward). To obtain these characteristics and metrics, it was necessary to analyze and process the data provided by the supplier. Once these characteristics were obtained, an analysis of the correlation between them was carried out, as well as the standardization of the data and the study, using the elbow method, of the number of clusters that would be necessary to use in the K-Means algorithm. With all this, the algorithm was applied to the data and the different resulting clusters were evaluated.

On the other hand, with the aim of offering a tool that facilitates the work done by the technical and analytical staff of soccer clubs, an application has been developed in Streamlit. This application allows for the visualization of reports with information related to soccer matches. It shows information such as the result of the match and team line-ups, as well as various types of visualization such as heat maps, passing and shooting maps, and maps that allow visualizing the passing network generated between the different players in the match.

In short, the development of this project has allowed me to go deeper into data analysis and application of machine learning algorithms in a field with so much potential as professional football.

6.3 Future work

In this section, some future lines of work will be named on which to iterate this project.

- Creation of a system for recommending players based on certain parameters, which helps the scouting work carried out by football clubs.
- Streamlit application deployment to make it accessible. As well as other types of improvements in terms of data visualization, such as the use of real-time data.
- Integration of biometric data analysis to better understand the workload performed by soccer players, which could lead to the creation of a predictive injury model.
- Creation of a model to simulate different game strategies and measure their impact on the outcome of matches.

APPENDIX A

Impact of this project

A.1 Social Impact

Data analytics and the use of machine learning algorithms as used in the development of this project can help democratize football knowledge, improving the accessibility of data allows both amateur analysts and the general public to access information that was previously only available to professionals with many resources. By providing more detailed information on the game, the general public can gain a deeper understanding of tactical issues.

From the point of view of professional clubs, the use of these types of developments and tools can level the playing field between clubs with large economic differences, since the use of data complements the scouting work carried out, making it possible to find promising players that smaller clubs can afford to buy. In addition, these tools also help the coaching staff prepare for matches and analyze opponents.

A.2 Economic Impact

The economic impact of the project that incorporates advanced data analysis and machine learning techniques in professional football is significant and includes several aspects. Such as the optimization of resources in clubs, improving the identification of talent which leads

to a better management of the transfer budget; the improvement of team performance can lead to a growth of popular interest generating an increase in the social mass of the club, as well as better income in terms of results in the different competitions.

A.3 Environmental Impact

This project does not represent a substantial change from an environmental point of view. The only important point in this respect is the amount of energy consumed in the development, deployment, and maintenance of such a project. Projects involving large amounts of data are usually based on a cloud infrastructure. Data centers used in the cloud have quite high energy consumption, which has a negative impact on the environment. This is something that needs to be worked on to improve the efficiency of these data centers.

A.4 Ethical Implications

The use of advanced data analytics and machine learning techniques in professional soccer, as explored in this project, has some important ethical implications that must be carefully considered. Data must be used in a fair and transparent manner, avoiding biases that may discriminate players. Additionally, decisions influenced by the analysis must be transparent to all involved; it is important to understand how and why such decisions have been made. Finally, a responsible use of the data should be made, avoiding the reduction of players to mere data sets, recognizing their value more than the value of the data.

APPENDIX B

Economic budget

B.1 Human resources

This section will provide an estimate of the human resources required to develop this project. In order to make this estimate, we have used the ECTS credits for the final master's degree project. There are a total of 30 ECTS, each of these credits corresponds to about 25 hours of work, making a total of about 750 hours of work. Assuming that the hourly rate for a junior telecommunications engineer is about 15 euros, this adds up to 11250 euros.

B.2 Physical resources

In order to develop the project, a MacBook Pro 2020 laptop computer with the following technical specifications has been used:

- Processor: 2 GHz Quad-Core Intel Core i5.
- Graphics: Intel Iris Plus Graphics 1536 MB.
- Memory: 16 GB 3733 MHz LPDDR4X.
- OS: macOS Sonoma 14.2.1.

The price of this laptop, if purchased in the year in which this project was developed, would be approximately 1200 euros.

B.3 Licenses

All the software used in this project is open source, so there has been no additional cost.

B.4 Total Budget

In short, the total cost of the project will be the sum of human resources and physical resources, which amounts to 12450 euros. Knowing that the project has been developed and commercialized in Spain, to this figure must be added the corresponding applicable VAT, which in this case is 21%. Therefore, the final budget will be 15064.5 euros.

APPENDIX C

Modelling K-Means

Here are the correlation matrices and the result of the analysis of the elbow method for each of the positions (goalkeeper, defender, midfielder and forward).

C.1 Goalkeeper

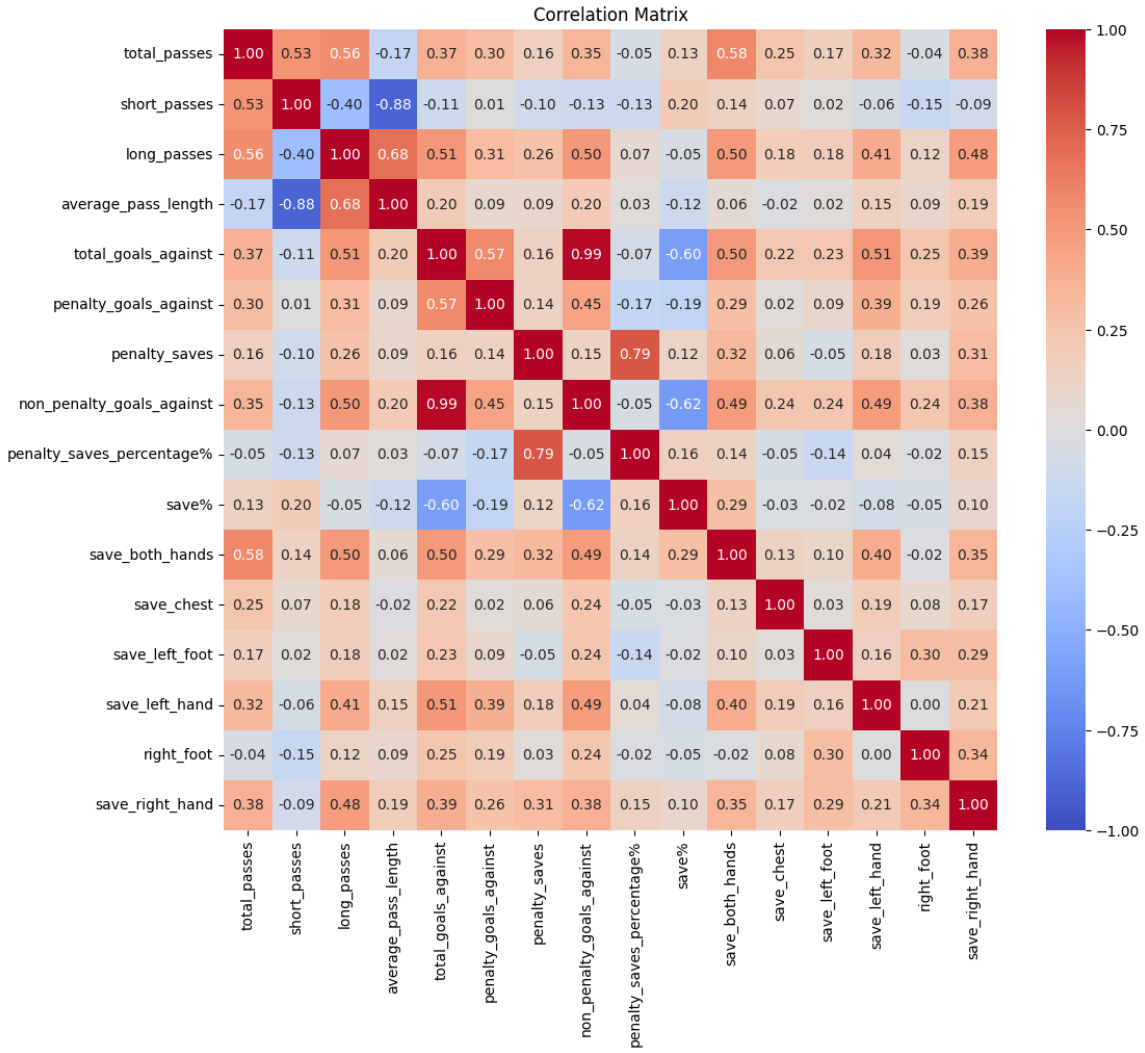


Figure C.1: Correlation matrix, goalkeepers

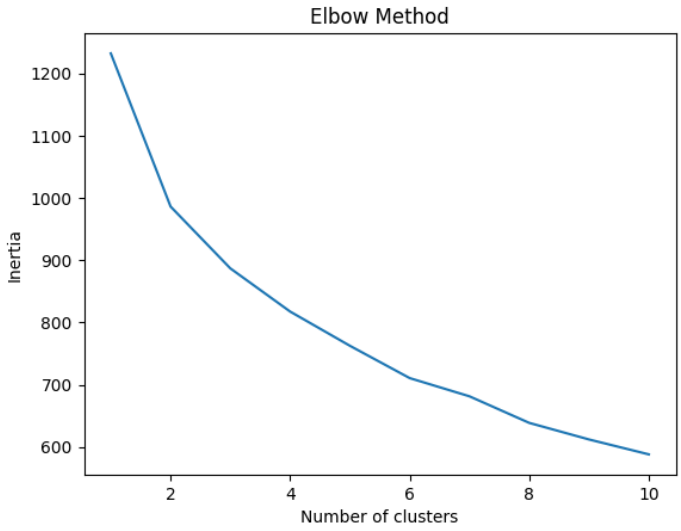


Figure C.2: Elbow Method, goalkeepers

C.2 Defender

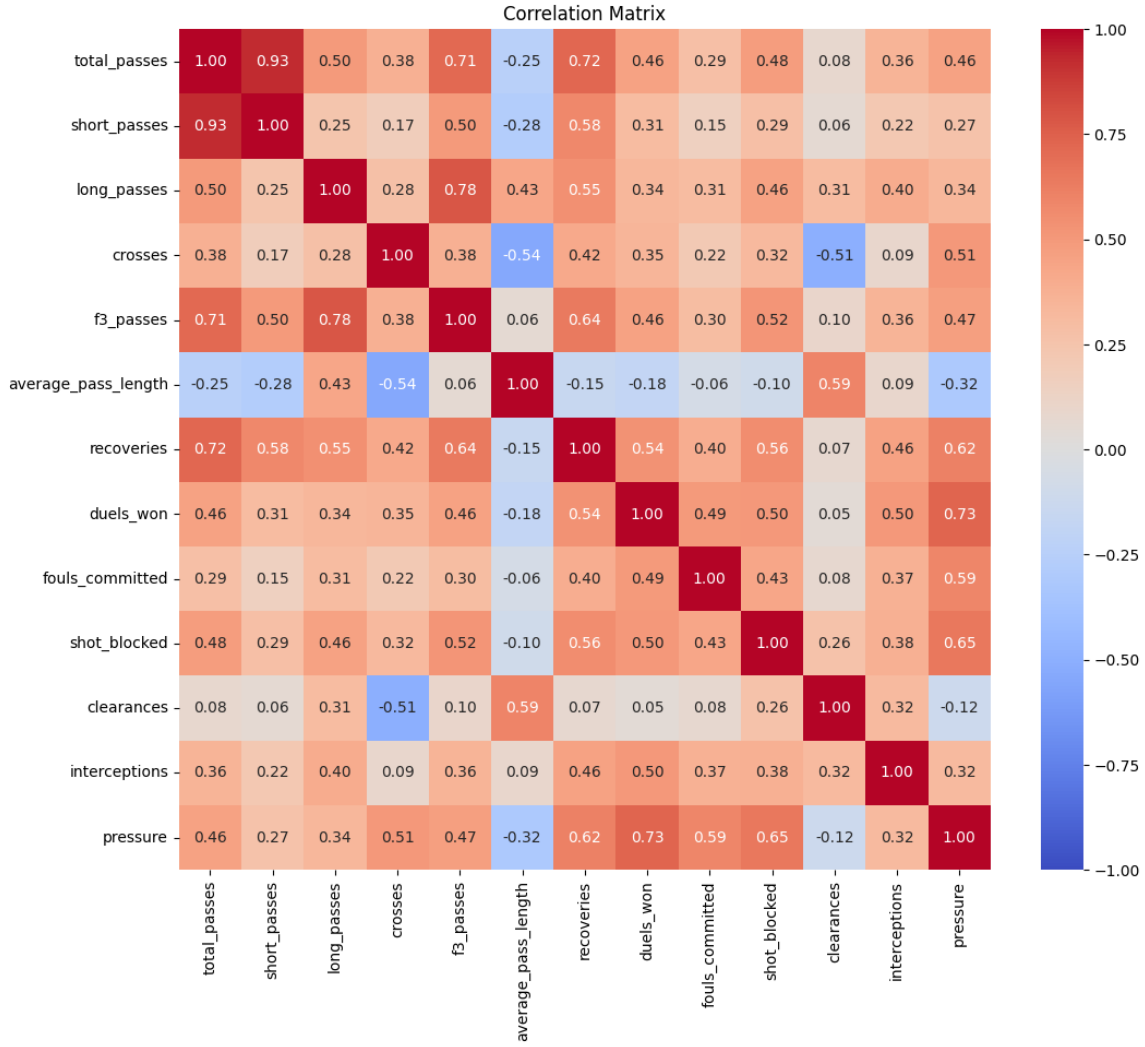


Figure C.3: Correlation matrix, defenders

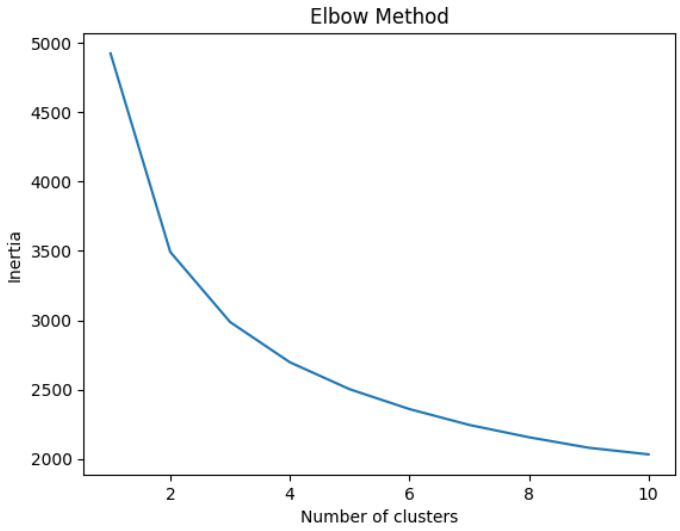


Figure C.4: Elbow Method, defenders

C.3 Midfielder

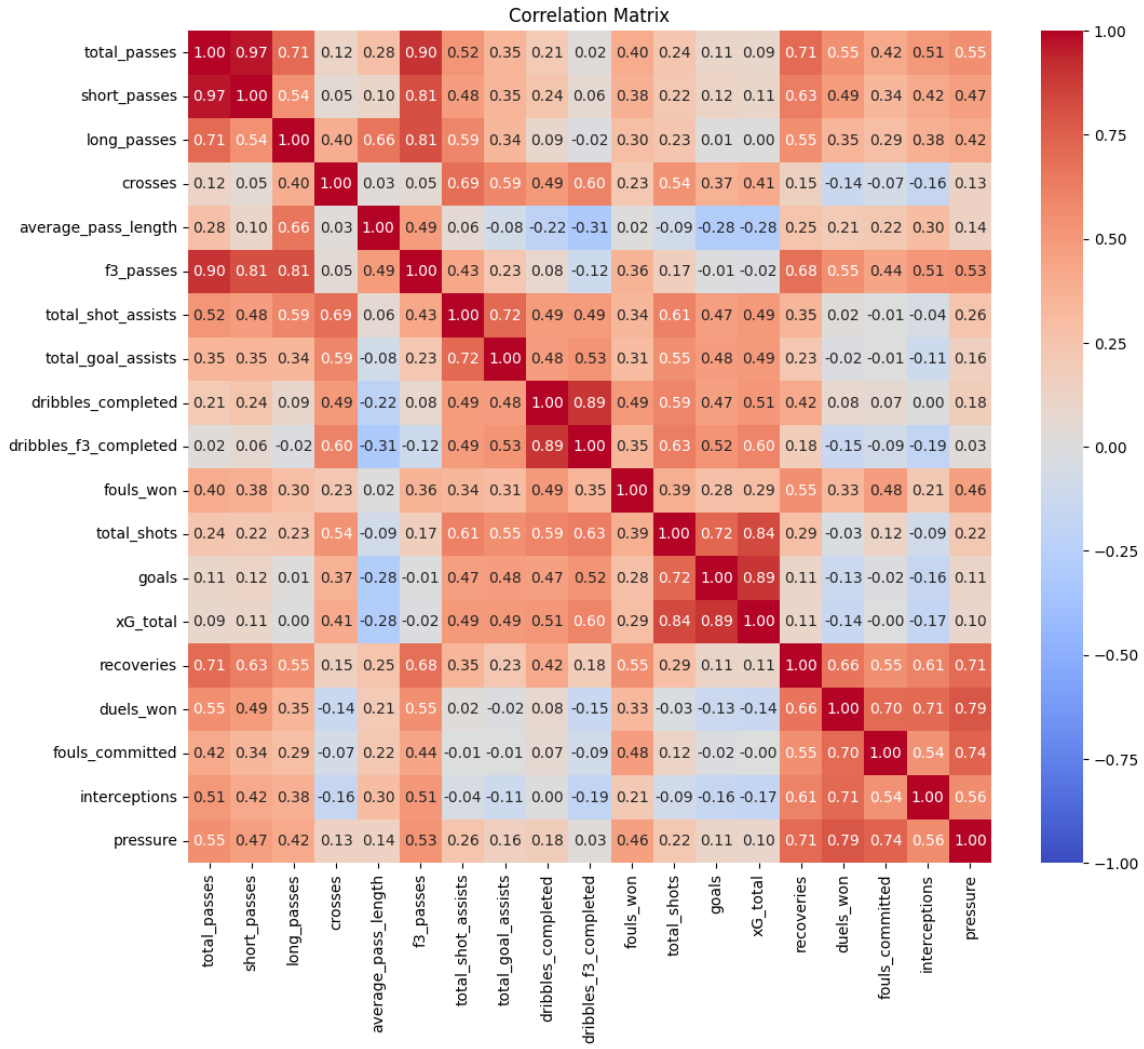


Figure C.5: Correlation matrix, midfielders

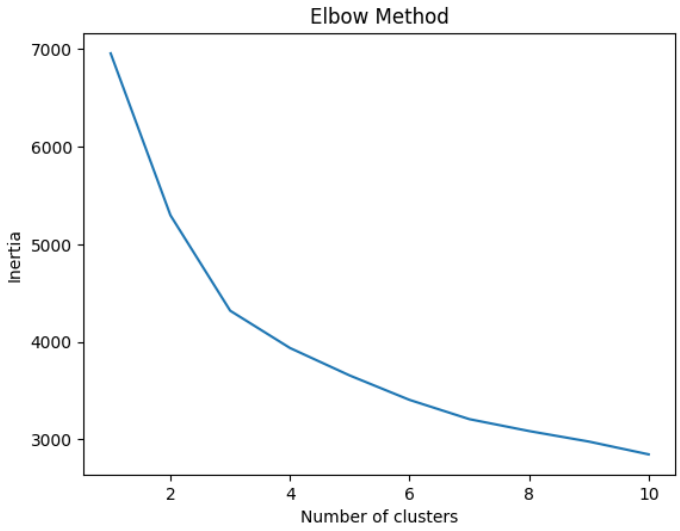


Figure C.6: Elbow Method, midfielders

C.4 Forward

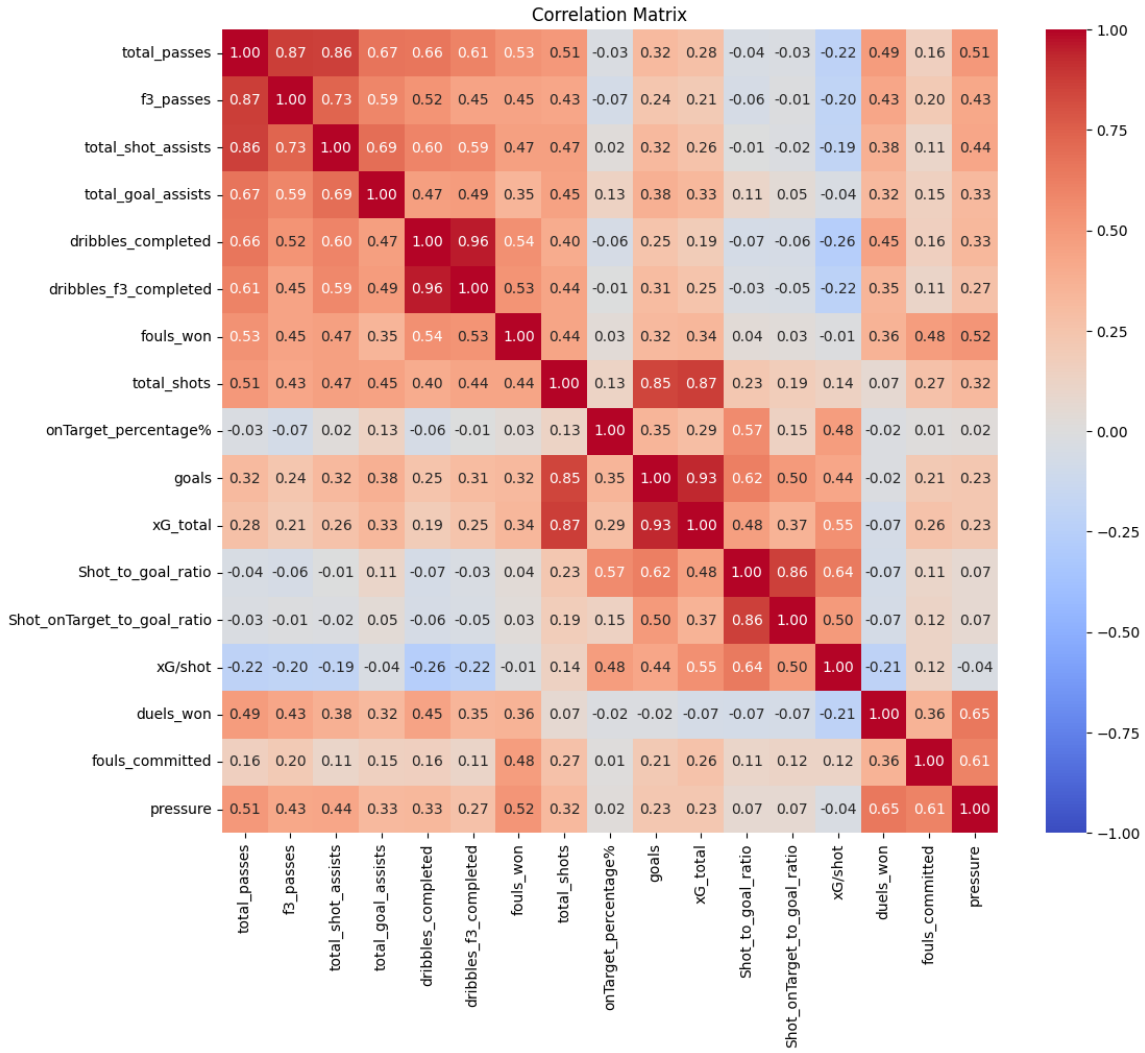


Figure C.7: Correlation matrix, forwards

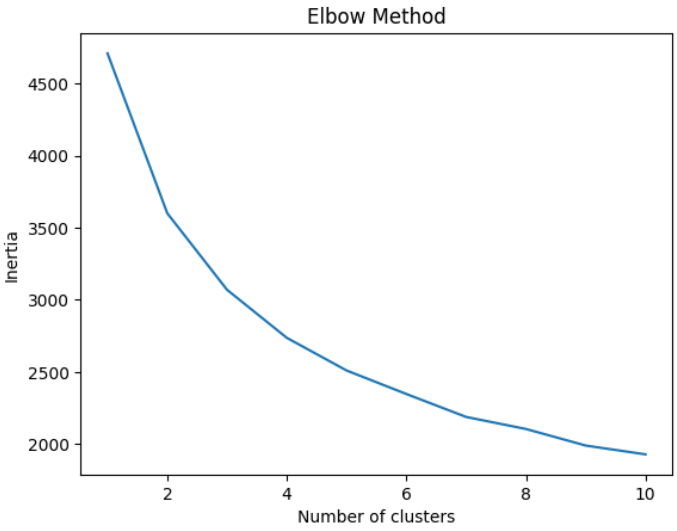


Figure C.8: Elbow Method, forwards

Bibliography

[1] Ricardo André. K-means for player clustering. Available at https://medium.com/@ricardoandreom/k-means-for-player-clustering-2d-3d-22df7d12f1a9. Access date: October 20, 2022.

[2] Rodrigo Aquino, Enrico F. Puggina, Isabella S. Alves, and Júlio Garganta. Skill-related performance in soccer: a systematic review. *Human Movement Special Issues*, pages 3–24, 2017.

[3] Rio Rizki Aryanto. Clustering nba player using k-means. Available at https://medium.com/nerd-for-tech/clustering-nba-player-using-k-means-7b568830edfd. Access date: October 20, 2021.

[4] Luca Carloni, Andrea De Angelis, Giuseppe Sansonetti, and Alessandro Micarelli. *A Machine Learning Approach to Football Match Result Prediction*, pages 473–480. Springer, 07 2021.

[5] Olivier Caya and Adrien Bourdon. A framework of value creation from business intelligence and analytics in competitive sports. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1061–1071, 2016.

[6] Left Field Football Consulting. Analytics in football: State of play, 2023.

[7] Karthik Garimella, Rishave Kumar, and K Kalaiselvi. Prerequisites for winning a league using data analytics. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, 11(4):116–139, 2021.

[8] Alexandre Gramfort, Mathieu Blondel, Olivier Grisel, Andreas Mueller, Eric Martin, Giorgio Patrini, and Eric Chang. Standard scaler. Available at https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Access date: November 24, 2023.

[9] Michael Hughes, Tim Caudrelier, Nic James, Athalie Redwood-Brown, Ian Donnelly, Anthony Kirkbride, and Christophe Duschesne. Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position. *Journal of Human Sport and Exercise*, 7(2):402–412, 2012.

[10] Sergio Llana. Passing networks in python. Available at https://github.com/Friends-of-Tracking-Data-FoTD/passing-networks-in-python/tree/master. Access date: October 13, 2023.

[11] Matplotlib. Available at https://matplotlib.org. Access date: September 28, 2023.

[12] Mplsoccer. Available at https://mplsoccer.readthedocs.io/en/latest/#. Access date: September 30, 2023.

[13] Numpy. Available at https://numpy.org. Access date: September 27, 2023.

[14] Pandas. Available at https://pandas.pydata.org. Access date: September 27, 2023.

[15] Luca Pappalardo, Paolo Cintia, Dino Pedreschi, Fosca Giannotti, and Albert-Laszlo Barabasi. Human perception of performance, 2017.

[16] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel. Scikit-learn. Available at https://scikit-learn.org/stable/. Access date: September 27, 2023.

[17] José Maria Pratas, Anna Volossovitch, and Ana Isabel Carita. Goal scoring in elite male football: A systematic review, 2018.

[18] Python. Python software foundation. Available at https://www.python.org. Access date: September 27, 2023.

[19] Sarah Rudd. A framework for tactical analysis and individual offensive production assessment in soccer using markov chains. In *New England symposium on statistics in sports*, 2011.

[20] Hugo Sarmento, Rui Marcelino, M. Teresa Anguera, Jorge CampaniÇo, Nuno Matos, and José Carlos LeitÃo. Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20):1831–1843, 2014. PMID: 24787442.

[21] Karun Singh. Introducing expected threat (xt), 2023.

[22] William Spearman, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. Physics-based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*, 2017.

[23] StatsBomb. About StatsBomb. Available at https://statsbomb.com/who-we-are/. Access date: September 18, 2023.

[24] Statsbomb. Statsbomb open data. Available at https://github.com/statsbomb/open-data. Access date: October 2, 2023.

[25] Statsbomb. What are expected goals (xg)? Available at https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/. Access date: October 11, 2023.

[26] Streamlit. A faster way to build and share data apps. Available at https://streamlit.io. Access date: September 27, 2023.

[27] David Sumpter. Centralization index. Available at https://soccermatics.readthedocs.io/en/latest/gallery/lesson1/plot_PassNetworks.html. Access date: October 13, 2023.

[28] David Sumpter. Statistical scouting. Available at https://soccermatics.readthedocs.io/en/latest/lesson3/ScoutingPlayers.html. Access date: October 14, 2023.

[29] Colin Trainor. Defensive metrics: Measuring the intensity of a high press. Available at https://statsbomb.com/articles/soccer/defensive-metrics-measuring-the-intensity-of-a-high-press/. Access date: October 14, 2023.

[30] Michael Waskom. Box plot. Available at https://seaborn.pydata.org/generated/seaborn.boxplot.html. Access date: January 10, 2024.