TÉCNICO LISBOA

Assessing Team Success in the Big-5 European Football Leagues

José Paulo Direito Fonseca

Thesis to obtain the Master of Science Degree in

Industrial Engineering and Management

Supervisors: Prof. Maria do Rosário de Oliveira Silva
Prof. João Manuel Ferreira Ribeiro

Examination Committee

Chairperson: Prof. Maria Margarida Catalão de Oliveira Pina
Supervisor: Prof. Joao Manuel Ferreira Ribeiro
Members of the Committee: Prof. Micael Santos Couceiro

November 2022

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

The present document is the final realization of my academic path that started in the Fall of 2017. Five years, several projects, tests, exams and some little-sleep nights later, I now realize that a lot of people contributed to making this work possible.

Firstly, I would like to thank Prof. Maria Oliveira and Prof. João Ribeiro for sharing their knowledge and guidance to make this dissertation possible.

Then, to all my friends with whom I had the opportunity to learn and share experiences over these last years, particularly Dinis, Miguel, and Isaac.

Also, a wholeheartedly thank you to all my family members I knew I could rely on, and particularly to primo Zé Alberto and prima Maria for the help, kindness, and generosity during these years.

Finally, and most importantly, to my parents, Paulo and Maria José, and sister Inês, the dearest people I hold in my heart, as they were always present in every defeat and victory throughout my life.

Abstract

This work aimed to identify a set of Key Performance Indicators suggestive of team success and also identify metrics that provide insights into the distinctions between the Big-5 European Football Leagues. Additionally, assess the performance of several classifiers in estimating the final ranking of teams for the 2021/22 season according to the three categories of success.

Data from the FBREF website respecting the Big-5 leagues' teams during a period of four consecutive seasons (2017/18 - 2020/21) was considered and analysed via Principal Components Analysis, Cluster Analysis and classification methods.

Significant differences were obtained between high-ranked teams and the remaining categories in terms of higher ball possession, presence in the attacking third of the pitch and higher number of passes leading to goal and shot attempts. Low-ranked teams displayed salient patterns in defensive statistics, despite some cases where these teams purposefully adopt a low-intermediate block as an attempt to explore the space behind the defensive line through counter-attack. Results from the same study suggest a higher strong overall presence in all areas of the pitch for English and German teams. Additionally, Italian Serie A clubs showed a prominent use of the defensive section of the field, while findings also suggest that together with Italian clubs, the Spanish teams committed the highest amount of infringements and received the highest amount of yellow and red cards.

Finally, results from the classification methods showed that ensemble techniques achieved the best results, with XGBoost leading with an accuracy of 86.7%.

Keywords: Key Performance Indicators; Principal Component Analysis; Classification; Big-5 European Football Leagues; Performance Analysis; Football.

Resumo

Esta investigação visa identificar um conjunto de Indicadores-chave de desempenho que evidenciem sucesso para as equipa e identificar métricas sugestivas de diferenças entre as cinco Grandes Ligas Europeias de Futebol. Adicionalmente, avaliar o desempenho de vários classificadores na estimativa da posição final das equipas para a época 2021/22 de acordo com três categorias de sucesso.

Foram considerados dados do website FBREF respeitantes às ligas referidas durante um período de quatro temporadas consecutivas (2017/18 - 2020/21), analisando-os através de Análise de Componentes Principais, Análise de Cluster e métodos de classificação.

Foram obtidas diferenças significativas entre as equipas de alto-nível e as restantes categorias em-termos de maior posse de bola, presença no terço ofensivo e maior número de passes que levam a tentativas de golo e remates. As equipas de baixo-nível apresentaram padrões salientes em estatísticas defensivas, excluindo casos em que estas equipas adoptam propositadamente um bloco intermédio-baixo como uma tentativa de explorar o espaço atrás da linha defensiva através do contra-ataque. Resultados do mesmo estudo sugerem uma presença global mais forte em todas as áreas do campo para as equipas Inglesas e Alemãs. Além disso, clubes da Serie A Italiana mostraram uma utilização proeminente da secção defensiva do campo, e que, juntamente com clubes da liga espanhola cometem a maior quantidade de infracções e recebem a maior quantidade de cartões amarelos e vermelhos.

Finalmente, os resultados dos modelos de classificação mostraram que os métodos ensemble alcançaram os melhores resultados, com o XGBoost a liderar com uma exatidão de 86,7%.

Palavras Chave: Indicadores-chave de desempenho; Análise de componentes principais; Classificação; 5 Grandes Ligas Europeias de Futebol; Análise de desempenho; Futebol.

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

AB Adaptive Boosting.

AI Artificial Intelligence.

BMI Body Mass Index.

BN Bayesian Network.

DS Data Science.

EPL English Premier League.

FIFA Fédération Internationale de Football Association.

FLO French League One.

GB German Bundesliga.

GNB Gaussian Naïve Bayes.

ISA Italian Serie A.

KNN K-Nearest Neighbors.

KPIs Key Performance Indicators.

LR Logistic Regression.

ML Machine Learning.

PC Principal Component.

PCA Principal Component Analysis.

RF Random Forest.

ROC Receiver Operating Characteristic.

SC Silhouette Coefficient.

SLL Spanish La Liga.

UCL UEFA Champions League.

UECL UEFA Europa Conference League.

UEFA Union of European Football Associations.

UEL UEFA Europa League.

XGB Extreme Gradient Boosting.

Chapter 1

Introduction

In this chapter, the context of the problem and the motivation for the study are introduced. Additionally, the purpose and objectives of the dissertation are presented and the structure of the document is outlined.

1.1 Contextualization and Motivation

Most industries throughout the world have adopted the expanding trend of employing data collecting and analysis to make better-informed decisions. Similarly, the sports industry has discovered in Data Science (DS) the power to improve various aspects that may affect all areas of the organization, which could ultimately provide them with the upper hand over its rivals through a competitive advantage(s).

In an industry where just the European football market size is estimated to be worth €27.6 billion as of 2020/21 [5], it is clear that the adoption of these practices might help sports stakeholders both inside and outside the field. Concerning the former, one area where DS is already being implemented is the analysis of match performance given that the identification of Key Performance Indicators (KPIs) most linked with team success could empower, for example, coaches to adjust game tactical strategy, and players' training. Concerning the latter, another area that has been receiving increased attention is the analysis of the distinctions between the different football leagues. Indeed, as clubs do not only compete in national competitions but must also participate in international exhibitions, understanding what characterizes each league could reveal important insights not only for coaches but also other decision-makers such as sports coordinators and heads of recruitment as to what are the best acquisitions for a team.

Nevertheless, the use of data and DS in sports has not come without its challenges [6], given that teams and game strategies are constantly evolving. Additionally, with data from both national and international competitions being produced every year, analysis of this type are always pertinent as they offer new fertile ground for insights to be derived.

1.2 Objectives

This dissertation aims at identifying a set of KPIs suggestive of team success and also identifying metrics that provide insights into the distinctions between the leagues, with the aim of characterizing them. For this purpose, this thesis analyses data from the FBREF website respecting the Big-5 leagues' teams during a period of four consecutive seasons (2017/18 - 2020/21) and targets the following objectives:

1. Identify a set of KPIs that distinguish teams on the basis of three categories of success according to the final ranking: clubs that win the championship and/or qualify for European competitions (high-ranked); clubs that safeguard their place in the league the next season (middle-ranked) and clubs that are relegated or qualify for the relegation play-off (low-ranked).
2. Identify a set of KPIs that differentiate teams on the basis of the league they are part of, with the purpose of characterizing them.
3. Construct a classification rule using data from the previous four seasons and evaluate its performance in estimating the final rankings for the 2021/22 season based on the success categories.

1.3 Structure

The present thesis is structured as follows:

- **Chapter 1 - Introduction**
  The contextual relevance and background of the problem investigated is provided, along with the dissertation objectives and structure.
- **Chapter 2 - Literature Review**
  A background is provided in terms of data and DS usage in football.
- **Chapter 3 - Methodology**
  A overview of the methodological approaches used to complete this dissertation is provided.
- **Chapter 4 - Business Understanding**
  Discusses the purpose of analyzing game-related KPIs and using classifiers.
- **Chapter 5 - Data Understanding**
  Investigates the raw data obtained from the various sources.
- **Chapter 6 - Data Preparation**
  Describes the actions taken to obtain the final dataset.
- **Chapter 7 - Modeling and Evaluation**
  Presents the results as well as its discussion.
- **Chapter 8 - Conclusion**
  Main conclusions of the thesis are discussed along with future work to be developed.

Chapter 2

Literature Review

In this chapter, a background is provided in terms of data and DS usage in football. Firstly, by describing when, how and why the handling of data began to be practised in sports generally, the chapter then explores pioneering methods and works for tracking some simple initial statistics and how the transition to the academic realm came to occur. Then, a more comprehensive and purposefully-oriented review of existing research is carried out in terms of investigations developed to distinguish teams based on their performance by looking into game-related statistics. Afterward, a revision of the literature that aims to compare and characterise the Big-5 European Football Leagues is carried out and, lastly, an examination of studies produced that focus on predicting football-related elements is performed.

2.1 Sports through the lens of Data Science

With the turn of the millennium, notably the late 2000s and early 2010s, as noted by [7], came noteworthy advancements in the fields of Artificial Intelligence (AI), DS, and Machine Learning (ML) empowered by advancements in infrastructure and computer science, which [8] defends to be caused by market forces and technological evolution. In essence, whether companies wish to focus on their business adversaries to study their processes and how they do things – outward perspective – or target their own activities, operations and mechanisms as a way to get insights into these internal processes with the aspiration of correcting, when needed, and improving them – inward perspective –, fulfilling the aforementioned requirements made possible for organisations across all industries to store and analyse complex data, guiding the way to more data-driven habits, decisions, and solutions. But what is ”Data”?

A formal definition of “Data” from the Online Cambridge Dictionary characterizes it as “information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer” [9] whilst the Online Oxford Learner’s Dictionary explains it as “facts or information, especially when examined and used to find out things or to make decisions” [10]. As made evident in both definitions, ”Data” has the underlying purpose of supporting and providing guidance towards solving the problem(s) at hand by means of better decisions.

Nowadays, the term has grown to be commonplace in every major sector as a result of it being associated with its ability to create competitive advantages, innovate and grow businesses. In fact, the value that these approaches bring in terms of business performance across all industries is unambiguous, even when the investment necessary is considered. This was precisely what [8] took under examination and, by looking into cases where these types of procedures had been implemented, the authors discovered that not only could organisations expect near but also long-term benefits in terms of overall growth. Likewise, as revealed by a study on data-oriented decision-making [11], the more data-driven an organization is, the more productive it is, and the better it performs on objective metrics of operational and financial success. Economist Erik Brynjolfsson and his colleagues from MIT and McKinsey's business technology office even went a step further and acknowledged that companies who used data-driven decision-making at a higher rate than the rest of their sector were, on average, 5% more productive and 6% more profitable than their rivals.

In sports, as of now, these advantages are also substantial and tangible as recognized by the myriad of success stories that proved the usefulness of data and its interpretation in this context. However, seeking to reap the benefits from statistical analysis, the path of implementing data gathering, assembling and how to do it – sports notation – would begin to be walked as early as the mid-nineteenth century [12].

With sports data usage history dating back to the 1860 publication of the baseball guide 'The Beadle's Dime Base-Ball Player' [13], in the book, the author succeeds at providing evidence that these practices emerged with the majority of the groundwork being created by sports journalists who attempted to convey their results through newspapers and specialized publications around the mentioned century. According with Simon Eaves' research [12], he emphasizes that even though some were designed for illustrative purposes, others saw the practices' promise as a tool for analysing sports performance with tennis match statistics surfacing as early as 1883 in England whilst by 1900, sharing sports statistics about game occurrences and events was already common practice. Indeed, throughout the world, there are recordings of cases like these. In the USA, for example, [14] carried a strokes analysis that tracked the number of aces, double faults and the returned in and out of court and total strokes played by 1891. Also that year, an assessment of a tennis game appeared in the New York Times that looked mainly into how players got aces (service that is delivered to the service box but is not touched by the opposition [15]). Likewise, such was the case with two tennis matches reports published in the New Zealand's Auckland Star (edition of 7^{th} November 1903) and the Australian Sydney Morning Herald (edition of 26^{th} March 1907) while records in France provide information about passing sequences leading to a goal, in football, and positional data of horses during a horse race. Looking back, these examples demonstrate that newspapers and similar types of publications provided the perfect and necessary initial fertile ground for these quantitative analysis to be developed over the following years, not only because they functioned as concrete, even though possibly not accurate in the beginning, records of events but also allowed for the practice of tracking statistics itself to propagate quite effectively.

Contrary to the growing prominence of sports-statistic in this format - newspapers - was the diminished attention that this matter got academically, around this time. Likewise, in football, the subsequent

decades also saw little investigation around related statistics with relevant work only being performed later in the century. In the view of [16], the scant amount of work carried during these years was one of the reasons that delayed the establishment of appropriate topic-oriented academic journals, around that time. Match analysis, however, has risen to prominence in the scientific literature since the 1990s thanks to the formation of international scientific societies, specialized journals, and world conferences like the International Society of Performance Analysis of Sport, Journal of Sports Science, and World Congress of Performance Analysis in Sport, as respective examples [3].

The popularity and rapid growth that these types of analysis experienced in the academic sphere in the years prior, along with the initially mentioned technological advancements in the years that followed the turn of the millennium (first paragraph), provided the ideal setting for these practices to be adopted even more conveniently in all sports throughout the world. This computer-powered statistical data analysis is known more simply as "Data Science", a term that has arisen in parallel with these breakthroughs. Respectively, the Online Cambridge Dictionary and the Oxford Learner's Dictionary define it as:

1. the use of scientific methods to obtain useful information from computer data, especially large amounts of data [17]

2. the scientific study of the creation, validation and transformation of data to create meaning [18]

One famous moment that is often mentioned as having revolutionised and popularised these practices came in 2002 when, faced with a budget lower than that of nearly every other team, the general manager of Major League Baseball Oakland Athletics, decided to study statistical information closely to assemble a team of athletes who, contrary to traditional measures, would play a steady, consistent, disciplined game, and ultimately perform better than flashier, more obviously gifted players. This pioneering work of basing and implementing decisions on the analysis of data rather than conventional thinking, later immortalized in the book "Moneyball: The Art of Winning an Unfair Game", provided concrete proof and ultimate confirmation of these practices, having an impact in all sports around the world [19].

2.2 Data and Data Science in Football

With three possible outcomes – win, defeat or tie –, association football, also known as "soccer" (American English term), is a team sport in which two teams of 11 players compete against one another to score goals. As simply as the definition might seem, nonetheless, it is usually hard to define or even comprehend the immense complexity of the world's most popular sport (since the late nineteenth century) due to its dynamism, intricacy and sophistication [20].

As mentioned previously, similarly to all other sports, the football and academic environment were shy in taking their first steps together, with significant research only being carried out later in the $20th$ century. However, one noteworthy exception is the work of Charles Reep, an acknowledged British football analyst, in 1968 entitled "Skill and Chance in Association Football" [21]. Often cited as one of the first academic research studies where data was taken into consideration to analyse the game, the

author concluded about the optimal number of passes leading up to a goal and the best way to score a goal, based on the analysis of more than 2000 games.

Notwithstanding subsequent studies having questioned Reep's conclusions by demonstrating that his results might not be in line with reality in every case, this analysis, regarded by several authors [12,22,23] as the earliest research developed in the fields of football analytics, was at the forefront and inspired succeeding research and practices carried out with the same purpose of comprehending the game. Thus, in the upcoming years, and then more vigorously in the early twentieth century, studies looking into the game from different perspectives were carried out, with match analysis has being reckoned as the most popular research topic [24].

To add to that, data analysis and DS have grown in popularity outside of academia as well with sports-organisations and more prominently clubs and federations opting and adopting solutions to track, quantify, and analyse individual and teams' performance outcomes internally. Whether the data is gathered and thus the property of the club – proprietary data –, a practice supported by [6], or third-party data, the reality is that almost every aspect of the game can be translated into a numerical and/or statistical format which allow insights to be derived. With all these advancements and new practices in latter years, the volume and storage requirements have escalated with [6, 7] even going further and saying that the sheer amount of data accessible for analysis might prove itself hard to handle, exceeding what most teams can now process - storing location, event, and video data from a single German Bundesliga (GB) season alone consumes 400 terabytes of storage space [25].

Given the ubiquitousness of the data, it is no wonder that [26] emphasizes in his research that, to serve as a base for interpretation and investigation of the relevant football domains, a multidisciplinary approach should be used to make sense of these complex data sets. Fundamentally, match performance and analysis in football is said to be dependent on the combination of tactical-technical, physical and mental elements [26–28]. However, even though these categories are a good way to organize the existing literature, their objectives can vary considerably. This was precisely the conclusion reached by [29] while conducting a comprehensive study of the existing literature on match analysis, later adapted by [3], who suggested that the best way to organise and explore existing literature around these topics would be to categorize it as shown in Figure 2.1.

Figure 2.1 distinguishes literature based on two levels of analysis: a first-order level based on the type of analysis undertaken (descriptive analysis, comparative analysis, and predictive analysis) and a second-order level based on the set of variables studied. The articles from the first set share the goal of describing players' activity patterns, while the comparative category focuses on establishing contrasts and/or connections among several domains. The last group, predictive analysis, gravitates towards the creation of models that allow the prediction of football-related elements.

Given the dissertation objectives explained in the introduction, the following chapters are organized according to research that aims to investigate KPIs associated with team success, the distinctions between leagues and pertinent classification and predictive analysis carried out in this field.

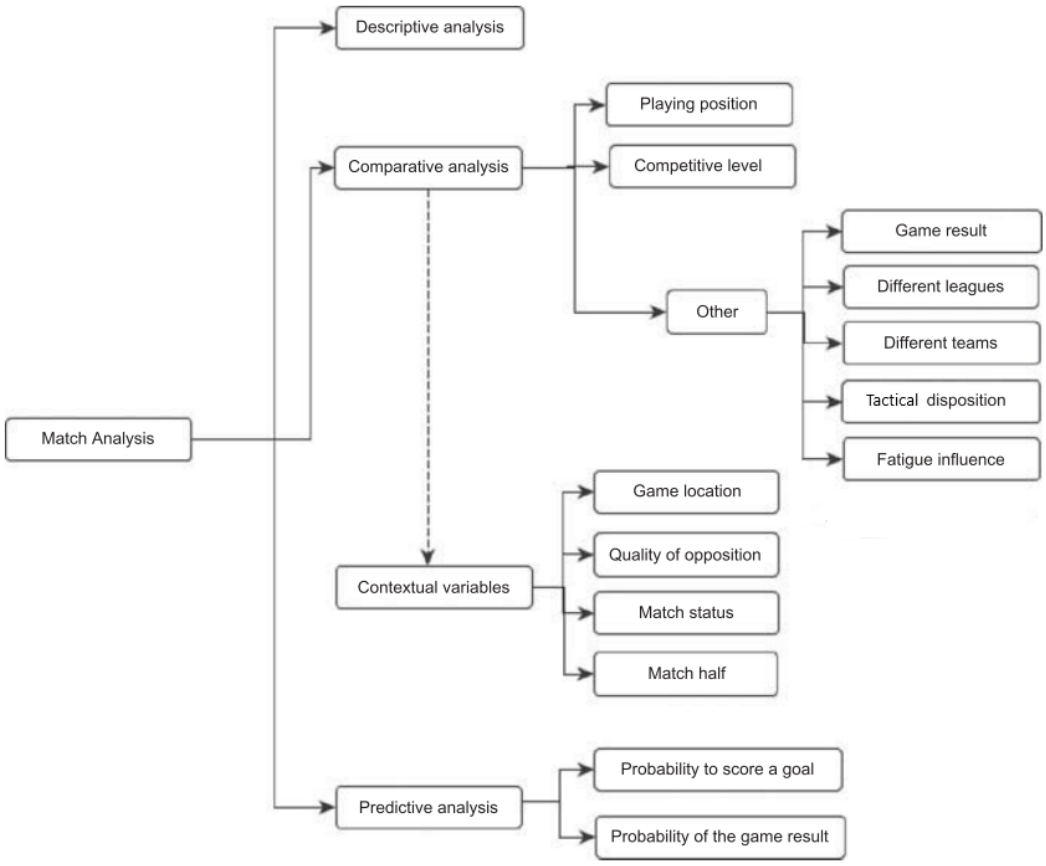

Figure 2.1: Scopes of match analysis (adapted from [3]).

2.2.1 Key Performance Indicators

The study of performance indicators from the technical, tactical, mental, and physical domains has drawn much interest, particularly from researchers and clubs who have tried to determine specific aspects that provide a portrait of successful performance. In the technical and tactical domains, studies conducted for this purpose have not only looked into games from international tournaments but also national leagues. Concerning the former, [30] showed that, at the Euro 2000, there were noticeable discrepancies between successful and unsuccessful teams in terms of ball possession, with the former registering higher levels of this metric. Likewise, by looking into a collection of games from the 1990 Italian and 1994 American World Cups, [31] suggested that longer passing combinations and a higher level of possession were the foundation of the successful teams (quarterfinalists) playing approach. Consistently, the authors found that unsuccessful teams' (first round losers) performances were marked by smaller passing patterns and more direct game strategies which, in turn, resulted in an inferior performance in terms of turning ball possession into shots on goal.

Curiously, contrasting results were found when [32] analysed the same data from the 1994 World Cup with the authors discovering no evidence that supported the relationship between ball possession and success in the competition. Equivalently, when studying the 1986 Fédération Internationale de Football Association (FIFA) World Cup eight years prior in Mexico, [33], coherently with results from [21], discovered that when penalties were not considered, approximately 80% of goals were preceded by two

passes or less with the minority developing after five or more passes. Similarly, results from analysing the 2001 Copa America competition data revealed that teams were less likely to take advantage of shooting opportunities and score goals the longer they had the ball in their control [34].

Defending the idea that equivocal conclusions had been reached by previous literature due to first, the small sample size of teams/games analysed and second, to the narrow scope of variables studied and the types of analysis carried, [35] investigated a total of 288 games from the UEFA Champions League (UCL) group stage over three seasons (2007/08 - 2009/10). After examining the data over a one-way analysis of variance (ANOVA) and a discriminant analysis, the authors concluded that higher mean values were reported in the effectiveness of converting shots to goals, ball possession and number of shots, shots on target, passes and successful passes performed for winning teams and higher values in terms of yellow and red cards for losing teams. The multivariate analysis further underlined the impact of ball possession and shots on goal in differentiating top teams. Identical conclusions in terms of these variables were reached by [36] after analysing matches from the 2014 FIFA World Cup competition in Brazil during the group stage. Additionally, other noteworthy findings from this study include variables such as shots taken from inside the opponent's area, short passes, tackles and aerial advantage as having a considerable effect on the likelihood of teams winning, with the number of red cards, dribbles and crosses metrics suggesting the opposite.

Other studies like [1, 37–40] have looked solely at one national league. In all cases, the authors made use of the season's final competition ranking as the method to establish the levels to distinguish the teams' measure of success, as some defend [40] would provide a more holistic understanding of the team and its performance rather than just looking at the three possible game outcomes.

With the objective of replicating results obtained by [30], [37] investigated 24 English Premier League (EPL) matches within the 2001/02 season (games from one national league). Conclusions reached validate the initial hypothesis with successful teams (first three) demonstrating a significantly longer possession than unsuccessful ones (bottom three) regardless of match status.

In their research, [38] examined 416 games from the Italian Serie A (ISA) league to contrast the technical and physical capabilities of the most and least successful clubs (ranked in the first and last five positions, respectively). The authors discovered that, as compared to the less successful teams, players from the more successful teams covered more distance with the ball in overall terms and in running in high intensity. Furthermore, they executed more successful short passes, tackles and dribbles besides registering higher values in terms of shots, and shots on target.

In line with these findings are the conclusions reached by [1] after studying a total of 380 games from the Spanish La Liga (SLL) during the 2008/09 season with the goal of discovering particular performance metrics that differentiate teams based on their final ranking. The metrics examined were initially divided into three categories: goals scored, offensive, and defence, however the one-way ANOVA study revealed only substantial insights for the first two. Concretely, considering a significance level of $p < 0.05$ (highlighted in bold in Table 2.1), respecting the first response variable (goals scored), the authors demonstrated that successful teams (four first-placed clubs) scored more goals, performed a greater number of shots and shots on target and had higher effectiveness (percentage of shots converted to

goal) than middle and bottom (four last-placed clubs) teams, who required more shots than any other group of teams to score a goal. Concerning the offensive metrics, statistically significant differences were also identified among top-ranked and middle teams, with the latter exhibiting a lower value in terms of ball possession and assists, as can be seen from Table 2.1.

Table 2.1: One-Way ANOVA results from the analysis of goals (1st group of variables), offense (2nd group of variables) and defense (3rd group of variables) in SLL during 2008/09 (adapted from [1]).

Variable	Mean values			F-value	p-value
	Top 4 clubs	Middle 12 clubs	Bottom 4 clubs		
Goals For	2.12	1.33	1.14	13.33	**0.000**
Goals Against	1.20	1.49	1.58	2.99	0.077
Total shots	16.25	12.41	12.93	9.09	**0.002**
Shots on goal	6.71	5.04	4.84	6.36	**0.009**
Shooting Accuracy	41.37	41.82	37.72	1.42	0.268
Shots for a goal	8	9.33	11.25	8.55	**0.003**
Assists	9.61	7.56	7.91	7.20	**0.055**
Crosses	29.06	28.56	28.78	0.05	0.948
Offsides Committed	2.66	2.68	2.54	0.09	0.915
Fouls Received	16.61	16.73	16.66	0.01	0.987
Corners	5.65	5.21	5.06	1.08	0.362
Ball Possession	55.57	48.34	49.04	6.14	**0.010**
Crosses Against	25.98	29.31	29.54	2.00	0.165
Offsides Received	2.90	2.44	3.02	1.48	0.255
Fouls Committed	15.60	17.22	16.31	1.65	0.221
Corners Against	4.90	5.42	5.18	0.92	0.416
Yellow Cards	2.73	2.95	2.94	0.53	0.596
Red Cards	0.22	0.27	0.23	0.59	0.567

Similar conclusions were reached in terms of effectiveness between successful and unsuccessful teams by [39] when assessing offensive actions from the Greek Super League 1 games over the course of ten seasons (1998/99 - 2008/09), with the former group (top-two teams) displaying increased action in the final/attacking area of the pitch with more shots performed inside penalty and assists made into this area when compared to the two-bottom teams.

Authors in [40], to identify key physical and technical performance variables related to team quality/strength, divided teams from the 2014 Chinese Super League into four groups: upper-ranked (1–4), upper-middle-ranked (5–8), lower-middle-ranked (9–12), and lower-ranked (13–16) teams. In terms of physical performance factors, the results were not highly expressive, with the sole notable finding being that the top-ranked category covered greater total distance when sprinting than the upper-middle-ranked group. Technically, teams in the top-ranked group had superior possession in the opponent's half, number of passes that entered the last third of the field, and penalty area. Similarly, middle-ranked teams outperformed lower-ranked teams in all prior criteria, with just possession time distinguishing them from lower-middle-ranked teams, leading the authors to conclude that teams should promote ball possession-focused styles of play over those that emphasize direct play.

When considering international tournaments, divergent conclusions were reached by [41] in terms

of the game's optimal strategy, after analysing almost 2000 goals or goal attempts events from where they happened in the opponent's half from the EPL. Main results reveal that around 30% of goals came as a consequence of set plays (situations where the ball is returned to open play typically by free-kicks, corners, throw-ins and penalties) and that the greatest portion of goals ($>$ 70%) occurred within the penalty area (estimation aligned with results from [42] that obtained a value of 79% for this statistic). Also noteworthy is, first, the finding that no relationship was found between the possession-related variables considered and the ability to score goals and second, that more than 70% of goals resulted from previously exchanging the ball four or less times, leading the authors to promote an offensive play method based on counterattack and/or fast attack.

Authors in [43], aimed at bridging the gap on the youth soccer league's limited research. They reached similar conclusions when identifying KPIs that would distinguish winning, drawing, and losing teams and, ultimately, determining the preferred ones in terms of forecasting the team's success. They did so by considering both defensive and offensive game-related statistics as well as the match location contextual variable from 46 matches played by one English Football League One soccer teams during the 2012/13 season. The results obtained were expressive in terms of the passes performed with the team performing a higher number in overall terms and in the opposing half when it lost compared to when they drew or won. Occurrences when the team drew seemed to only have a noteworthy impact on the successful passes percentage with this proportion being lower than matches where the team won. Thus, to be successful, these insights point toward a direct style of play with fewer passes with results also agreeing and emphasizing the relevance of performing fewer dribbles and more shots for the intended purpose. Nonetheless, the authors make the case that it is not reasonable to generalize these conclusions for every team, particularly given the highly disconcerting research. Thus, in their view, the teams' tactical decisions should be taken in line with the team's and their opponent's capabilities, meaning that a "direct" style of play may be preferable if the team's skill level is unsatisfactory to sustain meaningful possession.

Utilizing data from the same league with the addition of the two following seasons (2013/14 and 2014/15) to test their hypothesis on important metrics that distinguished between a successful and unsuccessful performances, as well as factors that best indicated success, [44] focused on contrasting the performances of solely one team across the periods of time considered. Oriented toward the goal of building a performance profile for a team and maybe measuring any tactical evolution, in total, 138 matches were investigated with results manifesting lower goal-scoring effectiveness and higher attempted and completed passes in the team's least fortunate season. These results lead the authors to advocate that teams should try fewer passes while making sure that more of these passes are successful and also focus on the quality (effectiveness) and not the quantity if the shots taken.

As is visible from the previous paragraphs, the scope of the investigations is complete in terms of competitions analysed - both international and national leagues are explored extensively even though literature is still limited in terms of female and youth leagues. The majority of literature to distinguish teams based on performance metrics, however, still focuses on the long-time studied metrics such as possession, the number of shots and passes, its effectiveness and other offensive statistics, with po-

larizing conclusions still being reached, leading the authors to advocate/discourage opposing tactical approaches to playing the same game.

Contrastingly, to this date, insufficient emphasis has been placed on other aspects of the game with limited research being available with the objective of looking into variables respecting goalkeeping, defensive actions, types of passes (rather than the number) and the location from where these took place in the game. Thus, it might be the case that by extending the sphere of analysis to include these types of statistics, other relevant patterns might emerge along side with an improved comprehension of the game.

2.2.2 Big-5 European Football Leagues Comparative Analysis

The literature is not vast in terms of comparative analysis that considers all five major leagues - EPL, ISA, SLL, French League One (FLO), GB - as the unit of investigation (Figure 2.1) with the objective of contrasting them and characterizing their differences, as agreed by [3] in their systematic review and as pointed out by [45,46]. Over the years, it is a fact that football-related research on a league level has not followed the more avidly analysed match performance statistics from individual competitions. If on one hand, this comes as no surprise since according to [10], these practices have proven to contribute to success inside the field, other justification, as defended by [47], for why investigations around this topic being withhold is the methodological barrier that affect researchers' capacity to perform an impartial and correct evaluation due to the inconsistency of how data is collected and lack of homogeneous analysis methods among leagues. However, there still is some interesting research worth looking into.

Based on the cultural, historical, and social differences that exist among all countries from Germany, Italy, Spain, France, and England, it is obvious and simple to argue that these backgrounds provide distinct environments for players to be formed and trained, as well as competitions to take place. As a result, distinct individual match-play styles emerge in the leagues, making certain players more fit than others in terms of their ability to compete in each one. This was precisely what [45] investigated. In their research, to find which league had the highest quality players, the authors looked into data from the EPL, ISA, SLL and GB during the 2001-2002 season. The knowledge retrieved considered not only the players' positional information but also their international status (nationality and FIFA world ranking, number of international appearances and international goals scored). Additionally, information about the players such as age, stature, body mass, and Body Mass Index (BMI) was gathered since the authors also hypothesized that there would be differences in the players' biometric/anthropometric data between the different positions, across these leagues. Indeed, results obtained not only validate the later premise by making distinctions evident across leagues (suggesting distinct physical demands in each one), but they also highlight patterns when the focus is placed on the four positional categories. Concretely, SLL contained the highest quality players according with the FIFA world rankings, with ISA and EPL being second and third, respectively. GB scored the lowest in the player quality metric but, in contrast, players from this league showed highest mean values in stature, body mass, and BMI of the four leagues.

Having taken on the challenge of analyzing football game summaries and media articles from four

countries (Italy was not considered in the analysis) to comprehend the factors involved in the establishment of national stereotypes, [48] reached similar conclusions, particularly in terms of the GB players. In the end, all European media descriptions analysed create the German depiction in the same terms, characterizing them as having aggressive power and substantial confidence in their strategy while also being efficient in executing their game plan.

In a similar comparative fashion, [46] proposed the analysis of performance variables across player positions such as technical actions and physical activity, for the SLL and EPL over the 2006-2007 season. Whilst some results demonstrate some similarities amongst the groups, others, show that unique characteristics exist in the two leagues. Specifically, in terms of technical performance, the first relevant insight was found in terms of ball possession with central attacking midfielders playing out the highest amount of time in control of the ball in SLL, whereas the same was true for this position as well as wide midfielders in the EPL. The disparity between the leagues was highlighted further when the amount of ball interactions was considered, with statistics revealing that wide midfielders in the EPL and forwards in SLL had 20% more ball contacts per possession than their counterparts. Contrastingly, there was no difference between the two divisions in terms of the quantity of heading disputes and the proportion of completed passes, apart from SLL strikers, who outperformed their English counterparts in both categories. Along the second vector of analysis – the physical performance –, the authors found significant disparities in total distance run across midfield roles, with defensive midfielders covering much longer total distances than centre offensive midfielders, notably in the EPL. However, when the total distance travelled by players in the leagues was considered, no distinction between the competitions could be made because these figures fell within the range of those reported in other professional European leagues. The authors did, however, found evidence that EPL players ran a much longer total distance in high intensity running than SLL players, regardless of location on the field. This latter conclusion constitutes the basis of a very well-known playing stereotype for the English top-tier national competition. Commonly perceived as the league with the harshest marking and quickest game among the five, the EPL significantly quicker and faster paced approach to game has been defended by [49–51] whose findings point in the same direction as the insights derived from previous results [46].

In their study, [50] examined statistics like the amount of fouls and cards between the 2007/08 and 2017/18 seasons in order to understand the evolution of the aggression profiles in the Big-5 European Football Leagues. Relevant findings show that whereas attempted tackles per foul have remained the same or even grown over the years, the number of fouls per game and each yellow card has declined in all competitions, leading the authors to conclude that top European soccer has gotten less aggressive over time. In addition, an interesting pattern was found in the EPL with the results showing it regularly had fewer total fouls and cards issued each game even though more tackles per foul were attempted, with writers characterising it as the most aggressive league.

In a comparative analysis, [51] studied the English, Italian and Spanish top-tier competitions only for one season - 2008/09 - with the objective of describing each one by looking into game-related indicators. As statistical studies, the authors applied multiple regression models even though the ANOVA study produced more interesting results. Indeed, and building on the EPL profile drawn before, the researchers

portray it as having both the quickest paced game and the hardest and most resilient approach to playing (opinion supported on the basis that they had the highest number of tackles). In contrast with this profile, but in line with the prevalent assumption, SLL was found to have a more intricate and skilled performing technique with results emphasizing their effectiveness in converting shots into goals (statistically significant higher values were found when compared with the EPL and ISA). Finally, authors still concluded that the results obtained do not point entirely in a different direction from the common belief that ISA is characterized by a cerebral and thoroughly measured playing style often characterized by making greater use of the defensive third of the pitch, since results from the analysis leads them to characterize it as the league with the best passing even though they showed a more strong offensive presence than was expected relative to the remaining ones.

Contrasting findings were found in respect to the Italian league by [47], when the authors decided to investigate the Big-5 European Football Leagues clubs' performances in UCL games over a period of nine editions (2009/2010 – 2017/2018), with the aim of discovering technical disparities. This different approach to comparing the leagues considered, in total, 20 performance measurements categorized into three broader categories: factors relating to goals, organization and passing, and defense. Notable findings revealed that most distinctions occurred when the German and Italian leagues were compared to the others. Concerning the Germans, results indicate that players from this league took more shots compared to their Spanish, English, and Italian counterparts and completed more long balls passes ($\geq$ 25 yards) than French players. Respecting the later, players in the Italian league had the lowest number of ball touches, passes and passing accuracy compared to any of the divisions. Likewise, when confronted with English players, the same was true for the quantity of dribbles made. On the other hand, players from the Italian league played more long balls per game than their French counterparts. However, because the study only examines teams participating in the UCL, the research is restricted in its ability to advance any conclusion made and generalize/extrapolate it to the leagues. Since only the top-ranked clubs at the end of the season, saving few exceptions, guarantee a spot to participate, it is fair to say that the results are adequate to make distinctions between the leagues' elite rather than a comprehensive and in-depth profile of each one.

A comparable research was recently conducted by [52], covering the same number of seasons (2009/10 - 2018/19), with the primary goal of analyzing trends among the Big-5 European Football Leagues in terms of goals scored. In a Kruskal-Wallis analysis of 18 offensive indicators, the authors found that superior conversion rates from corner kicks were achieved in the EPL while the same was true for throw-ins in SLL and for counterattacks in GB. ISA results show that it was the league where most penalties were scored with FLO demonstrating the weakest scoring capacity of all leagues.

With a slightly different aim than the previous research that, in their analysis, [53], the authors focus their attention in analysing whether or not scoring the first goal and the time at when it is scored had any consequence on the match outcome in the Big-5 European Football Leagues. For the purpose, data sourced included games from only one season (2014/15), and a tree and linear regression analysis were applied with results demonstrating that in 57.8% of the cases home teams scored first. GB teams lead the way (61.84%), followed by their French and British counterparts, with ISA teams demonstrating

the weakest home advantage value (56.47%) of the five. External aspects, such as the match location (home vs away), have indeed been linked to weight on the whether teams are successful, particularly on a match outcome basis, as other research has demonstrated [43, 54–56], with [57] going even further and suggesting that these effects are due to crowd, travel and familiarity factors.

As is visible from the previous paragraphs, the scope of the investigations carried towards comprehending the different leagues provide interest distinctions that could be interpreted more broadly as different game-approach stereotypes, saving some contradictory results, with the English league games demonstrating higher-paced and tougher events in contrast with the more well thought off and measured Spanish and recondite Italy approaches. GB teams seem to be distinguished by the strength and athletic presence of their players with [47] characterizing FLO as a blend between the ISA's defensive organization and EPL's inclination for physically capable players.

2.2.3 Classification Analysis

The first effort at forecasting the outcome of a match was made in 1997, and it was based on a Poisson distribution to derive probability for the goals scored [58]. Over the years, the advent of increasingly research being developed with the aim of understanding the game of football and metrics that play a more/less important role in the teams' performance has also motivated the usage of DS to build predictive models with ever-increasing accuracy, to the extent that enormous business models have been created around this matter (e.g. betting).

Baboota and Kaur in [59] touched on these issues while using ML techniques to forecast the outcomes of EPL games using data from 11 seasons (2005/06 - 2015/16). The methodological steps included feature engineering and feature selection, whose result was a feature set, constructed based only on the three best-performing variables (corners, shots on target and goals) since the authors believed these provided sufficient information regarding the team's level of superiority. In total, four ML models were tested (Gaussian Naïve Bayes (GNB), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB)) that outputted promising results after tuning their hyper-parameter, particularly when using the later two techniques whose accuracy values came close to matching the betting organization's models (in contrast to the benchmark values reported by the betting websites analysed, which claimed a performance of 0.2012, the gradient boosting method produced a performance of 0.2156 on the ranked probability score measure). As the authors had made evident in the introduction of the investigation, match outcome prediction is often challenging because of the draw possibility and the results obtained further corroborate this premise with this – drawing outcome – being the worst modelled category for all algorithms considered.

Earlier investigations had already tried different approaches such as [60] and [61] who advanced Bayesian Network (BN) models intending to forecast the outcome of matches (win, lose or draw) played by Tottenham Hotspur Football Club and Futbol Club Barcelona, respectively. In the former, the authors studied data from two seasons (1995/96 and 1996/97), which they divided into three distinct train and test data groups. First, they considered the 1995/96 season games as training data and the 1996/97 season

games as testing data. Then, they examined individual seasons, dividing the game information of each season into training and testing sets. Lastly, they conducted a cross-seasonal study that considered incomplete information from the first season as training data and forecast game results from the second as test data. In the end, the best performing model achieved an average accuracy of 59.21%, a value that was sufficient to outperform other techniques such as the K-Nearest Neighbors (KNN). The latter investigation analysed data from the 2008/09 season, where the author reached a staggering value in terms of mean accuracy (92%). Nonetheless, as part of the limitations of the study, authors assert that the scope of the studies – in both situations, just one team was included while training the models, and only one season was considered when training the second model –, poses a difficulty for further use since they are unlikely to perform well for further generalizations.

Still on a game-level prediction basis and expanding on previous research from [62], [63] considered data from different sports leagues: the football EPL, the Australian Football League (AFL), Super Rugby (SR) and Australian National Rugby League (NRL) with the aim of analysing the usefulness of neural networks in forecasting the outcome of various sports events. The authors decided to concentrate on objective data when modeling the feature space by taking into account variables such as the points scored by and against the team, the points won and lost in all previous games, and the team's current ranking, even though some subjective contextual data was used, such as the game's location and player availability. In comparison to rugby (average accuracy of 67.5% and 63.2% for Super Rugby and NRL, respectively) and Australian football (average accuracy values of 65.1% for AFL), results obtained show that football data was the most difficult to forecast (average accuracy values of 54.6%), with the authors emphasizing the models difficulty in modelling draws which comprised a quarter of the observations.

In [64], the authors used the decision tree algorithm on data gathered over three seasons (2015/16 - 2017/18) in order to assess its performance in anticipating results from the Chinese Football Association Super League. To predict the outcome of games (for example, between teams A and B), the authors built the feature space by focusing on the team's most recent match result, the outcomes of their most recent meetings (home and away), and each team's final ranking position in the previous season, with results demonstrating a 57.7% accuracy.

Research has also been conducted with the goal of predicting the classification of teams at season end, rather than on a game-outcome basis, despite the difficulty in doing so, as emphasized by [15] since it is contingent on the ranking of all other teams.

Authors in [65], based solely on the number of goals scored in home and away games by each team of the Croatian First Football League, developed a model based on a Poisson distribution with the aim of predicting the final ranking of teams (first, second and third most likely position) using only data from the first half of the season. The authors validated the model after having correctly predicted the final place of six of the ten teams in the league during the 2014/15 season. The writers then examined the partially available data from the 2015/16 season, given that the season had not yet ended as of the time of writing. A quick web check of the final rankings for this season reveals that the model accurately predicted the position of 50% of the teams.

In [2], authors conducted a more extensive analysis by considering data from 1200 Chinese Super

League matches (seasons 2014/15 - 2018/19). The investigation was predicated on 22 characteristics (Table 2.2) that looked at offensive and defensive performance and passing quality, which the authors fed into a Linear Support Vector Classifier with the aim of ranking the teams based on their performance and analyzing the factors that most impact game result. On top of the satisfactory prediction accuracy of 83% obtained, results also showed that the variables number of saves, shots on goal from inside the penalty area, and passing accuracy contributed the most to explain the success of teams with factors like open play shots on target, number of passes, and poor shot percentage emphasizing precisely the opposite.

Table 2.2: Variables used in [2] to predict the teams final ranking

Shots	Pass	Cross success	Interceptions
Shots on target	Pass success	Lost ball	Defensive Foul
Shot on target in penalty area	Pass attacking success	Tackles	Clearances
Shots opponent	Pass forward success%	Saves	Penalty
Bad shot%	Possession	Red card	
Shots on target opponent	Cross	Pen opponent	

In [23], rather than using data from only one country, the authors stretched the investigation to four of the Big-5 European Football Leagues: EPL, SLL, ISA, and FLO by making use of a broad range of classifiers: GNB, Decision Tree, RF, KNN, SVM (radial basis function and polynomial kernel) and XGBoost. For the intended purpose, 40 features were considered over four seasons (2015/16 to 2018/19), from where the first three constituted the training set. To simulate the games from the trial season in each competition, a model was developed which, through the resulting forecasted points, allowed to obtain predictions not only in terms of teams' final rankings but also in terms of whether teams ended up winning the championship, qualifying for European competitions and/or being relegated to a lower-tier league. Both resulting from the model that achieved the most consistent results - SVM with polynomial kernel (represents the similarity of vectors, or training samples, in a feature space over polynomials of the original variables) -, the highest accuracy was obtained for the EPL (57%) and the best result in terms of the RMSE metric was achieved in SLL, where the classifier correctly predicted the ranking of the first six clubs out of twenty in the league. Furthermore, findings collected show accuracies of 71% for both EPL and SLL and 57% for ISA and FLO in terms of forecasting the championship winner, with even higher average values for this statistic being found for teams that qualify for European championships (accuracies of 86%, 76%, 82%, and 46% for EPL, SLL, ISA, and FLO, respectively). The Relegated teams category was the most difficult to model, with an overall accuracy of 42%.

A recent and interesting approach to predict the final standings in football was advanced by [66], who seized the premature ending of seven main European competitions in the 2019/20 season due to the COVID-19 pandemic as a reference to guide the study. In their research, the authors use the team's performances in games (obtained solely through the matches outcome) before the season's halt to forecasts the remaining unplayed matchups through a statistical method. However, the results obtained were uninspiring when compared to the alternative model-based approach to predict the final standings, with the latter yielding better results in terms of accuracy than the former.

Other interesting predictive analyses have also been carried out, such as [67] which considered actual data from two bookmakers' forecasted odds as the groundwork to base their predictions of the outcome of games. In essence, they focused on the 1-X-2 classical bet type that is indicative of the home team winning, a draw and an away team winning, respectively, when analysing a total of 2615 games. On the premise that binary classifiers have better overall performance, the researchers focused on developing three different models, one per class type. Being so, for a given class (e.g. predicting home wins, equivalent to the 1 in 1-X-2) the new (binary) class variable takes the value one for the team in that class, and zero otherwise. Findings demonstrate promising values in terms of predicting wins for both types of teams: the home-wins ("1" in the 1-X-2 bet type) model was the best-performing classifier reaching an accuracy of 70.56% with the draw ("X" in the 1-X-2 bet type) classifier performing the worst from the three, further emphasizing the previously mentioned difficulty of the algorithms in modelling this category.

Chapter 3

Methodology

The purpose of the present chapter is to provide a description of the methodological options taken for the accomplishment of this dissertation. The variable respecting the categories of success is defined first, followed by a brief explanation about the methods and models employed.

3.1 How to measure teams' performance in football?

Contrary to some studies that analyse game-related statistics and whose basis to determine teams' performance is the outcome of games (namely win, draw or loss), a parallel aim of this study, when looking into the Big-5 European Football Leagues was to take a different perspective regarding this concept and evaluate a broader range of statistics on their capacity to express and determine distinctions in terms of the three possible outcomes for a team at the end of the season [1, 37–40].

Each European country has a top professional league, followed by minor divisions. Each one belongs to a larger regional organization, Union of European Football Associations (UEFA) for the leagues considered in this analysis, that oversees European-wide events such as the UCL, the UEFA Europa League (UEL), and, beginning in 2021, the UEFA Europa Conference League (UECL) [68]. Club's qualification for these events and presence or demotion to a lower-tier national competition depends on the ranking position at the end of the season. Hence, for this dissertation project, the distinction amongst the clubs was made considering three possible outcomes, as specified in Table 3.1.

Table 3.1: Categories of success.

Category	Short Denomination
Clubs that win the championship and/or qualify for European competitions	high-ranked
Clubs that safeguard their place in the league the next season	middle-ranked
Clubs that are relegated or qualify for the relegation play-off	low-ranked

One additional remark is that regarding the first category – high-ranked – there were no distinctions between direct qualifications and qualifications from the group stage – all teams were considered to belong to the same category. The same approach was adopted towards teams that were deemed

qualified for relegation matches (but were not directly relegated) – these were considered as belonging to the latter category.

3.2 Framework

The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is a process model for data mining that can be applied across industries, and it served as the methodology's backbone used to guide this investigation [69]. R. Daimler Chrysler launched this approach in late 1996, which has expanded over the years to include six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment [4]. The approach is summarized by the dynamic logical connection, represented in Figure 3.1.

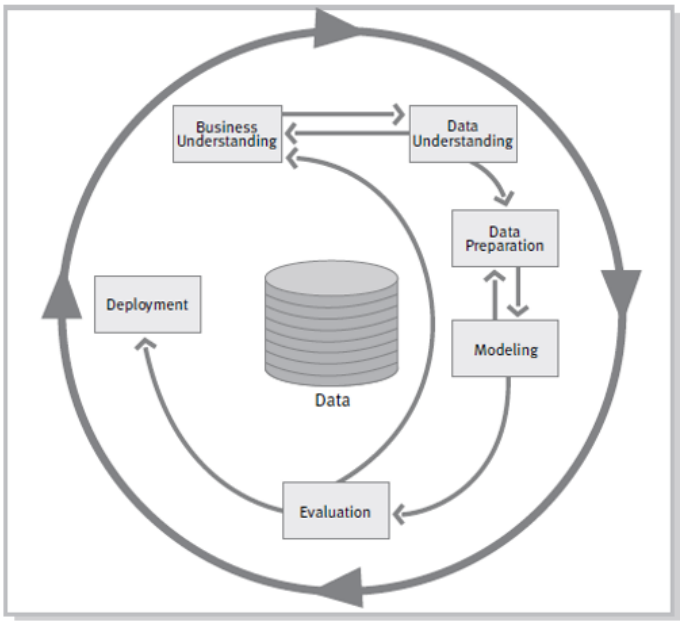

Figure 3.1: Phases of the CRISP-DM reference model (from [4]).

The first phase of the CRISP-DM process focuses on understanding the project goals from a business standpoint to subsequently translate this purpose into a data mining question and construct an initial plan tailored to meet these objectives. Chapter 4 targets this stage by providing a comprehensive explanation of why these types of research could be of interest to clubs and stakeholders and further outlining a course of action toward answering the proposed question based on available resources.

The Data Understanding phase comprises the earliest contact with the data. In this stage, activities such as collecting the initial data, checking its quality and deriving first insights are required to ensure its proper comprehension. Concretely, this stage was undertaken by exporting the relevant information from the FBREF website and further exploring the description of the variables available. Chapter 5 provides a more detailed explanation of the information retrieved: data for the Big-5 European Football Leagues during the 2017/18 – 2020/21 seasons for the first part and additional data relative to the same leagues for the 2021/22 season for the classification dimension of the research question and the procedures utilized in this stage.

Following the previous step, the Data Preparation phase is responsible for all operations related to obtaining the final dataset which will later be input into the modelling tool(s). Covered in Chapter 6, as listed by [70], these tasks include but are not limited to selecting relevant features, cleaning, constructing, integrating, and formatting the data. Fundamentally, the first was achieved through three steps: first, disregard every predictive statistic calculated by the platform; second, discard those that were deemed irrelevant for the study, and third perform an exploratory analysis to refine the dataset so it only included variables with increasing explanatory power. The next relevant steps included filling in the missing values, creating the success metric variable by manually introducing the information, reorganizing some metrics, and making the required merges. Subsequently, a preliminary data analysis was carried out to better understand some global and group properties of the dataset.

For this investigation, Python [71] was used as a programming language since, as evidenced by [72], it has over the years demonstrated great potential in the statistical analysis domain due to the rising reliability of its numerical libraries and quality of documentation. It was used in the Jupyter Notebook [73] environment: a free, browser-based program that allows coding, data treatment and analysis. For purposes of data pre-processing, features engineering and model creation, the Pandas [72], Numpy [74] and matplotlib [75] packages were used.

Covered in Chapter 7, the Modelling step includes selecting and implementing the appropriate data mining techniques for the project along with ensuring optimum calibration for their parameters [69,70,76]. As mentioned by [18], while working on a data mining problem, more than one technique is frequently required to achieve the final purpose of the study, which is exactly the case in the present research. Firstly, to perform the data analysis, the techniques employed to identify the KPIs that determine the success of teams included two unsupervised learning tools. Principal Component Analysis (PCA) was used first to filter down the most relevant performance indicators and then a Clustering Analysis using Hierarchical (Agglomerative) and Partitioning (K-Means) methods was applied. Then, regarding the classification dimension of the research question, several classifiers were tested.

The idea behind Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset while preserving as much of the original variability as possible. Fundamentally, principal components permit us to condense a large set of correlated variables into a smaller one of new variables - principal components - that together account for the total variability of the original set [77]. Given the different variable measurement scales and wide range of variability, the data was standardized. Mathematically, this can be achieved by subtracting the mean, $\bar{x}_i$, and dividing by the standard deviation, s_i , of the $i-th$ variable x_i, i.e. $z_i = (x_i - \bar{x_i})/s_i$.

Following this step, the eigenvalues, and eigenvectors of the correlation matrix, which includes entries associated with all possible pairings of the original variables, were calculated. In mathematical terms, if the initial data is defined as n observations with measurements on a set of p variables, $\mathbf{X} = (X_1, X_2, ..., X_p)^T$, then, essentially, the first component (Z_1) is the linear combination of the original variables with the highest variance, as in (3.1).

$$Z_1 = \phi_{11}X_1 + \phi_{12}X_2 + ... + \phi_{1p}X_p \tag{3.1}$$

The loadings of the first principal component will then be $\phi_1 = (\phi_{11}, \phi_{12}, ..., \phi_{1p})^T$ such that $||\phi_1|| = 1$ i.e. $\sum_{i=1}^{p} \phi_{1i}^2 = 1$.

It can be proved that ϕ_1 is the first eigenvector of the correlation matrix of X (as the data is standardized). Because of it, both ϕ_1 or $-\phi_1$ define the first vector of loadings.

Finally, the criteria used to select the number of PCs to be retained is the minimum number of PCs that accounted for at least 80% of the total variability.

Whilst PCA can be described as an unsupervised feature transformation method to find a lower dimensional representation of the original data that still captures a significant amount of the information, clustering techniques, also seek to simplify the data but, instead, the objective further includes finding patterns in the data by looking for smaller homogeneous groups in the observations. Thus, building on the insights from the previous analysis and complementing it, Hierarchical (Agglomerative) and Partitioning (K-means) methods were considered.

Contrary to the Hierarchical Clustering Divisive method, as the name suggests, the Bottom-up or Agglomerative Clustering approach starts by considering every data point as a cluster. From here, the algorithm evaluates all pairwise distances – inter-cluster dissimilarities – using a dissimilarity measure between objects and a linkage algorithm that will determine the merge process among clusters (Algorithm 1), yielding a tree representation of the datapoints [78], most commonly known as a dendrogram. The nature of the representation itself is extremely informative in terms of the number of clusters since, depending on the cut-off value, a different number of groupings can be identified.

Following this approach to finding clusters in the data means that once a data point enters a cluster, it can never leave/change clusters and can only be merged with other clusters. Conversely, this method is often considered since there is no need to pre-specify the number of clusters which might be an advantage in cases where there is no initial idea about a reasonable number of groups to partition the observations.

Algorithm 1 Hierarchical Clustering

1. Start with n observations and compute the dissimilarity matrix based on the pairwise dissimilarities between objects (using, for instance, the Euclidean distance). Consider each observation as a separate cluster.
2. For $i = n, n-1, ..., 2$:
   (a) Find the two clusters that are most similar. Merge them - their dissimilarity specifies the height in the dendrogram where the merge occurs.
   (b) Determine the updated pairwise dissimilarities amongst the remainder $i-1$ clusters and construct the new dissimilarity matrix.

First suggested by [79] in 1967 and improved subsequently by [80], K-Means Clustering Algorithm aims to find K groups of objects with minimum within cluster variability. Contrary to the Hierarchical approach, this algorithm belongs to a family of partitioning techniques that rely first and foremost on setting, a priori, the number of smaller groups – K (clusters) – that we wish to divide the entire (unlabelled) data into (Algorithm 2). After setting a value for K, a centroid is defined for each of the K clusters and

then, for each point in the dataset, the algorithm matches it with the nearest centroid. Then, each new centroid must be recalculated and each object is reallocated to the cluster which has the minor distance between the point and the cluster's centroid. This iterative process must then be continued until no more modifications to the K centroids' positions are made.

In addition to the need to prespecifying a number of clusters to partition the data, other drawbacks of this technique include the considerable sensitivity to the initial partition, where it can converge to a local optimum [81] that may yield unwanted solutions.

Algorithm 2 K-Means Clustering

1. Set the number of K clusters to be assigned.
2. Construct the initial partitioning, e.g. by randomly assigning each object to one cluster.
3. Repeat the cycle until the objects allocations into the clusters remain unchanged:
   (a) Calculate each cluster centroid (sample mean vector).
   (b) Allocate each individual observation to the cluster with the nearest centroid (using, for example, the Euclidean distance).

Its consideration here in the study follows the Hierarchical Clustering Agglomerative technique since it aims at building on the knowledge from the dendrogram but also to see if the classification of the teams into two groups was able to distinguish between high-ranked teams from middle-ranked and low-ranked teams (K = 2) and then, to examine how effectively the three preceding categories are distinguished, a model with three groups was evaluated (K = 3).

The outputs of the clustering models were evaluated through the accuracy, precision, recall and F1-score measures. The first indicates the percentage of correct predictions for the test results. The second - precision -, is classified as the percentage of relevant observations (true positives) among all the observations predicted to belong in a given class. And the third - recall –, is defined as the proportion of observations predicted to belong to a class compared to all of the observations that actually belong in the class. Finally, the F1-score aims combine the precision and recall metrics into a single metric.

Finally, the Silhouette Coefficient (SC) for the different clustering solutions were computed. Overall, it assesses the degree of separation between the groups [82]. Mathematically this can be achieved through Equation 3.2, for the $j-th$ observation, $j = 1, ..., n$ where $a(j)$ is the average distance between the $j-th$ observation and all the points in the same cluster; $b(j)$ is the shortest distance among the distances between $j-th$ observation and all the points in a distinct cluster. From here follows that a positive coefficient value suggests a correct assignment of that data point, while a negative one is indicative of a bad association of the $j-th$ observation to its cluster. Furthermore, the SC for the $j-th$ observation, $s(j)$, is defined as:

$$s(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}}, j = 1, ..., n \tag{3.2}$$

It can be easily proved that $-1 \leq s(j) \leq 1$ [82].

When using the methods, the original data which included all observations and features was used

and, in a separate analysis, data projected in the first 22 PCs was considered.

To answer the classification dimension of the research question, which aimed at classifying teams, in terms of the categories of success (Table 3.1), from all leagues given their performance in the 2021/22 season, information relative to this period for all leagues considered was added to the original dataset that served as the input for the PCA and Cluster Analysis. On that account, the four prior seasons worked as the train set while the newly added information constituted the test data. Since the possible outcomes of the success metric included three alternatives, the problem was kept as a multiclass or multinomial classification problem, as opposed to [67], for example, who decide to break down the problem to a binary classification one.

The models considered in this analysis were GNB, Logistic Regression (LR), KNN, Adaptive Boosting (AB), XGB, and RF. Nonetheless, before applying the methods, both the training and testing feature data were standardized to avoid poor model performance when attempting to discover patterns in the data and to guarantee that each feature is equally valuable to the study. To build the ML models, the Scikit-Learn package [83] was considered except for (extreme) gradient boosting that required the installation of the XGBoost Python module [84].

These python packages consider a set of default settings (hyperparameters) for all models but these are not guaranteed to be optimal for a problem. This obstacle can be minimized, to a certain extent, by tuning these parameters to values that optimize the algorithm's performance. Parallel to this objective, there was also the need to circumvent one of the most fundamental issues in ML that arises by merely testing each model on the training data – overfitting. Hence, the method considered for parameter tuning and avert models that scores exceptionally in the training set but are not able to hypothesise correctly to new data was cross-validation. The method used, K-fold Cross Validation [85], works by dividing the training set further into K number of subgroups and running several rounds of this procedure using various model settings combinations. The value chosen for K was 3. Finally, the methodologies available to run cross-validation on the training data included random, grid search or a combination of both and these were accounted for in some of the classification algorithms used.

The multivariate GNB model was the first classification model considered. By using the concepts of conditional probability [59], this simple yet efficient classification method takes a probabilistic approach by computing the posterior probabilities of a data instance belonging to a certain class. Let C be the class variable assuming values $1, 2, ..., c$ (in the present work $c = 3$) and $\mathbf{X} = (X_1, ..., X_p)^T$ be the random vector of the input variables and $\mathbf{x} = (x_1, ..., x_p)^T$ its realization. Assuming that $X_j|C = c$ follows a Gaussian distribution with mean value μ_j and variance σ_j^2,

$$f_{X_j|C=c}(x_j) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right), x_j \in \mathbb{R} \tag{3.3}$$

The hypothesis of conditional independence allows us to write:

$$P(C = c|\mathbf{X} = \mathbf{x}) = \frac{\prod_{j=1}^{p} f_{X_j|C=c}(x_j) \cdot P(C = c)}{f_{\mathbf{X}}(\mathbf{x})} \tag{3.4}$$

Given a new observation, $\mathbf{x}_0$, the model assigns it to the class with maximum posterior probability i.e. $c_0 = \arg\max_c P(C = c|\mathbf{X} = \mathbf{x})$.

The Logistic Regression (LR) [86] is used to predict a categorical or qualitative output variable, in contrast to linear regression, which is designed for a continuous response variable. Although it is most commonly used and best suited for situations with two answer classes, this classification approach may be modified to anticipate responses with more than two response classes. Like before, given $\mathbf{x}_0$, the object is assigned to the class with highest $P(C = c|\mathbf{X} = \mathbf{x})$. The multiple logistic functions are described mathematically for each explanatory variable $\mathbf{x}$ as in (3.5), which, after some modifications and taking the log transformation vector, results in (3.6).

$$p(X) = P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}} \tag{3.5}$$

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{3.6}$$

In the latter, the left-hand term is linear in $\mathbf{x} = (x_1, ..., x_p)^T$, and is referred to as the log-odds function or the logit function. In the end, the $\beta_0, \beta_1, ..., \beta_p$ coefficients reveal how much each explanatory variable adds to the odds of the response variable.

The KNN approach is a nonparametric (making no prior assumptions about the probability distribution of the observed data) algorithm used for classification and regression, first proposed by [87] in 1951. The letter "K" is a reference to the number of nearest neighbours to use as a proxy for similarity, the selection of which has a significant impact on the resulting KNN classifier, as stressed by [78].

Over the years, a respectable amount of research dedicated to improving this algorithm has been carried [88, 89] and, as demonstrated by [90], KNN performs best with a low number of features. Therefore, expectations for this model were not great.

The final three classifiers considered were the ensemble ML approaches Adaptive Boosting (AB), XGB, and RF. As the name suggests, ensemble methods are ML techniques that combine several base models to produce one final model. Thus, overall accuracy is frequently greater with ensemble approaches than with individual ML models since it is not enough for one model to be wrong for a prediction to be wrong; the majority of classifiers must be wrong [91].

Adaptive Boosting, commonly known as AdaBoost, was first established in the literature by [92] to enhance ML learners' performance. It is the first ensemble technique considered in this study which, like its counterparts, relies on aggregating results from multiple predictors to achieve their purpose. The difference between them lies, as pointed out by [59], in how this process is carried out with the present method operating according to the boosting principle (as suggested by the name). Concretely, this technique is described as the process of combining imperfect and moderately accurate classifiers (also known as weak learners) to get a very accurate prediction rule by training them sequentially on the complete dataset, with the model accuracy increasing with each iteration [92].

In this work, decision trees were chosen as weak learners.

The second ensemble prediction model considered was XGB. Like its previous counterpart, it oper-

ates under the boosting principle but employs a more regularized model formalization to control overfitting, leading, in general, to improved performance. Specifically tailored for speed, [84] reveal in detail in their study the novel design characteristics that provide XGBoost with a significant speed advantage over comparable ensemble algorithms.

The Random Forest Decision Tree Ensemble approach [93] was the last prediction model considered. Contrary to the previous two algorithms, this method follows the bagging principle which means that several decision trees on training samples are built in parallel – each classifier is trained simultaneously with a random subgroup of the data. To ensure that the trees do not look similar to each other, this ML technique employs a small adjustment while generating these decision trees by only considering a random subset of predictors to split candidates.

In the end, as briefly mentioned, the outputs of the ML models were evaluated through the 3-fold cross-validation mean accuracy. This technique used to assess the models' performances poses an advantage to simply looking at the proportion of correct predictions since, by training and testing the model on different blocks of data, an average of all results is taken and also provides insights into how well the model generalizes to new data.

Furthermore, the models' performances were also evaluated in terms of precision, recall and F1-score, as defined earlier. Additionally, to deepen our knowledge, Receiver Operating Characteristic (ROC) curves [94] were constructed to assess the performance of the best-performing classifiers. These curves plot, for various cut-off points, the true positive rate (recall or sensitivity) as a function of the false positive rate (specificity) - the proportion of observations from a class wrongly predicted (false negative) compared to all of the observations that actually belong in the class. Hence, they can provide a more comprehensive understanding of how the algorithms model each category since classifiers that give curves closer to the top-left corner, the more discriminating power the test has –, the curve of a test with perfect discrimination (no overlap between the two distributions) would go through the top left corner.

In the Evaluation stage, also integrated into the same chapter as the Modelling stage – Chapter 7 –, the results are analysed, interpreted, and compared to the established business objectives to see if they were reached.

Finally, despite the authors agreeing that the Deployment step's objective may vary considerably depending on the requirements [4], its main purpose is to put the knowledge captured by the previous phases into practice in a real context. Thus, since the present research was developed under the scope of a dissertation project, there will not be a deployment phase.

Chapter 4

Business Understanding

In this chapter, the context and motivation for the analysis of game-related performance indicators and the construction of classification models are introduced by exploring the advantages of carrying out research of this nature for the relevant stakeholders. It also describes a strategy for solving the requested question based on available resources.

4.1 Business Goals

The growing practice of implementing data collection and analysis towards better informed judgements has taken over most industries in the world. Likewise, the sports sector has also found in DS the wherewithal to improve, to certain extend, the myriad of aspects that might impact all domains of the business. As a company, the purpose of a football club is to be profitable through superior exhibitions of its team during the league season which, in the case of European teams, could ultimately mean placements in European competitions. According to Deloitte's Annual Review of Football Finance, together, the Big-5 European Football Leagues made a total of €15.6 billion in revenue in 2020/21, an increase of 3% from the previous year but a considerable decrease from the €17 billion in revenue before COVID–19 [5]. Hence, it is only natural for the team's decision-makers and managers ultimate goal to be aligned with increasing the club's performance.

A famous example was the German team's 2014 World Cup victory due to a incredibly effective strategy in which athletes and coaches were able to analyse data in real-time, helping them to identify their flaws and areas that needed improvement, through the usage of the SAP's Match Insights [61].

Like the German national team, federations and clubs are keen to absorb the wisdom that data and available studies can provide. As the review of the literature respecting performance indicators showed, the study of statistics to distinguish between successful and unsuccessful teams is nothing new to the football industry. Indeed, a substantial number of investigations focusing on international tournaments is already available, while others aim solely at analysing teams from national competitions. Research is also complete in terms of the number of teams and seasons considered, with some focusing their efforts on one team and/or season to others whose objective is to encase entire leagues and/or a collection

of competition editions (more than one season). Likewise, the scope of these investigations has been further expanded in terms of how researchers distinguish between successful and unsuccessful performances and/or teams, with some taking the outcome of games and others looking into the final position rankings for guidance.

However, literature also shows that arguably the most important element of these types of researches is the set of variables and statistics considered, as they could have the potential explanatory power to comprehend the game, at least partially (these metrics do not contemplate the spatial-temporal dynamics of football teams). In this regard, two problems were found. Firstly, it was the case that most analysis still look into a small number of variables. Then, it also fails to be comprehensive on all types of actions and interactions that can occur in a game, with most research overlooking and narrowing the analysis to metrics like possession and attacking statistics, placing little emphasis on goalkeeping, defensive actions, types of passes (rather than the number) and the location from where these took place in the game variables. For these reasons, and with the staggering amount of data generated every year, there is still room for building and improving the process of what has already been done, particularly towards extracting possible insights from underestimated dimensions of the game.

A review of the literature to compare the Big-5 European Football Leagues revealed several intriguing differences that may be regarded as various game-approach stereotypes. However, there is still a dearth of research available that aims to compare all five leagues in terms of performance indicators (probably for the reasons previously advanced in the beginning of subsection 2.2.2), a problem that has already been brought up by some researchers [3, 45, 46], with studies that aim to do so primarily focusing on league pairs or trios. Furthermore, results show that evidence suggests clearer distinctions amongst the EPL in terms of game pace and intensity [46, 49–51] and GB players' strength [45, 48], with conflicting findings being found, especially with the Spanish and Italian approaches to the game. Even though it was the subject of investigation in various studies, FLO had the least distinguishing qualities. Then, examining the predictive studies showed that the majority attempted to anticipate either the outcome of games or the team standings at the season's conclusion with a considerable amount of research focus in on the former due to betting. Furthermore, most studies [1, 21, 30–43] focused primarily on competition, with a small number of research seeking to broaden the analysis to include all Big-5 European Football Leagues [47, 50, 52, 53]. In the end, the accuracies gained were satisfactory in most circumstances, even though most models were constructed/trained using data from only one league, which likely limits their capacity to generate predictions when utilizing data from other competitions.

Even though the practice of using data did not come without its challenges given the complexity of the game (high amount of variables or "moving parts" and actions that can occur), which results in a sheer amount of data being generated at an increasingly fast pace, analysis such as the present that look into the key statistics that differentiate amongst the levels of success achieved in the end of the season and amongst the leagues might act as a compass for clubs in various dimensions.

Firstly, by having a comprehensive understanding of what better distinguishes teams in terms of the success metric considered, coaches could be able to better orient and define training practices and tactical strategies with focus on aspects that are demonstratively shown to given increased results.

On top of these insights, a deeper knowledge of the differentiation between the leagues in terms of performance indicators would be most useful to team managers and scouts since it would equip them with a clear idea of the different requirements across leagues thus indicating the most/least appropriate fits for the team in every decision. Additionally, physical coaches could also benefit from this information since the recovery from training and games is a very important element when it comes to sports, having a clear perception of the requirements of the several leagues and/or what parts of the body might be most sore and affected as a result of a specific game strategy might contribute to a more particularized and adequate rehabilitation regime.

Furthermore, the outcome of this analysis might still be useful for players given that it provides useful information regarding the areas that he/she should focus on and try to improve individually and a level of understanding that will allow them to better discern when confronted with real in-game situations.

Finally, the classification element of the research question might be useful for game and opposition analysts since, if robust enough, it could help forecast the team's and its adversary's expected final league result, in terms of the categories of success, by testing the model with partial data from the season. Having this information prior to the season end date would mean teams could implement opportune resolutions towards the club's end goal. One further observation is that it could also be useful for other entities such as betting organisations.

4.2 Available Data

Oriented towards answering the research question, it was clear that the required metrics had to include statistical information at the club level and across the Big-5 European Football Leagues.

Football data, as it was made evident previously, has, over the years, gained notoriety due to the myriad of dimensions it has revealed itself useful in the sport. In consequence, it is extremely valuable and, as a result, not easily accessible.

4.2.1 Performance Indicators

The FBREF website was the only source where the technical data with the necessary volume and degree of granularity was available. In terms of information accuracy, the website gets its data from StatsBomb, a very well-known data provider when it comes to football statistics. They claim "uncompromising accuracy" which most likely comes as a result of the way the information is gathered: the company focuses its efforts on collecting unique event-level data thus offering datasets with unmatched levels of granularity [95].

The site makes statistical information available at player, individual club, country and even match level, however, the relevant data is accessible through the "Competitions" tab. As the name suggests, the information in this section is organised according to the type of contest that teams participate in, such as the domestic leagues we are interested in. For each league, the available statistics are categorized into five groups prior to the 2017/18 season and eleven groups henceforth, as shown in Table 4.1.

Table 4.1: Groups of features availability for the leagues across seasons in FBREF website.

Category	until 2016/17	from 2017/18 on
Standard Stats	X	X
Goalkeeping	X	X
Advanced Goalkeeping		X
Shooting	X	X
Passing		X
Pass Types		X
Goal and Shot Creation		X
Defensive Actions		X
Possession		X
Playing Time	X	X
Miscellaneous	X	X

The variables included in each group are the same for every season even though the format in which the variables are expressed varies. They can be defined in terms of total number of times some action occurred, the percentage of successful actions, number of actions per game and even expected values (which results from an analysis performed by StatsBomb [96]).

4.2.2 Categories of Success

As introduced in the methodology in Chapter 3, the principle used to distinguish between successful and unsuccessful teams included the three conceivable scenarios for a club at the season's end. As this information was not included in the FBREF website, there was the need to look into alternative sources to understand how the distribution of teams occurred in terms of the categories defined in Table 3.1.

Teams' classification are organized into the three categories defined for all leagues. Firstly, as can be confirmed by the values in Table 4.2, every league counts with the participation of 20 clubs except for the GB, where only 18 dispute the championship. The smaller number of teams probably affected the decision of the number of clubs to demote to a lower tier, with only two teams (possibly three) losing their right to play in the following edition, compared to the majority of the leagues where this number is always three. FLO system for dismissing last-ranked clubs is the same as the German one, which is visible in Table 4.2 in the number between parenthesis - one club qualifies for relegation play-offs and must compete for its right to stay in the league.

In terms of UEFA events, the number of teams that have access to each tournament varies according to some factors. For the UCL, is determined by the UEFA coefficients of the member associations (leagues), which are calculated by the club's results throughout the previous five UCL and UEL seasons. The greater an association's coefficient, the more teams it has in the UCL and the fewer qualification rounds its clubs must participate in. Apart from FLO's club's yearly contribution of three, the four highest-ranked clubs in the remaining leagues are awarded a group stage place (see Table 4.2). Alternatively, clubs that win the UCL or UEL competitions are also awarded direct qualification for next year's edition.

Comparatively, a smaller number of teams participate in the UEL with the former leagues being

entitled to two places each, in normal circumstances. Additionally, several eliminated clubs from the first competition are still able to compete in this league.

Lastly, for the UECL, every league is entitled to a specific number of qualification(s) which also depends on the UEFA coefficients. For the relevant leagues, only one place is conceded.

Table 4.2: Big-5 European Football Leagues number of teams, usual UEFA competitions qualifiers and relegated teams or teams that qualify for relegation play-offs (between parenthesis).

	Premier League	Bundesliga	Ligue 1	Serie A	La Liga
Total nr of teams	20	18	20	20	20
Usual nr of UEFA qualifiers					
UCL	4	4	3	4	4
UEL	2	2	2	2	2
UECL	1	1	1	1	1
Total nr relegated teams (play-offs)	3(0)	2(1)	2(1)	3(0)	3(0)

To obtain this information and identify high-ranked teams, the institutional pages of the competitions were used which tracked and provide information of previous seasons participants. For the UCL, the website `https://www.uefa.com/uefachampionsleague/history/` was used whilst for the UEL, the information was found available in `https://www.uefa.com/uefaeuropaleague/history/`. Following the 2019/20 season, UEFA, the sport's governing body in Europe, decided to introduce a third tier to the two previously operating tournaments - the UECL. This annual competition was first contested in the 2020/21 season and, like its counterparts, the relevant information regarding the teams who participated could be found in `https://www.uefa.com/uefaeuropaconferenceleague/history/` .

The approach for mapping the two remaining categories was achieved by identifying the clubs that were relegated in the official league websites: `https://www.premierleague.com/tables` for EPL, `https://www.bundesliga.com/en/bundesliga/table` for GB, `https://www.ligue1.com/ranking` for FLO, `https://www.legaseriea.it/en/serie-a/archive` for ISA and `https://www.laliga.com/en-GB/laliga-santander/standing` for SLL.

Chapter 5

Data Understanding

The current chapter tries to investigate the raw data obtained from the various sources. First, a quick overview of how to access it is offered, followed by an examination of the variables included in each category.

5.1 Data Collection

In the FBREF website platform's home page, the first step to reach the relevant data is to hover over the "Competitions" tab. By doing so, a "Competitions Index" becomes available with an extensive list of league categories, being the relevant ones under the "Men's Big 5 European Leagues" title. By clicking on one competition, the website opens a page for that specific league where there is the possibility to navigate over the seasons by using the "Previous Season" and "Next Season" buttons. In addition, a secondary navigation bar (below the league identification) appears which includes, among other options, the "Squad and Player Stats" Tab. By hovering over this element, a list of eleven categories appears: Standard Stats, Goalkeeping, Advanced Goalkeeping, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions, Possession, Playing Time and Miscellaneous. Previously, it was shown that the availability of these categories depends on the season with the complete set only being available from the 2017/18 season (Table 4.1). This detail, associated with the fact that part of the research question included the classification analysis of the season 2021/22 meant that the suitable seasons to consider for the first part were 2017/18 until 2020/21. By clicking on one option, the website opens a different page showing two tables. The first one is under the title "Squad Stats", and this is the one that includes the relevant information.

The data was extrated by exporting it to CSV files and the complete statistical information that constituted the basis for this dissertation project was obtained by repeating this process across the five main leagues – GB, SLL, FLO, EPL and ISA – and the four seasons – 2017/18, 2018/19, 2019/20 and 2020/21.

Given that at the moment of extracting the data the 2021/22 season had not come to an end, there was the need for it to be extracted later so as to serve the purpose of validating the used classifiers.

5.2 Data Description

As briefly noted in Subsection 4.2.1, the statistical information in the 11 data categories was expressed in terms of the total number of times some action occurred, the percentage of successful actions, the number of actions per game and even some continuous variables as is the case with some statistics that measure the distance for some specific actions. Additionally, the website provides variables defined on an estimated value basis which were excluded from the beginning since it was not the scope of this investigation to analyse other models' forecasts.

In global terms, excluding repeated metrics across tables, the categories' initial statistical variables are presented and described in more detail in the following subsections.

5.2.1 Standard Stats

The Standard Stats category variables are listed in Table 5.1. First, it includes a few non-performance general statistics such as the number of players used by the team during the season, the team's player weighted mean age (weighted by minutes played) and the number of games played in the season, for example. Additionally, it considers game-related metrics like the number of penalty and non-penalty goals scored and the number of yellow and red cards. Moreover, it also includes some metrics defined as the number of actions per game like the number of assists, goals scored and assists, non-penalty goals and non-penaly goals, and assists per game performed by the team.

Table 5.1: Initial variables included in Standard Stats category.

Nr	FBref (Code) Denomination	NR	FR	FC
	# Pl (num_players_used)	X		
1	Age (average_age)			GS
	MP (matches_played)	X		
	Starts (starts)	X		
	Min (minutes_played)	X		
	90s (games_played)	X		
	G-PK (non_penalty_goals)		X	
2	PK (penalty_goals)			GSC
	Pkatt (penalty_kicks_attempted)		X	
3	CrdY (num_yellow_cards)			GS
4	CrdR (num_red_cards)			GS
5	Gls (goals_scored_per90)			GSC
	Ast (assists_per90)		X	
	G+A (goals_scored_and_assists_per90)		X	
	G-PK (non_penalty_goals_per90)		X	
	G+A-PK (non_penalty_goals_and_assists_per90)		X	

5.2.2 Goalkeeping

The second category, whose metrics are shown in Table 5.2, is named Goalkeeping. Excluding the wins, draws, and losses variables that measure the number of wins, draws, and losses of a team during the season, respectively, the remaining ones focus on actions that involved the goalkeeper and the adversary team's ability to score goals. The three initial variables were, nonetheless, considered since they might be useful when interpreting the results. Concretely, the group encompasses statistics such as the number of goals the team conceded during the season and goals conceded per game (GA90). In addition, it also measures the opponent's number of shots on target or framed with the goal (SoTA) and how many of these the goalkeeper and defenders successfully defended - Saves (num_saves). The former, when considered with the number of goals the team conceded (GA), allows the calculation of the team's successful save percentage - Save% (save%).

Additionally, the aggregation of variables still incorporates the proportion of games the team finished without conceding any goal - CS% (clean_sheets%).

The category is closed by some opponent's penalty kick metrics such as the number of attempted (Pkatt), scored (PKA), saved (PKsv) and missed (PKm) from the eleven-meter mark. When considered together, the first two allows the calculation of the team's successful penalty kicks save percentage.

Table 5.2: Initial variables included in the Goalkeeping category.

Nr	FBref (Code) Denomination	P&90	FR	FC
	GA (goals_against)	X		
6	GA90 (goals_against_per90)			G
7	Saves (num_saves)			G
	SoTA (shots_on_target_against)		X	
8	Save% (save%)			G
9	W (wins)			GS
10	D (draws)			GS
11	L (losses)			GS
	CS (clean_sheets)	X		
12	CS% (clean_sheets%)			G
	Pkatt (penalty_kicks_attempted_against)		X	
13	PKA (penalty_kicks_allowed_against)			G
14	PKsv (penalty_kicks_saved_against)			G
15	PKm (penalty_kicks_missed_against)			G
16	Save% (penalty_save%)			G

5.2.3 Advanced Goalkeeping

Still related to the goalkeeper and his actions in the game, the Advanced Goalkeeping category also targets variables concerned with this player's participation in the game, as seen in Table 5.3.

Firstly, it measures the number of free kicks, corner kicks and own goals the team conceded during the season. Secondly, it also incorporates metrics related to the number and types of passes performed

by the goalkeeper. Concretely, it assesses the number of passes performed by the goalkeeper, how many of these travelled longer than 40 yards (approximately 36.6 meters) - Cmp (launched passes) - and how many were successful - Cmp% (launched passes success%). Then, it looks into the number of throws (action where the goalkeeper returns the ball to the game by hand) and also some distance-based metrics that evaluate average pass and goal kick length - AvgLen (passes average length) and AvgLen (goal kicks average length), respectively. Additionally, other metrics appraise the goalkeeper's capacity to perform when under attack by the opposing team such as the number of crosses stopped - Stp (crosses stopped against) -, and an overall variable that measures the total number of defensive actions outside the penalty area - #OPA (num defensive actions outside penalty area).

Table 5.3: Initial variables included in Advanced Goalkeeping category.

Nr	FBref (Code) Denomination	P&90	NR	FR	FC
17	FK (free_kicks_goals_against)				G
18	CK (corner_kicks_goals_against)				G
	OG (own_goals_against)		X		
	Cmp (launched_passes)	X			
	Att (launched_passes_attempted)	X			
19	Cmp% (launched_passes_success%)				G
20	Att (passes_attempted_by_goalkeeper)				G
21	Thr (throws_attempted)				G
	Launch% (launched_passes%)			X	
22	AvgLen (passes_average_length)				G
23	Launch% (launched_goalkicks%)				G
	AvgLen (goal_kicks_average_length)			X	
	Opp (crosses_attempts_against)	X			
	Stp (crosses_stopped_against)	X			
24	Stp% (crosses_stopped_against%)				G
	#OPA (num_defensive_actions_outside_penalty_area)	X			
25	#OPA/90 (num_defensive_actions_outside_penalty_area_per90)				G
26	AvgDist (defensive_actions_average_distance)				G

5.2.4 Shooting

The fourth category is named Shooting and the variables included are in Table 5.4.

As the name suggests, the majority respect the number of shots performed by the team - Sh (shots) - and how many of these were on target - SoT (shots on target). Not including penalty kicks, the percentage of shots on target - SoT% (shots on target%) - measures the proportion of intentional shots that are on target with the goal. Alternatively, this measure is expressed on a ratio basis - SoT/90 (shots on target per90). The number of shots the team converted into goals is also considered - Gls (goals scored) -, along with how effectively they occurred - G/Sh (goals per shot). In addition, the number of free-kick shots - FK (free kick shots) - and the shot average distance from the goal - Dist (average dist from goal for all shots) - were also contemplated.

Table 5.4: Initial variables included in Shooting category.

NR	FBref (Code) Denomination	P&90	FR	FC
	Gls (goals_scored)	X		
	Sh (shots)	X		
	SoT (shots_on_target)	X		
27	SoT% (shots_on_target%)			GSC
	Sh/90 (shots_per90)		X	
28	SoT/90 (shots_on_target_per90)			GSC
	G/Sh (goals_per_shot)		X	
29	G/SoT (goals_per_shot_on_target)			GSC
30	Dist (average_dist_from_goal_for_all_shots)			GSC
31	FK (free_kicks_shots)			GSC

5.2.5 Passing

The next category is named Passing, and these variables, listed in Table 5.5, track the intentionally played balls between players of the same team [36].

Table 5.5: Initial variables included in Passing category.

NR	FBref (Code) Denomination	P&90	FR	FC
	Cmp (passes_completed)	X		
	Att (passes_attempted)	X		
32	Cmp% (pass_completion%)			P
	TotDist (passes_total_distance_travelled)		X	
33	PrgDist (passes_total_distance_travelled_twoards_opponent)			P
	Cmp (short_passes_completed)	X		
	Att (short_passes_attempted)	X		
34	Cmp% (short_passes_completion%)			P
	Cmp (medium_passes_completed)	X		
	Att (medium_passes_attempted)	X		
35	Cmp% (medium_passes_completion%)			P
	Cmp (long_passes_completed)	X		
	Att (long_passes_attempted)	X		
36	Cmp% (long_passes_completion%)			P
37	Assists (assists)			P

Adding to the more straightforward metrics that measure the total number of passes completed and their accuracy - Cmp (passes completed) and Cmp% (pass completion%), respectively -, this group also includes comprehensive information on the passes' length by recording their number of occurrences. Categorized into short, medium and long passes if occur at less than 15 yards (approximately 13.7 meters), between 15 and 30 yards and greater than 30 yards (approximately 27.4 meters), respectively, the variables included are expressive in terms of the number of attempts and successful passes of each type.

It also gives the precise measurement of the overall distance covered by the ball for all passes

- TotDist (passes total distance travelled) - and for passes performed towards the opponent's goal - PrgDist (passes total distance travelled towards opponent).

5.2.6 Pass Types

Still related to the way team mates exchange the ball, the Pass Types category, whose variables are in Table 5.6, complements the previous table information by targeting variables that provide information concerning the type of passes performed.

Table 5.6: Initial variables included in Pass Types category.

Nr	FBref (Code) Denomination	NR	FR	FC
38	Live (live_ball_passes)			P
	Dead	X		
39	FK (passes_from_free_kicks_attempted)			P
40	TB (in_depth_passes_completed)			P
41	Press (passes_under_pressure_comleted)			P
42	Sw (launched_passes_players)			P
43	Crs (crosses)			P
44	CK (corner_kicks)			P
45	In (inswing_corner_kicks)			P
46	Out (outswing_corner_kicks)			P
47	Str (straight_corner_kicks)			P
	Ground (ground_passes)		X	
48	Low (low_passes)			P
49	High (high_passes)			P
	Left (passes_attempted_left_foot)	X		
	Right (passes_attempted_right_foot)	X		
	Head (passes_attempted_head)	X		
	TI (throw_ins)	X		
	Other (passes_other_body_parts)	X		
	Cmp (passes_completed)	X		
	Off (offsided)	X		
	Out (out_of_bounds)	X		
50	Int (intercepted_passes_by_opponent)			P
51	Blocks (blocked_passes_by_opponent)			P

Firstly, the category distinguishes between passes performed during open ball plays - Live (live ball passes) - and passes performed when the ball is deemed temporarily not playable (e.g. throw-ins or free-kicks) - Dead (dead_ball_passes). Then, it also considers the number of in-depth passes completed (TB) and passes performed under pressure (Press). It further differentiates the height at which passes were performed by tracking the number of ground, low and high ball exchanges.

The group still includes information regarding the number and type of corner kicks (inswing, straight or outswing), accomplished by the team in the season. In addition, statistics related to the successful interceptions and blocks of passes from the opposing team.

5.2.7 Goal and Shot Creation

The seventh group of variables is named Goal and Shot Creation, and the metrics included are in Table 5.7.

Overall, these variables' objective is to track the team offensive performance in detail by providing information regarding either actions that lead or could have led to a goal. The website defines shot-creating action as *"offensive actions directly leading to a shot, such as passes, dribbles and drawing fouls"* and it is one of the metrics considered.

More granularly, it includes the number of these occurrences individually with the variables: dribbles, shots and faults that lead to shooting attempts.

Table 5.7: Initial variables included in Goal and Shot Creation Time category.

NR	FBref (Code) Denomination	P&90	NR	FC
	SCA (shot_creating_actions)	X		
52	SCA90 (shot_creating_actions_per90)			GSC
53	PassLive (passes_lead_to_shot_attempt)			GSC
	PassDead (dead_passes_lead_to_shot_attempt)		X	
54	Drib (dribbles_lead_to_shoot_attempt)			GSC
55	Sh (shots_lead_to_shoot_attempt)			GSC
56	Fld (fouls_lead_to_shoot_attempt)			GSC
57	Def (defensive_actions_lead_to_shoot_attempt)			GSC
	GCA (goal_creating_actions)	X		
58	GCA90 (goal_creating_actions_per90)			GSC
59	PassLive (passes_lead_to_goal)			GSC
	PassDead (dead_passes_lead_to_goal)		X	
60	Drib (dribbles_lead_to_goal)			GSC
61	Sh (shots_lead_to_goal)			GSC
62	Fld (fauls_drawn_lead_to_goal)			GSC
63	Def (defensive_actions_lead_to_goal)			GSC

5.2.8 Defensive Actions

The eighth group of variables is named Defensive Actions, and the metrics included can be seen in Table 5.8.

Overall, these variables' objective is to quantify actions such as the number of tackles, pressure, interceptions and blocks to opponents' shots.

This category further includes information regarding the location where these actions occurred, segregating them into performing in the defensive, middle and attacking third of the pitch. The pressure variables accounts for the number of times a team applies pressure to opposing player who is receiving, carrying or releasing the ball, and it is deemed successful when the squad gained possession within five seconds of applying it.

Table 5.8: Initial variables included in Defensive Actions category.

NR	FBref (Code) Denomination	P&90	NR	FR	FC
	Tkl (tackles)			X	
64	TklW (winned_tackles)				DA
65	Def 3rd (tackles_defensive_1/3)				DA
66	Mid 3rd (tackles_middle_1/3)				DA
67	Att 3rd (tackles_attacking_1/3)				DA
	Tkl (tackles_completed)	X			
	Att (tackles_attempted)	X			
68	Tkl% (tackles_completed%)				DA
	Past (tackles_uncompleted)	X			
	Press (pressure_to_opponent_attempted)	X			
	Succ (pressure_to_opponent_completed)	X			
69	% (pressure_to_opponent_completed%)				DA
70	Def 3rd (pressure_defensive_1/3)				DA
71	Mid 3rd (pressure_middle_1/3)				DA
72	Att 3rd (pressure_attacking_1/3)				DA
	Blocks (blocks)			X	
73	Sh (blocked_shots)				DA
74	ShSv (blocked_shots_on_target)				DA
75	Pass (blocked_passes)				DA
76	Int (interceptions)				DA
	Tkl+Int (tackles_plus_interceptions)		X		
77	Clr (clearances)				DA
78	Err (mistakes_leading_to_opponent_shot)				DA

5.2.9 Possession

The ninth category is called Possession, and its variables can be seen in Table 5.9.

Besides tracking one of the most famous statistics in football – possession –, it includes information relating to the total number of times players touched and carried the ball in their and the opposing's team's penalty area and each third of the pitch.

The type of carrying is also tracked, with the website making the distinction between carries and progressive carries (defined by the website as "Carries that move the ball towards the opponent's goal at least 5 yards, or any carry into the penalty area. Excludes carries from the defending 40% of the pitch.").

Additionally, it incorporates information concerning the dribbles, namely the number of attempted, successful and the proportion of successful ones.

Furthermore, the number of carries in attacking third - 1/3 (carries in attacking 1/3) - and carries into goal box - CPA (carries into goal box) - are also contemplated, along with variables such as the number of failed attempts to regain the ball - Mis (failed attempts to regain the ball) - and number of ball losses - Dis (ball losses).

Table 5.9: Initial variables included in Possession category.

Nr	FBref (Code) Denomination	P&90	NR	FR	FC
79	Poss (possession)				Pss
	Touches (touches)			X	
80	Def Pen (touches_in_defensive_penalty_area)				Pss
81	Def 3rd (touches_in_defensive_1/3)				Pss
	Mid 3rd (touches_in_middle_1/3)			X	
82	Att 3rd (touches_in_offensive_1/3)				Pss
	Att Pen (touches_in_attacking_penalty_area)			X	
	Live (live_ball_touches)			X	
	Succ (dribbles_completed)	X			
	Att (dribbles_attempted)	X			
83	Succ% (dribbles_completed%)				Pss
	#Pl (num_players_dribbled_past)		X		
	Carries (num_times_ball_was_carried)			X	
	TotDist (total_distance_ball_was_carried)			X	
	PrgDist (total_distance_ball_was_carried_towards_opponents_goal)			X	
84	Prog (num_times_ball_was_carried_towards_opponents_goal)				Pss
85	1/3 (carries_in_attacking_1/3)				Pss
86	CPA (carries_into_goal_box)				Pss
87	Mis (failed_attempts_to_regain_the_ball)				Pss
88	Dis (ball_losses)				Pss
	Targ (num_times_player_was_targeted_to_receive_pass)			X	
	Rec (num_times_player_successfully_received_pass)			X	
89	Rec% (passes_received%)				Pss
90	Prog (num_progressive_passes_received)				Pss

5.2.10 Playing Time

As the name implies, the Playing Time category, shown in Table 5.10, covers mostly measures that are not tied to any part of the game's performance but respect some features of the players' playing time and substitutions.

Additionally, it tracks the team's average points per match - PPM (average points per match).

Table 5.10: Initial variables included in Playing Time category.

Nr	FBref (Code) Denomination	NR	FC
	Mn/Start (minutes_per_match_started)	X	
	Compl (complete_matches_played)	X	
	Mn/Sub (minutes_per_substitution)	X	
	unSub (games_unused_substitute)	X	
91	PPM (average_points_per_match)		GS

5.2.11 Miscellaneous

The eleventh and final category is named Miscellaneous, and the metrics included can be seen in Table 5.11. Overall, these include information regarding aspects of the game such as the number of fouls committed and drawn, offsides and own goals. Additionally, reports on aerial disputes are also provided.

Table 5.11: Initial variables included in Miscellaneous category.

NR	FBref (Code) Denomination	P&90	NR	FC
92	Fls (fouls_commited)			GS
93	Fld (fouls_drawn)			GS
94	Off (offsides)			P
	PKwon (penalty_kicks_won)		X	
95	PKcon (penalty_kicks_conceded)			DA
96	OG (own_goals)			DA
97	Recov (num_loose_balls_recovered)			GS
	Won (aereals_won)	X		
	Lost (aerials_lost)	X		
98	Won% (%_aereals_won)			GS

Chapter 6

Data Preparation

This chapter describes the actions taken to obtain the final dataset, which will eventually be analysed. It specifically investigates the criterion selection procedure and offers information on the restructuring made to the data. A preliminary data analysis is accomplished.

6.1 Data Selection

To make the information available in the programming environment, the Pandas library [72], more precisely the read_csv() functionality, was used. Initially, the categories were treated individually, which meant the creation of 11 Jupyter notebook files. In the end, each of these files outputted a treated dataset that, when merged with every other, constituted the final structured data to be utilized or processed for further analysis.

However, the metrics used for the analysis did not include every initial statistic included in the original tables from Chapter 5. In addition to filtering down repeated variables, which were only considered once, this happened for three reasons.

Having excluded any expected value type of statistic, the format of the remaining variables was either the total number of times some action occurred, the percentage of successful actions, number of actions per game and even some continuous variables.

However, what happened in some cases was that some metrics were related, as shown in Table 6.1. The initial two variables from Table 6.1a can be considered as fitting in the first category as they result of simply counting the number successful and attempted launched passes, whereas the last is obtained through a simple calculation based on the first two. Likewise, in Table 6.1b, the only distinction between the statistics is that the second is the ratio between the first and the total number of games played by a particular team in a season. From here, the decision was to consider only percentage and ratio-based metrics (later from each category in the last example). The excluded variables as a result of this step are identified in Chapter 5 tables by the P&90 (Percentage & 90 ratio) column.

Secondly, according to the purpose of the study, there was the need to discard some variables given that they were judged irrelevant to the study. This exclusion resulted from the fact that several metrics

Table 6.1: Groups of related variables

(a)
launched_passes
launched_passes_attempted
launched_passes_success%

(b)
goals_against
goals_against_per90

stored by the website were not informative of any aspect of the game or respected and measured a feature of the match acknowledged as suitable for the present investigation. Similarly, in Chapter 5 tables, it is possible to identify these abandoned variables under the NR (Not relevant) column.

Finally, a correlation analysis to the remaining variables (Appendix A) allowed us to rule out some metrics, to include (almost) exclusively those with added explanatory capacity. The Pearson sample correlation coefficient determined the exclusion criteria and the threshold value considered was 0.9. Hence, if two or more variables had a value for this measure greater than this reference, the one considered to have the highest interpretative capacity was accounted for in the final analysis. Like the previous, it is possible to see which variables were abandoned for this reason in Chapter 5 tables under the FR (Feature Reduction) column.

6.2 Data Munging

Data Munging, or the purpose of treating and converting it into a different format so that it may be utilized and processed, constituted an important step towards obtaining the final dataset. Concretely, it comprehended making changes, and solving errors in the initial eleven datasets.

Regarding the former, the only noteworthy operation included converting the continuous variables represented in Table 6.2 from an initial yard basis description to meters (equivalency used: 1 yard = 0.9144 meters).

Table 6.2: Variables that were converted from yards to meters.

AvgDist (defensive_actions_average_distance)	Advanced Goalkeeping
Dist (average_dist_from_goal_for_all_shots)	Shooting
PrgDist (passes_total_distance_travelled_twoards_opponent)	Passing

The only significant error with the data occurred in the number of yellow and red card variables, where values from the GB, EPL and ISA for 2017/18 were missing. In total, the lacking information accrued to 58 values, which, given the low quantity, was resolved by manually filling in the values sourced from each league's official pages (Subsection 4.2.2).

As mentioned previously, none of the initial categories had information respecting the success metric target variable, which meant that the only approach to make it accessible for further analysis was to create it manually. Given that, in the end, to obtain the final dataset, the initial eleven had to be merged, the addition of the success metric variable was solely done in the first – Standard Stats – category's Jupyter Notebook. Indeed, the last step towards constructing the final dataset comprehended merging the eleven individually treated datasets. Furthermore, eleven groups to organize the data was unneces-

sarily high since, in some cases, the nature of the metrics was similar and related to the same aspect of the game. To minimize unnecessary complexity, the variables were restructured to only six categories, as seen in Table 6.3. In the end, information regarding the included variables and where they were considered is provided in Chapter 5 tables under the last column – FC (Final Category).

Table 6.3: Final variable's groups.

Final Category	**Initial Category**	**Nr of Variables**
General Stats (GS)	Standard Stats	3
	Goalkeeping	3
	Playing Time	1
	Miscellaneous	4
Goalkeeping (G)	Goalkeeping	8
	Advanced Goalkeeping	10
Passing (P)	Passing	6
	Pass Types	14
	Miscellaneous	1
Goal and Shot Creation (P)	Standard Stats	2
	Shooting	5
	Goal and Shot Creation	12
Defensive Actions (DA)	Defensive Actions	15
	Miscellaneous	2
Possession (Pss)	Possession	12
Total		**98**

The allocation of each variable was performed individually according to the category that made the most sense nonetheless, in general terms, categories related to similar aspects were merged. This happened in the Goalkeeping and Advanced Goalkeeping datasets since both provide goalkeeper and keeper-related actions statistics during the game. Likewise, both the Passing and Pass Types provide knowledge related to the act of passing during the game. And the Shooting and Goal and Shot Creation, since both provide offensive action information during the game. Additionally, the majority of statistics included in the Standard Stats, Miscellaneous and Playing Time datasets were not related to any specific aspect of the game and thus were grouped. In the end, 98 variables were accounted for in the final dataset.

6.3 Preliminary Data Analysis

In the end, the 490 observations considered across five leagues and five seasons in the final dataset accounted for a total of 135 different clubs competing across these events.

In terms of the club's representativity concerning the success metric categorical variable, it is clear from Figure 6.1 that more than half of the teams (51.4%) account for clubs that safeguard their right to compete in the league in the following season (middle-ranked teams). The category that includes championship winners and clubs that qualify for European competitions (high-ranked teams) comes

next with 34.9% of the teams, with clubs that are relegated and compete in the relegation play-offs (low-ranked teams) adding to just 13.7%.

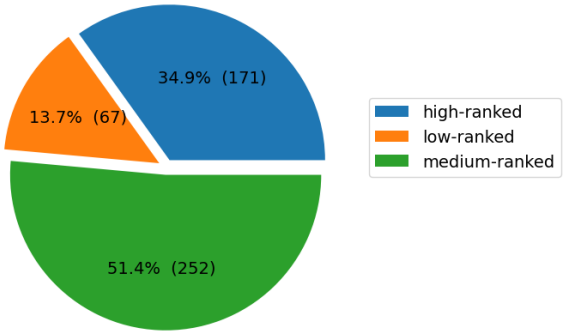

Figure 6.1: Distribution of clubs in terms of the categories of success.

Concerning the leagues, as supported by evidence from Figure 6.2, the GB was the only under-represented league with 90 teams, which is not surprising since only 18 squads compete yearly in the league, compared with the 20 that do so in the remaining ones considered, as previously emphasized in Subsection 4.2.2.

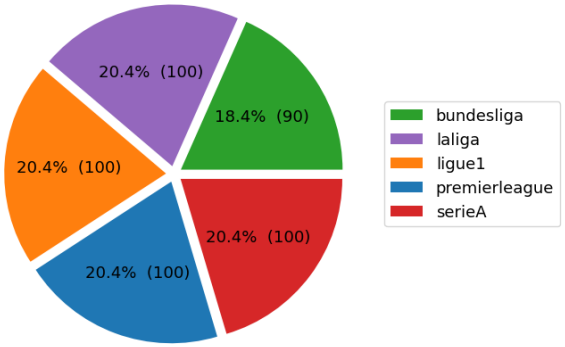

Figure 6.2: Distribution of clubs in terms of leagues.

Structured in a way that, throughout one season, each team faces every other team competing for the championship twice, the fact that only 18 teams compete in the GB has a direct impact on the total number of games at the end of the season.

Indeed, a preliminary analysis to the variable matches_played (from Table 5.1) comproves this fact as is suggested by the pattern of a consistently lower number of games in the German league, with most GB seasons totalling 306 match events compared to the 380 in the remaining leagues, visible in by Figure 6.3.

Curiously, when analysed in more detail, the histogram from this variable still suggests other interesting findings, explained below.

Firstly, and distinctive from the rest, is the number of matches played during the 2019/20 season in FLO, with the 279 games contested, instead of the supposedly 380 that should have taken place

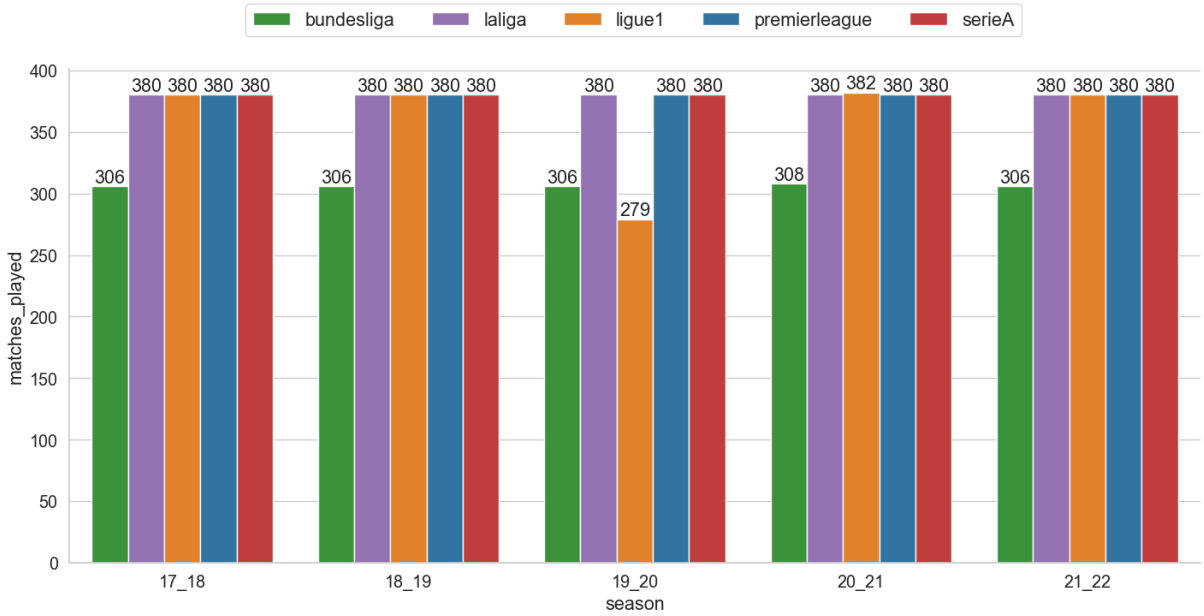

Figure 6.3: Distribution of games in terms of leagues and seasons.

that year. This value is, however, not surprising given the effect that the COVID-19 pandemic had on the leagues. More specifically, during the 2019/20 season, the organization in charge of overseeing France's top professional football leagues – Professional Football League (LFP) – announced a complete shutdown in operations across its top two tiers following the outbreak of COVID-19 in the country. Contrary to their counterparts in other European countries, FLO did not resume that year, resulting in fewer disputes for the teams competing for the championship.

In addition, the GB and the FLO have two additional reported games for the 2020/21 season, summing 308 and 382 events, respectively. After looking into these exceptions, the conclusion was that they resulted from the relegation play-off still implemented in these countries, where one team from the top tier competes with another from the subsequent tier in the league to be kept and promoted to the former, respectively. For some unknown reason, the FBref website only accounted this distinction (briefly mentioned in Table 4.2) for the 2020/21 year. Thus, as these games result in two additional opportunities for FC Nantes (from FLO) and FC Köln (GB), it would not be appropriate to compare these feature values (excluding percentage and ratio-based metrics) to their counterparts, as they would only have played the two scheduled disputes with every other team in the league. Hence, these observations were removed and not considered for further analysis.

The same issue applies to the metrics for German clubs as they have an overall smaller number of game disputes in the league. The decision, here, however, was not to exclude the league as a whole from the analysis but rather multiply all metrics from the features space, excluding percentage and ratio-based metrics, by a constant, as evidenced in (6.1).

$$UpdatedFeature = Feature * \frac{20}{18} \tag{6.1}$$

The multiplication will have the desired effect on variables that are defined continuously (some

distance-based metrics) but will, in turn, affect more deeply statistics that are expressed in terms of the total number of times some action occurred. For example, $3 * 20/18 = 3.33$ is impossible and unreasonable. Furthermore, as the effect that this multiplication has is considerably worse in statistics with lower values, the approach in these cases was to round to the nearest integer. The variables that track the number of wins, draws and losses (Table 5.2), required special attention since, upon multiplying the features' values, their sum did not accounted for 38 (the supposed number total games a team would play in a league with 20 contestants). Thus, in few these cases, the values for these variables were reasonably adjusted to ensure their sum yielded the desired value.

These decisions are not without its harm to the analysis and is, indeed, a limitation of the study since these values are estimations and do not represent the actual reality.

Concerning the exception of FLO in the 2019/20 season, the decision was to use the data without applying any correction/edition, as it was considered a completely different situation from the previous. Whether the relegation play-off games for the 2020/21 season in the German and French leagues and the number of teams in the GB are intrinsic league characteristics, the initial refereed occurrence respects an unexpected event.

Following these adjustments, the data was ready to be used in the models.

Chapter 7

Modelling and Evaluation

This chapter combines the presentation of the model's results as well as its discussion. Firstly, the outputs from the PCA are displayed together with an analysis of the most relevant metrics that influence teams' performance and on which KPIs suggest distinctions in the Big-5 European Football Leagues. Then, results from the Cluster Analysis are explored. Finally, the outputs of several classifiers are presented and investigated.

7.1 Principal Component Analysis

The eigenvalues and eigenvectors were obtained from the correlation matrix of the original statistics. Even though 98 variables were analysed, Table 7.1 only provides information for the first 24 and last three eigenvalues, to condense the most important results.

According to the findings, the first component, alone, explains roughly 28% of the total variance, followed by the second and third components which explain approximately 8% and 6%, respectively. Together, the first five components explain 50.895% of the total variance.

7.1.1 Key Performance Indicators and Football teams' performance

In Figure 7.1, the scores on the first and second PCs are plotted with the help of the categories of success to colour the datapoints. The figure suggests that the first PC does a good job at separating the high-ranked teams from the middle-ranked and low-ranked clubs but misses on a more detailed segregation between the two latter categories. More concretely, squads with positive scores on the first component tend to belong to the high-ranked teams group whilst squads with negative scores on the first component, saving a small number of exceptions, tend to be classified as either as clubs that ensure their right to compete in the league the next season (middle-ranked teams) or clubs that are relegated or qualify for the relegation play-offs (low-ranked teams). The latter – low-ranked teams – score the lowest in PC1, with middle-ranked teams scoring higher, on average, than this group.

Considering an additional restriction to plot the previous data – data points with an absolute scores in PC1 grater than 10 $(\hat{PC}_1 > 10)$ –, it is possible to identify the teams that have the highest scores in

Table 7.1: Eigenvalues and associated percentage of variance explained by the i PCs.

i	Eigenvalues (λ_i)	Cumulative Percentage of Total Variance
1	27.515	28.005
2	7.817	35.960
3	6.044	42.112
4	4.416	46.607
5	4.213	50.895
6	3.268	54.221
7	2.943	57.217
8	2.656	59.920
9	2.325	62.286
10	2.243	64.569
11	1.963	66.567
12	1.748	68.345
13	1.564	69.937
14	1.526	71.491
15	1.333	72.847
16	1.300	74.170
17	1.224	75.416
18	1.114	76.550
19	1.112	77.682
20	1.043	78.744
21	0.979	79.740
22	0.959	80.715
23	0.908	81.640
24	0.900	82.556
...	...	...
96	0.002	99.997
97	0.002	99.998
98	0.002	100

PC1. As expected, these teams correspond to some of the historically most successful teams in each league.

In the EPL, the teams identified were Manchester City FC in all seasons, Chelsea FC in the 2018/19 season and Liverpool FC in the 2018/19 and 2019/20 seasons. All cases are not surprising since Manchester City FC was champion in three out of the four seasons considered. Furthermore, besides winning the UCL (which in itself can be perceived as a more general insight into the team's competitiveness and performance rather than a concrete direct relationship between the quality of their work in the competition and its performance in the league), the 2018/19 season was one of the most competitive EPL competitions ever, with Liverpool FC finishing only one point short than the winner – Manchester City FC. The team would then win the league trophy next year.

In FLO, Paris Saint-Germain FC was the team that stood out in all seasons except in 2019/20. Despite being champion in every season but 202/21, when it finished just one point away from tying Lille, the absence of Paris Saint-Germain in the referred season might also be related to the premature

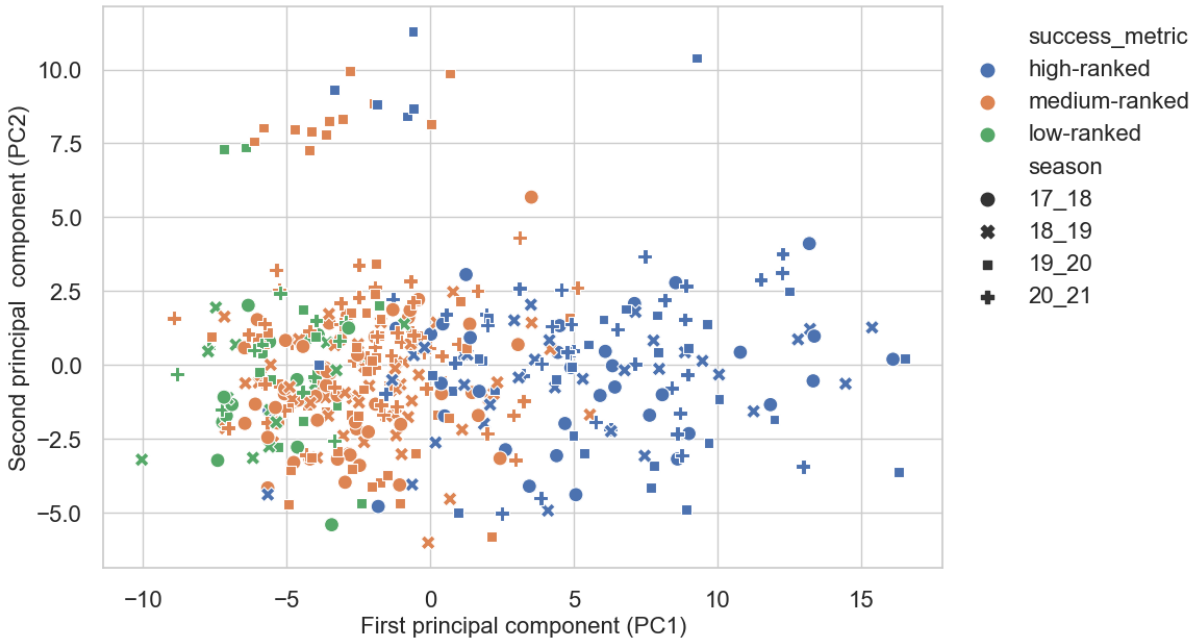

Figure 7.1: Scores on PC1 vs PC2.

ending of the league that year, as noted in the preliminary data analysis.

Having won the league title all years, Bayern Munich stands out in every season in GB. In SLL, Barcelona appeared for all seasons, while Real Madrid only stood out in the 2017/18 season, a year that it won the UCL. S.S.C. Napoli, in the season 2017/18, was also one of the teams that stood out, with the team finishing second, just 4 points ways from the champion Juventus, which is indicative of a strong performance from the team this season.

To understand why these teams, who are recurrently classified as high-ranked teams, achieve tremendous success, we must investigate the variables that have the most significant impact on this PC.

As the initial number of variables in the feature space was considerable, there was the need to consider a specific threshold value to filter the most significant metrics for each component. In all cases, the (absolute) value was 0.15.

First Principal Component Loadings

The variables whose loadings satisfy the former restriction are shown in Table 7.2, and they reveal that the factors that have the greatest effect on this component are organized into three main categories: Possession, Passing, and Goal and Shot Creation. Additionally, for each variable (i) and the PC, the Pearson sample correlation coefficient $(Cor(PC_1, X_i))$ and corresponding p-value are displayed under the ”Pearson” and ”p-value” columns, respectively.

Included in the first is the percentage of ball possession, the variable with the second highest absolute value, which means it was one of the statistics with the greatest effect on positive values in the first component and, for this reason, was a significant factor in the team's classification at the end of the season. Furthermore, since the coefficient is positive, this is suggestive of a direct proportionality association with the PC. The Pearson's sample correlation coefficient between the variable - posses-

Table 7.2: Loadings whose absolute value is greater than 0.15 for the first PC and Pearson sample correlation coefficient and p-value between each variable and the component.

Category	Variable	Loadings	Pearson	p-value
Pss	num_times_ball_was_carried_towards_opponents_goal	0.175	0.934	$< 10^{-8}$
Pss	possession	0.172	0.917	$< 10^{-8}$
Pss	carries_in_attacking_1/3	0.167	0.886	$< 10^{-8}$
Pss	touches_in_offensive_1/3	0.166	0.883	$< 10^{-8}$
Pss	carries_into_goal_box	0.163	0.869	$< 10^{-8}$
P	live_ball_passes	0.173	0.923	$< 10^{-8}$
P	assists	0.165	0.876	$< 10^{-8}$
P	long_passes_completion%	0.161	0.854	$< 10^{-8}$
P	passes_total_distance_travelled_twoards_opponent	0.157	0.838	$< 10^{-8}$
P	pass_completion%	0.156	0.833	$< 10^{-8}$
GSC	shots_on_target_per90	0.171	0.909	$< 10^{-8}$
GSC	goals_scored_per90	0.170	0.907	$< 10^{-8}$
GSC	goal_creating_actions_per90	0.170	0.906	$< 10^{-8}$
GSC	passes_lead_to_goal	0.169	0.897	$< 10^{-8}$
GSC	shot_creating_actions_per90	0.167	0.889	$< 10^{-8}$
GSC	passes_lead_to_shot_attempt	0.166	0.884	$< 10^{-8}$
GS	wins	0.166	0.883	$< 10^{-8}$
GS	average_points_per_match	0.164	0.871	$< 10^{-8}$

sion - and the PC ($r = 0.917$), indicative of a strong positive association between the two variables – higher values of possession are associated with higher values in the first PC -, validates this premise while the p-value ($p < 10^{-8}$) for testing the null hypothesis that this correlation is zero suggests that the findings are, indeed, statistically significant. These results seem to be in line with conclusions from [30, 31, 35, 37], thus emphasizing the role that ball possession plays in terms of team performance. Furthermore, other noteworthy metrics from the same category include the number of times the ball was carried into opponents' goal ($r = 0.921; p < 10^{-8}$), carries in attacking third ($r = 0.867; p < 10^{-8}$), into goal box ($r = 0.862; p < 10^{-8}$) and touches in offensive third ($r = 0.872; p < 10^{-8}$). Even though the former is calculated over the whole area of the pitch - players can carry the ball towards the opponent's goal in all areas of the field (defensive, middle, and attacking thirds), whilst the latter three are measured only in the attacking third, all statistics are distinctively attacking statistics. Hence, given that high-ranked teams score the highest on PC1 and the coefficient for these statistics indicates a direct proportionality relationship, these results are suggestive of higher values on these metrics for teams belonging to the high-ranked category, thus emphasizing that on top of being in control of the ball more often – a higher percentage of ball possession -, high-ranked teams also display a strong offensive presence and attitude. These results are in line with findings from [39] and [40], who argued that top-ranked teams displayed increased action in the final/attacking area of the pitch and had much more possession in the opponent's half.

From the Goal and Shot Creation category, variables such as the number of passes that lead to goal and number of passes that lead to a shot attempt stand out, as having a direct proportionality relationship

with the present component. As such, this indicates that high-ranked teams displayed higher values on these metrics, on average. Results from the same category also suggest, even though more obviously, that high-ranked teams looked to maximize the number of goals scored per game and goal creation actions per game, a conclusion drawn previously by [1,35,36,38,39,43], rather than the effectiveness of the attempts, as promoted by [44]. Also, when considered with the variables from the previous paragraph - higher values in ball possession -, these results seem to suggest that most high-ranked teams might have prioritized a positional attacking style, commonly employed by teams with good tactical-tecnhical skills, rather than a more direct method of play (counter-attack and fast attack), which is characteristically quicker (attacking time is shorter), involves less number of players, and the ball can be played in depth (counter-attack) or towards with and depth (fast-attack). Besides that, the circulation of the ball takes place all over the width and depth of the pitch, where teams try to unbalance the opponent and take advantage of the spaces they can create. The use of this method - possession-based -, in fact, leads to a higher number of passes made, with the ball being transported from creation areas to finishing areas.

When analysed in more detail, the possession (Figure 7.2) and the number of passes that lead to goal (Figure 7.3), versus the values on PC1, shine some additional light into the relevance that these metrics play in team's success.

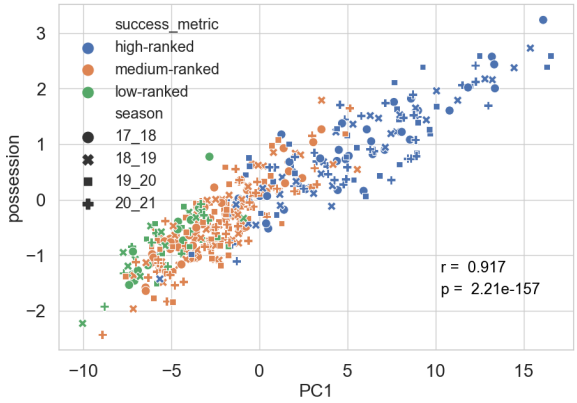

Figure 7.2: PC1 vs possession.

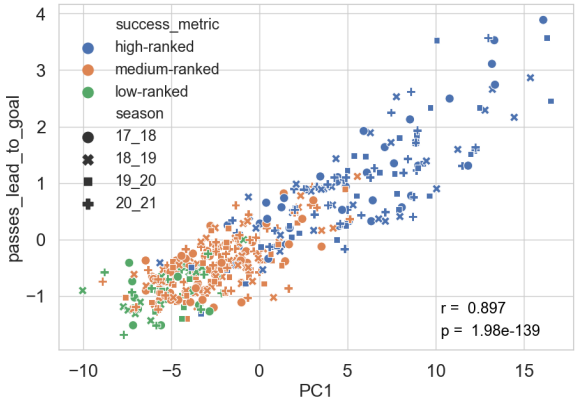

Figure 7.3: PC1 vs passes_lead_to_goal.

From the dispersion of points, there seem to be some distinction between groups, with the majority of high-ranked teams indeed showing higher values of these metrics, on average, thus corroborating that this pattern of play benefits teams. As expected, on the top right corner of each graph, the teams that stand out as having the highest values in PC1 as well as possession and number of passes that lead to goal are the same as the ones previously identified as having scored higher in PC1.

Contrary to these teams, distinctions are not so clear in the lower end of the high-ranked data points cloud given that some of these clubs show lower possession and passes leading to goals than some middle-ranked teams (datapoints from the two categories overlap approximately for PC1 values between zero and five). A closer examination to these points suggest that most cases are amongst the worst performing clubs from the high-ranked group and best-performing teams from the middle-ranked category, having finished just short of qualifying to European competitions. The most extreme from the high-ranked teams group include Eintracht Frankfurt e. V. (GB) who was last to qualify for European competitions

in both 2017/18 and 2018/19 seasons, Burnley FC (2017/18), Wolverhampton Wanderers FC (PL) and RDC Espanyol (SLL) in the 2018/19 season and Stade de Reims (FLO) and Granada CF (SLL) in the 2019/20 season, while the same happened for Torino FC (ISA) and West Ham United FC in the 2020/21, who were second-last to qualify. Curiously, Borussia Dortmund's performance during 2020/21 season was found to be substantially dependent in possession than their high-ranked teams counterparts, even though they finished third in the league - suggestive of a solid qualification for European competitions, as opposed to the expected less dominant performance.

On the other hand, the most extreme cases from the middle-ranked category include A.C. Milan during the 2018/19 season who, despite finishing fifth was unable to qualify over financial fair-play violation, and U.S. Sassuolo Calcio from the same league, having finished eight in both subsequent seasons (2019/20 and 2020/21), just below S.S.C. Napoli and A.S. Roma, the last qualified teams for European competitions, respectively. Qualifying in ninth, also just one position away from the European competitions, is TSG 1899 Hoffenheim in the 2018/19 season.

Not so clear is the distinction between middle-ranked and low-ranked clubs in the bottom left part of each graph, even though the former group still scores higher, on average, than the latter. Particularly in Figure 7.2, there are some green points in the upper part of the arrangement (in the bottom left corner), that seem to suggest that having high possession alone is not sufficient if the team is incapable of convert this advantage into meaningful opportunities. The most extreme cases include teams who were relegated third such as Fulham (PL) in the 2020/21 season and Empoli FC and U.S. Lecce (ISA) during 2018/19 and 2019/20 respectively. Other teams include SC Paderborn 07 (GB), relegated last in the 2019/20 season.

These exceptions to the rule, suggest that a possession-based style of play, as the previous findings seem to advocate, should be approached with caution as they are proof that if not well executed and implemented, the tactical approach might still not be sufficient to yield the desired results, in some cases. This conclusion, seems to be exactly in line with the vision of [43], where the authors defend that instead of being looked at the go-to play style for teams regardless of the context, coaches should align their tactical decisions (whether to follow a possession-based or, conversely adopt a direct style of play) for each game with their game strategy for each match under the team's and opponent's recognized capabilities.

These results, however, indicate that most dominant teams indeed have higher patterns of ball possession and that a style of play reliant on exchanging the ball more often with the intent of reaching the final third of the pitch must be adopted. Nonetheless, as academic consensus has not yet been reached regarding this topic, together with the insights from [43], the exceptions previously encountered, such as Borussia Dortmund -, are exactly proof that teams can also find success in a more direct playing style.

In the Passing category, statistically significant variables such as the assists ($r = 0.874; p < 10^{-8}$) stand out, which have previously been linked to success inside the field by [1, 39]. Additionally, the inclusion of the percentage of successful passes completed variable ($r = 0.829; p < 10^{-8}$) and the percentage of successful long passes completed ($r = 0.854; p < 10^{-8}$) in this set suggests that high accuracy when exchanging the ball is an important factor that distinguishes these teams, as highlighted

by [43] in their study, who promoted attention to carry out fewer passes while attempting to ensure these were successful.

Finally, as part of the General Statistics, the number of wins and the average points per match were identified as relevant variables in the team's classification at the end of the season, which is not surprising since these metrics track the results of the games – teams that win more games will have a higher number of recorded wins and a higher value in the average points per match statistic.

Still concerning Figure 7.1 on a success metric and season basis, the second component does not do a good job of distinguishing the categories. It does, however, identify a set of outlier points with abnormally high values corresponding to the 2019/20 season.

To understand how these values came to be, it's useful to know the variables that have the most impact in the second PC – variables with the biggest loadings.

Second Principal Component Loadings

The loadings whose absolute values are greater than 0.15 for the second PC, presented in Table 7.3, show that most variables, five out of twelve, belong to the defensive actions category, followed by four belonging to the passing category. The Possession and General Statistics categories also include two and one variables, respectively, with the number of loose balls, recovered recording the highest absolute weight – variable with the greatest effect on negative values in the second component. Additionally, for each variable (i) and the PC, the Pearson sample correlation coefficient $(Cor(PC_2, X_i))$ and corresponding p-value are displayed under the "Pearson" and "p-value" columns, respectively.

Table 7.3: Loadings whose absolute value is greater than 0.15 for the second PC and Pearson sample correlation coefficient and p-value between each variable and the component.

Category	Variable	Loadings	Pearson	p-value
Pss	ball_losses	-0.167	-0.465	$< 10^{-8}$
Pss	failed_attempts_to_regain_the_ball	-0.224	-0.626	$< 10^{-8}$
P	offsides	-0.150	-0.419	$< 10^{-8}$
P	low_passes	-0.159	-0.445	$< 10^{-8}$
P	high_passes	-0.239	-0.667	$< 10^{-8}$
P	blocked_passes_by_opponent	-0.244	-0.681	$< 10^{-8}$
GS	num_loose_balls_recovered	-0.315	-0.880	$< 10^{-8}$
DA	pressure_attacking_1/3	-0.189	-0.528	$< 10^{-8}$
DA	tackles_middle_1/3	-0.205	-0.572	$< 10^{-8}$
DA	winned_tackles	-0.213	-0.593	$< 10^{-8}$
DA	pressure_middle_1/3	-0.219	-0.611	$< 10^{-8}$
DA	blocked_passes	-0.255	-0.711	$< 10^{-8}$

Since the coefficients are all negative, this is suggestive of an inverse proportionality relationship between PC2 and these variables: it is expected for teams that score higher on PC2 to have lower values on these statistics. This relationship is confirmed by the Pearson's correlation coefficient between this component and number of loose balls recovered ($r = -0.880$).

After a more thorough study of the apparently atypical observations, it was found that not only did they occur in the same season, but they also belong to the same league – FLO.

Previously in Chapter 6 in Section 6.3, it was noted that during 2019/20, contrary to other leagues in Europe, FLO did not resume that year, which resulted in a lower number of games played for these teams. As a direct consequence, we can hypothesise that given a smaller number of games played, variables that are constructed on the basis of counting the number of times a specific action occurred, will show a smaller values. As an example, Figure 7.4 shows the values for the statistic number of loose balls recovered (variable with the highest absolute weight in PC2) are displayed across seasons, which seems to confirm this reasoning given the abnormally low value recorded for FLO in the 2019/20 season, despite fluctuations in the values from the other seasons. This pattern, suggestive of lower counting opportunities, is recurrent across all variables in Table 7.3 and since all these are negatively correlated with PC2, low values on these statistics would mean high values of PC2, thus justifying the outlier points.

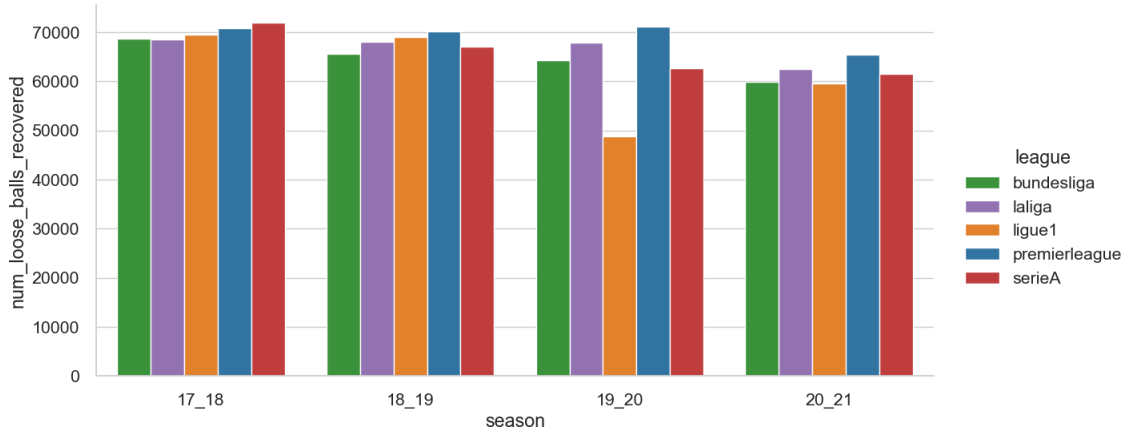

Figure 7.4: The sum of num_loose_balls_recovered for all teams in the different leagues across the seasons studied (2017/18 – 2020/21)

Results from the third PC, when plotted against PC1 (Figure 7.5), is marginally more expressive in terms of the distinction between middle-ranked and low-ranked teams. With PC1 results demonstrating, similarly to Figure 7.1, that high-ranked teams indeed scores higher in this component, the third PC helps in differentiating the middle-ranked and low-ranked categories more clearly with low-ranked teams scoring slightly lower than middle-ranked teams, on average.

To understand what are the most preeminent variables that dictate the behaviour of teams in this category, we must consider the third PC loadings.

Third Principal Component Loadings

The loadings whose absolute values are greater than 0.15 for the third PC are presented in Table 7.4. Additionally, for each variable (i) and the PC, the Pearson sample correlation coefficient $(Cor(PC_3, X_i))$ and corresponding p-value are displayed under the ”Pearson” and ”p-value” columns, respectively.

The variable with the highest absolute loading was the number of touches in the defensive penalty area, closely followed by the number of touches in the defensive third of the pitch, both part of the Pos-

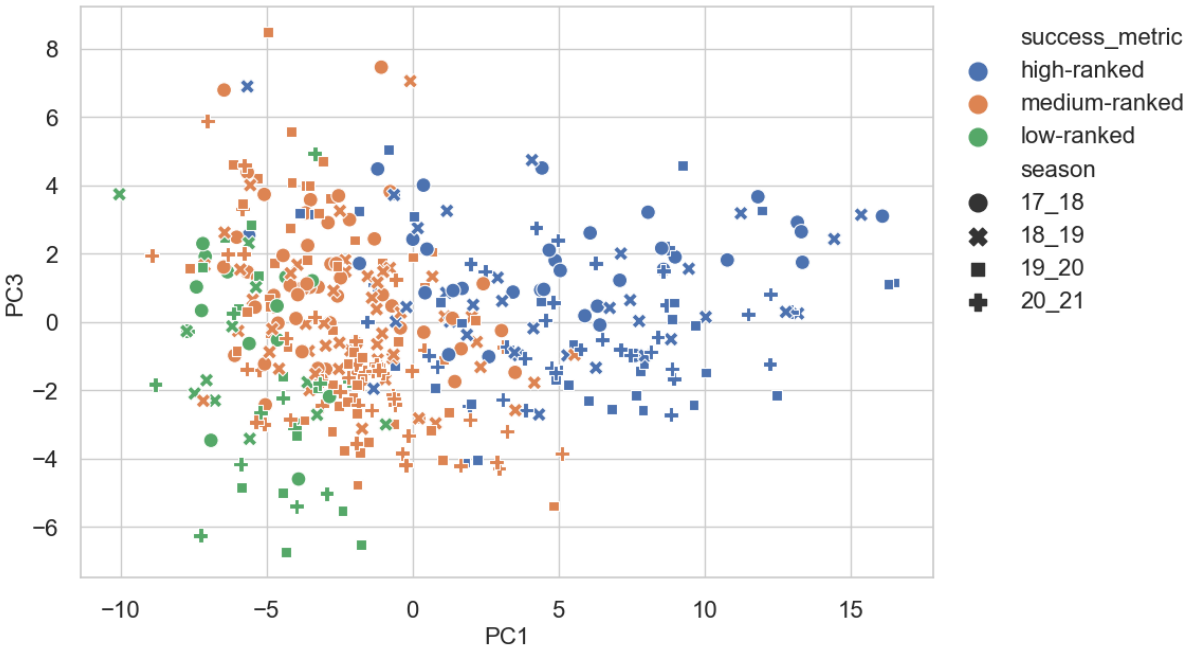

Figure 7.5: Scores on PC1 vs PC3.

Table 7.4: Loadings whose absolute value is greater than 0.15 for the third PC and Pearson sample correlation coefficient and p-value between each variable and the component.

Category	Variable	Loadings	Pearson	p-value
G	clean_sheets%	0.196	0.482	$< 10^{-8}$
G	launched_goalkicks%	0.190	0.468	$< 10^{-8}$
G	passes_average_length	0.174	0.429	$< 10^{-8}$
G	defensive_actions_average_distance	0.165	0.406	$< 10^{-8}$
Pss	touches_in_defensive_1/3	-0.296	-0.729	$< 10^{-8}$
Pss	touches_in_defensive_penalty_area	-0.333	-0.819	$< 10^{-8}$
P	medium_passes_completion%	-0.169	-0.416	$< 10^{-8}$
G	passes_attempted_by_goalkeeper	-0.153	-0.375	$< 10^{-8}$
G	corner_kicks_goals_against	-0.169	-0.417	$< 10^{-8}$
G	penalty_kicks_allowed_against	-0.170	-0.418	$< 10^{-8}$
G	num_saves	-0.192	-0.472	$< 10^{-8}$
G	goals_against_per90	-0.202	-0.497	$< 10^{-8}$
G	throws_attempted	-0.230	-0.566	$< 10^{-8}$
DA	penalty_kicks_conceded	-0.161	-0.395	$< 10^{-8}$
DA	blocked_shots	-0.201	-0.493	$< 10^{-8}$

session category. This finding suggests that these were the statistics with the greatest effect on negative values in the third component and, for this reason, were significant factors in the team's classification at the end of the season. Furthermore, since the coefficients and the Pearson sample correlation measure are negative for the former ($\hat{\phi}_{3,80} = -0.333; r = -0.819$) and latter ($\hat{\phi}_{3,81} = -0.296; r = -0.729$), from here follows that since low-ranked clubs tend to show lower values on PC3, a higher number of occurrences in these statistics can be expected.

This discovery is not surprising, particularly because it concurres with results from the first PC load-

ings analysis. More concretely, considering that some of the variables with the highest importance in PC1 suggest a strong offensive attitude, it follows that not-so-successful low-ranked clubs will most likely display more salient patterns in defensive statistics. Alternatively, however, this could mean that these teams purposefully adopt a low-intermediate block, inviting the higher level teams to move up the offensive lines (block climbing), in an eventual attempt to explore the space behind the defensive line, when regaining the possession of the ball, through counter-attack. However, as football is a game whose outcome depends on the performance of both teams, one reflection to capture here is that, even though it might be the case that in some situations, these statistics can work as an exact representation of a specific style/strategy of play, they closely inform us about the essence of the game itself: a higher collection of values from a defensive perspective indicates that teams that tend to be less successful – low-ranked teams – will gravitate towards a more defensive style of play, due to the, most likely, superior adversary they're facing, whilst high-ranked teams will most likely have a dominant offensive presence.

Besides the Possession group, results show that most variables, ten out of sixteen, belong to the Goalkeeping category, further emphasizing the influence of this aspect of the game for this PC. On the one hand, as having a direct proportionality relationship with the PC, the average distance of the defensive actions stand out, suggesting a low average distance in terms of these actions for low ranked teams as they score the lowest on PC3.

Likewise, these teams also seem to have a lower percentage of launched goal kicks variable (direct proportionality relationship) which, when considered with the passes attempted by the goalkeeper and throws attempted (inverse proportionality relationship) is compatible with the previous conclusions that these teams - low-ranked - lean, on average, towards a more defensive game development.

The two previous patterns in terms of ball possession and the goalkeeper's activity in the game are proven when related variables are considered (see Figure 7.6 and Figure 7.7).

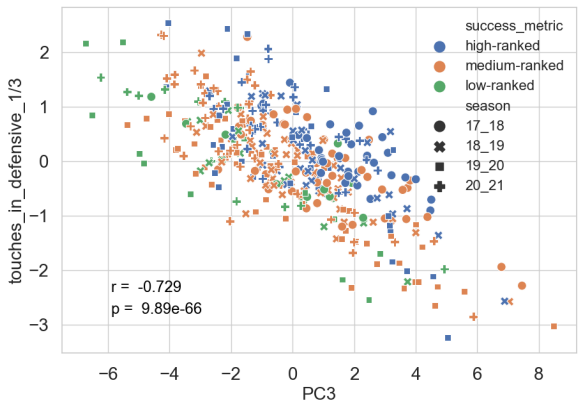

Figure 7.6: PC3 vs touches_in_defensive_third.

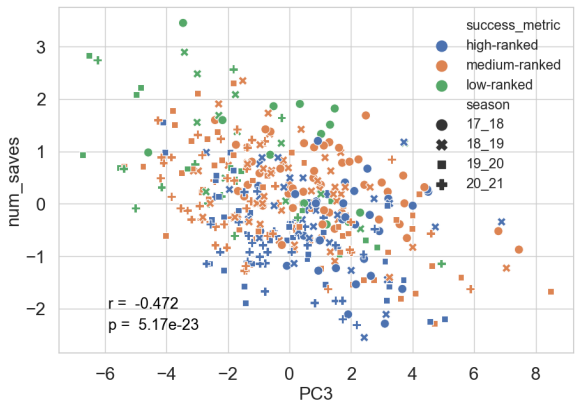

Figure 7.7: PC3 vs num_saves.

When plotted against PC3, the number of touches in defensive third variable, visible in Figure 7.6, demonstrates that excluding some datapoints, the majority of low-ranked clubs are in the top left corner, where the number of touches is higher.

The arrangement of datapoints in Figure 7.7, where the arrangement of points between PC3 and num_saves is considered, is consistent with these findings since it emphasizes the relevance and in-

creased importance of goalkeepers actions in low-ranked teams, as the majority of teams is organised around the top of the graph, with high-ranked teams displaying considerably lower patterns in terms of the number of saves.

Coherently, as a considerable part of the action and interaction between the squads is performed close to the goal of the less dominant team, the likelihood of fouls and corners in this area is higher, as is suggested by the number of penalty kicks and corners which the opposition was able to convert into a goal and from the penalty kicks conceded variable (defensive actions category).

7.1.2 What characterizes/distinguishes the Big-5 European Football Leagues

Concerning the leagues, results demonstrate interesting distinctions when the PCs are plotted against each other with the help of the league categorical variable to identify the points.

When coloured differently, Figure 7.8, an analysis to the first and second components reveals that disregarding the FLO 2019/20 season outliers, the arrangement of points for the SLL, ISA, and FLO leagues is fairly dispersed, not showing any clear patterns. On the other hand, the second PC makes the distinction between GB and EPL from the remaining leagues more evident, with clubs from the two formers consistently scoring lower than the latter group, on average (Table 7.5), across all seasons considered.

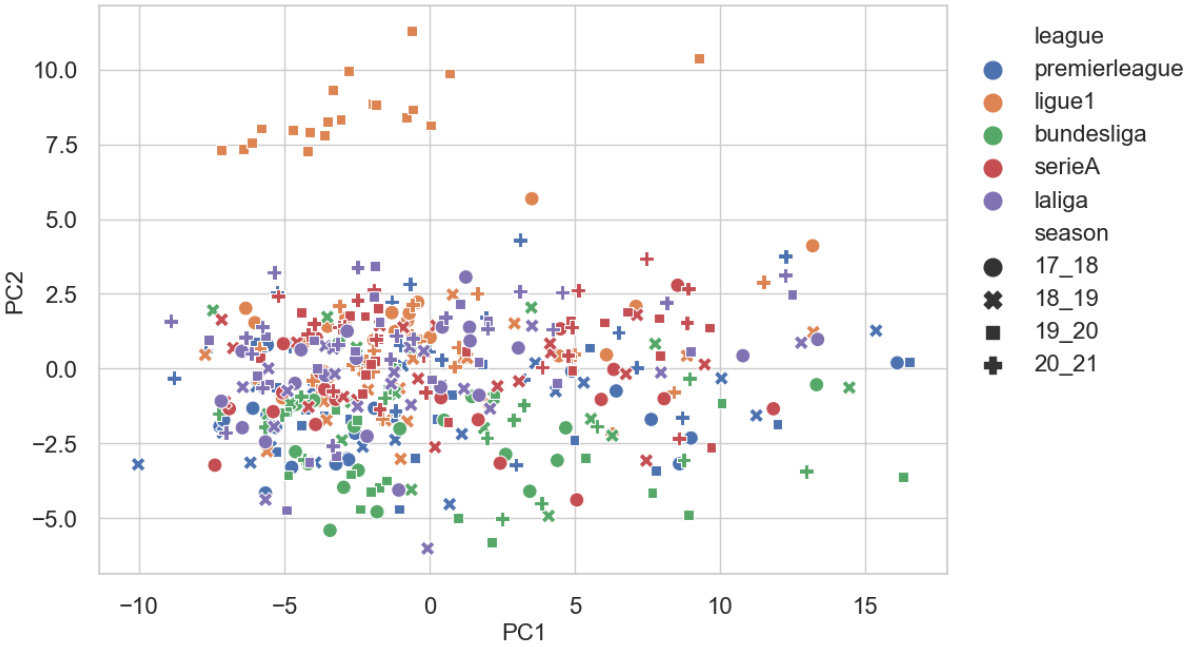

Figure 7.8: Scores on PC1 vs PC2, coloured according with the league.

Table 7.5: Mean and standard deviation of PC2 scores per league.

	Premier League	Ligue 1 (no outliers)	Bundesliga	Serie A	La Liga
Mean	-1.006	0.528	-2.124	0.192	0.165
Std	1.806	1.580	1.822	1.624	1.868

Indeed, specifically for the EPL, the English game is reportedly distinguished by a fast-paced approach whose method of play is supposed to even out the work rate, meaning that all players are required to exert the same amount of effort [97]. This characterization matches our findings since low values on PC2 indicate high values on metrics with an inverse proportionality relationship such as the level of pressure in the middle and attacking third and the number of tackles in the middle third of the pitch (Table 7.3). Above everything, higher values reported in these statistics suggest a style of play characterized by a strong presence in all areas of the pitch which most certainly requires the work and effort of all team members. Still, in the defensive actions category, the number of tackles won and blocked passes variables stand out along with the number of loose balls recovered (variable with the highest absolute weight) from the General Statistics category which is also indicative of a strong collective presence and cohesion – teams who are well positioned altogether will more readily pressure the opposing team and more conveniently regain ball possession.

Surprisingly, findings also point to strong team cohesion in the case of GB clubs, given the lowest mean value score in PC2, as opposed to [98] who stood by the idea that the German style of play relied more prominently on the athleticism of their players rather than emphasizing the collective performance. Hence, authors should not be quick to dismiss the technical and group quality of German teams over the proven [45] physical differences that differentiate these players.

These findings can be interpreted as a signal for the differentiation of the country's and its respective league's playing styles. When analysed in more detail, the arrangement of points between PC2 and the number of times the team pressured the opponent middle and attacking third, as visible in Figure 7.9 and Figure 7.10, corroborate this affirmation as visible from the left-most datapoint groupings for GB and EPL - teams from these leagues display the highest values, on average, in terms of pressure in middle and attacking thirds.

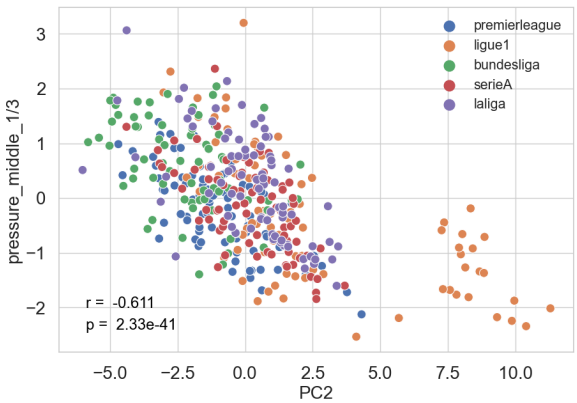

Figure 7.9: PC2 scores vs pressure_middle_third.

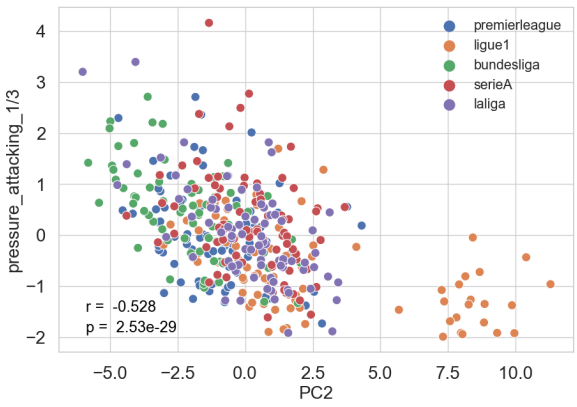

Figure 7.10: PC2 scores vs pressure_attacking_third.

Results from the third PC (Figure 7.11), do not provide unquestionable insights concerning the differentiation amongst the five leagues though they point to higher scores on PC3 from SLL points, on average (Table 7.6), and lower for GB, EPL and ISA (most expressively), across all leagues considered.

Following the statistics identified as having more significance in this loading vector (Table 7.4), the previous assertion suggests higher values on variables with an inverse proportionality relationship and

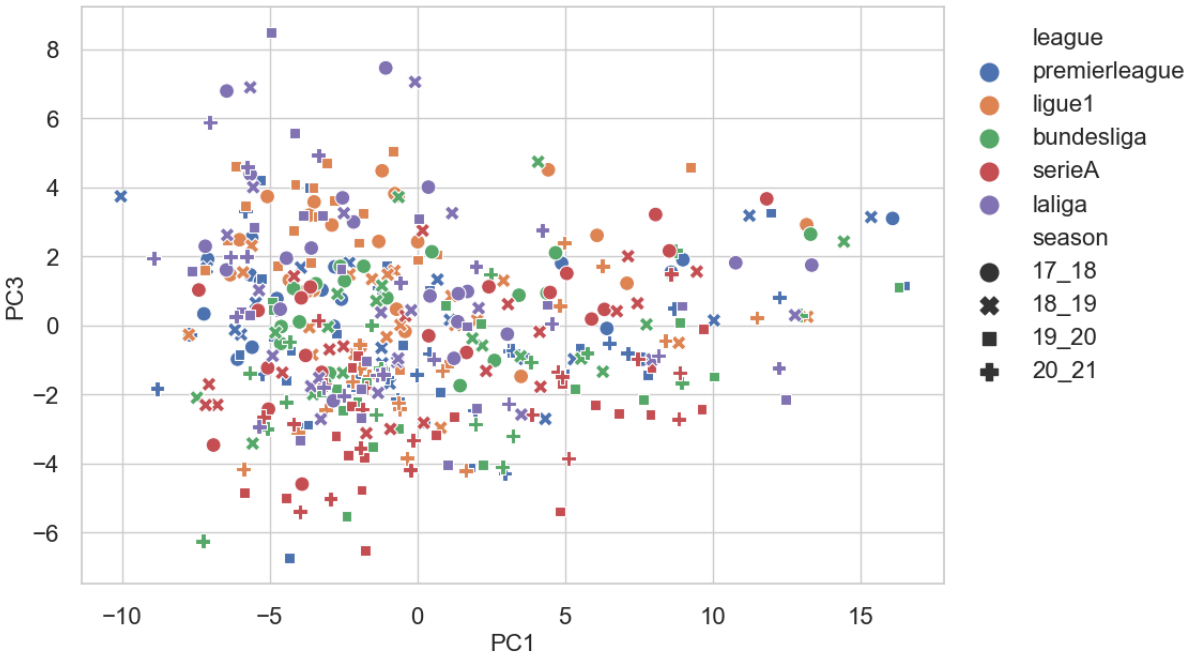

Figure 7.11: Scores on PC1 vs PC3, coloured according with the league.

Table 7.6: Mean and standard deviation of PC3 scores per league.

	Premier League	Ligue 1 (no outliers)	Bundesliga	Serie A	La Liga
Mean	-0.032	0.384	-0.630	-1.471	1.033
Std	1.989	2.129	2.039	2.165	2.764

lower values on variables with a direct proportionality relationship in the aforementioned metrics, on average, for teams belonging to the German, English and Italian leagues.

Previously, in the third PC loading analysis, it was found that in terms of success metric categories, low-ranked teams displayed slightly lower values on this component. From here, one important insight that deserved attention was the fact that this was probably a result of the dynamic of the game itself – in a game where one team dominates the game they will most likely have a strong attacking presence, leaving the other team to withhold itself to a defensive approach. The story told by the previous scatter plot points, nonetheless, in a different direction, with the data points expressing that this PC indeed has some explanatory capacity that helps distinguish this league from the remaining. As noted formerly, the relevance of the defensive variables identified for this PC is unequivocal which, alternatively to the last interpretation, can provide evidence of a deliberate approach to the game with the prevailing ambition of being methodical, specifically aiming to develop their attacks from the defensive section of the field.

This premise is further corroborated when the arrangement of points for PC3 and the number of touches in defensive third is considered, as seen in Figure 7.12, particularly for the ISA teams given the concentration of the datapoints towards the lower end of the data points cloud. As a result, as [45] concludes, the first Italian football league shows indeed patterns of a more cautious, disciplined, and cerebral style of play, specifically aiming to develop their attacks from the defensive section of the field.

Additionally, the fourth PC results give interesting insights into the divergence between the five

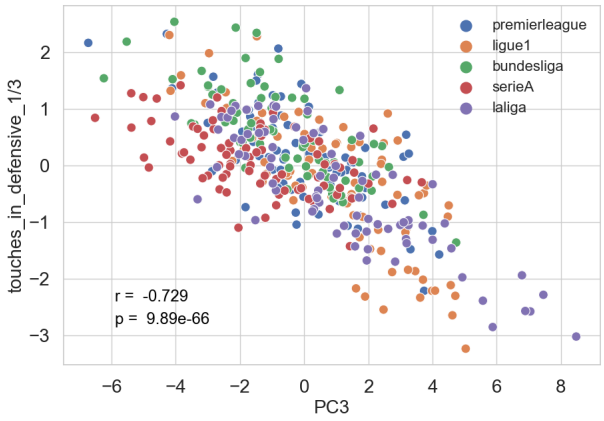

Figure 7.12: PC3 scores vs touches_in_defensive_third.

leagues, classifying more clearly the ISA and SLL as having lower scores and EPL as having higher scores on PC4, on average (Table 7.7), across all leagues analysed, as visible in Figure 7.13.

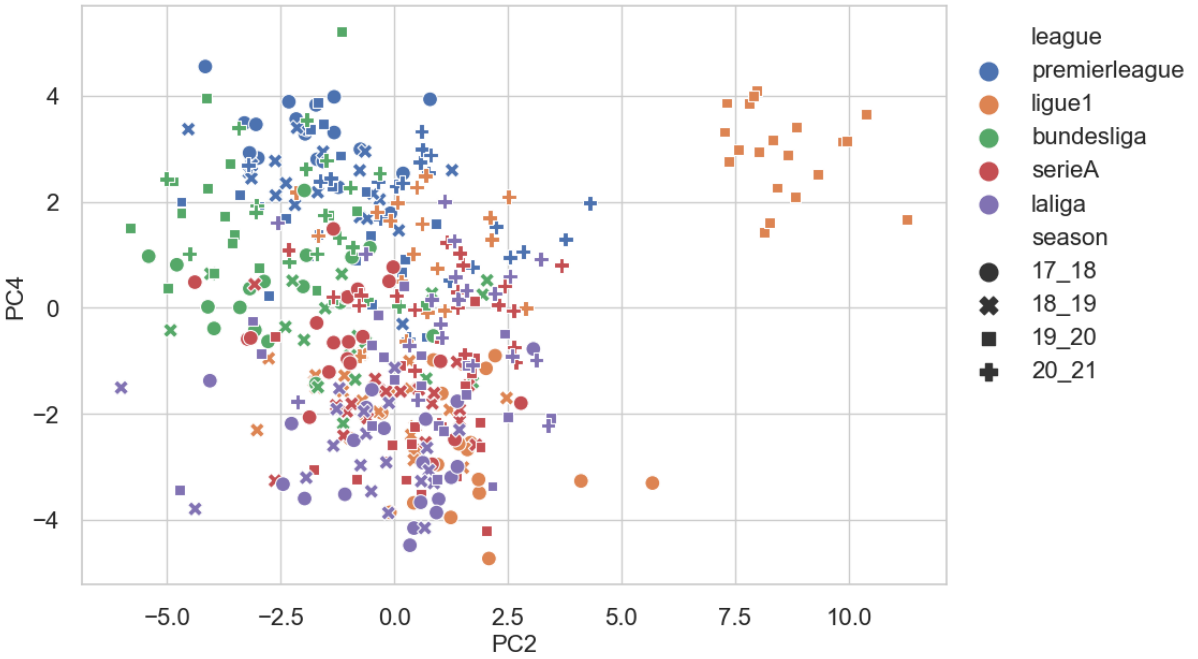

Figure 7.13: Scores on PC2 vs PC4, coloured according with the league.

Table 7.7: Mean and standard deviation of PC4 scores per league.

	Premier League	Ligue 1 (no outliers)	Bundesliga	Serie A	La Liga
Mean	2.272	-1.137	0.811	-1.128	-1.761
Std	1.083	1.846	1.391	1.299	1.505

More concretely, disregarding the FLO 2019/20 season outliers, when plotted together, the second and fourth component's results show that EPL clubs score the highest on PC4, on average, across seasons which is made evident by the relatively close concentration of points in the top left corner of the graph. Scoring slightly lower than the EPL values, on average, are GB clubs mostly from the 2019/20 and 2020/21 seasons, with teams from this league in the 2017/18 and 2018/19 seasons further displaying

lower values in this PC. FLO and ISA reveal significant variation in point distribution, with clubs from the former league scoring high, approaching EPL on PC4 in the 2020/21 season but scoring far lower in the 2018/19 and 2017/18 seasons. With a more consistent arrangement of points, SLL teams score the lowest on PC4, on average, across all seasons except for the last season considered - 2020/21.

To understand how these values came to be, it's useful to know the variables that have the most impact in the fourth PC – variables with biggest loadings.

Fourth Principal Component Loadings

The loadings for the fourth PC whose absolute values greater than 0.15, presented in Table 7.8, allow to make the distinction between several categories. Additionally, for each variable (i) and the PC, the Pearson sample correlation coefficient $(Cor(PC_4, X_i))$ and corresponding p-value are displayed under the "Pearson" and "p-value" columns, respectively.

Table 7.8: Loadings whose absolute value is greater than 0.15 for the third PC and Pearson sample correlation coefficient and p-value between each variable and the component.

Category	Variable	Loadings	Pearson	p-value
P	inswing_corner_kicks	0.278	0.600	$< 10^{-8}$
P	outswing_corner_kicks	0.226	0.486	$< 10^{-8}$
P	straight_corner_kicks	0.191	0.411	$< 10^{-8}$
DA	pressure_to_opponent_completed%	0.217	0.467	$< 10^{-8}$
DA	tackles_completed%	0.169	0.364	$< 10^{-8}$
G	penalty_kicks_allowed_against	-0.160	-0.345	$< 10^{-8}$
GSC	free_kicks_shots	-0.194	-0.418	$< 10^{-8}$
GSC	fouls_lead_to_shoot_attempt	-0.244	-0.525	$< 10^{-8}$
GS	num_red_cards	-0.170	-0.366	$< 10^{-8}$
GS	fouls_commited	-0.253	-0.545	$< 10^{-8}$
GS	num_yellow_cards	-0.283	-0.610	$< 10^{-8}$
GS	fouls_drawn	-0.300	-0.646	$< 10^{-8}$

From the table, it is visible that the variable with the highest absolute weight is the number of fouls drawn, belonging to the General Statistics, together with the number of yellow and red cards and the number of faults committed, still with considerable importance. Based on this information, the interpretation could be made that EPL clubs (highest scoring datapoints in the fourth component), will most likely display low values in these variables, as their negative coefficient denotes a negative correlation. A more thorough analysis to the arrangement of points when the number of fouls drawn to opponent and yellow cards are plotted against PC4, as visible in Figure 7.14 and Figure 7.15, demonstrate, that is precisely the case, as can be seen from the close concentration of English and German datapoints in the bottom right corner. Contrastingly, low reported values in PC4 suggest that SLL and ISA players tend to do the highest amount of infringements and, coherently, receive the highest number of yellow and red cards and have the highest amount of corners from the five leagues.

These findings seem to be in line with [50], who reported in their research that even though EPL is considered the most aggressive league to play in, it also has lowest number of transgressions from

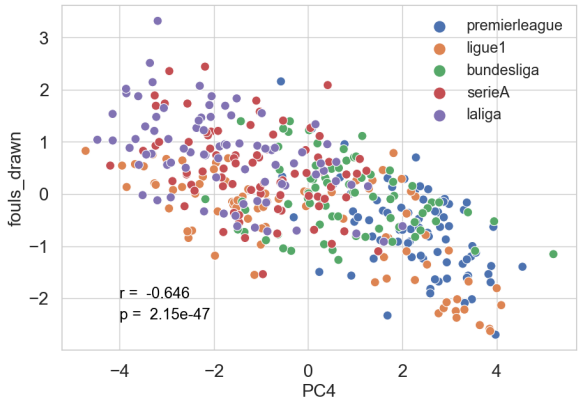

Figure 7.14: PC2 scores vs fauls_drawn.

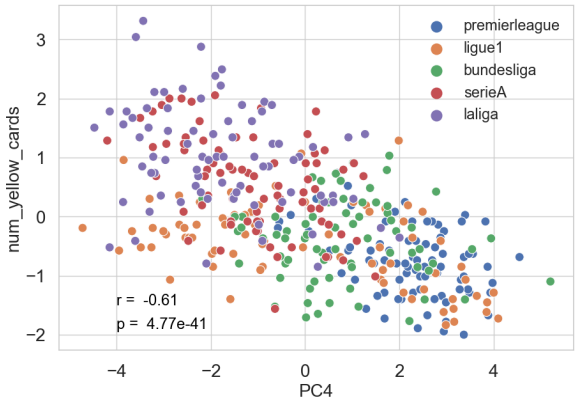

Figure 7.15: PC2 scores vs num_yellow_cards.

players and, consequently, where fewer cards are handed. GB follows, with SLL and ISA being the competitions where most infractions are committed, a conclusion also supported by both in Figure 7.14 and Figure 7.15. Respecting these conclusions, however, it is relevant to add that when it comes to fouls in football, the referee also has a prominent role to play, a commentary also made by [52], given the different levels of compliance they are willing to tolerate.

The (positively correlated) variables from the Defensive Actions category are in line with findings from the analysis of Figure 7.8, as their presence here suggests high values for these statistics in teams that score high on PC4. Thus, the fact that EPL teams are acknowledged as having high levels of cohesion when pressing, is coherent with a high percentage pressure on the opponent, as is corroborated by the scatter plot between PC4 and this variable in Figure 7.15 Likewise, for GB, these metrics further emphasize the importance that team effort also plays in the league's style of play.

As supported by the majority of literature, the previous paragraphs show that there is evidence supporting different characteristics across the Big-5 European Football Leagues.

Concretely, for the 2017/18 to 2020/21 seasons analysed, it was found that EPL teams have a strong overall presence in all areas of the pitch with teammates showing a strong cohesion amongst them both for pressing and thus regaining ball possession and goal-scoring opportunities. Even though not directly analysed, these findings seem to also be in line with the high-intensity patterns that several authors have characterized it as [46, 49–51]. Similar results that suggested a strong collective performance were obtained for the German league, as opposed to the emphasis that is placed on the strength and physical ability of the GB players. Results from the third PC help in distinguishing better ISA, with teams from this country showing patterns of well-thought-off plans of attack with cautious and cerebral approaches to the game stands out, specifically aiming to develop their attacks from the defensive section of the field. Finally, the fourth component revealed that SLL and ISA players committed the highest amount of infringements and, coherently, received the highest amount of yellow and red cards. Interestingly, this findings together with the fact that EPL showed the lowest reported number of fouls drawn and committed might be linked with the actual useful playing time in the leagues and, consequently, be related to the high pace of play reported in this league.

For subsequent analysis, there was the need to retain a number of PCs. This is a very well-known

challenge of PCA, which [99] tackles by providing a great review of known methods and courses of action to tackle this issue, namely:

1. Consider the required number of PCs whose cumulative percentage of total variance is above or equal a specific threshold.
2. Contemplate a cut-off value of 1 (PCs whose explained variance exceed 1 when the original variables are standardized and higher than the mean of the eigenvalues, otherwise) – Kaiser's rule [100].
3. Visually decide upon the number of PCs to consider with the help of the scree plot [101].
4. Other approaches: in the book, the author still acknowledges other alternatives such as distribution-based test tools (e.g., Bartlett's test), which, the author recognizes, frequently suggest retaining more variables than are necessary, and some statistically based methods (e.g., Eastment and Krzanowski's cross-validation technique).

Regarding the first option, despite the fact that a reasonable cut-off is often between 70% and 90%, as stated by [102], the percentage of total variance accepted is often dependent on the useful information about a certain data set and the study being performed. In line with this argumentation, [99] recommends for the threshold to be somewhat lower in cases where the number of statistics is very large since selecting the right number of components to achieve these percentages may still result in a considerable and impractical amount of eigenvectors for further analysis. Even though this was the case in this study, the percentage of variance was considered a priority.

Alternatively, the second option would mean 20 components to carry on the study while the scree plot, visible in Figure B.1 in Appendix B, suggests that a reasonable answer would gravitate towards 3 and 8.

In the end, a cut of value of 80% in terms of the cumulative percentage of total variance was considered, which meant that the first 22 PCs were considered to carry the analysis. The associated eigenvectors allowed us to obtain the scores matrix which represents the new transformed data.

7.2 Clusters Analysis

7.2.1 Hierarchical

To understand the extent to which the Clustering Analysis performed when modelling the outliers identified previously – FLO from the 2019/20 season -, the Cluster Analysis was initially performed with the complete initial information (i.e. all observations).

Results demonstrate that the best performing model considered the ward linkage approach and the Euclidean distance for the dissimilarity metric. Based on the hierarchical structure of the dendrogram and the highest vertical distance that does not intersect with any clusters, as seen in Figure 7.16, which suggested two clusters. The first subgroup is colored orange and the observations that integrate it

belong almost exclusively to the high-ranked teams category or teams from the middle-ranked category whose performance was close to the high-ranked teams. Curiously, there was one observation that stood out from the rest in this category because it corresponded to a low-ranked team. More concretely, the team was SLL's Rayo Vallencano in the 2018/19 season.

Logically, in terms of the categories of success, the second includes mostly observations from the middle-ranked and low-ranked categories.

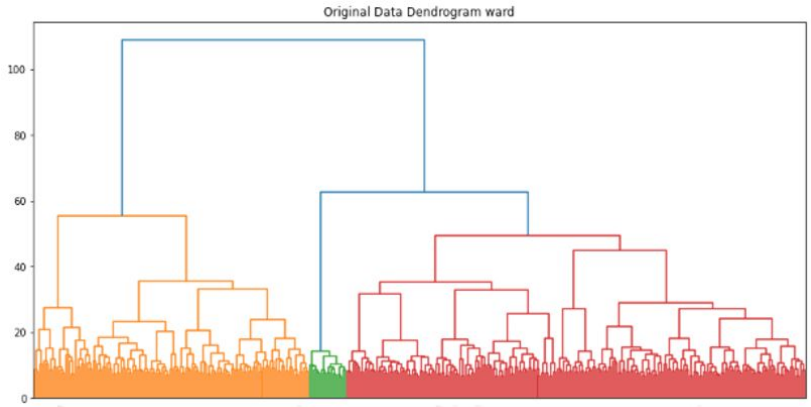

Figure 7.16: Dendrogram obtained from hierarchical clustering all input variables and observations with ward linkage and Eucledian distance.

Finally, even though previous results point to two clusters, the dendrogram also recognizes a small subgroup of observations, given that, from the nature of the analysis, the sooner (lower in the tree) fusions occur, the more similar the observations/groups of observations are to each other. Coloured in green, these observations, that correspond precisely to the outliers' points from the FLO 2019/20 season, corroborate our hypothesis that the model would be able to capture the effects from the COVID pandemic successfully. These results can be more easily comproved in Figure 7.17.

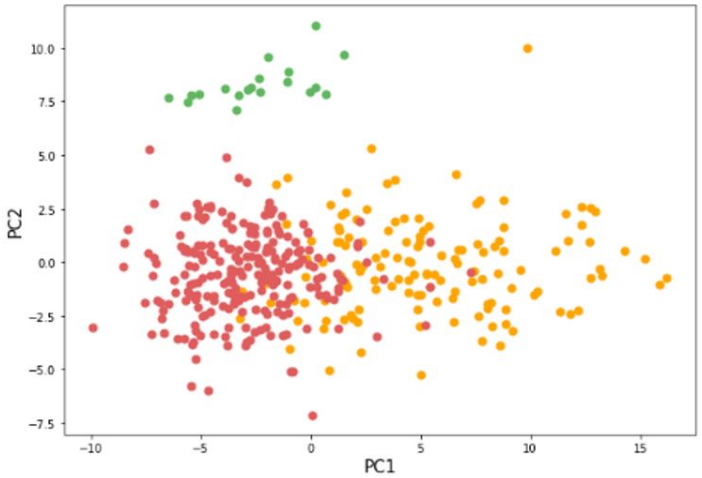

Figure 7.17: Two-dimensional representation (between PC1 vs PC2) of the clusters obtained from hierarchical clustering all input variables and observations with ward linkage and Eucledian distance.

7.2.2 k-Means

Figures in Appendix C show how each team was categorized using k-means Cluster Analysis and the appropriate number of clusters, as well as information on the SC for each point classified into cluster groups.

To evaluate the effectiveness of the models, we examined whether teams were properly sorted into separate categories. In Table 7.9, results revealed that both models, for three clusters $(K = 3)$, classified the same proportion of samples (55% of all samples) accurately but when only looking for the distinction between high-ranked teams from middle-ranked and low-ranked teams (K = 2), these values increased substantially to 80% and 86% when all input variables and observations (original data) were considered and using data projected in the first 22 PCs (PCA data).

On a more detailed level, for K = 2, using original data, we discovered that the majority of high-ranked teams (65%) was correctly categorized into the same group, with this percentage jumping to 88% in the case of middle-ranked and low-ranked clubs. Following a similar pattern, using PCA data, 68% of the high-ranked teams teams were accurately recognized with the other group including 96% correct classifications. Curiously, when considered as part of the same cluster, the middle-ranked and low-ranked categories shows higher precision values than the high-ranked teams when using the original data (0.83 for the former and 0.75 for the latter). Contrary to this behaviour, by using data projected in the first 22 PCs, the model performs worse when modelling the middle-ranked and low-ranked categories than the high-ranked teams one (precision values of 0.85 and 0.90).

For K = 3 using PCA Data, 71% of the low-ranked teams were accurately recognized, but only 60% of the high-ranked teams and 47% of the middle-ranked teams were correctly classified. These values drop significantly for the first two categories when Cluster Analysis was performed with the original data with only 52% and 51% of the teams being properly classified, respectively. On the contrary, the recall for the middle-ranked clubs increased 11 percentage points to 58%. When it comes to the precision metric, there is a predictable pattern, with the high-ranked teams category recording the fewest false positives and the low-ranked category recording the most, for both types of data. More specifically, for the aforementioned categories, the precision values for the original data and PCA models are staggering at 0.90 and 0.98, respectively, with these values dropping to 0.24 in both models for the low-ranked category, suggesting a lower capacity of the clustering algorithm to estimate teams from this category.

Considering our initial hypothesis that Cluster Analysis is more effective in distinguishing high-ranked teams from middle-ranked and low-ranked teams $(K = 2)$ than between all categories of success $(K = 3)$, results show that when middle-ranked and low-ranked teams are considered in the same cluster $(K = 2)$ while using PCA data, this distinction is indeed more effective. Similarly, when the differentiation between the three categories is analysed using this data, as opposed to utilizing original data for this number of clusters, the model appears to perform a better job, exposing the least adequate model to categorize the elements.

The SC results, visible in Table 7.10, demonstrate highest average values being achieved using the original data ($SC_{K=2} = 0.54$ and $SC_{K=3} = 0.48$) with the least compact clusters having occurred when considering $(K = 3)$ using PCA data.

Table 7.9: K-Means Clustering Results

	Original Data			PCA Data		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
$K = 2$						
high-ranked	0.75	0.65	0.70	0.90	0.68	0.77
middle + low-ranked	0.83	0.88	0.85	0.85	0.96	0.90
$K = 3$						
high-ranked	0.90	0.51	0.65	0.98	0.60	0.74
middle-ranked	0.59	0.58	0.59	0.62	0.47	0.53
low-ranked	0.24	0.52	0.33	0.24	0.71	0.36
Accuracy						
$K = 2$	0.80			0.86		
$K = 3$	0.55			0.55		

Table 7.10: Silhouette Coefficient for K-Means clustering algorithm with two $(K = 2)$ and three $(K = 3)$ clusters performed on all input variables and observations (original data) and using data projected in the first 22 PCs (PCA Data).

	$K = 2$	$K = 3$
Original Data	0.54	0.48
PCA Data	0.24	0.11

Knowing that higher values for this internal validation index correspond to a model with better defined clusters, and that according to [82], when SC has a value greater than 0.51, a reasonable structure has been discovered, the SC for both groups indicate that more reasonable cluster structures were found when using original data even though PCA data was unambiguously better in categorizing the elements in their correct group – higher recall values.

7.3 Classification Analysis

For experimental purposes in the classification, a total of six different classifiers were used, and the results of the mean test accuracy scores for each model are displayed in Figure 7.18.

The XGB algorithm performed the best with a mean test score of 86.68%, followed by the two other ensemble techniques RF and AdaBoost, with values for the previous measure of 81.66% and 80.59%, respectively. Scoring slightly lower than the LR algorithm, as the model with the lowest accuracy was the KNN technique (70.42%). The GNB, despite the model's assumptions, still performed remarkably well, with a mean accuracy score of 73.45%.

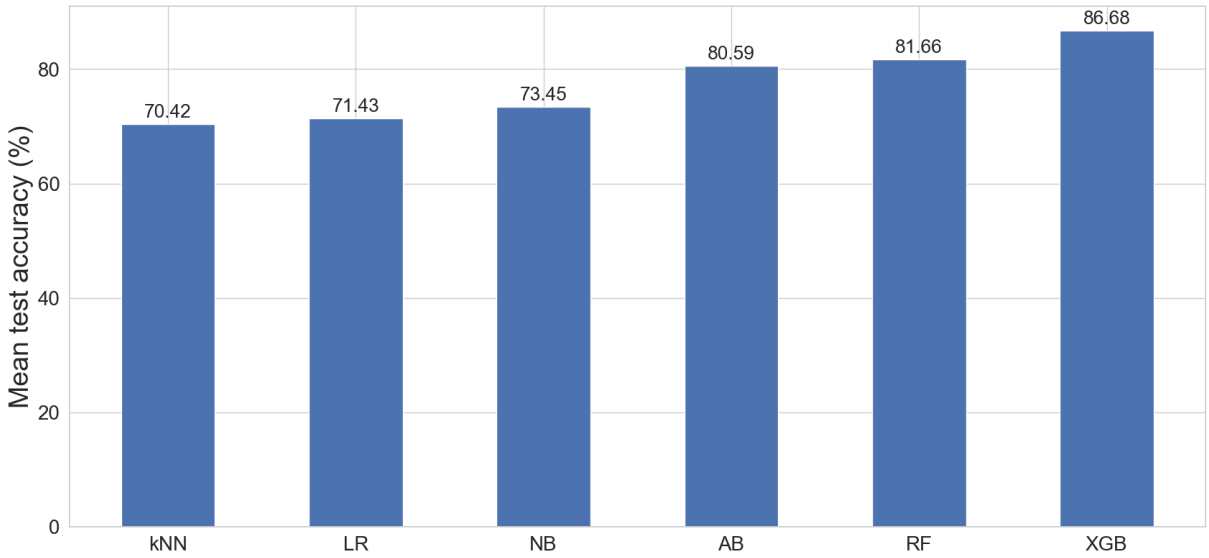

Figure 7.18: Mean test accuracy scores for the applied machine learning models

Gaussian Naïve Bayes

Table 7.11 displays the confusion matrix, precision, and recall data from our GNB algorithm. With a mean accuracy score of 73.45%, it is safe to claim that the results obtained from this model were satisfactory even though it was only the third lowest performing amongst all classifiers.

When compared, it performed a better job in forecasting data points belonging to the high-ranked and low-ranked categories, with significant recall value scores of 0.86 and 0.80.

In terms of high-ranked teams, the model wrongly classified Eintracht Frankfurt e. V., Freiburg and Union Berlin (GB), Nantes (FLO) and West Ham United FC (EPL), from where, two clubs stand out - Eintracht Frankfurt e. V. and Nantes. Having finished eleventh (bottom half) and ninth, based on the 2021/22 season results, the clubs should have indeed been classified as middle-ranked, were it not for the case that the previous year the German team had won the UEL, thus qualifying directly to the UCL the next season, while the French team were the 2021/22 Coupe de France winners, thus being awarded a qualification spot in the UEL group stage. This is, however, a limitation of the study given that the qualification for European competitions is not solely dependent on the national leagues' competitions performance. As such, teams that qualify and show a not as good as expected performance will tend to confuse the algorithm as they are hard to classify.

However, the value obtained in the latter - low-ranked - is the most surprising since it is the highest among all models (along with AdaBoost), only surpassed by the XGBoost algorithm. This indicates that our probabilistic algorithm modelled the occurrence of low-ranked clubs quite well and has a lower yield of false negatives in this case, with only three teams being mistakenly classified: Levante (SLL), Bordeaux (FLO) and Burnley FC (EPL). Not so promising is the low precision value obtained in classifying low-ranked teams (0.63), suggesting a high false positive rate. This is, however, understood when the recall value for the SPSN category is taken into account. With a value of 0.71 for this measure, the model still had some difficulties in drawing distinctions between the other remaining categories as is demonstrated by the seven (wrongly) predicted high-ranked and low-ranked clubs in each one.

All teams wrongly classified as high-ranked were indeed teams that performed well overall in the season, coming just one or two places short of qualifying for European competitions. Similarly, the same happened when the model attempted to draw the frontier between middle-ranked and low-ranked clubs, with the majority of the seven clubs ranking fifth and fourth last in the season ranking.

Table 7.11: Gaussian Naïve Bayes results (overall accuracy of 0.76).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	30	5	0
Actual middle-ranked	7	34	7
Actual low-ranked	0	3	12
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.81	0.86	0.83
middle-ranked	0.81	0.71	0.76
low-ranked	0.63	0.80	0.71

Logistic Regression

The second classifier considered was the multinomial LR.

In terms of hyper-parameter tuning, the relevant factors to consider included the maximum number of iterations (max_iter), solver, penalty and C. Whilst the first is self-suggestive and is related to the maximum number of iterations taken for the solver to converge, the following simply defines the algorithm used in the optimization problem. The latter two are intrinsically related since the penalty defines the form of regularization that will be used to reduce overfitting and C specifies the strength of this regularization.

Given the low amount of hyperparameters to tune, the only approach considered towards this goal was grid search, and the findings for C and max_iter are shown in Figure 7.19 in terms of overall accuracy. Results demonstrate that using the saga solver and l1 penalty method, results were better. Another insight that can be derived from this analysis is that the model's performance is somewhat invariant to the maximum number of iterations in the range considered.

After tuning the hyperparameters, results from the multinomial LR algorithm can be visualized in Table 7.12.

Overall, the algorithm did a fair job at modelling all categories, despite considerably lower recall values of 0.8 and 0.73 in classifying high-ranked and low-ranked data points than the GNB.

The model misclassified Koln (GB) in addition to the previous misclassified clubs in this league for the GNB model, Real Sociedad (SLL), the same club in the English league as before and Fiorentina (ISA), totalling seven wrongly forecasted high-ranked teams. For the low-ranked category, the model misclassified Metz (FLO) and Granada CF (SLL) in additionally to the previous two clubs wrongly classified - Bordeau and Burnley FC.

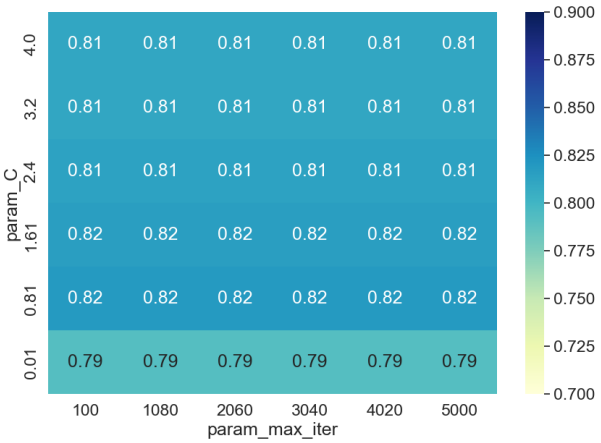

Figure 7.19: Hyperparameter tuning for the C and max_iter parameters.

Table 7.12: Logistic Regression results (overall accuracy of 0.80).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	28	7	0
Actual middle-ranked	6	39	3
Actual low-ranked	0	4	11
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.82	0.80	0.81
middle-ranked	0.78	0.81	0.80
low-ranked	0.79	0.73	0.76

Conversely, the model performed better in classifying actual middle-ranked teams with a total amount of nine false negatives (six were predicted as high-ranked and three as low-ranked teams) With a respectable value in terms of precision for this category of 0.79, the model did not do such a good job in terms of false negatives as is demonstrated by the confusion matrix cell respecting the Actual low-ranked and Predicted middle-ranked value of four.

K-Nearest Neighbors

The following algorithm considered was the KNN.

The main considerations addressed for hyper-parameter tuning included the number of neighbours (n_neighbors) and distance measure (p), which, according to [103] are the main factors that affect how effectively a KNN classifier performs. Additionally, the weights and leaf size were considered. While the first refers to the number of neighbours to employ by default for neighbour queries, p is simply a power parameter associated with how the distances are computed by the algorithm. Then, the weights parameter relates to the weight function used in classification while the last factor determines the minimum number of points in a node. To tune these parameters, only grid search was utilized, as in the prior

procedure.

The findings show that when p = 1 and the points were weighted by the inverse of their distance – closer neighbours of a query point had a stronger effect than neighbours further away –, the model produced superior results. Furthermore, the values of the hyperparameter optimization for the number of neighbours and leaf size, as shown in Figure 7.20 in terms of overall accuracy, reveal that seven is the most appropriate number of neighbours to consider, with the model's performance demonstrating an invariant behaviour concerning the leaf size factor.

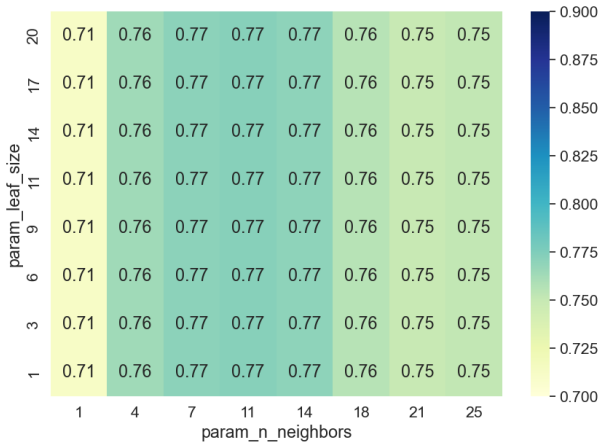

Figure 7.20: Hyperparameter tuning for the param_leaf_size and param_n_neighbors parameters

The outputted results from this approach displayed in Table 7.13, under the format of a confusion matrix and measures of performance table, suggest an equal capacity of the model to classify data points belonging to the high-ranked category and better performance in classifying middle-ranked clubs than the LR model, with a recall value of 0.85 (compared to the value of 0.81).

Table 7.13: K-Nearest Neighbors results (overall accuracy of 0.76).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	28	7	0
Actual middle-ranked	4	41	3
Actual low-ranked	0	10	5
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.88	0.80	0.84
middle-ranked	0.71	0.85	0.77
low-ranked	0.62	0.33	0.43

Conversely, the present model obtained the lowest precision values, when compared to the remaining classification models, in the middle-ranked and low-ranked categories. These suggest an uncommonly high amount of false positives but these values still pale in comparison to the recall value attained in the latter category. With a value of 0.33 for the ratio between the number of correctly classified low-

ranked clubs to the total amount of low-ranked clubs in the test set, this value underlines the model's poor performance in modelling this category (also made evident by the number of wrongly classified low-ranked clubs – ten – in the confusion matrix).

Adaptive Boosting

AdaBoost was the fourth algorithm taken into account. The learning rate (learning_rate), the number of estimators (n_estimators), and the algorithm were key considerations for hyper-parameter tuning. Whilst the first is related with the weight applied to each classifier at each boosting iteration, the second specifies the maximum number of estimators at which boosting is terminated while the last determines the boosting algorithm. Similar to the two previous models, grid search was the sole strategy taken into consideration to achieve this aim due to the small number of hyperparameters to adjust.

The model yielded improved results when the SAMME discrete boosting method was utilized, according to the findings. Additionally, results of the learning rate and number of estimators hyperparameter optimization, as shown in Figure 7.21, in terms of overall accuracy, point to increased performance for low values of the former component and large values of the later.

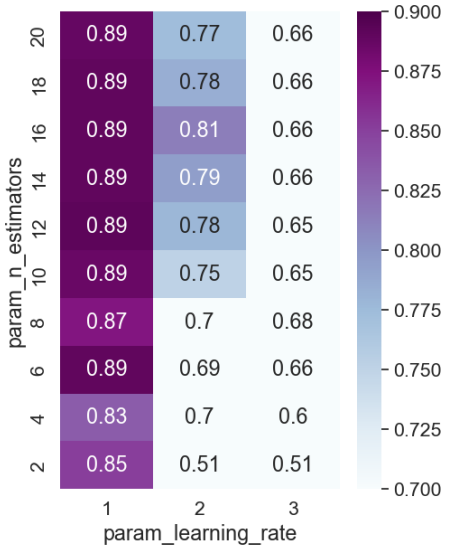

Figure 7.21: Hyperparameter tuning for the param_learning_rate and param_n_estimators parameters

An examination of the AdaBoost confusion matrix, precision, and recall results, shown in Table 7.14, revealed that the model was able to forecast all categories almost flawlessly.

Unlike most previous classifiers, where the low-ranked category usually has the lowest precision values, this model achieves a precision value of 0.92, the second highest from all models, only surpassed by the XGBoost. This emphasizes the model's low number of false positives in forecasting the remaining categories, with the model just mistakenly classifying one middle-ranked team as a low-ranked club - Unione Sportiva Salernitana 1919 (ISA). Having finished the fourth last (with just one additional point from the third last), the model's confusion is understandable as it came extremely close to being relegated.

As mentioned previously, this model is only capable of matching the GNB model in modelling the

low-ranked category, mistakenly classifying Granada CF (SLL), Burnley FC (EPL) and Hertha BSC (GB), with a recall value of 0.8.

It shows, however astonishing results when compared to the three previous in modelling the other two categories, with the algorithm just wrongly forecasting Eint. Frankfurt, Nantes and Manchester United FC. The later club from the English league is new to the group of misclassified teams and after finishing sixth in the final rankings (thus qualifying based on their performance in the league), it was a surprise to see the algorithm wrongly classifying this club.

Table 7.14: Adaptive Boosting results (overall accuracy of 0.89).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	32	3	0
Actual middle-ranked	4	43	1
Actual low-ranked	0	3	12
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.89	0.91	0.90
middle-ranked	0.88	0.90	0.89
low-ranked	0.92	0.80	0.86

Extreme Gradient Boosting

As the fifth classifier and the second ensemble approach studied was XGBoost. The key elements addressed for hyper-parameter tuning were gamma, lambda (reg_lambda), alpha (reg_alpha), and the learning rate. Gamma is the lowest loss reduction necessary to build another partition on a tree leaf node, whereas lambda and alpha are measures of regularization strength. Unlike earlier techniques, we used both random and grid search to improve our gradient boosting model due to the large number of hyper-parameters that needed to be tuned. Making use of the random search to narrow down the range of values analysed for each parameter in the grid search, the results of this hyper-parameter optimization analysis between the gamma and learning rate, and lambda and alpha coefficients are displayed in Figure 7.22 and Figure 7.23, respectively, in terms of overall accuracy.

According to Figure 7.22, the smaller the learning rate and the greater the gamma, the worse the model performed. More specifically, better results are accomplished for low gamma values (with the learning rate exhibiting rather uniform behaviour in this instance) or for gamma values in the range of 55 to 76 and learning rate values near 1. Displaying more stagnant values is the model performance concerning the various values of lambda (reg_lambda) and alpha (reg_alpha), with the model only showing a significantly homogeneous performance for the range of values considered.

From the XGBoost confusion matrix, precision, and recall are shown in Table 7.15. With a mean accuracy score of 87.78%, this model outperformed all other algorithms and, when tested, despite

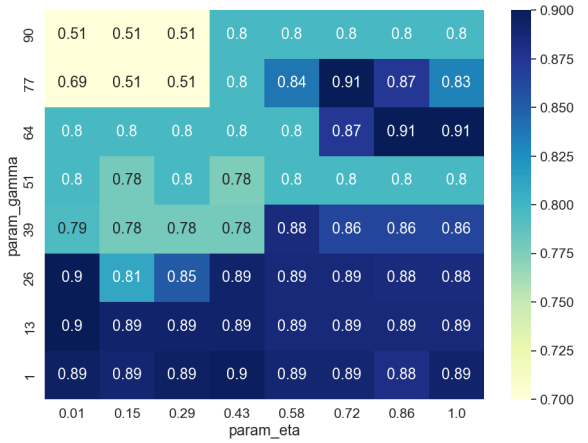

Figure 7.22: Hyperparameter tuning for the param_reg_alpha and param_reg_lambda

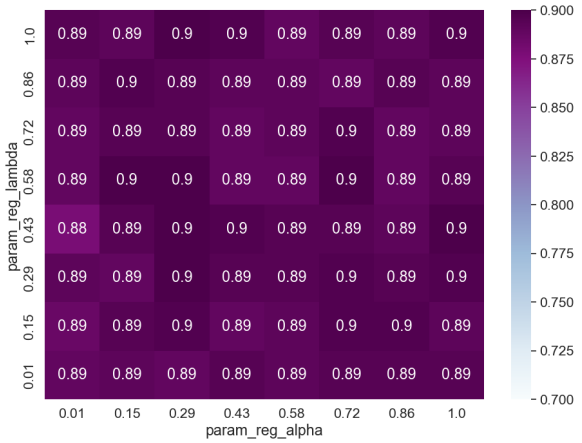

Figure 7.23: Hyperparameter tuning for the param_eta and param_gamma

performing unquestionably well and showing balanced results across all three categories, the model stands out in the classifications made for the high-ranked category. With a recall value of 0.94, the model's only mistakes included the aforementioned teams whose performance in the league didn't mirror the performances of teams that would have ended up being classified for UEFA competitions - Eintracht Frankfurt e. V. and Nantes.

Displaying an F1-score value of 0.90 for the previous category data points (the highest of all algorithms), the model suggests both a low false positive and false negative rate, as indicated by its precision and recall values of 0.93 and 0.87, respectively. Like the first ensemble technique, only Unione Sportiva Salernitana 1919 (ISA) was mistakenly taken by a low-ranked team even though it finished in the middle-ranked category, yielding a precision value of 0.93 to the low-ranked category.

Table 7.15: Extreme Gradient Boosting results (overall accuracy of 0.91).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	33	2	0
Actual middle-ranked	4	43	1
Actual low-ranked	0	2	13
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.89	0.94	0.92
middle-ranked	0.91	0.90	0.91
low-ranked	0.93	0.87	0.90

The model was still able to improve on the previous one (recall value of 0.87 when compared to the 0.8 value for the AdaBoost) in terms of modelling the low-ranked category, with the algorithm just mistakenly taking Hertha BSC (GB) and Granada CF (SLL) for middle-ranked teams even though they belong to the low-ranked category. Coherently, both these teams were third last in the championship with the first finishing with the same amount of points as the fourth last-placed club and the Spanish

team finishing just one point shy from their subsequent upper-ranked counterpart.

Also with a low false positive rate on low-ranked, are the only four teams mistakenly classified as high-ranked: Lens, Lyon, Strasbourg (FLO) and Atalanta (ISA). Concerning the latter, the model's confusion is understandable since the team finished just three points away from the seventh-ranked club. however, the former teams' inclusion in this group might be because Nantes (belonging to the high-ranked category) finished lower on the ranking than these, resulting in a precision value of 0.89 for the high-ranked category.

Random Forest

The last classifier considered was the RF. In terms of hyperparameter tuning, it took less time to find the ideal values than it did for the (extreme) gradient boosting model since there were a smaller number of parameters that needed to be tuned, even though both random and grid searches were examined. Explicitly, these included the criterion (function to measure the quality of a split), the minimum number of samples required to split an internal node (min_sample_split) and the maximum depth of the tree (max_depth).

Results demonstrate that when using the Gini Index as the splitting criterion, contrary to the Cross-Entropy, the results yielded by the model were slightly better. Furthermore, the findings for the max_depth and min_sample_split hyperparameters, shown in Figure 7.24, demonstrate that the model yielded improved accuracy for low values of the minimum number of samples needed to split an internal node and high values of the maximum depth of the tree.

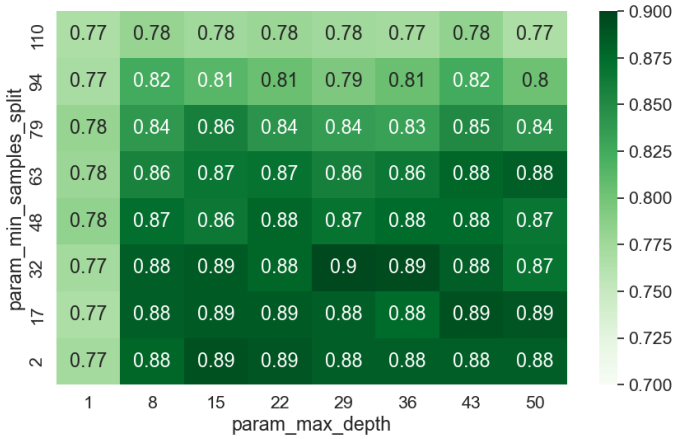

Figure 7.24: Hyperparameter tuning for the param_max_depth and param_min_samples_split parameters

This model achieved a mean test accuracy of 80.65%, scoring just below the best-performing model.

As expected, results from testing the algorithm, shown in Table 7.16 through its confusion matrix and measures of performance table, appear to be nearly identical to the XGBoost values, with the only difference appearing in the slightly lower number of true positives for the low-ranked category. This result suggests a slightly lower capacity for modelling low-ranked clubs, the most unlikely category to be part of, as is evident from the difference of nine percentage points in the F1-score. As the false positives occurrences did not change (only Unione Sportiva Salernitana 1919 was classified as a low-

ranked team, having finished in the middle-ranked category), this change must be due to a noteworthy higher amount of false negatives, as is corroborated by the recall values for both models (0.87 and 0.73 for XGBoost and RF, respectively). In addition to the two previously wrongly identified low-ranked teams in XGBoost - Hertha BSC (GB) and Granada CF (SLL) -, the classifier still placed Levante (SLL) and Burnely (EPL) in the same group.

Table 7.16: Random Forest results (overall accuracy of 0.89).

(a) Confusion matrix			
	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	33	2	0
Actual middle-ranked	4	43	1
Actual low-ranked	0	4	11
(b) Measures of performance			
	Precision	Recall	F1-score
high-ranked	0.89	0.94	0.92
middle-ranked	0.88	0.90	0.89
low-ranked	0.92	0.73	0.81

The classification results based on the test set showed that the worst performing models were the GNB, LR, and KNN, with the latter having the worse performance in modelling low-ranked teams - F1 score of 0.43 -, even though it did a fairly good job in the other categories. Surprisingly, the GNB performed better than initially expected. However, all three classifiers still had some difficulty in drawing a clear distinction between the categories as is emphasized by the high false positive incidence in terms of middle-ranked clubs.

Showing considerably improved overall performance were the ensemble techniques, as demonstrated by the consistently higher F1-scores. In the end, the XGBoost was the model that better classified teams in terms of the categories of success.

Teams such as Eintracht Frankfurt e. V. and Nantes were identified from the initial models as not complying with the pattern as they were always classified as a middle-ranked team even though they were high-ranked teams. As explained previously, this is a consequence of some exceptions regarding the UEFA competition qualification methods and their presence here shows that the model does not respond well in these cases.

Likewise, teams such as Burnley FC, Hertha and Granada CF were consistently classified as middle-ranked teams even though they were all low-ranked in the 2021/22 season. However, the classifiers' errors, in this case, are understandable as they were teams whose performance in terms of points obtained in the end came extremely close to the fourth-last teams of their respective league (last team belonging to the middle-ranked category).

ROC curves were constructed to expand the analysis of the three best-performing models. Because an individual error rate, like accuracy, is measured just for a specific classification threshold, examining a ROC curve might reveal itself helpful and preferable in some cases.

Results from the ROC curves, as shown in Figure D.1, Figure D.2, and Figure D.3 in Appendix D, demonstrate that the XGBoost classifier outperforms the AdaBoost and RF classifiers in terms of overall performance because its curves are placed the furthest away from the ROC space's 45-degree diagonal - higher proximity to the baseline random classifier (expected to yield points lying along the diagonal) signals a less accurate test. Coherently, the area under the ROC curve (abbreviated to AUC) is greater for the XGboost than for RF classifier, with the former displaying a value for the AUC in the micro-average ROC of 0.93, while the latter obtained a value of 0.92. As previously noted in Chapter 6 in Section 6.3, when it comes to the distribution of clubs in terms of the categories of success, there still is a considerable discrepancy in the group sizes with majority of clubs belonging to the middle-ranked category which means that the the micro average measure is indeed the measure to evaluate rather than the macro-average that treats all categories equally.

Based on the findings from the last paragraphs, two additional experiments were considered. Firstly, the classifiers were run to understand the extent to which they assign observations to categories when using data projected in the first 22 PCs. Second, Eint. Frankfurt and Nantes (previously demonstrated to affect all classifiers' performances) were removed to understand how their performance would improve.

Results, as displayed in Figure 7.25, showed that when using PCA data, the models scored consistently lower in all algorithms, with a more noticeable impact in the ensemble techniques. An analysis of the confusion matrices for these models revealed that in no-ensemble techniques, the models had a higher overall capacity to model middle-ranked teams, while, in turn, the capacity for all algorithms to model low-ranked teams clubs decreased considerably.

On the other hand, findings also show that, when not accounting for the two exceptions, all models returned improved performance in terms of mean test accuracy scores, with XGB reaching an astonishing mean test value score of 87.5%.

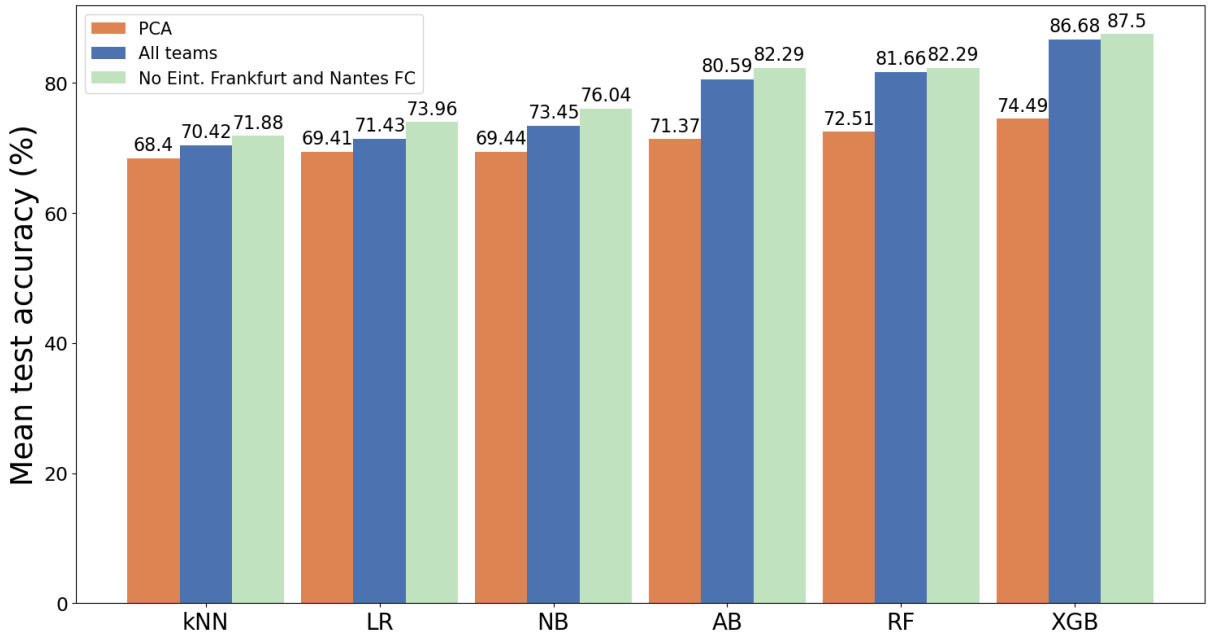

Figure 7.25: Mean test accuracy for the applied classifiers before (blue) and after (green) removing the Eintracht Frankfurt e. V. and Nantes and when using data projected in the first 22 PCs (orange).

Chapter 8

Conclusion

A summary of this dissertation will be presented in this Chapter, together with the closing observations. Furthermore, future avenues of research are also provided.

8.1 Final Remarks

Since the late nineteenth century, the world's most popular sport has experienced advancements both within and outside the field. Concerning the latter, it is true that even though the first steps in terms of data gathering were done in the beginning of the $20th$ century, only with the turn of the century came advancements in the fields of AI, DS, and ML that allowed more data-driven behaviours, and solutions.

According to the literature, investigations using these approaches have covered the technical-tactical, physical, and mental domains, with a focus on match analysis to understand what determines success in the sport. For this purpose, some academics focus their investigations on the three potential game outcomes (wins, draws, or losses), while others look at the team's overall rating when analysing different international and national competitions. However, it is the case that most of the analyses still look into a small number of variables and second, the majority of them also fail to be comprehensive on all types of actions and interactions that can occur in a game. Furthermore, it indicated that certain studies do hint at a level of differentiation across the Big-5 European Football Leagues, although research that explores all of them together is still scant. Finally, some academic work in predictive analysis has been developed, with the bulk of the study aimed at projecting either the final result of matches or the team's final ranking at the end of the season. Several studies show decent outcomes (depending on the algorithm), however, the majority of the models do not generalize well since they are trained using data from only one league and/or season.

Given these limitations, it was concluded that by broadening the scope of analysis to include performance indicators about goalkeeping, defensive actions, types of passes (rather than the number) and the location from which these occurred in the game, other relevant patterns could emerge not only in terms of success categories but also leagues.

In the end, 98 KPIs sourced from the FBREF website were investigated over the course of four

seasons (2017/18 - 2020/21) for the Big-5 European Football Leagues.

With this dissertation, in terms of the categories of success, the PCA scatters and first component loading's analysis demonstrated firstly that high-ranked teams displayed higher patterns of ball possession, and a playing style reliant on exchanging the ball more often with the intent of reaching the attacking third of the pitch. Indeed, these teams showed increased activity around the final third of the field (through the number of carries into the attacking third and goal box and the number of touches in the offensive third) when compared with medium and low-ranked teams. Secondly, it was found that high-ranked teams performed, on average, a higher number of passes leading to a goal and shot attempts. However, exceptions to this rule as Borussia Dortmund are proof that teams can also find success in a more direct method of play, or based on the counter-attack, which is characteristically quicker.

Thirdly, aiming at maximizing the goals scored and shot attempts per game seemed to be the rule amongst high-ranked clubs rather than trying to maximize the attempts' effectiveness. Conversely, passing accuracy when exchanging the ball stood out.

When analysed, the second component captured (more distinctively) the effects of the COVID-19 pandemic, particularly in the French league, given the abnormal values reported in the 2019/20 season.

The third component results emphasized the low-ranked team's salient patterns in defensive statistics, as they displayed the highest values in terms of the number of touches in the defensive third, despite some cases where these teams purposefully adopt a low-intermediate block, inviting the higher level teams to move up the offensive lines (block climbing), in an eventual attempt to explore the space behind the defensive line through counter-attack.

Concerning the distinctions between the leagues, the EPL team's strong overall presence in all areas of the pitch stands out, in line with the reported high-intensity patterns that commonly characterize it. Similarly, results suggested a strong collective performance for the German teams (as opposed to the usual emphasis that is placed on the strength and physical ability of these league players). Additionally, ISA clubs showed a prominent use of the defensive section of the field to develop their attacks, while the fourth component revealed that SLL and ISA players committed the highest amount of infringements and, coherently, received the highest amount of yellow and red cards. On the other side of the spectrum were EPL clubs with the lowest reported number of fouls drawn and committed, which may have some impact on useful playing time and consequently be related to the high pace of play reported in this league.

The Hierarchical clustering using ward's method and Eucledian distance conducted was successful in identifying the outliers from the French league during the 2019/20 season, whilst the k-means algorithm revealed that improved accuracy was achieved in distinguishing high-ranked teams from middle-ranked and low-ranked teams, rather than distinguishing the three groups.

Finally, results from the classification models showed that ensemble techniques achieved the best results, with XGBoost leading with an accuracy of 86.7%.

8.2 Future Avenues of Research

Following this dissertation, it would be interesting to comprehend the extent to which the consideration of the FLO 2019/20 season outliers impacted the results. In this matter, carrying the same analysis with statistics that have the same number of games as basis could provide information of whether some KPIs were under or overestimated in terms of their capacity to distinguish clubs in terms of the categories of success or even the leagues.

On the another hand, building on what has been done in this investigation, future work should aim at selecting, from the 98 KPIs analysed, the ones highlighted throughout the study with the purpose of studying it(them) more deeply.

Other future studies might focus on applying the classifiers to only a partially complete data set (for example mid-season) to determine the extent to which they are capable of predicting the teams' final rankings in terms of success categories before the season's end, as having this information could help teams implement opportune resolutions towards the club's end goal. However, in this case, there would be the need to predict what would occur for the remaining portion of the season by considering, for example, different scenarios (optimistic, pessimistic and realistic) based on a percentage based improvement of the variables. Finally, an in-depth analysis could be carried out to understand what leads to the correct or incorrect classification of teams from all classifiers studied.

Bibliography

[1] J. Lago-Ballesteros and C. Lago-Peñas, "Performance in team sports: Identifying the keys to success in soccer," *Journal of Human Kinetics*, vol. 25, no. 1, pp. 85–91, 2010.

[2] Y. Li, R. Ma, B. Gonçalves, B. Gong, Y. Cui, and Y. Shen, "Data-driven team ranking and match performance analysis in Chinese Football Super League," *Chaos, Solitons & Fractals*, vol. 141, pp. 1–9, 2020.

[3] H. Sarmento, R. Marcelino, M. T. Anguera, J. Campaniço, N. Matos, and J. C. Leitão, "Match analysis in football: a systematic review," *Journal of Sports Sciences*, vol. 32, no. 20, pp. 1831–1843, 2014.

[4] P. Chapman, *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS, 2000.

[5] T. Ajadi, A. Clarke, S. Dhillon, G. Gardner, D. Garg, T. Hammond, A. Malcolm, J. Pang, J. Pugh, and D. Jones, "Deloitte annual review of football finance 2022," 08 2022. Accessed: 06-09-2022.

[6] T. H. Davenport, "Analytics in sports: The new science of winning," *International Institute for Analytics*, vol. 2, pp. 1–28, 2014.

[7] P. Thakkar and M. Shah, "An assessment of football through the lens of data science," *Annals of Data Science*, vol. 8, no. 4, pp. 823–836, 2021.

[8] R. Bryant, R. H. Katz, and E. D. Lazowska, "Big-data computing: creating revolutionary breakthroughs in commerce, science and society," *CCC - Led White Papers*, pp. 1–7, 2008.

[9] Online Cambridge Dictionary, "Meaning of data." `https://dictionary.cambridge.org/dictionary/english/data`, 2022. Accessed: 2022-05-06.

[10] Oxford Learner's Dictionary, "Definition of data noun from the Oxford Advanced Learner's Dictionary." `https://www.oxfordlearnersdictionaries.com/definition/english/data?q=data`, 2022. Accessed: 2022-05-06.

[11] A. McAfee and E. Brynjolfsson, "Big data: The management revolution: Exploiting vast new flows of information can radically improve your company's performance," *Harvard Business Review (October 2021 edition)*, 2012.

[12] J. S. Eaves, “A history of sports notational analysis: a journey into the nineteenth century,” *International Journal of Performance Analysis in Sport*, vol. 15, no. 3, pp. 1160–1176, 2015.

[13] H. Chadwick, *Beadle’s dime base-ball player: a compendium of the game, comprising elementary instructions of this American game of ball, together with the Revised Rules and Regulations for 1860, rules for the formation of clubs, names of the officers and delegates to the general convention, c..* Johannsen Collection. Rare Books and Special Collections, Northern Illinois University.: Irwin P. Beadle Co., 141 William St., Cor. Fulton., 1860.

[14] J. T. Whittelsy, *Wright and Ditson’s Lawn Tennis Guide*. Boston, Mass.: Wright and Ditson’s, 1891.

[15] D. Whiteside and M. Reid, “Spatial characteristics of professional tennis serves with implications for serving aces: A machine learning approach,” *Journal of Sports Sciences*, vol. 35, no. 7, pp. 648–654, 2017.

[16] M. Hughes and I. M. Franks, “Notational analysis—a review of the literature,” *Notational Analysis of Sport*, pp. 71–116, 2004.

[17] Online Cambridge Dictionary, “Meaning of data science.” `https://dictionary.cambridge.org/dictionary/english/data-science`, 2022. Accessed: 2022-05-06.

[18] Data Science Association, “Data science code of professional conduct.” `https://datascienceassn.org/code-of-conduct.html`, 2022. Accessed: 2022-05-06.

[19] M. Hughes, T. Caudrelier, N. James, A. Redwood-Brown, I. Donnelly, A. Kirkbride, and C. Duschesne, “Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position,” *Journal of Human Sport and Exercise*, vol. 7, no. 2, pp. 402–412, 2012.

[20] R. Giulianotti and R. Robertson, “The globalization of football: a study in the glocalization of the ‘serious life’,” *The British journal of Sociology*, vol. 55, no. 4, pp. 545–568, 2004.

[21] C. Reep and B. Benjamin, “Skill and chance in association football,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 131, no. 4, pp. 581–585, 1968.

[22] B. Drust and M. Green, “Science and football: evaluating the influence of science on performance,” *Journal of Sports Sciences*, vol. 31, no. 13, pp. 1377–1382, 2013.

[23] V. C. Pantzalis and C. Tjortjis, “Sports analytics for football league table and player performance prediction,” in *11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–8, IEEE, 2020.

[24] T. Reilly and D. Gilbourne, “Science and football: a review of applied research in the football codes,” *Journal of Sports Sciences*, vol. 21, no. 9, pp. 693–705, 2003.

[25] R. Rein and D. Memmert, “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science,” *SpringerPlus*, vol. 5, no. 1, pp. 1–13, 2016.

[26] T. Stølen, K. Chamari, C. Castagna, and U. Wisløff, "Physiology of soccer," *Sports Medicine*, vol. 35, no. 6, pp. 501–536, 2005.

[27] C. Carling, T. Reilly, and A. M. Williams, *Performance assessment for field sports*. Routledge, 2008.

[28] B. Drust, G. Atkinson, and T. Reilly, "Future perspectives in the evaluation of the physiological demands of soccer," *Sports Medicine*, vol. 37, no. 9, pp. 783–805, 2007.

[29] R. Marcelino, J. Sampaio, and I. Mesquita, "Investigação centrada na análise do jogo: da modelação estática à modelação dinâmica," *Revista Portuguesa de Ciências do Desporto*, vol. 11, no. 1, pp. 481–499, 2011.

[30] C. Hook and M. Hughes, "Patterns of play leading to shots in Euro 2000," *Pass.com*, pp. 295–302, 2001.

[31] M. Hughes and I. Franks, "Analysis of passing sequences, shots and goals in soccer," *Journal of Sports Sciences*, vol. 23, no. 5, pp. 509–514, 2005.

[32] J. Stanhope, "An investigation into possession with respect to time, in the Soccer World Cup 1994," *Notational Analysis of Sport III*, pp. 155–162, 2001.

[33] E. Olsen, "An analysis of goal scoring strategies in the World Championship in Mexico, 1986," *Science and Football*, pp. 373–376, 1988.

[34] M. Hughes and S. Churchill, "Attacking profiles of successful and unsuccessful teams in Copa America 2001," in *Science and Football V: The Proceedings of the Fifth World Congress on Science and Football*, vol. 23, pp. 222–228, 2005.

[35] C. Lago-Peñas, J. Lago-Ballesteros, and E. Rey, "Differences in performance indicators between winning and losing teams in the UEFA Champions League," *Journal of Human Kinetics*, vol. 27, no. 1, pp. 135–146, 2011.

[36] H. Liu, M.-Á. Gomez, C. Lago-Peñas, and J. Sampaio, "Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup," *Journal of Sports Sciences*, vol. 33, no. 12, pp. 1205–1213, 2015.

[37] P. Jones, N. James, and S. D. Mellalieu, "Possession as a performance indicator in soccer," *International Journal of Performance Analysis in Sport*, vol. 4, no. 1, pp. 98–102, 2004.

[38] E. Rampinini, F. M. Impellizzeri, C. Castagna, A. J. Coutts, and U. Wisløff, "Technical performance during soccer matches of the Italian Serie A League: Effect of fatigue and competitive level," *Journal of Science and Medicine in Sport*, vol. 12, no. 1, pp. 227–233, 2009.

[39] V. Armatas, A. Giannakos, D. Skoufas, G. Zangelidis, S. Papadopoulou, and N. Fragkos, "Differences in offensive actions between top and last teams in Greek first soccer division: A retrospective study 1998-2008," 2009.

[40] G. Yang, A. S. Leicht, C. Lago, and M.-Á. Gómez, "Key team physical and technical performance indicators indicative of team quality in the soccer Chinese Super League," *Research in Sports Medicine*, vol. 26, no. 2, pp. 158–167, 2018.

[41] C. Wright, S. Atkins, R. Polman, B. Jones, and L. Sargeson, "Factors associated with goals and goal scoring opportunities in professional soccer," *International Journal of Performance Analysis in Sport*, vol. 11, no. 3, pp. 438–449, 2011.

[42] M. Acar, B. Yapicioglu, N. Arikan, S. Yalcin, N. Ates, and M. Ergun, "Analysis of goals scored in the 2006 World Cup," in *Science and Football VI*, pp. 261–268, Routledge, 2008.

[43] K. Harrop and A. Nevill, "Performance indicators that predict success in an English professional League One soccer team," *International Journal of Performance Analysis in Sport*, vol. 14, no. 3, pp. 907–920, 2014.

[44] C. S. Kite and A. Nevill, "The predictors and determinants of inter-seasonal success in a professional soccer team," *Journal of Human Kinetics*, vol. 58, no. 1, pp. 157–167, 2017.

[45] J. Bloomfield, R. Polman, R. Butterly, and P. O'Donoghue, "Analysis of age, stature, body mass, BMI and quality of elite soccer players from 4 European Leagues," *J Sports Med Phys Fitness*, vol. 45, no. 1, pp. 58–67, 2005.

[46] A. Dellal, K. Chamari, d. P. Wong, S. Ahmaidi, D. Keller, R. Barros, G. N. Bisciotti, and C. Carling, "Comparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga," *European Journal of Sport Science*, vol. 11, no. 1, pp. 51–59, 2011.

[47] Q. Yi, R. Groom, C. Dai, H. Liu, and M. Á. Gómez Ruano, "Differences in technical performance of players from 'the big five' European football leagues in the UEFA Champions League," *Frontiers in Psychology*, vol. 10, p. 2738, 2019.

[48] L. Crolley, D. Hand, and R. Jeutter, "Playing the identity card: Stereotypes in European football," *Soccer & Society*, vol. 1, no. 2, pp. 107–128, 2000.

[49] E. Rienzi, B. Drust, T. Reilly, J. E. x. L. Carter, and A. Martin, "Investigation of anthropometric and work-rate profiles of elite south american international soccer players," *Journal of Sports Medicine and Physical Fitness*, vol. 40, no. 2, p. 162, 2000.

[50] R. M. Sapp, E. E. Spangenburg, and J. M. Hagberg, "Trends in aggressive play and refereeing among the top five European Soccer Leagues," *Journal of Sports Sciences*, vol. 36, no. 12, pp. 1346–1354, 2018.

[51] J. Oberstone, "Comparing team performance of the English Premier League, Serie A, and La Liga for the 2008-2009 season," *Journal of Quantitative Analysis in Sports*, vol. 7, no. 1, 2011.

[52] C. Li and Y. Zhao, "Comparison of goal scoring patterns in "the big five" European Football Leagues," *Frontiers in Psychology*, vol. 11, pp. 1–7, 2021.

[53] C. Lago-Peñas, M. Gómez-Ruano, D. Megías-Navarro, and R. Pollard, "Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues," *International Journal of Performance Analysis in Sport*, vol. 16, no. 2, pp. 411–421, 2016.

[54] J. Gama, G. Dias, M. Couceiro, P. Passos, K. Davids, and J. Ribeiro, "An ecological dynamics rationale to explain home advantage in professional football," *International Journal of Modern Physics C*, vol. 27, no. 09, pp. 1–16, 2016.

[55] W. Tucker, D. S. Mellalieu, N. James, and B. J. Taylor, "Game location effects in professional soccer: A case study," *International Journal of Performance Analysis in Sport*, vol. 5, no. 2, pp. 23–35, 2005.

[56] C. Lago-Peñas, J. Lago-Ballesteros, A. Dellal, and M. Gómez, "Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league," *Journal of Sports Science & Medicine*, vol. 9, no. 2, p. 288, 2010.

[57] K. S. Courneya and A. V. Carron, "The home advantage in sport competitions: a literature review.," *Journal of Sport & Exercise Psychology*, vol. 14, no. 1, pp. 13–27, 1992.

[58] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, no. 2, pp. 265–280, 1997.

[59] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," *International Journal of Forecasting*, vol. 35, no. 2, pp. 741–755, 2019.

[60] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using bayesian nets and other machine learning techniques," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544–553, 2006.

[61] F. Owramipur, P. Eskandarian, and F. S. Mozneb, "Football result prediction with bayesian network in Spanish League - Barcelona team," *International Journal of Computer Theory and Engineering*, vol. 5, no. 5, p. 812, 2013.

[62] A. McCabe, "An artificially intelligent sports tipper," in *AI 2002: Advances in Artificial Intelligence*, pp. 718–718, Springer Berlin Heidelberg, 2002.

[63] A. McCabe and J. Trevathan, "Artificial intelligence in sports prediction," in *Fifth International Conference on Information Technology: New Generations*, pp. 1194–1197, IEEE, 2008.

[64] X. Tang, Z. Liu, T. Li, W. Wu, and Z. Wei, "The application of decision tree in the prediction of winning team," in *International Conference on Virtual Reality and Intelligent Systems (ICVRIS) 2018*, pp. 239–242, IEEE, 2018.

[65] D. Munđar and D. Šimić, "Croatian first football league: Prediction of teams' ranking in the championship," in *Proceedings of the ISCCRO-International Statistical Conference in Croatia ISSN*, vol. 2, pp. 15–23, 2016.

[66] P. Gorgi, S. J. Koopman, and R. Lit, “Estimation of final standings in football competitions with a premature ending: The case of covid-19,” *AStA Advances in Statistical Analysis*, pp. 1–18, 2021.

[67] K. Odachowski and J. Grekow, “Using bookmaker odds to predict the final result of football matches,” in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 196–205, Springer, 2012.

[68] J.-S. Geurts, “Master thesis - football players’ transfer price determination based on performance in the Big 5 European Leagues,” Master’s thesis, 2016.

[69] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM Process Model,” *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[70] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–39, Manchester, 2000.

[71] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: Create Space, 2009.

[72] W. McKinney *et al.*, “Data structures for statistical computing in Python,” in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, Austin, TX, 2010.

[73] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, “Using the Jupyter Notebook as a tool for open science: An empirical study,” in *ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2017*, pp. 1–2, IEEE, 2017.

[74] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, p. 357–362, 2020.

[75] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[76] U. Shafique and H. Qaiser, “A comparative study of data mining process models (KDD, CRISP-DM and SEMMA),” *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.

[77] J. E. Jackson, *A user’s guide to principal components*. John Wiley & Sons, 2005.

[78] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[79] I. MacQueen, “Some methods for classifiction and analysis of multivariate observations,” in *Proceedings 5th Berkeley Symposium on Mathematical Statistics Problems*, pp. 281–297, 1967.

[80] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society. Series C*, vol. 28, no. 1, pp. 100–108, 1979.

[81] T. Velmurugan and T. Santhanam, “Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points,” *Journal of Computer Science*, vol. 6, no. 3, p. 363, 2010.

[82] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in Python,” *the Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[84] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[85] D. Berrar, “Cross-validation.,” *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 542–545, 2019.

[86] A. M. El-Habil, “An application on multinomial logistic regression model,” *Pakistan Journal of Statistics and Operation Research*, pp. 271–291, 2012.

[87] E. Fix and J. L. Hodges, “Nonparametric discrimination: consistency properties,” *Randolph Field, Texas, Project*, pp. 21–49, 1951.

[88] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, “An improved k-nearest neighbor model for short-term traffic flow prediction,” *Procedia-Social and Behavioral Sciences*, vol. 96, pp. 653–662, 2013.

[89] K. Syaliman, E. Nababan, and O. Sitompul, “Improving the accuracy of k-nearest neighbor using local mean based and distance weight,” in *Journal of Physics: Conference Series: Journal of Physics: Conf. Series 978 (2018)*, vol. 978, pp. 1–6, IOP Publishing, 2018.

[90] S. Jiang, G. Pang, M. Wu, and L. Kuang, “An improved k-nearest-neighbor algorithm for text categorization,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[91] C. Wade, *Hands-On Gradient Boosting with XGBoost and Scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd, 2020.

[92] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[93] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[94] C. E. Metz, “Basic principles of ROC analysis,” in *Seminars in Nuclear Medicine*, vol. 8, pp. 283–298, Elsevier, 1978.

[95] O. El Yousfi, “Master thesis - design and implementation of a football data analysis application,” Master’s thesis, 2021.

[96] FBref, “xG explained.” `https://fbref.com/en/expected-goals-model-explained/`, 2000. Accessed: 2022-05-06.

[97] T. Reilly and A. Williams, *Science and Soccer*. Sport Science, Routledge, 2003.

[98] D. Harre, D. Harre, and J. Barsch, *Principles of sports training: Introduction to the theory and methods of training*. Ultimate Athlete Concepts, 2012.

[99] I. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, Springer, 2002.

[100] H. F. Kaiser, “The application of electronic computers to factor analysis,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 141–151, 1960.

[101] Y. Choi, J. Taylor, and R. Tibshirani, “Selecting the number of principal components: Estimation of the true rank of a noisy matrix,” *The Annals of Statistics*, pp. 2590–2617, 2017.

[102] F. A. Moura, L. E. B. Martins, and S. A. Cunha, “Analysis of football game-related statistics using multivariate techniques,” *Journal of Sports Sciences*, vol. 32, no. 20, pp. 1881–1887, 2014.

[103] S. B. Imandoust, M. Bolandraftar, *et al.*, “Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background,” *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605–610, 2013.

Appendix A

Variables Correlation Analysis results

In this Appendix, the results from the correlation analysis to the variables belonging to eight categories are presented. Described in more detail in Chapter 6, this process allowed to rule out some variables, to include (almost) exclusively those with added explanatory capacity in the final data set. The Pearson sample correlation coefficient determined the exclusion criteria, and the threshold value considered was 0.9.

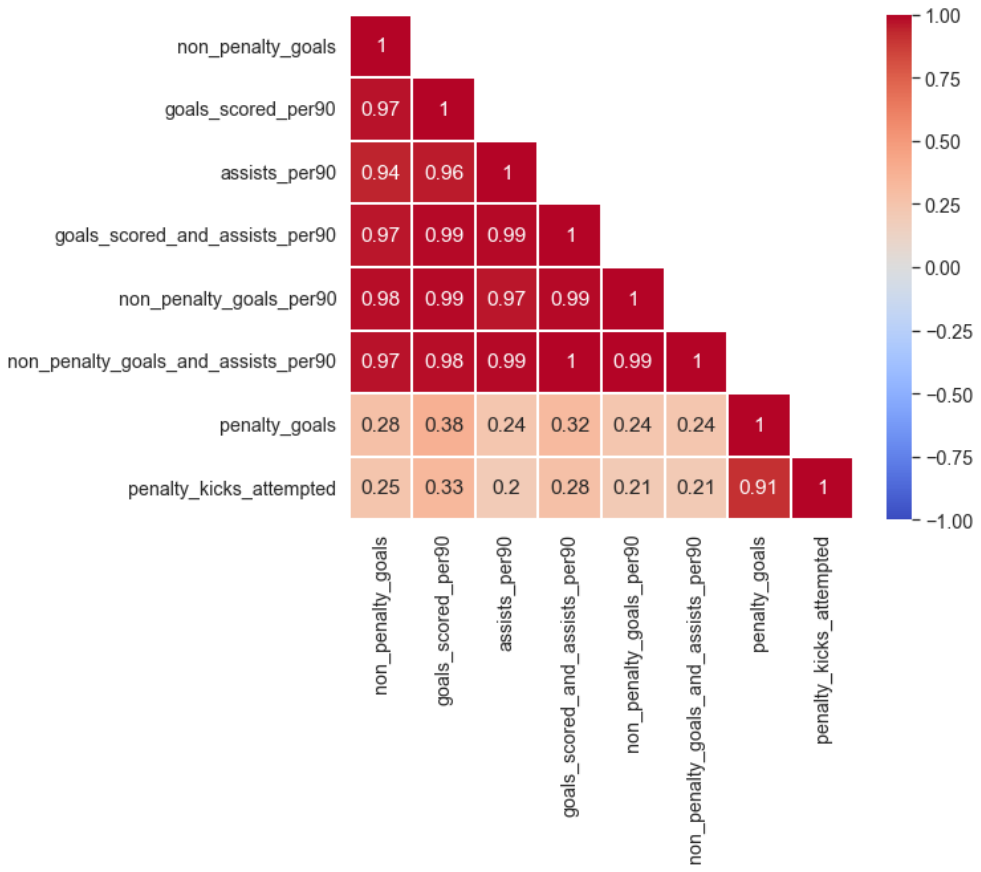

Figure A.1: Standard Stats correlation analysis (P $>$ 0.9)

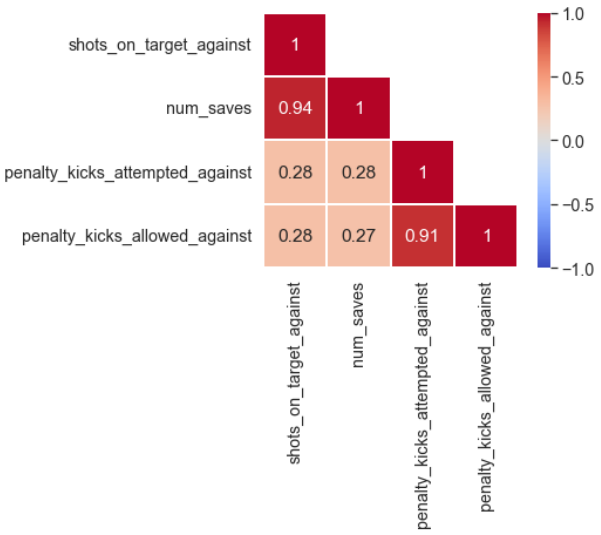

Figure A.2: Goalkeeping correlation analysis (P $>$ 0.9)

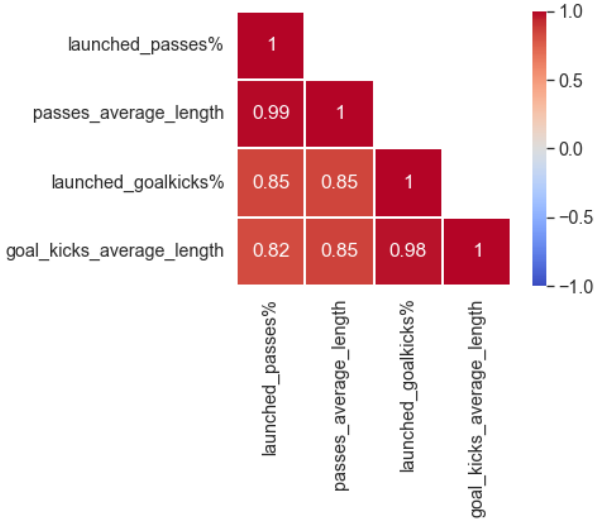

Figure A.3: Advanced Goalkeeping correlation analysis (P $>$ 0.9)

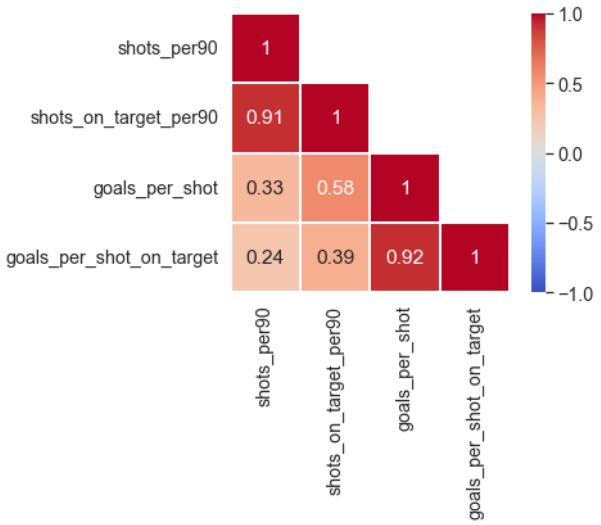

Figure A.4: Shooting correlation analysis (P $>$ 0.9)

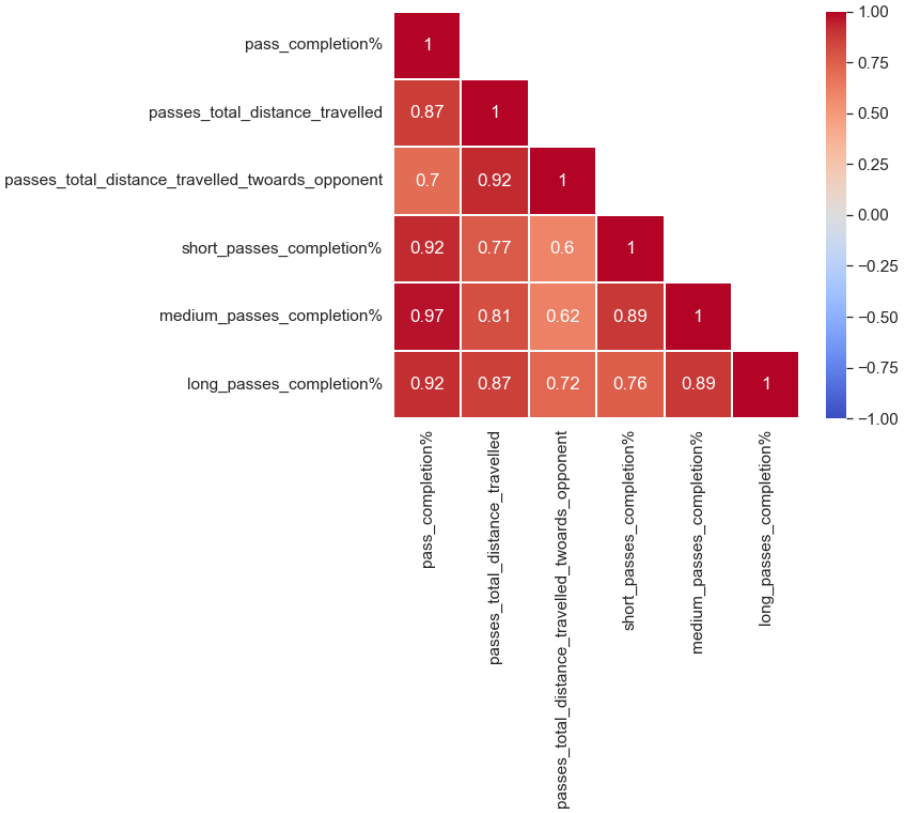

Figure A.5: Passing correlation analysis (P $>$ 0.9)

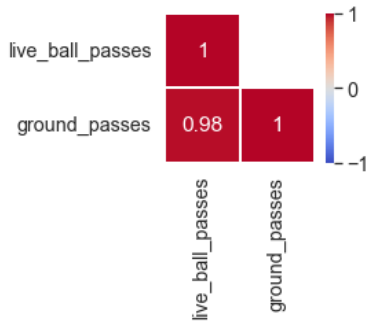

Figure A.6: Pass Types correlation analysis (P $>$ 0.9)

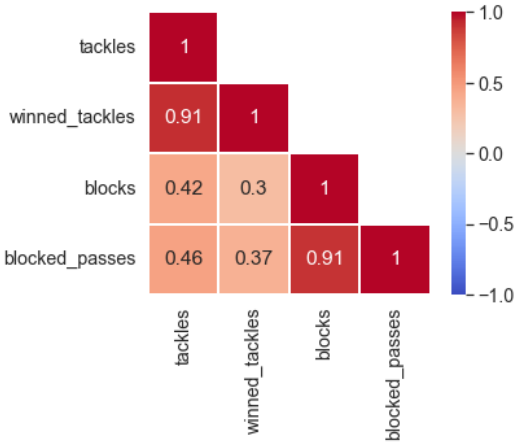

Figure A.7: Defensive Actions correlation analysis (P $>$ 0.9)

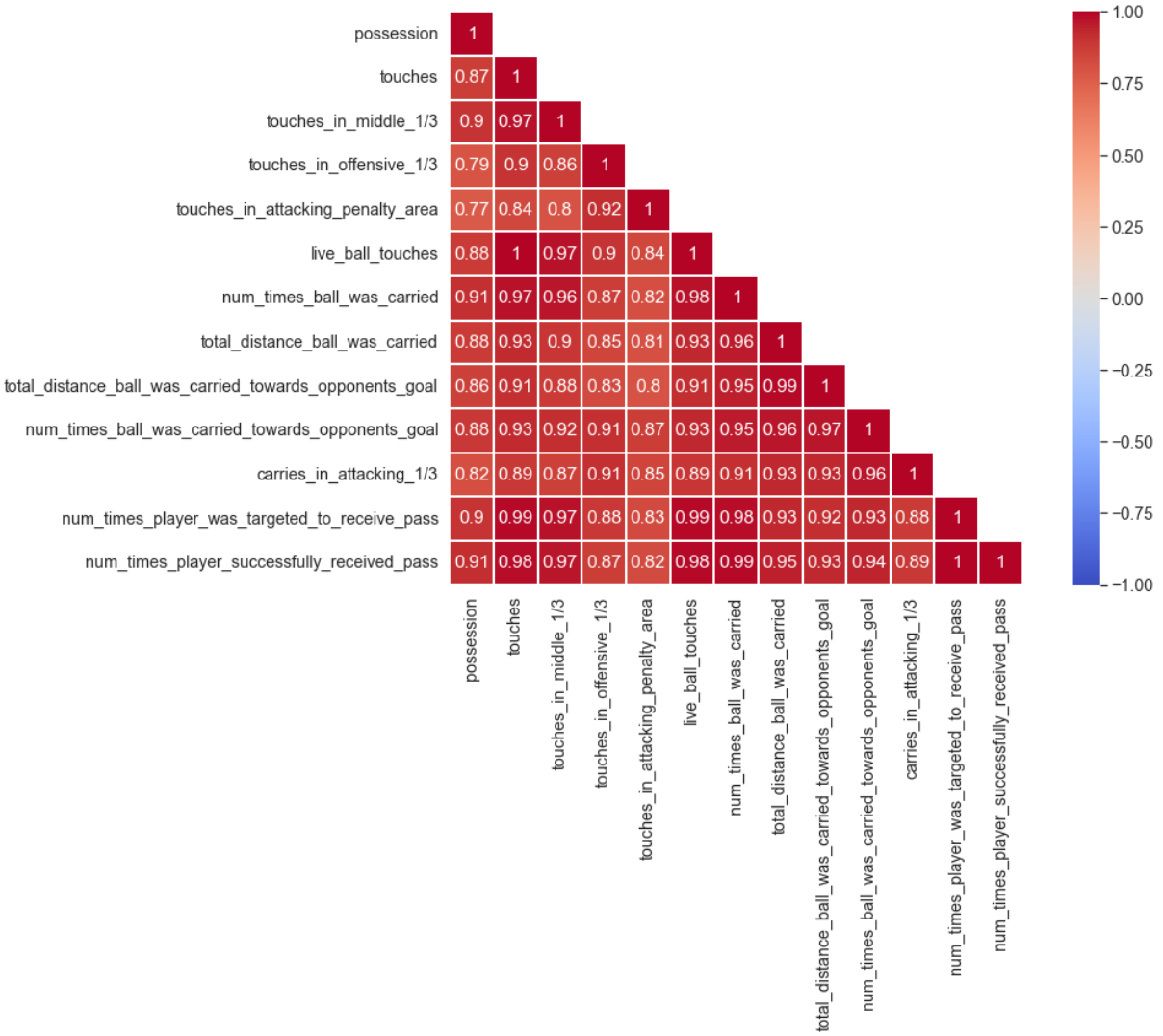

Figure A.8: Possession correlation analysis (P $>$ 0.9)

Appendix B

PCA additional results

In this Appendix, a section of the complete PCA scree plot is presented. Described in more detail in Chapter 7, this representation of the eigenvalues is useful when deciding upon the number of components to keep for further analysis.

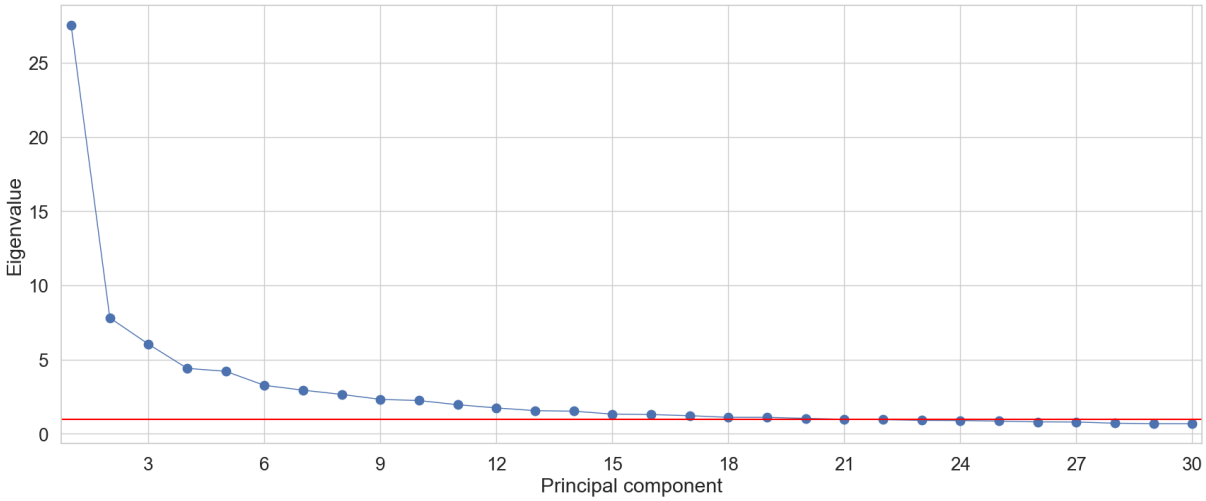

Figure B.1: Scree Plot Section (until PC 30)

Appendix C

Cluster Analysis additional results

In this Appendix, information about how each team was categorized using the K-Means clustering algorithm with two $(K = 2)$ and three $(K = 3)$ clusters performed on all input variables and observations (original data) and using data projected in the first 22 PCs (PCA Data).

Furthermore, information on the Silhouette Coefficient (SC) for each point classified into cluster groups for the K-Means clustering algorithm with two $(K = 2)$ and three $(K = 3)$ clusters performed on all input variables and observations (original data) and using data projected in the first 22 PCs (PCA Data).

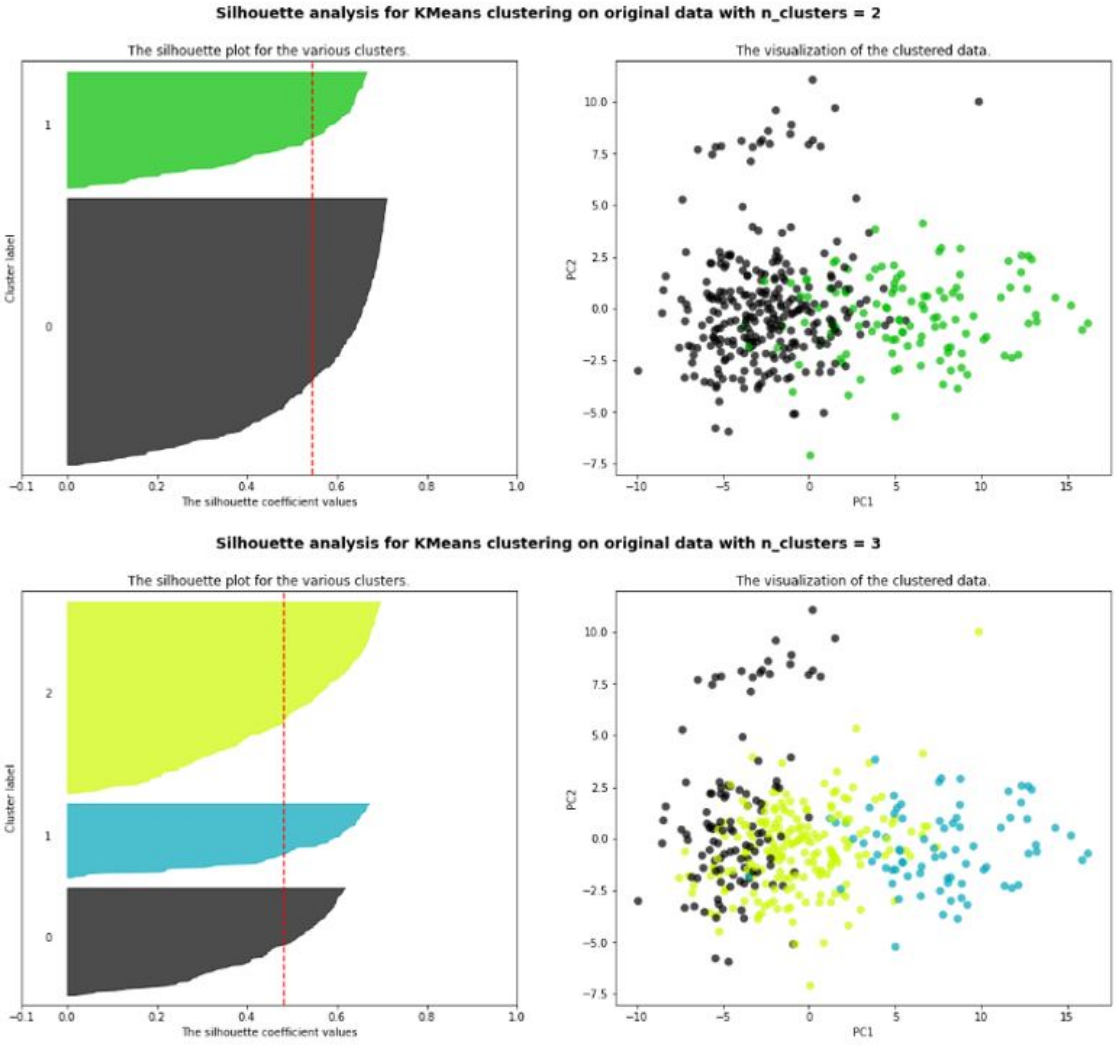

Figure C.1: K-Means clustering algorithm results using all the initial features and observations (original data) for $K = 2$ and $K = 3$.

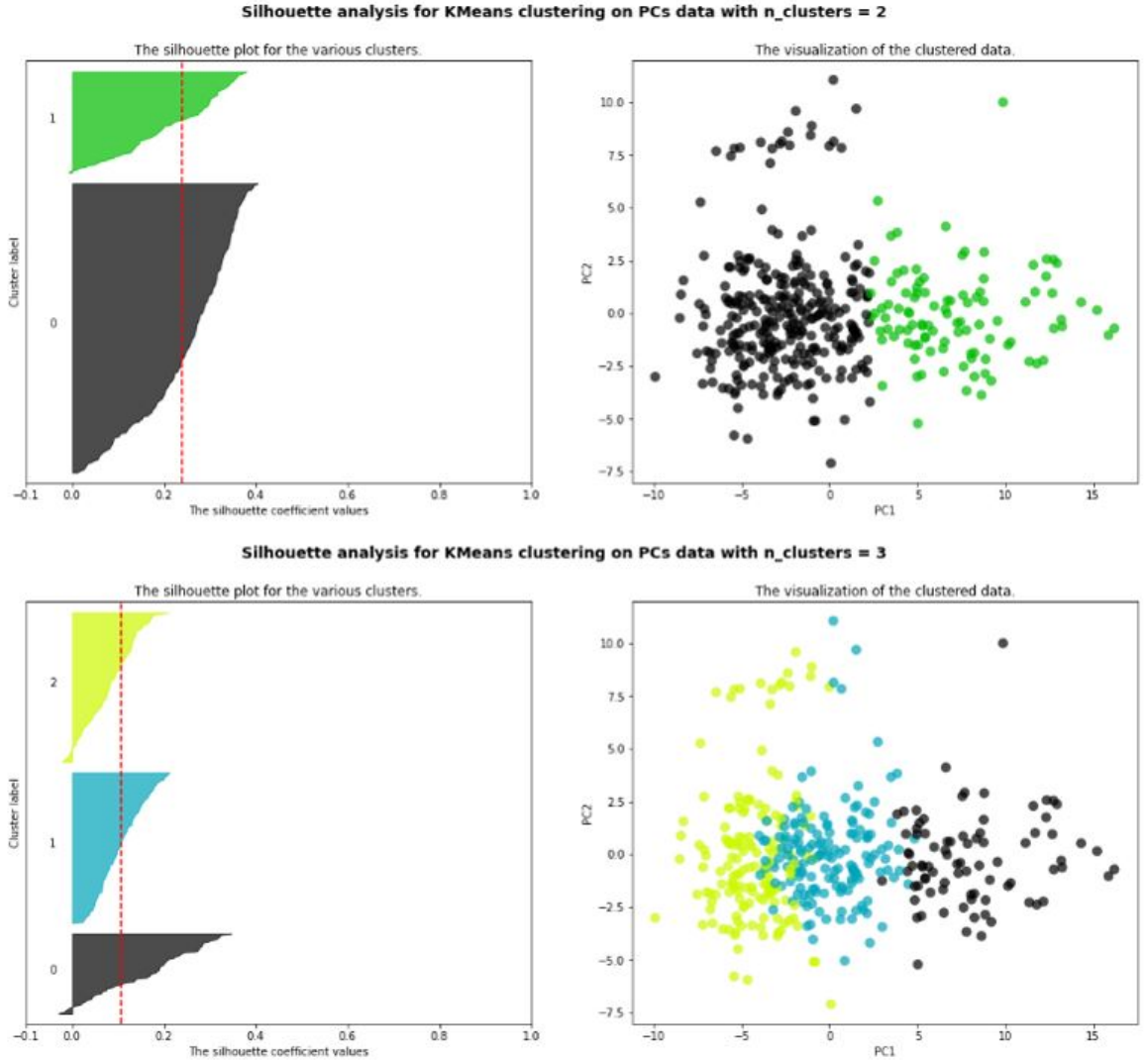

Figure C.2: K-Means clustering algorithm results using data projected in the first 22 PCs (PCs data) for $K = 2$ and $K = 3$.

Appendix D

Classifiers additional results

In this Appendix, the Receiver Operating Characteristic (ROC) curves are presented for AdaBoost, XGBoost and RF. Described in more detail in Chapter 7, they provide information regarding the performance of the classifiers used

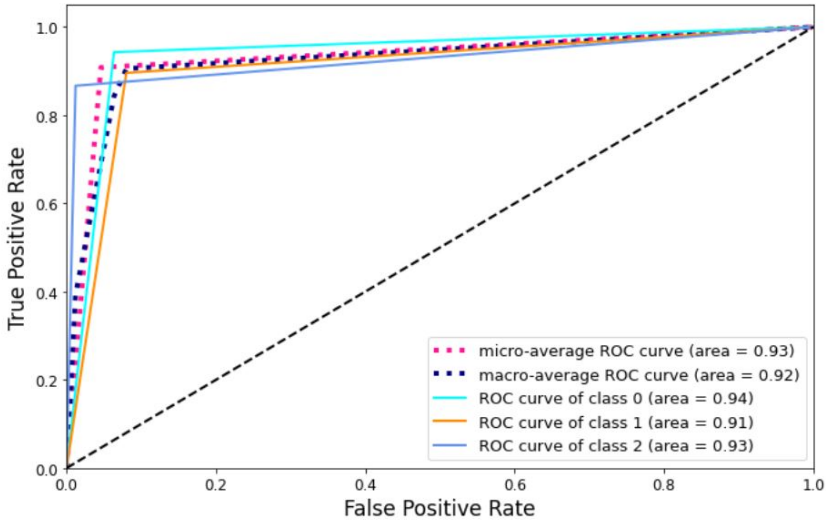

Figure D.1: Micro-average and macro-average ROC curves and ROC curves for the high-ranked (light blue), medium-ranked (yellow) and low-ranked (dark blue) categories of success for AdaBoost.

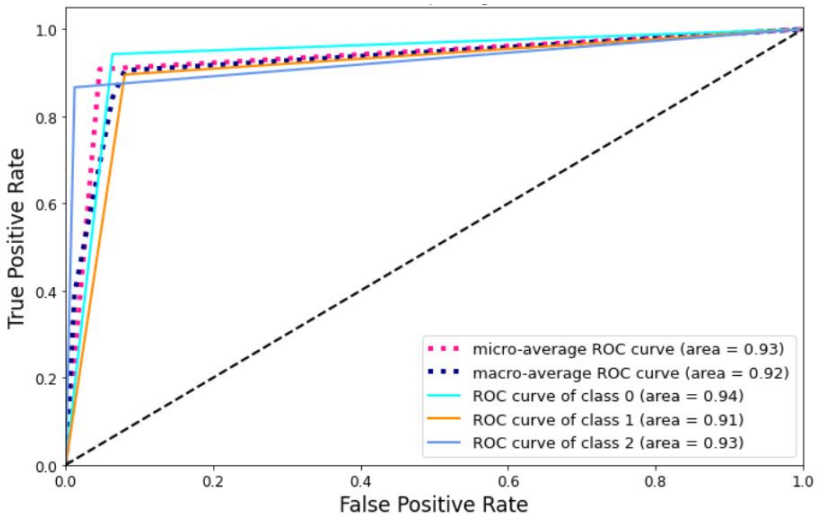

Figure D.2: Micro-average and macro-average ROC curves and ROC curves for the high-ranked (light blue), medium-ranked (yellow) and low-ranked (dark blue) categories of success for XGBoost.

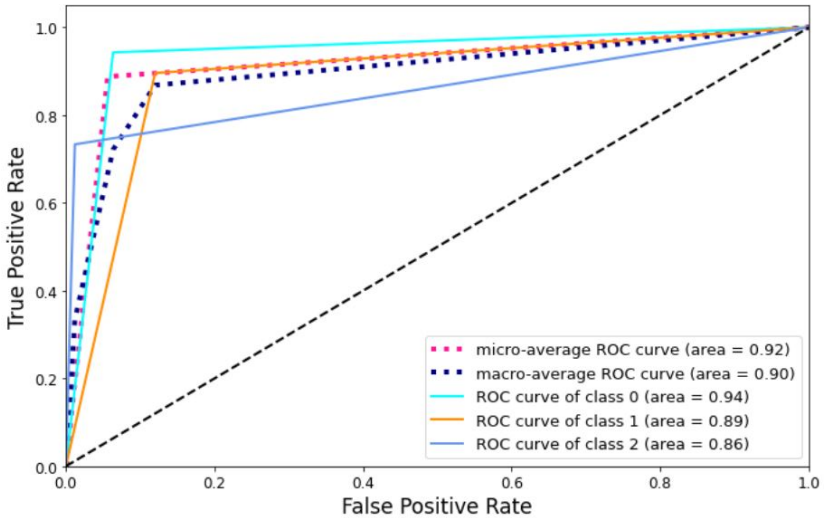

Figure D.3: Micro-average and macro-average ROC curves and ROC curves for the high-ranked (light blue), medium-ranked (yellow) and low-ranked (dark blue) categories of success for RF.