UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Machine Learning and Multivariate Statistical tools for Football Analytics

19 de junio de 2023

Autora: Pilar Malagón-Selma

Directores: Ana Debón Aucejo
Alberto J. Ferrer Riquelme

"Yo te alabo, Padre, Señor del Cielo y de la tierra, porque ocultaste estas cosas a los sabios y prudentes y las revelaste a los pequeños". Lc 10, 21

Aknoledgements
Agradecimientos

A Dios, por esta tesis que no es más mía que suya.

A mi marido, por ser el bastón que me mantiene erguida, la roca de mi descanso y el sol en mis días nublados.

A mi padre y mi madre, por su amor incondicional y por educarme en la cultura del trabajo y del esfuerzo.

A mi yaya, por su amor eterno.

A mis hermanos, amigos y comunidad, por la motivación, el ánimo y el sustento dado a lo largo de estos años.

To Prof. Maurizio, for accepting me as a visiting researcher in your group. Also, thanks to my colleagues Matteo, Riccardo and Mattia during these months in Brescia. Most of the work compiled here would have been impossible without your friendship, welcome and company.

A mis directores de tesis Dr. Alberto J. Ferrer y Dra. Ana Debón por su dedicación, colaboración y confianza durante este largo proceso de aprendizaje.

A Rafael Nadal, por ser una fuente de motivación e inspiración y por mostrar que el sacrificio y el trabajo duro tienen su recompensa.

Abstract

This doctoral thesis focuses on studying, implementing, and applying machine learning and multivariate statistics techniques in the emerging field of sports analytics, specifically in football. Commonly used procedures and new methods are applied to solve research questions in different areas of football analytics, both in the field of sports performance and in the economic field. The methodologies used in this thesis enrich the techniques used so far to obtain a global vision of the behaviour of football teams and are intended to help the decision-making process. In addition, the methodology was implemented using the free statistical software R and open data, which allows for reproducibility of the results.

This doctoral thesis aims to contribute to the understanding of the behaviour of machine learning and multivariate models for analytical sports prediction, comparing their predictive capacity and studying the variables that most influence the predictive results of these models. Thus, since football is a game of chance where luck plays an important role, this document proposes methodologies that help to study, understand, and model the objective part of this sport. This thesis is structured into five blocks, differentiating each according to the database used to achieve the proposed objectives.

The first block describes the most common study areas in football analytics and classifies them according to the available data. This part contains an exhaustive study of football analytics state of the art. Thus, part of the existing literature is compiled based on the objectives achieved, with a review of the

statistical methods applied. These methods are the pillars on which the new procedures proposed here are based.

The second block consists of two chapters that study the behaviour of teams concerning the ranking at the end of the season: top (qualifying for the Champions League or Europa League), middle, or bottom (relegating to a lower division). Several machine learning and multivariate statistical techniques are proposed to predict the teams' position at the season's end. Once the prediction has been made, the model with the best predictive accuracy is selected to study the game actions that most discriminate between positions. In addition, the advantages of our proposed techniques compared to the classical methods used so far are analysed. The database used for the analysis comprises quantitative variables that store cumulative information on the game actions performed by the teams throughout the 2018/2019 season.

The third block consists of a single chapter in which a web scraping code is developed to facilitate the retrieval of a new database with quantitative information on the game actions carried out over time in football matches. This block focuses on predicting match outcomes (win, draw, or loss) and proposing the combination of a machine learning technique, random forest, and Skellam regression model, a classical method commonly used to predict goal difference in football. Finally, the predictive accuracy of the classical methods used so far is compared with the proposed multivariate methods. The scraped database contains match-by-match statistics for the 2019/2020 and 2020/2021 seasons.

The fourth block also comprises a single chapter and pertains to the economic football area. This chapter applies a novel procedure to develop indicators that help predict transfer fees. Specifically, it is shown the importance of popularity when calculating the players' market value, so this chapter is devoted to propose a new methodology for collecting players' popularity information. The database of this block contains information similar to that of the second block but related to the players. In addition, this database has been completed with the proposed popularity indicators.

The fifth block reveals the most relevant aspects of this thesis for research and football analytics, including future lines of work.

Resumen

Esta tesis doctoral se centra en el estudio, implementación y aplicación de técnicas de aprendizaje automático y estadística multivariante en el emergente campo de la analítica deportiva, concretamente en el fútbol. Se aplican procedimientos comunmente utilizados y métodos nuevos para resolver cuestiones de investigación en diferentes áreas del análisis del fútbol, tanto en el ámbito del rendimiento deportivo como en el económico. Las metodologías empleadas en esta tesis enriquecen las técnicas utilizadas hasta el momento para obtener una visión global del comportamiento de los equipos de fútbol y pretenden ayudar al proceso de toma de decisiones. Además, la metodología se ha implementado utilizando el software estadístico libre R y datos abiertos, lo que permite la replicabilidad de los resultados.

Esta tesis doctoral pretende contribuir a la comprensión de los modelos de aprendizaje automático y estadística multivariante para la predicción analítica deportiva, comparando su capacidad predictiva y estudiando las variables que más influyen en los resultados predictivos de estos modelos. Así, siendo el fútbol un juego de azar donde la suerte juega un papel importante, se proponen metodologías que ayuden a estudiar, comprender y modelizar la parte objetiva de este deporte. Esta tesis se estructura en cinco bloques, diferenciando cada uno en función de la base de datos utilizada para alcanzar los objetivos propuestos.

El primer bloque describe las áreas de estudio más comunes en la analítica del fútbol y las clasifica en función de los datos utilizados. Esta parte contiene un estudio exhaustivo del estado del arte de la analítica del fútbol. Así, se

recopila parte de la literatura existente en función de los objetivos alcanzados, conjuntamente con una revisión de los métodos estadísticos aplicados. Estos modelos son los pilares sobre los que se sustentan los nuevos procedimientos aquí propuestos.

El segundo bloque consta de dos capítulos que estudian el comportamiento de los equipos que alcanzan la Liga de Campeones o la Europa League, descienden a segunda división o permanecen en mitad de la tabla. Se proponen varias técnicas de aprendizaje automático y estadística multivariante para predecir la posición de los equipos a final de temporada. Una vez realizada la predicción, se selecciona el modelo con mejor precisión predictiva para estudiar las acciones de juego que más discriminan entre posiciones. Además, se analizan las ventajas de las técnicas propuestas frente a los métodos clásicos utilizados hasta el momento. La base de datos utilizada para el análisis se compone de variables cuantitativas que almacenan información acumulada sobre las acciones de juego realizadas por los equipos a lo largo de la temporada 2018/2019.

El tercer bloque consta de un único capítulo en el que se desarrolla un código de *web scraping* para facilitar la recuperación de una nueva base de datos con información cuantitativa de las acciones de juego realizadas a lo largo del tiempo en los partidos de fútbol. Este bloque se centra en la predicción de los resultados de los partidos (victoria, empate o derrota) y propone la combinación de una técnica de aprendizaje automático, random forest, y la regresión Skellam, un método clásico utilizado habitualmente para predecir la diferencia de goles en el fútbol. Por último, se compara la precisión predictiva de los métodos clásicos utilizados hasta ahora con los métodos multivariantes propuestos. La base de datos contiene estadísticas partido a partido de las temporadas 2019/2020 y 2020/2021.

El cuarto bloque también comprende un único capítulo y pertenece al área económica del fútbol. En este capítulo se aplica un novedoso procedimiento para desarrollar indicadores que ayuden a predecir los precios de traspaso. En concreto, se muestra la importancia de la popularidad a la hora de calcular el valor de mercado de los jugadores, por lo que este capítulo propone una nueva metodología para la recogida de información sobre la popularidad de los jugadores. La base de datos de este bloque contiene información similar a la del segundo bloque pero relacionada con los jugadores. Además, esta base de datos se ha completado con los indicadores de popularidad propuestos.

En el quinto bloque se revelan los aspectos más relevantes de esta tesis para la investigación y la analítica en el fútbol, incluyendo futuras líneas de trabajo.

Resum

Aquesta tesi doctoral se centra en l'estudi, implementació i aplicació de tècniques d'aprenentatge automàtic i estadística multivariant en l'emergent camp de l'analítica esportiva, concretament en el futbol. S'apliquen procediments comunament utilitzats i mètodes nous per a resoldre qüestions d'investigació en diferents àrees de l'anàlisi del futbol, tant en l'àmbit del rendiment esportiu com en l'econòmic. Les metodologies emprades en aquesta tesi enriqueixen les tècniques utilitzades fins al moment per a obtindre una visió global del comportament dels equips de futbol i pretenen ajudar al procés de presa de decisions. A més, la metodologia s'ha implementat utilitzant el programari estadístic lliure R i dades obertes, la qual cosa permet la replicabilitat dels resultats.

Aquesta tesi doctoral pretén contribuir a la comprensió dels models d'aprenentatge automàtic i estadística multivariant per a la predicció analítica esportiva, comparant la seua capacitat predictiva i estudiant les variables que més influeixen en els resultats predictius d'aquests models. Així, sent el futbol un joc d'atzar on la sort juga un paper important, es proposen metodologies que ajuden a estudiar, comprendre i modelitzar la part objectiva d'aquest esport. Aquesta tesi s'estructura en cinc blocs, diferenciant cadascun en funció de la base de dades utilitzada per a aconseguir els objectius proposats.

El primer bloc descriu les àrees d'estudi més comuns en l'analítica del futbol i les classifica en funció de les dades utilitzades. Aquesta part conté un estudi exhaustiu de l'estat de l'art de l'analítica del futbol. Així, es recopila part de la literatura existent en funció dels objectius aconseguits, conjuntament amb una

revisió dels mètodes estadístics aplicats. Aquests models són els pilars sobre els quals se sustenten els nous procediments ací proposats.

El segon bloc consta de dos capítols que estudien el comportament dels equips que aconsegueixen la Lliga de Campions o l'Europa League, descendeixen a segona divisió o romanen a la meitat de la taula. Es proposen diverses tècniques d'aprenentatge automàtic i estadística multivariant per a predir la posició dels equips a final de temporada. Una vegada realitzada la predicció, se selecciona el model amb millor precisió predictiva per a estudiar les accions de joc que més discriminen entre posicions. A més, s'analitzen els avantatges de les tècniques proposades enfront dels mètodes clàssics utilitzats fins al moment. La base de dades utilitzada per a l'anàlisi es compon de variables quantitatives que emmagatzemen informació acumulada sobre les accions de joc realitzades pels equips al llarg de la temporada 2018/2019.

El tercer bloc consta d'un únic capítol en el qual es desenvolupa un codi de *web scraping* per a facilitar la recuperació d'una nova base de dades amb informació quantitativa de les accions de joc realitzades al llarg del temps en els partits de futbol. Aquest bloc se centra en la predicció dels resultats dels partits (victòria, empat o derrota) i proposa la combinació d'una tècnica d'aprenentatge automàtic, random forest, i la regressió Skellam, un mètode clàssic utilitzat habitualment per a predir la diferència de gols en el futbol. Finalment, es compara la precisió predictiva dels mètodes clàssics utilitzats fins ara amb els mètodes multivariants proposats. La base de dades conté estadístiques partit a partit de les temporades 2019/2020 i 2020/2021.

El quart bloc també comprén un únic capítol i pertany a l'àrea econòmica del futbol. En aquest capítol s'aplica un nou procediment per a desenvolupar indicadors que ajuden a predir els preus de traspàs. En concret, es mostra la importància de la popularitat a l'hora de calcular el valor de mercat dels jugadors, per la qual cosa aquest capítol proposa una nova metodologia per a la recollida d'informació sobre la popularitat dels jugadors. La base de dades d'aquest bloc conté informació similar a la del segon bloc pèro relacionada amb els jugadors. A més , aquesta base de dades s'ha completat amb els indicadors de popularitat proposats.

En el cinqué bloc es revelen els aspectes més rellevants d'aquesta tesi per a la investigació i l'analítica en el futbol, incloent-hi futures línies de treball.

Contents

List of Figures

List of Tables

Chapter 1

Justification, Objectives and Contributions

1.1 Sports Analytics

Sports analytics originated in 1858 when a sportswriter named Henry Chadwick developed the first statistician to quantitatively measure baseball players' and teams' performance: the box score[1]. Years later, in 1861, Chadwick wrote the first documented guide to sports analysis, "Beadle's Dime Base-Ball Player" (Chadwick 1860). This book advocated the need to analyse players' performance on the field to estimate players' abilities. However, it was not until the beginning of the 21st century that sports analytics came to the attention of analysts and companies. In 2003, Lewis (2004) wrote the true story of Billy Beane, the manager of the Oakland Athletics baseball team, who won the American League West title by building a team with limited financial resources. This book tells how the key to success was using a public database to find players whose market value was undervalued and whose skills were complementary to the rest of the team.

[1]The box score, commonly written in tabular form, summarises the players' game actions during a match.

In this context, researchers and analysts saw the need to implement information systems and mathematical algorithms to evaluate performance data. Thus, sports analytics emerged as a process of searching, interpreting and processing data that provide competitive advantages (Link 2018). According to Link (2018), the competitive advantages are related to both the sports and economic areas. Thus, the former allows for improving the performance and efficiency of the teams, while the economic part is related to income generation.

With the emergence of Industry 4.0[2], the possibilities of both areas have been greatly increased. Then, the necessity of having appropriate information systems to manage and analyse the high volume of data has driven the cohesion between data analysis and sports. This fact establishes sports analysis as an emerging field that provides researchers with the opportunity to study the intricacies of sports.

The areas of analysis, as well as the purpose of the studies, differ according to the sport and the available data. In the case of football, teams receive large amounts of data, either through external providers or as a result of player-related information they collect themselves (blood tests or global positioning systems (GPS) devices). Thus, in most of the cases, the nature of the data available determine the problem of study. Figure 1.1 shows the categorisation of sports analytics studies as a function of two levels of analysis: a first-order level, according to the nature of the data available (eventing, tracking, GPS and injuries data); and a second-order level, according to the main objective of the study.

[2] A new revolution that combines advanced production and operations techniques with intelligent technologies that will be integrated into organisations, people and assets (Cotteleer and Sniderman 2019).

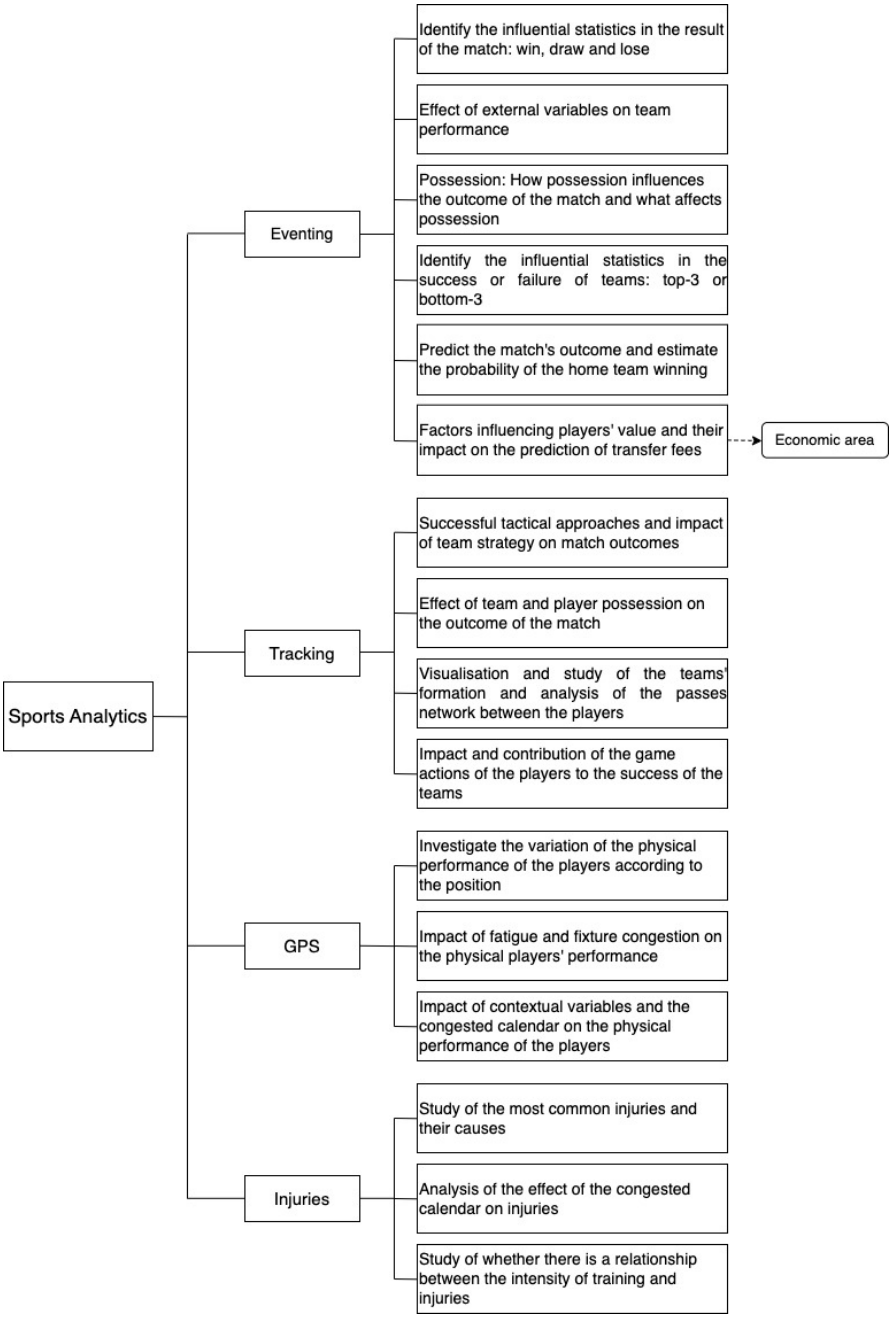


Figure 1.1: Categorisation of sports analytics studies as a function of two levels of analysis: the nature of the data available and the main objective of the studies.

1.1.1 Eventing

The eventing data are the measurable game actions, which are related to the ball, performed during a football match (e.g. number of goals, number of assists, number of crosses). This data can be obtained through sports analytics providers (Opta, Wyscout or STATS), football teams and open websites that store football statistics. The specific studies, corresponding objectives, samples, and methods carried out using eventing data are listed in Table 1.1.

According to Table 1.1, there is a wide range of possible analyses using eventing data: identifying the optimal game strategies for winning a match (Peñas et al. 2010; Lago-Peñas, Lago-Ballesteros, and Rey 2011; Castellano, Casamichana, and Lago 2012; Carpita et al. 2015; Liu et al. 2015a) and differentiating between successful (top-3) and unsuccessful (bottom-3) teams at the end of the season (Oberstone 2009; Schauberger, Groll, and Tutz 2017; Souza et al. 2019) are common objectives of many of the papers reviewed. Other studies are the analysis of the effect of possession on the match outcome (Lago and Martín 2007; Lago 2009; Collet 2013) as well as the study of the impact of external variables on the teams performance (Taylor et al. 2008; Lago-Peñas 2010). Regarding the economic area, the optimization of economic resources has been studied in depth (Müller, Simons, and Weinmann 2017; Hofmann et al. 2019; Singh and Lamba 2019).

Regarding the usual statistical methods, most researchers used similar methods to conduct the studies (linear and logistic regression, discriminant analysis and analysis of variance (ANOVA)), except for Carpita et al. (2015), who applied machine learning techniques (Table 1.1). In addition, Table 1.1 shows that in the case of studies related to predicting match outcomes (Karlis and Ntzoufras 2009; Carpita, Ciavolino, and Pasca 2019; Carpita and Golia 2021; Carpita, Ciavolino, and Pasca 2021) or players' valuation (Müller, Simons, and Weinmann 2017; Hofmann et al. 2019; Singh and Lamba 2019), researchers employed robust machine learning and multivariate methods to achieve their objectives.

Study	Objective	Sample	Methods
Variables related to the outcome of the match: win, draw or loss			
Peñas et al. (2010)	Identify match-related statistics that allow to discriminate between winning, draw and losing teams	380 matches of LaLiga (2008-2009)	Discriminant Analysis
Lago-Peñas, Lago-Ballesteros, and Rey (2011)	Identify match-related statistics that allow discriminating winning teams from drawing and losing teams	288 matches of UEFA Champions League (2007-2008 to 2009-2010)	ANOVA and Discriminant Analysis
Castellano, Casamichana, and Lago (2012)	Identify match-related statistics that best discriminate between winning, drawing and losing teams	177 matches of the World Cup (2002, 2006 and 2010)	ANOVA and Discriminant analysis
Carpita et al. (2015)	Discover the factors that lead to winning the match	1520 matches of the Serie A (2008-2009 to 2011-2012)	Random Forest, Neural Network, K-Nearest Neighbor, Naïve Bayes, and Multinomial Logistic Regression
Liu et al. (2015a)	Identify relationships between match-related statistics and the match outcome (win, loss and draw)	48 matches of Brazil World Cup (2014)	K-means cluster and cumulative logistic regression
Influence of external variables			
Taylor et al. (2008)	Examine the effects of match location, quality of opposition, and match status on performance statistics of a specific football team	40 matches of the same team which compete in the Premier League (2002-2003 and 2003-2004)	Log-linear and Logit modelling procedures
Lago-Peñas (2010)	Examine the effects of external variables on ball possession strategies	380 matches of LaLiga (2008-2009)	Pearson coefficients of variation (CV), Linear regression and Ramsey Regression Equation Specification Error Test (RESET)
Possession			
Lago and Martín (2007)	Influence of the match events on ball possession	170 matches of LaLiga (2003-2004)	Two linear regression models: additive and interactive model
Lago (2009)	Influence of the match location, quality of opposition, and match status on ball possession.	27 matches of LaLiga (2005-2006)	Linear regression and Ramsey Regression Equation Specification Error Test (RESET)
Collet (2013)	Influence on the possession of the match outcome	5484 matches in the Premier League, Serie A, Ligue 1, Bundesliga, LaLiga (2007-2008 to 2009-2010) and the Europa League (2009-2010). 111 matches of the African Cup of Nations (2010), AFC Asian Cup (2011), European Championships (2008), the FIFA Confederations Cup (2009). 177 matches of the World Cup (2002, 2006 and 2010)	Logistic regression
Variables related to the teams' success			
Oberstone (2009)	Identify game actions that determine the teams' success at the end of a season	20 teams of the Premier League (2007-2008)	Multiple linear regression
Schauberger, Groll, and Tutz (2017)	Identify the game actions that are connected to the success or failure of football teams	18 teams of the Bundesliga (2015-2016)	Bradley-Terry model (standard method for paired comparation)
Souza et al. (2019)	Analyse game actions to differentiate between the top-3 (Champions League positions) and the bottom-3 (relegated positions) teams ranked in LaLiga	48 teams 8 of LaLiga (2010-2011 to 2017-2018)	Student's t-test
Prediction of the match result			
Karlis and Ntzoufras (2009)	Predict the outcome of the matches	380 matches of the Premier League (2006-2007)	Skellam regression model
Carpita, Ciavolino, and Pasca (2019)	Use match-related statistics to estimate the win probability of the home teams	Data related to players, teams and matches of 10 European leagues (2009-2010 to 2015-2016)	Binomial logistic regression, Random Forest, Neuronal Network, K-Nearest Neighbor, and Naïve Bayes
Carpita and Golia (2021)	Use players-related statistics to predict the home team's win	28,000 players and about 21,000 matches of 9 European leagues (2008-2009 to 2015-2016)	Bayesian Network and Naïve Bayes
Carpita, Ciavolino, and Pasca (2021)	Use Clustering Latent Variables indicators to improve the match outcomes prediction	Data related to players, teams and matches of 10 European leagues (2009-2010 to 2015-2016)	Clustering Latent Vatiables and Skellam regression model
Players' valuation			
Müller, Simons, and Weinmann (2017)	Estimate players' market values	4,217 players who competed in the Premier League, Serie A, Ligue 1, Bundesliga, and LaLiga (2009-2010 to 2014-2015)	Multilevel regression method
Hofmann et al. (2019)	Impact of measures of player popularity on the market value	316 football players of the Bundesliga (2014-2015)	PLS-SEM model
Singh and Lamba (2019)	Find the factors that affect the players' market value and use them to predict their market value	Players who participated in UEFA European Football Championship	Decision Tree, Random Forest, Gradient Boost, Linear Regression, and Ridge Regression

Table 1.1: Empirical studies using eventing data

1.1.2 Tracking

Tracking data collects information about the movement of players (with and without the ball) on the pitch, i.e. spatio-temporal data that captures the position of players in every second of the match. Unlike eventing data, tracking data are more difficult to find as, in most cases, they are not available on public websites and can only be obtained from data companies or football teams. A summary of studies conducted with tracking data corresponding objectives, samples, and methods are shown in Table 1.2.

Study	Objective	Sample	Methods
Team tactics and strategy			
James, Mellalieu, and Hollely (2002)	Variation of the teams' strategies according to the nature of the competition (domestic and European competitions)	190 domestic games and 57 games in European competition of an Premier League team (2001-2002)	Chi-squared test
Tenga et al. (2010)	Examine the effects of counterattack or elaborate attack tactics on the possession outcome	163 matches of Norwegian Football League (2004-2005)	Multiple logistic regression, univariate analyses and multivariate analyses
Lucey et al. (2013)	Study the teams' aggressive or passive strategies depending on the venue (Home vs Away)	380 matches of the Premier League (2010-2011).	Calculate D-dimensional spatiotemporal feature vector x and k-Nearest Neighbor approach.
Kempe et al. (2014)	Evaluate the successful tactical approach: possession vs direct play	676 matches of the Bundesliga (2009-2010 and 2010-2011) and the South Africa (2010) World Cup.	ANOVA
Wang et al. (2015)	Application of an unsupervised approach to automatically differentiating tactical patterns	241 matches of LaLiga (2013-2014).	Team Tactic Topic Model (T3M), Gibbs sampling, and Latent Dirichlet Allocation (LDA)
González-Rodenas et al. (2020)	Investigate the combined effects of tactical and contextual indicators on achieving offensive and scoring game actions	380 matches of the Premier League (2017-2018)	Binary logistic regressions and Multilevel logistic regression
Possesion			
Jones, James, and Mellalieu (2004)	Analyse if possession is synonymous with success (win, draw or loss the match)	24 matches of the Premier League (2001-2002)	Mann-Whitney U non-parametric test
Link and Hoernig (2017)	Detection of the type of ball possession according to players' position	60 matches of the Bundesliga (2012-2013)	Rauch-Tung-Striebel (RTS) method and Bayesian Network
Positional and movement study			
Bialkowski et al. (2014)	Present an approach for the detection and visualization of the teams' formation.	380 matches (league and season omitted)	Expectation maximization algorithm and earth mover's distance
Goncalves et al. (2017)	Explore the relationship between passing networks and positioning variables to the match outcome.	44 players under-15 and under-17 age group	Closeness and betweenness centrality, two-step cluster analysis and Schwartz's Bayesian criterion
Memmert, Lemmink, and Sampaio (2017)	Overview of the current state of the application of data analysis techniques on position data	22 players (Bayern Müinch vs. FC Barcelona)	Inter-player coordination, inter-team and inter-line coordination artificial, and artificial neuronal network
Quantifying the game's actions			
Duch, Waitzman, and Amaral (2010)	Develop a network approach to quantify the contributions of individual players and the teams' performance	Players who participated in European Cup (2008)	Social network analysis and Monte Carlo methods
Cintia et al. (2015)	Analysis of the teams' passing behaviour and their relationship with the success	1446 matches of the Premier League, Serie A, Bundesliga, and LaLiga (2013-2014).	Weighted network, K-Nearest Neighbor, Random Forest, Logistic regression, Decision tree, Naïve Bayes, and Support Vector Machine
Link, Lang, and Seidenschwarz (2016)	Use of the goals probability to quantify the teams' attacking performance	64 matches of the Bundesliga (2014-2015)	ANOVA
Power et al. (2017)	How the value of a pass is estimated	726 matches of the Premier League (2014-2015 to 2015-2016)	Logistic Regression, supervised learning approach, and formation clustering method
Decroos et al. (2019)	Evaluate players' game actions based on their impact on the match result.	11565 matches of Spanish, English,German, Italian, French, Dutch, and Belgian Football Leagues (2012-2013 to 2017-2018)	SPADL (Soccer Player Action Description Language), CatBoost, Logistic regression, Random forest and XGBoost

Table 1.2: Empirical studies using tracking data

Currently, the technology available to clubs and data companies (e.g.GPS and heat maps) allows them to track player movements. Table 1.2 shows an excerpt of the research studies conducted to determine the effect of game strategies (James, Mellalieu, and Hollely 2002; Tenga et al. 2010; Lucey et al. 2013; Kempe et al. 2014; Wang et al. 2015; González-Rodenas et al. 2020), formation (Bialkowski et al. 2014; Goncalves et al. 2017; Memmert, Lemmink, and Sampaio 2017), and possession (Jones, James, and Mellalieu 2004; Link and

Hoernig 2017) on match results. Another emerging area of study is the quantification of game actions, i.e., how much affect the players' performance on the match result and which specific game actions have a greater impact on it (Duch, Waitzman, and Amaral 2010; Cintia et al. 2015; Link, Lang, and Seidenschwarz 2016; Power et al. 2017; Decroos et al. 2019).

In the case of tactical behaviour, multivariate logistic regression has been used to study the combined and interactive effects of different tactical variables (Tenga et al. 2010; González-Rodenas et al. 2020). Similarly, classical statistics techniques have been used to evaluate teams' tactical behaviour and to determine successful strategies (James, Mellalieu, and Hollely 2002; Jones, James, and Mellalieu 2004; Kempe et al. 2014; Link, Lang, and Seidenschwarz 2016). In the case of machine learning techniques, they have been primarily applied in the study of the game actions on the match results (Duch, Waitzman, and Amaral 2010; Cintia et al. 2015; Power et al. 2017; Decroos et al. 2019). In addition, specific statistical techniques of spatio-temporal data analysis have been applied to study the formation of teams and the movement of players (Bialkowski et al. 2014; Goncalves et al. 2017; Memmert, Lemmink, and Sampaio 2017).

1.1.3 Global positioning systems (GPS)

Football is a sport characterised by its high physical demands since, on average, football players cover distances between 10 and 13 km per match (Stølen et al. 2005). These distances are covered at different physical intensities: low (15 km/h), middle (20 km/h) and high (25 km/h). Additionally, according to Dolci et al. (2020), a player can perform between 600 and 650 accelerations during a match. Therefore, throughout this thesis, we will refer to the physical activity information of football players collected throughout matches as global positioning system (GPS) data (e.g. total distance covered, high-intensity running, jumping and jogging). These data, as well as tracking data, are managed by multi-camera computerized tracking systems: Prozone and Amisco are pioneering companies in this field. Mainly, researchers collaborating with football teams obtain players' physical activity information. Table 1.3 summarises the studies carried out with GPS data, with their corresponding objectives, samples and methods.

Study	Objective	Sample	Methods
Players' performance by position			
Di Salvo et al. (2007)	Examine the exercise patterns according to the position	300 players of LaLiga (2002-2003) and Championship League (2003-2004)	ANOVA
Bradley et al. (2010)	Determine the high-intensity activity patterns at different playing positions, and compare game fatigue in elite domestic and international football matches	100 domestic and 10 international players	Tukey's post hoc tests
Gregson et al. (2010)	Determine the between-match variability of high-speed running activities and the influence of playing position on this variability	485 players of the Premier League (2003-2004 to 2005-2006)	One-factor general linear model and t-test
Di Salvo et al. (2010)	Analyse the sprinting activities of different playing positions during European Champions League and UEFA Cup competitions	717 players of European Champions League and UEFA Cup (2002-2003 to 2005-2006)	Kruskal-Wallis test and Mann Whitney U-tests
Carling, Le Gall, and Dupont (2012)	Investigate the high-intensity activity patterns and the demands specific to the positional role in professional football match-play	20 players of the same team which compete in the Ligue 1 (2007-2008 to 2010-2011)	ANOVA
Di Salvo et al. (2013)	Compare the physical performances between Premier League and EFL Championship and analyse differences between positions and leagues	13991 players of the Premier League and 12458 of the Championship League (2006-2007 to 2009-2010)	ANOVA
Bradley et al. (2013)	Compare the match performance and physical capacity of players in the top three competitive standards by position	949 players of the Premier League, 867 of the Championship League, and 867 players of the Ligue 1	ANOVA
Martín-García et al. (2018)	Determine the most physically demanding position and moments of play in football players	23 players of LaLiga SmartBank (2015-2016)	Bonferroni test and Dunnett's T3 test
Altmann et al. (2021)	Examine the importance of the position and player in the physical match performance	25 players of the Bundesliga (2019-2020)	ANOVA and t-test
Fatigue			
Mohr, Krustrup, and Bangsbo (2003)	Assess physical fitness, match performance and development of fatigue of professional football players	18 players of the Serie A and Championship League	Student's paired t-test and Student's unpaired t-test
Rampinini et al. (2009)	Examine the technical and physical performance changes between the first and second half	186 players of the SerieA (2004-2005)	ANOVA, Student's paired t-test, and Student's unpaired t-test
Bradley and Noakes (2013)	Determine if football players fatigue or modulate high-intensity running and study factors that impact high-intensity running	1140 matches of the Premier League (2019-2020)	ANOVA, Bonferroni-corrected dependent and independent t-test
Smith et al. (2017)	Examine the impact of mental fatigue on players' speed and specific skills	14 male football players who compete in Belgian leagues	ANOVA and MANOVA
Coutinho et al. (2017)	Examine the effects of mental fatigue on players' physical and tactical performances	20 amateur youth football players	ANOVA, Mauchly's test, and Bonferroni post-hoc
Jones et al. (2019)	Investigate the effect of fixture congestion on the players' physical performance	515 matches of the Premier League (2015-2016 to 2016-2017)	Linear mixed model
Effect of contextual variables and the congested calendar			
Rampinini et al. (2007)	Examine the influence of the level opponent team, seasonal variations and the first-half activity on players' performance	20 players from the same team which compete in the Champions League	ANOVA and Bonferroni test
Lago et al. (2010)	Examine the influence of match location, quality of opposition, and match status on distance covered	27 matches from the same team which compete in LaLiga (2005-2006)	Multiple regression
Dellal et al. (2015)	Investigate the influence of playing multiple matches in a short period on physical activity, technical performance and injury rates	16 players who competed in the Ligue 1, Coupe de France and the Champions League (2011-2012)	ANOVA, Tukey's post hoc tests, Fisher's exact test, and Mann Whitney U-tests
García-Unanue et al. (2018)	Investigate the influence of match location, match period and opponent level on players' physical performance	14 matches of LaLiga SmartBank (2016-2017)	ANOVA and Bonferroni test

Table 1.3: Empirical studies using global positioning systems (GPS) data

As mentioned above, football is characterised by its high physical demands. However, the physical performance requirements of matches may vary depend-

ing on the players' position (Di Salvo et al. 2007; Gregson et al. 2010; Carling, Le Gall, and Dupont 2012; Bradley et al. 2013; Martín-García et al. 2018; Altmann et al. 2021) or the competition: domestic league or international matches (Bradley et al. 2010; Di Salvo et al. 2013) and European or UEFA Champions League (Di Salvo et al. 2010). Another important area of study, given the high physical demands of this sport, is the study of fatigue's impact on player performance (Mohr, Krustrup, and Bangsbo 2003; Rampinini et al. 2009; Bradley and Noakes 2013; Smith et al. 2017). In addition, it is highlighted that, in recent years, research has also considered the necessity of studying mental fatigue's effect on physical performance (Coutinho et al. 2017; Jones et al. 2019). As well as in the case of research which used eventing and tracking data, GPS data has also been used to analyse the impact of contextual variables and the congested calendar on the physical performance (Rampinini et al. 2007; Lago et al. 2010; Dellal et al. 2015; García-Unanue et al. 2018).

As for the statistical methods used for studying the players' performance by position, the effect of fatigue, and the contextual variables, it is highlighted that most researchers used similar classical methods (Table 1.3).

1.1.4 Injuries

As mentioned in the previous section, elite football is characterised by its high physical demands, even more so when football teams compete in many matches and competitions across a season. However, despite the injury's negative effect on the team and players' performance, only a limited number of analyses have studied the causes of injuries (Carling et al. 2016). The reason for the scarce research on this topic compared to the rest of the study areas (Table 1.1,Table 1.2, and Table 1.3) may be because player injury belongs to the player's personal domain. Therefore, this information is only available by collaborating with football teams or through the public news. Table 1.4 shows an excerpt from these studies with their corresponding objectives, samples and methods.

Study	Objective	Sample	Methods
Nature of injuries			
Hawkins and Fuller (1999)	Define the causes of injuries during competition	Player injuries of the Premier League between 1994 and 1997	Chi-square significance test
Hawkins et al. (2001)	Study the injuries sustained over two competitive seasons	6030 injuries of the Premier League between 19974 and 1999	Chi-square significance test and Student's t-test
Ekstrand, Waldén, and Hägglund (2016)	Analyse time trends in hamstring injury over 13 consecutive seasons	Players injuries in 36 teams from 12 European countries between 2001 and 2014	Linear regression with log-transformed
Effect of the congested calendar on injuries			
Dupont et al. (2010)	Analyse the effects of a congested calendar on physical performance and injury rate	32 players from the same team which compete in the Scottish Premier League (2007-2008 and 2008-2009)	Paired t-test, Fisher exact test and ANOVA
Bengtsson, Ekstrand, and Hägglund (2013)	Study the relationship between recovery time and congested calendar and injury rates and team performance	8029 injuries from the players who competed in 27 European teams from 10 countries (2001 to 2012)	Chi-square significance test
Carling et al. (2016)	Investigate injury epidemiology during short periods of a congested calendar	14 players from the same team which compete in the Ligue 1 (2009-20010 to 2015-2016)	Paired t-test
Effect of training intensity on injuries			
Owen et al. (2015)	To examine whether an increase in training volume affects the incidence of injury or the probability of injury	130 injuries from 23 players who competed in the same top-European team	Stepwise multiple linear regression and Chi-square significance test
Ehrmann et al. (2016)	Investigate the relationship between GPS variables measured in training and gameplay and injury occurrences	19 players of the A-League Men	ANOVA

Table 1.4: Empirical studies on the causes of injury in elite football

According to Table 1.4, researchers have analysed the most common injuries in elite football players (Hawkins and Fuller 1999; Hawkins et al. 2001; Ekstrand, Waldén, and Hägglund 2016). In addition, other topics studied have been the effects of the congested calendar (Dupont et al. 2010; Bengtsson, Ekstrand, and Hägglund 2013; Carling et al. 2016) and training intensity on injury risk (Owen et al. 2015; Ehrmann et al. 2016).

Regarding the statistical methods used, except in the case of Ekstrand, Waldén, and Hägglund (2016) and Owen et al. (2015), who applied regression techniques, the rest of the researchers applied similar methods: ANOVA, Student and Chi-square test.

1.2 Objectives of this thesis

This thesis addresses some of the previous competitive areas related to sports and economics. Specifically, the aims are to analyse the behaviour of successful and unsuccessful teams on the field to assist the boards in the decision-making process. According to the nature of the data available, the analysis performed is related to the study of measurable game actions (see Section 1.1.1). This thesis is focused on applying multivariate statistical and machine learning techniques to improve decision-making process by comparing their predictive capacity and

studying the variables that most influence the prediction results. Specifically, the main objectives of this thesis are:

1. Compare the effectiveness of the classical statistical techniques used so far with multivariate statistical and machine learning techniques.
2. Implement predictive models based on multivariate statistical and machine learning techniques.
3. Propose multivariate statistical and machine learning techniques to determine the variables that most influence the prediction results of these models.
4. Design, develop and propose a new methodology to calculate several indicators that summarise information about the popularity of players and that can be useful to predict their market value.

Based on the main objectives of the thesis, it has been organised as follows:

Chapter 2: Quality or chance? Application of Machine Learning and Multivariate Statistics techniques to improve the decision making process

One of the big challenges of data analytics in any research field is improving and assisting organisational decision-making. Especially in sports, this issue is of utmost importance, as management has to deal with high-impact decisions, such as whether to terminate or renew a coach's contract. Thus, this chapter addresses objective 2 through the following key points:

- Propose the most accurate multivariate statistical model to predict the position of teams at the end of the season.
- Adjust and evaluate the machine learning and multivariate statistical models by incorporating a data balancing technique to the double cross-validation.
- Compare the accuracy of the models as a function of data balancing (i.e., unbalanced vs balanced data).
- Study misclassified teams to determine the game actions that cause misclassification.

Chapter 3: Exploring the success of "Big Five" football teams with Multivariate Statistics techniques

Multivariate models were fitted to explore the essential game actions to understand the critical differences between successful and unsuccessful teams and obtain a global vision of football teams' behaviour in the field of play. Therefore, this chapter handles objectives 1, 2, and 3; specifically, it will focus on the following:

- Compare multivariate statistical techniques with the classical two-sample univariate tests, highlighting the advantages and disadvantages of the different models.
- Show the benefits of using the PCA in the preliminary exploratory data analysis.
- Compare three supervised multivariate techniques, and choose the best model to identify the most contributive game actions to the team's success.
- Compare our results with previous authors who carried out similar studies.

Chapter 4: Using the Skellam regression model in combination with the Random Forest algorithm to predict match results

A fairly common study in sports analytics is developing methodological treatments to make forecasts, especially for predicting the result of the matches. Thus, this chapter will address objectives 1, 2, and 3 through the following key points:

- Combine the Skellam Regression Model, an approach based on the double Poisson distribution, with a Random Forest algorithm to solve some weaknesses of the Skellam Regression Model.
- Study the predictive accuracy of the Skellam Regression Model.
- Study teams' performance in Season 2019/2020 compared to Season 2020/2021.
- Compare the predictive accuracy of the Skellam Regression Model in the different leagues.
- Compare the effectiveness of the Skellam Regression Model with multivariate statistical and machine learning techniques.
- Compare our results with previous authors who carried out similar studies.

Chapter 5: Development of popularity indicators with Google Trends to measure popularity influence on the market value of players

The ultimate chapter, related to the teams' economic area, consists of a methodological proposal to study the effect of popularity in predicting a player's market value. This chapter handles objectives 2 and 4; specifically, it will focus on the following:

- Construct several popularity indicators to measure the players' popularity from the information provided by Google Trends.
- Select the best method and model to predict the transfer fee of players.
- Study the effect and suitability of the proposed popularity indicators.
- Compare our results with previous authors who carried out similar studies.

1.3 Contributions

1.3.1 Articles in peer-reviewed journals

1. Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Modelos de machine learning y estadística multivariante para predecir la posición de los equipos de primera división". In: Journal of Sports Economics & Management 12.1, pp.3–22.

2. Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Exploring essential variables for successful and unsuccessful football teams in the "Big Five" with multivariate supervised techniques." In: Electronic Journal of Applied Statistical Analysis 15.1, pp. 249-276.

1.3.2 Conference contributions

1. Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Essential variables for successful and unsuccessful football teams with multivariate supervised methods". In: XXXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO). Granada, Spain.

2. Malagón-Selma, Pilar, Ana Debón, and Josep Domenech (2022). "Influence of popularity on the transfer fees of football players". In: 4th International Conference on Advanced Research Methods and Analytics (CARMA 2022). Valencia, Spain, pp. 101–108.

3. Malagón-Selma, Pilar, Ana Debón, and Josep Domenech (2022). "Influencia de la popularidad en la tarifa de transferencia de los jugadores de fútbol". In: XI Congreso Iberoamericano de Economía del Deporte (CIED 12). Toledo, Spain, pp.73–76.

4. Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2021). "Cómo alcanzar puestos de promoción y evitar puestos de descenso". In: XI Congreso Iberoamericano de Economía del Deporte (CIED 11). A Coruña, Spain, pp.77–80.

Chapter 2

Quality or chance? Application of machine learning and multivariate statistics techniques to improve the decision making process

Part of the content of this chapter has been included in:

1. *Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2021). "Cómo alcanzar puestos de promoción y evitar puestos de descenso". In: XI Congreso Iberoamericano de Economía del Deporte (CIED 11). A Coruña, Spain, pp.77–80.*

2. *Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Modelos de machine learning y estadística multivariante para predecir la posición de los equipos de primera división". In: Journal of Sports Economics & Management 12.1, pp.3–22.*

Abstract

This chapter aims to find which Machine Learning and Multivariate Statistics techniques have a better predictive capability when classifying the teams' positions at the end of the season. The teams who competed in the "Big Five" (Bundesliga, Premier League, LaLiga, Ligue 1 and Serie A) throughout the 2018-2019 season were studied. The misclassified teams by the best of the models, the Random Forest with balanced data, were analysed in-depth to determine the game's actions that caused the misclassification. According to the results, the effectiveness of shots on goal and possession are the variables in which the poorly classified teams differ the most concerning their actual position. In conclusion, this chapter presents the valuable and successful way in which machine learning and multivariate statistical techniques can discriminate between the bottom and the top teams.

2.1 Introduction

The sports industry, especially football, stands out for its vast revenues. Consultancy firm Deloitte, which annually produces a report on the highest-earning football teams, has recently published a report on the earnings of the top 20 Money League clubs in 2020/21 (Ajadi et al. 2022). This report highlights that in the last year, Manchester City topped the Money League with revenues of €644.9 million, followed by Real Madrid (€640.7 million, second) and Bayern Munich (€611.4 million, third). According to Ajadi et al. (2022), team revenues are primarily supported by three essential line items: match day revenues (around 1% of revenues), broadcast revenues (about 40%-50% of revenues) and commercial revenues (approximately 40%-50% of revenues). However, while it may appear that team profits may depend on external factors, the truth is that both strong on-field performance and a solid commercial profile support the teams' income. Because the fact is that, in football, broadcast and commercial revenues depend on the position in the standings at the end of the season. Therefore, the better the final position, the higher the team's income. For example, in the Premier League, the most equitable of the major European leagues, the winner (Manchester City) in the 2020/21 season earned €152.5 million, while the bottom-placed team earned €97 million (Premier 2021). Another example of income imbalance could be the Spanish league, where 90% of television revenues go to LaLiga (First Division) and the remaining 10% to LaLiga 2 (Second Division).

Therefore, given team performance's high impact on football clubs' economic situation, much research has been devoted to studying efficiency in football (Espita-Escuer and García-Cebrían 2008; Boscá et al. 2009; Zambom-Ferraresi

et al. 2017). Thus, in recent years, the game actions that have a greater impact on winning, losing, or drawing a football match have been studied in depth (Peñas et al. 2010; Castellano, Casamichana, and Lago 2012; Liu et al. 2015a). Likewise, there is also extensive literature that analyses the specific game actions that contribute to reaching the top positions in the ranking (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019).

However, it seems necessary to highlight that football is a sport in which chance plays an important role since it is not always the team that plays the best that wins. Thus, although football is a purely results-oriented sport, researchers and analysts should be aware that only it is possible to model the objective part of football (team performance) and, through this reality, try to help team management to judge coaches and players based on objective data beyond the result at the end of the season.

Under this context, this chapter acquires relevance by discriminating between objective reasons that are measurables, as is the case of performance variables, which influence the final ranking of football teams, and causes purely by chance since sometimes the ball goes in or does not go in. Thus the work aims to use machine learning and multivariate statistical techniques to predict the ranking of teams at the end of the season. Then, the method that best differentiates the positions will be proposed, and misclassified teams will be studied deeply through the game action identified by previous authors as statistically significant to discriminate between positions (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019). It is considered that the results and conclusions of this article can provide handy information to sports managers, who, through objective indicators, improve the decision-making process, such as the termination of a coach's contract or his renewal.

This chapter consists of five sections. Section 2.2 describes the database and explains how machine learning and multivariate statistical techniques have been applied. Section 2.3 presents the results of the prediction and analysis of the misclassified teams. Sections 2.4 and 2.5 comprise the discussion and conclusion, respectively.

2.2 Material and methods

This section presents the database and the statistical methods applied. First, exploratory data analysis was performed through the principal component analysis (PCA) (Wold, Esbensen, and Geladi 1987), an unsupervised analysis technique commonly used for data exploration. Next, five supervised analysis techniques were selected to determine the best predictive model. The models used were: Partial Least Squares Discriminant Analysis (PLS-DA) (Wold, Johansson, Cocchi, et al. 1993), Random Forest (RF), Classification and Regression Trees (CART) (Breiman et al. 1984), Naïve Bayes (Maron 1961), and K-nearest neighbours (K-NN). Once the prediction was made and the best model was chosen, the radar plot (Kolence and Kiviat 1973) was used to compare the statistics of the poorly predicted teams with the average of their observed position. In this way, it was possible to have an overview of the behaviour of the teams. The R computer program was used to analyze the database (R Core Team 2019). R is free software from which a wide variety of statistical and graphical methods can be performed. RStudio (RStudio Team 2020) is an integrated development environment used to program in R.

2.2.1 Database

The database collected contains the statistics on the performance of football teams competing in European leagues (LaLiga, Premier League, Bundesliga, Serie A and Ligue 1) throughout the 2018-2019 season, with 98 observations (the football teams), 48 explanatory variables (performance variables) and the response variable "ranking". Considering the proposed objective, the teams were labelled according to their final position in the national league. Three positions were defined: "top" for those clubs whose classification allowed them to participate in the Champions League (20 teams), "bottom" for clubs that were relegated to the second division (15 teams), and "middle" for the rest (63 teams). Table 2.1 shows the variables used to perform the predictive analysis.The game's actions were collected through the data sources FBref (fbref.com), WhoScored (www.whoscored.com) and Fichajes.net.

Table 2.1: Variables classified by type of game actions

Type of variables	Game actions and abbreviations
Variables related to defensive actions	Shots conceded blocked (SCB), Recoveries (R), Clean sheets (CS), Penalties conceded (PC), Interceptions (I), Shots conceded on target inside the box (SCTI), Shots conceded on the target outside the box (SCTO), Tackles won (TW), Tackles lost (TL), Yellow cards (YC), Clearances (Cl), Fouls conceded (FC), and Tackles accuracy (TA)
Variables related to offensive actions	Corners won (CW), Crosses unsuccessful (CU), Successful crosses (SC), Dribbles successful (DS), Fouls won (FW), Dribbles unsuccessful (DU), Crossing accuracy (CrA), and Dribbles accuracy (DrA)
Variables related to the goal	Goals accuracy (GA), Goals inside the box (GIB), Key passes (KP), Penalties took (PT), Direct free kick goals (DFKG), Shots off target (SOT), Shots blocked (SB), Shots accuracy (SA), Assists (A), and Shots on target (ST)
Variables related to passes and possession	Passing accuracy (PA), Average possession (AP), Passes unsuccessful opponents half (PUOpp), Passes successful opponents half (PSOpp), Successful longpasses (SLP), Unsuccessful shortpasses (PUS), Successful shortpasses (PSS), Longpasses success (LPS), Unsuccessful longpasses (ULP), Passes per 90 mins (P_90), Passing accuracy in opponents half (PAOppH), Passing accuracy in own half (PAOwnH), Aerial duel accuracy (ADA), Aerial duel lost (ADL), Duels Accuracy (Dl_A), Duels lost (DlL), and Duels won (DlW)

2.2.2 Unsupervised learning methods

Unsupervised learning is a machine learning approach characterised by the fact that it works on unlabeled data, i.e., the class to which the individuals belong is unknown and, therefore, the response variable is unknown. This fact allows the model to discover patterns or undetected information.

Principal Component Analysis

The exploratory data analysis was carried out by using PCA a technique developed by Pearson (1901), who argued that in much scientific research, it would be desirable to be able "to represent a system of points in the plane, three, or higher dimensional space by the "best-fitting" straight line or plane". Thus, this technique aims to find the subspace of the variable space where the variability is higher. Therefore, through the transformation of the original variables, a smaller number of variables, principal components (PCs), are obtained that are not correlated with each other and explain most of the variability of the data (Wold, Esbensen, and Geladi 1987). PCA is defined as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.1}$$

where $\mathbf{X}$ is a data matrix N x M with N observations and M variables, $\mathbf{T}$ is the score matrix N x A, which contains the projection of the N observations in the A subspace of the PCs, $\mathbf{P}$ is the loading matrix M x A which represents the linear combination of the variables in each PCs, and $\mathbf{E}$ is the residual matrix N x M. The number of PCs, A, is selected by the researcher and varies according to the studied problem. PCA obtains detailed multivariate information about each observation's value for each extracted PC, which allows us to represent and analyse the relationship between observations and variables.

An additional advantage to the PCA is that when it is fitted, PCA provides two valuables' statistics for outlier detection: squared prediction error (SPE) and Hotelling's T^2 (Ferrer 2007). Specifically, SPE detects anomalous observations, while Hotelling's T^2 detects extreme observations. The difference is that an observation with a high SPE might indicate that it is breaking the correlation structure (anomalous observation), and a high Hotelling's T^2 refers to an observation that may have extremely high or low values in certain variables (extreme observation).

SPE measures the squared Euclidean distance of an observation from the M-dimensional original variable space to the A-dimensional subspace of the PC. SPE is defined as:

$$SPE_n = \mathbf{e}'_n\mathbf{e}_n = (\mathbf{x}'_n - \hat{\mathbf{x}}'_{n,A})(\mathbf{x}_n - \hat{\mathbf{x}}_{n,A}) = \sum(x_{n,m} - \hat{x}_{n,m,A})^2 \tag{2.2}$$

where $\mathbf{x}'_n$ is the row vector of the original values of each *n-th* observation ($n = 1, \ldots, N$) in the M variables, and $\hat{\mathbf{x}}'_{n,A}$ is the row vector of predicted

values of that *n-th* observation in the M variables in the A PCs subspace. The SPE statistic can be modelled by the (noncentral) χ^2 distribution:

$$SPE \sim g\chi^2_h \tag{2.3}$$

Hotelling's T^2 measures the estimated Mahalanobis squared distance from the projection of the observations in the PC subspace to the centre of this subspace. The Hotelling's T^2 statistic is defined as:

$$T^2_n = \sum_{a=1}^{A} \left(\frac{t_{n,a}}{s_a} \right)^2, \tag{2.4}$$

where $t_{n,a}$ is the score value of the *n-th* observation in the *a-th* latent variable ($a = 1, \ldots, A$) and s_a is the variance of the *a-th* dimension. The Hotelling's T^2 statistic can be modelled as a Snedecor F-distribution:

$$T^2 \sim \frac{A(N^2-1)}{N(N-A)} F_{A,(N-A)} \tag{2.5}$$

For both statistics, the control limits are calculated for confidence level of 95% and 99%. Thus, it can be expected that 5% or 1% of the observations, respectively, have values slightly higher than these control limits. This event is called a false alarm rate. However, if an observation far exceeds any of these control limits, it will be considered an outlier and deeply analysed.

2.2.3 Supervised learning methods

Supervised learning is a machine learning approach that works on previously labelled data. Thus unlike unsupervised learning, supervised learning methods a priori know the class to which the individuals belong to and, therefore, the response variable is known. This fact defines its use to train the algorithms and to make predictions.

Partial least square discriminant analysis

PLS-DA (Barker and Rayens 2003) is a Partial least squares (PLS) variant for classification models. PLS regression is a supervised multivariate technique commonly used to model the internal relationship between two matrices, $\mathbf{X}$ (predictors) and $\mathbf{Y}$ (responses). Unlike PCA, which seeks to maximise the variance of $\mathbf{X}$, PLS models find A latent variables (LVs) in $\mathbf{X}$ and $\mathbf{Y}$ space by maximising their covariance. As a result, the dimensionality of the data set is significantly reduced, and a few LVs are obtained that explain the sources of variability in $\mathbf{X}$ space that are related to $\mathbf{Y}$ space (Wold, Esbensen, and Geladi 1986). Specifically, the first LV has more information than the second, the second LV more than the third, and so on (Höskuldsson 1988). Thus, in classification problems, the $\mathbf{Y}$ matrix is built with dummy variables (as many classes to be considered, in our case three classes: top, middle and bottom).

One of the main advantages over other predictive models is that PLS admit correlated regressors, giving rise to easily interpretable models.

Classification and regression trees

CART is the name given to the decision tree algorithm (Breiman et al. 1984). They are commonly used to reveal the hidden structure of the data and reduce the number of possible predictors. The model is structured on a sequence of questions from which the tree is created. The tree is built through "nodes" that divide the data according to their characteristics. The algorithm starts at the initial node that uses one of the explanatory variables to split the data set into two parts as homogeneous as possible. The tree construction continues up to the "leaf" node, where the data are classified by class and probability according to the path taken (Nisbet, Elder, and Miner 2009).This procedure creates a tree-based classification model which classifies observations (classification trees) or predicts values (regression trees) of a dependent variable from the values of the predictor variables.

Random Forest

The RF algorithm (Breiman 2001) is a machine learning ensemble method [1] that uses the bagging [2] technique to combine trees randomly and improves predictive ability. Succinctly, the process begins with random sampling (with

[1] Ensemble methods use multiple algorithms to achieve better prediction (Opitz and Maclin 1999).
[2] Performs repeated training of the data set through a random subset (Breiman 1996).

replacement) of N observations (approximately two-thirds of the data) from the original database that forms the training dataset to build the tree, and the rests are out-of-bag observations (OOB). Unsampled observations OOB (approximately the remaining one-third of the data) make up the test set and will be employed to calculate the prediction error. Each constituted dataset creates an independent decision tree with a subset of variables randomly selected for each tree. In addition, each tree grows deep and without pruning. Ultimately, the test observations are classified according to how most trees have predicted them to measure classification error.

Naive Bayes

Naïve Bayes is a technique based on Bayes' theorem, which explains how, from a known event that meets certain conditions, it is possible to know the probability that another event with similar characteristics will also happen. This technique is characterised by the "naive" assumption that the variables are independent of each other (Maron 1961).

K-nearest neighbours

K-NN is an algorithm for regression and classification problems. In this method, from the information obtained in the training stage, individuals are classified according to the majority class of K nearest neighbours. The value of K, which the analyst determines, indicates the number of observations the algorithm uses to classify an observation. The metric is usually the Euclidean distance (Altman 1992).

2.2.4 Majority weighted minority oversampling technique

As it was said previously, to carry out the study, teams were labelled based on their position in the ranking at the end of the season (Bottom, Middle, and Top). However, according to the criteria used, the dataset was unbalanced, i.e. the classes were not represented equally. The main disadvantage of working with an unbalanced database is that the algorithms tend to classify the teams in the majority class since it leads to minimising the error rate. Therefore, a technique to handle unbalanced datasets was used to solve possible bias problems and discriminate among minority classes. The majority weighted minority oversampling (MWMOTE) technique was selected after comparing

its performance with other methods and obtaining the best results (Barua et al. 2015).

The MWMOTE technique is divided into three stages. In the first stage, the technique identifies the minority class individuals and constructs a new data set from them. In the second stage, the importance (weight) of the individuals identified in the first stage is calculated. The weight is computed considering the following items: the minority class individuals close to the majority class group will have more weight than those further, the minority class individuals in scattered groups will have more weight than those in a dense group, the minority class individuals nearby to a dense majority class group will have more weight than those nearby to a scatter majority class group. Finally, MWMOTE computes the importance (weight) of the individuals as the product of the closeness factor and the density factor and converts each weight into a selection probability. In the third stage, this technique clusters the minority data set using a modified hierarchical clustering algorithm and generates new individuals by interpolation (Barua et al. 2015). Thus, the mwmote function of the imbalance R-package (Cordón et al. 2018) was used to build the balanced dataset following the specifications explained previously. First, the *KNoisy* argument filters out those individuals of the minority set only surrounded by individuals of the majority class, thus eliminating possible noise from the data and preventing the new data set from containing it. Second, the *KMajority* parameter detects the position of individuals bordering the majority class. From this information, weights are assigned because these observations are considered more challenging to learn than individuals surrounded by observations of the same class. Third, the *KMinority* parameter indicates the number of original samples necessary to create the synthetic individuals. Fourth, the parameter *cclustering* designates the space the new samples will occupy. The values used in the investigation changed for the bottom and top classes since, according to the bibliography, the values that offer the best results should be selected in each specific case (Barua et al. 2015). According to Japkowicz (2000) the number of synthetic samples created should be 200 per cent of the original. Thus, 30 synthetic observations were generated for the top teams and 24 for the bottom teams.

2.2.5 Validation of learning methods

The double cross validation (2CV) technique was used to fit and evaluate the models (see Figure 2.1). This technique was selected because it optimises the parameters and evaluates the model with non-overlapping data sets, thus avoiding over-fitting problems (Stone, 1974). In 2CV, two splits are carried out. First, in the VC2, the database is randomly divided into F subgroups. From this division (VC2), the training set is created with F-1 subgroups (80% of the data), which are used to obtain the optimal model, and the remaining subgroup (validation set) is reserved for validation (20% of the data). After the first split (VC2), the training set is again randomly separated (VC1). From the second division (VC1), two new sets are obtained: the calibration set, consisting of F-1 subgroups (80% of the data), and the test set, consisting of the remaining subgroup (20% of the data). The training set, in VC1, is used for hyperparameter optimisation: *mtry*, the number of variables in each tree in the RF algorithm; K, the number of nearest neighbours used by each observation for classification in the K-NN algorithm; in the Naïve Bayes classifier, the parameter *usekernel* to decide if using a kernel density estimate or a gaussian density estimate; the parameter *Cp* to decide when the classification tree should stop growing and thus stop adding variables to the decision tree; LV, the optimal number of latent variables to build the PLS-DA model. Therefore, in VC1, the optimal model to perform the prediction in the validation set (VC2) is calculated. This process is repeated F times, using non-overlapping data sets to train and validate the model, concluding when all individuals have been once in both groups (Westerhuis et al. 2008; Szymańska et al. 2012).

The oversampling technique was applied to the training set at each iteration. Thus, in VC1, the MWMOTE technique was used to create synthetic data, which were used together with the actual data to calculate the optimal parameters. Then, the optimal models[3], were used to carry out the prediction on the validation set that consisted only of actual data. Previous research advocates that this is the correct way to validate the results in the context of unbalanced classes (Santos et al. 2018).

Once the F repetitions in VC1 and VC2 are completed, the value of the Matthews Correlation Coefficient (MCC) obtained in each model loop is averaged (see Figure 2.1). This coefficient uses the result of the confusion matrix, a table reporting TP, FP, FN and TN totals, to calculate the quality of the prediction:

[3]The result of the optimal parameters calculated in VC1

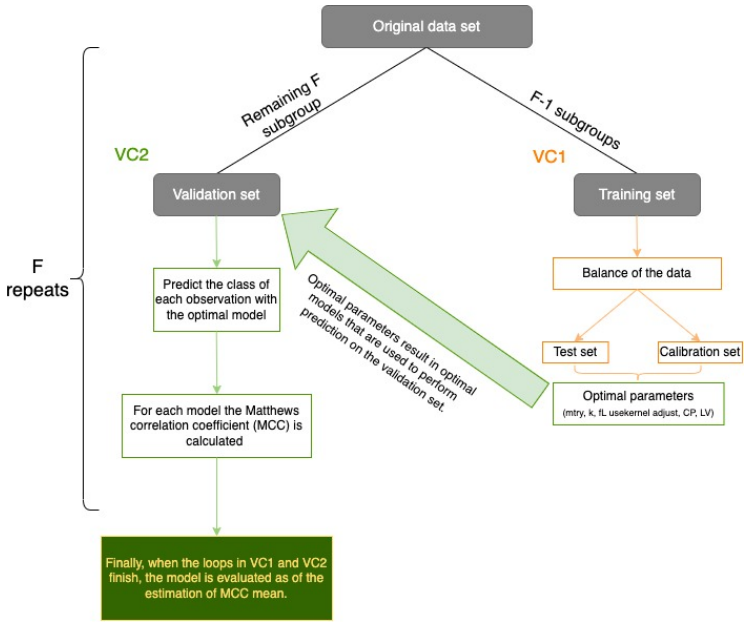


Figure 2.1: Diagram of the double cross validation used to evaluate the classification models.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FP)}} \tag{2.6}$$

where TP and TN are the numbers of true positives and negatives, respectively, and FP and FN are the number of false positives and false negatives, respectively (see Table 2.2). The MCC value can range from -1 to 1, where 1 represents a perfect prediction, and -1 indicates no relationship between the observed and the predicted (Matthews 1975). Therefore, MCC is used to evaluate the model with the best predictive capacity. A two-way ANOVA is then used to test whether statistically significant differences exist between using a balance or unbalance data set and between models. Thus, the test set is the block factor, and the model and balance are the main factors. Note that interaction was also considered.

Table 2.2: Confusion matrix showing the distribution of predictions at TP, FN, FP and TN for a classification model

		True/Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

2.2.6 Radar plot

Once the prediction has been made, the radar plot will be used to analyse misclassified teams. A radar plot is a handy tool for representing multivariate data in two dimensions (Saary 2008). These plots are characterised by their circular shape and the spoke, called radii, projected from the central point. The values of the variables are scaled to the radii's length and plotted on a two-dimensional plane (Budsaba, Smith, and Riviere 2000). Thus, the projection of the explanatory variables on a radar plot makes it possible to quickly and easily compare several observations simultaneously. In recent years this descriptive tool has gained notoriety in sports data analysis. Companies such as Opta (Carey and Sormaz 2019), Statbomb (Knutson 2020) or Driblab (Driblab 2020) often use these figures to evaluate teams and players. In addition, several researchers have started to illustrate performance variables using radar plots (Liu et al. 2016; Liu et al. 2015b; Oberstone 2009).

In our case, the game actions to be analysed are those highlighted in previous articles (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al.

2019) as statistically significant variables to differentiate between positions (see Table 2.3).

Table 2.3: Statistically significant variables (p-values<0.05) to differentiate among top, middle and bottom teams

Type of variables	**Oberstone (2009)**	**Lago-Peñas and Lago-Ballesteros (2010)**	**Souza et al. (2019)**
Variables related to defensive actions	YC and FC	-	SCTI, SCTO, R, YC,FC, and PC
Variables related to offensive actions	CW	-	CW, PT, and FW
Variables related to the goal	GA and ST	GA, A and ST	GA, DFKG, and SA
Variables related to passes and possession	PA, PSS, and LPS	AP	PA

Yellow cards (YC), Fouls conceded (FC), Shots conceded on target inside the box (SCTI), Shots conceded on the target outside the box (SCTO), Recoveries (R), Penalties conceded (PC), Corners won (CW), Penalties took (PT), Fouls won (FW), Average possession (AP), Passing accuracy (PA), Successful shortpasses (PSS), Longpasses success (LPS), Shots on target (ST), Goals accuracy (GA), Assists (A), Shots accuracy (SA), and Direct free kick goals (DFKG).

2.3 Results

The exploratory data analysis was carried out using PCA. Then, the prediction of the teams' position (bottom, middle or top) was calculated through five supervised learning models (PLS-DA, RF, CART, Naïve Bayes, and K-NN). Finally, the misclassified teams were studied to know the game actions that led to the error of the proposed models.

2.3.1 Unsupervised learning methods application

PCA was used as an unsupervised learning method to obtain a global view of the behavioural patterns of the observations. The PCA was obtained using the mixOmics R-package (Rohart et al. 2017). First, it was necessary to determine the number of PCs to obtain lower dimensional data while preserving as much variation in the data as possible. Figure 2.2 shows the evolution of the ratio of cumulative explained variance versus the number of PCs in the PCA model. Although there is no consensus on the best threshold for selecting CPs (Peres-Neto, Jackson, and Somers 2005; Jombart, Devillard, and Balloux 2010), we decided to keep nine CPs that accounted for 80% of the total variance. Then, SPE and Hotelling's T^2 with 95% and 99% control limits were used to check for anomalous or extreme observations, respectively.

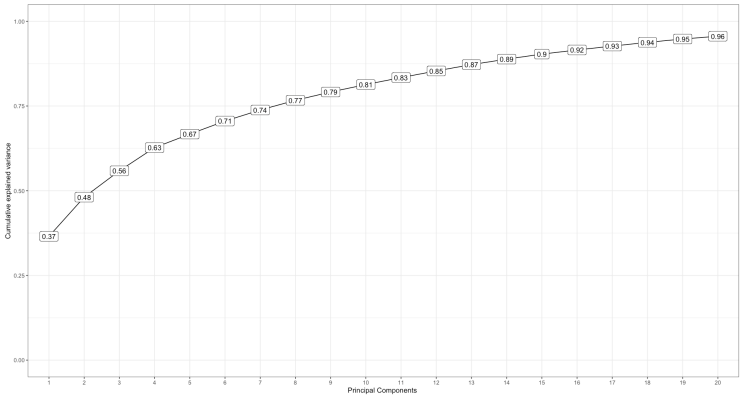


Figure 2.2: Cumulative explained variance ratio vs. the number of PCs.

Figure 2.6 shows the SPE with 95% and 99% control limits to verify that there are no anomalous observations. The five teams (23, 81, 84, 94, and 96) and the team (86) with SPE slightly higher than 95% and 99% control limits, respectively, are within the expected number of teams slightly trespassing these limits (0.05x98=4.9 on average; 0.01x98=0.98 on average).

Figure 2.4 shows the Hotelling's T^2 with 95% and 99% control limits to verify that there are no extreme observations. As in the case of the SPE (Figure 2.6), it is expected that some teams slightly trespass the 95% and 99% control limits.

Focusing on the first two PCs, Figure 2.5 shows the projection of the teams on the plane of the first two PC, which jointly explain 48% of the total variability.

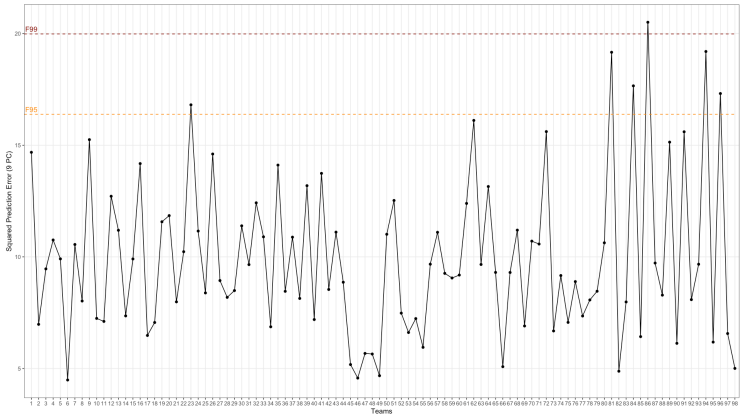


Figure 2.3: SPE of the PCA model with nine PCs for teams.

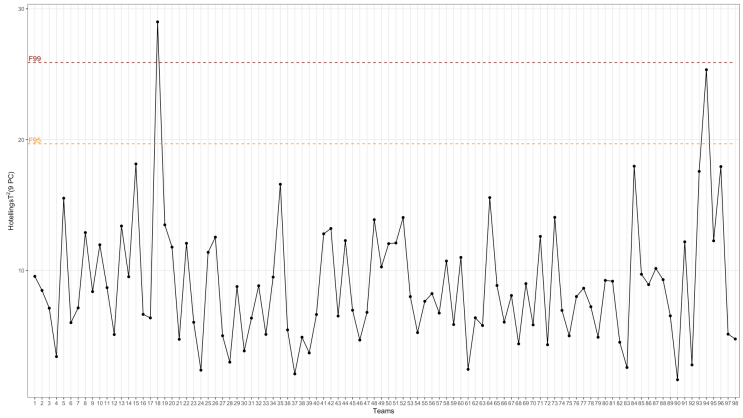


Figure 2.4: Hotelling's T^2 chart of the PCA model with nine PCs for teams.

The scores scatterplot (Figure 2.5) allows us to visualise the teams' relationship between groups and within groups.

Figure 2.5 shows the scores scatterplot of the first two PCs; the bottom teams are coloured in blue (circle shape), the middle teams in yellow (star shape) and the top teams are coloured in green (cross shape). Figure 2.5 reveals that although the teams appear clustered along the PC1 and the top and bottom teams are separated, the middle teams appear between the previous two, partially overlapping with both groups. In addition, the PCA scatterplot

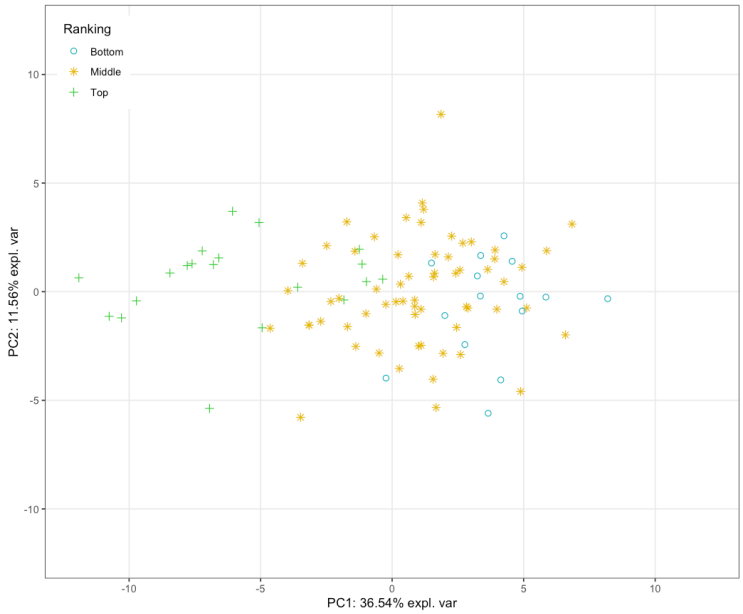


Figure 2.5: PCA scatterplot of team scores in the first two PCs (distribution of teams according to ranking; projected in PC1 / PC2) with indication of their position.

of team scores shows that the database is unbalanced, being the middle class the most numerous.

2.3.2 Supervised learning methods application

After applying the methodology described in Section 2.2.5, Table 2.4 shows the average Mathews Correlation Coefficient (MCC) of the F=5 replicates for each model obtained using the caret R-package (Kuhn 2020). This process was repeated on the unbalanced data to test which methodology was more efficient.

Table 2.4: MCC values of the supervised learning models for unbalanced and balanced data.

Model	CART	RF	K-NN	Naïve Bayes	PLS-DA
Imbalance	0.558	0.663	0.663	0.615	0.631
Balance	0.615	**0.722**	0.508	0.643	0.536

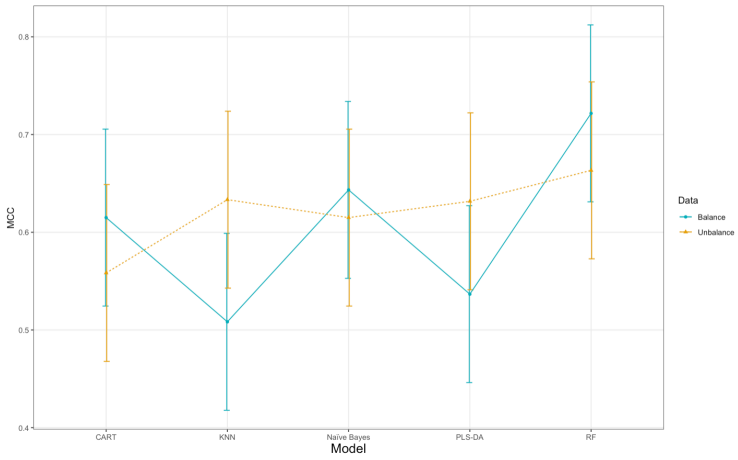


Figure 2.6: Multiple comparisons of the models (X-axis) vs. the MCC (Y-axis) as a function of the data balance. The dots indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences. The colour of the intervals indicates whether the MCC results correspond to a balance (blue) or unbalanced (yellow) data set.

Then, the two-way ANOVA was calculated from the MCC using two main factors: models and data set (balanced or unbalanced) and tested for statistically significant differences. The results indicated that the model factor approaches statistical significance ($pvalue$=0.0675). In addition, the factor interaction: model and data set (balanced or unbalanced) is not statistically significant ($pvalue$=0.1260).

According to Figure 2.6, the RF algorithm with the balanced data set was the model that provided the best results, as it has the highest mean MCC values. Furthermore, Figure 2.6 shows that the average MCC is statistically significantly higher in the RF algorithm with balanced data than in the PLS-DA and KNN with balanced data. Therefore, teams misclassified by the RF with balanced data will be studied in depth. Table 2.5 shows the confusion matrix of the RF algorithm for the F=5 replicates.

The results in Table 2.5 complete the information provided by the MCC values (Table 2.4). The main diagonal in Table 2.5 indicates the number of teams that have been correctly classified. The teams outside the diagonal are those whose prediction was wrong. Of the 60 middle teams, two were wrongly classified as

Table 2.5: General confusion matrix of the RF algorithm.

Predicted/Observed	Bottom	Middle	Top
Bottom	**7**	2	0
Middle	8	**55**	5
Top	0	3	**15**

bottom and three as top; of the 20 top teams, five were confused with middle teams; and of the 15 bottom teams, eight were classified as middle, being the class with the highest number of wrong classified teams.

These results, the confusion between teams of contiguous classes, were already intuited in the exploratory PCA analysis in Figure 2.5, especially highlighted between the bottom and middle classes. However, it is important to note that there has been no confusion between extreme classes, i.e., no top teams have been classified as bottom, nor vice versa.

2.3.3 Radar plot

Figure 2.7 shows the radar plot created using the plotly R-package (Sievert 2020). The variables that will be used to carry out the comparative analysis of the positions are shown on the axes of the radar plot. These variables were selected since, in previous articles, they were statistically significant in differentiating between positions.

Figure 2.7 highlights that top teams are those that carry out, on average, the highest number of offensive actions (CW and PT), more game actions related to the goal (GA, ST, SA, A and DFKG), complete more passes and have more possession (PA, AP, PSS and LPS). In addition, top teams performed, on average, fewer defensive actions (YC, FC, SCTI, SCTO, and PC), except for the variable number of recoveries (R), whose mean is higher for top teams. On the contrary, bottom teams perform, on average, fewer offensive actions (CW and PT), fewer game actions related to the goal (GA, ST, SA, A and DFKG), complete fewer passes and have less possession (PA, AP, PSS and LPS), and perform a higher number of defensive actions (YC, FC, SCTI, SCTO, and PC). Middle teams' averages in these variables are generally between the top and bottom teams. As exceptions, the average number of fouls received (FW) is the same for the three groups, and the average of the middle and bottom teams is almost the same in the case of fouls committed (FC) and yellow cards received (YC).

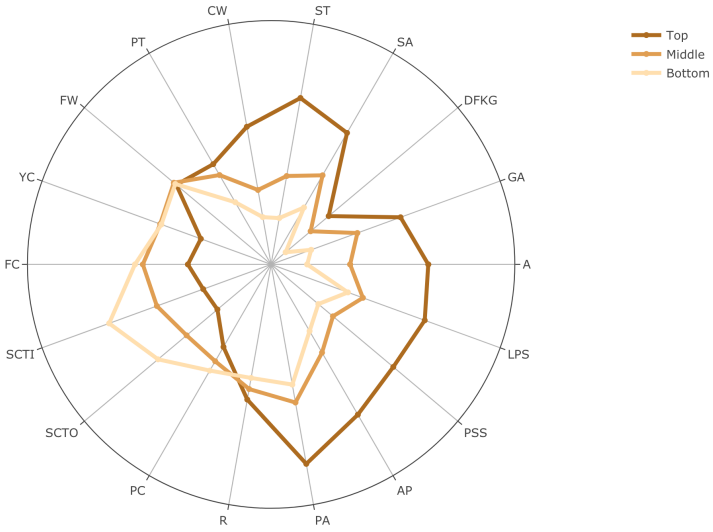


Figure 2.7: Radar plot to compare the mean values of statistically significant game actions to differentiate between positions of the bottom, middle and top teams.

To analyse the causes of the models' prediction errors and to go deeper into the behaviour of the teams, the teams misclassified by the RF algorithm will be studied in more detail. Thus, in the radar plot, the statistics of each mispredicted team and the average number of actions performed by the teams belonging to their observed position have been projected. The two middle teams classified as bottom (Figure 2.8), the eight bottom teams classified as middle (Figure 2.9), the three middle teams ranked as top (Figure 2.10) and the five top teams classified as middle (Figure 2.11) were analysed together.

Figure 2.8 shows the game actions of the two middle teams (observed position) that were classified as bottom (predicted position): Real Valladolid C.F. (purple dot) and Amiens Sporting Club (red dot). From the radar plot in Figure 2.8, it is observed that both teams executed fewer game actions related to the goal (GA, ST, SA and A) than the average of the middle teams. In addition, Amiens Sporting Club (red dot) executed fewer passes and had less ball possession (PA, AP, PSS and LPS), and carried out more defensive actions (PC, SCTO and FC) than the average of the middle teams.

Figure 2.9 shows the values of the performance variables of the eight bottom teams (observed position) that were classified as middle (predicted position):

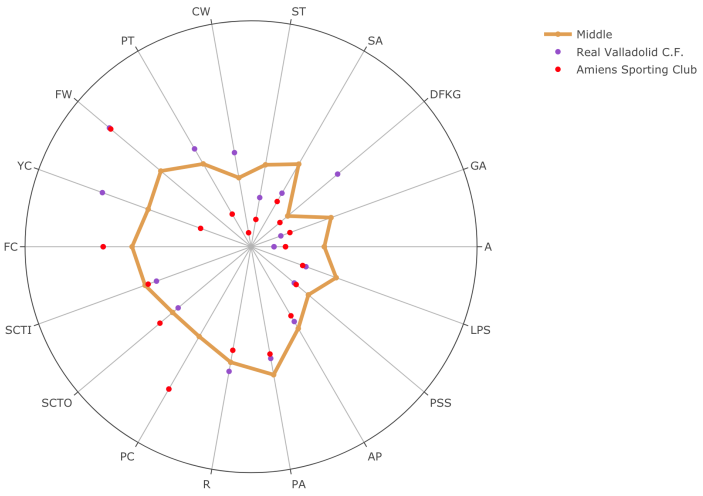


Figure 2.8: Radar plot for the comparison of teams misclassified as bottom with the mean values of the game actions (statistically significant to differentiate between positions) of the middle teams.

Rayo Vallecano de Madrid (maroon dot), Sociedad Deportiva Huesca (dark blue dot), Girona F.C (yellow dot), Dijon Football Côte d'Or (green dot), 1. F.C. Núremberg (black dot), Fulham (orange dot), En Avant de Guingamp (pink dot) and Cardiff City F.C. (light blue dot). Figure 2.9 shows that the teams that competed in LaLiga (Rayo Vallecano de Madrid (maroon dot), Sociedad Deportiva Huesca (dark blue dot) and Girona F.C. (yellow dot) performed a higher number of game actions related to the goal (GA, ST, SA and A) and had more possession (AP) than the average of the bottom teams. In addition, Rayo Vallecano de Madrid and Sociedad Deportiva Huesca received fewer shots from outside the box (SCTO) than the average of the bottom teams. Dijon F.C.O. (green dot) carried out a higher number of passes and had more ball possession (PA, AP and LPS), and performed a lower number of defensive actions (YC and SCTI) than the average of the bottom teams. 1. F.C. Núremberg (black dot) and Cardiff City F.C. also performed fewer defensive actions (PC, SCTO, SCTI, FC and YC) concerning the average of the bottom teams. The remaining teams, Fulham F.C. (orange dot) and En Avant de Guingamp (pink dot), presented differences with the bottom teams in the variables related to the goal (GA, ST, SA and A), the possession (AP) and in defensive actions (SCTI, SCTO).

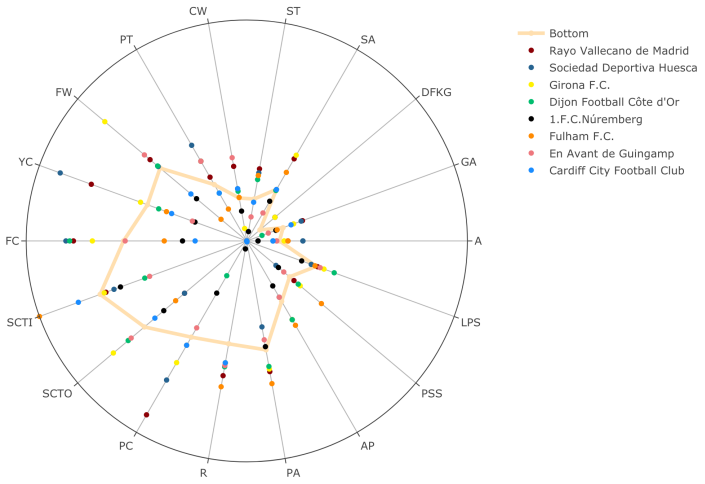


Figure 2.9: Radar plot for the comparison of the teams misclassified as middle with the mean values of the game actions (statistically significant to differentiate between positions) of the bottom teams.

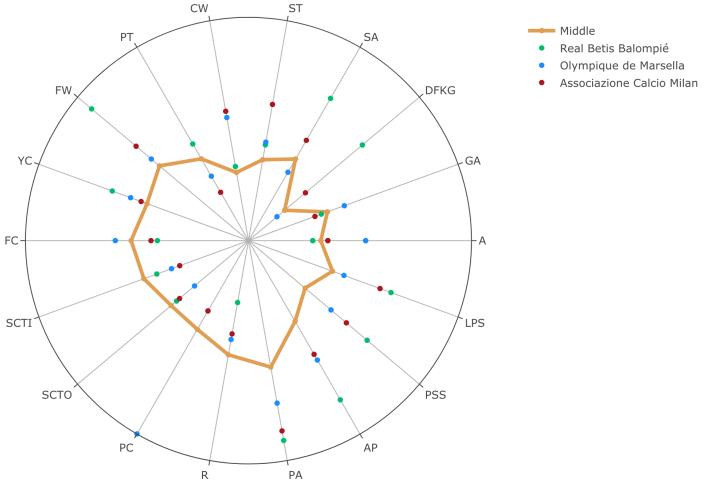


Figure 2.10: Radar plot for the comparison of the teams poorly classified as top with the mean values of the game actions (statistically significant to differentiate between positions) of the middle teams.

Figure 2.10 provides information on the performance variables of the three middle teams (observed position) ranked as top (predicted position): Real Betis Balompié (green dot), Olympique de Marseille (light blue dot) and Associazione Calcio Milan (maroon dot). Figure 2.10 highlights that the middle teams misclassified as top carried out fewer defensive actions (R, SCTI and SCTO), completed a high number of passes and had more possession (PA, AP, PSS and LPS) than the average of the middle teams. In addition, Olympique de Marseille (light blue dot) stands out for its high number of game actions related to the goal (GA and A). In the same way, the Real Betis Balompié (green dot) and Associazione Calcio Milan (maroon dot) also performed a high number of actions related to the goal (ST, SA and DFKG).

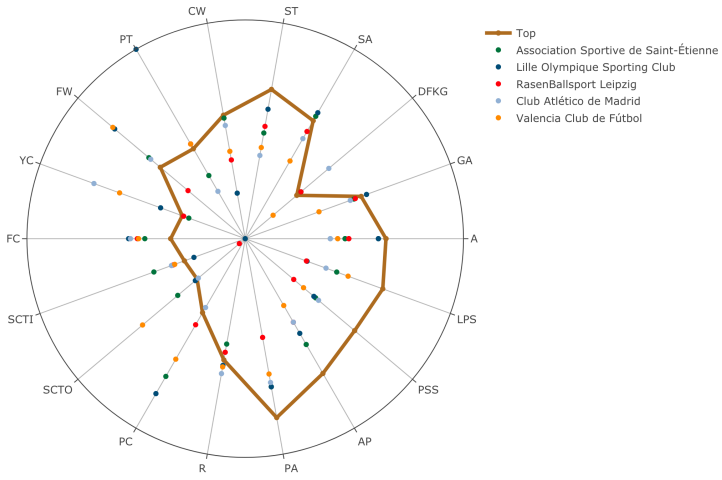


Figure 2.11: Radar plot for the comparison of the teams poorly classified as middle with the mean values of the game actions (statistically significant to differentiate between positions) of the top teams.

Figure 2.11 shows the game actions of the five top teams (observed position) that were classified as middle (predicted position): Association Sportive de Saint-Étienne (dark green dot), Lille Olympique Sporting Club (dark blue dot), RasenBallsport Leipzig (red dot), Club Atlético de Madrid (light blue dot) and Valencia Club de Fútbol (orange dot). Figure 2.11 highlights that the five misclassified teams executed, on average, fewer offensive actions (ST and CW), a lower number of passes, had less ball possession (PA, AP, PSS and LPS) and performed more defensive actions (FC) than the average of the top teams (observed position). In general, the misclassified teams had

similar defensive behaviour, as they received a high number of shots on goal from inside the box (SCTI) (Association Sportive de Saint-Étienne (dark green dot), Valencia Club de Fútbol (orange dot) and Club Atlético de Madrid (light blue dot)), and from outside the box (SCTO) (Association Sportive de Saint Étienne (dark green dot) and Valencia Club de Fútbol (orange dot)). Finally, in the case of Association Sportive de Saint-Étienne (dark green dot), Lille Olympique Sporting Club (dark blue dot), RasenBallsport Leipzig (red dot) and Valencia Club de Fútbol (orange dot), they also took a high number of penalties (PC).

2.4 Discussion

This chapter aims to use machine learning and multivariate statistics techniques to predict the final ranking position of first division football teams in the five major leagues (Premier League, Bundesliga, LaLiga, Serie A and Ligue 1) throughout the 2018/2019 season. In addition to proposing the best predictive model, this chapter analyses the misclassified ranked football teams by comparing the performance of the teams based on the game actions identified as statistically significant to discriminate positions in previous articles (Souza et al. 2019; Lago-Peñas and Lago-Ballesteros 2010; Oberstone 2009). This statistical analysis acquires relevance because it is a study conducted with data from teams that competed in the first division of the five most important leagues in the world. In addition, machine learning and multivariate statistics techniques have been used to perform the calculations, providing a different analysis method to those used in most previous studies. To the best of our knowledge, this analysis is the first to use machine learning and multivariate statistics methods to predict the final position of the teams that competed in the "Big Five". Concerning the study performed, the RF algorithm, using balanced data sets, has the highest mean MCC values to predict the final position of the football teams. Additionally, based on the results obtained, 2CV is recommended to avoid overfitting the data (see Table 2.4).

Even though the prediction of positions has not been perfect, it should be noted that no strange results have been obtained (Table 2.5), such as, for example: a top team being predicted as a bottom team (and vice versa); that a middle team ranked in the upper part of the table (e.g. fifth, sixth or seventh position) was predicted as a bottom team; or that a middle team ranked in the bottom half of the table (e.g. fifteenth, sixteenth or seventeenth position) was predicted as a top team. On the contrary, Real Valladolid C.F. and Amiens Sporting Club, which finished the season in sixteenth and fifteenth place (tied

on points with sixteenth), respectively, were predicted as bottom teams, being the teams that occupied the eighteenth, nineteenth and twentieth positions those relegated to the second division. In the case of Dijon F.C.O., which predicted position was middle instead of bottom, it finished the season in eighteenth place. However, in Ligue 1, occupying this position does not mean immediate relegation; instead, the team must participate in a play-off to opt to remain in the first division for the following season. Indeed, Dijon F.C.O. was not relegated. Similarly, Olympique de Marseille and Associazione Calcio Milan came in fifth place and were predicted as middle teams. In the same way, Valencia Club de Fútbol and Association Sportive de Saint-Étienne came in fourth place and were classified as top.

To further investigate the prediction errors, the variables identified in previous articles as statistically significant in differentiating between the position of the teams were used (Figure 2.8, Figure 2.9, Figure 2.10, Figure 2.11). To sum up, all misclassified teams showed effectiveness values (GA) different from the average of their observed position, except for teams wrongly predicted as middle. Most of the misclassified teams differed with the average of their observed position in the game actions related to passes and possession actions, highlighting the variables possession (PA) and passing effectiveness (PA). Regarding the defensive variables analysed, the number of shots received from outside the area (SCTO) and from inside the area (SCTI) were the game actions in which misclassified teams differed most from the average values of their observed position.

Football is a sport in which multiple factors interact, and chance plays an important role. Thus, the nature of football makes it difficult (and it would be a mistake) to judge the performance of teams based on a single indicator, such as the position on the table at the end of the season. Therefore, this analysis gains relevance by offering a methodology to model the objective part of football (team performance). Thus, the analysis methodology described is a valuable tool for sports managers since it quantifies the performance of coaches and players beyond the results at the end of the season.

The analysis of the incorrectly predicted teams (see Section 2.3.3) shows that sometimes the expected performance (position predicted by the predictive models) may differ from the observed performance (final position in the table). Thus, the essence of these discrepancies between predicted and observed positions may help sports managers' decision-making process. For example, according to our predictive model, a team that generates a reasonable number of scoring chances, concedes few shots and tends to have possession of the ball is expected to finish in mid-table positions. However, if instead of finishing

the season as a middle team, the team is relegated, it could be argued that this relegation could be the result of bad luck (chance factor) rather than the team's poor performance. Therefore, supposing that team management of this team (which has done well but has been relegated) has to decide whether to terminate the coach, they could find arguments to keep him/her in the position since, according to the analysis carried out, under normal conditions, it would be expected that his good indicators in the variables studied would translate into good sporting results.

This study demonstrates that in the decision-making process, observing a single indicator, such as the team's position at the end of the season, may not be the correct indicator to judge the performance of teams and coaches.

2.5 Conclusion

After discussing the results, it is concluded that data analysis techniques are a valuable tool for developing the team's strategy. Machine learning and multivariate statistical techniques can guide coaches and analysts who could check whether their teams' play is endangering their permanence in the first division or whether they are maintaining their place among top teams. In this context, analysis of the radar plot has shown, in a fast and intuitive way, that the performance of misclassified teams more closely resembles the average performance of the teams in which they have been predicted than those with which they share their observed position. In addition, this article is complementary to previous studies whose objective was to establish the game actions that most contribute to the success or failure of football teams.

This chapter can be of great use to sports managers, analysts and football experts, as it has been shown that the teams' performance does not consistently deliver the expected results. Therefore, this analysis suggests coaches and managers a new way to assess the performance of football teams at the end of the season, beyond the standings, and provide a tool to quickly and visually find the teams' weaknesses.

Chapter 3

Exploring the success of "Big Five" football teams with Multivariate Statistics techniques

Part of the content of this chapter has been included in:

1. *Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Essential variables for successful and unsuccessful football teams with multivariate supervised methods". In: XXXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO). Granada, Spain.*

2. *Malagón-Selma, Pilar, Ana Debón, and Alberto Ferrer (2022). "Exploring essential variables for successful and unsuccessful football teams in the "Big Five" with multivariate supervised techniques." In: Electronic Journal of Applied Statistical Analysis 15.1, pp. 249-276.*

Abstract

This chapter proposes three multivariate techniques to study the game actions that contribute the most to classifying the teams for Champions positions or descending to the Second Division at the end of the season. The game statistics of the teams that competed in the Bundesliga, Premier League, LaLiga, Ligue 1 and Series A during the 2018-2019 season are used. First, PCA is used for the detection of outliers and to obtain a preliminary vision of the behaviour of football teams. Statistically significant game actions were identified using supervised multivariate techniques: RF, PLS-DA and logistic regression. PLS-DA is the method that identifies more statistically significant variables when differentiating between the bottom and top teams. In addition, after comparing the results obtained with the univariate tests of two samples (such as the Student's t-test or the Mann-Whitney test), the advantages of using multivariate approaches to the univariate methods were confirmed. In conclusion, top teams stand out for being more offensive and stronger defensively. Contrarily, bottom teams are much less offensive and finish few offensive actions.

3.1 Introduction

Sports, especially football, have taken root as essential elements in society, both culturally and economically. Therefore, it is not surprising that data analysis has led to extensive literature investigating football characteristics in depth.

The game actions that significantly influence team success (Oberstone 2009; Schauberger, Groll, and Tutz 2017; Souza et al. 2019) and the influence of match events and game actions on the match's outcome have been extensively analysed (Lago and Martín 2007; Taylor et al. 2008; Lago 2009; Lago-Peñas 2010; Peñas et al. 2010; Lago-Peñas, Lago-Ballesteros, and Rey 2011; Castellano, Casamichana, and Lago 2012; Collet 2013; Liu et al. 2015a).

Likewise, the performance of football players as a function of their position during football matches has also been the subject of numerous studies (Carpita and Golia 2021; Carpita, Ciavolino, and Pasca 2021). According to the literature, researchers delving into this study area often find it difficult to compare their results, as there are no standard criteria for determining the number of positions in the field. However, regardless of whether researchers classify players according to their roles or spatial location, all studies have three classification groups in common: defender, midfielder and forward. Thus, players' physical activity (e.g., distance covered during high-speed and low-speed running) has been explored by video analysis according to player position. These studies

have been carried out in the leagues of several countries: LaLiga (Di Salvo et al. 2007), Premier League (Gregson et al. 2010), Serie A (Vigne et al. 2010), and Ligue 1 (Carling, Le Gall, and Dupont 2012); and in the Champions League (Rampinini et al. 2007; Bradley and Noakes 2013).

As reflected in Chapter 2, researchers have also studied the essential variables for determining the teams' classification at the end of a season (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019). These authors compared team rankings and identified variables that discriminate between positions. Specifically, researchers used univariate statistical techniques, Student's t-test (Souza et al. 2019) and one-way ANOVA (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010), to perform the analysis. However, univariate methods have a crucial disadvantage, as these techniques do not consider the relationship between variables. Therefore, it is impossible to know the correlations between statistically significant variables, being challenging to reach a global view of the behaviour of football teams on the field.

Oberstone (2009) also used points won during the 2007-2008 English Premier League season to conduct a multiple linear regression to identify statistically significant game actions for a team's success at the end of the season. However, he noted the need to use backward elimination in a stepwise regression to refine the analysis due to the possible existence of multicollinearity problems. A disadvantage of using this method is that correlated variables, important for discrimination between teams, could be eliminated, resulting in an unfinished interpretation of vital game actions.

This chapter is devoted to overcoming the limitations of previous studies (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019) by applying multivariate statistical techniques for analysing the game actions that have a greater contribution to the teams' success or failure. As in the previous studies (Liu et al. 2015a; Souza et al. 2019), the main objective is to determine which game actions contribute significantly to reaching the Champions League or Europa League positions or avoiding being relegated to the second division. In addition, this analysis can be helpful to coaches who could infer the results to consider the variables that would indicate a high probability of success or failure in future seasons.

This chapter consists of five sections. Section 3.2 is devoted to describing the database and summarising the methodology applied. Section 3.3 shows the game actions with the highest contribution to discrimination among each analysed group according to the proposed supervised learning methods. Furthermore, this section compares the classical two-sample univariate techniques

used so far with the proposed multivariate methods, highlighting the advantages and disadvantages of the different models. Sections 3.4 and 3.5 present the discussion and conclusion, respectively.

3.2 Material and methods

The following section presents the database and proposes different statistical methods to achieve the proposed objectives. Firstly, an exploratory analysis was carried out using the PCA (Wold, Esbensen, and Geladi 1987). PCA is a multivariate exploratory tool capable of showing the complex network relationships between observations and providing inside information on the behaviours of variables as a function of the studied classes (in our case, top and bottom teams). In addition, PCA is a valuable tool for detecting outliers that may compromise subsequent statistical analysis. Next, a confirmatory analysis was performed to determine which game actions contributed statistically significantly to the team's success at the end of the season. The multivariate models used were the following: RF (Breiman 2001), a machine learning algorithm widely used to date to identify important game actions within various sports (Carpita et al. 2015; Migliorati 2020; Smithies et al. 2021; Whitehead et al. 2021); PLS-DA (Wold, Johansson, Cocchi, et al. 1993), a method with outstanding results in most fields where it has been used, is especially suggested when working with many correlated variables (Gottfries et al. 1995; Worley and Powers 2013; Noçairi et al. 2016); and Logistic Regression (LR) (Nelder and Wedderburn 1972), a classical statistical model used as a benchmark. According to the literature, regression models have been used to explain the number of points won throughout the season (Oberstone 2009; Souza et al. 2019) to study the impact of ball possession on team success (Peñas et al. 2010; Collet 2013) and to predict the final score in a match (Lago 2009; Liu et al. 2016). As in the previous analysis, the free software R was used to conduct the research (R Core Team 2019).

3.2.1 Database

The database consists of 35 observations (top and bottom teams) corresponding to teams competing in European leagues (LaLiga, Premier League, Bundesliga, Serie A and Ligue 1) in the 2018-2019 season and 53 variables, of which 51 are explanatory variables. The remaining two are the name of the teams (categorical variable) and the final position in the league standings, either top or bottom (ordinal). The labelling of teams according to their final position was determined from previous research (Oberstone 2009; Lago-Peñas

and Lago-Ballesteros 2010; Souza et al. 2019). Thus, the "top" teams are those that qualified to play in the Champions League or Europa League (20 teams), while the "bottom" teams are those that were relegated to the second division (15 teams). To achieve the objectives of this chapter, in addition to using the variables shown in Table 2.1, four game actions directly related to the goal were included: Goals accuracy (GA), Goals inside the box (GIB), Goals outside the box (GOB) and Direct free kick goals (DFKG). Note that the data sources were the same as listed in Section 2.2.1.

3.2.2 Exploratory analysis

PCA

As explained in Section 2.2.2, PCA is one of the most widely applied multivariate statistical projection methods to reveal the internal structure of data. Specifically, the PCA model provides multivariate information through scatterplots of scores and loadings on the relationship between observations and variables for each PC extracted (Section 2.2.2).

3.2.3 Confirmatory analysis

Partial least square discriminant analysis

This chapter builds the PLS-DA model (see Section 3.2.3) Y matrix with two dummy variables (top and bottom). Thus, the dimension of X space is 51 (number of explanatory variables), and the dimension of Y space is 2 (number of labels). In PLS-DA, the variable influence on projection (VIP) is a measure that provides information about the contribution of each variable to the PLS-DA model (Eriksson et al. 2013). The VIP score is obtained through the following expression:

$$VIP_m = \sqrt{M \times \left(\frac{\sum_{a=1}^{A} w_{ma}^2 \times ssy_{total}}{ssy_{total}} \right)}$$

where w_{ma}^2 is the squared PLS weight of its respective *m-th* variable for each PLS *a-th* dimension, SSY_a is the sum of squares explained by that *a-th* dimension of the PLS, SSY_{total} is the total sum of squares explained by the PLS model, and M is the number of variables. Therefore, the result is a weighted

sum of the squared correlations between the PLS-DA components and the original *m-th* variable. The weights correspond to the percentage of variation explained by each component in the PLS-DA.

In addition to the VIP, which provides information about the contribution of each variable to the PLS-DA model, 95% jackknife confidence intervals can be calculated to select the statistically significant variables. The jackknife method is a resampling technique that estimates the variability of the PLS regression coefficients (Quenouille 1949). The procedure calculates the model's coefficients with N-1 observations (i.e. leaving each time a single observation out) as many times as N observations. Finally, the 95% jackknife confidence intervals for the PLS regression coefficients are calculated, being statistically significant those variables whose 95% jackknife confidence intervals do not contain the zero value.

Random Forest

RF is a data mining technique that stands out as an alternative to traditional classification trees. One of the reasons for its widespread use among sports analysts is that RF provides insight into the importance of variables in the classification models (see Section 3.2.3). Breiman (2001) proposed four ways to calculate the importance of variables, two of which were integrated into their randomForest R-package developed by Liaw and Wiener (2002): mean decrease accuracy (MDA) and mean decrease Gini (MDG). MDA is a measure obtained by calculating the increase in prediction error in OOB when the values of one variable in the training set are permuted, remaining the rest unchanged. In MDA, the greater the decrease in precision, the greater the importance of the variables. MDG is based on the Gini criterion. Each time the variable X_m is split, the impurity of the descendant nodes decreases. Thus, the impurity reaches the minimum (0) when the observations in the node split are classified into the same group. The MDG is obtained as a result of the sum of the Gini reductions for each variable, normalised by the number of trees. The greater the decrease in the Gini coefficient, the greater its importance.

Even though the MDA and MDG are practical measures of the contribution of variables to the model, a shortcoming of the library developed by Liaw and Wiener (2002) is that they did not include the calculation of the p-value that provides statistical significance. Thus, in this chapter two tests are used to determine the statistical significance of the variables. First, as proposed by Paluszyńska (2017), the p-value will be calculated through the one-sided binomial test, where the null hypothesis is that variable selection occurs by chance.

Thus, with the variable X_m randomly assigned to a number W of nodes, W can be modelled as a binomial statistical distribution $B(G, p)$, where G is the total number of nodes in the forest, and $p = 1/M$. Therefore, if the observed number of nodes where the variable X_m appears in the forest (r) is greater than the 95% percentile of the binomial distribution, with the p-value calculated as $P(W \geq r)$, the variable will be considered statistically significant. Second, this chapter proposes the calculation of the p-value using a permutation test. Permutation tests are non-parametric procedures that, like other hypothesis tests, represent statistical significance through the p-value (Edgington and Onghena 2007). These tests determine statistical significance by rearranging the labels of observations under the null hypothesis that the labels assigned to classes are interchangeable (Edgington and Onghena 2007). Thus, if a p-value of less than 0.05 is obtained, it is concluded that the original classification of each observation is relevant and, therefore, that the labels are not interchangeable (Knijnenburg et al. 2009). We propose calculating statistical significance from randomly reordered class labels (Tusher, Tibshirani, and Chu 2001; Subramanian et al. 2005). Specifically, the adjustment of 1000 RFs with and without the permuted response variable will be carried out. Thus, the p-value is calculated as the proportion of times that the MDA and MDG values obtained from permuted data are equal to or greater than the MDA and MDG got with the unpermuted response variable. In both cases (one-sided binomial test and permutation test), statistical significance was considered at p-values less than 0.05.

Logistic regression

LR model is a recognised model (Cox 1958) used in binary (1 or 0) response problems. LR uses the explanatory variables to model the probability that an observation belongs to a class. Thus, increasing or decreasing the value of the explanatory variables can influence the likelihood that an individual belongs to one class or another. However, a high correlation between explanatory variables can lead to multiple models with similar classification performance but different statistically significant regression coefficients, making interpreting the model complex. Therefore, to study the correlation between variables, the variance inflation factor (VIF) has been used to quantify the multicollinearity between these variables.

The VIF is expressed as:

$$VIF_m = \frac{1}{1 - R_m^2}$$

where R_m^2 is the determination coefficient for the regression of X_m on the remaining explanatory variables. To control for multicollinearity it was necessary to eliminate variables with a high VIF (note that depending on the context, the threshold may vary). Then, the stepwise variable selection method was used after applying the VIF and selecting the explanatory variables with the lowest VIF.

Resampling method for comparison of the model's performance

The performance of the model was evaluated using the cross validation (CV) technique. Therefore, in contrast to what was shown in Chapter 2, this analysis was carried out using a single CV. Note that a single CV was performed instead of double cross-validation due to the size of the dataset. In addition, it was taken into account that the subgroups were balanced. Therefore, for the CV, the subgroups were balanced by being composed of 4 top and 3 bottom teams. The evaluation and comparison of the performance of the test set were carried out using receiver operating characteristic (ROC) curves (Fawcett 2006), summarised as the area under the ROC curve (AUC). The AUC is a single scalar value between 0 and 1, representing a portion of the unit square area under the ROC curve. Therefore, the greater the AUC, the better the classification method. Then, a two-way ANOVA was used to test for statistically significant differences between the models, with the test set being the block factor and the model being the main factor.

Note that according to Refaeilzadeh, Tang, and Liu (2009), the CV not only assesses the prediction error but also allows the selection of models with generalisable results to other datasets.

Univariate approach: two-sample tests

As indicated in Section 1.2, one of the main objectives of this thesis and this chapter is to show the advantages of using multivariate techniques over classical univariate methods. Therefore, given that univariate techniques have been used to differentiate between successful and unsuccessful teams in this study area (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019), both approaches have been applied to facilitate their comparison.

Student's t-test was used to statistically contrast each univariate variable's effect on the teams' position at the end of the season (top versus bottom). Shapiro-Wilk test (Shapiro and Wilk 1965) was used to check that the nor-

mality condition were fulfilled, and the Levene test (Levene 1961) was used for homoscedasticity condition. If either assumption was rejected, the Mann-Whitney test (Wilcoxon 1945) or Welch's t-test (Welch 1947) was performed for a lack of normality and homoscedasticity, respectively. The threshold of statistical significance considered was 0.05.

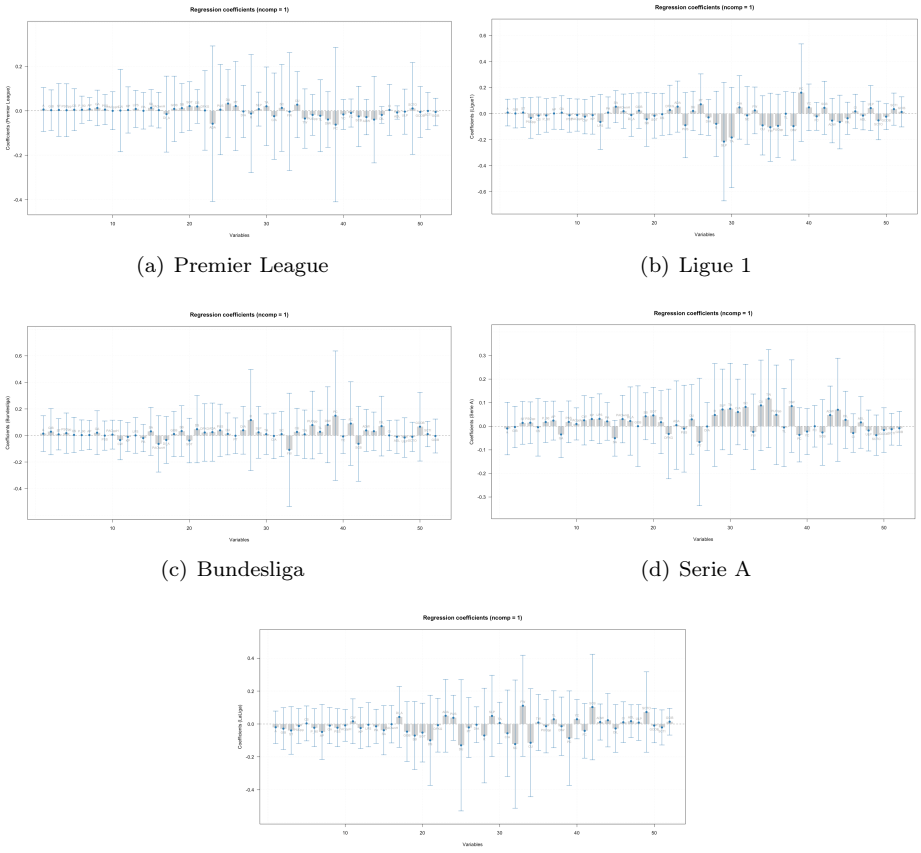
3.3 Results

The exploratory analysis of the data consisted of two parts: firstly, the data processing was carried out to verify that the distributions of the variables did not differ between leagues and, therefore, could be studied together. Secondly, PCA was used for the preliminary analysis of the data. Then, the selected models (PLS-DA, RF, and LR) were used to study the statistically significant variables to differentiate between top and bottom teams. Finally, the models were compared, and the best of them was selected.

3.3.1 Exploratory analysis

Although the analysis of the game actions that contribute most to the success or unsuccess of a football team has been studied before, few researchers have jointly studied data from European competitions (Collet 2013; Decroos et al. 2019). Therefore, before standardising the variables, it was verified that they had a similar distribution and that no vital information about playing style in the different leagues was lost during the normalisation process. Thus, a PLS-DA was performed, using leagues as the dependent variable, to calculate 95% jackknife confidence intervals and check whether any variable was statistically significant in discriminating between leagues, both for the bottom and top teams. Figures 3.1 and 3.2 show the 95% jackknife confidence intervals for the top and bottom teams in the five leagues, respectively, calculated from the mdatools R-package (Kucheryavskiy 2020).

Based on the results shown in Figures 3.1 and 3.2, it is concluded that none of the variables was statistically significant in discriminating between leagues, both for the bottom and top teams (p-value greater than 0.05). There are no statistically significant variables, as all jackknife confidence intervals contain the zero value. Furthermore, the mean and standard deviation were calculated (Tables A.1. and A.2.), and the boxplot was used to compare the standardised values of the playing actions differentiating by league and position (Figures A.1. and A.2.). After the above analyses, it was concluded that the behaviour of the



(a) Premier League

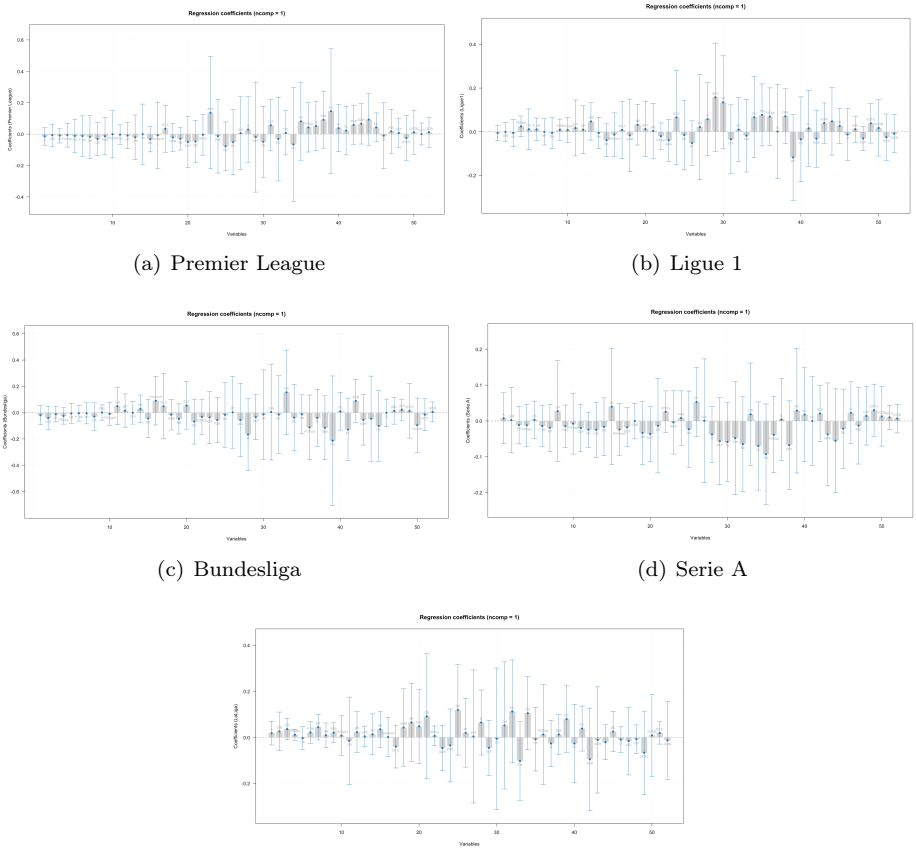
(b) Ligue 1

(c) Bundesliga

(d) Serie A

(e) LaLiga

Figure 3.1: PLS-DA regression coefficients with 95% jackknife confidence intervals for verifying no different behaviour on the top teams depending on the leagues



(a) Premier League

(b) Ligue 1

(c) Bundesliga

(d) Serie A

(e) LaLiga

Figure 3.2: PLS-DA regression coefficients with 95% jackknife confidence intervals for verifying no different behaviour on the bottom teams depending on the leagues

teams was quite similar, so it was possible to study all leagues together without the risk of losing important information during the normalisation process.

PCA

PCA was used in the exploratory analysis because it provides a global view of the relationships between observations and variables. Note that this is considered a decisive advantage compared to univariate techniques. First, the mixOmics R-package (Rohart et al. 2017) was used to find the number of PCs needed to summarise the data without losing important information. Thus, Figure 3.3 shows the evolution of the cumulative variance ratio explained according to the number of PCs selected. To explain 80% of the total variance (see Section 2.3.1), keeping seven PCs in the PCA model was necessary (Figure 3.3).

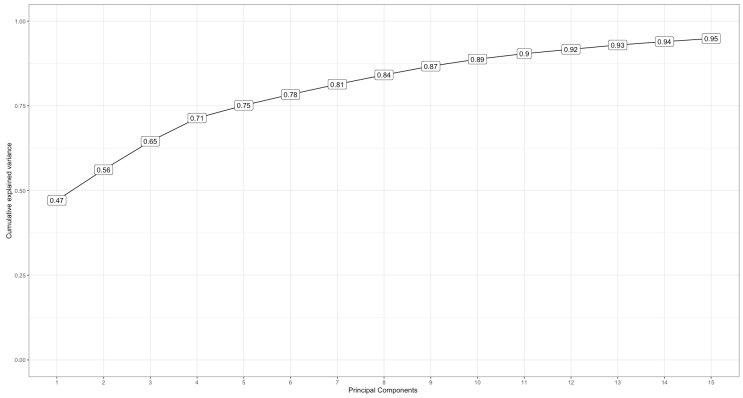


Figure 3.3: Cumulative explained variance ratio vs. the number of PCs

Figures 3.3 and 3.4 show the SPE and Hotelling's T^2 with 95% and 99% control limits used to verify no anomalous or extreme observations. The two teams (25 and 31), with SPE slightly above the 95% control limits, are within the expected number of teams slightly exceeding this limit (0.05x36=1.8 on average).

Once the number of required PCs had been established and verified that there are no anomalous or extreme data, the relationship between the teams and the variables was analysed. Figure 3.6 shows the scores scatterplot of the first two PCs, which together explain 56% of the total variability. The scores scatterplot (Figure 3.6) allows us to visualise the relationship between the

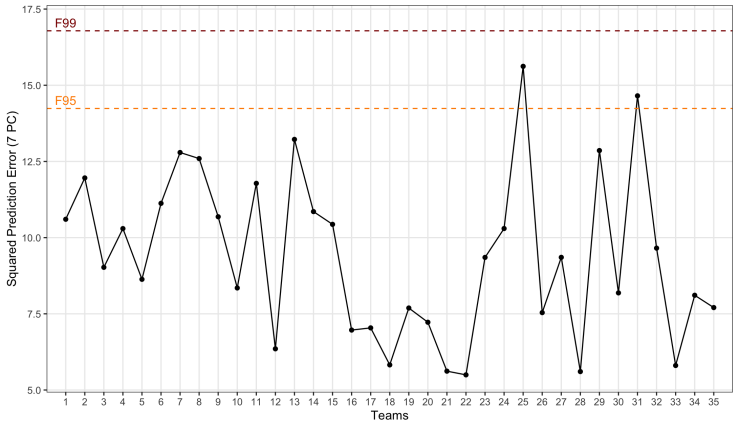


Figure 3.4: SPE of the PCA model with seven PCs for teams

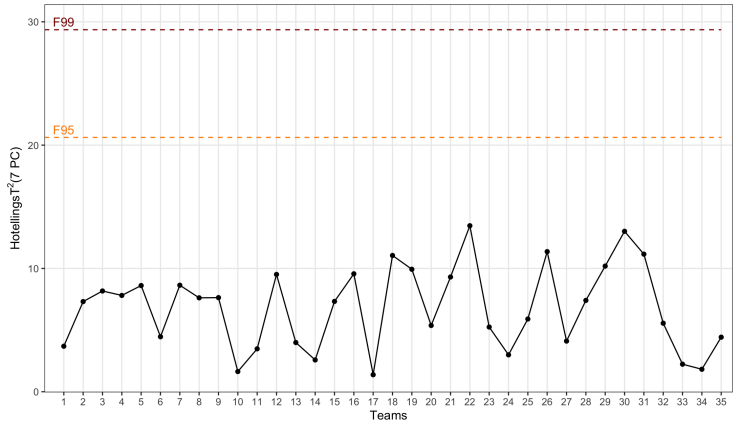


Figure 3.5: Hotelling's T^2 chart of the PCA model with seven PCs for teams

top teams (coloured in blue) and the bottom teams (coloured in red). In addition, Figure 3.6 allow us to see that the teams are not grouped according to their league but randomly within each top and bottom group. This is in line with the results of Figures 3.1 and 3.2 and justifies that all leagues can be analysed together. It is clear from Figure 3.6 that the teams are split by their position (top (right) and bottom (left) teams) along the first PC. Note that the first two PCs were chosen after performing all the score plots for each pair of PCs and finding that the first PC discriminated the most between teams (see Figures A.3. and A.4.).

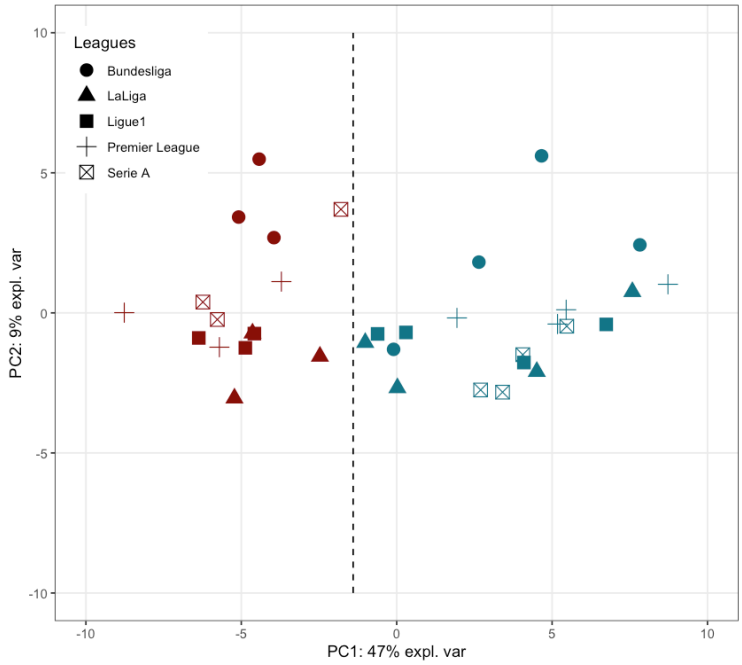


Figure 3.6: PCA scores scatterplot of the teams and leagues projected in the PC1/PC2 space: top teams in blue and bottom teams in red

Figure 3.7 shows the loadings scatterplot of the first two PCs, illustrating the relationship between the variables. The further a variable is from the origin of the coordinates, and the closer it is to a given PC, the greater the relationship with that PC, and vice versa. Thus, at both ends of the x-axis are the variables most highly correlated with PC1, the most negatively correlated in blue, and the most positively correlated in red.

Figure 3.7: PCA loadings scatterplot of the variables in the PC1/PC2 space sized by a variable's correlation strength to PC1. The colour of the dots indicates the negative (blue) or positive (red) correlation of the variables with PC1. Orange dotted arrow indicates the direction of the most discriminating PC

Figure 3.7 shows the game actions grouped into two clusters: in blue (dots on the left), the variables related to defensive actions (SCB, SCTO, SCTI, GCOB, GCIB, ADW, ADL, CI and I) and passes unsuccessful (ULP); and in red (dots on the right), those related to offensive actions (A, SA and ST), goals (GA and GIB) and passes and possession (AP, PA, P90, PAOwnH, LPS, PAOppH, PSS and PSOpp). Variables within a cluster will be positively correlated, while between clusters, they will be negatively correlated.

A comparison of Figures 3.6 and 3.7 can give a preliminary idea of the behaviour of the teams. Thus, by comparing Figures 3.6 and 3.7, one would expect the top teams (coloured in blue in Figure 3.6) to take higher values for the variables on the right of the loading scatterplot (Figure 3.7) and, therefore, lower values for the variables on the left. Conversely, the bottom teams (coloured in red in Figure 3.6) will take higher values for the variables on the left of the loading scatterplot (Figure 3.7) and lower values for those on the right. This exploratory analysis can provide some preliminary clues about the variables with discriminant behaviour between the top and bottom teams.

3.3.2 Confirmatory analysis

PLS-DA

In PLS-DA, the first step was to apply the perf function of the mixOmics R-package, which estimates the mean squared error of prediction and calculates the optimal number of latent variables needed to obtain the best predictive model. In this case, two PCs were extracted to obtain a goodness of fit (i.e. percentage explained of the total variance) of 55%. SPE and Hotelling's T^2 statistics were then analysed to check for anomalous and extreme data.

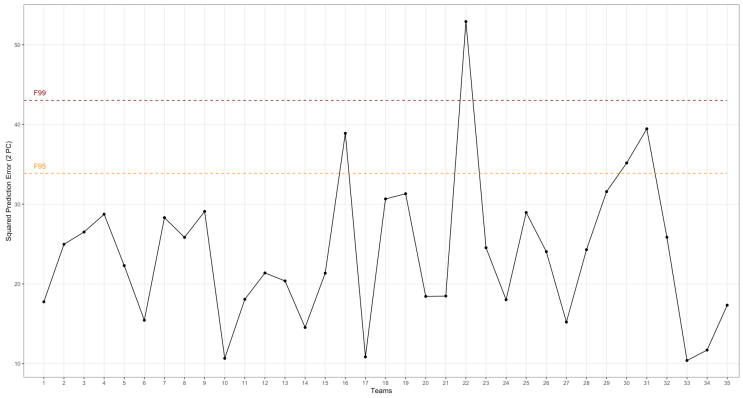


Figure 3.8: SPE of the PLS-DA model with two PCs for teams

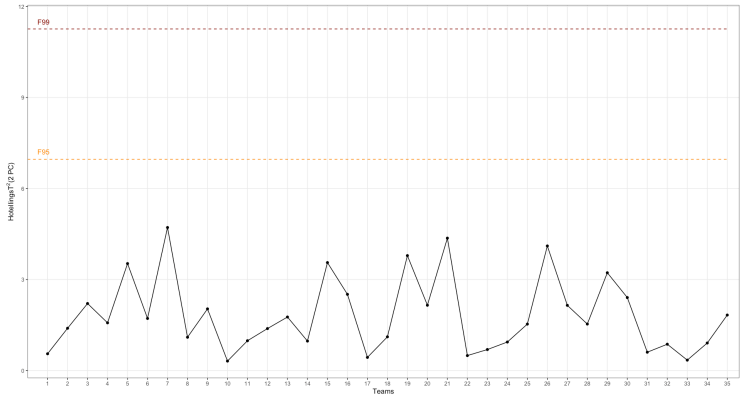


Figure 3.9: Hotelling's T^2 of the PLS-DA model with two PCs for teams

Figures 3.8 and 3.9 show the SPE and Hotelling's T^2 with 95% and 99% control limits. As in the case of the PCA, no anomalous or extreme observations were detected. Figure 3.10 shows the scores scatterplot of the two PLS-DA components coloured by their ranking label (top teams in blue and bottom teams in red). As shown, the first PLS-DA component clearly discriminates between positions.

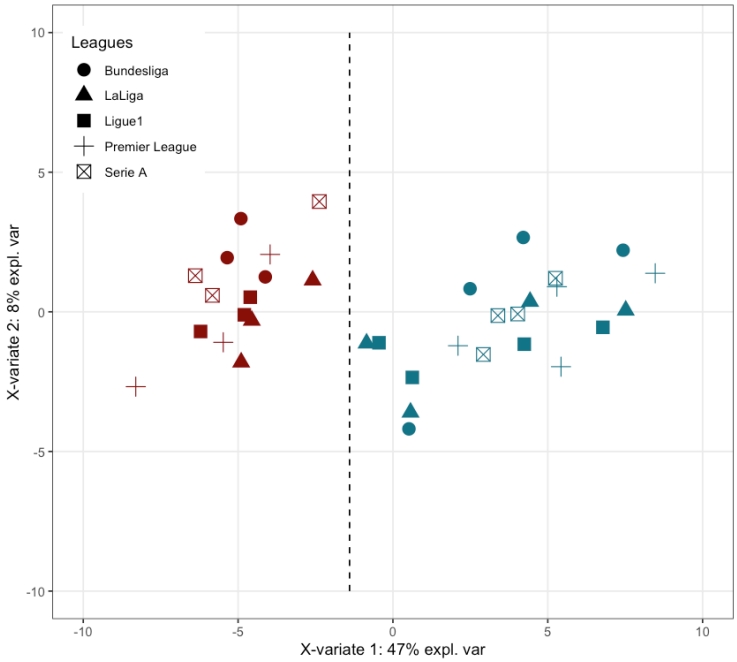


Figure 3.10: PLS-DA scores scatterplot of the distribution of the teams and leagues projected in the PLS-DA1/PLS-DA2 space: top teams in blue and bottom teams in red

In addition to the scores scatterplot, PLS-DA also allows representing the weightings scatterplot Figure 3.11), which reveals the relationship of the top and bottom teams with the explanatory variables. The further away a variable is from the centre, and the closer it is to the centroids of the top and bottom teams, the greater its contribution to the ranking of those classes (with higher values in the nearest category than in the opposite - the furthest one).

According to Figure 3.11, the higher the number of goals and shots conceded (SCTI, SCTO, GCIB, and GCOB), the higher the probability of belonging to a bottom team. Likewise, the higher the number of goals scored (GA), the

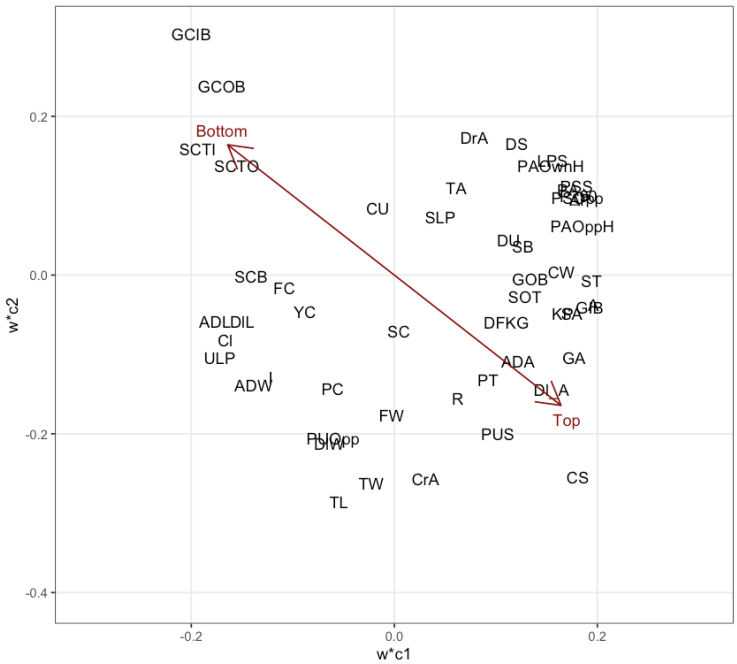


Figure 3.11: PLS-DA weightings scatterplot showing the relationship between the explanatory variables and the response variables in the PLS1/PLS2 space

percentage of one-on-one duels in which the player wins the ball (Dl_A), and the number of games the team finishes without conceding a goal (clean sheets, CS), the higher the probability of belonging to a top team.

The output of the PLS-DA model was used to obtain the VIP plot using the RVAideMemoire R-package (Hervé 2020). Figure 3.12 represents the contribution of each variable to the model. Following the approach of previous authors (Lazraq, Cléroux, and Gauchi 2003; Chong and Jun 2005; Sun et al. 2012), variables where the VIP was greater than one were selected as the most critical game actions to differentiate between top and bottom teams.

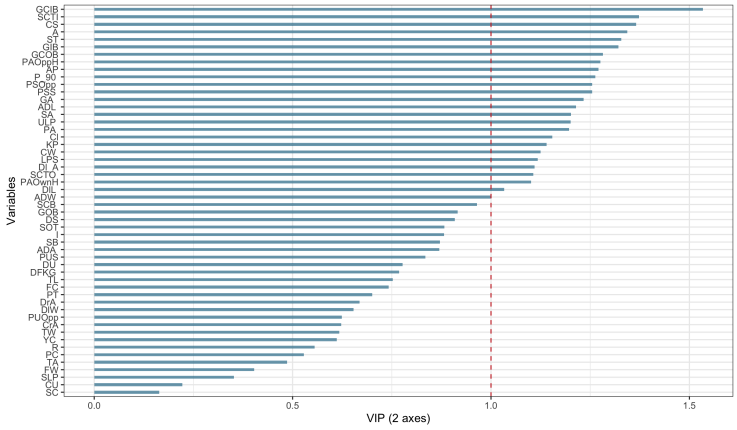


Figure 3.12: Importance of the variables in the model PLS-DA

Since the VIP only measures the importance of the variables, it was necessary to calculate the jackknife confidence intervals to select the statistically significant variables. Figure 3.13 shows the regression coefficients with 95% jackknife confidence intervals of the PLS-DA model for the bottom class, ordered from most negative to most positive values. Statistically significant variables are those whose jackknife confidence intervals do not contain the zero value.

Thus, variables with negative regression coefficients (such as CS, A, ST, GIB, GA, PSOpp, P90, AP and PSS) will take, on average, higher mean values in the top than in the bottom teams. Contrarily, variables with positive regression coefficients (such as SCTI, SCTO, GCIB, GCOB, ADL, Cl, DIL and ULP) will take, on average, higher mean values in the bottom than in the top teams. The PLS-DA regression coefficients of the top class are shown in the Appendix

(see Figure A.5.), as they are a specular image of the bottom class. The results of Figures 3.13 and A.5. are summarised in Table 3.1.

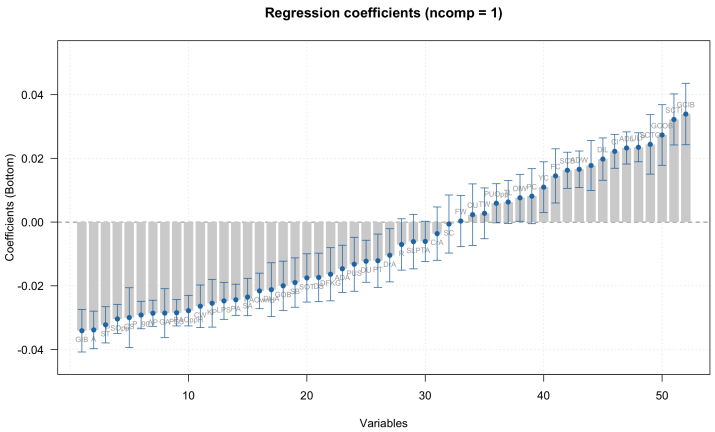


Figure 3.13: PLS-DA regression coefficients with 95% jackknife confidence intervals for the variables to predict the bottom teams

Note that the variables statistically nonsignificant for the jackknife confidence intervals (Figures 3.13 and A.5.), were not important for the VIP Figure 3.12. However, the jackknife confidence intervals identify some statistically significant variables that the VIP discarded (VIP values close to but lower than one). Finally, it is worth noting that, according to the results of the weighting plot (Figure 3.11), the VIP plot (Figure 3.12), and the jackknife confidence intervals (Figures 3.13 and A.5.), it is possible to confirm that the variables that contribute most to the discrimination between team positions coincide with those explained by the PC1 (Figure 3.7).

Random Forest

Liaw and Wiener (2002) implemented in therandomForest function of their randomForest R-package the calculation of the MDA and MDG measures (see Section 3.2.3). First, the RF algorithm optimises the hyperparameter mtry, the number of randomly selected variables in each tree, and calculates the nodesize, the minimum size of the terminal nodes. According to the results obtained in our analysis, the optimised values were 7 for the mtry and 9 for the nodesize. However, as indicated in Section 3.2.3, Liaw and Wiener (2002)

did not implement any test that calculated the statistical significance of the explanatory variables in their R-package. Therefore, two ways of performing the calculation are proposed: a one-sided binomial test and a permutation test. Figure 3.14 shows the results of the one-sided binomial test obtained using the randomForestExplainer R-package (Paluszyńska 2017). In Figure 3.14, the game actions are projected onto a variance significance scatterplot with MDA values on the x-axis and MDG values on the y-axis and labelled according to their p-value (red, green or blue).

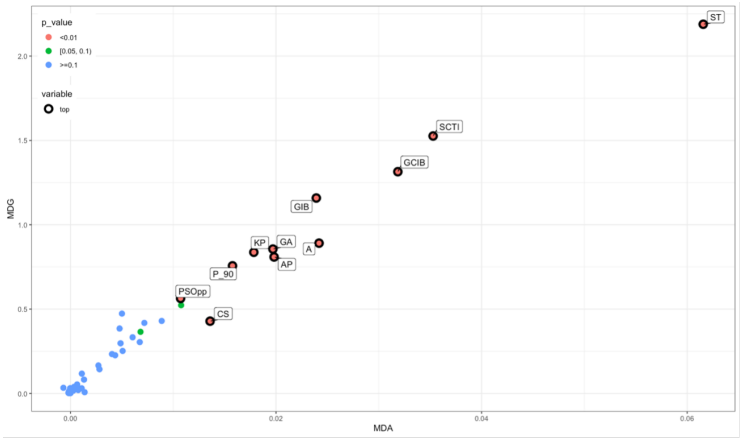


Figure 3.14: Multiway importance plot with mean decrease accuracy (MDA) and mean decrease Gini (MDG)

Figure 3.14 highlights the game actions with a p-value lower than 0.01 (in red), which coincides with the most important variables for both MDA and MDG: number of shots on goal (ST), shots conceded inside the box (SCTI), effectiveness (GA), assists (A), goals conceded inside the box (GCIB), number of goals inside the box (GIB), key passes (KP), average possession (AP), number of passes per 90 minutes (P90), successful passes in the opponent's half (PSOpp) and number of games the team has finished without conceding a goal (CS). Then, we used our proposed methodology to calculate the p-values of MDA and MDG using permutation tests. Table 3.1 shows the statistically significant variables for MDA and MDG permutation tests with a p-value lower than 0.05. According to Table 3.1, the MDG permutation test and the one-sided binomial test provide similar results. However, the MDA permutation, in addition to the variables selected by these approaches, found more game actions already selected from PLS-DA. These differences could be due to the MDG permutation test and the one-sided binomial test use the Gini criterion, while the

MDA permutation test measures the decrease in precision. According to the literature (Kim and Loh 2001; Strobl et al. 2007), the Gini index may be less reliable due to it benefits variables with many missing values or categorical cut-off points.

Logistic regression

The last model proposed was the LR, a classical statistical method used as a benchmark. First, as mentioned in Section 3.2.3, it was necessary to eliminate strongly correlated variables before fitting the LR. Thus, to mitigate multicollinearity, the vif_function (Beck 2013) was used to calculate the VIF value of all explanatory variables and eliminate the one with the highest VIF. This process was repeated until the remaining game actions had a VIF below the set threshold. Following previous researchers (Kutner et al. 2005; Sheather 2009; Johnston, Jones, and Manley 2018), three commonly used threshold values were selected: 2.5, 5 and 10. The output of the function was the name of the uncorrelated variables. Then, the stats R-package was then used to fit the model, and the stepAIC function was used to determine the most relevant variables according to the Akaike information criterion (AIC) (Akaike 1974).

Table 3.1, shows the statistically significant variables (p-values<0.05) for the LR models (thresholds 2.5, 5 and 10), PLS-DA and RF algorithm. According to Table 3.1, the LR is the model which selects a fewer number of statistically significant game actions. The reason is that, unlike PLS-DA and RF, LR penalises, through the VIF factor, the inclusion of collinear regressors. Therefore, the LR model tries to keep only uncorrelated or slightly correlated variables, regardless of whether the unselected variables are related to the response variable. For this reason, if the model contains collinear regressors, LR suffers from interpretation problems.

Resampling method for comparison of the model's performance

Once the prediction was made, the performance of the models was studied in depth to select the best one. For each CV test set, the AUC statistic was calculated using the ROCR R-package (Sing et al. 2005). Then, the mean model performance was calculated from the AUC and tested for statistically significant differences using a two-way ANOVA. The results of the two-way ANOVA indicated that the model factor was statistically significant (p-value$=0.0028$). Therefore, Fisher's 95% post hoc LSD interval test, implemented in the agricolae R-package (de Mendiburu 2021), was calculated to determine which models

Table 3.1: Comparison of the statistically significant variables (p-values<0.05) in the PLS-DA, RF and LR (thresholds 2.5, 5 and 10) models

Methods	**Variables related to defensive actions**	**Variables related to offensive actions**	**Variables related to the goal**	**Variables related to passes and possession**
PLS-DA: Jackknife intervals	SCTI[6], GCIB[4], CS[3], GCOB[3], SCTO[3], ADL[3], Cl[3], YC[2], ADA[2], SCB[1], I[1], FC[1], and ADW[1]	A[4], ST[4], KP[4], Dl_A[4], DlL[2], CW[2], SA[2], PT[1], SOT[2], SB[1], DrA[2], DS[1], DlW[1], and DU[1]	GIB[4], GA[4], DFKG[3], and GOB[1]	AP[4] P_90[4], PSOpp[4], PUS[4], PAOwnH[3], PSS[3], LPS[2], ULP[2], PA[2], and PAOppH[2]
RF: Binomial test	SCTI, GCIB, CS, and GCOB	A, ST, KP, and Dl_A	GIB and GA	AP, P_90, PSOpp, and PSS
RF: MDA	SCTI, GCIB, CS, GCOB, SCTO, ADL, and Cl	A, ST, KP, Dl_A, DlL, CW, and SA	GIB and GA	AP, P_90, PSOpp, PAOwnH, PSS, LPS, ULP, PA, and PAOppH
RF: MDG	SCTI, GCIB, and ADL	A, ST, and KP	GIB and GA	AP, P_90, and PSOpp
LR:10	SCTI, SCTO, ADA, Cl, TW[1], and TA[1]	Dl_A, SOT, and DrA	-	PUS, PUOpp[1]
LR:5	SCTI	CrA[1]	DFKG	PUS and PAOwnH
LR:2.5	YC	SOT and FW[1]	DFKG	PUS

The first time a variable appears, it is accompanied by a number that indicates the number of models for which it was statistically significant.

showed statistically significant differences. Since Fisher's post hoc 95% LSD interval test does not assume the AUC constraint (values between 0 and 1), the LSD intervals have been shortened.

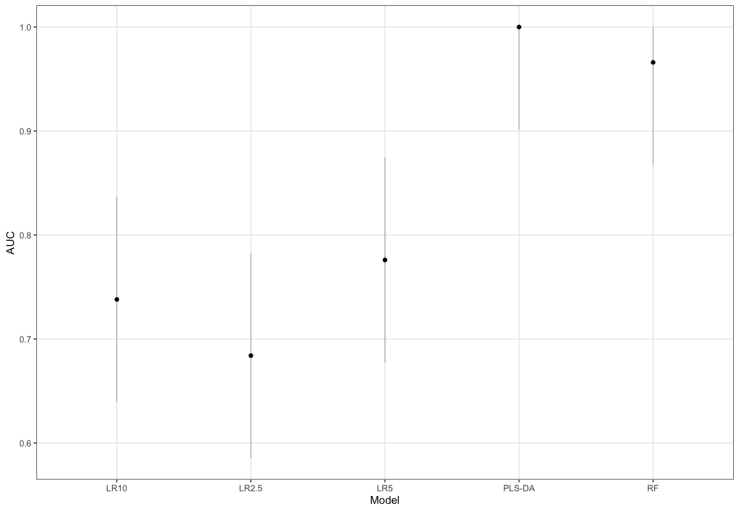


Figure 3.15: Multiple comparisons of the models (X-axis) vs. the AUC (Y-axis). The black points indicate the mean AUC for each model, and the intervals are based on 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences.

From Figure 3.15, it is possible to conclude that the average AUC of PLS-DA and RF is statistically higher than in the LR models. In addition, there is no statistically significant difference in the average AUC between PLS-DA and RF. Nevertheless, according to Table 3.1, PLS-DA is the method that selects more statistically significant variables to differentiate between successful and unsuccessful football teams.

Univariate approach: two-sample tests

This section is devoted to applying the univariate statistical two-sample tests. The Student's t-test, the Shapiro-Wilk test and the Levene test from the stats R-package (R Core Team 2019) were performed to check if the normality and homoscedasticity conditions were fulfilled. Thus, if the normality or homoscedasticity should be rejected, the Mann-Whitney test or Welch's approximation was used, respectively, instead of Student's t-test. Table 3.2 presents the statistically significant variables from the univariate analysis with a p-value less than 0.05.

Table 3.2: Statistically significant variables (p-values<0.05) for the two-sample test (top vs. bottom teams)

Type of variables	**Game actions**
Variables related to defensive actions	$SCTI^{B}$, $GCIB^{B}$, CS^{T}, $GCOB^{B}$, $SCTO^{B}$, ADL^{B}, Cl^{B}, YC^{B}, ADA^{T}, SCB^{B}, I^{B}, FC^{B}, and ADW^{B}
Variables related to offensive actions	A^{T}, ST^{T}, KP^{T}, Dl_A^{T}, DlL^{B}, CW^{T}, SA^{T}, PT^{T}, SOT^{T},SB^{T}, DS^{T}, and DU^{T}
Variables related to the goal	GIB^{T}, GA^{T}, $DFKG^{T}$, and GOB^{T}
Variables related to passes and possession	AP^{T}, P_90^{T}, $PSOpp^{T}$, PUS^{T} $PAOwnH^{T}$, PSS^{T}, LPS^{T}, ULP^{B}, PA^{T}, and $PAOppH^{T}$

^{T} indicates the variables that take higher mean values in the top teams than the bottom and ^{B} vice versa.

Furthermore, the results in Table 3.2 show that the statistically significant variables in the univariate analysis are consistent with the preliminary results of the PCA, ultimately confirmed by the PLS-DA.

3.4 Discussion

Even though the comparison between successful and unsuccessful teams has been investigated previously (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019), this analysis constitutes a novel approach by incorporating the game actions of teams that competed in the "Big Five" (Premier League, LaLiga, Bundesliga, Ligue 1 and Serie A) throughout the 2018-2019 season. To our knowledge, even though previous research has used data from European competitions (Collet 2013; Decroos et al. 2019), this study comprises the first analysis conducted jointly with match statistics from the world's top five leagues on this topic. In addition, classical two-sample univariate tests, commonly used in the literature, have been incorporated into this study to compare these approaches with the proposed multivariate statistical methods. In this sense, this chapter demonstrates the advantages of using PCA as a very effective technique to conduct preliminary exploratory data analysis. Instead

of analysing one variable at a time, as is done with univariate exploratory analysis techniques, PCA allows us to analyse all variables simultaneously and jointly interpret information from multivariate data. In particular, PCA is very effective in detecting outliers, finding observation patterns and visualising the relationship between variables. All this information provides a preliminary overview of the game actions that discriminate between the top and bottom teams, which were ratified by the results of the supervised multivariate techniques. Of the multivariate techniques studied, PLS-DA is the method that selects the largest number of relevant variables with statistically significant discriminant power to differentiate between top and bottom teams (these results support those obtained with PCA in the exploratory analysis of the data). Thus, together with RF, PLS-DA stands out as the technique with the best statistical classification performance.

According to the results obtained (see Table 3.1), the PLS-DA model found a high number of statistically significant defensive variables (SCTI, GCIB, CS, GCOB, SCTO, ADL, Cl, YC, ADA, SCB, I, FC and ADW). However, previous analyses only detected the goal average conceded per match (Oberstone 2009), shots conceded, recoveries (R), yellow cards (YC) and fouls conceded (FC) (Souza et al. 2019). Regarding the offensive variables, previous analyses differ in their results depending on the leagues, variables and the number of seasons employed (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019). In the case of PLS-DA, in addition to detecting the statistically significant variables studied by previous researchers, it also highlighted new variables not previously found (A, ST, KP, Dl_A, DlL, CW, SA, PT, SOT, SB, DrA, DS, DlW and DU). In the variables related to goals, our PLS-DA model found statistically significant all the variables analysed (GA, GOB, GIB and DFKG). However, previous studies only detected effectiveness (GA) (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019), goals outside the box (GOB) (Oberstone 2009) and free-kick goals (DFKG) (Souza et al. 2019). In the case of the variables related to passes and possession, the PLS-DA model identified the same statistically significant variables as previous researchers (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Peñas et al. 2010; Casal et al. 2019; Souza et al. 2019)x. In addition to these variables, the number of unsuccessful short passes (PUS), long passes success (LPS), unsuccessful long passes (ULP), and passing accuracy in its own half (PAOwnH) were detected as game actions with high discriminant power.

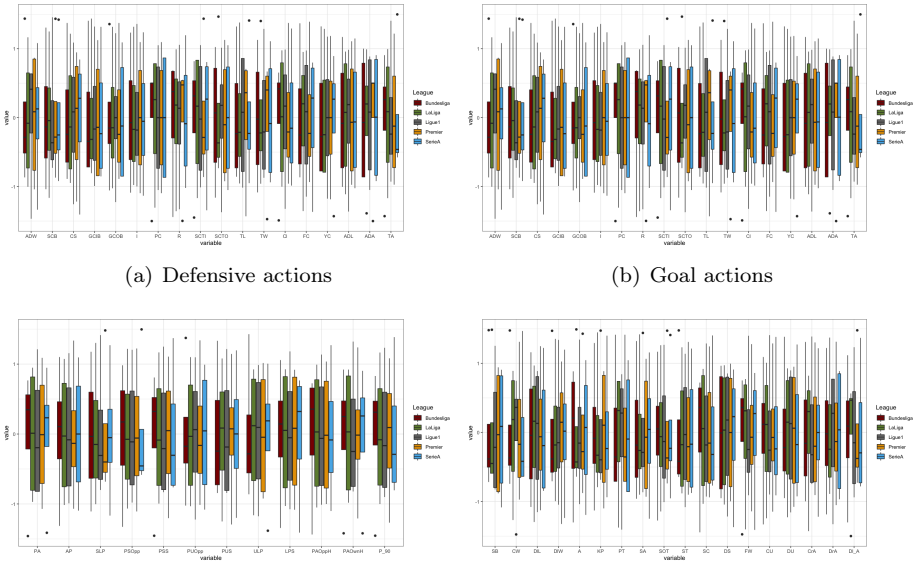
As for the commonly used method for data analysis in football, the two-sample Student's t-test (Rampinini et al. 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019), this univariate analysis is not very efficient. Firstly, it

is necessary to conduct as many statistical tests as there are game actions in the database, which may cause multiple comparison problems because of the high number of hypothesis tests conducted. Secondly, since the variables (i.e. game actions) are analysed independently (i.e. one at a time), the tests do not supply information on the relationships between game actions, which makes it challenging to obtain a complete view of the behaviour of football teams on the pitch. Moreover, most of the game actions that the Student's t-tests detected as statistically significant (see Table 3.2) were also identified by the multivariate models, in particular by PLS-DA (see Table 3.1), which also selected two additional game actions: DrA and DlW. Nevertheless, it should be noted that the univariate approach can be helpful only for testing the statistical significance of a single predictor but should not be used if the analysis aims to predict the classification of the teams.

3.5 Conclusion

This chapter significantly contributes to sports analytics, especially football analytics, as it proposes powerful multivariate techniques such as PCA and PLS-DA to incorporate into the analysis toolkit. Furthermore, this study shows the advantages of multivariate methods, which can be used for exploratory and confirmatory data analysis. Regarding the analysis of the game actions, it is highlighted that although football is a game in which the team that scores the most goals is the one that wins, not everything depends exclusively on the goals. Still, the defensive and offensive strategies have a great weight in the final result. It is important to note that since the game actions used in the analysis measure information accumulated until the end of the season (when the final ranking is already known), there is no value in using any predictive model to predict it. However, this information can be helpful for coaches, sports analysts and researchers to plan future seasons.

3.6 Appendix



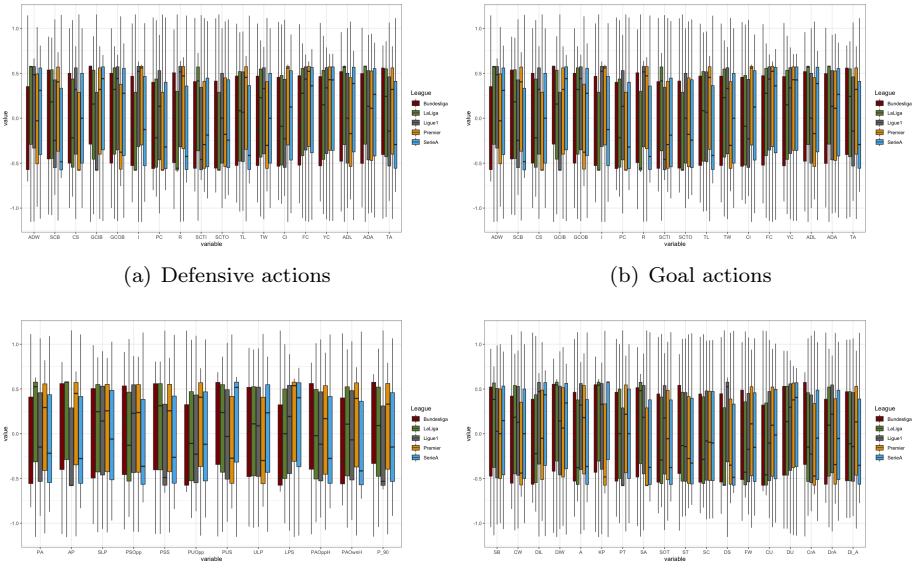
(a) Defensive actions

(b) Goal actions

(c) Passing and possession actions

(d) Ofensive actions

Figure A.1.: Boxplot with standardised values for the Top teams in each league



(a) Defensive actions

(b) Goal actions

(c) Passing and possession actions

(d) Ofensive actions

Figure A.2.: Boxplot with standardised values for the bottom teams in each league

Table A.1.: Mean and standard deviation of the variables for the top teams in the "Big Five"

	Premier		**Ligue 1**		**Bundesliga**		**Serie A**		**LaLiga**	
Variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
SCB	81.5	21.3	108.0	19.2	76.5	14.9	108.8	15.0	98.3	17.4
R	2333.8	75.3	2253.0	29.4	2166.3	111.9	2304.5	136.8	2273.0	38.3
CS	17.5	3.7	13.8	3.0	11.8	3.1	14.0	3.6	15.5	3.7
PC	3.0	1.8	7.5	2.4	5.8	0.5	4.5	0.6	5.8	2.9
I	338.0	21.9	380.5	41.6	373.5	34.5	381.8	61.6	410.8	39.6
SCTI	83.3	27.0	98.3	12.1	88.5	23.8	91.0	11.2	98.8	1.5
SCTO	34.3	12.2	45.8	4.2	33.5	7.3	39.5	2.4	44.0	9.6
TW	360.0	22.4	401.0	24.1	344.5	46.6	379.8	26.4	403.3	59.8
TL	232.5	29.0	259.5	6.4	206.0	40.8	219.5	15.3	233.8	45.8
YC	46.5	8.0	61.5	7.2	47.8	10.0	69.0	5.6	83.0	10.1
Cl	636.3	97.4	628.3	106.0	569.3	151.2	600.5	22.8	660.0	157.8
FC	339.5	26.2	462.0	17.5	359.0	77.5	448.5	28.2	448.8	28.6
TA	60.9	1.8	60.7	0.9	62.7	1.6	63.4	2.2	63.5	1.2
ADW	574.5	40.9	567.5	101.9	583.8	125.5	536.8	96.4	567.8	107.5
ADL	565.3	48.5	538.5	137.7	535.8	81.4	533.0	72.3	522.8	127.6
ADA	50.4	0.6	51.7	3.1	52.0	2.8	50.0	3.6	52.3	2.4
GCIB	27.0	7.1	35.0	6.1	32.8	9.4	31.5	6.5	33.8	7.1
GCOB	3.8	3.1	4.0	0.8	6.5	3.3	4.8	2.1	2.8	1.7
KP	400.8	49.9	364.5	51.9	341.0	54.1	463.3	33.4	356.5	48.7
PT	5.0	1.4	10.0	3.6	5.0	1.4	6.3	2.6	7.3	2.4
SOT	213.0	10.9	214.8	32.5	200.5	38.1	269.8	20.7	205.8	20.8
SB	169.0	31.8	129.0	27.9	119.5	10.5	163.3	8.5	114.5	21.2
CW	240.5	46.9	213.8	33.7	209.5	53.2	261.5	27.6	211.5	10.7
CU	395.0	41.9	344.3	61.6	311.3	69.4	498.8	85.5	335.3	109.8
SA	50.5	2.6	49.8	0.6	50.6	3.2	44.3	1.1	48.8	5.0
A	56.5	10.8	53.5	13.0	53.5	9.6	48.3	10.3	43.5	11.7
ST	218.5	31.7	213.5	33.8	204.8	35.4	214.5	17.7	198.8	44.0
SC	93.8	14.7	90.3	20.8	85.0	19.4	144.0	37.8	86.3	31.3
DS	403.0	46.2	398.5	72.0	375.5	66.8	340.3	40.4	404.0	57.6
FW	359.0	50.9	469.8	29.5	375.8	22.9	468.0	38.3	515.5	44.6
DU	278.0	24.2	280.0	23.4	257.3	28.7	254.8	30.0	263.5	46.9
CrA	19.3	3.9	20.8	3.4	21.4	1.4	22.2	1.6	20.4	1.7
DrA	59.1	3.8	58.5	2.5	59.2	2.7	57.2	1.9	60.6	2.5
Dl_A	50.7	1.0	50.9	1.5	51.7	1.4	51.0	1.4	52.4	1.0
DlL	1862.0	56.9	2002.3	104.4	1751.0	192.1	1846.5	147.5	1914.0	122.8
DlW	1918.8	92.8	2075.8	76.4	1868.8	161.3	1925.3	170.9	2106.5	127.4
GA	18.1	2.3	17.7	3.9	18.8	3.5	14.4	1.6	15.9	2.9
GIB	67.8	16.5	67.3	19.6	66.3	7.7	60.0	7.9	55.8	15.8
GOB	10.8	4.6	8.3	1.5	9.0	3.7	9.5	5.9	9.0	5.2
DFKG	1.3	1.0	2.3	1.7	1.8	1.3	1.3	1.9	2.8	2.4
PA	86.1	2.7	84.3	3.8	83.3	5.5	86.0	1.4	84.0	4.5
AP	63.0	3.7	57.0	6.1	58.3	7.0	58.0	1.8	54.8	8.6
PUOpp	3037.3	249.9	2757.8	202.6	2826.3	366.6	2833.0	155.4	2789.8	261.2
PSOpp	11654.5	2326.9	8591.3	2429.7	8671.8	1391.9	10027.5	964.1	9531.8	2649.3
SLP	1116.0	33.8	996.3	149.0	1070.5	188.3	1263.5	135.0	1258.3	212.5
PUS	2414.5	155.0	2086.5	86.0	2291.8	274.7	2117.8	179.1	2182.3	144.2
PSS	19705.3	2574.2	15597.0	3609.1	15419.0	3341.0	16941.3	1371.6	15913.5	4387.3
ULP	909.5	247.8	905.8	345.9	858.3	154.6	845.3	62.4	923.5	248.4
P_90	635.4	59.2	515.4	85.8	577.7	90.7	557.1	35.1	533.7	113.7
PAOppH	81.8	4.3	77.9	6.0	77.9	5.6	81.9	1.8	79.3	5.9
PAOwnH	91.7	1.6	91.6	2.1	90.3	4.0	91.3	1.2	90.4	2.5
LPS	55.7	6.4	53.7	8.5	55.4	8.2	59.8	3.5	57.9	10.1

Table A.2.: Mean and standard deviation of the variables for the bottom teams in the "Big Five"

	Premier		Ligue 1		Bundesliga		Serie A		LaLiga	
Variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
SCB	140.3	9.1	123.7	14.8	120.3	4.0	138.3	11.0	102.7	23.8
R	2319.3	48.2	2256.0	33.0	1969.3	74.5	2188.7	27.4	2278.3	24.0
CS	6.7	2.9	8.3	2.1	5.0	1.0	7.0	2.0	7.3	1.5
PC	7.3	0.6	5.7	2.5	4.7	2.1	5.7	2.1	9.3	1.5
I	497.7	56.0	463.7	37.0	403.3	52.8	393.7	68.4	434.0	1.7
SCTI	163.0	17.1	123.0	13.1	144.7	15.5	152.3	7.1	139.7	3.2
SCTO	48.0	5.6	62.0	1.0	54.7	12.9	71.0	8.2	52.0	13.9
TW	387.7	35.2	425.0	15.1	351.0	8.5	352.0	23.0	412.3	42.3
TL	265.7	29.4	264.3	10.0	218.3	20.6	229.7	16.1	245.7	16.0
YC	63.0	7.0	68.0	8.9	59.3	2.5	91.7	17.2	94.7	15.6
Cl	869.0	182.8	788.0	91.3	798.3	39.9	835.7	90.5	863.7	18.6
FC	405.7	27.5	539.3	47.5	396.3	23.6	535.3	68.4	561.0	21.6
TA	59.4	1.7	61.7	1.2	61.7	2.6	60.5	1.1	62.6	2.6
ADW	864.3	115.0	779.7	44.1	673.7	60.6	609.7	146.4	700.7	127.9
ADL	846.7	126.9	858.0	28.0	713.0	25.1	685.7	109.8	734.7	106.2
ADA	50.5	0.6	47.6	0.8	48.5	2.5	46.8	2.3	48.7	1.0
GCIB	64.3	8.3	49.3	4.0	60.7	2.3	58.3	3.8	55.3	10.6
GCOB	11.0	2.6	11.3	3.8	9.0	1.0	13.0	3.6	7.3	2.1
KP	290.7	32.5	265.7	55.8	256.7	27.1	289.7	14.4	293.0	29.5
PT	2.7	1.5	4.7	1.2	3.3	2.5	6.0	1.0	6.0	1.0
SOT	179.7	12.0	176.7	30.4	171.3	17.2	190.0	21.2	192.0	17.1
SB	114.0	2.0	106.7	12.5	99.7	3.5	112.3	18.1	101.7	24.4
CW	164.7	3.8	162.7	33.7	154.7	13.9	161.0	28.0	164.3	25.3
CU	367.0	94.8	392.7	64.3	303.3	83.0	433.7	48.0	450.0	8.7
SA	41.9	1.1	40.1	1.3	39.9	2.1	41.1	5.2	43.7	2.1
A	19.0	5.6	19.3	2.1	20.7	5.0	24.7	10.0	25.0	2.6
ST	130.0	14.4	118.3	23.2	113.3	8.6	132.3	19.3	148.3	2.5
SC	90.3	14.0	92.0	12.5	83.0	36.0	109.7	35.6	118.7	9.3
DS	283.3	63.5	323.7	48.3	256.7	19.8	249.0	73.9	388.7	126.7
FW	367.3	24.1	469.0	28.8	400.0	19.5	484.3	48.4	474.7	45.8
DU	215.0	2.6	246.7	24.8	208.7	47.8	200.7	30.4	270.7	46.8
CrA	20.0	2.0	19.1	2.2	21.0	3.5	19.9	3.5	20.8	1.6
DrA	56.4	5.0	56.6	6.0	55.6	4.1	54.9	5.0	58.3	4.9
Dl_A	49.7	0.8	48.9	1.2	49.6	2.4	48.1	0.5	49.1	0.6
DlL	2181.0	58.6	2328.3	48.2	1916.7	178.2	2050.7	168.4	2278.0	54.0
DlW	2155.3	121.7	2224.3	60.5	1879.7	68.4	1897.7	123.1	2195.7	15.3
GA	9.7	1.8	10.1	1.7	10.4	0.4	10.8	4.1	11.8	0.4
GIB	25.3	7.2	25.3	3.8	25.0	2.6	31.7	15.0	34.7	3.5
GOB	4.7	0.6	4.0	3.6	4.7	1.2	3.3	1.5	5.7	4.0
DFKG	0.3	0.6	0.3	0.6	0.3	0.6	0.3	0.6	1.0	0.0
PA	73.3	8.5	76.6	1.6	76.2	0.6	78.7	2.6	77.1	2.9
AP	43.3	8.1	45.3	2.3	44.3	2.1	44.0	3.6	46.7	2.3
PUOpp	3273.3	261.8	3043.3	167.8	2799.0	201.5	2834.3	113.6	2955.0	239.1
PSOpp	5805.3	1427.8	5373.7	599.2	4517.7	433.9	5744.3	1187.1	5287.7	160.0
SLP	1054.7	165.6	1184.0	127.0	980.0	49.5	1079.7	112.1	1103.0	60.9
PUS	2221.0	197.6	2060.3	140.6	1962.7	87.8	1999.3	68.8	1967.0	126.7
PSS	9876.0	3824.4	9784.3	880.2	9187.7	330.4	10829.7	1548.3	9980.3	1259.5
ULP	1456.0	166.7	1274.7	48.6	1214.7	54.6	1197.7	142.9	1306.3	78.4
P_90	384.4	102.6	376.4	23.5	392.5	13.2	397.5	39.3	377.8	30.6
PAOppH	66.3	7.7	67.4	1.7	64.9	1.7	70.5	4.8	68.4	3.0
PAOwnH	83.7	6.0	87.4	1.5	87.6	0.6	87.3	1.0	86.6	1.6
LPS	42.0	6.3	48.1	1.7	44.7	2.1	47.5	4.5	45.8	1.0

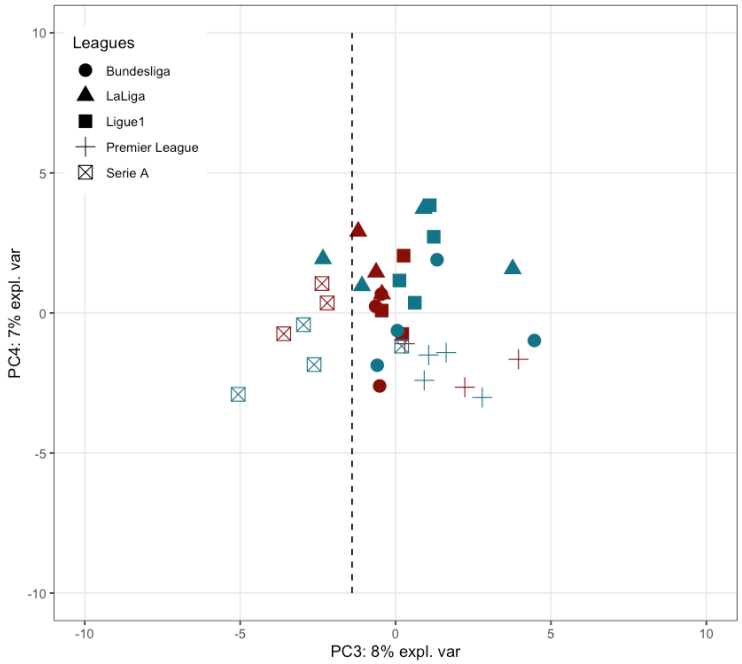


Figure A.3.: PCA scores scatterplot of the teams and leagues projected in the PC3/PC4 space: top teams in blue and bottom teams in red

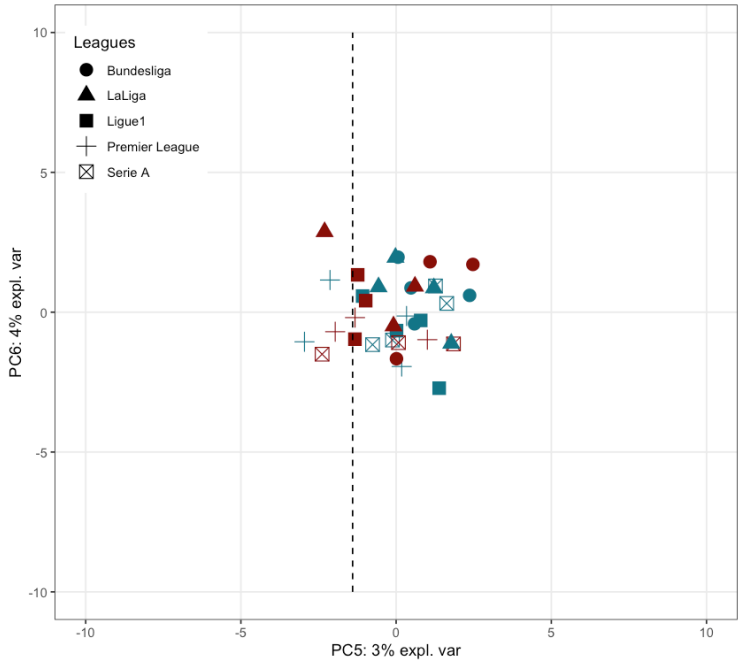


Figure A.4.: PCA scores scatterplot of the teams and leagues projected in the PC5/PC6 space: top teams in blue and bottom teams in red

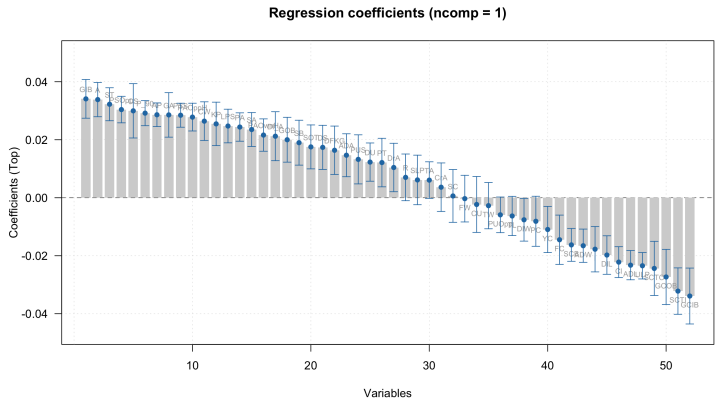


Figure A.5.: PLS-DA regression coefficients with 95% jackknife confidence intervals for the variables to predict the top teams

Chapter 4

Using the Skellam regression model in combination with the Random Forest algorithm to predict match results

Abstract

This chapter investigates the predictive ability of the Skellam Regression Model and compares it with the predictive accuracy of machine learning and multivariate analysis techniques. Data from the "Big Five" during the 2019/2020 and 2020/2021 seasons were used for the analysis. This chapter presents the procedure for building the dataset by web scraping. Furthermore, combining the RF and the Skellam Regression Model is proposed to perform the predictive analysis. Therefore, the RF will be used to select the explanatory variables to be included in the Skellam Regression Model to predict the match outcomes (win, draw or loss). The prediction results of the Skellam Regression Model are then compared with the predictive accuracy of the PLS-DA and the RF. Furthermore, it was analysed whether the predictive accuracy of the methods differed between leagues and seasons. The results showed that RF and PLS-DA have a higher predictive ability than the Skellam Regression Model.

PLS-DA provide the most accurate results. In addition, this study reveals some differences between leagues and seasons.

4.1 Introduction

Compared with other sports, such as American football (Boulier and Stekler 2003; Kuzmits and Adams 2008), baseball (Bennett and Flueck 1983; Lewis 2004; Barnes and Bjarnadóttir 2016) or basketball (Ibáñez et al. 2008; Zimmermann 2016; Zuccolotto, Manisera, and Sandri 2018), football stands out for its intricacy when it comes to being studied (Oberstone 2009). One of the reasons for this complexity is that, given the dynamic nature of football, several game actions often co-occur. On average, 1,682 events happen in the course of a football match (Pappalardo et al. 2019). In addition, match only stops at specific occasions, such as fouls, half-time or in case of substitution (each coach can use three substitution windows at any time during the match, allowing a maximum number of 5 substitutions). These qualities make football a game with countless challenges, especially if the aim is to predict the outcome of matches (Peñas et al. 2010; Lago-Peñas, Lago-Ballesteros, and Rey 2011; Liu et al. 2015a).

The study of the results of football matches is a recurring topic of analysis in which various methodological treatments have been used. On the one hand, several authors have applied discriminant analysis (Peñas et al. 2010; Lago-Peñas, Lago-Ballesteros, and Rey 2011; Liu et al. 2015a) to identify game actions and important contextual variables in discriminating between losing, drawing or winning a match. In addition, several approaches have been used to predict the outcome of a match. Carpita et al. (2015), Carpita, Ciavolino, and Pasca (2019), and Carpita and Golia (2021), used a wide range of machine learning techniques (RF, Neural Networks, Bayesian Network, K-NN and Naive Bayes) and regression models (Multinomial Logistic Regression and Binomial Logistic Regression). Schauberger, Groll, and Tutz (2017) applied the Bradley-Terry model with ordered response categories.

Skellam Regression Model (SRM), a classical model based on the double Poisson distribution, is another approach widely used for prediction purposes (Karlis and Ntzoufras 2009; Carpita, Ciavolino, and Pasca 2021; Pelechrinis and Winston 2021). Therefore, assuming that the outcome of a match can be explained as the difference in goals scored (home team goals minus away team goals), the SRM is used to model the goal difference. However, a disadvantage of the SRM is that, like the rest of the regression models, it does not work with correlated

variables. Thus, it is necessary to perform a variable selection procedure to deal with multicollinearity problems.

This chapter is dedicated to proposing the combined use of the RF algorithm and the SRM to predict the outcome of matches. Thus, the RF will be used to select the most influential variables for predicting the match outcome. Then, the selected variables will be introduced into the SRM to carry out the prediction. Specifically, this chapter has as main objectives: i) to analyse whether there were differences in the performance of the teams during the 2019/2020 and 2020/2021 seasons; ii) to study whether the predictive accuracy of the SRM varies between leagues and; concerning the main objectives of the thesis, iii) to compare the predictive accuracy of the multivariate PLS-DA and RF techniques with the SRM. In addition, the most meaningful game actions will be analysed to discriminate between winning, drawing and losing teams.

This chapter consists of five sections. Section 4.2 is devoted to describing the data collection and cleaning process, and the variable selection procedure. Section 4.3 offers the game actions with the highest contribution to discriminate between win, draw and loss results. Furthermore, this section compares the predictive results to the SRM and the proposed multivariate methods. Sections 4.4 and 4.5 present a discussion and the conclusions of the paper, respectively.

4.2 Material and methods

This section briefly describes the R code programmed for constructing a new dataset, explains the variable selection procedure, and presents the multivariate analysis.

4.2.1 Data collection and cleaning process

To conduct the study, a dataset was constructed with match-to-match performance information from teams competing in the "Big Five" (LaLiga, Premier League, Bundesliga, Ligue 1 and Bundesliga) over the 2019/2020 and 2020/2021 seasons. In this case, the source of the data was *FBref* (fbref.com).

The programmed web scraping[1] code provided a first dataset composed of $n = 7102$ observations and 349 variables. Of the 349 variables, the first seven

[1]"Web scraping is the practice of collecting data through any means other than programme interaction with the API". The data are obtained by programming automated code that queries a web server and then analyses that data to extract the necessary information (Mitchell 2018).

collected general information about the match: date, time, round, day, venue, result and opponent. The remaining 342 variables represented the game actions performed during the match. According to *FBref*, these explanatory variables are grouped into ten types: *standard stats*, *goalkeeping*, *advanced goalkeeping, shooting passing, pass types*, *goal and shot creation*, *defensive actions*, *possession*, *playing time*, and *miscellaneous stats*. In addition to the variables related to team performance, as a novel proposal to control the effect of possible changes in the coaches' strategy and to analyse the impact of unforeseen events on the match outcome (e.g. injuries or red cards), four variables, called "strategy proxy", were incorporated into the analysis: number of red cards (CrdR), the number of substitutions (changes), and the mean and standard deviation of playing time of players who played more than 60 minutes (Mean_60, SD_60).

Once the information for all the teams and matches was in the same data set, the first step was to merge the statistics for the home and away teams (henceforth H and Aw) in the same row. Thus, the game actions of both teams were unified into a single observation. Note that the statistics related to the H team were labelled with the superscript "H" and those of team Aw with "Aw". After this step, a data set of $n = 3551$ observations was obtained.

The second step consisted of extracting all variables related to goals, either directly (e.g. the number of goals from outside the box) or indirectly (e.g. assists). Post-match statistics (e.g. expected[2] goals or assists) were also eliminated. In addition, redundant variables resulting from merging the H and Aw team statistics into a single observation were removed (e.g. the number of fouls committed by team H and the number of fouls received by team Aw). Thus, after the cleaning process, the database had a total of 259 variables: the response variable, calculated as the goal difference between teams H and Aw (positive values indicated a win for team H, negative values a win for team Aw and 0 a draw), and 258 numerical variables collecting information about the game actions and the "strategy proxy" (see Table B.1.).

[2]Metric that calculates the goal probability of each shot on goal.

4.2.2 Variable selection procedure

As indicated in Section 4.1, a drawback of regression models is the need to refine them by eliminating strongly correlated variables. Thus, before conducting the analysis, it was necessary to select the explanatory variables to be introduced into the SRM. Since the stepwise variable selection method is not practicable with the SRM (Carpita et al. 2021; Carpita et al. 2015), the RF algorithm was used to select the variables that contribute most to the match outcome (in our case, goal difference).

RF is an ensemble learning method used in both classification and regression problems. As explained in Sections 2.2.3 and 3.2.3, RF is based on constructing multiple decision trees that are independent of each other and improve prediction. As in the case of classification problems, RF regression provides insight into the importance of variables. In randomForest R-package, Liaw and Wiener (2002) also included the measures proposed by Breiman (2001) to estimate the importance of variables in regression problems: IncMSE% and IncNodePurity. Similar to the MDA for classification, IncMSE% measures the increase in mean squared error (MSE) in OOB when the values of one variable in the training set are permuted, with all others remaining unchanged. Thus, the greater the increase in MSE, the greater the importance of the variable. IncNodePurity measures the increase in purity each time a node is split on a given variable. Thus, a decrease in node purity will imply an increase in MSE. IncNodePurity is calculated for each tree and normalised by the number of trees in which the variable appears. The higher the increase in purity, the higher the importance of the variable.

4.2.3 Skellam Regression Model

To answer questions i) and ii) posed in Section 4.1, after selecting the most important game actions (for each league and season), the SRM was used to predict the match outcome (goal difference).Note that according to Ley, Wiele, and Eetvelde (2017), who compared different classical prediction models, SRM is the best model for predicting league rankings.

Skellam Regression Model fitting

Considering that the match result can be summarised as the difference between the number of goals scored by the H team (H_G) and the number of goals scored by the Aw team (Aw_G), SRM is used to model the effect of some covariates (explanatory variables) on two statistically independent discrete random variables, H_G and Aw_G, each Poisson-distributed with expected values λ_H and λ_{Aw}, respectively, and their difference $Z = H_G$ - Aw_G (Skellam 1946). SRM follows the probability function (Skellam 1946):

$$P(z) = Prob(Z = z) = e^{\lambda_H + \lambda_{Aw}} \left(\frac{\lambda_H}{\lambda_{Aw}}\right)^{z/2} I_z\left(2\sqrt{\lambda_H \lambda_{Aw}}\right) \tag{4.1}$$

where $I_z(x)$ is the modified Bessel function. Note that, Equation 4.1 follows the exact distribution of the difference of two independent Poisson variables (Skellam 1946).

Therefore, we will assume the difference between H_G and Aw_G as the match outcome $Z = H_G - Aw_G$. Then, it will be considered a win H if Z is positive, a win Aw (or loss H) if Z is negative, and a draw if $Z = 0$.

Hence, following Karlis and Ntzoufras (2009), the SRM is

$$Z \sim Skellam(\lambda_H, \lambda_{Aw}) \tag{4.2}$$

$$log(\lambda_H) = \mu_H + \boldsymbol{x_H}\boldsymbol{\beta}_H \tag{4.3}$$

$$log(\lambda_{Aw}) = \mu_{Aw} + \boldsymbol{x_{Aw}}\boldsymbol{\beta}_{Aw} \tag{4.4}$$

where Equation 4.3 and Equation 4.4 refer to the logarithm of expected H_G and Aw_G, respectively; $\boldsymbol{\beta}_H$ and $\boldsymbol{\beta}_{Aw}$ are the vectors of regression coefficients of the explanatory variables (game actions) for H and Aw teams, respectively; and μ_H and μ_{Aw} are constant parameters that include the H and Aw effects of a match (i.e. $\mu_H > \mu_{Aw}$).

To answer question i) formulated in Section 4.1, the season effects for H, γ_H, and Aw teams, γ_{Aw}, have been included in Equation 4.3 and Equation 4.4. Thus, we set

$$log(\lambda_H) = \mu_H + s\gamma_H + \boldsymbol{x_H}\boldsymbol{\beta}_H \tag{4.5}$$

$$log(\lambda_{Aw}) = \mu_{Aw} + s\gamma_{Aw} + \boldsymbol{x_{Aw}}\boldsymbol{\beta}_{Aw} \tag{4.6}$$

where $s = 0$ for matches played during the 2019-2020 season, and $s = 1$ for matches played during the 2020-2021 season.

Following the procedure carried out by Carpita et al. (2015), the bootstrap approach was used to estimate the regression coefficient and the statistical significance (p-value) of the explanatory variables, including the seasonal effect. Thus, for every 1000 bootstrap replicates, the data set was randomly divided into a training set (75% of the data) and a test set (25% of the data). The regression coefficients and p-values obtained for each model replicate were then averaged.

Evaluation of the Skellam Regression Model

The second objective of this chapter was to assess the ability of the SRM to predict the match result (win, draw or loss) of the H and Aw teams, and to test whether it differed between leagues. Since in the SRM, the response variable is the match outcome measured as the goal difference (Z), the estimation of the probability P of a win ($Home_W$), loss ($Home_L$) or draw ($Home_D$) of the H team was calculated for both the predicted and observed goal difference. Thus, the threshold used to classify the matches in the three categories was: $P[Home_W] = P[Z \geq 0.5]$, $P[Home_L] = P[Z \leq -0.5]$, and $P[Home_D] = P[-0.5 < Z < 0.5]$. Then, to evaluate the performance of the SRM, the sensitivity, specificity and the MCC (see Section 2.3) of the model were calculated through the confusion matrix to measure the accuracy of the prediction.

- *Sensitivity* measures how effectively the model identifies (in the test set) true positives, e.g., how well the model predicts a win H when team H wins.

$$Sensitivity = \frac{TP}{TP + FP} x1000 \tag{4.7}$$

- *Specificity* measures how effectively the model identifies (in the test set) true negatives, e.g., how well the model predicts a non-win H (draw or loss) when team H does not win.

$$Specificity = \frac{TN}{TN + FP} x1000 \tag{4.8}$$

In this case, for each league and season, 100 bootstrap replicates were used (Carpita et al. 2015), each randomly splitting the dataset into 75% (training set) and 25% (test set). Then, in each bootstrap replication, the confusion matrix was calculated, and the values of the sensitivity, specificity and MCC of the SRM were computed and stored. Note that the probability of P of a win, loss or draw was calculated for both the predicted and observed goal difference in each loop.

4.2.4 Multivariate statistical analysis

As indicated in Chapter 3, one of the main objectives of this thesis is to show the advantages of using multivariate techniques. Therefore, the ultimate aim of this chapter is to test whether the PLS-DA and RF (see Sections 2.2.5 and 3.2.3) have a higher predictive accuracy than the SRM. Note that instead of using the variables H and Aw independently, the difference of the game actions were used as explanatory variables (except in the case of the "strategy proxies": Mean_60, SD_60, and changes). Furthermore, instead of using the goal difference as the response variable, the matches were labelled according to the result of the H team (Win, Draw, and Loss).

To assess the performance of the PLS-DA and RF, the same approach applied to the evaluation of the SRM was used (see Section 4.2.3). Then, the comparison of the performance of the models was carried out using the MCC. Two-way ANOVA was used to check if the models had statistically significant differences, indicating the test set as the blocking factor and the model as the main factor. Note that in order to conduct a correct model comparison, the same test and training sets were used in each repetition.

4.3 Results

This section presents the results obtained by selecting explanatory variables through the RF, using goal difference as the dependent variable. Then, after selecting the game actions to be introduced in the SRM, questions i) and ii) (see Section 4.1) will be answered by fitting and estimating the SRM. In addition, the performance of the SRM will be compared with the proposed multivariate statistical methods (PLS-DA and RF), in this case, using the match results (win, draw and loss) as the dependent variable. Finally, the most important variables for discriminating between win, draw and loss will be analysed.

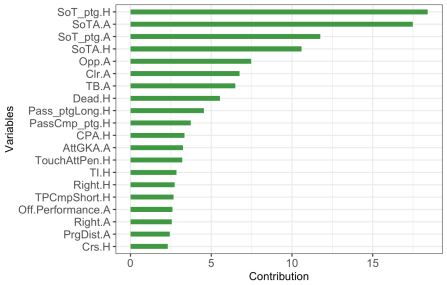
4.3.1 Variable selection procedure

Due to the nature of our data, we cautiously analyse the variables' selection in the RF algorithm. As was explained and shown in Chapter 3, the Gini index tends to benefit variables with many missing values or categorical cut-off points (Section 3.3.2). This fact has also been explained and verified by previous researchers (Kim and Loh 2001; Strobl et al. 2007; Sandri and Zuccolotto 2010). Therefore, to perform a more reliable variable selection procedure, we used the *impurity_ corrected* argument included in the ranger function of the ranger R-package (Wright and Ziegler 2017), which is based on a modified version of the method suggested by Sandri and Zuccolotto (2010). Figure 4.1. shows the average contribution of the 20 most important game actions for 100 repetitions after applying the *impurity_ corrected* importance measure in each league for both seasons.

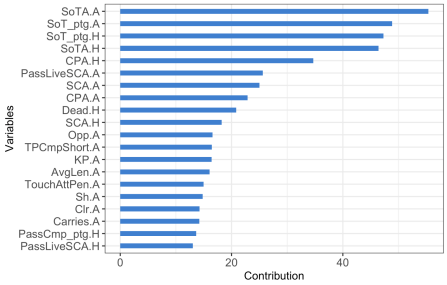
According to the results in Figure 4.1., each league's top ten game actions were selected for inclusion in the SRM. Note that if any of the selected variables correlated higher than 0.7, the one with the lowest correlation with the dependent variable (goal difference) was removed from the model. In addition to examining the correlation, the vif_function (Beck 2013) was used to calculate the VIF value of all explanatory variables and to ensure that in any case it was greater than 7 (Neter, Wasserman, and Kutner 1989; Allison 1999).

Thus, at the end of the process nine game actions of H teams and eleven actions of Aw teams were selected as the most important explanatory variables for predicting the match outcome (measured as goal difference) in the "Big Five". Table 4.1. shows the variables' name (*Variable*); the type of game action (*Type*), according to *Fbref*; the leagues in which the variable was important (*Leagues*); the teams for which this action contributed to the match outcome (*Teams*); and the variables' description (*Description*), also provided by *Fbref*.

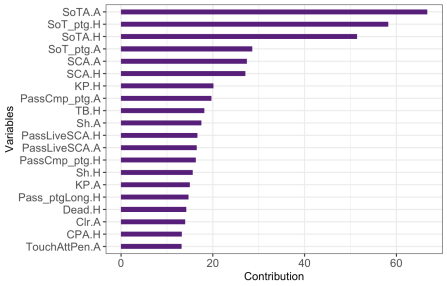
In addition to the variables selected by the RF, the "strategy proxies" for teams H and Aw were included in the SRM (see Section 4.2.1). Due to the high correlation between Mean_60, SD_60, only Mean_60 was included in the SRM.



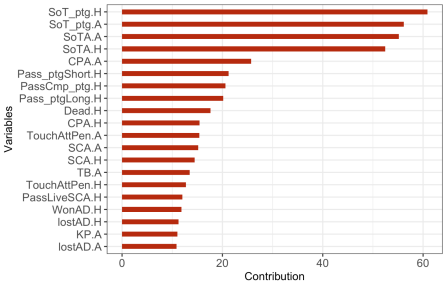
(a) La Liga.



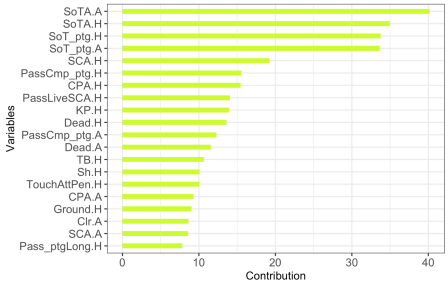
(b) Serie A.



(c) Premier League.



(d) Bundesliga.



(e) Ligue 1.

Figure 4.1.: Twenty most important explanatory variables in each league, according to the RF, for predicting goal difference (Z) - Seasons 2019/2020 and 2020/2021

Table 4.1.: Most influential explanatory variables to predict the goal difference (Z) and the corresponding league and team they belong to, according to the RF, after discarding variables with a correlation higher than 0.7 in each league for both seasons

Variable	Type	League	Teams	Description
SoTA	Goalkeeper	LaLiga Serie A Premier Bundesliga Ligue 1	H/Aw H/Aw H/Aw H/Aw H/Aw	Shots on target against
SoT_ptg	Shooting	LaLiga Serie A Premier Bundesliga Ligue 1	H/Aw H/Aw H/Aw H/Aw H/Aw	Shots on target percentage
Dead	Pass Types	LaLiga Serie A Premier Bundesliga Ligue 1	H H H H H/Aw	Dead-ball passes (Includes free kicks, corner kicks, throw-ins, and goal kicks)
PassCmp_ptg	Passing	Serie A Premier Ligue 1	H H Aw	Pass completion percentage
SCA	Goal and Shot Creation	Premier Bundesliga Ligue 1	H/Aw H H	Shot-Creating actions (Actions leading directly to a shot (passes, dribbles ,drawing fouls, etc.))
CPA	Possession	Serie A Bundesliga Ligue 1	H/Aw H/Aw H	Number of times a player controlled the ball inside the opponent's penalty area
PassLiveSCA	Goal and Shot Creation	Serie A	H/Aw	Completed live-ball passes that lead to a shot attempt
Pass_ptgLong	Passing	LaLiga Ligue 1	H H	Passes longer than 30 yards
Pass_ptgShort	Passing	Bundesliga	H	Passes between 5 and 15 yards
Clr	Defensive Actions	LaLiga Premier	Aw Aw	Clearence
TB	Pass Types	LaLiga	Aw	Completed passes sent between back defenders into open space
Opp	Goalkeeper	LaLiga	Aw	Opponent's attempted crosses into the own penalty area
TouchAttPen	Possession	Bundesliga	Aw	Touches in the opponent's penalty area

Sample size (n) and number of explanatory variables (k): LaLiga, n= 760 k=9; Serie A, n= 760 k=10; Premier, n= 760 k=9; Bundesliga, n= 612 k=10; Ligue 1, n=659 k=10.

4.3.2 Skellam Regression Model fitting

Firstly, the explanatory variables were standardised to answer question i) and to compare the teams' performance depending on the league and season. Additionally, the same regressors were introduced in all the models to make a more precise and forthright comparison between the estimated coefficients.

Table 4.2. shows the regression coefficients of the variables selected by the RF (Table 4.1.) and introduced into the SRM after discarding the variables with a correlation greater than 0.7 (Equation 4.5 and Equation 4.6). Note that the variables are highlighted with superscripts to differentiate the teams' performance (H and Aw). Thus, a negative coefficient on an explanatory variable indicates a negative relationship with the number of goals scored by this team and a positive correlation of this same variable with the number of goals scored by the opposing team, and vice versa (Pelechrinis and Winston 2021).

As for the results in Table 4.2., except for the Bundesliga, the H effect is influential in all leagues, i.e. the probability of scoring a goal decreases with statistical significance in the case of being an Aw team.

The only statistically significant variables for all leagues and teams are the number of shots on goal received (SoTA) and the accuracy of shots on goal (SoT_ptg). In the case of both SoTA and SoT_ptg, these game actions have a greater impact (negative in the case of SoTA and positive in the case of SoT_ptg) on the number of goals scored by Aw teams. There is an exception for the SoT_ptg variable in the case of Premier League, where the accuracy of shots on goal has a higher effect on the likelihood of scoring for H teams. Another exception occurs for SoT_ptg variable in the case of Series A (the positive effect is slightly higher for H teams).

In the case of the number of dead balls passes (Dead), this variable negatively affects both H and Aw. However, Table 4.2. shows a higher negative effect on the likelihood of scoring for H teams (except in the case of the Premier League). In addition, the Aw teams of LaLiga are the only ones in which the effect of this variable is not statistically significant. To interpret the impact of the variable Dead into the goals difference, it was necessary to analyse it more deeply. As it was shown in Table 4.1., the description of the variable Dead (provided by Fbref) highlights that it is not an isolated variables but stores information about four variables: free kick passes (crosses are not considered), throw-ins, goal kick (goalkeeper pass), and corners. Since, except for corner kicks, the rest of the game actions are not self-explanatory, i.e. the direct relationship of free kicks, throw-ins, and goal kicks on the probability of scoring

Table 4.2.: Regression coefficients and statistical significance of the most influential explanatory variables of the fitted SRM after discarding variables with a correlation higher than 0.7 - Seasons 2019/2020 and 2020/2021

Explanatory variables	**LaLiga**	**Serie A**	**Premier**	**Bundesliga**	**Ligue 1**
$SoTA^H$	-0.173***	-0.237***	-0.222***	-0.206***	-0.196***
$SoTA^A$	-0.334***	-0.251***	-0.302***	-0.279***	-0.345***
SoT_ptg^H	0.342***	0.397***	0.446***	0.375***	0.335***
SoT_ptg^A	0.450***	0.357***	0.402***	0.425***	0.367***
$Dead^H$	-0.227***	-0.268***	-0.232***	-0.228**	-0.294***
$Dead^A$	-0.106	-0.219***	-0.237***	-0.128**	-0.253***
SCA^H	0.106	0.153**	0.268***	0.132**	0.152**
SCA^A	0.117	0.182**	0.355***	0.144	0.204**
CPA^H	0.117*	0.142**	0.062	0.164***	0.126**
CPA^A	0.178**	0.180**	0.067	0.118*	0.085
$Pass\_ptgLong^H$	0.053	-0.053	0.009	-0.029	-0.057
$Pass\_ptgShort^H$	-0.054	0.047	-0.066	0.124*	0.053
Clr^A	0.101	0.085	0.087	-0.031	-0.048
TB^A	0.160***	0.020	0.073	0.114**	0.077
Opp^A	0.273***	0.097	0.123*	0.094	0.138*
$TouchAttPen^A$	-0.017	0.000	-0.010	0.098	0.014
$Intercept^H$	-0.063	-0.024	-0.053	0.177	0.013
$Intercept^A$	-0.257**	-0.331***	-0.392***	0.020	-0.576***
$S\_2020/2021^H$	-0.051	-0.183	-0.105	-0.381**	-0.299*
$S\_2020/2021^A$	-0.080	-0.083	0.137	-0.524***	0.190
$CrdR^H$	-0.151**	-0.220***	-0.011	-0.013	-0.156*
$CrdR^A$	0.005	-0.086	-0.107	-0.019	-0.224**
$Changes^H$	0.047	0.081	0.108**	0.178**	0.086
$Changes^A$	0.039	0.017	0.059	0.190**	0.055
$Mean\_60^H$	-0.023	0.012	0.081	-0.047	-0.094
$Mean\_60^A$	-0.038	-0.066	0.044	-0.024	-0.020

* Statistical significance (p-values): ***p <0.01,**p <0.05, *p <0.1. Note: explanatory variables are standardized.

a goal does not seem evident, a correlation analysis was performed to study the game actions preceding them. Thus, it was found that there is a high and positive correlation (above 0.6) between free-kick passes and goal kicks with fouls received and shots off target against, respectively. At the same time, both variables (fouls received and shots off target against) are negatively correlated with the probability of scoring a goal. Regarding the variable throw-ins is the opposite of out of bounds, and both are positively correlated (in this case, 0.41, probably because other variables are involved in the throw-ins action). This fact explains why throw-ins are negatively correlated with the probability of scoring goals, given that out of bounds are positively correlated. Note that the only variable positively correlated with the response variable is the number of corners.

Even though the RF identified, in all leagues, variables related to *passing* as important for predicting the outcome of matches (see Table 4.1.), only the variable Pass_ptgShort had a positive and statistically significant effect on the goal probability of the H teams in the Bundesliga.

On the other hand, the RF model only identified SCA (shot creation actions) and PassLiveSCA (live ball passes leading to a shot attempt) as important variables related to the *goal and shot creation*. Note that the variable PassLiveSCA was not included in the SRM (Table 4.2.) due to the high correlation between the two variables. Regarding the results shown in table Table 4.2., the strong impact of the SCA variable on the probability of scoring a goal in the Premier League stands out for both teams H and Aw. Note that, in the case of LaLiga, the non-statistical significance of the SCA variable in the SRM aligns with the results of the RF (Table 4.1.), which did not point to either SCA or PassLiveSCA as important variables.

According to Table 4.2., CPA (number of times a player controlled the ball inside the opponent's penalty area) has a positive and statistically significant effect on the likelihood of scoring a goal in all leagues and teams except the Premier League (both H and Aw) and the Aw teams of Ligue 1. These results are consistent with those presented by the RF (see Table 4.1.), as the RF did not select CPA as an influential variable in the case of the Premier League and Ligue 1. However, in the case of LaLiga, although the RF did not select CPA either, it was statistically significant in the SRM (both H and Aw).

Regarding Table 4.1., RF identified four important game actions for Aw teams: clearances (Clr), passes sent between the defenders' free space (TB), attempted crosses (Opp) and touches in the opposing penalty area (TouchAttPen). Regarding the results of the SRM (Table 4.2.), the increase in the number of TB

and Opp significantly increases the effect on the probability of scoring a goal. However, no statistically significant effect was found in any league in the case of Clr and TouchAttPen (although RF chose these variables (Table 4.1.)).

In the case of the "strategy proxy" game actions (Mean_60, CrdR, and changes), Mean_60 was the only variable that was not statistically significant for any league. CrdR showed a negative and statistically significant effect on H teams in LaLiga, Serie A and Ligue 1 and Aw teams in Ligue 1. On the contrary, the number of changes had a positive and statistically significant effect on the H teams from the Premier and H and Aw teams in the Bundesliga.

Table 4.2. shows that the H teams competing in Ligue 1 and the H and Aw teams competing in the Bundesliga performed worse in the 2020/2021 season than in the 2019/2020 season, as the season effect has a negative and statistically significant coefficient.

Table 4.3. shows the goodness of fit calculated through the Root Mean Square Error (RMSE), LogLikelihood, AIC and Bayesian Information Criterion (BIC).

Table 4.3.: Goodness-of-fit statistics of the SRM for the "Big Five" - Seasons 2019/2020 and 2020/2021

League	RMSE	LogLikelihood	BIC	AIC	R^2%
LaLiga	1.281	-904	1973	1810	36.2
Serie A	1.371	-957	2079	1916	44.0
Premier	1.412	-962	2090	1927	44.7
Bundesliga	1.465	-790	1740	1583	49.4
Ligue 1	1.297	-800	1760	1601	44.5

According to Table 4.3., the Bundesliga and Ligue 1 are the leagues in which the SRM best fits the data. (Bundesliga, BIC=1740 and AIC=1583 and Ligue 1, BIC=1760 and AIC=1601). However, LaLiga has the lowest RMSE (1.281).

4.3.3 Evaluation of the Skellam Regression Model fitting

Tables 4.4. and 4.5. show the average prediction statistics (sensitivity, specificity, and MCC) calculated from the confusion matrix results in the 100 replicates for each league and season (see Section 4.2.3).

Tables 4.4. and 4.5. show that the SRM has a high specificity for predicting losing teams (i.e. the model predicts a non-losing H (draw or win) when team H does not lose), although at the expense of a lower sensitivity (effectiveness

Table 4.4.: Sensitivity, Specificity, and MCC (Means and 95% Centred Intervals) of the SRM for the "Big Five" (75% training set and 25% testing set, 100 replications) - Season 2019/2020

Result	Index	LaLiga	Serie A	Premier	Bundesliga	Ligue 1
Win	Sensitivity	0.718 (0.705 - 0.731)	0.643 (0.629 - 0.656)	0.660 (0.647 - 0.674)	0.647 (0.630 - 0.663)	0.722 (0.710 - 0.737)
	Specificity	0.778 (0.767 - 0.790)	0.868 (0.860 - 0.876)	0.741 (0.730 - 0.753)	0.870 (0.861 - 0.881)	0.772 (0.760 - 0.784)
Draw	Sensitivity	0.484 (0.467 - 0.502)	0.558 (0.539 - 0.577)	0.487 (0.468 - 0.505)	0.593 (0.572 - 0.616)	0.481 (0.457 - 0.504)
	Specificity	0.631 (0.618 - 0.643)	0.651 (0.640 - 0.611)	0.676 (0.665 - 0.686)	0.688 (0.677 - 0.699)	0.667 (0.656 - 0.679)
Loss	Sensitivity	0.376 (0.356 - 0.397)	0.531 (0.516 - 0.546)	0.467 (0.447 - 0.488)	0.602 (0.584 - 0.620)	0.429 (0.407 - 0.452)
	Specificity	0.934 (0.927 - 0.940)	0.895 (0.887 - 0.903)	0.919 (0.913 - 0.925)	0.896 (0.888 - 0.905)	0.933 (0.927 - 0.938)
	MCC	0.346 (0.332 - 0.361)	0.374 (0.360 - 0.388)	0.336 (0.323 - 0.350)	0.423 (0.408 - 0.438)	0.373 (0.358 - 0.387)

Table 4.5.: Sensitivity, Specificity, and MCC (Means and 95% Centred Intervals) of the SRM for the "Big Five" (75% training set and 25% testing set, 100 replications) - Season 2020/2021

Result	Index	LaLiga	Serie A	Premier	Bundesliga	Ligue 1
Win	Sensitivity	0.590 (0.575 - 0.605)	0.617 (0.602 - 0.632)	0.614 (0.596 - 0.631)	0.631 (0.617 - 0.646)	0.581 (0.564 - 0.597)
	Specificity	0.830 (0.820 - 0.840)	0.824 (0.813 - 0.834)	0.851 (0.842 - 0.860)	0.806 (0.794 - 0.818)	0.875 (0.867 - 0.883)
Draw	Sensitivity	0.543 (0.527 - 0.561)	0.575 (0.557 - 0.593)	0.401 (0.380 - 0.422)	0.521 (0.498 - 0.545)	0.586 (0.566 - 0.606)
	Specificity	0.591 (0.579 - 0.602)	0.661 (0.651 - 0.670)	0.698 (0.689 - 0.708)	0.661 (0.650 - 0.673)	0.642 (0.633 - 0.651)
Loss	Sensitivity	0.418 (0.400 - 0.437)	0.494 (0.477 - 0.512)	0.620 (0.605 - 0.636)	0.514 (0.495 - 0.533)	0.571 (0.556 - 0.586)
	Specificity	0.934 (0.927 - 0.940)	0.895 (0.887 - 0.903)	0.919 (0.913 - 0.925)	0.896 (0.888 - 0.905)	0.933 (0.927 - 0.938)
	MCC	0.288 (0.274 - 0.302)	0.350 (0.334 - 0.360)	0.356 (0.342 - 0.370)	0.346 (0.330 - 0.361)	0.366 (0.354 - 0.378)

in predicting a losing H when team H loses). In the case of predicting the winning outcome, although, as with loss predictions, the specificity is higher than the *sensitivity,* these measures are slightly more balanced. In addition, the sensitivity for predicting the winning outcome is higher than for predicting the losing outcome, although the opposite is true for specificity. As for the results of the MCC (Tables 4.4. and 4.5.), although in all cases, the value is higher than 0, the results are not outstanding. Note that the MCC takes values between -1 and 1, so the value 0 indicates that the model does not predict better than chance (see Section 2.2.5).

The SRM performance was then calculated using the MCC for each league and season and tested for statistically significant differences using a two-way ANOVA. The results of the two-way ANOVA revealed statistically significant differences in the predictive accuracy of the SRM as a function of the league analysed in both the 2019/2020 season (p-value =< 2e-16) and the 2020/2021 season (p-value = 9.7e-15). Fisher's 95% post hoc LSD interval test available in the agricolae R-package (de Mendiburu 2021) was used to test between which leagues difference of predictive accuracy was statistically significant.

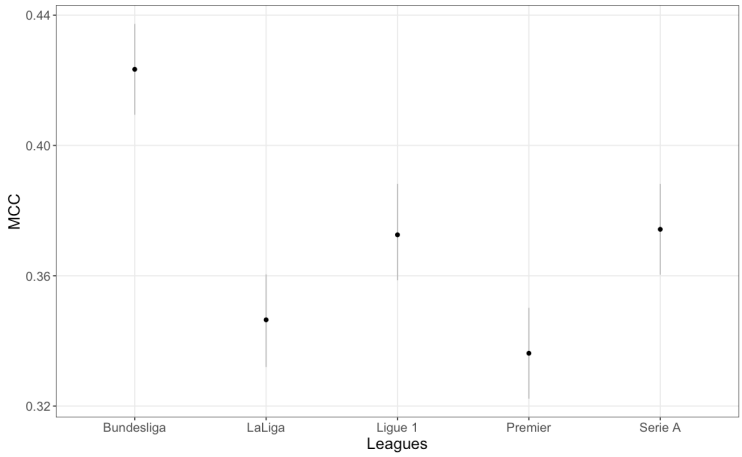


Figure 4.2.: Multiple comparisons of the leagues (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each league, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2019/2020

According to Figure 4.2., the average MCC of the Premier League and La Liga is not statistically significantly different, nor is it between Ligue 1 and Serie A.

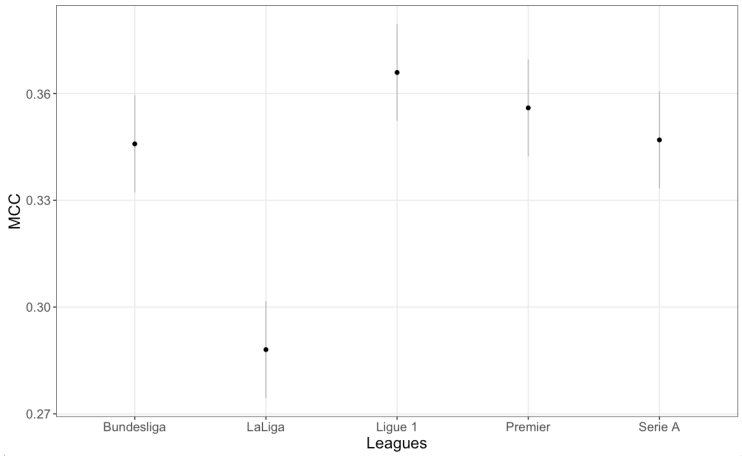


Figure 4.3.: Multiple comparisons of the leagues (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each league, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2020/2021

On the other hand, it is possible to conclude that the average MCC of the Bundesliga in the season 2019-2020 is statistically significant higher than in the rest of the leagues. In addition, Figure 4.3. shows that the average MCC of LaLiga in the 2020-2021 season is statistically significantly lower than the average MCC of the other leagues. Finally, Student's t-test was used to test whether there were differences in SRM prediction performance by season. The results showed that, except for Ligue 1 (p-value=0.477), in all other cases, the prediction results for the 2019/2020 season were significantly more accurate than those for the 2020/2021 season. These results agree with those shown in Figures 4.2. and 4.3. since the average value of MCC is higher in the 2019/2020 (Figure 4.2.) season than in the 2020/2021 season (Figure 4.3.) in all leagues.

4.3.4 Multivariate statistical analysis

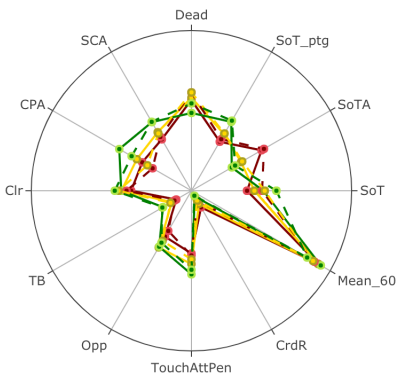
As indicated in Section 4.2.4, PLS-DA and RF were selected to show the advantages of using multivariate techniques and to study their predictive accuracy compared to SRM. In the case of PLS-DA, the first step was applying the vip function of the mixOmics R-package (Rohart et al. 2017) to select the most important variables. PLS-DA tends to give suboptimal results when it has regressors that are not statistically significant. Therefore, as we had a dataset

with a large number of non-statistically significant exploratory variables those variables whose 95% jackknife confidence intervals do not contain the value zero, or which have a high VIP where deleted from the model.

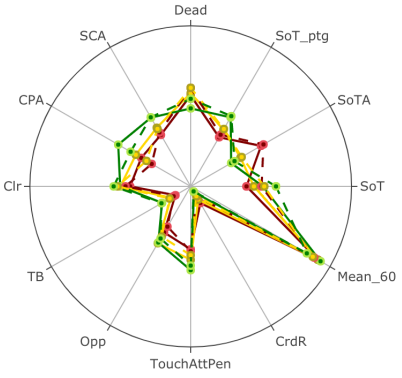
Additionally, to continue the study of the game actions that most contribute to the success or failure of football teams, the variables statistically significant (p-value<0.05) in the RF were also calculated using the randomForestExplainer R-package (Paluszyńska 2017). Thus, Tables B.2. and B.3. show the variables whose mean p-value across the 100 RF replicates was less than 0.05 and those selected by the VIP in the PLS-DA according to league and season.

To analyse the variables that contribute most to discriminating between winning, drawing and losing teams, Figure 4.4. shows the radar plots (Section 2.2.6) with the top ten variables selected by the VIP and statistically significant for RF in most leagues and seasons (see bold in appendix Tables B.2. and B.3.) and the top two selected "proxy variables" (see italics in appendix Tables B.2. and B.3.). Thus, Figure 4.4. shows five radar plots according to the league and the average number of actions performed by the teams during the 2019/2020 (solid line) and 2020/2021 (dashed line) seasons as a function of the outcome of the match: win (green), loss (red) and draw (yellow). Note that the values of the variables are scaled to the radii's length and plotted.

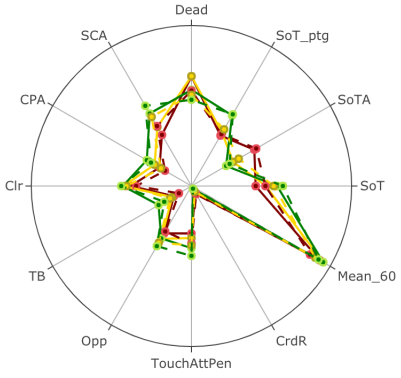
Figure 4.4. highlights that the winning teams carry out a higher number of actions related to shots (SoT, SoT_ptg and SCA). On the contrary, loser teams receive a higher number of shots on target (SoTA). It is highlighted that the VIP and RF select two variables related to controlling the ball inside the opposing penalty area (TouchAttPen and CPA). In both variables, the average of the winning teams is higher. In the case of defensive and goalkeeping actions (Opp and Clr), the winning teams performed a higher number of these actions, except in the case of Ligue 1 during the 2020/2021 season (Figure 4.6(e)), where the average Clr is higher for the draw teams. In the case of TB and Dead, the winning teams take higher values for the former and lower values for the Dead variable. It is noticeable that the teams that draw are the ones that perform a higher average number of dead-ball passes (Dead). In terms of the "strategy proxies" variables, the mean playing time for players who stay on the pitch for more than 60 minutes (Mean_60) is higher for winning teams, except in the case of Bundesliga and Ligue 1 in season 2020/2021 (Figures 4.6(d) and 4.6(e)). The average of red cards (CrdR) in the losing teams is higher. Note that, with a few exceptions, draw teams' averages in these variables are generally between the winner and loser teams.



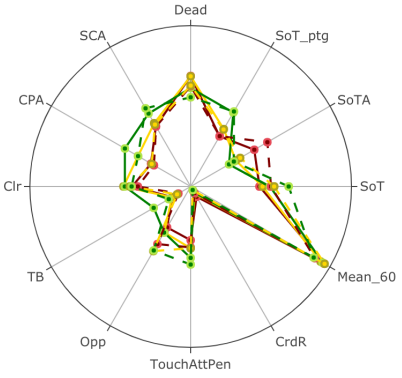
(a) LaLiga



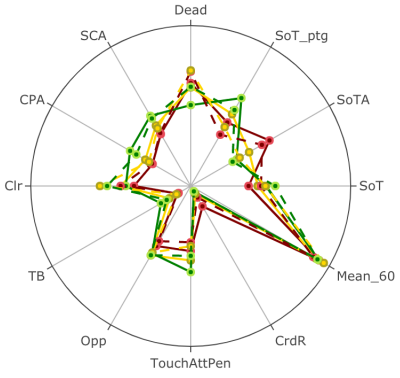
(b) Serie A



(c) Premier League



(d) Bundesliga



(e) Ligue 1

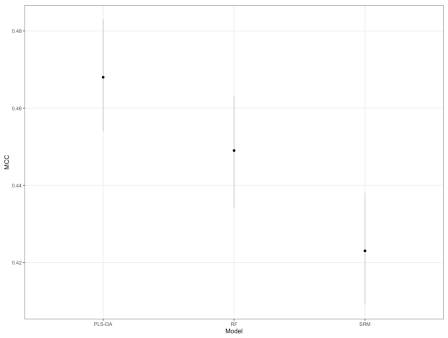
Figure 4.4.: Radar chart to compare the mean values of the main variables selected by PLS-DA and RF differentiating by season (2019/2020 (solid line) and 2020/2021 (dashed line)) and match result: win (green), loss (red) and draw (yellow)

After completing the prediction of all models (SRM, RF and PLS-DA), the mean model performances were calculated from the MCC and tested for statistically significant differences using the two-way ANOVA. Two-way ANOVA was conducted to compare if there were statistically significant differences between the models according to the league and the season studied. The results of the two-way ANOVA indicated that the model factor was statistically significant with a p-value less than 2e-16, except in the case of the Bundesliga in the 2019/2020 season (p-value=9.75e-5) and Serie A in the 2020/2021 season (p-value=3.1e-15). To inspect which models differed statistically from each other, Fisher's 95% post hoc LSD interval test, implemented in the agricolae R-package (de Mendiburu 2021) was performed.

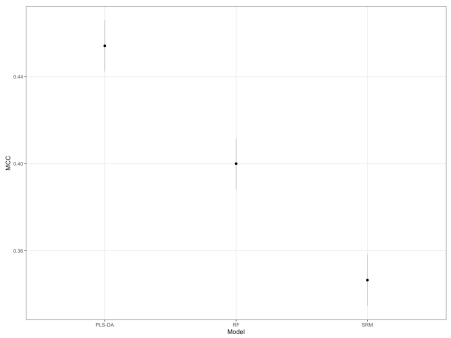
From Figures 4.5. and 4.6., it is possible to conclude that the average MCC is statistically higher in the PLS-DA and RF models than in the SRM (except in the case of LaLiga in the 2020/2021 season, where there was no statistically significant difference in the average MCC between RF and SRM). Furthermore, according to Figures 4.5. and 4.6., except in the case of LaLiga (2019/2020 season) and the Bundesliga (2020/2021 season), statistical differences were found between PLD-DA and RF, it is possible to conclude that PLS-DA has the highest mean MCC. For a more detailed analysis of the MCC results of the 100 replicates, a violin plot was also plotted in combination with the box plot obtained using the ggplot2 R-package (Wickham 2016) to reflect the distribution of MCC according to leagues and seasons studied (see Figures B.1. and B.2.). After the above analysis, it was concluded that the performance of the multivariate techniques (PLS-DA and RF) was better than that of the SRM.

4.4 Discussion

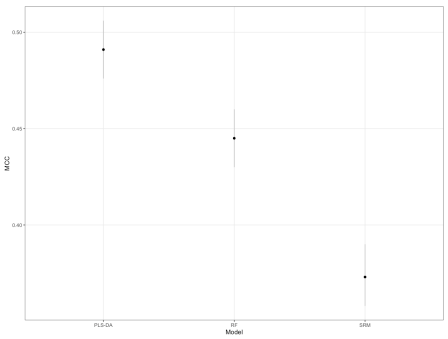
This chapter aims to use the SRM, a classical model based on the double Poisson distribution, RF algorithm and PLS-DA model to predict the results of football matches played in the Spanish, Italian, German, English and French leagues during the 2019/2020 and 2020/2021 seasons. The prediction performance of the SRM with PLS-DA and RF is then compared. To our knowledge, this study comprises the first analysis conducted to compare the evaluation of the SRM with machine learning and multivariate statistical techniques, specifically PLS-DA and RF, with match statistics from the top five leagues in the world. Notably, the three main tasks performed in this chapter were (i) to study whether there was a difference in the performance of the "Big Five" teams between seasons, (ii) to discover the league in which the SRM has higher



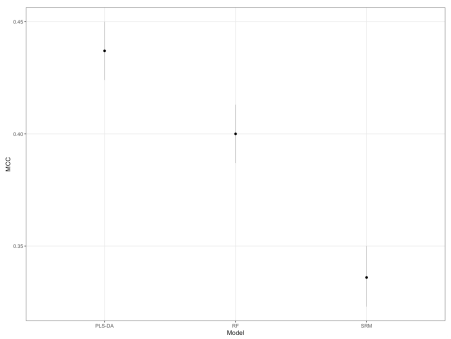
(a) LaLiga



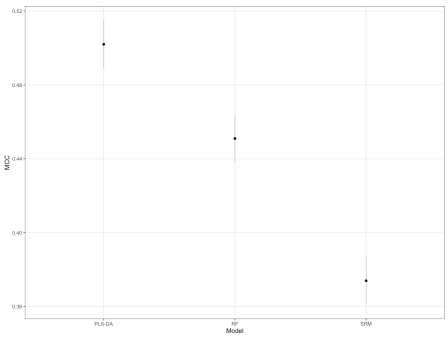
(b) Serie A



(c) Premier League

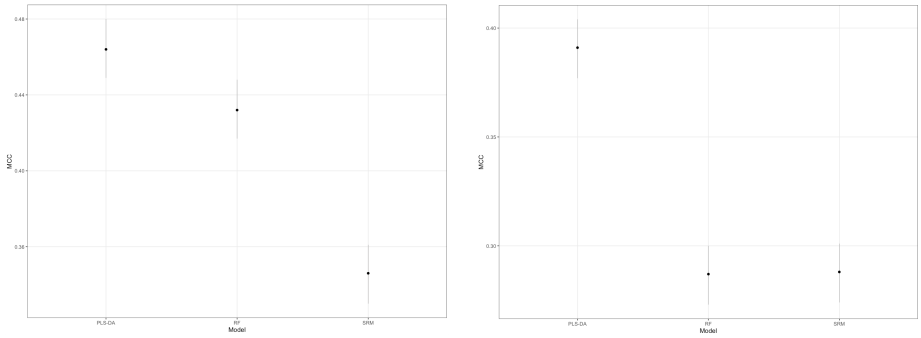


(d) Bundesliga



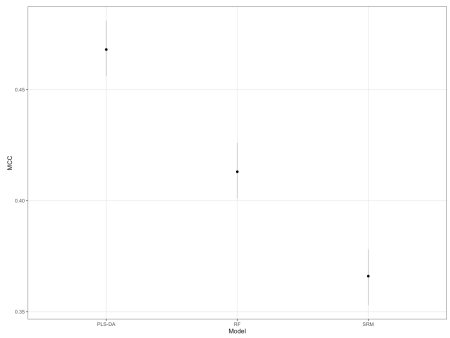
(e) Ligue 1

Figure 4.5.: Multiple comparisons of the models (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2019/2020

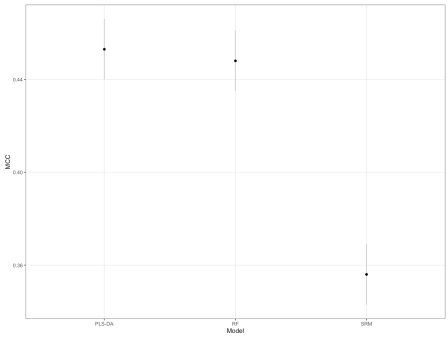


(a) La Liga

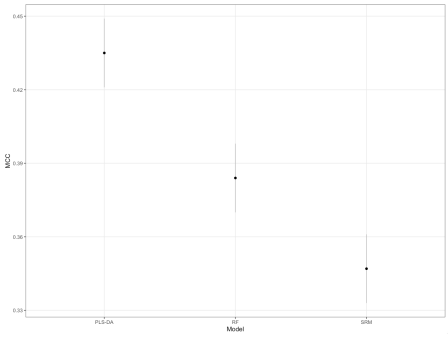
(b) Serie A



(c) Premier League



(d) Bundesliga



(e) Ligue 1

Figure 4.6.: Multiple comparisons of the models (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2020/2021

accuracy, (iii) to compare the predictive capability of the SRM with machine learning and multivariate statistical techniques. In addition, the most important or statistically significant game actions were studied to discriminate between winning, drawing and losing teams.

In the first step, refining the SRM to eliminate highly correlated variables was necessary. Variable selection was carried out using the ranger function of the ranger R-package, which uses the corrected Gini impurity RF measure. In addition to the variables selected for RF, three "strategy proxy" variables were used to control the (possible) impact of changes, red cards or injuries on team strategy during the match.

Regarding the analysis of the variables that have a greater impact on the discrimination between winning, drawing or losing teams or that increase the probability of scoring, the first conclusion is that H teams have a higher likelihood of scoring (see Table 4.2.). This is true for all leagues except the Bundesliga, where there is no evidence of a H effect on goal probability. On the other hand, shooting accuracy is essential for increasing the likelihood of scoring (see Table 4.2.) and winning the match (Figure 4.4.) in all leagues. According to Table 4.2., Aw teams receive a higher number of shots on goal, increasing the probability of obtaining a goal and losing the match (Figure 4.4.). In terms of game actions related to team shots and goals, the results of the study (Table 4.2.) showed that the impact of actions leading to a shot (SCA) on the probability of scoring is twice as high in Aw teams competing in the Premier League than in the other leagues. Furthermore, looking at the positive effect of the variable TB, which refers to completed passes sent between defenders at the back into open space, one could conclude that Aw teams in LaLiga and Bundesliga play counter-attacking football. According to Figure 4.4., on average, winning teams performed more TB and SCA in all leagues. According to Table 4.2., the variable dead-ball passes may indicate that Aw teams have a defensive strategy that reduces the goal probability of H and could lead to a draw (Figure 4.4.). Note that the results of the three models are consistent with each other and with previous research, which shows that H teams and winning teams have more aggressive behaviour (Bialkowski et al. 2014; Carpita et al. 2015). In addition, the results suggest that winning teams keep starting players on the pitch longer and that red cards increase the likelihood of losing a match (see Figure 4.4.).

Regarding question i), the Bundesliga and Ligue 1 are the only leagues that show a difference in performance between the 2019/2020 and 2020/2021 seasons. Thus, it is concluded that the performance of teams competing in LaLiga, Serie A and Premier League during both seasons was similar (see Table 4.2.).

Focusing on the predictive accuracy of the SRM in all leagues and taking into account the results of the two seasons, it can be concluded that, according to the MCC, the outcomes of the matches of the Bundesliga in season 2019/2020 were the best predicted by the SRM (MCC=0.423)(see Table 4.2. and Figure 4.2.). In the same way, it is worth noting the low average MCC in the case of LaLiga in the 2020/2021 season (MCC=0.288) (see Table 4.5. and Figure 4.3.).

Concerning question iii), this study highlights the better predictive accuracy of the multivariate techniques regarding the SRM. Additionally, throughout this chapter, it has been demonstrated some disadvantages of using the SRM. First, it forces us to use H and Aw game actions as explanatory variables. Second, as in all regression models, it is necessary to eliminate highly correlated variables. Therefore, when working with a large data set, a variable selection method should be used before using the SRM, verifying that the selected game actions are not highly correlated.

Finally, in terms of prediction accuracy, this study highlights PLS-DA as the model with the best statistical classification performance (Figures 4.5. and 4.6.). Even though PLS-DA does not provide an outstanding classification, it should be noted that the results obtained are in line with those obtained by previous researchers who already highlighted the difficulty of predicting draw outcomes (Karlis and Ntzoufras 2009; Carpita et al. 2015; Carpita, Ciavolino, and Pasca 2019; Carpita, Ciavolino, and Pasca 2021).

4.5 Conclusion

The results obtained throughout this chapter reinforce the reflections reached in Chapter 2 (Section 2.4) when predicting the final position of the teams (top, middle and bottom). Given that football is a sport in which chance and uncertainty (Carpita et al. 2015) play an important role and where many factors interact, it does not seem logical to judge the teams' performance based solely on match outcomes. Furthermore, it is also highlighted that, as investigated in Chapter 3, even though there may be differences in team strategies, styles or levels of competitiveness between leagues, all football leagues share fundamental game actions.

4.6 Appendix

Table B.1.: Variables classified by type of game actions and their corresponding description

Variable	Type	Description
SoT	Shooting	Shots total
SoT_ptg	Shooting	Shots on target percentage
FK	Shooting	Shots from free kicks
SoTA	Goalkeeper	Shots on target against
PassesLaunch	Goalkeeper	Passes completed (passes longer than 40 yards)
AttLaunched	Goalkeeper	Passes attempted (passes longer than 40 yards)
Comp_ptg	Goalkeeper	Pass completion percentage (passes longer than 40 yards)
AttNGK	Goalkeeper	Passes attempted (not including goal kicks)
Thr	Goalkeeper	Throws attempted
Launch_ptgNGK	Goalkeeper	Percentage of launched passes (passes longer than 40 yards)
AvgLenNGK	Goalkeeper	Average length of passes, in yards (not including goal kicks)
AttGKA	Goalkeeper	Goal kicks attempted
Launch_ptg	Goalkeeper	Percentage of launched goal kicks (passes longer than 40 yards)
AvgLen	Goalkeeper	Average length of goal kicks, in yards
Opp	Goalkeeper	Opponent's attempted crosses into penalty area
Stp	Goalkeeper	Number of crosses into penalty area which were successfully stopped by the goalkeeper
Stp_ptg	Goalkeeper	Percentage of crosses into penalty area which were successfully stopped by the goalkeeper
OPA	Goalkeeper	Number of Defensive actions outside of penalty area
AvgDist	Goalkeeper	Average distance from goal (in yards) of all Defensive actions
PassesAtt	Passing	Passes attempted
Pass_ptg	Passing	Pass completion percentage

Variable	Type	Description
TotDist	Passing	Total distance that completed passes have travelled in any direction, in yards
PrgDist	Passing	Total distance that completed passes have traveled towards the opponent's goal, in yards
PassShort	Passing	Passes completed (passes between 5 and 15 yards)
Pass_ptgShort	Passing	Pass completion percentage (passes between 5 and 15 yards)
PassMedium	Passing	Passes completed (passes between 15 and 30 yards)
PassesAttMedium	Passing	Passes attemped (passes between 15 and 30 yards)
Pass_ptgMedium	Passing	Pass completion percentage (passes between 15 and 30 yards)
PassLong	Passing	Passes completed (passes longer than 30 yards)
PassesAttLong	Passing	Passes attemped (passes longer than 30 yards)
Pass_ptgLong	Passing	Pass completion percentage (passes longer than 30 yards)
KP	Passing	Passes that directly lead to a shot (assisted shots)
1/3	Passing	Completed passes that enter the 1/3 of the pitch closest to the goal
PPA	Passing	Completed passes into the into the penalty area
CrsPA	Passing	Completed crosses into the into the penalty area
Prog	Passing	Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area
Live	Pass types	Live-ball passes
Dead	Pass types	Dead-ball passes (Includes free kicks, corner kicks, throw-ins, and goal kicks)

Variable	**Type**	**Description**
PassTypesFK	Pass types	Passes attempted from free kicks
TB	Pass types	Completed passes sent between back defenders into open space
Press	Pass types	Passes made while under pressure from opponent
Sw	Pass types	Passes that travel more than 40 yards of the width of the pitch
Crs	Pass types	Crosses
CK	Pass types	Corner kicks
In	Pass types	Inswinging corner kicks
OutCK	Pass types	Outswinging corner kicks
Str	Pass types	Straight corner kicks
Ground	Pass types	Ground passes
Low	Pass types	Passes that leave the ground, but stay below shoulder-level
High	Pass types	Passes that are above shoulder-level at the peak height
Left	Pass types	Passes attempted using left foot
Right	Pass types	Passes attempted using right foot
Head	Pass types	Passes attempted using head
TI	Pass types	Throw-Ins taken
Other	Pass types	Passes attempted using body parts other than the player's head or feet
Off	Pass types	Offsides
Out	Pass types	Out of bounds
Int	Pass types	Intercepted
Blocks	Pass types	Blocked by the opponent who was standing it the path
SCA	Goal and Shot Creation	Shot-Creating actions (Actions leading directly to a shot (passes, dribbles,drawing fouls, etc.))
PassLiveSCA	Goal and Shot Creation	Completed live-ball passes that lead to a shot attempt
PassDeadSCA	Goal and Shot Creation	Completed dead-ball passes that lead to a shot attempt (Includes free kicks, corner kicks, throw-ins, and goal kicks)
DribSCA	Goal and Shot Creation	Successful dribbles that lead to a shot attempt

Variable	Type	Description
ShSCA	Goal and Shot Creation	Shots that lead to another shot attempt
FldSCA	Goal and Shot Creation	Fouls drawn that lead to a shot attempt
DefSCA	Goal and Shot Creation	Defensive actions that lead to a shot attempt
Tkl	Defensive actions	Number of players tackled
TkW	Defensive actions	Tackles in which the tackler's team won possession of the ball
Def1/3	Defensive actions	Tackles in defensive 1/3
Mid1/3	Defensive actions	Tackles in middle 1/3
Att1/3	Defensive actions	Tackles in attacking 1/3
TklADri	Defensive actions	Number of dribblers tackled
Tk_plus_Drib	Defensive actions	Number of times dribbled past plus number of tackles
Tkl_ptg	Defensive actions	Percentage of dribblers tackled
Past	Defensive actions	Number of times dribbled past by an opposing player
PressPress	Defensive actions	Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball
PressSucc	Defensive actions	Number of times the squad gained possession withing five seconds of applying pressure
ptg	Defensive actions	Successful Pressure Percentage
PressDef1/3	Defensive actions	Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the defensive 1/3

Variable	**Type**	**Description**
PressDefMid1/3	Defensive actions	Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the middle 1/3
PressAtt1/3	Defensive actions	Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the attacking 1/3
BlockBlock	Defensive actions	Number of times blocking the ball by standing in its path
ShBlock	Defensive actions	Number of times blocking a shot by standing in its path
ShSv	Defensive actions	Number of times blocking a shot on target, by standing in its path
Pass	Defensive actions	Number of times blocking a pass by standing in its path
IntBlock	Defensive actions	Interceptions
Clr	Defensive actions	Clearence
Err	Defensive actions	Mistakes leading to an opponent's shot
Poss_ptg	Possession	Possession percentage
TouchTouches	Possession	Number of times a player touched the ball
TouchDefPen	Possession	Touches in defensive penalty area
TouchDef1/3	Possession	Touches in defensive 1/3
TouchMid1/3	Possession	Touches in middle 1/3
TouchAttMid1/3	Possession	Touches in attacking 1/3
TouchAttPen	Possession	Touches in attacking penalty area (does not include corner kicks, free kicks, throw-ins, goal kicks or penalty kicks)
TouchLive	Possession	Live-ball touches
DrbSucc	Possession	Dribbles completed successfully
DrbAtt	Possession	Dribbles attempted
DrbSucc_ptg	Possession	Percentage of dribbles completed successfully
PI	Possession	Number of players dribbled past

Variable	Type	Description
Megs	Possession	Number of times a player dribbled the ball through an opposing player's legs
Carries	Possession	Number of times the player controlled the ball with their feet
CarriesTotDist	Possession	Total distance, in yards, a player moved the ball while controlling it with their feet, in any direction
CarriesPrgDist	Possession	Progressive distance (Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal)
CarriesProg	Possession	Carries that move the ball towards the opponent's goal at least 5 yards, or any carry into the penalty area
Carries1/3	Possession	Carries that enter the 1/3 of the pitch closest to the goal
CPA	Possession	Number of times a player controlled the ball inside the opponent's penalty area
Miss	Possession	Number of times a player failed when attempting to gain control of a ball
CarriesDis	Possession	Number of times a player loses control of the ball after being tackled by an opposing player
Targ	Possession	Number of times a player was the target of an attempted pass
Rec	Possession	Number of times a player successfully received a pass
ReceivingProg	Possession	Progressive passes received
Rec_ptg	Possession	Passes received percentage
CrdY	Miscellaneous stats	Yellow Cards
SecondCrdY	Miscellaneous stats	Second Yellow Card
Fls	Miscellaneous stats	Fouls committed
Fld	Miscellaneous stats	Fouls drawn

Variable	**Type**	**Description**
Recov.Performance	Miscellaneous stats	Number of loose balls recovered
WonAD	Miscellaneous stats	Aerials won
LostAD	Miscellaneous stats	Aerials lost
WonAD_ptg	Miscellaneous stats	Percentage of aerials won

Table B.2.: Comparison of the important and statistically significant variables (p-values<0.05) in the PLS-DA and RF, respectively, for the "Big Five" (75% training set and 25% testing set, 100 replications). The variables in bold indicate the top ten variables selected by the VIP and statistically significant for RF in most leagues and seasons - Season 2019/2020

League	PLS	RF
LaLiga	AttGKA, Blocks, BlockBlock, **Clr**, **CPA**, Crs, Change_H, *CrdR*, CrdY, **Dead**, DribSCA, Err, In, KP, Mid1/3, OPA, **Opp**, Pass, PassLiveSCA, **SCA**, **SoTA**, **SoT_ptg**, Stp, **TB**, TI, Thr, TouchDefPen, **TouchAttPen**, and PassShort	AttGKA, AvgDist, AvgLen, **Clr**, **CPA**, Crs, **Dead**, Head, High, *Mean_60_H*, **Opp**, ptg, **SCA**, SD_60_H, **SoT_ptg**, **SoTA**, Stp_ptg, TI, Tkl_ptg, and TouchDefPen
SerieA	Carries, Carries1/3, **Clr**, **CPA**, *CrdR*, **Dead**, DefSCA, Err, Ground, KP, Launch_ptg, Live, **Opp**, PassCmp_ptg, Pass_ptgLong, Pass_ptgMedium, PassesAtt, PassLiveSCA, PPA, PrgDist, ptg, Rec, Right, **SCA**, **SoT**, ShSCA, **SoT_ptg**, **SoTA**, Targ, **TB**, TotDist, **TouchAttPen**, TouchLive, TouchTouches, PassLong, and PassShort	AttLauched, AvgLen, AvgLenNGK, **Clr**, **CPA**, Crs, **Dead**, KP, Launch_ptg, Launch_ptgNGK, *Mean_60_H*, **Opp**, PassLiveSCA, ptg, **SCA**, SD_60_H, **SoT**, **SoT_ptg**, **SoTA**, and **TouchAttPen**
Premier	Att1/3, Carries, Carries1/3, **Clr**, **CPA**, **Dead**, DefSCA, DribSCA, Err, FldSCA, Ground, KP, Live, Mis, **Opp**, Pass_ptgLong, PassLiveSCA, PPA, PrgDist, Prog, Rec, ReceivingProg, Right, **SCA**, **SoT**, ShSCA, **SoT_ptg**, **SoTA**, **TB**, TotDist, **TouchAttPen**, TouchLive, TouchMid1/3, TouchTouches, PassLong, PassShort, and WonAD_ptg	AvgDist, AvgLen, CarriesDis, **Clr**, Comp_ptg, **Dead**, Head, High, Launch_ptg, *Mean_60_H*, Mis, **Opp**, Out, Pass_ptgShort, PressSucc, **SoT_ptg**, **SoTA**, Stp_ptg, Sw, TI, Tkl, Tkl_ptg, and WonAD_ptg
Bundesliga	Carries, **Clr**, **CPA**, **Dead**, DefSCA, Err, Ground, High, KP, Live, PassCmp_ptg, Pass_ptgShort, PassesAtt, PassesAttMedium, PassLiveSCA, Poss_ptg, PPA, PrgDist, ptg, Rec, Right, **SCA**, **SoT**, **SoT_ptg**, **SoTA**, Targ, **TB**, TotDist, **TouchAttPen**, TouchDef1/3, TouchLive, TouchTouches, PassMedium, and PassShort	**Clr**, Comp_ptg, **CPA**, **Dead**, DrbSucc_ptg, High, *Mean_60_H*, Mis, Out, SD_60_H, **SoT_ptg**, **SoTA**, Tkl_ptg, and **TouchAttPen**
Ligue 1	AvgDist, AvgLenNGK, BlockBlock, **Clr**, **CPA**, *CrdR*, **Dead**, DefSCA, DribSCA, KP, Left, OPA, **Opp**, Pass_ptgShort, PassLiveSCA, PPA, **SCA**, SecondCrdY, **SoT**, **SoT_ptg**, **SoTA**, Stp, **TB**, Thr, TI, **TouchAttPen**, and PassShort	AvgDist, AvgLen, AvgLenNGK, BlockBlock, **Clr**, Comp_ptg, **CPA**, Dead, High, PassLiveSCA, **SCA**, **SoT_ptg**, **SoTA**, **TB**, TI, and Tkl_ptg

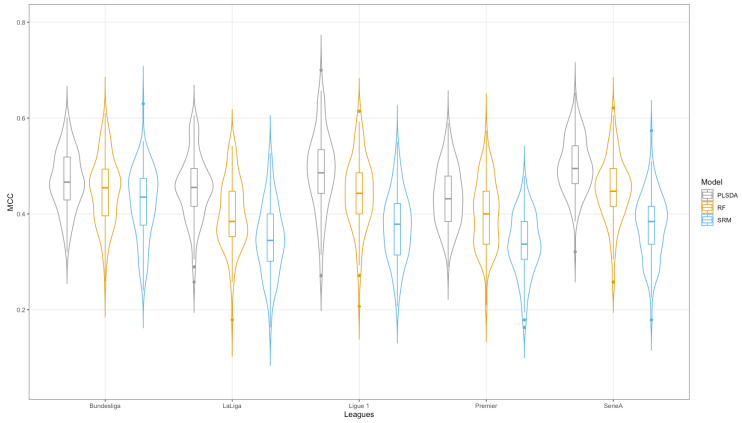


Figure B.1.: Violin plot in combination with the box plot to compare the distribution of the MCC (Y-axis) depending on the league and model: PLS-DA (grey), RF (yellow) and SRM (blue) - Season 2019/2020

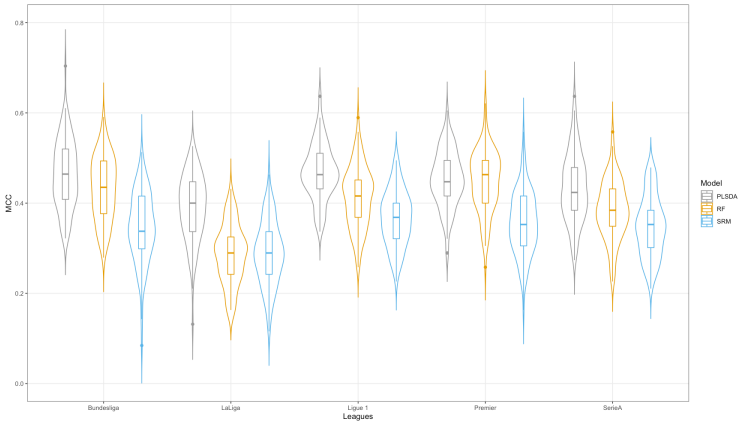


Figure B.2.: Violin plot in combination with the box plot to compare the distribution of the MCC (Y-axis) depending on the league and model: PLS-DA (grey), RF (yellow) and SRM (blue) - Season 2020-2021

Table B.3.: Comparison of the important and statistically significant variables (p-values<0.05) in the PLS-DA and RF, respectively, for the "Big Five" (75% training set and 25% testing set, 100 replications). The variables in bold indicate the top ten variables selected by the VIP and statistically significant for RF in most leagues and seasons - Season 2020/2021

League	**PLS**	**RF**
LaLiga	AvgDist, CarriesDis, Carries1/3, Change_H, CrdY, **Dead**, DefSCA, Err, Ground, Head, In, IntBlock, Left, Live, Megs, OPA, **Opp**, PassesAtt, Pass_ptgLong, PassLiveSCA, PassTypesFK, Poss_ptg, PPA, PressAtt1/3, Prog, PrgDist, Rec, SCA, SecondCrdY, Right, **SoT**, ShSCA, Str, **SoT_ptg**, **SoTA**, Stp, Stp_ptg, Str, Sw, Targ, **TB**, TotDist, TouchDefPen,TouchDef1/3, TouchLive, TouchTouches, PassMedium, PassShort, and WonAD_ptg	AvgLen, **Clr**, **Dead,** Head, Launch_ptg,**Opp**,PassTypesFK, PressPress, **SoT_ptg**, **SoTA**, **TB**, and TouchDefPen
SerieA	AvgLen, Blocks, CarriesDis, **Clr**, **CPA**, *CrdR*, Crs, **Dead**, DribSCA, High, KP, Launch_ptg, *Mean_60_H*, Mid1/3, Off, Off.Performance, OPA, **Opp**, Pass, PassLiveSCA, PPA, **SCA**, SD_60_H, **SoT**, ShSCA, **SoT_ptg**, **SoTA**, Stp, **TB**, Thr, Tkl, TklW, and **TouchAttPen**	AvgLen, **Clr**, Comp_ptg, **CPA**, Crs, **Dead**, High, Launch_ptg, *Mean_60_H*, **Opp**, Pass_ptgShort, PassLiveSCA, Recov.Performance, SCA, SD_60_H, **SoT_ptg**, and **SoTA**
Premier	**Clr**,**CPA**, *CrdR*, CrdY, Crs, PA, **Dead**, DefSCA, DribSCA, Err, Ground, High, KP, OPA, **Opp**, Pass_ptgLong, PassLiveSCA, PrgDist, SCA, **SoT**, ShSCA, **SoT_ptg**, **SoTA**, Stp, **TB**, **TouchAttPen**, TouchLive, TouchTouches, and PassShort	AttNGK, AvgDist, AvgLen, **Clr**, **CPA**, Crs, **Dead**, Head, High, *Mean_60_H*, **Opp**, PassLiveSCA, SCA, SD_60_H, **SoT_ptg**, **SoTA**, Stp_ptg, **TB**, and TouchDef1/3
Bundesliga	AvgLen, AvgLenNGK, CarriesOne_Third, **Clr**, **CPA**, **Dead**, DefSCA, DrbSucc_ptg, DribSCA, KP, Launch_ptg, Launch_ptgNGK, lostAD, Mid1/3, PassLiveSCA, PPA, **SCA**, **SoT**, ShSCA, **SoT_ptg**, **SoTA**, **TB**, **TouchAttPen**, TouchAtt1/3, WonAD_ptg, and WonAD	AttLauched, AttNGK, AvgDist, AvgLen, AvgLenNGK, **Clr**, **CPA**, **Dead**, DrbSucc_ptg, High, Launch_ptg, Launch_ptgNGK, Mid1/3, Pass_ptgShort, PassesLaunch, PassLiveSCA, **SCA**, SD_60_H, **SoT_ptg**, **SoTA**, and Tkl_ptg
Ligue 1	AttLauched, AvgDist, AvgLen, AvgLenNGK, CarriesDis, **Clr**, **CPA**, CrdR, Crs, **Dead**, DribSCA, Err, FldSCA, KP, Launch_ptg, Launch_ptgNGK, Mid1/3, OPA, **Opp**, Other, Pass, PassesLaunch, PassLiveSCA, PPA, **SCA**, **SoT**, **SoT_ptg**, **SoTA**, Sw, **TB**, Thr, TI, Tkl, and **TouchAttPen**	AvgDist, AvgLenNGK, **Clr**, **CPA**, **Dead**, Head, Launch_ptgNGK, *Mean_60_H*, **Opp**, Pass_ptgShort, PassesLaunch, PassLiveSCA, ptg, **SCA**, **SD_60_H**, **SoT_ptg**, SoTA, **TB**, Thr, and **TouchAttPen**

Chapter 5

Development of popularity indicators with Google Trends to measure popularity influence on the market value of players

Part of the content of this chapter has been included in:

1. *Malagón-Selma, Pilar, Ana Debón, and Josep Domenech (2022). "Influence of popularity on the transfer fees of football players". In: 4th International Conferenceon Advanced Research Methods and Analytics (CARMA2022). Valencia, Spain.*

2. *Malagón-Selma, Pilar, Ana Debón, and Josep Domenech (2022). "Influencia de la popularidad en la tarifa de transferencia de los jugadores de fútbol". In: XII Congreso Iberoamericano de Economía del Deporte (CIED 12). Toledo, España.*

Abstract

Google Trends helps to measure popularity, as it collects information from Google searches. However, sports analysts have not yet used Google Trends often or correctly. This research proposes a novel method of calculating indicators to measure the popularity of football players. Google Trends provides a time series as a normalised index, making it challenging to compare player popularity. Therefore, to get the popularity indicators correctly, it is necessary to request players simultaneously and build popularity levels. In addition, to compare them, it is essential to rescale the popularity of the players through a cumulative conversion factor. Once the indicators had been calculated, to verify their usefulness, we studied whether the performance of predicting the players' transfer fees improved when they were introduced into the models. The database consisted of 1,428 players who competed in LaLiga, Premier League, Bundesliga, Serie A and Ligue 1 during the 2018-2019 season. The results showed that the proposed popularity indicators improved the prediction and were influential in predicting the market value of the players. This study contributes to the literature with a practical guide to measuring popularity using Google Trends, which can be helpful for sports analysts and researchers in any field of study.

5.1 Introduction

Football is one of the most profitable businesses in the world. In fact, according to the consulting firm Deloitte (Bridge et al. 2023), the combined turnover of the top 3 clubs (Manchester City, Real Madrid, and Liverpool) was €2.146,5 million in income in the 2021/2022 season. However, these revenue amounts are preceded by significant investments by the teams in their main assets: footballers. One example is Manchester City which, according to CIES Football Observatory (Poli, Ravenel, and Besson 2022), invested €1.806 million in signing new players between 2013 and 2022. Therefore, given the impact of signings on the football club economy, one of the challenges in football today is the players' valuation.

In recent years, the number of researches that aim to predict the market value[1] of players has increased. Thus, the studies carried out so far use data analysis tools to try to minimise the deviation of the prediction error concerning the transfer fee [2].

[1]"estimate of the price a football team would be willing to pay for a player to sign a contract"(Steffen, Hans-Markus, and Henning 2014)

[2]"real transaction paid"(Müller, Simons, and Weinmann 2017)

Researchers from different fields of knowledge have begun to specialise in the valuation of players and to study the factors that affect the market value and, therefore, transfer fees. Three main groups of variables have been identified as essential in predicting the market value of players: characteristics, performance, and popularity.

In terms of player characteristics, previous research included *age*, *height*, *footedness*, *position* and *contract* (years to the end of the contract) in their analyses, as they considered that they could be essential in predicting market value (Feess, Frick, and Muehlheusser 2004; He, Cachucho, and Knobbe 2015; Müller, Simons, and Weinmann 2017; Behravan and Razavi 2021). According to their results, *age* (Steffen, Hans-Markus, and Henning 2014; Müller, Simons, and Weinmann 2017; Behravan and Razavi 2021; Felipe et al. 2020; Gyimesi and Kehl 2021) and *position* (Frick 2007; Müller, Simons, and Weinmann 2017; Singh and Lamba 2019) are the variables most researchers agree are influential. As for *height* (Bryson, Frick, and Simmons 2013; Behravan and Razavi 2021), *footedness* (Steffen, Hans-Markus, and Henning 2014; Bryson, Frick, and Simmons 2013) and *contract length* (Feess, Frick, and Muehlheusser 2004; Frick 2011), there is no consensus on whether they influence the prediction of market value.

The player's performance corresponds to the game actions carried out during a match. Researchers commonly find *playing time* to be a statistically significant variable for predicting market value (Feess, Frick, and Muehlheusser 2004; Müller, Simons, and Weinmann 2017; Singh and Lamba 2019). Concerning shots and goal creation variables, *goals* and *assists* stand out as game actions with positive statistical significance on market value. Recent studies also included in the prediction model both *shots* and *key passes*, although only *shots* was statistically significant (Behravan and Razavi 2021). *Interceptions*, *yellow* and *red cards*, and *fouls* are generally included defensive variables, although only *yellow cards* were statistically significant (Müller, Simons, and Weinmann 2017). Müller, Simons, and Weinmann (2017), Singh and Lamba (2019), and Behravan and Razavi (2021) found *passing accuracy* and *dribbles* were statistically significant game actions.

Even though player characteristics and performance store essential information to predict players' market value, with the advent of social media, a new challenge arises for researchers who have to find a method to store this information in measurable variables that allow them to study the (likely) effect of popularity on the transfer fees.

Therefore, considering that in most cases, the information related to popularity is open and easy to find, researchers have tried to measure it in different ways. Followers on *Facebook*, *Twitter*, and *Instagram* are the most widely used variables to study the effect of player popularity on market value (Müller, Simons, and Weinmann 2017; Hofmann et al. 2019). *Google hits*[3] (Garcia-del-Barrio and Pujol 2007; Steffen, Hans-Markus, and Henning 2014; Hofmann et al. 2019) or *Wikipedia views* (Müller, Simons, and Weinmann 2017; Singh and Lamba 2019) are other variables commonly introduced into predictive models to estimate market value and then predict transfer fees. In all cases, the variables used by the researchers were statistically significant and had a positive coefficient. Müller, Simons, and Weinmann (2017) also incorporated *Reddit posts*, *YouTube videos* and *Google Trends (GT)*. In the case of Müller, Simons, and Weinmann (2017), all variables showed a statistically significant effect except GT. GT is a very interesting tool that allows us to discover the interest that a subject or person has aroused over time worldwide. This measurement is obtained from the number of searches made in Google Search Engine (Rogers 2016). In this context, GT provides a granular time series based on measuring the number of times a player's name has been searched on Google. However, although this variable could be closely related to popularity, Müller, Simons, and Weinmann (2017) did not find it statistically significant. Specifically, Müller, Simons, and Weinmann (2017) used as a popularity variable the GT time series average (GTA) got from each player individually.

The non-statistical significance of the GTA is probably explained by the method used to calculate it. GT provides term-dependent normalised indices from 0 to 100. Therefore, since the values obtained are not time series of absolute searches, this procedure is unsuitable for direct player comparison.

Then, there is a gap in how GT can be used to quantify the influence of popularity on the market value of players. First, since GT contains information over time, it must be used and summarised appropriately. Second, some of the variables used in previous studies (probably) store duplicate information, e.g. visits to Wikipedia are closely related to Google searches, since most of the time, Google is the browser used to access Wikipedia. Third, although researchers have found that popularity has a statistically significant influence on market value, they have not quantified the specific impact of popularity on the transfer fees.

The main contribution of this chapter is to develop a novel way to obtain popularity indicators (PIs) of the GT time series. One of the advantages of

[3]Number of links reported by Google when entering the name of the players (Steffen, Hans-Markus, and Henning 2014; Hofmann et al. 2019)

GT is that it allows us to summarise and collect various variables related to Google: views from Wikipedia, YoTube or Google hits. Therefore, collecting information from GT simplifies data collection and avoids obtaining redundant information. In addition, the problem of missing values is circumvented.

To show the adequacy of the proposed popularity indicators, the market value prediction will be made using the information of the players who competed in the "Big five" European leagues during the season 2018-2019. Then, the error of the model will be evaluated using the transfer fee of the players sold during the summer market of that season. In addition, the popularity effect will be calculated by quantifying the decrease in the prediction error.

The rest of this chapter is organised as follows. Section 5.2 is devoted to explaining the construction of the PIs, presenting the database and the statistical methods and models used. Section 5.3 presents the results obtained and highlights the usefulness of the PIs developed. Sections 5.4 and 5.5 present the discussion and conclusion, respectively.

5.2 Material and methods

This section explains how GT information has been used to build the proposed PIs. Then, the database and statistical methods will be displayed to predict the market value. The multivariate techniques used to train the models are multiple linear regression (MLR)(Berry, Feldman, and Feldman D. 1985), a classic method used as benchmark, RF and gradient boosting machine (GBM)(Friedman 2001; Friedman 2002). As in previous chapters, the free R software was used to carry out the analysis (R Core Team 2019).

5.2.1 Development of popularity indicators

GT is a tool that provides information about the interest in a subject or person by means of a time series. The GT time series contains relative granular values (horly, daily, weekly, monthly, and yearly) of a topic normalised from 0 to 100. The time series reach 100 in the period when the subject has the maximum number of searches (Rogers 2016). In our context, even though the normalisation allows us to study the evolution of the player's popularity, it does not allow us to compare the popularity of the players since each player's series is individually normalised. According to Rogers (2016), additional topics can be added to put the popularity of the search topic into perspective. Therefore, using two players simultaneously, the time series provided by GT are jointly

normalised to the maximum of either. As a result, both series are obtained on the same scale and are now comparable. Henceforth, the time series obtained from GT after comparing two players will be called GTN. Note that since the main idea is to compare the popularity of all players, they must be on the same scale and therefore rescaled concerning the same player.

One problem with how GT provides the time series is that it gives integer values instead of real numbers. Thus, if two players with unequal popularity are compared, the GT reports a GTN of 0 for the most unpopular player. To solve this problem, this study proposes to select some players as references (Ref) according to their relative popularity. Considering that the notoriety of the players may depend on their position, the Ref will also be ranked by position.

Table 5.1. shows the Ref_{lj} where the row indicates level of popularity $l = 1, 2, 3$ and the column $j = for, mid, def$ the position of the players.

Table 5.1.: Reference players according to their popularity level and position

Popularity level	**Forward**	**Midfielder**	**Defender**
1	Cristiano Ronaldo	Paul Pogba	Sergio Ramos
2	Franck Ribry	Fabin Ruiz	Kamil Glik
3	Kevin Lasagna	Yannick Gerhardt	Diego Rico

For forwards, midfielders and defenders were chosen as Ref of the first level ($\text{Ref}_{1\text{j}}$), the most popular player compared to the rest. The Ref of the second level ($\text{Ref}_{2\text{j}}$) was the least popular player out of those whose GTA respect to $\text{Ref}_{1\text{j}}$ was 1. The same procedure was followed with $\text{Ref}_{3\text{j}}$ regarding $\text{Ref}_{2\text{j}}$. GTN($\text{player}_{\text{l,j}}$, Ref_{lj}) indicates the normalised time series of GT calculated by comparing $\text{player}_{\text{l,j}}$ with a specific Ref_{lj}.

At the end, it was necessary to rescale all the players concerning a single player $\text{Ref}_{1,\text{for}}$ who was the most popular player in the period studied. Therefore, a conversion factor $\text{CF}_{\text{l,j}}$ of each level $l = 1, 2, 3$ and position $j = for, mid, def$ was obtained as follows,

$$CF_{\mathrm{lj}} = \begin{cases} \dfrac{max(GTN(Ref_{\mathrm{1,j}}, Ref_{\mathrm{1,j}}))}{100} & l = 1 \quad j = for, mid, def. \\ \dfrac{max(GTN(Ref_{\mathrm{1,j}}, Ref_{\mathrm{l\text{-}1,j}}))}{100} & l = 2, 3 \quad j = for, mid, def. \end{cases} \tag{5.1}$$

Then, all players were rescaled concerning $\mathrm{Ref_{1,f}}$ via the cumulative conversion factor $\mathrm{CCF_{lj}}$ calculated from those factors $\mathrm{CF_{lj}}$ from Equation 5.1,

$$CCF_{\mathrm{l,j}} = \prod_{r=1}^{i} CF_{\mathrm{r,j}} \quad l = 1, 2, 3 \quad j = for, mid, def. \tag{5.2}$$

Equation 5.3 was defined to rescale player popularity to the corresponding $CCF_{\mathrm{l,j}}$ from Equation 5.2.

$$GTN(player_{\mathrm{l,j}}, Ref_{\mathrm{1,for}}) = GTN(player_{\mathrm{l,j}}, Ref_{\mathrm{l,j}}) \times CCF_{\mathrm{l,j}} \quad l = 1, 2, 3 \quad j = for, mid, def. \tag{5.3}$$

5.2.2 Database

The database consisted of 1428 players competing in European leagues (LaLiga, Premier League, Bundesliga, Serie A and Ligue 1) throughout the 2018-2019 season, and 41 explanatory variables grouped into three classes: *physical characteristics*, *performance*, and *popularity* (see Table 5.2.). The variables used in this chapter were collected through the data sources FBref *(fbref.com)*, WhoScored *(whoscored.com)*, Google Trends *(trends.google.es)*, and Transfermarkt *(transfermarkt.es)*. As for the model's evaluation, the model's training was carried out using the market value of the non-transferred players as dependent variables, specifically, the market value estimated by the Transfermarkt[4] website. In the test set, i.e. to evaluate the model's error, the response variable used was the transfer fee of the 193 players transferred in the summer market of the season 2018-2019 (Müller, Simons, and Weinmann 2017; Behravan and Razavi 2021).

[4]Transfermarkt is a website that estimates players' market value based on the "judgement" of some members. In recent years Transfermarkt has gained a high reputation within clubs and among researchers who consider the value given by the Transfermarkt platform as a valid method to estimate the market value Müller, Simons, and Weinmann 2017; Behravan and Razavi 2021; Singh and Lamba 2019.

Table 5.2.: Variables grouped by class used to estimate players' market value

Class	**Variable and label**
Dependent variable	Transfermarkt's market value (train set) Transfer fees (test set)
Player characteristics	Position (P), Footedness (FT), Age, Height (H), and Contract (Ct)
Player performance	Playing time (PT), Aerial duel accuracy (ADA), Tackles accuracy (TA), Shots intercepted (SI), Fouls (F), Yellow cards (YC), Red cards (RC), Goals (Gls), Shots (S), Shot accuracy (SA), Assists (A), Dribbles (Dr), Crosses (Cr), Corners (C), Passing accuracy (PA), Short passes accuracy (SPA), Long passes accuracy (LPA), Key passes (KP, passes that create shots for teammates), Progressive passes (PP, passes that move the ball to the opponent's goal), Deep passes (DP, passes into space between defenders), Passes in the penalty area (PPA), Passes in the last quarter of the opponent half (PLO), and Free kicks (FK)
Popularity Indicators	GTA, PC1, mean (Mn), median (Mdn), maximum (Max), minimum (Min), and Standard deviation (Sd)

Note that, in the case of PIs, GTN was calculated regarding the number of times each player had been searched worldwide in the category "Web search", i.e. any search by name, image, news, shopping or Youtube was considered. The analysis period covered was from 17 May 2018 (the date on which the first summer market opens, in this case, the Premier League) and 26 May 2019 (the date on which Serie A closes the summer market, the last league to close). Thus, the GTN time series, from which the PIs were calculated, consisted of 54 values referring to the weekly popularity of each player.

5.2.3 Methods

Multiple Linear Regression

MLR is a classical method where the dependent variable (Y) is predicted through a linear combination of explanatory variables ($X_1, \ldots, X_m$). MLR allows us to know the statistically significant effect of the explanatory variables on the dependent variable. However, before performing MLR, avoiding multicollinearity among the explanatory variables was necessary. Then, the stepwise variable selection method was used for variable selection from the AIC (Akaike 1974).

Random Forest (RF)

The supervised predictive learning method RF (see Sections 2.2.3 and 3.2.3) implemented in the randomForest R-package (Liaw and Wiener 2002), was used to predict the market value of players, since as explained in Section 4.2.2, RF can be used for both classification and regression problems. As presented in previous chapters, RF was also used to study the variables' importance.

Gradient Boosting Machine (GBM)

GBM is a supervised learning method that, like RF, can be used for regression and classification problems. GBM is a boosting method that arose from whether a cluster of several weak classifiers can result in a robust model (Kearns and Valiant 1988; Kearns and Valiant 1994). The GBM model is based on successively creating several predictors (clusters of trees) dependent on each other, shallow and weak. A weak predictor is one whose prediction error rate is slightly better than chance (Schapire 1990). Thus, the main idea is to train

models or trees sequentially so that each one adjusts for the errors of the previous model or tree.

According to Friedman (2001), the first step of the GBM algorithm is fitting the data to a simple decision tree (Equation 5.4), where x represents the explanatory variables and y the response variable.

$$y = f_1(x) \tag{5.4}$$

Then, instead of predicting the response variable, the following tree adjusts the pseudoresidual of the previous tree as a function of explanatory variables (Equation 5.5).

$$h_1(x) = y - f_1(x) \tag{5.5}$$

The next step is to update the original prediction (calculated in Equation 5.4) by adding the results of the adjusted tree $h_1(x)$, to the original one $f_1(x)$ (Equation 5.6).

$$f_2(x) = f_1(x) + h_1(x) \tag{5.6}$$

Note that, to avoid overfitting, the contribution from each new tree is scaled by using a learning rate (value between 0 and 1). This process continues until some criterion (in our case, cross-validation) indicates to stop. Then, the idea is that for each adjusted tree, the prediction obtained is relatively better than the one received by the previous one.

Therefore, boosted regression algorithm can be summarised as a stepwise additive model of b individual regression trees (Equation 5.7).

$$f(x) = \sum_{b=0}^{B} f^b(x) \tag{5.7}$$

In GBM, the variables' contribution is similarly calculated to the importance of RF (Section 4.2.2), even though GBM uses the whole training dataset, not only the OOBs. Concretely, the explanatory variables are randomly permuted (one variable at a time), and the decrease in predictive performance (MSE) is calculated. Then, the value for all trees and variables is averaged (the greater

the average reduction of the predictive accuracy, the greater the variables' contribution).

5.2.4 Models

As indicated in Section 5.1 after the construction of the proposed PIs, their contribution to market value and the extent to which they improved the prediction of transfer fees was studied. In addition, a comparison of the PIs with the GTA used by Müller, Simons, and Weinmann (2017) was carried out. Thus, three models were developed: model 1 (Equation 5.8), which served as the reference model, so no popularity variables were used to calculate the transfer fees; model 2 (Equation 5.9), which added the popularity variable GTA used by Müller, Simons, and Weinmann (2017) to the reference model; and model 3 (Equation 5.10), which included the proposed PIs to the reference model (see Section 5.2.1).

$$Transferfee_{\mathrm{i}} = f(characteristics_{\mathrm{i}}, performance_{\mathrm{i}}) \tag{5.8}$$

$$Transferfee_{\mathrm{i}} = f(characteristics_{\mathrm{i}}, performance_{\mathrm{i}}, GTA_{\mathrm{i}}) \tag{5.9}$$

$$Transferfee_{\mathrm{i}} = f(characteristics_{\mathrm{i}}, performance_{\mathrm{i}}, PIs_{\mathrm{i}}) \tag{5.10}$$

5.2.5 Repeated f-fold cross-validation

This chapter optimised the MLR, RF and GBM hyperparameters in training set by repeated f-fold CV. This validation technique differs from the one used in Chapter 3 (see Section 3.3.2) by replicating the multiple CV procedure.

Once the models were trained, and the hyperparameters optimised, the model was evaluated based on the RMSE of the transfer fees of the 193 players who remained in the test set and, therefore, were not used to build the model. RMSE is typically used to know the difference between the observed and predicted values by the regression models; its expression is:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_{\mathrm{i}} - y_{\mathrm{i}})^2}{n}}, \tag{5.11}$$

where y_n are the observed transfer fees and $\hat{y}_i$ the corresponding predicted transfer fees for each player $i = 1, \ldots, n$. Therefore, the higher the RMSE (Equation 5.11), the worse the prediction model fits the data.

5.3 Results

After applying the method described in Section 5.2.1, the GTN time series of each player were summarised into PIs. The proposed statistical models fitted by the MLR, RF and GBM methods were then used to perform the predictive valuation and study the effect of the PIs on the market value.

5.3.1 Development of popularity indicators

As shown in Section 5.2.1, the development of the PIs consisted of four parts: first, nine players were chosen as Refs concerning their popularity level and position. Secondly, the GTN time series of the remaining (i.e. non-reference) players were obtained by comparing them with the reference of their popularity level. Third, to rescale all players to a single player, their GTN time series values were multiplied by the CCF of their level and position.

Table 5.3. shows the CF and the CCF calculated for each level and position.

Table 5.3.: Conversion factor (CF) and cumulative conversion factor (CCF) for the players according to their popularity level and position

Popularity layer	Factor	Forward	Midfielder	Defender
1	CF	1	0.15	0.14
	CCF	1	0.15	0.14
2	CF	0.04	0.03	0.08
	CCF	0.04	0.0045	0.0112
3	CF	0.04	0.03	0.09
	CCF	0.0016	0.000135	0.001008

The weekly player popularity time series obtained from Equation 5.3 was summarized into six PIs: *mean*, *median*, *maximum*, *minimum*, and *standard deviation* and *First Principal Component (PC1)*. Specifically, the PC1 explained more than 50% of the variability. Note that the proposed PI can be used in cross-sectional studies.

5.3.2 Methods

After applying the methodology described in Section 5.2.3, Table 5.4. shows the estimated coefficients and the statistical significance of the regressors used in each proposed model (see Section 5.2.4). Note that, as indicated in Section 5.2.3, before fitting the MLR, the vif_function was used to remove variables with a VIF greater than 2.5 (Beck 2013). The stepAIC function to the stats R-package was then used to select the most relevant variables according to the AIC criterion (Akaike 1974).

Table 5.4.: Coefficients of the statistically significant variables (p-values<0.05) for the three models fitted by the MLR method.

Dependent variable: Market value			
Explanatory variables	**Model 1**	**Model 2**	**Model 3**
Intercept	845295 .	845295 .	805864 .
Age	-940347 ***	-940347 ***	-958674 ***
Ct	2767608***	2767608 ***	2860415 ***
PT	5151 ***	5151 ***	4908 ***
ADA	108214 **	108214 **	59561 .
F	-3613263 ***	-3613263 ***	-3297411 ***
Gls	52809577 ***	52809577 ***	43962548 ***
A	22451375 ***	22451375 ***	22354543 ***
Dr	784660 **	784660 **	530249 *
SPA	925507 ***	925507 ***	805051 ***
LPA	80557 *	80557 *	69613 .
DP	10234401 ***	10234401 ***	8457341 ***
FK	-4557891 ***	-4557891 ***	-4593996 ***
Min			9914083 ***
Mdn			103096 *

Significance codes: . $p < 0.1$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

For a better overview of the results in Table 5.4., Table 5.5. shows the performance of each method and model as measured by the RMSE.

According to Table 5.5., models 1 and 2 have the same RMSE, i.e. although model 2 includes the variable GTA, there is no difference between the performance of both models. This result is in line with what is shown in Table 5.4. because although in model 2, we included the variable GTA, the AIC discarded it in the step before the MLR fitting. In the case of model 3, of the devel-

Table 5.5.: RMSE for all methods according to the three models (€).

	MLR	**RF**	**GBM**
Model 1	17,101,619	16,721,201	17,190,763
Model 2	17,101,619	17,132,503	18,139,202
Model 3	16,231,400	11,961,370	11,515,548

oped PIs, Min and Mdn were statistically significant in predicting the players' market value (see Table 5.4.). Furthermore, since Min and Mdn are positive and statistically significant, one would expect that the higher the popularity, the higher the market value. According to Table 5.5., including PIs slightly improves the prediction error for the MLR method.

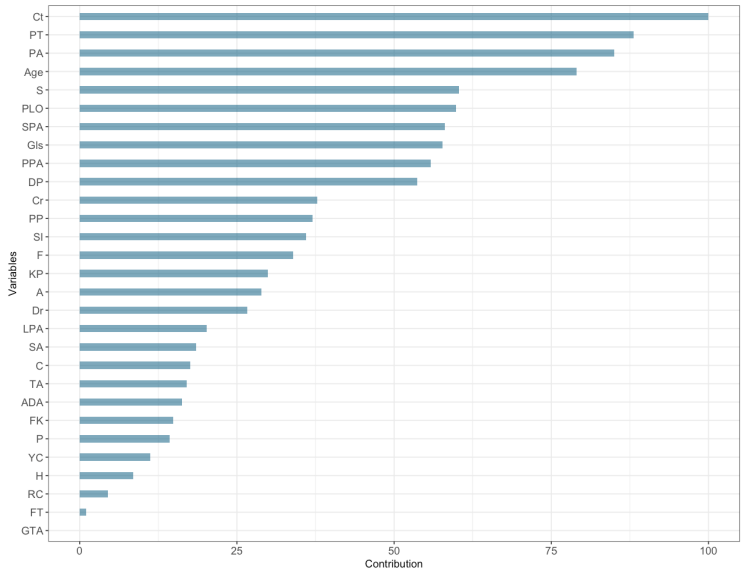
It is noteworthy that in the case of RF and GBM, Table 5.5. shows (barely) better accuracy for model 1 (does not include popularity-related variables) than for model 2 (consists of the GTA variable used by Müller, Simons, and Weinmann (2017)). In addition, model 3 of the GBM method has the lowest RMSE.

As in the MLR, neither in the RF nor in the GBM, GTA is not among the most important variables contributing to the prediction of the market value of the players (see Figures 5.1(a) and 5.2(a)). Conversely, several PIs are important for both the RF and the GBM (see Figures 5.1(b) and 5.2(b)).

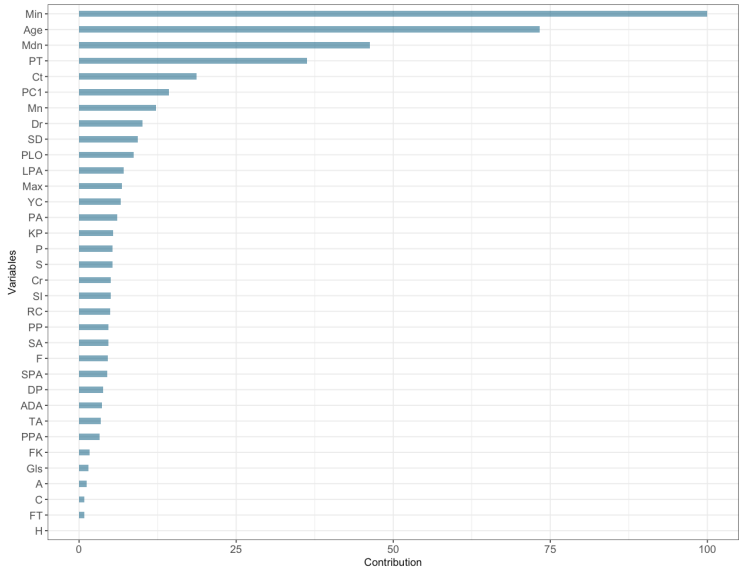
As for the results shown in Table 5.5. using the PIs (model 3) instead of the GTA (model 2), the prediction error decreased by €6,623,654, €5,171,133, and €870,219 for the GBM, RF, and MLR methods, respectively. Consequently, it is possible to conclude that obtaining GTN time series and using the proposed PIs, which summarise the popularity of players, is a valuable way to improve the prediction of transfer fees.

To deepen the prediction and learn about the PIs, the variables that contribute most to the player's market value were analysed. Mainly, the variables selected by the GBM in model 3 since the combination yielded the lower RMSE (Table 5.5.). Then, 5.2(b) below presents a bar chart with the variable names on the y-axis and the average contribution of the variables to the GBM on the x-axis. The contribution of the variables was computed using the gbm R-package, as explained in Section 5.2.3.

In light of 5.2(b), the most contributive variable to the GBM method in model 3 was Min. Likewise, this PI was also the most important variable in the case of the RF 5.1(b) and statistically significant in the MLR (see Table 5.4.).

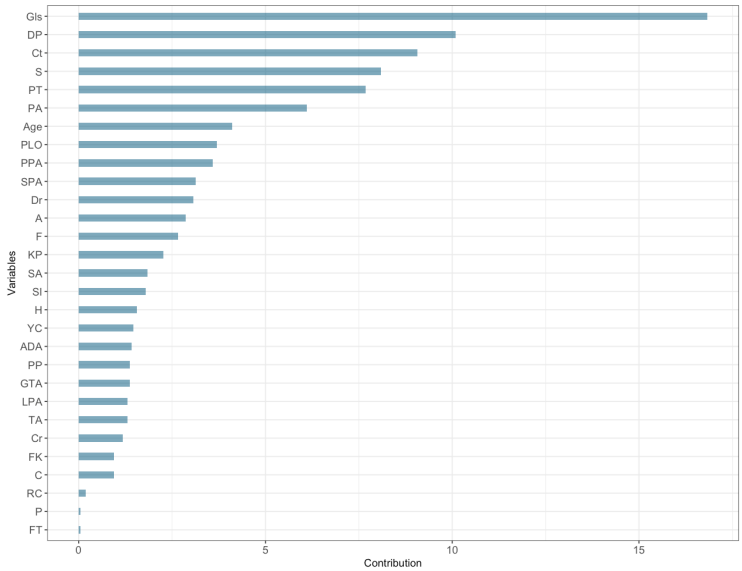


(a) Model 2.

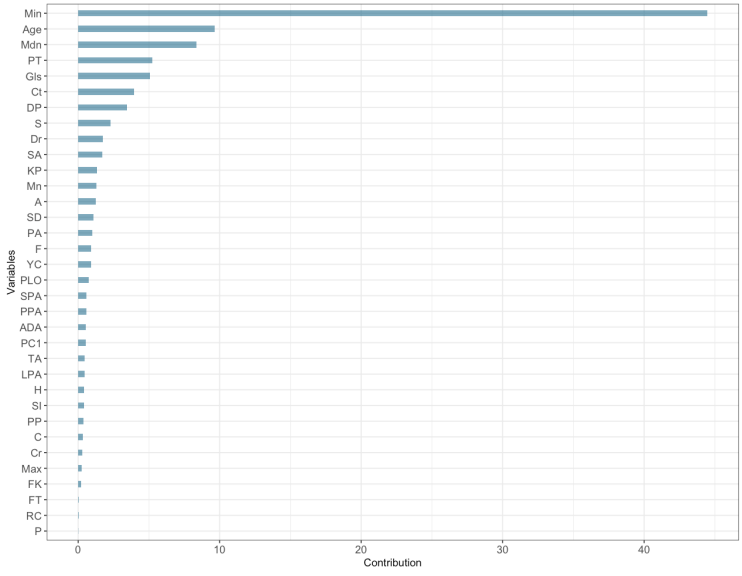


(b) Model 3.

Figure 5.1.: Contribution of variables fitted by the RF method.



(a) Model 2.



(b) Model 3.

Figure 5.2.: Contribution of variables fitted by the GBM method.

Therefore, Min stands out in all three models for its relevance in predicting the market value of the players.

Min contains the search value of the player in the week that was least searched. This PI could be important because this variable stores information on the number of "actual" or "unconditional" fans, i.e. those fans whose interest in the player is not conditioned by successful or unsuccessful actions on the pitch. Thus, the variable Min is not altered by variations in player popularity resulting from specific events. Furthermore, this reflection is supported by the results of the Max variable, which is the PI that contributes the least to the market value prediction in both GBM and RF (see Figures 5.1(b) and 5.2(b)). In contrast to Min, Max contains the value of searches of the player in the week that was most searched. Therefore, considering that Max and Min contain opposite information and that Min stores information about "unconditional" fans, the Max value indicates the number of "conditional" fans, i.e. those fans who have become interested in the player as a result of a prominent action (positive or negative). Thus, this reflection suggests Min is a robust PI.

As for the variables related to player characteristics, the results shown in 5.2(b) are in line with those of previous researchers (Feess, Frick, and Muehlheusser 2004; Frick 2011; Steffen, Hans-Markus, and Henning 2014; Müller, Simons, and Weinmann 2017; Behravan and Razavi 2021), who found Age and Ct to be statistically significant variables and for which H and FT were also located not to affect the market value of players. However, concerning the position variable (P), the results obtained differ from previous analyses (Frick 2007; Müller, Simons, and Weinmann 2017; Singh and Lamba 2019) that highlighted this variable as statistically significant.

In the case of the variables related to player performance, the most prominent variables of the GBM (see 5.2(b)) coincide with previous results: PT (Feess, Frick, and Muehlheusser 2004; Müller, Simons, and Weinmann 2017; Singh and Lamba 2019), Gls (Steffen, Hans-Markus, and Henning 2014; Müller, Simons, and Weinmann 2017; Singh and Lamba 2019), Dr (Müller, Simons, and Weinmann 2017; Singh and Lamba 2019; Behravan and Razavi 2021), and S (Behravan and Razavi 2021). Furthermore, it is highlighted that the DP variable also showed a high contribution to the GBM model. However, this result could not compare with previous studies since it has not been analysed.

5.4 Discussion

Even though previous research has used popularity variables to predict transfer fees (Garcia-del-Barrio and Pujol 2007; Steffen, Hans-Markus, and Henning 2014; Müller, Simons, and Weinmann 2017; Hofmann et al. 2019; Singh and Lamba 2019), this chapter proposes using GT, a Google tool, to construct PIs that summarise and explain players popularity. Although GT is a valuable tool that contains information on Google searches (by player name, images, shopping, news, and YouTube), it has been underutilised or misused. GT has several advantages, such as the overview of player popularity, as it provides information over time, has no missing values and can avoid multicollinearity problems without data loss.

To our knowledge, this study is the first to propose a methodology for developing PIs by adding additional players to rescale GT information. It also suggests ranking players according to their "popularity level" to facilitate the comparison of players of unbalance popularity. Then, to test whether the proposed PIs are a valuable tool for predicting transfer fees, a database with 1428 players who competed during the 2018-2019 season in the "Big Five" (Spanish, Italian, German, English and French) was used.

In addition to the proposed methodology, throughout this chapter, we test our theory that the way in which Müller, Simons, and Weinmann (2017) calculated the GTA variable may not be completely effective. Thus, in addition to checking the influence of the PIs created, the performance of the methods using GTA (model 2) vs using the proposed PIs (model 3) was also compared. The results shown in Table 5.4. regarding model 2 coincide with those obtained by Müller, Simons, and Weinmann (2017) since the AIC rejected introducing the GTA variable in the MLR. Therefore, this study agrees with the one carried out by Müller, Simons, and Weinmann (2017) since they did not find the GTA was statistically significant. However, the PIs introduced in model 3 were statistically significant, specifically Min and Mdn. GBM and RF support the MLR results and highlight the Min variable's importance in players' market value for predicting the transfer fees. Thus, Table 5.5. also noteworthy the improvement in the prediction error in all the models when including the PIs compared to not having them (model 1) or including the GTA variable (model 2). Therefore, evaluating the performance of the methods and models demonstrates the usefulness of including PIs in predicting transfer fees. According to Table 5.5., comparing the RMSE of the GBM models (the best method), model 3 (5.10) had the lowest prediction error (€11,515,548), which meant a difference of €5,675,215 compared to model 1 (5.8). Thus, it is concluded that

using proposed PIs is a valuable way to improve the prediction of transfer fees for today's players.

Although previous researchers included popularity variables in their studies, most did not quantify how much the prediction error of transfer fees decreased when using popularity variables compared with not using them. Garcia-del-Barrio and Pujol (2007) developed two models, one in which they did not take popularity variables into account and another in which they did. However, they measured the improvement of the model through the squared goodness of fit of R-squared, i.e. they did not consider the RMSE. Müller, Simons, and Weinmann (2017) created four multilevel regression models but did not compare the RMSE of these models with each other but the RMSE of the best with the RMSE of Transfermarkt. Singh and Lamba (2019), although they used RMSE to compare the models, was a log-transformed dependent variable and did not show a measure of error for market value. Hofmann et al. (2019) considered the effect of players' brand image on their popularity. Even though they did not investigate the impact on market value, they highlighted its statistically significant effect.

In summary, recent research finds that popularity variables positively and statistically significantly affect the market value and help predict transfer fees. However, the researchers did not quantify the specific impact of incorporating popularity variables into the prediction. Therefore, this chapter becomes relevant to incorporate into the literature a novel methodology and a comprehensive analysis of the influence of PIs in predicting transfer fees.

5.5 Conclusion

The central conclusion of this chapter is that the built PIs improved the prediction of transfer fees. Furthermore, it has been confirmed that popularity has a noticeable impact on the market value of players. The importance and statistical significance of Min are highlighted since it has a positive effect on transfer fee prediction. Furthermore, we would like to point out that this chapter potentially contributes not only to the literature related to player valuation but also to any field of research that studies the effect of popularity.

Chapter 6

General Conclusions

This thesis used several machine learning and multivariate analysis procedures to model and tackle sports analytics problems commonly solved by applying classical models. In addition, new methodologies were developed to handle specific issues related to the use of eventing data. After an initial introduction to the areas of analysis in football and their categorisation according to the nature of the available data, the different contributions were presented in four parts. Chapter 2: proposes, fits, and compares multivariate statistical methods to predict the ranking of teams at the end of the season and to study the most crucial game actions to discriminate the teams' positions. Chapter 3: compares and shows the advantages of multivariate statistical techniques concerning the classical two-sample univariate test. Then, the best model to extract the game actions that contribute most to the team's success is chosen. Chapter 4: compares a classical model based on the double Poisson distribution with multivariate statistical techniques to select the best predictive model to discriminate match results and study the important game actions for team success. Chapter 5: Designs, develops and proposes a new methodology to calculate several novel popularity indicators useful to predict the market value of players. Finally, the last part outlines the thesis's conclusions, relevance, and future lines.

6.1 Achievement of the objectives

This section summarises the main findings of this thesis, demonstrating that the main objectives have been achieved.

Objective 1: Compare the effectiveness of the classical statistical techniques used so far with multivariate statistical and machine learning techniques

Multivariate statistical and machine learning techniques revealed as powerful approaches to improve and assist organisational decision-making. Chapter 2 and Chapter 3 presents PCA as a powerful and effective technique for carrying out the data preliminary exploratory analysis. Univariate approaches are commonly used to study the significance of the game actions in the teams' success (Rampinini et al. 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019). However, these techniques are primarily inefficient with several disadvantages: it is necessary to carry out as many statistical tests as there are variables in the data set, which can lead to multiple comparison problems due to the large number of hypothesis tests completed. Moreover, it does not provide an overview of the teams' behaviour since the game's actions are studied independently. In contrast, using PCA makes it possible to study all variables simultaneously, which allows us to obtain an overview of the teams' behaviour according to their label classification. In addition, it will enable us to detect outliers. Chapter 4 proposes RF and PLS-DA as substitutes for SRM to predict the outcome of matches. SRM is a parametric model based on the double Poisson distribution, commonly used to predict the match's outcome (Karlis and Ntzoufras 2009; Carpita, Ciavolino, and Pasca 2021; Pelechrinis and Winston 2021). However, SRM has some drawbacks. Mainly, it is a model that does not work with correlated variables. So, since it is necessary to avoid multicollinearity, exists the risk of eliminating essential variables useful to understand the behaviour of teams. In addition, since it suffers from multicollinearity, it is necessary to perform an initial variable selection step before performing the predictive analysis. Moreover, since SRM does not calculate the match outcome directly but the goal difference of the matches from the Home and Away statistics, its use requires having a data set that includes both statistics independently. It was concluded that the performance of the multivariate techniques (PLS-DA and RF) were better than that of SRM. In particular, PLS-DA was the method that provided the best results.

Objective 2: Implement predictive models based on multivariate statistical and machine learning techniques

Throughout the chapters of this thesis, different predictive models of machine learning or multivariate statistics have been proposed and developed to achieve the objectives. The aim of Chapter 2 was to propose several machine learning and multivariate statistical techniques (PLS-DA, CART, RF, Naive Bayes, and K-NN) to predict the ranking of the teams at the end of the season (bottom, middle, and top). In addition, a data balancing procedure was also performed to check if it improved the prediction. The results showed that RF was the model with the best predictive accuracy and that working with balanced data gave a higher average outcome for the validation set.

Even though the objective of Chapter 3 was to determine the game actions that most contributed to discriminating between positions (bottom and top), the prediction accuracy of the proposed methods was calculated. As a result, both RF and PLS-DA stood out as the technique with the best statistical classification performance, which performed better than LR.

As previously explained, in Chapter 4, RF and PLS-DA were proposed to predict the outcome of matches. As a result, RF and PLS-DA stood out for their higher prediction accuracy. Finally, in Chapter 5, RF and GBM, two machine learning techniques, were used to predict the market value of players sold in the summer during a season. The predictive accuracy of both models was compared with that of MLR. Both proposed machine learning techniques had better predictive accuracy than MLR, with GBM highlighting as the best of the methods.

Objective 3: Propose multivariate statistical and machine learning techniques to determine the variables that most influence the prediction results of these models

Two chapters were launched to fulfill the third objective. Chapter 3 presents the advantages of PCA, an unsupervised learning helpful method for exploratory data analysis, and RF and PLS-DA for confirmatory analysis. As discussed above, PCA gives researchers an overview of the teams' behaviour, providing clues about the determinant game actions to discriminate between the bottom and top teams. Furthermore, regarding the multivariate techniques studied, PLS-DA and RF selected many statistically significant variables to discriminate between both positions. Moreover, these results, especially those obtained by PLS-DA, supported those obtained by PCA in the exploratory analysis of the data. Note that the results obtained by these techniques stood out above the univariate methods used so far (Oberstone 2009; Lago-Peñas and Lago-Ballesteros 2010; Souza et al. 2019). As explained in

Chapter 4, SRM does not work with multicollinearity. Therefore, it was necessary to carry out a variable selection procedure to deal with the problem of highly correlated explanatory variables. Thus, in this chapter, we proposed combining RF, a machine learning technique, and SRM. Thus, the selection of the most influential variables and, therefore, susceptible to being introduced in the SRM was chosen by the RF. Moreover, the important game actions to discriminate between winning, drawing, and losing teams selected by RF and PLS-DA were also analysed.

Objective 4: Design, develop and propose a new methodology to calculate several indicators that summarise information about the popularity of players and that can be useful to predict their market value.

In Chapter 5 the main contribution was dedicated to achieving objective 4. The main contribution of Chapter 5 was to develop a novel procedure to obtain and calculate PIs. PIs summarise player popularity through the GT time series provided by Google. This methodology simplifies data collection and avoids getting redundant information and missing values. In addition, it was demonstrated that GT had been underused or misused and showed the correct way to use it. Finally, using popularity indicators was found to improve the prediction of transfer fees.

6.2 Future Lines

This PhD dissertation opens some future lines:

- Predict the result of the teams' classification at the end of the season from the prediction of the results of the football matches using the accumulated evolution of the game actions as explanatory variables.
- Repeat the study carried out along this thesis but using tracking data. Since football is a dynamic sport, combining the knowledge acquired through the study of the eventing and tracking data would allow us to obtain a global view of the game actions' effect on match results or teams' success.
- Develop an ensemble method from the prediction of several models to classify match results according to the majority vote. In this way, the problem observed by previous researchers draws results (maybe) overcome.

- Even though Chapter 5 is considered a significant literary contribution regarding GT indicators, the prediction error is high. This high error is (probably) because of the size of the used sample. It is considered of great interest to repeat the analysis by including players competing outside the Big Five, as researchers have yet to take these players into account so far.

Bibliography

Ajadi, Theo et al. (2022). *Restart: Football Money League*. Tech. rep. Deloitte.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705.

Allison, Paul D. (1999). *Multiple Regression: A primer*. Thousand Oaks: Pine Forge Press.

Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185. DOI: https://doi.org/10.2307/2685209.

Altmann, Stefan et al. (2021). "Match-related physical performance in professional soccer: Position or player specific?" In: *PloS one* 16.9, e0256695. DOI: https://doi.org/10.1371/journal.pone.0256695.

Barker, Matthew and William Rayens (2003). "Partial least squares for discrimination". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 17.3, pp. 166–173. DOI: https://doi.org/10.1002/cem.785.

Barnes, Sean L. and Magrét V. Bjarnadóttir (2016). "Great expectations: An analysis of major league baseball free agent performance". In: *Statistical*

Analysis and Data Mining: The ASA Data Science Journal 9.5, pp. 295–309. DOI: https://doi.org/10.1002/sam.11311.

Barua, Sukarna et al. (2015). "MWMOTE–Majority Weighted Minority Oversampling Technique for imbalanced data set learning". In: *IEEE Transactions on knowledge and data engineering* 26.2, pp. 405–425. DOI: https://doi.org/10.1109/TKDE.2012.232.

Beck, Marcus (2013). "Collinearity and stepwise VIF selection". In: *Retrieved from http://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/*.

Behravan, Iman and Seyed Mohammad Razavi (2021). "A novel machine learning method for estimating football players' value in the transfer market". In: *Soft Computing* 25.3, pp. 2499–2511. DOI: https://doi.org/10.1007/s00500-020-05319-3.

Bengtsson, Håkan, Jan Ekstrand, and Martin Hägglund (2013). "Muscle injury rates in professional football increase with fixture congestion: An 11-year follow-up of the UEFA Champions League injury study". In: *British Journal of Sports Medicine* 47.12, pp. 743–747. DOI: https://doi.org/10.1136/bjsports-2013-092383.

Bennett, Jay M. and John A. Flueck (1983). "An evaluation of major league baseball offensive performance models". In: *The American Statistician* 37.1, pp. 76–82. DOI: https://doi.org/10.2307/2685850.

Berry, William D., Stanley Feldman, and Stanley Feldman D. (1985). *Multiple Regression in Practice*. Newbury Park: Sage. DOI: https://doi.org/10.4135/9781412985208.

Bialkowski, Alina et al. (2014). "Large-scale analysis of soccer matches using spatiotemporal tracking data". In: *2014 IEEE International Conference on Data Mining*, pp. 725–730. DOI: https://doi.org/10.1109/ICDM.2014.133.

Boscá, José et al. (2009). "Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football". In: *Omega* 37.1, pp. 63–78. DOI: https://doi.org/10.1016/j.omega.2006.08.002.

Boulier, Bryan L. and Herman O. Stekler (2003). "Predicting the outcomes of National Football League games". In: *International Journal of Forecasting* 19.2, pp. 257–270. DOI: https://doi.org/10.1016/S0169-2070(01)00144-3.

Bradley, Paul S and Timothy D Noakes (2013). "Match running performance fluctuations in elite soccer: Indicative of fatigue, pacing or situational influences?" In: *Journal of Sports Sciences* 31.15, pp. 1627–1638. DOI: https://doi.org/10.1080/02640414.2013.796062.

Bradley, Paul S et al. (2010). "High-intensity activity profiles of elite soccer players at different performance levels". In: *The Journal of Strength & Conditioning Research* 24.9, pp. 2343–2351. DOI: https://doi.org/10.1519/JSC.0b013e3181aeb1b3.

Bradley, Paul S et al. (2013). "Match performance and physical capacity of players in the top three competitive standards of English professional soccer". In: *Human Movement Science* 32.4, pp. 808–821. DOI: https://doi.org/10.1016/j.humov.2013.06.002.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140. DOI: https://doi.org/10.1007/BF00058655.

— (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

Breiman, Leo et al. (1984). *Classification and regression trees*. Monterrey, CA: Routledge. DOI: https://doi.org/10.1201/9781315139470.

Bridge, Tim et al. (2023). *Deloitte Football Money League 2023: Get up, stand up*. Tech. rep. Deloitte.

Bryson, Alex, Bernd Frick, and Robert Simmons (2013). "The returns to scarce talent: Footedness and player remuneration in European soccer". In: *Journal of Sports Economics* 14.6, pp. 606–628. DOI: https://doi.org/10.1177/1527002511435118.

Budsaba, Kamon, Charles E. Smith, and Jim E. Riviere (2000). "Compass plots: A combination of star plot and analysis of means to visualize

significant interactions in complex toxicology studies". In: *Toxicology Methods* 10.4, pp. 313–332. DOI: https://doi.org/10.1080/105172300750048764.

Carey, Mark and Mladen Sormaz (2019). "Clustering playing styles in the modern day full-back". In: *Retrieved from https://www.statsperform.com/resource/clustering-playing-styles-in-the-modern-day-full-back/*.

Carling, Chris et al. (2016). "The impact of short periods of match congestion on injury risk and patterns in an elite football club". In: *British Journal of Sports Medicine* 50.12, pp. 764–768. DOI: https://doi.org/10.1136/bjsports-2015-095501.

Carling, Christopher, Franck Le Gall, and Gregory Dupont (2012). "Analysis of repeated high-intensity running performance in professional soccer". In: *Journal of Sports Sciences* 30.4, pp. 325–336. DOI: https://doi.org/10.1080/02640414.2011.652655.

Carpita, Maurizio, Enrico Ciavolino, and Paola Pasca (2019). "Exploring and modelling team performances of the Kaggle European Soccer database". In: *Statistical Modelling* 19.1, pp. 74–101. DOI: https://doi.org/10.1177/1471082X18810971.

— (2021). "Players' role-based performance composite indicators of soccer teams: A statistical perspective". In: *Social Indicators Research* 156.2, pp. 815–830. DOI: https://doi.org/10.1007/s11205-020-02323-w.

Carpita, Maurizio and Silvia Golia (2021). "Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues". In: *Journal of Applied Statistics* 48.9, pp. 1696–1711. DOI: https://doi.org/10.1080/02664763.2020.1772210.

Carpita, Maurizio et al. (2015). "Discovering the Drivers of Football Match Outcomes with Data Mining". In: *Quality Technology & Quantitative Management* 12.4, pp. 561–577. DOI: https://doi.org/10.1080/16843703.2015.11673436.

— (2021). *Football Mining with R*. Elsevier.

Casal, Claudio A. et al. (2019). "Possession in football: More than a quantitative aspect–A mixed method study". In: *Frontiers in Psychology* 10, p. 501. DOI: https://doi.org/10.3389/fpsyg.2019.00501.

Castellano, Julen, David Casamichana, and Carlos Lago (2012). "The use of match statistics that discriminate between successful and unsuccessful soccer teams". In: *Journal of Human Kinetics* 31.2012, pp. 137–147. DOI: https://doi.org/10.2478/v10078-012-0015-7.

Chadwick, Henry (1860). *Beadle's Dime Baseball Player*. New York: Irwin P. Beadle & Co.

Chong, Il-Gyo and Chi-Hyuck Jun (2005). "Performance of some variable selection methods when multicollinearity is present". In: *Chemometrics and intelligent laboratory systems* 78.1-2, pp. 103–112. DOI: https://doi.org/10.1016/j.chemolab.2004.12.011.

Cintia, Paolo et al. (2015). "The harsh rule of the goals: Data-driven performance indicators for football teams". In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. DOI: https://doi.org/10.1109/DSAA.2015.7344823.

Collet, Christian (2013). "The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010". In: *Journal of Sports Sciences* 31.2, pp. 123–136. DOI: https://doi.org/10.1080/02640414.2012.727455.

Cordón, Ignacio et al. (2018). "Imbalance: Oversampling algorithms for imbalanced classification in R". In: *Knowledge-Based Systems* 161.329, pp. 329–341. DOI: https://doi.org/10.1016/j.knosys.2018.07.035.

Cotteleer, Mark and Brenna Sniderman (2019). *Forces of change: Industry 4.0*. Tech. rep. Deloitte.

Coutinho, Diogo et al. (2017). "Mental fatigue and spatial references impair soccer players' physical and tactical performances". In: *Frontiers in Psychology* 8, p. 1645. DOI: https://doi.org/10.3389/fpsyg.2017.01645.

Cox, David R. (1958). "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2, pp. 215–232. DOI: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

de Mendiburu, Felipe (2021). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.3-5.

Decroos, Tom et al. (2019). "Actions speak louder than goals: Valuing player actions in soccer". In: *Proceedings of the 25th ACM SIGKDD international cconference on knowledge discovery & data mining*, pp. 1851–1861. DOI: https://doi.org/10.1145/3292500.3330758.

Dellal, Alexandre et al. (2015). "The effects of a congested fixture period on physical performance, technical activity and injury rate during matches in a professional soccer team". In: *British Journal of Sports Medicine* 49.6, pp. 390–394. DOI: https://doi.org/10.1136/bjsports-2012-091290.

Di Salvo, Valter et al. (2007). "Performance characteristics according to playing position in elite soccer". In: *International Journal of Sports Medicine* 28.3, pp. 222–227. DOI: https://doi.org/10.1055/s-2006-924294.

Di Salvo, Valter et al. (2010). "Sprinting analysis of elite soccer players during European Champions League and UEFA Cup matches". In: *Journal of Sports Sciences* 28.14, pp. 1489–1494. DOI: https://doi.org/10.1080/02640414.2010.521166.

Di Salvo, Valter et al. (2013). "Match performance comparison in top English soccer leagues". In: *International Journal of Sports Medicine* 34.6, pp. 526–532. DOI: https://doi.org/10.1055/s-0032-1327660.

Dolci, Filippo et al. (2020). "Physical and energetic demand of soccer: A brief review". In: *Strength & Conditioning Journal* 42.3, pp. 70–77. DOI: https://doi.org/10.1519/SSC.0000000000000533.

Driblab (2020). "Football Player Analysis". In: *Retrieved from https://www.driblab.com/serviciosdriblab/ player-analysis/*.

Duch, Jordi, Joshua S Waitzman, and Luís A Nunes Amaral (2010). "Quantifying the performance of individual players in a team activity". In: *PLoS One* 5.6, e10937. DOI: https://doi.org/10.1371/journal.pone.0010937.

Dupont, Gregory et al. (2010). "Effect of 2 soccer matches in a week on physical performance and injury rate". In: *The American Journal of Sports Medicine* 38.9, pp. 1752–1758. DOI: https://doi.org/10.1177/0363546510361236.

Edgington, Eugene S. and Patrick Onghena (2007). *Randomization tests*. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781420011814.

Ehrmann, Fabian E. et al. (2016). "GPS and injury prevention in professional soccer". In: *The Journal of Strength & Conditioning Research* 30.2, pp. 360–367. DOI: https://doi.org/10.1519/JSC.0000000000001093.

Ekstrand, Jan, Markus Waldén, and Martin Hägglund (2016). "Hamstring injuries have increased by 4% annually in men's professional football, since 2001: A 13-year longitudinal analysis of the UEFA Elite Club injury study". In: *British Journal of Sports Medicine* 50.12, pp. 731–737. DOI: https://doi.org/10.1136/bjsports-2015-095359.

Eriksson, Lennart et al. (2013). *Multi-and megavariate data analysis basic principles and applications*. Vol. 1. Umetrics Academy. DOI: https://doi.org/10.1002/cem.713.

Espita-Escuer, Manuel and Lucía Isabel García-Cebrían (2008). "Measuring the efficiency of spanish firstdivision soccer teams". In: *European Sport Management Quarterly* 8.3, pp. 229–246. DOI: https://doi.org/10.1177/1527002503258047.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8, pp. 861–874. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

Feess, Eberhard, Bernd Frick, and Gerd Muehlheusser (2004). "Legal restrictions on buyout fees: Theory and evidence from German soccer". In: *Retrieved from https://ssrn.com/abstract=562445 or*

http://dx.doi.org/10.2139/ssrn.562445. DOI: https://doi.org/10.2139/ssrn.562445.

Felipe, Jose Luis et al. (2020). "Money talks: Team variables and player positions that most influence the market value of professional male footballers in Europe". In: *Sustainability* 12.9, p. 3709. DOI: https://doi.org/10.3390/su12093709.

Ferrer, Alberto (2007). "Multivariate statistical process control based on principal component analysis (MSPC-PCA): Some reflections and a case study in an autobody assembly process". In: *Quality Engineering* 19.4, pp. 311–325. DOI: https://doi.org/10.1080/08982110701621304.

Frick, Bernd (2007). "The football players' labor market: Empirical evidence from the major European Leagues". In: *Scottish Journal of Political Economy* 54.3, pp. 422–446. DOI: https://doi.org/10.1111/j.1467-9485.2007.00423.x.

— (2011). "Performance, salaries, and contract length: Empirical evidence from German soccer". In: *International Journal of Sport Finance* 6.2, p. 87.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine". In: *Annals of statistics* 29.5, pp. 1189–1232. DOI: https://doi.org/10.1214/aos/1013203451.

— (2002). "Stochastic gradient boosting". In: *Computational Statistics & Data Analysis* 38.4, pp. 367–378. DOI: https://doi.org/10.1016/S0167-9473(01)00065-2.

Garcia-del-Barrio, Pedro and Francesc Pujol (2007). "Hidden monopsony rents in winner-take-all markets? Sport and economic contribution of Spanish soccer players". In: *Computational Statistics & Data Analysis* 28.1, pp. 57–70. DOI: https://doi.org/10.1002/mde.1313.

García-Unanue, Jorge et al. (2018). "Influence of contextual variables and the pressure to keep category on physical match performance in soccer players". In: *PloS One* 13.9, e0204256. DOI: https://doi.org/10.1371/journal.pone.0204256.

Goncalves, Bruno et al. (2017). "Exploring team passing networks and player movement dynamics in youth association football". In: *PLoS One* 12.1, e0171156. DOI: https://doi.org/10.1371/journal.pone.0171156.

González-Rodenas, Joaquín et al. (2020). "Playing tactics, contextual variables and offensive effectiveness in English Premier League soccer matches. A multilevel analysis". In: *PLoS One* 15.2, e0226978. DOI: https://doi.org/10.1371/journal.pone.0226978.

Gottfries, Johan et al. (1995). "Diagnosis of dementias using partial least squares discriminant analysis". In: *Dementia and Geriatric Cognitive Disorders* 6.2, pp. 83–88. DOI: https://doi.org/10.1159/000106926.

Gregson, Warren et al. (2010). "Match-to-match variability of high-speed activities in premier league soccer". In: *International Journal of Sports Medicine* 31.4, pp. 237–242. DOI: https://doi.org/10.1055/s-0030-1247546.

Gyimesi, András and Dániel Kehl (2021). "Relative age effect on the market value of elite European football players: A balanced sample approach". In: *European Sport Management Quarterly*. 0.0, pp. 1–17. DOI: https://doi.org/10.1080/16184742.2021.1894206.

Hawkins, Richard D and Colin W Fuller (1999). "A prospective epidemiological study of injuries in four English professional football clubs". In: *British Journal of Sports Medicine* 33.3, pp. 196–203. DOI: http://dx.doi.org/10.1136/bjsm.33.3.196.

Hawkins, Richard D et al. (2001). "The association football medical research programme: An audit of injuries in professional football". In: *British Journal of Sports Medicine* 35.1, pp. 43–47. DOI: https://doi.org/10.1136/bjsm.35.1.43.

He, Miao, Ricardo Cachucho, and Arno J. Knobbe (2015). "Football player's performance and market value". In: *Proceedings of the 2nd Workshop of Sports Analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Porto, Portugal, pp. 87–95.

Hervé, Maxime (2020). *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*.

Hofmann, Julian et al. (2019). "Talent or popularity: What drives market value and brand image for human brands?" In: *Journal of Business Research* 124, pp. 748–758. DOI: https://doi.org/10.1016/j.jbusres.2019.03.045.

Höskuldsson, Agnar (1988). "PLS regression methods". In: *Journal of chemometrics* 2.3, pp. 211–228. DOI: https://doi.org/10.1002/cem.1180020306.

Ibáñez, Sergio J. et al. (2008). "Basketball game-related statistics that discriminate between teams? season-long success". In: *European Journal of Sport Science* 8.6, pp. 369–372. DOI: https://doi.org/10.1080/17461390802261470.

James, Nic, Stephen D Mellalieu, and Chris Hollely (2002). "Analysis of strategies in soccer as a function of European and domestic competition". In: *International Journal of Performance Analysis in Sport* 2.1, pp. 85–103. DOI: https://doi.org/10.1080/24748668.2002.11868263.

Japkowicz, Nathalie (2000). "Learning from imbalanced data sets: A comparison of various strategies". In: *AAAI workshop on learning from imbalanced data sets*. Vol. 68. AAAI Press Menlo Park, CA, pp. 10–15.

Johnston, Ron, Kelvyn Jones, and David Manley (2018). "Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour". In: *Quality & Quantity* 52.4, pp. 1957–1976. DOI: https://doi.org/10.1007/s11135-017-0584-6.

Jombart, Thibaut, Sébastien Devillard, and François Balloux (2010). "Discriminant analysis of principal components: A new method for the analysis of genetically structured populations". In: *BMC genetics* 11.1, pp. 1–15. DOI: https://doi.org/10.1186/1471-2156-11-94.

Jones, PD, Nic James, and Stephen D Mellalieu (2004). "Possession as a performance indicator in soccer". In: *International Journal of*

Performance Analysis in Sport 4.1, pp. 98–102. DOI: https://doi.org/10.1080/24748668.2004.11868295.

Jones, Ross Nicholas et al. (2019). "The influence of short-term fixture congestion on position specific match running performance and external loading patterns in English professional soccer". In: *Journal of Sports Sciences* 37.12, pp. 1338–1346. DOI: https://doi.org/10.1080/02640414.2018.1558563.

Karlis, Dimitris and Ioannis Ntzoufras (2009). "Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference". In: *IMA Journal of Management Mathematics* 20.2, pp. 133–145. DOI: https://doi.org/10.1093/imaman/dpn026.

Kearns, Michael and Leslie Valiant (1988). *Learning Boolean formulae or finite automata is as hard as factoring*. Tech. rep. TR 14-88. Massachusetts (USA): Harvard University Aiken Computation Laboratory.

— (1994). "Cryptographic limitations on learning Boolean formulae and finite automata". In: *Journal of the ACM (JACM)* 41.1, pp. 67–95. DOI: https://doi.org/10.1145/174644.174647.

Kempe, Matthias et al. (2014). "Possession vs. direct play: Evaluating tactical behavior in elite soccer". In: *International Journal of Sports Science* 4.6A, pp. 35–41. DOI: https://doi.org/10.5923/s.sports.201401.05.

Kim, Hyunjoong and Wei-Yin Loh (2001). "Classification trees with unbiased multiway splits". In: *Journal of the American Statistical Association* 96.454, pp. 589–604. DOI: https://doi.org/10.1198/016214501753168271.

Knijnenburg, Theo A et al. (2009). "Fewer permutations, more accurate P-values". In: *Bioinformatics* 25.12, pp. i161–i168. DOI: https://doi.org/10.1093/bioinformatics/btp211.

Knutson, Ted (2020). "StatsBomb presenta sus nuevas visualizaciones". In: *Retrieved from https://statsbomb.com/es/noticias/statsbomb-presenta-sus-nuevas-visualizaciones/*.

Kolence, Kenneth W. and Philip J. Kiviat (1973). "Software unit profiles & Kiviat figures". In: *ACM SIGMETRICS Performance Evaluation Review* 2.3, pp. 2–12. DOI: https://doi.org/10.1145/1041613.1041614.

Kucheryavskiy, Sergey (2020). "mdatools — R package for chemometrics". In: *Chemometrics and Intelligent Laboratory Systems* 198.

Kuhn, Max (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Kutner, Michael H. et al. (2005). *Applied Linear Statistical Models*. McGraw Hill Irwin, New York. NY, p. 409.

Kuzmits, Frank E. and Arthur J. Adams (2008). "The NFL combine: Does it predict performance in the National Football League?" In: *The Journal of Strength & Conditioning Research* 22.6, pp. 1721–1727. DOI: https://doi.org/10.1519/JSC.0b013e318185f09d.

Lago, Carlos (2009). "The influence of match location, quality of opposition, and match status on possession strategies in professional association football". In: *Journal of Sports Sciences* 27.13, pp. 1463–1469. DOI: https://doi.org/10.1080/02640410903131681.

Lago, Carlos and Rafael Martín (2007). "Determinants of possession of the ball in soccer". In: *Journal of Sports Sciences* 25.9, pp. 969–974. DOI: https://doi.org/10.1080/02640410600944626.

Lago, Carlos et al. (2010). "The effects of situational variables on distance covered at various speeds in elite soccer". In: *European Journal of Sport Science* 10.2, pp. 103–109. DOI: https://doi.org/10.1080/17461390903273994.

Lago-Peñas, Carlos (2010). "Ball Possession Strategies in Elite Soccer According to the Evolution of the Match-Score: The Influence of Situational Variables". In: *Journal of Sports Sciences* 25, pp. 93–100. DOI: https://doi.org/10.2478/v10078-010-0036-z.

Lago-Peñas, Carlos and Joaquín Lago-Ballesteros (2010). "Performance in team sports: Identifying the keys to success in soccer". In: *Journal of*

Human Kinetics 25.1, pp. 85–91. DOI: https://doi.org/10.2478/v10078-010-0035-0.

Lago-Peñas, Carlos, Joaquín Lago-Ballesteros, and Ezequiel Rey (2011). "Differences in performance indicators between winning and losing teams in the UEFA Champions League". In: *Journal of Human Kinetics* 27.1, pp. 135–146. DOI: https://doi.org/10.2478/v10078-011-0011-3.

Lazraq, Aziz, Robert Cléroux, and Jean-Pierre Gauchi (2003). "Selecting both latent and explanatory variables in the PLS1 regression model". In: *Chemometrics and Intelligent Laboratory Systems* 66.2, pp. 117–126. DOI: https://doi.org/10.1016/S0169-7439(03)00027-3.

Levene, Howard (1961). "Robust tests for equality of variances". In: *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292.

Lewis, Michael (2004). *Moneyball: The art of winning an unfair game*. New York: WW Norton & Company.

Ley, Christophe, Tom Van Wiele, and Hans Van Eetvelde (2017). "Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches". In: *Statistical Modelling* 19.1, pp. 55–73. DOI: https://doi.org/10.1177/1471082X18817650.

Liaw, Andy and Matthew Wiener (2002). "Classification and regression by randomForest". In: *R news* 2.3, pp. 18–22.

Link, Daniel (2018). *Data Analytics in Professional Soccer: Performance Analysis Based on Spatiotemporal Tracking Data*. Munich: Springer Fachmedien Wiesbaden GmbH.

Link, Daniel and Martin Hoernig (2017). "Individual ball possession in soccer". In: *PloS One* 12.7, e0179953. DOI: https://doi.org/10.1371/journal.pone.0179953.

Link, Daniel, Steffen Lang, and Philipp Seidenschwarz (2016). "Real time quantification of dangerousity in football using spatiotemporal tracking

data". In: *PLoS One* 11.12, e0168768. DOI: https://doi.org/10.1371/journal.pone.0168768.

Liu, Hongyou et al. (2015a). "Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup". In: *Journal of Sports Sciences* 33.12, pp. 1205–1213. DOI: https://doi.org/10.1080/02640414.2015.1022578.

Liu, Hongyou et al. (2015b). "Performance profiles of football teams in the UEFA Champions League considering situational efficiency". In: *International Journal of Performance Analysis in Sport* 15.1, pp. 371–390. DOI: https://doi.org/10.1080/24748668.2015.11868799.

Liu, Hongyou et al. (2016). "Technical performance and match-to-match variation in elite football teams". In: *Journal of Sports Sciences* 34.6, pp. 509–518. DOI: https://doi.org/10.1080/02640414.2015.1117121.

Lucey, Patrick et al. (2013). "Assessing team strategy using spatiotemporal data". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1366–1374. DOI: https://doi.org/10.1145/2487575.2488191.

Maron, Melvin E. (1961). "Automatic indexing: An experimental inquiry". In: *Journal of the ACM (JACM)* 8.3, pp. 404–417. DOI: https://doi.org/10.1145/321075.321084.

Martín-García, Andrés et al. (2018). "Positional differences in the most demanding passages of play in football competition". In: *Journal of Sports Science & Medicine* 17.4, p. 563.

Matthews, Brian W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2, pp. 442–451. DOI: https://doi.org/10.1016/0005-2795(75)90109-9.

Memmert, Daniel, Koen APM Lemmink, and Jaime Sampaio (2017). "Current approaches to tactical performance analyses in soccer using position data". In: *Sports Medicine* 47.1, pp. 1–10. DOI: https://doi.org/10.1007/s40279-016-0562-5.

Migliorati, Manlio (2020). "Detecting drivers of basketball successful games: An exploratory study with machine learning algorithms". In: *Electronic Journal of Applied Statistical Analysis* 13.2, pp. 454–473. DOI: https://doi.org/10.1285/i20705948v13n2p454.

Mitchell, Ryan (2018). *Web scraping with Python: Collecting more data from the modern web*. O'Reilly Media, Inc.

Mohr, Magni, Peter Krustrup, and Jens Bangsbo (2003). "Match performance of high-standard soccer players with special reference to development of fatigue". In: *Journal of Sports Sciences* 21.7, pp. 519–528. DOI: https://doi.org/10.1080/0264041031000071182.

Müller, Oliver, Alexander Simons, and Markus Weinmann (2017). "Beyond crowd judgments: Data-driven estimation of market value in association football". In: *European Journal of Operational Research* 263.2, pp. 611–624. DOI: https://doi.org/10.1016/j.ejor.2017.05.005.

Nelder, John Ashworth and Robert W.M. Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384. DOI: https://doi.org/10.2307/2344614.

Neter, John, William Wasserman, and Michael Kutner (1989). *Applied Linear Regression Models, 2nd Edition*. Irwin: McGraw-Hill.

Nisbet, Robert, John Elder, and Gary D. Miner (2009). *Handbook of statistical analysis and data mining applications*. Academic press. DOI: https://doi.org/10.1016/C2012-0-06451-4.

Noçairi, Hicham et al. (2016). "Improving stacking methodology for combining classifiers: Applications to cosmetic industry". In: *Electronic Journal of Applied Statistical Analysis* 9.2, pp. 340–361. DOI: https://doi.org/10.1285/i20705948v9n2p340.

Oberstone, Joel (2009). "Differentiating the top English premier league football clubs from the rest of the pack: Identifying the keys to success". In: *Journal of Quantitative Analysis in Sports* 5.3. DOI: https://doi.org/10.2202/1559-0410.1183.

Opitz, David and Richard Maclin (1999). "Popular ensemble methods: An empirical study". In: *Journal of artificial intelligence research* 11, pp. 169–198. DOI: https://doi.org/10.1613/jair.614.

Owen, Adam L. et al. (2015). "Heart rate–based training intensity and its impact on injury incidence among elite-level professional soccer players". In: *The Journal of Strength & Conditioning Research* 29.6, pp. 1705–1712. DOI: https://doi.org/10.1519/JSC.0000000000000810.

Paluszyńska, Aleksandra (2017). "Understanding random forests with randomForestExplainer". In: *The Comprehensive R Archive Network*.

Pappalardo, Luca et al. (2019). "A public data set of spatio-temporal match events in soccer competitions". In: *Scientific Data* 6.1, pp. 1–15. DOI: https://doi.org/10.1038/s41597-019-0247-7.

Pearson, Karl (1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. DOI: https://doi.org/10.1080/14786440109462720.

Pelechrinis, Konstantinos and Wayne Winston (2021). "A Skellam regression model for quantifying positional value in soccer". In: *Journal of Quantitative Analysis in Sports* 17.3, pp. 187–201. DOI: https://doi.org/10.1515/jqas-2019-0122.

Peñas, Carlos et al. (2010). "Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league". In: *Journal of Sports Science & Medicine* 9.2, p. 288.

Peres-Neto, Pedro R., Donald A. Jackson, and Keith M. Somers (2005). "How many principal components? Stopping rules for determining the number of non-trivial axes revisited". In: *Computational Statistics & Data Analysis* 49.4, pp. 974–997. DOI: https://doi.org/10.1016/j.csda.2004.06.015.

Poli, Raffaele, Loïc Ravenel, and Roger Besson (2022). *Financial analysis of big-5 league clubs' transfers*. Tech. rep. CIES Football Observatory.

Power, Paul et al. (2017). "Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1605–1613. DOI: https://doi.org/10.1145/3097983.3098051.

Premier, League (2021). "Premier League value of central payments to clubs 2020/21". In: *Retrieved from https://www.premierleague.com/news/2222377*.

Quenouille, Maurice Henri (1949). "Approximate tests of correlation in time-series". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 11.1, pp. 68–84. DOI: https://doi.org/10.1111/j.2517-6161.1949.tb00023.x.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rampinini, Ermanno et al. (2007). "Variation in top level soccer match performance". In: *International Journal of Sports Medicine* 28.12, pp. 1018–1024. DOI: https://doi.org/10.1055/s-2007-965158.

Rampinini, Ermanno et al. (2009). "Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue and competitive level". In: *Journal of Science and Medicine in Sport* 12.1, pp. 227–233. DOI: https://doi.org/10.1016/j.jsams.2007.10.002.

Refaeilzadeh, Payam, Lei Tang, and Huan Liu (2009). "Cross-Validation". In: *Encyclopedia of Database Systems*. Springer US, pp. 532–538. DOI: https://doi.org/10.1007/978-0-387-39940-9_565.

Rogers, Simon (2016). "What is Google Trends data? And what does it mean?" In: *Retrieved from https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8*.

Rohart, Florian et al. (2017). "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLoS Computational Biology* 13.11, e1005752. DOI: https://doi.org/10.1371/journal.pcbi.1005752.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA.

Saary, M. Joan (2008). "Radar plots: A useful way for presenting multivariate health care data". In: *Journal of Clinical Epidemiology* 61.4, pp. 311–317. DOI: https://doi.org/10.1016/j.jclinepi.2007.04.021.

Sandri, Marco and Paola Zuccolotto (2010). "Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms". In: *Statistics and Computing* 20, pp. 393–407. DOI: https://doi.org/10.1007/s11222-009-9132-0.

Santos, Miriam Seoane et al. (2018). "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]". In: *IEEE Computational Intelligence Magazine* 13.4, pp. 59–76. DOI: https://doi.org/10.1109/MCI.2018.2866730.

Schapire, Robert E. (1990). "The strength of weak learnability". In: *Machine learning* 5, pp. 197–227. DOI: https://doi.org/10.1007/BF00116037.

Schauberger, Gunther, Andreas Groll, and Gerhard Tutz (2017). "Analysis of the importance of on-field covariates in the German Bundesliga". In: *Journal of Applied Statistics* 45.9, pp. 1–18. DOI: https://doi.org/10.1080/02664763.2017.1383370.

Shapiro, Samuel Sanford and Martin B. Wilk (1965). "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3/4, pp. 591–611. DOI: https://doi.org/10.2307/2333709.

Sheather, Simon (2009). *A modern approach to regression with R*. Springer Science & Business Media. DOI: https://doi.org/10.1007/978-0-387-09608-7.

Sievert, Carson (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Sing, Tobias et al. (2005). "ROCR: visualizing classifier performance in R". In: *Bioinformatics* 21.20, p. 7881. DOI: https://doi.org/10.1093/bioinformatics/bti623.

Singh, Prabhnoor and Puneet Singh Lamba (2019). "Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players". In: *Journal of Discrete Mathematical Sciences and Cryptography* 22.2, pp. 113–126. DOI: https://doi.org/10.1080/09720529.2019.1576333.

Skellam, J. Gordon (1946). "The frequency distribution of the difference between two Poisson variates belonging to different populations". In: *Journal of the Royal Statistical Society. Series A (General)* 109.3, pp. 296–296.

Smith, Mitchell R et al. (2017). "Impact of mental fatigue on speed and accuracy components of soccer-specific skills". In: *Science and Medicine in Football* 1.1, pp. 48–52. DOI: https://doi.org/10.1080/02640414.2016.1252850.

Smithies, Tim D. et al. (2021). "A Random Forest approach to identify metrics that best predict match outcome and player ranking in the esport Rocket League". In: *Scientific Reports* 11.1, pp. 1–12. DOI: https://doi.org/10.1038/s41598-021-98879-9.

Souza, Diego Brito de et al. (2019). "An extensive comparative analysis of successful and unsuccessful football teams in LaLiga". In: *Frontiers in Psychology*, p. 2566. DOI: https://doi.org/10.3389/fpsyg.2019.02566.

Steffen, Herm, Callsen-Bracker Hans-Markus, and Kreis Henning (2014). "When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community". In: *Sport Management Review* 17.4, pp. 484–492. DOI: https://doi.org/10.1016/j.smr.2013.12.006.

Stølen, Tomas et al. (2005). "Physiology of soccer". In: *Sports Medicine* 35.6, pp. 501–536. DOI: https://doi.org/10.2165/00007256-200535060-00004.

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC Bioinformatics* 8.1, pp. 1–21. DOI: https://doi.org/10.1186/1471-2105-8-25.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. DOI: https://doi.org/10.1073/pnas.0506580102.

Sun, Xiao-Ming et al. (2012). "Combining bootstrap and uninformative variable elimination: Chemometric identification of metabonomic biomarkers by nonparametric analysis of discriminant partial least squares". In: *Chemometrics and Intelligent Laboratory Systems* 115, pp. 37–43. DOI: https://doi.org/10.1016/j.chemolab.2012.04.006.

Szymańska, Ewa et al. (2012). "Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies". In: *Metabolomics* 8.1, pp. 3–16. DOI: https://doi.org/10.1007/s11306-011-0330-3.

Taylor, Joseph B. et al. (2008). "The influence of match location, quality of opposition, and match status on technical performance in professional association football". In: *Journal of Sports Sciences* 26.9, pp. 885–895. DOI: https://doi.org/10.1080/02640410701836887.

Tenga, Albin et al. (2010). "Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches". In: *Journal of Sports Sciences* 28.3, pp. 245–255. DOI: https://doi.org/10.1080/02640410903502766.

Tusher, Virginia Goss, Robert Tibshirani, and Gilbert Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response". In: *Proceedings of the National Academy of Sciences* 98.9, pp. 5116–5121. DOI: https://doi.org/10.1073/pnas.091062498.

Vigne, Gregory et al. (2010). "Activity profile in elite Italian soccer team". In: *International Journal of Sports Medicine* 31.5, pp. 304–310. DOI: https://doi.org/10.1055/s-0030-1248320.

Wang, Qing et al. (2015). "Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA, pp. 2197–2206. DOI: https://doi.org/10.1145/2783258.2788577.

Welch, Bernard L. (1947). "The generalization of "Students's" problem when several different population variances are involved". In: *Biometrika* 34.1-2, pp. 28–35. DOI: https://doi.org/10.2307/2332510.

Westerhuis, Johan A. et al. (2008). "Assessment of PLSDA cross validation". In: *Metabolomics* 4.1, pp. 81–89. DOI: https://doi.org/10.1007/s11306-007-0099-6.

Whitehead, Sarah et al. (2021). "The use of technical-tactical and physical performance indicators to classify between levels of match-play in elite rugby league". In: *Science and Medicine in Football* 5.2, pp. 121–127. DOI: https://doi.org/10.1080/24733938.2020.1814492.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4.

Wilcoxon, Frank (1945). "Individual comparisons by ranking methods". In: *Biometrics* 1.6, pp. 80–83. DOI: https://doi.org/10.2307/3001968.

Wold, Svante, Kim Esbensen, and Paul Geladi (1986). "Partial least-squares regression: A tutorial". In: *Analytica Chimica Acta* 185, pp. 1–17. DOI: https://doi.org/10.1016/0003-2670(86)80028-9.

— (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52. DOI: https://doi.org/10.1016/0169-7439(87)80084-9.

Wold, Svante, Erik Johansson, Marina Cocchi, et al. (1993). "PLS: partial least squares projections to latent structures". In: *From 3D QSAR in Drug Design: Theory, Methods and Applications*. Kubinyi H (eds.). ESCOM Science Publishers, pp. 523–550. DOI: https://doi.org/10.1002/wics.51.

Worley, Bradley and Robert Powers (2013). "Multivariate analysis in metabolomics". In: *Current Metabolomics* 1.1, pp. 92–107. DOI: https://doi.org/10.2174/2213235X11301010092.

Wright, Marvin N. and Andreas Ziegler (2017). "Ranger: A fast implementation of random forests for high dimensional data in C++ and

R". In: *Journal of Statistical Software* 77.1, pp. 1–17. DOI: https://doi.org/10.18637/jss.v077.i01.

Zambom-Ferraresi, Fabíola et al. (2017). "Performance evaluation in the UEFA Champions League". In: *Journal of Sports Economics* 18.5, pp. 448–470. DOI: https://doi.org/10.1177/1527002515588135.

Zimmermann, Albrecht (2016). "Basketball predictions in the NCAAB and NBA: Similarities and differences". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9.5, 350?364. DOI: https://doi.org/10.1002/sam.11319.

Zuccolotto, Paola, Marica Manisera, and Marco Sandri (2018). "Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions". In: *International Journal of Sports Science & Coaching* 13.4, pp. 569–589. DOI: https://doi.org/10.1177/1747954117737492.

Abbreviations and acronyms

2CV: double cross validation, 27

ANOVA: analysis of variance, 4

Aw: away teams, 88

Aw_G: number of goals scored by the away teams, 90

AUC: area under curve, 52

CART: classification and regression trees, 20

CV: cross validation, 27

CF: conversion factor, 128

CCF: cumulative conversion factor, 129

FN: false negative, 29

FP: false positive, 29

GBM: gradient boosting machine, 131

GPS: global positioning system, 2

GT: Google Trends, 126

GTA: Google Trends time series average, 126

GTN: normalised Google Trends time series, 128

H: home teams, 88

H_G: number of goals scored by the home teams, 90

KNN: K-nearest neighbours, 20

LV: latent variables, 27

MDA: mean decrease accuracy, 50

MDG: mean decrease Gini, 50

MCC: Matthews correlation coefficient, 27

MLR: multiple linear regression, 127

MWMOTE: majority weighted minority oversampling, 25

OOB: out-of-bag, 25

PCA: principal component analysis, 20

PC: principal component, 22

PI: popularity indicator, 126

PLS: partial least squares, 24

PLS-DA: partial least squares discriminant analysis, 20

ROC: receiver operating characteristic, 52

Ref: Reference player, 128

RF: random forest, 20

SPE: squared prediction error, 22

TN: True negatives, 29

TP: true positive, 29

VC1: random division of the training set in the 2CV, 27

VC2: random division of the database in 2CV, 27

VIF: variance inflation factor, 51

VIP: variable influence on projection, 49

Z: goals diference, 90

Parameters and nomenclature

A: subspace of the PCs, 22

Cp: complexity parameter to control the growth of the decision tree, 27

E: residual matrix, 22

F: number of subgroups into which the database is divided in the CV technique, 27

j: position of the players, 128

K: number of nearest neighbours in the KKN, 27

l: level of the reference players popularity, 128

LV: number of latent variables in PLS-DA, 27

$mtry$: number of variables in each tree of the RF algorithm, 27

M: variables, 22

N: observations, 22

$nodesize$: minimum size of the terminal nodes of the RF algorithm, 67

P: loading matrix, 22

T: score matrix, 22

usekernel: logical, to estimate conditional class densities (kernel or density estimation) in Naïve Bayes, 27

X: predictors matrix N x M, 22

Y: response matrix, 24