Assessing Team Success in the Big-5 European Football Leagues

José Paulo Direito Fonseca
jose.paulo@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2022

Abstract

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

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

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

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

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

1. Introduction

1.1. Contextualization and Motivation

Similarly to most sectors throughout the world, the sports and football industry have discovered in data and Data Science (DS) the power to improve areas of the organization, which could ultimately provide them with the upper hand over its rivals through a competitive advantage(s).

In an industry where just the European football market size is estimated to be worth €27.6 billion as of 2020/21 [1], it is clear that the adoption of these practices might help sports stakeholders both inside and outside the field.

Nevertheless, the use of data and data analysis in sports has not come without its challenges [2], given that teams and game strategies are constantly evolving. Additionally, as data from both national and international competitions is being produced every year, analyses of this type are always pertinent as they offer new fertile ground for insights to be derived.

1.2. Objectives

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

1. Identify a set of KPIs that distinguish teams on the basis of three categories of success according to the final ranking: clubs that win the championship and/or qualify for European competitions (high-ranked); clubs that safeguard their place in the league the next season (middle-ranked) and clubs that are relegated or qualify for the relegation play-off (low-ranked).

2. Identify a set of KPIs that differentiate teams on the basis of the league they are part of, with the purpose of characterizing them.

3. Construct a classification rule using data from the previous four seasons and evaluate its performance in estimating the final rankings for the 2021/22 season based on the success categories.

2. Literature Review

2.1. Sports through the lens of Data Science

With the turn of the millennium came noteworthy advancements in the fields of Artificial Intelligence (AI), DS, and Machine Learning (ML) empowered by advancements in infrastructure and computer science.

The technological advances together with the developments in these areas made it possible for organisations in all industries to store and analyse complex data, guiding the way to more data-driven habits, decisions, and solutions. With the underlying purpose of supporting and providing guidance towards solving the problem(s) at hand, the value of data for companies and institutions has now grown to be commonplace in every sector as a result of it being associated with its ability to create competitive advantages, innovate and grow businesses.

In sports, as of now, these advantages are also substantial, and tangible as recognized by the myriad of success stories that proved the usefulness of data. However, the path of implementing data gathering, assembling and how to do it – sports notation – would begin to be walked as early as the mid-nineteenth century [3]. Back then, these practices emerged in work developed by sports journalists and would, over the following years, emerge throughout the world in sports.

Contrary to the growing prominence of sports-statistic in this format - newspapers - was the diminished attention that this matter got academically, around this time. Likewise, in football, the subsequent decades also saw little investigation around related statistics, with relevant work only being performed later in the century motivated by the establishment of appropriate topic-oriented academic journals around these years.

However, only with the advancements that came along the new millennium were the ideal settings in place for the computer-powered statistical data analysis practices, known more simply as “Data Science”, to be adopted even more broadly and conveniently in all sports throughout the world.

2.2. Data and Data Science in Football

With three possible outcomes – win, defeat or tie –, association football, also known as “soccer” (American English term), is a team sport in which two teams of eleven players compete against one another to score goals.

Often cited as one of the first academic research studies where data was taken into consideration to analyse the game, Charles Reep’s work entitled ”Skill and Chance in Association Football” carried in 1968 focused on the optimal number of passes leading up to a goal and the best way to score a goal, based on the analysis of more than 2000 games [4]. This research was at the forefront and inspired succeeding investigations carried out with the same purpose of comprehending the game, being match analysis, reckoned as the most popular research topic [5].

2.2.1 Key Performance Indicators

Aiming at determining specific aspects that provide a portrait of teams’ successful performance, the study of performance indicators has drawn much interest among researchers over the years, with studies that looked into international and national leagues over different periods being conducted.

Concerning international competitions, some investigations demonstrated that longer passing combinations and a higher level of possession were the foundation of successful teams. Contrastingly, findings from studies that analysed different sets of data deviate from the previous idea, discovering no evidence that supported the relationship between ball possession and success in the competition. Furthermore, contradictory results were also reached when analysing international competitions in terms of the optimal number of passes that lead to a goal [6], with authors demonstrating that approximately 80% of goals were preceded by two passes or less with the minority developing after five or more passes.

Other studies have looked solely at one national league. Whilst some research suggests, like the previous, that successful teams demonstrate a significantly longer ball possession than unsuccessful ones, findings from other investigations seem to place more emphasis on other performance metrics, such as a greater distance covered in running in high-intensity, and the execution of more dribbles and passes. Additionally, other investigations still suggest that successful teams performed a greater number of shots, shots on target, had higher effectiveness, and displayed increased action in the attacking area of the pitch.

2.2.2 Big-5 European Football Leagues Comparative Analysis

Not as avidly analysed as match performance statistics from individual competitions is literature that focuses on comparing the Big-5 European

Football Leagues However, football-related research carried towards comprehending what characterizes them provides interesting distinctions that could be interpreted more broadly as distinct game-approach stereotypes.

Overall, investigations seem to agree on the higher strength and athleticism of German Bundesliga (GB) players, while the English game is recurrently described as the most intense and highest paced-game. Concerning the Spanish La Liga (SLL), investigations highlight the more intricate and skilled performing technique of its players, with results emphasizing their effectiveness in converting shots into goals, while the Italian league is usually portrayed as the league most reliant on the defensive organization.

2.2.3 Classification Analysis

The first effort at forecasting the outcome of a match was made in 1997, and it was based on a Poisson distribution to derive probability for the goals scored [7]. Over the years, the advent of increasing research being developed to understand the game of football and metrics that play a more/less important role in the teams' performance has also motivated the usage of DS to build predictive models and employ classifiers with ever-increasing accuracy.

On the one hand, a substantial amount of investigations aims at predicting match outcomes despite the often challenging drawing possibility, as emphasized by every author, which hinders the performance of the algorithms in most cases. Indeed, the majority of results obtained showed that even though promising results were achieved concerning estimating wins and losses with the help of ML techniques, their accuracy was decreased by the algorithm's difficulty in modelling draws.

On the other hand, research has also been conducted to predict the classification of teams at the end of the season, rather than on a game-outcome basis, despite the difficulty in doing so since it is contingent on the ranking of all other teams. One noteworthy analysis was [8], whose authors, rather than using data from only one country, stretched the investigation to four of the Big-5 European Football Leagues by considering 40 features and making use of a broad range of classifiers: Naive Bayes, Decision Tree, Random Forest, KNN, SVM (radial basis function and polynomial kernel) and XGBoost. In the end, the best-performing algorithm was SVM with the polynomial kernel which obtained its highest accuracy for the English Premier League (EPL) (57%) and the best result in terms of the RMSE metric for SLL, where the classifier correctly predicted the ranking of the first six clubs out of twenty in the league.

3. Methodology

3.1. How to measure teams' performance in football?

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

Every nation in Europe has a top professional league, followed by minor divisions. They are all part of a bigger regional body, the Union of European Football Associations (UEFA) for the leagues under consideration in this research, that controls European-wide tournaments like the UEFA Champions League (UCL), UEFA Europa League (UEL) and UEFA Europa Conference League (UECL). The club's standing at the end of the season determines whether it qualifies for these events and whether it advances or drops down to a lower-level national tournament. As a result, for this work, the differentiation between the clubs was established in light of three potential outcomes, as noted in Table 1.

There were no distinctions between direct qualifications and group stage qualifications; all teams were regarded to be in the high-ranking category. The similar technique was taken with clubs that qualified for the relegation playoffs (but were not immediately demoted) - they were categorized as low-ranked.

3.2. Framework

The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework served as the methodology's backbone used to guide this investigation [9]. The approach is summarized by the dynamic logical connection, represented in Figure 1 between its six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

The first phase of the CRISP-DM process focuses on understanding the project goals from a business standpoint by providing a comprehensive explanation of why these types of research could be of interest to clubs and stakeholders. Additionally, it outlines a course of action toward answering the proposed question based on available resources.

The Data Understanding phase comprises the earliest contact with the data and encompasses activities such as collecting the initial data, checking its quality and deriving first insights. This stage

Table 1: Categories of success.

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

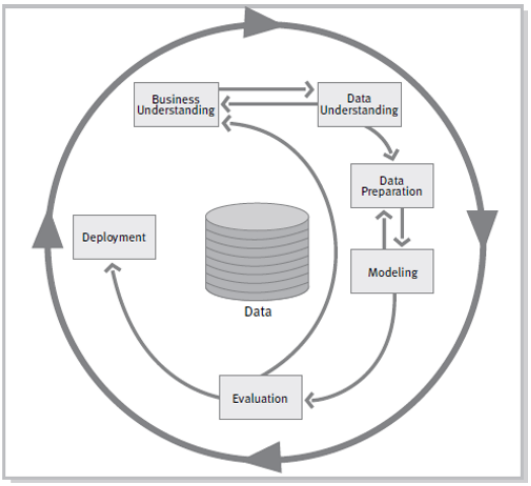

Figure 1: Phases of the CRISP-DM reference model (from [9]).

was undertaken by exporting the relevant information from the FBREF website and further exploring the description of the variables available.

Following the previous step, the Data Preparation phase is responsible for all operations related to obtaining the final dataset which will later be input into the modelling tool(s). In a starting step, there was the need to select the relevant features, which was achieved through three steps: first, disregard every predictive statistic calculated by the platform; second, discard those that were deemed irrelevant for the study, and third perform an exploratory analysis to refine the dataset so it only included variables with increasing explanatory power. Other tasks included cleaning, constructing, integrating, and formatting the data. Concretely, there was the need to fill in the missing values, create the success metric variable by manually introducing the information, reorganizing some metrics, and making the required merges. Subsequently, a preliminary data analysis was carried out to better understand some global and group properties of the dataset.

For this investigation, Python [10] was used as a programming language in the Jupyter Notebook [11] environment. For purposes of data preprocessing, features engineering and model creation, the Pandas [12], Numpy [13] and matplotlib [14] packages were used.

The Modelling step includes selecting and implementing the appropriate data mining techniques for the project along with ensuring optimum calibration for their parameters. To identify the KPIs that determine the success of teams, two unsupervised learning tools: Principal Components Analysis (PCA) and Clustering Analysis using Hierarchical (Agglomerative) and Partitioning (K-Means) methods.

The idea behind PCA is to reduce the dimensionality of a dataset while preserving as much of the original variability as possible. Given the different variable measurement scales and wide range of variability, the data was standardized. Mathematically, this can be achieved by subtracting the mean, $\bar{x}_i$, and dividing by the standard deviation, s_i , of the $i-th$ variable x_i, i.e. $z_i = (x_i - \bar{x}_i)/s_i$.

Following this step, the eigenvalues, and eigenvectors of the correlation matrix were calculated. In mathematical terms, if the initial data is defined as n observations with measurements on a set of p variables, $\mathbf{X} = (X_1, X_2, ..., X_p)^T$, then, essentially, the first principal component (Z_1) is the linear combination of the original variables with the highest variance, as in (1).

$$Z_1 = \phi_{11}X_1 + \phi_{12}X_2 + ... + \phi_{1p}X_p \qquad (1)$$

The loadings of the first principal component will then be $\boldsymbol{\phi}_1 = (\phi_{11}, \phi_{12}, ..., \phi_{1p})^T$ such that $||\boldsymbol{\phi}_1|| = 1$ i.e. $\sum_{i=1}^{p} \phi_{1i}^2 = 1$.

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

Clustering Analysis, also seek to simplify the data but, instead, the objective further includes finding patterns in the data by looking for smaller homogeneous groups in the observations. Thus, Hierarchical (Agglomerative) and Partitioning (K-means) methods were considered to see if the classification of the teams into two groups was able to distinguish between high-ranked teams from middle-ranked and low-ranked teams (K = 2) and then, to examine how effectively the three preceding categories are distinguished, a model with three groups was evaluated (K = 3).

The outputs of the clustering models were evaluated through the accuracy, precision, recall and F1-score measures. In addition, the Silhouette Coefficients (SC) for the different clustering solutions were computed [15].

When using the methods, the original data which included all observations and features was used and, in a separate analysis, data projected in the

first 22 PCs was considered.

To answer the classification dimension of the research question, which aimed at classifying teams, in terms of the categories of success (Table 1), from all leagues given their performance in the 2021/22 season, the four seasons studied worked as the train set. The models considered in this analysis were Gaussian Naïve Bayes (GNB), Logistic Regression(LR), K-Nearest Neighbors (KNN), Adaptive Boosting (AB), Extreme Gradient Boosting (XGB), and Random Forest (RF). Before applying the methods, both the training and testing feature data were standardized to avoid poor model performance when attempting to discover patterns in the data and to guarantee that each feature is equally valuable to the study. To build the machine learning models, the Scikit-Learn package [16] was considered except for (extreme) gradient boosting that required the installation of the XGBoost Python module [17].

K-fold Cross Validation [18] was the technique used to assess the mean accuracy of the techniques, and the value chosen for K was 3. Furthermore, the performances of the algorithms were also evaluated in terms of precision, recall and F1-score, and, to assess the performance of the best-performing ones, Receiver Operating Characteristic (ROC) curves were constructed.

In the Evaluation stage, the results are analysed, interpreted, and compared to the established business objectives to see if they were reached.

4. Business Understanding

4.1. Business Goals

As a company, the purpose of a football club is to be profitable through superior exhibitions of its team during the league season which, in the case of European teams, could ultimately mean placements in European competitions. Hence, it is only natural for the team's decision-makers and managers' ultimate goal to be aligned with increasing the team's performance.

As the review of the literature respecting performance indicators showed, a considerable amount of research has been carried with some focusing on international or national tournaments while others analyse a different number of teams and/or seasons. However, literature also shows that arguably the most important element of these types of research is the set of variables and statistics considered, as they could have the potential explanatory power to comprehend the game, at least partially (the majority of these metrics do not contemplate the spatial-temporal dynamics of football teams). In this regard, two problems were found. Firstly, it was the case that most of the analyses still look into a small number of variables and second, the majority of them also fail to be comprehensive on all types of actions and interactions that can occur in a game. The review of the Big-5 European Football Leagues' literature revealed that there is still a dearth of research available that aims to compare all five leagues in terms of performance indicators. Likewise, examining the predictive studies showed that the majority attempted to anticipate either the outcome of games or the team standings at the season's end for one competition, with a small number of research seeking to broaden the analysis to include all Big-5 European Football Leagues.

From a business perspective, investigations that aim to tackle the aforementioned challenges might reveal themselves useful for clubs at many levels. By having a comprehensive understanding of what better distinguishes teams in terms of the success metric considered, coaches could be able to better orient and define training practices and tactical strategies with a focus on aspects that are demonstratively shown to give increased results. Additionally, this knowledge would be most useful to team managers and scouts since it would equip them with a clear idea of the different requirements across leagues thus indicating the most/least appropriate fits for the team in every decision. Then, physical coaches could also benefit from this information since recovery from training and games is a very important element when it comes to sports. Finally, the classification element of the research question might be useful for game and opposition analysts.

4.2. Available Data

The FBREF website was the only source where the technical data with the necessary volume and degree of granularity was available. In terms of information accuracy, the website gets its data from StatsBomb, a very well-known data provider when it comes to football statistics.

The site makes statistical information available at club level categorized into five groups before the 2017/18 season and eleven groups henceforth, as shown in Table 2.

Information concerning the categories of success was sourced from different pages. To identify the high-ranked teams, the institutional pages of the UEFA competitions were used, while information on the remaining categories was obtained by identifying the clubs that were relegated, in the official league websites.

5. Data Understanding

5.1. Data Collection

The data was extrated by exporting it to CSV files and the complete statistical information that constituted the basis for this dissertation project respected the Big-5 European Football Leagues –

Table 2: Groups of features availability for the leagues across seasons in FBREF website.

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

EPL, GB, SLL, FLO and ISA – and the four seasons – 2017/18, 2018/19, 2019/20 and 2020/21. For the classification analysis, the 2021/22 season was considered.

5.2. Data Description

The statistical information was expressed in terms of the total number of times some action occurred, the percentage of successful actions, the number of actions per game and even some continuous variables as is the case with some statistics that measure the distance for some specific actions.

Excluding repeated metrics across tables, the categories' initial statistical variables were listed as in Table 3, and described in more detail.

Table 3: Initial variables included in Pass Types category.

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

6. Data Preparation

6.1. Data Selection

The metrics used for the analysis did not include every initial statistic from the original tables. In addition to filtering down repeated variables, there was the need to discard some variables that were not informative of any game-related aspect and hence judged to be irrelevant to the study (identified in Table 3 and similar tables by the Not relevant - NR - column). Then a correlation analysis to the remaining variables allowed to rule out some metrics, to include (almost) exclusively those with added explanatory capacity, as visible in Figure 2 example. The Pearson sample correlation coefficient determined the exclusion criteria and the threshold value considered was 0.9 and the variables abandoned for this reason can be identified in Table 3 and similar tables under the FR (Feature Reduction) column.

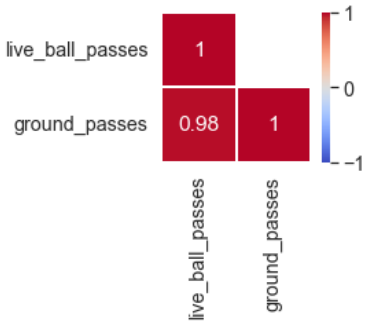

Figure 2: Pass Types correlation analysis ($P > 0.9$)

6.2. Data Munging

In order for the data to be utilized and processed, changes such as converting the continuous variables represented from an initial yard basis description to meters (equivalency used: 1 yard = 0.9144 meters) and filling a total of 58 missing values in the number of yellow and red cards variables had to be made.

In the end, the last step comprehended merging the eleven individual datasets, into only six categories, as seen in Table 4, since eleven was considered an unnecessarily high number of groups to organize the data (information regarding the updated categories is provided in Table 3 and similar tables under the last column – FC (Final Category)).

6.3. Preliminary Data Analysis

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

In terms of the club's representativity concerning the success metric categorical variable, it is clear from Figure 3 that more than half of the teams

Table 4: Final variable's groups.

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

(51.4%) account for middle-ranked clubs, followed by high-ranked with 34.9% and finally low-ranked, adding to 13.7% of the teams.

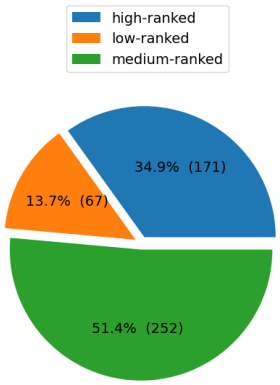

Figure 3: Distribution of clubs in terms of the categories of success.

Concerning the leagues, the GB was the only underrepresented league with 90 teams, which is not surprising since only 18 squads compete yearly in the league, compared with the 20 that do so in the remaining ones considered. This has a direct impact on the total number of games at the end of the season, with the underrepresented league totalling 306 match events per season compared to the 380 in the remaining. Here, the decision, was not to exclude the league as a whole from the analysis but rather multiply all metrics from the features space, excluding percentage and ratio-based metrics, by a constant, as evidenced in (2).

$$UpdatedFeature = Feature * \frac{20}{18} \quad (2)$$

The analysis of the games played by league and seasons further captured an abnormally low amount of games played in FLO during the 2019/20 season due to the COVID-19 pandemic as the league did not resume that year. However, the decision was to use the data without applying any correction/edition, as it was considered a completely different situation from the previous one because it respects an unexpected event.

7. Modelling and Evaluation

7.1. Principal Components Analysis

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

7.1.1 Key Performance Indicators and Football teams' performance

Figure 4 shows the dispersion of points when PC1 and PC2 are considered together.

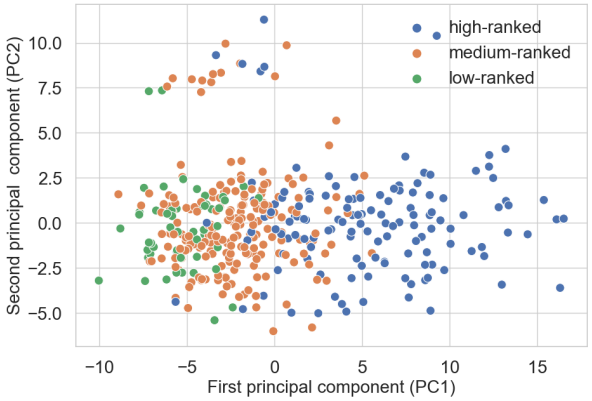

Figure 4: Scores on PC1 vs PC2.

Results suggest that the first Principal Component (PC) does a good job at separating the high-ranked teams from the middle-ranked and low-ranked clubs but misses on a more detailed segregation between the two latter categories. However, the latter – low-ranked teams – still score the lowest in PC1, with middle-ranked teams scoring higher, on average, than this group.

To understand what distinguishes the teams, the variables that have the most significant impact on this PC were analysed, considering a threshold value of 0.15 (in all cases) to filter the most important ones. Results reveal, as seen in Table 5, that the factors that have the greatest effect on this component belong to the Possession, Passing, and Goal and Shot Creation categories.

Included in the first and displaying a positive coefficient (suggestive of a direct proportionality association with the PC) is the percentage of ball possession, the variable with the second highest absolute value. Pearson's sample correlation coefficient between the variable and the PC ($r = 0.917$) validates this premise, while the p-value ($p < 10^{-8}$) for testing the null hypothesis that this correlation is

Table 5: Loadings whose absolute value is greater than 0.15 for the first principal component and Pearson sample correlation coefficient and p-value between each variable and the component.

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

zero suggests that the findings are, indeed, statistically significant. This is a testament to conclusions reached by previous researchers, thus emphasizing the role that ball possession plays in terms of team performance. Other noteworthy metrics from the same category include the number of times the ball was carried into opponents' goal, carries in attacking third and into goal box and touches in offensive third. Given that high-ranked teams score the highest on PC1 and the coefficient for these statistics indicates a direct proportionality relationship, these results are suggestive of higher values on these metrics for teams belonging to the high-ranked category, thus emphasizing that on top of being in control of the ball more often – a higher percentage of ball possession -, high-ranked teams also display a strong offensive presence and attitude.

From the Goal and Shot Creation category, variables such as the number of passes that lead to goal and number of passes that lead to a shot attempt stand out, as having a direct proportionality relationship with the present PC (indicative that high-ranked teams displayed higher values on these metrics, on average). When considered with the variables from the previous paragraph, these results seem to suggest that most high-ranked teams prioritized a positional attacking style with the circulation of the ball occurring all over the field, rather than a more direct method of play, or based on counter-attack, which is characteristically quicker and reliant on fewer passes.

Still concerning Figure 4, it is clear that the second PC does not do a good job of distinguishing the categories but succeeds, however, in identifying a set of outlier points with abnormally high values corresponding to the 2019/20 season.

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

Results from the third PC were marginally more expressive in terms of the distinction between middle-ranked and low-ranked teams, with the latter group scoring lower on average. When analysed, the third principal component's loadings demonstrate that the variable with the highest absolute loading was the number of touches in the defensive penalty area, closely followed by the number of touches in the defensive third of the pitch, both part of the Possession category and since the coefficients and the Pearson sample correlation measure are negative for the former ($\hat{\phi}_{3,80} = -0.333; r = -0.819$) and latter ($\hat{\phi}_{3,81} = -0.296; r = -0.729$), it follows that a higher number of occurrences in these statistics can be expected for low ranked clubs. In parallel, results show that most variables, ten out of sixteen, belong to the Goalkeeping category, further emphasizing the influence of this aspect of the game for this principal component.

7.1.2 What characterizes the Big-5 European Football Leagues

Concerning the leagues, an analysis to PC2 reveals that GB and EPL data points stand out as scoring lower than the remaining categories.

Reportedly distinguished by its fast-paced approach whose method of play is supposed to even out the work rate, this characterization of the English game is in line with the analysis of the second component loadings where the relevance of the level of pressure in the middle and attacking third and the number of tackles in the middle third of the pitch variables is highlighted.

Surprisingly, findings also point to strong team cohesion in the case of GB clubs, as opposed to most literature that defends a style of play more reliant on the athleticism of their players rather than their collective performance.

An analysis to PC2 and PC4, as in Figure 5, demonstrate that, disregarding the FLO 2019/20 season outliers, EPL clubs score the highest on PC4, on average, followed by GB clubs. Revealing significant variation in point distribution FLO and ISA clubs follow, with SLL teams scoring the lowest on PC4, on average.

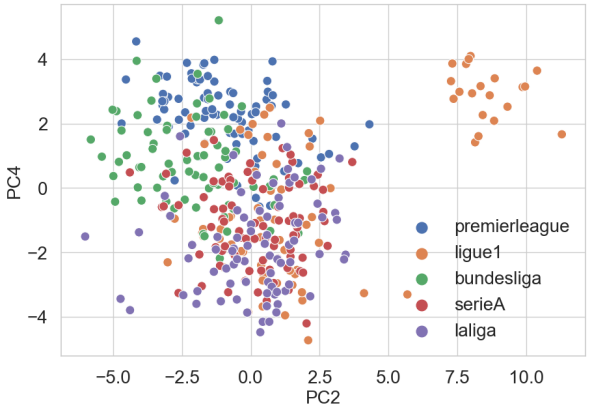

Figure 5: Scores on PC2 vs PC4, coloured according with the league.

An analysis to the fourth PC loadings demonstrate that SLL and ISA players committed the highest amount of infringements and, coherently, received the highest amount of yellow and red cards, which could be said to weight on the useful playing time and consequently be related to the high pace of play reported EPL (which scores on the other end of the spectrum in terms of the number of fauls drawn and conceded).

7.2. Classification Analysis

From the six classifiers used, results demonstrate that XGBoost was the algorithm that better estimated the category of success teams would finish at the end of the 2021/22 season, achieving a mean test accuracy of 86.68% (approximately 5% higher than the previous - RF) as it is visible from blue bars in Figure 6.

Additional studies using the data projected in the first 22 PCs revealed that every classifier performed worse (orange) and when Eintracht Frankfurt and Nantes FC were removed from the analysis (two clubs whose presence in the UEFA competitions was granted by other competitions rather than their performance on the national league), the classifiers' performance improved (green).

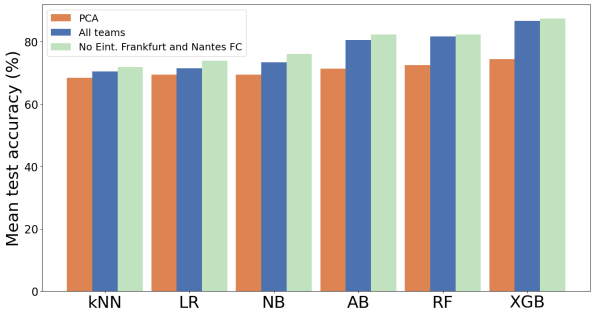

Figure 6: Mean test accuracy for the applied classifiers before (blue) and after (green) removing the Eintracht Frankfurt e. V. and Nantes and when using data projected in the first 22 PCs (orange).

8. Final Remarks and future work

8.1. Final Remarks

Since the late nineteenth century, the world's most popular sport has experienced advancements both within and outside the field. Concerning the latter, according to the literature, investigations have covered the technical-tactical, physical, and mental domains, with a focus on match analysis to understand what determines success in the sport. However, it is the case that most of the analyses still look into a small number of variables and second, the majority of them also fail to be comprehensive on all types of actions and interactions that can occur in a game. Furthermore, research indicates that certain studies do hint at a level of differentiation across the Big-5 European Football Leagues, although studies that explore all of them together are still scant. Given these limitations, it was concluded that by broadening the scope of analysis to include KPIs about goalkeeping, defensive actions, types of passes (rather than the number) and the location from which these occurred in the game, other relevant patterns could emerge not only in terms of success categories but also leagues. In the end, 98 KPIs sourced from the FBREF website were investigated over the course of four seasons (2017/18 - 2020/21) for the Big-5 European Football Leagues.

With this work, in terms of the categories of success, the PCA scatters and first PC loading's analysis demonstrated firstly that high-ranked teams displayed higher patterns of ball possession, and a playing style reliant on exchanging the ball more often with the intent of reaching the attacking third of the pitch and performed, on average, a higher number of passes leading to a goal and shot attempts. The third component results emphasized the low-ranked team's salient patterns in defensive statistics, despite some cases where these teams purposefully adopt a low-intermediate block as an attempt to explore the space behind the defensive line through counter-attack.

Concerning the distinctions between the

leagues, the EPL and GB team's strong overall presence in all areas of the pitch stands out, in line with the reported high-intensity patterns that commonly characterize it. Additionally, ISA clubs showed a prominent use of the defensive section of the field, while the fourth PC revealed that SLL and ISA players committed the highest amount of infringements and received the highest amount of yellow and red cards.

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

8.2. Future Avenues of Research

Following this dissertation, it would be interesting to comprehend the extent to which the consideration of the FLO 2019/20 season outliers impacted the results by carrying the same analysis with statistics that have the same number of games as basis. Additionally, future work should aim at studying the variables highlighted in this investigation more deeply. Finally, future studies might focus on improving the interpretation of the results of each classifier and on applying the classifiers to only a partially complete data set (for example mid-season) to determine their accuracy in predicting the teams' final rankings before the season's end.

References

[1] Theo Ajadi, Amy Clarke, Sumeet Dhillon, Grace Gardner, Dhruv Garg, Tom Hammond, Alasdair Malcolm, Jenny Pang, and Jamie Pugh. Deloitte annual review of football finance 2022, 08 2022. Accessed: 06-09-2022.

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

[3] J Simon Eaves. A history of sports notational analysis: a journey into the nineteenth century. *International Journal of Performance Analysis in Sport*, 15(3):1160–1176, 2015.

[4] Richard Giulianotti and Roland Robertson. The globalization of football: a study in the glocalization of the 'serious life'. *The British journal of Sociology*, 55(4):545–568, 2004.

[5] Thomas Reilly and David Gilbourne. Science and football: a review of applied research in the football codes. *Journal of Sports Sciences*, 21(9):693–705, 2003.

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

[7] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997.

[8] Victor Chazan Pantzalis and Christos Tjortjis. Sports analytics for football league table and player performance prediction. In *11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE, 2020.

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

[10] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Create Space, Scotts Valley, CA, 2009.

[11] Bernadette M Randles, Irene V Pasquetto, Milena S Golshan, and Christine L Borgman. Using the Jupyter Notebook as a tool for open science: An empirical study. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2017*, pages 1–2. IEEE, 2017.

[12] Wes McKinney et al. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[13] Charles R. Harris and K. Jarrod Millman. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[14] John D Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

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

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, et al. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[18] Daniel Berrar. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, 1:542–545, 2019.