NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

How does each team's playing style approach influence results?

A Machine Learning approach

Diogo Rajés do Passo Liladar

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

How does each team's playing style approach influence results?

A Machine Learning approach

by

Diogo Rajés do Passo Liladar

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Professor Roberto Henriques, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 14th 2024

ABSTRACT

More now than ever, analytics in sports refer to the ability to analyse and retain insights from data generated from broadcasting, sensors and players' trackers. This project aims to check whether a more offensive playing style approach achieves better outcomes and vice versa. The data chosen corresponds to two seasons of the Football's Portuguese premier league, and the project followed a two-phase approach, the first leveraging the use of supervised algorithms to perform interpretable predictions and the second using unsupervised algorithms for clustering. The best cluster solution reached resulted in two highly distinct clusters, displaying differences in the playing style approach, one clearly more offensive than the other, as measured on several match statistics. Furthermore, the clusters differed on background and environmental contexts. Additionally, the project provides a methodology and a framework to be applied in different contexts and data.

KEYWORDS

Football; Machine learning; XGBoost; Unsupervised algorithms; Supervised algorithms; Prediction; Clustering

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

IFAB	International Football Association Board
FIFA WC	FIFA World Cup
JSON	JavaScript Object Notation
CPU	Central Processing Unit
GPU	Graphics Processing Unit
RF	Random Forest
SSW	Sum of Squares Errors
T-SNE	T-distributed Stochastic Neighbor Embedding

1. INTRODUCTION

1.1. FOOTBALL

Football is undoubtedly the most popular sport globally (Blakemore, 2023). In comparison, for instance, the 2023 Superbowl between the Kansas Chiefs and the Philadelphia Eagles yielded 155 million viewers worldwide, while the 2023 Champions League Final between Manchester City and Inter Milano reached 450 million viewers, almost three times the Superbowl viewership (May et al., 2023). Moreover, the 2022 World Cup final, displaying Messi and Argentine against Mbappe's France, had around 1.5 billion viewers (May et al., 2023). The English Premier League alone broadcasted, in the year 2022, to 800 million homes in 188 countries (Evans, 2022), expecting to generate over six billion dollars solely on foreign rights sales for the three seasons starting in 2022/2023 (Evans, 2022).

Furthermore, football's global industry has been developing quickly to becoming one of the most relevant in the world as the top 32 clubs on aggregate's enterprise value reached 51.7 billion euros, a 98% increase in seven years (26.3 billion euros in 2016) (Football Benchmark, 2023).

1.2. FOOTBALL IN PORTUGAL

Portugal is no exception concerning football popularity as the sport is locally denominated as "the king sport". In terms of viewership, for instance, in 2023, the top ten television programs watched were all football games (*Futebol domina audiências*, 2024). Moreover, when comparing practitioners, according to Pordata's database[1], in the year before Covid – 2019, football practitioners were almost four times more than the second most played sport, which was handball.

Furthermore, the Portuguese League is described as a brilliant nursery of talent (Hawkey, 2020) and one of the main entrance points of European football for emerging football talents, most of them coming from South America, such as, most recently, Enzo Fernandez, Manuel Ugarte and Luiz Diaz, to mention one from each of the leading Portuguese clubs – SL Benfica, Sporting CP and FC Porto, respectively. Nonetheless, it is a great development league for many local-born talents such as João Felix, Bruno Fernandes and Ruben Dias, all playing in top 5 leagues and worldwide recognized today.

1.3. MACHINE LEARNING

In every sport, especially football, observational analysis was the most important for researchers and practitioners to study and innovate in training and team game preparation. However, more recently, due to significant technology developments obtained mainly in the

[1] Database created by Fundação Francisco Manuel dos Santos with European and Portuguese statistics on several different fields.

last decade, it was possible to get individual and team statistics, key performance indicators such as GPS tracking data, and video-based motion analysis, among many others (Herold et al., 2019). This mainly happened due to the change implemented by IFAB (The International Football Association Board, 2016), allowing players to use Electronic Performance & Tracking Systems (EPTS).

1.4. RESEARCH GAP/QUESTION

Machine learning is a recent tool in football, and its usage provides important knowledge to the most relevant stakeholders, such as coaches, scouts, and analysts whose objective resides on understanding how to leverage its teams winning possibilities. The focus of this project covers an important piece of such topic as its primary aim is to find if each team's game model influences the team's results, i.e., if, for instance, the teams that display a more offensive playing style approach, on average, win more times. It will, thus, attempt to answer wheather teams displaying distinct playing style approachs get different match outcomes. Therefore, in order to accomplish this, the project will be divided into two main parts. The first part will be based on using supervised models to predict the goal difference outcome of a given team on a given match and extracting the most essential features. Its main goal is to indentify what are the most relevant features for the model, which can be inferred as the team's statistics that impact the most the match outcome. Conversely, the second part will cluster the most relevant features generated in the first step, attempting to identify distinct group of teams performing similar playing style approachs. This will, consenquently, allow to acess whether team's clusters with distinct playing style differ on match outcome.

This project will focus on Portuguese teams but can later be replicated with/for other leagues. The methodology can provide a framework for many stakeholders' analysis or even shape current and future coaches' data-driven approaches.

Therefore, for an easier read, the outline of this report is arranged in different chapters. The second chapter mentions and discusses relevant literature review of papers that have been published concerning the topic of football performance prediction and tactical analysis. Followingly, the third chapter describes the methodology that has been followed and applied on the data exploration, preparation, engineering, the models implemented and the evaluation metrics used. The fourth chapter concerns the presentation of the results obtained alongside their discussion according to the models used and the evaluation metrics chosen. Finaly, the fifth chapter addresses the main findings, their limitations and future work suggestions.

2. LITERATURE REVIEW

This chapter will provide an overview of literature concerning machine learning approaches to football. The first part regards a compilation of distinct papers from several databases performed in 2019 relating to the usage of machine learning in football; and the remaining chapter is divided into two main topics: literature about predicting performance and literature on which the goal is to analyse team behaviour or formations using only event data, tracking data or both.

2.1. MACHINE LEARNING IN FOOTBALL

Despite being a growing field in Football, Herold et al. (Herold et al., 2019) compiled papers and studies using Machine Learning in elite professional male football. This compilation refers to the period between 1996-2018 and was retrieved from the following databases: PubMed, Web of Science, MEDLINE, SportDiscus, Researchgate, Elsevier and ProQuest, with the searches focused on keywords such as machine learning, football (or soccer) and performance analysis. The study searched for different traits to assess each paper found. The traits used were the machine learning strategy (supervised/unsupervised), specific machine learning techniques (e.g., neural networks, k-nearest neighbour), the input data utilized (event vs tracking data), and if the authors gave enough information to repeat the study. In this project, the most relevant distinction was in input data, on which event data can be described by measures collected from one or several matches such as proportions, frequencies, and other accumulated metrics.

On the other hand, tracking data always allows for the recording of the spatial locations of the players and the ball. The authors indicated that, from the papers analysed, the ones concerning event data usually used larger samples to build their models but then applied relatively simple regression models to build classification or prediction models. Regarding tracking data, the authors identified four primary categories: i) patterns of pass sequences on an individual level, ii) patterns of pass sequences on team level, iii) the time and space sequences on plays that either result in goal or goal-scoring opportunities and iv) regarding defending or regaining the ball. The papers that studied the first two categories did not produce reasonable conclusions as the evaluation metrics for the classification of passes were not great (mostly recall and precision) but have laid the methodology for relevant future work. Regarding the third, results and conclusions are centred on whether the number of shots or their location correlates with team performance success, while on the fourth resided on the greater quality of the top teams at effectively use and stop counter-attacks.

2.2. PERFORMANCE PREDICTION

Concerning performance prediction, the literature varies in terms of both end goals and techniques used. Papers are attempting to predict matches and championships' results using machine learning algorithms, as in the case of Joseph et al. (Joseph et al., 2006), where they

used Bayesian Nets and other algorithms to predict the outcome of matches of Tottenham Hotspur based on data from 1995-1997. Conclusions reached indicated that from the algorithms used, KNN performed better for full-season training data – averaging around 97% of correct scores – but with low scores when trained and tested only on cross-season data – i.e. model trained on periods of one particular season and tested on a period of other distinct season - and an expert Bayesian Net outperformed all others when using disjoint training and test data with around 59% correct predictions. In another study, Pappalardo and Cintia (Pappalardo & Cintia, 2018) attempted to analyse more than 6000 games and 10 million events – such as passes, dribbles, shots or fouls with the correspondent events' field location - in six European leagues collected over three seasons from 2013 to 2016. Here, their work was divided into three distant parts. The first regarded the relationship between these game events and the performance of each team on a season by applying a regression model and concluding that it explained around 65% of variance encountered on the final score. Then, on the second, they predicted the outcome of each game based on the match events of each team using a logit and a random forest model classifier, both yielding similar results, on which they observed that the strongest predictors were the number of passes, shots and goalkeeping actions alongside the fact that draws are hard to predict with a machine learning approach. In the last one, a ranking for each of the six analysed leagues was performed based solely on the synthetic outcome generated from the predictions of the match events. The predicted league rankings were measured by the correlation and group accuracy (accuracy of the league rankings as a whole) between the predicted and the actual rankings. It was found that both yielded similar results. Nevertheless, the authors stated that the simulation tends to underestimate the number of points for the teams at the top of the ranking and, on the other hand, to overestimate the number of points for the teams at the bottom of the ranking.

Yet, on the same topic, Moustakidis et al. (Moustakidis et al., 2023) attempted to predict team performance with explainable AI to identify key performance metrics. The paper used 160 event-related features from all the matches played in the eleven European countries' top divisions in the 2021/2022 season. Then, a ten-fold cross-validation strategy was implemented with XGBoost as the model and the R-squared as the selection criterion when compared with Support Vector Machine, Random Forest and K-Nearest Neighbours algorithms. The main objective was to predict the goal difference between teams in a match, for which the XGBoost yielded a root mean squared error (RMSE) of 32.09%, and to identify the contribution of each performance indicator to the match score both for the teams as a whole and for each team individually using Shapley Addictive explanations (SHAP). Overall results showed that offensive features were regarded as the most important, as features such as "*shots_per_quantity_of_possession_percent*", "*missed_chances*", "*entrance_to_the_penalty_box*", "*chances_percent_of_conversion*" and "*key_passes_accurate*" constituted the top 5 most essential features for the model.

The team-specific SHAP plot analysis also allowed us to identify distinctive match statistics from the coaches' signature game models, such as Klopp's on Liverpool or Guardiola's on

Manchester City. Performance is also the subject of a paper designed by Van Haaren et al. (Van Haaren et al., 2019), in which the authors indicate that ball event data can be used to analyse performance and teams' playing styles. Moreover, the paper suggests there are three different ways to collect data from a football match according to their granularity and availability: i) match sheet data which is the most simple but most available – mainly for free -, ii) ball event data including game events such as shots, passes, duels, attached with the respective player and location on the pitch and iii) the most granular but less available, tracking data comprising of cameras following all players and the ball at all times. Yet, the authors suggest three different ways to analyse team tactics and playing styles: the first through clustering the team's statistics, the second through pattern mining and the last through modelling the complete behaviour of a team, usually on a network-based approach. In this project, a mixture of the first two data collection types – match sheet data and ball event data - will be used alongside clustering of the essential features of the teams to identify playing styles.

2.3. TACTICAL ANALYSIS

Picking up on the third data collection point mentioned by Van Haaren (Van Haaren et al., 2019) regarding the camera data tracking type of collection, Wei. et al. (Wei et al., 2013) used 14 hours of continuous ball and player tracking data to analyse large-scale formations in football. On their paper, the first objective was to automatically detect game phases, meaning whether the ball was in play or the game was stopped for some reason, for which they developed a two-layer approach, the first to segment when the match was in play or stopped while the second focus on the stoppage moments to distinguish among the type of stoppages in a match. Using a decision tree classifier, results regarding the first layer were substantially good as tests on detection of in-plays / stoppages were higher than 95% on ball positions' data, averaged around 90% on player positions' data and 83% on the centroid of team positions' data while for the stoppages segmentation, similar results were computed regarding the ball position's observations being the one with higher scores. The second goal concerns the recognition of "role-representations" of players. For this part, the authors performed clustering techniques such as K-Means or Agglomerative Clustering of the frame 10s before a shot and 10s following a corner/free-kick. The results for one team determined that, for that team, offensive plays leading to a goal opportunity started on the left-hand side of the pitch for 33% of the plays, while 18% resulted from a corner taken on the right-hand side. This analysis was also conducted to assess the defensive process of the analysed teams. Their analysis can complement this project's results as they can provide better insights into each coach's/team's preferred game model or favourite plays.

Likewise, Cintia et al. (Cintia et al., 2015) proposed to model the game of a team as a network. Based on a dataset containing events from the 2014 FIFA World Cup and 2013/2014's Italian Serie A with the respective player and location on the pitch, the authors describe the performance of each team by a passing network using the following network measures of a

team – i) a measure of the passing volume, ii) a measure of passing heterogeneity and iii) a combination of the two measures by a harmonic mean. A simulation was then constructed with these measures for each game where the team with the highest measure wins. When compared with 48-26-26[2] and a historic based model that uses the last competition rankings or the FIFA ranking for the FIFA World Cup's predictions, the passing network model outscored both on the Italian Serie A, with a maximum prediction accuracy of 53% and was outscored by the ranking model for the FIFA WC, yielding a maximum prediction accuracy of 36%.

With similar data from the last papers but coming from different competitions, Bialkowski et al. (Bialkowski et al., 2016) attempted to discover formations in data through a role-alignment method based on the minimum entropy data partitioning method directly from player tracking data. Employing an expectation-maximization algorithm, in all similar to K-Means except that at each frame, players must be assigned to a unique role instead of being assigned to its closest cluster; and Hierarchical Clustering, the authors correctly classified 77.5% when assigning formations' patterns discovered in data to six different tactical formations given by a football expert. Following that, match statistics, ball occupancy, and team formations from the previous analysis were used to predict the identity of the analysed teams. The accuracy obtained in this paper was 17.1% from match stats only, 19.5% from just ball occupancy and 67.3% from solely formation analysis. When all combined, the authors reached 70.4% accuracy when testing the team identity prediction model.

All in all, this chapter demonstrated several distinct approaches already performed in football resorting to machine learning's tools and techniques. Moreover, on literature concerning performance prediction, papers utilized distinct machine learning models such as KNNs, XGBoost or Random Forest algorithms and different evaluation metrics as RMSEs, on the case of regressions, or accuracy on classification algorithms. Some of these even performed output explainability of the models used. All the mentioned items can be used as benchmark for the prediction part of this project as a direct comparision can be established. On the other hand, on literature regarding tatical analysis, different types of clustering analysis resorting to models such as K-means, Hierarchical Clustering and pass-networks, and evaluation metrics, namely the usage of accuracy, were performed attempting to find team identities. Some of these clustering techniques will be used also on the clustering part of this project. It is also noteworthy to ascertain that no authors used the Silhouette score from the papers analysed to access their clustering solutions.

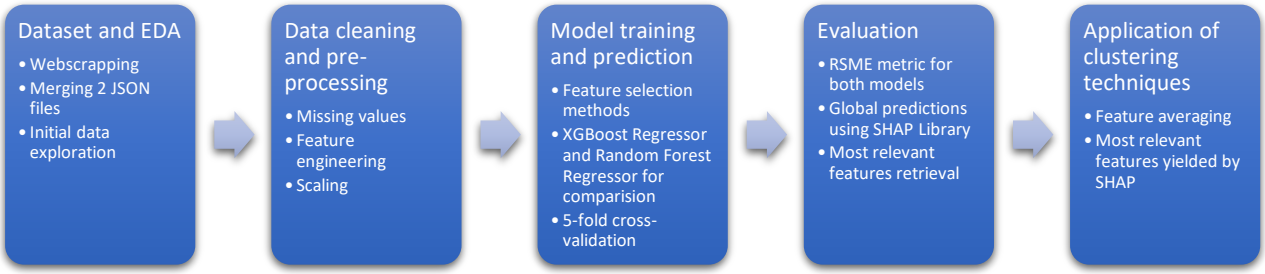
[2] Refers to a created model where the outcome of a match is randomly extracted from a probability distribution computed with the data of the paper, in which 48% is the probability of win from the home team, 26% probability of draw and 26% of a win from the away team.

3. METHODOLOGY

The current chapter will present the methodology approach followed in this project. First and foremost, the methodology chosen was the Cross Industry Standard Process for Data Mining (CRISP-DM), which is regarded as the most common methodology used for Data Science projects as it presents an easy and structured framework to tackle Data Science projects (Schröer et al., 2021). The CRISP-DM methodology follows six main steps:

- Business understanding
- Data access, exploration and understanding
- Data preparation
- Modelling
- Evaluation
- Deployment (not covered in this work)

Figure 1 - Methodology pipeline followed



3.1. BUSINESS UNDERSTANDING

As stated before, the project consists of two distinct but complementary parts. The first aims to create a model that predicts the goal difference of a given team in a game based on the match statistics of the team itself and later analyses the most important features the model used to yield the predictions (hereunder called prediction phase). The second phase of the project comprises the performance of a cluster analysis of different clusters, based on the most relevant features the model used for the goal difference predictions, to determine whether a team, for instance, that adopts a more offensive playing style wins, in fact, more games than other teams (hereunder called clustering phase).

3.2. DATA ACCESS, EXPLORATION AND UNDERSTANDING

For this project, the chosen data comprises match statistics for all games of the Portuguese Football League throughout the seasons 2021/2022 and 2022/2023. The data was scrapped from the internet. The web scrapping phase was performed through two JSON files for each collected match – the first with information about the match, such as the team name, competition or matchday and the second with match statistics - requiring additional pre-

processing to merge them. The web scrapping outcome was a dataset consisting of 40 features, 7 regarding the match information and the remaining factual statistics of each team in that specific game. Table 1 presents the dataset's features by category – for the full feature's breakdown please refer to Appendix A.

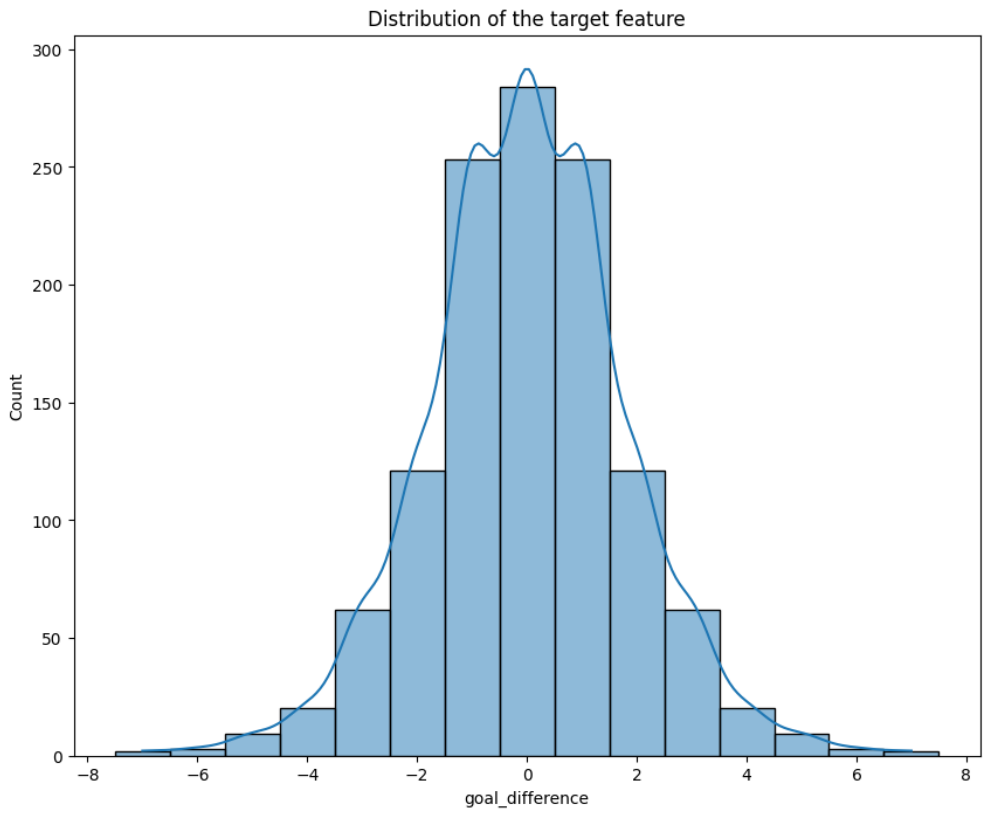
Table 1 - Number of features by category

Match information	7
Shooting features	7
Goalkeeping features	1
Goal-scoring opportunity features	2
Passing features	4
Dribbling features	1
Defensive features	3
Possession features	2
Offensive transition features	3
Other features	10

Regarding the number of observations, each data point corresponds to each team's performance statistics on a specific game, i.e. each match generates two distinct observations (rows), one for each team, which yields a total of 612 observations per season (1,224 observations in total). As observed in Appendix A, there are features in absolute and relative values, such as ball possession, crosses or dribbles, inherent to the statistic as it measures completion ratios. This last format of features required additional pre-processing before the prediction phase.

On an early exploratory data analysis, Figure 2 shows that the target feature on the used dataset follows a normal distribution, obviously symmetrical as one goal scored by one team is the same goal conceded by the other team, centred on the number 0, which is the most frequent goal difference outcome of each team.

Figure 2 - Distribution of the target feature



In Figures 3 and 4 concerning the target feature, one might observe that only four teams had positive goal differences in both seasons. According to the same figures, only SL Benfica, FC Porto, Sporting CP and SC Braga yielded positive goal difference. As mentioned before, SL Benfica, FC Porto and Sporting CP are considered the three greatest teams in the Portuguese league; therefore, it is unsurprising that they have greater goal differences than the others. Similarly, SC Braga presents a positive goal difference as it has been consolidating itself as the fourth leading team in Portugal, closing the gap to the first group. Only two teams, Vitória SC and Gil Vicente FC, had a positive goal difference in one season but could not maintain it in the following. On a brighter note, there was a positive evolution in FC Arouca and FC Vizela's team, significantly reducing its goal deficit from one season to the following. Apart from FC Arouca and FC Vizela in 2021/2022 and Portimonense in 2022/2023, teams with a negative goal difference larger than 20 goals ended up relegated.

Figure 3 - Total goal difference by team in season 2021/2022

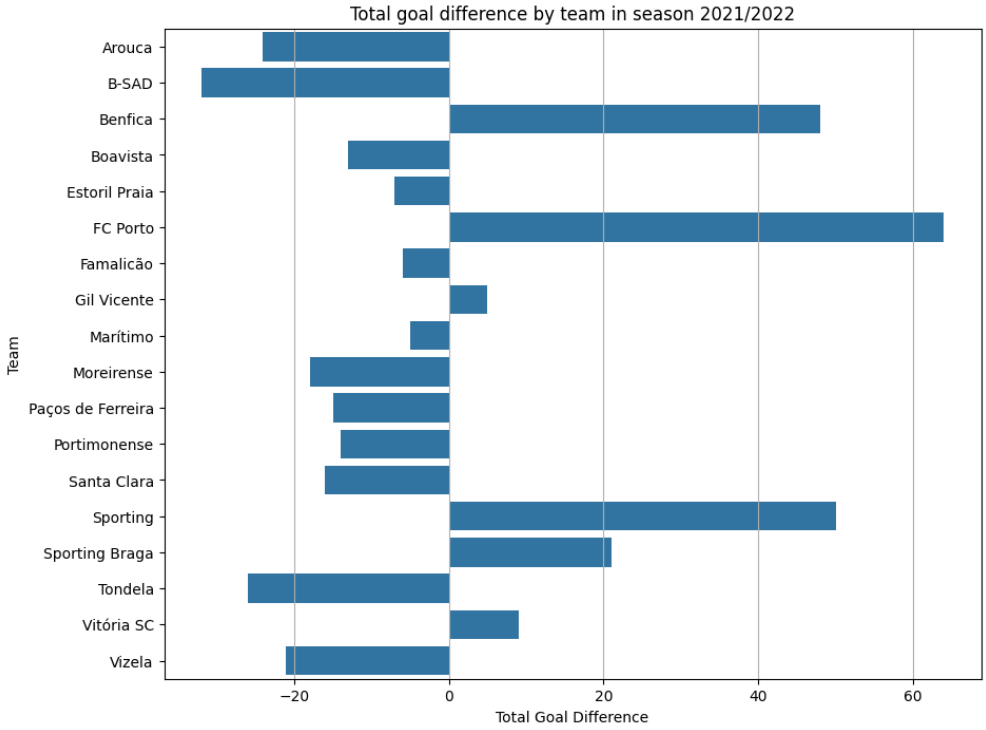
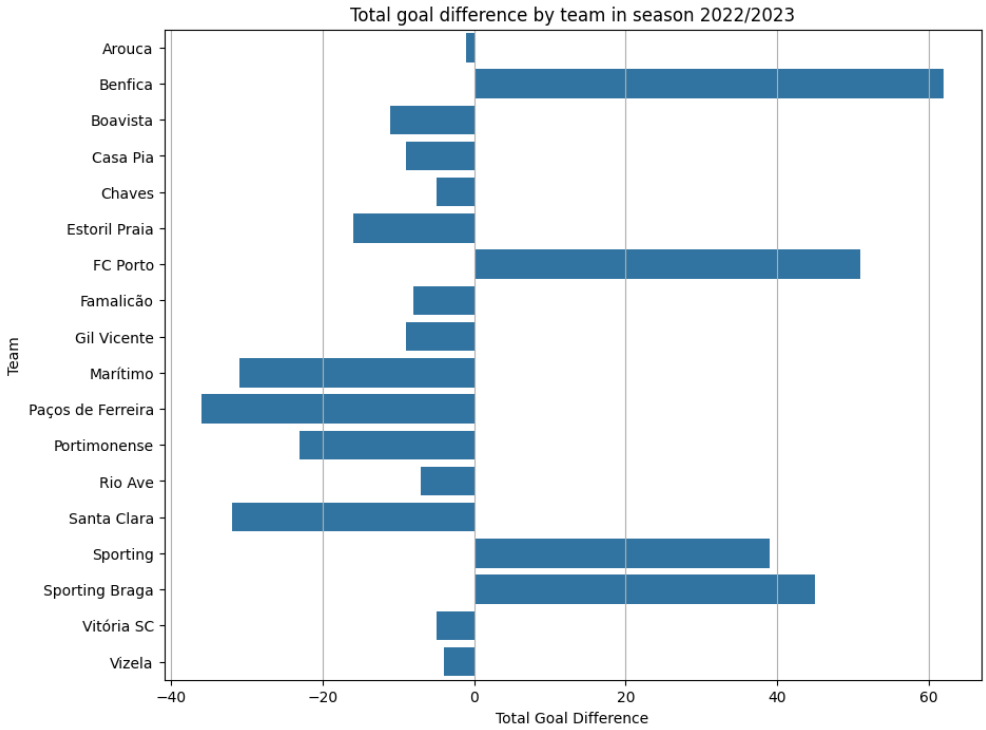


Figure 4 - Total goal difference by team in season 2022/2023



Another piece of information worth highlighting is that the sum of goal difference for all teams when comparing home and away matches in the combination of both seasons is favourable for the home team. This means that, according to the data used for this project, it is indeed an advantage to play at home, with the support of each club's fans. Table 2 below shows that

the home advantage got stronger from the first analysed season to the second as the sum of goal difference for all teams playing at home increased by 34 goals. One possible reason for the difference observed was that at the beginning of the first season, stadiums were not at full capacity due to the COVID-19 pandemic, initially at 33% and later at 50% (*O regresso dos estádios cheios*, 2021).

Table 2 - Total goal difference by location of the match

Total goal difference	Home	Away
In season 2021/2022	69	-69
In season 2022/2023	103	-103

3.3. DATA PREPARATION

After analysing the initial insights from the dataset, it was needed to perform several pre-processing steps to guarantee that there were not any inconsistencies, errors, or missing values and assess whether there was the need to perform data transformation to the input data, which most of the time is noisy, into a cleaner dataset to ensure the quality of the results yielded from the models used.

The main steps of this stage corresponded to i) Access any incoherencies or mistakes in the data and possible missing values; ii) The performance of feature engineering and transformations, and iii) Scaling and standardization of the features.

3.3.1. Inconsistences and missing values

The dataset assembled from the web scrapping process had several missing values spread over many features. However, these were all features that could not have happened in the course of a match. For instance, there are missing values on features such as counterattack goals (1076 missing values), red cards (868), shots to woodwork (636), and offsides, that presented a pattern on which it was considered safe to assume that the missing values corresponded to events that did not happen on that specific match, which were; therefore, all set to 0. Appendix B presents a heatmap of missing values and their location.

3.3.2. Feature engineering

Further than dealing with the missing data, several feature engineering was needed to clean the dataset. First and foremost, the scrapped dataset did not include the target feature, only the number of goals scored and conceded, from which the goal difference feature was computed. Moreover, many features were presented in relative values (i.e., in percentage) or with the number of correct attempts and the percentage of completion for a given feature (for instance, crosses or dribbles). The percentage sign was removed on the former, and the value was divided by 100. All the merged information was split and stored in separate features for the latter. This process added another 16 features to the initial dataset.

Later, in the clustering phase, all statistics were averaged by the team for both seasons to perform the clustering analysis. This step was performed to give meaning to the cluster results as they represent a numeric approximation of each team's overall game playing style approach in each season. Therefore, this transformation reduced the original dataset to 36 observations for the clustering phase, 18 data points per season.

3.3.3. Scaling and standardization

Machine learning algorithms can differ on several topics. One of the most important is the type of input data an algorithm can take and whether it only accepts normalized data. Data scaling or standardization is essential for several algorithms as their process includes the calculation of distances between different features. This process aims to prevent some features from dominating others. In the literature review, concerning this topic, Moustakidis et al. used a standard scaler to perform feature scaling on their dataset.

Moreover, in previous research, Ahsan et al. (Ahsan et al., 2021) analysed the impact of different scaling and normalization techniques on several algorithms in the context of heart disease prediction, including those used in the project. The main conclusions regarding XGBoost and Random Forest were that scaling does not provide significant improvements, with results being similar on scaled and not scaled datasets, independent of the technique used. However, to avoid some features dominating others during the training process of both phases, we decided to use it on the project. Therefore, in this project, standard scaling was used to normalize the dataset in the prediction and clustering phases.

3.4. MODELLING

The prediction part of the project was mainly focused on using the XGBoost and SHAP despite using an alternative algorithm for benchmark purposes. The alternative algorithm chosen was the Random Forest algorithm, as it generates outcomes on which the features that contributed more for the final output can be accessed.

Nevertheless, before the model implementation, two categories of feature selection techniques were applied: filter and wrapper. While filter methods evaluate each feature only based on its relationship to others and to the target feature, wrapper techniques assess subsets of features through iterative training and testing of models, considering the model's performance with various combinations. In this project, the Pearson correlation method was used as a filter method to discover if there are any redundant features, and the Recursive Feature Elimination (RFE) was used as a wrapper method.

From the analysis of the filter method, please refer to the correlation matrix in Appendix C, which shows four different correlated pairs that were observed. The first concerned features *Total shots* and *Shots inside box,* for which we created a feature computing the percentage of shots inside the box over the total number of shots. The same situation happened between *Big chances* and *Big chances missed*. The third pair regarded *Counter attacks* and *Counter*

attacks shots, resolved by dropping *Counter attacks* as the latter can be considered a good proxy of the first. The fourth was between the number of *Passes* and *Ball Possession*. It was decided to drop *Ball Possession* from the dataset as it is more correlated with the other features compared to the former.

One of the most common problems presented on data science projects is overfitting, in which the model performs well on the training data but generalizes rather badly to new data (Grus, 2015). To avoid this, one solution is to use k-fold cross-validation. This method partitions the original data into k independent and similar subsets, then trains in k-1 subsets and tests on the subset remaining for k number of times and averages the results (Larose & Larose, 2015). A 5-fold cross-validation implementation process was applied to avoid overfitting, using the Pipeline function of the Scikit-learn library. It also used the Randomized Search and the Grid Search algorithm to optimise the combination of hyperparameter values on the models used (Raschka & Olson, 2015).

To retrieve insights from the predictions performed during the prediction phase of the project, the Shappley Addictive explanation (SHAP) values were used. These values can be defined as a unified measure of feature importance and are yielded by attributing each feature, which changes the expected model prediction when conditioning on that feature (Lundberg & Lee, 2017). In this project, SHAP was used to quantify the impact of a given feature has on an individual prediction performed by the model, allowing for an individual analysis of the most important features for each team and the most essential features for the model overall (Moustakidis et al., 2023).

For the second part of the project, a combination of the most relevant features yielded from SHAP global value, according to each feature impact on the model, were used to cluster the average statistics of each team into different clusters using K-means and Hierarchical clustering as algorithms. DBScan was also used to help decide the optimal number of clusters, as this number is not defined beforehand in this algorithm. For this phase, three different approaches were followed: i) the first consisted of using the top 10 most essential features of the XGBoost model, ii) the second comprised of the usage of the combination of the top 10 most relevant features yielded by both models (XGBoost and Random Forest) and iii) the last used also the top 10 most essential features of the XGBoost. However, a correlation matrix was computed for these features, and one of the features from the heavily correlated pairs was discarded. It is also vital to highlight that if two or more features measured the same statistics and were on the top 10 for any approach, only one was accounted for (for instance, if passes and pass completion were present on top 10 most relevant feature, one of them was dropped for clustering purposes).

3.5. EVALUATION

Once both feature selection and modelling implementation for predictions are concluded, the Root Mean Squared Error (RMSE) metric was chosen to evaluate each model's accuracy. The

metric was selected for the results to be comparable with the results of Moustakidis et al. (Moustakidis et al., 2023).

For the clustering phase, the evaluation metric used will be the Silhouette scores, as this method can be used as a graphical tool to plot the measurements of how tightly grouped the observations on the clusters are (Raschka & Olson, 2015). Additionally, the sum of squared errors was computed for the clustering solutions and used as a secondary evaluation metric to access the quality of each solution (Humaira & Rasyidah, 2020).

3.6. TOOLS

For this project, Python was chosen for its efficiency, interpretability, and the number of already developed free libraries. These libraries include Pandas for data analysis and manipulation, NumPy for numeric computations, Scikit-learn for machine learning, and even Matplotlib or Seaborn for data visualization. Additionally, Jupiter Notebook and Google Collab were the web-based platforms used to develop the project. Concerning hardware, Table 3 presents the computer specifications used for the model training.

Table 3 - Computer specifications

Model	MacBook Pro
Chip	Apple M3
CPU	10-core
Memory	8GB LPDDR5
GPU	10-core

4. RESULTS AND DISCUSSION

This chapter presents the results of the prediction phase, the importance of the global model feature with SHAP, and the results of the clustering phase.

4.1. PREDICTION PHASE

As previously mentioned, the prediction phase consisted of predicting the goal difference of each team in each match to generate the most relevant statistics for the model to use in the clustering phase.

Therefore, in this phase, the main algorithm used was the XGBoostRegressor with a 5-fold cross-validation implementation to avoid excessive overfitting. The Random Forest algorithm was also computed as a benchmark algorithm with a 5-fold cross-validation implementation. Furthermore, both Randomized Search and Grid Search algorithms were used for hyperparameter optimization in both cases. Final RMSE results are shown below in Table 4, with the optimal hyperparameters yielded by the Grid Search algorithm for both models.

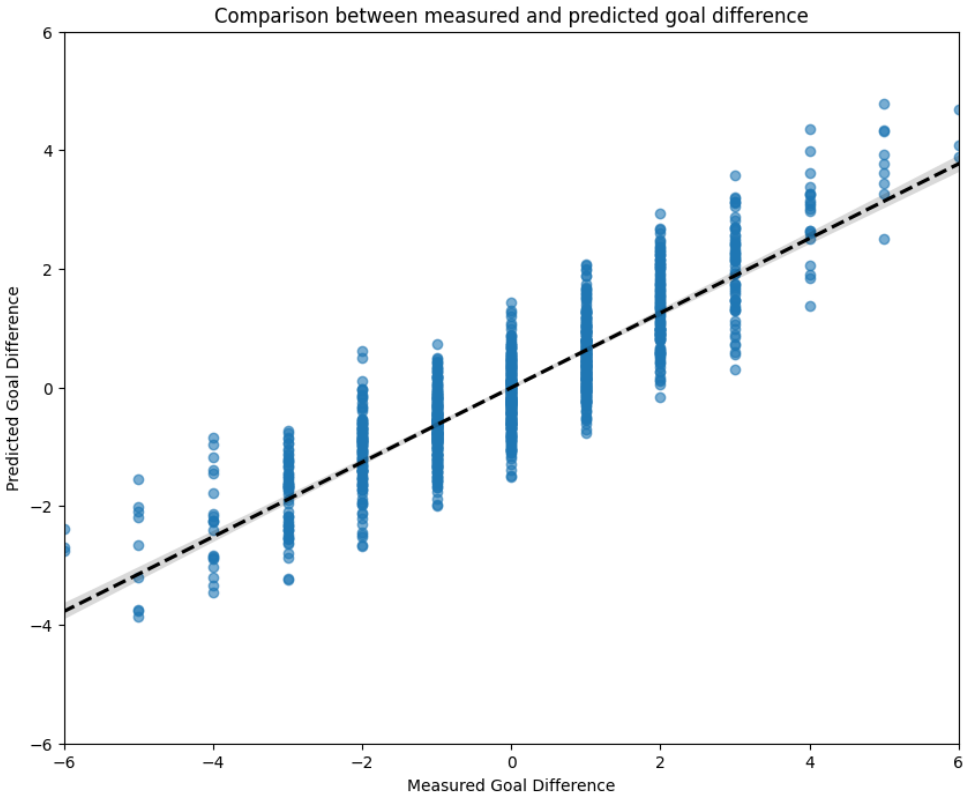
Table 4 - Machine learning model's parameters and performance

Model	**XGBoostRegressor**	**RFRegressor**
Hyperparameters	”Colsample_bytree”: 0.4 “Learning_rate”: 0.05 “Max_depth”: 4 “N_estimators”: 100 “Subsample”: 0.75	”Max_depth”: 5 “Min_samples_leaf”: 2 “N_estimators”: 150
Validation approach	5-Fold CV	5-fold CV
Performance (RMSE)	Train: 0.895 Test: 1.257	Train: 1.085 Test: 1.321

As observed in Table 4, the XGBoost provided lower RMSE training and test results compared with the benchmark algorithm, although by a small margin on the latter. It is worth noting that, despite cross-validation, the results obtained still presented overfitting when comparing the training against the test results, particularly in the first model. Figure 5 shows the comparison of the predicted goal difference per team against the actual goal difference scored with the XGBoost’s model. The results obtained from the prediction phase are lower than the ones Moustakidis et al. reached. This might be due to a lower number of data points shown

to the model, as the dataset used for their paper comprised only one season of data but encompassed matches from 11 different leagues (5,992 observations from 2,996 matches).

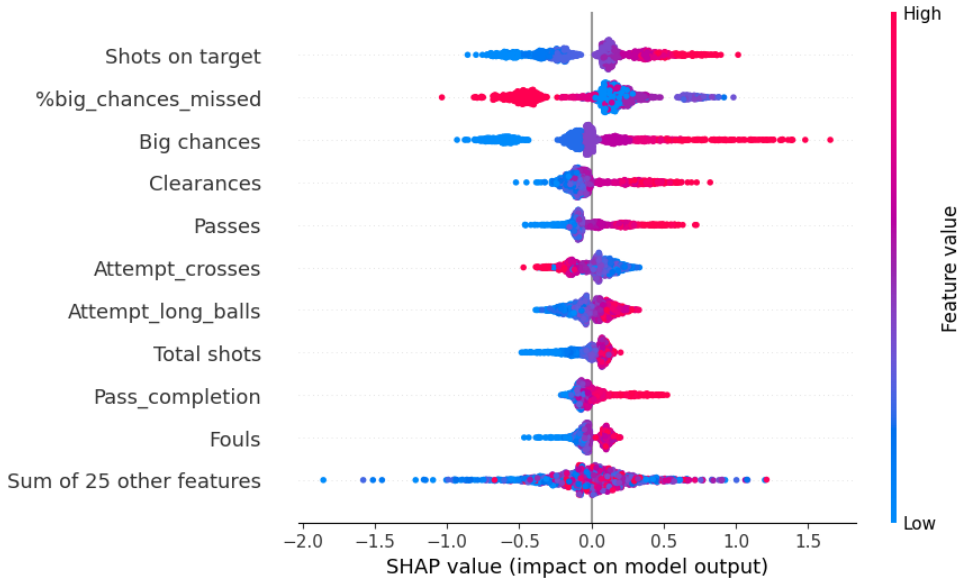
Figure 5 - Comparison between measured and predicted goal difference from XGBoost



The following step was using the SHAP library to extract the most relevant features for global predictions from the prediction model. Figure 6 presents a summary of how the top 10 features in the dataset impact the XGBoost model's output. Each dot on the plot represents the feature's impact on model output, and each colour represents the feature value.

The SHAP value summary (Figure 6) can already provide insights about the features that lead to a better (or worse) goal difference. For instance, starting on the most essential features, the higher the value of big chances and shots on target performed by a team, the better the goal difference outcome in that game. The high number of both big chances created or shots on target performed by a team can be associated with an offensive approach; therefore, from the first two features, we can infer that the offensive teams are, on average, closer to winning. Other playing style defining features are the number of passes and the pass completion, which, according to the figure, the higher the number of passes and percentage of passes completed, the better the outcome of the goal difference. The SHAP summary dot plot also shows that the number of clearances positively impacted the outcome of the goal difference, as did the number of long balls attempted and the number of fouls. On the other hand, attempted crosses contribute to a worse goal difference.

Figure 6 – XGBoost's SHAP summary dot plot



As observed in Figure 6, the ten most important features were mainly related to the offensive side of the game, which was expected due to the low number of defensive-related features in the dataset. Nevertheless, clearances are undoubtedly of high relevance for the model. Appendix D presents the same SHAP summary dot plot for the random forest model. Despite small differences in the order of importance of features for the model, nine out of the ten most relevant features are the same. As seen in Table 5 below, only the *Fouls* features are present on the XGBoost's model and not on the Random Forest's model, which Aerials Won are replacing. This increases the confidence level on the importance of the features for the clustering part.

Table 5 - Comparison between models' top 10 features

Feature	**Rank on XGBoost**	**Rank on RF**
Shots on target	1	2
% of Big Chances Missed	2	3
Big Chances	3	1
Clearances	4	4
Passes	5	5
Attempted Long Balls	6	6
Attempted Crosses	7	8
Pass Completion	8	10
Total Shots	9	9
Fouls	10	n.a.

Aerials Won	n.a.	7

There are a few relevant statistics that coincide with Moustakidis et al.'s work as, for their model, features such as 'missed chances', 'passes', and passes-related statistics were also highly relevant (note that the authors used 141 different features in their model). Moreover, Pappalardo and Cintia mentioned that the number of passes, shots and goalkeeping actions were the strongest predictors of a match outcome, which is consistent with the results from both models (except for the goalkeeping actions since the dataset obtained for this project does not encompasses such statistics but only goalkeeping saves). Nevertheless, these are the features on which the clustering phase will focus.

4.2. CLUSTERING PHASE

The following phase regards the clustering of the average statistics of each team using the most important features for the model computed on the previous phase.

As stated in the previous chapter, three approaches were considered for this part to assess which one computed the best result. The first approach used XGBoost's top 10 most relevant features, the second combined the ten most important features of both models, and the third used a correlation matrix to discard heavily correlated features. Table 6 presents the Silhouette scores, the sum of squared errors and the R2 scores per the number of clusters using both K-Means and Hierarchical Clustering. Despite not being presented in Table 6, DBScan was also performed solely to help decide the number of clusters chosen – which yielded two cluster solutions in all approaches.

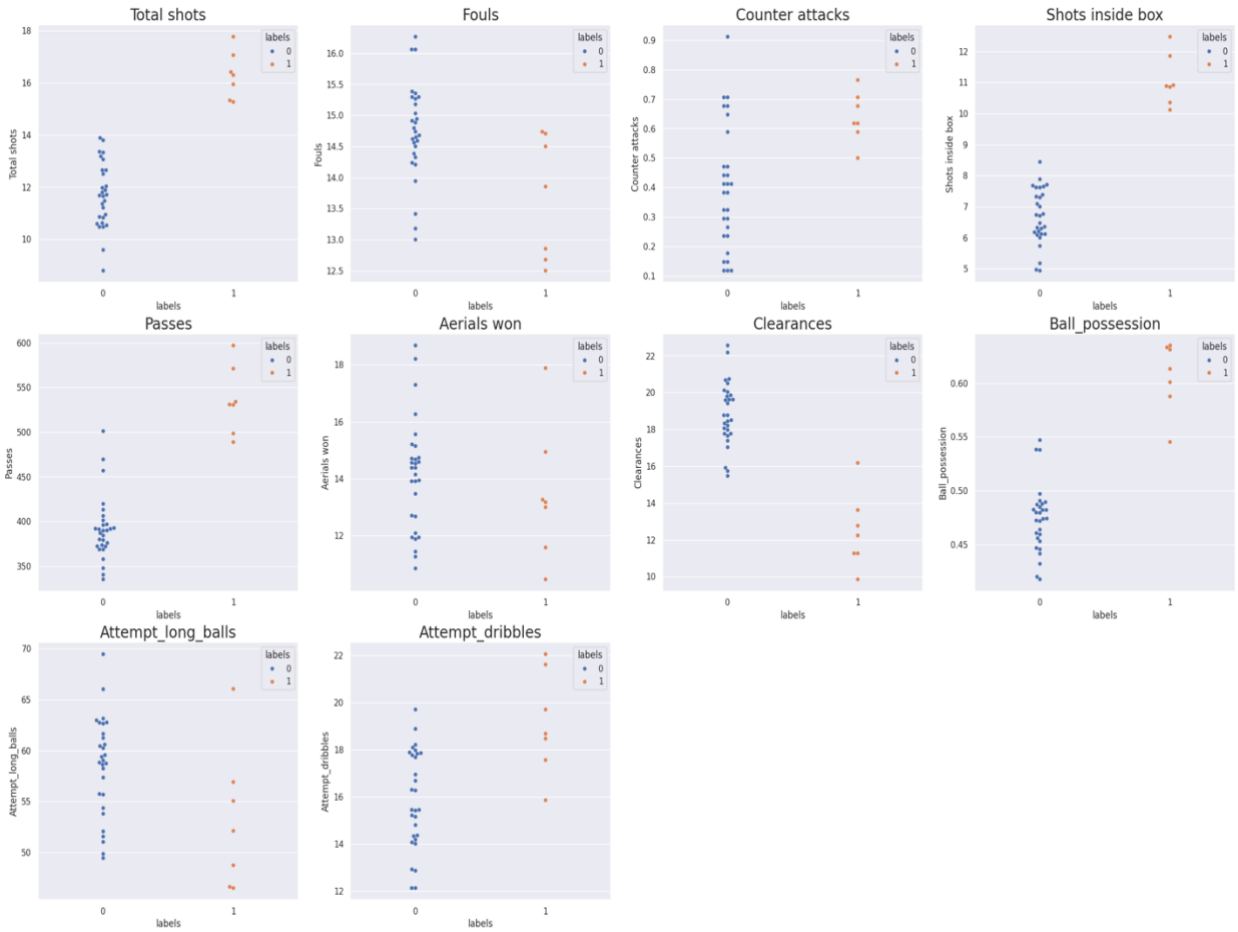
Table 6 - Silhouette, SSW and R2 scores for each algorithm per number of clusters

	Approach 1			**Approach 2**			**Approach 3**		
Model	**Silhouette score**	**SSW score**	**R2 score**	**Silhouette score**	**SSW score**	**R2 score**	**Silhouette score**	**SSW score**	**R2 score**
KM2C	0.544	45669	0.722	0.497	45811	0.722	0.479	45656	0.722
KM3C	0.243	44573	0.729	0.246	44304	0.725	0.230	44660	0.728
HC2C	0.488	30861	0.812	0.439	31004	0.812	0.479	45656	0.722
HC3C	0.410	23738	0.856	0.385	22296	0.864	0.216	45202	0.725

Comparing the scores from each approach, the first approach emerges as the clear winner as both the Silhouette score and sum of squared errors are better than the others for the same

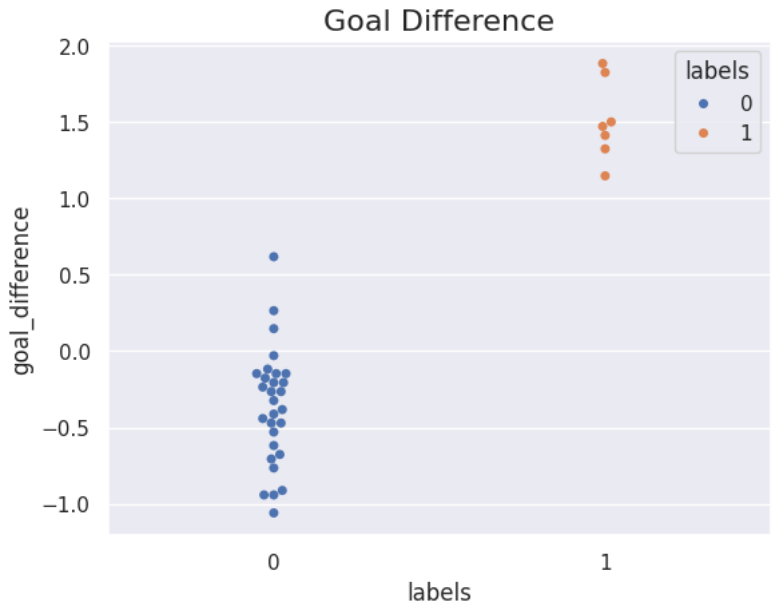
model and number of clusters. Therefore, the final clustering solution will be drawn from approach 1. Despite the features that were used to perform the clustering solution being the top 10 features generated by the XGBoost model – shown above in Figure 6 -, the features that will be compared and subject to analysis and discussion in this section will be the following T*otal shots, Fouls, Counter attacks, Shots inside the box, Passes, Aerials won, Clearances, Ball possession, Attempt long balls* and *Attempt dribbles* as these define better the playing style approaches of the different teams and coaches. Moreover, among the solutions of the first approach, it becomes clear that neither the K-means solution with three clusters nor the Hierarchical Clustering solution with three clusters produced good clustering solutions due to their Silhouette scores. Regarding the remaining solutions, even though the Hierarchical Clustering has a lower number on the sum of squared errors metric, not only the Silhouette score of the K-means is better, but the clusters concerning the playing style defining features are clearer defined as shown below and on Appendix E, which uses the T-SNE algorithm in an attempt to visualize high dimensionality clusters. Below, in Figure 7, are swarm plots of the K-means solution with two clusters that present each feature by cluster – Appendix F presents the swarm plots of the two clustered Hierarchical Clustering solution.

Figure 7 - Swarm plots from approach 1's K-means solution with two clusters presenting playing style features per cluster



Observing Figure 7, despite the unbalance in the number of teams in each cluster, one can conclude that there is one cluster – cluster 1 - where teams perform a higher number of shots in total and shots inside the opponent's box, complete a higher number of passes, on average shorter, with a higher ball possession per match; attempt more dribbles on average and have a lower number of clearances. By these statistics, one might infer that the teams composing this cluster hold the ball for longer periods than the opponent, play closer to the opponent's box, which allows for defending less time and upper in the field, thus the lower number of clearances, and necessity to commit fouls. Also, the number of attempted dribbles show a higher level of allowed creativity by these teams. Therefore, assuming these teams display a more offensive playing style approach is safe. As mentioned in the introduction section, the purpose of the project was to attempt to find if each team's game model has an influence on the results of the team, or put in other words, if a more attacking playing style would win more games and vice-versa, which on this project is measured through the average goal difference outcome of each team. Now that it is established that the teams in cluster 1 display, on average, a more attacking playing style, they should, in theory, also display a higher goal difference outcome. According to the swarm plot presented below in Figure 8, which shows the average goal difference by each team per cluster, the average goal difference outcome of the teams included in cluster 1 is significantly higher than the remaining.

Figure 8 - Swarm plot from approach 1's K-means solution with two clusters presenting goal difference outcome per cluster



Analysing the composition of each cluster, cluster 1 is composed of 7 teams: SL Benfica (both seasons), FC Porto (both seasons), Sporting CP (both seasons) and SC Braga (season 22/23). This outcome can be considered to be expected since, as previously stated on the data access, exploration and understanding section of the methodology, the first three teams are historically the best and strongest teams in Portugal. Braga, on the other hand, is not yet consistently performing as the other three teams since there are seasons in which the club is

competitive enough to almost challenge for the title and there are seasons on which it is fighting to obtain the fourth place. Concerning the difference of these four teams to the remaining, to put in perspective, the average squad value of cluster 1 amounts to around €333m[3] comparing with the average of €34m of the other cluster. These numbers tell us that both the amount of talent and the depth of the squad of these teams can enable them to perform a more offensive playing style approach.

Nevertheless, it is quite surprising that only one season of SC Braga is included on the cluster 1 since although it is not on the level of the other three teams, it still has a squad value greater than the remaining teams. The squad value decreases from 2021/2022 from €159m to €150m in season 2022/2023. One defining change that occurred from one season to the other was the coach, as Carlos Carvalhal, who coached the team for two seasons until the end of 2021/2022, was replaced by the second team coach Artur Jorge who deployed a more attacking oriented playing style, objectively resulting on more than doubling the average goal difference of the team – from 0.62 net goals per match to 1.32 net goals per match (please refer to Appendix G for playing style defining features comparison) – despite losing one of the main defenders (David Carmo to FC Porto for €20m) and one of the main strikers (Vitinha to Olympique Marseille for €32m) in January's transfer market.

Furthermore, it is important to highlight that the solution found (only two cluster solution) is, most probably, not well generalizable other than for the Portuguese football environment as the distance in both quality and talent between the top three (four including SC Braga) clubs and the remaining teams is extremely large. Nevertheless, we expected to have a more diverse set of remaining teams that would result in a different solution – for instance, a three-cluster solution with higher Silhouette scores – that would split cluster 0 into a more defensive or balanced playing style approach. However, when attempted to figure out why that is the case, it was found that there is a highly turnover of coaches on these clubs. For instance, out of the 29 teams in the cluster 0, only fourteen did not change their coach during the season. Moreover, there was one team that changed coach three times during the season (FC Paços de Ferreira 2022/2023), five teams changed coach twice and the remaining eight replaced their coach only once.[4] With these many alterations it might be difficult to find a playing style 'identity' as each coach might have a different one from the last or the next. Therefore, there is the possibility that the average statistics could incorporate several distinct playing styles and not one identifiable.

On the other hand, a specific playing style change was generated by a coach replacement that the best clustering solution did not differentiate. Here, we want to address the introduction of Roger Schimdt that replaced Nélson Verissimo in 2022/2023. SL Benfica changed its coach in the middle of 2021/2022 letting go Jorge Jesus and replacing him by the second team coach Nelson Verissimo that even reached the quarterfinals of the Champions League, however the

[3] Amounts of squad values retrieved from Transfermarkt.pt (*Transfermarkt.pt*, n.d.).

[4] Coaches' tenure was retrieved from ZeroZero.pt (*zerozero.pt*, n.d.).

team was not playing nearly as offensive than the gegenpress[5] Schimdt would have implemented on the next season. As presented in Appendix H, Schimdt's team performed, on average, almost two more shots (in total and inside the box), nearly 70 more passes per game and almost two percentage points more ball possession, less clearances and fouls. Also, the goal difference in 2022/2023 increased to 1.82 net goals per match from 1.41 net goals per match in the previous season. Hence, despite the observed differences, it is noteworthy that the best clustering solution did not distinguish among both seasons. Nevertheless, it is understandable on a two-cluster solution as the difference between these seasons is not as high as the differences between the top four clubs and the remaining.

[5] Gegenpress is a word derived from the German language that can be translated to "counter-pressing", which is a strategy used to disrupt the opponent team as soon as the possession is lost through aggressively pressing the ball and the opponents' players near the ball with several players. The main goal is to get the ball possession as soon as possible ("Counter-Pressing and the Gegenpress," n.d.).

5. CONCLUSIONS AND FUTURE WORKS

In this last section, a succinct summary of the work performed throughout this research project will be presented alongside the main conclusions. Moreover, the limitations surrounding the project and possible recommendations for future work will be provided.

The aim of this project was to find out, through a machine learning data-based approach, if the outcome of a given match was influenced by the playing style approach of that team, for instance, whether a more offensive minded team wins more games than a defensive one.

In order to get an answer to the above mentioned proposed goal, the CRISP-DM methodology was followed to structure the development of the project into the following distinct stages: business understanding, data access, exploration and understanding, data preparation, modeling and evaluation. Throughout this stages, the dataset was scrapped from the internet, then manipulated and prepared for the modeling and evaluation stages.

Moreover, this project consisted in two separate but complimentary parts with one main objective each. The first part consisted on the elaboration of a prediction model that attempted to predict the goal difference of a given team based on their match statistics for that game. The purpose and aim of this phase was to use model interpretability tools such as the SHAP library to identify what were the most important features (top 10) for the prediction model. The second part consisted of the usage of these most relevant features identified on the last phase to cluster the aggregated data, i.e. data averaged by team, using unsupervised models.

Regarding the results obtained, the first objective was met as it was possible to predict the goal difference for each team in a given match, even though with lower RMSE's scores than the literature reviewed. Furthermore, feature importance was performed using the SHAP library to identify what were the ten most relevant match statistics for the prediction model. Concerning the clustering phase, the best result was a two-cluster solution which, despite the existing unbalance between both clusters, it became clear that both represented completely different ways of playing style approaches, where the cluster 1 represented a significantly more offensive approach, supported by a higher number of shots (on target and in total), higher number of passes and pass completion rates, which usually infers domination over the opponent, and a lower number of attempted long balls. On the defensive end, cluster 1 has less clearances and a lower number of fouls per game. The identified playing style differences between both clusters meet the second main objective of the project, which was to attempt to identify distinct group of teams performing similar playing style approachs.

All in all, the average goal difference for the teams in cluster 1 is significantly higher than to the ones in cluster 0, meaning that in the past two Portuguese leagues, the most offensive teams got better match outcomes than the ones without an offensive playing style, which answers positively to the research gap identified as according to the results obtained, teams

with distinct playing styles generated different results, namely the more offensive the playing style, the best result a team obtained, on average.

5.1. LIMITATIONS AND FUTURE WORK

The positive answer to the research gap can be due to several different variables. Some of the reasons suggested in the results and discussions chapter concerns the large difference in talent or quality players among the teams in each cluster or due to the high rotation of coaches in the teams of cluster 0 that did not enable to identify specific approaches.

In future research projects, combining data from several leagues will prevent the exposure to the idiosyncrasies of each context or country, as teams from several environment and contexts are joined together. Furthermore, it would also be useful to have more teams to perform the clustering solution as, on this project, when aggregated, the dataset was reduced to 36 data points, therefore using either more teams or more seasons would definitely prevent this issue and also could yield great solutions with more clusters, thus identifying more and distinct playing styles. Besides, adding more teams or seasons would help on obtaining better results – on this project measured by RMSE - from the prediction model as it was what Moustakidis et al. did and indeed achieved better scoring metrics.

To sum up, this project provides a data-driven framework that can help on the identification and comparison of distinct playing style approaches and access whether they are yielding better results or not. It can be used and developed by relevant industry stakeholders for a better analysis of performance.

BIBLIOGRAPHICAL REFERENCES

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), Article 3. https://doi.org/10.3390/technologies9030052

Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., & Fookes, C. (2016). Discovering Team Structures in Soccer from Spatiotemporal Data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2596–2605. https://doi.org/10.1109/TKDE.2016.2581158

Blakemore, E. (2023, June 8). *Soccer is the world's most popular sport. But who invented it?* Premium. https://www.nationalgeographic.com/premium/article/soccer-world-cup-origins-mesoamerica-football-games-archaeology

Cintia, P., Rinzivillo, S., & Pappalardo, L. (2015). *A network-based approach to evaluate the performance of football teams*.

Counter-pressing and the gegenpress: Football tactics explained. (n.d.). *The Coaches' Voice*. Retrieved June 30, 2024, from https://www.coachesvoice.com/cv/counter-pressing-gegenpressing-football-tactics-explained-klopp-guardiola-bielsa-hasenhuttl/

Evans, S. (2022, August 16). *Premier League celebrates 30 year rise to global dominance | Reuters*. https://www.reuters.com/lifestyle/sports/premier-league-celebrates-30-year-rise-global-dominance-2022-08-16/

Football Benchmark. (2023). *The European Elite 2023*. Football Benchmark. https://www.footballbenchmark.com/documents/files/public/Football%20Benchmark%20Football%20Clubs'%20Valuation%20The%20European%20Elite%202023.pdf

Futebol domina audiências. (2024, January 8). FPF. https://www.fpf.pt/pt/News/Todas-as-notícias/Notícia/news/43072

Grus, J. (2015). *Data science from scratch: First principles with Python* (First edition). O'Reilly.

Hawkey, I. (2020, June 3). *Portugal's Primeira Liga remains a brilliant nursery of talent even as clubs struggle in Europe*. The National. https://www.thenationalnews.com/sport/football/portugal-s-primeira-liga-remains-a-brilliant-nursery-of-talent-even-as-clubs-struggle-in-europe-1.1028153

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, *14*(6), 798–817. https://doi.org/10.1177/1747954119879350

Humaira, H., & Rasyidah, R. (2020, January 1). *Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm*. https://doi.org/10.4108/eai.24-1-2018.2292388

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553. https://doi.org/10.1016/j.knosys.2006.04.011

Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics* (2. ed). Wiley.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.

May, J., Roche, C., & Reidy, P. (2023, June 13). *Champions League final vs Super Bowl: Which is the most watched sporting event?* Diario AS. https://en.as.com/soccer/super-bowl-vs-champions-league-final-which-is-the-most-watched-sporting-event-n/

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. *Future Internet*, *15*(5), Article 5. https://doi.org/10.3390/fi15050174

O regresso dos estádios cheios: DGS levanta todas as restrições à lotação nos recintos desportivos. (2021, September 30). Tribuna Expresso. https://tribuna.expresso.pt/coronavirus/2021-09-30-O-regresso-dos-estadios-cheios-DGS-levanta-todas-as-restricoes-a-lotacao-nos-recintos-desportivos-dc9338f4

Pappalardo, L., & Cintia, P. (2018). Quantifying the relation between performance and success in soccer. *Advances in Complex Systems*, *21*(03n04), 1750014. https://doi.org/10.1142/S021952591750014X

Raschka, S., & Olson, R. S. (2015). *Python machine learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Packt Publishing open source.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

The International Football Association Board. (2016). *130th Annual General Meeting of The International Football Association Board* (Minutes). The International Football Association Board. http://static-3eb8.kxcdn.com › documents

Transfermarkt.pt. (n.d.). Retrieved May 25, 2024, from https://www.transfermarkt.pt/

Van Haaren, J., Robberechts, P., Decroos, T., Bransen, L., & Davis, J. (2019). *Analyzing Performance and Playing Style Using Ball Event Data*.

Wei, X., Sha, L., Lucey, P., Morgan, S., & Sridharan, S. (2013). Large-Scale Analysis of Formations in Soccer. *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. https://doi.org/10.1109/DICTA.2013.6691503

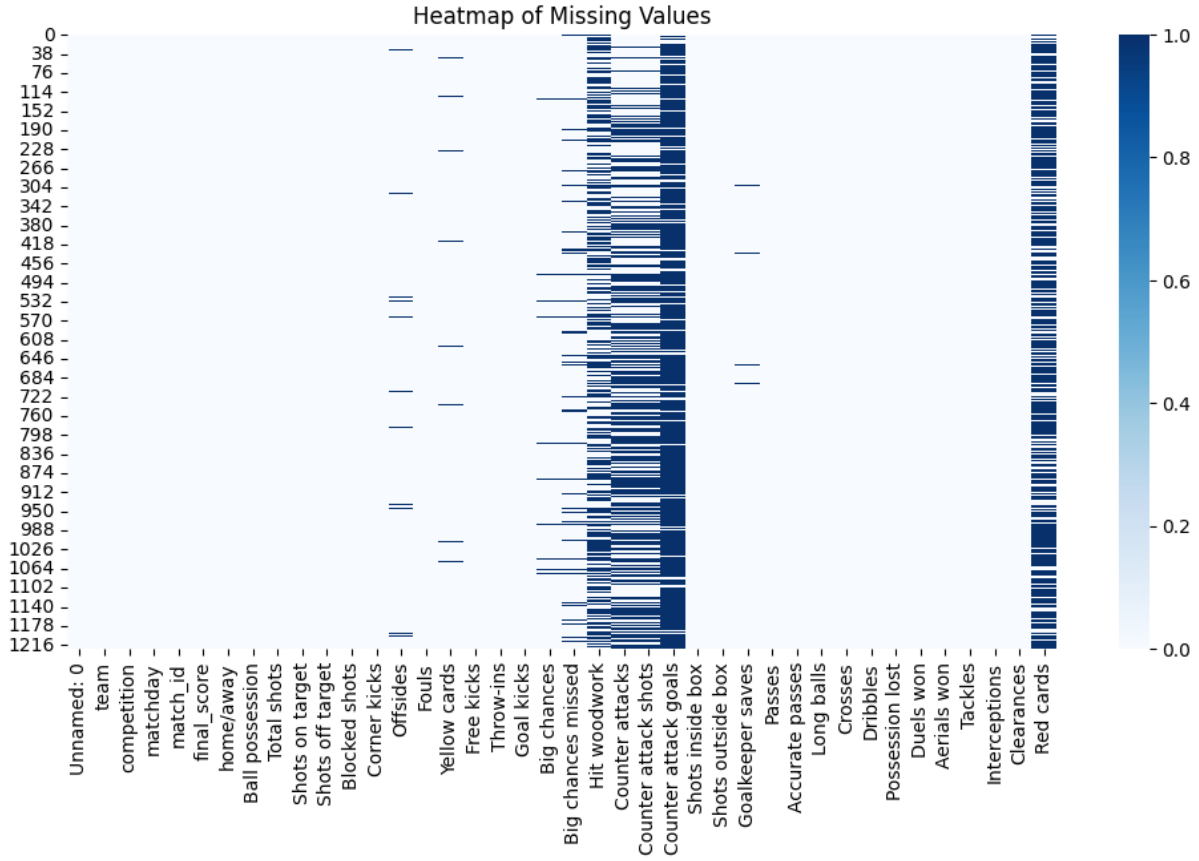
zerozero.pt: Porque todos os jogos começam assim... (n.d.). www.zerozero.pt. Retrieved May 25, 2024, from https://www.zerozero.pt/

APPENDIX A – INITIAL DATASET'S FEATURE LIST

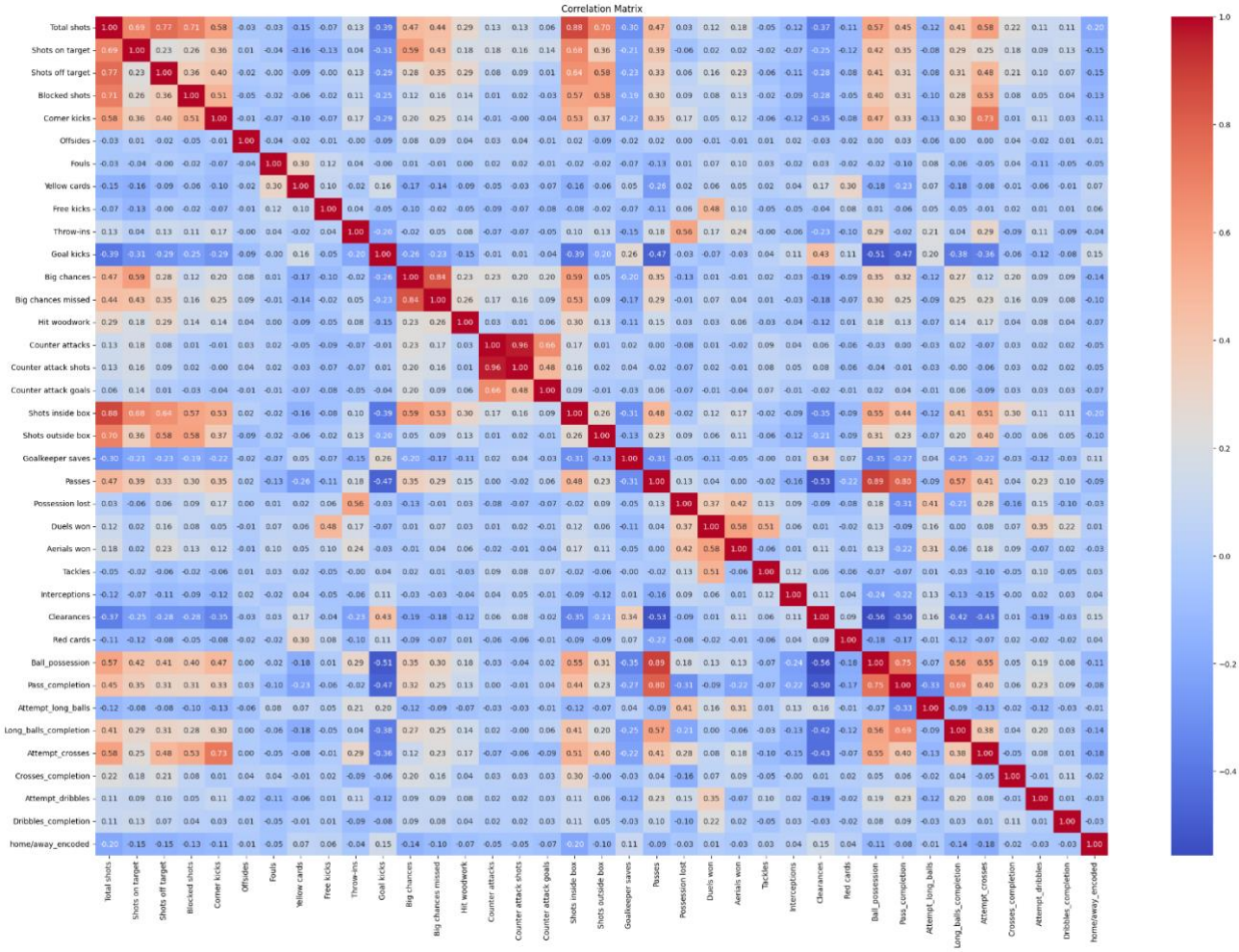
#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1224 non-null	int64
1	team	1224 non-null	object
2	competition	1224 non-null	object
3	matchday	1224 non-null	int64
4	match_id	1224 non-null	int64
5	final_score	1224 non-null	object
6	index	1224 non-null	object
7	Ball possession	1224 non-null	object
8	Total shots	1224 non-null	int64
9	Shots on target	1224 non-null	int64
10	Shots off target	1224 non-null	int64
11	Blocked shots	1224 non-null	int64
12	Corner kicks	1224 non-null	int64
13	Offsides	1188 non-null	float64
14	Fouls	1224 non-null	int64
15	Yellow cards	1208 non-null	float64
16	Free kicks	1224 non-null	int64
17	Throw-ins	1224 non-null	int64
18	Goal kicks	1224 non-null	int64
19	Big chances	1196 non-null	float64
20	Big chances missed	1100 non-null	float64
21	Hit woodwork	586 non-null	float64
22	Counter attacks	642 non-null	float64
23	Counter attack shots	642 non-null	float64
24	Counter attack goals	150 non-null	float64
25	Shots inside box	1224 non-null	int64
26	Shots outside box	1224 non-null	int64
27	Goalkeeper saves	1216 non-null	float64
28	Passes	1224 non-null	int64
29	Accurate passes	1224 non-null	object
30	Long balls	1224 non-null	object
31	Crosses	1224 non-null	object
32	Dribbles	1224 non-null	object
33	Possession lost	1224 non-null	int64
34	Duels won	1224 non-null	int64
35	Aerials won	1224 non-null	int64
36	Tackles	1224 non-null	int64
37	Interceptions	1224 non-null	int64
38	Clearances	1224 non-null	int64
39	Red cards	358 non-null	float64

*Note: The feature "Unnamed:0" is related with the extraction process required from the JSON file to distinct between home team and away team in each match.

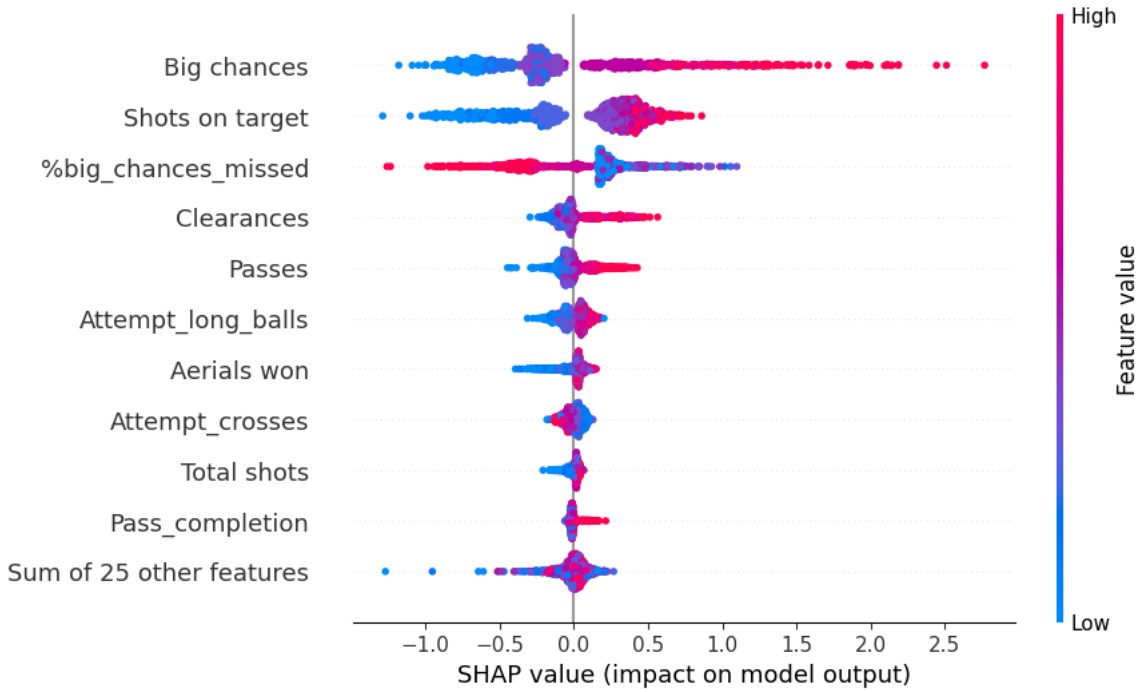
APPENDIX B – HEATMAP OF MISSING VALUES



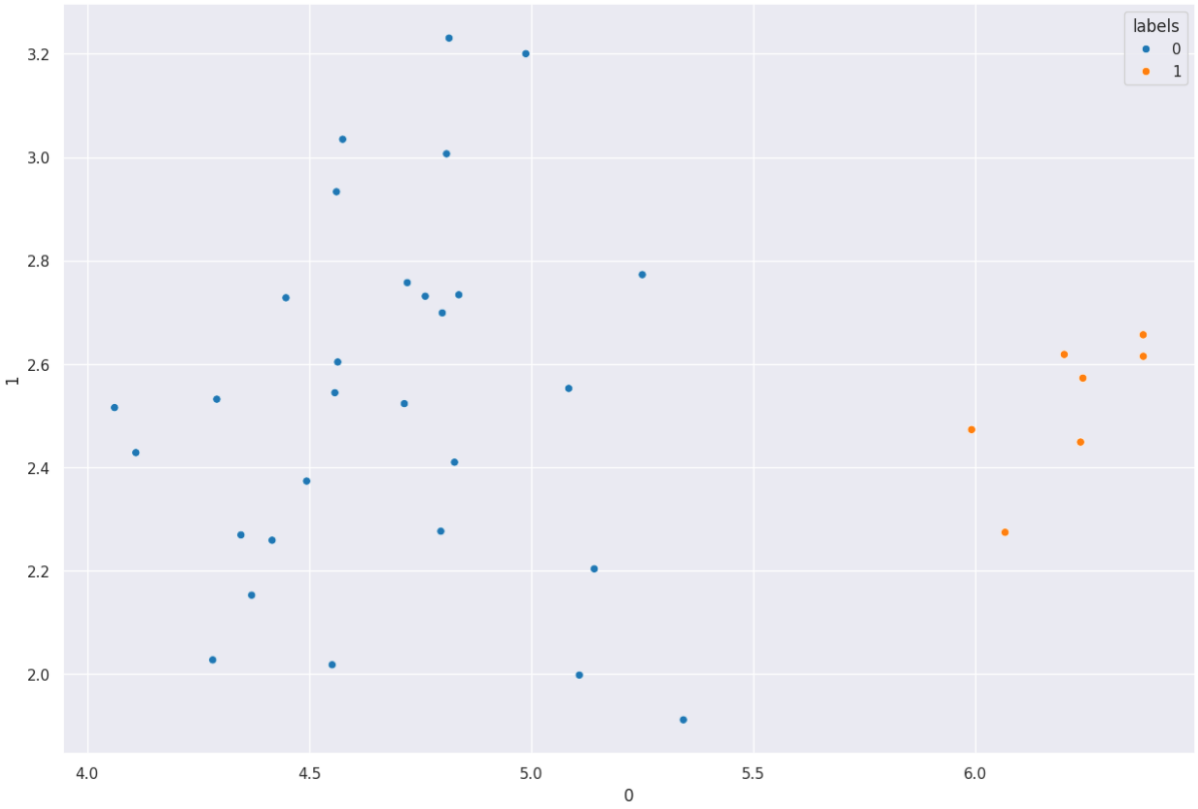
APPENDIX C – PEARSON'S CORRELATION MATRIX



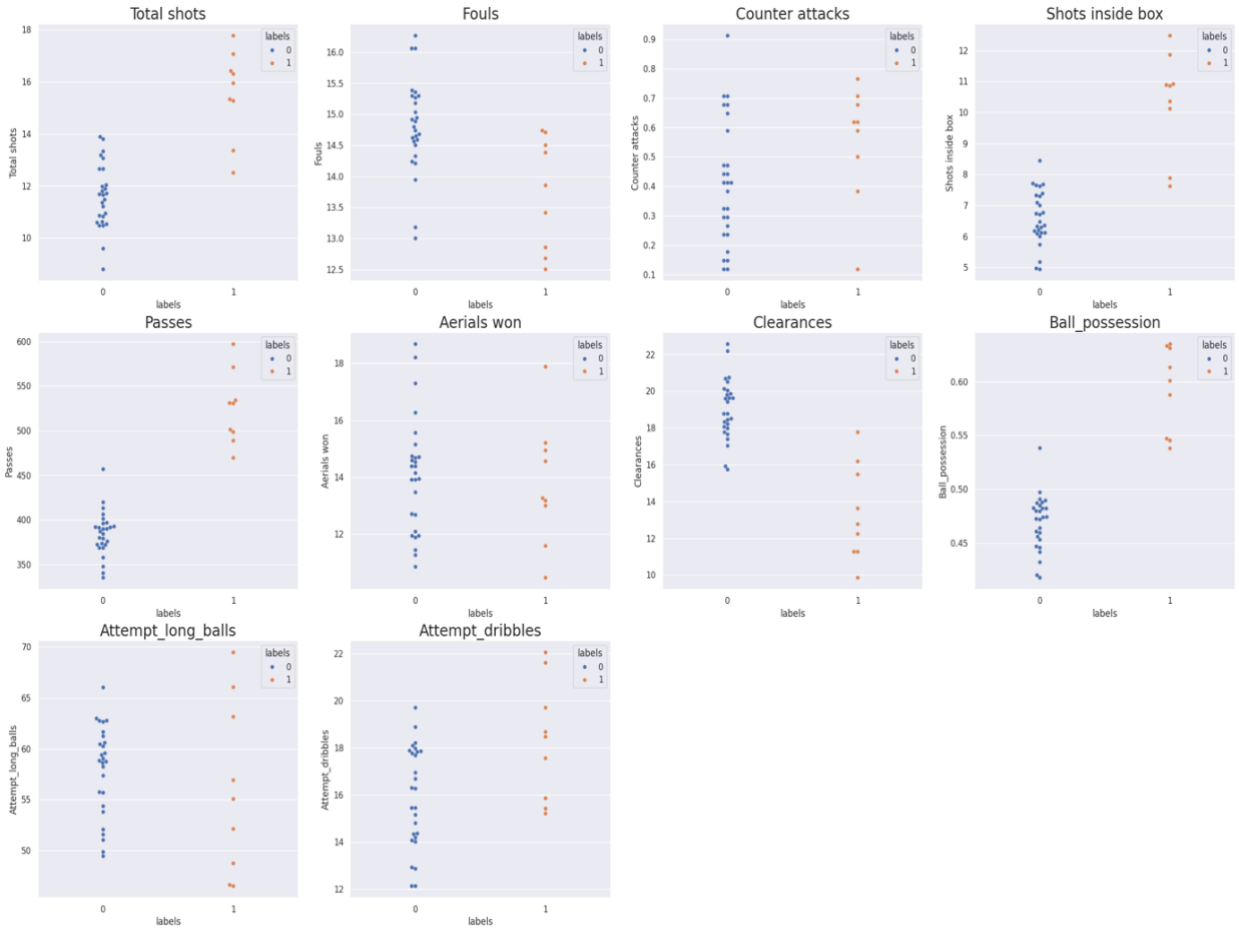
APPENDIX D – RANDOM FOREST'S SHAP SUMMARY DOT PLOT



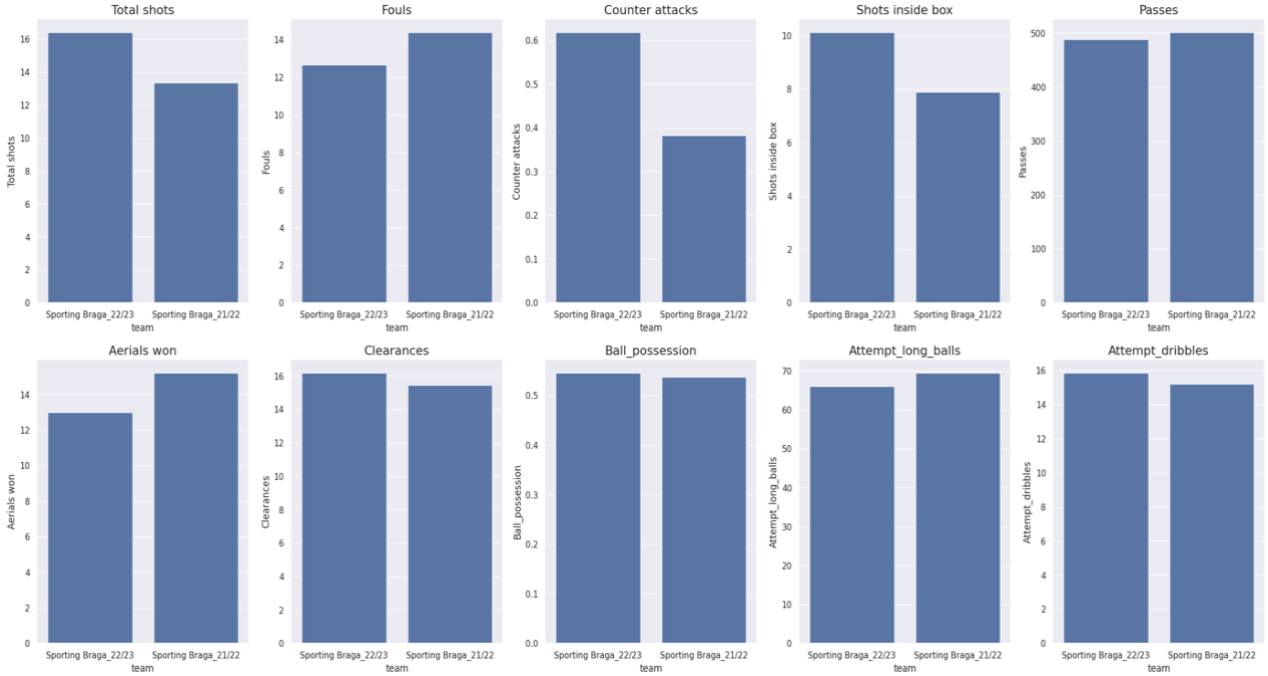
APPENDIX E – K-MEANS' TWO CLUSTER SOLUTION FOR APPROACH 1'S CLUSTER VISUALIZATION USING T-SNE



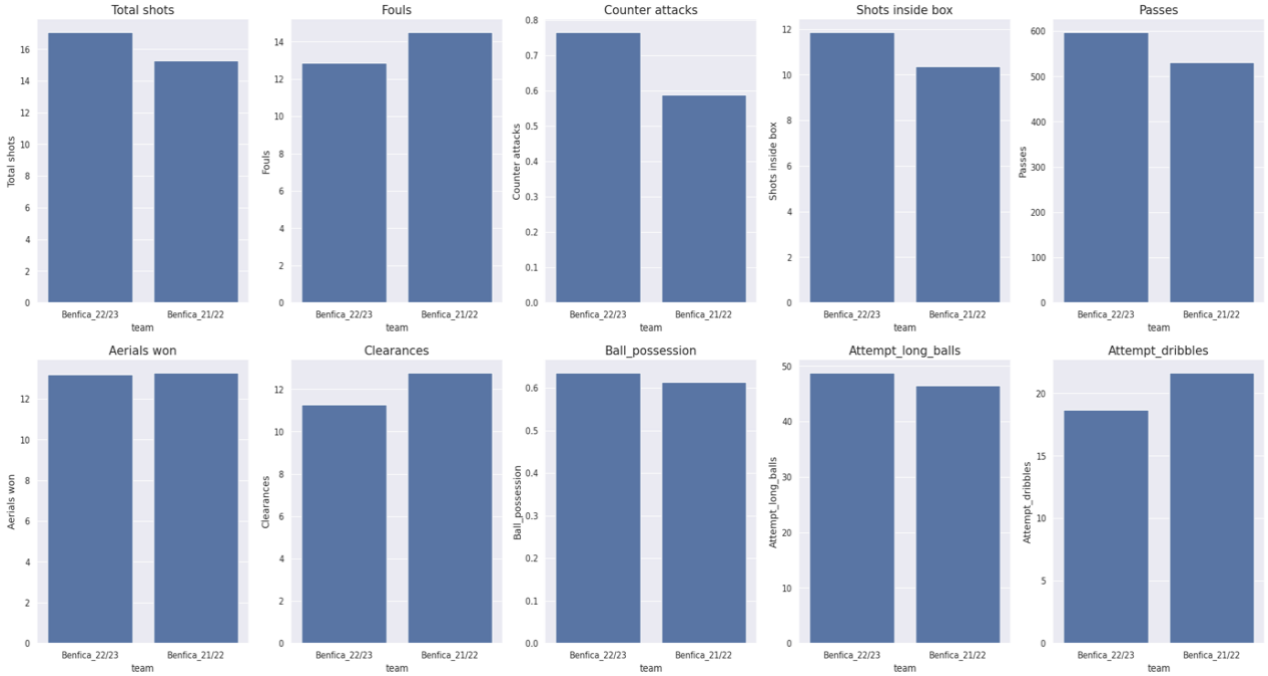
APPENDIX F – SWARM PLOTS FROM APPROACH 1'S HIERARCHICAL CLUSTERING SOLUTION WITH TWO CLUSTERS PRESENTING PLAYING STYLE FEATURES PER CLUSTER



APPENDIX G – BAR PLOTS COMPARING SC BRAGA'S TEAM PERFORMANCE IN BOTH SEASONS ON PLAYING STYLE FEATURES



APPENDIX H – BAR PLOTS COMPARING SL BENFICA'S TEAM PERFORMANCE IN BOTH SEASONS ON PLAYING STYLE FEATURES



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa