NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

# The Influence of Data Analysis on Football Teams to Increase Sports' Performance

Diogo Miguel Cavaco Pina Cabral

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

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

# THE INFLUENCE OF DATA ANALYSIS ON FOOTBALL TEAMS TO INCREASE SPORTS' PERFORMANCE

by

Diogo Miguel Cavaco Pina Cabral

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

**Advisor:** Vitor Manuel Pereira Duarte dos Santos, PhD

October 2022

# DEDICATION

Como não poderia deixar de ser, esta página terá de ser escrita na minha lingua materna, o português. Só assim os agradecimentos fazem sentido e chegam ao coração de quem têm de chegar.

Em especial a ti, Pai, pela crença e pela insistência no maior sucesso possível. Não se trata dos títulos e, muito menos, dos diplomas que me fizeste alcançar, mas sim do investimento que sempre quiseste fazer em mim.

A ti mãe, que amparaste todos os medos, meus e do pai, e nunca duvidaste do meu sucesso e das minhas capacidades.

A ti Tatiana, por nunca saíres, nem por um segundo do meu lado e seres a primeira a insistir e persistir para que concluísse este caminho, com o maior êxito possível, sem nunca pôr em causa o futuro da nossa relação.

A ti Sara e a ti André, por serem a minha irmã e cunhados mais velhos, a quem nunca faltou uma palavra amiga e um conselho, fruto da experiência, sempre com o simples objetivo de que o tivesse e tenha o maior sucesso possível.

A ti Laura, que entraste na nossa vida e a revolucionaste completamente e que, com um simples sorriso, sempre me mudaste o dia e, muitas vezes, foste a motivação que faltava.

À VTXRM, à minha equipa de projeto e, em especial, ao meu Team Leader, Miguel Figueira, que sempre foram e são apoio nos meus projetos académicos e acreditaram, também eles, no meu sucesso.

A si, Professor Vitor Santos, por garantir, desde o primeiro dia, que nunca me faltava orientação de qualidade e pela abertura ao tema, colocando sempre as minhas necessidades e vontades em primeiro lugar, em prol do meu sucesso.

E por fim, a dedicatória mais importante de todas. A ti, avó Maria. Como prometido, quer fosse uma pequena ou grande conquista, seria sempre por ti. Que onde quer que estejas, te tenha feito sorrir e te tenha orgulhado mais um bocadinho, porque esta foi mesmo das grandes!

A todos, muito obrigado!

Do vosso, Diogo.

# ABSTRACT

As we know, football is the most popular sport among the fans all over the world and, in today's world a very lucrative business for club owners and stakeholders, and sometimes its own supporters.

With the board and supporters' expectations being higher with the money spent on new players and conditions to attract valuable assets for the clubs, the teams tend to invest their money on infrastructures and other type of conditions for their players, including a better staff.

The teams' staff normally gather many data during the training sessions, other teams' observation, and post-match observations, meaning that the investment is now increasing on hiring new data analysts. Additionally, there are scouting teams that gather data as well. With that, the question that arises is how can football teams increase their performance, using data analysis?

The goal of this dissertation is to understand how the existing tools are helping teams improving their performance in and off the pitch and propose new ways on how future analysis can be conducted. To meet this goal, an extended systematic literature review will be taken, to present a discussion and conclusions on how data analysis can influence football clubs and players' performance.

# KEYWORDS

Data Analysis, Business Intelligence, Big Data, Data Science, Data Mining, Decision Support

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

In this section a background about football will be done, supporting the problem identification, giving a first draft of the first research questions that should be taken into account. It will also be presented the motivations of the study, as well as the objectives.

## 1.1. BACKGROUND AND PROBLEM IDENTIFICATION

In today's world, as we know, the sport that has the most supporters and viewers in the whole world is football, or soccer, as it is known in North America. With that, the investment in clubs has naturally grown, increasing development of many private soccer clubs all over the world (Baboota & Kaur, 2019) and consequently, the supporters, club boards, and owners started to put more pressure on teams to have better performance and exceed initial expectations. From a business perspective, creating new clubs with young players, loaning players from other clubs, picking specific positions, determining wages to players based on performance and international rankings is a complicated decision process (Rajesh et al., 2020).

The basis of football and what normally happens is that the better performance a team has, better the rewards the team gets. These rewards are either directly related with sports, gaining the right of playing in the best competitions with the best team of their continent, and potentially, the world, with teams of the same level, or other over-achieving teams, or related with higher monetary prizes.

With the objective of doing better and achieving board expectations, the money gained by clubs, that can be earned on the yearly competitions and on extra revenues (from merchandising, ticket sales, player sales, personal owner capital injections, bank loans, among others), is reinvested not just in good team managers or coaches or new players, but also in excellent infrastructures for first team, and youth teams, and hiring other staff members, also related with data analysis.

Since players' prices are exponentially growing, having teams spending more than 100 million in the transfer markets repeatedly over the years, teams with lower budgets are essentially betting on their youth systems to promote players to their first team and consequently increase their market value. This tendency is specially increasing in Portugal, since only three teams in our country can, in a smaller scale, follow the European tendency.

The valorization of players is, in the current times, increasing not just on the pitch, but also off it. In that basis, clubs are focusing on guaranteeing that players are focused on results, being both physically and psychologically well recovered. With the money spent on these areas and staff to do the daily squad monitorization, the club is building on their future, since the player can have a valorization of millions on the future and, for a club with small yearly budget, the strategy should be this one (Rajesh et al., 2020). The football culture in Portugal is exactly this one, and there are many examples of Portuguese players found and developed by Portuguese clubs that were sold for millions after some years on the first team. These homegrown players, if well exploited, can not only guarantee financial sustainability, but also good sportive results.

All of this daily processes in football generates millions of data per day, and even more per season, since we can look to many statistical indicators, which means that a good staff is also needed to analyze the raw data and transform it into physical outcomes, since a highly specialized coaching

provision has a significant and positive impact (Valenti et al., 2020) on decision making. The data received can come from training, opponent teams' analysis, previous matches' analysis, transfer targets' analysis and the youth teams' games and training. All of these indicators and all in-game actions should be included to create statistics and come out with a descriptive result (Tianbiao & Andreas, 2016). A season or even a game should be prepared having in mind the data from previous years, but also current years. Using feature engineering and exploratory data, the clubs can combine the most important factors and predict the results of the next match (Tianbiao & Andreas, 2016).

Therefore, the clubs are in much need of tools that can support them to make the best decisions as possible around all these questions, where the raw data can be imputed creating a solid output that could inform clubs if they are underperforming, on track to accomplish their objectives, or having a better performance than expected. Additionally, the tools should be helpful to staff teams to adopt new ways of increasing performance.

In the last years, football is improving the data-driven approach in terms of the decision, so the motivation for this research resides on the fact that data analysis on football is increasingly starting to be a key part of football. Since there are specific events to talk about data analysis and football related themes, and teams' staff are hiring more and more data analysts, to different and specific areas, data analytics has permeated the sports world (Corscadden et al., 2018). Although, there is new clear way on how data analysis is effectively influencing the players' performance and clubs' results, even though there are proofs that data science leads to improve business profits through a systematic enhancement to football datasets (Rajesh et al., 2020), but again, no clear information about sportive performance. Sometimes statistics and accuracy is often disparaged as a threat to football's romanticism (Spector, 2016), making this a much-criticized theme among the lovers of the sport.

The clubs do not explain clearly how they use data and what are the specific functions of their data related staff members. This raises many questions, such as:

- Does data analysis only affect training?
- What influence can it have on match preparation and during the match?
- What is the influence on the players, mentally? Are their decisions influenced on data inputs given by the coaching staff?
- Does data have any influence on injury prevention, or even on recovery itself?
- Does data influences decisions about young players promotion to the first team?
- Are players more likely to be signed for other clubs if they have a better performance, from the statistical point of view?

Some authors already tried to explain and study this subject, talking about some clubs as given examples, but as far as we know, never a Portuguese club was referenced, or anyone tried to understand how the evolution of data analysis is in football and performance.

## 1.2. MOTIVATION AND JUSTIFICATION

The study relevance of this dissertation, from the sports community point of view, resides on the fact that supporters can have a better overview on what is the day-to-day process of a team staff, that is

sometimes considered unnecessary. Additionally, to get to know the reasons that are behind the promotion of an academy player or the signing of certain, and sometimes, unknown player.

For the teams themselves, the contribution could be wider, in the sense that, according to the budget they have, the teams should adapt their analysis to their needs and resources. This will guarantee and improve their sustainability and sportive evolution throughout the years, showing audacity and no fear to risk on new methods to achieve results.

Scientifically, this research can help the data analysis world to be more associated with another successful application of data analysis methods, that is used in a numerous field of study. Additionally, this study can overcome many barriers in many other sports that can also, in the future, depend and see in data analysis a safe guarantee of performance improvement, having in mind that an exclusively quantitative approach does not allow definitive conclusions to be drawn on the complex relationship between elite sport policies and analysis outcomes (Valenti et al., 2020). For example, *Moneyball* changed baseball and almost all ball games (Kuper, 2013), consisting of a great case study on how crunching numbers can give you an edge.

### 1.3. Study Objectives

Since good decisions are critical for business success, and the world of football is increasingly becoming more and more a business, the goal of this master's degree dissertation would be to find out how can data analysis influence football teams to have better sports performance.

After the analysis of the questions and background problems earlier identified, it was decided that this research would just focus on in-game performance, match preparation and opponent team analysis.

To achieve this goal, the following intermediate objectives were defined:

- Understand how technologies can help staff gather raw data to be analysed.
- Understand how specific software helps staff teams to understand:
    a. Who are the players that are performing better on the team?
    b. Who are the players that are performing better on the opposing team?
- Perform and extensive and systematic literature review to understand what tools, how they are used, and their results.

## 2. METHODOLOGY

To achieve the goal of study a methodological path divided in three different phases will be followed: exploration phase, analytical phase, and conclusive phase. Each phase is divided in specific steps as identified in Figure 1 that will be able to reach the proposed specific goals.

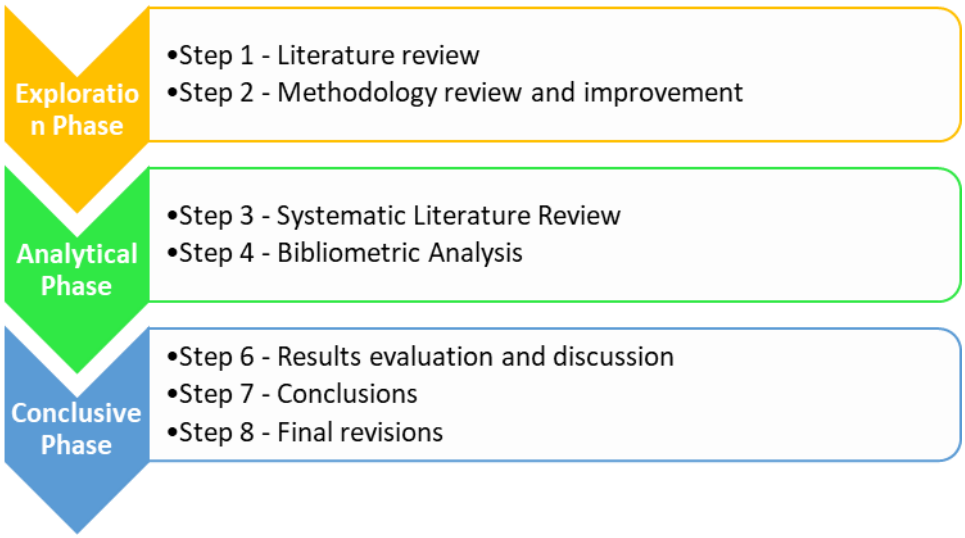


Figure 1 – Methodology Proposal.

In exploration phase, the literature review will be carried out, identifying the theoretical basis for this study, starting with football theoretical concepts, historical contexts and rules. After this introduction, a theoretical basis will also be given to data analysis concepts and its relation to sports.

In analytical phase, a systematic literature review will be conducted, identifying the important papers for this research. Then, the bibliometric analysis, consisting on the identification of the most used terms on papers. The papers will be divided into categories according to themes discussed on the manuscripts analysed. Lastly, the papers will be individually analysed to understand the outcome of the researches that were taken.

In the conclusive phase, the conclusions about previous works will be presented. In the end of this study, it will be identified the principal conclusions obtained, the identified future study recommendations.

# 3. LITERATURE REVIEW

In this chapter, the theoretical basis for the development of the master thesis will be presented.

The objective of the literature review is to give, firstly, an introduction to the major concepts of this research: football and data analysis. With that purpose, the literature review was organized in three parts.

In the first part, the football as sport is explained, alongside with its basic concepts and rules. More technical aspects of the game were also explained (tactics, performance, scouting, challenges and opportunities of the sport, and technologies used on the sport).

In the second part, the research focus on the concepts regarding data analysis.

The third and final part, show the relation of football and data analysis through technologies, algorithms, and tools, using a systematic literature review.

## 3.1. FOOTBALL

Football, or by the etymology of the noun "the action of kicking a ball", has its origins, as we know it today, on the 19th century, in England. The oldest known and official professional football blub is Notts County, formed on the year of 1982 and currently playing on the fifth tier of English Football.

The first competitions were born on the industrialization phase of history and led to larger groups of people meeting at places such as factories, pubs, and churches. In 1885, professional clubs were legalized and, three years later the English Football League, was created.

Although, the first major competition was the Football Association Challenge, as known as FA Cup, which still exists today. Initially, the players were only playing as a hobby, where the greatest part of the players, and specially the competitions' winners were the richest part of the population. The rules, back then, explicitly said that no player could have any type of remuneration from playing football.

In Portugal, the first match was played in the *Madeira* Island, in the end of the 10th century, induced by a British student that was living there. The first football club was only founded in 1893, by the name of *Futebol Clube do Porto,* and still runs on the topflight of Portuguese football. Although, some historians, said that the first club was founded on 1898, and that *F.C. Porto* was created lately, by the name of *Sport Clube Vianense*.

In the current days, and as talked on the background of this research, football is the sport with the most audience in the world. The financial investment is exponentially growing as well, to historical numbers. In fact, football has had (Doidge et al., 2019) a significant economical transformation. It attracts, from a positive side, much academic attraction, but also extreme groups that do not contribute to the popularity of the sport.

### 3.1.1. Concepts and Rules

There are some concepts and rules that should be presented before further explanations, to better understand the more technical details of this research.

Starting by the rules, the objective of football is simple: to score a goal in the opponents' net. There are three possible results on this sport: the victory, which goes to the team that scored most goals, the draw, when both teams score the same number of goals, and lastly, the loss, to the team that, scored the least goals. The main rules are:

- The field should have between 90 meters and 120 meters of length and between 45 and 90 meters of width, with the define markings for the boxes, halfway line, corner areas and penalty spots.
- Eleven players (ten field players and one goalkeeper) and a certain number of substitutes and substitutions allowed (depending on the rules of each competition), compose a team.
- The game must have at least three referees (a principal referee, two linesmen, and optionally a 4$^{th}$ official). Additionally, depending on the competition, there could be two more referees on the Video Assisting Referee (VAR) technology.
- The game has a minimum of 90 minutes plus added time. In certain competitions, if the final score, on the 90 minutes, is a draw, the game can have up to 120 minutes and, if in the end of that time, the score is still a draw, penalties are required to break a tie.

To conceptualize football, it is considered (FIFA, 2019) that football is part of our communities, meaning that every nationality, creed, ethnicity, education, gender, or religion is equally accepted. The growth of the sport will be guaranteed by these basic values, improving its quality and integrity.

Football is a complex game with various number of actions per training and game, and that it is (Constantinou & Fenton, 2017) the most popular sport in the world, leveraging the inspiration of many researchers to use football activities in their work.

### 3.1.2. Tactics

According to (Beal et al., 2020), there are multiple decisions that the manager must make before and during a football match. The success of these tactical wise decisions is measured by its positive or negative outcomes.

Teams normally adapt their tactics according to their probability to win, using, for instance, more reserved tactics to pick up points, in case they are the team with less chances of winning (Beal et al., 2020).

Therefore, the matches are prepared previously, and managers prepare, during training session for the upcoming matches, focusing on their team style, defining the team formation, and selecting the first team players. The tactics can have specific characteristics, like the mentality (offensive, defensive, counterattack, among others), the tempo, and the passing style, among other.

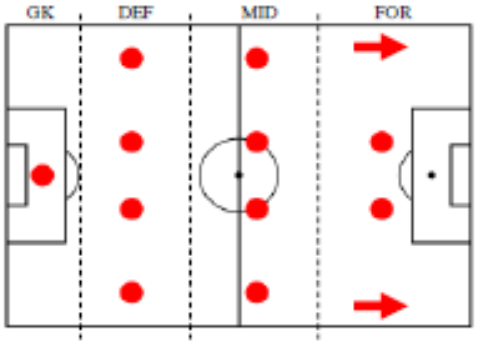


Figure 2 – Example of Team Formation (4-4-2) (Beal et al., 2020).

The managers also make many in-game decisions, where they can change their play style, team formation, and even change the players, according to the competition rules, as explained above. The objective of these decisions is to improve team's chances of winning the game.

Some of the key pre-game decisions that are made by both teams include (Beal et al., 2020):

- **Team Style:** relates to the teams overall use of different playing methods. In table 1, we can see some examples of playing styles.
- **Team Formation:** how players are organized on the pitch. In figure 2, we can see a formation example.
- **Selected Players:** 11 players thar are selected to play. Some players may play better in certain styles and formations.

Table 1 – Examples of Playing Styles (Beal et al., 2020).

| Style | Description |
|---|---|
| Tiki-Taka | Attacking play with short passes |
| Route One | Defensive play with long passes |
| High Pressure | Attack by pressuring the opposition |
| Park the Bus | A contained defensive style |

### 3.1.3. Performance

From business perspective, and defining the word performance, it can be seen as the ability of a specific company to implement an optimal organization, with the objective of meeting expectations of buyers or consumers. From a sports perspective, it is not very different, if we look at the consumers as the board or the team's supporters. This means that performance can be measured by the proposed objectives that were accomplished. Although, in sports that is not linear, since we can consider that a team had a high performance, but did not have the objective desired, since the team lost the game, for example.

In football, such as in many other sports and fields of study, the performance is a good key indicator to the success of business. Specifically for football, performance analysis is an integral part of the coaching process (Mackenzie & Cushion, 2013). To asset and evaluate football performance the most important variables are the physical conditions alongside with the technical and tactical performance (Rösch et al., 2000).

In this sport, measuring and manage performance is a complex process that involves optimizing many key factors for football players, for instance, physical performance, skill-based training, tactical training, minimizing risk of injuries, and providing psychological support (De Silva et al., 2018).

In (Rice, 2014) article, it is told, for example, that Brentford, a club from the English Premier League (by the time the article is written, the team was still on Championship, that corresponds to the English's Second Division), the key performance metric should be expected goals, that explains the quantity and the quality of the attacking team plays.

### 3.1.4. Challenges and Opportunities

Like every other study in the sports, there are challenges and opportunities associated with the study of players and teams' performance is no exception. The biggest opportunity found is that performance analysis is now widely accepted among coaches, athletes, and sports scientists as a valuable input into the feedback process (Mackenzie & Cushion, 2013). On the other hand, the biggest challenge is how to successfully predict the future performance, despite the inherent problems associated with investigating a multifaceted and often uncontrollable phenomenon.

According to reports from (Kuper, 2013), some English Premier League managers, even the most sceptic ones, started to understand the importance that data can have in football. The examples given are Roberto Mancini, while he was at Manchester City, and were the league winners, and David Moyes, currently at West Ham United. The questions that were put by the managers were about the opposing team's performance, and general statistics to see how they play. This consists into a very important opportunity to data analysis in sports since high profile managers use these techniques in their managerial staff.

On the other hand, states that football has alternative perspectives that influence team results and should be more explored, and sometimes being left out by data analysis, such as player recovery from fatigue and psychological mind-sets.

## 3.2. DATA ANALYSIS

In this chapter, the theoretical insights about data analysis will be given having in mind the techniques used on the papers and articles taken into consideration for this research. The data analysis concept is very complex, since there are many ways of reaching an objective with the dataset we are treating. According to (Haneem et al., 2017) the most researched topics on data analysis were: master data, data quality, business intelligence, business process, data integration, big data, data governance, information governance, data management, and product data.

### 3.2.1. Concepts

After the papers' analysis, the theoretical basis needed to better understand the algorithms, tools and technologies that will be presented later this research are:

- **Data Analysis:** systematic application of statistics and logical techniques to describe, treat and evaluate data.
- **Data Mining:** process used to take information from large sets of data, deriving patterns and trends existing in data, to define a data-mining model.
- **Social Network Analysis:** tools used to analyze relationships between people in groups. Its objective is to understand structures and interdependencies.
- **Data Analytics:** process to do an examination to datasets, with the objective of gathering conclusions about the information they contain.
- **Machine Learning:** subfield of artificial intelligence, defined has the capability a machine has to do the same as intelligent human behavior.
- **Feature Engineering:** step to transform raw data into features to be used in algorithms, such as predictive models.
- **Predictive Models:** these models are the outcome of the creation, during the feature-engineering step, of predictor variables selected for this predictive model.
- **Data Engineering:** step that allows data to be more useful and accessible, doing transformation to the raw variables.
- **Business Intelligence:** an infrastructure, of type procedural and technical, which collects, stores, and analyzes the produced data.
- **Big Data:** bigger and more complex datasets, which cannot be processed by traditional software and machines, due to its lack of capability to manage data.
- **Data Science:** interdisciplinary field of study, using data from various research to derive the meaning of the data.
- **Statistics**: branch of mathematics, that collects, analyzes, interprets, and presents numerical data.
- **Decision Support Systems:** software used to support courses of action, judgements, and determination in an organization or business.

### 3.2.2. Football Analytics – Concept

The study of (Babbar, 2019) refers sports analytics is a combination of data collections, forecasting the game and techniques to interpret the game strategy to improve a player's performance individually and for the team. Additionally, football can be used a support method to coaches' decisions.

With the use of the right algorithms and technologies, a huge number of aspects of the game can be captured, starting with the ball movement, and finishing on the players' heart rate and positioning. These tools allied with algorithms and software can be enablers so that coaches can alter their tactics, from game to game and during the game, to gain advantage to his direct competitor.

As we can see on figure 3, from (Morgulev et al., 2018), the information is gathered, with data from multiple resources, quantitative ou qualitative, then, by standardization, centralization and integration methods, data management occurs, and lastly, it is analyzed using metrics, providing outcomes that are presented to the decision-makers.

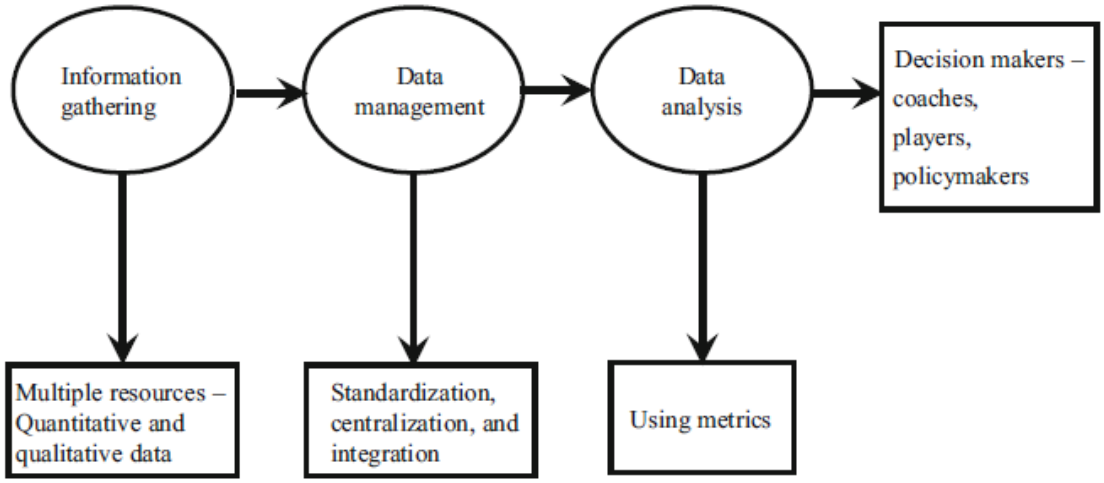


Figure 3 – A sports analytics framework (Morgulev et al., 2018).

According to study of (Morgulev et al., 2018), in football, in comparison with basketball, the scoring events are less and less frequent, so other moments of the game becomes more relevant to evaluate performance, for example, possession time, passing sequences, territory played spatial analysis, among others.

## 3.3. Systematic Literature Review

### 3.3.1. Materials and Methods

The systematic literature survey has the objective of providing an evaluation to the scientific papers, articles, and communities' contributions to the topic of research, in this case, the influence of data analysis in football teams' performance, by using a rigorous and auditable methodology based on the PRISMA approach.

The PRISMA method is composed by five phases, presented below, as follows:

1. Identification of relevant manuscripts of the domain.
2. Screening of titles, abstracts, papers without experiments, and position papers.
3. Eligibility analysis.
4. Full-text screening exclusion.
5. Final papers to be analyzed in detail.

To a more complete analysis on the quality of the papers, a bibliometric map – used to find relationships between terms, following three phases, evaluating the following quantities:

1. Words frequency.

2. Most common words.
3. Frequency of these common words in the final manuscripts of the study.

By following this method, this section will be organized by five steps: (1) our research questions, (2) followed paper search strategy, (3) bibliometric map, (4) inclusion and exclusion criteria, and (5) final paper selection.

### 3.3.2. Research Questions

As stated before, the main objective of this research is to find out how can data analysis influence football teams to have better sports performance. With the purpose of accomplish this objective, the following research questions were created:

RQ1: How can technologies help staff to gather raw data?

RQ2: How can a specific software help the staff to take conclusions about the gathered data?

RQ3: How can the treated data be used to improved teams' performance?

### 3.3.3. Search Strategy

A literature survey normally recommends searching various databases, to determine if similar work has already been performed, aiding in locating potentially relevant studies. In this study, the searches were done on the following repositories: (1) Scopus and (2) Web of Science.

The search query was then created, using alternative keywords, logically connected by 'OR' or 'AND' statements. The resulting search query utilized in the mentioned electronic repositories is depicted in Figure 4.

("football" OR "soccer") AND ("data mining" OR "decision support system" OR "business analytics" OR "data analysis" OR "machine learning" OR "business intelligence" OR "artificial intelligence" OR "predictive") AND ("performance")

Figure 4 – Search Query.

Figure 5 depicts the PRISMA flow chart, illustrating the five phases when filtering the manuscript set. The variable n corresponds to the number of papers at the end of each step.

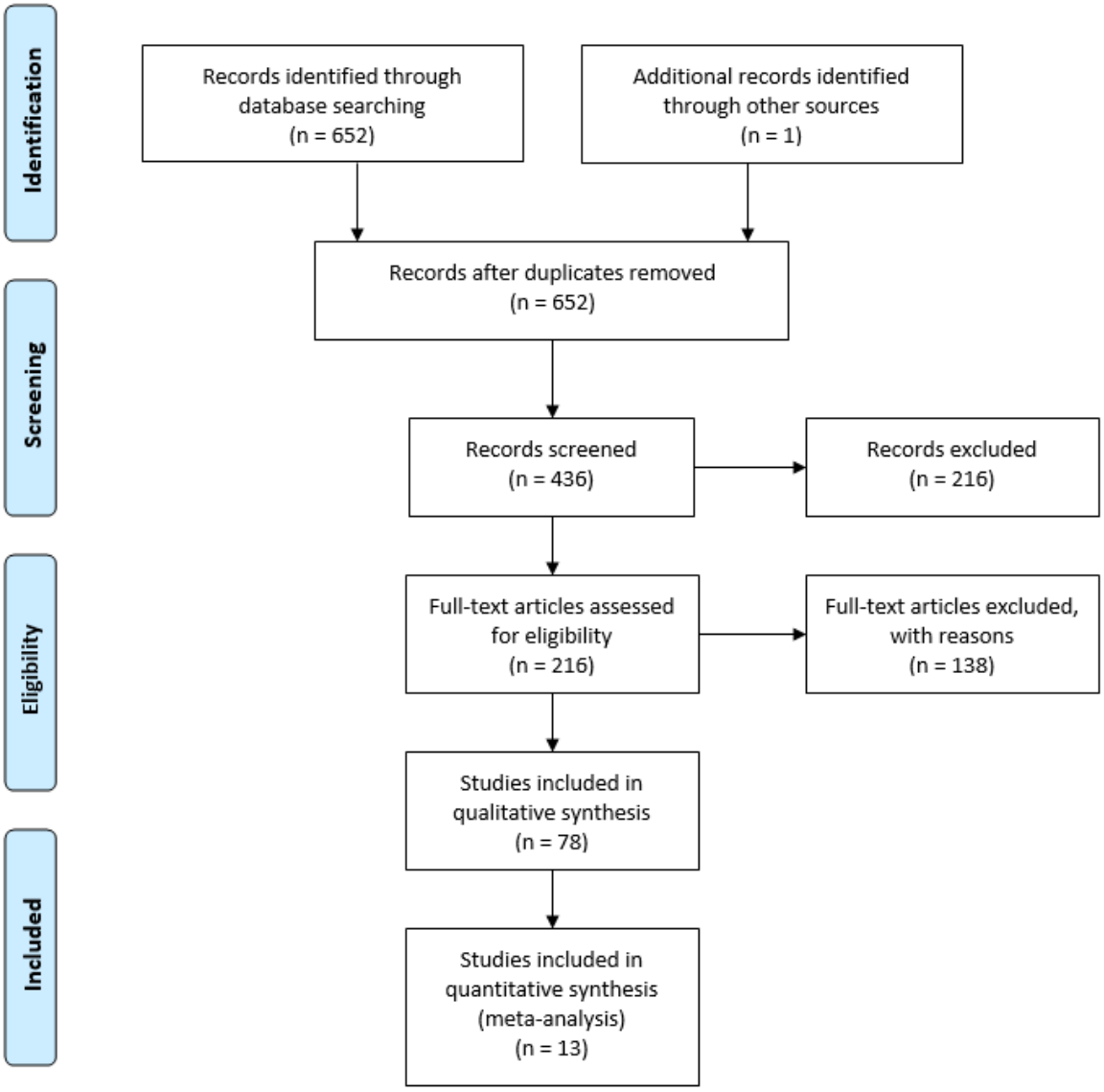


Figure 5 – PRISMA Flow Chart to analyse manuscripts.

In **Phase 1**, the search string was applied to all electronic repositories, for papers published between 2012 to 2021, which resulted in 653 publications.

In **Phase 2**, a 4-step approach was followed. In the first one, the manuscripts were excluded according to its titles, which narrowed the search to 436 papers. In the second one, the manuscripts were excluded based on the abstract, narrowing the search to 106 papers. Then, papers that present no experiment were excluded, and lastly position manuscripts were excluded. In the end of this phase, the number of research remaining was 78.

In **Phase 3**, manuscripts underwent a full-text reading and review, leading to 65 exclusions (the result of **Phase 4**).

### 3.3.4. Search Outcome

As result of the paper selection approach, the final list included 12 papers (**Phase 5**). These papers were then divided into 2 categories, as shown in the tables 2 and 3, presented below:

1. Post-Game Analysis
2. Pre-Game Prediction

Table 2 – Summary of papers in the category of Post-Game Analysis.

| Paper | Reference | Application | Data Dimensions | Method and Techniques | No. of Citations |
|---|---|---|---|---|---|
| S1 | (Tianbiao & Andreas, 2016) | Analysis of passing sequences on football games | Technical and Tactical Actions | Apriori Algorithm | 8 |
| S2 | (de Silva et al., 2018) | Player tracking for physical performance | Physical data about the game | Analysis of Variance | 16 |
| S3 | (Borges et al., 2019) | Tactical efficacy and offensive game processes | Offensive sequences | Correlation Network Analysis, Chi-Squared Test | 4 |
| S4 | (Pappalardo et al., 2018) | Performance Evaluation | Individual Performance in game | PlayeRank Framework, Clustering | 45 |
| S5 | (Li et al., 2020) | Data-Driven team ranking and match performance analysis | Match data | Team-Rank Framework, LSVC Model | 7 |
| S6 | (Zambom-Ferraresi et al., 2018) | Performance analysis | In-game statistics for Big-5 Leagues | Bayesian Model and Robustness Check | 2 |
| S7 | (Wu et al., 2020) | Big Data analysis of football matches | Player positions and passing processes | Social Networks, Clustering, Local Efficiency | 6 |
| S8 | (Beal et al., 2020) | Optimize football game tactics | Two teams' actions | Pre-Match Bayesian Game, In-Match Stochastic Game | 3 |

Table 3 - Summary of papers in the category of Pre-Game Prediction.

| Paper | Reference | Application | Data Dimensions | Method and Techniques | No. of Citations |
|---|---|---|---|---|---|
| S9 | (Gomes et al., 2015) | Decision Support System for Predicting Football Game Results | Statistical data of football games | CRISP-DM Methodology | 2 |
| S10 | (Baboota & Kaur, 2019) | Predict future matches results | Previous games' statistics | Predictive Models | 68 |
| S11 | (Arndt & Brefeld, 2016) | Predict future performance of football players | Performance of previous games | Multitask Regression | 9 |
| S12 | (Rajesh et al., 2020) | Data science to predict first team selection | Player position, nationality, overall rating, overall performance | Naïve Bayes Model, Random Forest Model, Decision Tree Model, SVC Model, Proposed Prediction of Team Players | 5 |

# 4. RESULTS, ANALYSIS, AND DISCUSSION

In this section, a visual analysis will be done to add information to the search outcome. Additionally, the topics that arose from the analysis will be analyzed in detail.

## 4.1. VISUALIZING ANALYSIS

This section has two main parts that corresponds to the bibliometric analysis and the analyses of the previous work, in detail. The first part shows the relationship between the common terms in intelligence, statistical techniques, and performance metrics, used on the selected papers. The second part has the objective of analyzing the existing algorithms, software and tools and answer the research questions.

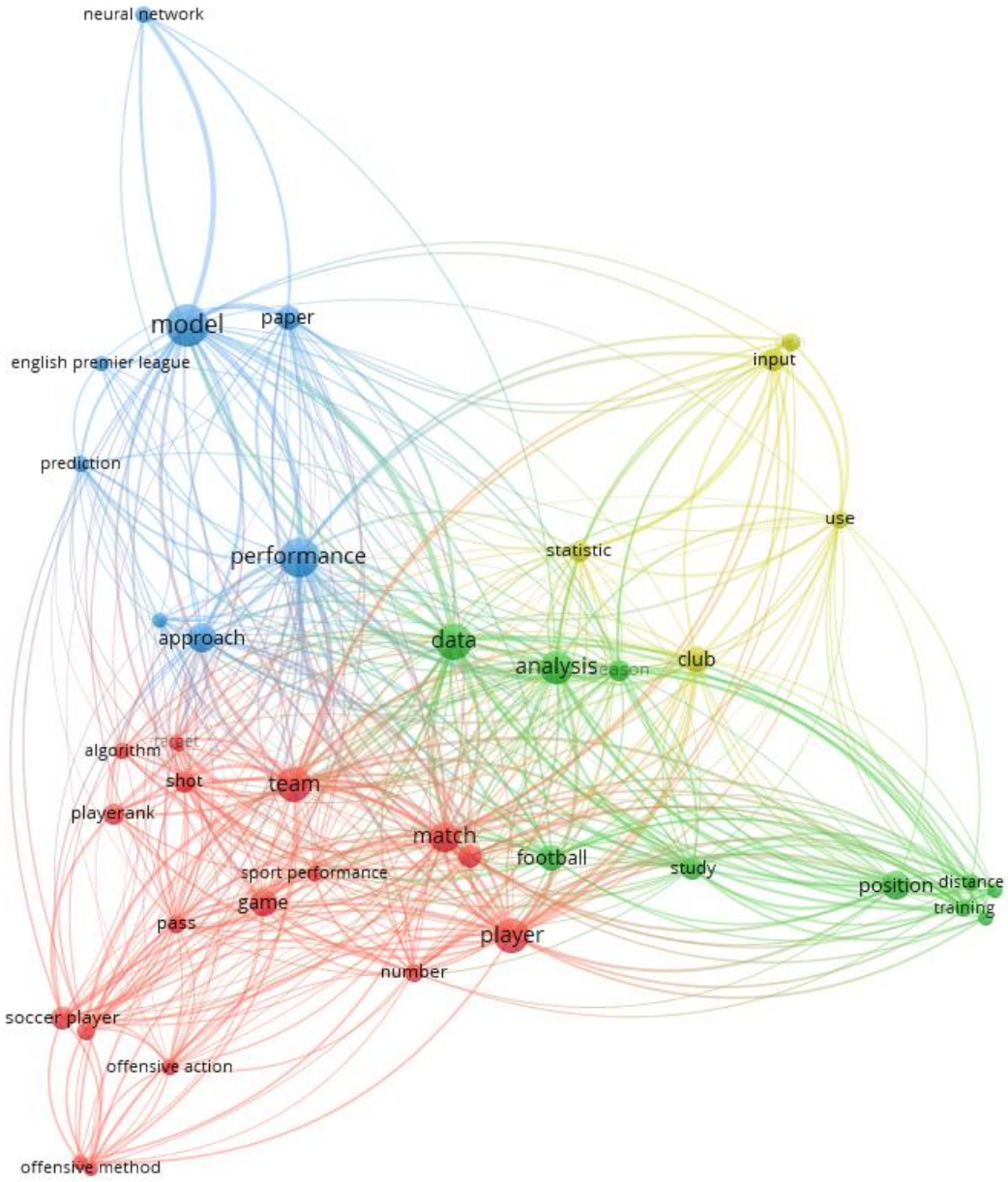


Figure 6 – The relationships between the common terms using bibliometric map.

The analysis that concluded on the visualisation available on Figure 6 was obtained only by searching on the title and the abstract, using a binary counting method, with a minimum threshold of three occurrences.

The largest nodes, meaning they are the most important nodes for each cluster in the network map are determined as “performance” and “model” (blue), “team”, “match”, “game” and “player” (red), “analysis” and “data” (green), “input” and “statistic” (yellow). Looking closer, we can conclude that all clusters are connected by the words “performance”, “data”, “football”, “analysis” and “statistic”.

## 4.2. Analysis Per Topic

### 4.2.1. Post-Game Analysis

(Tianbiao & Andreas, 2016) observed the 2012 Algarve Cup final match with structured observation methods. The method used consists on dividing the pitch into 30 zones, as we can see on Figure 7 below.

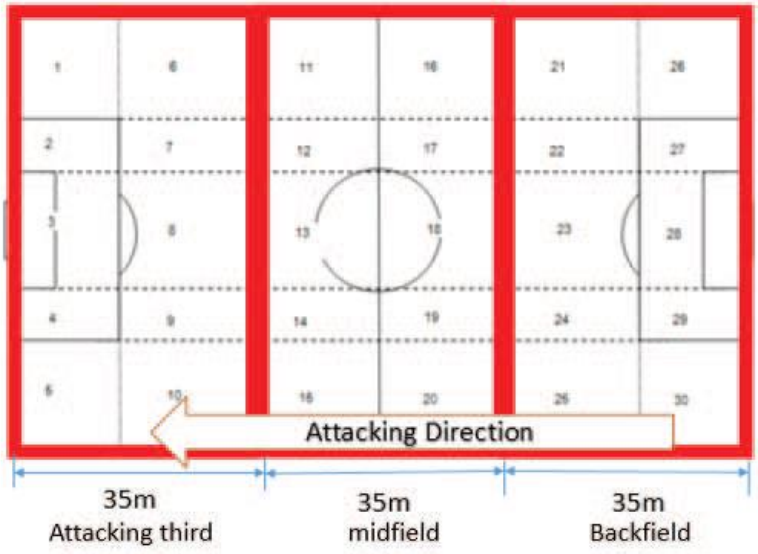


Figure 7 – 30 Zones divided football fields (Tianbiao & Andreas, 2016).

The model consists of defining, recording, cleaning, and processing data. All the aspects of the game were recorded according to Table 4. The data cleansing consisted on keeping only the last five elements of the controlling chain, support by the fact that (Hughes, 1990) most goals arise from last five transitions. For processing data, Microsoft SQL Data Mining Adding was applied the a-priori modified algorithm.

Table 4 – Data Definition (Tianbiao & Andreas, 2016).

| index | Start of Sequence | index | End of Sequence |
|---|---|---|---|
| 100 | kickoff | 111 | Goal |
| 101 | Free kick | 112 | Shoot but catched by goalkeeper |
| 102 | Throw in | 113 | Shoot bounced (by GK or goalframe) |
| 103 | Conner kick | 114 | Shoot out of goalframe |
| 104 | Goal kick | 116 | Be fouled |
| 105 | Attacking starts | 117 | be kicked out of touchline by opponent |
| 106 | Goalkeeper starts the ball | 118 | be kicked out of goalline by opponent |
| 107 | Referee whistles to restart the game | 119 | passing/shooting but intercepted |
| 122 | offside | 120 | out of touchline/goalline |
| 123 | Referee whistles to stop the game | 121 | dribbling, receiving but lost control |

The algorithm applied for this solution required two variables to be divided, so that the probability, or confidence could be calculated. This probability is calculated having in mind the "support", that, in data mining means that the pattern must have a certain frequency in the records to be taken into account (Tianbiao & Andreas, 2016).

For this study, the Association Rules were calculated, showing effective and general player combinations, both for first and second half. In Figure 8, we can see that the most frequent connections were between the numbers 4 and 14, and 7 and 13, this for the first half. In Figure 9, the same is represented, but for the second half of this game, this time with the most effective connection being between numbers 2 and 17, and 2 and 9. This connection is represented by a thicker black arrow that represent main game combinations that have very positive and strong influence to opportunity to score (Tianbiao & Andreas, 2016).

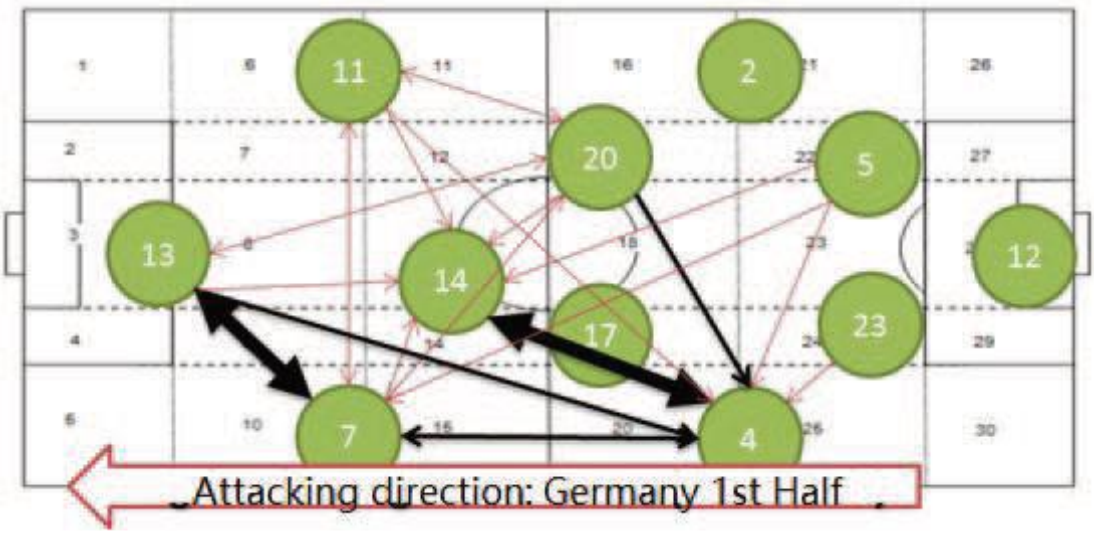


Figure 8 – Tendency Network 1st Half – Player Model (Tianbiao & Andreas, 2016).

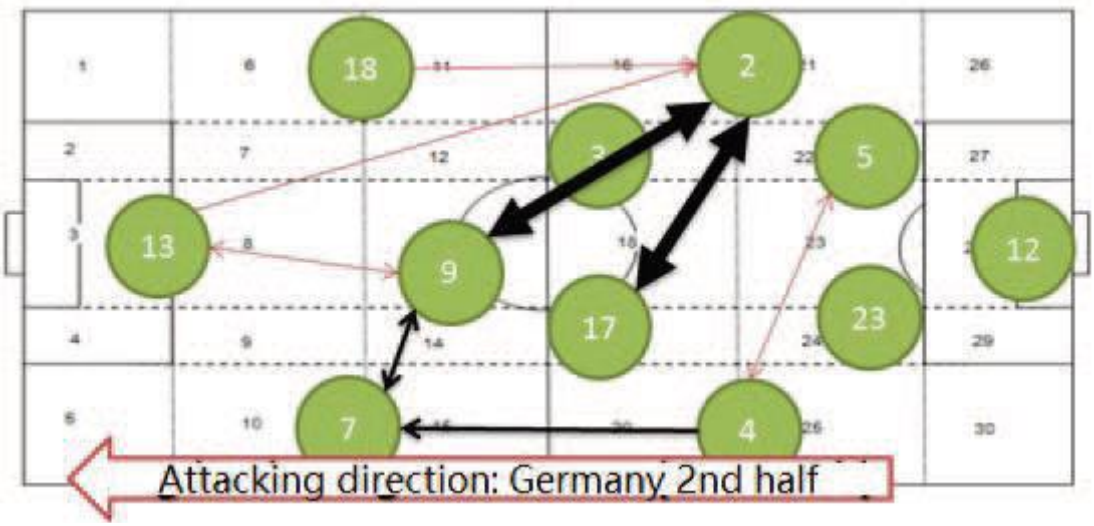


Figure 9 – Tendency Network 2nd Half – Player Model (Tianbiao & Andreas, 2016).

The results of this study and the application of a modified version of the Apriori algorithm allowed the analysts to understand that the team played better on the first half, with number 4 being the most influent player. On the second half, a decrease of quality was noted, although number 2 had the most positive impact.

To the authors of this research there is stillroom for improvement on the algorithm, so that the results could be closer to simulate the real game, optimizing that way, the match analysis.

For the next study, the used data source was collected and analyzed from 150 players, with ages between 18 and 23, across four different seasons, in the Chelsea Football Academy. The study's objective was to prove how player tracking data analytics could be a tool for Physical Performance Management (De Silva et al., 2018).

With the use of descriptive statistics, it was possible to compare the high-speed running activity demands between training and matches for different positions across the fields (fullbacks, center midfielders and center forwards). The results obtained stated that in training there are not major differences, while in game, more advanced positions require more high-speed runs.

(Costa, 2021) represents his weekly cycle, from an analysis and observation perspective. The three principal factors to prepare this cycle, should be:

- Team ideas and team's style of play.
- Our team's performance in the last few games.
- Opponent's observation.

The work of the analyst starts two days after the game. The week of work begins by analyzing the opponent, starting with a written report, by a personal analysis with all the information the analyst can get, not shared with the rest of the staff team. After the analysis of this report, the components are shown to the rest of the team. The report is (Costa, 2021) divided in two parts: qualitative and quantitative.

The complete structure of this report consists of:

- Quantitative analysis.
- Qualitative analysis.
- Defensive organization.
- Offensive organization.
- Defensive transition.
- Offensive transition.
- Set pieces.

**Quantitative Analysis**

The quantitative analysis, in the perspective of (Costa, 2021), is done with the objective of answering the questions: how the opponent scores goals? How the opponent suffers goals? What players are used? In which position?

The example given by (Costa, 2021) is the team of *Shakhtar Donetsk*, in the analysis made by the team headed by Luis Castro, the Portuguese coach, before the game against *Sport Lisboa e Benfica*, for the 2019/2020 season. The analysis was conducted taking into consideration all the games of the opponent team, including friendly games. The information gathered was:

- General data – pitch dimensions, titles, coach, and meteorological prevision. These data can be important conditions for the game.
- Squad list – the name of all players and its general characteristics (position, age, height, weight, preferred foot, and nationality).
- Tactical structure and first team used – used tactical systems, and in which competition and most used players in each tactical system.
- Used players, by minute – all minutes played by each player, and the position they played in, by game. This gives insights about the possible first team selection and the players that may come in during the game.
- Most used players – definition of the players that are most used in the first team selection.
- Goals scored analysis – areas of the field where the last pass is done, and the goal is scored. The moment when the goal is scored. The goal and assister. This maps and graphs the team's intentions and ways to get the goal can be easily predicted.
- Goals scored percentage, by game moment – the game moments considered are set-pieces, organized attack, and counterattack. This analysis can give important information on how the opponent build their goals.
- Assistance goals zones – this analysis can give the teams the information where the last pass, before the goal, is done, if the players prefer a more central zone of the field or work the ball from outside.

- Goal zones – where the goals are scored. This can help understanding if the goals come from working the ball into the box or from shots inside the box, for example. Additionally, is given the place where the penalties are scored and the best scorers of the team.
- Patterns encountered in the goals scored.
- The exact same analysis is conducted for suffered goals.
- Symmetric map of passes – this map shows the number of passes, and the most frequent combinations made during the game, between players. In this analysis, the number of failed passes is also taking into analysis.

**Qualitative Analysis**

This analysis is conducted by watching the opponents team opponent's observation. The examples of analysis used by (Costa, 2021) to work on this analysis are *Diego Simeone*'s *Atletico de Madrid* team, Paulo Fonseca's *Roma* team, *Jorge Jesus*'s *Benfica* team, Jurgen Klopp's *Liverpool*, and *Porto's Sérgio Conceição*.

The information gathered by this analysis is:

- First team and structural organization – this analysis is conducted with the objective of predicting the opponent's possible first team selection, and possible system variation during the whole game, considering the game context, and difficulty level.
- General collective principles – description of the principal game plan, with no variations considered.
- Sector principals and inter-sectors relation – the objective is to define if the whole team defends or some players have more liberty, if the called "space between lines", that its very important tactically, is given or not.
- Positioning – the positions occupied by the players in the first part of build-up of the opponent if it is more conservative or high pressure based.
- Opponent's strengths – this is a very important part for the analysis since it is important to understand the opponent's strengths to be easier to define the way of stopping them.
- Goalkeeper ball distribution – consists of the option the goalkeeper takes to start team's build up (long ball for the forwards, try to play with central defenders, for example).
- Defensive transition – this analysis is taken to define, when defending, which spaces should be occupied and more effectively taken, when starting the defensive

process, and the way the team should react when loses the ball (automatically press the player, give space to the opponents to start build up, for example).

- Offensive transition – the options taken by team the opponents when starting to attack (a more explosive option, through the pace of the attackers and long ball, a lower tempo and patient build-up or a higher tempo exploring spaces, for example).
- Set-pieces – this analysis will be important on the way the team defends and attacks in set-pieces (corners, free-kicks and even throw-ins).

After the elaboration of these two analyses, the synthetized report is sent to rest of the staff, passing only the information that is important for the coach to work with. Then the evaluation starts by being focused on the team we are working with, instead of the opponent. The first step is to analyse the training footage. The importance of this moment consists of:

- Visualize and register the posture of the staff.
- Visualize and register the posture of the players, specially in drills that needs more concentration.
- Visualize and register training exercises, and possible failures and successes.

The last steps to a successful week of work for the analyst is to give the outputs of the analysis, translated into decisions to the players and use that as strengths to use in-game. After the game itself, it is important to review all process and assess what needs to be changed in future weeks and games.

On other study with the objective (Borges et al., 2019) of analysing success and failure of offensive sequences, used in under 15 and under 17 soccer players. This is an observational study, focusing on static behavioural indicators. The sample included 218 offensive actions, selected from 28 matches.

The used software consisted of Match Vision Studio Premium®, enabling research to create a categorical matrix according to the variables to be analysed.

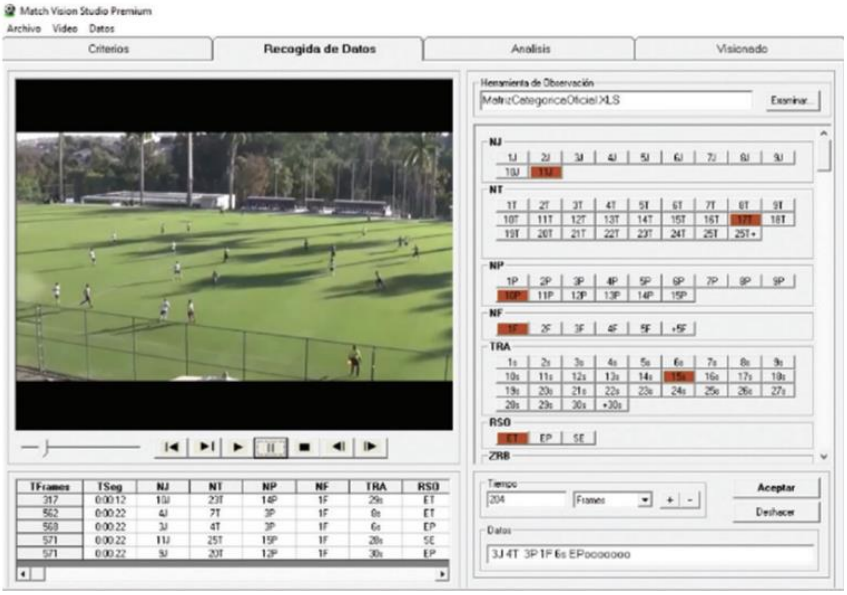


Figure 10 – Screen capture from Match Vision Studio Premium® (Costa, 2021).

The categories of offensive actions which ended in finalization were selected, reaching the variables used to create the categorical matrix, that are presented on Table 5.

Table 5 – Categories of offensive actions and results of the actions (Borges et al., 2019).

| Variables | Description |
|---|---|
| **Players Involved** | Number of players that were involved, touching the ball during the offensive action. |
| **Ball Touches** | Total number of ball touches performed by players during the offensive action. |
| **Passing** | Total number of passes made with any part of the body that was received by the attacking partner and continued the offensive phase of the team. |
| **Duration** | The duration of the offensive phase, from the interception of the ball, to the end of the offensive action (seconds). |
| **Corridor Changes*** | Number of times that the ball changed field corridors during the offensive action, taking into account the division of the field into 3 corridors (left, central, and right). |
| **Results of Offensive Actions** | **Description** |
| **Success** | Sequence finished in goal. |
| **Failure** | Sequence finished with kick out or goalkeeper's defense. |

In table 6, we can see the characterization of the offensive game methods and playing styles, during the offensive actions.

Table 6 - Characteristics of Offensive Game Methods (Borges et al., 2019).

| Counter attack (CA) | Quick attack (QA) | Positional attack (PA) |
|---|---|---|
| Ball recovered in any area of the playing field | Ball recovered in any area of the playing field | Ball recovered in any area of the playing field |
| Performs equal or less than 5 passes | Performs a maximum of 7 passes | Performs more than 7 passes |
| Offensive sequence duration equal to or less than 12 seconds | Offensive sequence duration equal to or less than 18 seconds | Offensive sequence duration exceeding 18 seconds |
| Opponent team advanced on the pitch and defensively unbalanced | Opponent team balanced defensively | Opponent team balanced defensively |
| Ball circulation more in depth than width | Ball circulation in width and depth | Ball circulation more in width than depth |
| High play intensity | High play intensity | Cadenced play intensity |

The (Borges et al., 2019)'s study results were that all offensive sequences ended in in shots according to the following variables: number of players involved, ball touches, passing, duration, and corridor change. With this context, the research suggests that all offensive methods adopted can be used to achieve success during a game of Under 15 and Under 17 football players.

According to (Pappalardo et al., 2018), the problem of evaluating the performance of soccer players is attracting the interest of many companies and the scientific community. All of this is because there is many ways and a massive amount of data available. Although there is a huge amount of websites and television broadcasters, such as Opta, WhoScored and Sky that use these data sets to compare teams and players' performance.

The study of (Pappalardo et al., 2018) presents the limitations of the existing approaches and developing PlayeRank, that is a new generation data-driven framework for the for the performance evaluation and the ranking of players in soccer.

For that, is important to understand, firstly, how the PlayeRank framework works. Figure 9 shows the schema of the framework, starting with a database of soccer-logs. The framework has three main phase: the learning phase, the rating and the ranking phases.

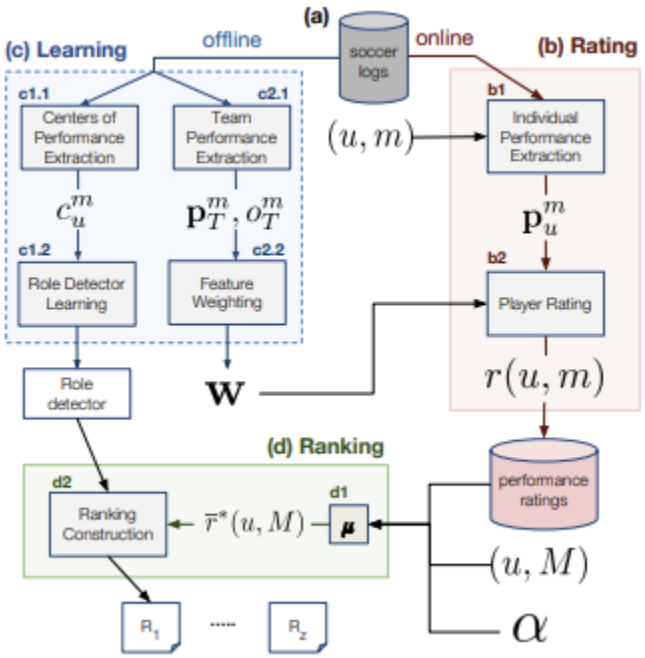


Figure 11 – Schema of PlayRank Framework (Pappalardo et al., 2018).

**Rating Phase:** The rating phase, that corresponds to step B on figure 11, corresponds to the computation of the performance rating, using each player u and every match m, that is available on the database. The phase consists of tow main steps (on Figure 11, shown as step B1 and step B2):

- **Individual Performance Extraction:** the model is given by the performance of a certain player u in a match m. The framework is designed to work with any set of features, thus giving to the user a high flexibility about the description and deployment of soccer performance.
- **Player Rating:** is given by the scalar product between the values of the features referring to a given match m and the feature weight w, that was computed during the learning phase.

With this, the performance is evaluated, given the following formula:

$$r(u,m) = \frac{1}{R} \sum_{i=1}^{n} w_i \times x_i.$$

Figure 12 - Performance Evaluation Formula (Pappalardo et al., 2018).

The study, in an experimental phase, was done using a massive database provided by *Wyscout*, consisting 31 496 332 events, capturing 16 619 matches, 296 clubs and 21 361 players, that do not include goalkeepers, in different seasons, from 18 different competitions around the world, including the Portuguese *Primeira Liga* (see table 7).

Table 7 - List of competitions from WyScout database (Pappalardo et al., 2018).

| competition | area | type | #seasons | #matches | #events | #players |
|---|---|---|---|---|---|---|
| La Liga | Spain | national | 4 | 1520 | 2,541,873 | 1264 |
| Premier League | England | national | 4 | 1520 | 2,595,808 | 1231 |
| Serie A | Italy | national | 4 | 1520 | 2,610,908 | 1499 |
| Bundesliga | Germany | national | 4 | 1124 | 2,075,483 | 1042 |
| Ligue 1 | France | national | 4 | 1520 | 2,592,708 | 1288 |
| Primeira Liga | Portugal | national | 4 | 1124 | 1,720,393 | 1227 |
| Super Lig | Turkey | national | 4 | 1124 | 1,927,416 | 1182 |
| Souroti Super Lig | Greece | national | 4 | 1060 | 1,596,695 | 1151 |
| Austrian Bundesliga | Austria | national | 4 | 720 | 1,162,696 | 593 |
| Raiffeisen Super League | Switzerland | national | 4 | 720 | 1,124,630 | 647 |
| Football Championship | Russia | national | 4 | 960 | 1,593,703 | 1046 |
| Eredivisie | The Netherlands | national | 4 | 1248 | 2,021,164 | 1177 |
| Superliga | Argentina | national | 4 | 1538 | 2,450,170 | 1870 |
| Campeonato Brasileiro Serie A | Brazil | national | 4 | 1437 | 2,326,690 | 1790 |
| UEFA Champions League | Europe | continental | 3 | 653 | 995,363 | 3577 |
| UEFA Europa League | Europe | continental | 3 | 1416 | 1,980,733 | 9100 |
| UEFA Euro Cup 2016 | Europe | continental | 1 | 51 | 78,140 | 552 |
| FIFA World Cup 2018 | World | international | 1 | 64 | 101,759 | 736 |
| | | | 64 | 19,619 | 31,496,332 | (*)21,361 |

The events record the following variables that can be confirmed on the example on figure 14, showing the simple pass done by *Rafinha*, from *Internazionale*, on the match against Lazio:

- A unique event identifier
- The type of event
- A timestamp
- The player related to the event
- The team where the player plays
- The match in which the event was observed
- The position on soccer field
- The event subtype
- List of tags


```
{"id": 253668302,
"eventName": "Pass",
"eventSec": 2.41,
"playerId": 3344,
"matchId": 2576335,
"teamId": 3161,
"positions": [{"x": 49, "y": 50}],
"subEventId": 85,
"subEventName": "Simple pass",
"tags": [{"id": 1801}]}
```


Figure 13 – Example of event in dataset (Pappalardo et al., 2018).

The events can also be shown as a pictorial representation, as shown in Figure 14. This example corresponds to the events produced by Lionel Messi during a match in the Spanish *La Liga*, in the 2015/2016 season.

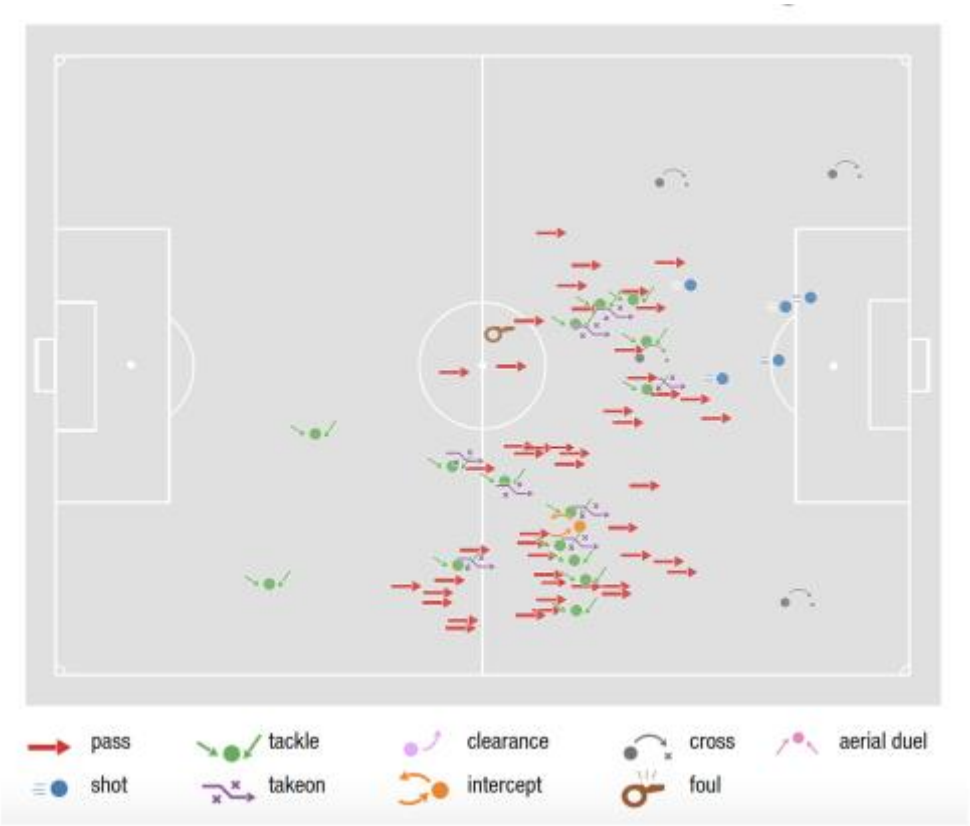


Figure 14 – Events observed for *Lionel Messi* during a Spanish *La Liga* match (Pappalardo et al., 2018).

The events and sub-events that were detailed by the dataset, can be seen detailed, on table 8, as well as the tags that give additional information to the event.

Table 8 – Event Types, Subtypes, and possible Tags (Pappalardo et al., 2018).

| type | subtype | tags |
|---|---|---|
| ***pass*** | cross, simple pass | accurate, not accurate, key pass, opportunity, assist, (goal) |
| ***foul*** | | no card, yellow, red, 2nd yellow |
| ***shot*** | | accurate, not accurate, block, opportunity, assist, (goal) |
| ***duel*** | air duel, dribbles, tackles, ground loose ball | accurate, not accurate |
| ***free kick*** | corner, shot, goal kick, throw in, penalty, simple kick | accurate, not accurate, key pass, opportunity, assist, (goal) |
| ***offside*** | | |
| ***touch*** | acceleration, clearance, simple touch | counter attack, dangerous ball lost, missed ball, interception, opportunity, assist, (goal) |

**Performance Extraction:** This phase corresponds to data transformation, where the 76 features were extracted to a range of [0, 1], to guarantee that all variables are expressed on the same scale. Then, a performance vector was built for a player in match with the count of the events of a given type, subtype and correspondent tags associated.

**Role Detection:** A role detection algorithm was run and discovered eight different clusters, corresponding to eight different roles the players have on pitch. Considering this analysis only considers 10 players per team, means that a role can have multiple players included on the team. The clusters (roles) can be found on figure 15 and its explanation on table 9.

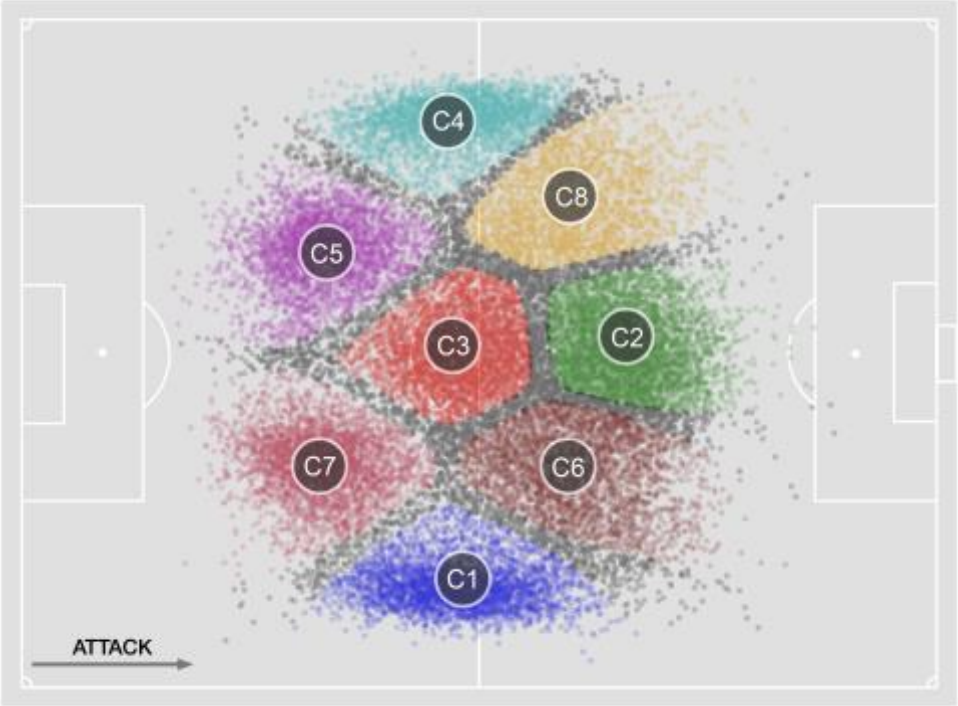


Figure 15 – Grouping of the centres of performance in the clusters (roles) (Pappalardo et al., 2018).

Table 9 – Interpretation of the eight clusters (roles) (Pappalardo et al., 2018).

| cluster | name | description | examples |
|---|---|---|---|
| C1 | right fielder | plays on the right side of the field, as a wing, back, or both | Sergi Roberto, Danilo |
| C2 | central forward | plays in the center of the field, close to the opponent's area | Messi, Suárez |
| C3 | central fielder | plays in the center of the field | Kroos, Pjanić |
| C4 | left fielder | plays on the left side of the field, as a wing, back, or both | Nolito, Jordi Alba |
| C5 | left central back | plays close to his own goal, preferably on the left | Bartra, Maguire |
| C6 | right forward | plays on the right side of the field, close to the opponent's area | Robben, Dembélé |
| C7 | right central back | plays close to his own goal, preferably on the right | Javi Martínez, Matip |
| C8 | left forward | plays on the left side of the field, close to the opponent's area | Neymar, Insigne |

According (Pappalardo et al., 2018), existing player ranking approaches report judgements that consist mainly of informal interpretations based on some simplistic metrics is important, instead of evaluate the goodness of ranking and performance evaluation algorithms in quantitative manner, through the help of human experts.

To validate the PlayeRank, a survey was created and submitted to three professional soccer talent scouts. The questionnaire consisted of a randomly selected 35% of players of the dataset, the generation of pairs involving 202 distinct players, with the objective of selecting the best player per pair.

The applications of PlayeRank are:

- **Retrieval of players:** searching players on the database throw the use of queries, that considers the events occurred during a match and their position on the field.

- **Versatility:** the role detector algorithm on the role detection phase of study, can give important information on the player's propensity to change role from match to match.

The conclusions given by (Pappalardo et al., 2018) is that the studied framework outperform existing approaches, being more concordant with professional soccer scouts. It can be used as a support tool in evaluating, searching, ranking, and recommending football players.

The tool can also be developed to be adapted to various and more sophisticated algorithms, to detect different player roles during the same match. Another improvement is to allow the use of different data sources, using out of possession and positioning on the pitch without the ball events. To conclude, table 10 shows the list of features used on this study.

Table 10 – List of 76 features extracted from database (Pappalardo et al., 2018).

| type | feature | type | feature |
|---|---|---|---|
| duel | duel-air duel-accurate | others on the ball | others on the ball-accelleration-accurate |
| | duel-air duel-not accurate | | others on the ball-accelleration-not accurate |
| | duel-ground attacking duel-accurate | | others on the ball-clearance-accurate |
| | duel-ground attacking duel-not accurate | | others on the ball-clearance-not accurate |
| | duel-ground defending duel-accurate | | others on the ball-touch-assist |
| | duel-ground defending duel-not accurate | | others on the ball-touch-counter attack |
| | duel-ground loose ball duel-accurate | | others on the ball-touch-dangerous ball lost |
| | duel-ground loose ball duel-not accurate | | others on the ball-touch-feint |
| foul | foul-hand foul-red card | | others on the ball-touch-interception |
| | foul-hand foul-second yellow card | | others on the ball-touch-missed ball |
| | foul-hand foul-yellow card | | others on the ball-touch-opportunity |
| | foul-late card foul-yellow card | pass | pass-cross pass-accurate |
| | foul-normal foul-red card | | pass-cross pass-assist |
| | foul-normal foul-second yellow card | | pass-cross pass-key pass |
| | foul-normal foul-yellow card | | pass-cross pass-not accurate |
| | foul-out of game foul-red card | | pass-hand pass-accurate |
| | foul-out of game foul-second yellow card | | pass-hand pass-not accurate |
| | foul-out of game foul-yellow card | | pass-head pass-accurate |
| | foul-protest foul-red card | | pass-head pass-assist |
| | foul-protest foul-second yellow card | | pass-head pass-key pass |
| | foul-protest foul-yellow card | | pass-head pass-not accurate |
| | foul-simulation foul-second yellow card | | pass-high pass-accurate |
| | foul-simulation foul-yellow card | | pass-high pass-assist |
| | foul-violent foul-red card | | pass-high pass-key pass |
| | foul-violent foul-second yellow card | | pass-high pass-not accurate |
| | foul-violent foul-yellow card | | pass-launch pass-accurate |
| free kick | free kick-corner free kick-accurate | | pass-launch pass-assist |
| | free kick-corner free kick-not accurate | | pass-launch pass-key pass |
| | free kick-cross free kick-accurate | | pass-launch pass-not accurate |
| | free kick-cross free kick-not accurate | | pass-simple pass-accurate |
| | free kick-normal free kick-accurate | | pass-simple pass-assist |
| | free kick-normal free kick-not accurate | | pass-simple pass-key pass |
| | free kick-penalty free kick-not accurate | | pass-simple pass-not accurate |
| | free kick-shot free kick-accurate | | pass-smart pass-accurate |
| | free kick-shot free kick-not accurate | | pass-smart pass-assist |
| | free kick-throw in free kick-accurate | | pass-smart pass-key pass |
| | free kick-throw in free kick-not accurate | | pass-smart pass-not accurate |
| | | shot | shot-shot-accurate |
| | | | shot-shot-not accurate |

In the (Li et al., 2020) research, a dataset of all 1200 matches from 2014 to 2018 Chinese Super League matches provided by the company Champion Tecnology Co. Ltd. The process started with the feature selection, creating a total of 164 match events from the cleaned raw data.

With the using of the team-rank framework, consisting of three main phases, shown on Figure 16: (a) the performance extraction phase chooses 22 features from the dataset and extracts the

performance vector and the match outcome. (b) The learning phase that solves a classification problem and learns the weight of each feature. (c) The rating and ranking phase rates the matches based on the feature weights and ranks team by their season average rating (Li et al., 2020).

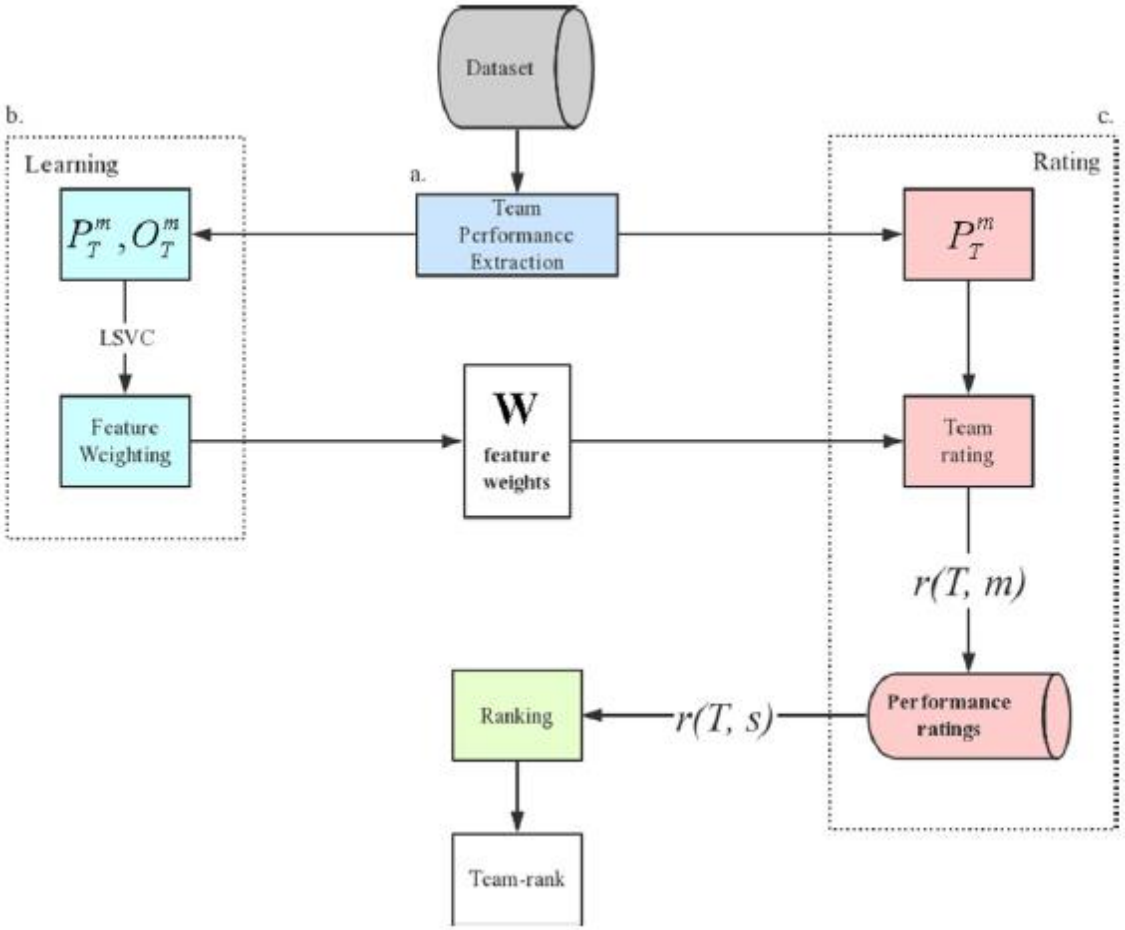


Figure 16 – Schema of the Team-Rank Framework (Li et al., 2020).

To validate the prediction accuracy of LSVC model, the simulation was made using two different outputs: (1) Team-rank, which is based on the calculated team performance within a single match or a season, and (2) Predicted Ranking, which the end-of-season ranking is given the actual performance of teams at each match.

(Li et al., 2020) presented the results using different methods.

**Descriptive Analysis of Selected Features**

On Table 11, we can see the average values and the result pf one-way ANOVA of each feature, that is represented on the same table.

Table 11 – Differences between winning, drawing and losing teams in game statistics (Li et al., 2020).

| Feature name | Win Mean (SD) | Draw | Loss | F | P |
|---|---|---|---|---|---|
| Shots | 13.33 (4.38) | 12.17 (4.22) | 11.75 (4.27) | 31.43 | <0.001 |
| Shots on target | 5.54 (2.39) | 3.98 (2.19) | 3.54 (2.16) | 188.95 | <0.001 |
| Shot on target in penalty area | 3.98 (1.96) | 2.63 (1.68) | 2.22 (1.64) | 233.55 | <0.001 |
| Penalty | 0.24 (0.46) | 0.13 (0.35) | 0.12 (0.35) | 22.83 | <0.001 |
| Bad shot% | 0.69 (0.15) | 0.78 (0.16) | 0.80 (0.16) | 115.24 | <0.001 |
| Pass | 382.23 (61.7) | 375.85 (56.80) | 386.69 (56.78) | 6.3 | 0.002 |
| Pass success | 296.86 (65.73) | 286.68 (60.08) | 295.28 (58.65) | 5.54 | 0.004 |
| Pass attacking success | 63.88 (20.32) | 59.96 (21.77) | 59.32 (23.21) | 10.96 | <0.001 |
| Pass forward success% | 0.69 (0.07) | 0.67 (0.07) | 0.66 (0.07) | 52.3 | <0.001 |
| Possession | 0.51 (0.07) | 0.5 (0.07) | 0.49 (0.07) | 5.82 | 0.003 |
| Cross | 14.67 (5.63) | 16.69 (6.29) | 16.63 (6.23) | 30.22 | <0.001 |
| Cross success | 4.48 (2.40) | 4.88 (2.70) | 4.52 (2.62) | 5.14 | 0.006 |
| Lost ball | 24.92 (6.91) | 24.43 (6.98) | 25.84 (7.48) | 7.78 | <0.001 |
| Tackles | 17.65 (6.19) | 16.46 (5.67) | 16.82 (5.37) | 8.88 | <0.001 |
| Saves | 2.34 (1.85) | 2.29 (1.88) | 2.61 (1.96) | 6.7 | 0.001 |
| Red card | 0.05 (0.21) | 0.08 (0.28) | 0.13 (0.35) | 18.2 | <0.001 |
| Pen opponent | 0.12 (0.35) | 0.13 (0.35) | 0.24 (0.46) | 22.83 | <0.001 |
| Interceptions | 20.4 (12.07) | 19.88 (11.47) | 18.61 (10.87) | 5.6 | 0.004 |
| Defensive Foul | 14.39 (5.25) | 13.64 (5.16) | 13.43 (5.11) | 8.16 | <0.001 |
| Clearances | 20.62 (8.15) | 19.79 (7.64) | 16.99 (6.92) | 54.59 | <0.001 |
| Shots opponent | 11.75 (4.27) | 12.17 (4.22) | 13.33 (4.38) | 31.43 | <0.001 |
| Shots on target opponent | 3.54 (2.16) | 3.98 (2.19) | 5.54 (2.39) | 188.95 | <0.001 |

**LSVC Model**

This model shows the Receiver Operating Characteristic Curve (ROC) of each fold and the mean ROC (Li et al., 2020). After training and validating of the model, the prediction accuracy is shown on Figure 17.

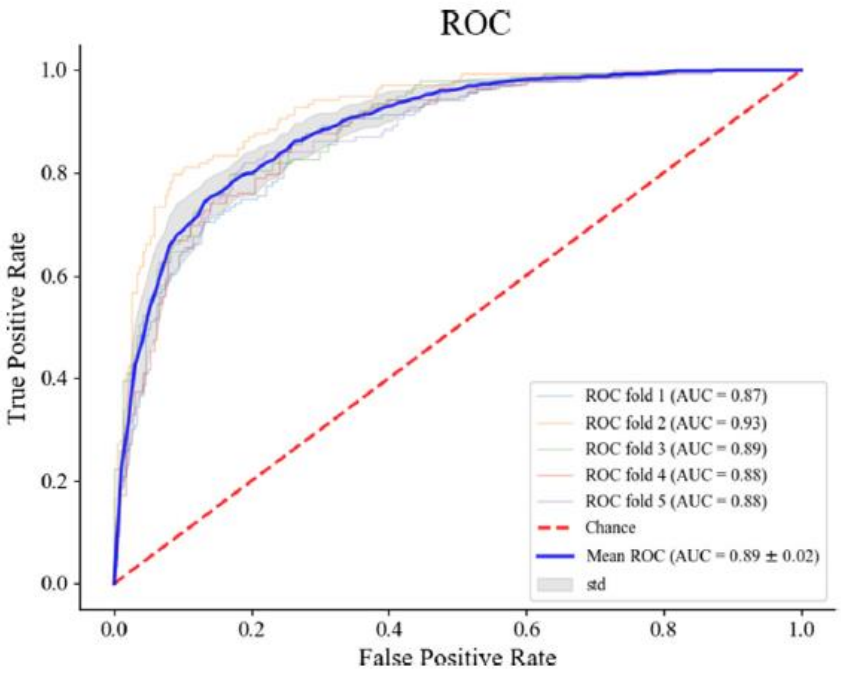


Figure 17 – Mean ROC and ROC of each validated fold note (Li et al., 2020).

The results and conclusions taken by the research of (Li et al., 2020) was that better performance does not mean a winning and high ranking does not always mean a better performance, but better teams could maintain a performance that have bigger chance to win.

On (Zambom-Ferraresi et al., 2018), the sample of data is composed of three seasons from the one starting in 2012 and the one finishing on 2015, considering the 'big five' leagues (English Premier League, Spanish *La Liga*, Italian *Serie A*, Deutch *Bundesliga*, and French *Ligue 1*), implying a coverage of 5532 games. The datasource is from the company OPTAPro, that is, according to (Liu et al., 32), one of the largest databases of European football.

The outcome variable of (Zambom-Ferraresi et al., 2018) study was the number of points of each team in each league and season. The raw data suffered, on the data transformation phase of analysis, a max-min method normalization, scaling the total points between 0 and 1. These arose from the problem that the number of teams playing in the leagues were not equal, changing the number of games and possible points.

The 24 variables taken into consideration by (Zambom-Ferraresi et al., 2018) research, was retrieved from the references that served as literature review for the study, and are represented on Table 13.

Table 12 – Definitions and descriptives statistics of the explanatory variables (Zambom-Ferraresi et al., 2018).

| Variable | Definition | Mean | STD | Expected Effect |
|---|---|---|---|---|
| Outcome variable | | | | |
| Sport performance | Normalized total points archived by clubs at the end of a season | 0.4 | | |
| A. Attack plays | | | | |
| Total shots attempted | Shot: An attempt to score a goal, made with any part of the body, either on or off target. The outcomes of a shot could be: goal, shot on target, shot off target, blocked shot, post | 367.80 | 60.83 | + |
| Shots on target | Total shots on target | 164.76 | 36.42 | + |
| Total passes, (excl. crosses, and corners) | Pass: An intentionally played ball from one player to another | 15,902.81 | 2739.50 | + |
| Passing accuracy (excl. crosses and corners) | Successful passes/total passes | 0.78 | 0.05 | + |
| Assists | The final pass or cross leading to the recipient of the ball scoring a goal | 34.13 | 12.66 | + |
| Crosses attempted | Any ball played into the opposition team's area from a wide position | 603.79 | 131.86 | + |
| Corners taken (incl. short corners) | A corner kick is a method of restarting play. It is awarded to the attacking team when the ball leaves the field of play crossing the goal line | 192.19 | 32.62 | + |
| Dribbles and runs attempted | An attempt by a player to beat an opponent in possession of the ball. A successful dribble: the player beats the defender; unsuccessful: the dribbler is tackled | 746.89 | 161.07 | + |
| Dribble and run success rate | Effective dribbles and runs with respect to the total number attempted | 0.45 | 0.07 | + |
| Long pass final third | A pass over 32 m on the final third of the field (attack of the reference team) | 931.76 | 156.55 | + |
| Through ball | A pass playing a player through on goal, which could lead to a goal scoring opportunity. The pass needs to split the last line of defence and plays the teammate through on goal. | 27.60 | 18.70 | + |
| Offsides | Being caught in an offside position resulting in a free kick to the opposing team | 88.71 | 19.14 | ? |
| B. Defence plays | | | | |
| Total shots conceded | Total shots attempted for the opposite team | 164.76 | 29.78 | − |
| Tackles attempted | The act of gaining possession from an opposition player when he is in possession of the ball | 754.60 | 79.47 | − |
| Tackled possession retained (%) | A tackle won when a player makes a tackle and possession is retained by his team | 0.23 | 0.03 | + |
| Recoveries | The event given at the start of a team's recovery of possession from open play. The defending team must have full control of the ball and must start a new passage of play. | 2071.00 | 297.60 | + |
| Recoveries in opp half | A recovery on the opposite team's field (attack of reference team) | 400.63 | 93.26 | + |
| Clearances, blocks, and interceptions | Attempts to get the ball out of the danger zone when there is pressure. A defensive block, blocking a shot going on target. An interception is given when a player intercepts a pass with some movement | 1750.02 | 246.97 | ? |
| Total fouls conceded | Any infringement that is penalised as foul play by a referee | 517.86 | 73.38 | − |
| Fouls conceded in danger area | Infringement that is penalised as foul play by a referee in the lower 1/3 | 106.59 | 18.87 | − |
| Yellow cards | Indicates that a player has been officially cautioned/penalised due to infringement. A player receiving two yellow cards in a match is sent off. | 75.59 | 20.28 | − |
| Red cards | A red card is shown by a referee to signify that a player has been sent off. | 4.46 | 2.66 | − |
| Saves made | The goalkeeper prevents the ball from entering the goal with any part of his body. | 112.22 | 20.79 | + |
| Catches | The goalkeeper catching a cross or a ball played into the area when there is pressure from the rival | 52.11 | 17.18 | + |

To analyse the determinants of sport performance (Zambom-Ferraresi et al., 2018) considered a linear regression model, which formula is represented on Figure 19. The equation denotes a dimensional vector, consisting of observation for normalized sport performance index for each team. This considers a matrix of the variables that were taken into consideration on Table 12.

$$y = \alpha\iota_{nt} + X\beta + \varepsilon$$

Figure 18 – Linear Regression Model Equation

After applying the Bayesian Model Averaging and a robustness check, the results obtained by the (Zambom-Ferraresi et al., 2018) research are that the results may consist on useful inputs for decision-making in football. Based on results observation, it is possible to conclude that assists and through balls are more important than dribbles, runs, and crosses. These can indicate that this should be the main focus on team's training. Additionally, the second conclusion, is related to team management, concluding that managers should hire players with skills and abilities that are more approximate with the determinants studied. The complete analysis for the determinants given by Table 13.

Table 13 – Robustness Check: Relative importance decomposition by league (Zambom-Ferraresi et al., 2018).

| Variable | Big five | Premier | La Liga | Serie A | Bundesliga | Ligue 1 |
|---|---|---|---|---|---|---|
| A. Attack | | | | | | |
| Total shots attempted | 0.062 | 0.078 | 0.066 | 0.066 | 0.066 | 0.057 |
| Shots on target | 0.098 | 0.156 | 0.085 | 0.085 | 0.092 | 0.067 |
| Total passes | 0.071 | 0.064 | 0.058 | 0.058 | 0.099 | 0.076 |
| Passing accuracy | 0.079 | 0.067 | 0.068 | 0.068 | 0.083 | 0.105 |
| Assists | 0.179 | 0.193 | 0.202 | 0.202 | 0.158 | 0.131 |
| Crosses attempted | 0.005 | 0.002 | 0.004 | 0.004 | 0.004 | 0.012 |
| Corners taken | 0.033 | 0.052 | 0.032 | 0.032 | 0.032 | 0.023 |
| Dribbles and runs attempted | 0.009 | 0.030 | 0.015 | 0.015 | 0.011 | 0.013 |
| Dribble and run success rate | 0.007 | 0.005 | 0.010 | 0.010 | 0.013 | 0.005 |
| Long pass final third | 0.015 | 0.031 | 0.005 | 0.005 | 0.005 | 0.025 |
| Through ball | 0.051 | 0.040 | 0.055 | 0.055 | 0.052 | 0.047 |
| Offsides | 0.014 | 0.003 | 0.054 | 0.054 | 0.006 | 0.013 |
| B. Defence | | | | | | |
| Shots conceded | 0.215 | 0.135 | 0.148 | 0.148 | 0.157 | 0.189 |
| Tackles attempted | 0.003 | 0.008 | 0.011 | 0.011 | 0.024 | 0.004 |
| Tackled possession retained % | 0.003 | 0.005 | 0.007 | 0.007 | 0.003 | 0.008 |
| Recoveries | 0.015 | 0.008 | 0.022 | 0.022 | 0.006 | 0.012 |
| Recoveries in opp. half | 0.011 | 0.027 | 0.020 | 0.020 | 0.008 | 0.012 |
| Clearances, blocks and intercept | 0.017 | 0.020 | 0.013 | 0.013 | 0.016 | 0.022 |
| Total fouls conceded | 0.007 | 0.007 | 0.016 | 0.016 | 0.022 | 0.009 |
| Fouls conceded (danger area) | 0.033 | 0.023 | 0.028 | 0.028 | 0.041 | 0.104 |
| Yellow cards | 0.005 | 0.002 | 0.027 | 0.027 | 0.036 | 0.003 |
| Red cards | 0.005 | 0.003 | 0.006 | 0.006 | 0.016 | 0.003 |
| Saves made | 0.058 | 0.040 | 0.046 | 0.046 | 0.024 | 0.049 |
| Catches | 0.003 | 0.005 | 0.004 | 0.004 | 0.024 | 0.009 |

On their research, (Wu et al., 2020) constrcuted a social network, having in mind playing positions and passing processes. The objetive was to quantify the importance of each of the positions on the pitch, using a high diversity of methods, such as degree of centrality, closeness centrality, betweeness centrality, eigenvector centrality, load centrality, reciprocity, and clustering.

The methodology used on the (Wu et al., 2020) study consisted of an undirected graph, using a set of points and an edge set. The adjancy matrix, for this, is only denoted if there are connected edges between nodes. Three definitions for this study were given as (1) the degree ($k_i$) of a node ($v_i$) in an undirected netwoek refers to the number of edges directly connected to the note. The in-degree (IDC) of a directed network refers to the number of edges pointing in to a node. The out-degree (ODG) refers to the number of edges pointing out of a node, formulated on Figure 19. (2) node distance is defined as the number of edges on the shortest path connecting these two nodes, represented by $d_{ij}$. (3) the average path lenght L on a network is defined as the average distance between two nodes, as formulated on Figure 20.

$$k_i = \sum_{j=1}^{N} a_{ij}$$

Figure 19 – Formula given by definition 1 (Wu et al., 2020).

$$L = \frac{2}{N(N-1)} \sum_{i \geq j} d_{ij}$$

Figure 20 – Formula given by definition 2 (Wu et al., 2020).

The following definitions are regarding the methods used presented early, to better understand the results phase of (Wu et al., 2020) research.

**Degree of Centrality (DC):** this method measures the overall level of connection a player has with his teammates. Out Degree Centrality (ODC) means the connection of a player passing the ball to his mates. In Degree Centrality means the connection of a player getting the ball from his mates (Wu et al., 2020).

**Closeness Centrality (CC):** reflects the degree to which a node is centered in network. It can be characterized by the average of shortest distance (Wu et al., 2020).

**Betweeness Centrality (BC):** refers to the shortest path (Wu et al., 2020).

**Eigenvector Centrality (EC):** evaluates the importance of nodes in the network (Wu et al., 2020).

**Reciprocity Centrality (LC):** measures the propotion of directed edges that are bidirectional (Wu et al., 2020).

**Clustering (CL):** measurement of the degree to which nodes in a network tend to cluster together ((Wu et al., 2020).

After applying the methods explained above to each position on the field (Wu et al., 2020) concluded that the advanced midfielder position was the first ranked under the most measures. Left Forward (LFW) and Right Forward (RFW) also got a highest rank in the BC, and LC measures, for LFW and in the CL measure for the RFW position, as shown on Table 14.

To conclude (Wu et al., 2020) stated that by analyzing the characteristics of passing network in different moments of the game, specially temporal and specific situations, the team's performance can be tracked.

On the (Beal et al., 2020) research, two formal models can be defined for the game of football: the prematch tactical decision making process, and the in-match decisions.

Since there are many unknown factors about the opposition, a Bayeasian game is used to model strategic decisions and make optimised decisions, by defining two teams, with a respective action set, that includes all tactical choices before the match. In this model, used by (Beal et al., 2020), it is assumed that both teams do not know each other teams' tactics, and both of them have acces to the

same assumptions. The three approaches made for this model are (1) best response – maximises the chances of a win, (2) spiteful – minimises the chances of losing the game, and (3) minimax – maximises the chances of team winning the game and the other team minimising the chances of winning.

Table 14 – Rank of Importance (Wu et al., 2020).

**Table 3**
Rank of Importance.

| | IDC | ODC | CC | BC | EC | LC | RC | CL | ODG | IDG | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **AMC** | 1 | 2 | 1 | 11 | 1 | 12 | 1 | 2 | 1 | 1 | 3.3 |
| **MC** | 3 | 3 | 4 | 8 | 4 | 8 | 2 | 7 | 2 | 2 | 4.3 |
| **LFW** | 2 | 6 | 2 | 1 | 2 | 1 | 7 | 14 | 8 | 3 | 4.6 |
| **DL** | 7 | 8 | 7 | 4 | 5 | 4 | 3 | 5 | 5 | 4 | 5.2 |
| **DMC** | 6 | 5 | 6 | 10 | 7 | 10 | 4 | 8 | 3 | 5 | 6.4 |
| **DR** | 9 | 7 | 9 | 5 | 10 | 5 | 6 | 9 | 6 | 7 | 7.3 |
| **RFW** | 4 | 12 | 3 | 13 | 3 | 11 | 5 | 1 | 13 | 9 | 7.4 |
| **DML** | 10 | 1 | 10 | 7 | 9 | 6 | 10 | 3 | 7 | 11 | 7.4 |
| **FW** | 5 | 10 | 5 | 3 | 8 | 3 | 11 | 12 | 11 | 8 | 7.6 |
| **MR** | 8 | 9 | 8 | 6 | 6 | 7 | 8 | 10 | 9 | 6 | 7.7 |
| **DC** | 11 | 4 | 12 | 2 | 14 | 2 | 9 | 13 | 4 | 10 | 8.1 |
| **AMR** | 12 | 11 | 11 | 9 | 11 | 9 | 12 | 6 | 10 | 12 | 10.3 |
| **DMR** | 14 | 14 | 14 | 12 | 13 | 13 | 13 | 4 | 12 | 14 | 12.3 |
| **ML** | 13 | 13 | 13 | 15 | 12 | 15 | 14 | 11 | 14 | 13 | 13.3 |
| **AML** | 15 | 15 | 15 | 14 | 15 | 14 | 15 | 15 | 15 | 15 | 14.8 |
| **GK** | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |

Ranks of IDC, ODC, CC, BC, EC, LC, RC, CL, ODG and IDG are in decreasing order, since smaller value indicates more importance in our calculation. The last column is the mean of ranks measured by IDC, ODC, CC, BC, EC, LC, RC, CL, ODG and IDG.

On the (Beal et al., 2020) research, two formal models can be defined for the game of football: the prematch tactical decision making process, and the in-match decisions.

Since there are many unknown factors about the opposition, a Bayeasian game is used to model strategic decisions and make optimised decisions, by defining two teams, with a respective action set, that includes all tactical choices before the match. In this model, used by (Beal et al., 2020), it is assumed that both teams do not know each other teams' tactics, and both of them have acces to the same assumptions. The three approaches made for this model are (1) best response – maximises the chances of a win, (2) spiteful – minimises the chances of losing the game, and (3) minimax – maximises the chances of team winning the game and the other team minimising the chances of winning.

On the other hand, the progress of the game has a lot of changes, in different factors, so (Beal et al., 2020), model the in-game tactical decisions as a stochastic game. In this case, two teams are defined, there is a set of states that represents the possible scoreline, and a set of strategies.

To solve the pre-match bayesian game, when predicting the opposition strategy, a cluster for the teams is created into different play styles categories, allowing (Beal et al., 2020) to evaluate the style

of a team. To learn the payoffs, the team's tactical style, potential formation and team strenght to give a probability of winning the game. Lastly, to optimse the pre-match tactics, the best decisions are tried to be found, finding the most probable oppositions tactics to happen.

To solve the in-match stochastic game, will allow teams to make in-match decisions, that can improve chances of winning their games. To optimise game tactics (Beal et al., 2020), the payoffs are pre-computed, and the optimised action is selected, using either aggressive approach or a more reserved approach.

The key observations of (Beal et al., 2020), is that away teams are more likely to minimise the chances of the home team of winning rather than maximise their own chances. On Figure 21, we can confirm that that using the spiteful approach, the chances of drawing is less than losing the game, in comparison to the other approaches. The in-game decisions and model is more challenging due to great uncertainty regarding probabilities.

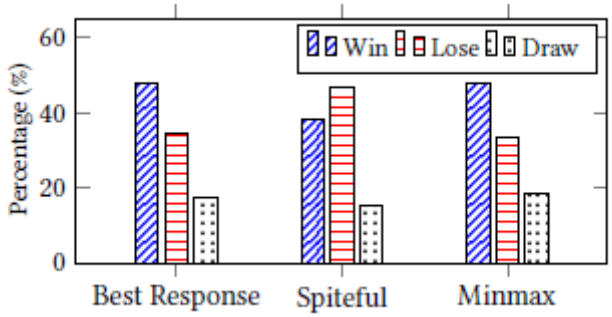


Figure 21 – Percentage of real-world results with close tactic selection (Beal et al., 2020).

### 4.2.2. Pre-Game Prediction

A different study, with the objective of predicting football match results using decision support systems, which used the CRISP-DM methodology, combined with Simon phases of decision-making and Design Science Research (DSR) methodology, as explained on Table 11 (Gomes et al., 2015).

The study and application of the methodology, for this study was applied in 4 phases:

**Phase 1:** identification of the problem. The conclusion was that betting systems are more profitable, since their users have a low margin of profit over them. Despite that, it corresponds to a semi-structured decision because it is necessary to complement the collected data with the information possessed by users.

**Phase 2:** definition of project goals and dataset collection. The dataset consisted on data from the last 13 seasons of the English Premier League, corresponding to statistical information for 4940 games. The data was treated by an extract, transform and loading (ETL) process, as showed on figure 16.

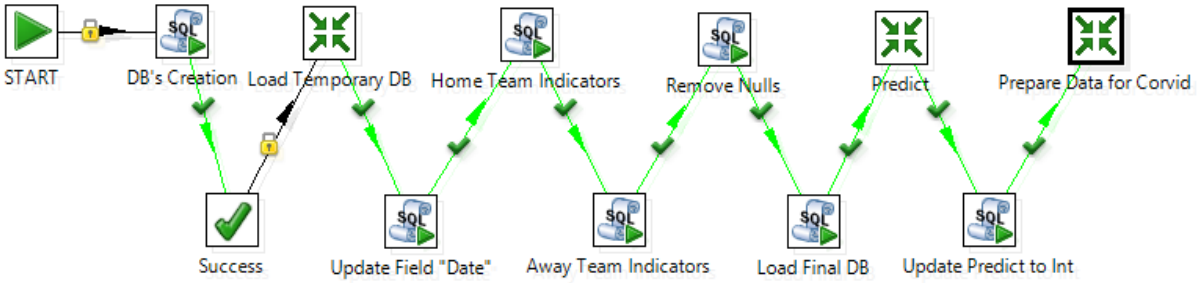


Table 15 – Combined Methodology (Gomes et al., 2015).

| | | Combined Methodology - DSR | | | |
|---|---|---|---|---|---|
| | | **Phase 1** | **Phase 2** | **Phase 3** | **Phase 4** |
| **CRISP-DM** | Business Understanding | X | | | |
| | Data Understanding | | X | | |
| | Data Transformation | | X | | |
| | Modelling | | X | | |
| | Evaluation | | | X | |
| | Deployment | | | | X |
| **Decision-making** | Intelligence | X | | | |
| | Design | | X | | |
| | Choice | | | X | |
| | Implementation | | | | X |

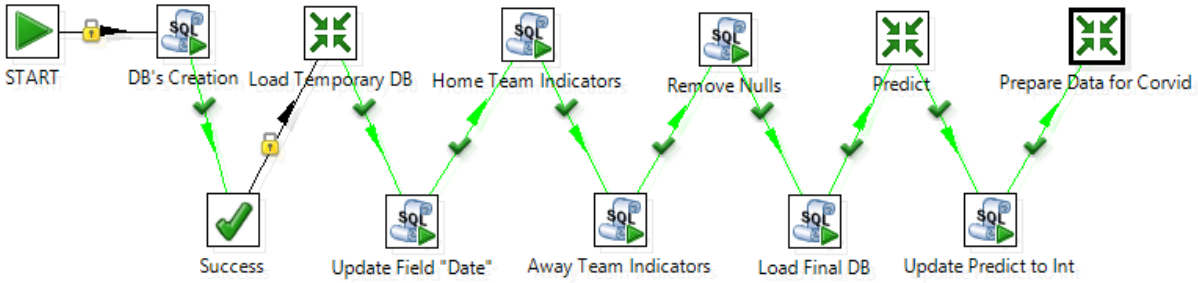


Figure 22 – ETL Process

This process consists of creating a temporary database and one to store the processed data. In this phase, new variables were created to add to the ones that already existed, and the existing nulls fields were removed.

The Data Mining models were induced using three different techniques, after loading the final dataset. This techniques were Naïve Bayes, Decision Trees and Support Vector Machine and two

sampling methods, 10-Folds Cross Validation and Percentage Split (66% percent to training and the rest for testing. The final variables used were:

- Home Team
- Away Team
- Average Goals Home / Away Team – stores the goal average of the home or away team, in matches disputed at home / away.
- Average Shots Home / Away Team – stores the shots average of the home or away team, in matches disputed at home / away.
- Average Shots on Target Home / Away Team – stores the shots on target average of the home or away team, in matches disputed at home / away.
- Home / Away Win Last Five – number of victories in the last five game of the home / away team.
- Home / Away Win Last Five Confrontation – number of victories in direct confrontation in the last five games in home / away games.

The target variable was FTR (Final Time Result).

**Phase 3:** as result, it is possible two obtain three distinct predictions (Home Win – "1", Draw – "2", Away Win – "3"). In this phase, the models were analysed, and the obtained results are shown on Table 12. By the table observation, we can see that although the models' accuracy is very identical, the model 3 is the best.

Table 16 – Models Evaluation (Gomes et al., 2015).

| Model | Sampling Method | Algorithm | Accuracy |
|---|---|---|---|
| **Model 1** | Percentage Split | NaiveBayes | 0.487 |
| **Model 2** | Percentage Split | J48 | 0.472 |
| **Model 3** | Percentage Split | LibSVM | 0.508 |
| **Model 4** | 10-Folds CV | NaiveBayes | 0.492 |
| **Model 5** | 10-Folds CV | J48 | 0.476 |
| **Model 6** | 10-Folds CV | LibSVM | 0.492 |

**Phase 4:** in this phase, the data treatment and transformation was performed, and the decision support system was developed. After the system creation, tests were performed four seven Premier League matches, assuming a bet of 100€ in each of the ten games by each round of matches. On Table 13, there is the system tests performed.

Table 17 – System Performed Tests (Gomes et al., 2015).

| Round | Percentage of Correct Bets | Return (Bets of 100€ by game) |
|---|---|---|
| **Round 5** | 80 % | 689 € |
| **Round 10** | 30 % | -418 € |
| **Round 15** | 40 % | 11 € |
| **Round 20** | 70 % | 713 € |
| **Round 25** | 60 % | 480 € |
| **Round 30** | 40 % | -245 € |
| **Round 35** | 60 % | 179 € |
| **Total** | **Average = 54,29%** | **1409 €** |

In total, in this simulation, the bet total 7000€ and obtained a return of 1409€, about 20%. In the Table 14, we can see a detailed example about the fifth round of the Premier League season.

Table 18 – Return in fifth Premier League round (Gomes et al., 2015).

| | Game | Result (1,2,3) | Corvid Output (1,2,3) | Return (Bets of 100€) |
|---|---|---|---|---|
| **Round 5** | Norwich x A. Villa | 3 | 3 | 220€ |
| | Liverpool x Southampton | 3 | 1 | -100€ |
| | Newcastle x Hull City | 3 | 2 | -100€ |
| | West Brom x Sunderland | 1 | 1 | 110€ |
| | West Ham x Everton | 3 | 3 | 140€ |
| | Chelsea x Fulham | 1 | 1 | 33€ |
| | Arsenal x Stoke | 1 | 1 | 40€ |
| | C. Palace x Swansea | 3 | 3 | 130€ |
| | Cardiff x Tottenham | 3 | 3 | 91€ |
| | Man City x Man Utd | 1 | 1 | 125€ |
| | | | **Total** | **689€** |

The conclusion taken from this study is that the Decision Support Systems have future to continue their research and development. In this case, adapting the results to performance, with this type of software and tools, staff teams can be more aware on what are their team chances on winning the game.

The statistical analysis performed on (Rajesh et al., 2020), follows qualitative and quantitative measures of attributes consider into account:

- Overall Distribution Value of each player according to overall rating, performance, and age factors.
- Number of players distribution from each nationality and their social impacts for investors to select a player as brand ambassadors.
- Comparing overall performance and potentially of players by nationality.

The comparison is based on a qualitative property, which is the player nationality, and multiple quantity properties (performance, potentiality, age, value, wage, position, crossing, finishing, heading accuracy, and short passing).

The dataset of (Rajesh et al., 2020) consists of around 18 thousand records. This database is divided into a training and test set, in a ratio split of 80:20 to implement classification algorithms. Using this algorithm in different models, we obtained the results showed on table 16.

Table 19 – Classifier Models and Evaluation Results ((Rajesh et al., 2020).

| S. No | Algorithm | Accuracy Score | F1 Score | Jaccard Similarity |
|---|---|---|---|---|
| 1 | Naive Bayes | 0.460748 | 0.516143 | 0.610273 |
| 2 | Random Forest | 0.832546 | 0.920906 | 0.856523 |
| 3 | Decision Tree | 0.766956 | 0.811075 | 0.735357 |
| 4 | SVC | 0.753468 | 0.809675 | 0.784768 |
| 5 | Proposed prediction of Team players | 0.783258 | 0.795274 | 0.764862 |

According to (Rajesh et al., 2020), a representation using a correlation matrix approach is beneficial, using the coeficcients between the players' skills and their performance, which will help predicting how the performance of a player with a skill set influeces the position. This representation is shown on Figure 19.

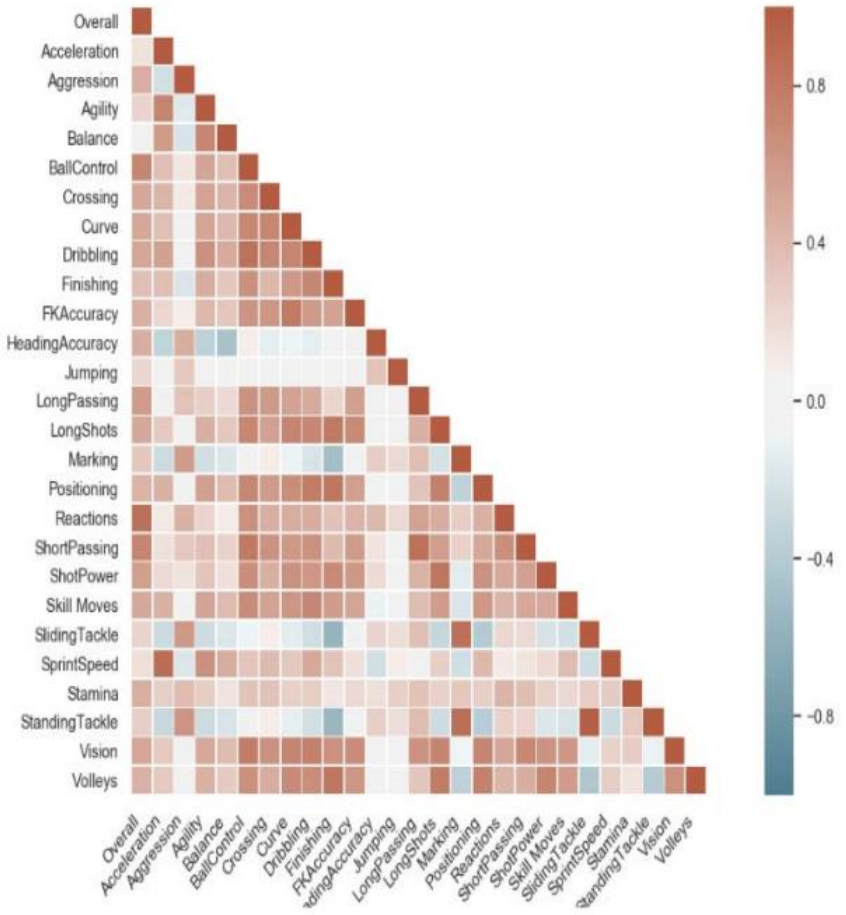


Figure 23 – Correlation between overall and performance at each skill (Rajesh et al., 2020).

From the analysis of the correlation matrix, (Rajesh et al., 2020) concluded that every variable (key influencer) is perfectly correlated with itself. Additionally, {Sprint Speed, Accelaration}, {Sliding Tackle, Marking}, {Standing Tackle, Marking}, and {Standing Tackle, Slide Tackle} have a high correlation.

After the application of the correlation matrix, along with machine learning classification algorithms, K-means clustering is implemented by (Rajesh et al., 2020). The objective of this algorithm application is to cluster player's position based on his overall performance. The method is concluded by using K as 3, performing clustering using centroids, iterative relocation technique to achieve intra cluster similarity and inter cluster dissimilarity. The results are shown on figure 20.

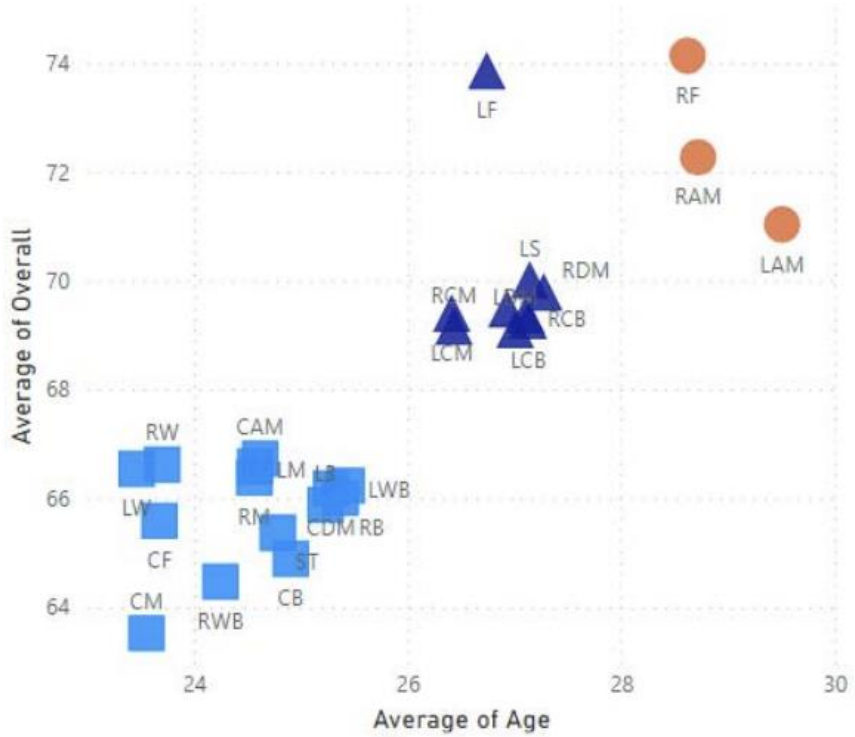


Figure 24 – Clustering of player's position based on overall performance and age (Rajesh et al., 2020).

The results of the study of (Rajesh et al., 2020) shown a reduction of 50% of players risk factors, having in mind the players performance, their rating, and a qualitative feature, that is the nationality.

On the (Arndt & Brefeld, 2016) study, multitask regression was used, with methods like multitask ridge regression, multitask support vector regression, and feature hashing. For every match day, the performance of player in previous game is encoded, and the rating for that player is denoted for that match day.

The approach of (Arndt & Brefeld, 2016) was to use the data from five different seasons (2009/10 to 2013/14) of the German *Bundesliga*, using the same database has the study previously explained, given by OPTA company. The features that describe the player activities from previous games were extracted, considering all previous games as a whole. The general statistics taken from this is that there are about 1000 players that have participated in, at least, one match, leaving 1350 single-task and 135 000 multitask features.

The base lines used for the performance approaches by (Arndt & Brefeld, 2016) were (1) predicting the average grade that is computed in the training data, and (2) use the multitask regression method.

To identify the most informative features for the prediction task (Arndt & Brefeld, 2016), in every round a single feature is removed for the single-task variant and 20 for the multi-task. Table 19 represents the most informative features, resulting on three conclusions:

1. The clubs analysed are important, meaning that playing against *Bayern München* and *Borussia Dortmund* increases the chance of getting a worst rating. On the other hand, playing for *Vfl Bochum* increases the probability of getting bad ratings.

2. Key players of their respective teams have higher chances of good rating, for example *Jefferson Farfan* and *Franck Ribery*.
3. Players' number of goals or dribbles, for example, correlates positively with higher ratings.

Table 20 – Most informative features (Arndt & Brefeld, 2016).

| | Feature | Weight |
|---|---|---|
| 1 | Next game opponent team Bayern München | 0.320 |
| 2 | Next game own team VfL Bochum | 0.314 |
| 3 | Player Jefferson Farfàn | –0.304 |
| 4 | Player Franck Ribèry | –0.303 |
| 5 | Goals | –0.298 |
| 6 | Dribbles | –0.295 |
| 7 | Shots opponent team total | 0.287 |
| 8 | Player Benedikt Höwedes | –0.281 |
| 9 | Player Roman Weidenfeller | 0.273 |
| 10 | Next game opponent team Borussia Dortmund | 0.270 |
| 11 | Player Naldo | –0.268 |
| 12 | Average rating | 0.258 |
| 13 | Player Arjen Robben | –0.251 |
| 14 | Player Marco Reus | –0.251 |
| 15 | Next game own team Borussia Dortmund | –0.244 |
| 16 | Aerials won own team total % | –0.241 |
| 17 | Player Mario Götze | –0.235 |
| 18 | Player Jaroslav Drobny | –0.233 |
| 19 | Player Martin Stranzl | –0.233 |
| 20 | Next game own team Bayern München | –0.225 |

The results obtained for the research taken by (Arndt & Brefeld, 2016) were that majority of players participated in only a few games, as shown on Figure 22, with the frequency statistics of players and games. Analysing player rates, is possible to conclude that very good and very bad ratings are not very common and that medium grades are a clear tendency, as shown on Figure 23. Figure 24 shows the frequent distribution of subsequent grades, that helps to explain the last conclusion taken, that medium grades are more common.

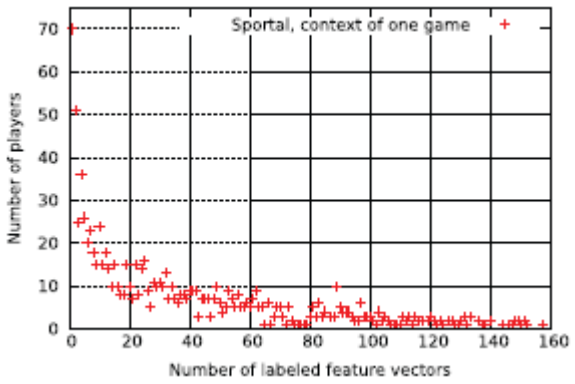


Figure 25 – Frequency distribution of games per player (Arndt & Brefeld, 2016).

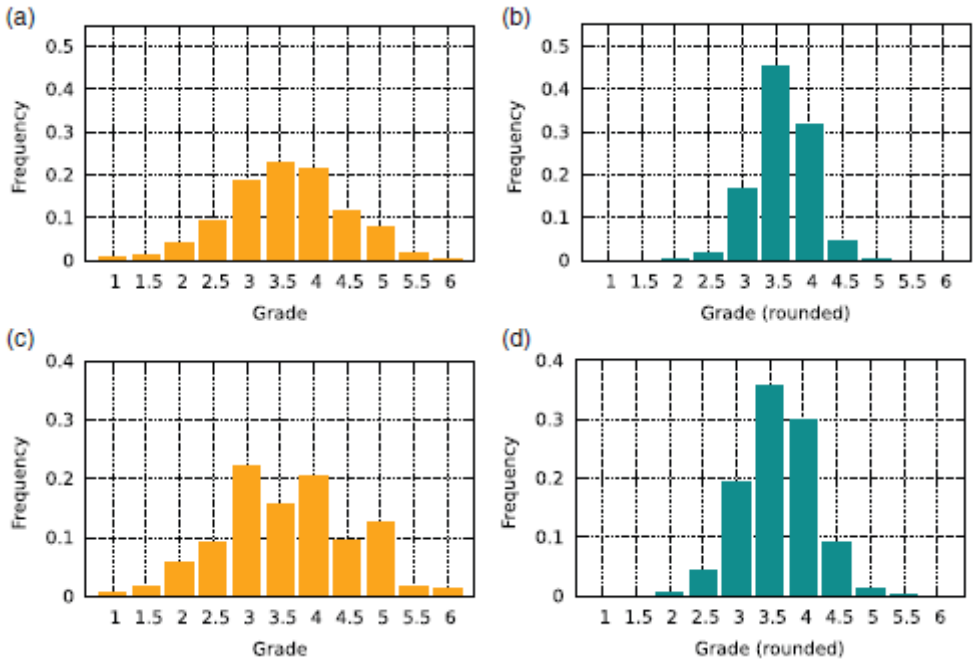


Figure 26 – Frequency distribution of actual (a, c) and predicted (b, d) grades (Arndt & Brefeld, 2016).

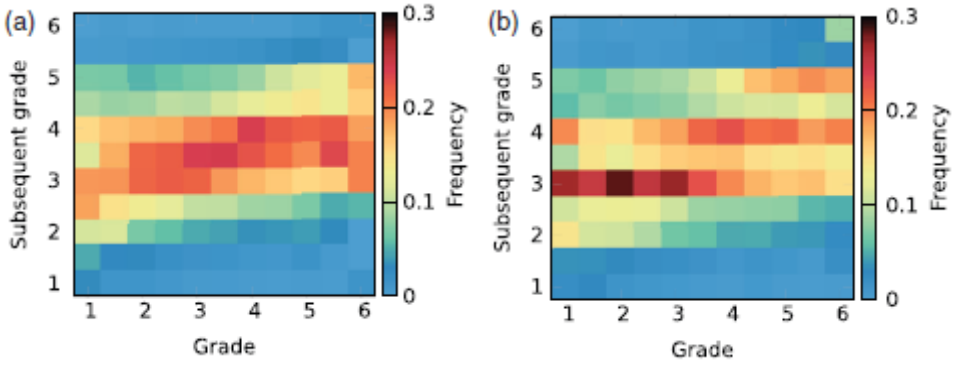


Figure 27 – Frequency distribution of subsequent grades (Arndt & Brefeld, 2016).

The dataset used on the (Baboota & Kaur, 2019)'s study, was obtained from a United Kingdom database, from 11 different seasons, from 2005 to 2016. The rating statistics used are from a FIFA database, with in the same range of seasons. On Table 21, we can see all the feature description taken from the feature-engineering phase of analysis.

Table 21 – Feature description (Baboota & Kaur, 2019).

| Feature name | Feature abbreviation | Class category |
|---|---|---|
| Home form | *HForm* | *Class A* |
| Away form | *AForm* | *Class A* |
| Home streak | *HSt* | *Class A* |
| Away streak | *ASt* | *Class A* |
| Past *k* home shots on target | *HSTKPP* | *Class A* |
| Past *k* away shots on target | *ASTKPP* | *Class A* |
| Past *k* home goals | *HGKPP* | *Class A* |
| Past *k* away goals | *AGKPP* | *Class A* |
| Past *k* home corners | *HCKPP* | *Class A* |
| Past *k* away corners | *ACKPP* | *Class A* |
| Home attack rating | *HAttack* | *Class A* |
| Away attack rating | *AAttack* | *Class A* |
| Home midfield rating | *HMidField* | *Class A* |
| Away midfield rating | *AMidField* | *Class A* |
| Home defence rating | *HDefence* | *Class A* |
| Away defence rating | *ADefense* | *Class A* |
| Home overall rating | *HOverall* | *Class A* |
| Away overall rating | *AOverall* | *Class A* |
| Home goal difference | *HTGD* | *Class A* |
| Away goal difference | *ATGD* | *Class A* |
| Home weighted streak | *HStWeighted* | *Class A* |
| Away weighted streak | *AStWeighted* | *Class A* |
| Form differential | *FormDifferential* | *Class B* |
| Streak differential | *StDifferential* | *Class B* |
| Past *k* shots on target differential | *STKPP* | *Class B* |
| Past *k* goals differential | *GKPP* | *Class B* |
| Past *k* corners differential | *CKPP* | *Class B* |
| Attack rating differential | *RelAttack* | *Class B* |
| Midfield rating differential | *RelMidField* | *Class B* |
| Defence rating differential | *RelDefense* | *Class B* |
| Overall rating differential | *RelOverall* | *Class B* |
| Goal difference differential | *GDDifferential* | *Class B* |
| Weighted streak differential | *StWeightedDifferential* | *Class B* |

On Table 21, we can see the existance of two different class categories – class A and class B, containing the individual features for both home and away team and the differential features, respectively. On (Baboota & Kaur, 2019) study, the feature set was tested with the Gaussian naïve Bayes, support vector machine, random forest and gradient boosting models.

On the (Baboota & Kaur, 2019) paper, is stated that the first nine seasons were set as training data and the remaining two seasons as test data, resulting on the accuracies shown on Figure 28 and, on Figure 29, we can see the feature importance.

The research from (Baboota & Kaur, 2019) stated that on the most influencing limitations was the fact injury information, presence of key players and psychological factors are not considered for this study. On the other hand, it was possible, through the study of the features the odds that each team add to win the game.

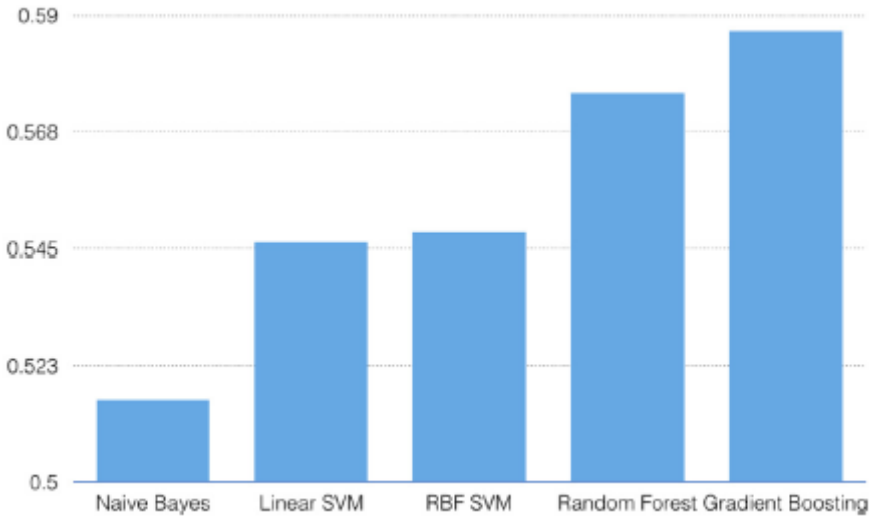


Figure 28 – Mean test accuracy scores of the different machine learning models (Baboota & Kaur, 2019).

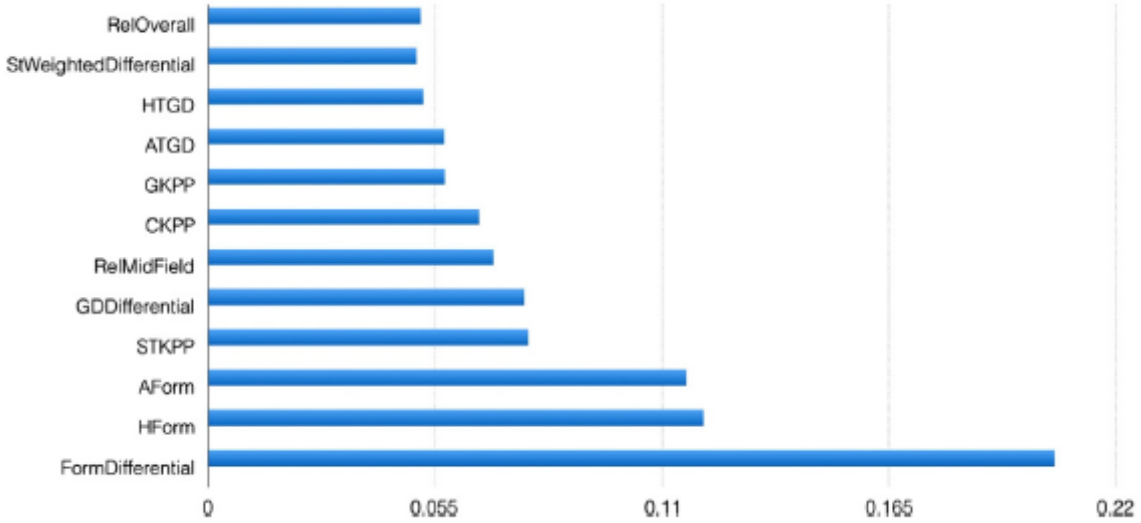


Figure 29 – Feature importance, recorded by the mean decrease in the Gini index (Baboota & Kaur, 2019).

# 5. CONCLUSIONS

## 5.1. SYNTHESIS OF DEVELOPED WORK

This paper presented a systematic literature review on the topic of the influence of the data analysis on football teams' performance, trying to find replies to the research questions, and finding the most used terminology. By the PRISMA approach to the systematic review, 653 papers were defined by the search query and, after the analysis ended with 12 papers. After this analysis, two topics arose from this literature review: (1) post-match analysis, and (2) pre-match prediction.

From all the analyzed studies, we could conclude that there are many conclusions that can be taken from these researches, in different topics, from training, passing from team selection to in-match decisions, with the use of different models and algorithms across the different studies.

To conclude, the study objectives were achieved, in the measure that, by the analysis of the studies and researches in the sections above, it is possible to state that staff are very dependent on the raw data collected previously, and during the game, conceding their conclusions to the managers and coaches to help on decision-making. Additionally, it was possible to understand which technologies and software are used to perform the data analysis, and to perform decision-making.

## 5.2. LIMITATIONS

The limitations found for this paper was the restriction of the studies' years (2012 to 2021) and the search keywords or search query itself. Additionally, a fixed number of researches databases were set.

To conclude, the principal limitation is lack of information sharing by the football clubs, keeping to themselves the tools used to perform data analysis, and specially the way those tools and analysis outputs are used to influence decision-making.

## 5.3. FUTURE WORK

For future work, the recommendation is to try to use different techniques of study and identify different critical factors to reveal defects on the analyzed studies.

Additionally, a close study near to professional football clubs to understand what tools, techniques and algorithms are used in real world, and how these models can improve their chances of winning and in-game performance of their players.

## BIBLIOGRAPHY


Costa, P. (2021). Microciclo do Analista: O trabalho do observador numa equipa de futebol (2nd ed.). Primebooks.

FIFA. (2019, September 29). FIFA. Retrieved from https://www.fifa.com/livingfootball

Liu, H., Hopkins, W., Gómez, M. A., & Molinuevo, J. S. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. International Journal of Performance Analysis in Sport, 803–821. https://doi.org/10.1080/24748668.2013.11868690

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Valenti, Maurizio, Scelles, Nicolas, Morrow, Stephen (2020). Elite sport policies and international sporting success: a panel data analysis of European women's national football team performance. In European Sport Management Quarterly (pp. 300-320). Taylor & Francis

Tianbiao, L., & Andreas, H. (2016). Apriori-based diagnostical analysis of passings in the football game. In Proceedings of 2016 IEEE International Conference on Big Data Analysis, ICBDA 2016. Institute of Electrical and Electronics Engineers Inc.

Corscadden, J., Eastman, R., Echelberger, R., Hagan, C., Kipp, C., Magnusson, E., Muller, G., Adams, S., Valeiras, J., & Scherer, W. T. (2018). Developing analytical tools to impact U.Va. football performance. 2018 Systems and Information Engineering Design Symposium, SIEDS 2018, 249–254.

Rajesh, P., Bharadwaj, Alam, M., & Tahernezhadi, M. (2020). A Data Science Approach to Football Team Player Selection. IEEE International Conference on Electro Information Technology, 2020-July, 175–183.

Spector, J. (2016). The game changer. How Michael Cox transformed the way we watch and talk about football. Nation, 303(7–8), 28–29.

Kuper, S. (2013). Match of the data - The statistical revolution comes to football. In New Statesman.

Constantinou, A., & Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. Knowledge-Based Systems, 124(February 2018), 93–104.

Beal, R., Chalkiadakis, G., Norman, T. J., & Ramchurn, S. D. (2020). Optimising game tactics for football. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2020-May(May), 141–149.

Mackenzie, R., & Cushion, C. (2013). Performance analysis in football: A critical review and implications for future research. Journal of Sports Sciences, 31(6), 639–676.

Rösch, D., Hodgson, R., Peterson, L., Graf-Baumann, T., Junge, A., Chomiak, J., & Dvorak, J. (2000). Assessment and evaluation of football performance. American Journal of Sports Medicine, 28(5 SUPPL.), 29–39.

Haneem, F., Ali, R., Kama, N., & Basri, S. (2017). Descriptive analysis and text analysis in Systematic Literature Review: A review of Master Data Management. International Conference on Research and Innovation in Information Systems, ICRIIS, 0–5.

De Silva, V., Caine, M., Skinner, J., Dogan, S., Kondoz, A., Peter, T., Axtell, E., Birnie, M., & Smith, B. (2018). Player Tracking Data Analytics as a Tool for Physical Performance Management in Football: A Case Study from Chelsea Football Club Academy. Sports, 6(4), 130.

Rice, B. X. (2014). The league tables lies. How a London football club uses data. February, 7–13.

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5(4), 213–222.

Borges, P. H., Garganta, J., Guilherme, J., de Oliveira Jaime, M., Menegassi, V. M., Rechenchosky, L., Teixeira, D.; Rinaldi, W. (2019). Tactical efficacy and offensive game processes adopted by Italian and Brazilian youth soccer players. Motriz. Revista de Educacao Fisica. https://doi.org/10.1590/s1980-6574201900020017

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2018). PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. https://doi.org/10.1145/3343172

Wu, Y., Xia, Z., Wu, T., Yi, Q., Yu, R., & Wang, J. (2020). Characteristics and optimization of core local network: Big data analysis of football matches. Chaos, Solitons and Fractals, 138.

Gomes, J., Portela, F., & Santos, M. F. (2015). Decision Support System for predicting Football Game result. Proceedings of the 19th International Conference on Computers, 348–353.

Arndt, C., & Brefeld, U. (2016). Predicting the future performance of soccer players. Statistical Analysis and Data Mining, 9(5), 373–382. https://doi.org/10.1002/sam.11321

Li, Y., Ma, R., Gonçalves, B., Gong, B., Cui, Y., & Shen, Y. (2020). Data-driven team ranking and match performance analysis in Chinese Football Super League. Chaos, Solitons and Fractals, 141. https://doi.org/10.1016/j.chaos.2020.110330

Zambom-Ferraresi, F., Iráizoz, B., & Lera-López, F. (2019). Are football managers as efficient as coaches? Performance analysis with ex ante and ex post inputs in the Premier league. Applied Economics, 51(3), 303–314. https://doi.org/10.1080/00036846.2018.1495821

Zambom-Ferraresi, F., Rios, V., & Lera-López, F. (2018). Determinants of sport performance in European football: What can we learn from the data? Decision Support Systems, 114, 18–28. https://doi.org/10.1016/j.dss.2018.08.006