MASTERARBEIT | MASTER'S THESIS

Titel | Title

Data Science in soccer: a case study of using machine learning to predict opponents' playstyles

verfasst von | submitted by

Alessandro Djogovic-Drexler BSc (WU)

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 977
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Business Analytics
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Arnold Baca

Abstract

Soccer is the most popular sport in the world and understanding of an opponents' playstyle plays a crucial role. While data science has been applied across a wide range of domains in soccer, the prediction of soccer clubs' playstyles remains a comparatively underexplored subject, as prior research focusses on individual players and performance metrics. Additionally, usually playstyles are only investigated among a single competition. This master's thesis addresses the gap in the literature by developing an end-to-end machine learning framework to identify and predict opponents' playstyles among the top 5 European leagues during the 2024/25 season. Hereby, the CRISP-DM methodology is used, along a Principal Component Analysis, unsupervised clustering with k-means and a supervised classification with XGBoost and Random Forest. Furthermore, the clustering identified four distinct and interpretable playstyles, namely Defensive & Structured, Balanced, Possession Play and Pressing/Aggressive Defending. The Random Forest model could outperform XGBoost and was able to predict playstyles on unseen data with an accuracy of 83.33%. A subsequent SHAP analysis confirmed that possession, attacking output and defensive actions are the most decisive features in distinguishing between playstyles. Moreover, a performance case study on FC Barcelona demonstrated with an OLS regression, that an opponent's playstyle does indeed have a statistically significant influence on a team's performance. Ultimately, the results of this thesis are showing that data science offers an insightful and interpretable approach to predict tactical behaviour in soccer.

Abstract (German)

Fußball ist die beliebteste Sportart der Welt und das Verständnis des gegnerischen Spielstils spielt dabei eine essenzielle Rolle. Während Data Science bereits in einer Vielzahl von Bereichen des Fußballs angewendet wurde, bleibt die Vorhersage der Spielstile von Fußballvereinen ein relativ untererforschtes Thema, da sich die bisherige Forschung hauptsächlich auf die Spieler an sich und Leistungskennzahlen konzentriert. Dazu werden Spielstile üblicherweise nur innerhalb eines einzelnen bestimmten Wettbewerbs untersucht. Diese Masterarbeit adressiert diese Forschungslücke. indem sie ein end-to-end Machine Learning Framework entwickelt, um die Spielstile von Fußballteams in den besten fünf Ligen Europas während der Saison 2024/25 zu identifizieren und vorherzusagen. Hierbei wird die CRISP-DM Methodik verwendet, mit einer Principal Component Analysis, einem unsupervised Clustering mit k-Means, sowie einer supervised Classification mithilfe von XGBoost und Random Forest. Außerdem konnte das Clustering vier voneinander unterscheidbare und interpretierbare Spielstile identifizierten, nämlich Defensive & Structured, Balanced, Possession Play und Pressing/Aggressive Defending. Das Random Forest Modell konnte besser als XGBoost performen und war dazu imstande Spielstile auf zuvor ungesehenen Daten mit einer Genauigkeit von 83.33% vorherzusagen. Eine darauffolgende SHAP Analyse konnte bestätigen, dass Ballbesitz, Offensive und defensive Aktionen die entscheidendsten Merkmale zur Unterscheidung zwischen Spielstilen sind.
Überdies zeigt eine Case Study anhand von FC Barcelona mithilfe einer OLS-Regression, dass die Spielstile von Gegnern tatsächlich einen statistisch signifikanten Einfluss auf die Performance eines Teams ausüben. Im Endeffekt zeigen die Ergebnisse dieser Masterarbeit, dass Data Science einen aufschlussreichen und interpretierbaren Ansatz zur Vorhersage von Taktik in Fußball bietet.

Acknowledgments

I extend my profound gratitude to my supervisor Univ.-Prof. Dipl.-Ing. Dr. Arnold Baca for his guidance, patience, and support throughout the development of this master's thesis. His continuous feedback was invaluable in shaping and completing this body of work. I am thankful for the opportunity to write my thesis about a topic I am truly passionate about.

My appreciation also goes to the University of Vienna for providing an environment that facilitated my academic journey.

Special thanks to my mother and my friends for their unconditional support and encouragement throughout the entire process. You have always been a source of motivation.

Thank you to everyone!

Table of Contents

List of Figures

List of Tables

1. Introduction

For many decades now, soccer has been the most popular sport in the entire world. Furthermore, coaches and teams are employing the concept of tactics to win a match, thus an understanding of the opponent's playstyle, which is dictated by a certain tactic, is crucial. However, teams often simply rely on the analysis of subject-matter experts. With the advancement of technologies and digitalization, new possibilities arise that can minimize human error and allow the subject matter to be approached more objectively. Every single aspect of a soccer match can be identified and turned into a data point. Obvious examples for this are goals, passes and shots, among others. But there are also often overlooked indicators such as turnovers, passes per defensive action and much more. As a whole, all of these different indicators characterize a playstyle. Furthermore, a playstyle of a team is essentially a pattern in the play that is repeated in specific situations of a soccer match. Some notable playstyles to be named are possession-based, pressing and positional play among others (Plakias et al., 2023).

Moreover, with the rise of Big Data, the integration of data science into sports has enabled the generation of new insights and more accurate data-driven predictions based on facts and science. A combination of these two different worlds enables the application of theoretical methods in a real-world context. In soccer, identifying an opponent's typical play pattern or their tactical tendencies is a crucial aspect of match preparation. Knowledge of the opponent's playstyle can be exploited by adjusting one's own tactics accordingly, making, this a highly relevant topic for research. Various analytical approaches can be applied to identify different teams' playstyles, most commonly through the analysis of historical event stream data, which captures on-ball actions occurring during a match, such as passes and shots (Clijmans et al., 2023).

In this context, this master's thesis investigates prior scientific research and summarize its most relevant findings in a literature review. In addition, in a case study a model to predict opponents' playstyles in soccer is established with the use of machine learning algorithms. Accordingly, the predicted playstyles will be further analysed to examine how a team's performance would vary when facing different types of opponents with differing playstyles.

Additionally, the machine learning model is based on available historical event stream data. Unsupervised learning, in particular clustering, is used to identify the different playstyles, followed by different data science methods such as Random Forest to ultimately predict a team's playstyle based on the available data.

2. Research Problem

As can be seen in research papers by Decroos & Davis (2020), who analysed the playstyle of players with match event data, or by Plakias et al. (2023), who conducted a comprehensive literature review of the identification of teams' playstyles with analytical means, there is a clear focus in the literature on the analysis of the soccer players themselves. However, using scientific means to research and predict playstyles of entire teams seems to be a rather neglected topic. There are cases of using clustering to analyse playstyles, but the approach of combining clustering algorithms and prediction algorithms can deliver new insights.

Accordingly, the case study of this master's thesis has the major aim to predict the playstyle of teams in soccer based on historical data from the top 5 leagues in Europe (England, Germany, Italy, Spain and France). Also referred to as the "Big Five" soccer leagues, they are highly regarded for their outstanding level of play (Li & Zhao, 2021). These top 5 leagues will be used to obtain a huge dataset, on which it is possible to perform analysis and gain meaningful insights. For this, a clustering model will be developed which a predictive model then uses to identify the distinct playstyles.

Beyond the identification and prediction of playstyles, further insights can be derived from the analysis of how teams perform against different opponents with their different playstyles. Immler et al. (2021) show that scoring influences the team's playing behaviour. Furthermore, the playstyle of opponents changes the use of attacking and defensive styles in a team. While the researchers focused on social network analysis for passing networks, their approaches also influence further analysis in this thesis. Due to the nature of the data used in this thesis, passing networks themselves were not applicable. However, other performance related analysis remains possible. An application is the analysis of how a team's performance would change when they face different playstyles (Immler et al., 2021).

In general, as the following literature review will show, there is a distinct lack of analysis being applied to several different soccer leagues at once. Many researchers focus on simply one competition or country. Therefore, this master's thesis aims to address this gap in the literature.

Furthermore, this master's thesis provides an extensive overview of the application of data science in soccer. While the identification of playstyles is the core of the case study, it represents just one of many areas in which data-driven methods can be applied in soccer.

2.1 Research Questions

Hence, this leads to the two main research questions of this thesis:

RQ1: *"How has Data Science been applied in different areas of soccer and what is its usage?"*

RQ2: *"Can a data-driven machine learning model reliably and accurately predict opponent's playstyles?"*

Hereby, the first research question is descriptive in nature and will be addressed through the literature review, which provides an extensive overview of how data science is applied across soccer, where playstyle identification is just one aspect. Followingly, the second research question encompasses the core aim of this thesis. It responds to the gap in the research, particularly the strong focus in the existing literature on individual players rather than teams and the lack of analysis across multiple leagues. With the combination of clustering and predictive algorithms across the domain of the top 5 European soccer leagues, it will be examined whether opponent's playstyles can be reliably and accurately predicted.

3. Literature Review

3.1 Methods

At the beginning a literature review was conducted using several keywords across several databases suited for scientific research, particularly ResearchGate, ScienceDirect, academia.edu, Springer and Google Scholar. The following keywords were used individually and as different combinations of each other:

- Sports Analytics
- Soccer Analytics
- Soccer data
- Football
- Playing Style
- Playstyle
- Tactics
- Data Science
- Machine learning
- Performance indicator

Consequently, this resulted in a large amount of available literature which fits the topic of this master's thesis, based on their keywords and titles. Hence, this amount of more than 100 different papers was further whittled down based on a multi-stage screening process. Firstly, a removal of duplicates was conducted followed by an abstract analysis to assure the relevance of the papers in line with the topic of this thesis. Specifically, papers were retained if the abstract aligned with the core focus of this thesis, even more so if they addressed the general application of data science in soccer or the analysis of playstyles. Moreover, since data science methods are becoming more popular across all sports, papers not focusing on soccer were largely excluded.

In general, the following selection criteria were applied:

- Focus on soccer as the primary researched sport
- An application of data science methods
- Relevance regarding playstyle- and performance analysis or predictive modelling techniques

Followingly, these exclusion criteria were used:

- Research paper focuses on other sports besides soccer with no possible transferability of used methods
- Lack of empirical evidence or irrelevant findings
- Insufficient analytical components

Consequently, this multi-stage screening process resulted in a final selection of 22 papers which were used in the literature review. Their research objectives and quality are in line with the chosen criteria.
In addition, one of the most often used methods to find new relevant literature was backward and forward search, which was applied to the most relevant papers. Backward search consists of screening the bibliography of research papers to identify other relevant research work, which did lay the foundation for the paper in question. In contrast, forward search tracks future citations from a research paper, meaning recent studies are built upon the findings of the selected research paper.

3.2 Literature

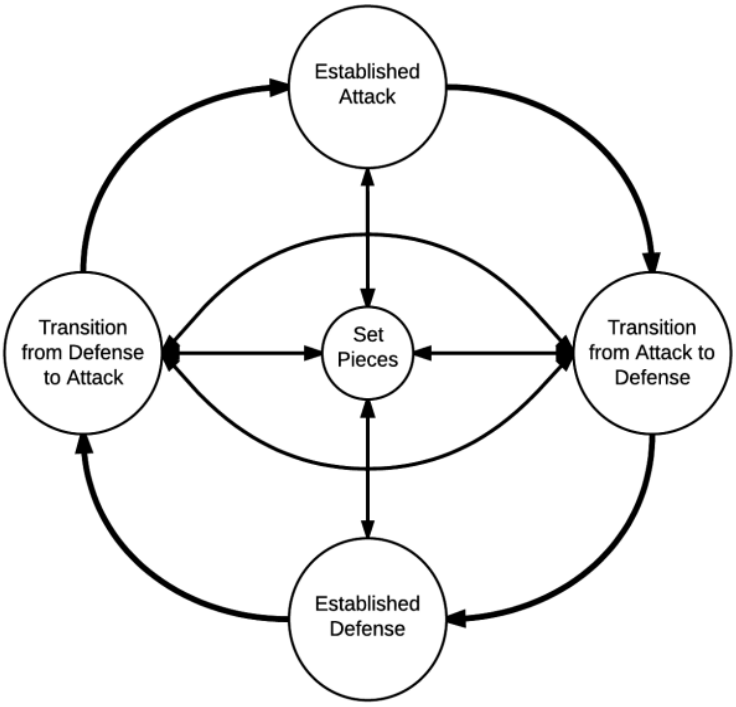
In general, the research with data science in soccer has expanded massively over the past years. The current existing literature can be conceptualized into different application areas in soccer, which will be highlighted in this literature review. Studies focusing on the identification of playstyles mainly use measurable performance indicators. This builds on other papers firstly identifying the most important performance indicators and finding new insightful ones across the data. With the advancement of AI, there is also now research about employing AI-assisted decision-making systems for tactics. Moreover, a large body of the current literature is focusing on the quantification of performance in soccer and the collection of available data. Thus, these papers also provide the foundation for analytical models which are applied in several areas like the identification of playstyles.

In the following, the structure of this literature review is organized thematically to address several different application domains of data science in soccer.
Moreover, each of these different sections focuses on one area, beginning with the fundamental concept of playstyles and their analysis (Section 3.2.1). Subsequently, specific analytical applications are looked into, like pass network analysis (Section 3.2.2), AI-assistant systems (Section 3.2.3) and how performance can be quantified as different indicators (Section 3.2.4). Practical considerations are also taken into account, namely data collection methods (Section 3.2.5) and challenges and constraints one encounters with available data (Section 3.2.6). Furthermore, there is a short overview about additional applications of data science in soccer (Section 3.2.7), followed by predictive modelling approaches (Section 3.2.8). Lastly, there is a discussion about the identified gaps in the literature which are the motivation of the presented research questions and how this thesis aims to address those gaps (Section 3.3).

3.2.1 Playstyles

One early work investigating how playstyles can be actually quantified and measured is by Hewitt et al. (2016). Notably, it is mentioned that often commentators and experts vaguely referred to teams with terms such as “Tiki-Taka” or “Total Football” without there being an actual consistent definition available. Thus, this is one of the earliest papers attempting to provide a framework which can classify specific patterns of play. In focus, they describe a playstyle as a pattern that is repeated in specific situational contexts in a game, whereas the measurement of variables remains stable for a specific playstyle.
For their framework, five different moments of play have been identified which are repeated phases during a game. Hence, these different moments are distinct periods which influence other elements in a soccer match. Patterns can be measured within these five, which enables the identification of playstyles from the different occurring play patterns. Namely these are: “Established Attack”, “Transition from Defense to Attack”, “Set Pieces”, “Transition from Attack to Defense” and “Established Defense” as evidenced in Figure 1.

Figure 1 *The five moments of play (Hewitt et al., 2016)*



Furthermore, the researchers already identified important metrics which contribute to a certain playstyle such as “Number of Passes”, “Goal Attempts”, and “Possession Level” for the moment of Established Attack.

Thus, this work did lay the foundation for several important research areas such as player recruitment depending on playstyles, training methodologies, evaluations of playstyles and more (Hewitt et al., 2016).

Peng et al. (2025) highlight in their recent research the usage of data science methods to analyse the evolution of playstyles, in particular in the UEFA European Championships. Moreover, data from 2012-2024 was used in the sample, which further consisted of various national teams, namely in total 33.

This study offers new insights, as it researches multiple different national teams with their own diverse playstyles, instead of focusing on a single domestic league. As aforementioned, there is currently only a relatively limited analysis available for international competitions.

Regarding the statistical analysis, among others, correlation matrices and multicollinearity criteria were used to improve and diminish the available indicators from the match data to solely relevant ones.
Principal component analysis (PCA) then identified five distinct playstyles in the data while accounting for a selected eigenvalue threshold. Followingly, the playstyles with their most influential factors were:

- Direct Attacking, identified mainly through total shots, key passes
- Possession Play, with total passes, pass success in percentage, possession in percentage being important factors
- Wide Play, which consists of crosses, corners, passes in the final third
- Aggressive Defending, including fouls conceded, yellow cards
- Discipline Defending, which includes tackle success in percentage with a negative loading of attempted tackles

Moreover, after controlling for factors like randomness and opponent quality, the results show a significant impact of the tournament year for the used playstyles. Meaning, there indeed has been an evolution in the development of soccer playstyles for European national teams. There has been a rise in the usage of possession-based playstyle over the years, whereas there is a noticeable decline in the occurrence for Wide Play and Direct Attacking.
Further analysis highlights that the quality of a team also influences the employed playstyles, as stronger teams are better at keeping possession and creating opportunities in a match to score goals. Overall, this paper offers conclusive evidence for the tactical evolution of soccer which can be uncovered by data science (Peng et al., 2025).

While the prior paper discussed the UEFA European Championship, Branquinho et al. (2024) evaluated the metrics of physical, technical and tactical performance in the FIFA World Cup in Qatar 2022. Their researched match data was more diverse, as this competition is played by the best national teams in the world from six qualifying soccer confederations across the globe. Consequently, there could be significant differences identified in styles of play between the different confederations (Branquinho et al., 2024).

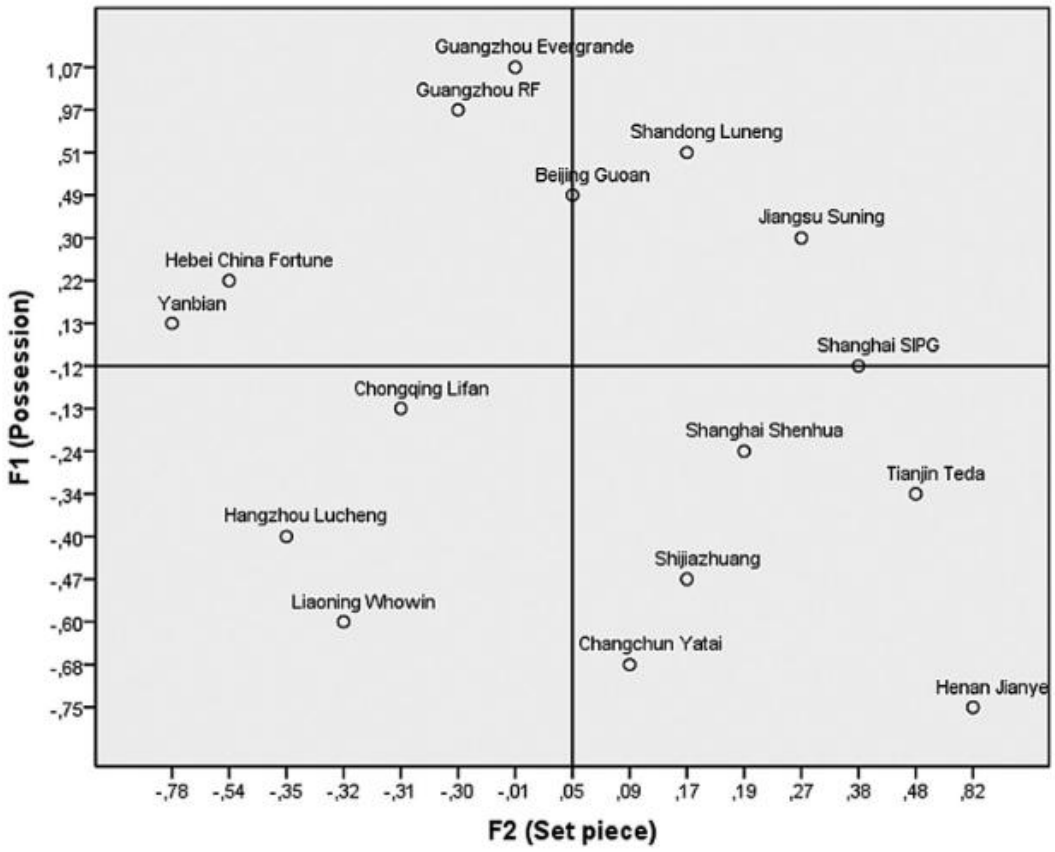
As established, playstyles play a crucial role in soccer. Research into how playstyles are related to match results has been conducted by Castellano and Pic (2019). In particular, they focused on the

Spanish LaLiga to identify different styles of play for the teams through performance indicators and to followingly associate the identified playstyles with the outcomes of matches. Sample data was taken from the match data of the 2016-2017 season of LaLiga.
Nine different performance indicators were extracted to identify offensive and defensive actions like possession, non-possession of the ball, dribbling and number of passes, among others. PCA was further applied to the mean performance indicator values to again extract the two most influential components that are able to explain about 80% of the variance and could assign one of four different playstyles to the teams as their predominant one, consisting of direct attack, elaborate attack, deep defending and high-pressure defending.
Finally, associating the different playstyles with the results of a match showed that there are significant differences between these four styles of play. Teams using elaborate attacking and deep defending won more games than expected, whereas direct attack paired with high-pressure defending ended up with more losses. A team using their preferred playstyle could also be linked with the result of a match.
Furthermore, it is worth noting that this study derived just two component factors, from which four different playstyles were proposed. This stands in contrast to previous works where some studies even used up to 62 variables and identified a considerably larger number of distinct playstyles. Even more, the first two components in those previous works captured considerably less variance than in this study (Castellano and Pic, 2019).

In a comparable approach, Lago-Peñas et al. (2017) also analysed match data, using the data from 240 matches from the 2016 Chinese Soccer Super League season to identify and measure distinct playstyles. In their used approach, the selected variables were drawn from prior relevant literature. In total, 20 different variables were chosen, among them different ones for ball possession, attacking, defending, passing, set pieces and more.
PCA was yet again applied in this research study on the available factors to reduce the dimensionality, resulting in five extracted factors with an eigenvalue above the value of 1. Moreover, they explain about 79.6% of the total variance. Followingly, the results are showing that it is possible to describe and measure the playstyles of different teams across an entire season. Figure 2 showcases the allocation of the different teams across the dimensions of the first PCA factor “Possession” and the second one, “Set piece”. Hence, identified playstyles focused on the different tactics of possession, set piece, counterattack and transitional play. The findings of Lago-Peñas et al.

(2017) again support the insight, that the top teams use a possession-focused playstyle as they prefer to control the game instead of letting their opponent have the ball and dictate the game. Correlating performance indicators also support their thesis that teams favouring possession are indeed technically superior to the teams they are facing as opponents (Lago-Peñas et al., 2017).

Figure 2 *The playstyles according to Factor 1 (Possession) and Factor 2 (Set piece) (Lago-Peñas et al., 2017)*



3.2.2 The analysis of passes in soccer

One other interesting application of data science in soccer is the usage for passing network analysis. For this Immler et al. (2021) researched passing data in the UEFA Champions League from three highly regarded coaches, namely Jürgen Klopp, Pep Guardiola and Mauricio Pochettino. Ultimately, the goal was to identify the passing dynamics these coaches use and how they differ between them.

For the analysis of the data, Social Network Analysis was applied, thus for each match the 11 most used players represented nodes, with passes represented as edges. Moreover, the resulting graph could be used to calculate different metrics like the density, average shortest-path length, mean centrality, centrality dispersion, local clustering coefficient and the largest eigenvalue.
To compare the resulting networks, statistical analysis was applied including ANOVA, which could identify significant differences across the three coaches in focus. Pep Guardiola's passing network demonstrated a greater overall volume of passes and a stronger reliance on key players compared to the networks of both Klopp and Pochettino.
PCA and k-means clustering was further used as analytical tools which could indeed identify that there are noticeable differences visible as clusters whereas Guardiola again differed the most from the others. This can directly be interpreted as visible differentiations in the coaching styles like a higher focus on possession from Guardiola. Furthermore, this paper showcases how data science can be used to characterize tactical behaviour from different elite coaches and their used playstyle systems (Immler et al., 2021).

Another interesting aspect of researching passes, is the usage of predictive analytics to make predictions for which soccer player will receive a pass. Sanyal (2021) developed a data-driven framework that can predict which teammate will be the first one to receive a pass with the data taken directly from a soccer broadcast. For that, a combination of two dependent models was created. An Opponent Proximity Model (OPM) is used to estimate the likelihood of an opponent intercepting a pass, which is achieved through using a triangle as a region of attack, consisting of the player, a possible nearest pass recipient and an opposing team player who can intercept the pass along an arrow line.
The Pass Region Model (PRM) is another model which defines a region that favours passes, which are going through areas where there is a high concentration of one's own teammates. OPM and PRM are accordingly weighted with a joint decision probability distribution to build a framework that is able to effectively capture both aspects of defensive pressure and offensive positioning.
Furthermore, for validation of this model a collection of 5000 frames from selected soccer video sequences, sourced from the Spanish LaLiga and English Premier League, was used. Hence, the algorithm used the coordinates of all players and the ball from the current video frame as its input and applied this through the framework to reach a joint decision. Accordingly, the result is the return of the ID of the player with the highest decision score as the predicted pass recipient.

Moreover, results show that the framework can attain approximately an accuracy of 73% in predicting the correct pass recipients in practice.
This framework outperforms two other recently published models. A possible explanation is that this combination approach incorporates regional information too. Moreover, an ablation study by the researchers confirms that an individual use of OPM and PRM does indeed perform worse than the proposed combined framework. Followingly, it is proposed to use this model to validate ball possession statistics during soccer games (Sanyal, 2021).

3.2.3 Developing AI assistants

There have been many publications about the analysis and theoretical usage of data science in the realm of soccer. However, one recent real-world application is an AI-driven tactical assistant that has been built by researchers in close collaboration with a world class soccer club, namely Liverpool FC. Identifying tactics used by other teams including the adequate response to the team is a crucial aspect for coaches.
Wang et al. (2024) built TacticAI, an AI soccer tactics assistant, and applied it on corner kicks, as these are a situation where coaches can directly intervene. Additionally, corner kicks were chosen for this AI tool, as they are frequently occurring situations whereas teams can use pre-planned methods for better opportunities to score a goal.
TacticAI uses raw, spatio-temporal player tracking data, which is translating the corner kicks into graphs, whereas the players represent nodes, players heights, weights, velocities, and positions among others become features. Geometric deep learning is processing these graphs and TacticAI is hence able to make predictions, like which player is the most likely to first reach the ball. Moreover, it can also act as a generative recommendation system, which suggests the coach changes to the position of a player to maximize the outcome of a shot taking place.
Although they achieved significant results with benchmarks in theory, validation by domain experts was necessary. Five soccer experts from Liverpool FC helped in evaluating the utility of TacticAI. Ultimately, the experts were unable to distinguish proposed AI adjustments to corners from real ones and in even 90% of the test cases the experts approved of the AI suggestions to improve a corner kick setup.
In conclusion, the researchers were able to demonstrate an AI assistant, with not only statistical evidence in theory, but also with an approval of domain experts in practice. This achievement

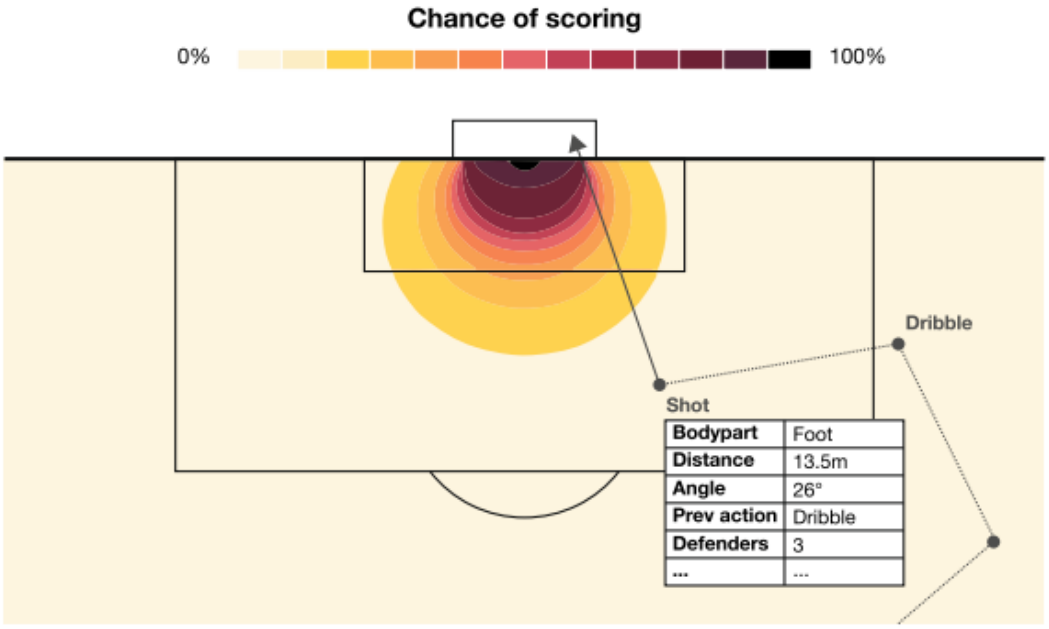
stands out by using generative modelling with a geometric deep learning platform for corner kicks and their tactical adjustment, whereas prior literature mostly focuses on other aspects of soccer like passes and shots (Wang et al., 2024).

3.2.4 Quantifying performance

An often-encountered implementation of data science in sports analytics is the quantification of performances. Hence, it is necessary for a machine learning model to be trained on large amounts of data, meaning lots of prior matches. In soccer, one application area is the investigation of the relationship between the number of shots taken and the resulting number of goals. It is possible to identify a correlation and to quantify areas on the pitch with their corresponding likelihood of a goal being scored from that place (StatsPerform, n.d.).

A binary class identifier uses features like the passage of play, the distance to the goal and the angle among others. This performance indicator is called expected goals (xG). Furthermore, this indicator is a more consistent type of measure for performances than the actual number of goals occurring in a match. Figure 3 shows how the indicator xG works with the respective chance of scoring using a heatmap on a soccer pitch (Davis et al., 2024).

Figure 3 *Expected goals (xG) is estimating the likelihood of a shot resulting in a goal, with the colors indicating a higher chance to score the closer a player is to the goal (Davis et al., 2024***)**



In general, performance indicators need to meet several different conditions. Followingly, they need to offer new insights that have not been previously available. In the case of xG this is clear, as this indicator offers insight into how often a player gets into shooting positions with a favourable chance to score a goal. Also, the performance indicator needs to be based on domain knowledge, meaning not only data science experts should be able to understand the indicator. Lastly, a performance indicator needs to be trusted, for that it needs to be robust and offer clear context where its application is useful (Davis et al., 2024).

3.2.5 Data collection

Broadly speaking, as evidenced by Davis et al. (2024), there are four different types of collected data in sports: physical data, match-sheet data, play-by-play data and optical tracking data.
In soccer, GNSS trackers with measuring technology are usually worn by the players. Moreover, these trackers are able to trace the covered distance, amounts of high-speed sprints, accelerations and much more. Paired with questionnaire data, they are also used to monitor and enhance the fitness levels and availability of players.

The category of the match-sheet data encompasses aggregated statistics of teams and players from the individually played matches. Therefore, different statistics like the number of passes, shots and assists are available, among others.
In contrast, play-by-play data is tracking actions related to the soccer ball, such as the start and end of actions, their location, their type and their result.
Lastly, optical tracking data can trace and report the location of all players on the field, including the ball, several times every single second. However, a setup like this is usually only used in top soccer leagues, due to their associated costs (Davis et al., 2024).

These performance indicators are also used in player recruitment, as it is possible to get a comprehensive overview of a player's profile. If the explicit data isn't available, they can also be estimated through machine learning models (Kovalchik, 2023).

Looking at the examples of the indicator of progressive passes, a pass that brings the ball closer to the opponent's goal, it is used to identify the players whose passes are disproportionately driving the ball forward. Moreover, this understanding of a player's profile can be used to understand if a player's skill is suited to the tactics of a team. Data science methods at the highest level of soccer are employed to discover patterns and to evaluate the efficacy of a tactic (Davis et al., 2024).

3.2.6 Challenges and constraints with data

One important factor to keep in mind is that due to the inherent nature of sports data there are many challenges for the analysis of it. Admittedly, the definite tactics employed by a soccer team are not known to third parties. Hence, only educated guesses are possible on a surface level view. Additionally, there is no definite variable which contributes to and can be assigned as a player's ratings in a soccer match. Meaning, assigned ratings from an analytical standpoint, cannot be truly objective ones. Moreover, the reasons for outcomes in a match are not always clear. The reason for a failed pass could be because of the mistake of the passer or the recipient. Likewise, a player with a high number of tackles in a match could be forced to do that, because they cover for another teammate. In conclusion, there is a lack of ground truth.
Additionally, one must account for noisy features. Especially regarding subjective data like questionnaires related to player health, the received data could be simply false. A reason is that players

could be motivated to make false statements as they have a reason to do so like wanting to play more games (Davis et al., 2024).

Also, play-by-play data can be faulty due to human error from the annotators. A bias can occur when the scorekeepers are employed by the teams themselves. For example, a cross accidentally resulting in a goal can be falsely labelled as a shot (Van Bommel & Bornn, 2017).

Small sample sizes can also pose a common issue, as usually a soccer team only plays a certain limited number of games in their own league resulting in 1500-3000 on-the-ball actions in a single game. Moreover, soccer players do not play every game in their league season. This issue is amplified by the fact that teams can significantly change their playstyles over seasons, while also player performance can widely differ. Tactics which highly influence the games also do change and evolve constantly (Davis et al., 2024).

Consequently, there can also be undetected bias occurring in the data itself. Better performing teams could simply have higher possession than others because of their ability to retain the ball for longer periods of time. This can become an issue when training models with data based on that, leading to the over- or undervaluation of some actions (StatsBomb, 2021).

3.2.7 Other applications of data science methods

While the prior highlighted literature mainly involved the discussion of playstyles, passing, tracking and performance indicators, data science methods are used for a wide range of data-driven insights in soccer nowadays.
A different use case is the research of Fernández-Cortés et al. (2024) who analysed if the crowd support in a soccer match at home, does offer an advantage to a team. Descriptive studies that have been analysed by them, do indeed confirm that there is a decrease in the home advantage occurring without fans. Even referees' decisions can be influenced by the crowd. However, it is noted that only conclusive evidence for professional men's leagues is available, while it is not for youth and amateur settings (Fernández-Cortés et al., 2024).

Another researched factor is the influence of playstyles on the match performance. In particular, Forcher et al. (2023) investigated the effect of an offensive playstyle on the physical and technical

match performance. They used the 2020/21 season from the German Bundesliga as their dataset for this analysis. PCA was applied while building on previous researchers' works. Followingly, this resulted in a playing-style-coefficient (PSC), with a high PSC indicating a focus on possession, whereas a lower PSC is focusing on counter-attacking (Forcher et al., 2023). In general, prior research has linked the influence of offensive playstyles with the ball possession in a soccer match. This relationship between these two factors is supported by the findings of Yi et al. (2019) who observed a significant positive relationship between them.

The application of a linear mixed model revealed that possession-oriented teams do indeed experience a higher physical load across soccer matches, like recording more accelerations ($p \leq 0.01$; $R^2 = 0.69$). Teams with a ball possession focused playstyle also play more horizontal passes and have a higher passing success rate.
Consequently, counter-attacking teams showed more sprinting distances per second in possession ($p \leq 0.01$; $R^2 = 0.14$), due to rapid transitions. So, offensive playstyles do indeed influence the physical and technical performance, whereas only a limited influence on the success of a team could be found, which was analysed based on the influence of the PSC on xG ($p < 0.01$), goals ($p = 0.14$) and points ($p = 0.36$) (Forcher et al., 2023).

3.2.8 Predictive Modelling in Soccer

In 2017 the Soccer Prediction Challenge was held to evaluate different modeling approaches for the prediction of soccer match outcomes. The provided dataset consisted of over 200,000 matches from the past 17 years, while also omitting relevant features such as the player names. Hubáček et al. (2019), who were the winners of the competition, constructed several new features for this, such as historical strength. In this quite extensive process, they added 66 new features in total.
For the prediction, they experimented with both relational and feature-based methods. Consequently, the relational model using the RDN-Boost algorithm performed worse than the feature-based approach using Gradient Boosted Trees. In particular, XGBoost was used which turned out to be the best performing model. In about 30% of the matches a draw was predicted, which is the least likely outcome. This is in line with the prediction of bookmaker odds, which do predict draws in about 25% of the matches. While this research work mainly focuses on the prediction of

match results, their extensive feature engineering process highlights its importance, which proves to be an interesting insight (Hubáček et al., 2019).

Mills et al. (2024) outlined a framework that compares the performance of different machine learning algorithms and deep learning models, which are used for predictive modelling use cases in soccer. Moreover, feature engineering is applied, as is the framework's application across three different international leagues to prove the robustness of their results. In particular, the Dutch Eredivisie, the Belgian Jupiler Pro League and the Scottish Premiership. This wider and more diverse dataset offers better generalizability across the results.
Another new addition are real-time features like half-time results and goals compared to previous studies, which usually solely use end-of-game features. In total, seven different models were tested, including Logistic Regression, XGBoost, Random Forest, SVM, Naive Bayes, Feedforward Neural Networks and Vanilla Recurrent Neural Networks.
Furthermore, the researchers addressed two different prediction tasks, the prediction of match outcomes (win at home, draw, win away) and the prediction whether the total goal count would be over or under 2.5. Extensive hyperparameter tuning with grid search, but also Bayesian optimization techniques were employed to find the best model configurations. Support Vector Machine Synthetic Minority Oversampling Technique (SVM-SMOTE), Near-Miss and Random-OverSampling were also used to address class imbalances in the dataset.
Ultimately, Feedforward Neural Networks had the highest accuracy in predicting match results while also being consistent across the three different leagues. For the number of goals to be predicted, Logistic Regression performed the best. However, the usage of a voting model combining Random Forest and XGBoost achieved an even higher accuracy than the individual models with 0.83. Accuracy was just one metric to compare the performance of the models, as F1-Score, Precision and Recall were also used.
In their conclusion, the researchers further point out that there is a lack of analysis across different leagues with different playstyles in the literature which is an area for future research. This further underlines the importance of robustness and generalizability in results which is a core theme of this master's thesis (Mills et al., 2024).

3.3 Identified gaps in the literature

The literature review of this thesis lays out the current status of data science applications in soccer. Furthermore, several notable gaps could be identified across different papers.
As can be seen by Decroos and Davis (2020), there is an abundance of literature available focusing on individual player analysis with match event data, whereas Plakias et al. (2023) presents a comprehensive literature review on a small selection of soccer team's playstyles. In the area of predictions using machine learning, there is numerous literature for the prediction of match results like Mills et al. (2024) available. Consequently, there is a distinct lack of research utilizing the capabilities of data science to predict playstyles of entire teams in a robust setting. Moreover, clustering algorithms to analyse playstyles have been employed previously (Castellano & Pic, 2019).

In addition, one can find a clear geographical limitation in the available research works. Often, the researchers focus on just a single competition in a country, which can be seen in cases from Lago-Peñas et al. (2017) with a focus on the Chinese Soccer Super League, or Castellano and Pic (2019), with just LaLiga in Spain.
This point is further emphasized by Mills et al. (2024), as even though they investigated three different leagues across Europe, in Belgium, the Netherlands and Scotland namely, they still bring up the point that future research work is needed which focuses on different leagues with more differing playstyles. Moreover, there is no research focusing on the top 5 ranked European soccer leagues, which offer the highest level of play in different regions with their own history and playstyles.

Accordingly, the current body of work is insufficient in the examination of different playstyles in soccer. This thesis addresses the mentioned gaps in the literature by incorporating the top 5 European soccer leagues, namely Germany (Bundesliga), England (Premier League), Italy (Serie A), Spain (LaLiga) and France (Ligue 1), which provides a more diverse and comprehensive dataset which is suited for this new methodological approach in analysis.

Another neglected aspect is that the relationship between performance outcomes and different playstyles is not explored. Immler et al. (2021) found evidence that scoring influences playing

behaviour and that tactics can change depending on the opponents' playstyles. However, their research focused solely on three elite coaches and their teams. Another example is that Forcher et al. (2023) investigated the effect of an offensive playstyle on the physical and technical match performance. Thus, there is also an identifiable gap for how soccer teams' performances changes when they face different playstyles.

Playstyle identification with the help of unsupervised machine learning algorithms is also a topic which can be further researched. Past research work used for example PCA to extract playstyles, but there is a lack of combination of clustering with predictive modelling. Hence, this thesis uses a clustering algorithm like k-means to identify playstyle clusters, with the addition of supervised learning algorithms such as Random Forest or XGBoost. This combination of clustering and predictive algorithms offers an unused and novel methodological approach to this subject matter.

4. Methodology

The methodology of this master's thesis follows the CRISP-DM methodology for conducting data science analysis. It is a standardized data mining framework providing a systematic approach to ensure well-defined results. It consists of various phases which will be adhered to in the following:

•Business Understanding
understanding the objective and the requirements
•Data Understanding
collecting the data and conducting first exploratory data analysis
•Data Preparation
cleaning the data and applying transformations
•Modelling
selectin the appropriate models for the clustering and the prediction
•Evaluation
evaluating the results
•Deployment
not applicable to this master's thesis

Overall, Visual Studio Code will be used as the IDE (Integrated Development Environment) of choice with Python 3.10.11 and different extensions like pandas, numpy, scikit-learn to analyse the data and apply the modelling steps. Additionally, visualizations will also be conducted with Python based on the matplotlib extension.

4.1 Business Understanding

The goal of this master's thesis is to research whether it is possible to accurately predict opponents' playstyles in soccer with the help of data science methods. Following the aforementioned CRISP-DM framework, this phase translates the overarching goal into actionable requirements which lay the foundation for the subsequent analysis.

Moreover, the first needed step is to research and identify a suitable dataset that provides reliable performance data for the top 5 European leagues. This is the basis for later modelling tasks. Firstly, an unsupervised clustering applied on the selected dataset is needed to identify distinct playstyles. Hence, this aggregated mass of data gets labelled by the clustering. Furthermore, based on the established playstyle clusters, a supervised prediction of the playstyles is conducted. Then, the results of the chosen models are compared with each other to determine the best overall model given the use case.

Beyond the clustering and the prediction, a performance case study will be conducted for a specific team. Hereby, it is examined how the performance changes when facing opponents of different playstyles.
For the clustering, the degree of separation between the playstyles is investigated, whereas for the predictions established metrics like accuracy will be used. Overall, it is important to establish clear criteria how the final results can then be evaluated and compared with one another. Consequently, these criteria determine whether the research questions of this thesis can be answered.

4.2 Data Understanding and collection

Sourcing the data from the appropriate sources is a crucial step in establishing a sound foundation for this master's thesis. In total the data was collected from three distinct selected sources for the 2024/25 soccer season.

Firstly, FBref.com was used, which is a website that provides a wide variety of soccer stats with public access. However, due to a change to the database in January 2026, where the company Opta announced that it terminated its data agreement with FBref.com, advanced stats that have been publicly available for free, were removed from public access (Twomey, 2026).

This change resulted in the research for other suitable soccer databases which can supplement the very basic data that is now just provided by FBref.com. A thorough selection process resulted in the inclusion of two other databases which provide advanced and reliable data across the top 5 European soccer leagues, namely the Premier League (England), Bundesliga (Germany), LaLiga

(Spain), Serie A (Italy) and Ligue 1 (France). The top 5 leagues were selected for this thesis, as they not only embody the pinnacle of club soccer, but also represent a more diverse dataset with different tactics and playstyles, which also addresses gaps in prior research. Additionally, as these are also the most popular leagues, there are more detailed and comprehensive datasets available.

Consequently, the second chosen database was StatsHub (statshub.com), which is a comprehensive sports statistics platform that also provides advanced statistics for soccer. Lastly, the data is supplemented by the small online database markstats.club. This addition further provides some more advanced metrics that can prove crucial for the data science methods to be used. An overview of the different data sources can be found in Table 1.

Table 1 *Data collection overview*

Source	**Feature Columns**	**Description**
StatsHub	69	Advanced metrics across all categories
Markstats	17	Advanced metrics such as field tilt or defensive line
FBref	18	General match data such as goals, clean sheets, penalties
Total	154 (including duplicates)	

Depending on the league, there are 18 or 20 teams with respectively 34 or 38 matches played across a season.

As the data was not available for download in all instances, web scraping with Python was used to scrape the individual tables and access the data. However, as the access of the scraper was blocked by both FBref and StatsHub, this could only be applied to Markstats. Followingly, in the two other instances a manual extraction of the data tables to Excel was applied with an aggregation done within the same tool. In the following, Figures 4, 5 and 6 will show an example of the available statistics from each database, respectively.

Figure 4 *Example of available statistics from FBref.com* **(FBref, n.d.)**

Squad Standard Stats 2024-2025 Premier League View Player Stats Glossary Toggle Per90 Stats

Squad Stats | Opponent Stats

				Playing Time				Performance								Per 90 Minutes				
Squad	# Pl	Age	Poss	MP	Starts	Min	90s	Gls	Ast	G+A	G-PK	PK	PKatt	CrdY	CrdR	Gls	Ast	G+A	G-PK	G+A-PK
Arsenal	25	25.8	56.9	38	418	3,420	38.0	67	55	122	65	2	2	70	6	1.76	1.45	3.21	1.71	3.16
Aston Villa	28	27.0	50.5	38	418	3,420	38.0	56	45	101	53	3	6	76	4	1.47	1.18	2.66	1.39	2.58
Bournemouth	29	25.1	48.5	38	418	3,420	38.0	57	41	98	51	6	7	97	3	1.50	1.08	2.58	1.34	2.42
Brentford	28	25.8	47.9	38	418	3,420	38.0	65	44	109	60	5	6	62	1	1.71	1.16	2.87	1.58	2.74
Brighton	32	24.8	52.3	38	418	3,420	38.0	64	41	105	57	7	7	78	3	1.68	1.08	2.76	1.50	2.58
Chelsea	29	23.7	57.1	38	418	3,420	38.0	61	47	108	57	4	5	101	2	1.61	1.24	2.84	1.50	2.74
Crystal Palace	29	26.2	42.8	38	418	3,420	38.0	49	38	87	46	3	4	80	4	1.29	1.00	2.29	1.21	2.21
Everton	26	28.0	40.9	38	418	3,420	38.0	39	27	66	37	2	2	82	2	1.03	0.71	1.74	0.97	1.68
Fulham	26	28.1	52.3	38	418	3,420	38.0	53	44	97	50	3	4	80	2	1.39	1.16	2.55	1.32	2.47
Ipswich Town	32	25.8	40.6	38	418	3,420	38.0	35	26	61	33	2	2	93	5	0.92	0.68	1.61	0.87	1.55
Leicester City	32	26.5	45.4	38	418	3,420	38.0	33	25	58	31	2	3	87	0	0.87	0.66	1.53	0.82	1.47
Liverpool	24	27.2	57.7	38	418	3,420	38.0	85	65	150	76	9	9	67	3	2.24	1.71	3.95	2.00	3.71
Manchester City	30	26.8	61.3	38	418	3,420	38.0	71	51	122	68	3	4	59	2	1.87	1.34	3.21	1.79	3.13
Manchester Utd	31	25.5	53.5	38	418	3,420	38.0	42	29	71	38	4	4	86	3	1.11	0.76	1.87	1.00	1.76
Newcastle United	24	27.4	51.3	38	418	3,420	38.0	66	50	116	61	5	6	68	1	1.74	1.32	3.05	1.61	2.92
Nottingham Forest	23	26.1	41.2	38	418	3,420	38.0	57	42	99	54	3	3	90	2	1.50	1.11	2.61	1.42	2.53
Southampton	36	25.2	48.5	38	418	3,420	38.0	25	16	41	25	0	2	89	3	0.66	0.42	1.08	0.66	1.08
Tottenham Hotspur	31	25.0	54.7	38	418	3,420	38.0	61	47	108	58	3	4	72	1	1.61	1.24	2.84	1.53	2.76
West Ham United	27	28.1	48.4	38	418	3,420	38.0	43	29	72	40	3	3	82	3	1.13	0.76	1.89	1.05	1.82
Wolves	32	26.9	48.1	38	418	3,420	38.0	53	42	95	53	0	0	79	2	1.39	1.11	2.50	1.39	2.50

Figure 5 *Example of available statistics from statshub.com (StatsHub, n.d.)*

Figure 6 *Example of available statistics from markstats.club (markstats, n.d.)*

TEAM	XG	XGA	XGDIFF	POS	FTILT	XT	XTA	XTDIFF	DLINE	ODLINE	BUILD	OBUILD
Liverpool	2.22	1.02	1.19	57.8	61.34	1.79	1	0.79	48.35	41.14	86.7	81.0
Arsenal	1.81	0.91	0.89	57.0	65.07	1.65	1.03	0.62	51.34	39.96	87.0	80.1
Man City	1.79	1.35	0.44	61.4	71.55	1.70	1.08	0.62	51.24	39.17	90.5	82.6
Bournemouth	1.77	1.34	0.43	48.5	52.13	1.47	1.45	0.02	47.64	43.54	78.2	78.5
Chelsea	1.71	1.36	0.35	57.0	60.75	1.51	1.14	0.37	47.58	44.52	86.6	81.8

4.3 Data Preparation

Due to the nature of the data, some general metrics like the number of matches played or the number of goals appear in multiple data sources. In the applied data cleaning process, duplicated and redundant features were excluded while merging all three different data sources. Further analysis of the data showed that there are no missing values across the dataset from FBref and markstats. Thus, there is no need to handle missing data or impute replacements.
However, regarding the source StatsHub, a few metrics are missing for various teams, which led to the total exclusion of the affected feature column, as an insertion of a replacement is unsuitable. Additionally, some variables need to be transformed due to the differing number of matches in the leagues. This ensures consistency for the metrics across the different leagues. Notably, no new features were engineered from the data.
In summary, all three data sources provide together a comprehensive dataset that is suitable for data analysis.

4.4 Feature Selection

In the following the process of the feature selection will be applied. As there are currently 154 different columns available, using all of them would risk overfitting. A result could still be achieved, but it would be tailored solely to this specific dataset and not applicable in general at all.

In the first step it was decided to remove all the data from the stats of “Squad Goalkeeping” and “Squad Playing Time” from the source FBref, as they are rather suited for individual analysis. This leads to the deletion of 31 columns. As aforementioned, one column provided by StatsHub has missing entries for some teams, which leads to its exclusion. Furthermore, domain knowledge was used to exclude other redundant features to achieve a drastically reduced dataset. Hence, this resulted in an initial raw dataset consisting of 64 feature columns, whereas four of them are identifiers. A clearer overview of the data structure is visible in Table 2.

Table 2 *Data overview after manual selection*

Number of teams	96
Number of columns	64
Total data points	6144

Furthermore, Figure 7 shows a full correlation matrix of the aforementioned 64 columns which have been selected.

Pearson's correlation matrix is a square symmetrical matrix which shows the pairwise correlation between different features. Moreover, the correlation denotes the linear association between two variables, indicating how strongly the movement of one variable is correlated to the other one. The coefficient ranges from -1 up to +1. Hence, the closer the value is to 0, the weaker the association and the less correlation there is. However, if the variables are at +1 or -1, they have a perfect positive or negative linear relationship. Subsequently, scores close to a perfect relationship are highly correlated too and can be candidates for feature reduction (Laerd Statistics, n.d.).

Hereby, the logic of dropping essentially all correlated features with a score over 0.95 was applied. An example output can be seen in Figure 8. Moreover, there are some exceptions like goals and assists, which do have a very high correlation score ($r = 0.97$), but they encompass different concepts in soccer. Additionally, domain knowledge was applied to further reduce some redundant features which are also highly correlated ($r > 0.85$ for all of them). Consequently, this feature reduction ultimately resulted in 39 (four of which act as identifiers), which are used in the following PCA.

Figure 7 *Pearson's correlation as a heatmap matrix*

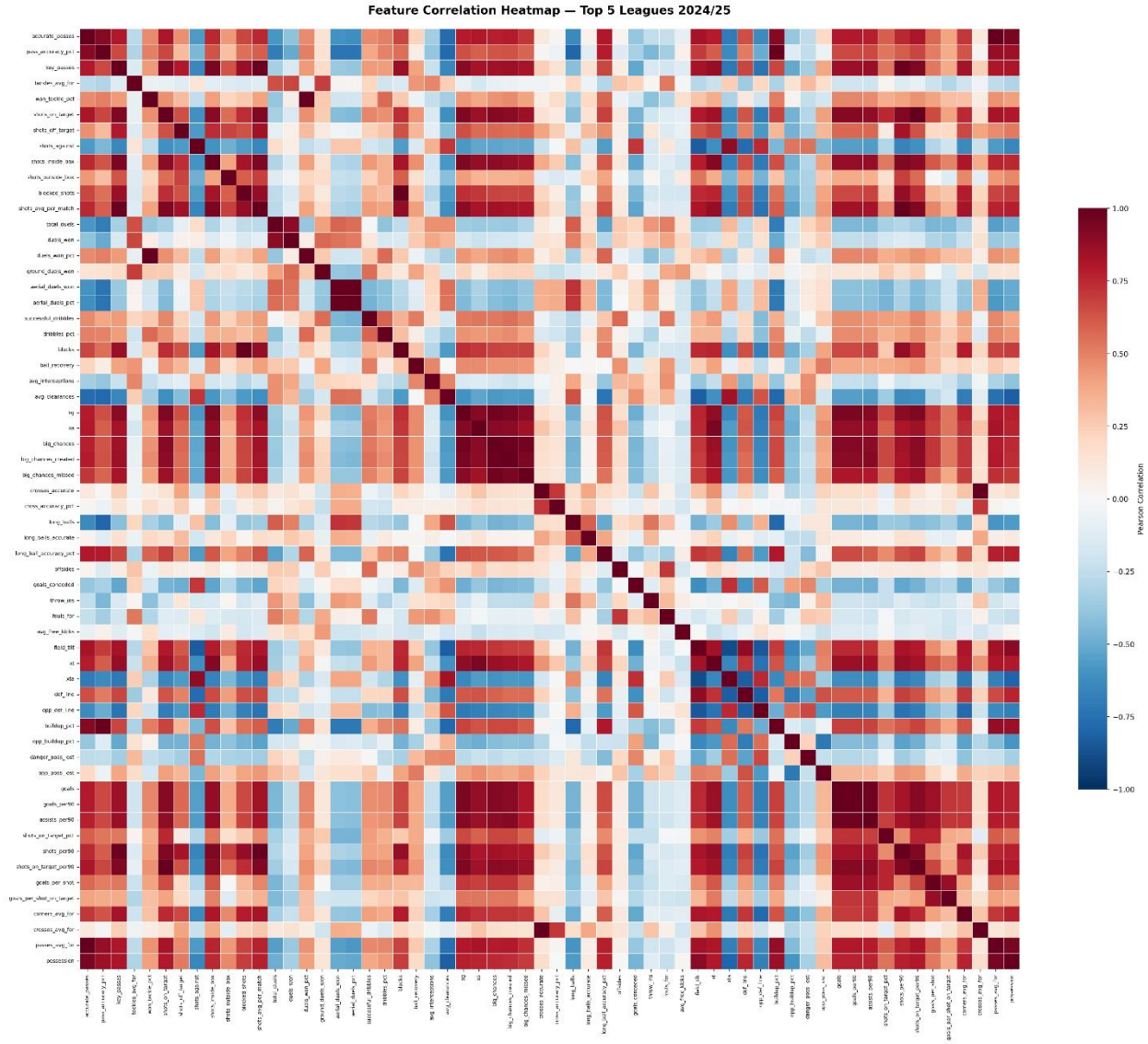


Figure 8 *Example excerpt of the highest correlated features from VS Code*

Feature 1	Feature 2	r
won_tackle_pct	duels_won_pct	1.000
blocked_shots	blocks	1.000
goals	goals_per90	1.000
crosses_accurate	crosses_avg_for	1.000
aerial_duels_won	aerial_duels_pct	0.996
accurate_passes	passes_avg_for	0.996
shots_avg_per_match	shots_per90	0.995
shots_on_target	shots_on_target_per90	0.994
pass_accuracy_pct	buildup_pct	0.990
key_passes	shots_avg_per_match	0.984
key_passes	shots_per90	0.980

Ultimately, Table 3 shows all of the selected features, including their domain for a better understanding of the data at hand.

Table 3 *All selected features*

Domain	**Feature**
Passing	Key Passes
Passing	Long Balls
Passing	Long Balls Accurate
Passing	Long Ball Accuracy %
Passing	Buildup %
Passing	Opponent Buildup %
Shooting	Shots Outside Box
Shooting	Shots On Target %
Attacking General	xG
Attacking General	xA
Attacking General	Big Chances Created
Passing	Cross Accuracy %
Passing	Corners Avg For
Miscellaneous	Avg Free Kicks

Defensive Actions	Tackles Avg For
Defensive Actions	Blocks
Defensive Actions	Avg Interceptions
Defensive Actions	Avg Clearances
Defensive Actions	Goals Conceded
Duels	Total Duels
Duels	Duels Won %
Duels	Ground Duels Won
Duels	Aerial Duels %
Duels	Successful Dribbles
Duels	Dribbles %
Duels	Fouls For
Possession	Possession
Possession	Field Tilt
Possession	xTA
Possession	Defensive Line
Possession	Opponent Defensive Line
Possession	Danger Poss Lost
Possession	Opponent Poss Lost
Miscellaneous	Ball Recovery
Miscellaneous	Throw Ins
Identifier	League
Identifier	Position
Identifier	Team
Identifier	Matches Played

4.5 Exploratory Data Analysis

Due to the nature of the data, the exploratory data analysis is conducted after the preparation, as otherwise this would not present an accurate overview of the different leagues. As aforementioned there is a different number of matches played within the top 5 leagues. Thus, a look at the data makes more sense in the context of this factor being accounted for.

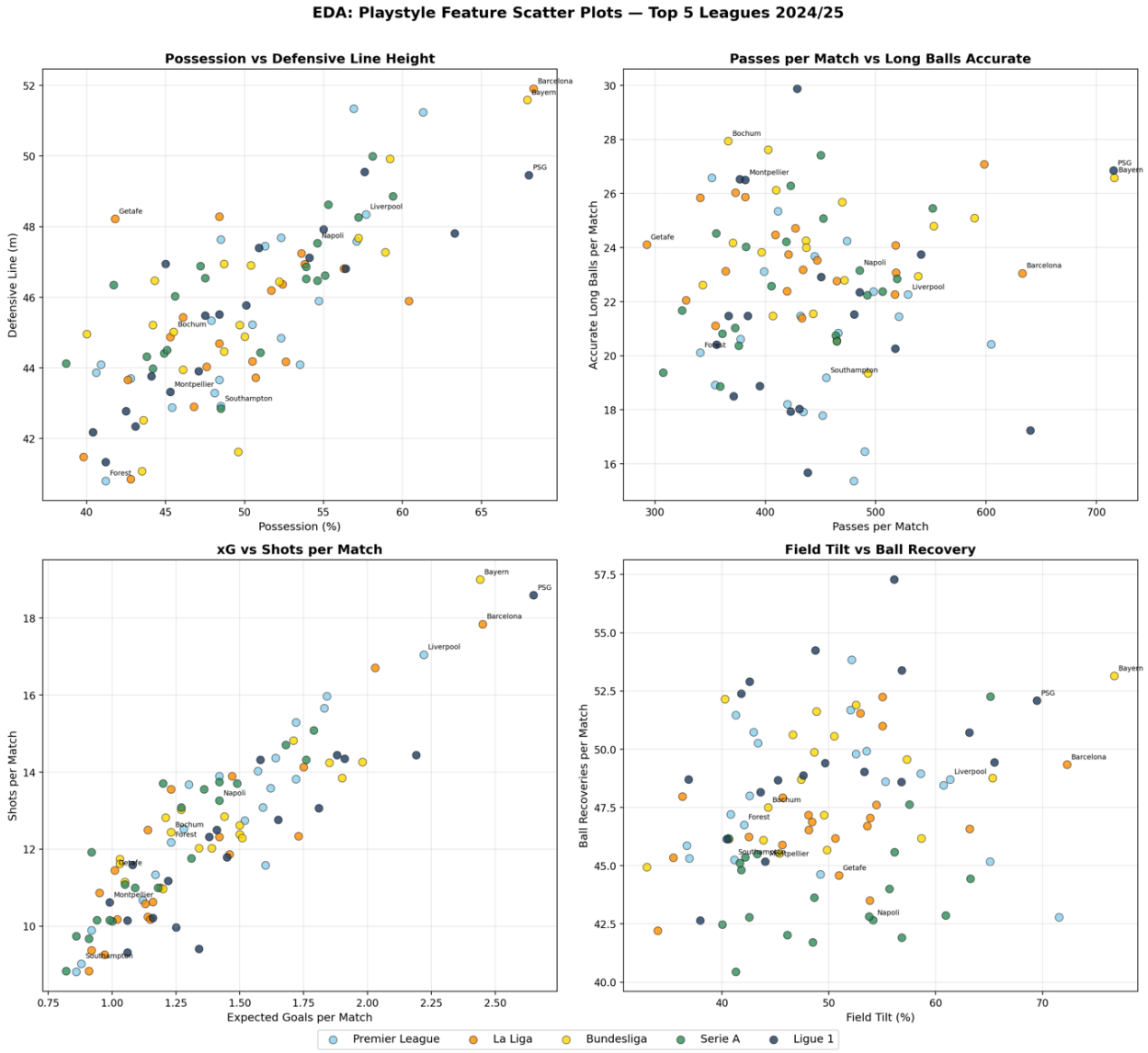
A first visualization of paired variables across a scatter plot delivers interesting insights. Figure 9 consists of four different plots incorporating all teams from the top 5 leagues in Europe. The upper left plots the teams with the features of possession and the defensive line height (this is the average distance in meters from a team's goal where their defensive actions occur), meaning this plot separates deep-block teams from team using a high defensive line. Visually one can immediately notice a strong positive correlation (correlation score of 0.76) between the variables. Teams who have a higher possession, tend to have a higher defending line. Very dominant possession teams with a high-line defence are the teams FC Barcelona and FC Bayern in the upper right corner. A deep-block defence can be especially noted in the lower left for teams like Nottingham Forest in the English Premier League.

Additionally, the next scatter plot visualizes the passes per match along successfully played long balls. Hereby, teams who use a rather more direct style of play versus teams who favour build-ups can be identified. PSG and FC Bayern are in the upper right corner with an extreme number of passes compared to the others, suggesting they favour a passing build-up approach in their style of play. Interestingly Southampton shows a lacklustre performance, but not the worst in their league, despite them finishing in last place with a huge margin to the second worst team. However, there is no correlation between these variables observable.

Furthermore, comparing the xG metric along the shots per match, shows the quality of the shots compared to their volume. Again, a strong positive correlation (correlation score of 0.89) is observable, meaning generally the more shots a team takes, the higher their number of expected goals is. FC Bayern, PSG and FC Barcelona are once again here the best performing teams across all leagues. They take the greatest number of shots while having the highest xG metric.

Lastly, plotting the field tilt % (the percentage of ball possession based on touches completed in the attacking third of the soccer pitch) versus the ball recovery shows a widely scattered result. On the one hand there are again the three dominant teams of FC Bayern, FC Barcelona and PSG in the upper right corner showing a high field tilt % and high recovery rates, meaning the majority of their possession occurs in the attacking third, while they are at the same time able to win the ball back quickly from their opponent. On the other hand, most of the other teams are clustering around the middle.

Figure 9 *EDA: Playstyle Feature Scatter Plots*



Furthermore, a box plot analysis of single variables across the top 5 leagues in this exploratory data analysis shows several notable patterns, as can be seen in Figure 10 and Figure 11. One of the most likely influential features is the ball possession as it has also been referenced in the literature before. Figure 11 shows the distribution of the possession for all distinct soccer teams in their respective leagues. Remarkably, the median across all leagues is very similar, but there is a wider spread in the Bundesliga and LaLiga. One can identify that the best performing teams are FC Barcelona (possession of 68.3%) and FC Bayern (possession of 67.9%).

Subsequently a look at the build-up visualization, which shows the number of successful passes completed outside the attacking third for a team, shows a bigger contrast between the leagues. Whereas the English Premier League records the highest median, the Bundesliga has visibly the lowest build-up percentage. An interesting negative outlier is Getafe from Spain, who record by far the lowest build-up value, which is in line with domain observations as they are known for their rather destructive playstyle.

Likewise, interceptions as seen in Figure 11, show a vastly different picture again across the top 5 leagues. Ligue 1 records by far the highest median (8.77 interceptions per game) in comparison to Serie A with the lowest median (7.37 interceptions per game).

What is more is, that in Figure 10 another important metric for the ball possession playstyle to look at, are the completed successful passes. There is not really a big difference across the medians of the top 5 leagues. However, in this metric there are more visible outliers appearing such as PSG in Ligue 1 or Manchester City in England. The best performing teams in their respective leagues are more successful across the board in this metric.

Lastly, the box plot for the xG metric highlights that the English Premier League has the highest median. Thus, one can infer that England has a higher attacking focus compared to the others. Notably Serie A records the lowest median with LaLiga. This is in line with Serie A being historically known for their focus on defence over attack.

Figure 10 *Box plots of Passes & xG for the top 5 leagues*

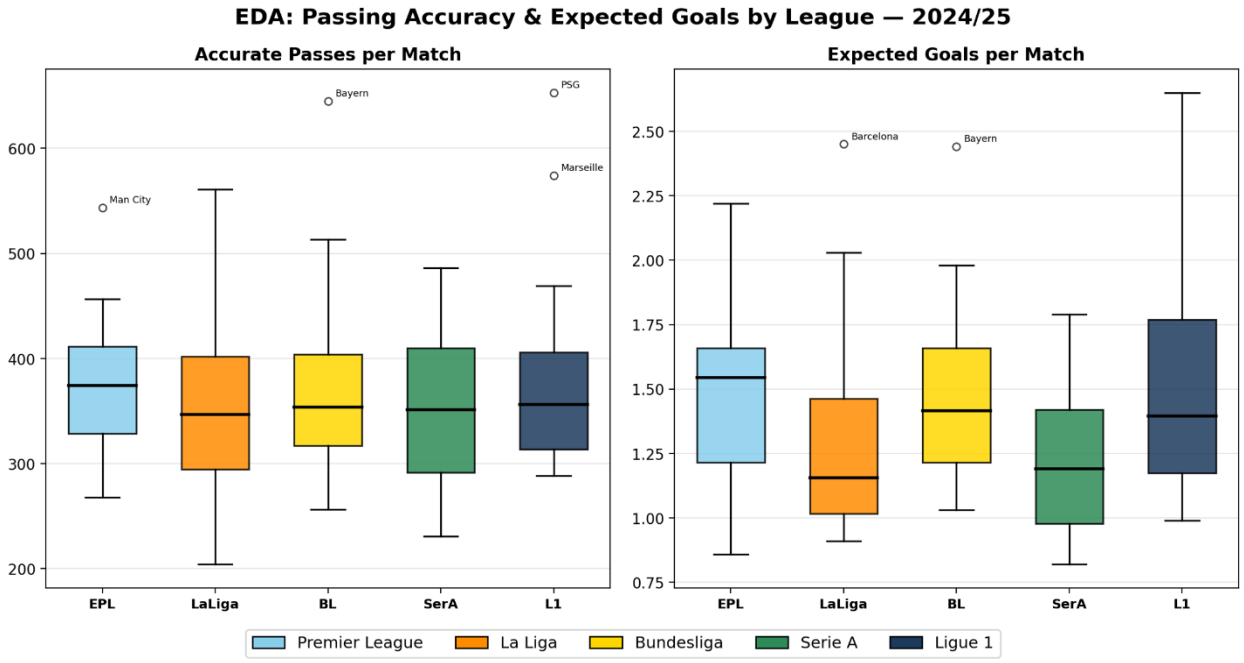
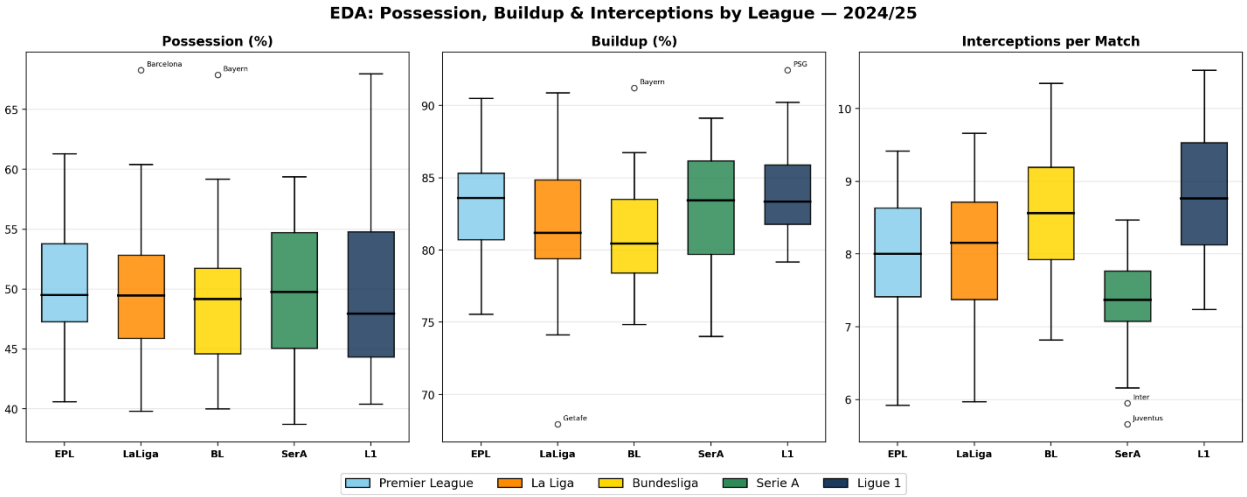


Figure 11 *Box plots of Possession, Buildup & Interceptions for top 5 leagues*



4.6 Modelling

The modelling phase of this thesis consists of a sequential approach. Firstly, Principal Component Analysis (PCA) will be applied to reduce the dimensionality of the cleaned dataset. Subsequently, clustering based on k-means is used to capture and identify distinct playstyles amongst the diverse selection of international soccer teams. Thus, the newly labelled data is the bases for which two different supervised machine learning algorithms will be used on. In particular, these

are XGBoost and Random Forest, whereas the results of those will be hereafter compared with one another.

Beginning with PCA, it is considered an unsupervised machine learning algorithm. According to Greenacre et al. (2022), principal components are essentially linear combinations of the original variables, which aim to reflect a maximum of the variance of the original data, while getting rid of the redundant variables. Each successive principal component is ordered by the amount of variance it is able to capture. Furthermore, due to the nature of the dataset, the original features have to be standardized. This is achieved through a z-score normalization, meaning the features have a mean of zero and a standard deviation of one. Afterwards, for the further modelling steps, a certain number of principal components will be selected based on a threshold of 80% for the variance that they need to capture. Another advantage of this approach is that PCA removes multicollinearity, which enables further insights into the data with two-dimensional visualizations. Generally, PCA is especially suited for datasets with a large number of correlated features, as the dimensionality can be reduced without losing much information (Greenacre et al., 2022).

In the next step, another unsupervised machine learning algorithm is applied, namely k-means. This algorithm labels the unstructured data by finding playstyle clusters. It offers flexibility, efficiency and ease of implementation as it is among the top 10 clustering algorithms used in data mining. Moreover, k-means operates by starting with a set number of clusters k, whereas k random points in the data are selected as their cluster centroids. The Euclidean distance from all the data points to these centroids is calculated and used to assign each data point to its closest centroid and thus to a specific cluster. After this first iteration, the cluster centroids are reassigned based on the average of the points within the cluster. Followingly, the distances are calculated again with a subsequent reassignment to the corresponding clusters. Once the set threshold of iterations is reached, or if there are no more changes in the assignment, the algorithm ultimately stops and outputs the final cluster assignment of the data. Hence, k-means is an integral part of the process, as the newly labelled data can subsequently be used as the training input to predict playstyles (Ikotun et al., 2023).

Therefore, XGBoost, a supervised machine learning algorithm, can be applied to the data for the prediction. XGBoost, namely Extreme Gradient Boosting, is essentially a gradient-boosted decision tree algorithm. Gradient boosting means one simple decision tree is built, which acts as the

base learner. The errors between the prediction and the actual data are calculated, whereas more decision trees are then iteratively added, each of them fitted to the gradient of the loss function to correct and improve the errors of the previous ones. Consequently, the algorithm stops once a threshold of decision trees is met or if the overall model can no longer be improved. Due to its sequential nature, XGBoost is particularly suited for large datasets, where it performs better, as in smaller datasets it is at risk of overfitting (Chen & Guestrin, 2016).

In addition, Random Forest is used as the other supervised machine learning algorithm, in order to compare the performance of the models with one another and to determine whether an accurate prediction of playstyles is possible. Similarly, decision trees are used, but hereby multiple decision trees are combined in one single prediction. The respective parameters determine the size of the nodes, the number of decision trees and features used. However, in contrast to XGBoost, Random Forest is using an ensemble of independent decision trees, with each of them being trained on a random subset of the original features. Hence, this results in a decorrelation of the decision trees, while an aggregation of all their outputs reduces the total variance and minimizes the risk of overfitting. Accordingly, the Random Forest algorithm is especially suited for smaller datasets, as it is able to generalize well even with limited information (Breiman, 2001).

Besides, to combat the problem of overfitting in the predictions, the method of cross validation was applied. It is a resampling procedure that is used while working with datasets limited in size. In this thesis a 5-fold cross validation is chosen as it is a popular choice in the application of data science. This means that the model is not only getting evaluated once, but actually over five different partitions of the dataset.
Additionally, the data is split into a 75% training dataset and a 25% test dataset for the algorithms, including stratified splitting to preserve the original cluster proportions across the different splits. The hyperparameters for the supervised algorithms aren't chosen manually, but with the help of GridSearchCV. This method is part of the scikit-learn python library and evaluates several possible combinations of different key hyperparameters, resulting in the best performance of the models regarding the accuracy (Jason Brownlee, 2023).

For further interpretability regarding the predictions, SHAP (SHapley Additive exPlanations) was used. SHAP is based on game theory, which is used to help explain the predictions made by

the supervised machine learning algorithms. Moreover, SHAP incorporates all possible feature combinations and then calculates the model's predictions with and without each feature, to output which features are the most influential for making the prediction.
In more detail, the SHAP value attributes to each feature the change in the expected model prediction that results from that one specific feature. Starting with the base value, this is the prediction the model computes without knowing any features. Each feature getting added then changes the value of the prediction until the final output is reached. As the order in how the features are added changes their contribution, the final SHAP values are calculated by averaging each feature's contribution to the prediction across all possible ordering iterations (Lundberg & Lee, 2017).

As for the last model, Ordinary Least Squares (OLS) regression is used in the performance case study analysis. OLS hereby investigates whether an opponent's playstyle does indeed have a statistically significant influence on the performance of a team. It operates by assuming a linear relationship between the dependent (target) variable and the independent (predictor) variables, which is represented through a straight line along the data. The differences between the observed values and the predicted values are the residuals. As positive and negative deviations would cancel each other out, the residuals are squared, whereas larger deviations are also penalized. Finally, the fitted line is the most optimal as it shows the least amount of total squared error. In the output, the coefficients denote the predicted change in the target variable per one-unit change in the corresponding predictor, while all other predictors stay constant (Burton, 2021).

4.7 Evaluation metrics

Furthermore, to sufficiently evaluate and compare the performance of the applied models, several established metrics were used. Regarding the prediction of playstyles, a confusion matrix is suitable, as this is a multi-class classification problem. The confusion matrix shows the predicted classes in a table against the actual cases in the data. Hereby, a comprehensive overview of the model where it made correct and incorrect predictions is offered. In more detail, there are four different outcomes. True Positives are correctly predicted positive cases; True Negatives, are correctly predicted negative cases; False Positives are negative cases that have been falsely

predicted as positive; and False Negatives are positive cases that have been falsely predicted as negative (Google Developers, 2026).

In addition, various other established metrics are used, for instance the accuracy. It measures the correctness of the entire model, meaning the correct predictions over the total predictions:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

However, it is to be noted, that accuracy should not be seen as the one defining metric. This metric can be misleading in imbalanced datasets, as a very high accuracy could be achieved by simply predicting the majority class. This is evidently the case in this dataset, as the to be predicted playstyle clusters do indeed differ in size. Other metrics will thus complement the evaluation (Mills et al., 2024).

In particular, the recall metric is used, it shows the proportion of the correctly classified positive cases over all actual positives in the data. Hence, it captures the effectiveness of the model for finding all relevant cases of a certain class:

$$\text{Recall} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Besides, the precision metric measures the proportion of the positively predicted cases, which are truly positive. The precision shows how reliable the positive predictions of a model are:

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

However, the precision and the recall metric have to be viewed as trade-offs between each other, as optimizing for a better recall score can lower the precision and vice versa. Therefore, the F1-score comes into play, as it combines both metrics into one measure through calculating the harmonic mean. This approach penalizes low values in either of the metrics more heavily than just an arithmetic mean:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$

A high F1-score indicates a good performance of both the precision and the recall. Thus, this metric is quite suitable for a multi-class classification problem like it is the case in this thesis (Mills et al., 2024).

Furthermore, besides the metrics used for the classification and the model performance, two other metrics are used in the empirical section. Firstly, the silhouette score is used to help in selecting the optimal number of clusters for k-means. This metric measures how similar a data point is to the other data within its own cluster compared to the data points of the nearest neighbouring cluster. Values range from -1 to 1, whereas a value close to 1 means that a point is similar to its own cluster and different from others, a value close to 0 means that the data is overlapping between clusters (GeeksforGeeks, 2026).
Moreover, in the case of the OLS regression in the performance case study, the R^2 metric is used. It measures the proportion of the variance from the target variable that can be explained by the predictor. Accordingly, a higher value means the statistical model can successfully explain a large proportion of the variance. The data points are close to the fitted regression line (Burton, 2021).

5. Results and Evaluation

Hereby, this following chapter outlines the results of the aforementioned analysis framework from the prior methodology section. First, a dimensionality reduction with Principal Component Analysis is applied, whereas the subsequent clustering with k-means identifies four distinct playstyle clusters across all 96 teams from the top 5 European soccer leagues in the 2024/25 season.

Followingly, two supervised machine learning algorithms are applied to the available data, particularly XGBoost and Random Forest. Moreover, their respective performance is evaluated through established classification metrics and a SHAP analysis.
In addition, a case study using FC Barcelona is performed. Hence, this performance analysis examines how the performance of one of the best soccer teams in the world varies while they face opponents of different playstyles. Besides, to ensure robust and reliable results throughout the predictive modelling phase, a 5-fold cross validation was applied. Similarly, the researchers Moustakidis et al. (2023) used XGBoost and Random Forest in a comparable soccer analytics context. However, due to the nature of the dataset in this thesis, a 5-fold cross validation was chosen over a 10-fold split as it was the case in the research paper, as this would have resulted in insufficiently small training partitions ultimately compromising the robustness of the results.

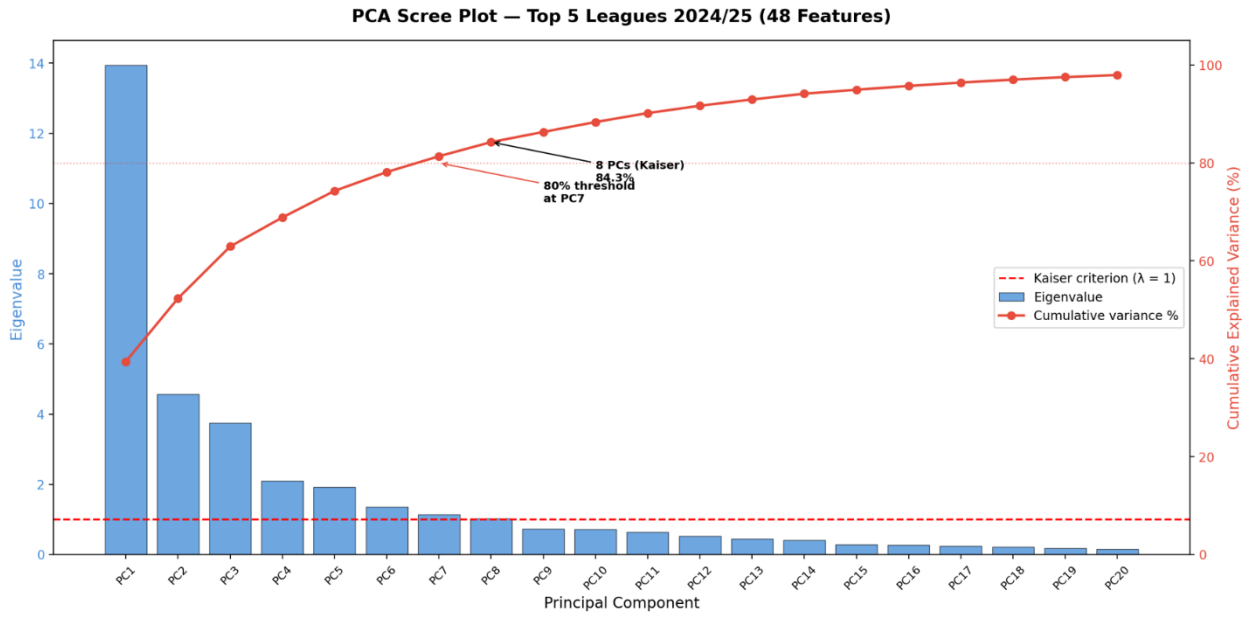
5.1 Principal Component Analysis

Following the aforementioned feature reduction based on Pearson's correlation coefficient, PCA is used in this next step on the remaining features before the clustering starts. PCA hereby further reduces the dimensionality of the dataset, resulting in an extraction of the most impactful and relevant features. As found in the literature, there are multiple options to choose from for how many final principal components will be selected for the final model. Following Castellano and Pic (2019), a threshold of the PCA capturing 80% of the total variance is used.
As highlighted in Figure 12, the scree plot shows that upon reaching n = 7 for the principal components, the threshold of 80% is met. The very first principal component (PC) contributes 39.43% of the variance as the most dominant dimension. Moreover, the PC2 and PC3 are capturing 12.92% and 10.60% of the variance respectively.

Additionally, accounting for the Kaiser criterion (as marked in Figure 12), where a principal component has an eigenvalue over 1, this would result in a total of 8 selected PCs, but the eighth component would only deliver a very marginal variance pay-off. Thus, to avoid overfitting the model, the final number of PCs stays at 7. This is again in accordance with the findings of domain experts as Lago-Peñas et al. (2017) used 5 different PCs. However, they only used 20 features in their dataset, which is a stark contrast to the 39 (4 of them act as identifiers) features used in this thesis.

Prior to the application of PCA, all of the features in the dataset were standardized using a z-score normalization. Hence, this results in each feature having a mean of zero with a standard deviation of one across the board. Using a method like this is necessary due to the different nature of the features and the scale of the values. Moreover, this prevents some features with disproportionately high or low values influencing the results of the PCA.

Figure 12 *PCA Scree Plot*



Looking at Table 4, it provides a detailed overview of the exact feature loadings for the two most influential principal components. Accordingly, PC1 represents a team's possession and attacking output accounting for the features with the highest loadings. Thus, it is labelled as *"Possession &*

Attacking Dominance". Evidently past research has identified possession as the most dominant indicator to differentiate between playing styles of teams as per Lago-Peñas et al. (2017) and Castellano and Pic (2019).

Looking at PC2, labelled *"Defensive Intensity"*, this component characterizes the defensive playstyle of a team.

Followed by PC3, which has negative loadings for successful dribbles, but positive loadings for throw ins and aerial duels, it showcases a trade-off between physicality and technicality of a team.

Next up is PC4, showing set pieces and the crossing threat of a team, while PC5 demonstrates direct play and long balls. Moreover, PC6 captures the difference of crossing versus dribbling as methods of ball progression with PC7 lastly encapsulating teams that avoid risky defensive actions. This can be seen through the high negative loading for possession lost during one's own defensive third.

Table 4 *Top five feature loadings for PC1 & PC2*

PC1	**Feature Loading**
field_tilt	+0.251
possession	+0.250
key_passes	+0.244
xg	+0.239
corners_avg_for	+0.232
PC2	**Feature Loading**
ball_recovery	+0.344
opp_poss_lost	+0.334
total_duels	+0.332
fouls_for	+0.315
tackles_avg_for	+0.265

5.2 Clustering with k-means

As aforementioned, after having established the seven retained principal components, the transformed dataset was used as the input for the following clustering analysis.
Hereby, k-means, an unsupervised machine learning algorithm, was applied to identify distinct playstyle clusters across the 96 different teams from the top 5 leagues in European soccer. K-means is a widely established and used algorithm in the literature as can be seen in the works of Plakias et al. (2023) and Immler et al. (2021).

However, one has to determine the optimal number of clusters. Followingly, two different methods were used to settle on an appropriate *k* of clusters as ultimately this hybrid approach ensures clustering quality and efficiency.
Firstly, the elbow method was used, which visualizes the inertia, the within-cluster sum of squares. The total of the within-cluster sum of squares is plotted against an increasing number of *k* clusters, which ideally shows an "elbow" point where the rate drastically starts to change. Hereby, it shows that while adding more clusters reduces the error, it has diminishing returns as the gains flatten and one risks an overfitting of the model.
Figure 13 showcases that there is not one drastic cut-off point in the used dataset, but the returns continually diminish.
On the other hand the silhouette score was used, which measures how similar a data point in one cluster is to one in the neighbouring cluster. Hence, it is used to paint a clearer picture of the separation between clusters. As evident in Figure 14, a $k = 2$ has the highest silhouette score followed by a sharp drop-off for the further number of clusters (GeeksforGeeks, 2025).

Figure 13 *Elbow method plot*

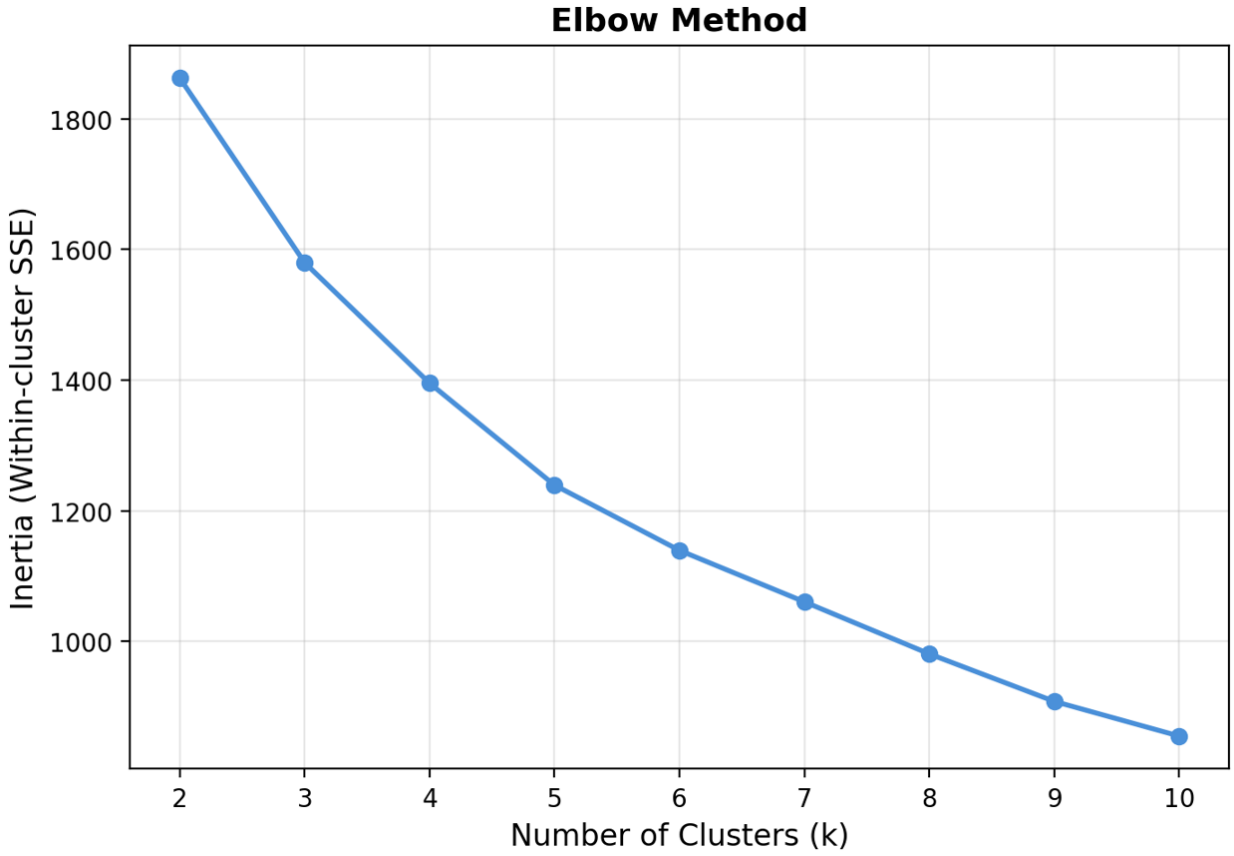
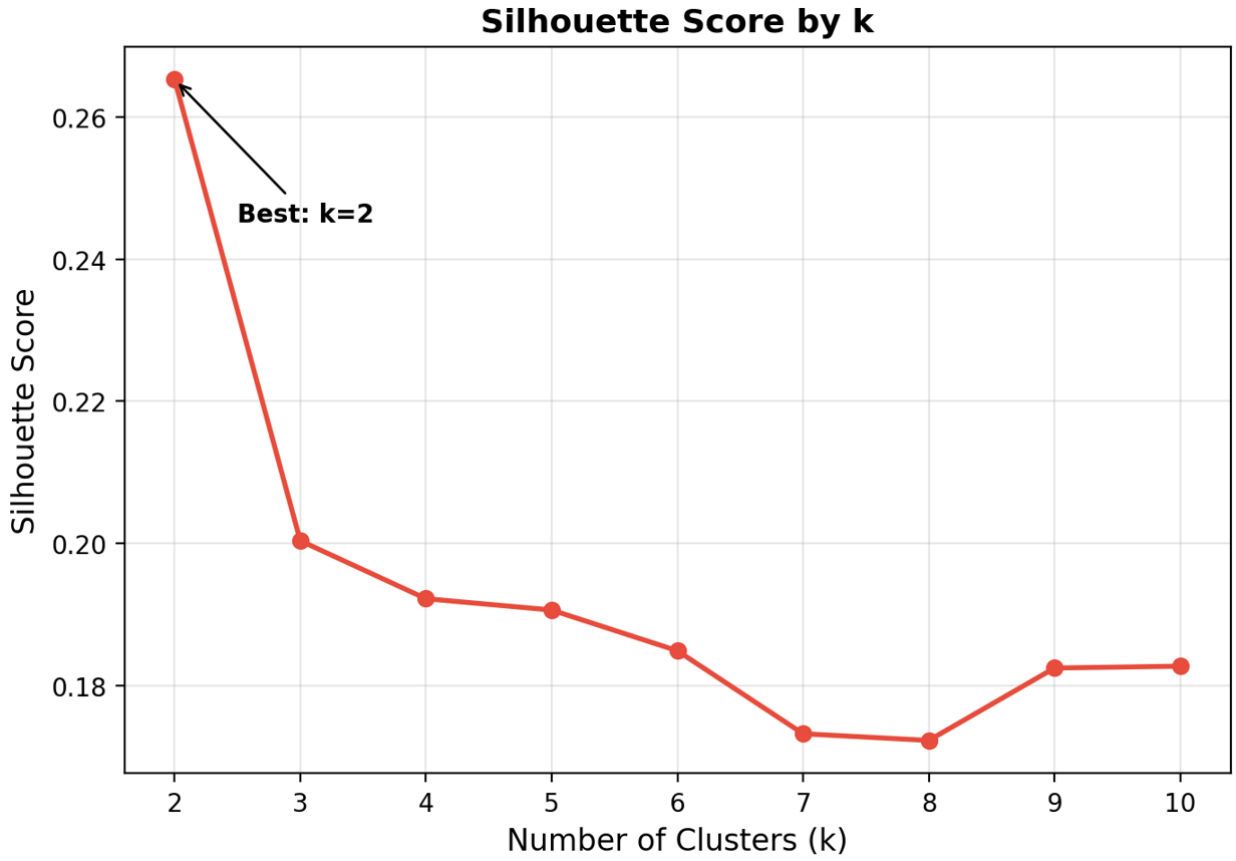


Figure 14 *Silhouette score plot*



However, while $k = 2$ is theoretically the best solution, this decision would ultimately result in several limitations for the analysis. Having just two clusters does not enable one to properly identify distinct playstyles and would further be an overfitting of the data in this case. The following analysis would be too generalised in its conclusions. Hence, this binary split resulting in dominant and dominated teams was deemed insufficient.
Looking at the silhouette scores, there is a further drop-off happening after $k = 5$. Additionally, Castellano and Pic (2019) identified four different playstyles in their research. Consequently, this led to the decision of testing both cases of $k = 4$ and $k = 5$ for the k-means algorithm.

The results for each respective value of k show that $k = 5$ is not an ideal solution, as there are smaller clusters which can lead to further complications in the next step, as they would be underrepresented in the training and test data split. Additionally, the clusters are heavily overlapping in multiple instances and a visual distinction is not explicitly visible. Furthermore, the possibility of overfitting must be accounted for here.

Thus, the decision was made to choose $k = 4$ as the final amount for the k-means algorithm. Moreover, this decision offers a more balanced and interpretable result, with a closer alignment to prior identified playstyles in the literature. The number of runs for the algorithm to iterate through its different centroid seeds was set to 20 to adequately capture the best possible result. Additionally, the random state was set as a fixed integer, which makes the randomness deterministic and ensures reproducibility for the results.

Figure 15 shows the identified four distinct clusters labelled with different colours. Their distribution can be seen in Table 5. Hereby the found playstyle clusters are named after their corresponding feature profile. Each of them features distinct playstyle features that make a meaningful interpretation possible. Cluster 1 records the lowest number of fouls, the lowest duels and tackles, which resembles more of a balanced defensive playstyle. Followed by cluster 2 which showcases direct play features like lots of long balls, but also low build-up and possession. Additionally, cluster 3 has the highest xG, xA, possession and key passes among all the clusters. This consists exclusively of dominant and possession-based teams. Lastly, cluster 4 records the highest number of tackles, the highest ball recoveries, and lots of interceptions per game including losing possession in dangerous positions quite often. This is a very aggressive playstyle with pressing.

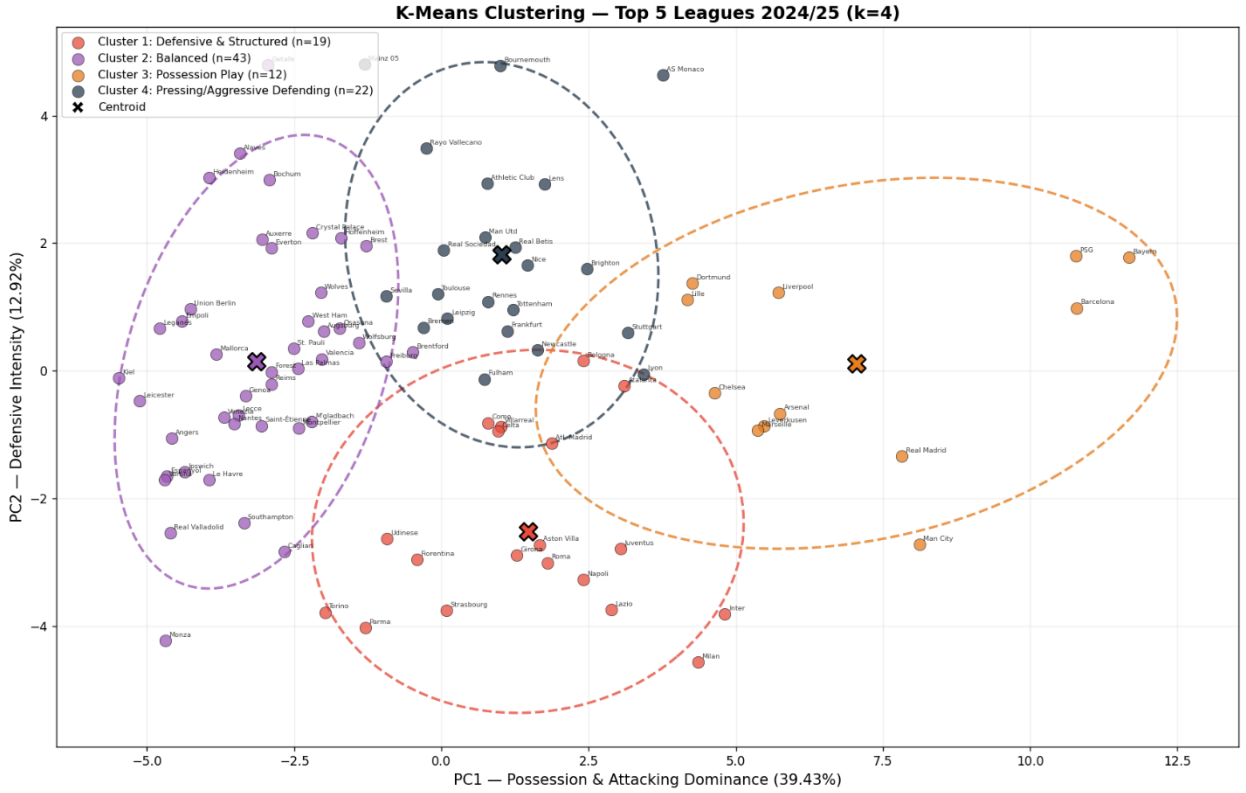
Looking at the research by Peng et al. (2025) there is a distinctive overlap in the features for the clusters of this thesis along their identified playstyle dimensions. Possession Play closely aligns with cluster 3, which was thus labelled *Possession Play* too. Aggressive Defending resembles cluster 4 the most, but in this instance other key pressing features like ball recoveries were captured too, leading to the naming of *Pressing/Aggressive Defending.* Moreover, cluster 2 has partial similarities to Direct Attacking, while hereby this cluster encapsulates less shot creation. Otherwise, many metrics lie in the middle of the range, thus leading to the label of *Balanced*. For cluster 1 then *Defensive & Structured* was chosen for its balanced defensive playstyle. Divergences from the literature were expected here, as the researchers focused on national soccer teams instead of club competitions.

Table 5 Cluster distribution

Cluster	N
1	19
2	43
3	12
4	22

To further highlight the distinct playstyles, dashed ellipses have been added in Figure 15. These represent confidence ellipses per each playstyle. Meaning, an ellipse denominates the spread of the dataset through calculating the covariance matrix of the data within each cluster. Eigenvectors determine the direction of the ellipse, in the example of cluster 3, which is oriented diagonally toward the upper right. This means that teams who have a high PC1, also tend to have a higher PC2 value. Furthermore, a standard deviation of the value 2 was applied as a parameter, resulting in an ellipse capturing the whole area within 2 standard deviations from a cluster centroid. Moreover, this clearly visualizes that there is a small overlap to a certain degree between the clusters.

Figure 15 *Results of clustering with k-means*



Consequently, looking more detailed at the results, various insights are available after investigating the data. Table 6 shows an overview of the respective mean features across a subset per playstyle. Additionally, Table 7 displays the distribution of the identified playstyle across the top 5 European leagues.

Notably, *Defensive & Structured* features almost exclusively teams from Serie A, with 13 out of 19 teams being from Italy. This observation is in line with the Italian league being historically known for a more defensive approach. Moreover, coaches widely characterize the Italian league by its tactical, defensively focused approach. There is even a term coined for this Italian playstyle, namely “catenaccio”, which represents defensive play and aggressive fouls (Sarmento et al., 2013).

As expected, the *Balanced* playstyle makes up the largest cluster in terms of sheer size. One cannot identify any extreme outliers in the features, except for the highest value for played long balls. However, likewise this playstyle also features the lowest amount of possession. Overall, the values are all very evenly distributed across the dataset.

Highlighting *Possession Play*, this playstyle features the most distinct observable feature values

like extremely high possession (61.38%), key passes (12.55), while also having the fewest number of conceded goals per game (1.12). As one would expect, this cluster consists of the teams which are the most dominant in their respective leagues like Bayern Munich, FC Barcelona, or Manchester City.

Besides, *Pressing/Aggressive Defending* consists of the teams with the most physical and aggressive playstyle. Not only does it record the highest physical actions and ball recoveries, but it also features the highest metric for dangerous possession lost by opponents, indicating that these teams are pressing up high on the pitch leading to turnovers.

Table 6 *Feature Mean Table*

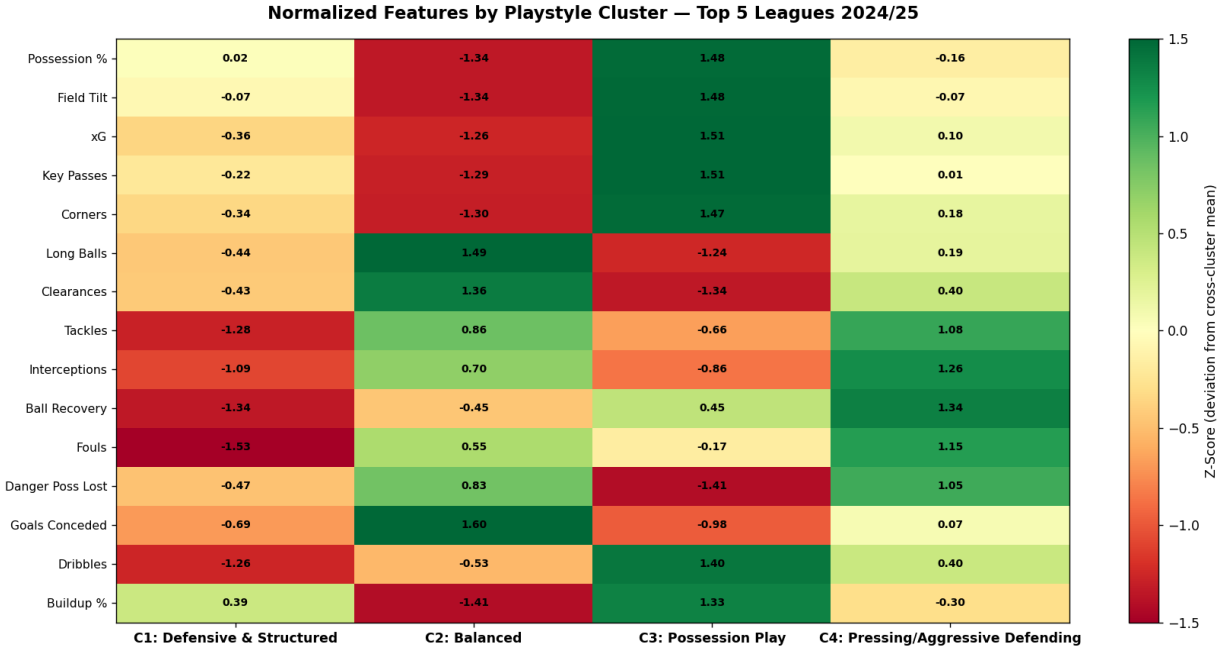
Feature	**Defensive & Structured**	**Balanced**	**Possession Play**	**Pressing/Aggressive Defending**
Possession %	52.77	44.71	**61.38**	51.72
Field Tilt	53.14	42.55	**65.97**	53.08
xG	1.41	1.12	**2.02**	1.56
Key Passes	9.72	7.97	**12.55**	10.09
Corners	4.86	4.15	**6.19**	5.24
Long Balls	45.48	**53.25**	42.25	48.00
Clearances	19.74	**26.02**	16.53	22.64
Tackles	14.81	17.00	15.44	**17.21**
Interceptions	7.03	8.45	7.21	**8.89**
Ball Recovery	44.91	46.80	48.73	**50.64**
Fouls	7.92	11.08	9.98	**11.99**
Danger Poss Lost	21.25	23.57	19.59	**23.96**
Goals Conceded	1.18	**1.65**	1.12	1.34
Dribbles	4.84	6.12	**9.47**	7.72
Buildup %	85.54	79.19	**88.86**	83.11

Table 7 *Distribution per league*

Cluster/League	**Defensive & Structured**	**Balanced**	**Possession Play**	**Pressing/Aggressive Defending**
Bundesliga	0	10	3	5
LaLiga	4	9	2	5
Ligue 1	1	8	3	6
Premier League	1	9	4	6
Serie A	13	7	0	0

Moreover, different features have different influences on how a playstyle belongs to a certain cluster. In Figure 16 the normalized feature values, with a z-score indicating the deviation from the mean across all clusters, are observable. Consequently, strong positive values (labelled as dark green) signalize that a feature value is in this particular cluster way above the average. Likewise, strong negative values (labelled as dark red) indicate a feature value far below the average. This heatmap further highlights the distinct features of each playstyle and enables an easier interpretation. Observations like *Possession Play* having high scores in features like xG, and possession; *Pressing/Aggressive Defending* having a high score in ball recovery are all in line with prior discussion of the findings.

Figure 16 *Normalized Feature Profile*



5.3 Predictive Modelling

Following the clustering and the identification of four different distinct playstyles, the supervised machine learning approach of classification was used to investigate if it is possible to accurately predict the playstyles from available performance features. Hereby, this algorithm tackles this problem as a multi-class classification, in which the predicted variable will be assigned to one of the four target clusters. Consequently, two different machine learning approaches are applied to the same problem. Afterwards, the performance of both XGBoost and Random Forest will be compared with each other. This is in line with prior research where Moustakidis et al. (2023) used both of these to perform classification tasks in the realm of soccer data analytics.

Furthermore, as aforementioned, the dataset was split into a training set and a test dataset with a ratio of 75/25. In explicit terms, this means 75% of the available data of the 96 teams were used as the training data for the algorithm, whereas the remainder is used for the performance evaluation. Likewise, stratified splitting was used to keep the cluster proportions intact across the splits. This step is needed, as highlighted earlier, since the identified four clusters are imbalanced.

5.3.1 Predictive Modelling with XGBoost

To identify the optimal selection of hyperparameters for XGBoost, GridSearchCV was used, which is part of the Python library scikit-learn. Hence, this method evaluates several possible combinations of different key hyperparameters that ultimately maximize the accuracy of the model. Additionally, 5-fold cross validation was applied, as laid out in the methodology section of the modelling steps. This ensures a more robust and reliable result.

Table 8 *Optimal Parameter Selection for XGBoost*

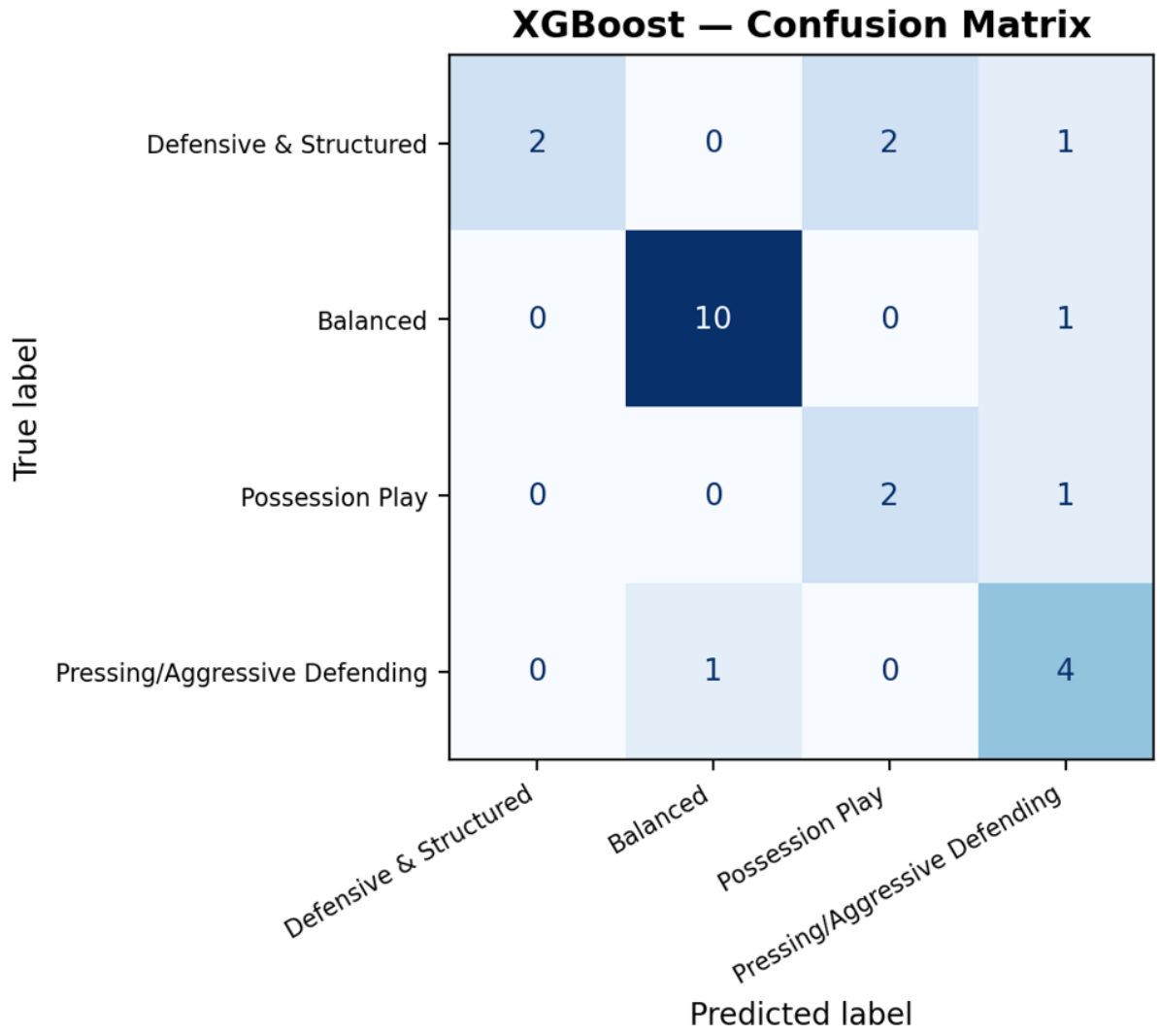
Parameters	XGBoost
colsample_bytree	1.0
learning_rate	0.01
max_depth	3
n_estimators	200
subsample	0.7

Moreover, Table 8 shows the identified optimal hyperparameters after the application of GridSearchCV. Hereby, colsample_bytree = 1.0 means that all seven principal components are used in each subsequent tree, also the low learning rate of 0.01 indicates that the model performs better with smaller incremental learning steps, as each tree only provides a small correction. Furthermore, a maximum tree depth of just 3 is selected as optimal. This makes sense to avoid overfitting in this dataset of 96 teams. Likewise, the optimal model uses 200 sequential trees with a subsample ratio of 0.7, so each tree is hence trained on a random 70% set of the training data.

Looking at the performance of the model, the results show that XGBoost reached an overall cross-validated accuracy of 90.48% using the optimized hyperparameters, whereas the accuracy for the test set is 75%. In more detail, Table 9 shows that the *Balanced* playstyle could be predicted with the highest accuracy out of all four distinct playstyles. Furthermore, this is also apparent in the confusion matrix in Figure 17. A reason for that could be, that this playstyle is also the largest cluster in terms of the number of teams in it.
Pressing/Aggressive Defending remarkably shows a recall value of 80% for its teams, meaning 4 out of 5 could be correctly identified, but there are multiple false positives in this category. Following that is *Possession Play*, which performs worse regarding the recall and the accuracy metrics.
Lastly, *Defensive & Structured* shows the highest precision score at 100%, as the model always predicted that playstyle correctly. However, only 40% of the teams could be captured by the model as seen through the recall value.

In general, the model is very effective at accurately predicting the various established playstyles across the top 5 European leagues.

Figure 17 *XGBoost Confusion Matrix*



Moreover, the SHAP analysis in Figure 18 shows that PC1 is the most influential input indicator, which was expected, as PC1 captures the most variance followed by the other PCs in descending order. PC1 especially contributed to the identification of *Balanced*, but *Defensive & Structured* the least out of the four playstyles. Additionally, PC2 is used by the algorithm the most for *Defensive & Structured* and *Pressing/Aggressive Defending*. The remaining components mostly contribute in small to moderate amounts to the classification task at hand.

Figure 18 *SHAP Analysis - XGBoost*

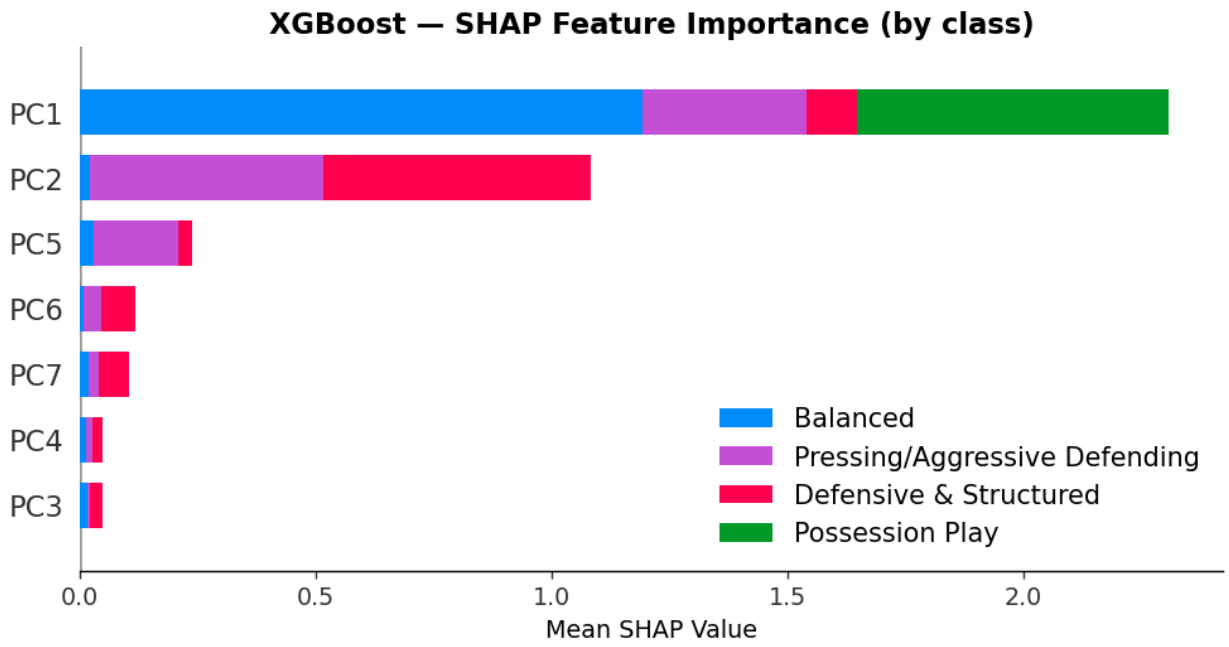


Table 9 *Classification Table of XGBoost*

Playstyle	Precision	Recall	F1-score
Defensive & Structured	1.00	0.4	0.56
Balanced	0.91	0.91	0.91
Possession Play	0.5	0.67	0.57
Pressing/Aggressive Defending	0.57	0.8	0.67
Overall Accuracy	0.90		
Test set accuracy	0.75		

5.3.2 Predictive Modelling with Random Forest

Likewise, as aforementioned, GridSearchCV was implemented to select the most optimal hyperparameters for this supervised machine learning algorithm. Consequently, the same approach of a 5-fold cross validation was used, which leads to robust and reliable results.

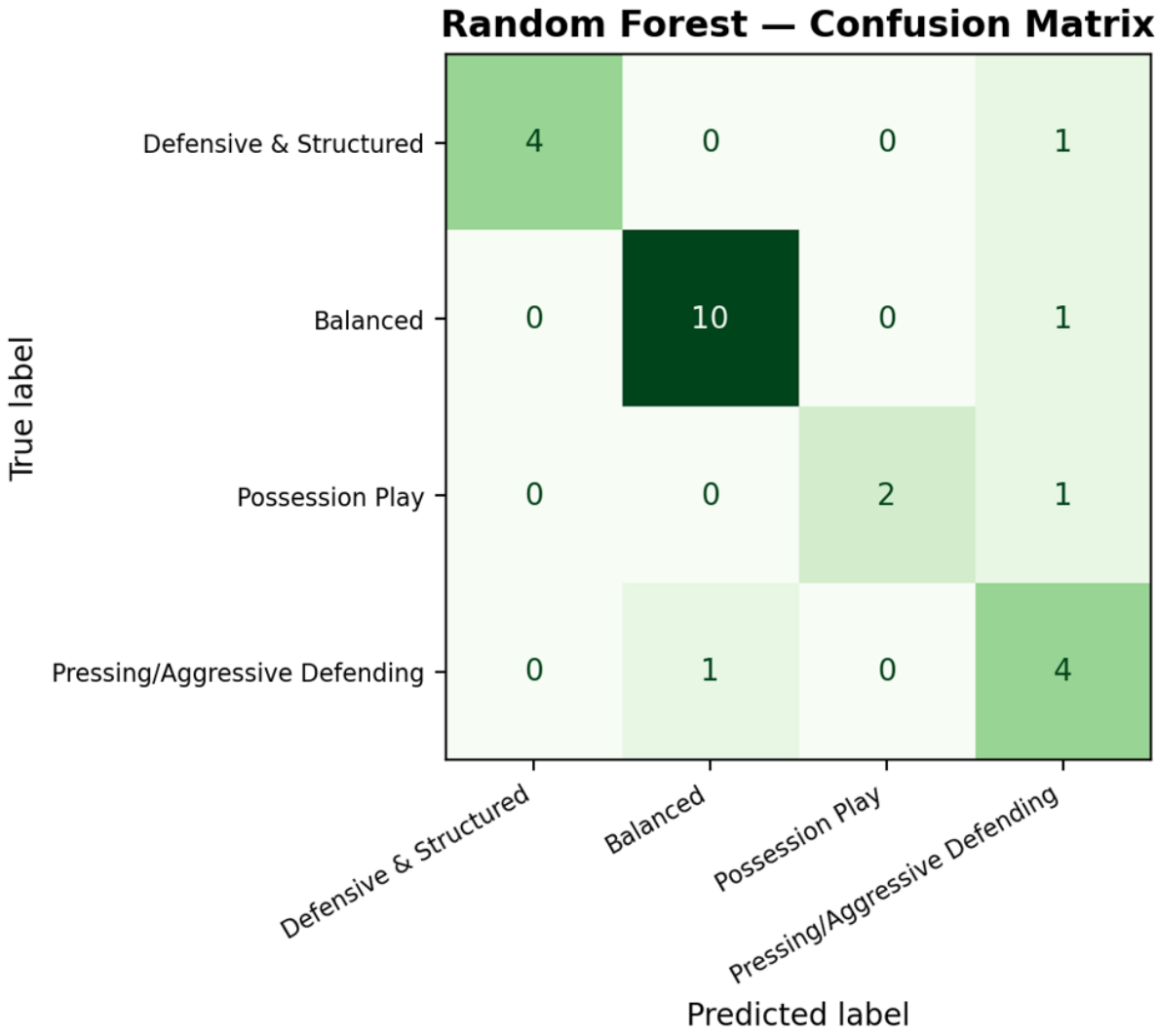
Table 10 *Optimal Parameter Selection for Random Forest*

Parameters	Random Forest
max_depth	None
max_features	sqrt
min_samples_leaf	1
min_samples_split	2
n_estimators	50

Moreover, Table 10 shows the final selected optimal parameters. There is no maximum depth constraint, which leads to the decision trees being able to be very specific in their final decisions as they terminate once all their leaves are either pure or the minimum split threshold is not met. Setting the number of estimators to 50, means that there are 50 different decision trees which are trained independently from each other. Additionally, there is a minimum of one sample per leaf. The square root of the total input features is set as the maximum for the split at each node in the tree. So, only a randomly determined subset of the PCs acting as the input features, is selected for the splitting. Consequently, this leads to more diversity in the ensemble of trees to avoid overfitting.

In fact, the results of the Random Forest model, as can be seen in Figure 19 and Table 11, demonstrate a cross-validated accuracy of 89.14% overall and an accuracy of 83.33% for the test set. Once again, the confusion matrix shows that the *Balanced* playstyle cluster could be predicted with the highest reliability boasting a precision of 0.91, meaning 91% of the teams predicted in that cluster, were actually part of that playstyle in the real data. Similarly, a recall of 91% indicates 10 out of 11 actual teams from *Balanced*, could be found through the Random Forest model.
Followingly, *Defensive & Structured*, achieved a perfect 100% precision score and an 80% recall metric. Four out of five teams assigned to this playstyle, were indeed correctly labelled.
Moreover, *Possession Play* could achieve a 100% precision score as well, but the prediction was lacking regarding the recall, with just 67%. One out of three teams was assigned to a wrong playstyle.
Lastly, *Pressing/Aggressive Defending* has the lowest precision and F1-score values across all playstyles. Clearly, the Random Forest model performed the worst in that regard.

Figure 19 *Random Forest Confusion Matrix*



Furthermore, the SHAP analysis in Figure 20 again shows similar insights to XGBoost. However, this was expected as the underlying logic for the PCs did not change. PC1 and PC2 are both by far the factors contributing the most to the classification logic of the algorithm. Although it is to be noted, that PC1 contributes the most to *Balanced*, whereas PC2 mainly contributes to the differentiation between *Defensive & Structured* and *Pressing/Aggressive Defending*. The rest of the PCs do only contribute in small margins, mainly due to these factors explaining less and less of the total variance.

Figure 20 *SHAP Analysis - Random Forest*

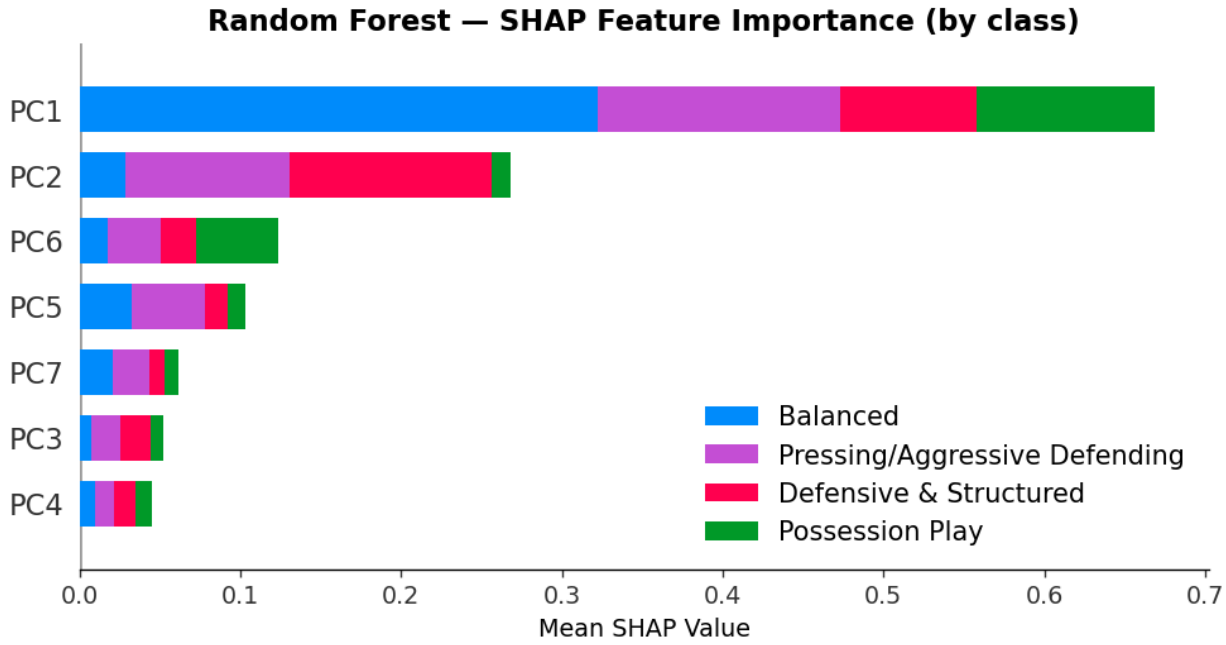


Table 11 *Classification Table of Random Forest*

Playstyle	Precision	Recall	F1-score
Defensive & Structured	1.00	0.8	0.89
Balanced	0.91	0.91	0.91
Possession Play	1.0	0.67	0.8
Pressing/Aggressive Defending	0.57	0.8	0.67
Overall Accuracy	0.89		
Test set accuracy	0.83		

5.4 Performance Analysis by opponent playstyle

For further analysis which builds upon the previously identified playstyles, a descriptive performance analysis was conducted using a single team of choice. FC Barcelona was chosen as it is one of the most well-known and successful soccer teams in the entire world. Furthermore, they are one of the few teams which are part of the *Possession Play* cluster. Evidently, they are

suitable for a case study on how the performance of an elite, possession-focused team changes based on the opponents they are facing by their respective playstyles.

5.4.1 Data collection

Furthermore, for this performance analysis a new additional dataset was needed to have access to individual match data from FC Barcelona across all 38 matches from the full LaLiga season in 2024/25. StatsHub, which already contributed a major portion of data to the previous playstyle analysis, offered the most complete and insightful dataset for this task. Hence, this database was chosen as the sole source. As aforementioned in the methodology section, web scraping is not possible on this website, thus it required a manual extraction of the data to Excel. Followingly, manual cleaning of the data was required too, including linking match statistics to the correct teams, while also ensuring consistency in the naming with the previous approach used in the playstyle clustering.
Figure 21 highlights a selection of the available performance metrics such as goals, expected goals, possessions, shots on target, passes and clearances, among others. Also, it is to be noted that the metrics of opponents were removed, as this performance analysis is built upon the previous playstyle clustering which matches the metrics of the corresponding playstyles to the teams.

Figure 21 *Data snippet from the individual match dataset*

Stat Type	AVG	FOR	AGT	0-3 A 25 May	2-3 H 18 May	0-2 A 15 May	4-3 H 11 May	1-2 A 03 May	1-0 H 22 Apr	4-3 H 19 Apr	0-1 A 12 Apr	1-1 H 05 Apr	4-1 H 30 Mar	3-0 H 27 Mar	2-4 A 16 Mar	4-0 H 02 Mar	0-2 A 22 Feb	1-0 H 17 Feb	1-4 A 09 Feb
Goals	3.71	2.68	1.03	3 0	2 3	2 0	4 3	2 1	1 0	4 3	1 0	1 1	4 1	3 0	4 2	4 0	2 0	1 0	4 1
Corners	10.89	6.97	3.92	6 2	11 3	3 3	10 2	8 4	13 2	4 1	8 4	9 3	10 2	8 4	6 6	12 0	3 5	4 5	6 5
Cards	4.13	1.74	2.39	0 1	2 0	1 5	4 3	2 3	0 1	2 2	1 2	2 1	0 1	1 3	3 1	0 1	1 1	2 3	3 5
Crosses	6.92	3.89	3.03	2 3	3 2	3 5	5 1	7 4	4 4	6 2	5 3	4 1	2 3	1 4	4 2	4 1	5 0	3 4	2 2
Big Chance Created	6.32	4.24	2.08	5 1	2 5	0 3	7 5	3 2	4 1	4 5	2 2	2 0	3 1	6 1	4 2	5 0	2 0	7 3	3 1
Big Chance Missed	3.74	2.45	1.29	2 1	2 2	0 3	4 2	2 2	4 1	1 3	2 2	1 0	1 0	3 1	3 0	2 0	0 0	6 3	0 0
Big Chance Scored	2.58	1.79	0.79	3 0	0 3	0 0	3 3	1 0	0 0	3 2	0 0	1 0	2 1	3 0	1 2	3 0	2 0	1 0	3 1
Expected Goals (xG)	3.53	2.43	1.10	3.49 1.20	1.37 1.75	0.63 1.39	4.26 2.74	1.93 0.59	3.45 0.25	2.46 1.97	1.35 0.72	1.31 0.31	2.63 0.49	3.72 0.40	1.26 1.36	2.85 0	1.47 0.47	2.56 0.97	1.93 0.79
Shots On Target	9.47	6.58	2.89	4 1	9 3	4 1	9 5	7 4	12 0	6 8	3 1	5 2	9 2	5 0	6 2	10 0	5 2	5 4	6 4
Shots In The Box	17.26	11.47	5.79	9 6	10 7	6 7	15 7	15 5	26 4	10 11	10 4	8 3	15 4	12 4	7 4	22 0	11 6	14 8	7 6
Total Shots	25.58	17.84	7.74	13 9	24 7	12 9	23 9	24 5	40 4	18 11	12 6	13 6	21 4	19 4	13 7	33 0	15 10	14 9	10 8
Shots Outside The Box	8.32	6.37	1.95	4 3	14 0	6 2	8 2	9 0	14 0	8 0	2 2	5 3	6 0	7 0	6 3	11 0	4 4	0 1	3 2
Clearances	40.74	14.61	26.13	14 19	10 35	13 16	9 28	9 36	20 45	12 31	30 27	6 44	10 29	13 22	21 22	5 33	24 24	18 11	14 29
Dispossessed	16.13	8.24	6.89	5 6	4 7	6 6	16 9	7 9	10 8	7 3	6 13	7 6	5 7	10 10	9 6	7 6	8 7	10 5	6 8
Errors Lead To Goal	0.50	0.08	0.42	0 1	0 0	0 0	0 3	0 1	0 0	1 0	0 0	0 0	0 0	0 0	0 1	0 1	0 1	0 0	0 0
Errors Lead To Shot	0.76	0.21	0.55	0 1	0 1	0 1	0 0	0 0	0 1	0 0	0 0	0 1	0 3	0 0	0 1	0 0	0 2	0 3	0 0

5.4.2 Performance Analysis of FC Barcelona

Following that, for each match the respective team was assigned to the playstyle cluster they were assigned to in the previous analysis. Accordingly, the distribution for LaLiga, considering the amount of matches they played against a specific playstyle can be seen in Table 12.

Table 12 *FC Barcelona playing matches against different playstyles - Overview*

Playstyle	Amount of played matches
Balanced	18
Pressing/Aggressive Defending	10
Defensive & Structured	8
Possession Play	2

It becomes apparent, that this is a much smaller sample size in comparison, as such there are only two matches played against teams of the playstyle *Possession Play*. This is due to only one other team besides FC Barcelona being part of that cluster, namely Real Madrid. Thus, the

following results should be seen with caution.

Quite several interesting insights can be derived from the descriptive analysis. The highest mean of expected goals (3.42) and shots on target (8.0) were recorded against *Possession Play*. This could be due to both teams relying on ball control and having high attacking output. Likewise, the possession of FC Barcelona dropped the most while playing against this playstyle.
In comparison, against *Pressing/Aggressive Defending*, FC Barcelona performed the worst in multiple important attacking metrics such as expected goals (2.18), total shots (15.9) and corners (6.3). Opponents in that playstyle cluster seem to be more effective in shutting down Barcelona's attacking capabilities than other teams. In addition, the mean number of fouls was also the highest recorded across the board at 10.6.
Against the *Balanced* playstyle, Barcelona's possession remained rather uncontested in comparison (72.83%) which is also reflected in the recorded passes mean value (666.17).
Lastly, facing the *Defensive & Structured* playstyle, there is an interesting difference between the high goals value (3.25) and low expected goals value (2.25), which is the lowest overall. One explanation is that Barcelona was just especially clinical in their chance conversion in these matches.
Moreover, Figure 22 plots the metrics of expected goals and possession, against the four distinct playstyles, including an error bar which shows the mean ± one standard deviation. Additionally, Figure 23 shows the performance of FC Barcelona as a bar chart with their respective average points per game per opposing playstyle.

Figure 22 *FC Barcelona performance metrics by opponent's playstyle*

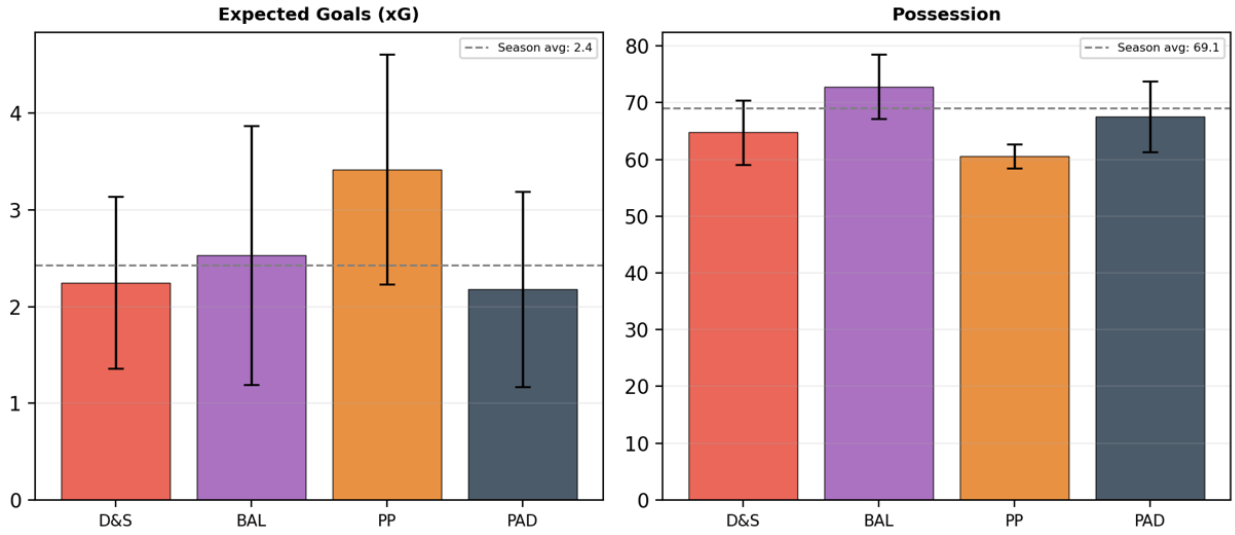
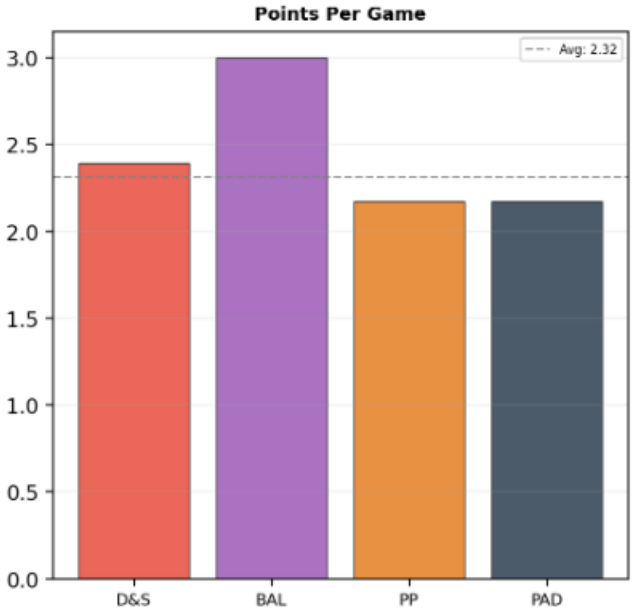


Figure 23 *FC Barcelona points per game by opponent's playstyle*



Moreover, an OLS regression model was applied to further investigate if an opponent's playstyle does have a statistically significant influence on FC Barcelona's performance. Hereby, the *Balanced* playstyle was used as the reference category, meaning the coefficients for the three other playstyles denominate the predicted deviation values from the chosen constant baseline. In addition, as there is data available for home and away matches, it was included as a binary control variable. Home advantage is a well-known and researched phenomenon in the sport of soccer. In

particular, emphasis was put on four distinctively selected target variables: expected goals, possession, passes and shots on target.

Figure 24 and Figure 25 showcase interesting results across all four chosen target variables. Looking at possession, it has the highest R^2 value across the board ($R^2 = 0.414$, Adjusted $R^2 = 0.343$), while also all three playstyles have statistically significant deviations with a p-value < 0.05. Furthermore, the regression shows that against *Defensive & Structured* opponents, possession will drop by 8.1% ($p = 0.001$), against *Possession Play* by 12.3% ($p = 0.005$) and against *Pressing/Aggressive Defending* by 5.3% ($p = 0.019$). Home advantage could also be identified as statistically significant leading to an increase in possession by 3.6% ($p = 0.049$).
For a possession-heavy team like FC Barcelona there are also significant predictions observable for the target variable passes. Playing against *Possession Play* leads to 199 fewer passes across a match ($p < 0.001$). Also, the R^2 is quite high as before for possession ($R^2 = 0.392$, Adjusted $R^2 = 0.318$). However, it should be noted again that the *Possession Play* playstyle was by far the smallest with $n = 2$ (whereas Real Madrid is the sole unique opponent).
Shots on target show a bit weaker explanatory R^2 (0.226) and no statistically significant deviations across the playstyles, except for the home advantage ($p = 0.035$).
Interestingly, the OLS regression performed the worst for the variable expected goals ($R^2 = 0.13$) with no statistically significant observations at all.

Figure 24 *OLS Regression - Predicted Means by Opponent Playstyle*

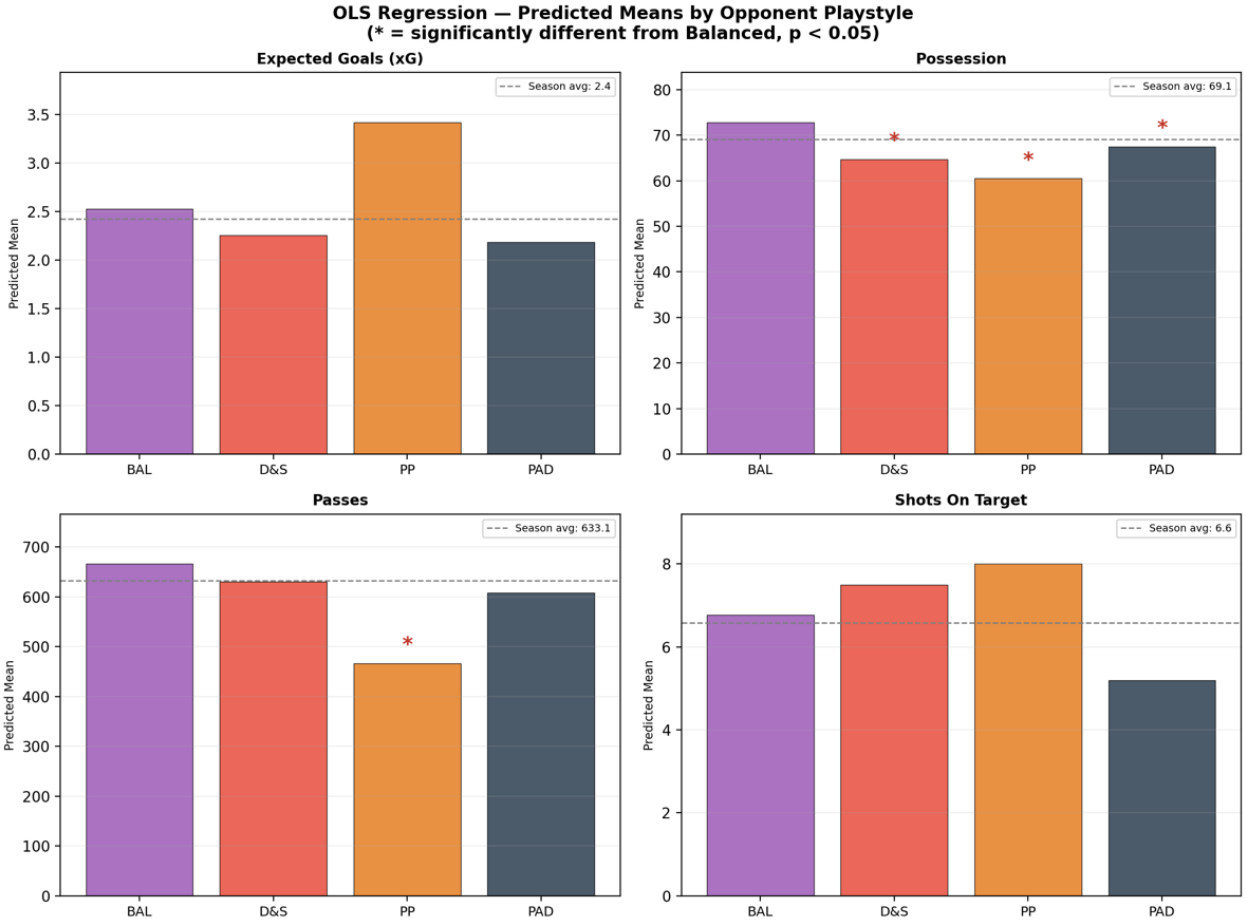
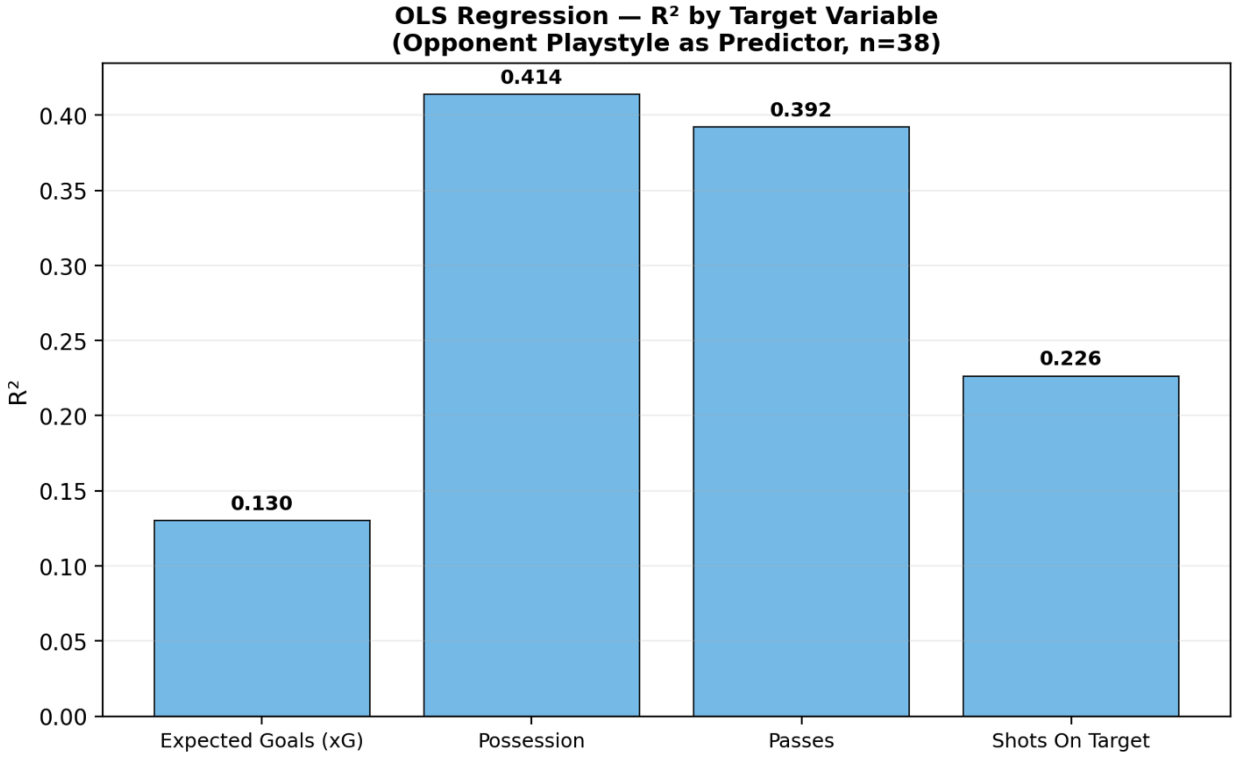


Figure 25 *OLS Regression - R² by Target Variable*



As aforementioned, the *Balanced* playstyle acted as the baseline, while the regression predicted the deviations for the other playstyles. Consequently, the fully detailed results of the OLS regression for each target variable, including coefficients, standard errors, t-statistics, p-values and confidence intervals are presented in Table 13, 14, 15 and 16.

Table 13 *OLS Regression - Target Variable: Possession*

Predictor	Coef.	Std. Err.	t	p	[0.025	0.975]
Constant	71.02	1.57	45.23	<0.001	67.82	74.21
Defensive & Structured	-8.08	2.33	-3.47	0.001	-12.83	-3.34
Possession Play	-12.33	4.09	-3.02	0.005	-20.65	-4.01
Pressing/Aggressive Defending	-5.33	2.16	-2.46	0.019	-9.74	-0.93
Home (vs Away)	3.63	1.78	2.04	0.049	0.01	7.25

R^2 = 0.414, Adjusted R^2 = 0.343. Reference category: Balanced. n = 38

Table 14 OLS Regression - Target Variable: Expected Goals (xG)

Predictor	Coef.	Std. Err.	t	p	[0.025	0.975]
Constant	2.23	0.33	6.79	<0.001	1.56	2.89
Defensive & Structured	-0.28	0.49	-0.57	0.576	-1.27	0.72
Possession Play	0.89	0.85	1.04	0.304	-0.85	2.63
Pressing/Aggressive Defending	-0.35	0.45	-0.77	0.449	-1.27	0.57
Home (vs Away)	0.61	0.37	1.63	0.113	-0.15	1.36

R^2 = 0.130, Adjusted R^2 = 0.025. Reference category: Balanced. n = 38.

Table 15 *OLS Regression - Target Variable: Passes*

Predictor	Coef.	Std. Err.	t	p	[0.025	0.975]
Constant	635.69	21.13	30.08	<0.001	592.70	678.68
Defensive & Structured	-35.92	31.38	-1.14	0.261	-99.76	27.93
Possession Play	-199.17	55.04	-3.62	<0.001	-311.16	-87.18
Pressing/Aggressive Defending	-57.27	29.13	-1.97	0.058	-116.53	1.99
Home (vs Away)	60.95	23.96	2.54	0.016	12.20	109.69

R^2 = 0.392, Adjusted R^2 = 0.318. Reference category: Balanced. n = 38.

Table 16 *OLS Regression - Target Variable: Shots on Target*

Predictor	Coef.	Std. Err.	t	p	[0.025	0.975]
Constant	5.88	0.72	8.18	<0.001	4.42	7.35
Defensive & Structured	0.72	1.07	0.68	0.503	-1.45	2.89
Possession Play	1.22	1.87	0.65	0.518	-2.59	5.03
Pressing/Aggressive Defending	-1.58	0.99	-1.59	0.121	-3.59	0.44
Home (vs Away)	1.79	0.82	2.20	0.035	0.13	3.45

R^2 = 0.226, Adjusted R^2 = 0.133. Reference category: Balanced. n = 38.

Ultimately, this analysis showed that the performance of FC Barcelona does indeed vary meaningfully while facing teams of different playstyles. Firstly, the descriptive analysis showcased notable differences across several important performance indicators. Followingly, the OLS regression could find statistically significant deviations across four key metrics while FC

Barcelona played against a different playstyle. Overall, this section illustrates that the prior playstyle clustering does translate into an analysis capturing performance changes for different opponents.

5.5 Feature-Level Interpretability Analysis

While the aforementioned classification models were trained on the output of the PCA, a further analysis was run based on the same two models but with the original selected dataset as the input. The reason for this, was to provide further interpretability. It is to be noted, that these two new models achieved a similar or slightly higher test accuracy across the board. Hence, these two new models further prove the robustness of the overall framework. Moreover, training the models on the original features before applying PCA, paints a much clearer picture regarding the SHAP analysis, as it does not solely show the few different principal components, but the direct name of the most influential metrics.

Looking at Figure 26 and Figure 27, the plots clearly show that possession is one of the most dominant features for the playstyles in both models. This finding is further supported by the literature, as both Lago-Peñas et al. (2017) and Castellano and Pic (2019) could identify possession as one key indicator to distinguish between different playstyles in soccer.
Likewise, the SHAP analysis shows that *key_passes* and *corners_avg_for* occur as some of the most important features across both models. These factors are linked to the attacking output of teams, which showcases that the abilities of a team to create chances and goal scoring opportunities are indeed relevant for differentiating between playstyles.
Additionally, as one would expect, defensive metrics, in particular *avg_interceptions* and *ball_recovery*, are the most influential metrics for the two identified rather defensive playstyles.

Figure 26 *SHAP Analysis - Random Forest (Original Features)*

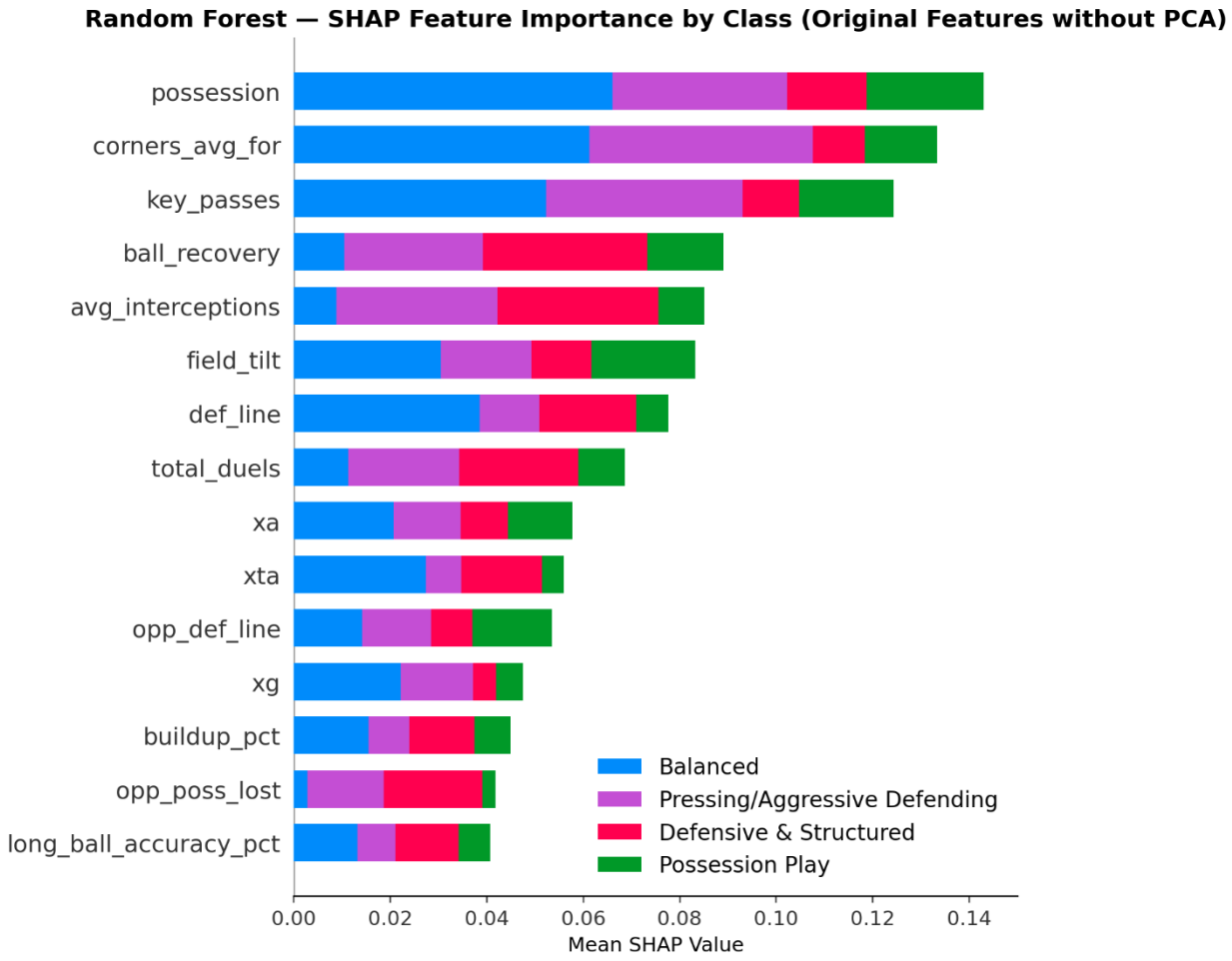
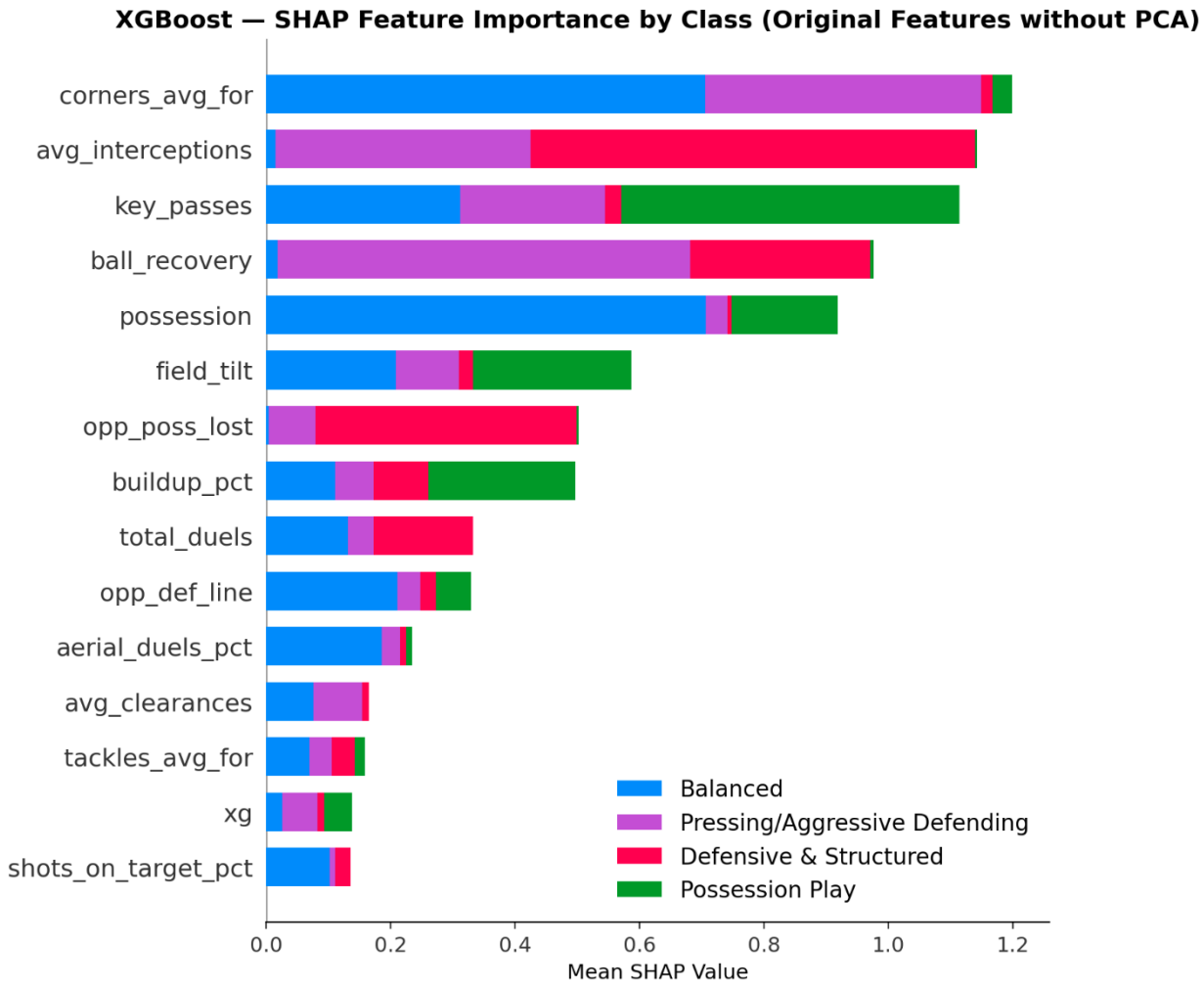


Figure 27 *SHAP Analysis - XGBoost (Original Features)*



In conclusion, this SHAP analysis on a detailed feature level offers much more interpretability compared to the previous models using the PCA results as the input. Possession, attacking metrics and defensive metrics are hereby the key differentiators across the four playstyles.

5.6 Model comparison

Now that both models, namely XGBoost and Random Forest, have been applied to the same classification task, a direct comparison between them offers meaningful insights into the performance and suitability of them for this type of problem. On the one hand, both models perform very similarly in terms of the overall cross validated accuracy with 90.48% and 89.14%

respectively, whereas on the other hand accounting for the test set accuracy too, makes a more nuanced interpretation possible.
In this metric Random Forest recorded an accuracy of 83.33%, whereas XGBoost achieved an accuracy of 75%. Given that a 5-fold validation was applied to investigate the ability of the models to generalize for unseen data, this metric clearly favours Random Forest as the superior model of choice.

Moreover, incorporating the F1-scores reinforces this judgement, as on a per-playstyle case, Random Forest performs far better in two out of four instances and performs equally in the other two. In particular, Random Forest recorded an F1-score of 0.89 for the playstyle Defensive & Structured, while XGBoost lags behind with 0.56. In the case of Possession Play, Random Forest also performed superiorly with 0.8 over 0.57. This consistent overperformance of Random Forest is noteworthy and underlines its robustness, as this affects the minority clusters, and such smaller clusters are usually harder to classify correctly.

Furthermore, the results are in line with the general characteristics of these supervised algorithms, as has been mentioned before in the Methodology section. Random Forest is indeed well suited for smaller and structured datasets, as it is an ensemble method producing a certain amount of independent decision trees, whose outputs are then aggregated with a majority voting system. This independence of the decision trees reduces the variance and diminishes overfitting in general. It is a machine learning algorithm that can generalize well.
In comparison, XGBoost is a sequential gradient boost algorithm that is operating iteratively. It achieves better and better results by correcting errors of prior computed trees. Evidently this algorithm is particularly well suited for enormous datasets, whereas it can tend to overfit in smaller datasets. While the overall accuracy was on an even playing field, the accuracy of the test set further supports this tendency. Additionally, the four playstyle clusters are imbalanced, which can further hinder the capabilities of XGBoost, as it operates through sequential learning and misclassifications of the smaller clusters can lead to mistakes in the later stages of the algorithm.

Generally, Random Forest could be identified as the better model of choice for the playstyle classification task in this thesis. It performs better with unseen data and provides more consistent results across the data. A clear overview of the model comparison can be seen in Table 17.

Table 17 *Model comparison*

Metric	Random Forest	XGBoost
Overall Accuracy	89%	90%
Test Set Accuracy	83%	75%
Best F1-Score	0.91 (Balanced)	0.91 (Balanced)
Worst F1-Score	0.67 (Pressing/Aggressive Defending)	0.56 (Defensive & Structured)

6. Discussion

Overall, the aim of this thesis was to investigate whether opponents' playstyles in soccer can be reliably and accurately predicted with the help of data science methods as can be seen in the research question:

"Can a data-driven machine learning model reliably and accurately predict opponent's playstyles?"

Hereby, this research question can be answered positively as the results presented in the previous chapters affirmatively confirm that this is indeed possible. By combining an unsupervised clustering of all soccer teams across the top 5 leagues with a supervised classification, four distinct and interpretable playstyles could be identified. The Random Forest model could predict these playstyles with a test accuracy of 83.33%. Followingly, this discussion section interprets the findings and not only relates them to the existing literature, but also outlines the new insights this thesis demonstrates.

Evidently, one major finding of the clustering analysis is the impactful role of possession in differentiating playstyles. Just the first principal component, labelled Possession & Attacking Dominance, accounted for 39.43% of the total variance by itself, which is substantially higher than the other ones. This finding is strongly supported by prior research as both Lago-Peñas et al. (2017) and Castellano and Pic (2019) identified possession as the most influential feature for distinguishing between playstyles in soccer. Additionally, the SHAP analysis operating on a feature level reinforced this observation, as possession was one of the most influential features for the classification task in the Random Forest and the XGBoost algorithm. Hence, the presented results of this thesis do indeed converge with prior existing literature, even though a completely different dataset, a larger number of features and a much broader geographical scope were used.

In addition, the analysis settled on four playstyle clusters as the best choice, which also corresponds to the findings of Castellano and Pic (2019), who could likewise identify four playstyles in their analysis of the Spanish LaLiga. Looking at the theoretical optimal solution of the silhouette score, the optimal *k* would be 2, but this binary split does not make sense practically, as the

analysis would not offer any meaningful insights. Thus, a *k* of 4 is a justifiable trade-off in a context where the results should be able to offer interesting insights and real-world applicability. Also, it is to be noted that the four identified distinct clusters partially overlap with the playstyle dimensions laid out by Peng et al. (2025), which also heavily influenced the naming conventions of the playstyles. However, some divergences could be observed, but this was expected and can be attributed to a differing methodology and an entirely different dataset consisting of national teams in international competitions, whereas this thesis focused on the top 5 leagues in Europe.

Interestingly, the clustering shows findings that do not just align with quantitative studies, but also widely regarded knowledge by domain experts. In particular, the *Defensive & Structured* playstyle cluster consists majorly of teams from the Italian Serie A, with 13 out of 19 teams. This is consistent with Serie A's historical reputation of having a more defensive tactical approach. Even more, the clustering corresponded to these established characteristics without any prior labelling (Sarmento et al., 2013).

Furthermore, comparing the predictive modelling of Random Forest and XGBoost yielded results that are in line with their theoretical characteristics. While both performed very well in regard to the cross-validated accuracy, Random Forest showed much better results in the test set with 83.33% against just 75%. Also, Random Forest recorded superior F1-scores. As Random Forest is an ensemble algorithm working with independent decision trees, it can generalize better on smaller datasets, whereas XGBoost works on a sequential basis and is at risk of overfitting with imbalanced datasets, which was the case here. Similarly, Moustakidis et al. (2023) also employed the use of XGBoost and compared it with other algorithms, amongst them Random Forest.

Besides the playstyle prediction, this thesis also provides a performance case study on FC Barcelona, which showcases that an opponent's playstyle does indeed have a statistically significant influence on a team's performance. Hereby, the OLS regression uncovered that possession varied meaningfully depending on the opponent. It also has the largest share of variance among all target variables ($R^2 = 0.414$). Likewise, the number of passes dropped substantially against opponents of the same playstyle FC Barcelona belongs to, which is *Possession Play*. Correspondingly Immler et al. (2021) observed that the playstyle of opponents changes the tactical behaviour of a

team. Forcher et al. (2023) found that an offensive playstyle influences the physical and technical match performance. Whereas in prior studies it was examined how a team's own playstyle influences their performance, this thesis adds another perspective building on the playstyle clustering. This outcome of how a team's performances changes based on their opponent's playstyles has received comparatively little attention in the literature at hand.

Looking back at the other research question of:

"How has Data Science been applied in different areas of soccer and what is its usage?"

The literature review in this thesis demonstrated that the application is not only very broad but has also been rapidly expanding over the years. Data science methods are employed across a wide range of domains, like the identification of playstyles, the analysis of passing networks, the development of AI-powered tactical assistants or the construction of new performance indicators such as expected goals. Across the literature, one consistent pattern emerges. Overall, data science is used to produce objective and measurable insights across all areas of the sport. Hence, this research question can be answered comprehensively.

6.1 Limitations

While several new contributions could be highlighted, the limitations of this thesis must be acknowledged too. Evidently, the first constraint is regarding the size of the dataset. The analysis was conducted on 96 different teams across one league season. However, this limitation can be attributed to the domain of soccer analytics in general. As evidenced in the literature, soccer teams only play a limited number of matches per season, and playstyles and other factors can change significantly across seasons. Thus, aggregating multiple seasons would risk obscuring any meaningful distinct playstyles and with them the ability to draw conclusions. Accordingly, limiting the dataset to the chosen size of one season is justified. Moreover, it should be pointed out that this thesis uses a considerably larger and more diverse dataset than other prior research as was the case with Castellano and Pic (2019). Still, in the realm of machine learning algorithms it is a smaller dataset which is reflected in the imbalance of the playstyle clusters. For example, *Possession Play* only consists out of 12 teams from 96.

This becomes apparent in the performance case study, as FC Barcelona faced only a single opponent from that playstyle, namely Real Madrid. Consequently, the findings from the OLS regression must be seen with caution, despite the statistical significance. In addition, the case study is based on just a single team, thus the findings cannot be generalized across all teams without further analysis.

Davis et al. (2024) discussed that there is no definitive ground truth in soccer, as the exact tactical plans of teams are unknown to outsiders. The presented findings are a data-driven approximation of the reality and not an objective fact.

6.2 Future Research

Hereby, this thesis offers several new possibilities for future research. One aspect that can be looked into, is that the used framework could be applied over multiple seasons individually to investigate the evolution of playstyles instead of capturing the static state of just one season. Each season would be looked at and analysed separately to then compare how the identified clusters change over time.

Additionally, this thesis only used two supervised machine learning algorithms. As evidenced in the literature, there is a wider range of algorithms applicable to this domain. Even more so, Mills et al. (2024) employed a voting ensemble combining algorithms to reach their best results. Looking into how new and different methods can be applied, could further improve the prediction of playstyles.
Likewise, the applied framework is lacking extensive feature engineering. Making use of feature engineering, supported by domain experts, could help researchers refine the clustering and analysis further.

Furthermore, as aforementioned the performance case study was applied on a single team. Naturally, conducting this analysis across all teams within a soccer dataset would allow for more generalizable conclusions regarding the performance changes per playstyles and whether a certain playstyle is better suited for playing against one another.

Finally, the hereby developed methodological framework is not limited to the used top 5 European leagues. Applying the same framework or enhancing it with the proposed improvements, on other competitions in soccer such as other leagues within or outside Europe, can lead to novel insights. Even more, it would test the robustness and transferability of this framework.

7. Conclusion

In conclusion, this master's thesis investigates the application of data science in soccer, with a focus on the prediction of opponents' playstyles. Regarding the first research question, a comprehensive literature review offers an overview of how data science is applied across a wide range of use-cases. This includes not only the analysis of passing networks but also the creation of AI-driven assistants.
The second research question can be positively answered with a developed and applied analysis framework that can reliably and accurately predict opponents' playstyles.

Through basing the analysis on existing literature and building upon it, an unsupervised k-means clustering with a following supervised classification was conducted across the top 5 European leagues. Consequently, four distinct playstyles could be identified, namely Defensive & Structured, Balanced, Possession Play and Pressing/Aggressive Defending. A comparison of the models showed that Random Forest was the superior one, which achieved an accuracy of 83.33%. Additionally, a SHAP analysis reveals possession as one decisive feature for distinguishing between playstyles, which is in line with the existing literature. Lastly, a performance case study conducted with FC Barcelona uncovers a significant statistical influence for changes in the performance based on the opponent's playstyle.

All things considered, this thesis presents a comprehensive analysis and adds several new contributions. Firstly, it addresses the geographical limitation and diversity that encompasses prior research, which has frequently focused on a single competition or a smaller subset, as it was the case for LaLiga in Castellano and Pic (2019) or Lago-Peñas et al. (2017) with the Chinese soccer league. Through looking at all top 5 European soccer leagues together at once, this thesis ultimately offers a more diverse and generalizable basis for further analysis. Mills et al. (2024) call out for future research to include more soccer leagues with different playstyles for more comprehensive insights, which this thesis directly addresses.
Moreover, the used methodology with a combination of unsupervised clustering and a following supervised prediction, which is supported by several evaluation metrics and a performance case study, provides a robust and novel end-to-end framework that has not been applied in this scope before.

Ultimately, this thesis demonstrates that the combination of unsupervised and supervised machine learning enables the prediction of playstyles and offers interesting new insights in the tactical domain of soccer. Beyond its academic contribution, the developed analytical framework provides a basis for future research to build upon.

Bibliography

Branquinho, L., França, E., Teixeira, J. E., Valente, N., Reis, T., Thomatieli-Santos, R. V., Forte, P., & Ferraz, R. (2024). Comparing physical, technical and tactical performances in the World Cup Qatar 2022. *Journal of Human Sport and Exercise*, *19*(2), 654–666. https://doi.org/10.55860/t8ybtk14

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burton, A. L. (2021). OLS (Linear) Regression. In J. C. Barnes & D. R. Forde (Eds), *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (1st edn, pp. 509–514). Wiley. https://doi.org/10.1002/9781119111931.ch104

Castellano, J., & Pic, M. (2019). Identification and Preference of Game Styles in LaLiga Associated with Match Outcomes. *International Journal of Environmental Research and Public Health*, *16*(24), 5090. https://doi.org/10.3390/ijerph16245090

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.48550/ARXIV.1603.02754

Clijmans, J., Van Roy, M., & Davis, J. (2023). Looking Beyond the Past: Analyzing the Intrinsic Playing Style of Soccer Teams. In M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, & G. Tsoumakas (Eds), *Machine Learning and Knowledge Discovery in Databases* (Vol. 13718, pp. 370–385). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26422-1_23

Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J., & Van Roy, M. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning*, *113*(9), 6977–7010. https://doi.org/10.1007/s10994-024-06585-0

Decroos, T., & Davis, J. (2020). Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams. In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, & C. Robardet (Eds), *Machine Learning and Knowledge Discovery in*

Databases (Vol. 11908, pp. 569–584). Springer International Publishing. https://doi.org/10.1007/978-3-030-46133-1_34

FBref. (n.d.). *2024-2025 Premier League Stats*. Retrieved 20 February 2026, from https://fbref.com/en/comps/9/2024-2025/2024-2025-Premier-League-Stats

Fernández-Cortés, Carlos D. Gómez-Carmona, Javier García-Rubio, & Sergio J. Ibáñez. (2024). Is crowd support important in professional football's home advantage? A systematic review based on covid-19 effect. *E-Balonmano Com Journal Sports Science*, *20*(2), 169–188. https://doi.org/10.17398/1885-7019.20.169

Forcher, L., Forcher, L., Wäsche, H., Jekauc, D., Woll, A., Gross, T., & Altmann, S. (2023). Is ball-possession style more physically demanding than counter-attacking? The influence of playing style on match performance in professional soccer. *Frontiers in Psychology*, *14*, 1197039. https://doi.org/10.3389/fpsyg.2023.1197039

GeeksforGeeks. (2025, July 5). *Elbow Method vs. Silhouette Score: Which is better?* https://www.geeksforgeeks.org/machine-learning/elbow-method-vs-silhouette-score-which-is-better/

GeeksforGeeks. (2026, January 19). *What is Silhouette Score*. https://www.geeksforgeeks.org/machine-learning/what-is-silhouette-score/

Google Developers. (2026, January 12). *Classification: Accuracy, recall, precision, and related metrics*. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Hewitt, A., Greenham, G., & Norton, K. (2016). Game style in soccer: What is it and can we quantify it? *International Journal of Performance Analysis in Sport*, *16*(1), 355–372. https://doi.org/10.1080/24748668.2016.11868892

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, *108*(1), 29–47. https://doi.org/10.1007/s10994-018-5704-6

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Immler, S., Rappelsberger, P., Baca, A., & Exel, J. (2021). Guardiola, Klopp, and Pochettino: The Purveyors of What? The Use of Passing Network Analysis to Identify and Compare Coaching Styles in Professional Football. *Frontiers in Sports and Active Living*, *3*, 725554. https://doi.org/10.3389/fspor.2021.725554

Jason Brownlee. (2023, October 4). A Gentle Introduction to k-fold Cross-Validation. *Machine Learning Mastery*. https://machinelearningmastery.com/k-fold-cross-validation/

Kovalchik, S. A. (2023). Player Tracking Data in Sports. *Annual Review of Statistics and Its Application*, *10*(1), 677–697. https://doi.org/10.1146/annurev-statistics-033021-110117

Laerd Statistics. (n.d.). Pearson Product-Moment Correlation. *Laerd Statistics*. Retrieved 6 March 2026, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Lago-Peñas, C., Gómez-Ruano, M., & Yang, G. (2017). Styles of play in professional soccer: An approach of the Chinese Soccer Super League. *International Journal of Performance Analysis in Sport*, *17*(6), 1073–1084. https://doi.org/10.1080/24748668.2018.1431857

Li, C., & Zhao, Y. (2021). Comparison of Goal Scoring Patterns in "The Big Five" European Football Leagues. *Frontiers in Psychology*, *11*, 619304. https://doi.org/10.3389/fpsyg.2020.619304

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.07874

markstats. (n.d.). *PREMIER LEAGUE 2024-2025*. Retrieved 20 February 2026, from https://markstats.club/england/2024-2025/teams

Mills, E. F. E., Deng, Z., Zhong, Z., & Li, J. (2024). Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques. *Journal of Big Data, 11*(1), 170. https://doi.org/10.1186/s40537-024-01008-2

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. *Future Internet*, *15*(5), 174. https://doi.org/10.3390/fi15050174

Peng, W., Ribeiro, J., Wang, M., & Gómez-Ruano, M. Á. (2025). Evolution of playing styles in UEFA European championships: Trends from 2012 to 2024. *International Journal of Performance Analysis in Sport*, 1–16. https://doi.org/10.1080/24748668.2025.2536954

Plakias, S., Moustakidis, S., Kokkotis, C., Tsatalas, T., Papalexi, M., Plakias, D., Giakas, G., & Tsaopoulos, D. (2023). Identifying Soccer Teams' Styles of Play: A Scoping and Critical Review. *Journal of Functional Morphology and Kinesiology*, *8*(2), 39. https://doi.org/10.3390/jfmk8020039

Sanyal, S. (2021). Who will receive the ball? Predicting pass recipient in soccer videos. *Journal of Visual Communication and Image Representation*, *78*, 103190. https://doi.org/10.1016/j.jvcir.2021.103190

Sarmento, H., Pereira, A., Matos, N., Campaniço, J., Anguera, T. M., & Leitão, J. (2013). English Premier League, Spaiṅs La Liga and Italýs Seriés A – What's Different? *International Journal of Performance Analysis in Sport, 13*(3), 773–789. https://doi.org/10.1080/24748668.2013.11868688

StatsBomb. (2021). *Introducing On-Ball Value (OBV)*. https://www.hudl.com/blog/introducing-on-ball-value-obv

StatsHub. (n.d.). *PREMIER LEAGUE*. Retrieved 20 February 2026, from https://www.statshub.com/league/premier-league/1

StatsPerform. (n.d.). Assessing The Performance of Premier League Goalscorers. *Assessing The Performance of Premier League Goalscorers*. Retrieved https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/

Twomey. (2026, January 28). FBref and Opta: The data break-up that sent soccer's analytics world into meltdown. *The Athletic*. https://www.nytimes.com/athletic/7002196/2026/01/28/fbref-opta-football-data-soccer-analytics/

Van Bommel, M., & Bornn, L. (2017). Adjusting for scorekeeper bias in NBA box scores. *Data Mining and Knowledge Discovery*, *31*(6), 1622–1642. https://doi.org/10.1007/s10618-017-0497-y

Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., Bachrach, Y., Elie, R., Wenliang, L. K., Piccinini, F., Spearman, W., Graham, I., Connor, J., Yang, Y., Recasens, A., Khan, M., Beauguerlange, N., Sprechmann, P., Moreno, P., … Tuyls, K. (2024). TacticAI: An AI assistant for football tactics. *Nature Communications*, *15*(1), 1906. https://doi.org/10.1038/s41467-024-45965-x

Yi, Q., Gómez, M. A., Wang, L., Huang, G., Zhang, H., & Liu, H. (2019). Technical and physical match performance of teams in the 2018 FIFA World Cup: Effects of two different playing styles. *Journal of Sports Sciences*, *37*(22), 2569–2577. https://doi.org/10.1080/02640414.2019.1648120

AI Acknowledgments

AI tools were used in a supporting capacity during the development of this thesis. Claude and Microsoft Copilot were used to assist with debugging and improving the Python code. Moreover, Claude was used to help with finding typographical and grammatical errors in the text. All AI suggestions were critically reviewed and verified by the author and in many cases substantially revised. The intellectual content and analysis of this thesis are the author’s own.

Appendix

Figure 28 *Additional clustering with k = 5*

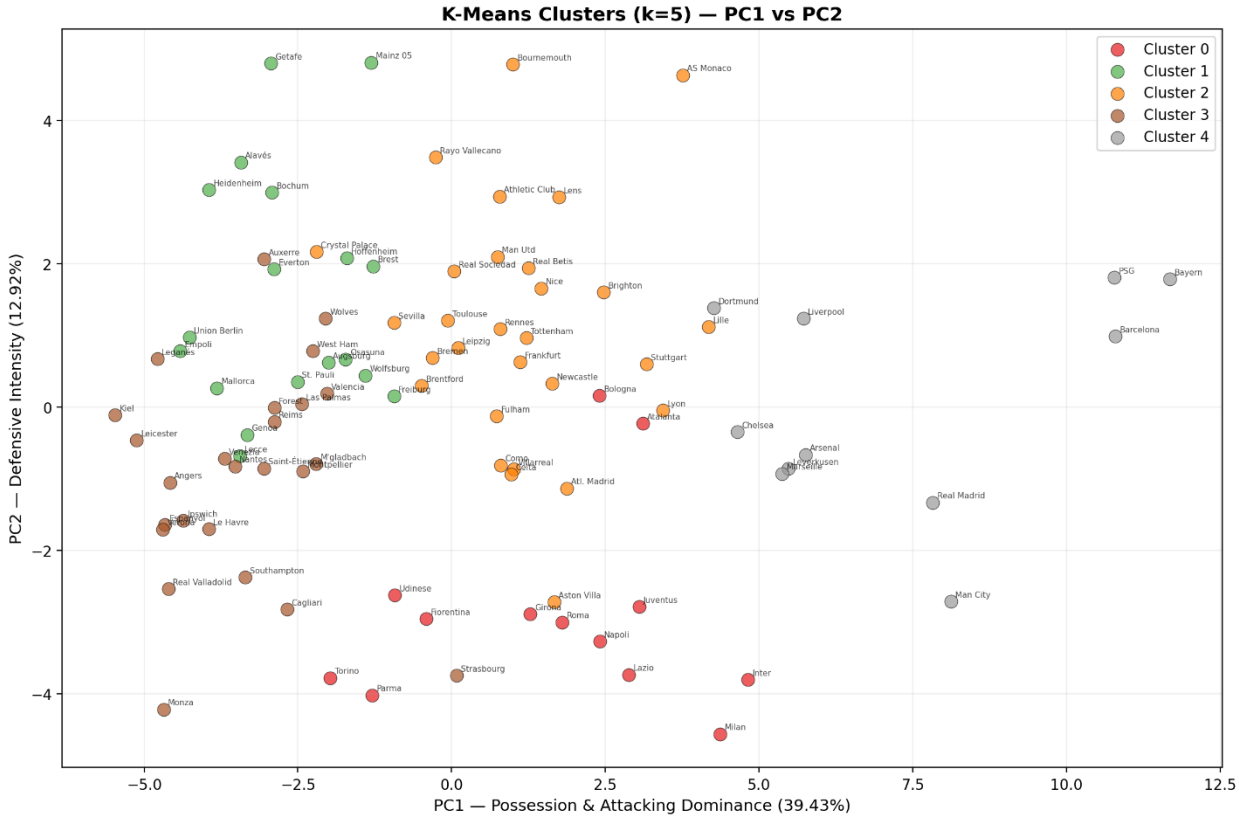


Figure 29 *Additional FC Barcelona performance metrics by opponent's playstyle*