Politechnika Łódzka
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej

Patryk Głuszek

234780

DIPLOMA THESIS

engineering (B.Sc.)

field of study Modelling and Data Science

Predictive analytics in football: machine learning approach to assessing team success factors

Institute of Applied Computer Science

Supervisor: prof. dr hab. Inż. Anna Fabijańska

ŁÓDŹ 2024

Table of Contents

Abstract

This thesis aims to explore critical factors leading to successful football team through applications of machine learning techniques as a predictive analytics tool. Random Forest model was built using data from historical matches in the Big Five European leagues. With Recursive Feature Elimination and hyperparameter tuning, the model recorded 70% prediction accuracy of the outcome of a match. Critical performance indicators such as Post-Shot Expected Goals, save percentage and Expected Goals are identified as the most significant predictors of success. Findings highlight the critical role of shot quality and goalkeeper performance in determining match outcomes. Moreover, they point out the necessity of balance between offense and defense. Insights include actionable recommendations for football teams, scouting divisions and managers, including focus on improving shot precision and enhancing goalkeeper training. These results contribute to the growing field of sports analytics, providing a data-driven foundation for strategic decision-making in football.

Keywords

Machine learning, predictive analytics

Streszczenie

Praca ma na celu zbadanie kluczowych czynników przyczyniających się do sukcesu drużyny w piłce nożnej, wykorzystując techniki uczenia maszynowego jako narzędzia analityki predykcyjnej. Wykorzystując historyczne dane z pięciu największych lig europejskich, opracowano model lasów losowych. Ulepszony poprzez rekurencyjną eliminację cech i strojenie hiperparametrów, model przeprowadza przewidywania z dokładnością wynoszącą 70%. Krytyczne wskaźniki wydajności takie jak Oczekiwane Gole po Strzale, procent obronionych strzałów i Oczekiwane Gole zostały zidentyfikowane jako najbardziej znaczące czynniki prognostyczne. Wyniki podkreślają znaczącą rolę jakości strzałów i skuteczności bramkarzy w określaniu wyników meczu. Ponadto, podkreślono również znaczenie równowagi pomiędzy atakiem i obroną. Uzyskane spostrzeżenia oferują praktyczne rekomendacje dla drużyn piłkarskich, działów skautingu i kierowników, w tym szczególny nacisk na poprawę precyzji strzałów i polepszenia treningu bramkarzy. Wyniki te przyczyniają się do rozwijającej się dziedziny analityki sportowej, zapewniając podstawy do podejmowania strategicznych decyzji w piłce nożnej.

Słowa kluczowe

Uczenie maszynowe, analityka predykcyjna

Introduction

Football, recognized as one of the most popular sports globally, places a significant emphasis on winning matches. Fédération Internationale de Football Association reports that this sport is played officially in over 200 countries, with approximately 1.3 billion fans worldwide (Constantinou, 2019). This discipline is currently undergoing a process of implementing analytics in decision-making. Analytical methods, commonly employed across various industries to foster competitive advantages, innovation, and growth, are now being adopted in football (Guimarães, 2018). Data can serve not only to reveal new economic opportunities but also could provide deeper scientific insights. The benefits of analytics in football are substantial due to several reasons. Firstly, football staff and team supporters are very analytical, which can be verified by their countless debates, statistical analyses and significance of betting industry. Additionally, the success attributed to analytics in various literature, along with the frequent exchange of knowledge through the movement of coaches and managers among teams, underscores the value of these approaches. (Davenport, 2014). One key application of analytics in football is the evaluation of match performance. If crucial team success factors were identified and optimized, teams, decision-makers and players could significantly benefit from this knowledge (Hughes et al., 2012).

The advancements of tracking technologies, along with more affordable and powerful data storage and processing systems, has extended the adoption of analytics in sports. While sports like baseball have traditionally embraced statistics to explain outcomes, football has maintained a conventional perspective, often resisting the notion that traditional methods might overlook key insights. However, larger clubs are increasingly acknowledging the utility of statistical analyses, even though some managers remain skeptical, preferring intuition over data (Kröckel, 2019).

Recent shifts toward scientifically-informed training and diverse tactical approaches in sports have led to unpredictable outcomes, highlighting the relevance of a team's overall strength and live performance. The diminishing influence of star athletes has further underline the need for strategic, data-driven approaches. Many leading clubs have started employing professional data analysts to develop corresponding strategies and tactics by

analyzing various indicators of both sides of the competition, in order to secure greater competitive advantage. Given football's status as a globally representative with intense competition, exploring the winning factors of football matches through machine learning techniques is becoming increasingly mature. This will provide objective guidance for the strategic and tactical development of participating teams (Yang, 2023). Publicly accessible data from football leagues and detailed player metrics from popular sports websites offer a rich datasets for training various machine learning models.

The primary aim of this thesis is to use machine learning models to identify the most significant factors contributing to the success of football teams using historical data analysis. Thereby by enhancing strategic decision-making processes, the goal is to provide guidance to scientifically formulate strategies. In the first chapter existing research on predictive analytics in football is explored to establish a theoretical foundation. Second chapter consist of description of the methodology and implementation process – data collection and scraping, data preprocessing, feature selection, development of the model, testing and hyperparameter tuning. Chapter 3 presents results, analysis of findings – discussion what these results mean in the context of the research and comparisons with existing literature.

1 Literature review

1.1 Development of performance analytics in football

Utilizing scientific methods to enhance decision-making in football and sports in general is not a new concept. Early applications of these methods emerged even before sports science was formally recognized as an academic field in universities (Reilly and Williams, 2003). By the 1980s, it became evident that football could benefit greatly from scientific insights to improve club organization. The initial integration of theoretical knowledge and football practice occurred in 1987 at the First World Congress of Science and Football (Reilly, et al., 2011). Clubs that recognized this opportunity early on experienced greater success compared to those that did not (Reilly and Williams, 2003).

In recent years, the accessibility of tracking and analytic technologies have increased. As a result, football clubs that initially hesitated to rely on data for decision-making lost their competitive edge. Then they accepted the fact that they could not resist predictive analytics. Football is a field with a huge amount of money involved, therefore organizations were changing their opinions and accepted analytics (Kröckel, 2019). The shift in development of data collection and advanced analytics contributed to the development of sports science in academia. By the late 1980s, only about thirty articles had been published in this field (Kuper and Szymanski, 2018). However, over the following twenty years, data availability increased dramatically, leading Coleman (2012) to identify 1146 articles across 140 sports and non-sports journals. Academic research on soccer has appeared in various magazines and papers, across different fields like economics, physics, operations research, psychology or statistics (Anderson and Sally, 2013). This marks a substantial rise in sports analytics publications, particularly at the start of the new century.

The evolution of predictive analytics in sports could be divided into four distinct stages, each with unique characteristics (Kröckel, 2019). Initially, it involved simple counting the frequency of certain actions. This was followed by qualitative assessment made by experts which were often subjective. The next stage saw a more refined analysis of performance indicators such as number of passes, player movements or running distances. Today, advanced dynamic tactical analysis is employed, utilizing advanced statistics on large datasets to uncover patterns and interactions within the game (Memmert and Raabe, 2017).

The transition towards more complex and data-driven approaches has significantly transformed the understanding of football. Early methods provided basic insights, but predictive analytics and advanced machine learning methods allow for a deeper understanding of the game. This progression highlights the importance of data and critical success factors in shaping modern sport strategies.

Since the 1950s observations and analyses were done mostly with the use pen and paper. However, the beginning of twenty-first century marked a technological revolution in this field. Carling, Williams and Thomas (2006) documented this transition in their book on soccer analysis, which still included guidelines for paper-based predictive analytics. The rapid development of video-based analysis and dynamic systems soon followed. Advancements in hardware and software enabled live statistical processing, benefiting all football staff and coaches (Carling, Williams and Thomas, 2006). The change into modern technology has significantly improved the precision and efficiency of football analytics. The integration of real-time data processing tools allows for more immediate insights, transforming the way teams strategize.

With the rise of the technology a whole industry of data providers had become to grow. The first company was Opta Sports, which was started in the 1990s by a group of sports consultants. Their goal was to create an index of player performance. When Opta started to work in the football industry, each game recording took four hours to code, using pen and paper. At the beginning noticed actions were simple – passes, shots and saves. Nowadays the level of details is much more massive. During the 2010 Champions League final Opta's staff recorded 2842 events, one every two seconds of the match. This is not the only company dealing with predictive analytics. Prozone, Amisco, Impire or StatDNA also work in the industry of collecting performance indicators. (Anderson and Sally, 2013) All of the companies are benefiting from the rise of football analytics, selling data to coaches, players, managers or betting industries.

Despite its potential, analytics was initially met with resistance in the football community. Many football managers viewed the sport more as an art form, emphasizing aesthetic experience over numerical analysis. They believed that their experience and intuition could not match with predictive analytics (Carling, Williams and Thomas, 2006). The skepticism

among professionals, who values their traditional methods, contributed to a slower adoption of data-driven approaches within the industry (Kuper and Szymanski, 2018). The reluctance to embrace analytics cause the delay in football development, as many clubs continued to operate under conventional management styles. Managers who relied only on their intuition and experience were hesitant to integrate data science into their decision-making process. As a result, those who dismissed predictive analytics fell behind their competitors who were using analytics for performance evaluation and strategic planning (Carling, Williams and Thomas, 2006). As data science and machine learning methods continue to prove their value, more clubs are beginning to recognize the benefits of integrating analytics into their acting. It is bridging the gap between sports and science.

The primary aim of the performance analysis is to enhance achievements through the use of objective measures (Kröckel, 2019). This process typically involves notational analysis which focuses on systematically recording and quantification of critical events (Carling, Williams and Thomas, 2006). This approach ensures consistent and valid quantification of performance, providing an objective evaluation of the game (Nevill, Atkinson and Hughes, 2008). The main goals of notational analysis include movement analysis, tactical and technical evaluation, database creation and providing feedback to coaches (Hughes and Franks, 2004). In recent years, notational analysis has significantly influenced coaching decisions. However, it faced criticism for its emphasis on documenting actions rather than understanding the underlying process. This method often isolates individual statistics without providing a comprehensive explanation (Travassos et al., 2013). Advanced machine learning techniques are now being used to address these limitations and offer deeper insights (Lees, 2002).

Predictive studies are instrumental in uncovering valuable insights into performance indicators that can influence future efforts to enhance performance (Sarmento et al., 2014). Despite this potential, Lepschy, Wäsche and Woll (2018) found that fewer than half of the studies on success factors in football employed predictive analytics. Their findings highlight a significant gap, indicating a need for more predictive analyses to gain a comprehensive understanding of the determinants of success in football. To bridge this gap, it is crucial to integrate predictive analytics more extensively in research related to football performance.

By doing so, researchers can identify patterns and trends that traditional methods might overlook, leading to more informed strategies and interventions. The adoption of these advanced analytical techniques could revolutionize the approach to studying and improving football performance, ultimately contributing to a higher level of success in the sport.

1.2 Analysis of performance indicators

The football industry frequently keeps technological advancements within clubs to maintain a competitive edge over rivals. Access to the latest advanced data is often privatized and requires substantial funding, creating potential barriers for academic research (Hewitt and Karakuş, 2023).

In predictive analytics, performance indicators serve as crucial variables and measures of key aspects of the game, though their interpretation can vary (O'Donoghue, 2009). These indicators have been extensively examined in academic literature related to football analytics. However, traditional statistical methods have been used to showcase their effectiveness in explaining sports performance, rather than machine learning algorithms (Kröckel, 2019). Despite extensive study, the evolution in this research area has been limited. There is minimal evidence proving that the findings from these studies have been implemented or utilized by coaches and football managers in practice (Mackenzie and Cushion, 2013). This gap suggests a disconnect between academic research and practical application, indicating a need for more innovative approaches and better integration of advanced analytics techniques in football management.

Predictive analytics aids coaches and managers in assessing football players' performance, yet there is still limited evidence on the effectiveness of predictive models in identifying key factors and team statistics that most influence winning (Gifford and Bayrak, 2020). This limitation suggests that while analytics can provide valuable insights, the precise determinants of success in football remain unclear. More research is needed to develop models that can reliably point the elements that have the greatest impact on a team's victory.

Mackenzie and Cushion highlighted several issues regarding the applicability of research results, focusing on two key methodological aspects: sample size and the definitions used for deriving results. James (2006) pointed out the lack of established knowledge about what constitutes a representative sample size in football predictive analytics. Mackenzie and Cushion (2013) noted that, of the 44 technical articles they reviewed, only 10 utilized a sample size of 100 or more games, which is likely insufficient given the number of games in a full season. Additionally, Castellano, Alvarez-Pastor and Bradley (2014) examined 38 studies that employed semi-automated tracking systems to quantify players' physical profiles. They found that half of the studies analyzed only one team, while the rest either included multiple clubs or did not specify the number of teams studies. Mackenzie and Cushion (2013) also raised concerns about the definitions used in analytics processes. Their review revealed that 79% of the technical papers did not fully define the variables involved, creating challenges for comparing new research with existing literature. This inconsistency in definitions and sample sizes hampers the ability to generalize findings and apply them effectively in practical settings.

Describing football accurately requires more than just a few variables. Liu et al. (2015) analyzed the FIFA World Cup 2014 and found that the majority of the 24 variables they examined had an impact on the match outcome. In contrast, most research studies tend to concentrate on only few variables (Lepschy, Wäsche and Woll, 2020). This discrepancy highlights the complexity of football and suggests that a more comprehensive approach is needed to capture the factors influencing game results effectively.

However as Yang (2023) suggests, random factors as player injury or status could not be influenced or estimated in advanced. Therefore they might be understood as unavoidable errors which should not be considered while analyzing performance indicators. For that matter weather, stadium state or other on site factors impacts both sides of the game equally, and are not factors which contribute to the win.

Moreover the biggest football leagues in the world, in Spain, Germany, France, England and Italy are very similar while comparing key performance indicators (Anderson and Sally, 2013). The authors of the book states that despite tiny differences, the biggest European leagues are highly alike. The most crucial elements of football are the same across various

countries and leagues. These differences are becoming bigger in case of lower level leagues, however at the top level nature of the game is uniform.

1.3 The use of data science techniques

Progress in data processing and predictive analytics have boosted data-driven decision-making. The need for productivity due to advanced data science methods has significantly altered the perception of individual and team performance in the 21st century (Fury, Oh and Berskon, 2022). Sophisticated statistical techniques are now essential for analyzing performance indicators across different games and teams. Furthermore, the rise of data science, along with big data, machine learning, and deep learning, has revolutionized the landscape of predictive analytics and impacted various sports to varying extents (Gifford and Bayrak, 2020). Sports institutes are now paying more attention to this research, recognizing its potential to provide a competitive advantage over rivals (Thakkar and Shah, 2021). This shift underscores the growing importance of integrating cutting-edge data science methods into sports strategy and performance analysis.

Over the years, numerous studies and their authors have stated that utilizing various machine learning algorithms reports in positive performance from their models. For example, Joseph, Fenton and Neil (2006) examined the effectiveness of expert-constructed Bayesian Networks in predicting match outcomes for Tottenham Hotspur Football Club, using data from 1995 to 1997. These findings suggest that sophisticated machine learning techniques can be successfully applied to forecast football match results, demonstrating the potential of advanced analytics in sports predictions.

Nunes and Sousa (2006) employed data mining techniques to uncover non-trivial patterns in datasets from various European championships. Their approach included data association rules, classification and visualization techniques. They asserted that their exploratory work validated several well-known patterns in football and demonstrated the effectiveness of their visualization methods. Study highlights the potential of data mining to provide deeper insights into football performance, reinforcing the value of predictive analytics in the sport.

Romero et al. (2021) analyzed the 2018 European Men's Handball Championship games to evaluate team performance through a weighted aggregation of statistical indicators. They employed principal component analysis to investigate the relationships between various game statistics and utilized a fuzzy multi-criteria decision-making method to predict the player of the match in each game. Similarly, Davoodi and Khanteymoori (2010) applied artificial neural networks to predict outcomes in horse racing. Using data from 100 races at the AQUEDUCT Race Track in New York, collected in January 2010, they demonstrated that neural networks are effective for predictions in the horse racing domain. As well Deshpande and Jensen (2016) analyzed the influence of NBA players on their winning games. They suggested a Bayesian linear regression model for assessing influence of every individual player. There exist also studies exploring predictive analytics in Australian football league, American football and rugby (Stefani, 1987; Croucher, 1995; Robertson, Woods and Gastin, 2015). These studies show the diverse applications of advanced analytical techniques in sports.

Delen, Cogdell and Kasap (2012) utilized eight years of data to develop classification and regression models using three data mining techniques: artificial neural networks, decision trees, and support vector machines. Their goal was to evaluate the predictive capabilities of these different methodologies. They found that classification models were more effective at predicting game outcomes compared to regression-based algorithms. Similarly, Maszczyk et al. (2014) conducted a study involving 116 javelin throwers to compare the accuracy of regression and neural models in predicting sports results. Their finding indicated that neural models outperformed regression models in prediction accuracy.

Kapadia et al. (2020) explored a range of machine learning classification techniques, including naïve bayes, random forest and k-nearest neighbor, to predict cricket match outcomes using historical data. They reported that tree-based methods, especially random forest demonstrated superior accuracy and precision. Similarly, Yezus (2014) applied four machine learning models – k-nearest neighbors, random forest, logistic regression and support vector machine – to predict English Premier League football match outcomes. The random forest model achieved the highest accuracy, averaging 63%.

Supporting these finding, Tüfekci (2016) carried out research using seventy different features to develop machine learning models for predicting match outcomes in the Turkish Super League. This study included support vector machines, bagging with REP tree, and random forest, using publicly available datasets from four seasons (2009 to 2013) encompassing 1222 matches. Performance indicators such as games played, games won, total goals, and goals per game were utilized. The random forest model outperformed the others with a 71% accuracy rate. These studies underscore the effectiveness of tree-based methods, particularly random forest, in sports predictions. Applying these techniques to football can enhance the identification of key success factors and improve the accuracy of predictive models, ultimately benefiting strategic decision-making in the sport.

1.4 Conclusions

Predictive analytics in football and sport in general is a growing field of science. Soccer benefits greatly from applications of these methods in decision-making processes and setting up strategies. With more advanced technology data collection process becomes easier and more detailed, allowing researchers to gather more specific performance indicators. Still there appears a need for more predictive analyses to understand the key determinants of soccer more clearly. Accurate elements of the game remain vague and more research regarding improving models is needed.

However there are some issues which it is recommended to take into consideration while building machine learning model. First one concerns sample size – there are almost 400 matches in single season therefore studies should analyze sufficient number of games and different clubs. As well used variables should be clearly defined to allow consistent way of comparing the results with existing work. Football is complex discipline, and more comprehensive approach – analyzing bigger number of variables is needed. Addressing these gaps, this study analyzes substantially larger sample size of 1683 matches, incorporating all clubs from the top five European leagues over two seasons. This broader approach not only enhances the representativeness of the data but also allows for a more comprehensive understanding of performance indicators across various teams. By doing so, this research provide more generalized and actionable insights, which can be effectively applied in real-

world football scenarios. In case of model selection, the random forest model outperforms others, tree-based methods appear to be the most effective in case of sports prediction. Applying these methods allows for identifying critical success factors more accurately.

2 Methodology and Implementation

2.1 Data collection

All historical data for the analysis was obtained using the SoccerData Python library – tool which is a collection of scrapers to gather football data from various sources. This library provide access to both current and archival match statistics, event streams, and forecasts. Performance indicators needed for analytic process were sourced from FBref website, which offers a comprehensive range of statistics from over 40 countries. These include data from domestic cups, super cups, youth leagues, and prestigious international tournaments such as UEFA Champions League. FBref, in collaboration with Opta, provides advanced analytical metrics, including Expected Goals (xG), Expected Assists (xA), progressive passes, and other sophisticated performance indicators. The decision about using FBref and SoccerData was made due to reliability, comprehensive coverage, and the richness of provided data, which is crucial for building robust predictive models.

To address the concerns highlighted by Mackenzie and Cushion (2013) regarding insufficient sample sizes, data was collected from the Big 5 European leagues, encompassing 1683 matches from 2021/2022 and 2022/2023 seasons. For testing the models' performance, data from 2023/2024 season have been gathered, resulting in substantial dataset of 863 match records. The training dataset is composed of 594 lost, 690 won and 399 drawn matches, while the testing dataset consist of 290 lost, 350 won, and 223 drawn games.

Collected data include wide variety of parameters that describe various aspects of the game. These parameters include shooting metrics, such as goals, shots on target, shot accuracy, and more advance metrics like Expected Goals. These indicators are crucial for assessing the quality of shooting opportunities and the likelihood of converting them into goals. Additionally, the dataset include passing metrics like total passes completed, pass completion percentage or number of specific types of passes such as crosses or long balls. Defensive actions are also well-represented in the data through metrics such as tackles, interceptions, blocks or clearances. These statistics are essential for evaluation a team ability to disrupt opponent's attacking play. Goalkeeping performance is measured using performance indicators like save percentage, goals allowed or average length of goal kicks,

highlighting the effectiveness of the goalkeeper in preventing the opponent from scoring. Possession metrics are included as well, such as the percentage of possession or the number of passes per possession phase, which help analyze a team's ability to control the game and dictating the pace of play. The dataset also includes metrics on aerial and ground duels, offering insight into the team's physical engagement and success in winning balls.

Larger datasets ensure that model generalize well across different conditions, improving predictive accuracy. By including a diverse set of parameters across different aspects of the game, analysis captures the complexity of football and allows for more practical insights for football teams.

2.2 Data Preprocessing

Match statistics in the datasets across various categories, including goalkeeping, shooting, passing, defense, and possession, were initially provided as separate Pandas DataFrames. These DataFrames were then flattened and merged into a single, larger DataFrame. After the preliminary steps of removing duplicate columns and rows with missing values, a total of 1683 match records have been obtained. Taking into consideration number of games in one full season (less than 400 matches) this dataset size appears sufficient for conducting accurate predictive analytics process.

In the next stage of data preprocessing, the relevance of certain variables that were presented both as raw counts and as derived percentage was assessed. For example, metrics such as the number of dribblers tackled (Challenges Tkl) and the percentage of dribblers tackled (Challenges Tkl%) represents similar aspects of gameplay, but they do so in different formats – one as an absolute number and the other as a percentage. The goal was to determine whether both types of variables should be included in the analysis or if one can be considered redundant due to potential multicollinearity. Therefore to evaluate this, Variance Inflation Factors (VIF) were calculated for each related pair of variables. VIF which is a statistical measure, quantifies the degree of multicollinearity among features. When two or more performance indicators are highly correlated, VIF assesses how much the variance of an regression coefficient is inflated due to this correlation (Repala, 2023). High value may

indicate that certain indicators are strongly related to other predictors or could even be redundant. In Table 2.1, the VIF values are presented for various pairs of variables. These pairs were analyzed to determine if both the raw count and percentage should be included in the final model.

Table 2.1 Variance Inflation Factors

Features	VIF
Challenges Tkl	1.31
Challenges Tkl%	
Crosses Stp	2.11
Crosses Stp%	
Launched Cmp	1.18
Launched Cmp%	
Long Cmp	2.06
Long Cmp%	
Medium Cmp	2.42
Medium Cmp%	
Short Cmp	2.38
Short Cmp%	
Standard SoT	1.48
Standard SoT%	
Take-Ons Succ	1.52
Take-Ons Succ%	
Take-Ons Tkld	1.35
Take-Ons Tkld%	
Total Cmp	4.36
Total Cmp%	

Generally, a VIF value greater than 5 is considered indicative of higher multicollinearity, whereas values closer to 1 suggest low or no multicollinearity (Repala, 2023). Presented in Table 2.1 variables Challenges Tackled (Challenges Tkl) and Challenges Tackled Percentage (Challenges Tkl%) have a VIF of 1.31, indicating a very low level of multicollinearity. This shows that both versions of variables can be included in the model without significant risk of redundancy. Similarly, the number of crosses into penalty area which were successfully stopped by the goalkeeper (Crosses Stp) and percentage of those crosses (Crosses Stp%) indicate a VIF of 2.11, which shows moderate correlation. However, this value is still within acceptable range, implying the both variables can provide unique and valuable information to the model. On the contrary, the pair Passes Completed (Total Cmp) and Pass Completion Percentage (Total Cmp%) shows the highest VIF value at 4.36. While this value indicates a

more visible degree of multicollinearity compared to the other pairs, it is not yet at a level that would require to remove one of these variables from the model. Overall, the analysis of VIF values suggests that while there is some multicollinearity present among the variables, it is generally within acceptable limits. This allows for keeping both raw count and percentage variables in the model, ensuring that the analysis captures a comprehensive view of the factors influencing match outcomes. By addressing multicollinearity through the calculation of VIF, the analysis is protected from the distortions that could appear from highly correlated variables.

Including statistics directly influencing final score, such as assists or goal-related indicators like goals per shot or goals per shot on target, could negatively affect the analysis of the results. Although these features might significantly increase accuracy, they do not facilitate drawing meaningful conclusions. They offer little insight into tactical changes or decision-making improvements, as it is difficult to derive objective conclusions from them. Similarly, columns containing information such as time, round, opponent or date are irrelevant and excluded for the same reason. As a result 20 unnecessary features, in addition to duplicates were removed from the analysis. The careful selection and removal of irrelevant features ensure that the model remains focused on variables that offer valuable insights, improving its utility for real-world applications. This led to a final set of 133 performance indicators relevant to the study, variables which allow for valuable conclusions, precious to the football clubs and management.

2.3 Feature Selection

To enhance the performance of the model in this project, feature selection process was implemented using Recursive Feature Elimination with Random Forest (RF-RFE). This crucial step involves choosing relevant subset of indicators, thereby reducing feature space, which boost the accuracy and predictive capabilities of the model. RFE was applied specifically to identify the most important features that would contribute to a more efficient and effective model. Feature selection not only enhances model performance but also improves efficiency, making the model more practical. Particularly when working with high-dimensional datasets, selecting the most relevant features is crucial to prevent overfitting.

Recursive Feature Elimination (RFE) is widely recognized method for feature selection that has demonstrated to effectively improve the accuracy and reliability of the models across various domains. Granitto et al. (2006) demonstrated the effectiveness of combining Random Forest with RFE for identifying relevant features. The study compared RF-RFE with methods like Support Vector Machine-RFE and found that RF-RFE outperformed others in selecting small subset of features that maintained positive results. This highlights RFE's ability to reduce the feature space while preserving essential information. Similarly, Zhang et al. (2016) applied recursive feature elimination in combination with Random Forest to select key features. By focusing on the most important variables, the algorithm significantly improved the model's accuracy and efficiency. The study emphasizes that RFE, when used with Random Forest, is effective in handling the imbalance in datasets. This method iteratively removes the least significant variables from the model until the desired number of features is achieved. During this process, the coefficients for each feature are calculated, and the feature with the lowest score is systematically eliminated (Kiptoon, 2023).

To assess the performance of the model in relation to the number of input features, an approach involving Recursive Feature Elimination with cross validation was implemented. This approach helps identifying the optimal number of features that should be retained for the best predictive performance. Initially, a list of models was created, each corresponding to a different number of features to select. The RFE method was applied to a Random Forest Classifier, with the number of features to retain varying from 3 to 20. For each model, the RFE recursively removed the least important features, based on the Gini importance derived from the Random Forest model, until only the specified number of variables remained. The next step involved evaluating the performance of each model. This was accomplished using 5-fold cross validation, a method that divides the dataset into five parts, training the model on four parts and testing it on the fifth. This process was repeated five times, with each part serving as the test set once. The mean and standard deviation of the accuracy – proportion of correctly classified cases – were calculated for each model across the five folds. The results of these evaluations were then compared to determine which number of features resulted in the highest model accuracy.

Table 2.2 Mean and standard deviation values of model accuracy for different number of features in RFE

Number of features	Accuracy	
	Mean	Standard deviation
3	0.604	0.026
4	0.671	0.014
5	0.674	0.015
6	0.686	0.021
7	0.690	0.023
8	0.682	0.032
9	0.700	0.014
10	0.695	0.022
11	0.682	0.016
12	0.688	0.021
13	0.690	0.013
14	0.690	0.025
15	0.693	0.015
16	0.687	0.019
17	0.694	0.013
18	0.692	0.011
19	0.693	0.014
20	0.685	0.014

The Table 2.2 shows the mean accuracy (probability that the model prediction is correct) and standard deviation for each model, with the number of features ranging from 3 to 20. The analysis shows that model performance improves as the number of features increases, peaking around 9 indicators with a mean accuracy of 0.700 and low standard deviation of 0.014. Beyond this point, adding more features does not significantly improve accuracy, indicating that the model is already capturing the most relevant information with 9 variables. This number of features was therefore selected for the final model to balance performance and complexity.

2.4 Model development and hyperparameter tuning

Random Forests, as an ensemble method, enhance the predictive power of decision trees by combining the outputs of multiple trees. Rather than depending on a single decision tree, Random Forests aggregate predictions from numerous trees, each trained on different subsets of the data, to produce a more accurate and reliable prediction (Shaik and Srinivasan, 2018) .

In this project three distinct Random Forests models, each configured in a different way were developed to evaluate their performance. The initial model utilized default hyperparameters, serving as a baseline. The second model incorporated a feature selection method, Recursive Feature Elimination with nine input variables. Model was created with a pipeline that first applied RFE to select a specific number of features and then trained a Random Forest model on the reduced feature set. Finally, the third model not only employed variable selection but also involved hyperparameter tuning through grid search, optimizing the model's configuration variables for better performance.

The hyperparameters explored in the Random Forest models included the number of trees in the forest (96, 128, 150 or 200), the method of sampling data points (with or without replacement), the maximum depth of each tree (expansion of nodes until all leaves are pure, 10 or 20) and the minimum number of samples required to split a node (2, 5 or 10) or to be retained in a leaf node (1, 2 or 4). The hyperparameter tuning process involved a 5-fold cross-validation, testing various model configurations. The best model was selected based on the highest accuracy achieved during training. Grid Search was employed to evaluate all possible combinations of hyperparameters. After fitting 5 folds for each 216 candidate models, Grid Search identified the optimal parameters, as shown in Table 2.3. The best model achieved a mean accuracy of 0.709, representing a slight improvement over the models that utilized only recursive feature elimination.

Table 2.3 Tunned hyperparameters

Hyperparameter	**Value**
Bootstrap	True
Number of decision trees in the forest	128
Maximum number of levels in each decision tree	None (nodes are expanded until all leaves are pure or until all leaves contain less than specified in min_samples_split parameter)
Minimum number of data points placed in a node before the node is split	10
Minimum number of data points allowed in a leaf node	1

2.5 Model Testing

The goal of developed predictive model in this thesis was to forecast the outcome of football matches. Specifically, the model is designed to predict one of three possible outcomes for a given match: win, loss or draw.

With the unseen data from the current season, all three models were evaluated for their performance. The same data preprocessing techniques were applied, enabling the preparation of the dataset, followed by predictions and the calculation of accuracy, precision, recall and f1-score across 863 match records. For a specific class, precision is the fraction of instances correctly classified as belonging to a given class out of all instance the model predicted to belong to that class. Recall is the fraction of instances in a class that the model correctly classified out of all instances in that class. F1 score is a metric that measure the harmonic mean of precision and recall. Reported averages include macro average (treating each class as equally important) and weighted average – by taking into account the balance of classes where each class's score is weighted by its presence in the true data sample. The classification report for the baseline model, presented in Figure 2.1, serves as a reference point for further analysis.

	precision	recall	f1-score	support
Loss	0.66	0.91	0.77	290
Draw	0.57	0.14	0.22	223
Win	0.75	0.87	0.80	350
accuracy			0.70	863
macro avg	0.66	0.64	0.60	863
weighted avg	0.67	0.70	0.64	863

Figure 2.1 Classification report of the default model

The model employing Recursive Feature Elimination (RFE), as presented in Figure 2.2, showed improved balance across all performance metrics. Specifically, there was a noticeable enhancement in precision, recall, and F1-scores for predicting draws, while maintaining similar levels of accuracy for loss and win predictions. The slight improvements observed in both macro and weighted averages suggest that the feature selection process positively impacted the model's overall performance.

	precision	recall	f1-score	support
Loss	0.71	0.87	0.78	290
Draw	0.52	0.37	0.43	223
Win	0.79	0.78	0.78	350
accuracy			0.70	863
macro avg	0.67	0.67	0.67	863
weighted avg	0.69	0.70	0.69	863

Figure 2.2 Classification report of the model with recursive feature elimination

When comparing the RFE-enhanced model to the model that also underwent hyperparameter tuning, the performance metrics – such as accuracy, precision, recall, and F1-score – remained mostly similar. Both models achieved accuracy of 0.70, indicating that the overall ability of the model to correctly classifies games results did not change.

	precision	recall	f1-score	support
Loss	0.70	0.88	0.78	290
Draw	0.51	0.32	0.40	223
Win	0.77	0.79	0.78	350
accuracy			0.70	863
macro avg	0.66	0.66	0.65	863
weighted avg	0.68	0.70	0.68	863

Figure 2.3 Classification report of the model with RFE and hyperparameter tuning

Confusion matrix for the baseline Random Forest model, shown in Figure 2.4 highlights the model's strong performance in predicting wins and losses, but also its significant difficulty in accurately predicting draws.

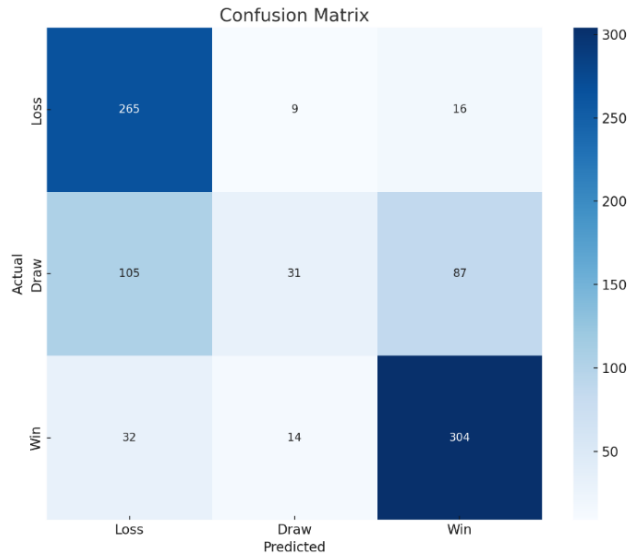

Figure 2.4 Confusion matrix of the default model

After applying recursive feature elimination, the model demonstrated better balance across all outcome categories, particularly improving its accuracy in draw predictions. Although there was a slight decrease in accuracy for win and loss predictions, the trade-off resulted in a more balance model overall.

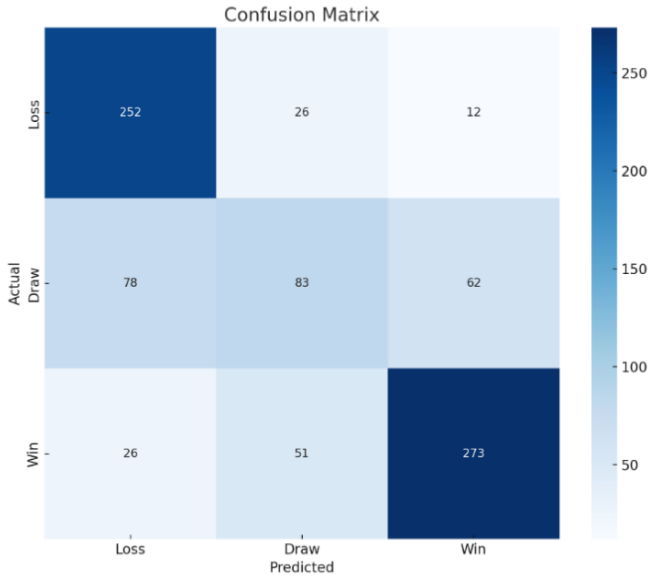

Figure 2.5 Confusion matrix of the model after recursive feature elimination

Following the application of both RFE and hyperparameter tuning, the model continued to perform well, with only marginal changes compared to the second matrix. The accuracy of win predictions remain strong, while there was a slight decline in the accuracy of draw predictions. The model appeared to have reached a stable state after the final hyperparameter adjustments.

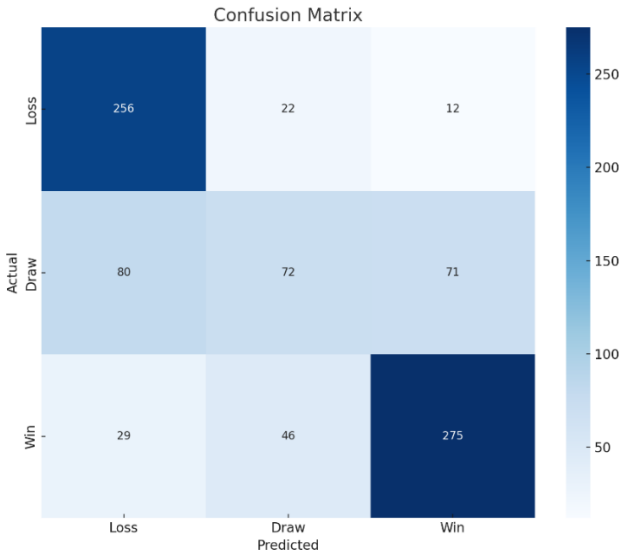

Figure 2.6 Confusion matrix of the model after RFE and hyperparameter tuning

Across all three models, the accuracy consistently remained at 0.70, indicating strong generalization capabilities on unseen data. This stability across all models highlights the strength of the Random Forest algorithm, making it a reliable choice for predictive analytics in football, especially with the precision, recall, and F1-scores for loss and win outcomes remaining consistent. The implementation of RFE notably enhanced the model's ability to predict draw outcomes, reflected by an increase in the F1-score for draw from 0.22 in the default model to 0.43 in the RFE model. By focusing on the most relevant features, the model became more balanced, reducing the severe misclassifications observed initially. Feature selection proved to be a critical factor in improving the model's ability to handle more complex predictions. The subsequent addition of hyperparameter tuning resulted in only minor changes to the model's metrics. While the F1-score for draws slightly decreased from 0.43 to 0.40, the overall accuracy and other performance metrics remained unaffected. This consistency suggests that the default hyperparameters were already well-suited for the

data and the predictive analytics process. The small variations observed could likely be attributed to the stochastic nature of the Random Forest algorithm, where inherent randomness in feature selection and decision tree construction can lead to slight performance differences. Despite this, the final model represents a well-balanced trade-off between accuracy and generalization and demonstrate the effectiveness of Random Forest approach.

Another valuable tool for evaluating classification models is the ROC Curve. This curve illustrates the separability of the classes across all possible thresholds, providing insights into how well the model is classifying each class (Trevisan, 2022). For evaluating the multiclass classification models, One vs Rest strategy is often adapted, when each class is treated as the positive class while the others are considered negative. This method reduces multiclass classification to a series of binary classifications, allowing for the generation of three distinct ROC curves corresponding to each game outcome.

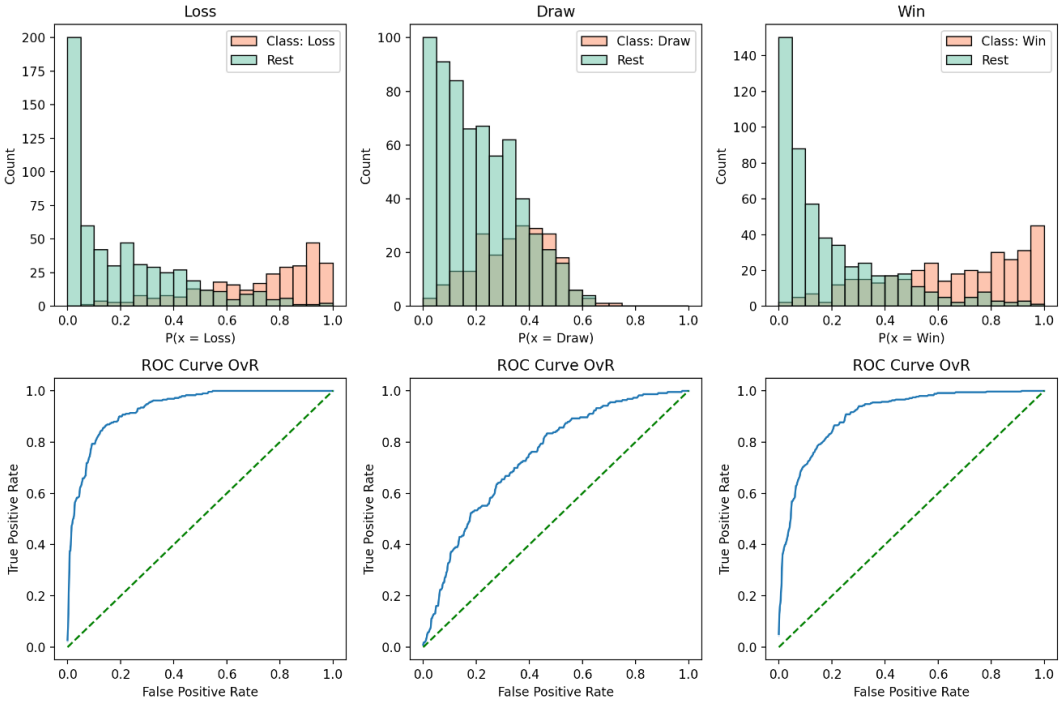

Figure 2.7 ROC Curves and histograms OvR

The final model demonstrated strong performance in predicting loss and win outcomes, as evidenced by the clear separation in the histograms and the high AUC values in the ROC curves in the Figure 2.7. The model consistently distinguished these outcomes with

confidence and accuracy. However, it struggled with the draw class. The histogram for this outcome revealed significant overlap with the other classes, and the ROC curve indicated lower sensitivity and a higher rate of false positives.

Additionally, the model's performance could be evaluated using the precision-recall curve, which highlight the trade-off between precision and recalls across different thresholds. This curve is particularly useful when dealing with imbalanced classes. The curve could also be summarized into a single metric – average precision – which represents the weighted mean of the precision achieved at each threshold value. High value of this coefficient represents satisfying balance between precision and recall (Saxena, 2022).

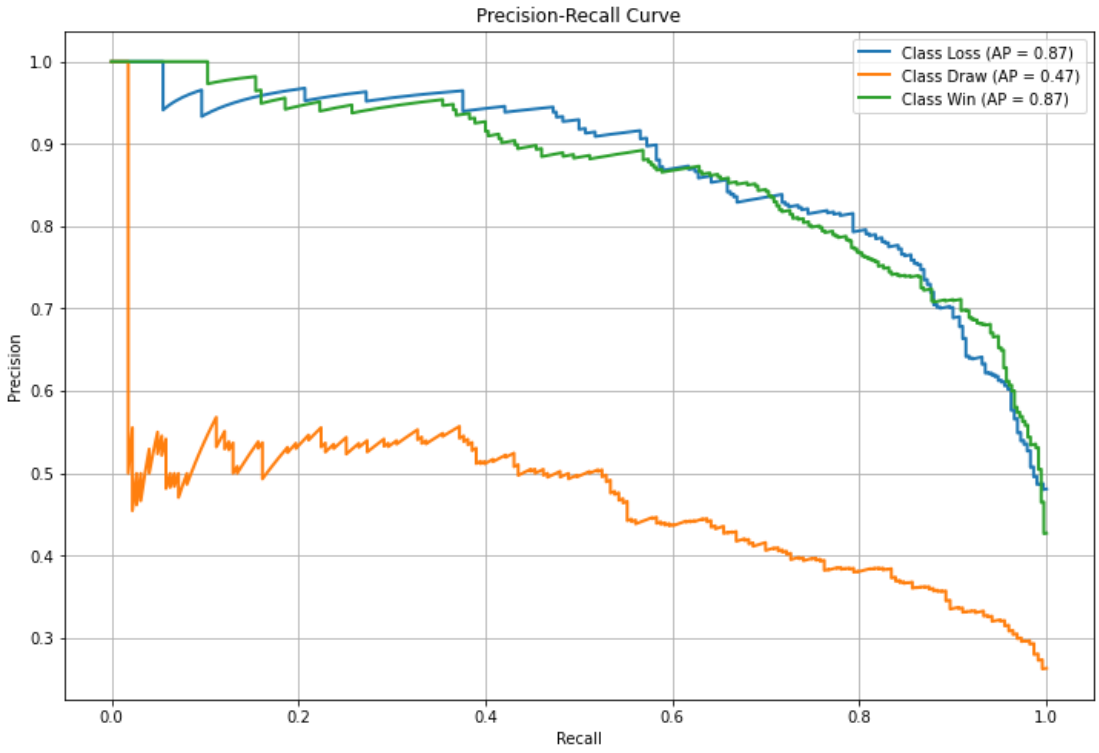

Figure 2.8 Precision-Recall Curve

As shown in Figure 2.8, for both loss and win outcomes, the model achieved high average precision values, approximately 0.87, indicating effective performance with a good balance between precision and recall. The curves for these outcomes were stable and remained high, reflecting consistent model performance across different decision thresholds. In contrast, the draw class showed weaker performance, with an average precision of 0.47. The fluctuations in the precision-recall curve suggest that the model struggled to maintain a

consistent balance for draw predictions, likely due to the fact that these outcomes are harder to classify accurately.

ROC and precision-recall curves provide a wide understanding of the model's performance, especially in terms of its ability to distinguish between different outcomes, which is crucial for making informed tactical decisions.

By showing one of the decision trees from the Random Forests models it is possible to visually illustrate how specific features and their values influence the model's decision-making process. For all three models, single exemplary decision tree was extracted and plotted. However, due to complexity and large structure as well as for clear and interpretable visualization their depths were limited to two levels.

The first decision tree, created with the default model without feature selection and hyperparameter tuning, presented in the Figure 2.9, prioritize total shots metric (Standard Sh) at the root level. As the tree branches out, the model considers percentage of shots that were on target and Post-Shot Expected Goals (Performance PSxG), reflecting a focus on offensive indicators. However the appearance of certain metrics – not present in the final results (Table 3.1) could indicate the initial model's broader focus. The model considered all available features without distinguishing between the most and least important ones, therefore it included metrics that appeared less relevant and negligible. Performance indicators like total shots could provide biased power early in the decision-making process, even if they were not the most critical factor overall.

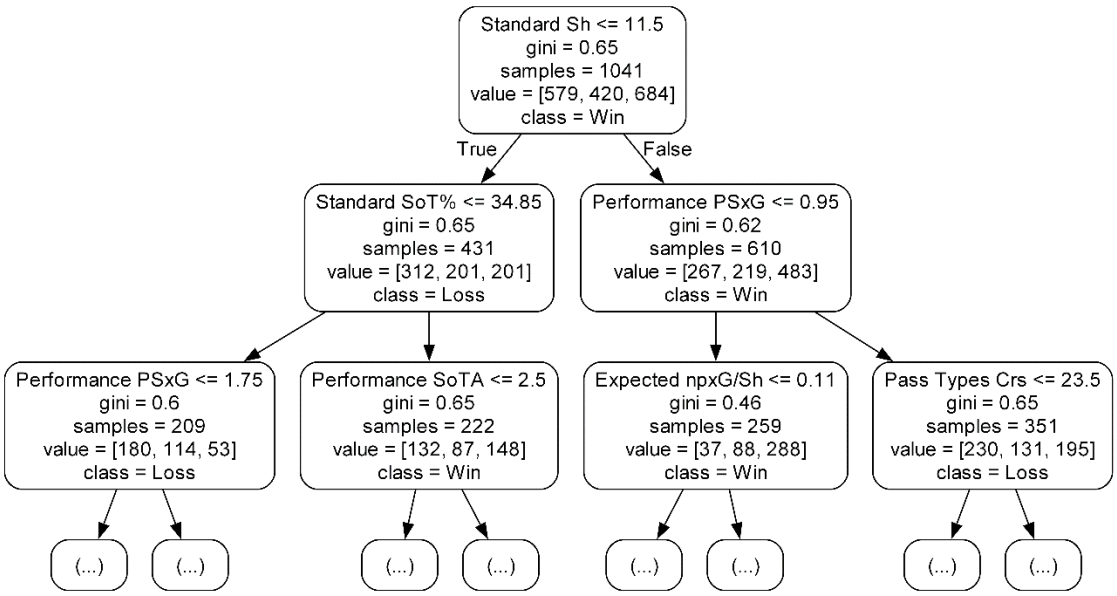

Figure 2.9 Visualization of decision tree derived from the default model

The second decision tree in the Figure 2.10, derived from the model with Recursive Feature Elimination showed focus on different indicators than the baseline model. The root node indicates Expected Goals (Expected xG), showing a higher emphasis on the quality rather than quantity of the shots. This result suggest that the model started to prioritize more sophisticated and precise metrics, rather than just measuring attempts. The tree also incorporated Post-Shot Expected Goals (Performance PSxG) earlier in the decision-making process, aligning with the feature importance analysis that indicated this metric as highly critical. The model is focusing on fewer and more impactful features, leading to more accurate and reliable outcomes.

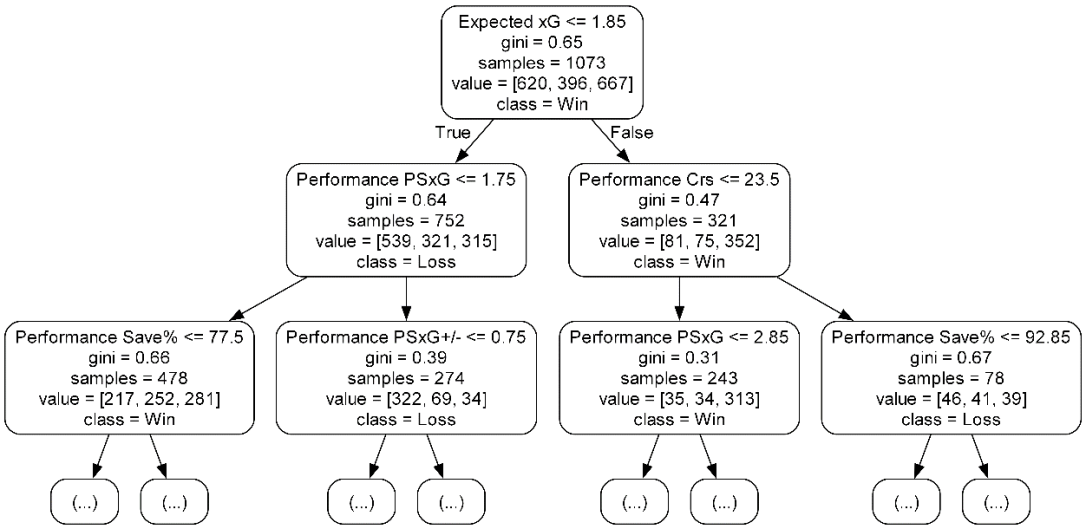

Figure 2.10 Visualization of decision tree derived from the model with feature selection

The third decision tree (Figure 2.11), generated from the model that utilized both RFE and hyperparameter tuning, shows further improvement. Similar to the second tree, Expected Goals remained at the root node, with lowered threshold of 1.65 reinforcing the fine-tuned importance of this metric. The graph continues to highlight the importance of Post-Shot Expected Goals, but also introduces save percentage (Performance Save%) earlier in the tree. This could indicate that defensive metrics have gained more significance after hyperparameter tuning.

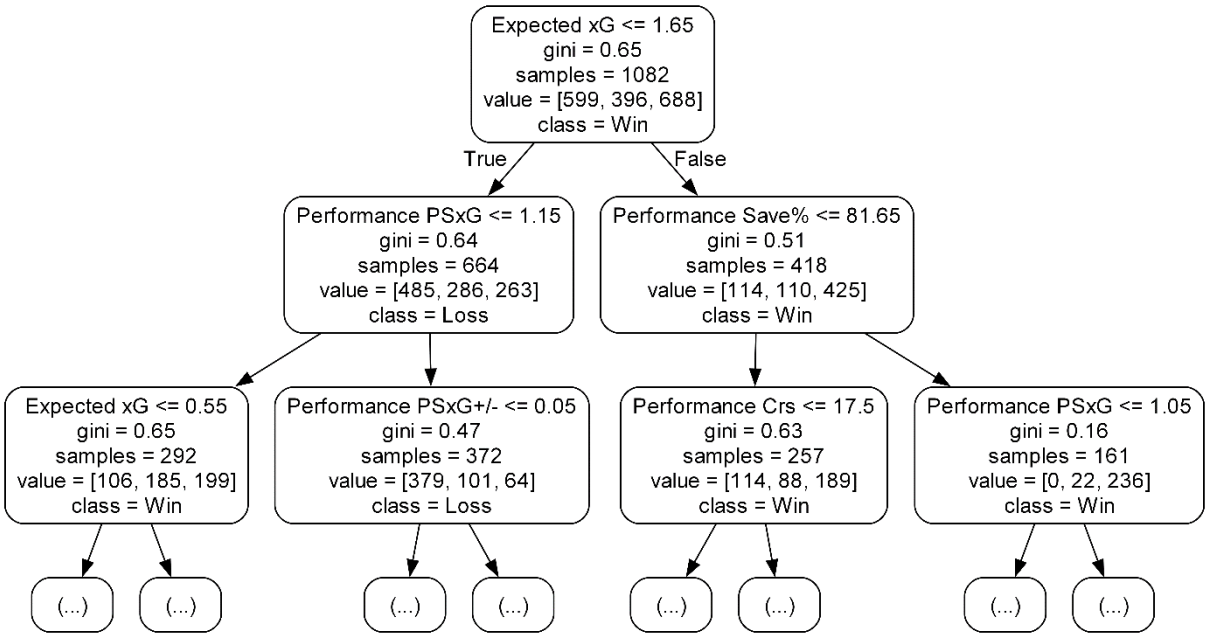

Figure 2.11 Visualization of decision tree derived from the model with feature selection and hyperparameter tuning

Across three decision trees, the evolution from the default model to the model with RFE and hyperparameter tuning shows the increasing importance of feature selection. The presence of certain metrics in the base tree but not in final results suggest that while these features may have some predictive power, they are not as strongly correlated with the outcomes as the final selected features. Optimization process through feature selection and hyperparameter tuning essentially filters out performance indicators that might be noisy or redundant, leading to more precise model. The root nodes' shift from total shots to Expected Goals underscores the model's progression from a broad focus on general shot metrics to a more precise emphasis on quality rather than quantity. Persistence of defensive metrics like save percentage indicate that both offensive and defensive factors are critical in predicting match outcomes.

Table 2.4 Maximum depths and node counts of exemplary decision trees

	Maximum depth	**Node count**
Default model	17	525
Model with RFE	19	641
Model with RFE and hyperparameter tuning	18	317

Maximum depths and node counts for presented exemplary trees are provided in Table 2.4. The first model with maximum depth of 17 and a node count of 525, indicates that it is relatively deep with moderate level of complexity. Second model, after applying Recursive Feature Elimination, shows that model is exploring more complex interactions between metrics. The higher node count of 641 indicates that the model is considering more decision points, likely due to more clarified set of features selected by feature selection algorithm. The third tree, has a slightly lower maximum depth of 18 compared to the model with RFE but significantly fewer nodes. The reduction in this parameter suggests that hyperparameter tuning reduced unnecessary complexity leading to more powerful model.

3 Results and discussion

3.1 Analysis of findings

Importance of features within Random Forest model could be assessed using a metric known as Gini importance. This method is widely recognized for its computational efficiency and straightforward interpretation. Gini importance measures the significance of a feature by evaluating the total decrease in node impurity that the feature contributes across all the trees in the forest. Essentially, the more a feature reduces Gini impurity during the splits where it is utilized, the higher its importance score (Menze et al., 2009). The computed values, as presented in Table 3.1, highlight the key factors that contribute to a teams' success in football matches. These findings offer valuable insights into the elements of gameplay and team performance that are most critical for achieving positive outcomes.

Table 3.1 Feature Importances

Feature	Gini importance
Post-Shot Expected Goals (PSxG)	0.179016
Save Percentage	0.175497
Expected goals (xG)	0.135553
Post-shot Expected Goals minus Goals Allowed	0.116842
Percentage of shots that are on target	0.105394
Non-Penalty Expected Goals	0.080260
Crosses	0.076662
Average length of goal kicks	0.067994
Pass completion percentage (passes longer than 30 yards)	0.062783

Among the features, Post-shot Expected Goals (PSxG) stands out with the highest importance score. This metric estimates the likelihood that a shot will result in a goal, taking into account factors such as the shot's location, angle, trajectory, speed and whether it was on target. (Çavuş and Biecek, 2022). The prominence of PSxG in the analysis suggests that the quality of shooting opportunities, particularly in terms of their placement and scoring potential, is the most decisive factor in determining the outcome of a match. Teams that consistently generate high-quality shooting opportunities are more likely to win.

Goalkeeper performance also emerges as a critical factor, particularly regarding their ability to save shots. Importance of the average length of goal kicks further underscores the role of the goalkeeper in influencing match outcomes. The strategic execution of goal kicks, whether to initiate an attack or maintain possession, plays a crucial role in shaping the flow of the game. This finding highlights the significant impact of strong goalkeeper on a team's chances of winning.

The Expected goals (xG) metric as the third most important performance indicator, reinforce the value of creating high-quality scoring opportunities. Additionally, Post-Shot Expected Goals minus Goals Allowed reflects both offensive and defensive capabilities, emphasizing the necessity of not only generating scoring chances but also effectively defending against opponent's attempts. A positive value in this metric indicates either favorable circumstances or an above-average ability to prevent goals. Shooting accuracy also plays a crucial role in determining success. Teams that consistently direct their shots on target have a higher likelihood of scoring, which directly correlates with winning matches.

The model also highlight the significance of other performance indicators. The ability to deliver precise and effective crosses into the opponent's penalty area is particularly important, as it contributes significantly to creating scoring opportunities. Additionally, accurately completing long passes, is vital for transitioning play and setting up goal-scoring opportunities, as reflected by the Expected Goals metric.

3.2 Comparison with the existing literature

Mao et al. (2016) identified five key variables that consistently influenced match outcomes in the Chinese Super League. Among these, three of them align closely with the results obtained in this study: shots on target, the number of shots, and cross accuracy. According to their findings, these factors significantly increase a team's likelihood of winning. Additionally, they identified the number of tackles and yellow cards as important factors, though these were not included in the current analysis.

Lepschy, Wäsche and Woll (2020) conducted an analysis of three seasons of the German Bundesliga, encompassing a total of 918 matches. They concluded that defensive errors had

the most significant impact on match outcomes. Additionally, they emphasized the importance of maintaining a balanced defense. Their study, similar to the findings in the thesis, highlighted the importance of shot frequency and quality as critical success factors. They observed a trend toward valuing accuracy over the number of shots, indicating that precise and well-placed shots are more beneficial. The researchers also noted that effective coordination, tactical awareness, and physical positioning are crucial for enhancing goal-scoring efficiency.

Geurkink et al. (2021) utilized tree-base machine learning methods to determine that shot-related variables were among the top predictors of success in the Belgian Pro League. Specifically, they found that the total number of shots on target from withing the attacking penalty box was the most effective predictor of match outcomes. This finding aligns with the identification of Expected Goals (xG) and Post-Shot Expected Goals (PSxG) as critical factors in this study, both of which assess the quality and likelihood of successful shots. Pratas, Volossovitch and Carita (2016) also emphasized the importance of shots on goal, particularly in determining the timing of the first goal in a match, which directly influences the chances of winning. Consistent with the findings presented in Table 3.1, these studies underscore the significance of shot-related metrics as vital indicators of success in football.

Anderson and Sally (2013) stressed the importance of balancing both offensive and defensive strategies. Their research revealed that, the value of clean sheet (a game without conceding a goal), is worth almost 2.5 points per game on average, suggesting that goalkeeping performance, particularly the ability to prevent goals highlights the critical importance of the save percentage metric.

In summary, the current study's results align well with existing literature, especially taking into account the importance of shot quality, defensive stability, and goalkeeping performance as team success factors. These findings not only reinforce existing research but also contribute to a deeper understanding of the performance indicators that most significantly influence match outcomes.

Developed model in this thesis achieved an accuracy of 0.70 and compares favorably with other studies in the domain of football match prediction. For example, Alfredo and Isa

(2019) applied tree-based models, including Random Forest, to predict match outcomes using data from ten seasons of the English Premier League. Their Random Forest algorithm achieved an accuracy of 68.55%, slightly lower than the accuracy obtained by the model in this thesis. This difference may indicate that the feature selection and tuning strategies employed in this project could offer a slight edge in predictive performance. Similarly, Hu and Fu (2022) explored the use of Random Forest for football match result prediction, achieving an accuracy of 66.7% on training data and 63.8% on test set. Additionally, Yezus (2014) employed four different machine learning models for football prediction and obtained a maximum accuracy with RF - 63%. These comparisons suggest that the methodological choices in feature selection and model tuning in this study were effective, leading to a model that performs at a competitive level.

Achieved accuracy of the model can also be effectively compared to those reported in studies with alternative machine learning algorithms for predictive analytics. For instance, Aithal and Bargavi (2023) assessed various classification algorithms, including logistic regression and support vector machines, reporting an accuracy of around 70% for their best-performing model, similar to the metric of the model in this thesis. Similarly, Raju et al. (2020) applied machine learning algorithms such as gradient boosting and neural networks to predict outcomes in the English Premier League, achieving a multi-class accuracy of 70.27%. Rahman et al. (2020) used deep neural networks to predict outcomes for the FIFA World Cup 2018, with an accuracy of 63.3%. Consistency in predictive performance across these different algorithms underscores that, despite variations in methodological approaches, the accuracy of the model developed in this thesis aligns well with other advanced predictive models in the field. The comparisons indicate that the Random Forest model developed in this study is competitive with, and in some cases exceeds, the performance of models reported in other works. Methodological choices of feature selection and hyperparameter tuning appears to have contributed positively for performance of the model, marking as a suitable tool for football match prediction.

3.3 Interpretation and recommendations

There are several practical recommendations that the model can give, according to its findings and Gini importance scores, for football coaches, staff, and decision-makers. The first recommendation is to prioritize shot quality in training sessions. Coaches need to prepare the players to take quality shots, considering shot placement and decision-making in critical goal areas. This could involve implementing structured training exercises that mimic high-pressure situations to enhance shot accuracy, or drills designed to consistently improve shot placement on target. The emphasis during shooting exercises should be on precision rather than power. Tactical approaches would create high likelihood chances to score, positioning players in perfect places for taking advantage of such chances.

Another key recommendation is to enhance goalkeeper training. By investing in specialized coaching aimed at improving save percentage, teams could significantly reduce the number of goals conceded. Training should focus on improving goalkeepers' reaction times, positioning, and decision-making under pressure, all of which are crucial for increasing the likelihood of winning matches. Regular performance reviews, including video analysis of saves and conceded goals, can also help identify and address weaknesses, further reducing the number of goals allowed. Additionally, goalkeepers should be trained to use goal kicks strategically, either to initiate quick counter-attacks or to maintain possession and control the pace of the game. Outfield players must also be aware of these strategies and position themselves accordingly to maximize the effectiveness of goal kicks.

Maintaining a well-balanced approach between offense and defense is vital for consistent success. Coaches should ensure that the team's formation and tactics are optimized to strengthen both attacking and defensive phases, especially during transitions. Defensive exercises should focus on preventing opponents from creating high-quality scoring opportunities by improving player positioning and situational awareness.

Optimizing crossing and passing strategies is also crucial. Confirming that players are in the right positions to convert these situations into goals can maximize the effectiveness of these plays. Particular attention should be given to improving the accuracy of long passes, especially those exceeding 30 yards, as they play a significant role in transitioning play and setting up scoring opportunities.

Insights derived from this analysis have the potential to greatly influence decision-making processes within football teams, impacting everything from tactical adjustments to player development and recruitment strategies. Coaches and analysts could benefit from these findings to make more informed decisions about game tactics. By concentrating on the most impactful aspects, such as shot quality and defensive actions, team can adjust their approach to maximize their chances of success.

Understanding the key performance indicators associated with winning allows teams to develop more targeted training programs. Player development could be customized to strengthen areas such as shot accuracy, goalkeeping, and passing efficiency, ensuring that resources are allocated toward improving the skills that have the most significant impact on match outcomes.

These insights could also redefine recruitment strategies by identifying players who star in the most critical areas, such as those with high shot quality, strong goalkeeping abilities or accurate long passing. By focusing on these attributes, teams can build a squad that is better equipped to achieve success. Furthermore, identifying players with high Expected Goals (xG) indicators, especially at young age, allows for more cost-effective signings. Recognizing undervalued players creates opportunities for both financial profit and enhanced team quality, which is especially advantageous for smaller clubs (Hewitt and Karakuş, 2023).

During matches, coaching staff and managers could apply these insights to make informed in-game decisions, such as substitutions or tactical adjustments. For instance, if a team is struggling to generate high-quality chances, adjustments can be made by substituting players or altering the formation to exploit the opponent's weaknesses. Ultimately, by incorporating these data-driven insights into their decision-making processes, football teams can improve their overall performance, leading to better results on the pitch. Ability to focus on the most critical aspects of gameplay guarantee that efforts are concentrated on areas that have the most substantial impact on success, thereby increasing the likelihood of winning.

Conclusions

This thesis aims to identify key success factors in football by analyzing historical data. With advancements in technology and improved access to data, the field of football analytics has benefit greatly from predictive analytics processes. However, there remains a need for more precise and comprehensive identification of the critical aspects of the game. In response to the concerns raised in the literature review – particularly regarding the clear definition of variables and the importance of a sufficient sample size and number of clubs – data was collected from three seasons across top five European leagues. Utilizing Random Forest model, along with recursive feature elimination and hyperparameter tuning, the model achieved an accuracy of 0.70.

Results revealed that Post-Shot Expected Goals, Save Percentage, and Expected Goals are among the most influential predictors of match success. These findings highlight the crucial role of shot quality and goalkeeper performance in determining the outcomes of football matches. Additionally, maintaining a balance between effective offense and organized defense appeared as a significant factor, particularly through the use of accurate passing and strategic goal kicks.

Implications of these findings suggest actionable changes that football clubs, coaches, and staff could implement, focusing on the identified critical performance indicators. By prioritizing shot quality and enhancing goalkeeper training, teams could significantly increase their chances of winning matches. Furthermore, recruitment and scouting divisions can leverage these insights to make more informed decisions, particularly by identifying players with high Expected Goals metrics, thereby enabling the construction of a more competitive squad.

List of tables

List of figures

References

Aithal, S. and Manju Bargavi, S.K. (2023) “A NOVEL APPROACH FOR PREDICTING FOOTBALL MATCH RESULTS: AN EVALUATION OF CLASSIFICATION ALGORITHMS,” *International Journal of Advanced Research*, 11(03), pp. 679–686. Available at: https://doi.org/10.21474/IJAR01/16481.

Alfredo, Y.F. and Isa, S.M. (2019) “Football Match Prediction with Tree Based Model Classification,” *International Journal of Intelligent Systems and Applications*, 11(7), pp. 20–28. Available at: https://doi.org/10.5815/ijisa.2019.07.03.

Anderson, C. and Sally, D. (2013) The numbers game: Why everything you know about soccer is wrong. New York: Penguin Books.

Castellano, J., Alvarez-Pastor, D. and Bradley, P.S. (2014) “Evaluation of Research Using Computerised Tracking Systems (Amisco® and Prozone®) to Analyse Physical Performance in Elite Soccer: A Systematic Review,” Sports Medicine, 44(5), pp. 701–712. Available at: https://doi.org/10.1007/s40279-014-0144-3.

Çavuş, M. and Biecek, P. (2022) “Explainable expected goal models for performance analysis in football analytics,” in 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 1–9. Available at: https://doi.org/10.1109/DSAA54385.2022.10032440.

Christopher, C., Williams, A. and Thomas, R. (2006) Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance. London: Routledge.

Coates, D., Frick, B. and Jewell, T. (2016) “Superstar Salaries and Soccer Success,” Journal of Sports Economics, 17(7), pp. 716–735. Available at: https://doi.org/10.1177/1527002514547297.

Coleman, B.J. (2012) “Identifying the ‘Players’ in Sports Analytics Research,” Interfaces, 42(2), pp. 109–118. Available at: https://doi.org/10.1287/inte.1110.0606.

Constantinou, A.C. (2019) “Dolores: a model that predicts football match outcomes from all over the world,” Machine Learning, 108(1), pp. 49–75. Available at: https://doi.org/10.1007/s10994-018-5703-7.

Croucher, J. (1995) “Scoring Patterns in Rugby League,” Teaching Statistics, 17(2), pp. 47–49. Available at: https://doi.org/10.1111/j.1467-9639.1995.tb00865.x.

Davenport, T.H. (2014) “Analytics in sports: The new science of winning,” International Institute for Analytics, 2, pp. 1–28. Available at: https://studylib.net/doc/8079544/analytics-in-sports--the-new-science-of-winning (Accessed: August 9, 2024).

Davoodi, E. and Khanteymoori, A.R. (2010) “Horse Racing Prediction Using Artificial Neural Networks,” Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, pp. 155–160. Available at: https://www.researchgate.net/profile/Alireza-Khanteymoori/publication/228847950_Horse_racing_prediction_using_artificial_neural_networks/links/53fc54590cf2dca8ffff0df8/Horse-racing-prediction-using-artificial-neural-networks.pdf (Accessed: August 9, 2024).

Delen, D., Cogdell, D. and Kasap, N. (2012) “A comparative analysis of data mining methods in predicting NCAA bowl outcomes,” International Journal of Forecasting, 28(2), pp. 543–552. Available at: https://doi.org/10.1016/j.ijforecast.2011.05.002.

Deshpande, S.K. and Jensen, S.T. (2016) “Estimating an NBA player’s impact on his team’s chances of winning,” Journal of Quantitative Analysis in Sports, 12(2), pp. 51–72. Available at: https://doi.org/10.1515/jqas-2015-0027.

Fury, M.S., Oh, L.S. and Berkson, E.M. (2022) “New Opportunities in Assessing Return to Performance in the Elite Athlete: Unifying Sports Medicine, Data Analytics, and Sports Science,” Arthroscopy, Sports Medicine, and Rehabilitation, 4(5), pp. e1897–e1902. Available at: https://doi.org/10.1016/j.asmr.2022.08.001.

Geurkink, Y. et al. (2021) “Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer,” Applied Sciences, 11(5), p. 2378. Available at: https://doi.org/10.3390/app11052378.

Gifford, M. and Bayrak, T. (2020) “What makes a winner? Analyzing Team Statistics to Predict Wins in the NFL,” AMCIS 2020 Proceedings, 35. Available at: https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/35/?utm_source=aisel.aisnet.org%2Famcis2020%2Fdata_science_analytics_for_decision_support%2Fdata_science_analytics_for_decision_support%2F35&utm_medium=PDF&utm_campaign=PDFCoverPages (Accessed: August 9, 2024).

Granitto, P.M. et al. (2006) “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” Chemometrics and Intelligent Laboratory Systems, 83(2), pp. 83–90. Available at: https://doi.org/10.1016/j.chemolab.2006.01.007.

Guimarães, J.H.M.M. (2018) Data analytics applied to football and football players. Available at: https://repositorio.ucp.pt/handle/10400.14/27626?locale=en (Accessed: August 13, 2024).

Hewitt, J.H. and Karakuş, O. (2023) “A machine learning approach for player and position adjusted expected goals in football (soccer),” Franklin Open, 4. Available at: https://doi.org/10.1016/j.fraope.2023.100034.

Hu, S. and Fu, M. (2022) “Football Match Results Predicting by Machine Learning Techniques,” in *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*. IEEE, pp. 72–76. Available at: https://doi.org/10.1109/ICDACAI57211.2022.00022.

Hughes, M. et al. (2012) “Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position,” Journal of Human Sport and Exercise, 7(2), pp. 402–412. Available at: http://www.redalyc.org/articulo.oa?id=301023544006 (Accessed: August 9, 2024).

Hughes, M. and Franks, I. (2004) “Notational analysis - A review of the literature,” in Notational Analysis of Sport, pp. 71–116.

James, N. (2006) “Notational analysis in soccer: past, present and future.,” International Journal of Performance Analysis in Sport, 6(2), pp. 67–81. Available at: https://doi.org/10.1080/24748668.2006.11868373.

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

Kapadia, K. et al. (2020) “Sport analytics for cricket game results using machine learning: An experimental study,” Applied Computing and Informatics, 18(3/4), pp. 256–266. Available at: https://doi.org/10.1016/j.aci.2019.11.006.

Kiptoon, D. (2023) Feature Selection in Machine Learning, Medium. Available at: https://medium.com/@jdkiptoon/feature-selection-in-machine-learning-20417d052b80 (Accessed: August 8, 2024).

Kröckel, P. (2019) Big Data Event Analytics in Football for Tactical Decision Support. Available at: https://www.researchgate.net/publication/337195840_Big_Data_Event_Analytics_in_Football_for_Tactical_Decision_Support/citations (Accessed: August 9, 2024).

Kuper, S. and Szymanski, S. (2018) Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the U.S., Japan, Australia, Turkey -- and Even Iraq -- Are Destined to Become the Kings of the World’s Most Popular Sport. Hachette UK.

Lees, A. (2002) “Technique analysis in sports: a critical review,” Journal of Sports Sciences, 20(10), pp. 813–828. Available at: https://doi.org/10.1080/026404102320675657.

Lepschy, H., Wäsche, H. and Woll, A. (2018) “How to be Successful in Football: A Systematic Review,” The Open Sports Sciences Journal, 11(1), pp. 3–23. Available at: https://doi.org/10.2174/1875399X01811010003.

Lepschy, H., Wäsche, H. and Woll, A. (2020) “Success factors in football: an analysis of the German Bundesliga,” International Journal of Performance Analysis in Sport, 20(2), pp. 150–164. Available at: https://doi.org/10.1080/24748668.2020.1726157.

Liu, H. et al. (2015) “Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup,” Journal of Sports Sciences, 33(12), pp. 1205–1213. Available at: https://doi.org/10.1080/02640414.2015.1022578.

Mackenzie, R. and Cushion, C. (2013) “Performance analysis in football: A critical review and implications for future research,” Journal of Sports Sciences, 31(6), pp. 639–676. Available at: https://doi.org/10.1080/02640414.2012.746720.

Mao, L. et al. (2016) “Identifying keys to win in the Chinese professional soccer league,” International Journal of Performance Analysis in Sport, 16(3), pp. 935–947. Available at: https://doi.org/10.1080/24748668.2016.11868940.

Maszczyk, A. et al. (2014) “Application of Neural and Regression Models in Sports Results Prediction,” Procedia - Social and Behavioral Sciences, 117, pp. 482–487. Available at: https://doi.org/10.1016/j.sbspro.2014.02.249.

Memmert, D. and Raabe, D. (2017) Revolution im Profifußball. Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: https://doi.org/10.1007/978-3-662-53910-1.

Menze, B.H. et al. (2009) “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” BMC Bioinformatics, 10(1), p. 213. Available at: https://doi.org/10.1186/1471-2105-10-213.

Nevill, A., Atkinson, G. and Hughes, M. (2008) “Twenty-five years of sport performance research in the Journal of Sports Sciences,” Journal of Sports Sciences, 26(4), pp. 413–426. Available at: https://doi.org/10.1080/02640410701714589.

Nunes, S. and Sousa, M. (2006) “Applying Data Mining Techniques to Football Data from European Championships.” Available at: https://www.semanticscholar.org/paper/Applying-Data-Mining-Techniques-to-Football-Data-Nunes-Sousa/fa0e6fc6d2143d0faca96a1f542772bef296667f (Accessed: August 9, 2024).

O’Donoghue, P. (2009) Research methods for sports performance analysis. Milton Park, Abingdon, Oxon: Routledge.

Pratas, J.M., Volossovitch, A. and Carita, A.I. (2016) “The effect of performance indicators on the time the first goal is scored in football matches,” International Journal of Performance Analysis in Sport, 16(1), pp. 347–354. Available at: https://doi.org/10.1080/24748668.2016.11868891.

Rahman, Md.A. (2020) “A deep learning framework for football match prediction,” *SN Applied Sciences*, 2(2), p. 165. Available at: https://doi.org/10.1007/s42452-019-1821-5.

Raju, M.A. *et al.* (2020) “Predicting the Outcome of English Premier League Matches using Machine Learning,” in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*. IEEE, pp. 1–6. Available at: https://doi.org/10.1109/STI50764.2020.9350327.

Reilly, T. et al. (2011) Science and Football: Proceedings of the first World Congress of Science and Football, Liverpool, 13-17th April 1987. Abindgon, Oxon: Routledge.

Reilly, T. and Williams, M. (2003) “Introduction to science and soccer,” in Science and Soccer. 2nd edn. London: Routledge, pp. 9–14.

Repala, S. (2023) Tackling Multicollinearity: Understanding Variance Inflation Factor (VIF) and Mitigation Techniques, Medium. Available at: https://medium.com/@satyarepala/tackling-multicollinearity-understanding-variance-inflation-factor-vif-and-mitigation-techniques-2521ebf024b6 (Accessed: August 9, 2024).

Robertson, S., Woods, C. and Gastin, P. (2015) “Predicting higher selection in elite junior Australian Rules football: The influence of physical performance and anthropometric attributes,” Journal of Science and Medicine in Sport, 18(5), pp. 601–606. Available at: https://doi.org/10.1016/j.jsams.2014.07.019.

Romero, F.P. et al. (2021) “A data-driven approach to predicting the most valuable player in a game,” Computational and Mathematical Methods, 3(4). Available at: https://doi.org/10.1002/cmm4.1155.

Sarmento, H. et al. (2014) “Match analysis in football: a systematic review,” Journal of Sports Sciences, 32(20), pp. 1831–1843. Available at: https://doi.org/10.1080/02640414.2014.898852.

Saxena, S. (2022) Precision-Recall Curve, Medium. Available at: https://pub.towardsai.net/precision-recall-curve-26f9e7984add (Accessed: August 15, 2024).

Shaik, A.B. and Srinivasan, S. (2018) “A Brief Survey on Random Forest Ensembles in Classification Model,” in International Conference on Innovative Computing and Communications. Chennai: Springer Nature, pp. 253–260. Available at: https://doi.org/10.1007/978-981-13-2354-6_27.

Stefani, R.T. (1987) “Applications of statistical methods to American football,” Journal of Applied Statistics, 14(1), pp. 61–73. Available at: https://doi.org/10.1080/02664768700000006.

Thakkar, P. and Shah, M. (2021) “An Assessment of Football Through the Lens of Data Science,” Annals of Data Science, 8, pp. 823–836. Available at: https://link.springer.com/article/10.1007/s40745-021-00323-2 (Accessed: August 9, 2024).

Travassos, B. et al. (2013) “Performance analysis in team sports: Advances from an Ecological Dynamics approach,” International Journal of Performance Analysis in Sport, 13(1), pp. 83–95. Available at: https://doi.org/10.1080/24748668.2013.11868633.

Trevisan, V. (2022) Multiclass classification evaluation with ROC Curves and ROC AUC, Medium. Available at: https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a (Accessed: August 15, 2024).

Tüfekci, P. (2016) “Prediction of Football Match Results in Turkish Super League Games,” in Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015. Springer, pp. 515–526. Available at: https://doi.org/10.1007/978-3-319-29504-6_48.

Yang, Y. (2023) “Research on the winning factors of football matches based on machine learning,” Academic Journal of Mathematical Sciences, 4(4), pp. 51–56. Available at: https://doi.org/10.25236/AJMS.2023.040408.

Yezus, A. (2014) Predicting outcome of soccer matches using machine learning. Saint-Petersburg. Available at: https://math.spbu.ru/SD_AIS/documents/2014-12-341/2014-12-tw-15.pdf (Accessed: August 12, 2024).

Zhang, C. et al. (2016) “Feature selection of power system transient stability assessment based on random forest and recursive feature elimination,” in 2016 IEEE PES Asia-Pacific

Power and Energy Engineering Conference (APPEEC). IEEE, pp. 1264–1268. Available at: https://doi.org/10.1109/APPEEC.2016.7779696.