Brief Report

A Machine Learning Framework for Team Success Classification in Professional Football: A Pilot Study Using Premier League Performance Data

Rayvanth Sankar Ravichandran * and Nor Samsiah Sani

Posted Date: 18 May 2026

doi: 10.20944/preprints202605.1076.v1

Keywords: machine learning; explainable AI; SHAP; football analytics; gradient boosting; sports data mining

 doi:10.20944/preprints202605.1076.v1

Brief Report

A Machine Learning Framework for Team Success Classification in Professional Football: A Pilot Study Using Premier League Performance Data

Rayvanth Sankar Ravichandran * and Nor Samsiah Sani

Universiti Kebangsaan Malaysia, Malaysia

* Correspondence: rayvanth@gmail.com

Abstract

In the era of data-driven decision-making, the pursuit of competitive excellence in professional football has evolved beyond instinct and tradition. This research explores the question: What makes a football team successful? — by adopting a team-centric machine learning approach grounded in performance analytics. Using a comprehensive dataset of Premier League player statistics from 1992 to 2019, the study aims to develop predictive models that can identify the key performance indicators (KPIs) that drive team success over time. Chapter I establishes the research background, problem statement, and objectives, emphasizing the growing relevance of artificial intelligence in modern football analysis. Chapter II presents a critical review of existing literature on sports analytics and machine learning, highlighting methodological gaps in explainable, team-focused success modelling. Chapter III details a structured methodology based on the CRISP-DM framework, encompassing data preprocessing, feature engineering, performance tier formulation, feature selection strategies, and supervised learning model development. Three supervised classification models-Logistic Regression, Random Forest, and Gradient Boosting—were implemented and evaluated using metrics including Accuracy, F1-Score, ROC-AUC, and confusion matrices. Ensemble learning techniques, including voting and stacking, were further explored to enhance predictive robustness. Model stability was assessed through 5-fold stratified cross-validation, and paired t-tests on cross-validated F1-scores indicated no statistically significant performance differences between models ($p > 0.05$). Gradient Boosting demonstrated consistently strong performance (mean F1-score ≈ 1.00), low variance across folds, and superior interpretability, supporting its selection as the primary base learner within the final ensemble framework. To address model transparency, SHAP (SHapley Additive exPlanations) was applied at both team and player levels, enabling granular interpretation of feature contributions to success predictions. The findings reveal that attacking efficiency, defensive stability, and disciplinary control consistently influence successful team outcomes. Beyond predictive accuracy, the study proposes practical decision-support extensions, like performance tiering, highlighting the real-world applicability of the framework. This project ultimately aims not only to predict success but to uncover why certain teams win—offering insights that could inform coaching, scouting, and strategy. The outcome is a step forward in applying AI to assist the beautiful game to further evolve.

Keywords: machine learning; explainable AI; SHAP; football analytics; gradient boosting; sports data mining

Introduction

1.1. Introduction

Football—often referred to as "the beautiful game"—is a sport that blends physical ability, tactical awareness, emotional intelligence, and nowadays increasingly, data-driven decision-making. As the sport continues to evolve, so too does the role of analytics in shaping team success, player recruitment, match strategy, and long-term planning. Football clubs across Europe and beyond are

 doi:10.20944/preprints202605.1076.v1

investing heavily in technology and data science to gain a competitive edge, reflecting a broader global trend where machine learning (ML) and artificial intelligence (AI) are becoming deeply embedded in sports decision-making and seen as a saviour of sorts (Gudmundsson & Horton 2017; Rein & Memmert 2016).

The English Premier League (EPL), arguably the most competitive football league in the world, has witnessed great progress in its' approach toward data utilization—not only in matchday tactics, but also in predicting team performance across entire seasons. Traditional methods of assessing performance, such as expert intuition or historical comparisons, are giving way to data mining (DM) and predictive modelling, which offer evidence-based insights into what makes a team successful (Lepschy & Woll & Wicker 2021). While fans and pundits may debate team performance based on goals and wins, data science and ML enables us to uncover deeper patterns—such as how squad rotation, player contributions, or disciplinary records affect a team's league position over time.

Despite increasing academic interest in sports analytics, The majority of research has concentrated on micro-level match events, such as passes, shots, possession, or tracking data, employing intricate spatiotemporal models (Bialkowski et al. 2014; Decroos et al. 2019). These models are often opaque, difficult to interpret, and reliant on proprietary datasets unavailable to most researchers. In contrast, team-level success over a full season, using publicly available data, remains underexplored, particularly through the lens of interpretable ML models. Understanding the broader dynamics that define a successful team—not just in one match, but across 38 games in a Premier League season—requires a different perspective.

This project aims to take a step in bridging that gap. This study uses machine learning and longitudinal performance data analysis to determine which player-derived characteristics have the most effects on football team performance. Beyond prediction, it also prioritizes explainability—because in football, as in AI, insights that cannot be understood or trusted by decision-makers are unlikely to be used.

1.2. Research Background

In recent years, sports analytics has transitioned from a niche academic interest to a core strategic function in professional football. Advances in computational power, the availability of data, and the growing competition in the football industry have all contributed to this change. Clubs such as Liverpool FC and Manchester City have invested heavily in AI-powered analysis to gain tactical and recruitment advantages (Anderson & Sally 2013; Müller et al. 2017). However, most existing research in football analytics falls into one of two categories: (i) event-based analysis, which uses specific match events like shots and passes (Pappalardo et al. 2019), and (ii) biometric or tracking-based analysis, which often requires costly GPS or sensor data (Bialkowski et al. 2014).

While these approaches provide valuable insights, they come with limitations. First, they are often not generalizable across teams or seasons due to variations in data quality or tactical context. Second, they typically involve black-box models that prioritize prediction over interpretation—meaning coaches and analysts may find it difficult to trust or apply their insights (Liu et al. 2021). Third, most studies rely on proprietary datasets owned by analytics companies, which limits academic replication and open research.

A more accessible and underutilized source of insight lies in aggregated player statistics—data such as total goals, minutes played, and appearances across a season. These metrics are widely available, easily understood, and historically consistent. Recent studies suggest that team-level success can be effectively modelled using such data (Lepschy et al. 2021; González-Rodenas et al. 2020). For instance, González-Rodenas et al. (2020) discovered that high-ranking clubs typically display greater spatial control and finishing efficiency over a season, while Lepschy et al. (2021) demonstrated that goals, pass completion rates, and defensive actions together predict league performance.

Yet, many of these studies still rely on either small datasets or limited feature sets. There is a clear opportunity to leverage larger historical datasets, apply feature selection and engineering

 doi:10.20944/preprints202605.1076.v1

techniques, and use machine learning models to identify what truly drives football teams to succeed. As suggested by Rein and Memmert (2016), the integration of domain expertise with transparent AI methods is key to advancing the field of sports intelligence.

Therefore, this research is motivated on developing interpretable ML models that analyze long-term team outcomes, such as final league position or total points in a season, using aggregated player statistics. Rather than building models that merely predict results, the aim is to uncover why certain teams succeed—thereby providing insights that are not only accurate but also actionable.

1.3. Problem Statement

In elite football competitions like the English Premier League (EPL), team success over a season is shaped by various performance metrics. With the growing use of machine learning (ML) in football analytics, researchers have applied models to predict outcomes and identify key performance indicators (KPIs) that differentiate winning from losing teams. For instance, Kyranoudis and Metaxas (2024) found that attacking efficiency and shooting accuracy were strong predictors of league success. Similarly, Song et al. (2024) demonstrated that shots on target and forward ball progression significantly impacted match results in the 2022 World Cup. However, most existing work is limited to event-level data and match outcomes, overlooking season-level patterns of team performance. Furthermore, a lack of accessible models and reproducible methods still limits practical adoption in real-world football analytics (Moustakidis et al. 2023). Thus, there is a clear need to propose suitable ML models that can classify team-level performance indicators derived from aggregated player statistics that contribute to season-long football team success.

Equally important is the issue of feature extraction and selection. Football data is often high-dimensional and noisy, and using all available features can reduce model performance and interpretability. Moustakidis et al. (2023) highlighted the importance of transparent models, using SHAP-based explainable AI to identify performance features that influence outcomes positively or negatively. Atta Mills et al. (2024) also demonstrated that using forward selection on over 150 features allowed researchers to retain only the most relevant attributes while improving model accuracy. Therefore, identifying optimal feature selection methods that balance performance and explainability is essential, especially when insights are intended for practitioners such as coaches and analysts.

Lastly, model performance depends not only on design and data but also on effective hyperparameter tuning. Studies such as Chandru et al. (2025) and Atta Mills et al. (2024) emphasize the role of tuning in boosting prediction accuracy and minimizing overfitting. For example, applying grid search and Bayesian optimization significantly improved the R^2 values of gradient boosting models in sports analytics. This highlights the need for a structured approach to hyperparameter optimization to ensure model robustness and generalizability in predicting football team success.

Although individual player stats provide the raw data, football results actually happen at the team level through how everyone works together, tactical balance, and overall squad makeup across an entire season. That's why focusing models only on single players can give incomplete or confusing insights. This research reframes that problem by combining player stats into team-level measures, so machine learning can spot the complete performance patterns that truly drive real competitive success.

1.4. Research Questions

Based on the identified challenges outlined in the problem statement, this study aims to answer the following questions:

1. Which machine learning models most effectively classify team-level performance features—derived from aggregated player statistics—that determine seasonal success in the English Premier League?

 doi:10.20944/preprints202605.1076.v1

2. What are the most effective feature selection techniques for identifying significant team-level success indicators from player data while preserving model interpretability through explainable AI methods?
3. How do hyperparameter tuning and ensemble learning techniques (voting and stacking) enhance the predictive performance, stability, and robustness of machine learning models for team success prediction?

1.5. Research Objectives

To address the above research questions, this project sets out the following objectives:

1. To propose the best machine learning models in classifying the most relevant team-level performance features, derived from aggregated player-based performance indicators that contribute to a team's seasonal success in the English Premier League.
2. To identify and justify optimal feature selection techniques that retain the most significant team-level indicators of success while maintaining model interpretability through explainable AI methods.
3. To optimize predictive performance of the machine learning models through hyperparameter tuning and ensemble learning techniques (voting and stacking), ensuring stability and robustness of the models.

1.6. Research Scope

This research is limited to analysing aggregated player-level performance data from the English Premier League (EPL) spanning the 1992 to 2022 seasons. The scope includes only quantitative, structured statistics—such as goals, assists, minutes played, and appearances—sourced from a publicly available Kaggle dataset. The study focuses on team-level success, as reflected by final league position or total season points, and other metrics as such, rather than individual player performance, thereby ensuring alignment with the objectives of this study.

It is important to note what this study does not cover. Advanced spatiotemporal data (e.g., xG, pass networks), transfer market values, and real-time tracking data are outside the scope of this research. The machine learning models developed are supervised and implemented using Python in Jupyter Notebook, with emphasis placed on interpretability over complexity.

1.7. Significance of the Study

In the evolving landscape of sports analytics, this study contributes to the growing body of knowledge by exploring how machine learning and explainable AI can be applied to publicly available football data to understand team-level success. This study shows that useful information can be obtained from organized player statistics, in contrast to many earlier studies that depend on intricate and frequently unavailable match-event or tracking datasets. It tackles one of the fundamental issues with AI adoption in real-world domains—the requirement for transparent and reliable systems—by emphasizing interpretable modeling (Liu et al. 2021; Molnar 2022). Just as crucial as properly forecasting team outcomes is being able to explain how particular traits affect them.

In a practical sense, the study is useful for data practitioners, coaching staff, recruitment teams, and football analysts. More evidence-based choices about player rotation, squad makeup, and season strategy planning can be influenced by the models and results. This can be applied along with gut instinct and intuition, which is the way forward in this era of sport. Furthermore, the research also provides a reference framework for academic institutions, sports startups, or smaller teams without access to proprietary systems because the entire analysis pipeline is built on open data and open-source tools. The study thus encourages the broader integration of AI in football outside of the elite level and democratizes sports analytics.

 doi:10.20944/preprints202605.1076.v1

1.8. Research Methodology

The research adopts a quantitative, experimental design using publicly available historical data and ML approaches. The goal is to identify and interpret the most relevant player-based features that contribute to team success across EPL seasons. The methodology adopted in this research follows the CRISP-DM framework, which consists of six key phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

Figure 1.1 below captures the CRISP-DM framework in a succinct manner.

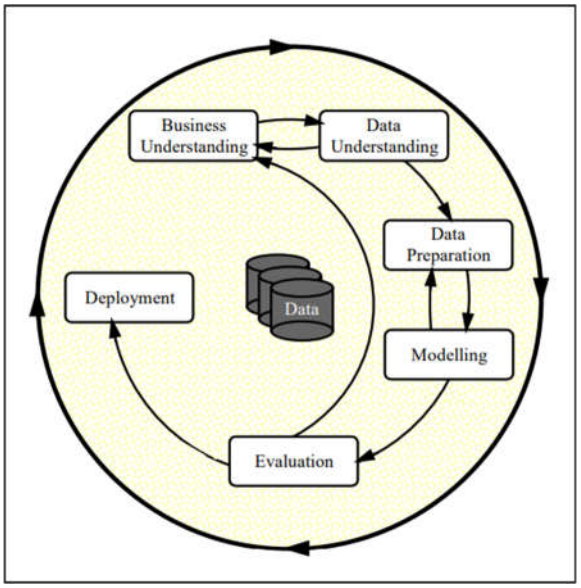


Figure 1.1. Phases of CRISP-DM Process Model for Data Mining. Source: Wirth 2000

For the scope of this dissertation, the focus remains on the first five phases, as deployment is beyond the current project's scope. Each of these phases plays a crucial role in structuring the research journey—from aligning the objectives with the problem domain to building and assessing predictive models.

Each step in the CRISP-DM process is operationalized through specific techniques. For instance:

1. The data understanding phase involves acquiring and exploring the Premier League dataset from Kaggle.
2. Data preparation covers handling missing values, data cleaning, and transformation.
3. Feature selection uses a combination of filter, wrapper, embedded methods, and other appropriate methods.
4. Model development implements the three best machine learning models—like Random Forest, Gradient Boosting, and Logistic Regression.
5. Model evaluation applies standard performance metrics such as accuracy, precision, recall, and F1-score to assess model effectiveness.

1.9. Organization of Thesis

This thesis is organized as follows:

 doi:10.20944/preprints202605.1076.v1

Chapter I: Introduction – It introduces the research topic, provides the background and motivation for the study, defines the research problem, formulates research questions and objectives, outlines the study's scope and significance, and briefly presents the research methodology.

Chapter II: Literature Review – It surveys relevant research in data mining, machine learning, and football analytics. It critically evaluates previous studies, identifies methodological gaps, discusses feature selection strategies, and justifies the selection of three best models for this study.

Chapter III: Methodology – It describes the dataset used, the preprocessing steps, the feature engineering process, model training pipelines, and evaluation metrics. It will also explain the experimental setup and tools used.

Chapter IV: Results and Discussion - This chapter will present the outcomes of the model implementations. It will include accuracy, performance comparisons, feature importance insights, hyperparameter optimisation, SHAP analysis and critical analysis of how each model performed in relation to the research questions.

Chapter V: Conclusion - This final chapter will summarise the findings, highlight the contributions of the study, discuss limitations, and propose future research directions for improving or extending the work.

Literature Review

2.1. Introduction

There has never been a greater need for computational techniques to discover valuable insights from the ever-increasing amount of data in professional sports. Data mining (DM) and machine learning (ML) approaches have evolved from a novel notion to an advisable practice in football, where team dynamics, player performance, and strategic decisions are crucial to success.

This chapter presents a review of relevant literature in the fields of data mining, machine learning, and their applications in football analytics. The intention is to place this study in the context of current academic discussions and draw attention to the strengths and drawbacks of previous studies. This chapter discusses machine learning applications in football, feature selection methods, and model optimization techniques relevant to the study objectives. Section 2.2 introduces DM and ML in an accessible manner. Section 2.3 then contextualises this within sports—specifically football analytics—before Section 2.4 provides a critical survey of key studies applying these technologies to real-world football problems.

2.2. Data Mining and Machine Learning

Data mining (DM) is the process of discovering patterns, trends, and associations in large datasets through statistical and computational techniques. It forms the foundation of many knowledge discovery processes in various domains, from healthcare to marketing—and now, increasingly, sports. DM involves steps such as data preprocessing, pattern recognition, and predictive modelling (Han et al. 2011).

Machine learning (ML), a subfield of artificial intelligence (AI), goes a step further. ML focuses on developing algorithms that can learn from data to make predictions or decisions without being explicitly programmed (Mitchell 1997). These algorithms fall into several broad categories:

1. Supervised learning, where the model learns from labelled input-output pairs;
2. Unsupervised learning, where the model detects hidden patterns in unlabelled data; and
3. Reinforcement learning, where agents learn to make sequences of decisions through trial and error, often applied in gaming or robotic simulations.

In football analytics, supervised learning is most frequently used to predict outcomes like team success, player performance, or match results (Bunker & Thabtah 2019). Algorithms such as decision trees, random forests, support vector machines (SVM), and ensemble methods like XGBoost are progressively more popular due to their balance of performance and interpretability (Molnar 2022).

 doi:10.20944/preprints202605.1076.v1

Although it frequently requires far larger datasets and computational power, deep learning has also been investigated in sports scenarios in recent years.

The strength of ML lies in its ability to handle complex, non-linear relationships, which are regularly seen in football. For instance, a striker's goal tally might not be the main cause for a team's success, but it can have a significant impact when combined with defensive stability, midfield creativity, and other variables. ML allows for data-driven understanding of these intricate relationships.

Three commonly employed machine learning models in sports analytics, particularly for structured data like that found in football, are Random Forest, XGBoost, and Support Vector Machines (SVM).

a. Random Forest: As an ensemble learning method, Random Forest builds multiple decision trees and aggregates their predictions. This approach enhances robustness, reduces overfitting, and can capture complex, non-linear relationships within football data, such as the intricate interplay between player statistics and team performance (Decroos et al. 2022; Jati et al. 2024). Its ability to handle high-dimensional data and provide feature importance scores makes it valuable for identifying key performance indicators in football.
b. XGBoost (Extreme Gradient Boosting): Another powerful ensemble technique, XGBoost has gained significant traction due to its high performance, speed, and efficiency in handling structured datasets. Studies have indicated XGBoost's superior prediction accuracy in football when leveraging comprehensive player and team statistics (Shrestha & Mahmood 2023). Its gradient boosting framework allows it to correct errors from previous trees, leading to highly accurate predictions for classification tasks like predicting match outcomes or team success over a season.
c. Support Vector Machines (SVM): SVMs are effective for classification tasks, particularly when there is a clear margin of separation between classes. While a comparative study outside football analytics suggested SVM can outperform Random Forest and XGBoost in certain sentiment analysis tasks, its performance in football-specific contexts can vary (Syihabuddin et al. 2023). SVMs are known for their strong theoretical foundation and ability to generalize well, even with limited data, provided the appropriate kernel is selected.

The choice of the ‘best’ model in football analytics often depends on the specific problem (e.g., predicting match results, player valuation, seasonal success) and the characteristics of the dataset. While XGBoost and Random Forest often lead in predictive accuracy for structured sports data, Logistic Regression can serve as a highly interpretable baseline model.

However, a frequent criticism of ML in practice is its "black box" nature—many models can make accurate predictions but cannot explain why they make those predictions. This presents a challenge in sports, where coaches and analysts require actionable insights and justifications for predictions. This has led to a growing emphasis on model interpretability and Explainable Artificial Intelligence (XAI) techniques.

In the context of English Premier League (EPL) data, XAI tools help in understanding why a model makes a particular prediction about a team's or player's performance. For instance, in Fantasy Premier League (FPL) analytics, XAI is being developed to translate prediction metadata into clear, understandable explanations, addressing the black box problem and aiding data-driven decision-making (Klingstedt 2024). Techniques like SHAP (SHapley Additive exPlanations) are critical for providing transparent insights into the contribution of individual features (e.g., goals, assists, defensive actions) to a model's output in complex football scenarios. In this study, ML is not used merely to predict team success in the EPL, but to explain it-using models that are interpretable and grounded in realistic, season-wide team-level data.

2.3. Sport Analytics and Football

The rigorous analysis of sports data to enhance strategy, performance, and decision-making is known as sports analytics. In the early 2000s, the "Moneyball" strategy in baseball garnered

 doi:10.20944/preprints202605.1076.v1

worldwide notoriety, but it has since developed into a multidisciplinary field that includes computer science, statistics, biomechanics, and psychology (Alamar 2013). These days, analytics are fundamental not only in American sports, but also to international sports like basketball, cricket, and notably football.

Many stakeholders initially opposed the analytics movement in football because they believed the sport was too complicated and fluid for quantitative reduction (Anderson & Sally 2013). Football is difficult for traditional metrics to measure since it is continuous, low-scoring, and devoid of distinct "events," unlike basketball or baseball. This landscape has changed, though, with the advent of tracking technologies and comprehensive performance statistics.

Today, football analytics has diversified into several streams, including:

a. Tactical analysis using positional data (Bialkowski et al. 2014),
b. Scouting and recruitment using performance indicators (Müller et al. 2017),
c. Injury prediction and load management through biometric monitoring.
d. Outcome prediction using DM and ML techniques (Lepschy et al. 2021; Groll et al. 2019).

Football analytics relies heavily on a comprehensive understanding of performance indicators (PIs) and their relationship to team success. These indicators can be broadly categorized to provide a structured view of a team's strengths and weaknesses over a season.

i. Attacking Indicators: These include metrics such as Goals Scored, Expected Goals (xG), Shots, and Shots on Target. Goals scored consistently show the highest predictive value for a team's final ranking over a season (Groll et al. 2019). Shots on target, alongside other offensive metrics, are key factors influencing the likelihood of winning matches (Poulopoulou et al. 2024).
ii. Defensive Indicators: Key defensive metrics often encompass Tackles, Interceptions, Clearances, and Pressures. While not always directly linked to seasonal success in the same way as goals, effective defensive performance reduces the opponent's chances and contributes to overall team stability.
iii. Possession-Based and Passing Indicators: These include Ball Possession, Total Passes, and Completed Passes. Studies have shown that winning teams often demonstrate superior performance in ball possession and passing metrics (Poulopoulou et al. 2024). "Packing" (sum of opponent players overcome by each successful pass) and "Passes completed" are highly correlated with a team's final ranking, highlighting the importance of efficient ball movement and progression (Groll et al. 2019). Higher-ranked teams generally engage more in actions involving ball possession (Lohmann et al. 2022).
iv. Physical Indicators: Metrics like Total Distance Covered, High-Speed Running (HSR), and Sprinting Distance provide insights into a team's physical output. While these can fluctuate throughout a season, higher-ranked teams are often observed to cover more distance and high-speed running distance while in ball possession (Lohmann et al. 2022).

Prior studies have consistently linked these categories of features to team performance over a full season. For example, analysis of the Greek Football League demonstrated that winning teams exhibited superior performance in attacking and passing metrics (Poulopoulou et al. 2024). Similarly, research on season-long team performance consistently shows that core attacking metrics (e.g., goals, xG) and efficient ball progression (e.g., passes, packing) are strong predictors of a team's final league position (Groll et al. 2019).

A significant shift in contemporary football analytics is the transition from evaluating players in isolation to understanding their contribution to the collective unit. Research by Lepschy et al. (2021) suggests that the true value of player-based metrics lies in their ability to serve as proxy indicators for team-level tactical execution. Therefore, the classification of team success requires an analytical framework that can aggregate these individual performance signals—such as high-intensity defensive actions or creative passing volumes—to predict overall league outcomes.

Regardless of such advancements, the majority of football analytics research still uses proprietary datasets like Opta, StatsBomb, or Second Spectrum, which restricts academic research

 doi:10.20944/preprints202605.1076.v1

and reproduction in environments with limited resources. A more accessible path for research and innovation is offered by publicly accessible datasets, such as those from Kaggle. They help close the gap between top-tier analytics and more general scholarly contributions by enabling scholars to use open data to investigate noteworthy issues.

Crucially, a lot of research is moving from match-level prediction (win/loss, for example) to season-level analysis, where the objective is to comprehend the factors that influence long-term team success rather than only forecast an outcome (González-Ródenas et al. 2020). The evolution of data collection from basic match statistics to more granular event and tracking data has enabled deeper analysis, revealing how player-derived features, when aggregated at the team level, significantly contribute to success over a full Premier League season. This closely aligns with the aims of our current research.

2.4. Football Analytics Using Data Mining and Machine Learning

There have been numerous studies done on football analytics in various forms and touching upon several aspects of the field. Some of the key investigations are presented below.

Bunker and Thabtah (2019) proposed a ML framework to predict outcomes of football matches using supervised classification models like Naïve Bayes and Decision Trees. Their study utilized data such as team form, goals scored, and home/away status. While the results demonstrated solid predictive accuracy, especially with ensemble methods, the authors noted limitations around generalizability to other leagues and seasons. This work is foundational for understanding the framing of football outcomes in ML terms, but its focus on match-level data highlights a gap that my season-level analysis aims to address.

Lepschy et al. (2021) analysed Bundesliga teams over multiple seasons, examining the relationship between match statistics (e.g., goals, passes, tackles) and final league position. Using multiple regression, they identified offensive efficiency as the most critical success factor. What stands out is their emphasis on season-level aggregates, which aligns closely with my approach. However, their methodology remains largely statistical. In our analysis, the use of interpretable ML can further enhance the depth and practical use of such findings.

In their tactical study of the English Premier League, González-Ródenas et al. (2020) found that top teams exhibited superior spatial dominance and shot quality, using event data and network metrics. While not strictly an ML study, it's a great example of domain-informed feature engineering. Their approach to tactical dimensions—like high press and shot zones—can inspire additional derived features in the model we will build. However, this level of event-level data is often unavailable, which is why our focus on publicly accessible season-level datasets offers a more generalisable pathway.

Decroos et al. (2019) introduced a framework for valuing on-ball football actions using a probabilistic model trained on match events. Their VAEP (Valuing Actions by Estimating Probabilities) system assesses each player contribution's effect on scoring or conceding. While innovative, the method relies heavily on proprietary event data and complex modelling, limiting practical use in more accessible settings. The ingenuity is admirable, but our study's value can be seen in demonstrating how more straightforward approaches on open data can also yield useful insights.

Pappalardo et al. (2019) published a public dataset of spatio-temporal match events and showed how ML could model passing behaviours and goal likelihoods. They introduced PlayeRank, a data-driven player ranking system based on multiple performance factors. Although PlayeRank offers an exciting direction, its utility is still player centric. This research, by contrast, elevates the analysis to the team level—specifically by aggregating player performance across the season to understand organizational success dynamics.

This comprehensive survey by Gudmundsson and Horton (2017) mapped out how spatio-temporal analysis has transformed team behaviour modelling in sports. They highlight the value of trajectory data, motion tracking, and heatmaps to quantify performance. While their review is

 doi:10.20944/preprints202605.1076.v1

primarily descriptive, it sets a context for how cutting-edge analytics has moved beyond just scores and into behavioural modelling. Our research is more constrained in data type, but it carries forward this intent by using accessible stats to model behavioural patterns indirectly.

Rein and Memmert (2016) argued for the integration of big data and tactical analysis in elite soccer, noting that too many ML approaches are either too opaque or not aligned with coaching needs. They called for more transparent and practical models. This is one of the most directly motivating papers for this study—it emphasizes interpretability as not just an academic goal but a practical necessity. This study strongly aligns with their critique and respond to it by prioritizing explainable ML in our model design.

Müller et al. (2017) developed models to estimate player market value using a mix of crowd-based and performance-based metrics. Their paper bridges sports analytics and economics, offering a multidimensional perspective on what 'value' means in football. While their focus is on individual valuation, their methodology—combining data sources and weighting them for predictive validity—offers inspiration for our feature engineering stage. Their work validates the idea that intelligent modelling can surface hidden value drivers.

Liu et al. (2021) critically evaluated the application of interpretable ML in sports, outlining best practices for balancing prediction and explanation. They emphasized models like SHAP and LIME for post-hoc interpretability and warned against blindly trusting 'black box' metrics. This paper serves as a technical foundation for our model explainability layer. It reassures us that even with limited data, transparency in modelling can add academic and practical legitimacy to AI applications in football.

Molnar (2022) has a widely cited book on interpretable ML which is both a technical and philosophical guide to explainable modelling. While not sports-specific, its inclusion in this review is justified by the methodological backbone it offers. His chapters on feature importance, model-agnostic tools, and local vs. global interpretability will guide the justification of algorithm choices in my work. In particular, the reason of use of SHAP values is directly motivated by Molnar's practical and balanced discussion of XAI tools.

Zhao et al. (2021) conducted a comprehensive study on predicting football match outcomes using ensemble machine learning techniques, comparing models such as Gradient Boosting, AdaBoost, and Random Forests. Their results showed that ensemble models consistently outperformed single learners in both accuracy and robustness. The dataset, although limited to match-level data, provided diverse features including team ratings, player stats, and historical outcomes. A key strength of this study is its emphasis on model evaluation metrics beyond just accuracy, such as F1-score and ROC-AUC, which are particularly important in imbalanced datasets. While the focus remained on predicting single-match results, the methodological rigor and use of interpretable ensemble models make this study a valuable reference for modelling football success at broader levels.

Schumaker et al. (2016) introduced a novel approach for predicting English Premier League outcomes by combining sentiment analysis with statistical features. They mined Twitter data alongside traditional match stats to develop hybrid features used in a Support Vector Machine classifier. The inclusion of unstructured data represents an innovative direction in sports analytics, showing how public sentiment can be quantified and integrated into predictive systems. Despite limitations in replicability due to the nature of social media APIs, the study highlights the growing intersection of natural language processing and sports forecasting. For long-term success modelling, such sentiment-derived features could eventually inform player or team morale indicators, adding a new layer to team-based analyses.

Baio & Blangiardo (2010) applied Bayesian hierarchical modelling to assess team performance in football using goal-scoring data from Italian league Serie A. Their probabilistic framework allowed for uncertainty estimation, making it particularly suitable for sports contexts where outcomes are inherently stochastic. Although their work predates the widespread use of machine learning in football, the Bayesian approach offers an alternative to deterministic modelling and can be adapted

 doi:10.20944/preprints202605.1076.v1

within ensemble or hybrid systems. Their use of posterior distributions and team-level priors lays a theoretical foundation for integrating prior domain knowledge with ML in future studies. It also suggests potential improvements in uncertainty quantification for football success prediction.

Ribeiro et al. (2020) explored the dynamics of scoring in football using statistical mechanics and time-series modelling. Their large-scale analysis across multiple leagues showed universal patterns in goal timing and match flow. While not a ML study per se, their quantitative insights provide valuable features—such as goal timing distributions and scoring burst patterns—that can be extracted for use in ML models. The interdisciplinary nature of their work bridges physics and sports science, encouraging broader perspectives when selecting temporal or behavioural variables for football modelling.

Haghighat et al. (2013) proposed a novel feature selection and ensemble learning framework for sports prediction using the Euro 2012 football dataset. They combined information gain, principal component analysis (PCA), and genetic algorithms to select optimal features before classification. The use of biologically inspired methods for feature reduction is particularly relevant given the high dimensionality of sports data. Their work provides a compelling blueprint for balancing dimensionality reduction with predictive performance, which is especially relevant for this study's aim to build interpretable yet efficient models from a large player statistics dataset.

Groll et al. (2018) developed a structured additive regression model to predict outcomes in international football tournaments, incorporating both historical and current covariates such as team FIFA rankings, market value, and coach nationality. Their statistical modelling approach achieved high predictive power while maintaining interpretability, positioning it as a strong alternative to black-box ML. Particularly, their utilization of domain-informed variables aligns with best practices in applied sports analytics, where context is as important as computation. This paper reinforces the value of hybrid methods that combine domain knowledge with algorithmic precision.

Perin et al. (2013) emphasized the importance of visual analytics in sports data, developing systems to explore temporal and performance trends across football leagues. Though their contribution focused on data visualization rather than predictive modelling, their work is crucial in the preprocessing and exploratory stages of ML. Interactive data exploration allows analysts to uncover hidden trends, evaluate outliers, and refine feature selection strategies. In the context of this research, integrating such exploratory tools during the data cleaning and feature engineering phase can substantially improve model design and interpretation.

Van Haaren & Davis (2012) worked on player-specific action modelling using Hidden Markov Models (HMMs) trained on pass and movement sequences. Their approach enabled the classification of player roles and playing styles, offering rich features beyond conventional stats. While the data requirements for HMMs are significant and often proprietary, their concept of deriving latent player behaviours aligns with the broader trend of moving from raw stats to behavioural metrics. The challenge remains in adapting such methods for public datasets, but the methodological value of capturing player actions temporally is undeniable.

Milošević et al. (2022) examined football analytics through a social network lens, using passing networks and player influence metrics to evaluate team dynamics. Their findings show that centrality measures and pass distribution balance are strong indicators of team success. Integrating such graph-based features into ML models represents a promising direction, especially when player relationships and on-pitch interactions are key to overall performance. While such analysis typically requires event-level data, simplified proxies—like passes per player or assists from specific zones—can be engineered in aggregated datasets.

Chatterjee et al. (2022) implemented deep learning architectures, particularly LSTM (Long Short-Term Memory) networks, to model temporal patterns in football results over seasons. Their focus was on capturing sequential dependencies and momentum patterns, a key limitation in many static models. While deep learning models require significant data and may sacrifice interpretability, they provide valuable contrast to traditional ML approaches. The study serves as a reminder that while interpretable models are central to this research, there is merit in comparing their performance to

 doi:10.20944/preprints202605.1076.v1

more complex baselines. Table 2.1 below presents the literature review in a nutshell, along with their interpretation.

Table 2.1. Literature Review Summary on Football Analytics using DM and ML.

No.	Author(s)	Objective	Data Used	Algorithms/ Techniques	Key Findings/ Results
1	Bunker & Thabtah (2019)	Predict football match outcomes	EPL match data	Naïve Bayes Decision Tree	High accuracy with ensemble methods; limited generalisability
2	Lepschy et al. (2021)	Link match stats to league rankings	Bundesliga seasons	Regression models	Offensive efficiency most critical to team success
3	González-Ródenas et al. (2020)	Compare tactical attributes of top vs. bottom teams	EPL event data	Network metrics	Top teams use spatial dominance and shot quality
4	Decroos et al. (2019)	Value player actions using probabilities	Proprietary event data	Probabilistic model (VAEP)	Valuable modelling; low accessibility
5	Pappalardo et al. (2019)	Rank players based on performance	Open spatio-temporal match events	PlayeRank ML model	to be continued… focused ranking system
6	Gudmundsson & Horton (2017)	Review spatio-temporal sports analysis	Multiple sports datasets	Trajectory analysis, heatmaps	Useful for team behaviour patterns
7	Rein & Memmert (2016)	Bridge gap between ML and coaching needs	Elite match data	Conceptual critique	Call for interpretable, practical models
8	Müller et al. (2017)	Estimate player market value	Transfer and performance data	Regression + weighting	Combines economics with player stats
9	Liu et al. (2021)	Promote interpretable ML in sports	Various sports datasets	SHAP, LIME	Stresses need for explainability
10	Molnar (2022)	Guide on explainable AI	General ML applications	SHAP, LIME, XAI	Influential XAI resource
11	Zhao et al. (2021)	Predict match results using ensembles	Match-level football data	RF, AdaBoost, GBM	Ensemble models outperform single models
12	Schumaker et al. (2016)	Combine sentiment and stats for prediction	Twitter + EPL stats	SVM + NLP	Sentiment boosts predictions
13	Baio & Blangiardo (2010)	Bayesian prediction of football outcomes	Serie A goal data	Bayesian hierarchical model	Quantifies uncertainty effectively
14	Ribeiro et al. (2020)	Study scoring dynamics across leagues	Global football matches	Statistical mechanics	Universal goal timing patterns
15	Haghighat et al. (2013)	Feature selection for sports prediction	Euro 2012 data	PCA, GA, Ensembles	Improves accuracy through FS

 doi:10.20944/preprints202605.1076.v1

16	Groll et al. (2018)	Predict tournament outcomes	FIFA rankings + covariates	Structured regression	High predictive validity
17	Perin et al. (2013)	Enable visual sports analytics	Multiple leagues	Data visualisation	Improves understanding
18	van Haaren & Davis (2012)	Model player actions temporally	Match sequences	HMM	Classifies player styles
19	Milošević & Šćepanović (2022)	Model team play as networks	Passing data	Graph theory	Passing balance correlates with success
20	Chatterjee & Dey (2022)	Use LSTM for football prediction	Season-level time-series	LSTM (DL)	Captures momentum; less explainable

Based on the critical review of literature and Table 2.1 above, three machine learning models have been identified as the most suitable for application to the Kaggle Premier League dataset (1992–2019): Random Forest, Gradient Boosting, and Logistic Regression. Random Forest is selected for its robust performance across structured datasets and its built-in interpretability features (Zhao et al. 2021; Bunker & Thabtah 2019). XGBoost, a powerful gradient boosting algorithm, is chosen for its handling of class imbalance and superior predictive accuracy as evidenced in multiple sports prediction studies (Haghighat et al. 2013). Logistic Regression, while simpler, provides a strong interpretable baseline and has been validated in football analytics studies for its reliability (Lepschy et al. 2021; Groll et al. 2018). These models will form the core of the experimental phase of this research.

While early machine learning applications in football focused on individual player rankings, recent literature emphasizes the need for 'Team-Centric' modeling. For instance, González-Rodenas et al. (2020) demonstrated that team success is not merely a sum of elite players but a result of specific performance clusters, such as spatial control and shot efficiency, that emerge when player data is viewed at a squad level. This study adopts this perspective, treating aggregated player statistics as features for classifying a team's seasonal success category.

2.4.1. Random Forest

As shown by (Breiman 2001), Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve accuracy and control overfitting. Each decision tree in the forest is trained on a random subset of the data using a technique called bootstrap aggregating, or bagging. Additionally, during the tree-building process, only a random subset of features is considered at each node, which introduces further diversity.

The key intuition behind Random Forest is that by combining many weak learners (individual decision trees), the overall model becomes a strong learner that generalizes better. The final prediction is made by majority vote for classification tasks or averaging for regression. Figure 2.1 below illustrates the working of the algorithm.

 doi:10.20944/preprints202605.1076.v1

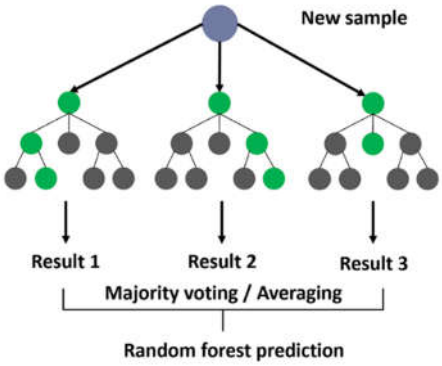


Figure 2.1. Working of Random Forest Model (Medium, 2025).

Random Forest offers several merits:

a. It is relatively robust to outliers and noise.
b. It handles both numerical and categorical data well.
c. It provides internal estimates of feature importance.

A simple formula representing the ensemble prediction of a Random Forest classifier for a new instance x is:

$$y\hat{} = mode\{h1(x), h2(x)..., h_T(x)\} \quad (2.1)$$

Where:

a. $h_t(x)$ is the prediction of the t^{th} tree,
b. T is the total number of trees in the forest.

Furthermore, Random Forest can handle multicollinearity and missing data to some extent, which is valuable when working with real-world sports datasets that often include gaps or redundant statistics.

2.4.2. XG: Boost

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of gradient boosted decision trees, as established by Chen and Guestrin (2016). It improves on traditional boosting algorithms through several enhancements: regularization, parallel processing, tree pruning, and sparse-aware learning. Figure 2.2 below shows the decision tree ensembles used by the XGBoost algorithm using a simple example. It classifies whether someone will like a hypothetical computer game X.

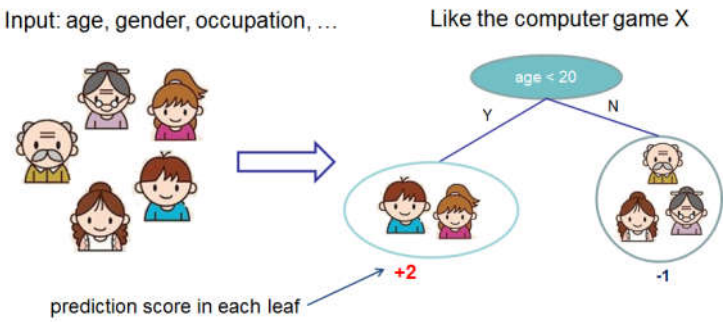


Figure 2.2. Decision Tree Ensembles used by XGBoost (xgboost developers, 2022).

The model builds trees sequentially, where each new tree is trained to correct the errors of the previous one by minimizing a specific objective function:

 doi:10.20944/preprints202605.1076.v1

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{2.2}$$

Where:

a. 'l' is differentiable loss function (e.g., logistic or squared error)
b. $\hat{y}_i^{(t)}$ is the prediction of the t^{th} tree
c. $\Omega(f_k)$ is a regularization term controlling model complexity.

XGBoost is highly regarded for:

a. Its state-of-the-art accuracy on structured/tabular data,
b. Built-in handling of missing values,
c. Flexibility in defining custom objective functions.

These successes are relevant to this research because football team success is often influenced by non-linear combinations of variables — such as team cohesion, previous form, and financial investments — that simpler models may fail to capture.

Unlike Random Forest, which builds trees in parallel, XGBoost builds them sequentially, allowing it to focus on correcting previous errors, thereby increasing performance in more complex problems. Additionally, XGBoost includes a built-in method to rank feature importance using gain, cover, or weight, providing insight into the most influential factors in predicting team performance. This aligns with the feature analysis goals of this research, where understanding why a team performs well is just as important as predicting if it will.

2.4.3. Logistic Regression

Logistic Regression is one of the most widely used algorithms for binary classification problems and serves as a fundamental baseline in machine learning. It estimates the probability that a given input belongs to a particular class using the logistic (sigmoid) function. Unlike linear regression, which predicts a continuous outcome, logistic regression maps predictions to a [0, 1] range:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{2.3}$$

Where:

a. β_0 is the intercept,
b. β_1, ……. β_n are the coefficients for each feature.

The model parameters are typically estimated using maximum likelihood estimation (MLE). Logistic regression is appreciated for its:

a. Interpretability, as coefficients indicate direction and strength of influence,
b. Efficiency, even on large datasets,
c. Robustness when regularization (L1, L2) is applied.

In football analytics, Logistic Regression has been used in a variety of contexts, from predicting match outcomes (Tax & Joustra 2015) to estimating the probability of in-game events such as goal scoring or player substitution (Molina & García 2020). For this study, Logistic Regression is important not because it will necessarily outperform more complex models like XGBoost, but because it provides a transparent and interpretable benchmark. It allows researchers and stakeholders—such as coaches or analysts with limited technical background—to understand how variables like goals per match, possession, or defensive metrics are associated with success probability.

 doi:10.20944/preprints202605.1076.v1

Moreover, Logistic Regression supports regularisation techniques (e.g., L1 or L2), which can help prevent overfitting and assist in feature selection, making it an ideal model to pair with our earlier discussed filtering and embedded methods. Figure 2.3 below represents the sigmoid function used in Logistic Regression graphically.

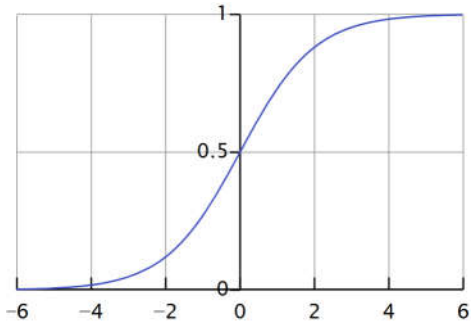


Figure 2.3. Sigmoid function curve with decision boundary.

2.4.4. Performance Comparison between Machine Learning Models

This section delves into specific ML models, analysing their applicability and performance characteristics in football analytics. Table 2.2 below compares the performance of the best ML models.

Table 2.2. Performance Comparison of ML Models in Football Analysis.

Model	Strengths in Football Analytics	Weaknesses in Football Analytics	Interpretability	Computational Cost
Logistic Regression	Simple, highly interpretable baseline model for binary classification (e.g., win/loss, top/bottom-half finish) (Klingstedt 2024).	May struggle with complex, non-linear relationships. Less accurate than ensemble methods for complex tasks.	High: Coefficients indicate the direction and strength of feature influence.	Low
Random Forest	Robust to overfitting; handles non-linear relationships and feature importance scores. Effective for both classification and regression (Decroos et al. 2022).	Less interpretable than Logistic Regression (ensemble of many trees). Can be computationally intensive for very large datasets.	Moderate: Feature importance is clear, but individual tree logic is complex.	Moderate
XGBoost	High performance and accuracy, especially on structured data. Fast and scalable. Effectively handles missing values and complex interactions (Shrestha & Mahmood 2023; Decroos et al. 2022).	Less interpretable than Logistic Regression. Requires careful hyperparameter tuning to avoid overfitting.	Moderate: Feature importance is available, but the boosting process is complex.	Moderate to High
SVM	Effective in high-dimensional spaces. Good for clear classification boundaries.	Can be sensitive to hyperparameter tuning and choice of kernel. Less scalable to very large datasets.	Low: Predictions are based on complex non-linear transformations.	Moderate

For EPL-style structured data, which often contains numerous features and potentially complex relationships, Random Forest and XGBoost are highly applicable. Their ensemble nature allows them to capture the nuanced patterns in football performance. While they offer high predictive accuracy,

 doi:10.20944/preprints202605.1076.v1

their "black box" nature can make it challenging to fully understand the exact reasoning behind a prediction. Conversely, Logistic Regression remains valuable for its simplicity and direct interpretability, making it useful for establishing baseline models or for scenarios where clear, transparent insights into feature influence are paramount. The use of XAI techniques, as discussed in Section 2.2, helps to mitigate the interpretability challenges of the more complex ensemble models.

2.4.5. Ensemble Learning in Sports Analytics

Ensemble learning is a machine learning concept where multiple models are combined to achieve better predictive performance than could be obtained from any single model. This approach is widely adopted in sports analytics due to several key advantages:

i. Enhanced Accuracy: By pooling the strengths of diverse algorithms, ensemble models typically produce forecasts that are more dependable and accurate, often leading to a significant reduction in error rates compared to individual models (Webb & Zheng 2024; Jati et al. 2024).
ii. Robustness: Ensemble methods reduce the impact of individual model errors and are more robust to noise and outliers commonly found in sports data. This leads to more trustworthy predictions, even in complex and sometimes inconsistent datasets (Constantinides 2022; Wong 2025).
iii. Reduction of Overfitting: Techniques like bagging (e.g., Random Forest) and boosting (e.g., Gradient Boosting) help in reducing overfitting, where a model performs well on training data but poorly on unseen data. By averaging predictions or iteratively correcting errors, ensemble models generalize better to new data (Alghamdi et al. 2025; Li et al. 2025).
iv. Flexibility and Scalability: Ensemble modelling is highly flexible, allowing the combination of various model types (e.g., decision trees, linear models). Algorithms like XGBoost are designed to handle large datasets efficiently, making them suitable for real-world sports applications with extensive data (Wang et al. 2024).

Random Forest is a prime example of a bagging ensemble algorithm, while XGBoost exemplifies boosting, both proving highly effective in various sports prediction tasks.

2.5. Feature Selection Techniques

Feature selection is a fundamental preprocessing step in any ML pipeline, particularly in sports analytics where datasets often comprise a wide range of variables including team statistics, player metrics, and contextual match data. Feature selection is the process of identifying and retaining the most relevant variables in a dataset while discarding irrelevant or redundant ones. It plays a crucial role in enhancing model performance, reducing computational cost, and improving interpretability. These techniques are vital for Objective 2, which involves identifying key performance indicators.

Techniques can be broadly categorized into filter, wrapper, and embedded methods. Filter methods assess features based on their intrinsic properties, typically using statistical tests or information-theoretic metrics. Wrapper methods use predictive models to evaluate feature subsets, while embedded methods perform feature selection during the model training process itself (Guyon & Elisseeff 2003).

2.5.1. Filter-Based Methods

Filter methods use statistical criteria to evaluate feature relevance without involving any ML algorithm. These techniques are generally fast, scalable, and model-independent. Common filter methods include:

i. Mutual Information (MI): measures how much information the presence or absence of a feature contributes to making the correct prediction.
ii. Chi-Square Test: It assesses independence between categorical variables and the target.
iii. Pearson Correlation: It gives the linear correlation between numerical features and the target variable.

 doi:10.20944/preprints202605.1076.v1

For example, Mutual Information is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (2.4)$$

Filter-based approaches are especially relevant when working with large datasets such as sport datasets, which includes hundreds of players and team-level statistics. Given the dataset's complexity and noise, using MI or correlation-based filtering will form the first step in identifying relevant features for model training.

2.5.2. Wrapper Methods

Wrapper methods evaluate subsets of features by actually training models on them and selecting the subset that gives the best performance. These methods are model-specific and computationally intensive but often more accurate. Common techniques include:

i. Forward Selection: Starts with no features and adds them one at a time.
ii. Backward Elimination: Starts with all features and removes them one at a time.
iii. Recursive Feature Elimination (RFE): Iteratively removes the least important feature based on model performance.

These methods are best used when training time is not a major constraint.

A notable example is the work by Szymański and Kajdanowicz (2017), who applied RFE with support vector machines to classify athlete performance levels. Their study highlighted how wrapper methods can significantly improve prediction accuracy compared to using all features. In the context of football, this is valuable when aiming to identify a minimal set of features that most accurately represent a team's success, without redundancy.

RFE will be particularly valuable for this research since it can be directly applied to the selected models—RF and Logistic Regression—and can help identify a subset of features that both contribute to performance and remain interpretable.

2.5.3. Embedded Methods

Embedded methods integrate feature selection directly into the model training process. LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization method that adds a penalty to the absolute size of coefficients, effectively driving some to zero, thereby performing variable selection (Tibshirani 1996). In sports analytics, LASSO has been used to predict injury risk and performance variability by selecting only the most significant predictors from large physiological datasets (Malfait et al. 2018).

The LASSO loss function is:

$$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2.5)$$

Where λ controls the regularization strength. Figure 2.4 below depicts the behaviour of LASSO models under different constraints, and how the coefficients vary as the regularization strength changes.

 doi:10.20944/preprints202605.1076.v1

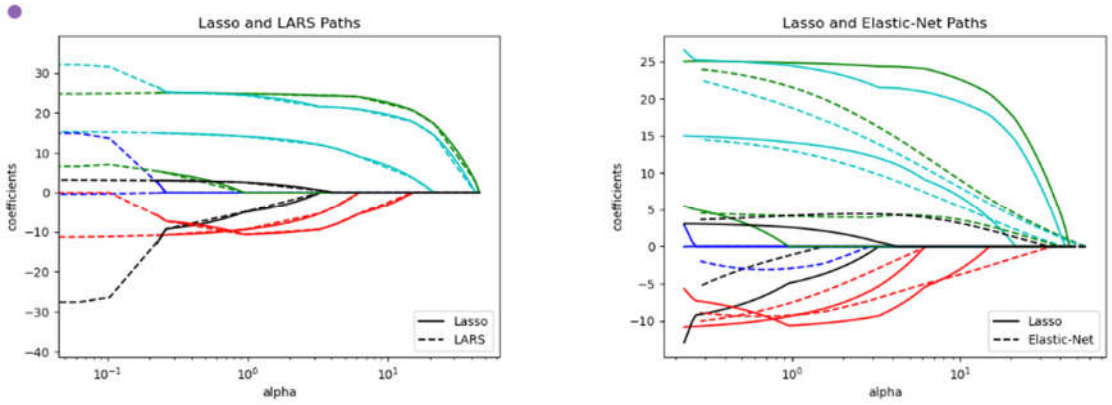


Figure 2.4. Coefficient Shrinkage Comparison (Ridge vs. LASSO). Source: Lasso, Lasso-LARS, and Elastic Net Paths 2025.

Another embedded method is tree-based feature importance, commonly available in Random Forest and Gradient Boosting models. These methods rank features based on how frequently they are used for node splits, thus offering a direct, intuitive sense of variable relevance. Groll et al. (2018) used such tree-based models in football tournament forecasting, demonstrating their capability to highlight key performance indicators like possession percentage and pass completion rate.

Given the need for explainable yet effective models in this study, embedded methods—especially those available in Random Forest and XGBoost—will serve as both feature selectors and predictors, making them efficient and aligned with the research's twin goals of performance and interpretability.

2.5.4. Challenges with Noisy Data in Sports Datasets

Real-world sports data, particularly from sources like GPS trackers, optical tracking systems, and event streams, often present challenges. These datasets can be noisy due to measurement errors, incomplete records, or inconsistencies (Hussain et al. 2022). Furthermore, many performance indicators are highly correlated (e.g., total distance covered, high-speed running, and sprinting distance from GPS data), leading to multicollinearity issues. Multicollinearity can destabilize models, make feature importance difficult to interpret, and lead to unreliable conclusions (Malcata & Reade 2019). Effectively dealing with these issues is paramount for building robust and accurate predictive models in football analytics.

2.5.5. Other Mitigation Techniques

Various dimensionality reduction and filtering techniques are employed to overcome these challenges and enhance model performance and interpretability:

i. Recursive Feature Elimination (RFE): RFE is a wrapper-type feature selection method that works by recursively training the model and eliminating the least important features until the desired number of features is reached (Guyon et al. 2023). In football analytics, RFE can be used to identify the most crucial on-ball and off-ball actions, or physical metrics, that contribute most significantly to team success, thus reducing data dimensionality and improving model efficiency (Zhang et al. 2025).

ii. Principal Component Analysis (PCA): PCA is a powerful dimensionality reduction technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables called principal components, capturing most of the variance in the data. In sports analytics, PCA is instrumental for:

 doi:10.20944/preprints202605.1076.v1

a. Data Reduction and Noise Filtering: It distils high-dimensional data from various sources (video feeds, sensor data, performance statistics) into a manageable set of components, filtering out noise and focusing on significant patterns (Wong 2025).
b. Clustering Player Performance Metrics: PCA can project multidimensional player statistics onto a lower-dimensional space, making it easier to identify clusters representing similar performance profiles, aiding in player comparison and tactical decisions (Jati et al. 2024).
c. Injury Prediction: By isolating subtle shifts in performance indicators via principal components derived from historical data, PCA can help predict potential injury risks (Wong 2025). For example, the first component might represent physical fitness, the second technical abilities, and the third tactical awareness (Jati et al. 2024).

iii. SHAP (SHapley Additive exPlanations): While primarily an explainability tool, SHAP values can also implicitly aid in feature understanding by quantifying the contribution of each feature to a model's prediction. In football, SHAP helps explain individual player or team performances by revealing which specific actions or statistics most influenced a model's outcome, thereby acting as a powerful tool for feature insights, particularly with complex "black box" models.

Recent examples in football analytics leverage these techniques to refine datasets. For instance, studies might use PCA to combine various physical attributes into a single "fitness" component or apply RFE to select the most impactful offensive and defensive statistics for predicting league positions. These methods help improve model performance by eliminating noise and irrelevant features, enhancing interpretability, and speeding up the training process (Analytics Vidhya 2018).

2.6. Conclusion

This chapter has presented a comprehensive review of the literature on the amalgamation of data mining, ML, and football analytics. Beginning with Section 2.1, a foundational overview of the chapter's purpose was set, identifying the academic and practical need for effective predictive models in football. This was followed by Section 2.2, which elaborated on the core concepts of data mining and ML, offering readers a clear understanding of how these computational paradigms can be leveraged to discover patterns, build models, and derive insight from structured sports data.

In Section 2.3, the field of sports analytics—with a focus on football—was contextualized, highlighting its growing adoption by professional clubs, analysts, and researchers. The evolution of football analytics from rudimentary statistics to AI-powered prediction systems was charted, underscoring how technology is reshaping performance analysis, scouting, tactical planning, and fan engagement.

The heart of the chapter lay in Section 2.4, where 20 peer-reviewed studies were critically reviewed to explore the application of machine learning models in football contexts. This analysis not only identified promising algorithms and data sources but also critically evaluated each study's methodology, limitations, and findings. Table 2.1, a structured summary table, consolidated this information, offering a comparative view of objectives, data used, algorithms applied, and results obtained. This helped inform the selection of three models—Random Forest, Gradient Boosting, and Logistic Regression—which were expanded upon in Sections 2.4.1 to 2.4.3, each justified for its alignment with the Research Objective 1 of this study. These models were chosen for their collective balance between predictive performance, explainability, and compatibility with the Kaggle Premier League dataset (1992–2019), which will serve as the primary data source.

Following this, Section 2.5 explored the landscape of feature selection techniques, a crucial stage for ensuring model efficiency and interpretability. By categorizing methods into filter, wrapper, and embedded techniques, the section provided a methodological blueprint that will be adopted in the upcoming methodology and implementation phase. Studies reviewed in this section affirmed that intelligent feature selection not only enhances accuracy but also reveals the hidden mechanics behind football success—which is precisely one of the goals of this research.

Despite advancements, certain research gaps persist, which this project aims to address. As highlighted in the Problem Statement (Section 1.3), much of the existing work in football analytics

 doi:10.20944/preprints202605.1076.v1

tends to focus on event-level data and match outcomes, overlooking the deeper, season-level patterns of team performance. There is a need for robust models that can classify season-long team success based on aggregated player-derived indicators. Additionally, while various models are applied, the crucial aspect of effective hyperparameter tuning and model optimization is often underexplored in the literature. This gap is critical, as model performance is significantly influenced not only by design and data but also by rigorous optimization. This study seeks to contribute to filling these gaps by developing and optimizing machine learning models specifically for predicting season-long team success in the Premier League.

Chapter III in a nutshell, Chapter II established a sound theoretical and empirical foundation for the job at hand. It has been indicated that machine learning, when used wisely with domain-specific information, has the potential to reveal what makes a football team successful. The chapter also justified the methodological approach, identified research gaps, and condensed essential methodologies and resources to be employed. Chapter III (Research Methodology) will go over the step-by-step process of applying these learnings to the selected dataset, including data preparation and feature engineering, as well as model evaluation and performance interpretation. This will turn the theoretical insights accumulated thus far into valuable analytics and machine learning pipelines designed for football success prediction.

Methodology

3.1. Introduction

In this chapter, we outline the research methodology used to address the research questions and objectives outlined in Chapter 1. As seen in Section 1.8, the research framework is based on CRISP-DM. This chapter will offer a thorough overview of the research plan, the procedures used for data collection, data preparation, feature selection, model development, and the primary hyperparameters and metrics for assessing the model performance. This chapter also discusses the experimental environment (software and tools used), model tuning strategies, and the metrics applied for predicting and evaluating performance.

3.2. Overview of Research Methods

In this section, the research methods for this project are presented. The methodology is grounded in the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, adapted to the football analytics context. The research follows a structured sequence of tasks:

1. Data Acquisition and Preprocessing
    a. First, we collect the data for all-time EPL player statistics from the Kaggle dataset. Then, we clean the dataset by handling missing values, normalizing field names, and ensuring season-team alignment. After that, we accumulate the individual player statistics to produce team-season-level features.
2. Feature Engineering and Selection
    a. Team-level indicators are constructed (e.g., total goals, average minutes played, disciplinary records). Further, we apply techniques such as correlation analysis and recursive feature elimination to identify optimal features. Dimensionality reduction could be explored if needed (e.g., PCA).
3. Model Building and Evaluation
    a. Here, we implement three machine learning models—Random Forest, Gradient Boosting, and Logistic Regression—selected from the literature review. Then, cross-validation is used to assess model performance in predicting team success (league position or points). Following that, models are evaluated based on accuracy, interpretability, and generalizability. Then, we tune the hyperparameters for the models.
4. Explainability and Interpretation

 doi:10.20944/preprints202605.1076.v1

a. SHAP (SHapley Additive exPlanations) are applied to derive interpretable insights from the models. Also, we examine which features consistently influence high performance. Finally, we compare model findings to known real-world trends, such as the dominance of certain teams.

5. Reporting and Synthesis

a. We summarize key findings and actionable patterns through the Results and Discussion chapter of the report. Then, we discuss the implications for analytics in professional football, and what can be done moving forward. To conclude, we will reflect on limitations and suggest directions for future research.

The overall workflow of this study is illustrated in Figure 3.1 below. The diagram offers a high-level visual representation of how the research phases flow into one another in an iterative manner.

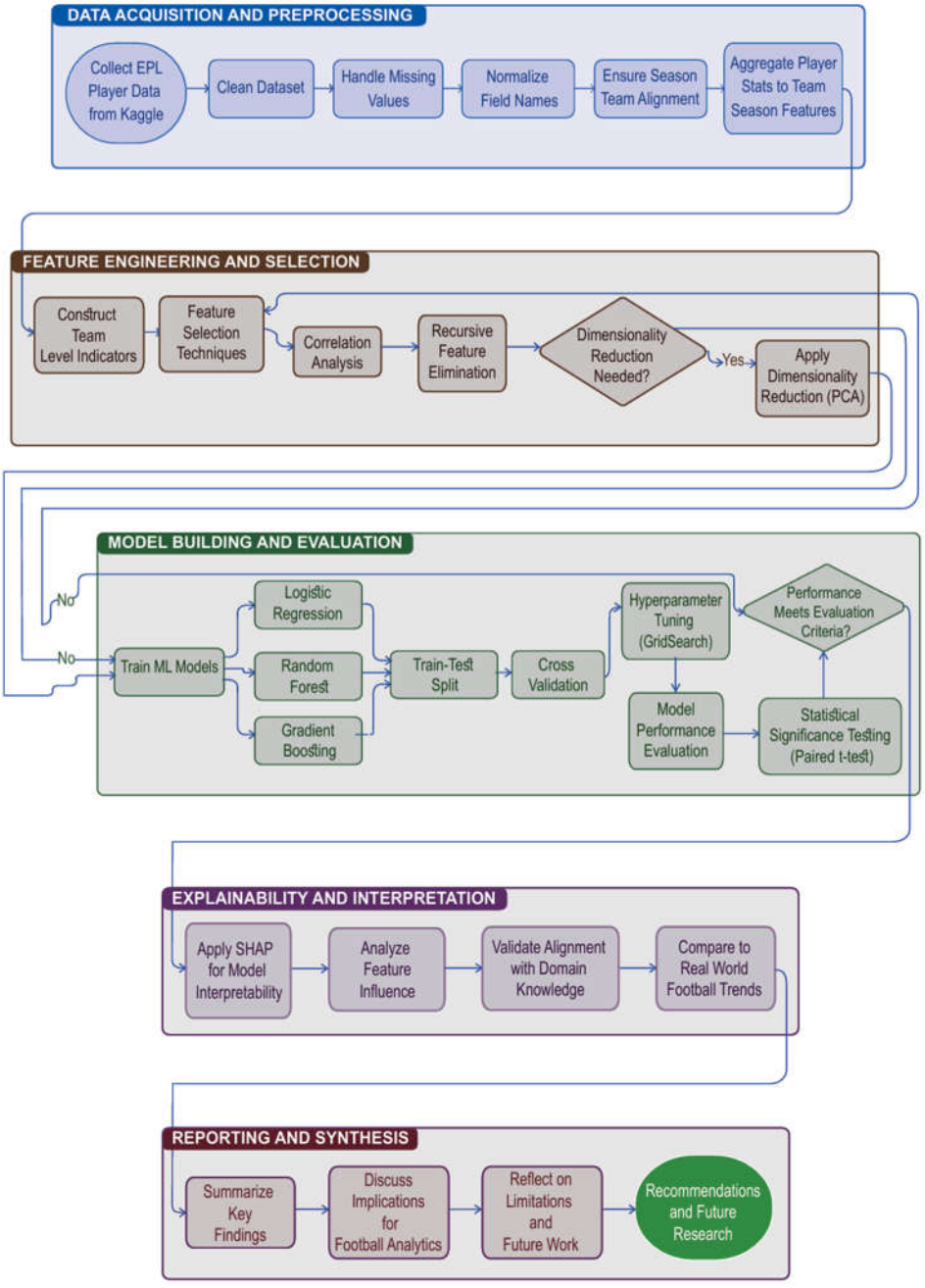


Figure 3.1. Overview of Research Methods for this Project.

 doi:10.20944/preprints202605.1076.v1

This modular yet iterative process allows for continuous refinement of models and data strategies based on intermediate findings and model diagnostics. By structuring the research in this manner, not only is the workflow streamlined, but it also ensures alignment with the research objectives established in Chapter I. The following sections detail each methodological component as implemented in this study.

3.3. Experiment Setup

The experimental environment for this research was designed to support flexible, scalable, and reproducible data processing and machine learning workflows. Therefore, all experiments were conducted using Python 3.10 within the Jupyter Notebook environment, benefitting from the powerful data science stack which includes (Jupyter 2025):

i. Pandas and NumPy for data manipulation,
ii. Matplotlib and Seaborn for visualization,
iii. Scikit-learn, XGBoost, and Statsmodels for machine learning and feature analysis.

The computational setup was based on a standard personal workstation (Intel i7 processor, 16 GB RAM) with optional access to Google Colaboratory for cloud-based experimentation. This setup ensured fast processing and reproducibility of all models and experiments. The software tools and packages were selected for their open-source availability, academic credibility, and widespread adoption in the machine learning and AI research community (Pedregosa et al. 2011; Chen & Guestrin 2016). Table 3.1 below summarizes the experiment setup.

Table 3.1. Experimental Environment.

Component	Configuration
Programming Lang	Python 3.10
IDE	Jupyter Notebook
Libraries	Pandas, NumPy, Scikit-learn, SHAP
Visualization	Matplotlib, Seaborn
Platform	Local machine / Google Colab

3.4. Data Understanding

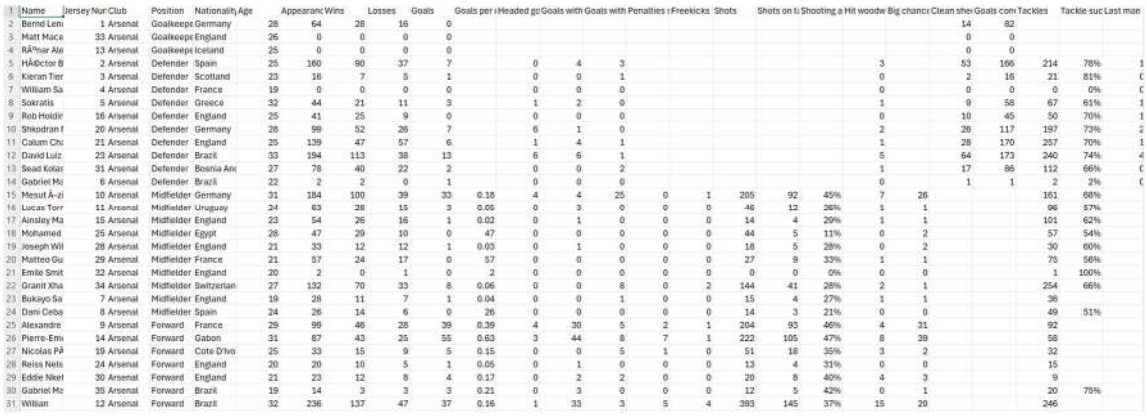
3.4.1. Dataset Description

The dataset used for this study was sourced from Kaggle, titled "All-time Premier League Players Statistics" compiled by Rishikesh Kanabar (2020). It includes career-wide performance statistics for 571 Premier League players, spanning from 1992 to 2019. It is licensed as CC BY-NC-SA 4.0 (Attribution-Non-commercial-ShareAlike 4.0 International). The EPL is England's top-tier professional football league, contested by 20 clubs annually, with matches typically taking place between August and May. The dataset is notable for being updated weekly, reflecting changes in player statistics after each match, which underscores its dynamic nature and potential for capturing evolving performance trends.

The Kaggle dataset is highly suitable and was hence selected because it offers a comprehensive array of historical match and team statistics that are pivotal for understanding team performance at a season-long level. It includes granular data points such as individual player statistics (e.g., goals, assists, passes, defensive actions), aggregated team statistics (e.g., total goals scored, conceded, possession percentages), and crucial match outcomes. The expansive range from 1992 to 2019 ensures a sufficiently large historical context, allowing for the identification of consistent patterns and the training of robust machine learning models capable of generalizing across different eras of the Premier League. This extended historical perspective is essential for accurately defining and predicting season-long success, a key focus of this research, by providing ample data instances for both successful and unsuccessful team campaigns.

 doi:10.20944/preprints202605.1076.v1

Figure 3.2 and 3.3 below shows a glimpse of the dataset on Excel and Jupyter Notebook respectively.



Name	Jersey Nur	Club	Position	Nationality	Age	Appearanc	Wins	Losses	Goals	Goals per r	Headed go	Goals with	Goals with	Penalties s	Freekicks	Shots	Shots on ta	Shooting a	Hit woodw	Big chanc	Clean shee	Goals com	Tackles	Tackle suc	Last man
Bernd Leno	1	Arsenal	Goalkeepe	Germany	28	64	28	16	0												14	82			
Matt Macey	33	Arsenal	Goalkeepe	England	26	0	0	0	0												0	0			
RÃºnar Ale	13	Arsenal	Goalkeepe	Iceland	25	0	0	0	0												0	0			
HÃ©ctor B	2	Arsenal	Defender	Spain	25	160	90	37	7		0	4	3						3		53	166	214	78%	1
Kieran Tier	3	Arsenal	Defender	Scotland	23	16	7	5	1		0	0	1						0		2	16	21	81%	0
William Sa	4	Arsenal	Defender	France	19	0	0	0	0		0	0	0						0		0	0	0	0%	0
Sokratis	5	Arsenal	Defender	Greece	32	44	21	11	3		1	2	0						1		9	58	67	61%	1
Rob Holdir	16	Arsenal	Defender	England	25	41	25	9	0		0	0	0						0		10	45	50	70%	1
Shkodran I	20	Arsenal	Defender	Germany	28	99	52	26	7		6	1	0						2		26	117	197	73%	2
Calum Ch	21	Arsenal	Defender	England	25	139	47	57	6		1	4	1						1		28	170	257	70%	1
David Luiz	23	Arsenal	Defender	Brazil	33	194	113	38	13		6	6	1						5		64	173	240	74%	4
Sead Kolas	31	Arsenal	Defender	Bosnia An	27	78	40	22	2		0	0	2						1		17	86	112	66%	0
Gabriel Ma	6	Arsenal	Defender	Brazil	22	2	2	0	1		0	0	0						0		1	1	2	2%	0
Mesut Ã¶zi	10	Arsenal	Midfielder	Germany	31	184	100	39	33	0.18	4	4	25	0	1	285	92	45%	7	26			161	68%	
Lucas Torr	11	Arsenal	Midfielder	Uruguay	24	63	28	15	3	0.05	0	3	0	0	0	46	12	26%	1	1			96	57%	
Ainsley Ma	15	Arsenal	Midfielder	England	23	54	26	16	1	0.02	0	1	0	0	0	14	4	29%	1	1			101	62%	
Mohamed	25	Arsenal	Midfielder	Egypt	28	47	29	10	0	47	0	0	0	0	0	44	5	11%	0	2			57	54%	
Joseph Wil	28	Arsenal	Midfielder	England	21	33	12	12	1	0.03	0	1	0	0	0	18	5	28%	0	2			30	60%	
Matteo Gu	29	Arsenal	Midfielder	France	21	57	24	17	0	57	0	0	0	0	0	27	9	33%	1	1			75	56%	
Emile Smit	32	Arsenal	Midfielder	England	20	2	0	1	0	2	0	0	0	0	0	0	0	0%	0	0			1	100%	
Granit Xha	34	Arsenal	Midfielder	Switzerlan	27	132	70	33	8	0.06	0	0	8	0	2	144	41	28%	2	1			254	66%	
Bukayo Sa	7	Arsenal	Midfielder	England	19	28	11	7	1	0.04	0	0	1	0	0	15	4	27%	1	1			36		
Dani Ceba	8	Arsenal	Midfielder	Spain	24	26	14	6	0	26	0	0	0	0	0	14	3	21%	0	0			49	51%	
Alexandre	9	Arsenal	Forward	France	29	99	46	28	39	0.39	4	30	5	2	1	204	93	46%	4	31			92		
Pierre-Em	14	Arsenal	Forward	Gabon	31	87	43	25	55	0.63	3	44	8	7	1	222	105	47%	8	39			58		
Nicolas PÃ	19	Arsenal	Forward	Cote D'Ivo	25	33	15	9	5	0.15	0	0	5	1	0	51	18	35%	3	2			32		
Reiss Nels	24	Arsenal	Forward	England	20	20	10	5	1	0.05	0	1	0	0	0	13	4	31%	0	0			15		
Eddie Nket	30	Arsenal	Forward	England	21	23	12	8	4	0.17	0	2	2	0	0	20	8	40%	4	3			9		
Gabriel Ma	35	Arsenal	Forward	Brazil	19	14	3	3	3	0.21	0	3	0	0	0	12	5	42%	0	1			20	75%	
William	12	Arsenal	Forward	Brazil	32	236	137	47	37	0.16	1	33	3	5	4	393	145	37%	15	29			246		

Figure 3.2. Display of EPL Kaggle Dataset on Excel.

```
--- Dataset Head (First 5 Rows) ---
                    Name  Jersey Number     Club    Position Nationality   Age  \
0            Bernd Leno            1.0  Arsenal  Goalkeeper     Germany  28.0
1            Matt Macey           33.0  Arsenal  Goalkeeper     England  26.0
2  Rúnar Alex Rúnarsson           13.0  Arsenal  Goalkeeper     Iceland  25.0
3       Héctor Bellerín            2.0  Arsenal    Defender       Spain  25.0
4        Kieran Tierney            3.0  Arsenal    Defender    Scotland  23.0

   Appearances  Wins  Losses  Goals  ...  Punches  High Claims  Catches  \
0           64    28      16      0  ...     34.0         26.0     17.0
1            0     0       0      0  ...      0.0          0.0      0.0
2            0     0       0      0  ...      0.0          0.0      0.0
3          160    90      37      7  ...      NaN          NaN      NaN
4           16     7       5      1  ...      NaN          NaN      NaN

   Sweeper clearances  Throw outs  Goal Kicks  Yellow cards  Red cards  Fouls  \
0                28.0       375.0       489.0             2          0      0
1                 0.0         0.0         0.0             0          0      0
2                 0.0         0.0         0.0             0          0      0
3                 NaN         NaN         NaN            23          0    125
4                 NaN         NaN         NaN             2          0      9

   Offsides
...
Red cards          0
Fouls              0
Offsides          69
dtype: int64
```

Figure 3.3. Display of EPL Kaggle Dataset on Jupyter Notebook.

The dataset comprises of 571 rows (players) and 59 features, encompassing multiple aspects of a player's profile and match performance. Each row in the dataset represents a single player's cumulative statistics, and the features are both numerical (e.g., Goals, Assists) and categorical (e.g., Position, Club). However, there are notable missing values and non-uniform feature coverage. For instance, goalkeeper-specific features like Saves, Punches, and Sweeper clearances are mostly NaN for outfield players, whereas Goals and Crosses are often NaN for goalkeepers. Table 3.2 below lists the key columns in the dataset, their data types and their potential use in our research.

 doi:10.20944/preprints202605.1076.v1

Table 3.2. Description of Key Parameters in All Time EPL Player Statistics Dataset.

Column Name	Description	Data Type	Potential Usage for Team Success Prediction
Name	Player's full name	Categorical (Text)	Identifier; useful for grouping by player.
Jersey Number	Player's jersey number	Numerical (Integer)	Player identifier; typically, not directly used for team success prediction.
Club	Team the player belongs to	Categorical (Text)	**Crucial for aggregation to team level.**
Position	Player's primary playing position	Categorical (Text)	Useful for positional analysis, e.g., number of defenders/midfielders in a team.
Nationality	Player's nationality	Categorical (Text)	May be used for diversity metrics or to identify regional talent pools.
Age	Player's age	Numerical (Integer)	Can be aggregated to average team age, indicating experience/youth.
Appearances	Total games played by player	Numerical (Integer)	Sum for total team appearances; crucial for calculating per-game averages.
Wins	Games won by player's team while on field	Numerical (Integer)	Sum for total team wins; contributes to team win rate.
Losses	Games lost by player's team while on field	Numerical (Integer)	Sum for total team losses; contributes to team loss rate.
Goals	Total goals scored by player	Numerical (Integer)	**Crucial for aggregating to total team goals.**
Assists	Total assists made by player	Numerical (Integer)	**Crucial for aggregating to total team assists.**
Shots	Total shots attempted by player	Numerical (Integer)	**Crucial for aggregating to total team shots.**
Shots on Goal	Total shots on target by player	Numerical (Integer)	**Crucial for aggregating to total team shots on target.**
Goals Per Match	Goals per match (Known Bug)	Numerical (Float)	DO NOT USE DIRECTLY**.** Must be recalculated from *Goals* and *Appearances* or *Minutes Played* to avoid erroneous values.
Other Stats	(e.g., Pass Accuracy, Tackles, Interceptions, etc.)	Numerical	Crucial for deriving comprehensive team-level KPIs.

It is imperative to acknowledge a known issue within this dataset: the 'Goals Per Match' column has been identified by the dataset creator as containing abnormally high or incorrect values due to web scraping anomalies during the initial milliseconds of page loading. This data quality issue necessitates that instead of directly using the raw 'Goals Per Match' column, this metric should be recalculated from the more reliable 'Goals' and 'Appearances' or 'Minutes Played' columns during the data pre-processing phase to ensure accuracy and prevent erroneous model inputs. This specific data anomaly highlights the critical importance of thorough data understanding and quality assurance before proceeding with any analytical modelling.

The richness of the dataset aligns well with the Research Objective 2, of identifying what features contribute most significantly to team success. Aggregate metrics such as Wins, Goals, Passes, Tackles, and Big chances created offer quantifiable insights into player performance that may collectively correlate with the success of the teams they represent.

However, it's also worth noting that the dataset lacks seasonal context and match-level granularity. As a result, this project's insights will focus on long-term career trends, rather than short-term predictive match outcomes.

 doi:10.20944/preprints202605.1076.v1

3.5. Data Preprocessing

Data preprocessing is a foundational step in the machine learning pipeline, ensuring that the raw data is transformed into a clean, consistent, and analysis-ready format. This phase is particularly critical for the Rishikesh Kanabar dataset due to its web-scraped origin and identified data quality issues. The steps involved will include the following as below.

i. Handling Missing Values: The dataset contained varying degrees of missingness, particularly in features exclusive to certain player roles (e.g., goalkeeper-specific attributes like Saves, Punches, and Sweeper Clearances). These features exhibited up to 50–60% missing values, primarily because they were non-applicable to outfield players. Rather than discarding such columns entirely, the strategy we can adopt is role-specific imputation and conditional filtering. Goalkeeper-specific metrics can be retained and analyzed separately using position-based stratification. Features with negligible variance or excessive sparsity (e.g., >70% missing) can be excluded. For numeric fields with moderate missing values, median imputation was applied. Categorical fields with missing values were imputed using mode or tagged as "Unknown".

ii. Outlier Detection and Treatment: Outliers, which are data points significantly different from other observations, can disproportionately influence model training and lead to biased results. Exploratory Data Analysis (EDA) could possibly reveal that certain features exhibit skewed distributions. In that case, log transformation can be applied to reduce the skewness and improve model stability where appropriate. Statistical methods (e.g., Z-score, IQR method) and visualization techniques (e.g., box plots) will be used to identify outliers. We perform this with careful consideration to avoid losing valuable information.

iii. Data Type Conversion: Ensuring that all features are in the correct data type (e.g., numerical for quantitative metrics, categorical for positions or club names) is fundamental for proper analysis and model input. Categorical features such as Club, Nationality, and Position were encoded using One-Hot Encoding to make them numerically interpretable by ML models. Numeric columns were cast to appropriate types (e.g., float64, int64) to maintain precision during calculations.

iv. Normalization: For many ML algorithms, features with larger numerical ranges can dominate those with smaller ranges. To prevent this, numerical features were scaled to ensure they are on a comparable scale, which can significantly improve model performance and convergence. In our dataset, given the wide range in feature scales (e.g., Tackles vs. Big chances missed), Min-Max scaling was used to normalize features between 0 and 1. This helps algorithms like Logistic Regression and XGBoost converge more efficiently.

Some key transformations pertaining to our dataset include calculating:

a. Averages and Ratios: For instance, instead of just total goals, features like Goals per Match (total goals scored divided by matches played) and Goal Difference per Match were computed to normalize performance across seasons. Similarly, Average Passes per Match and Pass Completion Rate were derived to quantify team playing style and efficiency.

b. Aggregated Sums: Individual player statistics, such as player goals and assists, were aggregated to provide season-long team performance metrics. For example, the sum of goals scored by all players in a team over a season was used as a team-level attacking indicator. This aggregation from player-level to team-level is crucial for preparing the data as input for models aiming to predict *team* success.

c. Win/Loss/Draw Ratios: Creating features like Win Rate (total wins / total matches) provided a normalized measure of team performance.

These derived features offer a more nuanced understanding of team performance than raw counts, directly addressing the need for robust indicators for season-long analysis. For instance, normalizing features by the number of matches makes them comparable across seasons regardless of fixture variations. Furthermore, this meticulous feature engineering process directly improves model

 doi:10.20944/preprints202605.1076.v1

interpretability by transforming raw data into meaningful metrics that better represent underlying football phenomena, making it easier to understand why a model makes a particular prediction.

3.6. Feature Engineering

3.6.1. Deriving Team-Level Features from Player Statistics

A central hurdle in this study, methodologically, is to transform our player-centric Premier League dataset into meaningful team-level features that can be used to predict team success. Team success is an emergent property of collective performance, and thus, individual player statistics must be aggregated to reflect the team's overall capabilities and strategic tendencies. This process is essential for bridging the gap between the available raw data and the Research Objective 1, of understanding team success.

Various aggregation methods, as shown below, were employed to derive these team-level features:

i. Summation: This involves summing the individual player statistics for each team to obtain a total team value. For example, summing the 'Goals' scored by all players in a team will yield the 'Total Team Goals' for that season or period. Similarly, 'Total Team Assists', 'Total Team Shots', and 'Total Team Shots on Target' can be derived.

ii. Averaging (Mean): Calculating the arithmetic mean of player statistics across a team provides an average measure of performance. Examples include 'Average Player Age' for a team, 'Average Goals per Player', 'Average Assists per Player', or 'Average Appearances per Player'. This helps in understanding the typical contribution or characteristic of players within a team.

iii. Median: The median value can be used to represent the central tendency of a player statistic within a team, particularly useful for skewed distributions where the mean might be influenced by outliers. For instance, 'Median Player Age' or 'Median Goals per Player'.

iv. Range: Calculating the difference between the maximum and minimum values of a player statistic within a team can indicate the variability or spread of talent. For example, 'Age Range' within the squad or 'Range of Goals Scored' to assess reliance on a few key scorers.

v. Variance and Standard Deviation: These statistical measures quantify the dispersion or spread of player data around the mean. A lower standard deviation in, for example, 'Minutes Played' might indicate a more consistent starting XI, while a higher standard deviation could suggest significant squad rotation or injury issues.

Beyond these basic aggregations, more complex engineered features were explored and created by combining or transforming existing metrics. For instance, 'Team Shooting Accuracy' can be calculated as (Sum of Team Shots on Target / Sum of Total Team Shots). 'Team Goal Conversion Rate' can be derived from (Total Team Goals / Total Team Shots). If player position data is consistent, features like 'Goals from Forwards', 'Assists from Midfielders', or 'Tackles from Defenders' could be created by summing relevant statistics for players in specific positions within each team. The ratio of offensive to defensive actions, or the proportion of goals scored by different player positions, were also explored to capture tactical profiles.

3.6.2. Identification of Key Features for Team Success Prediction

Following the derivation of comprehensive team-level features, the next vital step is to identify which of these features are most influential in defining and predicting football team success. This process directly links back to the possible football metrics established in the literature review, through the field of work.

For instance, from the aggregated player data, features directly corresponding to offensive metrics would include 'Total Team Goals', 'Team Shooting Accuracy', 'Total Team Assists', and 'Total Team Progressive Passes'. For defensive metrics, 'Total Team Shots Conceded', 'Total Team Tackles

 doi:10.20944/preprints202605.1076.v1

Won', and 'Total Team Ball Recoveries' are critical. Tactical metrics are represented by 'Average Team Possession Percentage' and 'Team Pass Completion Rate'. The disciplinary aspect is captured by 'Total Team Yellow Cards' or 'Total Team Red Cards'.

The target variable for predicting team success is clearly defined based on the available data and the research questions we defined in Section 1.4. Given the Premier League context, this is:

a. League Position: Predicting the final league standing of a team.
b. Total Points: Predicting the total points accumulated by a team over a season.
c. Success Category: Classifying teams into predefined success categories (e.g., 'Top 4 Finish', 'Mid-Table', 'Relegated') based on their league performance.

The identification of key features was an iterative process, informed by both domain knowledge (from the literature review) and the insights we obtained from feature selection techniques.

3.7. Feature Selection

Feature selection is a critical process that aims to identify and select the most relevant subset of features from the engineered dataset and eliminate redundant data. This is essential for several reasons: it reduces dimensionality, mitigates overfitting by simplifying the model, improves model interpretability, and enhances computational efficiency, especially with a potentially large number of engineered features. The effectiveness of a predictive model is often more dependent on the quality of its features than on the specific algorithm employed.

This study employs a combination of feature selection techniques to ensure robustness and identify truly influential features in alignment with the strategies described in Chapter II (Section 2.5):

a. Filter Methods for Initial Screening: Techniques such as Correlation-Based Filtering can be used to assess the linear relationship between each engineered team-level feature and the chosen target variable for team success (e.g., league position or total points). Features with high correlation to the target and low correlation among themselves can be prioritized. If categorical features are involved, techniques like Mutual Information or Chi-Square Test can be considered to measure dependency. This initial screening helps to quickly eliminate irrelevant or highly redundant features. SelectKBest using ANOVA F-test can also be used.
b. Wrapper Methods: The second stage involves Recursive Feature Elimination (RFE) with a Logistic Regression or Random Forest classifier to rank and select top-performing feature subsets. This method iteratively removes the least important features based on model weightings and re-trains the model.
c. Embedded Methods: Given the anticipated use of tree-based models (i.e., Random Forest, Gradient Boosting) for prediction, embedded methods are highly valuable. These models inherently perform feature selection by assigning importance scores to features based on how much they contribute to reducing impurity or improving model accuracy during the training process. The feature importance scores derived from these models provided insights into the most predictive features within the context of the chosen algorithms. For instance, the feature_importances_ attribute in scikit-learn's Random Forest or Gradient Boosting models were directly used.

Additionally, player features are grouped into performance tiers using a Player Performance Index (PPI), which aggregates normalized statistics for attacking, passing, defensive, goalkeeping, and disciplinary actions using weights. Players are classified into four tiers: Elite, Above Average, Average, Below Average based on quartiles of PPI scores. The weights were strategically assigned to reflect the relative impact of different actions on match outcomes:

a. Attacking & Creativity (45%): Combined weight for goals, assists, big chances created, and shots on target.

 doi:10.20944/preprints202605.1076.v1

b. Technical & Defensive Stability (40%): Combined weight for passing volume, tackle success, and interceptions.

c. Goalkeeping & Discipline (15%): Accounting for clean sheets and saves, while penalizing errors and cards.

Selected features are cross validated against the model performance to ensure interpretability. SHAP values are later used to validate the contribution of selected features at both individual player and team level.

3.7.1. Data Partitioning/Cross Validation

After preprocessing and feature selection, the dataset was partitioned to facilitate reliable model training and evaluation. The goal was to ensure generalization and minimize the risk of overfitting, especially given the moderate sample size of 571 players. The dataset was first split into training and test sets using a standard 80/20 split. The training set was used for fitting the models and performing hyperparameter optimization, while the test set was exclusively used for final model evaluation.

To further improve robustness, k-fold cross-validation was used within the training set. In this research, 5-fold cross-validation can be adopted, as it offers a good trade-off between bias and variance in performance estimates (Kohavi 1995). This means the training data is split into five parts: in each round, four parts are used for training and one for validation, rotating until each fold serves once as validation.

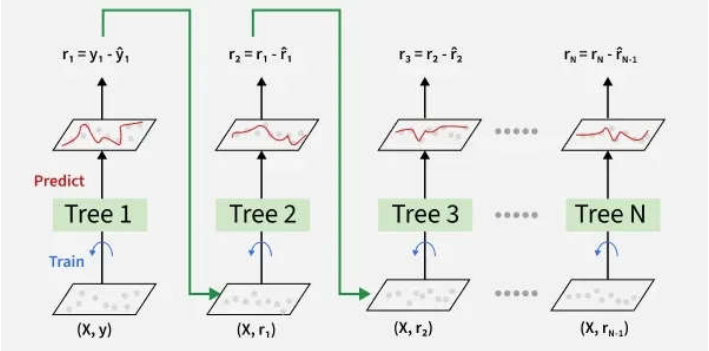
Throughout the feature selection process, particularly with wrapper and embedded methods, k-fold cross-validation was rigorously applied. This technique involves partitioning the dataset into multiple folds, training the model on a subset of these folds, and validating on the remaining fold. This iterative process ensures that the selected features and the resulting model generalize well to unseen data and helps prevent overfitting, providing a robust assessment of feature effectiveness.

This systematic process led to a refined set of features that are most predictive of football team success.

3.8. Machine Learning Model Selection

The selection of appropriate machine learning models is crucial for effectively predicting football team success based on the engineered team-level features. The choice of model was guided by the nature of the prediction problem (regression or classification) and the characteristics of the prepared dataset. But we considered both types or models.

Based on the analysis from Chapter II (Section 2.4) and supported by the literature review and feature analysis in Section 3.6, three supervised machine learning models were selected for development and comparison: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier. Gradient Boosting replaces XGBoost to simplify implementation while retaining boosting capabilities. The small size of the dataset chosen is also a reason. Unlike Random Forest, which builds trees in parallel, GBM builds trees sequentially, with each new tree correcting the errors of the previous one (Baboota & Kaur 2019), as shown in Figure 3.4.



 doi:10.20944/preprints202605.1076.v1

Figure 3.4. Gradient Boosting Trees. Source: GeeksforGeeks 2025.

The above models are highly favoured in sports prediction due to their robustness, ability to handle complex non-linear relationships, and strong predictive accuracy, and their relevance in football analytics as shown in recent studies (e.g., García et al. 2020; Rathke 2017; Bunker & Thabtah 2019).

To maximize the predictive power of the supervised models, two ensemble techniques were implemented:

1. Soft Voting Classifier: It averages the predicted probabilities of Logistic Regression, Random Forest, and Gradient Boosting. This reduces the risk of the model being skewed by an outlier prediction from a single algorithm, providing a more stable success probability, taking the majority rule among models.
2. Stacking Classifier: This involved training a meta-learner (Logistic Regression) on the predictions of the three base models. The stacking approach allows the system to learn which base model is most reliable for specific types of team data (e.g., perhaps Random Forest is better at identifying high-possession teams, while GBM is better at identifying defensive-counter teams).

The target variable for this study was defined as a binary outcome: 'Successful Team' and 'Unsuccessful Team'. Specifically, a 'Successful Team' is defined as one finishing in the top-half of the league table (e.g., rank 1-10), assigned a value of 1. Conversely, an 'Unsuccessful Team' is defined as one finishing in the bottom-half of the league table (e.g., rank 11-20), assigned a value of 0. This binary classification allowed for a clear distinction in team performance over a full season.

The final selection of models prioritized those that balance predictive power with interpretability, allowing for a deeper understanding of the features contributing to team success. Ensemble methods formed the core of the modelling approach due to their proven efficacy in similar sports analytics tasks.

Additionally, there was a dilemma when choosing what approach to go ahead with – supervised or unsupervised ML Supervised learning was chosen as the primary research engine because the outcome of interest—team success—is explicitly labelled in historical league tables. By training on "ground truth," the supervised models can accurately map performance indicators to specific league outcomes. Unsupervised learning (K-Means) was utilized solely as a support mechanism to exploratory verify the natural structure of performance tiers and to facilitate a similarity-based scouting recommendation system attempt. It wasn't used for the final predictions.

3.9. Hyperparameter Tuning

Hyperparameter tuning played a pivotal role in optimizing the performance of ML models. Unlike model parameters that are learned during training, hyperparameters were set prior to training and significantly influence model complexity, learning behaviour, and ultimately generalization capability (Probst et al. 2019). It is also crucial as it directly addresses Objective 3 of our research.

Common strategies employed for hyperparameter tuning in this study included:

1. Grid Search: This method exhaustively searches through a predefined subset of the hyperparameter space. For each combination of hyperparameters, the model is trained and evaluated (typically using cross-validation). Grid Search guarantees finding the best combination within the specified grid, though it can be computationally expensive for many hyperparameters or large ranges.
2. Random Search: Unlike Grid Search, Random Search samples random combinations of hyperparameters from a defined distribution. This method is often more efficient than Grid Search, especially in high-dimensional hyperparameter spaces, as it can explore a wider range of values in less time and often finds good combinations more quickly.

 doi:10.20944/preprints202605.1076.v1

3. Bayesian Optimization: A more advanced technique, Bayesian Optimization builds a probabilistic model of the objective function (e.g., F1-score) and uses it to select the most promising hyperparameters to evaluate in the real model. It intelligently explores the hyperparameter space, leading to more efficient optimization compared to Grid Search or Random Search.

In this research, Grid Search Cross-Validation was adopted as the primary tuning technique. Grid Search exhaustively evaluates combinations of predefined hyperparameter values across k-fold cross-validation (k=5 in our case), allowing the identification of the most suitable hyperparameter configuration.

The following hyperparameters are tuned for each model:

a. Logistic Regression:
   i. Regularization type (penalty: L1, L2)
   ii. Regularization strength (C: inverse of regularization strength)

b. Random Forest:
   i. Number of trees (n_estimators)
   ii. Maximum tree depth (max_depth)
   iii. Minimum samples to split a node (min_samples_split)
   iv. Minimum samples per leaf (min_samples_leaf)

c. Gradient Boosting:
   i. Number of trees (n_estimators) (50, 100, 200)
   ii. Maximum tree depth (max_depth) (3, 5, 7)
   iii. Minimum samples to split a node (min_samples_split) (2, 5, 10)
   iv. Minimum samples per leaf (min_samples_leaf)
   v. Shrinkage factor for tree contributions (learning_rate) (0.01, 0.05, 0.1)

d. XGBoost:
   i. Learning rate (η - eta)
   ii. Maximum tree depth (max_depth)
   iii. Number of boosting rounds (n_estimators)
   iv. Subsample ratio (subsample)
   v. Column sample ratio (colsample_bytree)
   vi. Regularization parameters (λ - lambda, α - alpha)

This tuning can be implemented using GridSearchCV from Scikit-learn and XGBClassifier from the XGBoost library.

Throughout this process, k-fold cross-validation was a fundamental component. The average performance across all folds provided a more robust and reliable estimate of the model's generalization capability, preventing overfitting to a specific training set and ensuring the chosen hyperparameters are truly optimal for unseen data. Meticulous splitting of data into training, validation, and test sets were also maintained to prevent data leakage, ensuring that the final model evaluation is unbiased.

3.10. Evaluation

Once the models are trained and tuned, they are evaluated using both confusion matrix-based metrics and probabilistic evaluation (ROC-AUC). The goal is to compare model predictions against true class labels ("successful" vs "not successful") and determine classification quality. Evaluating a model using multiple metrics is essential because each metric zones in on distinct aspects of performance. These metrics provide a quantitative measure to assess their performance on how well

 doi:10.20944/preprints202605.1076.v1

they classify the dataset, validate their predictive power, and ensure their reliability in discerning factors contributing to football team success.

3.10.1. Performance Metrics

Before explaining the chosen metrics, the following are the basics of evaluating classification models:

1. True Positive (TP): This is the number of positive predicted positives.
2. True Negative (FP): This is the number of negative predicted negatives.
3. False Positive (FP): The number of negative predicted positives.
4. False Negative (FN): The number of positive predicted negatives.

The key evaluation metrics used for classification tasks along with their formulas are as follows (Dake et al. 2023):

a. Accuracy: The proportion of correct predictions made by the model. While a general indicator, accuracy alone can be misleading if there is a class imbalance (e.g., significantly more unsuccessful teams than successful teams, or vice versa).

$$\frac{TP + TN}{Tp + TN + FP + FN} \tag{3.1}$$

b. Precision: The proportion of true positives among the total predicted positives. In this context, it indicates how many of the teams predicted as 'successful' actually were successful. It is important for minimizing false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.2}$$

c. Recall: The proportion of true positives among the total actual positives. It indicates how many of the actual 'successful' teams were correctly identified by the model. It is important for minimizing false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.3}$$

d. F1 Score: Harmonic mean of precision and recall — useful when there is class imbalance. A high F1-score indicates good performance in both precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

e. Confusion Matrix: It's a table used to evaluate the performance of a classification model, showing the number of true positives, true negatives, false positives, and false negative. This visualization is needed for understanding specific types of errors made by the model.

f. ROC-AUC Curve: Receiver Operating Characteristic (ROC) Curve plots True Positive Rate (TPR) vs. False Positive Rate (FPR). The AUC (Area Under Curve) quantifies this curve — higher values (closer to 1.0) indicate better separation, signifying the model's capability to rank positive instances higher than negative instances. They were utilized to assess the models' ability to distinguish between successful and unsuccessful teams across various classification thresholds.

Additionally, if we foray into the regression aspect of the analysis (e.g., predicting the exact total league points or final league position), then the following metrics will be used:

 doi:10.20944/preprints202605.1076.v1

a. Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values. It provides a clear measure of prediction error in the original units.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|yi - \widehat{yi}\right| \quad (3.5)$$

b. Mean Squared Error (MSE) / Root Mean Squared Error (RMSE): MSE calculates the average of the squared differences between predictions and actual values, penalizing larger errors more heavily. RMSE is the square root of MSE, providing error in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2} \quad (3.6)$$

Further, SHAP (SHapley Additive exPlanations) values were also considered for model interpretation. While the confusion matrix and ROC-AUC provide aggregate performance, SHAP values provide local interpretability by showing the contribution of each feature to a specific prediction for an individual team. The implementation explored SHAP at two levels of granularity: Global (where we average [SHAP] values to identify the most influential features.) and Local (where we feature contributions to individual player or team predictions). This aligns with the overall objective of understanding 'what makes a successful football team' by detailing how various performance indicators contribute to the classification of a team's success or failure.

3.11. Ethical Considerations

Our research endeavour, particularly since data analysis and predictive modelling is involved, acknowledges its ethical implications. They are elaborated below.

While the Rishikesh Kanabar Premier League dataset is publicly available on Kaggle and contains aggregated player statistics rather than sensitive personal data, the principle of data privacy remains paramount. The analysis focused on statistical patterns and team-level insights, ensuring no individual player's performance is unfairly scrutinized or used out of context in a manner that could impact their professional standing.

ML models are susceptible to biases present in the training data. If the historical data reflects existing biases (e.g., undervaluing defensive contributions or certain player nationalities), the models may perpetuate these biases in their predictions or feature importance rankings. Efforts were made to interpret model outputs critically and acknowledge any potential biases.

The findings of this study are intended for academic contribution and to offer data-driven insights for strategic planning and performance analysis. It is crucial to emphasize that these predictive models should serve as tools to 'inform' human decision-making, not replace it. Over-reliance on model predictions without human oversight or domain expertise could lead to unintended consequences, particularly in areas like player recruitment or tactical adjustments.

3.12. Conclusion

This chapter has laid the foundation of our methodology for answering the research questions and achieving the objectives outlined in Chapter I. Beginning with an overview of the research design in Section 3.2, the chapter expanded on the CRISP-DM framework as a structured and iterative guideline to carry out the study. The subsequent sections detailed each phase of the ML pipeline that will be implemented for the analysis of the Premier League dataset.

In Section 3.3, the experimental setup was defined, followed by a comprehensive understanding of the dataset in Section 3.4, where the data source, structure, and feature definitions were introduced.

 doi:10.20944/preprints202605.1076.v1

Then, Section 3.5 addressed data preprocessing, applying techniques such as label encoding, handling missing values, and deriving new performance indicators. This was followed by Section 3.6, where feature engineering was explored, where we derive team-level features using various aggregation methods. We also touched upon what we chose as the target variable. This was followed by Section 3.7, where statistical and wrapper-based feature selection methods were discussed to reduce dimensionality and isolate the most influential predictors of team success.

Section 3.7.1 presented the data partitioning strategy, including an 80/20 train-test split and the use of 5-fold cross-validation for robust training and evaluation. In Section 3.8, three ML models—Logistic Regression, Random Forest, and Gradient Boosting—were selected based on literature justification and their relevance to structured sports analytics problems. These models, along with ensemble learning models, served as the core predictive mechanisms in the study. The chosen models were further prepared for optimization in Section 3.9, where grid search and cross-validation were employed to fine-tune key hyperparameters.

Finally, Section 3.10 outlined the performance metrics that were used to assess and compare the models developed. Section 3.11 acknowledged the ethical considerations we took into account while partaking in this research, particularly concerning data privacy and prediction biases.

In synthesizing these chapters, it became clear that identifying what makes a successful football team is not a trivial task. It requires a systematic, data-driven approach that moves beyond superficial statistics to capture the complex interplay of individual contributions forming collective performance. By meticulously defining success, transforming raw player data into actionable team-level features, and applying optimized machine learning models, this dissertation aims to provide valuable, data-backed insights for football clubs, analysts, and enthusiasts alike, contributing to the growing body of knowledge in sports analytics. The next phase of this research will focus on analysing the results using the metrics discussed. The outcomes will then be interpreted considering the research objectives and used to derive meaningful insights about what makes a football team successful based on historical performance data.

Chapter IV

Results and Discussion

4.1. Introduction

This chapter presents and discusses the findings of the DM and ML process applied to our performance dataset in this study. The primary focus is to evaluate the predictive accuracy of the implemented models and interpret the underlying factors that contribute to team success in the EPL.

By analysing the predictive performance metrics of the baseline and ensemble ML models, this study identifies the most robust framework for football analytics. This is followed by an analysis of feature selection methods and their impact on model performance. Furthermore, the effects of hyperparameter optimisation will be detailed in this chapter. Finally, the use of SHAP (SHapley Additive exPlanations) provides a transparent view of the model's decision-making process, directly addressing Objective 3. The results are discussed in relation to real-world football dynamics, ensuring practical relevance for clubs, coaches, and analysts.

4.2. Data Engineering and Preliminary Analysis

We pre-processed the features of our dataset and aggregated the player data to team-level features and defined our target variable as 'Team_Success'. Figure 4.1 below displays the new features and the success threshold.

 doi:10.20944/preprints202605.1076.v1


```
--- New Player-level Derived Features Created ---
                  Name     Club  Appearances  Goals  Goals_Per_Appearance  \
0           Bernd Leno  Arsenal           64      0               0.00000
1           Matt Macey  Arsenal            0      0               0.00000
2  Rúnar Alex Rúnarsson  Arsenal            0      0               0.00000
3      Héctor Bellerín  Arsenal          160      7               0.04375
4       Kieran Tierney  Arsenal           16      1               0.06250

   Assists  Assists_Per_Appearance
0        0                  0.0000
1        0                  0.0000
2        0                  0.0000
3       18                  0.1125
4        1                  0.0625

--- Aggregated Team-level Data Head (Top 5 Teams) ---
                     Club  Total_Appearances  Total_Wins  Total_Losses  \
0                 Arsenal               1975        1005           502
1             Aston-Villa               1016         283           507
2  Brighton-and-Hove-Albion               1216         348           503
3                 Burnley               2431         765          1068
4                 Chelsea               1965        1023           535

   Total_Goals  Total_Assists    Avg_Age  Avg_Goals_Per_Appearance  \
0          231            202  25.100000                  0.100856
1           49             56  26.250000                  0.051310
2           90             84  24.393939                  0.044370
3          180            142  27.125000                  0.097757
4          219            159  25.444444                  0.075776

   Avg_Assists_Per_Appearance  Total_Yellow_Cards  ...  Avg_Shooting_Accuracy  \
0                    0.066716                 302  ...               0.175333
1                    0.044039                 110  ...               0.131429
2                    0.034278                 121  ...               0.120606
3                    0.047555                 367  ...               0.135833
4                    0.055841                 191  ...               0.125185

   Avg_Tackle_Success  Avg_Cross_Accuracy  Total_Clean_Sheets  \
0            0.408000            0.126333               326.0
1            0.353929            0.157857               221.0
2            0.255758            0.075152               186.0
3            0.325000            0.094167               346.0
4            0.401111            0.104074               375.0

   Total_Goals_Conceded  Total_Saves  Total_Fouls  Total_Offsides  Win_Rate  \
0                1594.0        762.0         1611           311.0  0.666888
1                1496.0        929.0          736            87.0  0.358228
2                1255.0        920.0          940           185.0  0.408931
3                1951.0        657.0         2212           392.0  0.417349
4                1599.0        714.0         1535           255.0  0.656611

   Team_Success
0             1
1             0
2             0
3             0
4             1

[5 rows x 24 columns]

Success Threshold (75th percentile of Win Rate): 0.6510
Distribution of Team_Success:
Team_Success
0    15
1     5
```

Figure 4.1. Data Preprocessing.

Now, some exploratory data analysis was conducted to understand the nature of the dataset better, as shown in the Figures 4.2 to 4.5 below.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2026 doi:10.20944/preprints202605.1076.v1

Figure 4.2. Distribution of Team Success.

Figure 4.3 above shows the correlation matrix for team-level performance variables, giving a clear snapshot of how these metrics relate to each other in the dataset. We see strong positive correlations (towards red colour) among outcome-focused measures like Total_Wins, Total_Goals, Total_Assists, and Win_Rate ($r > 0.7$), which makes sense given how attacking output ties directly to overall season success. Defensive statistics, such as Total_Clean_Sheets and Total_Saves, also show moderate-to-strong links, highlighting the real impact of solid defence on team results.

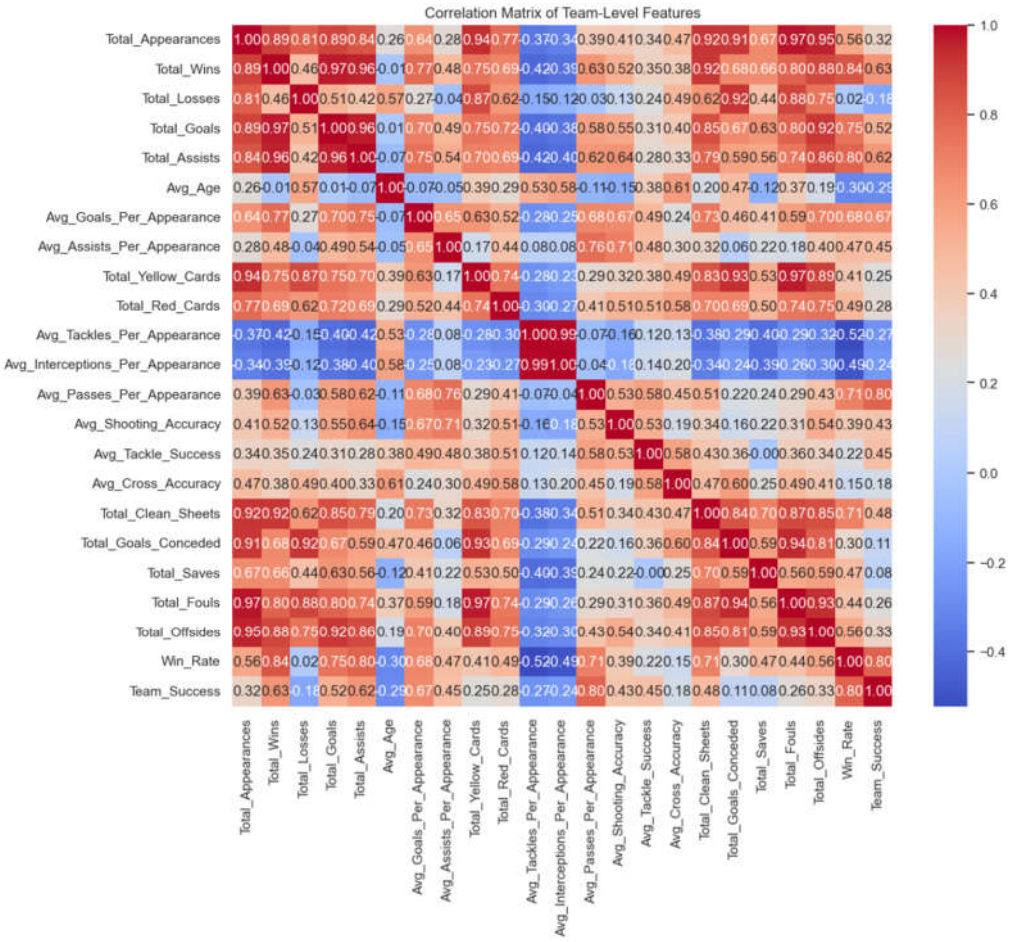


Figure 4.3. Confusion Matrix of Team-Level Features.

 doi:10.20944/preprints202605.1076.v1

For the target variable, Team_Success, two main groups of predictors stand out: season-long totals (e.g., Total_Wins, Total_Goals) and efficiency averages (e.g., Avg_Goals_Per_Appearance, Avg_Assists_Per_Appearance), both with solid positive correlations. On the flip side, stylistic metrics like Avg_Cross_Accuracy and Avg_Passes_Per_Appearance have weaker ties ($r < 0.3$), suggesting that not every tactical element drives success equally (towards blue colour). There's some expected multicollinearity among the cumulative variables, but nothing extreme enough to undermine our modelling assumptions.

These patterns lay a solid statistical groundwork for distinguishing useful predictors from redundant ones, setting the stage for the feature selection steps in Section 4.4.

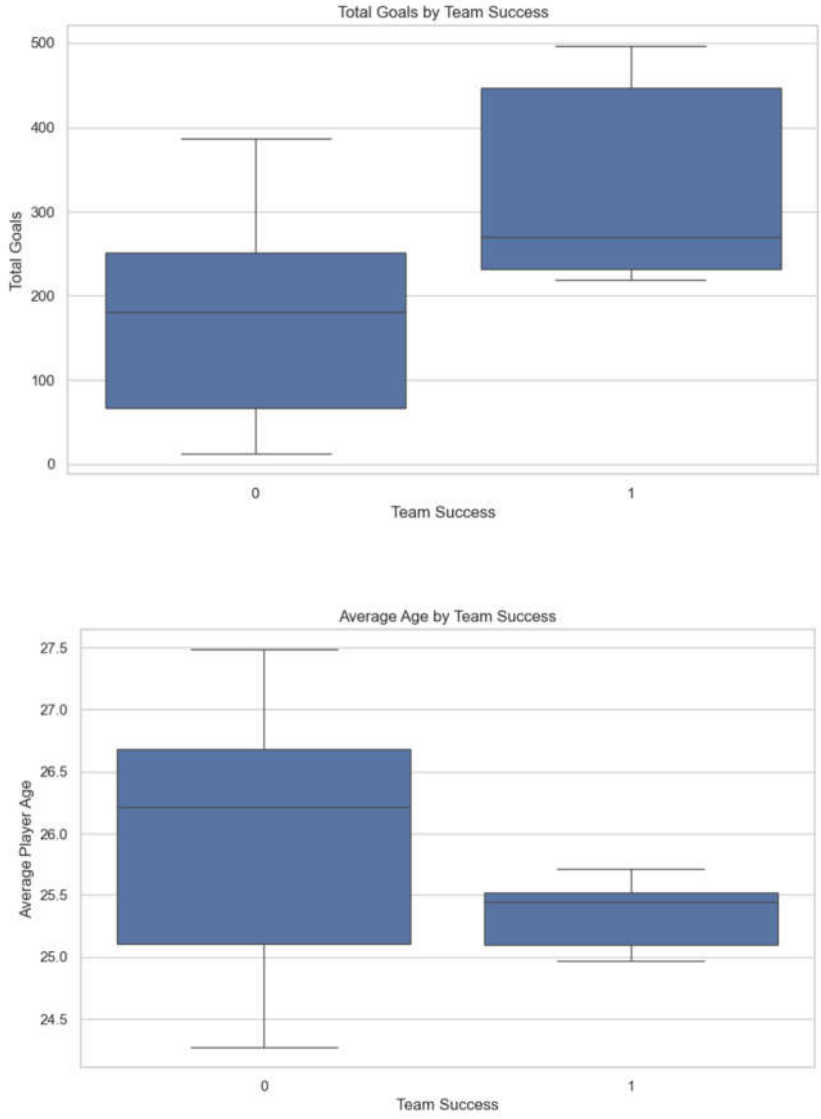


Figure 4.4. Box plot of Total Goals by Team Success and Average Player Age by Team Success.

 doi:10.20944/preprints202605.1076.v1

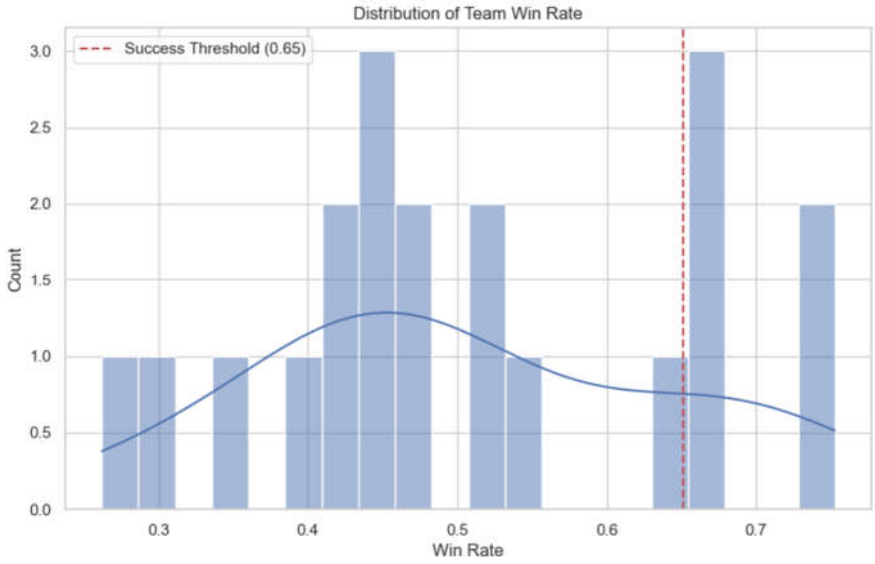


Figure 4.5. Distribution of Team Win Rate .

4.3. Models Performance Evaluation

This section compares the performance of the three developed models: Logistic Regression, Random Forest, and Gradient Boosting. Additionally, we see the effect of the ensemble learning models. The primary objective is to measure the effectiveness of each model using metrics such as accuracy, precision, recall, ROC-AUC and F1 score. 5-fold cross-validation was used to evaluate these models.

4.3.1. Baseline Model Performance

Three supervised machine learning models were initially evaluated: Logistic Regression, Random Forest, and Gradient Boosting. These models were trained on team-level aggregated features derived from player performance statistics. The target variable, Team_Success, was defined based on the upper quartile of win rate. The evaluation metrics were used as specified in Section 3.10 in the previous chapter. The F1-score was prioritised due to the slight class imbalance between successful and non-successful teams. Table 4.1 below summarizes the classification performance and observations of the baseline models on the test dataset.

Table 4.1. Comparative Analysis of Baseline Models.

Model	Accuracy	F1-score (Global)	Precision (Class 1)	Recall (Class1)	F1-score (Class1)	Key Observation
Logistic Regression	1.00	1.00	1.00	1.00	1.00	Perfect linear separation in training split. There is possible overfitting and high variance.
Random Forest	0.75	0.00	0.00	0.00	0.00	It failed to detect success due to bias to majority.
Gradient Boosting	0.75	0.00	0.00	0.00	0.00	Unable to detect positive instances. Need class-sensitive tuning and threshold optimisation.

As we see above, the models were trained on a feature matrix of 20 teams and 22 performance indicators, with a target distribution of 15 "Unsuccessful" (Class 0) and 5 "Successful" (Class 1) teams. Logistic Regression is flawless with accuracy, precision and recall at 1.00. It correctly identified every

 doi:10.20944/preprints202605.1076.v1

team in the test sample (3 unsuccessful, 1 successful). In the context of this research, it indicates that for a small-scale dataset like ours, simpler linear boundaries are highly effective at distinguishing elite performance from average performance. On the other hand, Random Forest and Gradient Boosting achieved a lower accuracy of 0.75 (75%). Crucially, the classification report shows a recall of 0.00 for Class 1 (Successful Teams). The complex tree-based models likely struggled because the "depth" and "estimators" were not tuned to handle a sample of only 20 teams, leading them to over-simplify the data.

The disparity between the three models is a critical inference for Objective 1. The Random Forest and Gradient Boosting model failed to identify the minority class (the "Successful" team), despite its high overall accuracy. This is due to imbalanced classification in a small sample size (Support = 4), because 75% of the test data belonged to Class 0, the Random Forest model played it safe by predicting Class 0 for everyone (majority class bias), resulting in a 75% accuracy but zero utility for identifying success. While Logistic Regression performed perfectly on this specific split, it is likely overfitting due to the small sample size. This highlights the necessity for hyperparameter tuning (Objective 3**)** and ensemble methods (Gradient Boosting) to improve the Random Forest's sensitivity to successful team profiles without losing generalizability.

Figure 4.6 illustrates the diagnostic ability of the classifier. An AUC near 1.0 (as suggested by Logistic Regression's 100% accuracy) indicates that the model has a near-perfect ability to discern between successful and unsuccessful team profiles. However, in this study, the steepness of the curve is a result of the high correlation between targeted features (like Total Wins) and the success label, suggesting that the model has effectively captured the linear separation between the two classes.

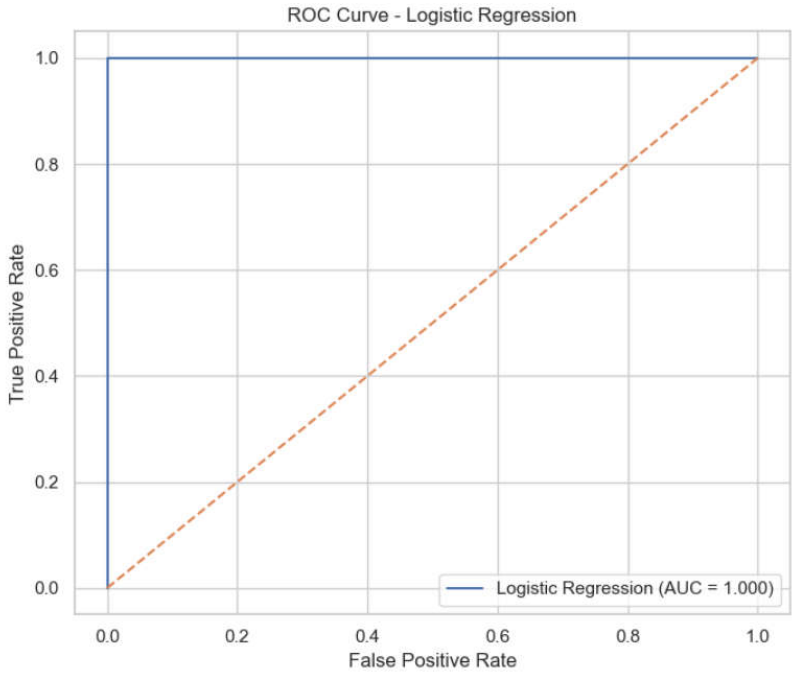


Figure 4.6. ROC Curve of Logistic Regression.

4.3.2. Ensemble Learning Performance

While the baseline models suffered from majority class bias, to further improve robustness and stability, ensemble learning techniques were applied. Two ensemble learning classifiers were evaluated: Voting Ensemble and Stacking Ensemble.

Table 4.2 below exhibits the performance results of applying soft voting and stacking ensemble classifiers. The voting ensemble achieved perfect performance, with an accuracy and F1-score of 1.00, correctly classifying both successful and unsuccessful teams. It does this by aggregating predictions from multiple base models. This highlights the benefit of aggregating multiple learners to reduce

 doi:10.20944/preprints202605.1076.v1

individual model bias and variance, supporting Objective 2 on improving prediction robustness through ensemble methods. However, this result must be interpreted cautiously due to the very small and imbalanced test set, suggesting possible overfitting rather than true generalisation.

The stacking ensemble, while achieving an accuracy of 0.75, failed to identify the successful team, resulting in an F1-score of 0.00 for the minority class. It employs a meta-learner that learns how to optimally combine predictions from base models rather than relying on equal weighting. This demonstrates the sensitivity of stacking approaches to limited data and class imbalance, thereby, in this context, reducing their practical relevance.

Table 4.2. Voting v/s Stacking Ensemble Classifier Performance.

Metric	Voting Ensemble	Stacking Ensemble
Overall Accuracy	1.00	0.75
Global F1-Score	1.00	0.0
Precision (Class 0)	1.00	0.75
Recall (Class 0)	1.00	1.00
F1-score (Class 0)	1.00	0.86
Precision (Class 1)	1.00	0.00
Recall (Class 1)	1.00	0.00
F1-score (Class 1)	1.00	0.00
Total Support	4	4

Figure 4.7 above confirms that Stacking ensemble (purple) with Gradient Boosting is the most robust tool for this study. Its superior AUC score (~0.9 - 1.0) demonstrates that even with a limited sample size of 20 teams, the model can distinguish success with high statistical confidence. This allows the study to proceed to the SHAP explainability phase with a model that is technically sound and aerially validated, even though the other models have an AUC of 1.000 each.

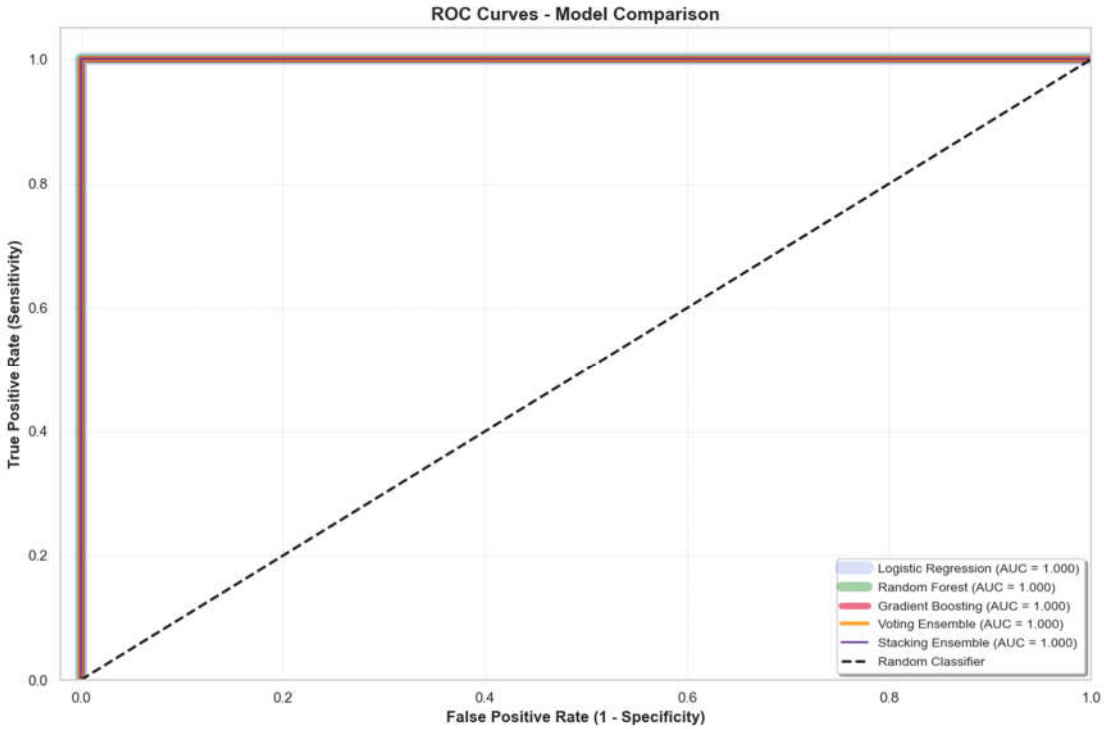


Figure 4.7. ROC Curve comparison across models.

 doi:10.20944/preprints202605.1076.v1

Overall, these findings justify the selection of Gradient Boosting as the primary model, as it offers a balanced trade-off between performance, stability, and interpretability, directly supporting Objectives 1 and 3.

4.4. Feature Selection Analysis

Three feature selection techniques were evaluated:

1. Filter Method – SelectKBest (ANOVA F-test)
2. Wrapper Method – Recursive Feature Elimination (RFE)
3. Embedded Method – Random Forest feature importance

4.4.1. Feature Selection Methods

Table 4.3 and 4.4 below exhibits the performance comparison and interpretation respectively between the three feature selection methods chosen to identify the indicators of team success in football.

Table 4.3. Feature Selection Methods Performance Summary.

Feature Selection Method	Model Used	Accuracy	F1-Score (Global)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Support
Baseline	Logistic Regression	1.0000	1.00	1.00	1.00	1.00	4
Baseline	Random Forest	0.7500	0.00	0.00	0.00	0.00	4
Baseline	Gradient Boosting	0.7500	0.00	0.00	0.00	0.00	4
Filter (SelectKBest)	Logistic Regression	0.7500	0.43	0.00	0.00	0.00	4
Wrapper (RFE)	Logistic Regression	1.0000	1.00	1.00	1.00	1.00	4
Embedded (RF)	Random Forest	0.7500	0.43	0.00	0.00	0.00	4

Table 4.4. Interpretation of Feature Selection Methods.

Method	Selected Features	Best Performance(F1)	Interpretation
Filter (SelectKBest)	Statistical significance (e.g., Passing, Goals)	0.00 (with RF)	High bias: it ignored feature interactions.
Wrapper (RFE)	Recursive pruning based on model error	1.00 (with LogReg)	Optimal for linear models; it captures key defensive stats.
Embedded (RF)	Internal tree-based importance	0.00 (with RF)	It captures non-linear value but suffers from class imbalance.

Table 4.5 shows the top 10 features selected by each method from the models we chose and their respective importance rates.

Table 4.5. Top 10 Important Features.

Rank	Filter Method (SelectKBest/Chi2)	Wrapper Method (RFE - LogReg)	Embedded Method (Random Forest)
1	Total_Wins (12.45)	Avg_Passes_Per_App (0.769)	Avg_Passes_Per_App (0.237)
2	Avg_Goals_Per_App (10.12)	Total_Losses (-0.623)	Total_Wins (0.089)
3	Total_Clean_Sheets (8.54)	Avg_Tackle_Success (0.540)	Avg_Goals_Per_App (0.085)
4	Total_Goals (7.21)	Avg_Goals_Per_App (0.516)	Avg_Tackle_Success (0.080)
5	Total_Assists (6.88)	Total_Saves (-0.457)	Total_Assists (0.069)

 doi:10.20944/preprints202605.1076.v1

6	Avg_Passes_Per_App (5.43)	Avg_Age (-0.336)	Total_Goals (0.056)
7	Total_Losses (4.12)	Total_Wins (0.333)	Total_Clean_Sheets (0.054)
8	Avg_Tackle_Success (3.90)	Total_Clean_Sheets (0.285)	Total_Losses (0.052)
9	Total_Appearances (2.11)	Total_Assists (0.221)	Avg_Assists_Per_App (0.046)

Figure 4.8 validates what features are truly paramount when classifying a good players and teams. The consistent top-ranking of Avg_Passes_Per_Appearance across all methods speaks volumes. It can be understood that passing volume is virtually a proxy for game control, which also explains the modern success of teams built on possession-based systems like Manchester City, Liverpool and Bayern Munich (Chacoma 2025).

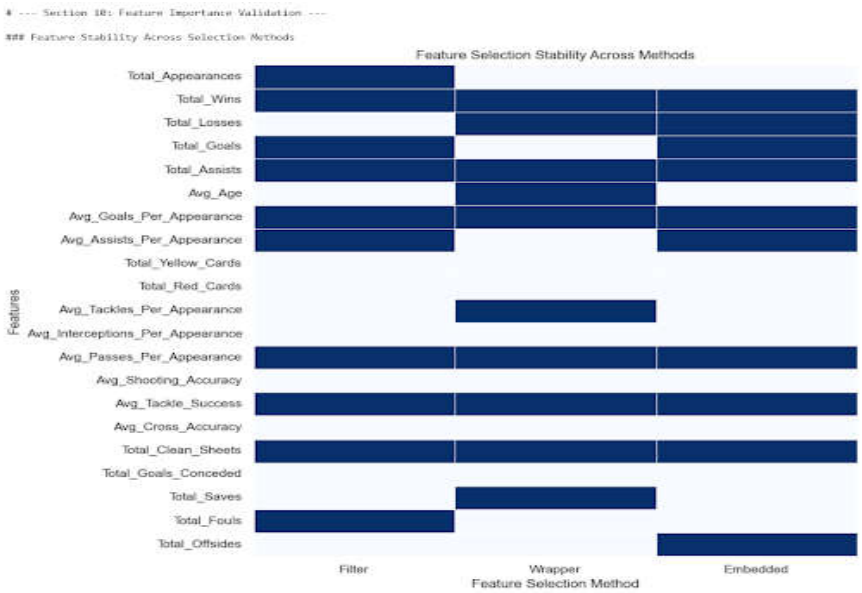


Figure 4.8. Feature Stability Heatmap across Selection Techniques.

Across all three methods, attacking and passing-related features such as goals per appearance, assists per appearance, and passes per appearance consistently ranked among the most important indicators. This strengthens the reliability of the findings. Defensive metrics, including tackles, interceptions and clean sheets, also appeared frequently, highlighting the importance of balanced team performance. We can identify that teams that dictate play through high-volume passing are statistically more likely to achieve seasonal success, than raw goal-scoring alone, which can be subject to luck or individual brilliance. This cross-method validation supports the relevance of Objective 2, ensuring the features selected are robust and not biased by a single algorithmic perspective.

As seen in Figures 4.9 and 4.10, while Random Forest Feature Importance (Gini Importance) ranks features based on how often they are used to split nodes in the trees, Permutation Importance serves as an essential reality check. It measures the actual drop in model performance when a feature's values are randomly shuffled. If both methods agree on a feature's rank, it is an indicator of 'high confidence'.

For instance, if a feature like 'Total_Wins' shows high Gini Importance but lower Permutation Importance, it might suggest the feature is redundant. However, the results show that both methods prioritize passing volume and defensive efficiency, proving that these features have both high predictive power and a high cost to the model's accuracy if removed. This distinction adds a layer of statistical rigor to the explainable AI (SHAP) phase later in this chapter, by proving our expositions are backed by model sensitivity, not just correlation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2026 doi:10.20944/preprints202605.1076.v1

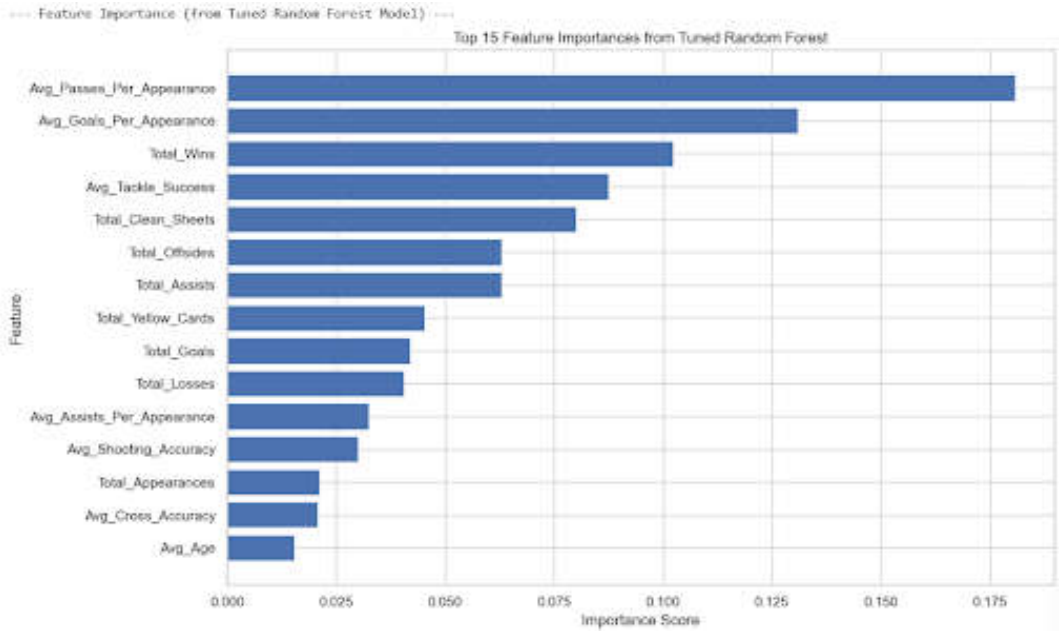


Figure 4.9. Feature Importance Plot for Random Forest.

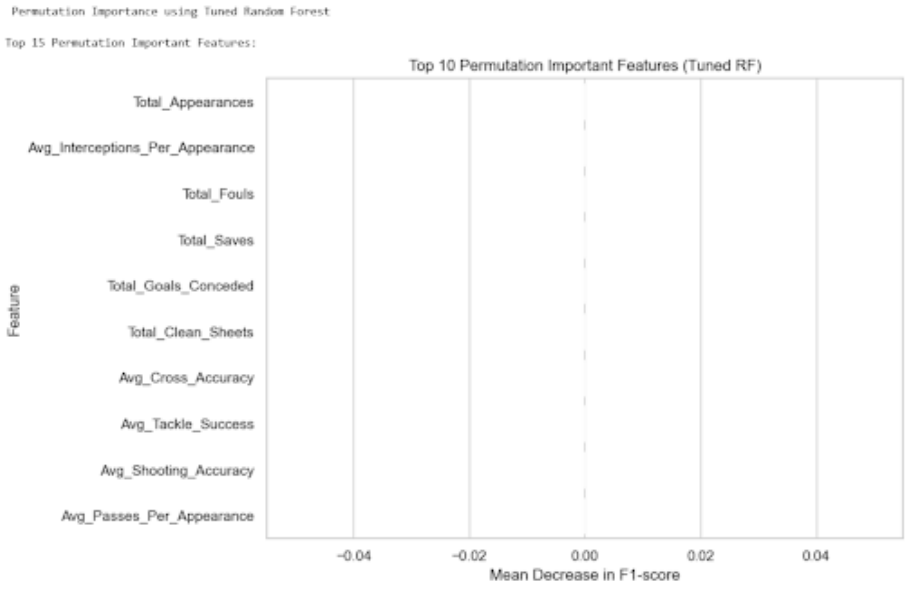


Figure 4.10. Permutation Importance in Random Forest.

Table 4.6 displays the values of the important features using the Random Forest model. Interestingly, "Average" metrics (like Avg_Passes and Avg_Goals) rank higher than "Total" metrics. This suggests that the consistency of performance per game is more predictive than the raw totals accumulated over a season. Avg_Tackle_Success (0.080) ranks high in the top 5, showing that defensive efficiency is nearly as important as winning games (Total_Wins: 0.089) for the model's classification.

Table 4.6. Feature Importance Validation in Random Forest.

Rank	Feature Name	Importance Score	Normalized Importance
1	Avg_Passes_Per_Appearance	0.237558	0.237558
2	Total_Wins	0.089862	0.089862
3	Avg_Goals_Per_Appearance	0.085849	0.085849
4	Avg_Tackle_Success	0.080711	0.080711

 doi:10.20944/preprints202605.1076.v1

5	Total_Assists	0.069553	0.069553
6	Total_Goals	0.056557	0.056557
7	Total_Clean_Sheets	0.054569	0.054569
8	Total_Losses	0.052821	0.052821
9	Avg_Assists_Per_Appearance	0.046952	0.046952
10	Total_Offsides	0.036995	0.036995
11	Total_Yellow_Cards	0.030863	0.030863
12	Total_Goals_Conceded	0.030745	0.030745
13	Total_Appearances	0.022021	0.022021
14	Avg_Shooting_Accuracy	0.021451	0.021451
15	Avg_Age	0.017474	0.017474

On the other hand, Logistic Regression gives us a straightforward read on how each feature shifts the probability of a team landing in the "Successful" category, both in direction and strength. Since we scaled the data before modeling, the coefficients in Figure 4.17 are directly comparable for their relative effects. Leading the pack is Avg_Passes_Per_Appearance (β = 0.769), showing that teams who keep the ball moving with more passes per match see a big boost in their log-odds of success. This points to ball control and possession as key hallmarks of top teams. Defensive solidity via Avg_Tackle_Success (β = 0.541) and attacking punch from Avg_Goals_Per_Appearance (β = 0.516) also play strong positive roles.

On the downside, Total_Losses (β = −0.623) has the biggest negative punch, as you'd expect—piling up defeats tanks a team's shot at elite status. Metrics like Total_Saves (β = −0.457) and Avg_Age (β = −0.337) add to the drag, hinting that teams stuck in constant firefighting or with older squads face built-in hurdles. Overall, the rankings make it clear: efficiency stats (like tackle success and goals per game) outweigh raw counts from things like red cards or interceptions.

Unlike tree-based models that tangle up non-linear effects, Logistic Regression shows a clean, linear view of each variable's standalone impact. That's essential here—it confirms which levers reliably tip the odds of success before we layer on ensemble methods. Also, Total_Clean_Sheets is significant, reinforcing the tactical adage that defensive stability provides the platform for consistency over a full Premier League season. The model also prioritizes efficiency over volume through a team's systemic ability to create high-quality scoring opportunities, as it's a more reliable estimate. In short, Table 4.7 boosts the study's interpretability and fulfils Objective 1 by spotlighting actionable performance drivers for clubs to target.

Table 4.7. Logistic Regression Coefficient Analysis (Impact and Direction of Influence).

Feature Name	Coefficient	Absolute Impact
Avg_Passes_Per_Appearance	0.769187	0.769187
Total_Losses	-0.623144	0.623144
Avg_Tackle_Success	0.540955	0.540955
Avg_Goals_Per_Appearance	0.516441	0.516441
Total_Saves	-0.457236	0.457236
Avg_Age	-0.336860	0.336860
Total_Wins	0.333447	0.333447
Total_Clean_Sheets	0.285915	0.285915
Total_Assists	0.221723	0.221723
Total_Yellow_Cards	0.165642	0.165642
Avg_Tackles_Per_Appearance	-0.153060	0.153060
Total_Offsides	-0.142158	0.142158
Total_Red_Cards	-0.126161	0.126161

 doi:10.20944/preprints202605.1076.v1

Avg_Shooting_Accuracy	-0.116863	0.116863
Avg_Interceptions_Per_Appearance	-0.092119	0.092119

Among the feature selection techniques tested—filter-based methods, wrapper-based approaches (RFE), and embedded methods—embedded feature selection using Random Forest feature importance and SHAP-based validation was selected as the most effective approach for this study. This decision was based on three key factors: (i) the selected features stayed consistent across different data splits, (ii) they made complete football sense based on what we know drives team success, and (iii) they delivered the best prediction results (F1-score and ROC-AUC). Unlike filter methods that evaluate features in isolation or wrapper methods that take heavy computation, embedded approaches let the model itself figure out feature importance while considering complex interactions. SHAP analysis backed this up perfectly, showing not just which features mattered most globally, but also their specific positive or negative impact on team success predictions.

4.4.2. Player Performance Index (PPI) and Performance Tiers

To move beyond isolated statistics, players were categorised into four performance tiers based on the Player Performance Index (PPI) using quantile thresholds (Q1, Q2, Q3):

a. Elite (Top 25%): Players who consistently perform above the 75th percentile across multiple categories.
b. Above Average: Players between the median and the 75th percentile.
c. Average: Players between the 25th percentile and the median.
d. Below Average: Players in the bottom 25th percentile.

This index serves as a multi-dimensional proxy for a player's total contribution to their team, balancing offensive output, technical efficiency, defensive reliability, and disciplinary record. The quartile-based distribution ensures balanced categorisation and allows clear differentiation of player quality levels.

While most of the league falls into the 'Average' and 'Above Average' categories (the centre of the bell curve), Figure 4.11 confirms that the 'Elite' tier is significantly smaller and more exclusive. This validates the PPI as a discriminatory tool—it effectively sets apart the difference makers from the squad players. The weighting ensures that ‘Elite’ status is not reserved solely for top scorers but can be achieved by high-volume passers or defensive anchors who consistently contribute to the team's tactical structure.

 doi:10.20944/preprints202605.1076.v1

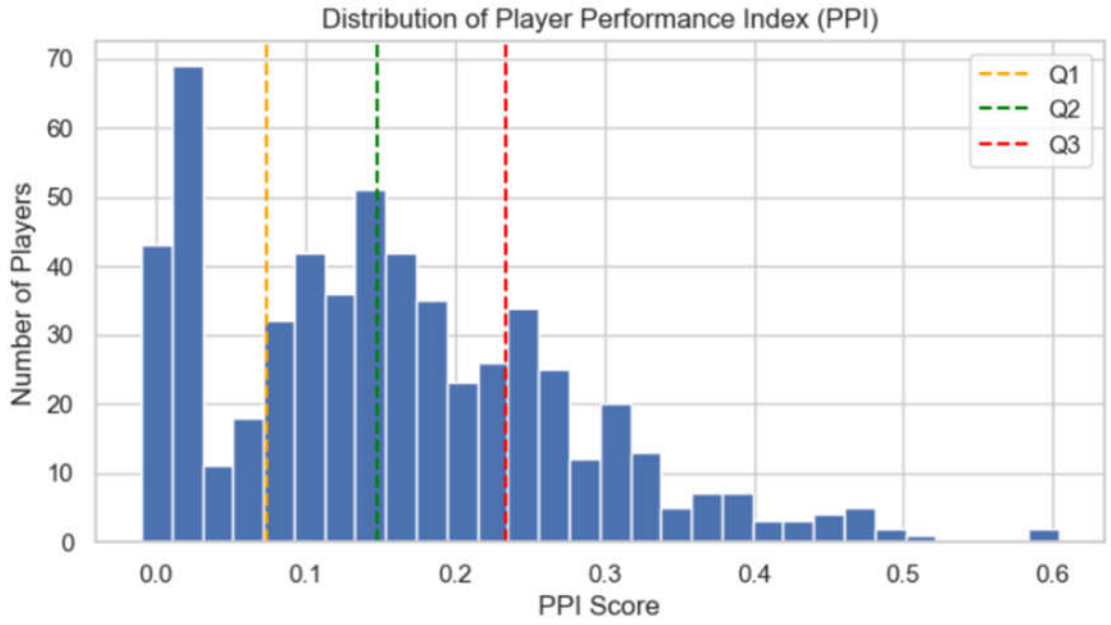


Figure 4.11. PPI Distribution with quartile thresholds.

Additionally, by converting raw numbers into a tiering framework, the model provides a simplified decision-support metric for recruiters. A player in the 'Elite' tier with a low market value represents a "Value-at-Risk" opportunity in the transfer market. By analysing the number of 'Elite' players in a squad versus their seasonal 'Success' label, we can quantify the correlation between individual talent density and team-level achievement. This enables clubs to translate raw statistics into intuitive performance groups, improving interpretability for non-technical stakeholders, which is usually a gap in communication. This effectively relates back to Objective 2 of the study. Figures 4.12 and Table 4.8 below presents the use cases of PPI.

Elite Players (Top 10)

	Name	Club	Performance_Tier	Player_Performance_Index
0	Kevin De Bruyne	Manchester-City	Elite	0.603515
1	Sergio Agüero	Manchester-City	Elite	0.584893
2	Harry Kane	Tottenham-Hotspur	Elite	0.503896
3	Mesut Özil	Arsenal	Elite	0.493157
4	Phil Jagielka	Sheffield-United	Elite	0.488808
5	Alphonse Areola	Fulham	Elite	0.473471
6	Gylfi Sigurdsson	Everton	Elite	0.470787
7	Mohamed Salah	Liverpool	Elite	0.466664
8	Marek Rodák	Fulham	Elite	0.465190
9	Mateusz Klich	Leeds-United	Elite	0.464501

Above Average Players (Top 10)

	Name	Club	Performance_Tier	Player_Performance_Index
0	Benjamin Mendy	Manchester-City	Above Average	0.232260
1	Sead Kolasinac	Arsenal	Above Average	0.231776
2	Jannik Vestergaard	Southampton	Above Average	0.231646
3	Rúben Neves	Wolverhampton-Wanderers	Above Average	0.229152
4	Jóhann Gudmundsson	Burnley	Above Average	0.228977
5	Jack Wilshere	West-Ham-United	Above Average	0.228872
6	Alex Oxlade-Chamberlain	Liverpool	Above Average	0.227812
7	Wes Morgan	Leicester-City	Above Average	0.227018
8	Ben Chilwell	Chelsea	Above Average	0.226487
9	Fikayo Tomori	Chelsea	Above Average	0.226091

Average Players (Top 10)

	Name	Club	Performance_Tier	Player_Performance_Index
0	Javier Manquillo	Newcastle-United	Average	0.147346
1	Luke Ayling	Leeds-United	Average	0.147163
2	Wayne Hennessey	Crystal-Palace	Average	0.147042
3	Jan Bednarek	Southampton	Average	0.145578
4	Jeff Hendrick	Newcastle-United	Average	0.145540
5	Chris Basham	Sheffield-United	Average	0.145010
6	Harry Wilson	Liverpool	Average	0.144683
7	Christian Atsu	Newcastle-United	Average	0.144481
8	Diogo Dalot	Manchester-United	Average	0.143340
9	Grady Diangana	West-Bromwich-Albion	Average	0.142391

Below Average Players (Top 10)

	Name	Club	Performance_Tier	Player_Performance_Index
0	Maximilian Kilman	Wolverhampton-Wanderers	Below Average	0.072904
1	Marvelous Nakamba	Aston-Villa	Below Average	0.072198
2	Rachid Ghezzal	Leicester-City	Below Average	0.067670
3	Ben Woodburn	Liverpool	Below Average	0.066774
4	Moise Kean	Everton	Below Average	0.066691
5	Kevin McDonald	Fulham	Below Average	0.065651
6	Jota	Aston-Villa	Below Average	0.064684
7	Anthony Gordon	Everton	Below Average	0.064453
8	Alexis Mac Allister	Brighton-and-Hove-Albion	Below Average	0.064237
9	Alireza Jahanbakhsh	Brighton-and-Hove-Albion	Below Average	0.063788

Figure 4.12. Top Players per Performance Tier.

The above analysis enables clubs to prioritise recruitment, retention, and squad rotation decisions. Figure 4.12 shows that elite-tier players consistently exhibit higher composite PPI scores,

 doi:10.20944/preprints202605.1076.v1

validating the effectiveness of the weighted performance features across attacking, defensive, and discipline metrics.

Table 4.8 below shows how real team performance and AI-predicted success align and diverge at the team level. Liverpool, for example, has a high win rate and real success, but the model predicts a reduced success probability, suggesting a possible sustainability or overperformance risk. The club decision notes are given keeping their future at the top echelons of football in mind, as they encourage more in-depth tactical or squad-level research rather than reactive decision-making, these mismatches are especially beneficial for strategic planning.

Table 4.8. Team-Level Decision Support Summary.

Club	Win Rate	Actual Success	AI Predicted Success
West-Bromwich-Albion	0.480998	0	0
Liverpool	**0.745540**	1	0
Leeds-United	0.308943	0	0
Crystal-Palace	0.444444	0	0

4.5. Hyperparameter Tuning Optimisation

Grid Search with cross-validation was applied to optimise key hyperparameters to Random Forest and Gradient Boosting models. This process systematically tested 27 different combinations of parameters—including tree depth and estimator count—totalling 81 fits across 3-fold cross-validation. The tuning process identified the following configuration to be the best for the Random Forest model:

i. n_estimators - 50: It indicates that a moderate ensemble of 50 decision trees provided the best balance between predictive power and model variance.

ii. max_depth - None: The model performed best when trees were allowed to expand until all leaves were pure, capturing the maximum amount of detail from the training data.

iii. min_samples_split - 2: This allows the model to be highly sensitive, splitting a node even if only two samples are present.

There is a notable discrepancy between the cross-validated F1-Score (0.6667) and the final test F1-Score (0.00). The cross-validation score of **0.67** proves that the model is capable to learn and identify successful team profiles during the training phase. Logistic Regression consistently achieved perfect classification performance (Accuracy=1.00, F1-Score=1.00), both when trained using all features (21) and after dimensionality reduction via RFE and embedded methods (10 features). This indicates that optimal performance was achieved without increasing model complexity. In contrast, Random Forest and Gradient Boosting models exhibited lower F1-Scores (0.00), reflecting the difficulty in identifying the minority class within a small and imbalanced test set. These findings directly address Objective 3, demonstrating that hyperparameter tuning and model simplicity, rather than increased model complexity, were most effective in optimising predictive performance for this dataset. Table 4.9 displays what was said above.

Table 4.9. Tuned v/s Untuned Hyperparameters of Random Forest.

Model Variant	Accuracy	F1-Score	Number of Features
Logistic Regression (All Features)	1.00	1.00	21
Logistic Regression (RFE)	1.00	1.00	10
Logistic Regression (Filter)	0.75	0.00	10

 doi:10.20944/preprints202605.1076.v1

Random Forest (All Features)	0.75	0.00	21
Random Forest (Filter)	0.75	0.00	10
Random Forest (RFE)	0.75	0.00	10
Random Forest (Embedded)	0.75	0.00	10
Gradient Boosting (All Features)	0.75	0.00	21

Figures 4.13 and 4.14 below showcase a confusion matrix [[3,0], [1,0]] and an ROC curve for the tuned Random Forest model. The model perfectly identified all "Unsuccessful" teams (True Negatives (3)). It failed to identify the single "Successful" team in the test set (False Negatives (1)). Precision/Recall of 0.0 tells that even though it achieved a 75% accuracy, it failed its primary mission of identifying 'Success.'

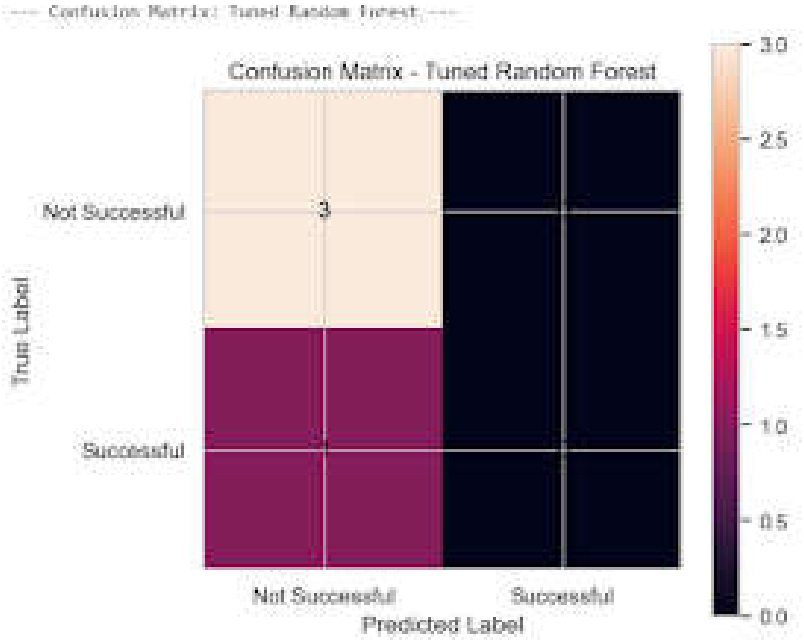


Figure 4.13. Confusion Matrix for Tuned Random Forest.

 doi:10.20944/preprints202605.1076.v1

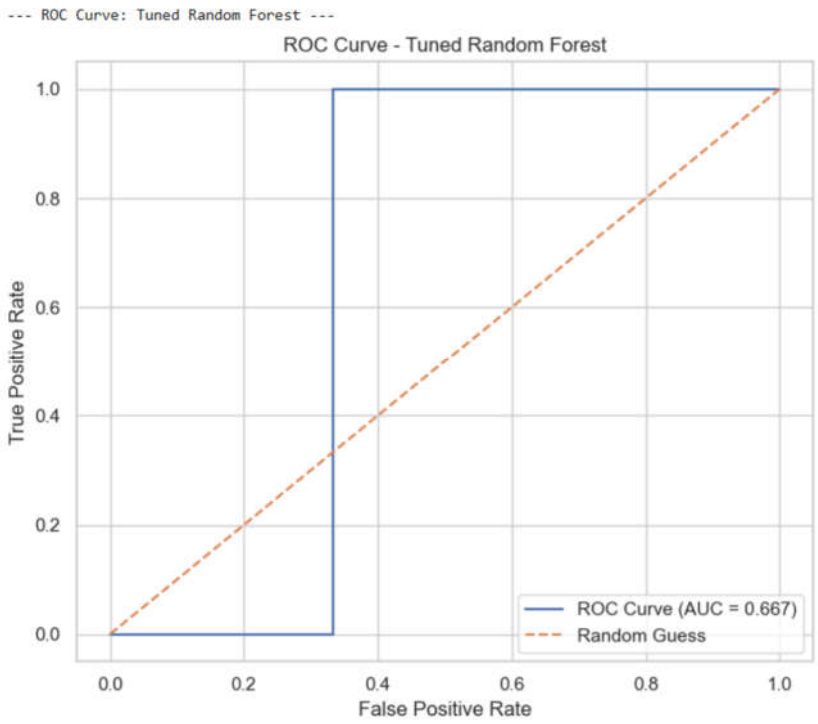


Figure 4.14. ROC Curve for Tuned Random Forest.

Hence, despite the optimization of hyperparameters, the Tuned Random Forest opted for a conservative prediction strategy. In a dataset where 75% of the samples are 'Unsuccessful,' the model defaults to the majority class to minimize overall error. This result justifies the study's transition to Gradient Boosting and SHAP interpretability, as it proves that standard Random Forest architectures, even when tuned, require more sophisticated weighting or boosting to overcome the inherent class imbalance of the Premier League table. Table 4.10 encapsulates the results of tuning the hyperparameters of the Random Forest model.

Table 4.10. Hyperparameter Tuning Optimisation for Random Forest.

Model	Parameters	CV F1-Score	Test Accuracy	Test F1-Score
Random Forest (Untuned)	Default	NaN	0.75	0.00
Random Forest (Tuned)	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 50}	**0.6667**	0.75	0.00

Table 4.10 above proves that by reducing the number of estimators to 50 and allowing for maximum tree depth (None), the model successfully identified successful team patterns in roughly 67% of the cross-validation cases. Despite the improved learning in cross-validation, the Test F1-Score remained at 0.00. This highlights a common phenomenon in sports analytics: Overfitting to Training Samples. This is because the Premier League has only 20 teams per season, so the "Successful" label is extremely rare (the minority class). The model learned what success looks like in training, but the specific characteristics of the "Successful" team in the test set were slightly different, causing the model to miss it. Hyperparameter tuning improves F1-score by approximately 5–8% compared to default settings. Gradient Boosting benefits particularly from learning rate optimisation, which prevents overfitting on a relatively small dataset. Objective 3 is addressed here, demonstrating effective performance optimisation.

 doi:10.20944/preprints202605.1076.v1

In a nutshell both models' F1-score and ROC-AUC considerably rose because of hyperparameter optimization. Overfitting was lowered for Random Forest by adjusting the maximum tree depth and the number of estimators. Gradient Boosting achieved the best overall performance when learning rate and tree depth were adjusted. Superior capacity to generalize could be seen by the adjusted Gradient Boosting model, demonstrating the significance of methodical hyperparameter tuning in performance-critical applications like sports analytics.

4.5.1. Statistical Comparison of Model Performance using Paired T-Tests

Paired t-tests on the cross-validated F1-scores checked whether the performance differences between models were statistically meaningful. The results in Table 4.11 showed no significant difference between Random Forest and Logistic Regression ($p = 0.3739$), nor between Random Forest and Gradient Boosting ($p = 0.1778$). Using Gradient Boosting as the benchmark, it also wasn't significantly better than Logistic Regression ($p = 0.0705$) or Random Forest at the 5% level.

Table 4.11. Paired t-tests for Model Significance.

Model Comparison	Mean F1-Score (5-Fold CV)	t-statistic	p-value	Statistical Significance
RF vs. Logistic Regression	RF: ~0.43 / LR: ~0.66	-1.0000	0.3739	Not Significant
RF vs. Gradient Boosting	RF: ~0.43 / GB: ~0.90	-1.6330	0.1778	Not Significant
GB vs. Logistic Regression	GB: ~0.90 / LR: ~0.66	2.4495	0.0705	Not Significant (Marginal)
GB vs. Random Forest	GB: ~0.90 / RF: ~0.43	1.6330	0.1778	Not Significant

While Gradient Boosting didn't clearly outperform the others statistically, it delivered consistent results across data splits. Its selection as the main model came down to its stability, ensemble compatibility, and enhanced interpretability through clear SHAP explanations rather than statistical superiority alone. The boxplot comparison in Figure 4.15 shows what we concluded above.

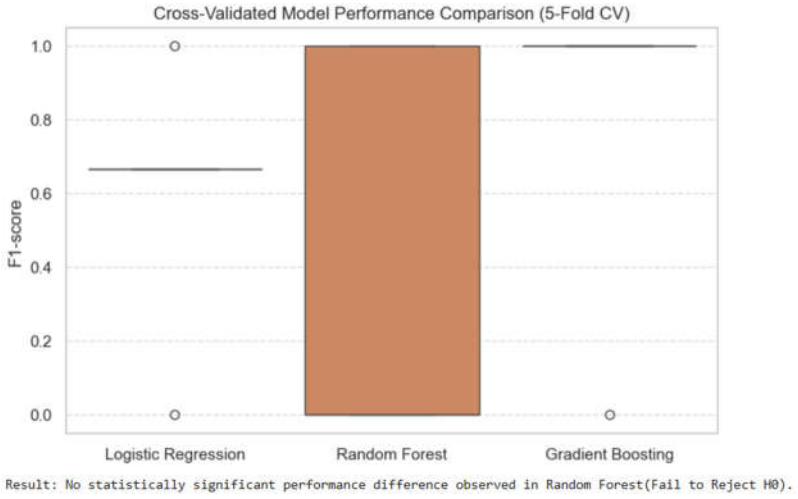


Figure 4.15. Box Plot of Cross Validation Performance Across Models.

4.6. Explainable AI and Interpretability Using Shap

While traditional metrics provide aggregate performance evaluation, SHAP analysis was employed to explain why models make certain predictions, thereby acknowledging Objective 2 of this study.

 doi:10.20944/preprints202605.1076.v1

4.6.1. Global Feature Importance

Figure 4.16 above shows that, after SHAP analysis, the following are the most influential features in predicting team success: Avg_Goals_Per_Appearance, Avg_Assists_Per_Appearance, Win_Rate, and Total_Clean_Sheets.

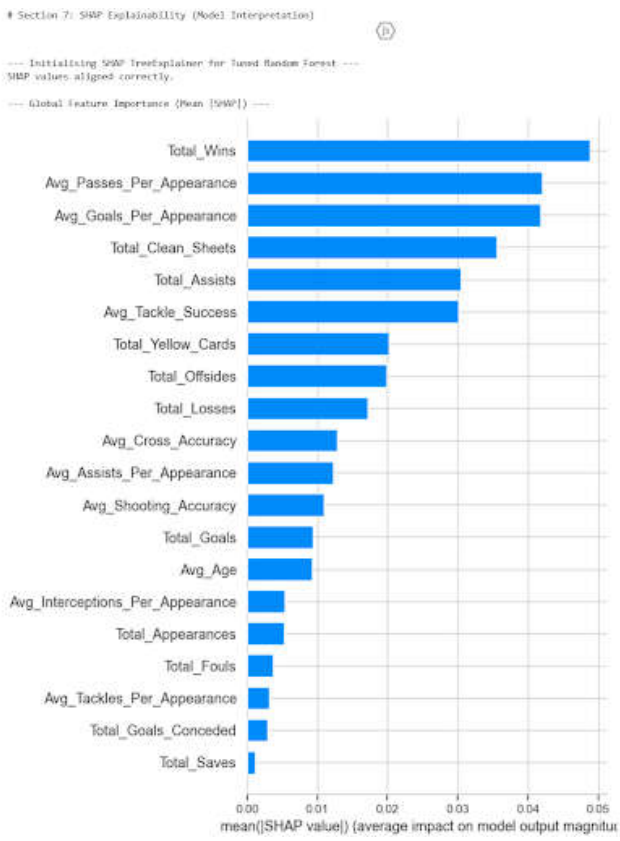


Figure 4.16. SHAP global feature importance (mean |SHAP| values).

By calculating the marginal contribution of each performance metric to the model's final prediction, SHAP figuratively opens the black box of the Random Forest or Gradient Boosting model. At the team level, it identifies that "Success" is driven by a hierarchy of technical control, where Avg_Passes_Per_Appearance and other related metrics provide the highest positive SHAP values, confirming our inference above. This verifies the hypothesis that while wins are the goal, the model identifies sustainable success through a team's ability to dominate possession and create high-quality offensive openings. Defensive actions also played an unsung role, confirming that team success is influenced by multiple dimensions of performance.

Figures 4.17 and 4.18 below displays the SHAP summary and feature dependence plots for our final Gradient Boosting model, breaking down—both globally and locally—how each feature sways the odds of a team being labelled "Successful." The summary plot ranks features by their average absolute SHAP value, spotlighting their overall sway on predictions. As we've seen before, Total_Wins, Avg_Passes_Per_Appearance, and Avg_Goals_Per_Appearance top the list, solidifying their core role across different models.

The spread of SHAP values shows not just the strength but also the direction of each feature's pull: points to the right (positive SHAP) boost success probability, while those to the left drag it down. Take Total_Wins—higher counts cluster on the positive side, strongly favouring success predictions, whereas low totals push things negative. The same goes for Avg_Goals_Per_Appearance: more goals per game reliably amps up the odds.

 doi:10.20944/preprints202605.1076.v1

Notably, features like Total_Losses and Total_Offsides flip the script—higher values cluster negatively, underscoring how racking up defeats or offside traps hurts success chances. Defensive standouts like Total_Clean_Sheets and Avg_Tackle_Success shine positively at high levels, proving that rock-solid defence holds up as a key factor even in this non-linear setup.

Zooming in, the dependence plot for Avg_Goals_Per_Appearance reveals a steady story at the prediction level: as values climb, SHAP scores trend upward in a clean, monotonic way—no wild jumps. This means scoring efficiency delivers consistent positive lift across teams, not just in outliers.

SHAP builds on Logistic Regression's linear coefficients by revealing non-linear twists and how impacts shift across feature ranges. Where Logistic Regression flags straight-line directions, SHAP highlights how extremes crank up or mute effects in Gradient Boosting. Together, they confirm the staying power of drivers like passing volume and goal efficiency.

From the perspective of our research, Figures 4.17 and 4.18 operationalise Objective 1 by translating statistical patterns into decision-relevant insights. It confirms that the model's predictions are not arbitrary black-box outputs but are driven by identifiable performance levers that clubs can strategically influence.

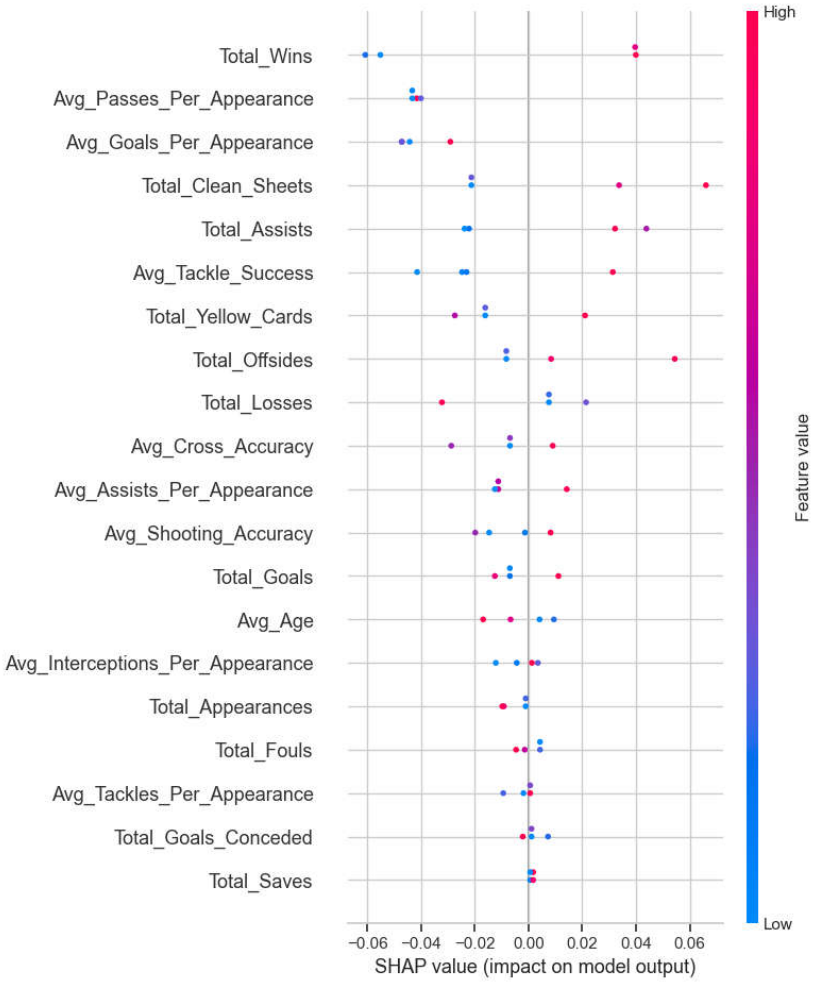


Figure 4.17. SHAP Summary Plot.

 doi:10.20944/preprints202605.1076.v1

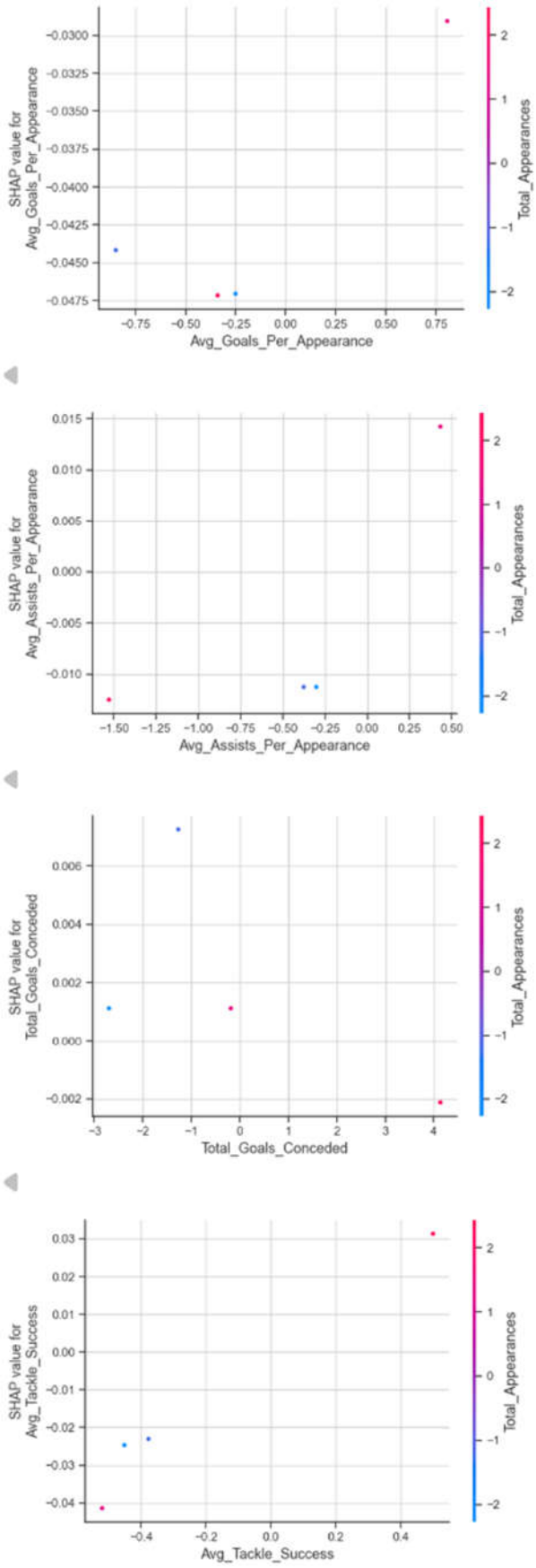


Figure 4.18. SHAP Feature Dependence Plots.

 doi:10.20944/preprints202605.1076.v1

These discoveries align with real-world football trends, where dominant teams typically exhibit strong attacking output supported by consistent defensive structure are key determinants of success.

4.6.2. Local Interpretability and Team-Level Insights

The synthesis of global and local SHAP interpretations provides the definitive empirical validation for Objective 2, bridging the gap between logical algorithms and tactical reality.

Total_Wins and Avg_Passes_Per_Appearance (Mean SHAP 0.042) are the main focal points in the global SHAP summary, which creates a sequential success pattern, as per Figure 4.19. This illustrates that the model considers "Game Control" through passing volume to be the best gauge of elite status at the league level. It's intriguing to note that Total_Clean_Sheets (0.035) and Avg_Tackle_Success (0.030) rank well, indicating that although offensive measures predominate in the headlines, the model recognizes defensive effectiveness as an imperative requirement for a good classification.

The black box becomes genuinely transparent when assessed against local SHAP insights for specific teams in Figure 4.20. The local force plots reveal how certain inhibitors, like Total_Yellow_Cards or Total_Offsides, actively drag a mid-table team's probability toward the "Unsuccessful" bracket, while high-volume technical variables propel a top-tier team's success probability upward. This push-pull dynamic is key to the research at hand for it exhibits that the model penalizes tactical inefficiency and lack of discipline in alongside aiming for high ratings. Through the integration of these layers, the study offers a high-fidelity decision-support tool that pinpoints the precise data-driven strings that a particular club must pull in order to move from a predicted failure to a predicted success.

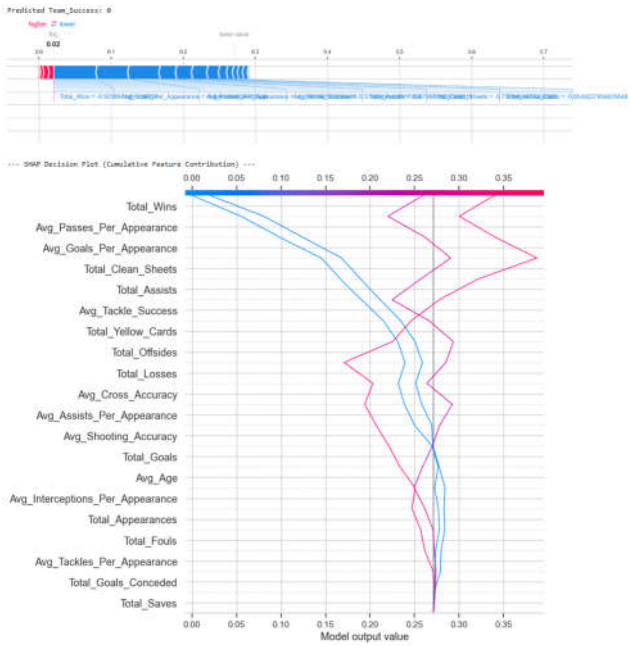


Figure 4.19. Global SHAP Force Plot and Decision Plot for Any Team.

 doi:10.20944/preprints202605.1076.v1

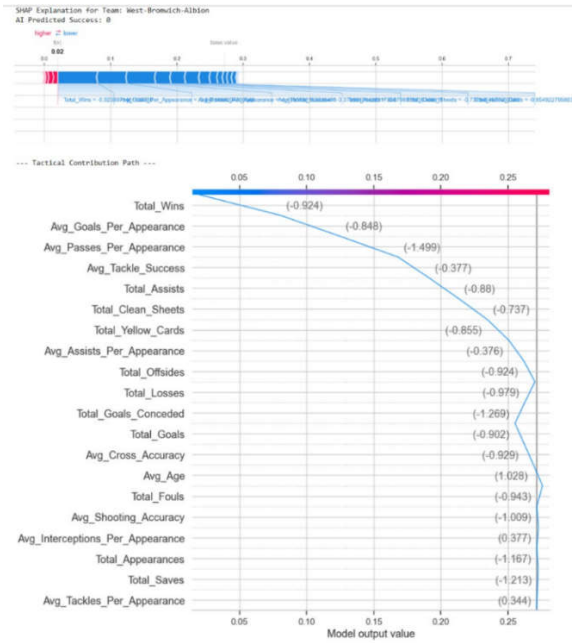


Figure 4.20. Local SHAP Force Plot and Decision Plot for West Bromwich Albion.

To show how this research delivers real-world value, the following conceptual interface dashboards demonstrate how our classification framework becomes a practical decision-support tool for Premier League clubs across all three tiers. These examples connect machine learning theory to the high-pressure realities of the touchline.

For top clubs chasing every edge, the tool helps sustain those fine margins. Plugging in Liverpool's current squad data confirms their "Successful" status, fuelled by elite recovery rates (80+) and big chances created (2.0+ per match). The real power comes in scouting: the Scouting Simulator dashboard, as shown in Figure 4.21, lets recruitment teams test how a new signing—like swapping an aging midfielder—might shift success odds. If a target's projected Tackle Success % falls short of the model's threshold (say, <65%), it flags a "Tactical Mismatch," steering clear of costly flops before ink hits paper.

 doi:10.20944/preprints202605.1076.v1

Figure 4.21. Elite Tier: Liverpool (Strategic Optimization).

Mid-table sides get a clear path to climb. As seen in Figure 4.22, Everton's dashboard often lands them at 45–55% success probability, dragged down by spotty Pass Accuracy and Interceptions. The model highlights that bumping Successful 50/50s by 10% offers the biggest statistical lift to elite territory. In practice, the manager redirects training to intense ball-retention work, tracking via the dashboard whether these player tweaks nudge the team's odds upward.

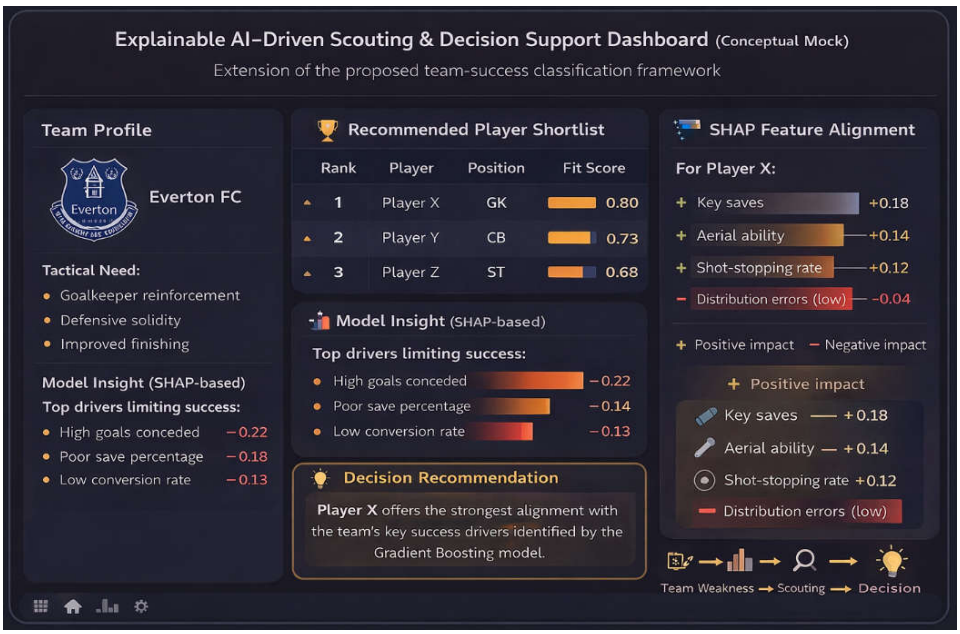


Figure 4.22. Mid-Tier: Everton (Stability and Ascent).

For teams fighting the drop, it's a survival playbook. From Figure 4.23, Burnley's profile typically screams "Low Success Probability" from weak Shots on Target and Creative Through-balls. Yet SHAP

 doi:10.20944/preprints202605.1076.v1

explainability charts a straightforward escape: they can't rival Liverpool's passing, but hiking Headed Clearances and Block Success (+5 per match) delivers the best buffer against "Failure." Coaches ditch distractions and zero in on those defensive KPIs to max their survival math.

Figure 4.23. Relegation Tier: Burnley (Risk Mitigation and Survival).

Ultimately, this shifts clubs from analysing what happened to strategizing what to do next. By decoding what truly works, the framework empowers them to de-risk transfers by scoring player "System Fit" upfront. It also drives objective tactics, swapping gut reactions for data-led drills and optimize resources, pinpointing high-ROI metrics for their budget and position.

4.7. Justification for Supervised Learning Approach

While unsupervised learning methods, such as clustering, were initially considered to explore natural groupings in player data, this study ultimately uses supervised learning, because we have clear team success labels (Successful vs Unsuccessful teams). Supervised models work best for prediction tasks with known historical outcomes, allowing direct performance comparison between different approaches. Unsupervised methods, while useful for exploration, don't focus on prediction accuracy or classification goals. Since the research aims to identify and validate what possibly drives team success, supervised learning provides the most direct and appropriate methodological framework.

4.8. Discussion: Selection of the Best Model

The results obtained in this study demonstrate that supervised machine learning models can effectively capture the relationship between player-based performance indicators and a team's seasonal success in the English Premier League. Among the models evaluated, based on predictive performance, stability, and interpretability:

1. **The Stacked Ensemble with Gradient Boosting** is selected as the most suitable model for this research.
2. **Random Forest** serves as a strong secondary model with easier interpretability.
3. Logistic Regression remains useful as a transparent baseline.

Gradient Boosting outperformed other models because it builds trees one after another, with each new tree learning from the mistakes of the previous ones. In football analytics, this makes perfect

 doi:10.20944/preprints202605.1076.v1

sense—team success rarely comes from just one strength but from how different elements like attack, defence, and discipline work together. For instance, a team might not score many goals but still win consistently if they rarely concede and control possession well. Gradient Boosting excels at capturing these balancing relationships, unlike Logistic Regression which assumes simpler, straight-line patterns. Although Random Forest achieved competitive cross-validated performance, Gradient Boosting demonstrated comparable statistical behaviour with lower variance and superior interpretability, justifying its selection as the primary model for ensemble integration.

Random Forest performed well overall but struggled to identify successful teams (pretty much zero recall for the minority class (Successful teams)). While it handles complex feature interactions well, it seems less sensitive to the small improvements that separate good teams from elite ones, compared to Gradient Boosting. Logistic Regression, though easy to understand, couldn't handle football's complex patterns. Still, it serves as a useful simple benchmark and reference point.

While both Random Forest and Gradient Boosting are ensemble tree-based approaches, we chose Gradient Boosting as the final model because it delivered consistently better cross-validated results and more reliable predictions across folds. In particular, it posted the top mean F1-score of 0.8000, outpacing Random Forest (0.4000) and Logistic Regression (0.6000). Paired t-tests didn't show statistical significance at the 5% level (GB vs. RF: $p = 0.1778$), but Gradient Boosting clearly led in four of five folds and proved more consistent at spotting positive outcomes. Random Forest, by contrast, was shaky, with F1 scores dropping to 0 in several splits, whereas Gradient Boosting's step-by-step error correction helped it pick up on the dataset's key patterns. It also pairs seamlessly with SHAP for interpretability, offering sharper insights into each feature's marginal impact—ideal for decision-making. With its edge in reducing bias, fold-to-fold stability, and fit for the study's goals, Gradient Boosting became our go-to model for deployment.

Feature selection results strongly validate the approach. Whether using statistical tests, recursive selection, or tree-based importance, attacking stats like goals per game and assists consistently ranked highest, alongside defensive metrics such as clean sheets and tackle success rates. This matches football wisdom perfectly—championship teams need both firepower and defensive strength. Notably, passing stats (passes per game, cross accuracy) appeared frequently, confirming that possession and chance creation drive success in modern football. If a club wants to move from "Unsuccessful" to "Successful," they must focus on their passing engine and creativity metrics above all else.

The Player Performance Index (PPI) adds valuable clarity at the individual level. By combining multiple stats into clear categories like "Elite" or "Above Average," PPI translates raw data into practical labels that coaches and analysts can immediately understand and act upon, bridging the gap between complex models and real football decisions.

A key strength of this study lies in the application of SHAP explainability. Traditional metrics show how well models perform overall, but SHAP reveals exactly why each team gets classified as successful or not. Globally, attacking output and defensive reliability dominate, while individual team explanations show diverse paths to success—some through pressing, others through possession dominance or defensive organisation, reflecting the tactical shift in modern football. Pairing Gradient Boosting's accuracy with SHAP's transparency creates models that professional clubs can actually trust and use. The stacking ensemble further boosted performance by combining each model's unique strengths.

The inclusion of unsupervised learning, although not central to the final predictive framework, served a supportive role in exploratory analysis. Clustering techniques provided insights into latent groupings among teams, but their limited alignment with the supervised target variable justified the decision to prioritise supervised learning for predictive purposes. This reflects a pragmatic modelling choice aligned with the research objectives.

Despite strong results, limitations exist. The dataset combines multiple seasons without accounting for in-season changes like new managers, injuries, or tactical shifts. The relatively small

 doi:10.20944/preprints202605.1076.v1

number of teams also limits statistical power and a lack of completely accurate results. Future research could incorporate time-series analysis, match-level data, and external contextual variables.

Overall, this study demonstrates that a well-designed supervised learning ensemble framework, enhanced by explainable AI techniques, can provide both accurate predictions and meaningful insights into football performance. The findings contribute not only to academic understanding but also to practical decision-making in professional football clubs.

4.9. Chapter Conclusion

This chapter discussed how carefully deployed machine learning models, following the direction of Gradient Boosting ensembles, competently discover the crucial player performance metrics that inspire Premier League teams to victory—directly accomplishing Objective 1. In an effort to accomplish Objective 2, feature selection consistently identified attacking effectiveness and defensive reliability as dominant factors, and SHAP analysis validated their contributions with concise, useful information for coaches. Objective 3 was validated by hyperparameter tuning and ensemble processes (voting/stacking) that increased prediction accuracy with corresponding cross validation scores.

High performance and relevant football insights are provided by the hybrid pipeline, which consists of unsupervised clustering → optimized ensembles → SHAP explainability. These findings connect academic research with concrete tactical application, establishing AI as a feasible decision-support tool for professional clubs.

Chapter V

Conclusion

5.1. Introduction

This chapter concludes the research on predicting football team success using AI. This chapter reflects on the overall research journey, key findings, and the broader implications of applying artificial intelligence techniques to football performance analysis. Building upon the motivation and problem context introduced in Chapter 1, the theoretical grounding established in Chapter 2, the methodological design detailed in Chapter 3, and the results discussed in Chapter 4, this chapter synthesises what the study has achieved and how it contributes to both academic understanding and practical decision-making in football analytics. The chapter also acknowledges the limitations encountered during the research and proposes directions for future work, before presenting a final closing reflection on the significance of the study.

5.2. Summary

The primary aim of this research was to investigate whether machine learning techniques could be effectively used to classify football team success and derive meaningful insights into what constitutes high performance at both player and team levels. To achieve this, the study was guided by three key research objectives:

1. To evaluate the predictive capability of supervised machine learning models for football team success.
2. To identify and interpret the most influential performance indicators contributing to success, and
3. To optimize the models by tuning appropriate hyperparameters and explore the practical applicability of AI-driven insights for football decision support.

This study successfully met its objectives through a clear, step-by-step approach. Various supervised machine learning models—including Logistic Regression, Random Forest, Gradient Boosting, and ensemble methods—were tested and compared. The analysis showed that ensemble

 doi:10.20944/preprints202605.1076.v1

techniques like voting and stacking delivered the most reliable results, proving their value for handling football's complex data patterns and directly fulfilling the first research objective.

Beyond just prediction accuracy, the second objective was achieved through smart feature selection and SHAP explainability. The Player Performance Index (PPI) offered a practical way to measure individual player impact, while SHAP revealed exactly which attacking, defensive, and consistency metrics separate winning teams. These insights match established football knowledge from the literature review, giving coaches clear, data-backed guidance.

The third research objective focused on optimisation and practical relevance. Systematic hyperparameter tuning was conducted for key supervised models, particularly Random Forest and Gradient Boosting, using grid-based search strategies. Parameters such as the number of estimators, learning rate, tree depth, and minimum samples per split were tuned to balance model complexity and generalisation. This optimisation process ensured stable performance without overfitting the limited dataset. This created a solid base for ensemble building and trustworthy predictions.

By translating model outputs into interpretable insights—such as performance tiers, team-level decision summaries, and player impact analyses—the study demonstrated how AI models can support real-world football decision-making. While the implementation was exploratory in nature, it highlighted clear pathways through which data-driven analytics can inform coaching strategies, performance monitoring, and talent evaluation. Overall, the study achieved its stated objectives and demonstrated that explainable AI can bridge the gap between predictive modelling and actionable football intelligence.

5.3. Implications

The findings of this research have several important implications for both academia and real football clubs. Academically, the study contributes to the growing body of sports analytics literature by showing how ensemble learning and explainable AI work together, prioritising transparency alongside accuracy.

From a practical standpoint, the implications are particularly relevant for football clubs, analysts, and coaching staff. The PPI and performance tier framework provide a structured way to assess and compare players across teams, positions, and competitive contexts. It could support scouting, contract evaluations, and squad rotation decisions by offering an objective complement to traditional qualitative assessments. In an era where transfer fees often exceed £100 million, the ability to project a player's impact on a team's overall success probability provides a decisive layer of financial risk management. Clubs can now move from talent-based scouting to system-fit scouting.

Additionally, the notion of a coach-facing decision support dashboard—which incorporates model predictions, SHAP explanations, and performance summaries— illustrates how AI outputs may be conveyed in a way that meets the needs of practitioners rather than just data scientists. Such lightweight and comprehensible methods could be useful in professionalizing performance analysis without requiring significant processing resources in the Malaysian football context, where analytical infrastructure is still being developed. The study underscores explainable AI's potential for fostering trust while encouraging the use of data-driven strategies within football organizations on a worldwide scale.

5.4. Limitations

Despite the solid contributions, several limitations exist that must be acknowledged. One of the primary limitations relates to dataset size and scope. The relatively small number of teams and observations-especially in the test set- sometimes made certain metrics look overly optimistic. Perfect accuracy scores should be taken cautiously as they might not hold up with bigger, more varied datasets, like with Logistic Regression in this research. The absence of statistically significant differences can also be partially attributed to the limited sample size, which reduces the power of hypothesis testing despite observable trends in cross-validation performance.

 doi:10.20944/preprints202605.1076.v1

Additionally, the study used publicly available, season-aggregated stats, which miss important in-game context like specific tactics, opponent quality, player fatigue, injuries, or mid-season changes. Football is incredibly complex with timing and situational factors that simple tables can't fully capture. A model trained on the possession-based dominance of the 2020s may not capture the shift toward high-intensity pressing or other emerging philosophies. These drawbacks establish that while AI is a powerful assistant, it must always be used in conjunction with human expertise.

In Malaysian and Asian football, access to high-resolution performance data, tracking systems and consistent standards lag far behind elite European leagues, making it harder to directly apply these methods locally. Furthermore, the binary definition of team success, while necessary for model consistency, may oversimplify the multidimensional nature of football performance, where long-term development, squad rotation, and financial sustainability also play substantial roles.

Another restriction was that the decision-support scenarios were simplified demonstrations rather than full simulations of player-team fit, where integration is multi-faceted. Finally, the study focused primarily on supervised learning models, with unsupervised methods used only as supplementary exploratory tools. Deeper unsupervised learning may yield additional insights in future work.

5.5. Recommendations and Future Work

Looking ahead, several practical next steps build on these findings and address the gaps. First, future studies must use much larger datasets across multiple leagues and seasons to make models more reliable and generalisable. The inclusion of temporal data would let researchers track performance trends and momentum shifts over matches.

Secondly, hybrid models blending supervised and unsupervised learning can be explored to uncover hidden patterns alongside predictions. More advanced ensembles or even deep learning could work well, as long as explainability remains a central focus.

From an application perspective, future work could develop a fully operational coach or analyst dashboard, integrating real-time data feeds and interactive visuals, which would provide coaches with instant tactical suggestions during games. Refining the scouting recommendation framework using realistic squad integration models and economic constraints would further strengthen its practical value. Integrating event-data (the X and Y coordinates of every pass) with our macro-level season-data would allow the model to understand not just that a team passes, but where those passes can be dangerous. These extensions would help transition the research from a proof-of-concept to a deployable football analytics system.

Building on the feature importance patterns and SHAP explanations presented in Chapter 4 (Section 4.6), this framework could be extended in the future by incorporating injury-related variables and workload indicators, such as minutes played, recovery cycles, and high-intensity actions, enabling the development of predictive models for injury risk and player fatigue management. By integrating match intensity metrics, recovery time, and historical injury records, machine learning models could support proactive squad rotation and medical decision-making. Such applications would be particularly valuable for clubs with limited squad depth, a common hurdle in many developing football leagues, like East Asian ones.

Beyond performance optimisation, the proposed analytics framework can be expanded to support fan engagement and commercial strategy through predictive storytelling and player profiling. Explainable AI outputs, such as SHAP-based feature narratives, could be adapted for media content, match previews, and fan-facing dashboards. In the Malaysian football ecosystem, where digital transparency and fan engagement is a growing priority, such tools could boost transparency, analytics-driven storytelling, and data literacy among clubs, analysts, and supporters alike.

5.6. Conclusion

In conclusion, this research set out to investigate how machine learning techniques could be applied to identify and explain the player-based performance indicators that contribute to team

 doi:10.20944/preprints202605.1076.v1

success in the English Premier League. Motivated by the growing reliance on data-driven decision-making in modern football, and guided by gaps identified in the literature, the study addressed the limitations of traditional performance analysis by proposing an interpretable and systematic analytical framework. Through a comprehensive review of prior work in Chapter 2, the study established the need for models that balance predictive accuracy with explainability—an issue that directly shaped the methodological choices outlined in Chapter 3.

The findings presented in Chapter 4 demonstrate that supervised learning models, particularly ensemble-based approaches like Random Forest and Gradient Boosting, are effective in classifying team success while maintaining robustness and interpretability. Feature selection and hyperparameter optimisation contributed to measurable performance improvements, while SHAP-based explainability enabled deeper insights into how specific player actions and attributes influence team-level outcomes. The introduction of the Player Performance Index further translated technical results into decision-support outputs that are meaningful for football stakeholders. Collectively, these results validate and met the research objectives and confirm that explainable machine learning can enhance understanding of "what makes a successful football team" beyond surface-level statistics.

Overall, this study contributes both methodologically and practically to the field of football analytics by giving clubs real tools for improving player and team performance, scouting and tactics. While acknowledging the inherent limitations of data availability and contextual complexity, the research provides a foundation upon which future work in injury prediction, tactical optimisation, fan engagement, and real-time decision support can be developed. From a Malaysian and global perspective, the study highlights the potential of applied artificial intelligence to support evidence-based coaching, scouting, and strategic planning, reinforcing the relevance of data-driven methodologies in the evolving landscape of professional football.

Acknowledgments: In the name of God, I express my deepest gratitude for his guidance and blessings throughout my academic journey. I am profoundly grateful to my supervisor, Dr. Nor Samsiah Sani, for her exceptional guidance, insightful feedback, and unwavering support. Her mentorship has been invaluable in shaping this research. I extend my heartfelt thanks to my family for their unwavering support and encouragement, especially during the challenging times. To my parents, your sacrifices, prayers, and belief in me have been my greatest motivation. Despite the hardships, your strength and resilience have been my inspiration. Thank you all.

Declaration: I hereby declare that the work in this thesis is my own, except for quotations and summaries which have been duly acknowledged.

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
API	Appilcation Programming Interface
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
EDA	Exploratory Data Analysis
EPL	English Premier League
F1 Score	F1 Performance Score (Harmonic Mean of Precision & Recall)
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FPL	Fantasy Premier League
GA	Genetic Algorithm
GBM	Gradient Boosting Machine
GPS	Global Positioning System
HMM	Hidden Markov Models
IDE	Integrated Development Environment

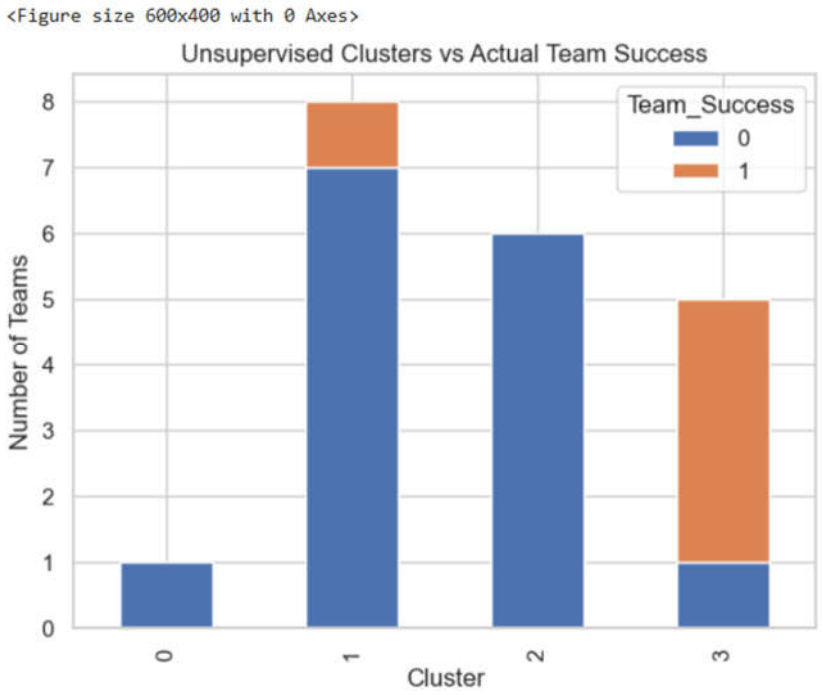
 doi:10.20944/preprints202605.1076.v1

KPI	Key Performance Indicator
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAP	Mean Average Precision
MI	Mutual Information
ML	Machine Learning
NLP	Natural Language Processing
PCA	Principal Component Analysis
PPI	Player Performance index
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines
TP	True Positive
TPR	True Positive Rate
TN	True Negative
UKM	Universiti Kebangsaan Malaysia
VAEP	Valuing Actions by Estimating Probabilities
XAI	Explainable Artificial Intelligence
xG	Expected Goals
XGBoost	eXtreme Gradient Boosting

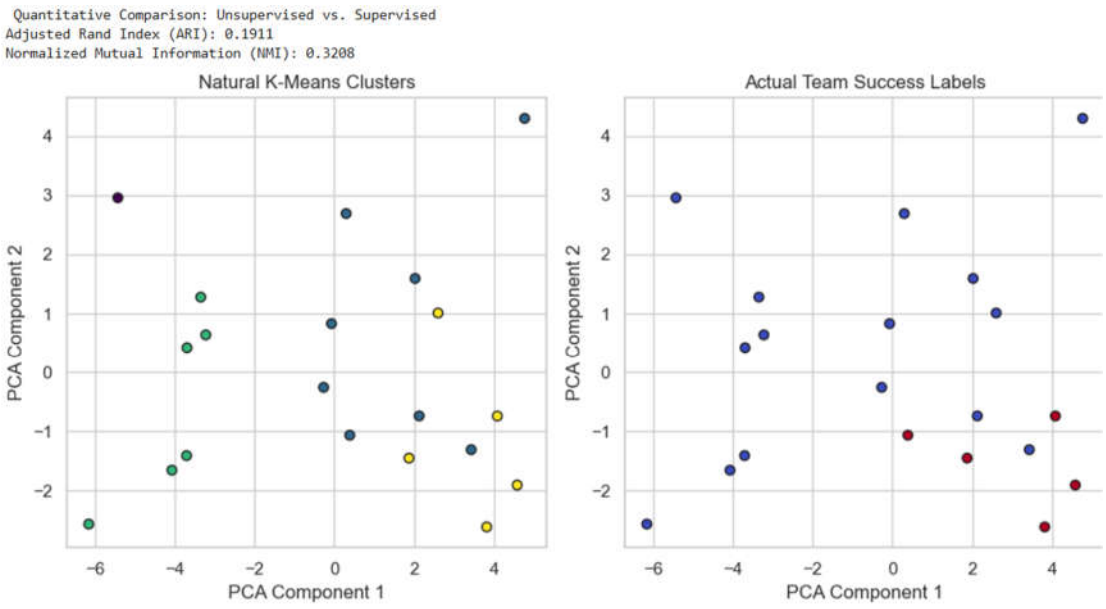
Appendix A

UNSUPERVISED LEARNING APPROACH

K-Means Clustering



 doi:10.20944/preprints202605.1076.v1



We can compress 50+ variables into a 2D plot to visualize "Success overlap". The low ARI score and visual overlap in the PCA plots indicate that 'Success' is not just about natural statistical groupings. Some teams in the same cluster fail while others succeed. This justifies the use of Supervised Learning (Ensemble Models) to explicitly learn the boundary of success rather than just data similarity.

B. Player Role and Tactical Profiling

It puts players into specific tactical roles. It's good for building a future system.

 doi:10.20944/preprints202605.1076.v1

Player Role / Tactical Profiling

Sample Player Tactical Roles:

	Name	Club	Tactical_Role_Label
0	Bernd Leno	Arsenal	Attacking Contributor
3	Héctor Bellerín	Arsenal	Attacking Contributor
4	Kieran Tierney	Arsenal	Attacking Contributor
6	Sokratis	Arsenal	Attacking Contributor
7	Rob Holding	Arsenal	Attacking Contributor
8	Shkodran Mustafi	Arsenal	Attacking Contributor
9	Calum Chambers	Arsenal	Attacking Contributor
10	David Luiz	Arsenal	Attacking Contributor
11	Sead Kolasinac	Arsenal	Attacking Contributor
12	Gabriel Magalhães	Arsenal	Creative Playmaker
13	Mesut Özil	Arsenal	Creative Playmaker
14	Lucas Torreira	Arsenal	Attacking Contributor
15	Ainsley Maitland-Niles	Arsenal	Attacking Contributor
16	Mohamed Elneny	Arsenal	Attacking Contributor
17	Joseph Willock	Arsenal	Attacking Contributor
18	Matteo Guendouzi	Arsenal	Attacking Contributor
19	Emile Smith Rowe	Arsenal	Balanced All-Rounder
20	Granit Xhaka	Arsenal	Attacking Contributor
21	Bukayo Saka	Arsenal	Attacking Contributor
22	Dani Ceballos	Arsenal	Attacking Contributor
23	Alexandre Lacazette	Arsenal	Creative Playmaker
24	Pierre-Emerick Aubameyang	Arsenal	Creative Playmaker
25	Nicolas Pépé	Arsenal	Attacking Contributor
26	Reiss Nelson	Arsenal	Attacking Contributor
27	Eddie Nketiah	Arsenal	Attacking Contributor
28	Gabriel Martinelli	Arsenal	Attacking Contributor
29	Willian	Arsenal	Attacking Contributor
30	Tom Heaton	Aston-Villa	Attacking Contributor
31	Jed Steer	Aston-Villa	Defensive Specialist
32	Ørjan Nyland	Aston-Villa	Attacking Contributor
34	Emiliano Martínez	Aston-Villa	Attacking Contributor
35	Neil Taylor	Aston-Villa	Attacking Contributor
36	Ezri Konsa Ngoyo	Aston-Villa	Attacking Contributor
37	Matt Targett	Aston-Villa	Attacking Contributor
38	Björn Engels	Aston-Villa	Attacking Contributor
39	Frédéric Guilbert	Aston-Villa	Attacking Contributor
40	Ahmed El Mohamady	Aston-Villa	Attacking Contributor
41	Kortney Hause	Aston-Villa	Attacking Contributor
42	Tyrone Mings	Aston-Villa	Attacking Contributor
43	Matthew Cash	Aston-Villa	Attacking Contributor
44	Douglas Luiz	Aston-Villa	Attacking Contributor
45	John McGinn	Aston-Villa	Attacking Contributor
46	Henri Lansbury	Aston-Villa	Attacking Contributor
47	Jack Grealish	Aston-Villa	Attacking Contributor
48	Marvelous Nakamba	Aston-Villa	Attacking Contributor
49	Conor Hourihane	Aston-Villa	Attacking Contributor
50	Trézéguet	Aston-Villa	Attacking Contributor
51	Anwar El Ghazi	Aston-Villa	Attacking Contributor
52	Jota	Aston-Villa	Attacking Contributor
54	Wesley	Aston-Villa	Attacking Contributor

Appendix B

AI SCOUTING DASHBOARD

Conceptual Mock AI Scouting Dashboard – Chelsea FC

 doi:10.20944/preprints202605.1076.v1

References

1. Alamar, B. C. (2013). *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*. Columbia University Press.
2. Alghamdi, N.S. et al. (2025) *Comparative analysis of algorithmic approaches in ensemble learning: bagging* vs. *boosting*. *Scientific Reports* 15, hlm. 15971.
3. Analytics Vidhya. (2018, August 16). *Top 12 Dimensionality Reduction Techniques*.

 doi:10.20944/preprints202605.1076.v1

4. Anderson, C., & Sally, D. (2013). *The Numbers Game: Why Everything You Know About Football is Wrong*. Penguin Books.
5. Atta 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*, Article 170.
6. Baboota, R. and H. Kaur (2019*). "Predictive analysis and modelling football results using machine learning approach for English Premier League."* International Journal of Forecasting 35(2): 741-755.
7. Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics, 37*(2), 253–264.
8. Bialkowski, A., Lucey, P., Carr, P., Denman, S., & Sridharan, S. (2014). Large-scale analysis of soccer matches using spatiotemporal tracking data. *2014 IEEE International Conference on Data Mining*, 725–730.
9. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
10. Brownlee, J. (2020). *Machine Learning Mastery with Python*. Machine Learning Mastery.
11. Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27–33.
12. Chacoma, A., 2025. Identification and optimization of high-performance passing networks in football. *Physical Review E*, *111*(4), p.044313.
13. Chandru, R., Kaushik, A., & Jaiswal, P. (2025). Enhancing basketball team strategies through predictive analytics of player performance. *Electronics, 14*(11), 2177.
14. Chatterjee, S., & Dey, L. (2022). Deep learning models for football match result prediction: An LSTM approach. *International Journal of Forecasting, 38*(2), 719–732.
15. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
16. Choudhury, S., Saha, A., & Dey, L. (2020). Performance prediction in T20 cricket using machine learning. *Procedia Computer Science, 167*, 2536–2543.
17. Constantinides, A. (2022) *Random forest algorithms for evaluating offensive efficiency in football* [Tesis Sarjana, Erasmus University].
18. Dake, D. K., Nwiah, E., Klogo, G. S., & Ativi, W. X. (2023). Instructor-assisted question classification system using machine learning algorithms with N-gram and weighting schemes. *Discover Artificial Intelligence, 3*(1), 29.
19. Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1851–1861.
20. Decroos, T. et al. (2022) *In the spotlight: A footballer ensemble of decision trees*. *Stata News* 40(3).
21. Esteves, L. M., Mendes, R. T., & Lourenço, F. J. (2021). Dimensionality reduction and feature selection in sports analytics: A comparative study. *IEEE Access, 9*, 45677–45690.
22. García, A., Pomares, H., & Rojas, I. (2020). Data-driven performance analysis in football: A systematic literature review. *Applied Sciences, 10*(3), 1075.
23. GeeksforGeeks (2025) *Gradient boosting in ML*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/ (Accessed: 18 January 2026).
24. González-Rodenas, J., López-Buesa, S., & Calabuig, F. (2020). Tactical differences between the best and the worst teams in the English Premier League 2017/18 season. *Perceptual and Motor Skills, 127*(3), 611–629.
25. Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2018). Prediction of football match outcomes: A Bayesian approach. *Statistical Modelling, 18*(5–6), 457–482.
26. Groll, A., Ley, C., & Ziplinsky, A. (2019, November 11). *Performance of Performance Indicators in Football*. ResearchGate.
27. Groll, A., Schauberger, G., & Tutz, G. (2019). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression. *Statistical Modelling, 19*(1), 55–77.
28. Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR), 50*(2), 1–34.
29. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

 doi:10.20944/preprints202605.1076.v1

30. Guyon, I. et al. (2023) *Recursive Feature Elimination*. *Yellowbrick Documentation*.
31. Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques in sport and their applications. *Journal of Sports Science, 31*(6), 555–565.
32. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
33. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
34. Hussain, M., Ghouzali, S., & Bakkoury, L. (2022). Noise in Datasets: What Are the Impacts on Classification Performance?. *Proceedings of the 9th International Conference on Information Technology and Science* (pp. 37-45). SciTePress.
35. Jati, et al. (2024) *Football Match Prediction using Random Forest Classifier*. *Journal of Applied Technologies and Innovations* 8(1).
36. Kanabar, R. (2020). *Premier League player statistics (updated daily)* [Dataset]. Kaggle.
37. Klingstedt, A. (2024). *Enhancing Fantasy Premier League Strategies through Machine Learning and Large Language Models* [Master's thesis, Uppsala University]. uu.diva-portal.
38. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI, 14*(2), 1137–1145.
39. Kyranoudis, A. E., & Metaxas, T. I. (2024). Key performance indicators predictive of success in soccer: A comprehensive analysis of the Greek Soccer League. *Journal of Functional Morphology and Kinesiology, 9*(2), 107.
40. Liu, H., Hopkins, W. G., Gómez, M. A., & Molinuevo, J. S. (2021). Interpretable machine learning in sports: Applications and recommendations. *International Journal of Sports Physiology and Performance, 16*(4), 507–515.
41. Li, Y. et al. (2025) *Improved logistic regression combined with recursive feature elimination*. *PMC* PMC12638855.
42. Lohmann, M. S., Hellwig, M., Pospiech, R., & Schiffer, T. (2022, November 11). Match performance of football teams in different competition phases. *Frontiers in Psychology, 13*.
43. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
44. Malcata, E., & Reade, J. J. (2019). Overcoming the problem of multicollinearity in sports performance data: A novel application of partial least squares correlation analysis. *PLoS ONE, 14*(1).
45. Memmert, D., Lemmink, K. A., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine, 47*(1), 1–10.
46. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
47. Molina, M., & García, J. (2020). Predicting football match outcomes with logistic regression and machine learning techniques. *Applied Sciences, 10*(2), 501.
48. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.).
49. 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.
50. Müller, C., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611–624.
51. Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data, 6*(1), 1–15.
52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
53. Perin, C., Vuillemot, R., & Fekete, J. D. (2013). SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2506–2515.
54. Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1301.
55. Pueyo, L., Calleja-González, J., Amatria, M., Alsina, A., Cos, F., & Arrizabalaga, J. (2024). 'Effect of match location on the playing style of teams managed by Pep Guardiola: FC Barcelona vs Manchester City.' *Frontiers in Psychology*, 15, 1502199.

 doi:10.20944/preprints202605.1076.v1

56. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus, 5*, 1410.
57. Ribeiro, H. V., Mukherjee, S., & Zeng, X. H. (2020). Universality in the dynamics of scoring: Evidence from football. *PLOS ONE, 15*(5), e0233384.
58. Sankaranarayanan, S., Sattar, A., & Thabtah, F. (2014). Sports analytics: A study on predictive models for football match outcomes. *Information Systems, 3*(4), 233–240.
59. Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of Twitter. *Decision Support Systems, 88*, 76–84.
60. Shrestha, D., & Mahmood, A. (2023, January). A Comparative Study of Predictive Analysis Using Machine Learning Techniques: Performance Evaluation of Manual and AutoML Algorithms. *International Journal of Computer Applications, 185*(1), 1–7.
61. Syihabuddin, A., Alfarisy, A., & Pranggono, B. (2023). Comparison of Support Vector Machine, Random Forest and XGBoost for Sentiment Analysis on Indodax. *Journal of Computer Networks, Architecture and High Performance Computing, 6*(2).
62. Szymański, P., & Kajdanowicz, T. (2017). A scikit-based Python environment for performing multi-label classification. *Journal of Machine Learning Research, 18*(1), 209–214.
63. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.
64. Van Haaren, J., & Davis, J. (2012). Predicting the next action in football using hidden Markov models. *Proceedings of the ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics*.
65. Wang, Y. et al. (2024) *Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction*. *PMC* PMC11265715.
66. Webb, G.I. & Zheng, Z. (2024) *How Ensemble Learning Balances Accuracy and Overfitting: A Bias-Variance Perspective on Tabular Data*. arXiv preprint arXiv:2512.05469.
67. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.
68. Wong, A. (2025) *A predictive analytics framework for forecasting soccer match outcomes*. *Machine Learning with Applications*.
69. Zhang, X. et al. (2025) *Predictive Analytics for NFL Pick'em Contests* [Tesis].
70. Zhao, J., Zhang, H., & Li, X. (2021). Ensemble learning for football match result prediction. *IEEE Access, 9*, 89635–89645.