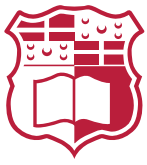
Artificial Intelligence for Team Sports

Darren Saliba

Supervisor: Dr. Charlie Abela

May 2023

Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Information Tech (Hons) - Artificial Intelligence.



L-Università ta' Malta
Faculty of Information &
Communication Technology

Abstract

Football is one of the world's most popular sports, with a massive fan base and a yearly revenue of billions of euros. Therefore, accurately predicting the outcomes of football matches has become a crucial task within the field of sports. It has always been a challenging task to predict the outcome of a football match, not only for fans but also for experts like bookmakers. There are multiple factors that can significantly influence the result, including the team's form throughout a season, weather conditions, and playing style. In this dissertation, we aim to provide a comprehensive overview of the different methods employed to predict football match outcomes through the implementation of machine learning algorithms, while also leveraging historical data. Machine learning models have proven to be highly effective in predicting the outcome of football matches since they take into account a wide range of factors. Furthermore, these models use historical data to uncover patterns and trends that can subsequently be used to make predictions. The goal of this dissertation is to predict the full-time result of a football match. A prediction can be classified into three possible outcomes: win, draw, or loss. The first step in predicting the outcome of a match is to collect and preprocess the data. The data collected focuses on the English Premier League, which is widely recognised as one of the most popular leagues in the world. The data is sourced from Football-Data, an open-source platform. In total, four machine learning algorithms are employed, Logistic Regression, Random Forest, Extreme Gradient Boosting, and Support Vector Machine. These algorithms are trained using an 80:20 ratio split. Initially, a baseline model is defined, employing manual feature selection and default parameters. The accuracies achieved of the models ranged between 49.5% and 55.5%, with the Logistic Regression model performing the best. Then, we conducted an optimisation procedure to fine-tune the parameters of the achieved models. This resulted in a 55% accuracy for the Support Vector Machine model. In the next experiment, we introduced feature selection and dimensionality reduction techniques, such as Forward Feature Selection, and Principal Component Analysis, whilst also keeping the default parameters for each model. The accuracies achieved ranged between 86% and 90%, with the top performer being the Random Forest model. Furthermore, another experiment is performed by combining these techniques with an exhaustive grid search to identify the optimal parameters for each model. The Extreme Gradient Boosting model achieved the best accuracy of 94%. Furthermore, besides accuracy, other evaluation metrics are considered to gain a more detailed understanding of the predictive performance of each model. We concluded that implementing appropriate techniques and selecting optimal parameters can significantly enhance predictive power.

Acknowledgements

My heartfelt appreciation and gratitude go to my supervisor, Dr. Charlie Abela. His invaluable guidance and support throughout this process have been instrumental in shaping my research journey and achieving the desired results. Also, his vast expertise and comprehensive understanding of the subject matter have significantly contributed to the formulation of my research ideas.

Words cannot express my appreciation towards my girlfriend for her unwavering support and reassurance throughout this journey. I would also like to express my heartfelt gratitude to my parents. Their unwavering presence has been a great source of inspiration and strength during the challenging and exhaustive days and nights of hard work. Also, their presence has been instrumental in keeping me motivated and dedicated throughout this endeavor.

Contents

List of Figures

List of Tables

List of Abbreviations

ML Machine Learning
BBC British Broadcasting Corporation
FTR Full Time Result
EPL English Premier League
ANN Artificial Neural Networks
KDD Knowledge Discovery in Databases
PCA Principal Component Analysis
FFS Forward Feature Selection
VAR Video Assistant Referee
UEFA Union of European Football Associations
IFAB International Football Association Board
OLR Ordered Logit Regression
SVM Support Vector Machine
LR Logistic Regression
XGB Extreme Gradient Boosting
RF Random Forest
FDUK Football-data.co.uk
CSV Comma-Seperated Values
MW Matchweek
HTGS Home Team Goals Scored
ATGS Away Team Goals Scored
HTGC Home Team Goals Conceded
ATGC Away Team Goals Conceded
FTHG Full Time Home Goals
FTAG Full Time Away Goals
HTP Home Team Points
ATP Away Team Points
HTGD Home Team Goal Difference
ATGD Away Team Goal Difference
DiffPts Difference in Points
TP True Positive
FP False Positive
FN False Negative
AUC Area Under Curve
ROC Receiver Operating Characteristic

1 Introduction

The ability to make accurate predictions has always been regarded as an essential human desire. Accurate predictions can be valuable in a variety of scenarios [1]. Examples of such predictions are commonly found in weather forecasts and stock market trends, where meteorologists and market analysts aim to predict future events. The goal in any scenario is to make informed decisions based on the most likely outcomes. The ability to make accurate predictions is not always possible due to the fact that the future is subject to constant change and impacted by a variety of factors. Predictions can also be made within sporting events. For example, predicting the outcome of a football match. The need for accurate football match prediction is driven by various factors. Betting companies, for example, rely on predictions to set odds and determine payouts. Fans use predictions to make informed decisions when placing bets in order to maximise their profits. This makes it an essential field of study.

Predicting the outcome of a football match is a difficult endeavor, requiring a thorough understanding of the sport. There are many factors that can influence the outcome of a football match, such as the quality of the teams, the form of the players, and the tactics used, among others [2–4]. Furthermore, external factors such as weather conditions and injuries can also have an impact on a match. A timeless example of football's unpredictable nature is when Leicester City, an emerging team, overcame all odds to win the 2015/2016 English Premier League (EPL) title over greater, more historic clubs [5]. This was a huge surprise for many fans and critics who had projected that the team would struggle. Many factors, both internal and external, contributed to this team's surprise triumph. This title victory emphasises the reality that football is a complex and unpredictable sport.

In a study conducted by J. M. Alberola and A. Garcia-Fornes [6], they shed light on the potential and capabilities of machine algorithms compared to humans. The study emphasised the limitations of human judgment, which can lead to misclassifying predictions. It intriguingly suggested that Machine Learning (ML) models are better suited and highly competitive in such scenarios. Furthermore, another research conducted by B. Fischhoff and P. Slovic [7] delved into the frequency of human error when individuals exhibit unwavering certainty in their knowledge. Surprisingly, the results uncovered a striking trend of people being consistently wrong, even in situations where they firmly believed they were right. As a result, ML algorithms possess the remarkable ability to uncover patterns and make accurate predictions by leveraging historical data. Considering there are always unpredictable events and factors that might influence the outcome of a match, it is important to approach football match prediction with caution and acknowledge that no prediction is ever 100% accurate [4].

1.1 Motivation

Football is one of the most popular sports in the world [2], and new data analysis tools can help us understand how sports teams perform and what elements influence their performance. The use of data analytics and ML technologies to forecast the outcomes of football matches is becoming increasingly popular. With the increasing access to detailed football statistics and the ability to differentiate between them, there is an opportunity to develop models that can accurately forecast match results [8].

According to Darren Small, the director of integrity at Sportradar, a prominent betting and sports data analysis company, states in an article [1] published by the British Broadcasting Corporation (BBC) in 2013 that the football betting industry is already estimated to range from 700 billion dollars to 1 trillion dollars. In addition, besides the economic opportunities, the development of such a model has the potential to make valuable contributions to the extensive field of sports analytics. By establishing a reliable prediction model, researchers can gain new and insightful understandings of the game, potentially uncovering hidden patterns and emerging trends. These insights can prove highly beneficial in assisting teams, coaches, and analysts in making informed decisions and strategic adjustments, ultimately leading to enhanced on-field performance. Moreover, by creating a dependable and accurate prediction model, researchers can provide valuable support to individuals involved in football betting, enabling them to reduce risks and make more well-informed decisions.

The development of a football match prediction model is a highly relevant and significant research area. By highlighting the growing interest in using data analytics to predict match outcomes, the potential contributions to the field of sports analytics, and the practical benefits for bettors, it establishes the importance of the research project and generates interest in its findings.

1.2 Aims and Objectives

In this dissertation, we explore the effectiveness of ML models in predicting football match results. More specifically, we aim to explore the suitability of ML techniques in predicting the full-time result (FTR) of the EPL. The result can be categorised as either a win, a draw, or a loss. The main objective of this research is to contribute to the development of reliable and efficient systems that accurately predict the outcome of football matches.

The following set of objectives have been identified:

- Objective 1 (**O1**):

[1] `https://www.bbc.com/sport/football/24354124`

  - We aim to evaluate and optimise the performance of ML models in accurately classifying the FTR of a football match. To achieve this, we will implement four ML algorithms and compare their predictive performance against each other to determine the most effective model for this task.

- Objective 2 (**O2**):

  - We also aim to build a dataset that can effectively capture the relevant information that may potentially have an effect on the FTR of a football match. To accomplish this, we aim to collect historical match data, including various match statistics such as goals scored, goals conceded, and current form, among other relevant factors.

- Objective 3 (**O3**):

  - In line with **O1**, we aim to test two feature selection approaches, manual selection, and data-driven selection combined with dimensionality reduction.

1.3 Why is the Problem Non-Trivial

Football match outcome prediction using ML is a non-trivial task due to the complexity of the sport. A match's outcome can be influenced by a variety of internal and external factors. Each match generates a large amount of data, including statistics on player and team performance, as well as other factors. Collecting and processing this vast amount of data in a meaningful way and also being able to identify the driving features that are most important is a complex task [9]. Furthermore, the relationship between input variables and output variables may not always be linear, making it difficult to develop accurate predictive models. As such, traditional linear models may not be as effective in representing the complex interactions between input and output variables. Overfitting poses another notable concern in football match prediction. It occurs when a model becomes excessively specialised on a certain dataset. Even though there are multiple techniques that can be employed to alleviate this concern, it is not always possible to eliminate such instance. Furthermore, another concern could be the class imbalance present within datasets [10]. Despite these challenges, the development of accurate predictive models for football match prediction is an important and valuable area of research that can have significant practical applications.

1.4 Proposed Solution

To address the challenge of predicting football match outcomes as a multiclassification problem, our proposed solution is to leverage ML methods and employ feature engineering and selection techniques. By utilising ML methods, we can effectively analyse large amounts of historical match data, uncovering meaningful patterns and relationships that enable accurate predictions [11]. We will be using a total of four ML algorithms, Random Forest (RF), Extreme Gradient Boosting (XGB), Logistic Regression (LR), and Support Vector Machine (SVM). These models will be tested on historical data pertaining to the EPL, gathered from football-data.co.uk (FDUK) [2]. The essence of our approach lies in employing feature engineering and feature selection techniques. Through feature engineering, we are able to create relevant and more informative features from the available data, which further enhance our model.

We have adopted two distinct approaches for our analysis. First, we employ manual feature selection, where the same set of features is chosen for all models. Initially, the models are trained using the default parameters. Then, hyperparameter optimisation is performed to maximise their performance. The second approach incorporates Principal Component Analysis (PCA), a dimensionality reduction technique, followed by Forward Feature Selection (FFS). This approach is implemented separately for each model, allowing us to tailor the process to meet the specific needs of each model. Similar to the first approach, we also perform hyperparameter tuning to further refine the models.

1.5 Document Structure

The rest of the thesis is structured as follows, in Chapter 2, we provide a thorough background research and literature review. The background research explains the techniques and other terminologies that will be used in this dissertation. The literature review, on the other hand, highlights the use of ML techniques in the field of football match predictions. Other approaches used to predict the outcome of football matches are also explored in the literature review. In Chapter 3, a detailed explanation of how the data is collected and formulated is given. Moreover, an in-depth explanation of the process taken to achieve the objectives that were set is given. In Chapter 4, we present and discuss the results of our findings. Furthermore, a detailed explanation of the obtained results from the different experiments is provided. In Chapter 5, we review the aims and objectives, and a summary of the findings is given. Furthermore, the limitations of this research, as well as recommendations for future works, are also discussed.

[2] https://www.football-data.co.uk/

2 Background and Literature Review

2.1 Overview of the Football Match

The game of football has a rich history that dates back to over a century ago. It originated in England, Europe and swiftly spread within a few years, becoming a highly popular sport on a global scale [12]. As a result, numerous leagues and tournaments were established that extended beyond Europe's borders.

In a professional setting, the game of football is contested by two teams, consisting of 11 players each. All players and members of the team are required to adhere to a set of well-established regulations. The regulations are set by the International Football Association Board (IFAB) [1]. IFAB formulates these regulations to ensure safety, fairness, and uniformity throughout the matches. The game lasts for a total of 90 minutes, split into two equal halves of 45 minutes each, with a halftime break of 15 minutes. Furthermore, an indefinite period of time is added at the end of each half, known as injury time. The added time compensates for any stoppages that may have occurred during a game. Stoppage occurrences can consist of substitutions, injuries, time-wasting, delay from Video Assistant Referee (VAR) checks, and others. Matches at a professional level are officiated by a total of four referees consisting of a referee, two assistant referees, and a fourth official.

The main objective of the game is to outscore the opponent by placing the ball within the opposing team's goalposts in order to take the win. If no team outscores the other, the game is declared a draw. A draw in a league season means that both clubs split the points, however, a winner must always be chosen in a tournament context. The majority of club competition games are played at one of the team's stadiums. The team playing in its stadium is regarded as the home team, while the other team is regarded as the away team. The main club competitions are played in a league format. A league is comprised of a predetermined number of teams that come from the same country. In this dissertation, the top league according to the Union of European Football Associations (UEFA) [2] [3] is taken into account. The league is the EPL which according to UEFA has dominated the rankings in the last decade. The EPL consists of a league structure of 20 teams. After the completion of each season, the three lowest-ranked teams are relegated to a lower league, while the top three teams from the lower league earn promotion to the EPL. Each team throughout the season faces the other on two occasions. On one occasion one team is the home team and then on the other, that team is the away team. The outcome of a match could be a win, loss, or tie. A win is worth three

[1] https://www.theifab.com/

[2] UEFA is the governing body of football in all of Europe.

[3] https://www.uefa.com/nationalassociations/uefarankings/country/

points, a tie is worth one point for each team and a loss does not award a point. The winner of the league is the team that manages to prevail over the others by accumulating the most points during a league season.

2.2 Match Events

During the course of a football match, a number of different events can occur [13]. Table 2.1 presents a subset of the most frequent and significant events that can be observed.

Event	Description
Pass	Player transfers the ball to a teammate.
Goal	The ball crosses the goal line between the goal posts.
Corner Kick	Awarded to a team when the ball crosses the opposing team's goal line outside the goal posts and is last touched by a defending player.
Free Kick	Awarded to a team when an opposing team's player commits a foul.
Penalty Kick	Awarded to a team when the opposing team commits a foul within their own penalty area.
Yellow Card	Shown to a player for committing a minor infraction. If a player receives two yellow cards, it results in a red card, and they are sent off the pitch for the remainder of the match.
Substitution	A player is replaced by another player.
Throw-In	Awarded to a team when the ball goes out of play over the touchline and is last touched by a player from the opposing team.
Offside	Occurs when a player is involved in active play and is closer to the opposing team's goal than both the ball and the second-to-last defender when the ball is played.

Table 2.1 Events frequently observed in a football match

2.3 Machine Learning Process

The ML process outlines a series of steps that must be taken to effectively develop and implement a ML model. This iterative process involves a series of interconnected tasks,

each playing a significant role. The main steps of the process consist of data collection and preprocessing, feature processing (which encompasses feature engineering and feature selection), and model training and evaluation. It is important to iterate and refine the steps as new insights emerge, fostering continuous improvement and fine-tuning of the model to achieve superior results.

2.3.1 Feature Processing

Feature Engineering

Feature engineering is an essential step in the ML pipeline that involves creating new features or transforming existing ones to extract meaningful patterns, relationships, and representations from the data, thereby enabling the models to better capture the underlying complexities of the problem at hand. The importance of feature engineering has been highlighted in prior research, namely [5]. Within this study, it has been established the quality of the results achieved is directly linked with the quality of the feature set employed. Another study conducted by Owramipur et al. [14], has also demonstrated that the quality obtained from the results is closely related to the quality of the feature set. In line with this, [5] further emphasises that one of the key aspects of their research was the process of feature engineering. While raw data often contains valuable information, the process of generating more features may provide a vaster feature set. When performed effectively, this approach has the potential to significantly enhance model performance, improve predictive accuracy, and enable the extraction of valuable insights from the data.

Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. According to research conducted by O. Gomez et al. [15], feature selection plays a crucial role in better understanding the data and improving the overall performance of predictive models. It is desirable to reduce the number of input variables to both reduce the computational cost and, in some cases, to improve the performance of the model by reducing overfitting [16]. The process of feature selection is driven by the understanding that not all features contribute equally to the prediction or classification task at hand. This technique has been applied in a considerable body of research focusing on this topic, as evidenced by the contribution of several studies [17, 18]. In their research [19], Guyon et al. provide a comprehensive analysis of the different feature selection algorithms. They categorise them into filter methods, wrapper methods, and embedded methods.

Forward feature selection (FFS) is a specific type of wrapper method used in feature selection. Wrapper methods, which rely on statistical measures, aim to rank and select the most relevant features for a given task. FFS starts with an empty feature set and gradually incorporates one feature at a time. During each iteration, the model's performance is assessed using the current subset of features, and the most valuable feature is selected based on its individual contribution to the model's predictive ability. An advantage of employing wrapper methods, including FFS, is their ability to capture feature interactions and consider the unique characteristics of the ML algorithm employed. By evaluating the model's performance at each iteration, FFS ensures that only the most relevant features are included in the final feature subset. A study conducted by A. Saifudin et al. [20] demonstrates the use of the FFS algorithm to enhance prediction accuracy and reduce computational complexity. The authors state that Forward Selection is a valuable approach to determine the influential features and improve classifier performance. They find that FFS helps in selecting the best features and enhancing the classifier's performance. However, it is important to note that wrapper methods can be computationally demanding since they require training and evaluating the model multiple times with different feature subsets. This computational expense arises due to the exhaustive search conducted by wrapper methods to identify the optimal feature subset.

2.3.2 Dimensionality Reduction

Principal Component Analysis (PCA) is often considered as a feature selection technique, although it serves a slightly different purpose compared to traditional feature selection methods. PCA is primarily used for dimensionality reduction, aiming to transform a high-dimensional dataset into a lower-dimensional space by identifying the principal components that capture the most significant variations in the data. While traditional feature selection methods focus on identifying the most relevant features, PCA focuses on capturing the underlying structure and reducing the dimensionality of the dataset. By projecting the original features onto a smaller set of principal components, PCA combines multiple correlated features into a reduced set of uncorrelated features, providing a more concise representation of the data while retaining its essential characteristics. According to a study conducted by L. Spagnol [21], PCA is widely implemented to reduce the dimensionality of data while preserving a significant portion of the variance. The author successfully retained 85% of the variance by employing PCA in their analysis. This demonstrates the effectiveness of PCA in reducing the complexity of high-dimensional datasets while retaining the crucial information needed for accurate analysis and modeling.

2.4 Literature Review

This section explores previous studies which employ a variety of approaches and techniques to forecast the outcome of football matches into the target type, as well as to examine the current status of research within this field. Making an accurate prediction of the outcome of a football match is fairly challenging, and as such, numerous studies have been carried out to determine the ideal criteria for foreseeing the outcome of football matches to be more precise and therefore enhance accuracy. According to Prasetio and Harlili's study [22], match predictions are often addressed as a classification problem in which only a single class is predicted. This class represents one of three possible outcomes: a win, a loss, or a tie.

Over the years, researchers have explored various approaches and methodologies within the field of football match outcome prediction. Some studies focused on the application of the Poisson distribution to model various aspects of the game. Other research endeavors took a different approach by employing team ratings to evaluate their performance. Additionally, certain research investigations delved into uncovering patterns and relationships by leveraging historical data through ML algorithms.

2.4.1 Poisson Distribution

In the past, a substantial amount of research has been conducted, that addressed statistical modeling and prediction in football. Research carried out by Maher [23] has analysed the development of a model based on Poisson distributions. The Poisson distribution is a statistical probability distribution that is widely used to model the number of times an event occurs in a given time or space interval. In his research, Maher introduced an attack and defense parameter for each team. The model was then developed using the values of the goals scored by both the home and away teams which follow the Poisson distributions. The method was unable to forecast football match scores or outcomes. Dixon and Coles [24] employed a similar approach in which they adopted a bivariate Poisson distribution for the number of goals scored by each team. These variables were also parameterised by features that pertain to the team's previous performances. The developed method was successful in generating match outcome probabilities. Furthermore, the model was used to devise a betting strategy that would generate positive expected returns over the bookmaker's odds. A study carried out by Rue and Salvesen [25] added the assumption that the attack and defense parameters vary over time. In order to update the estimates, they made use of Bayesian methods. Also, inferences were made through the use of Markov chain Monte Carlo iterative simulation methods. A few years later, Crowder et al. [26] successfully developed an algorithm that was discovered to be a computationally more efficient approach for updating the parameters

of the estimations. The aforementioned studies were primarily concerned with predicting the number of goals scored by a team and then using this information to assess the probabilities of a match's outcome. It is also possible to directly model the match results. This enables the development of alternative methodologies for predicting match outcomes that can improve the efficiency and accuracy of existing models.

2.4.2 Elo Rating Systems

The Elo rating system [4] is a statistical rating system used in sports to establish the relative ability level of a team. Arpad Elo, a physics professor at the time, initially developed this system in 1978. Primarily, this model was built to rank international chess players. It was later adapted to a variety of other sports, including football. Elo ratings may be employed in football match prediction to forecast the outcome of a match. It assigns each team a numerical ranking based on their previous performance and the strength of the opponents they have faced. A team with a higher rating is considered to be the stronger team and the favorite in the match. The greater the rating differential between the two teams is, the higher the chance that the stronger team will win. In the occurrence that the stronger team wins, the improvement in its rating is not as substantial as a lower-rated team winning against a higher-rated team.

C. H. Chen wrote a thesis [27] dedicated to developing an Elo rating system for the 2017 UEFA European Women's Championship. The thesis algorithm structure consists of three key components: Elo ratings for the team's strengths, ratings for player strengths, and simulation for the reliability of the ratings acquired in the previous two sections. The data used was gathered from UEFA's official website [5]. Moreover, the Elo updating formula was configured with an array of parameters. This was done in an attempt to identify the ideal combination of parameters that results in the lowest prediction error. During testing, the ideal combination was discovered. According to the simulation analysis of team Elo ratings, the estimated ratings were found to be highly reliable. On the other hand, the results of the other simulation study for estimating player strength do not appear to be as accurate. Chen provided an appropriate explanation for this. The tournament has a limited number of matches and therefore a limited number of data was available. Chen utilises the goals-based actual score function and other parameters to estimate the Elo ratings of teams. Furthermore, an ordered probit regression model based on the estimated Elo ratings was then used to predict the outcome of a match. The findings have shown that the goals-based actual score function yields better results when compared to the results-based actual score function. Chen pointed out that creating a reliable and unbiased measure for evaluating the productivity of players

[4] https://en.wikipedia.org/wiki/Elo_rating_system
[5] https://tinyurl.com/3dzue97h

is an exceedingly intricate and demanding task because only the team's overall performance can be directly observed. It is difficult to estimate the strengths of a player and their contribution to the team. This thesis has demonstrated that the Elo rating system can be implemented to measure the strengths and performances of teams and players participating in a tournament or league setting.

Arntzen and Hvattum [28] conducted a study in which they compared the performance of team ratings against individual player ratings in predicting the outcome of a football match. In other words, using ratings to predict the outcomes of matches. In this study, the Elo rating system is used to compute the individual team ratings. In total, two distinct models are used to generate the predictions. The first model is an ordered logit regression (OLR) model. This model provides the odds for the several possible outcomes of a match. There are three possible outcomes: home win, away win or draw. The Elo ratings are used by the OLR model to generate predictions. The second model relies on competing risk modeling. This model entails estimating the scoring rates of both teams. Through discrete event simulation, the estimated scoring rates are utilised to calculate the outcome of a match. The data used spans over fourteen seasons, from the 1993/1994 season to the 2007/2008 season pertaining to the top four divisions of English football. The dataset contains 30,524 matches, supplemented with odds data gathered from various bookmakers over the last eight seasons, for a total of 16,288 matches. The first two seasons are used for initial calculations of ratings, historic frequencies, and team performances. The subsequent five seasons were used for estimating the initial parameters for the OLR model. The final eight seasons were dedicated to testing. Based on the computational experiments conducted, it suggests that team ratings and player ratings perform nearly on par when it comes to predicting match results. Furthermore, when both ratings are used as covariates, the results obtained are significantly better than when the ratings are used alone. The loss functions that are applied to evaluate the prediction models are informational loss and quadratic loss. Using a combination of these two loss functions, both the accuracy of the model in predicting the correct class as well as the confidence of the predictions can be evaluated.

A study conducted by BenTaieb and Hamarneh [29] shows that combining loss functions can indeed improve generalisation of deep learning models which are a subset of ML. The study proposes three strategies for presenting the results: UNIT BET, UNIT WIN, and KELLY. The UNIT WIN strategy outperforms the others in terms of returns. This strategy places smaller bets when the odds are higher. According to Arntzen and Hvattum [28], these results could be interpreted as evidence in support of a favorite-longshot bias. This means that betting on favorites yields higher returns compared to underdogs. Furthermore, the authors emphasise the efficacy of Elo ratings in encoding information about past results, which can be considered a key takeaway.

2.4.3 Machine learning-based approaches

In recent years, there has been incredible progress in the field of ML. Its applications have spanned across a wide range of fields and as such, it has become an important tool for data analysis and interpretation, among others [30]. The literature on ML is broad and ever-changing and therefore a review of the literature on such methodologies can provide a comprehensive overview of the field's state of the art. Many experts in academia and industry have worked on the topic of football match prediction since it is both interesting and economically significant. Previous research on this topic can be classified into two categories: outcome-based studies and goal-based studies. Goal-based approaches seek to forecast the number of goals scored by each team in a given match, whereas result-based approaches seek to predict the outcome class [5].

Alfredo and Isa [31] discussed the use of tree-based model algorithms to predict football match outcomes. The algorithms employed in their research are C5.0, RF, and XGB. For their research, they acquired a dataset FDUK. The dataset consisted of over ten seasons of the EPL, which is the top English league. Furthermore, the prediction model was created using only football match statistics. In order to select the optimal attributes, feature selection is performed within the first feature set, which consists of 14 attributes. The best attributes are those that have the potential to significantly impact the prediction accuracy. The backward wrapper model is used to assist in selecting the optimal attributes. For training and testing, an 80:20 ratio split is adopted. From the results obtained, the RF algorithm produced the best results with a prediction accuracy of 68.55%. The other algorithms produced similar results. The C5.0 algorithm had an accuracy score of 64.87% while the XGB algorithm had an accuracy score of 67.89%.

Similarly, Baboota et al. [5] focused their research on developing a generalised predictive model capable of predicting the results of the EPL matches. The model's development is approached as a multi-class classification problem, with the model capable of predicting a result that falls into one of three categories: home win, away win, or draw. The data is obtained from a public United Kingdom-based source, known as Football UK [6]. Data spanning across eleven seasons is utilised, from the 2005 season to the 2016 season. Furthermore, to gather rating statistics, an online database [7] was scraped. The website includes information on an array of players from various leagues and teams throughout the world, including the EPL. The information includes their ratings, attributes, and other pertinent data. The authors also state that they were precluded from going before the 2005 season due to the limited match statistics available from those earlier seasons. The dataset was separated into nine seasons of training data and two seasons of testing data. A feature set is created using feature engineering and exploratory data analysis to

[6] http://www.bbc.com/sport/football/24354124
[7] https://www.fifaindex.com

determine the most important features for predicting the outcome of a football match, and those features are extracted to compile the feature set in order to create a highly accurate predictive system using ML. In total, four ML algorithms are employed which include, Gaussian Naive Bayes, SVM, RF, and XGB. Through the evaluation conducted it was found that the best-performing model was the XGB model, followed by the RF model. The outcomes of both models looked to be nearly identical, yet the XGB model outperformed the RF model when it came to modeling draws. Draws are considered to be the most unlikely outcome of a match. This is demonstrated by the fact that the XGB model has a higher recall score of 0.26 when compared to that of the RF which is 0.22, in the case of draws. This results in the XGB model having a higher F1 score of 0.31 than that of the RF model at 0.28. The most accurate model, which used XGB, had an accuracy of 58.5%.

Rana et al. [32] developed a Logistic Regression (LR) model to forecast the outcome of a match. In contrast to Baboota et al. [5], who approached the problem as a multi-classification problem, in this study, it is treated as a binary classification problem. The outcome of the match is either a win for the home team or a win for the away team. The data set is obtained from an online source, FDUK [8]. The data collected spans from the 2003-2004 season to the 2018-2019 season of the EPL. SVM, XGB, and LR are three ML techniques employed. These classifiers are used for primary classification and then the best out of the three is selected to predict the proper label. The authors also took into account how long it took for the model to train by keeping track of how much time it takes each model to train and generate its predictions. The model that made the most accurate predictions in the shortest amount of time is preferred. The LR model performed the best of the three, and it was subsequently utilised to forecast upcoming games. A training dataset of 4561 previously played fixtures was used to train the LR model. It only took 0.0156 seconds to train. Furthermore, the predictions were generated instantaneously, and the associated F1-Score for this model was 0.6119, with an accuracy of 65.63%. The final LR model obtained a prediction accuracy of 65.63%.

Prasetio et al. [33] created a model that predicted EPL match results using LR. The training and testing data spans six seasons from the 2010/2011 season to the 2015/2016 season. There are a total of 2,280 matches. The data has been obtained from FDUK [9]. Furthermore, statistics, such as each team's strength, are obtained from sofifa.com [10]. Sofifa.com is a website that offers data and statistics about football players, teams, and leagues. Moreover, it is updated on a regular basis to reflect the most recent changes within the football world. In total four input features are used, home offence, away offence, home defence, and away defence. The most relevant variables had

[8] https://www.football-data.co.uk/
[9] http://www.football-data.co.uk/
[10] http://sofifa.com/teams

been found to be, home defence and away defence. Prediction cannot be done using only these two variables and therefore four variables were used. The publication does not disclose how these offence and defence ratings were created. A training process flow is applied to estimate coefficients. To estimate the aforementioned coefficients a LR model is built using the Newton-Raphson algorithm. The match data was then pre-processed. Using the previously calculated regression coefficients, each match is then forecasted and compared to the actual outcome to determine predictive accuracy. An experiment is carried out in order to achieve the highest prediction accuracy. To investigate whether the split of the data has any effect on prediction accuracy, varying amounts of training and testing data are used. In total, four tests were carried out, yielding four distinct sets of model coefficients. The highest accuracy achieved was 69.5%. This was achieved by using training data from the 2010/2011 season up to the 2014/2015 season. The lowest accuracy achieved was 68.005% when the match records from the 2010/2011 season were not included in the training data of the model. The best-performing model has an accuracy of 69.5%. A binary LR model is used and therefore, this result was based on the omission of 21 draws. The acquired results showed a significant improvement over Snyder's model [34], with the prediction accuracy reaching 69.5% compared to their accuracy of 51%. Furthermore, in addition to achieving higher accuracy, the model created by Prasetio et al. uses fewer variables while still achieving a higher prediction accuracy.

Similarly, Snyder conducted a study [34] aimed at predicting the outcomes of EPL matches in the 2011/2012 season by using all matches in the 2010/2011 season. The study incorporated a variety of variables, including stadium capacity, and distance traveled by a team before a match, among others. Furthermore, statistics from the previous season, such as ranking, number of wins, draws, losses, goals conceded, goals scored, goal difference, points, and money spent on the 2011 summer transfer market are also included. LR is used to build the model. The prediction accuracy achieved is 51.06%. Synder sought to figure out what were the deciding factors in successfully predicting football matches. He determined that assessments of player performance in offence, defence, midfield, and goalkeeper were the most significant variables out of all the variables he employed.

Raju et al. [35] utilised a dataset sourced from DataHub [11]. The dataset is made up of data from five seasons, ranging from the 2014-2015 season to the 2018-2019 season of the EPL. In total, 1870 historical match data is made use of. Following data collection, extraneous attributes such as the date of the match were omitted. Ultimately, 23 features were chosen, such as home team goals scored and conceded per game at home as well as the winning percentages of both home and away teams. Raju et al. highlighted that the primary objective of this study is to produce the most significant

[11] https://datahub.io/

features through feature engineering. This is done to accurately predict the outcome of EPL matches in multi-class and binary-class instances. Five supervised ML algorithms were employed which are, support vector machine, LR, AdaBoost classifier, decision tree classifier, and naïve bayes. The LR model yielded the most accurate results with an accuracy score of 77.43% and an F1-Score of 77.5%. The model achieved an accuracy of 69.95% in multi-class and an accuracy of 77.11% in binary-class when utilising all features. On the other hand, the model achieved an accuracy of 70.27% in multi-class and an accuracy of 77.43% in binary-class using features selected through the feature selection strategy. The overall number of matches was 1870, with 861 home team victories, 565 away team victories, and 444 ties for multi-class. The overall number of matches was 1870, with 861 home team victories and 1009 away team wins for the binary class. When compared to other studies, such as the one conducted by Rana et al. [32], this model demonstrated a superior level of accuracy in binary-class prediction achieving an accuracy score of 77.43%. The accuracy of the model achieved by Rana et al. was 65.63%.

Igiri et al. [36] conducted a study in which they explored the potential of artificial neural networks (ANN) and LR techniques to predict the outcome of 110 matches in the 2014-2015 EPL season. In total, nine features that were deemed valuable were chosen at random. Through the experiments carried out the model illustrates that the LR algorithm performs better than the ANN and that the technique has a higher prediction accuracy. The proposed framework for this system is knowledge discovery in databases (KDD). A data mining tool with enhanced capabilities will be utilised to extract the players' and managers' indices and develop the model. The previous step aims to address the shortcomings of existing systems' implementation difficulties and low prediction rates. According to the results, better prediction accuracy is obtained when nine features are optimised by weighting. Initially, the accuracy of the model was 75.04% without any optimisation. However, when optimising the model by weighting, it achieved a higher accuracy of 85%. Interestingly, when utilising the ANN technique, the accuracy improved to 85%. On the other hand, the LR algorithm yielded a prediction accuracy of 93%.

Rodrigues et al. [37] conducted a study in which they sought to forecast EPL matches. The data used in this study pertain to a total of 1900 football matches spanning 5 seasons, from the 2013/2014 season to the 2018/2019 season. Upon analysing the data collected, it was found that the home team emerged victorious in 861 matches which is equal to 45.3% of the data. The away team won in 569 matches, equal to 29.9% of the data. The rest of the data shows that the matches resulted in a tie. The data was divided into training and testing sets. To ensure the credibility of the classification model, the test set comprises of all the games of a season, as team performance tends to fluctuate throughout the season. At the start of the season, teams may not perform up to

their usual standards, and towards the end of the season, they may be fatigued or have already achieved their objectives, resulting in a below-par performance. These variables can result in unforeseen outcomes, and therefore testing the classification model over an entire season is imperative. In an attempt to avoid overfitting, which would lead to poor predictive accuracy, four seasons were utilised for training purposes and a single season was used for testing purposes. The training seasons, which spanned from 2013/2015 to 2016/2017, encompassed a total of 1520 games, while the test season, 2018/2019, comprised of 380 games. To accurately compare the classification models, various metrics were utilised, such as model accuracy, among others. The SVM algorithm performed the best, achieving a success rate of 61.32%. Then, every possible combination of the 18 pre-selected variables to achieve the highest possible success rate was tested. However, due to a large number of variables, this would have resulted in a high number of combinations, making the approach unfeasible. Therefore, the most effective approach was to identify the most important variables from the 18 pre-selected options. This was done using the ”rfe” method of the ”caret” package in the R software. This algorithm starts by evaluating the relevance of the variables using all the variables provided, then proceeds through iterations in which certain variables are removed, leaving only the most essential in each iteration. Finally, the most essential variables are those used in the best-performing test. As a result, 11 attributes were identified from the initial 18, making the approach more manageable. In all situations, the models outperformed the baseline model, which had a success rate of 61.32%. The algorithms used to develop a model are SVM, RF, XGB, and RNA. The optimal model was achieved by examining combinations of 8 variables, including the 7 most relevant factors, for a total of 15 variables. The RF algorithm produced the best results, with an accuracy rate of 65.26%. The accuracy rose by a little over 4% as compared to the initial model. The reason for the increase in accuracy is due to the higher number of accurately predicted draws.

In this section, we conducted an analysis of the application of ML algorithms for predicting football math outcomes. We explored the techniques used, the experiments performed using different features, and the results obtained. Table 2.2 provides a comprehensive overview of previous research papers in this domain, showcasing the highest accuracies achieved and the specific ML algorithm employed. Furthermore, it showcases the algorithms employed by each research paper, beyond the main algorithm. The results demonstrate varying levels of accuracies achieved, ranging from 51.06% to 93%. Notably, Raju et al. [35], achieved a high accuracy of 77.43% using LR, while Igiri et al. [36] attained a more impressive accuracy score of 93% with the same algorithm. Among the various models used, LR, RF, XGB, and SVM emerge as the four most commonly employed. Of these models, LR is the most commonly utilised method, indicating its frequent application in previous research within the field. The results suggest that the choice of ML algorithm significantly impacts the accuracy achieved in the task. Further-

more, it is important to consider the specific context and dataset used in each study, as these factors can influence the performance of the algorithms.

Paper	Highest Accuracy	Algorithm	Other Algorithms
Alfredo et al. [31]	68.55%	RF	C5.0, XGB
Baboota et al. [5]	58.5%	XGB	Gaussian Naive Bayes, SVM, RF
Rana et al. [32]	65.63%	LR	SVM, XGB
Prasetio et al. [33]	69.5%	LR	-
Snyder [34]	51.06%	LR	-
Raju et al. [35]	77.43%	LR	SVM, AdaBoost, Decision Tree, Naive Bayes
Igiri et al. [36]	93%	LR	ANN
Rodrigues et al. [37]	61.32%	SVM	RF, XGB, RNA

Table 2.2 Summary of previous research using ML

2.4.4 Summary

In this chapter, we presented research that leveraged several statistical and ML methods to predict the outcome of football matches. Through our review, we discovered that ML algorithms yield more accurate findings when compared to alternative methods. Furthermore, we recognised that the selection of variables is critical in order to generate the most accurate model possible. However, the research also underlined the shortcomings of current prediction models, such as the absence of consideration for non-measurable aspects such as team morale and motivation.

Moving forward, this chapter serves as a crucial foundation for our methodology section, where we detail the specific statistical and ML methods employed. Our methodology section represents a pivotal phase in our research, where we employ specific ML methods, address feature selection challenges, and consider other techniques to contribute to the existing field of football match prediction.

3 Methodology

In this chapter, we will go through all of the procedures applied to carry out the investigations in order to achieve the research's main aim. Each choice's reasoning, as well as how the relevant data was gathered are explained. The data is also pre-processed to make it suitable for study and use as input to ML models. We explain the methods used in each experiment and offer all of the information and parameters that were established for each test.

After conducting an extensive review of the literature, we decided to implement and evaluate a total of four ML algorithms: SVM, LR, XGB, and RF. This selection was made based on the analysis of multiple research papers that explored various algorithms. Our objective was to build upon the findings of Igiri et al. [36], who achieved an outstanding accuracy score of 93% in their research. We aimed to surpass this level of performance and improve upon their work in our own study. As shown in Figure 3.1, our suggested approach follows a systematic flowchart. This flowchart depicts the complete process, from data collection through model selection and assessment, creating a solid foundation for accurate match outcome predictions based on previous data and relevant features.

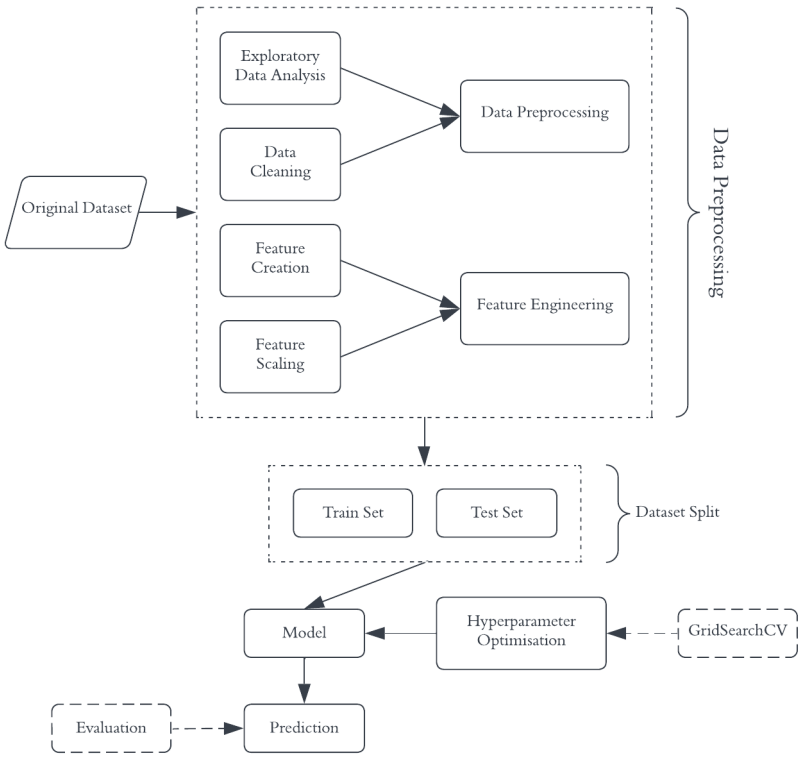


Figure 3.1 System Architecture

3.1 Football Match Dataset

This section addresses the dataset utilised for football match prediction. The data used for football match prediction was sourced from FDUK [1], an online platform. FDUK is a comprehensive and reliable repository of historical football data, encompassing vital information such as match results, player statistics, and team performance metrics. This wealth of data is essential for developing accurate prediction models. The collection consists of seasons represented as separate Comma-Seperated Values (CSV) files, where each file captures the data of individual matches throughout the year. Within each file, every row corresponds to a distinct match instance, containing comprehensive details, statistics, and results for each match. In previous research, researchers have leveraged the availability of data from FDUK to obtain the dataset needed for their research [14, 15]. Leveraging the resources available on FDUK enables us to access a wide range of information, thereby enhancing the accuracy and robustness of our football match prediction models.

The focus of our analysis revolves around the EPL, which holds a distinguished status as one of the top football leagues globally according to UEFA [2].Considering the context that different leagues have distinct play styles, team dynamics, and varying levels of competitiveness, our focused approach enhances the relevance and accuracy of the predictions. By tailoring our predictions to a specific league, we account for these unique characteristics and factors that influence match outcomes. We utilised data spanning approximately 20 seasons, starting from the 2000/2001 season. The decision to commence from this particular year stemmed from the limited availability of reliable and comprehensive data prior to that period. By selecting data from the season 2000/2001 onwards, we were able to collect a substantial amount of pertinent and detailed information.

It is fundamental to note that the 2020/2021 season has not been considered. Following a comprehensive examination of the season, it became clear that the absence of fans within the stadiums due to COVID-19 had a significantly greater impact on the FTR. Notably, away teams accounted for more wins than the home teams during the 2020/2021 season when compared to other seasons. Furthermore, considering the limited availability of data in such circumstances because of the fact that only a single season was affected by this phenomenon, it was deemed that including such data could introduce biases and affect the reliability of the models. Another critical factor we examined was the introduction of VAR in the 2019/2020 season. This revolutionary technology has significantly changed the way the game is played. To assess its potential influence on prediction models, the percentage split of match outcomes was exam-

[1] https://www.FDUK/

[2] https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2023

ined. Figure 3.2 illustrates the percentage distribution of match outcomes before the introduction of VAR, whilst Figure 3.3 displays the percentage distribution of the match outcomes after the introduction of VAR. The results displayed a relatively balanced distribution of the matches outcomes, which led us to incorporate data from both before and after the introduction of VAR into our experiments.

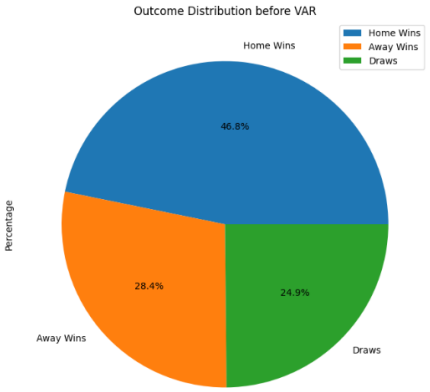


Figure 3.2 Before the introduction of VAR

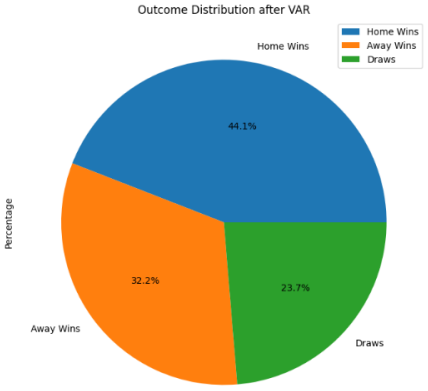


Figure 3.3 After the introduction of VAR

Before utilising this data, it was necessary to undertake a thorough process of cleaning and transforming it into a usable format. The process involved various steps aimed at ensuring the quality and reliability of the data. Once the cleaning process was completed, the data from each season was merged into a single dataset. This consolidation facilitated more efficient analysis and modeling, as it eliminated the need to work with separate datasets for each season. The resulting dataset served as a robust foundation for subsequent stages of analysis, including feature engineering, model selection, and performance evaluation.

3.2 Data Preparation and Preprocessing

3.2.1 Data Preprocessing

The data, which is initially stored in CSV format, is extracted and meticulously structured into separate pandas dataframes. By structuring the data into pandas dataframes, it becomes more accessible and amenable to subsequent data cleaning, feature engineering, and analysis. The tabular representation offered by dataframes enables us to efficiently handle the data, manipulating rows, columns, and cells with ease. The Date column in the data frame was excluded. This decision was made because the dates in the column had various formats, and instead of standardising them, a match week (MW) variable was introduced. The MW variable assigns a unique MW number to each match, enabling the analysis of performance trends and patterns over time. By examining team performance

across different MWs, it becomes possible to identify consistent patterns or fluctuations in their performance throughout the season. The MW column also enables the comparison of team performance at different stages of the season. The calculation of the MW number follows a simple logic: after every 10 matches, the MW number increases by 1. Since there are consistently 20 teams participating each year, and each match involves 2 teams, every 10 matches correspond to a new MW. This systematic approach ensures the accurate assignment of MW numbers, providing a valuable temporal context for further analysis and insights into team performance dynamics.

3.2.2 Feature Engineering

By leveraging the available data, it is possible to derive additional statistics that can significantly enhance the accuracy of predicting football matches. These statistics go beyond the basic information provided and delve deeper into the intricacies of the game, enabling a more comprehensive analysis. The implementation of the proposed methodology incorporates code snippets that were leveraged from a publicly available GitHub [3] repository, allowing for efficient utilisation of existing resources and enhancing the development process.

To facilitate the analysis of team performance, four new features are created: HTGS (Home Team Goals Scored), ATGS (Away Team Goals Scored), HTGC (Home Team Goals Conceded), and ATGC (Away Team Goals Conceded). These features capture the cumulative goals scored and conceded by each team throughout the season, providing valuable insights into their offensive and defensive capabilities. By examining these features, it becomes possible to evaluate the goal-scoring abilities, defensive strengths, and trends in scoring and conceding goals exhibited by each team. The calculation of these features involves determining the cumulative goals scored and conceded by each team for every MW. This is accomplished by computing the cumulative values and storing them in separate data frames, where each row represents a team and the columns correspond to MWs. The process entails iterating through the columns FTHG (Full Time Home Goals) and FTAG (Full Time Away Goals), extracting the goals scored and conceded values. These columns effectively represent the cumulative goals scored and conceded by the home and away teams up to a specific MW.

Moreover, two additional features are introduced: HTP (Home Team Points) and ATP (Away Team Points). These features represent the cumulative points earned by each team up until a specific MW. HTP captures the total points accumulated by the home team throughout the season, while ATP signifies the accumulated points for the away team. In this scoring system, a win is assigned 3 points, a draw receives 1 point, and

[3] https://github.com/llSourcell/Predicting_Winning_Teams/blob/master/Scraping%20and%20Cleaning.ipynb

a loss contributes 0 points. These features provide meaningful insights into the overall performance of each team and their ability to secure victories or draw matches. They offer a quantitative measure of team success and can be leveraged for various analytical purposes, such as comparing team strengths, identifying patterns in point accumulation, and evaluating the impact of points on the league standings. By incorporating HTP and ATP, a comprehensive assessment of team performance and league dynamics can be achieved.

In addition, another important aspect is the calculation of team form. This involves assessing the performance of each team based on their recent matches. Specifically, the form is evaluated for the five previous games played by each team. To capture this information, several features are created: HM1 to HM5 (representing the first home match result to the fifth home match result) and AM1 to AM5 (representing the first away match result to the fifth away match result). These features indicate the recent form of the home and away teams, respectively, based on their most recent match result (1 match ago). To obtain these features, the result of the previous match for each team is extracted and stored in a list. For example, HM1 and AM1 store the outcome of the most recent match for the home and away teams, such as 'W' for a win, 'L' for a loss, or 'D' for a draw. Similarly, HM2 and AM2 store the result of the second most recent match, and so on. By considering the past match results, these features provide insights into the immediate performance of each team. They offer a glimpse into a team's success, consistency, or momentum in their recent matches. This historical data can help assess the current state of each team, identify trends, and even make predictions about future match outcomes. Understanding the team's form is crucial for gaining a comprehensive understanding of their capabilities and evaluating their potential in upcoming matches.

Two additional features, HTFormPtsStr (home team form points) and ATFormPtsStr (away team form points), emerge from the previously constructed features. These features provide information about the recent form points of the home and away teams, respectively. To calculate these features, the code concatenates the individual match results of each team over their most recent five matches. For instance, if the home team's last five match results were 'WDLWL', the HTFormPtsStr value would be 'WDLWL', indicating a win, a draw, a loss, a win, and a loss in their last five matches. The same process is applied to the away team for the ATFormPtsStr feature. By incorporating the HTFormPtsStr and ATFormPtsStr features, valuable insights can be gained into the recent performance and form of both teams involved in a match. These features consider the sequence of match results and the total points earned by each team in their respective recent matches. They serve as indicators of the teams' success, consistency, or momentum in their gameplay.

Furthermore, the data frame undergoes a series of operations aimed at enriching

the dataset for further analysis and modeling. The initial set of operations calculates essential metrics such as goal difference, and point difference. These metrics, namely HTGD (Home Team Goal Difference), ATGD (Away Team Goal Difference), and DiffPts (point difference), provide valuable insights into team performance and historical positions. By incorporating these meaningful features, the dataset becomes more comprehensive, capturing important aspects like goal difference, point difference, and historical positions. In addition, a scaling technique is applied to several columns in the data frame based on the MW variable. This scaling process results in a normalised dataset that allows for consistent comparisons, improved model performance, and ensures fair evaluation. The columns that will undergo scaling, are 'HTGD', 'ATGD', 'DiffPts', 'DiffFormPts', 'HTP', and 'ATP'. Next, the MW column is converted to a float data type. Within a loop, each value in the specified columns is divided by the corresponding MW value. This scaling operation prevents features with large values from dominating the analysis or model training process, thereby enhancing the performance of models. Moreover, it enables fair evaluation by placing different features on a similar scale, ensuring their proportional contribution to the analysis. In summary, the performed operations on the playing statistics DataFrame, including the calculation of significant metrics and the scaling based on MW, substantially enhance the dataset for analysis and modeling purposes. These operations contribute to a more comprehensive dataset, enable consistent comparisons, improve model performance, and ensure fair evaluation of the features.

3.3 Home Advantage

The initial experiment conducted aimed to analyse the effectiveness of home advantage in football matches. As discussed in Chapter 2, many studies related to football match prediction revolve around identifying key statistics and performance indicators that correlate well with winning a football match. R. Pollard [38], remarked that home team advantage in football has been extensively researched, yet the precise causes and their impact on performance remain unclear. A team may benefit from home advantage due to the familiarity with their own stadium, pitch conditions, and surroundings, which can provide them with a sense of comfort and confidence. Additionally, the support and encouragement from their passionate home fans create an electrifying atmosphere, boosting team morale and motivation. To investigate the influence of home advantage on football match prediction, we divided the dataset into two subsets. One subset contained data from seasons with fans present in stadiums, while the other subset contained data from seasons when matches were played behind closed doors due to the COVID-19 pandemic. This division allowed for an analysis of how the absence of fans in stadiums during the COVID-19 seasons affected the home advantage phenomenon.

The subset representing the COVID-19-affected season contained data specifically from the 2020/2021 season. It is worth noting that the 2019/2020 season of the EPL faced disruptions caused by the COVID-19 pandemic. However, it is important to highlight that a significant number of matches were still able to take place with supporters in attendance before the suspension of the league occurred. Therefore, when examining seasons significantly impacted by the pandemic and where matches were played entirely without fans, our main focus shifts primarily to the 2020/2021 season. Upon observing Figure 3.5, it is evident that in the 2020/2021 season, the away team displayed a slightly higher win percentage as opposed to the home team. In contrast, all other seasons with spectators consistently demonstrated a higher win percentage for the home team. On average, across all seasons with fans present (all seasons used in the study except the 2020/2021 season), the home team recorded an average win rate significantly higher than the away team's average win rate, as can be observed in Figure 3.4. The results obtained from this experiment align with both our expectations and previous literature, confirming the prevailing notion that the home team tends to win more matches than the away team. This analysis contributes to the existing body of research by reinforcing the significance of home advantage in football and the role of fans in shaping match outcomes.

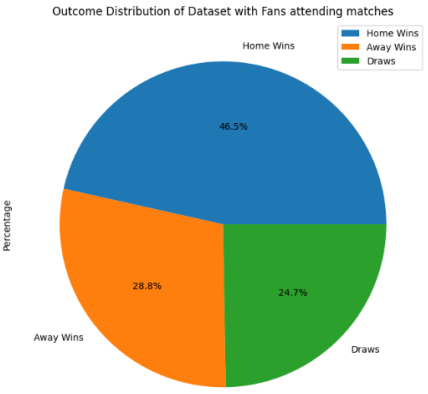


Figure 3.4 Distribution of Results with Fans

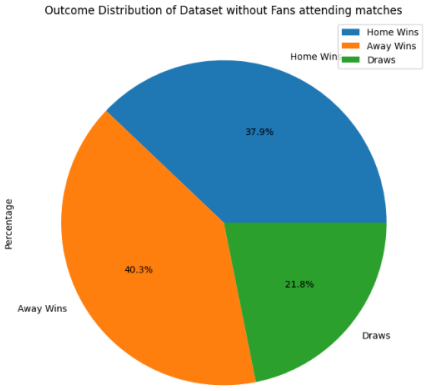


Figure 3.5 Distribution of Results without Fans

3.4 Feature Selection & Dimensionality Reduction

Two approaches, manual and data-driven, were used to identify informative and influential features that contribute significantly to the predictive task, aiming to achieve a superior model by discarding irrelevant or redundant ones.

3.4.1 Manual Feature Selection

During the manual feature selection process, a correlation matrix was utilised to identify variables that exhibited strong correlations. This approach involved visually examining the correlation matrix, which was computed by calculating correlation coefficients between all pairs of variables in the dataset. The resulting matrix was then transformed into a heatmap, which can be seen in Figure 3.6, where each cell represented the correlation between two variables. The examination of the heatmap involved identifying areas with high absolute correlation values, indicating variables that were strongly correlated with each other. Positive correlations, closer to 1, suggested that the variables increased or decreased together, while negative correlations, closer to -1, indicated an inverse relationship. Variables exhibiting high correlation coefficients were given priority during the feature selection process, as they indicated a strong relationship between them. Features that demonstrated strong correlations with the target variable were particularly emphasised, as they are likely to have a direct impact on the predicted outcome. Moreover, we conducted an examination of the presence of multi-collinearity, which refers to high correlations between variables. In such cases, we adopted the practice of selecting only one feature from a group of highly correlated ones to avoid redundancy and address potential issues associated with multi-collinearity.

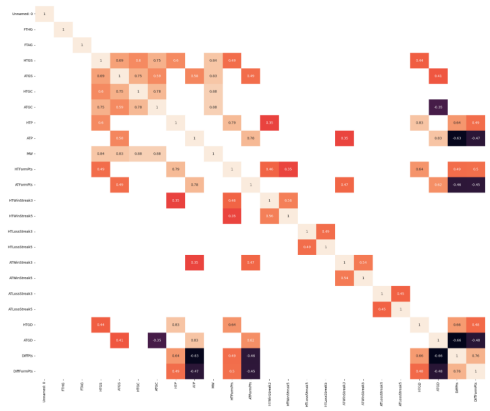


Figure 3.6 Correlation Matrix

3.4.2 Dimensionality Reduction - Principal Component Analysis

PCA is employed to reduce the dimensionality of the features and improve the performance of the classifiers. With 38 features available, PCA becomes particularly useful as it helps identify the most influential features while also reducing the complexity of the feature space. We applied PCA iteratively, starting with different numbers of features, on the training data. The transformed data was used to train a classifier, and its accuracy was assessed on both the training and testing data. This iterative process aimed to

identify the optimal number of features that resulted in the highest accuracy. We also retrieved the loadings of the principal components and determined the specific original features that contributed the most to each component. Finally, PCA was applied with the best number of features to the entire training dataset.

3.4.3 Data-driven Feature Selection - Forward Feature Selection

After applying PCA to reduce dimensionality and obtain a subset of the most suitable features, the next step involves utilising FFS. This iterative process aims to determine the optimal subset of features by progressively adding them to the model. In this case, FFS is performed on the PCA-transformed data. The 'auto' parameter is set to automatically determine the number of features, and the direction is set to 'forward', indicating that features are added sequentially. The transformed data is then fitted and transformed using FFS. The SequentialFeatureSelector methods are employed to obtain the indices and names of the selected features. Subsequently, a classifier is trained using these selected features. By combining PCA and FFS, the objective is to reduce the dimensionality of the data while retaining the most informative features for classification purposes. This approach has the potential to enhance the model's efficiency and accuracy by focusing on the most relevant aspects of the data.

3.5 Machine Learning Techniques

A goal outlined in Chapter 1 was to examine and compare the predictive performance of various ML techniques when applied to the task of classifying football match outcomes, in order to come up with a reliable model. Four ML algorithms are employed: SVM, XGB, LR, and RF. Each of these algorithms were trained and tested on the complete dataset, using an 80:20 dataset split. This involved using 80% of the data for training the models while reserving the remaining 20% for testing and evaluating the model's performance. By allowing the models to learn from a substantial amount of training data, we aimed to facilitate robust learning. Simultaneously, the testing data served as an independent set to assess the models' ability to generalise to unseen instances. This evaluation on unseen data provides valuable insights into the models' effectiveness and highlights potential areas for improvement.

3.5.1 Base Models

In the context of predicting football match outcomes, it is worth noting that a naive model that randomly guesses the result would achieve an accuracy of approximately 33.33%. This is because there are a total of three possible classes to be predicted [21].

In this study, we constructed the base model by carefully manually selecting the features through the manual feature selection process. Furthermore, initially, the models were trained using the default parameters provided by sci-kit learn. We further optimised the model by conducting an exhaustive grid search to finely tune its parameters. We carefully defined specific grids of hyper-parameters for each ML technique to be tested. By utilising the Grid Search [4] approach, our aim was to conduct a comprehensive analysis of the influence of hyper-parameters on the performance of each ML technique. Initially, our investigation aimed to answer whether the default hyper-parameters provided by sci-kit learn could produce a satisfactory predictive performance for classifying match results. Subsequently, we delved into analysing the impact of different hyper-parameters on the predictive capabilities of each ML technique. Our goal was to examine how the models' performance could be enhanced by finding the optimal set of hyper-parameters. Through manual feature selection we aimed to establish a solid foundation, a baseline model, against which we could evaluate the performance of our fine-tuned ML models. By comparing the base model with other models, we aimed to assess the incremental improvements achieved through automated feature selection and ML algorithms. Ultimately, through the base model, our objective was to surpass the predictive accuracy of a naive approach.

3.5.2 Fine-tuned model

After establishing the foundation with the base model, a combination of more relevant and effective techniques are strategically employed. These models aimed to address the limitations imposed by the manual feature selection process. To begin with, a more data-driven approach is implemented in the feature selection process. Both PCA and FFS selection are implemented for each distinct model in that order. First, through PCA we aimed to reduce the dimensionality and then through FFS, we aimed to identify the most important features. The models were developed from these identified features. The process was tailored for each model and therefore the number of features and also which features are chosen are different for each model. Additionally, we dedicated significant effort to implementing a suitable optimisation. This involved fine-tuning the parameters and selecting the appropriate ones, which may have the greatest impact, through the same process done for the base model. By incorporating these techniques, exploring alternative implementations, and conducting thorough optimisation, our primary objective was to develop an advanced model that surpasses the limitations of the base model. Ultimately, we aimed to achieve remarkable improvements in predictive accuracy and overall robustness.

[4] https://scikit-learn.org/stable/modules/grid_search.html

4 Evaluation

4.1 Base Models with Manual Feature Selection

The evaluation of the base models was carried out on a subset of the dataset, specifically 20% of the total matches, which corresponds to a sample size of 1444 matches. As previously mentioned, a naive model would typically achieve an accuracy of approximately 33% when predicting match outcomes, since the target vector would be made up of three distinctive classes. In contrast, the base models were developed with a focus on manual feature selection to improve upon this baseline performance. While accuracy is a commonly used metric for evaluating classification models, as seen in previous literature [32, 33], it may not always provide a complete picture of the model's performance. Accuracy is calculated by dividing the number of correctly predicted instances by the total number of instances, as illustrated in Equation 4.1.

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{True Positive + True Negative + False Positive + False Negative}} \tag{4.1}$$

Therefore, the F1-Score, which considers both precision and recall and is a measure of the tradeoff between these two metrics, is also taken into account. By calculating the harmonic mean of precision and recall, one can derive a score for the F1-Score, as illustrated by Equation 4.2.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.2}$$

Through a meticulous manual feature selection process, a set of 12 features were identified. These are illustrated in Table 4.1. The results obtained from the evaluation of the four ML models are displayed in Table 4.2. The displayed results, show the numerical values rounded to four decimal places. It is evident that the base models outperformed the naive model significantly, by over 20 percentage points, highlighting the effectiveness of selecting a feature set. Notably, these models exhibited a balanced accuracy between the training and testing sets, which is a highly desirable outcome. This balance indicates that the models were able to generalise well to unseen data, demonstrating their robustness and reliability. Furthermore, with there being a balanced performance on both the training and testing sets, it eliminates the possibility of overfitting within the models. Moreover, the training of these models employed the default parameters provided by scikit-learn.

Upon closer examination, it was observed that the LR model demonstrated satisfactory generalisation for both the home and away classes. A slight adjustment was

Features			
HomeTeam	AwayTeam	HTGS	ATGS
HTGC	ATGC	HTP	ATP
HTFormPts	ATFormPts	DiffPts	DiffFormPts

Table 4.1 Features Selected through Manual Feature Selection

Model	Accuracy	F1-Score
LR	0.5554	0.4838
RF	0.5014	0.4751
SVM	0.5485	0.4634
XGB	0.4952	0.4787

Table 4.2 Performance Metrics of Base Models

made to the maximum number of iterations parameter of the LR model because the default value was insufficient for the classifier to train effectively. The LR exhibited some shortcomings when it came to predicting the draw class. This observation is supported by the obtained precision scores. Similarly, the SVM, XGB, and RF models had similar shortcomings. The precision and recall values for the draw class were notably low, consequently affecting the F1-Score. Given the class imbalance in the dataset, this was anticipated.

Among the evaluated models, the LR model demonstrated the most promising results. This model achieved an accuracy of 55% and an F1-Score of 48%. When compared to the other models, this result is not significantly better, and all models exhibited a similar performance. This consistency across the models is advantageous as it provides reliable and consistent predictive performance, allowing for more confidence in the overall predictions. The worst-performing model was the XGB model. The model achieved an accuracy of 50% and an F1-Score of 48%. The obtained results were consistent with our expectations, as they reflected the inherent limitations of manual feature selection. These limitations can adversely affect the predictive capability of the models for several reasons. Moreover, the distribution of full-time match results aligned with our prior understanding, with the Home Win class representing the majority of instances, while the draw class accounted for the lowest number of instances.

4.1.1 Base model Optimisation

The objective of the subsequent experiment was to fine-tune the hyperparameters of each model and evaluate their influence on the predictive performance of each technique, using the same feature set. To achieve this, a dedicated parameter grid was established for each model. Additionally, a 5-fold cross-validation strategy was implemented

during the training process. In the appendix, a comprehensive overview of the results and optimal parameters obtained is given in Table B.1. Additionally, a summarised table, Table 4.3, is provided below for quick reference.

Model	Accuracy	F1-Score
LR	0.5429	0.4595
RF	0.5436	0.4622
SVM	0.5506	0.4683
XGB	0.5395	0.4585

Table 4.3 Performance Metrics of Optimised Base Models

The best-performing model was the SVM model based on its accuracy and F1-Score. It achieved an accuracy of 55% and an F1-score of 47%, showcasing a marginal improvement of 0.21% compared to the previous model's performance. Furthermore, the SVM model exhibited the highest precision among all of the models, achieving a precision of 0.65. This signifies that the SVM model had a lower rate of false positives (FP) compared to the other models. In contrast, the remaining models averaged a precision of 0.42, highlighting a relatively higher proportion of FPs in their predictions. This further emphasises the superior performance of the SVM model in correctly identifying positive instances. Furthermore, when evaluating recall, all of the models achieved similar results, suggesting comparable abilities to capture true positive (TP) instances. However, the SVM model's higher precision implies that it achieved a more favorable balance between precision and recall. By prioritising accuracy in positive predictions, while still maintaining a reasonable level of true positive captures, the SVM model exhibited a superior trade-off.

Despite experiencing a significant improvement of 4.22% in its performance, the XGB model persisted as the worst-performing model among all the tested models with an accuracy of 53.95% and an F1-Score of 45.85%. These findings indicate that, despite the optimisations made to its hyperparameters, the XGB model encountered inherent difficulties in achieving performance on par with the other models in the experiment. It suggests that there may be underlying factors or complexities specific to the XGB model that hinder its ability to reach comparable performance levels. With the implementation of the optimal parameters, the majority of the models demonstrated a slight improvement in performance compared to the previously defined models, with one exception being the LR model. Notably, the RF model exhibited a significant improvement of 4.22%, showcasing its potential for higher predictive accuracy. However, an intriguing observation was that the LR model actually exhibited poorer performance compared to the previously defined model. This may imply that the LR model might perform better when trained using the default parameters or that there may exist unexplored parame-

ter values that could yield better results.

The process of fine-tuning hyperparameters proved instrumental in gaining a deeper understanding of how the parameters of each model impact its predictive ability. The resulting slight improvement achieved by the optimised models underscores the significance of carefully selecting appropriate hyperparameters to maximise performance. Even though some algorithms did not display substantial changes with the optimised parameters, the experimentation process provided invaluable insights into the varying sensitivities of different models to hyperparameter adjustments.

4.2 Fine-tuned Models

In this experiment, a comprehensive and effective approach was implemented to enhance the accuracy and performance of the models. This approach involved combining PCA and FFS. Furthermore, the models were developed using the selected features derived from this approach, while keeping the default parameters. The evaluation of these models led to the generation of results, which are presented in Table 4.4.

Model	Accuracy	F1-Score
RF	0.9037	0.9024
LR	0.9965	0.9965
SVM	0.8601	0.8620
XGB	0.8843	0.8853

Table 4.4 Performance Metrics of Fine-tuned Models

The results obtained from this approach showcased significant improvements compared to the previously defined base model. This approach addressed the limitations associated with manual feature selection and allowed for a more systematic and data-driven feature selection process. A crucial aspect to highlight is that both PCA and FFS were implemented separately for each model. This approach acknowledged the fact that each model has its own specific requirements and dependencies on particular features. By customising the feature selection process for each model, we ensured that the selected features were aligned with the unique characteristics and needs of that particular model, thus improving the overall performance and predictive capabilities of the models.

The RF model has shown remarkable effectiveness in handling the task at hand. It has achieved an impressive accuracy of 90% and an F1 score of 90%. During the feature selection process, a total of 11 features were identified. The precision values further highlight the model's capability to precisely identify positive instances, with an overall precision of 0.9018. Notably, it is important to mention that the precision for

the draw class is slightly lower compared to the home and away classes, which is expected. The majority of the model's predictions align with the actual labels, showcasing its accuracy in making classifications. Moreover, when examining the confusion matrix depicted in Figure B.2, it becomes evident that most misclassifications specifically occur in the context of draw outcomes. In the confusion matrix, the values are encoded as follows: 0 denotes "away," 1 denotes "draw," and 2 denotes "home." Furthermore, the model exhibits a Standard Deviation (SD) of 0.1679, suggesting a low level of variability in its performance. As for the area under the curve (AUC) scores, both the home and away classes demonstrate exceptional discrimination power, boasting AUC scores of 0.99. However, the AUC score for the draw class is slightly lower at 0.95.

The SVM model obtained an accuracy of 86% and an F1-Score of 86%. Initially, 19 features were selected, which were then further refined to 9 features. The model achieved a precision score of 0.8658 and a recall score of 0.8601. Furthermore, it achieved a SD of 0.003745. These results suggest that the model consistently performed well on the testing data, exhibiting a reliable level of performance with relatively low variability. The AUC score for the home and away classes achieved an impressive value of 0.98. However, the AUC score for the draw class was comparatively lower, measuring at 0.82. This finding suggests that the model encountered difficulties in accurately predicting instances belonging to the draw class. A closer examination of the confusion matrix, displayed in Figure B.3, reinforces this observation, revealing that the model faced the most challenges when distinguishing draw outcomes.

With regards to the XGB model, initially, 26 features were selected, which were further refined to 13 features. The accuracy achieved by the XGB model is 88%. Furthermore, the F1-Score is also 88%. With a precision value of 0.8869, the model exhibited a high percentage of correctly predicted positive instances, while the recall value of 0.8843 emphasised its capability to identify relevant positive instances. The SD exhibited by the model is 0.1218. Additionally, the AUC score for the home class was an outstanding 0.99, indicating a high level of predictive performance. On the other hand, the draw class exhibited a slightly lower AUC score of 0.94 and the away class obtained a score of 0.98. Upon examining the confusion matrix depicted in Figure B.4, it becomes apparent that the XGB model demonstrated a higher proficiency in capturing draw instances when compared to the RF and SVM algorithms.

The LR model exhibited an impressive accuracy of 99% and an F1-Score of 99%, suggesting highly accurate predictions. This can also be observed in Figure B.1, which depicts the confusion matrix. However, such outstanding performance raises concerns about potential overfitting. The training set exhibited a SD of 0.2222, while the testing set displayed a SD of 0.2142. These relatively high SDs indicate significant variations in the model's performance across different subsets of the data, suggesting the presence of overfitting. To address the potential overfitting issue, several techniques were em-

ployed, including regularisation and the incorporation of a validation set. Despite these diligent attempts, finding an effective solution to mitigate overfitting proved to be challenging. In addition, other techniques such as undersampling and oversampling were taken into account. During the research, we came across the study of M. Heijboer [10], who highlighted that previous studies have shown that undersampling often results in a substantial loss of information, while oversampling techniques can increase the risk of overfitting. Additionally, a receiver operating characteristic curve (ROC) is plotted. The curve results in a straight line for the three distinct classes, as displayed in Figure 4.1. This suggests that the model's predictions do not exhibit any discriminatory power to differentiate between positive and negative instances. While a straight-line ROC curve does not indicate overfitting specifically, it does indicate that the model lacks predictive power and is not performing better than random chance. Furthermore, this is also backed up from the visualisation of the precision-recall curve, as depicted in Figure 4.2. It suggests that the model's predictions are random or close to random. Another explanation for the LR model's poor performance is that previous research primarily evaluated its effectiveness in binary classification scenarios. In a study conducted by Rana et al. [32], the LR model was specifically developed to predict home or away wins, disregarding the possibility of a draw outcome. Consequently, this limitation could contribute to the LR model's subpar performance.

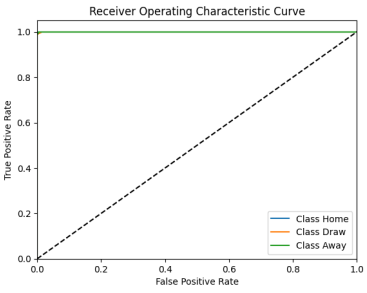


Figure 4.1 ROC Curve

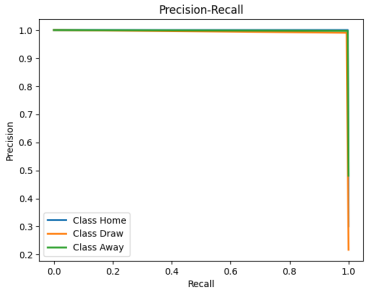


Figure 4.2
Precision-Recall Curve

The RF model emerged as the best-performing model, while the LR model performed the worst. The RF model achieved impressive results with an accuracy and F1-score of 90%, showcasing effective feature selection by identifying 11 features. The precision values indicated its ability to accurately identify positive instances. Furthermore, the RF model showcased strong discrimination power for the home and away classes with high AUC scores of 0.99 and a score of 0.95 for the draw class. However, despite the RF model's superiority, the XGB model excelled in modeling draw instances, even though it had lower accuracy and F1-Score. Similarly, Baboota et al. [5] discovered that both the XGB model and RF model yielded similar results, but the XGB model exhibited superior performance in modeling draw instances. This suggests that the XGB model could be applied effectively to model the outcomes in a better way.

4.2.1 Optimisation

Similarly, the following experiment focused on fine-tuning the hyperparameters of the previously defined models. Furthermore, a 5-fold cross-validation strategy is implemented during the training process. The table below, Table 4.5, showcases the accuracy and F1-Scores obtained by each model. The optimal parameters achieved by each model are then displayed in Table B.2.

Model	Accuracy	F1-Score
RF	0.8837	0.8851
LR	0.9938	0.9937
SVM	0.8740	0.8782
XGB	0.9440	0.9444

Table 4.5 Performance Metrics of Optimised Primary Approach

After refining the parameters of the LR model, it produced similar results. The ROC curve and Precision-Recall Curve remained unchanged, as can be seen in Figure B.9 and B.10, indicating that the LR model lacked the predictive power needed for the task at hand. Therefore, the LR model was not considered further in the evaluation. Instead, the focus shifted to the other models. The optimised RF model, resulted in an improved SD of 0.00396, indicating reduced variability in its performance. However, this improvement came at a slight cost to the model's accuracy and F1-score, which decreased to 89%. Despite this decrease, the model maintained a commendable precision of 0.89. Furthermore, it continued to exhibit strong discriminatory power with high AUC scores of 0.99 for both home and away classes and an AUC score of 0.93 for the draw class. A total of 10 features were identified. These findings highlight the continued effectiveness of the RF model despite the minor trade-off in accuracy. Upon optimising the SVM model, we observed a slight improvement in accuracy, which reached 87%, along with an F1-Score of 88%. The precision value of 0.89 indicated the model's ability to accurately classify positive instances. In total, 10 features were identified. The optimisation resulted in a higher SD of 0.010307, indicating increased variability in the model's performance across different subsets of the data. When evaluating the model's discriminatory power, we found that it achieved high AUC scores of 0.99 for the home and away classes. In terms of recall, the model achieved a value of 0.88, indicating its effectiveness in correctly identifying relevant positive instances. Specifically, the model accurately predicted 93% of instances belonging to the home class, 87% of instances belonging to the away class, and 79% of instances belonging to the draw class, as can be seen in Figure B.7.

The optimised XGB model exhibited a significant improvement in performance, achieving an impressive accuracy of 94% compared to the previous model. This signif-

icant increase of approximately 6% highlights the model's enhanced ability to correctly classify the FTR. Moreover, the SD decreased to 0.009086, indicating a higher level of stability and consistency in the predictions. A notable aspect of the XGB model's performance is its strong F1-score of 94%, which signifies a balanced combination of precision and recall across the classes. In particular, the XGB model showcased exceptional predictive power for the draw class, as evidenced by its impressive AUC score of 0.98 in the ROC curve analysis, as demonstrated in Figure 4.3. This indicates the model's ability to distinguish draw outcomes with a high degree of precision, outperforming other models in this regard. Moreover, Figure 4.4, provides additional evidence that the model managed a high percentage of correct predictions across all classes. Additionally, Figure 4.5, highlights the XGB model's ability to strike a balance between precision and recall across all of the classes.

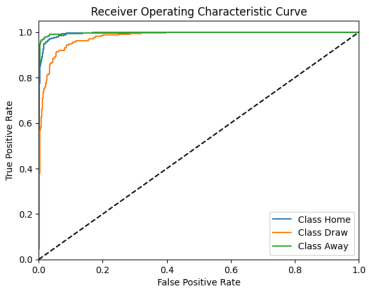


Figure 4.3 ROC Curve

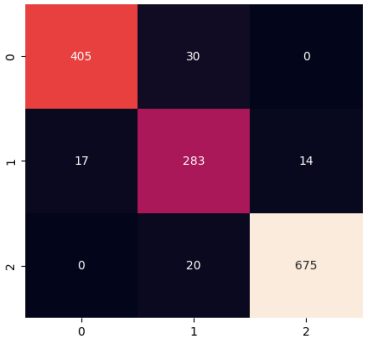


Figure 4.4 Confusion Matrix

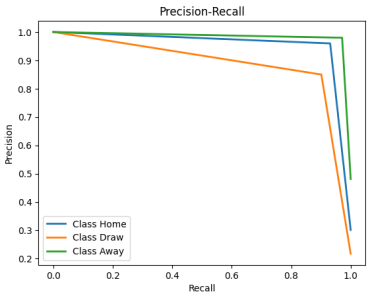


Figure 4.5 Precision-Recall Curve

4.3 Summary

The optimised XGB model emerged as the best model for the task. It achieved an impressive accuracy of 94%, outperforming other models by approximately 6%. The model's F1-score of 94% indicates a balanced combination of precision and recall. Notably, the XGB model demonstrated a reduced SD of 0.009086, highlighting its enhanced stability and consistency in predictions. Despite the presence of other models with better SDs, the combination of this model's accuracy, F1-score, and other metrics makes it more favorable and suitable for the task at hand. Furthermore, the model excelled in predicting draw outcomes with a high AUC score of 0.98, surpassing other models in this aspect. In addition, our model showed significant progress when compared to the results reported in previous research. Specifically, our optimised XGB model achieved an impressive 9% increase in accuracy compared to the LR model developed by Igiri et al. [36]. Also, it is worth noting that our model outperformed an XGB model mentioned in the literature [5] by over 20%.

5 Conclusion

5.1 Revisiting Aims and Objectives

This dissertation focused on investigating the effectiveness of ML techniques for predicting the outcomes of football matches in the EPL. The primary objective of this research was to contribute to the existing knowledge of reliable and efficient systems that can accurately predict the outcome of a football match by developing a highly accurate predictive model. The aims and objectives outlined at the beginning of this dissertation have been accomplished with success. The resulting framework demonstrates strong predictive capabilities for football match outcome prediction, providing valuable insights for decision-making in the field.

In pursuit of **O1**, our focus was to identify the most promising ML techniques that were suitable for our specific context. Thorough evaluation and careful consideration led us to identify the most suitable ML techniques, which were, LR, SVM, XGB and, RF providing a solid foundation for our subsequent analyses. Furthermore, having implemented all the necessary models, we conducted a rigorous assessment of the system's performance. Our primary objective was to evaluate its capability to successfully predict the outcomes of football matches. Through rigorous testing, optimisation, and analysis, we analysed the effectiveness and efficiency of our models, ensuring that they met the desired standards of accuracy and reliability. In pursuit of **O2**, we dedicated our efforts to the creation of a robust dataset. Through meticulous research, we thoroughly analysed and considered various publicly available datasets. Our aim was to identify the most suitable data source that encompassed essential information pertaining to the diverse factors influencing the FTR of football matches in the EPL. After careful consideration, we leveraged data from Football-Data. This particular dataset served as the fundamental building block for our research, providing us with crucial insights into the intricate dynamics shaping the outcomes of EPL matches. **O3** revolved around implementing feature engineering and feature selection techniques. To accomplish this objective, we capitalised on a dedicated repository specifically designed for the purpose of generating a multitude of features from datasets obtained from the same source. In addition, we employed methodologies, including manual feature selection, PCA for dimensionality reduction, and FFS for feature selection.

5.2 Limitations and Future Work

Although the implementation of ML techniques in predicting football match outcomes showcased promising outcomes, it is crucial to acknowledge the intrinsic limitations of

this dissertation. One of the limitations pertains to the narrow selection of features utilised in the predictive model. Football is a multifaceted sport, encompassing a wide range of variables such as team strategies, player injuries, tactical changes, and referee decisions, all of which significantly impact match results. ML models, by their nature, may face challenges in comprehensively incorporating all these factors into their predictions. Thus, it becomes imperative to explore alternative approaches that can effectively address the broader set of features and intricacies inherent in football matches.

In terms of future work, there are several intriguing avenues to explore in order to further enhance the project's effectiveness and applicability. One potential area is through leveraging deep learning architectures, such as graph neural networks, which can capture complex relationships and dependencies within the football data, thus improving prediction accuracy [39]. Incorporating explainable AI techniques, such as model-agnostic interpretability methods or rule extraction algorithms, can provide valuable insights into the factors driving the predictions, enabling stakeholders to understand the decision-making process. Another promising direction is the development of interactive and user-friendly interfaces or mobile applications, allowing users to access real-time predictions, visualise match insights, and customise prediction models according to their preferences. Furthermore, exploring transfer learning approaches, where knowledge learned from one league or season can be applied to another, can help overcome data limitations in smaller leagues or tournaments. By pursuing these future research directions, the project can significantly contribute to the advancement of ML techniques in predicting football match outcomes and provide valuable tools for decision-making in the football domain.

5.3 Final Remarks

This dissertation has successfully addressed the research objectives and provided valuable insights into the prediction of football match outcomes using ML techniques. By focusing on the specific context of the EPL, we conducted extensive research to identify the most promising ML model, namely XGB, for accurate and reliable predictions. Through leveraging the comprehensive dataset obtained from FDUK, we gained crucial insights into the intricate dynamics shaping these outcomes. The results obtained in this dissertation align with existing literature, further emphasising the suitability of ML techniques in addressing the multifaceted factors influencing the FTR of football matches. Furthermore, the findings and methodologies presented here provide a solid foundation for future research and the development of more accurate and reliable prediction models in the field of sports analytics.

References

[1] F. Petropoulos *et al.*, "Forecasting: Theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, Jul. 2022.

[2] J. Hucaljuk and A. Rakipovic, "Predicting football scores using machine learning techniques.," in *MIPRO, 2011 Proceedings of the 34th International Convention*, Jan. 2011, pp. 1623–1627.

[3] J. Lago Ballesteros and C. Peñas, "Performance in Team Sports: Identifying the Keys to Success in Soccer," *Journal of Human Kinetics*, vol. 25, Jan. 2010.

[4] A. Azeman, "Football Match Outcome Prediction by Applying Three Machine Learning Algorithms," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 1.1, pp. 73–77, Sep. 2020.

[5] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," *International Journal of Forecasting*, vol. 35, no. 2, pp. 741–755, Apr. 2019.

[6] J. M. Alberola and A. Garcia-Fornes, "Using a case-based reasoning approach for trading in sports betting markets," *Applied Intelligence*, vol. 38, no. 3, pp. 465–477, Apr. 2013.

[7] B. Fischhoff, P. Slovic, and S. Lichtenstein, "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology*, vol. 3, pp. 552–564, Nov. 1977.

[8] J. Castellano, D. Casamichana, and C. Lago, "The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams," *Journal of Human Kinetics*, vol. 31, pp. 139–147, Mar. 2012.

[9] E. Wheatcroft, "Forecasting football matches by predicting match statistics," *Journal of Sports Analytics*, vol. 7, no. 2, pp. 77–97, Jan. 2021.

[10] M. Heijboer, "Predicting Football Match Outcomes Using Machine Learning Algorithms," M.S. thesis, Tilburg University, 2022.

[11] S. Moustakidis, S. Plakias, C. Kokkotis, T. Tsatalas, and D. Tsaopoulos, "Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics," *Future Internet*, vol. 15, no. 5, p. 174, May 2023.

[12] G. Kitching, "The Origins of Football: History, Ideology and the Making of 'The People's Game'," *History Workshop Journal*, vol. 79, Apr. 2015.

[13] L. Pappalardo *et al.*, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, p. 236, Oct. 2019.

[14] F. Owramipur, P. Eskandarian, and F. Mozneb, "Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team," *International Journal of Computer Theory and Engineering*, pp. 812–815, Jan. 2013.

[15] O. Gómez, I. Cárdenas Pimentel, and L. Pacheco, *Prediction of Football Match Results Using Virtual Data*. Jun. 2023.

[16] E. Altug Çimen, "Prediction of the football match results with using machine learning algorithms," M.S. thesis, Çankaya University, 2019.

[17] R. Chauhan and H. Kaur, "Predictive Analytics and Data Mining: A Framework for Optimizing Decisions with R Tool," in *Advances in Secure Computing, Internet Services, and Applications*, Jan. 2013, pp. 73–88.

[18] Y. S. Taşpinar, İ. Çinar, and M. Koklu, "Improvement of Football Match Score Prediction by Selecting Effective Features for Italy Serie A League," *MANAS Journal of Engineering*, vol. 9, no. 1, pp. 1–9, 2021.

[19] I. Guyon and A. Elisseeff, "An Introduction to Feature Extraction," in *Feature Extraction*, vol. 207, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–25.

[20] A. Saifudin, Ekawati, Y. Yulianti, and T. Desyani, "Forward selection technique to choose the best features in prediction of student academic performance based on naïve bayes," Journal of Physics Conference Series, vol. 1477, Mar. 2020, p. 032 007.

[21] L. Spagnol, "Predicting the outcomes of football matches," M.S. thesis, University of Malta, 2022.

[22] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, Aug. 2016, pp. 1–5.

[23] M. J. Maher, "Modelling association football scores," *Statistica Neerlandica*, vol. 36, no. 3, pp. 109–118, Sep. 1982.

[24] M. J. Dixon and S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, no. 2, pp. 265–280, 1997.

[25] H. Rue and Ø. Salvesen, "Prediction and Retrospective Analysis of Soccer Matches in a League," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 49, no. 3, pp. 399–418, 2000.

[26] M. Crowder, M. Dixon, A. Ledford, and M. Robinson, "Dynamic Modelling and Prediction of English Football League Matches for Betting," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 51, no. 2, pp. 157–168, 2002.

[27] C. Heng Chen, "Elo rating system for uefa women's euro 2017," M.S. thesis, Leiden University.

[28] H. Arntzen and L. M. Hvattum, “Predicting match outcomes in association football using team ratings and player ratings,” *Statistical Modelling*, vol. 21, no. 5, pp. 449–470, Oct. 2021.

[29] A. BenTaieb and G. Hamarneh, “Uncertainty Driven Multi-loss Fully Convolutional Networks for Histopathology,” in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, vol. 10552, 2017, pp. 155–163.

[30] S. Sah, *Machine Learning: A Review of Learning Types*. Jul. 2020.

[31] Y. Alfredo and S. Isa, “Football Match Prediction with Tree Based Model Classification,” *International Journal of Intelligent Systems and Applications*, vol. 11, pp. 20–28, Jul. 2019.

[32] Sushant and D. Rana, “Premier League Match Result Prediction using Machine Learning,” Bachelor’s Thesis, Jaypee University of Information Technology, 2019.

[33] D. Prasetio and D. Harlili, “Predicting football match results with logistic regression,” in *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, Pages: 5, Aug. 2016.

[34] J. Snyder, “What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit,” Bachelor’s Thesis, May 2013.

[35] M. A. Raju, M. S. Mia, M. A. Sayed, and M. Riaz Uddin, “Predicting the Outcome of English Premier League Matches using Machine Learning,” in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dec. 2020, pp. 1–6.

[36] C. Igiri, “An Improved Prediction System for Football a Match Result,” *IOSR Journal of Engineering*, vol. 04, pp. 12–020, Dec. 2014.

[37] F. Rodrigues and Â. Pinto, “Prediction of football match results with Machine Learning,” *Procedia Computer Science*, International Conference on Industry Sciences and Computer Science Innovation, vol. 204, pp. 463–470, Jan. 2022.

[38] R. Pollard, “Home Advantage in Football: A Current Review of an Unsolved Puzzle,” *The Open Sports Sciences Journal*, vol. 1, Jun. 2008.

[39] A. Mirzaei, “Sports Match Outcome Prediction with Graph Representation Learning,” M.S. thesis, Isfahan University, 2022.

[40] X. Zou, Y. Hu, Z. Tian, and K. Shen, “Logistic Regression Model Optimization and Case Analysis,” in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019, pp. 135–139.

Appendix A Background Research

A.1 Machine Learning Algorithms

Within this section, an analysis of multiple ML algorithms has been done with the objective of identifying the various strengths and weaknesses of each algorithm in different scenarios. Understanding the strengths and weaknesses of the algorithms will aid in making informed decisions about which algorithm to use and also how to configure them to get the optimal performance.

A.1.1 Logistic Regression

LR [1] is a statistical analysis model used to forecast a dichotomous outcome. A dichotomous outcome can be represented as a yes or no response, 1 or 0. LR can be used to solve both binary and multi-class classification problems. Whereas the former yields a single outcome variable, the latter yields multiple outcome variables, each of which indicates a distinct class. LR is widely employed as a broad data processing method for binary classification and prediction tasks [40]. The objective of LR is to estimate the likelihood of an event occurring based on a set of predictors. It employs a logistic function, also referred to as the sigmoid function. This function is used to estimate the likelihood of the outcome variable having the value 1 by converting a real-valued number to a value in the range of 0 to 1. Its equation is:

$$S(x) = \frac{1}{(1 + e^{-x})} \tag{A.1}$$

S(x) is the likelihood of the dependent variable having a value of one. The variable x is a linear combination of the independent variables and their coefficients. LR provides various advantages over other classification methods, including simplicity, interpretability, and speed. Also, LR has the ability to model complex relationships between predictors and outcome variables. Moreover, it does also have disadvantages, such as the assumption of linearity between predictors and the logit of the outcome variable. The logit of the probability is modeled as a linear combination of the independent variables. The logistic function is then applied to the result to provide the projected probability.

[1] https://sparkbyexamples.com/machine-learning/logistic-regression-explained-with-examples/

A.1.2 Support Vector Machine

Support Vector Machine (SVM) [2] is a ML algorithm that is often applied in classification and regression tasks. It is used to analyse data in order to identify patterns and correlations between variables. SVM's fundamental notion is to identify the optimal boundary between multiple data points or classes. For binary classification, SVM aims to locate a boundary that partitions the data points into two groups or classes. This boundary is commonly known as the hyperplane. The equation of the hyperplane is given by:

$$w * x + b = 0 \tag{A.2}$$

The variable w denotes the weight vector perpendicular to the hyperplane, x denotes the input vector, and b denotes the bias term that shifts the hyperplane. SVM seeks the ideal values of w and b that reduce the classification error and maximise the margin. The margin can be defined as the distance between the hyperplane and the closest data points from either class. SVM is effective given that it can work with large, complex datasets. Also, it is adept at dealing with noisy data and can even provide accurate results, when the data is variable.

A.1.3 Random Forest

Random Forest (RF) [3] is used to address classification and regression problems. It is an ensemble learning approach, which implies that it aggregates the results of numerous individual models in an attempt to improve the accuracy of predictions. It is made up of multiple smaller decision trees that work together to create more accurate predictions. Each tree functions similarly to a flowchart, with each decision dependent on a set of input variables. RF works by building different decision trees using a random subset of the input variables and a random subset of the training data. Each sub-tree is unique, but they all collaborate to make predictions. When the algorithm is given a new input, each sub-tree makes its own prediction. The final prediction is then based on the sub-trees majority vote. This approach mitigates overfitting, a scenario whereby a model becomes too complex and performs excellently on the training data but poorly on unseen data. More accurate predictions on unseen data can be made by using separate trees that have been trained on different subsets of the data. Creating a single tree may not be accurate enough as it may fail to capture the important features and variable correlations. This is where RF comes into context. RF provides high accuracy and is also efficient when working with large data sets.

[2] https://shorturl.at/ayXZ4

[3] https://www.mygreatlearning.com/blog/random-forest-algorithm/

A.1.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) [4] is an ensemble learning method that enhances the accuracy of a model by combining the predictions of multiple decision trees. XGB iteratively constructs decision trees, with each new tree trained to rectify the mistakes of the previous tree. This is known as gradient boosting. It enables XGB to generate models that are extremely accurate in predicting the outcome. During each iteration, XGB computes the gradients and Hessians [5] of the loss function, with respect to the predicted values of the previous decision tree. The gradients and Hessians are then used as weights to train the subsequent decision tree, using gradient boosting. This approach allows XGB to learn complex correlations between input features and output variables while preventing overfitting. The final prediction is a weighted average of all the decision trees' predictions. XGB is a powerful algorithm that combines the benefits of gradient boosting and decision trees. It is capable of producing accurate and efficient models. One of the advantages of XGB is its ability to manage missing data and outliers. Additionally, it incorporates several other features, such as early stopping and cross-validation. These features aid in improving the accuracy and efficiency of a model. In summary, XGB is a versatile and effective ML algorithm, capable of producing accurate and efficient models. Furthermore, it is suitable for large datasets with intricate relationships between the input features and the output variables.

A.2 Evaluation Metrics

It is essential to evaluate the model's performance in order to determine whether it is appropriate for the intended use case. There are several measures for evaluating performance, and identifying the appropriate evaluation metric is vital. It is vital since each measure has its own strengths and limitations, and using an unsuitable metric may have an impact on the final model's performance and suitability for the intended task.

A.2.1 Accuracy

Accuracy is the measure of the effectiveness of a ML model. It can be defined as the number of correct predictions made by the model divided by the total amount of input data.

[4] https://shorturl.at/eBCK4
[5] https://shorturl.at/wDFJW

A.2.2 Precision

Precision quantifies the number of correct positive predictions made by the model, also known as TP, in comparison to the total number of positive predictions made by the model. It is useful in order to minimise false positives. False positives happen when the model incorrectly predicts a good outcome for a negative event. Precision is represented by EquationA.3

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{A.3}$$

A.2.3 Recall

Recall measures the number of TP made by the model in comparison to the total number of positive occurrences found within the dataset. It is advantageous to reduce the number of false negatives (FN) recognised by the model, especially in imbalanced datasets with an uneven class distribution. Reducing the number of FN would increase the model's ability to correctly classify positive cases, increasing the model's sensitivity. Furthermore, in order to provide a more comprehensive evaluation, it should be utilised in conjunction with other measures such as precision. Recall is represented by Equation A.4

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{A.4}$$

A.2.4 F1-Score

F1-Score combines both precision and recall metrics into a single metric. When compared to using accuracy, recall, or precision alone, this evaluation metric provides a more balanced and unified evaluation and comprehension of the final model's performance.

Appendix B Results

B.1 Results of Base Model Optimisation

Model	Parameters	Accuracy	Precision	Recall	F1-Score
LR	'C' = 0.001 'class_weight' = None 'fit_intercept' = True 'multi_class' = ovr 'penalty' = l1 'solver' = liblinear 'max_iter' = 10000	0.5429	0.4287	0.5429	0.4595
RF	'class_weight' = None 'criterion' = gini 'max_depth' = 5 'max_features' = sqrt 'min_samples_leaf' = 1 'min_samples_split' = 5 'n_estimators' = 300 'random_state' = 42	0.5436	0.4246	0.5436	0.4622
SVM	'C' = 0.01 'class_weight' = None 'gamma' = scale 'kernel' = linear 'shrinking' = False 'random_state' = 42	0.5506	0.6487	0.5510	0.4683
XGB	'colsample_bytree' = 0.8 'gamma' = 0.3 'lambda' = 1.0 'learning_rate' = 0.01 'max_depth' = 3 'n_estimators' = 300 'subsample' = 1.0 'seed' = 42	0.5395	0.4200	0.5395	0.4585

Table B.1 Base Model Optimisation Results

B.2 Confusion Matrices of Fine-tuned Models

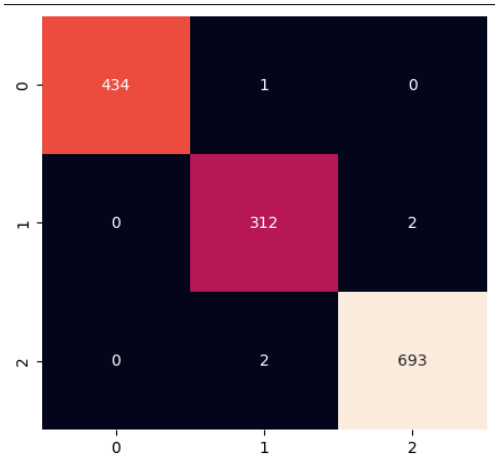


Figure B.1 LR

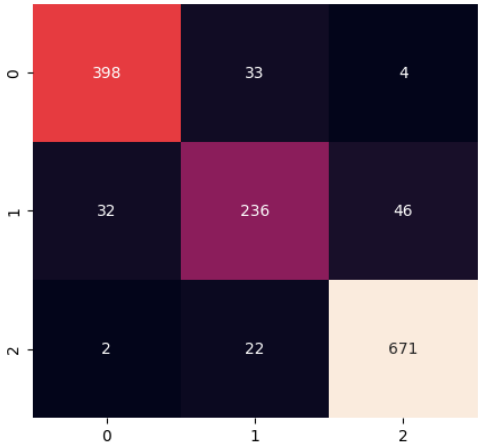


Figure B.2 RF

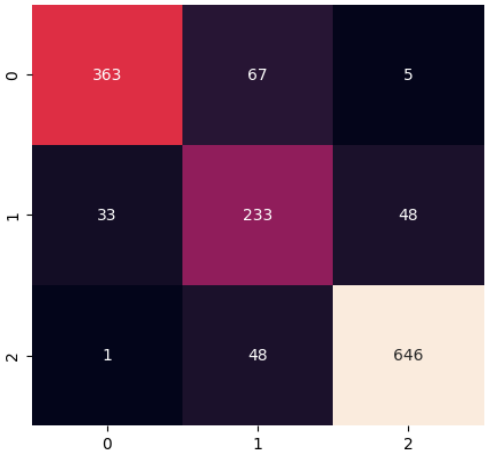


Figure B.3 SVM

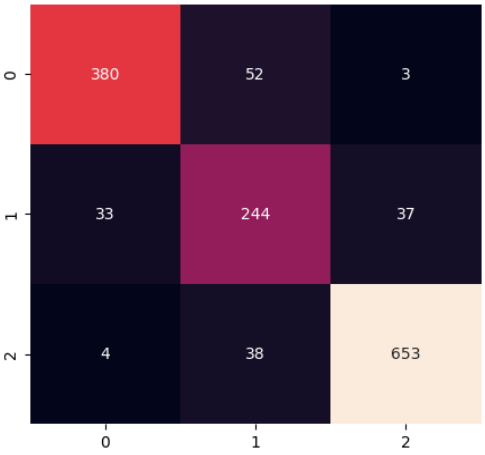


Figure B.4 XGB

B.3 Results of Fine-tuned models Optimisation

Model	Parameters	Accuracy	Precision	Recall	F1-Score
LR	'C' = 0.001 'class_weight' = None 'fit_intercept' = True 'multi_class' = ovr 'penalty' = l1 'solver' = liblinear 'max_iter' = 10000	0.9938	0.9938	0.9938	0.9937
RF	'class_weight' = None 'criterion' = gini 'max_depth' = 5 'max_features' = sqrt 'min_samples_leaf' = 1 'min_samples_split' = 5 'n_estimators' = 300 'random_state' = 42	0.8837	0.8872	0.8837	0.8851
SVM	'C' = 0.01 'class_weight' = None 'gamma' = scale 'kernel' = linear 'shrinking' = False 'random_state' = 42	0.8740	0.8881	0.8740	0.8782
XGB	'colsample_bytree' = 0.8 'gamma' = 0.3 'lambda' = 1.0 'learning_rate' = 0.01 'max_depth' = 3 'n_estimators' = 300 'subsample' = 1.0 'seed' = 42	0.9439	0.9454	0.0 9439	0.9444

Table B.2 Fine-tuned models Optimisation Results

B.4 Confusion Matrices of Optimised Fine-tuned models

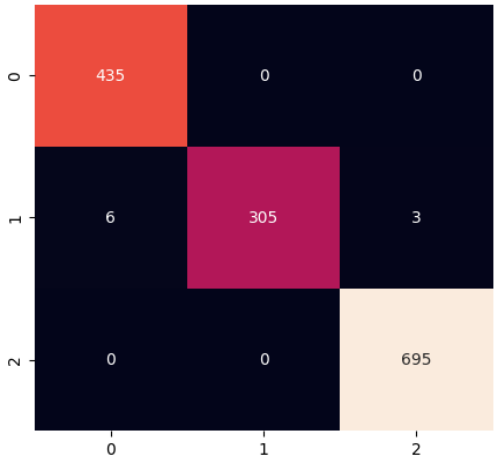


Figure B.5 LR

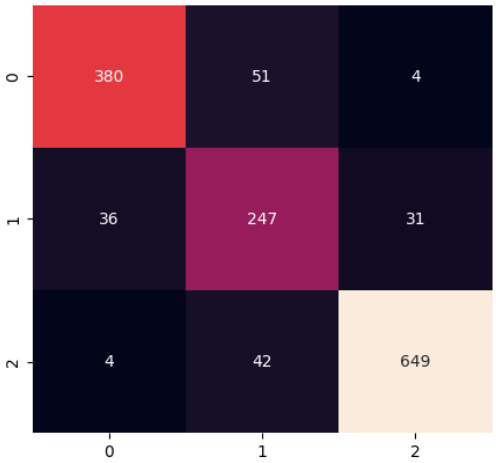


Figure B.6 RF

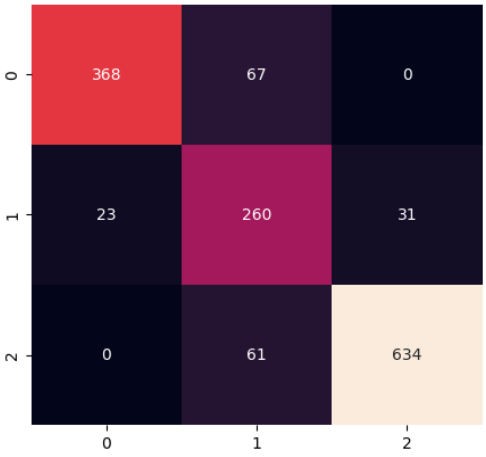


Figure B.7 SVM

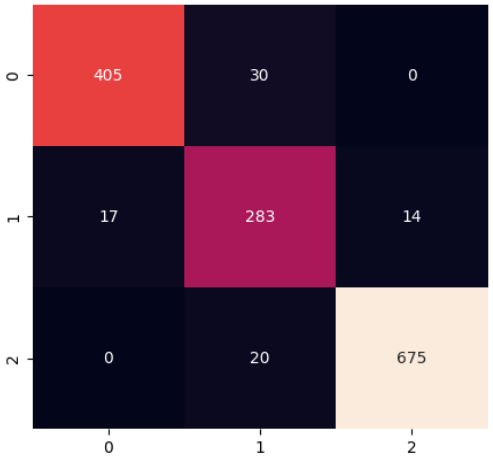


Figure B.8 XGB

B.5 Optimised LR model

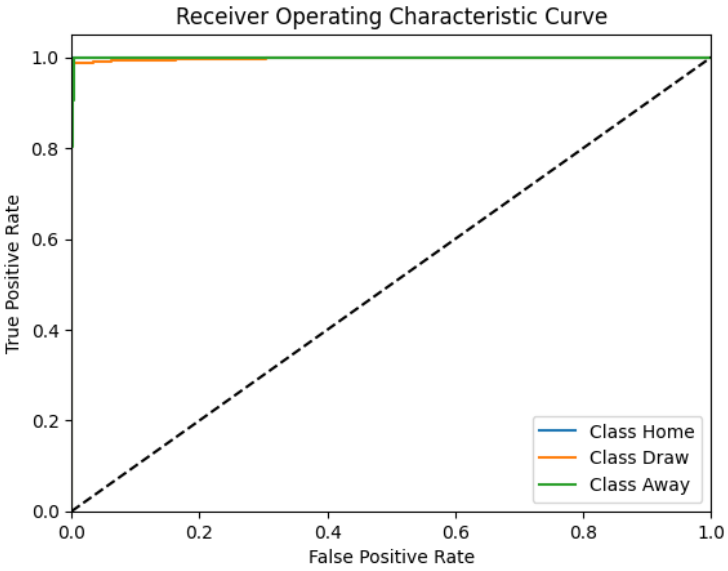


Figure B.9 Receiver Operating Characteristic Curve

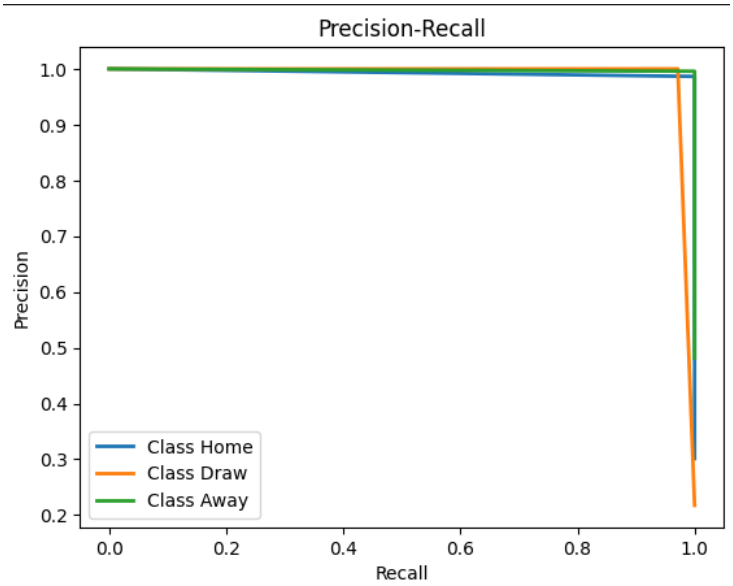


Figure B.10 Precision-Recall Curve