National College of Ireland

Exploring the Impact of Artificial Intelligence in Predicting English Premier League Football Matches

MSc Research Project
Data Analytics

Taiwo Mubarak Oladapo
Student ID: 22107312

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Taiwo Mubarak Oladapo
Student ID:	22107312
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr Catherine Mulwa
Submission Due Date:	31/01/2024
Project Title:	Exploring the Impact of Artificial Intelligence in Predicting English Premier League Football Matches
Word Count:	11731
Page Count:	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	TAIWO M. OLADAPO
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploring the Impact of Artificial Intelligence in Predicting English Premier League Football Matches

Taiwo Mubarak Oladapo
22107312

Abstract

Sports analytics, particularly in football match predictions, faces challenges due to the complex structure of the game. The variety of factors impacting match results is frequently too complicated for conventional methods to completely understand. By utilizing artificial intelligence (AI) techniques to improve prediction accuracy, this investigation seeks to close this gap. The significance lies in its potential to completely change how the sports sector makes strategic decisions across three distinct classes (home win, away win, and draw) and in identifying complex patterns and correlations in football match data. Accurate match predictions are advantageous not just to sports fans but also to sports betting markets, team management, and the whole sports analytics field. By utilizing five different machine learning models: Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), the study presents a thorough methodology. The novel aspect is how these models are compared with five metrics including accuracy, precision, f1-score, recall, and confusion matrix, with an emphasis on handling the complexity included in EPL match predictions. With an accuracy of 49.6%, XGBoost was the most effective model implemented. This demonstrates how AI can predict the results of EPL matches and how machine learning has the potential to perform better than more conventional techniques. While this research contributes to our understanding of artificial intelligence in sports analytics, certain issues remain, such as the exploration of real-time data integration, implementation of other algorithms, feature engineering, model building and the ongoing optimization required to improve multi-class prediction accuracy.

1 Introduction

As an area of research that was established in the 1950s, artificial intelligence (AI) is referred to as the capacity of a system to effectively absorb and learn from external data as well as incorporate the learning outputs to attain defined objectives and solve challenges through modification Kaplan and Haenlein (2019). It is part of a discipline in computer science that investigates how computers can be programmed to acquire knowledge, process information, and understand Ding (2019). Sports analytics have evolved tremendously under the influence of artificial intelligence (AI). Football game outcomes, or who will win the game, attract the interest of both academics and sports fans equally. The EPL is one of the largest and most watched football leagues in the world, with over 4.7 billion people watching these matches internationally. Due to its widespread popularity, many people are curious to share their predictions using a variety of methods

before the start of each match Raju et al. (2020). Using machine learning algorithms, data analysis techniques, and predictive modelling, AI has been demonstrated to be beneficial in various areas, including sports. Understanding how AI influences football match predictions can provide a crucial understanding of the benefits and limitations of these techniques in the realm of sports. Because it is a highly competitive league with a significant fan base, the English Premier League competition offers a great framework for testing whether AI approaches are effective at anticipating match results Fialho et al. (2019). Football match results are challenging to predict since so many factors influence the outcome, including player statistics, team dynamics, injuries, stadium conditions, and even weather conditions. To create intelligent AI prediction models, it is necessary to understand how each of these factors interacts to determine a game's winner.

1.1 Research Question, Objectives, and Contribution

The research question for this research project is: To what extent would artificial intelligence (AI), particularly machine learning models, help predict EPL matches? To address this research question, the following specific sets of research objectives were derived:

Objective 1- Literature Review: An extensive literature review will be conducted to enable an understanding of the current state of machine learning techniques used in sports analytics and football match prediction. With the help of this review, a basic understanding of the field is gained, as well as the methods and techniques that have been used in related research.

Objective 2- Data Pre-processing: The data would undergo preprocessing to ensure it is suitable for analysis. The dataset would be loaded into the coding environment, and missing values would be handled carefully. The data would also be split into training and testing sets. The model would use several variables, including team performance and historical data, to improve the accuracy of match predictions.

Objective 3- Implementation of AI Football Premier League Prediction models: Five Machine learning models will be applied to the Premier League match data. The machine learning models are listed below:

Objective 3.1- Random Forest

Objective 3.2- Logistic Regression

Objective 3.3- Decision Tree

Objective 3.4- Support Vector Machine (SVM)

Objective 3.5- Extreme Gradient Boosting (XGBoost)

Objective 4- Evaluation of the impact of AI model predictions on the markets for sports betting: Different evaluation metrics will be used to assess the developed models. The metrics include accuracy, precision, F1-score, recall, and confusion matrix. Prediction can influence betting patterns and odds variations, potentially altering the sports betting market. Both sports fans and industry stakeholders must understand this market effect since it can guide strategic and investment decisions.

Objective 5- Comparison of developed models as regards (Objective 3). The 5 models would be compared with each other to determine the best model that provides high accuracy in predicting English Premier League football matches.

Objective 6- Hyperparameter Tuning: Hyperparameter tuning will be performed on the best-performing model to optimize the model's performance. They would include the learning rate, sub-sample, n estimators, and max depth.

Objective 7- Comparison of developed models (Objective 4) against existing models:

The newly developed model would be compared against existing models to determine if there has been an improvement in the prediction of English Premier League football matches. This comparison would help identify the critical challenges in the football industry and sports analytics.

The primary contribution of this research is an approach that compares five classification machine learning models to predict Premier League match outcomes as either a win, draw, or loss. This research can assist betting companies and individuals in predicting match outcomes before the start of the game. Additionally, it aims to enlighten individuals about sports, contributing to the growth of the sports analytics field.

This document is organized as follows: The related work is discussed in Section 2. It contains earlier studies about this project, together with their conclusions, comparisons, and potential contributions to the body of knowledge. Knowledge Discovery in Databases (KDD) is the study approach that is presented in Section 3. It entails choosing the right models, pre-processing, transformation, and selecting appropriate AI algorithms. Section 4 describes the design specification. It includes the methods that have been used, like algorithms and model architecture. Section 5 presents the implementation. It includes the outputs that were produced, such as the data transformation, codes, etc. Section 6 discusses the evaluation. It comprises the result analysis and major inquiry findings, and Section 7 concludes the research project and discusses its future work.

2 Related Work

The literature study will go over what previous researchers have found out about using artificial intelligence (AI) to forecast results in English Premier League (EPL) matches. This research will be related to past work results, and this section will discuss the analysis. It will help identify any information gaps that this study can fill. The earlier study that is covered in this section will help to better comprehend the topic and provide insightful information for this research.

2.1 A Critique of Predictive Models in the English Premier League

By looking into how intelligent computing methods and machine learning techniques like Support Vector Machines (SVM), Decision Trees, and Random Forests can be used to predict soccer match outcomes, Madan et al. (2022) discussed the application of intelligent computing to sports administration as a means of resolving issues related to strategic planning and game tactics. To predict match results, the study selected six key variables from a dataset of 380 English Premier League soccer matches. Random Forest achieved the greatest accuracy score of 0.7631 in the trials, outperforming SVM and Decision Tree. The study concluded that decision-making and sports administration can be enhanced by intelligent computer techniques. Despite the fundamental unpredictability of sports, the level of accuracy reached in predicting soccer match results is remarkable. They suggested extending their strategy to other sports to improve outcome estimation.

By utilizing a logistic regression model, Rana et al. (2019) determined the probability of the home team winning an English Premier League football match. They initially categorized the match data using SVM, XGBoost, and logistic regression; the best three

method were then chosen to get the final forecast. Utilizing actual data from several football teams collected in the 2003–04 and 2018–19 seasons, they created a model that was 65.63% accurate in predicting match outcomes.

The English Premier League's (EPL) 2014–2015 season's full dataset was studied by Igiri and Nwachukwu (2014) using a mix of logistic regression and artificial neural networks. Their study produced very good findings, with an accuracy rate of 95%. Because they included post-match statistics, including shots, corners, fouls, and even betting odds, their approach is noteworthy. Their focus was on identifying which side was most likely to win based on the events of the match, as opposed to forecasting the results of upcoming games. Because of this, their study was distinct and very helpful in this instance.

A five-year plan for applying machine learning techniques to forecast football game outcomes in the Premier League and La Liga was outlined by Hu and Fu (2022). The method involves selecting features, preparing the data, and using supervised learning techniques such as logistic regression, gradient boosting decision trees, and random forests. With a training set accuracy of 66.7% and a test set accuracy of 63.8%, the Random Forest model outperforms the other models. Yet, factors including changing regulations, irregular team composition, and off-field circumstances significantly limit prediction accuracy.

Through an analysis of in-game statistics for English Premier League (EPL) football predictions, Pipatchatchawal and Phimoltares (2021) used the team and individual ratings from video games to enhance projections. They assessed the performance of two fusion-based models, the sequential and ensemble models, using data from EPL seasons 2010-2011 to 2015-2016. They found that when current match factors are considered together with the recent match results of opposing teams, forecast accuracy increases. The models also performed better when only two or three seasons of data were used as opposed to all five. Although the recommended models outperform comparable research models, they still fall short of high accuracy, possibly due to feature limitations. They noted that not many prediction models have been made public, maybe because football clubs hide their strategies, and they suggested that additional feature analysis and selection could increase the predictive model's efficiency.

A statistical model for football game prediction in the English Premier League was developed by Baboota and Kaur (2019). They utilized exploratory data analysis and software engineering to identify key factors for football match prediction and to generate a feature set. Using machine learning techniques, they planned to develop an incredibly comprehensive forecasting system. They found that the efficiency of the model was closely associated with the essential features they had identified. Their best model, supported by trends, achieved a performance measure of 0.2156 in the probability (RPS) meter for game weeks 6 to 38 under two seasons (2014–2015 and 2015–2016) of the English Premier League. However, compared to their model's predictions, betting businesses like Pinnacle Sports and Bet365 outperformed it, with RPS values for the same period of 0.2012. A smaller RPS value indicates a higher level of prediction accuracy. Consequently, despite the positive results, their model's accuracy was no more than that of bookies' predictions.

Driven by the desire for more accurate football predictions, Alfredo and Isa (2019) focused on using tree-based machine learning algorithms (C5.0, Random Forest, and Extreme Gradient Boosting) to predict the outcomes of English Premier League football matches. It not only offers a comprehensive review of relevant research, but it also emphasizes how important accurate projections are to both investors and fans. A source of

data providing various match statistics from ten seasons of Premier League matches is Footballdata.co.uk. By using backward wrapper feature selection, 10 features out of 14 were selected for the feature set. Three classification algorithms—Random Forest, Extreme Gradient Boosting, and C5.0—were used to provide predictions. After adjusting parameters, Random Forest performs best, with an accuracy of 68.55% on the validation dataset followed by the extreme gradient boosting technique (67.89%) and the C5.0 algorithm (64.87%), which had the lowest accuracy. A comprehensive assessment of the model's performance was conducted, utilizing several measures. The study showed that when it comes to football match results, decision tree-based models (Random Forest) provided superior forecast accuracy. Parameter modification was a crucial factor to consider in maximizing model performance.

By applying machine learning, Azeman et al. (2021) predicted the results of football matches during the English Premier League's 2005–2006 season. The multiclass decision forest and multiclass neural network methods were evaluated. The multiclass decision forest outperformed the multiclass neural network with an accuracy rate of 88% as opposed to 71%. The study that showed how well machine learning works to forecast football match outcomes revealed Decision Forest to be the optimal option.

Football analytics were examined by Vashist et al. (2022), with an emphasis on English Premier League match results and league winners. To anticipate match outcomes with an accuracy rate of 80%, it included more characteristics and used ensemble approaches in the first section. In the second section, the study developed a model that achieved an amazing 94.8% accuracy rate in predicting the league champion after each match week. The techniques used can be applied to different football leagues, and it includes an interactive Flask application to show these predictions.

Conducting a study to forecast the results of the 220 games in the English Premier League for the 2017–2018 season using the random forest classifier and the multilayer perceptron model, Pugsee and Pattawong (2019) used 1140 matches from the three previous seasons of the competition's data to train their algorithms. They created three distinct models, as opposed to one that might have predicted the tie, the home victory, and the away victory. The predictions were divided into three categories:" home wins,"" draws," and" away wins." The random forest classifier fared better than the other models, with accuracy rates between 79% and 81%. The precision (the accuracy of the positive predictions), ranged from 60% to 80%, while the recall (the proportion of real positive instances that were correctly recognized) varied from 40% to 88%.

In the context of football, Snyder (2013) used an innovative approach that combined a betting strategy with outcome prediction. They considered several variables, not just football-related ones, such as stadium capacity, the distance travelled by the other team, player statistics, and management data. They examined data from the 2010–2011 season using a logistic regression model to forecast the results for the next English Premier League season. 51.06% was the model's accuracy. It's noteworthy to note that the study found that the most significant predictors were player ratings and the results of the two prior games.

The English Premier League (EPL) season 2015–2016 was forecasted using a logistic regression model by Prasetio et al. (2016), the major source used for the research. To train their model, they used information from the same competition between 2010 and 2014. However, their methodology was designed to predict which team would win and did not take into consideration the possibility of a" draw. The ratings for the offensive and defensive performances of the home and away teams comprised the only four features

that made up their model. It's noteworthy that these components originated from the FIFA 5 video game. They claimed to have obtained a 69.5% accuracy rate with their top model.

By focusing on victories or draws both at home and away and analyzing the critical aspects impacting full-time outcomes, Jawade et al. (2021) investigated the usage of multiple machine learning algorithms to anticipate the Premier League 2017–18 season. The machine learning techniques they used were Multinomial Naive Bayes classifier, Random Forest, Logistic Regression, Support Vector Machine, and Linear regression. These models were applied using season-specific data from 2017–2018, with an emphasis on feature engineering and data preprocessing. Several machine learning techniques are tested, and their accuracy is compared. The outcomes from the Random Forest Classifier, Naive Bayes, and Linear SVC were initially subpar. Consequently, K-Nearest Neighbors (KNN) and Logistic Regression were used to predict the results; KNN proved to be most effective in this study. They emphasized the importance of traits, such as home-field advantage and previous team success, in improving forecast accuracy. It also highlighted how useless some algorithms such as Random Forest and Support Vector Machine are at forecasting football outcomes. The results show that combining key traits with K-Nearest Neighbors is the most accurate way to predict football game outcomes.

2.2 A Critical Review of Predictive Models in Football

A unique method was used by Hucaljuk and Rakipović (2011) to forecast football game outcomes. Rather than calculating the number of goals each team scored, they gathered their statistics through in-game match activities. They experimented with several machine learning techniques to produce these forecasts, and they assessed their models using standard procedures. The anticipated goal measure, which they also established, is an innovative method that evaluates a team's performance by considering the level of scoring opportunities rather than just the quantity of goals scored. Information on how a team's offensive and defensive ranks evolved during a game was merged with this statistic. They were able to create a classification model to forecast match results and a regression model to forecast score changes using this information. This is a noteworthy outcome in the field of football match prediction since the researchers discovered that their models almost exactly matched the accuracy of predictions given by bookies, performing well and being compatible with common tactics.

Examining the use of machine learning and candlestick charts made from betting market data to forecast the outcomes of NFL games was reviewed by Hsu (2020). Regression was used to predict winning/losing margins, while classification was used to predict wins and losses. Many machine learning algorithms were applied; regression performs better than classification, and the Random Subspace approach has the highest classification accuracy of 68.4%. While the results are only somewhat better than the betting market's estimates, they suggest that betting data's candlestick patterns could represent a valuable resource for predicting match outcomes. The research analyzes team predictions for the game's outcome, emphasizes the need to evaluate team dynamics individually, and considers both the offensive and defensive lines of attack.

To give innovative hybrid classification strategies for football match prediction, KINALIOĞLU and KUŞ (2023) introduced clustering and classification algorithms. They utilized data from over 6,000 European League football matches in addition to remarks left by supporters on social media. Empirical findings indicate that Random Forest (RF)

consistently produces strong outcomes across several match result categories. Football teams and websites that provide sports betting are only two of the places where these models could be useful. Expanding the dataset and further refining the methodology could be necessary to increase forecast accuracy, the study suggests. They concluded with a workable plan for improving the accuracy of football match outcome predictions using hybrid categorization approaches.

To forecast soccer match outcomes, Elmiligi and Saad (2022) proposed a hybrid approach that blends statistical models with machine learning. Data from over 200,000 soccer matches performed across several seasons were analyzed. As part of the study, two hybrid models were developed. Additionally, numerous prediction accuracy assumptions were explored, prediction coding formats were compared, feature engineering was examined, and the models were evaluated using a test dataset. With a prediction accuracy of 46.6% and a ranking probability score of 0.2176, the best-performing model stands out. They shed light on the difficult task of predicting soccer match outcomes and emphasized the importance of feature selection and model-building considerations in this field.

A method for predicting soccer matches that combines a Bayesian model based on team rankings with a machine learning model that uses historical match data was described. To assess their method, Hervert-Escobar et al. (2018) examined a dataset of over 200,000 soccer matches from different leagues across the globe, including the 2018 FIFA World Cup. As compared to other methods, the results demonstrated higher forecast accuracy. They discussed team morale, skill levels, league-specific rules, and the challenges of predicting soccer match outcomes. It included feature engineering and data preparation techniques including changing likelihood and giving teams a score. The results were analyzed using the Ranked Probability Score (RPS), with an average RPS of 0.2620 for league projections. The method's performance is shown to be either higher than or comparable to the DOLORES algorithm. The method yielded accurate predictions for match outcomes and team positions for the next round, with an average RPS of 0.2761 for the 2018 FIFA World Cup. Several recommendations for future study were included in the report's conclusion. These included looking into knowledge-based systems, developing prediction models for match scores, and using more advanced metrics and assessment techniques.

According to P et al. (2023), football (soccer) match results can be predicted by machine learning, specifically with the Gaussian Naive Bayes method. Football match prediction was underlined, along with its implications for betting and team development. The model achieved an accuracy rate of 85.43%, surpassing the score of 79.81% when it was applied to actual squad data and match results. A brief literature review that compiled prior studies on sports prediction was also provided, along with an explanation of the recommended system's design. This system includes the procedures for collecting, organizing, verifying, generating a model, and forecasting data. The importance of domain expertise, data preparation, and feature extraction was emphasized. The Gaussian Naive Bayes approach was chosen because it is suitable for text classification, convenient, and easy to use. They also revealed the software implementation that was used to determine a team's odds of winning a tournament.

A machine learning-based method for predicting football match outcomes was presented by Carloni et al. (2021). With the use of different machine learning algorithms, the system was able to extract data from the internet. They talked about the selection of appropriate variables and looked at a variety of machine learning models such as Na-

ive Bayes, Random Forest, K-Nearest Neighbors, Support Vector Machine and Artificial Neural Networks (ANNs). The most successful model was found to be the ANN model. The major emphasis of the study's system performance evaluation was the Return on Investment (ROI) metric. The study's positive return on investment (ROI) for the system points to potential profitability. The study further demonstrated the recommended machine learning system's superiority by contrasting it with a baseline methodology.

In conclusion, while AI has gone a long way in forecasting results in various fields, the EPL football game poses a special and dynamic challenge. The literature review has shown that while current AI models show potential, they frequently fall short of accurately representing all aspects of football dynamics, player performance variances, and unforeseen match occurrences.

3 English Premier League Football Methodology

This section outlines the methodology that will be used for this research. The objective is to present a thorough summary of all the different phases and procedures needed to handle the research issue. The English Premier League (EPL) adopts the Knowledge Discovery in Databases (KDD) process which includes data selection, pre-processing, transformation, mining, assessment, and interpretation. KDD was selected because it is a multi-stage, iterative technique used to extract valuable and non-trivial information from huge datasets Debuse et al. (2001). The methodology flow is shown in Figure 1

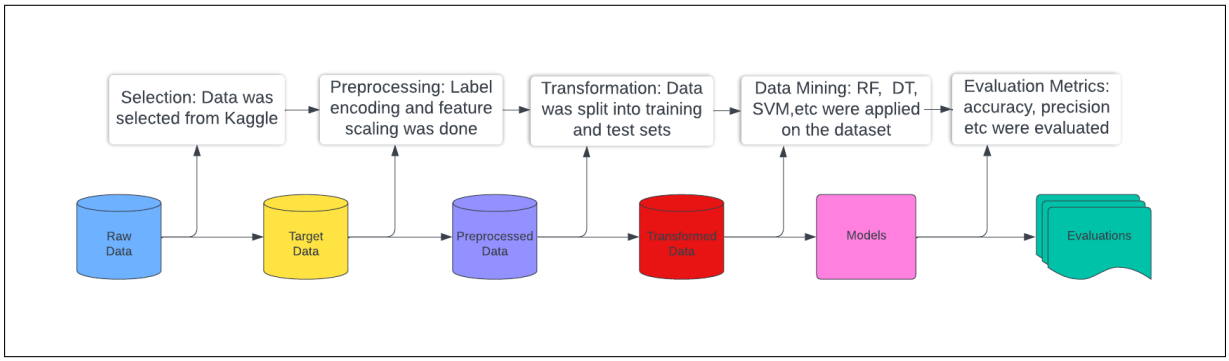

Figure 1: Process Flow Diagram for the Proposed Methodology

3.1 Data Selection

The data for this project is gathered from Kaggle. The dataset consists of 12,026 observations and 8 features. The following is a description of the features of the dataset:

Season End Year: This is the year that the football season concludes. It shows which season the match data is recorded for.

Wk: The symbol" Wk" denotes" Week," indicating a particular week of games in each season. It makes it easier to keep the matches' records and organization in order. For example, the first weekend of matches in a season would be designated as Wk 1, the second weekend as Wk 2, and so on.

Date: This feature shows the exact day that a certain match took place. It helps to analyze the matches in connection to time and arrange them in the correct order depending on their dates since it gives the precise day, month, and year of the match.

Home: The name of the team that is competing in the game in their home stadium is referred to in this feature. It helps in determining which team benefits from playing at home.

Home Goals: The number of goals the home team scored during a game is shown by this characteristic. It aids in determining how successfully the home team performed its attacking strategy and goal-scoring capabilities during a particular game.

Away Goals: The number of goals scored by the opposition in a game is shown by this characteristic. It aids in determining how successfully the opponent's team executed their attacking strategy and goal-scoring capabilities during a certain match.

Away: This feature indicates which team is playing on the other team's field or away from their own. It facilitates determining which team is visiting.

FTR: The term" Full-Time Result," or FTR, indicates how the game concluded. The following values can be used to characterize the match's outcome:

" H" (Home win): The home team emerged victorious in the match.

" A" (Away win): The away team emerged victorious in the match.

" D" (Draw): The game ended in a draw, and hence there was no winner.

3.2 Data PreProcessing

In this section, data preprocessing is carried out to ensure the data is accurate and suitable for analysis. For the preprocessing, necessary libraries were imported, and the data was loaded into the coding environment using the pandas libraries. A thorough examination of missing values in each column was conducted. Remarkably, the dataset exhibited no missing values across all columns, signifying a high level of data completeness. The data set consists of both numerical and categorical data types. The numerical data types include Season End Year, Wk, HomeGoals, and AwayGoals while the categorical data types include Date, Home, Away and FTR. The target feature" FTR" consists of three categories: H-Home, D-Draw, and A-Away.

3.3 Exploratory Data Analysis

The Exploratory Data Analysis is conducted to gain a pictorial view of the dataset. The dataset is visually and statistically explored to gain insights, discover patterns, and identify trends. Figure 2 presents the key findings from the dataset.

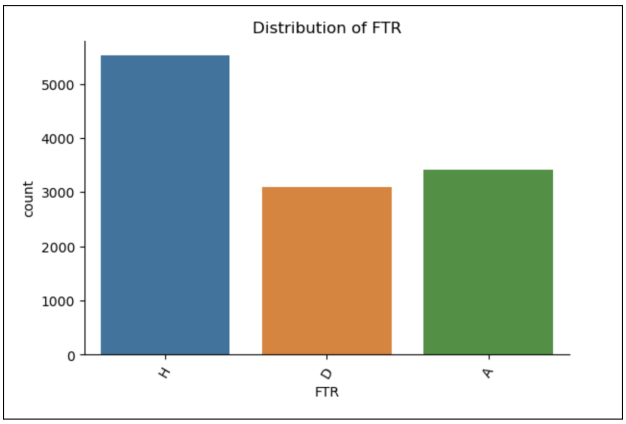

Figure 2: Distribution of FTR

The chart provides an overview of the frequency of various match outcomes. With 5519 occurrences, home victories are the most common, followed by away victories with 3410 occurrences and draws with 3097 occurrences. This visualization helps to comprehend the Premier League's general competitive balance across the given period better.

Figure 3 shows the percentage of total goals contributed to games played at both home and away. The results showed that 18332 goals (57.2%) were scored at home, and 13740 goals (42.8%) were scored away. This chart provides an accessible and graphically appealing summary of how goals are distributed in various match conditions.

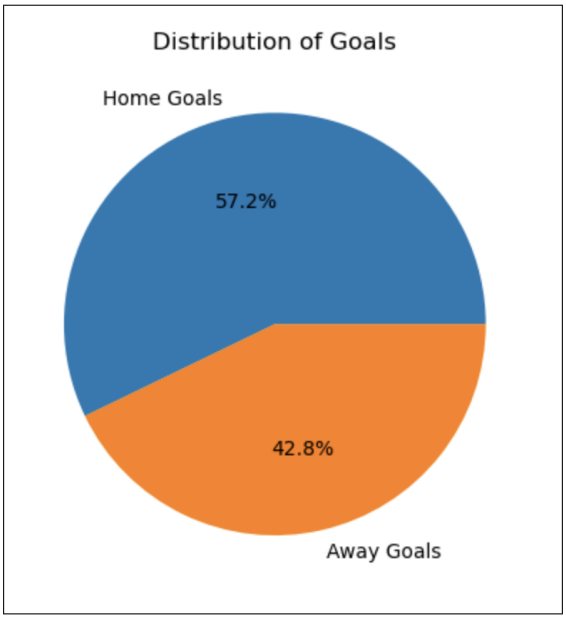

Figure 3: Distribution of goals

3.4 Data Transformation

Data transformation is the process of changing the source data to improve its fit for modelling or analysis. To improve the data's quality, accessibility, and comprehension, this procedure involves numerous steps. The pd. to datetime () method is used to convert the Date column to the datetime data type. Date-related actions, such as sorting and time-based filtering, were handled more easily due to this modification. Label encoding is used to transform the target (FTR), home, and away variables to numerical values. After conversion," FTR" displayed labels like" H,"" A," and" D," which stands for Home Win, Away Win, and Draw, respectively. The numerical values inside it were 0, 1, and 2, which correspond to these categories. When training machine learning models that need numerical input for the target variable, this numerical representation is frequently helpful. Two additional columns," DayCode" and" MonthCode," were added to the English Premier League dataset to improve its temporal features. This feature engineering aims to use the" Date" column to determine the day of the week and month for every match. For every match date, the" DayCode" column specifies a numerical code that corresponds to the day of the week. For every match date, the month's numerical representation is recorded in the" MonthCode" field. These modifications make it easier to include temporal

information within the dataset, enabling more complex analysis. The categorical data in the" Home" and" Away" columns were transformed into numerical representations to make it easier to integrate categorical data into the research. To provide connectivity with machine learning algorithms that rely on numerical input, this modification is essential. Following the encoding of the values in the" Home" column into numerical representations, the outcomes were recorded in a new column called" HomeCode." Now, a unique number code corresponds to each team name. The same encoding procedure was applied to the values in the" Away" column, and the resulting numerical representations were recorded in the" AwayCode" column. The incorporation of team information in a manner appropriate for analysis and modelling is made possible by this numerical representation. The average team statistics for the previous five games were calculated using rolling averages. Sorting the date and categorizing the teams is how this is done.

3.5 Data Mining

The transformed data is now divided in an 80% to 20% ratio into training and testing sets. To guarantee that the model's performance can be assessed on unobserved data, this is done. To find important patterns and insights, the data is analyzed using classification algorithms. Many machine learning classification techniques, including Decision Tree, Support Vector Machine (SVM), Random Forest, Logistic Regression, and Extreme Gradient Boosting (XGBoost), were used to predict the results. The efficacy of these algorithms in predicting EPL match outcomes and their ability to handle the complexity and dynamic of football match data were taken into consideration while choosing them. Using the training data, the models were assessed and trained. Methods for cross-validation and feature selection were used to improve the model's performance.

Random Forest: This is a tree classification that can be applied to forecast outcomes. This approach was introduced by Breiman (2001), which modifies the construction of the regression and classification trees. According to this strategy, subsets of predictors randomly selected at each node are separated into their best. This method will make the tree created robustly against overfitting, according to Breiman (2001).

Logistic Regression: This machine-learning technique determines the likelihood of an occurrence based on the values of the independent variables. This statistical method makes use of data sets that have one or more independent factors influencing the outcome Dreiseitl and Ohno-Machado (2002).

Decision Tree: Task classification and prediction can be accomplished with the decision tree prediction model approach. According to Dunham (2006), the Decision Tree divides the problem-finding area into a group of issues using the" divide and conquer" strategy. Converting the data table's form into a model tree is the decision tree procedure. As per Basuki and Syarif (2003), the model tree will produce simpler rules.

Support Vector Machine (SVM): A group of potentially extremely potent machine learning techniques is known as the support vector machine (SVM) Huang et al. (2002). The foundation of Support Vector Machine Learning (SVM) is the construction of an optimum hyperplane, sometimes called a decision boundary or optimal boundary, which optimizes the distance between the closest samples, or support vectors, to the plane and successfully divides classes Huang et al. (2002), Foody and Mathur (2004), Yang (2011).

Extreme Gradient Boosting (XGBoost): This was conducted by Tong He and Tianqi Chen. This approach was presented to address challenges in the Higgs boson ma-

chine learning competition. By examining ensembles from boosted trees, this technique is an advancement of the gradient boosting strategy. Good accuracy values and a quick training procedure are provided by this approach Cowan et al. (2015).

3.6 Model Evaluation

During the machine learning stage, it is crucial to assess how well a trained model performs on a test dataset. Usually, a variety of metrics appropriate for the kind of work being done—like classification—are used to do this. The output evaluation step is crucial because it enables the evaluation of the model's performance on data that was not previously known as well as its capacity to adjust to novel situations. It also enables the comparison of the performance of many models and selects the one that best suits the current situation. Among the metrics that are frequently used for classification problems are accuracy, precision, F1 score, recall, and confusion matrix.

4 The Design Specification

The design specification is an essential element that outlines the basic elements of the recommended solution when it comes to using artificial intelligence to predict English Premier League football matches. This section aims to give a thorough understanding of the methods, structures, and frameworks that constitute the implementation. Figure 4 shows the design architecture of this project.

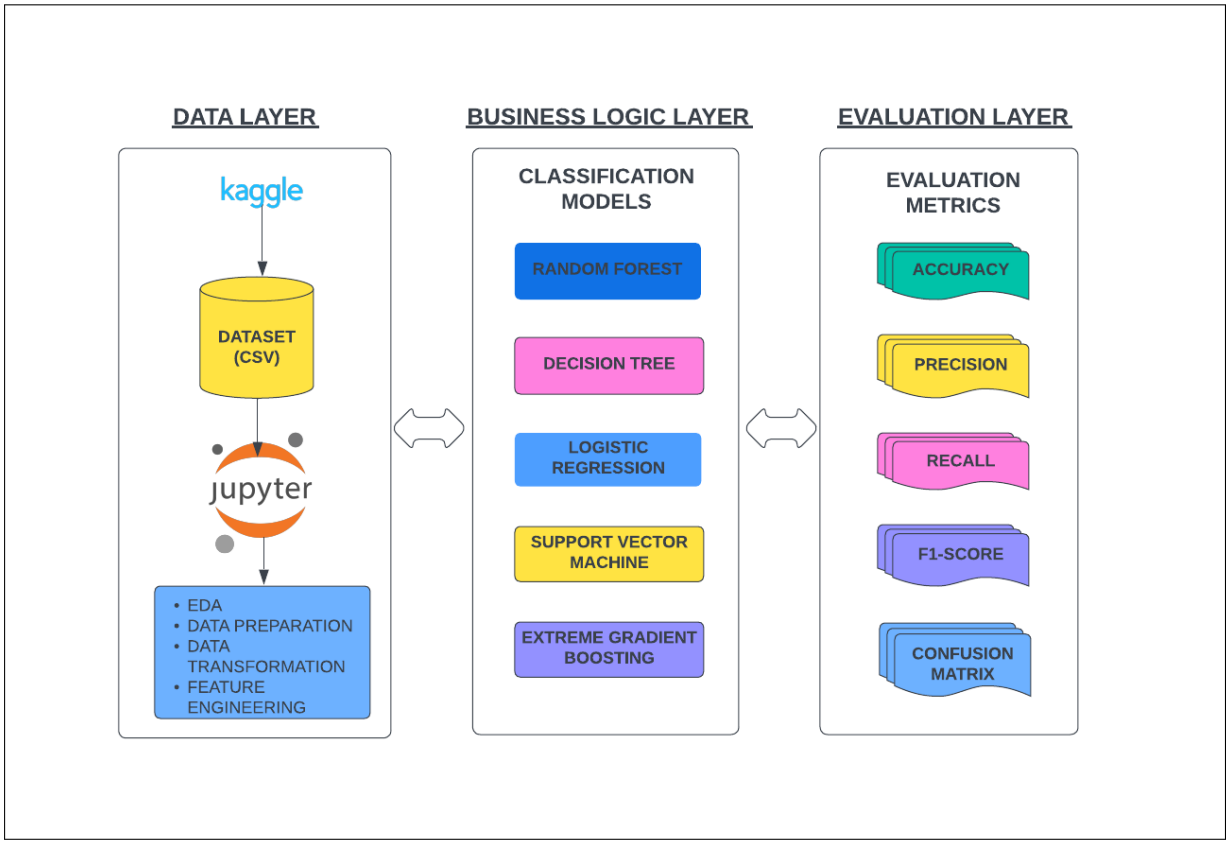

Figure 4: EPL Football Process Flow Diagram

Data was first collected from Kaggle for the design specification for English Premier League football match prediction. The CSV file containing the dataset is imported into a Jupyter Notebook. Preprocessing the data to handle missing values and exploratory data analysis were the next stages. To optimize the model's performance and raise the quality of the data, feature engineering was also applied to the dataset. After that, the dataset is split up into test and training sets. To predict match results, a variety of machine learning models were used, including logistic regression, decision forests, random forests, support vector machines, and XGBoost. Many measures, including recall, F1 score, precision, accuracy, and confusion matrix, are used to evaluate the performance of the model. To enhance the performance of the best model, hyperparameter tuning was also carried out. The accuracy of the model is often increased by adjusting the hyperparameters.

5 Implementation

This research is centred around the critical phase of model training and assessment. This section discusses and carefully implements the approaches described in Section 3.

5.1 Tools and Technologies

Python and Jupyter Notebook is the technologies and tools used for the implementations. To process, evaluate, and develop prediction models, the Python programming language, and associated libraries—Pandas, Matplotlib, and Scikit-learn are utilized.

5.2 Data Collection and Preparation

The CSV-formatted file includes historical match data spanning from 1993 to 2022. The dataset is free of missing values. There are 12026 items total, comprising 4 object kinds and 4 integer types. A mean of around 19.73 and a maximum of 42 matches each week indicate differences in match scheduling. Teams that played at home achieved 1.52 goals on average, while teams that played away achieved 1.14 goals on average. The goal totals for both home and away goals have a median of 1, and the range is 0 to 9. The variance in goal-scoring trends across games is indicated by the standard deviations for both home and away goals. The" FTR" consists of three values which are H (Home Win), A (Away Win) and D(Draw). Home Win has 5519 goals, Away Win has 3410 goals and Draw has 3097 goals. There are 50 distinct home teams with Manchester United, Arsenal, Liverpool, Everton, and Chelsea having played 595 matches each. Some teams have played a fair number of matches while some have played more often than others. Manchester United is the team with the most goals both home and away. This indicates that, whether they are playing at home or away, Manchester United has a solid goal-scoring record. 31,072 goals have been scored overall in the dataset (18,332 goals at home and 13,740 goals away).

5.3 Feature Engineering

The" Date" column is converted to a datetime format. The days of the week and the months are taken from the" Date" column and stored in two new columns called" Day-Code" and" MonthCode," respectively. The LabelEncoder is used to encode the" FTR" (Full-Time Result) column. The encoded values are then placed in a new column named"

Target." The numerical codes generated from the category columns" Home" and" Away" are saved in the" HomeCode" and" AwayCode" columns, respectively. Thirteen columns total, including the original ones and the new ones generated during preprocessing, make up the resultant data frame. The columns' data types are object, datetime, int64, int32, and int 8. For the chosen variables," HomeCode,"" AwayCode,"" DayCode," and" MonthCode" a correlation matrix is created towards the" Target" variable.

5.4 Modelling

The rolling average function is implemented and applied to the data frame creating new columns for the rolling average of selected features. The function takes a group (a subset of the Data Frame), columns to calculate rolling averages for (cols), and names for the new rolling average columns (new cols). It sorts the group by the" Date" column to ensure chronological order. It calculates rolling averages using a window size of 3 (rolling 3 matches) for the specified columns. The new rolling averages are added to the Data Frame in the columns specified by new cols. The function then drops any rows with missing values in the new rolling average columns. The modified group is returned. The rolling average's function is applied to each group formed by grouping the Data Frame by the" Home" column. The selected columns for rolling averages are" HomeGoals" and" AwayGoals." The new columns for rolling averages are named" HomeGoals rolling" and" AwayGoals rolling." The resulting data frame is stored in the variable EPL rolling. The index level 'Home' is dropped to flatten the DataFrame. The index is then reset to ensure consecutive integer indices. The resulting DataFrame EPL rolling has additional columns" HomeGoals rolling" and" AwayGoals rolling," representing the rolling averages of home and away goals, respectively, for each team. Rolling averages are useful for smoothing out fluctuations and identifying trends in time-series data. The applied rolling averages in this case will help analyze the teams' performance trends in terms of goals scored at home and away over consecutive matches.

5.5 Data Split

The dataset is split into a training set (train) and a testing set (test) by considering the date. Data up to April 1, 2017, is used for training, and data after that date is used for testing, which is in the ratio of 80% to 20%. The training set has 9516 rows and 15 columns. The testing set has 2360 rows and 15 columns. The predictors for the machine learning model are specified. These include HomeCode and AwayCode along with the new rolling average columns. Features (X_train and X_val) and labels (y_train and y_val) are defined for training and testing sets.

5.6 Machine Learning Classification Models

A variety of machine learning classification models were developed to assess the classification strength, as discussed in section 3.5. Each implemented model will be explained in detail here.

- Random Forest: Three parameters, n_estimators, min_samples_split, and random_state were applied to the model in the random forest ensemble learning (bagging) technique. To improve resilience against tiny, noisy subsets, it is recommended to include a considerable number of trees (100) in the forest and establish a minimum

number of samples necessary for node splitting (10). This will assist in limiting tree depth. Reproducibility is ensured via a fixed seed (42) which is essential for strong comparisons across various model assessments. Because of its innate resistance to overfitting and sensitivity to noisy data, the Random Forest model was selected. Stability and generalization across different contexts are achieved by the model by combining predictions from numerous trees.

- Logistic Regression: A kind of linear model that emphasizes simplicity and interpretability, the logistic regression employs default parameters. It is especially crucial in situations when transparency is required. To preserve model simplicity and provide resilience across a variety of datasets, default parameters are kept.

- Decision Tree: Among the decision tree types that use default parameters is the decision tree classifier. Decision trees are well-known for their iterative splitting method and are flexible enough to accommodate many types of data. To provide stable performance across many circumstances, the default parameters are kept in place to enable the model to capture complicated interactions without requiring unnecessary adjustment.

- Support Vector Machine: Because of its reliable performance in high-dimensional domains, the linear support vector classifier (a support vector machine) is used. To ensure that the model can grow and handle big datasets steadily, the parameter value dual=False was used for efficiency.

- Extreme Gradient Boosting: As an ensemble learning (boosting) technique, the XGBoost Classifier is known for its effectiveness and ability to improve robustness. Three parameters are used: random state, max depth, and n estimators. A total of 500 estimators were employed. By repeatedly improving the model, more boosting rounds increase resilience. To avoid overfitting and guarantee strong performance, a tree with a maximum depth of 7 was employed. To ensure consistency in assessments, a fixed seed of 42 random states was used. The selected parameters ensure consistent results across runs, prevent overfitting and balance the complexity of the model.

All these models work together to form a strong and dependable ensemble that can manage a variety of research settings because of their properly selected parameters.

5.7 Hyperparameter Tuning

Hyperparameters in machine learning are numbers that regulate the learning process and can significantly affect an algorithm's outputs. As a result, selecting the right hyperparameters is an essential component of machine learning. Choosing a candidate set of hyperparameters, testing each combination on a training set while using cross-validation, and choosing the one that optimizes a predetermined performance metric is a straightforward and popular method of hyperparameter tuning. For instance, the goal of a spam filter is often to maximize" accuracy," or the percentage of properly identified data. Instead, when using probabilistic forecasting, one may try to maximize a function of a probabilistic forecast and its result about a scoring criterion like the log loss Wheatcroft (2019). The hyperparameter was applied to the Extreme Gradient Boosting (XGBoost)

because it has the highest accuracy and precision. The process of hyperparameter tuning yields the optimal hyperparameters, which are printed. The process of hyperparameter tuning yields the best estimator. The training set is used to train the best model. Predictions for the test set are based on the training model. The precision score and accuracy score functions in scikit-learn are used to compute precision and accuracy scores. The tuned model's accuracy and precision scores on the test set are printed. The following hyperparameters were used.

learning rate: This hyperparameter moves the step size closer to the loss function minimum with every iteration. Although it requires more iterations, a lower learning rate could result in improved model performance. The following values were investigated: 0.0001, 0.001, 0.005, and 0.5.

RepeatedKFold: This cross-validation technique involves doing the K-Fold cross-validation procedure three times, or ten folds each. The dataset is divided into training and testing sets to do hyperparameter tuning.

subsample: It shows the percentage of the training set that is randomly selected to develop trees. Stochastic gradient boosting, which introduces unpredictability and can boost resilience, occurs when a value is less than 1.0. The values 0.5, 0.1, 0.2, and 0.9 were investigated.

n estimators: The number of boosting rounds or trees to be built. More trees generally lead to better performance, but there's a trade-off with computational cost. The values that were explored are 100, 400, 800, and 1000.

max depth: This is the maximum depth of a tree. It controls the maximum number of nodes from the root to the farthest leaf node. Deeper trees can capture more complex patterns but may lead to overfitting. The values that were explored are 4, 6, 8, and 10.

RandomizedSearchCV: This is a method for hyperparameter tuning that randomly samples a specified number of combinations from the hyperparameter space. It uses the XGBoost classifier (xgb), and the hyperparameter grid (parameters), it evaluates the models based on precision (macro-averaged) and performs 10x3 cross-validation.

6 Evaluation and Results of English Premier League Classification Models

Accuracy, confusion matrix, recall, precision, and F1-score are the evaluation metrics used in this study to gauge how well the classification models performed. For this investigation, the most important factor is accuracy. This is because data transformation involves creating additional variables that combine to generate the target variable" FTR" to increase prediction accuracy. Experiments are conducted to validate each model used in the modelling process, including the Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, and Extreme Gradient Boosting classifiers. The most accurate model for correctly and accurately predicting Premier League match results is determined by comparing the evaluation metrics. Using the confusion matrix established for this evaluation, an outcome is categorized as True Positive (TP) if the model correctly predicts a team's victory and the team wins; False Negative (FN) occurs when the model predicts a draw or a loss but the team wins, False Positive (FP) occurs when the model predicts a team will win but the team ends up with a tie or a loss, and True Negative (TN) occurs when the model accurately predicts that a team will not win (or draw or lose) and the team does not win.

Confusion Matrix: The performance of a classification system is shown in a table called a confusion matrix. Home Team Wins, Away Team Wins, and Draw are the three classifications that are used when predicting football matches. A confusion matrix for a multi-class classification issue is described with its constituent parts below.

True Positive (TP): A match outcome is predicted by the model accurately. For example, predicting that a home team wins a match, and eventually, the home team wins the match.

True Negative (TN): The model predicts that a match does not belong to a certain outcome class. For example, predicting a draw for a specific match and it turns out to be a draw.

False Positive (FP): The match outcome is incorrectly predicted by the model. For example, picking the home side to win a game only to have the other team win.

False Negative (FN): The model predicts the lack of a match outcome incorrectly. For instance, predicting that a game would conclude in a draw and one of the sides wins.

Accuracy: The accuracy score indicates how accurate the forecasts are generally. It is the proportion of situations that, in every case, were accurately predicted.

Precision: The correctness of the positive predictions is measured by precision. This is the ratio of correctly predicted positive observations to the total number of predicted positives.

Recall: This measures how well the model can identify and include all relevant cases. It's the proportion of all real positive observations to all positive observations that were accurately predicted.

F1-score: The precision and recall equivalent is represented by this. It enables recall and precision in a balanced manner. In the context of this project, a higher F1 score for a particular outcome class indicates a better balance between correctly predicting that outcome and not misclassifying other outcomes.

6.1 Experiment using a Random Forest Classifier

The Random Forest Classifier model developed is evaluated to understand its performance in predicting the outcome of premier league matches. Figure 5 shows the evaluation metrics of the Random Forest Classifier. Each metric is explained with its associated values.

Confusion Matrix: For Class 0 (Away Team Win), the model correctly predicted 256 instances where the away team wins (TP). It incorrectly predicted 86 instances as the away team wins when they belong to other classes (FP). There were 427 instances where the model failed to predict away team wins when they occurred (FN). The model accurately predicted 64 instances of draws (TP) for Class 1 (Draw). 120 occurrences were mispredicted as draws while they belong to different classes (FP). There were 351 instances where the model failed to predict draws when they occurred (FN). For Class 2 (Home Team Win), the model correctly predicted 767 instances where the home team wins (TP). In 115 cases, it projected the home side to win whereas in reality, they belong to different classes (FP). In 174 cases, the algorithm was unable to forecast home team victories when they happened (FN).

Accuracy: The accuracy of 46% suggests that, on average, the model correctly predicts the outcomes of matches 46% of the time. This accuracy level is considered modest and could potentially be improved.

Precision: For Class 0 (Away Team Wins), 47% of the instances predicted as away

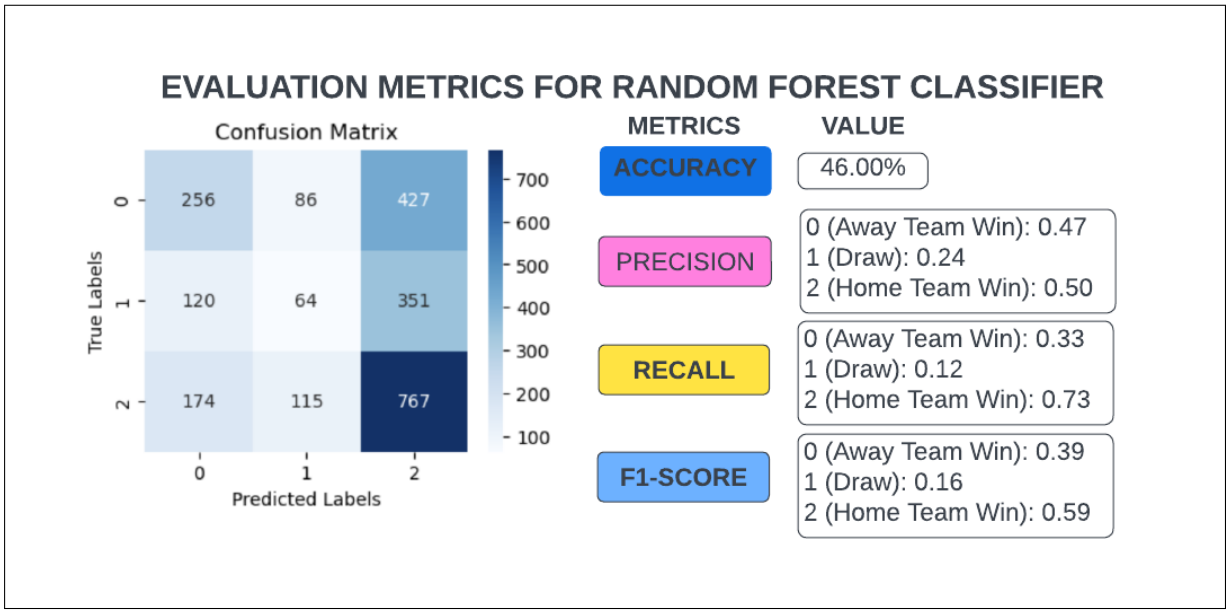

Figure 5: Evaluation Metrics of Random Forest Classifier

team wins are correct. For Class 1 (Draw), 24% of instances predicted as draws are correct. For Class 2 (Home Team Wins), 50% of the instances predicted as home team wins are correct.

Recall: For Class 0 (Away Team Wins), 33% of actual away team wins were correctly predicted. For Class 1 (Draw) 12% of actual draws were captured by the model. For Class 2 (Home Team Wins), 73% of actual home team wins were correctly predicted.

F1-Score: Class 2 has the highest F1-Score of 59%, indicating a better balance between precision and recall for predicting home team wins.

6.2 Experiment using Logistic Regression

The Logistic Regression model developed is evaluated to understand its performance in predicting the outcome of premier league matches. Figure 6 shows the evaluation metrics of the classifier. Each metric will be explained with its associated values.

Confusion Matrix: For Class 0 (Away Team Win), the model correctly predicted 106 instances where the actual outcome was an away team win (TP). The model did not provide any false positives when it came to away team wins (FP). The model failed to predict an away team win in 663 instances (FN). For Class 1 (Draw), the model did not correctly predict any draws (TP), as the count is 0. There were no cases when the model predicted a draw (FP) in error. In all 535 cases, the model was unable to forecast draws (FN). For Class 2 (Home Team Win), the model correctly predicted 968 instances where the actual outcome was a home team win (TP). No cases occurred when the model predicted a home team win when that was not the case (FP). The model failed to predict a home team win in 88 instances (FN).

Accuracy: The accuracy is 0.46 (46%), indicating that the model correctly predicts the outcomes for 46% of instances.

Precision: For Class 0 (Away Team Wins), 41% of the instances predicted as away team wins are correct. For Class 1 (Draw), the model doesn't correctly predict any draws.

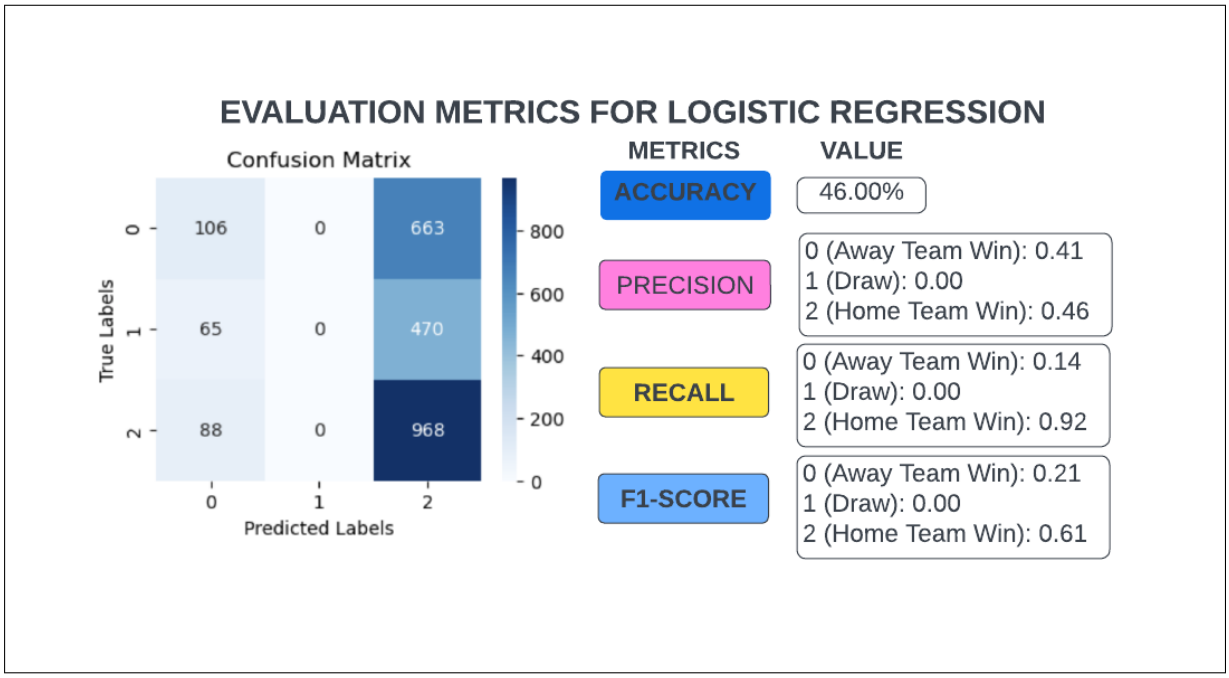

Figure 6: Evaluation Metrics of Logistic Regression

For Class 2 (Home Team Wins), 46% of the instances predicted as home team wins are correct.

Recall: For Class 0 (Away Team Wins), 14% of actual away team wins were correctly predicted. For Class 1 (Draw) the model misses all instances of actual draws. For Class 2 (Home Team Wins), 92% of actual home team wins were correctly predicted.

F1-Score: Class 2 has the highest F1-Score of 61%, indicating a better balance between precision and recall for predicting home team wins.

6.3 Experiment using Decision Tree

The Decision Tree model developed is evaluated to understand its performance in predicting the outcome of premier league matches. Figure 7 shows the evaluation metrics of the classifier. Each metric will be explained with its associated values.

Confusion Matrix: For class 0 (Away Team Wins), the model correctly predicted 266 instances where the actual outcome was an away team win (TP). It incorrectly predicted 188 instances as the away team wins when they belong to other classes (FP). The model failed to predict an away team win in 315 instances (FN). For Class 1 (Draw), the model correctly predicted 137 instances of draws (TP). It incorrectly predicted 158 instances as draws when they belong to other classes (FP). The model failed to predict draws in 240 instances (FN). For Class 2 (Home Team Wins), the model correctly predicted 523 instances where the actual outcome was a home team win (TP). It incorrectly predicted 259 instances as the home team wins when they belong to other classes (FP). The model failed to predict a home team win in 274 instances (FN).

Accuracy: The accuracy is 39%, indicating the proportion of correctly predicted outcomes across all classes.

Precision: For Class 0 (Away Team Wins), 38% of the instances predicted as away team wins are correct. For Class 1 (Draw), 23% of instances predicted as draws are

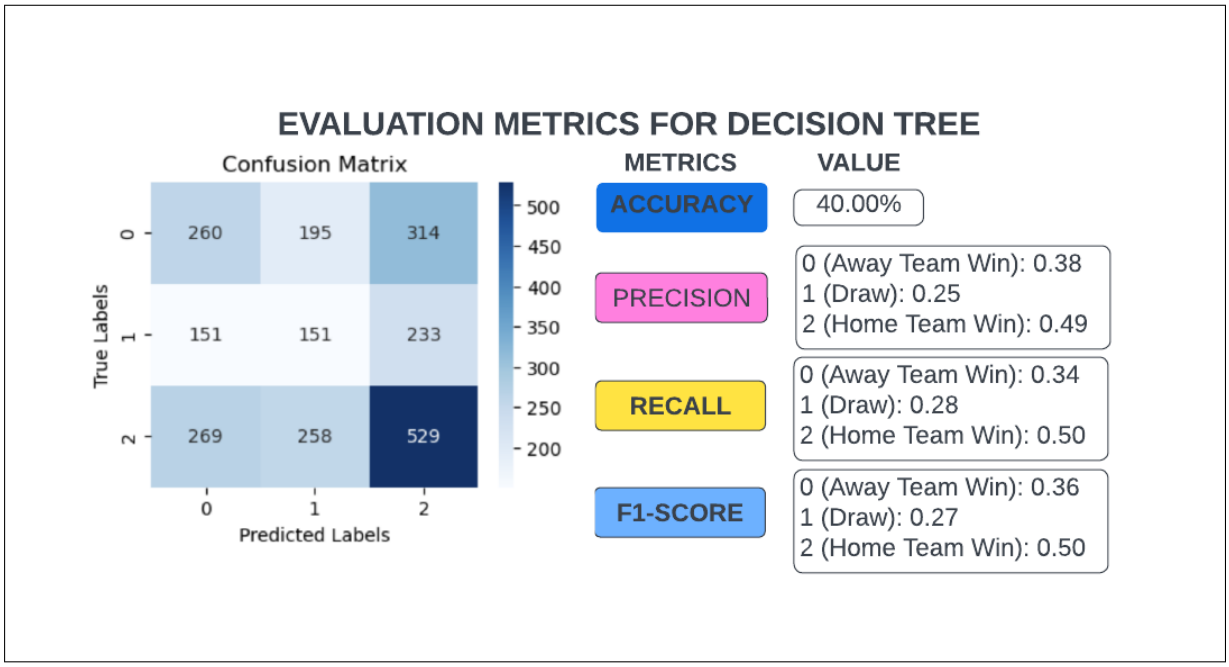

Figure 7: Evaluation Metrics of Decision Tree

correct. For Class 2 (Home Team Wins), 49% of the instances predicted as home team wins are correct.

Recall: For Class 0 (Away Team Wins), 35% of actual away team wins were correctly predicted. For Class 1 (Draw), the model captures only 26% of actual draws. For Class 2 (Home Team Wins), 50% of actual home team wins were correctly predicted.

F1-Score: Class 2 has the highest F1-Score of 49%, indicating a better balance between precision and recall for predicting home team wins.

6.4 Experiment using Support Vector Machine

The Support Vector Machine model developed is evaluated to understand its performance in predicting the outcome of premier league matches. Figure 8 shows the evaluation metrics of the classifier. Each metric will be explained with its associated values.

Confusion Matrix: For class 0 (Away Team Win), the model correctly predicted 102 instances where the actual outcome was an away team win (TP). The model did not provide any false positives when it came to away team wins (FP). The model failed to predict an away team win in 667 instances (FN). For Class 1 (Draw), the model did not correctly predict any draws (TP), as the count is 0. The model did not anticipate a draw (FP) in any of the cases. In all 535 cases (FN), the model was unable to forecast the draws. The algorithm accurately predicted 971 occurrences in Class 2 (Home Team Wins) where the actual result was a home team win (TP). No cases occurred when the model anticipated a home team win when that was not the case (FP). In 85 cases, the model did not correctly forecast a home-side victory (FN).

Accuracy: The accuracy is 45%, indicating the proportion of correctly predicted outcomes across all classes.

Precision: In Class 0 (Away Team wins), 41% of the cases that were projected to be away team victories turned out to be accurate. The model did not accurately anticipate

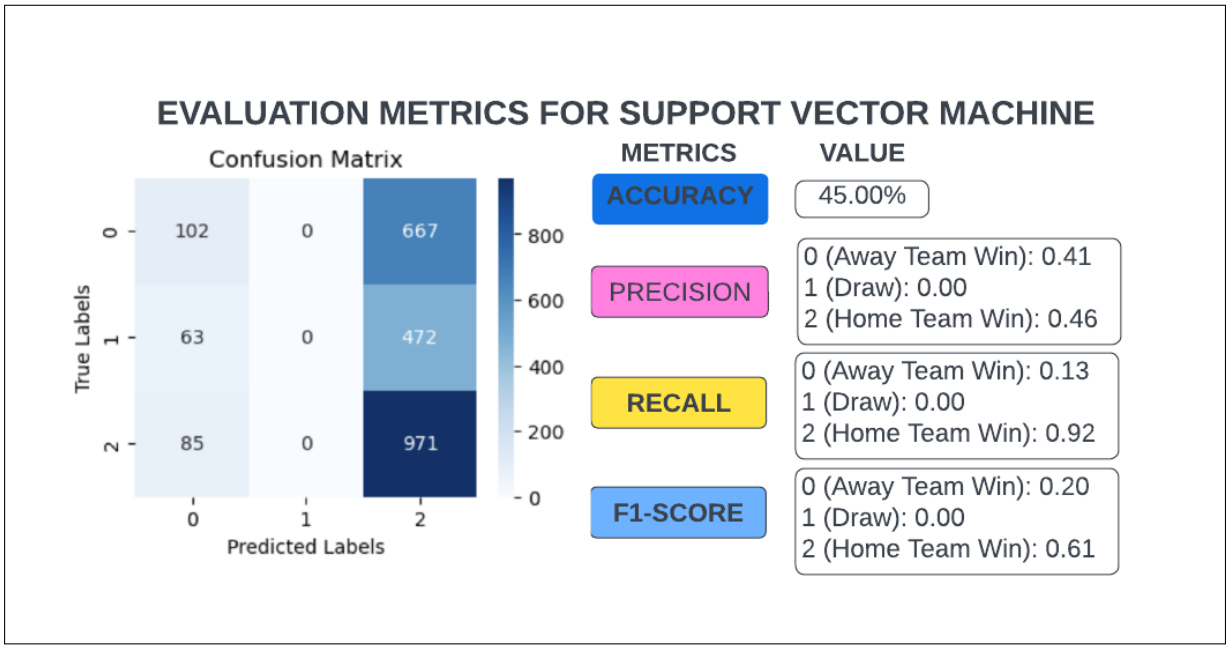

Figure 8: Evaluation Metrics of Support Vector Machine

any draws for Class 1 (Draw). 46% of cases projected as the home team wins in Class 2 (Home Team Wins) are accurate.

Recall: For Class 0 (Away Team Wins), 13% of actual away team wins were correctly predicted. For Class 1 (Draw), the model misses all instances of actual draws. For Class 2 (Home Team Wins), 92% of actual home team wins were correctly predicted.

F1-Score: Class 2 has the highest F1-Score of 61%, indicating a better balance between precision and recall for predicting home team wins.

6.5 Experiment using Extreme Gradient Boosting

The Extreme Gradient Boosting model developed is evaluated to understand its performance in predicting the outcome of premier league matches. Figure 9 shows the evaluation metrics of the classifier. Each metric will be explained with its associated values.

Confusion Matrix: The algorithm accurately predicted 291 cases for class 0 (Away Team Wins) where the actual result was an away team victory (TP). It incorrectly predicted 98 instances as the away team wins when they belong to other classes (FP). The model failed to predict an away team win in 380 instances (FN). For the Class 1 (Draw), the model correctly predicted 71 instances of draws (TP). It incorrectly predicted 137 instances as draws when they belong to other classes (FP). The model failed to predict draws in 327 instances (FN). For Class 2 (Home Team Wins), the model accurately predicted 749 occurrences in which a home team victory (TP) was the actual result. It incorrectly predicted 123 instances as the home team wins when they belong to other classes (FP). The model failed to predict a home team win in 184 instances (FN).

Accuracy: The accuracy is 0.47 (47%), indicating the proportion of correctly predicted outcomes across all classes.

Precision: For Class 0 (Away Team victories), 48% of the cases that were predicted to be away team victories turned out to be accurate. 24% of the occurrences that were

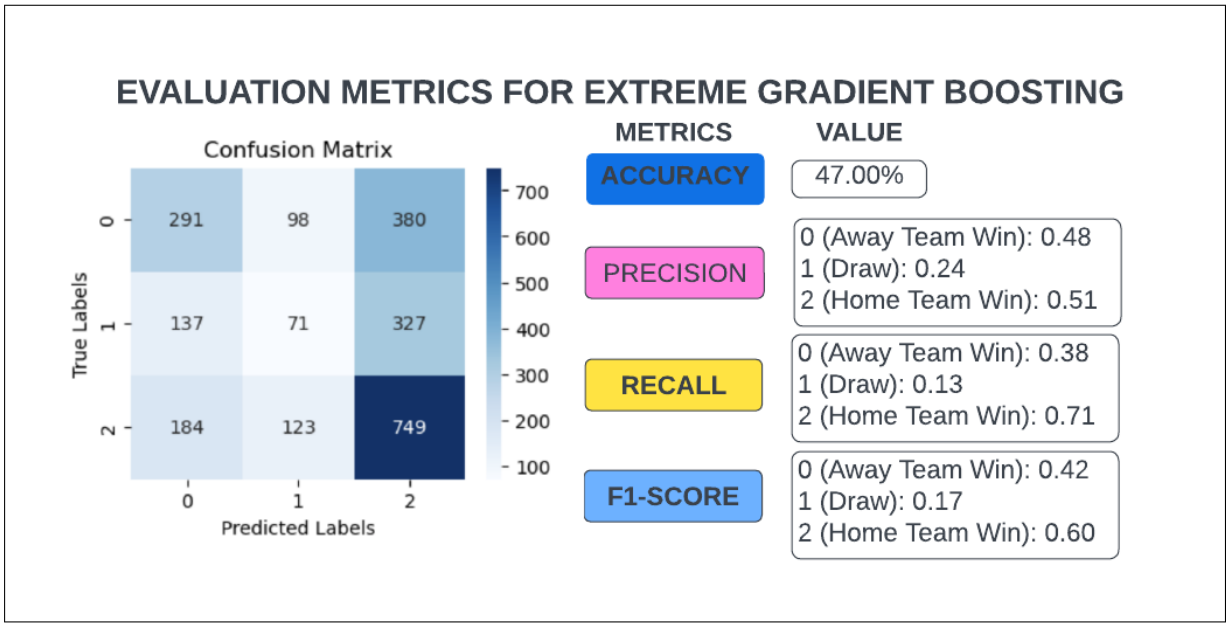

Figure 9: Evaluation Metrics of Extreme Gradient Boosting

predicted as draws for Class 1 (Draw) are accurate. 51% of the cases that were projected as home team wins in Class 2 (Home Team Wins) are accurate.

Recall: For Class 0 (Away Team Wins), 38% of actual away team wins were correctly predicted. For the Class 1 (Draw), the model captures only 13% of actual draws. For Class 2 (Home Team Wins), 71% of actual home team wins were correctly predicted.

F1-Score: Class 2 has the highest F1-Score of 60%, indicating a better balance between precision and recall for predicting home team wins.

6.6 Discussion

In this study, five machine learning models were evaluated for a classification task for the prediction of premier league football matches. The comparison of evaluation metrics of each model is shown in Figure 10. Three classes (0, 1, and 2) were used to test the initial model, the Random Forest. The model performed well in Class 2 in terms of precision, recall, and f1 scores, yielding an accuracy of 46%. The ability of Random Forest to manage complex connections and outliers is what makes it so powerful. Lower class 1 scores suggest that class imbalances may have an impact on the model. With a 46% accuracy rate, the second model, the logistic regression—performed well on Class 2 in terms of precision, recall, and f1 score. Because logistic regression is linear, it might not be as effective at capturing non-linear relationships as tree-based models. This explains why it has trouble handling some patterns in the data. The Decision Tree model, the third model, produced an accuracy of 40% and showed strong performance on Class 2 in terms of precision, recall, and f1 score. Decision trees are prone to overfitting, and noise in the data may be captured by the model, leading to performance deviations. Despite this, Decision Trees' simplicity makes them simple to understand and efficient to compute. With precision, recall, and f1 score all doing well in Class 2, the fourth model, the Support Vector Machine produced an accuracy of 45%. SVM may struggle with large datasets, despite its reputation for doing well in high-dimensional settings. Class 1's worse recall suggests that

MODELS	CLASS	PRECISION	RECALL	F-1-SCORE	ACCURACY
RANDOM FOREST	0	0.47	0.33	0.39	0.46
	1	0.24	0.12	0.16	
	2	0.50	0.73	0.59	
LOGISTIC REGRESSION	0	0.41	0.14	0.21	0.46
	1	0.00	0.00	0.00	
	2	0.46	0.92	0.61	
DECISION TREE	0	0.38	0.34	0.36	0.40
	1	0.25	0.28	0.27	
	2	0.49	0.50	0.50	
SUPPORT VECTOR MACHINE	0	0.41	0.13	0.20	0.45
	1	0.00	0.00	0.00	
	2	0.46	0.92	0.61	
EXTREME GRADIENT BOSTING	0	0.48	0.38	0.42	0.47
	1	0.24	0.13	0.17	
	2	0.51	0.71	0.60	

Figure 10: Comparison of Evaluation Metrics

there may be issues distinguishing class examples from other occurrences. With a 47% accuracy rate, the Extreme Gradient Boosting model is the final one to do well in Class 2. It also has good precision, recall, and f1 score. XGBoost can handle a broad range of data distributions and is resistant to overfitting since it combines the best aspects of boosting and bagging approaches. Given that it can recall more instances from Class 2, it appears to be proficient in that subject. With competitive precision, recall, F1 score, and accuracy across all three classes, Extreme Gradient Boosting seems to be the most successful overall. Hyperparameter tuning was performed on the best model with the highest accuracy which is Extreme Gradient Boosting (XGBoost). Robust model assessment is achieved by using repeated K-Fold cross-validation with 10 folds, 3 repeats, and 1 random state. The XGBoost model's hyperparameters are specified as these four sets of numbers: integers for the learning rate (0.0001,0.001,0.005,0.5); subsample (0.9,0.5,0.2,0.1); number of estimators (100,400,800,1000); and maximum depth (4,6,8,10) are all represented by four sets of integers. Using the RandomizedSearchCV module, the optimal set of hyperparameters is found. The hyperparameter tuning process identified 4 sets of hyperparameters as the best for the Extreme Gradient Boosting (XGBoost) model. The values include subsample (0.2), n estimators (100), max depth (6), and learning rate (0.005). Following the hyperparameter tuning, the accuracy of the model increased by 1.3% to a new value of 49.6%. This suggests that hyperparameter tuning is a very useful factor in the optimization of results. We can categorically state that the best-performing model is 49.6%, which is Extreme Gradient Boosting (XGBoost), which is the highest extent to which premier league matches can be predicted in this research. With the unpredictable outcomes in sports, the accuracy level attained is quite remarkable. In regards to the

existing models, This research builds upon Elmiligi and Saad (2022) work showcasing enhanced accuracy with XGBoost and providing valuable insights into diverse modelling approaches and their implications for soccer match prediction. This information aligns with the research question in Chapter 1, which asks to what extent artificial intelligence, and specifically machine learning models, helps predict EPL matches.

7 Conclusion and Future Work

This research project aimed to determine the potential impact of Artificial Intelligence (AI), more especially machine learning models, on the prediction of English Premier League (EPL) football match outcomes. To achieve this, a series of research objectives guided our efforts. The first objective included doing a thorough literature analysis to provide the groundwork for understanding machine learning techniques used in football match prediction and sports analytics. To make sure the dataset is ready for analysis, Objective 2 concentrated on data pre-processing. Five different machine learning models were implemented in Objective 3: Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Objective 4 assessed how predictions made by AI models affected sports betting markets. To get the best model performance, Objective 5 included hyperparameter tuning. Comparing the developed models to previous models and each other was the goal of objectives six and seven, respectively. The research effort has accomplished the stated goals and addressed the research topic with great success. A thorough grasp of current approaches was given by the literature study. Pre-processing the data made sure the dataset was solid for analysis. The performance of five machine learning models in forecasting the results of EPL matches varied in terms of accuracy. Evaluation metrics provided information on the effect on markets for sports betting. Hyperparameter tuning demonstrated an ongoing dedication to improving outcomes by further optimizing the top-performing model. Important discoveries show that every machine learning model has advantages and disadvantages. Interestingly, with an accuracy of 49.6% after hyperparameter tuning, Extreme Gradient Boosting (XGBoost) proved to be the most successful overall. The model performed well, especially when it came to forecasting Class 2 results. To improve prediction accuracy, further research in this area can involve real-time data integration, more variables, other AI algorithms and additional hyperparameter tuning. Offering betting companies and individuals predictive analytics services that give insightful information about the results of EPL matches has the potential to become a profitable venture.

8 Acknowledgement

I express my deep gratitude to Almighty Allah for His guidance throughout this research. Special thanks to my supervisor, Dr Catherine Mulwa, for her invaluable support. To my parents, siblings, and friends, your encouragement and understanding have meant the world to me. A heartfelt thank you to Olanrewaju Jayeoba, a friend who has been a brother, for his unwavering support.

References

Alfredo, Y. F. and Isa, S. M. (2019). Football match prediction with tree based model classification, *International Journal of Intelligent Systems and Applications* **11**(7): 20–28.

Azeman, A. A., Mustapha, A., Razali, N., Nanthaamomphong, A. and Abd Wahab, M. H. (2021). Prediction of football matches results: Decision forest against neural networks, *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1032–1035.

Baboota, R. and Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for english premier league, *International Journal of Forecasting* **35**(2): 741–755.

Basuki, A. and Syarif, I. (2003). Decision tree, *Surabaya: Politeknik Electronika Negeri Surabaya ITS* .

Breiman, L. (2001). Random forests mach learn 45 (1): 5–32.

Carloni, L., De Angelis, A., Sansonetti, G. and Micarelli, A. (2021). A machine learning approach to football match result prediction, *in* C. Stephanidis, M. Antona and S. Ntoa (eds), *HCI International 2021 - Posters*, Springer International Publishing, Cham, pp. 473–480.

Cowan, G., Germain, C., Guyon, I., Kégl, B. and Rousseau, D. (2015). Nips 2014 workshop on high-energy physics and machine learning, *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, Vol. 42, p. 134.

Debuse, J., de la Iglesia, B., Howard, C. and Rayward-Smith, V. (2001). *Building the KDD Roadmap*, Springer London, London, pp. 179–196.
URL: *https://doi.org/10.1007/978-1-4471-0351-6*$_1$2

Ding, P. (2019). Analysis of artificial intelligence (ai) application in sports, *Journal of Physics: Conference Series*, Vol. 1302, IOP Publishing, p. 032044.

Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review, *Journal of biomedical informatics* **35**(5-6): 352–359.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*, Pearson Education India.

Elmiligi, H. and Saad, S. (2022). Predicting the outcome of soccer matches using machine learning and statistical analysis, *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1–8.

Fialho, G., Manhães, A. and Teixeira, J. P. (2019). Predicting sports results with artificial intelligence–a proposal framework for soccer games, *Procedia Computer Science* **164**: 131–136.

Foody, G. M. and Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines, *IEEE Transactions on geoscience and remote sensing* **42**(6): 1335–1343.

Hervert-Escobar, L., Matis, T. I. and Hernandez-Gress, N. (2018). Prediction learning model for soccer matches outcomes, *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 63–69.

Hsu, Y.-C. (2020). Using machine learning and candlestick patterns to predict the outcomes of american football games, *Applied Sciences* **10**(13): 4484.

Hu, S. and Fu, M. (2022). Football match results predicting by machine learning techniques, *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, pp. 72–76.

Huang, C., Davis, L. and Townshend, J. (2002). An assessment of support vector machines for land cover classification, *International Journal of remote sensing* **23**(4): 725–749.

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques, *2011 Proceedings of the 34th International Convention MIPRO*, IEEE, pp. 1623–1627.

Igiri, C. P. and Nwachukwu, E. O. (2014). An improved prediction system for football a match result.

Jawade, I., Jadhav, R., Vaz, M. J. and Yamgekar, V. (2021). Predicting football match results using machine learning.

Kaplan, A. and Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence, *Business Horizons* **62**(1): 15–25.
URL: *https://www.sciencedirect.com/science/article/pii/S0007681318301393*

KINALIOĞLU, İ. H. and KUŞ, C. (2023). Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods, *International Journal of Nonlinear Analysis and Applications* **14**(1): 2939–2969.

Madan, K., Taneja, K. and Taneja, H. (2022). Intelligent computing based soccer sports management for effective estimation of match outcome, *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 660–664.

P, A. V., D, R. and S, S. N. S. (2023). Football prediction system using gaussian naïve bayes algorithm, *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pp. 1640–1643.

Pipatchatchawal, C. and Phimoltares, S. (2021). Predicting football match result using fusion-based classification models, *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–6.

Prasetio, D. et al. (2016). Predicting football match results with logistic regression, *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, IEEE, pp. 1–5.

Pugsee, P. and Pattawong, P. (2019). Football match result prediction using the random forest classifier, *Proceedings of the 2nd International Conference on Big Data Technologies*, pp. 154–158.

Raju, M. A., Mia, M. S., Sayed, M. A. and Riaz Uddin, M. (2020). Predicting the outcome of english premier league matches using machine learning, *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6.

Rana, D., Vasudeva, A. et al. (2019). Premier league match result prediction using machine learning.

Snyder, J. (2013). What actually wins soccer matches: Prediction of the 2011-2012 premier league for fun and profit.

Vashist, M., Bahl, V., Sengar, N. and Goel, A. (2022). Machine learning for football matches and tournaments, *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Vol. 1, pp. 489–496.

Wheatcroft, E. (2019). Interpreting the skill score form of forecast performance metrics, *International Journal of Forecasting* **35**(2): 573–579.

Yang, X. (2011). Parameterizing support vector machines for land cover classification, *Photogrammetric Engineering & Remote Sensing* **77**(1): 27–37.