Predicting Football Match Outcomes with Ensemble Machine Learning Models

Dhiren Dharmendra Patel

Applied Research Project is submitted for the degree of
Master of Science in Business Analytics
at Dublin Business School

Supervisor: **Dr. Syed Mustafa**

MAY 2024

Declaration:

‘I Dhiren Patel declare that this Applied Research Project that I have submitted to Dublin Business School for the award of is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.’

Signed: Dhiren Dharmendra Patel

Student Number: 20000619

Date: 16/05/2024

Acknowledgements

I would like to begin by thanking the Dublin Business School's Business Analytics department for giving me the opportunity to study and extend my knowledge and experience in the field of Business Analytics.

I want to express my gratitude to my supervisor, Dr. Syed Mustafa, for his consistent encouragement and advice in helping me complete this research study effectively. He was always accessible to answer any questions and provided detailed instructions on the processes of research project implementation during various meetings. It would have been difficult to accomplish the task without him.

Abstract

The study performed here focused on the prediction of the results of football matches using machine learning models. The machine learning models considered in the study were Random Forest (RF), Decision Tree (DT) and the Gradient Boosting classifier and these models were used to build an ensemble model which was the model that performed prediction based on voting. The data associated with football matches was used for training the ensemble model. The best 20 features from the data were selected using the Chi-square technique and the class imbalance in the dataset was solved using Synthetic Minority Oversampling Technique (SMOTE). The results of the study showed that the ensemble model showed an accuracy of 99.5% in predicting the results of football matches. The model was implemented as a desktop application that predicted if the outcome of the football match was a win, lose or draw for the home team.

Table of Contents

Chapter 1
Introduction

1.1 Overview

Football is the most popular sport in the world. It is also one of the largest revenue generating sports in the world (Ozanian, 2023). There are several ways people use football to earn money. Revenue can be generated by owning a football club, selling products associated with football and broadcasting the football matches. Another way in which a lot of people earn money is through betting. Betting is not legal in a large number of countries around the world (Etuk et al., 2022). However, it is legal in several countries where football is highly popular, like European countries. These countries have allowed people to bet on football matches. People are able to perform betting through officially registered gambling companies. Sports betting companies like Bet365 and William Hill have allowed people to bet on football matches online(Gomez-Gonzalez and Del Corral, 2018). A large amount of revenue is generated in association with sports betting and it was found that the global sports betting market had a value of 203 billion dollars in 2020(gitnux, 2023.). The sports betting market is also expanding rapidly and the revenue associated with sports betting is expected to increase by 86% in 2028(F,2021). Football betting plays a major part in the revenue being generated for sports betting around the world. In 2021 it was found that 70% of the revenue generated from sports betting worldwide is generated from football betting(Linder,2023). People place bets on the outcome of football matches, the goals scored by a team, the goals scored by a player etc. However, people lose a lot of money when they perform betting as all the betting is based on guesses by people. The loss of money and being addictive are the reasons betting is looked down upon. People will be so interested in the betting process that they lose track of the money that they have lost leading to a large financial loss. One way to reduce the loss of money is

by somehow predicting the outcome of matches in a systematic manner. Such prediction will help people lose less amount of money.

A large amount of data is being generated from football. These data include the data associated with the performance of the players, the entire team, positions of players on the pitch etc(Figure(1)). The data generated in football is used by countries and football clubs to improve the performances of the players. The data is also used by football clubs to find the best players or the players that fit their needs(Bogicevic, 2018). This data can also be used while performing betting as the historical data associated with the performance of teams and players can be analysed to find out if a team wins loses or draws a match. However manual analysis of the data is a tedious process and the human judgement based on the analysis of data has a high chance of becoming wrong as humans are always prone to error naturally. Analysing the data automatically and performing a prediction automatically makes the prediction process highly effective. One major technique that has been used for performing predictions based on data is machine learning. Machine learning has been successful in performing predictions based on the data generated in a number of sectors(Tercan and Meisen, 2022; Santangelo et al., 2023; Skorikov and Momen, 2020). As a large amount of data is being generated in football the historical data can be used by machine learning models for performing predictions

FIGURE (1): THE STATISTICS OF PLAYERS (BOGICEVIC, 2018)

Machine learning techniques have been used for solving many problems. Even if these machine learning techniques have been successful several methods have been identified to improve the performance of the machine learning models. These techniques ensure that the machine learning models show a good performance in prediction. One such technique is featuring selection. Feature selection involves the selection of the best informative features from the data for training and then machine learning models(Chen et al., 2020). Feature selection helps in improving the time taken for training the machine learning models and the performance of the machine learning models(Khaire and Dhanalakshmi, 2019). Feature selection improves the performance of the machine learning models but another technique that is useful for improving the performance of machine learning models is data balancing. The class imbalance presents in the dataset used for training the machine learning model is solved using data balancing(Mooijman et al., 2023). Data balancing helps in improving the performance of machine learning models as the class imbalance negatively affects the performance of the machine learning models. Another way to enhance the

performance of machine learning models is by using ensemble models. An ensemble model consists of several machine learning models and the final prediction of the ensemble model is based on the prediction made by each individual machine learning model present in the ensemble model. The ensemble model improves the generalizability of machine learning models and this leads to a better prediction performance (Feng et al., 2023). So in the study performed here a system is built for the prediction of the results of football matches using an ensemble machine learning model. In the study, the performance of the machine learning model is improved using feature selection and data balancing. Most of the club football matches are played in a home and away format. So the model proposed in the study predicts if the away team in a match wins, loses or draws.

1.2 Aim

The aim of the study is to build a system based on machine learning that is able to predict if an away team in a football match wins, loses or draws.

1.3 Objective

The objectives of the study are:

- Find a dataset containing the data associated with football matches.
- Handle the missing values and the unwanted elements in the data.
- Use the Chi Square algorithm to select the best features from the dataset.
- Solve the class imbalance in the data using the Synthetic Minority Oversampling Technique (SMOTE).
- Build an ensemble model that contains the Random Forest(RF), Gradient Boosting and Decision Tree(DT).

- Test the performance of the ensemble machine learning model in predicting the outcome of a football match.
- Implement the model as a desktop application that receives data associated with football teams and matches as input and generates an output if the away team wins, loses or draws.

1.4 Research questions

- Will the ensemble model proposed here be able to predict the outcome of football matches?
- Which are the best features associated with the data in the dataset?

1.5 Report structure

The first section of the report consists of the background details associated with the prediction of results of football matches using machine learning. The second section of the report consists of the details of the literature associated with machine learning-based prediction of the results of football matches. The third section of the report consists of the details of the methodologies used in the study. The fourth section consists of the details of the implementation of the system proposed in the study. The fifth section consists of the analysis of the results. The sixth section contains the discussion surrounding the results that were obtained in the study. The seventh and final section of the report consists of the aggregation of the main techniques and results of the study and the details about the future work.

Chapter 2
Literature review

The literature associated with the use of machine learning algorithms for the prediction of the results of football matches is discussed in this section.

2.1 Sports prediction using machine learning

Machine learning is used for the prediction of sports results in the study by(Bunker and Thabtah, 2019). The Artificial Neural Network(ANN) is used in the study for the prediction of results. The data containing the data associated with players, matches, stakeholders, managers, and bookmakers is used in this study. Feature subsets are created in this study to identify the best features in the study and use the features for training. The results of the study show that the best performance in prediction is shown by the ANN that has a backpropagation. However, a large computational burden is associated with building the ANN model.

The performance of the machine learning models in sport outcome prediction is improved by integrating machine learning algorithms with adaptive weighted features in the study by(Lu et al., 2021). The results of basketball games are predicted in the study and the machine learning algorithms considered in the study are stochastic gradient boosting (SGB), RF, Classification and regression trees (CART), extreme learning machine (ELM) and XgBoost. The data used in the study is the data collected from the NBA in the 2018-19 season. The dataset contains 15 features or variables. The best performance in prediction is shown by the SGB as it achieved a Root Mean Square(RMS) value of 11.558. It was found in this study that using the adaptive weighted features improved the performances of the machine learning models. However, the building of the model becomes a complex process when adaptive weighted features are involved in the study.

A review of the machine learning techniques used for the prediction of the results of the matches of team sports was proposed in the study by (Bunker and Susnjak, 2022). Different studies between 1996 and 2019 were analysed in this study. Feature selection and feature engineering were analysed in this study. The results of the study showed that feature selection methods and the feature set used for training machine learning models were important in ensuring that the performances of the machine learning models are good in predicting the results of team sports. The studies showed that machine learning techniques can be effectively used for the prediction of the results of team sports. However, this study only performed an analysis of the literature and no models for predicting the results of team sports was proposed in this study.

2.2 Football match prediction using machine learning

The tactics used in a football game were analysed using machine learning methods in the study by (Herold et al., 2019). The attacking play was the focus of the study. In this study, the ability of different machine learning models like neural networks, k-nearest neighbours (KNN) and logistic regression(LR). The study was carried out by analysing different literature associated with the analysis of data from football using machine learning. However, this study was only a survey of the literature and no actual analysis of the data associated with football was done using machine learning methods in this study.

Deep learning and Machine learning methods are utilised for the prediction of events in a football match and the outcome of a football match in the study by(Herbinet, 2018). A number of different deep learning and machine learning algorithms are considered in this study. The dataset used in the study contained statistics associated with football matches and has been obtained from Kaggle. An extra feature is created in this study for training the machine learning models. The parameters of the deep learning and machine learning models are optimised using OpenMOLE parameter

optimisation software. From the results of the study, it can be seen that the best performance in predicting the probability of a shot becoming a goal is shown by the Gaussian Naïve Bayes model as it achieved an accuracy of 50.5%, the best accuracy in predicting the goals in the match that are not caused by a shot is shown by the Random Forest(RF) as it achieves an accuracy of 51%, the best accuracy in predicting the general outcome of the match is shown by the linear Support Vector Classifier(SVC) as it achieved an accuracy of 51% and the best Mean Absolute Error(MAE) in predicting the match score is shown by the RF model as it achieved a MAE of 0.975. However, the data used in the study consisted of the data from only 2 seasons of 5 footballing leagues.

Machine learning is used for the prediction of the outcome of football matches in the study by(Rodrigues and Pinto, 2022). The machine learning algorithms considered in the study are, linear SVC, RF and KNN. The data associated with football matches like the venue, weather conditions and team news are used in the study. The performance of the team and the home advantage are considered important features that affect the result of a match in this study. The best performance in this study is shown by the RF. The study also shows that machine models like linear SVM are suitable for predicting the results of football matches. However, the issue of overfitting is associated with the RF model.

The prediction of the results of English Premier League(EPL) matches are done using machine learning algorithms in the study by(Choi, Lee Kien Foo and Chua, 2023). The machine learning algorithms considered in the study are LR, linear SVC, RF and Extreme Gradient Boosting(XgBoost). The data used in this study was collected from different online sources and it contained the data associated with EPL matches that took place over 10 seasons. The best features in the data are selected using the Boruta feature selection technique in the study. This study shows that the data collected from a large number of football matches may be imbalanced when used for

the prediction of the outcome of football matches. The data in the dataset used in the study is balanced using both stratified sampling and balanced sampling. The model proposed in the study performed both binary(win and loss) and multiclass(win, draw and loss) predictions. The best performance in multi-class prediction is shown by the RF. Meanwhile, the best performance in binary prediction is shown by the XGBoost model. This study showed the importance of feature selection and data balancing. However, the performance of the machine learning model is not improved using a technique like ensemble learning.

Football match outcomes are predicted using machine learning and deep learning algorithms in the study by(Carloni et al., 2021). The machine learning algorithms considered in the study include the LR, KNN, SVM, NB, RF and an ANN. The data used in the study is collected by web scraping. The best features from the dataset are selected in the study and for performing the features with the most importance feature importance, univariate selection and correlation matrix are used in the study and the best 31 features from the dataset are identified. The results of the study show that the best performance is shown by the ANN as it achieved an accuracy of 59%. However, the accuracy achieved by the ANN model is relatively low.

Machine learning models are used for predicting the results of football matches in the study by(Jawade et al,2021). The machine learning algorithms considered in the study include LR, RF, linear SVC, KNN , RF and Multinomial Naïve Bayes. The data used in the study is the data collected from the 2017-18 EPL season. The study focused on the features like team form, goals scored and conceded. The best performance in this study is shown by the KNN. This study shows that the KNN is highly effective in building a model that can predict the outcome of football matches. However, no technique that can improve the performance of the machine learning model in prediction is not used in the study.

The performances of different machine learning models are compared to find the best model that can be used for the prediction of the outcome of football matches in the study by(Sjöberg,2023). The machine learning algorithms considered in this study include RF, LR and NB. The data associated with the team attributes, player attributes and match statistics are used in the study. For the data associated with the player statistics, the data from the European Soccer database is used. The data from Fifaindex.com is used for getting the data associated with team attributes and the data associated with match statistics is obtained from DataHub.io. The predictions of the outcome of the football matches are made by focusing on player attributes, match statistics and team statistics. Based on the player attributes it is seen that the best performance in predicting the outcome of the football match is shown by the RF, the best performance based on the match statistics is also shown by the RF but based on the based on the team attributes the best performance is shown by the LR. It is also seen in the study that the NB showed balanced performances for all three types of attributes considered in the study. The best features from the respective datasets are used in the study. However, the performances achieved by the machine learning models built in the study vary based on the different kinds of attributes and no one model showcases the best performance in all of the three scenarios containing the different attributes.

The results in the English premier league are predicted and modelled using machine learning methods in the study by(Baboota and Kaur, 2019). Machine learning models like SVM, Gaussian naive Bayes, gradient boosting and RF were used in the study. The data used in the study was the data obtained from a public database based in UK and the data was collected between 2005 and 2016. Feature engineering was performed in this study to make meaningful features from the data. The results of the study show that the best performance was shown by the gradient boosting model as it achieved a ranked probability score (RPS) of 0.2156 for the game weeks between 6 and 38

over two seasons. However, the data used in the study was outdated as the data collected between 2005 and 2016 was used in the study.

Machine learning techniques were used for analysing matches and player attributes in the study by(Stübinger, Mangold and Knoll, 2019). An ensemble model was used in this study and the model consisted of the RF, Boosting, SVM and LR. The data used in the study consisted of from matches of the top five European football leagues collected between the years 2006 and 2018. The study also performed backtesting and statistical arbitrage trading. The results of the study show that the ensemble model achieves a good performance in predicting the results of the matches based on the attributes of players. On considering the performances of the individual machine learning models used in the study, the best performance was shown by the RF. However, the use of the ensemble model may result in high computational expenses.

Machine classifiers like NB, KNN, RF, SVM and LR were used for predicting the results of football matches in the study by (Usman Haruna et al., 2022). The data used in this study was the data from two seasons of the English Premier League, which were 2012 to 2013 and 2011 to 2012. The best features from the data were selected using the Sequential Forward Selection (SFS) technique The results of the study showed that the best performance was shown by the KNN as it achieved an accuracy of 83.95% in predicting the results of the matches. However, the data used in the study was outdated.

The performances of football teams in the long run were predicted using machine learning methods in the study by (Constantinou and Fenton, 2017). The method used for predicting the performances in study was the dynamic Bayesian networks (DBNs). The data used in the study was obtained from various sources and from the seasons 2000-01 to 2014-15 of European football. The data associated with features like managerial ability, squad instability, fatigue, stress, days with injury

etc, were used in this study. The results of the study show that DBN model achieves an average error value of 4.98 points per team per season. However, the DBN has issues in capturing the relationship between random variables.

Deep learning models like Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) were used for the prediction of the results of football matches in the study by (Tiwari, Sardar and Jain, 2020). The data used in this study was obtained from an online source and it was the data generated in the seasons between 2010/11 and 2017/18 of the English Premier League. New features that were not present in the dataset were created in this study like Home Team Goals Scored(HTGS) and Away Team Goals Scored(ATGS). One hot encoding was performed in the study to encode the label values. Hyperparameter tuning was done for different parameters of the models used in the study. The results of the study show that the best performance was shown by the LSTM as it achieved an accuracy of 80.75%. However, the study does not use techniques like data balancing and feature selection which may have improved the performances of the deep learning models considered in the study.

Machine learning algorithms like NB, Linear SVM, Artificial Neural Network(ANN), RF and LR were used for the prediction of the results of football matches in the study by(Zaveri et al.,2018). The data used in the study was the data generated from the five seasons of Spanish La Liga. The data used in the study was pre-processed and feature selection was performed. Statistical approaches were also used for the decision support systems in the study. The results of the study show that the LR shows the best performance as it achieves an accuracy of 71.63%. However, the data used in the study was outdated as the data generated between 2013 and 2018 was used in this study.

RF and SVM classifiers were used for the prediction of the results of the matches in the English Premier League (EPL) in the study by (Yonus Saiedy, Hemmat Qachmas and Faqiri, 2020). The data used in the study was obtained from matches in the English Premier League between the seasons 2013/14 and 2018/19. The data was obtained from different online sources. The feature engineering and feature selection techniques were used to improve the quality of the dataset. The results of the study show that the best performance was shown by the SVM as it achieved an accuracy of 54.3%. However, the accuracy achieved by the SVM model in this study is low.

DT, NB and RF algorithms were used for the prediction of the results of football matches in the study by(Stefano et al., 2020). The data used in the study contained different kinds of data associated with football matches and their results. The data used here contained the results of football matches including home losses, wins and draws from different football leagues. These data were analysed to get an idea about the impact of factors like infrastructure and the size of the country. The results of the study show that the best performance is shown by the DT as it achieved a Ranked Probability Score(RPS) of 0.45 and an accuracy of 35%. The results of the study shows that the DT can be used to predict the results of football matches based on the data associated with football matches. However, the accuracy achieved by the DT model is relatively low.

Machine learning algorithms like ANN, RF, KNN, LogitBoost, Bayesian Networks and NB were used for the prediction of the outcome of football matches in the study by(Hucaljuk and Rakipović,2011). The data utilised in this study was associated with 96 football matches from the group stages of the champions league. The data contained 20 features associated with each team and the features include the average of the goals received and scored, players who are injured, rankings and outcomes of the previous meetings between teams. The results of the study show that the best performance in the prediction of the results of football matches was shown by ANN trained

by utilising the backpropagation algorithm and containing 5 layers that are hidden as it achieved an accuracy of 68%. However, the deep learning model used in the study is computationally expensive.

Machine learning algorithms like LR was used for the prediction of the outcomes of English Premier League matches in the study by (Raju et al., 2020). The data used in the study was the data generated from the English Premier League matches in the seasons 2018-2019 and 2014-2015 which is the historical records for a total of 1870 football matches. The features associated with data were engineered, feature scaling was performed using min-max normalization and feature selection was performed using the Chi-Square feature selection technique. The results of the study shows that the LR model achieves an accuracy of 77.43% for binary class predictions and 70.27% for multi-class predictions. The study shows that the Chi-square feature selection method can be used effectively for selecting the best features from the dataset. However, the LR model assumes linearity between different features and this may lead to incorrect predictions by the model.

SVM was used with a Gaussian combination kernel for predicting the results of football matches in the study by(Igiri,2015). The data utilised in this study contained the data associated with the results of 16 matches in the English Premier League season from the 2014-2015 season. The data focused on the performances of the managers and the players. The data was pre-processed and the string to numerical transformations were performed on the data. The parameters of the SVM were optimised by adjusting the SVM complexity constant and kernel sigma values. The results of the study showed that the model utilised in the study achieved an accuracy of 53.3% with the model predicting eight out of the fifteen matches. However, the time taken for training was very high for the SVM model.

RF, NB, LR and voting classifiers between the NB and RF were used for the prediction of outcomes of football matches in the study by(Vaidya, Sanghavi, Gevaria, 2016). The data used in the study was collected between the seasons 2004-05 and 2014-15. The data contained features like shot ratio, goals conceded and scored, and form. The results of the study showed that the voting classifier achieved an accuracy of 47.11%. The accuracies achieved by the other machine learning models were between the range of 47 to 50% and these models had a mean absolute error of 0.37. The study shows that the voting classifier can be used to predict the outcome of football matches. However, the data used in the study was outdated.

2.3 Research gap

From the literature review, it is observed that the machine learning models are effective in the prediction of the outcome of football matches. The studies show that the data collected from a particular season of a club footballing competition can be used for training the machine learning models. The study shows that a class imbalance may be associated with the data from football and a data balancing technique has to be applied to solve the class imbalance. The importance of feature selection in the studies. However, none of the studies used ensemble models that show a better prediction performance than individual machine learning models.

In the study proposed here machine learning models are used for the prediction of the outcome of football matches. The performances of the machine learning models are improved using data balancing and feature selection. The final prediction result is obtained from an ensemble model that produces a result based on the predictions by the different machine learning models used in the study. Most of the studies used outdated data and, in the study, proposed here the latest data generated from football matches between the years 2021 and 2022 is used. The study proposed

here finds out if the ensemble models are effective in predicting the outcome of football matches. The study also helps in determining the best features in the dataset used in the study.

Chapter 3
Methodology

3.1 Dataset

The datasets used in the study were publicly available and free. It contained the data associated with different football matches in the English Premier League. The data associated with the football matches in the year 2021 and the year 2022 were obtained as separate CSV files (football-data.co.uk,2024). The different features associated with the data of the football matches were given as columns and the different matches between the teams were given as rows. The column 'FTR' represented Full Time Result which means that this column specifies who won the football match. The aim of the study performed here was to predict if an away team would win, draw or lose the match. So the column 'FTR' was used as the label of the data which can be used for prediction. The data in the columns were the letters 'H', 'A' and 'D'. The letter 'H' represented a win for the home team, the letter 'A' represented a win for the away team and the letter 'D' represented a draw. However, the values in the column 'FTR' are alphabets and these cannot be used to train machine learning models.

The data in the two different datasets was combined into a single CSV file and this final dataset was used to train the machine learning model.

3.2 Chi square algorithm

The Chi-square algorithm helps in determining the relevance of each features in the dataset. The Chi-squared test statistic is used in this algorithm to find the feature scores associated with the different features in the dataset. The feature score represents the relation between the target variable in the data and the different features in the data(Spencer et al,2020). A high feature score

shows that the target variable is highly dependent on a particular feature. The main reason for choosing the Chi-square method for determining the best features in this study was the ease of computation of the Chi-square statistic compared to the other methods that are used for determining the most relevant features from the data(McHugh, 2013). From the study of the literature, it was seen that the Chi-square is effective in selecting the best features from the dataset.

3.3 Synthetic Minority Oversampling Technique(SMOTE)

In SMOTE, rather than simply replicating data, it oversamples the minority class by generating artificial instances in the feature space constructed by the instance and its nearest neighbours. This method effectively circumvents the issue of overfitting (Gong and Gu, 2016). The fundamental concept of SMOTE is to create additional data samples in the minority class by interpolating between instances of this class that are closely located to each other. SMOTE enhances the count of instances in the minority class within an unbalanced dataset, thereby improving the classifier's ability to generalize effectively(Joloudari et al., 2023). If data balancing is done using SMOTE then the relevant information in the data is retained as oversampling is done by SMOTE to solve the class imbalance. In other techniques used for solving class imbalance like undersampling where the data associated with the class in majority is reduced, here there is a risk of important data being lost. However, in over sampling new data samples are created and no data is lost. So using the SMOTE helps in retaining the relevant data and solving class imbalance.

3.4 Decision Tree(DT)

DT is a progressive model that effectively and seamlessly integrates a sequence of fundamental tests, where each test involves comparing a numerical attribute to a benchmark value(Bahzad and Adnan,2021). A DT architecture consists of leaf nodes, branches and a root node. A root node, as

its name implies, is the highest node in a Tree and serves as the parent to all other nodes(Patel and Prajapati, 2018). A DT is a structure(figure(2)) where each node represents a characteristic (attribute), each connection (branch) represents a choice (rule), and each terminal point (leaf) represents a result (either categorical or continuous value) (Patel and Prajapati, 2018). The DT is chosen for this study because of its advantages like the ease with which data can be prepared utilising the DT, the non-linear relationships between the features in the data do not affect the performance of the DT (Kotu and Deshpande, 2019).

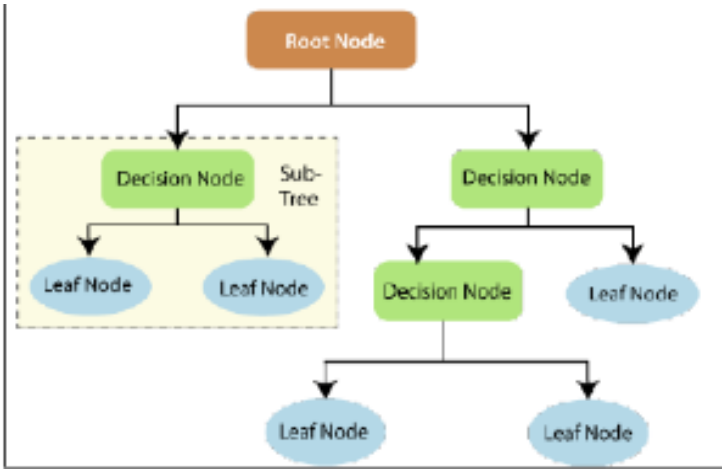


FIGURE (2): STRUCTURE OF THE DT(BAHZAD AND ADNAN,2021)

3.5 Random Forest(RF)

Random Forest(RF) combines multiple DTs to perform predictions. It consists of two steps which are building DTs that are individual and aggregation of predictions(S. A. A and H. F, 2023). Every node in the RF is separated using the best from the subsets that chosen randomly at the node(Huynh-Cam, Chen and Le, 2021). Every tree within the Random Forest is constructed using a bootstrap sample, which is randomly selected from the dataset. The construction process utilizes

the Classification and Regression Tree (CART) technique, and the Decrease Gini Impurity (DGI) is used as the standard for splitting(Couronné, Probst and Boulesteix, 2018). During the construction of each tree, at every division, only a specific quantity of randomly chosen attributes are evaluated as potential splitting candidates. The main reasons for choosing the RF algorithm for this study is that it can be understood easily. RF is flexible and it is able to rank the predictors that are the most significant(Loef et al., 2022).

3.6 Gradient Boosting

Gradient Boosting classifiers that utilize DT's for building the classifiers. The Gradient Boosting Classifiers employ a repetitive process where they enhance the overall performance accuracy by incorporating additional models to rectify the shortcomings of previous models (Bui et al., 2021). Gradient Boosting is a type of AdaBoost classifier. The Gradient Boosting classifiers usually utilize the DT as a weak learner. The Gradient Boosting utilizes the additive model strategy, along with the loss function and a weak learner (Md. Johir Raihan and Abdullah-Al Nahid, 2023). Within the additive model, Gradient Boosting constructs DTs in a manner that is iterative, either in stages or in an order that is sequential. Gradient Boosting utilizes a boosting method where the new model is educated using the residual errors from previous predictions. Gradient Boosting was used in this study because it boosts the performance of the model utilized for classification in the study.

3.7 Ensemble model

The ensemble model is a combination of several machine learning algorithms. Constructing an ensemble model, within this broad structure, involves determining the method for training the foundational classifiers and an appropriate procedure for combining the results of these foundational classifiers (Mohammed and Kora, 2021). In the ensemble model the predictions of

the different machine learning models that make up the ensemble model are analyzed. A voting based technique is performed to determine the classification which has occurred the most or has been done by the machine learning models in the ensemble model. The prediction which is the most in number or has received the most votes is produced as the result of the ensemble model. The ensemble models parallelize the training process and help in making the classification more accurate(Mohammed and Kora, 2021).

3.8 Performance metrics

The performance metrics of the ensemble model are determined to assess the performance of the model. The performance metrics that were found for the model proposed in the study are precision, accuracy, recall and F1-score. A True Positive(TP) classification is the classification by the football match result prediction model in which the model predicts that the home team in a football match will win, and they actually do win. A True Negative(TN) classification occurs when the model predicts that the home team will not win(ie., they either lose or draw) and the home team actually does not win. A False Positive(FP) classification is a kind of classification in which the model predicts that the home team wins but they actually do not win. A False Negative(FN) classification is the kind of classification in which the model predicts that the home team will not win but the home team actually wins.

Accuracy

Accuracy is defined as the ratio of the number of classifications that are accurate to the total number of classifications made by the ensemble model. It is the ratio of the sum of TP and TN to the sum of TP, TN, FP and FN.

Precision

Precision is defined as the ratio of the number of TP classifications to the sum of TP and FP classifications. Precision defines the quality associated with the positive predictions of the ensemble model built here.

Recall

Recall is the ratio of the TP classifications to the sum of TP and FN classifications. Recall is a measure of how the TP instances are computed from all the actual positive samples.

F1-score

F1-score is defined as the harmonic mean of precision and recall.

3.9 Ethical considerations

The study involved the building of a model that predicts the outcome of a football match based on the result achieved by the home team. The data used in the study was freely and publicly available. It contained only data associated with football matches does not contain any sensitive details like the personal information of people. No human subjects were a part of this study. This study does not bias against or favor any of the football teams that are considered in the dataset. The football teams are not discriminated based on financial status. This study does not encourage activities like sports betting in areas where it is illegal. The study does not slander any football club based on their poor results and this study is only aimed at developing a model that predicts the outcomes of football matches.

Chapter 4
Implementation

4.1 Tools

The model proposed in the study was built using Python. The code for developing the model was written in Visual Studio(VS) code. A PC that had 8 GB RAM and an i7 Intel processor was used for the development of the model. The libraries in Python were used for the implementation of machine learning models and the desktop application.

4.2 Combining the datasets

The two datasets were downloaded as two CSV files. The two datasets were then combined. For this, the data from the two CSV files were read. Then the data in the two datasets were concatenated into a single CSV file. The missing values in the dataset or the NaN values in the dataset were replaced with empty spaces(‘ ‘). As a new dataset was created using the data in the two CSV files, the index of the newly created dataset was reset. The existing indices in the two datasets were not considered in the newly created dataset. The index was reset for the newly created dataset because the previous datasets may have indices that are duplicated or non-sequential. The index of the new dataset now starts from 0 and is incremented sequentially. The newly created dataset is given in figure(3).

	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	\
0	E0	13/08/2021	20:00	Brentford	Arsenal	2	0	H	1	
1	E0	14/08/2021	12:30	Man United	Leeds	5	1	H	1	
2	E0	14/08/2021	15:00	Burnley	Brighton	1	2	A	1	
3	E0	14/08/2021	15:00	Chelsea	Crystal Palace	3	0	H	2	
4	E0	14/08/2021	15:00	Everton	Southampton	3	1	H	0	
..	..	...	...	...	...	...	...	..	...	
755	E0	28/05/2023	16:30	Everton	Bournemouth	1	0	H	0	
756	E0	28/05/2023	16:30	Leeds	Tottenham	1	4	A	0	
757	E0	28/05/2023	16:30	Leicester	West Ham	2	1	H	1	
758	E0	28/05/2023	16:30	Man United	Fulham	2	1	H	1	
759	E0	28/05/2023	16:30	Southampton	Liverpool	4	4	D	2	

	HTAG	...	AvgC<2.5	AHCh	B365CAHH	B365CAHA	PCAHH	PCAHA	MaxCAHH	\
0	0	...	1.62	0.50	1.75	2.05	1.81	2.13	2.05	
1	0	...	2.25	-1.00	2.05	1.75	2.17	1.77	2.19	
2	0	...	1.62	0.25	1.79	2.15	1.81	2.14	1.82	
3	0	...	1.94	-1.50	2.05	1.75	2.12	1.81	2.16	
4	1	...	1.67	-0.50	2.05	1.88	2.05	1.88	2.08	
..	...	...	...	...	...	...	...	...	...	
755	0	...	2.14	-1.00	2.02	1.77	2.10	1.81	2.17	
756	1	...	2.50	0.25	1.84	2.06	1.83	2.10	1.90	
757	0	...	2.51	-0.25	1.75	2.05	1.85	2.06	1.90	
758	1	...	2.95	-1.25	1.98	1.92	1.98	1.93	2.07	
759	2	...	3.22	1.25	1.82	2.08	1.85	2.07	1.96	

FIGURE (3): THE NEWLY CREATED DATASET

4.3 Pre-processing the dataset

The columns in the newly created dataset were loaded into a dataframe and these columns were reordered. The column ‘FTR’ is the class in the dataset, so this column was moved to the end of the dataframe. After reordering the columns the dataframe was saved. Then the columns that had null values in the dataset were found. These columns were then removed from the dataframe. The columns ‘Date’, ‘Div’, ‘Time’, ‘AHCh’ and ‘Ahh’ were removed from the dataframe. All the missing values in the dataset were then filled with an empty string(‘’). Now all the columns in the dataset contained string values, this operation was done to make sure that all the columns in the dataframe contained string values.

It was important to make sure that all the columns in the dataframe contained string values as the label encoding was the operation that was performed next. Label encoding was performed categorical or string values in the dataset were converted to numerical values. This process was done because the machine learning models learned better with numerical values than string values (Uddin, Ong and Lu, 2022). So the data in all the columns except the class or column 'FTR' were converted to numerical values. The label encoding was carried out using the method 'LabelEncoder()' imported from the Python library 'sklearn.preprocessing'. The data in the column 'FTR' was not converted to numerical values because an ensemble model with the voting classifier andstring values can be used as during voting for finding the prediction that was repeated the most. Comparing figures (3) and (4) it can be seen that the data in the dataset were converted to numerical values.

	HomeTeam	AwayTeam	FTHG	FTAG	HTHG	HTAG	HTR	Referee	HS	AS	...	\
0	3	0	2	0	1	0	2	16	8	22	...	
1	14	10	5	1	1	0	2	19	16	10	...	
2	5	4	1	2	1	0	2	6	14	14	...	
3	6	7	3	0	2	0	2	11	13	4	...	
4	8	18	3	1	0	1	0	0	14	6	...	
..	...	...	...	...	...	...	...	...	..	..	...	
755	8	2	1	0	0	0	1	22	13	7	...	
756	10	19	1	4	0	1	0	2	19	11	...	
757	11	21	2	1	1	0	2	23	13	16	...	
758	14	9	2	1	1	1	1	20	21	10	...	
759	18	12	4	4	2	2	1	7	15	30	...	

	AvgC<2.5	B365CAHH	B365CAHA	PCAHH	PCAHA	MaxCAHH	MaxCAHA	AvgCAHH	\
0	1.62	5	34	1.81	2.13	2.05	2.17	1.80	
1	2.25	35	4	2.17	1.77	2.19	1.93	2.10	
2	1.62	9	43	1.81	2.14	1.82	2.19	1.79	
3	1.94	35	4	2.12	1.81	2.16	1.93	2.06	
4	1.67	35	17	2.05	1.88	2.08	1.90	2.03	
..	...	...	...	...	...	...	...	...	
755	2.14	32	6	2.10	1.81	2.17	1.92	2.03	
756	2.50	14	35	1.83	2.10	1.90	2.14	1.81	
757	2.51	5	34	1.85	2.06	1.90	2.16	1.82	
758	2.95	28	21	1.98	1.93	2.07	1.98	1.97	
759	3.22	12	37	1.85	2.07	1.96	2.12	1.88	

FIGURE (4): THE DATA FRAME AFTER PRE-PROCESSING

4.4 Feature Selection

The best features or the most important features from the data were then selected. The best features were selected by utilising the Chi-square feature selection algorithm. Feature selection helped in selecting the best features from the dataset, this helped in reducing the training time of the machine learning model. The learning performance of the machine learning model is also enhanced when feature selection is performed (Miao and Niu, 2016).

The columns that contained NaN values were removed from the dataset. The column that contained the values of the different classes the column 'FTR' was not needed for feature selection, so all the other columns or the features in the dataframe were selected. Only the features were considered for feature selection. The 'k' value for the Chi-square method was set as 20 which meant that the best 20 features from the dataframe were selected. For performing Chi-square feature selection, the method 'SelectKBest()' was loaded from the library 'sklearn.feature_selection'. The method 'SelectKBest()' has two parameters: 'k' and 'score_func'. The parameter 'k' was set as 20 as the top 20 features had to be determined and 'score_func' was set as 'chi2' which represented the Chi-square algorithm. 'chi2' was imported from the library 'sklearn.feature_selection'. The Ch-square feature selection was successfully implemented and the best 20 features in the dataframe were successfully determined. The best 20 features in the dataframe are shown in figure(5).

```
Selected features: Index(['FTHG', 'FTAG', 'B365A', 'BWA', 'IWA', 'PSA', 'WHA', 'VCA', 'MaxA',
       'AvgA', 'P<2.5', 'BWCA', 'IWCH', 'IWCA', 'PSCA', 'WHCA', 'VCCA',
       'MaxCA', 'AvgCA', 'PC<2.5'],
      dtype='object')
```

FIGURE (5): THE BEST 20 FEATURES

A new dataframe was then created with only the best 20 features and the data associated with best 20 features from the dataset. The class column or 'FTR' was also added to the new datafarme, so the new dataframe contained the best 20 features from the dataset and the 'FTR' column which was the class.

4.5 Data balancing

The data in the column 'FTR' in the new dataframe was then analysed and it was found that the number of rows or data samples associated with the class 'H' was 347, the number of data samples associated with 'A' was 238 and the number of data samples associated with 'D' was 175 which meant that the number of data samples associated with the classes 'A' and 'D' were much lesser than the number of data samples associated with the class 'H'. This means that the data is imbalanced and if this data is used for training the machine learning model the predictions by the model will be biased in favour of the class in majority ie., 'H'. So data balancing was carried out using SMOTE. SMOTE was implemented using the method 'SMOTE()' which was imported from the library 'imblearn.over_sampling'. The 'SMOTE()' method had the parameter 'random_state' and the value of this parameter was set as the numerical value 42. The data was successfully balanced using SMOTE. After performing data balancing utilising SMOTE the number of data samples associated with 'H' was 347, the number of data samples associated with 'A' was 347 and the number of data samples associated with 'D' was also 347. This shows that synthetic samples were created for the classes in majority and the number of samples for every class was made equal and the class imbalance in the dataset was solved.

4.6 Model training

The RF, DT and Gradient Boosting classifier models were then trained using the data in the dataframe which was obtained after data balancing. The RF model was implemented using the method 'RandomForestClassifier()', the DT model was implemented using the method 'DecisionTreeClassifier()' and the Gradient Boosting model was implemented using the method 'GradientBoostingClassifier()'. The methods 'RandomForestClassifier()' and 'GradientBoostingClassifier()' were imported from the library 'sklearn.ensemble'. The method 'DecisionTreeClassifier()' was imported from the library 'sklearn.tree'. The data in the dataframe was separated into a training set and testing set. The training set was utilised for training the model and the testing set was used for testing the performance of the model. The testing set contained 20% of the data in the dataframe and the training set contained 80% of the data in the dataframe. Then using the DT, RF and Gradient Boosting models implemented here, an ensemble model was built by combining the three classifier models. The ensemble model was implemented as a voting classifier. The voting classifier was implemented using the method 'VotingClassifier()' which was imported from the library 'sklearn.ensemble'.

The parameters used for the method 'VotingClassifier()' were 'estimators' and 'voting'. The parameter 'estimators' defined the machine learning classifiers that are part of the ensemble model and the RF, DT and Gradient Boosting classifiers were initialised to this parameter. The parameter 'voting' was given the value 'hard' as the majority rule voting was utilised in this model. The ensemble model was the complied successfully and saved.

Chapter 5
Results

The performance metrics of the ensemble model were determined and the ensemble model that predicted the outcome of a football match for the home team was implemented as a desktop application.

5.1 Performance Metrics

Accuracy

The accuracy of the ensemble model was determined using the method 'accuracy_score()' imported from the library 'sklearn.metrics'. The data associated with the features and the data associated with the class from the testing set were passed as the parameters of the 'accuracy_score()' method and it was found that the ensemble model built here achieved an accuracy of 99.5%.

Precision

The precision of the ensemble model was determined using the method 'precision_score()' imported from the library 'sklearn.metrics'. The data associated with the features and the data associated with the class from the testing set were passed as the parameters of the 'precision_score()' method and it was found that the ensemble model built here achieved a precision of 99.5%.

Recall

The recall of the ensemble model was determined using the method 'recall_score()' imported from the library 'sklearn.metrics'. The data associated with the features and the data associated with the

class from the testing set were passed as the parameters of the 'recall_score()' method and it was found that the ensemble model built here achieved a recall of 99.5%.

F1-score

The f1-score of the ensemble model was determined using the method 'f1-score_score()' imported from the library 'sklearn.metrics'. The data associated with the features and the data associated with the class from the testing set were passed as the parameters of the 'recall_score()' method and it was found that the ensemble model built here achieved a recall of 99.5%.

5.2 Desktop application

The football match outcome prediction model was implemented as a desktop application. The desktop application was implemented using the different methods from the 'Tkinter' library in Python which was utilised for building the desktop application. The desktop application had only a single interface and this was where the prediction of the outcome of the football match was done. The interface for prediction was displayed when the desktop application built here was run. The interface contained a label with the text 'Features' and next to this label was an input field and this was where the features were given as input. The interface contains a black area below a label with the text 'Activity log'. The interface contained a button with the text 'Predict'(Figure(6)). The button was programmed in a way that when the feature values were given as the input and the button was clicked the trained ensemble model was loaded along with the saved label encoder and feature selector. The best feature after label encoding and feature selection were passed into the loaded ensemble model and the model predicted the result based on voting. The result was shown as text in a pop-up window in the same interface. The pop-up window displayed the name of the

team that won the match along with the name of the team that lost the match or if the result will be a draw(Appendix A).

FIGURE (6): THE INTERFACE OF THE DESKTOP APPLICATION

Chapter 6
Discussion and Evaluation

6.1 Discussion

The aim of the study was to build a model that was able to predict the outcome of football matches based on the results of the home team. An ensemble model was successfully built to perform the prediction and the model was implemented as a desktop application. Python was used to implement the model and the desktop application. The model showed a good performance in predicting the outcome of the football matches. The objectives of the study were achieved successfully. The objectives are:

- Find a dataset containing the data associated with football matches.
- Two datasets were obtained from an online source and were combined to create a dataset that was used for training the ensemble model.
- Handle the missing values and the unwanted elements in the data.
- The missing value and unwanted values were removed from the dataset.
- Use the Chi Square algorithm to select the best features from the dataset.
- The Chi Square algorithm was successfully implemented and the best 20 features from the data were found.
- Solve the class imbalance in the data using the Synthetic Minority Oversampling Technique (SMOTE).
- The SMOTE was successfully implemented and the class imbalance in the dataset was solved successfully.

- Build an ensemble model that contains the Random Forest(RF), Gradient Boosting and K Nearest Neighbour(KNN).
  - The ensemble model was built successfully by combining the RF, KNN and DT models.
- Test the performance of the ensemble machine learning model in predicting the outcome of a football match.
  - The performance of the ensemble model was assessed by determining the performance metrics of the ensemble model.
- Implement the model as a desktop application that receives data associated with football teams and matches as input and generates an output if the away team wins, loses or draws.
  - The trained ensemble model was successfully implemented as a desktop application and the desktop application displayed if the home team wins, loses or draws based on the input given to it.

From the existing literature studied it was seen that only some of the studies specified accuracy in their literature, so the performance of the model built here was compared with the existing models that predicted the results of football matches by utilising machine learning.

The study	**Accuracy(in %)**
Carloni et al., (2021)	59%
Usman Haruna et al., (2022)	83.95%
Tiwari, Sardar and Jain, (2020)	80.75%
Zaveri et al.,(2018)	71.63%
Yonus Saiedy, Hemmat Qachmas and Faqiri, (2020)	54.3%
Herbinet, (2018)	51%
Stefano et al.(2020)	35%
Hucaljuk and Rakipović(2011)	68%
Raju et al.,(2020)	70.27%
Igiri(2015)	53.3%
Vaidya, Sanghavi, Gevaria,(2016)	47.11%
The model proposed here	99.5%

TABLE (1): COMPARISON OF THE PERFORMANCE OF THE FOOTBALL MATCH RESULTS PREDICTION MODELS

From table(1) it can be seen that the accuracy achieved by the ensemble model built here was greater than the accuracies of all of the existing football match prediction models considered here. However, a direct comparison could not be made with any of the existing studies considered here as the existing studies did not use an ensemble model, except for the study by Vaidya, Sanghavi, Gevaria,(2016) that used the voting classifier. The study by Carloni et al., (2021) used the ANN, Usman Haruna et al., (2022) used the KNN, Tiwari, Sardar and Jain(2020) used the LSTM, Zaveri et al.(2018) used the LR, Yonus Saiedy, Hemmat Qachmas and Faqiri(2020) used the SVM, Herbinet(2018) used the SVC, Hucaljuk and Rakipović(2011) used the ANN, Raju et al., (2020) used the LR, Igiri(2015) used the SVM, Vaidya, Sanghavi, Gevaria, (2016) used voting and Stefano et al.(2020) used the DT. It can be seen that the accuracy achieved by the ensemble model was greater than the accuracies achieved by the deep learning models like LSTM and ANN. The voting classifier was used in the study by Vaidya, Sanghavi, Gevaria,(2016) but the accuracy achieved by the model was very low it may be due to the difference in the machine learning models considered for voting and the feature selection and data balancing technique not being used in the study by Vaidya, Sanghavi, Gevaria,(2016). The use of methods like feature selection and data balancing may also have led to the performance of the model built here being greater than the performance of the existing models that perform football match result prediction.

6.2 Evaluation

The building of the ensemble model was time consuming and relatively hard as the familiarity with implementing the ensemble model was less. The designing and programming of the desktop

application were also relatively hard. All the other methods that were a part of the study were implemented relatively easily.

The model built here was able to successfully predict the outcome of football matches but the study had some limitations. The study used data that was relatively new but considered only the data produced in two years. Data generated from a larger period could have been considered for the study. The study only considered the data from the English Premier League, the data from different leagues from different countries could have been used in the study.

Chapter 7
Conclusion and future enhancements

The study was done to build a model that was able to predict if the home team in a football match wins, loses or draws. The study was successful in building a model that predicted if the home team in a football match wins, loses or draws. The model proposed in the study was an ensemble model that contained machine learning models like DT, RF and Gradient Boosting. The data used in the study was the data associated with football matches obtained from an online source. The data was pre-processed, label encoded and the best 20 features in the data were determined. The class imbalance in the dataset was successfully solved using SMOTE. The ensemble model performed prediction based on voting in which the predictions by the RF, DT and Gradient Boosting models were considered for voting. The model that predicted the results of the football match was successfully built and its performance was assessed by determining its performance metrics. It was seen that the model built here achieved a value of 99.5% for the accuracy, precision, recall and f1-score. The ensemble model was successfully implemented as a desktop application and the desktop application successfully predicted the outcome of the football matches. The main aim of the study was successfully achieved and the research questions of the study performed here were also successfully answered, the questions were:

- Will the ensemble model proposed here be able to predict the outcome of football matches?
  - The ensemble model built in the study was able to successfully predict the results of football matches.
- Which are the best features associated with the data in the dataset?

- The best 20 features from the dataset were found using the Chi-square feature selection algorithm.

The model that predicted the results of football matches and the desktop application was built using Python.

In the future, the data produced in a longer timeframe can be used to improve the study. In the study performed here currently the voting based method is used for prediction in the ensemble model. In future studies, other kinds of techniques for prediction can be implemented in the ensemble model for prediction.

References

[1] 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), pp.741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003.

[2] Bahzad, J. and Adnan, M.A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2, pp.20-28. https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning.

[3] Bogicevic, M. (2018). Football: A Data-Driven Evolution. Medium. https://blog.cambridgespark.com/how-data-science-is-changing-the-world-of-football-64df28f36996.

[4] Bui, Q.-T., Chou, T.-Y., Hoang, T.-V., Fang, Y.-M., Mu, C.-Y., Huang, P.-H., Pham, V.-D., Nguyen, Q.-H., Ngoc, T., Pham, V.-M. and Meadows, M.E. (2021). Gradient Boosting Machine and Object-Based CNN for Land Cover Classification. *Remote Sensing*, 13(14), pp.2709–2709. https://doi.org/10.3390/rs13142709.

[5] Bunker, R. and Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review. *Journal of Artificial Intelligence Research*, 73, pp.1285–1322. https://doi.org/10.1613/jair.1.13509.

[6] Bunker, R.P. and Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1), pp.27–33. https://doi.org/10.1016/j.aci.2017.09.005.

[7] Carloni, L., De Angelis, A., Sansonetti, G. and Micarelli, A. (2021). A Machine Learning Approach to Football Match Result Prediction. *HCI International 2021 - Posters*, pp.473–480: https://doi.org/10.1007/978-3-030-78642-7_63.

[8] Chen, R.-C., Dewi, C., Huang, S.-W. and Caraka, R.E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1). https://doi.org/10.1186/s40537-020-00327-4.

[9] Choi, B., Lee Kien Foo and Chua, S.-L. (2023). Predicting Football Match Outcomes with Machine Learning Approaches. *Mendel*, 29(2), pp.229–236. https://doi.org/10.13164/mendel.2023.2.229.

[10] Constantinou, A. and Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, 124, pp.93–104. https://doi.org/10.1016/j.knosys.2017.03.005.

[11] Couronné, R., Probst, P. and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1).https://doi.org/10.1186/s12859-018-2264-5.

[12] Etuk, R., Xu, T., Abarbanel, B., Potenza, M.N. and Kraus, S.W. (2022). Sports betting around the world: A systematic review. *Journal of Behavioral Addictions*, 11(3). https://doi.org/10.1556/2006.2022.00064

[13] F., (2021). *16 Eye-Opening Sports Betting Statistics for a Luckier 2021*. https://playtoday.co/blog/sports-betting-statistics/

[14] Feng, W., Gou, J., Fan, Z. and Chen, X., 2023. An ensemble machine learning approach for classification tasks using feature generation. *Connection Science*, 35(1). https://doi.org/10.1080/09540091.2023.2231168.

[15] football-data.co.uk (2024) *England Football Results Betting Odds | Premiership Results & Betting Odds*. https://www.football-data.co.uk/englandm.php

[16] Gitnux.org. (2023). *Sports Betting Industry Statistics [Fresh Research]*. https://gitnux.org/sports-betting-industry-statistics/ (Accessed 28 March 2024).

[17] Gomez-Gonzalez, C. and Del Corral, J., 2018. The betting market over time: overround and surebets in European football. *Economics and Business Letters*, 7(4), pp.129-136. https://doi.org/10.17811/ebl.7.4.2018.129-136

[18] Gong, C. and Gu, L., 2016. A Novel SMOTE-Based Classification Approach to Online Data Imbalance Problem. *Mathematical Problems in Engineering*, [online] 2016(5685970), pp.1-14. https://doi.org/10.1155/2016/5685970

[19] Herbinet, C. (2018). *Individual project report department of computing predicting football results using machine learning techniques*. https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf

[20] Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C. and Meyer, T., 2019. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6), pp.798-817. https://doi.org/10.1177/1747954119879350

[21] Huynh-Cam, T.-T., Chen, L.-S. and Le, H., 2021. Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms*, 14(11), p.318: https://doi.org/10.3390/a14110318

[22] Jawade, I., Jadhav, R., Vaz, M.J. and Yamgekar, V., 2021. Predicting Football Match Results using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*. https://www.irjet.net/archives/V8/i7/IRJET-V8I730.pdf

[23] Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S. and Hussain, S., 2023. Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, 13(6), p.4006. https://doi.org/10.3390/app13064006

[24] Khaire, U.M. and Dhanalakshmi, R., 2019. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.06.012

[25] Kotu, V. and Deshpande, B., 2019. Classification. In: *Data Science*. pp.65-163. https://doi.org/10.1016/b978-0-12-814761-0.00004-6.

[26] Linder, J. (2023). *Football Betting Statistics: Market Report & Data*. Gitnux. https://gitnux.org/football-betting-statistics/

[27] Loef, B., Wong, A., Janssen, N.A.H., Strak, M., Hoekstra, J., Picavet, H.S.J., Boshuizen, H.C.H., Verschuren, W.M.M. and Herber, G.-C.M., 2022. Using random forest to identify longitudinal predictors of health in a 30-year cohort study. *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-14632-w

[28] Lu, C.-J., Lee, T.-S., Wang, C.-C. and Chen, W.-J., 2021. Improving Sports Outcome Prediction Process Using Integrating Adaptive Weighted Features and Machine Learning Techniques. *Processes*, 9(9), p.1563. https://doi.org/10.3390/pr9091563

[29] McHugh, M.L., 2013. The Chi-square test of independence. *Biochemia Medica*, 23(2), pp.143-149. https://doi.org/10.11613/bm.2013.018

[30] Md. Johir Raihan and Abdullah-Al Nahid, 2023. Classification of histopathological colon cancer images using particle swarm optimization-based feature selection algorithm. In: *Elsevier eBooks*. pp.61–82. https://doi.org/10.1016/b978-0-323-96129-5.00012-3

[31] Miao, J. and Niu, L., 2016. A Survey on Feature Selection. *Procedia Computer Science*, 91, pp.919-926. https://doi.org/10.1016/j.procs.2016.07.111.

[32] Mohammed, A. and Kora, R., 2021. An effective ensemble deep learning framework for text classification. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.11.001

[33] Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A. and Blokland, M., 2023. The effects of data balancing approaches: A case study. *Applied Soft Computing*, 132, p.109853. https://doi.org/10.1016/j.asoc.2022.109853

[34] Ozanian, M. (2023) 'World's Most Profitable Sports Teams: Cowboys Banked $1.2 Billion Over The Past Three Years', *Forbes*, 2 June. https://www.forbes.com/sites/mikeozanian/2023/06/02/worlds-most-profitable-sports-teams-cowboys-banked-12-billion-over-the-past-three-years/?sh=1b1f3f5e6e44.

[35] Patel, H.H. and Prajapati, P. (2018) 'Study and Analysis of Decision Tree Based Classification Algorithms', *International Journal of Computer Sciences and Engineering*, 6(10), pp. 74-78. https://doi.org/10.26438/ijcse/v6i10.7478

[36] Raju, M.A., Mia, M.S., Sayed, M.A. and Uddin, M.R. (2020) 'Predicting the Outcome of English Premier League Matches using Machine Learning', *IEEE Xplore*. https://doi.org/10.1109/STI50764.2020.9350327

[37] Rodrigues, F. and Pinto, Â. (2022) 'Prediction of football match results with Machine Learning', *Procedia Computer Science*, 204, pp. 463–470. https://doi.org/10.1016/j.procs.2022.08.057

[38] S. A. A, B. and H. F, E. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.1708.

[39] Santangelo, O.E., Gentile, V., Pizzo, S., Giordano, D. and Cedrone, F. (2023) 'Machine Learning and Prediction of Infectious Diseases: A Systematic Review', *Machine Learning and Knowledge Extraction*, 5(1), pp. 175-198. https://doi.org/10.3390/make5010013

[40] Sjöberg, F. (2023) *Football Match Prediction Using Machine Learning*. Åbo Akademi University, Finland Faculty of Science and Engineering (FNT). https://www.doria.fi/bitstream/handle/10024/187628/sjoberg_fredrik.pdf?sequence=2&isAllowed=y.

[41] Skorikov, M. and Momen, S. (2020) 'Machine learning approach to predicting the acceptance of academic papers', *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. https://doi.org/10.1109/iaict50021.2020.9172011

[42] Spencer, R., Thabtah, F., Abdelhamid, N. & Thompson, M. (2020) 'Exploring feature selection and classification methods for predicting heart disease', DIGITAL HEALTH, 6. https://doi.org/10.1177/2055207620914777

[43] Stefano, E., Farroco, L. de O., Lima, G.B.A., Sant'Anna, A.P., Gavião, L.O. and Principe, V.A. (2020) 'Decision trees for the prediction of outcome of soccer games - historical data analysis', *Brazilian Journal of Development*, 6(1), pp. 4719–4732. https://doi.org/10.34117/bjdv6n1-339

[44] Stübinger, J., Mangold, B. and Knoll, J. (2019) 'Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics', *Applied Sciences*, 10(1), p. 46. https://doi.org/10.3390/app10010046

[45] Tercan, H. and Meisen, T. (2022) 'Machine learning and deep learning based predictive quality in manufacturing: a systematic review', *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-022-01963-8

[46] Tiwari, E., Sardar, P. and Jain, S., 2020. Football match result prediction using neural networks and deep learning. *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. https://doi.org/10.1109/icrito48877.2020.9197811

[47] Uddin, S., Ong, S. and Lu, H. (2022) 'Machine learning in project analytics: a data-driven framework and case study', *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-19728-x

[48] Usman, H., Maitama, J.Z., Mohammed, M. and Raj, R.G., 2022. Predicting the outcomes of football matches using machine learning approach. *Communications in Computer and Information Science*, pp.92-104. https://doi.org/10.1007/978-3-030-95630-1_7

[49] Vaidya, S., Sanghavi, H. and Gevaria, K., 2016. Football Match Winner Prediction. *International Journal of Computer Applications*, 154(3). https://d1wqtxts1xzle7.cloudfront.net/72051622/ijca2016912066-libre.pdf?1633854276=&response-content-disposition=inline%3B+filename%3DFootball_Match_Winner_Prediction.pdf&Expires=1715439799&Signature=Y9B~UYOWb-q4nDXM37PDO1-Zlac-bJ9TD83aDgo0EpSo0SZab-CRYbutH9OhIcJUAtHMY3EgsQalzQrazRPcWVbe6bbhxqvq9Q6dljHret9WutLkMkxOmK84ft239atVw-Pqpw3wOOFZn~3D8j8bECuePd9EJpCatR4De0DxOar0wkixQ0a2UPMkKsfzw3y~EetgLfIzBRiWjB9efGW7tC-rYtYrsKDWjB3cFHgKul9hKEebSrcKDVXn2TRCoIu0PXWYiRDY9R-RxXPxgJI~Or8XhF0wxZaiFVBKbu8UoGoI7TUVY5uGK9FKUyKNtkhJQPnK-PES4REqVYbaVIelEg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

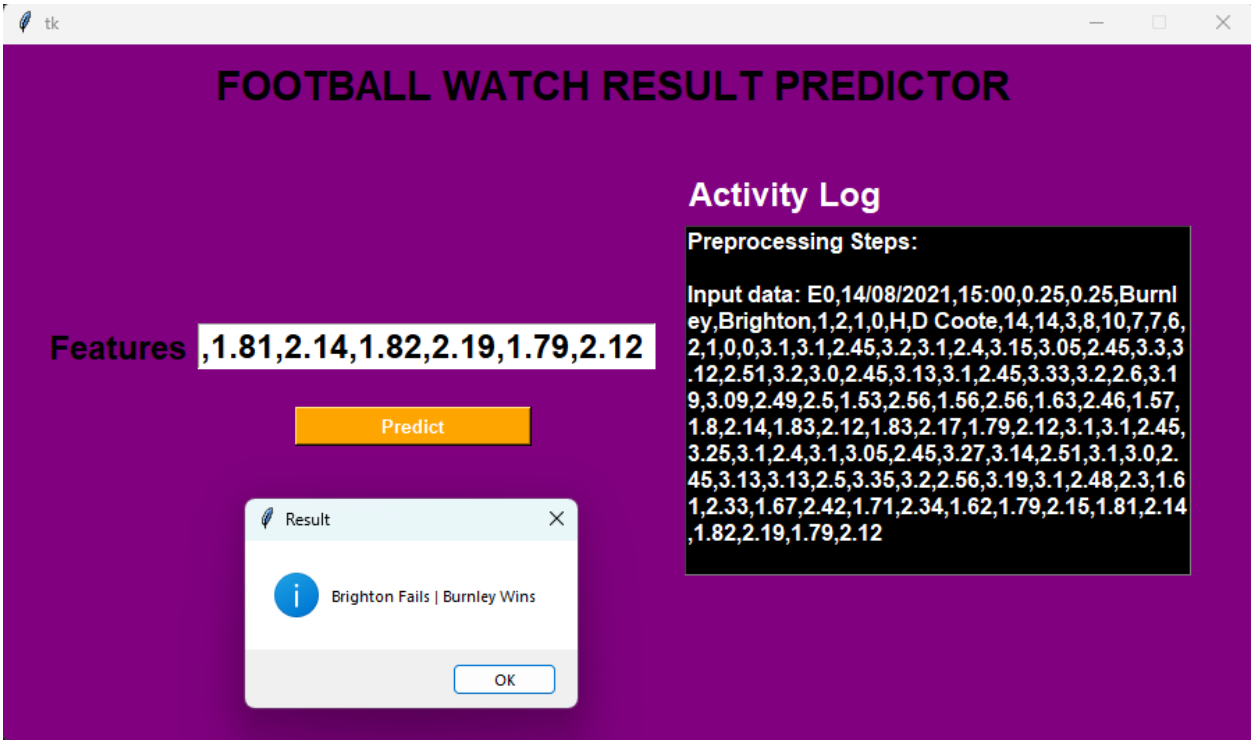
[50] Saiedy, Y., Qachmas, H., and Faqiri, A., 2020. Predicting EPL football matches results using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 5(3), pp.83–91. https://doi.org/10.33564/ijeast.2020.v05i03.013

[51] Zaveri, N., Tiwari, S., Shinde, P., Shah, U. and Kumar Teli, L., 2018. Prediction of Football Match Score and Decision Making Process. *International Journal on Recent and Innovation Trends in Computing and Communication*, 6(2). https://d1wqtxts1xzle7.cloudfront.net/56704483/1520497925_08-03-2018-libre.pdf?1527843802=&response-content-disposition=inline%3B+filename%3DPrediction_of_Football_Match_Score_and_D.pdf&Expires=1715179126&Signature=BOZa9LSaFHv01-vq-U1InHZdS9bW0omEI3DPJBEsVmTyBWXQrovqB055mLTSD2wa3jPk46MDxD9U49xns4MPseb1Yfzp-tbf7ZM8EWgsrzWuIK6VundY~Vg3c947qNmaEqOJnHMaeytwF9iQrMALH4uWD6jvnC7JftZd6UzchkFdYjg7MvNP4AyIAyTN3uC6DzL2Fiwvc5IggseBNsJ~lfWq0V2~J2QcGbshwZmYts9K8sIbOm~Xh8HA8~KZidOmhzUejbZQKRefbB8AsMEeXq~xX~KuycxAMVMKX1vUziYqDgDPz7caYp4n8PY8NTtKMPsWeTaamCxLMtthaPyBEA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Appendices

Appendix A

Predicted Output of the result:


tk
FOOTBALL WATCH RESULT PREDICTOR
Activity Log
Preprocessing Steps:
Input data: E0,14/08/2021,15:00,0.25,0.25,Burnley,Brighton,1,2,1,0,H,D Coote,14,14,3,8,10,7,7,6,2,1,0,0,3.1,3.1,2.45,3.2,3.1,2.4,3.15,3.05,2.45,3.3,3.12,2.51,3.2,3.0,2.45,3.13,3.1,2.45,3.33,3.2,2.6,3.19,3.09,2.49,2.5,1.53,2.56,1.56,2.56,1.63,2.46,1.57,1.8,2.14,1.83,2.12,1.83,2.17,1.79,2.12,3.1,3.1,2.45,3.25,3.1,2.4,3.1,3.05,2.45,3.27,3.14,2.51,3.1,3.0,2.45,3.13,3.13,2.5,3.35,3.2,2.56,3.19,3.1,2.48,2.3,1.61,2.33,1.67,2.42,1.71,2.34,1.62,1.79,2.15,1.81,2.14,1.82,2.19,1.79,2.12
Features
,1.81,2.14,1.82,2.19,1.79,2.12
Predict
Result
Brighton Fails | Burnley Wins
OK