PREDICTING ENGLISH PREMIER LEAGUE MATCHES USING MACHINE LEARNING AND CONDITIONAL PROBABILITY

MALAMBO MUTILA

A Final Year Research Project submitted in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science

ZCAS University

2023

DECLARATION

Name: Malambo Mutila

Student Number: 202202933

I hereby declare that this final year research project is the result of my own work, except for quotations and summaries which have been duly acknowledged.

Plagiarism check: %

Signature: Date: 31 December 2023

Supervisor Name: Dr. Aaron Zimba

Supervisor Signature:

Date: 31 December 2023

PREDICTING ENGLISH PREMIER LEAGUE MATCHES USING MACHINE LEARNING AND CONDITIONAL PROBABILITY

ABSTRACT

This research addresses the continuous challenge of accurately predicting football match outcomes, which is crucial for the sports betting sector's profitability and reliability. Considering the complex nature of factors influencing results, the study employs a quantitative approach and integrates Bayes' conditional probability theory using Gaussian Naïve Bayes. By leveraging English Premier League data from 11 complete seasons and half a season, the developed model achieves a training accuracy of 87% and an average testing accuracy of 85%. Comparative analysis with existing studies reveals competitive performance, albeit trailing certain advanced models. Despite the need for further refinement, the model offers a profitable avenue for betting markets, emphasizing the importance of ongoing enhancements in feature engineering. Overall, this research contributes to the field by providing a robust predictive model with potential implications for both bookmakers and punters.

Keywords: Naïve Base, Probability, English Premier League, Football, Betting

ACKNOWLEDGEMENT

Sincere appreciation to my mother, Esther Chimimba, for always being an inspiration to desire to learn more. I love you mom.

Special thanks to my supervisor for his guidance and selflessness.

THANK YOU.

DEDICATION

I dedicate this to all data lovers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN - Artificial Neural Networks

EPL - English Premier League

LSTM - Long Short-Term Memory Network

ML - Machine Learning

RBF - Radial Basis Function Kernel

SGD - Stochastic Gradient Descent

SVM - Support Vector Machines

PARX - Poisson AutoRegression with eXogenous covariates

ROI - Return on investment

TAN - Tree Augmented Naive Bayes

GBN-HC - General Bayesian Networks with Hill Climbing algorithm

OPERATIONAL DEFINITIONS

Bookmaker (bookie) – Betting company that facilitates gambling on sports events.

Punters (bettors) – Individuals or groups that make gamble on sports events on platforms provided by the betting companies (bookmakers/bookies).

CHAPTER 1

INTRODUCTION

1.1 Background to the study

Machine learning (ML) is the bedrock of adaptive systems. It enables computers to learn from experience, examples and analogies in order to continuously improve the performance of intelligent systems (Spearman, 2018; Michael, 2005). News websites use machine learning to suggests news articles to their readers, journals do the same for research papers, Spotify, Apple Music and YouTube implement machine learning to recommend music artists and songs to listen to (Carloni et al., 2021). Netflix recommends movies to its users using machine learning, and eBay and Amazon use machine learning to indorse products for purchase (Bologna et al., 2013). Therefore, it is only right that bookmakers and punters use machine learning in sports analytics.

The application of machine learning combines computer science, statistics and probability. One of the probability theories of use in machine learning is Bayes' theorem, a revolutionary probability theory that became the cornerstone of statistics in the 18th Century. The fundamental concept behind Bayes' theorem is the idea of conditional probability, which measures the likelihood of an event 'A' happening given that another event 'B' has already happened (Razali et al., 2018; Rahman et al., 2018). It is this mathematical framework that forms the basis of this study on the outcome of football matches in the English Premier League (EPL).

The English Premier League is broadcast to over 643 million homes in more than 212 regions and with a TV audience of at least 4.7 billion people worldwide, making it the world's most watched sports league (Baboota and Kaur, 2019). In 2021/2022 Premier League clubs generated a record revenue of 5.5 billion pounds (Deloitte, 2023). On the betting market the previous year, according to the Statista Research Department (2023), bets on Premier League matches were the highest in European football in 2020/2021 generating more than 68.5 billion euros worldwide.

The popularity of the English Premier League continues to grow each year. The enormous viewership, combined club revenue and betting market give modelling a predictive system for English Premier League matches high significant economic value and growing academic interest. This study aims to develop a predictive model for English Premier League matches that uses Bayes' conditional probability theory and machine learning to predictive the outcome of matches in an ongoing Premier League season.

1.2 Problem Statement

In an ideal environment where all factors affecting football match outcomes are finite and known, predictive models would achieve 100% accuracy by studying the impact of each factor on the coaches, players, teams, match officials, spectators, etc. and the match result itself. However, predicting outcomes of football matches continues to be a challenge due to the numerous known and unknown factors that affect match outcomes such as team morale, player form, skills, current score, weather, team composition, formation, etc. (Arabzad et al., 2014). There is also the possibility of overlooking significant factors that contribute to match outcomes and the likelihood of incorrectly applying the known factors when making predictions. Another challenge is the high competitive nature of the English Premier League and the high occurrence of weaker teams outscoring stronger teams, thereby causing upsets (Baboota and Kaur, 2019).

The solution to this array of problems is to use conditional probability to model the impact of each factor affecting the outcome of a football match to other factors present and available to us. For example, instead of only looking at how many games a team has won to try and predict if the team will win the next game, we can additionally look at how many games that team won given that the team’s key players were in the starting line-up. Consequently, “key” players would have to be determined by their contribution to the team not by their commercial value, popularity or performance in the previous season.

Not resolving this issue would have serious financial consequences. Both bookmakers and punters suffer significant monetary losses every day as a result of inaccurate football predictions. The inability to calculate odds that appropriately reflect the possibilities if various in-match events may result in bookmakers earning less money. Conversely, punters who place bets based on inaccurate forecasts risk suffering substantial financial losses. Additionally, the overall quality of the sports betting experience is diminished by the inaccuracy of match predictions, thereby deterring fans and lowering market participation. Therefore, overcoming the difficulty of accurately predicting football match results is crucial to improving the profitability and dependability of the sports betting sector to offering useful insights for all involved parties.

1.3 Aim

To develop a profitable predictive model that uses machine learning and Bayes' conditional probability to predict English Premier League results for sports betting.

1.3 Objectives of the study

The objectives of the study were:

1. To determine the predictive models and algorithms used in predicting football match outcomes.
2. To develop a machine learning model using conditional probability in order to predict outcomes of English Premier League matches.
3. To evaluate and validate the accuracy of the predictive model.

1.4 Research Questions

1. What predictive models and algorithms are used in predicting football match outcomes?
2. How can conditional probability be effectively employed to predict outcomes of English Premier League matches?
3. What metrics can be used to evaluate and validate the accuracy of the predictive model?

1.5 Scope and Limitation

The scope of this research encompasses an in-depth analysis of predictive models and algorithms commonly employed in predicting football match results, with a particular emphasis on their application in the context of the EPL.

The following limitations are appropriate to this study:

1. The study does not cover all potential factors influencing football match outcomes, as the complexity of the game involves a multitude of variables, some of which may be challenging to quantify or incorporate into the model.
2. The predictive model is designed to work within the framework of the English Premier League and may or may not be directly transferable to other football leagues or sports.
3. The model's predictive accuracy may be influenced by the quality and availability of historical data, which can vary in reliability and detail.

1.6 Significant of the Project

By shedding light on impact of overlooked factors on the final match outcome, this study will not only advance our understanding of the English Premier League but also provide valuable insights that can be used to guide sports analytics in Zambia.

Furthermore, the model could extend its applicability beyond the realm of football, offering promise in stock trading, forex, human behavioural changes, customer behaviour and online recommendation systems. This study's multi-faceted impact has the potential to reshape predictive modelling and decision-making processes across a spectrum of industries and domains.

1.7 Preliminary Sections of the Project Report

The remainder of the report is organised as follows: Chapter 2, the Literature Review, covers a broad review of the topic, a critical review of related works, and our proposed model, concluding with a comparison to related works. Chapter 3, Methodology, discusses the research design, adopted methods with justifications, the association of research methods with the project, research data and datasets, data collection methods, data analysis techniques, and ethical concerns. Chapter 4, Data, Experiments, and Implementation, addresses appropriate modelling, techniques, algorithms, and the main functions and models used to achieve the research objectives. Chapter 5, Results and Discussions, presents and analyses the research results, compares them to related work, and explores the implications. Finally, Chapter 6, Summary and Conclusion, summarises the main findings, contributions to the body of knowledge, acknowledges research limitations, and outlines potential directions for future work.

CHAPTER 2
LITERATURE REVIEW

2.1 General Background

The landscape of predictive modelling in sports, particularly football, has witnessed a paradigm shift with the integration of machine learning (ML) techniques. As outlined in Chapter 1, machine learning has become integral to various domains, ranging from news recommendations to music suggestions. Prior to current advancements in machine learning, predictive modelling in sports analytics relied heavily on statistical methods to analyse player performance, team dynamics, and match outcomes (Koopman and Lit, 2015; Nyquist and Pettersson, 2017). However, over time, sports analytics has evolved to incorporate machine learning techniques (Igiri, 2015; Vaidya, Sanghavi and Gevaria, 2016) . This chapter discusses the evolution of machine learning algorithms, models and algorithms used in the prediction of football matches, evaluation and validation metrics for these models and the conceptual framework that guided the proposed model.

2.2 Machine Learning in Predictive Sports Analytics

The integration of machine learning algorithms has enabled sports analysts to uncover patterns, trends, and insights from large datasets, contributing to more accurate predictions and informed decision-making in sports (Ren and Susnjak, 2022; Razali et al., 2018; Angelini and De Angelis, 2017). An example of such sports where predictive analysis has been studied is basketball (Smith et al., 2015; Cervone et al., 2016; Deshpande and Lam, 2018). Researchers have modelled different areas of the game such as shooting effectiveness using spatial modelling (Goldsberry, 2012), expected number of points per shot depending on distance and location (Shortridge et al., 2014), players' shooting habits (Miller et al., 2014), and players' contribution to the match outcome (Jensen, 2016).

Predictive analysis in football has mainly been used with the aim of predicting the number of goals scored in a football match and the outcomes of football matches (Ulmer, Fernandez and Peterson, 2013; Angelini and De Angelis, 2017; Constantinou, 2019; Alten-Ronæss, 2021; Ren and Susnjak, 2022). For example, Angelini and De Angelis (2017) used Poisson AutoRegression with eXogenous covariates (PARX) to predict the number of goals scored in a match and Alten-Ronæss (2021) looked at the use of a generalised Poisson model to predict English Premier League (EPL) results. Other researchers such as Tax and Joustra

(2015), Johnson and Brown (2018), and Nyquist and Pettersson (2017) focused on the application of neural networks in predicting match outcomes in football.

2.2.1 Evolution of Machine Learning Algorithms

Machine learning algorithms progressed from simplistic methods such as Linear Regression (Prasetio et al., 2016), K-nearest neighbors (Yezus, 2014; Brooks et al., 2016), and Decision Trees (YILDIZ, 2020) to advanced techniques with the adoption of Random Forests (Groll et al., 2019), Support Vector Machines (Hubáˇcek et al., 2019), Artificial Neural Networks (Arabzad et al., 2014; Bunker and Thabtah, 2019), and Boosting (Hubáˇcek et al., 2019). More recently, deep learning methods have gained prominence, employing Convolutional Neural Networks (Hsu, 2021) and LSTM models (Zhang et al., 2021; Malini and Qureshi, 2022). This evolution signifies a significant shift in the approach to sports analytics, marking a departure from traditional statistical methods toward more sophisticated and data-driven ML techniques.

2.3 Critical Review of Related Work

Constantinou, Fenton and Neil (2012) developed a Bayesian Network model called pi-football using key features such as strength, form, psychology and fatigue. The model aimed to create a profitable betting strategy and yielded predictions with a maximum profit levels of 8.86% to 35.63%, dependent on the bookmakers' odds.

Yezus (2014), targeting an accuracy of 70% had the aim of beating bookmakers in the sports betting market. The author applied KNN, random forest, logistic regression, and SVM on nine features and 640 observations in a three-class classification problem (home win, draw, away win). However, the study achieved an accuracy of 63.4% placing it below its intended target of 70%.

Ren and Susnjak (2022) discovered that predictions made over a large time span, for example, predicting the results of a game over 10 seasons, yield less accurate models than predicting the results of a game over one season. Ren and Susnjak (2022) hypothesized that the machine learning method used does not have a significant impact on the accuracy of the machine learning model. The authors used eight different models and their best result was a 70% accuracy using CatBoost. This was against their hypothesis as each machine learning model did significantly impact accuracy. The study’s use of data from 2001 to 2021 disadvantaged their models as the set of features for real-time data collected over this period changed significantly in 2006 when Opta came into play (Zaveri et al., 2018; Robberechts, Haaren and Davis, 2021).

Rahman et al. (2020) developed a deep learning framework using artificial neural networks (ANN) and long short-term memory (LSTM) networks. The study trained its using national team football results from 1872 to 2018 and ran predictions on the 2018 FIFA world cup group stage matches. An accuracy of 63.3% in predicting group stage matches was achieved. The authors faced challenges in the lack of accurate data and recommend the use of more relevant features.

Ulmer, Fernandez and Peterson (2013) implemented a three-class classification problem to predict the outcomes of English Premier League matches using machine learning. The study used gaussian naive bayes, multinomial naive bayes, a hidden Markov model, support vector machines (SVM). The models were evaluated using error rate and ROC curves. The best-performing model achieved an accuracy of 52%. The study under-predicted draws and their accuracy rates (50-52%) were relatively low. The authors believe this accuracy could be improved by using statistics and elements of previous matches such as corner kicks.

2.4 Comparison with related works

Owramipur, Eskandarian and Mozneb (2013) developed a Bayesian model with 92% in predicting matches in the 2008-2009 season in the Spanish La Liga. However, the authors focused on Barcelona’s 38 matches only and it is unknown how their model performs for all teams in the league.

Moura, Martins and Cunha, (2014) conducted an analysis of football statistics using principal component and cluster analyses of 2006 World Cup matches. Their model scored an accuracy of 70.3%. The results of the study could be improved by training and test on more recent matches as football advances every year and players change.

Igiri and Nwachukwu (2014) developed a predictive model using knowledge discovery in databases (KDD), artificial neural networks (ANN) and logistic regression (LR). The model produced 85% and 93% prediction accuracy when ANN and LR techniques were applied respectively. These results were later tested against a support vector machine (SVM) model and performed better as the SVM had an accuracy of only 53.3% . However, only the ANN could predict win, lose or draw, as the logistic regression could only predict a win or loss result.

Razali et al. (2017) tackled the challenge of predicting football match outcomes in the English Premier League (EPL) through the utilisation of Bayesian Networks (BNs). The authors used data from three consecutive seasons (2010-2011, 2011-2012, and 2012-2013) and implemented K-fold cross-validation to assess the performance of their model. The model achieved multi-class predictive accuracy of 75.09%. However, the study's reliance on data from a specific time frame (2010-2011 to 2012-2013) limited the generalisability of the model to current conditions, as team dynamics, player compositions, and other contextual factors in football can evolve over time.

Rahman et al. (2018) classified football match outcomes using only Bayesian methods which included Naive Bayes (NB), Tree Augmented Naive Bayes (TAN), and General Bayesian Network (K2). The study focused on three seasons (2014-2015, 2015-2016, and 2016-2017) and reported that the highest predictive accuracy of 90.0% on average was achieved by TAN, outperforming NB and K2. However, similar to previous studies, the applicability of the model to current EPL conditions may be limited due to its reliance on historical data within a specific timeframe.

In the exploration of the prediction of football match score and the decision-making process, Zaveri et al. (2018) examined 12 different attributes using various machine learning techniques, including Logistic Regression (LR), Random Forest (RF), Artificial Neural Network (ANN), Linear Support Vector Machine (Linear SVM), and Naive Bayes (NB). The study, spanning five years from La Liga seasons 2012-2013 to 2016-2017, reported the highest accuracy of 71.63% with LR. Notably, the accuracy specified in the table represents the average across home win, away win, and draw classifications. While the paper provides valuable insights into the prediction of football match scores, it is crucial to recognize potential limitations, such as the specific attributes chosen and the applicability of the model to different leagues or timeframes.

Stübinger, Mangold and Knoll (2019) focused was on minimising betting losses by predicting match results through the analysis of player characteristics. The authors incorporated attributes such as age, height, weight, ball skills, passing, shooting, defence, physical, and mental capabilities. Employing machine learning techniques such as Boosting, Random Forest, Support Vector Machine, and Linear Regression, the study opted for regression trees instead of classification. Noteworthy performance metrics included a Random Forest accuracy of 81.26% and an Ensemble accuracy of 81.77%. The research spanned five leagues from 2006 to 2018, offering a comprehensive timeframe. However, the study could benefit from addressing potential limitations, such as the generalisation of results across diverse betting contexts and the dynamic nature of player performance.

Razali et al. (2021) developed multiple Bayesian models to predict the outcome of major European league matches. Their models included Naive Bayes (NB), Tree Augmented Naive Bayes (TAN), and two General Bayesian Networks (GBN); GBN-K2 and GBN-HC. Notably, the NB produced the lowest accuracy at 72.78% while GBN-HC outperformed other models, achieving an accuracy of 92.01% while TAN scored 91.86%. The findings underscore the effectiveness of Bayesian models, particularly GBN-HC, in football match prediction. However, further exploration of TAN's applicability and its ability to handle minimal attribute dependencies in small datasets would contribute to the depth of the research.

Haruna et al. (2021) addressed the challenge of predicting football match outcomes, acknowledging a gap in specific features and classifiers that lead to good accuracy. They considered factors such as refereeing subjectivity, key players, home team advantage, coaching strategy, field dimension, and distance between teams. Employing Sequential Forward Selection (SFS) for feature selection, the study applied various classifiers including Logistic Regression, SVM, Random Forest, K-NN, and Naïve Bayes, with K-NN emerging as the top-performing classifier. The model achieved an accuracy of 83.95% in multi-class classification across two seasons. Despite this success, the study acknowledged the challenge of interpreting factors like refereeing subjectivity and other interactive elements. Additionally, the generalisability of the model to different leagues and seasons warrants consideration.

Muszaidi et al. (2022) employed a deep learning approach using Multilayer Perceptron (MLP) and Dense Neural Network (DNN) for the classification of English Premier League (EPL) match outcomes based on full-time results. The study focused on multi-class classification, comparing the performance of MLP and DNN. The results indicated that MLP outperformed the standard DNN, achieving an accuracy of 78.42% compared to 67.63% for the latter. Considerations may be given to potential limitations such as the interpretability of the deep learning models and the generalisability of results across different seasons or leagues.

Table 2.1: Summary of Related work using Bayesian Models

Study	Competition	Best Model/Algorithm	Evaluation and Validation Method	Performance
(Owramipur, Eskandarian and Mozneb, 2013)	Spanish La Liga (Barcelona only)	Bayesian network	Accuracy	92%
(Moura, Martins and Cunha, 2014)	World Cup 2006	Principal Component and Cluster Analyses	Accuracy	70.3%
(Igiri and Nwachukwu, 2014)	English Premier League	Artificial Neural Networks	Accuracy	85%
(Razali et al., 2017)	English Premier League	Naive Bayes	Accuracy	75.09% (3 - average)
(Rahman et al., 2018)	English Premier League	Tree Augmented Naive Bayes (TAN)	Accuracy	90.0% (3 - average)
(Zaveri et al., 2018)	Spanish La Liga	Logistic Regression	Accuracy	71.63%
(Stübinger, Mangold and Knoll, 2019)	England, France, Germany, Italy, Spain	Random Forest	Accuracy	81.26%
(Razali et al., 2021)	English Premier League	General Bayesian Networks + Hill Climbing algorithm (GBN-HC)	Accuracy	92.01%
(Haruna et al., 2021)	English Premier League	K-NN classifier	Accuracy	83.95%
(Muszaidi et al., 2022)	English Premier League	Multilayer Perceptron (MLP)	Accuracy	78.42%

***(3 - average) average across three seasons

2.5 Conceptual framework

This study adapted the SRP-CRISP-DM framework proposed by Bunker and Thabtah (2019). The SRP-CRISP-DM framework (Figure 2.1) focuses on result prediction for team based sports as opposed to individual/singles sports. The SRP-CRISP-DM framework consists of six steps, however, the conceptual framework guiding this study (Figure 2.2) implemented only the first five crucial steps solely for the development of the predictive model.

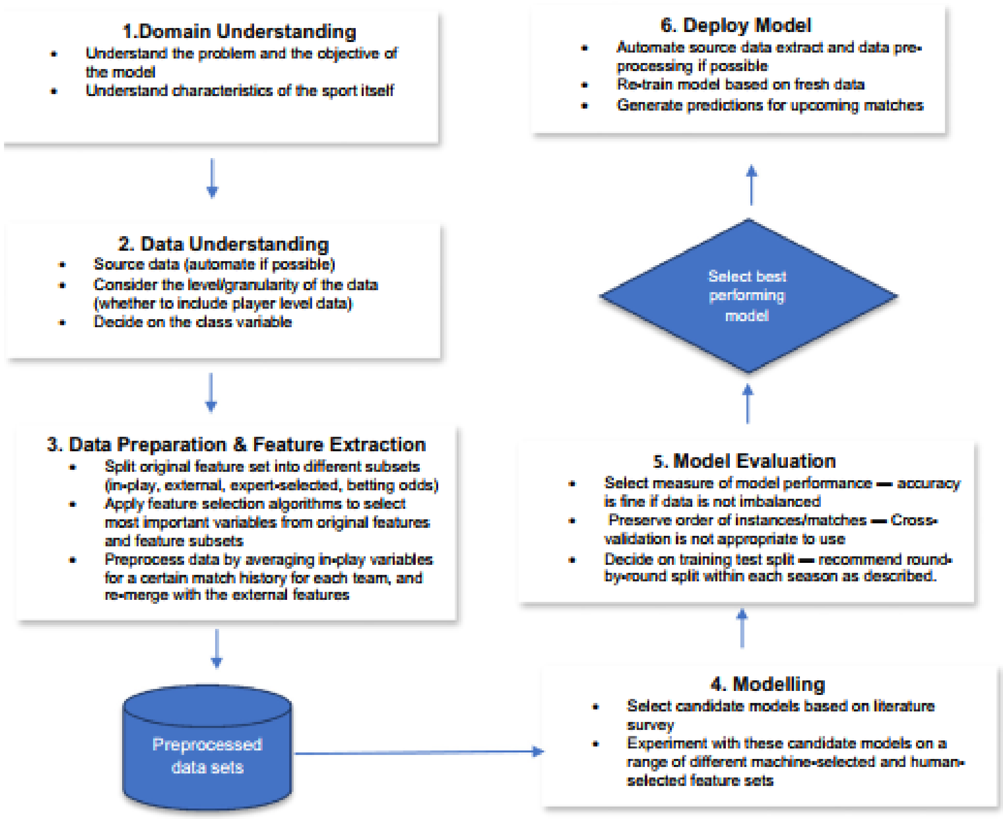

1.Domain Understanding
Understand the problem and the objective of the model
Understand characteristics of the sport itself
2. Data Understanding
Source data (automate if possible)
Consider the level/granularity of the data (whether to include player level data)
Decide on the class variable
3. Data Preparation & Feature Extraction
Split original feature set into different subsets (in-play, external, expert-selected, betting odds)
Apply feature selection algorithms to select most important variables from original features and feature subsets
Preprocess data by averaging in-play variables for a certain match history for each team, and re-merge with the external features
Preprocessed data sets
4. Modelling
Select candidate models based on literature survey
Experiment with these candidate models on a range of different machine-selected and human-selected feature sets
5. Model Evaluation
Select measure of model performance — accuracy is fine if data is not imbalanced
Preserve order of instances/matches — Cross-validation is not appropriate to use
Decide on training test split — recommend round-by-round split within each season as described.
Select best performing model
6. Deploy Model
Automate source data extract and data pre-processing if possible
Re-train model based on fresh data
Generate predictions for upcoming matches

Figure 2.1: Initial SRP-CRISP-DM framework (Bunker and Thabtah 2019)

2.5.1 Domain Knowledge

An in-depth exploration of the English Premier League's competitive landscape was undertaken, highlighting the challenges inherent in predicting match outcomes due to team dynamics, external influences, and the inherent randomness of the sport. The potential benefits of combining machine learning and conditional probability to improve prediction accuracy were subsequently explored.

2.5.2 EDA & Data Comprehension

A diverse range of data points encompassing historical match results, player statistics, and betting odds were gathered for further analysis. Exploratory data analysis techniques were then employed to uncover the distributions, relationships, and hidden patterns within this data. Key insights pertaining to factors influencing match outcomes and potential biases in betting odds were investigated from this comprehensive exploration.

2.5.3 Pre-processing and Feature Engineering

Prior to utilising the data in the chosen machine learning model, thorough cleaning and refinement processes were implemented to address missing values and ensure data integrity. Feature engineering techniques were subsequently employed to create and extract features that most effectively captured the essence of each match. These meticulously crafted features served as the foundation for the predictive model.

2.5.4 Modelling

Gaussian Naive Bayes, a machine learning algorithm adept at modeling binary outcomes such as wins and losses, was selected for the task of predicting match outcomes. This choice was further bolstered by the nature of Gaussian Naive Bayes being a conditional probability model. This was needed to account for the dynamic interplay of factors influencing each match. This hybrid approach combined the analytical power of machine learning with the adaptable nature of conditional probability for improved predictions.

2.5.5 Model Evaluation

A rigorous evaluation process employing metrics such as accuracy, precision, and recall was implemented to assess the model's performance on unseen data. Cross-validation techniques were utilised to ensure the model's generalisability and prevent overfitting to past results.

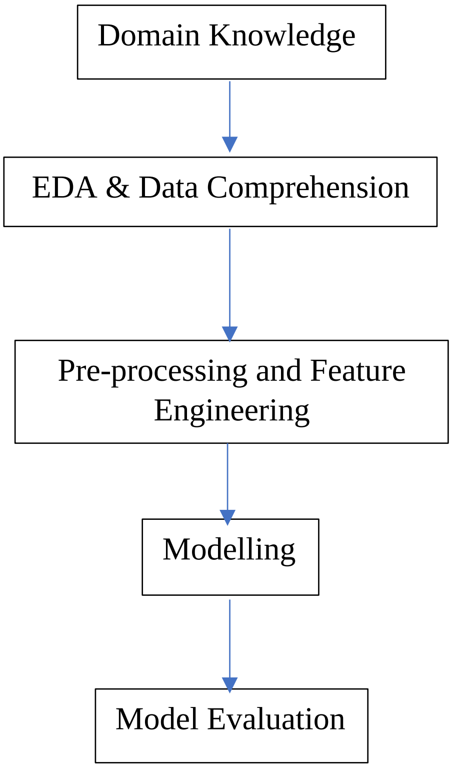


Figure 2.2: Conceptual Framework

2.6 Proposed Model

2.6.1 Conditional Probability

The study implemented a Gaussian Naïve Bayes classifier. The Naïve Bayes classifier is based on the concept of conditional probability as stated by Bayes' Theorem and makes a naive assumption that all variables are independent between predictors. Bayes' theorem calculates the posterior probability, $P(A|B)$ from P(A), P(B), and $P(B|A)$ (Equation 2.1).

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Where,

$P(A)$ is the prior probability of A.

$P(A|B)$ is the conditional probability of A given B.

$P(B|A)$ is the conditional probability of B given A.

$P(B)$ is the prior probability of data B.

2.6.2 Match Result (Double Chance)

Football results have three classes; home win, draw and away win. A number of studies that have attempted to predict match outcomes either underestimated draw (Ren and Susnjak, 2022) , neglected them (Ulmer, Fernandez and Peterson, 2013) or they did not predict them accurately (Igiri and Nwachukwu, 2014). To overcome this issue and develop a model for profitable sports betting, the study used a double chance strategy. Double chance covers two of the three possible outcomes into one outcome. This study's model (Figure 2.3) combined the home win and draw into one result, thereby creating two classes; 1 – home win or draw and 2 – away win.

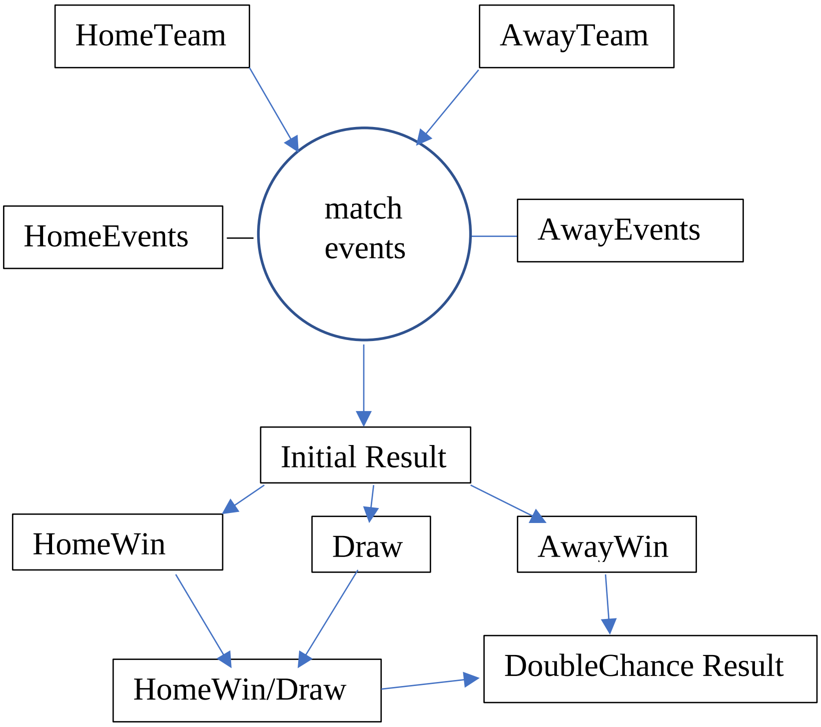


Figure 2.3: Proposed Model

2.6.2 Profitable Betting

According to Constantinou (2019), Carloni et al. (2021), and Constantinou et al. (2012), betting models are profitable if their win rate is at least 55-60%. Henceforth, to make a profitable model, the study's model had to have an accuracy of at least 60%.

2.7 Chapter Summary

This chapter explored the application of machine learning in predictive sports analytics. Various aspects, such as shooting effectiveness, points per shot, players' habits, and contribution to match outcomes, were covered. A critical review of related works summarized studies that employed Bayesian models, machine learning algorithms, and deep learning methods. Each study's competition, best model/algorithm, evaluation and validation method, and performance metrics were presented, offering a comparative analysis.

The conceptual framework of the study was introduced, adapting the SRP-CRISP-DM framework proposed by Bunker and Thabtah (2019). The framework emphasized domain knowledge, exploratory data analysis (EDA), and data comprehension, pre-processing, and feature engineering, modeling, and model evaluation. The proposed model, based on conditional probability using a Gaussian Naïve Bayes classifier, was detailed. The study addressed the challenge of predicting football match outcomes by implementing a double chance strategy, combining home win and draw into one class.

CHAPTER 3 - METHODOLOGY

3.1 Research design

The study used a quantitative research approach with an experimental research design to develop and evaluate predictive models for English Premier League (EPL) match outcomes (Kamiri and Mariga, 2021). The primary focus was on implementing Naïve Bayes' models, specifically Gaussian and Multinomial variations. The chosen research design allowed for a systematic examination of the predictive performance of these models across multiple seasons resulting into 22 models (Table 2).

To comprehensively assess the effectiveness of the models, a single-season approach was adopted. Eight complete seasons of the EPL were used for training, three were used for testing and half a season for the current 2023/2024 season was used for validation. Separate Naïve Bayes' models were developed for each of the eight seasons in the training set (Razali et al., 2017; Rahman et al., 2018). This approach facilitated a detailed evaluation of model performance over time, accounting for the dynamic nature of football competitions where team dynamics, player compositions, and external factors can evolve from season to season.

Furthermore, to explore potential enhancements in predictive accuracy, ensembles of the Naïve Bayes' models were constructed. Both Gaussian and Multinomial ensembles were created and compared against their respective single models. Ensemble methods leverage the collective wisdom of multiple models to improve overall prediction accuracy (Qamar, et al., 2016; Vashist et al., 2021), providing a robust approach to handling the inherent uncertainties in sports outcomes.

In addition to the single-season models, a concatenated approach was implemented (Ren and Susnjak, 2022). This involved developing Naïve Bayes' models using a combined dataset spanning all eight seasons. The concatenated models were then compared against their single-season counterparts to investigate the impact of incorporating historical data on prediction accuracy (Beal et al., 2021).

This research design allowed for the systematic collection and analysis of numerical data, which is essential for developing and evaluating predictive models in machine learning. The study also followed an exploratory research design to investigate and understand the predictive modelling of English Premier League football matches. The research design used elements of a longitudinal study, considering historical data and trends to inform the development and evaluation of the predictive model (Constantinou, 2019; Baboota and Kaur, 2019; Rahman et al., 2020).

Table 3.1: Model Comparison Schedule

Season	Single Model	
2012/2013	GNB_13	MNB_13
2013/2014	GNB_14	MNB_14
2014/2015	GNB_15	MNB_15
2015/2016	GNB_16	MNB_16
2016/2017	GNB_17	MNB_17
2017/2018	GNB_18	MNB_18
2018/2019	GNB_19	MNB_19
2019/2020	GNB_20	MNB_20
Ensemble		
Model	GNB13_20	MNB13_20
Concatenated	GNBconc	MNBconc
Ensemble2	GNB13_20, GNBcon	MNB13_20,MNBconc

****GNB: Gaussian Naïve Bayes model from each season.*

****MNB: Multinomial Naïve Bayes model from each season.*

****GNB12_20 & MNB13_20: Ensemble of models from 2012/2013 to 2019/2020.*

****GNBconc & MNBconc: Models developed when seasons were combined into one dataframe.*

3.2 Adopted Method

The study integrated Bayes' conditional probability theory into the predictive modelling process using Naïve Bayes models. This theoretical framework allowed the model to consider the likelihood of an event (match outcome) given the occurrence of another event (in-match factors, historical performance, etc.) (Diniz et al., 2019). By incorporating conditional probability, the model aimed to capture the interdependencies of various factors affecting football match results (Keshtkar Langaroudi and Yamaghani, 2019; Herold et al., 2019; Fahey-Gilmour, et al., 2019).

The combination of machine learning techniques and Bayes' conditional probability provides a comprehensive approach to address the complexity of football match prediction (Razali et al., 2021; Robberechts, Haaren and Davis, 2021). Machine learning algorithms offer the capacity to learn patterns and relationships from historical data, while Bayes' theorem provides a probabilistic framework for updating predictions based on new information

(Ćwiklinski, Giełczyk, and Choraś, 2021; Carloni et al., 2021). This dual approach enhanced the model's adaptability and robustness (Lee et al., 2022; Mohandas, Ahsan, and Haider, 2023).

The decision to employ a single-season approach stems from the findings of Ren and Susnjak (2022), who noted that predicting outcomes over longer time spans, such as 10 seasons, may result in less accurate models. The dynamic nature of football, influenced by factors like shots on target, referees, coaches, teams in the League and tactical shifts, suggests that focusing on individual seasons allows for a more deeper understanding of the stochastic nature of football.

The significance of utilising ensemble methods, as incorporated in this study, is supported by the work of Constantinou et al. (2012) and Stübinger, Mangold, and Knoll (2019). Constantinou et al. demonstrated that ensembles, specifically Gaussian Ensemble, can contribute to profitable betting strategies by leveraging diverse sources of information. Stübinger et al.'s exploration of player characteristics aligns with the idea that combining models can enhance predictive accuracy by capturing various facets of team dynamics.

The concatenated approach draws inspiration from the rationale presented by Rahman et al. (2018), who emphasized the importance of historical data in achieving high predictive accuracy. By amalgamating seven seasons of data, the concatenated models aim to provide a broader context for understanding match outcomes.

3.3 Justification of Adopted Method

The decision to employ multiple Bayesian model reflects a detailed approach to football match prediction, considering the details of individual seasons, long-term trends, and the potential interactions among different modelling perspectives. This strategy aligns with the dynamic and multifaceted nature of football, aiming to enhance the accuracy and adaptability of the predictive model.

3.3.1 Diverse Perspectives

Each single model for a specific season (e.g., GNB_13, MNB_13) captured the distinctions and dynamics of that particular season. Football is inherently dynamic, with team compositions, strategies, and external factors evolving from season to season. Employing individual models for each season allowed the model to learn and adapt to the unique characteristics of that period.

3.3.2 Adaptation to Seasonal Trends

Football leagues often witness trends and patterns that evolve over time. By developing separate models for each season, the approach attempted to acknowledge and adapt to the specific conditions prevalent in that particular timeframe. This was done to ensure that the model was attuned to seasonal variations, contributing to a more accurate representation of the stochastic nature of football.

3.3.3 Robustness Through Ensembles

The ensemble models (e.g., GNB13_20, MNB13_20) brought together the collective insights from multiple single models. This approach leverages the strengths of individual models while mitigating their weaknesses. Ensembles are known to enhance predictive accuracy and robustness by considering diverse perspectives and aggregating predictions, thereby minimising the impact of potential outliers or errors in individual models.

3.3.4 Long-Term Context with Concatenated Models

The concatenated models (e.g., GNBconc, MNBconc) provided a long-term perspective by combining data from multiple seasons. This aligns with the understanding that historical data contributes valuable context and trends that may influence current match outcomes. The concatenated approach aimed to capture overarching patterns and dependencies that transcend individual seasons, contributing to a complete understanding of the predictive outcomes.

3.3.5 Exploration of Model Combinations

The Ensemble2 models (e.g., GNB12_20+GNBconc, MNB13_20+ GNBconc) explored combinations of single models and concatenated models. This exploration allowed for a detailed investigation into the relationship between short-term and long-term factors. It added a layer of density to the analysis, offering insights into whether certain combinations yield superior predictive performance compared to standalone models or concatenated approaches.

3.4 Association of Research Method to Project

The adoption of predictive modelling within this study was aligned with the main objective of crafting a profitable predictive model for English Premier League matches. This choice facilitated the collection of historical match data to construct a model proficient in predicting match outcomes. The research method, executed with a machine learning framework, entailed the systematic collection and careful analysis of quantitative data. This covered historical match results and betting odds, aligning seamlessly with the selected quantitative research design for the study. This design was chosen deliberately to ensure that the predictive model adhered to a data-driven paradigm, underscoring its capacity to produce profitable numerical predictions.

The integration of predictive modelling, influenced by the principles of Bayes' conditional probability and machine learning, aligned with the pioneering work of Razali et al. (2017), Robberechts, Haaren, and Davis (2021), and Constantinou (2019). These studies collectively emphasized the efficiency of such methods in the world of football match prediction, especially when dynamic variables influence outcomes.

Bayes' conditional probability was employed thoughtfully, following the insights of Diniz et al. (2019), Keshtkar Langaroudi and Yamaghani (2019), and Herold et al. (2019). This framework was instrumental in capturing the detailed associations between factors influencing football match results. By considering the likelihood of events given the occurrence of others, the model aimed to discern patterns that are not captured in simplistic analyses.

The utilisation of machine learning algorithms, as demonstrated by Ćwiklinski, Giełczyk, and Choraś (2021) and Carloni et al. (2021), complemented Bayes' conditional probability by giving the model the capacity to distinguish patterns and relationships from historical data. This dual approach, with roots in both statistics and machine learning, ensured the adaptability and robustness of the predictive model, as corroborated by Lee et al. (2022) and Mohandas, Ahsan, and Haider (2023).

The decision to opt for a single-season approach, informed by Ren and Susnjak's (2022) insights, allowed for a granular examination of each season's peculiarities. In football, where the playing field is influenced by many factors, such an approach afforded a deeper understanding of the stochastic nature of the sport.

Ensemble methods, a key facet of this research methodology, found validation in Constantinou et al.'s (2012) demonstration of their profitability, especially the Gaussian Ensemble. The incorporation of diverse models was used to enhance the predictive accuracy and robustness of the model, echoing the findings of Stübinger, Mangold, and Knoll (2019) who researched into the effects of player characteristics on match outcomes.

The concatenated approach, inspired by Rahman et al.'s (2018) emphasis on historical data, was seamlessly integrated into the research design. This approach sought to provide a broader contextual perspective by combining eight seasons of data, a methodology deemed essential for achieving high predictive accuracy.

The meticulous alignment of the research method with the project's objectives, incorporating Bayes' conditional probability, machine learning, and strategic modelling approaches, laid the groundwork for a comprehensive, data-driven, and adaptable predictive model for English Premier League match outcomes. The chosen methodology drew from a rich review of prior research, connecting various techniques to navigate the difficulties of football match prediction.

3.5 Research Data and Datasets

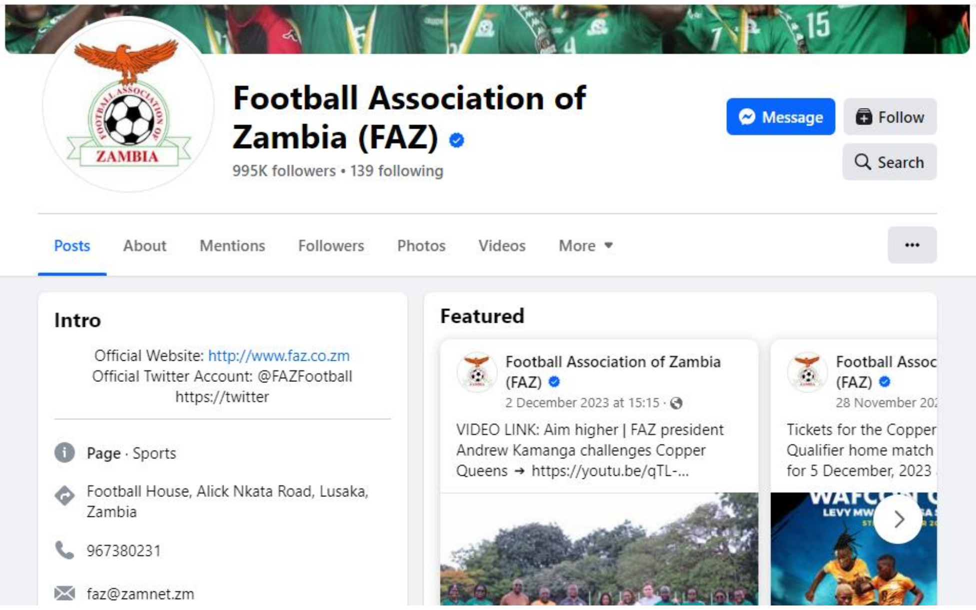
3.5.1 Choice of Football Competition

The choice of the English Premier League (EPL) over the Zambian Super League (ZSL) for this study was supported by several factors, mostly centered on data availability, depth, and the international prominence of the EPL, which provides for comparative research.

3.5.1.1 Data Availability and Quality

The English Premier League, being the most globally recognised and followed football league (Deloitte, 2023), boasts a wealth of comprehensive and high-quality data (Baboota and Kaur, 2019). Organisations such as Opta, renowned for their in-depth data collection processes, provide a rich dataset covering various facets of EPL matches. This includes not only full-time results but also detailed in-match statistics, player performances, and contextual information crucial for predictive modelling (Statista Research Department, 2023). The availability of such extensive and granular data gave the study the chance to train a robust and accurate predictive model.

The Zambian Super League lags behind not only in popularity, which for the EPL has resulted into research and the existence of comparative models, but also in the availability of private and public data. This judgment was made after thorough investigation that included a review of the Football Association of Zambia’s (FAZ) Facebook page (Figure 3.1) to determine the association’s official website (Figure 3.2) as two similar domains exited but only one had available information (Figure 3.3). A search for the existence of a website for the Zambian Super League was made which found the website inactive. With that, other credible sources of information on the ZSL where sought after. These included, ESPN, Sofascore and Flashscore.



Football Association of Zambia (FAZ)

995K followers • 139 following

Intro

Official Website: http://www.faz.co.zm
Official Twitter Account: @FAZFootball
https://twitter

Page · Sports

Football House, Alick Nkata Road, Lusaka, Zambia

967380231

faz@zamnet.zm

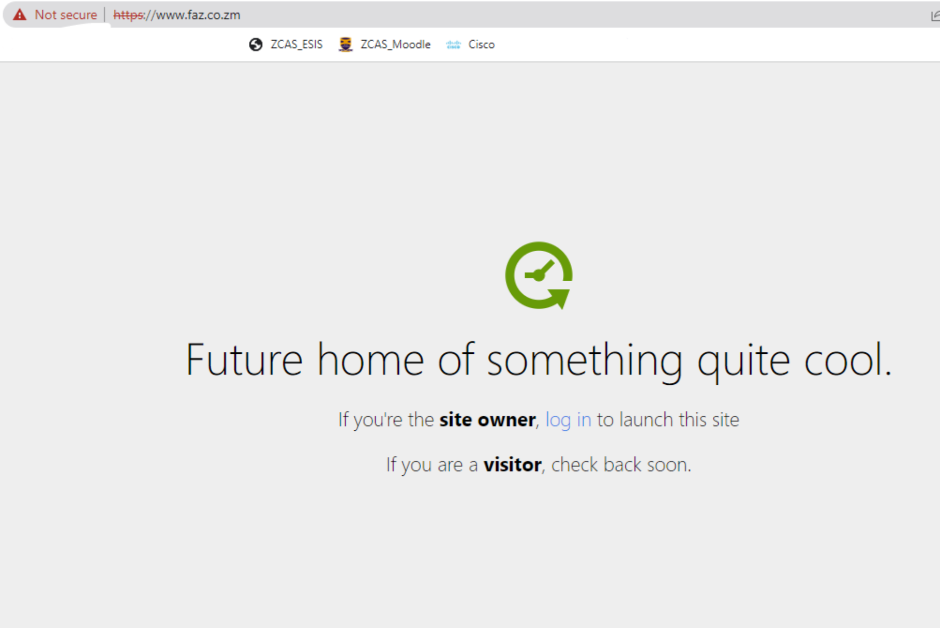
Featured

Football Association of Zambia (FAZ)
2 December 2023 at 15:15 ·
VIDEO LINK: Aim higher | FAZ president Andrew Kamanga challenges Copper Queens → https://youtu.be/qTL-...

Football Assoc
(FAZ)
28 November 202
Tickets for the Copper
Qualifier home match
for 5 December, 2023

Figure 3.1: FAZ Facebook showing Official Website

The official website for the Football Association of Zambia was determined to be www.faz.co.zm, however, the website at the time was running but had no available information as shown in Figure 3.2.



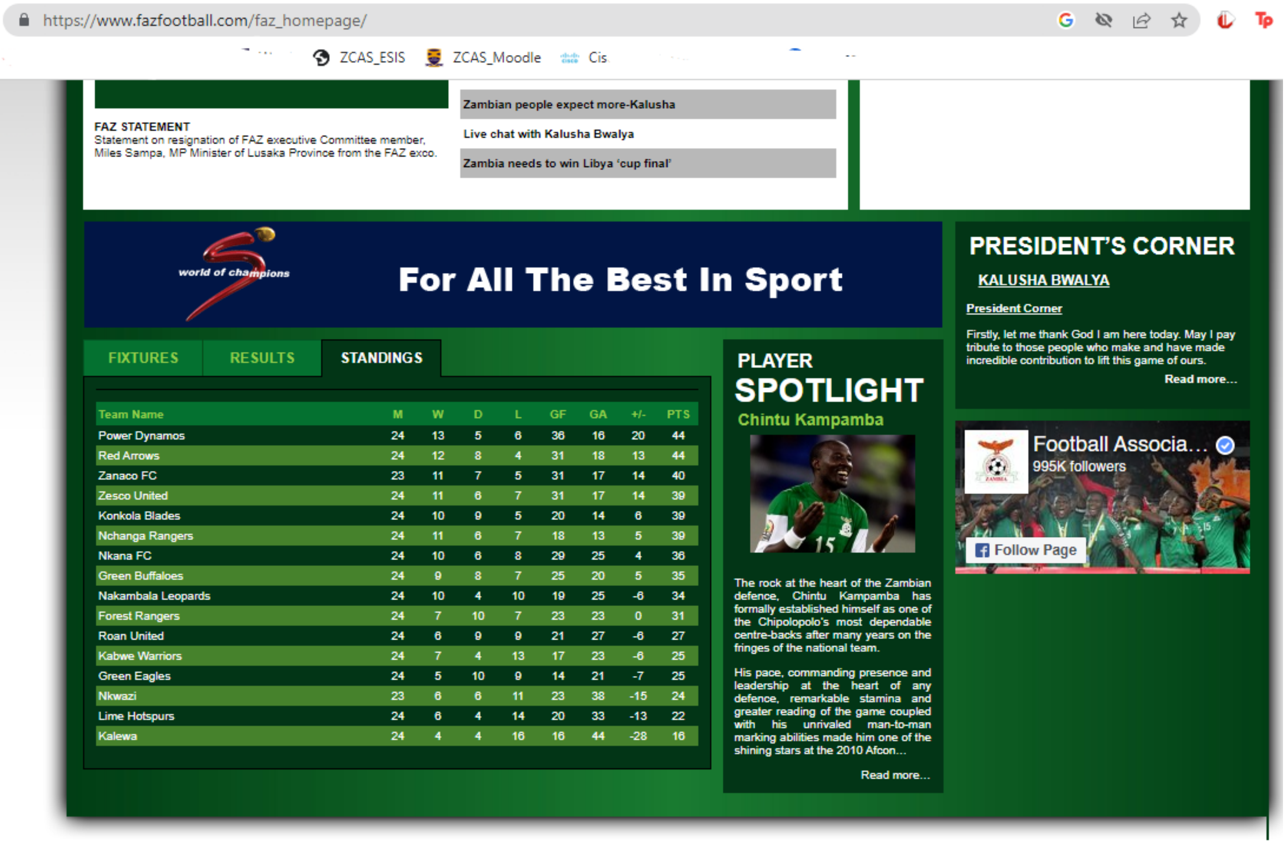
Future home of something quite cool.

If you're the **site owner**, log in to launch this site

If you are a **visitor**, check back soon.

Figure 3.2: FAZ Website Inactive

Another FAZ related domain (www.fazfootball.com) was discovered to be active (Figure 3.3). The website was also reviewed for match-related data on the Zambian Super League. However, only table listings for the previous season 2022/2023 where available but incomplete and only up to match 24. Despite having a tab for results, the previous results were not available as the tab's button was not functional.


https://www.fazfootball.com/faz_homepage/
ZCAS_ESIS
ZCAS_Moodle
Cis

FAZ STATEMENT
Statement on resignation of FAZ executive Committee member, Miles Sampa, MP Minister of Lusaka Province from the FAZ exco.

Zambian people expect more-Kalusha

Live chat with Kalusha Bwalya

Zambia needs to win Libya 'cup final'

world of champions
For All The Best In Sport
FIXTURES
RESULTS
STANDINGS

Team Name	M	W	D	L	GF	GA	+/-	PTS
Power Dynamos	24	13	5	6	36	16	20	44
Red Arrows	24	12	8	4	31	18	13	44
Zanaco FC	23	11	7	5	31	17	14	40
Zesco United	24	11	6	7	31	17	14	39
Konkola Blades	24	10	9	5	20	14	6	39
Nchanga Rangers	24	11	6	7	18	13	5	39
Nkana FC	24	10	6	8	29	25	4	36
Green Buffaloes	24	9	8	7	25	20	5	35
Nakambala Leopards	24	10	4	10	19	25	-6	34
Forest Rangers	24	7	10	7	23	23	0	31
Roan United	24	6	9	9	21	27	-6	27
Kabwe Warriors	24	7	4	13	17	23	-6	25
Green Eagles	24	5	10	9	14	21	-7	25
Nkwazi	23	6	6	11	23	38	-15	24
Lime Hotspurs	24	6	4	14	20	33	-13	22
Kalewa	24	4	4	16	16	44	-28	16

PLAYER SPOTLIGHT

Chintu Kampamba

The rock at the heart of the Zambian defence, Chintu Kampamba has formally established himself as one of the Chipolopolo's most dependable centre-backs after many years on the fringes of the national team.

His pace, commanding presence and leadership at the heart of any defence, remarkable stamina and greater reading of the game coupled with his unrivaled man-to-man marking abilities made him one of the shining stars at the 2010 Afcon...

Read more...

PRESIDENT'S CORNER

KALUSHA BWALYA

President Corner

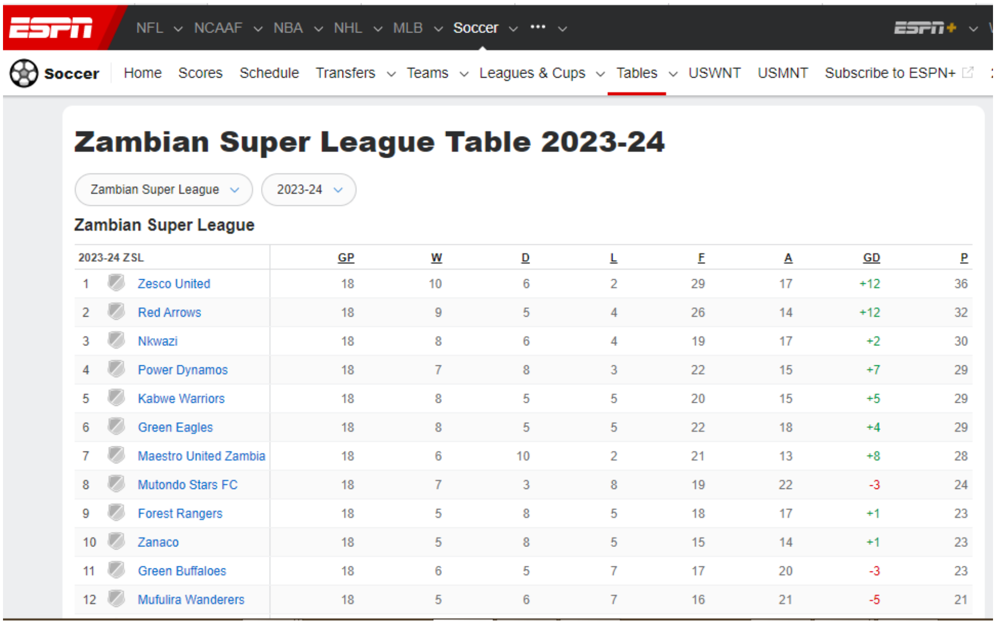
Firstly, let me thank God I am here today. May I pay tribute to those people who make and have made incredible contribution to lift this game of ours.

Read more...

Figure 3.3: FAZ Related Website with Outdated and Incomplete Information

Unlike the English Premier League (www.premierleague.com), a website for the Zambian Super League was not found despite the domain name www.zambiasuperleague.com being registered. The alternative domain "zambiasuperleague.co.zm" was also not in existence.

Other credible sources of sports information; ESPN (Figure 3.4), Sofascore and Flashscore (Figure 3.5) contained Zambian Super League information but this was limited only to table standings, fixtures and previous results. Neither of these sources had collected match-related data such as starting line-ups, possession, shots on target, number of yellow cards, referees, etc. However, Sofascore and Flashcore, in addition to full-time results and table standings also showed which players scored in a given match and the time of the goal. This information could used as part of a set of other useful features. Without a complete set using this information would be challenge as no starting line-ups are provided so the impact of a player being in a given match would be difficult to determine.



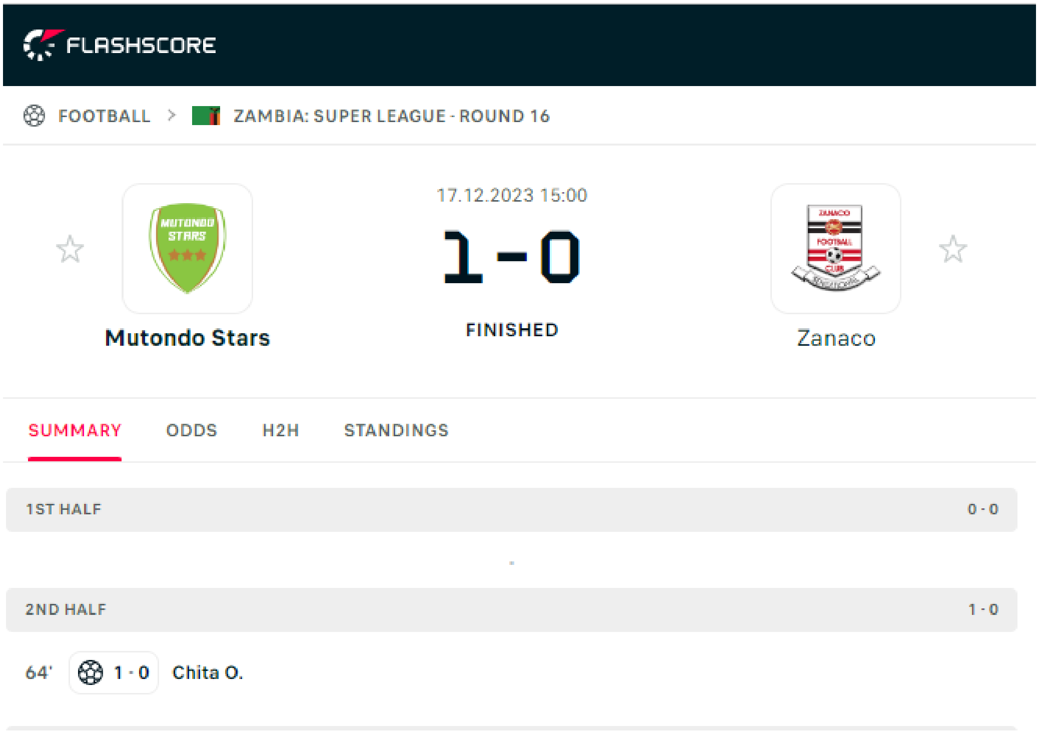
Zambian Super League Table 2023-24

Zambian Super League ⌄ 2023-24 ⌄

Zambian Super League

2023-24 ZSL		GP	W	D	L	F	A	GD	P
1	Zesco United	18	10	6	2	29	17	+12	36
2	Red Arrows	18	9	5	4	26	14	+12	32
3	Nkwazi	18	8	6	4	19	17	+2	30
4	Power Dynamos	18	7	8	3	22	15	+7	29
5	Kabwe Warriors	18	8	5	5	20	15	+5	29
6	Green Eagles	18	8	5	5	22	18	+4	29
7	Maestro United Zambia	18	6	10	2	21	13	+8	28
8	Mutondo Stars FC	18	7	3	8	19	22	-3	24
9	Forest Rangers	18	5	8	5	18	17	+1	23
10	Zanaco	18	5	8	5	15	14	+1	23
11	Green Buffaloes	18	6	5	7	17	20	-3	23
12	Mufulira Wanderers	18	5	6	7	16	21	-5	21

Figure 3.4: Zambian Super League on ESPN



FOOTBALL > ZAMBIA: SUPER LEAGUE - ROUND 16

17.12.2023 15:00

Mutondo Stars 1-0 Zanaco

FINISHED

1ST HALF 0 - 0

2ND HALF 1 - 0

64' 1 - 0 Chita O.

Figure 3.5: Zambian Super League on Flashscore

3.5.1.2 Opta's Impact on Data Quality

Opta, a leading sports data provider, has been integral in revolutionising football analytics. Their partnership with the English Premier League has resulted in a sophisticated and detailed dataset that extends beyond mere match outcomes. Opta's contribution includes real-time tracking of player movements, team dynamics, shots on target, possession percentages, and various other in-depth metrics. This depth of data is pivotal for developing good predictive models, going beyond traditional full-time results to capture the intricate dynamics of football matches. At the time of writing, no such organisation exists in Zambia, which could be one of the reasons why the ZSL does not have quality data.

3.5.1.3 Global Relevance and Generalisability

The English Premier League's global popularity ensures a diverse and extensive fanbase. Analysing a widely followed league enhances the generalisability of the study's findings. The global interest in the EPL also means that the study's outcomes could have broader implications and applications, contributing to the existing body of knowledge in football analytics. Consequently, the study’s findings could also inform on the essential features that should be collected for the Zambian Super League.

3.5.1.4 Comparative Studies and Benchmarks

The EPL serves as a benchmark for football analytics due to its competitive nature, featuring top-tier teams and world-class players. Many research studies and benchmarking initiatives in the field of football prediction models use the EPL as a reference point. Choosing the EPL allows for the study's outcomes to be contextualised within the broader landscape of football analytics, facilitating comparisons with other research efforts.

A Google Scholar search with the string “English Premier League” limited to papers between 2013 – 2023 on yielded over 16,800 results on a range of topics. A search for “Zambian Super League” yielded only four (4) results between 2013 – 2023 and none of them were on sports analytics. According to Opta rankings of the top 100 football leagues in the world, the English Premier League, German Bundesliga, Spanish La Liga, Italian Serie A and French Ligue 1 are the five (5) most popular leagues in the world (The Analyst, 2023). Table 3.2 shows a comparison in individual search results in Google Scholar’s search engine for indexed literature on these leagues and the Zambian Super League.

Table 3.2: Google Scholar Search Results for Literature on the ZSL

League Name	Google Scholar (2013 - 2023)			
	"League Name"	"Machine Learning" AND "League Name"	"Predictive" AND "League Name"	"Prediction" AND "League Name"
English Premier League	16,800	2,040	2,970	4,120
German Bundesliga	4,590	560	718	1,050
Spanish La Liga	1,680	276	295	398
Italian Serie A	3,390	372	529	766
French Ligue 1	1,010	145	177	256
Zambian Super League	4	0	0	0

*** As last searched on 31 December, 2023

Each league was run in four Boolean searches, first with its name only (e.g. "Spanish La Liga"), second with its name and the string "Machine Learning" (e.g. "Machine Learning" AND "Spanish La Liga"), third using "Predictive" and the league's name (e.g. "Predictive" AND "Spanish La Liga"), and lastly using the league's name and "Prediction" (e.g. "Prediction" AND "Spanish La Liga"). As expected, the English Premier League topped all the other leagues in each category. This made the EPL a rich source of comparative models, techniques and benchmarks that can be applied to other leagues, including the Zambian Super League. A summation of these results in a stacked chart (Figure 3.6) further amplifies the lack of literature on the Zambian Super League.

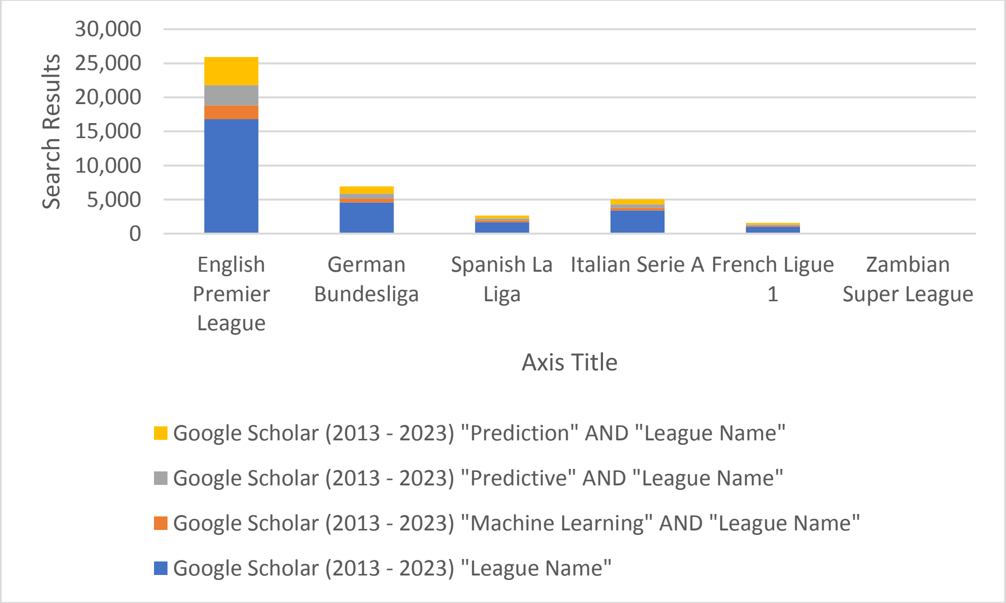


Figure 3.6: Cumulative Google Scholar Results Literature on the ZSL

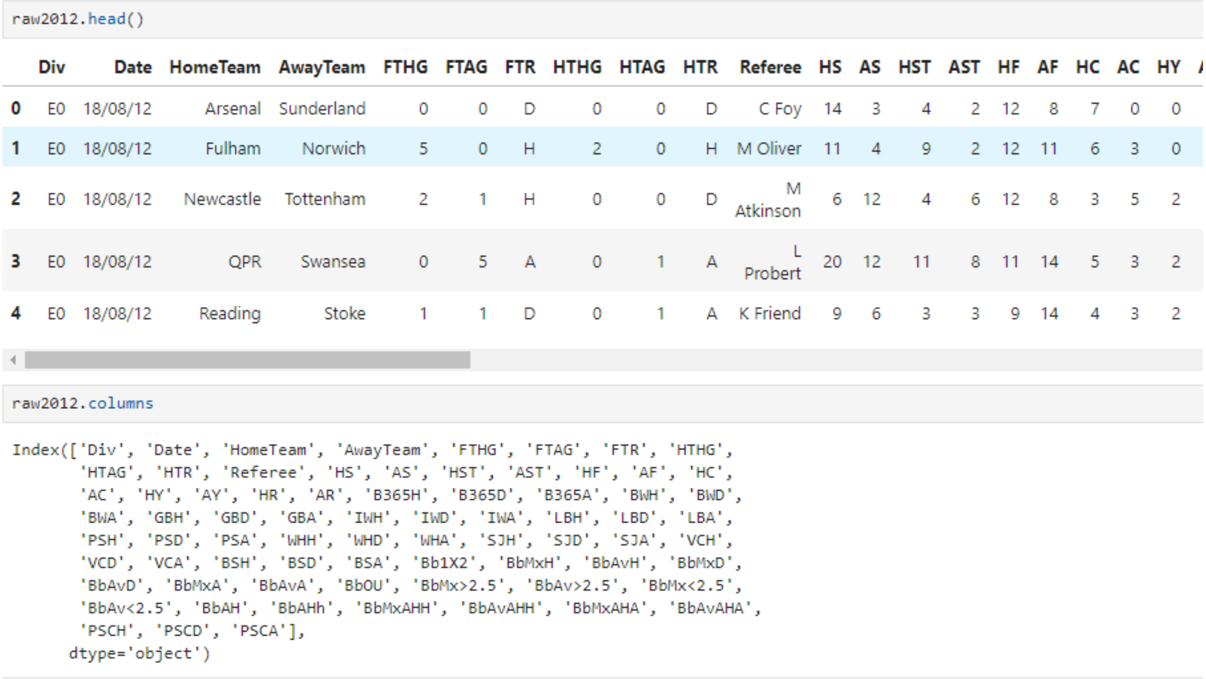
3.5.1.5 Proficiency in Predictive Modelling

The EPL has been a focal point for numerous predictive modelling studies (Figure 3.6), resulting in a wealth of literature, methodologies, and best practices. Leveraging this existing knowledge base provided a solid foundation for the study, allowing it to draw insights from prior research, implement proven techniques, and contribute to the ongoing body of knowledge in the field.

While the Zambian Super League holds its own significance, the scarcity of detailed and varied data, particularly match statistics beyond full-time results, poses a significant challenge for a predictive modelling study. Opta's extensive coverage of the English Premier League, capturing diverse aspects of the game, makes it a more conducive choice for a research endeavour aiming to develop a sophisticated and profitable predictive model. The decision aligns with the practical need for robust data sources to support the study's objectives and ensure the reliability and effectiveness of the developed model.

3.5.2 English Premier League Datasets

English Premier League data was collected for 11 complete seasons (2012/2013 to 2022/2023) and half a season (2023/2024). The data (Figure 3.7) comprised match-related facts and betting odds (Esme and Kiran, 2018) collected from www.football-data.co.uk.



```
raw2012.head()
```

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY
0	E0	18/08/12	Arsenal	Sunderland	0	0	D	0	0	D	C Foy	14	3	4	2	12	8	7	0	0
1	E0	18/08/12	Fulham	Norwich	5	0	H	2	0	H	M Oliver	11	4	9	2	12	11	6	3	0
2	E0	18/08/12	Newcastle	Tottenham	2	1	H	0	0	D	M Atkinson	6	12	4	6	12	8	3	5	2
3	E0	18/08/12	QPR	Swansea	0	5	A	0	1	A	L Probert	20	12	11	8	11	14	5	3	2
4	E0	18/08/12	Reading	Stoke	1	1	D	0	1	A	K Friend	9	6	3	3	9	14	4	3	2

```
raw2012.columns
```

```
Index(['Div', 'Date', 'HomeTeam', 'AwayTeam', 'FTHG', 'FTAG', 'FTR', 'HTHG',
       'HTAG', 'HTR', 'Referee', 'HS', 'AS', 'HST', 'AST', 'HF', 'AF', 'HC',
       'AC', 'HY', 'AY', 'HR', 'AR', 'B365H', 'B365D', 'B365A', 'BWH', 'BWD',
       'BWA', 'GBH', 'GBD', 'GBA', 'IWH', 'IWD', 'IWA', 'LBH', 'LBD', 'LBA',
       'PSH', 'PSD', 'PSA', 'WHH', 'WHD', 'WHA', 'SJH', 'SJD', 'SJA', 'VCH',
       'VCD', 'VCA', 'BSH', 'BSD', 'BSA', 'Bb1X2', 'BbMxH', 'BbAvH', 'BbMxD',
       'BbAvD', 'BbMxA', 'BbAvA', 'BbOU', 'BbMx>2.5', 'BbAv>2.5', 'BbMx<2.5',
       'BbAv<2.5', 'BbAH', 'BbAHh', 'BbMxAHH', 'BbAvAHH', 'BbMxAHA', 'BbAvAHA',
       'PSCH', 'PSCD', 'PSCA'],
      dtype='object')
```

Figure 3.7: Initial Dataset for EPL Season 2012/2012

The initial data had over 62 - 106 features which were then reduced to 15 after an evaluation of the model performance’s and feature importance. The 15 features use are shown in Table 3.3.

Table 3.3: Key Features used for Modelling

Feature	Description
HomeTeam	Home Team
AwayTeam	Away Team
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HC	Home Team Corners
AC	Away Team Corners
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards
FTR or Result	Full Time Result (H=Home Win, D=Draw, A=Away Win)

3.6 EDA and Data Comprehension

As guided by the conceptual framework in Chapter 2, domain knowledge was built through the review of literature on similar research. This was followed by data collection, and then the next stage of the framework guided the exploratory data analysis (EDA) and data comprehension. This section discusses the EDA & data comprehension process.

3.6.1 Data Collection Methods

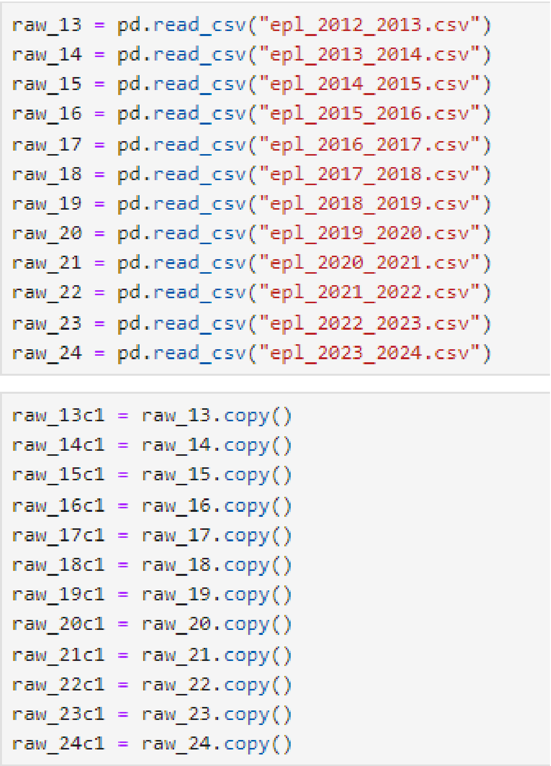
The data was collected in structured comma-separated value files (CSV) from Football-Data, a reliable source that provided well-organised data, minimising the need for extensive pre-processing and cleaning. This streamlined approach ensured the integrity of the dataset for robust analysis.

3.6.2 Overview of Data Analysis Techniques

For data analysis, Python3 was the primary programming language employed using a Jupyter Notebook. A combination of widely used libraries and packages facilitated various aspects of the analysis process. Pandas was instrumental for handling CSV files and manipulating dataframes, while Sci-kit Learn was employed for the actual modelling process. The creation of model ensembles for each season during training utilised Scipy. Matplotlib and Seaborn were employed for the visualization of tables and figures, providing a clear representation of the obtained results. The preservation of models for future reference was achieved through the use of Pickle. Lastly, NumPy played a vital role in array manipulation, contributing to the efficiency of data processing throughout the analysis. This comprehensive set of tools and techniques ensured a rigorous and systematic approach to both data collection and subsequent analysis in the development of the predictive model.

3.6.3 Exploratory Data Analysis

The CSV files were imported and read as Pandas dataframes, after which a copy of each dataframe was made to preserve the original form. The EDA also involved initial pre-processing.

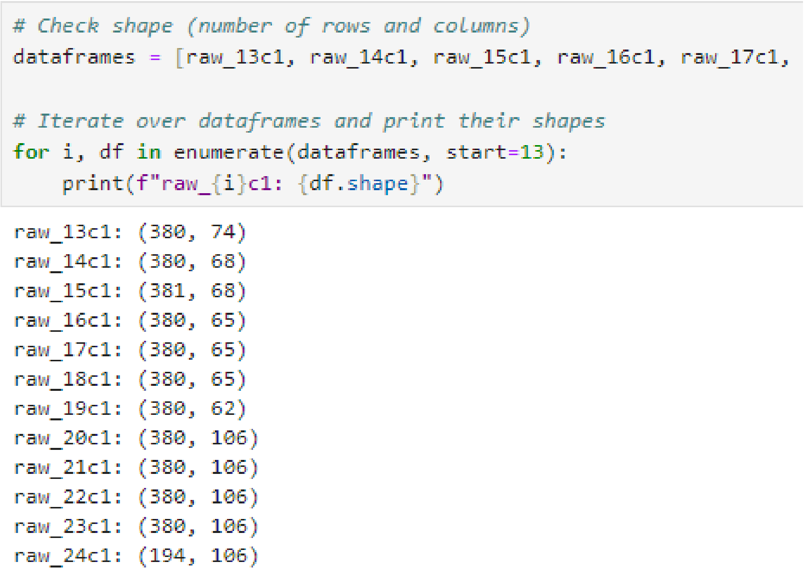


```
raw_13 = pd.read_csv("epl_2012_2013.csv")
raw_14 = pd.read_csv("epl_2013_2014.csv")
raw_15 = pd.read_csv("epl_2014_2015.csv")
raw_16 = pd.read_csv("epl_2015_2016.csv")
raw_17 = pd.read_csv("epl_2016_2017.csv")
raw_18 = pd.read_csv("epl_2017_2018.csv")
raw_19 = pd.read_csv("epl_2018_2019.csv")
raw_20 = pd.read_csv("epl_2019_2020.csv")
raw_21 = pd.read_csv("epl_2020_2021.csv")
raw_22 = pd.read_csv("epl_2021_2022.csv")
raw_23 = pd.read_csv("epl_2022_2023.csv")
raw_24 = pd.read_csv("epl_2023_2024.csv")
```

```
raw_13c1 = raw_13.copy()
raw_14c1 = raw_14.copy()
raw_15c1 = raw_15.copy()
raw_16c1 = raw_16.copy()
raw_17c1 = raw_17.copy()
raw_18c1 = raw_18.copy()
raw_19c1 = raw_19.copy()
raw_20c1 = raw_20.copy()
raw_21c1 = raw_21.copy()
raw_22c1 = raw_22.copy()
raw_23c1 = raw_23.copy()
raw_24c1 = raw_24.copy()
```

Figure 3.8: Importing CSV Files

The dataframes grouped into a list and were checked for their shape (rows, columns) to ensure consistency in their features. This exploration revealed that some datasets had different numbers of features ranging from 62 – 106 and the dataframe for the 2014/2015 season had one additional row (Figure 3.9).



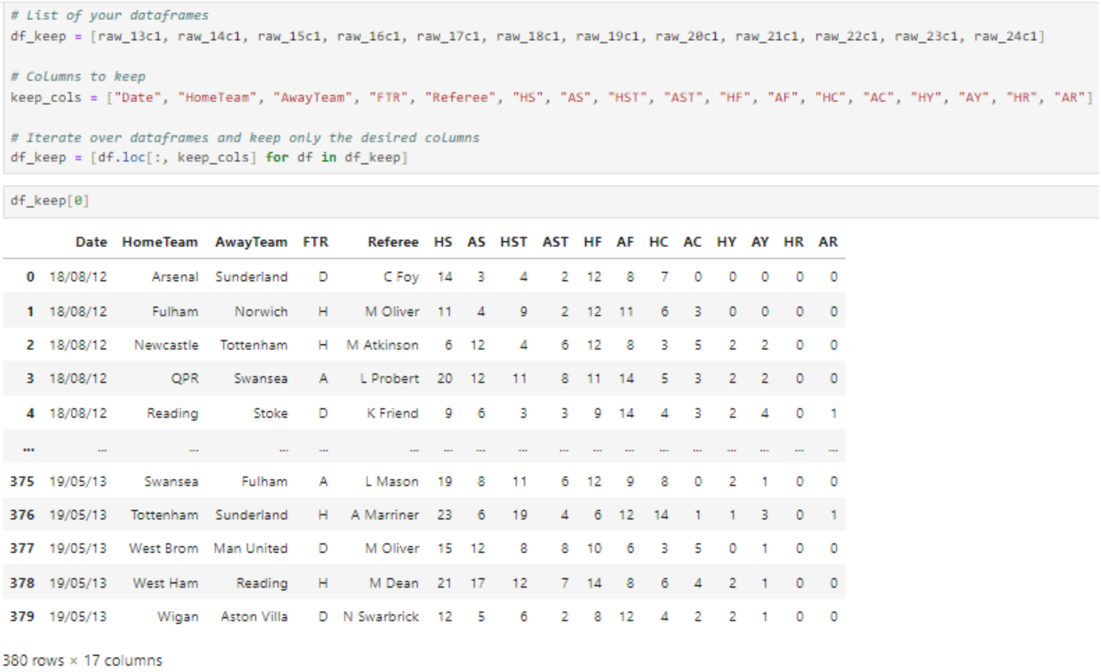
```
# Check shape (number of rows and columns)
dataframes = [raw_13c1, raw_14c1, raw_15c1, raw_16c1, raw_17c1,

# Iterate over dataframes and print their shapes
for i, df in enumerate(dataframes, start=13):
    print(f"raw_{i}c1: {df.shape}")
```

```
raw_13c1: (380, 74)
raw_14c1: (380, 68)
raw_15c1: (381, 68)
raw_16c1: (380, 65)
raw_17c1: (380, 65)
raw_18c1: (380, 65)
raw_19c1: (380, 62)
raw_20c1: (380, 106)
raw_21c1: (380, 106)
raw_22c1: (380, 106)
raw_23c1: (380, 106)
raw_24c1: (194, 106)
```

Figure 3.9: Initial Rows and Columns in Dataset

To understand the data, each dataframe was previewed to compare the features, identify the different ones and take note of the indices of the features the study wanted to keep. At inspection, the Div column and columns with scores were dropped. Keeping the scores would have "told" the model what the result was. The study needed the model to predict the results using match-facts expect from the scores. Rows with at least one missing value were also dropped. Dataframes were held in lists for easier manipulation.



```
# List of your dataframes
df_keep = [raw_13c1, raw_14c1, raw_15c1, raw_16c1, raw_17c1, raw_18c1, raw_19c1, raw_20c1, raw_21c1, raw_22c1, raw_23c1, raw_24c1]

# Columns to keep
keep_cols = ["Date", "HomeTeam", "AwayTeam", "FTR", "Referee", "HS", "AS", "HST", "AST", "HF", "AF", "HC", "AC", "HY", "AY", "HR", "AR"]

# Iterate over dataframes and keep only the desired columns
df_keep = [df.loc[:, keep_cols] for df in df_keep]
```

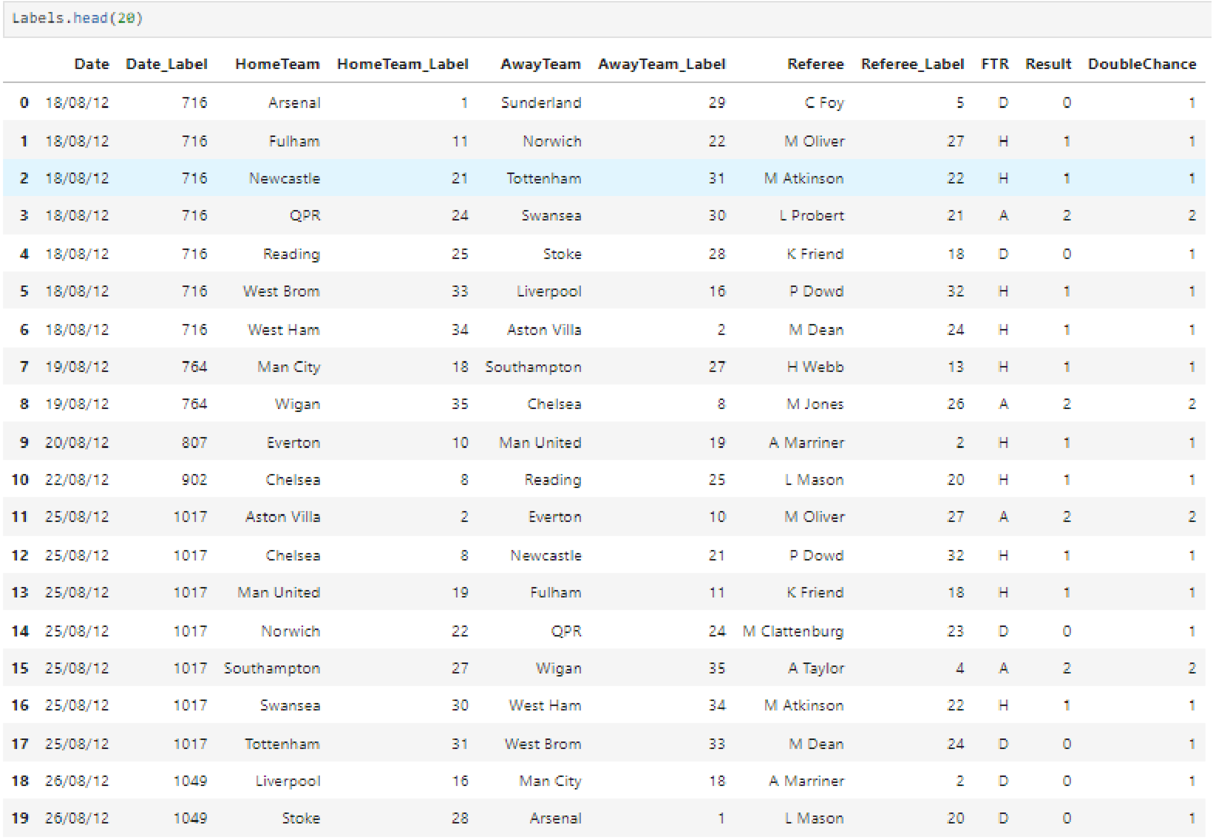
```
df_keep[0]
```

	Date	HomeTeam	AwayTeam	FTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
0	18/08/12	Arsenal	Sunderland	D	C Foy	14	3	4	2	12	8	7	0	0	0	0	0
1	18/08/12	Fulham	Norwich	H	M Oliver	11	4	9	2	12	11	6	3	0	0	0	0
2	18/08/12	Newcastle	Tottenham	H	M Atkinson	6	12	4	6	12	8	3	5	2	2	0	0
3	18/08/12	QPR	Swansea	A	L Probert	20	12	11	8	11	14	5	3	2	2	0	0
4	18/08/12	Reading	Stoke	D	K Friend	9	6	3	3	9	14	4	3	2	4	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
375	19/05/13	Swansea	Fulham	A	L Mason	19	8	11	6	12	9	8	0	2	1	0	0
376	19/05/13	Tottenham	Sunderland	H	A Marriner	23	6	19	4	6	12	14	1	1	3	0	1
377	19/05/13	West Brom	Man United	D	M Oliver	15	12	8	8	10	6	3	5	0	1	0	0
378	19/05/13	West Ham	Reading	H	M Dean	21	17	12	7	14	8	6	4	2	1	0	0
379	19/05/13	Wigan	Aston Villa	D	N Swarbrick	12	5	6	2	8	12	4	2	2	1	0	0

380 rows × 17 columns

Figure 3.10: Initial Key Features

After the shape and features of each dataframe was consistent throughout the dataset, the 12 dataframes were concatenated to a new dataframe to preserve the cleaned dataframes. The concatenated dataframe was used to develop numerical labels for categorical features; Date, HomeTeam, AwayTeam, Referee and Full Time Result (FTR). The FTR was used to create two additional columns "Result" and "DoubleChance". In the Result column, Values of 1 were assigned to instances denoting HomeWin (H), 0 to those indicating a Draw (D), and 2 to occurrences signifying an AwayWin (A). For the DoubleChance column both HomeWin (H) and Draw (D) were encoded as 1, capturing the combined probability of either a HomeWin or Draw. In parallel, the AwayWin (A) was represented by the numerical value 2.



```
Labels.head(20)
```

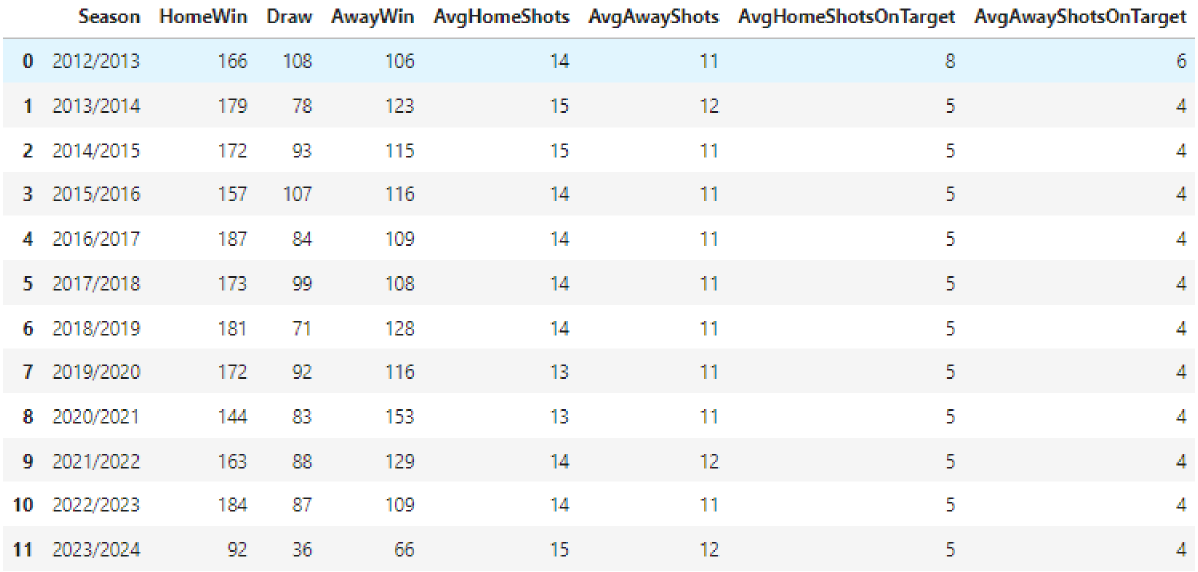
	Date	Date_Label	HomeTeam	HomeTeam_Label	AwayTeam	AwayTeam_Label	Referee	Referee_Label	FTR	Result	DoubleChance
0	18/08/12	716	Arsenal	1	Sunderland	29	C Foy	5	D	0	1
1	18/08/12	716	Fulham	11	Norwich	22	M Oliver	27	H	1	1
2	18/08/12	716	Newcastle	21	Tottenham	31	M Atkinson	22	H	1	1
3	18/08/12	716	QPR	24	Swansea	30	L Probert	21	A	2	2
4	18/08/12	716	Reading	25	Stoke	28	K Friend	18	D	0	1
5	18/08/12	716	West Brom	33	Liverpool	16	P Dowd	32	H	1	1
6	18/08/12	716	West Ham	34	Aston Villa	2	M Dean	24	H	1	1
7	19/08/12	764	Man City	18	Southampton	27	H Webb	13	H	1	1
8	19/08/12	764	Wigan	35	Chelsea	8	M Jones	26	A	2	2
9	20/08/12	807	Everton	10	Man United	19	A Marriner	2	H	1	1
10	22/08/12	902	Chelsea	8	Reading	25	L Mason	20	H	1	1
11	25/08/12	1017	Aston Villa	2	Everton	10	M Oliver	27	A	2	2
12	25/08/12	1017	Chelsea	8	Newcastle	21	P Dowd	32	H	1	1
13	25/08/12	1017	Man United	19	Fulham	11	K Friend	18	H	1	1
14	25/08/12	1017	Norwich	22	QPR	24	M Clattenburg	23	D	0	1
15	25/08/12	1017	Southampton	27	Wigan	35	A Taylor	4	A	2	2
16	25/08/12	1017	Swansea	30	West Ham	34	M Atkinson	22	H	1	1
17	25/08/12	1017	Tottenham	31	West Brom	33	M Dean	24	D	0	1
18	26/08/12	1049	Liverpool	16	Man City	18	A Marriner	2	D	0	1
19	26/08/12	1049	Stoke	28	Arsenal	1	L Mason	20	D	0	1

Figure 3.11: Categorical Value Labels

3.6.3.1 Descriptive Statistics

A major component of EDA is descriptive statistics as it helps understand the data. These insights provide a more detailed understanding of the performance trends across different seasons of the EPL, emphasizing the variability and competitiveness of the League.

The descriptive statistics on the study’s dataset revealed that in each season the number of home wins was more than the number of draws or away wins (Figure 3.12). This showed the significance of a team playing at home as it creates a home advantage. The lowest number of home wins (144) in a complete season was in the 2020/2021 season that had EPL matches without fans in attendance due to COVID-19 restrictions. The absence of supporters in the stadium could have impacted the feeling of being at home. This was also noted from the increase in away wins (153), the most across the 11 complete seasons.



	Season	HomeWin	Draw	AwayWin	AvgHomeShots	AvgAwayShots	AvgHomeShotsOnTarget	AvgAwayShotsOnTarget
0	2012/2013	166	108	106	14	11	8	6
1	2013/2014	179	78	123	15	12	5	4
2	2014/2015	172	93	115	15	11	5	4
3	2015/2016	157	107	116	14	11	5	4
4	2016/2017	187	84	109	14	11	5	4
5	2017/2018	173	99	108	14	11	5	4
6	2018/2019	181	71	128	14	11	5	4
7	2019/2020	172	92	116	13	11	5	4
8	2020/2021	144	83	153	13	11	5	4
9	2021/2022	163	88	129	14	12	5	4
10	2022/2023	184	87	109	14	11	5	4
11	2023/2024	92	36	66	15	12	5	4

Figure 3.12: Match Results and Average Shots

The 2016/2017 season stands out as one of the most competitive seasons with a high count of home wins (187), suggesting strong home performance, but also a substantial count of draws (84) and a reasonable count of away wins (109), indicating a balanced and competitive season. The average number of shots and shots on target for home and away teams showed a clear correlation with the match outcomes. Throughout all the seasons, the home teams registered more shots and shots on target than the away teams. Evidently pointing to a higher probability of scoring if more shots are taken.

3.6.3.2 Correlation and Multicollinearity

In assessing the predictive model's strength, a comprehensive investigation of correlation and multicollinearity among features was completed to ensure the model's reliability in capturing interdependencies. This EDA step aimed to understand the strength and nature of associations between features. It should be noted that Naive Bayes models are generally considered less sensitive to multicollinearity compared to some other machine learning algorithms (Haruna et al., 2021; Ćwiklinski, Giełczyk, and Choraś, 2021). This is because Naive Bayes models operate under the assumption of independence between features, given the class label. Therefore, even if there is some degree of multicollinearity among predictors, Naive Bayes tends to handle it reasonably well (Beal et al., 2021; Muszaidi et al., 2022).

The correlation matrix (Figure 3.13) summarises the observation from Figure 3.12. The number of shots by the home teams (HS) positively correlated (0.91) with the outcome of the home team winning (HomeWin). The away teams' outcome of winning (AwayWin) also correlated positively (0.71) with the number of shots they took (AS).

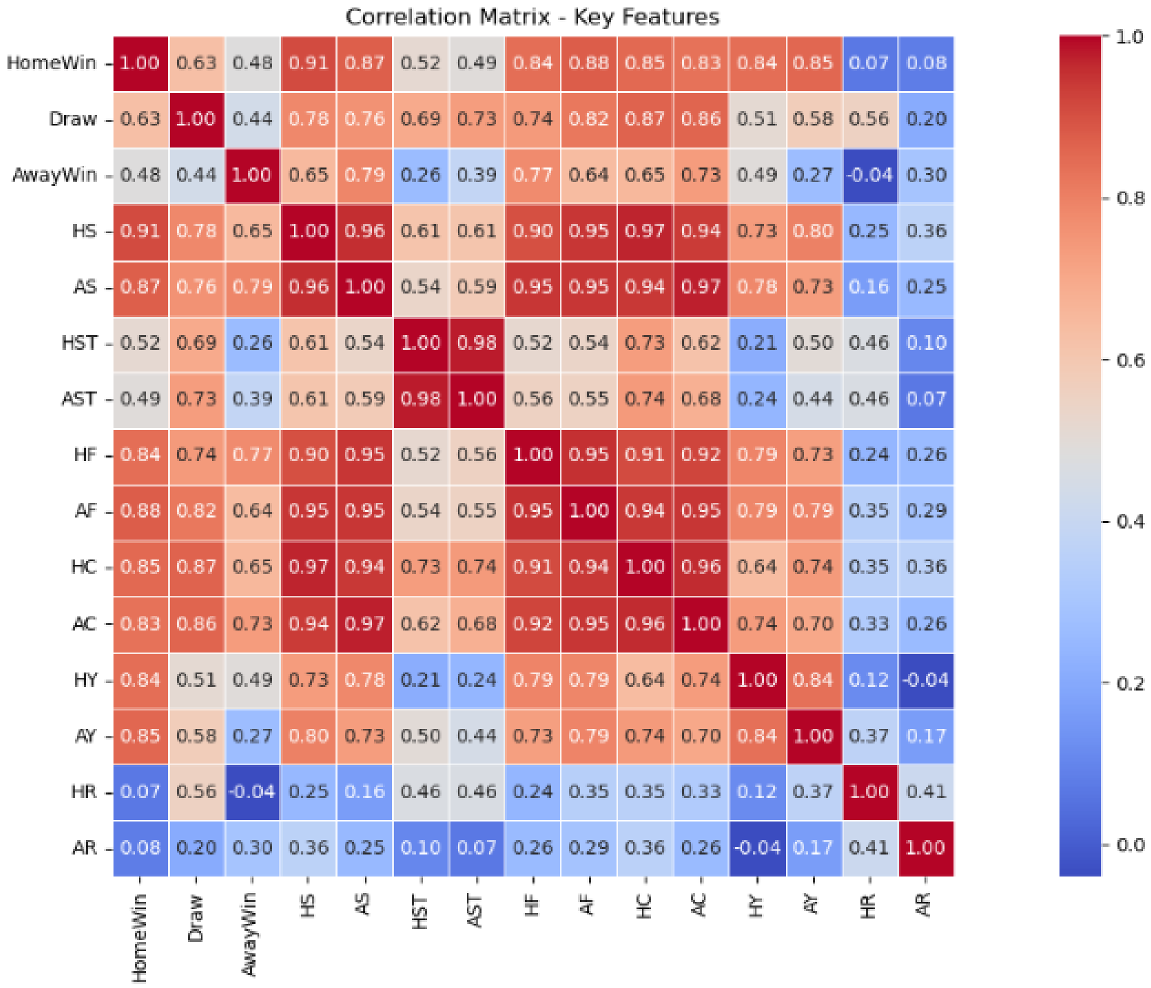


Figure 3.13: Correlation between Results and Key Features

3.7 Ethical Concerns

The data collection process involved using publicly available information from reputable sources. The study prioritised data privacy and compliance with relevant regulations throughout the research process. No private information was collected about the players or any other individuals involved in the football matches.

3.8 Chapter Summary

This chapter discussed the research design, methodology, and data-related aspects. The study adopted a quantitative research approach, emphasizing predictive modelling techniques and incorporating Bayes' conditional probability theory for a comprehensive analysis. The dataset, covering 11 complete seasons and half of the 2023/2024 season, was introduced. Data collection involved a combination of CSV downloads and Python3, along with various libraries, was used for data analysis. The chapter also highlighted the ethical considerations taken into account during the research process.

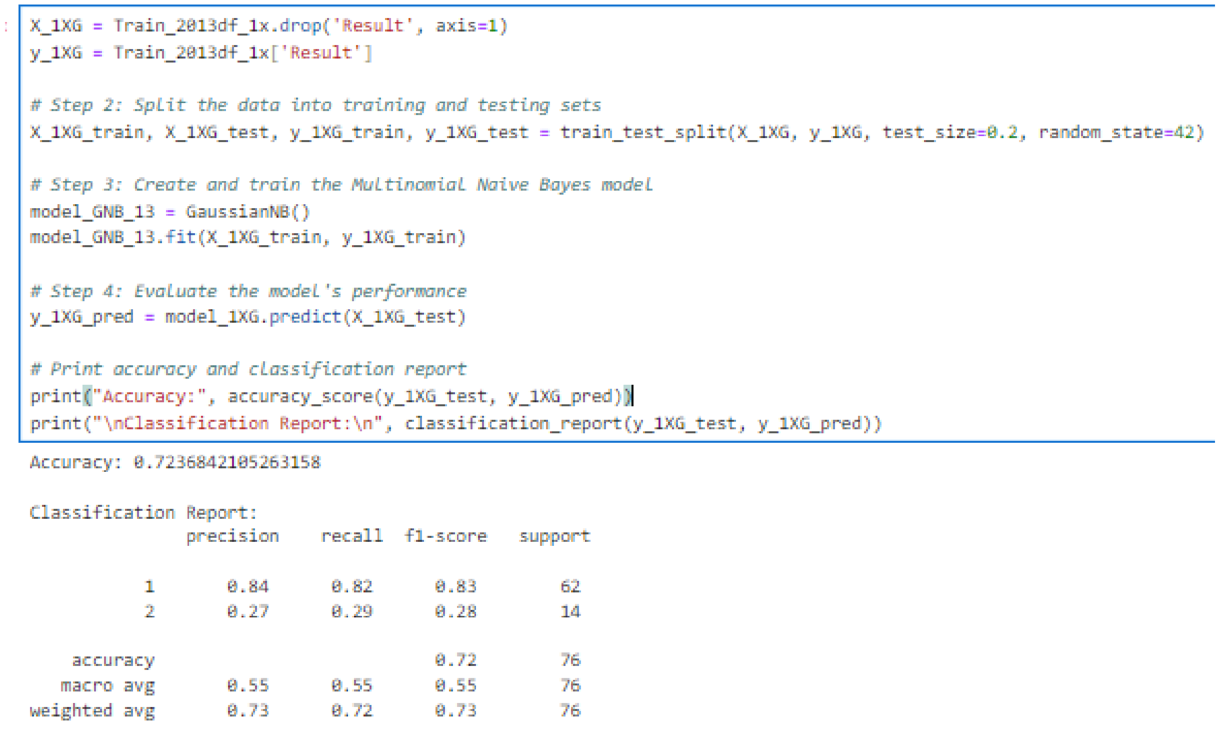
CHAPTER 4

DATA, EXPERIMENTS, AND IMPLEMENTATION

4.1 Appropriate Modelling

The modelling approach adopted in this project drew inspiration from other researchers (Razali et al., 2018; Rahman et al., 2018) who highlight the significance of Bayesian methods in predictive modelling. It also took in consideration the importance of incorporating betting odds into the modelling process (Leushuis, 2018), providing valuable insights into the dynamic nature of odds in football prediction. The integration of historical data and bookmakers' odds, as indicated by (Egidi, Pauli, and Torelli, 2018), is considered a crucial component for accurate football match prediction.

Gaussian Naïve Bayes (Figure 4.1) was used and compared with Multinomial Naïve Bayes (Figure 4.6). The Gaussian NB performed better than the Multinomial NB in the initial stages.



```
X_1XG = Train_2013df_1x.drop('Result', axis=1)
y_1XG = Train_2013df_1x['Result']

# Step 2: Split the data into training and testing sets
X_1XG_train, X_1XG_test, y_1XG_train, y_1XG_test = train_test_split(X_1XG, y_1XG, test_size=0.2, random_state=42)

# Step 3: Create and train the Multinomial Naive Bayes model
model_GNB_13 = GaussianNB()
model_GNB_13.fit(X_1XG_train, y_1XG_train)

# Step 4: Evaluate the model's performance
y_1XG_pred = model_1XG.predict(X_1XG_test)

# Print accuracy and classification report
print("Accuracy:", accuracy_score(y_1XG_test, y_1XG_pred))
print("\nClassification Report:\n", classification_report(y_1XG_test, y_1XG_pred))
```

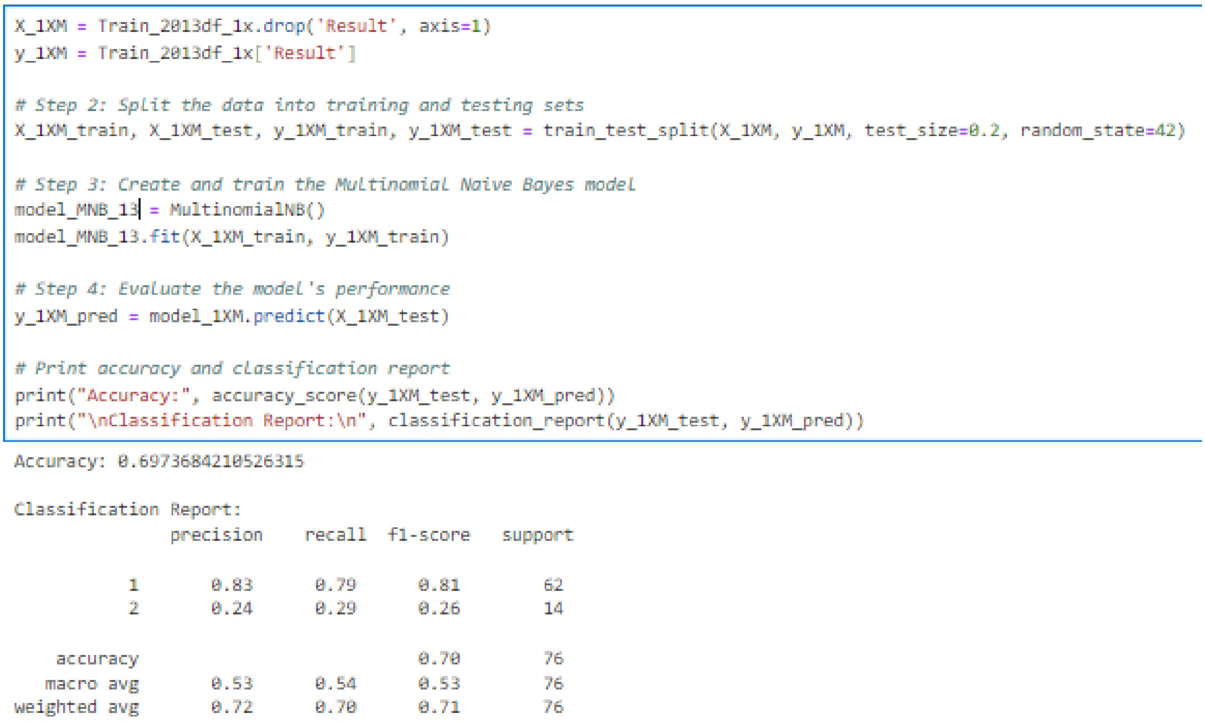
```
Accuracy: 0.7236842105263158

Classification Report:
               precision    recall  f1-score   support

           1       0.84      0.82      0.83        62
           2       0.27      0.29      0.28        14

    accuracy                           0.72        76
   macro avg       0.55      0.55      0.55        76
weighted avg       0.73      0.72      0.73        76
```

Figure 4.1: GaussianNB_12 Model for 2012/2013



```
X_1XM = Train_2013df_1x.drop('Result', axis=1)
y_1XM = Train_2013df_1x['Result']

# Step 2: Split the data into training and testing sets
X_1XM_train, X_1XM_test, y_1XM_train, y_1XM_test = train_test_split(X_1XM, y_1XM, test_size=0.2, random_state=42)

# Step 3: Create and train the Multinomial Naive Bayes model
model_MNB_13 = MultinomialNB()
model_MNB_13.fit(X_1XM_train, y_1XM_train)

# Step 4: Evaluate the model's performance
y_1XM_pred = model_1XM.predict(X_1XM_test)

# Print accuracy and classification report
print("Accuracy:", accuracy_score(y_1XM_test, y_1XM_pred))
print("\nClassification Report:\n", classification_report(y_1XM_test, y_1XM_pred))
```

```
Accuracy: 0.6973684210526315

Classification Report:
               precision    recall  f1-score   support

           1       0.83      0.79      0.81        62
           2       0.24      0.29      0.26        14

    accuracy                           0.70        76
   macro avg       0.53      0.54      0.53        76
weighted avg       0.72      0.70      0.71        76
```

Figure 4.2: MultinomialNB_18 Model for 2018/2019

4.2 Techniques and Algorithms

In the project's implementation, the code involved the application of Gaussian Naive Bayes and Multinomial Naive Bayes models for training and evaluation. The utilisation of label encoding for categorical variables and the introduction of a double chance strategy were pivotal steps in addressing the challenges of football match prediction.

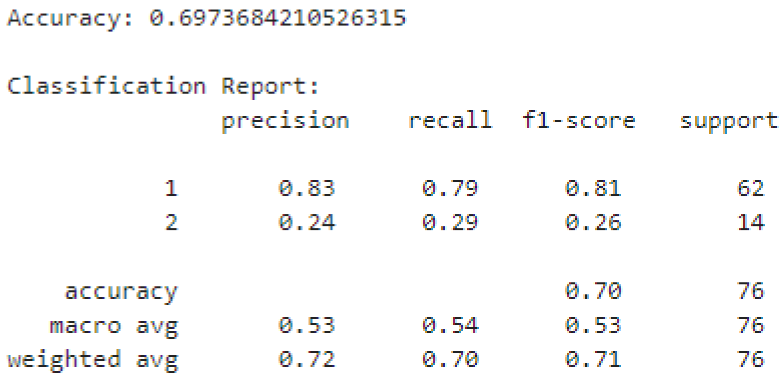
A single model was developed for each season trained and an ensemble was developed from 9 seasons (9 models). 2 seasons were left for testing without exposure to the model and the current season 2023/2024 was used as validation. The ensemble models were developed using Scipy's mode method and applied using majority vote.

The study went through the follow techniques:

Data Collection and Preparation: Historical data spanning multiple seasons (2012/2013 to 2023/2024) was collected and consolidated into a comprehensive training set. Features such as bookmakers' odds were considered and manipulated for relevant analysis.

Feature Selection and Engineering: Irrelevant features, including certain odds and match-related details, were excluded. Label encoding was applied to categorical variables, and the dataset was structured for modelling.

Model Training and Evaluation: Multinomial Naive Bayes and Gaussian Naive Bayes models were implemented and evaluated for predicting match outcomes. The performance of these models was assessed using accuracy and classification reports (Figure 4.7).



```
Accuracy: 0.6973684210526315

Classification Report:
              precision    recall  f1-score   support

           1       0.83      0.79      0.81        62
           2       0.24      0.29      0.26        14

    accuracy                           0.70        76
   macro avg       0.53      0.54      0.53        76
weighted avg       0.72      0.70      0.71        76
```

Figure 4.3: Initial Performance of Ensemble Model before Optimisation

Double Chance Strategy: To address the challenge of predicting football match outcomes, a double chance strategy was introduced, combining the classes of home win and draw. The implementation of the double chance strategy is increased the accuracy to 73% from 55%.

4.3 Double Chance Model

The primary objective of the study was to develop a profitable predictive model for English Premier League matches using machine learning and Bayesian conditional probability. The objectives of the study were:

1.***To determine the predictive models and algorithms used in predicting football match outcomes.***

The study uncovered a number machine learning models and algorithms used in predicting football outcomes and reviewed the accuracy of these models in Chapter 2.

2.**To develop a machine learning model using conditional probability in order to predict outcomes of English Premier League matches.**

The study successfully developed a predictive model using Bayes theorem based on conditional theorem. It forecast the outcomes of matches using match-related events as independent variables.

3. **To evaluate and validate the accuracy of the predictive model.**

The study uncovered multiple ways of evaluating the accuracy of the model, the most common being accuracy, as the number of correct predictions, followed by return on investment or profitability. The profitable range was found to be 55-60% accuracy and the accuracy of similar studies as shown in Chapter 2 ranges from 70.3% to 92%.

4.4 Chapter Summary

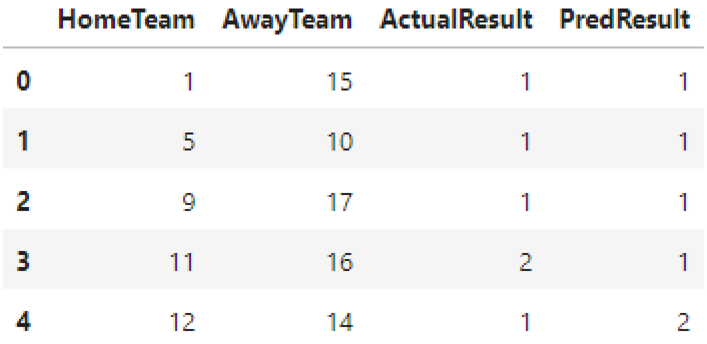
Chapter 4 presented the modelling techniques, algorithms, and experiments conducted to develop a robust predictive model for English Premier League matches. Leveraging insights from existing research and incorporating innovative approaches, the study not only built and evaluated predictive models but also introduced a double chance strategy to enhance accuracy. The chapter highlighted the careful process of data collection, preparation, and the significance of feature engineering in the context of football match prediction. The ensemble model's initial performance was discussed, laying the foundation for further optimization and validation. The chapter encapsulated the hands-on journey of implementing, refining, and validating the developed predictive model.

CHAPTER 5

RESULTS AND DISCUSSIONS

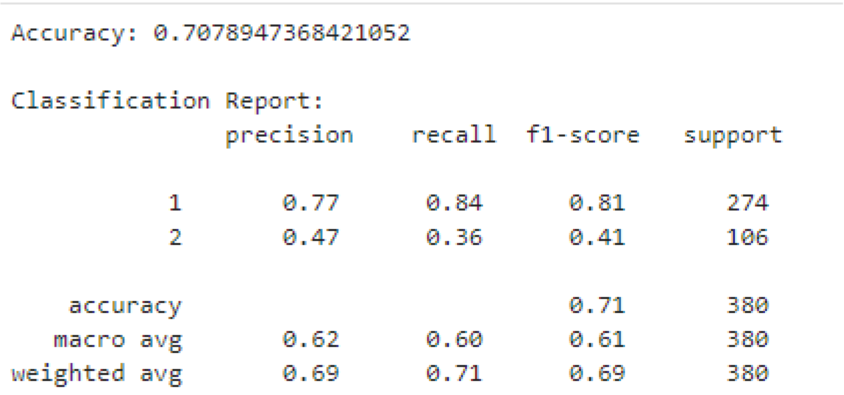
5.1 Results Presentation

The best performing model achieved an accuracy of 87% during training and an average of 84% on unseen data during testing. The least performance was 71% (Figure 5.2)



	HomeTeam	AwayTeam	ActualResult	PredResult
0	1	15	1	1
1	5	10	1	1
2	9	17	1	1
3	11	16	2	1
4	12	14	1	2

Figure 5.1: Comparing Predictions and Actual Result



```
Accuracy: 0.7078947368421052

Classification Report:
              precision    recall  f1-score   support

           1       0.77      0.84      0.81       274
           2       0.47      0.36      0.41       106

    accuracy                           0.71       380
   macro avg       0.62      0.60      0.61       380
weighted avg       0.69      0.71      0.69       380
```

Figure 5.2: Average accuracy tested on 2.5 seasons

5.2 Analysis of Results

The model performed well especially given that other models that used Naïve Bayes did not do so well or recorded similar results.

5.3 Comparison to Related Work

In comparison to studies using other techniques such as ANN, Tree Augmented Naive Bayes (TAN) and GBN-HC, the model lags behind. However, it performed better than studies that used the same model.

Table 5.1: Results Comparison

Study	Best Model/Algorithm	Performance
(Razali et al., 2021)	General Bayesian Networks + Hill Climbing algorithm (GBN-HC)	92.01%
(Owramipur, Eskandarian and Mozneb, 2013)	Bayesian network	92%
(Rahman et al., 2018)	Tree Augmented Naive Bayes (TAN)	90.00%
(Igiri and Nwachukwu, 2014)	Artificial Neural Networks	85%
This Study	Gaussian Naïve Bayes (Ensemble)	85%
(Haruna et al., 2021)	K-NN classifier	83.95%
(Stübinger, Mangold and Knoll, 2019)	Random Forest	81.26%
(Muszaidi et al., 2022)	Multilayer Perceptron (MLP)	78.42%
(Razali et al., 2017)	Naive Bayes	75.09%
(Zaveri et al., 2018)	Logistic Regression	71.63%
(Moura, Martins and Cunha, 2014)	Principal Component and Cluster Analyses	70.30%

5.4 Implications of Results

More work needs to be done to improve the model. Possibly by including more features. However, the model provides a profitable way to win on the betting market.

5.5 Chapter Summary

This chapter presented the results of the study and the implications of the results as compared to other studies.

CHAPTER 6
SUMMARY AND CONCLUSION

6.1 Summary of Main Findings

In this study, we employed a Naive Bayes approach to predict the outcomes of Premier League matches, achieving a noteworthy accuracy of 85%. The main findings of our research highlight the effectiveness of the chosen predictive modelling technique in the context of football match outcomes. Through careful analysis of historical data and relevant features, our model demonstrated a commendable ability to make accurate predictions.

6.2 Discussion and Implications in Relation to Objectives

The primary objective of this study was to develop a predictive model for Premier League match outcomes using Naive Bayes. The achieved accuracy of 85% indicates a promising capability to forecast the results of football matches. This has significant implications for various stakeholders, including sports analysts, betting enthusiasts, and football clubs seeking insights into match dynamics. The successful application of Naive Bayes in this domain underscores its potential for predicting complex and dynamic events.

In relation to the objectives, the study delves into the importance of feature selection, data pre-processing, and model training parameters. Additionally, it discusses the interpretability of Naive Bayes results and their practical applications in the realm of sports analytics.

6.3 Contribution to the Body of Knowledge

This research contributes to the existing body of knowledge in sports analytics by demonstrating the efficacy of Naive Bayes in predicting Premier League match outcomes. The study provides valuable insights into the factors influencing football match results and showcases a practical application of machine learning techniques in the sports domain. The success of the predictive model serves as a reference point for future studies exploring the application of different algorithms and methodologies in sports prediction.

Furthermore, the study contributes to the understanding of the limitations and challenges associated with sports prediction models, paving the way for further advancements in this field.

6.4 Limitations of the Model

Despite the promising results, it is crucial to acknowledge the limitations of the predictive model. The accuracy of 85% signifies a robust performance, but it leaves room for improvement. The model's performance may be affected by factors such as the dynamic nature of football, unforeseen events, player injuries, and team dynamics, which are challenging to capture comprehensively in a predictive model.

Additionally, the study relied on historical data, and the predictive power of the model may diminish over time due to changes in team strategies, player transfers, and other dynamic factors inherent in the sports environment.

6.5 Future Works

To enhance the predictive accuracy and robustness of the model, future research could explore the incorporation of more advanced machine learning algorithms, ensemble methods, and deep learning techniques. Additionally, including real-time data and leveraging more granular features could further improve the model's ability to adapt to changing conditions in the football landscape.

Furthermore, extending the scope to consider other football leagues, tournaments, or even different sports would contribute to a more comprehensive understanding of sports prediction models. Collaborations with domain experts, such as football analysts and coaches, could provide valuable insights and refine the model for practical applications.

6.6 Chapter Summary

In summary, this chapter encapsulates the key findings, implications, and contributions of the study. The Naive Bayes model exhibited a commendable accuracy of 85%, providing valuable insights into the predictive modelling of Premier League match outcomes. The limitations of the system were acknowledged, and recommendations for future research were outlined, aiming to further advance the field of sports analytics. Overall, this study lays the foundation for future endeavours in refining and expanding predictive models for sports events.

REFERENCES

Alten-Ronæss, J. (2021). Predictions of football results in the Premier League and its use on the betting market (Master's thesis, NTNU).

Angelini, G., & De Angelis, L. (2017). PARX model for football match predictions. Journal of Forecasting, 36(7), 795-807.

Arabzad, M., Tayebi, E., Araghi, S., Sadi-Nezhad, N. G. (2014). Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. Journal of Applied Research on Industrial Engineering, 1(3), 159-179.

Baboota, R., & 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.

Bologna, C., De Rosa, A. C., De Vivo, A., Gaeta, M., Sansonetti, G., & Viserta, V. (2013). Personality-based recommendation in e-commerce. CEUR Workshop Proceedings, 997.

Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(5), 338–349.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27–33.

Carloni, L., De Angelis, A., Sansonetti, G., & Micarelli, A. (2021). A Machine Learning Approach to Football Match Result Prediction. Department of Engineering, Roma Tre University.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry. (2016). A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes. Journal of the American Statistical Association, 111, 585–599.

Chance, D. (2020). Conditional probability and the length of a championship series in baseball, basketball, and hockey. Journal of Sports Analytics, 6(2), 111-127.

Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. Machine learning, 108(1), 49-75.

Constantinou, A. C., Fenton, N. E., and Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems, 36, 322-339.

Cortez, A., Trigo, A., & Loureiro, N. (2022). Football match line-up prediction based on physiological variables: a machine learning approach. Computers, 11(3), 40.

Deloitte (2023). Premier League clubs' revenue. Source: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance-premier-league-clubs.html

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA Player's Impact on his Team's Chances of Winning. Journal of Quantitative Analysis in Sports, 12, 51–72.

Diniz, M. A., Izbicki, R., Lopes, D., & Salasar, L. E. (2019). Comparing probabilistic predictive models applied to football. Journal of the Operational Research Society, 70(5), 770-782.

Ellefsrød, M. B. (2013). The betting machine: Using in-depth match statistics to compute future probabilities of football match outcomes using the Gibbs sampler (Master's thesis, Institutt for datateknikk og informasjonsvitenskap).

Esme, E., & Kiran, M. S. (2018). Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm. International Journal of Machine Learning and Computing, 8(1), 26-32.

Fahey-Gilmour, J., Dawson, B., Peeling, P., Heasman, J., & Rogalski, B. (2019). Multifactorial analysis of factors influencing elite Australian football match outcomes: a machine learning approach. Journal of Sports Analytics, 18(3).

Głowania, S., Kozak, J., & Juszczuk, P. (2023). Knowledge Discovery in Databases for a Football Match Result. Electronics, 12(12), 2712.

Goller, D., Knaus, M. C., Lechner, M., & Okasa, G. (2021). Predicting match outcomes in football by an Ordered Forest estimator. A modern guide to sports economics, 335.

Goldsberry, K. (2012). Courtvision: New Visual and Spatial Analytics for the NBA. In 2012 MIT Sloan Sports Analytics Conference.

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. Journal of quantitative analysis in sports, 15(4), 271–287.

Haruna, U., Maitama, J. Z., Mohammed, M., & Raj, R. G. (2021, November). Predicting the outcomes of football matches using machine learning approach. In International Conference on Informatics and Intelligent Applications (pp. 92-104). Cham: Springer International Publishing.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & 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), 798-817.

Hubáˇcek, O., Šourek, G., & Železny, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, 108(1), 29–47.

Hsu, Y.-C. (2021). Using convolutional neural network and candlestick representation to predict sports match outcomes. Applied Sciences, 11(14), 6594.

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

Igiri, C. P. (2015). Support vector machine–based prediction system for a football match result.

Johnson, R., & Brown, S. (2018). Neural Networks for Predicting Soccer Game Outcomes. International Journal of Sports Data Science, 2(2), 112-125.

Jensen, S. T., B. B. McShane, and A. J. Wyner. (2009). Hierarchical Bayesian Modeling of Hitting Performance in Baseball. Bayesian Analysis, 4(4), 631–652.

Kamiri, J. and Mariga, G., 2021. Research methods in machine learning: A content analysis. Int. J. Comput. Inf. Technol, 10(2), pp.78-91.

Keshtkar Langaroudi, M., & Yamaghani, M. (2019). Sports result prediction based on machine learning and computational intelligence approaches: A survey. Journal of Advances in Computer Engineering and Technology, 5(1), 27-36.

Khan, S., & Kirubanand, V. B. (2019). Comparing machine learning and ensemble learning in the field of football. International Journal of Electrical and Computer Engineering, 9(5), 4321.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. Journal of the Royal Statistical Society Series A: Statistics in Society, 178(1), 167-186.

Koopman, S. J., & Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. International Journal of Forecasting, 35(2), 797-809.

Lee, J., Kim, J., Kim, H., & Lee, J. S. (2022). A Bayesian Approach to Predict Football Matches with Changed Home Advantage in Spectator-Free Matches after the COVID-19 Break. Entropy, 24(3), 366.

Leushuis, C. (2018). Beating the Odds-A State Space Model for predicting match results in the Australian Football League. MaRBLe, 2.

Malini, P., & Qureshi, B. (2022). A deep learning framework for temperature forecasting. In 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), pages 67–72. IEEE.

Michael, N. (2005). Artificial Intelligence: A Guide to Intelligent Systems. Pearson Education Limited. England.

Miller, A., L. Bornn, R. Adams, and K. Goldsberry. (2014). Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball. In International Conference on Machine Learning, pp. 235–243.

Mohandas, A., Ahsan, M., & Haider, J. (2023). Tactically Maximize Game Advantage by Predicting Football Substitutions Using Machine Learning. Big Data and Cognitive Computing, 7(2), 117.

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. Journal of sports sciences, 32(20), 1881-1887.

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. Future Internet, 15(5), 174.

Muszaidi, M., Mustapha, A. B., Ismail, S., & Razali, N. (2022, June). Deep Learning Approach for football match classification of English Premier League (EPL) based on full-time results. In Proceedings of the 7th International Conference on the Applications of Science and Mathematics 2021: Sciemathic 2021 (pp. 339-350). Singapore: Springer Nature Singapore.

Nestoruk, R., & Slowinski, G. (2021). Prediction of Football Games Results. In CS&P (pp. 156-165).

Nyquist, R., & Pettersson, D. (2017). Football match prediction using deep learning. Department of Electrical Engineering, CHALMERS UNIVERSITY OF TECHNOLOGY, Gothenburg, Sweden 2017.

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with Bayesian network in Spanish League-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 812.

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

Rahman, M. A. A., Mustapha, A., Razali, N., & Fauzi, R. (2018). Bayesian approach to classification of football match outcome. International Journal of Integrated Engineering, 10(6).

Razali, N., Mustapha, A., Utama, S., & Din, R. (2018, May). A review on football match outcome prediction using Bayesian networks. In Journal of Physics: Conference Series (Vol. 1020, No. 1, p. 012004). IOP Publishing.

Razali, N., Mustapha, A., Mustapha, N., & Clemente, F. M. (2021). A Bayesian approach for major European football league match prediction. International Journal of Nonlinear Analysis and Applications, 12(Special Issue), 971-980.

Ren, Y., & Susnjak, T. (2022). Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index. arXiv preprint arXiv:2211.15734.

Robberechts, P., Van Haaren, J., & Davis, J. (2021, August). A bayesian approach to in-game win probability in soccer. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3512-3521).

Qamar, U., Niza, R., Bashir, S. and Khan, F.H., 2016. A majority vote based classifier ensemble for web service classification. *Business & Information Systems Engineering*, *58*, pp.249-259.

Santos-Fernandez, E., Wu, P., & Mengersen, K. L. (2019). Bayesian statistics meets sports: a comprehensive review. Journal of Quantitative Analysis in Sports, 15(4), 289-312.

Spearman, W. (2018, February). Beyond expected goals. In Proceedings of the 12th MIT sloan sports analytics conference (pp. 1-17).

Statista Research Department (2023). Total amount wagered on European soccer worldwide in the 2020/2021 season, by league. Source: https://www.statista.com/statistics/1263462/value-betting-on-european-soccer/#:~:text=When%20looking%20at%20the%20global,reached%20roughly%2042.1%20billion%20euros.

Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics. Applied Sciences, 10(1), 46.

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. Transactions on knowledge and data engineering, 10(10), 1-13.

The Analyst. 2023, August 8. *The Strongest Leagues in World Football: How the Opta Power Rankings Can Help*. Available at: https://theanalyst.com/eu/2023/08/the-strongest-leagues-in-world-football-opta-power-rankings/ [Accessed 20 December 2023]

Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the English Premier League. Doctoral dissertation, Ph. D. dissertation, Stanford.

Vaidya, S., Sanghavi, H. and Gevaria, K., 2016. Football match winner prediction. Int. J. Comp. Appl, 154, pp.31-33

Vashist, M., Bahl, V., Goel, A. and Sengar, N., 2021. Full Time Result Prediction using Ensemble Techniques. Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146, 7(3), pp.38-42

YILDIZ, B. F. (2020). Applying decision tree techniques to classify European football teams. Journal of Soft Computing and Artificial Intelligence, 1(2), 86–91.

Zhang, Q., et al. (2021). Sports match prediction model for training and exercise using attention-based LSTM network. Digital Communications and Networks.