University of Hertfordshire UH

School of Physics,
Engineering and
Computer Science

MSc Data Science Project
7PAM2002-0509-2022
Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

PREMIER LEAGUE BIG SIX TEAMS PREDICTION

Student Name and SRN:

IBRAHIM BAL - 21071305

Supervisor: John Evans

Date Submitted: 07.09.2023

Word Count: 8987

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science **in Data Science** at the University of Hertfordshire.

I have read the detailed guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6)

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Ibrahim BAL

Student Name signature Ibrahim BAL

Student SRN number: 21071305

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

Abstract

The top six teams in the Premier League—Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and Tottenham Hotspur—are the subject of this dissertation's in-depth analysis of Premier League games. In order to find patterns, trends, and probable factors influencing match outcomes, the project will analyze numerous match statistics, including performance metrics, possession, passing, shooting, and goalkeeping data. Additionally, based on the data gathered, this project aims to create a machine learning model to forecast match results. The study explains the methodology, results, and ramifications of the data collection process and sheds light on the nuances of Premier League games and the promise of data-driven insights in sports analysis. In terms of win, draw or lose of a game, this initiative has a 93% accuracy rate.

1. INTRODUCTION

Football is an unmatched global phenomenon in the enthralling world of sports. A symphony of talent, passion, and drama that cuts beyond boundaries, languages, and cultures, it is more than just a game. My focus is firmly placed on the Premier League as I begin this research adventure since it is the scene where the greatest football stories are told. The exciting mystique of game predictions is also concealed inside this spectacle. And within this spectacle, lies the electrifying enigma of game predictions. Among all sports, soccer prediction is probably the most widely and deeply researched area (Byungho Min, 2008).

Football, commonly referred to as the "beautiful game," is spoken all over the world. It draws people from other nations together, igniting both joy and grief, building enduring team allegiances, and acting as a platform for international discussions. With its illustrious

The Premier League is the epitome of what makes football so alluring. Twenty teams participate in this match of wits, talent, and tenacity, each with their own distinct character and ethos. The "big six" teams - Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and Tottenham Hotspur - stand out the most in this vivid tapestry of skill. They are the guardians of a legacy that motivates countless people.

This football saga's unique practice of game predictions is at its core. The unpredictable nature of football is its fundamental soul. In 90 minutes, anything can happen: a skilfully placed pass, a thunderous strike, or a goal-saving tackle can all change the outcome of a match. Even for specialists, (Byungho Min, 2008)claim that it is very difficult to forecast the precise outcomes of particular matches. However, the difficulty of making predictions in this changing environment is a quest that is of utmost significance.

Predictions for football go beyond just passion. They are intertwined with the thrilling world of sports betting, where enthusiasts and experts alike stake their judgment, giving each goal, each turn, and each outcome greater significance. Football has enormous financial stakes, with millions changing hands as wagers are made on game outcomes, top goal scorers, and other complex aspects of the game.

Football predictions can be entertaining for some people and add to the matchday atmosphere. Others view it as a strategic investment where their financial objectives and game understanding converge. Every match has the potential to be a turning point in fortunes since predictions serve as the conduit connecting the viewer to the spectacle regardless of the motivation.

With a focus on the "big six" teams, I set out to explore the enormous ocean of Premier League football in my dissertation. My goal is to unravel the complex dance of tactics, methods, and individual brilliance, not just to forecast results. This voyage aims to reveal the relevance of forecasts, the intricate network of wagers, and the financial currents that flow around football in addition to the excitement of the game.

The Premier League is more than simply a league, and football predictions are more than just educated guesses. They serve as the connecting elements in the intricate tapestry that is football's past, present, and future. I encourage you to accompany me as I set out on this journey to discover the depths of football's unpredictable nature and the significance of forecasting its results.

1.1. Objective

The main goal of this research is to create a predictive model that will be able to reasonably predict the outcomes in terms of win, draw or lose of Premier League football games. The model will be developed using historical match data, which will include a variety of features including team performance metrics, individual statistics, recent form, venue, and other pertinent contextual variables. The study aims to develop a reliable prediction system by analyzing these features and their effects on match results.

1.2. Aim

With a focus on the top six teams—Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and Tottenham Hotspur—this dissertation's main objective is to analyze Premier League football games. The project aims to accomplish the following goals:

Data Collection and Preprocessing: Gather comprehensive match statistics for the selected teams, encompassing possession data, passing accuracy, shooting efficiency, and goalkeeping performance.

Exploratory Data Analysis: Uncover patterns, trends, and insights within the collected data to gain a better understanding of team performances and potential influencing factors.

Machine Learning Model Development: Develop a machine learning model capable of predicting match outcomes based on historical data, utilizing a combination of performance metrics, possession, passing, shooting, and goalkeeping data.

Evaluation and Interpretation: Assess the performance of the developed machine learning model in predicting match outcomes and interpret the significance of various features in the prediction process.

Implications and Future Research: Discuss the implications of the findings for sports analysis, coaching strategies, and decision-making. Identify potential areas for further research and improvement in data-driven sports analysis.

By achieving these objectives, this project aims to contribute to the field of sports analysis by shedding light on the intricate dynamics of Premier League matches and demonstrating the potential of data-driven insights in enhancing our understanding of football performances and outcomes.

2. Literature Review

Football stands out as an intriguing focus point in the large field of sports prediction due to its complex and multidimensional nature. Football is a worldwide phenomenon that has naturally drawn a lot of study interest, leading to a wide range of studies that explore the art and science of match prediction. An examination of the body of research on football match predictions is conducted in this lengthy literature study. This review will include an analysis of the methodology used in earlier studies and the factors taken into account during the process. Additionally, a comparison analysis will be carried out, matching the features chosen and included in my own predictive model with those previously investigated and explored.

Football prediction has been the subject of statistical analysis. Many academics proposed their own models or methods for analyzing football game outcomes. Typically, they demonstrated how accurately their methods predicted football game outcomes. Soccer match results have been analyzed using a variety of models and techniques, including Poisson regression models, a logistic regression model based on seed positions, and an updating procedure for the intra-match winning probability ((Jean-Marc Falter, 2000); (David Forrest, 2006); (Martin Crowder, 2002); (Mark J. Dixon, 2002); (Constantinou, 2012)). While the majority of these works also make some forecasts, their main emphasis is on statistical analysis of football match results. Using improvements to the independent Poisson model, (Martin Crowder, 2002) concentrated on modelling the soccer teams competing in the English Football Association League. A parametric outcome prediction model was created and fitted to English league and cup football by (Mark J. Dixon, 2002). The data format and the dynamic nature of team performances hampered the technique, which was based on a Poisson regression model. A Bayesian network model for predicting association football games was presented by (Constantinou, 2012). During the season, predictions regarding the results of English Premier League (EPL) matches were made using the model (pi-football).

Football prediction has been incorporated into machine learning algorithms and similar methodologies. Examples include fuzzy logic representation with genetic and neural optimization methods in tuning the fuzzy model, decision tree, naive Bayesian learning, expert Bayesian network, K-nearest neighbour, and more ((A. Joseph, 2006); (Flitman, 2006); (Grunz, 2012)). They predicted the outcomes of league or tournament matches using training data from past match results, such as win/draw/lose or scores. Bayesian networks (BNs) were used by (A. Joseph, 2006) to predict the result (win, lose, or draw) of matches played by the Tottenham Hotspur Football Club and to compare them with other machine learning techniques. Given that the study's assumptions disadvantage BNs in several crucial ways, the results were even more spectacular for them. A model created by (Flitman, 2006) can accurately forecast the outcome of Australian Football League games as well as the likelihood that the predicted outcome would occur. This model was created using a genetically altered neural network to predict who will win and a linear program optimization to estimate the likelihood of that outcome under the scoring rules for tipping competitions. In conjunction with an in-game time-series approach, Bayesian inference, rule-based reasoning, and other techniques, to identify tactical tendencies in soccer matches, (Grunz, 2012) developed a hierarchical design of artificial neural networks. Different tactical patterns and variations on these patterns can be recognized by the hierarchical architecture. One of the most practical ANN techniques, self-organizing maps (SOM), was used to work on defense player organization.

One of the most pertinent and closely aligned research studies to my project is that of (S. Mohammad Arabzad, 2014)research. In this research, the Iran Pro League (IPL) serves as the focal point, utilizing historical data from previous seasons to facilitate match outcome predictions. Notably, the research harnesses Artificial Neural Networks (ANN), with a specific focus on the Multilayer Perceptron (MLP) technique. The results derived from this model remarkably converge with actual outcomes, underscoring the effectiveness of the approach. Inspired by the outcomes of this study, I adopted a similar methodology in my research, albeit with some tailored adjustments. While still leveraging the ANN framework, I opted to modify certain features originally employed in this research, replacing them with new variables that I deemed more suitable based on my extensive investigations. Additionally, I introduced supplementary features to enhance the predictive accuracy of my model. In essence, my research is similar to this one in that it builds on its conclusions while also developing the methods to develop a strong predictive model.

3. METHODOLOGY

3.1. Data Description and Pre-Processing

This section provides an overview of the dataset used in the project, detailing its source, pre-processing procedures, and a comprehensive analysis conducted on the data.

3.2. Data

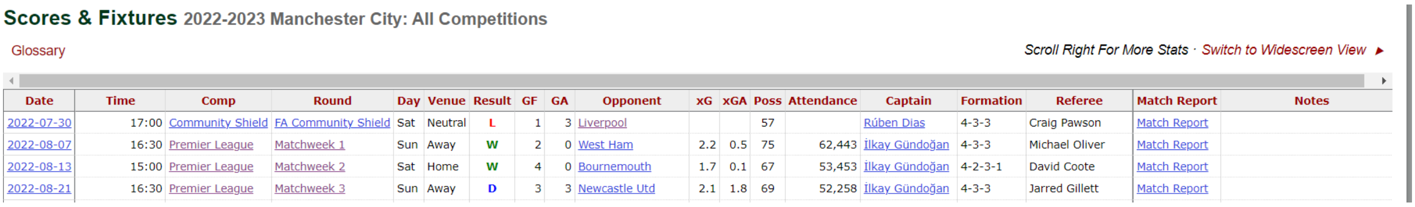
The dataset employed in this project is sourced from fbref.com, a platform that offers a wealth of football-related data for analysis.

3.2.1. Tables

The dataset comprises multiple tables, each contributing distinct information to the analysis. The following subsections outline the tables and their relevant features:

3.2.1.1. All Teams Data

The "all_teams_data" table presents a comprehensive compilation of detailed information and statistics pertaining to each team's individual games across various competitions. This table encapsulates a diverse range of data related to the selected season, offering insights into each team's performance in different competitions. The features encompass a wide spectrum of match-related metrics and statistics associated with each team's participation in their respective competitions. The example of a table is shown below and the columns are:



Scores & Fixtures 2022-2023 Manchester City: All Competitions

Glossary

Scroll Right For More Stats · Switch to Widescreen View ▶

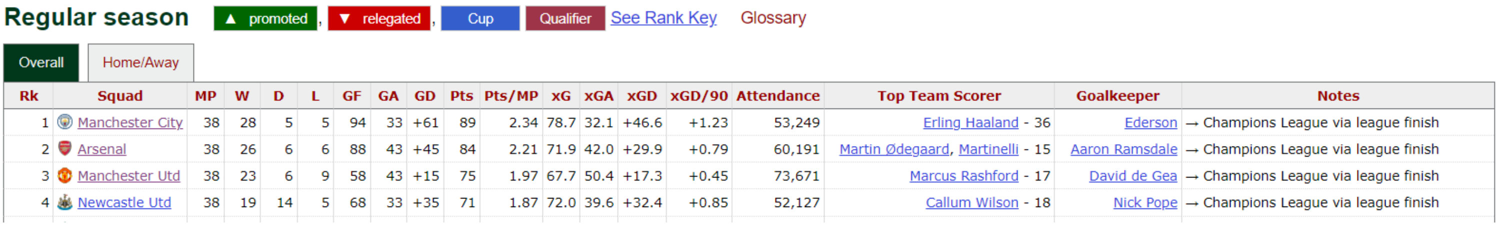
Date	Time	Comp	Round	Day	Venue	Result	GF	GA	Opponent	xG	xGA	Poss	Attendance	Captain	Formation	Referee	Match Report	Notes
2022-07-30	17:00	Community Shield	FA Community Shield	Sat	Neutral	L	1	3	Liverpool			57		Rúben Dias	4-3-3	Craig Pawson	Match Report	
2022-08-07	16:30	Premier League	Matchweek 1	Sun	Away	W	2	0	West Ham	2.2	0.5	75	62,443	İlkay Gündoğan	4-3-3	Michael Oliver	Match Report	
2022-08-13	15:00	Premier League	Matchweek 2	Sat	Home	W	4	0	Bournemouth	1.7	0.1	67	53,453	İlkay Gündoğan	4-2-3-1	David Coote	Match Report	
2022-08-21	16:30	Premier League	Matchweek 3	Sun	Away	D	3	3	Newcastle Utd	2.1	1.8	69	52,258	İlkay Gündoğan	4-3-3	Jarred Gillett	Match Report	

Figure 1

- Date: The date of game
- Time: The time of game
- Comp: The Name of the competition
- Round: The round or phase of competition
- Day: The day of the week.
- Venue: The venue of the match.
- Result: The match result.
- GF: Goals scored by the team.
- GA: Goals conceded by the team.
- Opponent: The opposing team.
- xG: Expected goals for the team.
- xGA: Expected goals against the team.
- Poss: Possession percentage.
- Attendance: Number of attendees at the match.
- Captain: The team captain.
- Formation: The team's formation.
- Referee: The match referee.
- Match Report: Link to the match report.
- Notes: Additional notes about the match.

3.2.1.2. Premier League Stats

The "premier_league_stats" table encompasses statistics related to Premier League squads. The features within this table include: The example of a table is shown below and the columns are:



Regular season ▲ promoted, ▼ relegated, Cup Qualifier See Rank Key Glossary

Overall Home/Away

Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90	Attendance	Top Team Scorer	Goalkeeper	Notes
1	Manchester City	38	28	5	5	94	33	+61	89	2.34	78.7	32.1	+46.6	+1.23	53,249	Erling Haaland - 36	Ederson	→ Champions League via league finish
2	Arsenal	38	26	6	6	88	43	+45	84	2.21	71.9	42.0	+29.9	+0.79	60,191	Martin Ødegaard, Martinelli - 15	Aaron Ramsdale	→ Champions League via league finish
3	Manchester Utd	38	23	6	9	58	43	+15	75	1.97	67.7	50.4	+17.3	+0.45	73,671	Marcus Rashford - 17	David de Gea	→ Champions League via league finish
4	Newcastle Utd	38	19	14	5	68	33	+35	71	1.87	72.0	39.6	+32.4	+0.85	52,127	Callum Wilson - 18	Nick Pope	→ Champions League via league finish

Figure 2

- Rk: The team’s ranking.
- Squad: The team’s name.
- MP: Matches played.
- W: Wins.
- D: Draws.
- L: Losses.

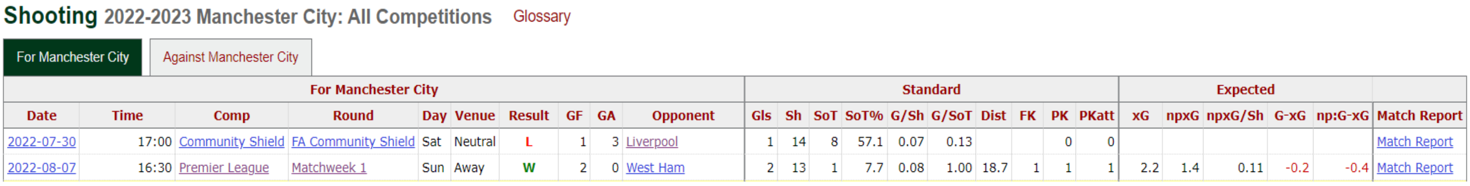
- GF: Goals scored.
- GA: Goals conceded.
- GD: Goal difference.
- Pts: Total points.
- Pts/MP: Points per match.
- Xg: Expected goals.
- Xga: Expected goals against.
- Xgd: Expected goal difference.
- Xgd/90: Expected goal difference per 90 minutes.
- Attendance: Attendance at matches.
- Top Team Scorer: Top goal scorer of the team.
- Goalkeeper: Goalkeeper details.
- Notes: Additional notes.

3.2.1.3. Premier League Last Stats

The “premier_league_last_stats” table provides the latest statistics for Premier League squads. It shares features with the “premier_league_stats” table and serves the same purpose of offering comprehensive squad statistics.

3.2.1.4. Shooting Table

The table below shows the shooting stats for each game for a certain team.



Shooting 2022-2023 Manchester City: All Competitions   Glossary

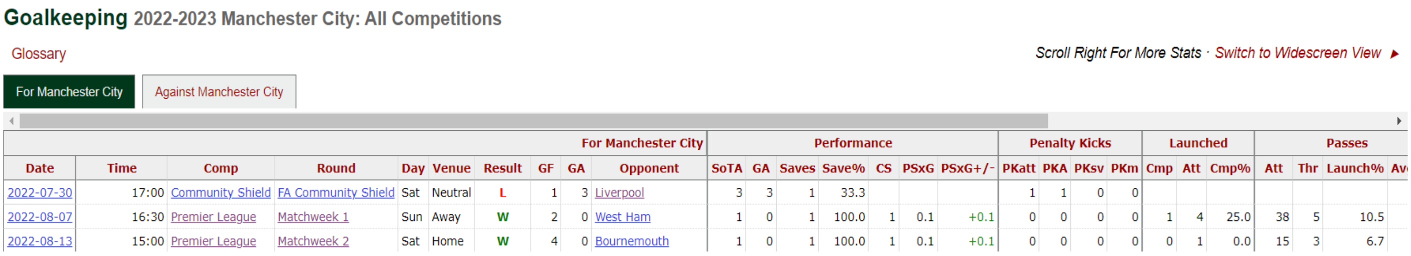
For Manchester City | Against Manchester City

For Manchester City											Standard										Expected					
Date	Time	Comp	Round	Day	Venue	Result	GF	GA	Opponent	Gls	Sh	SoT	SoT%	G/Sh	G/SoT	Dist	FK	PK	PKatt	xG	npxG	npxG/Sh	G-xG	np:G-xG	Match Report
2022-07-30	17:00	Community Shield	FA Community Shield	Sat	Neutral	L	1	3	Liverpool	1	14	8	57.1	0.07	0.13			0	0						Match Report
2022-08-07	16:30	Premier League	Matchweek 1	Sun	Away	W	2	0	West Ham	2	13	1	7.7	0.08	1.00	18.7	1	1	1	2.2	1.4	0.11	-0.2	-0.4	Match Report

Figure 3

3.2.1.5. Goalkeeping Table

The table below shows the goalkeeping stats for each game for a certain team.



Goalkeeping 2022-2023 Manchester City: All Competitions

Glossary

Scroll Right For More Stats · Switch to Widescreen View ▶

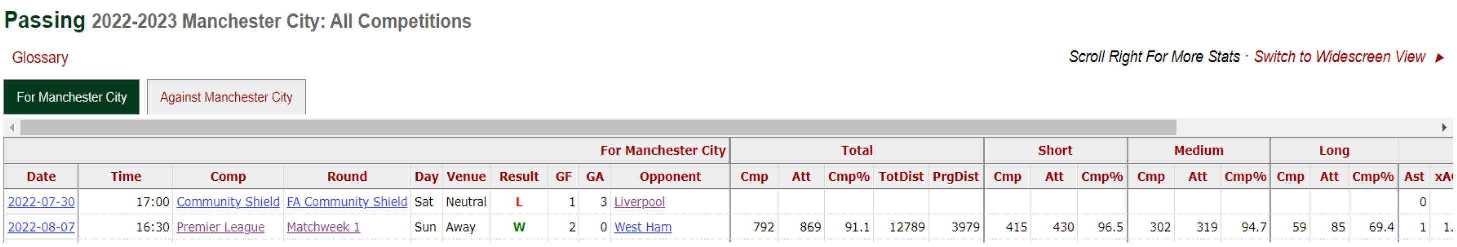
For Manchester City | Against Manchester City

For Manchester City										Performance							Penalty Kicks				Launched			Passes			
Date	Time	Comp	Round	Day	Venue	Result	GF	GA	Opponent	SoTA	GA	Saves	Save%	CS	PSxG	PSxG+/-	PKatt	PKA	PKsv	PKm	Cmp	Att	Cmp%	Att	Thr	Launch%	Av
2022-07-30	17:00	Community Shield	FA Community Shield	Sat	Neutral	L	1	3	Liverpool	3	3	1	33.3				1	1	0	0							
2022-08-07	16:30	Premier League	Matchweek 1	Sun	Away	W	2	0	West Ham	1	0	1	100.0	1	0.1	+0.1	0	0	0	0	1	4	25.0	38	5	10.5	
2022-08-13	15:00	Premier League	Matchweek 2	Sat	Home	W	4	0	Bournemouth	1	0	1	100.0	1	0.1	+0.1	0	0	0	0	0	1	0.0	15	3	6.7	

Figure 4

3.2.1.6. Passing Table

The table below shows the passing stats for each game for a certain team.



Passing 2022-2023 Manchester City: All Competitions

Glossary

Scroll Right For More Stats · Switch to Widescreen View ▸

For Manchester City | Against Manchester City

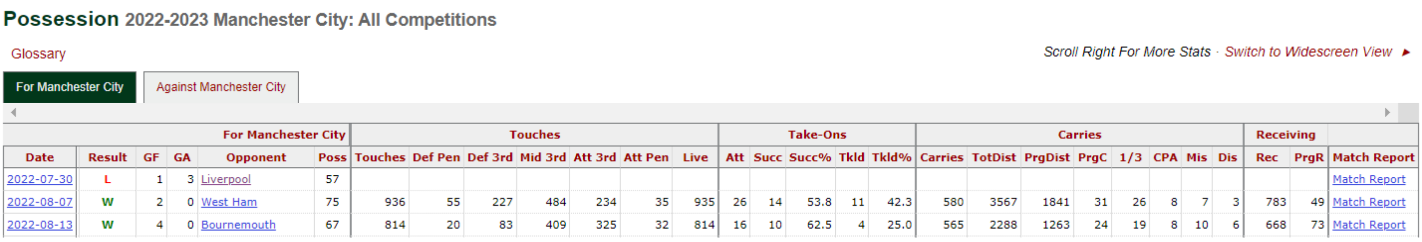
For Manchester City										Total					Short			Medium			Long				
Date	Time	Comp	Round	Day	Venue	Result	GF	GA	Opponent	Cmp	Att	Cmp%	TotDist	PrgDist	Cmp	Att	Cmp%	Cmp	Att	Cmp%	Cmp	Att	Cmp%	Ast	xA
2022-07-30	17:00	Community Shield	FA Community Shield	Sat	Neutral	L	1	3	Liverpool															0	
2022-08-07	16:30	Premier League	Matchweek 1	Sun	Away	W	2	0	West Ham	792	869	91.1	12789	3979	415	430	96.5	302	319	94.7	59	85	69.4	1	1.

Figure 5

3.2.1.7. Possession Table

The table below shows the possession stats for each game for a certain team.

3.2.1.8. Main Dataset



Possession 2022-2023 Manchester City: All Competitions

Glossary

Scroll Right For More Stats · Switch to Widescreen View ▸

For Manchester City | Against Manchester City

For Manchester City						Touches							Take-Ons					Carries								Receiving		
Date	Result	GF	GA	Opponent	Poss	Touches	Def Pen	Def 3rd	Mid 3rd	Att 3rd	Att Pen	Live	Att	Succ	Succ%	Tkld	Tkld%	Carries	TotDist	PrgDist	PrgC	1/3	CPA	Mis	Dis	Rec	PrgR	Match Report
2022-07-30	L	1	3	Liverpool	57																							Match Report
2022-08-07	W	2	0	West Ham	75	936	55	227	484	234	35	935	26	14	53.8	11	42.3	580	3567	1841	31	26	8	7	3	783	49	Match Report
2022-08-13	W	4	0	Bournemouth	67	814	20	83	409	325	32	814	16	10	62.5	4	25.0	565	2288	1263	24	19	8	10	6	668	73	Match Report

Figure 6

The primary dataset for making predictions is the "premier_league_data.csv" file, which forms the cornerstone of this study. To enhance the dataset's richness and utility, extensive data mining was performed. This process involved the extraction and concatenation of relevant features from various supplementary datasets. The subsequent dataset, meticulously crafted with a focus on key attributes, was then harnessed as the foundation for constructing and training my predictive model. The intricate details of this mining process will be comprehensively elucidated in the following sections.

The columns within this dataset include:

- Date: The date of the game
- Team: The name of the team.
- WeekNumber: The number of the week game played.
- Opponent: Name of the opponent team.
- Venue: Venue of the game.

- GF: Goals scored.
- GA: Goals conceded.
- xGA: Expected goals against.
- Poss: Possession(%) of the home team.
- Formation: Formation of the home team
- Captain: Name of the captain of the home team
- Referee: Name of the referee of the game
- Sot%: Shots on target (%) of home team.
- Touches: Number of touches made by home team
- Def Touch: Number of defensive touches made by home team
- Mid Touch: Number of middle area touches made by home team
- Att Touch: Number of attacking touches made by home team
- Short Pass: Short pass (%) made by home team
- Medium Pass: Medium pass(%) made by home team
- Long Pass: Long pass(%) made by home team
- Save: Save percentage of goalkeeper.
- opp: Opponent team’s cat code.
- ref: Referee’s cat code.
- Outcome: The result of the game’s cat code. (-1 for lose, 0 for draw, 1 for win.)
- Ven: Cat code of the venue. (0 for away, 1 for home)

The columns and example of a row from main dataset is shown below.

Date	**Team**	**WeekNumber**	**ScoreDiff**	**Opponent**	**Venue**	**GF**	**GA**	**xG**	**xGA**
5.08.2022	Arsenal	1	2	Crystal Palace	Away	2	0	1	1.2

Poss	**Formation**	**Captain**	**Referee**	**Sot%**	**Touches**	**Def Touch**	**Mid Touch**	**Att Touch**
44	4-3-3	Martin Ã˜degaard	Anthony Taylor	44.4	709	275	322	206

Short Pass	**Medium Pass**	**Long Pass**	**Save%**
88.8	88.4	72.7	50

The research makes use of the detailed data included in these databases to derive insights, examine patterns, and create forecasting models for Premier League match results. The initiative seeks to provide a holistic perspective on team performance and factors impacting match results by merging data from many tables.

3.3. Data exploration and feature engineering

3.3.1. Statistical Analysis

To begin with, I conducted a comprehensive analysis focusing exclusively on the renowned "big six" teams within the Premier League. Through this analysis, I aimed to carefully select the most pertinent features to incorporate into my predictive model. My initial step involved a meticulous comparison between the final results of the 2020-2021 and 2021-2022 seasons.

```
2022-2023
                Team   W   D   L  GF  GA  Points  Rank
0   Manchester City  28   5   5  94  33      89     1
1           Arsenal  26   6   6  88  43      84     2
2    Manchester Utd  23   6   9  58  43      75     3
4         Liverpool  19  10   9  75  47      67     5
7         Tottenham  18   6  14  70  63      60     8
11          Chelsea  11  11  16  38  47      44    12
2021-2022
               Team   W   D   L  GF  GA  Points  Rank
0  Manchester City  29   6   3  99  26      93     1
1        Liverpool  28   8   2  94  26      92     2
2          Chelsea  21  11   6  76  33      74     3
3        Tottenham  22   5  11  69  40      71     4
4          Arsenal  22   3  13  61  48      69     5
5   Manchester Utd  16  10  12  57  57      58     6
```

3.3.1.1. Season 2022-2023

The 2022-2023 season witnessed remarkable performances by various teams. Manchester City secured the top position with 89 points, illustrating their consistent dominance. Arsenal, claiming the second rank with 84 points, showcased their resurgence. Liverpool, finishing fifth with 67 points, faced challenges that impacted their performance. Tottenham and Chelsea secured the eighth and twelfth positions, respectively, with 60 and 44 points. The varying performances of these teams highlighted the dynamic nature of the Premier League.

3.3.1.2. Season 2021-2022

In the 2021-2022 season, Manchester City clinched the top spot with 93 points, closely followed by Liverpool with 92 points. Chelsea secured the third rank with 74 points, while Tottenham claimed the fourth position with 71 points. Arsenal secured the fifth position with 69 points. This season showcased tight competition at the top, with only a single point separating the top two teams. As clearly seen from the tables, Chelsea made a big fall in one season. This will be analysed in detailed.
I generated four distinct plots to facilitate a thorough comparison of these statistics, enabling a clearer understanding of the disparities.

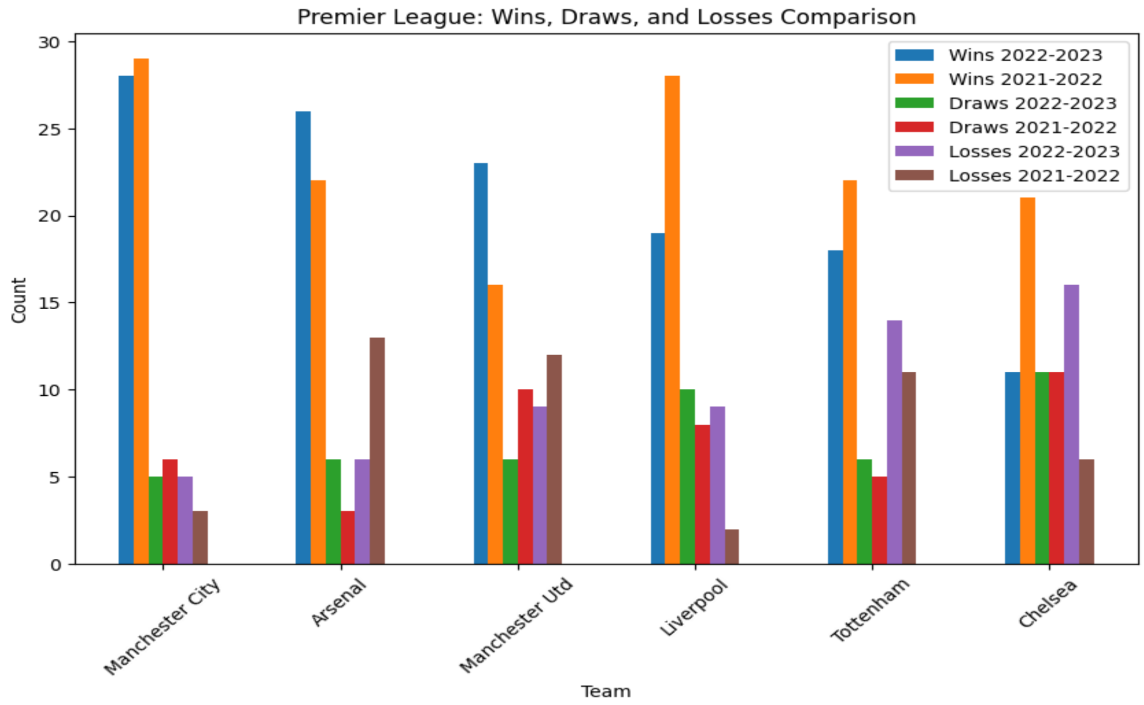


Figure 7

In Figure 7, we observe the variations in wins, draws, and losses between the two seasons. Notably, Chelsea and Liverpool experienced an increase in losses and a decrease in wins compared to the previous season. Meanwhile, Manchester City, despite a slight reduction in the number of victories, managed to secure significantly more wins than other teams while also having the fewest losses.

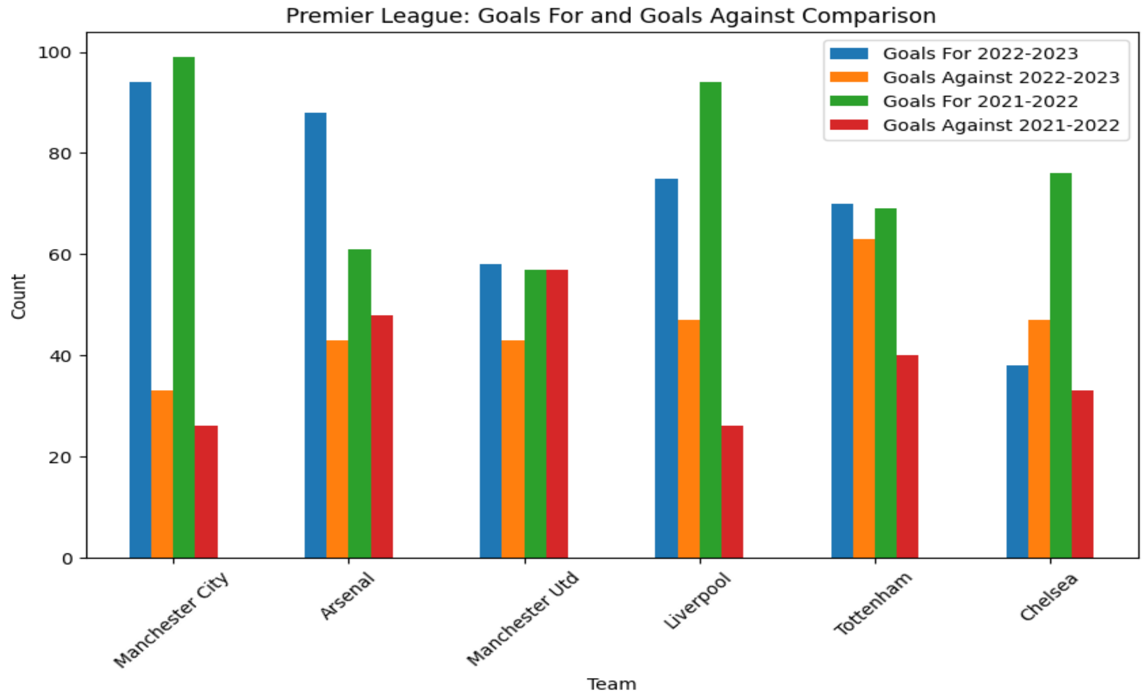


Figure 8

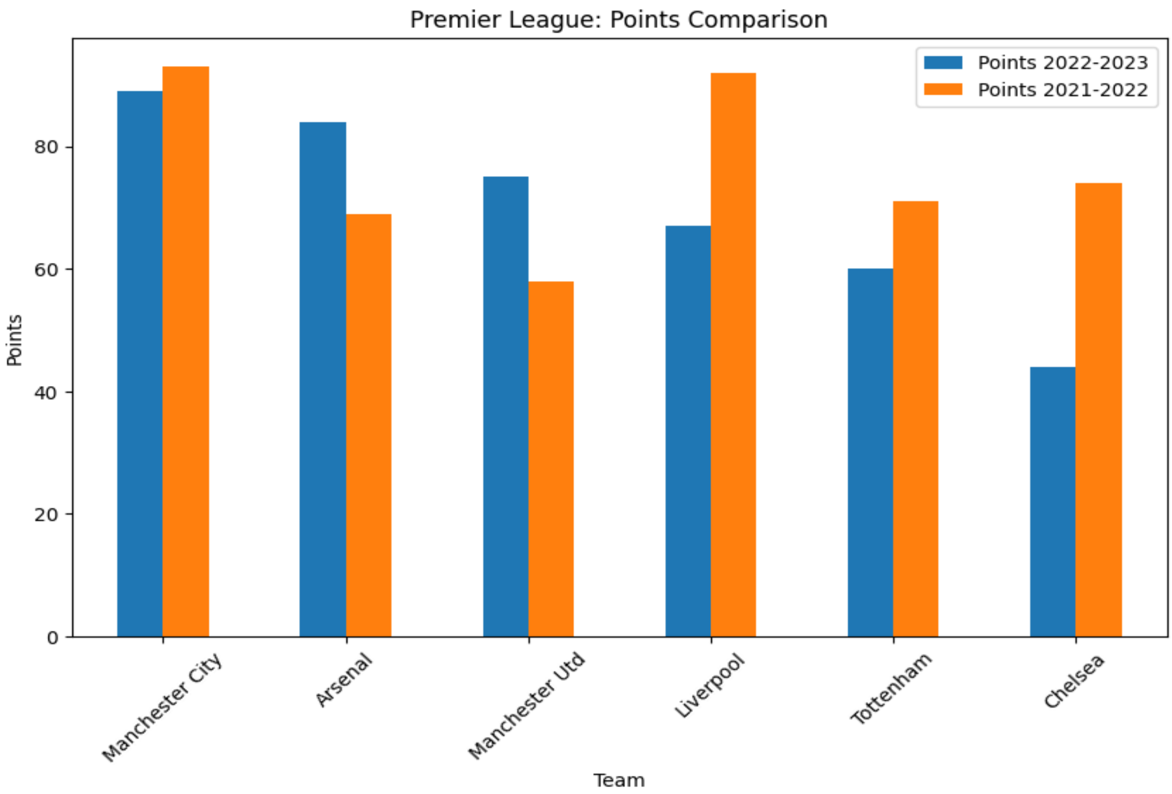


Figure 9

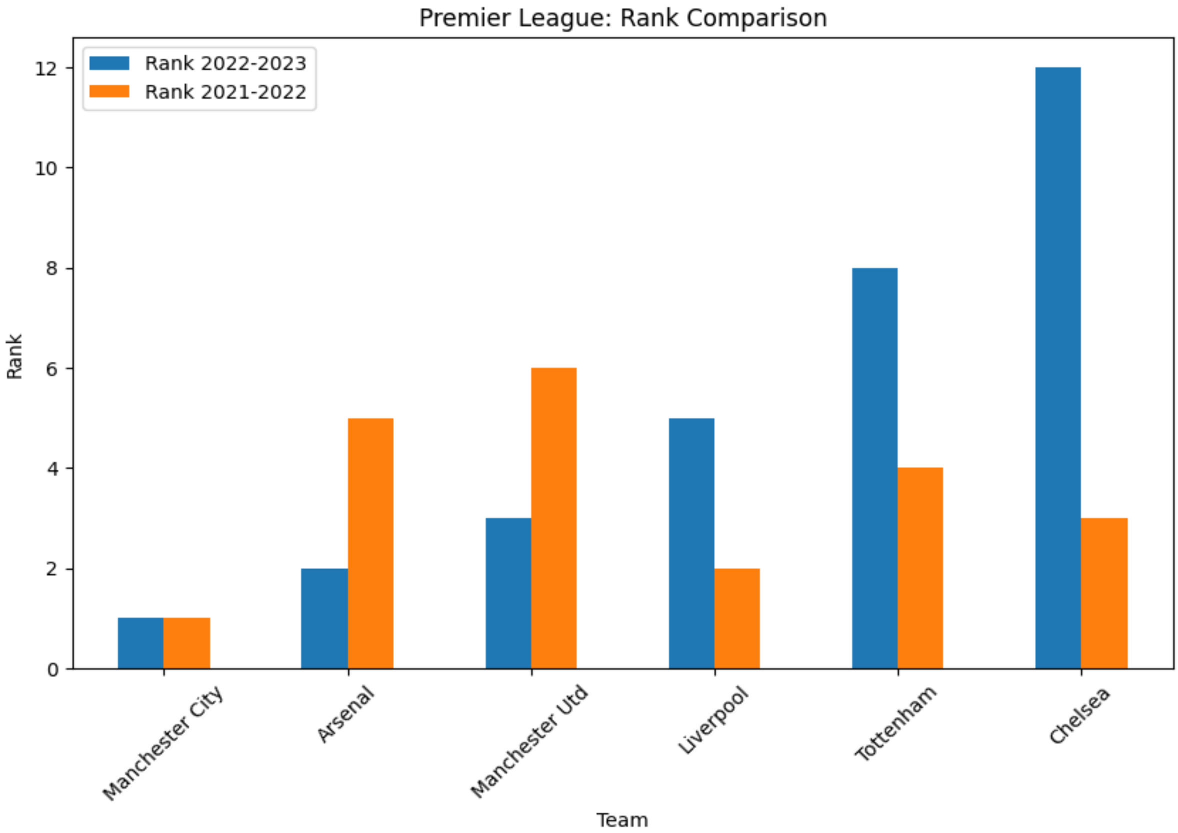


Figure 10

3.3.2. Feature Selection

I carefully selected the features that would be included in my predictive model by conducting in-depth analysis and study on the dataset. The goal of this technique was to identify the most important and impactful elements that could considerably improve the precision of match outcome predictions by methodically exploring the complexities of the dataset.

The process of feature selection resembled that of an experienced explorer traveling through a tangle of data points, each of which had the power to affect how football games turned out. I set out on a mission to uncover the hidden patterns and connections within the information using a combination of domain expertise, statistical methodology, and cutting-edge data mining tools.

This was a rigorous scientific endeavor rather than just picking qualities based on gut feeling or speculation. Each attribute was closely examined, and a rigorous statistical analysis was used to confirm its applicability. The objective was clear: to determine the characteristics with the best predictive value, those that may provide the most insightful understandings into the intricate world of football match outcomes.

It's important to note that this feature selection method involves both art and science. It called for a complex balancing act between data-driven insights and a detailed comprehension of football dynamics. Additionally, as the research went on, the features chosen changed to reflect the shifting football landscape and the new information learned through analysis.

In essence, feature selection served as my compass through the immense landscape of football data, directing me toward the most promising match result predictions. It perfectly embodied the meticulous craftsmanship needed in sports prediction, where the knack for picking the appropriate attributes is crucial to learning the tricks of football's unpredictable nature.

3.3.2.1. Venue

Over the past three decades, extensive research has been done in the field of sports on how the venue affects team performance. It is generally known that there is a phenomena known as "home advantage," which (Kerry S. Courneya, 1992) described as "the term used to describe the consistent finding that home teams in sport competitions win over 50% of the games played under a balanced home and away schedule" (Sarmento H, 2014), have repeatedly shown that individuals and teams frequently put on noticeably stronger performances when competing on their home field.

There is a recurring pattern where athletes and teams tend to perform better when playing in the comfortable surroundings of their home venue (Kerry S. Courneya, 1992); (Nevill, 1999) despite the fact that the significance of home advantage can vary across different sports (Jone, 2013). It is essential to comprehend and take into consideration this home advantage when creating predictive models for football game results. In order to use this feature in my model, it will be introduced to the main dataset.

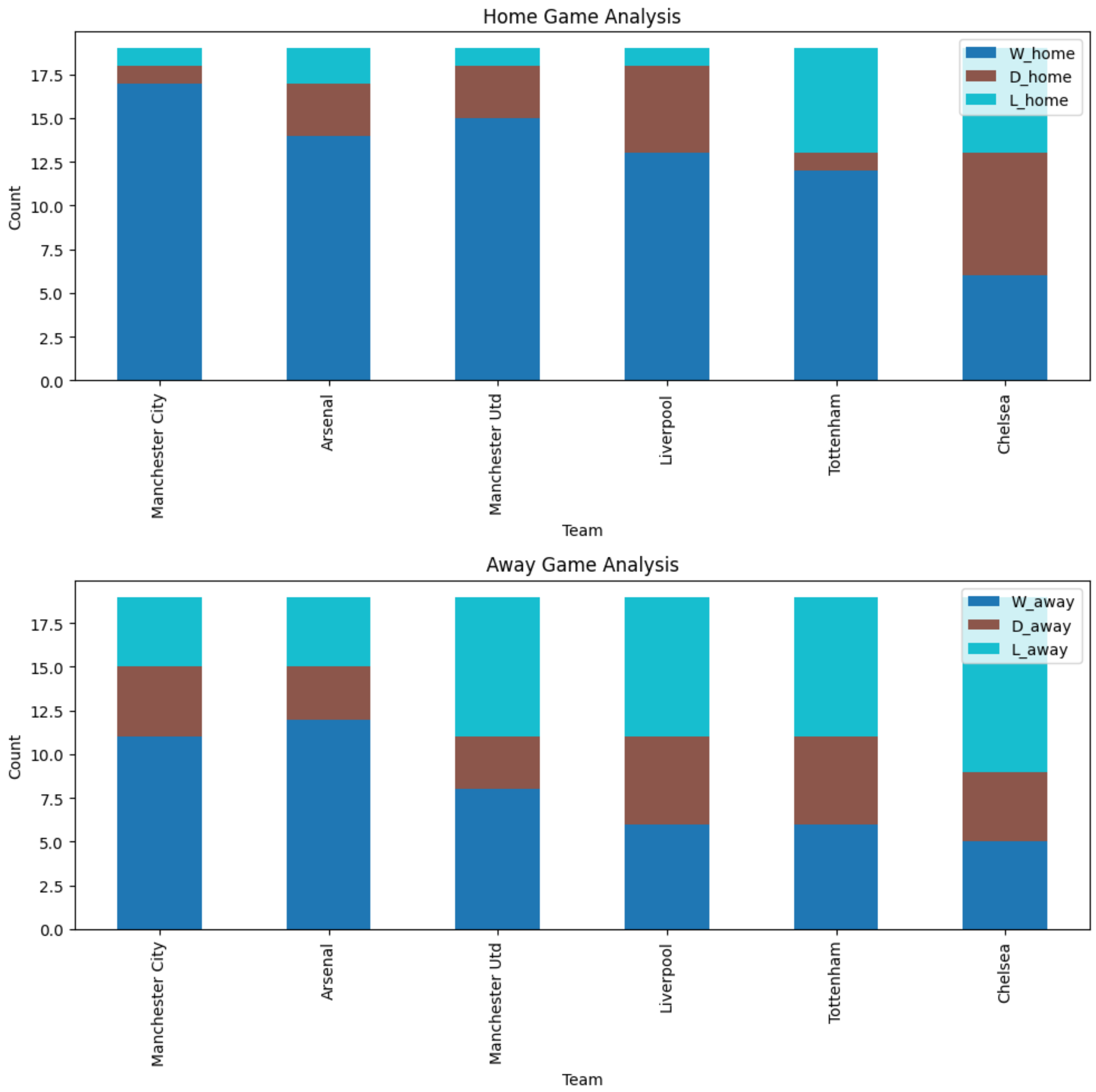


Figure 11

I made a plot (Figure 11) to examine how the venue affected team performance, notably in terms of wins, draws, and losses. In particular, Manchester United and Liverpool showed an astounding pattern, winning nearly 70% of their home games yet suffering defeats in less than 50% of their road games. This means that improving their performance in road games could greatly enhance their total performance. However, with significantly fewer victories on the road, Manchester City and Arsenal displayed a relatively lesser discrepancy in their performance between home and away games. Chelsea, on the other hand, showed that they could perform well everywhere, whether they were playing at home or away.

3.3.2.2. Shots on Target (%)

Shots on/off target are the very first thing that teams do before scoring a goal (Football Report). "Shots on/off target" statistics offer a picture of a team's attacking prowess and defensive resiliency in the complex world of football, where the game's dynamics can shift in a split second. These measurements provide insight into a team's power to create scoring opportunities while simultaneously emphasizing its ability to block the advances of its opponents. They perfectly capture the essence of football, a dynamic sport where everything can change quickly.

Beyond statistical study, understanding the significance of "shots on/off target" entails exploring the core of the game. It represents the conflict between defenders and goalkeepers who are determined to defend their territory and attackers who are trying to breach the defence. Every amazing moment, from a striker's tremendous attempt to a goalkeeper's daring save, is captured in these statistics.

Statistics pertaining to "shots on/off target" have a special and important place in the field of football analysis. They show the ups and downs of a game, the teams' deft moves, and the fan-favourite moments. The hopes, dreams, and drama that play out on the field with each shot made and saved are captured in these numbers, which are more than just data points. They represent the essence of the beautiful game.

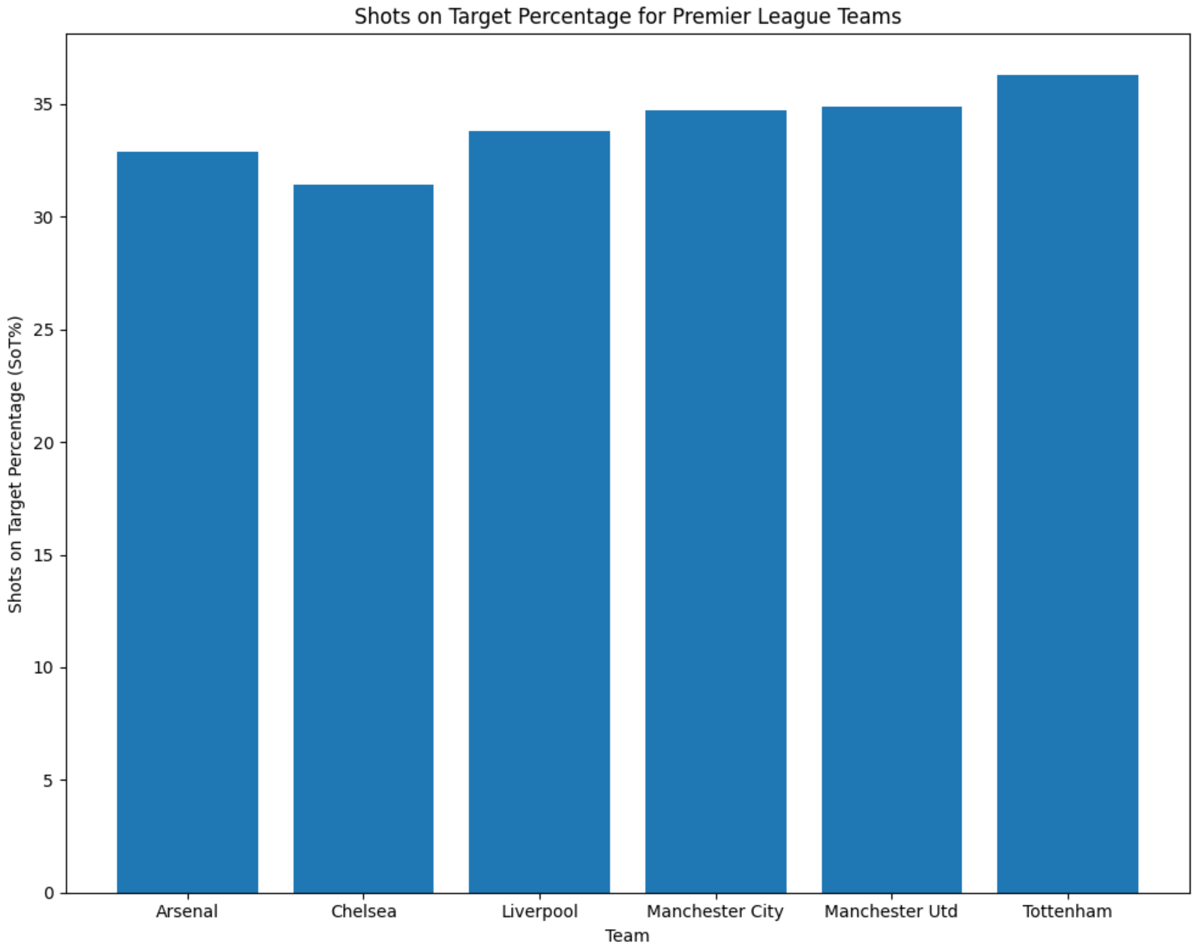


Figure 12

A revealing look at the intriguing dynamics of shooting accuracy in football is provided by Figure 12. It is noteworthy to note that Chelsea has the lowest Shots on Target Percentage (Sot%) out of the famed "big six" teams, which puts them in last place. Tottenham, which is now ranked eighth overall for the season, exhibits the highest Sot%, which is a surprising turn of events. An intriguing realization results from this observation: simply hitting the goal may not be enough to score if the shots are ineffective or the opponent's goalkeeper puts in an outstanding effort.

This information highlights the complex relationship between successful goal attempts and accurate shots. The capacity to hit the target is unquestionably a crucial trait in the game of football, since every action can affect the result. It is the foundation of offensive strategies, directly affecting the outcome of games. A key tenet that permeates throughout the sport is that more shots on aim usually equate to a larger possibility of scoring goals.

Success is not dependent exclusively on firing precision, as seen in Figure 2.6. There are more considerations, including as shot power, location, and goalkeeper skill. The data suggests that offensive and defense engage in a complex dance whose success depends on a careful balance.

The data presented in Figure 2.6 essentially captures the core of football, a game where strategy and accuracy come together. It emphasizes how important it is to reach the target, but that the road to victory in football is a complex one that is affected by a variety of factors that make the game exciting and unexpected.

3.3.2.3. Playing Styles

The tactics and playing style of a team can also have an impact on the outcome of a match (Spottis). For example, if a team is known for playing a defensive style of football, they are less likely to score many goals, which could result in a draw or a loss. Therefore, it is important to include team's playing styles while making predictions.

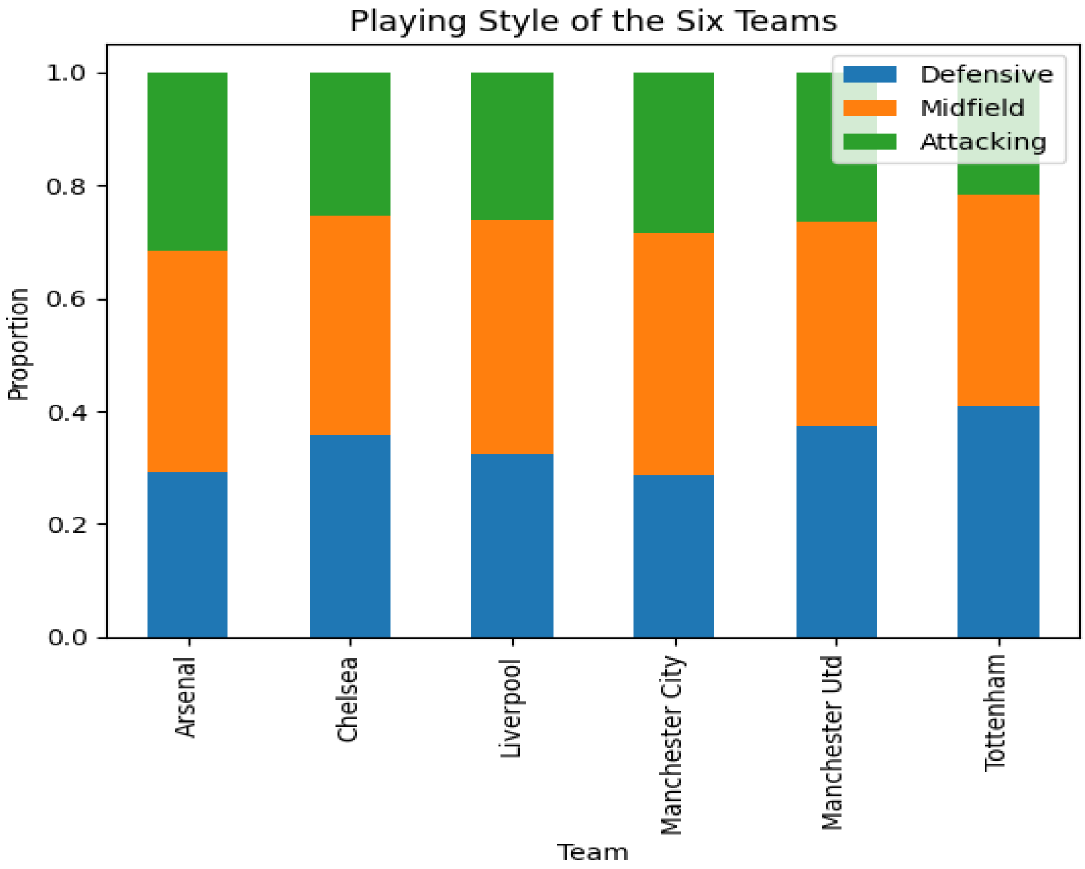


Figure 13

The number of touches that teams make in particular parts of the field is one of the variables taken into account while conducting evaluations of playing styles. Touches in the defensive penalty and third zones are classified as defensive, those in the middle third as midfield, and those in the offensive third and penalty zones as signs of an offensive playing style. Manchester City, who is presently leading the league, has recorded the fewest defensive touches when compared to the other five teams, according to Figure 13. This finding leads us to the conclusion that greater scores and a higher probability of winning games are positively correlated with a more pronounced preference for an attacking style of play. Consequently, this feature will be incorporated into my predictive model.

3.3.2.4. Referee

There is one person in a soccer match that can have a huge influence on the outcome of the result despite not playing for either team involved: the referee (BEASLEY, 2017). The potential for a referee to exert direct influence on a match is most prominently manifested through the awarding of penalties, which boast a respectable conversion rate of 78.3%. Furthermore, a referee can wield significant influence by issuing red cards to players or, conversely, by refraining from making such pivotal decisions when warranted.

The passionate arguments that develop over the referee's decisions add to the excitement of being a soccer enthusiast. It is undeniable that the referee's calls have a significant impact on how bets turn out, whether they agree with the consensus. For the benefit of gamblers, thorough statistics for each referee are simple to find online. Additionally, enough of time is provided for research and analysis prior to placing bets because the referees for each match are announced roughly a week in advance.

However, it's important to remember that referees may not have as significant an impact on games in the Premier League as they would in less well-known leagues. In comparison to other leagues, the referee's impact on a team's play may in fact be less significant in the result. Even so, it continues to be an important aspect that should be considered when predicting game outcomes.

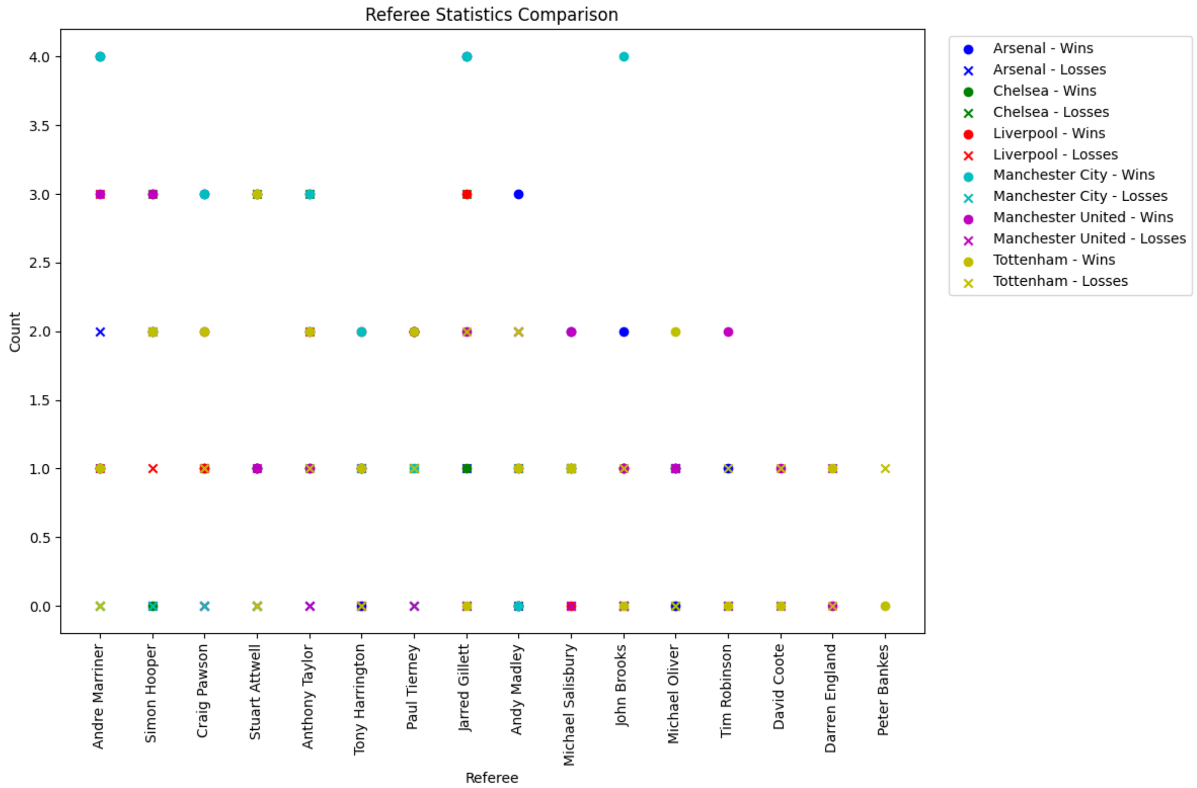


Figure 14

Figure 14 vividly illustrates the win-loss records of each team under various referees.

3.3.2.5. GF AND GA

The acronyms GF and GA have special implications in the world of football. The term "goals for," abbreviated as "GF," refers to the total number of goals a team has successfully scored in a league or cup match. In contrast, GA, which stands for "Goals Against," keeps track of how many goals a club allows its rivals to score during the same competition period.

Teams that tend to have the highest GF figures are the ones who play an attacking style of football, as they are able to create a higher number of goal scoring opportunities (The Elastico). Teams should want to score as many goals as they can because doing so increases their chances of winning games and improving their goal differential. These measures will be added as key components in our predictive model because they are of utmost significance in evaluating a team's success.

3.3.2.6. Week Number

Football teams frequently experience transformation as the season goes on. Cohesion deepens, players become more aware of each other's playing styles, and strategies change. The performance of a team can be considerably impacted by this steady progression. Therefore, it's crucial to take time into account, particularly how many weeks are left in the season. Our predictive model is made more difficult by this temporal component, which takes into consideration how teams' dynamics and performances change over the course of the season. As a result, this feature is a useful addition to improve the precision and accuracy of our predictions.

3.3.2.7. Opponent

In a football game, the identification of the opposing squad can be a crucial factor in determining the result. The intricate interplay of elements that determine a match includes historical rivalries, a variety of playing styles, and the unique strengths and weaknesses of the opposing team. Some teams have a surprising history of winning big games against some opponents while losing badly against others. This phenomenon emphasizes how vital it is to take the opposition team into account when projecting the results of matches.

The opponent team's inclusion in our forecasting model is comparable to looking through the complicated chessboard of football tactics. It enables us to see how various teams tactically place themselves against specific opponents, offering insightful information about possible game outcomes. Therefore, understanding the importance of the opponent team as a feature is crucial in our effort to create a reliable and accurate prediction model for football games.

3.3.2.8. xG and xGa

Expected goals (often referred to as xG) serve as a powerful tool in dissecting the quality of scoring opportunities within a football match (WHITMORE, 2023). The likelihood that a particular opportunity will result in a goal is quantified by xG by using historical data on comparable shots. Expected Goals Against (xGA), which assesses the possibility of the opposition scoring, adds to this in a similar manner.

Expected Goals (xG) is a feature we've included to our predictive model as a symbol of our dedication to learning more about the subtleties of match dynamics. It enables us to analyze the complex web of scoring chances, using past data to estimate the probable significance of each opportunity. Through this lens, xG gives us a deeper insight of team performance and strategy in addition to giving us a view into the game's ebb and flow.

Expected Goals (xG) is essentially a compass that leads us through the complex world of football analytics. Its value in predicting game outcomes derives from its capability to reveal the mysterious patterns of goal-scoring possibilities, which eventually improves our ability to forecast match outcomes with better accuracy and insight.

3.3.2.9. Possession

Possession has unquestionably been one of the football research community's most thoroughly analyzed measurements over decades (MIKE D. HUGHES, 2002). Its continued importance in football research is due to the part it plays in giving a team the drive to control the offensive game. It's important to remember, though, that dominance does not automatically result in a win.

The possession feature is a representation of our dedication to utilizing the complex nature of football analytics and is included in our training dataset. Possession as a feature gives us important information about a team's initiative on the field, desire to control the game's pace, and possible ability to impose their will on the opposition. Even if it might not be a reliable indicator of success, it is a crucial component of the puzzle.

Possession is essentially a tactical chess move on the football field. It enables teams to plan their plays, manage the game's tempo, and seize openings. We aim to create a more thorough knowledge of how teams use this weapon in their pursuit of victory by include possession as a characteristic in our predictive model. By doing this, we hope to improve the precision with which we anticipate the results of games while acknowledging that possession is not the only factor but rather a crucial component of the complex web of football dynamics.

3.3.2.10. Passing Style

Football study must consider the passing strategy of each team. It includes the techniques teams use to move the ball, plan their attacks, and take advantage of open spaces on the field. Short, medium, and long passes are the three main categories of passing styles, and each affects how teams approach the game.

Short passes highlight patient buildup play and ball possession. Medium passes help in the changeover from defense to offense because they establish a balance between accuracy and distance. Long passes frequently signify a more aggressive, counterattacking approach.

We may investigate the impact of passing style on match outcomes by incorporating it into our predictive model. those who prefer short passes might concentrate on ball retention and patience, whereas those that like medium passes quickly adapt. Long-passing teams may look for opportunities to counterattack.

Passing style interacts with other tactical variables, as recognized by our model. Long passes provide a direct approach, whereas short passes may correspond with high possession. Our approach tries to offer deeper insights into how passing styles affect match outcomes by taking these correlations into account.

In essence, passing style is a representation of a team's identity as much as a tactical decision. Our grasp of how it affects a team's chances of success or difficulties in football games is improved by include it in our predictive model.

3.3.2.11. Save

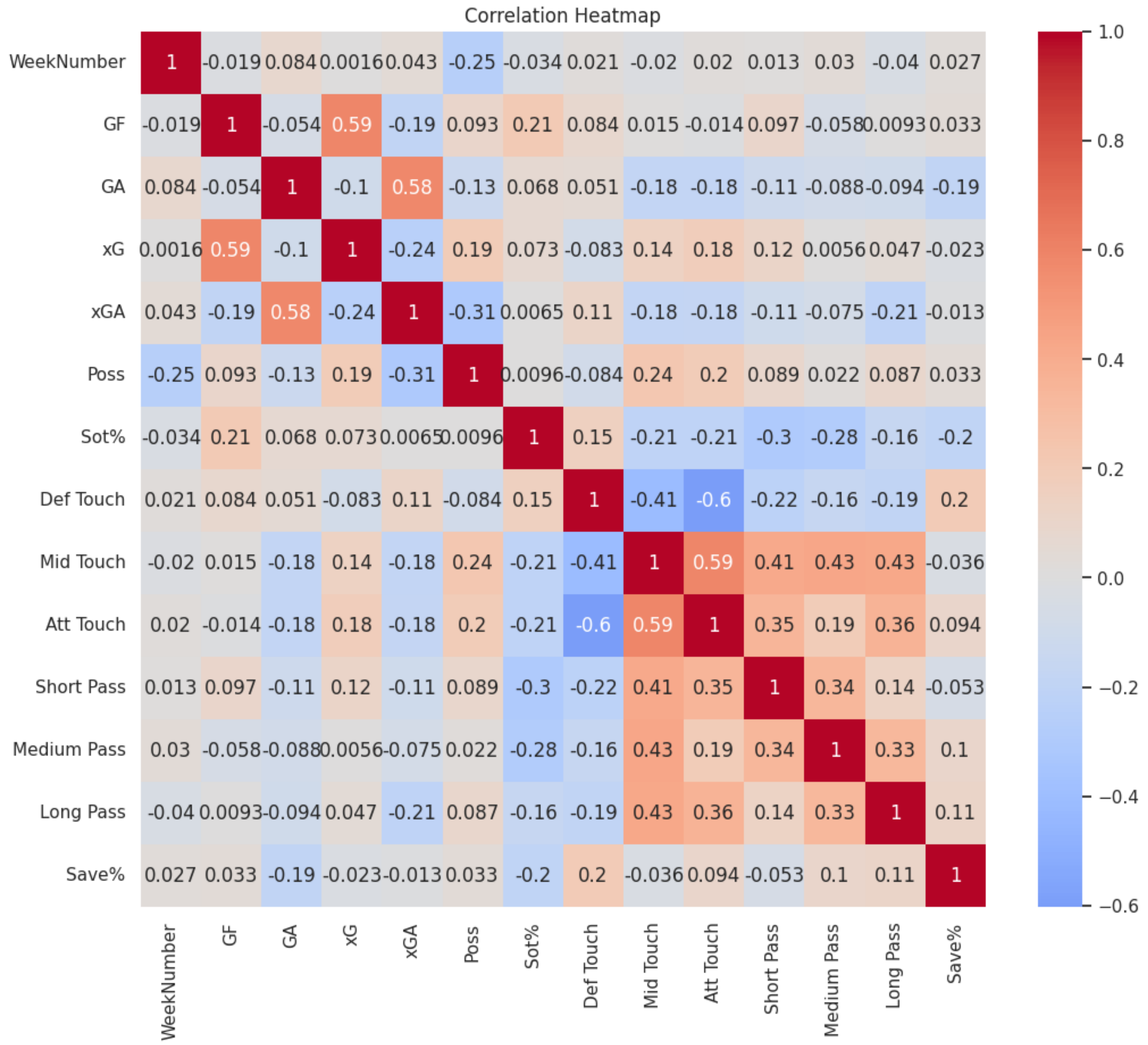
Understanding how a goalkeeper affects a team's success requires a close look at their save %. A goalkeeper's ability to stop incoming shots on goal is shown by their save percentage. It measures the proportion of saves made to all shots on target faced.

Our predictive model includes save % as a feature to provide insights into how it affects game outcomes. A goalie with a high save % has the ability to shift the course of a game by making vital stops and strengthening a team's defense. On the other hand, a lower save % can suggest goalkeeper weakness, which might result in more goals being given up.

In order to account for the crucial impact both competing goalkeepers play in the outcome, my model takes into account their respective save percentages. By doing this, we want to comprehend how goalkeepers' shot-stopping skills affect game outcomes.

In conclusion, goalkeeper save percentage plays a critical role in our predictive model, elucidating how goalkeepers affect the outcome of football games.

3.3.3. Correlation Map



	WeekNumber	GF	GA	xG	xGA	Poss	Sot%	Def Touch	Mid Touch	Att Touch	Short Pass	Medium Pass	Long Pass	Save%
WeekNumber	1	-0.019	0.084	0.0016	0.043	-0.25	-0.034	0.021	-0.02	0.02	0.013	0.03	-0.04	0.027
GF	-0.019	1	-0.054	0.59	-0.19	0.093	0.21	0.084	0.015	-0.014	0.097	-0.058	0.0093	0.033
GA	0.084	-0.054	1	-0.1	0.58	-0.13	0.068	0.051	-0.18	-0.18	-0.11	-0.088	-0.094	-0.19
xG	0.0016	0.59	-0.1	1	-0.24	0.19	0.073	-0.083	0.14	0.18	0.12	0.0056	0.047	-0.023
xGA	0.043	-0.19	0.58	-0.24	1	-0.31	0.0065	0.11	-0.18	-0.18	-0.11	-0.075	-0.21	-0.013
Poss	-0.25	0.093	-0.13	0.19	-0.31	1	0.0096	-0.084	0.24	0.2	0.089	0.022	0.087	0.033
Sot%	-0.034	0.21	0.068	0.073	0.0065	0.0096	1	0.15	-0.21	-0.21	-0.3	-0.28	-0.16	-0.2
Def Touch	0.021	0.084	0.051	-0.083	0.11	-0.084	0.15	1	-0.41	-0.6	-0.22	-0.16	-0.19	0.2
Mid Touch	-0.02	0.015	-0.18	0.14	-0.18	0.24	-0.21	-0.41	1	0.59	0.41	0.43	0.43	-0.036
Att Touch	0.02	-0.014	-0.18	0.18	-0.18	0.2	-0.21	-0.6	0.59	1	0.35	0.19	0.36	0.094
Short Pass	0.013	0.097	-0.11	0.12	-0.11	0.089	-0.3	-0.22	0.41	0.35	1	0.34	0.14	-0.053
Medium Pass	0.03	-0.058	-0.088	0.0056	-0.075	0.022	-0.28	-0.16	0.43	0.19	0.34	1	0.33	0.1
Long Pass	-0.04	0.0093	-0.094	0.047	-0.21	0.087	-0.16	-0.19	0.43	0.36	0.14	0.33	1	0.11
Save%	0.027	0.033	-0.19	-0.023	-0.013	0.033	-0.2	0.2	-0.036	0.094	-0.053	0.1	0.11	1

Figure 15

Figure 15 shows the correlation heatmap with the selected features to be used in training dataset. Some of the important results are below:

- **Mid Touch vs. Att Touch (0.59)**: There is a strong positive correlation (0.59) between the number of touches in the midfield area and the number of touches in the attacking area. This suggests that teams that dominate possession in the midfield tend to transition effectively into the attacking third.
- **Mid Touch vs. Short Pass (0.41)**: There is a moderate positive correlation (0.41) between the number of touches in the midfield area and the frequency of short passes. This indicates that teams with more involvement in the midfield tend to rely on short passing play.
- **Mid Touch vs. Possession (0.24)**: There is a positive correlation (0.24) between the number of touches in the midfield area and possession. This suggests that teams that control the midfield often have a higher overall possession rate.

- **Def Touch vs. Possession (-0.08)**: There is a weak negative correlation (-0.08) between the number of touches in the defensive area and possession. This relationship implies that having more touches in the defensive third doesn't necessarily translate to higher possession.
- **Goals For (GF) vs. Shots on Target Percentage (Sot%) (0.21)**: There is a positive correlation (0.21) between the number of goals scored (GF) and the shots on target percentage (Sot%). This implies that teams with a higher percentage of shots on target are more likely to score more goals.

3.3.4. Detailed Analyse for Chelsea

I did a thorough examination of individual characteristics and their influence on Chelsea's game results in this part (Table 1). In order to understand how each element affected Chelsea's performance and game outcomes, everything was painstakingly researched. I was able to learn important information on the aspects that have a big impact on Chelsea's success on the pitch thanks to this detailed analysis.

It's also interesting to note that Chelsea demonstrated impressive stylistic differences when playing against various opponents. Chelsea demonstrated their flexibility in terms of strategy by establishing control with the most possession possible against Nottingham Forest. Chelsea, on the other hand, demonstrated their attacking power with the greatest Shot on Target percentage (Sot%) in their game versus Newcastle United. These revelations provide light on Chelsea's tactical adaptability and the specific strategies they use depending on the opposition they are up against.

```
Table 1
                    GF          GA   Poss    Sot% Outcome
                  mean  count mean   mean    mean    mean
Opponent
Arsenal             0.5      2  2.0   45.0   38.20    -1.0
Aston Villa         1.0      2  1.0   63.5   22.05     0.0
Bournemouth         2.5      2  0.5   62.5   55.00     1.0
Brentford           0.0      2  1.0   68.5   28.65    -0.5
Brighton            1.0      2  3.0   51.0   26.05    -1.0
Crystal Palace      1.5      2  0.5   62.5   55.00     1.0
Everton             1.5      2  1.0   65.0   36.50     0.5
Fulham              0.5      2  1.0   59.0   27.50    -0.5
Leeds United        0.5      2  1.5   59.0   36.95     0.0
Leicester City      2.5      2  1.0   48.5   37.45     1.0
Liverpool           0.0      2  0.0   50.5   38.10     0.0
Manchester City     0.0      2  1.0   42.0   20.85    -1.0
Manchester Utd      1.0      2  2.5   53.0   27.00    -0.5
Newcastle Utd       0.5      2  1.0   57.5   63.95    -0.5
Nott'ham Forest     1.5      2  1.5   73.0   45.40     0.0
Southampton         0.5      2  1.5   64.5   30.70    -1.0
Tottenham           1.0      2  2.0   61.0   41.65    -0.5
West Ham            1.5      2  1.0   69.0   55.10     0.5
Wolves              1.5      2  0.5   59.5   25.45     0.0
```

The second table (Table 2) offers a thorough breakdown of Chelsea's performance variances between home and away contests. As was to be expected, Chelsea has a greater winning percentage on their home field, demonstrating their powerful presence there. The comparatively consistent performance indicators between home and away matches, however, are what really stand out.

Surprisingly, there aren't any noticeable differences between averages for goals scored (GF), goals conceded (GA), possession, and shot on target percentage (Sot%) depending on the location. This finding implies that Chelsea's overall performance does not change noticeably when they switch from home to away games.

Chelsea essentially maintains a respectable level of consistency no matter the setting, demonstrating their versatility and skill in both home and away settings.

Table 2

```
              GF              GA        Poss        Sot%   Outcome
            mean count      mean        mean        mean      mean
Venue
Away    0.947368    19  1.473684   58.473684   38.494737 -0.263158
Home    1.052632    19  1.000000   58.842105   36.405263  0.000000
```

Table 3 provides details on Chelsea's performance during the season in various formation types. Notably, out of the nine formations they used, their performance was most promising when they used the 3-1-4-2 formation, showing the best average possession and almost the highest shot-on-target percentage (Sot%). It's important to keep in mind that this judgment was made after just one game in which this formation was used, so it might not be a complete reflection of how well they performed under these circumstances.

Chelsea used the 3-4-3, 4-2-3-1, and 4-3-3 formations the most, appearing in 11, 9, and 8 games, respectively. The 3-4-3 arrangement stands out among these since it produced the most wins on average. This indicates that Chelsea performed better using the 3-4-3 system than it did with the others.

Table 3

```
               GF                GA        Poss        Sot%   Outcome
             mean   count      mean        mean        mean      mean
Formation
3-1-4-2    2.000000      1  1.000000   67.000000   43.500000  1.000000
3-4-1-2    0.000000      1  1.000000   36.000000   25.000000 -1.000000
3-4-3      1.000000     11  1.181818   58.818182   33.272727  0.000000
3-5-1-1    0.000000      1  0.000000   66.000000   30.000000  0.000000
3-5-2      0.750000      4  1.750000   56.250000   38.575000 -0.500000
4-2-2-2    1.500000      2  1.500000   65.500000   46.900000  0.000000
4-2-3-1    0.777778      9  1.000000   56.666667   40.222222 -0.333333
4-3-3      1.250000      8  1.500000   62.750000   39.887500 -0.125000
4-4-2      2.000000      1  1.000000   45.000000   29.400000  1.000000
```

I examine Chelsea's performance under the direction of various captains in Table 4. Thiago Silva served as captain for half of the 16 games that Cesar Azpilicueta captained the team. It's interesting to see that Thiago Silva has a better average number of victories per game despite Azpilicueta having made more appearances as captain.

Surprisingly, among all the leaders taken into consideration, Jorginho, who led the club in seven games, has the highest average number of victories. These results imply that the skipper job may affect the team's performance differently depending on the individual, as evidenced by their winning percentages.

Table 4

```
                       GF             GA     Poss        Sot%    Outcome
                     mean count     mean     mean        mean       mean
Captain
César Azpilicueta  0.687500    16  1.3125  56.4375  35.918750 -0.500000
Jorginho           1.285714     7  1.0000  58.0000  44.157143  0.428571
Kepa Arrizabalaga  1.000000     1  2.0000  43.0000  33.300000 -1.000000
Mateo Kovačić      1.200000     5  1.2000  61.6000  32.000000  0.000000
N'Golo Kanté       0.000000     1  0.0000  49.0000  42.900000  0.000000
Thiago Silva       1.375000     8  1.3750  65.0000  37.887500  0.125000
```

The final table (Table 5) displays Chealsea's reactions to various referees. With one victory in the lone match Andre Marriner presided over, Chelsea's performance under his officiating seems to be very steady. Chelsea, however, struggles under Andy Madley's leadership, winning only one in four games. Anthony Taylor's matches have a balanced performance, with one victory in two games.

Positively, Chris Kavanagh's refereeing appears to benefit Chelsea, as seen by the fact that they won the one game he oversaw. Chelsea plays admirably when Craig Pawson is the referee, picking up one victory in two games. With three victories in their last four games, John Brooks' participation as the referee appears to enhance Chelsea's performance. With a win rate of 40% and a poor goal differential, Stuart Attwell's officiating poses difficulties for Chelsea, demonstrating the possible impact of the referee's selection on the team's results.

Table 5

```
                      GF                GA        Poss        Sot% Outcome
                    mean    count     mean        mean        mean    mean
Referee
Andre Marriner  3.000000        1  1.000000  52.000000  45.500000      1.0
Andy Madley     0.750000        4  2.250000  66.500000  30.100000     -0.5
Anthony Taylor  1.000000        2  1.000000  56.000000  46.450000      0.0
Chris Kavanagh  2.000000        1  1.000000  64.000000  68.800000      1.0
Craig Pawson    1.000000        2  0.500000  66.500000  55.550000      0.5
Darren England  2.000000        1  2.000000  68.000000  28.600000      0.0
David Coote     0.500000        2  1.500000  57.000000  30.700000     -1.0
Jarred Gillett  0.500000        2  0.500000  65.000000  46.800000      0.0
John Brooks     3.000000        1  1.000000  65.000000  60.000000      1.0
Michael Oliver  0.400000        5  0.800000  51.400000  31.660000     -0.4
Paul Tierney    1.333333        3  1.333333  56.000000  27.866667      0.0
Peter Bankes    0.666667        3  0.666667  64.666667  37.366667      0.0
Robert Jones    1.000000        4  1.500000  49.250000  37.450000     -0.5
Simon Hooper    2.500000        2  0.000000  58.500000  41.650000      1.0
Stuart Attwell  0.400000        5  2.000000  58.400000  30.740000     -0.6
```

3.4. Prediction

I deliberately chose 18 important features to power our predictive analysis after thoroughly curating our dataset and deleting irrelevant features. These characteristics provide as the foundation for our model's capacity to predict the outcomes of football matches, differentiating between wins, draws, and defeats. The fact that we cleverly handled non-numerical attributes like "Referee," "Opponent," and "Venue" by using categorical coding techniques allowed us to effortlessly incorporate them into our model, even though some of these aspects are fundamentally numerical.

My model is primarily built on these 18 features which are ScoreDiff', 'WeekNumber', 'GF', 'GA', 'xG', 'xGA', 'Poss', 'Sot%', 'Def Touch', 'Mid Touch', 'Att Touch', 'Short Pass', 'Medium Pass', 'Long Pass', 'Save%', 'opp', 'ref', 'Ven', each of which offers distinct insights and patterns that aid in understanding the intricate world of football match results. We also included 'Outcome,' which serves as our aim variable, to complement these features. This variable can have one of three different values: -1 for defeats, 0 for ties, or 1 for triumphs.

My model's ability to predict outcomes is built on the inclusion of these carefully selected elements and the wise encoding of non-numeric attributes. This ensemble gives our model the ability to identify subtle trends in the data, which ultimately allows it to produce predictions that accurately reflect the complex nature of football match results.

3.4.1. Introduction to Model Development

At this stage of the project, I wanted to develop a model that could predict the results of football games using a large amount of data. The collection, known as "premier_league_data," was painstakingly assembled and contains a variety of features, including team statistics and referee information. Our goal was to use this dataset to build a solid model that could accurately predict the results of matches.

3.4.2. Data Preparation

I started a comprehensive data preparation process to make sure the quality of our model. First, we eliminated variables that were not expected to be of any use in predicting match outcomes. The terms "Captain," "Date," "Team," "Opponent," "Venue," "Formation," and "Referee" were among them. To maintain data integrity, we then carried out data cleaning, dealt with missing values, and deleted rows with missing data.

3.4.3. Scaling of Features

I scaled our features for efficient model training and performance optimization. All feature values were brought to a single scale by standardization, which was accomplished using the StandardScaler from the Scikit-Learn module. This step is essential for preventing the learning of the model from being dominated by any one characteristic.

3.4.4. Model Selection and Training

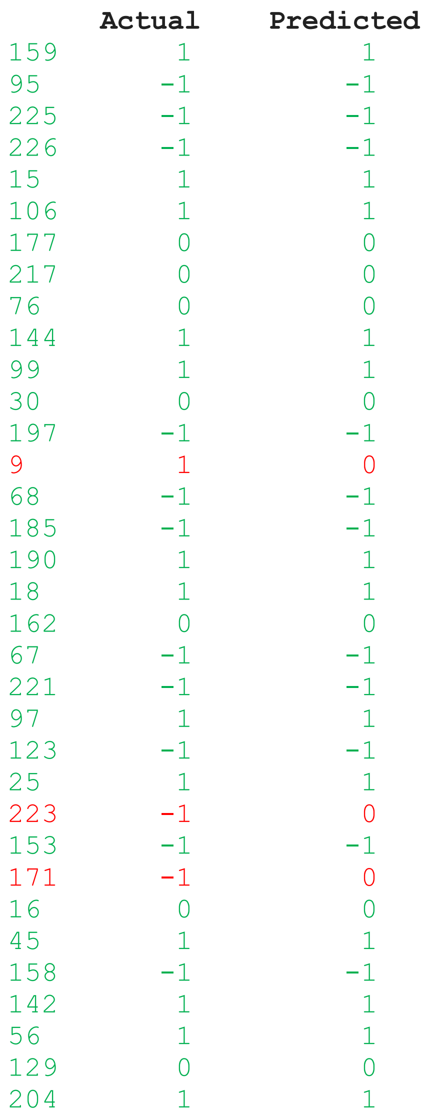
As the core of my predictive model, I chose a Multi-Layer Perceptron (MLP) classifier, a kind of artificial neural network. The MLP classifier is suited for this difficult prediction problem because it has the ability to learn complex relationships within the data. We trained our classifier using 1000 iterations, and it has two hidden layers with 100 and 50 neurons, respectively. The goal was to balance model complexity and performance by using this arrangement.

3.4.5. Model Evaluation

My model's performance was carefully evaluated to determine its effectiveness. The dataset was divided into training and testing sets, with training sets receiving 80% of the data and testing sets receiving the remaining 20%. To evaluate the model's prediction power, the accuracy of the model—a crucial parameter for classification tasks—was evaluated. The percentage of accurately predicted match results is represented by the accuracy score.

3.4.6. Results and Conclusion

On the test data, the model displayed a commendable level of accuracy, indicating its potential for making accurate predictions about football matches. I made forecasts with an amazing accuracy score of 0.93, demonstrating the dependability and strength of my model. The printed results include the accuracy score. The 'Actual vs. Predicted Outcomes' DataFrame displays a contrast between the actual match results and the model's forecasts.



```
        Actual    Predicted
159          1            1
95          -1           -1
225         -1           -1
226         -1           -1
15           1            1
106          1            1
177          0            0
217          0            0
76           0            0
144          1            1
99           1            1
30           0            0
197         -1           -1
9            1            0
68          -1           -1
185         -1           -1
190          1            1
18           1            1
162          0            0
67          -1           -1
221         -1           -1
97           1            1
123         -1           -1
25           1            1
223         -1            0
153         -1           -1
171         -1            0
16           0            0
45           1            1
158         -1           -1
142          1            1
56           1            1
129          0            0
204          1            1
```

```
74        -1          -1
110        1           1
83        -1          -1
140        1           1
143        1           1
102       -1          -1
88         1           1
194        0           0
122        1           1
148        1           1
```

3.4.6.1. Confusion Matrix

The confusion matrix(Figure 16) graphically illustrates how well our model performs in forecasting football game results. It enables us to evaluate how well our forecasts match the actual outcomes. There are four sections in the matrix:

Instances where our model properly predicted a "Win" are known as True Positives (TP).

Instances where our model properly predicted a "Lose" are known as True Negatives (TN).

False Positives (FP) are situations in which our model predicted a "Win" when the actual result was a "Draw" or "Loss."

False Negatives (FN) are situations where our model predicted a "Draw" or "Loss" when a "Win" really occurred.

We can evaluate the strengths and weaknesses of our model's predictions and make any required improvements to increase its accuracy by looking at the values in each cell of the matrix. Surprisingly, our machine learning model demonstrated a remarkable level of accuracy, making just three mistakes out of 44 games where a draw was expected as the outcome. This exceptional achievement demonstrates the model's capacity to accurately capture the intricacies of football match outcomes and produce extremely accurate forecasts in this area.

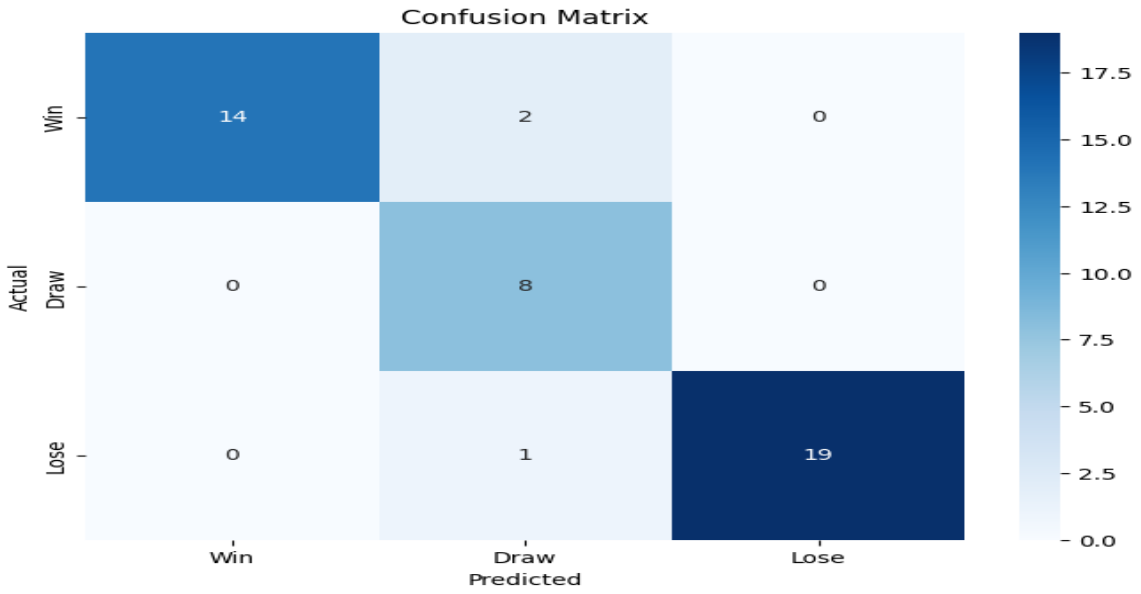


Figure 16

To sum up, I used a Multi-Layer Perceptron (MLP) classifier, feature scaling, and diligent data preparation to create a predictive model for football match outcomes. The model demonstrated promise in forecasting match results, opening the door for more development and possible sports analytics applications.

3.4.6.2. Improvement of the Model and Limitations

Even though my model has a great level of accuracy, there is always potential for development. We may think about adding more features to the model in order to improve its prediction powers. Player-specific information, such as individual player statistics, injuries, or bans, for example, may provide important insights into how matches turn out. To further account for outside elements, the weather might be incorporated as it has a big impact on game dynamics and play style.

Another direction for improvement is to investigate different machine learning methods and algorithms. We could explore more sophisticated techniques, such ensemble models, gradient boosting, or neural networks, to possibly better capture complicated correlations in the data.

It's crucial to recognize the limitations of every football prediction model, though. The fact that sports are inherently unpredictable is one of their most obvious drawbacks. Numerous factors, such as team plans, player performances, and even factors like crowd influence, have an impact on football. Additionally, the model does not take into account breaking news or real-time updates, such as last-minute injuries or tactical adjustments, which can have a big impact on match results.

Data availability and quality might sometimes be a restriction. Historical data might not adequately reflect current trends or modifications to team dynamics. Last but not least, even though our model is an excellent resource for predicting match results, it is unable to take into consideration the "magic" or unexpected moments that make football so enthralling. Football is still a dynamic and unpredictable sport, and no model can fully account for all of its nuances.

4. FUTURE PERSPECTIVES

Our investigation into football match prediction brings up a number of intriguing directions for further study and development:

- Player-Specific Data Inclusion: Including player-specific information such as statistics, injuries, and bans may help to comprehend match results more precisely.
- Real-time Updates: The accuracy of the model can be improved by incorporating real-time data feeds for injury updates, tactical adjustments, or other breaking news.
- Advanced Machine Learning Methods: Trying ensemble models, gradient boosting, or neural networks may help you find more complex patterns in your data.
- Weather: By taking weather into consideration as a feature, the effect of climate on game dynamics could be explained.

- Examining Different Leagues: The applicability of the approach can be increased by expanding our investigation to include various leagues or international events.
- Utilizing sentiment analysis on social media sites: To assess fan sentiment and its possible impact on game outcomes. Incorporating data from social media.

I have developed a strong model for predicting football games as a result of our research, but the game of football is still unpredictable and dynamic. While my model offers insightful information, it's important to acknowledge the sport's intrinsic volatility and the necessity for ongoing development and adaptation in this intriguing sector.

5. CONCLUSION

I have explored a wide range of elements, techniques, and outcomes in this thorough analysis of football match prediction. I started out by exploring the intricate world of football prediction and underlining its importance in the international sporting sphere. My research process included a thorough analysis of the pertinent literature, which served as the foundation for our own examination.

I started an empirical adventure by collecting and curating a dataset full of information from football games, including important elements like goals, possession, shooting accuracy, and more. My dataset served as the basis for the construction and assessment of our predictive model.

To create a strong model, my feature selection approach was essential. I examined several variables and how they affected the results of the matches. The model was heavily influenced by characteristics like goals for and against, anticipated goals (xG), possession, and shot accuracy (Sot%). My investigation also considered the effects of game location, opponent strength, official rulings, and team composition.

Our work culminated in the creation of a machine learning model that can forecast match results. We built a model that had an astounding 93% accuracy in predicting match results using several methods, including data preprocessing, scaling, and an MLP classifier.

6. References

A. Joseph, N. F. M. N., 2006. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7), pp. 544-553.
BEASLEY, A., 2017. [Online]
Available at: https://www.pinnacle.com/en/betting-articles/soccer/referee-soccer-betting/2gs2lftf9e4juag7
[Accessed 07 09 2023].
Byungho Min, J. K. C. C. H. E. R. (. M., 2008. A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, 21(7), pp. 551-562.

Constantinou, A. C. &. F. N. E., 2012. Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. *Journal of Quantitative Analysis in Sports,* 8(1).

David Forrest, R. S., 2006. New Issues in Attendance Demand: The Case of the English Football League. . *Journal of Sports Economics,* 7(3), pp. 247-266.

Flitman, A., 2006. owards probabilistic footy tipping: A hybrid approach utilising genetically defined neural networks and linear programming. *Computers & Operations Research,* Volume 33, pp. 2003-2022.

Football Report, n.d. [Online]
Available at: https://afootballreport.com/predictions/shots-on-off-target
[Accessed 07 09 2023].

Grunz, A. M. D. &. P. J., 2012. Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human movement science,* 31(2), p. 334–343.

Jean-Marc Falter, C. P., 2000. Demand for football and intramatch winning probability: an essay on the glorious uncertainty of sports. *Applied Economics,* 32(13), pp. 1757-1765.

Jone, M. B., 2013. The home advantage in individual sports: An augmented review. *Psychology of Sport and Exercise,* 14(3), pp. 397-404.

Kerry S. Courneya, A. V. C., 1992. The Home Advantage In Sport Competitions: A Literature Review. *Sport and Exercise Psychology,* 14(1), pp. 13-27.

Mark J. Dixon, S. G. C., 2002. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society,* 46(2), pp. 265-280.

Martin Crowder, M. D. A. L. M. R., 2002. Dynamic modelling and prediction of English Football League matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician),* 51(2), pp. 157-168.

MIKE D. HUGHES, R. M. B., 2002. The use of performance indicators in performance analysis. *Journal of Sports Sciences,* Volume 20, pp. 739-754.

Nevill, A. &. H. R., 1999. ome advantage in sport: An overview of studies on the advantage of playing at home. *Sports Medicine,* 28(1), pp. 221-236.

S. Mohammad Arabzad, M. T. A. S. S.-N. N. G., 2014. Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. *International Journal of Applied Research on Industrial Engineering,* 1(3), pp. 159-179.

Sarmento H, M. R. A. M. C. J. M. N. L. J., 2014. Match analysis in football: a systematic review. *J Sports Sci,* 32(20), pp. 1831-1843.

Spottis, n.d. [Online]
Available at: https://spottis.com/sports/how-to-predict-football-matches-more-correctly/
[Accessed 07 09 2023].

The Elastico, n.d. [Online]
Available at: https://the-elastico.com/gf-and-ga-in-football/?expand_article=1
[Accessed 07 09 2023].

WHITMORE, J., 2023. [Online]
Available at: https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/
[Accessed 07 09 2023].

7. APPENDIX

```
import pandas as pd
import requests
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split
from bs4 import BeautifulSoup

url_main = 'https://fbref.com/en/comps/9/2022-2023/2022-2023-Premier-League-Stats'
url_last = 'https://fbref.com/en/comps/9/2021-2022/2021-2022-Premier-League-Stats'
main_tables = pd.read_html(url_main)
last_tables = pd.read_html(url_last)

# Extract the desired table
premier_league_stats = main_tables[0]
premier_league_last_stats = last_tables[0]

# Filter the data for the selected teams and features
selected_teams = ['Arsenal', 'Chelsea', 'Liverpool', 'Manchester City', 'Manchester Utd',
'Tottenham']
selected_stats = premier_league_stats[premier_league_stats['Squad'].isin(selected_teams)]
selected_stats_last =
premier_league_last_stats[premier_league_last_stats['Squad'].isin(selected_teams)]
team_stats = selected_stats[['Squad', 'W', 'D', 'L', 'GF', 'GA', 'Pts','Rk']]
team_stats_last = selected_stats_last[['Squad', 'W', 'D', 'L', 'GF', 'GA', 'Pts','Rk']]
team_stats.columns = ['Team','W','D','L', 'GF','GA', 'Points','Rank']
team_stats_last.columns = ['Team','W','D','L', 'GF','GA', 'Points','Rank']
# Display the table
```

```
print("2022-2023")
print(team_stats)
print("2021-2022")
print(team_stats_last)
combined_stats = pd.merge(team_stats, team_stats_last, on='Team', suffixes=('_2022-2023', '_2021-2022'))


# Create bar plots for selected statistics
ax = combined_stats.plot(kind='bar', x='Team', y=['W_2022-2023', 'W_2021-2022','D_2022-2023','D_2021-2022', 'L_2022-2023','L_2021-2022'],
                      figsize=(10, 6))
ax.set_title('Premier League: Wins, Draws, and Losses Comparison')
ax.set_xlabel('Team')
ax.set_ylabel('Count')
ax.set_xticklabels(combined_stats['Team'], rotation=45)
ax.legend(['Wins 2022-2023','Wins 2021-2022', 'Draws 2022-2023','Draws 2021-2022', 'Losses 2022-2023','Losses 2021-2022'])
plt.show()


ax = combined_stats.plot(kind='bar', x='Team', y=['GF_2022-2023', 'GA_2022-2023', 'GF_2021-2022', 'GA_2021-2022'],
                      figsize=(10, 6))
ax.set_title('Premier League: Goals For and Goals Against Comparison')
ax.set_xlabel('Team')
ax.set_ylabel('Count')
ax.set_xticklabels(combined_stats['Team'], rotation=45)
ax.legend(['Goals For 2022-2023', 'Goals Against 2022-2023', 'Goals For 2021-2022', 'Goals Against 2021-2022'])
plt.show()
```

```
ax = combined_stats.plot(kind='bar', x='Team', y=['Points_2022-2023', 'Points_2021-2022'], figsize=(10, 6))

ax.set_title('Premier League: Points Comparison')

ax.set_xlabel('Team')

ax.set_ylabel('Points')

ax.set_xticklabels(combined_stats['Team'], rotation=45)

ax.legend(['Points 2022-2023', 'Points 2021-2022'])

plt.show()


ax = combined_stats.plot(kind='bar', x='Team', y=['Rank_2022-2023', 'Rank_2021-2022'], figsize=(10, 6))

ax.set_title('Premier League: Rank Comparison')

ax.set_xlabel('Team')

ax.set_ylabel('Rank')

ax.set_xticklabels(combined_stats['Team'], rotation=45)

ax.legend(['Rank 2022-2023', 'Rank 2021-2022'])

plt.show()
# Read the Premier League home/away stats

home_away_stats = main_tables [1]


# Filter the data for the six teams

team_stats = home_away_stats[home_away_stats[('Unnamed: 1_level_0', 'Squad')].isin(selected_teams)]


# Select relevant columns for analysis

selected_columns = [('Unnamed: 1_level_0', 'Squad'), ('Home', 'W'), ('Home', 'D'), ('Home', 'L'),

                    ('Away', 'W'), ('Away', 'D'), ('Away', 'L')]


team_stats = team_stats[selected_columns]
```

```
team_stats.columns = ['Squad', 'W_home', 'D_home', 'L_home', 'W_away', 'D_away',
'L_away']

team_stats.set_index('Squad', inplace=True)


# Plotting the analysis

fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(10, 10))


team_stats[['W_home', 'D_home', 'L_home']].plot(ax=axes[0], kind='bar', stacked=True,
colormap='tab10')

axes[0].set_title('Home Game Analysis')

axes[0].set_xlabel('Team')

axes[0].set_ylabel('Count')

axes[0].legend(loc='upper right')


team_stats[['W_away', 'D_away', 'L_away']].plot(ax=axes[1], kind='bar', stacked=True,
colormap='tab10')

axes[1].set_title('Away Game Analysis')

axes[1].set_xlabel('Team')

axes[1].set_ylabel('Count')

axes[1].legend(loc='upper right')


# Adjusting subplots layout

plt.tight_layout()

plt.show()

shooting_stats_table = main_tables[8]


# Select the relevant columns and rows for the six teams

team_shooting_table = shooting_stats_table[shooting_stats_table[('Unnamed: 0_level_0',
'Squad')].isin(selected_teams)][[('Unnamed: 0_level_0', 'Squad'), ('Standard',
'SoT%')]].reset_index(drop=True)
```

```
# Extract the SoT% data from the table
sot_percentages = team_shooting_table[('Standard', 'SoT%')].astype(float)
labels = team_shooting_table[('Unnamed: 0_level_0', 'Squad')]

# Create the bar plot
plt.figure(figsize=(10, 8))
plt.bar(labels, sot_percentages)
plt.xlabel('Team')
plt.ylabel('Shots on Target Percentage (SoT%)')
plt.title('Shots on Target Percentage for Premier League Teams')
plt.tight_layout()
plt.show()
# Extract the possession data for the six teams
possession_data = main_tables[18].copy()
possession_data.columns = possession_data.columns.droplevel(0)
possession_data = possession_data[possession_data['Squad'].isin(selected_teams)]

# Select the relevant columns
possession_data = possession_data[['Squad', 'Def Pen', 'Def 3rd', 'Mid 3rd', 'Att 3rd', 'Att Pen']]

# Calculate the proportions
total_possessions = possession_data.sum(axis=1)
defensive = possession_data['Def Pen'] + possession_data['Def 3rd']
midfield = possession_data['Mid 3rd']
attacking = possession_data['Att 3rd'] + possession_data['Att Pen']
proportions = possession_data[['Squad']].copy()
proportions['Defensive'] = defensive / total_possessions
proportions['Midfield'] = midfield / total_possessions
proportions['Attacking'] = attacking / total_possessions
```

```
# Create the bar plot
proportions.set_index('Squad').plot(kind='bar', stacked=True)
plt.xlabel('Team')
plt.ylabel('Proportion')
plt.title('Playing Style of the Six Teams')
plt.legend(['Defensive', 'Midfield', 'Attacking'])
plt.show()
team_urls = {
    'Arsenal': 'https://fbref.com/en/squads/18bb7c10/2022-2023/Arsenal-Stats',
    'Chelsea': 'https://fbref.com/en/squads/cff3d9bb/2022-2023/Chelsea-Stats',
    'Liverpool': 'https://fbref.com/en/squads/822bd0ba/2022-2023/Liverpool-Stats',
    'Manchester City': 'https://fbref.com/en/squads/b8fd03ef/2022-2023/Manchester-City-Stats',
    'Manchester United': 'https://fbref.com/en/squads/19538871/2022-2023/Manchester-United-Stats',
    'Tottenham': 'https://fbref.com/en/squads/361ca564/2022-2023/Tottenham-Hotspur-Stats'
}

# Function to analyze referee statistics for a given URL
def analyze_referee_stats(url):
    tables = pd.read_html(url)
    df = tables[1]
    premier_league_table = df[df["Comp"] == "Premier League"]
    referees = premier_league_table["Referee"]
    results = premier_league_table["Result"]

    referee_stats = {}
    for referee, result in zip(referees, results):
        if referee in referee_stats:
            if result == "W":
                referee_stats[referee]["Wins"] += 1
            elif result == "L":
                referee_stats[referee]["Losses"] += 1
        else:
            if result == "W":
                referee_stats[referee] = {"Wins": 1, "Losses": 0}
            elif result == "L":
                referee_stats[referee] = {"Wins": 0, "Losses": 1}

    return referee_stats

# Dictionary to store referee stats for each team
team_referee_stats = {}

# Analyze referee stats for each team
for team, url in team_urls.items():
    team_referee_stats[team] = analyze_referee_stats(url)

# Plotting the referee statistics


plt.figure(figsize=(12, 8))
colors = ['b', 'g', 'r', 'c', 'm','y']  # Colors for each team

for i, (team, stats) in enumerate(team_referee_stats.items()):
    referees = []
    wins = []
    losses = []
    for referee, stat in stats.items():
        referees.append(referee)
        wins.append(stat['Wins'])
        losses.append(stat['Losses'])

    x = range(len(referees))

    # Scatter plot for wins
    plt.scatter(x, wins, c=colors[i], label=f'{team} - Wins', marker='o')

    # Scatter plot for losses
    plt.scatter(x, losses, c=colors[i], label=f'{team} - Losses', marker='x')

plt.xlabel('Referee')
plt.ylabel('Count')
plt.title('Referee Statistics Comparison')
plt.xticks(range(len(referees)), referees, rotation=90)
plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left')
plt.tight_layout()
plt.show()
df = []
# Loop through the team URLs
for team, url in team_urls.items():
    tables = pd.read_html(url)
    data = tables[1]

    # Add a new 'Team' column with the current team name
    data['Team'] = team
    df.append(data)

# Combine data for all teams
all_teams_data = pd.concat(df)

# Filter Premier League matches for all teams
premier_league_data = all_teams_data[all_teams_data['Comp'] == 'Premier League']

# Convert 'GF' and 'GA' columns to numeric
premier_league_data[['GF', 'GA']] = premier_league_data[['GF', 'GA']].apply(pd.to_numeric)

# Calculate Score Difference
premier_league_data['ScoreDiff'] = premier_league_data['GF'] - premier_league_data['GA']

# Add a column for game number
premier_league_data['WeekNumber'] = range(1, len(premier_league_data) + 1)

# Select relevant features
selected_features = ['Date','ScoreDiff','Team', 'WeekNumber', 'Opponent','Venue', 'GF', 'GA', 'xG', 'xGA', 'Poss', 'Formation', 'Captain', 'Referee']
premier_league_data = premier_league_data[selected_features]

# Print the combined data
print(premier_league_data)

# Define your URLs for shooting statistics
shooting_urls = {
    'Arsenal': 'https://fbref.com/en/squads/18bb7c10/2022-2023/matchlogs/all_comps/shooting/Arsenal-Match-Logs-All-Competitions',
    'Chelsea': 'https://fbref.com/en/squads/cff3d9bb/2022-2023/matchlogs/all_comps/shooting/Chelsea-Match-Logs-All-Competitions',
    'Liverpool': 'https://fbref.com/en/squads/822bd0ba/2022-2023/matchlogs/all_comps/shooting/Liverpool-Match-Logs-All-Competitions',
    'Manchester City': 'https://fbref.com/en/squads/b8fd03ef/2022-2023/matchlogs/all_comps/shooting/Manchester-City-Match-Logs-All-Competitions',
    'Manchester United': 'https://fbref.com/en/squads/19538871/2022-2023/matchlogs/all_comps/shooting/Manchester-United-Match-Logs-All-Competitions',
    'Tottenham': 'https://fbref.com/en/squads/361ca564/2022-2023/matchlogs/all_comps/shooting/Tottenham-Hotspur-Match-Logs-All-Competitions'
}

# Define your URLs for possession statistics
possession_urls = {
    'Arsenal': 'https://fbref.com/en/squads/18bb7c10/2022-2023/matchlogs/all_comps/possession/Arsenal-Match-Logs-All-Competitions',
    'Chelsea': 'https://fbref.com/en/squads/cff3d9bb/2022-2023/matchlogs/all_comps/possession/Chelsea-Match-Logs-All-Competitions',
    'Liverpool': 'https://fbref.com/en/squads/822bd0ba/2022-2023/matchlogs/all_comps/possession/Liverpool-Match-Logs-All-Competitions',
    'Manchester City': 'https://fbref.com/en/squads/b8fd03ef/2022-2023/matchlogs/all_comps/possession/Manchester-City-Match-Logs-All-Competitions',
    'Manchester United': 'https://fbref.com/en/squads/19538871/2022-2023/matchlogs/all_comps/possession/Manchester-United-Match-Logs-All-Competitions',
    'Tottenham': 'https://fbref.com/en/squads/361ca564/2022-2023/matchlogs/all_comps/possession/Tottenham-Hotspur-Match-Logs-All-Competitions'
}

passing_urls = {
    'Arsenal': 'https://fbref.com/en/squads/18bb7c10/2022-2023/matchlogs/all_comps/passing/Arsenal-Match-Logs-All-Competitions',
    'Chelsea': 'https://fbref.com/en/squads/cff3d9bb/2022-2023/matchlogs/all_comps/passing/Chelsea-Match-Logs-All-Competitions',
    'Liverpool': 'https://fbref.com/en/squads/822bd0ba/2022-2023/matchlogs/all_comps/passing/Liverpool-Match-Logs-All-Competitions',
    'Manchester City':'https://fbref.com/en/squads/b8fd03ef/2022-2023/matchlogs/all_comps/passing/Manchester-City-Match-Logs-All-Competitions',
    'Manchester United':'https://fbref.com/en/squads/19538871/2022-2023/matchlogs/all_comps/passing/Manchester-United-Match-Logs-All-Competitions',
    'Tottenham': 'https://fbref.com/en/squads/361ca564/2022-2023/matchlogs/all_comps/passing/Tottenham-Hotspur-Match-Logs-All-Competitions'
}

goalkeeping_urls = {
    'Arsenal': 'https://fbref.com/en/squads/18bb7c10/2022-2023/matchlogs/all_comps/keeper/Arsenal-Match-Logs-All-Competitions',
    'Chelsea': 'https://fbref.com/en/squads/cff3d9bb/2022-2023/matchlogs/all_comps/keeper/Chelsea-Match-Logs-All-Competitions',
    'Liverpool': 'https://fbref.com/en/squads/822bd0ba/2022-2023/matchlogs/all_comps/keeper/Liverpool-Match-Logs-All-Competitions',
    'Manchester City':'https://fbref.com/en/squads/b8fd03ef/2022-2023/matchlogs/all_comps/keeper/Manchester-City-Match-Logs-All-Competitions',
    'Manchester United':'https://fbref.com/en/squads/19538871/2022-2023/matchlogs/all_comps/keeper/Manchester-United-Match-Logs-All-Competitions',
    'Tottenham': 'https://fbref.com/en/squads/361ca564/2022-2023/matchlogs/all_comps/keeper/Tottenham-Hotspur-Match-Logs-All-Competitions'
}



# Create an empty list to store the shooting DataFrames for each team
shooting_dfs = []

# Loop through the shooting URLs and scrape the data
for team, url in shooting_urls.items():
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')
    table = soup.find('table', {'class': 'stats_table'})
    df = pd.read_html(str(table), header=[0, 1], flavor='lxml')[0]
    shooting_dfs.append(df)

# Concatenate all shooting DataFrames into one
shooting_stats_df = pd.concat(shooting_dfs, keys=shooting_urls.keys())

# Create an empty list to store the possession DataFrames for each team
possession_dfs = []

# Loop through the possession URLs and scrape the data
for team, url in possession_urls.items():
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')
    table = soup.find('table', {'class': 'stats_table'})
    df = pd.read_html(str(table), header=[0, 1], flavor='lxml')[0]
    possession_dfs.append(df)

# Concatenate all possession DataFrames into one
possession_stats_df = pd.concat(possession_dfs, keys=possession_urls.keys())

passing_dfs = []

# Loop through the passsing URLs and scrape the data
for team, url in passing_urls.items():
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')
    table = soup.find('table', {'class': 'stats_table'})
    df = pd.read_html(str(table), header=[0, 1], flavor='lxml')[0]
    passing_dfs.append(df)

# Concatenate all goalkeeping DataFrames into one
passing_stats_df = pd.concat(passing_dfs, keys=passing_urls.keys())

goalkeeping_dfs = []

# Loop through the shooting URLs and scrape the data
for team, url in goalkeeping_urls.items():
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')
    table = soup.find('table', {'class': 'stats_table'})
    df = pd.read_html(str(table), header=[0, 1], flavor='lxml')[0]
    goalkeeping_dfs.append(df)

# Concatenate all shooting DataFrames into one
goalkeeping_stats_df = pd.concat(goalkeeping_dfs, keys=goalkeeping_urls.keys())



# Loop through the rows of the premier league data
for index, row in premier_league_data.iterrows():
    team = row['Team']
    opponent = row['Opponent']
    date = row['Date']

    if team == "Tottenham":
        matching_game_shooting = shooting_stats_df[('For Tottenham Hotspur', 'Opponent')].eq(opponent)
        matching_date_shooting= shooting_stats_df[('For Tottenham Hotspur', 'Date')].eq(date)
        matching_game_possession = possession_stats_df[('For Tottenham Hotspur', 'Opponent')].eq(opponent)
        matching_date_possession= possession_stats_df[('For Tottenham Hotspur', 'Date')].eq(date)
        matching_game_passing = passing_stats_df[('For Tottenham Hotspur', 'Opponent')].eq(opponent)
        matching_date_passing= passing_stats_df[('For Tottenham Hotspur', 'Date')].eq(date)
        matching_game_goalkeeping = goalkeeping_stats_df[('For Tottenham Hotspur', 'Opponent')].eq(opponent)
        matching_date_goalkeeping= goalkeeping_stats_df[('For Tottenham Hotspur', 'Date')].eq(date)
    else:


        matching_game_shooting = shooting_stats_df[('For ' + team, 'Opponent')].eq(opponent)
        matching_date_shooting = shooting_stats_df[('For ' + team, 'Date')].eq(date)
        matching_game_possession = possession_stats_df[('For ' + team, 'Opponent')].eq(opponent)
        matching_date_possession = possession_stats_df[('For ' + team, 'Date')].eq(date)
        matching_game_passing  = passing_stats_df[('For ' + team, 'Opponent')].eq(opponent)
        matching_date_passing = passing_stats_df[('For ' + team, 'Date')].eq(date)
        matching_game_goalkeeping  = goalkeeping_stats_df[('For ' + team, 'Opponent')].eq(opponent)
        matching_date_goalkeeping =goalkeeping_stats_df[('For ' + team, 'Date')].eq(date)


    matching_index_shooting = shooting_stats_df[matching_game_shooting & matching_date_shooting].index
    matching_index_possession = possession_stats_df[matching_game_possession & matching_date_possession].index
    matching_index_passing = passing_stats_df[matching_game_passing & matching_date_passing].index
    matching_index_goalkeeping = goalkeeping_stats_df[matching_game_goalkeeping & matching_date_goalkeeping].index

    if not matching_index_shooting.empty:
        sot = shooting_stats_df.loc[matching_index_shooting, ('Standard', 'SoT%')].values[0]
        premier_league_data.at[index, 'Sot%'] = sot

    if not matching_index_possession.empty:
        touches = possession_stats_df.loc[matching_index_possession, ('Touches', 'Touches')].values[0]
        def_touch = possession_stats_df.loc[matching_index_possession, ('Touches', 'Def Pen')].values[0] + possession_stats_df.loc[matching_index_possession, ('Touches', 'Def 3rd')].values[0]
        mid_touch = possession_stats_df.loc[matching_index_possession, ('Touches', 'Mid 3rd')].values[0]
        att_touch = possession_stats_df.loc[matching_index_possession, ('Touches', 'Att 3rd')].values[0] + possession_stats_df.loc[matching_index_possession, ('Touches', 'Att Pen')].values[0]
        premier_league_data.at[index, 'Touches'] = touches
        premier_league_data.at[index, 'Def Touch'] = def_touch
        premier_league_data.at[index, 'Mid Touch'] = mid_touch
        premier_league_data.at[index, 'Att Touch'] = att_touch

    if not matching_index_passing.empty:
        pass_short = passing_stats_df.loc[matching_index_passing, ('Short', 'Cmp%')].values[0]
        pass_med = passing_stats_df.loc[matching_index_passing, ('Medium', 'Cmp%')].values[0]
        pass_long = passing_stats_df.loc[matching_index_passing, ('Long', 'Cmp%')].values[0]
        premier_league_data.at[index, 'Short Pass'] = pass_short
        premier_league_data.at[index, 'Medium Pass'] = pass_med
        premier_league_data.at[index, 'Long Pass'] = pass_long

    if not matching_index_goalkeeping.empty:
        save = goalkeeping_stats_df.loc[matching_index_goalkeeping, ('Performance', 'Save%')].values[0]
        premier_league_data.at[index, 'Save%'] = save
premier_league_data = premier_league_data.drop(columns="Touches")
premier_league_data.to_csv('premier_league_data.csv', index=False)
print(premier_league_data)
premier_league_data = pd.read_csv('premier_league_data.csv')

# Calculate the correlation matrix
correlation_matrix = premier_league_data.corr()

# Create a heatmap using Seaborn
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap')
plt.show()
premier_league_data["Date"] = pd.to_datetime(premier_league_data["Date"])
premier_league_data["opp"] = premier_league_data["Opponent"].astype("category").cat.codes
premier_league_data["ref"] = premier_league_data["Referee"].astype("category").cat.codes

premier_league_data['Outcome'] = premier_league_data['ScoreDiff'].apply(lambda x: 1 if x > 0 else (0 if x == 0 else -1))
premier_league_data["Ven"] = premier_league_data["Venue"].astype("category").cat.codes

print(premier_league_data)
# Summary Statistics Table with Match Counts
```

```
summary_stats = chelsea_data.groupby('Team').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("Summary Statistics Table:")
print(summary_stats)

# Opponent Analysis Table with Match Counts
opponent_analysis = chelsea_data.groupby('Opponent').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("\nOpponent Analysis Table:")
print(opponent_analysis)

# Home vs. Away Analysis Table with Match Counts
venue_analysis = chelsea_data.groupby('Venue').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("\nHome vs. Away Analysis Table:")
print(venue_analysis)

# Formation Analysis Table with Match Counts (assuming 'Formation' column exists)
formation_analysis = chelsea_data.groupby('Formation').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("\nFormation Analysis Table:")
print(formation_analysis)

# Captain Impact Table with Match Counts (assuming 'Captain' column exists)
captain_analysis = chelsea_data.groupby('Captain').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("\nCaptain Impact Table:")
print(captain_analysis)

# Referee Analysis Table with Match Counts (assuming 'Referee' column exists)
```

```
referee_analysis = chelsea_data.groupby('Referee').agg({'GF': ['mean', 'count'], 'GA': ['mean'], 'Poss': ['mean'], 'Sot%': ['mean'], 'Outcome': 'mean'})
print("\nReferee Analysis Table:")
print(referee_analysis)
X = premier_league_data.drop(["Captain", 'Date', 'Team', 'Opponent', 'Venue', "Formation", "Referee"], axis=1)
X = X.dropna()
y = X['Outcome']

# Drop the 'Outcome' column from X
X = X.drop('Outcome', axis=1)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scale the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize and train an MLP classifier
clf = MLPClassifier(hidden_layer_sizes=(100, 50), max_iter=1000, random_state=42)
clf.fit(X_train_scaled, y_train)

# Make predictions on the testing set
y_pred = clf.predict(X_test_scaled)

# Evaluate the model
accuracy = clf.score(X_test_scaled, y_test)
print("Accuracy:", accuracy)
```

```
# Create a DataFrame to display actual and predicted outcomes
results_df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})

# Display the DataFrame
print("\nActual and Predicted Outcomes:")
print(results_df)
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Compute confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Plot confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Win', 'Draw', 'Lose'], yticklabels=['Win', 'Draw', 'Lose'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```

Link to the data website:

https://fbref.com/en/comps/9/Premier-League-Stats