Shape

Description automatically generated with medium confidence

DS7010

DISSERTATION

Harshil

Predictive Modelling of Football Match Results Using Supervised Learning and Deep Learning

Submitted by: Harshil Goradia

Student ID: 2218896

Submitted to: Dr Yang Li

**Table of Contents**

[**Abstract** 4](#_Toc134218015)

[**Chapter 1: Introduction** 5](#_Toc134218016)

[*1.1 Aims & Objectives* 5](#_Toc134218017)

[*1.2 Project Summary* 5](#_Toc134218018)

[*1.3 Expected* *Output of Practical* 5](#_Toc134218019)

[*1.4 Resources Required* 6](#_Toc134218020)

[*1.5 Nature and Sources of the Data* 6](#_Toc134218021)

[**Chapter 2: Literature Review** 7](#_Toc134218022)

[*2.1 Project Background* 7](#_Toc134218023)

[*2.2 Football* 7](#_Toc134218024)

[*2.3 Data-Technology in Sports* 8](#_Toc134218025)

[*2.4 Machine Learning* 9](#_Toc134218026)

[*2.5 Football Match Prediction* 9](#_Toc134218027)

[**Chapter 3: Methodology** 12](#_Toc134218028)

[*3.1 Data Collection* 12](#_Toc134218029)

[*3.2 Final Dataset* 12](#_Toc134218030)

[*3.3 Exploratory Data Analysis (EDA)* 16](#_Toc134218031)

[*3.4 Data Pre-Processing* 16](#_Toc134218032)

[**Chapter 4: Results** 18](#_Toc134218033)

[*4.1 Exploratory Data Analysis Results* 18](#_Toc134218034)

[*4.1.1 Visual Representations* 19](#_Toc134218035)

[*4.1.2 Hypothesis Testing* 27](#_Toc134218036)

[*4.2 Data Pre-Processing Results* 30](#_Toc134218037)

[*4.3 Machine Learning Results* 31](#_Toc134218038)

[**Chapter 5: Conclusion and Discussion** 40](#_Toc134218039)

[**References** 42](#_Toc134218040)

[**Appendix** 44](#_Toc134218041)

**Table of Content Tables**

[*Table 1. Top 5 European Leagues (Source- https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2023) 9*](#_Toc134195147)

[*Table 2 Attributes of the Dataset 17*](#_Toc134195148)

[*Table 3. Chi-square Test Results. 30*](#_Toc134195149)

[*Table 4. Two-Sample T-tests' Results 31*](#_Toc134195150)

**Table of Content Figures**

[*Figure 1. Distribution % of Home wins, Away wins and Draws. 20*](#_Toc134195190)

[*Figure 2. Distribution of total number of Home wins, Away wins and Draws. 21*](#_Toc134195191)

[*Figure 3. Distribution of average shots and average shots on target 22*](#_Toc134195192)

[*Figure 4. Distribution of wins, losses and draws for top teams. 23*](#_Toc134195193)

[*Figure 5. Number of Home goals scored based on Matchdays. 24*](#_Toc134195194)

[*Figure 6. Number of Away goals scored based on Matchdays. 24*](#_Toc134195195)

[*Figure 7. Home Teams' average expected goals during per match VS Average number of goals scored per game. 25*](#_Toc134195196)

[*Figure 8. Away Teams' average expected goals per match VS Average number of goals scored per game. 26*](#_Toc134195197)

[*Figure 9. Home Team's Average xG vs Average XGA 27*](#_Toc134195198)

[*Figure 10. Correlation Matrix. 28*](#_Toc134195199)

[*Figure 11. SVC confusion matrix. 33*](#_Toc134195200)

[*Figure 12. Gaussian confusion matrix. 34*](#_Toc134195201)

[*Figure 13. Confusion Matrix For XGB. 35*](#_Toc134195202)

[*Figure 14. Confusion matrix for Fast AI. 37*](#_Toc134195203)

[*Figure 15. Accuracy comparison. 38*](#_Toc134195204)

# **Abstract**

This project tries to build a football match prediction machine learning model. Feature engineering, visualisation, model training, and evaluation focus on the top 5 European football leagues. The practical component will produce a football match prediction model based on previous data. The dataset includes win, loss, and draw match data along with match statistics and team statistics. Training and testing data were pre-processed from the dataset. Neural networks, XGBoost, and Support Vectors may be used to learn from past data and predict. On the test dataset, the team must assess the model's accuracy, precision, recall, and F1 score. The practical/investigative component should yield sports betting and team management insights from the predictive model and supporting data. This project seeks to create an accurate football match prediction algorithm for betters, fantasy leagues, clubs, and fans. The study lacks external factors such as injuries, weather conditions, and team spirit. Future research can solve these limitations while improving model's accuracy.

# **Chapter 1: Introduction**

## *1.1 Aims & Objectives*

The aim of the project is to create a machine learning model that can accurately predict the outcomes of the football match, which can be used not only by betters or fantasy leagues but also by the teams and the fans as well.

## *1.2 Project Summary*

This project’s primary purpose is to create a model for predicting football matches using a variety of machine learning techniques. But the goal is to create a model that can predict the outcomes of matches in the top 5 European leagues with more accuracy and can be reliable. In this project, we will be talking about what is football, what are the top five leagues in European football, how have the teams been performing in these leagues, how is the current football prediction working, and entities that can benefit by using these kinds of models. The algorithms may also be used to forecast match outcomes as well as other factors like the number of goals scored because they were created using past data from football leagues. The project is made up of various components, including feature engineering, visualizing the football matches’ stats from the top five leagues for past 5 years, model training, and evaluation. The feature engineering module develops new features by using the data that is already accessible, such as match statistics, team standings, and previous match results. The best machine learning algorithm, hyperparameters, and performance metrics—accuracy, precision, recall, and F1 score—must be chosen. The final model can predict football game outcomes, which is useful for betting or predicting team performance.

## *1.3 Expected* *Output of Practical*

As part of its practical/investigative component, a study on predictive modelling of football match results using machine learning algorithms is anticipated to generate a predictive model for football match outcomes. This model would be used to predict the result of a game based on historical data. The model may use machine learning methods like neural networks, XGBoost, or Support Vectors to learn from past data and make predictions. To create this prediction model, the user would first need to collect and pre-process the relevant historical data, such as match results, team statistics, and previous match data. The user then has to develop and train the machine learning model, selecting the appropriate technique and hyperparameters depending on how well it worked with the training and validation sets of data. The team would also need to evaluate the model's performance using metrics like accuracy, precision, recall, or F1 score on a different test dataset. Insights and suggestions for sports betting or team management based on the study of the predictive model and the supporting data are another expected result of the practical/investigative component. The emphasis of this output will be on how forecasts can be used to guide choices in a range of contexts, including betting and team management. The user would need to analyse the prediction model's output and pinpoint the key characteristics or elements that have a significant impact on the match's outcome in order to generate these insights and suggestions.

## *1.4 Resources Required*

For the project, a dataset of previous football game outcomes, team statistics, and other relevant elements like home advantage, etc., is required. Online databases, sports analytics companies, and open libraries provide this kind of data. The project required data preparation, machine learning, and data visualization tools. Projects like this use Excel, Tableau, Jupyter Notebook, and machine learning libraries like scikit-learn, TensorFlow, FastAI, or PyTorch. A knowledge of basic supervised machine learning, neural networks and model evaluation. Domain knowledge is necessary to confirm the model's accuracy or relevance. In general, a project involving predictive modelling of football match outcomes using machine learning algorithms would call for a combination of technical, analytical, and domain-specific resources.

## *1.5 Nature and Sources of the Data*

The online resource for machine learning called Kaggle and website such as https://footystats.org/ and https://www.football-data.co.uk/ provides access to football match data. Datasets on different football leagues and competitions, such as the English Premier League, La Liga, Serie A, Bundesliga and Ligue 1, are available in these repositories. Football match data, such as scores, player statistics, and team performance, is widely available online for free. These websites include Football-Lineups.com, ESPN, and BBC Sport. To make sure the data sets collected from various sources are pertinent, accurate, and consistent, they must be carefully chosen and pre-processed. Overall, the nature and sources of the dataset(s) would be crucial to the success of a study on utilising machine learning algorithms to forecast football match outcomes.

# **Chapter 2: Literature Review**

## *2.1 Project Background*

The use of technology in the sports industry has significantly increased in recent years. The industry has benefited from the possibilities that technology offers by making continuous improvisations and inventions. "Moneyball", a movie featuring Brad Pitt, sparked a spike in the use of data and statistics to forecast outcomes and make critical decisions in sports, was a significant turning point in this progression. The movie gave the motivation needed to start a project aimed at leveraging data science to create a predictive model for football match results. The goal of this research is to develop a cutting-edge model for forecasting football match results using the data and technologies already available.

## *2.2 Football*

Football has two teams consisting of eleven-player each. The other team's goal is scored by kicking the ball into the goal. Win, draw, or lose, a game can end in three ways. In 90 minutes, the team with the highest score wins (Sfeir, 2011). It is a straightforward and affordable game, which explains why it attained such global acclaim. The most recent World Cup was held in Brazil in 2014, and more than 3.5 billion people watched it on television (FIFA, 2014). 4 billion people watched the London 2012 Olympic Games (Olympic, 2012). The World Cup had 32 nations competing in a single sport, whereas 204 nations competed in 26 different sports at the Olympic Games. This distinction across both of these happenings is significant (Martins, R.G et al., 2017).

According to the number of players and viewers, football is the most watched sport in the world. The sport may be played practically everywhere, from official grounds (pitches) to sports facilities, roads, school campuses, gardens, or seaside, thanks to its basic rules and necessary equipment. A total number of viewers of more than 26 billion people watched football's premier competition, the quadrennial season World Cup finals, in 2010. According to FIFA, there were nearly 250 million soccer athletes and over 130 Cr "interested" in the sport at the turn of the twenty-first century. (Alegi, P. Christopher , Joy, . Bernard , Giulianotti, . Richard C. , Rollin, . Jack and Weil, . Eric, 2023). As we talk about modern football, the most famous and top-rated leagues provide a huge amount of entertainment to the football fans around the world. FIFA World Cup is always important and exciting for the fans, but it happens every four years, so the main focus is always around the leagues for the rest of the duration. I have mentioned the top 5 leagues in Europe at the bottom.

|  |  |  |
| --- | --- | --- |
| **Rank** | **League Name** | **Country** |
| 1 | English Premier League (EPL) | England & Wales |
| 2 | La Liga | Spain |
| 3 | Bundesliga | Germany |
| 4 | Seria A | Italy |
| 5 | Ligue 1 | France |

Table 1. Top 5 European Leagues (Source- [https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2023](https://www.uefa.com/nationalassociations/uefarankings/country/%23/yr/2023))

## *2.3 Data-Technology in Sports*

There has been a rise in adoption of technology within different sports (individual and team both), as the sports industry has gained significantly from the new breakthroughs in technology. Effects of application of data technology on athlete’s performance, fan engagement, and managing big sporting organisations as well as overall sports industry are being reviewed in this section of the literature review.

Buchheit, Simpson, Al Haddad, Bourdon, & Mendez-Villanueva in 2012, discussed the use of data technology in the sports industry in their paper while stating that the incorporation of performance monitoring gadgets to track the abilities of athletes has been one of sports' most important technological developments. Heart rate tracking devices, GPS devices, and accelerometers have been used to gather information on athlete performance, giving coaches and athletes useful insights into topics like fatigue management, minimising injury, and training optimisation.

Technology has also had an impact on how sports organisations are managed, with the introduction of tools for data analytics enabling organisations to choose wisely when it comes to player hiring, contracts for sponsorship, and price tags for tickets. Additionally, the utilisation of video analysis techniques has made it possible managers and scouts to pinpoint player performance weak spots, allowing them to choose teams and strategies with greater knowledge (Hughes & Franks, 2004). But on the other end, Steadman, I., 2013 believe that even though technology has many advantages in sports, some people are worried that it might also have drawbacks. One of the biggest worries is the possibility for technology to provide particular athletes or teams an unfair advantage, along with worries about privacy and data security.

Fan involvement is another area where technology has made a big difference. Sports organisations can now communicate with fans in fresh and creative ways just by taking advantage of social networking services like Twitter and Facebook, which offer in-the-moment information, BTS materials, and exclusive conversations, said by Hays, S., Page, S.J. and Buhalis, D., 2013. Additionally, the introduction of VR and AR technology has improved the viewing experience by enabling spectators to engage with sporting events in novel and immersive ways (Kim, D. and Ko, Y.J., 2019).

Overall, the performance of athletes, fan engagement, and management of sports organisations have all been significantly impacted by the applications of data-technology in sports. Despite worries about any possible adverse impacts of technology, modern sports cannot function without it because of the advantages it offers in regard to data analysis, player performance monitoring, and fan engagement.

## *2.4 Machine Learning*

Creating mathematical and statistical frameworks that allow devices to improve their performance on a particular activity without being explicitly taught is known as Machine Learning (ML), which is a segment of Artificial Intelligence (AI). Machine learning algorithms analyse data to find patterns which are then used to create predictions or alternatives (D. Buursma, 2011). As processing power and data collecting have improved over the past few years, machine learning has become more and more popular.

One of the most popular kinds of ML techniques is supervised learning that trains an algorithm to produce forecasts based on input data using a labelled dataset. Unsupervised learning is a part of machine learning that finds insights and trends in the data without any specific instructions. Machine learning (ML) can be employed in a wide range of industries, like healthcare, robotics, picture and audio recognition and natural language processing (Musa, R.M., et al. 2020). In sports, ML has been utilised for tasks like both individual and collective performance analysis, fitness forecasting, and game outcome prediction.

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., and Honrao, V. (2013) employed data mining categorization to predict student success based on past performance. They examined student data including gender and test grades. They used ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms to predict newly accepted students' exam performance. 75.145% correct. The prediction parameters could not train the fresh dataset when input into the web application because this work was not dynamic.

## *2.5 Football Match Prediction*

Shin and Gasparyan, in 2014 predicted football results using real-world data and FIFA 2015 game data. They forecasted match outcomes better by combining real data with FIFA 2015 player trait data, such as heading, passing, shooting, and strength. They also argued that collecting data from computer games might save a lot of time and effort because real-world data is expensive to calculate or gather.

D. Buursma, 2011, January, devised a football bookie "beat" approach. He identified crucial football match prediction traits and calculated their probabilities to find profitable bets. Seven machine learning algorithms—MultiClassClassifier, RotationForest, LogitbooST, BayesNet, Naive Bayes, and Home Wins—classified the games as home wins, draws, or away wins. His algorithm was 55% accurate. He conceded his strategy failed to "beat" the bookmakers. Predictions were inaccurate, limiting research. Later, the author proposed a better system that included match-related bookings, team players, managers, and more.

Football match outcome prediction has been a popular research topic for several years, with many studies using machine learning algorithms to predict the result of football matches. One early study by (Constantinou, A.C., Fenton, et al., 2013) used Bayesian networks to predict the outcome of English Premier League matches, achieving an accuracy of 69.1%.

 Farzin et al. 2013, used the Bayesian Network Model to forecast the outcomes of games of football between the Barça team in the La Liga (domestic league of Spain) for 2008–2009. They categorised the project's data set into two categories: Weather, five-match history, outcomes against/for team, home game, and players' psychological state are examples of non-physiological elements. Physiological considerations include average player age and the amount of injured principal players, the typical match each week, the performance of the key performers, the performance of every player, and the typical number of goals scored in each home and away game).

By considering all of the games from season 2010–2011, (Snyder 2013) did research to anticipate the Barclays Premier League (now EPL) matches for season 2011–2012. He considered a number of variables like the stadium's capacity, the team's travel distance before the game, and the team's previous season's performance such as ranking, the number of wins, draws, losses, goals scored and conceded, goals difference per game, points, money spent on player salaries, money spent on the 2011 preseason transfer window as well as the number of league games the manager of the team had control over. He used logistic regression to create the model and its success rate for forecasting was 51.06%. He also made an effort to predict what determined the outcome of football games. He came to a conclusion that the two previous games, a few repeating incidents, player appraisal in strike, defence, the players in the middle of the field and keepers were the most crucial variables out of all the variables he used.

Goals, shorts, corners, odds, attack strength, players' performance index, managers' performance index, managers' victory, and teams' win streak are only a few of the nine groups of attributes from home and away statistics that Igiri, C.P., & Nwachukwu, E.O employed in 2014. In his study, he used logistic regression and ANN. He developed the algorithm and tested it using data from the 2014–2015 English Premier League matches, and the results showed that logistic regression had an incredibly high forecasting accuracy of 95%. One of the techniques Reddy, V., & Movva, Sai V. K. (2014), employed to forecast matches in the English Premier League for the 2012–2013 season also involved logistic regression. According to his study's findings, the key factors were the quantity of away goals scored, the quantity of red cards issued, the position of the home team, and the number of shots on target.

(Rahman, M.A., 2020) Deep neural networks (DNNs) and artificial neural networks (ANNs) have attracted increased attention in recent years for their potential use in football match prediction. In order to process a dataset that included positions, team performances, and all prior global football match results, amid other things, this study constructed an effective framework using DNNs and ANNs. To provide prediction values, the dataset was split into training, validating, and testing sections. With a success rate of 63.3%, the suggested DNN architecture performed admirably in terms of forecasting the 2018 FIFA World Cup matches. With the right datasets and more precise team information, this reliability can be improved. This hypothesis' conclusion is that deep learning can be used to accurately forecast the results of football games or any other type of sport. More data on each club, player, and match is preferred for the predictor to perform more accurately.

As new horizons emerge alongside changes in our evolving world comes a need for adaptation . Technology has transformed several areas such as sports bringing with it both immense possibilities as well as numerous challenges Sports now rely heavily on large quantities of data harnessed by machine learning algorithms making athletes' performances transparent at all times. Team performances improve through progressive analysis which demands critical consideration from all parties. Even spectators benefit from innovation since they have access to much richer experience before during and after engagements. Sporting entities should adhere responsible practices which promote fairness inclusion openness regarding affirmative action.

# **Chapter 3: Methodology**

## *3.1 Data Collection*

One cannot underestimate the massive appeal that football holds in every corner of the world which offers a vast pool from which data can be studied to extract useful trends. Researchers frequently resort to credible sources such as footystats.org or footballdata.uk for reliable soccer statistics analysis. The latter website with relevant historical data together with currently occurring soccer match facts drawn from numerous leagues globally comprising detailed metrics such as league table positions, individual players' stats plus game results. This information can be utilised to forecast results as well as identify behaviours and trends in team and player performance. The data collected from this source was for the top 5 European leagues for the past 5 seasons i.e., 17/18, 18/19, 19/20, 20/21 and 21/22 with most of the features of the final dataset. The attributes from this website were leagues from which the team is coming (EPL, La Liga, etc.), date of the match, Home Team, Away Team, number of goals scored by home team in the match, number of goals scored by away team before full-time, number of goals scored by home team and away team before half-time, Full-time result (H, A, D), Half-time result, Home and Away team’s shots taken, Home and Away team’s shots on target, number of fouls by home and away team and number of yellow and red cards for home and away team. The data on the website was for each season in an excel format, so had to download the data for each season from each league, that is 25 different excel books. But there were some issues with data from Ligue 1 for the season 2019-2020 due to which I couldn’t use that data, now that makes 24 excel sheets.

Another website that provides football data is footystats.org, however it places a stronger emphasis on statistics and analytics. The website offers information on more than 500 football leagues and tournaments throughout the world, as well as club statistics, fixture outcomes, and other things. To analyse this data, Footystats.org provides a variety of tools and features, including having the capacity to evaluate teams and players, follow trends throughout years, and visualise data using charts and graphs. The attributes which were scarped from footystats.org were expected goals and expected goals against for both home and away team across all the five leagues for the past 5 years. The scarped data was stored in the excel sheets which contained the rest of the data for the respective season.

Both footballdata.uk and footystats.org offer a plethora of data that may be utilised for a range of tasks, including forecasting match results, and evaluating player performance. Teams of devoted enthusiasts gather and curate the data in an effort to give their users accurate and trustworthy information. Analysts and researchers may make informed judgements that increase performance and result in better outcomes by utilising this data to get insightful knowledge about the world of football.

## *3.2 Final Dataset*

The data from the 24 excel sheets that were downloaded from footballdata.uk website was uploaded to a jupyter notebook. The data had some attributes like betting odds for home team and away team, odds for a team scoring more than 2.5 goals or a team conceding more than 2.5 odds, odds from different betting sites, etc, which were deleted as they were not useful for this project as per me. The scraped data from footystats.org was joined to the match statistics data sheets as per the respective season and the league using excel’s best ever function “VLOOKUPS”. The scraped data had club names and their “xG” and “xGA”, i.e., expected goals and expected goals against, so here the club name’s column was used a primary key and joined with the analysis dataset on Home Team and Away Team as the foreign keys.

The dataset was not yet ready, 4 new columns were added calculated using full-time home goals scored and full-time away goals scored. The new columns were called Home Team Goals scored/ conceded and Away Team Goals scored/ conceded. These columns were than used to create more two columns called as Home team goal difference which is (Home Team Goals Scored – Home Team Goals Conceded) and the column was Away Team goal difference (Away Team Goals Scored – Away Team Goals Conceded). Later, the four columns were dropped as we had the goal differences for all the teams playing away or home game. Similarly, accumulating points for each team from the 5 leagues from their previous games, two columns were added. These attributes were total points for a team in each season. If the team wins, they get 3 points, whereas the opponents go home empty ended but if there’s no winner amongst both the teams and the match is drawn both the teams get a point each. So, using this logic total points were calculated for each team.

Using jupyter notebook’s ‘for loops’, more 10 columns were added to the dataset. The past 5 match results for each team that is Home and Away. If the team had won 2 previous games and lost the one before that and won 2 consecutive games before that game, the values in the five columns would be “W”, “W”, “L”, “W” and “W”. Similarly, more 5 columns were calculated for the opponents. So, a total of 10 columns were created with having values as “w” for a win, “L” for a lose, “D” for a draw and “M” for no previous record. Likewise, more columns were regarded in terms of shots and shots on target for past 2 games for both the Home and Away Team. Instead of using only shots from the current game, we can use the average of shots taken and average of shots on target from the past 2 matches were used. As average of shots taken and average of shots on target were calculated, average of red cards from past 2 matches along with average corners from those previous games were calculated as well.

After creating columns by taking average of the statistics, it was time to create attributes that would help us any interesting streaks made by a club before a match either it be losing or winner for straight 3 or 5 games. Hence, 8 more new columns were built from the dataset, i.e., Home team’s win streak 3 and 5, Home team’s losing streak 3 and 5 and similarly for Away team as well. More two columns like difference in the points of Home Team and Away team, and difference in the points from their 5 previous games which is also known as difference in form points were created. The final dataset which was ready for analysis and interpret the trends and training for forecasting was ready. All the attributes are mentioned and explained below in Table 2, with the data types of their values. The total size of the data was 8750 rows and 47 columns. Few of the columns that were not relevant for the analysis were dropped and the dataset was left with 47 columns out of which one was the dependent variable.

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Information on Columns** | **Data Types** |
| Date | Date of the match | Object |
| Matchday | Matchday number (1-38) | Integer 64 |
| HomeTeam | Team playing on Home ground | Category |
| H\_points | Home Team’s points | Integer 64 |
| HM1 | Home Team’s previous match result | Category |
| HM2 | Home Team’s 2nd previous match result | Category |
| HM3 | Home Team’s 3rd previous match result | Category |
| HM4 | Home Team’s 4th previous match result | Category |
| HM5 | Home Team’s 5th previous match result | Category |
| HTWinStreak3 | Number of time Home Team won 3 consecutive matches | Integer 64 |
| HTWinStreak5 | Number of time Home Team won 5 consecutive matches | Integer 64 |
| HTLossStreak3 | Number of time Home Team lost 3 consecutive matches | Integer 64 |
| HTLossStreak5 | Number of time Home Team lost 5 consecutive matches | Integer 64 |
| HAvgShots | Average Home Team shots in previous 2 games | Integer 32 |
| HAvgShotsOnT | Average Home Team shots on target in previous 2 games | Integer 32 |
| HAvgCorner | Average Home Team corners in previous 2 games | Integer 32 |
| HAvgRed | Average Home Team red cards in previous 2 games | Integer 32 |
| HTGD | Home Team goal difference  (total goals scored – total goals conceded) | Integer 64 |
| FTHG | Full time Home Team Goals | Integer 64 |
| HxG | Home Team average expected goals for the season | Float 64 |
| HxGA | Home Team average expected goals against for the season | Float 64 |
| AwayTeam | Team playing on Away ground | Category |
| A\_points | Away Team’s points | Integer 64 |
| AM1 | Away Team’s previous match result | Category |
| AM2 | Away Team’s 2nd previous match result | Category |
| AM3 | Away Team’s 3rd previous match result | Category |
| AM4 | Away Team’s 4th previous match result | Category |
| AM5 | Away Team’s 5th previous match result | Category |
| ATWinStreak3 | Number of time Away Team won 3 consecutive matches | Integer 64 |
| ATWinStreak5 | Number of time Away Team won 5 consecutive matches | Integer 64 |
| ATLossStreak3 | Number of time Away Team lost 3 consecutive matches | Integer 64 |
| ATLossStreak5 | Number of time Away Team lost 5 consecutive matches | Integer 64 |
| AAvgShots | Average Away Team shots in previous 2 games | Integer 32 |
| AAvgShotsOnT | Average Away Team shots on target in previous 2 games | Integer 32 |
| AAvgCorner | Average Away Team corners in previous 2 games | Integer 32 |
| AAvgRed | Average Away Team red cards in previous 2 games | Integer 32 |
| ATGD | Away Team goal difference  (total goals scored – total goals conceded) | Integer 64 |
| ATHG | Full time Away Team Goals | Integer 64 |
| AxG | Away Team average expected goals for the season | Float 64 |
| AxGA | Away Team average expected goals against for the season | Float 64 |
| DiffPts | Difference in points between Home and Away Team | Integer 64 |
| DiffFormPts | Difference in points between Home and Away Team from 5 previous games | Integer 32 |
| FTR | Full Time Result | Category |

Table 2 Attributes of the Dataset

## *3.3 Exploratory Data Analysis (EDA)*

Exploratory Data Analysis is the process where an enthusiastic analyst interprets range of the data that he/she is dealing with, using few methods he/she explores the dataset. They try to get meaningful insights from the data. Visualizing the data (scatter plots, bar graph, histograms, box plots, etc.), looking at the measures of the central tendency of the attributes, checking correlations between the independent variables with the dependant variable or of two independent variables, running some statistical tests like chi-square test, t-tests, etc., are handful of methods used while exploring the data. EDA for this project was a performed using various tools like, Excel, Python (Jupyter Notebook) and Tableau. Visualization had a major stake in the process for the dataset. Few statistical tests were performed as well on the data which are explained in the *Results* chapter.

The goal of running EDA for the data was to get better understanding of the dataset, find patterns, discover any relationships between the variables, look out for any outliers and also any points that may affect the accuracy of the model that will be developed. Exploring Data Analysis played a crucial part of the machine learning project as it guided to create a hypotheses question and perform subsequent statistical models and tests.

## *3.4 Data Pre-Processing*

Data pre-processing is the part of the model where the raw data is prepared for training for machine learning. A number of techniques/ methods are employed like data cleaning where the raw data is cleaned, missing data is handled, outliers are treated then comes data manipulation where data is manipulated for creating new attributes. Feature engineering is an important method to select the variables that have high relationship with the independent variable. The next step is scaling the data or normalizing the data that helps to get the data into same scale, that helps while training the model. Pre-processing data enables us to improve the accuracy and efficacy of our machine learning models and ensure that their findings are based on reliable and applicable data.

*3.5 Machine Learning*

Machine Learning (ML) is the final part of the model creation project. Machine Learning is a subset of Artificial Intelligence (AI) that help computer systems to get learn from the data with the help of various algorithms. Basically, computer systems use statistical techniques to interpret the patterns and relations between in the data attributes and use those insights to forecast or make decisions for the latest data. Machine learning aims to build computer systems that can learn from the experience and improve over time just like how people do.

Machine Learning is a wide concept with many branches to it. Machine Learning can be divided into three different branches, the branches are called Supervised Learning, Unsupervised Learning and Reinforcement Learning. In this project it was supervised learning that was used. Algorithms like XGBoost, GaussianNB and Support Vector Classifier were the top performing model. Later, the best performing model was tuned with hyper-parameters.

Deep learning (DL) is another branch Machine Learning (ML) which uses many layers of neural networks to build a model. Neural networks are developed in such a way that learn in hierarchical manner to forecast, which is similar to how a human brain works. Fast.ai was used as a practical deep learning model which turned out to the best performing model out of all the other models.

# **Chapter 4: Results**

The findings and results of the study carried out to evaluate the accuracy with which different machine learning models perform in predicting the results of football matches are shown in the Results part of this project. The outcomes of the training, including the models' accuracy, precision, recall, and F1 scores, were carefully analysed in this part. This main objective of this project was to find out the accuracy percentage with which various machine learning algorithms were able to predict football game outcomes. Football is one of the most popular sports in the world, and forecasting match results can give sports analysts, betting companies, and football fans insightful data which makes this a crucial area of research. In order to accomplish it, a dataset of football match statistics from various sources, such as different footballing websites and betting companies was created. The dataset consists statistics of various events that took place during the entire match, such as number of goals scored by home and away teams, shots both the teams, shots on target for both the teams, yellow cards and red cards for both the teams, and other related data. In order to prepare the data for machine learning algorithms, few pre-processing steps like cleaning, normalizing, and encoding  data were taken.

After pre-processing the data, I used the fast.ai library to train and test the data before employing a number of supervised machine learning models on the dataset. Building and training deep learning models, such as neural networks is made simple by the library's user-friendly interface. In order find the ideal settings for each model, I also experimented with a few hyperparameters, including the learning rate, batch size, and number of epochs.

Accuracy, precision, recall, and F1 score are some of the evaluation criteria employed in this study. Precision shows the proportion of true positive predictions out of all positive predictions while accuracy measures the percentage of correct predictions generated by the models. The proportion of true positives among all actual positives is measured by recall, and the F1 score is the harmonic mean of precision and recall. These measures offer an in-depth review of the models' effectiveness and can assist in evaluating the efficiency of different techniques. A complete review of each model's performance, including its accuracy, precision, recall, and F1 score, is provided in the *Results* section. For help in visualizing the accuracy of the models, the section also offers visual representations of the data, such as bar graphs and confusion matrices. Lastly, I have discussed over the results' importance and the way it affects predicting football games.

Overall, this section delivers insightful information about the fact that various machine learning models achieve in predicting football match outcomes. The results may benefit football enthusiasts, betting companies, and sports analysts to improve their predictions and knowledge of the variables affecting game results. The section also outlines future directions for study in this area and highlights the potential of machine learning algorithms in sports analysis.

## *4.1 Exploratory Data Analysis Results*

As above mentioned, Exploratory Data Analysis plays the most crucial part in a predictive machine learning project, the main focus should be laid on to this section of the project. This is the section where the real understanding of dataset is gained. By performing Exploratory Data Analysis, I found some meaningful insights from the data, not only that it also helped to find the patterns and trends in the football matches. Most of the “decision making” factors or knowledge comes from EDA. By looking at the dataset, the creator of the project or the researcher or a company can find out answers to their business question from this part of the project or the research.

### *4.1.1 Visual Representations*

The EDA for a football match prediction dataset, starts with looking at the number of home wins and the percentage of home wins. As the dataset consisted of 8750 different match data from past 5 years of the top 5 leagues of Europe, total number of games played equals to 8750, out of which total number of matches where home team came out of as the winner was 3781. To calculate the percentage of home wins, total number of home wins was divided by total number of games player multiplied with 100. Employing the formula percentage of home win equals to 43.21%. Likewise, for total number of away wins and the percentage of away wins can be calculated. The total number of away wins turned out to be 2763 and using the same formula as for home win percentage, the percentage of away wins popped out to 31.58%. The total number of draws was also calculated on the basis of the same mathematics which turned out to 2206 matches were drawn where both teams scored same number of goals, or both the team could not manage to score a single goal and the percentage of draws equals 25.21%.

***Home/ Away wins percentage = ()\*100***

The diagrams below show the distribution percentage of Home wins, Away wins and Draws and the figure 2 shows the distribution of total number of Home Wins, Away Wins and Draws.

Chart, pie chart

Description automatically generated

Figure 1. Distribution % of Home wins, Away wins and Draws.

Chart, bar chart

Description automatically generated

Figure 2. Distribution of total number of Home wins, Away wins and Draws.

The second visualization was for average shots and average shots on target. Boxplot was chosen to visualize these attributes to check their range. All four of these variables were created from other variables, like average shots for home team was created using shots by home team in their previous 2 games which was then divided by two to get the average and reduce the number of columns. The remaining three variables were built in the same way as the average shot by home team. The maximum range for average number of shots for both home and away team is ~21 which is represented by the upper whisker of the box plot. The median of the average shots by both the team is approximately 11 and their 25th percentile and 75th percentile are 9 shots and 14 shots on average per game respectively, whereas their lower limit (minimum) goes down to 2 shots per game. There are few anomalies detected which were not treated for any of these four attributes as there are quite a lot of matches where a team tries to score more than 25 times in a match. On the other end, the range of average shots on target for both teams were 0 to 13 shots on target and the median was approx. 4 shots on target for home and away teams. Again, there were some outliers for average shots on target for both the teams, but they were not handled as there are days when a team is playing way too better than their opponents, so they get to try more than usual (Stronger team vs Weaker team). This visual representation shows that less than half of the shots taken were actually on target and rest of the shots were either blocked or not the goal.

Chart, box and whisker chart

Description automatically generated

Figure 3. Distribution of average shots and average shots on target

The figure 4., below shows the distribution of number of home wins, away wins and draws for top teams. The top teams were calculated by grouping teams and points scored. The data shows that the top teams were "Man City", "Juventus", "Liverpool", "Barcelona", "Inter", "Real Madrid", "Napoli", "Ath Madrid", "Paris SG" and "Chelsea" according to total points they scored. They have played all the five seasons for their respective leagues and have consistently finished in top positions which makes it obvious that these teams have gained more points. Manchester City leads the top teams with more than 140 wins, approximately 19 draws and 24 losses. Then comes Juventus followed by Liverpool with 125 and 130 wins respectively. The team with lowest number of wins amongst the top teams is Chelsea with just 101 wins from the five English Premier League seasons. The team with lowest number of losses is Liverpool with only 20 losses in 5 premier league seasons instead of having the lowest number of losses they have only managed to crown themselves as champions of England once and that was 19/20 season. The top team that faced the most losses during the five seasons of this project was 50 times for Atletico Madrid from Spanish League (La Liga). This graph helps to find out top performing teams around top 5 leagues of Europe.

Chart, bar chart

Description automatically generated

Figure 4. Distribution of wins, losses and draws for top teams.

Figure 5 (Home teams) and Figure 6 (Away Teams) depicts Home and Away Goals scored on each matchdays respectively. On matchday 1 more than 390 goals were scored by home teams whereas away teams could only manage to bag approximately 330 goals. There have been many ups and downs for both Home and Away teams during 38 matchdays. The highest number of goals scored by Home Team was 395 on 30th matchdays, on the other hand, 3 matchdays later on 33rd matchdays they could only manage to score 270 goals which is the lowest amongst all the matchdays. Away teams’ best performance on the terms of goals scored was on matchdays 2 with a total of ~330 goals scored but they could only score 208 goals on 32nd matchdays which is the least number of goals scored by them on any matchdays.

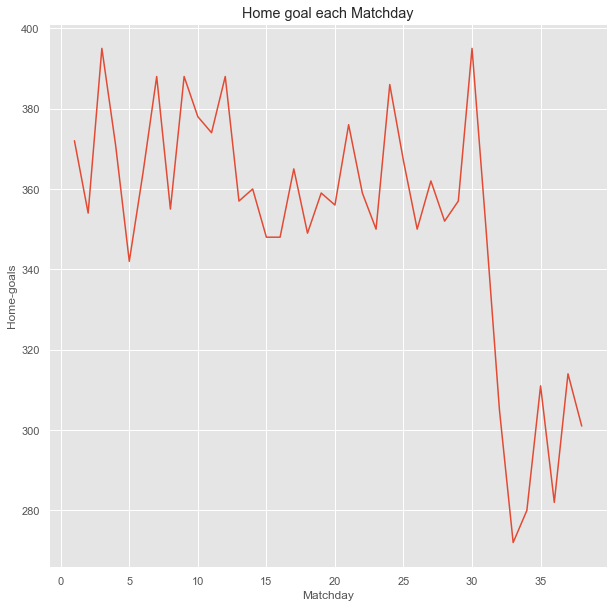


Figure 5. Number of Home goals scored based on Matchdays.

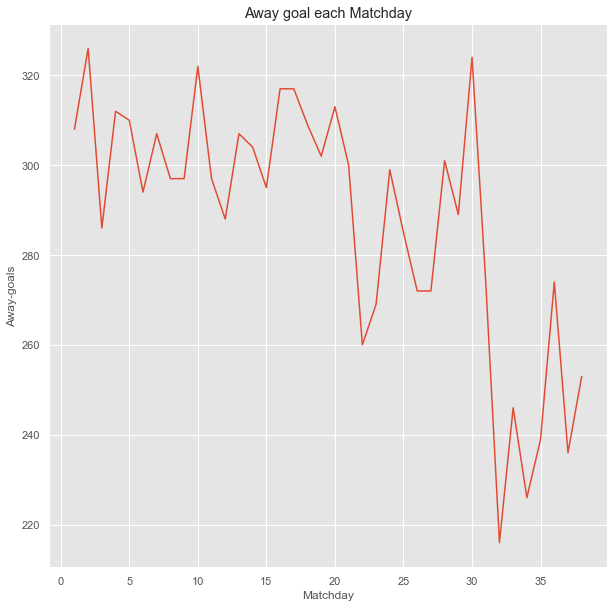


Figure 6. Number of Away goals scored based on Matchdays.

The Figure 7. shows comparison between the average number of goals scored by Home teams in a single match against average expected goals to scored by the Home Teams in a match for five different seasons. The home teams were expected to score 1.52 goals per game during 17/18 season, whereas they could only score 1.5 goals per match on an average. They couldn’t meet the average expected number of goals per match during that season. They surpassed the average expected number of goals during the next two consecutive seasons by scoring 1.52 goals and 1.58 goals per match on an average respectively, while they were only expected to score 1.5 goals per match on an average during the 18/19 season and for the 19/20 season, they were expected to score 1.51 goals per match on an average. The 20/21 season was not great season as Home Teams lacked 0.3 goals per match on average than the average expected goals to be scored per match. However, the next season they scored as many goals as they were expected to score per match on an average.

Likewise, the Figure 8. represents the average expected number of goals to be score by away team per match vs how many average goals they scored per game. The away teams managed to surpass the average expected goals to be scored per match by scoring approximately 0.6 more goals per match on average for the year 2017. 2020 was the only year when the away team could not manage to score as many goals on average per match as they were expected to score but they were only behind 0.1 goals per match on average. The rest of the years they managed to easily surpass the average expected goals to be scored in a match by score more goals in a match in reality on average. They achieved the highest average goals scored in a match during the last season with scoring 1.35 goals on average per match. However, 2017 still remains the best year as per performance as they managed to score 0.6 more goals than average expected number of goals per match.

Chart, histogram

Description automatically generated

Figure 7. Home Teams' average expected goals during per match VS Average number of goals scored per game.

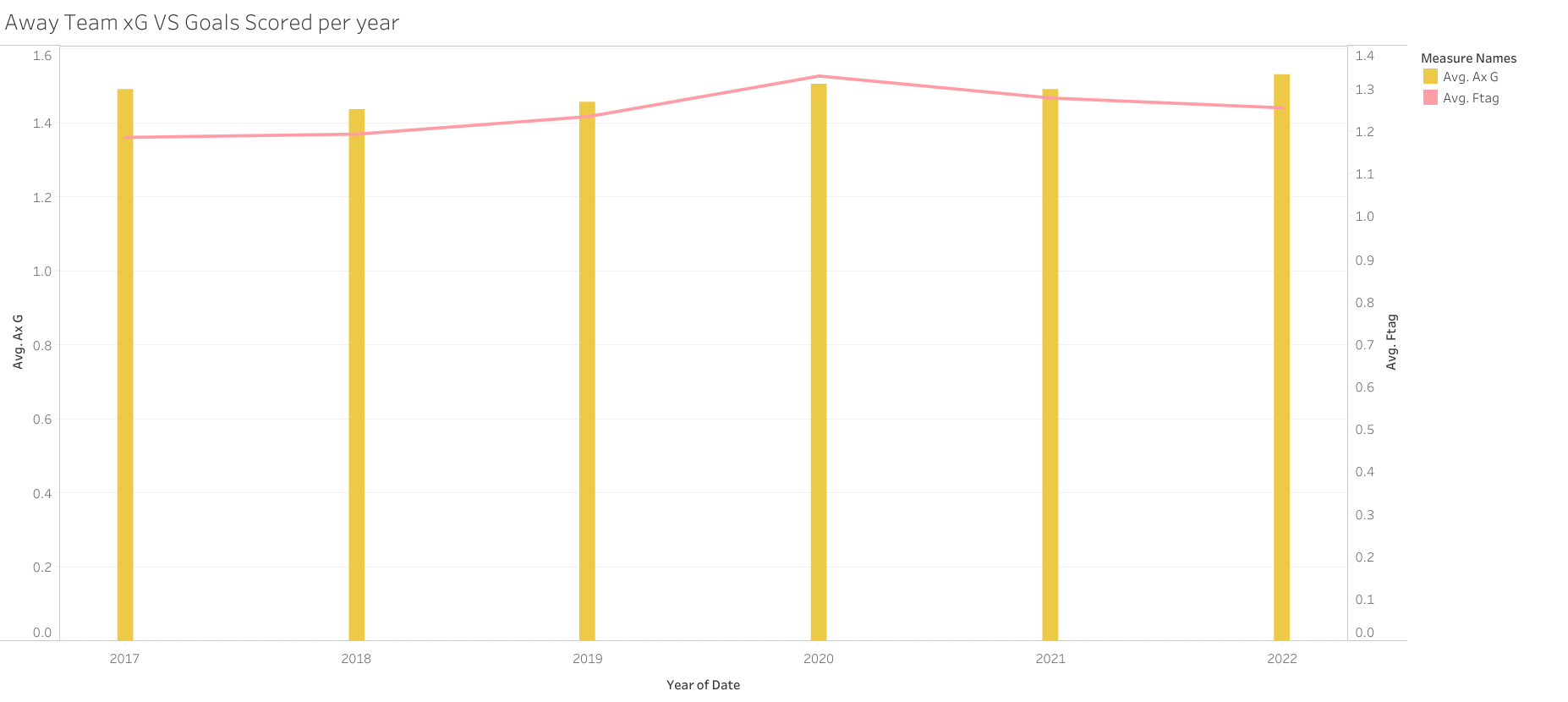


Figure 8. Away Teams' average expected goals per match VS Average number of goals scored per game.

The Figure 9. represents the Home Teams’ average expected goals vs average expected goals against. The black lines represent the average xGA for all the home teams and the orange bars represent the average expected goals for the home teams. The team with the highest average expected goals during the 5 seasons was Bayern Munich which 2.4 xG. They were followed by Manchester City with slight difference of 0.1 xG. But Man City are ahead in having the lowest average xGA with only 0.9 xGA. Many teams have less average xG than their average expected goals against, which makes them weaker teams and there are high chances that they score less than they concede which results in losing the match and maybe relegation as well.

Chart, bar chart

Description automatically generated

Figure 9. Home Team's Average xG vs Average XGA

Correlation Matrix, in the below figure we can check the relationship of independent variables with other independent variables. To interpret the findings from the confusion matrix, try look out for darker boxes which indicate a high positive or a negative relationship between two variables. Here in this figure a dark shade of orange indicates a high positive correlation between the variables and darker shade of blue represents a high negative correlation between the variables. Home average shots and Home average shots on target are highly correlated, as the teams tries to take more shots the percentage of shots being on target increases as well. Hence there is a positive correlation. Similarly, the away shots are positively correlated to away shots on target.

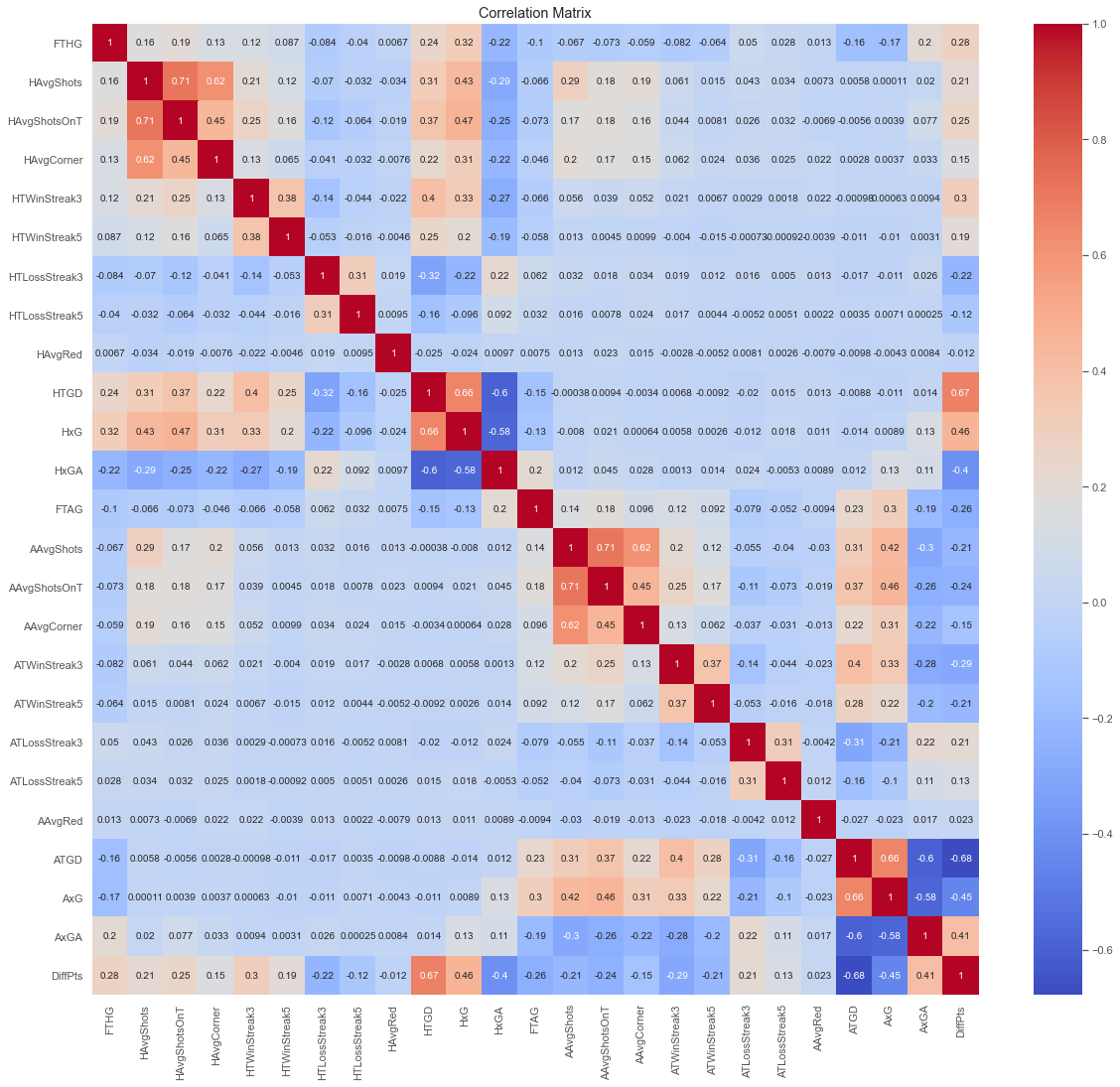


Figure 10. Correlation Matrix.

### *4.1.2 Hypothesis Testing*

**Chi-Square Tests:**

The chi-square test is a non-parametric test that used for a pair of different purposes: and they are, to test the null hypothesis that there is no link between multiple groups, populations, or criteria (to validate the independence within two factors), and the other is to determine the probability that the observed data distribution corresponds to the expected distribution (i.e., to analyse the goodness-of-fit). It is used to analyze categorical data, such as the fact that a patient is male or female, if they smoke or not, etc. Hence it is not meant to analyze continuous or parametric data like height measured in centimetres or weight measured in kilogrammes (Sharpe, D., 2015).

A picture containing text, watch, clock, gauge

Description automatically generated

Equation 1. Chi-Square Test (Sharpe, D., 2015)

**FTR and HM1**

The first two variables that were statistically tested were FTR and HM1. These two factors were tested using chi square test. After testing the results, the p-value turned out to be 9.85908114924601e-52 and the chi-statistics was 252.8967. So, it is very unlikely that the relationship between the two factors (FTR and HM1) could have occurred by accident. In other words, p-value being less than alpha shows that the relationship between these factors is real and significant. So, there is plenty of evidence to support the alternative hypothesis that the result of a football game (FTR) is related to the result of the previous match result of the Home Team and this link is not basically the result of chance.

Chi-square test results:

Chi-square statistic: 252.8967378181011

P-value: 9.85908114924601e-52

Degrees of freedom: 6

**FTR and AM1**

The next variables that were statistically tested were FTR and AM1. After testing the results, the p-value turned out to be 3.9764178571030157e-41 and the chi-statistics was 203.1882. So, it is very unlikely that the relationship between the two factors (FTR and AM1) could have occurred by accident. In other words, p-value being less than alpha shows that the relationship between these factors is real and significant. So, there is plenty of evidence to support the alternative hypothesis that the result of a football game (FTR) is related to the result of the previous match result of the Away Team and this link is not basically the result of chance.

Chi-square test results:

Chi-square statistic: 203.18823285829524

P-value: 3.9764178571030157e-41

Degrees of freedom: 6

Similarly, more chi-square tests were conducted using different variables from the dataset. The table below shows the result from all the chi-square tests conducted during the project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable 1** | **Variable 2** | **P value** | **Chi-square statistics** | **Result (Accept/ Reject Null Hypothesis)** |
| FTR | HM1 | 9.85908114924601e-52 | 252.8967378181011 | Reject |
| FTR | HM2 | 4.987901859324699e-15 | 79.29835201269347 | Reject |
| FTR | AM1 | 3.9764178571030157e-41 | 203.18823285829524 | Reject |
| FTR | AM2 | 1.0354771652672775e-09 | 53.2694201863767 | Reject |
| FTR | HM5 | 1.0460892701114172e-19 | 101.82063405353365 | Reject |
| FTR | AM5 | 1.2081288105474381e-17 | 91.92161346184085 | Reject |

Table 3. Chi-square Test Results.

**T-Tests:**

The two-sample t-test compares the means of two independent variables to calculate a t-statistic and a p-value. If the null hypothesis is true, then the p-value is the probability of detecting such a difference or one even more significant.

**FTHG and FTAG**

The first two variables that were being tested were FTHG (Full-time Home Goals scored) and FTAG (Full-time Away Goals scored). Where the null hypothesis was set as there is no significance difference between the means of goals scored by home and away teams. After running the two-sample t-tests, the p-value turned out to be less that alpha which was predetermined as 0.05 or 5%. As the p-value is smaller than alpha, the null hypothesis is rejected in support of the alternative hypothesis because it is highly unlikely that the difference or a more extreme difference between the means of goals scored would be observed if the null hypothesis were true.

**HAvgShots and AAvgShots**

The next two variables that were being tested were HAvgShots (Home Average Shots) and AAvgShots (Away Average Shots). Where the null hypothesis was set as there is no significance difference between the means of shots taken by home and away teams. After running the two-sample t-tests, the p-value turned out to be less that alpha which was predetermined as 0.05 or 5%. As the p-value is smaller than alpha, the null hypothesis is rejected in support of the alternative hypothesis because it is highly unlikely that the difference or a more extreme difference between the means of goals scored would be observed if the null hypothesis were true.

Similarly, more two-sample t-tests were conducted using different variables from the dataset. The table below shows the result from all the two-sample t-tests conducted during the project.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable 1** | **Variable 2** | **P value** | **Result (Accept/ Reject Null Hypothesis)** |
| FTHG | FTAG | 2.114606899851079e-49 | Reject |
| HxG | AxG | 1.0 | Accept |
| HxGA | AxGA | 0.9999999999999488 | Accept |
| HAvgShots | AAvgShots | 0.004949315437917223 | Reject |
| HAvgShotsOnT | AAvgShotsOnT | 0.11755027607528284 | Accept |
| HAvgRed | HAvgRed | 0.7129253344918329 | Accept |
| HAvgCorner | AAvgCorner | 0.07540265608307488 | Accept |

Table 4. Two-Sample T-tests' Results

## *4.2 Data Pre-Processing Results*

Data Pre-processing is the section of the project where the data is been treated for preparing it for training on machine learning models, i.e., any new columns added using the available data, encoding the categorical columns and normalization. These are common techniques used while processing the data to gain better and reliable accuracy of the model.

Many new columns were created to gain better insights from the data and also make the model more efficient. Columns like HTGD (Home Team Goal difference) and ATGD (Away Team Goal difference) were created using FTHG and FTAG. Furthermore, many columns like H\_points and A\_points, HM1, HM2, HM3, HM4, HM5, AM1, AM2, AM3, AM4 and AM5, HAvgShots and AAvgShots, HAvgShotsOnT and AAvgShotsOnT, HAvgRed and AAvgRed, HAvgCorner and AAvgCorner, Home and Away Winning/ Lossing Streaks and Difference in points were being created and the attributes that were not required after getting new columns were dropped from the dataset.

As the data was ready, the only two steps before machine learning were encoding the data and applying normalization to the data. Hence, label encoding was used to turn categorical variables into numbers. All the categorical variables were now converted into numbers so the whole data was ready for normalization. Thus, using MinMaxScaler function the data was normalized.

The data was yet not ready for putting it into training as there were too many attributes that could affect the accuracy of the model. Hence feature selection was applied and the only the features that had an impact on the dependent variable were selected. After selecting the features using SelectKBest, the data was ready for training employing various classification models.

## *4.3 Machine Learning Results*

This is the section of the project where the prediction is done. The data is first divided into the training data and testing data and then different model are used to predict the future trends or future outcomes. For this project the prediction of the football match outcomes using the past 5 seasons data from top 5 European leagues has been done using three classification models like Support Vector Classifier, Gaussian NB and XGBoostClassifier. Not only classification models have been used but also a Deep learning algorithm is used which was a good performing model alongside the classification models.

**Support Vector Classifier**

A support vector machine (SVM) is a type of supervised machine learning model that solves classification issues by applying classification techniques (Han J, et al. 2011)(Cortes C, et al. 1995)( Abdullah DM, Abdulazeez AMJQAJ, 2021). An SVM model can categorise incoming text after being given sets of training data that have been labelled for each category. They are more efficient and perform better with fewer samples (in the thousands) than more recent algorithms like neural networks. Due to the fact that classification issues frequently only have access to datasets with a few thousand or so tagged samples at most, the technique is ideally suited for these types of problems (Nisbet R, et al. 2009)( Naveed N, Jaffar AJIJoPS , 2011)( Kamalakannan J, et al. 2015).

After performing support vector classifier on the football match prediction dataset, the accuracy of the model turned out to be 0.81. For more better understanding of the classification report for support vector classifier has been mentioned below. The precision for away wins is 0.87, and for draws It is 0.65 and it is the maximum for home wins with 0.88.

**Accuracy: 0.8110624315443593**

**Report: precision recall f1-score support**

**0 0.87 0.79 0.83 573**

**1 0.65 0.71 0.67 472**

**2 0.88 0.89 0.88 781**

**accuracy 0.81 1826**

**macro avg 0.80 0.80 0.80 1826**

**weighted avg 0.82 0.81 0.81 1826**

The terms "True" and "False" refer to the model's right and wrong predictions, respectively while "Positive" and "Negative" represent the predicted class labels in a confusion matrix. So the True Positive for class 0 (Away wins) is 455, False Positive is 106+12 = 118, True Negative is 1199 and False Negative is 66 which is a sum of 54+12. Similarly, TP, FP, TN and FN for class 1 (Draws) is 333, 257, 1236 and 139 respectively. While the class 2 (Home wins) had a TP of 693, FP of 160, TN of 1079 and the last FN is 88. This is how the confusion matrix looks for SVC.

A picture containing timeline

Description automatically generated

Figure 11. SVC confusion matrix.

The model was also treated with cross-validation process. When the cross-validation technique was used on Support Vector Classifier with 5 folds the mean of the score turned out to be 0.788 (78.8%).

**CODE FOR SUPPORT VECTOR CLASSIFIER:**

*svc = SVC()*

*# fit the classifier on the training data*

*svc.fit(X\_train, y\_train)*

*# make predictions on the test data*

*svc\_pred = svc.predict(X\_test)*

*# evaluate the performance of the classifier*

*accuracy = accuracy\_score(y\_test, svc\_pred)*

*report = classification\_report(y\_test, svc\_pred)*

*print(f"Accuracy: {accuracy}")*

*print(f"Report: {report}")*

*cm = confusion\_matrix(y\_test, svc\_pred)*

*# Plot confusion matrix*

*sns.heatmap(cm, annot=True, cmap='Blues', fmt='g')*

*plt.xlabel('Predicted')*

*plt.ylabel('True')*

*plt.show()*

**Gaussian Naïve Bayes**

A special version of the Naive Bayes algorithm is called the Gaussian NB which is used for classification (Witten IH, Frank EJASR, 2002). This technique is most helpful when the attributes possess continuous values or when every feature has a Gaussian distribution just like normal distribution (Karabatak, M., 2015). Maximum likelihood estimation is used to determine the parameters x and y in equation, where x is the variable. The mean and variance are calculated once the data has been divided into classes (Witten IH, Frank EJASR, 2002).

A picture containing text, clock, gauge

Description automatically generated

Equation 2. Gaussian Naive Bayes Equation (Khorshid, S.F., et al. 2021)

After performing Gaussian Naïve Bayes on the football match prediction dataset, the accuracy of the model turned out to be 0.6736 which 67.36%. For more better understanding of the classification report for Gaussian has been mentioned below. The F1 scores for away wins is 0.72, and for draws It is 0.54 and it is the maximum for home wins with 0.75.

**Accuracy: 0.6736035049288062**

**Report: precision recall f1-score support**

**0 0.76 0.68 0.72 573**

**1 0.46 0.64 0.54 472**

**2 0.82 0.69 0.75 781**

**accuracy 0.67 1826**

**macro avg 0.68 0.67 0.67 1826**

**weighted avg 0.71 0.67 0.68 1826**

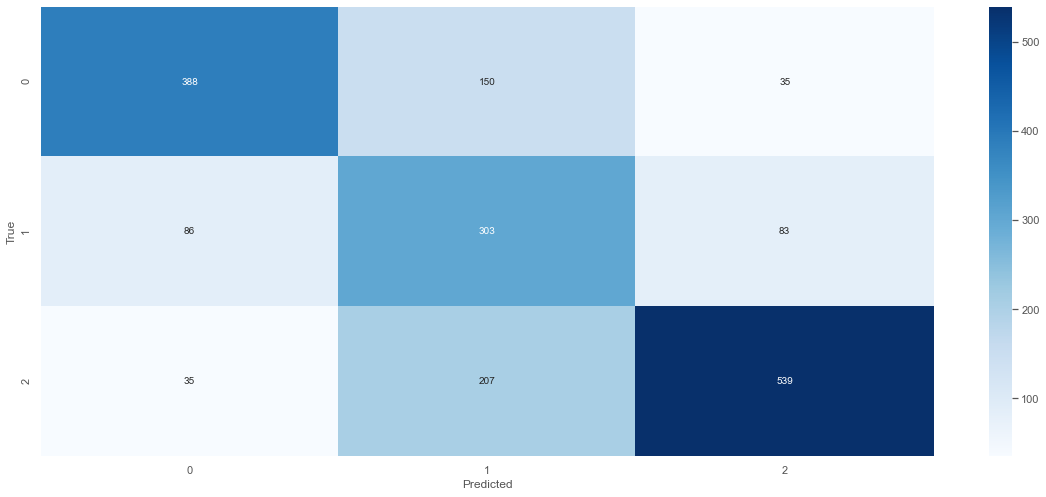


Figure 12. Gaussian confusion matrix.

The True Positive for class 0 (Away wins) is 388, False Positive is 185, True Negative is 1132 and False Negative is 121. Similarly, TP, FP, TN and FN for class 1 (Draws) is 303, 169, 1169 and 185 respectively. While the class 2 (Home wins) had a TP of 539, FP of 242, TN of 1152 and the last FN is 118. This is how the confusion matrix looks for Gaussian NB. The model was also treated with cross-validation process. When the cross-validation technique was used on GaussianNB with 5 folds the mean of the score turned out to be 0.663(66.3%).

**CODE FOR GAUSSIAN NAYE BAYES:**

*# create a Gaussian Naive Bayes classifier*

*gnb = GaussianNB()*

*# fit the classifier on the training data*

*gnb.fit(X\_train, y\_train)*

*# make predictions on the test data*

*y\_pred = gnb.predict(X\_test)*

*# evaluate the performance of the classifier*

*accuracy = accuracy\_score(y\_test, y\_pred)*

*report = classification\_report(y\_test, y\_pred)*

*print(f"Accuracy: {accuracy}")*

*print(f"Report: {report}")*

*cm = confusion\_matrix(y\_test, y\_pred)*

*# Plot confusion matrix*

*sns.heatmap(cm, annot=True, cmap='Blues', fmt='g')*

*plt.xlabel('Predicted')*

*plt.ylabel('True')*

*plt.show()*

**XGBoost Classifier**

A key part of XGBoost is the way it boosts gradients. Unlike other gradient boosting algorithms, XGBoost doesn't add weak learners one at a time. Rather, it applies a multiple-threaded technique that maximises the CPU core's speed and efficiency. Sparse aware version also handles missing data values efficiently, employs a block structure to build trees in parallel, and uses continuing training to develop a model based on incoming data. For classification, regression, and predictive modelling, XGBoost works better on ordered or tabular information (Ramraj, S., Uzir, N et al. 2016).

After performing XGBoost Classifier on the football match prediction dataset, the accuracy of the model turned out to be 0.99 which 99%. For more better understanding of the classification report for XGBoost has been mentioned below. The F1 scores for away wins is 0.99, and for draws It is 0.99 and it is the maximum for home wins with 1.

**Accuracy: 0.9906900328587076**

**Report: precision recall f1-score support**

**0 0.98 1.00 0.99 573**

**1 1.00 0.96 0.98 472**

**2 0.99 1.00 1.00 781**

**accuracy 0.99 1826**

**macro avg 0.99 0.99 0.99 1826**

**weighted avg 0.99 0.99 0.99 1826**

Graphical user interface

Description automatically generated with low confidence

Figure 13. Confusion Matrix For XGB.

This is how the confusion matrix looks for XGBoost Classifier. The True Positive for class 0 (Away wins) is 573, False Positive is 0, True Negative is 1236 and False Negative is 17. Similarly, TP, FP, TN and FN for class 1 (Draws) is 455, 11, 1209 and 6 respectively. While the class 2 (Home wins) had a TP of 781, FP of 0, TN of 1229 and the last FN is 0. The model was also treated with cross-validation process. When the cross-validation technique was used on XGBoost Classifier with 5 folds the mean of the score turned out to be 0.994(99.4 %). This model was achieved after hyper-parameter tuning for XGBoost. The best performing parameters were looked out for and then the model was trained setting those parameter values and it turned out to be the best performing model for out prediction which is football match outcomes.

**CODE FOR XGBOOST:**

*# fit the classifier on the training data*

*xgb.fit(X\_train, y\_train)*

*# make predictions on the test data*

*xbg\_pred = xgb.predict(X\_test)*

*# evaluate the performance of the classifier*

*accuracy = accuracy\_score(y\_test, xbg\_pred)*

*report = classification\_report(y\_test, xbg\_pred)*

*print(f"Accuracy: {accuracy}")*

*print(f"Report: {report}")*

*cm = confusion\_matrix(y\_test, xbg\_pred)*

*# Plot confusion matrix*

*sns.heatmap(cm, annot=True, cmap='Blues', fmt='g')*

*plt.xlabel('Predicted')*

*plt.ylabel('True')*

*plt.show()*

**Fast AI**

According to (Howard, J. and Gugger, S., 2020), Fast.ai is a DL package that offers researchers low-level aspects that may be combined to create unique methods as well as practitioners’ high-level aspects that can quickly and effortlessly offer cutting-edge outcomes in typical deep learning domains. It looks for accomplish both without making significant concessions in terms of use, adaptability, or performance. This is made possible by a carefully designed layering architecture that uses decoupled abstractions to describe common underlying patterns of numerous deep learning and data processing approaches. Ready-to-use functions that train models for a variety of applications are powered by a high-level API and provide customisable models with sane defaults (Howard, J. and Gugger, S., 2020). A hierarchy of lower-level APIs that offer composable building components is created on top of it. Beginners and professionals who are primarily interested in using existing deep learning techniques will likely find the more complex aspects of the API to be most helpful. Vision, text, tabular and time-series analysis, as well as collaborative filtering, are the four core application domains covered by its succinct APIs. (Howard, J. and Gugger, S., 2020)

Chart

Description automatically generated with medium confidence

Figure 14. Confusion matrix for Fast AI.

The True Positive for class 0 (Away wins) is 559, False Positive is 14, True Negative is 1232 and False Negative is 32. Similarly, TP, FP, TN and FN for class 1 (Draws) is 451, 20, 1341 and 13 respectively. While the class 2 (Home wins) had a TP of 768, FP of 24, TN of 1031 and the last FN is 3. This is how the confusion matrix looks for Fast AI looks. Once the model was trained and tuning was done with different learning rates while, the best result was achieved when the learning rate was set to 1e-01 and epoch at 10. The accuracy of the model was 0.976451 with 0.215764 training loss and the validation loss was 0.082190. The Fast AI was the second-best performing model for predicting football match outcomes.

A new dataset was downloaded for the current season of the same leagues which was used to predict the outcomes of the future matches. Data pre-processing was done as per needed and then the outcomes were predicted for the data.

**CODE FOR FAST AI:**

*data = (TabularList.from\_df(finaldata, cat\_names=cat\_names, cont\_names=cont\_names, procs=procs)*

*.split\_by\_idx(list(range(6924,8750)))*

*.label\_from\_df(cols=y\_names)*

*.add\_test(test\_fastai)*

*.databunch())*

*data.train\_dl = data.train\_dl.new(shuffle=False)*

*learn = tabular\_learner(data, layers=[200,100], metrics=accuracy)*

*learn.model*

*learn.lr\_find() # find the optimal learning rate*

# **Chapter 5: Conclusion and Discussion**

The football match outcome prediction has been a hot topic since the last century. There are many researchers, sports analysts, data experts or mathematicians/ statisticians who have been trying to the predict the match outcomes of football games. Not only football games but various other sports’ outcomes are tried to being predicted. But none of the research or models have achieved the top success in predicting the outcomes. The fun of watching sports as a fan is directly proportional to make assumptions while watching, so this kind of projects are going to disappoint the fans, but the sports analysts, coaches and betting companies are the ones that can make the most out of it. Models like this can be used by sports analysts and coaches to make tactical changes or find out the problem where they are lacking just by referring trends and patterns for the team. Betting companies can use this model to set the bets on the teams.

The model that has been created in this project was created using various team statistics and events of the matches from the past 5 seasons of top five European leagues. The visual representation of the dataset gives a lot of meaningful insights that helps to find out the top performing teams at home and away matches, the amount of the goals that teams have been scoring during the seasons, trend in average number of shots taken, trend in average shots on target, etc., this insights can be used by the opponents or the team itself to improve the performance and increase their chances of winning. The hypotheses that were being raised gave an effective understanding of the relations between the variables and if there were any significant differences between them or their means. The machine learning models were fantastic as they performed really well and can predict the outcomes effectively. The SVC model performed average with 81.11% accuracy. Gaussian NB was the worst performing model with only 67.36% accuracy. XGBoost and Fast AI a deep learning model, were the best models in terms of accuracy as they achieved 99% accuracy and 97.64% accuracy. Hyper parameter tuning was used to get the influencing parameters for the model. The figure below shows the comparison of the accuracy of the models.

Chart, bar chart

Description automatically generated

Figure 15. Accuracy comparison.

Predicting football game results has a number of restrictions. The following are some of the significant restrictions:

1. Limited data: A number of variables like player performance, team strategy, the environment, and injuries may affect the outcome of a football game. Data on these variables is limited and particularly for matches in low level leagues. This can make it challenging to correctly forecast a game's result.
2. Randomness: Football games seem to be random events that can never be predicted. Even the best teams occasionally lose to an inferior opponent and the opposite is also true. Football is enjoyable to watch because of this randomness, but it also makes it challenging to accurately forecast results.
3. Human errors: Human mistakes occur at every level of the game of football for example the referees, coaches, and players on the pitch. Predictive models may have difficulty considering these human mistakes, which can have a major effect on a game's outcome.
4. Unforeseen events: Unexpected occurrences like an important player having injuries or sudden shifts in team tactics may also drastically impact the results of matches. It can be difficult to accurately predict these events which makes it tough to create predictive models.
5. Non-linear connections: When it comes to match results there can be a non-linear relationship among the various elements which means that even a small change to a single factor can have a big effect on the other. Hence, it may be challenging to accurately predict the result of a game using just one aspect or variable.

Hence, despite of the fact that predictive models can prove to be helpful in predicting the football match outcomes, it is crucial to be aware of these limitations and proceed with caution.

Machine learning techniques for football match prediction have tremendous potential. The future prospects of football match prediction include:

1. More data: Machine learning algorithms predict match results using prior match data, team statistics, and player data. The models can also include weather, injuries, and suspensions, which can dramatically affect match outcomes. This data improves model accuracy and reliability.
2. Real-time prediction: Machine learning models use past data to make predictions before the match. With real-time data like live scores, injuries, and substitutions, models can forecast match results. This may benefit in-game betting and tactical adjustments.
3. Deep learning models: CNNs and RNNs have showed promise in other machine learning applications. These football match prediction methods may outperform typical machine learning algorithms.
4. Including player-specific data: Machine learning models currently use team statistics but not player statistics. Player statistics like passing accuracy, goals scored, and tackles made can improve models.
5. Explaining predictions: Machine learning models are called black boxes because they are hard to understand. It would be helpful to construct models that explain their forecasts and explain why a certain team is more likely to win. This aids betting and team management.

# **References**

Abdullah DM, Abdulazeez AMJQAJ. Machine Learning Applications based on SVM Classification A Review. 2021;1(2):81-90.

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R. and Honrao, V., 2013. Predicting students' performance using ID3 and C4. 5 classification algorithms. arXiv preprint arXiv:1310.2071.

Alegi, P. Christopher, Joy, Bernard, Giulianotti, Richard C., Rollin, Jack, and Weil. Eric (2023, April 21). football. Encyclopedia Britannica. <https://www.britannica.com/sports/football-soccer>

Buchheit, M., Simpson, M.B., Al Haddad, H., Bourdon, P.C. and Mendez-Villanueva, A., 2012. Monitoring changes in physical performance with heart rate measures in young soccer players. European journal of applied physiology, 112, pp.711-723.

Buursma, D., 2011, January. Predicting sports events from past results Towards effective betting on football matches. In Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland (Vol. 21). sn.

Constantinou, A.C., Fenton, N.E. and Neil, M., 2013. Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. Knowledge-Based Systems, 50, pp.60-86.

Cortes C, Vapnik VJMl. Support-vector networks. 1995;20(3):273-297.

Farzin, O., Parinaz, E., &Faezeh, S. M. Football result prediction with Bayesian network in Spanish league-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 2013, 812-815

Han J, Kamber M, Pei JJTMKSiDMS. Data mining concepts and techniques third edition. 2011;5(4):83-124.

Hays, S., Page, S.J. and Buhalis, D., 2013. Social media as a destination marketing tool: its use by national tourism organisations. Current issues in Tourism, 16(3), pp.211-239.

Howard, J. and Gugger, S., 2020. Fastai: A layered API for deep learning. Information, 11(2), p.108.

Hughes, M. and Franks, I., 2005. Analysis of passing sequences, shots and goals in soccer. Journal of sports sciences, 23(5), pp.509-514.

Igiri, C.P., & Nwachukwu, E.O. (2014). An Improved Prediction System for Football a Match Result. IOSR Journal of Engineering Volume 04 Issue 12, pp12-20

Kamalakannan J, Thirumal T, Vaidhyanathan A, MukeshBhai KD. Study on different classification technique for mammogram image. in 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], 2015;1-5:IEEE.

Karabatak, M., 2015. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, pp.32-36.

Khorshid, S.F., Abdulazeez, A.M. and Sallow, A.B., 2021. A comparative analysis and predicting for breast cancer detection based on data mining models. Asian Journal of Research in Computer Science, 8(4), pp.45-59.

Kim, D. and Ko, Y.J., 2019. The impact of virtual reality (VR) technology on sport spectators' flow experience and satisfaction. Computers in human behavior, 93, pp.346-356.

Martins, R.G., Martins, A.S., Neves, L.A., Lima, L.V., Flores, E.L. and do Nascimento, M.Z., 2017. Exploring polynomial classifier to predict match results in football championships. Expert Systems with Applications, 83, pp.79-93.

Musa, R.M., Majeed, A.P.A., Kosni, N.A. and Abdullah, M.R., 2020. Machine learning in team sports: performance analysis and talent identification in Beach Soccer & Sepak-takraw. Springer Nature.

Naveed N, Jaffar AJIJoPS. Malignancy and abnormality detection of mammograms using DWT features and ensembling of classifiers. 2011;6(8):2107- 2116.

Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications. Academic Press; 2009.

Rahman, M.A., 2020. A deep learning framework for football match prediction. SN Applied Sciences, 2(2), p.165.

Ramraj, S., Uzir, N., Sunil, R. and Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, *9*(40), pp.651-662.

Reddy, V., & Movva, Sai V. K. (2014). The Soccer Oracle: Predicting Soccer Game Outcomes Using SAS® Enterprise Miner™. SAS® GLOBAL FORUM. Washington, D.C.

Sfeir, M.N., 2011. Laws of the Game (adapted from FIFA 2010-11). World Literature Today, 85(3), p.38.

Sharpe, D., 2015. Chi-square test is statistically significant: Now what? Practical Assessment, Research, and Evaluation, 20(1), p.8.

Shin, J. and Gasparyan, R., 2014. A novel way to soccer match prediction. Stanford University: Department of Computer Science.

Snyder, J., 2013. What actually wins soccer matches: Prediction of the 2011-2012 Premier League for fun and profit.

Steadman, I., 2013. Big data and the death of the theorist. Wired Online, 25, p.2013.

Witten IH, Frank EJASR. Data mining: practical machine learning tools and techniques with Java implementations. 2002;31(1):76-77

Zheyuan, F., Yuming, K. and Xiaolin, L., 2013. Chess game results prediction system.

# **Appendix**

Click on the text below to check out the data used for the project and also the code for the project.

[Dissertation Python Script and Data](https://uelac-my.sharepoint.com/:f:/g/personal/u2218896_uel_ac_uk/Eo2kEBmGyOJCp-klq7JgCiABk8RqkHkbZTnezg50T50q_w?e=t0xKnM)