

Using Machine Learning to Analyse Football Teams and Predict the Outcome of a Football Match

Adish Golechha^{1,*}, Akshat Muke²

Abstract

Football, as one of the most popular sports on the planet, has always attracted a large number of fans. Over 150 million men and women of all ages play it in over 200 countries. Modern football has seen a paradigm shift from being just one of the most physical sports to now being one of the most complex sports due to the involvement of multiple new factors such as Home games, away games, form of an individual, total shots taken, total shots blocked, weather conditions etc. In recent years, the importance of these statistics in football and its usage has been seen by all. In this paper, football statistics have been used to predict the outcome of football matches via various machine learning algorithms as well as by analyzing them. A thorough examination of these statistics enables a manager to devise strategies for team management as well as upcoming matches. To calculate accuracy, three Machine Learning algorithms were used: Logistic Regression, Support Vector Machine, and Multinomial Nave Bayes. All three of them were compared at it was observed that the most efficient accuracy was that of SVM which was (61.29%). It was also noted that the accuracy achieved by Logistic Regression (60.9%) was very close to that of SVM

Keywords: Football Match, Machine Learning Algorithms, Support Vector Machine, Logistic Regression, Multinomial Naive Bayes

INTRODUCTION

Sporting activities have always piqued the curiosity of a diverse audience. Football is one of the most popular sports. A match can have three outcomes: home, draw, or loss. Because of the popularity of games and the large number of possible outcomes, predicting game outcomes is a particularly intriguing task.

It becomes difficult to predict the final outcome because the way the team plays on a given day is determined by a variety of factors, including the team's current form, recent team meetings, rivalries, offensive and defensive skills, individual abilities of key players, and even the psychological impact of fans in the stands. Football is a game in which the average number of goals scored is small (two to three each game), which implies that an individual's genius or folly can determine the ultimate outcome. As a result, selecting the features and classification method that will enable prediction is a significant difficulty.

***Author for Correspondence**
Adish Golechha

^{1,2}Student, Department of Electronics and Communication,
Nirma University, Ahmedabad, Gujarat, India

Received Date: April 29, 2022
Accepted Date: May 16, 2022
Published Date: May 20, 2022

Citation: Adish Golechha, Akshat Muke. Using Machine Learning to Analyse Football Teams and Predict the Outcome of a Football Match. Journal of Communication Engineering & Systems 2022; 12(1): 30–40p.

The bookies have their own prediction model and they get it right about 53% of the time.

Football pundit Mark Lawrenson had prediction accuracy of 52.6%. The aim of this project is to build a system to predict the outcome using machine learning algorithms like Support Vector Machine, Naïve Bayes and logistic Regression. The main goal of the paper is to develop a model that can predict the outcome of football matches with a 60% accuracy.

The outcome of a match is determined by a variety of factors, which needs many numbers of tests to identify the optimal subset that has the largest impact on the match's final conclusion. Three classifiers were tested in order to acquire the best possible prediction properties. The final step in the study was to select the classifiers that produced the greatest results in predicting match outcomes.

DATA COLLECTION

For this research, two datasets will be utilized. The first dataset comes from Kaggle's data science community [3]. This dataset is used for exploratory data analysis. However, this dataset includes a few missing data points that are critical for prediction, so the second dataset which will be utilized is from Football Data [4]. This dataset contains more seasons and is being combined with the initial dataset to create a collective set of data spanning 22 football seasons, as well as to develop new features that aid in prediction.

Initially, Exploratory data analysis is conducted on a dataset obtained from the Kaggle web-site. This dataset is in SQLite format and has tables for Country, League, Match, Player, Player Attributes, Team, Team Attributes, and Sequences. It has information on around 25000 matches, 10000 players, and 11 European countries with their Players and Teams qualities derived from EA Sports' FIFA video game series, betting odds from up to 10 providers. The dataset used for prediction contains more minute-by-minute statistics from matches played between 2000 and 2022.

LITERATURE REVIEW

Several efforts have been made to improve the accuracy of football match prediction results. The researchers offered numerous models by utilizing various ML techniques.

Razali et al. [16] created a Bayesian hierarchical model that forecasts football outcomes. Their approach was based on the goals scored by both sides in each encounter.

Min et al. [17] presented a dynamic algorithm for predicting football match outcomes. The FRES system is a dynamic structure that consists of two primary components: equation based theorems and the Bayesian-network element. As a result, the FRES approach is a hybrid of two techniques that work in tandem to forecast the results of football games. Furthermore, the FRES method is introduced in-game time-series approach, which allows for more practical prediction. Nonetheless, for the FRES programmed to be well-controlled, it necessitates a certain level of professional experience.

Constantinou [18] They have created a pi-rating football prediction model to forecast the outcome of football matches, whether home win, draw, or away win, for English premier league games during the 2010 - 2011 seasons, which combines objective and subjective information such as team performance, team form, psychological impact, and muscle soreness.

EXPLORATORY DATA ANALYSIS

Explanatory Data Analysis, or EDA, is a technique used by data scientists to examine and explore data sets and characterize their important aspects, often utilizing data visualization techniques. The main reason why this technique is used is the fact that this technique involves statistics and data visualization to not only analyses but also to identify trends in data sets. Significance of EDA is that it helps to detect anomalies in the data set, variables of data set are better understood by it and it helps to identify errors in dataset. Explanatory data analysis is being used in this case to determine which league is the most diverse and which is best suited for further investigation. This also determines which set of features would go in handy for analysis of a particular league in each season, the total number of goals scored by all teams in all the separate leagues is plotted.

It is observed that the Spanish teams score more than the teams in the Premier League and the Bundesliga combined. However, the total number of goals in the Bundesliga is modest due to the

league's small size (only 18 teams compete). As a result, each Bundesliga team plays only 34 matches every season, whereas each Premier League squad plays 38 matches. So the overall number of games played in the Premier League and La Liga is 380, whereas the total number of games played in the Bundesliga is 306.

Clubs in the Bundesliga score significantly more points per game on average than teams in other leagues. Bundesliga teams averaged more than three goals per game in 2013/2014. (967 goals in 306 matches- 3.16 per match).

- Average goals in the English Premier League is 2.7
- Average goals in the Spanish La Liga is 2.8
- Average goals in the Germany Bundesliga is 2.9
- Entropy is used to determine predictability of the leagues.
- The disorder is measured using entropy, which is calculated using bet 365 odds. The higher the entropy value, the more uncertain the results of the match.

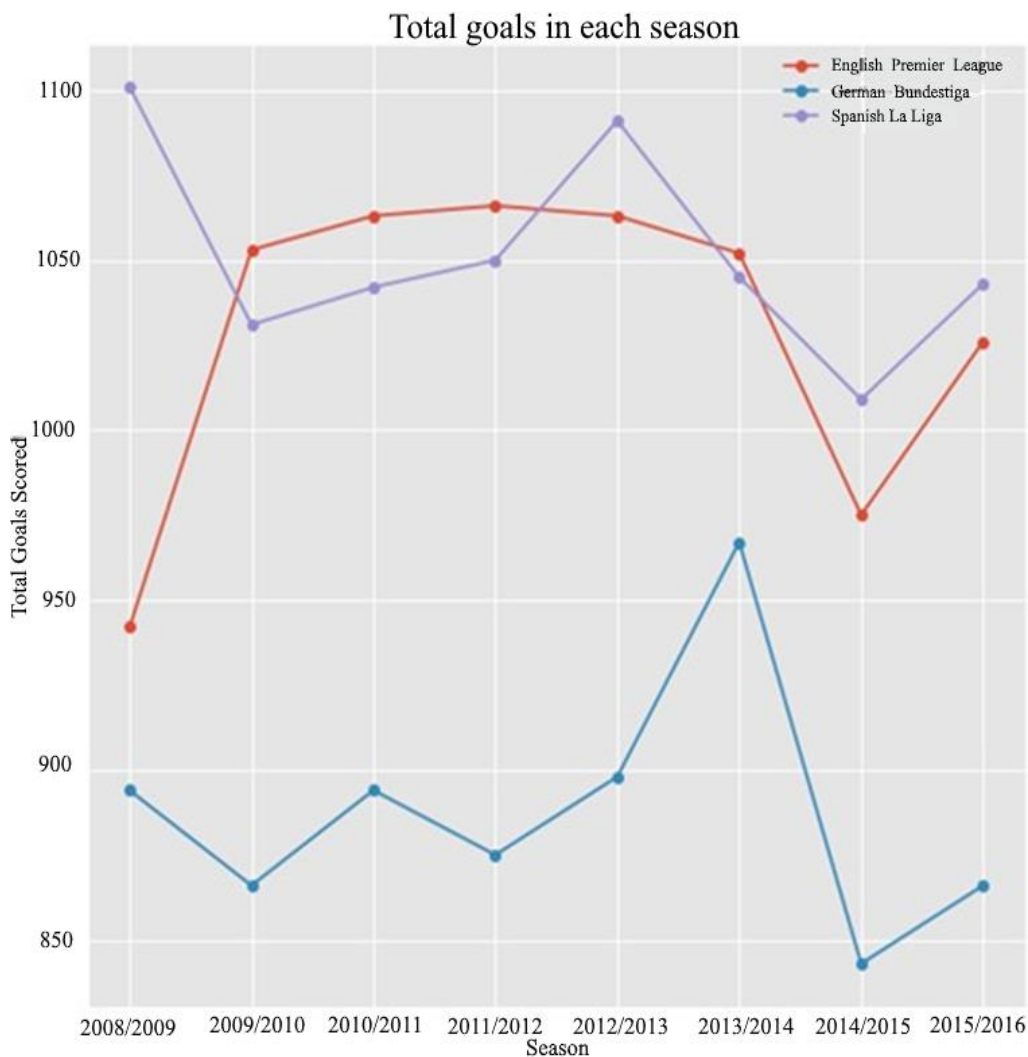


Figure 1. Total goals.

Bayern Munich and Barcelona have won the Bundesliga and La Liga six times each since 2008. Since 2012, the top three teams in La Liga have been the same. Since 2008, the rivalry for the title has always been amongst Barcelona, Real Madrid, and Atletico Madrid, with Real Madrid winning twice and Atletico Madrid winning once.

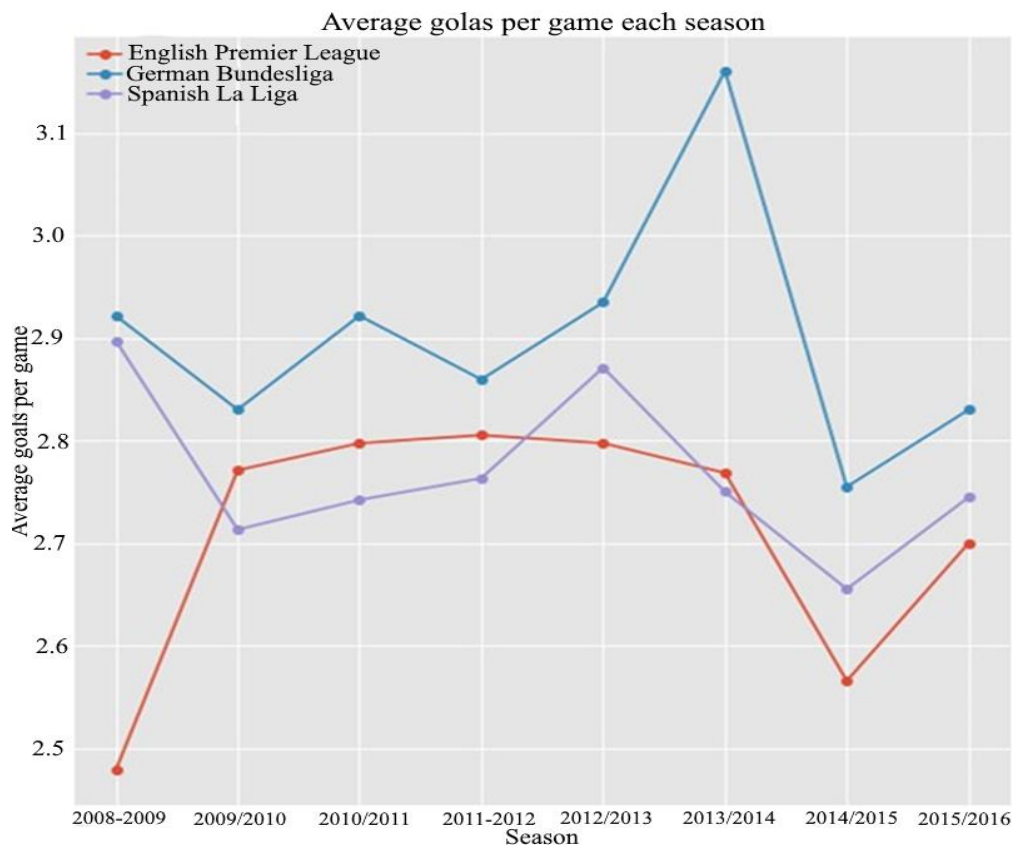


Figure 2. Average goals.

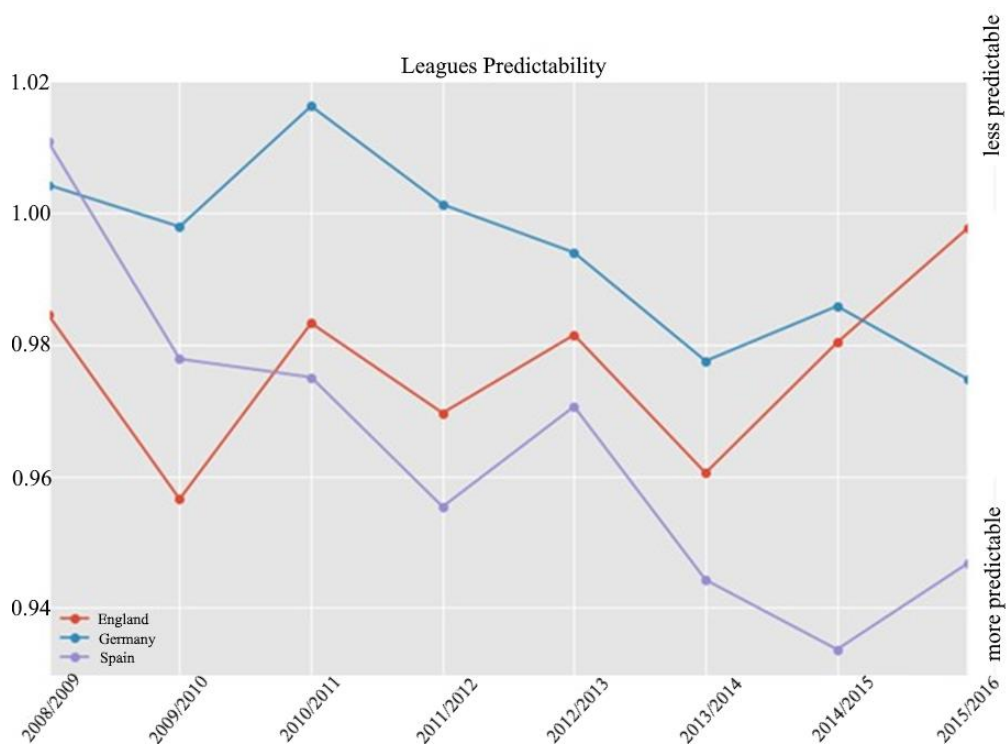


Figure 3. Predictability of leagues.

Bayern Munich has won the Bundesliga five times in a row since the 2012–13 season. Borussia Dortmund also performs well, having won the league twice in 2009–10 and 2010–11.

Since 2008–09, Manchester United has won the Premier League three times (08–09,10–11,12–13), Manchester City has won twice (11–12,13–14), Chelsea has won three times (09–10,14–15,16–17), and Leicester City has won once (15–16). The previous season saw Leicester City promoted to the Premier League, where it finished 14th. Even so, it won the league the following season, demon-starting how impossible it is to predict the Premier League.

The English Premier League is more volatile than the other leagues. As a result, this league will be used for match prediction and analysis. It can be easily observed that the home advantage in the Premier League is dwindling with each passing season. Initially, home matches were a very good element in determining the match's winner. The matches have become extremely competitive, and hence now outcome cannot be anticipated just on the basis of home advantage.

MODEL DESIGN

Feature Selection

A variety of elements influence the eventual outcome of a football match, including the team's current form, home court advantage, overall team quality, individual player quality, and so on. The problem of feature selection can be tackled in two ways.

The first method assumes that there is some prior knowledge about the problem (in this example, football matches) and that those features are selected that are believed to have the greatest influence on the ultimate result. The other approach assumes little to no knowledge of the problem and, as a result, selects all elements that may affect the outcome and then attempts to identify which have the greatest impact progressively. The first technique is adopted in this paper due to the enormous number of features that could affect the outcome of the match and the numerous difficulties in gathering them.

For the purpose of this study, a two-step approach has been followed, first the following features have been selected:

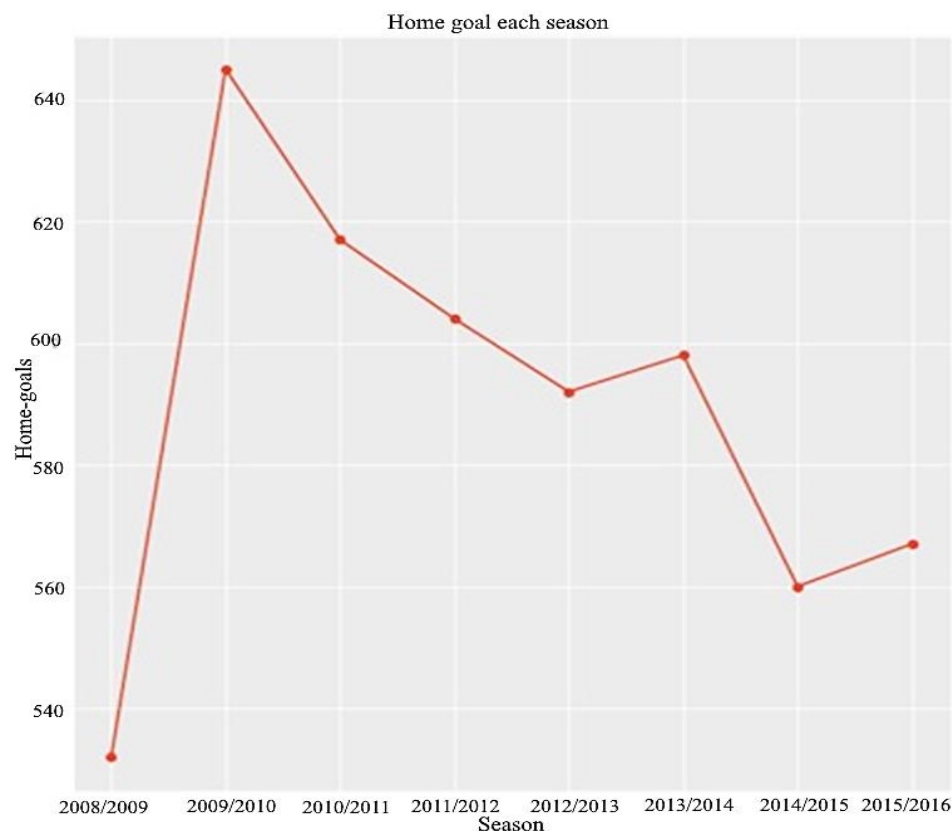


Figure 4. Home goals.

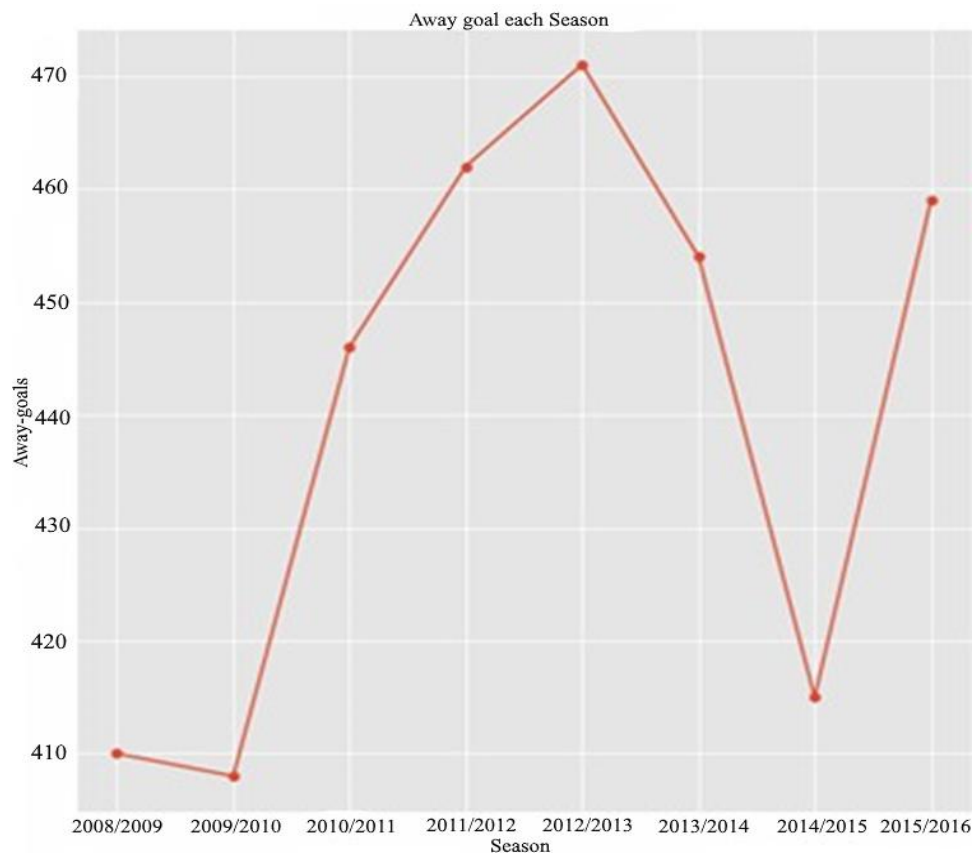


Figure 5. Away goals.

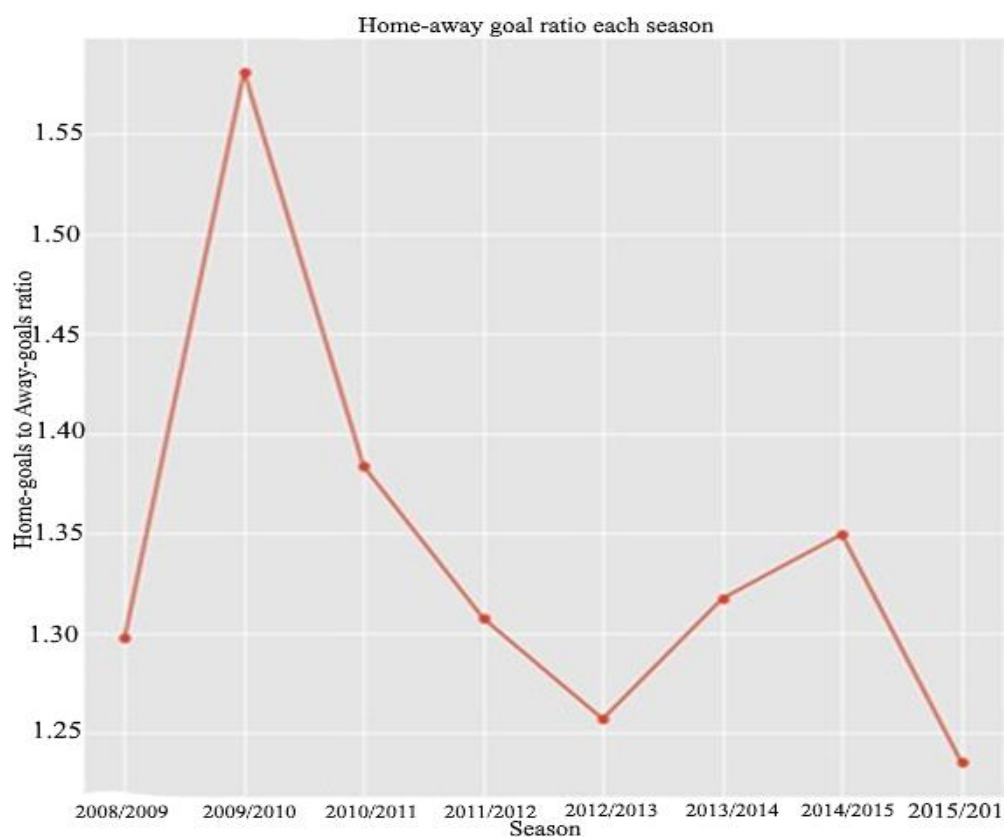


Figure 6. Home-away goal ratio.

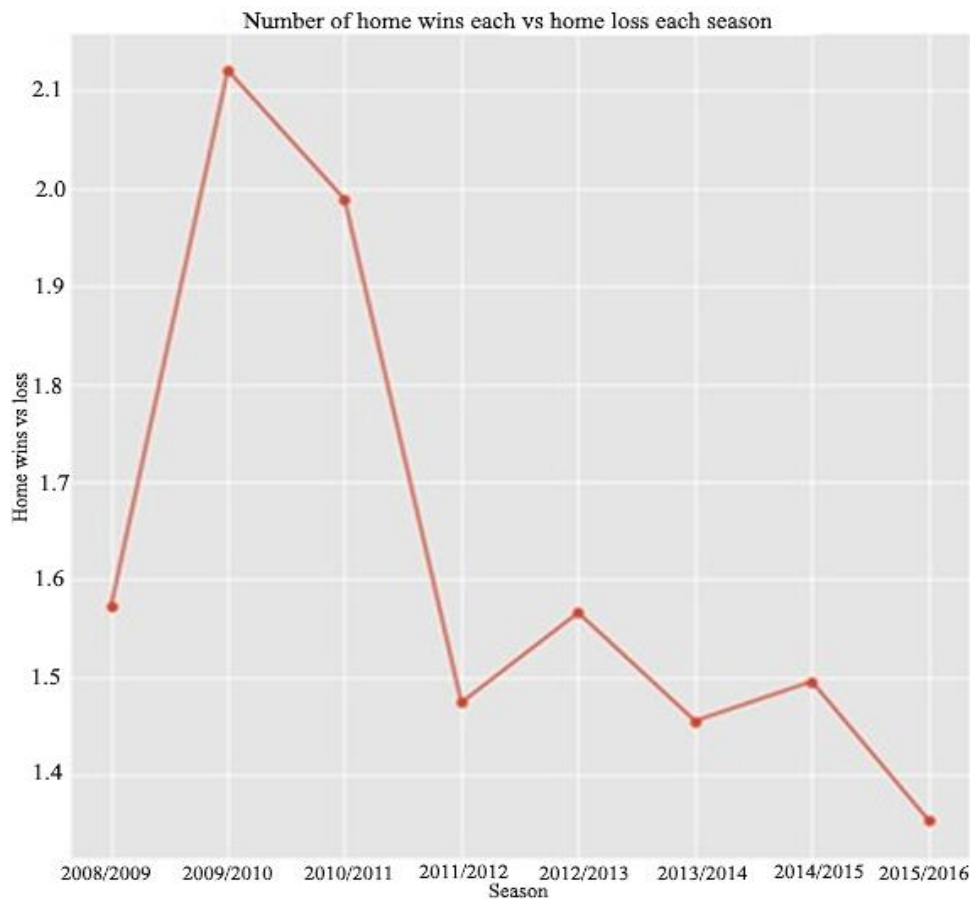


Figure 7. Win-loss ratio at home.

- *FTHG*: Full Time Home Goal
- *FTAG*: Half Time Away Goal
- *FTR*: Full Time Result
- *HS*: Home Shots
- *AS*: Away Shots
- *HF*: Home Team Foul
- *AF*: Away Team Foul
- *HC*: Home Team Corner
- *AC*: Away Team Corner
- *HAS*: Home Attacking Strength
- *AAS*: Away Attacking Strength
- *HDS*: Home Defensive Strength
- *ADS*: Away Defensive Strength

Feature Engineering is used to find the best combination of the above features because a smaller number of well-weighted features has a more balanced impact on the predictions. To reduce the dimensions, eight new features were defined from the above features. Each team's attacking strength is calculated at home and away based on goals scored at home and away compared to the number of matches they have played. Aggregate number of corners, fouls and shots taken by the home and away team have also been calculated. This will be a feature in the proposed prediction model. The above features are selected because they cover all the aspects of a football game except the Player-specific statistics. The prospect of covering Player-specific features and Team Psychological Related features can be explored in the future.

Learning Algorithms

Support Vector Machine

Support vector machines [5] (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

Support vector machines have the following advantages: They are effective in high-dimensional spaces. When the number of dimensions exceeds the number of samples, the method remains effective. It also saves memory by using a subset of training points in the decision function (called support vectors). The decision function can be specified with different Kernel functions. Common kernels are provided, but custom kernels can also be specified. The disadvantages of support vector machines are as follows: if the number of features is much greater than the number of samples, avoid over-fitting when selecting Kernel functions, and regularization term is critical. SVMs do not directly provide probability estimates; instead, they use an expensive five-fold cross-validation procedure. The goal of this classification algorithm is to increase the distance between the spots. As a result, categorizing between points becomes simple, as does classifying classes. As a result, Support Vector Machine (SVM) divides data in a manner it can augment the margins among the features in the model.

Logistic Regression

When the dependent variable is dichotomous, the appropriate regression analysis is logistic regression [6] (binary). The logistic regression, like all regression analyses, is a predictive analysis. Logistic regression is a statistical method for describing and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

The goal of this classification algorithm is to increase the distance between the spots. As a result, categorizing between points becomes simple, as does classifying classes. As a result, Support Vector Machine (SVM) divides data in a manner it can augment the margins among the features in the model.

The dependent variable should be dichotomous in nature, according to the major assumptions of logistic regression (e.g., presence vs. absent). Outliers in the data should be avoided, which can be determined by converting the continuous predictors to standardised scores and removing values less than -3.29 or greater than 3.29 . The predictors should not have high correlations (multicollinearity). A correlation matrix among the predictors can be used to assess this. According to Tabachnick and Fidell (2013), the assumption is met as long as the correlation coefficients between independent variables are less than 0.90 . The task of estimating the log odds of an event is at the heart of the logistic regression analysis. Logistic regression mathematically estimates a multiple linear regression function defined as:

$$\begin{aligned} &\text{logit}(p) \\ &\text{for } i = 1 \dots n. \end{aligned}$$

Multinomial Naive Bayes

Multinomial Naive Bayes algorithm [7] is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm predicts the tag of a text, such as an email or a newspaper article, using the Bayes theorem. It computes the likelihood of each tag for a given sample and outputs the tag with the highest likelihood.

The Naive Bayes classifier is a collection of many algorithms that all share one common principle: each feature being classified is unrelated to any other feature. The presence or absence of one feature has no effect on the presence or absence of another. The benefits of employing Naive Bayes classifiers include the fact that they are extremely scalable when presented with vast amounts of data: they take an approximately linear time to train when adding new features in the future prospects of the model. However, even if the classifier is strong enough to disregard the assumption, the estimated probabilities are known to be slightly wrong.

Implementation

The software system was written in Python and made use of libraries such as numpy, matplotlib, seaborn, pandas, sqlite, scipy, and sklearn. Metrics, preprocessing, naive bayes, model selection, linear models, svc, and other techniques in sklearn were also employed. Initially, only the first four variables were used to run, and the additional features were saved for the second iteration. This estimator, as seen on Sklearn, scales and translates each feature independently so that it falls within the training set's given range, e.g. between zero and one.

RESULT

Performance Evaluation

K-fold cross validation score will be used to measure how accurately the model is predicting.

To check how much of an impact the engineered features have on the model prediction, initially only the four primary features that included the goal statistics were used.

Later after adding the features of shots, corners and fouls it was seen that there was significant increase in accuracy and Support Vector Machine model performing better after adding the features.

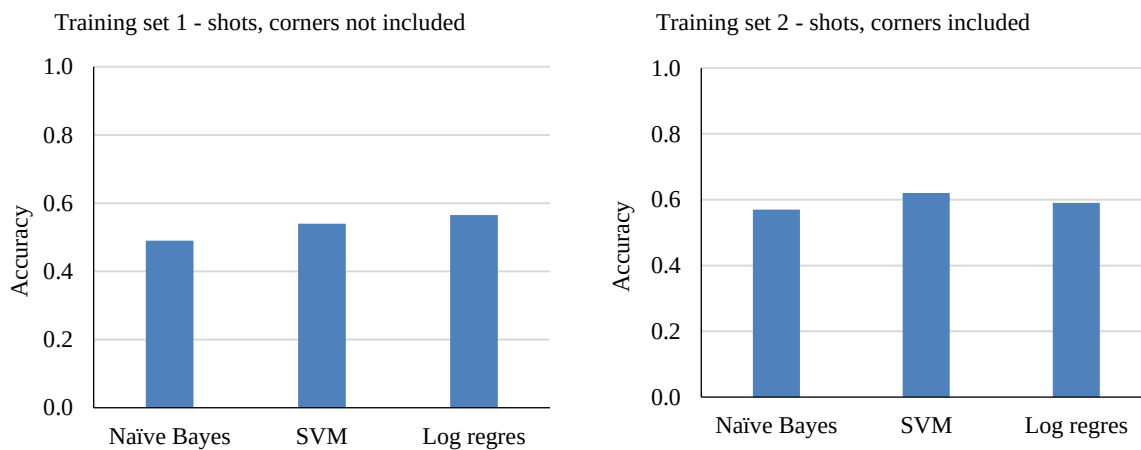


Figure 8. Cross-validation scores.

Features were also normalized so as to check whether it improves the outcome.

After normalization, it was seen that the performance of Naive Bayes and SVM decreases, and the performance of Logistics regression increases. When findings are examined in terms of accuracy, it is clear that the basic set (one without normalization) outperforms the created normalized set. As a result, it can be concluded that normalization works better for some models while others predict better results without it.

It's strange to see logistic regression do significantly better than the other models after normalization, mostly because, as seen, the data obviously did not have trends and it was a complex problem, and strangely, a very complicated problem was solved by a "linear" technique, possibly the simplest here. One reason for this could be that the features of each model were not investigated in this technique. Because SVM contains a large number of kernels, experimenting with different kernels may be a good way to improve this model's performance.

From these results it is observed that how much of an impact, feature and classifier selection has on the prediction capabilities.

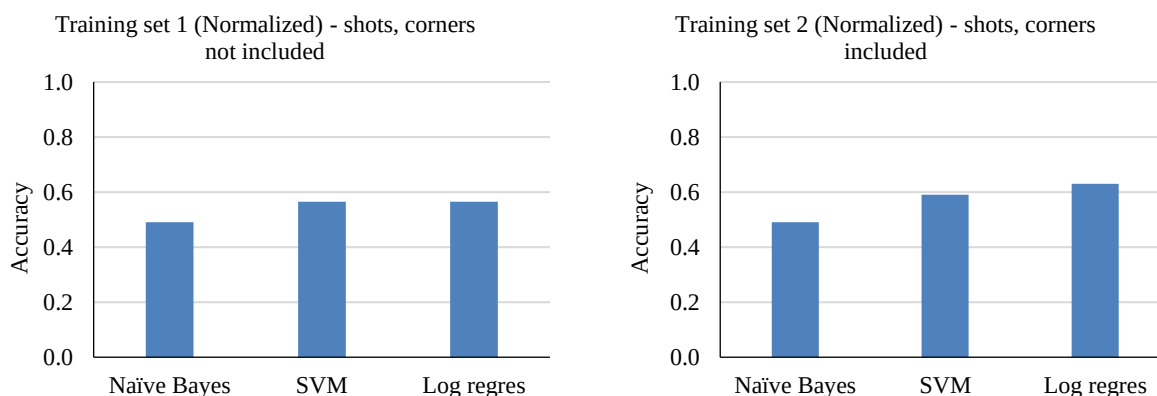


Figure 9. Cross-validation scores after normalization.

FINAL ASSESSMENT

In general, the outcomes obtained can be deemed satisfactory, especially when artificial neural networks are used. The results are much better than the reference technique (classification in the most common class). Even if direct and objective comparison with the results of similar articles is not possible, it can be demonstrated that the approach used delivers similar or slightly better results than those of other articles in the best-case scenario.

CONCLUSION

In this research paper, most suitable machine learning model has been proposed for analysis of football teams as well as match prediction. It was discovered that, of the three proposed models, Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes, SVM provides the highest accuracy when calculated without normalization. After normalization, improvement of accuracy can be seen in case of Logistic regression. The betting accuracy of bookmakers was roughly 53%, whereas the model generated with SVM had an accuracy of 61.29%. It can also be noted that when important factors such as corners and shots are included in the features, better output is observed than the case when they are not used. If more features could be added, such as team or player forms, rivalries between the clubs, injuries in both of the teams and fan mood, the accuracy would improve even more. Twitter analysis can also be used to make predictions. Also, alternative models, such as xgboost, can also be used with the same properties to enhance accuracy.

ACKNOWLEDGMENT

The work presented in this paper is completely ours and all the credit for any references is diligently given to proper sources from which they were referred. We would like to thank Professor Manish Patel for their guidance and wisdom that they provided us with throughout this research paper.

REFERENCES

1. Klaus Rosmanitz, European Football-Soccer, English Online AT, Accessed on: April 19, 2022. [Online]. Available: <https://www.english-online.at/sports/soccer/european-football.htm>
2. Frank Keogh, Football betting-the global gambling industry worth billions, BBC, 3rd October, 2013, Accessed on: April 19, 2022. [Online]. Available: <https://www.bbc.com/sport/football/24354124>
3. Hugo Mathien, European Soccer Database, Kaggle, October 2016, Accessed on: April 19, 2022. [Online]. Available: <https://www.kaggle.com/datasets/hugomathien/soccer>
4. Football Data Co. UK, Historical Data in England Football, Accessed on: April 19, 2022. [Online]. Available: <https://www.football-data.co.uk/>
5. Sci-kit Learn, Support Vector Machines, Accessed on: April 19, 2022. [Online]. Available: <http://scikit-learn.org/stable/modules/svm.html>
6. Statistics Solutions, What is Logistic Regression? Accessed on: April 19, 2022. [Online]. Available: <https://www.statisticssolutions.com/what-is-logistic-regression/>

- <http://www.statisticssolutions.com/what-is-logistic-regression/>
7. Sci-kit Learn, Naive Bayes, Accessed on: April 19, 2022. [Online]. Available: <http://scikit-learn.org/stable/modules/naive-bayes.html>
 8. Stylianos Kampakis, Andreas Adamides, Using Twitter to predict foot- ball outcomes, Accessed on: April 19, 2022. [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1411/1411.1243.pdf>
 9. Matheus Kempa, Machine Learning Algorithms for Football Predictions, Medium, Accessed on: April 19, 2022. [Online]. Available: <https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8>
 - [10]Md. Ashiqur Rahman, A deep learning framework for football match pre- diction, SpringerLink, Accessed on: April 19, 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s42452-019-1821-5>
 11. Mark Lawrenson, How do handicaps predict outcomes?, Pinnacle, September 12 2013, Accessed on: April 19, 2022. [Online]. Available: <https://www.pinnacle.com/en/betting-articles/Soccer/Mark-Lawrenson-vs-Pinnacle-Sports/VGJ296E4BSYNURUB>
 12. Siraj Raval, Predicting Winning Teams, GitHub, Accessed on: April 19, 2022. [Online]. Available: <https://github.com/llSourcecell/Predicting-Winning-Teams/blob/master/Prediction.ipynb>
 13. Yoni Lev, The Most Predictable League, Kaggle, Accessed on: April 19, 2022. [Online]. Avail- able: <https://www.kaggle.com/code/yonilev/the-most-predictable-league/notebook>
 14. N. Razali, A. Mustapha, F. M. Clemente, M. F. Ahmad, M. A. Salamat. Pattern Analysis of Goals Scored in Malaysia Super League 2015, Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, no. 2, pp. 718–724, 2018.
 15. B. Min, J. Kim, C. Choe, H. Eom, R. B. McKay. A compound framework for sports results prediction: A football case study, Knowledge-Based Systems, vol. 21, no. 7, pp.551–562, 2008.
 16. C. Anthony, N. Fenton, and M. Neil. Pi-football: A Bayesian network model for forecasting Association Football match outcomes, Knowledge-Based Systems, vol. 36, pp. 322-339, 2012.