Crowd Sourced Football Fortune: Unveiling Epl Match Winners With Machine Learning

Vengadaprasath .V.M[1], Dr. S. Manju[2]
[1]Dept of Computer Applications (MCA)
[2]Associate Professor, Dept of Computer Applications (MCA)

Abstract- *Machine Learning (ML) is a crucial area of intelligent methods that has shown a lot of promise, particularly in the areas of categorization and prediction precision. Sports forecasting has become a key application for ML-driven predictive modeling, with significant financial ramifications. Football is the most popular sport and is practiced in more than 190 countries, making it unique among other sports due to its intricacy and dynamic nature. In this study, we explore the use of ML approaches to forecast the English Premier League (EPL) champion. Our primary objective is to forecast the full-time result (FTR) of football matches, a determinant of the winning team. We leverage a range of algorithms including Support Vector Machines, Random Forest, and Naïve Bayes to train on historical data, ultimately selecting the model with the highest accuracy for predicting EPL match outcomes. The dataset utilized spans multiple seasons, sourced from [6], offering valuable insights into the intriguing domain of "Winning Team Prediction using Machine Learning" in the context of football.*

Keywords- Machine learning, soccer game, prediction, classification, accuracy.

I. INTRODUCTION

Accurately predicting the outcomes of soccer seasons has evolved into a high-stakes industry, with millions of dollars at play and a dedicated following of gamblers and passionate fans alike seeking more precise forecasts and probabilities. This growing interest in soccer match predictions extends from managerial teams, who use data analytics to analyze squad performance and enhance game strategies, to devoted fans eager to anticipate their favorite team's results. The strategic analytics employed in the sport has become a critical determinant of a team's success. Soccer prediction has emerged as an intriguing research challenge due to the multitude of factors influencing match outcomes, including home and away goals, team rankings, match conditions (day or night), teamwork, player skills, home-field advantage, and weather conditions. The primary aim of this project is to provide an accurate dataset for soccer matches and generate efficient predictions for upcoming matches, particularly by focusing on the Full- Time Result (FTR) as our class label, denoting outcomes as Home, Away, or Draw.

II. RELATED WORK

Numerous studies and surveys have explored the domain of sports prediction, particularly in the context of forecasting winning teams. An initial study by [1] focused on predicting winning teams in NBA matches. They employed algorithms including Linear Regression, Maximum Likelihood Classifier, and Multilayer Perceptron (Back Propagation) to achieve varying prediction accuracies. Linear Regression, in particular, outperformed others with a 68% prediction rate. However, this model called for a larger dataset and feature classification for further improvements

In a similar vein, [2] tackled the prediction of winning teams in NBA matches, using a dataset spanning 32 seasons, enriched with over 50 features. Feature selection played a crucial role in this study. Various machine learning algorithms, such as Extra Trees, Gradient Boosting, k-nearest neighbors, Logistic Regression, SVM, Neural Networks (MLP algorithm), and Non-Linear SVM, were implemented. SVM emerged as the most accurate, achieving a 71% accuracy rate. Notably, Gradient Boosting demonstrated better classification than Random Forest or SVM.

Regarding soccer, [3] investigated match outcome prediction in the English Premier League (EPL) for the 2015–16 season using Support Vector Machine (SVM) and libsvm. Although a 64.7% accuracy rate was achieved, the model had trouble predicting games later in the season and had trouble working with big datasets. The authors proposed that including more information from prior seasons as well as more statistics would improve accuracy. [4] created a model to forecast soccer match outcomes in the Barclays' English Premier League utilizing factors including away team goals, venue, scores, and home team data. Support Vector Machines (SVM), XGBoost, and Logistic Regression were used as machine learning classifiers. To improve model accuracy, future studies may use sentiment analysis, player metrics for each individual, and social media fan posts.

Additionally, [5] proposed a logistic regression model for estimating 2015/16 Barclays Premier League match results with an accuracy of approximately 71%. Their dataset focused on just four variables: Away Offense, Home Defense, Home Offense, and Away Defense. Despite the limited variables, this system demonstrated strong predictive accuracy, offering insights into match outcomes, odds, and regression coefficients These related works contribute to the ongoing exploration of sports prediction, offering insights and opportunities for improving the accuracy of predictions in the field.

Comparison of Alogrithms:

SR. No	ALGORITHMS	ACCURACY	MERITS	DEMERITS
1	Linear Regression Maximum Likelihood, Back Propagation	67.89%	Linear Regression is easy to implement and training can be done in short duration	Limited Dataset . Feature classification wasn't done
2	Support Vector Machine,KNN, Gradient Boosting , Neural networking	71.00%	Dataset had enough instances. Feature Classification done	Lack of complexity of algorithm used
3	Support vector Machine Libvism ,NN(sklearn implementation)	67.00%	Risk of over-fitting is less in SVM	poor-accuracy in testing large dataset
4	Support Vector Machine XGBoost, Logistic Regression	68.55	XGBoost - scalable and accurate	Slower training XGBoost
5	Logistic Regression	69.00%	Easy to Implement	Complex Algorithm, Overfitting is more

III. PROPOSED APPROACH

This section introduces the model. In this work, a predictive model for predicting football game results in the English Premier League is introduced. A variety of machine learning classifiers are used to train the model utilizing historical data spanning numerous seasons. These classifiers will be thoroughly assessed in order to determine which one has the best prediction accuracy. The model's accuracy in forecasting match results will then be further improved through optimization efforts. Home Win (H), Away Win (A), and Draw (D) are included in the predictive labels.

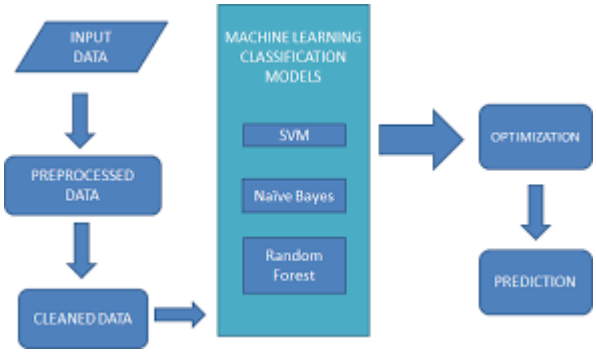

Figure 1 . Framework Diagram

a. Dataset Overview:

Our predictive analysis will rely on historical data drawn from recent seasons' matches. We have sourced an extensive dataset from [6], encompassing a wealth of information, ranging from past games to the most current ones. This dataset is rich in detail, comprising approximately 67 attributes per season, including key elements such as Home team, Away team, scores, and venue, among others.

To streamline our analysis, we have narrowed down our focus to a select set of 7 to 12 attributes that will serve as the primary predictors for match outcomes. The dataset encompasses a substantial sample size of 4000 instances.

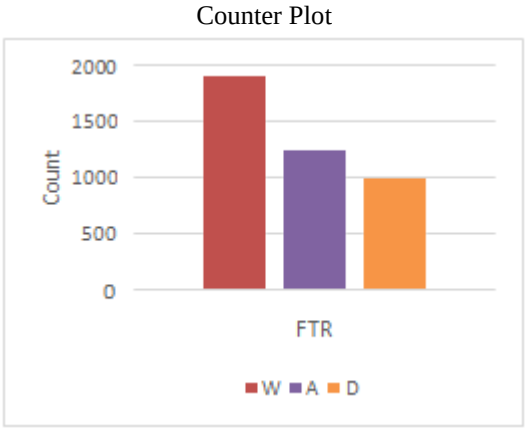

Figure 2. Counter Plot

b. Preprocessing of Data:

Numerous attributes for each season are included in the dataset that was obtained, some of which are useless or less important for outcome prediction. As a result, a data cleaning procedure has been used to save only the attributes that are pertinent to our prediction goals.

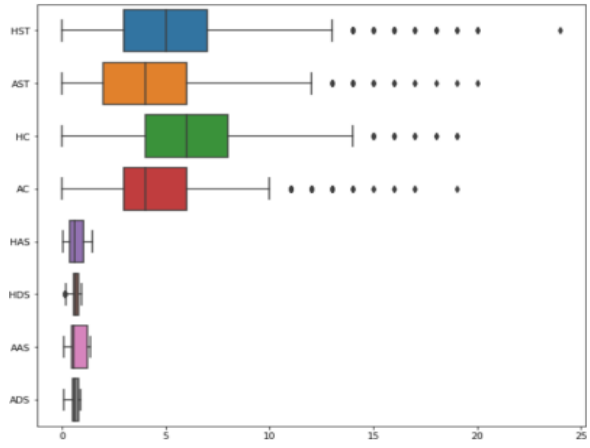

Figure 3. Boxer Plot

c. Data Division:

To fit the model and obtain the desired result, we divided the data into training and testing portions. They were divided into two groups, 80 and 20 percent each.

d. Modelling:

We have incorporated the Naïve Bayes Classifiers, Random Forest, and Support Vector Machine (SVM) algorithms into this system

Naïve Bayes Classifiers: The Naïve Bayes Classifier is a powerful probabilistic model frequently employed in various classification tasks. It leverages Bayes' Theorem, a fundamental principle in probability theory, to make predictions based on available data. Specifically, it calculates the probability of a certain event or hypothesis (e.g., event A) occurring given the occurrence of another event or evidence (e.g., event B). This probability estimation is crucial for classifying data points into predefined categories or classes. What distinguishes the Naïve Bayes Classifier is its "naïve" assumption of feature independence. It assumes that each feature used for prediction is unrelated to all other features, simplifying the complex task of probability estimation. While this independence assumption may not always hold true in real-world scenarios, the Naïve Bayes Classifier often performs surprisingly well, particularly when applied to text classification tasks, spam email detection, and sentiment analysis. In the context of your dataset, where you are predicting the Full-Time Result (FTR) of football matches using various parameters [7], the Naïve Bayes Classifier would estimate the probability of each possible outcome (e.g., Home Win, Away Win, or Draw) based on the values of these parameters. This probabilistic approach allows you to make informed predictions about the likely outcome of each match in the English Premier League.

Random Forest :Random Forest, a robust machine learning ensemble method, is composed of a large number of individual decision trees that operate collectively to enhance predictive accuracy. Ensembles, by definition, harness the strength of multiple learning algorithms to achieve superior predictive performance compared to any single algorithm operating independently. Within the Random Forest framework, each individual decision tree independently produces a class prediction, and the ultimate model prediction is determined by selecting the class with the highest vote count [8]. This approach not only improves predictive accuracy but also enhances the model's resilience against overfitting, making it particularly valuable in various classification and regression tasks.

Support Vector Machine (SVM): Support Vector Machines (SVMs) represent a class of Machine Learning models employed for both regression analysis and classification tasks. These models are categorized under supervised learning within the realm of Machine Learning and find extensive application in classification problems. At their core, Support Vector Machines operate on the concept of identifying the optimal hyperplane that effectively partitions a given dataset into two distinct sections [9]. This hyperplane, also known as the decision boundary, is strategically positioned to maximize the margin between data points of different classes, thus enhancing SVM's capacity to make precise predictions, especially in scenarios with complex data distributions.

IV. EXPERIMENT

We experimented using data from recent English Premier League seasons to determine the maximum level of prediction accuracy. The main goal was to determine whether the amount of training data has an impact on prediction accuracy. As shown in Tables 4 and 5, we separated the dataset into separate training and testing sections.

Eight important qualities, including home attacking strength (HAS), home defensive strength (HDS), away attacking strength (AAS), away defensive strength (ADS), as well as home and away corners and shots on target, are included in the training dataset, which makes up 80% of the overall dataset. We may assess each team's relative attacking strength using this set of characteristics.The testing dataset, on the other hand, makes up 20% of the total dataset and contains the same eight characteristics: home attacking strength (HAS), home defensive strength (HDS), away attacking strength (AAS), away defensive strength (ADS), home shots on target (HST), away shots total (AST), home corners (HC), and away corners (AC). These testing characteristics make it easier to evaluate the efficacy and performance of models.

Data Visualization:

The employment of a heatmap is one of the extensively used techniques for visualizing the dataset [12]. The use of heatmaps enables one to better understand the correlations present in the dataset. They use a color- coding scheme to indicate different values, allowing us to determine how strongly two traits are related.

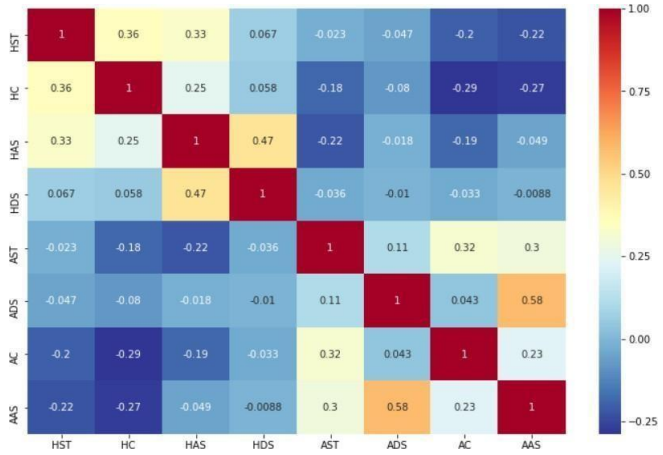

Figure 4 Visualisation using heatmap

We applied feature scaling to our dataset and created a heatmap to clarify the interrelationships between the features. Away Defensive Strength (ADS) and Away Attacking Strength (AAS), in that order, are the characteristics most strongly connected with Home Shots on Target (HST), according to the heatmap's color-coding. The heatmap then emphasizes how home attacking strength (HAS) and home defensive strength (HDS) relate to HST.

The accuracy of the Nave Bayes and Random Forest models was 56% and 60%, respectively, while the accuracy of the Support Vector Machine (SVM) was 64.7%. These findings show that SVM performed better in terms of accuracy than the other models.

In a different experiment, the accuracy of the Support Vector Machine was 6%, which was much better than the accuracy of Random Forest (58%) and Naive Bayes (60%). We used the GridSearchCV function from the sklearn toolbox [13] to adjust the model parameters and improve accuracy. These results confirm that SVM routinely outperforms competing models in terms of accuracy.

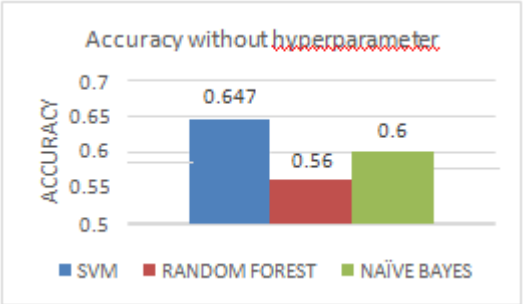

Figure 5.

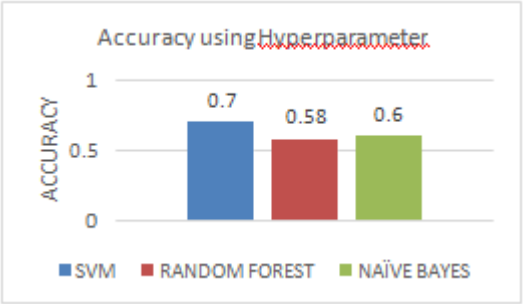

Figure 6.

V. RESULT ANALYSIS:

Our investigation aims to forecast the result of the game, and when combined with training data, the SVM algorithm provided the most accurate results. So, during the preprocessing stage, we normalized the dataset. The goal of normalization is to scale all of the features in the training dataset to the same value. The purpose of normalization is to change the values of the dataset's numeric columns to a standard scale without distorting the variations in the value ranges [11]. Additionally, SVM hyperparameter tuning was conducted because it had already demonstrated the highest accuracy.

The hyperparameter optimization ultimately resulted in an improvement in accuracy for the dataset for SVM, which is 64.7%.

VI. CONCLUSION

The goal of this project is to create a categorization model that may be used to forecast the results of English Premier League (EPL) games. We gained insights into the crucial roles played by important variables, notably the attacking and defensive strengths of the home and away teams, through data visualization. But it soon became clear that relying alone on these four characteristics wouldn't be enough to make reliable forecasts. Our research showed that, when compared to historical season data, contemporary season data maintains greater importance and influence for forecast

purposes. Utilizing current data in this situation is crucial for improving our model's capacity for prediction. Additionally, we found that the addition of extra feature attributes, including corners and shots on target, considerably increases the general accuracy of our forecasts. These factors add important context to the analysis, allowing for a more accurate assessment of match outcomes.

REFERENCES

[1] Renator Amorim Torres ”Prediction of nba games based on machine learning methods”. University of Wisconsin, Madison, 2013.

[2] Weronika Swiechowicz , Jacob Perricone, Ian Shaw “Sports Data Mining: Predicting Results for Professional Basketball Games”, Stanford University,CA,CS229 Autumn 2016.

[3] Steffen Smolka ,”Beating the bookies :Predicting the outcome of soccer games”, Stanford University,CA,CS229 Autumn 2017

[4] 1.Anand Ganesan, 2.Harini M , 1Student, 2Assistant Professor, “ENGLISH FOOTBALL PREDICTION USING MACHINE LEARNING CLASSIFIERS “, International Journal of Pure and Applied Mathematics, Volume 118 No. 22 2018, 533-536,SRM UNIVERSITY 2018

[5] Darwin Prasetio , Dra Harlili Predicting football match results with logistic regression, International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) 2016.

[6] football-data . (2023). data | football- data, [online] Available at: http://www.football-data.co.uk/ [Accessed on 7 Aug. 2023].

[7] Rahman, M. M., Faruque Shamim, M. O., & Ismail, S. An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach. International Conference on Innovations in Science, Engineering and Technology (ICISET) 2018

[8] Oughali, M. S., Bahloul, M., & El Rahman, S. A. Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models. International Conference on Computer and InformationSciences(ICCIS) 2019

[9] Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms. 4th International Conference on Electrical Engineering and Information & Communication Technology(iCEEiCT) 2018

[10] Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms. 4th International Conference on Electrical Engineering and Information & Communication Technology(iCEEiCT) 2018

[11] statisticshowto.datasciencecentral. (2015).Normalized | statisticshowto.datasciencecentral. [online] Available at : https://www.statisticshowto.datasciencecentral.com/normalized/ [Accessed on 10 Aug. 2022].

[12] likegeeks. (2019). Seaborn-heatmap- tutorial | likegeeks. [online] Available at: https://likegeeks.com/seaborn-heatmap- tutorial/ [Accessed on 15 Jan 2020].

[13] Towardsdatascience. [online] Hyperparameteer-Turing-c5619e7e6624 | towardsdatascience. [online] Available at https://towardsdatascience.com/hyperpara meter-tuning-c5619e7e6624