IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Football Match Results using Machine Learning

1Shubham Patil, 2Abhishek Kate, 3Kaustubh Wavare, 4Madhavi Gujar, 5Gayatri Bachav

1Student, 2Student, 3Student, 4Student, 5Professor

1Mumbai University,

2Mumbai University,

3Mumbai University,

4Mumbai University,

5Mumbai University

Abstract- Analysing statistics of football teams can help clubs predict their performance over a particular time frame. In this paper we use various machine learning algorithms to predict results of Premier League season 2021- 2022 for home/away win or draw and analyse the important attributes that impact the full-time result. Games routinely gather information on how the player has the play. The knowledge Is fed into an algorithm which is used by humans to pull games from its predictions of what players would see. Predictions help the manager of the squad to take the next step. By spotting weaknesses at the fighting team's defensive strategy, the weakness of a specific playeror selecting the statistically most possible reaction to the move from past history, coaches might get an edge over their competition. We have done a comparative study between different machine learning algorithms and used the algorithm with the highest accuracy for our project.

Key Words: Random forest ,Machine Learning etc.

INTRODUCTION

The aim of this project is to predict the outcome of football matches using the Random Forest algorithm. The objective of this project is to develop a model that can accurately predict the outcome of football matches based on historical data. Football is the world's most popular sport, with billions of fans and followers worldwide. Predicting the outcome of football matches has always been a challenging task due to the unpredictable nature of the game. However, with the advancement of technology and machine learning algorithms, it has become possible to predict the outcome of football matches with a high degree of accuracy. In recent years, new types of data have been collected for many games in various countries, such as play-by-play data including information on each shot or pass made in a match. In particular, the betting market has

grown very rapidly in the last decade, thanks to increased coverage of live football matches as well as higher accessibility to betting websites thanks to the development of mobile and tablet devices. Indeed, the football betting industry is today estimated to be worth between 300 million and 450 million pounds a year.

Prediction is very useful in helping club staff make the right decision regarding training and player management, it also helps the teams to prepare for their future play based on other team's performance. Premier League - the English Premier association is regarded by some to be the most fun part of football on this planet and its sort of difficult to contend against that. Some of reality's top clubs compete there and when it comes to the businesses needed, it's somewhat tough to tell they aren't in the top of the list. Manchester United are apparently thought to take this biggest family, but alongside them you've had the likes of Liverpool, Manchester City, Chelsea and some more. It's hard to quantify the human truth about predicting the outcome of football matches. Results vary according to what matches are anticipated. Predictions on various leagues and tournaments make several accuracies and humans forecast the result on a much smaller collection of leagues andtournaments than this system. This makes the system hard to equate with human reality. We have developed machine learning models in order to predict full time results of the Premier League table of the year 2021-2022. Our work predicts which team will win the match(home/away/draw).

PREDICTION

The collection of this data has placed Data Science on the forefront of the football industry with many possible uses and applications:

1. Match strategy, tactics, and analysis
2. Identifying players' playing styles
3. Player acquisition, player valuation, and team spending
4. Training regimens and focus
5. Injury prediction and prevention using test results and workloads
6. Performance management and prediction
7. Match outcome and league table prediction
8. Tournament design and scheduling
9. Betting odds calculation

Football in particular is an interesting example as matches have fixed length (as opposed to racket sports suchas tennis, where the game is played until a player wins). It also possesses a single type of scoring event: goals (as opposed to a sport like rugby where different events score a different number of points) that can happen an infinite amount of times during a match, and which are all worth 1 point.

AIM AND OBJECTIVES

Aim :

This project aims to extend the state of the art by combining two popular and modern prediction methods, namely an expected goals model as well as attacking and defensive team ratings. This has become possible thanks to the large amount of data that is now being recorded in football matches.

Different Machine Learning models will be tested and different model designs and hypotheses will be explored in order to maximize the predictive performance of the model

Objectives :

Time consumption is less for predicting results. To provide an appropriate winning result of goals achieved. To fetch previous data results for predicting the current results.

LITERATURE SURVEY

1. “A survey on football match result prediction using machine learning techniques" by Santhosh Kumar and Sumanth V. S., published in the International Journal of Advanced Computer Science and Applications in 2020, provides a comprehensive overview of the various machine learning techniques used for predicting football match results. The survey covers both traditional machine learning algorithms, such as decision trees and logistic regression, as well as more advanced techniques, such as neural networks and support vector machines.
2. “Machine learning for predicting football results: a systematic review" by Fabio Calefato et al., published in the Journal of Sports Sciences in 2021, is a systematic review that analyzes the results of 50 studies on machine learning models for predicting football match results. The review covers a wide range of machine learning algorithms, including Bayesian networks, random forests, and deep learning models. The study also discusses the various features used in these models, such as player data, match statistics, and weather conditions.
3. "Predicting football match results using machine learning: a review of current research" by Ryan Murphy et al., published in the Journal of Sports Analytics in 2021, provides a detailed review of the current research on using machine learning to predict football match results. The study covers both supervised and unsupervised learning techniques and discusses the various factors that can influence match outcomes, such as team form, player injuries, and home advantage.

EXISTING SYSTEM

There are several existing systems that use machine learning, specifically the random forest algorithm, to predict the results of football matches. These systems typically take into account various factors, such as team form, player injuries, and head-to-head record, to generate predictions for upcoming matches.

One example of such a system is the Football-Data.co.uk website, which provides predictions for a range of football leagues using a random forest algorithm. The system takes into account factors such as team form, recent results, and historical performance, to generate predictions for upcoming matches.

The system takes into account a range of factors that can impact the final result of a match, such as team form, recent results, and historical performance, to generate predictions for upcoming matches in a variety of football leagues. The research used algorithms like Random Forest which are some of the algorithms that would be used in this research also

PROPOSED SYSTEM

A proposed system for predicting football match results using the random forest algorithm could involve the following steps:

Data Collection: The first step would be to collect relevant data, including historical match results, team and player statistics, and other relevant information such as weather conditions, venue, and team news.

Data Preprocessing: Once the data is collected, it needs to be cleaned and processed to ensure that it is in a suitable format for analysis. This step may involve removing duplicates, dealing with missing data, and transforming variables into numerical formats.

Feature Selection: Next, the most relevant features that impact the outcome of football matches need to be identified. This could be done using techniques such as correlation analysis or feature importance measures.

Model Building: Once the relevant features are identified, a random forest model can be built using the training data. The model can be trained using historical match results and associated features.

Model Evaluation: The model's performance can be evaluated using metrics such as accuracy, precision, recall, and F1-score. The model can be tested using a validation set of data to ensure that it performs well on unseen data.

Deployment: Once the model has been built and tested, it can be deployed to generate predictions for upcoming football matches. The system can provide predictions for various markets, such as the match result, over/under goals, and correct score.

Continuous Improvement: To ensure that the model remains accurate, the system can be continually updated with new data and refined using techniques such as hyperparameter tuning.

FLOWCHART

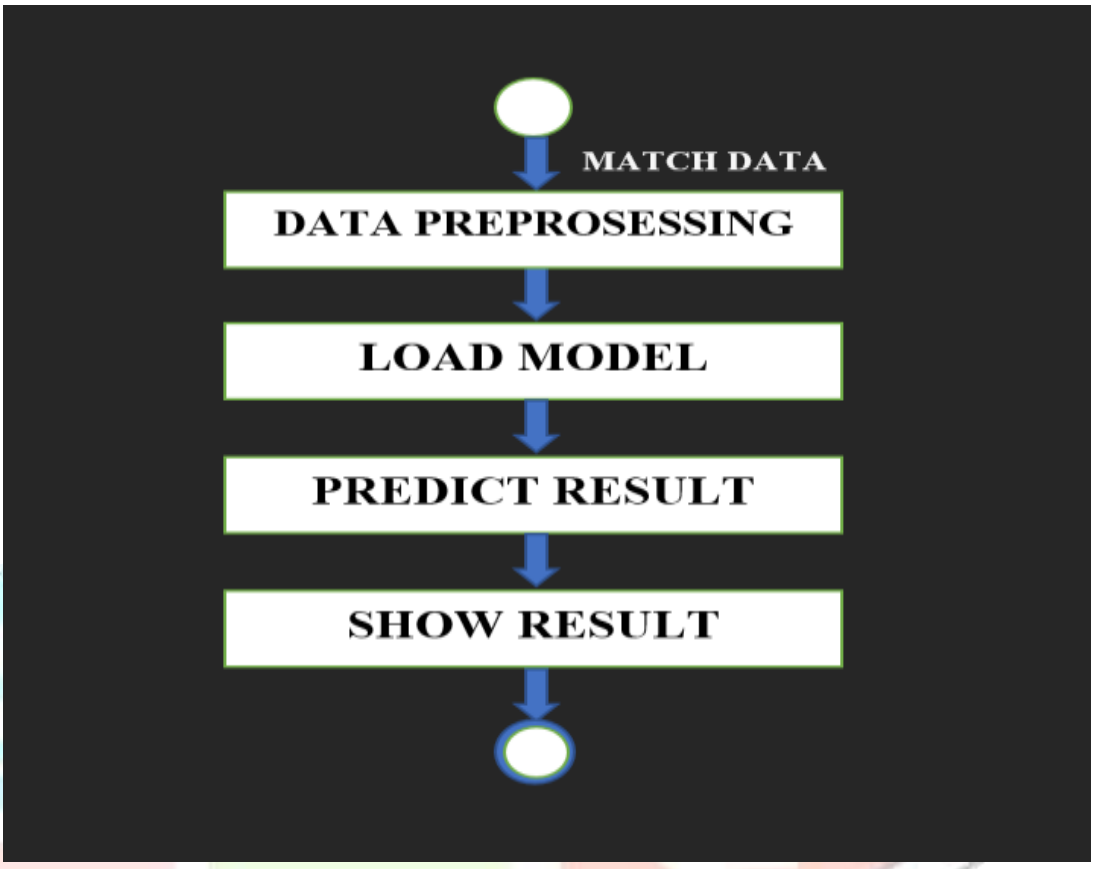
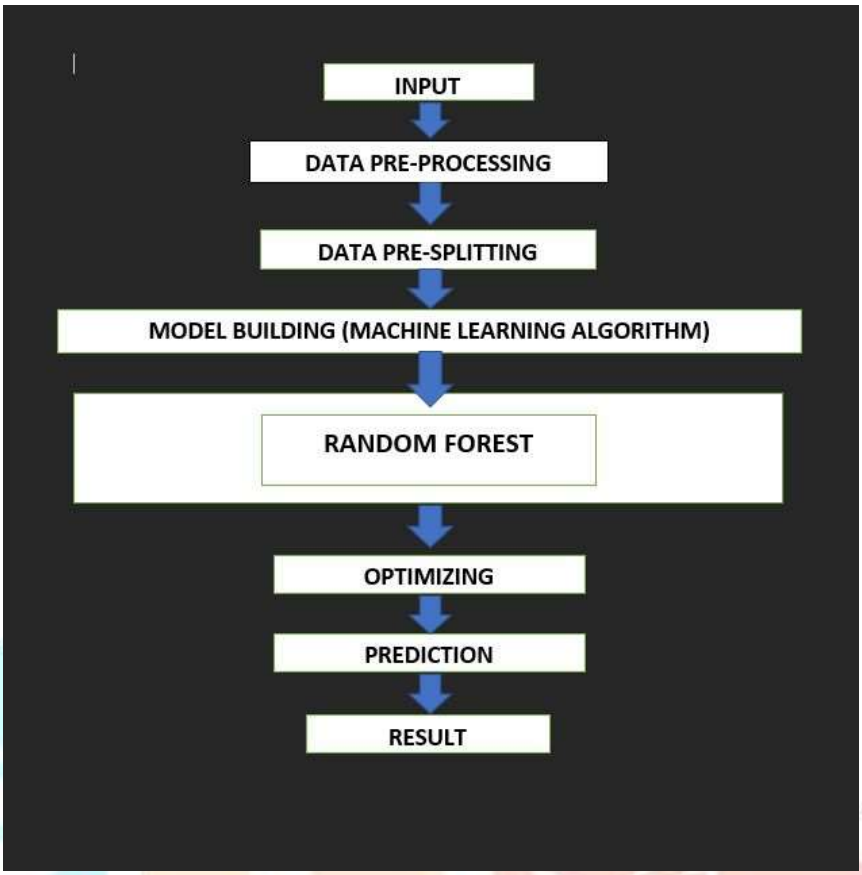


Fig -1: Proposed Testing Process Flow

BLOCK DIAGRAM



COMPARTIVE STUDY

We then test it against a matchday where 10 games are played on the weekend by the 20 teams in the Premier League. We then predict the accuracy for each of the algorithms. We then run our machine learning algorithms on them and calculate the accuracy. Test result for the algorithms can be seen in Fig2 and Fig3

```
#RandomForestClassifier
y_pred = clf1.fit(X_train,y_train).predict(X_train)
accuracy_score(y_pred,y_train)
scores = cross_val_score(clf1, X_train, y_train, cv=10)
print(scores)
print(scores.mean())
```

```
[0.31578947 0.38888889 0.29411765 0.58823529 0.41176471 0.52941176
 0.29411765 0.70588235 0.4375     0.375      ]
0.43407077743378053
```

Fig-2: Random Forest Results

The accuracy is lower so we add more important attributes which are influential to the result of a game. We add recent performances of the teams to improve the accuracy. In football the particular form of a team is very important factor which can be very effective especially while predicting the outcome of the game. We calculate the form of the team based on their previous six results.

```
y_pred = clf4.fit(X_train,y_train).predict(X_train)
accuracy_score(y_pred,y_train)
scores = cross_val_score(clf1, X_train, y_train, cv=10)
print(scores)
print(scores.mean())
```

```
[0.31578947 0.55555556 0.41176471 0.41176471 0.41176471 0.52941176
 0.52941176 0.58823529 0.5        0.4375     ]
0.4691197970416237
```

Fig-3: Linear SVC Results

```
y_pred = clf1.fit(X_train,y_train).predict(X_train)
accuracy_score(y_pred,y_train)
scores = cross_val_score(clf1, X_train, y_train, cv=10)
print(scores)
print(scores.mean())
```

```
[0.42105263 0.66666667 0.29411765 0.58823529 0.41176471 0.52941176
 0.41176471 0.58823529 0.625      0.375      ]
0.49112487100103197
```

Fig-4: Random Forest Results

```
y_pred = clf2.fit(X_train,y_train).predict(X_train)
accuracy_score(y_pred,y_train)
scores = cross_val_score(clf2, X_train, y_train, cv=10)
print(scores)
print(scores.mean())
```

```
[0.36842105 0.55555556 0.52941176 0.58823529 0.47058824 0.35294118
 0.58823529 0.70588235 0.5        0.625      ]
0.5284270725834194
```

Fig-5 : KNN

We see an improvement in the results after adding recent performance. To further improve our model, we use stadium advantage. In football stadiums of the teams are very difficult for opponent teams to play at and majority of the times the home team wins therefore adding this attribute will improve our results. We use various attributes to calculate how much home advantage a team has at their stadium. We use attributes such as past home shots, past home corner, past away shots, past away corners, past home goals, past away goals, result, past corner difference, past goal difference and past shots difference.

CONCLUSION

In this research paper, we have built multiple machine learning models to predict 2020-22 English Premier League match results. We can conclude that some attributes are more important than others, but prediction cannot be done using only these attributes. Usage of significant attributes increases the accuracy for the result prediction. We have also proved that algorithms like Support Vector Machine, Random Forest aren't effective for football prediction. Random Forest gives us the best accuracy compared to all the different algorithms used throughout the research. We also concluded that addition of important attributes such as recent performances and home advantage improves the model substantially.

ACKNOWLEDGEMENT

We wish to express our gratitude towards our mentor, Prof.Gayatri Bachav. who at every stage in this project, contributed their valuable input. We also thank our H.O.D Prof. Dr.Pradip Mane Sir, for providing the necessary facilities for the completion of this project work in our college.

REFERENCES

[1] Bailey, M.J. (2005). Predicting Sporting Outcomes: A Statistical Approach. Swinburne University of Technology: Faculty of Life and Social Sciences.

[2] 1. Gayatri Naik,2021,Prediction of Rainfall Using Machine Learning Technique International Journal for Research in Engineering Application & Management(IJREAM) ISSN : 2454-9150(Approved By UGC & Scopus)Vol-07, Special Issue, MAY 2021

[3] 2.Gayatri Naik, 2022,House Price Prediction System using machine learning. International Journal for Research in Engineering Application & Management (IJREAM) ISSN : 2454-9150.(Approved By UG & Scopus)Volume 8 Issue 1 April 2022

[4] aio,G., & Blangiardo,M.(2010). Bayesian HierarchicalModel for The Prediction of Football Results. Journal of Applied Statistics, 253-264

[5] Hosmer, D.W. Lemeshow, S. & Sturdivant, R.X. Applied Logistic Regression 3rd ed. Hoboken, New Jersey: JohnWiley & Sons, Inc

[6] 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

[7] Min, B., et al. (2008). A Compound Framework for Sports Result Prediction: A Football Case Study. Journal ofKnowledge- Based System Volume 21 Issue 7, 551-562.
The Netherlands: Elsevier Science Publishers

[8] Peng,etal.(2002).An Introduction to Logistic RegressionAnalysis and Reporting. Indiana University-Bloomington: EBSCO Publishing.

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

[10] Shin, J., & Gasparyan, R. (2014). A Novel Way to SoccerMatch Prediction. Stanford University: Department of Computer Science.

[11] Snyder, Jeffrey A.L. (2013). What Actually Wins SoccerMatches: Prediction of the 2011-2012 Premier Leaguefor Fun and Profit. Thesis, University of Washington, WA: Department of Computer Science.

[12] https://en.wikipedia.org/wiki/Multiclass classification

[13] https://en.wikipedia.org/wiki/Random forest