International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 442-446, May-2025
https://doi.org/10.47001/IRJIET/2025.905049

Machine Learning Driven Football Predictions

[1]Vivek Patil, [2]Akash Shetty, [3]Soham Tonape, [4]Prof. D.G. Modani

[1,2,3]Student, Department of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India
[4]Asst. Professor, Department of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India

***Abstract* - In this study, we analyzed player performance in 864 Qatar Stars League (QSL) matches (2012-2019) to determine key factors influencing match outcomes. Using a machine learning framework, we classified match results and identified performance metrics that distinguish winning teams from losing ones. Logistic regression emerged as the top model, achieving over 80% accuracy. Key features included opponent analysis, player market value prediction, player profiling, tactical pattern analysis, injury prevention, and team performance metrics. Notably, defenders' roles and fair play significantly impacted match outcomes, and player performance from the last five seasons provided strong predictive power for future matches. Feature Selection: Multiple feature selection methods were used to identify critical performance metrics that contribute to match outcomes, improving the accuracy of the prediction model. Defensive Importance: The analysis highlighted the significant role of defenders, indicating their crucial influence on match results, challenging the common focus on attacking players. Fair Play Impact: Teams that played fair, committing fewer fouls and receiving fewer cards, were more likely to win, showcasing the impact of discipline on success. Historical Data Utility: The model demonstrated that performance data from the last five seasons provides enough predictive power to forecast the winner in upcoming matches. Model Generalization: The machine learning framework showed strong potential to be applied to other leagues and competitions, given its robust predictive accuracy.**

Keywords: Logistic Regression, Match outcome prediction, Machine learning models, Feature selection, Tactical analysis, Player market value prediction.

I. INTRODUCTION

Football, also known as soccer, is one of the most popular and passionately followed sports worldwide, engaging millions of fans and generating significant economic value. From casual fans to professional analysts, predicting the outcome of football matches has always been an area of interest. With the growth of technology and the availability of vast amounts of sports data, machine learning (ML) has emerged as a powerful tool for extracting insights and making informed predictions. This project seeks to leverage ML techniques to forecast football match outcomes with the goal of achieving high prediction accuracy and providing meaningful insights into the game.

The unpredictable nature of football makes it a particularly challenging domain for predictive modeling. A wide range of factors such as team performance, player statistics, injuries, home/away status, and even psychological or weather conditions can influence the outcome of a match. Traditional methods like betting odds or expert opinions often fail to incorporate all these aspects systematically. This is where machine learning offers a major advantage—by analyzing large, multidimensional datasets and uncovering complex patterns that may not be immediately visible to human analysts.

To overcome this difficulty, the clubs recruit scouts of vast experience and regional understanding to identify players. The AIFF is trying to improve the situation by collaborating with various clubs and companies that make it possible to teach the coaches who may be inexperienced by bringing in connecting sessions with the experienced ones, hosting various tournaments at school, city, district, state level, establishing football academies and community initiatives.

The proposed model is aimed specifically at the grass root level players of India, further scaling to other soccer leagues. The system is trained as per the in-game values of the 2017 version of EA Sports FIFA. The reason for choosing values based on a game is that it seemed to be the only source for a reliable, near accurate and open form of data available for football players spanning across several leagues. Moreover, the very nature of the game being a team based sport makes it difficult to analyze the players due to their dependencies on the skillset of other team members, varying positions, formations, club budget, competitiveness in the league and injuries across their career span. Our model is designed to estimate the performance value of the player based on the attributes and skill sets that the player possesses. Coaches can then take advantage of this performance value and train the player, reshuffle the team, recruit, and loan or sell the player. Another value added to this process is the market value of the player obtained through the performance value of the player. However, there will be an approximate deviation in that value by a certain amount due to irregularities in the demand for a particular position, club budget, contract period, injuries and current on-field performance.

II. LITERATURE REVIEW

Bayesian hierarchical model for the prediction of football results

Authors: G. Baio and M. Blangiardo.

In recent years, the issue of football modeling has become more and more common and several various models have been introduced to predict the features that lead a team to lose or win a game or to forecast a result for a specific game. To meet these two objectives and check the Bayesian hierarchical model focused on data from the 1991-1992 Italian Series A championship. They recommend a more complex blend model that matches the observed data to solve the over-reduction problem generated by the Bayesian hierarchical model. The Italian Series A championship 2007-2008 is an illustration for checking its results.

Predicting football scores using machine learning techniques

Authors: J. Hucaljuk and A. Rakipovid

The key aim of the paper is to test numerous methods for machine learning to forecast the outcome and result of football matches by utilizing in-game match activities rather than the number of goals scored by each side. They have tested different model architecture theories and analyze the efficiency of their models against benchmark techniques. In this paper, they have established an' anticipated objective' measure to help us assess the success of a team rather than use the specific achieved goals. This measure is paired with the measurement of an offensive and defensive team ranking update during-game to construct a classification model that predicts the outcomes of future matches and a regression model that predicts future matches. The efficiency of their models correlates well with the current mainstream strategies and is close to that of bookmakers.

Predicting the Dutch Football Competition Using Public Data: A Machine Learning Approach

Authors: N. Tax and Y. Jousts

A framework for the Dutch Eredivisie focused on public data is defined in this article. A systematic literature review described the variables of predictive utility for the match outcomes. Candidate characteristics have been created. Self-made public data collection, consisting of 13 Dutch Eredivisie match data seasons, was accompanied by modeling preparation. A variety of variations have been evaluated on public data testing developed in the dimensional reduction techniques and classification algorithms. A mixture of PCA (with a difference of 15 percent) and a Naive Bayes or Multilayer Perceptron classifier obtained the best detection precision for the public data feature collection. Models for betting odds and a hybrid feature set (common data union and wagering odds features) have been created. Check McNemar has found no substantial gap in the accuracy of the model with the lowest accuracy hybrid function setting and the low precision betting odds, but the findings do lift the supposition that a mixed combination of betting odds and public data will defeat the bookmaker. The results obtained can be seen as a positive sign that competitive structures for supporting betting decisions based on open data can be created.

Predictive analysis and modelling football results using machine learning approach for English Premier League

Authors: Baboota, Rahul & Kaur, Harleen

This paper demonstrates their research in developing a common statistical model for the English Premier League games. They have built a feature collection using software engineering and an exploratory data review, which evaluates the main factors for predicting football match outcomes and thus develop a highly detailed predictive framework by machine learning. They have demonstrated that their model's success is highly based on important characteristics. In the EPL aggregated during two seasons (2014-and 2015–2016) their best model with the gradient boosts achieved a performance of 0.2156 in the probability (RPS) metric for the game week 6 to 38 whereas, the betting organizations they consider (Bet365 and Pinnacle Sports) received RPS value of 0.2012 for the same period. Because the low RPS value reflects a higher predictive accuracy, given encouraging performance, their model did not surpass the forecasts of the bookmaker.

Effects of expertise on football betting

Authors: Khazaal, Y., Chatton, A., Billieux, J

The goal of this analysis was to determine whether football experts could forecast football match scores than non-experts. The precision of football match prognoses does not seem to be affected by experience, age, and sex. The assumption that soccer expertise enhances betting skills is, therefore, nothing more than a cognitive delusion known as the "illusion of control;" gamblers may profit from psychological therapies that work on the illusion of control ties that their perceptions have between betting skills and soccer expertise. The practice that needs to be taken into account is the public safety agenda to discourage football gambling.

III. SYSTEM DESIGN

System design is the blueprint for developing the ML Driven Football Predictions application. It outlines the structure, components, data flow, and interactions within the

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 442-446, May-2025
https://doi.org/10.47001/IRJIET/2025.905049

system to ensure seamless prediction, analysis, and user experience.

The proposed ML model identified several features, including:

1. Opponent Analysis Task: Analyze the opponent's formation, style of play, attacking/defensive tendencies, and set-piece strategies.
2. Player Transfer Market Value Prediction Task: Predict the market value of a football player based on their performance, age, position, contract length, and more.
3. Player Performance Profiling Task: Understand key player tendencies, such as strengths, weaknesses, and style of play (e.g., right-footed, aerial threat).
4. Tactical Pattern Analysis Task: Identify the team's formation (e.g., 4-4-2, 3-5-2) and how it changes during different phases of the game (e.g., attack, defense, transition).
5. Injury Prevention Task: Analyze player workload (training and match intensity, frequency) to predict the risk of overuse injuries.
6. Team Performance Metrics. Interestingly, we revealed that the defenders' role could not be ignored for match results, and playing fair games improves the chance of winning matches in QSL.

We also showed that players' performance metrics from the last five seasons would provide sufficient discriminative power to the proposed ML model to predict the match-winner in the upcoming season. The proposed ML model will support the players, coaching staff, and team management to focus on specific performance metrics that may lead to winning a match in QSL.

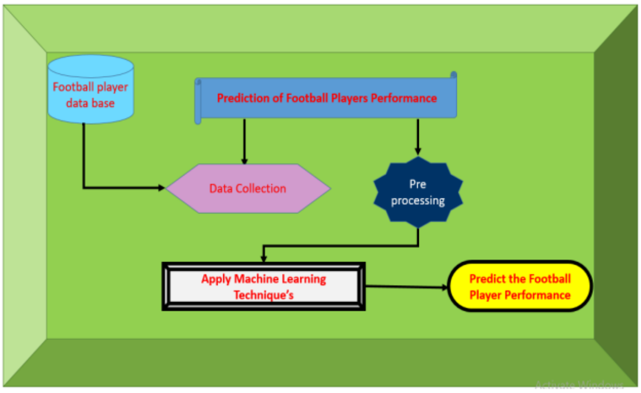
Football player data base
Prediction of Football Players Performance
Data Collection
Pre processing
Apply Machine Learning Technique's
Predict the Football Player Performance

Figure 1: System Architecture Diagram

IV. IMPLEMENTATION DETAILS

The implementation of the **ML Driven Football Predictions system** is carried out in a systematic manner following modular development practices. Each module — from data preprocessing to model training and prediction — is implemented and tested independently, then integrated into the complete pipeline. This section explains the step-by-step process followed during the actual development of the system.

Step-by-Step Implementation Phases

Phase 1: Data Collection and Integration

- Collected football match data (e.g., team stats, match outcomes, goals, possession) from online public datasets (Kaggle, football APIs).
- Stored the dataset in CSV format for ease of use with Pandas.
- Ensured data consistency by merging different files using unique match IDs or dates.

Phase 2: Data Preprocessing

- **Handled missing values** using imputation (mean/mode) or dropped incomplete records.
- **Converted categorical variables** (e.g., team names) into numerical values using Label Encoding or One-Hot Encoding.
- **Normalized numerical features** (e.g., possession %, passes completed) using Min-Max scaling.
- **Split dataset** into training and testing sets using an 80:20 ratio.

Phase 3: Feature Engineering

- Created new features like:
  - **Goal Difference (GD)** = Goals Scored – Goals Conceded
  - **Recent Form Score** = Numerical weightage of past 5 results
  - **Win Rate**, **Draw Rate**, **Loss Rate**
  - **Home/Away performance indicators**
- These features improved model understanding of team strength beyond raw data.

Phase 4: Model Selection and Training

Tested and compared the performance of different ML algorithms:

Model	Implementation
Logistic Regression	scikit-learn
Random Forest Classifier	scikit-learn

- Used **GridSearchCV** for hyperparameter tuning.
- Trained each model with the training set and validated with the test set.

Phase 5: Prediction and Evaluation

- Used trained models to predict the outcome (Win, Draw, Loss) for new match inputs.
- Evaluated models using:
  - **Accuracy Score**
  - **Confusion Matrix**
  - **Precision, Recall, F1 Score**

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 442-446, May-2025

https://doi.org/10.47001/IRJIET/2025.905049

Phase 6: Visualization and Insights

- Used **Matplotlib** and **Seaborn** for plotting:
  - Team-wise performance
  - Feature importance scores
  - Confusion matrix heatmaps
  - ROC Curves
- This provided analytical support to understand predictions better.

Phase 7: User Interface (Optional)

- Developed a basic Flask-based web app that allows users to:
  - Input team names and match conditions
  - View predicted result with confidence score
  - Display performance metrics of the model
- CLI (Command Line Interface) was also provided for simple interaction during testing.

Integration and Testing

- Each module was tested individually (unit testing).
- Then integrated into a pipeline:
  1. Load data → 2. Preprocess → 3. Feature Engineering → 4. Predict → 5. Show Results
- Performed **end-to-end testing** to verify consistency and correctness of predictions.

V. CONCLUSION

In conclusion, the ML-Driven Football Predictions project effectively tackles the complexities faced by modern football clubs in assessing performance, predicting outcomes, and managing player fitness. By leveraging machine learning, the system transforms vast amounts of data into actionable insights, allowing coaches, management, and analysts to make more informed decisions. From evaluating opponent strategies to predicting player market values and assessing injury risks, the system provides a comprehensive solution for improving team performance and safeguarding player welfare. Ultimately, this project empowers clubs to adopt data-driven strategies that enhance both immediate match results and long-term success in player management.

Additionally, the implementation of this ML-driven system marks a significant step towards the integration of advanced analytics in football. By continuously learning from new data, the model evolves, offering increasingly accurate predictions that reflect the dynamic nature of the sport. This adaptability ensures that teams can stay ahead of trends and make real-time adjustments to their strategies. Moreover, by identifying potential risks, such as player injuries or suboptimal performance, the system helps to minimize disruptions, ensuring players remain in peak condition and that teams can maintain a competitive edge over time. Ultimately, this project represents the future of football management, where technology and data work hand-in-hand to drive success.

REFERENCES

[1] G. Baio and M. Blangiardo. "Bayesian hierarchical model for the prediction of football results." University College London Department of Statistical Sciences, Gower Street, London WC1 6BT

[2] J. Hucaljuk and A. Rakipovid. "Predicting football scores using machine learning techniques." University of Zagreb, Faculty of Electrical Engineering and Computing Unska 3, 10000 Zagreb, Croatia

[3] N. Tax and Y. Jousts. "Predicting the Dutch Football Competition Using Public Data: A Machine Learning Approach."

[4] Baboota, Rahul & Kaur, Harleen. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2018.01.003.

[5] Khazaal, Y., Chatton, A., Billieux, J. Effects of expertise on football betting. Subst Abuse Treat Prev Policy 7, 18 (2012). https://doi.org/10.1186/1747-597X-7-18

[6] Kampakis, Stylianos and Andreas Adamides. “Using Twitter to predict football outcomes.” ArXiv abs/1411.1243 (2014): n. pag.

[7] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[8] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, 2016, pp. 1-5.

[9] N. Ancona, G. Cicirelli, A. Branca and A. Distante, "Goal detection in football by using support vector machines for classification," IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001, pp. 611-616 vol.1.

[10] Yang, Feng-Jen. (2018). An Implementation of Naive Bayes Classifier. 301-306. 10.1109/CSCI46756.2018.00065.

AUTHORS BIOGRAPHY

Vivek Patil, Student, Department of Computer Engineering, PES’s Modern College of Engineering, Pune, Maharashtra, India.

Akash Shetty, Student, Department of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India.

Soham Tonape, Student, Department of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India.

Prof. D.G. Modani, Asst. Professor, Department of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India.

Citation of this Article:

Vivek Patil, Akash Shetty, Soham Tonape, & Prof. D.G. Modani. (2025). Machine Learning Driven Football Predictions. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(5), 442-446. Article DOI https://doi.org/10.47001/IRJIET/2025.905049

****