Scoring the Future: XGBoost for Predictive Analysis of FC Barcelona Matches

Harsh Dipdatt Patil
*Dept. of **Computer Science and Applications***
Dr. Vishwanath Karad MIT World Peace University
Kothrud, Pune 411038, India
Email: workidharsh29@gmail.com

Ved Niteen Raut
*Dept. of **Computer Science and Applications***
Dr. Vishwanath Karad MIT World Peace University
Kothrud, Pune 411038, India
Email: vedraut49@gmail.com

Anurag Das
*Dept. of **Computer Science and Applications***
Dr. Vishwanath Karad MIT World Peace University
Kothrud, Pune 411038, India
Email: anuragdas00@gmail.com

***Abstract*—Predicting the outcome of football is complicated because of the dynamic nature of the game. In this research we have focuses on one specific team Fc Barcelon, which is one of the successful football clubs, founded in 1899. This research focuses on advancement of machine learning to predict Fc Barcelona's match outcome, with using self-made dataset from the past records of 2023/24 and 2024/25 seasons. Two ensemble-based machine learning algorithms were used, Random Forest and XGBoost to classify match outcome into Win, Draw or Loss. The models were trained and evaluated using an 80-20 train-test split, with hyperparameter and cross validation. Our model achieved an accuracy of 72.22%. This research highlights the potential of machine learning in sports analytics, offering valuable insights for fans, analytics and strategists.**

Keywords—xgboost, random forest, machine learning, accuracy, prediction.

I. INTRODUCTION

Fc Barcelona, commonly known as Barca, is one of the best globally recognized football club in Spain. Founded in 1899, the club has established itself as the strongest club in both domestic and international football. Fc Barcelona has won multiple league titles, UEFA Champions League trophies and Copa del Rey titles. Also known for their attacking style of play, strong youth academy "La Masia" and legendary players like Lionel Messi, Xavi and Andres Iniesta. Their matches attract millions of spectators, making them an ideal team for predictive analysis in sports, therefore predicting Fc Barcelona's match outcome presents an exciting challenge. Traditional methods for predicting football match results often rely on expert's opinions, fan debates and historical trends. However, these approaches struggle to process large volumes of historical and real-time data efficiently, limiting their predictive accuracy. This research aims to bridge the gap by developing a machine learning model capable of predicting Fc Barcelona's match outcome, by utilizing statistical and performance-based features from the 2023/24 and 2024/25 seasons. The primary objective of this study is to develop and evaluate a machine learning model that can predict whether Fc Barcelona will Win, Draw or Loss a match.

Literature Review :

We have reviewed several different research papers, we found that several studies have explored different types of machine learning techniques for football match prediction, utilizing various datasets, methodologies and algorithms to improve accuracy.

In [1], the authors explore the application of machine learning models to forecast outcomes in the English Premier League. The methodology they used was based on Utilized a Poisson distribution probabilistic classifier based on the 'expected goals' metric to predict match outcomes. This approach achieved an accuracy of 52.3%.

The authors of [2], apply deep neural networks to forecast football match outcomes. Analyzing data from 24,760 matches across 13 leagues, they introduce features like relative attack, defense, and midfield power. Their model achieved a 52.8% accuracy over 2,589 matches, suggesting that football outcomes possess a degree of predictability.

In Ref. [3], the authors Applied Support Vector Machines (SVM), Logistic Regression, and Multinomial Naïve Bayes on football statistics and achieved the highest accuracy at 61.29%. In [4], the authors aimed to find the score of the football match using machine learning techniques. They implemented various machine learning algorithms to predict match results based on historical data, and the success rates of this research are between 50% and 60%. In Ref. [5], the authors applied algorithms including Naïve Bayes, K-Nearest Neighbors, Random Forest and Support Vector Machines to forecast football outcomes using historical match data and player attributes. The study achieved a 65.26% accuracy with the Random Forest model, highlighting the potential of machine learning in sports betting. The authors of [6] applied Support Vector Machines (SVM) to forecast football match results. Their model achieved over 55% accuracy in predicting wins and losses, but struggled with draws, attaining less than 15% accuracy.

Research Gap:

Despite various studies applying machine learning to predict football match outcomes, most studies achieve accuracy below 68% and most of studies rely on generic datasets. Our research aims on one specific team "Fc Barcelona" and to enhance prediction accuracy, unlike other studies that rely on pre-existing or generic dataset our model is based on self-made dataset.

II. PROPOSE METHODOLOGY

Our Research proposes a machine learning based methodology to predict the match outcomes of Fc Barcelona using match level statistics from the 2023/24 and 2024/25 seasons. The research implements Random forest and XGBoost algorithms trained on a self-created dataset. Tools and Libraries used in our research are "Python 3" on Jupyter Notebook, "Scikit-learn" for Random Forest, for preprocessing and metrics, "XGBoost" for gradient boosting model and "NumPy & Pandas" for data manipulation. The methodology follows these key steps:

A. Data Collection

The most important part of our research lies in the creation of a custom dataset. Unlike many studies that reply on pre-existing or generic datasets, our research leverages a hand crafted dataset developed specifically for Fc Barcelona's performance across the 2023/24 and 2024/25 season. The data collection process was entirely manual and derived from the highly reputed sports statistics platform called fbref.com [7], which is recognized for its comprehensive and reliable football analytics.

Source of Data:

Fbref.com [7] is a leading platform maintained by Sports Reference LLC, which provides in depth football statistics, advanced metrics and analytical breakdowns for teams and players across major leagues worldwide. The platform compiles data from multiple verified sources, ensuring that the statistics are trustworthy and detailed enough for academic-level analysis. For our research, only official FC Barcelona matches were considered. Our Dataset contain a total of 95 matches, includes all competition FC Barcelona participated in 2023/24 and 2024/25 season, since 2024/25 season currently going on so we only have data till the date of 05-03-2025. Each match entry in the dataset represents a single row.

Data Collection Process:

The Data Collection Process was time consuming, it took us 3 to 4 weeks to complete. Each match was accessed individually through fbref.com, where multiple performance metrics were noted. These metrics were then logged in a structured spreadsheet format using Microsoft Excel.

Basic Match Information: This includes MATCH_ID, DATE, COMPETITION, ROUND, HOME_OR_AWAY, and OPPONENT. These fields establish the context of the match, such as the round number and whether the game was played at home or away.

The following are features which we used in our model, statistical indicators directly related to in-game performance were recorded:

1. GOALS_SCORED: Total number of goals scored by FC Barcelona.
2. GOALS_CONCEDED: Total number of goals conceded by FC Barcelona.
3. POSSESSION: Possession percentage controlled by Barcelona.
4. SHOTS_ON_TARGET: Total number of on-target shots attempted.
5. PASS_ACCURACY: The pass success rate, also recorded as in percentage.
6. EXPECTED_GOALS: An advanced metric that estimates the quality of goal scoring chances based on historical shot data.
7. MATCH_OUTCOME: Match results either Win, Draw or Loss.

To maintain data integrity, cross verification was performed after initial entry. Each data point was double checked against the original source on fbref.com. Where inconsistencies were found (e.g., in possession percentages or xG values), the data was corrected and revalidated. This two-step validation ensured that the dataset remained consistent, accurate and free from manual entry errors.

B. Data Preprocessing

Since the dataset was self-created and manually entered in Microsoft Excel half of the preprocessing part was done, the only part was left is to convert "String" datatypes into "integers", to prepare the data for modelling. Following steps were performed:

1. Datatype Conversion : Percentage columns (POSSESSION, PASS_ACCURACY) were converted from strings (e.g., "75%") to float values (e.g., 0.75).
2. Categorical Encoding: HOME_OR_AWAY mapped to binary values (Home: 0, Away: 1). OPPONENT encoded using Label Encoder to numerical labels. MATCH_OUTCOME mapped to {Win: 0, Draw: 1, Loss: 2}
3. Feature Scaling : StandardScaler was used to normalize numerical variables like POSSESSION, PASS_ACCURACY, and EXPECTED_GOALS.

This stage ensured the dataset was clean, structured, and suitable for analysis with machine learning algorithms

C. Feature Engineering and Selection

We engineered 4 more features to enhance our model and to increase accuracy of our machine learning model, the following features were engineered:

Rolling Averages:

1. AVG_GOALS_SCORED_LAST_5: Mean goals scored in the last 5 matches.
2. AVG_GOALS_CONCEDED_LAST_5: Mean goals conceded in the last 5 matches.

Formula we used:

$$Rolling\ Average_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i} \qquad where\ k = 5$$

Where,

- t is the current match,
- k is the window size (5, in our case),
- x_{t-i} is the value at time t – i (e.g. goal scored).

Performance Metrics:

GOALS_DIFF: Difference between actual goals scored and expected goals (EXPECTED_GOALS).

Interaction Features:

POSSESSION_SHOTS: Product of possession percentage and shots on target.

Streak Features:

1. WIN_STREAK: Count of wins in the last 3 matches.
2. LOSS_STREAK: Count of losses in the last 3 matches

Feature Selection:

To reduce dimensionality and improve model interpretability, feature selection was performed using Recursive Feature Elimination (RFE). RFE works to fitter the model and remove the least important features based on the model's weight until the most significant subset is identified.

This process helped eliminate irrelevant features, ensuring that only the most predictive variables were used for training. The following graph shows the top 10 feature importances by RFE:

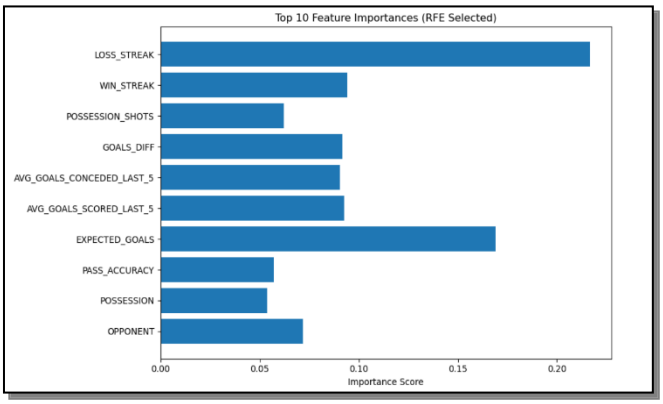

Fig. 1. Top 10 Feature Importants (RFE Selected)

D. Model Development

In this phase, the primary focus was on selecting that can be suitable for our model. The selected algorithms needs to handle both numerical and categorical features, manage non linear relationships and provide interpretable results. Based on that requirements Random Forest Classifier was chosen and to improve accuracy we chose XGBoot Classifier too.

Random Forest Classifier:

Random Forest is aa machine learning algorithm used for classification problems that operates by constructing a multitude of decision trees during training. This method offers several advantages:

- Robustness to overfitting, especially with a relatively small number of instances like the one used in this study.
- Internal feature selection, which highlights the most informative features for prediction.
- Handling of missing data and noise, making it reliable even with slightly imperfect datasets.

The model's performance improves with the number of trees (n_estimators) and can be further fine-tuned by adjusting the maximum depth (max_depth) and minimum samples required to split an internal node (min_samples_split).

XGBoost Classifier:

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of the gradient boosting framework. It builds models in a sequential manner, where each new tree attempts to correct the errors of the previous ones. Key features that make XGBoost suitable for this task include:

- Regularization to reduce overfitting, especially important for a dataset with limited observations.
- Built-in cross-validation and early stopping, which makes model evaluation more dynamic and adaptive.
- High predictive power in classification problems involving complex interactions among features.

E. Model Training

Once the model was selected training process was started, in this phase involved data splitting, validation, hyperparameter tuning and final model fitting.

Train Test Split:

The dataset was divided into two distinct subsets:

- Training Set (80%): Used to train the model and perform cross validation.
- Testing Set (20%): Reserved for evaluating the model's generalization ability on unseen data.

This stratified split ensured that all outcome classes (Win, Draw, Loss) were proportionally represented in both sets, maintaining class balance

Cross Validation:

To avoid the risk of overfitting and to assess model consistency, K-Flod Cross-Validation was employed with k=5. The training data was split into five equal parts, during each iteration, four parts were used for training and one for validation. The process was repeated five times, and the average of the evaluation metrics was taken. This approach ensures the model's performance is not overly reliant on a particular subset of the data. The following is the formula for our Cross-Validation 5-fold CV implementation:

$$CV_{score} = \frac{1}{k}\sum_{i=1}^{K} score_i$$

Where:

- CV_{score} is average cross-validation score,
- k is the number of folds.(in your case k = 5),
- $score_i$ is performance score (like accuracy or F1-score) from the i^{th} fold.

Hyperparameter Tuning:

To optimize model performance, GridSearchCV was utilized for systematic hyperparameter tuning. A predefined range of values was tested for each key parameter:

- For Random Forest: n_estimators, max_depth, min_samples_split, and criterion.
- For XGBoost: learning_rate, max_depth, n_estimators, and subsample.

GridSearchCV evaluated each combination using cross-validation and selected the configuration that achieved the best score, primarily using accuracy and F1-score as the selection criteria

III. Simulation Results

In Simulation results we have tested our model, calculated accuracy, confusion matrix and tested our model in real time results.

Accuracy:

After fitting the model we tested its accuracy, our model achieved an accuracy of 72.22% on the test set, demonstrating its ability to correctly classify match outcomes (Win, Draw, Loss) in approximately 7 out of 10 cases. The following formula was used:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Confusion Matrix:

The purpose of confusion matrix is to show counts of true vs predicted labels to identify misclassifications patterns.

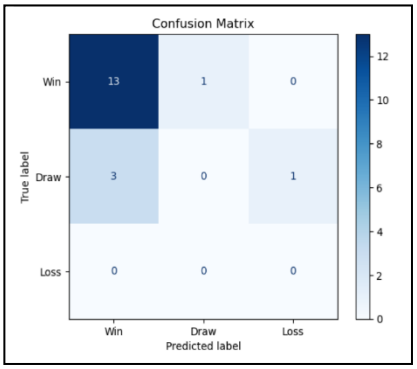

Fig. 2. Confusion matrix

From figure 2, Diagonal cells show correct predictions and Off-diagonal cells show misclassifications (e.g., Draws predicted as Wins).

Performance Metrics by Class:

To highlight precision, recall and F1-score for each class, the following table is the output table after calculating performance metrics.

TABLE I. Performance Metrics Output

Class	Precision	Recall	F1-Score	Supports
Win	0.81	0.93	0.87	14
Draw	0.00	0.00	0.00	4
Loss	0.00	0.00	0.00	0
Macro Avg	0.27	0.31	0.29	18
Weighted Avg	0.63	0.72	0.67	18

The model performed exceptionally well in predicting "Win" but struggled with "Draw" and "Loss" due to class imbalance and limited samples for these outcomes. Following Graph show Performance Metrics by Class (Precision, Recall and F1-score):

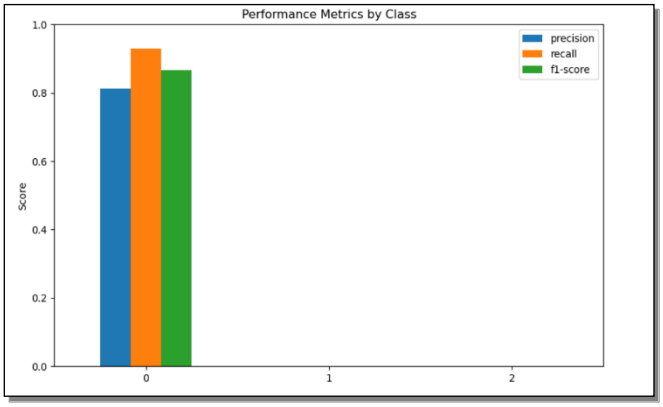

Fig. 3. Performance Metrics by Class

Multi-Class Precision Recall curve:

To visualize the trade-off between precision and recall for each class (Win/Draw/Loss) at varying probability thresholds. Higher Average Precision indicates better class separation. Formula for Average Precision (AP):

$$AP_k = \sum_n (R_n - R_{n-1})\ P_n$$

Here:

- AP_k is Average Precision at cut-off rank k,
- n is the index of the ranked item (from 1 to k),
- R_n is recall at position n,
- R_{n-1} is recall at position $n-1$,
- P_n is precision at position n.

The following Graph shows the trade-off between precision and recall for each class:

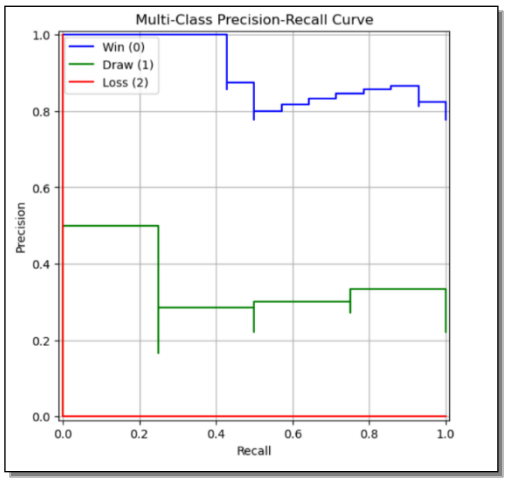

Fig. 4. Multi-Class Precision-Recall Curve

- Win(Blue): High AP suggests strong predictive power.

- Draw(Red): Low AP reflects class imbalance or model confusion.

Cross-Validation:

The model showed consistent performance with a mean accuracy of 73.52% (±9.63%) across 5 folds, indicating robustness against overfitting.

Hyperparameter Tuning:

The best hyperparameters for the XGBoost model were identified through GridSearchCV:

- Learning Rate: 0.2
- Max Depth: 5
- N_Estimators: 50
- Subsample: 1.0
- Colsample_bytree: 0.9
- Gamma: 0.1

Interactive Prediction Example:

The model was tested interactively to predict the outcome, here is the screenshot of the output from jupyter notebook:

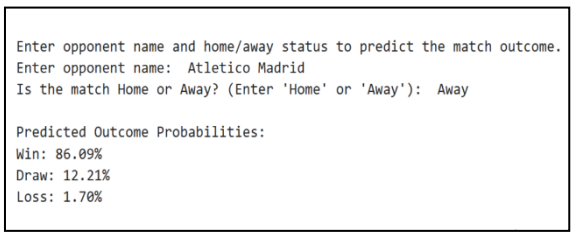

Fig. 5. Output of the Code

Predicted Probabilities:

- Win: 86.09%
- Draw: 12.21%
- Loss: 1.70%

Real time result came out on 17 March 2025 Fc Barcelona won the match by final score of 2 - 4.

IV. CONCLUSION

This research successfully demonstrates the application of machine learning techniques specifically Random forest and XGBoost classifiers to predict the outcome of Fc Barcelona's football match using a self-made dataset compiled from season 2023/24 and 2024/25. Our model achieved an accuracy of 72.22%, which is very good improvement over several existing studies in the literature that reported accuracies below 68%. While the model's predictive power is promising, there are still more ways for enhancement. Incorporating additional features like player availability, opponent form, weather conditions and tactical changes could further improve accuracy and real-world applicability. Additionally, increasing the dataset size and improving class balance especially for draw outcomes will likely lead to more robust results. In conclusion, this research highlights how data science and machine learning can effectively support outcome prediction in football, not only enhancing fan and analyst engagement but also providing valuable insights for coaching staff and sports strategists.

V. REFERENCES`

[1] V. Chang, K. Hall, and L. M. T. Doan, "Football results prediction and machine learning techniques," *Int. J. Bus. Syst. Res.*, vol. 17, no. 5, pp. 565–586, 2023, doi: 10.1504/IJBSR.2023.133178..

[2] Luiz LE, Fialho G, Teixeira JP. Is Football Unpredictable? Predicting Matches Using Neural Networks. *Forecasting*. 2024; 6(4):1152-1168. https://doi.org/10.3390/forecast6040057.

[3] Adish Golechha, Akshat Muke. Using Machine Learning to Analyse Football Teams and Predict the Outcome of a Football Match. Journal of Communication Engineering & Systems. 2023; ():-. Available:https://journals.stmjournals.com/joces/article=2023/view=90291.

[4] Hucaljuk, Josip & Rakipovic, Alen. (2011). Predicting football scores using machine learning techniques.. 1623-1627.

[5] Rodrigues, Fátima & Pinto, Ângelo. (2022). Prediction of football match results with Machine Learning. Procedia Computer Science. 204. 463-470. 10.1016/j.procs.2022.08.057.

[6] M, Jaeyalakshmi & S, Indrajith & C, Hirthik & K, Kaushiik & S, Eaknath. (2023). Predicting the outcome of future football games using machine learning algorithms. 1-7. 10.1109/RMKMATE59243.2023.10370000.

[7] FBref, "Barcelona match statistics – 2023/24 and 2024/25 seasons," 2024. [Online]. Available: fbref.com