Predicting Football Player Performance Using ML

1. ABSTRACT

This project explores a machine learning-based approach to predicting football player performance for the upcoming season using player data from the FIFA 23 dataset. The primary objective is to assist decision-making in football scouting, fantasy football strategies, and player development analysis. The project utilizes player attributes such as overall rating, potential, shooting, passing, dribbling, defending, and pace to create a custom performance score and classify players as likely to perform well or not. Three models—random forest, linear regression, and logistic regression—are used to predict player performance, with the performance of each model evaluated using metrics such as mean absolute error (MAE), R^2 score, accuracy, and F1-score. Furthermore, an interactive web-based dashboard is developed using Streamlit, enabling users to input player data, simulate future performance outcomes, and visualize top-performing young players (wonderkids) and undervalued talents (underrated gems). The project demonstrates the practical potential of machine learning in football analytics, providing clubs, analysts, and fans with valuable insights into player performance predictions.

1. INTRODUCTION

It has emerged as a tool for helping clubs, scouts, and analysts use numbers to make better decisions on potential scores and rather performance outputs. Ever since the advent of data, predicting football player performance has been gearing towards giving indications as to which players would fit recruitment decisions and fantasy football tactics and into development plans. This project also focuses on prediction based on the performance indicators such as overall rating and potential of football players and other metrics predicting whether or not one will be successful in a season. Machine learning-based models are used here to get the forecast about how a player will perform in that season considering the historical metrics. A web-based dashboard has also been developed to show performance and allow user-based prediction created through input data.

2. LITERATURE REVIEW

Machine learning applications in sports analytics, particularly football performance prediction, have been subject to several studies. Studies from the past typically deal with either regression or classification modeling for the interpretation of a match outcome or define a specific player's performance. However, all these studies usually do not provide a comparative study between models or any visual analytics to provide interesting characteristics in the player traits. For example, several studies cover player development or prediction in a match, yet very few afford a user-friendly interface for clubs and analysts to interactively dig through the data. Recently, "wonderkids" who are younger players and have the potential to develop their skills have been identified in the context of player prediction. Similarly, "underrated gems" are the ones that define the players whose performance exceeds the expectations determined by their market value. Hence, in filling up the existing research gaps, the current project integrates these concepts and provides an interactive tool.

1. Baboota & Kaur (2019) – *Predictive analysis and modeling football results using machine learning approach for English Premier League*

This paper applied various machine learning algorithms including Logistic Regression, SVM, and Random Forest to predict football match outcomes based on team stats. The focus was primarily on match-level outcomes rather than individual player performance. It demonstrated that ensemble models like Random Forest offered better accuracy (~70%) compared to traditional algorithms.

> Insight: ML is effective in sports outcome prediction but lacks focus on individual player performance.

2. Danisik, Lacko & Farkas (2018) – *Football match prediction using player attributes*

This study utilized FIFA player statistics and tried to correlate them with match outcomes. They used regression models to predict match scores and team strengths. However, player-level insights were not deeply explored, and performance score estimation for individuals was not their main goal.

> Insight: They validated that FIFA stats are a reliable proxy, but didn't model player performance directly.

3. Harshavardhan & Varma (2022) – *Prediction of Football Players' Performance using Machine Learning and Deep Learning Algorithms*

   This project proposed a data-driven approach to predict football player performance based on attributes extracted from FIFA 17 datasets. The system employed classical machine learning techniques like Decision Trees, ID3, and Boosting Algorithms to build regression and classification models. The focus was on identifying talented players at grassroots level by analyzing skills, overall ratings, and market values. Visualization techniques and custom scoring metrics were introduced to assess players holistically. However, the models were limited to basic algorithms and did not incorporate modern model tuning techniques or real-time interactivity.
   Insight: Offers a solid foundation for player performance

prediction using classical ML, but lacks hyperparameter tuning, latest datasets, and practical deployment in a user-facing tool.

Research Gaps Identified

1. Lack of Player-Centric Focus:
   Most previous works focus on team-level predictions (win/loss, match score), but very few address individual player performance in a structured, predictive manner.
2. Outdated or Limited Datasets:
   Many models rely on older FIFA datasets (2017 or earlier) or static stats. Real-world data evolution is not reflected.
3. No User-Interactive Systems:
   Prior research did not implement interactive dashboards or tools for real-time prediction or decision support for coaches, scouts, etc.

Justification for Chosen Approach

My project addresses these gaps by:

- Using FIFA 23 data for recent and relevant player statistics.
- Creating a custom Player Performance Score, which aggregates core skill metrics.
- Applying machine learning models (Random Forests, SVM) with hyperparameter tuning to accurately predict performance.
- Building a Streamlit-based web app, enabling real-time prediction using both player data and manual input.
- Including intelligent features like Wonderkids discovery and Underrated Gems, aiding real-world decision-making.

This approach not only offers greater accuracy, but also practical usability, making it more impactful than previous academic works.

4. METHODOLOGY

Dataset

The FIFA 23 dataset, which contains detailed player statistics, was chosen for this project. The dataset includes attributes such as player name, overall rating, potential, shooting, passing, dribbling, defending, and pace. These features were selected as they are highly indicative of a player’s overall performance and future potential.

5. DATA PREPROCESSING

Before training the models, several preprocessing steps were performed:

Label encoding: Categorical variables (such as player position) were label-encoded to convert them into numerical data.

a. Standardization: Continuous variables like player attributes were standardized to ensure they are on the same scale, improving model performance.

b. Feature engineering: A custom "performance score" was created by combining key player attributes such as overall rating, potential, and key skills like shooting and passing.

c. Handling missing data: Any missing values in the dataset were imputed using the mean value of the respective attribute.

6. MODELS USED

Three machine learning models were employed:

a. Random forest regressor: A robust ensemble method that captures non-linear relationships and interactions between features.
b. Linear regression: A basic model used for regression tasks to predict continuous variables.

c. Logistic regression: Used for classification tasks to predict whether a player will have a successful season (binary classification).

□

7. HYPERPARAMETER TUNING

The random forest models were tuned using RandomizedSearchCV to optimize hyperparameters and improve model performance.

Evaluation metrics

For regression tasks, the following metrics were used:

a. Mean absolute error (MAE): To measure the average magnitude of the errors.

R^2 score: To evaluate the proportion of the variance explained by the model.

For classification tasks, the following metrics were used:

b. Accuracy: To measure the percentage of correctly predicted outcomes.

c. F1-score: To evaluate the balance between precision and recall.

Implementation & Results

Model performance

The models were evaluated based on the following performance metrics:

d. Random forest regressor:

  i. MAE: 1.28 ○ R^2: 0.87

e. Grid search random forest regressor:

  i. MAE: 1.15 ○ R^2: 0.89

f. Linear regression: ○ MAE: 2.10 ○ R^2: 0.72

g. Random forest classifier:

  i. Accuracy: 0.93

h. Grid search random forest classifier:

  i. Accuracy: 0.95

i. Logistic regression classifier:

▯

i. Accuracy: 0.88

8. VISUALIZATIONS

a. Feature importance: Visualized the importance of different features (e.g., overall, potential) in the random forest model.

Wonderkids: A chart was created to identify the topperforming young players based on their potential.

b. Underrated gems: A scatter plot showing players whose performance is high relative to their market value.

9. DISCUSSION

The results indicate that random forest models outperformed linear regression models in both regression and classification tasks. GridSearchCV further optimized the performance of the random forest models, improving accuracy and R² scores. The ability of random forest to capture non-linear relationships and interactions between player attributes was crucial in predicting player performance. The visualization tools helped identify valuable insights, such as discovering promising wonderkids and identifying underrated gems, whose market value does not reflect their true performance.

While the models performed well, there is room for improvement. One limitation is the lack of external match data (e.g., real-time player performance during matches) to refine predictions. Additionally, more granular features such as injury history or match ratings could enhance model accuracy. Future work could explore deep learning models for even richer predictions.

10.CONCLUSION & FUTURE SCOPE

This project demonstrates the practical use of machine learning models in predicting football player performance with

▯

high accuracy. The interactive dashboard enhances the accessibility of this tool, enabling users to input player data, simulate future performance outcomes, and visualize key insights. The project can be extended to include real-time match data or predictive features for team-level performance.

Future developments could integrate external APIs for live player statistics, include deep learning models for more complex patterns, or expand the dashboard to support more granular metrics like player injuries or form over time.

11. REFERENCES

a. FIFA Player Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset

b. Constantinou, A. C., & Fenton, N. E. (2012). Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. Journal of Sports Sciences.

c. Scikit-learn documentation: https://scikit-learn.org/

d. Streamlit documentation: https://docs.streamlit.io/

e. Harshavardhan, D., & Rushendra Varma, M. L. S. (n.d.). Prediction of Football Players Performance Using
Machine Learning and Deep Learning Algorithms. Sathyabama Institute of Science and Technology.

f. N. Danisik, P. Lacko and M. Farkas, "Football Match Prediction Using Players Attributes," 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia, 2018, pp. 201-206, doi: 10.1109/DISA.2018.8490613. keywords: {Sports;Biological neural networks;Training;History;Predictive models;Games;Neurons},

g. Baboota & Kaur football results using machine learning *approach for English Premier League* (2019) – *Predictive analysis and modeling https://doi.org/10.1016/j.ijforecast.2018.01.003.*