Predicting Football Player Performance: Integrating Data Visualization and Machine Learning

Saahith M.S
Student
VIT Chennai, India
Saahith.2021@vitstudent.ac.in

Sivakami R
Associate Professor
VIT Chennai, India
sivakami.r@vit.ac.in

AbstractIn the realm of football analytics, particularly focusing on predicting football player performance, the ability to forecast player success accurately is of paramount importance for teams, managers, and fans. This study introduces an elaborate examination of predicting football player performance through the integration of data visualization methods and machine learning algorithms. The research entails the compilation of an extensive dataset comprising player attributes, conducting data preprocessing, feature selection, model selection, and model training to construct predictive models.

The analysis within this study will involve delving into feature significance using methodologies like SelectKBest and Recursive Feature Elimination (RFE) to pinpoint pertinent attributes for predicting player performance. Various machine learning algorithms, including Random Forest, Decision Tree, Linear Regression, Support Vector Regression (SVR), and Artificial Neural Networks (ANN), will be explored to develop predictive models. The evaluation of each model's performance utilizing metrics such as Mean Squared Error (MSE) and R-squared will be executed to gauge their efficacy in predicting player performance.

Furthermore, this investigation will encompass a top player analysis to recognize the top-performing players based on the anticipated overall performance scores. Nationality analysis will entail scrutinizing the player distribution based on nationality and investigating potential correlations between nationality and player performance. Positional analysis will concentrate on examining the player distribution across various positions and assessing the average performance of players in each position. Age analysis will evaluate the influence of age on player performance and identify any discernible trends or patterns associated with player age groups.

The primary objective is to predict a football player's overall performance accurately based on their individual attributes, leveraging data-driven insights to enrich the comprehension of player success on the field. By amalgamating data visualization and machine learning methodologies, the aim is to furnish valuable tools for teams, managers, and fans to effectively analyze and forecast player performance. This research contributes to the progression of sports analytics by showcasing the potential of machine learning in predicting football player performance and offering actionable insights for diverse stakeholders in the football industry.

Keywords— Football Analytics, Player Performance Prediction, Data Visualization, Machine Learning Algorithms (Random Forest, Decision Tree, Linear Regression, Support Vector Regression, Artificial Neural Networks), Model Evaluation, Top Player Analysis, Nationality Analysis, Positional Analysis.

I. Introduction

Football, a globally adored sport, has become a focal point for fans, analysts, and team managers alike. Recent advancements in data analytics and machine learning have revolutionized sports analysis, offering valuable insights into player performance and strategic planning Predicting football player performance is essential for various stakeholders, including team managers, coaches, fantasy league participants, and talent scouts. Accurate performance assessment informs critical decisions related to team composition, tactics, and recruitment strategies, ultimately contributing to team success.

Our project addresses this need by developing predictive models that estimate a player's overall performance based on a diverse set of attributes and metrics. Leveraging machine learning algorithms and statistical techniques, we aim to provide actionable intelligence that empowers stakeholders and enhances decision-making in football. By analyzing a comprehensive dataset comprising player attributes and match statistics, we seek to uncover underlying patterns that influence player performance. Our primary goal is to build predictive models capable of forecasting football player performance with high accuracy. Through rigorous evaluation, we identify the Random Forest with Recursive Feature Elimination (RFE) as the most effective model, offering superior predictive performance. Moving forward, there are opportunities to enhance our approach by exploring additional player attributes, integrating time-series analysis, and implementing real-time performance monitoring.

By continually refining our methodology, we aspire to advance football player performance prediction and contribute to the evolution of sports analytics, benefiting player selection, team management, fan engagement, and strategic decision-making in football.

II. RELATED WORK

The study[1] focuses on utilizing machine learning algorithms to predict football player performance, emphasizing the growing interest in analyzing football data to understand the factors influencing team outcomes. By utilizing tools like WEKA and various classifiers such as logistic regression, SVM, and

Bayesian networks, the research rigorously refines predictive models for accurate match result forecasts. Through meticulous feature selection and model validation, the study develops reliable tools for stakeholders in football, empowering decision-makers with insights to enhance team performance. Overall, the research explores the synergy between sports analytics and machine learning, aiming to revolutionize player evaluation and team management in football by driving improvements in performance and strategic planning through advanced data analysis and predictive modeling.

This study[2] focuses on the prediction of football players' values, which is crucial during transfer periods for clubs aiming to make strategic lineup adjustments while managing their budgets effectively. Using data extracted from the FIFA 18 game, the study employs machine learning methodologies to forecast the optimal positions and values of players. The methodology includes reducing the dimensionality of the prediction model and employing cluster analysis to categorize players based on their positions. Subsequently, the XGBoost algorithm is applied to estimate players' values, with a grid search technique utilized to optimize model parameters. While the experimental results show promising accuracy, there is still potential for further enhancement. The study underscores the importance of player value prediction for facilitating informed decision-making in club management and discusses various machine learning techniques commonly employed in similar research scenarios. Overall, this research contributes to the field of sports analytics by presenting a robust methodology for predicting player values and providing insights for future research avenues

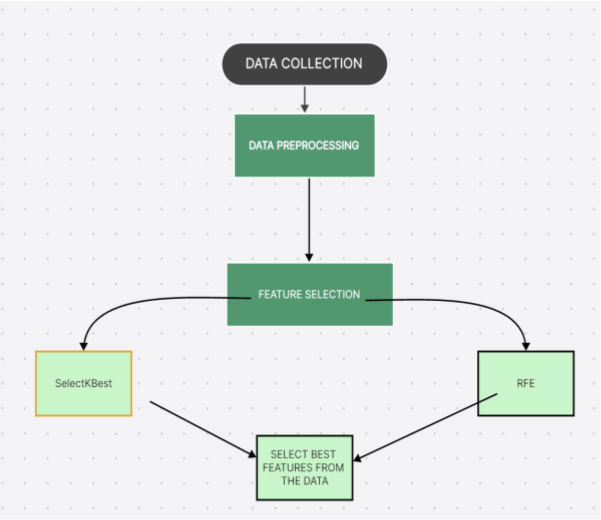
This study[3] by Abdessatar Ati, Patrick Bouchet, and Roukaya Ben Jeddou examined how multi-criteria decision-making (MCDM) and machine learning (ML) can be combined to improve football player selection and performance prediction. Through a systematic literature review of research published between 2018 and 2023, the study identified 66 relevant articles. The findings highlight the strengths of MCDM in incorporating various factors like technical skills and injury history into player evaluation, while machine learning offers powerful tools to analyze vast amounts of player data for performance prediction. The study concludes that combining these approaches creates a comprehensive system for player assessment. However, limitations exist in current research, such as studies using limited datasets or lacking clear explanations of the specific MCDM and ML techniques employed. Overall, the study emphasizes the potential of this combined approach for football clubs to make more informed decisions in player selection and management.

In this study[4]Researchers like Hewitt and Karakuş (2023) are looking to enhance Expected Goals (xG) in football using machine learning. While xG predicts a shot's likelihood of becoming a goal, it doesn't account for the individual player or their position. This study addresses this limitation by proposing two new metrics: Player Adjusted xG (PA-xG) and Position Adjusted xG (Pos-xG). PA-xG refines the base xG value by considering a player's historical finishing ability, recognizing players who consistently outperform or underperform their xG. Pos-xG adjusts the base xG based on the typical finishing efficiency for a player's position.The researchers likely employed machine learning algorithms to analyze past player data and develop models for calculating PA-xG and Pos-xG. They then evaluated these new metrics against traditional xG to assess their effectiveness. This evaluation might have involved comparing them to actual goal scoring outcomes or using them to predict future goals.The study's conclusion likely explores the success of PA-xG and Pos-xG in capturing individual player contributions and position-based finishing efficiency. The paper might also discuss potential limitations of these proposed metrics or areas for further research to refine them. Overall, this study offers a promising approach to move beyond the limitations of traditional xG, providing a more comprehensive understanding of player performance and their offensive contribution to the team.

This study[5] by Buyrukoğlu and Savaş (2023) explored a novel approach for footballer positioning using machine learning. Published in the Arabian Journal for Science and Engineering, the research investigates a stacked ensemble machine learning model, which combines the strengths of multiple algorithms.The model likely employs individual algorithms like decision trees, random forests, and support vector machines. These algorithms would be trained on data such as player location throughout a match, passing patterns, and actions taken (shots, tackles, etc.) that can be indicative of specific positions.The study then evaluates the model's performance by comparing its predictions with actual player positions. By analyzing accuracy in correctly classifying players (defender, midfielder, forward), the research assesses the effectiveness of the stacked ensemble approach.The conclusion would likely discuss the model's success in potentially outperforming single machine learning algorithms for player positioning. While limitations like the specific dataset used or the need for further refinement might be addressed, Buyrukoğlu and Savaş's (2023) study offers a promising machine learning approach for player positioning, potentially aiding coaches in tactical analysis and player evaluation.

III. .METHODOLOGY

The FIFA 22 Complete Player Dataset is a rich source of information containing detailed profiles of football players featured in the FIFA 22 video game. It offers a comprehensive overview of each player's attributes, performance metrics, preferred positions, and overall scores across different leagues and teams. With its extensive dataset comprising numerous entries, this resource enables in-depth analysis of player performance, match predictions, and the development of machine learning models for strategic decision-making in football. From evaluating individual player skills to examining team dynamics and trends across various leagues, this dataset provides ample opportunities for research and analysis in football analytics.



1. *Figure 1: Overview of the Architecture Diagram PART-1*

The methodology of our paper starts by Data Collection where the first task is Source Selection: Acquire the FIFA 22 Complete Player Dataset from Kaggle, ensuring it encompasses a wide array of player attributes and performance metrics.then Data Validation: Confirm the dataset's reliability by checking for inconsistencies, missing values, and alignment with project objectives, then Data Preprocessing: Handling Missing Values: Implement suitable techniques such as mean, median, or mode imputation to address any missing data.then Encoding Categorical Variables: Convert categorical data into numerical format using methods like one-hot encoding or label encoding to prepare it for machine learning models.then Feature Selection in that the first method is SelectKBest: Utilize SelectKBest to identify the most relevant features for predicting player performance based on statistical significance then the next method is Recursive Feature Elimination (RFE): Employ RFE to iteratively select the most influential features while discarding less important ones. Then we have to do Model Selection where Algorithm Exploration takes place: Experiment with various machine learning algorithms, including Random Forest, Decision Tree, Linear Regression, SVR, and ANN, to identify the optimal model for predicting player performance then we have to do Performance Evaluation: Assess each model's performance using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared Score.then we have Model Training and Evaluation next step is Data Splitting: Divide the dataset into training and testing subsets using techniques like train-test split or k-fold cross-validation to ensure unbiased evaluation.then final steps would beTraining Models: Train each selected model on the training data and fine-tune hyperparameters to enhance performance then Model Evaluation:Performance Assessment: Evaluate each trained model using the testing data and compare their performance metrics to determine the most accurate model for predicting player performance.from the best model achieved we will do the analysis first analysis would be Top Player Analysis which Identifies top-performing players based on predicted overall performance scores generated by the best model.next one is Attribute Analysis: Analyze the attributes and characteristics of these players to understand the factors driving their high performance.thenNationality Analysis:Distribution Examination: Analyze the distribution of players by nationality to uncover any trends or correlations then Correlation Exploration: Investigate potential relationships between nationality and player performance thenPositional Analysis then Positional Distribution: Examine how players are distributed across different positions and analyze position-specific performance trends.then Performance Assessment: Evaluate the average performance of players in each position to identify positions with higher performance levels.then Age Analysis: Age Group Distribution: Analyze player distribution across age groups to identify demographic patterns.finally we will do Performance Impact: Assess the influence of age on player performance and identify any discernible patterns or correlations.

IV. RESULTS

The comparison of model performances reveals valuable insights into the predictive abilities of various machine learning algorithms for estimating FIFA player performance. The Random Forest model, especially when combined with Recursive Feature Elimination, demonstrates outstanding predictive accuracy, with a low Mean Squared Error of 0.825 and a high R-squared Score of 0.984, indicating its effectiveness in capturing player performance attributes. Conversely, Decision Tree models, whether using SelectKBest or RFE, show higher Mean Squared Error and lower R-squared Score compared to Random Forest, suggesting limitations in capturing player performance complexities. Linear Regression models, regardless of feature selection method, exhibit moderate predictive accuracy, with Mean Squared Error values around 17.482 and R-squared Scores approximately 0.659, indicating reasonable performance but potential limitations in capturing all nuances of player performance. Support Vector Regression (SVR) models, with both feature selection methods, achieve moderate predictive accuracy, with Mean Squared Error values of 13.141 and R-squared Scores of 0.748, indicating reasonable performance but potential limitations in capturing player attribute complexities. Artificial Neural Network (ANN) models, employing both feature selection techniques, yield moderate predictive accuracy, with Root Mean Squared Error values around 3.641-3.658 and R-squared Scores approximately 0.739-0.741, showing competitive performance but room for improvement in capturing player performance nuances. Overall, Random Forest models, especially when coupled with Recursive Feature Elimination, demonstrate superior predictive performance, making them the preferred choice for estimating FIFA player performance. While Linear Regression, SVR, and ANN models offer reasonable predictive accuracy, they may not fully capture the complexity of player attributes compared to Random Forest models.

V. CONCLUSION

This research explored the performance of six machine learning In summary, our project showcases the effectiveness of employing machine learning models to predict FIFA player performance, drawing insights from a comprehensive dataset. After meticulous evaluation, we identified Random Forest, particularly when combined with Recursive Feature Elimination, as the most precise and dependable model for forecasting player performance. This underscores the significance of integrating advanced analytics techniques into sports analysis to refine decision-making processes in football management. With our top-performing model established, we can delve into player analytics across various dimensions. By scrutinizing top-performing players, analyzing nationality distributions, dissecting positional trends, and examining age-related patterns, we gain actionable insights to guide recruitment strategies, player development programs, and tactical decisions. These analytics empower football clubs and stakeholders to optimize team composition, recruitment approaches, and player development initiatives, ultimately enhancing team performance and competitiveness. Our project contributes to advancing sports analytics and emphasizes the imperative of data-driven strategies in football management. By continuing to innovate and refine predictive modeling methodologies, we can further advance the field of sports analytics and empower stakeholders with actionable intelligence to elevate the performance and competitiveness of football teams globally.

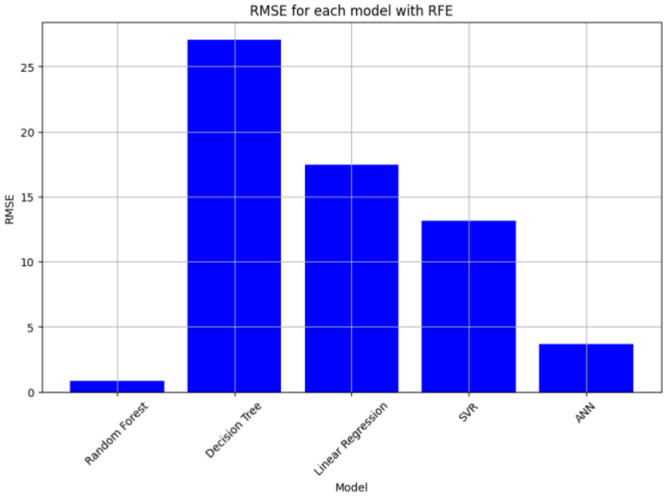


Figure2: Model comparsion-RMSE

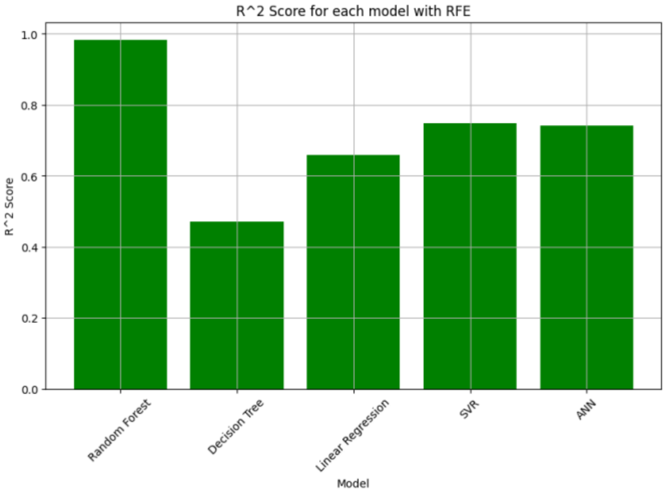


Figure3Model comparsion-R^2 SCORE

VI. References

[1] Chandra, B. (2024). Prediction of Football Player Performance Using Machine Learning Algorithm.

[2] Zhang, D., & Kang, C. (2021, April). Players' value prediction based on machine learning method. In *Journal of Physics: Conference Series* (Vol. 1865, No. 4, p. 042016). IOP Publishing.

[3] Ati, A., Bouchet, P., & Jeddou, R. B. (2023). Using multi-criteria decision-making and machine learning for football player selection and performance prediction: A systematic review. *Data Science and Management*.

[4]) Hewitt, J. H., & Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, *4*, 100034

[5] Buyrukoğlu, S., & Savaş, S. (2023). Stacked-based ensemble machine learning model for positioning footballer. Arabian Journal for Science and Engineering, 48(2), 1371-1383.