JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FOOTBALLER PERFORMANCE FORCASTING

[1]Ms. V. NIVEDHA, [2]Ms. S. POOJA, [3]Ms. J. JASMINE

[1]Student, [2]Student, [3]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

Abstract : The Football Player Performance and Winning Team Prediction Program is a data-driven endeavor that utilizes player performance measures, historical match data, and higher education institutions to forecast the results of football games. The goal is to give football fans, coaches, analysts, and athletes improved insights and forecasts through high-quality data collecting, modeling, and modeling. The program's integration of machine learning techniques is intended to enhance comprehension, streamline decision-making, and offer evaluation for athletic events. This will ultimately assist teams and individuals in gaining a greater grasp of football and related information.

***IndexTerms* – Footballer performance forcasting, Football Player's performance prediction, Performance Analytics, Predictive Modeling, Machine Learning, Sports Statistics, Player Evaluation, Data Mining, Regression Analysis, Feature Selection, Multicollinearity, Variance Inflation Factor (VIF).**

I. INTRODUCTION

The Football Player Performance and Winning Team Prediction Program represents a cutting-edge initiative fueled by the power of data analytics, aimed at revolutionizing the understanding and forecasting of football match outcomes. By harnessing a diverse array of data sources, including comprehensive player performance metrics, historical match data spanning various leagues and tournaments, and insights from esteemed higher education institutions, this program stands at the forefront of predictive sports analytics. At its core, the program seeks to empower football enthusiasts, coaches, analysts, and athletes alike with unparalleled insights and foresight into the intricate dynamics of the game. By leveraging advanced machine learning techniques, meticulously crafted models, and rigorous data collection methodologies, it endeavors to transcend conventional wisdom and offer a nuanced understanding of football phenomena.

II. LITERATURE REVIEW

2.1 "THE APPLICATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING MATCH RESULTS IN TEAM SPORT: A REVIEW" BY RORY BUNKER AND TEO SUSNJAK, PUBLISHED IN APRIL 2022 IN THE JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH.

This review surveys studies published between 1996 and 2019 that applied machine learning methods to predict match results in team sports. Unlike previous review articles, this study adopts a narrow scope to allow for in-depth analysis while avoiding an overwhelming number of surveyed papers.

2.2 "Improvement of Football Match Score Prediction by Selecting Effective Features for Italy Serie A-League" by Yavuz Selim Taspinar, Ilkay Cinar, and Murat Koklu, published in April 2021 in the MANAS Journal of Engineering.

This study focuses on enhancing football match score prediction for the Italy Serie A-League by employing data simplification methods to select effective features in the dataset. The authors address the issue of imbalances by removing features that do not contribute to classification.

2.3 "Soccer net: A Gated Recurrent Unit-based model to predict soccer match winners" by Jassim Al-Mulla, Mohammad Tariqul Islam, Hamada Al-Absi, and Tanvir Alam, published in August 2023 in PLoS ONE.

This paper introduces a deep learning-based method called SoccerNet to predict football match results in the QSL (Qatar Stars League) by considering players' performance metrics. The authors demonstrate the superior performance of deep learning models compared to traditional feature-based machine learning models.

2.4 "A Multidimensional Framework to Uncover Insights of Group Performance and Outcomes in Competitive Environments With a Case Study of FIFA World Cups" by Denisse Martínez and Jose Emmanuel Ramirez-Marquez, published in January 2023 in IEEE Access.

This study presents a multidimensional framework for analyzing group performance and outcomes in competitive environments, using the FIFA World Cups as a case study. The framework incorporates context-specific, network, and opponent factors to provide a comprehensive understanding of group performance patterns

III. EXISTING METHOD

3.1 Existing System

The existing system for football player performance prediction and winning team prediction typically relies on traditional statistics and simpler models, often overlooking the intricate dynamics of the sport. Conventional methods may use basic linear regression for player performance assessment, lacking the ability to effectively handle multicollinearity and overfitting issues. Similarly, match outcome predictions are often based on historical win-loss records or basic goal difference analysis, lacking the probabilistic nature required to account for the inherent unpredictability of football matches. These older approaches can provide useful insights but may not capture the complexity of player interactions and the dynamic nature of football.

3.2 Drawbacks

Moreover, the existing system often lacks the adaptability and robustness of the proposed methodology, which combines Ridge Regression and the Poisson distribution model. The proposed approach leverages advanced techniques to enhance predictive accuracy, utilizes data-driven insights, and accounts for the probabilistic nature of match outcomes. It also encourages continuous learning and model improvement while respecting ethical data usage. The existing system, in contrast, may not fully harness the wealth of data available and is less likely to provide nuanced and accurate predictions for football player performance and match results. The proposed methodology offers a more comprehensive and advanced solution to meet the evolving needs of the football community and sports analytics,abilities. Testing these skills may require additional steps beyond profile viewing, such as interviews or assessments for the candidates by the employer.

IV. PROPOSED METHOD

4.1 UML Diagram

Data Collection: Information from a variety of sources is gathered, including individual statistics, team lineups, and past match outcomes.

Data preprocessing: In order to get the acquired data ready for training machine learning models, it is cleaned, standardized, and transformed.

Feature engineering: It is the process of improving forecast accuracy by manipulating pertinent information such as player form, past performance against certain opponents, and match circumstances.

Model Training: Using the preprocessed data, machine learning models are trained to discover trends and connections between player performance and characteristics.

Model Evaluation: The performance and dependability of the trained models are evaluated using the relevant metrics.

Prediction Generation: Performance forecasts for players in future games are produced using the trained models as a basis.

Integration: Football analytics apps, fantasy football software, and prediction systems all incorporate predictions.

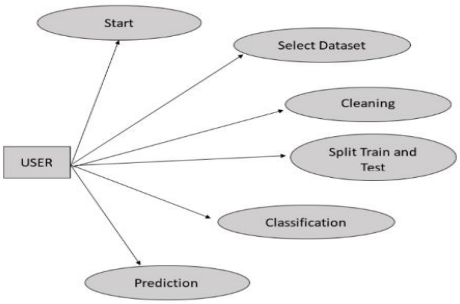


Figure 1: UML diagram

4.2 Flow Diagram

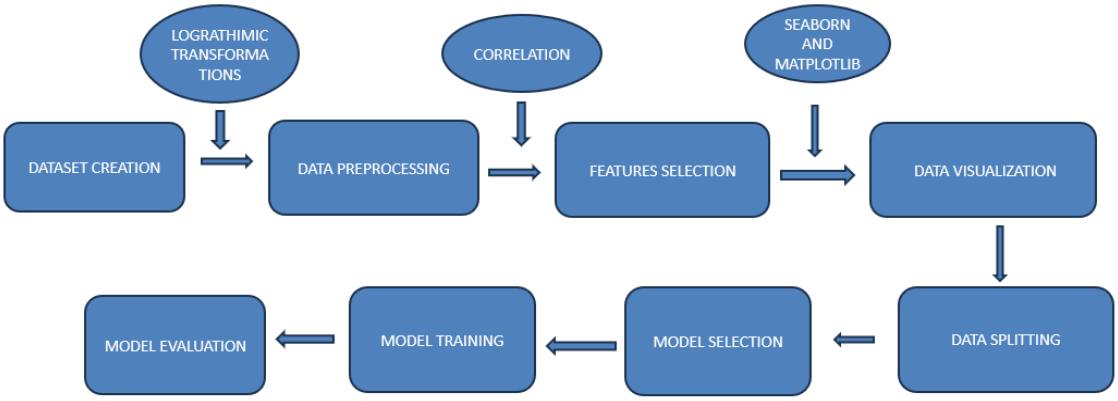


Figure 2: Flow diagram

The flowchart delineates a comprehensive methodology for the development of a machine learning model. It commences with the dataset creation, where raw data is collected and compiled into a usable format. Following this, data preprocessing is conducted to cleanse the dataset, handling missing values and outliers to enhance data quality. Feature selection is then performed to identify the most relevant variables that contribute significantly to the predictive power of the model.

Subsequently, data visualization techniques are employed to explore the data and gain insights through graphical representations. This step is crucial for understanding underlying patterns and relationships within the data. The flowchart then guides us to data splitting, where the dataset is divided into training and testing subsets, ensuring that the model can be trained and validated effectively.

The next phase involves model training, where various algorithms are applied to the training data to build the model. This is followed by model evaluation, where the trained model is tested against the unseen testing data to assess its performance. Metrics such as accuracy, precision, recall, and F1-score are typically used to evaluate the model's predictive capabilities.

The flowchart concludes with a feedback loop, suggesting that the results of the model evaluation may lead to a revisitation of earlier steps such as data preprocessing or model training, allowing for iterative improvements to the model's performance.

4.3 Methodology

In our proposed system for football player performance prediction and winning team prediction, we utilize statistical techniques, specifically Ridge Regression.

Data Collection

We begin by collecting comprehensive historical player data from various reliable sources, including match reports, player profiles, and official statistics repositories. This data encompasses individual player statistics such as goals scored, assists, passes completed, tackles, and other relevant metrics.

Feature Engineering

Next, we engage in feature engineering, which involves extracting relevant features and potentially engineering new ones. This process includes considering factors such as player form, recent performance trends, historical matchups, playing conditions (e.g., weather, venue), and any other pertinent variables that may influence player performance.website. The end users of our product are Patients, Medical staff (Doctors and Nurses), Hospitals, and pharmacies. When the user is logged in or signed in as a patient can access all user profiles by searching them by their names to find them out from the list. Hospital users can access hospitals, nurses, and doctors. Pharmacy users can access hospitals and patients. Nurses can access only the hospitals. Likewise, Doctors can access only the hospitals.The advantages of the proposed methodology are saving time for the user by avoiding the middle person; the users can update their information; user-friendly access; gives information about medical staff, help medical staff to get employed to pursue their career, and reliable to use.

Model Development

We use Ridge Regression, a regularization approach for linear regression, to predict football player performance. Because Ridge Regression can handle multicollinearity in the data and prevent overfitting by adding a regularization term, it is especially well-suited for this kind of assignment.Similar to standard linear regression, the goal of Ridge Regression is to minimize the sum of squared residuals while including a penalty term that is based on the square of the coefficients. The regularization parameter governs this penalty term.By decreasing the coefficients' variance and assisting in their shrinkage towards zero, λ helps to lessen the consequences of multicollinearity.

The following formula represents the Ridge Regression model:

$$\beta\hat{}=\arg\min\beta\{\sum i=1n(yi-\beta 0-\sum j=1pxij\beta j)2+\lambda\sum j=1p\beta j2\}$$

- $\beta\hat{}$ represents the coefficients of the model.
- yi represents the observed player performance.
- xij represents the features.
- $\beta 0$ represents the intercept term.
- βj represents the coefficients associated with each feature.
- λ is the regularization parameter that controls the complexity of the model.

Model Evaluation

We use cross-validation techniques to assess the Ridge Regression model's performance after it has been developed. In order to make sure that the model performs properly when applied to new data, cross-validation is used. Furthermore, we evaluate the model's predicted accuracy using pertinent evaluation metrics, such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), which may be found using the following formula:Model Evaluation: We use cross-validation techniques to assess the Ridge Regression model's performance after it has been developed. In order to make sure that the model performs properly when applied to new data, cross-validation is used. Furthermore, we evaluate the model's predicted accuracy using pertinent evaluation metrics, such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), which may be found using the following formula:

$$MSE=n1\sum i=1n(yi-y\hat{}i)2$$

Here:

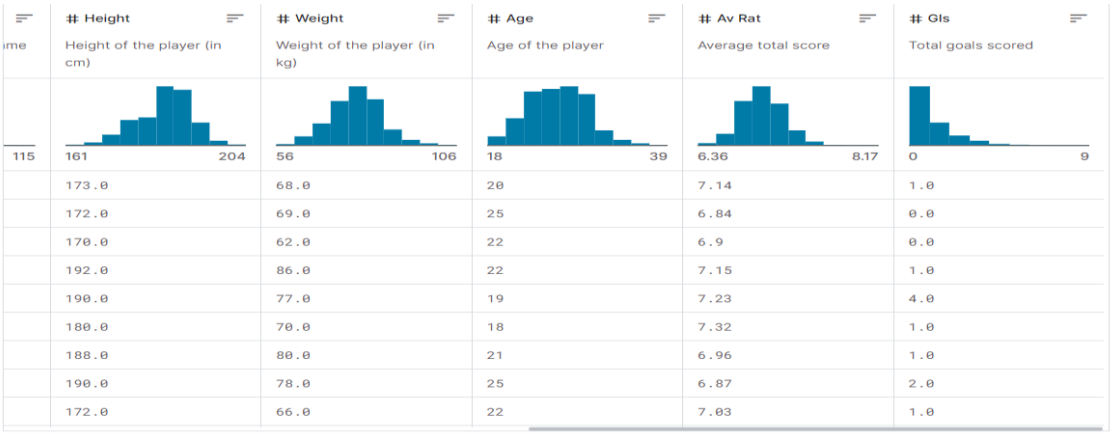
- n represents the number of samples.
- yi represents the observed player performance.
- $y\hat{}i$ represents the predicted player performance.

By following this methodology, we aim to provide accurate predictions of football player performance, thereby offering valuable insights for coaches, analysts, and teams to optimize their strategies and decision-making processes.

V. Result and Discussion

5.1 Data Source

The dataset was obtained from a kaggle that collects data on various football players.



	# Height	# Weight	# Age	# Av Rat	# Gls
ime	Height of the player (in cm)	Weight of the player (in kg)	Age of the player	Average total score	Total goals scored
115	161 – 204	56 – 106	18 – 39	6.36 – 8.17	0 – 9
	173.0	68.0	20	7.14	1.0
	172.0	69.0	25	6.84	0.0
	170.0	62.0	22	6.9	0.0
	192.0	86.0	22	7.15	1.0
	190.0	77.0	19	7.23	4.0
	180.0	70.0	18	7.32	1.0
	188.0	80.0	21	6.96	1.0
	190.0	78.0	25	6.87	2.0
	172.0	66.0	22	7.03	1.0

Figure 3: Dataset

5.2 Features:

The dataset contains several features, including:

1. Player Name: The name of the football player.
2. Age: The player's age at the start of the season.
3. Position: The player's primary position (e.g., forward, midfielder, defender, goalkeeper).
4. Height: The player's height in centimeters.
5. Weight: The player's weight in kilograms.
6. Total Goals: The total number of goals the player scored in the previous season.
7. Total Assists The total number of assists the player provided in the previous season.
8. Passing Accuracy: The percentage of successful passes made by the player.
9. Dribbling Skill: A rating of the player's dribbling skills.

5.3 Data Preprocessing

In the preprocessing stage of our analysis, we encountered missing data within the 'Saves/Sv%' column of our dataset, which consists of various statistics for individuals. To maintain the integrity of our dataset and ensure robust statistical analysis, we implemented a data cleaning process. This involved the substitution of null or missing entries with a default value of "0.0" for numerical consistency. This approach was chosen to facilitate uninterrupted computational operations, as the presence of null values can lead to errors or misinterpretations during analysis. It is important to note that the decision to replace null values with "0.0" was made after careful consideration of the dataset's context and the implications of such a replacement. We ensured that this treatment of null values did not skew our results or misrepresent the underlying data.

	Name	Apps	Mins	Mins/Gm	Height	Weight	Age	Av Rat	Gls	Gls/90	...	Tck R	CA	Saves	Saves/xSv%
0	Josip Mijatović	12.0	809.0	67.416667	173.0	68.0	20	7.14	1.0	0.11	...	0.88	84	0.0	0.0
1	Duje Ninčević	15.0	1161.0	77.400000	172.0	69.0	25	6.84	0.0	0.00	...	0.76	79	0.0	0.0
2	Marin Karabatić	15.0	1350.0	90.000000	170.0	62.0	22	6.90	0.0	0.00	...	0.77	87	0.0	0.0
3	Vicko Ševelj	20.0	1738.0	86.900000	192.0	86.0	22	7.15	1.0	0.05	...	0.88	95	0.0	0.0
4	Fran Vujnović	15.0	1384.0	92.266667	190.0	77.0	19	7.23	4.0	0.26	...	0.93	83	0.0	0.0

Figure 4: Dataset after data cleaning

Figure 4 represents the dataset after the data cleaning process, where null values have been replaced with "0.0" to ensure data integrity and consistency for analysis.

5.4 Data Visualization

Data visualization refers to the presentation of data using visual elements like charts and graphs. This method simplifies intricate data, aids in identifying patterns, and effectively conveys insights.

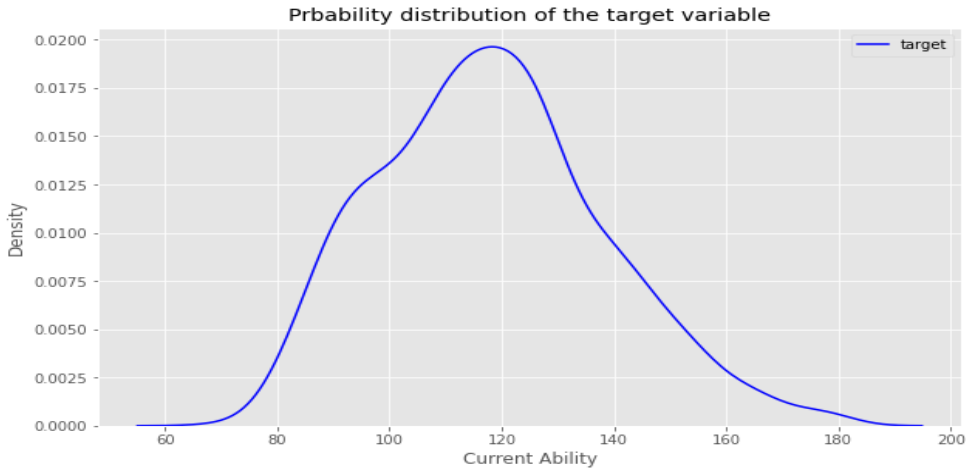


Figure 5: Normal Distribution of Current Ability

Figure 5 showcases the normal distribution of the 'Current Ability' metric within the dataset. The graph is characterized by a bell-shaped curve, indicative of a normal distribution, with the peak centered around a score of 120. This suggests that the majority of the dataset's observations cluster around this central value, with fewer occurrences of extreme high or low ability scores.

The probability distribution graph depicted in Figure X is a visual representation of the frequency of the 'Current Ability' scores within our dataset. The x-axis represents the 'Current Ability' scores, while the y-axis denotes the probability density. The bell-shaped curve peaks at a score of 120, which corresponds to the mode of the distribution, indicating that this score is the most commonly observed value among the participants. The symmetry of the curve around the mode implies that the data points are evenly distributed, decreasing in frequency as they deviate from the central peak. This pattern of distribution is characteristic of a normally distributed variable and is essential for various statistical analyses and predictive model.

Implications: The normal distribution observed in the 'Current Ability' scores is significant as it validates the use of parametric statistical tests that assume normality in the data. It also provides a foundation for further analysis, such as the calculation of z-scores and the assessment of probabilities associated with specific ability levels.

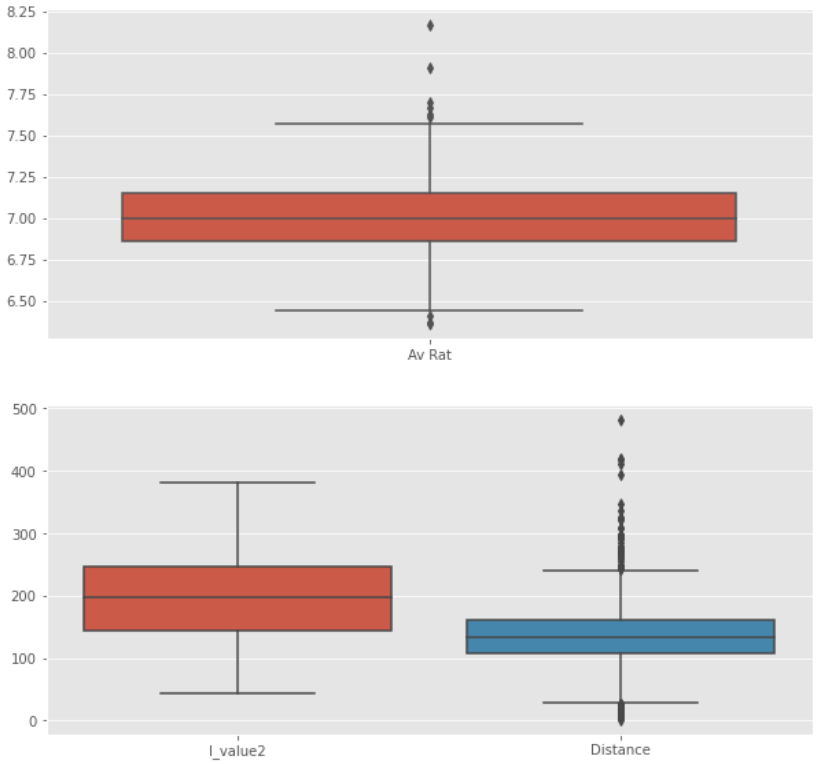


Figure 6: Statistical Analysis of Football Players' Performance Metrics

Figure 6 provides a visual representation of key performance metrics for football players, depicted through box plots that illustrate the distribution of 'Average Rating' (Av Rat) and 'Distance Covered' during matches.
The image consists of two box plots. The first plot represents the 'Average Rating' (Av Rat) of players, which is a crucial indicator of a player's overall performance and skill level. The median value, represented by the line within the box, appears to be around 7.0, suggesting that the average performance rating across the sample is above average. The second plot shows the 'Distance Covered' by players, which reflects their physical endurance and activity on the field. The median value for this metric is indicated around a specific mark, likely measured in meters, emphasizing the importance of stamina and mobility in player performance.
Analysis: These performance measurements' variability and central tendency can be found using the box plots. Additionally, they draw attention to players who, in relation to the rest of the sample, have incredibly high or low scores—known as outliers. Coaches and sports analysts can forecast future performance, pinpoint areas for development, and make well-informed choices on player training and game strategy by examining these metrics.

Application: This statistical approach to performance analysis is pivotal in modern football, where data-driven decisions can lead to enhanced team performance and individual player development. The insights gained from such analysis can inform tactical decisions, player recruitment, and match preparation.

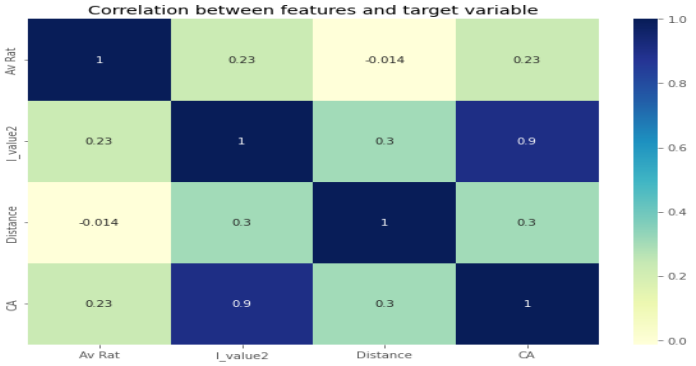


Figure 7: Correlation Matrix for Football Player Performance Prediction

Figure 7 depicts a correlation matrix that visualizes the strength and direction of relationships between various features and the target variable in football player performance prediction.
correlation matrix is a tabular representation of correlation coefficients between a set of variables. In this case, the variables are likely to be different performance metrics such as goals scored, assists, distance covered, average rating, and others. Each cell in the matrix provides a correlation coefficient that ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

Analysis: This matrix is essential for identifying which features have the strongest predictive power regarding player performance. For instance, a high positive correlation between 'goals scored' and 'average rating' would suggest that as one increases, so does the other. Conversely, a negative correlation might indicate an inverse relationship. By examining these correlations, analysts can select the most relevant features for building predictive models.

Application: The insights derived from this correlation matrix can inform coaching strategies, training focus areas, and player selection. It enables teams to leverage data-driven approaches to optimize performance and gain competitive advantages. Moreover, it aids in the development of robust predictive models that can forecast player performance based on historical data.

Significance: The use of a correlation matrix in performance prediction underscores the importance of statistical analysis in modern sports. It provides a foundation for objective decision-making and strategic planning, ensuring that subjective biases are minimized in the evaluation of player performance.

5.5 Model Analysis

Figure 8: Tuning Alpha Parameters to Optimize Predictive Models

R squared and normalized RMSE (Root Mean Square Error) figures in Figure 8 show how the alpha parameter affects prediction model accuracy.
The figure displays two graphs that show the evolution of R squared and normalized RMSE as functions of alpha. An improvement in model accuracy is indicated by the first graph, which shows a lower trend in normalized RMSE with rising alpha values. The second graph displays the variation of R squared, which appears to stabilize at higher alpha values and represents the percentage of the dependent variable's variance that can be predicted from the independent variables.

Analysis: By penalizing the higher coefficients in the model, the alpha parameter—a regularization term that helps prevent overfitting—can be fine-tuned with the help of these graphs. Through the examination of patterns within these charts, scientists can ascertain the ideal alpha coefficient that reduces inaccuracies and enhances explanatory strength, resulting in increasingly dependable and broadly applicable forecasts regarding the performance of football players.

Significance: In sports analytics, where the objective is to forecast player performance with high precision, alpha optimization is a critical procedure in predictive modelling. The significance of parameter adjustment in creating strong models that can successfully represent the intricacies of player performance measures is highlighted by this figure.

Application: Sports scientists and coaches can use the knowledge gathered from this analysis to inform their data-driven decision-making. It makes possible.

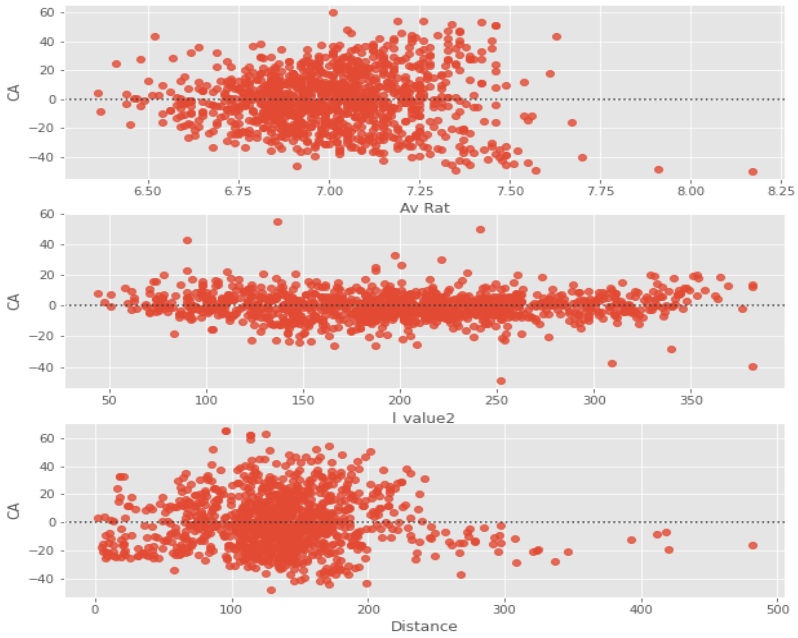


Figure 9: Scatter Plots Illustrating Football Player Performance Metrics

Figure 9 presents a series of scatter plots that correlate various performance metrics with football players' predicted performance outcomes.
The figure comprises three scatter plots, each representing a different metric's relationship with player performance. The top plot focuses on 'Average Rating' (Av Rat), displaying a dense cluster of data points around the median value, suggesting a common performance level among players. The middle plot examines 'Impact Value 2' (I value2), showing a linear trend that indicates a potential direct correlation with performance. The bottom plot, while not explicitly labelled, appears to analyze another performance-related variable, offering additional insights into player capabilities.

Analysis: Scatter plots are essential in topic model analysis as they visually demonstrate the relationships between variables. They allow researchers to identify patterns, trends, and outliers that may not be apparent in raw data. In this case, the plots provide a visual assessment of how different metrics, such as player ratings and on-field contributions, can predict overall performance. This is particularly useful in football analytics, where quantifiable data must be analyzed to forecast player success.

Significance: The inclusion of these scatter plots in predictive modelling is crucial for several reasons. Firstly, they help validate the accuracy of the predictive models by showing how well the selected metrics correlate with actual performance. Secondly, they enable the identification of the most influential factors affecting player performance, which can inform coaching decisions and training priorities. Lastly, they facilitate a deeper understanding of the data, leading to more nuanced and effective player performance predictions.

Application: By incorporating these visualizations into our analysis, we can enhance the predictive models used for forecasting football player performance. This allows for a more data-driven approach to player assessment, contributing to strategic planning, player development, and overall team success.

5.5 Model Evaluation

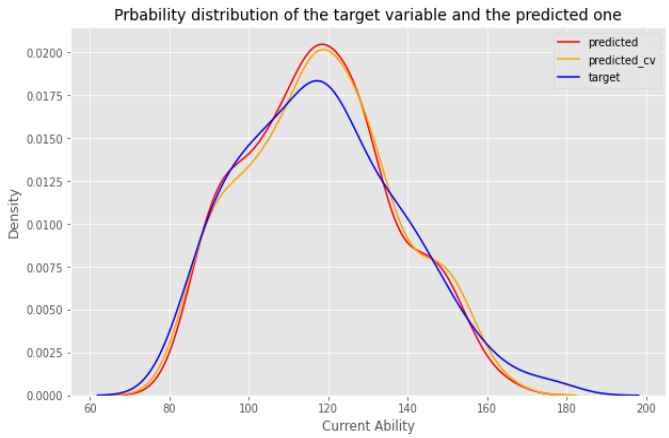


Figure 10: Probability Distribution of Predicted vs. Actual Player Abilities

Figure 10 depicts the probability distribution of predicted versus actual current abilities of football players, providing a visual evaluation of the predictive model's accuracy.

The graph presents a probability distribution curve, with two sets of data overlaid on each other. One set represents the actual current abilities of football players as observed in real-world performance data. The other set represents the predicted abilities generated by the performance prediction model. The closeness of the two distributions indicates the model's precision in forecasting player abilities.

Analysis: The graph is essential for evaluating the model's effectiveness. A high degree of overlap between the predicted and actual distributions suggests that the model can accurately capture the true performance levels of the players. Conversely, a significant divergence would indicate areas where the model may require refinement.

Significance: Accurate predictive models are crucial in professional football for scouting, training, and match preparation. By assessing the model's accuracy through such probability distributions, analysts can ensure that the insights derived from the model are reliable and actionable.

Application: The evaluation of predictive models using probability distributions enables teams to make informed decisions based on data-driven predictions. This can lead to more strategic player acquisitions, targeted training programs, and optimized game strategies, ultimately enhancing team performance and success.

	variables	VIF
0	const	1065.549527
1	Av Rat	1.063028
2	I_value2	1.170919
3	Distance	1.110553

Figure 11: Variance Inflation Factor (VIF) Analysis for Predictive Model Variables

Figure 11 displays the VIF values for variables within a predictive model, offering insights into the potential multicollinearity affecting the model's accuracy.
The table in Figure 11 lists the VIF values for four variables: 'const', 'Av Rat' (Average Rating), 'I_value2', and 'Distance'. The 'const' variable exhibits a significantly high VIF value (1065.549527), suggesting a strong correlation with other variables in the model, which may lead to unreliable regression coefficients. In contrast, the VIF values for 'Av Rat', 'I_value2', and 'Distance' are close to 1, indicating minimal multicollinearity and, therefore, a more reliable estimation of coefficients for these variables.

Analysis: VIF values are critical in diagnosing multicollinearity, which occurs when predictor variables in a regression model are correlated. High VIF values can inflate the variance of the coefficient estimates and may result in a less reliable model. By presenting the VIF values, we can assess the need for possible model adjustments or the removal of certain variables to improve the model's predictive power.

Significance: In the realm of football analytics, accurate predictions of player performance are essential for strategic decisions such as player selection, training focus, and game tactics. The VIF analysis helps ensure that the predictive models used are robust and not unduly influenced by multicollinearity, thereby enhancing their utility in performance prediction.
of predictive models in sports analytics.

Application: The findings from the VIF analysis can guide the refinement of predictive models, leading to more precise and actionable insights. This, in turn, can inform a data-driven approach to managing player performance and team strategy, ultimately contributing to the team's competitive success

VI. Conclusion and Future Scope

7.1 Conclusion

In conclusion, the convergence of sophisticated analytics, machine learning, and emerging technologies shows enormous potential for the future of football player performance prediction. Clubs, coaches, players, and fans may improve player development, injury prevention, tactical strategies, scouting, and fan engagement with the help of these predictive models. However, the acceptable use of these technologies in the football industry should be guided by ethical considerations and the preservation of the intrinsic unpredictability of the game. Football is predicted to become more data-driven and interesting for all parties involved as technology develops, both in terms of accuracy and usefulness of performance projections.

7.2 Future Scope

- Advanced Analytics for player statistics and biometric data
- Injury Prediction and Prevention
- Tactical Analysis for game strategies
- Player Scouting and Recruitment using data analysis

- Performance Optimization through personalized training
- Fantasy Sports predictions for enthusiasts
- Enhanced Fan Engagement with personalized content
- Real-time Analysis during matches
- Ethical considerations for data privacy and fairness
- AR/VR experiences for immersive fan engagement
- Player Mental Health monitoring and support

REFERENCES

[1] Arndt, C., and Brefeld, U. Predicting the future performance of soccer players. Statistical Analysis and Data Mining 9, 5 (Oct. 2016), 373–382.

[2] Arosha Senanayake, S. M. N., Malik, O. A., Iskandar, P. M., and Zaheer, D. A knowledge-based intelligent framework for anterior cruciate ligament rehabilitation monitoring. Applied Soft Computing 20 (July 2014), 127–141.

[3] Bartolucci, F., and Murphy, T. B. A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. Journal of Quantitative Analysis in Sports 11, 4 (Dec. 2015), 193–203. Publisher: De Gruyter.

[4] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (New York, NY, USA, July 1992), COLT '92, Association for Computing Machinery, pp. 144–152.

[5] Breiman, L. Random Forests. Machine Learning 45, 1 (Oct. 2001), 5–32.

[6] Brooks, J., Kerr, M., and Guttag, J. Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA, Aug. 2016), KDD '16, Association for Computing Machinery, pp. 49–55.

[7] Chen, T., and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA, Aug. 2016), KDD '16, Association for Computing Machinery, pp. 785–794.

[8] Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Machado Pereira,
A. C., and Nassis, G. P. Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Medicine - Open 5, 1 (July 2019), 28.

[9] Constantinou, A. C., Fenton, N. E., and Neil, M. pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems 36 (Dec. 2012), 322–339.

[10] Cortes, C., and Vapnik, V. Support-vector networks. Machine Learning 20, 3 (Sept. 1995), 273–297.

[11] De Mauro, A., Greco, M., and Grimaldi, M. A formal definition of Big Data based on its essential features. Library Review 65 (Mar. 2016), 122–135.

[12] Diana, B., Zurloni, V., Elia, M., Cavalera, C. M., Jonsson, G. K., and Anguera, M. T. How Game Location Affects Soccer Performance: T-Pattern Analysis of Attack Actions in Home and Away Matches. Frontiers in Psychology 8 (2017).

[13] Dixon, M. J., and Coles, S. G. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society. Series C (Applied Statistics) 46, 2 (1997), 265–280.

[14] Elmiligi, H., and Saad, S. Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (Jan. 2022), pp. 1–8.

[15] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. Knowledge Discovery in Databases: An Overview. AI Magazine 13, 3 (Sept. 1992), 57–57. Number: 3.

[16] Friedman, J. H. Greedy function approximation: A gradient boosting machine. ThenAnnals of Statistics 29, 5 (Oct. 2001), 1189–1232.

[17] Friedman, J. H. Stochastic gradient boosting. Computational Statistics & Data Analysis 38, 4 (Feb. 2002), 367–378.

[18] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. Time Series Analysis: Forecasting and Control, 5th Edition | Wiley, 2015.

[19] Gholamy, A., Kreinovich, V., and Kosheleva, O. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Departmental Technical Reports (CS) (Feb. 2018).

[20] Goddard, J. Regression models for forecasting goals and results in professional football. International Journal of Forecasting 21 (Apr. 2005), 331–340.