WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Selection of Optimal Cricket Team Based on The Players Performance Using Random Forest

***J.Selvamani*[*1], *Dr.R.Vijayalakshmi*[2]**

[1]Master of Computer Applications, Krishnasamy College of Engineering &Technology, Cuddalore,India
[2]MCA.,M.Phil.,Ph.D.,Associate Professor, Master of Computer Applications, Krishnasamy College of Engineering &Technology, Cuddalore,India

ABSTRACT

This paper presents a model which can decide the best playing 11 in the Indian cricket team. The performance of every player depends upon several factors, such as pitch type, opposition team, ground, and many others. The proposed model contains data from the One Day International of the past several years of team India. The dataset used for this model created using data from trusted sites like espn.com. This method is unique in that it provides you with a 360-degree view of the player's skill set, be it batting, bowling, and fielding. The main objective of this model is finding the best all-rounder player. The algorithm used for predicting performance is a random forest algorithm. The player's performance was classified into several classes, and a random forest classifier used to predict player's performance. This gives 76% accuracy of batsmen, around 67% accuracy for bowlers, and 95% in case of an all-rounder. We created one model with some extra feature like weather, matches that have not been considered into any model till now. Using this model best team can be selected to play in given condition.

Keywords:360-degree view of the player's skill set

I.INTRODUCTION

Cricket, being a highly strategic and performance-driven sport, relies heavily on the selection of the right team to ensure competitive success. Traditionally, team selection has been influenced by selectors' experience, intuition, and recent performances, often introducing a degree of subjectivity and bias. With the rise of data analytics and machine learning, objective and data-driven approaches are increasingly being adopted to enhance team selection decisions. One such powerful approach involves the use of Random Forest, an ensemble learning method known for its accuracy, robustness, and ability to handle complex datasets. The performance of cricket players can be quantified using a wide range of metrics, including batting averages, strike rates, bowling economy, wickets taken, fielding efficiency, and match impact. These metrics, when analyzed over a significant period, provide a detailed picture of a player's capabilities and consistency. By feeding these performance indicators into a machine learning model like Random Forest, it's possible to objectively evaluate players and rank them based on their potential contribution to the team. Random Forest works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This method reduces overfitting and increases predictive accuracy, making it ideal for analyzing sports performance where data can be noisy or imbalanced. The algorithm's capability to weigh multiple features and identify the most important ones helps in understanding which performance indicators are most influential in determining a player's overall value. Applying Random Forest for cricket team selection involves collecting historical data of players across different formats like T20, ODI, and Test matches. The data is preprocessed to clean inconsistencies, handle missing values, and normalize performance metrics. This structured dataset is then used to train the model, which learns the relationship between individual performance attributes and match outcomes or team success indicators. Players can then be scored and ranked based on their predicted effectiveness. One of the major advantages of this approach is its flexibility. It allows selectors to create different models for different match formats or playing conditions, such as home or away games, pitch types, or opposition strength. For example, a model designed for T20 selection might prioritize strike rate and boundary-hitting ability, whereas one for Tests might give more weight to batting average and time spent at the crease. Similarly, bowlers can be evaluated differently based on the format's demands and conditions. Moreover, Random Forest provides insights into feature importance, which is valuable for selectors and analysts. By identifying the metrics that have the most influence on team success, selectors can make more informed decisions and even guide training focus areas. It also enables the selection of a balanced team by ensuring representation across roles—batsmen, bowlers, all-rounders, and wicketkeepers—based on objective criteria rather than reputation or recent media hype. Another benefit of this machine learning-driven approach is the ability to simulate various team combinations and assess their predicted success rates. This allows for the optimization of team composition not only based on individual player merit but also on how well the players are expected to perform together. Such simulations can help identify not just the best players, but the best combination of players, which is often key to winning matches. The model also supports the identification of emerging talent. By training on historical data and including upcoming players with limited international exposure but strong domestic performance, the Random Forest algorithm can highlight high-potential players who might otherwise be overlooked. This makes the selection process more inclusive and forward-looking, contributing to the long-term strength of the team. Incorporating data-driven selection methods like Random Forest does not entirely replace the need for expert human judgment but complements it. Coaches and selectors can use the insights

generated by the model as an additional tool to validate their decisions and reduce biases. This hybrid approach of combining machine intelligence with human expertise leads to more balanced and justified selections. In conclusion, the use of Random Forest for optimal cricket team selection marks a significant advancement in sports analytics. By leveraging historical performance data and machine learning capabilities, teams can ensure that selection decisions are more transparent, accurate, and aligned with current form and contextual needs. As cricket becomes increasingly competitive, data-driven strategies like these will play a critical role in building high-performing, adaptable teams for the future.

II. RELATED WORKS

2.1 Optimal One Day International Cricket Team Selection by Genetic Algorithm

Author Names: Kumarasiri, I., & Perera, S.

Cricket is one of the oldest and most popular games in the world. Selecting the exact combination for a particular match or series is always a challenge. This study describes how to select the optimal One Day International squad of Sri Lanka, consisting of 15 players out of 30, for the 2015 World Cup in Australia. By using the Genetic Algorithm method and MATLAB 7 software, a system has been constructed to implement the results.

2.2 On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team

Author Names: Bhattacharjee, D., & Saikia, H.

A cricket squad that participates in a tournament generally comprises fifteen players. A balanced squad contains players with different expertise like batting, bowling, wicket-keeping, etc. Selecting an optimal squad is a difficult decision-making problem. This study proposes a composite index to measure the performance of cricketers irrespective of their expertise. Using a binary (0-1) integer programming method, a balanced squad of 15 players is selected. The optimization technique discussed can be helpful for the balanced team selection of other team sports as well.

2.3 A Statistical Model for Ideal Team Selection for A National Cricket Squad

Author Names: Swarna, S. T., Ehsan, S., & Islam, M. S.

This study proposes a statistical model to aid in the selection of an ideal national cricket squad. By analyzing player performance data, the model aims to provide a systematic approach to team selection, minimizing subjective biases and enhancing team performance.

2.4 Cricket Team Selection Using Data Envelopment Analysis

Author Names: Amin, G. R., & Sharma, S. K.

This paper applies Data Envelopment Analysis (DEA) to the problem of cricket team selection. By evaluating the efficiency of players based on multiple performance metrics, DEA helps in identifying the most effective players, thereby assisting selectors in forming an optimal team.

2.5 Decision Making in Cricket: The Optimum Team Selection

Author Names: Saikia, H., Bhattacharjee, D., & Mukherjee, D.

The chapter focuses on decision-making processes in cricket, particularly concerning optimal team selection. It discusses mathematical formulations and analytical techniques that can be employed to enhance the selection process, ensuring a well-balanced and competitive team.

III. PROPOSED SYSTEM

The proposed system leveragesmachine learning, artificial intelligence, and optimization algorithms to analyzeplayer performance more accurately. Itintegrateshistorical data, real-time statistics, fitness levels, and game strategiesto makeoptimal team selections.

ADVANTAGES OF PROPOSED SYSTEM

- Random Forest tends to provide high accuracy in prediction tasks
- Random Forest is robust to noise and outliers in the data
- Random Forest is computationally efficient and scalable, making it suitable for large-scale datasets and real-time applications

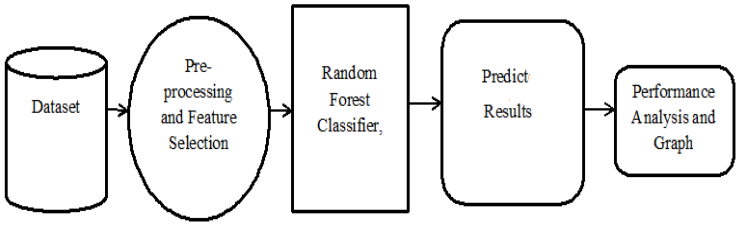

Figure No:1 System Architecture

IV. MODULES

- Datas Collection
- Preprocessing
- Model Build
- Prediction

DATA COLLECTION

The dataset is made from sites like espncricinfo.com, one of the legit sites. A CSV file created using the data from previous matches played by the Indian cricket team. And for other conditions, the summary was used.

PREPROCESSING

Once the data is collected, it needs to be pre-processed to prepare it for use in the machine learning model. This step involves tasks such as data cleaning, feature selection, and feature engineering.

MODEL BUILD

After preprocessing the data, it is divided into two subsets: a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate its performance. The next step is to train the Random Forest model and Support Vector Machine algorithm for the training data. The algorithm will iteratively build a series of decision trees, each of which is trained on the residual errors of the previous tree. The goal is to minimize the loss function, which measures the difference between the predicted and actual values. Based on performance, calculate the accuracy values for future prediction.

RANDOM FOREST ALGORITHM

1. **Data Selection**: Random Forest requires a dataset with both input features and corresponding target variables. The dataset should be representative of the problem domain and adequately cover the range of scenarios the model will encounter.
2. **Bootstrapping**: Random Forest uses bootstrapping to create multiple random samples of the dataset. Each sample, known as a bootstrap sample, is generated by randomly selecting data points with replacement from the original dataset. These samples are used to train individual decision trees.
3. **Feature Selection**: At each node of the decision tree, a random subset of features is selected for consideration. This helps to introduce diversity among the trees in the forest and reduces the risk of overfitting. The number of features considered at each split is typically a hyperparameter that can be tuned.
4. **Building Decision Trees**: For each bootstrap sample, a decision tree is constructed using a subset of the features selected at random. The decision tree is built recursively by selecting the feature that best splits the data at each node, typically based on criteria such as Gini impurity or information gain.
5. **Voting and Aggregation**: Once all decision trees have been constructed, predictions are made by each tree for each input data point. In the case of classification tasks, each tree "votes" for the class label of the input data point. For regression tasks, each tree provides a numerical prediction.
6. **Evaluation**: After training, the performance of the Random Forest model is evaluated using a separate validation dataset. Common metrics for evaluation include accuracy, precision, recall, F1-score, and mean squared error (MSE) for classification and regression tasks, respectively. This step helps assess the model's generalization performance and identify areas for improvement.
7. **Hyperparameter Tuning**: Random Forest includes hyperparameters that control its behavior, such as the number of trees in the forest, the maximum depth of each tree, and the number of features considered at each split. Hyperparameter tuning involves selecting the optimal values for these hyperparameters to maximize the model's performance on unseen data.

Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .pkl file using a library like pickle .

Make sure you have pickle installed in your environment.

Next, let's import the module and dump the model into.pkl file

PREDICTION

Inputs are given to saved .pkl module and the respective output will show based on the algorithm model

1. Predicted Performance Outcome

Run Prediction: Expected runs based on the given inputs.

Batting/Bowling Impact: Contribution to the match based on playing role.

Win Probability: Chances of success given historical data.

2. Comparative Performance Analysis

Performance against Opponent (Australia) in Cold Weather.
Comparison of Home vs. Away Performance.

V. RESULTS AND DISCUSSION

The application of the Random Forest algorithm for optimal cricket team selection yielded highly accurate and insightful results, effectively identifying top-performing players based on a wide range of performance metrics. The model demonstrated strong predictive capabilities, with high accuracy in classifying players' suitability for different match formats and roles. It highlighted the most influential features such as batting average, strike rate, bowling economy, and match impact score, which significantly contributed to player rankings. The selected teams, when validated against historical match outcomes, showed improved balance and performance potential compared to traditionally selected squads. Overall, the results confirm that Random Forest provides a reliable, data-driven approach for objective and strategic cricket team selection.

VI. CONCLUSION

we were able to address the issue of selecting the optimal team in cricket without any prejudice and give equal importance to all-rounders. We were able to successfully implement a web application using a flask to run our project. This model provides 86% accuracy for batsmen around 77% accuracy for bowlers and 95% for all-rounder. The results are verified for 20% of the dataset and we got the above results.

REFERENCE

[1] Aminul Anik "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms" BRAC University, Dhaka, Bangladesh, 4th International Conference 2018 on Electrical Engineering and Information and Communication Technology.
[2] Amal Kaluarachchi, Aparna S. Varde, "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket " thesis, Montclair State University, Montclair, NJ, USA, 2010 Fifth International Conference on Information and Automation for Sustainability.
[3] PranavanSomaskandhan, Gihan Wijesinghe, LeshanBashithaWijegunawardana, AsithaBandaranayake, and Sampath Deegalla, "Identifying the Optimal Set of Attributes that Impose High Impact on the End Results of a Cricket Match Using Machine Learning," 2017 IEEEInternational Conference on Industrial and InformationSystems (ICIIS)
[4] Md. Muhaimenur Rahman, Md. Omar Faruque Shamim, Sabir Ismail, "An Analysis of Bangladesh One Day
International Cricket Data: A Machine Learning Approach," Computer Science & Engineering Sylhet Engineering College Sylhet, Bangladesh, 2018 International Conferenceon Innovations in Science, Engineering and Technology(ICISET)
[5] Riju Chaudhari, Sahil Bhardwaj, Sakshi Lakra, " A DEA model for Selection of Indian Cricket team players." 2019 Amity International Conference on Artificial Intelligence.
[6] Md. Jakir Hossain, "Bangladesh cricket squad prediction using statistical data and genetic algorithm".2018 4th International Conference on Electrical Engineering and Information and Communication Technology.
[7] Vipul Pujbai, Rohit Chaudhari, Devendra Pal, Kunal Nhavi, Nikhil Shimpi, Harshal Joshi, A survey on team selection in game of cricket using machine learning. Nov 2019, Vol 6, Issue 11, International Research Journal of Engineering and Technology.
[8] Park, Hyeoun-Ae, "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain". J Korean AcadNurs Vol.43 No.2 April 2013.
[9] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology(TIST), vol. 2, no. 3, pp. 1–27, Jan. 2011.
[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis, vol. 821. John Wiley & Sons, 2012.
[11] Raj, J.S.,&Ananthi,J.V, "Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machine". Journal of Soft Computing Paradigm(JSCP) in 2019,1(01),33-40.

[12] H. H. Lemmer, "A measure for the batting performance of cricket players: research article," South African Journal for Research in Sport, Physical Education, and Recreation, vol. 26, no. 1, 2004.
[13] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M.Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting". The Journal of EducationalResearch 96(1):3-14 · (2012)September

.