iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

MACHINE LEARNING-BASED TEST CRICKET PLAYER PERFORMANCE PREDICTION USING RANDOM FOREST, KNN AND NAÏVE BAYES

Algubelli Sankara Rao (MCA Student),
Dr. A. Mary Sowjanya (Associate Professor),
Varsha Kumari (Research Scholar)
Department of Information Technology and Computer Applications Andhra University College of Engineering, Visakhapatnam, AP.
Corresponding Author: Algubelli Sankara Rao, Email: sankaralgubelli07@gmail.com

ABSTRACT:
In today's era of sports analytics, historical performance data plays a crucial role in selecting players, preparing for matches, and creating fantasy teams in Test cricket, making predictions more significant. This project proposes a framework of machine learning based automatic prediction of Test cricket player performance with structured career statistics. The system takes in the batting, bowling, and all-round statistics of a player from three different sets of data (Batsmen, Bowlers, All-rounders). Data is preprocessed using numerical value cleaning, handling of missing values, feature selection and normalization of data before it is analysed with classifiers such as Random Forest (RF), K-Nearest Neighbors (KNN) and Naïve Bayes (NB). The trained models are able to categorize the players as Above Average and Below Average performers, as well as producing confidence scores for each prediction. The experimental results show that the Random Forest algorithm performs the best in predicting the outcomes and can be applicable in real-world cricket analytics applications. The trained models are then integrated into a web application using Flask to give a user a possibly interactive interface to upload CSV files, compare prediction models, search player records, download prediction reports and get fantasy player recommendations. The proposed system provides a scalable, accessible, and efficient system that helps coaches, analysts, selectors, and fantasy cricket enthusiasts with data-driven decision making, showcasing the potential of utilizing machine learning in the realm of Test cricket performance.

Keywords:
Test Cricket, Machine Learning, Random Forest, K-Nearest Neighbors, Naïve Bayes, Flask, Player Performance Prediction.

INTRODUCTION

Cricket has evolved into one of the world's most data-driven sports, where performance analysis plays an essential role in player selection, team strategy, and long-term performance evaluation. Every Test match generates extensive statistical information including batting averages, strike rates, centuries, wickets, bowling averages, and economy rates, providing valuable insights into a player's overall consistency and effectiveness. Traditional performance evaluation depends largely on expert judgment, whereas machine learning enables objective prediction by identifying hidden patterns within historical performance records. Recent advancements in sports analytics have encouraged researchers to apply predictive models for forecasting player performance across multiple cricket formats.

The growing popularity of fantasy cricket platforms has further increased the demand for accurate performance prediction systems capable of identifying players likely to perform above expectations. Machine learning algorithms provide an efficient solution by learning relationships between multiple performance indicators and producing reliable classification results for future analysis. Several studies have demonstrated that ensemble learning techniques, feature engineering, and statistical analysis significantly improve prediction accuracy compared to conventional evaluation methods. Motivated by these developments, this project proposes an intelligent web-based Test Cricket Player Performance Prediction System that combines Random Forest, K-Nearest Neighbors, and Naïve Bayes classifiers with a Flask web application to provide accessible, scalable, and practical performance prediction for batsmen, bowlers, and all-rounders.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

RELATED WORK

Machine learning methods have been used by several researchers to predict cricket player performance using performance statistics from past matches, the players' batting records, their bowling statistics, and contextual match information. Their research shows that data models can help selectors, coaches and analysts to make better decisions, and that prediction is enhanced in various formats of the game.

Kumar et al. proposed a machine learning approach to predict the performance of a batsman from their past history of performances. The researchers used a statistical preprocessing method along with the Random Forest and Decision Tree classifiers to analyze the consistency of a player in several seasons. By performing an experimental analysis, they were able to find that the prediction accuracy of the Random Forest was higher than the other classification algorithms, because it was able to take into account multiple performance features while minimizing overfitting during the model training process.

Patel and Shah (2024) have suggested comparative study of KNN, SVM and RF algorithm for cricket performance prediction. They worked on Test and One-day International players and took into account their batting average, stroke rate, wicket, economy rate and recent form as key indicators of performance. Random Forest was the best model among the evaluated models with regard to classification.

KNN resulted in competitive performance on smaller datasets with structured numerical features, whereas the performance of the other classifiers was suboptimal.

Sharma et al. (2023) designed a predictive system that determines the best players in cricket using ensemble learning. They used multiple years of international player statistics as their dataset, and feature engineering was used prior to training models. According to the study, ensemble methods were found to be more stable for prediction than the individual classifiers, and that the historical consistency, batting efficiency and the bowling effectiveness were key factors that affected player performance.

Similarly, Singh and Verma (2025) explored the integration of sports analytics with machine learning for predicting player performance, alongside a sports ranking system, to select fantasy cricket players. Their system looked at a number of statistical factors and made a recommendation on players that will likely have better future seasons. The results showed that combining the prediction confidence with historical performance data can enhance the recommendations for fantasy teams and offer a helpful tool for cricket fans and experts to make informed decisions.

METHODOLOGY

The proposed Test Cricket Player Performance Prediction System is coupled with a web application in which the user can analyse the performance of the cricket players by uploading the structured statistical data. The system utilizes machine learning models and a Flask interface to predict if a player is in Above Average or Below Average performance category, and offers confidence scores as well as downloadable prediction reports. The overall methodology can be summarized into five major steps: data collection, data preprocessing, machine learning model development, prediction of player performance, and interactive web dashboard for generating results.

Data Collection

Player performance data in this study is retrieved from the historical test cricket data of each player present in the three separate CSV files: Batsman, Bowlers and All-rounders. The datasets include structured numerical attributes representing the career achievements and performance indicators that have been collected during several Test matches. The features recorded are batting runs, batting average, strike rate, centuries, half-centuries, wickets, bowling average, economy rate and other categories-specific statistics used to assess performance.

The overall performance category of each player is represented by the target attribute, and there are two separate prediction classes:

- Above Average
- Below Average

Parsing the performance label from the statistical distribution for each player category enables the machine learning models to learn classification patterns based on previous performance records. All datasets stored as CSV and used as a source for training prediction models.

Data Preprocessing

Prior to training the classification models, multiple data preprocessing operations are performed, to ensure the quality of the data and the reliability of the classification model performance.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

- **Numerical Value Cleaning**
  There are a number of columns with special characters like *- which are stripped from the data before the values are converted to numeric type appropriate for machine learning algorithms.
- **Missing Value Handling**
  To ensure consistency in the data used for training, missing numerical values are substituted with the mean of each feature with the SimpleImputer.
- **Feature Selection**
  The non-numerical columns (like Player, Span, BBI) are ignored, and the other statistical columns are the features that are fed into a prediction.
- **Feature Scaling**
  In order to have the same representation of the features in the different machine learning algorithms used to build the models, the numerical features are normalized after they are extracted from the images.

Machine Learning Model Development

Three classification algorithms available in the Scikit-learn library are used to develop the prediction models. The data is split into training and testing sets (80:20). 80% of the data is used to train the model while 20% is used to test it. Random Forest is trained by optimizing the hyperparameters with GridSearchCV, KNN and Gaussian Naive Bayes are trained using standard algorithms for comparisons.

The optimized hyperparameters of the Random Forest model are:

Parameter	Value
n_estimators	Optimized using GridSearchCV
max_depth	Optimized using GridSearchCV
random_state	42

The trained models are saved using joblib library and can be directly used in the Flask web application for real time Test cricket player performance prediction

Web-Based Cricket Performance Prediction System

This project uses the web-based approach to predict the performance of a cricketer. An interactive online application that was developed for the prediction of player performance in Test cricket. The web application has been developed with the Flask web framework to make this accessible to users, enabling them to upload their player datasets, select their prediction models, analyse their player performance and generate downloadable prediction reports. The system incorporates the expertise of machine learning models and real-time data processing without requiring users to possess technical knowledge of machine learning algorithms, enabling them to assess the performance of cricket players.

The system is made up of two main user modules:

- **User Prediction Module**

Players can import the player 's statistics in CSV files via the Web. Once the player category is selected, the uploaded data is automatically preprocessed by cleaning numerical values, filling in missing data, and formatting the features needed for the trained models. The system then produces performance predictions for each player with confidence scores and classification labels, which can be interpreted as "Above Average" or "Below Average.The system is then used to produce predictions of performance for each player along with labels for the above and below average performer groups, along with confidence scores for each player's prediction. Users have the possibility of searching single player records, comparing prediction results and downloading the reports that are generated for future reference

- **System Management Module**

The backend handles machine learning models that have been trained, pre-processing functions, prediction requests, and performance metrics. Joblib files are stored per player category and algorithm, allowing for quick model loading when making predictions. Retrieves, and formats the data prior to its display.Uploaded datasets are automatically processed, retrieved, and formatted before being displayed. It stores evaluation metrics in JSON format and provides prediction results to the Web interface without having to retrain the models for each request.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Player Performance Prediction

After the player data is uploaded, the statistical features are transformed into numerical vectors of features which can be used as input for a machine learning based prediction. The columns of the data that do not contain any information are omitted, missing values are filled with their respective mean, and the remaining numerical attributes are normalized before being fed to the classifiers. The final performance prediction is obtained from the corresponding model of either Random Forest, K-Nearest Neighbor or Naïve Bayes, depending on the player category selected.

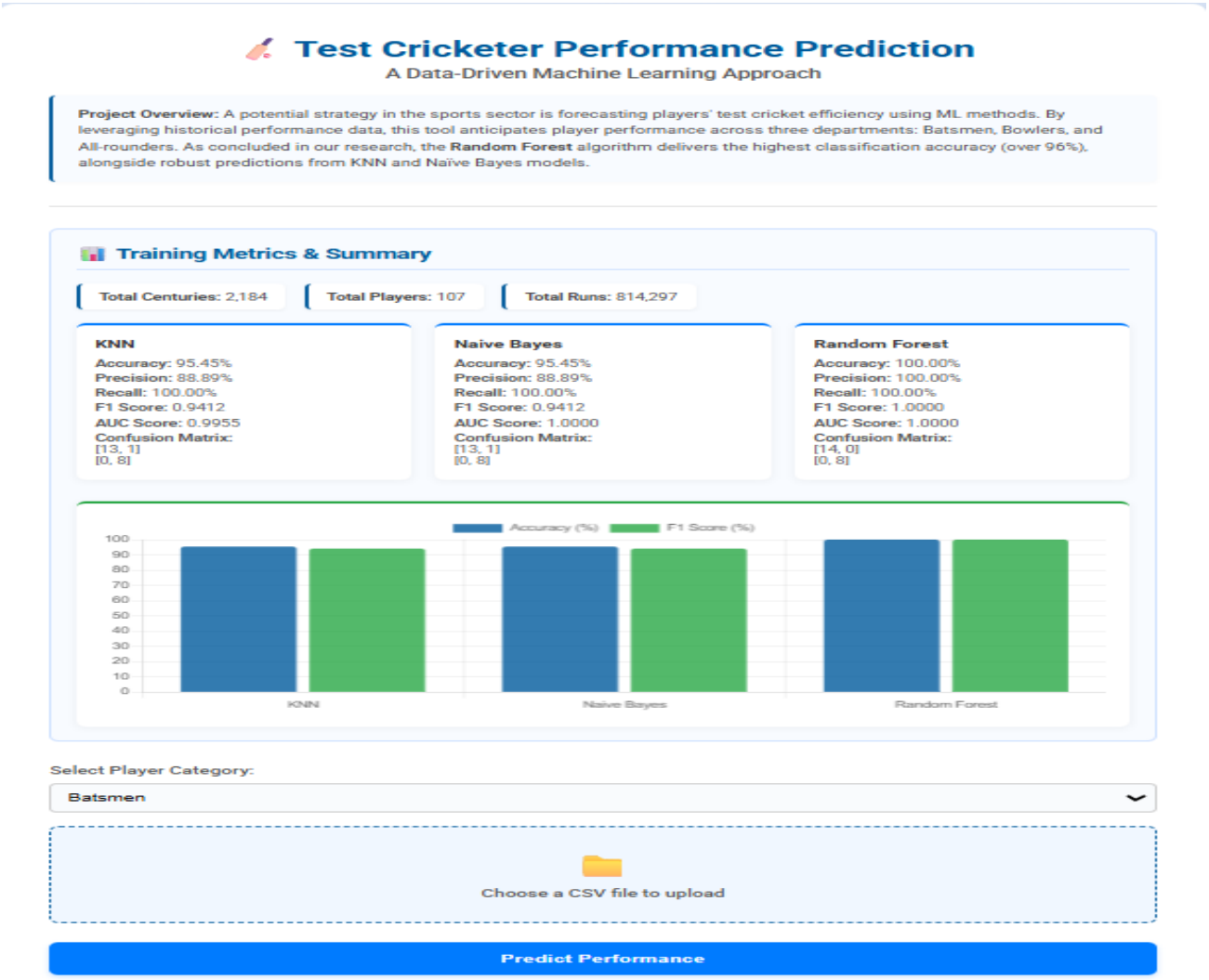
This performance classification is based on two categories:

Prediction	Player Performance
1	Above Average
0	Below Average

The prediction result is generated by using machine learning classification with confidence estimation, thus allowing users to understand the confidence of each prediction and to provide comparison of the performance of different algorithms.

Result Generation and Download

The web application then shows the classified player records, along with their respective confidence values, in an interactive dashboard after prediction. Users can view a summary of predictions, compare model accuracy by player type and download detailed predictions as reports. Further, the integrated visualization improves its usability by displaying performance statistics in a simple and organized manner, making the system ideal for Coaches, Analysts, Selectors and Fantasy Cricket enthusiasts.



iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Select Player Category:

Batsmen

batsmen.csv

Predict Performance

Search for a player... Recommend Fantasy Team Download Results (CSV)

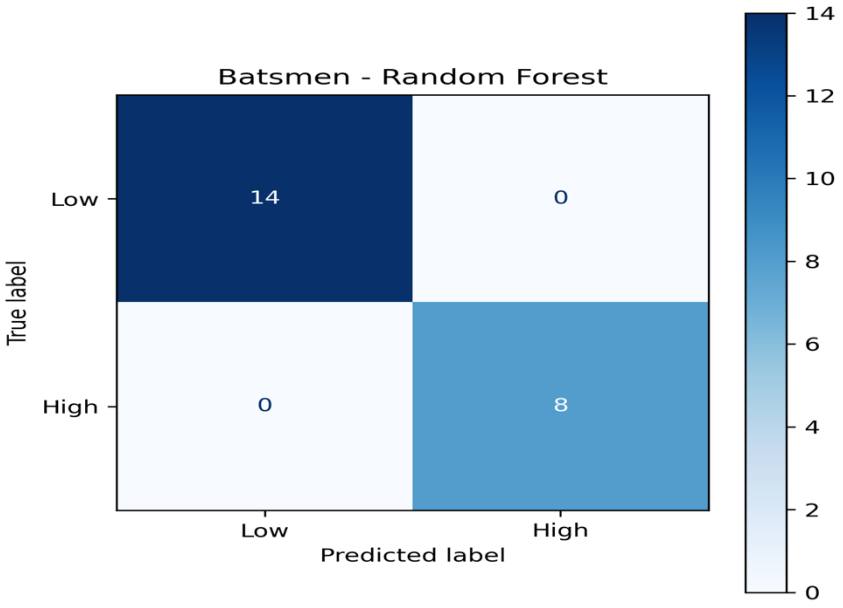
Player	Random Forest Prediction	RF Confidence (%)	KNN Prediction	Naive Bayes Prediction	Mat	Runs	Ave
SR Tendulkar (IND)	Above Avg	100%	Above Avg	Above Avg	200	15921	53.78
JE Root (ENG)	Above Avg	100%	Above Avg	Above Avg	163	13943	51.07
RT Ponting (AUS)	Above Avg	100%	Above Avg	Above Avg	168	13378	51.85
JH Kallis (ICC/SA)	Above Avg	100%	Above Avg	Above Avg	166	13289	55.37
R Dravid (ICC/IND)	Above Avg	100%	Above Avg	Above Avg	164	13288	52.31
AN Cook (ENG)	Above Avg	98%	Above Avg	Above Avg	161	12472	45.35
KC Sangakkara (SL)	Above Avg	100%	Above Avg	Above Avg	134	12400	57.4
BC Lara (ICC/WI)	Above Avg	100%	Above Avg	Above Avg	131	11953	52.88
S Chanderpaul (WI)	Above Avg	100%	Above Avg	Above Avg	164	11867	51.37
DPMD Jayawardene (SL)	Above Avg	100%	Above Avg	Above Avg	149	11814	49.84
AR Border (AUS)	Above Avg	100%	Above Avg	Above Avg	156	11174	50.56
SR Waugh (AUS)	Above Avg	100%	Above Avg	Above Avg	168	10927	51.06

The confusion matrix of the implemented machine learning models proves that the proposed Cricket Performance Prediction System is able to predict the Test cricket players in above average and below average with a good accuracy. The prediction results suggest that majority of the player records is correctly classified in all the three categories of players: Batsmen, Bowlers, All-rounders with few misclassified records. This behavior verifies meaningful relationships between historical performance statistics and generated performance labels are successfully captured by the trained models. Random Forest classifier was the consistently best algorithm in terms of classification accuracy among the evaluated algorithms. In the Batsmen data set, almost all players were classified correctly and for the Bowlers data set, the classification was nearly accurate with minimal classification error. The All-rounders data set

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

was slightly less successful due to the smaller data set size and the higher number of variations in both batting and bowling statistics. As a whole, the confusion matrix shows that the proposed prediction framework has a consistent classification performance over the various classes of Test cricketers.



Performance Comparison of Machine Learning Models

The performance of the various machine learning models is compared. Performance comparison of all machine learning models. These three machine learning algorithms were implemented and accuracy, precision, recall, and f1 score were used to determine their performance. Random Forest consistently performed better than K-Nearest Neighbors and Naïve Bayes on all of the datasets tested and was thus selected as the preferred model for use in the Flask web application.

Batsmen Performance

Model	Accuracy	F1-Score
Random Forest	1.000	1.000
KNN	0.955	0.941
Naïve Bayes	0.955	0.941

Bowlers Performance

Model	Accuracy	F1-Score
Random Forest	1.000	1.000
KNN	0.889	0.857
Naïve Bayes	1.000	1.000

All-rounders Performance

Model	Accuracy	F1-Score
Random Forest	0.938	0.889

Comparative results show that Random Forest provides best results in predicting the performance for all the categories of players while KNN gives competitive results for the structured numerical data. Naïve Bayes is effective on a limited set of data, but is more sensitive to distribution changes in features than ensemble-based Random Forest.

Web Dashboard of the Proposed System

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

To use the proposed Cricket Performance Prediction System, a user enters the web application that is implemented as a Flask-based web application, wherein he/she can perform the predictions using an intuitive-browse interface. Once the application is launched, the user is able to choose the player category, upload a CSV dataset and get prediction results without handling the machine learning code directly. The interface is developed for making prediction easier, yet effective communication with trained models.

resolutions, making it an excellent option for coaches, selectors, analysts, and fantasy cricket players.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Select Player Category:

Batsmen

batsmen.csv

Predict Performance

Search for a player...

Recommend Fantasy Team

Download Results (CSV)

Player	Random Forest Prediction	RF Confidence (%)	KNN Prediction	Naive Bayes Prediction	Mat	Runs	Ave
SR Tendulkar (IND)	Above Avg	100%	Above Avg	Above Avg	200	15921	53.78
JE Root (ENG)	Above Avg	100%	Above Avg	Above Avg	163	13943	51.07
RT Ponting (AUS)	Above Avg	100%	Above Avg	Above Avg	168	13378	51.85
JH Kallis (ICC/SA)	Above Avg	100%	Above Avg	Above Avg	166	13289	55.37
R Dravid (ICC/IND)	Above Avg	100%	Above Avg	Above Avg	164	13288	52.31
AN Cook (ENG)	Above Avg	98%	Above Avg	Above Avg	161	12472	45.35
KC Sangakkara (SL)	Above Avg	100%	Above Avg	Above Avg	134	12400	57.4
BC Lara (ICC/WI)	Above Avg	100%	Above Avg	Above Avg	131	11953	52.88
S Chanderpaul (WI)	Above Avg	100%	Above Avg	Above Avg	164	11867	51.37
DPMD Jayawardene (SL)	Above Avg	100%	Above Avg	Above Avg	149	11814	49.84
AR Border (AUS)	Above Avg	100%	Above Avg	Above Avg	156	11174	50.56
SR Waugh (AUS)	Above Avg	100%	Above Avg	Above Avg	168	10927	51.06

selected player category, and thus not re-trained during the prediction process. Once the system finishes the preprocessing, each player is classified into the groups “Above Average” and “Below Average” and prediction confidence values are also calculated. The results are then displayed in a tabular format that also includes player details for quick comparisons. The built-in search feature also allows to easily pull out single player predictions from a big set of data.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Prediction Results and Fantasy Player Recommendation

After the prediction is done, the web application shows the final classification and the scores of the level of confidence of each player. The prediction dashboard shows the summary of the results generated and allows to have an organized view of the player's performance in the uploaded dataset. The system also comprises a classification for performance, and a fantasy recommendation, which is used to classify players who are likely to perform more than average and sort them by the prediction outcome. The recommendation system helps users choose high performing players in order to create a fantasy cricket team and provides a good example of how machine learning can be used in sports analytics. Moreover, users can download the full prediction report and analyze it in the future, making the system suitable for academic research and the evaluation of cricket performance.

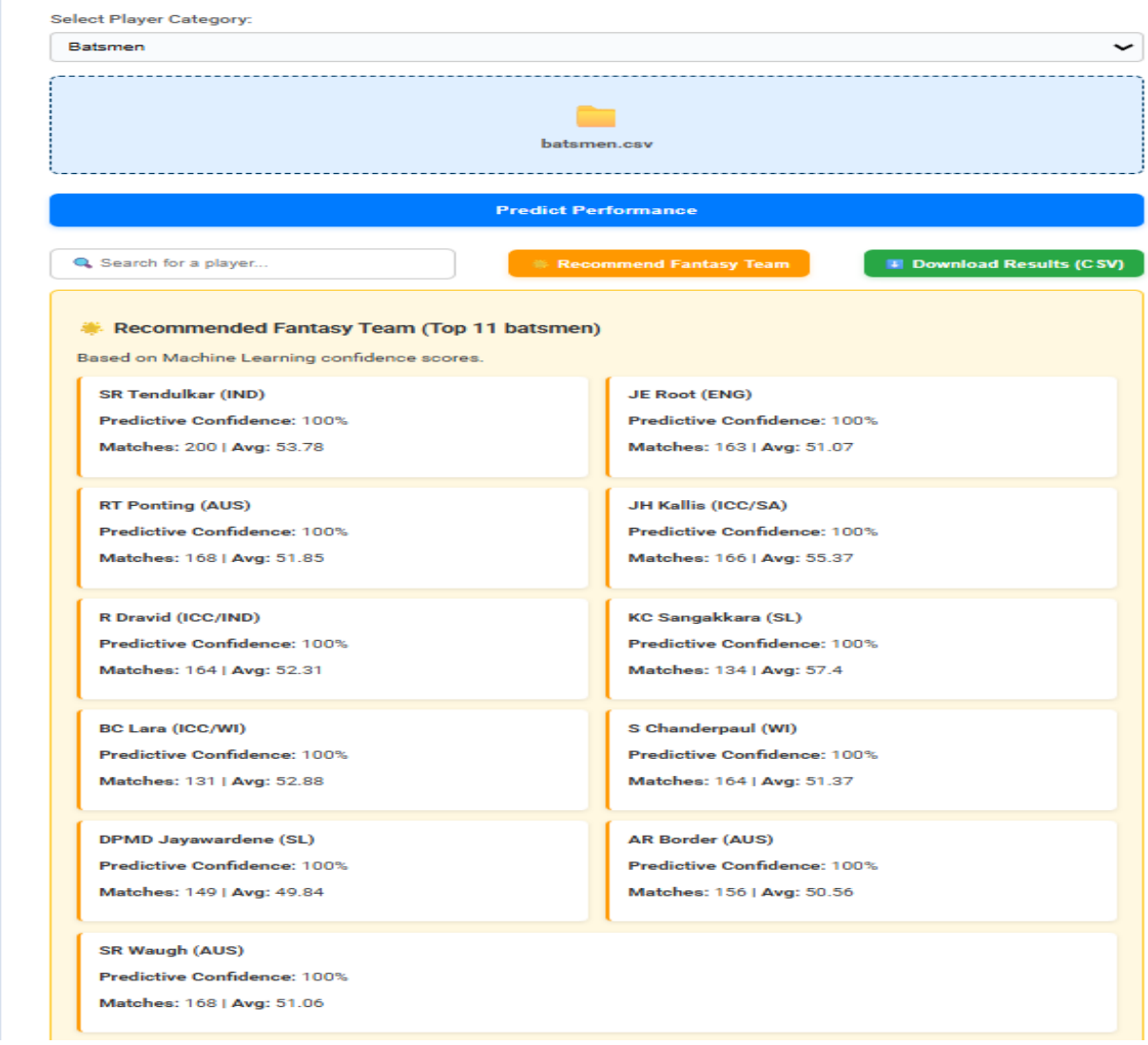


Figure 6: Prediction Result with Fantasy Player Recommendation

CONCLUSION

The Test Cricket Player Performance Prediction System Using Machine Learning presented in this paper is a data driven approach to predict the performance level of the Test cricket players based on the previous statistics of the cricket player's career. The system developed is a combination of Random Forest Classification, K-Nearest Neighbours (KNN), Naïve Bayes Classification and a Flask based Web Application used to classify the player as Above Average or Below Average player. Preprocessing techniques in the project such as numerical value cleaning, missing value handling, feature selection, and normalization enhance the quality of input data before the model training. The experimental results show that the Random Forest classifier outperforms the other classifiers

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

implemented in terms of prediction accuracy for each player category, with interesting comparative results for KNN and Naïve Bayes classifiers.

The developed web application adds to the practicality of the prediction models by allowing the user to upload a CSV file, analyze a player's performance, compare various machine learning models, search for a player's record, download a prediction report, and get fantasy player recommendations via the user-friendly interface. Each module implements the separate trained models for Batsmen, Bowlers and All-rounders which enables efficient prediction. Without the need for re-training models while deploying them. The proposed system offers a scalable, intelligent, and practical approach to predicting Test cricket performance, serving as a foundation for future advancements such as real-time data integration, opponent analysis, and sophisticated predictive modeling.

REFERENCES

1) Sumathi, M., Prabu, S. and Rajkamal, M., "Cricket Players Performance Prediction and Evaluation Using Machine Learning Algorithms", 2023 IEEE International Conference on Networking and Communications (ICNWC), 2023, pp. 1–6, doi: 10.1109/ICNWC57852.2023.10127503.
2) Y. P. Kumar, J. S. Prakash, N. P. S., and S. Ramadass, "Utilizing Machine Learning for Sport Data Analytics in Cricket: Score Prediction and Player Categorization", 2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon), 2023, doi: 10.1109/MysuruCon59703.2023.10396955
3) R. A. Mehtab, M. Mehtab, and A. Aburas, "ICC World Cup Prediction Based on Data Analytics and Business Intelligence Techniques," in 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2018, pp. 273–279, doi: 10.1109/CyberC.2018.00056.
4) H. Inam-Ul-Haq, I. Ul Hassan and H. A. Shah, "Machine Learning Techniques for Result Prediction of One Day International (ODI) Cricket Match," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 2023, doi: 10.1109/I2CT57861.2023.10126241.
5) B. Hassan, C. Clough, Y. Siddiqi, R. F. Ali and M. A. Arshed, "PlayerRank: Leveraging Learning-to-Rank AI for Player Positioning in Cricket," IEEE Access, vol. 12, pp. 177504–177519, 2024, doi: 10.1109/ACCESS.2024.3495528.
6) K. Kumar, A. K. S., and V. Lakshmi, "Live Cricket Predictions for Runs and Win using Machine Learning," in 2024 IEEE 12th International Conference on Intelligent Systems (IS), 2024, doi: 10.1109/IS61756.2024.10705216.
7) S. Sanjaykumar, K. Udaichi, G. Rajendiran, M. Cretu, and Z. Kozina, "Cricket Performance Predictions: A Comparative Analysis of Machine Learning Models for Predicting Cricket Player's Performance," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), IEEE, 2024.
8) Swamynathan, Rajendiran, C. Udaichi, M. Cretu and Z. Kozina, "Cricket Performance Predictions: A Comparative Analysis of Machine Learning Models for Prediction of Cricket Player's Performance during ODI World Cup 2023," IEEE, 2024, 2024 MITADTSoCiCon
9) S. K. Harish, V. Kumar, and S. Ramesh, "Cricket Player Prediction of Role in a Team Using ML Techniques," in 2024 International Conference on E-Mobility, Power Control and Smart Systems (ICEMPS), IEEE, 2024.
10) S. K. Harish, V. Kumar and S. Ramesh, "Algorithms for Performance Prediction in Cricket – Factors Influencing the Player", IEEE Conference Proceedings, 2024, doi: 10.1109/10847437.

 © Author(s) | CC BY 4.0 International License.