Machine Learning Based Selection of Optimal Sports team based on the Players Performance

M R Kavya[1], Sai Akshay R[2], Rose Priyanka[3], Pavan N[4], Ravi Ranjan Kumar Singh[5]

[1,2,4,5] Students, Department of Computer Science & Engineering, T John Institute of Technology, Bengaluru, India
[3] Asst. Professor, Department of Computer Science & Engineering, T John Institute of Technology, Bengaluru, India

---***---

ABSTRACT – *In this essay discusses a methodology that can choose the Indian cricket team's best starting 11. Each player's performance is influenced by a number of Variables, including the ground, the opponent's team, and the type of pitch. Data from Team India's recent One Day International appearances are included in the suggested model. This model's dataset was developed utilising information from reliable websites like espn.com. This approach is unique in that it provides a complete picture of the player's skill set, including batting, bowling, and fielding. The best player must be identified as a key component of this strategy. Using the random forest approach, performance can be predicted. A random forest classifier was used to forecast the player's performance after classifying the player's performance into different categories. This model predicts batters with a 76 percent accuracy, bowlers with a 6 7 to 69 percent accuracy, and allrounders with a 95 percent a ccuracy. A model is created with a few extra features, such as weather and played matches, which are not included in any other models. This methodology enables the best team to be chosen to compete under certain circumstances.*

Key Words: **All-rounder, Random forest, Player selection, Team selection, Cricket, Machine Learning, SVM, Logistic Regression**

1. INTRODUCTION

As machine learning aims to focus on larger and complex tasks, the importance of providing a potential amount of relevant data as become the crucial part of the field[1]. The second most viewed television programme is cricket. In southeast Asian nations like India, Pakistan, Bangladesh, and Sri Lanka, this sport is incredibly popular. The choice of playing 11 is currently one of the key challenges. Due to the Indian team's recent success, cricket is loved by zealots of all ages and socioeconomic backgrounds. However, there are many misunderstandings regarding squad combinations before to a game, such as which player to pick or bench for the forthcoming match, which batter to place in which position, or which bowler to choose. The proposed model intents to predict the run of a batsman he would score in the next match and also runs a bowler might give in the next match[1].In a variety of sports, machine learning is utilised to forecast the outcomes of matches [2]. From this, the desire to assess each player's performance in a particular match and choose the best starting 11 arose. This uses a technique called as Data Envelopment Analysis (DEA), to calculate the efficiency of players based on their past performances and selects the best ones to form a winning team[5].

2. LITERATURE REVIEW

Research has been conducted in the past to learn more about this subject in-depth, and the studies devoted to it are covered in length below.

[1] In order to determine a player's performance using SVM, Aminul Islam Anik et al. Suggested considering the balls faced, ground, pitch, opposition, and position. This model accurately analyses many variables and how they affect batsmen's ability to score runs. The number of balls that players confront is the main determinant, but it is problematic since it is impossible to determine a player's total number of shots before the game. The analysis for all-rounders has been left out, despite the paper having done the appropriate amount of work on bowlers and batters.

[2] Amal Kaluarachchi employed Bayesian classifiers in Machine learning to forecast how elements like home field advantage will impact the game's outcome. Using this concept, one of the variables that impacts how well the players perform is the home and away performance.

[3]Pranavan Somaskandhan et al.; used machine learning to examine the collection of factors that have a significant impact on a game' s outcome. When the attributes of high individual wickets, number of bowled deliveries, number of thirtys, total wickets, wickets in the power play, runs in death overs, dots in middle overs, number of fours, and singles in middle overs were combined, the maximum accuracy was

attained. The property, as previously noted, provided an SVM accuracy of 81 percent.

[4] Data from the Bangladesh One Day International Cricket was examined by Md. Muhaimenur Rahman. The study was separated into three parts: before the game began, after one inning, and after the loss of wickets. When they used Decision Tree, their accuracy was 63.63 percent at the start of the game, 72.72 percent in the first and second innings, 81.81 percent, and 80 percent and 70 percent for fall of wicket analysis.

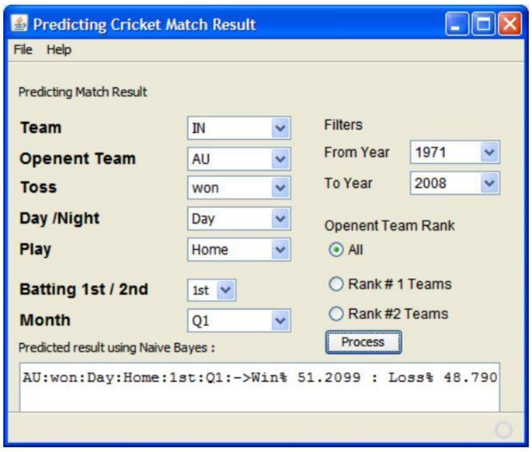
[5] To assess the effectiveness of players, Riju Chaudhari et al. employed a DEA (Data Envelopment Analysis). The player's performance in test matches is documented in the article. Due of the possibility of circumstances like match-fixing in T20. Every player may have to bat during a test series, thus the bowler with the highest batting strike rate is given preference. This work stands out from others since it takes a novel approach rather than directly utilising machine learning techniques.

[6]Md. Jakir Hossain chose the best players from among the 30 members of the Bangladesh cricket team using a genetic algorithm. In order to select elite players, the article combines statistical analysis with a genetic algorithm. Out of the 30C14 total solutions, each potential solution was treated as a chromosome. Using a statistical approach, player ratings were taken into consideration. The final fitness score is determined by taking into account variables such as the average player rating, the number of bowlers, batters, allrounders, and wicket keepers, the number of quick and spinners, and the number of right- and left-handed players.

[7] Vipul Punjabi and his team utilised the naive Bayes class ifier to estimate how many runs the batsman will score and how many wickets the bowler would take. Different categories are used to group runs scored and wickets takes IPL match records were utilised to create the dataset for this. This study uses a surprisingly small number of input features, which greatly limits the model's potential.

3. PROPOSED METHODOLOGY

The dataset was created using information from previous ODI matches and is covered in more detail in the following section of this paper. The performance of batsmen is categorised into different groups based on the amount of runs they score, much as the number of wickets a bowler claims is categorised into different groups. In order to design defensive strategy for the players' upcoming game, reverse data mining technique is employed to identify the weak points of the other teams. The CricAI software tool was created using the findings of our in-depth examination of actual data. Due to the fact that Bayesian classifiers produced the greatest results in our investigation, it is a Java implementation of Naive Bayes.



Predicting Cricket Match Result

Predicting Match Result

Team: IN
Openent Team: AU
Toss: won
Day /Night: Day
Play: Home
Batting 1st / 2nd: 1st
Month: Q1

Filters
From Year: 1971
To Year: 2008

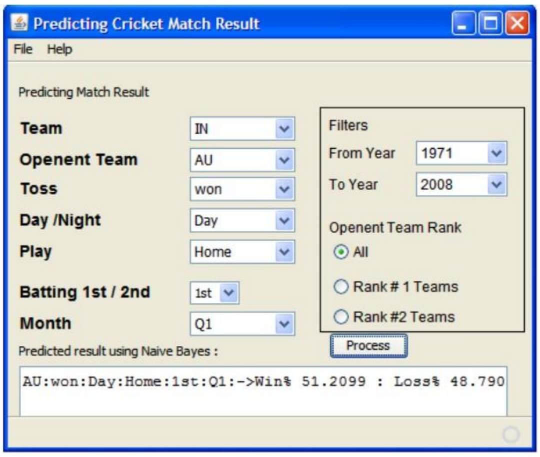
Openent Team Rank
● All
○ Rank # 1 Teams
○ Rank #2 Teams

Predicted result using Naive Bayes :

```
AU:won:Day:Home:1st:Q1:->Win% 51.2099 : Loss% 48.790
```

Fig 1. Snapshot of CricAI

In order to get better results, the CricAI tool has certain filters. Currently, the talent levels of all international teams are nearly equal. Therefore, winning a game may depend on the factors and choices made by the individuals (such as choosing to bat first and winning the toss). However, certain teams were far more talented than the rest in the early stages of the game right after the game was launched in the 1970s. Consequently, adding those matches in the calculation could have a negative impact on the outcomes. This disadvantage is mitigated by the Year filter, as seen in Figure 2 [3].



Predicting Cricket Match Result

Predicting Match Result

Team: IN
Openent Team: AU
Toss: won
Day /Night: Day
Play: Home
Batting 1st / 2nd: 1st
Month: Q1

Filters
From Year: 1971
To Year: 2008

Openent Team Rank
● All
○ Rank # 1 Teams
○ Rank #2 Teams

Predicted result using Naive Bayes :

```
AU:won:Day:Home:1st:Q1:->Win% 51.2099 : Loss% 48.790
```

Fig 2. Using filters in CricAI

Selecting the opponent's rank is another filter. The team rankings are determined by the International Cricket Council (ICC). This aids in improving judgement while determining the match winning criterion. For instance, against Rank 2 teams like Canada and the Netherlands, Sri Lanka has never lost a game.

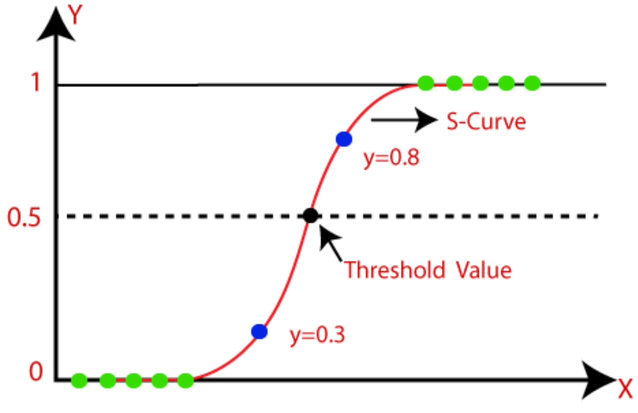


Fig 3. Logistic Function

4. ALGORITHMS AND TECHNIQUES

LOGISTIC REGRESSION : Binary classification tasks frequently use the logistic regression method. The softmax function is employed in the case of multi-class classification in place of the sigmoid function. [8] G(z)=1/(1+ez) is the formula for the logistic regression's hypothesis function.

SUPPORT VECTOR : The hyper plane assists in distinguishing between several classes in a support vector classifier. In order to separate non-linear data, several kernels can be used to map them to higher dimensions [9]. The data could be successfully classified by many hyperplanes. The hyperplane that best depicts the greatest difference or margin between the two classes is a logical option. As a result, the hyperplane is selected to maximise the distance from it to the closest point.

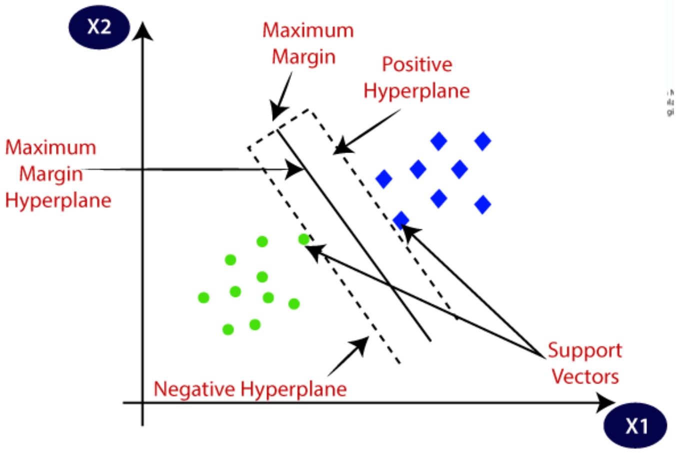


Fig 4. SVM Hyperplane

RANDOM FOREST : An ensemble learning technique called Random Forest uses the combined output of many decision trees to learn classification, regression, and other tasks. Random forests are used to naturally order the relevance of variables in a classification or regression issue. According to the graph of the dataset supplied below, the majority of batsmen score only a little number of runs.

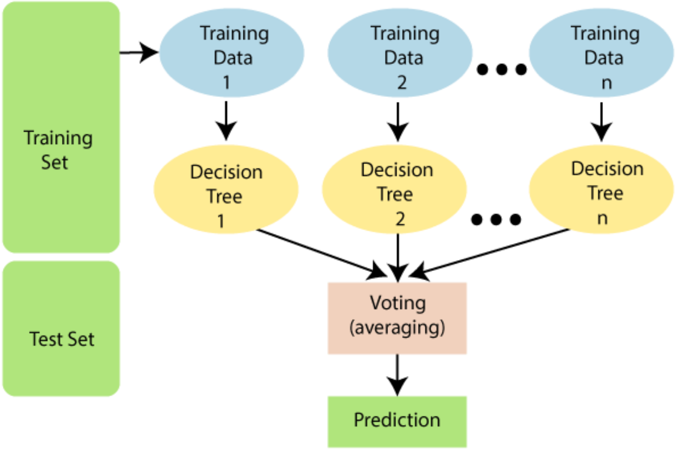


Fig 5. Random Forest Model

5. DATA DESCRIPTION

DATA COLLECTION: The dataset was created using legitimate websites like espncricinfo.com. A CSV file that was produced utilising information from Indian cricket team's prior games. And the summary was applied to various circumstances.

FEATURE SELECTION: Opponents, Runs Scored, Strike Rate, and Overall Average were among the features that were included in the models that had previously been developed for choosing the best team. The following characteristics are taken into account while evaluating a player' s performance. Position, matches played, runs, strike rate, ground, home-away, 50s, 100s, overall average, pitch, opposition, and weather are batting attributes.

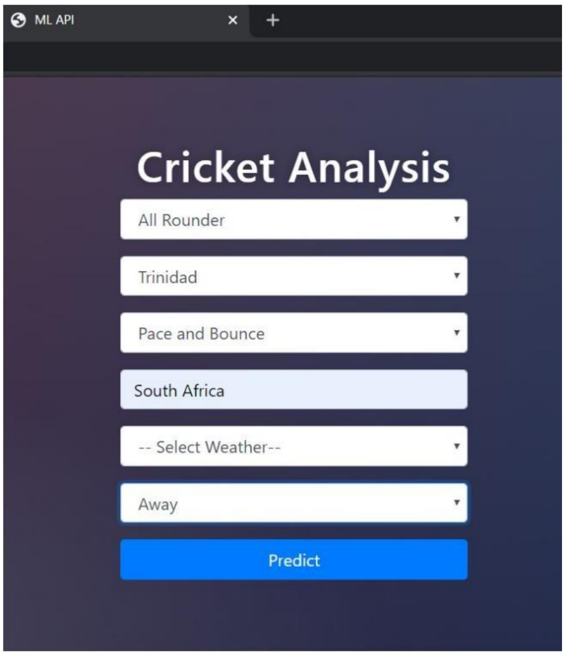
BOWLING ATTRIBUTES: Matches played, Wickets, Average, Economy, Strike Rate, Ground, Pitch, Opponent, Weather, and Home Away are bowling attributes. Matches played, wickets, runs scored, strike rate, average, ground, homeaway, opponent, weather, and pitch are all aspects of an all-rounder.

The accuracy of these models was insufficient because they took only few features into account. As a result, a model is created with some additional aspects, such as matches played, the field, the weather, and so forth.

These models solely took into account bowlers and batters while constructing an 11-player squad. However, as compared to bowlers and batsmen, all-rounders will always receive lower scores. All-rounders are therefore taken into account when developing the model to produce an official team.

6. IMPLEMENTATION

Utilizing the Flask API for implementation. The model is trained in a Jupyter notebook using the python scikit-lean library's available methods. The screenshot of our implementation and a representative outcome are shown below. The same input parameters are used for all three types of players: bowlers, allrounders, and batsmen.



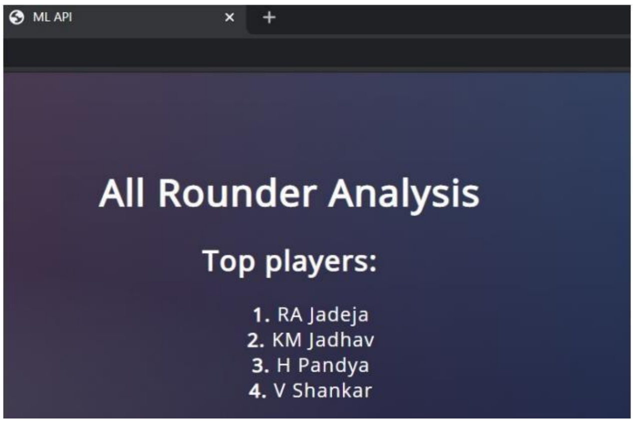
Cricket Analysis

All Rounder
Trinidad
Pace and Bounce
South Africa
-- Select Weather--
Away

Fig 6. Input Display

The sample output on clicking the predict button is given below.

Several algorithms were tried to predict the outcomes using the Python scikit learn module. Different methods were employed to predict the classes, including decision trees, logistic regression, SVM classifiers, and random forests, with random forests producing the best outcomes.

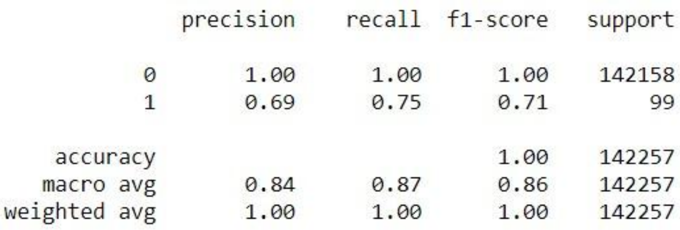


All Rounder Analysis

Top players:

1. RA Jadeja
2. KM Jadhav
3. H Pandya
4. V Shankar

Fig 7. Output display for All-rounders



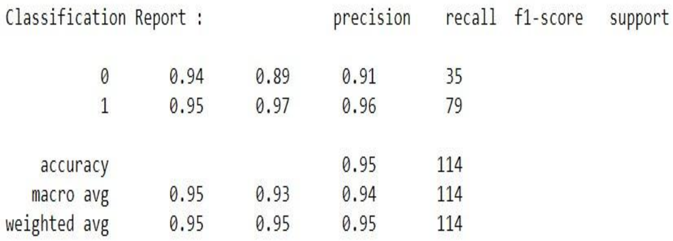
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    142158
           1       0.69      0.75      0.71        99

    accuracy                           1.00    142257
   macro avg       0.84      0.87      0.86    142257
weighted avg       1.00      1.00      1.00    142257
```

Fig 7.1 . Random Forest Classification report

The support vector classifier gives following report



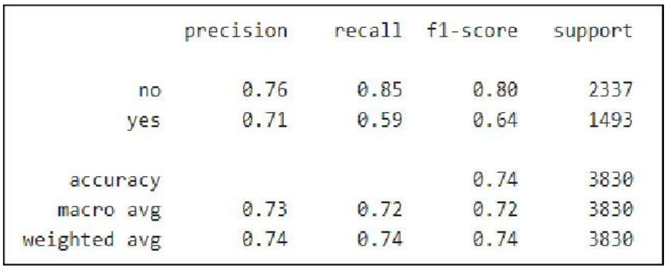
```
Classification Report :              precision    recall  f1-score   support

           0       0.94      0.89      0.91        35
           1       0.95      0.97      0.96        79

    accuracy                           0.95       114
   macro avg       0.95      0.93      0.94       114
weighted avg       0.95      0.95      0.95       114
```

Fig 7.2. Support Vector Classification report

While the results using the logistic regression algorithm were the poorest



```
              precision    recall  f1-score   support

          no       0.76      0.85      0.80      2337
         yes       0.71      0.59      0.64      1493

    accuracy                           0.74      3830
   macro avg       0.73      0.72      0.72      3830
weighted avg       0.74      0.74      0.74      3830
```

Fig 7.3. Logistic Regression Classification Report

Therefore, the random forest method is used to continue the process.

7. CONCLUSION

With knowledge about the number of dots, remaining over , remaining wickets, and strike rate, this analysis could be conducted in between innings of play, which would enable players to make better-informed decisions about where to play. Due to the fact that these variables affect the game's outcome in milliseconds, significant research might be done to create a model that is advantageous. A software tool called CricAI is developed, which can be used in real world scenarios to predict the chances of victory in a given match

using attribute and filters. Comparison of machine learning methods showed that the best method for solving the issue is classification. Testing different classifiers on actual data demonstrated that Naive Bayes performs well for the relevant datasets.

8. ACKNOWLEDGEMENT

We sincerely appreciate Ms. Suma R, the department head of computer engineering, as well as our project coordinators Dr. John T. Mesia Dhas and Ms. Rose Priyanka for their inspiration, ongoing support, and insightful suggestions during our research on "Machine Learning Based Selection of Optimal Sports team based on the Players Performance", which would have seemed challenging without them.

9. REFERENCES

[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] Pranavan Somaskandhan, Gihan Wijesinghe, Leshan Bashitha Wijegunawardana, Asitha Bandaranayake, 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 IEEE International Conference on Industrial and Information Systems (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 Conference on 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] 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.

[9] 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.

[10] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll , “An Introduction to Logistic Regression Analysis and Reporting”. The Journal Of Educational Research 96(1):3-14 (2012) September.