CRICKET TEAM PREDICTION USING MACHINE LEARNING TECHNIQUES

Nilesh M. Patil[1], Bevan H. Sequeira[2], Neil N. Gonsalves[3], Abhishek A. Singh[4]
[1,2,3,4]Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, India.
[1]nilesh.patil@fragnel.edu.in, [2]bevan.sequeira@gmail.com
[3]neilgonsalves.ng@gmail.com, [4]abhishek48298@gmail.com

Abstract*— *Player selection is an essential task for any sport and similarly for the game of cricket as well. The players' performance varies on various factors. The team management and the captain selects eleven players for each match from the entire squad. Review of different attributes and players' results is considered to pick the best eleven players. By scoring runs each batsman contributes and each bowler contributes by taking wickets and awarding minimum runs. This project aims to predict team success based on the player's past records. Acquisition of the players 'results individually and their contribution to the team i.e. Best batting performance among the batsmen available, best bowling performance among the available bowlers and best all-rounder performance will be a great help in selecting the eleven players. We used the Random Forest Algorithm and Decision Tree classifiers to produce the problem's prediction models. It was found that the Random Forest classifier is the most reliable for the problems proposed.

Keywords*— *Random Forest, Prediction, Decision Trees, Classifier.

I. INTRODUCTION

Selecting the best players for any match in any sport is nothing but predicting the players' performance. Each individual player performs differently for every match. Player selection is very important in the game of cricket as exactly 11 players are selected at the beginning of the match and are fixed for the entire match unless in case of injury. However, the substituted players in such cases have limitations. Players' performance can be predicted by analyzing their past records. Cricket players' abilities and performance is measured in different ways. Batsmen are measured on the basis of batting average, batting strike rate, number of centuries and so on, while bowlers are measured on the basis of bowling average, bowling strike rate, economy rate etc. The players' recent results, however, is a key factor considered during the selection process. Batsman / bowler are also taken into consideration for forecasting his success in the upcoming match at a given venue.

II. LITERATURE SURVEY

Muthuswamy and Lam [1] anticipated Indian bowlers' success against seven international teams that the Indian cricket community plays most frequently against. They used back propagation group and radial base network feature to predict how many runs a bowler is likely to acknowledge and how many wickets a bowler is likely to absorb a particular ODI match.

WikramaSinghe [2] anticipated batsmen's overall success in test series by using a hierarchical linear model.

Barr and Kantor [3] have outlined a framework for determining and deciding on batsmen in limited cricket overs. They specified a new measure P(out) i.e. probability to have out and used a two-dimensional graphic illustration with strike rate on one axis and strike rate on another.

Iyer and Sharda [4] used neural networks to predict players' success by separating batsmen and bowlers into three groups separately – winner, slight and failure. They recommend that the player wants to be included in the crew to play World Cup 2007 based on the amount of times a player has earned different ratings.

By comparing the strengths of the two teams, Jhanwar and Paudi [5] predict the outcome of a cricket match. For this they assessed the success of each team's individual players. They developed algorithms to version batsmen and bowlers 'results, where they assess a player's ability by analyzing his overall career output after which his recent performances.

Pandey and Niravkumar [6] forecast the success of players in Cricket's Game Using Machine Learning by predicting the runs they will make.

Kalpdrum Passi and Niravkumar Pandey [7] used a regression model to predict the success of Players in One Day International Cricket Matches according to their role in play.

Stylianos Kampakis and William Thomas [8] used ML to estimate English County's outcome of twenty over Cricket matches using past records from players.

Nischal S [9] estimated the results of the Cricket Matches based on previous match winning probabilities.

Sethuraman K and Parameswaran Raman [10] concentrated on learning how to evaluate scoring patterns to develop an Algorithm to predict the outcome in Cricket's game.

III. PROPOSED SYSTEM

The proposed system unlike the currently available systems focuses on predicting the best playing 11 instead of only predicting the performances of a particular player. The system makes use of Random Forest Algorithm which has been found to be the best and most adaptable for this model. The performance of the player is not only to be predicted but depending on the previous records and other parameters a decision has to be passed whether the player is an ideal player to be included in the team or not. The model completely eradicates biased selections and gives the best decision. Once a player or a set of players are selected the system first analyses the available positions in the team i.e. if one batsman and one bowler in selected the system will first decide the number of available positions for batsmen, bowlers and all-rounders. Deriving these numbers, the procedure of decision making begins for all the players in the dataset. The decision trees are to be designed in such a way wherein the team would comprise of balanced number of batsmen, bowlers and all-rounders. Finer details such as type of bowler-seam/fast/spin, batsmen hand-right-handed/left-handed etc. are also to be looked after. The team should also have a wicket-keeper who has best stats as a wicket keeper as well has made remarkable

batting performances. The model could be trained to also decide the performance of the predicted team depending on the collective analysis of individual players' performances which could virtually posturize the situations o the field during the match. This would not only help the players but also the team and the coaches as they could know where and how their team has to focus to get better results for the match.

Several additional features could be added in the future such as the analysis of venues and the payer/team performances. A summarized study of player fitness cold also be introduced so as to know for how long and how many matches a particular player performs consistently without any injuries etc. The proposed system could be very beneficial for domestic tournaments where few player of the team are retained or be in a contract for a duration of certain years. In such cases it becomes difficult for analyst to manually study the statistics of each player and find out which players would form the best team along with the retained players.

IV. SYSTEM DESIGN AND ALGORITHM USED

The figure 1 below depicts the architectural block diagram of the proposed system. It basically depicts what features for selection of players will be given to the Django Server and how random forest classifier will be used for prediction of player.

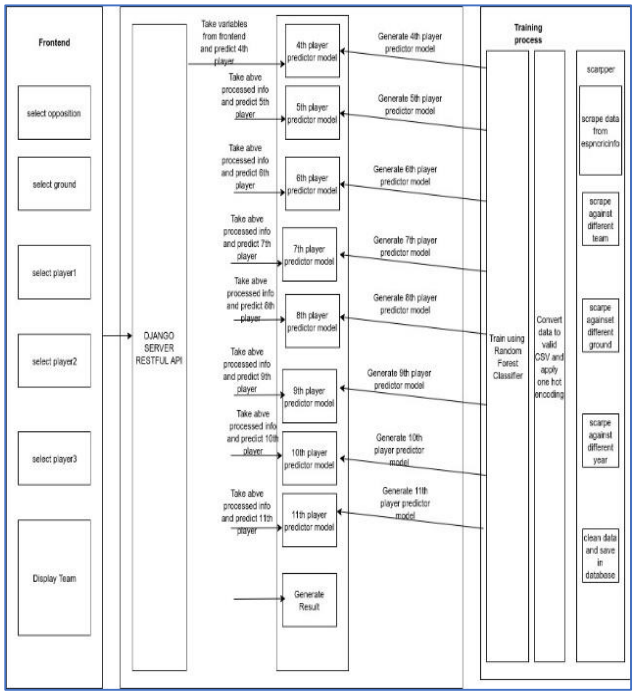

Figure 1 – Block Diagram

Figure 2 explains the Use Case Diagram for the proposed system. It shows the functionality performed by both the user and the server in our proposed system to recommend the player based on the feature vectors provided to the random forest classifier. Figure 3 gives the overview of the sequential flow of actions taken by the system to display the recommended output on the output screen after performing the training and testing phase of the classifier. All the feature sets are given to the Django server where entire pre-processing takes place.

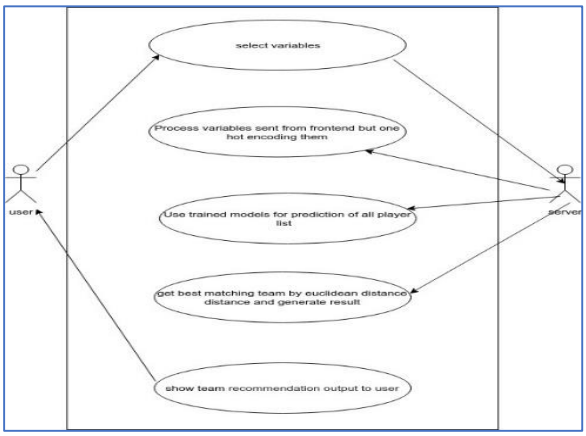

Figure 2 – Use Case Diagram

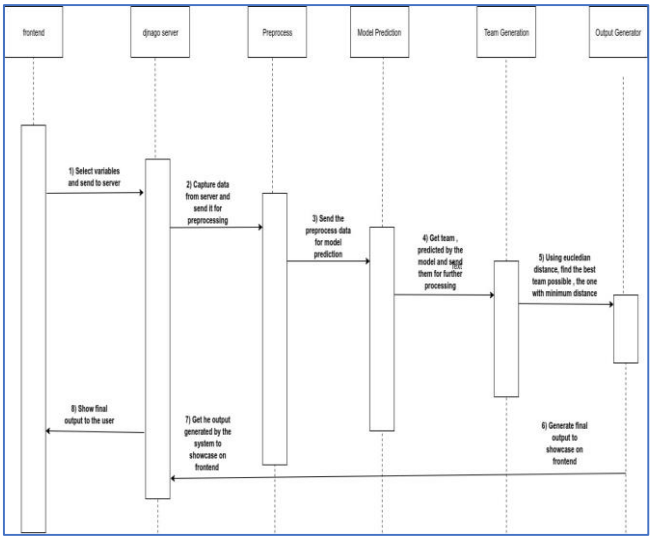

Figure 3 – Sequence Diagram

ALGORITHMS:

1. DATA SCRAPING (Web Scraping)

Web Scraping (Screen Scraping, Web Data Extraction, Web Filtering, etc.) is a method used to retrieve information from websites from which the data is processed to a local file in your laptop or database.

Many websites show a lot of useful data but only viewable. There is no provision for the saving for personal use of a copy of that data. The other alternative is to copy and paste the data manually, which is not technically feasible, however, because it can take a lot of time. Web scraping automates this operation, so that data can be easily retrieved without wasting time as shown in figure 4 instead of manually copying the data.

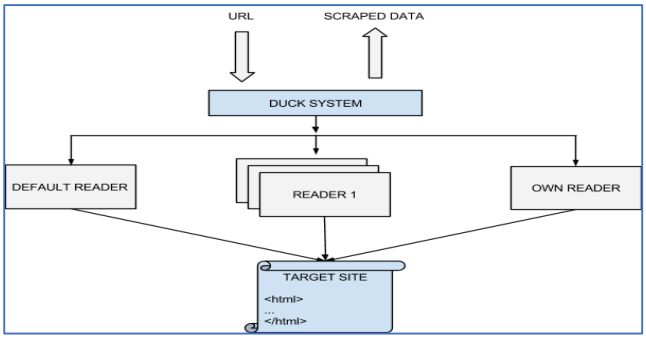

Figure 4 – Framework for Data Scraping

2. Random Forest Algorithm:

Random forests or random decision forests are a learning method for classification, regression and other procedures that operate by building a network of decision trees at the time of class (classification) or mean prediction (regression) training and performance mode of the individual trees. Random forests correct the habit of decision trees to overfit to their training range.

Preliminaries:

a. Decision Tree Learning

The use of Decision trees for various machine learning tasks is a common increase. Tree learning acts as a data mining method, since under scaling it is invariant and many other primary value transformations. This is an element of inspecting models with irrelevant functionality. They're correct though. It is precisely trees which grow very deep that must learn irregular patterns. Random forest summarizes more than one deep decision trees as shown on exceptional grounds in figure 5 with the objective of reducing the variance. However, this does cost a slight increase in bias and some lack of operability, but ultimately improves the output in the final model.

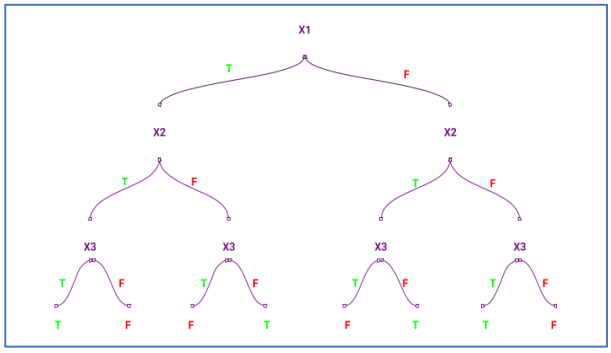

Figure 5 – Decision Tree Structure

b. Extra Trees

Another additional randomization stage generates trees that are highly randomised. Compared to ordinary random forests, there are differences: first, each tree is educated using the entire sample, and second, the top-down splitting is randomized within the tree learner. A random cut-point is chosen instead of determining the near-by cut-factor for every key under consideration. This value is extracted from an even distribution between the appropriate range (in the training set of the tree). Of all the splits generated randomly, the yield that has the highest score is chosen to divide the node. With this parameter, the default values are [root n] with classification, and n for regression, where n is the number of functions in model.

3. RANDOM FOREST CLASSIFIER

Random forest consists of a huge range of character decision trees that function as an ensemble as depicted in figure 6. Each individual tree within the random forest gives out a category prediction and the elegance with the majority becomes our model's prediction (see discern below).

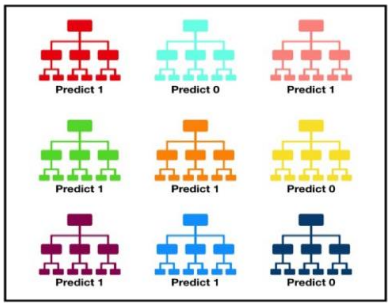

Figure 6 – Random Forest Structure

Working of Random Forest

The pseudocode for random forest algorithm is as follows:

Creation Pseudocode:

- Pick "k" from the total features of "m" by chance, where k<=m.
- Choose the best split point and measure the "d" node among the functions "k."
- Perform node splitting into daughter nodes again using the strongest break.
- Repeat the three steps above before hitting "1" number of nodes.
- Repeat above four steps to develop forest to establish 'n' trees.

Figure 7 – How Random Forest Works

Prediction Pseudocode:

- Take the test features and use the rules of will decision tree generated at random to predict the final outcomes and store the predicted outcome (target).
- Calculate the votes for each of the planned goals.
- Consider the predicted target strongly voted as the very last prediction from the algorithm of random wooded forest.

V. RESULT ANALYSIS

Landing Page/ Welcome Screen

Figure 8 – Landing Page

The Landing page as in figure 8 consist of the user assistance named "CrickBot" and the "Get Started" button.

User Assistance (CrickBot)

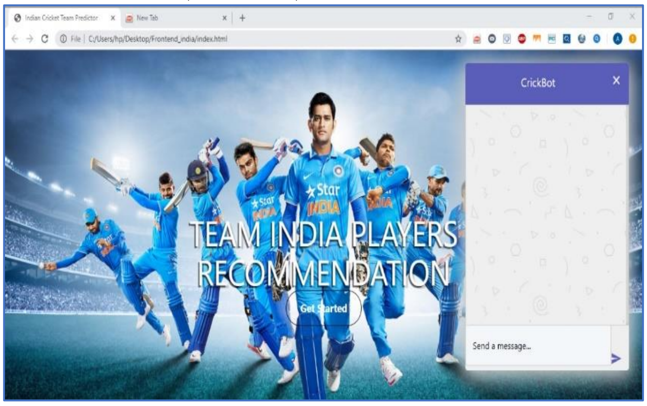

Figure 9 – Chatbot Screen

Figure 9 shows "Crickbot" that is a basic chatbot which provides the users assistance and help to use the system.

Running the Server

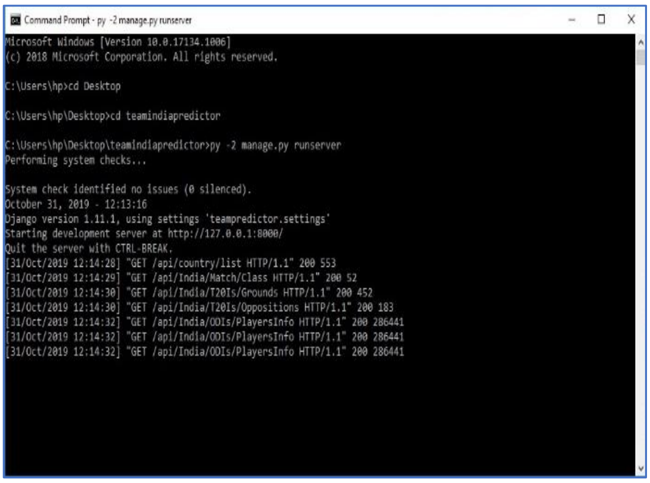

Figure 10 – Running the Server

The 'manage,py' file is run to initiate the server. All the URL mappings are checked and then the server connects the frontend and backend. It is by default run on port 127.0.0.1:8000

Selecting the Opposition team and the Match Venue

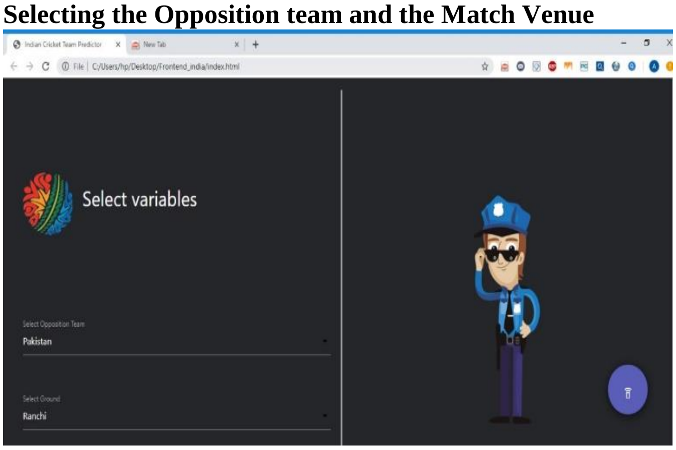

Fig. 5.4 – Selecting Opposition Team and Match Venue

The user has to select the Opposition Team from the list as it is an important parameter to analyze the records and predict the team.

The user has to also select the venue of the match as it classifies if the match is a home match or a away match for the team. It also considers the ground conditions based on the past performances.

Selecting the 3 players manually

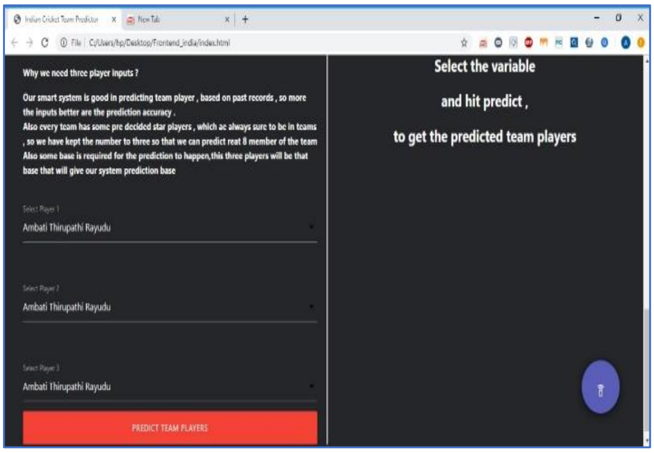

Figure 11 – Selecting 3 players manually

Every team has some pre-decided star players, which are always sure to be in team, so we have kept the number to three so that we can predict remaining 8 members of the team. Also some base is required for the prediction to happen, these three players will form that base which will give our system a prediction.

Our smart system is good in predicting team players, based on past records, so more the inputs better is the prediction accuracy.

Prediction Process

On hitting the 'PREDICT TEAM PLAYERS' button after the entering the above parameters, the system starts processing.

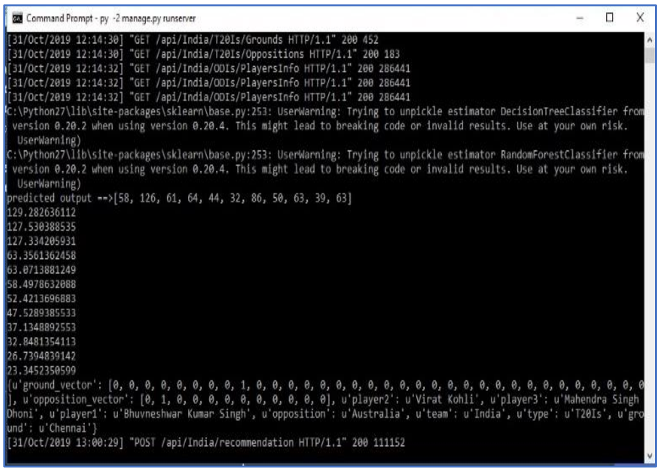

Figure 12 – Calculations of Euclidean Distances

Predicted Team

The system gives a team of 11 players based on their past records and considering the opposition team, venue etc.

Below are few examples of the format how the recommended players and their stats are displayed as shown in figure 13.

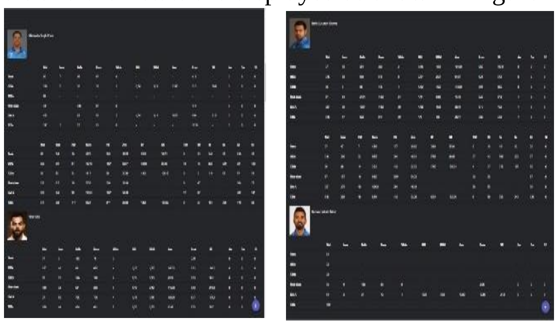

Figure 13 – Player Statistics

VI. CONCLUSION AND DISCUSSION

Conclusion

The selection of the right players for any match plays an significant role in the game outcome. A good estimate of the runs a batsman scores, and how many wickets a bowler is likely to take in a match, lets team management pick the best players for each match. Such findings, however, could only be extracted from the data that would come from various sources. We modelled the 11 members team in this project based on the stats and characteristics of the players. Different other factors that influence players' performance could be known as weather, or the design of the ground that could not be included in this study due to data unavailability. Random Forest proved to be the most accurate classifier with optimum precision for the datasets. Different tests can be performed for all game formats. The models for such formats can be designed to reverse the cricketers' vital characteristics; e.g. batsmen need stamina and the ability to play longer innings in test matches, while in T20 matches score more runs in fewer overs. Bowlers will also have better wicket-taking skills in test matches as well as a higher economy rate, which means they will concede less runs in T20 matches. In addition, attempts can be made to boost classifier accuracies and include other parameters such as location wise results, opponent team, ground conditions, player health, match fees etc.

Discussion

The models for these formats can be configured to demonstrate the players' main qualities required; for example, batsmen should have stamina and the ability to play longer innings. Similarly, bowlers should have greater potential to take wickets and a higher rate of performance, i.e. to concede less runs in T20 matches.

Many external factors impacting player performance such as climate or ground design could not be included in this framework due to data unavailability but could be researched, evaluated and introduced to enhance the framework.

To improve the dataset data could be taken from 2018 and 2019 into account by scraping them from the ESPN website and also possibly using the players' data to assess the quality of each recommended team player to summarize the overall performance of the team.

Example: Analysing the score for the innings or the number of wickets the bowlers could take in that match etc.

We could also try using more models together to gain improved accuracy and efficiency. Trying complex Machine Learning algorithms like Xgboost and fine-tuning the hyper parameters etc.

Moreover, attempts can be made towards the improvement of accuracies of the classifiers for various formats of matches.

REFERENCES

[1] S. Muthuswamy and S. S. Lam, "Bowler Performance Prediction for One-day International Cricket Using Neural Networks", Proceedings of IIE Annual Conference and Expo 2008, pp.1391-1395/

[2] I. P. Wickramasinghe, "Predicting the performance of batsmen in test cricket", Journal of Human Sport & Exercise, vol. 9, no. 4, pp. 744-751, May 2014.

[3] G. D. I. Barr and B. S. Kantor, "A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket", Operational Research Society, vol. 55, no. 12, pp. 1266-1274, December 2004.

[4] S. R. Iyer and R. Sharda, "Prediction of athlete's performance using neural networks: An application in cricket team selection", Expert Systems with Applications, vol. 36, pp. 5510-5522, April 2009.

[5] M. G. Jhanwar and V. Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach," in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016), 2016.

[6] Pandey, Niravkumar, "Predicting Players' Performance in the Game of Cricket Using Machine Learning" in Master of Science

[7] Kalpdrum Passi and Niravkumar Pandey, "Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning", Department of Mathematics and Computer Science, Laurentian University, Sudbury, Canada

[8] Stylianos Kampakis and William Thomas "Using ML to Predict the Outcome of English County twenty over Cricket Matches", University College London, 2015

[9] Nischal S "Predcting the outcome of Cricket Matches using AI", November 2017. https://medium.com/@xtaraim/predicting-the-outcome-of-cricket-matches-using-ai-14f61aed7b97

[10] Sethuraman K, Parameswaran Raman, "Learning Algorithm for Prediction in the game of Cricket", https://people.ucsc.edu/~praman1/static/pub/ML_Project_CS7641.pdf