7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

A Distinctive System for Cricket Predictions Using Machine Learning and Data Analysis

Atharva Karnik[1], Harshal Patil[2], Rushikesh Lokhande[3], Vinit Raut[4]
Department of Computer Engineering, Mumbai University, MUMBAI

Abstract— *This system proposes statistical modelling and machine learning approach to predict a team of XI players, individual performance of players and outcome of a cricket match. The system mainly focuses on two major methods i.e. Data Analysis and Predictions. For prediction, the proposed system will be using Three Algorithms as follows: Prediction of best possible team will be done by "K-means" clustering algorithm on the basis of overall statistics, location-wise statistics, opposition-wise statistics, year-wise statistics and the most important is recent performance with other features. This is because K-means is simple and effective, as the similar data will be stored in the group, clustering of data will become easier and effective. Prediction of individual performance of a player will be done by using "Random forest" algorithm. For predicting this, the system will be using data of predicted possible team. Prediction of winnability of a team will be done by using "Naïve Bayes" algorithm. For predicting this, the system will be using data of predicted best team and their predicted individual player's performance. Proposed system will update dataset after completion of each match, for giving better results. Proposed system will use dataset from www.cricsheet.org and www.espncricinfo.com*

Keywords: Machine Learning, Data Analysis, Clustering, Random Forest, Naïve Bayes.

I. INTRODUCTION

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed. Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics.

Selecting the best players for a particular match in any sport involves predicting the players' performance. Players' performance varies with the team they play against and the ground on which they play the match. Player selection is particularly more important in the game of cricket as the 11 players selected at the beginning of the match are fixed unless in case of injury. Moreover, the substituted players in such cases have limited privileges. Players' performance can be predicted by analyzing their past statistics and characteristics. Cricket players' abilities and performance can be measured in terms of different stats. Batsmen's statistics include batting average, batting strike rate, number of centuries etc. Whereas bowlers' statistics are measured by bowling average, bowling strike rate, economy rate etc. Other characteristics of batsmen include, batting hand of the batsman, the position at which the batsman bats etc. and those of bowlers include, the type of bowler, bowling hand of the bowler etc. Moreover, recent performances of the batsman/bowler, the performance of the batsman/bowler against a particular team and the performance of the batsman/bowler at a given venue are also taken into account for predicting his performance in the upcoming match.

II. LITERATURE REVIEW

An extensive online search produced very few articles related to the predictions in the game of cricket using machine learning and data analysis. The chapter contains literature survey of some papers which are related to the proposed system. Furthermore, the chapter contains an analysis table for all the papers discussed in literature survey.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Following are some papers related to the proposed system:

This paper defined the optimal set of various attributes that imposes high impact on outcome of cricket match. **[1]** In this paper data is collected from all past IPL matches. They have Extract and refine necessary data and considering Innings level structural data. Various relative indices are calculated using Extra-tree classifier. Higher the relative index implies that higher that attribute contributes towards the end results. For identifying different feature combinations, they have used Support Vector Machine algorithm (SVM) which gives better accuracies than naïve Bayes algorithm. SVM is a more advanced algorithm used in classification problems and has several parameters to pass.

This system applied AI Methodology for Automated Selection of Playing XI in IPL Cricket match using K mean clustering algorithm. **[2]** In this paper Elbow Method is used to determine the number of clusters K. using k mean clustering algorithm batsmen and bowlers are categorized into different clusters. 38 features are calculated for batsmen and 37 features are calculated for bowlers. Cluster Based Index (CBI) is developed and computed in order calculate the ranks of players within that cluster. ReliefF algorithm is used to determine the relative weightage of each feature within that cluster.

This paper defined the increased prediction accuracy for the individual performance of the players in the game of cricket using machine learning algorithms. **[3]** In this paper, system predicts how many runs a batsman will score and how many wickets a bowler will take in the upcoming match. They have experimented with four supervised machine learning algorithms and compared their performance. Random forest algorithm gives high accuracy than naïve Bayes, SVM, and decision trees. Random Forest builds the most accurate prediction models for both batting and bowling in all the cases. Also, the accuracy of the models increases as we increase the size of the training dataset for all algorithms. Classification and Regression Trees (CART) technique is used to grow the trees.

This paper defined the strategy for team selection in IPL 9 matches using random forest algorithm. **[4]** In this paper Team Selection is done using Heuristic function based on performance indices. The first step is to identify the factors and their weightages for creating a Performance Index for ranking the batting and bowling performances. MVPI (Most Valuable Player Index) of players are calculated. Recursive Feature elimination algorithm is used to get the important features. They have given different features and their weightages for different categories for the batsmen and bowlers and are according to the requirements of their respective roles.

This system predicts the outcome of a cricket match by comparing the strengths of the two teams. **[5]** For this, they measured the performances of individual players of each team. They developed algorithms to model the performances of batsmen and bowlers where they determine the potential of a player by examining his career performance and then his recent performances. KNN algorithms is used for predicting outcome of match. For predicting outcome of match, important three features are used as venue, toss and relative strength of teams. They haven't considered the recent performance of bowler in calculating his bowling score.

This system defines role of external factors on outcome of a One Day International Cricket (ODI) Match and also predicts the outcome of match using SVM and naïve Bayes algorithm. **[6]** In this paper SVM gives better accuracy than naïve Bayes algorithm. Logistic regression was used for identifying the significance of features and to determine the role played by individual as well as different combination of features. Using k-fold cross validation technique confusion matrix is generated which is used for predictive analysis.

This paper defines a new machine learning based Deep Performance Index for ranking IPL T20 cricketers. **[7]** In this paper deep performance index (DPI) is created to calculate performance of batsmen and bowlers in T20 cricket. The Recursive Feature elimination algorithm is used for extracting the important features and their relative importance towards designing the DPI. They have given many notations for describing metrics for performance evaluation of T20 players. 10-fold cross-estimation is performed and a ranking procedure is employed to complete the selection process.

This paper predicted the performance of batsmen in a test series using a hierarchical linear model. **[8]** This model is also used for data reduction. In this paper a three level HLM is used for the analysis of players by considering various aspects of players on

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

different levels such as some anthropometric characteristics such as the height and the handedness of the Player and variables like rank of team, location etc.

This system applied impact of power play overs on the outcome of Twenty20 cricket match. **[9]** This paper is important in developing the strategy of T20 cricket matches.so the team can arrange its batting order or shuffle their bowlers in such a way that they can attack their opponent in the power play overs. the team better in performance in both the skills during the power play is expected to win the match and vice-versa. This paper gives various methods and formulae to calculate the performance of both the teams during powerplay overs. To study the impact of power play on the outcome of Tewnty20 matches all the complete matches of four previous seasons of Indian Premier League (IPL) were considered. The details of the match information, score etc. necessary for computation is collected from tournament pages of the website www.espncricinfo.com for the respective seasons.

This system is analysing IPL match results using data mining algorithms. **[10]** In this paper two approaches are used for making balanced datasets such as oversampling and under sampling. As data mining algorithms on balanced datasets gives better accuracies. Dataset comprises of 12 attributes and 578 entities which are in Comma Separated Value (CSV) format. There are various methods used on data such as data cleaning, data transformation, and data pre-processing. The model which is used to analyze the results of matches was built successfully with accuracy rate of 97% for the balanced dataset using the classifiers i.e. after oversampling the imbalanced IPL dataset.

This paper defined cricket score prediction system using clustering algorithm. **[11]** This system is essential for making strategic decisions. It is a holistic approach as it takes in current input from user. database is updated after each prediction. The system works efficiently with a huge dataset of two thousand rows. This system focuses on only the performance of the player and is very quick at processing due to clustering of data. Currently the system processes and predicts data for IPL held on Indian cricket grounds.

This paper defined winning prediction analysis in One-Day-International (ODI) cricket using machine learning techniques. **[12]** In this paper two algorithms are used for the predictive analysis. k-mean clustering algorithm is used for selecting the playing 11 of both the teams. Logistic regression is used for winning prediction analysis. In order to check whether the aspects of the main prediction are right system take an example of historical match between two teams that is INDIA VS AUSTRALIA. Both the algorithms give better accuracies.

III. ANALYSIS

Table 1 gives the analysis of techniques and methods used in literature survey on a distinctive system for cricket predictions using machine learning and data analysis.

TABLE 1
ANALYSIS TABLE

Sr. No.	Title	Advantages	Open Challenges
1	Identifying the Optimal Set of Attributes that Impose High Impact on the End Results of a Cricket Match Using Machine Learning. [1]	This paper is helps to identify the optimal set of attributes by using extra tree classifier.	Using clustering algorithm, players could be divided into similar groups based on their performance which could be used to predicting playing 11 team. A mathematical relationship can be modelled between the winnability of a team and the players' performances

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

2	AI Methodology for Automated Selection of Playing XI in IPL Cricket. [2]	This paper uses k-mean clustering algorithm for classifying players into various clusters and a Cluster Based Index (CBI) is defined in order to calculate the ranks of players within that cluster. number of clusters is determined by Elbow method.	Use deep learning algorithms for team selection problem.
3	Increased prediction accuracy in the game of Cricket using machine learning. [3]	This paper is using random forest algorithm to predict runs and wickets. attributes are rated for better calculation process.	A model for test and T20 can be shaped.
4	Team Selection Strategy in IPL 9 using Random Forests Algorithm. [4]	This paper is using heuristic function for team selection along with random forest algorithm, so accuracy will be more. they are considering new kind of attributes for evolving cricket and also they obtained MVPI for each player.	Use Genetic algorithm.
5	Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. [5]	This paper is using 3 algorithms of modeling batsman, bowler and relative strength of two teams for analysing form of the player.	They haven't consider the recent performance of bowler in calculating his bowling score.
6	Role of External Factors on Outcome of a One Day International Cricket (ODI) Match and Predictive Analysis.[6]	This paper compares two algorithms on the proportion of accuracy. SVM has much greater accuracy than naïve Bayes.	Improve naïve Bayes algorithm for better accuracy results.
7	A new machine learning based D.P.I for ranking IPL T20 cricketers. [7]	Total MVPI will be sum of batting points and bowling points. Fielding points are not considered. In MVPI, Runs Scored by a batsman or wickets taken by a bowler dominates the value. There are some other factors that should be considered, otherwise only top order batsman will top the list.	Details can be enhanced with categorization of players according to their perceived role.
8	Predicting the performance of batsmen in test Cricket. [8]	This paper is using HLM(hierarchical linear model) for the purpose of data reduction and model the performance of batsman in test cricket	Consider additional attributes which could affect performance of batsman in test cricket as this paper is using attributes like height, handedness, rank.
9	Impact of Power Play Overs on the Outcome of Twenty20 Cricket Match. [9]	This paper is important in developing the strategy of T20 cricket matches. so the team can arrange its batting order or shuffle their bowlers in such a way that they can attack their opponent in the power play overs. the team better in performance in both the skills during the power play is expected to win the match and vice-versa.	This methodology can be applied in one day cricket.
10	Analyzing IPL match results using data mining algorithms [10]	This paper uses two approaches for making balanced datasets such as oversampling and undersampling. data mining algorithms on balanced datasets gives better accuracies.	Use other techniques for balancing datasets.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

11	Cricket score prediction system (csps) using Clustering algorithm. [11]	This paper is using holistic approach. database is updated after each prediction.	the system can be upgraded to encompass ODI and test match formats as well as on grounds around the world.
12	Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques. [12]	This paper is using logistic regression algorithm for predicting outcome of match.	Predictions can also be made when the match is abandoned due to rain, bad light, etc.

IV. PROPOSED SYSTEM

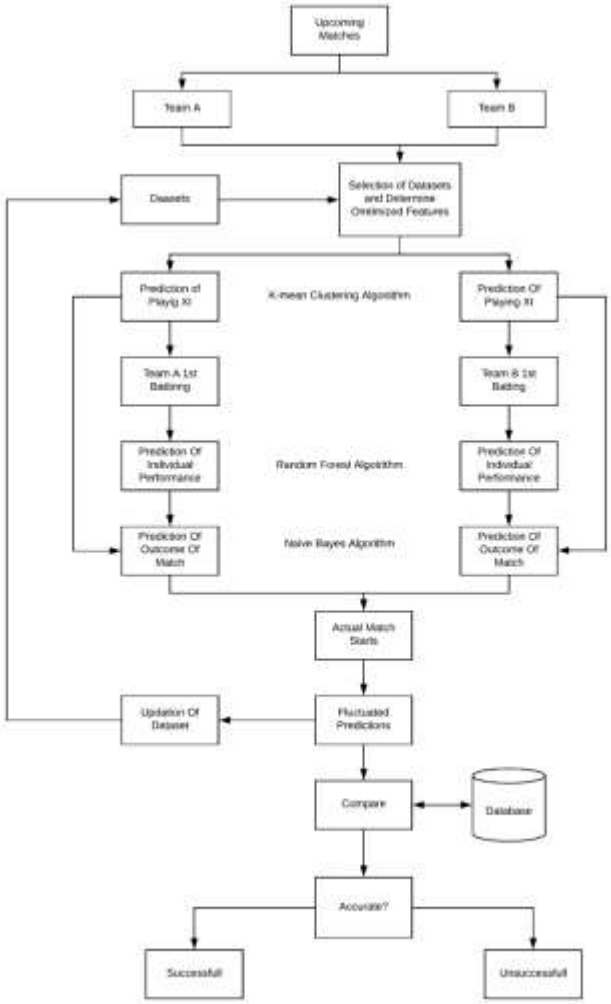

Fig 1: System Flow Diagram

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

The proposed system filters data from dataset and then determines optimized features. After this, proposed system aims to predict Best Possible Playing XI for the participating teams using K-Means Clustering Algorithm. This system will predict performance of individual player by using Random-Forest algorithm. Proposed system then predicts Outcome of the upcoming match.

V. CONCLUSION

There are many systems that perform predictions of upcoming cricket match. Existing system has predicted possible playing XI team by using clustering algorithm, but it has not used internal factors to a great extent and also not used external factors. Due to which the user may not get a proper prediction accuracy. Existing system have considered indices like MVPI and DPI to assess the player and also it keeps using the same dataset without being updated. Hence, the system will not get proper updated data. Proposed system introduces a model that can quantify the performance of batsmen and bowler using their past statistics. Proposed system introduces new measures based on raw attributes, that represent different aspects of both batsmen and bowlers' performance. Proposed system will update dataset for better results. As any system is not able to predict the toss, the proposed system gives two possibilities that one of the either teams bat first and according these possibilities, further predictions will take place. Filtering of datasets is done.

ACKNOWLEDGEMENTS

We would like to express a deep sense of gratitude towards our guide Prof. Janhavi Sangoi and Co-Guide Prof. Vinit Raut, Computer Engineering Department for their constant encouragement and valuable suggestions. The work that we are able to present is possible because of their timely guidance and support.

REFERENCES

[1] Pranavan Somaskandhan and Gihan Wijesinghe, "Identifying the Optimal Set of Attributes that Impose High Impact on the End Results of a Cricket Match Using Machine Learning" Department of Computer Engineering, University of Peradeniya, ICIIS'2017 1570379780.

[2] C.Deep Prakash, C. Patvardhan, C. Vasantha Lakshmi, "AI Methodology for Automated Selection of Playing XI in IPL Cricket" IJETSR, Volume 4, Issue 6 June 2017, ISSN 2394 – 3386.

[3] Kalpdrum Passi and Niravkumar Pandey, "INCREASED PREDICTION ACCURACY IN THE GAME OF CRICKET USING MACHINE LEARNING" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018.

[4] C. Deep Prakash, C. Patvardhan, C. Vasantha Lakshmi, "Team Selection Strategy in IPL 9 using Random Forest Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 139 – No.12, April 2016.

[5] Mehvish Khan, Riddhi Shah, "Role of External Factors on Outcome of a One Day International Cricket (ODI) Match and Predictive Analysis", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015.

[6] Madan Gopal Jhawar, Vikram Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2016) Conference Center, Riva del Garda, Report No: IIIT/TR/2016/-1.

[7] C. Deep Prakash, C. Patvardhan, Sushobhit Singh, "A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers", International Journal of Computer Applications (0975 – 8887) Volume 137 – No.10, March 2016.

[8] INDIKA PRADEEP WICKRAMASINGHE, "Predicting the performance of batsmen in test Cricket", Eastern New Mexico University, United States.

[9] Dibyojyoti Bhattacharjee, Manish Pandey, Hemanta Saikia, Unni Krishnan Radhakrishnan, "Impact of Power Play Overs on the Outcome of Twenty20 Cricket Match", Bhattacharjee, D., et al. (2016). Ann Appl Sport Sci, 4(1) : 39-47.

[10] Shimona.S, Nivetha.S, Yuvarani.P "ANALYZING IPL MATCH RESULTS USING DATA MINING ALGORITHMS", International Journal of Scientific & Engineering Research Volume 9, Issue 3, March-2018 ISSN 2229-5518.

[11] Prof. Preeti Satao, Ashutosh Tripathi, Jayesh Vankar, Bhavesh Vaje, Vinay Varekar, "CRICKET SCORE PREDICTION SYSTEM (CSPS) USING CLUSTERING ALGORITHM", ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME- 3, ISSUE-4, 2016

[12] Abhishek Naik, Shivanee Pawar, Minakshee Naik, Sahil Mulani "Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques", Volume 3 Issue 2 April – 2018.

[13] http://www.espncricinfo.com, last accessed on 20th Feb 2019.