Forecasting the Outcome of the Next ODI Cricket Matches to Be Played

Md. Minhazul Abedin, Silvy Rahman Urmi, Md. Towfiqul Islam Mozumder, Md. Samiur Rahman, Adnan Firoze

Abstract: ***Cricket is one of the most popular sports in the contemporary world. The ebullience of securing victory in a particular match has motivated the rudimentary part of this research aspect. Winning in Cricket depends on various aspects like weather, track records of the performances of players, performances at a specific venue, match experiences, performance against a specific team and the current form of the team and the player. In this paper the main goal is to estimate a win prediction for ODI (One Day International) cricket match. For purposes of model building, various method has been applied retrospectively to the data that are already obtained from previously played matches. From them, Random Forest has given the best outcome of 92.6%.***

Keywords: Cricket, Data Mining, Random Forest, K-Nearest Neighbor, SVM, Decision Tree.

I. INTRODUCTION

This research focuses on forecasting the outcome of forthcoming ODI cricket matches to be played .ODI cricket matches are considered to be one of the most popular sports in the ongoing era of sports [1].This is a game of in total 100 overs, with each team having 50 overs consisting of both batting and bowling. Let's dive deep into the research aspect, the data set involves the overall records of more than 3000 ODI cricket matches over the previous 10 years. Two approaches have been followed to test and train the data set which are classified as the classifier approach of prediction algorithm and the regression approach of the prediction algorithm. Different prediction algorithms produced varied results in terms of accuracy and acceptance. The dependency of forecast is applicable on real-time and not invariant which will depend on the contemporaneous condition of the matches [2]. We've used several machine learning algorithms to develop the prediction methodology for our work. We've come up with four respective algorithms namely Random Forest (RF), Support Vector Machine (SVM), Decision tree and KNN (K-Nearest Neighbor). Out of the four distinguished machine learning algorithms we have observed that Random Forest has yielded the maximum accuracy. In the subsequent sections of this paper, there will be a detailed discussion about the aforementioned aspect.

Manuscript published on November 30, 2019.
* Correspondence Author

Minhazul Abedin*, Electrical & Computer Engineering, North South University, Dhaka, Bangladesh. Email: abedin.minhazul@northsouth.edu
Silvy Rahman Urmi, Electrical & Computer Engineering, North South University, Dhaka, Bangladesh. Email: silvyrahman.urmi@northsouth.edu
Md. Towfiqul Islam Mozumder, Electrical & Computer Engineering, North South University, Dhaka, Bangladesh. Email: towfiqul.mozumder@northsouth.edu
Md. Samiur Rahman, Electrical & Computer Engineering, North South University, Dhaka, Bangladesh. Email: samiur.rahman09@northsouth.edu
Adnan Firoze, Electrical & Computer Engineering, North South University, Dhaka, Bangladesh. Email: adnan.firoze@northsouth.edu

II. METHODOLOGY

We have developed a solution where we have predicted the outcome of the next ODI matches based on features such as precisely Match Consequence (Win/Loss), Score of Individual Team (SIT), Loss of Total Wickets (LTW), score above 300 or not, were carefully chosen as the predictor variables [4]. We have used supervise machine learning algorithms to build our model and predict the outcome. The data set we used contains vast records having distinguished attributes specifically the host team's name, the invited team's name, match outcome, consequence of the toss, difference by which the team has won, venue, match schedule. The data we're dealing with to estimate the outcome have been collected from the ESPN and CricInfo website.

Machine learning approach is being applied intuitively to help construct the model which yield to foresee the outcome. Distinguished mechanisms have been implemented due to the statistical perfection outlined in schemes, assumptions, and implementation details. It is a fact that one cannot deny of is that machine learning is further distributed into supervised learning. In this paper we have used supervised learning algorithms such as Random Forest, K-Nearest Neighbor, Support Vector Machine, and Decision Tree. Now a days Scikit learns library [3] frequently be used and is helpful to tweak different aspects of an algorithm. As a result, the Scikit learn library is used to call the classification methods.

The approach at a glance to estimate the result would be comprised of the succeeding steps. Team features, batsmen features, and bowling features has been drawn from the collected data set. Then the analysis has been taken place on team features, batsmen features as well as bowler features with the help of Multiple Potential Features Selection Method.

After that we've compared the outcome to determine the best feature selection methods between team features, batsmen features, and bowling features with a variety of different predictive algorithms, parameters, and feature sets to be analyzed, and optimized. The scope of selecting the highest scoring features based on the analysis has been accomplished. Then the most promising algorithm has been run on the newly obtained feature set. The consequences of the analysis stated above would yield optimized model accuracies.

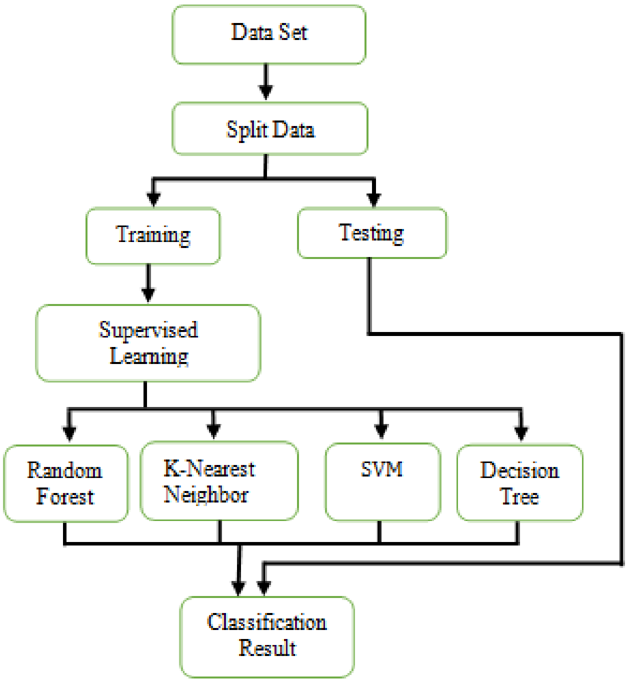

Fig. 1. Architecture of the Cricket Prediction Model using Machine Learning Algorithm

III. RELATER WORK

Two methods have been proposed for the prediction of the outcome of a particular match in this paper. First method which predicts the score of first innings on the basis of the current run rate and number of wickets fallen. The second method predicts the score of the match in the second innings evaluating the same aspects as of the previous method along with the aim that is given to each batsman of the batting team. To implement this two methods Linear Regression Classifier and Naïve Bayes Classifier have been used for the first and second innings [5].

Predicting the performance of each player is the main target of this paper e.g. runs that will be scored by each batsman and wickets that will be taken by each bowler. Naive Bayes, Random Forest Classifier, Multi-class SVM and Decision Tree Classifiers have been used to create the prediction model for number of runs and number of wickets. Among these three classifiers, Random Forest Classifier have been produced the most accurate and precise result [6].

The main purpose of this paper is to predict the outcome of ODI cricket matches. Career statistics and team performance have been utilized in order to train the models. Four types of machine learning algorithms (Decision tree, Support Vector Machine, Logistic Regression, Bayes Point Machine Binary Classification Model) have been used and even get compared to know the algorithm that even can give the most efficient result among that four. And according to this paper, Bayes Point Machine Binary Classification Model give the most accurate result [7].

Forecasting the outcome of the ODI cricket matches and the performance of individual players is the main purpose of this paper. Characteristics of individual players have also been evaluated in order to predict the outcome of a match. SVM, Random Forest, Logistic Regression, Decision Trees, K-Nearest Neighbor (KNN) classifier has been used to get the result. Among them K-Nearest Neighbor (KNN) gives the best result in comparison to other classifiers [8].

In this paper, they have discussed how to predict the outcome of a T20 match. Statistics which contain the information about player's performance and team have been analyzed initially to formulate the prediction model which focused on the fact of win or lose during the toss, player's rating, weather overcast, team's ranking, remaining wickets of a team, required run rates per over. This prediction model is based on multi-layer perception with adjustable factor weightage and this method have been assessed on historical features of ball by ball match. This new method produced 85% accurate prediction before match and 89% accurate prediction during match [9].

A tool was developed based on classification to predict the outcome in ODI Cricket by A. Kaluarachchi, and Aparna S. Varde. Different classification techniques (Naïve Bayes, Decision Trees, Bagging and Boosting) were used for comparison & it was performed using Receiver Operating Characteristics Curve (ROC) and Root Mean Squared Error (RMSE) among these classification techniques, Naïve Bayes came up with the highest ROC and lo west RMSE for better learning [10].

A method has been defined to enhance the accuracy of prediction of a cricket match by Kalpdrum Passi and Niravkumar Pandey through comparing four multi-cast classification algorithms including Naïve Bayes, Decision Trees, Random Forest and SVM. Weka and Dataiku tools were used for predictive analytics. Random Forest gave 90.74% accuracy for predicting runs and 92.25% accuracy for predicting wickets taken by a bowler for the given data sets [11].

In this paper, they explain how to identify and quantify the general measure of the significance of a match or tie in a tournament or competition. Their estimation is based on the results of all other games, some played, some predicting, and takes into account the effect of a single game at the end. They use the logistic regression in order to predict the matches and the Monte Carlo simulation to quantify and apply this experiment in soccer games. A model class and a number of possible predictor variables are to be defined. They assume that the outputs of matches are independently given the predictor values to simplify this prediction problem. Then a single generalized linear model can be used, and as the result can be clearly interpreted as winning, drawing or losing it is normal to concentrate on models that are appropriate for a range of responses [12].

In this paper a model with two methods is proposed, predicting first entry score not only based on the current ratio, but also the number of wickets dropped and the position of the match as well as batting equipment.

The second method forecasts the match results in the second round taking into account the same characteristics as the previous method and the target set for the batting team. The two techniques are being applied for the first intake and second intake using the Linear Regression Classifier and Naive Bayes Classifier. The error of the linear regression classification prediction was compared to the current score projection method by comparison of the actual ODI cricket results. The error in the Linear Regression Classification has been found to be less than the Current Run Rate Method to predict the final score in any match situation. The exactness of the Naive Bayes to forecast the match results also varies from 70% to 91% when the match progresses [13].

In this article, newly implemented encoding method for the convolutional neural network (CNN) is able to increase the accuracy of 80%. The encoding approach that this study uses is based on the player's power level. After encoding of the NBA event data, the deep learning model was employed and the desired results were obtained. Although the results of the experiments are less than those of a neural network, the experiment results are closer to realities by coding according to the capacities level of each participant [14].

IV. ALGORITHMS

A. Random Forest

Random Forest is one of the machine learning algorithms that is flexible and yields a better result most of the time than most other algorithms. It inherently constructs multiple Decision Trees and merge them together to get a more accurate, and precise prediction. Nowadays it is considered to be one of the eminent and most used algorithms. The main fact behind this consideration is that it can serve for both the purpose of classification and regression. Random Forest is a supervised learning algorithm. From the respective name, we get a flavor of how the philosophy works for Random Forest. It creates an imaginary forest by an ensemble of Decision Trees and most of the time trained with the "bagging" method. This method is a combination of learning models which enhances the overall result. Random Forest is considered to be a Meta estimator that suits a variety of Decision Tree Classifier on several sub-sample of data set. To increase predictive accuracy, the specified averaging technique has been used. Sub-sample would not differentiate from the original sample copy. Generating Decision trees using Random Forest involve a data set P of p tuples. Likewise, a set of Decision Trees can be generated from the data set. Traversing the k iteration, a training data set would need to be sampled with the data set generated earlier. To split at the node, a trivial amount of properties from the available properties are selected arbitrarily as feasible candidates [3]. Then CART (Classification and Regression Trees) is applied to enlarge the trees. The trees are then letting to be emphasized maximally and are not allowed to be trimmed. CART is a non-parametric Decision Tree induction technique that may create classification and regression trees [15].

B. K-Nearest Neighbors

K-Nearest Neighbor is one of the most straightforward machine earning algorithm which works on least distance for query indemnification to the training samples to regulate K-Nearest Neighbor [16]. K-Nearest Neighbor classifier classify with two stages first is the fixation of the nearest neighbor and the second is ascertainment the class using this neighbor. Let's say we have a training data set M made up of some training example. Each of the training example is leveled with a class level. The classifier identifies the K-Nearest Neighbor with distance matrices. There are lots of way to determine the classes of unknown example

Q. the most useful technique is to define the common class among the nearest class of quarry. Some noise reduction technique in K-Nearest Neighbor that improves the accuracy of the model.

C. Support Vector Machine

SVM (Support Vector Machine) is one of the most frequently used machine learning techniques [17]. It is an elegant, powerful and highly accurate intelligent classification algorithm. SVM is highly preferred algorithm as it produces significant accuracy with less computation power. SVM can be used for classification and regression test. But it is widely used in classification objectives. The objective of support vector machine (SVM) algorithm is to find a hyper plane in an N-dimensional space (N-number of features) that distinctly classifies the data points. It is an automated and efficient deterministic learning algorithm which provides the benefit of logical and verifiable results. It has the advantage over the other techniques of converging to the global optimum, and not to a local optimum that depends on the initialization and parameters affecting rate of convergence.

D. Decision Tree

The Decision Tree algorithm unravel problems by using tree illustrations. Each internal node and leaf node correspond to an attribute and to a class. There are several steps involved for this algorithm, splitting, pruning, tree selection. Splitting is the partitioning of the data sets to subsets. Pruning is the process of making some branch nodes to the leaf nodes and removing those from the main branch. And the tree selection is a process of finding the smallest or shortest tree that fits the data [18].

V. RESULT AND ANALYSIS

Using numerous algorithms and numeric options and therefore the results of the matches we tend to assessed an intensive number of binary classifiers utilizing their sickie-learn executions to make supervised classification models as Random Forest, K-Nearest Neighbor, Support Vector Machine (SVM) and Decision Tree Classifier [3].We split our data sets into training and testing data sets to find the combination that produce the best accuracy score. We have used various supervised machine learning algorithms such as Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree classifier to classify our model.

The table given below shows the accuracy rate of the different supervised machine learning algorithm depending on with the data split percentage.

Table- I: Accuracies for Different Algorithms with Data Splits Percentages.

Classifier	Accuracy (%)					
	60% train 40% test	65% train 35% test	70% train 30% test	75% train 25% test	80% train 20% test	90% train 10% test
Random Forest	86.42	91.26	90.93	91.83	90.67	92.61
K - Nearest Neighbor	72.4	72.09	72.8	71.43	72.02	73.73
SVM	83.01	84.46	87.54	88.44	89.83	91.96
Decision Tree	82.59	88.34	86.4	88.44	90.68	92.31

As we can see, Random Forest classifier builds the most accurate prediction model with 92.61% of maximum accuracy [15]. The accuracy of the model increases as we increase the size of the training data set. Random Forest predicts the winning parentage of a team with height accuracy of 92.61% when we use 90% of data to train the model and lowest accuracy of 86.42% when we used 60% of data to training. Decision tree preforms rationally well with maximum accuracy 92.31 and minimum accuracy 82.58 depending on with the data set split percentage. Support vector machine (SVM) Prediction model predict with maximum accuracy of 91.96% when the data split percentage is 90% and minimum accuracy of 83.01% when the data split percentage is 60% [18]. The forth prediction model k-nearest neighbor (KNN) predicts with maximum accuracy of 73.73% and minimum accuracy of 72.40% along with the data split percentage [16].

The performance comparison of the different classification model was shown using the Receiver Operator Characteristics Curve (ROC). The area under the ROC curve is the measurement different prediction models accuracy. The Y-axis of the curves show true positives and X axis shows false positive [19].

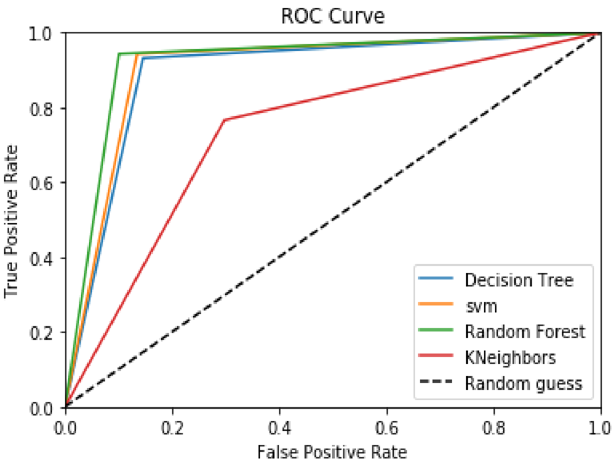

Fig. 2. ROC Curve for Different Algorithms

VI. CONCLUSION

The main purpose of this paper is to develop a model to predict the outcome of the upcoming ODI cricket match [20]. We have used the data of previous matches played between the oppositions in order to design our model. We have used different supervised algorithms such as Random Forest, k-Nearest Neighbor, SVM, and Decision Tree to design this model. By changing data set split percentage we have received maximum 92.61% prediction accuracy from our designed model with Random Forest classifier [11]. Random Forest provides a useful way to assign the winning probabilities for the competing teams in the ODI matches. The best part of this paper is that we can predict the cricket matches winning team just by using the existing technology and the computers.

REFERENCES

1. K. Desai and S. Doshi, “Predicting Outcome of ODI Cricket Games,” *International Journal of Engineering Research & Technology*, 2015.
2. S. Kumar and S. Roy, “Score Prediction and Player Classification Model in the Game of Cricket Using Machine Learning,” *INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH*, 19-Sep-2017.
3. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, pp. 93–117, Oct. 2011.
4. A. Kaluarachchi and A. S. Varde, “CricAI: A classification based tool to predict the outcome in ODI cricket,” *2010 Fifth International Conference on Information and Automation for Sustainability*, 2010.
5. R. A.Lokhande and P. M. Chawan, “Prediction of Live Cricket Score and Winning,” *International Journal of Trend in Research and Development*, vol. Volume 5(4), Jul. 2018.
6. F.Monir, M.K.Hasan, S.Ahmed and S. Md quraish, “predicting a T20 cricket result while the match is in progress” *Doctoral dissertation, BRAC University*, 2015.
7. G. J. Harshit, “A Review Paper on Cricket Predictions Using Various Machine Learning Algorithms and Comparisons Among Them,” *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*.
8. M. G. J. Jhawar and V. 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*, 2016.
9. M. Yasir and L. I. CHEN, “Ongoing Match Prediction in T20 International,” *IJCSNS International Journal of Computer Science and Network Security*, 2017.
10. Kaluarachchi and A. S. Varde, “CricAI: A classification based tool to predict the outcome in ODI cricket,” *2010 Fifth International Conference on Information and Automation for Sustainability*, 2010.
11. K. Passi and N. Pandey, “Predicting Players Performance in One Day International Cricket Matches Using Machine Learning,” *Computer Science & Information Technology*, 2018.
12. Scarf, Phil & Shi, Xin. (2008). The importance of a match in a tournament. Computers & Operations Research. 35. 2406-2418. 10.1016/j.cor.2006.11.005.
13. T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, Faridabad, 2015, pp. 60-66. doi: 10.1109/ICSCTI.2015.7489605
14. S. LIN, M. CHEN and H. CHIANG, "Forecasting Results of Sport Events Through Deep Learning," *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, Chengdu, 2018, pp. 501-506.
doi: 10.1109/ICMLC.2018.8526954

15. A. Liaw and M. Wiener, “Classification and Regression by Random Forest,” vol. Vol. 2/3, Dec. 2002.
16. J. A. M. E. S. M. KELLER and M. I. C. H. A. E. L. R. GRAY, “A Fuzzy ΛΓ-Nearest Neighbor Algorithm,” *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*.
17. R. Burbidge, M. Trotter, B. Buxton, and S. Holden, “Drug design by machine learning: support vector machines for pharmaceutical data analysis,” *Computers & Chemistry*, vol. 26, no. 1, pp. 5–14, 2001.
18. S. R. S. Safavian and D. Landgrebe , “A Survey of Decision Wee Classifier Methodology,” *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, vol. VOL. 21, 1991
19. J. A. Hanley and B. J. Mcneil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
20. A. B. Bandulasiri, “PPredicting the Winner in One Day International Cricket,” *Journal of Mathematical Sciences & Mathematics Education*, vol. Vol. 3, No. 1

AUTHORS PROFILE

Md. Minhazul Abedin is currently a student of North South University studying Computer Science Engineering. Where he is going to complete his under graduation by 2020. He has expert knowledge on C, C++, Java, shell script, assembly language and Python. He has a strong interest in the field of Machine Learning, Artificial Intelligence, Data Science, Computer security and information assurance. He has already worked on a paper on Bird Species Classification from an Image Using VGG-16 Network, in proceeding of the 2019 7th International Conference on Computer and Communications Management (pp. 38-42). ACM.

Silvy Rahman Urmi is currently a final year student of North South University doing BS in Computer Science Engineering and will be graduating in 2021. She is an ardent programmer in software engineering and programming languages such as C, C++, Java and Python. She has a strong interest in the field of Machine Learning, computer security and information assurance. Her research interest includes areas of Data Science, Machine Learning, computer security, operating systems and mobile security. She has already worked on a paper on distributed system of hash matching, has been published in International Research Journal of Engineering and Technology.

Md. Towfiqul Islam Mozumder is currently perusing my Undergraduate in Computer Science & Engineering .From 1st July 2019 to 30th September 2019, he worked for completing his Internship at "Transcom Limited" as a Web Developer. He has earned the prestigious Google Web Academy Certificate in January 2017. He has completed The Hour of Code from NSU ACM Student Chapter. He also recognized from Bangladesh Summit Featuring Google for Education. Recently he is working on a project named “Identification of Violence in Real-time, and Transmission of Detected Violent Exposure to the Nearest Law Enforcement Agency in Real-time.

B Md. Samiur Rahman is currently pursuing my Undergraduate in Computer Science & Engineering in North South University. He has worked as a Research Assistant under the Supervision of Professor Dr. Tim CW. Chen. He is looking forward to publishing this paper in this renowned conference. He has worked as a Software Development Engineer for completing my Internship at MySoft Limited. He has enough skill and experience on analog circuit design, Java, C, C++, Php, assembly language and artificial intelligence also. He is looking forward to work on IoT devices and social projects to help the society.

Mr. Adnan Firoze was a Faculty Member at North South University, Bangladesh and formerly a Teaching Fellow at the Computer Science Department in Columbia. He has completed his Dual M.S. in computer science and journalism. He has worked at Computer Vision and Cybernetics Group. His interdisciplinary research works are based on digital image processing, machine learning, fuzzy logic, neural networks and data mining.His previous research works have appeared in numerous prestigious conferences and journals. His present research is based on real time triangulation of mass calamities using NASA’s satellite imagery and also perception of visual data by artificial intelligence. A more ambitious research he has undertaken is detecting and classifying violent action in surveillance and cell phone videos.

He has research papers published in International Journal and Prestigious conferences.