Prediction of Outcome of ODI Cricket Matches Using Machine Learning

Prof. Pallavi[1], Kiran V Kulkarni[2]
[1, 2] Dept of Computer Science & Engineering
[1, 2] Atria Institute of Technology Bangalore, India

Abstract- *This paper presents a method that is aimed towards predicting the outcome of ODI cricket matches by implementing machine learning algorithms. Cricket is one of the most popular sports in Australia, Caribbean, UK and South Asian nations with a net fan base of around 3 billion. This form of sports is widely recognized by the International Cricket Council and is played by over 125 countries. There are lots of pre-game and in-game attributes which decides the outcome of a cricket match. Pre-game attributes like the venue, past track-records, innings(first/second), team strength etc. and the various in-game attributes like toss, run rate, wickets remaining, strike rate etc. influence the result of a match in a dominant manner. Based on these results CricAI: Cricket Match Outcome Prediction System has been developed. The designed tool takes into consideration the pre-game attributes like the ground, venue (home, away, neutral) and innings (first/second) for predicting the final result of given match.*

Keywords- Decision Tree Classifier, Machine learning, Multilayer Perceptron Classifier, Features

I. INTRODUCTION

Cricket is a bat and ball game which is played between two teams having 11 players each. Each team comes to bat and has a single inning in which it seeks to score as many runs as possible, while the other team fields. At the end, the team who scores the most runs and takes most wickets will become victorious. It is hard to predict the game of Cricket until the last ball is bowled because various natural factors affect the outcome of the game. The widely played formats in Cricket are T20, One-day International, Test Match.

The study is focused on one of the most popular formats which is the One-day International. The scheduled duration of the game is the prime difference between these three formats, which directly modifies the number of deliveries each team get to play in their respective innings.

Test cricket format is the longest and is considered as the primary format of the game. Match duration goes on for five days in which each team get to play two innings each. T20 is the shortest internationally recognized format of this game, where each team innings consists of 20 overs. This is more of an" explosive" compared to the other two formats.

The study is focused on the most popular format of Cricket, One Day Internationals or the ODIs. The outcome of One Day Internationals is influenced by a varied number of features and can be predicted like all other games. The best attributes or factors that influence the match outcome need to be found. The factors considered for analysis include:

- Teams Past Performance: This factor captures the historic outcomes of all the matches played between the teams.
- Ground: This plays a vital role as teams have great track records on particular grounds and carry psychological superiority over the other.
- Innings: This factor determines which team batted first & which batted second.
- Home Game Advantage: This is achieved by using venue feature, which determines whether a particular ground is home/away/neutral for each of the playing teams.

Both of the classifiers are trained on the basis of these factors. For predicting the final result of cricket matches two supervised classification techniques - Decision Trees and Multilayer Perceptron Networks have been used. Comparative study is done between both the classifiers and final results are summarized in this paper.

A desktop app cricAI was developed which could predict the outcome of any cricket match when given certain factors as input. This could help the support staff, governing bodies and cricketers to plan in advance which could have a massive impact on the result of the game.

The rest of this paper is organized as follows. Section 2 explains the approach which has been taken into account for the proposed analysis. Section 3 deals with the comparative analysis of both the classifiers used. Section 4 presents the other related works in this domain. Section 5 gives the conclusions and the future scope associated with this approach.

Since, multiple independent attributes need to be dealt with, therefore clustering them after finding similarity patterns doesn't seem feasible, due to which clustering doesn't make any reasonable contribution to this research.

II. APPROACH FOR ANALYSIS

A. Data Collection

Data was extracted from [3] by running a scraping script in a justified manner, sending 1 request per second.

TABLE I: SCRAPPED DATASET FORMAT

MatchI	Team1	Team2	Winne	Margin	Ground
ODI#1	Austra	Englan	Austra	5wicke	Melbour
ODI#2	Englan	Austra	Englan	6wicke	Manches
ODI#3	Englan	Austra	Austra	5wicke	Lord's

Dataset comprises of all the ODI matches from Jan 5, 1971, to Oct 29, 2017. A total of 3933 ODI match results were scrapped. The collected dataset was subjected to cleaning process where some of the matches were deleted from the analysis. Since it's not possible to foresee the impact of nature on cricket, matches which either ended up in a tie/draw or interrupted by rain, were being removed from the dataset. Matches of special teams like World XI, Asia XI & Africa XI were also removed.

The dataset was also replicated two times by swapping the team positions i.e. a game between team 1: India and team 2: Sri Lanka was also replicated as team 1: Sri Lanka and team 2: India. For further making the dataset suitable for input to the various machine learning classifier models, the continuous dataset was converted into a categorical dataset, using dummy variables.

Innings feature was determined by first translating Column:

Margin into Column: Winner Innings using:

⋄ Win by Wickets =⇒ Winner Innings: 2

⋄ Win by Runs =⇒ Winner Innings: 1

Further, Using Column: Winner and the generated Column: Winner Innings, the innings of each team per match were acquired.

Venue feature was determined by using Column: Winner and scrapped data frame from [3] which provided the names of cricket grounds in all countries. Combining both of these, Column: Host Country was generated, which was used to get venue of a match with respect to both the teams.

The dataset was saved in comma separated format. A total of 7494 match records were used for the analytical study which was further divided into the testing and training data.

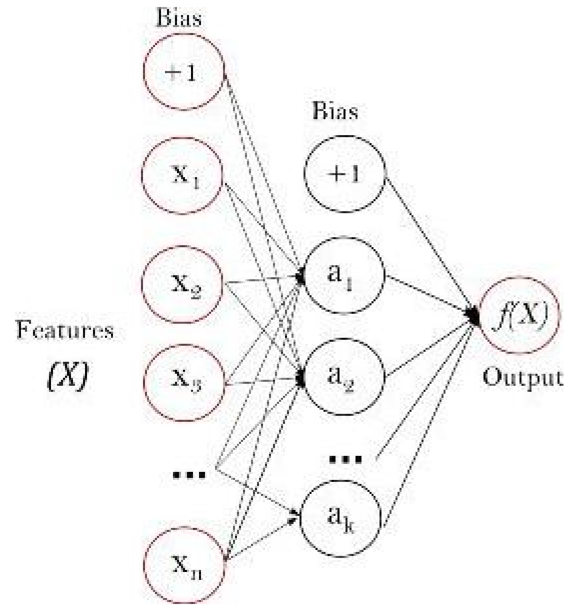


Fig.1: Multilayer Perceptron Network

- Training Dataset Size: 5620
- Testing Dataset Size: 1874

A. Multilayer Perceptron Networks

MLP Network is a type of supervised learning algorithm which learns a function

$$f(.) : R^n \rightarrow R^t \quad (1)$$

by using some training dataset, where *t* is the total number of output units and *n* is the total number of input units. Given features set $X = x_1, x_2, ..., x_m$ and a target *y*, MLP Network can be trained to be a non-linear function approximator for classification as well as regression. The core difference between MLP Networks and Logistic regression is that in the former one there can be hidden layers, which are actually one or more nonlinear layers. Fig 1. shows a Multilayer Perceptron Network with only 1 hidden layer.

Input layer is the leftmost layer representing the input features, consists of a set of neurons.

$$x_i | x_1, x_2, \ldots, x_m \quad (2)$$

Values from the previous layers are transformed using weighted linear summation by the neurons of the hidden layer,

$$w_1x_1 + w_2x_2 + . . . + w_mx_m \quad (3)$$

followed by a non-linear activation function acting on its output. The last hidden layer further transfers these values towards the output layer which transforms these intermediate values into the final output values.

MLP Classifier [4] is implemented using a multi-layer perceptron (MLP) algorithm in which back propagation is used for training. More precisely, some form of gradient descent is actually used to train the dataset, and such gradient values are computed using back propagation.

MLP trains using two input arrays: array X of size(n samples, n features); and array y of size (n samples). All feature vectors comprises of the training samples are held in X & the target values(class labels) for respective training samples are held in y.

Currently, only the cross-entropy loss function is supported by the MLP Classifier [4], using which the estimated probabilities can be derived by running predict proba function. MLP Classifier also supports multi-class classification in which any input feature set can belong to more than one class which makes it quite suitable for this approach.

- Advantages of Multilayer Perceptron Networks:
  - ◦ MLP Networks are capable to run all types of non-linear models.
  - ◦ MLP Classifier uses back propagation so, it learns and improvise itself with passage of time.
- Disadvantages of Multilayer Perceptron Networks:
  - ◦ They use a black box model, interpretation of results may become difficult.
  - ◦ MLP Networks requires a large number of hyper-parameters & thus proper tuning of the number of epochs, hidden neurons and layers is required.

A. Decision Trees

Decision trees are also a type of supervised machine learn- ing techniques where according to a certain parameter input training data is continuously split up. Any decision tree can be explained using two of its entities, decision nodes and leaf nodes. The leaves denote the final outcomes or the overall decisions made and the data is split using some entropy calculation at the decision nodes. Decision trees (DTs) can be used for both classification as well as regression problems. The entire goal is to create a supervised model which can predict the value of any input target variable by making use of the prominent decision rules formulated from the training dataset features.

Given, x_iϙ R^n, *i=1,..., l* are the training vectors and yϙ R^l is the target vector, recursive partitioning of entire dataset is done by the decision tree such that data samples with same target labels get in a single group. Let *Q* represents the data at node

m. For each candidate node, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets using split $\theta\ \overline{(j,t_m)}$ which consists of a feature *j* and threshold t_m,

$$Q_{left}(\theta) = (x, y)|x_j <= t_m \quad (4)$$
$$Q_{right}(\theta) = Q/Q_{left}(\theta)$$

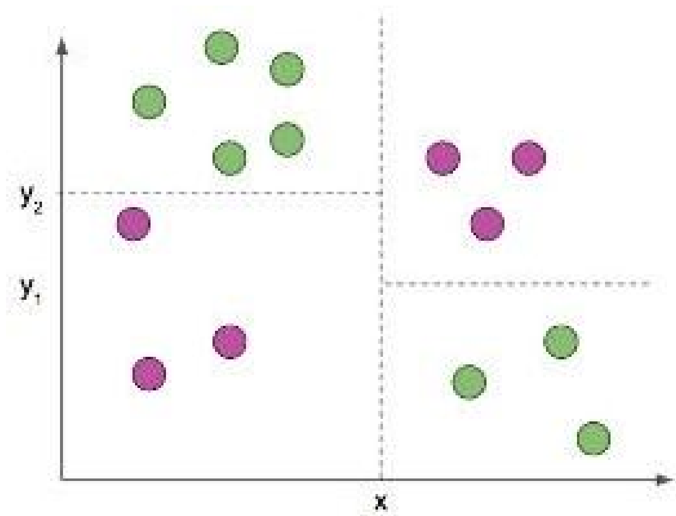


Fig. 2: Decision Tree

An impurity function *H()* is used to compute the impurity at *m* , whose choice depends on the task under consideration (regression or classification).

$$G(Q, \theta) = n_{left}/N_m * H(Q_{left}(theta)) + n_{right}/N_m * H(Q_{right}(theta)) \quad (6)$$

Select the parameters that minimizes the impurity

$$\theta^* = argmin\ G(Q, \theta) \quad (7)$$

Continue partitioning recursively for the subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached,

$N_m < min_{samples}$ or N_m=1.

- Advantages of Decision Trees:

◦ Decision Trees are simple enough to understand, interpret its outcome and visualize the results.

◦ Able to handle both numeric as well as categorical data and also multi-output problems.

◊ The white box model is followed up. If some situation is observable in the model, then its explanation is easily explained using the logic of boolean algebra.

- Disadvantages of Decision Trees:

◦ Sometimes complex trees are created which are not able to generalize the data well. Decision Trees are prone to over-fitting.

◦ Decision Trees are usually very unstable and even small modifications in the data might lead to an entirely different tree being generated.

◦ For the cases, where some classes dominate creation of biased Decision Tree takes place.

III. RESULTS AND OBSERVATIONS

A. Performance Measures

To evaluate classifier performance in a well effective manner, the performance measure needs to be defined. Efficiency and goodness of any classifier is measured by various defined performance measures which is itself a single index.

A comparative analysis of the classifiers has been performed considering the following performance measures:

- Accuracy Score: This compares the actual outcomes with the predicted outcomes of the classifier for a given input dataset. For best accuracy score, the set of actual true labels in testing dataset must match the corresponding set of predicted labels.

For measuring the success of the prediction, precision-recall is a useful index. In information retrieval, result relevancy is measured using precision, while recall is a measure of the total number of truly relevant results which were returned.

- Precision Score: This is defined as the number of true positives (T_p) divided by the number of true positives plus the number of false positives (F_p)

$$P = \frac{T_p}{T_p + F_p} \qquad (8)$$

The precision is the ability of the classification model for not labelling a negative sample as a positive one. Best value: 1 and Worst value: 0.

- Recall Score: This is defined as the number of true positives (T_p) divided by the number of true positives plus the number of false negatives (F_n)

$$R = \frac{T_p}{T_p + F_n} \qquad (9)$$

The recall is the ability of the classification model of finding all the possible positive samples. Best value: 1 and Worst value:0

- F1 Score: This is defined as the interpretation of a weighted average of the recall score and precision score of a classifier. Numerically it is equal to harmonic mean of the precision-score and recall-score.

$$F1 = 2\frac{P*R}{P+R} \qquad (10)$$

It is also known as the F-measure or balanced F-score. Both precision and recall have an equal relative contribution to the F1 score.

- Average Precision Score: This is defined as the weighted mean of precision achieved at each threshold value, summa- rized using precision-recall curve:

$$AP = \Sigma_k (R_k - R_{k-1}) P_k \qquad (11)$$

where R_k and P_k are the recall and precision at the k^{th} threshold.

B. Comparative Analysis

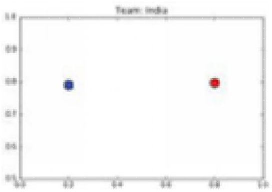
TABLE II: COMPARISION OF ACCURACY SCORES

Multilayer Perceptron Classifier	Decision Tree Classifier
0.574	0.551

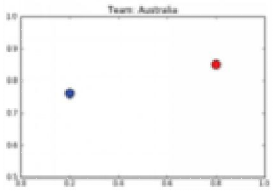
India, Australia and Pakistan were selected randomly and the match records of these 3 teams were separated to obtain their performance measure separately.

TABLE III: SAMPLE DATASET DISTRIBUTION

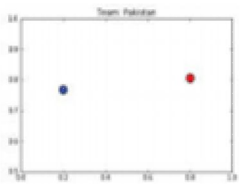
Team Name	Training Dataset Size	Testing Dataset Siz e
India	1320	440
Australia	1288	430
Pakistan	1281	427



(a) P = 0.791, R = 0.797

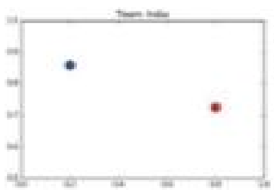


(b) P = 0.760, R = 0.850

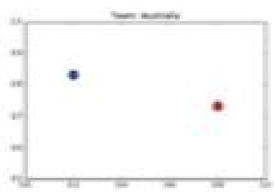


(c) P = 0.767, R = 0.806

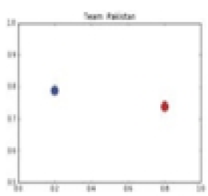
Fig. 3: Precision-Recall Scatter Plot for MLP Classifier



(a) P = 0.859, R = 0.726



(b) P = 0.830, R = 0.733



(c) P = 0.789, R = 0.739

Fig. 4: Precision-Recall Scatter Plot for DT Classifier.

IV. RELATED WORK

From the literature survey, it was observed that game of cricket has very few machine learning related work done on it. Despite sharing numerous features with other sports like baseball, game of cricket is unique of its type and thus an - independent analysis is required.

Statistical approach is the base of majority of the analytical studies & research done on cricket.

TABLE IV: SIMULATION RESULTS

Classifier	Performance Measure	India	Australia	Pakistan
MLP Classifier	Recall Score	0.797	0.850	0.806
	Precision Score	0.791	0.760	0.767
	F1 Score	0.794	0.803	0.786
	Average Precision Score	0.744	0.749	0.724
Decision Tree Cl assifier	Recall Score	0.726	0.733	0.739
	Precision Score	0.859	0.830	0.789
	F1 Score	0.787	0.779	0.763
	Average Precision Score	0.785	0.779	0.719

Prediction of the outcome of an in-progress game in one- day international cricket was conducted by Bailey and Clarke [5]. WASP(Winning and Score Predictor), 2012 is a product grounded on the theory of dynamic programming, by Dr Scott Brooker and Dr Seamus Hogan at the University of Canterbury in New Zealand.

Neeraj Pathak & Hardik Wadhwa conducted a similar com- parative analysis of match outcomes using the classification models: support vector machines, random forests and naive bayes[6]. Preeti Satao and Team predicted the score of cricket match using clustering techniques[7].

In Parag Shah, Mitesh Shah[8] and Amal Kaluarachchi, Aparna S. Varde[9], they explored the statistical significance of a range of factors & game-attributes which explain the outcome of a cricket match. In particular, home crowd ad- vantage, match type (day-night/day), past performance of the team against each other & game plan (batting first or fielding first) were the key interests in their investigation.

Madan Gopal Jhanwar and Vikram Pudi used a supervised learning approach from some team composition perspective for predicting the result of an one-day international (ODI) cricket match. Their work suggested that one of the distinctive features for predicting the winner is the relative team strength of both the competing teams. Swetha and Saravanan.KN analyzed the factors that cricket game depends on and decides winning[1].

V.CONCLUSION

In this study, a comparative analysis of the predictions generated by 2 different supervised classification models was performed for the same input dataset.

The proposed approaches are better than the statistical approach as unlike statistics which uses mathematical equa- tions to formalize the relationships between variables, these approaches require no prior assumptions regarding the data variables and their underlying relationships. During training phase, data needs to be fed in and the algorithm after processing the data discovers patterns and finally makes predictions for freshly generating data.

The major contributions of this study are:

- Comparative analysis of performance measure of two different supervised learning techniques.
- Analyzing all the factors which strive to affect the final outcome of the game of cricket.
- Design & development of a desktop application which can be used to predict the chances of winning, using input attributes.

As future course of work, this analytical study can be expanded further in terms of the team composition perspective. Also, the relevance of considering 1980s match data equivalent to the 2017s match data also need to be analyzed and worked upon. This methodology and technique can also be applied to predict the outcomes of games like hockey and football.

VI. ACKNOWLEDGEMENT

This research was supported by the Department of Com- puter Science and Engineering, AIT, India. I am grateful to all my friends who provided support and insight which assisted me a lot in carrying out this research. I also thank all of them for their worthy comments & criticism which helped me to refine the content and improvise the paper.

REFERENCES

[1] Swetha and S. KN, ”Analysis on Attributes Deciding Cricket Winning”, International Research Journal of Engineering and Technology, vol. 04, no. 03, pp. 1105-1107, 2017.

[2] M. Khan and R. Shah, ”Role of External Factors on Outcome of a One Day International Cricket (ODI) Match and Predictive Analysis”, Inter- national Journal of Advanced Research in Computer and Communication Engineering, vol. 04, no. 06, pp. 192-197, 2015.

[3] ESPNcricinfo - Cricket Teams, Scores, Stats, News, Fixtures, Results, Tables”, ESPNcricinfo, 2017. [Online]. Available: http://www.stats.espncricinfo.com.

[4] “Documentation scikit-learn: machine learning in Python scikit- learn 0.20.1 documentation”, Scikit-learn.org, 2018. [Online]. Available: https://scikit-learn.org/stable/documentation.html.

[5] M. Bailey and S. Clarke, ”Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress”, Journal of sports science & medicine, vol. 05, no. 04, pp. 480-487, 2006.

[6] N. Pathak and H. Wadhwa, ”Applications of Modern Classification Techniques to Predict the Outcome of ODI Cricket”, Procedia Computer Science, vol. 87, pp. 55-60, 2016.

[7] P. Satao, A. Tripathi, J. Vankar, B. Vaje and V. Varekar, ”Cricket Score Prediction System (CSPS) Using Clustering Algorithm”, International Journal of Current Engineering and Scientific Research, vol. 03, no. 04, pp. 43-46, 2016.

[8] P. Shah and M. Shah, ”Predicting ODI Cricket Result”, Journal of Tourism, Hospitality and Sports, vol. 05, pp. 19-20, 2015.

[9] A. Kaluarachchi and S. V. Aparna, ”CricAI: A classification based tool to predict the outcome in ODI cricket,” 2010 Fifth International Conference on Information and Automation for Sustainability, Colombo, 2010, pp. 250-255.

[10] M. Jhawar and V. Pudi, ”Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach”, European Confer- ence on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Riva del Garda, 2016