Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction using Machine Learning

Shashank Singh
B. Tech Student
Department of CSE
Inderprastha Engineering College, Ghaziabad, India

Yash Aggarwal
B. Tech Student
Department of CSE
Inderprastha Engineering College, Ghaziabad, India

Kumud Kundu, PhD
Assistant Professor
Department of CSE
Inderprastha Engineering College, Ghaziabad, India

ABSTRACT

The ICC Men's T20 Cricket World Cup 2020 is scheduled to be hosted by Australia in the month of October and November 2020. Machine Learning in sports analytics is now a days actively applied for prediction of winners. The work presented in this paper aims to predict the winner of the upcoming seventh version of ICC Men's T20 world cup using Random Forest Classifier, Naïve Bayes, KNN, Logistic Regression, Decision Tree, SVM, Bagging Classifier, Extra Trees Classifier, Voting (HARD & SOFT) training. All these approaches are tested on the different available historic data of international cricket matches played between different countries from 2005 to March 2020. Unstructured historic cricket statistics is picked from ESPN and Cricbuzz websites. Experimental results prove that all approaches are able to imbibe the extracted patterns from the various set of matches performed and hence is found suitable to predict the winner of the ICC Men's T20 Cricket World Cup 2020. A comparative study is also presented for the predictions made through different approaches.

General Terms

Machine Learning, Match Outcome.

Keywords

Cricket analytics, Winner Prediction, Classification.

1. INTRODUCTION

In India, Cricket is one of the most popular sport that has huge fan base. This sport is played at the national as well as international level. At international level, Indian national team plays different bilateral matches with the national team of other countries as well as it plays in the World Cup tournaments. Cricket's World Cup tournaments are organized by governing body: International Cricket Council (ICC). Cricket world tournaments are open to all its council members. The highest-ranking teams as per the ICC ranking list qualify to the world cup tournaments automatically. However, the remaining teams are picked through the ICC world cup qualifier matches. Cricket world cup tournaments is currently being held in two formats: T20 world cup tournament and ODI cricket world cup tournament. First format of cricket international championship is held every two years and another one is held every four years. The ICC Men's T20 World Cup (earlier known as ICC World Twenty20) is the international championship played by the sixteen teams. Out of sixteen teams, ten teams qualify directly into the tournament on the basis of the ranking furnished by the ICC while the other six other teams are chosen through the T20 World Cup Qualifier. The next edition of T20 world cup tournament is planned to be held in Australia from 18 October to 15 November 2020. As per the ICC ranking of 31 December 2018, the top nine ranked teams, along with host country Australia, had qualified directly for the 2020 tournament. Out of those ten qualifier teams, the top eight ranked teams had qualified for the Super 12 stage of the tournament while the teams of Sri Lanka and Bangladesh were placed in the group stage of the competition. Six other teams who had qualified for the tournament through the ICC T20 World Cup Qualifiers also include the United Arab Emirates and Nepal.

Recently, machine learning has been applied abundantly for analyzing the sport outcomes based on the data gathered from the past played games. The learning process involves training the model based on previous matches played, then the developed model gets evaluated on an independent future match to measure its effectiveness [1]. However, doing analysis and predictions from the huge volume of unstructured data generated from the matches played during a tournament poses a big challenge in itself. This paper presents the methodology of predicting the winner of the upcoming men's T20 Cricket World Cup 2020 using machine learning. The prediction accuracies of various machine learning algorithms are compared to study their performance.

The paper is organized as follows. Section 2 provides a brief literature survey of the approaches used in cricket analytics. Section 3 describes our methodology for predicting the winner of the tournament. In Section 4 results are presented while Section 5 presents the conclusion of the paper and puts forth the further direction of work that can be done to improve results further.

2. LITERATURE SURVEY

Since past two decades, Machine Learning has become one of the active ingredients in improving the performance of computer system. In the field of sports, Machine learning has paved its path in the prediction of performance of team as well as team members. Pathak et al. [2] developed COP (Cricket Outcome Predictor) where they applied Naïve Bayesian classifier for the prediction of the probability of win/loss of an ODI match. Passi et al. [3] applied four machine learning algorithms: Naïve Bayes, Decision Trees, Random Forest and Support Vector Machine for predicting the runs scored by a batsman and wickets taken by a bowler in ODI match. They found among the four machine learning algorithms that they considered, Random Forest turns out to be the most accurate classifier for the prediction of the runs scored by a batsman and the prediction for predicting wickets taken by a bowler. Jayalath [4] studied the effect of predictors like: home-field advantage, coin toss result, bat-first or second, day vs day-night game etc that affect the prediction of ODI outcome using ‘classification and regression tree’ CART-based approach. They found that the home-field advantage factor impacts the winning probability significantly for various teams like India, South Africa, Sri Lanka, New Zealand, and Pakistan. It was further found that for the South African team, home ground factor affects the winning probability more than the other teams. Viswanadha et al. [5] modelled the parameters affecting relative strength of the two teams playing the T20 match. In their work, they estimated the team’s strength by modeling the potential of the individual participating players by utilizing player’s recent performance statistics. The modelled features were evaluated using several supervised learning algorithms for the prediction of the T20 match winner. Rupai et al. [6] studied the impact of bowling performance on the match winning probability. They utilized various machine learning based approaches to predict the bowler’s performance in various instances of ODI matches. Wickramasinghe [7] utilized three machine learning classifiers, namely Naive Bayes (NB), k-nearest neighbours (kNN), and Random Forest (RF) for the classification of all-round player’s performance in ODI matches. Modekurti et al. [8] developed a deterministic model for the determination of the target score of the team batting first in T-20 Indian Premier League (IPL) match.

3. PROBLEM FORMULATION & PROPOSED SOLUTION

To study the classification accuracy of popular machine learning algorithms for the prediction of winner of upcoming ICC T20 Men’s world cup. Outcome of each machine learning algorithm deals with the hypothesis question: ‘‘***H=Computed feature score is probable to put forth as a future ICC T20 tournament winner or not?***’’.

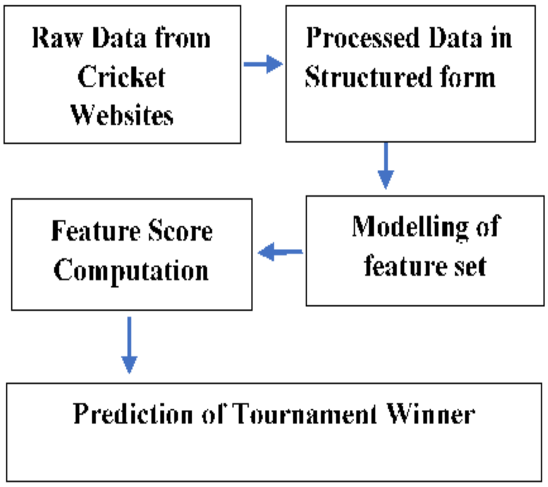

Raw Data from Cricket Websites
Processed Data in Structured form
Modelling of feature set
Feature Score Computation
Prediction of Tournament Winner

Fig 1: Proposed Solution

Hypothesis testing incorporates six phases i.e. data acquisition, data validation (if required), feature modelling, extraction of correlations among the modelled feature set, design of: train data set, test data set, validation data set, data model execution with selected machine learning algorithm. With regards to research, the data is known to be validated as both data sources are trustworthy. there is no invalid data to remove. The dataset acquired from ESPNcricinfo [9] can be classified as enterprise datasets as the data structure is predefined and the dataset is pre-validated.

3.1 Datasets

This paper is adopting datasets from different online available resources ESPNcricinfo, Cricbuzz. Most of the datasets that it will be using would be of the structured type, as the cricket statistics are data that conform to a data model and is often not stored in tabular form. Following datasets are used to derive experimental results:

3.1.1 T20 Stats Dataset

Table 1 below shows an example of the results of a statistics query team in T20s for the past 15 years (2005 to present time) from ESPNcricinfo [10]. The dataset T20 stat shows the data of every single T20 match played from 2005 to 2020 and contains a total of 1070 records. This table stores useful information such as the two teams, the winner, and the margin of the win, the venue, and the date.

Table 1. Historical T20 matches

Date	Team_1	Team_2	Winner	Margin	Ground
Feb 17, 2005	New Zealand	Australia	Australia	44 runs	Auckland
Jun 13, 2005	England	Australia	England	100 runs	Southampton
Oct 21, 2005	South Africa	New Zealand	New Zealand	5 wickets	Johannesburg
Jan 9, 2006	Australia	South Africa	Australia	95 runs	Brisbane
Feb 16, 2006	New Zealand	West Indies	tied	NaN	Auckland
...	...	...	...	...	...
Feb 26, 2020	South Africa	Australia	Australia	97 runs	Cape Town
Feb 27, 2020	Kuwait	U.A.E.	U.A.E.	102 runs	Al Amerat
Feb 29, 2020	Thailand	Singapore	Singapore	43 runs	Bangkok

3.1.2 T20 Fixtures Dataset

Detailed fixtures of the T20 World Cup are stored in T20 Fixtures dataset. This dataset stores information like date, venue, team name, group, round. T20 fixtures dataset is derived from the Cricbuzz website [11] which has the detailed list of fixtures of upcoming T20 world cup. The fixture data available at Cricbuzz website was in raw format, that’s why the data was converted into the structured format. Table 2 shows an example of this. It presents every fixture record with attributes like date, team name.

Table 2. Fixtures of the Forthcoming T20 World Cup 2020

Date	Location	Team_1	Team_2	Group
18-Oct-20	Simonds Stadium	Sri Lanka	Ireland	Group A
18-Oct-20	Simonds Stadium	Papua New Guinea	Oman	Group A
19-Oct-20	Bellerive Oval	Bangladesh	Namibia	Group B
19-Oct-20	Bellerive Oval	Netherlands	Scotland	Group B
20-Oct-20	Simonds Stadium	Ireland	Oman	Group A
20-Oct-20	Simonds Stadium	Sri Lanka	Papua New Guinea	Group A
21-Oct-20	Bellerive Oval	Namibia	Scotland	Group B
21-Oct-20	Bellerive Oval	Bangladesh	Netherlands	Group B
22-Oct-20	Simonds Stadium	Papua New Guinea	Ireland	Group A

3.1.3 ICC Ranking Dataset

All 16 teams participating in the upcoming world cup in 2020 have their respective rankings allotted by the ICC [12]. ICC Team Rankings is computed from the score points scored by a team on the basis of the results of the matches played by a team over the last 3−4 years. For instance, the ICC rankings of 16 teams is shown in Table 3.

Table 3. ICC Rankings and Points [12]

Position	Team	Points
1	Australia	278
2	England	268
3	India	266
4	Pakistan	260
5	South Africa	258
6	New Zealand	242
7	Sri Lanka	230
8	Bangladesh	229
9	West Indies	229
10	Afghanistan	228
12	Ireland	190

3.1.4 Master Dataset

Apart from the standard datasets, another dataset is designed as a master data set based on T20 stats dataset and ICC rankings dataset. This data set stores the computed feature score from the modelled feature set for every team. This particular dataset is built in order to summarize every vital feature of a cricket team like number of T20 matches played till date, winning percentage in those matches, no. of matches played in Australia, etc. Table 4 shows the feature score of every team while prioritizing every variable according to its importance. Features in the Mater Dataset are:

- ICC Ranking Rating (T20)
- Overall Matches Played (T20)
- Overall Winning Percentage (T20)
- Mean Batting Average
- Average Bowling Economy
- Matches Played in Australia (T20)
- Winning Percentage in Australia (T20)
- Matches Played after ICC World Cup 2019
- Winning Percentage after ICC World cup 2019

Table 4. Master Dataset

Team	IRR	OMP	OWP	MBA	ABE	MPA	WPA
Australia	278	125	55.73	25	2.64	43	62.79
England	268	117	52.21	30	1.74	8	12.5
India	266	134	65	33	2.84	9	55.55
Pakistan	260	151	62.33	29	3.12	4	0
South Africa	258	121	58.75	25	1.79	7	28.57
New Zealand	242	131	50.78	24	1.8	3	0
Sri Lanka	230	128	47.61	15	2.24	9	55.55
Bangladesh	229	96	34.04	23	1.91	0	0
West Indies	229	124	46.63	20	1.77	3	33.34
Afghanistan	228	81	68.71	25	3.06	0	0
Ireland	190	98	46.15	28	2.56	0	0
Scotland	182	65	47.58	29	2.14	0	0
Papua New Guinea	179	26	68	23	4.47	0	0
Netherlands	178	75	54.86	25	2.81	0	0

3.2 Modelling of Feature Set

- ICC Ranking Rating (IRR): This particular feature shows the current ICC T20 Team ranking of all the teams participating the T20 World Cup 2020.
- Overall Matches played (OMP): It shows the total matches played till date from 2005 till 2020.
- Overall Winning Percentage (OWP): This feature is calculated by dividing the Overall Matches Played by the Overall matches won (OMW) by the team i.e.

OWP = OMP/OMW……………… (1)

- Mean Batting Average (MBA): It shows the average no. of runs a batsman score. This average is calculated on the basis of maximum runs scored by the best 3 batsmen of a team for last 2 years.
- Average Bowling Economy (10 - BE): It shows the average number of runs a bowler concedes in an over. This feature is also calculated by the max. no of runs by the best 3 bowlers for the last 2 years.
- Matches played in Australia (MPA): It shows the number of T20 matches played by the particular team in Australia.
- Winning percentage in Australia (WPA): This feature is calculated by dividing the MPA by the number of matches won by the team in Australia (MWA). This feature is used since the T20 World Cup 2020 will be played in Australia i.e.

WPA = MPA/MWA……… ...(2)

- Matches played after 2019 World Cup (MPN): It shows the number of matches played by any team after the 2019 World Cup. This particular feature is used to analyze the team performance and confidence during this time.
- Winning Percentage after 2019 World Cup (WPN): This feature is calculated by dividing MPN by the matches won after 2019 World Cup (MWN).

WPN = MPN/MWN ……………… .(3)

- The final feature score is computed by the following formula:

Feature Score (FS) = ICC ranking*a + OMP*b + OWP*c + MBA*d + ABE*e + MPA*f + WPA*g + MPN*h + WPN*i ……………………… …(4)

Where a, b, c, d, e, f, g, h, i are empirical variables whose values are experimentally taken as: a = 0.25, b = 1.5, c = 1.5, d = 1.25, e = 1.25, f = 1.5, g = 1.5, h = 1.25, i = 1.25. These empirical variables are used to prioritize the individual features described above in the calculation of the final feature score.

3.3 PREDICTION ALGORITHM

To predict the winner of tournament ICC Men’s T20, popular machine learning classifier like: Random Forest, Naïve Bayes, KNN, Logistic Regression, Decision Tree, SVM, Bagging Classifier, Extra Trees Classifier [13] have been applied. Following tasks are performed on datasets before applying prediction algorithm:

- Importing the required Python Libraries
- Importing and reading the **T20 Stat Dataset** containing information about the historical T20 matches

played till date.

- Importing and reading the **T20 Fixture Dataset** having the fixtures for the upcoming ICC T20 World Cup 2020.
- Narrowing down the T20 Stat Dataset to those teams which will be playing the World cup.
- Dropping attributes or columns like **Ground** and **Margin** that will not affect match outcomes.

Table 5. Dataset with Dropped Attributes

S.No	Team_1	Team_2	Winner
1	New Zealand	Australia	Australia
2	England	Australia	England
3	Australia	South Africa	Australia
4	New Zealand	West Indies	tied
5	England	Sri Lanka	Sri Lanka
…	…	…	…
1060	South Africa	England	South Africa
1061	South Africa	England	England
1062	South Africa	England	England

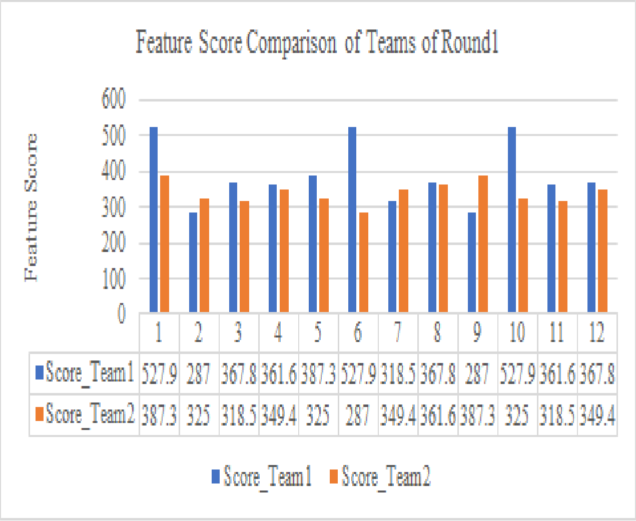
4. EXPERIMENTAL RESULTS

This section presents the experimental results with the above-mentioned datasets and the prediction algorithms.

4.1 Prediction on the basis of Feature Score Ranking

Table 6 below shows the Prediction Set where the 'team feature score' is embedded with its name

Table 6: Prediction of Round 1 on the Basis of Feature Score

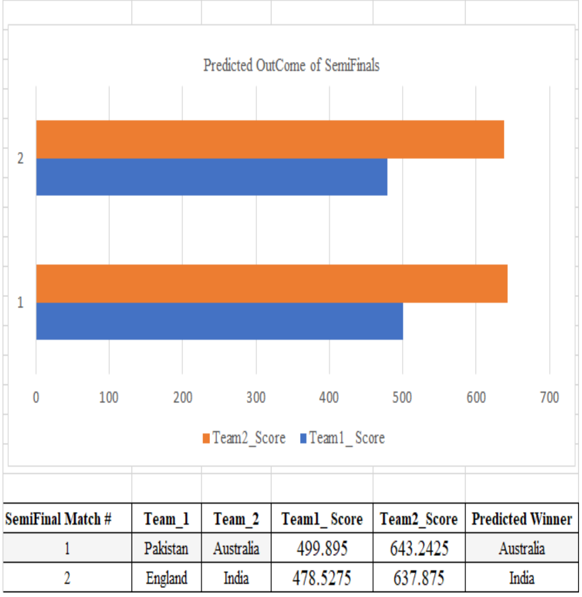


	1	2	3	4	5	6	7	8	9	10	11	12
Score_Team1	527.9	287	367.8	361.6	387.3	527.9	318.5	367.8	287	527.9	361.6	367.8
Score_Team2	387.3	325	318.5	349.4	325	287	349.4	361.6	387.3	325	318.5	349.4

Team_1	Team_2	Predicted Winner
Sri Lanka	Ireland	Sri Lanka
Papua New Guinea	Oman	Oman
Bangladesh	Namibia	Bangladesh
Netherlands	Scotland	Netherland
Ireland	Oman	Ireland
Sri Lanka	Papua New Guinea	Srilanka
Namibia	Scotland	Scotland
Bangladesh	Netherlands	Bangladesh
Papua New Guinea	Ireland	Ireland
Sri Lanka	Oman	Srilanka
Netherlands	Namibia	Netherland
Bangladesh	Scotland	Bangladesh

4.2 Match Outcome Prediction

Figure 2 below shows the snippet of the Prediction of the Group Matches of the ICC T20 World Cup 2020 on the basis of their respective Feature Score.

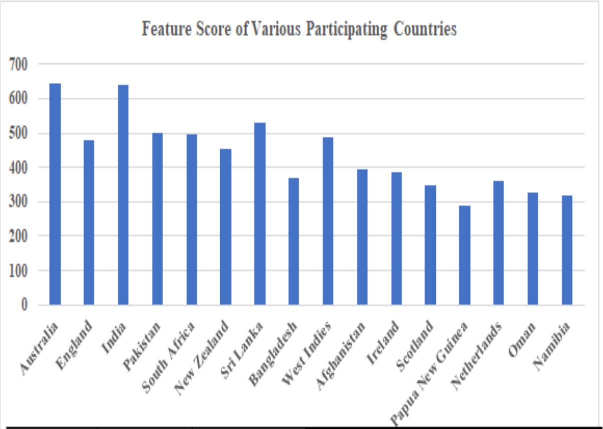


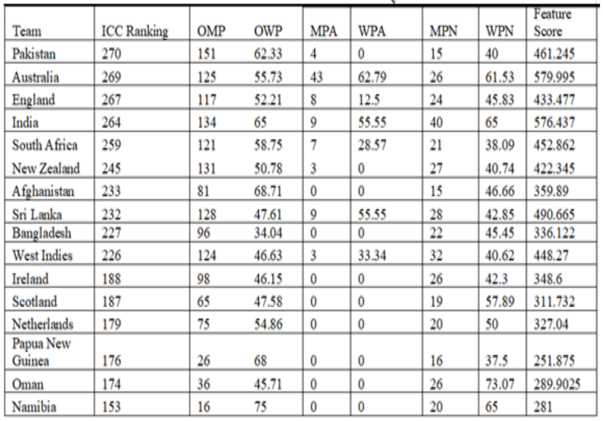
SemiFinal Match #	Team_1	Team_2	Team1_Score	Team2_Score	Predicted Winner
1	Pakistan	Australia	499.895	643.2425	Australia
2	England	India	478.5275	637.875	India

Fig 2: Predicted Winner of Semi Finals

4.3 Feature Score Comparison

Figure 3 below shows the comparison of all the teams participating in the T20 World Cup 2020. As shown below Australia and India have the best and the 2nd best feature score respectively.





Team	ICC Ranking	OMP	OWP	MPA	WPA	MPN	WPN	Feature Score
Pakistan	270	151	62.33	4	0	15	40	461.245
Australia	269	125	55.73	43	62.79	26	61.53	579.995
England	267	117	52.21	8	12.5	24	45.83	433.477
India	264	134	65	9	55.55	40	65	576.437
South Africa	259	121	58.75	7	28.57	21	38.09	452.862
New Zealand	245	131	50.78	3	0	27	40.74	422.345
Afghanistan	233	81	68.71	0	0	15	46.66	359.89
Sri Lanka	232	128	47.61	9	55.55	28	42.85	490.665
Bangladesh	227	96	34.04	0	0	22	45.45	336.122
West Indies	226	124	46.63	3	33.34	32	40.62	448.27
Ireland	188	98	46.15	0	0	26	42.3	348.6
Scotland	187	65	47.58	0	0	19	57.89	311.732
Netherlands	179	75	54.86	0	0	20	50	327.04
Papua New Guinea	176	26	68	0	0	16	37.5	251.875
Oman	174	36	45.71	0	0	26	73.07	289.9025
Namibia	153	16	75	0	0	20	65	281

Fig 3: Feature Score Comparison of Top 16 Teams

4.4 Prediction Comparison

Figure 4, 5, and 6 below show the prediction comparison results of top three instances with the modelled feature set. In every instance Random Feature algorithm is able to give more score and hence a better picture of prediction outcome.

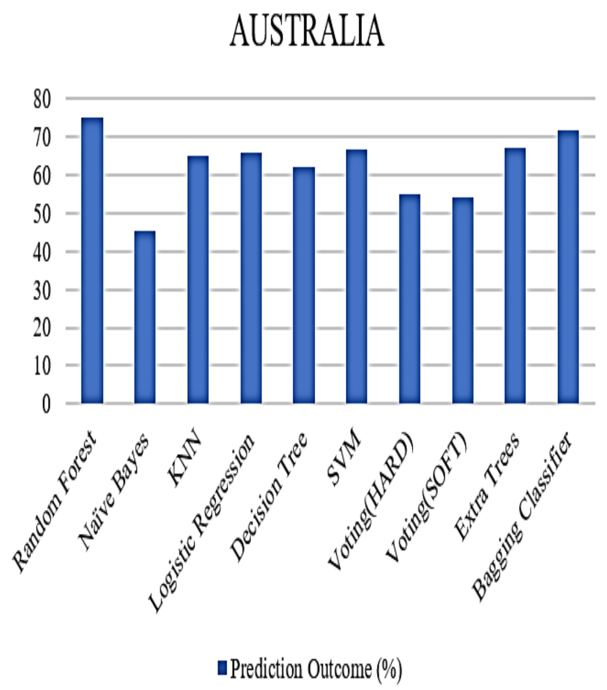


Fig 4: Accuracy Comparison of Prediction Algorithms for Australia

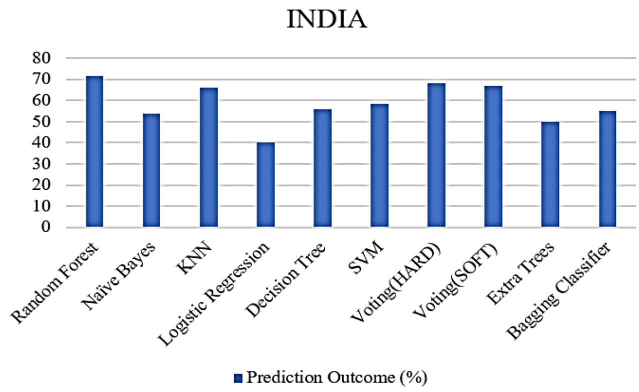


Fig 5: Accuracy Comparison of Prediction Algorithms for India

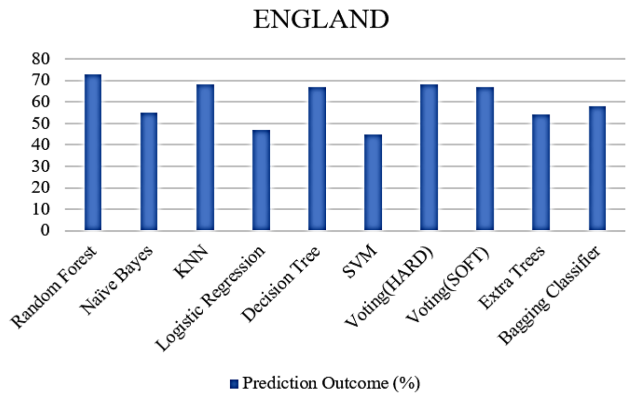


Fig 6: Accuracy Comparison of Prediction Algorithms for England

4.5 T20 World Cup Winner Prediction

A system has been created using machine learning in which team with higher Feature Score and previous T20 World Cup records are provided with higher priority in winning a particular match. The team with the higher Feature Score and previous record is predicted to win the world cup. As a result, prediction indicates that Australia has the highest chances to win the world cup and India being in the next possibility being with the 2nd highest Feature score. India trails by a difference of 5.3675 behind Australia in feature score.

5. CONCLUSION

This study is an exploratory study aimed to predict the winner of upcoming ICC T20 men’s tournament by utilizing few prominent parameters associated with T20 previous matches. A feature score was computed from the selected parameters to test the formulated hypothesis of predicting probable winner. It is apparent from the accuracy comparison chart that Random Forest algorithm gives the highest prediction as it is able to find the most important feature from the selected feature set and is less prone to overfitting.

6. REFERENCES

[1] Munir, F., Hasan, M.K., Ahmed, S., Md Quraish, S., 2015. Predicting a T20 cricket match result while the match is in progress (Doctoral dissertation, BRAC University).

[2] Pathak, N., and Wadhwa, H. 2016. Applications of modern classification techniques to predict the outcome of ODI cricket. *Procedia Computer Science*, *87*, 55-60.

[3] Passi, K., & Pandey, N. 2018. Increased prediction

accuracy in the game of cricket using machine learning. *arXiv preprint arXiv:1804.04226*.

[4] Jayalath, K. P. 2018. A machine learning approach to analyze ODI cricket predictors. *Journal of Sports Analytics*, *4*(1), 73-84.

[5] Viswanadha, S., Sivalenka, K., Jhawar, M. G., & Pudi, V. 2017. Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths. In *MLSA@ PKDD/ECML,* 41-50.

[6] Rupai, A. A. A., Mukta, M. S. H., & Islam, A. N. 2020. Predicting Bowling Performance in Cricket from Publicly Available Data. In Proceedings of the International Conference on Computing Advancements, 1-6.

[7] Wickramasinghe, I. Classification of All-Rounders in the Game of ODI Cricket: Machine Learning Approach.

[8] Modekurti, D. P. V. Setting final target score in T-20 cricket match by the team batting first. *Journal of Sports Analytics*, (Preprint), 1-8.

[9]]https://stats.espncricinfo.com/ci/engine/records/index.html?id=89;type=trophy.

[10]]https://www.cricbuzz.com/cricket-series/2798/icc-mens-t20-world-cup 2020/stats#!/?statsType=mostRuns&seriesType=WCT20&seriesId=2798.

[11] https://www.cricbuzz.com/cricket-series/2798/icc-mens-t20-world-cup-2020/matches.

[12] https://www.icccricket.com/rankings/mens/ team-rankings/t20i