JOURNAL OF STATISTICS, OPTIMIZATION AND DATA SCIENCE

VOL NO.2 , ISSUE 1 (JUNE 2024), ,PAGE NO.8-15 ,ISSN:2583-9462

Prediction of Outcome of Twenty20 (T20) Cricket Using Classification Techniques

Kumarapandiyan G[a], Keerthivarman S[b], Aswini R[c],Malolan T A[d] and Prajesh S[e]

[a,c,d,e] *Department of Statistics, Madras Christian College, Chennai, India.*
[b]*CRICVIZ, In-ground Data Collection Analyst, United Kingdom.*

ARTICLE HISTORY
Compiled June 2024

ABSTRACT
In the field of sports analytics, cricket stands out as a fascinating domain where data science ignites unparalleled curiosity and drives innovation. In our study, we delve into the dynamic realm of Twenty20 (T20) cricket, leveraging advanced data science techniques to predict match outcomes with unprecedented accuracy. Harnessing the inputs—such as first-inning performance, venue characteristics, toss outcomes, match timing, and partnership scores—we endeavor to forecast match results midway through the game. Our model draws strength from a comprehensive training dataset spanning T20 matches from 2011 to 2021, ensuring robustness and reliability in its predictions. Employing the machine learning algorithms like Random Forest, Naïve Bayes, and Support Vector Machine (SVM), we meticulously analyze and compare their efficacy in this predictive task. Notably, the Random Forest classifier emerges as a beacon of performance, excelling across diverse scenarios with its adept handling of both continuous and categorical variables. We have meticulously crafted an intuitive dashboard using R-Shiny. This user-friendly interface empowers to make informed decisions swiftly, transforming raw data into actionable intelligence. In essence, the paper not only showcases the symbiotic relationship between cricket and data science but also underscores the transformative prospective of predictive analytics in shaping the future of sports strategy and decision-making.

KEYWORDS
Runs, Accuracy, Team, Kappa value, Prediction and Outcome

1. Introduction

Cricket has become a part of social lives for the past three centuries, attributable to the exhibition of peculiarities like teamwork, strategic plays, and simultaneous usage of various physical skills of the player. The T20 format of cricket was introduced in 2003, in which each of the two teams will have one inning each with a maximum of twenty overs aside; the bowler is restricted with a maximum of four overs each. It has a shorter span of three hours compared to the other two forms of cricket and was introduced to create a fast-paced game that would attract more viewers. Although the

CONTACT Kumarapandiyan G[a]. Email:kumarapandiyan@mcc.edu.in

rules are the same as for the test and One Day International (ODI), it is often considered by many as the most uncertain format of the game, encouraging both layman and professional analysts to develop their interest in the process of prediction. This format of the game has taken the configuration of Indian Premier League and Syed Mushtaq Ali Trophy in India, T20 Blast in England, and similar other domestic competitions. Being a multiplayer outdoor game, one may observe both internal factors as well as external factors that affect the game. Internal factors include player performance and team performance as a whole. External factors may be enlisted as fall of wickets, advantages from home ground, the outcome of tosses, dew factor caused by the time of the match, and the rest. The International Cricket Council has five Asian countries, two European nations, three African countries along with two Australian countries as its permanent members. The fourteen countries considered under analysis are, Afghanistan, Australia, Bangladesh, England, India, Ireland, Netherlands, New Zealand, Pakistan, Scotland, South Africa, Sri Lanka, West Indies, and Zimbabwe. These are the teams with the highest rankings rated by the International Cricket Council (ICC).

2. Review of Literature

Predicting the result of a match is a crowd puller in literature. Many problems have been addressed in predicting the outcome of matches using Machine Learning (ML). We can discuss a few important works in literature. Kamble (2021) have included a detailed discussion over win percentage prediction of 50 over matches based on players alone. Mittal et al. (2021) aim to compare the accuracy of various machine learning techniques from their past usage and find the most appropriate one, considering the available data provision, expense. Kapadia et al. (2019) approach Indian Premier League (IPL) as a sport invoking considerable profit in India. The best predictive model concerning the accuracy, precision, and recall evaluation measures has been discussed.

There have also been approaches such as that of Wickramasinghe (2020a) to apply machine Learning for player classification, wherein Random Forest was proven to be the best approach. Raju (2021)‘s approach is a more direct way to analyse data on matches won and lost based on features like ground and month. Along with wise, a big data approach for match analytics by Awan et al. (2021), uses linear regression for score analysis and spark framework for a quantitative approach and considered multiple approaches in ML, and tries to perform a comparative analysis.

Raja et al. (2021) compared k-nearest neighbors (K-NN), Random Forest, Gradient boosting, and decision tree classifier to compare accuracy and error. Sudhamathy and Meenakshi (2020) aim at understanding the dataset of the past 10 year’s history of the IPL data and four different machine learning algorithms working principles and their implementation in ‘R’ have been described. Singh et al. (2020) has utilized structured data to predict the outcome of T20 matches; along with the popular machine learning techniques, it utilizes voting and bagging and compares their accuracy.

Pathak and Wadhwa (2016) closely analyses the three modern classification techniques of Naïve Bayes Random Forest and SVM to create a Cricket Outcome Predictor (COP) using the R software. They predicted the outcome of the ODI matches and compared the balanced accuracy and Kappa statistic of Naïve Bayes, Random Forest, and Support Vector Machine Classification.

Shenoy et al. (2020)'s approach is more player-performance oriented and brings in the Elo-based approach. They have considered the strike rate, not outs, Economy, Maiden Over, No balls, and wides to rate the player. But the authors have agreed that having a limited dataset and not having to test the accuracy of various approaches can act as a disadvantage in the project.

For further discussion about the data analytics approach in cricket, one may refer to Wickramasinghe (2020b), Passi and Pandey (2018), Sinha (2020), Goel et al.(2021), and Kumarapandiyan and Keerthivarman (2019).

The above mentioned works have motivated us to predict the outcome of the international T20 match once the first innings done using the Random forest, SVM and Naïve Bayes. Further we intended to create dashboard to predict the outcome of T20 matches using the above mentioned models.

3. Data and Methodology

Totally six hundred sixty (660) T20 matches have been played in the last ten years. For the analysis, we have considered six hundred thirty-six (636) T20 matches that took place from January 2011 to November 2021which involved full member nations of the International Cricket Council (ICC). We left out twenty-four (24) matches which are rain interrupted and ended with super over. Details from these matches can be found via the Archive link at the CricInfo website http://statts.espncricinfo.com.We have considered the Outcome of the matches as the dependent variable and the following variables as our independent variables.

Venue of the match: The players of a team are observed to get advantages of playing on certain grounds, based on whether the ground is at their home country or a country of their comfort level. Under this assumption, grounds can be Home (H), Away (A), or Neutral (N). This factor has a minimal weight-age to the outcome.

Mode of the match (Day/ Night): The performance of a team can be affected by the presence of dew, pitch type, and other climatic changes that might differ for day and night. So the game played during days may differ from its peer conducted at night.

Toss Outcome: A toss is done at the beginning of every match. A team will know about its inner structure and whether its strength occurs in its batsmen or bowlers. Hence getting the advantage of toss, would mean an opportunity to optimize their performance.

Partnership Score: This type of prediction is structured to occur after the first innings of a match. The bowling team of a match aims to hit the bails placed upon the stumps guarded by the batsmen. The outcome of a match can also depend on the highest score attained by two-wicket partnerships.

Totalscore: The score of the team that bats first is taken as our total score.

In this study, we have used three different machine learning techniques namely Random Forest, Support vector machine (SVM), and Naïve Bayes. We can discuss about three techniques.

Random Forest: Random forest algorithm is a combination of Decision Trees (DT). There are two types of learning algorithms in machine learning – supervised and unsupervised learning. The Random forest algorithm is a typical example of the supervised learning algorithm. It is constructed from DT and combines many algorithms to classify the value under the various category or predict the outcome as a numeric variable. It predicts based on the average from the output of DT. The random forest algorithm makes it based on ensemble learning. It uses a combination of multiple classifiers to solve complex problems, hence ensuring improvise the overall model's accuracy. Random forests are often defined as a collection of DT and use the average of prediction from all trees to make the final prediction. So, increasing the number of trees can also increase the accuracy and prevents the problem of overfitting.

Support Vector Machine: Similar to Random Forest, Support Vector Machine (SVM) is useful in both regression and classification. However, classification problems find the SVM method to be more useful. In a classification problem, items are segregated using hyperplanes, which graphically segregate the data based on similarities and differences between them. The vectors that lie close to the hyperplane are called support vectors. The objective of an SVM algorithm is to maximize the distance between vectors and hyperplane, called margin. A longer margin would imply a better predicted model. In R software, the e1071 library contains the SVM.

Naïve Bayes: Naïve Bayes is an algorithm useful in the case of binary and multiclass classification. These are also known as simple Bayes or independence Bayes algorithms. The working principle behind this model is the Bayes principle. The principle assumes that the existence of a feature is distinct to other features present in the data. That is, the features are independent. The concept of conditional probability is used to measure the chance occurrence of an event on the condition of occurrence of another event. Byes theorem expresses the conditional probability of a class of events, C-k given that an event X occurs as,

$$P(C_k \mid X) = \frac{P(X \mid C_k) \cdot P(C_k)}{P(X)}$$

The technique is more efficient in the case of categorical variables as compared to numerical variables. It gives predictions based on the probability of the presence of the objects.

4. Analysis and Interpretation

Factors like venue of the match, day/night mode, toss outcome, Partnership score (first and second wicket), Total score of first innings are utilized to predict an outcome- win or lose after the first innings, using each of these three algorithms. After which we developed a dashboard called T-20 Fore-teller using R shiny to predict the outcome. As stated earlier, we have included the Naïve Bayes, Random Forest, and SVM algorithms to predict the outcome of this model for each team.

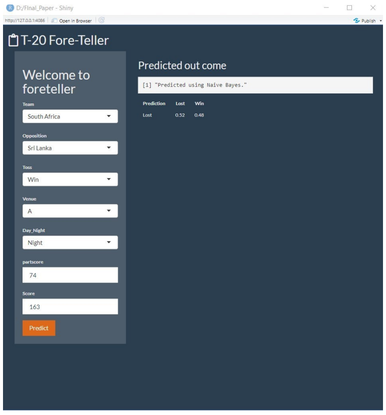
Table 1. Accuracy and Kappa value of three models for each team.

Teams	Accuracy			Kappa		
	Naive Bayes	Random Forest	SVM	Naive Bayes	Random Forest	SVM
Afghanistan	0.6000	0.9091	0.9091	0.2000	0.6207	0.0000
Australia	0.5000	0.7500	0.6667	0.1379	0.5000	0.3333
Bangladesh	0.6000	0.5833	0.5000	0.2000	0.1176	-0.2857
England	0.6000	0.5000	0.4167	0.0909	0.0000	-0.1667
India	0.6429	0.6842	0.7895	0.2857	0.3523	0.5682
Ireland	0.8333	0.5714	0.4286	0.5714	0.1600	0.0000
New Zealand	0.7692	0.6667	0.5000	0.4179	0.2603	-0.1408
Pakistan	0.8571	0.6842	0.7895	0.6500	0.3523	0.5529
South Africa	0.4375	0.4500	0.4500	-0.2414	0.0179	0.1129
Sri Lanka	0.6667	0.6250	0.8125	0.4000	0.2381	0.6129
West Indies	0.7273	0.5000	0.6875	0.4762	0.0154	0.4118
Zimbabwe	0.4000	0.5455	0.5455	0.0909	0.0351	-0.2791
Netherlands	0.6000	0.3333	0.6667	0.0000	-0.5000	0.0000
Scotland	0.4000	0.2000	0.2000	-0.3636	0.0000	0.0000
Average	0.6167	0.5716	0.5973	0.2082	0.1550	0.1228

The accuracy can be defined as a metric used to evaluate the goodness of a binary classifier. From table 1, we infer that the mean values of the accuracy of Naïve Bayes, Random Forest, and SVM are 0.6167, 0.5716, and 0.5973 respectively. The Kappa value represents how well various independent predictors agree with each other about their predictions. A higher kappa value implies better prediction models. A negative value in the kappa table implies an agreement worse than expected or disagreement. The mean value of Kappa statistics is, 0.22083, 0.1550, and 0.1228 for Naïve Bayes, Random Forest, and SVM respectively. The accuracy ranges from 0.2 to 0.9091 whereas the Kappa value ranges from -0.1408 to 0.6207.

5. T-20 Fore-Teller

Using the R- shiny, a dashboard was developed for each of the three methods, Naïve Bayes Random Forest and SVM. To predict the outcome, the user may input the team, its opposition; toss outcome, day/night mode, partnership score, and the runs scored by the team in the first innings. The system outputs the probability to win and losing. We have considered two bilateral series for predicting outcome. One series played between Sri Lanka (SL) and South Africa (SA) in September 2021 in Sri Lanka and another one series played between India (IND) and New Zealand (NZ) in November 2021 at India The snapshots of the predicted outcome of the matches between South Africa and Sri Lanka held on 10-Sep-2021 have been given below for three models.



T-20 Fore-Teller

Welcome to foreteller

Team
South Africa

Opposition
Sri Lanka

Toss
Win

Venue
A

Day_Night
Night

partscore
74

Score
163

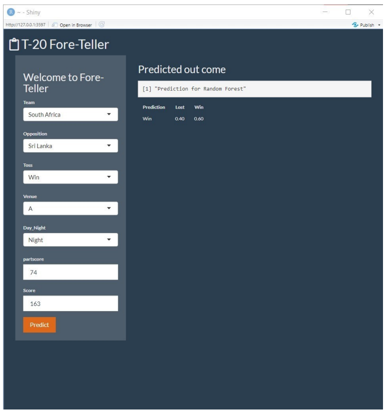
Predict

Predicted out come

```
[1] "Predicted using Naive Bayes."
```

Prediction	Lost	Win
Lost	0.52	0.48

Figure 1. Snapshot of the predicted outcome of the match between South Africa and Sri Lanka using Naïve Bayes in T-20 Fore-Teller Dashboard



T-20 Fore-Teller

Welcome to Fore-Teller

Team
South Africa

Opposition
Sri Lanka

Toss
Win

Venue
A

Day_Night
Night

partscore
74

Score
163

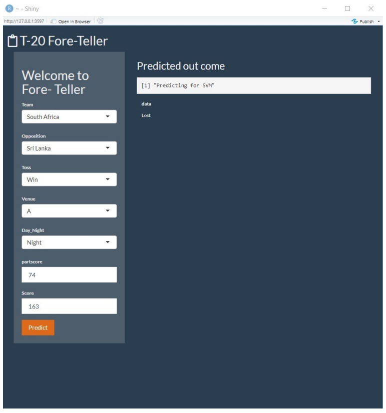
Predict

Predicted out come

```
[1] "Prediction for Random Forest"
```

Prediction	Lost	Win
Win	0.40	0.60

Figure 2. Snapshot of the predicted outcome of the match between South Africa and Sri Lanka using Random forest in T-20 Fore-Teller Dashboard



T-20 Fore-Teller

Welcome to Fore- Teller

Team
South Africa

Opposition
Sri Lanka

Toss
Win

Venue
A

Day_Night
Night

partscore
74

Score
163

Predict

Predicted out come

```
[1] "Predicting for SVM"
```

data
Lost

Figure 3. Snapshot of the predicted outcome of the match between South Africa and Sri Lanka using SVM in T-20 Fore-Teller Dashboard

Table 2. : Predicted and Actual outcome of matches between South Africa (SA) vs Sri Lanka (SL) and New Zealand (NZ) vs India (IND)

Date	Team	Oppos. Team	Naïve Bayes			Random Forest			SVM Probabilty Prediciton.	SVM Pred Outcome.	Actual Outcome
			Win Prob	Lose Prob	PredOutcome	Win Prob	Lose Prob	PredOutcome.			
10-09-2021	SA	SL	0.48	0.52	Lost	0.60	0.40	Win	Lost	Lost	Win
12-09-2021	SL	SA	0.02	0.98	Lost	0.09	0.91	Lost	Lost	Lost	Lost
14-09-2021	SL	SA	0.05	0.95	Lost	0.06	0.94	Lost	Lost	Lost	Lost
17-11-2021	NZ	IND	0.45	0.55	Lost	0.39	0.61	Lost	Lost	Lost	Lost
19-11-2021	NZ	IND	0.34	0.66	Lost	0.08	0.92	Lost	Lost	Lost	Lost
21-11-2021	IND	NZ	0.21	0.79	Lost	0.63	0.37	Win	Win	Win	Win

The above table represents the successful outcome predicted by the classifiers. We can infer that Random Forest prowess in predicting the outcome when the variables are categorical and continuous. We conclude that Random Forest is the best fit in such a scenario.

6. Conclusion

We conclude that the Random Forest classifier is a suitable classifier when variables are continuous and categorical. We measured the accuracy as well as the Kappa value for the fourteen countries. From Table 1, we could conclude that mean values of the accuracy of Naïve Bayes, Random Forest and SVM are 0.6167, 0.5716, and 0.5973 respectively and the mean value of Kappa statistics are, 0.22083, 0.1550, and 0.1228. Among the three classifiers, Random Forest performs the best, although the variance between the mean Kappa values is very small. The algorithm exhibits the best performance. The below table shows the mean of all the classifiers:

Table 3. Mean value of accuracy for Naïve Bayes, Random Forest, and SVM.

Model	**Accuracy**
Naïve Bayes	0.6167
Random Forest	0.5716
SVM	0.5973

The maximum level of accuracy of 91% can be seen in the case of Random Forest and SVM for Afghanistan. The maximum level of disagreement on the prediction occurs in the case of Scotland for the Naïve Bayes algorithm.

References

[1] Awan, M. J., Gilani, S. A. H., Ramzan, H., Nobanee, H., Yasin, A., Zain, A. M., & Javed, R. (2021). Cricket match analytics using the big data approach. *Electronics*, 10(19), 2350.

[2] Goel, R., Davis, J., Bhatia, A., Malhotra, P., Bhardwaj, H., Hooda, V., & Goel, A. (2021). Dynamic cricket match outcome prediction. *Journal of Sports Analytics*, 7(3), 185-196.

[3] Kamble, R. R. (2021). Cricket Score Prediction Using Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(1S), 23-28.

[4] Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2022). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*, 18(3/4), 256-266.

[5] Kumarapandiyan, G., & Keerthivarman, S. (2019). A Statistical Analysis for Predicting the Top Performing Players during the ODI Cricket World Cup 2019 using Principal Component Analysis. *International Journal of Statistics and Reliability Engineering*, 5(2), 69-76.

[6] Mittal, H., Rikhari, D., Kumar, J., & Singh, A. K. (2021). A study on Machine Learning Approaches for Player Performance and Match Results Prediction. *arXiv preprint arXiv:2108.10125.*

[7] Passi, K., & Pandey, N. (2018). Increased prediction accuracy in the game of cricket using machine learning. *arXiv preprint arXiv:1804.04226.*

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

[9] Raja, M. A. M., Manasa, V. V. L., Reddy, D. S. N., & Sundari, K. S. (2021). Applying Data Science for Cricket Predictions. *Annals of the Romanian Society for Cell Biology*, 1853-1863.

[10] Raju, U. N. (2021). Getting Useful Information from Cricket Data Set Using Data Analytics Techniques Odi Cricket Team Performance Analysis Using Data Mining Classification Techniques. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(9), 684-693.

[11] Singh, S., Aggarwal, Y., & Kundu, K. (2020). Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction using Machine Learning. *International Journal of Computer Applications*, 975, 8887.

[12] Shenoy, A. V., Singhvi, A., Racha, S., & Tunuguntla, S. (2022). Prediction of the outcome of a Twenty-20 Cricket Match: A Machine Learning Approach. *arXiv preprint arXiv:2209.06346.*

[13] Sinha, A. (2020). Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020.

[14] Sudhamathy, G., & Meenakshi, G. R. (2020). Prediction on IPL Data Using Machine Learning Techniques in R Package. *ICTACT Journal on Soft Computing*, 11(1), 2199-2204.

[15] Wickramasinghe, I. (2020a). Classification of All-Rounders in the Game of ODI Cricket: Machine Learning Approach. *Athens Journal of Sports*, 7, 1-13.

[16] Wickramasinghe, I. (2020b). Naive Bayes approach to predict the winner of an ODI cricket game. *Journal of Sports Analytics*, 6(2), 75-84.