Impact of Batsman on India's Performance:

S Surya*, I Nilavan, Sumegh Sadashiv GONUGADE, G Nithish KANNA and Vergin Raja SAROBIN M

Vellore Institute of Technology, Chennai, India

(* Corresponding author's e-mail: suryasunder72@gmail.com)

ABSTRACT:
"Batsman win matches" is a phrase that suits Indian Cricket Team in particular. This work requires a dataset which is not readily available. Hence, dataset has been synthesized by scraping from websites- ESPN cricinfo and Howstat. This formulated dataset named Crickipedia has been subjected to the following Machine Learning (ML) algorithms: Support Vector Machine (SVM), Logistic Regression, Decision Tree, Artificial Neural Network (ANN) to analyze the past performance of the batsman and their impact on India's victory. Decision Tree algorithm gave excellent results with an accuracy of 90.91% for this proposed methodology. This analysis can be used to measure the impact of any batsmen on Team India's performance. This work will be focusing on T-20 Internationals in particular.

KEYWORDS:
Sports Analytics, Dataset Formulation, Machine Learning Models, Data Visualization, Performance Evaluation

1. INTRODUCTION:

Cricket is a sport rich of numbers and stats. Indian Cricket has produced a huge number of game-changing batsmen over the time. These star batsmen have been very impactful to Indian Cricket Team for winning Bilateral trophies, Tri-series trophies, Asia cups, and finally the most enormous World Cups. Some of the star batsman Indian cricket have produced are Sachin Tendulkar, Rahul Dravid, Sourav Ganguly, Virender Sehwag, Yuvraj Singh, Mahendra Singh Dhoni, Virat Kohli, Rohit Sharma, Suresh Raina, etc. These players happen to score more than five thousand plus international runs, scored tough runs for India in crucial games, and have won individual awards like man of the match and man of the tournament. So, it can be said that India is a land of batsman as it has produced tremendous batsman time and time again.

The traditional method (ODI and Test formats) of recording performance which have been used for decades mainly takes only batting average and number of runs scored into account. In the T20 game which evolved in recent decades, this traditional method which uses the above features (average and runs scored by the batsman) will not be efficient to measure the performance of the batsman. Considering the flaws of this method, this work will be tilted towards the T20 format where Strike Rate carries higher significance and lesser average will still suffice to make an impact. To be precise, Batting Strike Rate of 130+ and average of around 30 is considered the benchmark in an impactful T20 innings.

Considering the attributes used to compute the impact in other papers, this work will also include certain important features required for a T20 innings that have not been included in other papers. Some of the aspects this work will include that improves the methods used in other works are: 1. Batting Strike Rate Benchmark (130+), 2. Batting Average Benchmark (around 30), 3. Balls Faced per match, 4. Opposition Team and 5. Venue and conditions of the match. This method can be applied to any batsman of Indian Team. It analyses the player's stats in and out, and determines their impact on the team.

Virat Kohli, a stalwart of Indian cricket has scored more than 23,000 international runs, a stat which only Sachin Tendulkar has achieved in the history of Indian cricket. Virat Kohli is also the first Indian cricketer to cross 10,000 T20 runs. This paper will analyze Virat Kohli's stats in particular and determine his impact on Team India.

2. LITERATURE SURVEY:

The literature survey comprises papers speaking about all aspects of cricket analysis. The topics included are predicting player's performance, predicting match outcome, helping in player selection for team and the comparative study of various machine learning approaches.

In [1] the authors used the ball-by-ball dataset of the entire IPL T20 to predict the outcome of IPL matches using machine learning. They used complex statistical formulas and visualization techniques to rank players. The prediction of IPL match outcomes is crucial nowadays for sponsors and traders. The outcome of an IPL match is dependent on the individual performance of each player in the team. The authors [2][3][4] used various machine learning algorithms and concluded that Random Forest Classifier has proven to be best in predicting the match outcome for a team and also for predicting the performance of the batsman as well as a bowler. In the approach of [5] previous year IPL matches were fed into machine learning-based prediction approaches like the Naïve Bayes network with the help of the KNIME tool. The paper [6] proposed a Deep Mayo Predictor model for predicting the outcomes of the matches in IPL 9 being played in April – May 2016. Out of 56 matches played in the league stage of the IPL IX tournament, the predictor model can correctly predict the outcomes of 39 matches. The outcome of research [7] stated that in the game of ODI apart from the player's performance other traditional factors like toss result, the benefit of home–field, chasing and duration of the game (day/day – night) also influence the outcome of the match. The Classification and Regression Tree (CART) approach gave interesting and novel interpretations of these predictors. The paper [8] found that Random Forest Classifier gives the best result. The researchers incorporated the relative strength between the two teams playing the match by estimating the individual participating player's potential. They applied numerous supervised learning algorithms to predict the winner of the match.

The research work in [9] proposed that the prediction of a player's performance just based on scores and wickets can be misleading sometimes. The influence of ground and weather conditions is inevitable. These factors should be considered by the team while selecting players for the team. In this regard, the use of a Weighted Random Forest Classifier with hyperparameter tuning predicted good accuracy compared to other models. Apart from the Random Forest Classifier techniques like Stacking also provided good results in predicting player's performance using the ball-by-ball data of the previous matches [10]. The researchers use various machine learning algorithms in the aspect of sports analysis. The authors [11] compared the performance of different machine learning algorithms in predicting the outcome of a cricket match or in analyzing the player's performance. They concluded that each algorithm is with it its merit and demerit. The metrics of the neural network seem to increase with the increase in dataset size. The researchers use different sources of datasets for their prediction purposes. The work of [12] uses player's career statistics and recent performances to predict the outcome of the world cup match. The K-Nearest Neighbor (KNN) algorithm yielded better results compared to other classifiers like Naïve Bayes, and Support Vector Machine (SVM). And in [13] researchers used longitudinal data of test cricket to predict the performance of batsmen. They applied the technique of the Hierarchical Linear Regression Model to infer results. The authors found that the handedness (left / right-hand batsman), and match location (HOME / AWAY) have a significant influence on a player's performance. They also mentioned the advantage batsman get in IPL matches by playing in different pitch conditions and against international bowlers. The selection of team players for one game or a tournament is very crucial as it decides the performance of the team. The paper [14] implemented neural network models to categorize players as bowlers, and batsmen, assess their performance and then finally decide whether to select them or not. They have used past performance data for analysis. The researchers in [15] took into account the possibility that players might have higher statistics if they had continuously faced relatively weaker players. To consider this factor also, they gave a score to each player which changes with the performance of the opposing player. The Adaboost classifier showed the best results among the other classifiers used. The authors have stated the limitation of the dataset was a huge hurdle.

The literature survey helped in selecting the right algorithms to be applied to the data. And to lay out the performance analysis framework to compare the different algorithms used. The inputs from the survey also guided in formulating the dataset according to the aim of the study.

3. PROPOSED WORK:

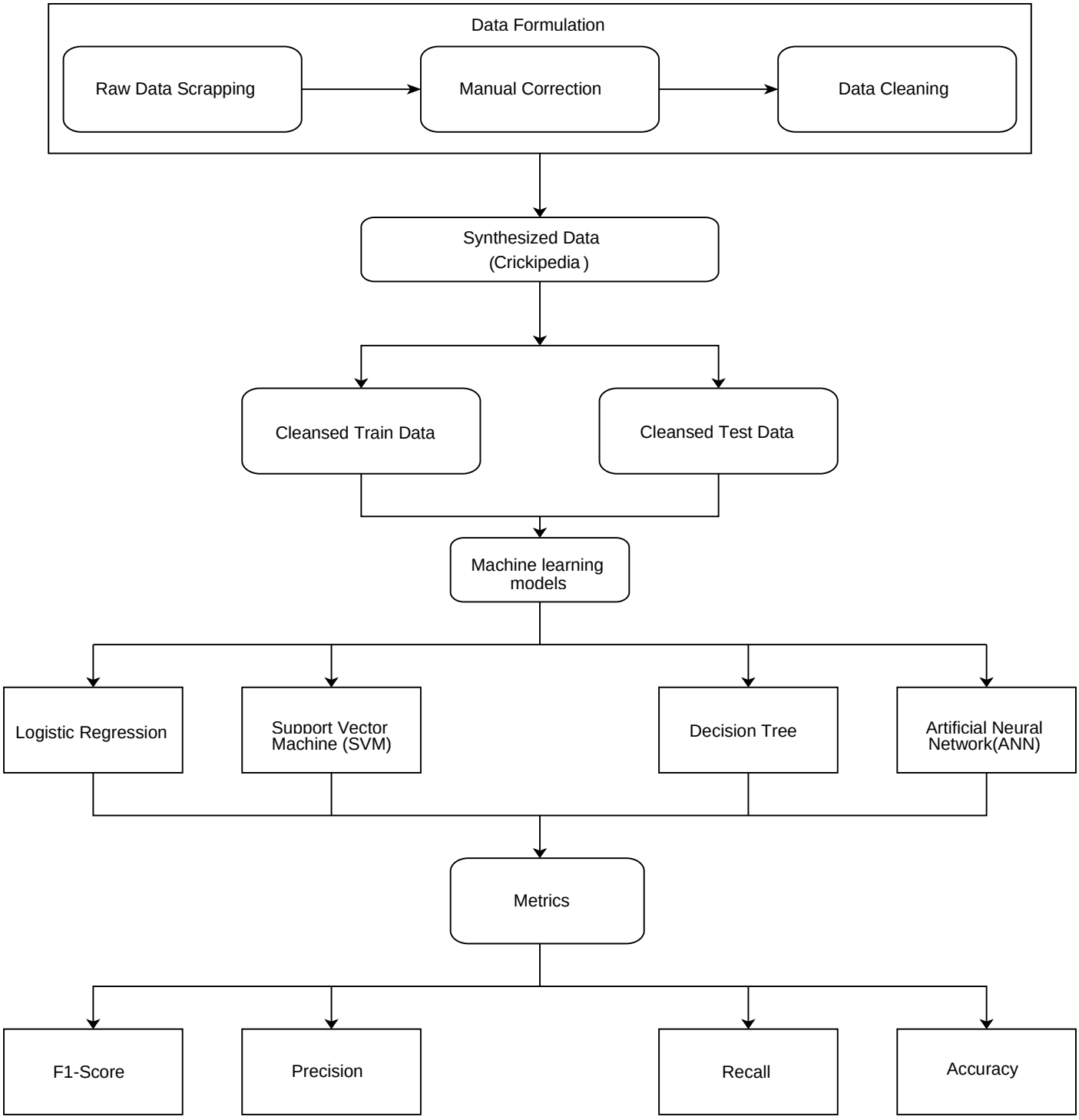

Figure 1 Proposed Framework for Player Performance Analytics

The proposed framework as mentioned in Figure 2 is implemented in this paper considers different statistical models to give a comparative analysis on the model accuracy. The Machine Learning algorithms implemented in this paper are- Support Vector Machine (SVM), Logistic regression, Decision tree, Artificial Neural Network (ANN). Further this work includes computation of metrics (F1-score, Accuracy, Precision, Recall) for each of the algorithms to analyze Virat Kohli's impact on India's performance and compare with other models. This paper also includes visualization with different graphs in order to understand trends in his dismissal and impact of his performance on India's Victory/Defeat.

3.1. DATA FORMULATION

Crickipedia dataset consisting of Virat Kohli's match by match stats, has been formulated by scraping raw data from ESPN cricinfo and howstat (the cricket statistician) and subjected to manual correction by removing data points where matches played vs single opposition<5. Further, Data cleaning is performed by introducing <NA> values to the matches where batsman did not bat (DNB) and Imputation of the data points according to the attribute type (Qualitative or Quantitative). Finally, cleansed data is split into training and testing sets to be supplied to the ML models.

Crickipedia dataset consists common attributes like Runs scored, Balls faced, Time spent on crease and Strike Rate. It also considers certain unique attributes like Home/Away, Opposition Country, Toss Factor and Venue. For the ML models, the dataset can be split into:

- Input Features- Runs, Balls faced, Time spent on crease (in minutes), Mode of dismissal, Bowler Type Dismissed, Opposition country, Ground and Home/Away
- Target Variable- Team India's Victory/Defeat.

3.2. MACHINE LEARNING METHODS

3.2.1. SVM

Support vector machines, or SVMs, are one of the most popular supervised learning algorithms used for both classification and regression problems. However, it is mainly used for machine learning classification problems. Dismissal is the y label and the classification (out / not- out) of the specified x features. The goal of the SVM algorithm is to create optimal lines or decision boundaries that can divide n-dimensional space into classes so that new data points can be easily placed in the correct category in the future. This optimal decision boundary is called a hyperplane. SVM selects extrema/vectors to help create hyperplanes. These extreme cases are called support vectors, and the algorithm is called a support vector machine**.**

Figure 2 Linearly Separated SVM region

Figure 2 shows plot of Support Vector Machine with linear boundary considering Runs and Balls Faced as input features.

3.2.2. Decision tree

Decision trees are a supervised learning technique that can be used for both classification and regression problems, but they are mostly suitable for solving classification problems. It is a tree-structured classifier, with internal nodes representing characteristics of the dataset, branches representing decision rules, and each leaf node representing a result. In this dataset Decision Tree is used as classifier. Match Result is the y-label and for the given x- features does the classification (Win / Loss).

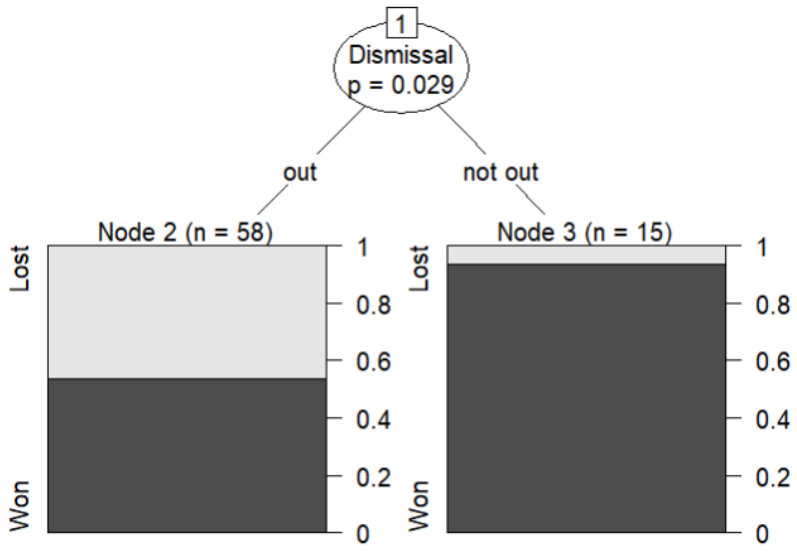

Figure 3 Decision tree for Dismissal v/s India's Win/Loss

Figure 3 is a decision tree drawn using ctree() method in R. It clearly shows India's probability of winning, when Virat Kohli gets out is around 0.5, but when Virat Kohli stays not-out the probability is as high as 0.9. This clearly concludes Virat Kohli is an impact player for team India.

3.2.3. Neural network

A neural network is a set of algorithms that attempt to discover underlying relationships in a set of data through a process that mimics how the human brain works. In this sense, neural networks refer to systems of neurons that are organic or artificial in nature. Neural networks can adapt to changing inputs. This way the network produces the best possible results without redesigning the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is rapidly gaining popularity in the development of trading systems.

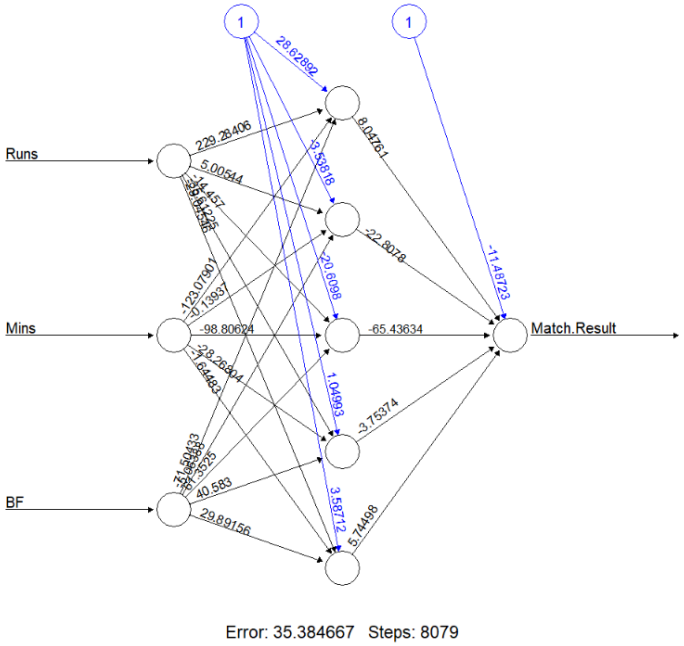

Figure 4 Artificial Neural network designed using forward propagation

Figure 4 is drawn using neuralnet() function in R. It takes Runs, Mins and Balls Faced in input layer (three neurons in the input layer), 5 neurons in the hidden layer and the output layer predicts Win/ Loss for team India.

3.2.4. Logistic regression:

Logistic regression is one of the most popular machine learning algorithms that falls under supervised learning techniques. It is used to predict a categorical dependent variable using a given set of independent variables. Logistic Regression predicts outputs for categorical dependent variables. Therefore, the results must be categorical or discrete. Can be yes or no, 0 or 1, true or false, and so on. But instead of giving exact values as 0 and 1, it gives probability values between 0 and 1.

Logistic Regression is very similar to Linear Regression except in how it is used. Linear regression is used to solve regression problems and logistic regression is used to solve classification problems.

Logistic Regression, instead of fitting a regression line, fits an "S" shaped logistic function that predicts the two maximum values (0 or 1). Logistic Regression takes the sigmoid function(refer Eq. (1)) as its activation function and computes the probability for the respective input features. If f(x) < 0.5, it predicts 0 else it predicts 1.

$$f(x) = \frac{1}{1+ e^{-x}} \quad (1)$$

3.3. DATA VISUALIZATION

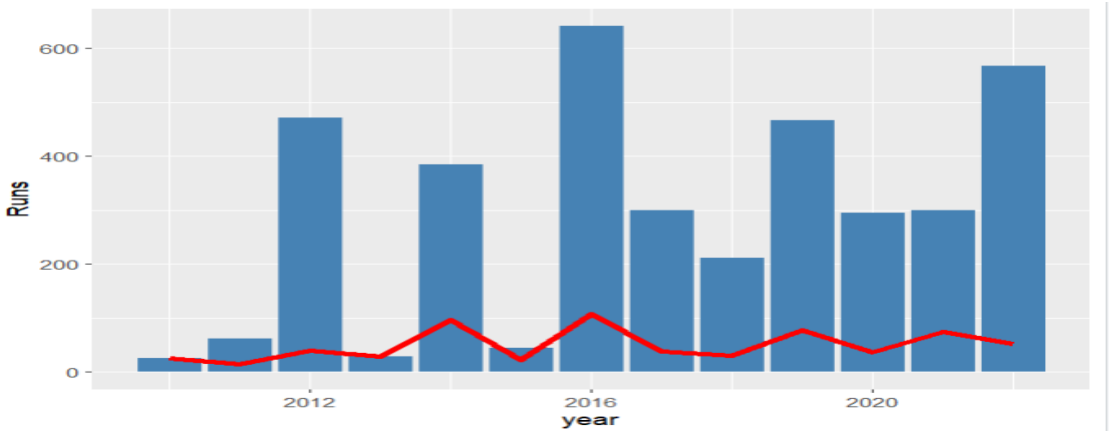

Figure 5 Kohli's Career Batting – T20 Internationals

Figure 5 depicts that Kohli scored more runs in the years 2016 and 2022 and scored less runs in the year 2010,2013 and 2015. Looking closely at the line graph it can be observed that his average rises in the years 2014, 2016, 2019 and 2021.

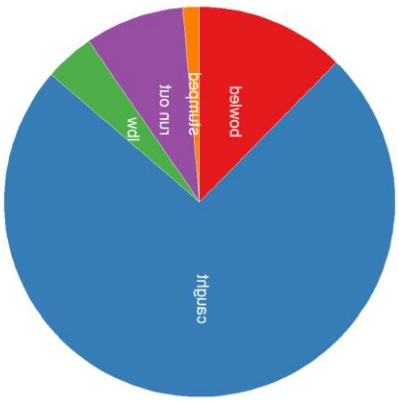

Figure 6 Pie-Chart of Kohli's dismissal mode

Figure 6 visualizes that Kohli is more prone to getting out "caught"(This mode of dismissal has become a problem for him of late as he gets out caught behind edging balls outside off stump and also getting strangled down the leg stump). He usually doesn't get "stumped" by stepping out to the spinner (as he is a good player of spin). Since he

is quick between the wickets, he doesn't get run out often.

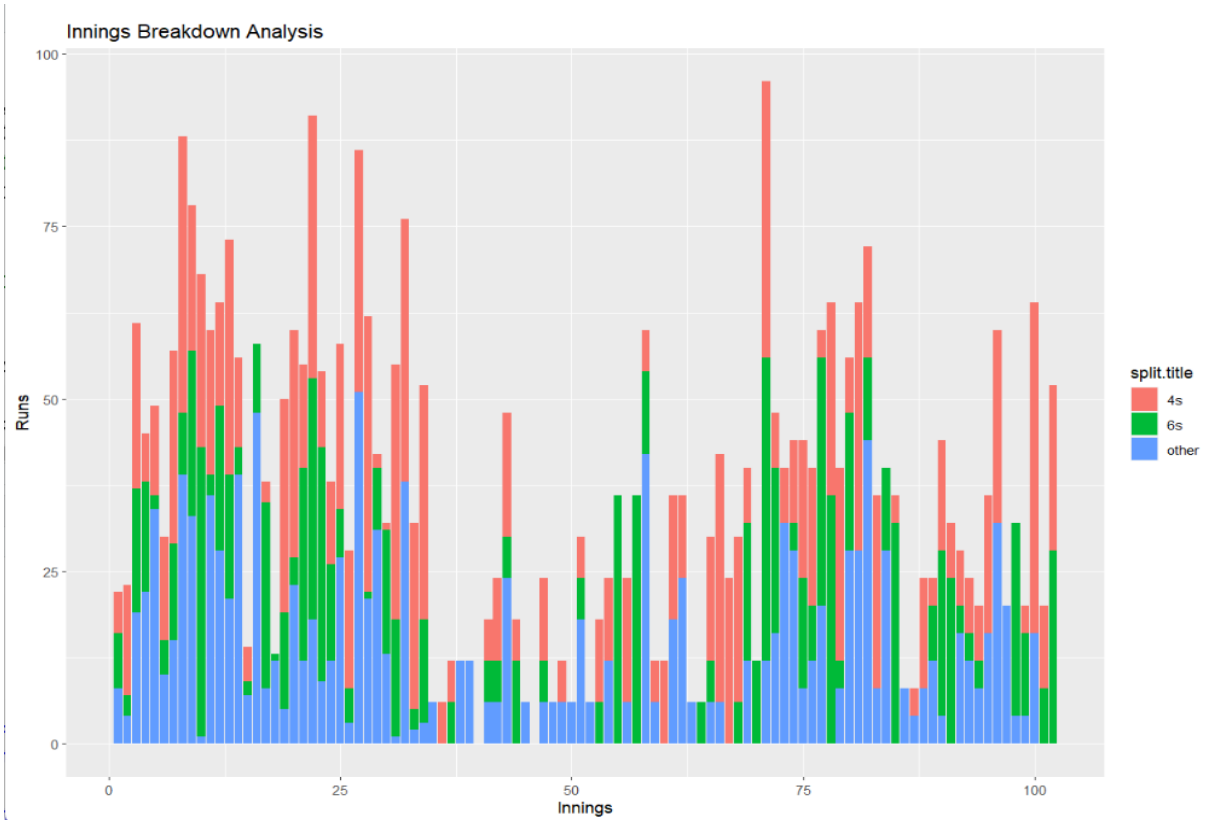

Figure 7 Innings Breakdown Analysis

On visualizing **Figure 7,** it can be concluded that Kohli's running between the wickets (blue stack) has been fluent throughout his career. In the innings 25-80, he has been in good form and his boundary percentage (4s-red stack and 6s- green stack) has been high. Recently, his boundary percentage is low and also his strike rate in T20 internationals has come down (major concern for team India).

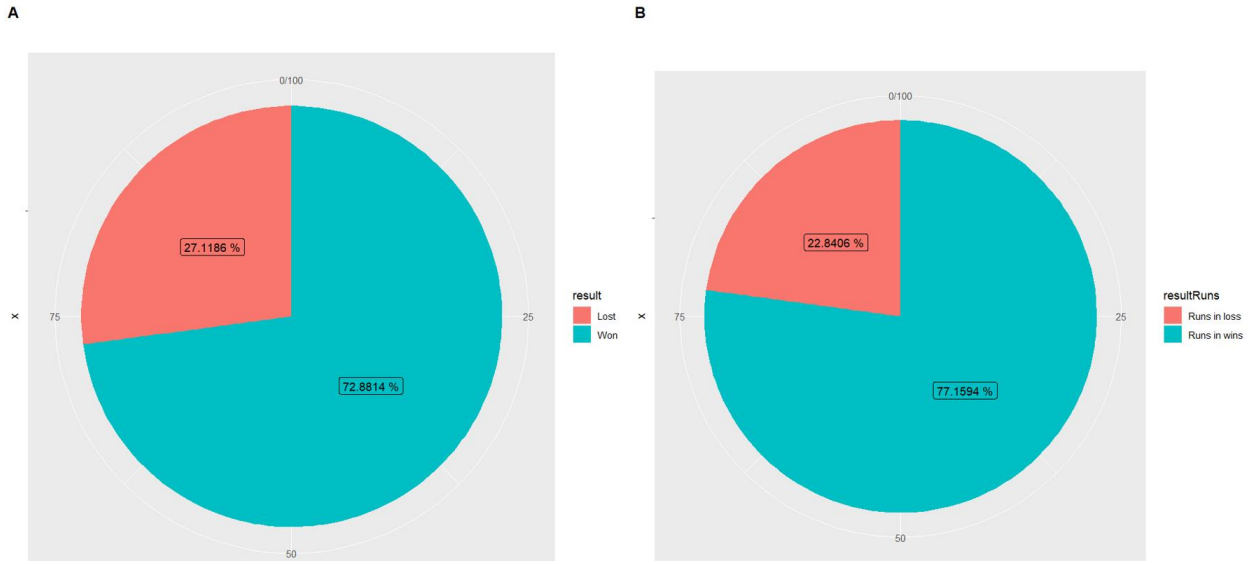

Figure 8 Pie chart of Kohli's contribution in India's Away games

In **Figure 8**, Graph – A depicts India's away wins as percentage:72.8%. Graph – B depicts Kohli's runs in wins and losses. Evidently, his total runs as percentage in wins is 77%. Therefore, it can be concluded that he plays a huge factor in India's away wins.

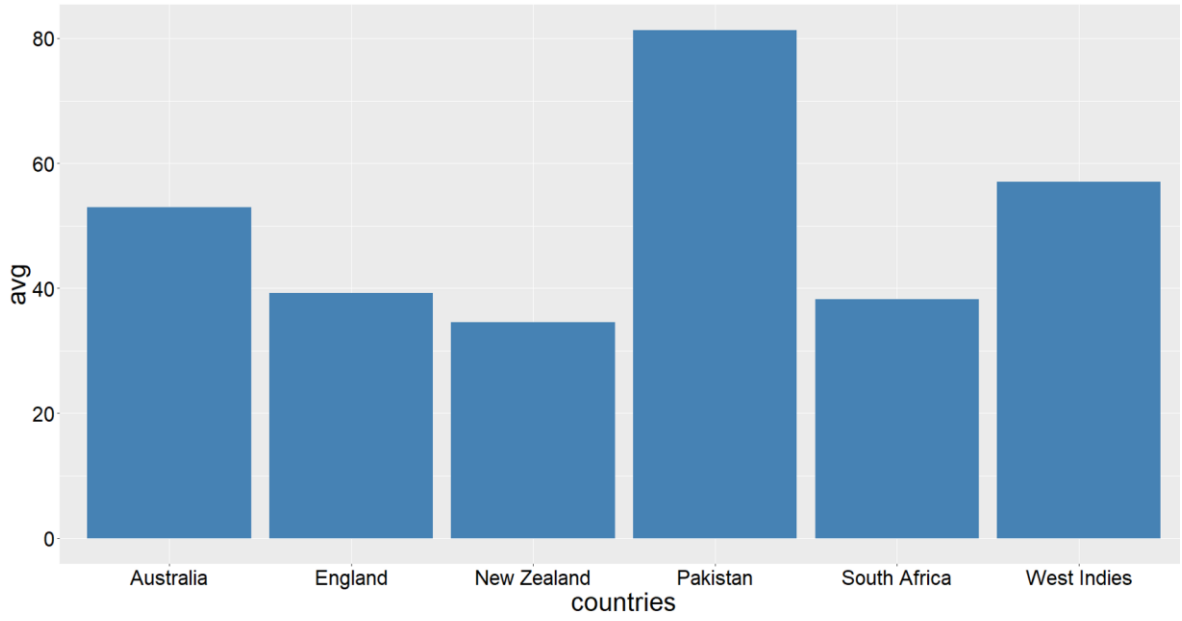

Figure 9 Kohli's average in T20 Internationals against top 6 cricketing nations

Figure 9 is a bar graph that compares his average against top 6 cricketing nations- Pakistan, West Indies, Australia, South Africa, England and New Zealand. Kohli's average against Pakistan is tremendous and hence it can be concluded that he certainly dominates Pakistan. Against the SENA countries (South Africa, England, New Zealand, Australia), his record against Australia is good and poor against New Zealand.

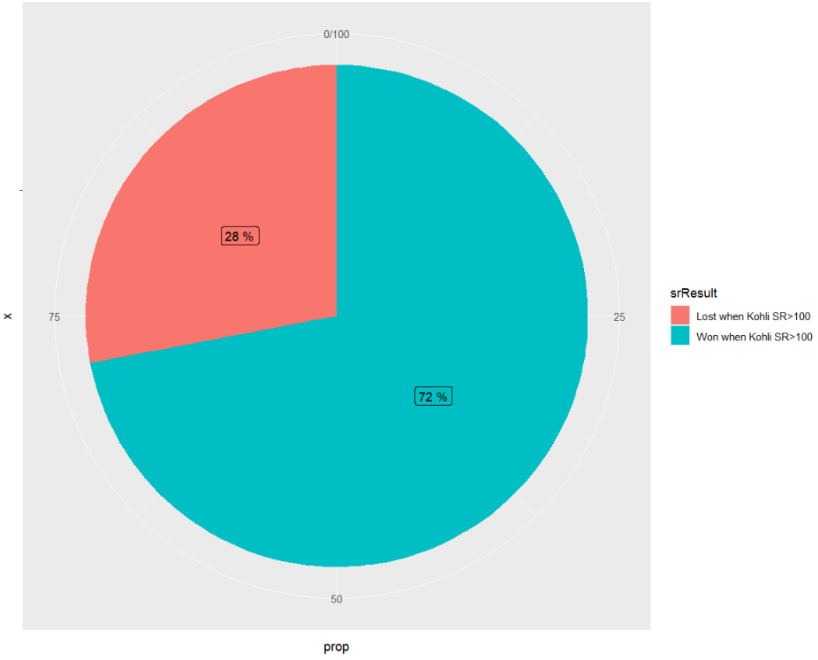

Figure 10 India's win% when Kohli's SR>100

Figure 10 depicts a Pie Chart on India's number of win and losses for those matches with Kohli's Strike Rate is greater than 100. Clearly, when his strike rate is greater than 100 (i.e., when he is batting very well), India's win percentage is 72.

4. RESULTS AND DISCUSSION

4.1. Precision and Recall

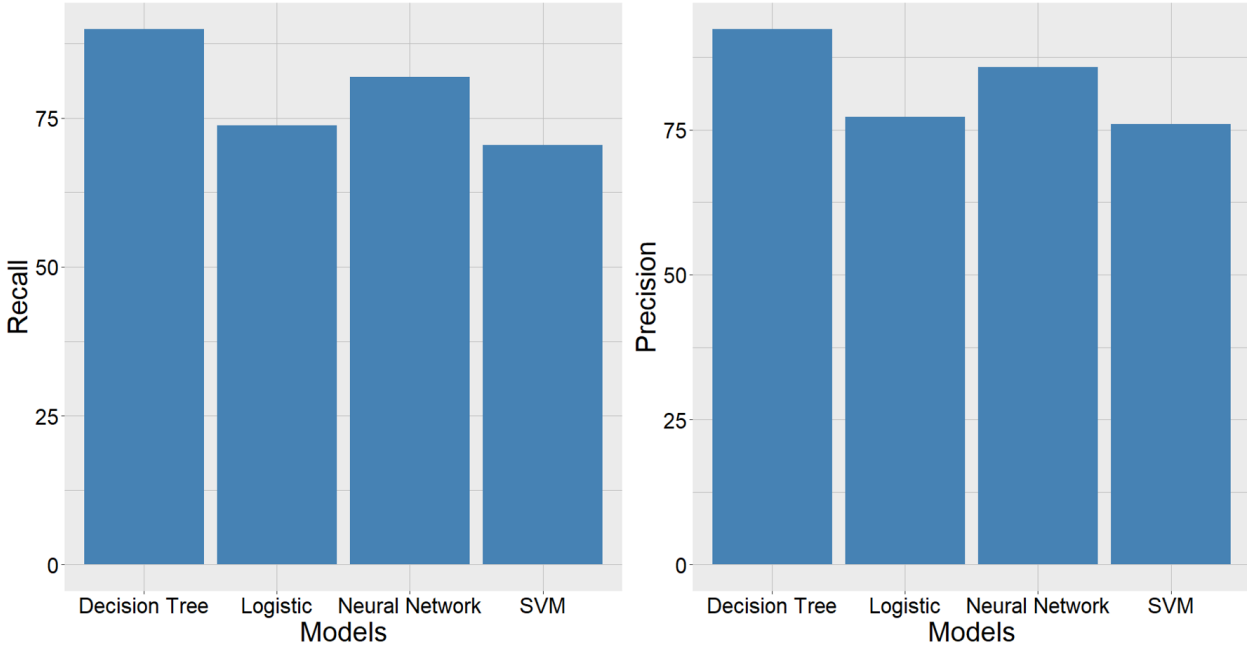

Figure 11 Precision and Recall score of various ML models used

Precision tells about how relevant the ML model classifies the data points. From the Figure 11, Decision Tree classifier model gives the best precision score. Recall tells about how many correct positive predictions made by the model to the total number of positive predictions. In this work, Decision Tree gives the best recall value followed by Neural Network. SVM and Logistic regression gives similar recall values.

4.2. F1 Score and Accuracy

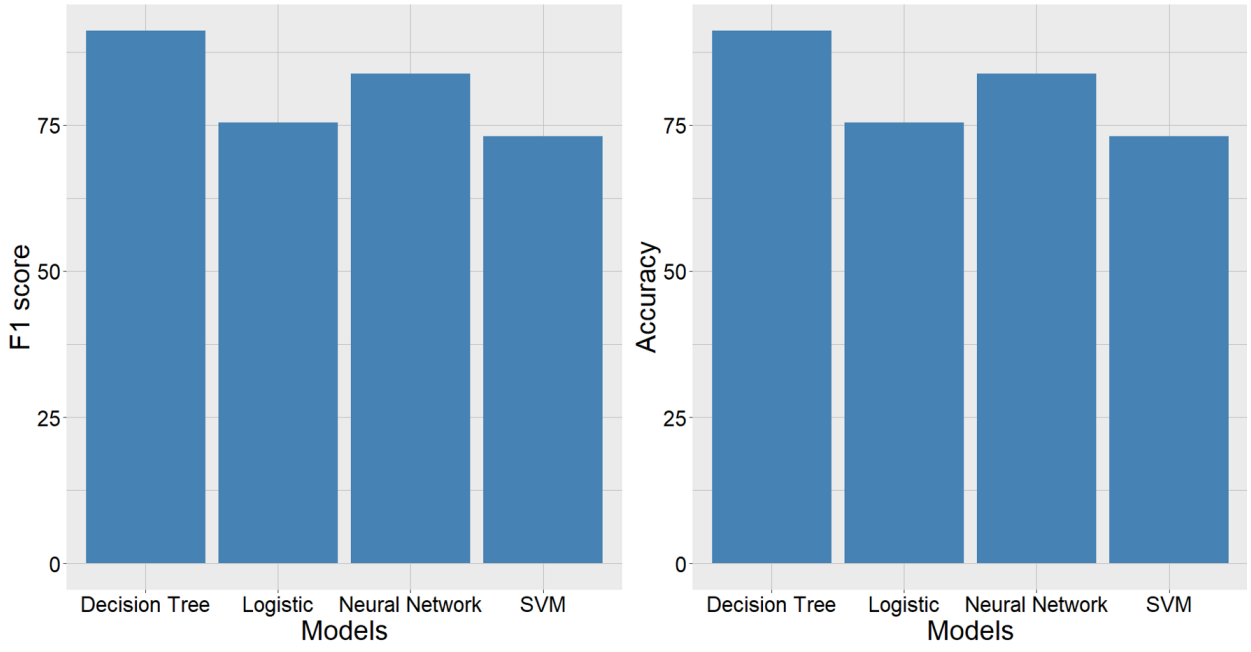

Figure 12 F1 Score and Accuracy value of various ML models used

F1 Score is the harmonic mean of Precision and Recall values. It specifies the number of correct predictions across the dataset made by the model. Accuracy denotes the number of correctly classified data points over the total number

of data points. Accuracy is the most important parameter to be considered while comparing the performance of different ML models. **Figure 12** concludes that Decision Tree is the best suitable data model for this dataset, which gives an accuracy of 90.91% and the highest F1 Score among the other models.

Table 1 Machine Learning models with its metrics

Algorithms	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Logistic	77.23	73.72	75.43	68.33
Decision Tree	92.34	89.89	91.09	90.91
Neural Network	85.82	81.93	83.82	78.12
SVM	75.98	70.43	73.09	73.54

Table 1 includes the combined scores of all the computed metrics for each of the ML algorithms. The results interpret that Decision Tree algorithm gives the best values for all the metrics and is the most suitable model for this work.

5. CONCLUSION AND FUTURE WORK

This work is an analysis on performance of Virat Kohli, one of the star of Indian cricket in T20 Internationals. Finally, the results of this work indicates that he has played a major part for Team India in T20 Internationals from 2014 – present day continuing his stellar form. From the four used ML models, Decision Tree and Neural Network have yielded good results for our dataset. Logistic Regression and SVM models haven't been up to the mark in this work. For getting a better understanding of the dataset, various parameters have been visualized related to Virat Kohli's impact on Team India in the form of Bar chart, Stacked Bar chart, Line chart and Pie chart. This methodology can be used for any batsman to find their impact on the team.

Further, this work can be extended to other formats namely One-day Internationals, Test matches and franchise cricket like IPL, BBL, SA-T20, etc. Also, it can be used for any International/ club team to view it's performance based on their impact players. Various advanced visualization techniques can be implemented using software like Tableau for better understanding of the dataset and results. Deep Learning models can also be used for improved performance over the traditional ML models increasing the interpretability.

\

6. REFERENCES

[1] VS, A.K., Mishra, A.S. and Valarmathi, B., 2020. Comprehensive Data Analysis and Prediction on IPL using Machine Learning Algorithms, International Journal on Emerging Technologies 11(3): 218-228.

[2] Srikantaiah, K.C., Khetan, A., Kumar, B., Tolani, D. and Patel, H., 2021, September. Prediction of IPL Match Outcome Using Machine Learning Techniques. In 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021) (pp. 399-406). Atlantis Press.

[3] Kapadia, K., Abdel-Jaber, H., Thabtah, F. and Hadi, W. (2022), "Sport analytics for cricket game results using machine learning: An experimental study", Applied Computing and Informatics, Vol. 18 No. 3/4, pp. 256-266.

[4] Passi, K. and Pandey, N., 2018. Increased prediction accuracy in the game of cricket using machine learning, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, arXiv:1804.04226.

[5] Sinha, A., 2020. Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020. *Preprints* 2020, 2020100436 (doi: 10.20944/preprints202010.0436.v1)

[6] Prakash, C.D., Patvardhan, C. and Lakshmi, C.V., 2016. Data analytics based deep mayo predictor for IPL-9. International Journal of Computer Applications, 152(6), pp.6-10.

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

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

[9] Kapadiya, C., Shah, A., Adhvaryu, K. and Barot, P., 2020. Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique. Int. J. Eng. Adv. Technol, 9(3), pp.3406-3409.

[10] Patel, N. and Pandya, M., 2019. IPL Player's Performance Prediction. International Journal of Computer Sciences and Engineering, 7, pp.478-481.

[11] Mittal, H., Rikhari, D., Kumar, J. and Singh, A.K., 2021. A study on machine learning approaches for player performance and match results prediction. arXiv preprint arXiv:2108.10125.

[12] Bhatia, P., Rane, M.A. and Katraj, P., 2020. ICC T20 Cricket World Cup Prediction Based Data Analytics and Data Mining Technique. JETIR May 2020, Volume 7, Issue 5.

[13] Wickramasinghe, I.P., 2014. Predicting the performance of batsmen in test cricket. Journal of Human Sport and Exercise, 9(4), pp.744-751.

[14] Iyer, S.R. and Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), pp.5510-5522.

[15] Shenoy, A.V., Singhvi, A., Racha, S. and Tunuguntla, S., 2022. Prediction of the outcome of a Twenty-20 Cricket Match. arXiv preprint arXiv:2209.06346.