DATA VISUALIZATION
FINAL REPORT
IMPACT OF BATSMAN ON TEAM PERFORMANCE

NAME : SURYA S
REG.NO : 20BCE1071
NAME : MELVIN DAVID
REG.NO : 20BCE1463

NAME : NILAVAN I
REG.NO : 20BCE1080

SUBMITTED TO:
Dr. Joshan Athanesious J

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

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. Visualization for various parameters of this formulated dataset named Crickipedia will be performed using Tableau Software in integration with Python (Tabpy). Further, 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 Limited Overs Internationals (LOI) in particular.

KEYWORDS:
Visualization, Performance, Average, Dismissal, Algorithms

OBJECTIVE:
The objective of this work is to examine and evaluate the performance of individual batsmen on a cricket team using data visualization and statistical analysis. This project considers various performance metrics such as runs scored, strike rate, average, and other relevant statistics.

The aim is to identify the strengths and weaknesses of individual batsmen in the team, their contributions to the team's overall performance, and areas for improvement. This analysis can help the team management make informed decisions about player selection, batting order, and other strategic decisions.

In this project performance of individual batsmen is visualized and analyzed across different matches and tournaments, identifying trends in his performance, and predicting future performance based on historical data.

METHODS USED:

1. **Bar charts:** Bar charts are a simple and effective way to compare the performance of a batsmen in different matches across the years. This work includes plot of the runs scored by the player on the y-axis and the year on the x-axis. This compares the contribution of the player to the team's overall performance.

2. **Line charts:** Line charts can be used to show the trend of a batsman's performance over time. You can plot the runs scored by the batsman in each match on the y-axis and the match dates on the x-axis. This can help you identify patterns in the batsman's performance

3. **Bubble charts:** Bubble charts can be used to show the relationship between different variables, such as runs scored and strike rate. You can plot the runs scored on the x-

axis, the strike rate on the y-axis, and the size of the bubble can represent the number of matches played. This can help you identify players who score runs quickly but may not have played as many matches.

4. **Pie charts:** A pie chart is a circular graph that is divided into sectors, with each sector representing a proportion of the whole. The size of each sector is proportional to the quantity it represents. It could be an effective way to visually represent the impact of Virat Kohli on team India by highlighting the different aspects of his contribution and their relative importance.

OUTCOME:

The outcome of this project investigating the impact of batsmen on team performance are:

- **Identifying key performance metrics:** The project analyses metrics which are most relevant for measuring the impact of batsmen on team performance. This could involve analyzing data on batting averages, strike rates, runs scored, partnerships, and other indicators.

- **Quantifying the impact of individual batsmen:** The project could use statistical techniques such as regression analysis or machine learning algorithms to estimate the contribution of individual batsmen to team performance. This could help identify which players are most valuable to their teams and which factors are most strongly associated with success.

- **Examining contextual factors:** The project could also explore how contextual factors, such as the match format, pitch conditions, and opposition strength, affect the impact of batsmen on team performance. This could involve comparing performance across different conditions and identifying patterns of performance.

- **Informing team selection and strategy:** The findings of the project could be used to inform team selection and strategy, by identifying which batsmen are most effective in different situations and how teams can optimize their batting order and partnerships.

Overall, the outcome of the project would depend on the specific goals and research questions, but could potentially provide valuable insights into the role of batsmen in team performance and inform decision-making in cricket.

SCOPE:

Some potential scopes of this work are:

- Batting statistics: This would involve analyzing individual player statistics such as batting average, strike rate, number of boundaries, etc. to assess their overall batting performance.

- Impact on team score: This would involve analyzing the impact of individual batsmen on the team's overall score. For example, how much does the team score increase or decrease when a certain batsman is playing?

- Match outcomes: This would involve analyzing the impact of individual batsmen on the team's chances of winning matches. For example, does a team have a better chance of winning when a certain batsman is playing?

- Role in the team: This would involve analyzing the specific role that a batsman plays in the team, such as opening batsman, middle-order batsman, or finisher, and how their performance in that role affects the team's overall performance.

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 its performance based on their impact players. Deep Learning models can also be used for improved performance over the traditional ML models increasing the interpretability.

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.

The researchers in [1] present a novel approach to sports data visualization, which combines traditional visualization techniques with machine learning algorithms and big data analytics. The authors describe the challenges associated with sports data visualization and suggest that the hybrid approach can overcome these challenges by providing more accurate and timely insights. The article provides detailed descriptions of several case studies where the hybrid design approach was applied to different sports, which should be of interest to researchers and practitioners working in the field of sports analytics. The authors of [2] and [3] highlight the importance of data analysis and visualization in sports and gaming. The papers focus on the use of data visualization techniques to analyze the performance of Indian Premier League (IPL) teams and batsmen, highlighting the importance of the toss and providing insights into the factors that contribute to success. The papers also provide an overview of the methodologies used to measure expert performance in serious games, emphasizing the importance of data analytics in gaining insights into player behavior and decision-making.

In [4] 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 [5][6][7] 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 [8] 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 [9] 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 [10] 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 [12] 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 [13]. The researchers use various machine learning algorithms in the aspect of sports analysis. The authors [14] 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 [15] 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 [16] 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 [17] 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 [18] 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. MATERIALS AND METHODS:

3.1. Info about models

Machine Learning models used in this work are:

3.1.1. 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).

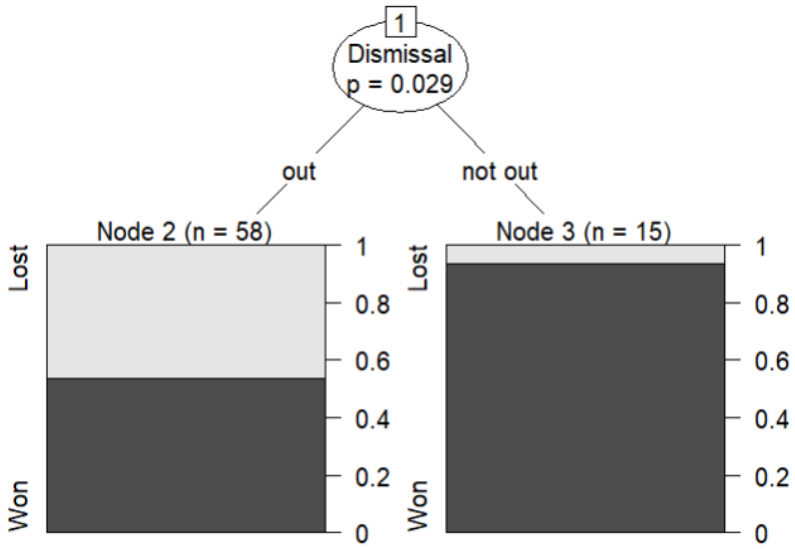


Figure 1 Decision tree for Dismissal v/s India's Win/Loss

Figure 1 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.

3.1.2. 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.

3.1.3. Artificial Neural Network (ANN)

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.

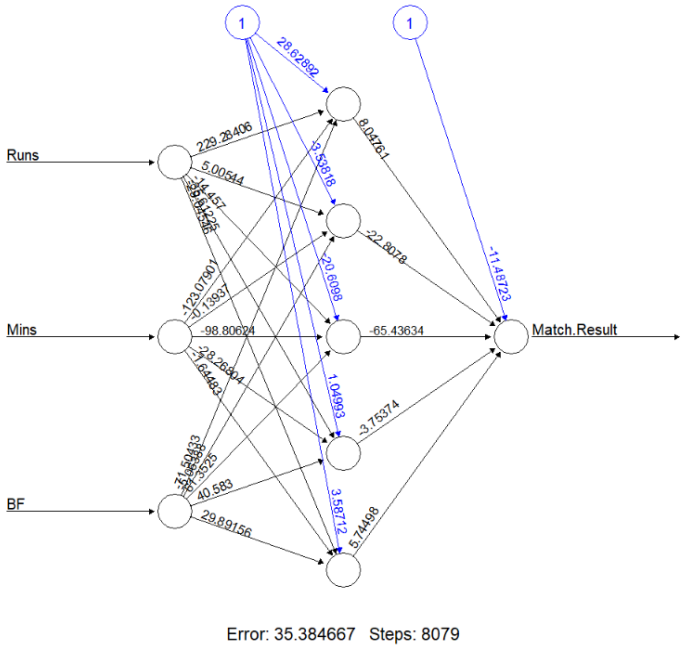


Figure 3 Artificial Neural network designed using forward propagation

Figure 3 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.

3.1.4. 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.

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

3.2. Dataset

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.

3.3. Architecture and explanation

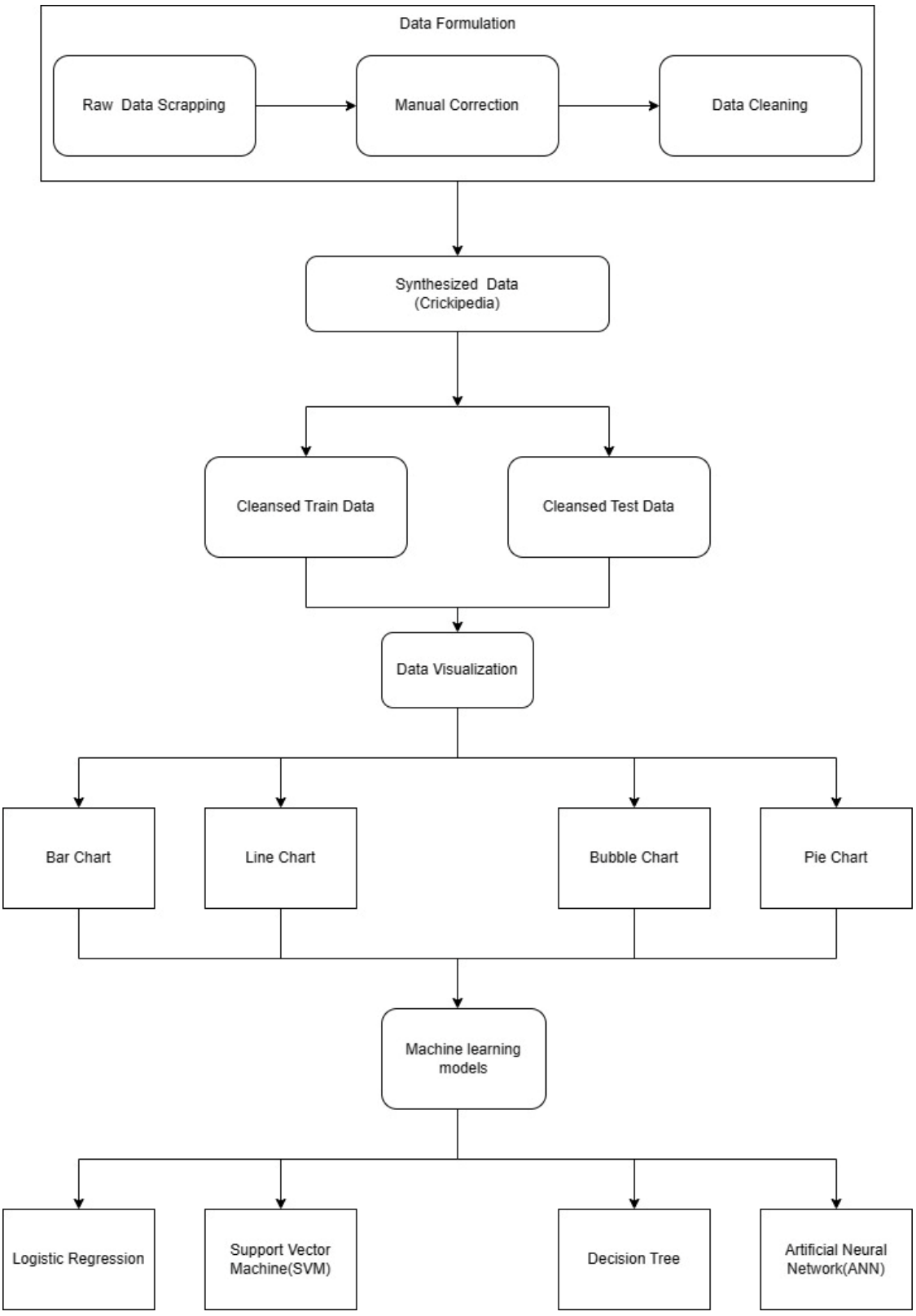


Figure 4 Proposed Methodology

- **Data Pre-processing:**
  Dataset for this work is not readily available. It is scarped from various websites followed by manual correction and data cleaning. This process is clearly described in the above subheading- DATASET.

- **Data Visualization:**
  Before proceeding to model building and deciding the Machine Leaning model to be selected, it is necessary to get a clear idea of the dataset and hence data is visualized clearly in this project using bar chart, stacked-bar chart, line chart, pie chart and bubble chart. All the visualizations along with explanation is explained in the results section clearly.

- **Model Selection and Building:**
  After visualizing the data, Classification models such as Logistic Regression, SVM Classifier, Decision Tree and Artificial Neural Network are applied on this dataset for various features. Further, each model is compared using appropriate performance metrics such as Precision, Recall and F1-score.

4. PROPOSED WORKS:

4.1. Novelty

In cricket, the performance of the batsmen is crucial for a team's success. The ability to score runs and build partnerships is essential for setting a challenging target or chasing down a total. However, the impact of individual batsmen on team performance is not always straightforward. Factors such as pitch conditions, opposition quality, and the role of the batsman within the team can all play a role in determining their influence on the outcome of a match. This project aims to explore the relationship between batsmen and team performance by analyzing data from a range of matches and examining the key factors that contribute to a batsman's impact on the game. By understanding the impact of batsmen on team success, we can gain insights into how teams can optimize their strategies and improve their chances of winning matches.

4.2. Project contributions

Surya S

- ODI Analysis and Visualization
- Decision Tree
- SVM

Nilavan I

- T20 Analysis and Visualization
- Neural Network

Melvin David

- T20 Analysis and Visualization
- Logistic Regression

5. RESULTS AND DISCUSSION:

5.1. Results

5.1.1. Precision and Recall

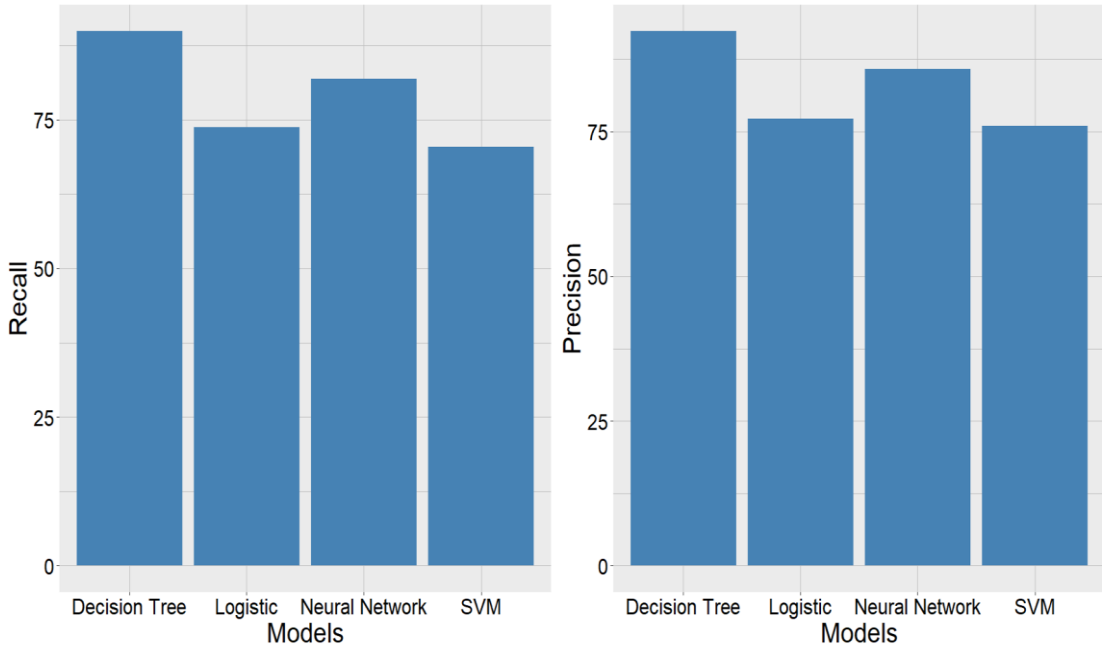


Figure 5 Precision and Recall score of various ML models used

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.

5.1.2. F1-score and Accuracy

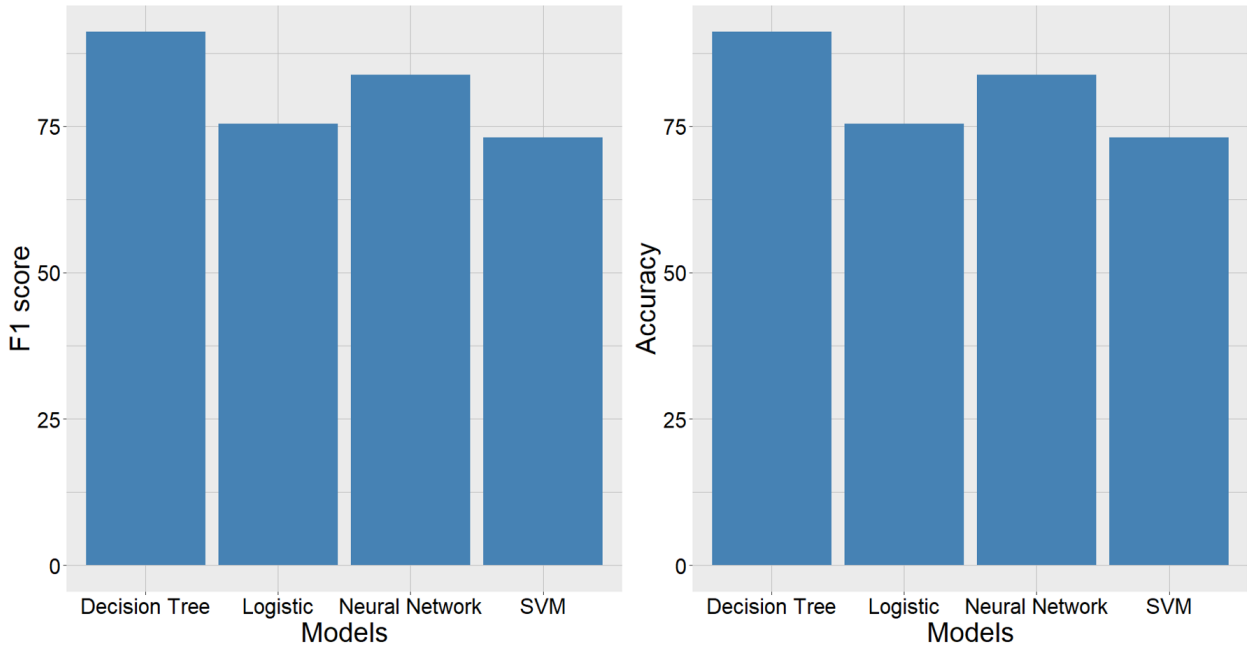


Figure 6 F1 Score and Accuracy value of various ML models used

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.

5.2. Figures and Explanation

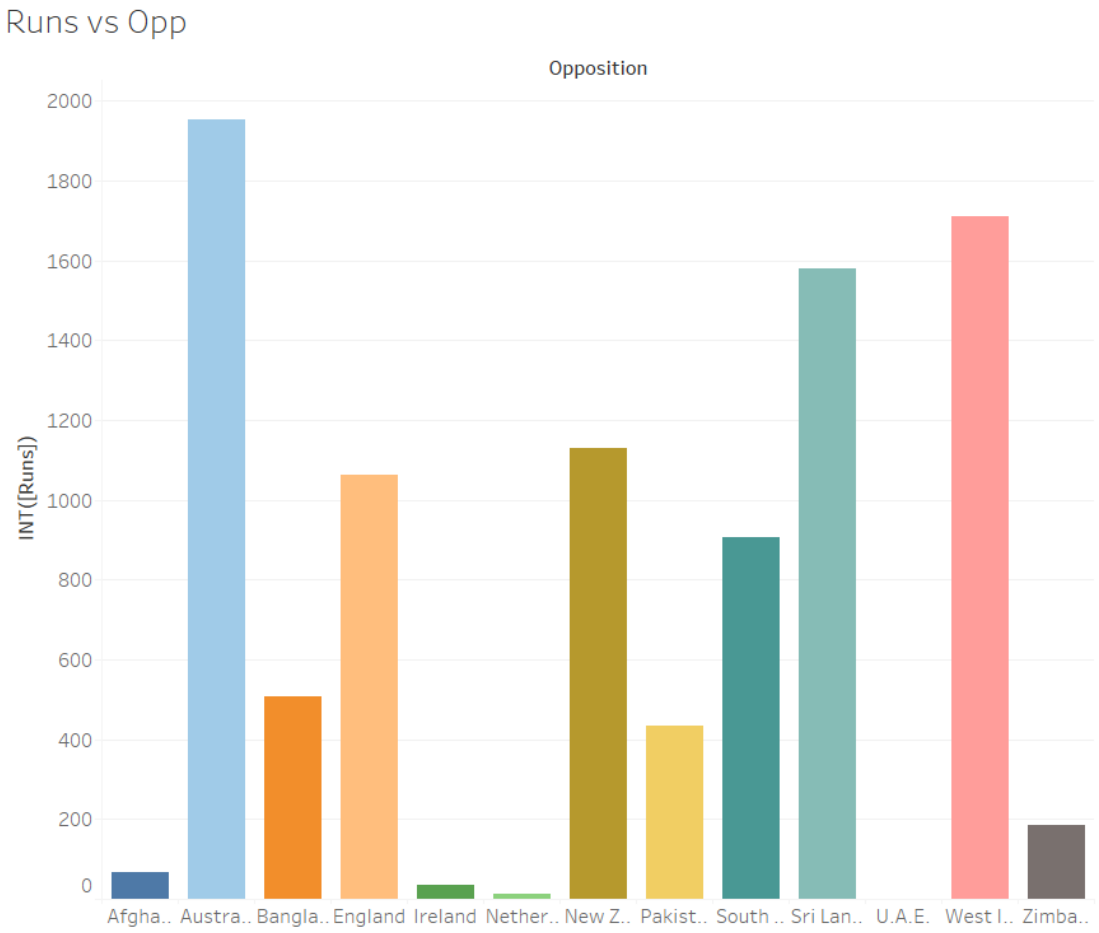


Figure 7 Runs vs Opposition

As we know, Virat Kohli has been a nemesis to Australia and has scored the highest runs in his career against Australia as visualized in **Figure 7.** Following Australia, he has scored many runs against Sri Lanka and West Indies.
So, it can be said that Virat Kohli is a major impact player for Team India when playing against Australia, Sri Lanka and West Indies.

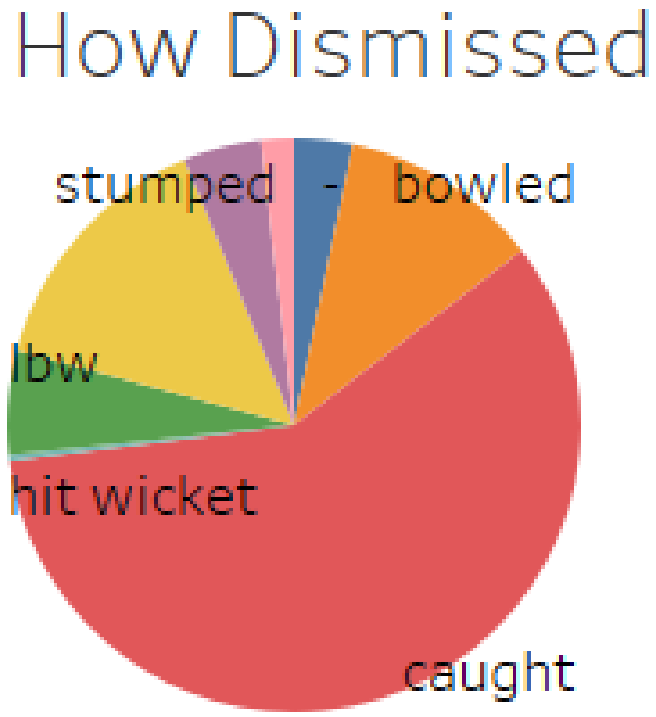


Figure 8 How dismissed

As most of the bowlers from different countries know that Virat Kohli edges ball outside off and the pie chart in **Figure 8** showing that "caught" has been the major form of dismissal. He has rarely been run out indicating that he is very fit cricketer and runs well between the wickets.

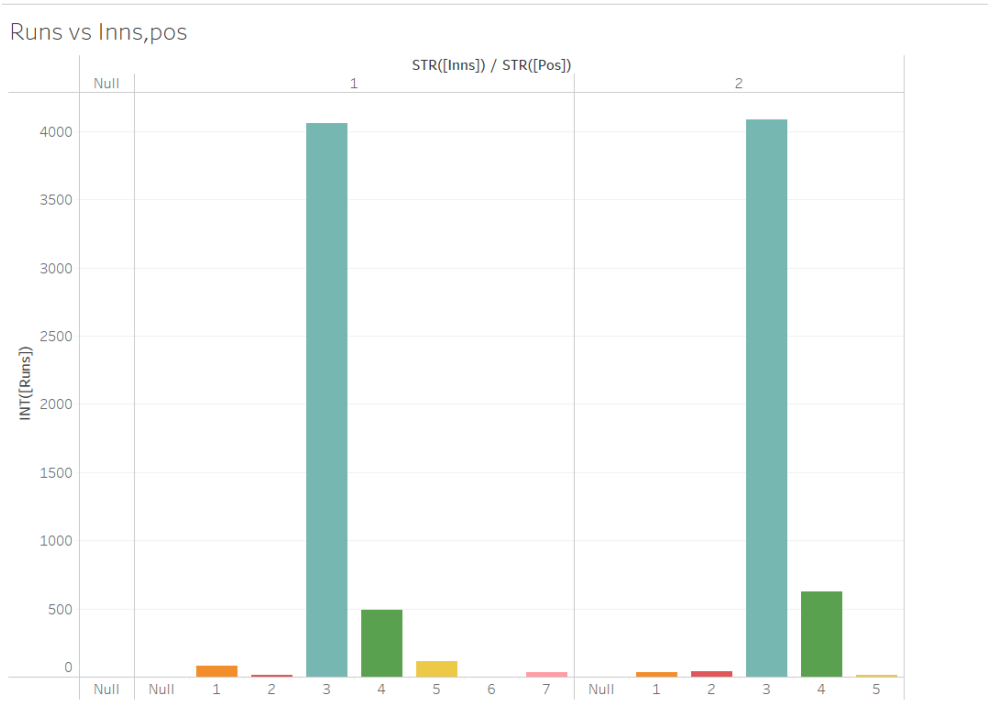


Figure 9 Runs vs Inns pos

From the graph in **Figure 9,** Virat Kohli has scored most runs in his career at number 3 position and owns the position as no other player in team India has scored as much as him. It can also be seen that he has a lot of runs in either of the Innings at number-3 position.

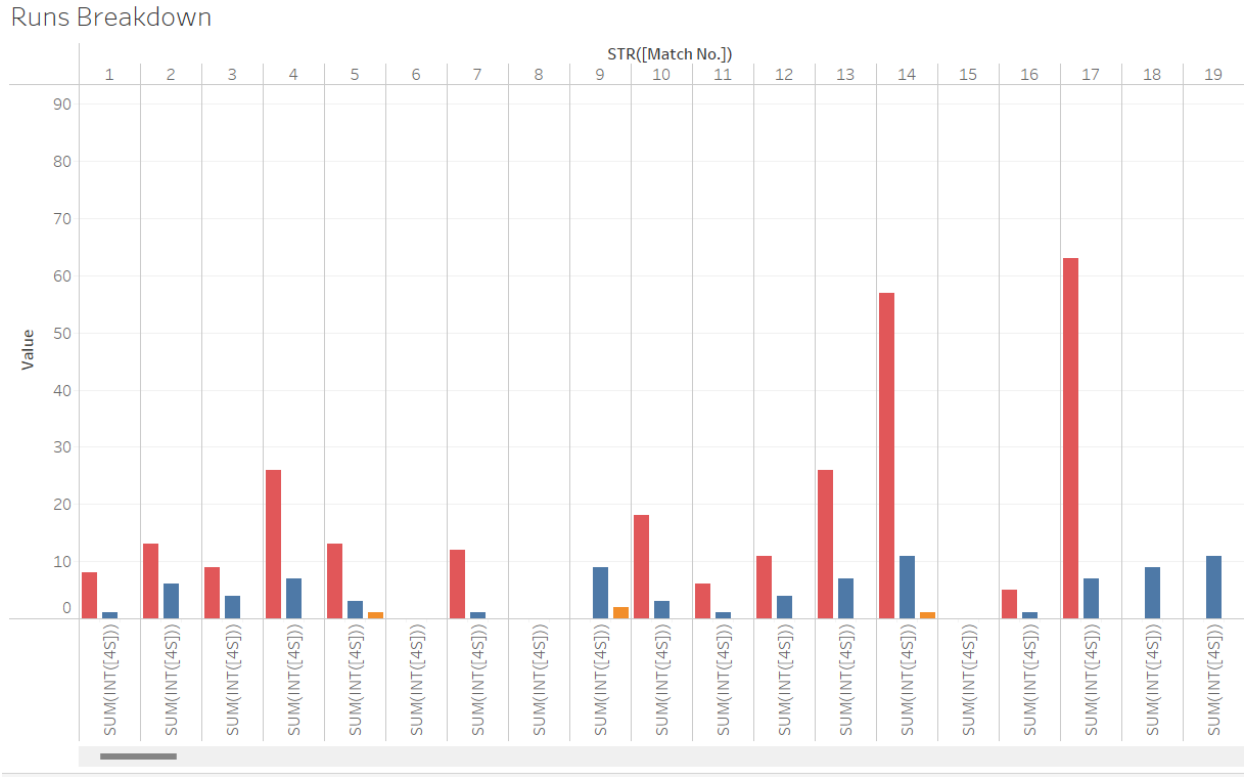


Figure 10 Runs Breakdown

Figure 10 shows that running between the wickets (Red bar) has been Virat Kohli's main form of runs hence it can be concluded that he is a fit cricketer. The sixers (Orange bar) have been very limited in this format. But the fours (Blue bar) have come as and when required around 8-10 each innings

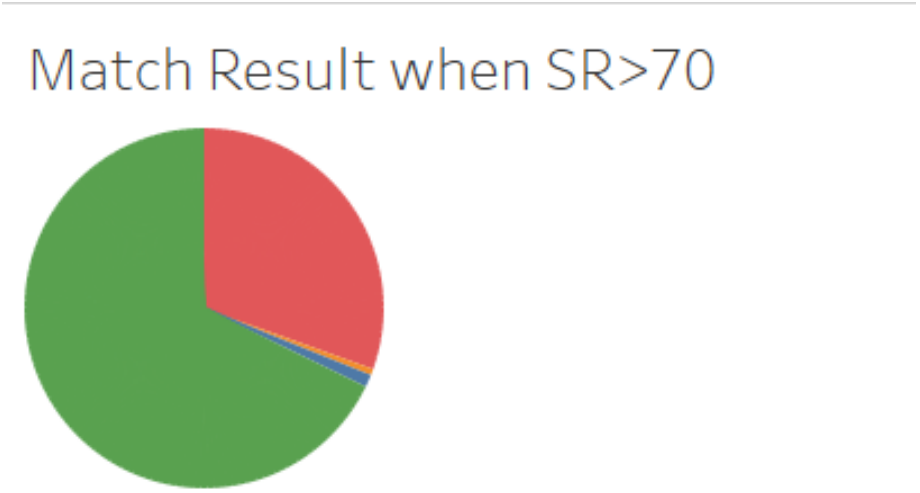


Figure 11 Match Result when SR>70

Virat Kohli always looks for needs of the team and the pie chart from **Figure 11** shows that India wins more than 70% of matches when Virat scores at a Strike rate of greater than 70.

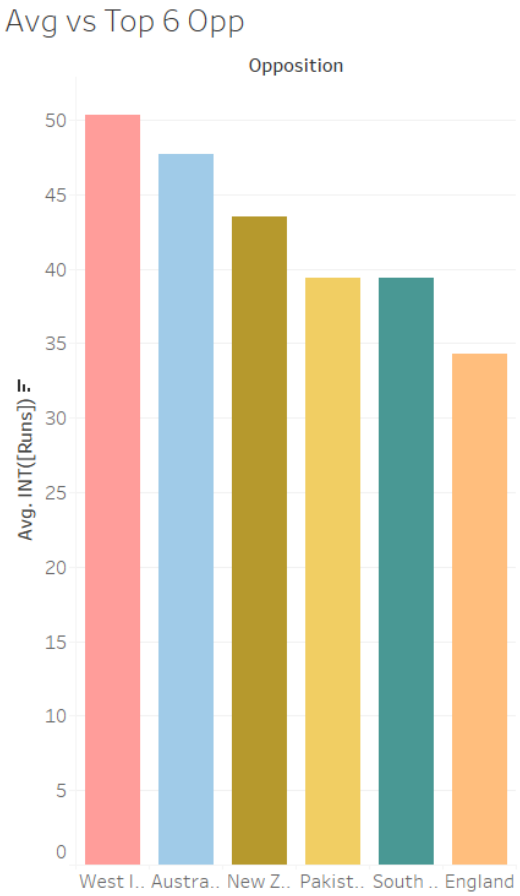


Figure 12 Average vs Top 6 Opposition

From the **Figure 12** , Virat Kohli has scored against West Indies at an average greater than 50. He also has a very good average of 48 against his favourite opponent Australia.

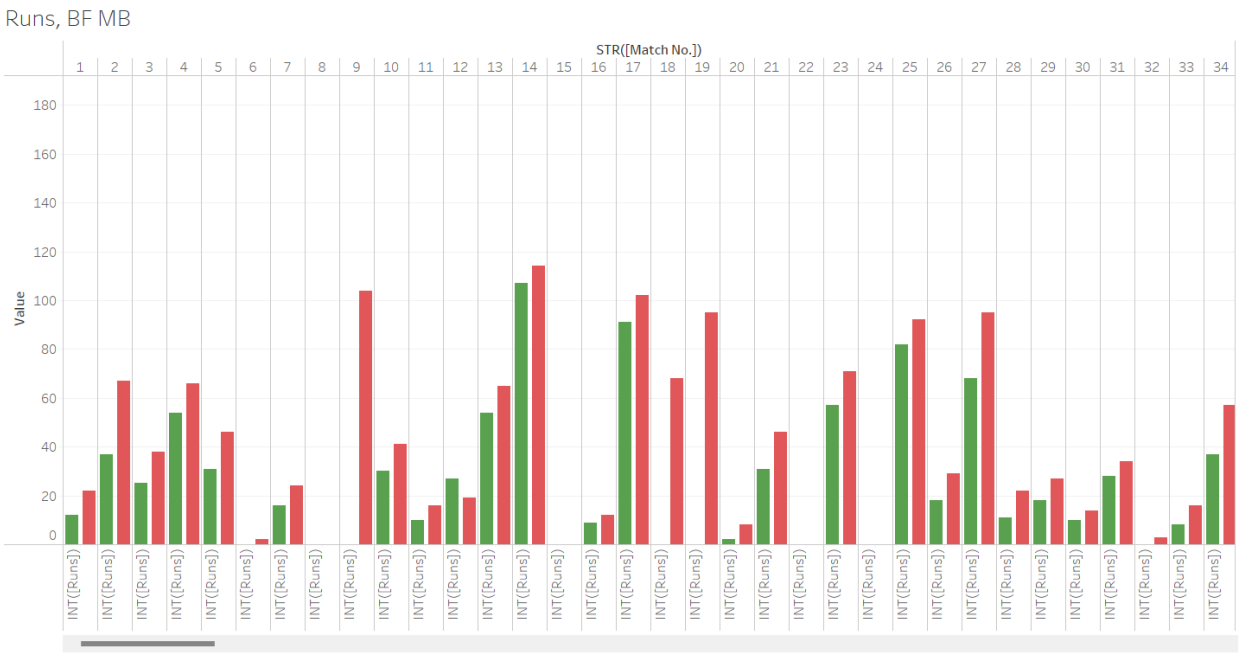


Figure 13 Runs, BF and MB

Figure 13 shows that in the prime of his career, there has been a run each ball he's faced. But, lately due to lack of form he has tried to score more runs in less number of balls and evidently the scores have been low.

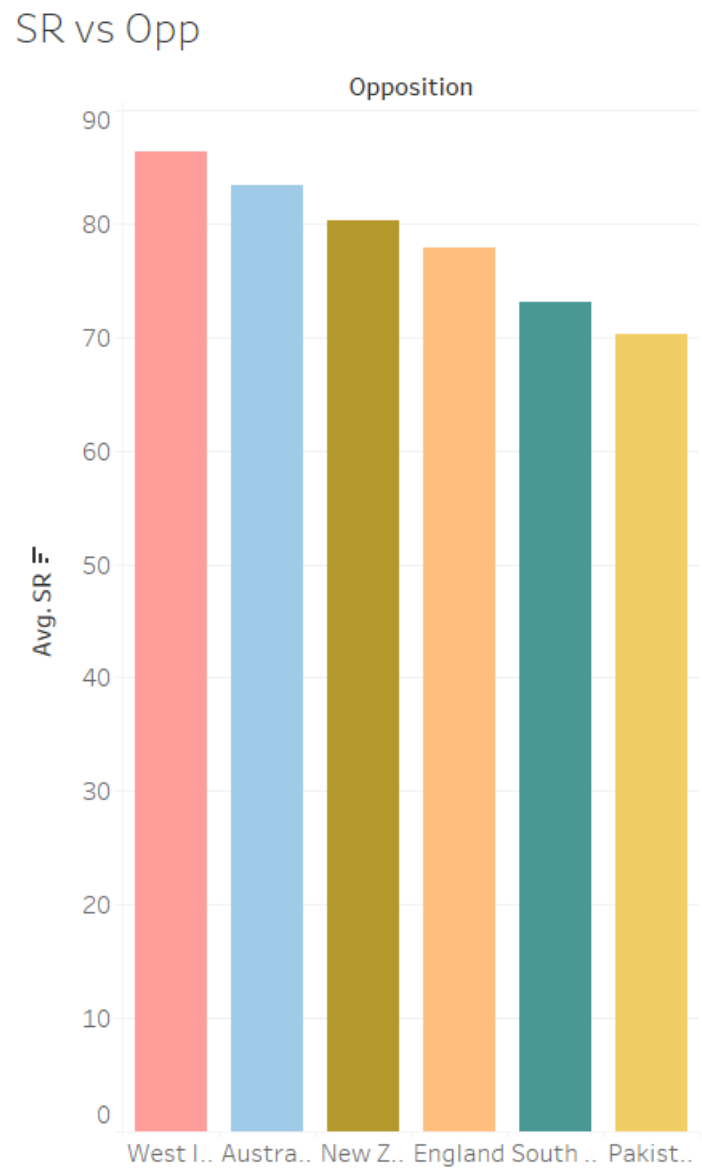


Figure 14 SR vs Top 6 Opposition

From the **Figure 14,** Virat Kohli has scored at a brisk pace against West Indies at a Strike rate nearly touching 90s. He also has a good Strike rate greater than 80 against Australia and New Zealand.

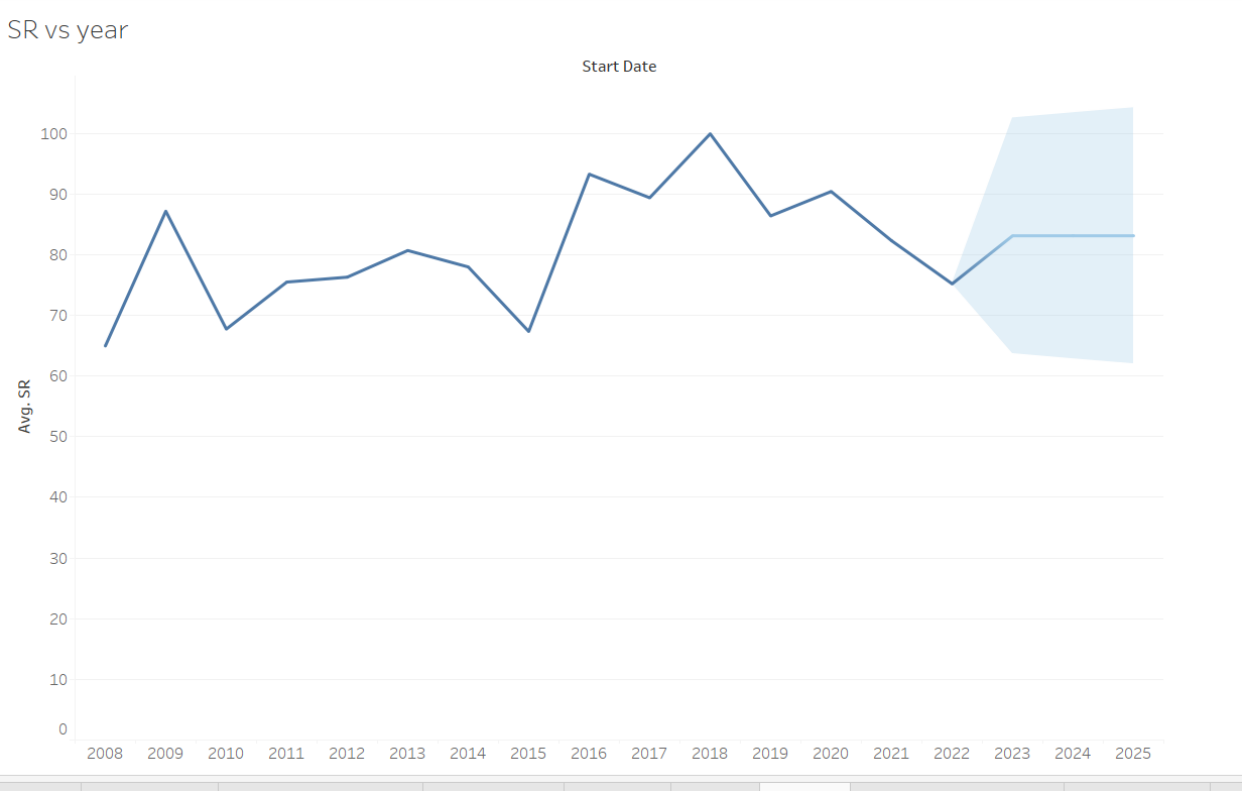


Figure 15 SR vs Year

Figure 15 depicts that from 2016-2019, the runs have come at a quick pace as he was in the form of his life. The Strike rate drastically reduces in 2021 and our forecast says that his Strike rate might average in the mid-80s in the upcoming years.

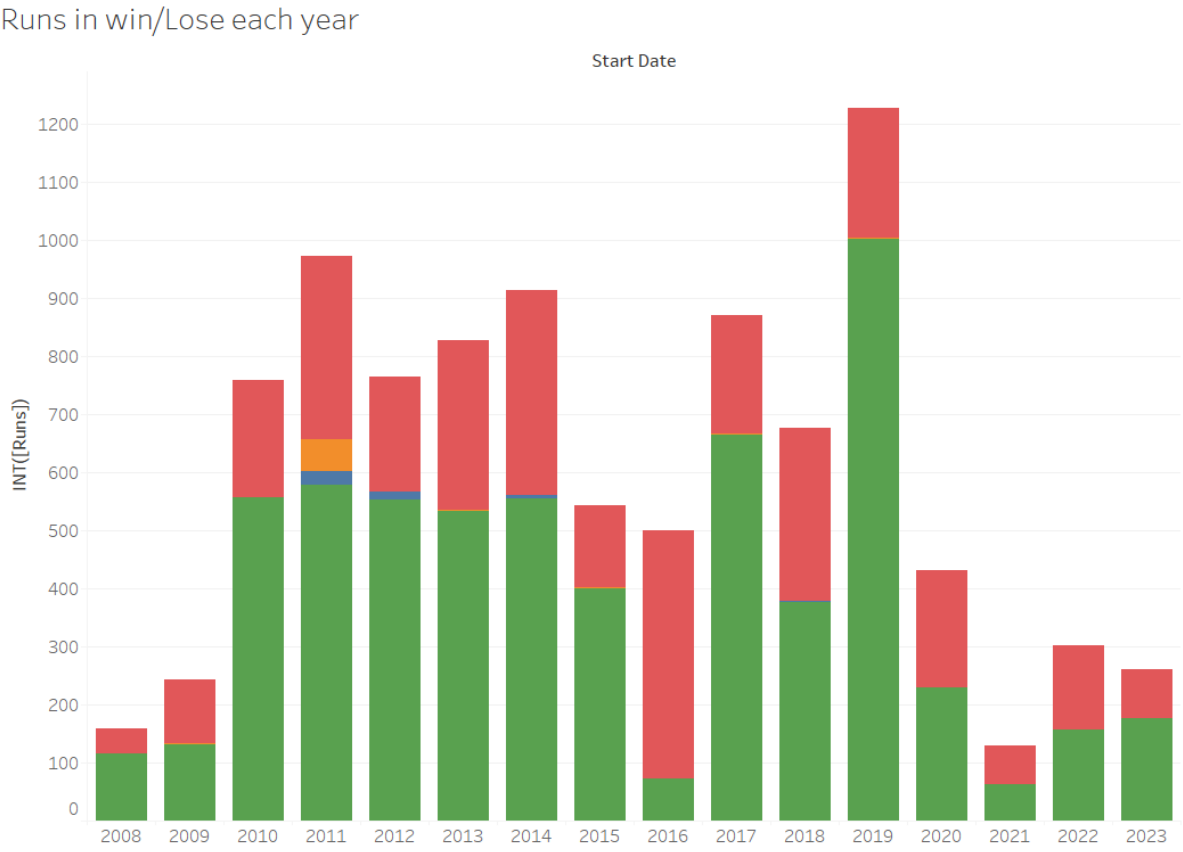


Figure 16 Runs in Won/Lose each year

Clearly from the **Figure 16**, we can see a lot of green colored stacked bar indicating that most of Virat's runs have come in Wins for Team India. In 2019, he has more than 1000 runs in Wins which is terrific.

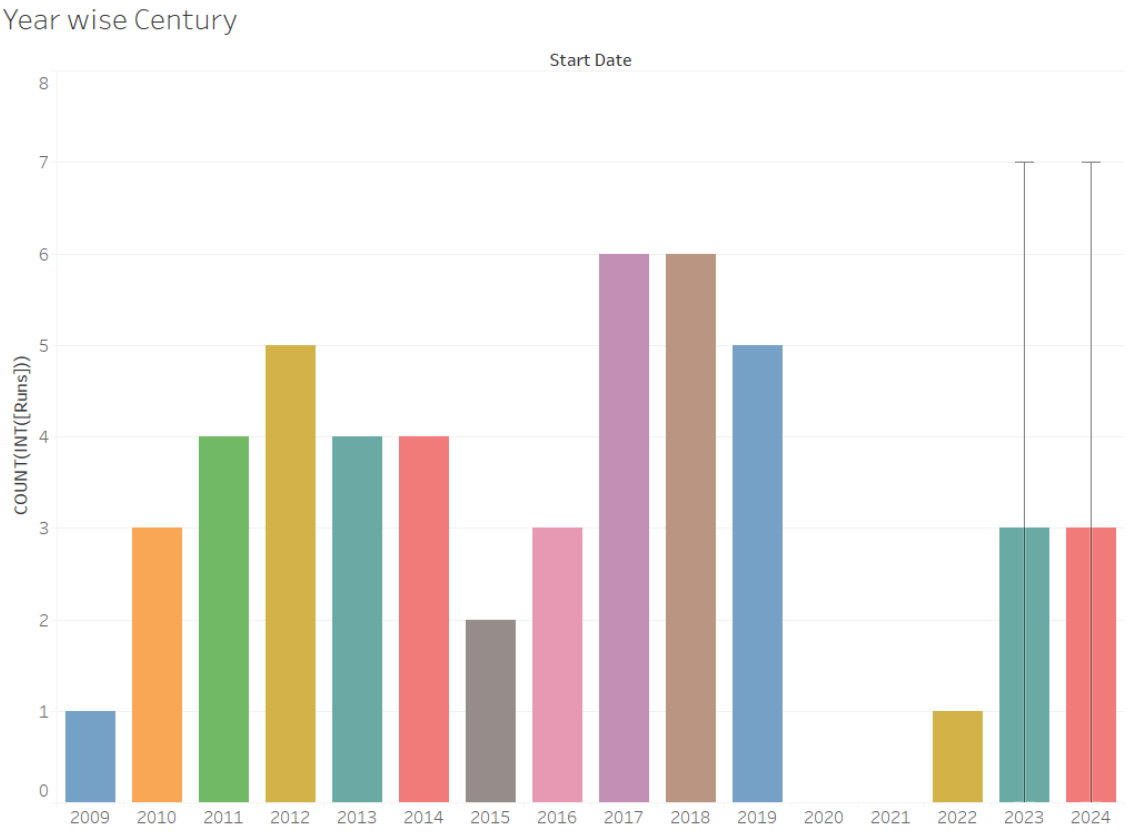


Figure 17 Year-wise century

Figure 17 shows that from 2012-2019, Virat Kohli has been scoring centuries for fun (39 out 46 centuries). Clearly, in 2017,2018 and 2019 he has been in prime form with 14 centuries in 3 years. Due to certain factors like COVID, he has not been in great form and centuries haven't come in the years 2020 and 2021. He scored 1 century towards the end of 2022. And the forecast predicts that he would score 3 in the upcoming two years each.

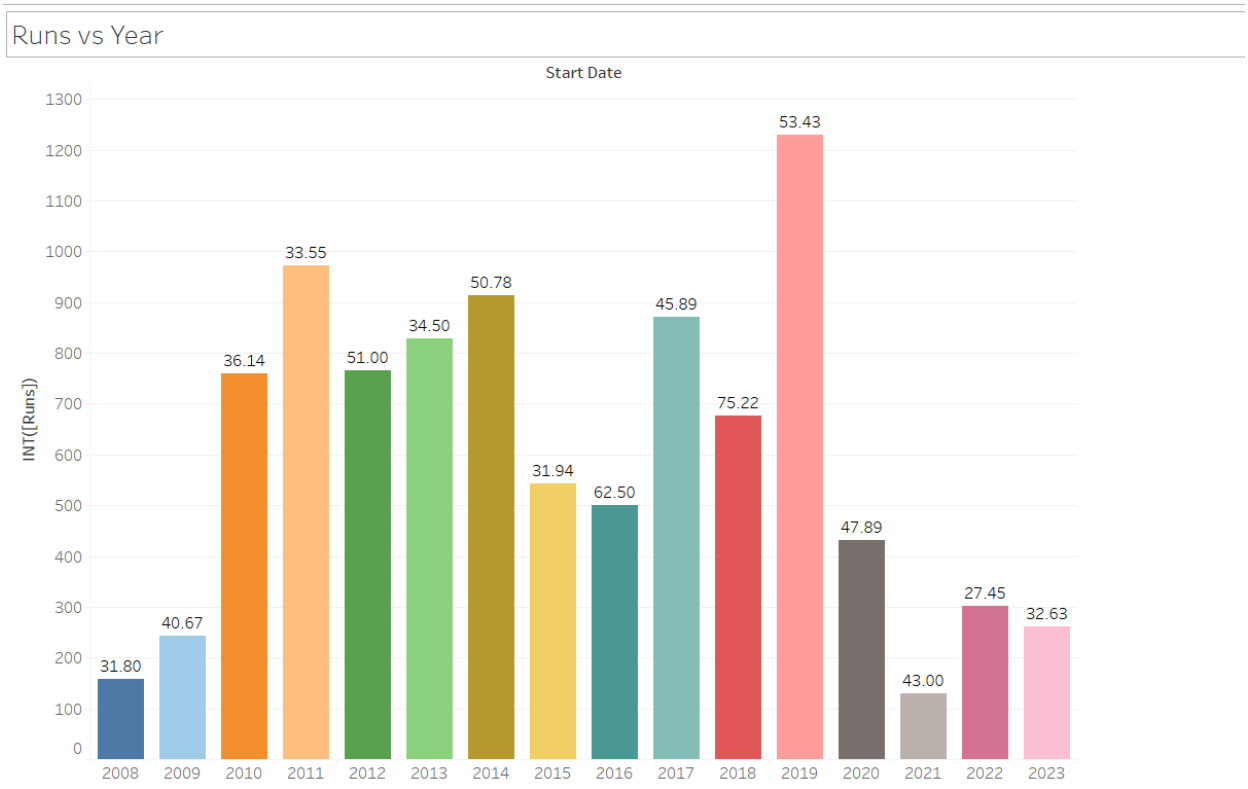


Figure 18 Runs vs Year

From 2012-2019, Runs have always been greater than 700 in most of the years with a healthy average of above 50 in most years as seen in **Figure 18.**

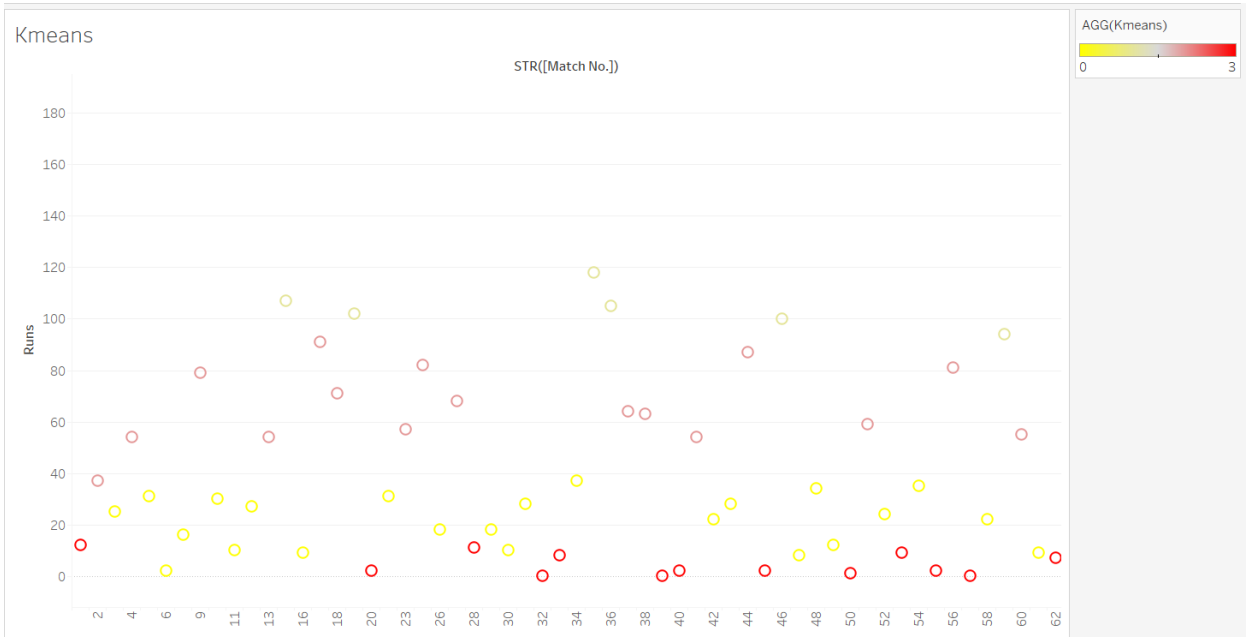


Figure 19 Kmeans Algorithm

In the above **Figure 19**, Runs vs Match No. is plotted using bubble chart and kmeans algorithm is applied using the Tabpy open-source tool (which is an integration of Python and Tableau). The features considered for the kmeans algorithm are Runs, Balls Faced and Strike Rate. From the visualization, it can clearly be seen that it has been grouped into 4 clusters based on the number of runs, balls faced and strike rate.

6. CONCLUSION:

This work is an analysis on performance of Virat Kohli, one of the stars of Indian cricket in Limited Over Internationals (LOI). 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. 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, Bubble chart and Pie chart using Tableau and also integrated with Python (Tabpy).

Further, 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. This methodology can be used for any batsman to find their impact on the team.

7. REFERENCES:

[1] Liu, A., Mahapatra, R.P. and Mayuri, A.V.R., 2021. Hybriddesign for sports data visualization using AI and big data analytics. Complex & Intelligent Systems, pp.1-12.

[2] Kanungo, V. and Tulasi, B., 2019. Data visualization and toss related analysis of IPL teams and batsmen performances. International Journal of Electrical and Computer Engineering, 9(5), pp.4423-4432.

[3] Loh, C.S. and Sheng, Y., 2015. Measuring expert performance for serious games analytics: From data to insights. Serious games analytics: Methodologies for performance measurement, assessment, and improvement, pp.101-134.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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).

[12] 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.

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

[14] 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.

[15] 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.

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

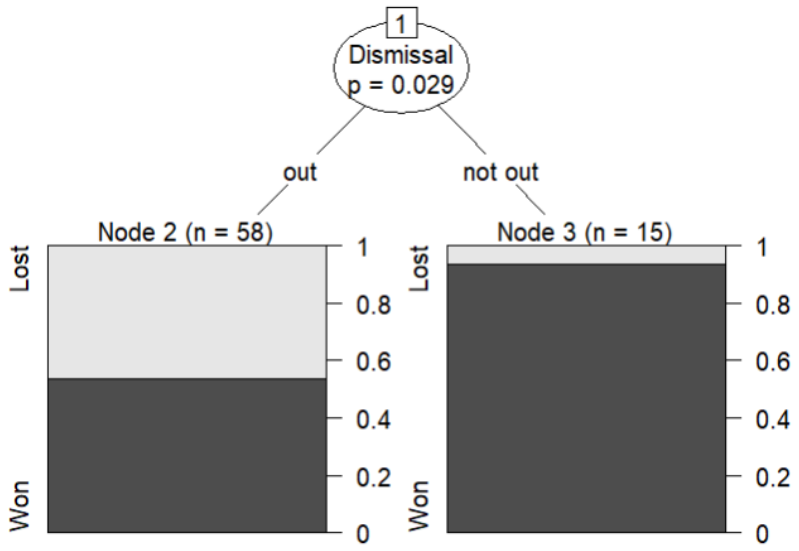
[17] Iyer, S.R. and Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), pp.5510-5522.

[18] Shenoy, A.V., Singhvi, A., Racha, S. and Tunuguntla, S., 2022. Prediction of the outcome of a Twenty-20 Cricket Match. arXiv preprint arXiv:2209.0634.

APPENDIX:

R Implementation of Decision Tree:

```
df=read.csv("D:\\Surya\\6th_sem\\DV_theory\\J-comp\\R_implementation\\VK_dataset.csv",stringsAsFactors=T)

df$Runs=as.numeric(df$Runs)

df$BF=as.numeric(df$BF)

df$SR=as.numeric(df$SR)
df$vs=as.factor(df$vs)

df$Start.Date <- as.Date(df$Start.Date,format = "%d-%b-%y")

# extract the year and convert to numeric format
df$year <- as.factor(format(df$Start.Date, "%Y"))

sample_data = sample.split(df, SplitRatio = 0.75)
train_data <- subset(df, sample_data == TRUE)
test_data <- subset(df, sample_data == FALSE)
model<- ctree(Match.Result ~ Runs + SR+BF+Inns  + Home.Away + vs + year + Dismissal+bowler.type.dismissed, train_data)
plot(model)
```




```
# testing the people who are native speakers
# and those who are not
```

```
predict_model<-predict(model, test_data)

# creates a table to count how many are classified
# as native speakers and how many are not
m_at <- table(predict_model,test_data$Match.Result)
m_at

##
## predict_model Lost Won
##          Lost    6   0
##          Won     3  18

ac_Test <- sum(diag(m_at)) / sum(m_at)*100
print(paste('Accuracy for test is found to be', ac_Test))

## [1] "Accuracy for test is found to be 90.6666666666667"

Accuracy2=round(ac_Test,2)

predict_model=as.character(predict_model)

dec_f1=F1_Score(predict_model,test_data$Match.Result)
dec_f1=round(dec_f1,2)

dec_pre=Precision(predict_model,test_data$Match.Result)
dec_pre=round(dec_pre,2)

dec_call=Recall(predict_model,test_data$Match.Result)
dec_call=round(dec_call,2)

stat2=rbind(dec_f1,dec_pre,dec_call,Accuracy2)
colnames(stat2)="Scores"
row.names(stat2)=c("F1 Score","Precision","Recall","Accuracy")
stat2

##           Scores
## F1 Score    0.80
## Precision   1.00
## Recall      0.67
## Accuracy   90.67
```