Ishan Bhatt

Student id: 180006050

Course Title: Data Analytics and intelligence (563,686)

**Special Thanks to mr. fadi fayez**

Prediction on Cricket

**Executive Summary / Abstract:**

**Predicting the result of any sport prerequisites detailed knowledge of the selected dataset and the methods used to build the classifiers on the data set. Cricket experts say that it is a gentlemen game. One of the world-renowned sports is cricket, and machine learning plays a major part in predicting a cricket's result. There are other factors, such as a venue, toss, weather condition, team combination and current form of players, which affect the match situation. This study investigates which model is best adapted to predict the match outcomes in terms of accuracy, precision and rate of recall. Following application of the feature selection on the training data set, Naïve Bayes’, Random forest, and Decision Tree algorithms were taken into account to achieve a better prediction. From the algorithm, we can conclude that it was found that some algorithms were able to predict the reasonable quantity of accuracies and the outcomes.**

**Keywords: Cricket, Modelling, Winner Prediction, Supervised Learning**

|  |  |
| --- | --- |
| Table Contents | Pages |
| Introduction | 1 |
| Problem Statement | 2 |
| Literature review | 2 |
| Data Description | 4 |
| Data tools | 5 |
| Result Analysis | 6 |
| Conclusion and Future works | 9 |
| Appendices | 10 |
| References | 14 |

**1. Introduction**

Cricket is a bat and ball sport in which two teams, each of 11 players and has many formats such as ODI, Test, and T-20 match. The batting squad has two players on the pitch at a specified moment, and the bowling squad has 11 players on the field, one of whom is a bowler. Each player is running between the pitches or hitting boundaries (4 or 6 runs) on the batting team score. The score of the team is the total score for each player. Each team has 10 wickets, so the bowling team's goal is to get as quickly as possible the 10 wickets. By scoring more runs, the target is to outperform the other team [10].

It is a huge market for business all over the world. Analysing this game could, therefore, be considered one of the most interesting projects in data mining. There are hundreds of statistics from which to compare and analyse, each of which is important in its way. As cricket is a massive money market, there is a considerable demand for the game winner's forecast that can be used for betting purposes.

We wanted to take advantage of the reality that this collection of data gives us ball information about each game. This was a vital piece of data that enabled various predictions and findings. Because we had performed information on the team matches, the winner could be predicted based on previous games and whether it is home ground or not. Many characteristics provided as a location in the dataset, ball by ball performance to batsman ball, is used to predict the winning team of the game.

**A. Project Scope**

This project aims to explore the connection between different attributes such as Home ground or Away ground, venue, and weather condition for cricket players and the use of this value in predicting match outcomes. To accomplish this, we will use a cricket dataset to construct a range of distinct predictive models and assist them both for accuracy, as well as how essential the forecast is to the different attributes.

**B. Goals**

The main reason for doing these tests is to understand the more factors of cricket game and how we can improve predict the game's winner by using some past information of the players and attributes. An additional goal is to increase the understanding of what makes a great cricket player, as we examine the predictive models in more depth to look at how the various features impact the ability to win matches.

**C. Objectives**

This project aims to predict the winner of cricket matches correctly from the various attributes before the game begins. We are especially interested in the rate of accuracy and precision on the forecast over other factors.

**2. Problem Statement:**

The issue we are trying to analyse with this experiment is one of prediction, in which, we want to understand how precise the dependent variables affect the results and from this, check the accuracy of prediction result which is going to win the game.

The significance of solving this issue originates from both the immediate consequence but also from the more indirect data acquired from understanding what influences the game and its relative usefulness in predicting the outcome.

**3. Literature Review:**

Prediction in sports analytics is not easy to part as there are many attributes from this prediction is possible. However, history said that very few individual and professional show interest of prediction on cricket compare to other sports as it is not an easy task to predict score or performance on any game. There are many formats of cricket. Moreover, all have different overs, and some of have unique rules. Nowadays, among these formats, the T-20 match is the most popular all around the world. Before the T-20, ODI cricket match is trendy. Also, many countries introduce T-20 leagues, which is more attractive to the world of cricket. Also, in cricket, Duckworth-Lewis-Stern (DLS) method used if any interruption is occurred such as rain which is based on statistical calculations. Some cricket data shows that some players are not scored well if they are not played in their home pitch.

Moreover, some teams manage efficiently, and they performed well. Therefore, some point, home pitch or away pitch is also affected to player’s performance. This research paper includes whether home ground or away ground is affected by the team performance or not.

Very few have worked on predicting the outcomes as Scott Brooker, and Seamus Hogan predicts the performance of the cricket matches by using various statistical methods. It measures how well the average batting squad will do under specified circumstances against the average bowling team and the present state of the match. Also, measures winning probability by using cricket data information, for example, ball remaining, current score and target (Singh, T., Singla, V., & Bhatia, P., 2015).

However, another study examines the prediction of the perfect cricket team by using various analysis such as player efficiency and performance. This analysis calculates the performance of players according to various criteria; it predicts the best team (Amin, G. R., & Sharma, S. Kumar, 2012).

**4. Data Description:**

In this project, there are two datasets one contains match details of over 6000 T20 matches, including innings scores and results and second one, Name and winners of all T20 series. However, we do not include the second dataset as it does not require to predict the winner.

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Description | Type |
| Match\_id | Unique ID | Numeric |
| Series\_id | Unique ID | Numeric |
| Match\_details | Team Name and venue including date | String |
| result | Match Result | String |
| Scores | summary of scores of both innings | String |
| date | date of match | Time and Date |
| venue | Name of Stadium | String |
| round | Tournament Type whether it is semi-final or final etc. | String |
| Home | home or designated home team | String |
| Away | away or designated home team | String |
| winner | Home team or Away team | String |
| win\_by\_runs | Winning margin | Numeric |
| win\_by\_wickets | Wicket margin of winning team | Numeric |
| balls\_remaining | Remaining ball | Numeric |
| innings1 | team batting first | Numeric |
| innings1\_runs | first innings score | Numeric |
| innings1\_wickets | first innings wickets | Numeric |
| innings1\_overs\_batted | actual length of first innings | Numeric |
| innings1\_overs | maximum length of first innings | Numeric |
| innings2 | batting second | Numeric |
| Innings2\_runs | second innings score | Numeric |
| Innings2\_wickets | second innings wickets | Numeric |
| Innings2\_overs\_batted | actual length of second innings | Numeric |
| Innings2\_overs | maximum length of second innings | Numeric |
| D/L\_method | To determine winner. | String |
| target | rain-adjusted target if D/L method used | String |

**A. Data Cleansing:**

DC is one of the vital data mining steps. It involves processing the data given by correcting or removing redundant data and fixing the missing or incomplete data with fair values. This may require changing the primary resource for information mining requirements to fit better.

**B. Data Preparation:**

In this dataset, we used a CSV format, which is used in a structured format to store data. For each component, these files have specific tags, and the information about each component is stored in the corresponding tags. For each component, it creates mappings and has values that can be strings, numeric, list, and so forth.

It is described as a method to remove redundancies or duplicities, if available and convert raw information into an accessible format that can fit into a model. The dataset selected to predict the result contained a few redundant values that were useless to create the issue while uploading. Many columns contain information which is not useful in our prediction; therefore, we removed those values and successfully stored the file in WEKA and evaluate the classifier output.

**C. Algorithm Used:**

**a. Naive Baye’s (NB):** NB is one of the learning that is performed by probabilistic methods. NB is a straightforward method for building classifiers: models that assign class labels to problem cases, represented as feature value vectors, where class labels are taken from some finite set. This implies that even with high-dimensional data points and a significant amount of information points, NB classifiers can perform well (Naive Bayes for classification, 2019).

**b. Random Forest (RF):**

It is made up of a significant amount of individual decision trees that function as an outfit. Each tree spreads a class forecast in the RF, and the class with the most votes becomes the basis of our model. It can be used for problems with classification and regression. There is a straight correlation between the number of trees in the forest and the outcome it can achieve, however, the maximum the number of trees, the more precise the result. Also, it can manage values that are lacking and also handle the overfitting issue (machine learning, 2017).

**c. Decision Tree (DT):**

It is possible to use the DT algorithm to solve regression and classification issues. It is to make a training model that can be used to predict the class or particular variable factors by learning from previous information. Compared to other classification algorithms, the level of understanding of the DT algorithm is very simple (Decision Tree, 2017).

**5. Data Tools:**

**A. WEKA**

It is a unique tool with all the machine learning algorithms based on Java. It also offers multiple techniques for selecting features.

**B. Language R:**

It is an open-source programming language. It is mostly used in developing statistical software and data analysis. It is most popular because of the easy UI and varieties of syntax.

**C. Tableau:**

Tableau is one of the best visualisation tools which can connect almost any databases and servers. So, we can analyse the information rapidly. It is simple to use as we only drag and drop the data and can even share the visualisation with anyone. Not only this, but we can also generate a dashboard as per clients requirement. Technology is moving faster nowadays, and data visualisation is the best way to present data. From this, all executives can easily understand the data, whether they are a technical person or a non-technical person. It is, however, a way of converting raw data into meaningful information.

**6. Result Analysis:**

In this dataset we select the entire two leagues as there are many leagues in this dataset one is Final round and the second one is South Group. In this primary dataset, DT, RF, NB classification methods were applied to the primary data. The data run using two testing options: 10-Fold Cross-validation and percentage Split - 20%.

**i.) Performance Measures (PM):**

Confusion Matrix (CM): It is regarded one of the most critical performance measures to calculate the accuracy, precision and recall as we can see from the matrix how many cases are counted as TP /TN / FP / FN.

However, the formula of CM is TP+TN/ (TP+TN+FP+FN).

Precision and Recall are the two other parameters of PM.

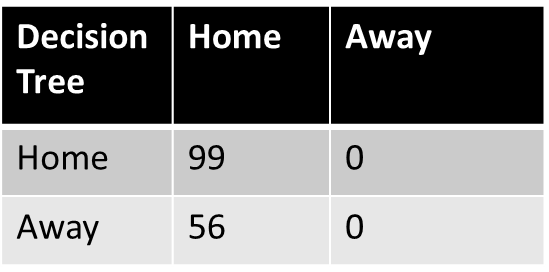
Precision = TP/TP+FP

Recall = TP/TP+FN

**A. Dataset (Final Round):**

**1. DT:**

**10-Fold Cross-validation (CV):**



In this, accuracy turned out to be 63.871 per cent, which is considered a good percentage and using this techniques betting can be possible.

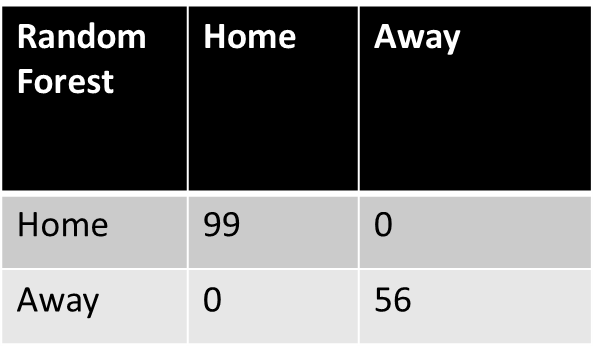
**Test Data (20% per cent split)**

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **Home** | **Away** |
| Home | 81 | 0 |
| Away | 43 | 0 |

The accuracy of these methods is pretty good, which is 65.3226 % and from this test options betting will be possible for the upcoming matches.

**2. RF:**

**10-Fold CV:**



In this, accuracy turned out to be 100 per cent, which is considered a good percentage; but, using this technique, betting cannot be possible. This problem called overfitting.

**Test Data (20% per cent split):**

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **Home** | **Away** |
| Home | 80 | 1 |
| Away | 42 | 1 |

In this, accuracy turned out to be 65.3226 %, which is considered a good percentage and using this technique, betting can be possible.

**3. NB:**

**10-Fold CV:**

|  |  |  |
| --- | --- | --- |
| **Naive Bayes** | **Home** | **Away** |
| **Home** | **46** | **53** |
| **Away** | **29** | **27** |

The accuracy of these methods is pretty good, which is 47.0968 %, and this is not considered as the best model; from this test options betting will not be possible for the upcoming matches.

**Test Data (20 per cent split):**

|  |  |  |
| --- | --- | --- |
| **Naive Bayes** | **Home** | **Away** |
| Home | 57 | 24 |
| Away | 22 | 21 |

In this, accuracy turned out to be 62.9032 percentage, which is considered a good percentage and using this technique, betting can be possible.

**B. Dataset (South Group):**

**1. DT:**

**10-Fold CV:**

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **Home** | **Away** |
| **Home** | **164** | **0** |
| **Away** | **157** | **0** |

In this, accuracy turned out to be 51.0903 % per cent, which is not considered as a good percentage; but using this technique, betting can be possible.

**Test Data (20% per cent split):**

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **Home** | **Away** |
| **Home** | **127** | **0** |
| **Away** | **130** | **0** |

In this, accuracy turned out to be 49.4163 % per cent, which is a little bit below from the perfect model and using this technique, betting cannot be possible.

**2. RF:**

**10-Fold CV:**

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **Home** | **Away** |
| **Home** | **102** | **62** |
| **Away** | **102** | **55** |

In this, accuracy turned out to be 48.9097 %, which is not a good model and using this technique, absolutely betting is not advisable.

**Test Data (20 per cent split):**

|  |  |  |
| --- | --- | --- |
| **Naive Bayes** | **Home** | **Away** |
| **Home** | **126** | **1** |
| **Away** | **125** | **5** |

The accuracy of these methods is pretty good, which is 50.9728 % and from this test options betting will be possible for the next matches.

**3. NB:**

**10-Fold CV:**

|  |  |  |
| --- | --- | --- |
| **Naive Bayes** | **Home** | **Away** |
| **Home** | **79** | **85** |
| **Away** | **79** | **78** |

In this, accuracy turned out to be 48.9097 per cent, which is not considered as the best model and using this technique, betting cannot be possible.

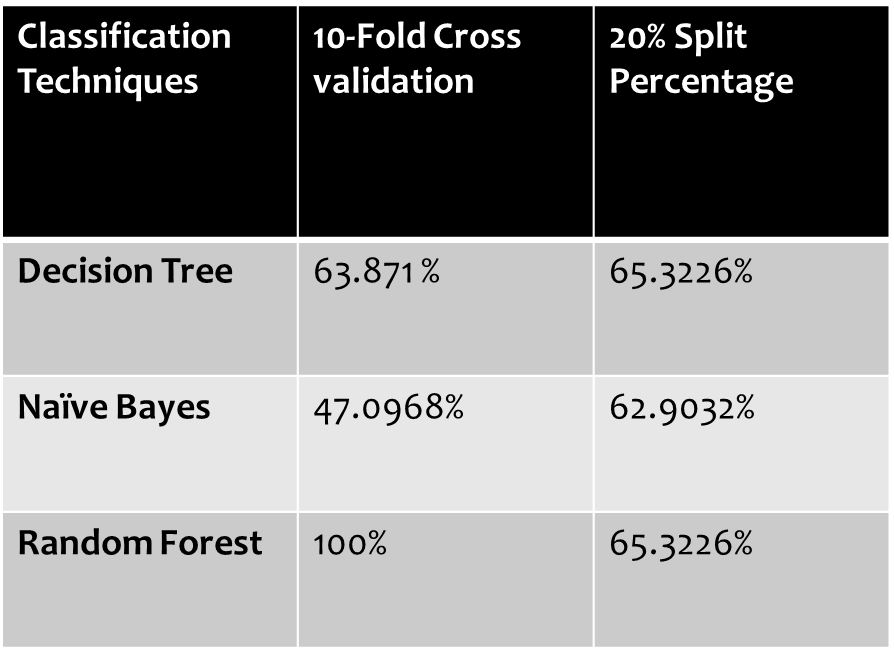
**Test Data (20% per cent split):**

|  |  |  |
| --- | --- | --- |
| **Naive Bayes** | **Home** | **Away** |
| **Home** | **72** | **55** |
| **Away** | **59** | **71** |

The accuracy of these methods is pretty good, which is 55.642 % and from this test options betting will be possible for the upcoming matches.

**ii.) Result Comparison:**

**1. Dataset: Final Round**

****

All classification techniques’ accuracy per cent varies from value to value. According to the result, we can say that Decision Tree and Random Forest produce a precise result. However, Random Forest’s output is 100% which in reality; it is not possible.

**2. Dataset: South Group**

|  |  |  |
| --- | --- | --- |
| **Classification Techniques** | **10-Fold Cross validation** | **20% Split Percentage** |
| **Decision Tree** | **51.0903%** | **49.4163%** |
| **Naïve Bayes** | **48.9097%** | **55.642%** |
| **Random Forest** | **48.9097%** | **50.9728%** |

Compare to the final round dataset; this dataset has not called as an accurate result. However, from all these results, Decision Tree and Naïve Bayes, predict precise value.

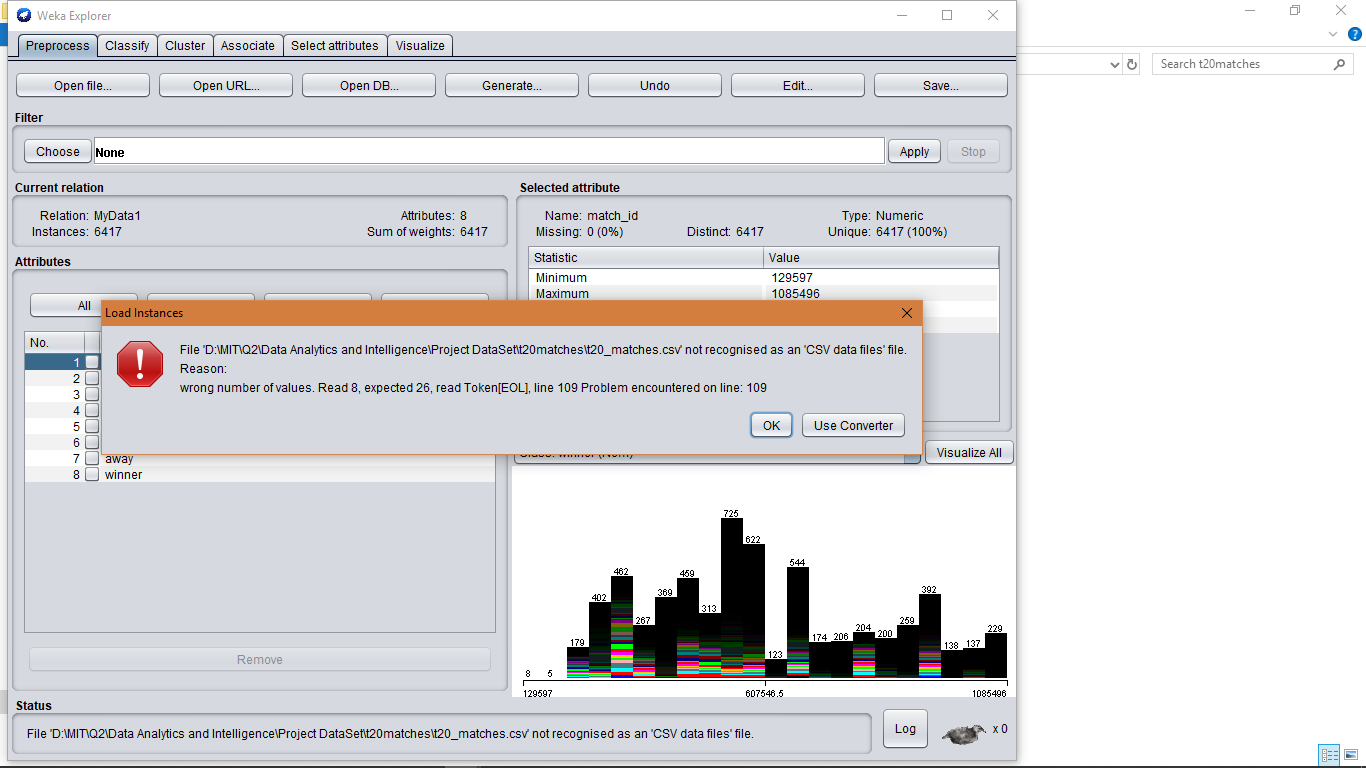
**Conclusion and Future work:**

From this various practical analysis by using Weka, various prediction method achieved different results, in which, some models can be used to predict the future matches, and some cannot be considered as accurate models. It can be seen from the experimental study that all models anticipated the various amount of the accuracy in which some models predict a precise value and some models are not. However, the percentage from 60% to 80% can be used as a prediction for betting purpose.

However, many models did not achieve the best result, and this needs to be precise in future work, and also we can run some different algorithm to achieve the best-predicted model. In the future, the focus will be on improving all models ' accuracy, especially a model which cannot be able to perform well. In addition, adding the other features such as the weather, toss, and pitch, will be utilised and use this for better prediction.

**Appendices**

**A. Error in Weka:**



After exporting data from Language R, the error is solved.

This is the code of Language R:

**Dataset – Final Round**

**getwd()**

**cricket1 <- read.csv(file.choose(), sep = ",")**

**write.csv(cricket1, file = "cricketFil.csv",row.names=FALSE, na="")**

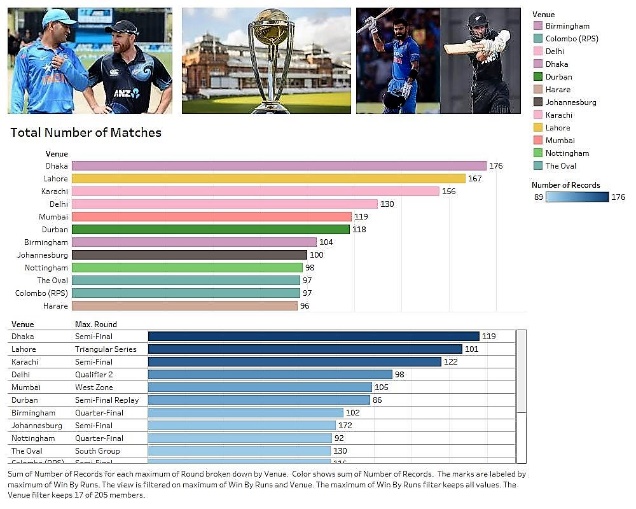
**Dataset – South Group**

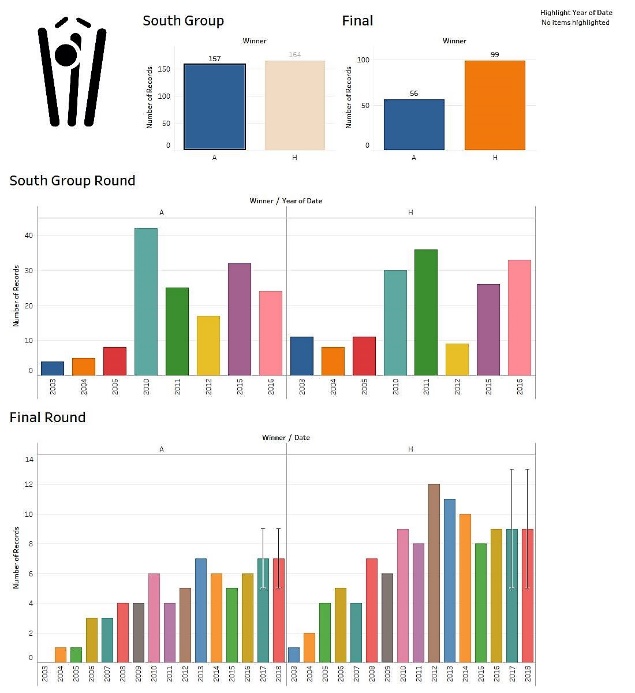
**cricket3 <- read.csv(file.choose(), sep = ",")**

**write.csv(cricket3, file = "cricketFil1.csv",row.names=FALSE, na="")**

This command not only export the file but also remove the empty variable if it is available.

**B. Tableau Dashboard:**

****

****

**References:**

**1. Cricket team selection using data envelopment analysis. (2014, January 21). Retrieved from https://www.tandfonline.com/doi/abs/10.1080/17461391.2012.705333**

**2. Score and winning prediction in cricket through data mining - IEEE Conference Publication. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/7489605/**

**3. Decision Tree Algorithm works. (2017, April 21). Retrieved from https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works**

**4. The duckworth-lewis-stern method works. (n.d.). Retrieved from https://www.espncricinfo.com/story/\_/id/19577040/how-duckworth-lewis-stern-method-works**

**5. random forest algorithm works in machine learning. (2017, October 1). Retrieved from https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/**

**6. Naive Bayes for classification. (2019, June 18). Retrieved from https://discuss.analyticsvidhya.com/t/how-to-decide-when-to-use-naive-bayes-for-classification/5720**

**7. Soni, D. (2018, May 16). Introduction to Naive Bayes Classification. Retrieved from https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54**

**8. Synced. (2018, June 25). Random Forest Algorithm Works in Machine Learning. Retrieved from https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674**

**9. Yiu, T. (2019, June 12). Understanding Random Forest. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2**

**10. Introduction to Cricket. (n.d.). Retrieved from https://www.checkfit.co.uk/introduction-to-cricket/**