The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics

Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David

Abstract: With the evolution in the field of Data Sciences, every business firm is adapting latest technologies to grow their business. There are competitions in delivering better management, better quality of evaluations and better services in the market. The only possible way to meet all these qualities is to conduct analysis of data with purity and more accurately. Machine learning is the emerging field to predict future outcomes with existing data and based on these predictions better decisions can be made. Cricket is a well-known game that played and watched around a globe in 104 countries. Many of these cricket fans want their team to perform good and declare as a winner. To make sure their team's win, team should work on their strengths and team performances. Predicting winner of a cricket match depends on many factors like batsman's performances, team strengths, venues and weather conditions etc. In this research various features have been analyzed to predict the match winner of the game. This research paper is about prediction of an IPL match winner before the match started. The winner of IPL is predicted by training machine learning models on the selected features. For this purpose of model building, different machine learning algorithms has been applied on test and training datasets of different sizes which are Random Forest, SVM, Naive Bayes, Logistic Regression and Decision Tree. The prediction model will have benefits for cricketing boards like evaluating the team's strength and cricket analysis. For gambling applications and match reporting media this model will be a blessing of disguise.

Index Terms:Cricket, Data Analysis, Data Science, Machine Learning, Model Classifiers, Modelling, Prediction, Prediction Models

1. INTRODUCTION

SPORTS statistical analysis use in sports has been growing quickly year by year. Due to which the ways in which game strategies are formed or the player's evaluation criteria has been changed but also has the got the more interest of audience towards cricket. Now Cricket has become one of the most followed team games in the world with billions of fans all across the globe. Cricket is a sports game that played globally across 106-member states of the International Cricket Council (ICC), which has 1.5 billion worldwide fans according to ICC. However, much of the global finance and interest is focused upon the 10 full ICC member nations and more specifically upon 'the big three' of England, Australia and India. Cricket has many evolutions over time. Today, there are three major formats in which cricket is being played internationally, One Day Internationals (ODIs) and the T20 cricket and Test Matches. Besides these international cricket matches, T20 League cricket is getting attention in the fans due to its shortest format and the most exciting format of the game. Indian Premiere League is one of most popular t20 cricket league in the world. Ever since its inception in 2007, IPL has been a huge success and has become an industry with investment of billion-dollars. Similarly, England's county cricket (t-twenty blast), Big bash, PSL and BPL are other big leagues who are investing a lot of money to promote their franchise-based cricket. In franchised-cricket every team wants to win and improve the team performance. For this purpose, every team needs a better management panel to handle the responsibilities of complete franchise, team selection committee who will select the best possible team with good players such as to select the best batsmen from the draft by looking at their past performances. Indian Premier League is a domestic competition played in India in April and May every year between eight teams. Eight teams participate in this competition every year. More than 150 players are selected by each team. Each team consist of 11 players, four overseas players and seven local players. Every team's performance based on the key performances of players, team conditions and other important aspects which decides the team's performances in a cricket match. The model will be built on all the possible factors affecting the outcome of cricket match. Ground impacts, team quality and home field advantage were observed to be essential in by the Nagel kerke R2 and AIC analysis. This might be on the grounds that the ICC rating assesses result (win, draw, misfortune) alongside the success edge, wickets and adversary rating. Winning the hurl was likewise considered in the model fitting however was observed to be insignificant. The playing conditions differ from ground to ground and nation to nation. For instance, playing conditions in Wankhede at Mumbai are very not quite the same as in Leeds at Headingly [14]. Player performances decides the win factor of a team. Player performances matters a lot as every team depends on their player to perform good and perform according to match situation. In selecting the lineup for the team, the player performance is taken as a major factor. Batsman performances in recent matches tells about their form, ability to score runs with a healthy strike rate which is a need of twenty-twenty cricket nowadays. Pitch Conditions are very important in cricket game. There are several kinds of pitches on which cricket has been played. Every ground and his own pitch conditions known for bowling pitches or batting pitches. A match's outcome can also be affected by bad weather. Weather conditions also plays a role in deciding results of a match. Players having good batting averages and consistent performances in the recent matches are the ones on which teams rely on. Because they can play a major role in posting a good target score and in chasing, by handling pressure situations. Sometimes, matches are interrupted by rain or any other miscellaneous circumstances. To reset the target in interrupted matches, there is an approach used name as Duckworth-Lewis or D/L method D/L [8]. Multiple Linear Regression is a valuable method to allot the winning probabilities to the contending groups in One Day International matches. With the utilization of D-L approach, this procedure can be promptly adjusted to deliver 'in the run' forecasts. While a conclusive investigation of the productivity of the betting

- *Daniel Mago Vistro, Asia Pacific University, Malaysia, daniel.mago@apu.edu.myFaizan Rasheed, Asia Pacific University, Malaysia, TP049340@apu.edu.my*
- *Leo Gertrude David, Asia Pacific University, Malaysia, leo.gertrude@apu.edu.my*
- *Leo Gertrude David, Kumaraguru College of Liberal Arts and Science, leodavid@kclas.ac.in*

market is yet to be directed, starter proof propose punters might be inclined to over or under estimate the genuine likelihood of the contending groups as the diversion advances [2].

2 Related Works

With the evolution of Cricket, it became a very hot topic for sports analysts. A lot of research has been made on cricket but due to inconsistent and complicated data sets, they could not get breakthrough in predicting match winner accurately. There are many techniques that has been used in predicting match winner like KNN, Logistic Regression, SVM, Naïve Bayes but nobody has achieved the accuracy. According to Ahmed & Nazir [1] they implemented different statistical approaches for formation of datasets and tried various classification techniques to predict the winner of One Day Cricket (50 over) match. He has predicted the winner with 80 % accuracy. Shah [14] predicted One Day International match results by using data of ICC match ratings, ICC ranking points for batsmen and bowlers, home factor, ICC rating differences and ground effects on the match. They implemented Logistic Regression on this data and achieved accuracy in predicting the results of matches 74.9% and in 81% matches they predicted the winner team correctly. Jhanwar [5] predicted 71% accuracy in predicting winner of the One Day International cricket match. He used binary classification models such as Logistic Regression, KNN, Random Forest and Decision trees. Cross validation procedure was not carried out. Jhawar [6] have done research on predicting the winner of the match at end of the over, player's performance recent and past performance and other statistics' which are necessary for predicting the winner of the match has been used. First challenge is to estimate the score that first team will score at the end of first innings. In Features combination to predict the match outcome, is relative strength of Team B divided by relative strength of Team A is successful in measuring and comparing the strength of the playing teams. By Random Forest classifier R.F.C. accuracy of 84% has been achieved. Jhanwar [5] analyzed the performances of the One Day International matches played from 2006 and 2016 and accuracy stated that 86% is achieved that top 3 positions of batsman are hot for the man of the match award which is better to previous search and models Random Forests, Decision Trees, KNN and Logistic Regression are the techniques used to predict player performances in a match. Yasir [16] predicted outcome of cricket match and for the winner prediction techniques, he proposed a method for predicting the team results and elaborated the working of method which is by using properties of dynamic team for the winner's prediction like player's history, weather conditions, ground history and winning percentage. He applied this technique on 100 matches and got 85 % prediction.

2.1 Factors to Anticipate Cricket winner

Winning a cricket match depends on multiple factors like batting, bowling, fielding, team performances and player performances. To predict the winner of a cricket match is never an easy task. But there are always some kind of unique aspects or match conditions that may favor to some team and sometime does not such as home advantage, Key Players, Pitch Conditions and weather conditions [8].

2.2 Cricket Winner Prediction Models

Machine learning has become a vast field that is consist of many domains' statistics such as artificial intelligence, information technology, and others. Many problems can be solved by Machine learning model. In the advance era of today, the machines can now work as a human brain because machine learning has been so much evolved. It is learning of computers by creating algorithms which tells the computer how to learn which includes finding the patterns using statistical approaches or similarities in the data. Machine learning algorithms has proved prediction very easy by using classification function to relate the values of attributes in the dataset [11].

2.2.1. Naïve Bayes

Naïve Bayes works on the Bayes probability theorem with the assumption that all the features are independent of class label (predicted variable) which may be a wrong assumption. Naive Bayes model used in conjunction with recursive feature elimination [10].

2.2.2. Decision Tree Regressor

Decision Tree Regressor has been used to check the overfit by learning from the noise of data using tree node system. If max depth of tree is high, decision tree regressor take details from training data's noise. Decision Trees classification works on tree node principal in which instances are sorted into tree node system. By this hierarchy complex decision-making system are break-down into smaller simpler decisions which provides a simple solution that is easy to implement [9].

2.2.3. Support Vector Machine (SVM)

Support Vector Machine has been proven to be most used component classifier of Ada Boosting for different prediction techniques like image recognition, medical health diagnosis and facial recognition. SVM classifier on given Training data, outputs an optimal hyperplane by which new examples can be categorized. Hyperplane is a plane that divides line into two parts where in each class lay in either side. SVM's optimization measured by Regularization parameters. Regularization parameter tells about the SVM Optimization [3]. SVM is a category of supervised machine learning algorithms which has to be trained with pre-defined output class. The SVM classifier on given Training data, outputs an optimal hyperplane by which new examples can be categorized. Hyperplane is a plane that divides line into two parts where in each class lay in either side [12].

2.2.4. Random Forest Classifier

Random Forest classifier is a method used for regression and classification techniques. In the Random Forest Classifiers, to classify a new instance, there are number of trees in working randomly in a forest putting input vector down and duty of every tree is to give a class label or target variable as a vote for the class. And which node has highest votes will be chosen by Random Forest Classifier. To increase the accuracy predicted and to control the over-fitting, Random forest uses estimation and averaging approach on the sub-samples of dataset that is done by fitting various number of decision tree classifiers. The sub-samples taken for this are remain equal to original input size [15]. Random forest is a versatile mechanism enough to deal with both supervised classification and regression tasks. For the datasets under experimentation,

DBDP approach achieves accuracy of that of the original Random Forest in a smaller number of trees, and the reduction in size achieved is in the range of 52% to 87% [7].

2.2.5. Accuracy Score

To optimize a model's performance, it should be ensuring that proper selection of features is under training of generative classifier. To calculate the model's performance or model's accuracy confusion matrix is a matrix which gives the comparison between the predicted class and the actual class into classification report [4].

3 Research Methodology

Methodology is a process in which data is selected, transformed and prepared for the calculations needed to generate useful insights [13]. For this research methodology is SEMMA modeling.

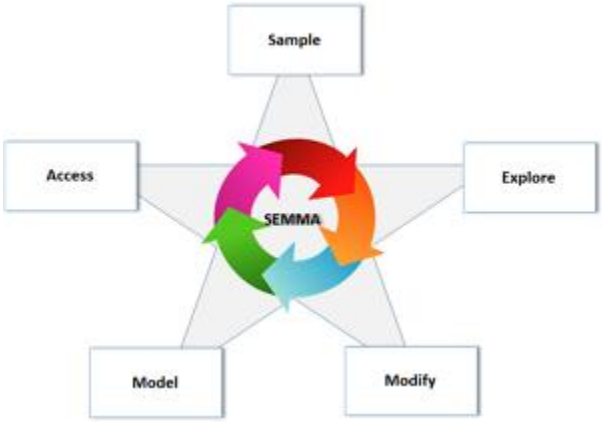

Figure 1: *SEMMA Methodology*

3.1 SEMMA

The SEMMA process was developed by the SAS Institute that considers a cycle with 5 stages for the process: Sample, Explore, Modify, Model, and Assess. Data mining is the process of discovering predictive information from the analysis of large databases. Python is used for the data mining of the following steps: There should be one informational dataset which contains enough information to fulfill the purpose of data mining and should be able to do calculations on it to generate useful insights. Target variable on which all the analysis will performed should be there in the dataset. This progression includes the utilization of information planning devices for information import, union, consolidation, filtering, connection and sifting, just as measurable examining systems. Finding patterns between data points by concatenating different options, finding correlations and relations between attributes. This step includes the exploration of data which includes checking missing values, inconsistencies, exploring variable's distributions, techniques for the determination of variables and finding factors. Purification of data includes treating missing values if there is any, removing outliers and transforming variables for getting the normal distributions of variables. This step is very important in modelling as its about modification of data. If the data will not be good, then good results cannot be generated. Using Artificial Intelligence techniques to generate useful insights, this includes training of AI models on selected data to generate results in desirable way. This step includes implementing suitable machine learning models according to nature of data for the forecasting of values of target variable. Last step is evaluation of implemented models. Checking the fitness's of models whether the model is overfit or underfit and comparing performances of models by different statistical techniques. If the model, is not appropriate and not giving the best results then try different techniques to make it appropriate.

3.2. Data Visualizations

Visualizations are important part of any research to understand the business and behavior of data in a way that how different attributes are relating to target variable and what attribute should be the point of focus. Visualizations of data give valuable meaning insights. By the visualizations every end user can easily represent the data into understandable interactive graphic. Cubes will be generated related to different aspects of data. There are various visual analytic tools to create visualizations but as this research has been done in python so visualizations will also be made in python programming using mat plot lib libraries. As the topic of this research is to predict the winner of match so all the cubes will be related about how different attributes of data are interacting with match winner variable.

4 Results and Discussions

4.1 Model's Implementation – Decision Tree Classifier

Decision Tree works on flow chart tree like structure having nodes, branches and leafs. Node represents attributes of dataset; branches are represented by decision rules and outcome of the model is represented by trees. The node on the top is called as root node and partitioning is done by it in recursive manner. With the structure of tree like flow chart it helps to make decisions. In machine learning decision trees are like white box which take a part in logics of internal-decision making which cannot be find in the black box type of algorithms like neural networks. Decision tree's time complexity can be found by number of observations and number of features in the dataset. Decision trees are non-parametric and high dimensional data can easily handle by the Decision Trees. The splitting of records in Decision trees are done by Attribute Selection Method and then splitting the data into smaller portions of data and recursively tree building process continue and end when every record plotted successfully.

Figure 2: Confusion matrix by Decision Trees

The above confusion matrix of Decision Tree model has successfully predicted the values of 'winner' by 76.9% accuracy. It may not be enough for our model as XGBoost predictions was over 90 % so we need to fine tune our Decision Tree model for better results.

Parameter's Tuning:
As the results by Decision Tree model were not perfect according to the requirements so we need to fine tune the parameters of Decision Trees. A machine learning model consists of various parameters which decide how different computations will be performed in selected models. Usually the predictions of data are made by parameters that has been already set by default in models but in some cases the results are not good enough because of different nature of data. So, if parameters are set according to requirements of data then the computations performed result in terms of better performances of model. In case of Decision Tree modeling, the maximum depth value describes how deep the tree will be. If set to be a larger value, Decision Tree model will be deeper and will cover more details about data by splitting more. The max_depth has been set to 33 in this case.Criterion of the Decision tree was Gini before, but it was not good for information gain. Now it has been changed to entropy for measuring the impurity and information gain.

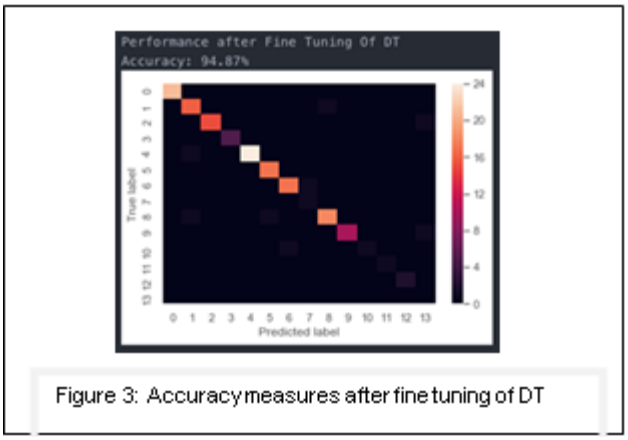

Figure 3: Accuracy measures after fine tuning of DT

Decision Tree model's performance has been changed and it has successfully predicted the winner by 94.87%. It means that tuning of parameters has made the model better and more accurate.

4.2 Model's Implementation – Random Forest Classifier
Random forest is one of the best machine learning algorithms that produces best results without parameter tuning frequently and very hard to beat in terms of performances. It's very easy to use because of its hyper parameters gives best results with default values. It avoids overfitting problem. It works both for classification and regression problems. Random forest is mixture of multiple Decision Trees that combine together to give better results. Most frequent method for training in Random Forest is bagging method and idea of this method is to combination of learning methods to enhance performance of model for the better predictions. The basic difference between the Decision Trees algorithm and random forest classifier is that in decision trees some set of rules needed to be set before applying model features and target variable and in Random Forest there is no need to set any kind of decision rules. Another difference is that sometimes deep decision trees has to suffer problems of overfitting whereas in Random Forest it prevents the overfitting. But Random Forest sometimes makes slower computation because it consists of subtrees which does not work every time. Allow else, Random Forest has various parameters to increase the model's performance like n_estimators , min_sample_leaf and max_features. Model's speed can be increased by setting hyper parameters such as n_jobs, for example, it can be set to 1 for using only one processer. By the random_state hyper parameter output of model can be made replicable and the last one is oob_score used for validation. Random Forest Classifier will be used in this research according to nature of our problem.

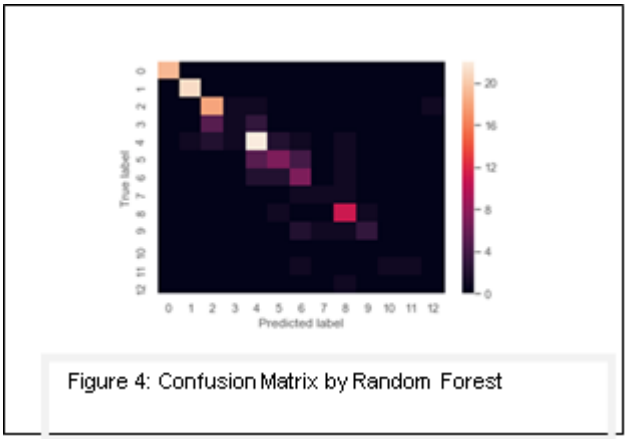

Figure 4: Confusion Matrix by Random Forest

The above confusion matrix showing that our model has predicted True label with 71 % accuracy. The performance of model was not good enough, but we can make it better with fine tuning of parameters.

Parameter's Tuning
Usually Random Forest is like a black box to which inputs are given and predictions has been made by Random Forest without knowing that what are the computations are going to take for this process. This Black Box Classifiers have several levers which we can tune to get better results. Parameter's tuning is necessary sometimes to achieve good results such as in our research 71 % is not enough so by tuning of parameter and to get better results parameter's will tune with random values. The first parameter tuned for this purpose is increasing the number of estimators from 100 to 1500. This may slow the model for milliseconds but make computations more stable and stronger, n_jobs set to 1 so that 1 processor will be used at a time and maximum depth of trees has been set to 565.

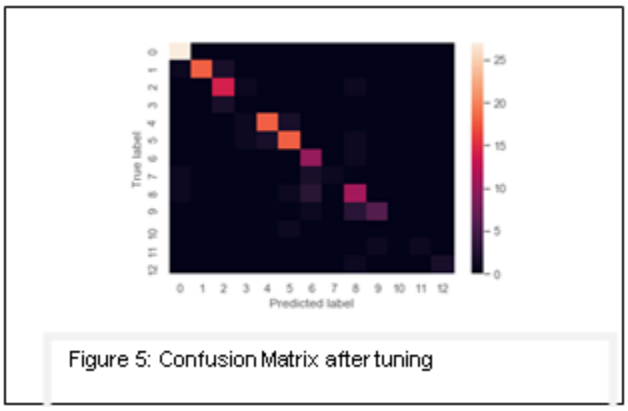

Figure 5: Confusion Matrix after tuning

Now, it can be seen that performance of confusion matrix has been improved by tuning of the parameters. And accuracy of the model after tuning the parameters of Random Forest has been increased to 80%.

4.3 Model's Implementation – XGBoost Classifiers

XGBoost is a machine learning algorithm used for prediction purposes. Engineering goal of XGBoost is to make very fast calculations by tree algorithms. It's also called gradient boosting as it supports the machine learning rate. Model's performance of XGBoost always remains good and very quick. XGBoosts prove its self the best in the machine learning whether it is the prediction of ad-click rates or the classification of higher energy physical events, It has proved its speed and performance. XGBoost can also deal with the fitting of models whether it is over fitting or under fitting as it supports different regularization techniques and can handle many other issues at the same time like handling sparse data, weighted quantile sketching, provides block structures for parallel learning, out of core computing and cache awareness. For the implementation of XGBoost data needed to be split in X and Y for training and test sets which we have done already. XGBoost used for both regression and classification tasks. It gives a wrapper class which allow models to how to treat, like behave as regressor or classifier in the scikit-learn framework.

Accuracy:
Accuracy measures tells about performance of the model by comparing results of model's performance on training's dataset with test dataset in the form of accuracy. The accuracy we got by XGBoost Classifier is 94.23 % which means that our model has behaved so well.

Confusion Matrix
It is also known as error matrix in machine learning and specially designed for classification problems as it shows the behaviors of models in which model got confused while it was doing predictions.

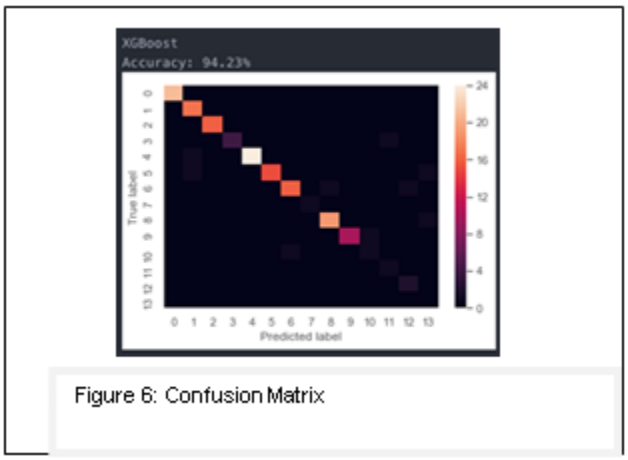

Figure 6: Confusion Matrix

The figure above is a plot of confusion matrix table describing the XGBoost classifier's performance on test data which contains true values with respect to our training data on which model has been trained. Our model has performed very well in predicting the winner.

5 Summary and Conclusion

The objective of this research was to predict the match winner of IPL using historical data of IPL from season 2008 to 2017. To conduct the analysis and predicting the winner of IPL various branches of Data Science has been converged including Pre-Processing of data, Visualizations of data, preparation pf data, feature selection and implementing different machine learning models for the predictions. SEMMA methodology has been selected for conducting the analysis of IPL T20 match winner dataset. Preprocessing has been done on the dataset to make it consistent by removing missing value, encoding variables into numerical format. Best features were selected by visualizing attributes of data with target variable. On selected features several machine learning models has been applied on the to predict the winner and the results were outstanding. First of all, Decision Tree model was applied which predicted the match winner with good accuracy 76.9%. To improve the performances of model, we fine-tuned the parameters of Decision Tree model and achieve good results. Model performance was enhanced by 76 % to 94%.

Models	Accuracy
Decision tree classifier	94.87%
Random forest classifier	80.76%
XGBoost classifier	94.23%

Table 1: Classifiers performance percentage

Then we applied Random Forest model on the selected features and the predicted the winner with 71% accuracy which was not good enough, so Random Forest Model was also tuned by parameter's tuning and results got better with 80 % accuracy. In the last XGBoost machine learning model was applied, and results were outstanding. The accuracy achieved by XGBoost was 94.23 % without tuning of parameters. The table above is explaining the performances of our classifiers in predicting the winner. In cricketing field, to achieve the full convergence into data science world, it would require a lot of additional data to meet the full picture of analysis, i.e. every

computation needs to perform very well, data needs to meet all the business problems and business systems. The prediction of winner produced through this project required a lot of domain information and expertise for observations and their relations to the winning team.

ACKNOWLEDGMENT

First of all, we would like to express our deepest gratitude to the God who has given us the knowledge and wisdom so that we can finish this research. Thank you for our family and colleagues who in a way always support and encourage us to work hard.

REFERENCES

[1] Ahmed, W. & Nazir, K., 2015. A Multivariate Data Mining Approach to Predict Match Outcome in One-Day International Cricket. 10.13140/RG.2.2.30683.46880.

[2] Bailey, M. & Clarke, S. R., 2006. Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress. Department of Epidemiology & Preventive Medicine, Monash University, Australia Swinburne University of Technology, Melbourne, Australia. .

[3] Firat, . U. S., Vargeloğlu, O. B. & Bingol, S., 2016. A Literature Review of Adabost and SVM Techniques. Conference: 3rd International Management Information Systems Conference, At İzmir Turkey.

[4] Hossin, M. & Sulaiman, M., 2015. A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. International Journal of Data Mining & Knowledge Management Process (IJDKP) , 5(2).

[5] Jhanwar, G. M., 2017. Quantitative Assessment of Player Performance and Winner Prediction in ODI Cricket. International Institute of Information Technology Hyderabad - 500032, INDIA.

[6] Jhawar, M. G., Viswanadha, S., Sivalenka, K. & Pudi, V., 2017. Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths.. Conference: Machine Learning For Sports Analytics at ECML-PKDD .

[7] Kulkarni, V. & Sinha , P., n.d. Effective Learning and Classification using Random Forest Algorithm. International Journal of Engineering and Innovative Technology (IJEIT).

[8] Lokhande, A., Chawan, R. &. & Pramila &, S., 2018. Prediction of Live Cricket Score and Winning.. Computer and IT Dept, Veermata Jeejabai Technological Institute, Mumbai, India , 5(4)(2394-9333).

[9] Mitchel, M. T., 1997. Machine learning.. Burr Ridge, IL: McGraw Hill, 45, 1997.

[10] Murphy, K. P., 2006. Naive bayes classifiers.. University of British Columbia.

[11] Nasteski & Vladmir, 2007. An Overview of the Supervised Machine Learning Methods. Faculty of Information and Technology.. Faculty of Information and communication Technologies.

[12] Patel,S.,n.d.MachineLearning101[Online]

[13] Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[14] Shah, P. & Shah, M., 2015. Predicting ODI Cricket Result. ISSN (Paper) 2312-5187 ISSN (Online) 2312-5179 An International Peer-reviewed Journal , Volume 5.

[15] Asare-Frempong, J. and Jayabalan, M., 2017. Predicting customer response to bank direct telemarketing campaign. In 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T) (pp. 1-4). IEEE.

[16] Yasir, M. et al., 2017. Ongoing Match Prediction in T20 International. IJCSNS International Journal of Computer Science and Network Security.