ANALYSING AND PREDICTING THE RESULTS OF IPL MATCHES BASED ON VARIOUS FACTORS.

Anusha C
PES1UG19CS074
Computer Science and Engineering
Pes University
Bangalore, India
anusha20c@gmail.com

Thrupthi M S
PES1UG19CS541
Computer Science and Engineering
Pes University
Bangalore, India
thrupthisomashekar@gmail.com

Divya K C
PES1UG19CS147
Computer Science and Engineering
Pes University
Bangalore, India
divyachandrashekhardiv@gmail.com

***Abstract* — We performs in depth analysis on the matches played during the Indian Premier League (IPL). Cricket, especially the T20 format, has maximum uncertainty, where a single over can change the momentum of the game. IPL Data Analysis is all about the analysing the data that is already present in data set using data science, machine learning and R. This is an application design for the purpose of analysing the data by fetching the attribute from the dataset and predicting the future of the match. EDA will help us to find patterns in data and determining relationships in data and. Visualizing the trends in performance of the team and predicting what can be the results in the next coming matches. This will helps to identify the team that has more chances to identify the team that has more chances to win the upcoming seasons.**

In this project we try to evaluate the effects of these major factors on win count:
1.LOCATION OF THE MATCH
2.TOSS WINNING
3.PERFORMANCE IN PREVIOUS MATCHES

I. INTRODUCTION

Cricket is the biggest tournament played in all most all the countries. It is the game between two teams in which each team has 11 players the final result will be either loss or win or at the rare cases points will be shared with both teams which mean no team has lost or won. Sometimes the game is unpredictable because of that game keeps on changing each and every time.

The Indian Premier League (IPL) is a Twenty20 cricket league tournament held in India contested during April and May of every year where top players from all over the world take part. The IPL is the most-attended cricket league in the world and ranks sixth among all sports leagues. The madness of cricket in people is like anything by looking into this, the main objective of our work is predicting the match result before the game starts based on the past statistic data that is present in the

form of data set. in this the study of Indian Premier League (IPL) is done using that past 12 seasons played till the date

in cricket and IPL, Data Science is used in a somewhat unique and interesting manner. Data analysis is very important in IPL to predict the match result.

You won't believe that IPL teams have started hiring proper companies who are experts in such Data Analysis. **Performance Analytics Companies** that analyze how good players are, and develop strategies for that players. These Data Analysis companies analyze data about players in detail to understand who is good at what aspect.

In one interview, **Virender Sehwag** encapsulated the importance of Data Science very nicely. He said that ***"Every game you play, they will record your good performance, your bad performance, you played against which bowler, you scored against which team and which bowler, and the whole data will easily show you that you are good against Pakistan but you're not good performed against Bangladesh, you're good against South Africa but you're not good against England. In 2003 when our computer analytics guy come in and he showed me videos and different kinds of data analysis, I got amazed!!"***

Machine Learning techniques are also used to predict the match results. Different models are created with the help of programming and computers in which, inputs like the position of a player, location of the match, the weather of the day, etc. are all added as variables and on the basis of previous matches, these models predict the future results of the matches. If you provide the data input of the previous matches, such as the venues of the matches as well as teams that played, players that were present as well as the type of players that were present, then in the future it could be predicted the result of the matches presently being played.

Obviously, it will not be 100% accurate but it could be quite useful.

II. LITERATURE REVIEW

In cricket, to predict an outcome of a match, the primary task is to extract out the essentials factors (features) which affect result of a match. Interesting works have been done in the field of predicting outcome in cricket.

Author [1] has analyzed the factors like home field advantage, winning the toss, game plan (first batting or first fielding) and the effect of Duckworth Lewis method [2] for one-day cricket format. Furthermore, Bailey and Clarke mention in their work [3] that in one-day cricket format, home ground advantage, past performances, venue, performance against the specific opposition, current form are statistically significant in predicting total runs and predicting the outcome of a match.

The software used for modelling is Anaconda and Python libraries like pandas, NumPy and IPython to work with the data structure and applying algorithms [3,4]. The main result obtained was based on the impact of toss winner and resultant match winner. The predictive model considered the innings score at regular intervals and the final scores to predict the match result. The model predicted score and run rate projected score were quite near to the final score, in particular the score predicted by the model was more accurate to the actual score. When no feature selection was applied to the dataset the model's accuracy was not satisfactory, i.e. slightly above 50%.

They considered the T20 International match data along with IPL data till 2015 as the training data set. In depth analysis was conducted by segmenting the data on the basis of venue, one team against all other teams, batting first and so on. Decision Tree was applied to predict the match outcome, and produced models with around 78% accuracy for the team that bats first and 75% when it bats second. IG technique was used for feature selection.

Authors [6] performed a comparative analysis of various regression and classification models in prediction of a football game and the results showed that classification-based models outperformed regression-based classification models. In cricket, Author [7, 8] proposed an adoption of h-index and PageRank to rank cricket teams and models to predict best suitable team line-up for a particular game using statistical modeling and network centrality-based approach. Authors [9–10] worked on identifying the role of external factors in the outcome of an ODI. The key features under consideration were home field advantage, winning the toss, game plan (batting first or fielding first), match type (day or day and night), competing team, current form etc. Author [11] used logistic regression technique to explore the statistical significance of various features and to build a model for result predictions in ODI.

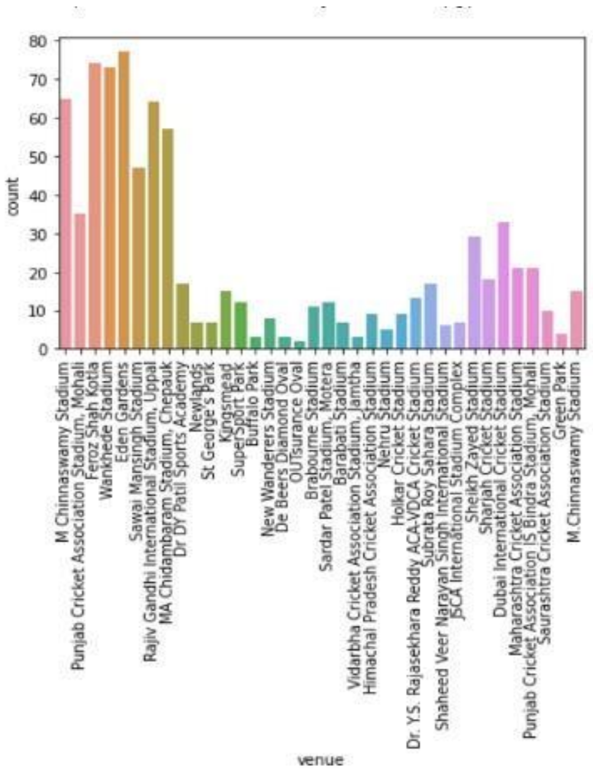
Live Cricket Score and Winning Prediction work [12] describes about the building of the model which predicts the score for the chasing team and will estimate the score of the second innings of match. The proposed work uses the concepts of Linear Regression, Naive Bayes Classifier and Reinforce Learning Algorithm. The factors such as toss result, ranking of the team, home team advantages are considered.

we are building a model for predicting the winning team based on the factors like home match, toss winning, performances in previous matches and team plan (batting or fielding).

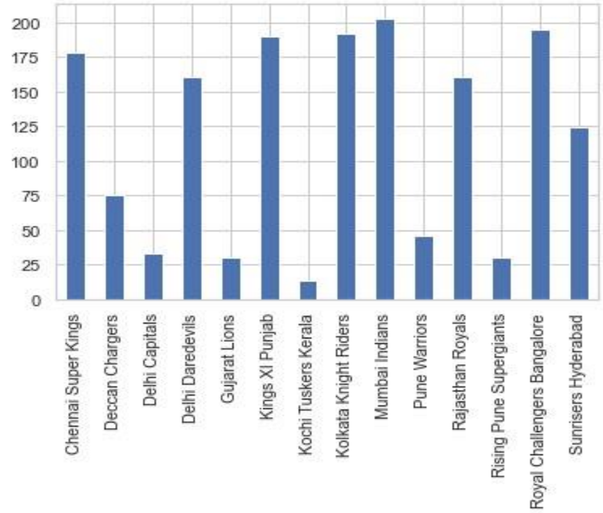
III. INITIAL INSIGHTS /DATA VISULAIZATION

The data of IPL matches collected from 2008 to till 2019.This was analysed and visualized. The dataset had records of matches covering almost 12 seasons. There are two datasets "matches.csv" and "deliveries.csv" are used to analyse and to visualize the data. The source of the data is https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020 .

We can notice the list of teams that played between the period 2008 to 2020. If you are a pro IPL fan then you will see some old team names on the list which are not playing these days but they contributed some valuable information in the IPL history.

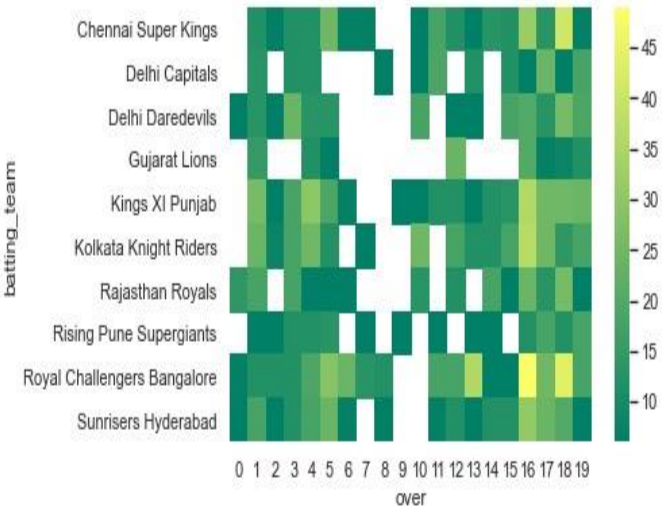


As you see that Eden Garden is the fan-favourite ground of IPL, nearly 80 matches are hosted there.



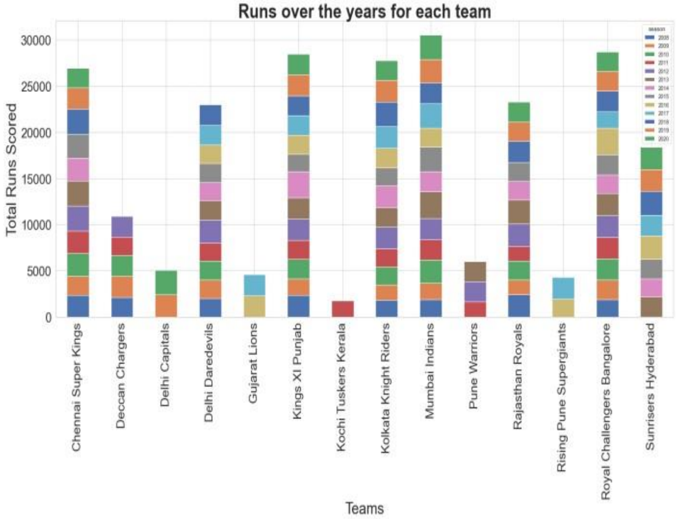
We count the value of each team playing in column one and add the count value of each team from team two to get the desired output. For example, if CSK played 90 times from team one and 85 times from team 2 then the total of175 matches are shown in the above graph. As you clearly see that Mumbai Indians played the highest number of matches in the IPL.

Over wise batting performance of each team in IPL (2008– 2020)



As you clearly see that if you are playing against MI or CSK then you have to play with your best bowling attack line-up from 1st over. MI's batsmen are silent in over 2 and 3 after that they went on rampage mode against their opponent. The same goes for CSK and RCB too. This data is not only helpful from the bowling team's perspective but also the batting team. If you are a team manager and you see using this data that your team is not performing well in the death overs then you probably focus to buy a good finisher in the next auction. As you see from the above heatmap the most of the team is lagging in finish the map, except CSK and MI. Which bowlers have performed well over last few seasons and can be considered further for upcoming auctions?

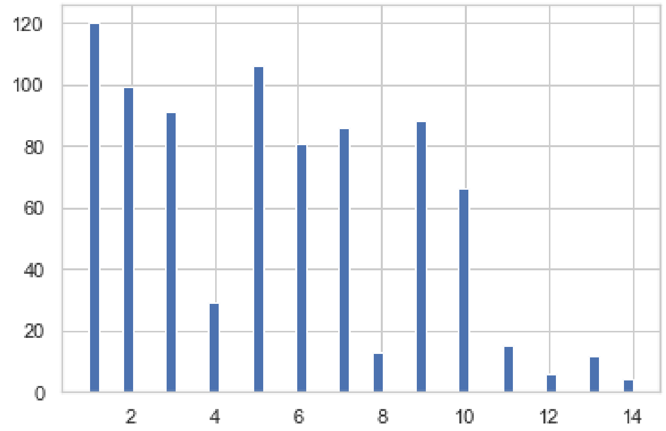
Runs scored by all the teams across seasons (2008-2019)



IV. MODEL BUILDING

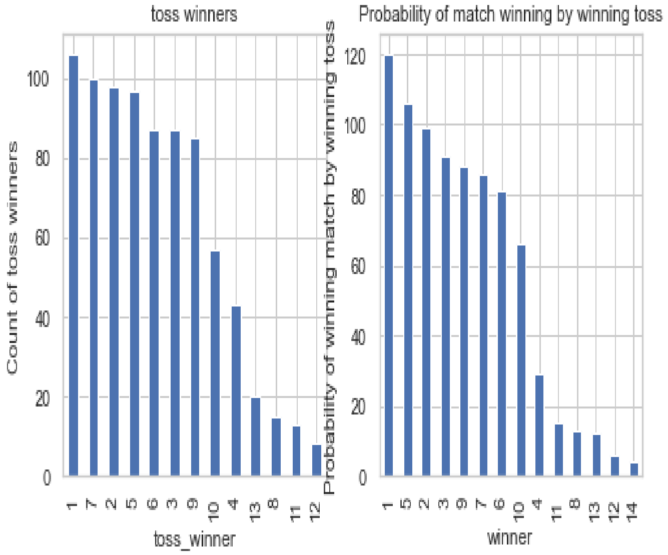
The data used for processing is the data of IPL matches collected from 2008 to till 2019. In two datasets we are only using "matches.csv" to predict the winning probability of the teams .the dataset contains the information of id, city , date , player of the match , venue, teams that are playing, toss winning team their decision, winner of the match and umpires . From the data we have calculated the number of matches won by the each teams and number of matches that are draw. There are total 13 teams participating in the IPL every year. The highest number of matches are won by the Mumbai Indians that is 120. Second highest is the Chennai super kings by winning 106 matches and the least is Kochi Tuskers Kerala by winning 6 matches. RCB as won 91 matches in the IPL which is in the 4th position after KKR(Kolkata knight Riders). For easy processing we have assigned the team names to numbers from 1 to 13 as there is 13 teams. 'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13 for further processing we are using these key values instead of team names.

Then we searched for the missing values. There were some null values in the city column. City is null, this is mainly for Dubai stadium. Hence update the City as Dubai and null for sharjah stadium which is also in dubai . we updated this values in the file. And we checked if there is any null vales but there were 0 null values in each column and Then we calculated some statistical values on the dataset.



We calculated the number of matches won by each team by plotting in a above graph.

Here we are building a model to predict the winning probabilities of the team based on the features like toss winning and venue of the match . we calculated the whether the team who won the toss also won the match for that we used the values like how many times the team as won the toss and the number of matches it as won and plotted the graph.



In the first graph the X- axis the team numbers while the y-axis gives the count of number of tosses won by them and in the second graph X-axis is same while Y-axis gives probability of each team to win the match by winning the toss. In both the graphs for the team number 1 that is Mumbai Indians won most toss and also most matches. In the most matches where they won the match toss their decision on fielding or batting choice or the game plan helped to win the matches.

Building model :Here we are building a predictive model for our dataset to predict the winning probabilities of the team based on the features like toss and venue. Did the winning of the toss and their decision on the game plan as an effect on the match winning and the place they played their match did have any influence on their match w by keeping all this factors we are constructing the model.

We used the classification model to get the accuracy and cross validation score of the different models we used here. We implemented this model using scikit learn module. The Scikit-learn Python library, initially released in 2007, is commonly used in solving machine learning and data science problems—from the beginning to the end. The versatile library offers an uncluttered, consistent, and efficient API and thorough online documentation. Classification models belong to the class of conditional models, that is, probabilistic models that specify the conditional probability distributions of the output variables given the inputs. The models we used for prediction are RandomForestClassifier and LogisticRegression() These were the models are imported from Sklearn module.

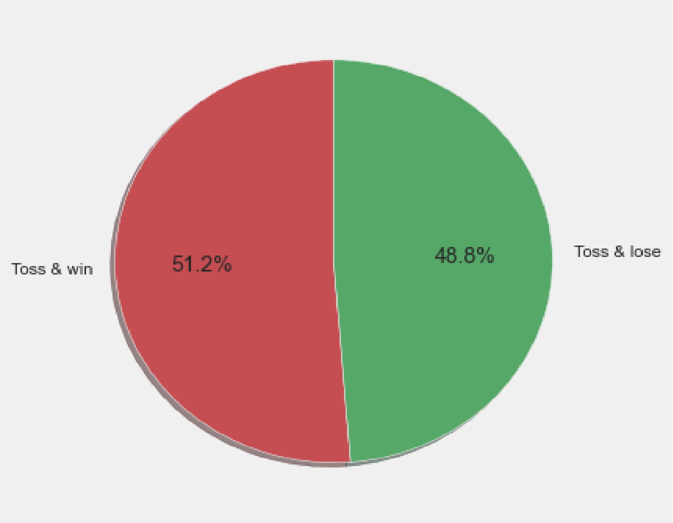
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. . In random forests each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features. The purpose of these two sources of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yield decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model. The model is easy to implement and has high flexibility.

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. This implementation can fit binary, One-vs-Rest, or multinomial logistic regression with optional ℓ1, ℓ2 or Elastic-Net regularization. The solvers implemented in the class logistic regression are "liblinear", "newton-cg", "lbfgs", "sag" and "saga". logistic regression implements Logistic Regression with built-in cross-validation support, to find the optimal c and l1_ratioparameters according to the scoring attribute.We used both logisticRegression and randomForestClassifier model and passed into classification model separately. For LogisticRegression model we got 24.020R% as accuracy and 22.672% as Cross-Validation Score. When we passed RandomForestClassifier we got 87.868% accuracy and 48.284% Cross-Validation Score. So we are using RandomForestClassifier model for our predictions.

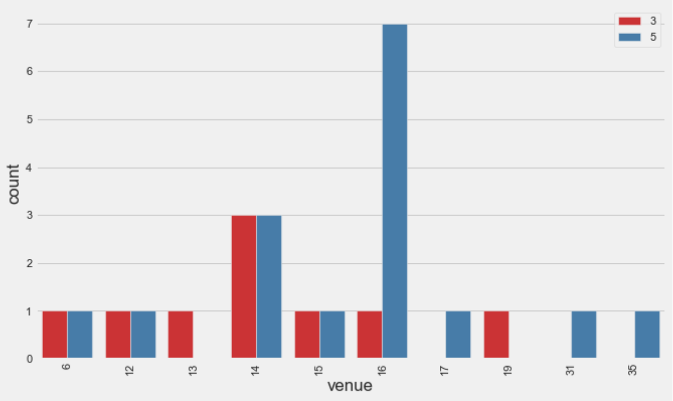
Evaluaton: After building the model we passed the parameters to the model such as who won the toss and the venue , city , toss decision to the model by selecting two teams at a time to predict who is most probable to win. When we passed RCB and KKR as two teams and the toss winner as RCB we got RCB as result and when we passed DC and DD as two teams playing with each other and toss winner as DC we got DD as result . In first case the team who won the toss as won the match and in the second case the team who lost the toss as won the match. When we calculated the feature importance from the model we got toss_winner as 0.191249,venue as 0.145619,city as 0.133680 and toss_decision as 0.071210 all of these doesn't have significant importance or does not influence the winning of the match at great extent. And both doesn't guarantee a match win. But from the analysis toss winning and venue of the match influence the match winning to small extent.

V. RESULTS

From the above prediction on features, we notice toss winner has least chances of winning matches. Toss winning does not gaurantee a match win from analysis of current stats and thus prediction feature gives less weightage to that.



The above graph shows the percentage of winning teams when they won the toss which is in red colour and the percentage of the losing teams when they won the toss. But from the analysis toss winning and venue of the match influence the match winning to small extent.From the analysis winning the task sometimes help the team to win the match and losing the toss also sometimes helps in winning.



In the above graph top 2 team analysis based on number of matches won against each other and how venue affects them is analysed.Previously we noticed that CSK won 79 , RCB won 70 matches now let us compare venue against a match between CSK and RCB we find that CSK has won most matches against RCB in MA Chidambaram Stadium, Chepauk, Chennai RCB has not won any match with CSK in stadiums St George's Park and Wankhede Stadium, but won matches with CSK in Kingsmead, New Wanderers Stadium.It does prove that chances of CSK winning is more in Chepauk stadium when played against RCB.Proves venue is important feature in predictability.

The Random Forest Regressor is unable to discover trends that would enable it in extrapolating values that fall outside the training set. When faced with such a scenario, the regressor assumes that the prediction will fall close to the maximum value in the training set. The model works well when the data has a non-linear trend and extrapolation outside the training data is not important.

The model fails when data is in time series form. Time series problems require identification of a growing or decreasing trend that a Random Forest Regressor will not be able to formulate. When the model fails we can use linear model such as SVM regression, Linear Regression, etc. we can Combine predictors using stacking. For example, you can create a stacking regressor using a Linear model and a Random Forest Regressor. Since Random Forest is a fully nonparametric predictive algorithm, it may not efficiently incorporate known relationships between the response and the predictors. The response values are the observed values Y1, . . . , Yn from the training data. RERFs are able to incorporate known relationships between the responses and the predictors which is another benefit of using Regression-Enhanced Random Forests for regression problems.

VI. CONCLUSIONS

In this project we tried to take into account the different factors like position of a player, location of the match, the weather of the day, etc. are all added as variables and on the basis of previous matches and possible trends in the performance of players individually and their contribution to the performance of a team as a whole. we tried to predict the probabilities of the influence of some features like toss winning and venue will help in winning the match. After building the model we found that the features doesn’t have significant importance or does not influence the winning of the match at great extent. And both doesn’t guarantee a match win. But from the analysis toss winning and venue of the match influence the match winning to small extent.From the model we found that Toss winning does not gaurantee a match win from analysis of current statitics and thus prediction feature gives less weightage to that. While venue proves the important feature in predictability for some teams.

VII. Contribution of each team member

Thrupthi M S : In phase one EDA part , code , reportand in phase 2 - code .

Anusha C : In phase one code, report and in phase 2 report.

Divya KC : report in phase I and 2.

V.REFERENCES

[1] A. Bandulasiri, “Predicting the winner in one day international cricket,” Journal of Mathematical Sciences & Mathematics Education, vol. 3, no. 1, pp. 6–17, 2008.

[2] F. C. Duckworth and A. J. Lewis, “A fair method for resetting the target in interrupted one-day cricket matches,” Journal of the Operational Research Society, vol. 49, no. 3, pp. 220–227, 1998

[3] M. Bailey and S. R. Clarke, “Predicting the match outcome in one day international cricket matches, while the game is in progress,” Journal of sports science & medicine, vol. 5, no. 4, p. 480, 2006

[4] W. McKinney, Python for data analysis: Data wrangling with Pandas, NumPy, and IPython, O’Reilly Media, Inc., 2012.

[5] Munir, F., Hasan, M.K., Ahmed, S., Md Quraish, S., 2015. Predicting a T20 cricket match result while the match is in progress (Doctoral dissertation, BRAC University).

[6] Delen, D., Cogdell, D., Kasap, N.: A comparative analysis of data mining methods in predicting NCAA bowl outcomes. Int. J. Forecast. 28(2), 543–552 (2012)

[7] Daud, A., et al.: Ranking cricket teams. Inf. Process. Manag. 51(2), 62–73 (2015)

[8] Saqlain, S.M., Usmani, R.S.A.: Comment on “ranking cricket teams”. Inf. Process. Manag. 53(2), 450–453 (2017)

[9] Bailey, M., Clarke, S.R.: Predicting the match outcome in one day international cricket matches, while the game is in progress. J. Sports Sci. Med. 5(4), 480 (2006)

[10] Morley, B., Thomas, D.: An investigation of home advantage and other factors affecting outcomes in English one-day cricket matches. J. Sports Sci. 23(3), 261–268 (2005)

[11] Bandulasiri, A.: Predicting the winner in one day international cricket. J. Math. Sci. Math. Educ. 3(1), 6 (2006)

[12] Rameshwari Lokhande and P.M. Chawan, “Live Cricket Score and Winning Prediction”, International Journal of Trend in Research and Development, Vol. 5, No. 1, pp. 3032, 2018.