e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

IPL CRICKET SCORE AND WINNING PREDICTION USING MACHINE LEARNING TECHNIQUES

Nikhil Dhonge[*1], Shraddha Dhole[*2], Nikita Wavre[*3], Mandar Pardakhe[*4], Amit Nagarale[*5]

[*1,2,3,4]Student, Department of Electrical Engineering,
MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India

[*5]Professor, Department of Electrical Engineering,
MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India

ABSTRACT

As cricket is the mostly played game. There are so many series are played in country one of them is Indian Premier League (IPL). Now it is conducted among 8 teams. In these papers the model has been proposed that has two methods the first one is prediction of score and the second one is team winning prediction. In these the score prediction includes linear regression, lasso regression and ridge regression whereas in winning prediction SVC classifier, decision tree classifier and random forest classifier are used. The model used the supervised machine learning algorithm to predict the winning. Random Forest Classifier used for good accuracy and the stable accuracy so that desired predicted output is accurate.

Keywords: Linear Regression, Ridge Regression, Naive Bayes, Random Forest Classifier, IPL Winning Prediction, IPL Score Prediction.

I. INTRODUCTION

Cricket was introduced to North America via the English colonies as early as the 17th century.as Cricket is most popular game. Most of the countries involved in it. In 2008 Indian Premier League, BCCI was established so many betting s were played on it like dream 11. So, there is huge demand for the algorithms that predicts the best result of score and winning team that is more important. Machine learning is the best way for prediction. All algorithms can be classified as reinforced, unsupervised, supervised learning. These algorithms used based on the application and the result achieved.

Problems of the supervised machine learning algorithms can be divided into problems of regression and classification. Output is the major problem with classification. output is a category, such as “green” or “pink” or “disease” and “no disease”. The major problem in the Regression, when real value is the desired output. Other common types of problems built on classification and Regression include recommendations and predictions for a series of time series. In unsupervised learning the purpose of unsupervised learning is to model the structure or distribution of data to learn more about data.

The algorithms used to predict the IPL first Inning Match Score are linear, lasso and ridge regression and for the IPL Match Winning Prediction, the classifier used here are SVC classifier, decision tree classifier and most important Random forest classifier. In the Linear Regression, labelled data is given to the machine learning model and the labelled data is already known. Linear regression used for the continuous values prediction than classification of the object. Mmulticollinearity in the data can be analyzed with the help of ridge regression. The Random forest algorithm plays an important role in winning prediction. Random forest classifier creates multiple decision trees and find out the individual output. It combines all the results together and give the results with more accuracy. It can be used as both classification and regression

II. LITERATURE SURVEY

The research paper of G.Sudhamathy helps to understand the different machine learning algorithms working principal and their implementation . It creates the Model and Training dataset and helps to predict with the help of the model created.

The model classifies the data and compares the results and get accuracy which is the important one [5].As in the dataset there are many parameters are present. Out of them which parameters are helpful in the project. The factors affecting concept was taken by Maheshwari in their prediction of live cricket score paper from that we get to know the main factors which required for the prediction of score and the prediction of winning team[6].[2]The role of classification is clarify in the paper of Tejinder Singh it gives proper information or use of naive bias and linear regression. They gives the proper knowledge of data collection and preparation also how to train the data

and test the data is given by them which is more helpful. The support vector machines brief idea is been taken from Aminul Islam Anik paper which is about players performance in this paper the idea about SVM system is given in detailed where the player performance prediction is given by collecting the old information or data[7]. From the literature survey it is concluded that the machine learning is need of prediction.

III. DATASET FEATURES

The approach over here we are using is ML based. So the basic requirements of an ML algorithm is dataset, training of that dataset using the algorithm and testing the model. So, we have imported dataset from Kaggle. Later on calculating the accuracy and improving the accuracy by using Random Forest classifier for winning prediction and Linear Regression for score prediction

Score Prediction:- For conducting our research, we collected data on all the IPL matches played in 2008. The dataset consists of 76015 numbers of rows. Dataset consists 15 columns over which we applied feature selection techniques and selected 8 features in which 7 are input feature and 1 is our target variable. The attributes selected were bat team, bowl team, overs, runs, wickets, runs in prev 5, wickets in previous 5 for score prediction.

Winner Prediction:- For conducting our research, we collected data on all the IPL matches played since 2008 till 2019. The dataset consists of 757 numbers of rows. Dataset consists 17 columns over which we applied feature selection techniques and selected 5 features in which 4 are input feature and 1 is our target variable. The attributes selected were team1, team2, winner, toss decision, toss winner for winning prediction. Each team was analyzed individually against every other team.

Table 1 : Dataset Attributes and their values

Attributes	**Values**
Batting Team	Batting Team Name among 8 teams in IPL
Bowling Team	Bowling Team Name among 8 teams in IPL
Overs	Value > 5 Over
Runs	0-300
Wickets	0-10
Run Scored in last 5 overs	0-300
Wickets fall in last 5 overs	0-10
Total Runs	0-300

Table 2: Dataset Attributes and their values

Attributes	**Values**
Team1	Team1 Name among 8 teams in IPL
Team2	Team2 Name among 8 teams in IPL
Toss Winner	Name of team
Toss Decision	“bat” and “field”
Winner	Winner team name

A. Feature Selection

Feature Selection is process in which we select an optimal set of features from input features set by using feature selection techniques. By removing redundant features, we reduce dimension of data and we can improve time and space complexity of data. Feature selection improves the performance of model and saving time and space.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

IV. BLOCK DIAGRAM AND METHODOLOGY

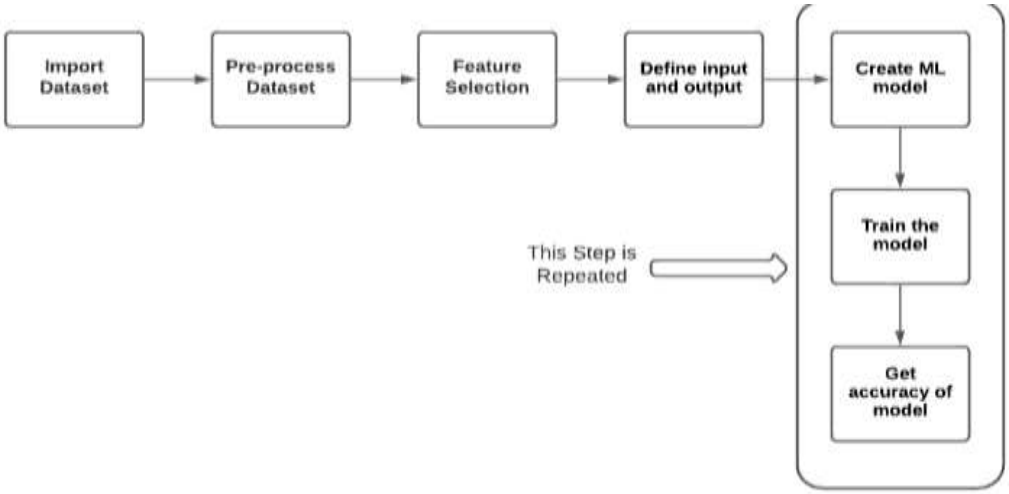


Figure 1: Block diagram

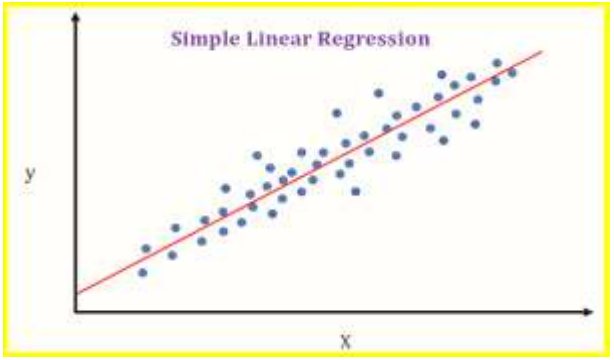
The first step is to import the dataset using the pandas library and then further preprocess the dataset by checking the null values and replace it with the mean or median values of the respective column. The categorical data in the columns is mapped into numerical values. After that the feature selection techniques are applied to the dataset and select only the optimal set of features. The set of input features (X) and the output (Y) are defined in the dataset. The input features are independent of each other and the output feature depends on the input features. Then the library is imported, and the ML Model is created, and the train-split-test method is used to separate the data into training data and testing data and then train the ML Model with training data and predict using test data. Then the accuracy of the model is calculated by simply taking the ratio of the predicted testing data and the actual testing data. This method is repeated with each ML Algorithm and the accuracy of each algorithm is calculated. Finally, the accuracy of the algorithms is compared and then which of the algorithms is the best for this dataset is determined.

V. ALGORITHMS

We tried to use two machine learning techniques: regression and classification. Selected algorithms from each technique were trained then.

Regression: Regression analysis uses various algorithm for the computation and based on that it predicts the continuous value. There are certain set of variables are used for the input and the continuous range value is the target variable. Based on the application different regression algorithms are used. There are different regression techniques. Out of which the linear, ridge and lasso algorithms are used for predicting the score .

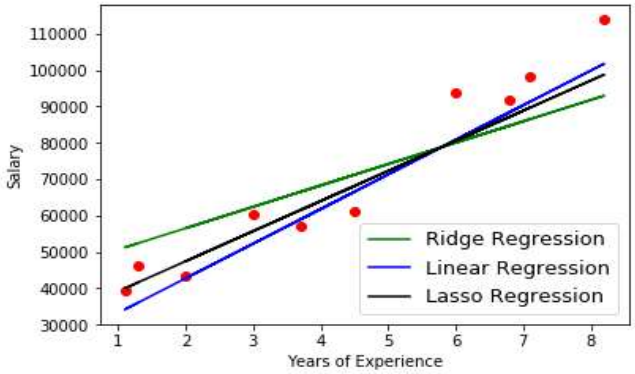
Linear regression: To predict the continuous values, Linear regression is used. Certain known parameters are given to the machine learning algorithms, it predicts the continuous values as output. It cannot used for the classification problems. The proposed model predicts the score using the Linear Regression.



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

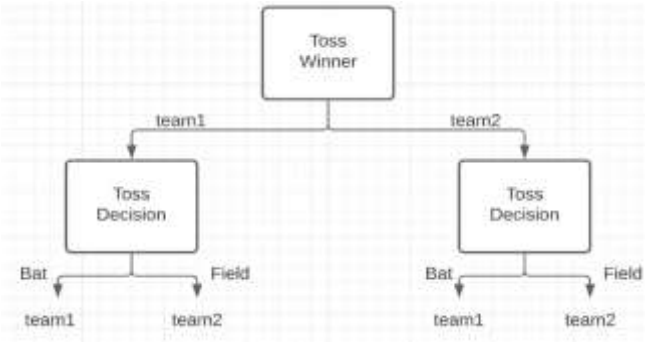
Ridge regression: Ridge regression is also used to predict the continuous values. When the variables used for the prediction greater than the observations of when multicollinearity present in the data, ridge regression is used set has multicollinearity (correlations between predictor variables).

Lasso regression: Lasso regression is a type of linear regression that used for predicting the continuous values. Shrinkage is used in the lasso regression. When data values focus towards central point shrinkage occurs. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models.

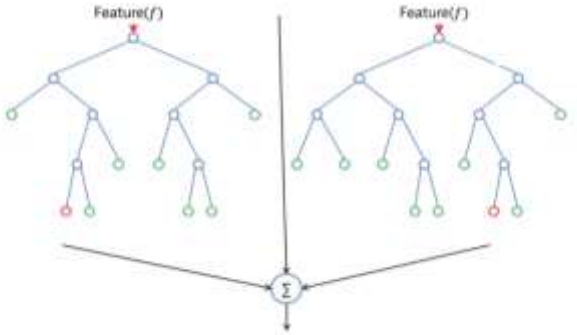


Classification: Classification is used, when the target variable represents particular category. Proposed system used classification for the winning prediction of the IPL matches such as "Winner" or "Looser". Winning prediction is the binary classification.

Decision tree: It is supervised machine learning algorithm. It is used for classification problem. It splits the data based on the entropy and information gain. It looks like flow chart. It creates the various categories within the categories. It splits the data of the high information gain first. It created a "Toss Winner" as root node and split the data further.

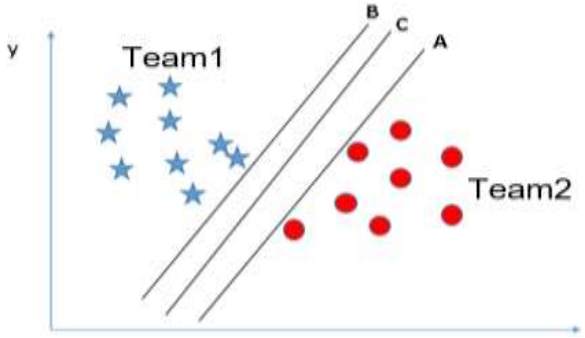


Random forest: Random forest classifier creates multiple decision trees and find out the individual output. It combines all the results together and give the results with more accuracy. It can be used as both classification and regression.



SVM: we create a graph of the space equal to number of features present and plot the data into it. Each value showed using dot at each particular coordinate. Classification is done by separating each category with the help of the Hyperplane. It creates support vectors at each side to increase accuracy.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com



VI. COMPARATIVE ANALYSIS OF ALGORITHMS

A. Score Prediction algorithms

It is found that for the score prediction the linear regression is giving the more accuracy as compared to Ridge regression and Lasso regression

Table 1: Accuracy of the Score Prediction Models

Algorithm	Accuracy
Linear regression	80.92
Ridge regression	80.84
Lasso regression	80.45

For the score prediction the linear regression gives the highest accuracy result as we see. So, the formula of linear regression for getting theoretical result is as follows.

y = Bo + B1 * x ... (linear regression equation)

So here, y is the dependent variable

x is independent variable

B0 is bias coefficient &

B1 is coefficient of x..

Thus, we are using Cost function which helps to get the most accurate possible values for B0 and B1. So, as we need the best values for B0 and B1 we have converted it into minimization problem where it minimizes the errors between the predicted score and actual score.

B. Winning Prediction algorithms

In case of winning prediction of the random forest algorithm has the highest accuracy among SVC and decision tree. Random forest is giving the best result with 90%, 80%, 75%, 70% variable dataset

Table 2: Accuracy of the Winning Prediction Models

Algorithms	Accuracy(%) (with 90 % training data)	Accuracy(%) (with 80 % training data)	Accuracy(%) (with 75 % training data)	Accuracy(%) (with 70 % training data)
SVC	43	54	55	52
Decision Tree	61	57	55	55
Random Forest	76	70	72	74

Splitting of the data by the decision tree can be decided using Entropy. the data. entropy is used as a way to measure how “mixed” a column is. Specifically, entropy is used to measure disorder.

$$E(S) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

The main key is the Information gain which used by Decision Tree Algorithms to create Decision Tree. Decision Trees algorithm will always try to maximize Information gain. An attribute with highest Information gain will tested/split first.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} Entropy(S_v) \frac{|S_v|}{|S|}$$

VII. IMPLEMENTATION OF THE GUI

The Graphical User Interface is developed for the machine learning models using the Flask Framework. For the backend of the site Python is used. The site can be used to predict the IPL match score with the help of last 5 overs of the data. We can also predict the Winner of the match with the data of just Toss Winner and Toss Decision.

All the input information necessary for the model for the prediction is provided to the model. The calculation is not stored in the system because all calculations computed at real time. We implemented it that way as we can add change more attribute to the system with minor changes to the program.

A. Score Prediction

The GUI required at least 5 overs of the data to predict the score as shown in the Fig 1.1. Model require the input data of Batting team, Bowling team, Over, Runs, Wickets, Run Scored in last 5 overs, Wickets fall in last 5 overs to predict the score of the match as shown in the Fig 1.2.

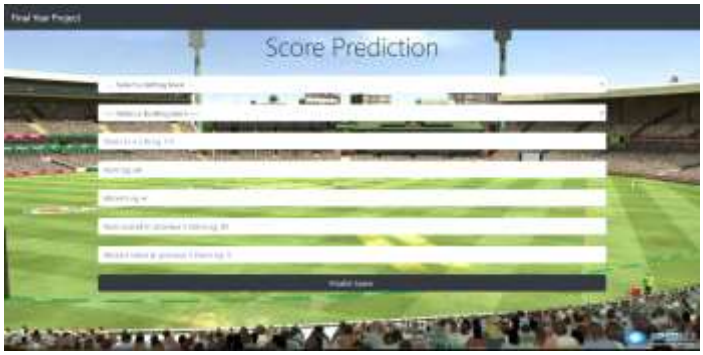


Fig 1.1: Score Prediction Model UI

Fig 1.2: Input to Score Prediction Model

The output we got from the model is not exact predicted output. So, to increase the accuracy of the model we add and subtract 3 to give the range of score as shown in Fig 1.3. So, our model works in majority of the cases.

Fig 1.3: Score Prediction result

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

B. Winning Prediction

The GUI required team1, team2, Toss Winner, Toss Decision data to predict the Winner of the IPL match as shown in the Fig2.1. Model require the input data as shown in the Fig2.2.

Fig 2.1: Winning Prediction UI

Fig 2.2: Input to Winning Prediction Model

The output we got from the model is shown in the Fig 2.3.

Fig 2.3: Winning Prediction result

VIII. CONCLUSION

This paper will give the important information regarding IPL score prediction and winning prediction system, that which parameters are required also the classifiers and algorithms. it helps in mathematical operation. Using all the information we have developed a website. for that the important work we have to do for the model is comparative analysis of machine learning techniques that is for score prediction the regressions and for winning prediction the analysis of classifiers. In Score Prediction analysis accuracy of Linear Regression is more than Ridge and Lasso Regression and in winning prediction analysis among SVC, Decision tree classifier and Random forest classifier, we got Random forest classifier accuracy more than other 2, with all 90%, 80%, 75%, 70% training data

ACKNOWLEDGEMENTS

We would like to thank Dr Rushikesh Borse, professor, MIT Academy of Engineering, Alandi for motivating and guiding us for this paper presentation.

IX. REFERENCE

[1] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, India, 2015, pp. 60-66, doi: 10.1109/ICSCTI.2015.7489605.

[2] J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.

[3] A. Kaluarachchi and S. V. Aparna, "CricAI: A classification based tool to predict the outcome in ODI cricket," 2010 Fifth International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 2010, pp. 250-255, doi: 10.1109/ICIAFS.2010.5715668.

[4] A. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 500-505, doi: 10.1109/CEEICT.2018.8628118.

[5] N. Rodrigues, N. Sequeira, S. Rodrigues and V. Shrivastava, "Cricket Squad Analysis Using Multiple Random Forest Regression," 2019 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India, 2019, pp. 104-108, doi: 10.1109/ICAIT47043.2019.8987367.

[6] M. Jhawar and V. Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Riva del Garda, 2016

[7] A. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 500-505, doi: 10.1109/CEEICT.2018.8628118.

[8] Rameshwari Lokhande, P. M. Chawan "Live Cricket Score and Winning Prediction" **Published** in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-5 | Issue-1 , February 2018, URL: http://www.ijtrd.com/papers/IJTRD12180.pdf

[9] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.

[10] A. Basit, M. B. Alvi, F. H. Jaskani, M. Alvi, K. H. Memon and R. A. Shah, "ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-6, doi: 10.1109/INMIC50486.2020.9318077.

[11] A. Bandulasiri, "Predicting the Winner in One Day International Cricket", Journal of Mathematical Sciences & Mathematics Education, Vol. 3, No. 1.

[12] Analysis and Prediction of Cricket Statistics using Data Mining Techniques Anurag Gangal VESIT, Mumbai Abhishek Talnikar VESIT, Mumbai Aneesh Dalvi VESIT, Mumbai Vidya Zope VESIT, Mumbai Aadesh Kulkarni VESIT, Mumbai.

[13] S. Agrawal, S. P. Singh and J. K. Sharma, "Predicting Results of Indian Premier League T-20 Matches using Machine Learning," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 67-71, doi: 10.1109/CSNT.2018.8820235.