e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

A COMPARATIVE STUDY OF MACHINE LEARNING MODELS IN IPL SCORE PREDICTION

Ann Thomas*[1], Dr. Jibrael Jos*[2]

*[1,2]Department of Data Science, Christ (Deemed to be University), Lavasa Campus Pune,
Maharashtra, India.

ABSTRACT

Cricket is a popular sport around the globe. With a perpetual increase in the popularity and advertising associated with it, forecasting the IPL matches is becoming a need for the advertisers and the sponsors. IPL has quickly risen to become cricket's most profitable competition. In a cricket match, the scoreline frequently depicts the team's chances of victory based on the present match circumstances. Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. This proposed paper is specifically concentrating on enactment and measuring the difference between the models to foretell the score of an IPL match. This work focuses on finding meaningful information about the IPL Teams by using python. The dataset is loaded and a set of pre-processing is done followed by feature selection. Four machine learning algorithms Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regressor are applied and the loss functions are compared to measure the errors, mean squared error, mean absolute error and root mean square error. The linear Regression model gives the lowest error and thus turns out to be the best model compared to the other four.

Keywords: Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor.

I. INTRODUCTION

Cricket is a prominent sport in India, and is played almost everywhere in the country. Indian Premiere League is one of most popular T20 cricket league in the world. 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. The major formats in which cricket is being played internationally, One Day Internationals (ODIs) and the T20 cricket and Test Matches Ever since its inception in 2007, IPL has been a huge success and has become an industry with investment of billion-dollars. 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.

II. EXISTING WORK

Vistro et al. [1] published the paper The Cricket Winner Prediction with Application of Machine Learning and Data Analytics on 2019. The data taken was the historical data of IPL from season 2008 to 2017. This research paper is about prediction of an IPL match winner before the match started. Predicting the winner of a cricket match depends on factors like batsman's performances, team strengths, venues and weather conditions. In this research various features have been analysed to predict the match winner of the game. The SAS Institute created the SEMMA process, which follows a five-stage cycle: Sample, Investigate, Modify, Model, and Evaluate. The XGBoost machine learning model gave the best accuracy of 94.23 % without tuning of parameters. The Decision Tree model predicted an accuracy of 76.9%, after fine-tuning of the parameters model performance was enhanced by 76 % to 94%. Random Forest model predicted the winner with 71% accuracy, parameter's tuning resulted with 80 % accuracy.

Pandey et al. [2] published the paper Predicting Players Performance in One day International Cricket matches using Machine Learning was published in the year 2018 and the batting matches played from January 14, 2005 to July 10, 2017 bowling - matches played from January 2, 2000 to July 10, 2017. This paper attempts to predict the performance of players as how many runs will each batsman score and how many wickets will each bowler take for both the teams. Number of runs and number of wickets are classified in different ranges as they are targeted as classification problems. Naïve bayes, random forest, multiclass SVM and decision tree classifiers are used to generate the prediction models for both the problems. Random Forest produces the best accurate prediction models for both batting and bowling.

Kapadia et al. [3] published the paper Sport analytics for cricket game results using machine learning on 2019. The goal of this study is to compare and contrast several machine learning algorithms for forecasting the outcome of IPL cricket matches. Intelligent models are created that forecast the outcome of a match based on the impact of the home ground and the toss winner. The winning team considers weather, pitch, and outfield conditions while deciding whether to bat or field first in order to gain a strategic edge. The impact of home ground is depicted in one model, while the effect of toss decision is considered in the other. Influential features of the dataset have been identified using filter-based methods including Correlation-based Feature Selection, Information Gain (IG), Relief and Wrapper. On the Toss featured subset, none of the considered machine learning algorithms performed well in producing accurate predictive models. Machine learning techniques including Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN) and Model Trees (classification via regression) have been adopted to generate predictive models from distinctive feature sets derived by the filter-based methods. Random Forest performed better in terms of accuracy, precision and recall metrics when compared to probabilistic and statistical models.

Passi et al. [4] published paper Increased Prediction accuracy in the game of cricket using machine learning on 2018. The data of batting matches played was collected from January 14,2005 to July 10,2017 and the bowling matches data was from January 2, 2000 to July 10, 2017.Any sport's most crucial responsibility is player selection, and cricket is no exception. This paper tries to forecast player performance, such as how many runs each batsman will score and how many wickets each bowler will take for both teams. Both issues are characterized as classification problems, with the number of runs and wickets falling into distinct ranges. Predicts the performance of players separately. Naïve bayes, random forest, multiclass SVM and decision tree classifiers are used to generate the prediction models for both the problems. Random Forest classifier was found to be the most accurate for both the problems. With an accuracy of 90.74% for predicting runs scored by a batsman and 92.25% for predicting wickets taken by a bowler.

Dhonge et al. [5] published paper IPL cricket score and winning prediction using machine learning technique on 2021.The data of IPL matches collected from 2008 till 2019. In this paper the model has two methods the first one is prediction of score and the second one is team winning prediction. The score prediction includes linear regression, lasso regression and ridge regression. In Score Prediction analysis accuracy of Linear Regression is more than Ridge and Lasso Regression. For winning prediction SVC classifier, decision tree classifier and random forest classifier are used. Random Forest Classifier gives good accuracy. Linear regression outperforms Ridge and Lasso regression in Score. Prediction analysis, and Random Forest classifier outperforms SVC, Decision tree classifier, and Random Forest classifier in Winning Prediction analysis, with all 90 percent, 80 percent, 75 percent, and 70 percent training data.

PriyaIyer et al. [6] published paper Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms on 2020.This paper has intended on analyzing the results of the IPL match during the year 2008-2019 by applying the data mining algorithms for existing data, and predicted the new data for the year 2020.In this paper, Prediction of IPL2020 is done on the basis of survey, and analysis are done based on data mining algorithms. Logistic Regression, SVM are the different types of algorithms used for model building to predict the outcome of an ODI match considering all the external features like home-field advantages, winning the toss, game plan, venue and season. SVM was proved to be a better model. Naïve bayes gives an accuracy of 35.91% for the bowler who takes the greatest number of wickets. Random Forrest gives an accuracy of 82.73% for the most favourite team. It focuses on analyzing the results of IPL matches using existing data mining algorithms on both balanced and skewed datasets.

Maginmani et al. [7] published paper Predicting accuracy of players in the Cricket using Machine Learning on 2020. This paper predicts the performance of the players. Here the number of runs and number of wickets are classified into dissimilar range using different classifier algorithm. Decision tree, Naive Bayes, Random Forest and Multiclass SVM classifiers to Predict for these 2 problems. The player's performance is influenced by a variety of elements, including the venue where the game is being played, previous records, present form, average rate, strike rate, runs scored at a certain venue, number of innings played against opposing teams, and so on. The two problems are included in the prediction model as a goal as a classification problem, where the number of "runs" and "wickets" are classified into dissimilar ranges using different classifier algorithms. The classifiers 'Decision tree,' 'Naive Bayes,' 'Random Forest,' and 'Multiclass SVM' to create an effective model for Predicting these two problems. Random forest classifier produces more accurate results than the other three classifier algorithms, whereas SVM produces the least helpful results.

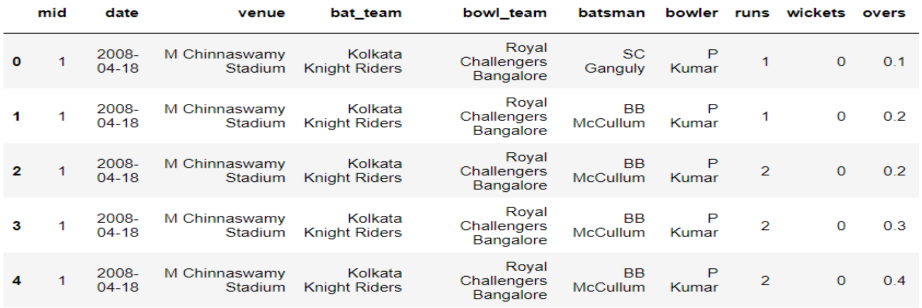
Research paper [1] published in 2019 is about prediction of an IPL match winner before the match started. The XGBoost machine learning model gave the best accuracy of 94.23 %. The paper [2], [4] which is published in the year 2018 and [7] in the year 2020, attempts to predict the performance of players as how many runs will each batsman score and how many wickets will each bowler take for both the teams. Random Forest classifier produces the best accurate prediction models for both batting and bowling in all the three papers. Paper [4] gives an accuracy of 90.74% for predicting runs scored by a batsman and 92.25% for predicting wickets taken by a bowler. In paper [5] published in 2021, the model has two methods first one is prediction of score and the second one is team winning prediction. Random Forest Classifier gives the best accuracy in winning team prediction. Linear regression outperforms Ridge and Lasso regression in Score. Paper [3] published in 2019 is a comparative study of several machine learning algorithms to forecast the outcome of a match based on the impact of the home ground and the toss winner. Paper [6] published in the year 2020 analyze the results of the IPL match by applying the data mining algorithms for existing data, and predicted the new data for the next year. Prediction is done on the basis of survey, and analysis are done based on data mining algorithms.

Paper Title	**Author(s)**	**Machine Learning Technique**	**Year**
[1]	Vistro et al.	Random Forest, Decision Tree and XGBoost	2019
[2]	Pandey et al.	Naïve bayes, random forest, multiclass SVM and decision tree classifiers	2018
[3]	Kapadia et al.	Naïve Bayes, Random Forest, K-Nearest Neighbour (KNN) and Model Trees (classification via regression)	2019
[4]	Passi et al.	Naïvebayes, random forest, multiclass SVM and decision tree classifiers	2018
[5]	Dhonge et al.	SVC classifier, decision tree classifier, random forest classifier, linear regression, lasso regression and ridge regression	2021
[6]	PriyaIyer et al.	Logistic Regression, SVM and Naïve Bayes	2020
[7]	Maginmani et al.	Decision tree, Naive Bayes, Random Forest and Multiclass SVM classifiers	2020

Different Machine Learning approaches utilized in the papers are listed in the table above along with the year of publishing.

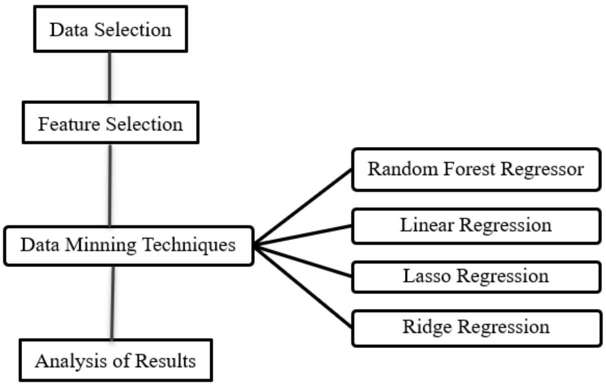
III. DATASET

The dataset that we are considering is the historical IPL data that was collected from Kaggle. The data contains all the details regarding the score of the first innings.



	mid	date	venue	bat_team	bowl_team	batsman	bowler	runs	wickets	overs
0	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	P Kumar	1	0	0.1
1	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	1	0	0.2
2	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.2
3	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.3
4	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.4

IV. MODEL ARCHITECTURE



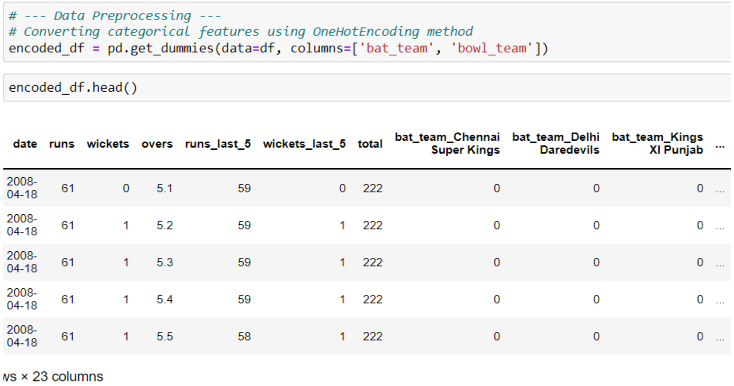
Data selection is choosing the right data for the process. After choosing the apt dataset we have to work on the dataset before building the model. Variable selection or attribute selection are other terms for feature selection. It's the process of automatically selecting the attributes in the data that are most important to the predictive modelling task at hand. Feature selection methods keep the attributes present in the data while including and excluding them and creating a precise predictive model. We used a simple correlation plot using seaborn library to investigates which all features are varying with target column in other ways we are looking into the correlation between variables. If two variables have a linear relationship, correlation can tell you how strong that relationship is. This makes sense as a beginning place because we're usually seeking for connections, and correlation is a rapid method to grasp the data set we're dealing with. By feature selection method we came to the conclusion of choosing most relevant variables.

The prediction was done by building different Machine Learning models that include Linear Regression, Lasso Regression, Ridge Regression and Random Forest Regressor. Linear Regression reveals a linear relationship, it can tell you how the dependent variable's value changes as the independent variable's value changes. Lasso regression is a regularization approach for lowering model complexity. It's identical to the Ridge Regression, only the penalty term only contains absolute weights instead of a square of weights. It can lower the slope to zero because it uses absolute data. Ridge regression is a classification approach that works in part because unbiased estimators are not required. Ridge regression reduces the residual sum of squares of predictors in a given model to the smallest possible value. The fundamental idea of a Random Forest Regressor is to aggregate several decision trees to decide the final outcome rather than relying on individual decision trees.

A. Model Building and Training

We have two sorts of data in our system: training and testing. The training data set consists of a large number of training samples, each with an input and output vector. We quantify each sample into a vector of integer variables using feature extraction methods.

1. Data Encoding: As a first step in building the Model, the dataset is pre-processed and cleared of all null values. The feature extraction is done in such a way that only the relevent features are selected.
2. Training and Test Set: The dataset is split into train and test in a ratio of 77:33. The training data set is made up of several training samples. The testing data set consists of several testing samples.
3. Formatting data: The data points are entered and formatted using excel for analysis and model building the formatted data is converted to comma separated values. Machine learning models can understand categorical values and majority of our columns are categorical in nature so it is important for us to ensure to convert categorical or string values to numerical to ensure better performance from models. Here we used one hot encoding technique so that the as shown in the below figure to ensure that our data is interpretable by models.



```
# --- Data Preprocessing ---
# Converting categorical features using OneHotEncoding method
encoded_df = pd.get_dummies(data=df, columns=['bat_team', 'bowl_team'])
```

```
encoded_df.head()
```

date	runs	wickets	overs	runs_last_5	wickets_last_5	total	bat_team_Chennai Super Kings	bat_team_Delhi Daredevils	bat_team_Kings XI Punjab	...
2008-04-18	61	0	5.1	59	0	222	0	0	0	...
2008-04-18	61	1	5.2	59	1	222	0	0	0	...
2008-04-18	61	1	5.3	59	1	222	0	0	0	...
2008-04-18	61	1	5.4	59	1	222	0	0	0	...
2008-04-18	61	1	5.5	58	1	222	0	0	0	...

ws × 23 columns

Cleaning the Data: Data cleaning is the process of identifying, deleting, and/or replacing inconsistent or incorrect information from the database. This technique ensures high quality of processed data and minimizes the risk of wrong or inaccurate conclusions. Delete all Formatting. Are various steps involved in cleaning, our data was made in a consistent and proper way in order to avoid these steps.

- Get Rid of Extra Spaces.
- Select and Treat All Blank Cells.
- Convert categorical into Numbers.
- Remove Duplicates.
- Highlight Errors.
- Organizing the date column

Are various steps involved in cleaning, our data was made in a consistent and proper way in order to avoid this step.

Model Building: The last step is to build the Model for implementation. Here 4 machine learning models, Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor were used for comparison. All the models were compared and the one with more accurate result was chosen.

B. Model Validation

The process of validating that a model accomplishes its intended goal is known as model validation. Comparison of model predictions and coefficients with theory, as well as the collecting of additional data to check model predictions, are all methods for determining the validity of regression models. Data splitting or cross-validation, in which a portion of the data is used to estimate the model coefficients and the remaining data is used to test the model's prediction accuracy, and comparison of findings with theoretical model calculations. When collecting new data to test the model is impractical, data splitting is found to be an efficient approach of model validation.

C. Model Scoring

After training the data, the model was tested on unseen data. When forecasting a numeric value such as a height or a dollar amount, we don't want to know if the model predicted the value exactly (this may be impossible in practice); instead, we want to know how near the predictions were to the expected values. Error is a metric that measures how close forecasts were to their predicted values on average. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are three error metrics that are often used to evaluate and assess the performance of a regression model (MAE).

1. Root Mean Square Error : Residuals are a measure of how distant the data points are from the regression line; the RMSE is a measure of how spread out these residuals are. In other words, it shows how closely the data is clustered around the best-fit line. Root means the square error is often used to evaluate experimental results in climatology, forecasting, and regression research.
2. Mean Absolute Error: The level of mistake in your measurements is known as absolute error. It's the distinction between what's measured and what's "true".

D. Predictive model

The models are from sklearn library which is a Python machine learning package. Linear Regression, Lasso Regression, Ridge Regression and Random Forest Regressor are among the regression algorithms included, and it's built to work with Python's NumPy and SciPy libraries.

A method of modelling the relationship between one or more variables is linear regression. The model's ability to predict outputs for inputs it has never seen before. The goal of LR is to locate the line that best fits the data points on the plot so that we can roughly forecast where the prediction will land. Linear Regression works by attempting to determine the weights that lead to the best-fitting line for the given input data. The best-fitting line is chosen based on the lowest cost.

Ridge Regularization works by penalizing the model for higher slope values, therefore decreasing the variance between different samples from the dataset.

Lasso regression penalizes less important features of your dataset and makes their respective coefficients zero, thereby eliminating them. Thus, it provides you with the benefit of feature selection and simple model creation. So, for dataset with high dimensionality and high correlation, lasso regression can be used.

A Random Forest is an ensemble technique that solves regression and classification problems by combining many decision trees using a technique called Bootstrap and Aggregation, also known as bagging. Random Forest is a fundamental learning paradigm that employs many decision trees. The dataset is sampled at random for row and feature sampling, resulting in sample datasets for each model. This section is called Bootstrap. Every decision tree has a high variance, but when we combine all of them in parallel, the total variance is low since each decision tree is perfectly trained on that specific sample data, and so the outcome is based on multiple decision trees rather than one. The majority voting classifier is used to generate the final output. The ultimate output in a regression problem is the average of all the outputs. Aggregation is the name of this section.

V. COMPARATIVE STUDY RESULTS

Performance metrics for Regression problems are MSE, MAE and R2. Mean Absolute Error (MAE) is the simplest error metric used in regression problems It's essentially the total of the absolute differences between the expected and actual values multiplied by the average. In other words, we can use MAE to determine how far the forecasts were off the mark. MSE is like the MAE, but the only difference is that, instead of using the absolute value, it squares the difference between actual and expected output values before summing them up. Here Linear Regression model gives better result than all the other models as it gives the lowest MSE and MAE values.

Model name	**MAE**	**MSE**	**RMSE**
Linear Regression	12.11	251.01	15.84
Lasso Regression	12.21	262.37	16.19
Ridge Regression	12.12	251.03	15.84
Random Forest Regressor	13.69	332.72	18.24

Table1.2 MAE, MSE and RMSE tells how well each model performed in predicting the score. The performance of all four models in predicting the score of the IPL match. For the given dataset Linear Regression gives the best accuracy as it offers the lowest result.

VI. CONCLUSION

The goal of this research is to use historical data to forecast the score of IPL match. As the popularity of the IPL and the advertising associated with it grows, advertisers and sponsors will need to forecast IPL matches. The Indian Premier League season of 2015 contributed $11.5 billion to the country's GDP. The IPL 2020 set a new viewership record with 31.57 million average impressions and a 23 percent increase in overall consumption over the previous season. The projections attract the audience's interest, which leads to increasing economic growth in India. In this system the past data is taken into consideration and the data is split in to train and test on the basis of year. The data till 2016 is taken as training data and the details of 2017 to test. Four models were built and the results were compared, Linear Regression gave the lowest value for MAE, MSE and RMSE.

Here the past data is taken into consideration and the data is split in to train and test on the basis of year. The data till 2016 is taken as training data and the details of 2017 to test. Four models Linear Regression, Lasso Regression, Ridge Regression and Random Forest Regressor were built and the results were compared. Linear Regression gave the lowest value for MAE, MSE and RMSE, hence concluded as the model with best result.

VII. FUTURE WORK

In future we can collect more data and add the details to predict the score of the upcoming IPL matches. Other machine learning models could be created. More data related to teams and players recent performance could make it more effective. ICC rankings and team composition could be included.

VIII. REFERENCES

[1] Kalpdrum Passi and Niravkumar Pandey, “Predicting players' performance in one day international cricket matches using machine learning” (2018)

[2] Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David ,“The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics”(2019)

[3] Kumash Kapadia, Hussein Abdel-Jaber, Fadi Thabtah, Wael Hadi, “Sport analytics for cricket game results using machine learning: An experimental study”(2019)

[4] Kalpdrum Passi and Niravkumar Pandey “Increased prediction accuracy in the game of cricketusing machine learning”(2018)

[5] Nikhil Dhonge, Shraddha Dhole, Nikita Wavre, Mandar Pardakhe, Amit Nagarale, “ipl cricket score and winning prediction using machine learning technique” (2021)

[6] Priyanka S, Vysali K, Dr K B PriyaIyer, “Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms”, (2020)

[7] Mr. Mujamil Dakhani, Umme Habiba Maginmani, “Predicting accuracy of players in the Cricket using machine learning” (2020)

[8] Abdul Basit, Muhammad Bux Alvi, Fawwad Hassan Jaskani, Majdah Alvi , “ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques”(2020)

[9] Mr. Suyash Mahajan, Ms. Gunjan Kandhari, Ms. Salma Shaikh, Ms. Rutuja Pawar, Mr. Jash Vora, Ms. A. R. Deshpande. “Cricket Analytics and Predictor”(2019)

[10] Stylianos Kampakis, William Thomas , “Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches” (2015)

[11] Sohail Akhtar, Philip Scarf and Zahid Rasool, “Rating players in test match cricket”(2014)

[12] Madan Gopal Jhanwar and Vikram Pudi, “Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach” (2015)

[13] H.V Ramachandra, R.R.Kamble, Nidhi Koul, Kaustubh Adhav, Akshay Dixit, RutujaPakhare, “Predicting Cricket Score By Using Machine Learning Concepts”(2020)

[14] Rabindra Lamsal and Ayesha Choudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”(2020)

[15] Pallavi Tekade, Kunal Markad, Aniket Amage, Bhagwat Natekar, ”Cricket match outcome prediction using machine learning”(2020)

[16] B V S Sai Praneeth, V Srighan Reddy, P Jayanth, K Jeevan Reddy, “Cricket analysis using machine learning”(2021)

[17] Vipul Punjabi, Rohit Chaudhari, Devendra Pal, Kunal Nhavi, Nikhil Shimpi, Harshal Joshi, “A survey on team selection in game of cricket using machine learning”(2019)

[18] Nandkishor Patil , Dilip Dalgade, “Cricket prediction using random forest regression”(2021)

[19] Waqar Ahamed, “A Multivariate Data Mining Approach to Predict Match Outcome in One-Day International Cricket”(2015)

[20] Jayalath, Kalanka P, “A machine learning approach to analyze ODI cricket predictors”(2018)

[21] Amal Chaminda Kaluarachchi, Aparna S Vard, “A classification based tool to predict the outcome in ODI cricket”(2010)

[22] C. Deep Prakash, C. Patvardhan, Sushobhit Singh, “A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers”(2016)

[23] Manage, Ananda B.W, Kafle, Ram C, Wijekularathna, Danush K, “Classification of all-rounders in limited over a machine learning approach”(2021)

[24] Kishan Kanhaiya, Rajat Gupta, Arpit Kumar Sharma, ”Cracked cricket pitch analysis(ccpa) using image processing and machine learning” (2019)

[25] Nilesh M. Patil, Bevan H. Sequeira, Neil N. Gonsalves , Abhishek A. Singh , “Cricket team prediction using machine learning techniques.”