iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35360

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Analysis of IPL Match Results using Machine Learning Algorithms

N. Lokeswari[1], L. Kalyan Pavan[2], K. Vandana[3], S.S.R. Rohan[4], S. Pavan[5]
[1] *Assistant Professor,* [2,3,4,5] *Student, Department of Computer Science,*
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

Abstract: ***Indian Premier League (IPL) is a famous Twenty-20 League conducted by The Board of Control for Cricket in India (BCCI). It was started in 2008 and successfully completed its thirteen seasons till 2020. IPL is a popular sport where it has a large set of audience throughout the country. Every cricket fan would be eager to know and predict the IPL match results.A solution using Machine Learning is provided for the analysis of IPL Match results. This paper attempts to predict the match winner and the innings score considering the past data of match by match and ball by ball. Match winner prediction is taken as classification problem and innings score prediction is taken as regression problem. Algorithms like Support Vector Machine(SVM),Naive Bayes, k-Nearest Neighbour(kNN) are used for classification of match winner and Linear Regression, Decision tree for prediction of innings score. The dataset contains many features in which 7 features are identified in which that can be used for the prediction. Based on those features, models are built and evaluated by certain parameters. Based on the results SVM performed.***

Keywords : IPL, Machine Learning, Match winner prediction, Score Prediction, SVM, kNN, Naive Bayes, Decision tree.

I. INTRODUCTION

Sports have gained much importance at both national and international level. Cricket is one such game, which is marked as the prominent sport in the world. T20 is one among the forms of cricket which is recognized by the International Cricket Council (ICC). Because of the short duration of time and the excitement generated, T20 has become a huge success. The T20 format gave a productive platform to the IPL, which is now pointed as the biggest revolution in the field of cricket. IPL is an annual tournament usually played in the months of April and May. Each team in IPL represents a state or a part of the nation in India. IPL has taken T20 cricket's popularity to sparkling heights .

It is the most attended cricket league in the world and in the year 2010, IPL became the first sporting event to be broadcasted live. Till date, IPL has successfully completed 13 seasons from the year of its inauguration . Currently, there are 8 teams that compete with each other, organized in a round robin fashion during the stages of the league. After the completion of league stages, the top 4 teams in the points table are eligible to the playoffs. In playoffs, the winner between 1st and 2nd team qualifies for the final and the loser gets another opportunity to qualify for the finals by playing against the winner between 3rd and 4th team. In the end, the 2 qualified teams played against each other for the IPL title. The significance is that IPL employs television timeouts and therefore there is no time constraint in which teams have to complete the innings. .

In this paper, we have examined various elements that may affect the outcome of an IPL match in determining the runs for each ball by considering the runs scored by the batsman in the previous ball as the labeled data. The suggested prediction model makes use of SVM and KNN to fulfill the objective of the problem stated. Few works have been carried out in this field of predicting the outcomes in IPL. In our survey, we found that the work carried out so far is based on Data Mining for analyzing and predicting the outcomes of the match. Our work novelty is to predict runs for each ball by keeping the runs scored by the batsman in the previous ball as the observed data and to verify whether our prediction fits into the desired model.

II. RELATED WORK

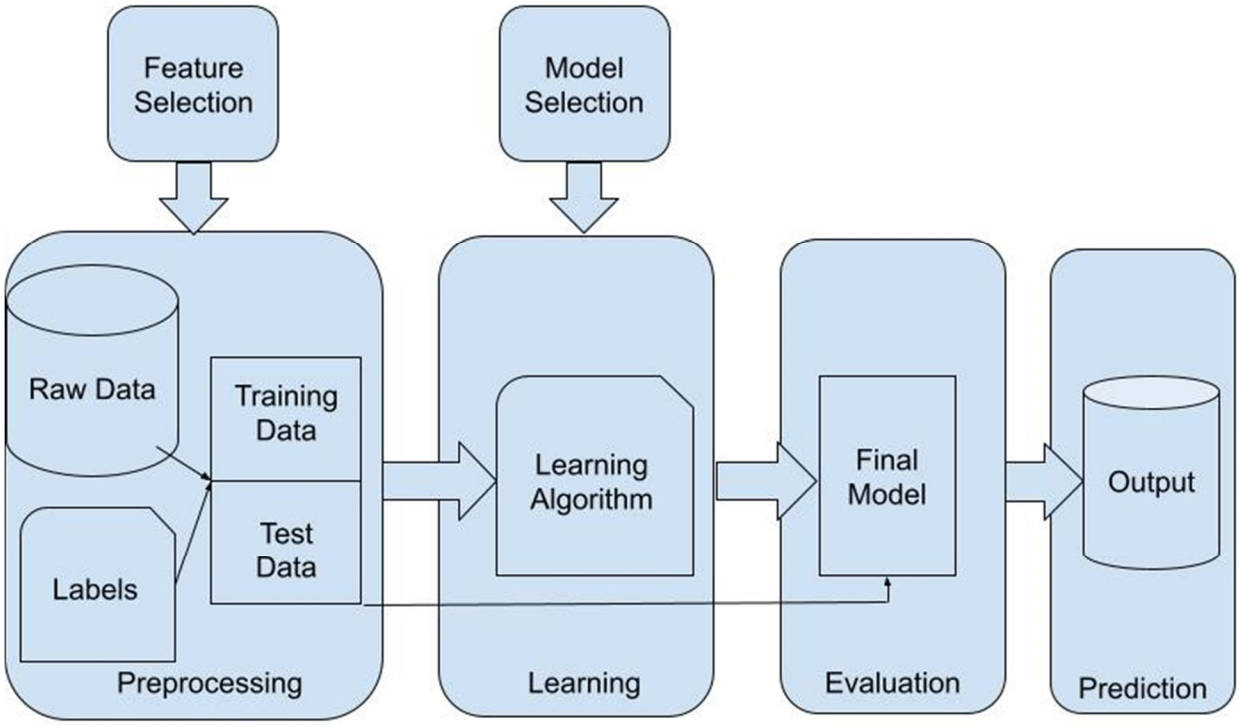
Rabindra Lamsal et al [1] aim to identify an optimal set of attributes and found 6 attributes which significantly influence the result of an IPL Match. The attributes include home team, away team, the toss winner, toss decision, the venue. Various classification based machine learning algorithms were trained on the IPL Dataset. Multilayer Perceptron outperformed other classifiers by predicting 43 out of 60, 2018 IPL Matches.

Gagana S et al [2] developed a new model for predicting runs by considering previously scored runs by batsmen. Research in the paper concludes that RNN and HMM give the best prediction accuracy for predicting runs in the IPL. Similarly this work can be negotiated for other formats of the game like test cricket, ODI matches and T20 matches.
Ch Sai Abhishek et al [5] aim to predict the winner in sports, cricket in particular. A generic classification model was designed to measure the points earned by each team based on their past performances.Random forest and Decision tree provided the highest accuracy of 89.151%.

III. METHODOLOGY

Machine learning has given computer systems the abilities to automatically learn without being explicitly programmed. In this we have used three machine learning algorithms (SVM, KNN, Naïve Bayes) .So, it can be described using the architecture diagram. The architecture diagram is shown as follows:

- IPL Data set
- Data Cleaning
- Pre processing the data
- Build Models
  Support Vector Machine
  KNN classifier
  Naive Bayes
- Test data
- Evaluate Performance



A. The architecture diagram is defined with the flow of process which is used to refine the raw data and used for predicting the IPL's data.

B. The next step is preprocessing the collected raw data into understandable format.

C. Then we have to train the data by splitting the dataset into train data and test data.

D. The IPL's data is evaluated with the application of a machine learning algorithm that is Support Vector Machine, KNN and Naïve Bayes algorithm and the classification accuracy of this model is found.

E. After training the data with these algorithms we have to test on the same algorithms.

F. Finally, the result of these three algorithms is compared on the basis of classification accuracy.

IV. MODULES

We have divided the whole project into 2 modules:

- Score Prediction
- IPL Match Prediction

For each of the modules, the sub module division is as follows:

A. Data set Collection:

The goal of this step is to identify and obtain all data-related problems. In this step, we need to identify the different data sources, as data can be collected from various sources such as files and database. The quantity and quality of the collected data will determine the efficiency of the output. The more data, the more accurate the prediction will be. We have collected our data from the Kaggle website. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training. It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.

B. Data Pre-Processing:

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It also includes encoding the values to numerical form which are in different formats other than numerical(like string, date etc..). It also includes feature selection so that we consider the features in which the output feature will be depending. It also includes standardization for the data set. In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose, if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as: training set and test set.

1) *Train Model:* Here, in this step we select the machine learning techniques such as Classification, Regression etc. then build the model using prepared data, and evaluate the model. We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and features.

2) *Test Data:* Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

C. Model Building(SVM, KNN Algorithm, Naive Bayes) and Testing:

1) Support Vector Machine(SVM):

a) Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges.

b) However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in N-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate.

c) Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Support Vectors are simply the coordinates of individual observation.

d) The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

e) Support Vector Machine algorithm works with categorical variables such as 0 or 1, Yes or No, True or False, Spam or not spam, etc.

2) KNN Algorithm:

a) K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

b) The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.

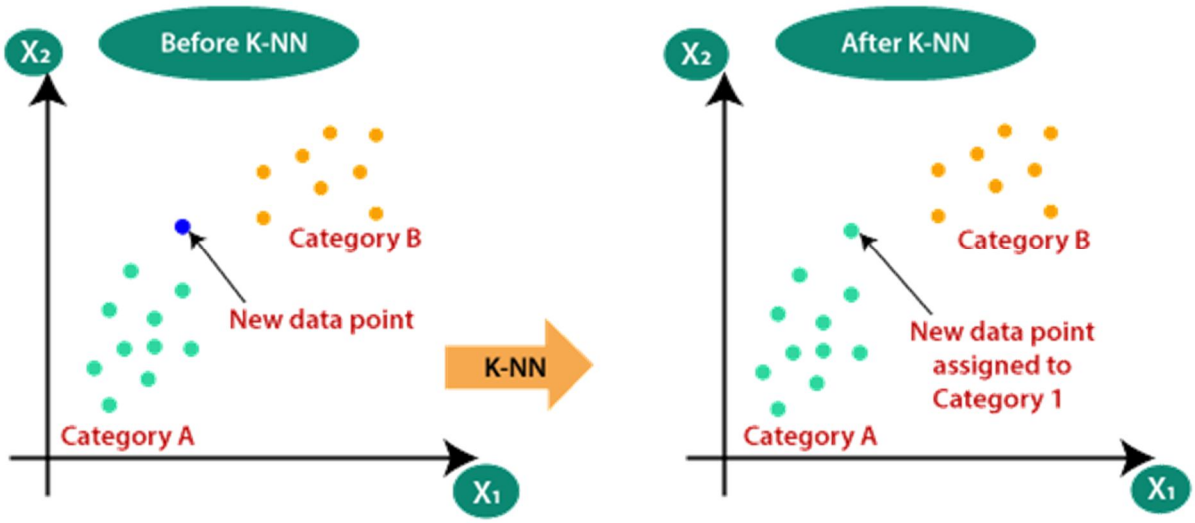
c) K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.

d) The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

e) K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

f) It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

g) KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



Steps for KNN:

Step 1: Load the data.

Step 2: Initialize the value of k.

Step 3: For getting the predicted class, iterate from 1 to total number of training data points.

- Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.

- Sort the calculated distances in ascending order based on distance values.
- Get top k rows from the sorted array.
- Get the most frequent class of these rows.
- Return the predicted class.

3) Naive Bayes :

a) Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

b) It is mainly used in text classification that includes a high-dimensional training dataset.

c) The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

d) It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

e) Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

P(h|D)=P(D|h)P(h)/P(D)

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

V. PERFORMANCE MEASURE

As the model is built for Prediction the next step is to measure the performance of the model. To evaluate the model, some of the standard measures such as precision and recall are used. Confusion matrix is used for the calculation of these measures. Accuracy is the measure of total correct predictions to that of total predictions.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

2019 Matches :

Method	**Precision**	**Recall**	**F1 Score**	**Accuracy**
SVM Linear	0.65	0.95	0.77	0.65
SVM RBF	0.90	0.95	0.92	0.90

Naive Bayes	0.65	0.70	0.68	0.58
KNN	0.78	0.84	0.81	0.75

2020 Matches :

Method	**Precision**	**Recall**	**F1 Score**	**Accuracy**
SVM Linear	0.51	0.90	0.65	0.51
SVM RBF	0.49	0.97	0.65	0.48
Naive Bayes	0.48	0.40	0.44	0.48
KNN	0.56	0.67	0.61	0.56

VI. CONCLUSION

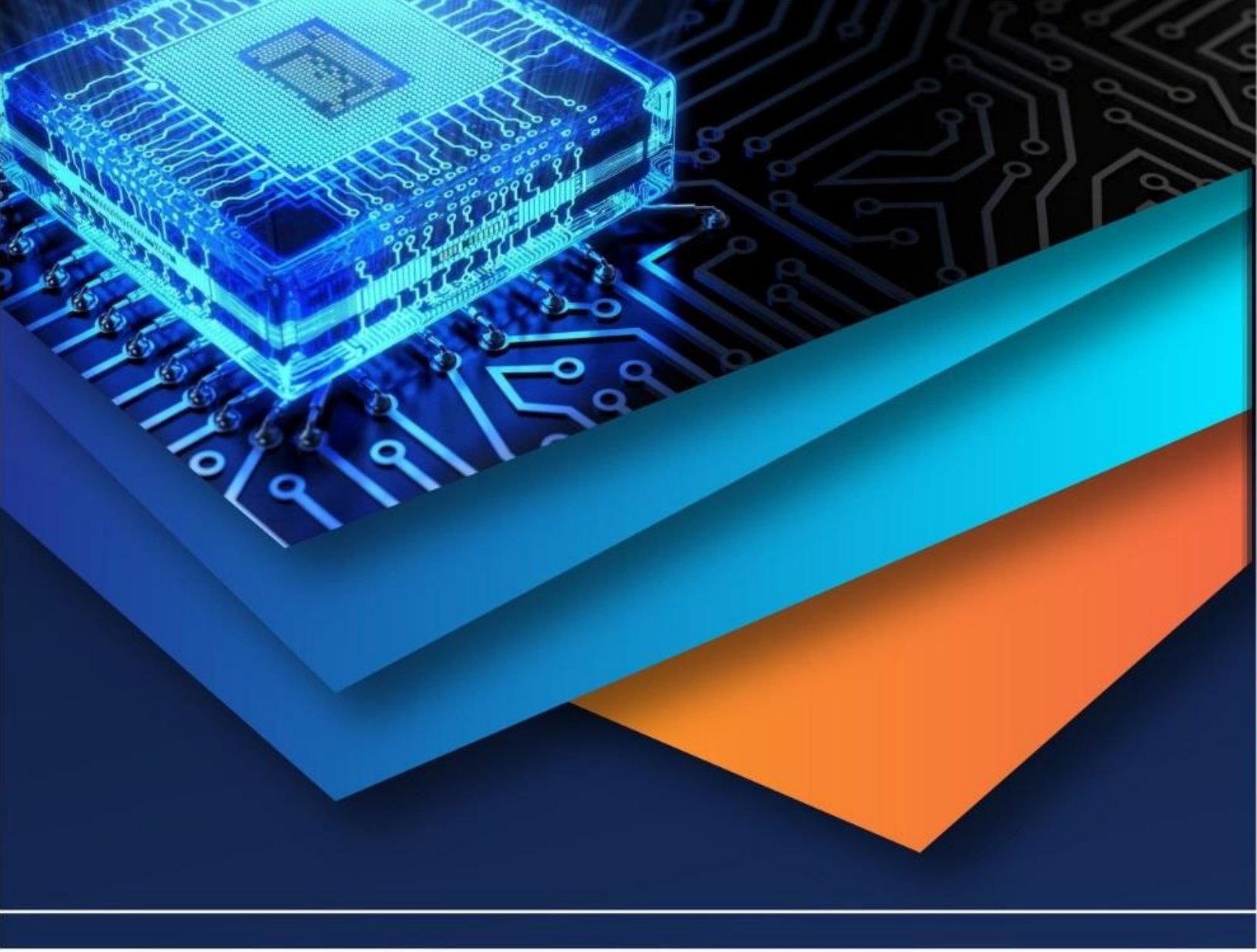
Support Vector Machine(SVM), Naive Bayes, k-Nearest Neighbour(kNN) algorithms are implemented on the input data to assess the best performance. These methods are compared using performance metrics. According to the analysis of metrics, Support Vector Machine(SVM) gives a better accuracy score on test data than the other two algorithms.

VII. FUTURE SCOPE

At present, the data is limited to match and score. It doesn't have details about the players and their stats. There is a great scope for applying this concept to the players and their stats data and can find the batting order and bowling order of a particular match. It will be helpful to franchise people who are at decision making level.

REFERENCES

[1] Gagana S, K Paramesha, "A Perspective on Analyzing IPL Match Results using Machine Learning", IJSRD - International Journal for Scientific Research & Development| Vol. 7, Issue 03, 2019 | ISSN (online): 2321-0613.
[2] Rabindra Lamsal and Ayesha Choudhary, "Predicting Outcome of Indian Premier League(IPL) Matches Using Machine Learning", arXiv:1809.09813v5 [stat.AP] 21 Sep 2020.
[3] Shubhra Singh, Parmeet Kaur, "IPL Visualization and Prediction Using HBase", Procedia Computer Science 122 (2017) 910–915.
[4] Pallavi Tekade, Kunal Markad, Aniket Amage, Bhagwat Natekar, "Cricket Match Outcome Prediction Using Machine Learning", International Journal Of Advance Scientific Research And Engineering Trends, Volume 5, Issue 7, July 2020, ISSN (Online) 2456-0774.
[5] Ch Sai Abhishek, Ketaki V Patil, P Yuktha, Meghana K S, MV Sudhamani, **"**Predictive Analysis of IPL Match Winner using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S, December 2019.
[6] https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY