IPL SCORE PREDICTION USING MACHINE LEARNING

[1]Mr. Chandan Kumar, [2]Dr. Shambhu Kumar Singh
[1]Research Scholar, [2]Assistant Professor & Head
CSE Department
Sandip University, Madhubani, BIHAR, INDIA
chandankumar9597k@gmail.com, shambhu.singh@sandipuniversity.edu.in

ABSTRACT

Cricket is a very familiar and exciting sport that people of all age groups are insane to see and play. For many it's a billion-dollar market as they speculate financially, hoping to be able to earn profit in the form of gambling and various other ways. accuracy so that desired predicted output is accurate.

The Indian Premier League (IPL) is one of the several series that are contested in the nation. A model with two techniques has been proposed. The first is a scoring prediction, and the second is a prediction of the team winning. In this project, a model using machine learning algorithms is proposed to predict the score of each match and winning team based on past datasets available from 2008 to 2019 IPL matches in Kaggle. This proposed methodology includes the following steps like Pre-processing of collected datasets, Feature selection from raw data, Conversion of categorical data into numerical data, Partitioning of samples into training and test samples, Training, and classification. Few machine learning algorithms like Support Vector Machine, Random Forest, Naive Bayes were already used in previous papers. In this project, algorithms like Lasso Regression, Ridge Regression, and Random Forest regression models are proposed for a score prediction, and SVM(Linear, RBF), Logistic Regression classifier is for the match-winning prediction. The accuracy of the above machine learning algorithms is used to predict the winner of an IPL match along with its Precision, Recall and F-Measure measured and the model with better accuracy is considered.

Linear regression, logistic regression, decision trees, random forests, gradient boosting regressors, extra tree regressors, and XGB regressors are employed in these for score prediction. This study gathers and analyses IPL data spanning multiple years, including player, match, team, and ball-to-ball information, to generate several conclusions that help improve a player's performance. To forecast the winner, the model employed a supervised machine learning technique. For high accuracy, Extra tree regressor used for good accuracy with 90%.

1. INTRODUCTION

Cricket is the most widespread and much-loved game of everyone. it's delighted in by the overall population of all age mass because it is an exceptionally fascinating and suspicious game. Cricket is also referred to as the Game of Uncertainty and there is no precise forecast that a selected team would win in any given conditions. Finally, a team wins which multiplies the energy of every team member. There turn into a major jam of cricket darlings within the stadium and television rooms to see the cricket at whatever point i.e., a world level, national level, or any test match. The magnetism of cricket

has also included businessmen that became a source of income for them as they gamble over their favorite teams.

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. The popularity of cricket increased when ICC (International Cricket Council) started the concept of fast cricket within the sort of twenty-20(T-20) matches. In 2007, the first twenty-20 world cup was held within South Africa that was won by India which increased the popularity of this game in India. BCCI (The Board of Control for Cricket In India) cashed the chance and created a league referred to as the Indian Premier League (IPL) in 2008 and got it approved by ICC. IPL is one of the finest twenty- 20 cricket competitions in the present cricketing world that is based on the EPL (English Premier League) league and NBA (National Basketball Association) Basketball League [5]. During its first edition, IPL gained huge popularity which opened avenues for many stakeholders. In every IPL season 8 teams play with one another within the first stage, after the first stage 4 teams attend the eliminator round (next stage) and after the eliminator round 2 teams attend the final match and at last, there will be one winner. Each team is owned by a franchise that is owned by a group of people. These franchises hire players, evaluate them on the idea of their national, international, T-20 experience and performance, and hire them at the time of auctions. Results of each match within the IPL depend on the varied conditions like venue, player performance, toss, performance in power-play, etc. Results of a match can only be predicted to some extent if previous player performance, venue, and other match-related data are available. In this paper, the authors predict the results of IPL match using three machine learning algorithms namely SVM (Linear, RBF), Logistic Regression classifier on the idea of previous data available.

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. Multicollinearity 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.

2. LITERATURE SURVEY

According to the findings of the literature review, there is a need for a machine learning model that can predict the outcome of an IPL match before it starts. The Twenty20 format of cricket, more than any other, sees a lot of changes in the game's momentum. A game can be radically changed by an over. As a result, forecasting the outcome of a Twenty20 game is a difficult undertaking. Developing a prediction model for a league that is entirely based on auction is also a challenge.

IPL matches cannot be anticipated just based on statistics derived from historical data. Players are bound to change teams as a result of player auctions, which is why the continued performance of each player must be taken into account while constructing a prediction model.

This research paper 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. As in the dataset there are many parameters are present. Out of them which parameters are helpful in the project. The factor's 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. The role of classification gives proper information or use of naive bias and linear regression. They give 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. Using machine learning techniques like Decision Tree, SVM, Decision Tree, logistic regression, random forest classifier, and k-nearest, the authors of discovered and noted some things. The random forest classifier beats every algorithm in this experiment by accurately predicting the outcome with the highest accuracy of 88.10%. This work has examined and analyzed IPL score prediction in Understanding the IPL data set from the previous ten years is the goal of this endeavor. Understanding the operation and use of the four distinct machine learning algorithms is beneficial work utilizing machine learning algorithms in Each player's point total was utilized to determine each team's relative strength. Using the IPL dataset created for this purpose, several classification-based machine learning algorithms were trained. The research focused on predicting the winner for an IPL match using machine learning and utilizing the available historical data of IPL from season 2008- 2019. 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. This will make things easier so that anyone checks the match prediction just by using their mobile or PC. The proposed LR algorithm shows better results as compared to the other previous ML algorithms. When the actual scores and the predictions were compared in , the findings showed a strong association between the two. The average impact factor of the team based on featuring players is taken into consideration in order to predict the result of a match using player performance data and the history of IPL cricket. The accuracy of Linear Regression in Score Prediction Analysis is higher than that of Ridge and Lasso Regression. In this study, we present the feedback analysis for tweets following IPL-2020 matches and examine the team's level of fame during the competition.

DATASET FEATURES

The approach over here we are using is ML based. So, the basic requirements of an ML algorithm are 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.

Attributes	**Values**
Batting Team	Batting Team Name among 12 teams in IPL
Bowling Team	Bowling Team Name among 12 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 1: 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

Table 2: Dataset Attributes and their values

3. METHODOLOGY

The proposed method consists of five sub modules, namely, loading the dataset, pre- processing, feature selection, classification using various algorithms and comparison of algorithms as shown in Fig.1.

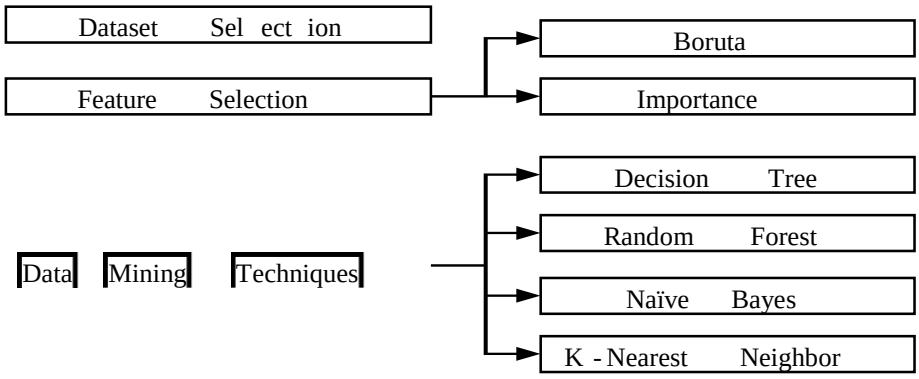

Figure 1: Methodology Diagram

LOADING THE DATASET

The dataset name is matches.csv (IPL Matches data from 2008 to 2017) whose size is 117,096 bytes and it is taken from the Kaggle Repository. The number of attributes is 18 and total number of records is 637. The Attributes of the dataset is id, season, city, date, team1, team2, toss_winner, toss decision, result, dl_applied, winner, win_by_runs, win_by_wickets, player_of_match venue, umpire1, umpire2, umpire3.

The dataset is loaded into the R Tool and command read.csv() is used to upload the data and this data is stored in the dataset named IPL data.

DATA PRE-PROCESSING

Data Pre-Processing plays a vital role in machine learning. It transforms raw data into a useful data format. Commonly it is used as a preliminary step to clean the data. Data Pre-Processing transforms the data into a format for more easily and error free processing for the classification. The dataset is first processed to remove the null attributes and the records that contain the NA attributes. The attribute umpire3 is removed initially as it had no values. The fields date and player_of_match are converted to numeric fields. Records with NA in the winner and player_of_match are removed. The levels in the winner fields are also dropped to make it a non-factor variable. These pre- processing has to be done before the feature selection and classification techniques.

FEATURE SELECTION

Feature selection is the use of specific attributes in the dataset to maximize efficiency Feature selection is also known as variable selection. It is important phase in machine learning because it significantly improves the performance by eliminating redundant and irrelevant features and also at the same time speeding up the learning task. Feature selection is done using two functions namely the Boruta() and the importance() functions. The Boruta() function is in the Boruta package and the importance() function is in the randomForest package. The Boruta function a narrow – down search for relevant features by comparing with original attributes. The importance is achievable at random estimated using their permuted copies, and progressively eliminating all irrelevant features stabilize that test. The importance() function is the function of extraction for variable importance measured as produced by random Forest. With the Boruta() function, date, dl_applied, umpire2 are confirmed as unimportant.

With the importance() function, umpire1, umpire2, venue, result and dl_applied are with least Mean Decrease Accuracy. Hence, the fields umpire1, umpire2, venue, dl_applied and result were removed by comparing both the algorithms.

CLASSIFICATION

In Machine Learning, classification is an important technique to classify different classes. It is a supervised learning method in which the computer program learns from the training data, and uses this learning to classify new data. Here four different classification algorithms are applied, namely, Decision Tree, Random Forest, Naive
Bayes and K-Nearest Neighbour.

Decision Tree

Decision Tree is one of the supervised learning algorithms which is used for both classification and Regression. A Decision Tree is a graph that uses a Tree based method to illustrate every possible outcome of a decision. A decision tree is a decision support tool which uses a tree-like structure and their possible consequences. The packages caret and rpart.plot are used from which the functions rpart(), createDataPartition(), trainControl(), train(), prp() and predict() are function are used to get the result of decision tree algorithm.

Random Forest

Random forest is a supervised learning method. In the random forest classifier, the more the number of the trees the more the best accuracy for the model. Random Forest is also an ensembled based method used for classification, regression and other tasks. The package randomForest is used which contains the functions sample(), randomForest() and plot() that are used to obtain the results of the Random
Forest algorithm.

Naive Bayes

A Naive Bayes classifier is a supervised learning algorithm which works based on Bayes' theorem. Naive Bayes classifiers uses conditional probability theorem to classify the data. Bayes classifiers will assume whether strong or naïve independence between attributes of data points. These classifiers are broadly used in text categorization based problems because they are easy to carry out. Naive Bayes is also known as independence Bayes. The naïve Bayes () function in the naïve Bayes
Package of R is used to obtain the results.

K-Nearest Neighbor

KNN is the simplest classification algorithm. A k-nearest- neighbour is a classification algorithm that attempts to determine how near the group of data points are around it. Used for both classification and regression base problems. The package RWeka and the function IBk() are used to achieve the results of this algorithm.

4. CONCLUSION

This work aims at understanding the dataset of past 10 years history of the IPL data. It helps to understand the four different machine learning algorithms working principal and their implementation

in R. 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. It takes into consideration the measures accuracy, error rate, precision, recall, sensitivity and specificity. Based on this the best algorithm is selected as Random Forest. This work focuses on exploring IPL data and presenting its insights as graphical representation and comparative analysis. By making use of this,

Indian Premier League and the fan followers can take decisions on the team's performance and predict the trophy winners that will lead to success in future

REFERENCES

- S. Abhishek, Ketaki V. Patil, P. Yuktha and S. Meghana, Predictive Analysis of IPL Match Winner using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering, Vol. 9, No. 1, pp. 430435, 2019.
- Sanjay Gupta, Hitesh Jain, Asmit Gupta and Hemant Soni, "Fantasy League Team Prediction", International Journal of Research in Science and Engineering, Vol. 6, No. 3, pp. 97- 103, 2017.
- Pabitra Kumar Dey, Gangotri Chakraborty, Purnendu Ruj and Suvobrata Sarkar, "A Data Mining Approach on Cluster Analysis of IPL", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. 351-354, 2012.
- Raza Ul Mustafa, M. Saqib Nawaz, M. Ikram Ullah Lali, Tehseen Zia and Waqar Mehmood, "Predicting the Cricket Match outcome using Crowd Opinions on Social Networks: A Comparative Study of Machine Learning Methods", Malaysian Journal of Computer Science, Vol. 30, No. 1, pp. 63-76, 2017.
- 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. 30- 32, 2018.
- Vignesh Veppur Sankaranarayanan and Junaed Sattar,"Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction", Proceedings of SIAM Conference on Data Mining, pp. 1-7, 2014.
- Parag Shah, "Predicting Outcome of Live Cricket Match using Duckworth-Lewis Par Score", International Journal of Latest Technology in Engineering, Management and Applied Science, Vol. 6, No. 7, pp. 72-75, 2017.
- Amal Kaluarachchi and S. Aparna, "A Classification based Tool to Predict the outcome in ODI Cricket", Proceedings of 5th International Conference on Information and Automation for Sustainability, pp. 233-237, 2010.
- C. Deep Prakash Dayalbagh, C. Patvardhan and C. Vasantha Lakshmi, "Data Analytics based Deep Mayo Predictor for IPL-9", International Journal of Computer Applications, Vol. 152, No. 6, pp. 6-11, 2016.
- Jayshree Hajgude, Aishwarya Parameshwaran, Krishna Nambi, Anupama Sakhalkar and Darshil Sanghvi, "IPL Dream Team-A Prediction Software Based on Data Mining and Statistical Analysis", International Journal of Computer Engineering and Applications, Vol. 9, No. 4, pp. 113-119, 2015.
- Sonu Kumar and Sneha Roy, "Score Prediction and Player Classification Model in the Game of Cricket using Machine Learning", International Journal of Scientific and Engineering Research, Vol. 9, No. 2, pp. 237-242, 2018.