International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

ANALYSIS AND WINNER PREDICTION OF CRICKET MATCH

Aditi Dalvi[1], Sneha Singh[1], Nitish Patel[1], Dr. Rahul Khokale[2]

Aditi Dalvi, Student, Dept. of I.T. VPPCOE & VA, Mumbai University Mumbai, India, avu4f1819079@pvppcoe.ac.in
Sneha Singh Student, Dept. of I.T. VPPCOE & VA, Mumbai University Mumbai, India, vu4f1819072@pvppcoe.ac.in
Nitish Patel Student, Dept. of I.T. VPPCOE & VA, Mumbai University, Mumbai, India vu4f1819073@pvppcoe.ac.in
Dr. Rahul Khokale HOD, Dept. of I.T. VPPCOE & VA, Mumbai University Mumbai, India Prof.rahulkhokale@pvppcoe.ac.in

ABSTRACT

In current time, machine learning is the most famous field to predict future output for making better decisions based on these predictions. Cricket is a well-known game that is watched and played around the world in more than 100 countries. Many of these cricket fans want their team to perform well and become the winner of the match. To make sure their team's win, team's have to work on their performance and area of strength. Similarly predicting winner of a cricket match depends on many factors like toss, team strengths, venues and weather conditions, etc. This research paper is about doing exploratory data analysis on cricket dataset and also predict IPL match winner. The winner of IPL match is predicted by training machine learning models on the selected features. For the purpose of model building, various machine learning algorithms have been used and applied on the test and training datasets of different size which are Random Forest, SVM, Linear Regression, Logistic Regression and Decision Trees. This model is very useful for legal betting applications, match reporting media and cricket enthusiasts. Exploratory data analysis on cricket dataset will be beneficial for cricket team management or analytics team for evaluating the team's strength.

***Keywords**: Cricket, EDA, Winner Prediction, Data Mining, Random Forest Algorithm*

1. INTRODUCTION

Use of statistical analysis in sports has been growing rapidly year by year. Due to which not only the ways in which game strategies were made or the player's rating criteria has been affected but also there has been a tremendous increase in the number of audiences that watches cricket.

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

Now, Cricket has become one of the most followed team games on the globe with billions of fans all across the world. Cricket is a team sport that is played globally across 106-countries, which are members of the International Cricket Council (ICC), which has 2.50 billion worldwide fans according to ICC. Today, there are three major formats in which cricket is being played all over the world, One Day Internationals (ODIs), T20 cricket and Test Matches. Besides these international cricket matches, T20 being the shortest and the most exciting format of the game, T20 League cricket is catching more attention of the fans.

Indian Premier League (IPL) is one of the most popular T20 cricket league in the world. In franchised-cricket every team participates to win and improve their team's performance. For this purpose, every team needs a better management panel to handle the responsibility of a franchise, team selection committee that will select the best possible team with good combination of players such as to select the best all-rounder from the draft by looking at their past performances. Indian Premier League is an international competition played in India which starts from the end of March to May every year among eight teams but from year 2022 number of teams has been increased to ten. 18 to 25 players are selected by each team. Every playing team consist of 11 players, four overseas players and seven local players. Every team's performance is based on toss, venue, conditions and other important factors which decides the team's performances in a cricket match.

In our project we have done analysis of IPL matches and also predicted winner of an IPL match using two different methods.

2. LITERATURE SURVEY

Cricket Score Prediction (Prasad Thorat,Vighnesh Buddhivant,Yash Sahane - 2021)

Objective: Prediction of Cricket Scores for Live IPL Matches. Dataset used: Dataset is taken from Kaggle Website. Factors used for Prediction: Runs scored in 5 overs and wickets taken. Algorithms used: Linear Regression. Future Scope: Predicting the chasing probability.

Analyzing and Predicting Outcome of IPL Cricket Data (D. Jyothsna, K. Srikanth - 2019)

Objective: To analyze the cricket data and predict the Player's performance. Dataset used: The seven IPL seasons real time dataset is taken in CSV format. Factors used: Toss winners

Algorithms used: Linear Regression, Decision tree, K-means, Logistic Regression. Future Scope: To predict the winning teams of next IPL Matches.

Predicting and Analysing the Performance of IPL Cricket Using Regression Models (Kasukurti Raviteja, Ganesh Kumar Macha, Dr. Gr Anantharaman - 2019)

Objective: Predicting the outcome of IPL Cricket Match. Dataset used: The required dataset is taken from an internet web source. Factors used: Toss, Over by Over: runs scored, wickets taken, extras and boundaries. Algorithms used: SVM, Random Forest, Decision Tree classifiers. Future Scope: The hyper parameters of the classifier may be fine-tuned in future to get better performance.

Cricket Score and Winning Prediction using Data Mining (Akhil Nimmagadda, Nidamanuri Venkata Kalyan, Manigandla Venkatesh, Nuthi Naga Sai Teja, Chavali Gopi Raju - 2018)

Objective: Predicts the score in each of the innings and finally the match. Dataset used: Dataset is taken from cricinfo website. Factors used: For score prediction - run rate, For winner prediction - toss winner. Algorithms used: Random forests, Logistic Regression, Multiple Linear Regression. Future Scope: More features can be added to improve the performance.

Sports analytics for cricket game results using machine learning: An experimental study (Kumash Kapadia, Hussein Abdel-Jaber, Fadi Thabtah, Wael Hadi - 2019)

Objective: Predicting the outcome of IPL Cricket Match. Dataset used: IPL Cricket Matches for 10 years (2008-2017) downloaded from Kaggle website. Factors used: Home ground, Toss winner. Algorithms used: Naive Bayes and KNN. Future scope: To make the prediction model adapt at addressing the entire match scenario.

Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Strengths(Sasank Viswanadha, Kaustubh Sivalenka, Madan Gopal Jhawar, Vikram Pudi - 2017)

Objective: Dynamic Winner Prediction in a Twenty20 cricket match. Dataset used: Dataset is taken from cricinfo and cricsheet website. Factors used: Target left, Wickets remaining, Overs Left, Relative Team strength Team A/Team B. Algorithms used: SVM, Random Forest, Decision Trees, KNN, Logistic Regression. Future Scope: To further make the model prediction adapt at addressing the entire match scenario.

3. PROPOSED SYSTEM

The IPL is one of the most-attended cricket league in the world and it ranks sixth on all sports league. Therefore, the proposed system focuses on analyzing and predicting the results of the IPL matches by applying various algorithm in data mining. Accuracy of all the algorithms that we have used will be compared and the one that is giving the highest accuracy will be used in the model. Data set has been collected for the IPL matches - Ball by Ball and IPL match dataset from Cricinfo and cricsheet. The modules contain the following steps, Data cleaning is process of removing the incomplete and unimportant data, missing information and also correct records. Data transformation is known for finding the inaccurate records and replacing it with converting one form of data into another form of data. It is considered to be both simple and complexes based on data between initial data and final data. Data pre-processing is a technique that involves transforming the imprecise form of data onto correct form. Real world data is often represented to be in form of inappropriate and inconsistent records. Attributes which are not important for the prediction of the match have been excluded from the dataset. As we are going to predict the winner of the match, we have to create the testing and training dataset with pre-processing. It includes attributes of team1, team2, win_by_run, win_by_wicket, toss winner, winner, etc. The attribute winner is divided in training set to foreknow the result of the match. We are also going to predict winner based on the target and score in the second innings after at least 5 overs.

In this project we will be creating a website where deep analysis of IPL can be done, prediction using 2 different methods is done by applying Machine Learning approaches (and eventually comparing them) for predicting which team will win the match based on input features and different statistics can also be observed. The user will input details for prediction using the Graphical User Interface. We are using Streamlit framework for this project.

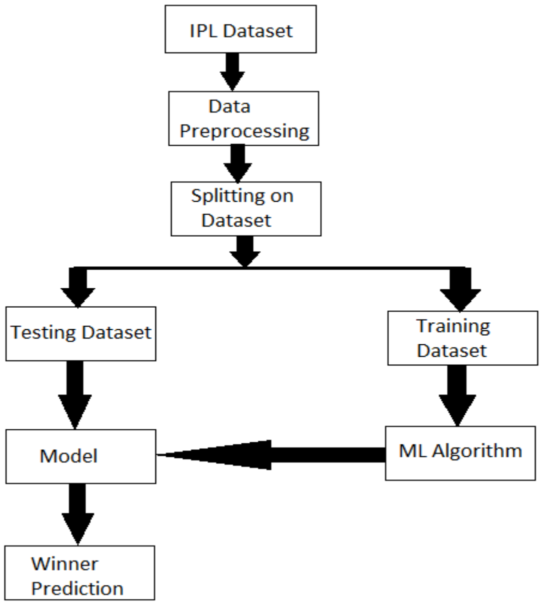

Fig.1: Use case diagram (winner prediction)

4. EXPLORATARY DATA ANALYSIS

The fig 3 shows how winning the toss has affected the results of a particular team

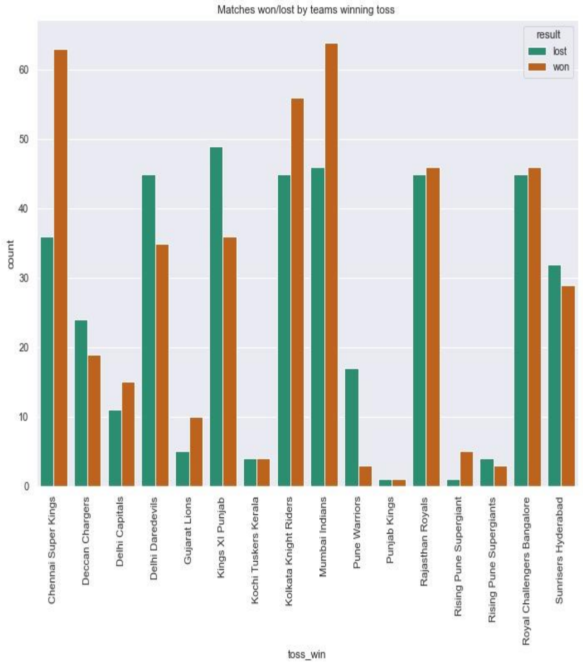

Fig.2: Matches loss/win by teams winning toss

In fig 3 there is graph which is showing the no. of matches that won by different teams in ipl till now on the basis of that we can find which team is stronger.

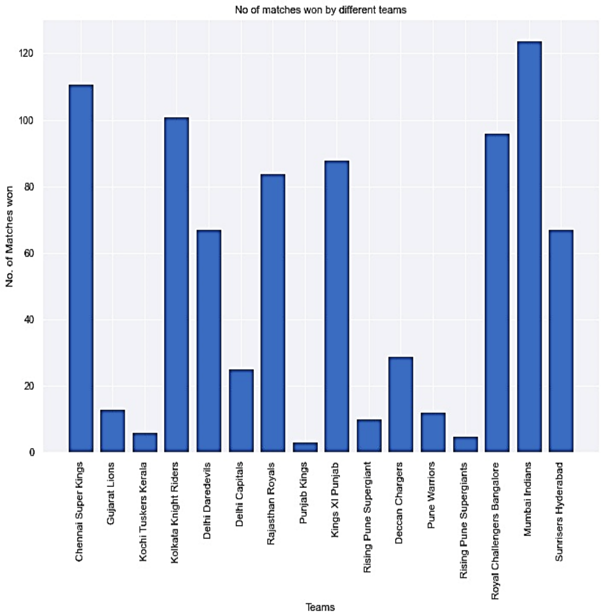

Fig.3: Total no. of matches won by different teams

Win Percentage of CSK at different venues. This analysis has been done for all the teams playing IPL

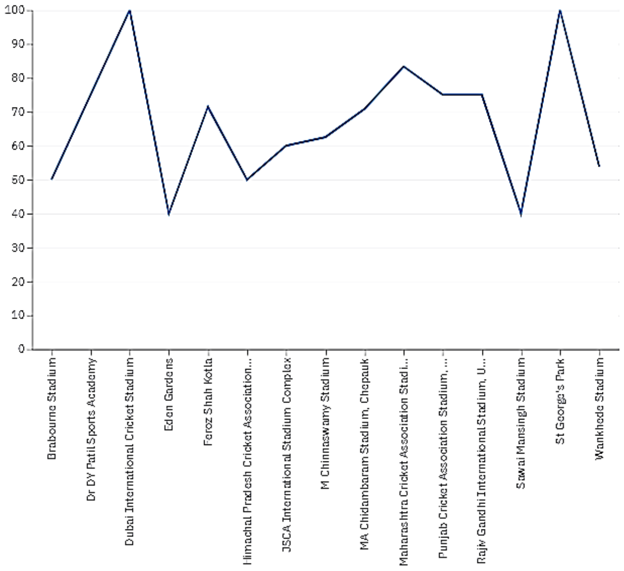

Fig.4: Win Percentage at a Venue

Modelling the dataset:

We will be using linear regression model, random forest regression and decision tree model for the prediction. the model with the highest accuracy will be selected for the prediction.

Machine Learning Algorithms used for prediction of match winner using score:

For evaluation of the models we have used evaluation metrics. We are using MAE(mean absolute error) and R squared value as evaluation metrics.

MAE: It represents the average of absolute difference between actual and predicted values in the dataset.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | x_i - x |$$

Where,

|xi-x|=absolute error

N= Data set size

R squared: This is used to determine how well a line fits to a dataset of observations especially when comparing models.

$$R^2 = \frac{SSR}{SST}$$

Where,

SSR= sum of squared regression

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$\hat{y}_i$= predicted value of y for observation i

$\bar{y}$ =mean of y value

SST= Total variation in the data

$$SST = \sum_i (y_i - \bar{y})^2$$

yi= value for observation i

$\bar{y}$ =mean of y value

Linear Regression: Linear Regression is a machine learning algorithm i.e., based on supervised learning that performs regression task. A regression models a target prediction value based on independent variables provided.

MAE value= 12.1347

R2 score= 0.751587

Accuracy= R2 score * 100 = 75.1587

Random Forest: Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a legion of decision trees at training time. Random Forest Regression is a supervised learning algorithm where ensemble learning method is used for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

MAE value= 13.8396

R2 score= 0.664979

Accuracy= R2 score * 100 = 66.4979

Decision Tree algorithm: Decision Tree algorithm in machine learning is one of the most popular algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well dividing for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

MAE value= 16.8462

R2 score= 0.483469

Accuracy= R2 score * 100 = 48.3469

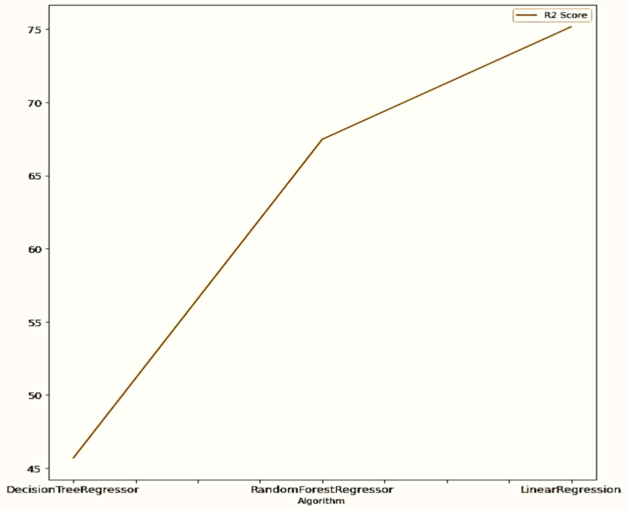

Fig.5: Popular Supervised Learning Algorithms Compared based on R2 Score

Machine Learning Algorithms used for prediction of

match winner using toss:

For evaluation of the models we have used score

Method from sklearn library.

Random Forest: Random decision forests is an

Ensemble learning method for classification,

regression and other tasks that operates by

constructing a legion of decision trees at training time.

For classification tasks, the class selected by most trees is the output of the random forest.

Accuracy= 88.2282%

Support Vector Machine (SVM): Support Vector

Machine (SVM) is a supervised machine learning

algorithm which is used for both classification and

regression. The objective of SVM algorithm is simple

i.e., to find a hyperplane in an N-dimensional space

that distinctly classify the data points.

Accuracy=27.8240%

Logistic Regression: Logistic Regression is used to

Estimate discrete values (usually binary values like 0/1)

From a set of independent variables. It helps to

Predict the probability of an event by fitting data to

A logit function. It is also called logit regression.

Accuracy=28.6563%

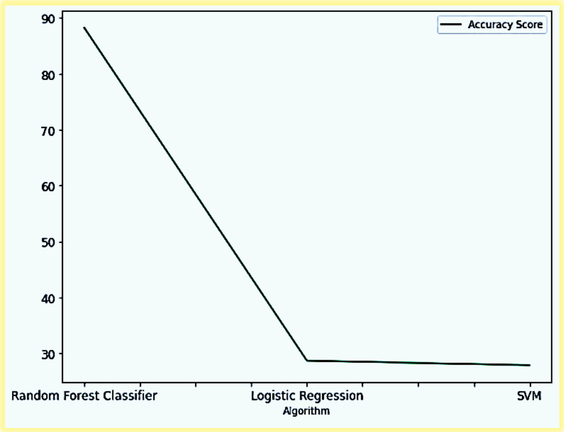

Fig.6: Popular Supervised Learning Algorithms Compared based on Accuracy Score

5. CONCLUSION AND FUTURE SCOPE

This project proves that machine learning is extremely useful in predicting winner of the IPL match. The algorithm used in this experiment has performed really well using the available attributes. The analysis done in our project can be useful for cricket enthusiasts and also help the team management to take better decisions.

We have built 2 machine learning models using- Linear Regression and Random Forest Classifier. Random Forest Classifier is used to predict the winner based on toss and venue whereas Linear Regression is used to predict winner based on the target and runs scored by the chasing team in at least 5 overs. As we know that T20 is a dynamic game, so we have done prediction of the winner considering the dynamic nature of the format.

In the future we can consider Player's Performance as one of the attributes as we know that player's form is also one of the most important factor for a team to win a match. Accuracy of a model would increase if we consider player's performance.

REFERENCES

[1] Viswanadha Sasank, Sivalenka Kaustubh, Madan Gopal Jhawar, and Pudi Vikram: **Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths.** (Machine Learning for Sports Analytics at ECML-PKDD)

[2] D. Jyothsna , K. Srikanth : **Analyzing and Predicting outcome of IPL Cricket Data.** (International Journal of Innovative Research in Science, Engineering and Technology)

[3] Shimona S. , Nivetha S., Yuvarani P. : **Analyzing IPL Match Results using Data Mining Algorithms.** (International Journal of Advance Research and Development)

[4] Raviteja Kasukurti, Macha Ganesh Kumar, Dr. Granantharaman : **Predicting and Analyzing the Performance of the IPL Cricket Using Regression Models.** (Complexity International Journal)

[5] Prasad Thorat, Vighnesh Buddhivant, Yash Sahane : **Cricket Score Prediction** (International Journal of Creative Research Thoughts)