IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE ANALYSIS OF AN IPL MATCH USING MACHINE LEARNING

Kiran Gawande, Prof. Sumit Harale, Simran Pakhare

Department of Computer Engineering, Indira College of Engineering & Management,

Parandwadi, Pune

Abstract: Prediction of outcome of a match using machine learning algorithms is a crucial aspect in cricket. Records of the past performance of players and other related data are often analyzed to make models that predicts the winning team. This model is created using the machine learning algorithms like Decision Tree, Naive Bayes, Logistic Regression, Random Forest, SVM, K-Nearest neighbor and their results are often compared supported the evaluation measures as accuracy, precision, recall, f1 score and support. For prediction we've used a number of the features like City, Team1, Team2, Toss winner, Winner for prediction with higher accuracy. For this process we took dataset from Kaggle.com, imported necessary python libraries then we visualized data. We trained ML model and predicted results. For prediction we studied various classification algorithms to get best accuracy.

Keywords: **Machine Learning, Random Forest, Classification Algorithm, Cricket Prediction**

I. INTRODUCTION:

Machine Learning is often used effectively over various times in sports, both on-the-field and off-the-field. When it's about on-the-field, machine learning applies to the analysis of a players fitness level, design of offensive tactics, or decide shot selection [3]. It's also utilized in predicting the performance of a player or a team, or the result of a match. On the opposite hand, the off-the-field scenario concerns the business perspective of the game which incorporates understanding sales pattern (tickets, merchandise) and assigning prices accordingly. The most focus is that the healthy growth in business and profitability of the team owners and other stakeholders. On-the-field analytics generally make use of supervised machine learning algorithms, example: (i) regression for calculating the fitness of a player, (ii) classification for predicting an outcome of a match; while off-the-field analytics concerns around performing sentiment analysis to know people's opinion a few players or a team or a sport league. At the present, Twitter has become one among the first sources of knowledge for sentiment analysis.

The use of analytical methods in various aspects of cricket including results prediction is extremely important. There's an enormous demand for the algorithm that best predicts the results of cricket due to its popularity and large amount of cash involved within the game. Thus, the analysis of IPL results becomes more important. Prediction of outcome of a match using machine learning algorithms is a crucial aspect in cricket. Records of the past performance of players and other related data are often analyzed to make models that predicts the winning team. This model is often created using the machine learning algorithms like Decision Tree, Naive Bayes and K-Nearest neighbor and their results are often compared supported the Evaluation Measures as accuracy, precision, recall, sensitivity and error rate.[3] Since the dawn of the IPL in 2008, it attracted viewers everywhere the world. High level of uncertainty and last moment nail biters has urged fans to watch the matches. Within a quick period, IPL has become the absolute best revenue generating league of cricket. Data Analytics has been a neighborhood of sports entertainment for an extended time. During a match, we'd have seen the score line showing the probability of the team winning supported this match situation. We've implemented a forecasting model to predict IPL match

Results. Random Forest model is employed for match analysis. We've used a number of the features like City, Team1, Team2, Toss winner, Winner for prediction with higher accuracy. The forecasting of sports prediction has not only used as entertainment, but also helps for gameplay assessment of players, teams, leagues, and thus the associated results.[7] Furthermore, unfolding deciding of respective coaches and staffs, financial success by enlarging revenue of stadium are also emitted.

II. PROBLEM DEFINITION:

The proposed system implemented a forecasting model to predict IPL Match Results. Random Forest model is employed for match analysis. We've used a number of the features like City, Team1, Team2, Toss winner, Winner for prediction with higher accuracy.

III. LITERATURE SURVEY:

Siddharth Sinha et al [1] This model is employed for predicting the result of the match supported historic data. During the extraction of features various features has been involved but most vital features has been taken during prediction. They also made a team structure in terms of slots which defines most vital slots contributing to match winning and a ranking system for the players through their performance statistics. They used K-means to cluster all players consistent with their performance and KNN (K-nearest neighbor) is employed to seek out interchangeable player to a specific player. SVM model was trained using linear, polynomial and RBF (Radial Basis function). Thus, they preferred SVM with RBF kernel for prediction.

Rameshwari Lokhande et al [2] The problem of churning is addressed by using interactive models of Predictions where a user predicts the results of each game in order to be rewarded which would further help him strengthen his Fantasy squad. The project thus, aims not only to attract more users to this game that is Fantasy Cricket, but also aims at improving the general attraction to the Premier League. This happens because in a predictive model, a user makes a prediction on every game, and ends up watching that game to check if his prediction is going right.

Rabindra Lamsal & Ayesha Choudhary et al [3] It includes the varied factors that influence the result of an Indian Premier League matches were identified. The seven factors which significantly influence the results of an IPL match include the house team, the away team, the toss winner, toss decision, the stadium, and therefore the respective teams' weight. Hence designing machine learning model for predicting the match outcome of an auction based 2020 format premier league with the accuracy of 72.66% & F1 score of 0.72 is very satisfactory at this stage.

Rajesh Goel et al [4] They considered 692 matches of all the seasons (2008–2019) to train their model and all the matches of season (2019) to test. In future we can start right from the beginning of the first inning. Although our study is done on IPLT twenty matches only, the however similar approach could be applied to predict outcome in other versions of Cricket matches as336 well i.e., test cricket and ODI matches.

Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David et al [5] the target of this research was to predict the match winner of IPL using historical data of IPL from season 2008 to 2017. SEMMA methodology has been selected for conducting the analysis of IPL T20 match winner dataset. Preprocessing has been done on the dataset to form it consistent by removing missing value, encoding variables into numerical format. Best features were selected by visualizing attributes of knowledge with target variable. On selected features several machine learning models has been applied on the to predict the winner and therefore the results were outstanding. Decision Tree model was applied which predicted the match winner with good accuracy 76.9%.

Srikantaiah,Aryan Khetan1 et al [6] They are trying to find out the match winner of an IPL match based on the stadium they are choosing and the toss decision using machine learning techniques like SVM, Random Forest, Logistic Regression etc.

IV. PROPOSED SYSTEM:

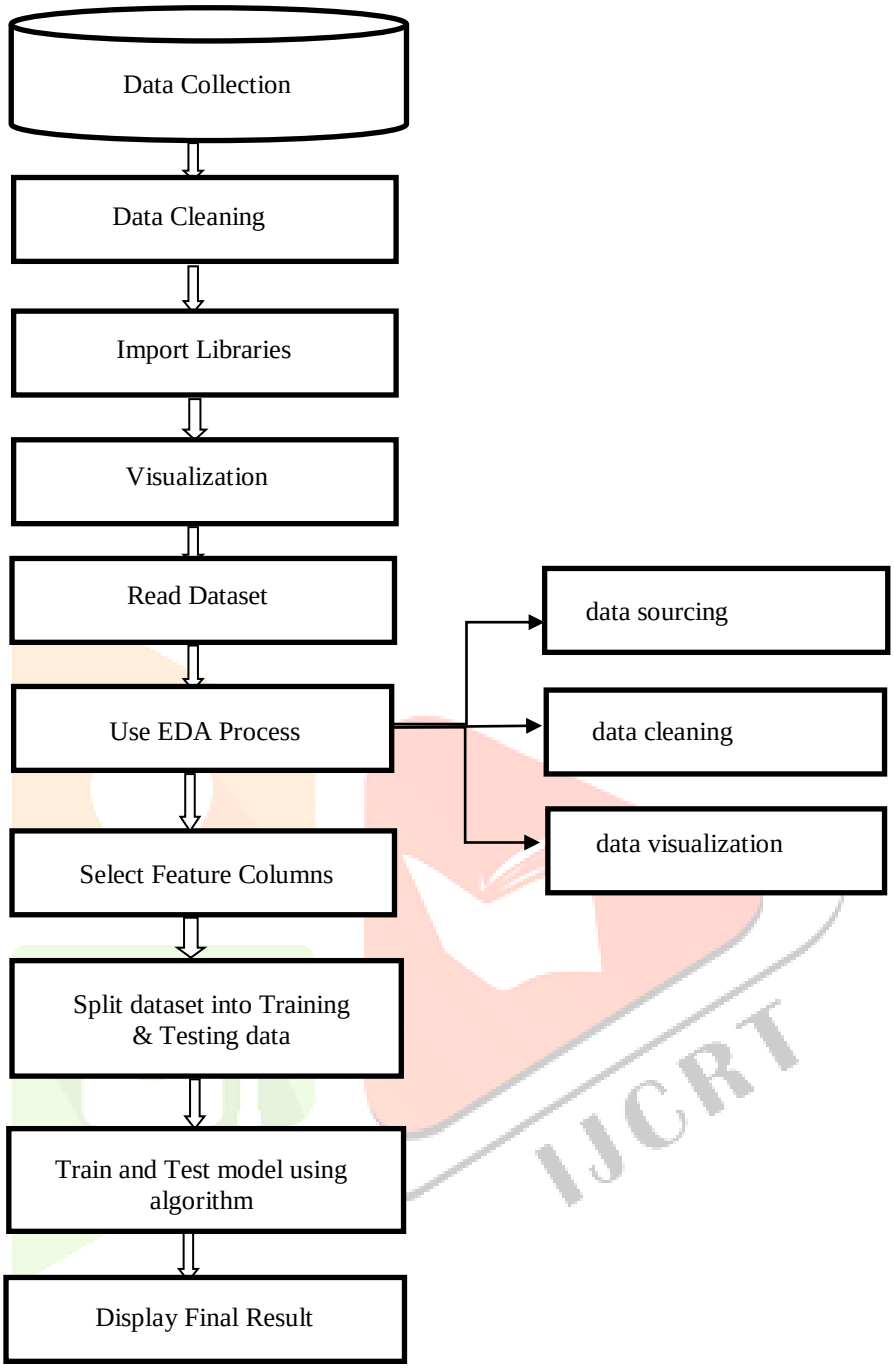

Fig: System Architecture

In first stage we collected dataset of records of previous matches played between 2008 to 2016 from Kaggle.com. Then we imported necessary libraries that are Pandas, Matplotlib, Seaborn.

After that we performed visualization of data to understand it better, we visualized data using matplotlib and seaborn libraries in bar graph format. In next stage we read dataset and displayed it in tabular format.

After that we used EDA process (Exploratory Data Analysis) to draw meaningful pattern and insights to prepare dataset for analysis by removing irregularities in the data.

In EDA we performed data sourcing for finding and loading data into our system, we used data.info () which loads all information about data i.e., count of entries, data columns, non-null counts & datatype of each column. After that we described data, cleaned data. The process of cleaning data includes checking null values and delete them, finding unnecessary columns and remove those columns to avoid inappropriate prediction. Then we performed data visualization of cleaned data.

Next stage is selection of feature columns in our case feature columns are team1, team2, toss winner, city and winner, we used team1, team2, city and toss winner as independent variable or we can say as input variables and winner as dependent variable i.e., output variable.

In next stage we have split dataset into training and testing dataset, we have train ML model using 80% training data and test using 20% testing data. By applying algorithm system will predict result.

V. METHODOLOGY:

5.1 Loading Dataset

The dataset name is matches.csv (IPL Matches data from 2008 to 2016) whose size is 132 kb and it is taken from the Kaggle Repository. The number of attributes is 17 and total number of records is 757. 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.

5.2 Data Collection and Cleaning:

The dataset is downloaded from the given site; https://www.kaggle.com/robinreni/cricket-data-set. The collected data preprocessed first so that we can remove unwanted data.

There are some significant steps in data preprocessing in Machine Learning:

1)Acquire the dataset- Acquiring the dataset is the first step in data preprocessing in machine learning. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. There are several online sources from where you can download datasets, we have downloaded data set from kaggle.com.

2)Import all the crucial libraries- Importing all the crucial libraries is the second step in data preprocessing in machine learning. The three core Python libraries used for this data preprocessing in Machine Learning are: NumPy – it is used for inserting any type of mathematical operation in the code. Pandas – we have used pandas for importing and managing dataset for e.g., we have given input as pd. read_csv file. Matplotlib– Matplotlib is a Python 2D plotting library that is used to plot any type of charts in Python so we use to display charts as result.

3) Import the dataset: In this step, need to import the datasets that we have gathered. Before importing datasets, set the current directory as the working directory. Then

a) Save Python file in the directory containing the dataset.
b) Go to File Explorer option in Jupyter IDE and choose the required directory.
c) Now, click on the F5 button or run option to execute the file.
d) Cleaning Data set: We are using pd. read csv () to read the CSV file.

5.3 Feature Engineering:

After data cleaning the key-features extraction is considered as the important part of the feature engineering. Only key features that are quite natural for predicting the cricket match score are extracted and understand. Here are some of the main features used that are listed as follows: 1) City 2) Team1 3) Team2 4) Toss Winner 5) Winner.

5.4 Build Prediction Model:

The system builds a prediction model by using Random Forest. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting and takes the average to improve the predictive accuracy of that dataset. We have trained the model with the city, Team1, Team2, Toss Winner, Winner and so on to predict the future of the cricket match.

VI. MATHEMATICAL MODEL:

When performing Random Forests based on classification data, we are often using the Gini index, or the formula used to decide how nodes on a decision tree branch.

$$\textbf{Gini} = 1 - \sum_{i=1}^{C} (Pi)^2$$

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, *pi* represents the relative frequency of the class you are observing in the dataset and *c* represents the number of classes. You can also use entropy to determine how nodes branch in a decision tree.

$$\textbf{Entropy} = \sum_{i=1}^{C} - p_i * (P_i)$$

Entropy uses the probability of a certain outcome in order to make a decision on how the node should branch. Unlike the Gini index, it is more mathematical intensive due to the logarithmic function used in calculating it.

VII. USED ALGORITHM:

7.1 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 ensemble-based method used for classification, regression and other tasks. The package random Forest is used which contains the functions sample (), random Forest () and plot () that are used to obtain the results of the Random Forest algorithm.

VIII. RESULT:

ANALYSIS OF ALGORITHMS:

Sr.No	Algorithm Used	Accuracy
1.	Naive Bayes	50%
2.	Support Vector Machine (SVM)	55%
3.	Decision tree	48%
4.	Random Forest	60%

IV. ANALYSIS OF DATASET:

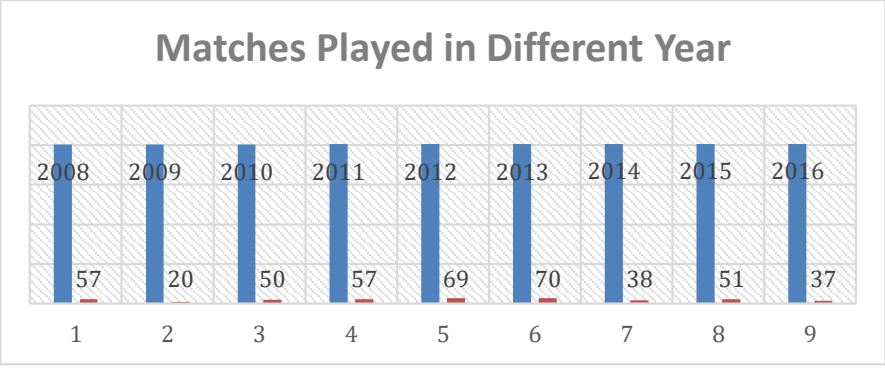

Fig 1: Matches Played in Different Year

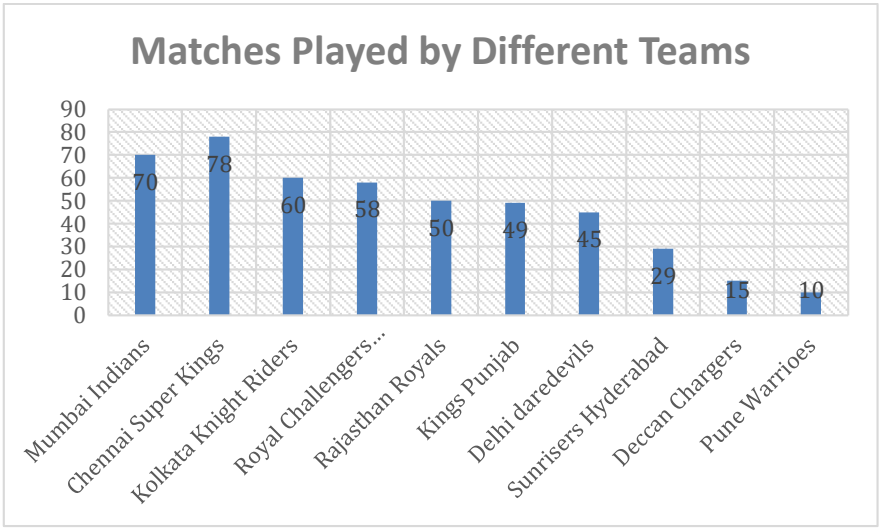

Fig 2: Matches Played by Different Teams

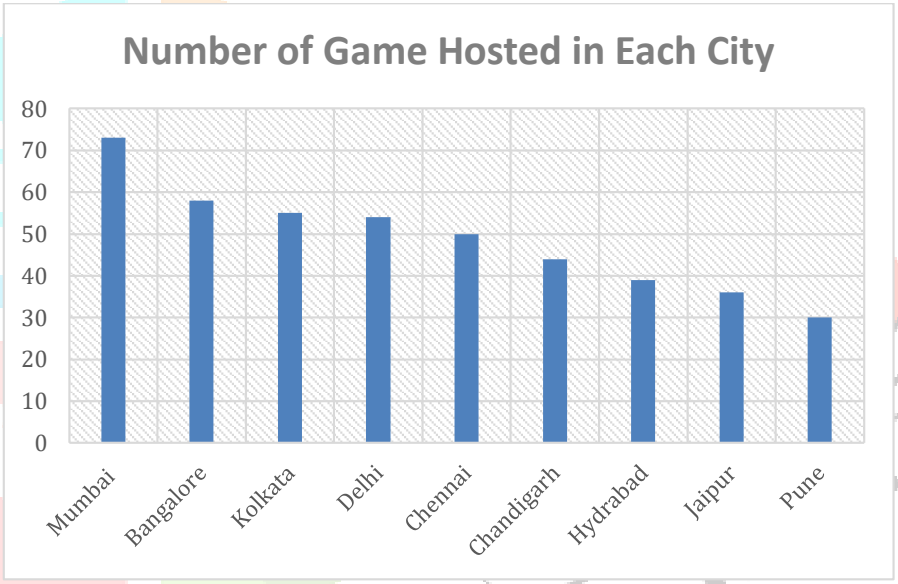

Fig 3: Number of Game Hosted in Each City

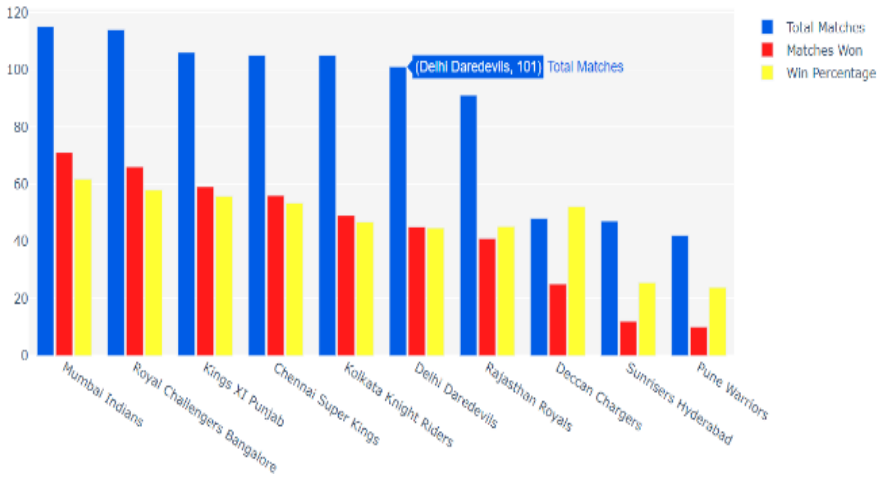

Fig 4: Match Played, Wins & Win Percentage

X. CONCLUSION:

The research focused on predicting the winner for an IPL match using machine learning and utilizing the available historical data of IPL from season 2008-2016. In this process, different Data Science methods were adopted to conduct the study, including data mining, visualization, preparation of database, feature engineering, applying the Analytic hierarchical process, creating prediction models, and

Training classification techniques. The IPL dataset was gathered and pre-processed. The missing values were removed, and variables were encoded into the numerical format to make the dataset uniform. The important features were then derived from data using the domain knowledge to extract raw data features via data mining techniques, and the results were derived from the model. As the dataset that is available for IPL is limited and small, multiple levels of features were created to make sure that the derived model is not underfit. Almost every feature that can affect the result of a match was derived. A number of machine learning models were applied to the selected features to predict the IPL match results.

XIII. REFERENCES:

1) Siddharth Sinha "IPL Win Prediction System To Improve Team Performance using SVM" IJFGCN Vol. 13, (2020).

2) Rameshwari Lokhande "Prediction of live cricket score and winning" IJRTD Volume 5(4), (2018).

3) Rabindra Lamsal and Ayesha Choudhary "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning" ResearchGate (2020).

4) Rajesh Goel "Dynamic Cricket match outcome prediction" Journal of Sports Analytics 2021.

5) Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David "Cricket winner prediction with application of machine learning and data analytics" IJSTR Volume 8, Issue 09, September 2019.

6) Srikantaiah, Aryan Khetan1, Baibhav Kumar, Divy Tolani, Harshal Patel "Prediction of IPL Match Outcome Using Machine Learning Techniques." Intelligent Computing Communication & Security (ICIIC 2021)

7) Nikhil Dhonge, Shraddha Dhole, Nikita Wavre "IPL Cricket Score & Winning prediction using machine learning ticniques" IRJMETS

8) 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.

9) 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.

10) 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.

11) 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. 430-435, 2019.

12) 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.

13) 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.

14) 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.