ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 9, Issuse.1 March 2021

Using Machine Learning to Analyze and Predict the Indian Premier League Winner

DR SAROJ KUMAR ROUT

Abstract

The Indian Premier League, which began in 2008, is the most popular Twenty20 cricket tournament in the world. Predicting the outcomes of the IPL is crucial if it is to be more successful, given the league's massive audience. Time series analysis and other machine learning algorithms and techniques that need less domain expertise may be used to anticipate future outcomes. Using Machine learning techniques, one must analyze data by looking back at the past and drawing inferences. Since there is so much interest in each new IPL season and its eventual champions, the proposed approach for forecasting matches must be reliable. Commercial sectors also use data analytics to arrive at the most accurate judgments. In this study, we use machine learning methods to predict outcomes based on a variety of factors, including match location, whether or not the batting team won the toss, and batsman strike rate. Data sets from the last seven years are collected using the aforementioned criteria, and the collected data is then preprocessed. In this case, we accurately predicted outcomes using Machine Learning Algorithms like Random Forest and Logistic Regression. The data has to be thoroughly explored and analyzed before any predictions can be made.

I INTRODUCTION

Specifically, real-world engineering challenges may be handled by using the branch of AI known as machine learning. This method relies only on data learning, in which a computer takes in information from the past in order to make predictions about the

Future. Decision trees, heuristic learning, knowledge acquisition, and mathematical models are all assets to the field of Machine Learning. As a result, it's reliable and efficient. Cricket may be played in a variety of formats, including the traditional Test match, the shorter Twenty20 International, and the longer one-day format. There are several such events, and the IPL is one of the most well-known. Cricket League played in India with the goal of fostering new, talented cricketers. Games are played over the course of 20 over's. The League was held annually in India during the months of March, April, or May. An auction determines which eight towns will be represented by the eight teams. To win the trophy, these groups must defeat one another. Many factors, including the teams' fortune and the individual players' abilities, will determine the outcome of the game. The forecast may also be affected by the result of the previous day's match. More people's interest and attendance means more money for the organizers and other participants. The magnitude of the data we collect for analysis and the records that are taken to anticipate the result are two factors that contribute to the reliability of the data.

PROFESSOR, Mtech,Ph.D
Department of CSE
Gandhi Institute for Technology, Bhubaneswar.

The use of Machine Learning Analytics for studying and forecasting sporting events has become the most effective method currently available. Using these methods, success has been attained in a number of different sports. Machine learning makes use of a wide variety of methods, each of which is deployed in order to make predictions about a given dataset using the given parameters. A total of 636 records were considered, and the models that suited them were used. Training the models and doing data preprocessing on the data eliminates the noise. Half of the data is used to train models, while the other half is utilized to assess the accuracy of those models. Accuracy is a metric that may be used to determine whether or not the model has correctly predicted the target variable.

II LITERATURE SURVEY

Considering how well followed the IPL is, there is a great deal of effort put into attempting to foretell its outcomes. Random Forest, Support vector machine, KNN, Logistic Regression, Naive Bays, etc. are only few of the models that have been employed in published works. Many studies have been published on IPL; however inaccurate findings have been avoided by using inconsistent data. Using Base, the authors of [1] are able to make a prediction on the winner, which is then used in conjunction with machine learning methods. This resource is useful for planning and deciding which players to bid on in the next auction.

The study [2] takes into account the total strength of a team by measuring its members' individual contributions. Multivariate regression is a method that takes into account the relative importance of each player based on their previous performances. To make accurate predictions, six distinct machine learning models were developed and fine-tuned. The most precise results were obtained using Random forest. The purpose of this study is to use machine learning methods to analyze the data and choose a winner. [3] Existing data mining methods are used to evaluate the results of an IPL on both symmetric and asymmetric data sets. Oversampling is used to correct for the imbalance in the dataset before any algorithms are applied to it. Accuracy of results is employed as the performance measure in this case, and algorithms are used for the calculation.

The purpose of this study is to utilize historical data to make predictions about the player's performance in terms of ball-to-ball analysis. The information from the preceding IPL is extracted, sorted, and examined. [4] Specifically, they have built models of the issue using recurrent neural networks and the Hidden Markova model. With these models, we have seen improved precision and efficiency.

In order to foretell the results of the game, the authors of this article used the Deep mayo predictor model. [5] Different parts of this model are built using advanced analytics techniques in machine learning. For all 30 matches, the Deep mayo prediction model has shown excellent accuracy. The effectiveness of the model may be improved by utilizing a more extensive dataset. In this study, we apply a multivariate regression method to evaluate the team's performance. [6] The odds from previous games are used to forecast the winner of the future match. The model was fit to seven dataset variables, and predictions were made. This article employs many distinct machine learning model types.

In this study, social media tweets are used as a data source. In this case we also make use of machine learning models, one of which is based on data from

Twitter (see [7]). Incorporating social media data like as tweets, these models have shown to be the most accurate after being updated after each of the match's 10 over’s. This study relies on reliable data provided by logistic regression and support vector machine.

III PROPOSED SYSTEM

Elements of the Proposed System include: • Data Collection • Data Cleaning

Information Display

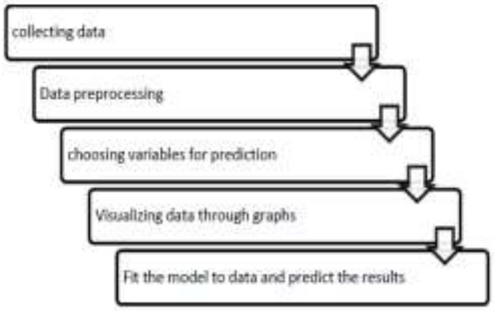

collecting data
Data preprocessing
choosing variables for prediction
Visualizing data through graphs
Fit the model to data and predict the results

Fig 1: process of predicting IPL team

3.1 Data Collection

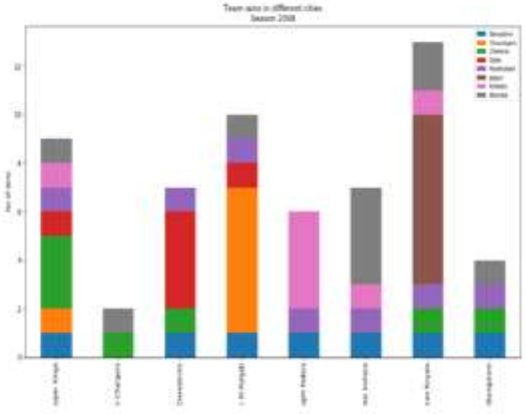
The Analysis uses data collected over the previous seven years (2008–2017) to choose which variables to use. To get the data in the right format, we utilize a library called pandas. Data analysis and manipulation may be performed using this open-source software. There were around 636 matches included for analysis and prediction.

3.2 Data Cleaning

In order to get accurate results, data must be "cleaned," or corrected, by deleting any false information. It is necessary to clean the gathered data sets of the noise they include, which includes null values and inappropriate values. So, we zero out any missing information and sort everything into neat little columns for easy analysis.

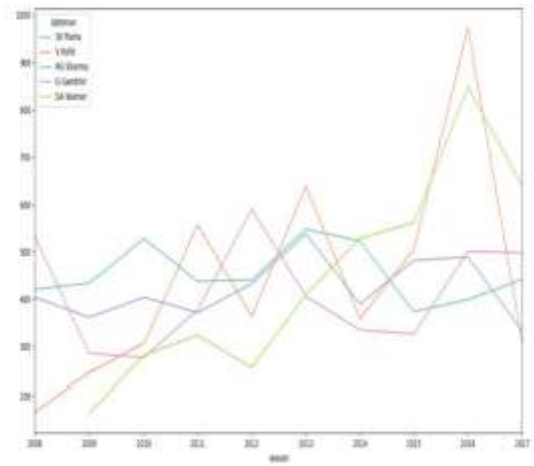
3.3 Data Visualization

Collecting data and using it to create visuals helps people grasp complex concepts. Here, we use the Matplotlib Library to display graphs of home-team wins, away-team losses, and overall winning percentage by city, stadium, and player from 2008 through 2017. Fig. 1 displays a graph of 2017 win percentages and city-specific win percentages for professional sports teams.



Team victories broken down by arena (Fig. 2).

A graph depicting the players' success rate from previous seasons is shown below. One of the most important factors in determining a team's success in the next season is the player's strike rate. If a player has a high strike rate, it's likely that his or her club will be competitive. Figure 2 below displays the success rate.



See Fig. 3 for a look at how often a certain player has scored in previous years.

We can't fully grasp the answer without the data visualization. All of the following charts were created using data from prior IPL seasons, which ran from 2008 to 2017.

IV PREDICTING THE RESULTS WITH THE HELP OF MODELS

After the necessary data is fitted into the necessary models for the forecast, the prediction is made. Both the Random Forest and support vector machine models are used, however the latter has shown worse results in this case.

4.1 The Random Forest Classifier the Random Forest classifier is a supervised learning technique that may be used for classification and regression. Data samples are processed using decision trees in a random forest, which finds the optimal answer. Primarily, it is used to issues of categorization. Here, the random forest has provided the highest accuracy for the considered variables. The random forest is an efficient tool for dealing with the issue of over fitting the data.

4.2 Classification Using a Support Vector Machine

The classification and regression issues are best suited to the Support Vector Machine classifier. It is a supervised learning algorithm that is taught new information by looking at examples from the past. The points for the classification process are put in a hyper -plane using a support vector method. Classification is carried out in the n-dimensional space because the hyper plane effectively distinguishes between the two groups. Support vector machines are most useful for classification, where they excel.

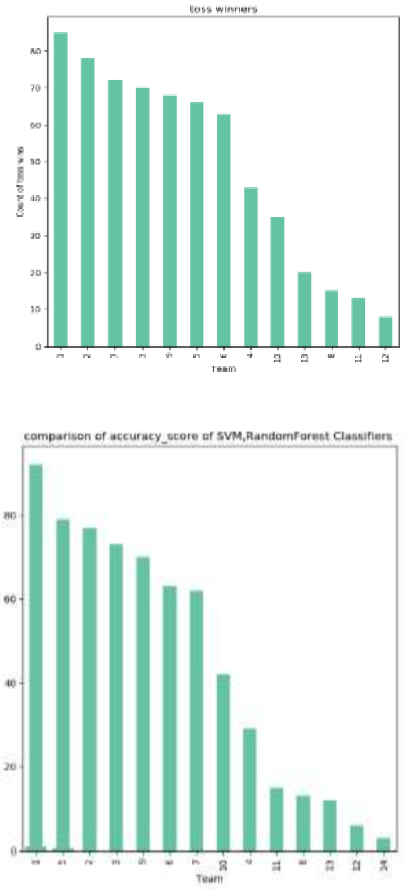
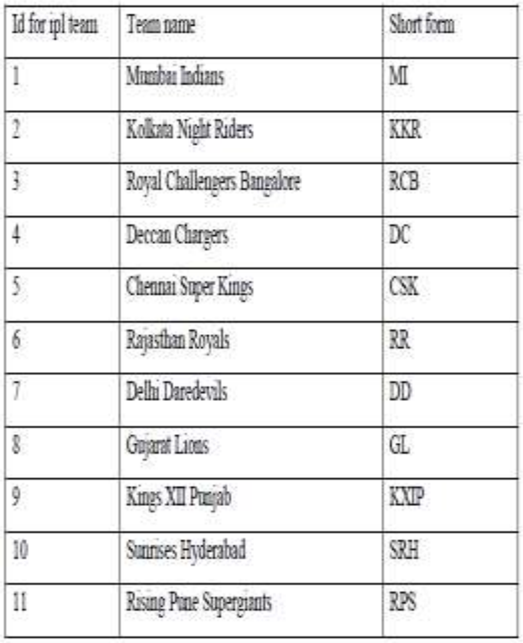


Figure 4: A Head-to-Head Evaluation of Random Forest and Support Vector Machines for Predicting the Winners of Coin Tosses

When calculating the outcomes, we included in the team that won the coin toss as one of the parameters using both the Random forest and support vector machine techniques. In fig. 3, we can see the forecast made by the algorithms using the historical data and the attribute toss winning.

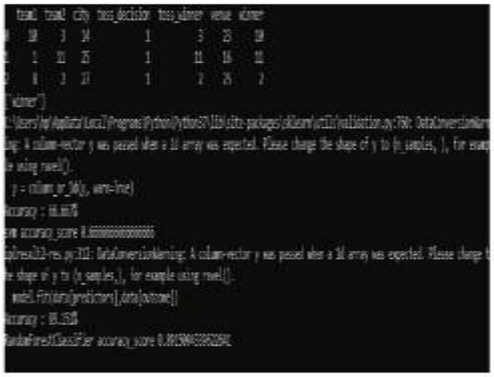
V. Results

In this case, we used random forest and support vector machine methods. With the settings we've used, random forest has shown to be the most effective. Accuracy of 89% was achieved using the random forest, whereas accuracy of 66% was achieved using the support vector model. Here, we've considered the outcomes of a coin toss, the bettor's choice, and the location.



Id for ipl team	Team name	Short form
1	Mumbai Indians	MI
2	Kolkata Night Riders	KKR
3	Royal Challengers Bangalore	RCB
4	Deccan Chargers	DC
5	Chennai Super Kings	CSK
6	Rajasthan Royals	RR
7	Delhi Daredevils	DD
8	Gujarat Lions	GL
9	Kings XII Punjab	KXIP
10	Sunrises Hyderabad	SRH
11	Rising Pune Supergiants	RPS

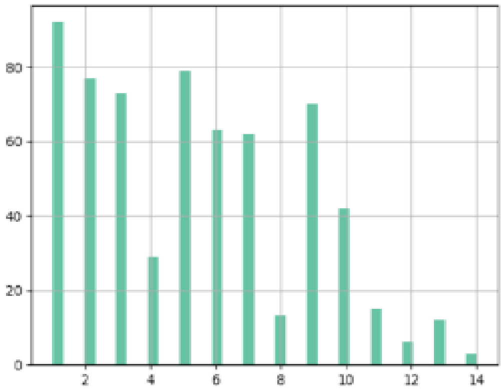
In this table, the various teams are given with the numeric values that represent them. In addition to the parameters, the venue was also encoded. Varying places have different financial worth.



```
   team1  team2  city  toss_decision  toss_winner  venue  winner
0     10      3    14              1            3     23      10
1      1     11    25              1           11     16      11
2      8      3    17              1            2     25       2
['winner']
C:\Users\[illegible]\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\utils\validation.py:760: DataConversionWarn
ing: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for examp
le using ravel().
  y = column_or_1d(y, warn=True)
Accuracy : 66.667%
svm accuracy_score 0.6666666666666666
iplresult2-res.py:111: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change t
he shape of y to (n_samples,), for example using ravel().
  model.fit(data[predictors],data[outcome])
Accuracy : 89.153%
RandomForestClassifier accuracy_score 0.8915966386[illegible]
```

What we found and what we learned, shown in Fig.

It can be seen in the first record of fig. 5 that SRH is represented by the number 10 in team1, whereas RCB is shown by the number 3 in team2. There are two possible values for toss decision: 1 for batting and 0 for bowling. The number 23 under "Venue" is the unique identifier for this location. If the toss determines that a team will play at a given location, and that team has a perfect record there, then the value 10 will be used.



Prediction outcome shown in Fig. 6

Final results for predicting the winning team from all of the collected data are shown in Fig. 6. Due to the inclusion of data from 2008–2017, the projection indicated that the Mumbai Indians, with the enablement of 1, will be the next IPL champions.

Conclusion:

The squad you choose and how well they play together will have a significant impact on whether or not you win the game. There are a number of variables that might affect the outcome of a cricket match than just how well each team performs. The unpredictable nature of each game makes it difficult to make accurate predictions for the IPL. The paper's primary motivation comes from the author's attempt to foretell the future by analyzing historical facts. In this study, we apply three distinct categorization techniques to make our forecasts. Python scripts serve as the implementation's primary tool. The accuracy of random forest was 89%, which was higher than the accuracy of support vector system (66%). Using this data, we can better anticipate future winners and choose winning squads.

References:

[1] Passim, Kalpdrum & Pander, Niravkumar. (2018) "Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning" 111-126. 10.5121/csit.2018.80310.

[2] I. P. Wickramasinghe et. al, "Predicting the performance of batsmen in test cricket," Journal of Human Sport & Exercise", vol. 9, no. 4, pp. 744-751, May 2014.

[3] R. P. Schumaker, O. K. Solieman and H. Chen, "Predictive Modeling for Sports and Gaming" in Sports Data Mining, vol. 26, Boston, Massachusetts: Springer, 2010.

[4] J. McCullagh, "Data Mining in Sport: A Neural Network Approach," International Journal of Sports Science and Engineering, vol. 4, no. 3, pp. 131-138, 2012.

[5] Bunker, Rory & Thabtah, Fadi. (2017) "A Machine Learning Framework for Sport Result Prediction. Applied Computing and Informatics", 15. 10.1016/j.aci.2017.09.005. [6] Ramon Diaz-Uriarte and Sara, "Gene selection and classification of microarray data using random forest, BMC Bioinformatics", doi:10.1186/1471-2105-7-3

[7] Rabindra Lamsal and AyeshaChoudhary, "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning"

[8] Akhil Nimmagadda et. Al, "Cricket score and winning prediction using data mining", IJARnD Vol.3, Issue3.

[9] Ujwal U J et. At, "Predictive Analysis of Sports Data using Google Prediction API" International Journal of Applied Engineering Research", ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2814-2816.

[10] Rameshwari Lokhande and P.M.Chawan, "Live Cricket Score and Winning Prediction", International Journal of Trend in Research and Development, Volume 5(1), ISSN: 2394-9333.