ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

Analyzing and Predicting outcome of IPL Cricket Data

D. Jyothsna [1], K. Srikanth [2]

Assistant Professor, Department of Computer Science and Engineering, JBIET, Hyderabad, Telangana, India[1]

Assistant Professor, Department of Computer Science and Engineering, JBIET, Hyderabad, Telangana, India [2]

ABSTRACT*:* In today's date data analysis is need for every data analytics to examine the sets of data to extract the useful information from it and to draw conclusion according to the information. Data analytics techniques and algorithms are more used by the commercial industries which enables them to take precise business decisions. It is also used by the analysts and the experts to authenticate or negate experimental layouts, assumptions and conclusions. In recent years the analytics is being used in the field of sports to predict and draw various insights. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcome of matches is very important for all stake holders.

In this paper, the past seven years data of IPL containing the players details, match venue details, teams, ball to ball details, is taken and analyzed to draw various conclusions which help in the improvement of a players performance. Various other features like how the venue or toss decision has influenced the winning of the match in last seven years are also predicted. Various machine learning and data extraction models are considered for prediction are Linear regression, Decision tree, K-means, Logistic Regression. The cross validation score and the accuracy are also calculated using various machine learning algorithms. Before prediction we have to explore and visualize the data because data exploration and visualization is an important stage of predictive modeling.

KEYWORDS*:* Machine Learning, Random Forest Classifier, Visualization, Cricket, Prediction

I. INTRODUCTION

Machine learning is a branch of artificial intelligence that aims at solving real life engineering problems. It provides the opportunity to learn without being explicitly programmed and it is based on the concept of learning from data [1]. It is so much ubiquitously used dozen a times a day that we may not even know it. The advantage of machine learning (ML) methods is that it uses mathematical models, heuristic learning, knowledge acquisitions and decision trees for decision making. Thus, it provides controllability, observability and stability.

The game of cricket is played in various formats, i.e., One Day International, T20 and Test Matches. The Indian Premier League (IPL) [7] is a Twenty-20 cricket tournament league established with the objective of promoting cricket in India and thereby nurturing young and talented players. The league is an annual event where teams representing different Indian cities compete against each other. The teams for IPL are selected by means of an auction. Players' auctions are not a new phenomenon in the sports world. However, in India, selection of a team from a pool of available players by means of auctioning of players was done in Indian Premier League (IPL) for the first time. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcome of matches is very important for all stake holders. This, in turn, is dependent on the complex rules governing the game, luck of the team (Toss), the ability of players and their performances on a given day. Various other natural parameters, such as the historical data related to players, play an integral role in predicting the outcome of a cricket match. A way of predicting the outcome of matches between various teams can aid in the team selection process [5]. However, the varied parameters involved present significant challenges in predicting accurate results of a game. Moreover; the accuracy of a prediction depends on the size of data used for the same. The tool presented in this paper can be used to evaluate the performance of players. This tool provides a visualization of players' performance. Further, several predictive models[3] are also built

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

for predicting the result of a match, based on each player's past performance as well as some match related data. The developed models can help decision makers during the IPL matches to evaluate the strength of a team against another. The contributions of the presented work are as follows:

- To provide the statistical analysis of players based on different characteristics
- To predict the performance of a team depending on individual player statistics
- To successfully predict the outcome of IPL matches

II. LITERATURE SURVEY

Kalpdrum Passi and Niravkumar Pandey discussed about the prediction accuracy in terms of runs scored by batsman and the no. of wickets taken by the bowler in each team[1]. I. P. Wickramasinghe proposed a methodology to predict the performance of batsman for the previous five years using hierarchial linear model [2]. R.P.Schumaker et. al, discussed about different statistical simulations used in predictive modeling for different sports[3]. John McCullagh implemented neural networks and datamining techniques to identify the talent and also for the selection of players based on the talent in Australian Football League[4]. Bunker et. al, proposed a novel sport prediction framework to solve specific challenges and predict sports results [5]. Ramon Diaz-Uriarte et. al, investigated the use of random forest for classification of microarray data and proposed a new method of gene selection in classification problem based on random forest [6]. Rabindra Lamsal and Ayesha Choudhary, proposed a solution to calculate the weightage of a team based on the player's past performance of IPL using linear regression [7]. Akhil Nimmagadda et. Al, proposed a model using Multiple Variable Linear Regression and Logistic regression to predict the score in different innings and also the winner of the match using Random Forest algorithm [8]. Ujwal U J et. Al, predicted the outcome of the given cricket match by analyzing previous cricket matches using Google Prediction API[9]. Rameshwari Lokhande and P.M.Chawan came up with live cricket score predicton using linear regression and Naïve Bayes classifier [10]. Abhishek Naik et. Al, proposed a new model used matrix factorization technique to analyze and predict the winner in ODI cricket match [11]. Esha Goel and Er. Abhilasha discussed the improvements in Random Forest Algorithmand described the usage in various fields like agriculture, astronomy, medicine, etc. [12].Amit Dhurandhar and Alin Dobra proposed a new methodology for analysing the error of classifiers and model selection measures to analyse the decision tree algorithm [13]. H. Yusuff et. Al, performed logistic regression using mammograms to find the accuracy with valid samples [14].

III. METHODOLOGY

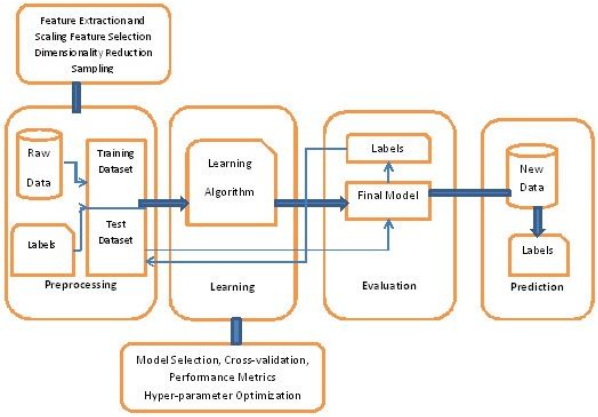

Fig. 1 System Architecture

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

The methodology consists of 4 main stages- Data Preprocessing, Data Cleansing, Data Preparation, Encoding the data. Initially, the seven IPL seasons real-time dataset is taken in CSV format. In data preprocessing phase, the data is incomplete, noisy and inconsistent. Data is to be filled with missing values and correct the inconsistencies. In data cleansing phase, data validation is done by maintaining consistency across the dataset and data enhancement is done by adding related information to the dataset [9]. The data preparation is significant for achieving optimal results. This involves choosing an outcome measure to evaluate different predictor variables. In data Encoding phase, label each term with short names and encode them as numerical values for predictive modeling as implemented below.

```
matches.replace(['MumbaiIndians','KolkataKnightRiders','RoyalChallengersBangalore','DeccanChargers','Chennai SuperKings','RajasthanRoyals','DelhiDaredevils','GujaratLions','KingsXIPunjab','SunrisersHyderabad','RisingPuneSupergiant','KochiTuskersKerala','PuneWarriors'],['MI','KKR','RCB','DC','CSK','RR','DD','GL','KXIP','SRH','RPS','KTK','PW' ],inplace=True)
 encode={
'team1':{'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW': 13},
 'team2':{'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13}
,
 'toss_winner':{'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13},
 'winner':{'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13,'Draw': 14}}
```

The test dataset is taken as input to the learning algorithm, evaluates different scenarios like toss winner, toss decision and team winner. Thus, new data is obtained with final prediction.

The scikit-learn LabelEncoder is a python library which converts categorical variables to numeric variables and a predictive model is created using a generic function called class_model that takes parameters model (algorithm), data, predictors input, and outcome predictable feature. Different multi-classification algorithms such as Logistic Regression[14], Decision Tree[13] and Random Forest[6][12] are implemented to predict the accuracy and cross-validation score. Out of these classification algorithms, Random Forest algorithm is proved to be the best accurate classifier.

Import models from scikit learn module:

```
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import KFold  #For K-fold cross validation from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
```

Generic function for making a classification model and accessing performance:

```
def classification_model(model, data, predictors, outcome): model.fit(data[predictors],data[outcome])
predictions = model.predict(data[predictors])
accuracy = metrics.accuracy_score(predictions,data[outcome])
print('Accuracy : %s' % '{0:.3%}'.format(accuracy))
kf = KFold(data.shape[0], n_folds=5)
error = []
for train, test in kf:
train_predictors = (data[predictors].iloc[train,:])
train_target = data[outcome].iloc[train]
model.fit(train_predictors, train_target)
error.append(model.score(data[predictors].iloc[test,:], data[outcome].iloc[test]))
print('Cross-Validation Score : %s' % '{0:.3%}'.format(np.mean(error)))
model.fit(data[predictors],data[outcome])
```

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

IV. RESULTS AND DISCUSSIONS

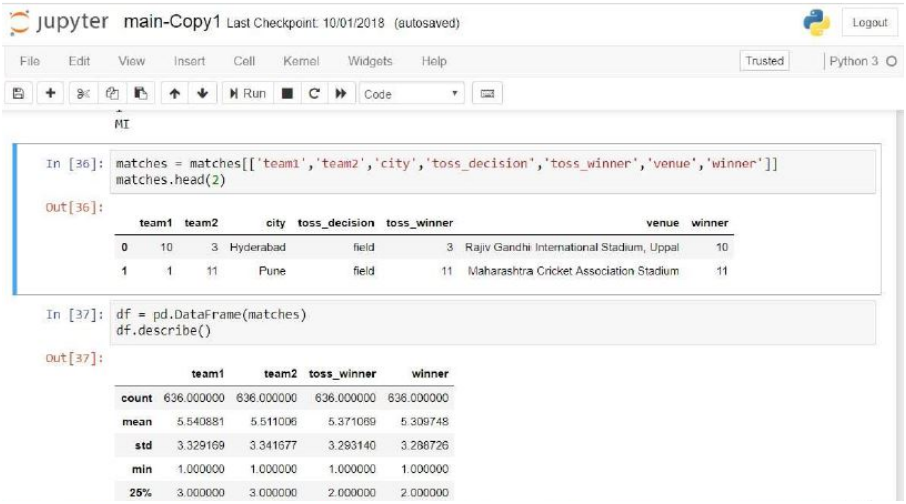

Fig.2.Probability of toss winning- 'head'

Discussion: This screenshot takes the different fields i.e, team1, team2, city, toss_decision, toss_winner, venue, winner and predicts the winner based on the toss_winner.

Random Forest Accuracy

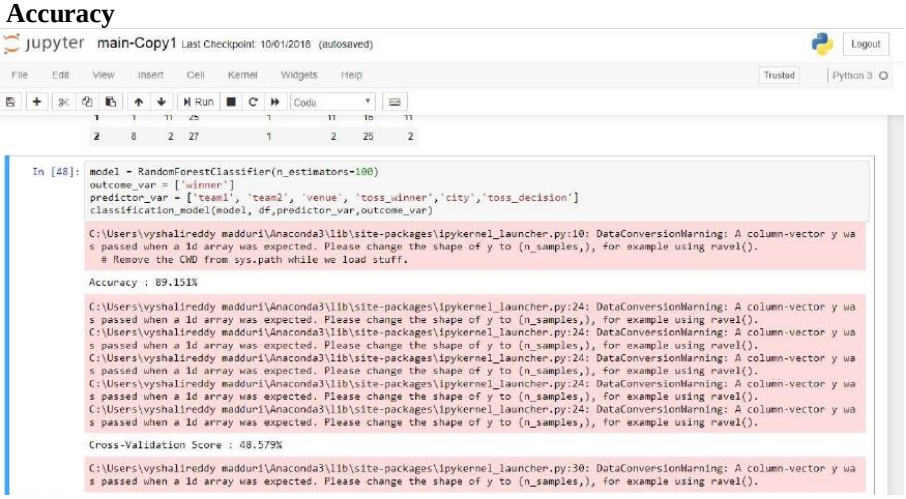

Fig. 3 Random Forest Accuracy and cross validation score

Discussion: This screenshot shows the accuracy and cross validation score of Random Forest Classifier algorithm by using a generic function classification model.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

Visualizations

Fig. 4 Visualizations

Discussion: This screenshot is a histogram that shows two different visualizations. First is toss winners and count of toss winners. Second is probability of match winning by winning toss.

V. CONCLUSION AND FUTURE SCOPE

Selection of the best team for a cricket match plays a significant role for the team's victory. The main goal of this paper is to analyze the IPL cricket data and predict the players' performance[1]. Here, three classification algorithms are used and compared to find the best accurate algorithm. The implementation tools used are Anaconda navigator and Jupyter. Random Forest is observed to be the best accurate classifier with 89.15% to predict the best player performance. This knowledge will be used in future to predict the winning teams[8][11] for the next series IPL matches. Hence using this prediction, the best team can be formed.

REFERENCES

[1] Passi, Kalpdrum & Pandey, 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.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

[11] Abhishek Naiket. Al, "Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques", IJETCS, vol. 3, issue 2, ISSN:2455-9954, April 2018.
[12] Esha Goel and Er. Abhilasha, " Random Forest: A Review", IJARCSSE, Volume 7, Issue 1, DOI: 10.23956/ijarcsse/V7I1/01113, 2017.
[13] Amit Dhurandhar and Alin Dobra, " Probabilistic Characterization of Random Decision Trees", Journal of Machine Learning Research, 2008.
[14] H. Yusuff et. Al, "BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION", IJRRAS, Vol. 10, Issue 1, 2012.