Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms

Article · February 2020

DOI: 10.22214/ijraset.2020.2121

1 author:

Sachi Priyanka
M.O.P.Vaishnav College for Women
1 PUBLICATION **0** CITATIONS

SEE PROFILE

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms

Priyanka S[1], Vysali K[2], Dr K B PriyaIyer[3]
[1,2]*Student, BCA, Department of Computer Science, M.O.P. Vaishnav College for Women, Chennai.*
[3]*Associate Professor, BCA, Department of Computer Science, M.O.P. Vaishnav College for Women, Chennai.*

Abstract: ***Cricket is one of the famous outdoor sports that contain a large set of statistical data in real world. As IPL games rise in popularity, it is necessary to examine the possible predictors that affect the outcome of the matches. In this paper, the several 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 player's performance. It focuses on measuring the outcome of Indian Premier League (IPL) matches by applying the existing data mining algorithms to the balanced as well as imbalanced dataset. This model is very much popular in predictive modelling. Currently, in Twenty-Twenty (T20) cricket matches first innings score is predicted on the basis of current run-rate which can be calculated as the amount of runs scored per the number of over's bowled. It includes factors like number of wickets fallen, venue of the match, toss and predicts the score in each of the innings and finally the winner of the match using Random Forest algorithm. In this paper, Prediction of IPL2020 are done on the basis of survey, and analysis are done based on data mining algorithms.***

Keyword: Data Mining, Prediction, T20, IPL, Decision Tree, Naïve Bayes, SVM – Support Vector Machine, Random Forest.

I. INTRODUCTION

Data mining tools predict the future trends and behaviours, which gives an opportunity to predict the outcome of an IPL (Indian Premier League) match using data mining algorithms. Data mining algorithms have been applied to the IPL dataset and the knowledge from each algorithm has been obtained and analyzed thoroughly as the results are obtained with good accuracy performance. Cricket is one of the most popular sports. The International Cricket Council (ICC) [10] out listed 106 cricket playing nations representing 10 belongs to the full members, 37 of them are associates, and the remaining 59 are considered to be affiliate members. The game of cricket is played in various formats, i.e., One Day International, T20 and Test Matches. The Indian Premier League (IPL) [9] is a Twenty-20 cricket tournament league established with the objective of promoting cricket in India and thereby nurturing young and talented players. 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. 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. A way of predicting the outcome of the matches between various teams can aid in the team selection process [10]. The tool presented in this paper can be used to evaluate in the performance of players. This tool provides a visualization of players' performance.

The result has been predicted using the algorithm approaches and have analyzed the results of the IPL match using the above approaches. Some of the popular variables considered in cricket literature are home-field advantage, coin-toss result, bat-first or second. Thus we measure the outcome of the IPL matches using the data-mining algorithms.

II. LITERATURE REVIEW

A. Describes about significant challenges that we face for accurate prediction including the various parameters which affect the outcome of the match. The ball movement gets changed from every over, so it is considered being important to predicting the outcome of each match on every ball. Here they had developed a model that predicts the match result of every ball played. \

B. Explains about the concept of identifying rising stars in cricket domain using some techniques. Rising stars can be predicted by both bats as well as bowling teams. Distinct features like concept of co-players, team and opposite teams are presented with their mathematical formulation.

C. Explained the outcome of ODI match depends on various factors. The list of key features is home-field advantages, winning the toss, game plan, venue and season. Logistic Regression, SVM are the different types of algorithm used for model building. Logistic Regression is applied for data that had been already obtained from previous matches. SVM used for predictive analysis. It was found that SVM was proved to be a better model based on both the parameters used to predict accuracy and model outcome.

D. 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.
E. Came up with live cricket score prediction using linear regression and naïve bayes classifier.
F. Proposed a new methodology for analyzing the error of classifiers and model selection measures to analyze the decision tree algorithm.
G. Proposed a model used matrix factorization technique to analyze and predict the winner in ODI cricket match.
H. Proposed a solution to calculate the weight age of a team based on players' past performance of IPL using linear regression.

III. ANALYSIS

1) *Rapid Miner:* it is a fast mineworker could be an information software system platform developed by the corporate of an equivalent name that gives AN integrated setting for information preparation, machine learning, deep learning, text mining and predictive mining.
2) *Decision Tree:* A decision tree could be a structure that has a root node, branches, and leaf nodes. Each internal node denotes a check on associate attribute, every branch denotes the result of a check, and each leaf node holds a label. The uppermost node within the tree is the root node. The general motive of the decision tree is to form a training model which may use to predict class or value of target variables by learning decision rules inferred from prior data.

Here in Fig-1, The dataset named match_data that has 746 rows and 24 attributes are connected and processed by using decision tree with a set role operator and displayed a decision tree graph as shown in Fig-2.

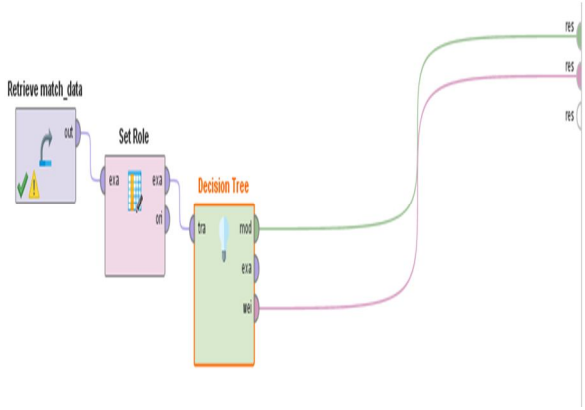


Fig-1:process

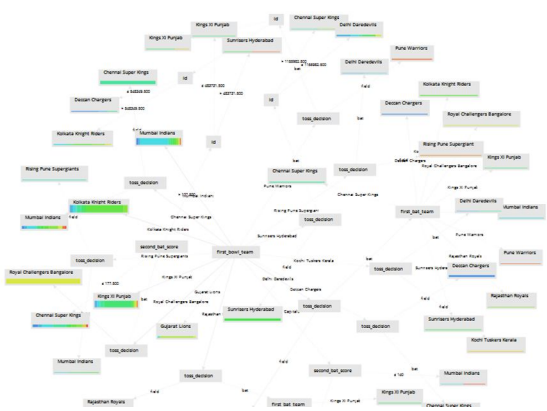


Fig-2:decision tree

3) *Naive Bayes:* Naïve Bayes algorithm is a classification technique based on bayes theorem with an assumption of independence among predictors. In simple terms, naïve bayes it has been successfully used for many purposes, but it works particularly well with NLP problems. This algorithm is used to predict the tag of a text.

Here in Fig-3, the dataset named match_data that has 746 rows and 24 attributes are connected and processed by using naïve bayes with set role operator and displayed using simple distribution in Fig-4 and displayed using a graph in Fig-5 and descritize as shown.

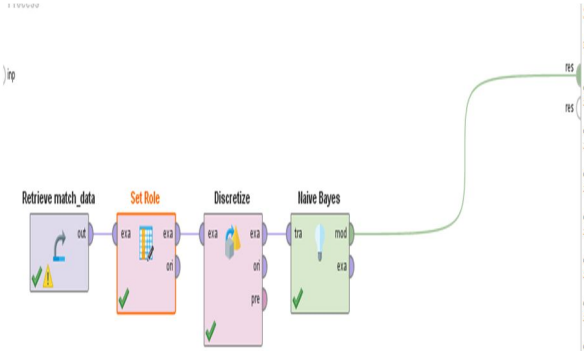


Fig-3: process

SimpleDistribution

Distribution model for label attribute team1

Class Royal Challengers Bangalore (0.130)
23 distributions

Class Delhi Daredevils (0.111)
23 distributions

Class Mumbai Indians (0.113)
23 distributions

Class Rising Pune Supergiants (0.009)
23 distributions

Class Chennai Super Kings (0.119)
23 distributions

Class Kings XI Punjab (0.113)
23 distributions

Class Kolkata Knight Riders (0.114)
23 distributions

Class Sunrisers Hyderabad (0.071)
23 distributions

Fig-4: Distributions of the graph

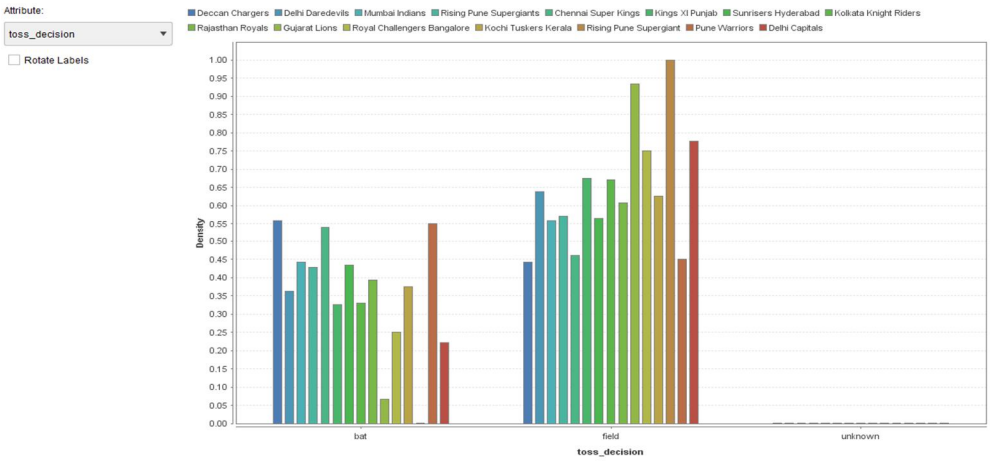


Fig-5: The distribution graph representing the toss decision.

4) Svm-Support Vector Machine: SVM are supervised learning methods used for classification and regression tasks that originated from statistical learning theory. As a classification method, SVM is a global classification model that generates partitions and usually employs all attributes.

Here in Fig-6 the dataset named match_data that has 746 rows and 24 attributes are connected and processed by using SVM with set role operator and displayed using as weighted visualization graph shown in Fig-8 and support vector visualization graph as shown in Fig-7.

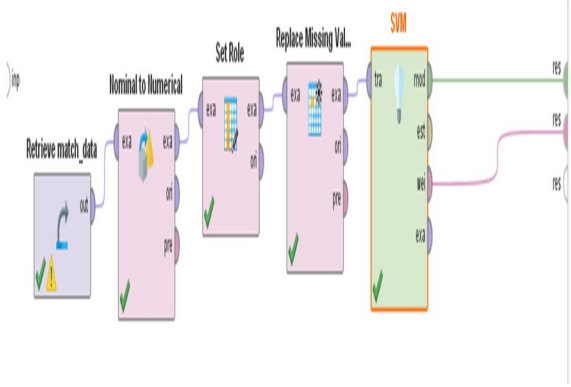


Fig-6:process

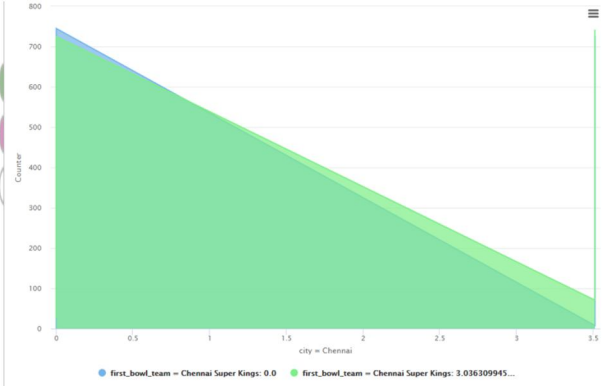


Fig-7:SVM graph

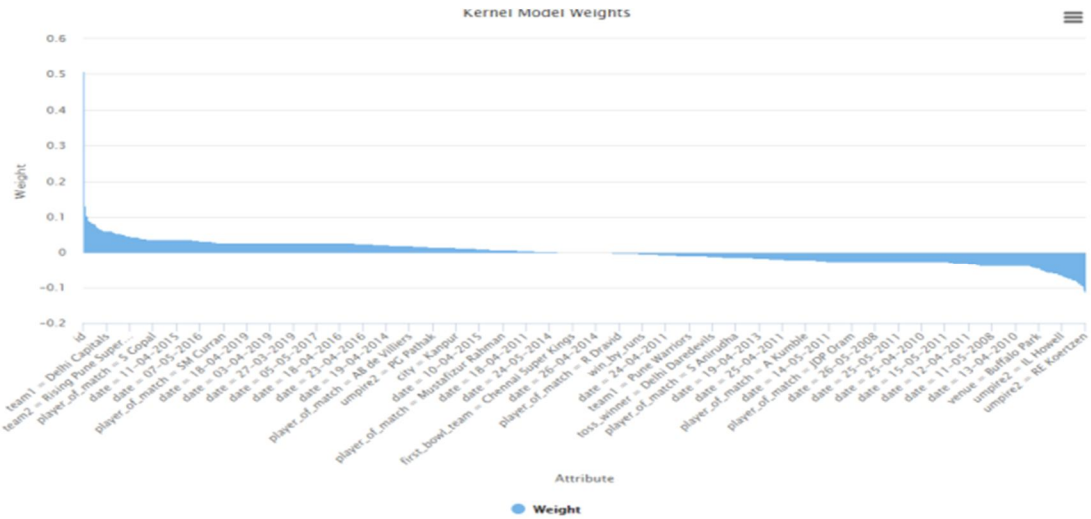


Fig-8: weighted graph

5) *Random Forest:* A random forest could be a meta computer that matches some identifiable call tree on numerous sub-samples of the dataset and use averaging to enhance the predictive accuracy over-fitting. Each call tree is made by using a random set of the training information.

Here in Fig-9, the dataset named match_data that has 746 rows and 24 attributes are connected and processed using Random forest algorithm with set role operator and displayed in the form of a tree as shown in Fig-10.

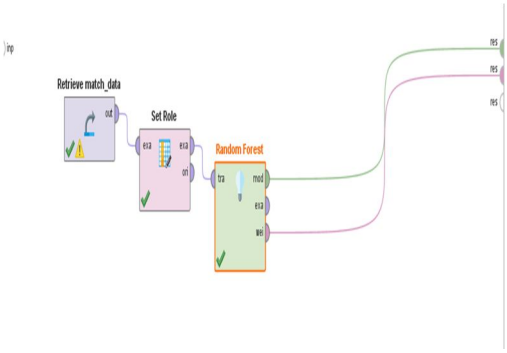


Fig-9: Process

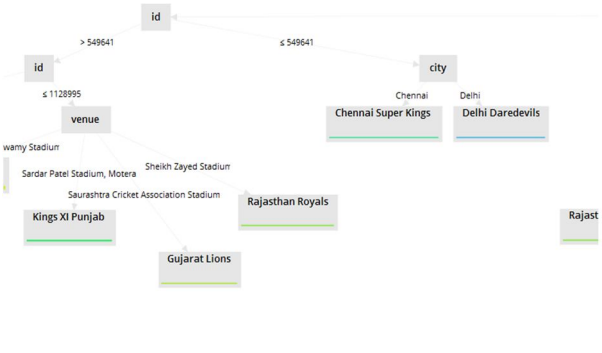


Fig-10: Random Forest Tree

Using GOOGLE FORMS did a survey for next IPL 2020 .107 people have attended the survey and answered on basis of the questions. Using those responses a new dataset named SURVEY FORM having 107 rows with 12 attributes. The dataset are predicted and analyzed using rapid miner tool and prediction are done using cross validation operator with each algorithm. The prediction results are as follows.

6) *Decision Tree:* Using decision tree , Cross validation process are done as shown in Fig-11. And accuracy table is been displayed as shown in Fig-12 . the survey chart is shown in Fig-13 and the performance vector is shown in Fig-14.

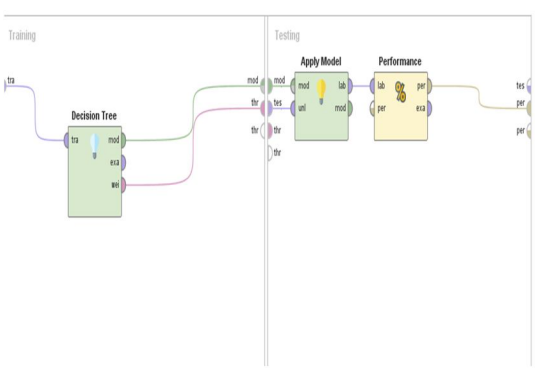


Fig-11:Cross validation

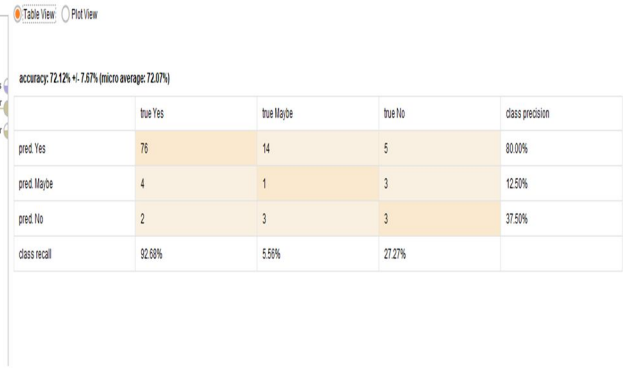


Fig-12:Accuracy table

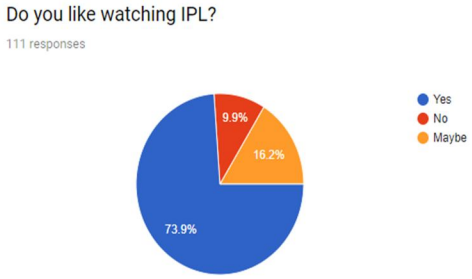


Fig-13: Survey chart

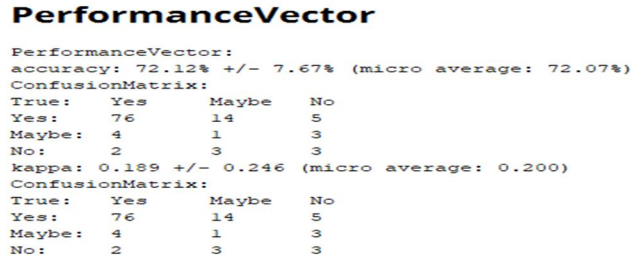


Fig-14:Performance vector of the accuracy tabl

The above figure shows the cross validation process followed by its accuracy percentage table of the graph shown above. It gives the accuracy of 72.12%.the other related measures are kappa 0.189,and predicting by means of which out of 100% responses 73.9% of people are responded that they like watching IPL.

7) Naïve Bayes: Using naïve bayes, cross validation process are done as shown in Fig-15, and the accuracy table are shown in Fig-16,the survey chart is displayed as shown in Fig-17 and performance vector of the accuracy table as shown in Fig-18.

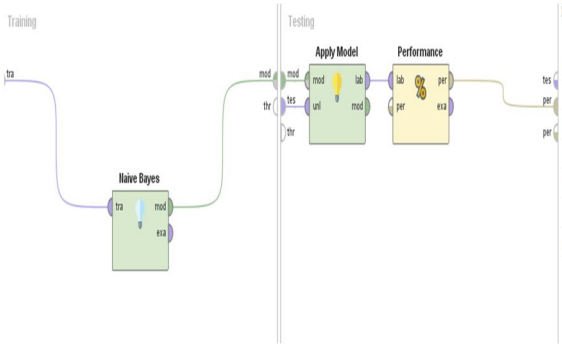
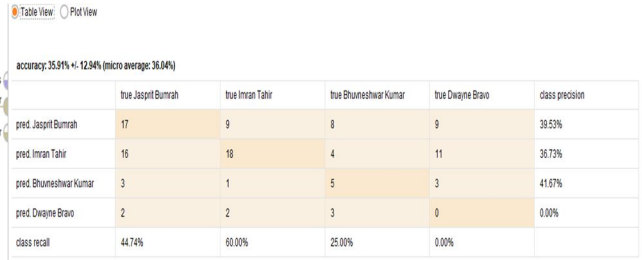


Fig-15:Cross validation



	true Jasprit Bumrah	true Imran Tahir	true Bhuvneshwar Kumar	true Dwayne Bravo	class precision
pred. Jasprit Bumrah	17	9	8	9	39.53%
pred. Imran Tahir	16	18	4	11	36.73%
pred. Bhuvneshwar Kumar	3	1	5	3	41.67%
pred. Dwayne Bravo	2	2	3	0	0.00%
class recall	44.74%	60.00%	25.00%	0.00%	

Fig-16:Accuracy Table

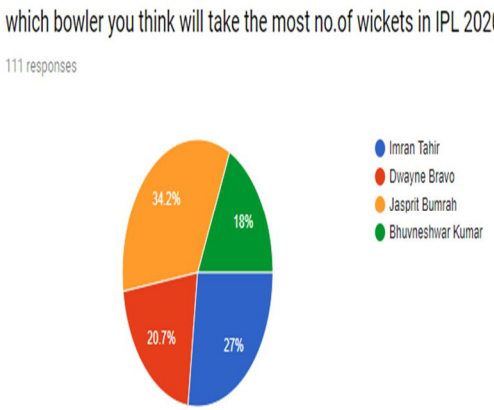


Fig-17:Survey Chart

```
PerformanceVector:
accuracy: 35.91% +/- 12.94% (micro average: 36.04%)
ConfusionMatrix:
True:   Jasprit Bumrah  Imran Tahir     Bhuvneshwar Kumar        Dwayne Bravo
Jasprit Bumrah: 17      9       8       9
Imran Tahir:    16      18      4       11
Bhuvneshwar Kumar:      3       1       5       3
Dwayne Bravo:   2       2       3       0
kappa: 0.098 +/- 0.176 (micro average: 0.106)
ConfusionMatrix:
True:   Jasprit Bumrah  Imran Tahir     Bhuvneshwar Kumar        Dwayne Bravo
Jasprit Bumrah: 17      9       8       9
Imran Tahir:    16      18      4       11
Bhuvneshwar Kumar:      3       1       5       3
Dwayne Bravo:   2       2       3       0
```

Fig-18:Performance vector

The above figure shows the cross validation process followed by its accuracy percentage table with graph shown above. It gives accuracy of 35.91%. The other related measures are kappa 0.098, and predicting by means of which out of 100% responses 34.2% of them are predicting that Jaspirit Bumrah will take the most number of wickets in IPL2020.

8) Random Forest: With the help of random forest algorithm, cross validation process is done as shown in Fig-19, and accuracy table is displayed as shown in Fig-20 . the survey chart used for the cross validation is shown in Fig-21, and performance vector of the accuracy table is displayed as shown in Fig-22.

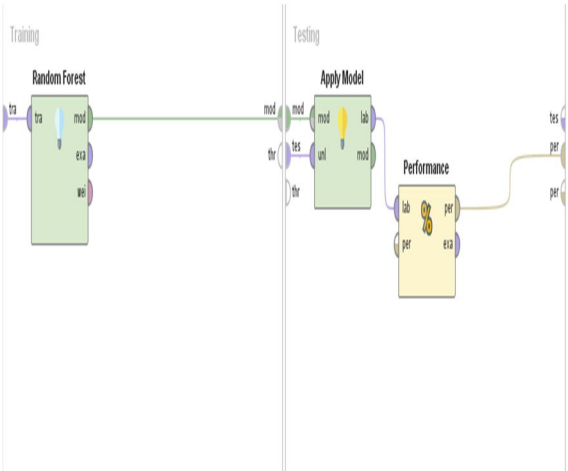
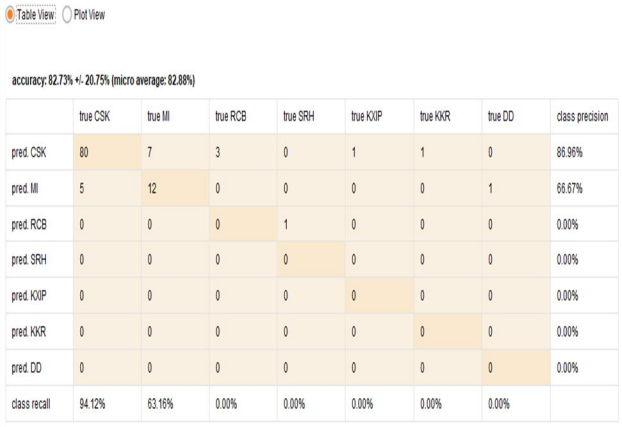


Fig-19:Cross Validation



	true CSK	true MI	true RCB	true SRH	true KXIP	true KKR	true DD	class precision
pred. CSK	80	7	3	0	1	1	0	86.96%
pred. MI	5	12	0	0	0	0	1	66.67%
pred. RCB	0	0	0	1	0	0	0	0.00%
pred. SRH	0	0	0	0	0	0	0	0.00%
pred. KXIP	0	0	0	0	0	0	0	0.00%
pred. KKR	0	0	0	0	0	0	0	0.00%
pred. DD	0	0	0	0	0	0	0	0.00%
class recall	94.12%	63.16%	0.00%	0.00%	0.00%	0.00%	0.00%	

Fig-20:Accuracy table

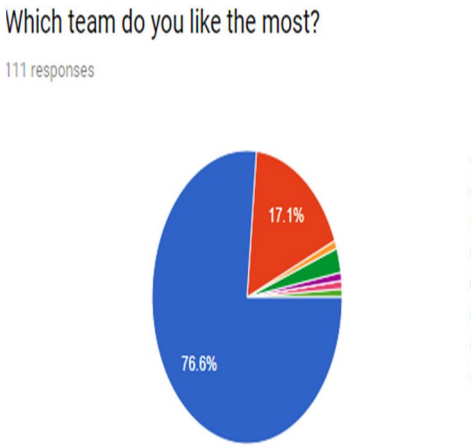


Fig-21: Survey chart

PerformanceVector

```
PerformanceVector:
accuracy: 82.73% +/- 20.75% (micro average: 82.88%)
ConfusionMatrix:
True:   CSK   MI   RCB   SRH   KXIP   KKR   DD
CSK:    80    7    3     0     1      1     0
MI:     5     12   0     0     0      0     1
RCB:    0     0    0     1     0      0     0
SRH:    0     0    0     0     0      0     0
KXIP:   0     0    0     0     0      0     0
KKR:    0     0    0     0     0      0     0
DD:     0     0    0     0     0      0     0
kappa: 0.493
ConfusionMatrix:
True:   CSK   MI   RCB   SRH   KXIP   KKR   DD
CSK:    80    7    3     0     1      1     0
MI:     5     12   0     0     0      0     1
RCB:    0     0    0     1     0      0     0
SRH:    0     0    0     0     0      0     0
KXIP:   0     0    0     0     0      0     0
KKR:    0     0    0     0     0      0     0
DD:     0     0    0     0     0      0     0
```

Fig-22: Performance vector

The above figure shows the cross validation process followed by its accuracy percentage table with the graph as shown above. It gives the accuracy of 82.73%. The other related measures are kappa 0.493, and predicting by means of which out of 100%responses 76.6% like CSK team the most.

IV. CONCLUSION

This paper has intended on analyzing the results of the IPL match during the year 2008-2019 by applying the data mining algorithms for existing data, and predicted the new data for the year 2020 and applied data mining algorithms for the proposed data. The Implementation tools used are Rapid Miner Studio Version-9.3.This knowledge will be used in future to predict the winning team. Hence using this prediction, the best team can be formed.

REFERENCES

[1] Parag Shah, “Predicting Outcome of Live Cricket Match Using Duckworth-Lewis Par Score”, Publisher: International Journal of Latest Technology in Engineering, Management & Applied Science, Volume VI, Issue VIIS, July 2017.
[2] Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Husain Dawood, and Yixian Yang , “Prediction of Rising Stars in the Game of Cricket”, Publisher: IEEE Access, Issue March 4 2017.
[3] Mehvish Khan, Riddhi Shah, “Role of External Factors on Outcome of One Day International Cricket (ODI) Match and Predictive Analysis”, Publisher: International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015.
[4] Akhil Nimmagadda et. Al, “Cricket score and winning prediction using data mining”, IJARnD Vol.3, Issue3.
[5] 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.
[6] Amit Dhurandhar and Alin Dobra, “ Probabilistic Characterization of Random Decision Trees”, Journal of Machine Learning Research, 2008.
[7] IJETCS, vol. 3, issue 2, ISSN:2455-9954, April 2018.
[8] Rabindra Lamsal and AyeshaChoudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”
[9] Rabindra Lamsal and AyeshaChoudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”
[10] 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.
[11] Hiremath, G., Venkatesh, H. and Choudhury, M. (2019), "Sports sentiment and behavior of stock prices: a case of T-20 and IPL cricket matches"