Analysis and Prediction of Indian Premier League using Data Science

1st Vinayak Srivastava
Centre for Machine Intelligence and Data Sciences
IIT Bombay
Mumbai, India
Email: 200100169@iitb.ac.in

2nd Vineet Pravin Ghule
Centre for Machine Intelligence and Data Sciences
IIT Bombay
Mumbai, India
Email: 20D070089@iitb.ac.in

3rd Vinayak Gautam
Centre for Machine Intelligence and Data Sciences
IIT Bombay
Mumbai, India
Email: 200020161@iitb.ac.in

4th Ayush M Gopal
Centre for Machine Intelligence and Data Sciences
IIT Bombay
Mumbai, India
Email: 200020004@iitb.ac.in

***Abstract*—Cricket is perhaps the most popular sport in India and also one of the most popular sports in the world. Of the various cricket tournaments held around the world, the Indian Premier League (IPL) is the most famous domestic league. In this project, a two large datasets containing ball by ball data as well as match by match data (2008-2020) are analysed. Various batting and bowling statistics are analysed. We also analysed some general team statistics which are important in cricket. We also analysed some relevant data related to different venues. Finally, we use a machine learning algorithm that predicts the result of a match based on factors such as toss results, quality of the teams, venue etc. Three Machine Learning Frameworks are used - Support Vector Machine, MLP Classifier, Random Forest Classifier.**

***Index Terms*—Cricket, sport, Indian Premier League, datasets, Machine Learning**

I. Introduction

The T20 format of cricket was first introduced in 2003 and has since become extremely popular among cricket fans because of it's shorter and faster format. The Indian Premier League is an annual T20 tournament organized by the BCCI. At the same time, sports analytics is an emerging field. Using the large amount of data to make predictive models helps teams make imformed decisions about their upcoming games. Thus analyzing IPL data becomes important. At the same time, large amounts of data are available for the purpose. So, it also becomes important to clean the data and filter out the important data which would help us to make a better model.

II. Background and Prior Work

A. About Cricket

Cricket is a bat-and-ball sport played between two teams comprising eleven players each, played on a field. The centre of the field has a 22-yard pitch having a wicket at either end. The wicket consists of three stumps and two bails placed on the stumps. A player from the fielding team, called the bowler bowls the ball from one end of the pitch to another. The batting team scores runs when the batsman at the striker's end hits the ball with the bat and the batsmen run between the wickets. Other ways of scoring runs are boundaries (fours and sixes), extras (wides, no balls, byes, leg byes). The bowling side can take wickets in any of the following ways - bowled, caught, leg before wicket (LBW), run out and stumped. There are three formats of cricket - One Day International (ODI), T20, Test Cricket.

B. About the Indian Premier League

The Indian Premier League is an annual professional T20 cricket tournament, organized by the Board of Control for Cricket in India (BCCI). The league was founded in 2007 and is usually held between March and May every year. The league consists of eight teams who play against each other twice in a home-and-away round-robin format in the league phase. The top four teams then qualify for the playoffs. The top two teams compete against each other in the "Qualifier 1" match - the winner of the match qualifies for the IPL final while the loser gets another chance to qualify for the final if they win the "Qualifier 2" match. The other playoff game (between the third and fourth placed teams) is called the "Eliminator" - the winner of this match qualifies for the "Qualifier 2" match and the loser gets the eliminated from the tournament. The winner of "Eliminator" and loser of "Qualifier 1" face off in "Qualifier 2" and the winner qualifies for the final. The winner of the final is crowned IPL Champion for that year.

C. Literature Review

G. Sudhamathy and G. Raja Meenakshi[1] have used four machine learning algorithms - Decision Tree, Naive Bayes, K-Nearest Neighbour and Random Forest to predict IPL data and compare the results. They have also measured accuracy, error rate, precision, recall, sensitivity and specificity.
Amala Kaviya V.S., Amol Suraj Mishra and Valarmathi B.[2] have portrayed the results of using a detailed ball-by-ball

dataset of all the matches played in the history of IPL and doing a comprehensive analysis of various aspects regarding measures involved in the game along with pragmatic visualizations. They have also ranked the players based on Player Ranking Index using machine learning techniques.

III. Data and Methodology

A. Datasets

The datasets used are available on www.kaggle.com [3] and [4]. The datasets which have been used for the purpose of this project contain data for individual matches as well as ball-by-ball data. The data contains useful features like toss decision, winner, result margin, whether D/L was used or not, umpires etc. We have added another feature to the dataset - 'Quality' which would help us in making better predictions.

B. Some Basic Batting Analyses

Let us look at some basic batting statistics to understand who are the best batsmen in IPL history.

1) *Players with the maximum number of runs* - Fig. 1 shows the top 10 batsmen for number of runs scored from 2008-2020. Virat Kohli is at the top of the list and by a fair margin (510 runs).

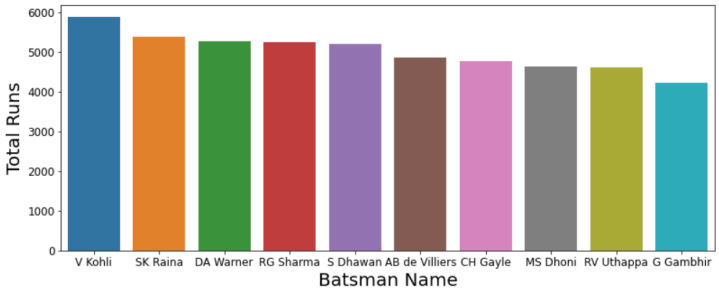

Fig. 1. Number of Runs Scored by each Batsman

2) *Players with the maximum number of sixes* - Fig. 2 shows the top 10 batsmen for number of sixes scored from 2008-2020. Chris Gayle completely dominates this category. He has scored 114 sixes more than the second placed AB De Villiers. Chris Gayle has also hit the maximum number of sixes in an innings (17) and also has the highest number of centuries (6) in IPL history. He has also scored the highest individual score in an innings (175).

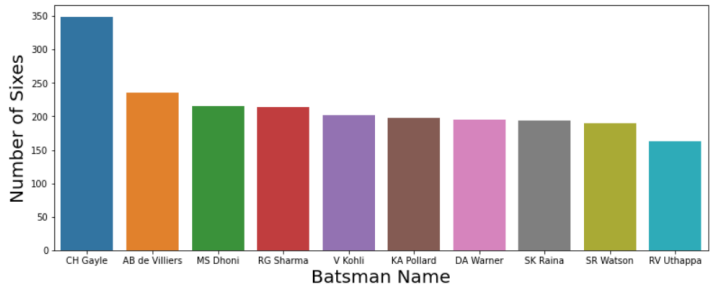

Fig. 2. Number of Sixes Scored by each batsman

3) *Hard-Hitting Ability* - The hard-hitting ability of a batsman is calculated as follows:

$$\text{Hard-hitting ability} = \frac{\text{Number of Fours and Sixes}}{\text{Number of Balls Faced}}$$

We have only considered those batsmen who have faced atleast 50 balls (which is almost equal to the length of a normal inning from a batsman). From Fig 3. we see that Luke Wright has the best hard hitting ability.

batsman	Hard Hitting Ability
LJ Wright	0.301587
SP Narine	0.270506
Kamran Akmal	0.269231
AD Russell	0.265306
MJ Lumb	0.257576
AC Blizzard	0.252747
KK Cooper	0.242857
V Sehwag	0.240044
BCJ Cutting	0.232877
N Pooran	0.222910

Fig. 3. Hard-Hitting Ability

4) *Finishing Ability* - The finishing ability of a batsman is calculated as follows:

$$\text{Finishing ability} = \frac{\text{Number of Not out innings}}{\text{Total number of innings}}$$

From Fig 3. one can see that Iqbal Abdulla has the best finishing ability in IPL history.

batsman	Finishing Ability
Iqbal Abdulla	0.988372
RD Gaikwad	0.982659
MN van Wyk	0.977778
PD Collingwood	0.974843
PV Tambe	0.974359
B Sumanth	0.973684
DJ Harris	0.970874
JP Duminy	0.970833
LMP Simmons	0.969248
HM Amla	0.968900

Fig. 4. Finishing Ability

Some interesting observations from our analyses:

- Looking at the hard-hitting ability, most of the players in the top 10 are all-rounders, which is expected since

most all-rounders usually bat during the last 5-6 overs and this is time when the batting team usually scores highest number of runs. Also, Virendra Sehwag being on this list is commendable, as he is an opener.

- M.S. Dhoni is widely considered by cricket experts as the greatest finisher in IPL history. However, statistically he doesn't even make the top 10 in terms of finishing ability. Why is this the case? Most of the batsmen on this list haven't played a large number of innings. For example, Iqbal Abdulla has only played 13 innings and he was not out in 11 of them. Further, Finishing ability in a sporting sense also includes many other factors.

C. Some Basic Bowling Analyses

1) *Most Wickets in IPL History* - From Fig. 5 it can be seen that L. Malinga is the leading wicket taker in IPL, followed by A. Mishra and Piyush Chawla.

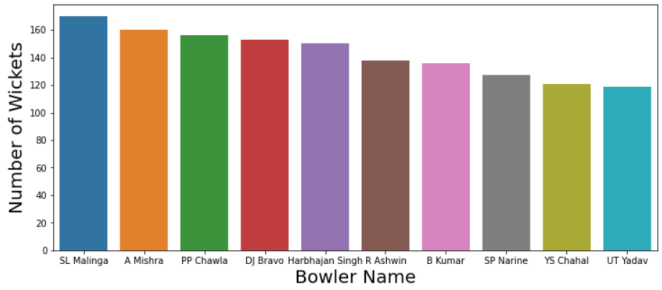

Fig. 5. Number of Wickets

2) *Most Dot Balls in IPL History* - From Fig. 6 it can be seen that Harbhajan Singh bowled the maximum number of dot balls (no runs scored) in IPL. This shows that he wasn't the best wicket taker but he was really good at saving runs.

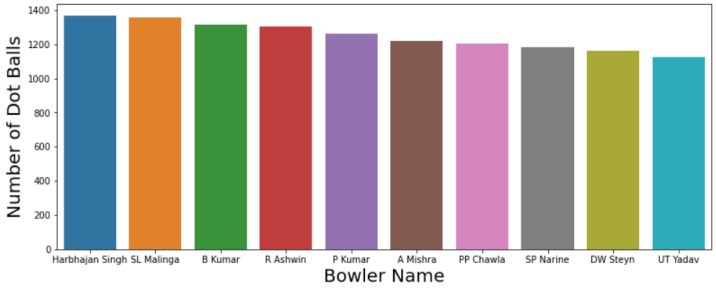

Fig. 6. Number of Dot Balls

3) *Economy* - Economy of a bowler is calculated as:

$$\text{Economy} = \frac{\text{Number of runs conceded}}{\text{Number of overs bowled}}$$

We have only considered those bowlers here who have bowled atleast 336 bowls. 336 is calculated as $4 \times 6 \times 14$ since a bowler can bowl a maximum of 4 overs in an innings and there are 14 matches in the league phase of the tournament. So a bowler should have bowled for atleast a full season. From Fig. 7 we find that Rashid Khan has the best economy in IPL history. He might not be the greatest bowler but he has certainly conceded runs at a slow rate. Harbhajan Singh not being in the top 10 can be explained as follows - he does have the most dot balls in IPL history, but not having a good economy means that the balls which aren't dot balls are usually expensive, i.e., he concedes a lot of runs as well. So, bowling a lot of dot balls doesn't indicate whether a bowler concedes less runs or not.

bowler	Economy
Rashid Khan	6.100671
M Muralitharan	6.209258
SW Tait	6.258427
A Kumble	6.268566
DW Steyn	6.303163
DP Nannes	6.391872
RE van der Merwe	6.408791
MA Starc	6.431373
SL Malinga	6.441829
R Ashwin	6.458070

Fig. 7. Economy

D. Most Man of the Match Awards

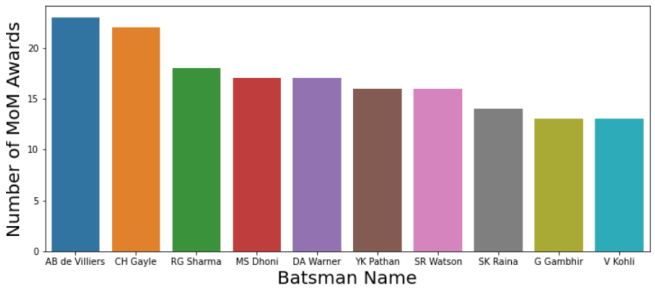

Fig. 8. Number of MoM Awards

AB de Villiers and Chris Gayle have the maximum number of MoM awards as is evident from Fig. 8. Both play for Royal Challengers Bangalore (RCB). However, RCB are not a very successful team implying they rely on their batsmen to produce moments of individual brilliance and the bowlers don't perform well enough.

E. Some general Team Analyses

1) *Instances of extra runs* - The most common form of extra runs are wides followed by legbyes as shown in Fig. 9.

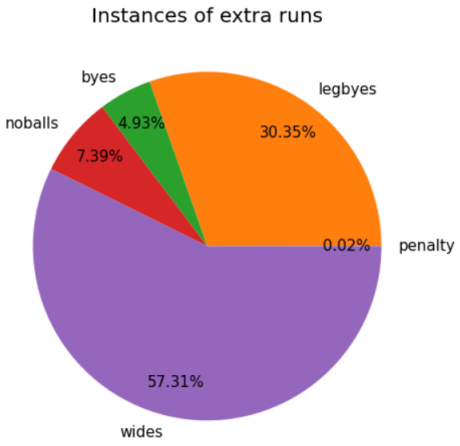

Fig. 9. Instances of extras

2) *Instances of wickets* - The most common form of wickets is 'caught' as shown in Fig. 10. This is expected as the easiest way of getting a wicket is forcing the batsman to play a risky shot.

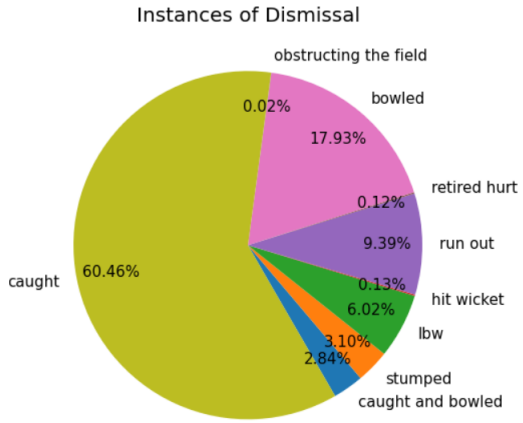

Fig. 10. Instances of wickets

3) *Change in run rate vs change in wicket rate* - Run rate here is calculated as the number of runs scored per ball and wicket rate is the number of wickets per ball. We plot Δ(Run Rate) versus Δ(Wicket Rate) as shown in Fig. 11. It seems that there is a negative linear correlation between the two variables. A high wicket rate in the first innings implies either a low run rate in the first innings or a high run rate in the second innings.
Further the eliminator matches or the playoffs are close to (0,0) on the graph indicating that these are close contests which is how it should be.

F. Home Advantage Analysis

Do teams get an advantage playing at home? One would logically think that they should. Let us see what the data says. In Fig. 12 we have plotted the data for certain matches and see the percentage of home team wins. Indeed it does seem that

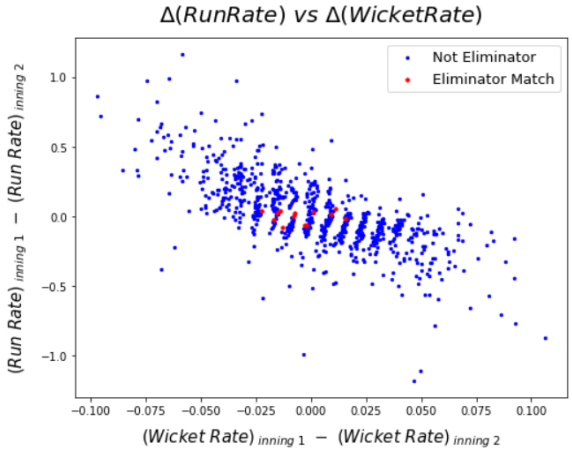

Fig. 11. Δ(Run Rate) vs Δ(Wicket Rate)

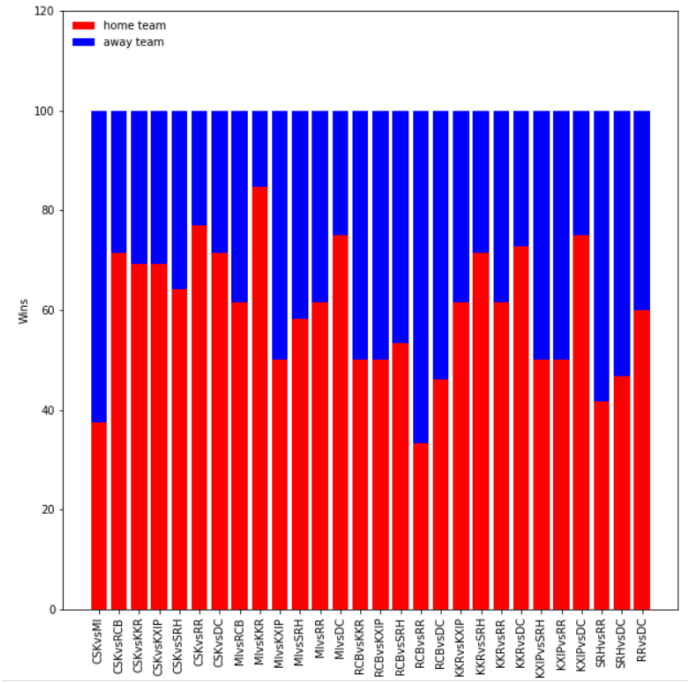

Fig. 12. Home Advantage Analysis

for most fixtures, home advantage is present. However, certain fixtures like CSK vs MI, RCB vs DC, SRH vs RR don't seem to have such a case. This could be because of similar quality of the teams.

G. Venue Analysis

1) *Total number of matches played* - M. Chinnaswamy Stadium in Bengaluru has hosted the maximum number of matches (83) followed by Eden Gardens, Kolkata (79).
2) *Toss Results* - The M. Chinnaswamy Stadium as well as Eden Gardens, Kolkata have seen the toss winners win the match on maximum occasions (43). Across the

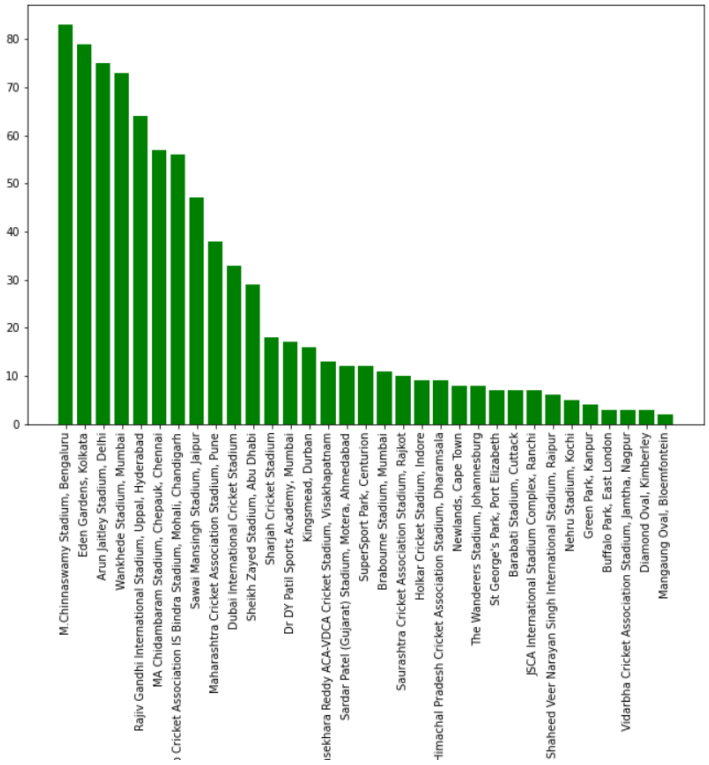

Fig. 13. Number of matches at each venue

various venues, it doesn't seem that toss results affect the result in a big way.

3) *Home Advantage* - Eden Gardens, Kolkata boasts the maximum number of home wins (46) followed by Wankhede Stadium, Mumbai (44).
4) *Toss Decisions* - It seems that batting first in the M. Chinnaswamy Stadium is an advantage (37 wins for teams batting first), as is the case with Wankhede Stadium (36 wins for teams batting first). Also balling first at Eden Gardens (47 wins for teams batting first) and Chinnaswamy stadium (44 wins for teams batting first) seems better.
5) *Most Runs, Wickets and Boundaries* - M. Chinnaswamy Stadium has produced the maximum percentage of runs, wickets and boundaries. We have shown the distribution of runs among the stadiums in Fig. 14

IV. Experiments and Results

A. Predictions

Before making a model to make predictions, it is important to clean the data. This includes taking care of null values, multiple names referring to the same team etc. So we take care of that first. In addition, to make predictions, we introduce an extra parameter, 'Quality' which is basically the win percentage of the team across all the seasons (2008-2020). The features we use to make our model are:

- venue

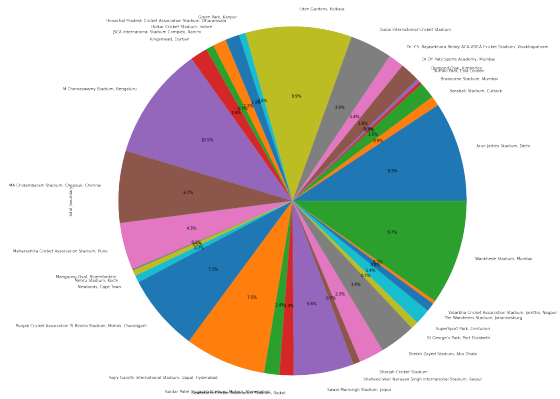

Fig. 14. Distribution of Runs

- team1
- team2
- toss_team1
- toss_bat
- method
- Quality1
- Quality2

The first 3 features are self-explanatory. 'toss_team1' represents whether team1 won the toss or not. 'toss_bat' represents whether the toss decision was batting or not. 'method' represents whether D/L was used or not. 'Quality1' and 'Quality2' represent the quality of team1 and team2 respectively.

1) *MLP Classifier* - A multi-layer perceptron (MLP) is a class of feedforward artificial neural network. The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons. A MLP which solves classification problems is called MLP Classifier. We have used the MLP Classifier provided by

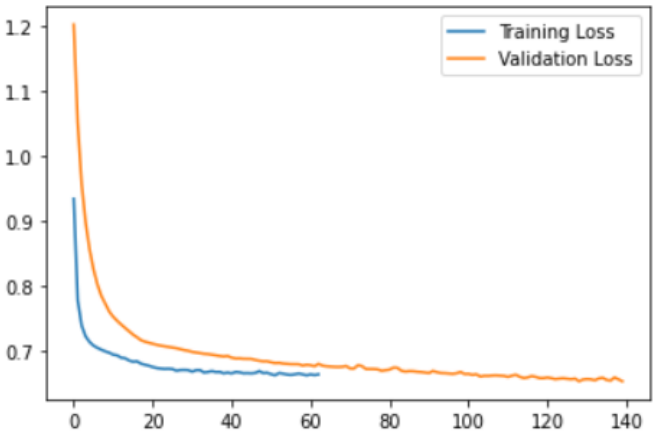

Fig. 15. Training and Validation loss plotted for MLP Classifier against epochs

scikit-learn. There are 2 hidden layers in the MLP of size 6 each, with the initial learning rate being equal to 0.003

and a batch size of 15 is used. 'Relu' activation is used. The metrics we will use are accuracy score and f1 score. *Results* - A validation accuracy of 68.03%, test accuracy of 60.66% and a f1 score of 69.23% on the test data are produced. The validation loss and training loss have been plotted against number of epochs in Fig. 15.

2) *Support Vector Machine* - SVM or Support Vector Machine is a linear model used to solve classification and regression problems. It can solve linear and non-linear problems . The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. We use 'rbf' kernel, value of gamma as 0.1 and C as 5. We use the same metrics as the previous case.
Results - A validation accuracy of 65.57%, test accuracy of 59.84% and a f1 score of 69.57% on the test data are produced.
3) *Random Forest Classifier* - Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest produces a class prediction and the class with the most votes becomes our model's prediction. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models operating together will outperform any of the individual constituent models. We have used a Random Forest Classifier with number of estimators as 12, maximum depth as 5 and maximum number of features as 5. Again, we will be using the same metrics to evaluate our model. *Results* - A validation accuracy of 58.20%, test accuracy of 62.30% and a f1 score of 68.92% on the test data are produced.

B. Results

Given below is a summary of all the frameworks and their results:

ML Framework	Validation Score	Test score	F1 Score
MLP Classifier	0.680328	0.606557	0.692308
SVM Classifier	0.655738	0.598361	0.695652
Random Forest Classifier	0.581967	0.622951	0.689189

Fig. 16. Summary of Various models and results

V. Learning, Conclusions and Future Work

This project provides valuable insight about various metrics and statistics important to understand quality of individual players and teams. Some of the most important players were seen during MoM analysis. Various metrics related to batsmen were seen and it was observed that some batsmen are better at hitting the ball harder whereas others are more consistent. Similarly, bowling data was also analyzed. A negative correlation between run rate and wicket rate was observed. We also observed that teams usually get an advantage when playing on their home ground. We also analysed the data related to different venues and provided some insight on the runs scored, results etc. at each venue. Predictions about the results of matches were made based on toss decisions, venue, quality of teams etc. The best model had nearly 62 % accuracy with nearly 70 % f1 score. Future work could include using recurrent neural networks to make better predictions.

Contribution of Team Members

Vinayak Srivastava - Individual player analysis, ML model used to make predictions and writing report.
Vineet Pravin Ghule - General statistics analysis, run rate vs wicket rate analysis and ML model used to make predictions.
Ayush M Gopal - Analysis of home advantage and making of demo video.
Vinayak Gautam - Analysis of data for different venues.

Acknowledgment

We want to thank all the professors for this course - Prof. Amit Sethi, Prof. Manjesh Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan who gave us the opportunity to work on this project and learn about the application of data science in sports

References

[1] PREDICTION ON IPL DATA USING MACHINE LEARNING TECHNIQUES IN R PACKAGE G. Sudhamathy and G. Raja Meenakshi Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, India, 2020. pp 2199-2204, DOI: 10.21917/ijsc.2020.0313

[2] AmalaKaviya V.S., Mishra, A. S. and Valarmathi B. (2020). Comprehensive Data Analysis and Prediction on IPL using Machine Learning Algorithms. International Journal on Emerging Technologies, 11(3): 218–228.

[3] IPL Complete Dataset (2008-2020): https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020.

[4] Indian Premier League (IPL) - All seasons https://www.kaggle.com/rajsengo/indian-premier-league-ipl-all-seasons

[5] IPL Data Analysis and Visualization Project using Python : https://machinelearningknowledge.ai/ipl-data-analysis-and-visualization-project-using-python/

[6] The Data Science behind IPL : https://www.analyticsvidhya.com/blog/2021/05/the-data-science-behind-ipl/

[7] Cricket:https://en.wikipedia.org/wiki/Cricket, Indian Premier League: https://en.wikipedia.org/wiki/Indian_Premier_League