International Journal on Emerging Technologies **11**(3): 218-228(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Comprehensive Data Analysis and Prediction on IPL using Machine Learning Algorithms

Amala Kaviya V.S.[1], Amol Suraj Mishra[2] and Valarmathi B.[3]
[1]Member of Technical Staff - Grade 2, VMware India Pvt. Ltd., Bangalore (Karnataka), India.
[2]Member of Technical Staff - Grade 2, NetApp, Bangalore (Karnataka), India.
[3]Associate Professor, Department of Software and Systems Engineering, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore (Tamilnadu), India.

(Corresponding author: Valarmathi B.)

(Received 28 January 2020, Revised 01 April 2020, Accepted 03 April 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: A detailed analysis of the complete IPL dataset and visualization of various features necessary for IPL evaluation is performed. Many machine learning algorithms have been used to compare and predict the winner between any two teams. Few models exist that try to rank players either based on simple formulae or based on few mathematical models. Efficiency was very low, in the absence of valuable data sets in large proportions. This is because enough data was not available when these models were suggested. T20 game has its own requirements which weren't satisfied by current models. In this paper, we 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. We faced issues with ranking the players and we overcame that by modelling their strength and weakness against a particular opponent, their performance on a particular pitch, etc. details which can be of great benefit and can give the team a winning edge to a large extent. We have also ranked the players, based on the Player Ranking Index using machine learning techniques. The accuracy of predictions have increased upto 81% using the proposed system (Comprehensive data analysis on IPL (CDAI)) causing a hike of 12% compared to the existing system (Deep mayo predictor (DMP)).

Keywords: BA, IPL, MVPI, ODI, PRI, T20.

Abbreviations: BA, Batting Average, IPL, Indian Premier League; MVPI, Most Valuable Player Index ; ODI, One Day International **;** PRI, Player ranking index; T20, Twenty-20.

I. INTRODUCTION

Cricket is a thrill to both play the game and to watch it and its importance is no less than any sporting event. Particularly after the advent of IPL, it gained huge popularity among people of all age groups, throughout the universe. On one hand where it is said that cricket is totally unpredictable, whereas on the other hand, this is also very true that, cricket matches results heavily rely on the past statistical data. Hence there is a need for an accurate prediction model, which could provide comprehensive analysis on players (his standards, strength, weakness), teams and could also predict higher chances of one team winning over the other. It could be of great help to team owners (who purchase players for their teams), captain and coaches to make the right selection for playing 11, to invest in the right team for betting and lastly for the people who are curious about IPL and its statistics. So far, no such application has been proposed or developed in the past. This application is one try to fill up the gap. Thus, an application which could analyze and see the existing data and also could make predictions on future matches would actually do wonders as far as IPL is concerned.

Few days back a prediction was that the succeeding ability of huge, web scale datasets, as a substitute for difficulties in models. And we got the detailed 10 seasons IPL dataset, of 636 matches played so far in IPL from the cricsheet website. This dataset if analyzed properly can do huge wonders. How analyzing data has done wonders in the field of the stock market, etc. in a similar way, an application which would do detailed analysis on players would be of great benefit. This motivated us to make an application which can do comprehensive analysis, visualization along with the prediction in every possible way and give the user detailed information.

Few models exist that attempt to rank players either based on simple formulae or based on few mathematical models. Few models try to predict the winner. Considering efficiency, it is very low, in the absence of enough data set. Because, the time when those models were suggested, enough data wasn't available to train the models. Most of the models made by using ODI cricket dataset too, along with T20 dataset, as T20 dataset alone wouldn't be enough for the need of prediction. But it had a loophole/shortcoming that the ODI performance of players was not equivalent or relevant to the performance rate of players in T20. Both the formats and its requirements are way different. These little variations found, creates the need to rank them using actual IPL/T20 data which are available now. The disadvantages of the system include i) Low efficiency ii) Incorrect prediction method iii) Incomplete functionality iv) Less options for analysing v) Less usage of graphs for output. Contrast to all the other attempts, which just concentrated on one of the aspects (either batsman characteristics or bowler characteristics), this paper will do a comprehensive analysis on all possible aspects of the IPL. It will be a 1 stop solution for any analysis needed. Besides, it will also be able to rank the players, not only with their batting average, but using a lot of parameters, and thus much more accurate and it will be in sync with their present form.

Initially, it will be able to read and it also derives batsman specific, bowler specific, 1 team detailed and 2 team specific data separately and saves them in separate files. Apart from these, it also contains special functions 3 for batsman analysis, bowler analysis, 1 team detailed analysis, 2 specific team analysis and particular match analysis. All these will be possible to do, using a web interface. Besides this, it also ranks players based on a combination of many factors. Through different rankings, we can analyze the same player's versatility. And these rankings are used to predict the players of teams, playing opposite each other, and predict the outcome of a match using our proposed approach CDAI and the Player Ranking index.

Advantages of the proposed system include (i) to analyze the player, it takes into account all the teams that he played for (ii) It takes into account, ball by ball details from all the 10 seasons 636 matches (iii) It has the option of both visualization and tabular output for a few functions (iv) It can be used in future also, if new seasons yaml data files are made available (v) It could be of great help to team owners (who purchase players for their teams in auction every year) (vi) It could be of great help to captain and coaches to make the right selection for playing 11 (vii) It could be of great help to invest in the right team for betting (viii) It could be of great help to lastly for the people who are curious about IPL and its statistics. In short, CDAI will be able to provide a beneficial prediction for analysing player performances and match results expectations using the varied machine learning algorithms analysed in this paper.

II. LITERATURE SURVEY

Surveying deep into analyzing cricket gives us the following insights. Dynamic programming models were used by Clarke to suggest the batting strategies which were optimal [1]. He suggested that it is the ball by ball nature of the cricket that makes it suitable for dynamic programming. As a part of his findings, it was able to suggest few computations at any stage of the innings, along with a few extra estimates like he runs to be scored totally, etc. Normal batting averages face a drawback related to the player not being out in a match. To overcome that, Kimber and Hansford and also Damodaran came up with the idea of alternate batting averages methods. To deal with situations when the batsman has not been out yet in the one day matches [2, 3]. A method for prediction of matches based on strike rates and batting average was suggested by Kantor and Barr respectively [4]. Test matches were explored on the basis of batting average by Borooah and Mangan respectively [5]. To increase the efficiency in the batting order, an approach based on mathematical models was applied by Clark and Norman and Bukeit and Ovens respectively [6, 7]. The mathematical modelling method also finds other applications in terms of likelihood of capability of one team to beat the other, besides finding the most efficient and effective batting order amongst the 11 players available in the team. Duckworth/Lewis percentage values were analyzed by Lewis respectively [8]. Duckworth Lewis method is of very great importance in cricket, during the times of rain to declare the outcome of a match, and also give targets when only shorter durations are left.

After the immense popularity of Test and ODI, in 2005, came the era of T20, where each team is supposed to play for a limited 20 overs. Since it came into existence, it spread across the world very fast and gained popularity very quickly because of the dynamic and unpredictable nature of the game respectively. In this format of the game, selectors prefer slow-consistent-higher average players rather a faster strike-rate player.

So, some new work was needed in this new dimension of cricket. From the dynamic batsmen who can score most of their runs in boundaries, to having bowlers who can bring in quick wickets. So, new prediction models were in need which would consider these factors. Since April 2008, IPL has started. The league, which was founded by the Board of Control for Cricket in India (BCCI) in 2008, has come a long way to 2017 currently playing through 10 seasons, 637 matches. It has gained a lot of popularity since the time it came into existence. The most interesting aspect of IPL is being its dynamic nature season by season. Every season the team goes through auction and the players keep changing. So, for the formation of teams, in order to decide which players are better to bag at the auction, a lot of work was done. A generic model for the valuation of players based on their past record was suggested by Parker *et al.*, respectively [9]. Lenten *et al.*, (2012) suggested a hedonic model to accomplish the same [10]. A lot of existing attributes were combined by Rastogi and Deodhar (2009) to suggest a pricing model, whether the bid would go in profit or loss for the owner [11]. But all the above work had a big drawback in them. All these analyses were done using the player's ODI profiles, as not much T20 data was available then, in the early days of IPL. Strike rate, Batting Average, no of 4s and 6s, etc. were some common attributes used to rank players into different classes and gave each class a certain valuation. And it was seen that players' prices in the actual auction were very much consistent with the class in which the model classified the players in. Season by season those models kept improving as more and more data kept on increasing and better algorithms were proposed.

To fill in the gaps that were prevalent in the existing models, some more work was done. Singh (2011) proposed a model to assess if the player was actually worth the price we bought him for [12]. Input parameters for his model included a wage bill of the player, the wages of the support staff for him and other miscellaneous expenses bored for the player from the team. Output parameters were based on the points awarded to him by various rankings, his net run rate across the tournament, the various profits and revenues that were collected. Graphical methods were used to analyze batsmen and bowler performance in all forms of cricket by Van Staden [13].

Sabermetrics style of principle to analyse batting performance in cricket was suggested by Lakkaraju and Sethi respectively [14]. Cricket carries a lot of similarities in itself from baseball. Because, a lot of work and discussions are already available on baseball. This method of Sabermetrics, it does deal essentially with the application of statistical methods to make predictions on the game of baseball. This paper tries to apply similar approaches and techniques to the game of cricket. Performance analysis using batting and bowling averages, strike rate and economy rates were suggested by Lemmer [15-17] respectively. While dealing with strike rates we do come across a peculiar anomaly. A particular player may have better strike rate because his matches must have been on easier pitches and this counterpart who would have played it on difficult pitches. A normalization technique is needed before we compare them. All these factors were covered in the work above. All round performances of a player were evaluated by Saikia and Bhattacharjee [18]. The Bayesian approach of classification was used for the classification of all-rounders in IPL, based on how good they were. It was suggested on classifying the all-rounders, as a good performer, all-rounder batsmen, all-rounder bowler, and below average performer as all-rounders are very good assets. Strategy to find the most valuable player in the tournament (MVP) using a decision tree approach was suggested by Khandelwal et al respectively.

In the initial stages, the models couldn't give a very efficient prediction. Most of the models made by using ODI cricket dataset too along with T20 dataset, as T20 dataset alone wouldn't be enough for the sake of prediction [16-19, 21]. But it had a loophole that the ODI performance of players was not equivalent or relevant to the performance rate of players in T20. Both the formats and its requirements are way different. These little variations found, creates the need to rank them using actual IPL/T20 data which are available now. And also a new approach was needed for IPL specific prediction of matches.

Nimmagadda *et al.*, (2018) proposed a model which is used to predict the score in each of the innings using Multiple Variable Linear Regression along with Logistic regression and the winner of the match using the Random Forest algorithm [22]. Kapadia *et al.*, (2019) used the significant features of the dataset to have been distinguished utilizing filter-based techniques including Correlation-based Feature Selection, Information Gain (IG), ReliefF and Wrapper [23]. AI systems including Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN) and Model Trees have been used to predict the models. Rupai *et al.*, (2020) used several classifiers to predict the bowling performances from ODI matches [24].

This paper attempts to fulfil all those needs. From providing an interactive and user-friendly portal, which would provide very advanced functionalities in order to perform detailed exploratory analysis on all dimensions of matches, batsman, bowlers, etc. To possess, the ability to rank players well, using the novel ranking approach and another one which is done using advanced techniques. It also possesses the feature to predict the outcome of a match, based on the players who are part of the current playing 11. This paper will be beneficial for 4 categories of people:

– Team owners to have a detailed idea of a player's history and his ranking to help in deciding how far is it worth going to purchase him, how to make the right selection and combination of teams.

– Coaches and team captains themselves have a good understanding about their foes and make plans with the right combination of their playing 11 (at particular venue) to overshadow and accordingly to beat their opponents.

– People who are betting on IPL matches. To help them with decision making, which team is stronger and has got higher chances of winning a match, etc. For them to invest in the right team and maximize their profit.

– Last but not the least, regarding the people who are interested in IPL cricket and are curious to explore its statistics as their past time.

III. THE PROPOSED SYSTEM

The complete work done has been compactly organized into this architecture. It first begins with the processing of datasets and loading it in the backend. Then user interface is provided with different functionalities, which can be performed on the player / match. It can also be used to perform prediction.

We have implemented the following modules for analysis, prediction, ranking and visualization.

– Processing of datasets
– Batsmen performance analysis
– Bowler performance analysis
– Match analysis
– Head-on-head analysis of teams
– Team overall performance analysis
– Ranking of teams
– Match prediction
– User interface creation

The below diagram illustrates on the various modules of the proposed system. Modules of our proposed system are demonstrated in Fig. 1.

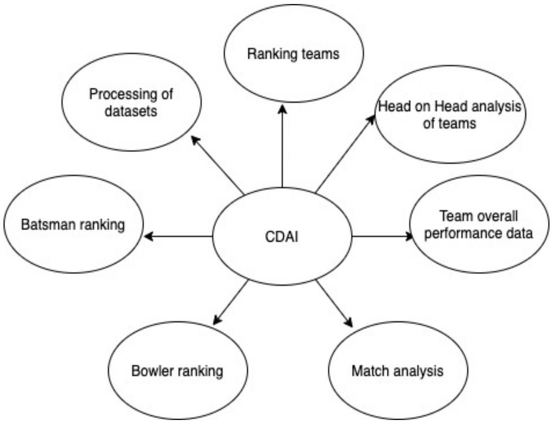


Fig. 1. Modules of the system.

A. Processing of datasets

This module's functionality is to get the IPL data ready and correctly formatted with apt data type for the rest of the project to function. We are using the dataset obtained from a cricsheet website which is presently in yaml (a specific type of xml format) (containing complete ball by ball detail). It reads each match's yaml data file and processes it, and saves match-wise complete ball by ball details in native R data frame with the correct data types assigned. Native R data frame because that will make further data reading and processing, much faster and efficient. Next from each match wise data frame, it also extracts and generates a separate data frame, which would contain the entire batsman related (team wise), bowler related (team wise), inter-team related and a particular team's entire career details. This module constitutes the heart of the project. And it's very important for all the data frames to be generated and placed in the right location, for the rest of the modules to work properly.

B. Batsmen performance analysis

This module provides the analyst with an ability to do a comprehensive analysis of a batsman profile. Initially, it extracts details of all the IPL teams a particular batsman played for (as it is highly probable for the player to have played for more than 1 team). Then after establishing the complete batsman profile, it can perform wide multifarious analysis and visualizations. A subset of them includes functionalities like plotting the runs of batsman against deliveries played by him, analysis of the various ways he got out, analysis of his batting average and strike rates, runs scored by him, venue wise etc.

C. Bowler performance analysis

This module provides the analyst with an ability to do a comprehensive analysis of a bowler profile. Initially, it extracts details of all the IPL teams a particular bowler played for (as it is highly probable for the player to have played for more than 1 team). Then after establishing the complete bowler profile, it can perform wide multifarious analysis and visualizations. A subset of them includes functionalities like mean economy rate of bowler, mean runs given by him, his wicket type plot, how well he has performed against a particular opposition, how well he has performed at a particular venue etc.

D. Match analysis's objective

This module is to analyze a single match completely. Apart including the basic functionalities to view the batting and bowling scorecard of a match, it is also embedded with advanced analysis and visualization functionalities. A subset of them includes analysis of the best batting partnership of each team in that match, how well particular batsmen have performed against a particular bowler and vice-versa, a few batsmen and bowler specific functions and vice versa, the match worm graph of two teams seeing how they have played etc.

E. Head on head analysis of teams

This module is used to compare and contrast only two teams, by analyzing all matches they played in the past, against each other. This feature would be of great help in decision making for both the teams whenever they come face to face against one another. It also offers a wide variety of functionalities.

A subset of which includes best batting partnerships team wise when they played in the past, the detailed batting and bowling scorecards, how well particular batsmen have performed against particular bowlers when those two teams played, win loss analysis, etc.

F. Team overall performance analysis

This module is used to analyse a team's performance as a whole. It does a comprehensive analysis on all the matches played by a particular team in its entire history by applying a wide variety of functions on it. This feature would be a very important and main deciding factor while accessing the standards of a team on whole and choosing favorites. A subset of them includes best batting partnerships in the history of the team, overall batting and bowling scorecard of the team, best batsmen of the team versus best bowlers of the tournament, best bowlers in the team versus best batsmen of the tournament etc.

As far as ranking is concerned, 3-3 modes of ranking are available for batsmen and bowlers. The first and most basic one is using batting average. The second one is done using MVPI (most valuable player index) ranking score suggested by Rediff. This was proposed by Rediff sports for giving useful insights about players.

For the third kind of ranking, it uses the parameters listed below for batsmen and bowlers, we generate the PRI of batsmen and bowlers and rank them. More details about all the rank generation will be discussed in later sections.

For batsmen, they are:

– Hard-hitter
– Finisher
– Fast-scorer
– Consistent
– Running-between-wickets

For bowlers, they are:

– Economy
– Wicket-taker
– Consistent
– Big-wicket-taker
– Short-performance-index

Now, a very important aspect is the ability to predict which team among the 2 playing teams would win a match. Likelihood value would be of great impact for a variety of things as discussed in previous sections. In IPL, players aren't constantly a part of a single team, because they keep changing based on a particular season's auction. The only thing that remains with a player, is his performance, how well he played across his previous seasons, no matter whichever team he was in. Based on this particular aspect, we use his PRI and perform the computation. More details about the match prediction will be discussed in later sections.

We have a user interface created for all the modules. It is an interactive shiny web app, whose front end and back end are purely written in R. It performs all the functionalities mentioned in the previous modules. It contains 3 input fields. First is the module to analyze, then, is to select the particular functionality to be analyzed and lastly to select the particular player to be analyzed for. The computation goes on in the backend. And the output gets displayed in the graph or tabular form in the front end.

G. Ranking

Ranking is done in 3 ways each for batsman and bowlers as mentioned in previous modules. They are explained below in detail.

(i) **Batting average ranking:** Here the batsman is ranked in descending order according to their batting average.

BA = (TR/TM)

– TR is the total runs scored by the batsman
– TM is the total matches played by the batsman.

(ii) **Batman MVPI ranking:** This is a better model of ranking compared to batting average which takes into consideration both batting average and batting strike rates respectively. So, we get a better measure of ranking for limited over IPL-T20 cricket [20].

MVPI = ((MR/TMR) + (MSR/TMSR)) * TR

where

– MR is the batting average of particular batsman

– TMR is the average of all batsmen in the IPL

– MSR is the mean strike rate of particular batsman

– TMSR is the mean strike rate of all the batsmen in the IPL

– TR is the total runs of the batsman

(iii) **Batsman PRI (Player ranking index):** This is the best model of ranking or we can say it is an improvisation over MVPI ranking. It takes into account 5 different parameters. All which matters, the most in T20 cricket. When it comes to batsman the measures like how hard can he hit the ball (being good at hitting 4s and 6s), capability of staying not out, capability of not wasting any deliveries, consistent performance and finally his running between the wickets. Using all these measures, we train a random forest model with predictors as these measures and the outcome being MVPI and we generate the PRI score and rank the batsmen.

The PRI is found using five parameters for batsmen. The parameters for batsmen include:

– Hard-Hitter = ((4*Four + 6*Six) / Balls played by batsman)

– Finisher = (Count of matches being not out/ Total count of innings played)

– Fast-Scorer = (Player batting strike rate)

– Consistent = (Player batting average)

– Running-Between-Wickets (RBW) = ((Run scored by the player) – (4*Fours+ 6*Sixes)/Number of balls faced without boundary)

(iv) **Bowling average ranking:** Here the bowlers are ranked in descending order according to their bowling average.

BOA = (TW/TM)

– TW is the total wickets taken by a bowler

– TM is the total matches played by a bowler

(v) **Bowling MVPI ranking:** This is a better model of ranking compared to bowling average which takes into account both bowling average and bowling economy rate respectively. So, we get a better measure of ranking for limited over IPL-T20 cricket [20].

MVPI = ((MW/TMW) + (TMER/MER)) * TW

where

– MW is the mean wickets taken by the bowler

– TMW is the mean wickets taken by all the bowlers in the tournament

– TMER is the mean economy rate of all the bowlers in the tournament

– MER is the average economy rate of the bowler

– TW is the total wickets taken by the bowler.

(vi) **Bowling PRI (Player ranking index):** This is the best model of ranking or we can say it is an improvisation over MVPI ranking. It takes into account 5 different parameters. All which matters, the most in T20 cricket. As far as bowlers are concerned, the measures like economy, wicket taker, consistent, big wicket taker and short performance. Using all these measures, we train a random forest model with predictors as these measures and the outcome being MVPI and we generate the PRI score and rank the batsmen.

The PRI is found using five parameters for bowlers. The parameters for bowlers include:

– Economy = (Runs conceded by player/ (Count of balls bowled/6))

– Wicket-Taker = (Count of balls bowled / Count of wickets taken)

– Consistent = (Runs conceded by the bowler/ Count of wickets taken)

– Big-Wicket-Taker = (Count of four wickets or five wickets or six wickets taken/ Count of innings played)

– Short-Performance = ((Count of total wickets – 4*Count of four wicket haul – 5* Count of times five wicket haul - 6*Count of six wicket haul) / (Count of total played innings /Count of times four (or) five (or) six wicket hauls totally))

H. Prediction

For prediction we make use of PRI generated in the previous sections. PRI are generated separately for batsmen and bowlers. Every player who ever played in the history of IPL surely would have a PRI. In the absence of corresponding batting/bowling records, he is assigned the last rank. The rank differences of playing 11 in the rival teams are the basic idea to make the predictions.

Prediction is made on two sets of data.

– Training data – Season 1 to Season 8 IPL data. Tested on – Season 9.

– Training data – Season 1 to Season 9 IPL data. Tested on – Season 10.

The first data is used to show significant difference compared to the existing models. Second is to predict the matches in the recent IPL.

Steps for predictive model used with the second training data is as follows:

– For a particular match, for both the teams separately, for each player, we need to find the batting PRI and bowling PRI for each player respectively.

– For batting and bowling PRI separately, we find differences between corresponding player's batting and bowling PRI.

– So, apart from the 22 columns (11 batting PRI and 11 bowling PRI) for a particular match, we add a 23rd column containing the match result, 1 if team-1 wins and 2 if team-2 wins.

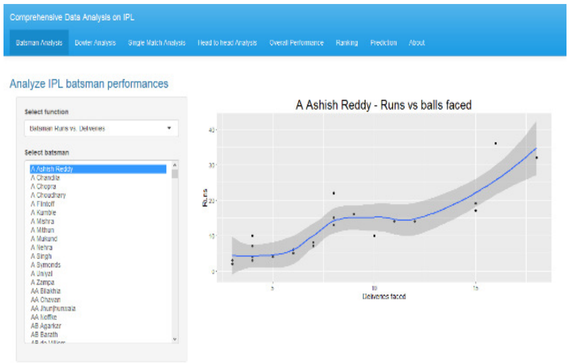
– Now we train various models over this dataset constructed. Which will be discussed in the below sections. After this we would have our prediction model ready.

Now when we are predicting a match's outcome, we generate the same data of 22 rows for that match and predict which among team 1 or team 2 would be the winner. Prepare a test set with these selected 22 features for 58 matches of season 10.

We use various algorithms for training which include support vector machine, sequential minimal optimization, Instance based learning in parameter k, Random Forest, JRIP reduced error pruning algorithm, J48 decision tree algorithm, Flexible Discriminant Analysis, Mixture discriminant analysis, C5.0 decision tree algorithm and naïve Bayes classifier.

IV. RESULTS AND DISCUSSION

A. Website Interface



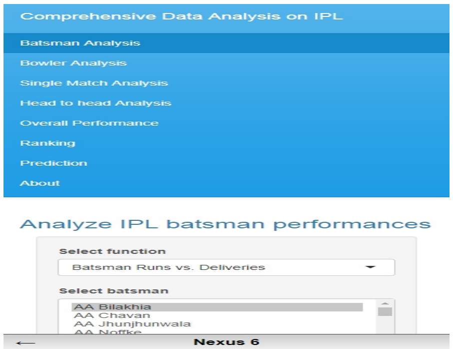
Analyze IPL batsman performances

Select function: Batsman Runs vs. Deliveries

Select batsman: A Ashish Reddy

Fig. 2. Website Web view.

Fig. 2 shows the website when it opens. The user needs to select 3 things, the module, function, and batsman.



Analyze IPL batsman performances

Select function: Batsman Runs vs. Deliveries

Select batsman: AA Bilakhia

Fig. 3. Website Mobile view.

Fig. 3 Shows how the website looks when it is launched on mobile phones.

B. Batsman Analysis Module

In Fig. 4, in the batsman analysis tab, we have selected the function, 'dismissals of batsman' and selected 'MS Dhoni'. It makes us a pie chart of his various dismissals all throughout his career.

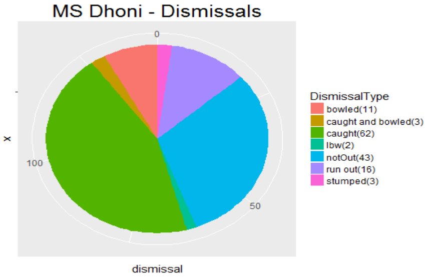


Fig. 4. Type of dismissal – MS Dhoni.

Using this chart, we can conclude that most of his dismissals have been through catch out. In Fig. 5, we analysed batsman runs vs dismissals for MS Dhoni and plotting a regression line through it. We can observe that as the amount of balls increase, strike rate goes higher and higher for Dhoni.

In Fig. 6 we are using a decision tree to predict what will be runs scored by the batsman having the balls faced as a predictor.

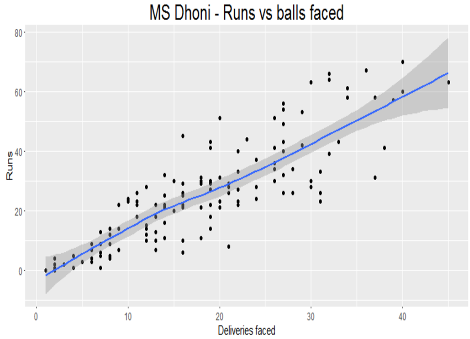


Fig. 5. Runs vs Balls faced – MS Dhoni.

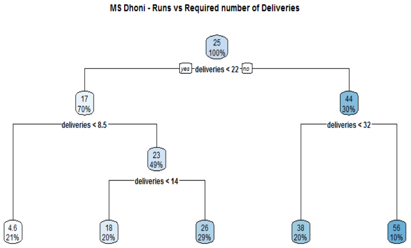


Fig. 6. Runs vs required no of deliveries – MS Dhoni.

C. Bowler Analysis Module

In Fig. 7, the bowler's average wickets, as a function of time throughout his career can be seen.

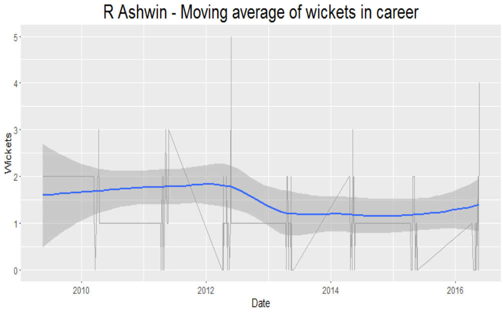


Fig. 7. Moving average of wickets in a career – R Ashwin.

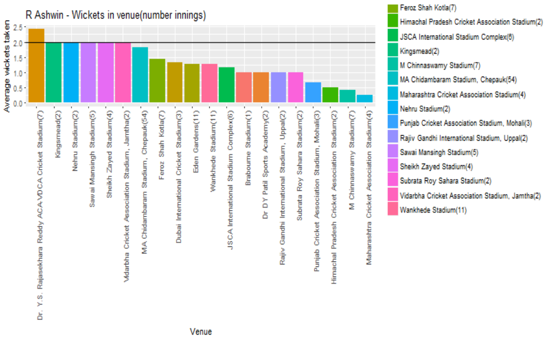


Fig. 8. Wickets by venue – R Ashwin.

In Fig. 8 we are analysing average wickets of a bowler at a particular venue. We can see that R Ashwin has the highest average wicket of 2.5 at the ACA-VDCA stadium which is at Visakhapatnam.

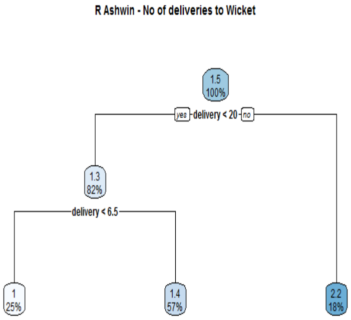
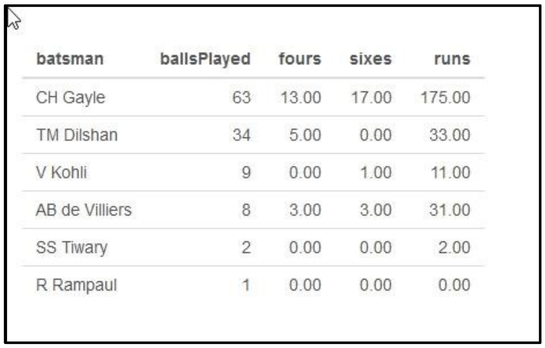


Fig. 9. No. of deliveries to wicket – R Ashwin.

In Fig. 9 we are using a decision tree to predict what will be wickets taken by a bowler having the deliveries bowled as a predictor.

D. Match Analysis Module



batsman	ballsPlayed	fours	sixes	runs
CH Gayle	63	13.00	17.00	175.00
TM Dilshan	34	5.00	0.00	33.00
V Kohli	9	0.00	1.00	11.00
AB de Villiers	8	3.00	3.00	31.00
SS Tiwary	2	0.00	0.00	2.00
R Rampaul	1	0.00	0.00	0.00

Fig. 10. Match scorecard – RCB vs PWI.

In Fig. 10, the match score card for a particular match. We have selected the same historic match in which Chris Gayle made a knock of 175 in 63 deliveries.

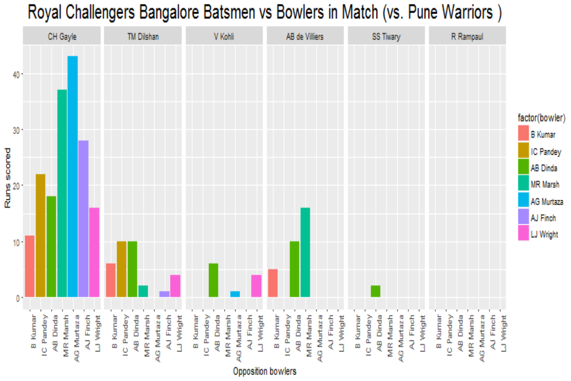


Fig. 11. Batsmen vs Bowlers – RCB vs PWI.

In Fig. 11, for a particular match, we analysed how the batsman of a particular team played against the bowlers of the opposite team. It is observed that Gayle spared none of the bowlers and scored as high as 48 runs against AG Murtaza.

E. Two team analysis module

In Fig. 12, we did a head on head analysis for 2 arch rivals, Chennai Super Kings and Mumbai Indians. We can see that Suresh Raina has the highest score and he has made the highest partnership with MS Dhoni as such.

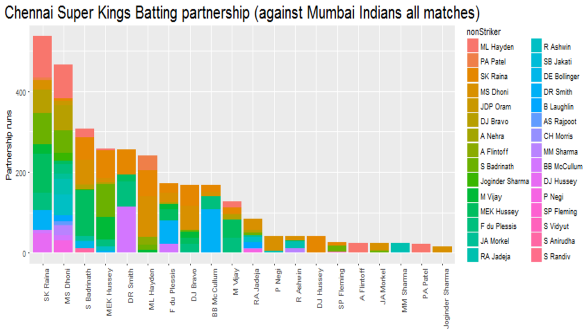


Fig. 12. CSK batting partnership against MI.

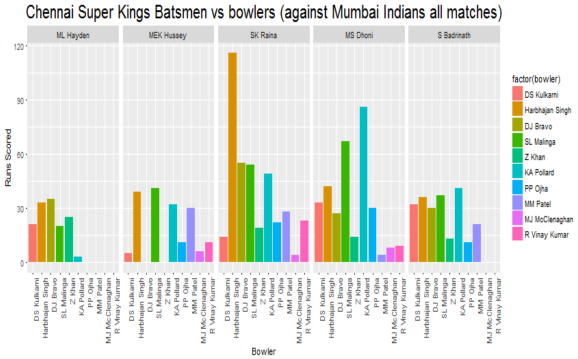


Fig. 13. CSK batsman vs MI bowlers.

In Fig. 13, between Chennai Super Kings and Mumbai Indians, we analysed the best batsman of CSK vs best bowler of MI. And we can see that Suresh Raina has hit Harbhajan Singh the most, who is also the top bowler in the opposition side.

F. Team overall performance module

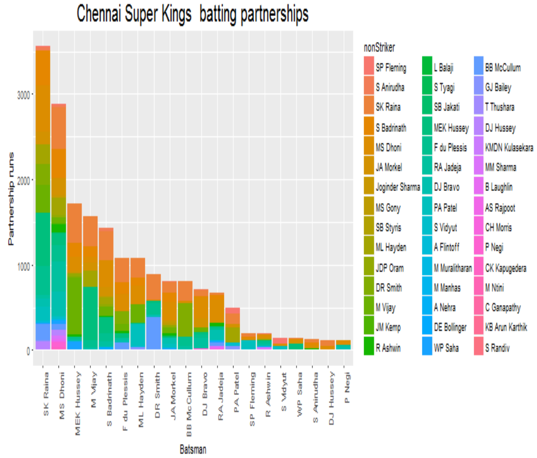


Fig. 14. CSK batting partnerships.

In Fig. 14, we tried to see Chennai Super Kings, top batsman, with whom they shared their best partnerships. In Fig. 15, we saw the performance of top batsman of Chennai Super Kings, Suresh Raina, against the top bowlers of IPL.

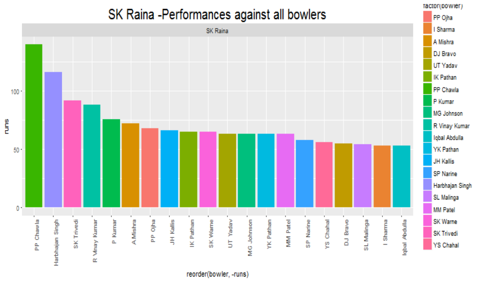


Fig. 15. SK Raina performance against all bowlers.

G. Ranking

In Table 1, the batsman is ranked according to their batting averages. Chris Gayle is ranked first with a batting average of 36.25. Only the top 10 batsmen are shown.

In Table 2, we are ranking players according to MVPI (most valuable player index) formula. Here we see that David Warner is ranked first and Virat Kohli is ranked second. Only the top 10 batsmen are shown.

Table 1: Batting average Ranking.

Batsman	Matches	Total runs	Mean runs	Mean SR	Rank
CH Gayle	100	3626.00	36.26	133.73	1
ML Hayden	30	1077.00	35.90	128.97	2
DA Warner	114	4014.00	35.21	126.24	3
SE Marsh	67	2320.00	34.63	117.61	4
MEK Hussey	57	1930.00	33.86	105.81	5
V Kohli	140	4331.00	30.94	115.66	6
AM Rahane	97	2895.00	29.85	102.81	7
SR Tendulkar	76	2221.00	29.22	108.15	8
AB de Villiers	117	3393.00	29.00	133.71	9
S Dhawan	123	3544.00	28.81	113.70	10

Table 2: Batsman MVPI Ranking.

Batsman	Matches	Total runs	Mean runs	Mean SR	MVPI	Rank
DA Warner	114	4014.00	35.21	126.24	17186.33	1
V Kohli	140	4331.00	30.94	115.66	16504.32	2
SK Raina	154	4408.00	28.62	122.96	16266.30	3
CH Gayle	100	3626.00	36.26	133.73	16127.22	4
RG Sharma	151	4109.00	27.21	113.48	14270.49	5
G Gambhir	144	4010.00	27.85	109.31	13970.35	6
RVUthappa	141	3744.00	26.55	123.91	13196.52	7
AB de Villiers	117	3393.00	29.00	133.71	13004.87	8
S Dhawan	123	3544.00	28.81	113.70	12796.71	9
MS Dhoni	134	3394.00	25.33	131.70	11883.46	10

Table 3: Batsman PRI ranking.

Batsman	Hard hitter	Finisher	Fast scorer	Consistent	RBW	MVPI	PRI	Rank
DA Warner	0.94	-0.57	0.75	2.73	0.41	17186.33	13633.77	1
V Kohli	0.47	-0.55	0.45	2.18	0.23	16504.32	13238.66	2
CH Gayle	1.66	-0.66	0.93	2.77	-0.48	16127.22	12533.84	3
SDhawan	0.40	-0.57	0.42	1.92	0.29	12796.71	12523.88	4
SK Raina	0.67	-0.49	0.64	1.91	0.66	16266.30	12119.40	5
G Gambhir	0.37	-0.59	0.31	1.92	0.33	13970.35	11856.81	6
RG Sharma	0.55	-0.54	0.41	1.79	0.36	14270.49	11313.47	7
RV Uthappa	0.66	-0.67	0.67	1.67	0.38	13196.52	11194.82	8
AB de Villiers	0.91	-0.08	0.92	1.95	0.71	13004.87	10086.89	9
MEK Hussey	0.32	-0.54	0.21	2.50	0.36	7636.65	9172.40	10

Table 4: Bowling mean wickets ranking.

Bowler	Matches	Total wickets	Mean wickets	Meaner	Rank
SL Malinga	108	169.00	1.56	6.72	1
A Nehra	87	121.00	1.39	7.72	2
MJ McClenaghan	39	54.00	1.38	8.64	3
Sandeep Sharma	55	75.00	1.36	7.82	4
SP Narine	80	109.00	1.36	6.33	5
DJ Bravo	103	137.00	1.33	8.08	6
YS Chahal	55	72.00	1.31	8.07	7
MG Johnson	46	60.00	1.30	8.01	8
B Kumar	89	116.00	1.30	7.13	9
P Awana	33	43.00	1.30	8.33	10

Table 5: Bowler MVPI ranking.

Bowler	Matches	Total wickets	Mean wickets	Meaner	MVPI	Rank
SL Malinga	108	169.00	1.56	6.72	563.44	1
DJ Bravo	103	137.00	1.33	8.08	385.04	2
A Nehra	87	121.00	1.39	7.72	355.65	3
Harbhajan Singh	131	134.00	1.02	7.09	344.54	4
SP Narine	80	109.00	1.36	6.33	344.30	5
B Kumar	89	116.00	1.30	7.13	339.22	6
A Mishra	123	133.00	1.08	7.70	338.77	7
R Vinay Kumar	102	125.00	1.23	8.24	331.82	8
PP Chawla	126	132.00	1.05	8.04	324.05	9
Z Khan	94	112.00	1.19	7.39	306.38	10

Table 6: Bowler PRI ranking.

Bowler	Pbwer	Pbwa	Pbwsr	Bwt	Shortperf	MVPI	PRI	Rank
SL Malinga	-0.80	-0.42	-0.57	1.57	1.47	563.44	343.81	1
DJ Bravo	-0.24	-0.43	-0.41	0.01	1.34	385.04	284.66	2
A Nehra	-0.37	-0.37	-0.40	0.38	1.36	355.65	272.20	3
Harbhajan Singh	-0.72	0.07	-0.18	0.10	0.54	344.54	230.01	4
SP Narine	-0.98	-0.23	-0.50	1.97	0.74	344.30	228.34	5
B Kumar	-0.67	-0.27	-0.42	0.36	1.15	339.22	233.35	6
A Mishra	-0.49	-0.08	-0.21	0.33	0.61	338.77	279.15	7
R Vinay Kumar	-0.18	-0.27	-0.26	0.26	0.99	331.82	245.40	8
PP Chawla	-0.47	-0.10	-0.21	0.12	0.61	324.05	276.52	9
Z Khan	-0.49	-0.12	-0.24	0.32	0.91	306.38	227.67	10

In Table 3, we are ranking batsman according to PRI (Player ranking index) formula. Here also we see that David Warner and Virat Kohli are ranked 1 and 2 respectively. Only the top 10 batsmen are shown.

In Table 4, bowlers are ranked according to their mean wickets. SL Malinga rules this table with an average of 1.56 being on top. Only the top 10 bowlers are shown.

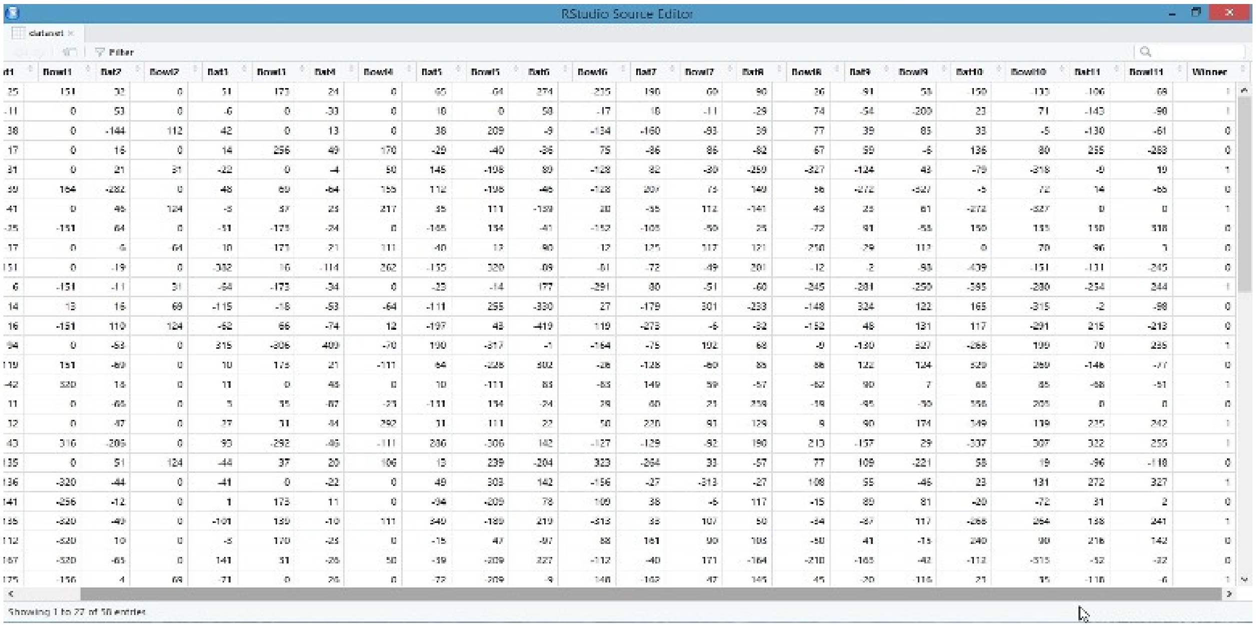
In Table 5, we are ranking players according to MVPI (most valuable player index) formula. Here also we see that SL Malinga is ranked first. Only the top 10 bowlers are shown.

In Table 6, we are ranking batsman according to PRI (Player ranking index) formula. Here also SL Malinga retains his rank 1 respectively. Only the top 10 bowlers are shown.

H. Prediction

Feature table of batsman and bowler rank differences (which will be used for further prediction) is generated in Table 7. Bat1 to Bat11 is the batting rank difference. Bowl1 to Bowl11 is the bowling rank difference. W stands for winner. Training set contains 550 rows approximately. First 22 rows of predicted outcomes for IPL season 10 are shown as samples in Table 7.

Table 7: Feature table for prediction.



d1	Bowl1	Bat2	Bowl2	Bat3	Bowl3	Bat4	Bowl4	Bat5	Bowl5	Bat6	Bowl6	Bat7	Bowl7	Bat8	Bowl8	Bat9	Bowl9	Bat10	Bowl10	Bat11	Bowl11	Winner
25	151	32	0	51	173	24	0	65	-64	274	-235	190	69	90	26	91	53	-150	-133	-106	68	1
-11	0	53	0	-6	0	-30	0	18	0	58	-17	18	-11	-29	74	-54	-209	23	71	-143	-90	1
38	0	-144	112	42	0	13	0	38	209	-9	-134	-160	-93	39	77	39	85	33	-5	-130	-61	0
17	0	16	0	14	256	49	170	-29	-40	-36	75	-86	86	-82	67	59	-6	136	80	255	-283	0
31	0	21	31	-22	0	-4	50	145	-198	80	-128	82	-30	-250	-327	-124	43	-79	-318	-9	19	1
39	164	-282	0	48	69	-64	155	112	-198	-46	-128	207	75	140	56	-272	-527	-5	72	14	-65	0
-41	0	46	124	-5	37	23	217	35	111	-159	20	-55	112	-141	43	25	61	-272	-527	0	0	1
-25	-151	64	0	-51	-175	-24	0	-165	154	-41	-152	-105	-50	25	-72	91	-58	150	135	150	518	0
-17	0	-6	-64	10	171	21	111	40	12	90	12	125	117	121	250	-29	112	0	70	96	7	0
151	0	-19	0	-382	16	-114	262	-155	320	-89	-61	-72	-49	201	-12	-2	-93	-409	-151	-131	-245	0
6	-151	-11	31	-64	-173	-34	0	-23	-14	177	-291	80	-51	-60	-245	-281	-250	-385	-280	-254	244	1
14	13	16	69	-115	-18	-53	-64	-111	255	-330	27	-179	301	-233	-148	324	122	165	-315	-2	-98	0
16	-151	110	124	-62	66	-74	12	-197	43	-419	119	-273	-6	-32	-152	48	131	117	-291	215	-213	0
94	0	-53	0	315	-306	409	-70	190	-317	-1	-164	-75	192	68	-9	-130	327	-268	199	70	235	1
119	151	-69	0	10	175	21	-111	64	-228	302	-28	-128	-60	85	86	122	124	529	269	-146	-77	0
-42	320	18	0	11	0	48	0	10	-111	83	-83	149	59	-57	-62	90	7	68	85	-68	-51	1
11	0	-66	0	5	35	-87	-25	-111	154	-24	29	60	21	258	-19	-95	-50	558	205	0	0	0
12	0	-47	0	27	31	84	292	11	111	22	50	220	91	129	9	90	174	349	119	225	242	1
43	316	-206	0	93	-292	-46	-111	286	-306	142	-127	-129	-92	190	213	-157	29	-337	307	322	255	1
135	0	51	124	-44	37	20	106	13	239	-204	323	-264	33	-57	77	109	-221	58	19	-96	-118	0
136	-320	-44	0	-41	0	-22	0	49	303	142	-156	-27	-313	-27	108	55	-46	23	131	272	327	1
141	-256	-12	0	1	173	11	0	-94	-209	78	109	38	-6	117	-15	89	81	-20	-72	31	2	0
135	-320	-40	0	-101	139	-10	111	340	-189	219	-313	33	107	50	-34	-87	117	-288	264	138	241	1
112	-320	10	0	-5	170	-25	0	-15	47	-97	88	161	90	103	-50	41	-15	240	90	216	142	0
167	-320	-65	0	141	31	-26	50	-19	-209	227	-112	-40	171	-164	-210	-165	-42	-112	-515	-52	-22	0
175	-156	4	69	-71	0	26	0	-72	-209	-9	148	-162	47	145	45	-20	-116	21	35	-118	-6	1

Showing 1 to 27 of 50 entries

Table 8 shows the prediction for IPL season 10. These are the predictions made by JRIP algorithm, which is found to be performing far better than the others. The table displays predictions of the top 10 matches (sample output). DMP (Deep mayo predictor) is the existing model proposed by Prakash *et al.,* CDAI (Comprehensive data analysis on IPL) is the proposed system.

In the first set of predictions in our proposed systems, for which a similar attempt was made by Prakash *et al.*, [21], for predicting Season 9 IPL results, where they made their training set with international T20+IPL Season 1-8 dataset. Their model was able to predict 39 out of 58 matches with an accuracy of 69.68%. Whereas, our model which was built only with IPL season 1-8 dataset is able to successfully predict 47 out of 58 matches outcomes. A comparison of both the model's accuracy is given below for reference.

Fig. 16 shows the accuracy comparison of predictions made by the existing system and proposed system respectively by using the SVM algorithm in our first set of predictions. The existing system has an accuracy of 69.64% in contrast to the proposed system, which has an accuracy of 81.03%.

In the second set of predictions in our proposed system, we predicted for season 10 using various algorithms. This is the first attempt for IPL 10. Our model is built using Season 1 to Season 9 IPL dataset in this case. Comparing the results of the model to the actual IPL 10 match outcomes. The accuracy comparison between each algorithm used for predicting season 10 results of matches is shown in Fig. 17.

Table 8: Prediction.

Match	Winner	Prediction	Result
Delhi Daredevils-Gujarat Lions-2017-05-04.RData	Delhi Daredevils	Delhi Daredevils	TRUE
Delhi Daredevils-Kings XI Punjab-2017-04-15.RData	Delhi Daredevils	Delhi Daredevils	TRUE
Delhi Daredevils-Kolkata Knight Riders-2017-04-17.RData	Kolkata Knight Riders	Kolkata Knight Riders	TRUE
Delhi Daredevils-Mumbai Indians-2017-05-06.RData	Mumbai Indians	Mumbai Indians	TRUE
Delhi Daredevils-Rising Pune Supergiants-2017-05-12.RData	Delhi Daredevils	Delhi Daredevils	TRUE
Delhi Daredevils-Royal Challengers Bangalore-2017-05-14.RData	Royal Challengers Bangalore	Delhi Daredevils	FALSE
Delhi Daredevils-Sunrisers Hyderabad-2017-05-02.RData	Delhi Daredevils	Sunrisers Hyderabad	FALSE
Gujarat Lions-Delhi Daredevils-2017-05-10.RData	Delhi Daredevils	Gujarat Lions	FALSE
Gujarat Lions-Kings XI Punjab-2017-04-23.RData	Kings XI Punjab	Kings XI Punjab	TRUE
Gujarat Lions-Kolkata Knight Riders-2017-04-07.RData	Kolkata Knight Riders	Kolkata Knight Riders	TRUE

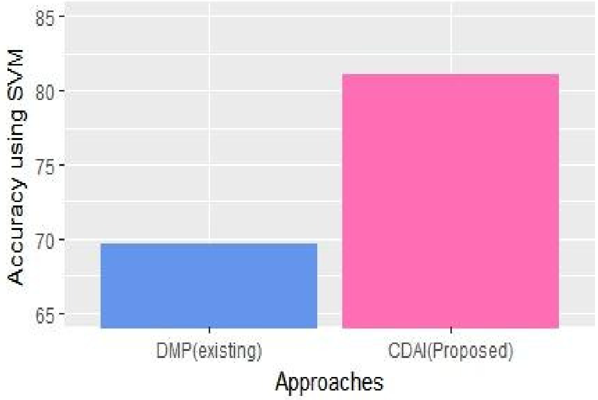


Fig. 16. CDAI system's proposed algorithms accuracy comparison.

As it is a binary classification problem, random forest and other tree based algorithms are outperformed by the likes of JRIP and SVM. Amongst all the algorithms we have applied, JRIP seems the most promising. With an accuracy of 75.86%, for predicting 44 out of 58 matches of IPL 10 correctly.

Then SVM and FDA also gave good results with an accuracy of 72.41% respectively, for predicting 42 out of 58 matches of IPL 10 correctly. Rest all the algorithm results are shown.

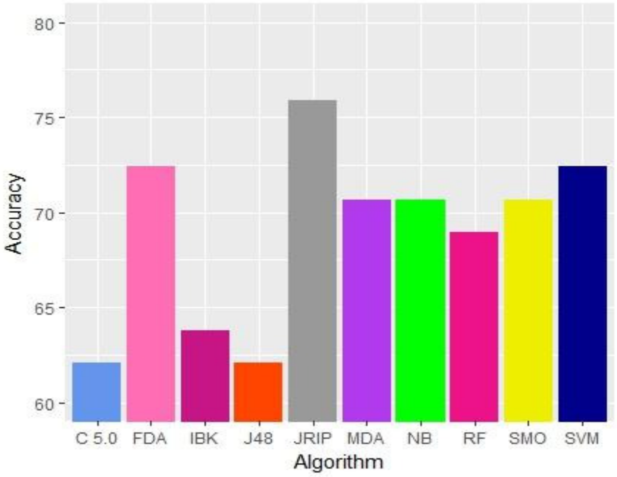


Fig. 17. SVM accuracy comparison for both approaches of existing and proposed system.

V. CONCLUSION

The approach has brought out analysis and visualization of various aspects of IPL matches in all the possible ways and gives useful results to the user. This information is of great value. It could be of great help to team owners (who purchase players for their teams in auction every year), captain and coaches to make the right selection for playing 11, to invest in the right team for betting and lastly for the people who are curious about IPL and its statistics.

VI. FUTURE SCOPE

In future, making minor changes the model can also be made to work with the ODI and test matches. The international matches can be analysed in a similar way and more visualizations can be added to the functions. The system can also be made to adapt more file formats of data for better analysis of varied forms of data collected.

Conflict of Interest. There is no conflict of interest involving the content enlisted in the given paper.

REFERENCES

[1]. Clarke, S. R. (1988). Dynamic programming in one day cricket - optimal scoring rates. *Journal of the Operational Research Society*, *50,* 536 – 545.

[2]. Kimber, A. C., & Hansford, A. R. (1993). A Statistical Analysis of Batting in Cricket. *Journal of Royal Statistical Society*, *156*, 443 – 455.

[3]. Damodaran, U. (2006). Stochastic Dominance and Analysis of ODI Batting Performance: The Indian Cricket Team, 1989-2005. *Journal of Sports Science and Medicine*, *5*, 503 – 508,

[4]. Barr, G. D. I., and Kantor, B.S..A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket.*Journal of the Operational Research Society*, *55*, 1266-1274.

[5]. Borooah, V. K., & Mangan, J. E. (2010). The Bradman Class: An Exploration of Some Issues in the Evaluation of Batsmen for Test Matches 1877–2006. *Journal of Quantitative Analysis in Sports*, *6*(3): 14-22.

[6]. Norman, J., & Clarke, S. R. (2004). Dynamic programming in cricket: Batting on sticky wicket. *Proceedings of the 7th Australasian Conference on Mathematics and Computers in Sport*, 226–232.

[7]. Ovens, M., & Bukeit, B. (2006). A mathematical modeling approach to one day cricket batting orders. *Journal of Sports Science and Medicine*, *5*, 495-502.

[8]. Lewis, A. (2008). Extending the Range of Player-Performance Measures in One-Day Cricket. *Journal of Operational Research Society*, *59,* 729-742.

[9]. Parker, D., Burns, P., & Natarajan, H. (2008). Player valuations in the Indian Premier League. *Frontier economics Journal*, *68,* 68-76.

[10]. Lenten, L. J., Geerling, W., & Kónya, L. (2012). A hedonic model of player wage determination from the Indian Premier League auction: Further evidence. *Sport Management Review*, *15*(1), 60-71.

[11]. Rastogi, S. K., & Deodhar, S. Y. (2009). Player pricing and valuation of cricketing attributes: exploring the IPL Twenty20 vision. *Vikalpa*, *34*(2), 15-23.

[12]. Singh, S. (2011). Measuring the Performance of Teams in the Indian Premier League. *American Journal of Operations Research*, *1,* 180-184.

[13]. Van, Staden, P. (2009). Comparison of Cricketers' Bowling and Batting Performance using Graphical Displays.*Current Science*, *96,* 764-766.

[14]. Lakkaraju, P., & Sethi, S. (2012). Correlating the Analysis of Opinionated Texts Using SAS® Text Analytics with Application of Sabermetrics to Cricket Statistics. *Proceedings of SAS Global Forum*, 1-10.

[15]. Lemmer, H. (2004). A Measure for the Batting performance of Cricket Players. *South African Journal for Research in Sport, Physical Education and Recreation*, *26*, 55-64.

[16]. Lemmer, H. (2008). An Analysis of Players' Performances in the First Cricket Twenty20 World Cup Series. *South African Journal for Research in Sport, Physical Education and Recreation*, 30, 71-77.

[17]. Lemmer, H. (2012). The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket.*Journal of Sports Science and Medicine*, *10*, 630-634.

[18]. Saikia, H., & Bhattacharjee, D. (2011). A Bayesian Classification Model for Predicting the Performance of All-Rounders in the Indian Premier League. *Vikalpa*, *36*(4), 51-66.

[19]. Khandelwal, M., Prakash, J., & Pradhan, T. (2015). An Analysis of Best Player Selection Key Performance Indicator: The Case of Indian Premier League (IPL). *Advances in Intelligent Systems Technologies and Applications*, 173-190.

[20]. http://www.rediff.com/

[21]. Prakash, C. D., Patvardhan, C., & Lakshmi, C. V. (2016). Data Analytics based Deep Mayo Predictor for IPL-9. *International Journal of Computer Applications*, *152*(6), 6-10.

[22]. Nimmagadda., A., Kalyan, N. V., Venkatesh, M., Teja, N. N. S., & Raju, C .G. (2018). Cricket score and winning prediction using data mining. *Int. J. Adv. Res. Development*, *3*(3), 299-302.

[23]. Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2019). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*, 1-6.

[24]. Rupai, A. A. A., Mukta, M, & Islam, A.K.M.N., (2020). Predicting Bowling Performance in Cricket from Publicly Available Data. *International Conference on Computing Advancements,* 1-6.

How to cite this article: 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.