# ClassicXI – Optimal IPL Team Selection based on Player Rating and Venue Based Analysis

Lavitra Kshitij Madan
CSE Dept
PES University
Bangalore, India
lavitra.kshitij@gmail.com

Ritik Hariani
CSE Dept
PES University
Bangalore, India
ritikhariani@gmail.com

Arjun Chengappa
CSE Dept
PES University
Bangalore, India
arjunchang25@gmail.com

Aditya G Burli
CSE Dept
PES University
Bangalore, India
adityaburli06@gmail.com

***Abstract*—Indian Premier League (IPL) is one of the most prestigious T20 cricket tournaments across the world. This paper focuses on suggesting the best playing XI given a squad of 15 players. We propose a model that rates all the players in the IPL and based on their predicted rating, suggests the best playing XI combination the team can select for a given match at the given venue. Our model experiments with various distance measures and other statistics to provide the best possible playing XI. It also provides insights on toss decision at a given venue based on history.**

***Keywords—Cricket, IPL, Exploratory Data Analysis, Modelling, Rating, Playing XI***

## I. INTRODUCTION

The Indian Premier League (IPL) is an annual T20 tournament held in India. It was started in 2008 by the BCCI and it is now in its 13th season. It is considered to be the best T20 tournament in the world. The success of the IPL can be attributed to a number of things, but one substantial factor to it was the quality of cricket it has produced over the years, hence providing quality entertainment for the audience. The IPL teams are owned by franchises that participate in auctions to purchase players- both domestic and international. The auctions are held every year, where players released from the teams and other uncapped players are put up to bid. A team is allowed to have a maximum of 25 players in its squad. The playing eleven of a team must have a maximum of four overseas players. The auctions are held every year, where players released from the teams and other uncapped players are put up to bid.

Every team spends a large amount of time and money to decide which eleven players to select from the 25 in the bag. The T20 format is a peculiar one, where the game is slightly skewed in favour of the batsmen compared to the bowlers. Shorter boundaries, a 6 over power play, day/night conditions etc. make up the reasons for that. Hence, there is no doubt that the teams give emphasis on the strength of their batting line up. Teams tend to have a mix of batsmen that can score at a good strike rate and those that bring stability to prevent a collapse. The order in which the two kinds of batsmen are sent also makes a huge difference. The opening pair generally has one batsman that can be explosive and another that takes a few balls to get going.

The middle order is considered to be very crucial in T20 and it generally has a combination of experienced batsmen that can take the game till the end and batsmen who can score big from the first ball. All-rounders are also known as game changers in the T20 format, as they can contribute with either their batting or bowling. Teams prefer to have a mix of batting and bowling all-rounders, and it is the all-rounders of a team that are generally juggled around in accordance to the match being played.

Since the game is skewed in favour of batting, it also becomes very important for a team to have a potent and balanced bowling line up. In the modern game, the diversity of bowlers is far greater. The line-up needs to have a mix of spin and seam bowling; impact bowlers that can either take wickets with the new ball, take control of the middle overs or restrict the opposition in the death overs. In the IPL, teams also have to decide on which four overseas players get their place in the playing eleven.

Given all this, we tend to notice that the process of selecting a playing eleven requires the teams to analyse all of their players individually to figure out whether they make the cut and where they are to be placed. Data Analytics and mathematical modelling can be extensively used in the process. This paper performs various exploratory data analysis (EDA) on a couple of datasets- the ball by ball dataset of the IPL and the match dataset to derive at required intuitions and deeper understanding of the game that will facilitate in the final effort to build a comprehensive model to compute the playing eleven of an IPL team given a particular squad. EDA is performed on different aspects of the game such as- analysis of fielding and dismissals, toss analysis, average of batsmen against a team, average wickets of a bowler against a team etc. The data gathered from this has been used to compute the ratings of all the players, which in turn will helped us in deciding the playing eleven for the given squad.

## II. RELATED WORK

A lot of statistical analysis has been done over the last decade on cricket. There has been an increase in the attention of performing analysis on the Indian Premier League due to the large sum of money invested in it. This had led to researchers taking the initiative of predict match outcomes, predicting auction sale prices of various cricketers. The respective IPL teams' management also include of data analysts to provide input to the team's

owners and coach on strategies on which player to choose and for what price. They also play a significant role in suggesting improvements in the various areas where performance is weak such as fielding, a batsman's strike rate, a bowler's economy, player's fitness levels based on the consumption of food and his physical training activities.

A method as devised by [1] ranks players based on the following factor - Deep Performance Index method to evaluate a player. The algorithm considers context of the player in terms of indices like most valuable player index for batsman and for bowler. The important context introduced is the winning contribution ratio of the player. An approach for evaluating a player based on the venue of match, overall average, current season average, and additional parameters. Batting Rating Points System (**BRPS**) = **WCR** + **AVI** + (**BCSA** / **BOA** * **BCSSR** / **BOSR**) + **HPR**,

where, WCR is the Winning Contribution Ratio, AVI is the Average Venue Index, BCSA is the Batsman Current Season Average, BOA is the Batsman Overall Average, BCSSR is the Batsman Current Season Strike Rate, BOSR is the Batsman Overall Strike Rate, HPR is the Hitting Power Ratio

Bowling Rating Points System (**BoRPS**) = **WCR** + **AVI** + (**BoOA** / **BoCSA** * **BoOE** / **BoCE**) + **WMR**,

where, WCR is the Winning Contribution Ratio, AVI is the Average Venue Index, BoCSA is the Bowler Current Season Average, BoOA is the Bowler Overall Average, BoOE is the Bowler Overall Economy, BoCE is the Bowler Current Season Economy, WMR is the Wickets Matches Ratio. The values of BPRS vary from 3.47 to 8.42 and that of BoRPS vary from 2.58 to 6.54.

Although, to apply the above algorithms the following assumptions were necessary: Factors involving team support and other factors such as fielding support or assists were ignored (Team Contribution). [1] Only considered IPL performance data up to IPL 7 and the overall T20 career data up to the end of IPL 8 of all the players participating in IPL. Batsman scoring more than 500 runs in twenty-20 internationals with a strike rate exceeding 100 and who have played in at least 25 matches were considered for batting ranking. Applying these criteria, 89 batsmen were considered for this rating system.

[2] provides insights about modelling batsmen, bowlers and ultimately modelling teams (rating) and suggesting a 'Playing XI', given a squad/pool of players using a few features. [2] propose novel methods to model batsmen, bowlers and teams, using various career statistics and recent performances of the players. They also propose that their model can predict the winner of ODI cricket matches, by using a novel dynamic approach to reflect changes in player combinations. A lot of assumptions have been made in the modelling step of batsmen, bowlers and the teams as a whole. [2] use features such as **Matches Played** (Matches played by the player), **Batting Innings** (Matches in which the player batted), **Batting Average** (Runs scored divided by the times the player got out), **Num Centuries** (Times the player scored ≥ 100 runs in a match), **Num Fifties** (Times the player scored ≥ 50 but less than 100 runs in a match), **Bowling Innings** (Matches in which the player bowled), **Wkts Taken** (Wickets taken by the player), **FWkts Hauls** (Times the player has taken ≥ 5 wickets in a match), **Bowling Average** (Runs conceded by the player per wicket taken) and **Bowling Economy** (Economy Average runs conceded by the player per over). Many important features such as strike rate of batsmen, dot balls of bowler and partnership within the team have been ignored. Another important concept is that all-rounders have not been considered separately but they are taken as batsmen or bowlers. Hence, our paper aims to also model all-rounders by using a similarity measure (equally perform well as a batsman or a bowler) to improve the playing XI prediction quality. Our model provides more value to recent performances as the players learn and grow. Although [2] has been implemented for ODI cricket matches, IPL T20 would be a more challenging task and hence would be more valuable in terms of need in today's world. [2] claim that the k-Nearest Neighbor (kNN) algorithm used provides better results with an accuracy of 71% as compared to other classifiers models that have an accuracy of 56% and 63%.

An insight is provided by [3] on how to tackle modelling all-rounders which was not taken care of in [2]. The paper takes into consideration whether the player was in the playing eleven for at least 5 matches in the IPL, the player has bowled for at least 10 overs in IPL, the player has faced at least 100 balls in the IPL, and, only a total of 35 players have been considered for the research. Therefore, it does not represent the whole population of players. From this paper we learn that use of Wald statistic, Step-wise multinomial logistic regression (SMLR) and the naïve Bayesian classification model for forecasting the expected class of all-rounders based on the significant predictors can be a method to approach our problem. [3] based on the above method have claimed that all-rounders have been successfully classified into 4 non-overlapping categories, namely- Performing all-rounders, batting all-rounders, bowling all-rounders and Under-performer with an accuracy of their model at 66.7%.

## III. METHODOLOGY

### *A. Dataset*

One of the datasets used in this project is taken from Kaggle[1]

This consists of two .csv files deliveries.csv and matches.csv. The file deliveries.csv contains the ball by ball data of every match since 2008. Data includes columns such as bowling team, batting team, over, ball, batsman etc. This dataset has a total of 21 columns and 179078 rows. The file matches.csv contains details about the match as a whole. Data includes columns such as season, city, teams, toss_winner, toss_decision, etc.

The second dataset used was obtained from the official IPL website[2]. This dataset consists of the playing XI details for every match that has been played in the IPL from 2008 to 2020 and also includes the respective player stats such as the player's full and short name, date of

[1] https://www.kaggle.com/nowke9/ipldata

[2] https://www.iplt20.com/

birth, nationality and their skillset. The winner, score of both teams and date of every match was also obtained.
This dataset consisting of 167422 lines of data was extracted in JSON format and was converted into a Dataframe.

### *B. Pre-processing*

The file deliveries.csv has at least 170244 NaN values in the attribute's fields: player_dismissed, dismissal_kind and fielder after cleaning.
The file matches.csv was fact checked and cleaned. All the empty places were replaced with NaN to avoid discrepancies and have an overall sense of consistency. Some stadium names in this dataset had minute spelling mismatches. These were found when the unique values of the stadium attribute were sorted. For example, we found two names that pointed to the same stadium '**M. Chinnaswamy Stadium**' and '**M Chinnaswamy Stadium**'. This was resolved by individually replacing all the errors using the replace function of the 'pandas' library. In some of the records it was observed that the number of innings was reported to be 5. This is not possible in T20 and is obviously an error in data entry. There were also cases where a match was reported to have 4 innings but weren't reported as super-overs. These errors were individually looked into and rectified. A few discrepancies regarding no balls and wide runs format were observed due to the ever-changing laws of cricket. These were simply handled by a series of well-informed if-else statements.
Using the JSON object obtained, two new data frames were created. One which contains data regarding the player information and the other contains number of bowlers / batsmen in each team in each game.

## IV. DATA ANALYSIS

To understand the data, and the features that will be needed to play an important role in our model, we have performed Exploratory Data Analysis and Visualizations to select out the important features as shown here and in the *Appendix* as well.

The following graph Figure 1. shows the head to head data analysis for the team Royal Challengers Bangalore (number of wins, losses and number of tied matches as respective percentages against every other team they have faced in the Indian Premier League)

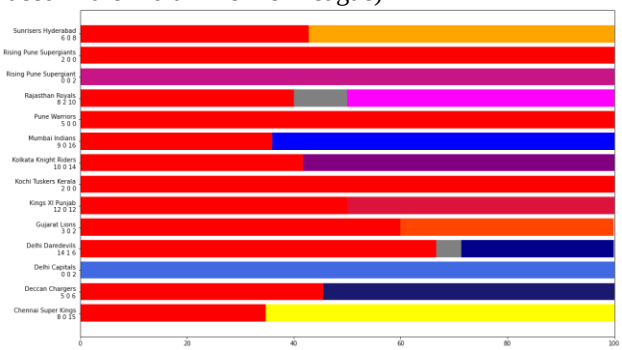


Figure 1. Head to Head data for RCB vs other teams

Figure 2. shows the toss analysis for the various venues on which the matches were played. It shows the win prediction possibility based on the data over 11 seasons of the IPL. It shows the probability that a team would likely win at a particular venue based on the toss decision (either choosing to bat or bowl first).

| | bat | field |
|---|---|---|
| Rajiv Gandhi International Stadium, Uppal | 0.21 | 0.42 |
| Maharashtra Cricket Association Stadium | 0.50 | 0.68 |
| Saurashtra Cricket Association Stadium | 0.00 | 0.57 |
| Holkar Cricket Stadium | 0.00 | 0.88 |
| M Chinnaswamy Stadium | 0.44 | 0.55 |
| Wankhede Stadium | 0.50 | 0.51 |
| Eden Gardens | 0.43 | 0.63 |
| Feroz Shah Kotla | 0.47 | 0.55 |
| Punjab Cricket Association IS Bindra Stadium, Mohali | 0.35 | 0.54 |
| Green Park | NaN | 1.00 |
| Sawai Mansingh Stadium | 0.32 | 0.68 |
| MA Chidambaram Stadium, Chepauk | 0.61 | 0.38 |
| Dr DY Patil Sports Academy | 0.43 | 0.60 |
| Newlands | 0.75 | 0.33 |
| St George's Park | 0.43 | NaN |
| Kingsmead | 0.60 | 0.60 |
| SuperSport Park | 0.50 | 0.83 |
| Buffalo Park | 0.67 | NaN |
| New Wanderers Stadium | 0.00 | 0.50 |

Figure 2. Toss Analysis for win based choosing to bat or bowl first

## V. PROPOSED WORK

We propose a model/algorithm that rates all the players to have played in all the previous IPL seasons and based on their predicted rating suggests the best playing XI combination the team can choose. The model experiments with various distance measures and other statistics such as venue, toss decision to provide the best prediction. Our model builds upon [2] by applying a more detailed algorithm that involves the addition of more features and increasing the weightage given to every feature. We also suggest Team 1 whether to bat or field first for (Team 1 vs Team 2) at a *given venue* by analyzing the previous Head to Head data as shown above in Section IV. Our model consists of 2 algorithms: One for Modelling Batsmen, other for Modelling Bowlers.

***Modelling Batsman***: Batsmen in any team play the most significant role in deciding the result of the match. Without them either setting a high target or chasing the target strategically, winning the match would be a very challenging task. Hence modelling the batsmen and choosing them by their rating would be essential for team selection. We rate the batsman by scoring him based on his all-time performance and recent performance as shown in Figure 4. The input features to this model are:

1. *Matches: Number of innings player batted in*
2. *Batting Average: Total Runs divided by (Number of Innings - Not Outs)*
3. *Strike Rate: (Total Runs / Balls faced) * 100*
4. *Number of Fifties*
5. *Number of Hundreds*
6. *Number of 4s and 6s.*

**Algorithm 1:** Modelling Batsmen

**Input:** Players data from IPL 2008 to 2019 seasons, IPL Career Stats
**Output:** Batsmen Rating
**for** *all players in the dataset* **do**
  /* All time Form ( IPL 2008 to 2019 ) */
  matches = Number of innings batted in (Min Max Normalization);
  achievementWeight = (20 * Hundreds) + (10 * Fifties) + (3 * Sixes) + Fours;
  careerStatScore = (0.3 * achievementWeight) + (0.55 * Batsman Average) + (0.15 * StrikeRate);
  OverallCareerScore = (matches * careerStatScore);
**end**
**for** *all players in the recent dataset* **do**
  /* Recent Form ( IPL 2018 to 2019 ) */
  matches = Number of innings batted in (MinMax Normalization);
  achievementWeight = (20 * Hundreds) + (10 * Fifties) + (3 * Sixes) + Fours;
  StatScore = (0.3 * achievementWeight) + (0.55 * Batsman Average) + (0.15 * StrikeRate);
  RecentScore = (matches * StatScore);
**end**
**for** *all players in the merged dataset* **do**
  /* Calculating Batsman Rating */
  OverallCareerScore=OverallCareerScore/max(OverallCareerScore);
  RecentScore = RecentScore / max(RecentScore);
  BatsmanRating = (0.4 * OverallCareerScore) + (0.6 * RecentScore);
  Perform MinMax Normalization(Batsman Rating)
**end**

Figure 4. Modelling Batsmen

The pseudo code of the algorithm to model the batsmen for the playing XI is given in Algorithm 1. The first **for** loop calculates the players *all-time performance*. Higher the *matches* the more batting experience the batsman withholds. Lower value of *matches* indicates that the batsman probably bats later down the order than at the top. *achievementWeight* takes into account the hundreds and the *Fifties* scored and more weightage has been given to *Hundreds. careerStatScore* accounts for all the career statistics and most weightage of 0.55 has been assigned to *BatsmanAverage* because it signifies his potential to score more runs without being out. Higher the value, the higher the batsman's rating. Similar potential is seen with *StrikeRate.* The second **for** loop calculates a score based on the *recent performance* of a batsman similar to *all-time performance* but for the years 2018 and 2019. The third **for** loop calculates the *BatsmanRating* as a 40-60 combination of *all-time* and *recent performance* respectively where more importance is awarded to *recent performance*. The ratings are normalized using MixMax normalization to ensure that they lie in the same range of [0,100].

| | batsman_rating |
|---|---|
| SK Raina | 100.000000 |
| V Kohli | 95.546684 |
| RG Sharma | 85.973005 |
| CH Gayle | 85.805574 |
| MS Dhoni | 83.395706 |
| ... | ... |
| Abdur Razzak | 0.000000 |
| H Brar | 0.000000 |
| C Nanda | 0.000000 |
| S Tyagi | 0.000000 |
| B Stanlake | 0.000000 |

516 rows × 1 columns

Figure 5. Batsmen ratings of all players in the IPL

***Modelling Bowlers:*** The bowlers in each team play a comparatively crucial role in deciding the outcome of the match. We rate the bowler by scoring him based on his all-time performance and recent performance as well as shown in Figure 6. The input features to this model are:

1. *Matches: Number of Innings player bowled in*
2. *Bowling Economy: Number of Runs conceded per over bowled*
3. *Bowling Average: Number of Runs conceded per wicket taken*
4. *4 wicket Hauls: Number of matches where 4+ wickets were taken*
5. *5 wicket Hauls: Number of matches where 5+ wickets were taken*
6. *Wickets Taken: Total wickets taken by the bowlers*
7. *Maiden overs: Number of overs in which no run was conceded*
8. *Number of dot balls: Totals balls on which no run was scored*

**Algorithm 2:** Modelling Bowler

**Input:** Players data from IPL 2008 to 2019 seasons, IPL Career Stats
**Output:** Bowler Rating
**for** *all players in the dataset* **do**
  /* All time Form ( IPL 2008 to 2019 ) */
  matches = Number of innings bowled in (Min Max Normalization);
  wicketWeight = (30 * 5WicktHaul) + (20 * 4WicketHaul) + (10 * WicketsTaken);
  achievementWeight = (5 * MaidenOvers) + dotBalls;
  careerStatScore = (Bowling Average) + (Bowling Economy) + (1 / 1 + achievementWeight);
  OverallCareerScore = (matches * wicketWeight) / (careerStatScore)
**end**
**for** *all players in the recent dataset* **do**
  /* Recent Form ( IPL 2018 to 2019 ) */
  matches = Number of innings bowled in (Min Max Normalization);
  wicketWeight = (30 * 5WicktHaul) + (20 * 4WicketHaul) + (10 * WicketsTaken);
  achievementWeight = (5 * MaidenOvers) + dotBalls;
  StatScore = (Bowling Average) + (Bowling Economy) + (1 / 1 + achievementWeight);
  RecentScore = (matches * wicketWeight) / (StatScore)
**end**
**for** *all players in the merged dataset* **do**
  /* Calculating Bowler Rating */
  OverallCareerScore=OverallCareerScore/max(OverallCareerScore);
  RecentScore = RecentScore / max(RecentScore);
  Bowler Rating = (0.35 * OverallCareerScore) + (0.65 * RecentScore);
  Perform MinMax Normalization(Bowler Rating)
**end**

Figure 6. Modelling Bowlers

The pseudo code of the algorithm to model the bowlers for the playing XI is given in Algorithm 2. The first **for** loop calculates the players *all-time performance*. Higher the *matches* the more bowling experience the bowler withholds. Lower value of *matches* indicates that the bowler probably is an all-rounder or part-time bowler who bowls occasionally. *wicketWeight* takes into account the *5WicketHauls*, *4WicketHauls* and *WicketsTaken* taken by the bowler and more weightage has been given to 5 wicket hauls as it would require exceptionally well and strategic bowling to achieve this feat. achievementWeights takes into account the number of *MaidenOvers* bowled and the total number of *dotBalls* bowled by the bower and more weightage has been given to *MaidenOvers*. *careerStatScore* accounts for all the career statistics and equal weightage has been assigned to *Bowling Average* and *Bowling Economy* because it signifies number of runs, they have conceded per wicket taken and the average number of runs conceded for each over bowled by the

bowler. Lower the value, the higher the bowler's rating. The second **for** loop calculates a score based on the *recent performance* of the bowler similar to *all-time performance* but for the years 2018 and 2019. The third **for** loop calculates the BowlerRating as a 35-65 combination of *all-time* and *recent performance* respectively. The ratings are normalized using MixMax normalization to ensure they lie in the same range of [0,100].

| | bowler_rating |
|---|---|
| Rashid Khan | 100.000000 |
| JJ Bumrah | 93.357948 |
| RA Jadeja | 81.674093 |
| Harbhajan Singh | 79.861995 |
| A Mishra | 79.323012 |
| ... | ... |
| SS Mundhe | 0.000000 |
| SS Agarwal | 0.000000 |
| SR Tendulkar | 0.000000 |
| SPD Smith | 0.000000 |
| AM Rahane | 0.000000 |

Figure 7. Bowlers ratings of all players in the IPL

## VI. EXPERIMENTATION

After a detailed analysis we concluded that the best criterion to choose the Playing XI out of a given squad of 15 players based on the Batsman Rating and Bowler Rating was by using the following combination:

1. *Top 3* batsman chosen based on their batsman rating as specialist batsmen in the team
2. *Top 3* bowlers chosen based on their bowler rating as specialist bowlers in the team
3. The remaining 5 players in the team were chosen based on the *closeness* of their ratings with respect to a perfect all-rounder's rating of [100,100] along with all-time *venue-based* analysis. The data pertaining to a stadium and every match that was played on its pitch was analysed. With the help of this data, we could perform analysis on the combination of selecting the remaining 5 team members for a given venue. The possibilities can be choosing 1 batsman and 4 bowlers, 2 batsmen and 3 bowlers, 3 batsmen and 2 bowlers or 4 batsmen and 1 bowler. Looking at history of the playing XI's chosen in the venue, we can determine whether the stadium is batting biased pitch or a bowling biased pitch. From this, we analysed the combination decisions taken before and based on that we suggest on how the remaining 5 need to be selected.

**Input**: 15 players squad given: SP Narine, A Mishra, DJ Bravo, RA Jadeja, SL Malinga, CH Gayle, AB de Villiers, MS Dhoni, SK Raina, V Kohli, S Gopal, Yuvraj Singh, PP Ojha, MP Stoinis, DA Miller.
**Venue**: M Chinnaswamy Stadium

The measures of closeness were determined by experimenting with the following distance measure techniques:

1. Euclidian distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$ [3]

2. Manhattan distance

$$Mdist = |x2 - x1| + |y2 - y1|$$ [4]

3. Cosine distance

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

The cosine distance is then defined as

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity}$$ [5]

4. Chebyshev distance

$$d(x, y) = \max_{i=1}^{m}|x_i - y_i|$$ [6]

5. Canberra distance

$$d(x, y) = \sum_{i=1}^{m}\frac{|x_i - y_i|}{|x_i + y_i|}$$ [7]

*Playing XI Predictions*

| | Players | batsman_rating | bowler_rating |
|---|---|---|---|
| 0 | SK Raina | 100.000000 | 1.987405 |
| 1 | V Kohli | 95.546684 | 0.060746 |
| 2 | CH Gayle | 85.805574 | 0.821980 |
| 3 | MS Dhoni | 83.395706 | 0.000000 |
| 4 | AB de Villiers | 80.892032 | 0.000000 |
| 5 | SP Narine | 20.635550 | 74.252936 |
| 6 | RA Jadeja | 19.106095 | 81.674093 |
| 7 | DJ Bravo | 18.009862 | 71.519727 |
| 8 | S Gopal | 2.872895 | 70.881631 |
| 9 | A Mishra | 2.241545 | 79.323012 |
| 10 | SL Malinga | 0.380898 | 64.601210 |

Figure 8. Using Euclidian distance

| | Players | batsman_rating | bowler_rating |
|---|---|---|---|
| 0 | SK Raina | 100.000000 | 1.987405 |
| 1 | V Kohli | 95.546684 | 0.060746 |
| 2 | CH Gayle | 85.805574 | 0.821980 |
| 3 | MS Dhoni | 83.395706 | 0.000000 |
| 4 | AB de Villiers | 80.892032 | 0.000000 |
| 5 | SP Narine | 20.635550 | 74.252936 |
| 6 | RA Jadeja | 19.106095 | 81.674093 |
| 7 | DJ Bravo | 18.009862 | 71.519727 |
| 8 | S Gopal | 2.872895 | 70.881631 |
| 9 | A Mishra | 2.241545 | 79.323012 |
| 10 | SL Malinga | 0.380898 | 64.601210 |

Figure 9. Using Manhattan distance

[3] https://www.i2tutorials.com/what-is-the-difference-between-euclidean-manhattan-and-hamming-distances/
[4] https://www.codespeedy.com/compute-manhattan-distance-between-two-points-in-cpp/
[5] https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm
[6] http://www.molmine.com/magma/analysis/distance.htm
[7] http://www.molmine.com/magma/analysis/distance.htm

| | Players | batsman_rating | bowler_rating |
|---|---|---|---|
| 0 | SK Raina | 100.000000 | 1.987405 |
| 1 | V Kohli | 95.546684 | 0.060746 |
| 2 | CH Gayle | 85.805574 | 0.821980 |
| 3 | MS Dhoni | 83.395706 | 0.000000 |
| 4 | AB de Villiers | 80.892032 | 0.000000 |
| 5 | SP Narine | 20.635550 | 74.252936 |
| 6 | RA Jadeja | 19.106095 | 81.674093 |
| 7 | DJ Bravo | 18.009862 | 71.519727 |
| 8 | S Gopal | 2.872895 | 70.881631 |
| 9 | A Mishra | 2.241545 | 79.323012 |
| 10 | SL Malinga | 0.380898 | 64.601210 |

Figure 10. Using Cosine distance

| | Players | batsman_rating | bowler_rating |
|---|---|---|---|
| 0 | SK Raina | 100.000000 | 1.987405 |
| 1 | V Kohli | 95.546684 | 0.060746 |
| 2 | CH Gayle | 85.805574 | 0.821980 |
| 3 | Yuvraj Singh | 31.416723 | 4.695564 |
| 4 | SP Narine | 20.635550 | 74.252936 |
| 5 | RA Jadeja | 19.106095 | 81.674093 |
| 6 | DJ Bravo | 18.009862 | 71.519727 |
| 7 | MP Stoinis | 8.454313 | 2.554853 |
| 8 | S Gopal | 2.872895 | 70.881631 |
| 9 | A Mishra | 2.241545 | 79.323012 |
| 10 | SL Malinga | 0.380898 | 64.601210 |

Figure 11. Using Chebyshev distance

| | Players | batsman_rating | bowler_rating |
|---|---|---|---|
| 0 | SK Raina | 100.000000 | 1.987405 |
| 1 | V Kohli | 95.546684 | 0.060746 |
| 2 | CH Gayle | 85.805574 | 0.821980 |
| 3 | Yuvraj Singh | 31.416723 | 4.695564 |
| 4 | SP Narine | 20.635550 | 74.252936 |
| 5 | RA Jadeja | 19.106095 | 81.674093 |
| 6 | DJ Bravo | 18.009862 | 71.519727 |
| 7 | MP Stoinis | 8.454313 | 2.554853 |
| 8 | S Gopal | 2.872895 | 70.881631 |
| 9 | A Mishra | 2.241545 | 79.323012 |
| 10 | PP Ojha | 0.093018 | 14.382160 |

Figure 12. Using Canberra distance

The predicted Playing XI obtained from Euclidian, Manhattan and Cosine as the closeness measures provided the same result. Whereas, while using Chebyshev distance we find that 'MS Dhoni' was replaced by 'Yuvraj Singh' and 'AB de Villiers' was replaced by 'MP Stoinis'. Although while using Canberra distance measure, along with 'MS Dhoni' being replaced by 'Yuvraj Singh' and 'AB de Villiers' being replaced by 'MP Stoinis', 'SL Malinga' was replaced by 'PP Ojha'. On the basis of domain knowledge, we understand that the results provided by Euclidian, Manhattan and Cosine distance are similar to the real-world statistics and hence, are more accurate for the purpose of distance measure in our work.

***Final Predicted Playing XI using Euclidian distance along with venue-based combination***

| | Players | Nationality | batsman_rating | bowler_rating |
|---|---|---|---|---|
| 0 | SK Raina | Indian | 100.000000 | 1.987405 |
| 1 | V Kohli | Indian | 95.546684 | 0.060746 |
| 2 | CH Gayle | West Indian | 85.805574 | 0.821980 |
| 3 | MS Dhoni | Indian | 83.395706 | 0.000000 |
| 4 | AB de Villiers | South African | 80.892032 | 0.000000 |
| 5 | SP Narine | West Indian | 20.635550 | 74.252936 |
| 6 | RA Jadeja | Indian | 19.106095 | 81.674093 |
| 7 | DJ Bravo | West Indian | 18.009862 | 71.519727 |
| 8 | S Gopal | Indian | 2.872895 | 70.881631 |
| 9 | A Mishra | Indian | 2.241545 | 79.323012 |
| 10 | SL Malinga | Sri Lankan | 0.380898 | 64.601210 |

Figure 10. Final Selection of Playing XI squad

**Output**: Playing XI team
Players selected are mentioned below:
Top 3 Batsmen: SK Raina, V Kohli, CH Gayle
Top 3 Bowlers: RA Jadeja, A Mishra, SP Narine
Middle 5: MS Dhoni, AB de Villiers are 2 batsmen (batting inclined), DJ Bravo, S Gopal and SL Malinga are the 3 bowlers (bowling inclined).

Combination of 5 middle order players predicted for Venue 'M Chinnaswamy Stadium': 2 Batsmen, 3 Bowlers indicating that the stadium is batting biased. Hence out of the 5 chosen, 2 players with good batting rating and 3 players with good bowling rating were selected.

*Toss Decision*

We also suggest Team 1 whether to bat or bowl first for (Team 1 vs Team 2) at a *given venue* by analyzing the previous Head to Head data as shown above in Section IV.

```
team1 = "Royal Challengers Bangalore"
team2 = "Sunrisers Hyderabad"
venue = "M Chinnaswamy Stadium"
bat_or_field(df2, team1, team2, venue)
```

```
Field
```

Figure 11. Toss decision for Team 1 at a given venue against Team 2 by analyzing previous Head to Head data as well as Toss Analysis

Figure 11. shows that for a match between 'Royal Challengers Bangalore' and 'Sunrisers Hyderabad' at 'M Chinnaswamy Stadium' if 'Royal Challengers Bangalore' wins the toss, it would be preferable to choose to field first. This data concurs with Figure 2. which indicates that there is a 55% chance of winning a match at 'M Chinnaswamy Stadium' when the team chose to 'field' first against 45% when the team chose to 'bat' first.

## VII. CONCLUSION AND FUTURE WORK

The proposed work lies in the domain of statistical analysis and performance evaluation. We conclude that our model can be used as a suggestion for the team management in the real-world to aid them in selecting the Playing XI from a given squad of 15 players. It also aids the team management on advising the captain whether to choose to 'bat' or 'field' first at a given venue against a particular team based on the previous Head to Head data and toss analysis.

The novelty in our work is that we have included more features than pre-existing player rating models. Theoretically, the combination of selecting 3 specialist batsmen, 3 specialist bowlers along with 5 all-rounders (1 batsman and 4 bowler, 2 batsmen 3 bowlers, 3 batsmen 2 bowlers or 4 batsmen and 1 bowler) provides an optimum/balanced skill set to a given team for a given match. As our model considers the all-time performance as well as the recent performances of the players (IPL season 2018 and 2019) we can expect sufficient and satisfactory prediction results to aid the team. The suggestion for toss decision is also a key factor that needs to be taken into consideration for constructing a match winning team.

Our future work would consist of taking our model to the next step of predicting outcome of matches as well as analyzing partnerships between various batsmen as a feature input.

## REFERENCES


[1] Vaibhav Khatavkar, Parag Kulkarni. "Context Based Cricket Player Evaluation Using Statistical Analysis". In, International Journal of Knowledge Based Computer Systems 7 (1), June 2019, 01-0

[2] Madan Gopal Jhawar, Vikram Pudi. "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach". European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2016) Conference Center, Riva del Garda.

[3] Hemanta Saikia, Dibyojyoti Bhattacharjee "On Classification of All-rounders of the Indian Premier League (IPL): A Bayesian Approach". Vikalpa: The Journal for Decision Makers, Oct 2011.


## APPENDIX

This section will include the visualization and EDA performed to understand the importance of various features and depending on that deciding the amount of weightage each feature deserves.

| | Runs | Balls | Dismissals | caught | bowled | run out | lbw | caught and bowled | stumped | retired hurt | hit wicket | obstructing the field | Innings | Strike Rate | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DA Warner | 4741.0 | 3345.0 | 108.0 | 69.0 | 23.0 | 4.0 | 4.0 | 2.0 | 5.0 | 0.0 | 1.0 | 0.0 | 1132.0 | 141.733931 | 43.898148 |
| S Dhawan | 4632.0 | 3714.0 | 131.0 | 79.0 | 28.0 | 8.0 | 8.0 | 2.0 | 5.0 | 1.0 | 0.0 | 0.0 | 1407.0 | 124.717286 | 35.358779 |
| MC Henriques | 969.0 | 760.0 | 34.0 | 20.0 | 4.0 | 1.0 | 2.0 | 3.0 | 4.0 | 0.0 | 0.0 | 0.0 | 613.0 | 127.500000 | 28.500000 |
| Yuvraj Singh | 2765.0 | 2143.0 | 107.0 | 84.0 | 10.0 | 2.0 | 5.0 | 2.0 | 3.0 | 0.0 | 1.0 | 0.0 | 793.0 | 129.024732 | 25.841121 |
| DJ Hooda | 535.0 | 426.0 | 32.0 | 20.0 | 6.0 | 2.0 | 1.0 | 0.0 | 2.0 | 0.0 | 1.0 | 0.0 | 403.0 | 125.586854 | 16.718750 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| A Turner | 4.0 | 11.0 | 3.0 | 2.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 11.0 | 36.363636 | 1.333333 |
| H Brar | 22.0 | 14.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.0 | 157.142857 | NaN |
| S Rutherford | 82.0 | 55.0 | 5.0 | 5.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 55.0 | 149.090909 | 16.400000 |
| P Raj | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.000000 | NaN |
| S Singh | 17.0 | 18.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 18.0 | 94.444444 | 17.000000 |

516 rows × 15 columns

Figure A1. shows the batting record of all batsman in IPL

Figure A1. was calculated based on the new dataset made from the files obtained from the dataset on Kaggle.

Figure A2. shows the number of dismissals categorized into 3 categories – Caught, Stumped, Run out. Based on this, we can perform analysis on the number of times a batsman has got 'out' on each type of dismissal to understand his weakness and the area on which he must work up on.

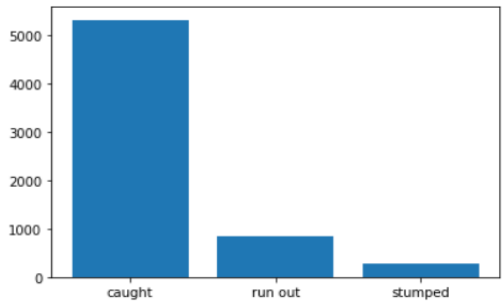


Figure A2. Dismissal Analysis

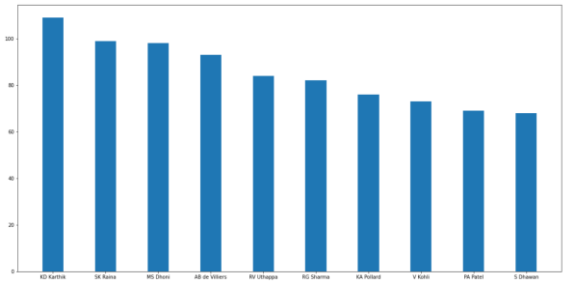

Figure A3. Most Caught Dismissal analysis of 10 batsmen dismissal analysis of 10 players in the history of IPL

The drawback in the Figure A3 and Figure A4 are that it depends on the number of matches a certain player has been chosen to play in. For example, MS Dhoni being one of the players who has participated in the most matches among the 10 players named in the figures as can be seen in Figure A4.

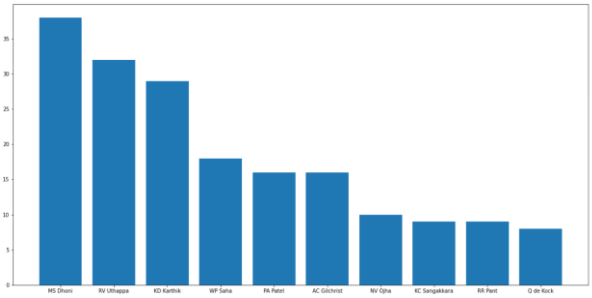

Figure A4. Most Stumped Dismissal analysis of 10 batsmen in the history of IPL

## CONTRIBUTION OF EACH TEAM MEMBER

The whole project was divided into 3 major sections and for each section was led by 3 different team members.

1. Pre-processing Visualization/EDA
   - Section head: Lavitra Kshitij Madan
   - Other contributors: Ritik Hariani, Arjun Chengappa, Aditya Burli
2. Modelling Algorithm and Design
   - Section head: Ritik Hariani
   - Other contributors: Arjun Chengappa, Lavitra Kshitij Madan
3. Playing XI Selection and Toss Decision model
   - Section head: Arjun Chengappa
   - Other contributors: Lavitra Kshitij Madan, Ritik Hariani