**How Data Analytics is used in performance-analysis of cricket player for making good team prediction based on player statistics**

**Introduction:**

Cricket is the most popular game among many countries in the world. It originated from England and is being played since 19th century. Cricket is a game played between two teams having eleven players aside [1]. Teams play cricket in three formats like TEST, ODI, and T-20. T-20 format is the most interesting short-game in cricket and played within 3 hours. It is important for any international cricket board to win the game. Cricket Board of different countries, Team Managements, Coaches will prefer player analysis, which helps in making a good team for their country. Building a team involves selecting good Batsman, Bowlers, All-rounders and a Wicket keeper [1] [9].

The present research uses different Data analytics tools (AHP- TOPSIS, AHP-COPRAS, MCDA (multiple criteria decision analysis) [1], Factor analysis [8] Naïve Bayesian [7], Statistical modelling with Hadoop [9]) in choosing good Batsmen and bowlers for a team and a new method which involves a two-stage regression and OWA method [2]. The literature review [5][6] uses Graphical methods to analyze Batting and Bowling Performances. [1] uses MCDA approach for Bowling analysis whereas [8] uses factor analysis approach for Player analysis. These analytical methods are helpful to selectors where they can pick the best eleven players that includes Batsman and Bowlers. Making a team prediction not only include batting and bowling performances but also wicket keeping and fielding: as these performances are very useful in cricket to stop the runs scoring by the opponents in the ground. There are many studies in analyzing the player performances of Batting and Bowling in cricket. But there is no analysis that put together to analyze all the methodologies that compare Batting and bowling performances of the players.

The purpose of this Review of literature is to fill the gap in the literature, by comparing methodologies that analyze the Batting and Bowling performances and also suggesting the need to analyze the wicket Keeping and Fielding performances of players. My primary research question is how Data Analytical Methods are used in analysis of players performance in Cricket to make a good team prediction.

**Player Analysis:**

This section introduces what does player analysis mean and explain the concept of player analysis in cricket and also show which parameters are required to analyze the players performances by extracting Data sets of players.

Player analysis mean analyzing the player performances to what extent the player is good in his game (cricket). It is done based on collecting players past performances data like number of matches played, number of matches batted, Number of matches bowled. For measuring the Batting performances, we need five parameters like Highest Individual Score, number of centuries scored, number of fifties scored, average batting performance, batting strike rate, number of fours and sixes hitten [2] [7] [8]. Out of these parameters strike rate (scoring more runs in less balls) is considered the most important parameter in analyzing for T-20 format [2].

Bowlers are divided in to two types like Pace Bowlers and Spin Bowlers [1]. Pace bowlers will bowl medium-fast at an average speed of 100-145 Km/hr and it is easy for the batter to play whereas Spin Bowlers will bowl at a speed of 60-90 Km/hr and it is hard for batters to predict the ball whether it comes on-to the bat or not. However, Bowling performances is analyzed by taking six parameters which include Bowling economy rate, strike rate, total number of wickets taken, total number of overs bowled, total number of matches played [2] [8]. For All-rounders both Batting and Bowling performances are drawn for making the analysis.

**Data analytical Methodologies used for player analysis:**

This section will introduce and compare the Methodologies used in analyzing the player performances of Batting, Bowling and All-Rounder.

**Batting Analysis:**

Batting analysis of a player is done in different methodologies, for instance [2] suggests two stage regression method for analyzing batting performances. OWA (ordered weighted average) operator and Regression methods are used for ranking the parameters of player batting analysis. This method uses the parameters Average, strike rate, Highest score, fours, sixes to analyze the batting performances. OWA parameter has two important measures dispersion and orness [2]. Dispersion is used to find probability distribution of a batsman and orness is used as a unit interval that can make a decision maker to rank the batsman [2]. The application of this OWA parameter is to generate the OWA weights. By different generated OWA weights and its uncertainty levels, the ranking given for batting is not sensitive to the change of weights [2].

Another way to analyze the batting performances is [11] uses Individual and Moving Range control charts for analyzing the batting performance of players. Data of players is collected and analysis is done with Statistical Tool R where p-plot and Histogram and control charts are displayed as output for player performances [11]. This method also uses basic sensitizing rules for calculating the batting performances to check whether in control or not. In contrast to [2] and [11],[4] used different Batting parameters in analyzing the batting performance like continuous adjusted average ( the performance of the player is consistent), Batsman impacting score (performance of a player in particular match), Runs impact score (ratio of single player to all the players in the team) , Strike rate impact score (strike rate of a player is greater than the mean strike rate of match), pressure impact score( It is the position when player plays after number of good wickets fallen and there is a less chance of winning) and chasing impact score ( playing in the second innings to his best to win the team). [2] used 40 players for analysis, [11] used only 2 players for analyzing batting performances whereas [9] can analyze unlimited players by using statistical modelling approach for measuring the performances of batting by adopting to Hadoop and Hive environments. Thus, Average, strike rate, Highest score are the important parameters used in analyzing the batting performances of a every player [2] [11] [9].

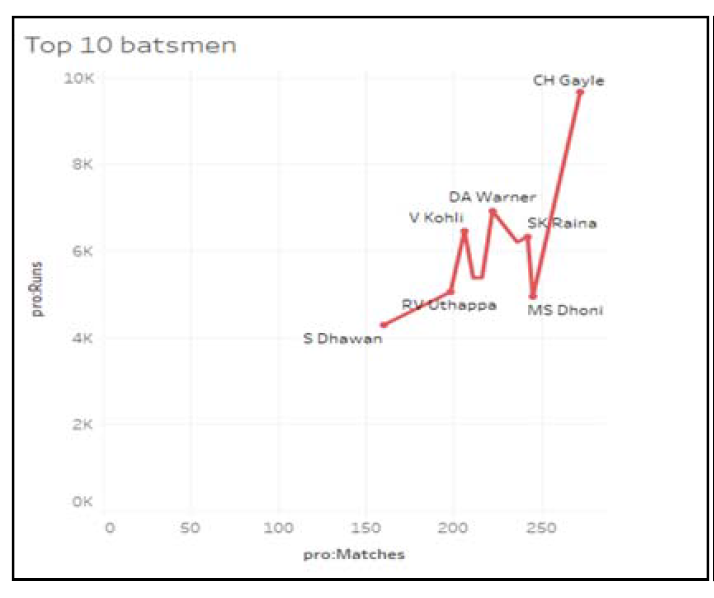


Figure 1: Batting analysis of players using H-base [3]

**Bowling Analysis:**

Bowling analysis is done by considering the parameters like Bowling Economy rate, Bowling Average rate, Bowling strike rate, Total number of wickets taken, Total number of overs played, Total number of matches played [1]. The MCDM (Multiple criteria decision analysis) approach is used for evaluating bowler’s performances [1]. Using MCDM Bowling analysis is done in considering two formats i.e. Fast bowlers and Spinners. MCDM approach can analyze the large number of parameters and offer valuable approach for decision makers.

For instance [1] uses MCDM to analyze all the six parameters to measure the bowling performances. MCDM uses two methodologies AHP-Topsis and AHP-copras [1]. MCDM is suitable for finding the raw scores in decision matrix and helps to do sensitivity analysis which helps in the sequence of ranking the player performances [1]. Factor analysis, Naïve Bayesian and Statistical modelling with Hadoop are the methodologies also used in evaluating the bowling performances of a player [8] [7] [9]. In contrast to [1][7] and [9], [8] et al. considers three different Bowling parameters like Bowling economy rate as TR/O (TR is total numbers of runs conceded and O is number of overs bowled), Bowling Average as TR/W (TR is total number of runs conceded and W is the number of wickets taken), Bowling strike rate as TB/W (TB is total number of balls bowled and W is number of wickets taken by player) for analyzing the bowling performances.

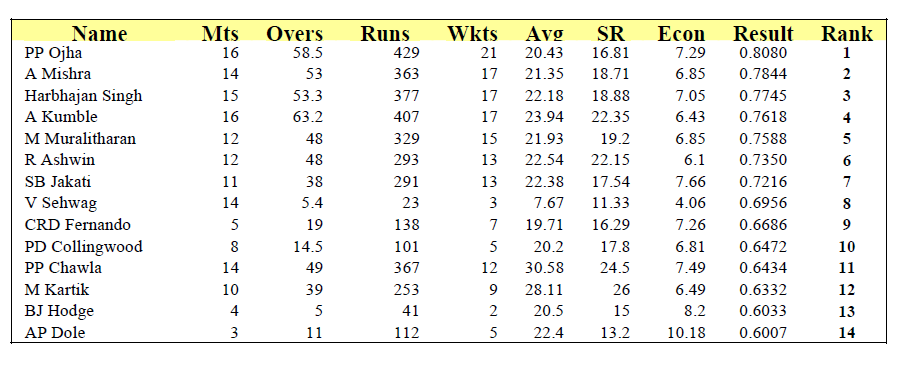


Figure 2: Bowling analysis of players by MCDM approach [1]

**All-rounder performance’s:**

Allrounder performances is defined as Both Batting and Bowling performances of a player. Factor analysis is the new statistical technique used in analyzing the data of player performances like Batting, Bowling and All-rounders [8]. For measuring both batting and bowling performances of players we can adopt to factor analysis method. In this method, data of only 85 players can be considered for analysis, Factor analysis is also known for multivariate data analysis [8]. Factor analysis helps in finding the accurate results even by considering various dimensions of Batting and Bowling. The purpose of factor analysis is we can maximize the variance of the players and it uses factor estimated method done on the basis of principal component analysis where data doesn’t require normal type and specific common factors [8].

Bayesian approach is another method used for measuring the performance of All-rounders [7]. The purpose of using naïve Bayesian model is to predict the player performances. This methodology splits all-rounders into four types Performer(All-rounder with strike rate above median and economy rate below median), Batting All-rounder(All-rounder with both strike rate and economy rate above median), Bowling All-rounder( All-rounder with both strike rate and economy rate below median), under performer(All-rounder with strike rate below median and economy rate above median).It can analyze many number of independent continuous variables and assumes the effect of variable values in a given class is independent of other variables such assumptions are called class conditional independence, these assumptions are not accurate but naïve Bayesian classification model is used in finding accurate results for small data sets like few players performances not more than 40 players [7] et al.

In contrast to [7] and [8] where these methods can analyze the data of just 85 and 40 players, the statistical modelling with Hadoop can analyze unlimited players by adopting the data bases like Hadoop and Hive platforms to make a good team prediction [9]. Statistical modelling is done on player stats of the matches he played in cricket. This model analysis all the performances of player like Batting, Bowling and All-rounders. Hadoop is an open-source platform where large amounts of data is stored and manipulated and the purpose of choosing the Hadoop is where it can form all statistical methodologies and the data of the player is generated in future where it can be added and used for analysis purpose [9]. Hive is a data warehouse which process the data stored in hadoop for analyzing and it also used to analyze the big data [9] et al. The results obtained with this methodology is 91 % accurate when compared to other methodologies.

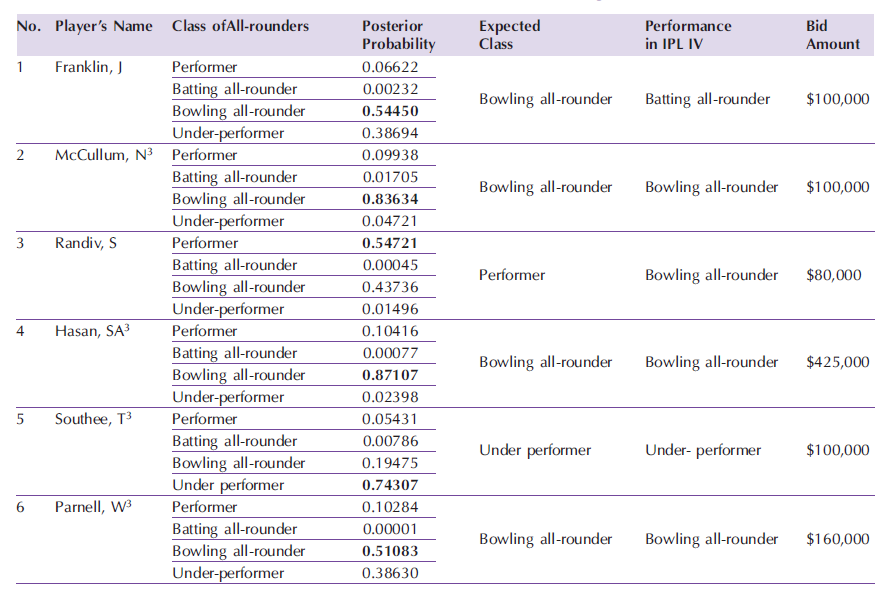


Figure 3: All-rounder performances of player using Bayesian approach [7]

**Player Analysis and its Implementations:**

This section represents how player analysis can help selectors, franchise and cricket boards to make good team prediction.

Player Analysis help franchises, sponsors, cricket selectors and cricket-boards of different countries to make a good team prediction. Indian premier league is a T-20 format cricket League which is played every year in India since 2008. The management of 8 different teams, auction the players to make a good team. They rely on player analysis for good decision making in team, which player to be auctioned, and at what price the player to be auctioned [1][3]. Sponsors for advertising their products prefer player analysis and pick a player at the beginning stage of his career, which helps them in future for the endorsements [4]. Cricket team selection committees also prefer player analysis to make the probable eleven key players for the upcoming series and major league tournaments [9] [10]. [1][2][3] [9] [10] have learned player analysis in cricket is very important for making decisions in player selection, franchises to auction and make a team, sponsors for their endorsements, Team Managements for picking best eleven players.

**Discussions and its Future work:**

In this paper, I have put together all the methodologies (OWA, MCDM approach, Factor analysis, Bayesian, statistical modelling) that have been used in analyzing the player batting, bowling performances. In addition to Naïve Bayesian and Factor Analysis (where these methods analyzed only data sets of 35 and 85 players), the statistical modelling with Hadoop is best method which can predict the accurate results in making a good team prediction and analyze large data sets ( Ex: 10000 players) by adopting the data bases like Hadoop and Hive platforms [10]. Cricket boards, Franchises of IPL, and sponsors who prefer these player analyses for selecting Good players and for their endorsements.

There are many articles which analyze both the bowling and batting performances for making a good team prediction, but there are no articles till now which focus on analyzing the parameters of wicket-keeping and Fielding performances of players. Most of the Franchises in IPL lose matches because they do not focus on selecting the player based on their wicket-keeping and fielding performances. Therefore, I suggest that both wicket-keeping and fielding performances must be taken into consideration while making a good-team prediction. There is more research available to do in future by analyzing the parameters of both wicketkeeping and fielding performances of players because there are many matches that are lost due to Fielding (Dropping catches and Miss-Field) and wicket keeping (Missing Stump outs, Run outs).

Also, I suggest the scholars to derive a method that can have 100% accuracy rate, as the highest accuracy (as of now) is 91% achieved by using statistical modelling (with Hadoop) method. There is also loss (data is not recorded) of players data while analyzing their performances, which in-turn effects the players career through this player analysis. There is more data being brought by the players in upcoming matches, more data mean more analysis can be done in future.

# References

|  |  |
| --- | --- |
| [1] | P. K. Dey, D. N. Ghosh and A. C. Mondal, "A MCDM approach for evaluating Bowlers performance in IPL," *Journal of Emerging Trends in computing and Information sciences,* 2011. |
| [2] | A. R and S. K. Sharma, "Measuring Batting Parameters in Cricket: A two stage regression-OWA method," *www.elsevier.com/locate/measurement,* 2014. |
| [3] | S. Singh and P. Kaur, "IPL Visualization and Prediction Using HBase," *Information Technology and Quantitative Management,* 2017. |
| [4] | A. Kumar and D. Sindhu, "Reflecting Against Perception: Data Analysis of IPL Batsman," *International Journal of Engineering Science Invention,* vol. ISSN, pp. 2319-6734, 2014. |
| [5] | A. Kimber, "A graphical display for comparing bowlers in cricket," *Teaching statistics,* pp. 84-86, 1993. |
| [6] | P. Van staden, "Comparision of Cricketers bowling and batting performances using graphical displays," *Current science,* pp. 764-766, 2009. |
| [7] | H. Saika and D. Bhattacharjee, "On classification of All-rounders of the Indian Premier League: A Bayesian Approach," *VIKALPA,* vol. 36, 2011. |
| [8] | S. K. Sharma, "A Factor Analysis Approach in performance analysis of T-20 cricket," *Journal of Reliability and Statistical Studies,* vol. 6, no. 1, pp. 69-76, 2013. |
| [9] | S. Agarwal, L. Yadav and S. Mehta, "Cricket Team Prediction with Hadoop: Statistical Modelling Approach," *Information Technology and Quantitative Management,* 2017. |
| [10] | F. Ahmeda, K. Deba and A. Jindal, "Multi-objective opimization and decision making approaches to cricket team selection," *www.elsevier.com/locate/asoc,* 2011. |
| [11] | M. Daniyal, T. Nawaz, I. Mubeen and M. Aleem, "Analysis of Batting Performance in Cricket using individual and Moving Range (MR) control charts," *International Journal of sports Science and Engineering ,* vol. 6, no. 4, pp. 195-202, 2012. |