2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

Predicting Optimal Cricket Team using Data Analysis

Ranjana Jadhav
Department of Information Technology
VIT, Pune
Pune, India
ranjana.jadhav@vit.edu

Bhargav Pawar
Department of Information Technology
VIT, Pune
Pune, India
bhargav.pawar18@vit.edu

Nishant Bhat
Department of Information Technology
VIT, Pune
Pune, India
nishant.bhat18@vit.edu

Shyam Kawale
Department of Information Technology
VIT, Pune
Pune, India
shyam.kawale18@vit.edu

Abhijit Gawai
Department of Information Technology
VIT, Pune
Pune, India
abhijit.gawai18@vit.edu

2021 International Conference on Emerging Smart Computing and Informatics (ESCI) | 978-1-7281-8519-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ESCI50559.2021.9396861

***Abstract*: Cricket has always been an integral part of the Indian culture. In India, Cricket is a religion followed by close to 1 billion hardcore cricket fans. Team selection plays one of the most important roles in deciding the outcome of a match. There are various factors that affect the performances of players including form, venue, previous records. Generally, fifteen or sixteen players form a cricket squad. A squad contains different players whose expertise range from batting, bowling, and wicket keeping. The squad needs to be well balanced in its construction for it to play to its maximum potential. Taking into consideration all these factors, deciding you're playing 11 is easier said than done. A human selection committee will invariably suffer from the shortcomings of unfair and biased judgment, human error, and overlooking of certain important aspects. Here's where data science and optimization come into the picture. Data Science is a growing field that is playing a pivotal role in many fields and cricket is no exception. Cricket is a sport which can be associated with a lot of data parameters due to the nature of the game. The insights that one can get from this data can provide people with a lot of information which in turn can help them make decisions. This data can be processed to get key information about the parameters that affect the selection process. In this project we are considering around 35 parameters that affect the player performances which in turn will help to build an optimized squad to maximize the performance of individuals.**

Keywords: Data Science, Parameters, Optimized Squad.

I. INTRODUCTION

T-20 format of cricket is slowly becoming the most popular format of the game. The cricket world is dominated by T-20 leagues. Whether it is in casual conversation or public debate, team selection is always an important factor. Each cricket board appoints a committee of selectors generally comprising of veteran cricketers. The selectors use their experience and expertise to select the team based on recent performance of the available cricketers. The responsibility of the selectors is not only to choose the best players but a balanced team with in-form players from all the different expertise like specialist batsmen, spin bowlers, fast bowlers, all-rounders and wicket keepers. The conventional team selection process, although used quite successfully, has got several shortcomings. But because of human errors, some parameters of players might be ignored. There is a possibility of the selection board becoming partial to certain players, who are selected in spite of sub-par performances in matches. Also, due to the vast pool of players to pick from, a few talented players who deserve a berth in the national team may get overlooked. Therefore, a need arises to devise an automated system that would select the best players in the national team, essentially to do away with the discrepancies of the manual selection process. Such an automated system would typically be software that would grade a player on pre-defined set of conditions and generate an output that consists of a team which is most suitable for that particular scenario. The various parameters considered for selection process are venue of the match, the opponent and batsmen and bowler specific statistics also the required team composition. Using these parameters our system would predict a team that will be best suited for the tournament. Even fantasy 11 games are gaining popularity with close to a million people being active on such sites. Such an automated system can also be used by the users to make an optimal team which will give them an advantage while playing such games.

II. LITERATURE REVIEW

Kamble et al. [1] presented a selection procedure. The selection procedure used analytical hierarchical process to choose some subset of players. A universal set of cricketers was taken comprising of batsmen, bowlers, all-rounders and wicket keepers.

Barr and Kantor et al. [2] used portfolio analysis to determine the set of batsmen who are supposed to be more suitable for a given one-day squad.

Gerber and Sharp et al. [3] proposed an integer programming technique for selecting a limited over squad or 15 players rather than playing 11 players. The method included to collect data from 32 South African players to select the ODI squad of 15 players.

Extending the same idea as et al. [3], Lourens et al. [4] selected an optimal Twenty20 South African cricket side based on performance statistics of a host of players who participated in the SA domestic Pro20 cricket tournament.

Using an integer programming, Brettenny et al. [5] selected players for a fantasy league cricket team under certain pre-specified bugetary constraints, but with a progressive approach. This optimal team at each stage of the

tournament, considered the performance of available cricketers up to the previous match.

Though most authors used the binary integer programming tool for the purpose of the optimized team selection, yet they used different tools for measuring the performance of cricketers. Some authors used the traditional statistics like batting average, strike rate, etc. to quantify the performance of cricketers, while many authors tried to combine some traditional measures to a completely refine the statistic to evaluate performance of player.

Lemmer et al. [6] devised a measure known as combined bowling rate (CBR) to measure the performance of bowlers by combining the three traditional bowling statistics *viz.* bowling average, economy rate and bowling strike rate. Harmonic mean was also used to combine the traditional bowling statistics. A measure of batting performance or (BP) to assess the batsmen's performance in limited over cricket was developed.

III. RESULT AND ANALYSIS

Sr. no	Previous Paper	This Paper
[1]	Analytical Hierarchical process (AHP) is used.	Statistics and Weights are used to measure performance of players.
[2]	Only Batsmen were selected.	Complete balanced team including batsman, bowler all-rounder and wicketkeeper selected.
[3]	Squad (15) players are selected for ODI.	Team (11) players are selected for Twenty-Twenty.
[4]	Squad (15) players are selected for Twenty-Twenty.	Team (11) players are selected for Twenty-Twenty.
[5]	Weather parameters are not used.	More parameters like Weather, Consistency are used.
[6]	Combined bowling rate (CBR) (by combining bowling average, economy rate and bowling strike rate) used to measure the performance of bowlers.	Bowling average, economy rate and bowling strike used separately.

IV. METHODOLOGY

A. Data and tools:

We have collected the data from various sites including Kaggle, Espncricinfo, howstat and cricmetric. The data varies from player statistics, team statistics also classifying them on the basis of various factors that we are considering which affect the player and team performances. We have considered the data in between the timeframe of 2008-2018. We imported all the data into excel sheets and used pandas' libraries to process and manipulate the data. We have uploaded our database into fire base in order to access the same on our app.

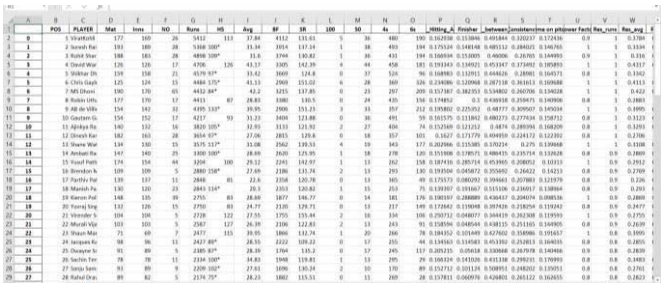

Fig. 1. Batting Database

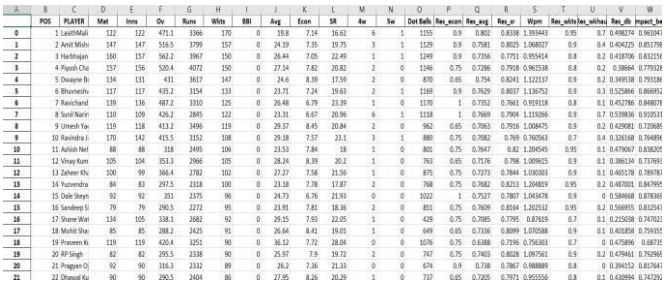

Fig. 2. Bowling Database

B. Data pre-processing:

1) Batting Parameters:

1. Innings Played: This states the number of matches played by the batsmen. This implies the experience of the batsmen. The more the number of matches played by the. batsmen, the better it is for the team.

2. Runs: This indicates the runs scored by the batsmen. It directly corelates to the performance of the batsmen over the years.

3. Consistency: This parameter tells you about the consistency showed by the batsmen in his performances. It is not always about scoring 50s and 100s but more about how frequently can the player maintain his form and performances over the course of time and how even the gap is between these performances.

4. Average: This is basically the number of runs per matches played by the player. The better the average the more can the team depend on the batsmen.

5. Strike Rate: This parameter gives the scoring rate per 100 balls faced. It indicates how quickly can a batsman score his runs.

6. No of 50s and 100s: This gives the sum total of the half centuries and centuries. It indicates the capability of the batsmen to play a big inning.

7. Hard Hitting Ability: Today hard hitting has become a very important part of the game, especially in the T-20 format of the game. How much a team scores in the death overs plays an important role in deciding the result of the game.

8. Finisher: This parameter focuses on the ability of the batsmen to finish out the innings and take the team home.

9. Running between the wickets: This parameter considers the ability of the players to rotate the strike and keep the scoreboard ticking. It is one of the most underrated aspects of the game vital in situations where the team has lost some early wickets.

10. Power Factor: This focuses on the ability of the batsmen to play the big innings as well as maintain a health run rate.

11. Time on pitch: This factors in your ability to play the big innings and build crucial partnerships for the team.

2) Bowling Parameters:

1. Innings Played: This states the number of matches played by the bowler. This implies the experience of the bowler. The more the number of matches played by the bowler, the better it is for the team.
2. Wickets: This is a simple parameter indicating the number of wickets picked by the bowler.
3. Average: This gives the ratio of the number of runs conceded per wickets taken. A lower bowling average indicates that the bowler is doing well.
4. Economy Rate: Economy rate gives the number of runs a bowler is conceding per over. Lower the economy rate, the better it is for the bowler.
5. Dot Balls Index: This parameter factors in the ability of the bowler to stop the batsmen form scoring runs. It is a very underrated factor of the game as studies have shown that a healthy dot ball record often leads to wickets.
6. 4 and 5 Wicket hauls: It factors in the ability of the bowler to play a match changing role in the game and how frequently they do it. A bigger wicket haul puts the opposition.
7. Strike Rate: This parameter states the wicket taking ability of the bowler per bowls bowled. The lower the strike rate, more effective the bowler. Strike rate even though more efficient in the longer formats still plays an important role in the T-20 format.
8. Pitch Impact: This parameter factors in the average economy rate of the bowlers at that particular venue and then compares it with the individual statistics of the bowlers thus giving us a far better idea as to which bowlers are suited to which pitches.

3) All-Rounders Parameters:

1. For all-rounders we are considering a combination of the above-mentioned parameters. Importance is given to the individual parameters depending on the pitch condition and whether it favours the bowling or batting all-rounder.

4) Wicket-Keeping Parameters:

1. Dismissals/Match: This parameter gives us the number of dismissals a keeper is enforcing in a single match.
2. Stumpings: It factors in the number of stumpings enforced by the keeper.
3. Catches: This states the number of catches taken by the keeper.
4. Mistake factor: This is a combination of dropped catches and missed stumpings. The less the mistake factor, the better the keeper.

In the modern game it is not only necessary for a keeper to be good behind the wickets but also be handy with the bat especially in the shorter formats. In the parameters for a wicket keeper, we are also considering the above-mentioned batting parameters so as to select a player who is adequate in both the departments.

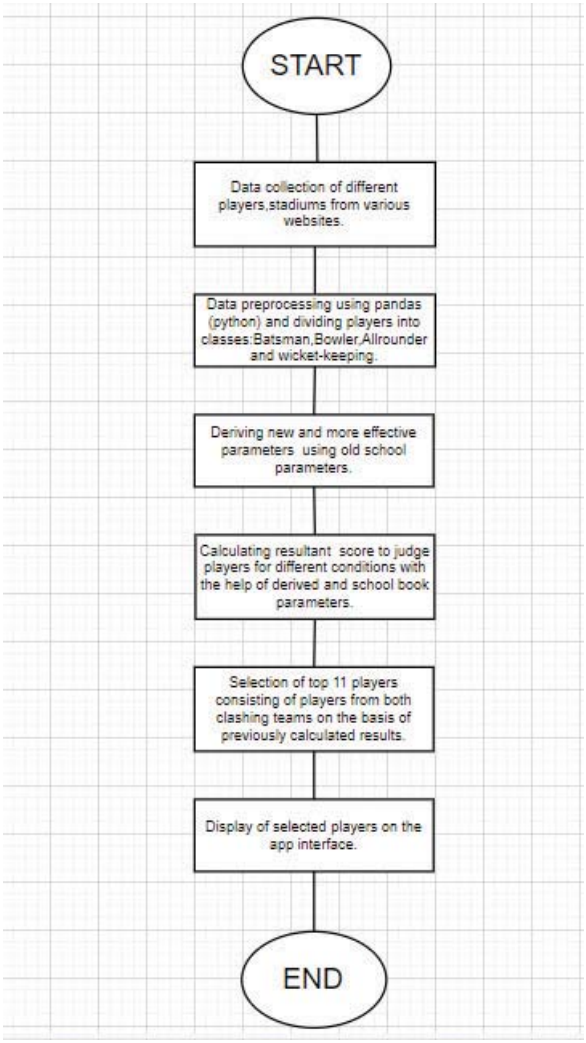

Fig. 3. Flowchart

C. Calculating derived attributes:

As the attributes that are generally available like strike rate, average, wickets are one dimensional and don't tell the whole picture so we have calculated derived attributes such that with the combination of the above-mentioned parameters will give us a clearer picture.

1) Batting:

For batting we have calculated quite a few derivedparameters such as:

1. Hard Hitting Ability: It gives us a clearer picture how much a batsman is capable of using the long handle or how much he can deliver to increase the current run rate.
   Hard Hitting Ability= (Number of fours + Number of Sixes)/Balls Faced.
2. Finisher :-It tells us about the match finishing capability of the player that how many times he can take his team over the line or possibility of taking his team to a respectable total while he is at the crease. Finisher=Number of not outs/total number of innings.
3. Running between wickets: It shows how much the batsman can be able to steal runs when big runs (4s,6s) are hard to come. We can also get a hint of

fitness measure of a particular player. Running Between wickets= (Total runs-4*Number of fours-6*Number of sixes)/ (Balls faced-Number of fours - Number of Sixes).

4. Consistency: Just like average it tells how much a player is capable of scoring big runs. Consistency= (Amount of runs/Number of innings)/100.
5. Time on pitch: It gives us information about how much a team can rely on a particular player when wickets are falling from the other end. Time on pitch= ((Balls Faced/Number of Inns) * 0.7087)/100.
6. Power Factor: It gives us a measure of potential that a batsman has of changing the course of game on his own.
   Power Factor=1 if (Avg+ Strike Rate) >170
   Power Factor=0.9 if (Avg+ Strike Rate) >=160 and <170
   Else Power Factor=0.8.

2) Bowling:

Derived parameters for bowling are:

1. Resultant economy: Assigning a certain value to a bowler based on his overall economy. Res_econ= 1 if economy<7

   Res_eon=0.9 if economy<7.5 and >7
   Res_econ=0.75 if economy<8 and >7.5
   Res_econ=0.65 if economy<8.5 and >8
   Else Res_econ=0.55.

2. Resultant average and Resultant strike rate, which simply tells about average and strike of a bowler.

   Resultant average=(100-average)/100
   Resultant strike rate= (100-strike rate)/100.
3. Wickets per innings: - It tells how much a bowler is

   capable of taking wickets at crucial times
   Wickets per match=Number of wickets/Total innings.
4. Dot Ball Index: - It tells how much dot balls a bowler can bowl in a particular match and it shows how a bowler can keep opposition under pressure.

   Dot ball index=dot balls/number of innings
5. Pitch Impact: how a particular venue will affect the economy of a particular bowler.

   Pitch impact (Bengaluru)= ((8.3675/Econ) +(29.125/Avg))/2.75
   Pitch impact (Chennai)= ((7.2625/Econ) +(26.875/Avg))/2.45
   Pitch impact (Mumbai)=. ((7.805/Econ) +(27.3125/Avg))/2.55

3) Wicket-Keeping:

There is only one derived parameter particularly for wicket keeping while mostly the derived and old school parameters of batman are used: -

1. Dismissals per innings: - It tells us about the number of catches taken and stumpings done by the wicket keeper in an innings.
2. Dismissals per match= (number of catches +number of stumpings) Total innings.

D. Resultant Formula and weight calculation:

In order to calculate the resultant formula, we have to assign different weights to different parameters such that they will give us a proper score to select the top-notch players in our team.

While selecting Batsman we have given it a total weight of 550 and assigned weightage to different parameters from the total weight on the basis of the venue and the importance of the particular parameter

e.g. Res_ben:-which is resultant score for classifying players when venue is Bengaluru.

Res_ben= (30/550)*Res_runs+(65/550)*Res_avg")+(100/550)*Res_SR +(50/550)*Res_big_score+(110/550)*Hard_Hitting_Ability +(80/550)*Finisher+(30/550)*Running_between_wickets+(35/550)+Consistency+(15/550)*Time on pitch+(35/550)*Power Factor.

As stadium in Bengaluru is a batting pitch so more emphasis has been given on strike rate, hard hitting ability and finishing ability and lesser weight is given to others. In the same way the weightage on other venues vary as the conditions vary e.g., pitch in Chennai is bowling track so more emphasis is given on average, running between the wickets and consistency lesser on other parameters and in case of Mumbai it is average for both batting and bowling so mean weightage of batting and bowling pitches is given to the parameters.

Similarly, in case of the bowlers the total weightage is 550 and weightage to parameters is given on the venue. e.g., if the match is in Chennai which is a bowling pitch so more weightage is given to wicket taking ability that is average and Resultant wickets. So the formulae for calculating the score is

Res_chen=(90/550)*Res_econ+(95/550)*Res_avg+(80/550) *Res_sr+(95/550)*Res_wkts+(20/550)*Res_wkhaul+(80/550)*Impact_chen+(70/550)*Res_db+(20/550)*Res_exp

In the same way we have calculated the resultant score of the Bengaluru and Mumbai pitch for bowlers by just little variation in the weightage of the parameters. As Bengaluru is batting wicket so more emphasis on dot balls bowled and the economy of the bowler and in case of Mumbai average weightage of the tracks is taken.

Now the next we calculate the score for the wicketkeepers. Wicketkeepers parameters mainly consist of batting parameters. So, we taken in account batting performance of wicket-keeper on a particular venue and his wicket-keeping skills.

e.g Res_mum=0.6*Bat_mum+0.4*Res_wktk

Here resultant score for wk-keeper in Mumbai is calculated by giving 60% weightage to his batting in Mumbai and 40% to his wicket-keeping skills.

In the same way the scores of both Chennai and Bengaluru is calculated by just changing the batting record of the player.

In the case of allrounders the parameters of both batting and bowling are combined and their weightage is changed according to the pitch

1. ARes_ben=0.6*RES_BEN+0.4*BRES_BEN
2. ARes_chen=0.4*RES_CHE+0.6*BRES_CHE
3. ARes_mum=(RES_MUM+BRES_MUM)/2

As discussed earlier Bengaluru pitch rewards batting more so 60% weightage given to batting and only 40% to bowling and in Chennai its vice-versa. In Mumbai its average for both so mean of both batting and bowling is taken there.

V. RESULT

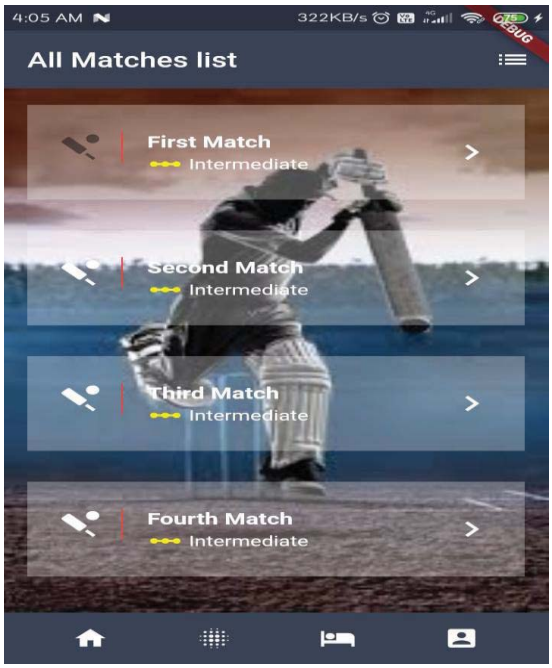

Fig. 4. User Interface

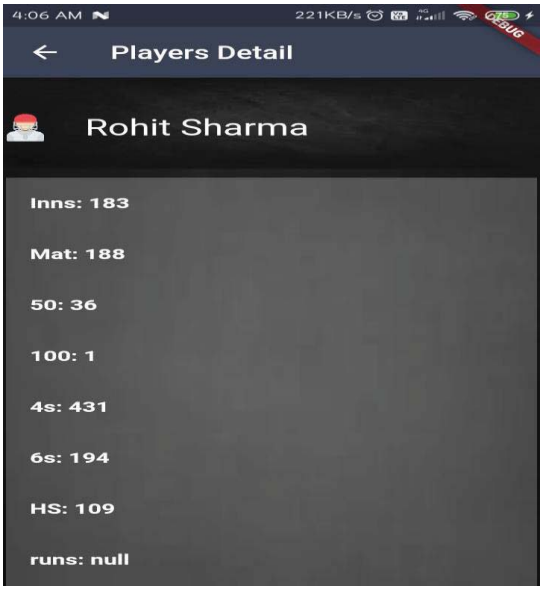

Fig. 5. Player Details

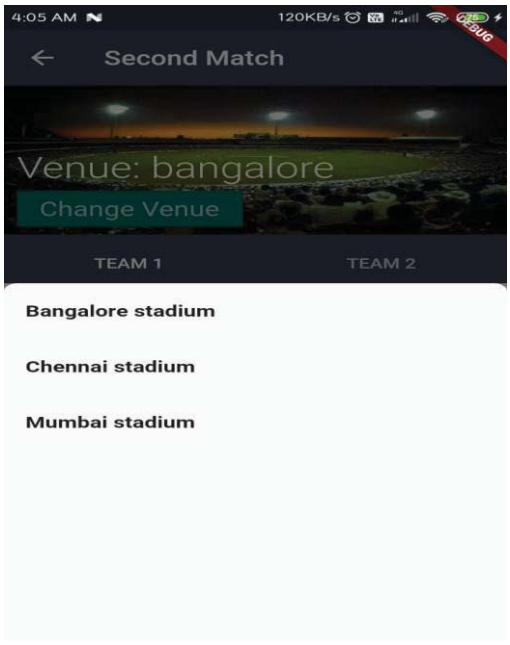

Fig. 6. Venues and pitch Impact

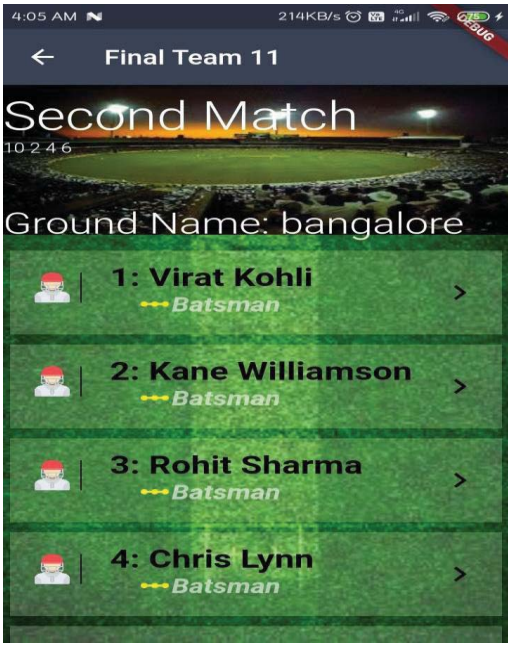

Fig. 7. Final Team

VI. CONCLUSION

Selecting an optimal starting team is a difficult task. Various factors have to be considered barring subjectivity. An individual up for a selection in the squad needs to be screened for his form, the oppositions strength and even whether that players style suits the pitch pattern. Such difficulties are removed using the automated mechanism mentioned in the paper. This will be a winning formula considering the parameters encompass all the shortcomings of the squad along with its strengths. Such a mechanism can also prove to be useful while playing fantasy 11 game considering that the factors can be applied to both the teams making the necessary changes in the database.

VII. Future Scope

This sort of mechanism can also be applied to a variety of games considering the necessary data is made available. Almost all team games can use this depending on the variation of parameters in those particular games. Machine Learning algorithms can be introduced to predict the outcome of matches. Emotional parameters can also be considered bringing into the game a human quotient which at most of the times is lacking. We can also consider different weather patterns and their effect on the pitch given the necessary data is made available.

Acknowledgment

Our team would like to take this opportunity to express our gratitude to the following professionals: Firstly, we would like to thank the Honorable Director of Vishwakarma Institute of Technology, Prof. (Dr.) R.M. Jalnekar Sir for including the concept of Engineering Design and Innovation (EDI) in our syllabus, and our HOD Premanand Ghadekar for giving us this opportunity to work on this innovative project Finally, we would specially like to thank our guide Mrs. Ranjana Jadhav for her valuable support and guidance.

References

[1] Kamble, A., Rao, R., Kale, A., and Samant, S. (2011), Selection of cricket players using analytical hierarchy process, International Journal of Sports Science and Engineering, 5(4), 207-212.

[2] Barr, G. D. I. and Kantor, B. S. (2004), A criterion for comparing and selecting batsmen in limited overs cricket, Journal of the Operational Research Society, 55(12), 1266-1274.

[3] Gerber, H. and Sharp, G. D. (2006), Selected a limited overs format of cricket squad. They choose using an integer programming model, South African Journal for Research in Sport, Physical Education and Recreation, 28(2), 81-90.

[4] Lourens, M. (2009), Integer optimization for the selection of a Twenty20 cricket team, Nelson Mandela Metropolitan University, Faculty of Science, South Africa: Unpublished Master of Science Dissertation in Mathematical Statistics.

[5] Bretteny, W. (2010), Integer optimization for the selection of a fantasy league cricket team, South Africa: M.Sc. Dissertation, Nelson Mandela Metropolitan University.

[6] Lemmer, H. H. (2002), The combined bowling rate was taken as a measure of bowling performance in cricket, The South African Journal for Research in Sport, Physical Education and Recreation, 24(2), 37-44.