Optimal IPL Playing 11 Team Selection

Nilesh Ravichandran
Dept. of Computer Science and Engineering
PES University
Bengaluru, India
niljams2002@gmail.com

Nihal Chengappa P.A
Dept. of Computer Science and Engineering
PES University
Bengaluru, India
nihalchengappa2625@gmail.com

Pranav Rao Rebala
Dept. of Computer Science and Engineering
PES University
Bengaluru, India
pranav.rebala11@gmail.com

***Abstract*—Since the inception of the Indian Premier League (IPL), the demand for quality data analysts has increased. The task to pick an ideal playing 11 from a huge squad is a real challenge and picking the best team can greatly improve a team's chances of winning. This project aims to predict the best possible playing 11 from a given set of players taking into consideration a number of factors like runs scored, runs conceded, strike rate, dot balls bowled etc. and this is essential as it can help pick an optimal eleven. Not only are the best batsmen and bowlers predicted but also the all-rounders which give a real balance to the team. Three models have been tested- XGBoost, Linear Regression and Random Forest. Linear Regression yields better results since it gives the least RMSE value.**

***Index Terms*—IPL, All-rounders, Runs scored, Runs conceded, RMSE, Linear Regression, XGBoost, Random Forest and Playing 11**

I. Introduction

The IPL first came into existence in 2008 and is presently in its 13th season. It is held every year across India and during this time, almost no international cricket takes place, enabling the players across the globe to feature in this league. The success of the IPL can be ascribed to a number of factors but a significant reason to it being the quality of cricket it has produced, consequently giving quality amusement for the group of onlookers/viewers. It is presently considered to be the leading T20 competition in the world. Not only international but also domestic cricketers feature in the IPL.

A lot of background analysis is required to be done by the franchises to pick a playing 11 from the squad. The IPL is played in a T20 format (20 overs per innings) with the game favouring the batsmen. Power play constitutes the first six overs and the final five comprise the death overs. This makes it quintessential for teams to focus on and improve the quality of their batting line up. The teams would generally prefer to have a mix of right and left handedness to unsettle the bowling team and versatility is a boon in this format.

Being a T20 game, the batsmen play with a high strike rate right from the start with the middle order providing the necessary impetus towards the end of the innings.

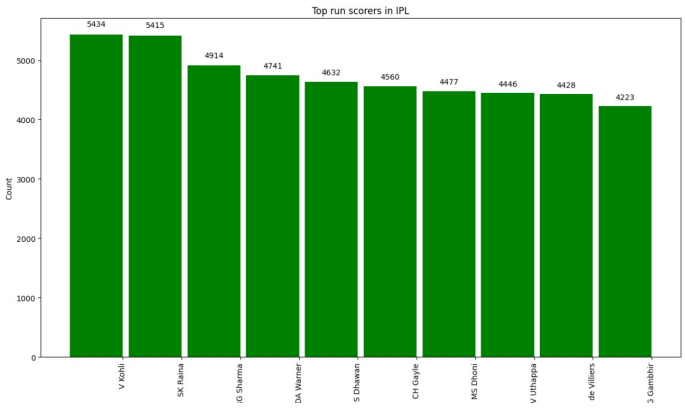

Fig. 1. Top run scorers in IPL

It can be inferred that the top run scorers are the ones who have a significant number of games under their belt indicating that they score those runs at a decent or quick rate (else the teams would have dropped them for slow run scoring).

Players who can bat and bowl- all-rounders, are vital. They provide the necessary balance to the team with the lower order expected to just make it towards the end of the innings as their main role is in the bowling department.

Among the bowlers, good spinners and pace bowlers are crucial- Swing/Pace bowlers with the new ball upfront and then spinners to control the middle overs followed by death over specialists are a key. The control of economy is the key since lesser the runs conceded, better is the result.

Three models have been tested to predict how much a batsman would score and how many runs a bowler would concede. The models are- Linear Regression, XGBoost and Random Forest. The dataset is taken from Kaggle- deliveries.csv which consists of the ball by ball data of all the matches from 2008 to 2019. All-rounders are analysed in a manner in which players with data who have batted and bowled are taken in the list (can be the whole dataset itself or from a selective list of players). The comparison of statistics of the batting and bowling will help in picking the right all-rounders.

II. RELATED WORK

R. Dutt and T. A. Kusupati et al. [1] have performed IPL Player Selection using Fuzzy Logic. Here, a system is developed for IPL auction that makes use of fuzzy logic and provides teams with a list of the five best players for each position based on their requirements which will help buy the desired players. The dataset named grip is taken. The result helps to identify a player as either a batsman, a bowler, or an all-rounder.

M. Ramalingam and S. Gokul et al. [2] have performed "Efficient Player Prediction and Suggestion using Machine Learning for IPL Tournament". The datasets have been used from kaggle. A module to analyze player performance in various matches is present which is assigned to the eight IPL teams. The player's performance was examined independently for bowling and batting while Random Forest was used to select the teams. The algorithm helps tackling regression and classification problems. The results suggests names of players under each team in the order of the implemented teams list. Here, each team has batsmen and bowlers according to their rating but does not take all rounders into consideration.

Jhanwar and Pudi et al. [10] provide insights on modelling batsmen and bowlers and eventually modelling groups (rating) and proposing a 'Playing XI', given a squad/pool of players employing a few features. It proposes strategies to demonstrate batsmen, bowlers and teams, utilizing different career insights and recent performances of the players. A parcel of assumptions has been made within the modelling step of batsmen, bowlers and the teams as a whole. Another critical concept is that all- rounders have not been considered independently. Although [10] has been executed for ODI cricket matches, IPL (T20) would be a more challenging task. Here, k-Nearest Neighbor (kNN) algorithm is utilized which gives a better accuracy of 71% as compared to other classifiers models that were used- 56% and 63%.

Gerber and Sharp et al. [11] proposed an integer programming strategy for selecting a limited over squad or 15 players instead of playing 11 players. The method includes- gathering information from 32 South African players to choose the ODI squad of 15 players and also describes the ability-indexing used and discusses the results of an empirical study conducted using the given statistics.

Barr and Kantor et al. [12] utilize new graphical representation with Strike rate and Probability of getting out, similar to the risk–return framework used in portfolio analysis to obtain insights on the batting performance, particularly in the context of the one-day game. A two-dimensional framework is developed for selection criterion for batsmen, which combines the average and the strike rate.

III. PROPOSED SOLUTION

In order to select the best playing 11 for an IPL team, the proposed solution involves- selecting batsman and bowlers (depends on how many batsmen and bowlers a team wishes to have) as well as the all rounders(after combining the batting and bowling statistics) where the data accumulated over the performances over the past decade (specifically 2008- 2019) is used.

The modelling for the best batsmen and the best bowlers is performed, but preprocessing is performed before the model building-

A. Pre-processing:

The file deliveries.csv has at least 431000 null values in the attribute's fields: player_dismissed, dismissal_kind and fielder. All the null values are replaced with NaN to avoid errors and make the dataset consistent overall. The columns bye_runs, legbye_runs, is_super_over are not useful for the predictions and hence are dropped.

In some of the records it is observed that the number of innings was reported to be 5. This is faulty in a T20 as there are only 2 innings and is an error in data entry. However, since the number of innings don't contribute to the predictions, that column is dropped as well. Exploratory Data Analysis and Visualizations are performed to select the important features and this would aid in the creation of better models

B. FOR BATSMEN: Model Building

The first model needed to be built is for predicting the best batsmen. The model predicts the runs scored by a batsman by taking the following features-

1. Balls faced
2. No of innings played
3. Total runs scored
4. Total 4s
5. Total 6s
6. Average score per game
7. Strike rate
8. Total 100s
9. Total 50s

This is done by creating a data frame with the summary statistics of each batsman having the columns- balls faced, no of innings played, total runs scored,4s,6s, average runs, strike rate, total 100s, total 50s and match_runs which is the average runs scored by the batsman in the last 5 matches to get the recent form.

While computing the total innings played by a batsman, values greater than the mean innings played by all the batsmen is considered since some batsmen who come down the order may play fewer matches. Also, people who have played very few games and have performed well will lead to a biased

calculation as it is harder to score over many number of games and consistency in run scoring is essential.

Algorithm:

A sample of the final batsman data frame who have played more than the mean number of games as shown below-

batsman	balls_faced	innings	runs	4s	6s	AVG	SR	match_runs	100s	50s
DPMD Jayawardene	1522	78	1808	200.0	40.0	23.18	118.79	12.4	1.0	11.0
CH Gayle	3131	124	4560	376.0	327.0	36.77	145.64	13.0	7.0	35.0
CH Morris	339	37	520	37.0	27.0	14.05	153.39	12.8	0.0	2.0
DT Christian	384	34	448	22.0	19.0	13.18	116.67	5.0	0.0	0.0
MK Tiwary	1489	83	1697	156.0	40.0	20.45	113.97	11.8	0.0	6.0

Fig. 2. Sample of statistics of 5 players having played greater than mean number of innings

Fig. 2. match_runs is the target variable; splitting the remaining features into train and test datasets in the ratio of 0.8:0.2.

Models for batsmen:

1. Linear Regression:

Features: balls faced, no of innings, total runs,4s,6s,average runs, strike rate, 100s, 50s

Target: Match_runs(average runs scored in last 5 matches to get the recent form as it is better to have a player who recently has scored runs than one who has scored years back).

This model is then tested against the test data and its accuracy is calculated using RMSE. This gives an RMSE score of 3.65 which is fairly good. Hence, proving that this model gives very low error and high accuracy.

2. XGBoost:

Features: balls faced, no of innings, total runs,4s,6s,average runs, strike rate, 100s, 50s

Target: Match_runs(average runs scored in last 5 matches)

Similarly this is then tested against the test data and its error is calculated using RMSE which yields a score of 5.39 which is good, however, linear regression gives better results.

3. Random Forest:

Features: balls faced, no of innings, total runs ,4s,6s,average runs, strike rate, 100s, 50s

Target: Match_runs(average runs scored in last 5 matches) This model yields a RMSE score of 6.91 which is higher than both linear regression and XGBoost and hence gives the highest error.

Hence, Linear Regression model is used for the predictions.

Predictions:

The Algorithm used to predict the Batsman's scores in the following game is predicted as shown in the figure Fig. 3

To predict the top N batsmen, a prediction function is created using the algorithm shown in in Fig. 3. In the function, the argument takes a list of N different batsmen who are to be compared (entire dataset or any list can be mentioned, has no restrictions). It first runs a for loop which predicts the runs scored for the N batsmen that are passed as arguments to it. It is then followed by graphical comparisons after which the batsmen to be chosen are decided.

Algorithm 1: Algorithm to predict batsmans' scores

Result: Predicted Batsmans' Score

batsmanscore ← []

for *players in batsmanlist* **do**

  pred ← lr.predict(x.loc[(players)].values.tolist())

  batsmanscore[players] = pred

end

top ← $sorted(batsmanscore, key \leftarrow batsmanscore.get, reverse \leftarrow True)[:]$

Fig. 3. Algorithm to predict batsman's runs

C. FOR BOWLERS:

Model Building:

The second model required is the bowlers model. This model predicts the economy of a given bowler accurately- lower the economy- better the result. The features of this model include-

1. Balls bowled
2. Wickets
3. Overs
4. Runs conceded
5. Bowlers economy

Bowlers data frame:

bowler	balls_bowled	wickets	overs	runs_conceded	bowl_econ
MA Starc	612	34.0	97	725	7.47
R Ashwin	3016	125.0	489	3391	6.93
KH Pandya	968	39.0	159	1159	7.29
SR Watson	2137	92.0	341	2751	8.07
R Dhawan	513	18.0	82	655	7.99

Fig. 4. Sample of statistics of 5 players having bowled overs greater than mean number of overs

The total runs conceded by the bowler in the last 5 matches is calculated. Using this, the bowler's recent economy is calculated which serves as the target variable for the model.

Now, the recent economy variable is taken as the target and while the remaining features are split into train and test datasets in the ratio of 0.8:0.2.

Models for bowlers:

1. Linear Regression :

Features: balls bowled, wickets, overs , runs conceded, bowler economy

Target: Recent economy (average economy of bowler in last 5 matches)

This model is then tested against the test data and its accuracy is calculated using RMSE value.

The RMSE value is found to be 1.76 which proves that the model is giving low error.

2. Random Forest :

Features: balls bowled, wickets, overs , runs conceded, bowler economy Target: Recent economy (average economy of bowler in last 5 matches)

Similarly, this too is tested against the test data and its RMSE yields a score of 1.9 which is good. However, linear regression gives better results.

3. XGBoost :

Features: balls bowled, wickets, overs , runs conceded, bowler economy Target: Recent economy (average economy of bowler in last 5 matches)

RMSE in XGBoost yields a score of 4.54 which is higher than both linear regression and Random Forest and hence gives the highest error.

Thus, Linear Regression model will be used for the bowlers as well.

Predictions:

The Algorithm used to predict the Bowlers' scores in the following game is predicted as shown in the figure Fig. 5

Algorithm 1: Algorithm to predict bowler's economy

Result: Predicted Bowler's Economy

bowlerseconomy ← []

for *players in bowler list* **do**

  pred ← lr.predict(x.loc[(player)].values.tolist())

  bowlerseconomy[player] = pred

end

top ← $sorted(bowlerseconomy, key \leftarrow bowlereconomy.get, reverse \leftarrow True)[:]$

Fig. 5. Algorithm to predict bowler's economy

To predict the top N bowlers, a prediction function is created using algorithm as shown in Fig. 5. In this function, the argument takes a list of N different bowlers to compare (list can be the entire dataset or a selective list). It first runs a for loop which predicts the economy for the N bowlers that are passed as arguments to it. It is then followed by graphical comparisons from which, bowlers to be picked for the team are chosen.

D. FOR ALL-ROUNDERS:

The graphical analysis is done on both- batsmen and bowlers and the inferences are combined to get the all-rounders. Example: If 2 all rounders are required then a rough estimation of the top 2 players who feature in both combined, are considered.

EXPERIMENTAL RESULTS

Through the data analysis and models, prediction of the top batsmen and bowlers for the team is done. Now a team is built and the composition is up to the team's discretion. The all-rounders are the ones who have a good record in both - batting and bowling. For example: Building a team of 5 batsmen,4 bowlers and 2 all-rounders. For this, the players in the squad (or auction pool) are to be known (or can be the entire dataset too) from which they have to be selected.

Squad/Auction pool: SK Raina, V Kohli, RV Uthappa, Yuvraj Singh, PA Patel, R Dravid, G Gambhir, KD Karthik, RG Sharma, V Sehwag, SR Watson, P Kumar, R Ashwin, IK Pathan, A Mishra, AB Agarkar, AB Dinda, AD Mathews, AD Russell and B Kumar

Now, all the players who can bat are put in the predict_batsman function: V Kohli, RV Uthappa, Yuvraj Singh, PA Patel, R Dravid, SK Raina, G Gambhir, KD Karthik, RG Sharma, V Sehwag, SR Watson, IK Pathan, AD Mathews and B Kumar

And, all the players who can bowl are put in the predict_bowler function: SK Raina, Yuvraj Singh, SR Watson, P Kumar, R Ashwin, IK Pathan, A Mishra, AB Agarkar, AB Dinda, AD Mathews, AD Russell and B Kumar

The predictions yield the following graphs:

Batsmen:

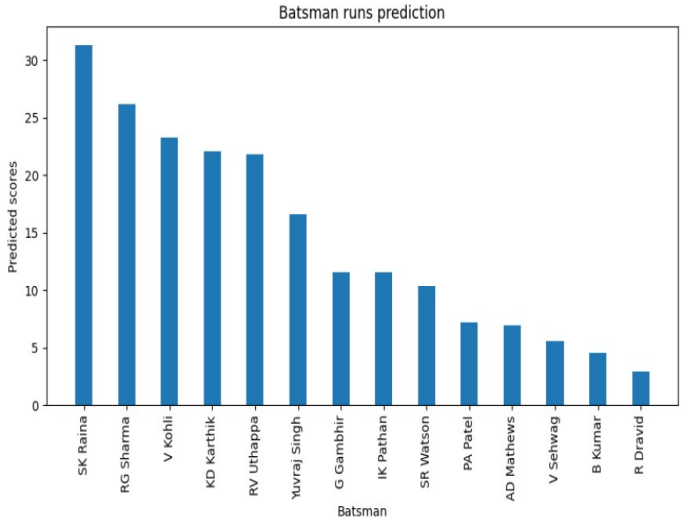

Fig. 6. Batsman's runs prediction

Bowlers:

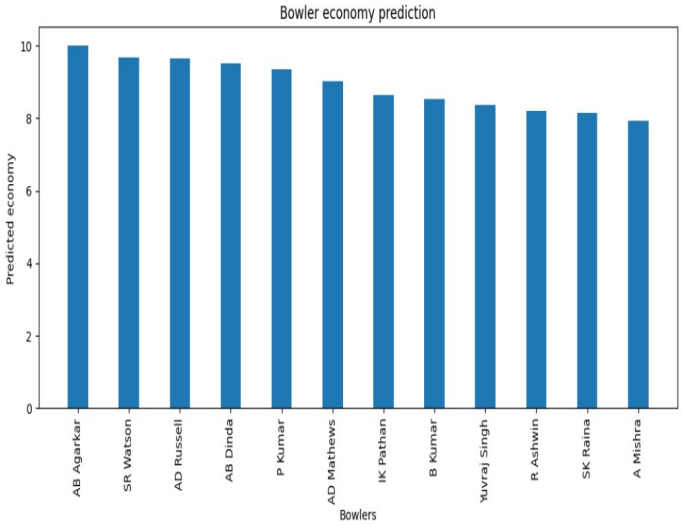

Fig. 7. Bowler economy prediction

On analyzing the graphs (Fig. 5 and Fig. 6) the following conclusions can be made:

a. The top 2 all-rounders : SK Raina, Yuvraj Singh

b. The top 5 batsmen are: RG Sharma, V Kohli, KD Karthik, RV Uthappa, G Gambhir

c. The top 4 bowlers are: A Mishra, R Ashwin, B Kumar, IK Pathan

The all-rounders: SK Raina and Yuvraj Singh are chosen because they have the best combined record of batting and

bowling and the team requires only 2 all-rounders so only the top two all-rounders are chosen.

Playing XI:

RG Sharma, G Gambhir, V Kohli, RV Uthappa, KD Karthik, SK Raina, Yuvraj Singh, A Mishra, R Ashwin, B Kumar, IK Pathan

This is not a fixed format and as per the requirements, the team is free to chose any number of batsmen, bowlers and all-rounders with a total of 11 players. Thus the ideal optimal 11 players for the team are chosen. Not only are the batsmen and bowlers being accounted for but also, the required number of all- rounders too - through graphical analysis.

CONCLUSIONS

After comparing the RMSE (Root Mean Squared Values) values from the three implemented models, Linear Regression yields the best result. The model is present in the range of analyzation of data and prediction of performance of players with the conclusion of the model being- it can serve as a recommendation on a real-world level and can propose the best possible playing eleven for the team, it predicts how much a batsman would score in the next match taking into consideration a number of factors like runs, strike rate, average, balls faced, boundaries, 50s, 100s and innings and for the bowlers on the other hand, it predicts their expected economy in the next match which is pivotal through factors such as balls bowled, wickets taken, runs conceded and economy from the training data because lesser the runs conceded, better is the outcome. Similarly, it is done for all-rounders too where the data of players who have batted and bowled is fed and the best set of the players are chosen with a good combined record and hence strengthen the balance of the team giving more depth in both departments.

Major Advantage:

A fixed number of batsmen, bowlers or all-rounders for a particular match need not be selected, i.e., the team can decide what combination to play, i.e., how many batsmen, all-rounders and bowlers to play depending on the conditions- the team can play 3 bowlers and 5 batsmen with 3 all-rounders if it's a bowling pitch or 4 batsmen and 5 bowlers with 2 all-rounders if it's a batting pitch etc. Another advantage is that the model can take a specific list of players which the team wants, it doesn't have to take all the players from the dataset, so if 2 batsmen from a list of 10 are to be selected then 10 players' statistics can be fed and predicted accordingly and similarly for the bowlers. Hence, prediction of the optimal playing eleven can be expected to help the team.

FUTURE WORK

The model predicts bowlers, batsmen and all-rounders but the future work is to take this to the next stage by not just picking batsmen as batsmen but as wicketkeepers, openers etc. and similarly for the bowlers as spinners and pacers. Another addition to the future work is to consider the historical record of the playing venue to determine what decision to make on winning the toss.

APPENDIX

This mainly consists of the Exploratory Data Analysis and data visualization. It is the base for any model and helps to analyze and comprehend the different features in the dataset.

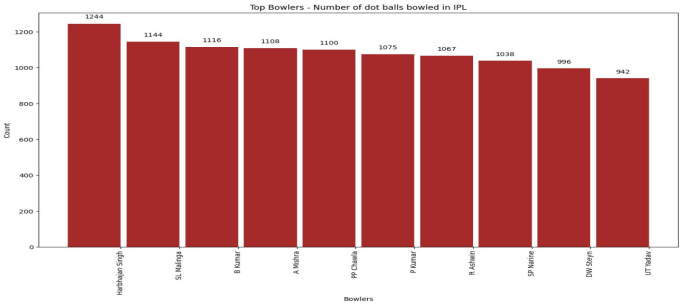

Fig. 8. Bowlers who have bowled the most dot balls

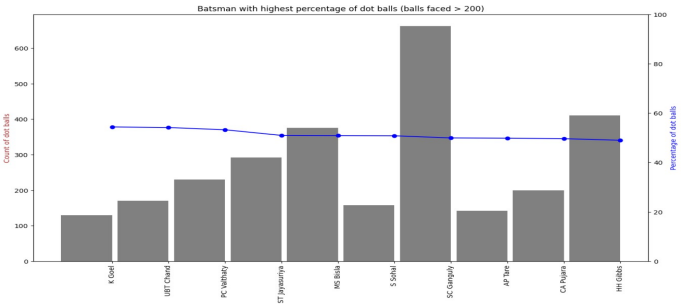

Fig. 9. Batsman with highest percentage of dot balls

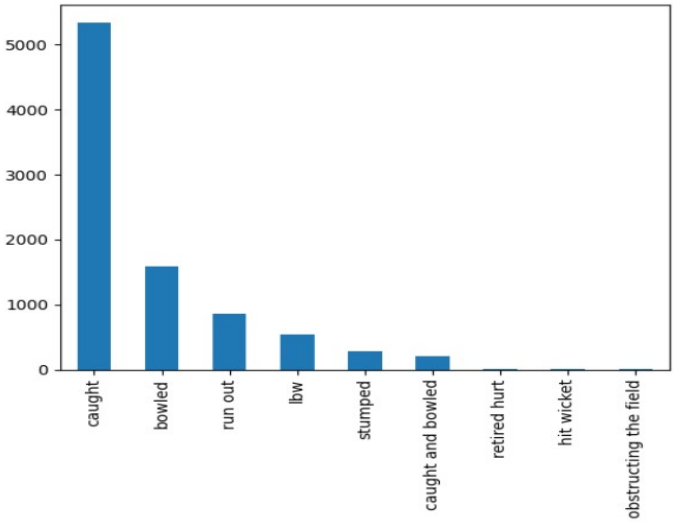

Fig. 10. Different modes of dismissals

REFERENCES

[1] R. Dutt, T. A. Kusupati, A. Srivastava and D. Hore, "IPL Player Selection using Fuzzy Logic," 2022 IEEE Industrial Electronics and Applications Conference (IEACon), 2022, pp. 180-184, doi: 10.1109/IEACon55029.2022.9951755.

[2] M. Ramalingam, S. Gokul, L. S. Mythravarshini and K. S. Harine, "Efficient Player Prediction and Suggestion using Machine Learning for IPL Tournament," 2022 International Mobile and Embedded Technology Conference (MECON), 2022, pp. 162-167, doi: 10.1109/MECON53876.2022.9752414.

[3] D. Hasanika, R. Dilhara, D. Liyanage, A. Bandaranayake and S. Deegalla, "Data Mining System for Predicting a Winning Cricket Team," 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS), 2021, pp. 92-97, doi: 10.1109/ICIIS53135.2021.9660702.

[4] M. K. Mahbub, M. A. M. Miah, S. M. S. Islam, S. Sorna, S. Hossain and M. Biswas, "Best Eleven Forecast for Bangladesh Cricket Team with Machine Learning Techniques," 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2021, pp. 1-6, doi: 10.1109/ICEEICT53905.2021.9667862.

[5] Mittal, H., Rikhari, D., Kumar, J., Singh, A. K. (2021). A study on machine learning approaches for player performance and match results prediction. arXiv preprint arXiv:2108.10125.

[6] Patil, N. M., Sequeira, B. H., Gonsalves, N. N., Singh, A. A. (2020). Cricket Team Prediction Using Machine Learning Techniques. Available at SSRN 3572740.

[7] M. M. Rahman, M. O. Faruque Shamim and S. Ismail, "An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach," 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2018, pp. 190-194, doi: 10.1109/ICISET.2018.8745588.

[8] M. J. Hossain, M. A. Kashem, M. S. Islam and M. E-Jannat, "Bangladesh Cricket Squad Prediction Using Statistical Data and Genetic Algorithm," 2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT), 2018, pp. 178-181, doi: 10.1109/CEEICT.2018.8628076.

[9] Swarna, S. T., Ehsan, S., Islam, M. (2017). A statistical model for ideal team selection for a national cricket squad. arXiv preprint arXiv:1702.02089.

[10] Jhanwar, M. G., Pudi, V. (2016, September). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In MLSA@ PKDD/ECML.

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

[12] Barr, G. D. I., 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.