Selection of optimal cricket team based on the players performance

Prof. Monali Shetty
Computer Engineering
Fr. CRCE, Mumbai
shettymonalin@gmail.com

Sankalp Rane
Computer Engineering
Fr. CRCE, Mumbai
sankalprane1998@gmail.com

Chaitanya Pandita
Computer Engineering
Fr. CRCE, Mumbai
chaitanyapandita97@gmail.com

Suyash Salvi
Computer Engineering
Fr. CRCE, Mumbai
suyash.salvi1998@gmail.com

Abstract: This paper is about a model that can select best playing 11 in the Indian cricket team. The performance of each player depends on several factors like the pitch type, the opposition team, the ground, and several others. The proposed model contains data from the One Day International of the past several years of team India. The dataset used for this model created using data from trusted sites like espn.com. This method is distinct in the sense that it gives you a 360-degree view of the player's skill set, be it, batting, bowling, and fielding. The vital part of this model is to find the best all-rounder player. Random forest algorithm used for predicting performance. The player performance classified into several classes, and a random forest classifier used to predict the player's performance. This model gives 76% accuracy for batsmen, around 67% accuracy for bowlers, and 95% for an all-rounder. We created a model with some extra features like weather, matches played that have not considered in any existing model. Using this model, the best team can be selected to play in given conditions.

Keywords: All-rounder, random forest, player selection, team selection, cricket, machine learning, SVM, logistic regression

1. INTRODUCTION

Cricket is the second most-watched program on television. The popularity of this sport is soaring high in southeast countries like India, Pakistan, Bangladesh, and Sri-Lanka. One of the major issues now is playing 11, that should be selected. People of different ages and backgrounds are zealot fans of cricket because the Indian team is doing well in the recent past. However, many confusions arise before a match about team combinations, i.e., which player to select or drop for the next game, which batsman should play in which position or which bowler should be picked for the upcoming game. Machine learning is used to predict the match results in many different sports [2]. From here, the motivation emerged to evaluate the performance of the player in a specific match and selecting the best playing 11 according to it using machine learning [3], as it will remove the proclivity towards one particular player and benefit our cricket team. The thing that makes this model unique is that it takes into consideration if a player is an all-rounder. All-rounders, when compared based only on one of their attributes, may not get a place in the team.

2. LITERATURE REVIEW

To know in-depth this topic, previous research is done in this field, and the studies dedicated to this are discussed in detail here.

[1] Aminul Islam Anik et al. proposed using the balls faced, ground, pitch, opposition, and position to find the performance of a player using SVM. This model does a proper analysis of various factors and their effect on the runs scored by batsmen. The primary factor is balls faced by the players, but a drawback is that the number of shots faced by a player cannot known before the match. The paper has done the right amount of work on batsmen and bowlers, but the analysis for all-rounder is left.

[2] Amal Kaluarachchi used Bayesian classifiers in Machine learning, to predict how the factors like home game advantage affect the outcome of the match. Using this idea, we used the home and away as one of the parameters affecting the players performance.

[3]Pranavan Somaskandhan et al.; analyzed the set of attributes that impose high Impact on the outcome of a game using machine learning. When the attribute combination of high individual wickets, number of bowled deliveries, number of the thirties, total wickets, wickets in the power play, runs in death overs, dots in middle overs, number of fours and singles in middle overs highest accuracy obtained. The attribute, as mentioned above, gave an accuracy of 81% using SVM.

[4]Md. Muhaimenur Rahman analyzed Bangladesh One Day International Cricket data. They divided the study into three sections, i.e., at the start of the game, after one inning and after the fall of wickets. They used Decision Tree and got an accuracy of 63.63% at the beginning of the game, 72.72%, 81.81% in first and second innings, and 80% and 70% for fall of wicket analysis.

[5]Riju Chaudhari et al. used a DEA(Data Envelopment Analysis) for measuring the efficiency of players. The paper takes records of players performance in test matches. Since there can be factors such as match-fixing in T20. Since in test series, every player might have to do batting a bowler with a higher batting strike rate is given preference. This paper is very different from other paper since it does not take direct machine learning techniques and instead uses a unique approach.

[6]Md. Jakir Hossain used a genetic algorithm on 30 players in the Bangladesh cricket team to select top players. The paper combines statistical analysis with a genetic algorithm to choose top tier players. Every possible solution out of ${}^{30}C_{14}$ total solutions taken as a chromosome. The ratings of players considered using a statistical method. The final fitness value calculated using factors like the sum of a rating of players, number of bowlers, batsmen, all-rounders, and wicket keepers, number of spins and fast bowlers, and number of right and left-handed players.

[7]Vipul Punjabi and team used naive bayes classifier to predict the runs scored by batsmen and wickets taken by the bowler. The runs scored and wickets taken classified into different categories. The dataset used for this is taken from records of IPL matches. This paper takes remarkably few input features, thus reducing a lot of potential of the model.

3. PROPOSED METHODOLOGY

Data of the past ODI matches used to create the dataset, which is mentioned in detail in the next section of this paper. The performance of batsmen, according to the runs scored is divided into various classes, the same done for the number of wickets taken by the bowler.

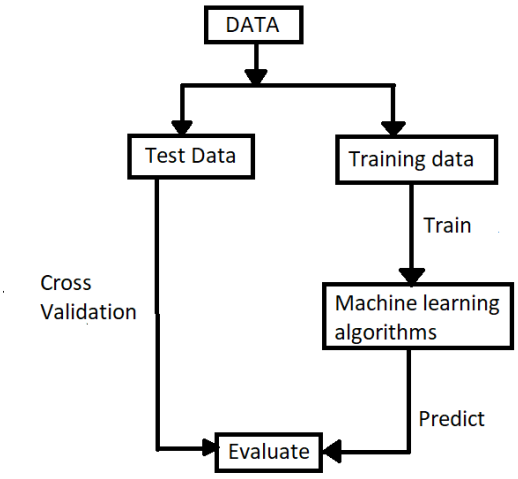

Fig.3.1: Proposed Model

The dataset which was created was split into 2 parts, 80% of the dataset was used for training the model and 20% of the dataset was used to evaluate the results (refer fig 3.1). Various algorithms like logistic regression, SVM and random forest were used to get the results which is discussed in details in the next section of this paper.

4. ALGORITHMS AND TECHNIQUES

LOGISTIC REGRESSION: Logistic regressor usually used for binary classification tasks. In the case of multi-class classification, softmax function used in place of the sigmoid function. [8] The hypothesis function for logistic regression is given as $g(z)=1/(1+e^{-z})$.

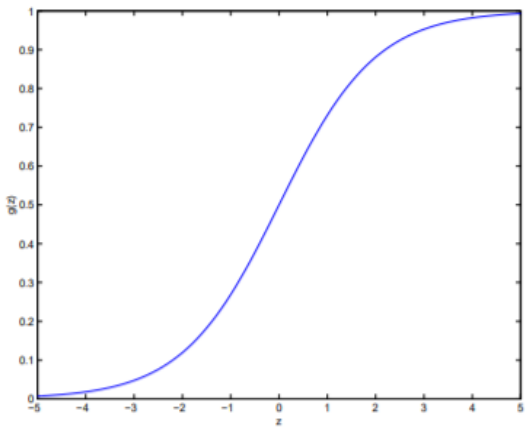

Fig.4.1 Graph of g(z) vs z

The plot (refer Fig.4.1) for g(z) tends towards one as z tends to infinity. And it tends towards 0 as z tends to negative infinity.

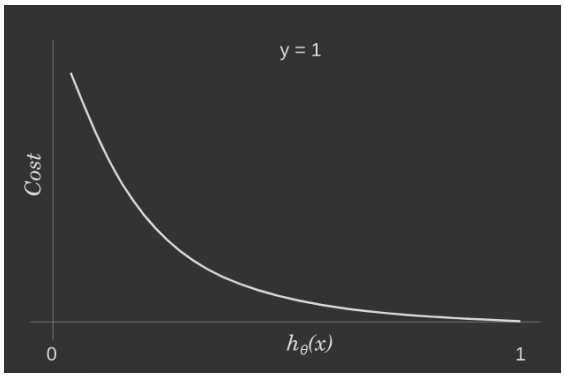

Fig.4.2. Cost function vs h(x) when y=0

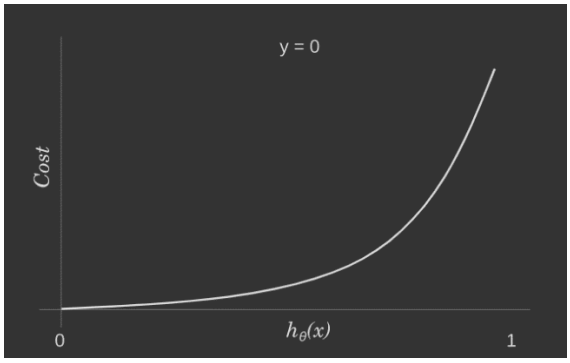

Fig.4.3 Cost function vs h(x) when y=0

The cost function can be written in one line as [8]:

$cost(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$

SUPPORT VECTOR: In support vector classifier, we hyperplane to distinguish different classes. There are different kernels to separate non-linear data by mapping them to higher dimensions [9]. Many hyperplanes might classify the data successfully. One reasonable choice as the best hyperplane is the one representing the most significant separation or margin between the two classes. So, we choose the hyperplane such that the distance from it to the nearest point maximized.

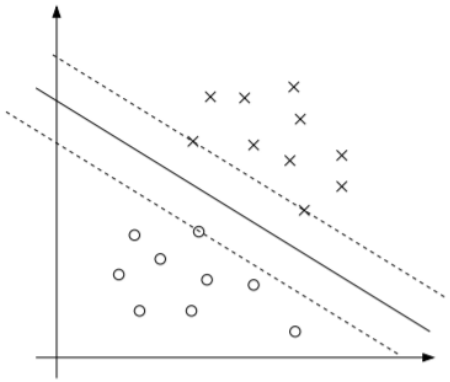

Fig.4.4 SVM Hyperplane

The solid line (refer Fig 4.4) represents one of the possible hyperplanes.

RANDOM FOREST: Random Forest is an ensemble learning method for classification, regression, and other tasks by taking a combination of results of several decision trees.

Random forests used to rank the importance of variables in a regression or classification problem in a natural way.

Most of the batsmen score a meager amount of runs which is given by the graph of the dataset provided below (Fig 4.5):

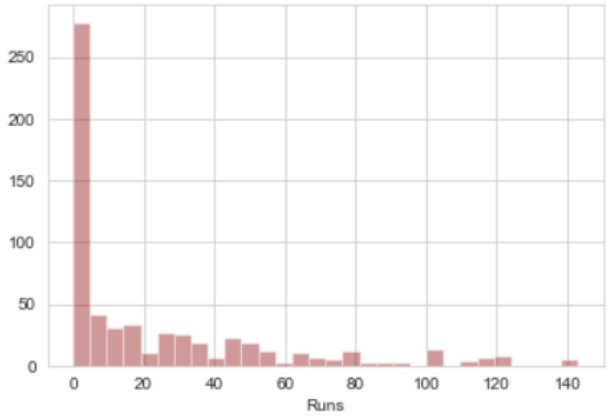

Fig.4.5. Distplot of Runs

There is a strong relationship between strike rate and performance of players. The higher the result, the better is the performance (refer Fig 4.6).

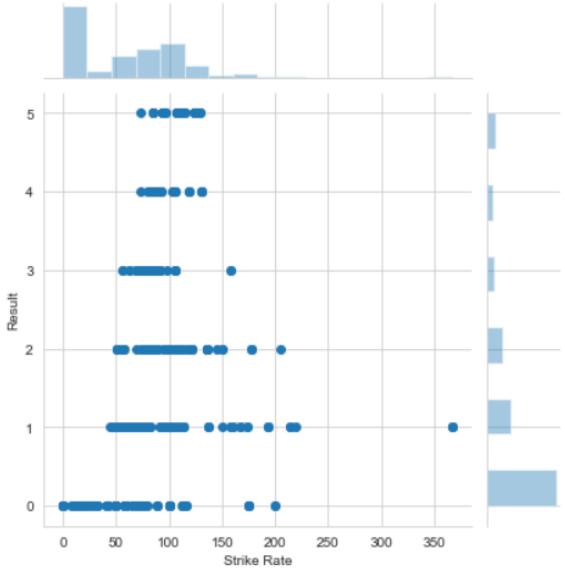

Fig.4.6 Joint plot of Result vs Strike Rate

5. DATA DESCRIPTION

Data Collection: The dataset is made from sites like espncricinfo.com, one of the legit sites. A CSV file created using the data from previous matches played by the Indian cricket team. And for other conditions, the summary was used.

Sr.No.	Position	Name	Matches Played	Runs	Fours	Sixes	Strike Rate	Ground	Home Away
1	1	RG Sharma	218	18	2	0	52.9	Trinidad	Away
2	2	S Dhawan	133	2	0	0	66.7	Trinidad	Away
3	3	V Kohli	239	120	14	1	96	Trinidad	Away
4	4	RR Pant	12	20	2	0	57.1	Trinidad	Away
5	5	SS Iyer	9	71	5	1	104.4	Trinidad	Away
6	6	KM Jadhav	68	16	2	0	114.3	Trinidad	Away
7	7	RA Jadeja	156	16	1	0	100	Trinidad	Away
8	8	B Kumar	114	1	0	0	50	Trinidad	Away
9	9	M Shami	70	3	0	0	60	Trinidad	Away
10	10	KK Ahmed	11	0	0	0	0	Trinidad	Away
11	11	YS Chahal	50	0	0	0	0	Trinidad	Away
12	1	RG Sharma	218	10	2	0	166.7	Trinidad	Away
13	2	S Dhawan	133	36	5	0	100	Trinidad	Away
14	3	V Kohli	239	115	14	0	115.2	Trinidad	Away

Fig. 5.1 A part of the Batsman dataset

The datasets: Figure 5.1 is a small sample of the batsmen dataset. To process this data, one hot encoding used. The player names and his stats are hardcoded in the backend.

After the stats entered in the API, all players' performance is recorded in a dictionary using for loop. The dictionary then sorted in the reverse order to give the names of the top players.

There are eight opposition teams and 15 grounds taken into consideration. Below is a part of the dataset after pre-processing with pandas library.

	Position	Matches Played	50s	100s	Overall Average	Bay Oval	Chandigarh	Delhi	Edgbaston	Headingley	..
0	1	218	42	27	48.53	0	0	0	0	0	..
1	2	133	27	17	44.50	0	0	0	0	0	..
2	3	239	54	43	60.31	0	0	0	0	0	..
3	4	12	0	0	22.90	0	0	0	0	0	..
4	5	9	4	0	49.43	0	0	0	0	0	..
5	6	68	6	2	42.97	0	0	0	0	0	..
6	7	156	11	0	30.84	0	0	0	0	0	..
7	8	114	1	0	14.22	0	0	0	0	0	..
8	9	70	0	0	7.56	0	0	0	0	0	..
9	10	11	0	0	4.50	0	0	0	0	0	..

Fig.5.2 Part of Batsmen dataset

Feature Selection: The previously created models for selection of optimal team contained features like Opponents, Runs Scored, Strike Rate, and Overall Average. We have considered following attributes to measure player's performance.

Batting Attributes: Position, Matches Played, Runs, Strike Rate, Ground, Home Away,50s, 100s, Overall Average, Pitch, Opponent and Weather.

Bowling Attributes: Matches Played, Wickets, Average, Economy, Strike Rate, Ground, Pitch, Opponent, Weather, Home Away.

All-rounder Attributes: Matches Played, Wickets, Runs Scored, Strike Rate, Average, Ground, Home Away, Opponent, Weather and pitch.

Since these models considered very few features, the accuracy was not good enough. Hence, we created a model with some extra features like matches played, pitch, weather.

These models only considered batsmen and bowler's to create a team of 11 players. But if bowlers and batsmen compared with all-rounders, all-rounders will always get fewer ratings. Thus we considered all-rounders while creating our model to generate an official team.

6. IMPLEMENTATION

For implementation, we use Flask API. We use a jupyter notebook to train the model and use the available algorithms in the scikit lean library in python. Below is the screenshot of our implementation and a sample result. The input parameters, which are the same for batsmen, bowlers, and all-rounders, are taken into consideration.

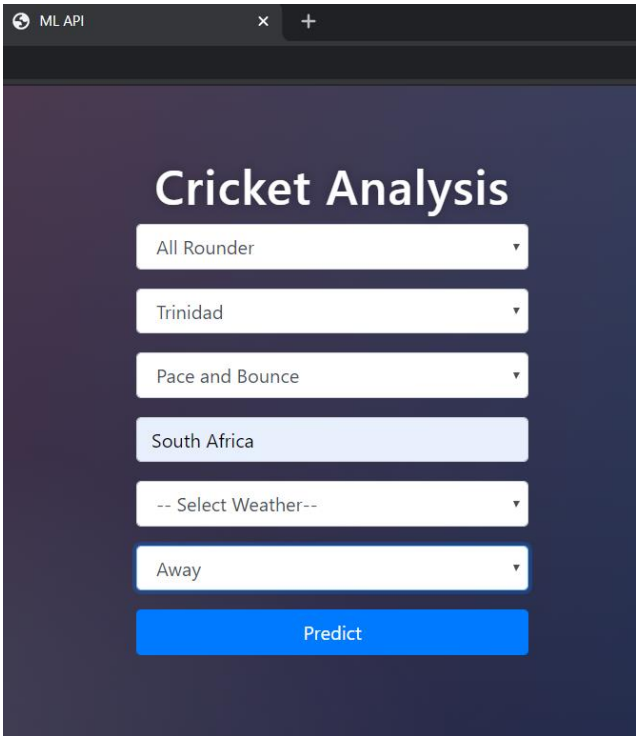

Fig.6.1: Input display

The sample output on clicking the predict button is given below in Fig.6.2.

Using the scikit learn library in python, various algorithms used to predict the results. Different techniques like logistic regression, SVM classifier, decision tree, and random forest used to predict the classes, out of which Random forest gave the best results.

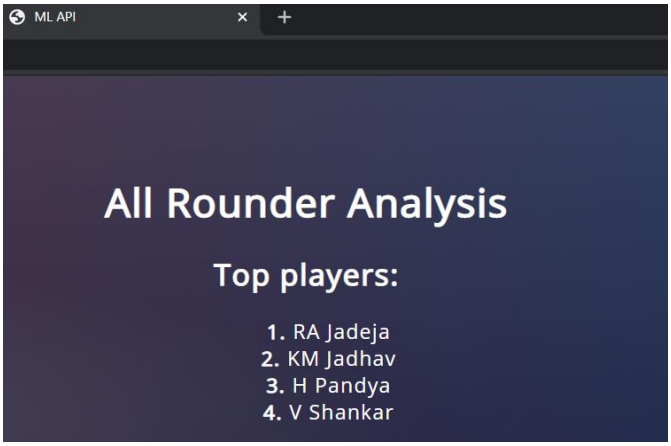

Fig.6.2: Output display for all-rounders

When we used the random forest classifier with a test size of 20%, we get the following results for the batsmen dataset.

	precision	recall	f1-score	support
0	0.83	0.90	0.86	63
1	0.59	0.57	0.58	23
2	0.62	0.67	0.64	12
3	1.00	0.36	0.53	11
4	0.60	0.60	0.60	5
5	0.75	0.86	0.80	7
micro avg	0.75	0.75	0.75	121
macro avg	0.73	0.66	0.67	121
weighted avg	0.76	0.75	0.74	121

Fig.6.3 Random Forest Classification Report

The support vector classifier gives the below classification report (refer fig 6.4). Class 0 had 55 values predicted correct out of total 63 values. But the class 4 and class 5 gave bad results. Only 1 out of 5 predictions were correct for class 4 and 3 out of 7 predictions were correct for class 5.

	precision	recall	f1-score	support
0	0.76	0.87	0.81	63
1	0.54	0.61	0.57	23
2	0.67	0.67	0.67	12
3	0.50	0.09	0.15	11
4	0.25	0.20	0.22	5
5	0.60	0.43	0.50	7
micro avg	0.68	0.68	0.68	121
macro avg	0.55	0.48	0.49	121
weighted avg	0.66	0.68	0.65	121

Fig.6.4 Support Vector Classification Report

While the logistic regression algorithm gave the worst results (refer fig 6.5). Class 4 had no correct predictions while only 2 out of 7 predictions were correct for class 5.

	precision	recall	f1-score	support
0	0.72	0.84	0.77	63
1	0.53	0.39	0.45	23
2	0.23	0.25	0.24	12
3	0.40	0.18	0.25	11
4	0.00	0.00	0.00	5
5	0.29	0.29	0.29	7
micro avg	0.57	0.57	0.57	121
macro avg	0.36	0.33	0.33	121
weighted avg	0.55	0.57	0.55	121

Fig.6.5 Logistic Regression Classification Report

So we went ahead with the random forest algorithm.

7. CONCLUSION AND FUTURE SCOPE

In this article, we were able to address the issue of selecting the optimal team in cricket without any prejudice and give equal importance to all-rounders. We were able to successfully implement a web application using a flask to run our project. This model provides 76% accuracy for batsmen (refer Fig 7.1) around 67% accuracy for bowlers (Fig 7.3) and 95% for all-rounder (Fig 7.2). The results are verified for 20% of the dataset and we got the above results.

	precision	recall	f1-score	support
0	0.83	0.90	0.86	63
1	0.59	0.57	0.58	23
2	0.62	0.67	0.64	12
3	1.00	0.36	0.53	11
4	0.60	0.60	0.60	5
5	0.75	0.86	0.80	7
micro avg	0.75	0.75	0.75	121
macro avg	0.73	0.66	0.67	121
weighted avg	0.76	0.75	0.74	121

Fig.7.1: Classification report for batsmen

For batsmen the dataset (refer fig 7.1) was verified for 121 values. 55 out of 63 total values were correctly identified for class 0, 15 out of 23 values were correct for class 1. 8,5,3 and 6 values gave correct predictions for class 2, class 3, class 4 and class 5 respectively.

	precision	recall	f1-score	support
0	0.98	0.96	0.97	46
1	0.93	0.96	0.95	28
micro avg	0.96	0.96	0.96	74
macro avg	0.95	0.96	0.96	74
weighted avg	0.96	0.96	0.96	74

Fig 7.2: Classification report for all-rounders

For all rounder (refer fig 7.2) 44 out of 46 values were correct for class 0. Only 1 out of 28 value was not identified correctly. Both the precision and recall was good for all rounders.

	precision	recall	f1-score	support
0	0.71	0.77	0.74	196
1	0.63	0.61	0.62	79
2	0.55	0.50	0.52	34
3	0.80	0.40	0.53	10
4	0.65	0.54	0.59	28
micro avg	0.67	0.67	0.67	347
macro avg	0.67	0.56	0.60	347
weighted avg	0.67	0.67	0.67	347

Fig.7.3: Classification report for bowlers

The bowler dataset (refer fig.7.3) was the largest in which 347 values were tested. 150 from 196 values were predicted correctly for class 0. 48 values were predicted correct in class 1 while 17 values were predicted correct for class 2. Class 3 and class 4 were verified for 10 and 28 values out of which 4 and 15 were correctly identified.

This analysis could be done between the game where the number of dots, remaining overs, number of wickets left, and strike rate are known, which could help the players decide the position they should play, giving even better results. As these factors determine the game's outcome within split seconds, a lot of work could be done in these dynamic factors leading to a beneficial model.

7. REFERENCES

[1] Aminul Anik “Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms” BRAC University, Dhaka, Bangladesh, 4th International Conference 2018 on Electrical Engineering and Information and Communication Technology.

[2] Amal Kaluarachchi, Aparna S. Varde, "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket " thesis, Montclair State University, Montclair, NJ, USA, 2010 Fifth International Conference on Information and Automation for Sustainability.

[3] Pranavan Somaskandhan, Gihan Wijesinghe, Leshan Bashitha Wijegunawardana, Asitha Bandaranayake, and Sampath Deegalla, "Identifying the Optimal Set of Attributes that Impose High Impact on the End Results of a Cricket Match Using Machine Learning," 2017 IEEE International Conference on Industrial and Information Systems (ICIIS)

[4] Md. Muhaimenur Rahman, Md. Omar Faruque Shamim, Sabir Ismail, "An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach," Computer Science & Engineering Sylhet Engineering College Sylhet, Bangladesh, 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)

[5] Riju Chaudhari, Sahil Bhardwaj, Sakshi Lakra, " A DEA model for Selection of Indian Cricket team players." 2019 Amity International Conference on Artificial Intelligence.

[6] Md. Jakir Hossain, "Bangladesh cricket squad prediction using statistical data and genetic algorithm".2018 4th International Conference on Electrical Engineering and Information and Communication Technology.

[7] Vipul Pujbai, Rohit Chaudhari, Devendra Pal, Kunal Nhavi, Nikhil Shimpi, Harshal Joshi, A survey on team selection in game of cricket using machine learning. Nov 2019, Vol 6, Issue 11, International Research Journal of Engineering and Technology.

[8] Park, Hyeoun-Ae, “An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain”. J Korean Acad Nurs Vol.43 No.2 April 2013.

[9] C. C. Chang and C. J. Lin, “LIBSVM: a library for support vector machines,” ACM transactions on intelligent systems and technology(TIST), vol. 2, no. 3, pp. 1–27, Jan. 2011.

[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis, vol. 821. John Wiley & Sons, 2012.

[11] Raj, J.S.,& Ananthi,J.V, “Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machine”. Journal of Soft Computing Paradigm(JSCP) in 2019,1(01),33-40.

[12] H. H. Lemmer, "A measure for the batting performance of cricket players: research article," South African Journal for Research in Sport, Physical Education, and Recreation, vol. 26, no. 1, 2004.

[13]Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M.Ingersoll, “An Introduction to Logistic Regression Analysis and Reporting”. The Journal of Educational Research 96(1):3-14 · (2012)September.