2022 3rd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET) 26-27 February 2022, Chittagong, Bangladesh

A Machine Learning Approach to Analyze the Performance of Bangladesh Cricket in T20

Tamal Joyti Roy[1], Md. Ashiq Mahmood[1], Aninda Mohanta [2], Diti Roy[1], Jannatun Tuba Jyoti[3] and Pran Krishna Ghosh[3]
[1]Institute of Information and Communication Technology
[1]Khulna University of Engineering & Technology (KUET), Khulna,Bangladesh
[2]Institute of Engineering & Management, Kolkata
[3]North Western University,Khulna,Bangladesh
tjroy13june@gmail.com,ashiqmahmoodbipu@gmail.com,anindamohanta21@gmail.com,
dity267@gmail.com, jannatuntuba@gmail.com,ghoshpk333@gmail.com

***Abstract*—Bangladesh is still an underachiever in terms of T20 cricket. Bangladesh has ranked in the least on the table of T20 international ranking. Our study has performed an analysis with K-means clustering to build a framework that shows the overall T20 international progression of Bangladesh cricket team players with the comparison of their playing years and other T20 cricket involvement. We have conducted our experiment with the help of a supervised machine learning algorithm. We have chosen our data set by taking player career statistics, who had debuted for Bangladesh between January 2005 to December 2020 in both test and T20 cricket. Our proposed framework and algorithm have tried to establish a connection of reasoning why Bangladesh cricket men team is not doing well in T20 cricket, what could go wrong. We have found a significant relationship between the playing span of any particular player and his connection with T20 international cricket as well as other T20 cricket. We have been able to group all the Bangladeshi T20 international cricketers into four groups on basis of their performance in T20.**

***Index Terms*—K-Means,Machine Learning,Bangladesh Cricket T20,Cricket,Clustering**

I. Introduction

There is an agreement of proficient estimation that cricket may have been thought up in the time of Saxon or Norman eras by progenies existing in the Weald, an area of impenetrable forests and defrayals in south-east England. The first locus of cricket played as a grown-up sport remained in 1611, and in the same era, a lexicon distinct cricket as a youngster's sports. It is thought that cricket may have come from bowls, by the interference of a batsman trying to stop the ball from attaining its board by drumming it absent [1]. Cricket is flowed by almost four billion people. Cricket is played on any spherical or elliptical-shaped grounds. There is a four-sided zone in the center of the field, named the pitch, that is the utmost significant share of the arena. altogether the bowling and batting take place. Customary woody twigs, called stumps and bails, are located on together trimmings of the terrain. A cricket ball, prepared of timber and leather, and a ligneous bat are obligatory to play cricket. The batsmen wear a full set of protecting equipment. Time of the old-style procedure of cricket, a match is frolicked among two teams, with 11 players on each team, in an innings set-up. One team befits to bat while the other side bowls and fields. The set of all-overs that a side bat for is called an innings. The team that bats first, tries to notch as many runs as likely. Afterward, the two teams take frolicked their innings, the team employing the most runs is declared the winner. Cricket was familiarized to North America by the English gatherings as primary as the 17th century, and time within the 18th period it inwards in other portions of the sphere.

It was familiarized with the West Indies by migrants and to India by British East India Company tars. It inwards in Australia practically the moment a settlement started in 1788 and it touched New Zealand and South Africa in the initial eons of the 19th century. At present moment there are many cricket formats is ongoing like-ODI, TEST, T20, Five5, and the most recent edition T10. Bangladesh is a cricket-playing nation no doubt about that. Bangladesh is a good side in one-day international cricket(ODI), but not good in test and T20. In this paper we showed an approach to why Bangladesh is not doing well in T20 cricket, utilizing the K-means clustering algorithm [2].

II. Related Works

In this literature [3] a new kind of K-means clustering was proposed. Named U-K-Means clustering, that automatically selects the number of K.Another studies showed [4] They had shown in what way neural network and K-Means could be used in a mixture, in time of clustering algorithms used to expressively signify data in an expressive path to output best batsman/bowler who supposed to be directed followed by another match condition. For predicting the winner of a particular cricket match with help of machine learning algorithms like logistic regression, SVM,k-NN, and random forest classifiers were applied [5], most of the cases for those algorithms used in the past scenario for forecasting the present match. The objective of this paper is to predict the winner of a cricket match through various machine learning algorithms like logistic regression, support vector machines, k-NN, Naive Bayes, and random forest classifiers applied to the past match performance records and improve the accuracy of the model. Nowadays people are in social media, their

978-1-6654-8397-1/22/$31.00 ©2022 IEEE

chirrup can be analyzed with their respective location that helped by the GPS, DBSCAN is used in this literature with the help of cosine similarity [6]. For improving, huge data sets efficacy this literature showed a significant procedure [7] which can be used for big data analysis with the help of the clustering method. Cricket is all about broadcasting, for that the resolution accuracy is needed to understand the fine details, this method [8] showed how the keyframe of video in cricket matches can be improved with the help of hybrid clustering frameworks. Cricket is highly dependent on weather, bad weather can hamper cricket-playing, for this, a beautiful study showed how to predict the weather before a cricket match [9]. This experimental analysis showed how the unsupervised machine learning algorithms with the help of cricket data and K-means clustering algorithm can predict a game result [10]. A fuzzy techniques combined with K-means clustering used for the IPL data set that clusters IPL data with three clusters [11].

III. Proposed Methodology

A K-means clustering algorithm attempts to set alike entries in the method of clusters. The quantity of groups is signified by K. Corresponding to the X and Y-axis in a two-dimensional vector system. We did a pre-process of our data, dividing our data variables in different segments and four attributes. "Fig. 1" is showing the overall working procedure of our experimental analysis. We prepared our data set with four columns. Name of the players, Number of T20 international matches played, Number of other T20 played, and the corresponding players playing span. We have chosen who debuted for Bangladesh between 2006 January to December 2020. We have chosen 53 players who debuted both for the test and T20 international for Bangladesh. Table 1 showing our sample data set.First of all, we have removed the strings and other lexical symbols. Placed data in the CSV file and saved it for analysis. We have used the python3 programming language for cluster analysis and other analytical experiments. We have conducted a correlation and regression analysis. We used the Jupyter notebook for our programming environment. "Fig. 2" and "Fig. 3" showing the relation of linear regression among the Number of T20 international played and Playing span, and Number of T20 international played, and the Number of other T20 Played. Table II is showing the after transposing data set results.

TABLE I: Sample Data Set

SL	Name	Number of T20 international played	Number of T20 played	Playing Span (years)
1	Abdur Razzak	34	90	19
10	Raqibul Hasan	5	28	15
20	Sohag Gazi	10	83	9
.	.	.	.	.
.	.	.	.	.
53	Musfiqur Rahim	86	202	16

A. Regression

Linear Regression: Linear regression refers to the continuity of sets that is outputted with a reference value.
Logistic Regression: Logistic regression analysis can give an output that is constant, more accurate in measure. It provides a curve that is like the alphabet S-shaped. The value of that can vary between o and 1 which means there always be a probability of this algorithm. Among three types of logistic regression, we used multi nominal and ordinal.
First of all, we remove the strings and other lexical symbols. Placed data in the CSV file and saved it for analysis.

Algorithm 1 Proposed Algorithm for finding the clusters

1: Start
2: Read data set
3: calculate the ca,msc and ra
4: Split_ds(X,Y)(t=0.30)
5: initialize xtrain,xtest,ytrain,ytest
6: if$(xtest, ytest < 0.30)$ then
7: goto step 4
8: else goto setp:6
9: Linearregressionanalytic(X,Y)
10: fittingthedataset(xtrain,ytrain)
11: make centroid_initial with $k = D_{n+1}$
12: begin iteration
13: for reoccurance_n:
14: goto step:8
15: initialize $k = D_{n+1} + 1$
16: End

We have used the python3 programming language for cluster analysis and other analytical experiments. We have conducted a correlation and regression analysis. We used the Jupyter notebook for the evaluation. "Fig. 1" showing the experimental flowchart. The relation of linear regression between the Number of T20 international played and Playing span, and Number of T20 international played, and the Number of other T20 played showing in the "Fig. 2" and "Fig. 3" . Table II is showing the after transposing data set results. The data set combined 25%,50%, and 75%. The mean, standard deviation. minimum values and maximum values also showed. Table III showing the axis predicting values. We split our data set for testing and training. We used machine learning under the supervision class, the regression analysis. The X and Y predicted values are shown in the data size column. Each array is 2x2 in size that means 2-dimensional. We have proposed an algorithm for our experiment, showing under the notation of the Algorithm.In the first observed cluster, we have evaluated that the all data from the csv files grouped in three sections. The mean square error was zero for the first observed , the second observed cluster showed exact same three grouping with the mean square error zero.This equation $Y_{D_{i+1}} = f(C_i, \beta) + E$ followed the linear regression method. Here the E is the error in minimum level. The Y is the dependent variable and the D independent.

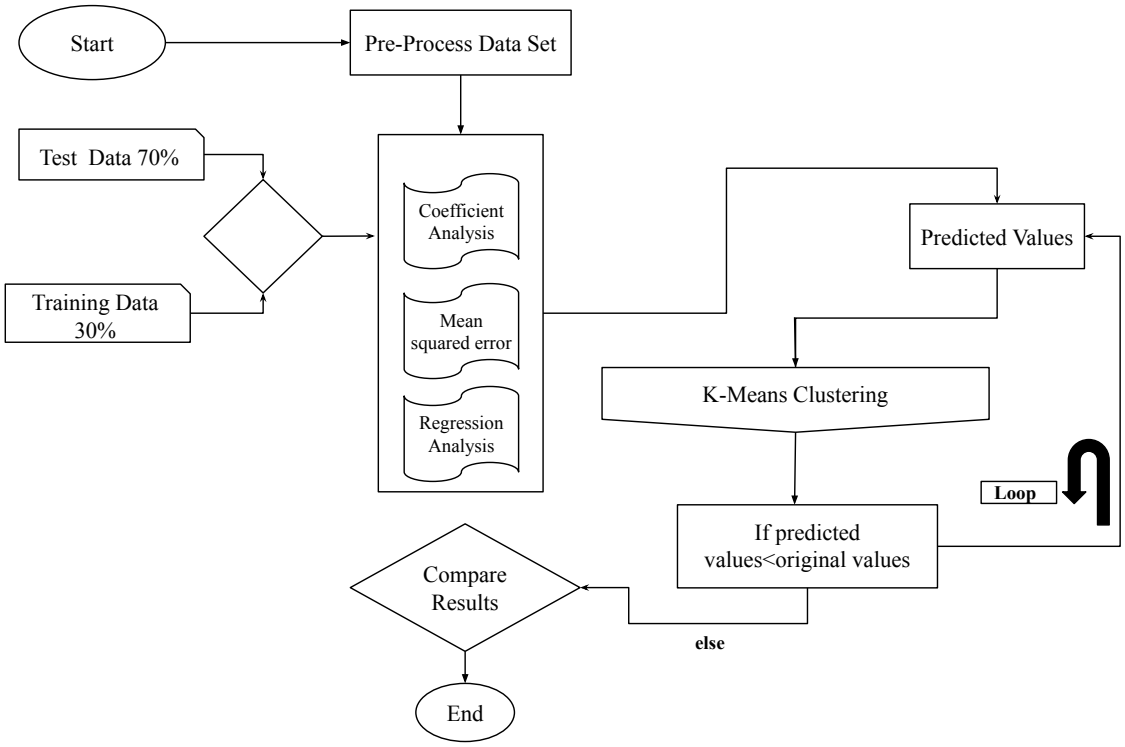

Fig. 1: Experimental Flow chart

TABLE II: Transposing the data set values

	Count	Mean	Std	Min	25%	50%	75%
Number Of T20 Played	53	16.1320	22.62	0	1.0	7	19
Number Of Other T20	53	82.943	61.83	1	31	81	110
Playing Span	53	11.54	3.880	4.0	9.0	12.0	15.0

$$MSE_n(D1, D2) = \frac{1}{M.N} \tag{1}$$

$$MSE_{n+1} = \sum_{c=1}^{N}\sum_{v=1}^{M}[D1(c,v) - D2(c,v)]^2 \tag{2}$$

From the equation 1 and 2 we have now merged the n to $n+1^{th}$ item

$$F = \frac{1}{M.N} \times \sum_{c=1}^{N}\sum_{v=1}^{M}[D1(c,v) - D2(c,v)]^2 \tag{3}$$

B. Data Analysis

$$Y_{D_1} = c + xK \tag{4}$$

$$Y_{D_{n+1}} = c + x(n+1)K \tag{5}$$

$$Y_{D_i} = f(C_i, \beta) + E \tag{6}$$

TABLE III: Axes Predecting values

Predecting Axis	Data Size
Y	([1, 0, 2, 1, 2, 0, 1, 2, 1, 2, 0, 1, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 2, 1, 2, 1, 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 0])
X	([1, 1, 0, 1, 2, 0, 1, 2, 1, 2, 0, 1, 1, 1, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 2, 1, 0, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 2, 1, 2, 1, 0, 2, 2, 1, 1, 1, 0, 2, 0, 2, 2, 0])

TABLE IV: Test set values

Variable	Test data values
x	[4.671,67,58],[2.83,61,58],[1.66,90,52.25],[5.62,54,58,], [2.67,80,37.9],[1.875,60,23.78]
y	Predicted Axis
Intersecting	9.690392959913552
Coefficient of X	[0.0077651 0.01960725]
Mean_squared	14.128094949699031

$$Y_{D_{i+1}} = f(C_i, \beta) + E \tag{7}$$

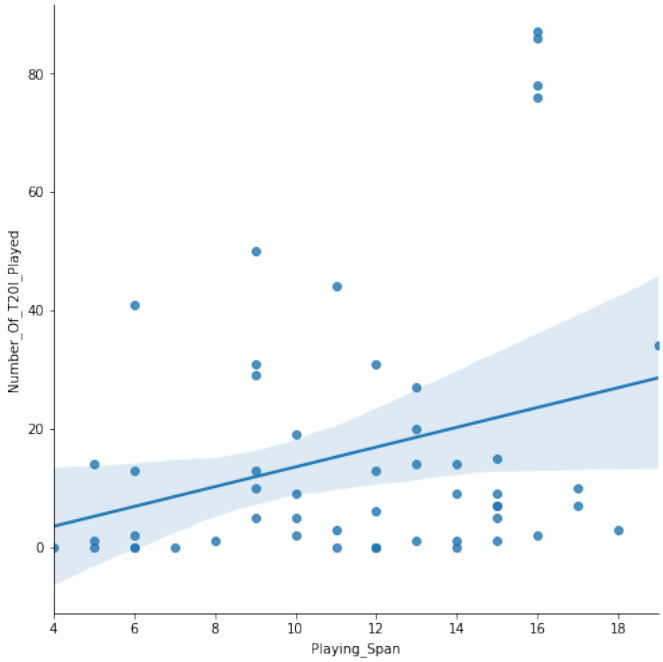

Fig. 2: First regression with Number of T20 played vs playing span

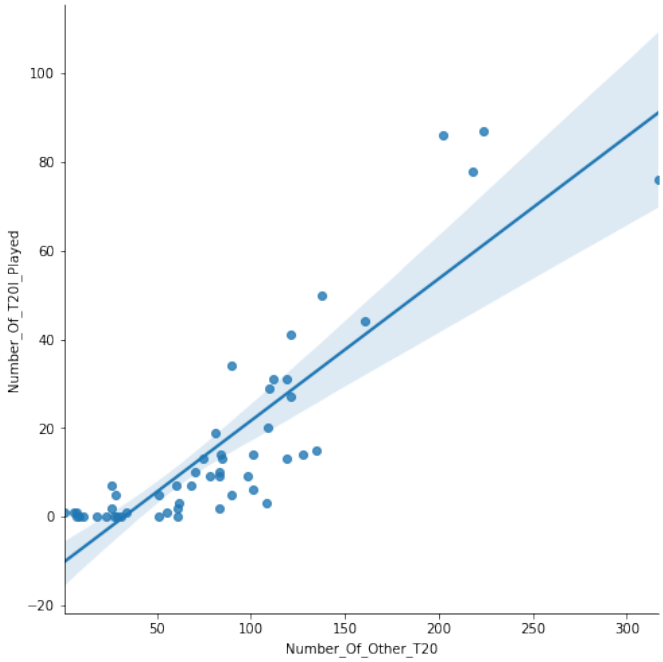

Fig. 3: Second regression with Number of other T20 played vs playing span

$$TotalPlayersinClusters = \frac{\sum_i \lambda_i^3}{6} \tag{8}$$

$$NumberofPlayersInitialdata = \frac{\text{Sum}D_2\text{-Total}D_2}{2} \tag{9}$$

$$C(D1 + D2) = \frac{\sum_i \lambda_i^3}{\text{Sum}D_2\text{-TotalData}D_1} \tag{10}$$

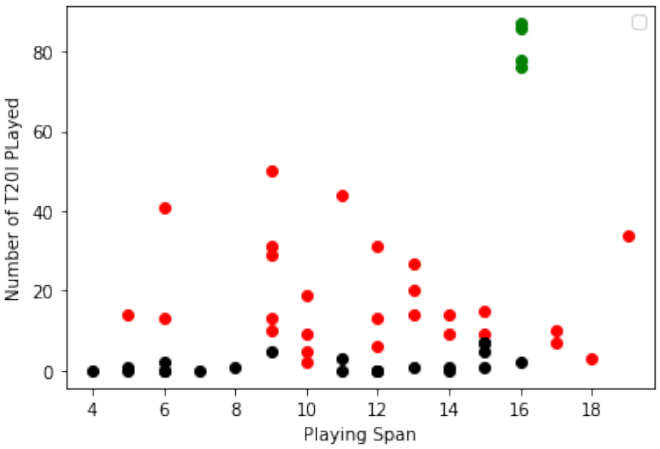

Fig. 4: First Observed Clustering

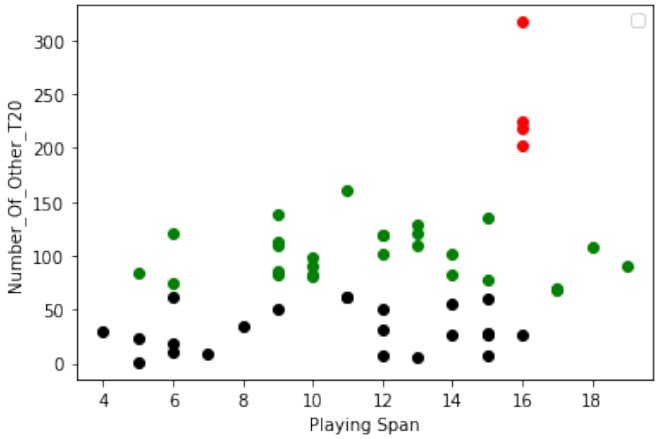

Fig. 5: Second Observed clustering

Algorithm 2 K-Means Procedure

1: Start
2: Initialize μ=first centroid, D1 random
3: for any value in K, K=1 to K do
4: compute distance σ= max-min
5: ω=probability
6: ω_{i+1}=normalize the distribution
7: $\mu = C_m$ centroid distributed
8: end
9: repeat if new centroid is measured

IV. Experimental Results

We used the Jupyter notebook python3 environment for our evaluation of the study. “Fig. 8” showing the confusion matrix of our data set. Our whole data set is visualized in “Fig. 6” and “Fig. 7” in scatter plots. Our test set is showing in table IV.K-means clustering was performed in several steps. Algorithm I

is proposed by us for this experimental analysis. As it's an unsupervised machine learning algorithm we wanted to test some values first. That's why without any assigned values we plotted the present data set to understand it. "Fig. 3" and showing the initial cluster results without assigning any k value. "Fig. 9"are the results of our K-means clustering. "Fig. 4" was obtained by assigning k=3 in the experiment and "Fig. 4" obtained by assigning k=4.The equation $C(D1 + D2) = \frac{\sum_i \lambda_i^3}{\text{Sum}D_2\text{-TotalData}D_1}$ showed how the clusters were emerged. "Fig. 6" took the k+1 iteration till the fixed clusters were observed. The X(cross mark) showing the centroid of both clusters.For the total player this $TotalPlayersinClusters = \frac{\sum_i \lambda_i^3}{6}$ equation have been followed. The whole Bangladesh cricket team data set is divided into four clusters. We analyzed the final cluster results and scrapped the data from them to new decision set. We assigned and then processed those data and finalize them in "Fig. 6", the results of the cluster. From our experiment, there are four groups emerged and we divided those groups in terms of clustering. The cluster 1 having players like Shakib Al Hasan, Tamim Iqbal, Musfiqur Rahim, Mahmudullah.

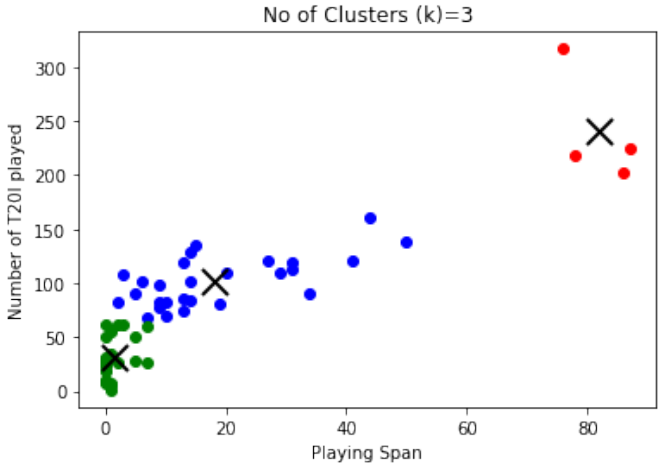

Fig. 6: K-Means Clustering of data set where k=3

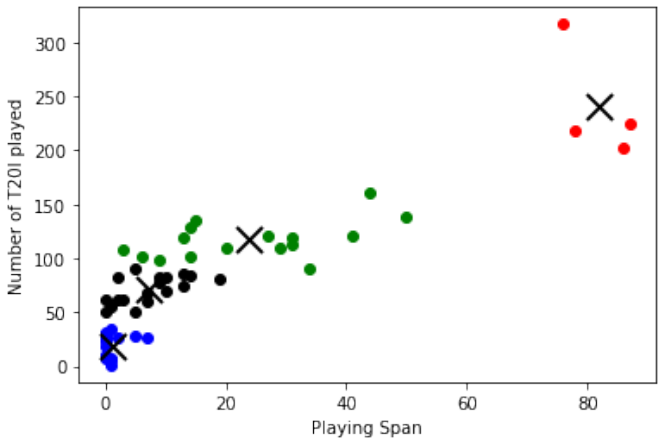

Fig. 7: K-Means Clustering of data set where k=4

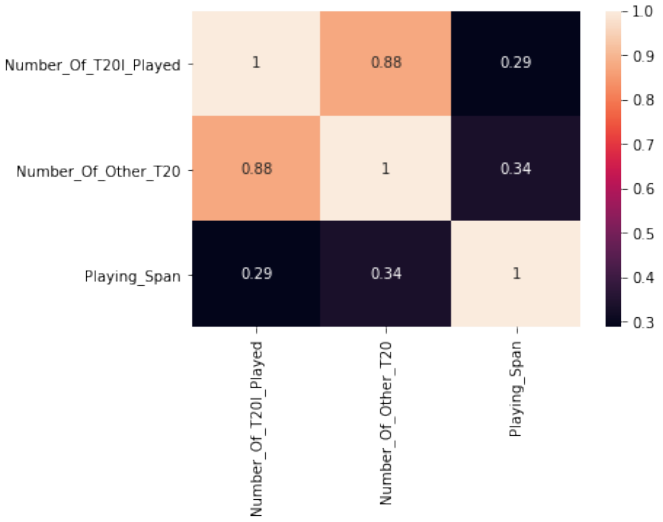

Fig. 8: Confusion Matrix

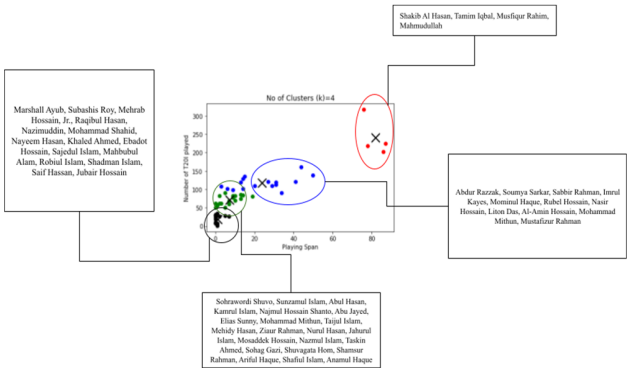

Fig. 9: Bangladesh Cricket Clustering with the player's name

Cluster 2 having players like Abdur Razzak, Soumya Sarkar, Sabbir Rahman, Imrul Kayes, Mominul Haque, Rubel Hossain, Nasir Hossain, Liton Das, Al-Amin Hossain, Mohammad Mithun, Mustafizur Rahman. Cluster 3 having players like Sohrawordi Shuvo, Sunzamul Islam, Abul Hasan, Kamrul Islam, Najmul Hossain Shanto, Abu Jayed, Elias Sunny, Mohammad Mithun, Taijul Islam, Mehidy Hasan, Ziaur Rahman, Nurul Hasan, Jahurul Islam, Mosaddek Hossain, Nazmul Islam, Taskin Ahmed, Sohag Gazi, Shuvagata Hom, Shamsur Rahman, Ariful Haque, Shafiul Islam, Anamul Haque. And last but not least cluster 4 having players like Marshall Ayub, Subashis Roy, Mehrab Hossain, Jr., Raqibul Hasan, Nazimuddin, Mohammad Shahid, Nayeem Hasan, Khaled Ahmed, Ebadot Hossain, Sajedul Islam, Mahbubul Alam, Robiul Islam, Shadman Islam, Saif Hassan, Jubair Hossain. "Fig. 8" is the confusion matrix of our data set."Fig. 9" showed how the cluster divided our Bangladesh cricket data set and cluster them.

V. CONCLUSION

Our study can not predict how Bangladesh cricket team players will perform in the future but surely can focus the lights on where things are going wrong. We have collected all the valid data for our experiments. No missing values were conducted. Furthermore, the K-means clustering with the help of supervised machine learning algorithms proposed a new framework for analyzing cricket nations' performance. We took 53 players for our experiment and if this study is conducted with more players with another team like India, Australia, or England then there will also be some outcomes. The limitation of our experiment was there were not enough debuting players for Bangladesh if we compared with other teams, that's why the clustering iteration was done with N=53. The final thoughts and limitation are- if the time range could be lengthy the there would have more clusters of players. The more clusters mean we could have divided the players into many categories. This could be helped in more accurate analysis. Future work will be added having analyze new international teams with more players.

REFERENCES

[1] "International Cricket Council." https://www.icc-cricket.com/about/cricket/history-of-cricket/early-cricket (accessed Jul. 03, 2021).

[2] "A Simple Explanation of K-Means Clustering and its Adavantages." https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/ (accessed Jul. 03, 2021).

[3] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[4] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[5] P. Jhansi Rani, A. Vidyadhar Kamath, A. Menon, P. Dhatwalia, D. Rishabh, and A. Kulkarni, "Selection of Players and Team for an Indian Premier League Cricket Match Using Ensembles of Classifiers," Jul. 2020, doi: 10.1109/CONECCT50063.2020.9198371.

[6] N. Pandey, "Density based clustering for Cricket World Cup tweets using Cosine similarity and time parameter," Mar. 2016, doi: 10.1109/INDICON.2015.7443520.

[7] S. S. Desai and J. A. Laxminarayana, "WordNet and Semantic similarity based approach for document clustering," in 2016 International Conference on Computation System and Information Technology for Sustainable Solutions, CSITSS 2016, Dec. 2016, pp. 312–317, doi: 10.1109/CSITSS.2016.7779377.

[8] S. Premaratne and L. Jayaratne, "A Novel Hybrid Adaptive Filter to Improve Video Keyframe Clustering to Support Event Resolution in Cricket Videos Event Resolution in Cricket Videos View project Education Data Mining For Studying The Impact of Learning View project," Int. J. Eng. Adv. Technol., no. 8, pp. 2249–8958, 2019, doi: 10.35940/ijeat.F1005.0986S319.

[9] M. K. Nallakaruppan, S. Nazz, K. Madhuvanthi, S. Karthikeyan, and M. Medarametla, "Predicting the weather for uninterrupted cricket matches and outdoor sports events," in Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019, Jan. 2019, pp. 451–458, doi: 10.1109/CONFLUENCE.2019.8776929.

[10] K. Parameswaran, "Vector quantization, density estimation and outlier detection on cricket dataset," 2013, doi: 10.1109/ICCCI.2013.6466249.

[11] "Fuzzy Clustering Technique in IPL database," Int. J. Publ. Gr., vol. 2, no. 7, pp. 259–262, 2012, Accessed: Jul. 03, 2021. [Online].