Machine Learning Based Yield Prediction of Agricultural Crop

Team QuadCore,   December 23, 2018

B. M. Raihanul Haque
Section - 2
ID: 1512756042
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
raihan011235@gmail.com

Yameen Irteza Hossain
Section - 2
ID: 1611135042
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
yameen.irteza@gmail.com

Md. Julhas Hossain
Section - 2
ID: 1513156642
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
julhas78@gmail.com

Md. Monzurul Islam
Section - 2
ID: 1511383042
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
Monzurul.islam@northsouth.edu

***Abstract*—Being the primary source of food in Bangladesh, predicting yield of rice successfully is necessary to achieve maximum profit and meet the demands of carbohydrates of general people by providing adequate amount of rice. This low-lying riverine country has tropical monsoon climate characterised by heavy seasonal rainfall, high temperatures, high humidity etc. and these parameters have huge impact on rice production. People involved in food industry need to choose certain portion of land carefully which will lead to desirable production and keeping that in mind, it is required to develop an efficient system for predicting future yield of agricultural crops i.e. rice. Five machine learning models have been incorporated such as Linear Regression, Artificial Neural Network, Self-organizing Map, Random Forest and XGB Regressor on three types of rice of Bangladesh (Aus, Aman and Boro). Correlation between past environmental patterns and crop production rate is considered to train the models and they are tested using unknown climatic variables later on. There is scope for extensive use of these methods despite the fact, three types of rice have been analyzed initially.**

***Index Terms*—Bangladesh rice yield, prediction, climate, linear regression, artificial neural network, self-organizing map (SOM), random forest, xgb regressor.**

I. INTRODUCTION

Situated in the South Asian territory, the economy of Bangladesh is vastly dependent on Agriculture. The reason behind this dependency is this country has fertile soil due to rich alluvium, ample water supply and sub tropical climate which is extremely suitable for crop production. This is why agriculture is the largest employment sector in Bangladesh which is, according to 2015 statistical yearbook of agriculture, more than 45 percent of total labour force. Agriculture sector contributes about 17 percent to the country‘s Gross Domestic Product (GDP). In 2012, a total of 33,889,632 metric tons of rice have been produced and market value of the rice in international market was $8,649,167,000.

In past few years, behavior of weather has changed and this has caused unpredictability in production. This is why historic data has been collected to analyze the pattern of production rate. Aman, Aus and Boro are the 3 major types of crops in Bangladesh. Correlation of sowing and harvesting time with seasonal variation is a primary factor for crop yield.

The main objective of this research is to develop efficient models that will predict rice production better. On one side, we have seven input parameters such as rainfall, wind speed, cloud coverage etc. and on the other side we have predicted output. All the data has been collected from Bangladesh Bureau of Statistics (BBS) and Bangladesh Meteorological Department (BMD). Total number of five machine learning algorithms have been implemented on the data set which will be discussed in the later part of the paper.

The rest of the paper is organized as follows: in section 2, similar works have been considered and evaluated. The research methodology is described in section 3. Results have been presented in section 4. Section 5 provides the insights about the scope of future work. Conclusion has been drawn in section 6.

II. Related works

The goal was to find the works that addresses co-relation between climactic variable and crop production.

Anjela Diana Corraya, and Sonia Corraya, two of the researchers in Bangladesh have worked on price and yield prediction before and presented it on the paper entitled ”Regression based Price and Yield Prediction of Agricultural Crop”. Their primary source of data was Bangladesh Rice Research Institute (BRRI). For interpolation Auto Regressive Moving Average model (ARMA) was used but since it causes fluctuation in the output the authors have decided to use weighted linear regression which has produced better output of accuracy 78.75% and 83.55% for predicting price and yield respectively.

In their work ”Machine Learning in Agriculture: A Review”, Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson and Dionysis Bochtis have generalized different machine learning techniques in the realm of agriculture. They have collected 40 different paper and sorted out the algorithms that have been used. For instance, to predict yielding Bayesian models, support vector machine (SVM) etc. are used.

”Machine Learning Facilitated Rice Prediction in Bangladesh” is conducted by Mohammad Motiur Rahman, Naheena Haq and Rashedur M Rahman and they worked on past 20 year data. For clustering, SOM has been used and for classification, some of the algorithms that have been implemented are non linear regression, regression tree, ensemble learning etc. Based on that they have got average yield, minimum yield maximum yield. Error obtained in each method was compared using root mean square error (RMSE).

M. T. Shakoor, K. Rahman, S. N. Rayta and A. Chakrabarty in their paper ”Agricultural production output prediction using Supervised Machine Learning techniques,” approached the problem with the supervised machine learning technique the authors found the profitable agricultural crops. The research provides a list of profitable crops in a particular area using decision making algorithms. The research only focuses on the six major crops for ten regions. The prediction was made using the k-Nearest Neighbor and Decision Tree Learning, ID3 (Iterative Dichotomiser 3) algorithms. The result of this analysis is proven to be reliable as in many of the cases the resultant error was below 10%.

”Crop Prediction System using Machine Learning” is written by Prof. D.S. Zingade1, Omkar Buchade2, Nilesh Mehta, Shubham Ghodekar and the paper is about presenting an android application based system which predicts the most profitable crops, in the current weather and soil conditions where the weather data is obtained from repositories.The weather data is collected from the Indian Meteorological Department for parameters like temperature, rainfall and others which gives insights to the crops and mulitple linear regression has been used for this prediction, which gives the prediction to the user about the multiple suggestion of crops conferring to the duration of crop.

The goal of this paper named ”Smart Farming System: Crop Yield Prediction Using Regression Techniques” by Ayush Shah, Akash Dubey, Vishesh Hemnani, Divye Gala, D. R. Kalbande is to identify a relationship between yield(dependent variable) and other independent variables such as temperature, humidity, rainfall. The Yield and Weather data is collected from the United States Department of Agriculture for the state of Iowa. Three algorithms are used: Multivariate Regression, Support Vector Machine Regression and Random Forest Application, the RMSE, MAE and MdAE values are calculated for each algorithm applied and are compared among each other where the graphs show the SVM is the one with the most accuracy for predicting the crop yield having a maximum R-squared value of 0.98.

Konstantinos G. Liakos and Patrizia Busato introduced all the method can be used in agriculture sector for machine learning. There are many methods and algorithm in machine learning. Its quite difficult to choose witch way we should go. For solving this problem this paper introduced many algorithm only for agriculture site to make our decision easy. They discussed about Similarity-based versus knowledge-based. Which help us to choose the basic our algorithm. There are many more, noise-tolerant versus exact, top-down versus bottom-up, supervised versus unsupervised, interactive versus non-interactive, single- versus multi-paradigm. This paper if full of paper reference and gave us idea about different feature and algorithm about machine learning in agriculture field.

TRobert J. McQueen and Stephen R. Garner introduced machine learning algorithm based on India. This paper is based on Indias agriculture site. As Indias weather is also changing they introduced this paper for better output in agriculture site by predicting weather and agriculture growth. Machine learning methods frames different models based on previous investigation which can then be used to anticipate new data. The model built is a result of a learning process that extracts useful information about the data generation process of the system using the previous observations.

III. Methodology

Past data of rice production has been collected from Bangladesh Bureau of Statistics (BBS) and and previous data of climate have been gathered from Bangladesh Meteorological Department (BMD). The algorithm used for the clustering is known as Self Organizing Map (SOM). For classification, five algorithms have been used which are, Linear Regression, Artificial Neural Network, Self-organizing Map, Random Forest and XGB Regressor. For both classification and clustering past 20 years of climatic data and 20 years rice production has been used. Entire data set has been divided into three subset i.e. the training set, cross validation set and test set. After the successful implementation of the algorithms mentioned earlier error has been calculated using RMSE method.

A. Data Preprocessing

Data collected from BBS and BMD was not compatible to each other.

1) : There were lots of missing data points. If there were too many empty cells in a column they were ignored.

2) : For interpolation, mean value was considered.

3) : Incompatibility was seen in two different data set as BBS has published rice production data from the middle of one year to the middle of another year, while BMD does it from the start of the year to the end of same year. Mean value has been incorporated to resolve this issue.

B. Clustering

The input features are rainfall (millimeters), cloud coverage (octas), hourly sunshine, percentage humidity, wind speed, maximum and minimum temperature. The values of these seven parameters have been averaged out for 16 different regions. Then in order to find cluster of region having the same climatic pattern self organizing map (SOM) is used.

In figure-1, SOM training iterations progress is shown. In every iteration, distance from each nodes weights to the samples represented by that node is reduced.

In figure-2, visualization of the count of how many samples are mapped to each node on the map and whether the sample is relatively uniform or not.

In figure-3, visualization is of the distance between each node and its neighbours is revealed.

In figure-4, 5 and 6, visualization of the distribution of a single variable across the map is represented which is also known as heatmaps.

In figure-7, the total within-groups sums of squares vs the number of clusters in a K-means solution is shown which can suggest the appropriate number of clusters.

Figure 8 shows 5 different cluster representing group of different regions. For example, Dinajpur, Faridpur and Jessore belongs to the same cluster.

All these led to the decision how many clusters are required.

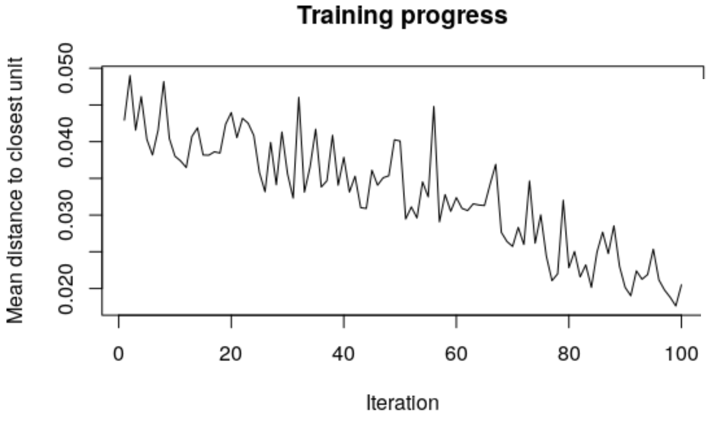

Fig. 1. SOM training iterations progress

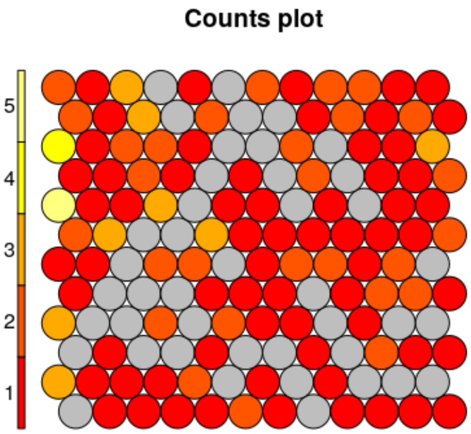

Fig. 2. Count of samples which are mapped to each node on the map

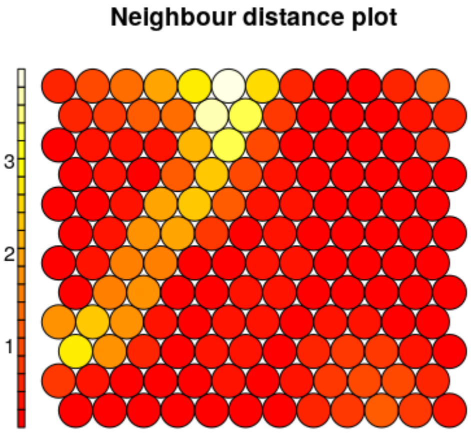

Fig. 3. Distance between each node and its neighbours

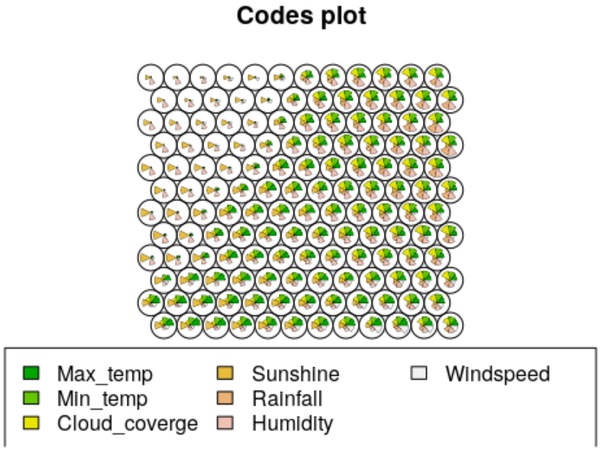

Fig. 4. Node weight vectors

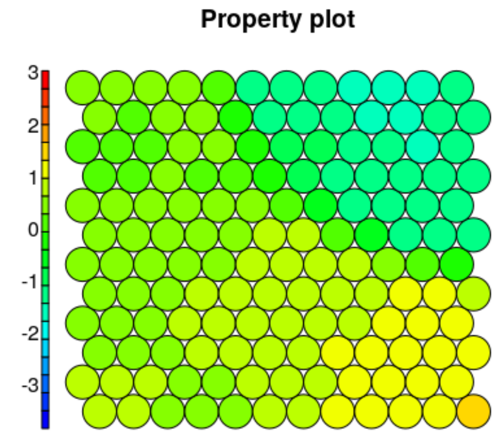

Fig. 5. Heatmaps

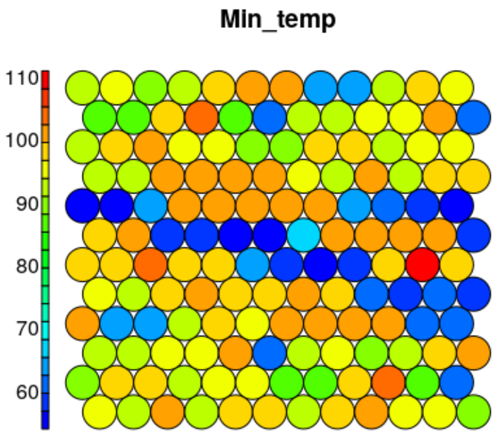

Fig. 6. Heatmaps

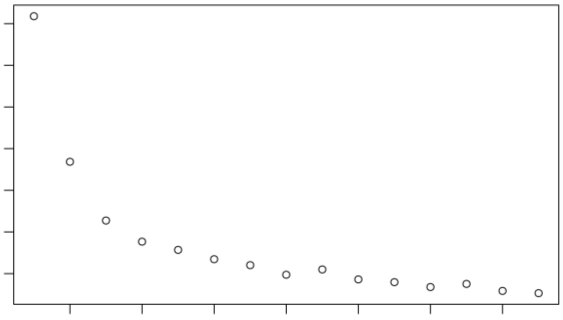

Fig. 7. K-means

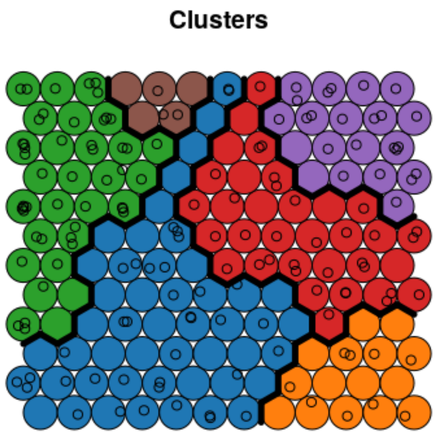

Fig. 8. SOM Clusters

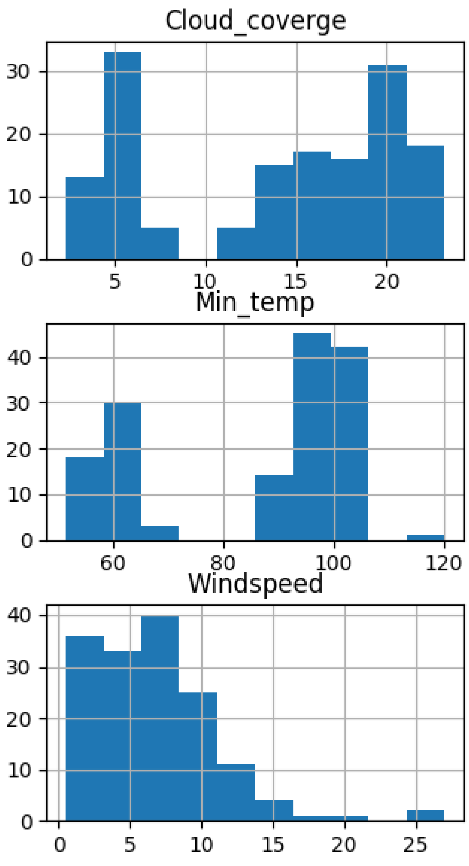

Fig. 9. Histograms

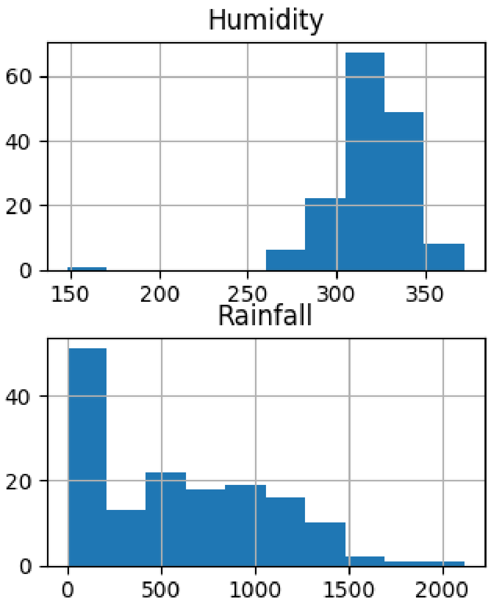

Fig. 10. Histograms

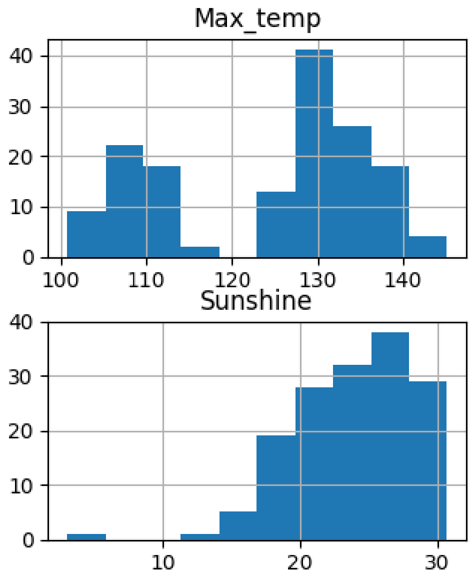

Fig. 11. Histograms

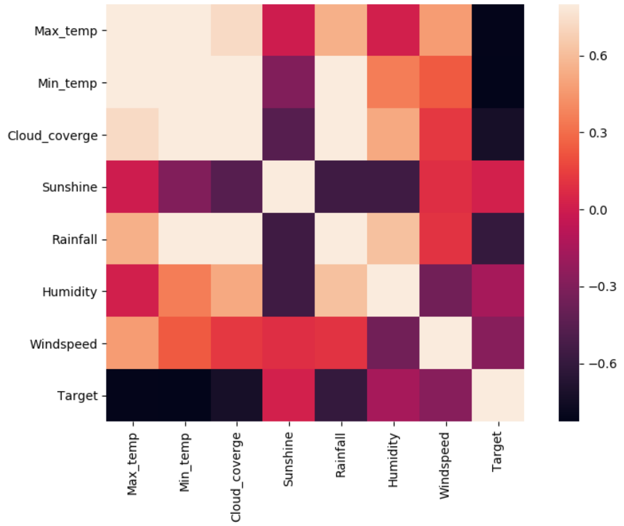

Fig. 12. Correlation between input parameters

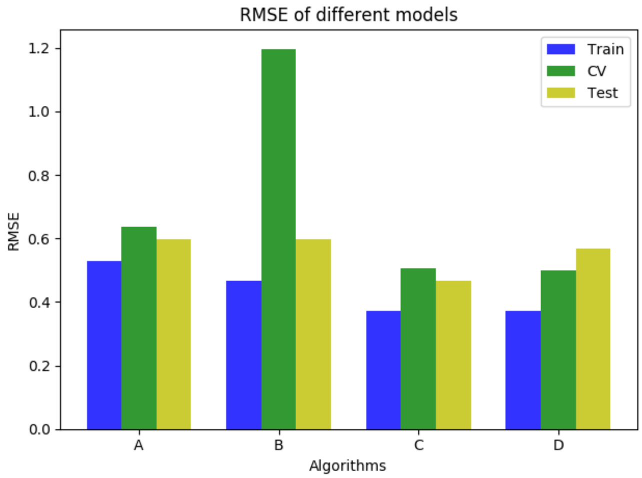

Fig. 13. RMSE of designed models

C. *Classification*

1) Linear regression: Here, different theta values have been set to find out correlation between independent and dependent variable. Normal equation is used to find optimal theta value.

2) Artificial neural network: The main idea involves mimicking biological neuron by setting up input layer, hidden layer and output layer. Figure 9, 10 and 11 is representing the histograms of seven input parameters. It is evident that all the input features form skewed shaped and u-shaped curves. Figure 12 is revealing the relation of each parameter with every other parameter. Dark squares mean that corresponding parameters are strongly correlated.

3) Random forest: Multiple decision trees have been constructed to build the forest where leaf nodes indicates completion of classification. Here, output variables are considered as a continuous instead of nominal.

4) XGBoost Regressor: XGBoost algorithm is designed to conduct a multiclass classification. It is an implementation of gradient boosted decision trees. Which means, this algorithm gives rise to a prediction model in the form of an ensemble of several weak prediction models.

IV. Result

RMSE method is used to calculate error of each model. Difference between the predicted output and the observed output is calculated and then the value is squared. After that it is averaged, and is being divided by the number of observation. Finally, square root of the immediate value is calculated. This process is implemented for 3 types of error; training error, testing error and cross-validation error as the data set was divided into three sets. Figure 13 is showing the comparison between different RMSE values. Here, A = linear regression, B = Artificial neural network, C = Random forest and D = XGB regressor.

A comparison between observed and predicted value is given below.

	Linear Regression	ANN	Random Forest	XGB
Average value in Metric Tonne/hectare (observed)	2.45	2.45	2.45	2.45
Average value in Metric Tonne/hectare (predicted)	2.41	2.45	2.47	2.47
Maximum yield in Metric Tonne/hectare (observed)	4.44	4.44	4.44	4.44
Maximum yield in Metric Tonne/hectare (predicted)	3.48	2.59	2.92	3.19
Minimum yield in Metric Tonne/hectare (observed)	0.86	0.86	0.86	0.86
Minimum yield in Metric Tonne/hectare (predicted)	1.82	1.34	1.07	0.89
R-squared score (in percentage)	73-75	54-80	80-87	80-87

The r-squared value is based one majority train, cross-validation and test set values which produces higher positive values.

Maximum yield value was recorded in Dhaka back in 2010. Minimum yield value was recorded in Tangail back in 2009.

By observing the predicted value it is evident that all the models did a fair job in predicting rice yield. Random forest and XGB regressor did slightly better than other algorithms.

V. Further Scope

Prediction can be improved by incorporating new features such as soil pH and level of nitrogen content. The quality of climactic data may play a vital role as well. This research is not only limited to rice prediction rather this highly effective classifier can be trained to predict the yield of other crops like wheat, jute, potato, sugarcane and so on.

VI. Conclusion

Predictive power of different classifiers is evident. These predictions can help farmers and entrepreneurs to battle against unfortunate events as they can be evaluated to eliminate uncertainty. As the classifier produced somewhat similar results, they indicates that they can be used to develop strong models for crop prediction.

References

[1] Mohammad Motiur Rahman, Naheena Haq, and Rashedur M Rahman, “Machine Learning Facilitated Rice Prediction in Bangladesh,” 2014 Annual Global Online Conference on Information and Computer Technology, Louisville, KY, 2014, pp. 1-4.

[2] Anjela Diana Corraya and Sonia Corraya, “Regression based Price and Yield Prediction of Agricultural Crop,” in International Journal of Computer Applications (0975 - 8887) Volume 152 - No.5, October 2016.

[3] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson and Dionysis Bochtis, “Machine Learning in Agriculture: A Review,” sensors, MDPI, Received: 27 June 2018; Accepted: 7 August 2018; Published: 14 August 2018.

[4] en.wikipedia.org, ’Agriculture in Bangladesh’, 11 December, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Agriculture_in_Bangladesh. [Accessed: 17- Dec- 2018].

[5] bbs.portal.gov.bd, ’Yearbook-2015’, 2015. [Online]. Available: http://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/1b1eb817_9325_4354_a756_3d18412203e2/Yearbook-2015.pdf. [Accessed: 17- Dec- 2018].

[6] M. T. Shakoor, K. Rahman, S. N. Rayta and A. Chakrabarty, ”Agricultural production output prediction using Supervised Machine Learning techniques,” 2017 1st International Conference on Next Generation Computing Applications (NextComp), Mauritius, 2017, pp. 182-187.

[7] Prof. D.S. Zingade1, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar and Chandan Mehta, ”Crop Prediction System using Machine Learning”, Department of Computer Engineering, All India Shri Shivaji Memorial Society’s Institute of Information Technology.

[8] Ayush Shah, Akash Dubey, Vishesh Hemnani, Divye Gala and D. R. Kalbande, ”Smart Farming System: Crop Yield Prediction Using Regression Techniques,” Ray K. (eds) Proceedings of International Conference on Wireless Communication. Lecture Notes on Data Engineering and Communications Technologies, vol 19. Springer, Singapore.

[9] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson and Dionysis Bochtis, Machine Learning in Agriculture: A Review, sensors, MDPI, 14 August 2018.

[10] TRobert J. McQueen, Stephen R. Garner, Craig G. Nevilla-Manning, Ian H. Witten, Applying Machine Learning to Agricultural Data, International Journal of Computer Applications , Volume 262 - No.3, February 2017.

[11] shanelynn, ’Self-Organising Maps for Customer Segmentation using R’, R-bloggers, February 3, 2014 [Online]. Available: https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/. [Accessed: 17- Dec- 2018]

[12] Mohammed Ma'amari, 'Deep Neural Networks for Regression Problems', Towards Data Science, September 29, 2018 [Online]. Available: https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33. [Accessed: 17- Dec- 2018]