Predicting Early Crop Production by Analysing Prior Environment Factors

Tousif Osman, Shahreen Shahjahan Psyche, MD Rafik Kamal, Fouzia Tamanna, Farzana Haque, and Rashedur M. Rahman(✉)

Department of Electrical and Computer Engineering, North South University, Plot-15, Block-B, Bashundhara, Dhaka, Bangladesh
{tousif.osman,shahreen.psyche,rafik.kamal,fouzia.tamanna, rashedur.rahman}@northsouth.edu, soniya.farzana@gmail.com

Abstract. Bangladesh has an agriculture dependent economy and hence prediction of agricultural production is of great importance to us. In this research we develop a model that considers and analyzes weather and climate prior to specific crop plantation and maps a correlation between these two. It allows us to provide information about the crop state, in quantity and quality with the possibility of early warnings so that timely interventions can be undertaken. The approach advocated in this paper is to help the people with food security and early warning system.

Keywords: Data mining · Adaptive learning · Machine learning · Prediction · Agriculture · Soft computing · Environment

1 Introduction

The economy of Bangladesh is highly depended on the agriculture but agriculture faces a lot of challenges each year. One of the problems agriculture face is crop yield prediction. Primarily the crop yield depends on weather, planning of harvest operation, diseases and pests. It is important for the farmer to decide which crops will have the best production before the harvest takes place. Timely and accurate crop yield prediction helps the farmer to ensure maximum production of crops for present and future. We have constructed a model in this research considering environmental variables on which crop production depends directly and indirectly. It uses the value of sunshine, cloud, wind-speed, humidity, temperature and rainfall for computing the production of certain crop. Extracting knowledge from this raw data is very difficult. But one can get the knowledge of the data to predict the major agricultural production which is depended on climate and weather by data mining methods and techniques. So, we have designed our system by which farmers can produce more crops.

2 Related Work

A good number of researches has been done to predict crop yield. We have studied and incorporated ideas and methods from those researches in to our

M. Akagi et al. (eds.), *Advances in Information and Communication Technology*, Advances in Intelligent Systems and Computing 538, DOI 10.1007/978-3-319-49073-1_51

system. In this section we will mention few research works and their approaches those we have considered while doing our research and building the system.

In paper [1], authors have found the yield of crops of Bangladesh using regression tree, neural network, ensemble learning, and linear/nonlinear regression. To find the similar characteristic of environment between the regions they have used k-means clustering and self-organizing map (SOM) and they mentioned two prominent regions. After doing this, they have taken average of 4 month's weather attributes values as input and yield of rice as output of the same time. In next research [2], the authors described Naive Bayes techniques for classification of agricultural land soils. The research utilized data collected from seven commonly occurring soil types and analyzed with agricultural data set using data mining techniques. The Naive Bayes helped to classify the soil based on texture of soil profiles and also classification worked with a simple probabilistic classifier based on independence assumption. In paper [3] the authors have mainly pointed on data mining techniques to predict major crop yield with existing data. The authors in [4] proposed a system named GZ-Agri GS. The local farmers can easily gain the knowledge about scientific guidance for crop management decisions from the online service. This research uses Domain knowledge and model with GIS technology to gain useful results. In the research paper [5], the authors discussed a system that is applicable for the real-life data quality control and assurance of reliable and error-free agro-meteorological data. The authors design a prototype system for detecting abnormal weather prediction.

3 Theory

The main goal of this research is to predict the production of some major crops of Bangladesh for upcoming season. We have used 10 years monthly environmental data and crop production data to model our system. Major components of the research have been described in the following three subsections.

3.1 Prediction Hypothesis

We want to construct a system that depend on prior environmental parameters (i.e., sunshine, rainfall, soil salinity). Yield of crops depends on present environment parameters. However if we consider the fact, most of the environmental parameters are depended on each other and previous environmental parameters we can directly correlate crop yield with prior seasons environment. This is our research's basis on which we have constructed our system. We will prove the hypothesis in our research.

3.2 Data Attributes

Six components of regional weather data of Bangladesh is archived and publicly available. They are temperature, rainfall, humidity, sunshine, cloud coverage and wind speed as monthly average of last 60 years. We have collected this data from

Bangladesh Agricultural Research Council (BARC) [6]. Data for 6 crops were publicly archived by Bangladesh bureau of Statistics (BBS) [7]. Those are Aus Rice, Amon Rice, Boro Rice, Jute, Potato, and Wheat, having 2313 entries of last 10 to 8 years. All the data attributes are described briefly below:

Temparature is an attribute of weather. It is a vital factor for the production of crops. Its meta attributes are shown in Fig. 1. During the model generation we have considered past 3 months weather data from the plantation month of specific crops. So in the data preprocessing stage 3 relevant consecutive months' attributes has been selected. **Rainfall** is another vital attribute for production of crops. Rainfall and all other environmental data has same table structure as temperature but the rainfall measurements are taken in millimeter. **Humidity** is the next attribute we have considered. This is the parameter of water vapor in the atmosphere. This is also an important factor for production. Here all the measurement is done in percentile. **Sunshine** is an attribute of weather. Sunshine plays a significant role for the production of crops. All the measurement of here has been done as sunshine hours. **Cloud Coverage** is one of the environmental parameters. It does not directly effect crop yield but it has indirect impact according to our in Hypothesis Sect. 3.1. The measurement has been done in Oktas. **Wind Speed** is one other parameter of weather. Like rest of the environmental data it has a similar table structure. All the measurement of temperature has been done in Knot.

Fig. 1. Temperature meta data on the database.

Crop Area is one of the crop attributes. Crop data for all 6 crops has similar table structure. Crop table has several attributes. Among them there are few that represent the area of crop cultivation lands but in different units (i.e. Acres, Hectors etc.). We have considered only one attribute, Area Hectors (Ahectors). One of the important issues while predicting the production is the area of the land which is used for production. This represents the total area under a certain region. Figure 2 is the data table for the crop Jute. Table structure for the rest of the crops is similar. **Crop Production** is our primary attribute. Our research goal is to predict crop production for unseen data. This represents the production of previous years on different regions. In Fig. 2 we can see that there are few attributes that represent production but in different units (i.e. Ton, Maunds), like crop area. Among them we have selected production in Metric Ton (M.Ton). **Region** is basically one of the crucial attributes. Bangladesh has been divided into separate regions. All the calculation will be done based on this region as yield varies based on region. We have analyzed and produced model and data representation based on separate regions. **Production Year** is the year of production for a crop. Here this attribute represents different years. For the research we have taken production year from 2007 to year 2013.

Fig. 2. Jute data table in the database.

3.3 Training the Model

One of the main objectives of this research is to construct a learning model for crop yield. We have constructed the system to predict production in terms of numerical data (M.Ton). We have applied Linear Regression and Neural Networks to train our the model. Finally we have compared methods in terms of error rate and prediction accuracy. As both of those models operate on numerical data only we cannot consider regions, which is a categorical data. But we cannot ignore the region while predicting crop yield. To resolve this we have designed the system to generate smaller models for all regions.

Linear Regression(LR) is a mathematical term that is used in statistical measurement to determine the relationship between a dependent variable and to its corresponding one or more independent variables. In data-mining linear regression is a numerical procedure to forecast the outcome of some dependent variable. **Neural Network**(NN) in Computer Science, is a computational model that utilizes the notion of back propagation to design a model. The number of hidden layers by default depends the number of attributes and also the number of classes for the given data set. If number of hidden layer is μ, number of attributes is α and number of classes is β then, the number of hidden layer in a neural network by will be:

$$\mu = \left[\frac{\alpha + \beta}{2}\right] + 1 \tag{1}$$

Machine Learning algorithm has been used improve the performance of the models. We have used Bootstrap Aggregating (Bagging) in our research. It is a machine learning algorithm which learns by splitting the training set. Bagging trains n number of models on the basis of new n numbers of training set. The algorithm will take the average and produce the final result. In our system, the value of n is 5 and data has split ratio 9:1.

4 Methodologies

We have used the concepts explained in the theory section to construct our system. Initially data is loaded from the database and preprocessed all the separate data tables to create a single table. Afterwards data is passed through a loop and splitted with respect to regions. This smaller dataset then is passed through Learning and bagging algorithms to train the model. Performance testing and prediction of unseen data is done in the loop. Finally all the resultant data are put together at the end of the loop and returned as the final result. The system has been built using RapidMiner Studio. Figure 3 shows major components of the main process of our system. All the major components of the system and its working will be explained in this section.

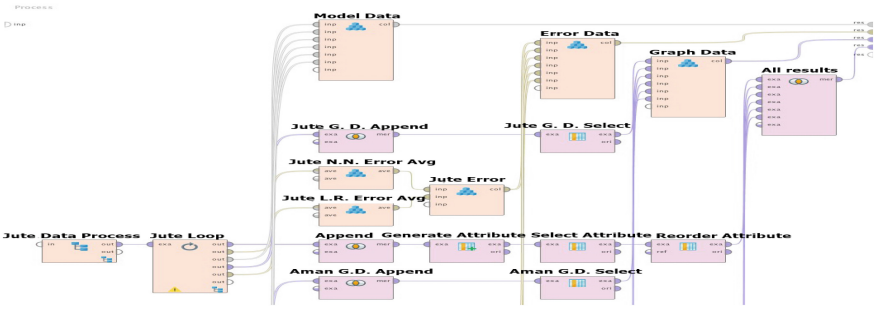

Fig. 3. Partial Design of the System.

4.1 Data Preprocessing

First step of the system is the data preprocessing. In the Fig. 3 the left most block named as Jute Data process is the starting point of the system. This block is a sub process which preprocesses the data. Figure 4 shows major components of this process. This block preprocesses jute data only. At the beginning all separate data tables are loaded and select operators are used to select the proper weather attributes. In case of Jute, it is planted on the month of March so the region, year and January to March weather data attributes are selected. Then those weather attributes are renamed and given a generic name and all weather tables are joined together. Finally the crop production table is joined with the previously joined table and Set Role operator has been used to identify which data attribute we want to predict.

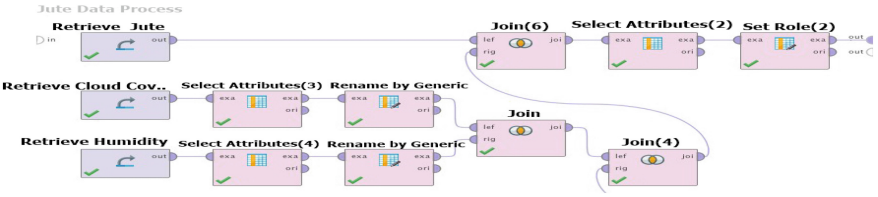

Fig. 4. Components of Data preprocessing sub process.

Figure 5 shows the final joined table for jute after preprocessing the data. In this sub process loading and joining of all the weather data has not been shown as it is a repeated work. Rest of the crops has similar preprocessing blocks that are not shown in Fig. 3. Those blocks also work in the same way but they operate on relevant data tables and attributes.

M.Ton	Year	Region	AreaH	cloud_m1	cloud_m2	..	..	..	humid_m3

Fig. 5. Preprocessed Jute data table.

4.2 Looped Subprocess

After preprocessing, the data is fed in a looped sub process. As explained in the Training Model Sect. 3.3. In the Fig. 3 we can see this loop block for jute named Jute Loop just after the Jute Data sub process. This loop iterated over all unique values of the Region values that is with all the region names. Figure 6 shows the loop process flow for jute preprocessed data table.

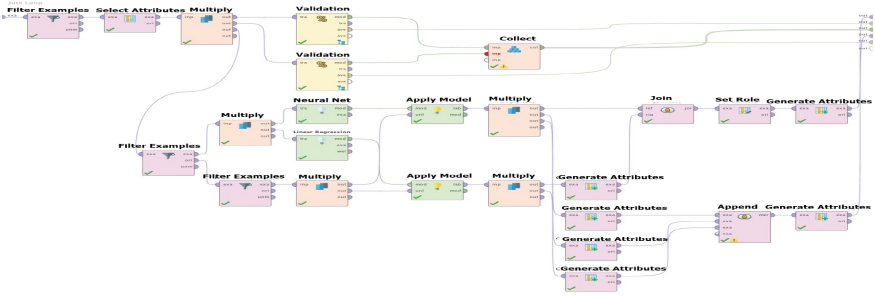

Fig. 6. Loop process for Jute.

For each iteration of the loop the name of a region is made available as a macro. Inside Loop process data is cloned in two data sets to perform testing and performance generation and Predict an unseen data in this case 2009 production of each crops. At first preprocessed data is split regionally and categorical attribute will be removed from the sub-tables because one table corresponds to only one region and as the methods we are using to train the model accepts numerical data only. After that sub-tables are cloned into two sets. One set is used to train two separate model using LR and NN. After training the model crop yield is predicted. Figure 7 shows the NN for Jute. NN model has one hidden layer. Both of the model are wrapped inside Bagging block which applies Bagging Algorithm. Another one of the cloned data sets are feed in the testing branch of Fig. 6. Tenfold method is used to measure the performance of the model. Finally the performance result of both algorithms are grouped in and collected in a data collection. At the end of the loop the categorical regional attribute is regenerated and added to the resulting table.

4.3 Data Collection and Produce Result

From Fig. 3 we can see several data collector is used to collect the results. During this step resulting data is processed so that we can plot and compare the results.

Fig. 7. Neural Network model for "Jute".

5 Results and Data Representations

In this research we have worked with six crops and applied our model on all six crops using both methods. In this section results and findings of our research has been described.

5.1 Testing Results

We have applied Root Mean Squared Error (RMSE) method using split testing algorithm for error calculation. Error rate of all crops using both methods has been listed and compared in the following table (Table 1).

5.2 Tabular Prediction of Result by Production on 2009

To test the result of our system we removed the data of 2009 while training the model and the predicted the result. Figure 8 shows the output of few crops as tabular format for both prediction models.

Table 1. Result Error

Crop	*Neural Net RMSE*	*Linear Regression RMSE*
Jute	1.943 +/- 0.00	0.992 +/- 0.00
Aman	0.248 +/- 0.078	0.334 +/- 0.202
Aus	0.106 +/- 0.00	0.173 +/- 0.00
Boro	0.236 +/- 0.00	0.244 +/- 0.00
Potato	8.136 +/- 0.00	2.983 +/- 0.00
Wheat	0.363 +/- 0.00	0.318 +/- 0.00

M.Ton	Neural Net	Linear Regr...	Crop Name	Region	Year
3.309	3.495	3.479	Boro Rice	Bhola	2009
4.009	4.153	3.991	Boro Rice	Jessore	2009
2.326	2.016	2.026	Boro Rice	Patuakhali	2009
3.987	3.986	3.922	Boro Rice	Bogra	2009
3.777	3.895	3.893	Boro Rice	Dinajpur	2009
13.172	10.840	12.330	Potato	Cox's Bazar	2009
20.392	16.583	14.369	Potato	Comilla	2009
19.245	19.653	16.163	Potato	Dhaka	2009
17.942	17.687	17.347	Potato	Barisal	2009
11.120	18.618	19.831	Potato	Bhola	2009
0	13.687	13.450	Potato	Jessore	2009
0	19.854	31.455	Potato	Patuakhali	2009
11.673	15.194	14.822	Potato	Bogra	2009
11.183	12.847	15.649	Potato	Dinajpur	2009

Fig. 8. Jute prediction result for year 2009.

5.3 Bar Charts of Results by Production on 2009

We have represented all results of all region as bar chart for year 2009 in Figs. 9, 10, 11, 12, 13 and 14.

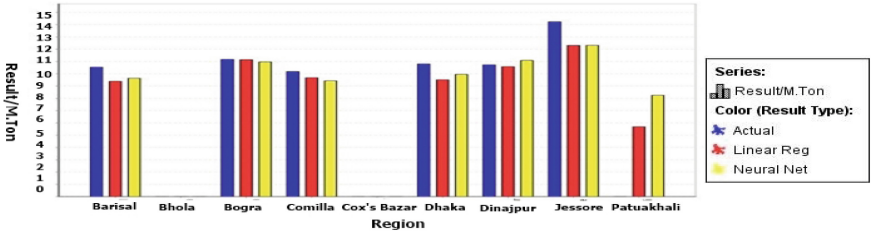

Fig. 9. Jute Prediction for 2009.

5.4 Research Finding

From all three research results, RMSE, Tabular result and Graphical we can see that crop production can be predicted considering prior environmental attributes with good accuracy. RMSE of all crops is relatively low and hence this satisfies our hypothesis. Furthermore from the comparison we can see that NN provides slightly better results. However NN takes more time to train the model.

6 Future Work

In this research we have considered 6 attributes to predict the yield of a crop. However correlation of environment attributes has little dependency with the

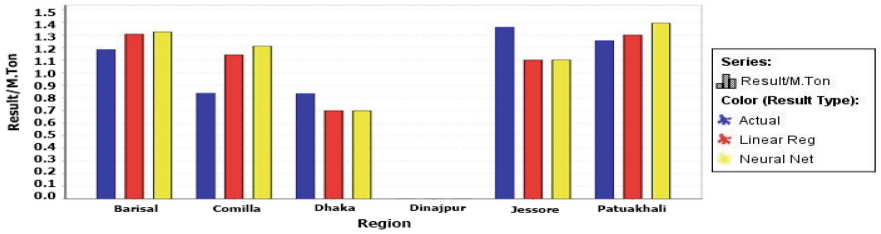

Fig. 10. Aus Prediction for 2009.

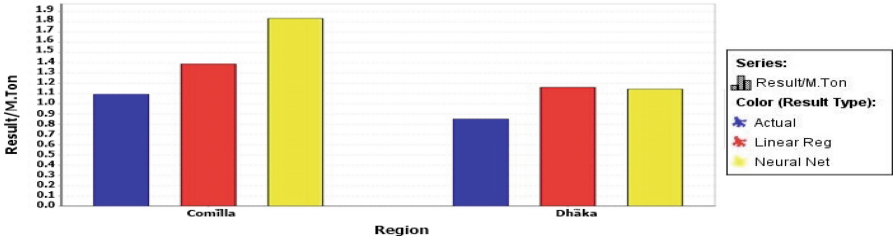

Fig. 11. Amon Prediction for year 2009.

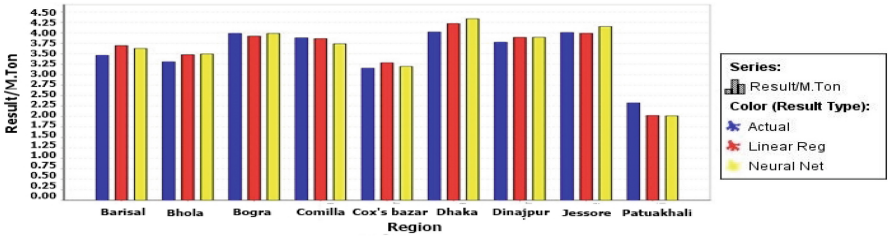

Fig. 12. Boro Prediction for year 2009.

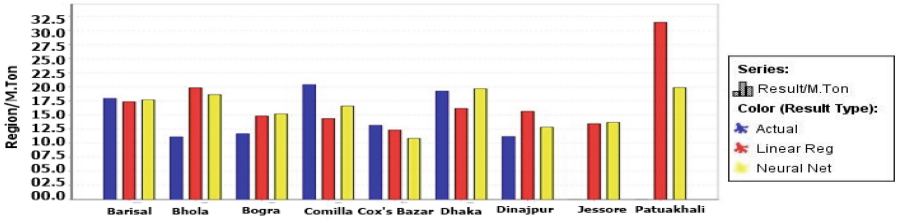

Fig. 13. Potato Prediction for year 2009.

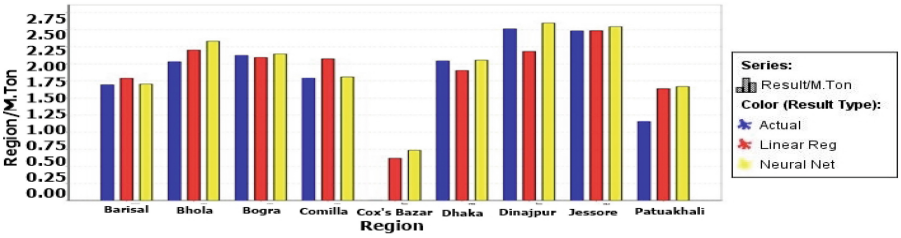

Fig. 14. Wheat prediction for year 2009.

attributes of previous months. We plan to collect more environment attributes and integrate in the system. Currently we are not considering already identified factors for predicting crop production that are used by farmers and environmentalist. We plan to include that knowledge in the system for better prediction.

Acknowledgments. We would not complete our research paper without having the support of the organizations named Bangladesh Agricultural Research Council and Bangladesh Bureau of Statistics. We would like to extend our sincere gratitude to those organizations.

References

1. Rahman, M., Haque, N., Rahman, R.: Application of Data Mining Tools for Rice Yield Prediction on Clustered Regions of Bangladesh. In: 17th IEEE International Conference on Computer and Information Technology (ICCIT), pp. 8–13. Daffodil International University, Dhaka, Bangladesh (2016)
2. Bhargavi, P., Jyothi, S.: Applying Naive Bayes Data Mining technique for classification of agricultural land soils. In: International Journal of Computer Science and Network Security, pp. 117–122 (2009)
3. Ramesh, D., Vardhan, B.: Data mining techniques and applications to agricultural yield data. Int. J. Adv. Res. Comput. Commun. Eng. **2**, 3477–3480 (2013)
4. Sha, Z., Zhang, M.: Development of web-based decision support system for field-based crop management. In: Geographic Information Systems, pp. 1–4 (2007)
5. Mateo, M., Leung, C.: Design and development of a prototype system for detecting abnormal weather observations. In: Proceedings of the 2008 C3S2E conference on - C3S2E 2008 (2008)
6. Bangladesh Agricultural Research Council (BARC)-Government of the People's Republic of Bangladesh, Barc.gov.bd (2016). http://barc.gov.bd/. Accessed 01 May 2016
7. Bangladesh bureau of Statistics (BBS). http://www.bbs.gov.bd/. Accessed 01 May 2016