Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh

A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman
Department of Electrical and Computer Engineering, North South University
Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh.
ahnafshakil@gmail.com, navid3893@gmail.com, mnhr7494@gmail.com, mtkabir285@gmail.com, gamer.eyes01603@gmail.com, hridoyrahman611@gmail.com, rashedur.rahman@northsouth.edu

***Abstract*—Agricultural crop production depends on various factors such as biology, climate, economy and geography. Several factors have different impacts on agriculture, which can be quantified using appropriate statistical methodologies. Applying such methodologies and techniques on historical yield of crops, it is possible to obtain information or knowledge which can be helpful to farmers and government organizations for making better decisions and policies which lead to increased production. In this paper, our focus is on application of data mining techniques to extract knowledge from the agricultural data to estimate crop yield for major cereal crops in major districts of Bangladesh.**

Keywords—data mining; crop analysis; yield prediction; clustering; K-means; K-NN; linear regression; neural net.

I. INTRODUCTION

Crop yield prediction is an important area of research which helps in ensuring food security all around the world. Bangladesh has rich soil which is very ideal for cultivation and it is one of the top rice producing countries in the world. In 2012 it produced a total of 33,889,632 metric tons of rice and it proved to be the 6th highest crop producing country in the world [9]. In order to take full advantage of the soil and sub-tropical climate of Bangladesh, farmers need to know exactly when to plant crop seeds. The entire economy also depends on the produce from harvesting annually.

Different districts in Bangladesh have varying climates and so it is very important to consider environmental factors of these separate areas. This will help to choose the best districts for cultivation of different type of crops. Rainfall also varies from district to district and this has a huge impact on farming because while too little or too much rain can kill crops, the proper amount of rain leads to an ideal crop yield. With rainfall comes humidity and since rainfall varies from district to district so does humidity. Humidity causes changes in the level of water that can be absorbed by atmosphere which can cause crops to remain too wet or too dry and so to get proper yield, a district with an ideal average annual rainfall and humidity is required.

Finally, the most important part of farming has to be considered, pesticides. Without pesticides crops would die significantly more due to insects and other pests leading to a sudden drop in yield. Too much pesticides may affect the crop on its own while too little may not get rid of pests. So, the amount of pesticides required by crops is a very important parameter.

In our research, we have considered the effects of environmental(weather), biotic(pH, soil salinity) and area of production as factors towards crop production in Bangladesh. Taking these factors into consideration as datasets for various districts, we applied clustering techniques to divide regions; and then we apply suitable classification techniques to obtain crop yield predictions.

II. RELATED WORK

Ramesh and Vardhan [1] deal with the challenge of predicting the yield of various crops. One approach to this problem is to employ data mining techniques. In this paper, different types of data mining methods were applied and then evaluated on the datasets we prepared.

In [2], Diepeveen and Armstrong discuss about various crop related data that is supplied to farmers to make better decisions to enhance yield and profits. While this may give the advantage of a particular crop species over others, the data is generalized and may not apply to others. There are data mining application that can process the data and improve the quality and reliability of this dataset for different farming situations. The challenge is identifying key attributes that affect crop yield, such as geographic location, soil type, seasonal conditions, nutrition, grain yield and quality, sowing and harvest data and tolerance to environmental stress. In this paper data mining techniques were used to help growers find the combination of traits required to identify high performance species. Several techniques were used over different geographical locations.

In [6] Murynin et al. study the dependency between the prediction and the accuracy of the forecast. The linear model is selected as a basic approach of yield prediction. Then, the model is extended with non-linear attributes in order to improve the accuracy of the prediction. The extensions take

SNPD 2015, June 1-3 2015, Takamatsu, Japan

into consideration long-term technological advances in agricultural productivity as well as regional variations in yields. The accuracy of the model has been estimated based on the time period between the moment of the forecast formation and the time of harvest.

III. MOTIVATION

Bangladesh has a sub-tropical climate which is suitable for rice cultivation. In 2012, Bangladesh yielded a total of 33,889,632 metric tons of rice. The market value of the rice in international market was $8,649,167,000. Bangladesh was 6th in 2012 among the rice producing countries [3].

A farmer must have a good understanding of the soil type, the biotic factors governing the soil and also a thorough knowledge about the traditional agricultural practices to gain maximum crop yield. Such practice may include harrowing and plowing using inputs such as fertilizers, insecticides and herbicides [6].

Bangladesh has a primarily agrarian economy. Most Bangladeshis earn their living from agriculture. Agriculture is the single largest contributing sector of the economy since it comprises about 30% of the country's GDP and employs around 60% of the total labor force. The performance of this sector has a highly significant impact on major macroeconomic objectives like employment generation, poverty alleviation, human resources development and food security. Although rice and jute are the major crops, maize and vegetables are of greater importance. Due to the spread of irrigation networks, some wheat producers have switched to cultivation of maize which is used mostly as poultry feed. Tea is grown in the northeast. Bangladesh's fertile soil and normal ample water supply is suitable for rice to be grown and harvested three times a year in many regions [5].

IV. DATA SET

The dataset used in this research has been collected from BARI (Bangladesh Agricultural Research Institute).

All the data were in pdf format which were converted to rtf format using miscellaneous tools and tricks. A lot of pre-processing was required to handle missing values, noise and outliers.

From the dataset, we have preprocessed and selected only the attributes which are important for our research: rainfall, maximum and minimum temperature, humidity, irrigated area for all districts; and cultivated area for every crop considered according to the districts.

One further environmental attribute: sunshine and two further biotic attributes – soil salinity and soil pH were considered for our research. These data were collected from the Bangladesh Agricultural Research Council (BARC) website [9].

After the necessary formatting and preprocessing of the datasets, the finalized version of our data contained a total of 15 districts for the time periods of 2009-10 and 2010-11.

The crop yields were selected for the following crops which have been considered for our project:

- Rice- AMON
- Rice- AUS
- Rice- BORO
- Potato
- Wheat

V. INPUT VARIABLES

From the vast initial dataset, we selected a limited number of important input variables which have the highest contribution to agricultural produce. All the inputs were considered for the two-year periods of 2009-10 and 2010-11.

a) The environmental variables:

i) Rainfall: The average yearly rainfall was considered by calculating average from the monthly rainfall (mm) of each district. Usually, the year that contains the highest average rainfall should provide for maximum crop yield in that year.
ii) Humidity: Similar to the way we collected the rainfall data, we also calculated and obtained the average yearly humidity for each district in percentage.
iii) Max Temperature: variation in temperature through the year puts a great impact in that year's crop production. Hence we consider both the maximum as well as the minimum temperature in our research.
iv) Min Temperature: The average yearly minimum temperature (considered in Celsius).
v) Average Sunshine: The amount of sunshine received on areas each year greatly effects the production of green crops as it directly affects the photo-synthesis process in plants. This attribute was considered in hours as a yearly average for each district.

b) The biotic input attributes:

i) Max pH: Maximum pH of a district's soil. pH is a scale attribute for farmers to keep track for how acidic the soil is. This scaled is defined by a value of 7, where soil pH above 7 meaning alkaline and below 7 meaning acidic. Crop production is highly affected by the variations of pH in soil.
ii) Min pH: Minimum pH of a district's soil.
iii) Soil Salinity: Taken as MMHOS/cm, the ranges were (<2), (2-4), (4-8) and (8-15). Soil salinity defines the amount or content of salt in soil. Soil salt content is increased by the process of salinization. Too high soil salinity can cause a detrimental effect towards crop production and yield. We calculated total areas (in hectares) under different salinity ranges for each of the 15 districts.

c) Area central input attributes:

i) Irrigated Area: The amount of production of a crop will depend on the actual area of land that has been irrigated throughout the year. Hence irrigated area was considered in our research for the selected districts (in hectares).

ii) Cultivated Area: The area that has been used to cultivate each crop also regulates the amount of production of the crop. Areas were taken in hectare unit.

VI. METHODOLOGY

The method of our research is initially divided into two major parts: Clustering and Classification.

A. Clustering of the selected districts:

In our research, we have considered a total of 15 districts of Bangladesh. In order to group the districts into distinct clusters, the assumption that we had to use was that the districts containing the similar values of relevant attributes should belong to the same cluster. According to this assumption, we categorized our selected attributes for the consideration of clustering the districts as follows:

1) Cluster Type-1 is based on the following attributes: Rainfall, minimum temperature, maximum temperature, humidity and sunshine. These are the environmental or climatic attributes considered for our research. The degree of similarity of the collection of these attributes should indicated distinct clusters for the selected districts.
2) Cluster Type-2 is based on the following attributes: soil pH and soil salinity. As discussed earlier, these biotic factors contribute largely towards the prediction of the crops. Similarity of the values of these attributes should also indicate separate clusters.
3) Cluster Type-3 is based on irrigated area. Clustering is based on the area attributes for each district was considered because we can obtain the separate clusters based on distinct ranges of areas that were irrigated for each district.
4) Cluster Type-4 is based on the individual crop yields of Amon, Aus, Boro, potato and wheat. This type of clustering was considered in order to classify the districts into separate clusters with similar crop yields- and after analysis of the results, to see whether they exhibit a pattern related to effects from the selected attributes.

K-Means Clustering:

The K-means clustering algorithm is used to produce non-hierarchical groups of similar points in the data based on the centroid. For our research, *k*-means clustering was used upon the selected districts according to the categorized types mentioned previously. Using our finalised dataset of 2009-10, we implemented *k*-means clustering technique through RapidMiner Studio software. Clustering results were separately written to Excel files for each cluster type (1 to 4) for the convenience of result showcasing and analysis.

B. Prediction of crop yields using classification techniques:

In our research, we determined prediction results for yields of selected crops for the selected districts in Bangladesh. The predictions results were obtained according to the selected input attributes using appropriate classification and regression models in RapidMiner.

For the purpose of use in learning models, two time periods of our dataset were considered as follows:

- 2009-10: Training Dataset
- 2010-11: Testing Dataset

The following classification/regression models were used to obtain the crop yield prediction results:

a) Linear Regression: It is a statistical measure that can be used to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables (regular attributes). If independent variable contains multiple input attributes like in our research (rainfall, sunshine hours, humidity, pH etc), then it is termed as multiple linear regressions. Linear regression provides a model for the relationship between a scalar variable and one or more explanatory variables. This is done by fitting a linear equation to the observed data [7].

b) *k*-NN: The k-nearest neighbour algorithm compares a given test example with training examples which are similar. Each example denotes a point in an n-dimensional space. Thus, all of the training examples are saved in an *n*-dimensional pattern space. K is a positive integer, usually small. For our purpose, the basic *k*-NN algorithm was applied. It first finds the *k* examples from the training set that are closest to the unknown example. Then it takes the most common occurring classification for the *k* examples [7].

c) Neural Net: An artificial neural network (ANN) is a mathematical model or computational model inspired by the structure and functional aspects of biological neural networks for instance in our brains. In most cases an ANN is an adaptive system that modifies its structure based on external or internal information that flows through the network during the learning phase. The basic neural network model consists of three layers: the input layer, the hidden layer and an output layer [7].

VII. RESULTS

A. Clustering Results:

In RapidMiner, X-Means operator was used to provide *k*-Means clustering according to the previously mentioned

clustering of 4 types depending on differently grouped attributes. We obtain the following results:

1) *Cluster Type-1:*
Based on the weather attributes.
2 clusters
Cluster_0: Chittagong, Rangamati, Sylhet
Cluster_1: Comilla, Dhaka, Faridpur, Mymensingh, Tangail, Barisal, Jessore, Khulna, Bogra, Dinajpur, Rajshahi, Rangpur

2) *Cluster Type-2:*
Based on soil-salinity and soil-pH.
2 clusters
Cluster_0: Chittagong
Cluster_1: Rangamati, Comilla, Sylhet, Dhaka, Faridpur, Mymensingh, Tangail, Barisal, Jessore, Khulna, Bogra, Dinajpur, Rajshahi, Rangpur

3) *Cluster Type-3:*
Based on irrigated area (in hectares).
3 clusters
Cluster_0: Chittagong, Rangamati, Comilla, Sylhet, Dhaka, Faridpur, Mymensingh, Tangail, Barisal, Khulna, Bogra
Cluster_1: Rajshahi
Cluster_2: Jessore, Dinajpur, Rangpur

4) *Cluster Type-4:*
Based on individual crop yields.
3 clusters
Cluster_0: Chittagong, Sylhet, Barisal, Khulna, Rajshahi
Cluster_1: Rangamati, Dhaka, Faridpur, Mymensingh, Tangail
Cluster_2: Comilla, Jessore, Bogra, Dinajpur, Rangpur

B. *Prediction Results:*

Results from performing the algorithms on the data set of different districts such as Chittagong (d1), Rangamati (d2), Comilla (d3), e.t.c is shown in Fig.1 – Fig.5 bellow.

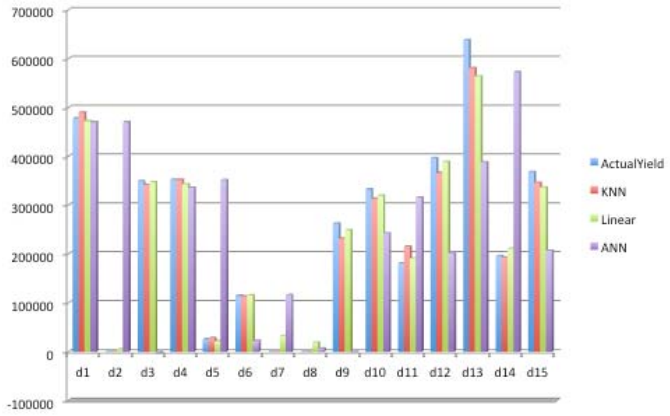

FIG. 1. Actual vs. Predicted yield of all models of Amon

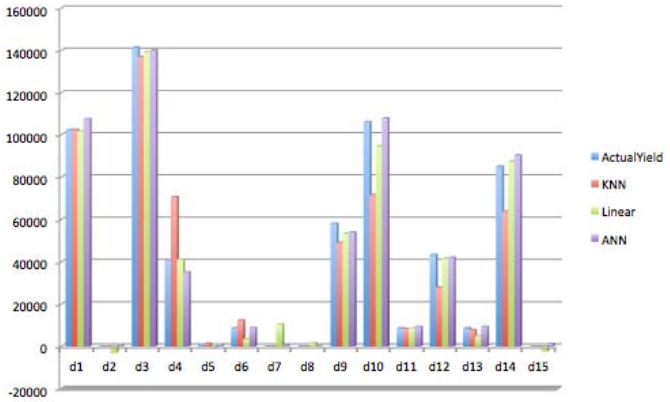

FIG. 2. Actual vs. Predicted yield of all models of Aus

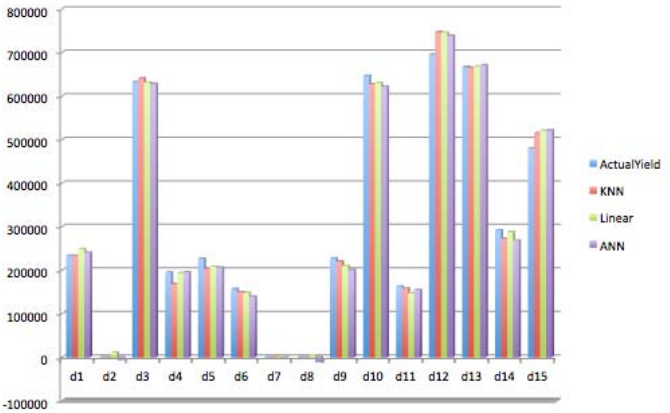

FIG. 3. Actual vs. Predicted yield of all models of Boro

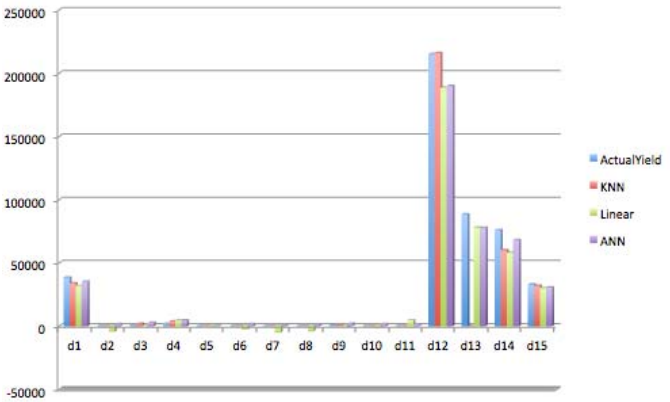

FIG. 4. Actual vs. Predicted yield of all models of Potato

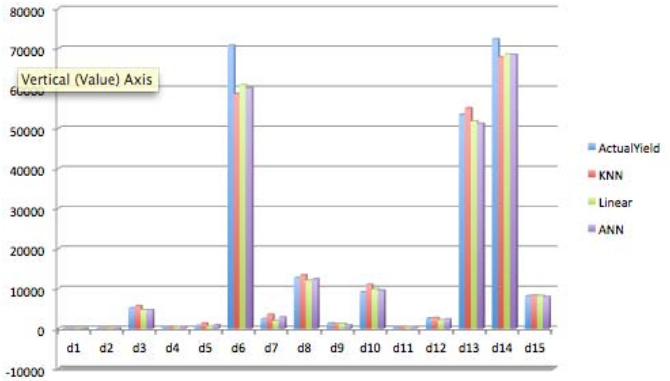

FIG. 5. Actual vs. Predicted yield of all models of Wheat

VIII. Results Analysis:

A. *Clustering result analysis:*

We applied clustering to find if any strong correlation exists between crop yield and different attributes (i.e. weather attributes, soil PH and soil salinity attributes and area cultivated attribute). Although there are some similarity between the clusters obtained using different attributes with the clusters obtained according to crop yields, we did not find any exact or strong correlation between crop yield with weather/ soil/ cultivate area attribute. This might be due to the fact that there were some missing values in our data sets and the size of our data set was comparatively smaller than required to find any strong correlation

B. *Analysis of prediction results from different models:*

From all the results obtained it is clearly seen that the accuracy lies within the range of 90 to 95 percent. Each of the techniques used gives a prediction with a slightly varying accuracy. However, due to the small training set, the prediction was not as accurate as expected and sometimes anomalies were experienced. For example, from FIG. 2 to FIG. 5, in some cases, if the actual yield was 0 (zero), our models sometimes erroneously predicted some nonzero value for the predicted yield.

If our training dataset were large enough (containing all the data about all 64 districts), avoiding this problem would've been possible.

C. *RMSE comparison:*

Root Mean Square Error (RMSE) is used to describe how well a machine learning algorithm performs on a certain data set.

	Linear	KNN	Neural
Amon	24270.2	22578.36	24791.96
Aus	4754.944	13873.16	2788.586
Boro	19653.76	20122.28	21128.44
Potato	9279.495	23264.65	7553.811
Wheat	2776.443	3414.207	2996.593

FIG. 6. RMSE of different models.

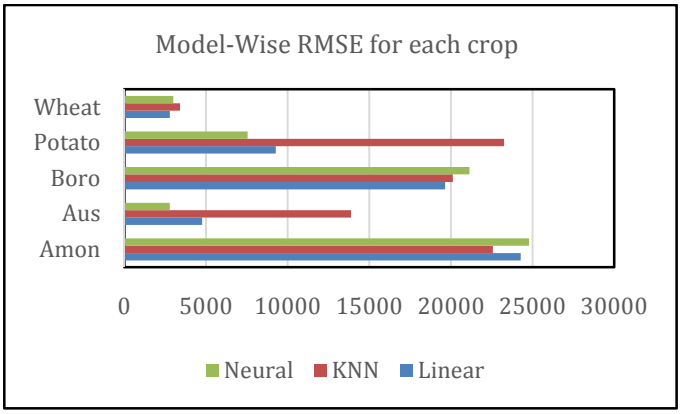

FIG. 7. Model-Wise Root Mean Squared Error for each crop

From the RMSE comparison it's clearly shown that different model provide the better result for the different crops. ANN provides better prediction for some of the crops, which have more missing values than others, for example wheat, potato and Aus. Linear regression provide better performance of predicting boro and amon.

IX. Recommendation System

After getting all the result tables and charts, we have written a program which takes into account the necessary tables from our results to post process the data and give the best three possible crops in order of preference to choose from for farming across all the major agricultural districts. If there are not any feasible choices the program simply outputs 'NONE'. These recommendations are based on a combination of annual yield of that crop species per hectare area of a district and its net worth.

X. Future Works

In this paper we take into account 5 environmental variables, 3 biotic variables and 2 area related variables to determine the crop yield in different districts. In the near future, geospatial analysis will be added to our data processing model to improve accuracy and also implement a better accountability of geographical data. Furthermore, the recommendation system will be enhanced and incorporate the time between seeding and harvesting for different crop species across

different districts. This will help make a system that will be used by the government or some other authoritative body that will notify farmers/growers when to plant appropriate seed in appropriate area at what time of year for the highest estimated yield and profitability based on the processing of past data.

XI. CONCLUSION

In our research we have found that the accurate prediction of different species of crop yields across several districts could help a lot of farmers and others alike. A farmer could plant different crops in different districts based on simple predictions made by this research and if that does take into effect, each and every farmer would get a chance at increasing their profits and increasing the country's overall produce. Also, using a better dataset for this research will lead to even better predictions and recommendations as the recommendation engine is basing its decision based on the predictions.

REFERENCES

[1] D Ramesh , B Vishnu Vardhan. "Data Mining Techniques and Applications to Agricultural Yield Data". International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013,pp.3477-3480.

[2] D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining" World Conference on Agriculture, Information and IT, 2008.

[3] Mohammad Motiur Rahman, Naheena Haq and Rashedur M Rahman "Comparative Study of Forecasting Models on Clustered Region of Bangladesh to Predict Rice Yield", 17th. IEEE International Conference on Computer and Information Technology (ICCIT), Dhaka, 2014.

[4] http://books.irri.org/0471097608_content.pdf

[5] http://www.assignmentpoint.com/science/zoology/agriculture-sector-of-bangladesh.html

[6] Alexander Murynin, Konstantin Gorokhovskiy and Vladimir Ignatie "Efficiency of crop yield forecasting depending on the moment of prediction based on large remote sensing data set" retrieved from http://worldcomp-proceedings.com/proc/p2013/DMI8036.pdf

[7] Ye, Nong; Data Mining: Theories, Algorithms, and Examples, CRC Press, 2013.

[8] http://docs.rapidminer.com/studio/

[9] http://www.barcapps.gov.bd/dbs/index.php

[10] http://www.faostat.fao.org/site/339/default.aspx