Automated Farming Prediction

Talha Siddique, Dipro Barua, Zannatul Ferdous, Amitabha Chakrabarty
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka-1212, Bangladesh
talhasiddique265@gmail.com,dipro.cse029@gmail.com,zferdous83@gmail.com, amitabha@bracu.ac.bd

***Abstract*—Farming is a predominantly manual process. The incorporation of any form of automation through the means of machine learning algorithms is still in incipient stage. This paper aims to introduce a fundamental approach to inaugurate the use of machine learning systems in the farming process. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop for a particular land. Here, the best crop signifies the crop which had the most increase in terms of yield per unit area compared to previous years. This will ensure proper crop allocation throughout the country since optimal production for each crop will be secured. It will also increase the farmer's revenue margin. The study focuses on six major crops of Bangladesh which are Aman rice, Aus rice, Boro rice, Potato, Wheat and Jute. The algorithms that were used are Multiple Linear Regression (MLR) and K-Nearest Neighbor Regression (KNNR). Multiple Linear Regression (MLR) gave the most accurate results during the analysis and was incorporated into an android application. The android application system is also able to prepare a schedule of the complete farming process for a chosen crop, e.g. the correct time to apply fertilizers, irrigation, etc.**

Keywords*—Multiple Linear Regression (MLR) Analysis; K-Nearest Neighbor Regression (KNNR); prediction; android application; fertilizer suggestion; dependent variable; independent variables

I. Introduction

In developed countries, farming is done through very advanced processes. On the other hand, our country is making very little headway against the current technological advancement. Still, the crop cultivation is a long manual process in our country. The automated process of farming is the beginning of a new era in Bangladesh that will be suitable for the farmers who seek experts to take suggestion about the appropriate crop on specific location of their land and don't want to forget any step of the cultivation throughout the process. Although, the opinion from experts is the most convenient way, this application is designed to give accurate solution in fastest manner possible. This research's main objective is to bring farming process a step closer to the digital platform. It is an android application [7] in which a decision-making algorithm [5] can predict which crop will be cost effective for that particular land. It will predict in terms of percentage change in yield per unit area, based on previously collected yield rates from the locations. Then the application will give suggestion regarding the cultivation process of that crop. This will ensure proper crop allocation throughout the country. Because, in this agro based country, three fifths of Bangladeshis are employed [6] in the agriculture sector. So this will increase our total growth of crop production. It will also contribute to increase the farmer's income by suggesting the accurate crop according to the location which is very effective to inspire future generation on investing in farming sector. The application will focus on six major crops [3] of Bangladesh. Our goal is to make a project prototype that will be easy to operate even for amateurs in technology usage. To make sure that a cost-effective farming solution is given to the farmer, taking cultivation location [3] (region/district) as the input; our app suggests the most effective crop type to be cultivated. The decision as to which crop type is the most cost-effective, is made through a decision-making algorithm [5]. The decision-making algorithm which is incorporated is linear regression [5]. In agro based country like Bangladesh, agriculture is one of the most important sources of income [6]. For making it more user friendly, this app will be available in Bengali. This application's GUI will be easily manageable by our farmers. The total system is focused on the climate and geographical condition of Bangladesh. To describe the functionality of this system; at first, farmer gives the perimeter of land in input area and the district (from dropdown menu) for which he wants the suggestion of best crop. Then the best crop's name will be shown in the screen. If the suggested crop is chosen, the entire procedure of cultivation will be shown to him. Then the notification of irrigation, fertilization, cutting crop and other important events will be shown up timely in a calendar form. If a symptom of any problem related to the lack of a fertilizer type is showing up, then farmer will check whether the fertilizer routine is maintained properly or not. Then app will show the correct amount of fertilizer intake. The crop zone is divided according to the division and districts. The data of crops of total seven divisions - Chittagong, Sylhet, Dhaka, Barisal, Khulna, Rajshahi and Rangpur along with three other large districts – Faridpur, Bogra and Comilla; will be stored in database system. The dataset consists of information on six major crops of Bangladesh, their yield rate, maximum temperature, minimum temperature, year range, region and rainfall. From these values, this algorithm gives a prediction result. The past fourteen years of Bangladesh have been considered for making this dataset, to ensure learning and training of the algorithm and increasing the accuracy rate of the prediction.

In this paper, Section II discusses the literature review. Section III explains our contribution. In Section IV experimental results and analysis are explained. In Section V the System Overview is discussed.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

II. Literature Review

Multiple Linear Regression has been used extensively in both agricultural farming and other areas. For example, Vinciya and A. Valarmathi performed an analysis to classify between organic forming, inorganic forming and real estate data to make a prediction of the crop yield based on the data set [9]. In this research they used Multiple Linear Regression (MLR) for the selected region [9]. In multiple linear regressions, there was an equation for prediction and the estimation of the residuals to find the difference between actual values and the fitted values. The use of the variance was to calculate the mean-squared error [9]. In addition to this work, to predict the crop yield and the best crop suggestion, Ashwani Kumar and her co-author used a new algorithm named "Agro Algorithm" in Hadoop platform and used Hadoop framework for handling large data [10]. They also considered the soil type and weather in their approach [10]. First, the data is accessed from the Hadoop distributed file system and the data is normalized [10]. Then the data was classified and from classification on the basis of disease, it gives the prediction about soil and crop [10].Furthermore, Snehal and Sandeep discussed about using Artificial Neural Network (ANN) for crop yield prediction [11]. In this case, they used feed forward back propagation neural network [11]. Where feed forward identify the pattern and back propagation is for comparison with the actual and gained output. They used the number of parameters like PH, Nitrogen, Temperature, and Rainfall for accurate prediction [11]. In the paper, DataMining in Agriculture on Crop Price Prediction: Techniques and Applications, the authors tried to discuss about the various data mining approach in agricultural field [12]. They discuss about the K-Means, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machines and the necessity of Price Prediction of crops according to the current market price policy [12]. In addition, in terms of android application developed for agricultural sector, the application developed by Santosh and Sudarshan shows the market and weather [13]. This approach also used android platform to sell farmer's product in global market and make larger profit [13]. The application also supports multiple languages. Koli and Jadhav conducted a research about agricultural decision support system in android platform [2]. This research is designed to find the maximum crop yield in minimum usage of cost [2]. This app gives prediction based on the water availability, average temperature, location of farm, average ph and so on [2]. To predict this they used Decision Support System (DSS) and AI [2]. Although, agriculture is the largest employment sector in Bangladesh, research of such nature has not been undertaken yet. This acted as the motivation behind our work.

III. Our Contribution

Two supervised machine learning algorithm has been applied here. One is multiple linear regression with the association of some independent variables i.e. rainfall and temperature of certain location and give prediction based on yield rate per unit area and another is KNN regression (K-Nearest Neighbor regression). The dataset obtained, was divided into two sets: learning set and testing set. These two algorithms are implemented and applied on the learning dataset and the value predicted by the algorithm was compared against the testing set in order to determine the accuracy of the model. The algorithms applied in the dataset and their respective pseudo codes are discussed here.

A. Algorithms

The system gives the prediction of the crop that had the most increase in output per unit area. Multiple Linear Regression algorithm and KNNR were both implemented separately and their results were compared to determine which one of the two outperformed the other in terms of accuracy. Multiple Linear Regression have been successfully implemented for such purposes in previous research works as well [9, 12]. The upcoming discussion is about these algorithms' background and implementation.

B. Multiple Liner Regression

- *Supervised Machine Learning:* These forms of algorithms consist of a target variable (or dependent variable) which is to be predicted [5] from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

- *Background Study:* As it has been said earlier, this algorithm is quite familiar in the field of prediction and classification based research. In the Exploiting Data Mining Technique for Rainfall Prediction research, MLR was selected [8].The general equation of MLR is as follows:

$$Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\cdots + \beta_n x_{in}$$

Where, i= 1, 2, 3…n ….. (1)

Y is the predicted outcome of this equation. The parameters, β_0, β_1, β_2 … β_n are the regression parameters. The formula of correlation is,

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} - \sqrt{n(\sum y^2) - (\sum y)^2}}$$

- *Implementation on dataset:* In this research QR Decomposition [4, 14] method has been used. By solving two matrices formed by dependent and independent variables, the desired outcome was obtained. (1) is expressed in matrix form below:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} 1 & x_{11} x_{12} & \cdots & x_{1k} \\ 1 & x_{21} x_{22} & \cdots & x_{2k} \\ \vdots & \vdots \quad \vdots & \ddots & \vdots \\ 1 & x_{n1} x_{n2} & \cdots & x_{nk} \end{bmatrix} B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Now this entire equation was solved using QR Decomposition. Since the implementation was done using java, the QR Decomposition class in the Java Matrix package was used to serve this purpose.

- *Pseudo code:*

```
Regression (double [] [] x, double [] y)

        Matrix x ← new Matrix(x)
```

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

```
// create matrix from vector
    Matrix y← new Matrix(y,N)
//find the least square solution
QRDecomposition qr← new QRDecomposition(x)
    beta ←qrsolve(y)
// Mean of y[] values
    For i=0 to N
        Sum += y[i]
        Mean ←$\frac{sum}{N}$
//Total variation to be accounted for
    For i=0 to N
    Dev ← y[i]-mean
Sst+= dev*dev
// variation accounted for
    Matrix residuals ←x.times(beta) . minus(y)
    SSE ←x.times(beta) . minus(y)
```

Here, in 2d matrix x, the independent variables are rainfall and temperature. The matrix y represents the dependent variables and is the yield per unit area. Then these matrices were solved using the QR Decomposition process and the total variation was also determined. The desired equation with the β variables was obtained. The rainfall and temperature values were then inserted, to give the predicted outcome.

C. KNN Regression:

In KNN regression, first Euclidean distances are considered and then the distance level is measured [1]. Based on the optimal number of K, it finds the nearest possible values. It calculates the inversed distance average with its neighbor [1]. In this research, firstly, Euclidean measures of the targeted year and training year's rainfall and average temperature had been measured. Then, using the measure value, matches were determined. The most appropriate yield match is selected for the result. The pseudo code has been discussed below:

- *Pseudo code:*

```
    // calculating the Euclidean Distance
    For i=0 to Yield.length
    D[i]= distance(rainfall, temp, rainfall[i], temp[i])
    End For
    //Finding the indices of 'k' no. mean Euclidean Distances
    FindMinIndex (k, d, minIndex)
//calculating the predicted yield
Yield ←calculateAverage (y,minIndex)
//Calculate percentage if error
Error ←percentageErrors (y1,yield)
```

In this pseudo code, firstly Euclidean distances are measured. Then it calculates the corresponding minimum yield rate by indices. Using these yield values, it calculates the average and then calculates an error percentage by using the actual yield per unit area from the test data set.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To determine which algorithm works better, both MLR and KNN have been applied to calculate the yield prediction. Ten regions of our country were chosen as the work domain. The regions which were considered are: Barisal, Bogra, Comilla, Dhaka, Dinajpur, Faridpur, Khulna, Rajshahi, Rangpur and Sylhet. Among these regions, the considered crops were: Aus, Aman, Boro, Wheat, Jute and Potato. The variables chosen for the calculation, were: crop yield rate per year, average rainfall per cropping season of a crop and average temperature per cropping season of a crop.

In this analysis, from year 2000 to 2011; data of 12 years are considered as learning data. And from year 2012 to 2014; data of 3 years are used for testing purpose.

The dataset we have used in this research is fully authentic. These data were scattered when we attempted to collect them. Some of the numbers were missing and so were some of the sources. The dataset from year 2008 to year 2014 can be found online, in the website of Bangladesh Bureau of Statistics (BBS) [3]. We had to collect data of the other years, from 2000 to 2007, from yearly books of agricultural statistics for our country, from different government organization's libraries. Data were all in hard copy form. We did a huge amount of data entry to convert them to our convenient form. And then the dataset was stored in a dedicated backend database.

These data were verified by experts from Department of Agricultural Extension (DAE) Bangladesh [16]. People from Agricultural Information Service (AIS) Bangladesh [17], helped us collecting these data.

In the graphs below, yield rate has been considered in X axis and crop type has been considered in Y axis. The results of 2013 and 2014 are shown in the graphs below (Fig. 1)

A. Khulna Region

For Khulna region, year 2013:

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

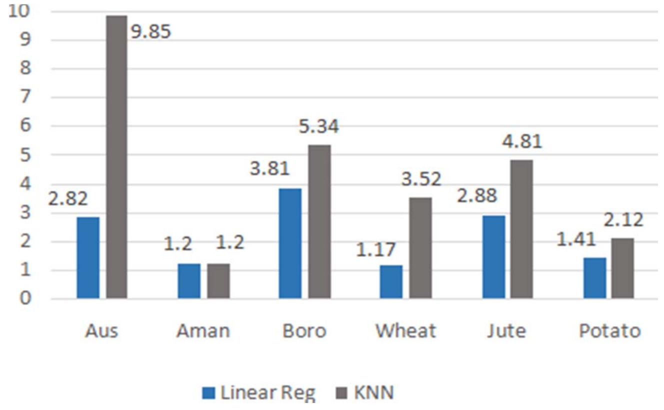

Fig. 1. Resulting percentage error for Khulna in 2013.

In this graph, it is clearly observable that for Aus, Boro, Wheat, Jute and Potato; MLR gives better prediction. For Aman rice, both KNN and MLR give similar prediction. MLR for Aus, gives error rate of a 2.82%, where KNN gives 9.85% error rate. In case of Aman, both give 1.2% error rate. MLR gives prediction for Boro with 3.81% error rate, while KNN gives 5.34% error. For Wheat, MLR gives 1.17% error in prediction which is quite satisfactory; on the other hand, KNN gives 3.52% error rate. In case of Jute, MLR gives result with 2.88% error rate and KNN gives 4.81% error rate. For Potato, MLR gives 1.41% error rate and KNN gives 2.12% error in prediction for yield.

For Khulna region, year 2014 (Fig. 2):

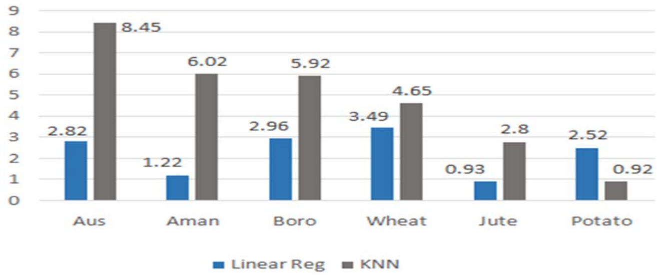

Fig. 2. Resulting percentage error for Khulna in 2014.

In year 2014, for all crops except Potato, MLR gives better prediction with less error rate. KNN gives better prediction for Potato which has 0.92% error rate.

B. Sylhet Region

For Sylhet region, year 2013 (Fig. 3):

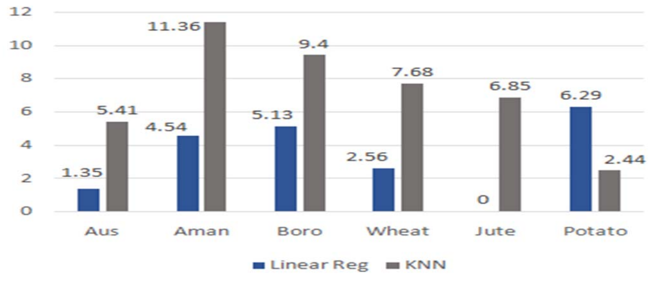

Fig. 3. Resulting percentage error for Sylhet in 2013.

In Sylhet region, for all crops except Potato, MLR gives better prediction. MLR accurately predicts for Jute with 0% error.

For Sylhet region, year 2014 (Fig. 4):

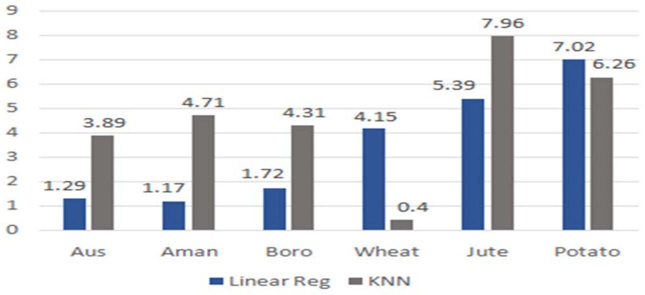

Fig. 4. Resulting percentage error for Sylhet in 2014.

Here, for all crops except Wheat and Potato, MLR gives better prediction. For Wheat and Potato, KNN gives better result with 0.40% and 6.26% error correspondingly.

C. Faridpur Region

For Faridpur region, year 2013 (Fig. 5):

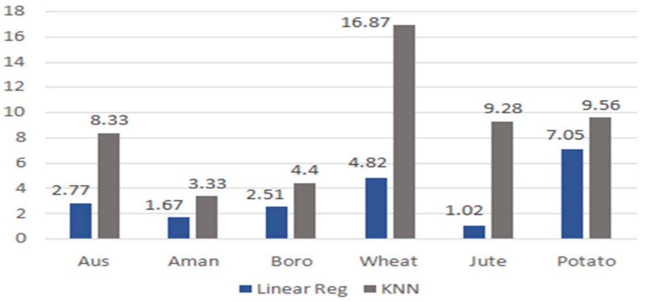

Fig. 5. Resulting percentage error for Faridpur in 2013.

Here, from this graph, it can be claimed that for all crops, MLR gives better prediction in Faridpur region, for the year 2013.

For Faridpur region, year 2014 (Fig. 6):

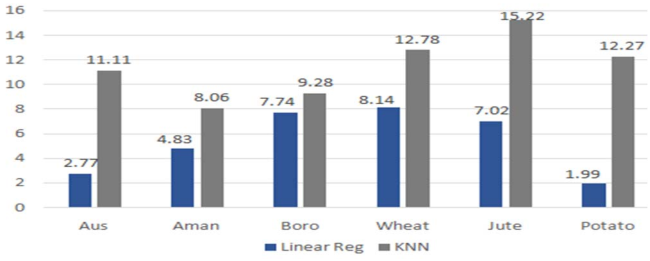

Fig. 6. Resulting percentage error for Faridpur in 2014.

In the year 2014, MLR gives better prediction with less error rate for all the crop types.

D. Rangpur Region

For Rangpur region, year 2013 (Fig. 7):

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

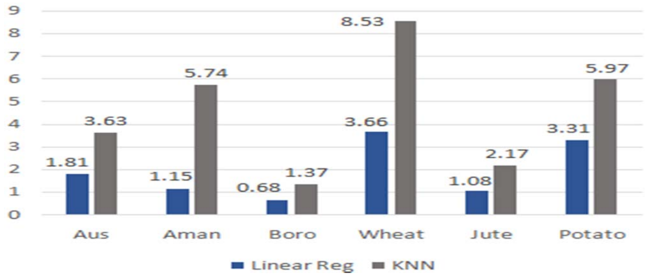

Fig. 7. Resulting percentage error for Rangpur in 2013.

From this graph, it is clearly observable that MLR gives better prediction for all crops in Rangpur region, for the year 2013. Out of the crop types, MLR predicts with below 1% error rate for Boro. Also, error rate in predicted result for Aus, Aman and Jute are almost fully negligible.

For Rangpur region, year 2014 (Fig. 8):

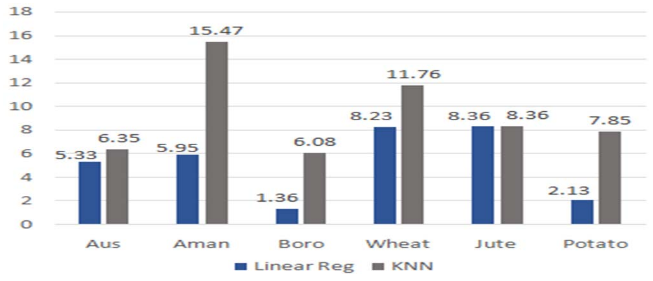

Fig. 8. Resulting percentage error for Rangpur in 2014.

In Rangpur region, 2014, for Aus rice, MLR barely wins against KNN where MLR gives 5.3% error and KNN gives 6.35% error rate. For Jute, they both predict with same error percentage – 8.36%. For rest of the crops, MLR gives way better prediction than KNN.

- *Regression Plot:*

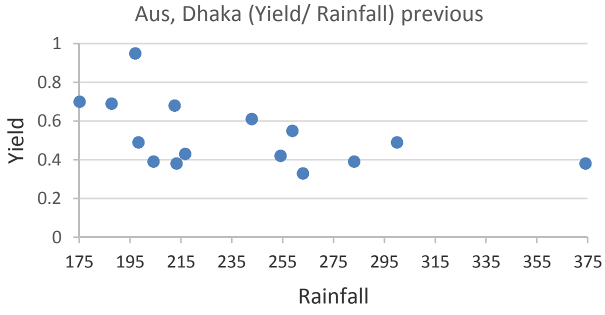

Fig. 9. Regression plot for Aus - raw data of 'Yield vs Rainfall'.

In the regression scatter plot above (Fig. 9), we can see different data points symbolizing 'Yield vs Rainfall' criteria, which was actually collected from BBS through AIS, were more scattered for using MLR.

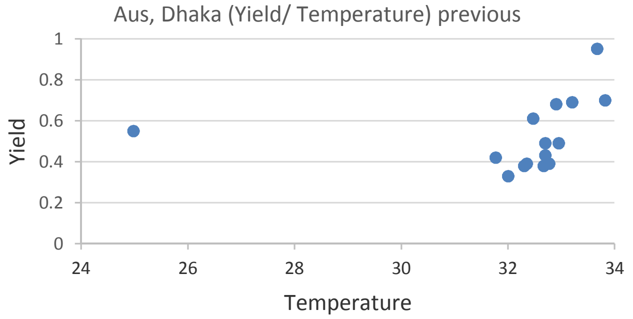

Fig. 10. Regression plot for Aus - raw data of 'Yield vs Temperature'.

Similarly, we can see that the data points of 'Yield vs Temperature' criteria, which was actually collected from BBS through AIS, were more scattered for using MLR (Fig. 10).

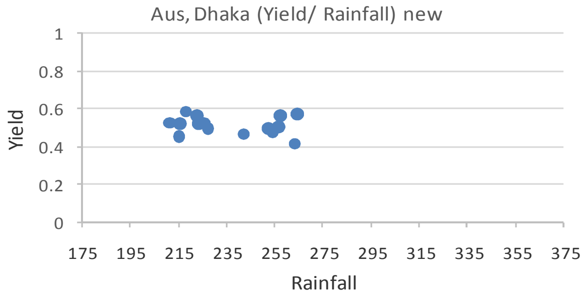

Fig. 11. Regression plot for Aus modified data for 'Yield vs Rainfall'.

After adjusting, by omitting the outliers (extreme points) and using synthetic data, the data points of 'Yield vs Rainfall' criteria become adjustable to draw a line on that scatter plot (Fig. 11).

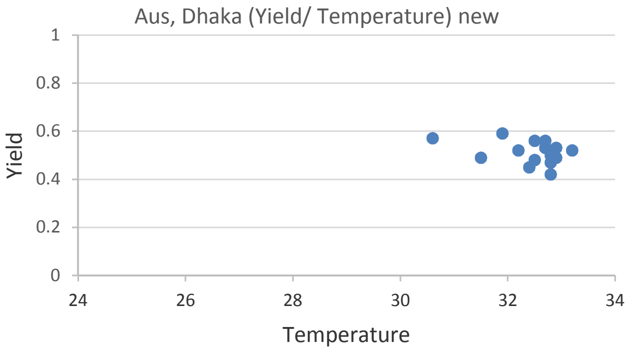

Fig. 12. Regression plot for Aus - modified data for 'Yield vs Temperature'.

Similarly, after adjusting, by omitting the outlier (extreme points) and using synthetic data, the data points of 'Yield vs Temperature' criteria become adjustable to draw a line on that scatter plot (Fig. 12).

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

V. System Overview

In the proposed system, primary user is the farmer or someone interested in farming. The user can sign up his/her credentials which will be the mobile number and a password. The user will then give land location and the date on which s/he plans to plant the crop, as input. The system will take into account, the farming season and determine which of the six crops can be cultivated during the season or time frame. Then the system will find the location of the given region and predict the percentage increase or decrease in yield per unit area of the determined crops for that particular year and region. The user can then select the predicted crop or any of the other suggested crops. Based on user's choice, the system will now suggest the appropriate resources needed and create the schedule of entire farming process.

In this system, there will be a local server in the android phone of the user. Server side from the administrator perspective – main server will do the calculations beforehand. And then evaluated results will be saved in the local server for different input sets. It will reduce the load of the main server dedicated for this system and by storing data in local database; the procedure will cause less internet cost.

A. The Entire System

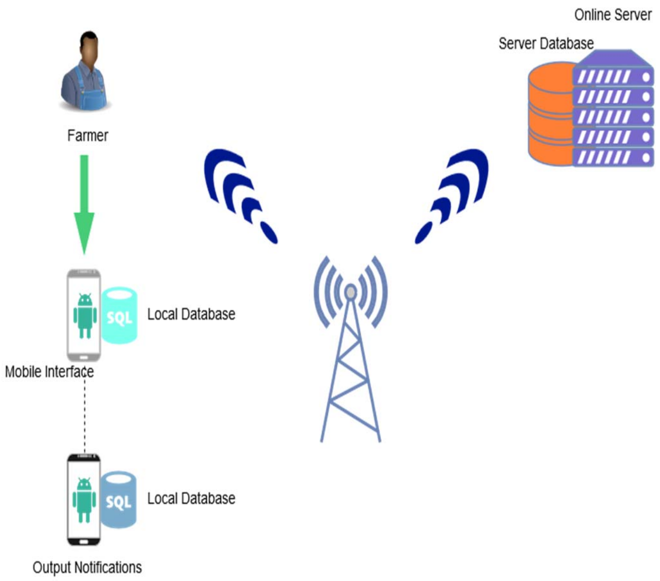

Fig. 13. The System Overview.

Here, the program starts by taking the input from user. Then the chosen decision making algorithm runs in the background of the app. Then upon selecting a crop, it will generate the suggestion and make the schedule (Fig. 13).

B. Work Flow (Fig. 14)

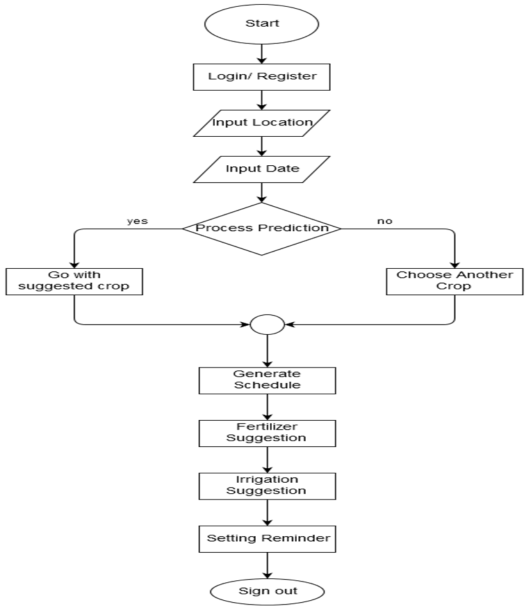

Fig. 14. Work Flow of the System

VI. Conclusion

Being dependent on agriculture for a long time, our country has not seen much collaboration between technology and agriculture so far. There are some websites and also a few mobile applications already in use, for agriculture in this country. But we aim at a future where almost everyone uses a smart phone. We intend to make cropping prediction and cultivation procedures digital. Our system is first of its kind in our country. Agricultural Extension of our country also welcomes this initiative and thinks that this will be a standard to be followed in the future, by other countries as well. Our step is very little but we hope that this is the beginning to something big.

References

[1] K-Nearest Neighbors Algorithm [Online]. Available at:https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#k-NN_regressionAccessed: 27 Nov. 2016.

[2] P.Koli,V. D. Jadhav, "Agriculture Decision Support System as Android Application", International Journal of Science and Research (IJSR),Vol. 4, issue. 4, pp. 903-906, April.2015.

[3] Yearbook of Agricultural Statistics - (2008-2015) [Online] http:bbs.gov.bd Accessed: 27 Nov. 2016.

[4] "The QR decomposition of a matrix,". [Online]. Available: https://inst.eecs.berkeley.edu/~ee127a/book/login/l_mats_qr.html. Accessed: Dec. 12, 2016.

[5] Sunil Ray. (2015) Analytics Vidhya [Online]. Available at: https://www.analyticsvidhya.com/blog/ 2015/08/common-machine-learning-algorithms/ Accessed: 15 August. 2016.

[6] Trading Economics. (1994-2016) Trading Economics [Online]. Available at: http://www.tradingeconomics.com/bangladesh/gdp-growth (Accessed: 15 August.2016).

[7] IRRI. (2012) IRRI. Available at: http://irri.org/news/media-releases/feeding-rice-just-got-easier-with-smartphones Accessed: 2 Nov. 2016

[8] Nikhil,Dr.Kanwal, "Exploiting Data Mining Technique for RainfallPrediction," International Journal of Computer Science and Information Technologies (IJCSIT),vol. 5, no. 3,pp. 3982-3984, 2014.

[9] P. Vinciya, Dr. A. Valarmathi, "Agriculture Analysis for Next Generation High Tech Farming in Data Mining, "International Journal of Advanced Research in Computer Science and Software Engineering(ijarcsse), vol. 6, issue. 5, pp.481-488, May.2016.

[10] Ashwani Kumar,SwetaBhattachrya, "Crop yield prediction using AgroAlgorithm in Hadoop," International Journal of Computer Science and Information Technology & Security (IJCSITS)**,** vol. 5, no. 2, pp. 271-*274,* April.2015**.**

[11] Miss.Snehal, Dr.Sandeep, "Agricultural Crop Yield Prediction Using ArtificialNeural Network Approach," International Journal of InnovativeResearch In Electrical, Electronics, Instrumentation And Control Engineering (ijireeice), vol. 2, Issue 1, pp. 683-686,Jan.2014.

[12] Manpreet, Heena, Harish, "DataMining in Agriculture on Crop Price Prediction: Techniques and Applications," International Journal of Computer Applications, vol. 99, no. 12, pp.1-3, August.2014.

[13] Santosh,Sudarshan, "A Modern Farming Techniques using Android Application," International Journal of Innovative Research in Science,Engineering and Technology (IJIRSET), vol. 4, issue 10, pp. 10499-10506, Oct.2015.

[14] Algorithm (Multiple Linear Regression) [Online]. Available at: http://www.originlab.com/doc/Origin-Help/Multi-Regression-Algorithm Accessed: 27 Nov, 2016.

[15] Department of Agricultural Extension, Bangladesh [Online]. Available at: http://www.dae.gov.bd.Accessed: 28 Nov, 2016.

[16] Agricultural Information Services, Bangladesh [Online]. Available at: http://www.ais.gov.bd. Accessed: 1 Dec, 2016.

978-1-5090-6435-9/17/$31.00 ©2017 IEEE