ISSN No:-2456-2165

Application of Data Mining in Crop Yield Precision

Hasi Saha
Assistant Professor, Department of CSE
HSTU, Dinajpur, Bangladesh

G C Saha*
Assistant Professor,
Department of CSIT
BSMRAU, Gazipur, Bangladesh

Masum Billah
Assistant Professor,
Department of CSIT
BSMRAU, Gazipur, Bangladesh

Najmul Hossain
HSTU, Dinajpur, Bangladesh
Student ID: 1402065

Suraiya Yasmin
Assistant Professor, Department of CSIT
BSMRAU, Gazipur, Bangladesh

Abstract:- Since Bangladesh is an agrarian nation, its economy for the most part relies upon farming yield growth and social agro industry items. Agriculture is to a great extent affected by some profoundly flighty parameters such as temperature and rainwater in Bangladesh. Growth of agriculture also depends on weather parameters like temperature, rainfall, humidity as well as various soil parameters, soil moisture, surface temperature and crop rotation. Since, now-a day's Bangladesh is quickly progressing into specialized advancement therefore technology will end up being beneficial to agriculture that will expand crop efficiency which brings about better productions to the farmers. One of those is also an essentially significant task in agricultures' yield prediction. Before the cultivation process the research suggests different area based beneficial crops. It recommends some crops for a specific territory of land that are financially savvy for farming. Here, the study considered six main crops which names are rice, wheat, maize, potato, pulses, and oil seeds in order to achieve these results. Using Supervised Machine Learning, we find out the prediction through analyzing a static arrangement of information. The static dataset contains previous year's data of those crops according to the area which are taken from the Yearbook of Agricultural Statics in Bangladesh. To obtain this prediction a comparative analysis between Multiple Linear Regression (MLR) and K-Nearest Neighbor Regression (KNNR) has been done in this research. To guarantee learning and preparing of the algorithm and expanding the exactness pace of expectation, we utilized past ten years (2006-2015) dataset and for the case of testing we used one year (2016) dataset for computing accuracy.

Keywords:- *Data Mining; Multiple Linear Regression; K-Nearest Neighbor Regression; Crop Yield; Precision.*

I. INTRODUCTION

Farming is one of the most noteworthy occupations rehearsed in Bangladesh. It assumes a significant contribution in generally speaking advancement of the nation and also the broadest economic sector. In Bangladesh, about 71% of the land in the nation is utilized for farming so as to get the job done the requirements of 163 million populaces. Therefore, agricultural transformation is very critical and accordingly will lead the farmers' profit. In the past, harvest expectation was accomplished by considering the farmers knowledge on a specific crop and field. Now-a-days the situations are changing very rapidly day by day and farmers need to cultivate more and more crops on their land. But farmers need more information about the new harvesting crops and they are not totally aware of the advantages they get while cultivating them. By taking into consideration various environmental condition, the agriculture profitability can be expanded by comprehension and anticipating crop execution. The main task of this study is to introduce a learning instrument that can help in captivating choices to make the agriculture more proficient and productive through technology. The studies give a rundown of beneficial harvesting crops in a specific region of land utilizing decision making algorithms [1]. To see the difference in error percentages of the prediction which will analyze the generation yield of various harvests in earlier years, a comparative analysis between two regression algorithms in data mining such as Multiple Linear Regression (MLR) and k-Nearest Neighbor regression (KNNR) was done. The research's goal is to assist farmers with maximizing their overall revenue by giving forecasts on crops that will give the maximal yield in a characterized zone. The further two extended regions named Dinajpur and Rangpur along with six major crops are focused by this research. The dataset covers the information on crop yield rate per acre (Ton), average of minimum and maximum temperature, humidity, rainfall, range of various year and region. The algorithms provide the outcome which forecasts the most gainful crop in a particular region at a point of time by analyzing these data. Data of previous eleven years is being utilized by the algorithms for the provision of learning and consequence analysis to test the accuracy of the prediction.

II. REVIEW OF RELATED LITERATURES

The field of automated learning is reached out by many researchers. In Bangladesh, it's challenging to expect the exact outcome since dataset is not well structured here and the area of research becomes new to this country. The researchers Vinciya and Valarmathi inspected the ordering between inorganic, real estate, and organic data to mark guess outcome in India [2]. For removing commit data and to foresee they utilized data mining innovation, for the chose area they utilized Multiple Linear Regression in this exploration. Different Linear Regression was utilized for speculation of forecast model and choice supervised learning tree algorithm algorithm was utilized for expectation. The planned misfortune was figured from preparing data from arrangement classifications. There was an equation for prediction to discover the contrast between genuine values and the fitted values and the assessment of the residuals in various linear regressions. It is used to determine the mean squared error. Ashwani Kumar et al. developed an algorithm named "Agro Algorithm" in Hadoop platform to determine the best crop suggestion and predict the crop yield and also used for Hadoop framework for control big data [3]. They considered the weather and soil type. Snehal and Sandeep used Artificial Neural Network (ANN) for crop yield forecast where they utilized feed forward back engendering neural system [4]. Manpreet et al., attempted to examine about the different information mining approach in farming field in agricultural data mining in on Crop Price Prediction: Applications and Methods [5]. K-implies, K-Nearest Neighbor, Support Vector Machines, Artificial Neural Networks, and the need of value expectation of yields were examined in here as per the present market value strategy.

Moreover, [6] used Multiple–linear regression to design representations of reflectance spectra that were greatest greatly associated with an interest for determining soil properties [6].

The paper titled "applying data mining techniques in the field of agriculture and allied sciences" an effort has been prepared to evaluation the studies on use of data mining methods in the field of precision agriculture. Some procedures, such as ID3 algorithms, the k nearest neighbour, artificial neural networks, the k-means, and support vector machines used in the field of agriculture were untaken. Data mining in agricultural use is a comparatively novel method for forecasting of agricultural crop managing [7]

In this line, [8] have done a research on agricultural production output prediction using supervised machine learning techniques. They used Decision Tree Learning-ID3 (Iterative Dichotomiser 3) and K-Nearest Neighbors Regression algorithms to discover the patterns in the data set as part of Supervised Machine Learning techniques. The researchers did not exclude the outliers from the considered dataset that may impacts in predicting results.

Besides this, their research is restricted to some limited dataset so they suggested for future directions of adding extra data that may be investigated with further machine learning methods to produce crop predictions with enhanced precision. Therefore, this study is implemented for leveraging the identified research gap.

III. DATA MINING IN PRECISION AGRICULTURE

Data mining is a discipline that is defining the methodologies being used recently in the field of database technology, statistics, machine learning, pattern recognition and other areas [9]. Data mining procedures target finding those examples in the information that are both significant and intriguing for crop managing. Yield expectation is a particular agricultural issue ordinarily happening. As right on time as could reasonably be expected, a farmer might want to know how a lot of yield he is going to anticipate. The capacity to anticipate yield used to depend on farmers' long haul information on specific fields, harvests and atmosphere conditions [10]. The methods from the field of machine learning, artificial intelligence, statistics along with the database system, that implements pattern recognition in large dataset [11]. To transform bulk amount of data into an understandable format that can be used in different analysis in the main objective of data mining.

IV. SUPERVISED MACHINE LEARNING

These types of algorithms predict target variable (subordinate variable) from a predetermined arrangement of indicators (autonomous factors). Utilizing this prearrangement of factors we generate a capability that diagrams input to favored harvest. Until the model achieves a foreseen degree of accuracy on the precision information, the training procedure remains.

V. MULTIPLE LINEAR REGRESSION (MLR)

Linear regression is a one kind of supervised machine learning technique [12]. From a given dataset containing observations on a dependent and an independent variable it evaluates the relationship between these two for a particular population sample [9]. The only differential fact in multiple linear regressions is that it evaluates the relationship between two or more independent variable and a dependent variable and [13].

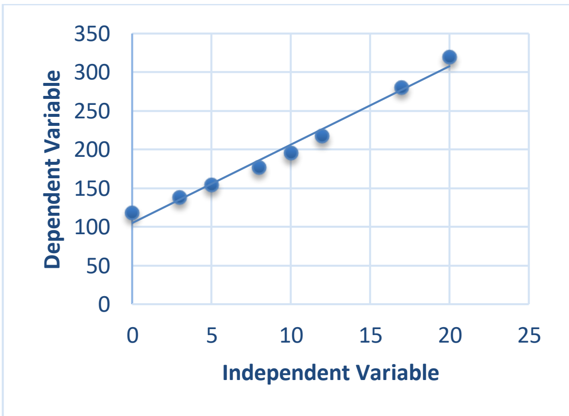

Fig 1:- Regression Model

A best fit line is drawn to establish a relationship between the independent and dependent variables are given in the above regression model graph. From the following equation the line is obtained as:

$$y = \beta_0 + \beta_1 * x$$

Where,
x= independent variable,
y = dependent variable,
β_0 = intercept,
β_1 = slope of the line.

Multiple linear regression has multiple independent variables so the equation which is to obtain the best fit line is following [9]:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n$$

Where, x_i = independent variable, y = dependent variable, β_0 = constant term, β_i = co-efficients relating y and x_i. The mean value of error term 0 is assumed.

VI. K-NEAREST NEIGHBOR'S REGRESSION

In K-Nearest Neighbor's Regression algorithm, to identify the classification of the test cases a correspondence norm between cases in the learning data set and test data set is done. To identify the class in which the test cases belongs, firstly Euclidean distance of all cases are determined and then the k nearest learned instances voted. Depending on the quantity of K that is optimum it finds the adjacent possible values.

Euclidean distance formula:

$$d[i] = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Where, x and y are two predictors in the model.

VII. PROPOSED SYSTEM ARCHITECTURE

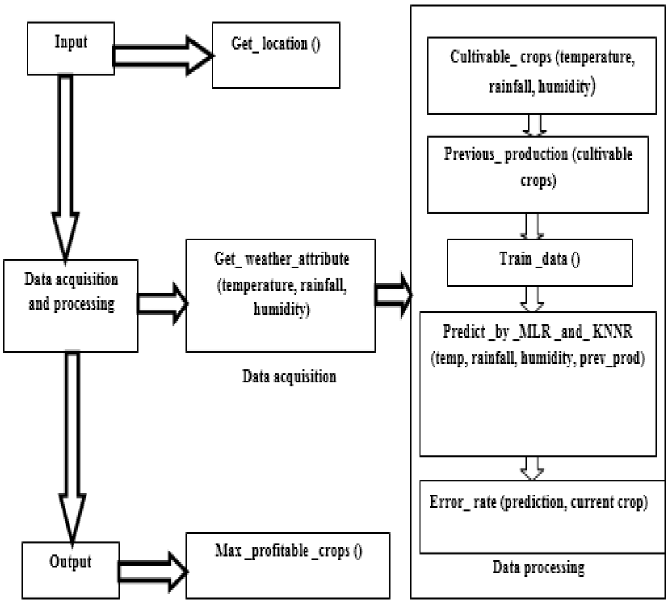

Fig 2:- Architecture of System Prediction

Input: To predict the crop accurately, the expectation of crop harvest is subject to various factors, for example, climate characteristics and past year crop forecast. Area of client is taken as a contribution to the system because of all those factors is location contingent.

Data acquisition: The system quarries the weather attributes in the corresponding area from the weather reservoir dependent on the present client locality.

Data processing: If proper conditions are met then a harvest can be cultivable. These conditions are including comprehensive parameters related to temperature, rainfall and humidity. Then these constraints are weighed and the productive crops are discerned. Multiple Linear Regression and K-Nearest Neighbor regression are used to predict the crop by the system. Past year crop production is used for prediction i.e.: the process is as identify the substantive weather parameters and comparing it with current conditions which will predict the crop in a practical manner and more accurately.

Output: Using our regression algorithms: Multiple Linear Regression and K-Nearest Neighbors Regression the system predict the most profitable crop. Then, the output will be provided to the farmer with a rank of profitable crop imparting the crop duration.

VIII. RESULTS AND ANALYSIS OF EXPERIMENTS

Here, for the analyses we used data bearing four characteristics- average (min+max) temperature, humidity, rainfall which are taken during the harvesting period and crop return ratio per year of six main crops taken from two main areas of Bangladesh. Dinajpur and Rangpur regions are taken into consideration. The six major crops named Rice, wheat, maize, potato, pulses, and oil seeds are considered into this analysis. To obtain the prediction the data of 11 years from every area was used. For the cases of data learning from year 2006-2015 were deliberated and for accuracy analysis the data from year 2016 were used. The relative error finding methodology was used for finding error and its formula is as following,

$$\%Error = \frac{True\ value - predicted\ value}{True\ value} * 100$$

As in the majority of the cases the resultant mistake was underneath 10% the aftereffect of this study is demonstrated to dependable. Because of irregularity in the indicators and furthermore because of playing out the investigation without overlooking any anomalies are the reasons for inability to foresee a nearby incentive in different cases. This expectation model can make increasingly exact and progressively dependable through further modification. An example of a graph mentioned below shows the result of this comparative analysis.

ISSN No:-2456-2165

A. Result of analysis for Dinajpur region

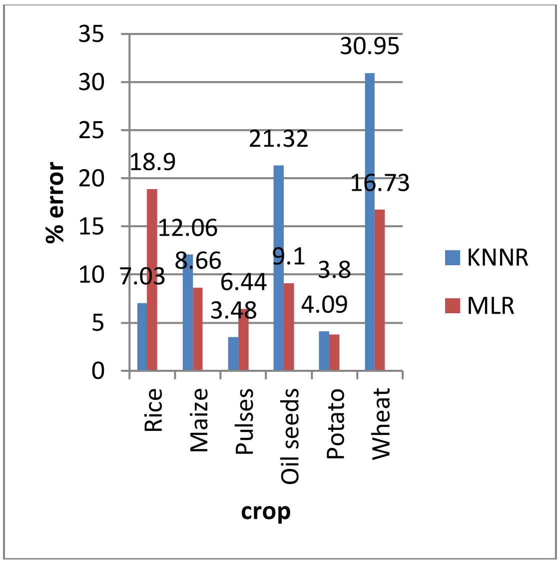

Fig 3:- Showing resultant percentage error for Dinajpur region in the year 2014

We can see that from the given graphs during the year 2016 MLR gives better prediction for the crops Maize, Oil seeds, Potato and Wheat than KKNR algorithm and on the contrary KNNR gives better prediction for the crops Rice and Pulses.

Depending on the seasons, hence following the above graph we applied KNNR algorithm for predicting rice yield rate and for the potato crops we applied MLR to predict. After the application of the algorithms we preferred farmer the best crops during the season should be cultivated by taking account on their yield rate based on their current market price. Same process will be done for other crops as well.

B. Result of analysis for Rangpur region

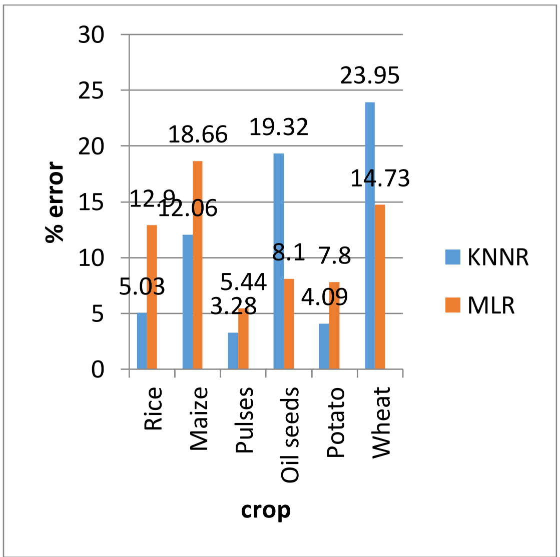

Fig 4:- Showing resultant percentage error for Rangpur region in the year 2014

In the same way, the results from the given graphs during the year 2016 MLR gives better prediction for the Oil seeds and Wheat than KNNR algorithm and on the contrary KNNR gives better prediction for the crops Rice, maize, pulses, and potato.

Hence, according to the result of experiment its suggested to apply KNNR algorithm for predicting rice, maize, pulses, and potato yield rate and for the oil seeds and wheat crops we may apply MLR to predict. After these precision directions, the farmers can choose the best crops during the season that should be cultivated by taking account on their yield rate based on their current market price. Same process can be applied for other crops in different regions as well.

IX. CONCLUSION AND FUTURE WORKS

The data related to previous year production and weather is taken into consideration into our proposed system. After taking account the data our system suggest which is the paramount gainful crops that can be planted in the relevant ecological circumstance. The system helps farmer in choice making of which crop to consider for plantation since the system lists out all possible crops. This system will assist the farmer to get insight into the demand and the budget of numerous crops in market by considering the past production of data. A famer having a particular land can acquire to recognize about the crop that may never have been planted as most extreme sorts of harvests will be secured under this system.

At present no systems are developed which can be recommends crops through considering numerous factors such as Phosphorus, Potassium nutrients in soil, Nitrogen and also weather constituents including temperature, rainfall and humidity. An Android based application is suggested by our future work which can be unerringly predict the most Economical crop to the farmer. With the help of GPS the user location will be identified. Then, the beneficial crop in the conforming position is identified from the weather and soil database according to the user location. The past year production database are compared with these soils to identify the most economical crop in the specified area. The result will be directed to the user's android phone after this handling is done at server side. In order to precise crop prediction, the earlier production of the crops is also taken into considerations. For the extrapolation system location is the only input. According to the user necessity the utmost producible crop will be recommended depending on the numerous scenarios and additional filters.

ACKNOWLEDGMENT

The authors wish to thank Dr.Ruzinoor Che Mat, School of Creative Industry Management & Performing Arts (SCIMPA), Universiti Utara Malaysia, Malaysia, Kedah 06010 for his patronage to the dissemination of knowledge, his valuable instructions, able guidance and keen interest throughout the research. The author is also grateful for his encouragement, precious comments and valuable suggestion in preparation for this work.

REFERENCES

[1]. S. Ray, "Essentials of Machine Learning Algorithms (with Python and R Codes)," in Analytics Vidhya, 2015 [Online]. Available at: https://www.analyticsvidhya.com /blog/2017/09/common-machine-learning algorithms/(Accessed: 12 February 2018).

[2]. 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, vol. 6, issue. 5, pp. 481-488, May .2016.

[3]. Ashwani Kumar, Sweta Bhattachrya, "Crop yield prediction using Agro Algorithm in Hadoop," International Journal of Computer Science and Information Technology & Security (IJCSITS), vol. 5, no. 2, pp. 271-274, April.2015.54.

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

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

[6]. J. A. Thomasson, R. Sui, M. S. Cox, & A. Al–Rajehy. (2001). SOIL REFLECTANCE SENSING FOR DETERMINING SOIL PROPERTIES IN PRECISION AGRICULTURE. Transactions of the ASAE, 44(6). doi:10.13031/2013.7002

[7]. YETHIRAJ N G. “APPLYING DATA MINING TECHNIQUES IN THE FIELD OF AGRICULTURE AND ALLIED SCIENCES”, International Journal of Business Intelligents ISSN: 2278-2400, Vol 01, Issue 02, December 2012.

[8]. Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017, July). Agricultural production output prediction using supervised machine learning techniques. In 2017 1st International Conference on Next Generation Computing Applications (NextComp) (pp. 182-187). IEEE.

[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, vol. 6, issue. 5, pp. 481-488, May .2016.

[10]. Ruß, G., & Brenning, A. (2010, June). Data mining in precision agriculture: management of spatial information. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 350-359). Springer, Berlin, Heidelberg.

[11]. S. Veenadhari, D. Bharat Mishra, and D. C. Singh, "Soybean productivity Modeling using decision tree Algorithms," International Journal of Computer Applications, vol. 27, no. 7, pp. 11–15, Aug. 2011.

[12]. S. Yang, “A Search Algorithm and Data Structure for an Efficient Information System,” University of Wisconsin Madison, Wisconsin.

[13]. M. Tranmer and M. Elliot. "Multiple linear regression." The Cathie Marsh Centre for Census and Survey Research (CCSR) (2008).