Predicting Agricultural Productivity through Machine Learning

Submitted by:

Shiyam Talukder	**16101243**
Habiba Jannat	**16101191**
Sukanta Saha	**15301104**
Katha Sengupta	**16101280**

Supervised by:

Dr. Muhammad Iqbal Hossain

Assistant professor

Department of computer science and engineering

Abstract

Bangladesh is an agricultural country. As the economy is based on the agriculture highly, there should be a progress in this sector. To make a progress in the agriculture the productivity must be increased. In these days, the productivity is low due to various factors. One of them is not finding suitable crop for a particular land. In this way, the crops are not produced at the maximum amount. Hence, productivity of agriculture depends on multiple parameters on the basis of location. The suitable crop for a particular location is necessary for agriculture to bring the most productivity. Here we have designed a system that predicts productivity of the crop with given parameter, also recommend the suitable crop. For the prediction we have used multiple machine learning algorithms and for recommendation we will use content-based filtering system.

1. Introduction

The development of a country depends on the economy and our economy depends on the agriculture. The importance of agriculture in our economy is very significant to eradicate poverty since 2000 and from 2005 to 2010, the agriculture has been responsible to remove 90 percent poverty in rural areas [1]. Previously Bangladesh has been self-dependent for ensuring staple food to feed the population. In these days, due to the population increase, the agricultural land has been 70.6% (with 59.65% arable land and only 6.38% permanent cropland) in 2016 whereas in 2005 it has been 71.5% [2]. Apparently, the change is not noticeable but within few decades the percentage will decrease more and more. Hence, the production of food will decrease to feed this vast population.

In recent years, Bangladesh is struggling more for the inefficiency of food. For instance, rice is the staple food in Bangladesh. Most of the lands are used to cultivate rice. Still, to meet up with the crisis, the import of the rice stood at 22.59 lakh ton in July-December of fiscal 2017-2018, which is the highest since 1998-99 when a record was 30.67 lakh ton [3]. Due to the climate change the productivity of land has been changed. Flood, heavy rainfall, less cold winter and every year with record amount of temperature in summer has a good deal of effect in cultivable lands. Every year, the temperature in summer is increasing at the rate of 1 degree Celsius. By the year 2050, the increasing temperature will be effective on GDP by decreasing 6.7% [4]. The GDP of Bangladesh depends on agriculture with the population of 47% involved in it [5]. Hence, if the agriculture field is neglected, then the population will suffer.

In Bangladesh, the agriculture is modernizing day by day as well as technology. Still the prediction in agriculture has been insufficient. The productivity has been very unpredictable. Most of the farmers depend on their previous experience to cultivate their land. As they are dependent highly on climate for taking decision for agricultural production so if there is change in climate pattern, they do not know which crop to produce. In these years, the productivity of the crop fluctuates due to the changing parameters of climate. Therefore, they cannot produce the desired the maximum amount of crop that could be produced in that area. In a specific area, there is a suitable crop accordingly for the best productivity. To solve this problem, we are going to develop a system which will take a particular district, season, rainfall, humidity and temperature as parameters and apply machine learning algorithms like naïve Bayes classifier, logistic regression, support vector machine, random forest and K-nearest neighbor to predict the productivity of the crops. Furthermore, our system will recommend particular crop which is best suitable for the specific area through Content-based filtering system. Content-based filtering, also referred to as cognitive filtering, recommends items based on a comparison between the content of the items and a user profile. The content of each item is represented as a set of terms. The user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the user. In this case, this method can be used by replacing the user with a specific area and the crops as item to recommend the suitable crop.

It is very challenging for the farmers to expect the maximum yield of any crop. Hence, they need to be informed for the predicted crop amount and also suitable crop for their land. Therefore, we believe that, with the help of our system the productivity will increase as the inputs of the parameter will predict the suitable crop for any area which will bring a solution to the underproduction of the staple foods.

2. Related works

In [6], researcher used various techniques like MLR, logistic regression, exponential smoothing models, Markov chain model for forecasting crops per count, forewarning low or high crop yields. Here, they take plant characters such as height, number of ear heads and effect of weather on crops, past values of production into consideration for forecasting crop production. Researchers developed a system where they put the location, soil attributes, Weather attribute, pH level of the water and soil. They also take the previous year production value to predict which crops will be produced better in that area. Using this algorithm, they can predict what crop will be suitable for that Area. However, their System cannot predict the numerical amount of the production and the individual effect of the soil attributes, weather and water and sudden variation of attributes.

In [7], Authors, use techniques like K-Nearest Neighbor, Support Vector Machine, and Least Squared Support Vector Machine for providing comparative study of various algorithm. In this paper they apply these algorithms on datasets and it shows the accuracy of each algorithm to train the datasets and also mean squared error at the cross-validation phase which shows the rate of production and these models can handle structural accident minimization. They predict the future production in categorized way. However, it requires a huge amount data for finding variance and providing a good result.

In [8], Authors propose an android base application which predicts most suitable crops according to current environmental situation by using multi-linear regression on the weather, soil and past production data. This app uses the location of a place and recommends the best suitable crops from the past data collected from the weather station. According to their model they try to predict the production from soil and rainfall using regression. Then try to incorporate previous production using a multi-linear regression for better accuracy. However, as they use only regression and multi-linear regression, they cannot compare with other algorithm to check if they can improve their accuracy.

In [9], S. S. Dahikar and S. V. Rode implemented the ANN algorithm to determine the suitable crop for the specific soil type. As the soil have a lot of parameter such as pH, nitrogen, temperature and the effect of the climate on the soil, they have predicted one particular crop to be yielded in one particular type of soil depending on the parameters. Furthermore, they have claimed that Artificial Neural Networks (ANNs) works better on this type problem than statistical models as complex neural systems with many inputs has shown better accuracy.

In [10], Mr. V. Lamba and Dr. V. S. Dhaka discussed the techniques of wheat yield prediction comparing Artificial Neural Network (ANN) with other models such as Multiple Linear Regression (MLR), Logistic regression, Time series etc. With their research they have claimed the accuracy and efficiency of ANN better than others. They have used different types of models on different aspects. For Multiple Linear regression and logistic regression model they predicted crop yields through plant characters and pest count. Moreover, amongst the Time Series Models they used exponential smoothing models and auto-regressive integrated moving average models for predicting the area/production of the crops. Finally, they have implemented the probabilistic model Markov chain to forecast the techniques in agriculture.

In [11], authors have compared different AI models for the best prediction model for the Midwestern US. Input dataset was from satellite-based vegetation indices and meteorological and hydrological variables. They have examined the effect of phenology in three periods, and one database was selected as the best months to predict crop and soybean yields. Using the DNN model they have performed an optimization process for the accuracy and they have found it has outperformed rest five of them with prediction error around 7.6% for corn and 7.8% for soybean which is much less than other methods.

In [12], authors have built a recommender system called PRES with object-oriented programming language Java. It has implemented the content-based filtering by vector space model where documents and profiles has been represented as vectors. Based on the profiles and recently visited pages, the documents have been matched. Hence, the user matches with the preferred document through previous interactions and preferences. This system has shown how users find their preferred pages with the algorithm implied.

In [13], authors have applied collaborative filtering that offers recommendation to the users by developing a system with the help of big data for the complexity. They have clustered the data by the characteristic similarities using an agglomerative hierarchical clustering algorithm, then applied Pearson correlation coefficient for the similarity (Collaborative filtering). With this system, farmers buy the essentials and sell the products through application.

In [14], M. Kuanr, B. K. Rath and S. N. Mohanty built a crop recommender system for farmers using fuzzy inference model. They included inference engine and knowledge-based engine together with input variables as temperature, humidity and rainfall. They have collected the data based on questionnaire assessment then created database to predict the outcome of suitable crop. The system takes the farmer's location and preferences for their crop for season and suggested whether the farmer should go for that crop or not. They applied cosine similarity to find similar farmers then used the inference rules for the recommendation. They have also proposed pesticides, fertilizers and other seeds.

3. Work Plan and Proposed Approach

Several Machine Learning Algorithms are not appropriate for the prediction of Agricultural productivity and this is because those are not efficient enough for the accuracy of prediction and they do not have exact information about the crop that are not present in their datasets. In our proposed model, we have used few of the algorithms that are well effective for our dataset to implement a system. Firstly, we have collected a dataset of Agriculture based on different parameters and then these data are pre-processed to extract the actual data needed in the form to work on from the raw data. This process is also obtained by Data reduction method where few of the parameters i.e. Temperature, Humidity, Rainfall and Season are taken as attribute selection so that our data could be easily trained and give the best result. Then the data are split and sent to the prediction model, where 70% data is sent to Training and 30% to Test. These data are trained with the Algorithms that are selected to give the best accuracy and one of those are chosen to be used for testing. After that, the rest of the pre- processed 30% is tested on the built system to predict the crops that can be useful for the suitable parameters of the place and recommend the suitable crop to grow for Agricultural Productivity. The algorithms we have used here are Naïve Bayes Classifier, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Logistic Regression

Naïve Bayes Classifier Algorithm is a supervised learning algorithm which is used to classify data into predefined classes. It is a method to predict the probability of different class based on various attributes using conditional probability and the Bayes Theorem: P(A|B) = $P(B|A)\ P(A)\ /P(B)$. This algorithm is mostly used in text classification and with problems having multiple classes. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature which means they are independent among predictors. It is mostly used in content arrangement, Spam separating, Sentiment Analysis. In our Model, we have used this algorithm to predict the accuracy of our Crop and got precision of around 65%.

K-Nearest Neighbor (KNN) is a type of another supervised learning model that classifies data points based on the points that are most similar to it. KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification. It is also a non-parametric learning algorithm because it does not assume anything about the underlying data. These two properties make this algorithm simple and easy. Moreover, it calculates the distance between test data and each row of training data with the help of some method such as Euclidean, Manhattan or Hamming distance and Euclidean is commonly used. It then finds the probability of these points being similar to the test data and classifies it based on which points share the highest probabilities. In our proposed Model, we used KNN and got an accuracy of 59.99%.

Support Vector Machine (SVM) algorithm is also a supervised learning algorithm based on structural risk minimization which minimizes expected error of a learning model and for classification. It is used to find a hyperplane in an N-dimensional space (N number of features) based on training data that distinctly classifies the data points. SVM produces significant accuracy with less computation power.

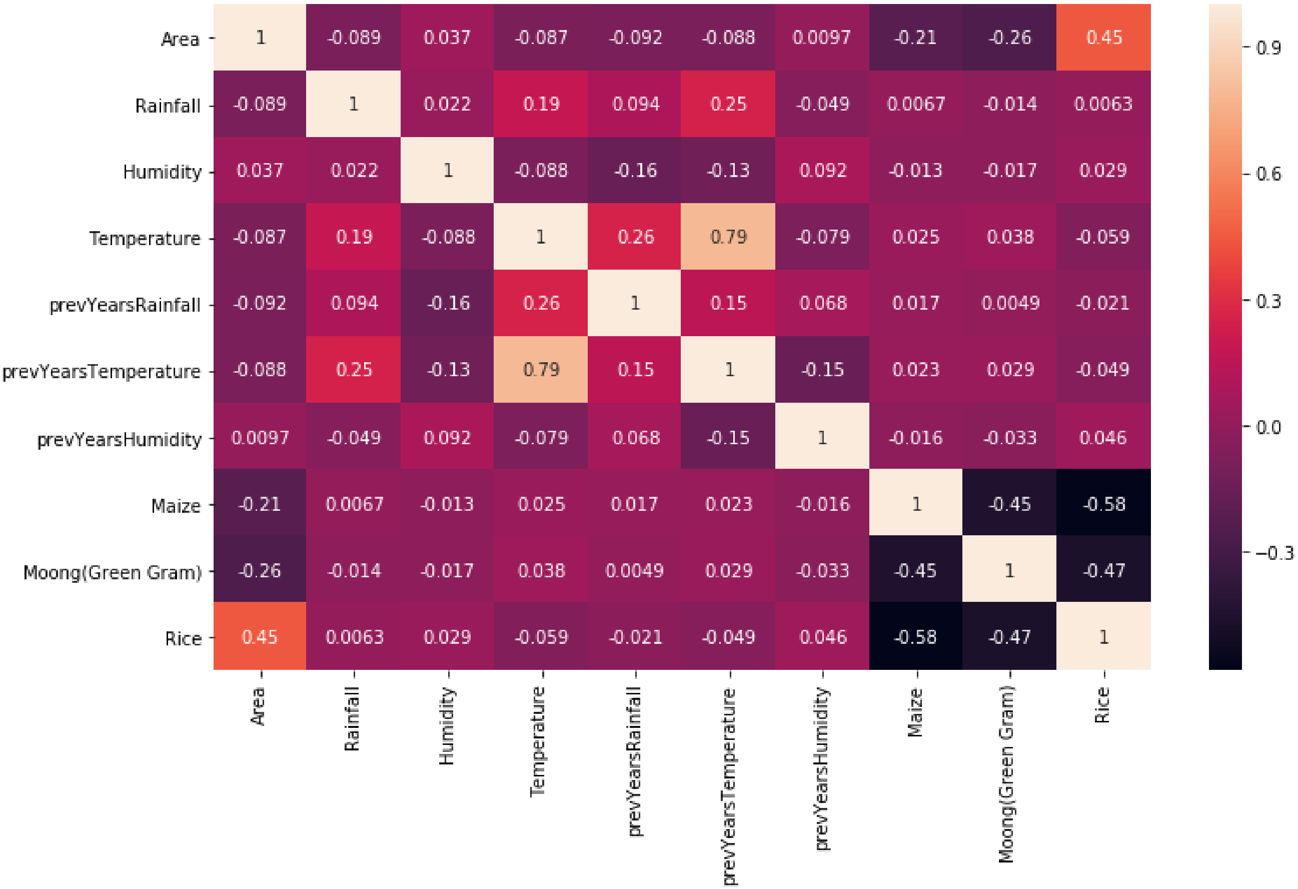
Logistic Regression is the statistical model which is used to estimate the probability of the categorical values and it is also used as a classification model for prediction. It describes the relationship between one dependent binary variable and one ratio-level independent variables. It works with binary value either with 1 or 0 by the probability equation p =1/ 1-e^-z where z = mx + c which does not give a linear curve. It has a threshold value from which it predicts to put 1 or 0 where it is appropriate. If the value is higher than threshold it is 1 otherwise it is 0. In our Model, we have implemented this to predict the crop for the agricultural suitable parameters.

Random Forest algorithm is another supervised classification algorithm which make a forest with many trees based on which a decision is made. The higher the number of trees in the forest, the greater is the accuracy. a Random Forest is built as an ensemble of Decision Trees and it is trained by the bagging method. In simple words, Random forest by making many small decision trees and combing them all to a forest.

We will proceed with our Recommendation System which is the most important part of our System. It will recommend the crop which will be suitable for the agricultural parameters like temperature, season etc.

4. Dataset Description

The dataset we have collected there are 11691 samples where a total of 12 columns and around 11691 rows. The columns were “State Name”, “District Name”, “Crop Year”, “Season”, “Area”, “Rainfall”, “Humidity”, “Temperature”, “Previous Year’s Rainfall”, “Previous Year’s Humidity”, “Previous Year’s Temperature” and “Crop”. Below is the Heat map of the dataset.



	Area	Rainfall	Humidity	Temperature	prevYearsRainfall	prevYearsTemperature	prevYearsHumidity	Maize	Moong(Green Gram)	Rice
Area	1	-0.089	0.037	-0.087	-0.092	-0.088	0.0097	-0.21	-0.26	0.45
Rainfall	-0.089	1	0.022	0.19	0.094	0.25	-0.049	0.0067	-0.014	0.0063
Humidity	0.037	0.022	1	-0.088	-0.16	-0.13	0.092	-0.013	-0.017	0.029
Temperature	-0.087	0.19	-0.088	1	0.26	0.79	-0.079	0.025	0.038	-0.059
prevYearsRainfall	-0.092	0.094	-0.16	0.26	1	0.15	0.068	0.017	0.0049	-0.021
prevYearsTemperature	-0.088	0.25	-0.13	0.79	0.15	1	-0.15	0.023	0.029	-0.049
prevYearsHumidity	0.0097	-0.049	0.092	-0.079	0.068	-0.15	1	-0.016	-0.033	0.046
Maize	-0.21	0.0067	-0.013	0.025	0.017	0.023	-0.016	1	-0.45	-0.58
Moong(Green Gram)	-0.26	-0.014	-0.017	0.038	0.0049	0.029	-0.033	-0.45	1	-0.47
Rice	0.45	0.0063	0.029	-0.059	-0.021	-0.049	0.046	-0.58	-0.47	1

Fig. 1: Heat-map of the input data

Here the input dataset has been manipulated for the optimal accuracy, the columns “State Name”, “District Name”, “Crop Year” has been eliminated as they are object type that machine learning algorithm cannot work on this type of data. The column “Season” is also string type, we have turned it into integer value (0 and 1). The process of our activities is shown in the following flowchart:

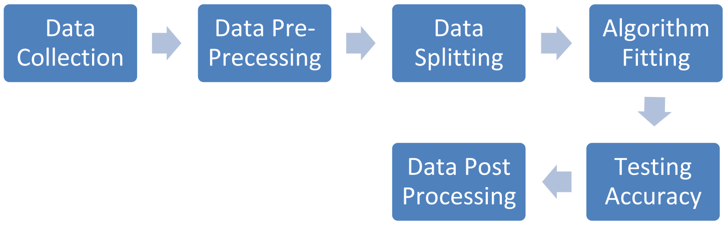


Fig.2: Workflow diagram

Furthermore, we have selected the column “Crop” as target file while the input parameters has been “Season”, “Area”, “Rainfall”, “Humidity”, “Temperature”, “Previous Year’s Rainfall”, “Previous Year’s Humidity”, “Previous Year’s Temperature”.

5. Experimental Analysis

We have performed multiple machine learning algorithms and found the following results:

Table 1. Experimental results

Name of the Algorithm	Rice	Moong(Green Gram)	Maize	Accuracy
1. Naïve Bayes Classifier	88%	42%	43%	60.00%
2. K-Nearest Neighbor (KNN)	66%	46%	40%	51.00% (n_neighbors=18)
3. Support Vector Machine (SVM)	60%	60%	45%	55.00%
4. Logistic Regression				51.00%
5. Random Forest	63%	50%	50%	55.00% (n=60)

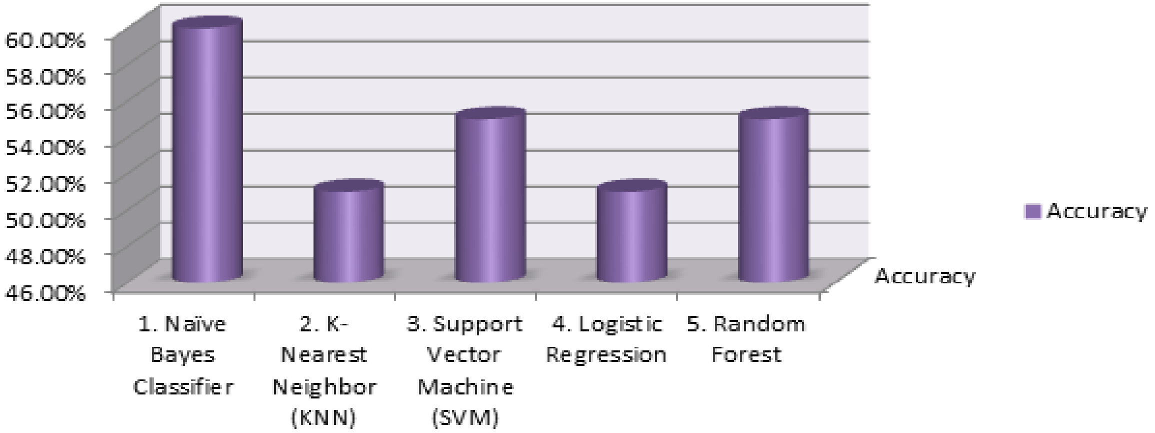


Fig.3: Accuracy bar chart for different algorithms

6. Conclusion

In the proposed model, multiple methods such as naive bayes, logistic regression classifiers, SVM, Random Forest and KNN are used. The dataset used in the proposed model has 11691 samples and the dataset was labeled. For logistic regression, the accuracy score is 51%, Naïve Bayes algorithm 60%, KNN algorithm 51% accuracy, Random Forest 55%and SVM 55%. Five methods are compared to predict the suitable crop. After comparing all the five methods, we found that Naïve Bayes Classifier got the highest accuracy of 60%. Naïve Bayes Classifier is better than other algorithm as it is incredible information driven, self-adaptable, flexible computational instrument having the capacity of catching nonlinear and complex fundamental attributes of any physical process with a high level of accuracy. Now, our outcome will need more works to improve the technique and accuracy of our proposed model.

References:

[1] “Bangladesh: Growing the Economy through Advances in Agriculture”, 2016. [Online]. Available: https://www.worldbank.org/en/results/2016/10/07/bangladesh-growing-economy-through-advances-in-agriculture. [Accessed: Oct. 9, 2016]

[2] “Bangladesh - Agricultural Land (% Of Land Area)”, 2015. [Online]. Available: https://tradingeconomics.com/bangladesh/agricultural-land-percent-of-land-area-wb-data.html. [Accessed: 2015].

[3] S. Parvez, "Rice imports hit two-decade high", Jan. 04, 2018. [Online]. Available: https://www.thedailystar.net/business/economy/bangladesh-rice-imports-hit-two-decade-high-in-2017-18-fiscal-year-1514755. [Accessed: Jan. 04, 2018].

[4] "Bangladesh: Rising Temperature Affects Living Standards of 134 Million People", Sep. 26, 2018. [Online]. Available: https://www.worldbank.org/en/news/press-release/2018/09/26/bangladesh-rising-temperature-affects-living-standards-of-134-million-people. [Accessed: Sep. 26, 2018]

[5] "Bangladesh - Agriculture Equipment and Inputs", Oct. 12, 2018. [Online]. Available: https://www.export.gov/article?id=Bangladesh-Agricultural-Sector. [Accessed: Oct. 12: 2018].

[6] D.S. Zingade et al., "Crop Prediction System using Machine Learning," International Journal of Advance Engineering and Research Development, vol. 4, 2017.

[7] A. Kumar, N. Kumar and V. Vats, "EFFICIENT CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS," International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 6, June 2018.

[8] D.S. Zingade et al., "Machine Learning based Crop Prediction System Using Multi-Linear

Regression," International Journal Of Emerging Technology and Computer Science, vol.3, no. 2, April 2018.

[9] S. S. Dahikar and S. V. Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach", International Journal Of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 2, no. 2, January 2014.

[10] V. Lamba and V. S. Dhaka, "Wheat Yield Prediction Using Artificial Neural Network and Crop Prediction Techniques (A Survey)," International Journal For Research In Applied Science and Engineering Technology (IJRASET), Vol. 2, no. 9, September 2014

[11] N. Kim, K. J. Ha, N. W. Park, J. Cho, S. Hong and Y.W. Lee, "A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015," ISPRS International Journal of Geo-Information, vol. 8, no. 5, p. 240, 2019.

[12] R. V. Meteren and M. V. Someren, "Using Content-Based Filtering for Recommendation," University of Amsterdam, Roeterstraat 18, The Netherlands, 2000

[13] A. A. Chirde and U. K. Biradar, "A Survey on Collaborative Filtering in Accordance with the Agricultural Application," International Journal of Computer Applications (0975 – 8887), International Conference on Advances in Science and Technology, 2014

[14] M. Kuanr, B. K. Rath and S. N. Mohanty, "Crop Recommender System for the Farmers using Mamdani Fuzzy Inference Model", International Journal of Engineering & Technology, vol. 7, 2018.