BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Agricultural Yield and Profit Prediction Using Data Analysis Techniques

Authors
Fazal Mahmud Hassan
Shamma Binta Zakir
Anik Das
Syed Maruf Hossain

Supervisor
Hossain Arif
Assistant Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

I would like to dedicate this thesis to my loving parents, supervisor, faculty members and friends. Without their constant support, this would not have been possible …

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Fazal Mahmud Hassan
Student ID: 14301007

Shamma Binta Zakir
Student ID: 14301092

Anik Das
Student ID: 14301002

Syed Maruf Hossain
Student ID: 13301107

Supervisor:

Hossain Arif
Assistant Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled Agricultural yield and profit prediction using data analysis techniques
Submitted by:
Fazal Mahmud Hassan Student ID: 14301007
Shamma Binta Zakir Student ID: 14301092
Anik Das Student ID: 14301002
Syed Maruf Hossain Student ID: 13301107
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of B.sc in Computer Science and Engineering

1.

Hossain Arif
Assistant Professor
University Building 5, Level 4, BRAC University

2.

Md. Abdul Mottalib, Ph.D
Professor and Chairperson
University Building 5, Level 4, BRAC University

Acknowledgements

At the very beginning we would like to thank the Almighty who enabled us to go on with our research work and give our best efforts to bring it to a conclusion.

Secondly, we would like to thank none other than our supervisor, Hossain Arif sir who supported and guided us since the very beginning of our research. Without his proper guidance it would have not been possible for us progress and finish this work. Whenever we hit a wall, he encouraged us to overcome it and offered any kind of help that was at his disposal. We, the entire team is grateful to him for his excellent supervision, guidance and encouragement that pushed us to successfully conduct and finish our research work.

Also, we would like to thank all the government officials who has given us access to all the data they have recorded over the course of time. Without these data, conduction our research would not have been possible by any means.

We would like to extend our gratitude to our parents, our friends and all well-wisher who had been supportive and helpful through the entire journey. We would like to acknowledge the aid of Big Data Research Team who has provided us with informative support and resources that enabled us to achieve our goal.

Last but not the least, we thank BRAC University and every single person associated with this organization for providing us the opportunity of conducting this research and for giving us the opportunity to complete our Bachelor degree.

Abstract

This research proposes a model that will enable us to predict the agricultural outcome of a country. The prediction will include the agricultural yield for the upcoming years as well as predict the profit margin for particular crops. Although some researches have been done in this field but most of those are in Bangladesh perspective. In our country, being the oldest form of raising economy, agriculture is a field that has not yet been blessed by modern technology or data analysis. This prediction is made by analyzing the dataset of certain vital parameters for rice production, such as Temperature, Humidity, Sunshine and Area using different Regression Analysis and Support Vector Machine (SVM) techniques. The dataset containing the history of Rice Production, Price Diversity and Fertilizer Usage, taken from the yearbook of Bangladesh Agricultural Development Corporation (BADC), Bangladesh Rice Research Institute (BRRI), Ministry of Agriculture (Bangladesh) and by some primary data collection from Gazipur and Pabna, is also analyzed to calculate the profit per year along with comparing the accuracy of the prediction. This proposed model aims to pave the way for data science to touch the sector that keeps our economy running aiming to maximize the production of crops which will result in more profit for the farmers as well as contribute to the economy with the ancient form of revenue collection.

Table of contents

List of figures

List of tables

Chapter 1

Overview

1.1 Introduction

Bangladesh, our motherland is a country that has been gifted by nature through its fertile land. But through the passage of time, the fertility has reduced to a certain extent. Although our agricultural production is satisfying, it still fluctuates. Therefore, predicting how much crop our country can produce in any given economic year is an urgent requirement. The model we are proposing will focus on predicting the agricultural production of our country more specifically the production of rice. Given the specific data that is required by the model, the model can then analyze and form a prediction for the production. Being able to predict the agricultural outcome not only helps us to be prepared for any kind of shortage, but also it helps to maximize the profit for the farmers. Farmers, who are the solely responsible to produce the food we eat are not well compensated. Very often it happens that they do not get the price they deserve just because we are not able to know what the total production might be and what the estimated price would be. Therefore, the model we are proposing will predict the rice production of our country for the upcoming economical year based on multiple factors including humidity, minimum and maximum temperature, sunshine, year-base production of rice, price of rice in the local market and so on. If we are able to generate a well enough prediction, we can then know how the price of each unit of crop is going to be throughout the entire year. This will help the farmers to get the price they deserve and maximize their profit. Therefore, the government can take necessary steps to prevent a chaos and import rice that are necessary.

We plan to achieve our goal by applying data analysis mainly. Our system will require historical data to work with. The data about the related fields are collected from various sources in order to analyze and generate prediction.

1.2 Motivation

These days, data are being collected and are being used in almost every fields. These data that are collected from individuals or organizations are being used for many things. Some of the data are being used in order to generate personalized advertisement over the internet, some are being used to generate probability or predictions about important events, some are being used in order to gain profit in business. However, in our country there is one sector that is not utilizing the advancement of data science and this sector is agriculture. To be more exact, even if a researcher wants to do some research work in this field it is very hard to find any useful data to work with. The least amount of data that can be found are not in any usable format and requires a lot of processing before it can be used in any kind of data science application. Although, the possibility of applications of data science in this sector is very promising as till date very few good researches have been done in this sector in perspective of our country. Taking this into consideration, it worked as our main motivation behind our research. We wanted to work and contribute in this field which has not been utilizing the resources that modern computer and data science has to offer. We wanted to develop a system at a initial state that can predict the agricultural outcome of a country given that it has proper input data. Correct and useful data are hard to find in this regard. Most of the data are not even digitized. Therefore, we wanted to work in this sector. So that we can contribute in making a good and useful data set so that any researcher who would want to work int this in the future would have a decent data to work with. This will make the process on contribution in this sector much more effective and would encourage and motivate the authority to collect more relative and useful data regarding this sector. This was our motivation behind putting our effort in this sector that has been deprived from the blessing of modern data science and analysis. A successful implementation of the proposed model will enable us to predict the future outcome of agricultural production which will tell the if authority if there is going to be any kind of scarcity of any crops. If so, then the authority, in this case the government cane take advanced steps in order to overcome this issue by suitable means.

1.3 Contribution Summary

As agriculture sector is the main contributor behind our economy, knowing the future outcome will not only help us to know about the food production of the country but will also enable us to have an idea about the major portion of our economy. Our proposed model will predict the future outcome of any agricultural product if data of certain parameters are given as input. The output data will give us a prediction about the production. This will give ideas

about where we should put more emphasis in our production. The proposed model is initially tested on rice production but it is suitable to work with any other crops to. Also, during our research, we have created a data set from scratch and unusable data that can be used for future researches. The data we initially got was not usable and most of them were in hard copy. We had to manually digitize the data. After that we have performed multiple filters and cleaning to the data set and came up with a usable data set for our research. This data set that we developed can be used to run multiple algorithms. So we would like to mention the data set as our contribution along with our main proposed model.

1.4 Thesis Outline

In the later part of the paper, the chapters of the thesis are organized as follows:

- Chapter 2 of this paper presents the background information regarding this approach. Also, the algorithms that were planned to be used and were finally used in order to generate the required prediction.
- Chapter 3 presents the proposed model of our research that includes detailed information about the parameters or attributes that were used to generate prediction. Also, this holds discussion about the final data set along with the data collection process and accuracy of the implemented algorithm.
- Chapter 3 proposed model
- Chapter 4 presents our experimental setup,analysis and findings in details.
- Chapter 5 concludes and talks about the future work plan for the research.

Chapter 2

Background Study

2.1 Literature Review

Prediction, or predicting the future data to be more precise is now one of the most explored topic in computer science. In data science, a major part is mining of data and predicting the future outcome of any certain field. Many algorithms have already been developed in order to achieve this. These predictions are being made by using algorithms in almost all the major fields. Companies and organizations are using these predictions to estimate their future sales or production so that they can take proper steps to maximize their profit and grow their business. Already researchers are using data analysis techniques in order to make predictions [1]. Predicting the future outcome, these researches have also been extended to the field of agriculture. Researchers have already tried to predict the future yield of crops applying the techniques of data science and machine learning.

Previously, researches have used BP Neural Network approach to achieve this same endeavor [2]. According to the researchers, BP Neural network approach has three layers that need to be implemented. They are known as input layer, hidden layer and output layer. The hidden layer itself has many multiple layers that do not have any individual connections but has only connection to adjacent layers. This approach of predicting the outcome is quite difficult to implement and increases complexity. Also, researchers had found some limitations for this approach.

Furthermore, researchers have also applied supervised machine learning techniques in order to achieve accurate prediction[3]. Multiple learning algorithm have been used in their approach. Researchers applied K-Nearest Neighbour or KNN, Decision tree algorithm ID3 which is also known as Iterative Dichotomiser 3 in order to analyze the production of the previous year. However, this method of implementations also had their limitations. The main drawback is that the data set has to be very well defined in order to apply these algorithms.

In perspective our country, finding a very well-defined data set is not possible. In fact, the available data are very much scattered and not of any particular format. So applying these techniques require a heavy data preprocessing which is not efficient. Our model looks for algorithms that do not require very well-structured data.

Researchers have applied decision tree model in order to make prediction about objects that are not discrete and real but structured[4] . They calculated the expected loss or profit from their training data classification. Previous work mainly focused on a framework based on information dimension system[5] .

Researchers who applied BP neural network has found that it has shortcomings of its own[2]. These are mainly seen when dealing with large number of multidimensional data. When the types of data sets are found, the networks structure becomes hard to design and it requires more training time as the convergence time become longer. Also, it is prone to fall in local minimum state[2].

Researchers also tried to apply time series analysis in order to predict the price[5]. Researches were conducted to predict and analyze the fluctuation price of vegetables and crops. Their study aimed to identify the time series function that would identify trends and seasonality of individual vegetables and investigate factors of supply and demand[6].

Furthermore, many researchers have also proposed model that utilized APC and CV in order to determine indicators that showed the fluctuating prices[6]. ARIMA and GARCH has also been applied by researchers as time series models[6].

Researches also used satellite data in order to predict the temperature in the areas where crops are produced[7]. But the major limitation for this approach is that this model will not be accessible to many because it requires access to real time satellite data. Additionally, the accuracy of this approach turned out to not satisfactory.

In this paper, we propose a model or a system that will make prediction about the future production of crops and their price and profit margin. In order to do so, we implement the regression algorithm and Support Vector Machine (SVM) algorithm. This model implements multiple methods of regression analysis to determine the accurate outcome. Also, this approach will give us overall idea about the most suitable algorithm for this use case as we will be implementing multiple algorithm on the same data set. Therefore, we can compare and determine which algorithm delivers the most accurate result and in a more efficient manner. These methods can work with data set that are structured and well defined. It is not possible to find such data set in our country so we have manually collected all data ourselves and processed them in order to make them suitable for our algorithms.

2.2 Algorithms

To implement our research on predicting agricultural yield for any particular crop using multiple factors of production, different algorithms were required. Therefore, we have implemented multiple algorithms in order to predict the outcome. Multiple algorithms have been used because not all algorithms were suitable for all the use cases and also, we wanted show a comparison between different algorithms. In our research work, we have used linear regression, support vector machine (also widely known as SVM), and random forest.

2.2.1 Regression Analysis

Regression analysis a method in which observations of events are used to determine and quantify relationship between one variable and some other independent variable, if any. The dependent variable or the target variable is a field in the data set and also it is widely referred to as dependent variable in regression analysis. Though regression analysis is a part of statistical science but it is implemented in numerous fields of data science and machine learning. The base of regression analysis is based on these two variable types. Researchers often have used regression analysis in order to generate a model for analyzed data. For the goal of making predictions based on multiple factors, there are many types of regression types that are available and widely used. However, all the regression techniques are based on mainly three main features or parameters[8]. These are the main features of any model that drive the regression algorithm towards prediction. The three parameters are:

1. Total number of independent variables
2. Type of all the different variables
3. Shape of the regression line.

In short it can be said that, regression analysis is used with a view to estimating or predicting the average value of the dependent variable in terms of the known or fixed values of the independent variable(s). Dependent variables are usually denoted by “Y” and independent variables are denoted by “X”.

Linear Regression

In order to make any kind of model for data prediction, one of the most widely used method is linear regression. The dependent variable, which is one of the key factors of regression method, is usually continuous in nature. However, the independent variable can either be both

of continuous or discrete. The regression line that is produced is linear by nature, hence the name linear regression. Also, the variables have linear relations between them. Using linear regression method, relation between dependent variable (X) and independent variable (Y) is established. The relation is established using a best fitted straight line. This is represented by the equation below:

$$Y = a + bx + e \tag{2.1}$$

In the equation given above,

1. represents intercept
2. represents the regression coefficient
3. e represents the error term

We will now discuss all these terms, in order to have better idea about them.

1. Intercept (a): if we consider a function y = f (x) then, the value of y will be the intercept value given that the value of x is zero.
2. Regression coefficient (b): For every one unit of change in “x”, the change of value seen in y is known as the regression coefficient.
3. Error term (e): This is the error that might have been generated by using the sample data or set of data. This error might generate wrong idea about the population.

After predicting the model there can be minimum error. Therefore, we can calculate the error from the following equation:

$$Y(pred) = bo + b1x \tag{2.2}$$

The values b0 and b1 must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error[8].

$$Error = \sum_{i=1}^{n} (Actual\ output - Predicted\ output) \tag{2.3}$$

Then, the intercept calculation with the equation below[8]:

$$bo = y' - b1x' \tag{2.4}$$

Lastly, the regression coefficient equation is[8]:

$$b1 = \frac{\sum_{i=1}^{n}(Xi - X')(Yi - Y')}{\sum_{i=1}^{n}(Xi - X')^2} \tag{2.5}$$

From these equations a perfect model can be plotted to get the best outcome possible. Linear regression is mostly used technique both in machine learning and data science so this should be calculated without any error.

2.2.2 Support Vector Machine

One of the prominent machine learning algorithms is support vector machine, which also widely known as SVM in short[8]. This is a supervised machine learning technique. This is used both for classification or regression. The portion of this which is used for regression is known as "Support Vector Regression or SVR". There are some minor differences between SVM and SVR. First of all, the output is a real number so it is hard to predict it in hand. When it is the case of regression, there is a margin of tolerance (epsilon) which is set for the approximation. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated[9] . There are two types of SVR. They are

1. Linear Support Vector Regression
2. Non-Linear Support Vector Regression.

Linear SVR

The following graph represents the model of Linear SVR in Fig. 2.1. https://www.overleaf.com/project/5bfc1e
This is the equation for the Linear SVR[8]:

$$y = \sum_{i=1^N} (a_i - a_i^*).(x_i, x) + b \tag{2.6}$$

Non Linear SVR

The following graph in Fig. 2.2 represents the non linear SVR model.

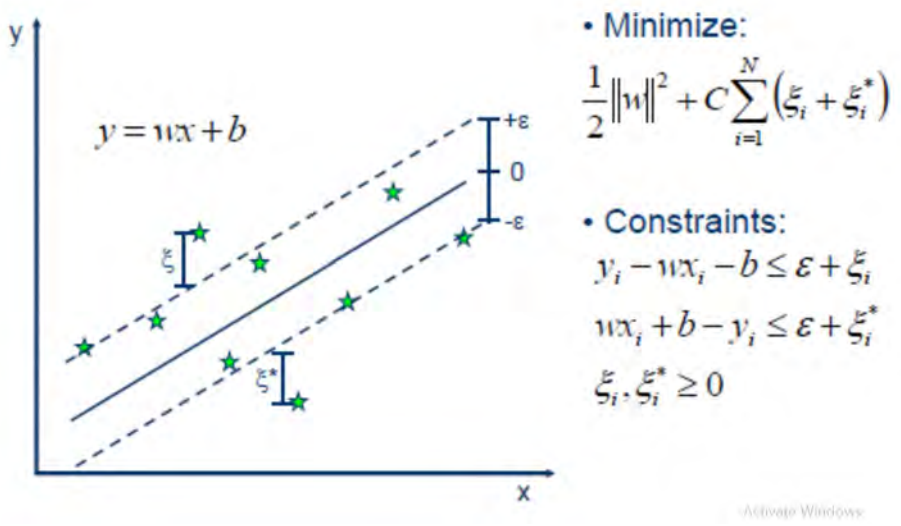


Fig. 2.1 Linear SVR[8]

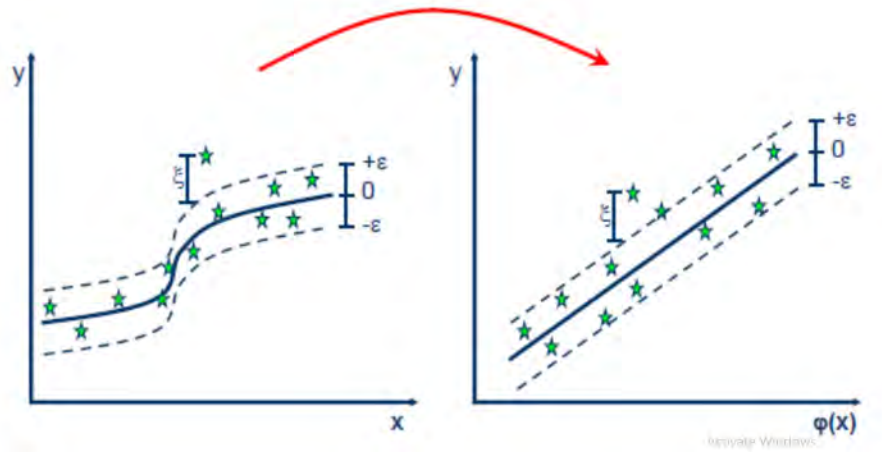


Fig. 2.2 Non Linear SVR[9]

The following equations are the formula of the non linear SVR[9]

$$y = \sum_{i=1}^{N} (a_1 - a_1^*).(\phi(x_i), \phi(x)) + b \tag{2.7}$$

$$y = \sum_{i=1}^{N} (a_1 - a_1^*).K(x_i,) \tag{2.8}$$

2.2.3 Random Forest

Random Decision Forest or Random Forest[10] is one of the popular methods that are used by researchers in order to generate predictive models. This method can be used for both regression problem and classification. This is a very good machine learning algorithm which predicts outcomes with higher accuracy most of the times even when data sets that does not have proper parameter tuning. Therefore, it has a simplicity compared to other algorithms and this algorithm is very much popular and used very often. As the name suggests, it creates a forest in random manner. The forest created by this algorithm consists of Decision trees trained with some method. In this approach, forest of all multiple decision trees are generated. After that, there are merged together in order to produce even more accurate predictions. The higher the number of decision trees that are generated, the higher the stability of the predictions. The same can be said about accuracy. When constructing a random tree in the Support Vector Regression approach, there are mainly three choices that are to be made. They are:

1. The method that is used in order to split the trees
2. Type of predictor that are to be used in each of the leaves
3. Method that is going to be used in order to push randomness into the trees

The algorithm can be represented using the following diagram in Fig. 2.3

Tuning Parameters

For tuning the parameters there are some steps that is followed widely. The steps are given below[11]:

Num Trees: The final number of trees in a model mainly depends on the total number of num trees. The accuracy of the model depends on the number of num trees also. Higher number of num trees result in higher accuracy. But on the other hand, both the training time and prediction time grows linearly with the number of num trees increasing.

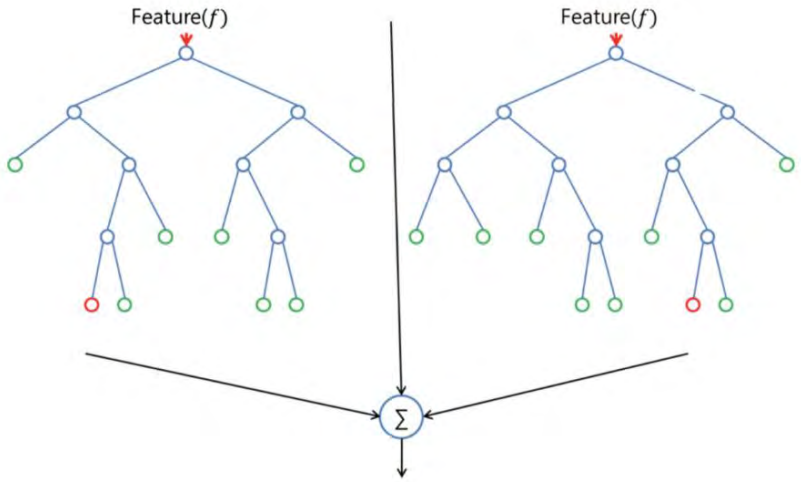


Fig. 2.3 Random Forest Model[10]

Max depth: This parameter is used in order to restrict the depth of each individual tree so that data over-fitting can be avoided.

Step size: Step size is also sometimes termed as shrinkage. This appears as in the equation. It works in the same learning rate of the gradient descent procedure where the smaller value takes more iterations to reach the equal level of training error of a large step size. In this case, there is a trade off between number of iterations ad step size.

Min Loss Reduction: This is another criteria of pruning for construction of decision tree. For a node split, this restricts the reduction of loss function. Here, simpler trees are produced by larger values.

Min Child Weight: This is another criteria of pruning for decision tree construction. This corresponds to the minimum number of observations required for a leaf node, in classification problem. This also generates simpler trees for larger values.

Row subsample: At every individual iteration, it only accesses and uses fraction of data. It has similar characteristics to mini batch stochastic gradient descent which reduces the computation cost for each iteration along with producing a more robust model.

Column subsample: Picks any subset of the column in order to use at any iteration..

Though decision tree and random forest are almost the same, there is a main difference in the concept. In decision tree entire data set and all the features are used whereas in random forest specific features and rows are used. After that average of the features and rows are used to

build multiple decision trees and then average the results.After a large number of trees are built using this method, each tree "votes" or chooses the class, and the class receiving the most votes by a simple majority is the "winner" or predicted class.

Chapter 3

Proposed Model

3.1 Overview

In order to predict the outcome, we initially wanted to use a couple of algorithms. However, some of them had to be vetted out because of their lower efficiency and accuracy. During our implementation, we made a guess that Linear regression would perform well. However, our practical finding was completely opposite. The efficiency and accuracy of linear regression was well below the acceptable threshold of error. However, Random Forest and SVM has performed well and had very good accuracy. We will compare all the algorithms and their outcomes in later part of the paper. Before even selecting the algorithm to we work with, we had to collect the data. In our case, getting data for the research work was very hard because in our country records are not maintained regarding this. Majority of the data are not in any kind of machine readable format or even on the internet. Therefore, we had to collect the data manually and reconstruct them to a machine readable format.

3.2 Workflow

Our research work has been divided into a few major parts. All these parts together represent our research model. The procedure we have taken are as follows:

1. Data collection
2. Data processing
3. Feature extraction
4. Feature scaling

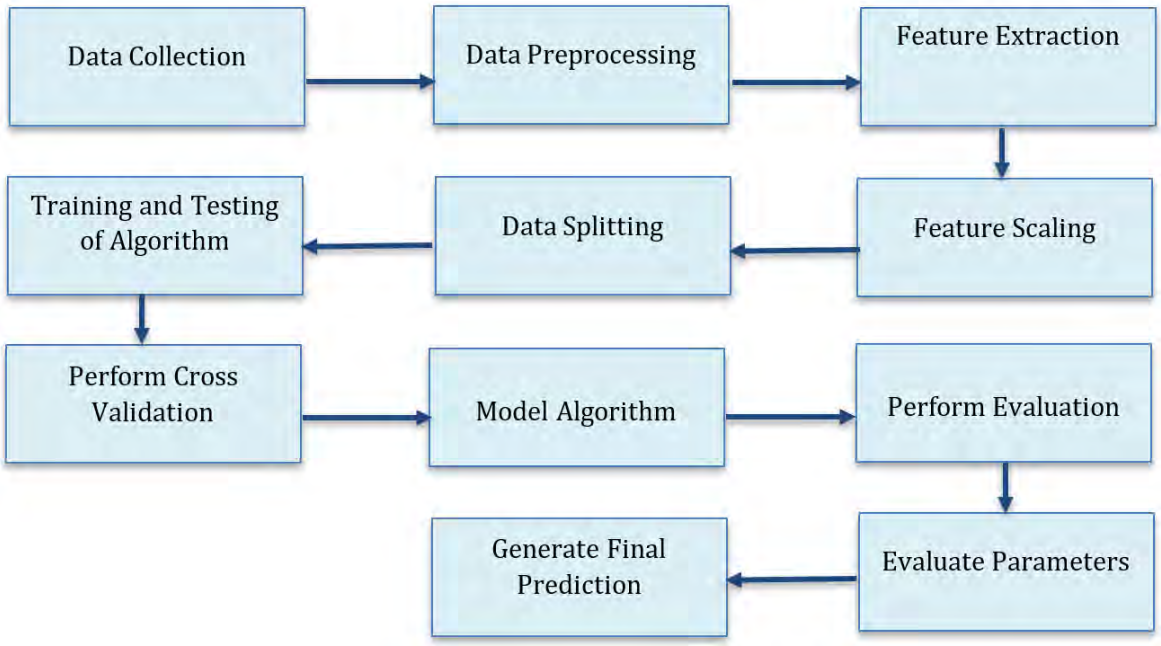


Fig. 3.1 Work Flow of the Research

5. Data splitting
6. Training and testing algorithms
7. Perform cross Validation
8. Perform evaluation of algorithms
9. Predict parameters
10. Generate final prediction

Fig. 3.1 illustrates our work flow for the research. Now in the later subsections, we will talk about each of the parts of our work flow.

3.2.1 Data collection

This was the very first step taken in our model. However, before going for the data collection process we had to understand what are the data that we need in order to make a successful implementation of our proposed model. In order to do so we had to perform background study on factors of agricultural production. After reading multiple research papers and methods of cultivation, we finalized the required attributes for our research. Unfortunately, necessary data were not available in any of the existing open databases. Datasets were also

not found. Data which were made available online were not suitable to be used in our model as most of them were hand written and then scanned. This not readable by machine. Also, accurate and consistent data were not found. Therefore, we had to do primary research in order to have the required data. Data were collected from the field. This included multiple government facilities including Bangladesh Agricultural Development Corporation (BADC), Bangladesh Rice Research Institute (BRRI) and Bangladesh Meteorological Department (BMD, Widely known as "Abohawa Office").

3.2.2 Data preprocessing

The second important step in our work flow was to preprocess the data. This is done so that the machine can understand the content of the data. Usually the step includes finding, removing or replacing data that are not accurate, incomplete or not related to the model. However, in our circumstance, this step included digitizing the data from hard copies, merging multiple files into one, converting all data into one type of unit system, finding missing values or entries, formatting entries, eliminating unnecessary values along with screening and cleaning. Our process took longer as we have collected data from the year 1972, some of the entries were missing and majority of the data were in hard copies. Also, as we have gathered data from multiple sources, some of the entries had conflicting values. In case of these types of values, we had to figure out which values are more reliable. Also, there were a lot of redundant entries and they needed to be filtered out. Unnecessary entries in the collected data were also filtered out during this process. Data needed to be converted into the same units of measurement from multiple formats. After getting through all the screening and cleaning process, the data needed to be formatted in excel and then it was converted to CSV. CSV file was used in the model.

3.2.3 Feature Extraction

Feature extraction is a process in which the number of features in a large data set which describes the data set are reduced in order to reduce computing complexity. Because having a large number of variables in a data set might cause overfit to training samples and result in poor generalization for new samples. However, the final data set did not require this because it was manually constructed which avoided having unnecessary variable the first place.

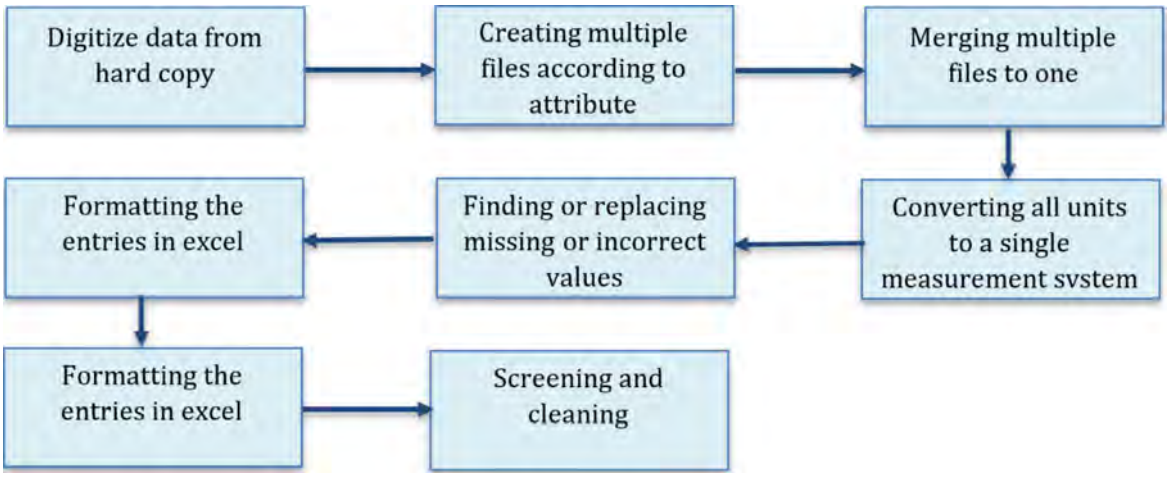


Fig. 3.2 Data preprocessing steps

3.2.4 Feature Scaling

This is a method in which ranges of independent variables or features are standardized. This is also widely known as data normalization and this was taken care of data processing step. We have utilized Standard Scalar library which is a part of Scikit Learn package to scale our data.

3.2.5 Data splitting

Collected data were split into two parts initially. The split was made for training and testing of the system. The process is considered essential for any supervised machine learning or data science application. Because the main outcome mostly depends on the accuracy of the result given by the machine or model. Firstly, we formatted our data in such a way so that it has equal or close number of rows for all the attributes that are being used. This was done in order to avoid any kind of anomaly. Also, this reduced the processing time to a great extent as many unused sets of rows were removed and the machine did not need to process the unneeded rows. This also helped the algorithm to run on our less powerful machine with considerable efficiency. Usually, the norm which is followed is that the data are split in to a ratio of 80:20. We followed the same norm which means that 80 percent of the total data were used as training set to train the system and rest 20 percent of the were used as testing set. After that we have done a split of 75:25 and followed the same process. This showed higher accuracy, therefore this ratio was used in the final model. We have made utilization of libraries, like Scikit learn, that are widely used and available in python programming language in order to achieve our goal.

3.2.6 Training and testing of algorithm

As mentioned before, data has been divided into two parts. One of the parts is for training the algorithm and the other is for testing it. Upon successful training and testing of the algorithm using the divided data set. This process is followed for all the algorithms that are to be implemented. This was implemented using TrainTestSplit library of the Scikit Learn package.

3.2.7 Performing Cross Validation

Cross validation is a process in which the data are compared with the test data for validation. Data cross validation is part of TrainTestSplit library which is a part of Scikit Learn Package. This step is done to find out whether the model is over-fitted. Besides there is another library, named CrossValScore, inside Scikit Learn Model selection package which is used to evaluate cross validation score. This method has been used to find out the mean of cross validation scores of multiple train test split.

3.2.8 Perform Evaluation of Algorithms

In order to evaluate performance of any regression algorithms some matrix are considered such as Variance and Mean Square Error. Here, Variance states the distance between observed values and the average of predicted values. Better model is subjected to have lower variance in terms of performance evaluation.
Another matrix is Mean Square Error (MSE). MSE mentions the average of the square errors where error is the difference between actual values and predicted values. Better model is subjected to have lower mean square error.
The performance evaluation of the algorithms is shown in bar charts centering different independent features mentioning algorithms used. Fig. 3.3 represents performance of Random Forest, Linear Regression and SVR algorithms applied on Minimum Temperature which is an independent variable for predicting rice production. In the figure, separatelye and grey represents Linear R barsegression,Random Forest and SVR respectively. Accuracy is represented by y-axis in percentage.
Fig. 3.4 shows performance of Random Forest, Linear Regression and SVR algorithms applied on Maximum Temperature which is also an independent variable for predicting rice production. In the figure, colored blue, orange and grey represents Linear R barsegression,Random Forest and SVR respectively. Accuracy is represented by y-axis in percentage.
Fig. 3.5 illustrates performance of Random Forest, Linear Regression and SVR algorithms

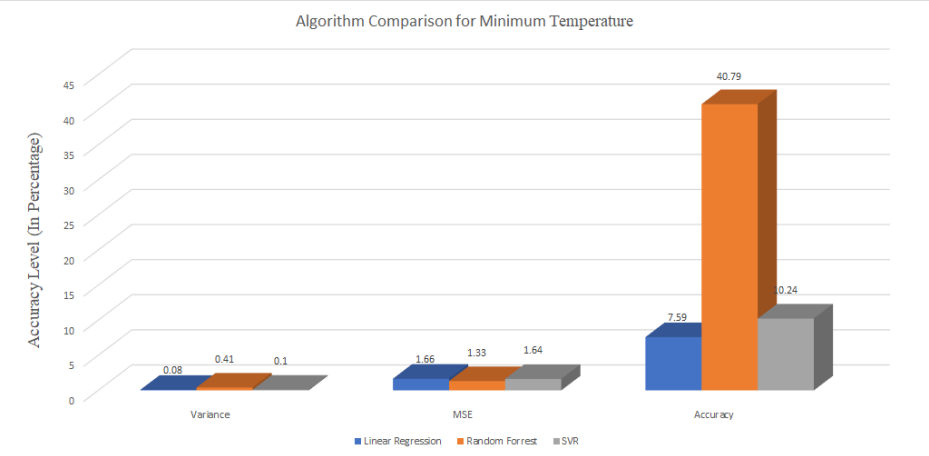


Fig. 3.3 Algorithm Evaluation of Minimum Temperature

applied on Yearly Rainfall which is also an independent variable for predicting rice production. In the figure, colored blue, orange and grey represents Linear R barsegression,Random Forest and SVR respectively. Accuracy is represented by y-axis in percentage.
Fig. 3.6 illustrates performance of Random Forest, Linear Regression and SVR algorithms applied on Yearly Humidity which is also an independent variable for predicting rice production. In the figure, colored blue, orange and grey represents Linear R barsegression,Random Forest and SVR respectively. Here accuracy is represented by y-axis in percentage.

3.2.9 Predicting Parameters

After evaluation of the tested and applied algorithm, we took the best performing algorithm in to consideration for rest of the research. This algorithm was used to make prediction about the selected parameters for the research. All the parameters that have been selected for the research has been run through the algorithm in order to generate prediction. Predictions of each of the feature or attribute were saved in CSV for future implementation.

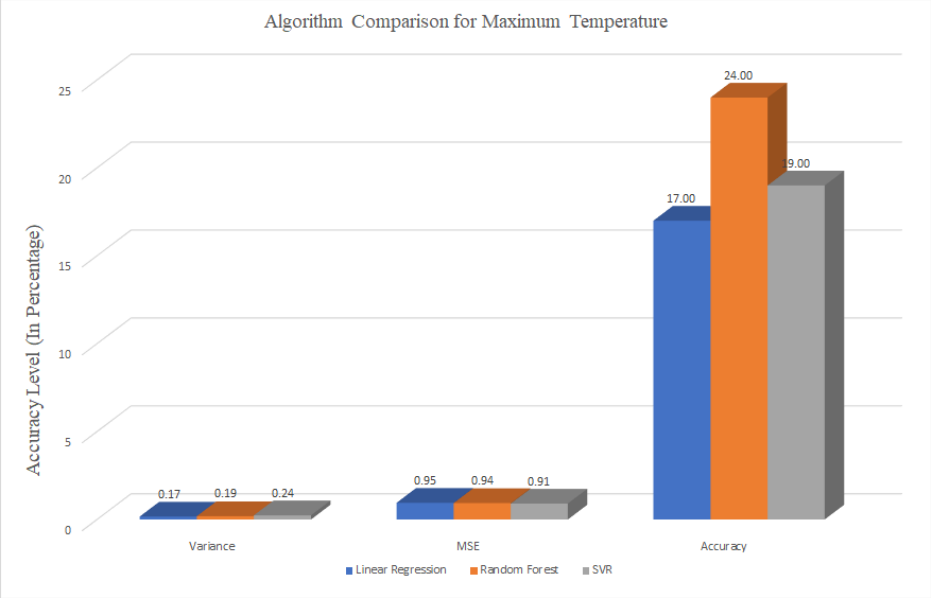


Fig. 3.4 Algorithm Evaluation of Maximum Temperature

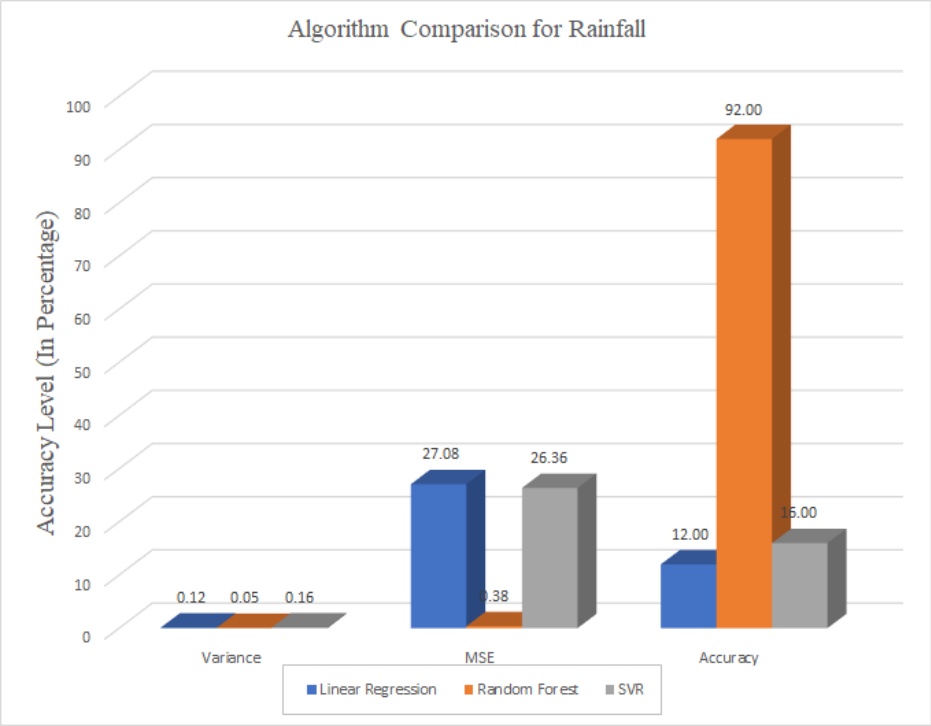


Fig. 3.5 Algorithm Evaluation of Yearly Rainfall

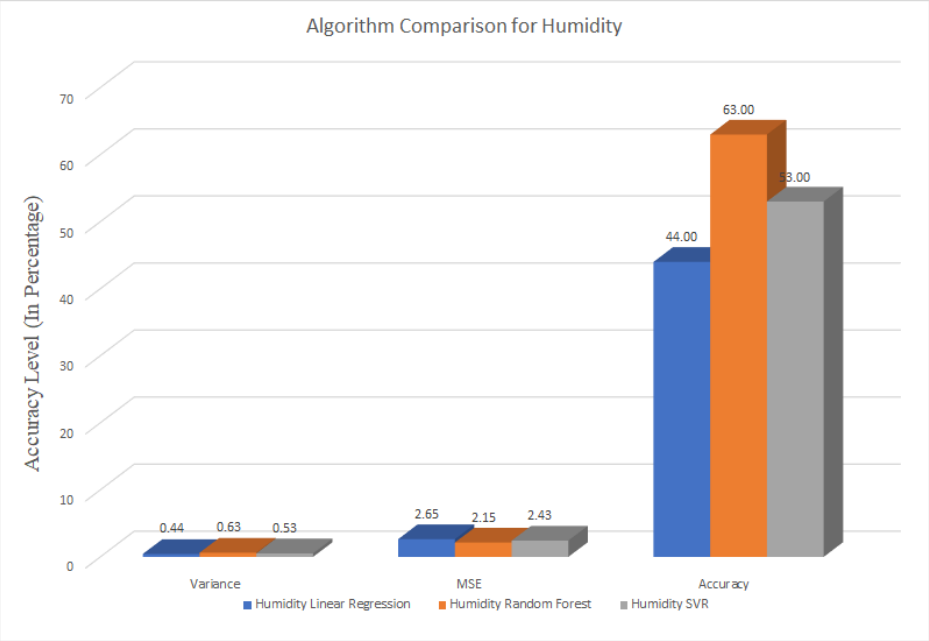


Fig. 3.6 Algorithm Evaluation of Yearly Humidity

3.2.10 Generate Final Prediction

Final prediction of our model was generated by running the best performing algorithm on the predicted parameters.

3.3 Parameters or Attributes

Parameters or attributes that has been used in this research are basically the factors of production of agricultural products. Agricultural production depends on these factors. Changes in these factors will have meaningful impact on the yearly agricultural outcome of any place or country. The attributes or parameters that we chose to work with mainly depended on the availability of the data. We avoided any attributes for which any data were not present or was not possible to collect. the attributes that have been used in our research are as follows:

1. Temperature
   (a) Minimum temperature
   (b) Maximum temperature
2. Humidity
3. Total mount of rainfall

4. Total amount of production from previous years

5. Total amount of land area

In this section of this paper, we will have a short description about the attributes and reasons behind choosing them for our research study.

3.3.1 Temperature

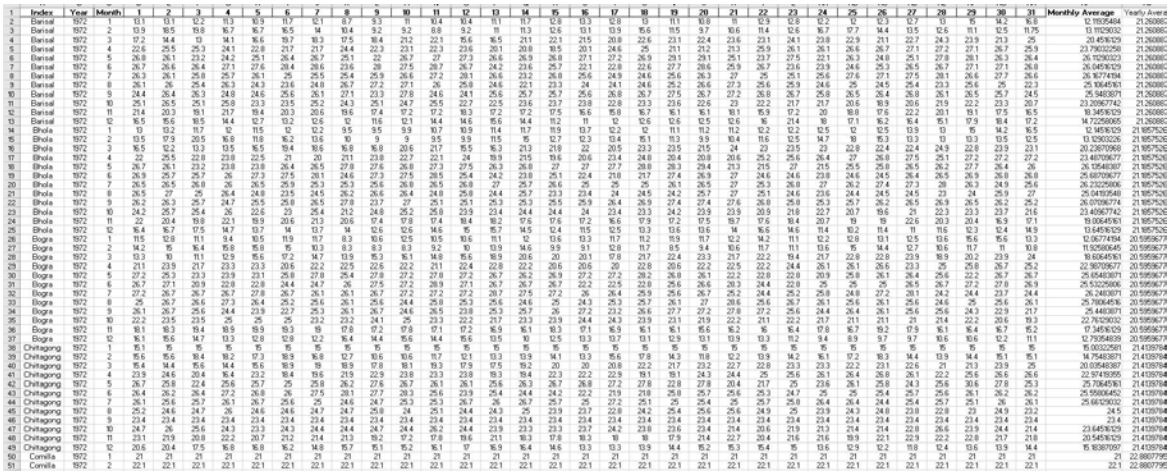
When it comes to growing any kind of plants, one of the most important key factor is the temperature. Temperature is one of the major part of the environment that has impact on the growth of any kind of plant or crop even for that matter[12]. If the temperature is too extreme, that can either be very cold or very warm, researchers have found that in [13] the crop that are to be produced will not grow at the expected rate or even grow at all. If there are massive temperature change during harvesting period, the crops that already have grown might end up dying[14]. Therefore, one of the most important attribute for our research was temperature. We have collected mainly two types of temperature data. Our data was mostly collected from BRRI and Bangladesh Meteorological Department. Our data started from the year 1972 and went up to 2014. Data collected about the temperature are as follows:

Minimum Temperature Data

This represents the minimum data of certain selected areas. The record of data of every single day of any given year. We have firstly calculated a monthly average of minimum temperature for the areas that have been selected then the average was made for entire year from 1972 to 2014. The following figure (Fig. 3.3) shows minimum temperature data on a daily basis that is being calculated for average values in Microsoft Excel. We will later see the processed data in other sections of this paper.

Maximum Temperature Data

This attribute is same as the minimum temperature data but in this case it had the maximum temperature of any given day. Fig. 3.4 Shows portion of the processed data in Microsoft Excel and Fig 3.9 represents the plot of the temperature data. The data was processed, formatted and merged in Microsoft Excel and then it was converted to CSV.



	Index	Year	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	Monthly Average	Yearly Avera
2	Barisal	1972	1	13.0	13.1	12.2	11.3	10.9	11.7	12.1	11.7	9.3	11	10.4	10.4	11.1	11.7	12.8	13.3	12.8	13	11.1	10.8	11	12.9	12.8	12.2	12	12.3	12.7	13	15	14.2	16.8	12.11935484	21.260886
3	Barisal	1972	2	13.9	13.5	13.8	13.7	13.7	14.5	14	13.4	13.2	9.2	9.8	9.2	11	11.3	12.6	13.1	13.1	13.9	15.6	11.5	9.7	10.6	11.4	12.6	13.7	17.7	14.4	13.5	12.6	11.1	12.5	11.75	13.1129032	21.260886
4	Barisal	1972	3	17.2	14.4	13	14.1	16.6	19.7	18.3	17.5	18.4	21.2	22.1	15.6	16.5	21.1	22.1	21.5	20.8	22.6	23.1	22.4	23.6	23.1	24.1	23.8	22.9	21.1	22.7	24.3	23.9	21.3	25	20.4516129	21.260886
5	Barisal	1972	4	22.6	25.5	25.3	24.1	22.8	21.7	21.7	24.4	22.3	23.1	22.3	23.6	20.1	20.8	19.5	20.1	24.6	25	21.1	21.2	21.3	25.9	26.1	26.1	26.6	26.7	27.1	27.2	27.1	26.7	25.9	21.79032258	21.260886
6	Barisal	1972	5	26.8	26.1	23.2	24.2	25.1	26.4	26.7	25.1	22	26.7	27	27.3	26.6	26.9	26.8	27.1	27.2	26.9	29.1	29.1	25.1	23.7	27.5	22.1	26.3	24.8	25.1	27.8	28.1	26.3	26.4	26.7290323	21.260886
7	Barisal	1972	6	26.7	26.6	26.4	27.1	27.6	28.4	28.6	23.6	28	27.5	28.7	26.7	24.2	23.6	25.7	22.1	22.8	22.6	27.7	28.6	25.9	26.7	23.6	23.9	24.6	25.3	26.5	26.7	27.1	27.1	26.8	26.0451613	21.260886
8	Barisal	1972	7	26.3	26.1	25.8	25.7	26.1	25	25.5	25.4	25.9	26.6	27.2	28.1	26.6	23.2	26.8	25.6	24.9	24.6	25.6	26.3	27	25	25.1	25.6	27.6	27.1	27.5	28.1	26.6	27.7	26.6	26.1677494	21.260886
9	Barisal	1972	8	26.1	26	25.4	26.3	24.8	23.6	24.8	26.7	27.2	27.1	26	25.8	24.6	22.1	23.3	24	24.1	24.6	25.2	26.6	27.3	25.6	25.9	24.6	25	24.5	25.4	23.3	25.6	25	22.3	25.1064516	21.260886
10	Barisal	1972	9	24.4	26.4	26.3	24.8	24.6	25.6	26.1	27.1	23.3	27.8	24.6	24.1	25.6	25.7	25.7	25.6	26.8	26.7	27.5	26.7	27.2	26.8	26.7	25.8	26.5	26.4	26.8	26.1	26.5	25.7	24.5	25.9483871	21.260886
11	Barisal	1972	10	25.1	26.5	25.1	25.8	23.3	23.5	25.2	24.3	25.1	24.7	25.5	22.7	22.5	23.6	23.7	23.8	22.8	23.3	23.6	22.6	23	22.2	21.7	21.7	20.6	18.9	20.6	21.9	22.2	23.3	20.7	23.20967742	21.260886
12	Barisal	1972	11	21.4	20.3	19.1	21.7	19.4	20.3	20.6	19.6	17.4	17.2	17.2	18.3	17.2	17.2	17.5	16.6	15.8	16.7	16.1	16.1	18.1	15.9	17.2	20	19.8	17.6	22.2	20.1	19.1	17.5	16.5	18.3458129	21.260886
13	Barisal	1972	12	16.5	15.6	18.5	14.4	12.7	13.2	12.6	12	11.6	12.1	14.4	14.6	15.6	14.4	11.2	11	12	12.6	12.6	12.5	12.6	16	21.4	18	17.1	16.2	16.4	15.1	17.9	18.4	17.2	14.72258065	21.260886
14	Bhola	1972	1	13	13.2	11.7	12	11.5	12	12.2	9.5	9.5	9.9	10.7	10.9	11.4	11.7	11.9	13.7	12.2	12	11.1	11.2	11.2	12.2	12.2	12.5	12	12.5	13.9	13	15	14.2	16.5	12.1458129	21.1857526
15	Bhola	1972	2	13.5	17.9	20.5	16.8	11.8	16.2	13.6	10	9	9	9.5	9.9	11.5	15	12.7	12.3	13.4	15.1	11.3	9.9	10.4	11.6	12.5	14.7	18	15.3	13.3	13	13.3	13.5	12.5	13.12903226	21.1857526
16	Bhola	1972	3	16.5	12.2	13.3	13.5	16.5	19.4	19.6	16.8	16.8	20.6	21.7	15.5	16.3	21.3	21.8	22	20.5	23.3	23.5	21.5	24	23	23.5	23	22.8	22.4	22.4	24.9	22.8	23.9	23.1	20.2387068	21.1857526
17	Bhola	1972	4	22	25.5	22.8	23.8	22.5	21	20	21.1	23.8	22.7	22.3	24	19.9	21.5	19.6	20.6	22.7	24.8	20.4	20.8	20.6	25.2	25.6	26.4	27	26.8	27.5	25.1	27.2	27.2	27.2	23.46709677	21.1857526
18	Bhola	1972	5	26.7	26.1	23.2	23.8	23.8	26.4	26.5	27.8	27.6	26.8	27.3	27.5	26.3	26.8	27	27	27.7	28.8	28.3	29.4	21.3	21.5	27	21.5	25.5	25.8	26.5	26.2	27.7	26.4	26	26.1354839	21.1857526
19	Bhola	1972	6	26.9	25.7	25.7	26	27.3	27.5	28.1	24.6	27.3	27.5	28.5	25.4	24.2	23.8	25.1	22.4	21.8	21.7	27.4	26.9	27	24.6	24.6	23.8	24.6	24.5	26.4	26.5	26.9	26.8	26.8	25.68709677	21.1857526
20	Bhola	1972	7	26.5	26.5	26.8	26	26.5	26.9	25.3	25.3	25.6	26.8	26.5	26.8	27	25.7	26.6	25	25	25	26.1	26.5	27	25.3	26.9	27	26.2	27.4	27.3	28	26.3	24.9	25.6	26.2325806	21.1857526
21	Bhola	1972	8	26.5	27	25	26.4	24.8	23.5	24.5	26.2	26.6	26.4	24.8	25.8	24.4	25.7	23.3	23.4	24	24.5	24.2	25.7	27	25.1	24.6	23.6	24.4	24.5	24.5	23	24	25.9	27	25.04193548	21.1857526
22	Bhola	1972	9	26.2	26.3	25.7	24.7	25.5	25.8	26.5	27.8	23.7	27	25.1	25.1	25.3	26.3	25.5	25.9	26.4	26.9	27.4	27.4	27.6	26.8	25.8	25.3	25.7	26.2	26.5	26.9	26.5	26.2	25.2	26.07096774	21.1857526
23	Bhola	1972	10	24.2	25.7	25.4	26	22.6	23	25.4	21.2	24.8	25.2	25.8	23.9	23.4	24.4	24.4	24	23.4	23.3	24.2	23.9	23.9	20.9	21.8	22.7	20.7	19.6	21	22.3	23.3	23.7	21.6	23.40967742	21.1857526
24	Bhola	1972	11	22	20.4	19.8	22.1	19.9	20.6	21.3	20.6	17.4	17.8	17.4	18.4	18.2	17.6	17.6	17.2	16.6	17.9	17.2	17.5	19.7	17.6	18.4	20.7	19	19	22.6	20.3	20.4	16.9	17.1	19.00645161	21.1857526
25	Bhola	1972	12	16.4	16.7	17.5	14.7	13.7	14	13.7	14	12.6	12.6	14.6	15	15.7	14.5	12.4	11.5	12.5	13.3	13.6	13.6	14	16.6	14.6	11.4	10.2	11.4	11	11.6	12.3	12.4	14.9	13.64516129	21.1857526
26	Bogra	1972	1	11.5	12.8	11.1	9.4	10.5	11.9	11.7	8.3	10.6	12.5	10.5	10.6	11.1	12	13.6	13.3	11.7	11.2	11.9	11.7	12.2	14.2	11.1	12.2	12.8	13.1	12.5	13.6	15.6	15.6	13.3	12.06774194	20.5995677
27	Bogra	1972	2	14.2	15	16.4	15.8	15.8	15	10.3	8.3	8.3	8.3	9.2	10	13.9	14.6	9.9	9.1	12.8	11.7	8.5	9.4	10.6	11.7	11.1	13.6	15	14.4	12.7	10.6	11.7	11	10.8	11.92580645	20.5995677
28	Bogra	1972	3	13.3	10	11.1	12.9	15.6	17.2	14.7	13.9	15.3	16.1	14.8	15.6	16.9	20.6	20	20.1	17.8	21.7	22.4	23.3	21.7	22.2	19.4	21.7	22.8	22.8	23.9	18.9	20.2	23.9	24	18.60645161	20.5995677
29	Bogra	1972	4	21.1	23.9	21.7	23.3	23.3	20.6	22.2	22.5	22.6	22.2	21.1	22.4	22.8	22.2	20.6	20.6	20	22.8	20.6	22.2	22.5	22.2	24.4	25.1	25.1	25.6	23.3	25	25.8	26.7	25.2	22.36709677	20.5995677
30	Bogra	1972	5	27.2	25.3	23.3	23.9	23.1	25.8	27.8	25.4	27.8	27.2	27.8	27.2	26.7	26.2	26.9	27.2	27.2	28.2	26.8	26.1	22.2	22.8	22.8	20.9	25.8	26.1	26.4	25.6	22.2	26.7	26.7	25.65483871	20.5995677
31	Bogra	1972	6	26.7	27.1	20.9	22.8	22.8	24.4	24.7	26	27.5	27.2	28.9	27.1	26.7	26.7	26.7	22.2	22.5	22.8	25.6	26.6	28.3	24.4	22.8	25	25	25	26.5	26.7	27.2	27.8	26.9	25.53225806	20.5995677
32	Bogra	1972	7	27.2	26.7	26.7	26.7	27.8	26.7	26.1	26.7	27.2	27.2	26.7	27.5	27.2	26	26.4	25.9	25.6	26.7	25.2	26.2	24.4	25.2	25.8	24.8	27.2	27.2	28.1	24.2	24.4	23.7	24.4	26.2483871	20.5995677
33	Bogra	1972	8	25	26.7	26.6	27.3	26.4	25.2	25.6	26.1	25.6	24.4	25.8	25.3	25.6	24.6	25	24.3	25.3	25.7	26.1	27	28.6	25.6	26.7	26.1	25.6	26.1	25.6	24.6	25	25.6	26.1	25.78064516	20.5995677
34	Bogra	1972	9	26.1	26.7	25.6	24.4	23.9	22.7	25.1	26.7	24.6	25.5	23.8	25.3	25.7	26	27.2	23.2	26.6	27.7	27.2	27.8	27.2	25.6	24.4	26.4	26.1	25.6	25.6	24.3	22.9	21.7	25.44483871	20.5995677
35	Bogra	1972	10	22.2	23.5	23.5	25	25	23.2	23.2	24.1	25	23.3	22.2	21.7	23.3	23.9	24.4	24.3	23.9	23.1	21.9	22.2	21.1	22.2	21.7	21.1	21	21	21.4	22.2	20.6	19.3	22.76129032	20.5995677
36	Bogra	1972	11	18.1	18.3	19.4	18.9	18.9	19.3	19	17.8	17.2	17.8	17.1	17.2	16.9	16.1	16.3	17.1	16.9	16.1	16.1	15.6	16.2	16	16.4	17.8	16.7	19.2	17.9	16.1	16.4	16.7	15.2	17.3458129	20.5995677
37	Bogra	1972	12	16.1	15.6	14.7	13.3	12.8	12.2	16.4	14.4	15.6	14.4	15.6	13.5	10	12.5	13.3	13.7	13.1	12.9	13.1	13.9	13.3	11.2	9.4	8.9	9.7	9.7	10.6	10.6	10.2	11.1	12.79354839	20.5995677
38	Chittagong	1972	1	15.1	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15.00322581	21.4139784
39	Chittagong	1972	2	15.6	15.6	16.4	16.2	17.3	16.9	16.8	12.7	10.6	10.6	11.7	12.1	13.3	13.9	14.1	13.3	15.6	17.8	14.3	11.8	12.2	13.9	14.2	16.1	17.2	18.3	14.4	13.9	14.4	15.1	15.1	14.75483871	21.4139784
40	Chittagong	1972	3	15.4	14.4	15.6	14.4	15.6	16.9	19	18.9	17.8	18.1	19.3	17.9	17.5	19.2	20	20	20.8	22.2	21.7	23.2	22.7	22.8	23.3	23.3	22.2	23.1	22.6	21	21.3	23.9	25	20.03548387	21.4139784
41	Chittagong	1972	4	23.9	24.6	20.4	16.4	23.2	18.4	19.6	21.9	22.9	23.8	23.3	23.8	19.3	19.4	22.3	22.2	22.9	19.1	19.1	24.3	24.4	25	25.6	26.1	26.4	26.8	26.1	22.2	25.6	26.6	26.6	22.97419355	21.4139784
42	Chittagong	1972	5	26.7	25.8	22.4	25.6	25.7	25	25.8	26.2	27.6	26.7	26.1	26.1	25.6	26.3	26.7	26.8	27.2	27.8	22.8	27.8	20.4	21.7	25	23.6	26.1	25.8	24.3	25.6	30.6	27.8	25.3	25.78064516	21.4139784
43	Chittagong	1972	6	26.4	26.2	26.4	27.2	26.8	26	27.5	28.1	27.7	28.3	25.6	23.9	25.4	24.4	24.2	22.2	21.9	21.8	25.8	25.7	25.6	25.3	24.7	25	25	25.4	25.7	25.6	26.1	26.2	26.2	25.55666452	21.4139784
44	Chittagong	1972	7	26.1	25.6	25.7	26.1	26.7	25.6	25	24.6	24.7	25.3	25.3	26.7	26	26.7	25.7	25	27.2	25.1	25	25.4	25	25.7	25.8	26.4	26.4	24.4	25.4	25.7	25.1	26	26.1	25.66129032	21.4139784
45	Chittagong	1972	8	25.2	24.6	24.7	26	24.6	24.6	24.7	24.7	25.8	24	25.1	24.4	24.3	25	23.9	23.7	22.8	24.2	25.4	25.6	25.6	24.9	25	23.9	24.3	24.8	23.8	22.8	23	24.9	23.2	24.5	21.4139784
46	Chittagong	1972	9	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	23.4	21.4139784
47	Chittagong	1972	10	24.7	26	25.6	24.3	23.3	24.3	24.4	24.4	24.7	24.4	26.2	24.4	23.9	23.3	23.3	23.7	24.2	23.8	23.6	23.4	21.4	20.6	21.9	21.3	21.4	21.4	22.8	26.6	23.9	24.4	21.4	23.64516129	21.4139784
48	Chittagong	1972	11	23.1	21.9	20.8	22.2	20.7	21.2	21.4	21.3	19.2	17.2	17.8	19.6	21.1	19.3	17.6	18.3	19	18	17.9	21.4	22.7	20.4	21.6	21.6	19.9	22.1	22.9	22.2	22.8	21.7	21.8	20.5458129	21.4139784
49	Chittagong	1972	12	20.6	20.4	17.5	16.8	16.8	16.2	14.8	15.7	15.1	15.2	16.1	17	16.9	16.4	14.6	13.3	13.3	13.9	14.4	15.2	15.3	15.4	15	13.6	12.9	12.2	11.8	12.4	13.6	13.9	14.4	15.16387097	21.4139784
50	Comilla	1972	1	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	22.8807795
51	Comilla	1972	2	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.1	22.8807795

Fig. 3.7 Screen Capture of Temperature Data in Excel

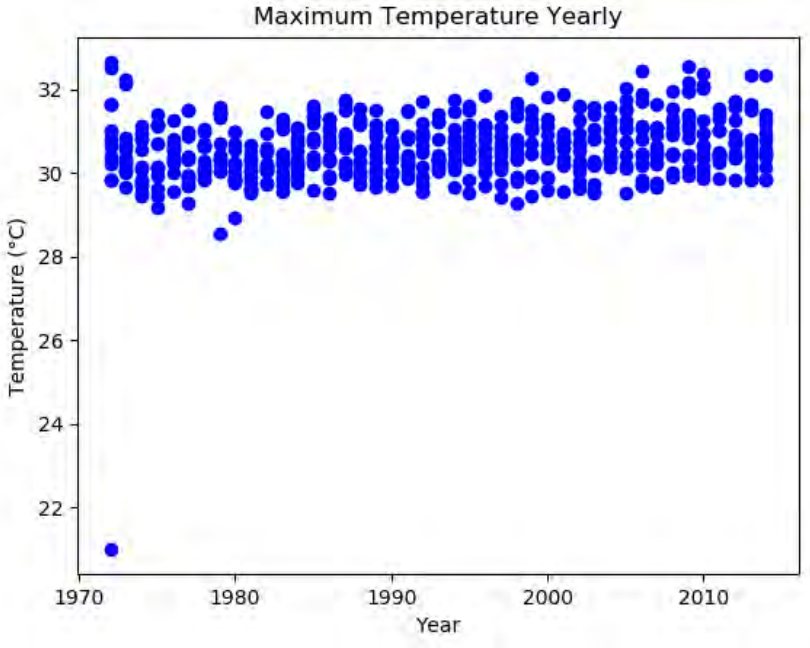


Fig. 3.8 Plot of All Maximum Temperature Values(Before Running algorithm)

1	area	year	month	max_temp	Yearly Avg
2	Barisal	1972	1	25.93	30.8024138
3	Barisal	1972	2	25.56428571	30.8024138
4	Barisal	1972	3	32.39333333	30.8024138
5	Barisal	1972	4	33.15714286	30.8024138
6	Barisal	1972	5	33.54642857	30.8024138
7	Barisal	1972	6	32.65666667	30.8024138
8	Barisal	1972	7	31.75357143	30.8024138
9	Barisal	1972	8	29.89655172	30.8024138
10	Barisal	1972	9	32.52857143	30.8024138
11	Barisal	1972	10	31.52666667	30.8024138
12	Barisal	1972	11	29.87333333	30.8024138
13	Barisal	1972	12	27.73636364	30.8024138
14	Bhola	1972	1	25.97741935	30.4135942
15	Bhola	1972	2	26.08965517	30.4135942
16	Bhola	1972	3	31.89354839	30.4135942
17	Bhola	1972	4	32.56206897	30.4135942
18	Barisal	1973	5	31.64516129	30.3064021
19	Barisal	1973	6	30.57777778	30.3064021
20	Bhola	1972	7	31.27096774	30.4135942
21	Bhola	1972	8	29.33870968	30.4135942
22	Bhola	1972	9	32.34137931	30.4135942
23	Bhola	1972	10	31.94516129	30.4135942
24	Bhola	1972	11	30.41785714	30.4135942
25	Bhola	1972	12	28.07419355	30.4135942
26	Bogra	1972	1	25.2	31.008333
27	Bogra	1972	2	25.75517241	31.008333
28	Bogra	1972	3	33.30967742	31.008333
29	Bogra	1972	4	36.41428571	31.008333
30	Bogra	1972	5	36.30645161	31.008333

Fig. 3.9 Processed Temperature data in Excel before being converted in to CSV format

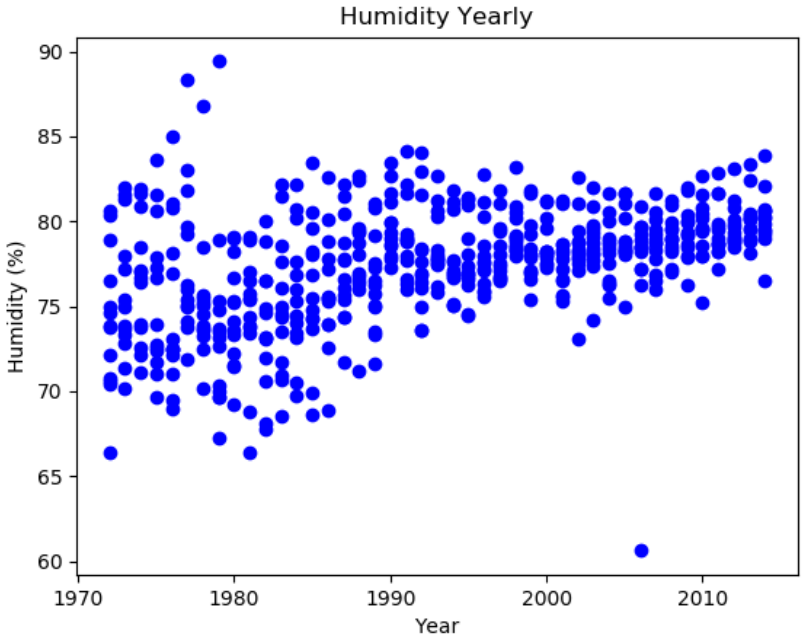


Fig. 3.10 Humidity data before using in algorithm

3.3.2 Humidity

Humidity is a term used to describe the amount of water vapour present in air. However, This is another important element of weather and environment and also it is one of the important factors or agricultural production. In many researches it has been found that humidity has impact on growth and survival of crops and other plants[15] [15]. Some crops cannot survive if the humidity of the air is too low, meaning the air is dry or if its too high. Therefore, we have also included this attribute as a parameter for our research. Humidity is usually measured in percentage because it represents the amount of water that is present in unit of air. Our humidity data started from the year 1972 and ranged up until the year 2014. Data was collected for specific selected areas that have higher production of our selected crop. As same as the temperature data, humidity data was also averaged on a monthly basis and then it was converted in to a yearly average. The final yearly average was used in our algorithm. As done before, data was processed, formatted and merged in Microsoft Excel and then it was converted to CSV. Fig. 3.5 shows the overall plot for all the used humidity data.

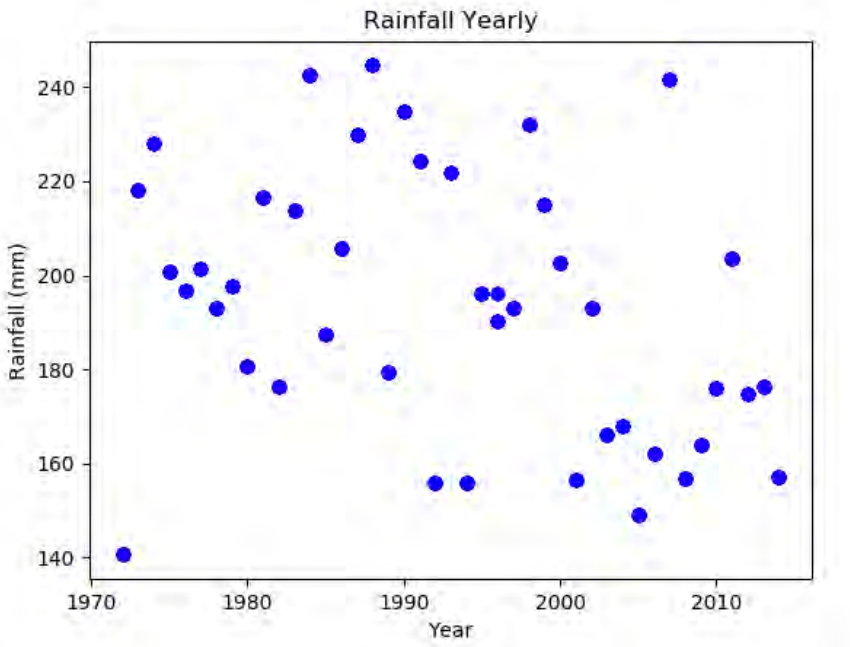


Fig. 3.11 Plot of all rainfall data

3.3.3 Total Amount of Rainfall

Total amount of rainfall represents the rainfall occurring inside the country for any given year. Rainfall is also a factor of crop production. Appropriate amount of rainfall can be very much beneficial for harvest or cultivation [16][16] . However, if the rainfall is too much it ca do the complete opposite resulting in flood which instantly kills the production as water gets stuck to the roots of the cross and stays there.

On the other hands, some crop needs to have extreme irrigation and these mainly depends on heavy rain for their successful cultivation. [17][17]. Fig. 3.6 represents the plot of rainfall data before running in any algorithm.

3.3.4 Area and Amount of Production

The most important factor of production is area. Without land, we cannot have any kind of agricultural production. Therefore, the amount of land is a major and maybe the most important factor of production. Not only that, the quality of the soil also falls in to this category. The total amount of production was also found from this same source and this made our work more easier and reliable as the total amount of production and total land area are collected from the same source. It has been found that the data was originally collected

Table:1.2.8											
Sl. No.	Districts Name	1981-82									
		Broadcast Aman		Local Transplant		HYV		Pajam		Total	
		Area (acres)	roduction (tons)	Area (acres)	roduction (tons)	Area (acres)	roduction (tons)	Area (acres)	roduction (tons)	Area (acres)	roducti n (tons)
1	Bandarban	0	0	0	0	0	0	0	0	0	0
2	Chittagong	8690	3195	187585	94895	144600	116710	210770	137520	551645	352320
3	Comilla	528100	221365	109340	51695	103665	69885	170945	98975	912050	441920
4	Khagrachari	0	0	0	0	0	0	0	0	0	0
5	Noakhali	55295	22690	466325	184850	33255	23420	162275	75890	717150	306850
6	Rangamati	475	210	10475	5060	54095	31580	22050	13530	87095	50380
7	Sylhet	422000	175185	636990	280820	28600	19965	50000	33065	1137590	509035
8	Dhaka	418560	198825	144625	64660	37135	24775	52290	26915	652610	315175
9	Faridpur	620900	179520	43545	17310	6865	4955	12280	8490	683590	210275
10	Jamalpur	82420	34760	384150	157360	42200	22695	26440	13295	535210	228110
11	Kishoregonj	90390	38020	300430	115890	46100	45235	140890	108075	577810	307220
12	Mymensingh	27295	10530	377065	138525	95770	72125	145555	90905	645685	312085
13	Tangail	239405	90325	173625	67360	20580	14120	19155	12835	452765	184640
14	Barisal	178380	75495	673655	275700	67370	52175	3165	1940	922570	405310
15	Jessore	349135	123775	255165	104055	45040	29900	15895	9910	665235	267640
16	Khulna	81135	31925	801165	359080	68950	55070	33975	19295	985225	465370
17	Kushtia	64000	19845	38325	16965	38000	31760	145	70	140470	68640
18	Patuakhali	0	0	665875	298690	16920	14385	0	0	682795	313075
19	Bogra	1425	525	499330	250400	114425	96685	5365	3545	620545	351155
20	Dinajpur	4210	1730	801740	387615	55010	44725	19905	8775	880865	442845
21	Pabna	325640	125255	186785	87835	17025	12905	1425	1295	530875	227290
22	Rajshahi	365480	165015	667240	313520	47990	33495	2060	1630	1082770	513660
23	Rangpur	56065	28960	1150140	633805	145405	127350	37660	31405	1389270	821520
	Bangladesh	**3919000**	**1547150**	**8573575**	**3906090**	**1229000**	**943915**	**1132245**	**697360**	**14853820**	**094515**

Fig. 3.12 Scanned document on area and amount of Production

to generate a report about area wise rice production. Therefore, this document or data was a very suitable resource for our research work. Areas that we have selected are those where the most amount of rice is produced. We have selected to work with only a few major rice producing areas of Bangladesh because the data available for the area is very much noisy and they are not in any kind of machine readable format. Processing this data was most time consuming and including more areas would have consumed more resource while running the algorithm which our machine would not be able to handle because of not having enough primary memory and CPU processing power. The data was manually converted to a machine readable format. Fig. 3.4 represents the scanned document that was collected as the resource for our information about total area and total production. The highlighted rows represents the areas that have been selected, filtered and processed for our research and algorithm. Also, Fig. 3.5 represents the processed and merged data set with area and production information.

Sl No.	Districts	Broadcast Aman		Local Transplant		HYV		Total	
		Area (acres)	Production (tons)	Area (acres)	Production (tons)	Area (acres)	Production (tons)	Area (acres)	Production (tons)
1	Bandarban	0	0	0	0	0	0	0	0
2	Chittagong	8690	3195	187585	94895	144600	116710	551645	852320
3	Comilla	528100	221365	109340	51695	103665	69885	912050	441920
4	Khagrachari	0	0	0	0	0	0	0	0
5	Noakhali	55295	22690	466325	184850	33255	23420	717150	806850
6	Rangamati	475	210	10475	5060	54095	31580	87095	50380
7	Sylhet	422000	175185	636990	280820	28600	19965	113759	0509035
8	Dhaka	418560	198825	144625	64660	37135	24775	652610	815175
9	Faridpur	620900	179520	43545	17310	6865	4955	683590	210275
10	Jamalpur	82420	34760	384150	157360	42200	22695	535210	228110
11	Kishoregonj	90390	38020	300430	115890	46100	45235	577810	807220
12	Mymensingh	27295	10530	377065	138525	95770	72125	645685	8712085
13	Tangail	239405	90325	173625	67360	20580	14120	452765	184640

Table 3.1 Unclassified document on area and amount of Production

The next part of the research paper describes our experimental setup, implementation of algorithm and research results.

Chapter 4

Experimental Setup and Analysis

Our research of agricultural yield and profit prediction which applies multiple data analysis algorithm requires heavy computational capability. Our data set had used multiple libraries from the Scikit Learn package. As we have already talked before, our research works tends to implement multiple algorithms and the initial data set was large and contained over a large number of rows. First of all, to begin our research we need to initialize our data.

4.1 Data Preparation

As we said before, our data set was very large. Therefore, we needed to reduce the size of our data set. In order to do this, we averaged out the data for every day of a month initially to a month and then averaged that out to a yearly basis. This reduced the amount of data required to being processed down to a great extent. To put that into perspective, if we consider all days of a month then total number of days we get in a year is 365. We had considered data from 1972-2000 which is a span of 28 years. So the total number of data for only one area would be (365x28).
We have considered to work with 13 areas. So that resulted in a total number of (365x28x13)=132860 However, the technique that we used generated only one average value for one year. Which means for 28 years and 13 areas the data resulted in (28x13)=364, this is for each of the attributes that we considered.
Therefore, for each attribute, the reduction had a ratio of approximately 365:1. This reduced the size of our data set to a great extent which made it easier to process in our machines. Fig. 4.1 represents the final prepared data set in Microsoft Excel. It was then converted to CSV for to be used in Python. Fig. 4.2 shows portion of the CSV file.

Table 4.1 (in page 33) shows a sample dataset used to test models accuracy.

1	Year	District	Rice Type	Area(acress)	g. Min Tem	g. Max Tem	roduction	vg. Humidit	Rainfall
2	1972	Barisa	Aush	56121	21.26	30.80	22875	78.89	140.55
28	1973	Barisal	Aush	570405	21.16	30.31	218675	81.29	218.2234
29	1973	Bhola	Aush	288995	21.77	30.22	93545	82.01	218.2234
30	1973	Bogra	Aush	316915	20.95	30.22	111860	75.40	218.2234
31	1973	Chittagong	Aush	219465	21.69	30.01	159125	78.00	218.2234
32	1973	Comilla	Aush	255630	20.85	30.07	255630	73.32	218.2234
33	1973	Dinajpur	Aush	455420	19.64	30.08	167430	71.34	218.2234
34	1973	Faridpur	Aush	572305	19.88	30.84	176270	77.20	218.2234
35	1973	Ishurdi	Aush	108965	20.32	30.66	53135 s	72.82	218.2234
36	1973	Jessore	Aush	689660	20.85	32.25	263875	73.81	218.2234
37	1973	Khulna	Aush	263875	21.60	30.35	263875	70.16	218.2234
38	1973	Mymensingl	Aush	767605	20.09	32.13	319040	74.98	218.2234
39	1973	Rajshahi	Aush	425855	20.41	29.68	154075	73.69	218.2234
40	1973	Rangpur	Aush	989435	20.48	30.57	295575	81.59	218.2234
41	1973	Barisal	Aman	655175	21.16	30.31	233550	81.29	218.2234
42	1973	Bhola	Aman	711710	21.77	30.22	277315	82.01	218.2234
43	1973	Bogra	Aman	619855	20.95	30.22	265960	75.40	218.2234
44	1973	Chittagong	Aman	569045	21.69	30.01	298870	78.00	218.2234
45	1973	Comilla	Aman	948270	20.85	30.07	407610	73.32	218.2234
46	1973	Dinajpur	Aman	741400	19.64	30.08	327010	71.34	218.2234

Fig. 4.1 Final Data Set in Excel

```
Year,District,Rice Type,Area(acress),Avg. Min Temp,Avg. Max Temp,
Rice Production (tons),Avg. Humidity,Rainfall
1972,Barisal,Aush,561215,21.26,30.80,228750,78.89,140.55464
1972,Bhola,Aush,463375,21.19,30.41,143625,80.61,140.55464
1972,Bogra,Aush,21065,20.60,31.01,58020.00,70.62,140.55464
1972,Chittagong,Aush,197320,21.41,30.22,95810.00,76.52,140.55464
1972,Comilla,Aush,502410,22.88,30.63,169950.00,73.76,140.55464
1972,Dinajpur,Aush,303370,19.64,29.83,103045.00,70.41,140.55464
1972,Faridpur,Aush,435560,19.86,31.65,117600.00,74.60,140.55464
1972,Ishurdi,Aush,324625,19.76,32.67,88315.00,66.36,140.55464
1972,Jessore,Aush,626820,20.46,31.00,203210.00,72.13,140.55464
1972,Khulna,Aush,93475,21.12,21.00,40365.00,73.88,140.55464
1972,Mymensingh,Aush,720075,20.05,29.83,222270.00,74.99,140.55464
1972,Rajshahi,Aush,442885,20.82,31.65,132530.00,70.75,140.55464
1972,Rangpur,Aush,916865,16.07,32.52,278480.00,80.36,140.55464
1972,Barisal,Aman,606070,21.26,30.80,221715,78.89,140.55464
1972,Bhola,Aman,736150,21.19,30.41,281485,80.61,140.55464
1972,Bogra,Aman,598645,20.60,31.01,297250,70.62,140.55464
1972,Chittagong,Aman,546425,21.41,30.22,299685,76.52,140.55464
1972,Comilla,Aman,946545,22.88,30.63,431115,73.76,140.55464
1972,Dinajpur,Aman,792830,19.64,29.83,285855,70.41,140.55464
1972,Faridpur,Aman,592870,19.86,31.65,127105,74.60,140.55464
1972,Ishurdi,Aman,479555,19.76,32.67,140690,66.36,140.55464
1972,Jessore,Aman,574785,20.46,31.00,222240,72.13,140.55464
1972,Khulna,Aman,761105,21.12,21.00,217995,73.88,140.55464
```

Fig. 4.2 Final data Set in CSV

Year	District	Rice Type	Area (sq.km)	Avg. Min Temp (°C)	Avg. Max Temp (°C)	Rice Production (tons)	Avg. Humidity (%)	Rainfall (mm)
1998	Rajshahi	Aush	229670	21.41	30.66	131960	78.35	232.12
1997	Ishurdi	Aush	70120	20.72	30.87	21480	76.48	192.99
1986	Faridpur	Aman	669500	20.33	31.10	255720	75.52	205.63
1972	Jessore	Aman	574785	20.46	31.00	222240	72.13	140.55
1976	Faridpur	Aush	585000	21.64	30.83	171920	72.49	196.862
1981	Bhola	Aman	747285	20.98	30.07	419850	79.11	216.56
1978	Ishurdi	Aush	367750	20.80	30.63	113160	70.19	193
1990	Rangpur	Aush	356415	19.73	30.44	156665	79.91	234.85

Table 4.1 Portion of test dataset

Table 4.2 (in page 34) illustrates a portion of training dataset which was used to train model about the co-relations of features so that it can make predictions about the future outcome.

4.2 Result

We have applied multiple algorithm in our data set in order to determine and compare the performance level of the algorithm. Among the three algorithms that were applied, Linear Regression had shown the lowest accuracy and the worst performance. Linear regression did not have acceptable due to the fact that the data we worked with were completely random and had no pattern in general. Therefore, the algorithm could not generate any acceptable results. The algorithm would have worked well if the data had any pattern to them.
Both Random Forest and Support Vector Regression had performed well in our situation and among them Random Forest had the most overall accuracy throughout the process. Random Forest had performed well for all the attributes followed by Support Vector Regression. The accuracy level of the algorithms are shown in table 4.3 below (in page 34)

In this part, we will see the figures that have been plotted after generating prediction. In these figures blue dots represent the actual input value and red dots represent the predicted values.

Fig 4.3 represents the actual value and the predicted value by using blue and red dots

Year	District	Rice Type	Area (sq.km)	Avg. Min Temp (°C)	Avg. Max Temp (°C)	Rice Production (tons)	Avg. Humidity (%)	Rainfall (mm)
1973	Barisal	Aush	570405	21.16	30.31	218675	81.29	218.22
1973	Chittagong	Aush	219465	21.69	30.01	159125	78.00	218.22
1978	Dinajpur	Aush	458780	21.55	30.72	167045	73.66	193.00
1980	Comilla	Aman	914885	20.94	28.95	482810	76.64	180.71
1977	Jessore	Aman	621995	20.18	31.50	280945	74.10	201.34
1979	Mymensingh	Aman	628310	22.56	30.05	361440	73.64	197.64
1981	Khulna	Aman	989470	20.62	29.58	508675	78.88	216.56
1974	Barisal	Aman	588305	16.26	29.79	136430	81.66	228.01
1976	Ishurdi	Aush	342900	20.54	31.28	105410	68.98	196.86
1977	Rangpur	Aush	994935	21.00	29.91	334430	83.05	201.34
1977	Jessore	Aman	621995	20.18	31.50	280945	74.10	201.34
1976	Barisal	Aman	607385	16.55	30.41	299115	81.07	196.86
1984	Comilla	Aush	246355	21.33	30.53	110890	76.85	242.64
1988	Bogra	Aush	77900	21.86	31.20	26423	78.08	244.65
1985	Ishurdi	Aush	369935	21.09	31.62	104791	69.91	187.52
1994	Rangpur	Aush	263500	20.45	30.10	144330	77.74	155.73
1994	Rangpur	Aush	263500	20.45	30.10	144330	77.74	155.73
2000	Rajshahi	Aman	970660	22.26	30.41	800330	78.06	202.79
1999	Barisal	Aush	354310	14.35	31.15	125450	81.64	214.98

Table 4.2 Portion of training dataset

Item	**Linear Regression**	**Random Forest**	**Support Vector Regression (SVR)**
Minimum Temperature	7.59	40.79	10.24
Maximum Temperature	16.55	24.21	19.07
Yearly Rainfall	11.96	91.82	16.07
Yearly Humidity	43.85	62.91	52.53
Rice Production	N/A	68.49	59.78

Table 4.3 Algorithm Accuracy Comparison

respectively for humidity prediction using Random Forest algorithm. In this figure x-axis represents year and y-axis represents corresponding humidity in percentage (%).

Fig. 4.4 represents the actual value and the predicted value by using blue and red dots

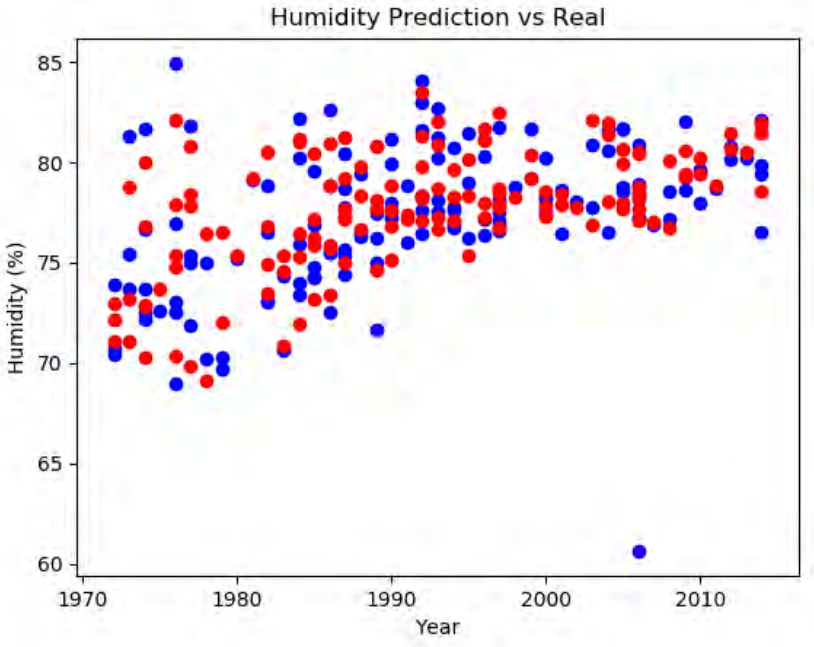


Fig. 4.3 Humidity prediction using Random Forest

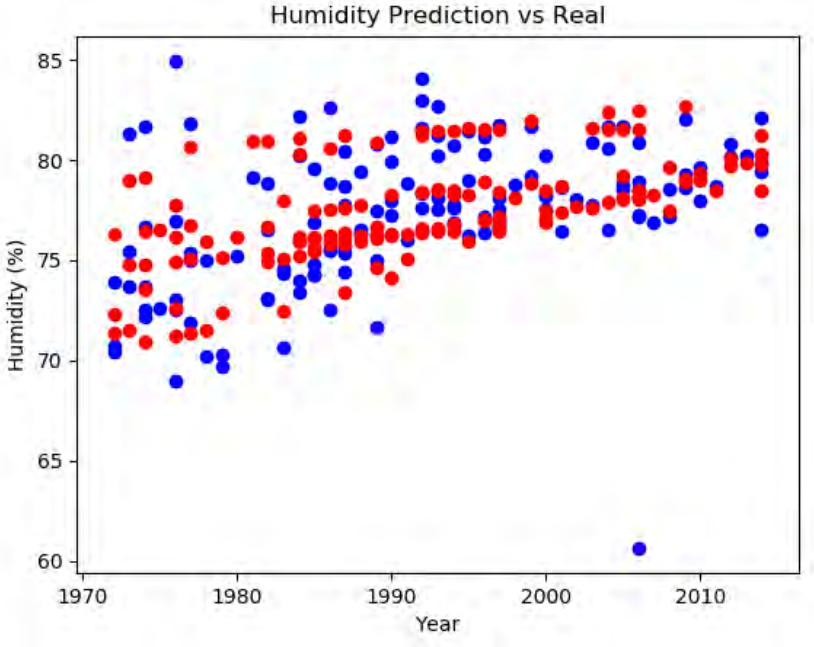


Fig. 4.4 Humidity prediction using SVR

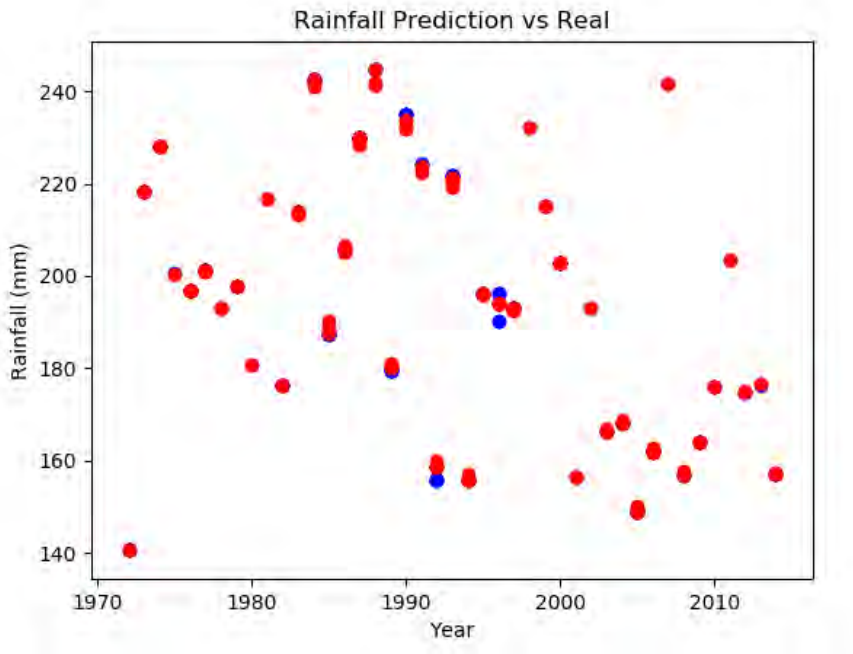


Fig. 4.5 Rainfall prediction using Random Forest

respectively for humidity prediction using Support Vector Regression (SVR). In this figure, x-axis represents year and y-axis represents corresponding humidity in percentage (%).

Fig. 4.5 represents Rainfall prediction using Random Forest algorithm. Where blue dots represent actual values and red dots represent predicted values sequentially. In this figure, x-axis represents the year and y-axis represents corresponding rainfall amount in millimeter (mm).

Fig 4.6 represents Rainfall prediction using Support Vector Regression (SVR). Where blue dots represent actual values and red dots represent predicted values. In this figure x-axis illustrates year and y-axis illustrates corresponding rainfall amount in millimeter (mm).

Maximum Temperature prediction using Random Forest algorithm is represented in Fig. 4.7 . Here blue dots show actual values and red dots show predicted values. x-axis represents year and y-axis represents corresponding temperature in degree Celsius (°C) in this figure.

Fig. 4.8 represents Maximum Temperature prediction using Support Vector Regression (SVR). Where blue dots represent actual values and red dots represent predicted values where x-axis represents year and y-axis represents corresponding temperature in degree Celsius (°C).

Fig. 4.9 states Minimum Temperature prediction using Random Forest algorithm. Here blue dots represent actual values and red dots represent predicted values. In the figure x-axis represents year and y-axis represents corresponding temperature in degree Celsius (°C).

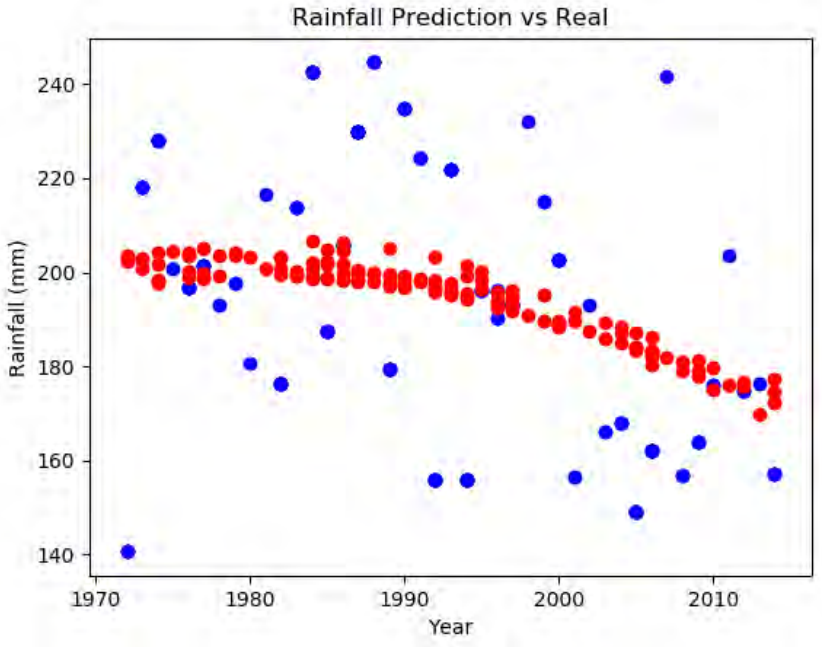


Fig. 4.6 Rainfall prediction using SVR

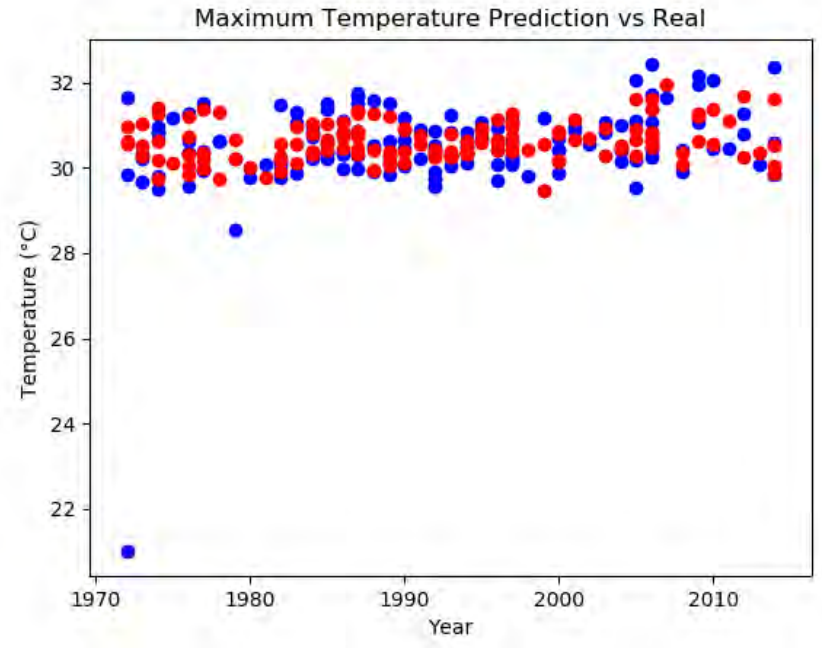


Fig. 4.7 Maximum Temperature prediction using Random Forest

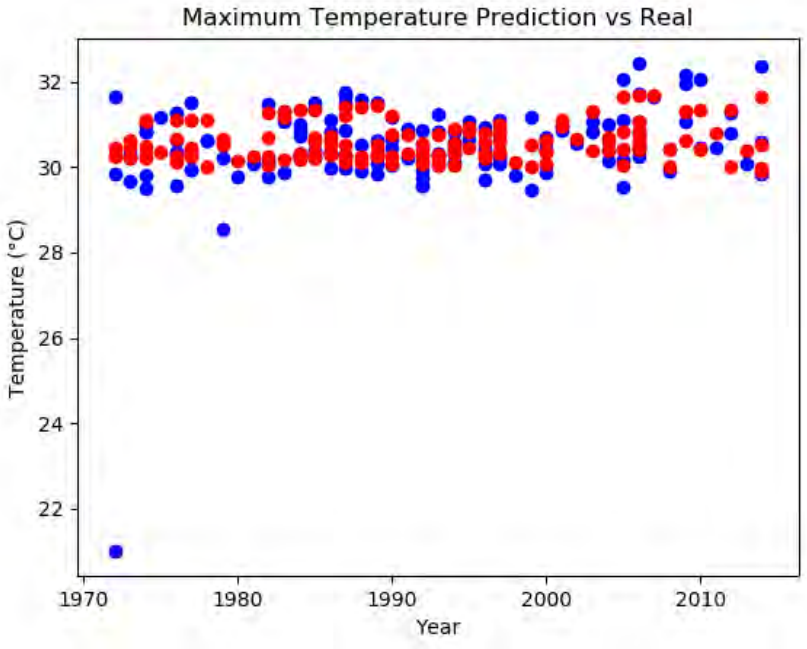


Fig. 4.8 Maximum Temperature prediction using SVR

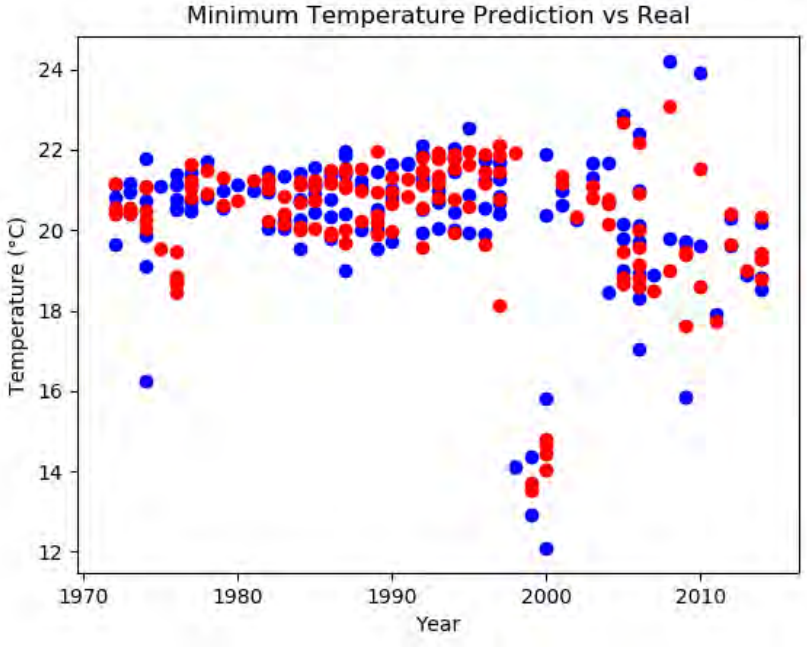


Fig. 4.9 Minimum Temperature prediction using Random Forest

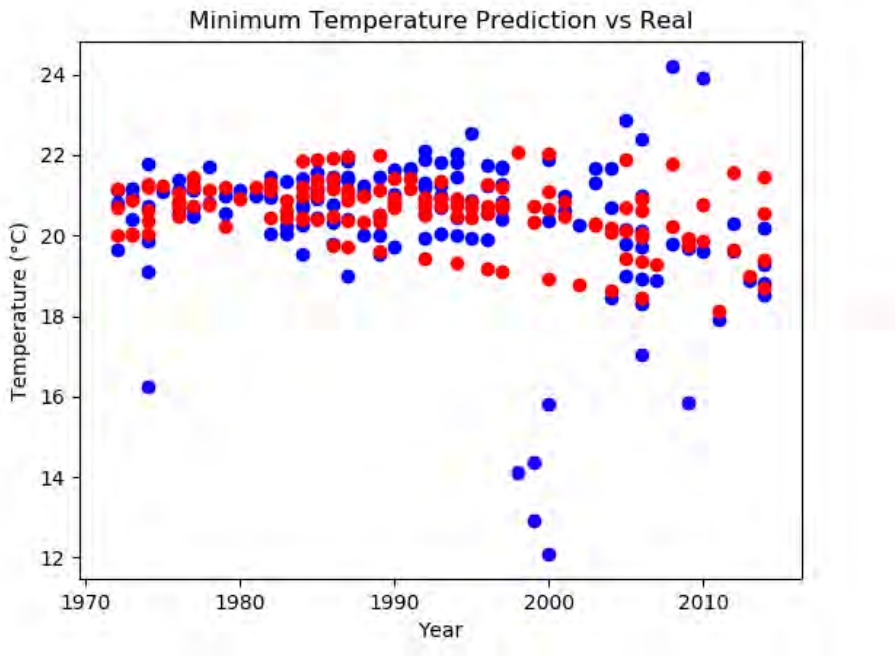


Fig. 4.10 Minimum Temperature prediction using SVR

Fig. 4.10 represents Minimum Temperature prediction using Support Vector Regression (SVR). Where blue dots represent actual values and red dots represent predicted values. In the figure x-axis represents year and y-axis represents corresponding temperature in degree Celsius (°C).

Rice production prediction using Support Vector Regression (SVR) is represented in Fig. 4.11 where blue dots represent actual values and red dots represent predicted values. In this figure x-axis represents the year and y-axis represents corresponding amount of rice production in Tons.

Table 4.4 shows the predicted production of Aus rice for the year 2019 to 2024 in Dinajpur district where Table 4.5 represents the predicted production of Aus rice for the year 2019 in the research areas.

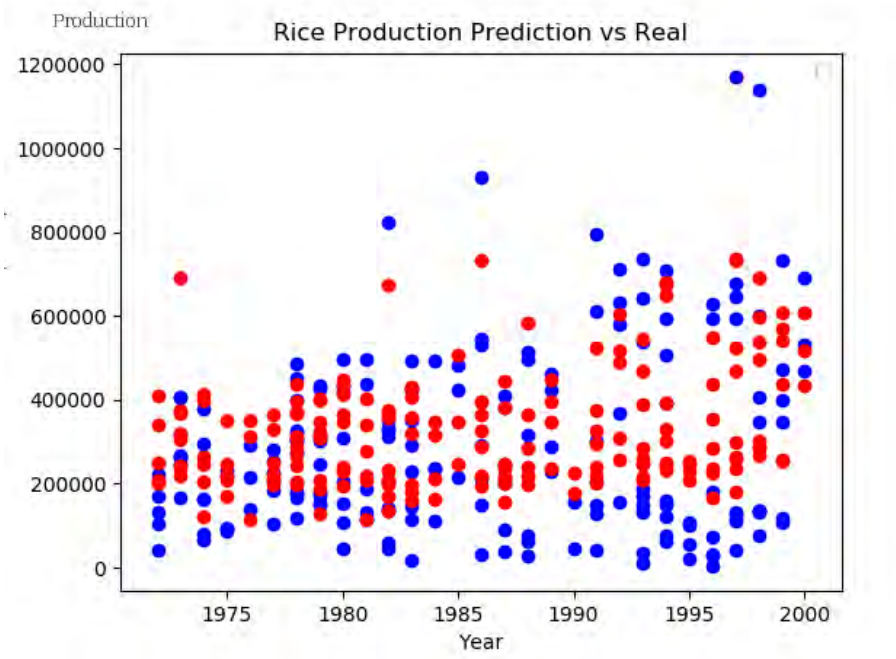


Fig. 4.11 Final Rice prediction using SVR

Area	Year	Rice variant	Predicted Production (Tons)
Dinajpur	2019	Aus	103585
	2020		98562
	2021		114520
	2022		108568
	2023		100685
	2024		99587

Table 4.4 Prediction of Aus Rice for the year 2019 to 2024 in Dinajpur

4.3 Findings

After finishing the research, key findings are that the algorithm performed poorly in many cases because the collected primary data that have been used are noisy to a great extent. Even after cleanup the data has some problems because the data that have been collected by the officials were not consistent and correct. However, the model that we have implemented would work well if correct data are given as input. After generating the final prediction, we found that the production slows down in a very lower ratio with time. Results could have

District	Rice variant	Predicted Production in 2019 (Tons)
Barisal	Aus	45902
Bhola		110751
Bogra		53845
Chittagong		115896
Comilla		153820
Dinajpur		103585
Faridpur		11238
Ishurdi		34825
Jessore		98005
Khulna		4360
Mymensingh		63764
Rajshahi		110630
Rangpur		3568

Table 4.5 Prediction of Aus Rice for the year 2019

been better if we could manage specified and some more attributes of particular crops.
Since this research work is done with real time data, collected from different areas and related online platforms, some odd experiences have to be faced. For example, the supply of Fertilizer and seed, yearly subsidy for crops could have been considered. But proper data set for all areas could not have been managed due to official mismanagement. Furthermore, the collected subsidy and supply of fertilizer and seed sometimes changes mysteriously bypassing the yearly target according to political situation.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this modern world, almost every sector is being enlightened by different technological innovations and findings. Bangladesh is also moving forward with these blessings although the most significant economic resource of our country, agriculture, is lying behind. In this research, the model proposed, analyses some real time data of Aus and Aman rice, two major types of rice and predict the yield, by which the profit intended for future production can be calculated. Linear Regression, Random Forest and Supervised Machine Regression algorithms has been used to demonstrate the prediction. Using multiple algorithms helped to understand which algorithm is more suitable for this system. Throughout this research and field work, it is found that the unavailability of expected data for different agricultural crops is the major problem for such analytically implementation. In future, this model can be implemented as web and mobile applications so that the agriculture department of the country and field workers can use the predicted result before planing of a production and have a convenient profit margin.

5.2 Future Work

The model currently works with temperature, humidity, rainfall, area and total amount of production as the working parameter. However, there are other factors of productions also. In the future, the plan is to integrate the other factors that have impact on the total outcome. Also, in the current model we have predicted only the production and the prediction of profit could not have been done due to data not being available. Therefore, we will try to collect the data and include that part in the model as well. This model was implemented only on the

production of rice to begin with. However, it is also suitable for other crops or products as well. The only drawback here is that we do not have any data set ready for other crops. In fact, the research took so long to come to a conclusion because of the lack of proper data set. Therefore, our focus will also be on creating a proper database for the research and make it open source and available to the public. Doing so would enable any researcher who wants to contribute in this field, have easier access to data. Future researchers would not need to go through the lengthy and hectic process of primary data collection and data processing. Our current data set can be considered as the base of the future work and any future contribution can be made based on this structure. We plan to develop an API that would give easier access to the collected data for anyone to use. We hope to contribute in the betterment of the country by doing and carrying our research forward in the future.Our current data set can be considered as the base of the future work and any future contribution can be made based on this structure. We are planning to develop an API that would give easier access to the collected data for anyone to use. We hope to contribute in the betterment of the country by doing and carrying our research forward in the future.

References

[1] J. Ming, L. Zhang, J. Sun, and Y. Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 224–227, April 2018.

[2] W. Minghua, Z. Qiaolin, Y. Zhijian, and Z. Jingui, "Prediction model of agricultural product's price based on the improved bp neural network," in *2012 7th International Conference on Computer Science Education (ICCSE)*, pp. 613–617, July 2012.

[3] M. T. Shakoor, K. Rahman, S. N. Rayta, and A. Chakrabarty, "Agricultural production output prediction using supervised machine learning techniques," in *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, pp. 182–187, July 2017.

[4] P. Vinciya and D. 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, no. 5, pp. 481–488, 2016.

[5] W. Lin, G. Dean, and C. Moore, "An empirical test of utility vs. profit maximization in agricultural production," *American Journal of Agricultural Economics*, vol. 56, no. 3, pp. 497–508, 1974.

[6] H. Lee and A. Moon, "Development of yield prediction system based on real-time agricultural meteorological information," in *Advanced Communication Technology (ICACT), 2014 16th International Conference on*, pp. 1292–1295, IEEE, 2014.

[7] S. Bagis, B. B. Ustundag, and E. Ozelkan, "An adaptive spatiotemporal agricultural cropland temperature prediction system based on ground and satellite measurements," in *2012 First International Conference on Agro- Geoinformatics (Agro-Geoinformatics)*, pp. 1–6, Aug 2012.

[8] S. Swaminathan, "Linear regression—detailed view," *Towards Data Science*, 2014.

[9] D. S. Sayad, "Support vector machine - regression (svr)," *Predicting the Future*, 2011.

[10] N. Donges, "The random forest algorithm," *Towards Data Science*, 2015.

[11] T. M. L. Platform, "Gradient boosted regression trees," 2013.

[12] R. Mendelsohn, W. D. Nordhaus, and D. Shaw, "The impact of global warming on agriculture: A ricardian analysis," *The American Economic Review*, vol. 84, no. 4, pp. 753–771, 1994.

[13] W. Wang, B. Vinocur, and A. Altman, "Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance," *Planta*, vol. 218, no. 1, pp. 1–14, 2003.

[14] A. Challinor, T. Wheeler, P. Craufurd, C. Ferro, and D. Stephenson, "Adaptation of crops to climate change through genotypic responses to mean and extreme temperatures," *Agriculture, ecosystems & environment*, vol. 119, no. 1-2, pp. 190–204, 2007.

[15] R. Grange and D. Hand, "A review of the effects of atmospheric humidity on the growth of horticultural crops," *Journal of Horticultural Science*, vol. 62, no. 2, pp. 125–134, 1987.

[16] N. Dastane *et al.*, *Effective rainfall in irrigated agriculture*, vol. 25. Food and Agriculture Organization of the United Nations Rome, 1974.

[17] D. Pimentel, J. Houser, E. Preiss, O. White, H. Fang, L. Mesnick, T. Barsky, S. Tariche, J. Schreck, and S. Alpert, "Water resources: agriculture, the environment, and society," *BioScience*, vol. 47, no. 2, pp. 97–106, 1997.