

# **Bangladeshi Paddy Yield Prediction Using Machine Learning**

**BY**  
**Md. Istiaq Ahmed**

**ID: 191-15-12714**

**AND**

**Sadia Afrin Santa**

**ID: 191-15-12445**

**AND**

**Md. Abu Sufian Tutul**

**ID: 191-15-12784**

This Report Presented in Partial Fulfillment of the Requirements for the Degree of  
Bachelor of Science in Computer Science and Engineering

Supervised By

**Md. Abbas Ali Khan**

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

**Md. Azharul Islam Tazib**

Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2023**

## **APPROVAL**

This Project titled “**Bangladeshi Paddy Yield Prediction Using Machine Learning**”, submitted by Md. Istiaq Ahmed, ID: 191-15-12714, Sadia Afrin Santa, ID: 191-15-12445 and Md. Abu Sufian Tutul, ID: 191-15-12784 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on *24 January, 2023*.

### **BOARD OF EXAMINERS**

**Chairman**

---

**Dr. Touhid Bhuiyan**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Abdus Sattar**  
**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Fatema Tuj Johra**  
**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**External Examiner**

---


**Dr. Dewan Md Farid**  
**Professor**

Department of Computer Science and Engineering  
United International University

## DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Md. Abbas Ali Khan, Assistant Professor, Department of CSE**, and co-supervision of **Md. Azharul Islam Tazib, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



22.01.2023

---

**Md. Abbas Ali Khan**

Assistant Professor  
Department of CSE  
Daffodil International University

Submitted by:



---

**Md. Istiaq Ahmed**

ID: 191-15-12714  
Department of CSE  
Daffodil International University



---

**Sadia Afrin Santa**

ID: 191-15-12445  
Department of CSE  
Daffodil International University



---

**Md. Abu Sufian Tutul**

ID: 191-15-12784  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First of all, we want to render our gratitude to the Almighty Allah for the enormous blessing that makes us able to complete the final thesis successfully.

We are really grateful and express our earnest indebtedness to **Md. Abbas Ali Khan, Assistant Professor**, Department of CSE Daffodil International University, Dhaka, Bangladesh. The Profound Knowledge & intense interest of our supervisor in the field of “*Machine Learning & Deep Learning*” made our way to carry out this thesis very smoothly. His remarkable patience and dedication, scholarly guidance, continual encouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, and great endurance during reading many inferior drafts and correcting the work to make it unique paved the way of work very smooth and ended with a great result.

We would like to express our gratitude wholeheartedly to **Prof. Dr. Touhid Bhuiyan**, Professor, and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank the fellow Daffodil International University student, who participated in this discussion during the completion of this work.

We would like to express our immense thanks to the Different food application for visible user original reviews as a result we collected raw data to make our work possible.

We would also like to thank the people who provide the data done by us to collect the market real information.

Finally, we must acknowledge with due respect the constant support and passion of our parents and family members.

## **ABSTRACT**

Rice is the staple food for Bangladesh's population of 135 million. It provides two-thirds of the region's total calorie supply and half of its total protein consumption, supports nearly half of all rural jobs, and accounts for almost half of all land area. The rice industry accounts for 16% of Bangladesh's GDP and 50% of the agricultural GDP. Most of the country's 13 million farming households produce rice. The area used for growing rice, at over 10.5 million hectares, has been fairly stable over the previous three decades. In Asia, rice takes up around 75% of all farmland and 80% of all irrigated land. Accordingly, rice is extremely important to the diets of the people of Bangladesh. Several rice types are the primary focus of this study. Aus, Aman, and Boro are their names. Yield Predictions for Aus, Aman, and Boro Rice via Data Mining and ML.

Six different regression methods were used to forecast the harvest of these plants. Ridge regression, linear regression, random forest regression, boosted regression, decision tree regression, and neural network regression are only some of the regression methods we've tested. In addition, seven different algorithms were evaluated for their ability to estimate Rice yield, and Random Forest Regression emerged as the clear victor. Our findings will pave the way for more precise estimates of future Bangladesh rice, wheat, and potato harvests.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Acknowledgments	iv
Abstract	v
List of Figure	viii
List of Table	ix

## **CHAPTER**

<b>CHAPTER 1: INTRODUCTION</b>	<b>PAGE NO.</b>
	<b>1-4</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Problem Definition	2
1.4 Research Questions	3
1.5 Research Methodology	3
1.6 Research Objective	3
1.7 Report Layout	
<b>CHAPTER 2: BACKGROUND</b>	<b>4-8</b>
2.1 Introduction	4
2.2 Related Work	4
2.3 Summary	8

<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>9-18</b>
3.1 Introduction	9
3.2 Data collection	10
3.3 Data Preprocessing	10
3.4 Algorithm Selection	14
3.5 Algorithm Implementation	16
3.6 Evaluation	16
<b>CHAPTER 4: RESULT ANALYSIS</b>	<b>19-28</b>
4.1 Introduction	19
4.2 Experimental Result	19
4.3 Analysis	25
<b>CHAPTER 5: SUMMARY, CONCLUSION, AND FUTURE WORK</b>	<b>29-30</b>
5.1 Conclusion	23
5.2 Future Work	23
5.3 Limitation	23
<b>REFERENCES</b>	<b>31</b>
<b>APPENDIX</b>	<b>33</b>
<b>PLAGIARISM REPORT</b>	<b>34</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO.</b>
Figure 3.1: Methodology diagram	9
Figure 3.2: Dataset Connection	12
Figure 3.3: Correlation of data	13
Figure 3.4: Representation of Yield vs Crop name	17
Figure 3.5: Representation of Yield vs Division	18
Figure 4.1: Aman Real vs. Aman Predicted	25
Figure 4.2: Aus Real vs. Aus Predicted	26
Figure 4.3: Boro Real vs. Boro Predicted	27



## LIST OF TABLE

<b>TABLE</b>	<b>PAGE NO.</b>
Table 3.1 Correlation analysis of the dataset	12
Table 3.2 Level Encoding Representation of Data	14
Table 4.1 Show the Gradient Boosting Algorithm Accuracy	19
Table 4.2 Show the MLP Regressor Algorithm Accuracy	20
Table 4.3 Show the Linear Regressor Algorithm Accuracy	21
Table 4.4 Show the Decision Tree Algorithm Accuracy	22
Table 4.5 Show the RF Regressor Algorithm Accuracy	23
Table 4.6 Show the Ridge Regressor Algorithm Accuracy	24

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

The agricultural industry plays a vital role in Bangladesh's economy because of the country's reliance on agriculture. Maintaining a sustainable agricultural system that is both economical, risk-free, and ecologically friendly is essential if citizens are to have access to enough food in the future. Primary foods consist of rice, jute, fish, and other fruits and vegetables. Wheat harvests have become more fruitful in recent times. The rising demand for chicken has resulted in a higher maize harvest. Textiles, leather goods, earthenware, and ready-to-wear apparel are also domestically produced. Several varieties of rice paddies can be found in Bangladesh. Aus, Aman, and Boro are the most typical kinds of paddies. Potatoes are a significant crop in Bangladesh, alongside rice and wheat. Vegetable use accounts for the vast majority of its consumption in Bangladesh. It is the country's fourth most important crop because of its nutritional value and high output. Therefore, it is crucial to provide careful attention to agricultural aspects to maximize each crop's yield. Boro produced 54% of the nation's rice in 2020-2021, with Aman producing 38% and Aus producing 8%, as reported by BBS. This year, the Department of Agriculture and Environment predicts that 14,000 acres of land will produce 34,000 tons of Aus and that 2,000,000 tons of Boro will be harvested from 47.54 million hectares of land. [1]. Regarding rice production, Bangladesh is in the top five. Despite a decline in arable land, rice harvesting has increased since 1971. Between 1995 and 2010, the annual rice harvest rose from 2.70 to 4.29 t/ha. Global rice output doubled from 26 ml t in 1995 to 50 ml t in 2010 [2]. it is the main food for the vast majority of Bangladesh's 149 million inhabitants. In 2009, the annual per capita consumption of milled rice was found to be 173.3 kg. By 2009, just 69.6% of Americans' calorie consumption came from rice, down from 74.8% in 1995. Protein consumption as a percentage of the population also decreased from 65.3% to 56.2% during this time period, with rice being the primary source. Because of its long rice production history, Bangladesh can now supply all of its own rice needs. From 1995 to 2009, imports dropped from 1 million t to 0.017 million t, then rose to 0.66 million t the

following year. The exportation of rice began in the new millennium. In order to keep domestic rice prices stable, some are imported. Efforts are being made by the government to increase rice production and decrease rice imports. Farmers that specialize in rice receive subsidies on agricultural inputs to keep prices manageable. In 2010, \$712,000,000 was allocated for subsidies. A government input distribution card provided small and marginal farmers with access to financial subsidies that could be used to purchase things like energy, fuel for irrigation, fertilizer, and other government aid. The primary rice-growing regions in Bangladesh are the country's highlands, irrigated lowlands, rain-dependent lowlands, stagnant medium-deep water, salty regions, and tidal non-saline regions. Before the onset of the monsoons, between March and May, Bangladesh receives about 400 millimeters of rain, allowing farmers to cultivate drought-resistant crops.

## **1.2 Motivation**

Rice is a typical food among us. The love of fish and rice is a sure sign that a person is an authentic Bengali. One of our mainstay foods is rice. If we can accurately anticipate the yield of crops like rice, we will be able to meet our needs. By utilizing advanced technologies, we can predict the upcoming harvest. So, we considered using a machine-learning approach to make amends.

## **1.3 Problem Definition**

In today's advanced information and communications technology (ICT) world, the concept of "machine learning" is of the utmost significance. The development of our agriculture industry will be facilitated by the application of machine learning. In order to come up with a workable solution, it is vital to define the problems that exist and the requirements that are associated with them in this area. In addition to the use of machine learning in the agriculture sector, it is essential to have an understanding of the policies of the government, as well as the standards of the technology industry and the various instructional choices.

## **1.4 Research Questions**

The following is a list of the primary questions that are the subject of this research:

- What aspects of the climate should be taken into account while determining crop yield?
- Which type of Rice would produce the highest amount of grain?
- Which algorithmic processes would be utilized?

## **1.5 Research Methodology**

In the methodology section of our research article, we described how we gathered data, preprocessed that data, categorized the data that resulted from the data collection, selected algorithms, implemented those algorithms, and then tested those algorithms. The conclusion of this section will conclude with a definition of the output of the suggested model.

## **1.6 Research Objective**

Utilizing Data Mining and Machine Learning Techniques is the way we plan to accomplish our goal of determining the yield of some of the most important crops in Bangladesh, including Aus, Aman, and Boro.

## CHAPTER 2

### BACKGROUND STUDY

#### 2.1 Introduction

We haven't been able to identify a reliable way to forecast crop production in our region, and there is no technology available that is capable of doing so either. The existing status of yield loss in Bangladesh's agriculture business and the implementation of machine learning are hence the frameworks in which we are discussing this topic. Machine learning is a subfield of artificial intelligence that enables computers to learn on their own and progress over time with minimal or no assistance from humans. The process of teaching self-operating computers to retrieve data and draw conclusions based on that data is referred to as machine learning. It can be difficult to find a coherent meaning for the term "learning" when it is applied to machine learning algorithms because there are several ways to extract knowledge from data depending on how the machine learning algorithm is created. This makes it challenging to find a meaning for the term "learning." It is important to have access to a considerable amount of data that displays a consistent result for a given set of inputs in order to learn anything new. The algorithm's performance will improve in direct proportion to the number of examples it has to learn from. This is the situation due to the fact that each input and output pair is embedded within a problem domain, which may be depicted by a line, a cluster, or some other statistical depiction. We used these algorithms in order to get the best possible results.

#### 2.1 Related Study

The use of machine learning to solve forecasting difficulties is becoming increasingly common. Utilizing machine learning to take action against yield prediction has been the subject of a great deal of thought. The use of machine learning has significantly facilitated the simplification of this method.

According to a study by Sung-Ju Jang et al. [3], manufacturing efficiency is a crucial element in determining a company's competitiveness in the semiconductor industry.

Examining the performance of the wafer maps before production and optimizing the wafer maps is one of the most important strategies for raising productivity. Numerous metrics, including gross dies, shot counts, lithographic performance rates, minimum film objective (MFO), cost, and so on, can be used to measure the productivity of wafer maps. In this study, which can be accessed here, they provide a better way for calculating agricultural yields using machine learning. Their method uses a spatial correlation that is independent of process characteristics between wafer position and die-level yield differences from a wafer test. By incorporating spatial modeling of these characteristics, the yield forecast has improved significantly. The trial findings also show that wafer maps may be produced using the proposed return model and method with productivity gains of up to 8.59 percent.

This study by Niketa Gandhi et al. [4] provides an overview of the machine-learning techniques that have been used in the rice-growing regions of India. A sizable amount of India's food production comes from cereal grains like rice, wheat, and various legumes. Favorable climatic conditions are crucial for rice-growing areas to be long-lasting and fruitful. The WEKA approach was used to apply the SMO classifier to a dataset of 27 districts in Maharashtra, India, and the study results are reviewed in this article. The dataset used to anticipate rice crop yields were collected from publicly accessible Indian government statistics. For the same dataset, the experimental findings demonstrated that several alternative methods significantly outperformed SMO.

Economic growth and food security in agricultural-based nations depend critically on careful agricultural planning, as noted by Rakesh Kumar et al. [5]. Inputs include production rates, market trends, and government policies. Research employing statistical or mathematical methods has been conducted on a wide variety of areas relevant to agricultural planning, including the prediction of crop yield rates, the prediction of weather, the classification of soil, and the categorization of crops. When more than one time of year is suitable for planting a crop on a given plot of land, it might be difficult to decide which crop to plant. This study suggests the CSM solves the yield selection problem and increases seasonal plants per output in order to maximize national economic growth. It has also been suggested that the proposed technique could increase crop net yields.

According to Anshal Savla et al. [6], precision agriculture necessitates using agricultural technology at the forefront of its field. In this particular piece of research, they went over a variety of classification algorithms that are applicable to data mining. These algorithms are then applied to a data collection that has been compiled over the course of time in order to make projections regarding the yields of soybean crops. In addition, a comparison of classification algorithms is carried out in order to determine which algorithm, in light of the various classification methods, is most suited for estimating yield.

The Collaborators Yogesh Gandge et al. [7] Farming accounts for a disproportionate share of the economy. Extreme weather events, such as flooding and drought, can have a negative impact on agricultural output in India, as they do in other countries. Soil quality, pH, EC, N, P, and K are precise crop predictions that can only be made by carefully analyzing a small number of factors. Since crop prediction calls for a large number of databases, this prediction method is a top pick for data mining. They can get useful insights from massive datasets by using data mining techniques. In this paper, we take a look at the many data mining strategies that have been used to make predictions about harvest yields. The success of a system designed to estimate crop yields is highly dependent on the precision with which features are extracted, and classifiers are applied. This paper compiles the outcomes of numerous agricultural output forecast algorithms used by different authors, together with details regarding their accuracy and recommendations.

The method of yield prediction developed by Monali Paul et al. [8] is widely used by farmers today to select the most productive crops for planting. Thus, estimating future harvests is an exciting challenge. In the past, a farmer's success depended on his or her prior knowledge and experience in a given land and crop. Here, we provide a data mining approach to determining the likely composition of soil-related variables. The expected category yield will be an indication of crop success. Using the Naive Bayes classifier and the K-Nearest Neighbor methods, the problem of predicting agricultural yields is formally recognized as a classification law.

Mohammad Motiur Rahman et al. [9] have shown that these limestone features have a significant impact on environmental variables such as erosion, wind direction, and humidity. A diverse scenery may be seen in Bangladesh, which is situated in the Himalayan

foothills. As a result of long-term human habitation, microregions have formed. Each of these places has its own distinct microclimate. In order to do this, the owner of a food business must deliberately consider the parts of a property that would produce the greatest profit. This study makes an effort to estimate future agricultural production using machine learning techniques. The simulations were subsequently "trained" on the relationship between current weather patterns and crop success. We'll then put the simulations to the test to see how well they can predict climatic variables that are yet unknown.

S. Bhanumathi et al. [10] discussed the assessment of crop productivity; data mining is a more recent academic discipline. Improved crop yields are a major concern in the farming industry. There is no farmer who doesn't wonder about his potential harvest. Investigate the many extraneous factors, such as location and pH, that are factored into soil alkalinity calculations. Percentages of nutrients like nitrogen (N), phosphorus (P), and potassium are calculated using third-party methods, such as APIs for the environment and temperatures, type of soil, the nutritional content of that soil, the amount of precipitation in that region, and the soil conditions (K). To construct a model, we will look at a number of these observables and train the data with a number of machine-learning strategies. The module increases agricultural productivity and farmer revenue by employing a model that reliably estimates crop production and provides the typical consumer with an appropriate fertilizer proportion based on environmental and field-specific characteristics.

Ratchaphum Jaikla et al. [11] created a method for estimating rice yield through crop yield forecasting. Most writers have tried to predict rice yields with high accuracy, but conventional approaches are time-consuming and often inaccurate. The purpose of this essay is to employ the SVR, which is the most popular image prediction model, to devise a strategy for estimating rice production. This article employs a three-stage prediction method using estimates of soil nitrogen, measurements of the weight of mosaic viruses, and predictions of rice yields. They compare and contrast these findings with those of commercial implementations, such as the DSSAT4 software used to put the Crop Yield Model into action (CSM-Rice simulation model). The outcomes are consistent with the CSM-Rice simulation model, suggesting that their approach is effective. Their model's error is likewise within reasonable ranges.



M. M. Hasan et al. [12] illustrate how the price of rice in Bangladesh has fluctuated throughout time. They tried to guess what the price of rice would be in the future to slow down the rate of change. For this, they use classifiers that have been around for a long time, like KNN, Naive Bayes, DT, SVM, and RF. The highest accuracy achieved by the Random Forest algorithm is 98.17 percent. Based on our knowledge and the research gap in previous works, we present a baseline method using regression algorithms to estimate yield estimation for aus, aman, and boro accounting for common yield-impacting parameters.

### **2.3 Summary**

The prior research conducted in this field is the topic of conversation in Chapter 2, which may be located by clicking this link. a wide range of approaches to the work that are related to machine learning and regressor algorithm, in addition to a specific study that is based on the English language.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

We are aware that data mining is a method that integrates machine learning, statistics, and database management systems. This method is used for the purpose of extracting and detecting patterns in big data sets. Therefore, in order to accomplish our research assignment, we followed the six steps of the Data Mining process.

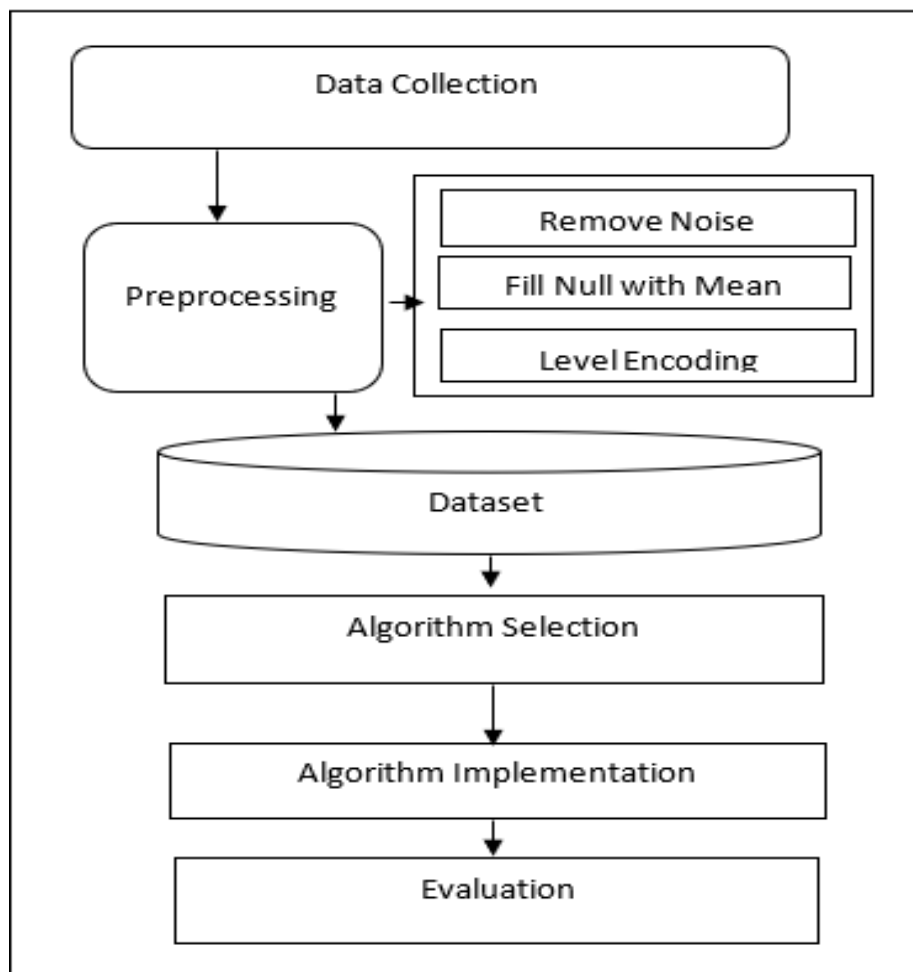


Figure 3.1: Methodology Diagram

While we were preprocessing the data, we observed three efforts: Remove Noise, Fill the Null with Mean, and Level Encoding. First, we gathered our data from several websites. Next, we preprocessed the data by following these steps: Remove Noise, Fill the Null with Mean, and Level Encoding. After that, we created our dataset, and after that, we chose our seven Machine Learning Regression techniques that need to be applied. Afterward, we are ready to move on. After that, we put our selected algorithms into action and, ultimately, as assessors well they worked. We have provided a condensed explanation of each of them below.

### **3.2 Data Collection**

When conducting research, data collection is typically one of the more challenging tasks. The website of Bangladesh's Agricultural Research Council provided us with the majority of the information that we used in this study (<http://www.barc.gov.bd/>). Our data can be broken down into two distinct categories. The first part was utilized for educational and research purposes, respectively. The other component was utilized for the purpose of making projections. The accumulation of data is a challenging effort that needs to be taken on by each and every interpretation. We were able to obtain the information that we required for our study by visiting the website of the Ministry of Agriculture in Bangladesh. After the data construction was finished, we subsequently split our dataset in two according to the instructions provided. The initial portion was utilized for the purposes of training and testing, while the remaining amount was utilized for the purposes of forecasting. In the first phase of the research, we selected 1460 daily prices from the years 2021 and 2022 in eight different cities, namely Barisal, Chattagram, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur, and, Sylhet, all of these cities are in Bangladesh. These were the points on which we began. In addition, for the section where we discussed developing forecasts, we utilized the average of 120 various daily costs for all cities in the year 2022.

### **3.3 Preprocessing**

These three stages have been traversed by our team while working on the preprocessing part of the Data Mining process. A rundown of the procedures can be found as follows:

## Remove Noise:

Our collection of data contained the variables Area, Production, Yield, and Production as independent variables. However, this is the only factor that we have considered: yield. As a result, we made the decision to halt production.

## Simply substitute the mean for the null:

Find the value that is averaged out over all of the criteria that were selected for each category. We have certain data relevant to each district that can be viewed for each division. In order to complete the task of filling in the value of the missing parameter, there is a certain technique that must be carried out.

The algorithm looks like this. The following tasks need to be completed in their entirety:

### Top:

- Place each of the districts into the appropriate category.
- Determine which of the divisions does not have a full complement of districts and write down the name of that division.
- Figure out which districts within each division are currently accepting new students.
- Calculate the mean score across all of the districts that are openly accessible for the given year.
- Enter an estimated value for this year for each of the districts that are included in a certain division.
- Remain Topping through 2022.

The outcomes of the correlation analysis performed on our dataset can be seen in Table 3.1, which can be found below. According to the information contained in this table, the value of Crop-name is the most significant one since it has the potential to have a higher impact on the variable that is being measured than any of the other variables.

Table 3.1 Correlation analysis of the dataset

	District	Division	Year	Crop name	Avg WindSpee d	Avg Sunshine	AvgMax Temp	Avg MinTemp	Avg Cloud Coverage	Avg Humidit y
0	Bagerhat	Khulna	2012	Aus	1.35	6.03	31.59	21.64	3.08	79.59
1	Bandarban	Chattagram	2012	Aus	1.21	5.75	30.59	21.68	3.31	74.92
2	borguna	Barisal	2012	Aus	1.08	6.02	31.05	21.85	3.48	82.76
3	Borishal	Barisal	2012	Aus	1.08	6.02	31.05	21.85	3.48	82.76
4	Bhola	Barisal	2012	Aus	1.08	6.02	31.05	21.85	3.48	87.76

In this particular scenario, the maximum temperature, humidity, and district all have a negative association with the variable that we are looking at. The amount of cloud cover, the lowest temperature, the number of daylight hours, the average wind speed, and the year all have a positive link with one another, and they are all very close to one another as well. Because of this, not a single attribute was removed from our dataset. This was the direct result of this.

### 3.4 Dataset Connection

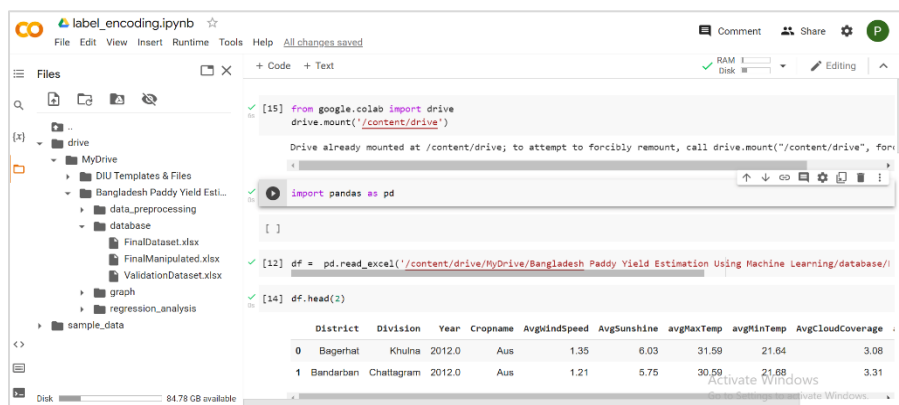


Figure 3.2 dataset connection

Figure 3.2 represent the dataset connection. As we used colab as our IDE. So we need to give permission from Gmail. We performed each of these functions very carefully.

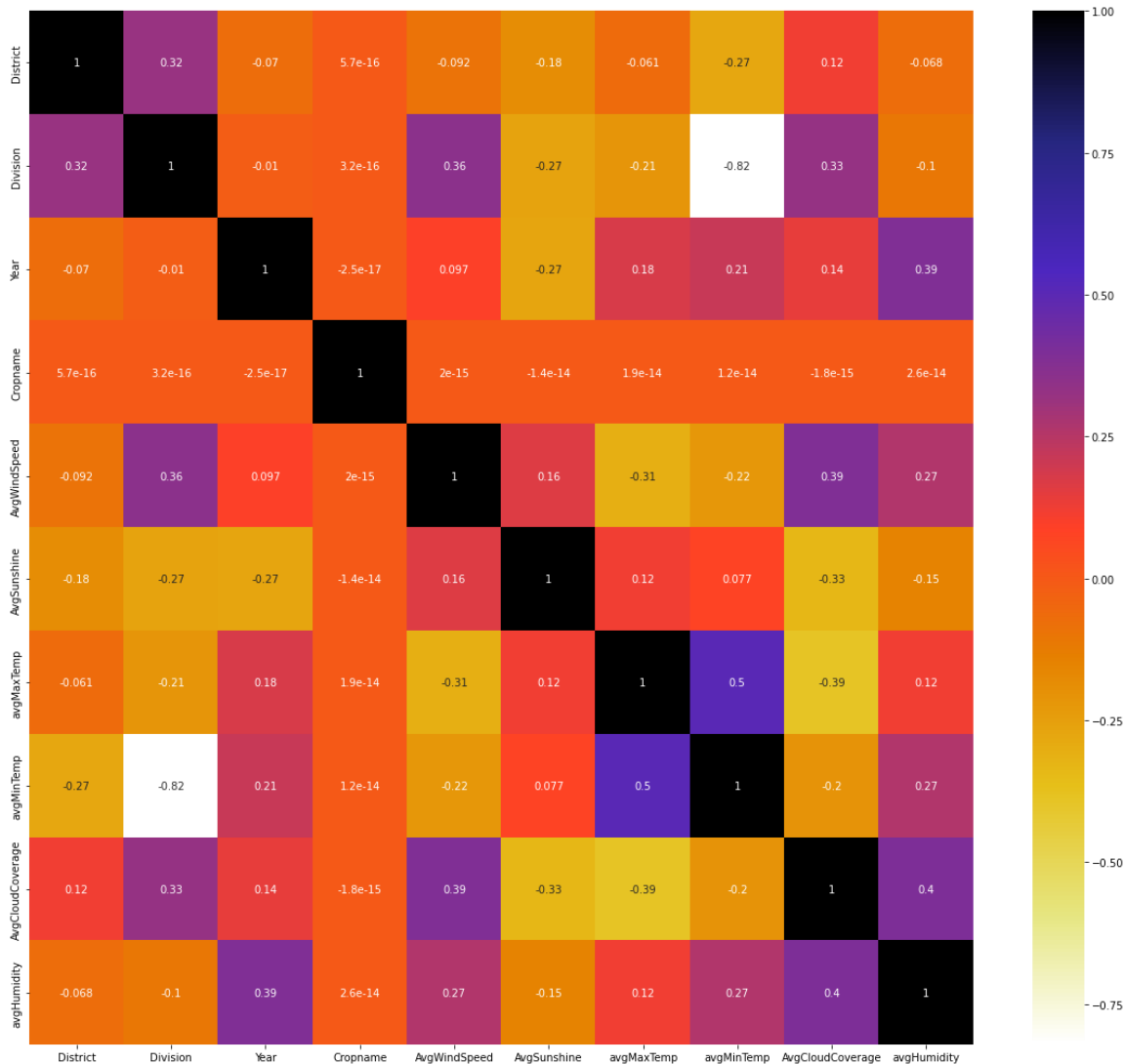


Figure 3.3: Correlation of data

The correlation of our data levels is represented here in figure 3.3. Each and every analysis result gained accurately displays the value. There is no inconsistency in the value of our data at all. Every level received a score lower than 0.5. If any level has an occurrence rate that is greater than 0.5, then we will have to adjust the level or get rid of it. However, as there are no levels, we do not require any level names to be changed.

The graph demonstrates that we have done a good job of analyzing our data and that our data itself is in good shape.

## Level Encoding:

We have gained control of 64 of our nation's districts. We are aware that the computer is not capable of understanding strings. Because of this, we decided to turn them into numbers, which range from 0 to 63. In addition, crop names were encoded using the numbers 1, 2, 3, and 4, which are respectively represented by Aus, Aman, and Boro. Level Encoding is a valuable technique for the analysis of large amounts of data, which helps machines better comprehend the processes.

Table 3.2 Level Encoding Representation of Data

	District	Division	Year	Crop name	Avg WindSpeed	Avg Sunshine	AvgMax Temp	Avg MinTemp	Avg Cloud Coverage	Avg Humidity
0	0	3	0	1	1.35	6.03	31.59	21.64	3.08	79.59
1	1	1	0	1	1.21	5.75	30.59	21.68	3.31	74.92
2	2	0	0	1	1.08	6.02	31.05	21.85	3.48	82.76
3	3	0	0	1	1.08	6.02	31.05	21.85	3.48	82.76
4	4	0	0	1	1.08	6.02	31.05	21.85	3.48	87.76

## 3.5 Algorithm Selection

Our work was constructed with the yield-dependent regression technique as its basis. To get good results quickly, we used seven of the most popular machine learning techniques. Gradient boosting, decision tree, Lasso, linear, neural network, random forest, support vector machine (SVM), and support vector machines (RF) are all names for these algorithms. Using this technique, we found the algorithm with the most precise outcomes. The discipline of machine learning makes use of a boosting technique called gradient boosting. An instance of boosting is gradient boosting. It is based on the premise that by integrating the previous model with the current best available model, the average prediction

error can be decreased. Defining the results that should be expected from the next model is the first step in reducing the likelihood of making mistakes.

A decision tree is a tool for making choices based on a tree-like decision model and its various consequences, such as the likelihood of events, the number of resources required, and the expected value of the final product. In particular, the following elements make up a decision tree: As such, this is one way to illustrate an algorithm that consists just of directives.

Common in statistical analysis, the linear viewpoint is used in linear regression (or "LR" for short). It's a visual depiction of the relationship between a continuous answer and a set of categorical predictors. A basic linear regression model is employed when there is only one candidate explanatory variable. Multiple linear regression is employed when there are several possible explanations for a given trend.

In statistics and machine learning, Lasso is a regression analysis technique used for variable selection and regularization, both of which improve the statistical model's interpretability and reliability. Lasso can be used to select which variables to include in an analysis, in addition to regularization.

Artificial neural networks, or "neural networks," refer to computer architectures that simulate the behavior of organic neural networks seen in animal brains. The term "neural network" is commonly used to describe such systems. In principle, an ANN is nothing more than a network of computers programmed to perform a computation in the same manner that your brain's neurons would. Random Forests are a group learning technique for classifying, performing regression, and carrying out other tasks by means of the construction of many decision-making trees during training and then class generation, the average forecast for individual trees. The procedure of working can also be approached using random forests.

Machine learning classification and regression analysis can be carried out by means of g. Support Vector Machines (SVM), are supervised learning models that employ learning approaches to probe data.



### **3.6 Algorithm Implementation**

After putting algorithms into place, we were able to execute Gradient Boosting and obtain the best possible accuracy with a data consumption rate of only thirty percent. The performance of the next six algorithms was also rather impressive. We came to the conclusion that the Gradient Boosting technique would provide the most accurate projection of the yield, thus we used it.

### **3.7 Evaluation**

At first, we acquired the information we needed for our research from a source that was easy to access. Following that, we estimated our data using seven different machine-learning methods to determine which of them was the most successful. And noted for possible application in the future. Each and every one of our studies will be delivered in order to ensure that the data preparation we conduct results in a good and reliable dataset.

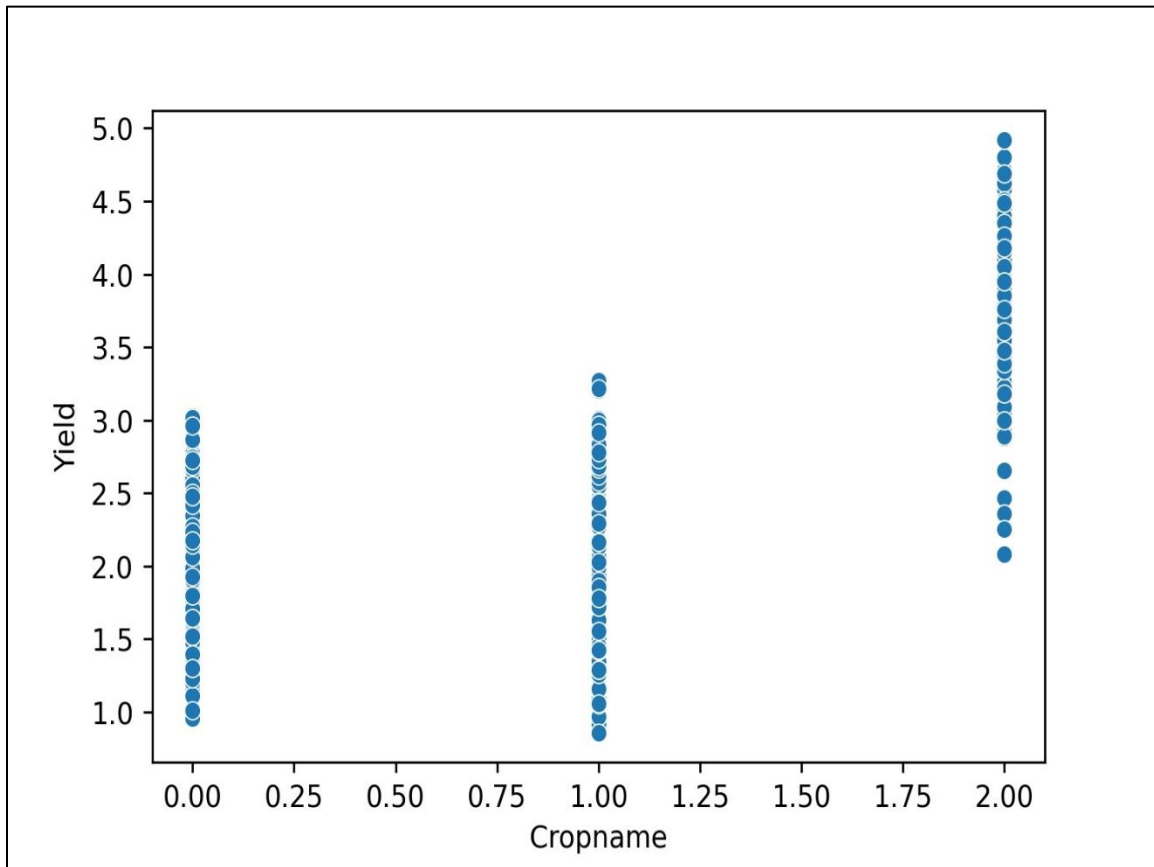


Figure 3.4 Representation of Yield vs Crop name

This graph depicts the yield vs the various crops. The three crops known as Aus, Amon, and Boro, can be seen in this picture. Boro yielded the most favorable results overall. Boro has a total of 2.00, and the others are as good. The count begins with 0 for Aus and 2 for Amon. Boro came up on top, despite the fact that there is not a significant gap between them and the other two.

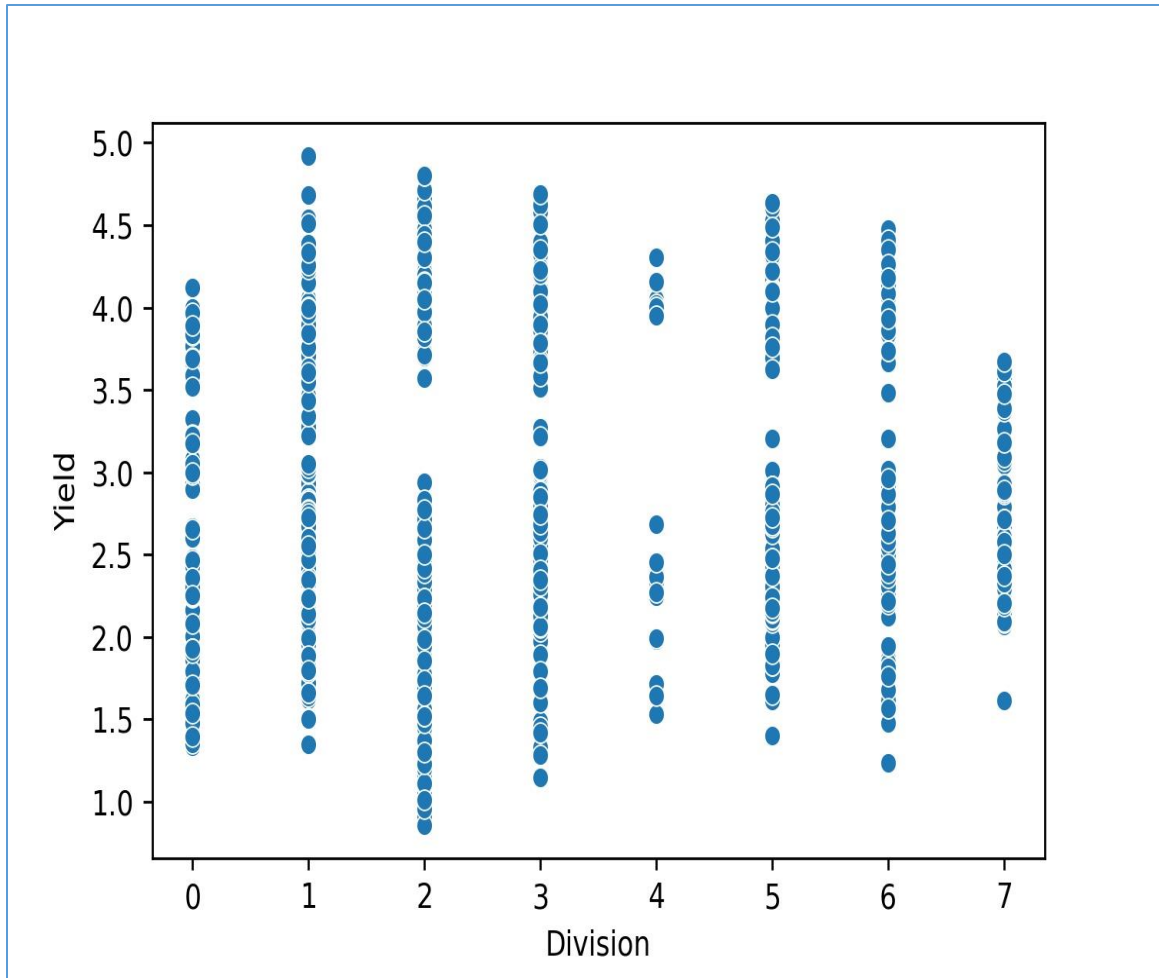


Figure 3.5 Representation of Yield vs Division

This graph illustrates the divisional outcome compared to the yield. Before we implemented level encoding, and as a consequence of that, each division was represented as a numeric number. Here, the totals from 0 to 7 of eight different divisions are displayed. After examining them, we discovered that the number 2 division obtained the best result. The other division likewise did very well in the scoring.

## CHAPTER 4

### EXPERIMENTAL RESULT ANALYSIS

#### 4.1 General Considerations

In this study, we have generated results by applying various machine learning regression algorithms to various types of parameters or variables (such as independent variables and dependent variables), compared these results with one another, and then analyzed the results in terms of their mean squared error and their R2 score.

#### 4.2 Result Analysis

Here we use test data from 30% to 70%. And use six different regressor algorithms for analysis results. Most of the algorithms gained the best result at 30% of R2\_Score.

#### Gradient Boosting

Table 4.1 Show the Gradient Boosting Algorithm Accuracy

Parameter	Gradient Boosting				
	30%	40%	50%	60%	70%
MAE	0.26	0.27	0.27	0.28	0.30
MSE	0.11	0.13	0.12	0.14	0.15
RMSE	0.33	0.36	0.35	0.37	0.39
R2_Score	0.85	0.84	0.85	0.83	0.81

Here, in Table 4.1, for the 30% to 70% data usage rate of the Gradient Boosting algorithm, we can see that the best value for RMSE (Root MeanSquare Error) is 0.33, which is given by Gradient Boosting Regression, and the value of R2 score for Gradient Boosting is 0.85 at 30% of test data. This is the best value for R2\_score, which is given by Gradient Boosting Regression.

## MLP Regressor

**Table 4.2 Show the MLP Regressor Algorithm Accuracy**

Parameter	MLP Regressor				
	30%	40%	50%	60%	70%
MAE	0.51	0.51	0.51	0.51	0.51
MSE	0.39	0.37	0.39	0.38	0.40
RMSE	0.63	0.61	0.62	0.62	0.63
R2_Score	0.55	0.56	0.53	0.54	0.52

Table 4.2 shows that for the MLP algorithm's 30%-70% data usage rate, Gradient Boosting Regression has the best RMSE (Root Mean Squared Error) value of 0.63 and the best R2 score of 0.55. Grad's MLP Regression algorithm has offered this as the most accurate result for the R2 score variable, and both of these values are better than the value of MAE (Mean Absolute Error), which is 0.65.

## Linear Regressor

**Table 4.3 Show the Linear Regressor Algorithm Accuracy**

Parameter	Linear Regressor				
	30%	40%	50%	60%	70%
MAE	0.52	0.51	0.51	0.51	0.51
MSE	0.39	0.37	0.38	0.38	0.38
RMSE	0.62	0.61	0.62	0.62	0.62
R2_Score	0.56	0.55	0.54	0.53	0.53

Table 4.3 shows that for the Linear Regressor algorithm's 30% to 70% data utilization rate, the best RMSE (Root Mean Square Error) value is supplied by Linear Regressor Regression, and the greatest R2 score is given by Linear Regressor as well (0.56). When compared to the MAE (Mean Absolute Error) value of 0.39 supplied by Grad The Linear Regression technique, this is the most accurate result for the R2 score variable.

## Decision Tree Regressor

**Table 4.4 Show the Decision Tree Regressor Algorithm Accuracy**

Parameter	Decision Tree Regressor				
	30%	40%	50%	60%	70%
MAE	0.52	0.51	0.51	0.51	0.51
MSE	0.39	0.37	0.38	0.38	0.38
RMSE	0.62	0.61	0.62	0.62	0.62
R2_Score	0.56	0.55	0.54	0.53	0.53

With a data utilization rate of 30% to 70% (as shown in Table 4.4), Decision tree Regression provides the best RMSE (Root Mean Square Error) value and Decision tree also provides the highest R2 score (0.56). By using the Decision Tree approach, this is the best possible outcome for the R2 score metric.

## Random Forest Regressor

**Table 4.5 Show the Random Forest Regressor Algorithm Accuracy**

Parameter	Random Forest Regressor				
	30%	40%	50%	60%	70%
MAE	0.24	0.25	0.26	0.27	0.30
MSE	0.11	0.10	0.12	0.13	0.15
RMSE	0.33	0.33	0.34	0.34	0.39
R2_Score	0.87	0.87	0.85	0.87	0.81

In Table 4.5, we see that when the data utilization rate is between 30% and 70%, Random Forest Regression yields the best RMSE (Root Mean Square Error) value, and it also yields the highest R2 score, at 0.87 at 30% of test data. The R2 score achieved here is the highest one can achieve utilizing the Random Forest methodology.



## Ridge Regressor

**Table 4.6 Show the Ridge Regressor Algorithm Accuracy**

Parameter	Ridge Regressor				
	30%	40%	50%	60%	70%
MAE	0.52	0.51	0.51	0.51	0.51
MSE	0.39	0.37	0.38	0.38	0.38
RMSE	0.62	0.63	0.62	0.62	0.62
R2_Score	0.56	0.56	0.54	0.54	0.53

In Table 4.5, we see that when the data utilization rate is between 30% and 70%, Ridge Regression yields the best RMSE (Root Mean Square Error) value, and it also yields the highest R2 score, at 0.56 at 30% of test data. The R2 score achieved here is the highest one can achieve utilizing the Ridge Regression methodology.

For all of the result analyses, Random Forest gained the best result for all six algorithms. Random Forest contains 0.87 R2\_score.

## 4.2 Analysis

We made an effort to identify the factors that led to the low crop yield in this region, and we used a graphical representation to examine the difference between the actual value and the value that had been projected.

Aman Real data vs. Aman Predicted data

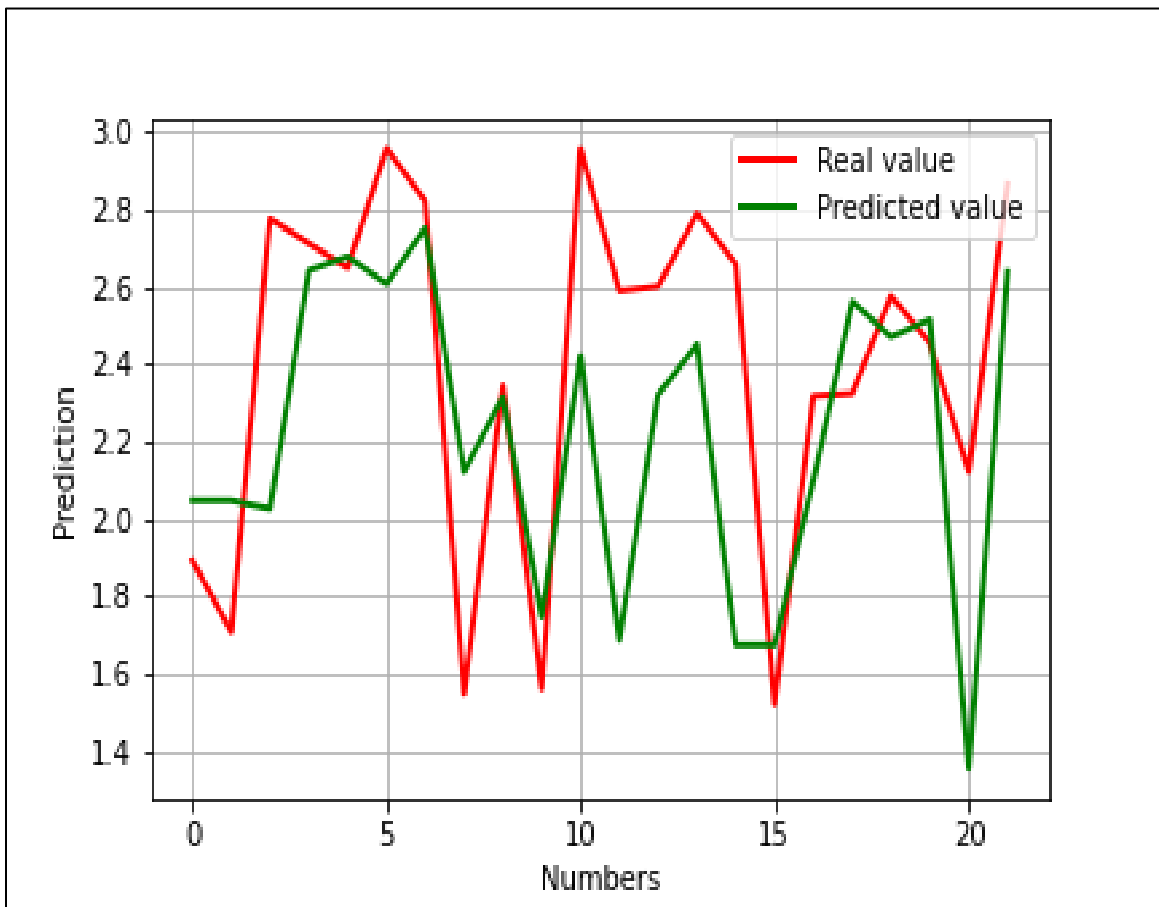


Figure 4.1: Aman Real vs. Aman Predicted

The line chart illustrates the relationship between Aman's current value and what was expected of him (rice). The actual values are displayed in red, whereas the expected values are shown in blue in this graph. The majority of the statistics are virtually identical to one another. There are simply a couple of minor mistakes that don't really matter. Both lines

will occasionally cross one another, while on other occasions, they will combine to form a continuous zigzag line. The data shown here does not reveal any signs of overfeeding or substantial volatility in the actual or predicted values. Based on this study, our data is quite clear, and there will be no great difficulties in making selections for upcoming jobs.

#### Aus Real data vs. Aus Predicted data

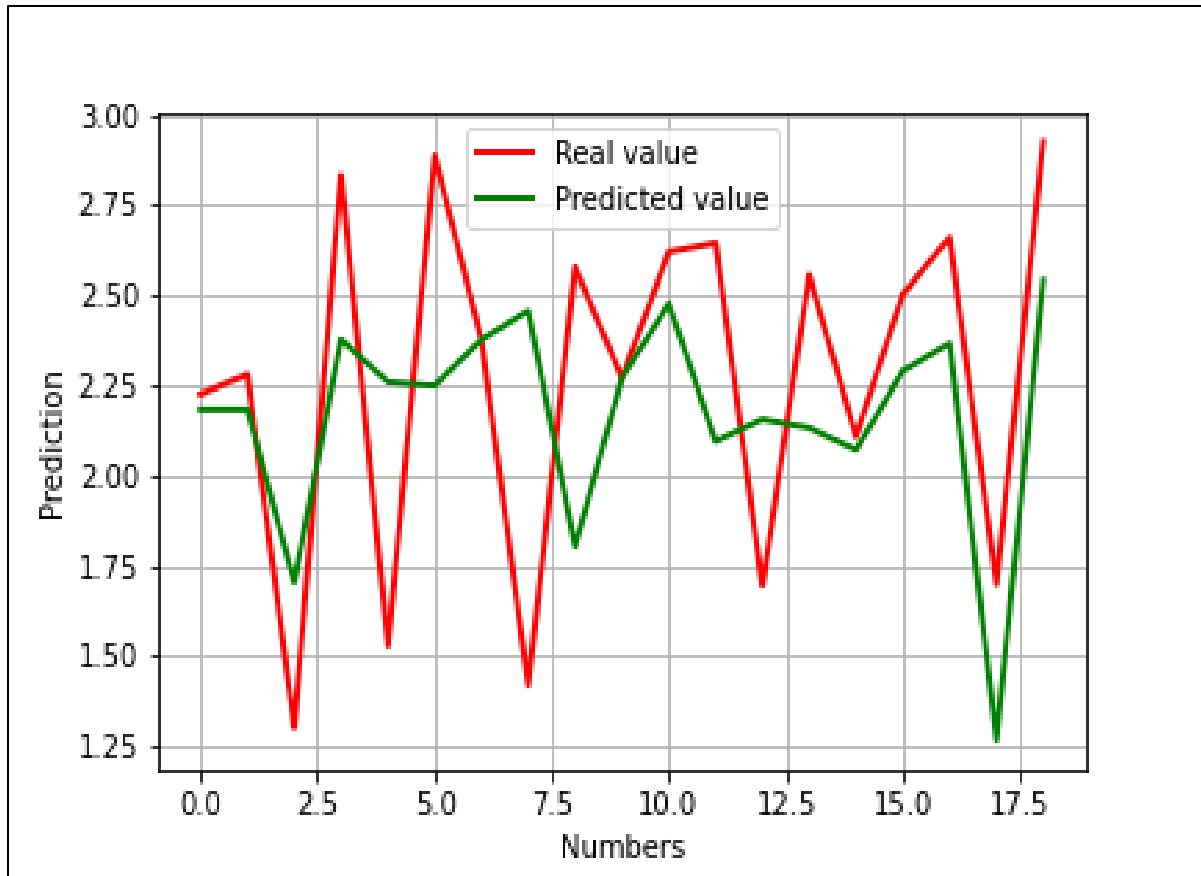


Figure 4.2: Aus Real vs. Aus Predicted

The chart shows a comparison of the actual value of Australia to the value that was anticipated for the country (rice). In this particular illustration, the values that actually occurred are depicted by the red color, while the values that were predicted are shown by the blue hue. When the graph first began, there was one discrepancy between the actual data and the forecasted data, as far as we can tell. After extending both lines in the same

manner, you should now have a smooth zigzag line. And the graph finishes with its recovery at the conclusion. The final conclusion drawn from the graph is that there are just one difference between the actual and anticipated data. When choosing a database for work in the future, there are not many mistakes. Here, no evidence of overfeeding or significant volatility in either the actual or anticipated values is presented.

### Boro Real vs. Boro Predicted

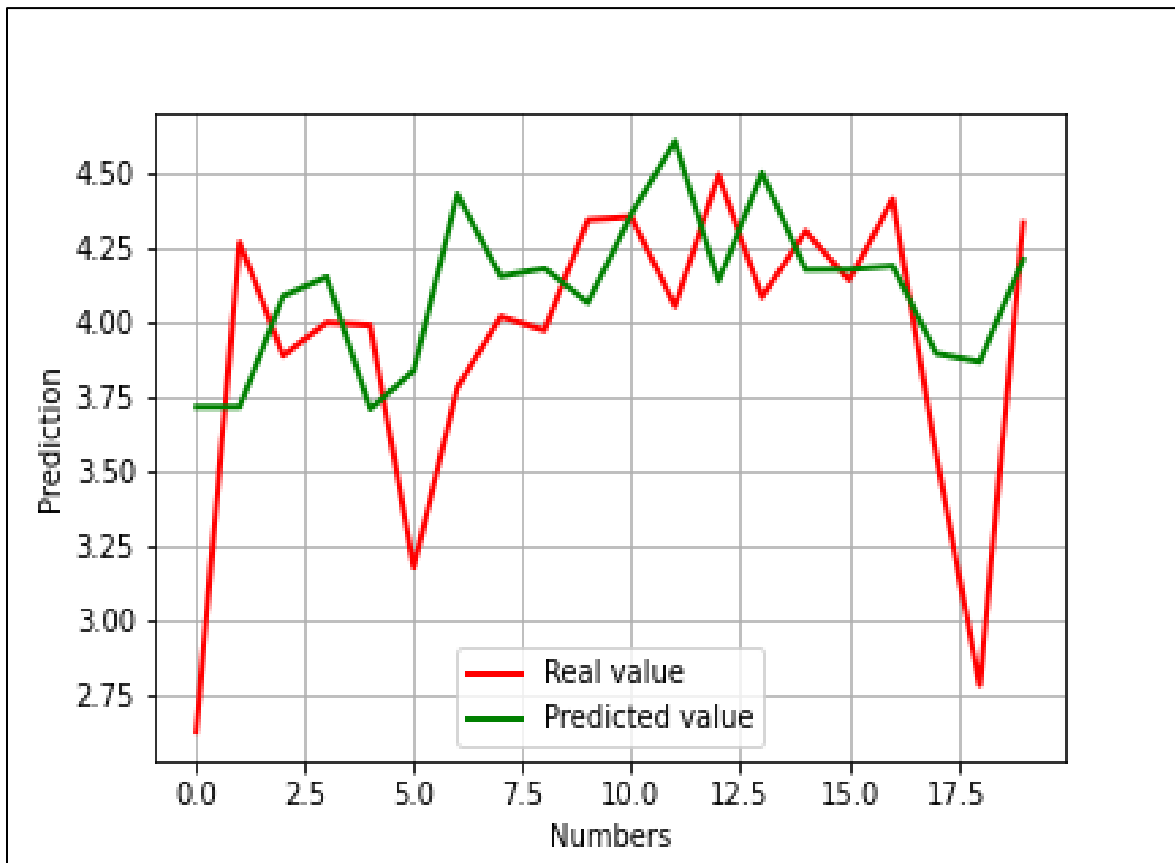


Figure 4.3: Boro Real vs. Boro Predicted

The graph presents a comparison of the actual value of Boro to the predicted value of the asset (rice). In this particular illustration, the values that actually occurred are depicted by the red color, while the values that were predicted are shown by the blue hue. As can be seen in the graph, both of the lines are practically identical, with the exception of a few

locations where they diverge slightly. Both lines will zigzag over one another sometimes, and in other situations, they will join forces to form a single continuous line that zigzags. There are not many options for errors to make when selecting a database to use for work in the future. In this case, there is no evidence of either significant fluctuation in the actual or predicted values, nor is there any evidence of overfeeding.

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 Conclusion

Bangladesh is primarily self-sufficient in the cultivation of rice and potatoes. Rice accounts for one-sixth of government revenue in Bangladesh and half of the agricultural GDP. In 2021, Bangladesh will require 27.26 million metric tons of rice. Reduced rice cultivation would leave a land area of 10.28 million hectares. There is a need to raise rice output from its current 2.74 t/ha to 3.74 t/ha. Our research shows that the yield of rice (Aus, Aman, and Boro) and potatoes can be forecast using a variety of climatic data from different locations. According to our findings, Random Forest is superior to the other six approaches for predicting future rice and potato yields. When compared to other types of rice, Boro produces more grain. To maximize harvest, plant more Boro. The findings of this study will aid farmers in making more accurate weather forecasts and so reducing the likelihood of future financial losses. The effects of this will be better able to stimulate our thought processes. Any issues can be resolved by increasing output. It is possible to increase yield by using machine learning to predict future results. A growing number of farmers see success with machine learning. Our success depends on our ability to increase output from the land. We incorporated seven widely used Machine Learning regression methods into our proposed model.

#### 5.2 Future Work

In the future, we will take those extension steps,

- Three different varieties of rice have been used in our experiments. We plan to use wheat, oil seeds, maize, legumes, etc., in our future work.
- We intend to improve our data collection primarily in the future by amassing information from all years.
- Making a user-friendly Android app for the masses.
- We will also develop a web base system that will provide suggestions.

### **5.3 Limitation**

We did our best to figure out how to achieve the most favorable conclusion imaginable, but there were still a few obstacles in our way. Inaccessibility to data was one of the constraints (i.e., soil data). We could have acquired more data, but it wasn't easy to do so due to the limits imposed by the majority of online government sites. We could have collected more data.

## REFERENCES

- [1] Anwar Ali, 'Save Boro first, then Aush', <https://www.thedailystar.net/frontpage/news/save-boro-first-then-aush-1893349>, April 16, 2020.
- [2] Ricepedia, 'Recent developments in the rice sector', <http://ricepedia.org/bangladesh>, September 2012.
- [3] S. Jang, J. Kim, T. Kim, H. Lee and S. Ko, "A Wafer Map Yield Prediction Based on Machine Learning for Productivity Enhancement," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 4, pp. 400-407, Nov. 2019, doi: 10.1109/TSM.2019.2945482.
- [4] N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 2016, pp. 1-5, doi: 10.1109/JCSSE.2016.7748856.
- [5] R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Avadi, India, 2015, pp. 138-145, doi: 10.1109/ICSTM.2015.7225403.
- [6] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture," 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2015, pp. 1-7, doi: 10.1109/ICIIECS.2015.7193120.
- [7] Y. Gandge and Sandhya, "A study on various data mining techniques for crop yield prediction," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, India, 2017, pp. 420-423, doi: 10.1109/ICEECCOT.2017.8284541.
- [8] M. Paul, S. K. Vishwakarma and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 2015, pp. 766-771, doi: 10.1109/CICN.2015.156.



- [9] M. M. Rahman, N. Haq and R. M. Rahman, "Machine Learning Facilitated Rice Prediction in Bangladesh," 2014 Annual Global Online Conference on Information and Computer Technology, Louisville, KY, USA, 2014, pp. 1-4, doi: 10.1109/GOCICT.2014.9.
- [10] S. Bhanumathi, M. Vineeth and N. Rohit, "Crop Yield Prediction and Efficient use of Fertilizers," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.
- [11] R. Jaikla, S. Auephanwiriyaikul and A. Jintrawet, "Rice yield prediction using a Support Vector Regression method," 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Krabi, Thailand, 2008, pp. 29-32, doi: 10.1109/ECTICON.2008.4600365.
- [12] M. M. Hasan, M. T. Zahara, M. M. Sykot, A. U. Nur, M. Saifuzzaman and R. Hafiz, "Ascertaining the Fluctuation of Rice Price in Bangladesh Using Machine Learning Approach," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225468.

## **APPENDIX**

The first was to outline the procedures for the analysis, which presented a number of difficulties. The report was the first. Furthermore, no progress has been made in this area previously. Indeed. It wasn't your typical job. We couldn't find someone who could help us that much. Another stumbling block was data collection, which proved to be a huge issue for us. We created a data-gathering corpus because we couldn't locate an open-source Bangladesh text pre-processing program. We've begun manually collecting data. Furthermore, classifying the various postings is a difficult task. We might be able to achieve it after a long time of hard labor.

# Bangladeshi Paddy Yield Prediction Using Machine Learning

ORIGINALITY REPORT

<b>2</b> %	%	%	<b>2</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

<b>1</b>	<b>Submitted to Daffodil International University</b>	<b>2</b> %
	Student Paper	

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		