

USE OF MACHINE LEARNING AND WEATHER  
PREDICTIONS TO BRIDGE THE GAP BETWEEN  
CONSUMER DEMANDS AND THE MARKET  
REQUIREMENTS OF FOOD CROPS

by

Raad Hasan

19301278

Romaiya Hossain

17101468

Farah Ahmed

17101280

Md. Iftekhar Islam

16201072

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

Department of Computer Science and Engineering  
Brac University  
June 2021

© 2021. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

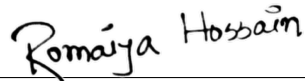
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

## Student's Full Name & Signature:



---

Raad Hasan  
19301278




---

Romaiya Hossain  
17101468



---

Farah Ahmed  
17101280



---

Md. Iftekhar Islam  
16201072

# Approval

The thesis/project titled “USE OF MACHINE LEARNING AND WEATHER PREDICTIONS TO BRIDGE THE GAP BETWEEN CONSUMER DEMANDS AND THE MARKET REQUIREMENTS OF FOOD CROPS” submitted by

1. Raad Hasan (19301278)
2. Romaiya Hossain (17101468)
3. Farah Ahmed (17101280)
4. Md. Iftekhar Islam (16201072)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 6th, 2021.

## Examining Committee:

Supervisor:  
(Member)



---

Dr. Md. Khalilur Rahman  
Associate Professor  
Department of Computer Science and Engineering  
BRAC University

Program Coordinator:  
(Member)

---

Dr. Md. Golam Robiul Islam  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)



---

Sadia Hamid Kazi  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

# Abstract

Bangladesh, being an Agro-Economy based country, is lacking the technological improvements regarding Supply-Demand ratio of Food Supply to the citizens. Since the recent price hike of onions in 2019, it is evident that such technological disadvantages bring severe losses to the projection of demand of food items and can bring significant hindrance to the economy of the country, adversely affecting the daily lives of the people. Therefore, we propose this paper to bridge the gap between Production, Consumption, Import-Export and Supply-Demand ratio using Linear Regression, Polynomial Regression, and Random Forest Techniques on datasets including and not limited to Average Min-Max Monthly Temperature, Average Monthly Rainfall and Humidity, Land Area-Production-Yield ratios, and Daily Consumption etc. This static dataset contains previous years' data taken from the Yearbook of Agricultural Statistics and Bangladesh Agricultural Research Council of those crops according to the area. The six year production period of 2013 to 2018 and the weather data set of 2008-2018 of Bangladesh have been taken into account in the formation of this dataset to corroborate learning and training of the algorithms and elevating the accuracy rate of the projection. The Cultivation Area estimates acquired are to be cross-verified with Satellite Images.

**Keywords:** Machine Learning; Weather Prediction; Land Area vs; Crop Production Prediction

## **Dedication**

We dedicate this research and all our efforts to the people suffering the unfortunate events brought on by the pandemic Covid-19. Their struggles, sorrows and strengths shall always glow brighter than the lanterns and lighten all darkness. We also dedicate our efforts to the proud farmers who work night and day to bring the food to our plates. Most importantly, we dedicate this paper to our friends; with their persistence, all our efforts has come to fruition.

## **Acknowledgement**

Firstly, all praise to the Almighty for whom we have completed our thesis amidst numerous hardships and tests of strengths. Secondly, we thank our Research Supervisor Dr. Md. Khalilur Rahman in his absolute and incomparable guidance and support through our journey.

And finally, to our parents who have always motivated us and inspired us to work. With their prayers, support, inspiration and investment, we are standing almost at the end point of our graduation.

# Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Dedication	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	2
<b>2 Background Study</b>	<b>3</b>
2.1 Literature Review . . . . .	3
2.2 Algorithms . . . . .	5
2.2.1 Decision Tree . . . . .	5
2.2.2 Random Forest . . . . .	7
2.2.3 Linear Regression . . . . .	8
2.2.4 Polynomial Regression . . . . .	9
<b>3 Methodology</b>	<b>10</b>
3.1 Data Collection . . . . .	10
3.2 Work Plan . . . . .	13
<b>4 Implementation &amp; Analysis</b>	<b>15</b>
4.1 Data Processing . . . . .	15
4.2 Algorithm . . . . .	19
4.2.1 Linear Regression . . . . .	25
4.2.2 Polynomial Regression . . . . .	26
4.2.3 Decision Tree Regression . . . . .	27
4.2.4 Random Forest Regression . . . . .	28
4.3 Result Analysis . . . . .	29
4.3.1 Linear Regression . . . . .	29
4.3.2 Polynomial Regression . . . . .	31

4.3.3	Decision Tree Regression . . . . .	34
4.3.4	Random Forest Regression . . . . .	36
<b>5</b>	<b>Conclusion</b>	<b>40</b>
5.1	Overview . . . . .	40
5.2	Research Challenges . . . . .	41
5.3	Contribution & Impact . . . . .	44
5.4	Future Work . . . . .	45
	<b>Bibliography</b>	<b>46</b>
	<b>Bibliography</b>	<b>46</b>



# List of Figures

2.1	Decision Tree . . . . .	6
2.2	Random Forest . . . . .	7
2.3	Linear regression R-squared Function . . . . .	8
2.4	Linear regression cost function . . . . .	9
3.1	View of unprocessed dataset . . . . .	10
3.2	Extracted and cleaned dataset . . . . .	11
3.3	Extracted and cleaned dataset(contd.) . . . . .	12
3.4	Workflow Diagram . . . . .	14
4.1	Names and Numbers . . . . .	16
4.2	Production dataset; ‘String’ replaced with ‘Int’ . . . . .	17
4.3	Major Dataset . . . . .	18
4.4	Median Dataset . . . . .	18
4.5	Minor Dataset . . . . .	19
4.6	Loading Median Dataset . . . . .	19
4.7	Median Dataframe . . . . .	20
4.8	Loading Major Data frame . . . . .	20
4.9	Loading Minor Data frame . . . . .	21
4.10	Major Data frame . . . . .	21
4.11	Minor Data frame . . . . .	22
4.12	Setting ‘x/y’ Parameter . . . . .	22
4.13	Major data frame train-test-split . . . . .	23
4.14	Median data frame train-test-split . . . . .	23
4.15	Minor data frame train-test-split . . . . .	23
4.16	View of Major, Median and Minor Dataframes . . . . .	23
4.17	Code for Linear Regression . . . . .	25
4.18	Code for Polynomial Regression . . . . .	26
4.19	Code for Decision Tree Regression . . . . .	27
4.20	Code for Random Forest Regression . . . . .	28
4.21	Crop101_Major Linear Regression Result . . . . .	29
4.22	Crop101_Median Linear Regression Result . . . . .	30
4.23	Crop101_Minor Linear Regression Result . . . . .	30
4.24	Crop101_Major Polynomial Regression Result . . . . .	31
4.25	Crop101_Median Polynomial Regression Result . . . . .	32
4.26	Crop101_Minor Polynomial Regression Result . . . . .	33
4.27	Crop101_Major Decision Tree Regression Result . . . . .	34
4.28	Crop101_Median Decision Tree Regression Result . . . . .	34
4.29	Crop101_Minor Decision Tree Regression Result . . . . .	35

4.30	Crop101_Major Random Forest Regression Result . . . . .	36
4.31	Crop101_Median Random Forest Regression Result . . . . .	37
4.32	Crop101_Minor Random Forest Regression Result . . . . .	38
5.1	Population Growth-Pulse Consumption-Production Deficit . . . . .	41
5.2	Results of prediction on Average Humidity . . . . .	43
5.3	Satellite Images . . . . .	44

# Chapter 1

## Introduction

### 1.1 Motivation

Bangladesh has a diverse rural population that produces a wide range of agricultural products. Bangladesh is primarily an agricultural country. Because of its very fertile land and favorable weather Crop varieties flourish in this country. In 2019, Agriculture accounts for approximately 12.68% of the country's GDP and employs more than 45% of the total labor force. However, productivity has not been as high as in developed countries. This is due to Bangladesh's lack of technological advancement required to feed such a large population, but still for export purposes. Then one thing comes to our mind: if all the farmers could have access to modern technological facilities, our country would be far ahead because they are the one who are pushing our economic growth rate and working hard for our country's better future. Our motivation arises to apply artificial intelligence in the farming sector. We thought to introduce machine learning into farming, where the machine will predict what and how much crops should and shouldn't be cultivated based on previous data. We gathered nearly 20 thousand databases of crop production and area based on 2013 to 2017 year ends and set them to our machine-learning model to predict how much crop should be cultivated in a specific area where agricultural production will increase with minimal effort for our farmers and also farmers will not have to face losses.

Another reason behind our motivation to work on this topic was businessmen in Bangladesh have a propensity to raise the price of essential goods by creating a drastic shortage of supply in the market through an unethical practice known as "artificial market." This problem has become a very critical issue during this pandemic and also in Ramadan month artificial marketing is very common. The Indian government unexpectedly imposed a ban on onion export to Bangladesh in 2019, resulting in a price spike. According to government statistics, the demand for onions of the whole country is 24 lakh tons per year, with local producers meeting the majority of our 16 lakh-tons requirement. The only way to meet the 8 lakh-tons shortage is to import onions from other countries, especially India. But still the price of onions went up to 50-70tk per Kg. To break the business syndicates we came up with this idea to help our farmers to predict how much crops they need to cultivate to fulfill our needs.

Another motivation was we wanted to work with a particular crop which has a major effect on our agricultural sector for example rice, pulse, wheat. After researching we found out a vast amount of work on our major crop Rice but we did not find any past work on pulse. So, we got our motivation to work on the second major crop of Bangladesh which is “Pulse”. But we came up with such an algorithm where the machine will predict what and how much crops should and shouldn't be cultivated based on previous data for any crop.

## 1.2 Problem Statement

Agriculture is considered to be the lifeline of the Bangladesh Economy. Cereal crops are the mainstay of Bangladesh Agriculture. It is said that Bangladesh is blessed by nature through its favorable weather and fertile land. But through the passage of time, fertility has reduced to a certain extent. In recent years, Bangladesh's agriculture sector has improved; the growth rate in crop yields has gone up. Even though Bangladesh has achieved significant progress in agriculture, with an expanding population, planning for future cereal production and demand is crucial to meeting the food security challenges in Bangladesh. Reducing the food insecurity problem at national level requires that the production, cultivation and marketing system to be efficient. In a subsistence agrarian economy of Bangladesh, domestic food production has an important role to play in the quest for food security. To facilitate this planning, projections of future supply and demand for cereals are critical. Moreover, there is a lack of knowledge and training among the farmers too. We live in the 21st century but our farmers still depend on their conventional methods and experience. Agriculture work in Bangladesh mostly has been done using backdated methods which prolongs the production of crops and fails to meet profitable standards, further affecting our economy unfavorably. Our farmers face numerous challenges, including lack of knowledge about soil, temperature, humidity which in turn leads to decreased earnings, higher production cost, and problems in achieving product quality. Moreover, in the current state of the world, the probabilities of global warming and environmental pollution induced natural calamities cannot be neglected. Along with that, threats of many emergencies, such as pandemics, political issues, are constantly looming over us. To keep ourselves prepared to face such difficulties, it is important to ensure the production of crops that not only fulfills our demands, but also aids in survival of the victims of such incidents. However, the core limitation of this thesis was data collection. We had to store these data into numerous formats to perform the training and test on them with the machine learning algorithm. The year wise data collection was strenuous. There was an absence of a digital approach to preserve the agricultural data in various government organizations as well. We had to overcome those difficulties. Simultaneously, another hindrance was working with bulk amounts of data where we faced several obstructions as well. The food insecurity of Bangladesh is not simply an economic problem. To be able to reduce the food insecurity problem at national level requires the food production and marketing system to be efficient. Therefore, it is really important to predict, with accuracy and precision, the amount of crop that can be produced in our country in any economic year to meet market demand.

# Chapter 2

## Background Study

### 2.1 Literature Review

Machine learning is the branch of computer science that is utilized to construct algorithms that exhibit self-learning properties i.e. learning which is done by the machine itself subsequently the term “Machine Learning”. It is seen as one of the noteworthy zones beneath Artificial Intelligence. To appear, the machine must translate and analyze the input. After analyzing it the result information separated from basically taking after the instructions on that information. Generally, that is the thing that machine learning algorithms do. It makes a difference to create superior choices within the future based on the cases that we provide. The basic point is to allow the computer to learn automatically without human intercession or offer assistance and control activities thus. It may be a major field in computer science that is being utilized in different shapes of dynamic technological development programs all around the world. These studies have also been expanded to the realm of agriculture in order to predict future outcomes. Researchers have already attempted to estimate future crop yields using data science and machine learning approaches. Machine learning algorithms are already being used by researchers to generate predictions. [8] Author K. Kaur uses machine learning in several applications in Indian agriculture in another study [1]. The numerous uses of machine learning techniques in agriculture, such as crop selection and crop prediction, are listed in this study. Crop Disease Prediction, Yield Prediction, Weather Forecasting, Smart Irrigation System, Crop Disease Prediction, Deciding the Minimum Support Price. These approaches will increase field productivity while reducing farmers’ input efforts. Along with advancements in farming tools and technology, relevant and accurate knowledge of several topics also plays an important role. [2] Satellite data was also used by researchers to predict the temperature in crop-growing areas [7]. However, because this strategy requires access to real-time satellite data, it will be inaccessible to the majority of people. Furthermore, this method’s accuracy was found to be insufficient. The weather has a significant impact on agricultural production. It has a significant impact on a crop’s growth, development, and yields, as well as the occurrence of pests and diseases, water requirements, and fertilizer requirements. In another paper they mainly focused on the K-means clustering and Regression model to predict rainfall throughout our country so that people could predict the possibilities of flooding earlier and could take proper risk management systems [10]. As Bangladesh has floods almost every year so it is very important

to take precautions before it hits. They have used ENSO (El Nino Southern Oscillation) index over others (AO, AAO, NAO) as it has a correlation coefficient of zero when conducting stimulation. ENSO helps to measure sea surface temperature and atmospheric pressure. They have contacted BMD (Bangladesh Meteorological Division) to select their regions based on previous records from the year 1980-2016 which are Dhaka, Khulna, Bogra, Barisal, Bhola and Cox's bazaar. They have shown rainfall comparisons between the months of June, July, August, September and October. They got closer values to the actual ones eliminating the points that didn't form clusters. They have mainly focused on 35 stations across Bangladesh. Although they mentioned degree tree but did not show any graph for the technique.

However, this paper models the agricultural yield and profit margin prediction [3] of Pulses for upcoming years in Bangladesh perspectives. The paper uses Decision Tree Algorithm, Linear Regression Analysis Technique, Support Vector Machine, and Random Forest to achieve the results with Decision Tree providing the highest accuracy factor. This paper achieves Crop Selection and Yield prediction methodologies with maximizing Yield with minimum Effort. The weather data collected consists of Monthly Average Humidity, Maximum-Minimum Temperature, and Humidity over 35 weather stations of Bangladesh from the year 2008-18. The production data consists of District/Division wise Area-vs-Production data of 8 breeds of pulses ranging from the year 2013-18. Using datasets of max-min Temperature, average Rainfall, Land Types, Chemical Fertilizer Usage, Soil Types, Soil Moisture, Soil Texture, Soil Consistency and Soil Consistency of Agricultural Zone [6]- 46 collected from Soil Resources Development Institute (SRDI), Bangladesh Agricultural Research Council (BARC) fed into algorithms of Deep Neural Network, Artificial Neural Network, Support Vector Machine, Logistic Regression and Random Forest on 3 major variants of Rice, Wheat, Jute and Potato. The paper proposes ANN to provide the highest average accuracy of 95% over other algorithms. The datasets used ranges from Temperature, Humidity, Sunshine, Price Diversity and Area allocated for production collected from Bangladesh Agricultural Development Corporation (BADC) Bangladesh Rice Research Institute (BRRI), and Ministry of Agriculture respectively for the areas Gazipur and Pabna to calculate profits per year and compare accuracy [5]. This paper projects the Food Production, Requirement and Surplus/Deficit situation using a rough and crude estimation method called Ohkawa's Equation commonly used in Business/Economic Modeling. The paper takes into consideration the Annual Growth Rates of Land Area-Production-Yield Rates, Imports, Food Aids, Safety Nets, Domestic Procurement and Public Distribution, Post-Harvest Losses, and Food Requirements to build a relationship between Production Growth, Consumption Growth and Supply-Demand Gap. The dataset contains details on crops' yield per hectare (Metric. Ton), average of minimum and maximum temperature, rainfall, year range, and region. Analyzing this data, the algorithm gives the result which predicts the most profitable crop in a particular region at a point in time. For the accuracy of the prediction, data of the past twelve years is being used by the algorithms for the purpose of learning and result analysis. The app also provides basic instructions on the cultivation process, irrigation, fertilization and harvesting for each crop cited by the Department of Agricultural Extension, Bangladesh which will facilitate the farmers in using scientific farming methodologies and reaching the preferred output mark. The datasets had been

collected from Bangladesh Agricultural Research Institute (BARI) [9], Bangladesh Bureau of Statistics (BBS), Ministry of Agriculture, and Bangladesh Bank Reports for income elasticity and FAOSTAT for food requirements. Analyzing this data, the algorithm gives the result which predicts the most profitable crop in a particular region at a point in time. For the accuracy of the prediction, data of the past twelve years is being used by the algorithms for the purpose of learning and result analysis. The paper aims to maximize the economic growth in terms of production and profit for both the country and farmers alike. This paper directly correlates the Production, and Supply-Demand situations of food cereals of Bangladesh [4]. Taking into consideration the Area-Production-Yield Ratios, Seasonal Acreage and Yield, Total Fertilizer Sales and Price, Wholesale Rice Price, and Monsoon Rainfall, and Population Expansion; the paper uses QUAIDS Model to project the rise in Demand and essential Supply of food grains 3 variants of Rice (Local and HYV) and Wheat with an average error of 10% [4]. The datasets have been collected from the Bureau of Statistics [6], Bangladesh Meteorological Department, Ministry of Finances and Planning, and UN Department of Economic and Social Affairs (UN DESA) for population expansion. We want to provide something to the farmers that will help them to continue their production without any hassle as well as the consumers will get an unstoppable supply of products so that none will be affected by the natural disasters that occur every year in Bangladesh. We want to ensure that the supply and the demand of the products are cohesive with one another.

## 2.2 Algorithms

### 2.2.1 Decision Tree

The Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm is used to solve regression and classification problems. Decision Tree Analysis is a general, predictive modeling tool that has applications spanning several different areas. A decision tree (also referred to as a classification tree or a reduction tree) is a predictive model that maps from observations about an item to conclusions about its target value.[1] The general motive of using the Decision Tree is to create a training model to predict class or value of target variables by learning decision rules inferred from preliminary data(training data). It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Attributes of Decision Tree:

Root Node: The root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Final output node of a decision tree is called the leaf node. The tree cannot be segregated further after getting a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions. Branch/Subtree: A tree formed by splitting the tree.

Pruning: Pruning is the process of removing unwanted branches from the tree.

Parent/Child node: The tree's root node is called the parent node, and other nodes are called the child nodes.

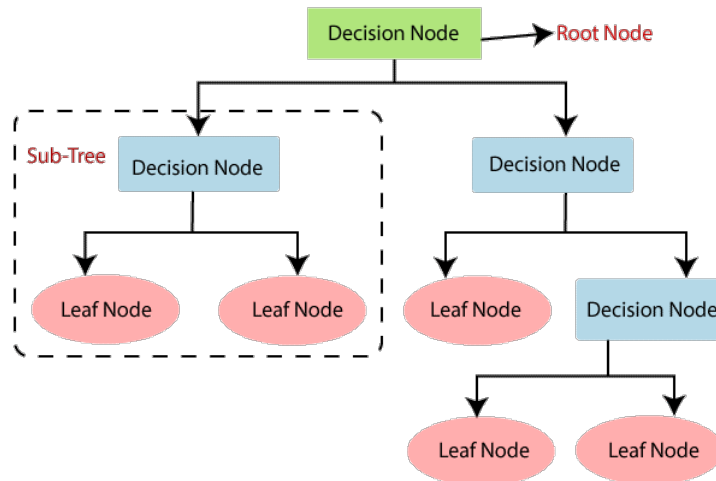


Figure 2.1: Decision Tree

#### Decision Tree Algorithm Pseudocode

1. The Decision tree begins with the root node, which contains the complete dataset.
2. Feature values are preferred to be categorical. If the values are continuous, then they are discretized before building the model. [3]
3. Recursively makes new decision trees using the subsets of the dataset. Continue this until a stage where it cannot be further classified and call it the final node as a leaf node.
4. Some statistical approaches are used in order to place the attributes as root or the internal node.

Except having over-fitting issues, Decision tree is the most easily understandable and most widely used algorithm. The understanding level of the Decision Trees algorithm is easy compared with other classification algorithms. Decision Tree can represent any Boolean function on discrete attributes. It is best for non-linear decision-making with a simple linear decision surface



## 2.2.2 Random Forest

Random Decision Forest or Random Forest is a supervised learning algorithm and one of the well-known methods that are utilized by researchers to generate predictive models. This method can be utilized for both regression problems and classification. It is mostly used for its simplicity and diversity. Random forest ensembles Decision Tree. It builds numerous decision trees and blends them to induce a more precise and steady prediction. While growing the trees, rather than searching for the foremost important feature while splitting a node, it searches for the best feature among a random subset of features. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting [1].

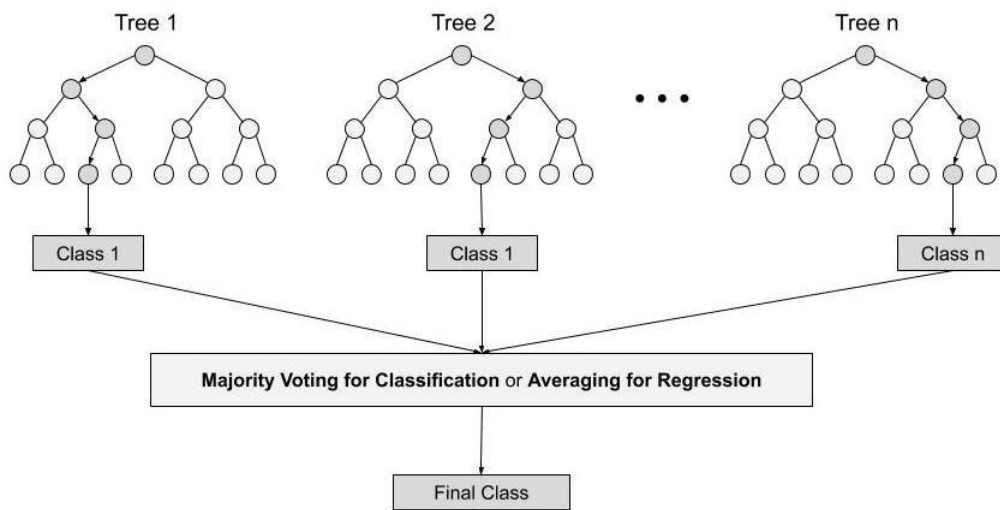


Figure 2.2: Random Forest

Random Forest works in two-phase first is to create the random forest by combining  $N$  decision tree, and the second is to make predictions for each tree created in the first phase. The Working prepare can be clarified like below:

Step-1: Select arbitrary  $K$  data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number  $N$  for decision trees that you want to build.

Step-4: Repeat Step 1 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Since Random forest combines numerous trees to anticipate the class of the dataset, it is conceivable that a few choice trees may foresee the correct output, whereas others may not. But together, all the trees foresee the correct output. The main

limitation of random forest is that an expansive number of trees can make the algorithm too slow and ineffectual for real-time predictions

### 2.2.3 Linear Regression

Developed by the field of statistics but proven to be very effective in the field of machine learning, Linear Regression is one of the most well-known models which predicts the output for the set of values (y) given that the input value is (x) and the output value is (y). As such both the input and output variables are numeric.

If the method consists of a single variable (x) then it is referred to as the simpler regression. Whereas, with multiple variables, the method is basically known as multiple linear regression.

A factor is assigned to each input value which is known as the coefficient of the linear equation. When there are multiple outputs in higher dimensions, the line is called hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients.

The process to find the best optimized model among various options can be achieved by the R-squared method. The R-squared model is a statistical method that measures the strength of the relationship between the independent and dependent variables where the high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model. It is also called the coefficient of determination. The formula is given below:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Figure 2.3: Linear regression R-squared Function

Linear regression plots a straight linear line in a diagram of scattered recorded points. The equation for the straight line or the best fit line is where one or more independent variables can be used to calculate the value of a dependent variable. The best fit line is found out by decreasing the average distance of original value to the points on the linear equation. This distance is called the cost function. The formula is given below:

$$Cost\ function(J_{\theta}) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad (4.2)$$

Figure 2.4: Linear regression cost function

To prevent the increase of R- squared value in a regression analysis, the unimportant variables are removed from the analysis with the development of the regression model. Therefore, the removal process is important in order to provide optimal and accurate results of the analysis.

## 2.2.4 Polynomial Regression

Polynomial regression is a type of regression analysis where it shows the relationship between the independent and dependent variables are represented by an nth degree polynomial. The Polynomial Regression equation is given below:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \dots + b_nx_1^n$$

Polynomial provides the best approximation of the relationship between the dependent and independent variable. Polynomial can accommodate a broad range of curvature and it fits with the method of least squares. In machine learning, it's also known as the special case of Multiple Linear Regression since we transform the Multiple Linear regression equation into Polynomial Regression by adding some polynomial terms. It's a linear model that's been modified a little to improve accuracy. The training dataset for polynomial regression is nonlinear in nature. To fit into the complicated and nonlinear functions and datasets, linear regression model is used. When we apply a linear model to a linear dataset, we get a decent result, as we have seen in Simple Linear Regression, but when we apply the same model to a non-linear dataset without any modifications, we get a drastic result. As a result of the increased loss function, the error rate will be high, and accuracy will be reduced. Therefore we need the Polynomial Regression model when data points are organized in non-linear form.

Steps for Polynomial Regression:

The main steps involved in Polynomial Regression are given below:

- Data Pre-processing
- Build a Linear Regression model and fit it to the dataset
- Build a Polynomial Regression model and fit it to the dataset
- Visualize the result for Linear Regression and Polynomial Regression models.
- Predicting the output.

# Chapter 3

## Methodology

### 3.1 Data Collection

The primary data is collected from the Yearbooks of Agricultural Statistics 2013/14 to 2018/19 as published by Bangladesh Bureau of Statistics (BBS). We have been presented with Area vs. Production Yield of 136 crops ranging from years 2013 to 2018, Average Monthly Rainfall, Humidity, and Max-Min Temperatures of 35 stations across the years 2008 to 2018. All the data have been presented in tabulated forms in .pdf files, which were then processed to .xl file format.

Area vs. Production							Weather													
District/Division	2013-14			2014-15			Name of stations	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	
	Area (acres)	Yield per acre (KG)	Production (MT)	Area (acres)	Yield per acre (KG)	Production (MT)														
1	Barguna	280	211	59	335	343	115	Dinejpur	78	69	64	69	89	81	82	86	81	85	74	78
2	Barisal	1051	328	345	1014	321	325	Rangpur	80	72	66	70	82	81	81	84	82	83	77	80
3	Bhola	280	782	219	364	610	222	Rajshahi	79	74	64	62	80	83	83	86	84	86	76	80
4	Jhalokati	280	335	87	288	347	100	Dogra	76	67	63	---	81	79	81	85	82	83	78	---
5	Patuakhali	1150	230	265	1455	250	364	Mymensingh	78	70	74	77	84	82	83	85	85	85	81	82
6	Projpur	-	-	-	-	-	-	Sylhet	74	62	60	69	83	81	83	85	83	81	74	75
7	Bandarban	-	-	-	6	667	4	Srimangal	79	72	67	72	83	82	83	85	85	85	81	83
8	Brahmanbaria	-	-	-	-	-	-	Ishwardi	81	74	63	63	81	83	82	85	84	85	78	81
9	Chandpur	-	-	-	-	-	-	Dhaka	65	55	55	63	78	76	77	80	81	78	66	72
10	Chittagang	80	325	26	73	342	25	Comilla	74	68	75	78	85	81	83	85	84	85	78	81
11	Comila	-	-	-	-	-	-	Chandpur	77	70	70	76	85	82	85	86	85	84	75	79
12	Cox's Bazar	-	-	-	-	-	-	Jessore	77	71	65	67	79	81	82	84	84	85	75	78
13	Feni	57	421	24	48	417	20	Fardpur	76	69	65	68	84	82	83	85	85	85	76	79
14	Khagrachhari	-	-	-	-	-	-	Madaripur	79	72	71	76	86	84	85	87	86	86	77	82
15	Laksmipur	31	161	5	30	167	5	Khulna	83	77	70	71	82	84	90	91	85	86	75	78
16	Noakhali	145	359	52	148	378	56	Satkhira	76	72	71	70	80	81	83	85	85	83	76	79
17	Rangamati	-	-	-	-	-	-	Barisal	79	76	75	75	86	84	87	87	88	87	80	82
2	Chittagang	313	342	107	305	361	110	Bhola	78	73	74	76	86	84	87	89	87	86	80	83
								Feni	77	70	72	76	86	83	85	86	86	86	79	80
								M court	73	67	71	75	85	82	86	87	85	85	76	80
								Hatiya	76	70	76	80	86	86	86	90	88	88	81	82
								Sitakunda	75	70	76	78	86	85	85	86	87	85	78	81
								Sandip	74	68	76	77	86	84	85	86	88	87	80	83
								Kuludkia	81	80	81	87	87	89	89	88	88	89	77	76
								Cox's Bazar	70	66	75	77	85	85	86	86	85	83	73	73
								Rangamati	75	62	63	68	79	82	83	85	84	86	82	82
								Takhal	64	62	74	77	85	87	89	87	87	84	78	71
								Patuakhali	77	71	73	77	87	86	89	89	90	89	79	81
								Khajupara	71	65	72	---	85	84	86	86	88	88	78	78
								Sayedpur	78	69	64	70	82	82	82	85	82	84	79	82
								Tangal	79	73	67	68	81	80	81	84	83	84	77	81
								Mongla	74	67	70	73	83	85	87	87	86	87	77	79
								Chudanga	79	74	65	65	81	84	86	87	87	88	82	82
								Chittagang	72	64	77	77	85	84	84	86	84	85	78	77
								Ambagan(Ctg.)	69	56	70	72	82	83	82	84	85	85	75	77

Figure 3.1: View of unprocessed dataset

The dataframes were then converted to .csv files with Microsoft Excel.

Thereafter, we had formulated 5 .csv files-  
1. Area vs. production (2013-2018)

2. Average monthly Humidity (2008-2018)
3. Average monthly Rainfall (2008-2018)
4. Average monthly Maximum Temperature (2008-2018)
5. Average monthly Minimum Temperature (2008-2018)

Following the conversion of the .pdf files, the new .xl files were in need of heavy clean-up. This encompasses removal of null values, appropriately dispatching missing values, and detethering and removal of joint rows and columns.

Area vs. Production					Average Humidity			
District	Crop	Year	Area(Acres)	Production(Ton)	Station	Year	January	February
Barguna	Gram	2013	200	59	Dinajpur	2008	80	74
Banshal	Gram	2013	1051	345	Rangpur	2008	83	75
Bhola	Gram	2013	280	219	Rajshahi	2008	79	73
Jhallokati	Gram	2013	260	87	Bogra	2008	80	72
Patuakhali	Gram	2013	1150	265	Mymensingh	2008	84	75
Firojpur	Gram	2013	0	0	Sylhet	2008	78	69
Barishal(Div)	Gram	2013	3021	975	Srimangal	2008	83	76
Bandarban	Gram	2013	0	0	Ishurdi	2008	81	75
Brahmanbaria	Gram	2013	0	0	Dhaka	2008	70	62
Chandpur	Gram	2013	0	0	Comilla	2008	80	73
Chattogram	Gram	2013	80	25	Chandpur	2008	78	73
Cumilla	Gram	2013	0	0	Jessore	2008	80	76
Cox's Bazar	Gram	2013	0	0	Faridpur	2008	78	74
Feni	Gram	2013	57	24	Madaripur	2008	78	73
Khagrachhari	Gram	2013	0	0	Khulna	2008	80	74
Laksmipur	Gram	2013	31	5	Satkhira	2008	79	75
Noakhali	Gram	2013	145	52	Barisal	2008	81	77
Rangamati	Gram	2013	0	0	Bhola	2008	83	79
Chattogram(Div)	Gram	2013	313	107	Feni	2008	80	74
					M.court	2008	81	75
					Hatiya	2008	81	73
					Sitakunda	2008	81	75

Average Rainfall				Average Max Temp			
Station	Year	January	February	Station	Year	January	February
Dinajpur	2008	33	1	Dinajpur	2008	23.5	28
Rangpur	2008	36	1	Rangpur	2008	23.3	27.9
Rajshahi	2008	26	0	Rajshahi	2008	24.5	29.5
Bogra	2008	27	0	Bogra	2008	24.8	29.2
Mymensingh	2008	30	5	Mymensingh	2008	24.9	28.8
Sylhet	2008	19	35	Sylhet	2008	26.6	30.5
Srimangal	2008	32	24	Srimangal	2008	26.2	29.9
Ishurdi	2008	37	2	Ishurdi	2008	24.8	29.4
Dhaka	2008	23	56	Dhaka	2008	25.9	29.7
Comilla	2008	30	11	Comilla	2008	26	28.9
Chandpur	2008	39	22	Chandpur	2008	26.2	29.9
Jessore	2008	70	35	Jessore	2008	27.3	30.8
Faridpur	2008	42	45	Faridpur	2008	25.7	29.7
Madaripur	2008	40	15	Madaripur	2008	26.5	29.9
Khulna	2008	66	36	Khulna	2008	26.2	29.8
Satkhira	2008	92	39	Satkhira	2008	26.4	29.8
Barisal	2008	41	24	Barisal	2008	26.7	30
Bhola	2008	23	15	Bhola	2008	26.3	29.6
Feni	2008	29	16	Feni	2008	26.3	29.1
M.court	2008	27	16	M.court	2008	26.4	29.6
Hatiya	2008	41	23	Hatiya	2008	27	30.1
Sitakunda	2008	32	5	Sitakunda	2008	28.2	30.9

Figure 3.2: Extracted and cleaned dataset

Average Min Temp				
Station	Year	January	February	
Dinajpur	2008	11.8	12.3	
Rangpur	2008	12.1	12.7	
Rajshahi	2008	11.1	12	
Bogra	2008	12.5	13	
Mymensingh	2008	13.1	13.8	
Sylhet	2008	14.4	14	
Srimangal	2008	11.2	11.1	
Ishurdi	2008	11.2	11.8	
Dhaka	2008	14.5	15.2	
Comilla	2008	13.2	14.1	
Chandpur	2008	14.4	14.9	
Jessore	2008	11.2	13.1	
Faridpur	2008	13.3	13.7	
Madaripur	2008	13.7	14.4	
Khulna	2008	13.8	15.3	
Satkhira	2008	12.9	14.5	
Barisal	2008	13.2	14.2	
Bhola	2008	13.7	14.7	
Feni	2008	14	14.8	
M.court	2008	14.8	15.4	
Hatiya	2008	14.1	14.5	
Sitakunda	2008	12.4	13.1	

Figure 3.3: Extracted and cleaned dataset(contd.)

Furthermore, we have intended the use of the Predictive Algorithms-

1. Supported Linear Regression
2. Polynomial Regression
3. Decision Tree Regression
4. Random Forest Regression

Finally, the results of the regression models are to be compared for maximum accuracy, and the algorithms with highest accuracies are to be processed for calculating the Consumption to Market Requirements.

## 3.2 Work Plan

Our prime objective is to forecast the production of the crops in the upcoming years. Therefore, we must obtain the most accurate prediction algorithm. The dataframes are in need of further processing, i.e, the prediction algorithms cannot process ‘String’ type variables. So, the districts must be converted to ‘Integer’ types. This infers to assigning numbers corresponding to the districts, stations, and crop types.

Next, a crop type is to be selected for processing into the predictive algorithms. Case in point, Pulse type Gram labelled as Crop101. But, the Area and Production data values range from 0 to over 10,000 depending upon crop types and region. Therefore, after a crop type has been selected, it has to be split into further Major, Media, and Minor subsets. These subsets are to be individually fed into the algorithms to obtain most precise results.

These results are compared against the others to select the most effective algorithm, and used to calculate the Production of said crop corresponding to the Year. Then the production data is to be used to calculate the consumption to demand gap.

The flowchart as shown summarizes the workflow procedures.

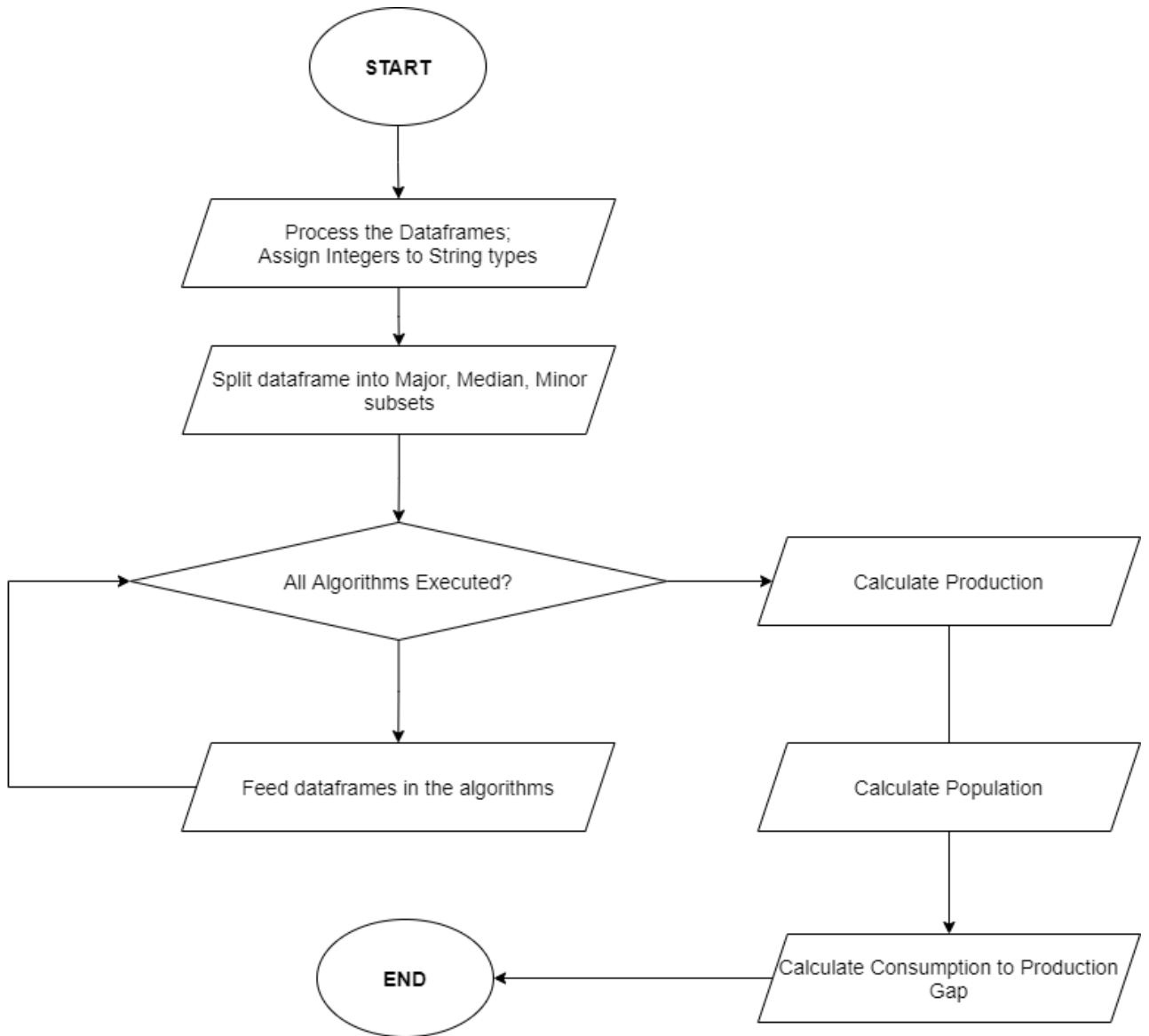


Figure 3.4: Workflow Diagram



# Chapter 4

## Implementation & Analysis

### 4.1 Data Processing

Firstly, the dataframes were converted from .pdf to .xl format. Following the conversion, the dataframes required cleanup. The .xl files showed merged rows and columns, multiple inputs in a single cell etc that had to be sorted manually via inspection. Null values were eliminated, missing values were appropriately assigned. For example, in the Area vs. Production datasets, districts with no production of certain crops were marked with the character '-'. These conditions were replaced with integer '0'.

Other cases in the Weather datasets, Stations with missing records of meteorological data had marked the cells with '-'. For such cases, the average of the data for the Station for all of the designated months has been used to replace the '-' assigned to the month's data in scope.

In the next step of the data processing, all the names of the districts and stations were assigned predetermined numeric values as shown in the following image.

District	Number		Crop	Number
Barguna	1		Gram	101
Barishal	2		Arhar	102
Bhola	3		Kheshari	103
Jhallokati	4		Mashkalai	104
Patuakhali	5		Mashur	105
Pirojpur	6		Motor	106
Barishal(Div)	7		Mung	107
Bandarban	8		Other	108
Brahmanbaria	9			
Chandpur	10			
Chattogram	11			
Cumilla	12			
Cox's Bazar	13			
Feni	14			
Khagrachhari	45			

Figure 4.1: Names and Numbers

A total of 72 district names (64 districts, 8 divisions), 8 variants of Pulses, and 35 stations were to be reassigned. This operation was made successful by iterating multiple replacement functions provided by Microsoft Exel application. The following image shows numeric representation of the 'String' data types that are now assigned 'Integer' values to be processed by the algorithm.

District	Crop	Year	Area	Production
1	101	2013	280	59
2	101	2013	1051	345
3	101	2013	280	219
4	101	2013	260	87
5	101	2013	1150	265
6	101	2013	0	0
7	101	2013	3021	975
8	101	2013	0	0
9	101	2013	0	0
10	101	2013	0	0
11	101	2013	80	26
12	101	2013	0	0
13	101	2013	0	0
14	101	2013	57	24
45	101	2013	0	0
16	101	2013	31	5
17	101	2013	145	52
18	101	2013	0	0
19	101	2013	313	107
20	101	2013	4	1
21	101	2013	427	156
22	101	2013	41	15
23	101	2013	99	41
24	101	2013	70	21
25	101	2013	7	2

Figure 4.2: Production dataset; ‘String’ replaced with ‘Int’

In the next step, the Area and Production of test Crop 101 over the years 2013 to 2018 are to be extracted into a .csv file. This new dataframe has to be parted into 3 ranges Major, Median, and Minor. For test Crop 101, the first operation is sorting the dataframe in the ascending order of ‘Year’ and ‘Area’.

Next, we are to set ranges to Minor, Median, and Major classes. We have judged the distribution of values of the Crop 101 dataframe, and deduced-

1. Minor range: 0-199 acres
2. Median range: 200-999 acres
3. Major range: 1000+ acres

In conclusion of which, the datasets are ready to be processed into the algorithms.

1	District	Crop	Year	Area	Production
2	2	101	2013	1051	345
3	5	101	2013	1150	265
4	26	101	2013	1532	512
5	48	101	2013	2711	1037
6	37	101	2013	2788	998
7	53	101	2013	2836	695
8	7	101	2013	3021	975
9	55	101	2013	3984	1604
10	57	101	2013	7676	2645
11	2	101	2014	1014	325

Figure 4.3: Major Dataset

1	District	Crop	Year	Area	Production
2	44	101	2013	204	77
3	29	101	2013	215	111
4	56	101	2013	229	99
5	60	101	2013	245	88
6	4	101	2013	260	87
7	1	101	2013	280	59
8	3	101	2013	280	219
9	19	101	2013	313	107
10	33	101	2013	320	112
11	40	101	2013	356	145

Figure 4.4: Median Dataset

17	70	101	2013	0	0
18	71	101	2013	0	0
19	20	101	2013	4	1
20	38	101	2013	6	1
21	25	101	2013	7	2
22	27	101	2013	7	3
23	31	101	2013	9	2
24	36	101	2013	9	3
25	61	101	2013	11	6
26	35	101	2013	21	8
27	62	101	2013	22	9

Figure 4.5: Minor Dataset

## 4.2 Algorithm

We start by loading the datasets in the algorithms. First we upload the Crop 101 Median dataset labelled C101\_Median.

```
# import datasets

datasets = pd.read_csv("/content/C101_Median.csv")

df = datasets

df.index = df.index + 1

print(df)
```

Figure 4.6: Loading Median Dataset

Which displays us with the contents of the ‘df’ dataframe that we shall be working on.

	District	Crop	Year	Area	Production
1	44	101	2013	204	77
2	29	101	2013	215	111
3	56	101	2013	229	99
4	60	101	2013	245	88
5	4	101	2013	260	87
..	...	...	...	...	...
75	3	101	2018	275	168
76	34	101	2018	289	158
77	43	101	2018	351	156
78	44	101	2018	384	116
79	66	101	2018	424	186

Figure 4.7: Median Dataframe

Addition of +1 to 'df.index' as shown in Figure 4.2.1 results in the Median Dataframe index to start from 1 instead of 0 as displayed in Figure 4.2.2. The application of indices is significant, since it is used for visual ease of interpretation by human users, also to identify and/or concat/merge in various levels of subset dataframes are generated to operate the predictive algorithms. This shall be further displayed and discussed in Section 4.3 Result Analysis.

Loading operations similar to the action in Figure 4.2.1 are to be taken for Crop 101 Major and Minor Dataframes labelled C101\_Major, and C101\_Minor, on separate algorithms.

```
# import datasets

datasets = pd.read_csv("/content/C101_Major.csv")
df = datasets
df.index = df.index + 1
```

Figure 4.8: Loading Major Data frame

```
# import datasets

datasets = pd.read_csv("/content/C101_Minor.csv")
df = datasets
df.index = df.index + 1
```

Figure 4.9: Loading Minor Data frame

Which displays the Major and Minor dataframes with (54 rows;5 columns), and (293 rows;5 columns).

	District	Crop	Year	Area	Production
1	2	101	2013	1051	345
2	5	101	2013	1150	265
3	26	101	2013	1532	512
4	48	101	2013	2711	1037
5	37	101	2013	2788	998
6	53	101	2013	2836	695
7	7	101	2013	3021	975
8	55	101	2013	3984	1604
9	57	101	2013	7676	2645
10	2	101	2014	1014	325
11	5	101	2014	1455	364
12	26	101	2014	1554	531
13	48	101	2014	2049	921
14	53	101	2014	2710	662
15	37	101	2014	3102	1200

Figure 4.10: Major Data frame

	District	Crop	Year	Area	Production
1	6	101	2013	0	0
2	8	101	2013	0	0
3	9	101	2013	0	0
4	10	101	2013	0	0
5	12	101	2013	0	0
..	...	...	...	...	...
289	41	101	2018	90	42
290	24	101	2018	91	36
291	23	101	2018	93	44
292	39	101	2018	125	44
293	54	101	2018	159	99

Figure 4.11: Minor Data frame

Our target is to isolate data of Year 2018 for Training and Testing. We shall first set the 'x' parameter to consist of 'District', 'Crop', 'Year', and 'Area'. And, the 'y' parameter is to contain 'Production'. This step is common for all the 3 subset data frames.

```
x = df[['District', 'Crop', 'Year', 'Area']]
y = df['Production']
```

Figure 4.12: Setting 'x/y' Parameter

In the Major, Median, and Minor dataframes, indices 46, 69, and 243 mark the respective starting index of Year 2018. Henceforth, we shall begin Train-Test Splitting by calculating the percentage of data in the year 2018 as opposed to the past years 2013/17. This action corresponds to-

1. Major Dataframe- 16.6667% [Figure 4.2.8]
2. Median Dataframe- 13.92405% [Figure 4.2.9]
3. Minor Dataframe- 17.4061% [Figure 4.2.10]



```
xTrain, xTest, yTrain, yTest = train_test_split( x, y,
test_size=0.166666666, shuffle = False )
```

Figure 4.13: Major data frame train-test-split

```
xTrain, xTest, yTrain, yTest = train_test_split( x, y,
test_size=0.13924050, shuffle = False )
```

Figure 4.14: Median data frame train-test-split

```
xTrain, xTest, yTrain, yTest = train_test_split( x, y,
test_size=0.174061433, shuffle = False )
```

Figure 4.15: Minor data frame train-test-split

‘Shuffle’ command is set to ‘False’ for the purpose of maintaining the ascending order of indices and ‘Area’ parameter to provide visually inspectable displayable output.

Table 4.16 displays the ‘y\_test’ dataframes that are the constituent subsets of Major, Median, and Minor subset dataframes of the district wise Area vs. Production record of Pulse variant Gram labelled Crop 101 over the years 2013 to 2018.

Major dataframe y_Test		Median dataframe y_Test		Minor dataframe y_Test	
		69	79	282	27
		70	85	283	29
46	370	71	90	284	25
47	803	72	107	285	24
48	493	73	86	286	32
49	462	74	104	287	31
50	474	75	168	288	36
51	820	76	158	289	42
52	1441	77	156	290	36
		78	116	291	44
		79	186	292	44
				293	99

Figure 4.16: View of Major, Median and Minor Dataframes

In the following sections we shall discuss our proposed algorithms-

1. Linear Regression
2. Polynomial Regression
3. Decision Tree Regression
4. Random Forest Regression

The blocks of codes that are to be discussed are reusable for all degrees of Major, Median and Minor dataframes for all Crop Variations. The only change required is to set the index at designated lines to the starting index of Major, Median or Minor dataframe in scope.

## 4.2.1 Linear Regression

```
# Linear Regression

from sklearn.linear_model import LinearRegression

model = LinearRegression()

model.fit(xTrain,yTrain)

prediction = model.predict(xTest)

pred1 = pd.DataFrame(prediction)

pred1.index = pred1.index + 46

pred1.columns = ['Pred. Production']

diff1 = pd.DataFrame(prediction)

diff1 = ((prediction - yTest)**2)

diff1 = diff1**0.5

diff1.columns = ['Difference']
```

Figure 4.17: Code for Linear Regression

In this section we see the implementation of Linear Regression for the C101.Major dataset. The changes between Major, Median, and Minor datasets is the addition of +46, +69, and +243 respectively to 'pred1.index' at line 6.

Linear Regression appeared to be the least complex code to implement. In the first line we call in the module to initiate linear regression. We generate the object 'model' of class 'LinearRegression()' and use 'model.fit()' function to fit the xTrain and yTrain variables into the linear regression algorithm for learning. Next we use 'model.predict()' function fed 'xTest' to predict the 'Production' of Crop 101 in the year 2018, which is then stored into 'prediction'. The result of prediction data in 'prediction' are in list form and starting index from 0. We have converted the results into a dataframe labelled 'pred1', and added +46 to 'pred1.index', and set column name to 'Pred. Production'. The index of 'Pred. Production' dataframe starts from 46, which will be the merge location when we concat the dataframes 'xTrain', 'yTrains', 'pred1', and 'diff1' which is the root squared difference between the predicted results and actual 'Production' data.

## 4.2.2 Polynomial Regression

In this section we see the implementation of Polynomial Regression for the C101\_Median dataset. The changes between Major, Median, and Minor datasets is the addition of +46, +69, and +243 respectively to 'pred2.index' at line 9.

```
# Polynomial Regression

from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(degree = 3)

Xpoly = poly.fit_transform(xTrain)

poly.fit(Xpoly, yTrain)

lin2 = LinearRegression()

lin2.fit(Xpoly, yTrain)

poly_prediction = lin2.predict(poly.fit_transform(xTest))

pred2 = pd.DataFrame(poly_prediction)

pred2.index = pred2.index + 69

pred2.columns = ['Pred. Production']

diff2 = pd.DataFrame(poly_prediction)

diff2 = ((poly_prediction - yTest)**2)

diff2 = diff2**0.5

diff2.columns = ['MSD']
```

Figure 4.18: Code for Polynomial Regression

Firstly, we import 'PolynomialFeatures' functions from 'sklearn.preprocessing' class. This is for our aid in setting the degree of polynomial curve, fitting, training, and prediction. In the parameter 'poly', we had set the 'degree=3', which shall generate a curve with 3 maximum-minimum points. The following lines are the actions required to fit 'xTrain' in the polynomial learning methods by converting the dataframe to a polynomial function. As we store the result of the regression in 'poly\_prediction', we receive the results as a 'List' type. Therefore, we convert the result into the dataframe 'pred2' and add +69 to its index, which is the starting index for the C101\_Median range for Year 2018. Finally, we change the column name and calcu-

late the root squared difference.

### 4.2.3 Decision Tree Regression

In this section we see the implementation of Decision Tree Regression for the C101\_Minor dataset. The changes between Major, Median, and Minor datasets is the addition of +46, +69, and +243 respectively to 'pred3.index' at line 7.

```
# Decision Tree Regressor

from sklearn.tree import DecisionTreeRegressor

d_regressor = DecisionTreeRegressor(random_state =1)

d_regressor.fit(xTrain, yTrain)

d_prediction = d_regressor.predict(xTest)

pred3 = pd.DataFrame(d_prediction)

pred3 = pred3.dropna()

pred3.index = pred3.index +243

pred3.columns = ['Pred. Production']

diff3 = pd.DataFrame(d_prediction)

diff3.columns = ['MSD']

diff3 = ((d_prediction - yTest)**2)

diff3 = diff3**0.5

diff3.columns = ['MSD']
```

Figure 4.19: Code for Decision Tree Regression

Another simple implementation of a very effective algorithm. Firstly, we import the 'DecisionTreeRegressor' as object 'd\_regressor', where we pass in the dataframes 'xTrain', and 'yTrain' to fit the prediction algorithm and train itself. Then, we feed 'xTest' into function 'd\_regressor.predict()', which returns the 'List of lists' type result of the predictive algorithm is stored into the object 'd\_prediction', which is further converted to the dataframe 'pred3'. the dataframe column is set to read 'Pred. Production', and 'pred3.index' is added +243, that corresponds to the starting index for the year 2018 of C101\_Minor. Next, the root squared difference 'diff3'

is calculated and formed into a dataframe.

## 4.2.4 Random Forest Regression

In this section we see the implementation of Random Forest Regression for the C101\_Minor dataset. The changes between Major, Median, and Minor datasets is the addition of +46, +69, and +243 respectively to 'pred4.index' at line 9.

Implementation of Random Forest Regression requires finesse and understanding of how the algorithm operates to maximize the output efficiency.

```
from sklearn.ensemble import RandomForestRegressor

rf_regressor = RandomForestRegressor(max_depth=242, random_state=6,
n_estimators=250)

rf_regressor.fit(xTrain, yTrain)

rf_prediction = rf_regressor.predict(xTest)

pred4 = pd.DataFrame(rf_prediction)

pred4 = pred4.dropna()

pred4.index = pred4.index + 243

pred4.columns = ['Pred. Production']

diff4 = pd.DataFrame(pred4)

diff4 = ((rf_prediction - yTest)**2)

diff4 = diff4**0.5

diff4.columns = ['MSD']
```

Figure 4.20: Code for Random Forest Regression

In Random Forest Regression, we first establish the operational parameters 'n\_estimators' to 250, which shall generate 250 Decision Trees in the ensemble. Each Decision Tree in the Forest of depth 242, as per the amount of rows present in the 'xTrain' dataset of C10\_Minor. These parameters are set for maximizing the accuracy of the prediction. Lower estimates result in lower accuracy, and higher estimates provide no further significance to the accuracy.

All the other operations are similar to Decision Tree discussed in section 4.2.2, as in, fitting and training 'rf\_regressor' with 'xTrain', and 'yTrain', making prediction

results with 'xTest' by 'rf\_prediction', forming dataframe 'pred4' and adding +243 to 'pred4.index'. The root squared difference 'diff4' is then calculated.

### 4.3 Result Analysis

In this section, we shall discuss the results obtained from the algorithms of section 4.2. Among the 4 algorithms, Decision Tree and Random Forest Regressor appear to display the most accurate results. The lesser amount of data points available has greatly affected the training capacity of all the regression models; the contrast is evident from the amount of data points present in each Major, Median, and Minor subsets of Pulse type Gram labelled as Crop 101.

#### 4.3.1 Linear Regression

	District	Crop	Year	Area	Production	Pred. Production	Production
46	5	101	2018	946	370	437.477592	67.477592
47	2	101	2018	989	803	452.428185	350.571815
48	48	101	2018	1276	493	624.185645	131.185645
49	26	101	2018	1285	462	604.017444	142.017444
50	53	101	2018	1638	474	783.030174	309.030174
51	37	101	2018	2041	820	936.356286	116.356286
52	7	101	2018	2495	1441	1096.031982	344.968018
53	55	101	2018	3053	1514	1384.803320	129.196680
54	57	101	2018	5246	2243	2316.246669	73.246669

Figure 4.21: Crop101\_Major Linear Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production	
69	19	101	2018	219	79	104.116011	25.116011
70	51	101	2018	222	85	98.732263	13.732263
71	21	101	2018	223	90	105.263252	15.263252
72	60	101	2018	243	107	105.062368	1.937632
73	4	101	2018	244	86	116.916263	30.916263
74	40	101	2018	246	104	110.324433	6.324433
75	3	101	2018	275	168	129.185377	38.814623
76	34	101	2018	289	158	128.287268	29.712732
77	43	101	2018	351	156	150.573498	5.426502
78	44	101	2018	384	116	163.211505	47.211505
79	66	101	2018	424	186	174.274459	11.725541

Figure 4.22: Crop101\_Median Linear Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production	
279	14	101	2018	40	18	16.397723	1.602277
280	1	101	2018	41	14	16.035498	2.035498
281	22	101	2018	47	20	19.909192	0.090808
282	29	101	2018	55	27	23.791412	3.208588
283	33	101	2018	59	29	25.763062	3.236938
284	52	101	2018	62	25	28.219100	3.219100
285	11	101	2018	68	24	28.305755	4.305755
286	65	101	2018	79	32	36.354289	4.354289
287	17	101	2018	87	31	36.877037	5.877037
288	56	101	2018	88	36	39.691039	3.691039
289	41	101	2018	90	42	39.638482	2.361518
290	24	101	2018	91	36	39.031931	3.031931
291	23	101	2018	93	44	39.834512	4.165488
292	39	101	2018	125	44	54.630414	10.630414
293	54	101	2018	159	99	70.228897	28.771103

Figure 4.23: Crop101\_Minor Linear Regression Result



Linear Regression appears to be the second least accurate among the four algorithms. This is primarily due to the straight line that is projected along the data points.. Even if there were numerous data points available, Linear Regression would still provide the least accuracy since the data points are distributed over a wide range of 'x' variables. A straight fitting line shall always overfit and/or underfit the data. As seen on index 47 of Figure 4.3.1.1, the predicted production is 452.43 tonnes where the actual production stands at 803 tonnes, which is 350.57 tonnes greater. Even though District 2 has seen 5 repetitions from years 2013/17, the straight lines generated by the linear regression algorithms have undermined the production of District 2 greatly. C101\_Minor has the largest dataframe, and projects the results clearly. It shows evidence that, with lower values of 'Area', the line is overestimating the predicted results. And with increasing value of 'Area', the line begins to underestimate the predicted results. By sheer chance, the addition of the excess production predicted by the algorithms outweigh the underestimated production of District 2, such feats shall never be relied upon when the requirements of a nation are in scope.

### 4.3.2 Polynomial Regression

District	Crop	Year	Area	Production	Pred. Production	Production
46	5	101	2018	946	370	138.044662 231.955338
47	2	101	2018	989	803	421.840454 381.159546
48	48	101	2018	1276	493	275.570526 217.429474
49	26	101	2018	1285	462	79.249146 382.750854
50	53	101	2018	1638	474	208.387619 265.612381
51	37	101	2018	2041	820	465.639709 354.360291
52	7	101	2018	2495	1441	717.863602 723.136398
53	55	101	2018	3053	1514	917.026459 596.973541
54	57	101	2018	5246	2243	1606.308685 636.691315

Figure 4.24: Crop101\_Major Polynomial Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production
69	19	101	2018	219	79	96.683562
70	51	101	2018	222	85	99.949862
71	21	101	2018	223	90	97.936480
72	60	101	2018	243	107	89.356276
73	4	101	2018	244	86	102.621121
74	40	101	2018	246	104	110.916980
75	3	101	2018	275	168	111.024263
76	34	101	2018	289	158	113.683960
77	43	101	2018	351	156	131.336868
78	44	101	2018	384	116	138.368670
79	66	101	2018	424	186	155.824639

Figure 4.25: Crop101\_Median Polynomial Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production	
275	59	101	2018	26	15	11.330832	3.669168
276	63	101	2018	38	12	16.324701	4.324701
277	35	101	2018	39	18	15.227820	2.772180
278	46	101	2018	39	16	16.030297	0.030297
279	14	101	2018	40	18	15.558183	2.441817
280	1	101	2018	41	14	18.224078	4.224078
281	22	101	2018	47	20	17.197040	2.802960
282	29	101	2018	55	27	20.035479	6.964521
283	33	101	2018	59	29	21.852807	7.147193
284	52	101	2018	62	25	26.319756	1.319756
285	11	101	2018	68	24	22.658733	1.341267
286	65	101	2018	79	32	38.092686	6.092686
287	17	101	2018	87	31	27.758784	3.241216
288	56	101	2018	88	36	40.903505	4.903505
289	41	101	2018	90	42	36.243374	5.756626
290	24	101	2018	91	36	30.647133	5.352867
291	23	101	2018	93	44	30.993921	13.006079
292	39	101	2018	125	44	53.428153	9.428153
293	54	101	2018	159	99	94.868395	4.131605

Figure 4.26: Crop101\_Minor Polynomial Regression Result

The least accurate algorithm. However, the previous statement is partially correct, since Polynomial Regression appears to display far superior performance for C101\_Minor dataframe. The falter in C101\_Major and C101\_Median is that, the range defined is of too broad spectrum, and the 3rd degree curve generated by polynomial regression, for larger values of ‘Area’, intersects points few and far between the actual points of ‘Production’. This is the reason why Polynomial Regression shows large discrepancies for smaller datasets with broad ranges of ‘x’ values. However, the case for C101\_Minor is quite satisfactory. This dataframe is of a tighter range 0 to 199 acres of ‘Area’, and consists of 293 rows, which is almost 6 times the data available compared to C101\_Major, and C101\_Minor dataframes. The curve generated by the polynomial regression for this dataframe follows through very closely to the actual data points.

### 4.3.3 Decision Tree Regression

District	Crop	Year	Area	Production	Pred. Production	Production	
46	5	101	2018	946	370	367.0	3.0
47	2	101	2018	989	803	385.0	418.0
48	48	101	2018	1276	493	542.0	49.0
49	26	101	2018	1285	462	453.0	9.0
50	53	101	2018	1638	474	412.0	62.0
51	37	101	2018	2041	820	844.0	24.0
52	7	101	2018	2495	1441	1000.0	441.0
53	55	101	2018	3053	1514	1502.0	12.0
54	57	101	2018	5246	2243	2161.0	82.0

Figure 4.27: Crop101\_Major Decision Tree Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production	
69	19	101	2018	219	79	134.0	55.0
70	51	101	2018	222	85	86.0	1.0
71	21	101	2018	223	90	90.0	0.0
72	60	101	2018	243	107	95.0	12.0
73	4	101	2018	244	86	87.0	1.0
74	40	101	2018	246	104	88.0	16.0
75	3	101	2018	275	168	132.0	36.0
76	34	101	2018	289	158	100.0	58.0
77	43	101	2018	351	156	112.0	44.0
78	44	101	2018	384	116	116.0	0.0
79	66	101	2018	424	186	176.0	10.0

Figure 4.28: Crop101\_Median Decision Tree Regression Result

	District	Crop	Year	Area	Production	Pred. Production	Production
243	6	101	2018	0	0	0.0	0.0
244	9	101	2018	0	0	0.0	0.0
245	10	101	2018	0	0	0.0	0.0
246	12	101	2018	0	0	0.0	0.0
247	13	101	2018	0	0	0.0	0.0
248	18	101	2018	0	0	0.0	0.0
249	28	101	2018	0	0	0.0	0.0
250	32	101	2018	0	0	0.0	0.0
251	45	101	2018	0	0	0.0	0.0
252	50	101	2018	0	0	0.0	0.0
253	62	101	2018	0	0	0.0	0.0
254	69	101	2018	0	0	0.0	0.0
255	70	101	2018	0	0	0.0	0.0
256	68	101	2018	2	3	4.0	1.0
257	47	101	2018	3	1	1.0	0.0
258	67	101	2018	3	1	4.0	3.0
259	31	101	2018	4	1	1.0	0.0
260	36	101	2018	4	2	1.0	1.0
261	58	101	2018	4	2	1.0	1.0
262	27	101	2018	5	2	1.0	1.0
263	42	101	2018	5	2	1.0	1.0
264	71	101	2018	5	4	4.0	0.0
265	25	101	2018	6	2	2.0	0.0
266	8	101	2018	7	3	3.0	0.0
267	20	101	2018	7	3	2.0	1.0
268	38	101	2018	10	4	4.0	0.0
269	61	101	2018	10	7	6.0	1.0
270	16	101	2018	17	3	4.0	1.0
271	30	101	2018	19	7	9.0	2.0
272	45	101	2018	23	8	11.0	3.0
273	49	101	2018	24	10	11.0	1.0
274	64	101	2018	24	11	12.0	1.0
275	59	101	2018	26	15	15.0	0.0
276	63	101	2018	38	12	11.0	1.0
277	35	101	2018	39	18	18.0	0.0
278	46	101	2018	39	16	18.0	2.0
279	14	101	2018	40	18	25.0	7.0
280	1	101	2018	41	14	25.0	11.0
281	22	101	2018	47	20	18.0	2.0
282	29	101	2018	55	27	24.0	3.0
283	33	101	2018	59	29	28.0	1.0
284	52	101	2018	62	25	28.0	3.0
285	11	101	2018	68	24	25.0	1.0
286	65	101	2018	79	32	37.0	5.0
287	17	101	2018	87	31	30.0	1.0
288	56	101	2018	88	36	37.0	1.0
289	41	101	2018	90	42	37.0	5.0
290	24	101	2018	91	36	41.0	5.0
291	23	101	2018	93	44	41.0	3.0
292	39	101	2018	125	44	78.0	34.0
293	54	101	2018	159	99	100.0	1.0

Figure 4.29: Crop101\_Minor Decision Tree Regression Result

First of the 2 most accurate algorithms. Rarely displays large discrepancies for smaller dataframes with broad range distribution of C101\_Major and C101\_Minor.

Perfectly accurate predictions have been achieved for some data points of Crop101\_Median, which is due to the range being tighter than Crop101\_Major; respectively, 200 to 999 acres as opposed to 1000+ which exceeds 5000 acres of cultivated 'Area'. The previous statement is further certified by figure 4.3.3.3, which displays the regression result for C101\_Minor. This dataframe consists of a tight range 0 to 199 acres, and 293 rows of data points. The algorithm has truly liberated its potential and provides the most accurate and desirable predictions. The modal class stands at less the 3 tonne difference with the most repeated numbers, in ascending order, are 0,1,3. Decision Tree is a prime candidate for future developments.

#### 4.3.4 Random Forest Regression

46	5	101	2018	946	370	398.936	28.936
47	2	101	2018	989	803	392.440	410.560
48	48	101	2018	1276	493	495.848	2.848
49	26	101	2018	1285	462	473.964	11.964
50	53	101	2018	1638	474	503.136	29.136
51	37	101	2018	2041	820	771.340	48.660
52	7	101	2018	2495	1441	974.108	466.892
53	55	101	2018	3053	1514	1450.988	63.012
54	57	101	2018	5246	2243	2322.272	79.272

Figure 4.30: Crop101\_Major Random Forest Regression Result

District	Crop	Year	Area	Production	Pred. Production	Production
69	19	101	2018	219	79	112.738000
70	51	101	2018	222	85	95.102952
71	21	101	2018	223	90	95.116952
72	60	101	2018	243	107	95.604533
73	4	101	2018	244	86	89.345333
74	40	101	2018	246	104	101.810724
75	3	101	2018	275	168	152.523429
76	34	101	2018	289	158	123.627086
77	43	101	2018	351	156	143.015000
78	44	101	2018	384	116	123.516000
79	66	101	2018	424	186	165.336000

Figure 4.31: Crop101\_Median Random Forest Regression Result

	District	Crop	Year	Area	Production	Pred. Production	Production
243	6	101	2018	0	0	0.000	0.000
244	9	101	2018	0	0	0.000	0.000
245	10	101	2018	0	0	0.000	0.000
246	12	101	2018	0	0	0.000	0.000
247	13	101	2018	0	0	0.000	0.000
248	18	101	2018	0	0	0.000	0.000
249	28	101	2018	0	0	0.000	0.000
250	32	101	2018	0	0	0.000	0.000
251	45	101	2018	0	0	0.000	0.000
252	50	101	2018	0	0	0.000	0.000
253	62	101	2018	0	0	0.000	0.000
254	69	101	2018	0	0	0.004	0.004
255	70	101	2018	0	0	0.000	0.000
256	68	101	2018	2	3	1.856	1.144
257	47	101	2018	3	1	1.044	0.044
258	67	101	2018	3	1	2.948	1.948
259	31	101	2018	4	1	1.164	0.164
260	36	101	2018	4	2	1.044	0.956
261	58	101	2018	4	2	2.296	0.296
262	27	101	2018	5	2	1.256	0.744
263	42	101	2018	5	2	1.132	0.868
264	71	101	2018	5	4	3.284	0.716
265	25	101	2018	6	2	2.020	0.020
266	8	101	2018	7	3	2.936	0.064
267	20	101	2018	7	3	2.132	0.868
268	38	101	2018	10	4	3.936	0.064
269	61	101	2018	10	7	5.336	1.664
270	16	101	2018	17	3	4.540	1.540
271	30	101	2018	19	7	8.760	1.760
272	45	101	2018	23	8	9.596	1.596
273	49	101	2018	24	10	9.848	0.152
274	64	101	2018	24	11	12.236	1.236
275	59	101	2018	26	15	13.080	1.920
276	63	101	2018	38	12	11.704	0.296
277	35	101	2018	39	18	16.488	1.512
278	46	101	2018	39	16	15.248	0.752
279	14	101	2018	40	18	15.300	2.700
280	1	101	2018	41	14	17.040	3.040
281	22	101	2018	47	20	20.660	0.660
282	29	101	2018	55	27	25.296	1.704
283	33	101	2018	59	29	26.912	2.088
284	52	101	2018	62	25	25.492	0.492
285	11	101	2018	68	24	25.064	1.064
286	65	101	2018	79	32	34.088	2.088
287	17	101	2018	87	31	29.140	1.860
288	56	101	2018	88	36	33.916	2.084
289	41	101	2018	90	42	35.432	6.568
290	24	101	2018	91	36	36.548	0.548
291	23	101	2018	93	44	37.616	6.384
292	39	101	2018	125	44	67.348	23.348
293	54	101	2018	159	99	93.560	5.440

Figure 4.32: Crop101\_Minor Random Forest Regression Result



The second most accurate predictive algorithm. Since this Random Forest generates 250 Decision Trees, the results are deemed to be more accurate as opposed to Decision Tree Regression. However, the claim deserves more consideration. Because, both Decision Tree Regression and Random Forest Regression may seem to display nearly the same results for C101\_Major and C101\_Median dataframes. In reality, Decision Tree displays more precision and accuracy for broad range : low datapoint datasets as the 2 aforementioned subset dataframes. Random Forest displays otherwise, since its results are derived from 250 other decision trees, and is outmatched by the precision of a single decision tree. But, in case of the tighter range : higher datapoint dataset like C101\_Minor, Random Forest and Decision Trees are both extremely viable, desirable, and potent algorithms.

# Chapter 5

## Conclusion

### 5.1 Overview

In this paper, we have established the essential methods to project the production of crops in the upcoming years. The proposed methods are commercially applicable to all variants of crops that are cultivated in Bangladesh and/or any other country. Bangladesh Bureau of Statistics (BBS), Bangladesh Agriculture Research Centre (BARC), Bangladesh Meteorological Department (BMD), and all other associated and affiliated institutes already possess the necessary production and weather data in .xls or .csv format. The transposition operation discussed in section 3.1 is not necessary in commercial approaches. We shall summarize the critical processes as follows-

1. Appropriately dispatch null and missing values. [Section 3.1]
2. Assign predetermined 'Integer' values to all 'String' type data points. [Section 4.1]
3. Sort dataframe in the ascending order of 'Year' and 'Area' or 'Production' [Section 4.1]
4. Designate appropriate ranges to the data point indices and split the dataframe to subset dataframes. [Section 4.1]
5. Sort the resultant subset dataframes in the ascending order of 'Year' and 'Area'. [Section 4.1]
6. Process the subsets into the algorithms individually. [Section 4.2]
7. Tabulate and analyze the result provided by the prediction algorithms. [Section 4.3]

In the next step, we process the consumption of Pulse variant Gram labelled Crop 101 for the year 2018.

According to United Nations' World Population Prospects- in the year 2017, the population of Bangladesh stood at 156.5 millions, with a population growth of 1.08%. According to FAOSTAT 2018, a total of 5.4 kg per capita per year pulses are consumed.

	2017	2018
Population (millions)	156.5	169.3
Pulses Produced (tonnes)	389,806	394,890
Pulses consumed (tonnes)	845,100	914,200
Deficit (tonnes)	455,294	519,310

Figure 5.1: Population Growth-Pulse Consumption-Production Deficit

In 2018, a total of 394,890 tonnes of pulses have been produced, and 914,200 tonnes were consumed. The numbers stated are not absolute. Primarily because, the consumption rate of pulses 5.4 kg per capita per year is a mean estimation. Secondly, approximately 15% of total production is allocated to several safety net protocols and food storage programmes. Bangladesh is being bombarded with tropical storms and flash floods every year, therefore, the food storage programs are constantly in effect. Apart from that, several trade procedures pertain to export/import of several food/cash crops. The deficit in consumption is then to be considered to estimate the amount of pulses that needs to be imported from international sources.

## 5.2 Research Challenges

Our research is a fruitful product of manual labor through several difficulties that we had to validate with appropriate solutions. The list below portrays some of the challenges-

1. There is nor .xls or .csv formatted production data available online, so we had to collect datasets from Yearbooks of Agricultural Statistics 2013/19.
2. All the data were in .pdf format, so we had to formulate several methods and approaches to extract the data with Microsoft Excel.
3. The data extracted from .pdf files into .xls were beyond unorganized. There had been several rows between rows, columns between columns, merged rows and columns, several data points inserted in a single cell, etc. The dataset demanded manual scrutiny and cleanup.
4. Since the machine learning algorithms cannot process character type data, 'District' and 'Crop' data as 'Strings' are required to be converted to 'Int' types. This designated label encoding function is extremely complex via the algorithm, and built-in swift label encoding protocols randomly assign integers to strings. These 'int' values are randomized in every iteration, making identification and analysis extremely difficult by humans. And so, 3702 rows of data had to be manually inspected and label assigned via multiple built-in replacement methods and iterations via Microsoft Excel before processing into the algorithms.
5. The datapoints in 'yTrain' dataset fluctuate within a very broad spectrum, so, one algorithm shall not suffice for all variants of a crop. Therefore, each crop variant

must be individualized. 6. Even for single crop variants, the ‘Area’ and ‘Production’ datasets are on a broad range, and operating the predictive algorithms on the broad dataset displays ridiculously inaccurate results. Therefore, the dataset for a single variant must be split into tight ranged subset (Major, Median, Minor), and each subsets must be operated on individually. For commercial use, a single, more sophisticated algorithm may be professionally developed.

7. Due to the non-availability of elaborated production data for the years prior to 2013, the smaller datasets of Median and Major ranges are greatly affected by inaccuracies of predictive algorithms. The elaborate datasets in question are present in the archives of affiliated authorities, but are not made accessible to the public/ for research purposes.

8. The total number of Weather Stations provided measure half of the District count, 35 stations as opposed to 75 individual ‘District’ data points. Moreover, not all weather stations correspond to the district of production of Crop 101. The resultant dataset generated for Weather to Production correspondents was miniature, shall provide extremely inaccurate and skewed results, and not a feasible commercial approach. The same procedure for the projection of crop production discussed in Section 5.1 can be repurposed for accurate projection of future meteorological effects. A part of which has been our preliminary research observation with crude unoptimized algorithm, trained with the year 2008 and tested on the year 2009 on Average, which still processed comparably desirable results.

Linear Regression					Polynomial Regression				
Station+Year	Actual	Prediction	Difference		Station+Year	Actual	Prediction	Difference	
0	2011	83	77.509587	5.490413	0	2029	79	78.292224	0.707776
1	2022	80	78.340019	1.659981	1	2011	83	76.976306	6.023694
2	2024	81	78.491006	2.508994	2	2033	82	79.457118	2.542882
3	2026	76	78.641994	-2.641994	3	2038	83	80.003509	2.996491
4	2026	81	78.641994	2.358006	4	2039	82	79.923840	2.076160
5	2044	82	80.000682	1.999118	5	2042	81	79.184593	1.815407
6	2032	79	79.094957	-0.094957	6	2010	81	77.574193	3.425807
7	2014	80	77.736069	2.263931	7	2031	85	78.924058	6.075942
8	2028	78	78.792982	-0.792982	8	2042	80	79.184593	0.815407
9	2032	81	79.094957	1.905043	9	2038	76	80.003509	-4.003509
10	2012	83	77.585081	5.414919	10	2032	81	79.206300	1.793700

Decision Tree					Random Forest				
Station+Year	Actual	Prediction	Difference		Station+Year	Actual	Prediction	Difference	
0	2010	73	81.0	-8.0	0	2015	80	74.605000	5.395000
1	2044	82	82.0	0.0	1	2008	76	77.484500	-1.484500
2	2014	80	80.0	0.0	2	2019	83	35.378500	47.621500
3	2026	81	76.0	5.0	3	2037	81	78.593571	2.406429
4	2037	81	79.5	1.5	4	2025	80	78.971833	1.028167
5	2026	75	76.0	-1.0	5	2035	82	80.237429	1.762571
6	2016	73	70.0	3.0	6	2032	82	80.231000	1.769000
7	2036	83	79.5	3.5	7	2038	79	80.766833	-1.766833
8	2024	81	78.0	3.0	8	2016	70	73.050000	-3.050000
9	2035	82	80.5	1.5	9	2037	81	78.593571	2.406429
10	2034	75	82.5	-7.5	10	2036	80	80.858905	-0.858905

Figure 5.2: Results of prediction on Average Humidity

Of the four algorithms, Linear and Polynomial regressions provide the maximum accuracy and least variance. This is primarily due to the tight variance in humidity data points across the ‘Stations’. If the humidity had been highly fluctuating data points similar to ‘Area vs. Production’ datasets, then-

Firstly, the dataframe again required to be split into subset dataframes. Secondly, due to the fluctuations, lines and curves generated by Linear and Polynomial Regression shall greatly underestimate/overestimate the resultant predicted data similar to results in Section 4.3. Decision Tree and Random Forest regression shall again prove worthy contenders.

9. From our ventures into Satellite Image Analysis, we have drawn that- not all satellites in Earth orbit cover the Land area of Bangladesh. Whichever satellites do cover an area, consecutive images are at a minimum of one year apart. Whatever images are found, only the vegetated areas and farmlands can be identified. Closer

zoom distorts the resolution of images to the extent that the crop types cannot be identified. Better resolution of the areas are zoomed images from farther away, for which, again, the crop types cannot be identified. The use of satellite image processing is currently deemed unworthy contender for commercial developments, and purely for research purposes. Even so, our recommendation is to apply UAV image stitching methods proposed by the paper ‘Pilot Study on the Development of a Low Cost Land Observation and Geo-information Retrieval System for Emergency Monitoring of Agricultural Crop and Disaster-Induced Damages’ supervised by Associate Professor Dr. Md. Khalilur Rahman (ph.D) from the Department of Computer Science and Engineering of BRAC University. The paper presents far more valuable insight and feasible methods of using images and real-time meteorological data recording and utilization, commercial use, and for further research purposes.

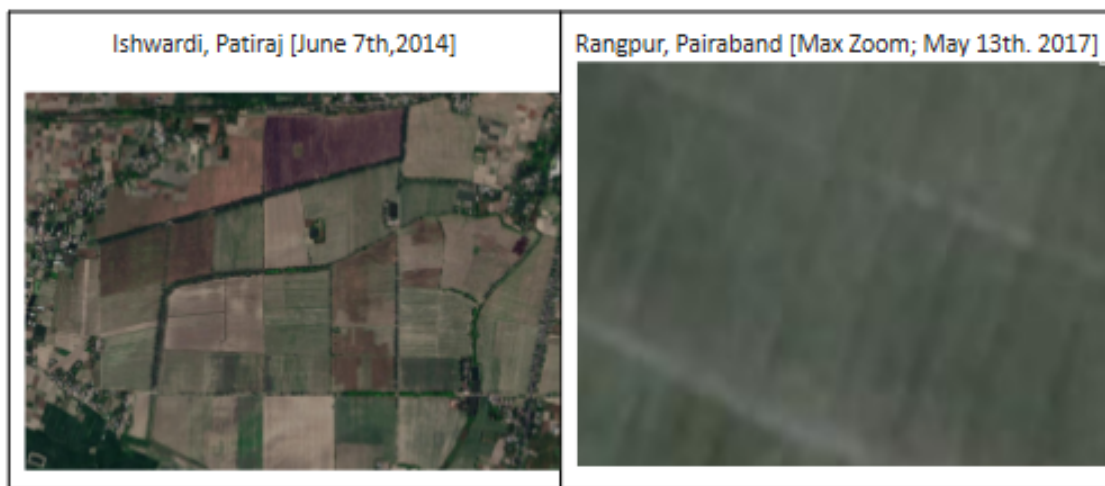


Figure 5.3: Satellite Images

### 5.3 Contribution & Impact

Agriculture provides 12.9% of the Gross Domestic Production of Bangladesh economy as of the year 2019, and so, 16.5 million farming families and their lives are directly affiliated with the harvest. Every year, tropical storms and flash floods destroy thousands of tonnes of crops, and the effects are drastical on the hard laboring farmers, as well as the population of Bangladesh alike. With the repeated strikes of the global pandemic COVID-19, the state of trades and markets have been upturned and deemed in a chaotic neutral state. The lower income households and farmers’ families are intensely affected by the conditions of the markets. Closure of trade routes and transportations oftentimes strike the prices of daily products and produce to unobtainable states. The research was initiated by the price hike of onions in 2019, due to the lack of necessary tools and knowledge to estimate market demands and public consumption. And now, the market prices of several essential commodities have hiked up, mostly due to the several conditions encompassing Supply and Demand. Our market, and our people, are in desperate need of a tool that can project the production, consumption, and the import/export states of food

crops and essential commodities. Therefore, we have proposed this research for the national administration to utilize and project the aforementioned quotas, plan on the food security and availability status, and stabilize the Supply-Demand ratios and prices. Since the proposed methodology is applicable for all variants of crops, and on all measures of lands, the method can be integrated by all nations of the world in their designated agriculture units. Utilizations of such modern tools can, in fact, stabilize the global economy.

## 5.4 Future Work

Our proposed method and algorithms appear to have several rooms for further developments, especially for commercial purposes.

A far more sophisticated method can be applied while encoding labels of ‘Districts’ and ‘Stations’ through the algorithms instead of manually assigning integer values in the dataframe before feeding into the algorithms. This one method can substantially minimize the complexity of building and operating the predictive systems.

The splitting of dataframes to ranged subsets can be approached with further analysis to mitigate the effects of variance between data points, which greatly affect the accuracy of the predictive algorithms. For the essential professionally designed commercial use of the methods, we suggest collecting the restricted archives of the elaborated production data prior to the year 2013.

In Section 2.1, we have addressed several papers that had considered various attributes such as soil type, moisture and mineral contents, temperature, fertilizer use etc. Also, papers that proposed use of android applications to estimate time between cultivation and harvesting. We have discussed the various approaches to estimate the household and per capita consumptions accurately. All such methodology can and must be incorporated to fortify the predictive capabilities of our proposed methodology.

Similar methods are to be followed for meteorological data to project the changes in weather, and Variable Analysis, and Codependency etc features to be analyzed to build the relation between weather situations and its effects on crop production.

Finally, a more sophisticated and results yielding approach should be devised in order to record, retrieve and analyze images of land areas being cultivated and harvested.

# Bibliography

- [1] Kaur, K. (2016). Machine learning: applications in indian agriculture. *Int. J. Adv. Res. Comput. Commun. Eng.(IJARCCE)*, 5(4).
- [2] Jain, N., Kumar, A., Garud, S., Pradhan, V., and Kulkarni, P. (2017). Crop selection method based on various environmental factors using machine learning.
- [3] ] Hassan, Mahmud, F., Zakir, Binta, S., Binta, Shamma, A., ... Maruf, S. (2018, December 1). Agricultural yield and profit prediction using data analysis techniques. Retrieved from <http://dspace.bracu.ac.bd/xmlui/handle/10361/11449>
- [4] Supply and Demand for Cereals in Bangladesh, 2010-2030. (2012). Retrieved from <http://ebrary.ifpri.org/utills/getfile/collection/p15738coll2/id/126957/filename/127168.pdf>
- [5] ] Amin, A., Farhana, I., Afee, Nobi, T. (2018, December 1). Crop prediction based on geographical and climatic data using machine learning and deep learning. Retrieved from <http://dspace.bracu.ac.bd/xmlui/handle/10361/11429>
- [6] <http://bbs.gov.bd/>
- [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- Geoinformat-ics (Agro-Geoinformatics), pp. 1–6, Aug 2012
- [8] Alif, Al Amin, Shukanya, Israt Farhana Afee, Tasnia Nobi (2018) Crop prediction based on geographical and climatic data using machine learning and deep learning. Retrieved from <http://dspace.bracu.ac.bd/xmlui/handle/10361/11449>
- [9] Begum, M., Dhaese, L. (2010). Supply and demand situations for major crops and food items in Bangladesh. *Journal of the Bangladesh Agricultural Univer-sity*, 8(1), 91–102. doi: 10.3329/jbau.v8i1.6405
- [10] Tasnim, F., Ahmed, Rifat, Islam. (2017). Prediction of rainfall using data mining techniques. Retrieved from <http://dspace.bracu.ac.bd/xmlui/handle/10361/9489>
- [11] November, M. R., Rubayet, M. (2019, November 29). Lessons from the onion crisis. *The Business Standard*. <https://www.tbsnews.net/opinion/lessons-onion-crisis>.



- [12] Linear Regression in Machine learning - Javatpoint. [www.javatpoint.com](http://www.javatpoint.com). (n.d.). <https://www.javatpoint.com/linear-regression-in-machine-learning>.
- [13] Brownlee, J. (2020, August 14). Linear Regression for Machine Learning. Machine Learning Mastery. <https://machinelearningmastery.com/linear-regression-for-machine-learning/>.
- [14] Krishna, A. (2019, April 2). Machine Learning Algorithms-Linear Regression. Medium. <https://medium.datadriveninvestor.com/machine-learning-algorithms-linear-regression-f89ab64ac490>.
- [15] Machine learning Polynomial Regression - Javatpoint. [www.javatpoint.com](http://www.javatpoint.com). (n.d.). <https://www.javatpoint.com/machine-learning-polynomial-regression>.
- [16] Donges, N. (n.d.). A complete guide to the random forest algorithm. Built In. <https://builtin.com/data-science/random-forest-algorithm>.
- [17] Machine Learning Random Forest Algorithm - Javatpoint. [www.javatpoint.com](http://www.javatpoint.com). (n.d.). <https://www.javatpoint.com/machine-learning-random-forest-algorithm>.
- [18] Random Forest. The Science of Machine Learning. (n.d.). <https://www.ml-science.com/random-forest>.
- [19] Machine Learning Decision Tree Classification Algorithm - Javatpoint. [www.javatpoint.com](http://www.javatpoint.com). (n.d.). <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>.
- [20] Decision Tree Introduction with example. GeeksforGeeks. (2021, March 17). <https://www.geeksforgeeks.org/decision-tree-introduction-example/>.
- [21] Linear Regression: Simple Steps, Video. Find Equation, Coefficient, Slope. (2020, July 06). Retrieved October 04, 2020, from <https://www.statisticshowto.com/probability-andstatistics/regression-analysis/find-a-linear-regression-equation/>
- [22] Technologies, M. (2017, October 05). Polynomial Regression: Polynomial Regression Formula - Tutorial. Retrieved October 04, 2020, from <https://mindmajix.com/polynomialregression>
- [23] Li, L. (2019, May 16). Classification and Regression Analysis with Decision Trees. Retrieved October 04, 2020, from <https://towardsdatascience.com/https-medium-com-lorrrli-classification-and-regression-analysis-with-decision-trees-c43cdb58054>
- [24] Schott, M. (2020, February 27). Random Forest Algorithm for Machine Learning. Retrieved October 04, 2020, from <https://medium.com/capital-one-tech/random-forestalgorithm-for-machine-learning-c4b2c8cc9feb>
- [25] GDP share of agriculture by country, around the world. (n.d.). Retrieved October 04, 2020, from [https://www.theglobaleconomy.com/rankings/Share\\_of\\_agriculture/](https://www.theglobaleconomy.com/rankings/Share_of_agriculture/)

- [26] Plecher, P., 28, J. (2020, July 28). Bangladesh - share of economic sectors in the gross domestic product2019.RetrievedOctober04,2020,from <https://www.statista.com/statistics/438359/share-of-economic-sectors-in-the-gdp-inbangladesh/>