

Bachelor of Science in Computer Science & Engineering



**Paddy Yield Prediction under the Influence of
Weather using Machine Learning**

by

Md. Junayed Hossain

ID: 1504112

Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

May, 2021

Paddy Yield Prediction under the Influence of Weather using Machine Learning



Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Md. Junayed Hossain

ID: 1504112

Supervised by

Muhammad Kamal Hossen

Associate Professor

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

The thesis titled ‘**Paddy Yield Prediction under the Influence of Weather using Machine Learning**’ submitted by ID: 1504112, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Muhammad Kamal Hossen

Associate Professor

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Md. Mokammel Haque

Professor & Head

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Abu Hasant Mohammad Ashfak Habib

Professor

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

I owe thanks to a number of people without whom this project would not be possible. First and foremost, I would like to thank my supervisor, Associate Professor Muhammad Kamal Hossen, for his continuous support and assistance. His dedication and curiosity to do unique things have helped me to do and fully understand this project and its scope of work and had patience, understanding, and willingness to let me find my own wings, have all contributed to my humble development of this project and training the dataset. I am grateful to him for his constant encouragement and his stimulating ideas. I would like to thank my defense committee members, for providing me with their valuable advice and feedback. I am incredibly grateful to Prof. Dr. Assaduzzaman, Head, Dept. of CSE, CUET for his outstanding support through my undergraduate education.

Abstract

Agriculture has a big influence in the food sector as well as in the economic sector of a country. But nowadays, it is not as influential as it was before. In the agricultural field, paddy is one of the most important crops which has the most demand in our country. In order to have a better production of paddy we have made an approach to develop a model that would help them know the production of the paddy before they should cultivate and they also can know about the upcoming temperature and rainfall. A model name SARIMA has been used for weather forecast and random forest algorithm that has been used to predict the production of paddy. As these models have better accuracy that's why this has been chosen. Finally, based on the predicted output, this model will an overview of the seasonal production of paddy for every district individually.

Table of Contents

Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Introduction	1
1.1.1 Problem Statement	1
1.1.2 Objectives	2
1.2 Framework/Design Overview	3
1.3 Difficulties	3
1.4 Applications	3
1.5 Motivation	4
1.6 Contribution of the thesis	4
1.7 Thesis Organization	5
1.8 Conclusion	5
2 Literature Review	6
2.1 Introduction	6
2.1.1 Machine Learning	6
2.1.1.1 Regression	6
2.1.1.2 Classification	7
2.1.2 Regression Model	7
2.1.2.1 Linear Regression	7
2.1.2.2 Decision Tree	7
2.1.2.3 K-Nearest Neighbours	7
2.1.2.4 Random Forest	8
2.1.2.5 Gradient Boosting Regression	8
2.2 Related Literature Review	8
2.3 Conclusion	10
2.3.1 Implementation Challenges	10

3	Methodology	11
3.1	Introduction	11
3.2	Diagram/Overview of Framework	11
3.3	Detailed Explanation	11
3.3.1	Implementation	11
3.3.1.1	System Requirements	12
3.3.1.2	Hardware Requirements	12
3.3.1.3	Software Requirements	13
3.3.2	Implementation Details	13
3.3.2.1	Data Collection	13
3.3.2.2	Data Cleaning	14
3.3.2.3	Data Preprocessing	14
3.3.2.4	Applying Haversine Formula	16
3.3.2.5	Feature Extraction	19
3.3.2.6	Regression Algorithm	20
3.3.2.7	Temperature and Rainfall Prediction	20
3.4	Conclusion	21
4	Results and Discussions	22
4.1	Introduction	22
4.2	Dataset Description	22
4.3	Impact Analysis	23
4.3.1	Social and Environmental Impact	23
4.3.2	Ethical Impact	23
4.4	Evaluation of Framework	23
4.5	Evaluation of Performance	27
4.5.1	Experimental Results	29
4.6	Conclusion	31
5	Conclusion	32
5.1	Conclusion	32
5.2	Future Work	32

List of Figures

1.1	Design of the Paddy Yield Prediction	3
3.1	Schematic Diagram of the Proposed System.	12
3.2	Production of paddy	14
3.3	Temperature	15
3.4	Rainfall	16
3.5	Metro Stations	17
3.6	Assigning Nearest Stations	18
3.7	Preprocessed Dataset	19
3.8	Heatmap	20
4.1	Linear Regression	24
4.2	Decision Tree Regression	24
4.3	KNN Regression	25
4.4	Random Forest Regression	25
4.5	Gradient Boosting Regression	26
4.6	Overview of Temperature	26
4.7	Overview of Rainfall	27
4.8	Weather prediction	30
4.9	Production	31

List of Tables

3.1	Overview of Dataset	14
4.1	Dataset Summary	23
4.2	Predicted Temperature and Rainfall	27
4.3	Comparison of Algorithm	29
4.4	Predicted Result	30

Chapter 1

Introduction

1.1 Introduction

Agriculture is the foundation of all economies. Advances in agriculture are expected to meet the needs of a country like Bangladesh, which has an ever-increasing demand for food due to the rising population. Agriculture has been considered the primary and foremost culture practiced in Bangladesh since ancient times. It is Bangladesh's most important source of jobs. This sector's success has a huge effect on major macroeconomic goals including job creation, poverty reduction, human capital growth, food security, and so on. Agriculture has been considered one of the most important occupations in our world. It is the backbone of our economy and contributes to the country's overall growth. In the agricultural field, paddy is one of the most important parts. So, predicting the production of paddy is going to make a good impact on our agriculture and economic system.

1.1.1 Problem Statement

Training machines to learn and generate models for future prediction is a commonly used science. Machine learning is the term for this process. As an agricultural country, our economy is heavily reliant on this sector. Paddy is one of the most important of them, having played an important role in the growth of our economic system. Nearly 60 percent of the country's land is used for agriculture, to meet the needs of a billion people. Rural areas are home to more than 70 percent of Bangladesh's population and 77 percent of its workers. Agriculture employs nearly half of all Bangladeshi employees, and two-thirds of those in rural areas, and about 87 percent of rural households rely on agriculture for at

least part of their income (World Bank, 2016). The agriculture sector contributes about 13.82 percent to the country's Gross Domestic Product (GDP) and employs more than 45 percent of the total labor force (BBS, 2017). Paddy is a major cereal crop in agriculture that contributes to national food security and socio-economic development [1]. Timely harvesting of paddy is very important to reduce losses affecting the total yield. Due to the unavailability of a mechanical harvesting system, a significant amount of field losses of paddy every year have been occurred due to natural calamities and shortage of time during the harvesting period. Nowadays, timely harvesting of paddy is a big challenge due to shortage of labor and high wages of labor. Yet evidence indicates a progressive shrinking of rural labor availability, as workers migrate to cities or abroad to engage in more remunerative employment. As there are 3 types of paddy which Aus, Aman, and Boro. Three of these paddies have a different season of harvest. So, different climate plays a different role on the production of paddy. The change of climate has become a bigger issue for the harvest of paddy. Because of increasing climate vulnerability, it is a great challenge to keep the pace of food production for the exponential growth of the population in Bangladesh. For traditional paddy harvesting methods, a significant amount of field losses has occurred every year[2]. These are one of the biggest challenges to overcome so that we get rid of the economic loss and increasing the food security of our own people.

1.1.2 Objectives

Our aim is to give our farmers knowledge of their production of paddy so that they can be familiar with their yield rate of paddy depending on the influence of the weather. With the change of the atmosphere how production changes will be noticed as well. Besides this, the outcomes of this systems are

- To predict the yield of paddy.
- To get an overview of every type of paddy's production.
- To know the influence on the production under changing weather.

1.2 Framework/Design Overview

The general design of whole system is consist of some components. The design is geven below: For building this system the datasets were the priority. After that,



Figure 1.1: Design of the Paddy Yield Prediction

the datasets were preprocessed and features were extracted from datasets. Regression models were used to train and test. Finally, the output can be generated from the model.

1.3 Difficulties

Collecting the dataset of the 64 districts is always going to be a big challenge Without standard dataset machine learning cannot operate it is expected to. So, collecting the data has been the prime task. The more standard dataset the better model will become. Otherwise, For getting a better output more accurate dataset is required. To make the dataset more accurate we have to label the datasets by assigning numerical values. The reason behind labeling the dataset is because we have used supervised algorithms. After assigning numerical values we have normalized the dataset so that value of every attribute has equal influence for making the system and output much better. If the data is wrong then it will bring wrong output as well as the performance of the system will also decrease. That's why we have collected datasets from various reliable sources.

1.4 Applications

Agriculture is becoming less popular nowadays. The prominent reason behind this is less use of scientific technology. That's why we are trying to use the machine learning technique so that agricultural systems get accustomed to scientific

technology. That's why the paddy yield prediction system is being developed. The applications are given below:

- Awareness about production.
- Weather monitoring.
- Development of agricultural economic system.
- Digital Agricultural industry creation.

1.5 Motivation

As our agricultural system is not as developed as expected as the country are. So, making this system developed and accustomed to scientific tools and technology is the main motivation. By developing this system we can also become enrich in our economic system. As the population is increasing at a very large rate. So keep up with the food requirements of our increasing popularity. Another motivation behind this that our farmers are deprived of the facility that they deserve as well as they are not well equipped with modern technology. By having a much more developed system that can connect our agricultural system with science and the farmers can have a better facility. Paddy yield prediction can become a step forward for making this dream true.

1.6 Contribution of the thesis

Contributions of this work are given below:

- We have developed the previous year's paddy yield of datasets of 3267.
- We have developed rainfall and temperature datasets as well.
- We have designed and developed a can predict the weather forecast and paddy yield as well.
- We have used regression algorithms for predicting the paddy yield.
- We have evaluated system performance according to the accuracy of the model.

1.7 Thesis Organization

The entire report is represented in five different chapters. We can outline the report structure as follows:

- Chapter 1 gives an overview of paddy yield prediction using machine learning and discusses the motivation behind this project. It also lists the objectives of the project and the challenges we may face in this work.
- Chapter 2 gives an overview of the literature review and related works and implementation challenges.
- Chapter 3 gives an overview of the proposed methodology, Data processing, regression algorithm, training set, testing set, and Implementation.
- Chapter 4 gives an overview of the datasets being used to train the model as well as the impact of the thesis and evaluation of performance and framework. It also gives an overview of experimental analysis, results, and evaluation.
- Chapter 5 gives the conclusion to this work and how it should be improved further in the future.

1.8 Conclusion

From the above discussion, we have got to know about the agricultural system and our objectives and opportunities of machine learning in paddy yield prediction. We have also discussed the limitations and the challenges we are going to face ahead. We will try our best to overcome those challenges and fulfill our objectives.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we will shortly describe machine learning and regression analysis and learn about regression algorithms such as K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting Machine learning method which is really useful for calculating numerical and continuous values which are important to understand. This chapter also contains a brief discussion on related previous works that are already implemented, their limitations.

2.1.1 Machine Learning

Machine learning is the study of computer algorithms that develop themselves over time as a result of experience and data. Machine learning is a form of data analysis that automates the development of analytical models. It's a branch of artificial intelligence focused on the premise that computers can learn from data, recognize patterns, and make decisions with little to no human input [3]. In this proposed system we are going to use supervised learning. There are two types of supervised learning which are classification and regression.

2.1.1.1 Regression

Using training data, regression produces a single output value. This is a probabilistic interpretation that is determined by taking into account the strength of association between the input variables [4].

2.1.1.2 Classification

It entails classifying the information. You may use classification to decide whether or not an individual would default on a loan if you are considering extending credit to them. Binary classification occurs when a supervised learning algorithm labels input data into two distinct groups. Multiple classifications refer to the division of information into more than two classes.

2.1.2 Regression Model

In this proposed system, regression models have been used to predict the outcome. The regression algorithms such as K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting are being used to predict the outcome.

2.1.2.1 Linear Regression

A linear approach to modeling the relationship between a scalar response and one or more explanatory variables is known as linear regression in statistics. Easy linear regression is used when there is just one explanatory variable; multiple linear regression is used when there is more than one.

2.1.2.2 Decision Tree

The Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree algorithm, unlike other supervised learning algorithms, can also be used to solve regression and classification problems. The aim of using a Decision Tree is to build a training model that can be used to predict the target variable's class or value by learning basic decision rules inferred from the data.

2.1.2.3 K-Nearest Neighbours

KNN regression is a non-parametric approach that approximates the relationship between independent variables and continuous outcomes by combining observations in the same neighborhood in an intuitive manner. The analyst must set the size of the neighborhood, or it can be chosen using cross-validation (which we will see later) to find the size that minimizes the mean-squared error. Although

the approach seems appealing at first, it soon becomes inefficient as the number of independent variables grows.

2.1.2.4 Random Forest

A random forest is a meta estimator that uses averaging to improve predictive accuracy and control over-fitting by fitting a number of classifying decision trees on different sub-samples of the dataset. If `bootstrap=True` (default), the sub-sample size is limited by the `max samples` parameter; otherwise, the entire dataset is used to construct each tree. Each tree in a random forest ensemble is constructed from a sample drawn with replacement from the training set. The aim of these two sources of randomness is to reduce the forest estimator's variance. Individual decision trees do, in reality, have a lot of variation and are prone to overfitting. Forests with inserted randomness produce decision trees with decoupled prediction errors. Some errors can be eliminated by averaging such forecasts. By mixing diverse trees, random forests minimize variation, often at the expense of a small increase in bias. In practice, the variance reduction is often important, resulting in a better overall model.

2.1.2.5 Gradient Boosting Regression

Gradient Boosted Regression extends the concept of boosting to every differentiable loss function. It's an off-the-shelf technique that works for both regression and classification problems in several fields, including Web search ranking and ecology.

2.2 Related Literature Review

There is a variety of work on yield prediction such as rice yield prediction, crop yield prediction in many different countries. But significantly in Bangladesh, there isn't much research work available on paddy yield prediction. In the proposed system [5] S.R. Rajeswari et al. has done a project which aims to show practical and experimental results to improve the crop yield production thus resulting in profitability to the farmers. But there will be numerous difficulties in

executing technological arrangements in agriculture as it is an enormous division. In this paper, [6] P. Priya et al. focus on predicting the yield of the crop based on the existing data by using the Random Forest algorithm. Real data of Tami Nadu were used for building the models and the models were tested with samples. This prediction will help the farmer to predict the yield of the crop before cultivating it in the agriculture field. But there is a chance of improvement in accuracy using a more efficient algorithm. Data mining focuses upon methodologies for extracting useful knowledge from data and after preprocessing of data applied k nearest neighbor algorithm using Data Mining using soil nutrients, fertilizers nutrients, rainfall and temperature [7] by I. Sirur et al. . But this method is not efficient for a larger dataset. A predictive analysis model [8] was designed by I. T. Gubbi et al. that will help the farmers to choose whether the particular crop is suitable for specific rainfall and crop price values. This predictive analysis is a branch of data mining that predicts future probabilities and trends. This approach is to increase the net yield rate of the crop, based on rainfall. A deep neural network (DNN) approach [9] that S. khaki et al. has made to know crop yield prediction. the model was found to have a superior prediction accuracy, with a root-mean-square-error (RMSE) being 12 percent of the average yield and 50 percent of the standard deviation for the validation dataset using predicted weather data. With perfect weather data, the RMSE would be reduced to 11 percent of the average yield and 46 percent of the standard deviation. They also performed feature selection based on the trained DNN model, which successfully decreased the dimension of the input space without a significant drop in the prediction accuracy. A research [10] in which S. Wolfert et al. have presented a survey on a smart farm. The precision Agriculture model is a personalized solution for farmers to analyze and manage variability within fields for profitability. Random forest (RF) was compared to multiple linear regression (MLR) for predicting crop production by J. H. Jeong et al. [11]. They demonstrated that RF outperforms MLR in terms of crop yield prediction. S. Veenadhari et al. [12] developed a software tool called crop advisor as a user-friendly website for forecasting the effect of climatic parameters on crop yield. S. Afrin et al. [13] based on the study of four of Bangladesh's most important crops. They looked at medium high land and high land soil assets in

28 Bangladeshi sub-districts, as well as climate data and crop growth over the last six years. For the study, data mining techniques such as PAM, DBSCAN, K-means, CLARA, and four linear regression methods were used.

2.3 Conclusion

As the things that have been explained above are going to be implemented in our proposed system. As there hasn't been much work done in agriculture in Bangladesh so there is plenty of opportunities for us to improve our designed model.

2.3.1 Implementation Challenges

The dataset is the first and foremost and most important prerequisite for implementing a machine learning algorithm. The more well-equipped the dataset, the more precise the result. If the dataset is right, the model can produce much better results. As a result, one of the most difficult tasks is to ensure that the dataset is correct. If the dataset is inaccurate then the model will show an incorrect result. As well as the model performance will deteriorate as well. After overcoming these difficulties we have been able to develop our proposed model.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we will discuss our proposed methodology and learn about every module of our system. We will find a short mathematical insight into the machine learning algorithm used in our system. Finally, the overall implementation of our system.

3.2 Diagram/Overview of Framework

For developing the system regression models are being used. In this system, producing the output is the production/yield of paddy based on input as rainfall and temperature of different districts of Bangladesh. Figure 3.1 shows the schematic diagram of the proposed system.

3.3 Detailed Explanation

For the betterment of the system and making data more understandable, raw data is being processed first. For processing the dataset, some steps were needed as data preprocessing, feature extraction, and other steps as well. In this chapter, a brief explanation will be given of described steps.

3.3.1 Implementation

For predicting the yield of paddy the most challenging task is collecting the dataset of different districts. We have tried our best to make sure that enough datasets are available so that the designed system works efficiently. In this chapter, we

have given brief information about our project and some snapshots as well. The setup of the whole system is given below:

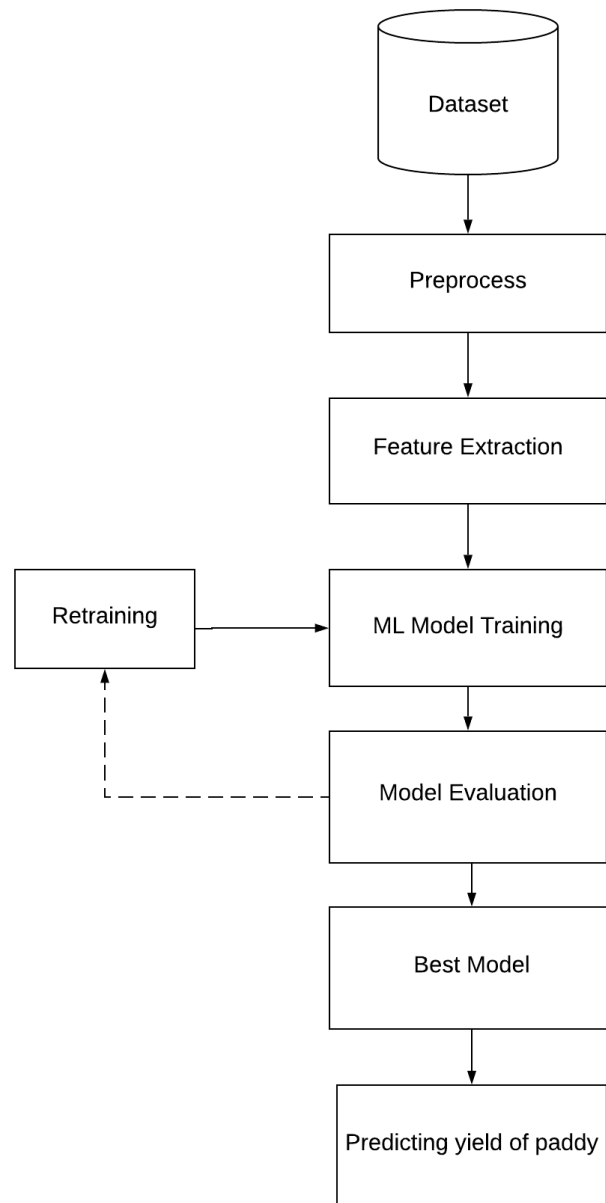


Figure 3.1: Schematic Diagram of the Proposed System.

3.3.1.1 System Requirements

To implement this system some hardware and software tools are needed. Required hardware and software tools are listed below.

3.3.1.2 Hardware Requirements

- Personal Computer

- Intel Core i5 CPU
- Physical Memory 4GB

3.3.1.3 Software Requirements

- Operating System: Microsoft Windows 10
- Software Libraries: Python 3.6.6
- Other: notepad++
- Jupiter Notebook.

3.3.2 Implementation Details

To implement the system, firstly the dataset is collected. Since the datasets are labeled datasets, they are sorted in a folder depending on their label. Then passed through preprocessing steps. Each step is described below.

3.3.2.1 Data Collection

Dataset is the most influential part for building the whole system. We have collected 3 types of a dataset of previous years.

- Production of paddy based on districts
- Rainfall of each station
- Temperature of each station

These datasets are collected from several reliable sources. The data collecting sources are khamarbari, Bangladesh Agricultural Research Council (BARC), and Bangladesh Agriculture University(BAU). The datasets of temperature and rainfall have been collected from the Bangladesh Metrological department.

Table 3.1: Overview of Dataset

Dataset Name	Year	Source
production of paddy	2012-2019	Khamarbari, BAU and BARC
Rainfall	1960-2019	Bangladesh Metrological Department
Temperature	1960-2019	Bangladesh Metrological Department

3.3.2.2 Data Cleaning

To ensure accuracy, we had to go through data cleaning step. It is a vital step for the system. For Data cleaning, data preprocessing is one of the key point.

3.3.2.3 Data Preprocessing

As we have 3 types of datasets. For getting the whole dataset in a frame we have to combine all data into a file. Initial dataset looks like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Zilla/Division	Year	Acres	Hectares	Acres(Maund)	Hectare (M.Ton)	Production (M. Ton)									
1	Bandarban	2012	0	0	0	0	0									
2	Bandarban	2013	235	95	18.42	1.699	162									
3	Bandarban	2014	512	207	19.03	1.755	364									
4	Bandarban	2015	557	225	20.19	1.862	420									
5	Bandarban	2016	560	227	20.3	1.872	424									
6	Bandarban	2017	158600	64182	12.53	1.156	74168									
7	Bandarban	2018	120158	48625	15.89	1.466	71270									
8	Bandarban	2019	0	0	0	0	0									
9	Chattagram	2012	80947	32757	16.68	1.539	50399									
10	Chattagram	2013	78239	31662	16.83	1.558	49326									
11	Chattagram	2014	69636	28180	21.05	1.942	54716									
12	Chattagram	2015	69858	28160	19.99	1.844	51923									
13	Chattagram	2016	68061	27543	19.35	1.785	49159									
14	Chattagram	2017	74850	30290	16.07	1.482	44892									
15	Chattagram	2018	63076	25525	16.56	1.527	38990									
16	Chattagram	2019	0	0	0	0	0									
17	Cox's Bazar	2012	10122	4096	21.76	2.007	8222									
18	Cox's Bazar	2013	11657	4717	20.77	1.916	9038									
19	Cox's Bazar	2014	11288	4566	20.17	1.86	8499									
20	Cox's Bazar	2015	11963	4841	19.51	1.8	8712									
21	Cox's Bazar	2016	12246	4956	19.63	1.811	8973									
22	Cox's Bazar	2017	10823	4380	18.23	1.682	7366									
23	Cox's Bazar	2018	9782	3959	15.78	1.456	5762									
24	Cox's Bazar	2019	0	0	0	0	0									
25	Comilla	2012	20545	8314	18.41	1.698	14118									
26	Comilla	2013	23936	9686	17.03	1.571	15216									
27	Comilla	2014	27924	11300	15.2	1.402	15843									
28	Comilla	2015	27202	11008	16.7	1.54	16957									
29	Comilla	2016	21217	8586	16.85	1.554	13345									
30	Comilla	2017	19435	7565	18.03	1.663	13081									
31	Comilla	2018	14567	5895	15.56	1.435	11687									
32	Comilla	2019	45200	18291	11.98	1.105	20213									
33	Chandpur	2012	10630	4302	19.41	1.79	7702									
34	Chandpur	2013	12580	5091	16.75	1.545	7865									
35	Chandpur	2014	10554	4271	15.8	1.457	6224									

Figure 3.2: Production of paddy

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...																	
A1	Area																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
23	Barisal	1954	17.9	19.9	26.6	28.8	29.2	28.2	27.5	28.1	28.3	27.4	23.2	20.9	25.5		
24	Barisal	1955	18.1	23.3	27.1	30.4	30.7	28.1	28.4	28.4	28.7	26.6	22.2	19.9	26		
25	Barisal	1956	19.5	20.8	26.4	28.9	28.9	27.6	28.1	28	27.9	27.4	23.7	19.8	25.6		
26	Barisal	1957	19.6	20.8	26.3	30.2	30.8	29.8	28.3	28.6	28.8	27.5	23.6	20.6	26.2		
27	Barisal	1958	20.4	22.1	26	30.2	30.4	30.2	28.2	28.3	28.7	28.4	24.9	21	26.6		
28	Barisal	1959	19.4	20.8	26.3	29.5	29.9	28.9	28.3	28	27.4	26.7	22.8	19.4	25.6		
29	Barisal	1960	18.2	22.1	25.8	30.1	30.4	29.5	27.8	28.6	28.1	27.6	23.2	20.3	26		
30	Barisal	1961	19.5	20	26.7	28.7	29.5	28.6	28.8	28.5	28.4	27.5	23.4	17.8	25.6		
31	Barisal	1962	17.7	21.5	28.3	29.4	29.7	28.9	29.1	28.1	28.6	27.2	23.4	19.9	26		
32	Barisal	1963	18.5	23.5	27	29	29	28.9	28.2	28.3	28.9	27.5	23.6	20.1	26		
33	Barisal	1964	37.8	21.6	26	28.5	29.9	29.1	28.2	28.8	29.2	28.2	25.8	22.7	28		
34	Barisal	1965	22	23.5	26.6	29.3	30.4	29.2	28.2	28.3	28.6	28.2	26.1	22.9	26.9		
35	Barisal	1966	22	25.8	27.9	30	30.9	28.6	29.5	28.8	28.7	27.8	26	21.3	27.3		
36	Barisal	1967	20.9	23	24.5	27.5	29	28.6	27.9	27.6	27.4	26.2	21.2	19.5	25.3		
37	Barisal	1968	17.7	20.9	25.6	27.6	29.1	27.1	28	27.8	28.4	26.2	22.9	18.4	25		
38	Barisal	1969	17.3	21.6	26.1	28.3	29.3	28.3	27.8	27.3	28.3	27.1	23.1	18.8	25.3		
39	Barisal	1970	17.6	21.6	26	28.3	29.3	28.2	28	28.1	27.7	26.8	22.7	18.3	25.2		
40	Barisal	1971	18.6	20.8	24.7	25.8	28.9	27.7	27.8	26.7	28.1	27.9	23.8	20.9	25.1		
41	Barisal	1972	18.3	19.7	25.8	27.8	29.4	29	28.6	27.3	28.6	27	23.5	19.9	25.4		
42	Barisal	1973	19.3	22.6	24.8	29.3	27.7	28.6	28.6	27.8	27.5	27.2	23.1	18.8	25.4		
43	Barisal	1974	17.9	20.8	25.5	27.7	28.3	28.3	26.9	28	27.6	27.7	24.7	18.1	25.1		
44	Barisal	1975	17.9	21.8	26	28.5	28.5	28.4	27	27.7	28.3	27.2	22.4	18	25.1		
45	Barisal	1976	18.4	21.8	26.7	28.5	28.1	27.9	27.7	27.4	28.1	27.2	24.9	18.9	25.5		
46	Barisal	1977	17.8	21	26.6	26.7	27.5	27.6	27.9	28	28.4	26.6	24.1	19.2	25.1		
47	Barisal	1978	17	20.8	24.8	27.2	28.1	28.3	27.9	28.3	27.5	27.8	24.4	19.2	25.1		
48	Barisal	1979	19.1	20.4	25.8	28.8	30.3	28.6	28.4	28	28.2	27.6	25.8	20	25.9		
49	Barisal	1980	18.2	21.6	26.2	29.5	28.5	28.4	28.1	28.4	28.6	26.8	23.2	19.8	25.6		
50	Barisal	1981	18.5	21.5	24.5	26.6	28	28.8	27.7	28.7	28.3	27.5	23.3	18.8	25.2		
51	Barisal	1982	19.1	21.5	24.9	27.6	29.9	28.2	28.8	27.8	28.6	27.7	22.5	18.8	25.5		
52	Barisal	1983	18.3	20.8	26.3	28.3	28.4	29	28.6	28.3	28.4	27	23.7	18.2	25.4		
53	Barisal	1984	17.4	19.6	26	28.1	28.8	27.8	27.7	27.8	27.8	27.4	22.3	18.9	25		
54	Barisal	1985	18.6	21	27.1	28.9	28.1	28.7	27.7	28.4	27.8	27	22.6	19.4	25.4		
55	Barisal	1986	17.7	21.1	26.6	28.1	28.2	28.9	27.7	28.8	27.1	26.5	23.6	19.6	25.3		
56	Barisal	1987	18.1	21.5	25.9	28.1	29.3	29.7	27.7	28.1	28.5	27.3	23.9	19.6	25.6		
preprocessed_avg_temp																	

Figure 3.3: Temperature

FileHomeInsertPage LayoutFormulasDataReviewViewTell me what you want to do...

A1																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Area	Year	January	February	March	April	May	June	July	August	September	October	November	December			
2	Barisal	1953	10	3	3	46	339	715	526	166	174	45	0	0			
3	Barisal	1954	0	3	56	14	70	500	260	360	144	169	1	6			
4	Barisal	1956	17	0	20	92	343	784	319	497	700	184	46	0			
5	Barisal	1957	104	13	0	1	22	244	564	286	216	52	0	0			
6	Barisal	1958	0	12	46.222	129	289	171	414	268	321	208	0	0			
7	Barisal	1959	24	5	59	59	258	377	751	621	608	475	0	0			
8	Barisal	1960	0	0	36	13	324	433	716	823	650	316	79	0			
9	Barisal	1961	8	43	51	44	303	553	244	529	206	108	0	0			
10	Barisal	1962	0	44	0	1	134	325	421	229	333	154	262	0			
11	Barisal	1963	0	0	260	54	252	232	367	234	274	385	0	0			
12	Barisal	1964	0	0	0	0	47	225	318	254	141	214	78	0			
13	Barisal	1965	0	44	34	53	76	370	185	346	179	171	1	16			
14	Barisal	1966	19	0	0	1	175	416	328	434	311	317	26	53			
15	Barisal	1967	42	0	92	105	71	159	292	228	547	342	0	0			
16	Barisal	1968	0	14	86	64	246	847	422	331	159	231	12	0			
17	Barisal	1969	0	0	107	82	55	347	487	546	366	45	22	0			
18	Barisal	1970	0	24	5	91	124	408	530	317	571	253	90	0			
19	Barisal	1971	7	1	0	104	267	503	250	233	394	152	114	0			
20	Barisal	1972	0	58	0	121	130	262	248	458	90	29	0	0			
21	Barisal	1973	0	12	81	42	508	502	392	266	398	308	171	197			
22	Barisal	1974	2	0	149	187	265	289	684	238	361	153	18	0			
23	Barisal	1975	0	1	46.222	96.781	230	353	321	351	300.875	150	44	0			
24	Barisal	1976	0	20	2	100	164	330	350	379	305	147	23	25			
25	Barisal	1977	9	36	0	307	157	557	299	138	237	69	53	0			
26	Barisal	1978	0	0	14	242	387	631	409	403	300	127	3	0			
27	Barisal	1979	1	14	30	38	67	453	559	642	316	127	31	47			
28	Barisal	1980	0	55	47	13	235	272	526	297	192	282	0	0			
29	Barisal	1981	24	43	115	412	280	433	521	217	173	71	0	33			
30	Barisal	1982	0	57	17	178	65	465	223	593	219	6	34	11			
31	Barisal	1983	8	39	18	92	318	292	399	575	312	192	48	19			
32	Barisal	1984	16	0	1	102	283	1025	448	431	224	118	0	1			
33	Barisal	1985	1	2	45	64	226	322	291	308	180	198	8	0			
34	Barisal	1986	6	0	7	102	181	254	381	252	526	135	274	1			
	preprocessed_rainfall																

Figure 3.4: Rainfall

For combining the dataset of rainfall and temperature into the same file we have to use the haversine formula because the dataset of rainfall and temperature are based on stations but datasets of production of paddy are based on districts. So, we have assigned the longitude and latitude of each district in the dataset, and using data of longitude and latitude we have been able to assign the nearest stations to the desired districts.

3.3.2.4 Applying Haversine Formula

For applying the haversine formula, we need a dataset consists of all station names and corresponding longitude and latitude. The dataset is given below:

FileHomeInsertPage LayoutFormulasDataReviewViewTell me what you want to do...

A1

Station

	A	B	C	D	E	F	G	H	I	J	K
1	Station	Latitude	Longitude								
2	Barisal	22.701	90.3535								
3	Bhola	22.1785	90.7101								
4	Bogra	24.8481	89.373								
5	Chandpur	23.2321	90.6631								
6	Chittagong	22.3621	91.811								
7	Chittagong	22.2352	91.7914								
8	Chuadanga	23.6161	88.8263								
9	Comilla	23.4607	91.1809								
10	Cox'sBazar	21.4272	92.0058								
11	Dhaka	23.8103	90.4125								
12	Dinajpur	25.6221	88.6438								
13	Faridpur	23.6019	89.8333								
14	Feni	23.0159	91.3976								
15	Hatiya	22.2824	91.0969								
16	Ishurdi	24.129	89.0715								
17	Jessore	23.1778	89.1801								
18	Khepupara	21.9938	90.2292								
19	Khulna	22.8456	89.5403								
20	Kutubdia	21.8167	91.8583								
21	Madaripur	23.1649	90.194								
22	MajdiCoul	22.8694	91.0969								
23	Mongla	22.4942	89.6016								
24	Mymensingh	24.7471	90.4203								
25	Patuakhali	22.3586	90.3317								
26	Rajshahi	24.3745	88.6042								
27	Rangamati	22.7324	92.2985								
28	Rangpur	25.7439	89.2752								
29	Sandwip	22.4919	91.4209								
30	Satkhira	22.3155	89.1115								
31	Sayedpur	25.783	88.8983								
32	Sitakunda	22.6171	91.6809								
33	Srimangal	24.3065	91.7296								
34	Sylhet	24.8949	91.8687								

bd_metro_station

Figure 3.5: Metro Stations

The Haversine formula is described below:

$$dlon = long2 - long1; dlat = lat2 - lat1; a = \sin(dlat/2)^2 + \cos(lat1) * \cos(lat2) * \sin(dlon/2)^2; c = 2 * \text{asin}(\sqrt{a});$$


After applying haversine formula the output looks like this:


	Zilla/Division	Latitude	Longitude	Station
0	Bandarban	22.195327	92.218377	Chittagong(Patenga)
1	Bandarban	22.195327	92.218377	Chittagong(Patenga)
2	Bandarban	22.195327	92.218377	Chittagong(Patenga)
3	Bandarban	22.195327	92.218377	Chittagong(Patenga)
4	Bandarban	22.195327	92.218377	Chittagong(Patenga)
5	Bandarban	22.195327	92.218377	Chittagong(Patenga)
6	Bandarban	22.195327	92.218377	Chittagong(Patenga)
7	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
8	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
9	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
10	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
11	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
12	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
13	Chattagram	22.335109	91.834073	Chittagong(Ambagan)
14	Cox's Bazar	21.564100	92.028200	Cox'sBazar
15	Cox's Bazar	21.564100	92.028200	Cox'sBazar
16	Cox's Bazar	21.564100	92.028200	Cox'sBazar
17	Cox's Bazar	21.564100	92.028200	Cox'sBazar
18	Cox's Bazar	21.564100	92.028200	Cox'sBazar
19	Cox's Bazar	21.564100	92.028200	Cox'sBazar
20	Cox's Bazar	21.564100	92.028200	Cox'sBazar
21	Comilla	23.468275	91.178814	Comilla
22	Comilla	23.468275	91.178814	Comilla
23	Comilla	23.468275	91.178814	Comilla


Figure 3.6: Assigning Nearest Stations

After Assigning the dataset we have got a final dataset. That looks like this:

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

 Cut


 Copy


 Format Painter


Clipboard


Calibri


11


 **B**

 *I*





 U










Font







Alignment

General

\$ % '  

Number

 Conditional Formatting

 Format as Table

Styles

Normal

Bad

Neutral

Calculation

A1 Unnamed: 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Unnamed: Zilla/Divisi	Year	Acres	Hectares	Acres(Mau	Hectare (N	Production	Type	Latitude	Longitude	Station	Rainfall	Temperature			
2	0	Bandarbar	2012	0	0	0	0	Aman	22.19533	92.21838	Chittagong	809	98.8			
3	1	Bandarbar	2013	235	95	18.42	1.699	162	Aman	22.19533	92.21838	Chittagong	542	100.1		
4	2	Bandarbar	2014	512	207	19.03	1.755	364	Aman	22.19533	92.21838	Chittagong	467	102.4		
5	3	Bandarbar	2015	557	225	20.19	1.862	420	Aman	22.19533	92.21838	Chittagong	543	102.8		
6	4	Bandarbar	2016	560	227	20.3	1.872	424	Aman	22.19533	92.21838	Chittagong	661	104.6		
7	5	Bandarbar	2017	158600	64182	12.53	1.156	74168	Aman	22.19533	92.21838	Chittagong	702	104.6		
8	6	Bandarbar	2018	120158	48625	15.89	1.466	71270	Aman	22.19533	92.21838	Chittagong	460	101.5		
9	7	Bandarbar	2019	0	0	0	0	0	Aman	22.19533	92.21838	Chittagong	677	101		
10	8	Chattagar	2012	80947	32757	16.68	1.539	50399	Aman	22.33511	91.83407	Chittagong	669	99.2		
11	9	Chattagar	2013	78239	31662	16.83	1.558	49326	Aman	22.33511	91.83407	Chittagong	708	99.8		
12	10	Chattagar	2014	69636	28180	21.05	1.942	54716	Aman	22.33511	91.83407	Chittagong	463	101.4		
13	11	Chattagar	2015	69858	28160	19.99	1.844	51923	Aman	22.33511	91.83407	Chittagong	752	100.9		
14	12	Chattagar	2016	68061	27543	19.35	1.785	49159	Aman	22.33511	91.83407	Chittagong	639	103.1		
15	13	Chattagar	2017	74850	30290	16.07	1.482	44892	Aman	22.33511	91.83407	Chittagong	832	102.3		
16	14	Chattagar	2018	63076	25525	16.56	1.527	38990	Aman	22.33511	91.83407	Chittagong	672	98.9		
17	15	Chattagar	2019	0	0	0	0	0	Aman	22.33511	91.83407	Chittagong	719	101.3		
18	16	Cox's Baza	2012	10122	4096	21.76	2.007	8222	Aman	21.5641	92.0282	Cox'sBaza	579	101		
19	17	Cox's Baza	2013	11657	4717	20.77	1.916	9038	Aman	21.5641	92.0282	Cox'sBaza	657	101.5		
20	18	Cox's Baza	2014	11288	4566	20.17	1.86	8499	Aman	21.5641	92.0282	Cox'sBaza	305	103		
21	19	Cox's Baza	2015	11963	4841	19.51	1.8	8712	Aman	21.5641	92.0282	Cox'sBaza	892	101.9		
22	20	Cox's Baza	2016	12246	4956	19.63	1.811	8973	Aman	21.5641	92.0282	Cox'sBaza	768	104		
23	21	Cox's Baza	2017	10823	4380	18.23	1.682	7366	Aman	21.5641	92.0282	Cox'sBaza	722	104.1		
24	22	Cox's Baza	2018	9782	3959	15.78	1.456	5762	Aman	21.5641	92.0282	Cox'sBaza	752	100.7		
25	23	Cox's Baza	2019	0	0	0	0	0	Aman	21.5641	92.0282	Cox'sBaza	972	103.4		
26	24	Comilla	2012	20545	8314	18.41	1.698	14118	Aman	23.46827	91.17881	Comilla	398	96.4		
27	25	Comilla	2013	23936	9686	17.03	1.571	15216	Aman	23.46827	91.17881	Comilla	382	97.8		
28	26	Comilla	2014	27924	11300	15.2	1.402	15843	Aman	23.46827	91.17881	Comilla	360	98.3		
29	27	Comilla	2015	27202	11008	16.7	1.54	16957	Aman	23.46827	91.17881	Comilla	432	100		
30	28	Comilla	2016	21217	8586	16.85	1.554	13345	Aman	23.46827	91.17881	Comilla	444	101.2		

updated.csv

Figure 3.7: Preprocessed Dataset

3.3.2.5 Feature Extraction

The improved dataset after pre-processing has a lot of distinctive properties. The feature extraction method extracts the aspect (adjective) from the dataset. There some features that have no influence over the output. That's why we removed those values from the dataset. For recheck, we have created a heatmap using the correlation of the column values.

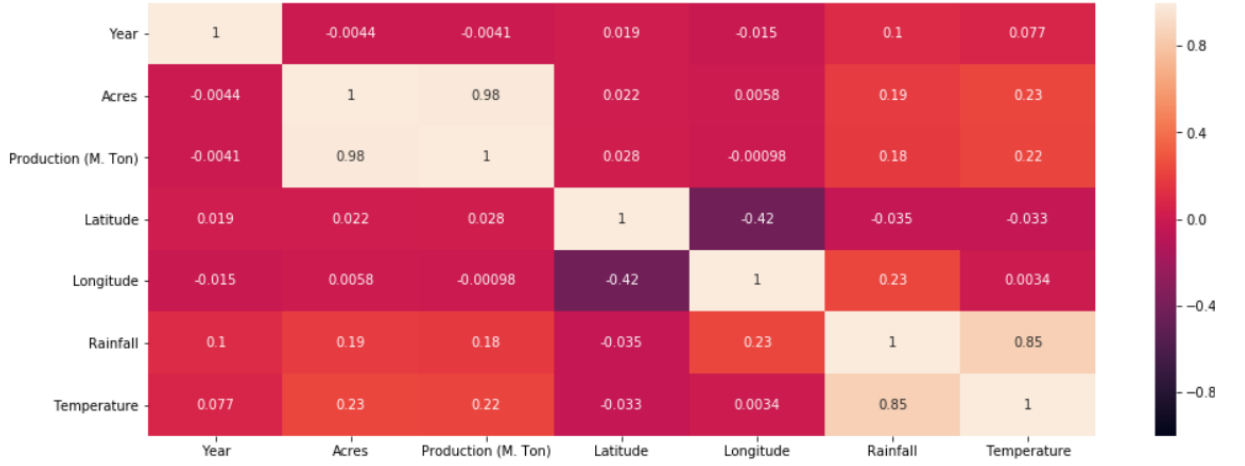


Figure 3.8: Heatmap

From the heatmap, we can see that two attributes which are longitude and latitude have much less correlation and its value are also negative. That means it has no influence over the output. So, that's why we removed them as well.

3.3.2.6 Regression Algorithm

For training the dataset, we used 5 regression algorithms to build a model to test the dataset because using one algorithm doesn't justify the outcome. The algorithms are Linear Regression, KNN Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. All 4 algorithms have an accuracy of over 90 percent. But among them, the Gradient Boosting Regression algorithm got the best accuracy which is 94.57 percent but its mean absolute and mean squared error rate is high. But on the other hand, Random Forest regression has an accuracy of 94.38, and it's mean absolute and mean squared error rate is less. So Random Forest regression is much more suitable than others.

3.3.2.7 Temperature and Rainfall Prediction

When the user selects a venue, the proposed model extracts the temperature and rainfall for that location from datasets. Temperatures and precipitation differ month to month. Furthermore, the data is a one-dimensional time sequence. For future forecasting, the Autoregressive Integrated Moving Average (ARIMA)

forecasting model can be used. As both temperature and rainfall are seasonal in nature, we have used SARIMA which can capture the seasonal pattern more precisely. SARIMA(p, d, q) can be used to specify the SARIMA model [14] (P, D, Q, s). Where p, d, and q stand for autoregressive, differencing, and moving average words, respectively. Seasonal autoregressive terms, seasonal order of differencing, seasonal moving average terms, and seasonal duration are all denoted by P, D, Q,s.

3.4 Conclusion

In this chapter, we have given a brief description of our implementation. During the implementation, we had to face some difficulties and were able to overcome all of those. After overcoming all those difficulties we have been able to get accuracy over 90 percent on all algorithms. Random Forest algorithm has very good accuracy and less error rate.

Chapter 4

Results and Discussions

4.1 Introduction

In this chapter, we will give an overview of the dataset that is used to train the model as well as the impact i.e. social, environmental, and ethical impact of the thesis and evaluation of performance and framework. We will also discuss the experimental analysis, result, and evaluation.

4.2 Dataset Description

For implementing any machine learning algorithm, the first and foremost priority is the dataset. The bigger the dataset the more accurate the build model will become. Starting the project dataset was the first requirement but unfortunately, we did not have any dataset available so that we can build a model to predict the outcome. To have a well-equipped dataset is very much of a challenge for us. building a dataset that contains all the necessary contents that were very tough because we have to go to some places to collect the data. We started collecting the datasets in August 2020 and ended in January 2021. As we need a lot of information so collecting and building the dataset required a lot of time. The data collecting procedure was totally manual. First of all, we labeled our datasets by assigning numerical values to each district. After that, we normalized our datasets. We also need to create the dataset of metro stations for assigning nearest to each district. The description of the datasets are given below:

Table 4.1: Dataset Summary

Dataset Name	Year	Attribute	Quantity
production of paddy	2012-2019	8	3267
Rainfall	1960-2019	3	1879
Temperature	1960-2019	3	1879

Using this dataset and the key attributes we have been able to build a model that can predict the production of paddy.

4.3 Impact Analysis

The impact of the prediction model of paddy yield can have a huge impact on agricultural life which can be socially and environmentally influential.

4.3.1 Social and Environmental Impact

With the help of this predictive model the persons who are associated with the agricultural activity, they can be very much benefited. They will be able to update with weather and that will make them cautious about their products and will be able to take necessary steps to prevent their expected loss as well.

4.3.2 Ethical Impact

This predictive model does have some ethical sides as well. Ethical impact means that how good or bad is it for the people who are connected to this project. By implementing this system our farmers can be very benefited and our agricultural system will be developed as well.

4.4 Evaluation of Framework

in this section, we are going to discuss the regression algorithm and its accuracy and how the algorithms have worked on trained values and tested values by showing the graph. The first algorithm that is being used is linear regression.

Its accuracy is 93.502 but it has a more mean absolute error and mean squared error. The plotting of the actual value and predicted value of this algorithm is given below:

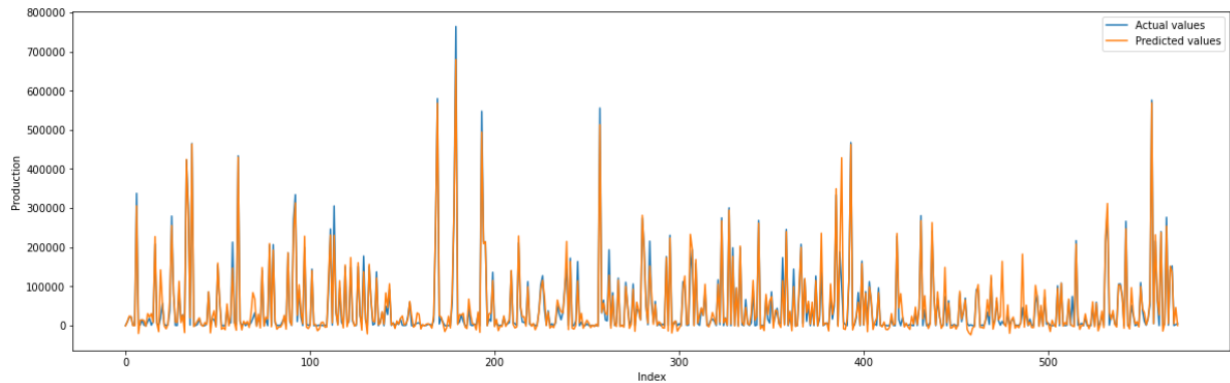


Figure 4.1: Linear Regression

The second algorithm that is being used is the decision tree. Its accuracy is 94.046 and it has less mean absolute error as well. The plotting of the actual value and predicted value of this algorithm is given below:

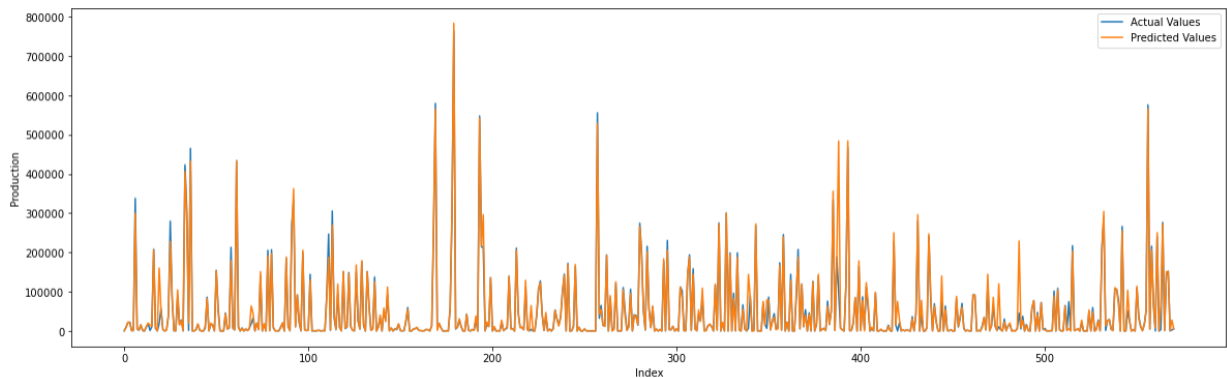


Figure 4.2: Decision Tree Regression

Third algorithm that is being used is KNN regression. It's accuracy is 94.046 percent. The plotting of the actual value and predicted value of this algorithm is given below:

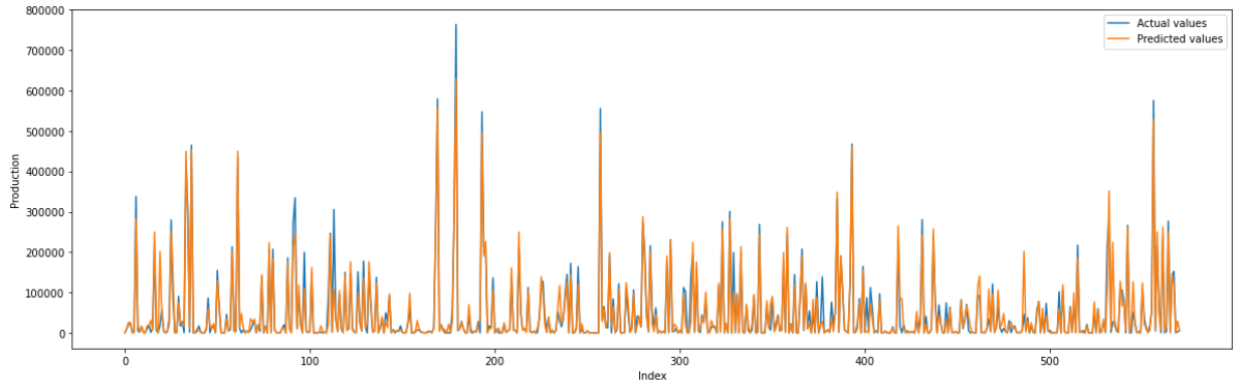


Figure 4.3: KNN Regression

The fourth algorithm that is being used is Random forest regression. Its accuracy is 94.38 which is better than the decision tree and its error rate is less than the decision tree as well. The plotting of the actual value and predicted value of this algorithm is given below:

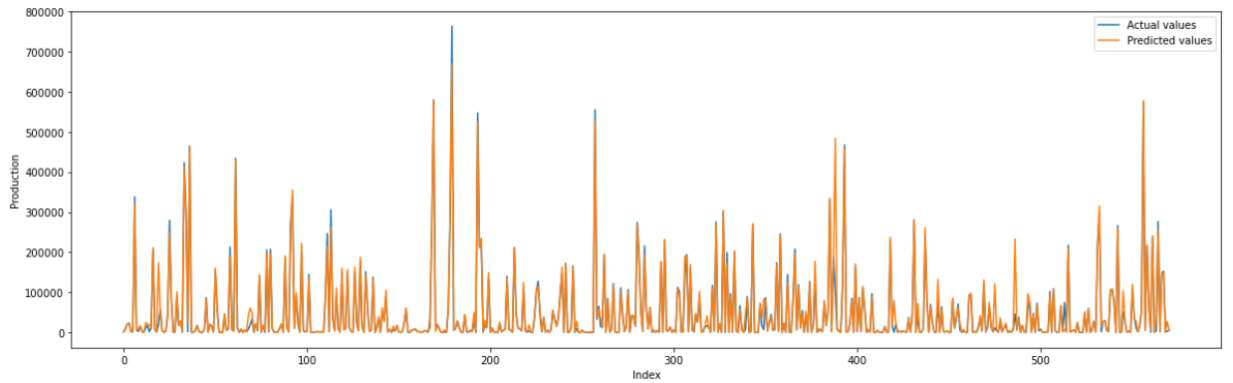


Figure 4.4: Random Forest Regression

The fifth algorithm that is being used is Gradient Boosting Regression. Its accuracy is 94.57 which is better than the decision tree but its error rate is higher than the decision tree. The plotting of the actual value and predicted value of this algorithm is given below:

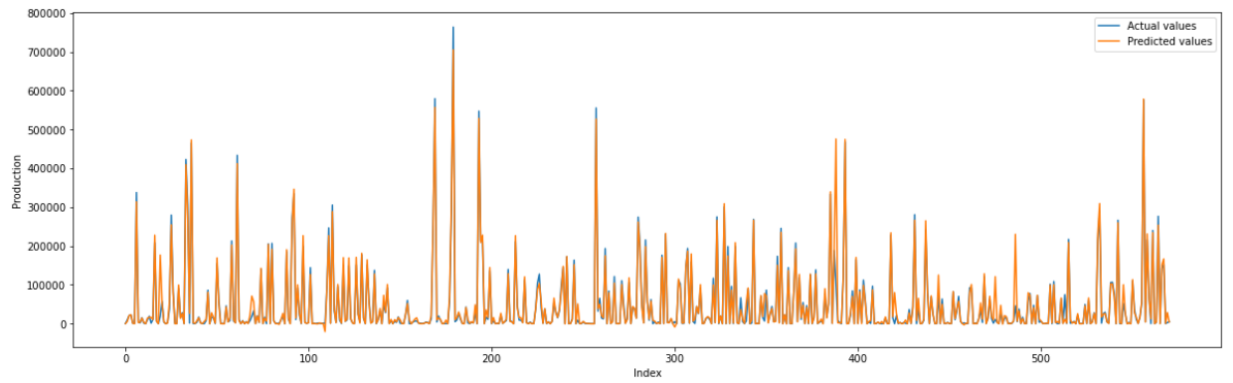


Figure 4.5: Gradient Boosting Regression

After building the model we next move to predict the forecast of the weather. For that, we have used the SARIMA model to generate a temperature and rainfall of a particular district.

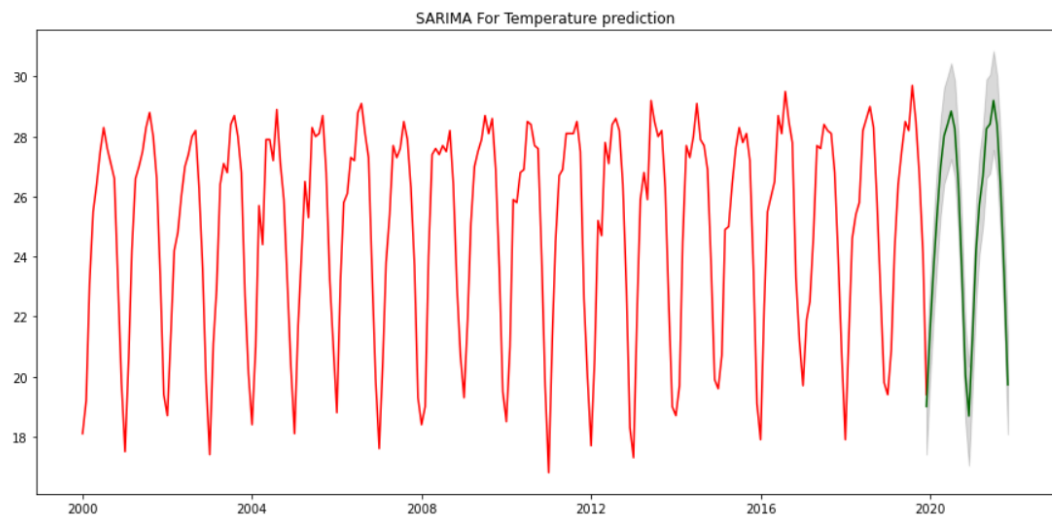


Figure 4.6: Overview of Temperature

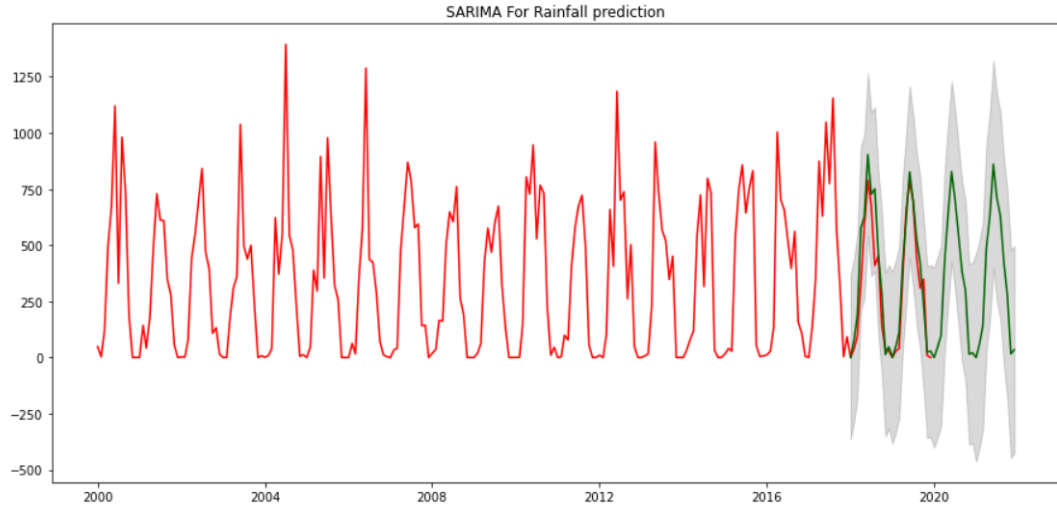


Figure 4.7: Overview of Rainfall

The predicted temperature and rainfall is given in the following table.

Table 4.2: Predicted Temperature and Rainfall

Month	Temperature	Rainfall
Jan	21.516	78.968
Feb	23.640	186.381
Mar	25.360	576.063
Apr	26.999	626.020
May	28.028	902.772
Jun	28.419	726.270
Jul	28.844	750.222
Aug	28.267	458.031
Sep	26.432	285.817
Oct	23.538	13.380
Nov	20.070	47.362
Dec	18.688	0

4.5 Evaluation of Performance

For implementing our system we have used a regression model. So, we need to know how to build a regressor and measure the quality of the regressor as well.

In this context, an error is defined as the difference between the actual value and the value that is predicted by the regressor. For evaluating we have used some matrices. They are :

- Mean absolute error
- Mean squared error
- Median absolute error
- Explained variance score
- R2 score

Mean absolute error: The absolute difference between the real or true values and the expected values is known as error. The sum of the absolute errors of all the data points in a dataset is called the mean absolute error. If the result has a negative symbol, it is ignored in absolute difference.

$$MAE = Truevalue - Actualvalue \quad (4.1)$$

MAE takes the average of this error from every sample in a dataset and gives the output.

Mean squared error: The sum of the squares of the errors of all the data points in a dataset is called the mean squared error. It is one of the most widely used metrics.

$$MSE = \frac{1}{N} \sum_{i=1}^n (actualvalues - predictedvalues)^2 \quad (4.2)$$

Median absolute error: The median absolute error is the average of all errors in a dataset. The key benefit of this metric is that it is unaffected by outliers. In contrast to a mean error metric, a single bad point in the test dataset does not distort the entire error metric.

Explained variance score: The difference between a model and real data is measured using an explained variance. To put it another way, that's the portion of the overall variance in the model that is clarified by variables that are present

rather than error variance. This score measures how well our model can calculate the value of the variation in our dataset. Our model is perfect if it receives a score of 1.0.

R2 score: The coefficient of determination is also known as the R2 score. This metric indicates how well a model matches a specific dataset. It shows how closely the regression line corresponds to the real data values. The R squared value ranges from 0 to 1, with 0 indicating that the model does not match the data and 1 indicating that the model matches the dataset perfectly.

4.5.1 Experimental Results

Here, we will give the results that we got from using our trained model. We have build models using four algorithms. The performance of all models has been very much satisfying.

Table 4.3: Comparison of Algorithm

Algorithm	Mean absolute error	Root Mean squared error	Median absolute error	R2 score
Linear Regression	11910.893	24053.590	6326.0	0.935025
Decision Tree	6068.9765	23024.962	616.0	0.94046
KNN	11406.685	25908.714	2239.8	0.92461
Random Forest	5880.020	22357.754	771.01	0.94386
Gradient Boosting Regression	6829.375	22076.62	1508.481	0.94578

Here we have produced the desired outcome using the Random Forest algorithm. Because it has better accuracy and its mean absolute and mean squared error is less as well. The outcome has been given in the table below:

Table 4.4: Predicted Result

District	Year	Acre	Type	Rainfall	Temperature	Production
Sylhet	2020	205803	Aman	346.560	88.729	708262
Sylhet	2021	170951	Aman	287.462	69.381	603717
Sylhet	2020	168398	Aman HYV	3809.878	210.599	601596
Sylhet	2021	175620	Aman HYV	3373	69.381	614361
Sylhet	2020	1544	Aus	3005.286	112.291	0
Sylhet	2021	16158	Aus	2665.9426	112.5525	21539
Sylhet	2020	70908	Aus HYV	3005.2865	112.29136	245116
Sylhet	2021	74876	Aus HYV	2665.9426	112.5525	267458
Sylhet	2020	43895	Boro	1467.43397	97.515	169449
Sylhet	2021	27185	Boro	1188.5001	75.999	82975
Sylhet	2020	7937	Boro Hybrid	1388.4659	97.51501	20799
Sylhet	2021	4851	Boro Hybrid	1138.8611	75.99923	21898

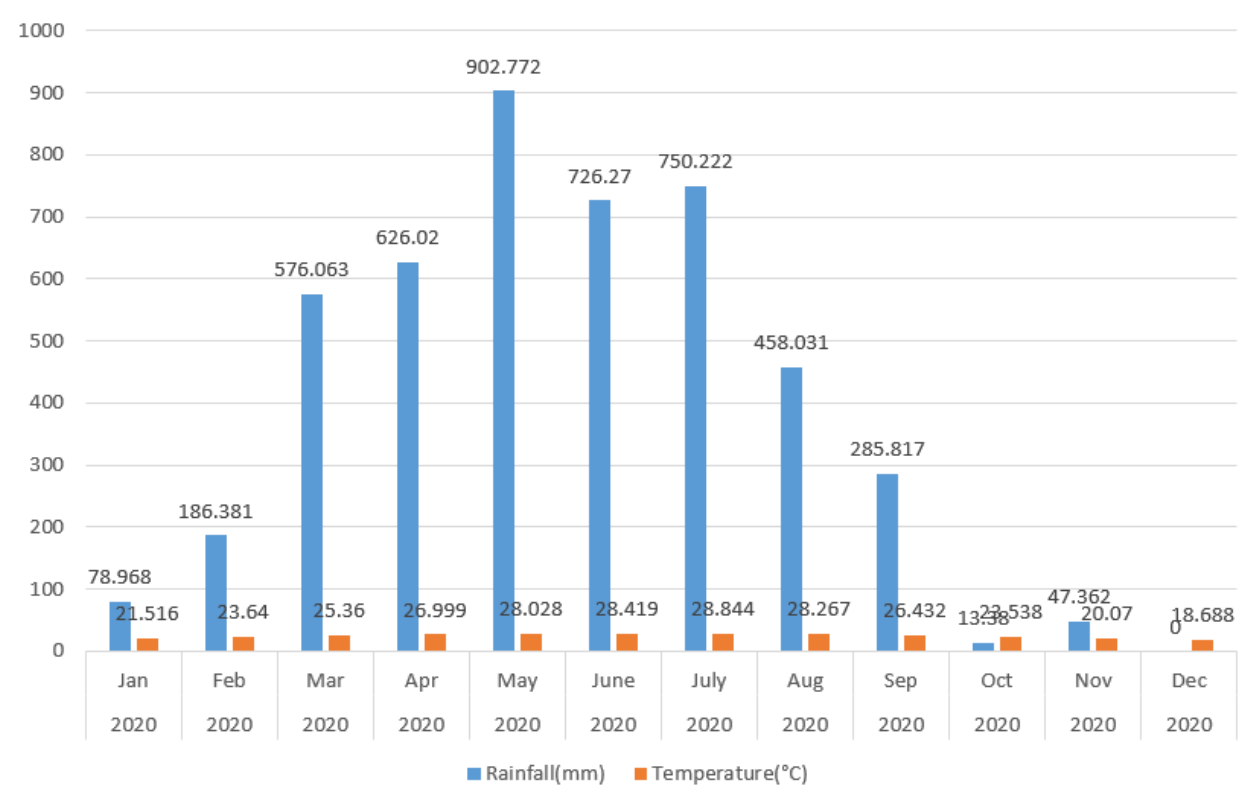


Figure 4.8: Weather prediction

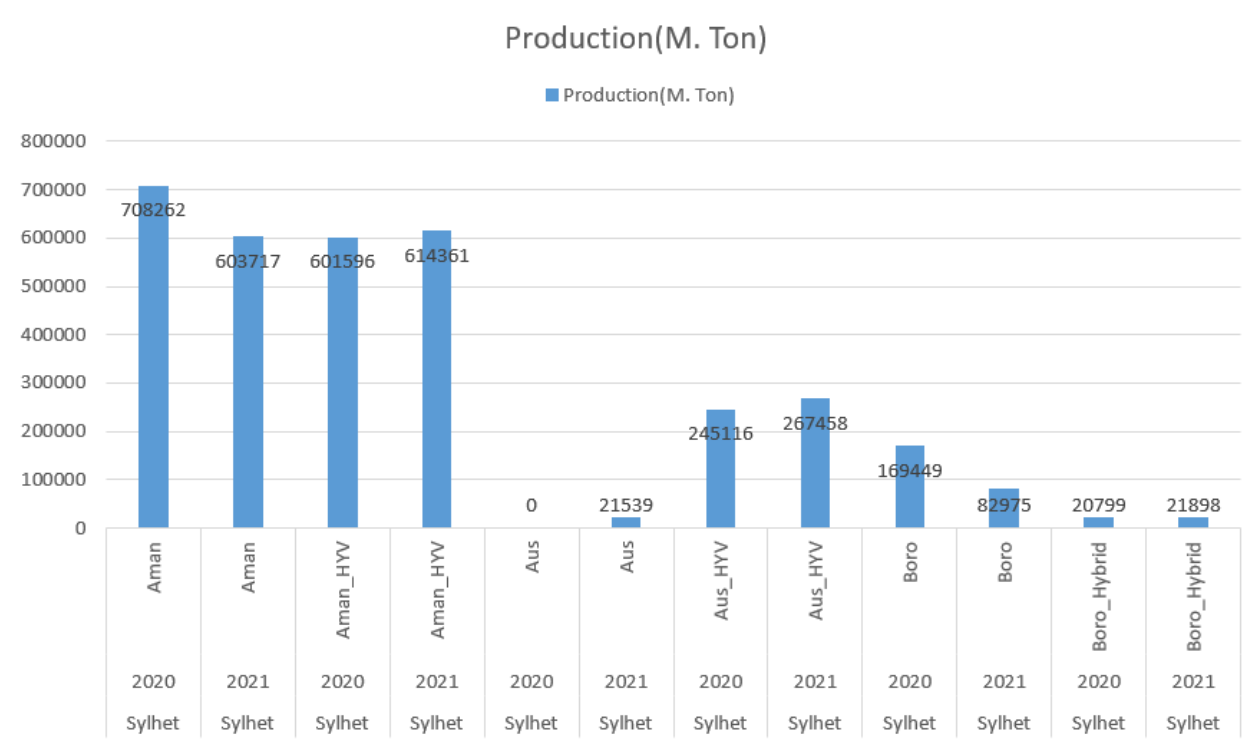


Figure 4.9: Production

4.6 Conclusion

In this chapter, we learned about the dataset used to train the model, as well as the thesis' effect (social, environmental, and ethical), as well as the performance and system evaluation. We also spoke about the experimental results, interpretation, and assessment. After evaluating the whole system we have got the best model which was built by using Random Forest regression. Its accuracy is 94.38 and contains less error rate as well.

Chapter 5

Conclusion

5.1 Conclusion

In this work, we have implemented a model that would predict the yield of paddy for upcoming years. As in the revolutionized world, everything is being developed by scientific technology. But the Agricultural field is one of the fields that has not been able to develop as much we expected to happen. That's why we have done this work so that our farmers are being able to be updated. If they are become much known about the production then it will help not only country as well as our people in the food sector as well. As the yield rate is dependent on the weather. So we have used the SARIMA model to predict the weather forecast so that they can be able to be aware of the temperature and rainfall. As our population is growing every day that's why more foods are being needed. One of the most required food is rice which comes from paddy. This model has been developed so that our production of paddy can fill the requirements of our countrymen as well to make our agricultural field much more scientific and advanced.

5.2 Future Work

In our proposed system, we have developed a model which would predict the production of paddy and also predicted the weather forecast. But there is always some scope of work that can be done in the future. Some of them are :

- The model has been developed by using machine learning. So the outcome of the system will not be 100 percent perfect. There is always a chance to get better results and accuracy.

- This model can be useful for any web-based application or android application. Turning this model into an application could be better work for the future.

References

- [1] M. K. Hasan, M. R. Ali, C. K. Saha, M. M. Alam and M. E. Haque, ‘Combine Harvester: Impact on paddy production in Bangladesh,’ *J. Bangladesh Agric. Univ.*, vol. 17, no. 4, pp. 583–591, 2019, ISSN: 1810-3030. DOI: 10.3329/jbau.v17i4.44629 (cit. on p. 2).
- [2] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, ‘Crop yield prediction using machine learning algorithms,’ in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, IEEE, 2019, pp. 125–130 (cit. on p. 2).
- [3] H. Salpekar *et al.*, ‘Design and implementation of mobile application for crop yield prediction using machine learning,’ in *2019 Global Conference for Advancement in Technology (GCAT)*, IEEE, 2019, pp. 1–6 (cit. on p. 6).
- [4] A. Shah, A. Dubey, V. Hemnani, D. Gala and D. Kalbande, ‘Smart farming system: Crop yield prediction using regression techniques,’ in *Proceedings of International Conference on Wireless Communication*, Springer, 2018, pp. 49–56 (cit. on p. 6).
- [5] S. R. Rajeswari, P. Khunteta, S. Kumar, A. Raj Singh and V. Pandey, ‘Smart farming prediction using machine learning,’ *Int. J. Innov. Technol. Explor. Eng.*, vol. 6, no. 4, pp. 190–194, 2019, ISSN: 22783075 (cit. on p. 8).
- [6] P. Priya, U. Muthaiah and M. Balamurugan, ‘Predicting yield of the crop using machine learning algorithm,’ *International Journal of Engineering Sciences & Research Technology*, vol. 7, no. 1, pp. 1–7, 2018 (cit. on p. 9).
- [7] I. Sirur, K. B, S. P, M. K. M and R. Anjum, ‘Paddy Yield Prediction Model Using Data Mining Techniques,’ no. Ncicnda, pp. 474–477, 2018. DOI: 10.21467/proceedings.1.71 (cit. on p. 9).
- [8] O. C. By I T Gubbi, B. T. R, B. K. S, A. K. T, A. Professor and U. Students, ‘National Conference on Technology for Rural Development (NCTFRD-18) Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning,’ pp. 61–65, 2018. DOI: 10.18231/2454-9150.2018.0805 (cit. on p. 9).
- [9] S. Khaki and L. Wang, ‘Crop yield prediction using deep neural networks,’ *Front. Plant Sci.*, vol. 10, no. May, pp. 1–10, 2019, ISSN: 1664462X. DOI: 10.3389/fpls.2019.00621. arXiv: 1902.02860 (cit. on p. 9).
- [10] S. Wolfert, L. Ge, C. Verdouw and M.-J. Bogaardt, ‘Big data in smart farming—a review,’ *Agricultural systems*, vol. 153, pp. 69–80, 2017 (cit. on p. 9).

- [11] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy *et al.*, ‘Random forests for global and regional crop yield predictions,’ *PLoS One*, vol. 11, no. 6, e0156571, 2016 (cit. on p. 9).
- [12] S. Veenadhari, B. Misra and C. Singh, ‘Machine learning approach for forecasting crop yield based on climatic parameters,’ in *2014 International Conference on Computer Communication and Informatics*, IEEE, 2014, pp. 1–5 (cit. on p. 9).
- [13] S. Afrin, A. T. Khan, M. Mahia, R. Ahsan, M. R. Mishal, W. Ahmed and R. M. Rahman, ‘Analysis of soil properties and climatic data to predict crop yields and cluster different agricultural regions of bangladesh,’ in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, IEEE, 2018, pp. 80–85 (cit. on p. 9).
- [14] S. Wang, J. Feng and G. Liu, ‘Application of seasonal time series model in the precipitation forecast,’ *Mathematical and Computer modelling*, vol. 58, no. 3-4, pp. 677–683, 2013 (cit. on p. 21).