Crop Yield Prediction
Using Machine Learning And Deep learning

by

Sarna Saha
22141051
Md. Asiful Islam
17201077
Nishat Anjum
18101431
Mahmudul Hasan Mitul
18101066

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
January 2023

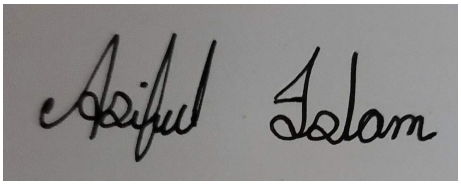
Declaration

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.

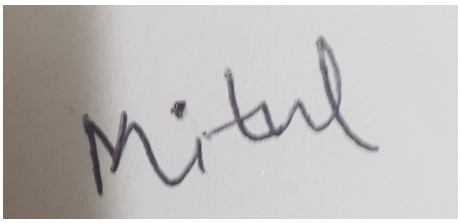
Student's Full Name & Signature:

Sarna Saha
22141051



Md. Asiful Islam
17201077

Nishat Anjum
18101431



Mahmudul Hasan Mitul
18101066

Approval

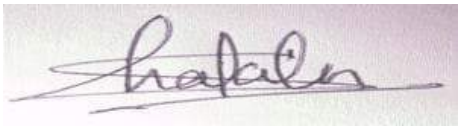
The thesis/project titled "Crop Yield Prediction Using Machine Learning And Deep Learning" submitted by

1. Sarna Saha (22141051)
2. Md. Asiful Islam (17201077)
3. Nishat Anjum (18101431)
4. Mahmudul Hasan Mitul (18101066)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 17, 2023.

Examining Committee:

Supervisor:
(Member)



Shakila Zaman
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Arif Shakil
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

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

Abstract

Bangladesh is an agrarian country. Though, a substantial portion of our economy and workforce depends directly or indirectly on agriculture. However, due to climate change, floods, insufficient incentives, and less grist our farmers are getting demotivated in farming. As a result, more and more farmers are leaving the agriculture sector every year and this can cause devastating effects for Bangladesh. Moreover, there is little or no research on improving Bangladesh agriculture using cutting-edge machine learning techniques. So, this research works on Crop yield prediction Using Machine learning and deep learning. This work explores the different state-of-the-art machine learning and deep learning techniques and relevant dataset to develop an effective Crop yield prediction system for Bangladeshi farmers. So that our farmers can decide which crop to cultivate for gaining the maximum yield by following our prediction system.

Keywords: Crop Yield Prediction; Artificial Intelligence; Machine Learning; Deep Learning; Neural Network; Classification Models; Recurrent Neural Network

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 Shakila Zaman mam, and Co-Supervisor Arif Shakil sir, for their absolute and incomparable guidance and support throughout our journey. Thirdly, we thank our special collaborator Md. Hasibur Rahman Niloy from Bangladesh Agricultural University for his special contribution to our Dataset. And finally, to our parents who have always motivated us and inspired us to work. With their prayers, support, inspiration, and investment, we are almost at the end of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

API Application Programming Interface

CNN Convolutional Neural Network

FN False Negative

FP False Positive

GRU Gated Recurrent Unit

KNN k-nearest neighbors algorithm

LDA Linear Discriminant Analysis

$LSTM$ Long short-term memory

$MM - MTL$ Multi Modal Meta Multi Task

NLP Natural Language Processing

RNN Recurrent Neural Network

SMO Sequential Minimal Optimization

SVM Support Vector Machine

TN True Negative

TP True Positive

Chapter 1

Introduction

1.1 Introduction

Bangladesh has been an agricultural country since its dawn. However, with the rise of industrialization, agriculture is on the decline. But still, agriculture accounts for 42.7% of the country's employment and 14-2% of the GDP [1] which is one of the highest.

Moreover, in the past decade, Bangladesh has become self-sufficient in most of the essential crops. It was possible due to extensive research and the green revolution. As a result, the production of the essential crops has increased 5-10 fold [2].

However, one thing that has not gone up is the condition of the farmers. It is still the same as in the 90s. Their land has not increased their wealth. As a result, more and more farmers are shifting from farming to other sectors. According to a survey by the daily star, the number of farmers has fallen from around 73 percent of the population in 1983/84 to nearly 42.7 percent by 2022 [3]. As a developing nation, it is an alarming signal. Because if we lose self-sufficiency in the crop sector, our food import cost will increase. Moreover, it might as well cause famines.

The main reason behind farmers leaving agriculture is not getting enough prices for their crops. For example, rice takes 3-4 months to grow and 1 kg of rice will require 3,000 liters of water to grow, not to mention the price of fertilizer and insecticides. Furthermore, growing rice is a labor-intensive task too. But farmers only get 36 tk for one Kg of rice [4]. That might be ok for large farms which account for only 2 percent of the farms in Bangladesh. On the other hand, 98 percent of the farmers have limited farmland and labor. So, with lesser pay, it's not possible for them to keep up with the price hikes. Moreover, there are middlemen involved in the trading of crops. The middleman takes a large chunk of the profit of the farmers. Lastly, there is a labor shortage in rural areas. So, more and more farmers are leaving their forefather's professions. Moreover, Bangladesh is one of the top producers of rice, potato, jute, and fish [5], but still, it has very little farm productivity.

To solve the issue, farming productivity must be increased so that farmers do not have to leave their profession and can get more pay from their limited farmland with less labor. From our research, we have found, precision agriculture can solve the

productivity issue and increase the profit of our farmers.

In order to support management decisions based on estimated variability for improved resource use efficiency, productivity, quality, and profitability, precision agriculture is a management strategy that collects, processes, and analyzes temporal, spatial, and individual data with the help of different state of the art machine learning models. It then combines this data with other information and finds out which crop is better for production in a certain area. However, agriculture is different in every country and continent. As a result, not all precision agriculture techniques work for our country. Furthermore, it is crucial that the predictions made by precision agriculture techniques have to be precise and error-free. Otherwise, it might result in heavy capital loss and labor waste.

Various studies are already done on precision agriculture to achieve a precise and effective system for crop yield recommendation. One such method is developing a custom dataset and testing it on different machine learning and deep learning models. This study develops a system that uses a custom dataset tailored to Bangladeshi crop data and uses this dataset to train and test state-of-the-art machine learning classifiers and custom deep learning models and compare the results.

1.2 Problem Statement

Crop yield prediction is a multiclass classification problem. As our deep learning and machine learning models have to predict from multiple crops based on the soil and weather data. Solving this issue by using deep learning and machine learning approaches with Bangladeshi crop data is one of the toughest tasks.

As the extreme effect of climate change, the temperature, humidity, and rainfall data drastically change every season. As a result, it is hard to collect and select relevant data which can be used for better prediction outcomes.

On the other hand, the crop yield processes and strategies might vary with time as these strategies are nonlinear in nature. Moreover, they might be intricate due to the integration of correlated factors. They can also be impacted by non-arbitrate runs and external aspects.

Furthermore, for prediction models, the novelty of data is quite important. However, finding accurate data from the Bangladesh Agricultural University is extremely difficult. As the data is taken on pen and paper. So, there is a high chance that there are so many errors in those data. Again, we need to transform the analog data into a digital format to train our models which is an extremely difficult task as the handwriting was not clear enough.

Another issue for us is the lack of knowledge about crops and the cultivation process. As all of us are from CSE background. So, we have no idea about how the cultivation system of our country works and without proper knowledge, we would not be able to develop and train our models. So, we had to study different crops, their cultivation process, and the optimal environment for them in order to develop our model.

Lastly, the Bangladeshi agriculture process is quite different from other countries and there is little or no previous study on crop yield prediction in Bangladesh agriculture. As a result, we are going through a lot of trial and error phases for developing the most suitable crop yield prediction system for Bangladesh.

1.3 Research Objectives

The principal objective of this study is to develop and compare multiple cutting-edge machine learning classifiers and deep learning models to find out the best solution for Crop yield prediction. So, the aims of this research are

1. To profoundly understand the proposed Machine learning classifiers and deep learning models.
2. Train our proposed state-of-the-art models with large-scale environment data containing nitrogen, phosphorus, and potassium content in the soil, temperature, humidity, rainfall, and pH value.
3. Test and compare the classifiers and models to figure out the best-performing model for crop yield prediction.
4. Evaluate the proposed model using different evaluation matrixes.

Lastly, the proposed system would not only bring solid outcomes in the training dataset but also in unseen data.

Chapter 2

Literature Review

2.1 Related works

Babu et al. [6] explain the specifications and preparation required to design a system for precision farming. It studies the foundations of precision farming in detail. In order to exert some control over unpredictability, this research proposes a system that applies Precision Agriculture techniques to small, open farms at the level of the solo farmer and crop. The model's overall goal is to use the most accessible technology, such as SMS and email, to provide consultancy services to even the tiniest farmer at the level of the smallest plot of crops. This model was created to account for the situation in Kerala State where the typical farm size is smaller than in any other state. As a result, this model can be applied elsewhere in the world with some minor upgrades.

Khedr et. al [7] aims to find a solution to Egypt's food security issue. It suggests a structure that would forecast production and import for that specific year. WEKA builds the prediction using Artificial Neural Networks and Multilayer Perceptron. As a result, the researchers would be able to monitor the quantity of cultivation, import, requirement, and availability at the conclusion of the procedure. Therefore, it would be easier to decide whether or not food needs to be imported further.

Ahamed et al. [8] outlines multiple classification techniques for the data set on liver disease. Because accuracy depends on the dataset and the learning process, the study underlines the importance for accuracy. These disorders were categorized using classifiers such VFI, ZeroR, ANN, and Naive Bayes. Then the efficacy and error rates were compared. The models' performance was evaluated in terms of precision and computational efficiency. All classifiers, with the exception of naïve bayes, had improved predictive performance, it was determined. The proposed algorithms with the highest accuracy are Multilayer Perceptron.

Yang et al. [9] attempt to find a solution to the critical ensemble learning problem of classifier selection. There has been a method developed for choosing the best models from a pool of classifiers. Their method seeks to increase performance and accuracy. Based on categorization precision and accuracy, the SAD approach was suggested. The relationship between the most accurate and relevant classifiers is discovered using Q statistics. The ensemble was created by combining the classi-

fiers that weren't selected. The goal of this measure is to ensure that the ensemble performs better and is more diverse. There were other ways found, including the No selection algorithm, the Selection by Accuracy, Selection by accuracy, and Diversity algorithms. Finally, it is implied that SAD functions more effectively than others.

Kumar et al. [10] demonstrates the significance of crop choice, and elements influencing crop choice are examined, such as production rate, market price, and governmental policy. This study suggests a crop selection method (CSM), which fixes the crop selection issue and raises the crop's net yield rate. It proposes that a season's worth of crops be chosen while taking the weather, crop type, soil type, and water density into account. The accuracy of CSM depends on the expected value of key factors. Consequently, a prediction approach with increased performance and accuracy must be included.

Savla et al. [11] explored assortment algorithms in detail, including how well they function in predicting yield in precision husbandry. These algorithms are used to predict the production of a soybean crop using data that has been gathered over a number of years. In this study, Random Forest, SVM, Bayes, Neural Network, Bagging, and REPTree were employed as the yield prediction techniques. The result reached, in the end, is that, of the algorithms mentioned above, bagging has the lowest error deviation with a mean absolute error of 18985, making it the best approach for yield prediction.

The soil datasets in the study of Paul et al. [12] are examined, and a projected categorization is made. It is determined that the crop yield is a classification rule from the expected soil category. For predicting agricultural yield, naive Bayes and KNN algorithms are employed. The indicated future study involves developing effective models utilizing other classification methods, such as principal component analysis and support vector machines.

In their study, Dahikar et al. [13] explored different feed-forward back propagation artificial neural networks for crop yield prediction. They created their own dataset. Their dataset contained features like phosphate, PH, potassium, nitrogen, organic carbon, calcium, magnesium, sulfur, manganese, copper, iron, depth, of soil along with the type of soil and climate contains like humidity, temperature, and rainfall. After that, they trained their custom ANN model with the dataset features and ANN with zero, one, and two hidden layers and found satisfactory results.

In their work, Panda et al.[14] used NDVI, GVI, SAVI, and PVI measurements from aerial images to develop models for predicting corn crop yield before harvest. The authors used data mining techniques, and satellite and drone images to create the dataset. Then they transformed the image data into numbers using different encoding methods. Lastly, they fed their data into different BPNN crop yield prediction models. From their research on corn, they have bought extraordinary results and hope to apply this method to other crops as well.

Chapter 3

Dataset Description

3.1 Data Collection

The data collection process was one of the toughest for us till now. Because for our prediction system to be precise and efficient we would need accurate data. Moreover, we need specific soil attributes to train our models which is hard to get because the BAU authority keeps the soil data in analog form. But we need the data in digital CSV format. So, our data collection process was slow and we could not manage a substantial amount of data.

Moreover, as the weather in Bangladesh is not diverse and farmers tend to follow other farmers to grow a certain type of crop in one district only. For example, Rajshahi and Chapainawabganj have similar weather and most mangos are produced there. As a result, the data points collected on a single crop are in close proximity.

However, general crop research data were collected from Bangladesh Agriculture University. After collecting the data on different crops we labeled each data point.

3.2 Dataset Analysis

The fruits and crops examined in our models are Banana, Watermelon, Mango, Rice, Pigeon Peas, Moth Beans, Black gram, Lentil, Pomegranate, Muskmelon, Papaya, Coconut, Maize, Mung bean, Cotton, and Jute. The number of data points for each crop available in the dataset is shown in the below table.

Crop and Fruit	Data Points
Banana	280
Watermelon	280
Mango	260
Rice	240
Pigeon Peas	240
Moth Beans	240
Blackgram	240
Lentil	240
Pomegranate	240
Muskmelon	240
Papaya	240
Coconut	240
Maize	220
Mungbean	220
Cotton	220
Jute	220
Total	3860

Table 3.1: Data Points on different crops of Bangladesh

Below here there is a small sample of our Dataset

	N	P	K	temperatu	humidity	ph	rainfall	label
0	90	42	43	20.87974	82.00274	6.502985	202.9355	rice
1	85	58	41	21.77046	80.31964	7.038096	226.6555	rice
2	60	55	44	23.00446	82.32076	7.840207	263.9642	rice
3	74	35	40	26.4911	80.15836	6.980401	242.864	rice
4	78	42	42	20.13017	81.60487	7.628473	262.7173	rice
5	69	37	42	23.05805	83.37012	7.073454	251.055	rice
6	69	55	38	22.70884	82.63941	5.700806	271.3249	rice
7	94	53	40	20.27774	82.89409	5.718627	241.9742	rice
8	89	54	38	24.51588	83.53522	6.685346	230.4462	rice
9	68	58	38	23.22397	83.03323	6.336254	221.2092	rice
10	91	53	40	26.52724	81.41754	5.386168	264.6149	rice
11	90	46	42	23.97898	81.45062	7.502834	250.0832	rice
12	78	58	44	26.8008	80.88685	5.108682	284.4365	rice
13	93	56	36	24.01498	82.05687	6.984354	185.2773	rice
14	94	50	37	25.66585	80.66385	6.94802	209.587	rice
15	60	48	39	24.28209	80.30026	7.042299	231.0863	rice
16	85	38	41	21.58712	82.78837	6.249051	276.6552	rice
17	91	35	39	23.79392	80.41818	6.97086	206.2612	rice
18	77	38	36	21.86525	80.1923	5.953933	224.555	rice
19	88	35	40	23.57944	83.5876	5.853932	291.2987	rice
20	89	45	36	21.32504	80.47476	6.442475	185.4975	rice
21	76	40	43	25.15746	83.11713	5.070176	231.3843	rice
22	67	59	41	21.94767	80.97384	6.012633	213.3561	rice
23	83	41	43	21.05254	82.6784	6.254028	233.1076	rice
24	98	47	37	23.48381	81.33265	7.375483	224.0581	rice
25	66	53	41	25.07564	80.52389	7.778915	257.0039	rice
26	97	59	43	26.35927	84.04404	6.2865	271.3586	rice
27	97	50	41	24.52923	80.54499	7.07096	260.2634	rice

Figure 3.1: Custom Dataset Sample

In the above sample, the data points are on nitrogen, phosphorus, pH, temperature, rainfall, and humidity. In the last column, we can view the labeled data according to the corp.

Crop and Fruit	Data Points
Banana	180
Watermelon	180
Mango	160
Rice	140
Pigeon Peas	140
Moth Beans	140
Black gram	140
Lentil	140
Pomegranate	140
Muskmelon	140
Papaya	140
Coconut	140
Maize	120
Mung bean	120
Cotton	120
Jute	120
Total	2260

Table 3.2: Unique Data Points on different crops of Bangladesh

3.3 Data Cleaning

After labeling the data we looked for certain errors in the dataset. We have found errors like duplicate data, symbols, punctuation, wrong input etc. Then we started the data cleaning process.

At first, we looked whether there were null values or not using the useful function Assert [15]. Then we removed the null values using the null value operation. Secondly, we checked whether there are duplicate values or not using the useful function Assert. Then we eliminated all duplicated rows with drop_duplicates. Lastly, we checked whether all the values are unique or not with the help of library functions. After cleaning the data we have got the following 2260 unique values

	Unnamed: 0	N	P	K	temperature	humidity	ph	rainfall	label
0	0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...	...	...	...	...	...	...	...	...	...
3855	2095	87	44	43	23.874845	86.792613	6.718725	177.514731	jute
3856	2096	88	52	39	23.928879	88.071123	6.880205	154.660874	jute
3857	2097	90	39	37	24.814412	81.686889	6.861069	190.788639	jute
3858	2098	90	39	43	24.447439	82.286484	6.769346	190.968489	jute
3859	2099	84	38	43	26.574217	73.819949	7.261581	159.322307	jute

3860 rows × 9 columns

Figure 3.2: Before cleaning the dataset

	N	P	K	temperature	humidity	ph	rainfall
count	2260.000000	2260.00000	2260.000000	2260.000000	2260.000000	2260.000000	2260.000000
mean	56.308407	45.29646	33.338053	26.948501	75.470661	6.449085	101.765575
std	37.656752	22.30688	13.143532	4.223495	15.953257	0.777079	60.751083
min	0.000000	5.00000	15.000000	18.041855	30.400468	3.504752	20.000000
25%	23.750000	25.00000	21.000000	24.415225	63.386295	6.000000	54.028381
50%	46.500000	47.00000	32.000000	26.795475	80.719235	6.427890	91.880393
75%	91.000000	61.00000	46.000000	29.030294	88.717139	7.000000	131.069138
max	140.000000	95.00000	55.000000	44.000000	99.981876	10.000000	298.560117

Figure 3.3: After cleaning the dataset

Figure 2 and 3 demonstrates the change in our dataset after the data-cleaning process.

After cleaning the data and making it more efficient for our models to understand, we started the feature selection process.

3.4 Feature Selection

For our models to be more efficient and provide better recommendations we selected certain features of our dataset. The crop data in the figure contains the optimal level of nitrogen, phosphorus, pH, temperature, rainfall, and humidity needed to harvest the crops in order to achieve high yield. The below table provides an initial idea of the data

	N	P	K	temperature	humidity	ph	rainfall
count	2260.000000	2260.00000	2260.000000	2260.000000	2260.000000	2260.000000	2260.000000
mean	56.308407	45.29646	33.338053	26.948501	75.470661	6.449085	101.765575
std	37.656752	22.30688	13.143532	4.223495	15.953257	0.777079	60.751083
min	0.000000	5.00000	15.000000	18.041855	30.400468	3.504752	20.000000
25%	23.750000	25.00000	21.000000	24.415225	63.386295	6.000000	54.028381
50%	46.500000	47.00000	32.000000	26.795475	80.719235	6.427890	91.880393
75%	91.000000	61.00000	46.000000	29.030294	88.717139	7.000000	131.069138
max	140.000000	95.00000	55.000000	44.000000	99.981876	10.000000	298.560117

Figure 3.4: Different Features of The dataset

From our initial research on different crops, we have found that Nitrogen helps the growth of leaves on the plant whereas Phosphorus accounts for root growth and flower and fruit development.

On the other hand, Potassium helps the overall functions of the plant perform correctly. Moreover, different plants require different optimal temperatures, humidity, pH, and rainfall. This is why we have chosen these seven features for our analysis.

All the above criteria in the soil are required for crops to grow perfectly and maximize the profit of the farmers.

3.5 Exploratory Data Analysis

After cleaning the data and selecting certain features from the data we initiated the EDA process. It will help us better understand the data and also help us understand the correlation between different data points. As a result, we will be able to understand patterns and detect errors when training and testing our models.

The below graphs show a glimpse of our analysis

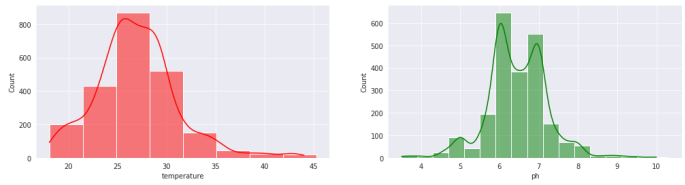


Figure 3.5: Predicted Temperature and pH level for selected crops

The above graph indicates Temperatures mostly ranged from 20 to 42 degrees Celsius. Meaning almost all the crops in our dataset grow within this temperature range. Moreover, a high number of the crops in the dataset needed an optimal temperature of 25 degree Celsius and a pH level of 6-6.5 to grow.

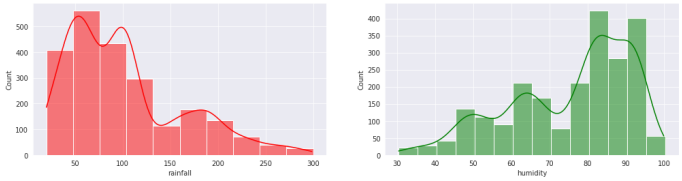


Figure 3.6: Rainfall and humidity level for selected crops

In figure 6, we find that the rainfall level ranges between 50-280 which is normal for our country. Rainfall is a very important feature of our dataset. As nowadays our country gets rain almost every season. Moreover, it is a great source of water for our farmers and also washes away the pests and germs from the plant. Most of our crops prefer a rainfall of 100 mm and humidity between 90-100 to grow.

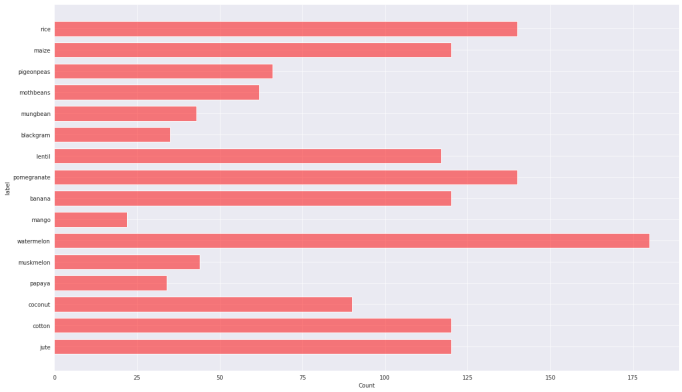


Figure 3.7: Data points on different crops

Figure 7 shows the unique data points available on each crop after cleaning the data.

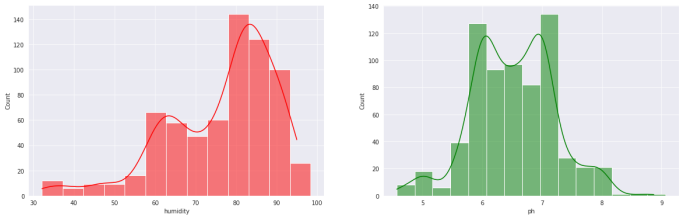


Figure 3.8: Humidity vs pH

In figure 8 we can understand more crops grow between 80-90% humidity which is mid-level and also medium pH level of 6-8 is also good for crops.

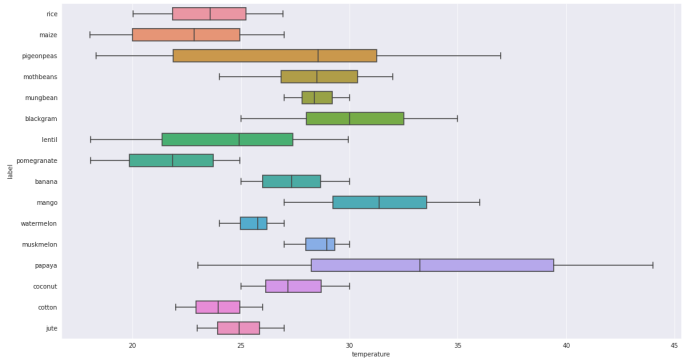


Figure 3.9: Optimal Temperature for selected crops

In figure 9 we can see the predicted optimal temperature of our models for different crops from the data points on different crops.

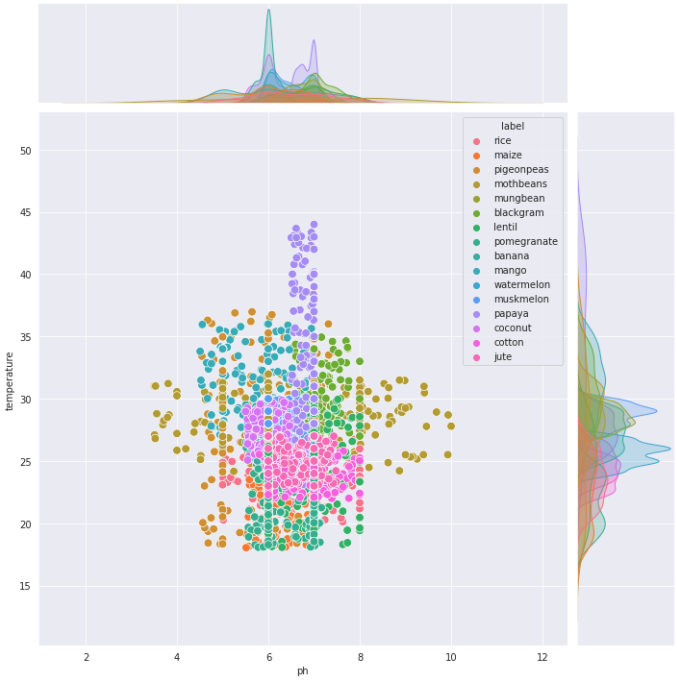


Figure 3.10: Optimal Temperature vs pH level for selected crops

Figure 10 demonstrates the Optimal Temperature vs pH level for our selected crops. The dense area is in the middle which proves that most of the selected crops in the dataset grows better within 22 degree Celsius to 35 degree Celsius and at a pH level of 5-8.

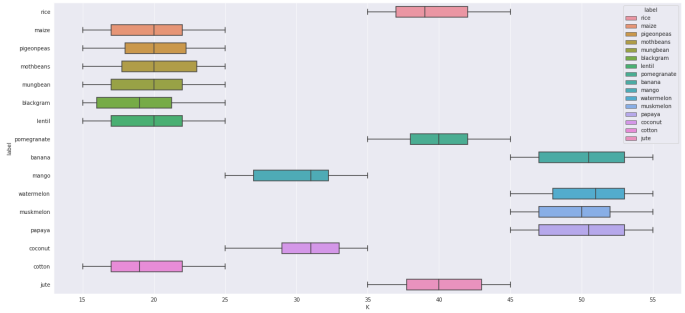


Figure 3.11: Optimal potassium level in soil for selected crops

In figure 11 we can see the required potassium level for each crop in our dataset.

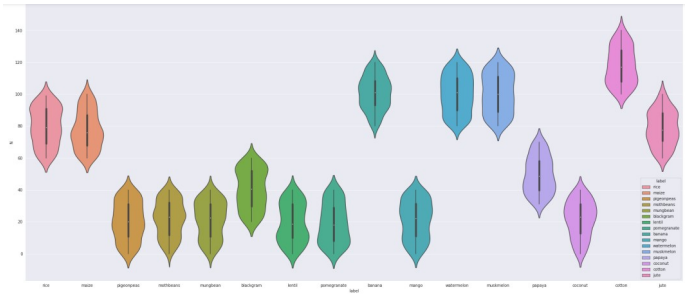


Figure 3.12: Optimal Nitrogen level in soil for selected crops

In figure 12 we can see the required Nitrogen level for the crops in the dataset.

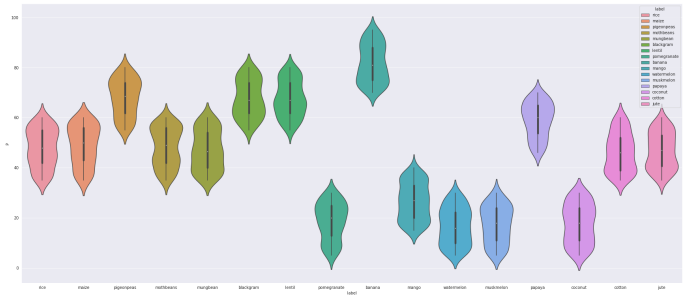


Figure 3.13: Optimal Phosphorus level in soil for selected crops

Figure 13 demonstrates the required Phosphorus level for the crops in the dataset.

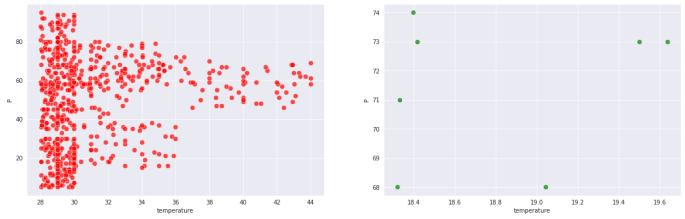


Figure 3.14: Selected crops growth in different temperature levels

Figure 14 shows that most of our selected crops have a mid-level optimal temperature to grow.

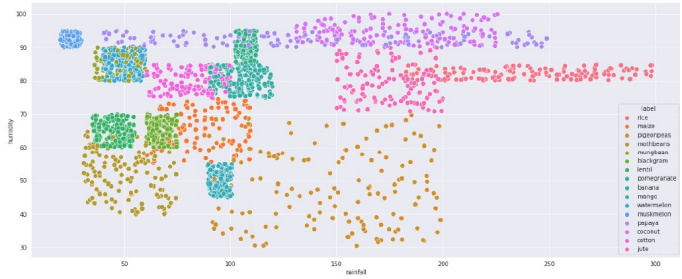


Figure 3.15: Optimal Rainfall and Humidity for selected crops

In figure 15 we can see the optimal rainfall and humidity of different crops in our dataset.

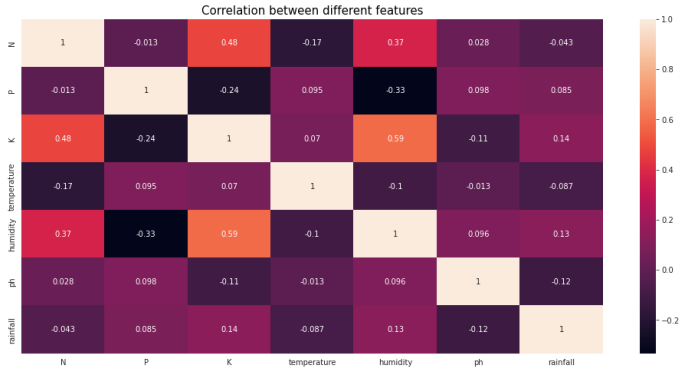


Figure 3.16: Correlation between different features

Figure 16 demonstrates the correlation between different features of our dataset.

Chapter 4

Research Methodology

4.1 Work Plan

This flowchart describes our work procedures step by step In the initial research

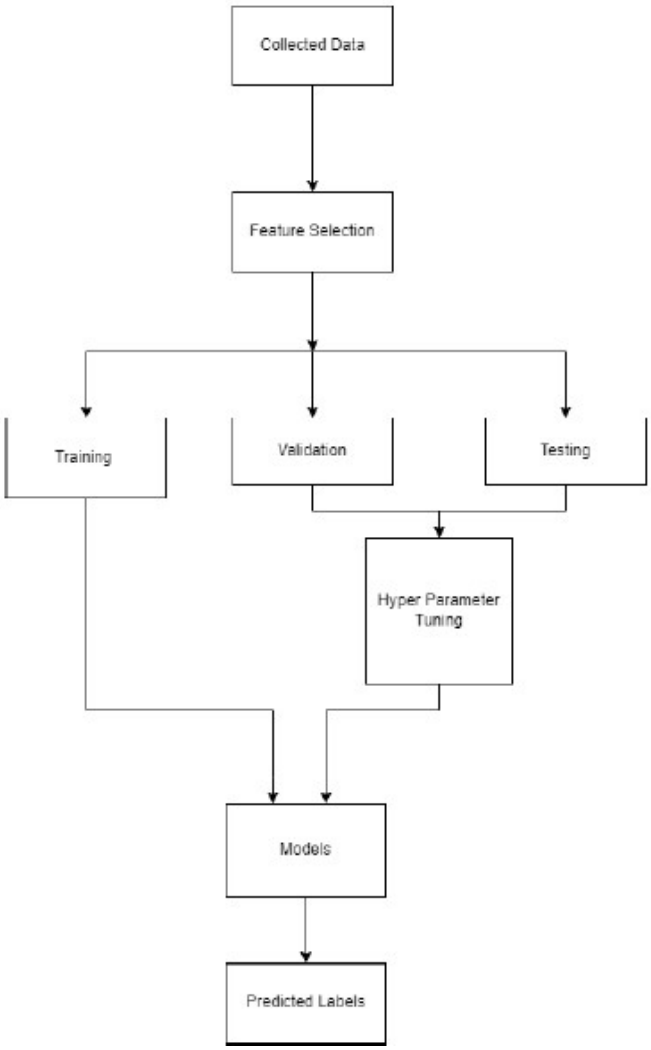


Figure 4.1: Work Plan

period, we explored different approaches to crop yield prediction. According to our research, we developed our own methodology for this research. Then we started

collecting data from Bangladesh Agricultural University.

Once the dataset was ready, we initiated the data-cleaning process and then started the feature selection. After feature selection, we split the data into training, validation, and testing. We also tuned some parameters of our deep learning models. Lastly, we trained and tested our models and bought predicted results.

Chapter 5

Model Specification

We have tested a handful of models. However, Decision Tree, Logistic Regression, SVM and our custom MLP have bought the best results. The model descriptions are given below.

5.1 SVM

SVM is a supervised machine learning classifier that can be used to solve binary classification, multiclass classification, or regression problems. The full form of SVM is Support Vector Machine. SVMs are particularly well-suited for classification of complex but small- or medium-sized datasets.

In the case of classification, the algorithm tries to find the hyperplane that maximally separates the two classes, which is done by finding the one with the largest margin between the nearest points in each class. Points that are nearest to the hyperplane are called support vectors[16], and the distance between the nearest points and the hyperplane is called the margin. In the case of regression, the algorithm tries to find the hyperplane that minimally separates the data points, which is done by finding the one that minimizes the distance between the points and the hyperplane.

One advantage of support vector machines is that they can be highly effective even when the number of dimensions is very high. Another advantage is that it can be used with both linear and nonlinear data by using the kernel trick, which allows them to learn more complex relationships in the data.

There are some limitations to SVM, however. One limitation is that they are sensitive to the choice of kernel and the kernel's hyper-parameters, which can be difficult to tune. Another limitation is that it can be inefficient to train on very large datasets.

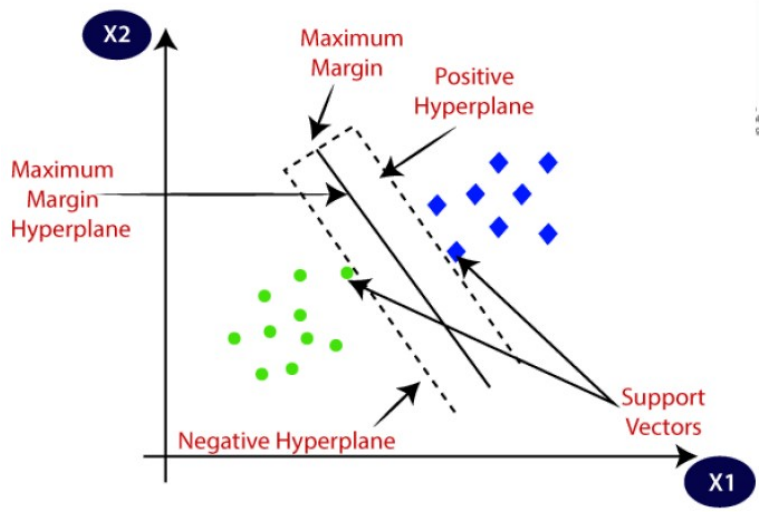


Figure 5.1: Support Vector Machines

5.2 Logistic Regression

Logistic regression is a statistical model that is used for binary classification, i.e. it is used to predict the probability of a binary response typically labeled as ”0” or ”1” based on one or more predictor variables also known as features.

In logistic regression, the log odds of the binary response are modeled as a linear combination of the predictor variables. The model estimates the probability of the binary response using the sigmoid function, which maps the log odds to a value between 0 and 1.

The coefficients which are known as weights or parameters of the logistic regression model are estimated using maximum likelihood estimation, which involves finding the values of the coefficients that maximize the likelihood of the observed data. The maximum likelihood estimate of the coefficients can be found using an iterative optimization algorithm [17], such as gradient descent. One of the strengths of logistic regression is that it can handle both continuous and categorical predictor variables. It is also relatively easy to implement and interpret, and it is robust to noise in the data.

However, logistic regression has some limitations. It assumes that the relationship between the predictor variables and the binary response is linear, which may not be the case in practice. It also assumes that the errors in the model are independent and identically distributed, which may not always be true.

Despite these limitations, logistic regression is a widely used and effective method for binary classification and is a good starting point for many machine-learning tasks.

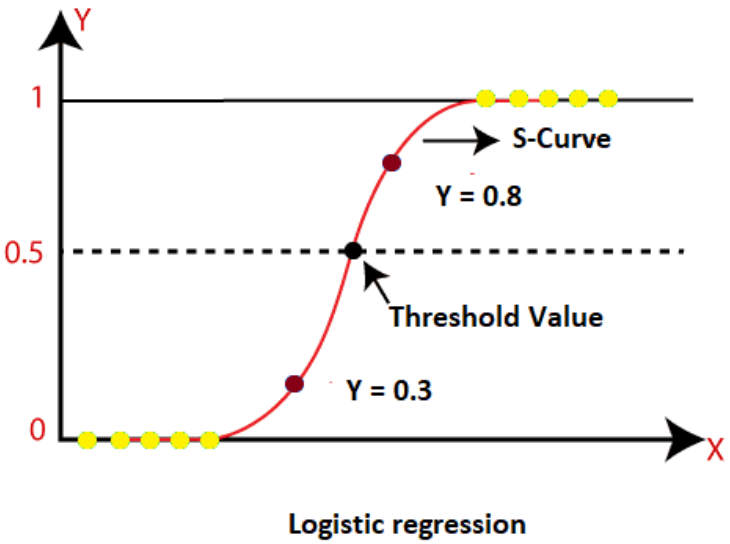


Figure 5.2: Logistic Regression

5.3 Decision Tree

Decision tree is a machine learning algorithm that belongs to the supervised learning category. It is a flowchart-like tree structure, where an internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to make decisions based on the features of the data, splits the data based on a decision rule, and generates sub-nodes. The process continues until it reaches a leaf node, which is the final decision.

In case you want to build a decision tree to predict whether an individual has diabetes or not based on certain features such as age, BMI (Body Mass Index), and blood pressure.

First, you need to select the feature that will be used to split the data at the root node. You can choose any of the three features - age, BMI, or blood pressure - as the root node. Let's say you choose age as the root node.

Next, you need to decide on a decision rule for the root node. For example, you can say that if the age of an individual is less than 30, then the individual does not have diabetes (outcome 0). If the age is greater than or equal to 30, then the individual has diabetes (outcome 1).

Now, the data will be split into two groups based on the decision rule - one group for individuals with age less than 30 and the other group for individuals with age greater than or equal to 30.

For the group with age less than 30, you can build a sub-node with BMI as the feature and a decision rule that says if the BMI of an individual is less than 25, then the individual does not have diabetes (outcome 0). If the BMI is greater than or equal to 25, then the individual has diabetes (outcome 1).

Similarly, for the group with age greater than or equal to 30, you can build a sub-node with blood pressure as the feature and a decision rule that says if the blood pressure of an individual is less than 120, then the individual does not have diabetes (outcome 0). If the blood pressure is greater than or equal to 120, then the individual has diabetes (outcome 1).

This process continues until you reach a leaf node, which is the final decision. The decision tree algorithm uses a greedy approach to make decisions, meaning it chooses the feature and decision rule that gives the best split at each step without considering the long-term impact of the decision on the final outcome.

Decision trees are simple to understand and interpret, and they can be used for classification or regression tasks. They are also easy to implement, but they can be prone to overfitting if the tree becomes too complex. To avoid overfitting, you can prune the tree or use ensemble methods such as random forests or gradient boosting.

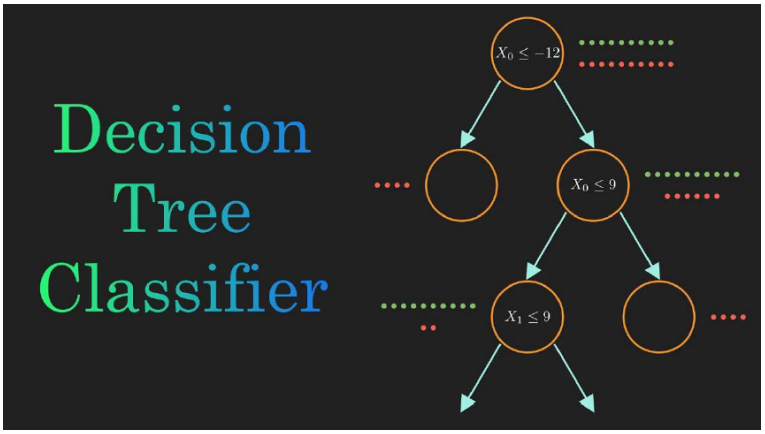


Figure 5.3: Decision Tree

5.4 Multi-layer Perceptron

The basic units of computation in an ANN [18] are called ”neurons,”. Multilayer Perceptron is a type of artificial neural network consisting of multiple layers of interconnected ”neurons” . Each successive layer processes the data using weighted connections between the neurons and an activation function after receiving the raw input data from the input layer. The model’s ultimate output is provided by the output layer.

In Multi-layer Perceptron, the layers between the input and output layers are called hidden layers, and the neurons in these hidden layers use the weighted connections and activation function to transform the input data into a representation that is useful for predicting the output. The weights of the connections between the neurons are adjusted during training, allowing the network to ”learn” to map the input

data to the desired output.

Multi-layer Perceptron is used for a variety of tasks, including image classification, language translation, and even playing games. They are particularly useful for tasks that require the modeling of complex relationships between the input data and the output.

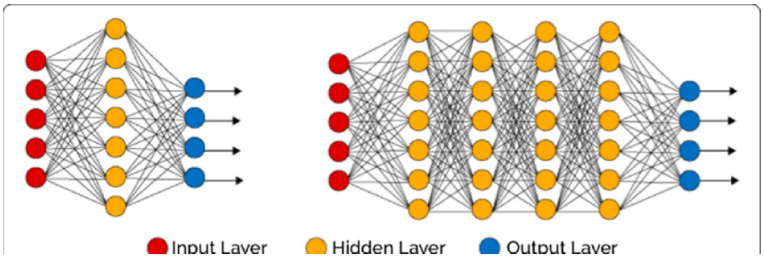


Figure 5.4: Custom Multilayer Perceptron

Chapter 6

Experiment Setup

We have trained and tested our data with various machine learning classifiers and our custom deep learning model.

6.1 Machine Learning Approach

For our machine learning approach, we tried SVM, Logistic Regression, and Decision tree.

After cleaning the dataset we looked for empty rows and columns and dropped them. Then we rechecked the dataset for duplicate values. Once we are done checking the dataset we split our cleaned data into 80:20 ratios for training and testing. 80 percent of our data is used for training while 20 percent was used for testing.

Then we started to train our data and once training was done we fed our selected models the test data which bought back test results.

After testing we have also initiated the cross-validation process [19] with cv=05 for our machine-learning models to detect overfitting.

6.2 Deep Learning Approach

Setting up the deep learning model is a bit more complex than the machine learning models. As we can not feed labeled words and alphabets to the deep learning models. So, first, we encoded our labels into numbers. All the numbers in the dataset were transformed into StandardScaler.

6.2.1 Model Setup Of Multi-layer Perceptron

After the data preparation, building the model is next. we trained and tested our data with various Multi-layer Perceptron setups. From there we have found that Multi-layer Perceptron with two hidden layers brings the most efficient result. So, the proposed model is made of three MLP layers.

In Keras, an MLP layer is referred to as dense, which stands for the densely connected layer.Both the first and second MLP layers are identical in nature with 22 units and 64 units, followed by the Rectified Linear Unit (ReLU) activation.

$$Relu(z) = max(0, z) \tag{6.1}$$

Multi-layer Perceptron is a sequential model [20]. The output layer has 16 units followed by a softmax activation layer. The 16 units correspond to the 16 possible labels, classes, or categories.

Softmax Function

$$X_i = \frac{e^{x_i}}{\sum_{j=1}^{N-1} e^{x_j}} \; j = 1, ..., n \tag{6.2}$$

The equation is applied on all N = 16 outputs, for i = 0, 1 … 15 for the final prediction. The idea of softmax is surprisingly simple. It squashes the outputs into probabilities by normalizing the prediction. Here, each predicted output is a probability that the index is the correct label of the given input features. The sum of all the probabilities for all outputs is 1.0.

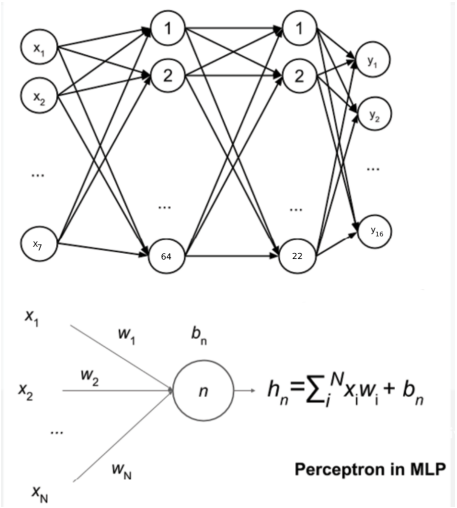


Figure 6.1: Custom Multilayer Perceptron

First, we set the input shape to 7 as there are 7 features of our data. Then the first hidden layer transformed our 7 features into 22 neurons. The second hidden layer took the 22 neurons as input and transformed them into 64 neurons. In the last output layer, the 64 neurons were taken as input and provided an output of 16. As we have 16 crops to predict.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 22)                176

 dense_7 (Dense)             (None, 64)                1472

 dense_8 (Dense)             (None, 16)                1040

=================================================================
Total params: 2,688
Trainable params: 2,688
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 6.2: Model summary of Custom Multilayer Perceptron

Chapter 7

Result and Discussion

We have tried both machine learning and deep learning approaches. The test results are then evaluated in four assessment criteria. They are **Accuracy:** It is the number of accurate predictions out of total predictions.

$$\frac{\text{Number of Correct Prediction}}{\text{Total number of predictions made}}$$

F1 Score: The F1 Score summarizes a model's predictive effectiveness by merging two previously opposing variables, accuracy, and recall.

$$2 * \frac{\text{precision * recall}}{\text{precision + recall}}$$

Precision: Precision determines the standard of positive predictions determined by the models. It's derived by dividing the total number of positive forecasts by the number of genuine positives.

$$\frac{TP}{TP+FP}$$

Recall: It estimates a model's ability to determine positive samples. It is estimated within the range of [0,1].

$$\frac{TP}{TP+FN}$$

7.1 Machine Learning Models

For our Machine Learning classifiers, we have observed the Precision, Recall, F1 score, and Test accuracy for Decision Tree, logistic regression, and, SVM

The below tables describe that in a supervised environment, our chosen models gained accuracy in selecting the accurate crop.

7.1.1 Test Result Of Decision Tree

Crops	Precision	Recall	F1-Score	Support
Banana	1.00	1.00	1.00	34
Cotton	1.00	1.00	1.00	25
Maize	1.00	1.00	1.00	22
Coconut	1.00	1.00	1.00	29
Mango	1.00	1.00	1.00	32
Jute	0.80	0.96	0.87	25
Moth beans	1.00	0.64	0.78	28
Black gram	0.84	1.00	0.92	27
Mung bean	1.00	1.00	1.00	33
Lentil	0.81	1.00	0.89	21
Muskmelon	1.00	1.00	1.00	30
Papaya	1.00	1.00	1.00	30
Pigeon peas	1.00	1.00	1.00	31
Pomegranate	1.00	1.00	1.00	24
Rice	0.96	0.79	0.86	28
Watermelon	1.00	1.00	1.00	33
Accuracy			0.96	452
Macro avg	0.96	0.96	0.96	452
Weighted avg	0.97	0.96	0.96	452

Table 7.1: Evaluation Matrix of Decision Tree

7.1.2 Confusion Matrix Of Decision Tree

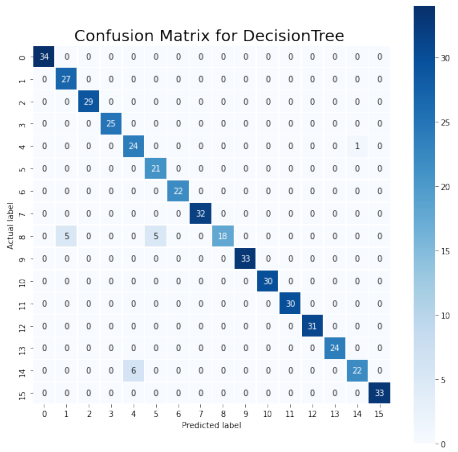


Figure 7.1: Confusion Matrix of Decision tree

This confusion matrix of the Decision tree classifier shows the summary of correct and incorrect predictions by the model. For example, in the case of Bananas, it predicted all the results correctly. On the other hand, in the case of Jute, it predicted 21 corrects and 5 wrongs.

7.1.3 Test Result Of Logistic Regression

Crops	Precision	Recall	F1-Score	Support
Rice	0.87	0.93	0.90	28
Muskmelon	1.00	1.00	1.00	30
Pigeon peas	0.97	1.00	0.98	31
Watermelon	1.00	1.00	1.00	33
Jute	0.88	0.88	0.88	25
Lentil	0.91	0.95	0.93	21
Banana	1.00	1.00	1.00	34
Maize	0.94	0.77	0.85	22
Black gram	0.96	0.85	0.90	27
Mango	1.00	1.00	1.00	32
Coconut	1.00	1.00	1.00	29
Moth beans	0.80	0.86	0.83	28
Cotton	0.89	0.96	0.92	25
Mung bean	0.94	0.97	0.96	33
Papaya	1.00	0.93	0.97	30
Pomegranate	1.00	1.00	1.00	24
Accuracy			0.95	452
Macro avg	0.95	0.94	0.94	452
Weighted avg	0.95	0.94	0.94	452

Table 7.2: Evaluation Matrix of Logistic Regression

7.1.4 Confusion Matrix Of Logistic Regression

Figure 7.2 demonstrates the confusion matrix for the logistic regression classifier. This means, it shows the summary of correct and incorrect predictions by the model. For example, the 15th crop of our dataset is watermelon and in the test data there are 33 samples of watermelon and logistic regression has predicted all of them to be correct. On the other hand, In case of 6 or lentils, logistic regression has predicted 17 corrects 4 wrongs among which it predicted 3 samples to be coconut and 1 to be cotton.

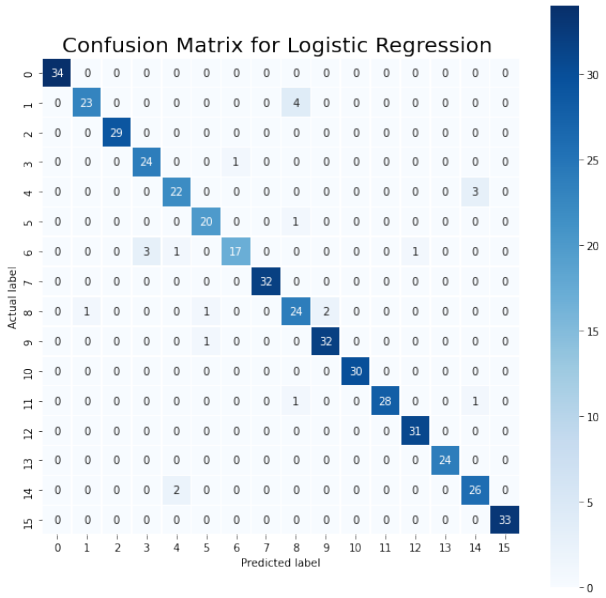


Figure 7.2: Confusion Matrix of logistic regression

7.1.5 Test Result Of SVM

Crops	Precision	Recall	F1-Score	Support
Banana	0.09	1.00	0.16	34
Blackgram	1.00	0.04	0.07	27
Coconut	0	0	0	29
Cotton	1.00	0.04	0.08	25
Jute	1.00	0.04	0.08	25
Lentil	0	0	0	21
Maize	0	0	0	22
Mango	1.00	0.31	0.48	32
Mothbeans	0	0	0	28
Mungbean	0	0	0	33
Muskmelon	1.00	0.30	0.46	30
Papaya	0	0	0	30
Pigeonpeas	0	0	0	31
Pomegranate	1.00	0.08	0.15	24
Rice	0	0	0	28
Watermelon	1.00	0.94	0.97	31
Accuracy			0.20	452
Macro avg	0.48	0.17	0.15	452
Weighted avg	0.46	0.20	0.17	452

Table 7.3: Evaluation Matrix of SVM

7.1.6 Confusion Matrix Of SVM

The matrix demonstrates the confusion matrix for the SVM classifier. This means it shows the summary of correct and incorrect predictions by the model. As this

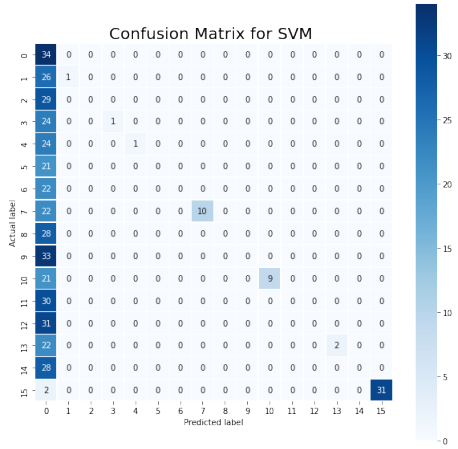


Figure 7.3: Confusion Matrix of SVM

model has the lowest accuracy with our dataset so, it has only done well in just a few test data samples like 15 or watermelon.

7.1.7 Cross Validation

The Below table shows the cross-validation score for our chosen classifiers.

Model	Cross Val Score	Cross Val Score	Cross Val Score	Cross Val Score	Cros
Decision Tree	0.97566372	0.94690265	0.9579646	0.96238938	0.982
Logistic Regression	0.94911504	0.94690265	0.95353982	0.94911504	0.962
SVM	0.15486726	0.17035398	0.18141593	0.2079646	0.199

Table 7.4: Cross Validation Score Of Different Models

From the above table, we can see the cross-validation score of our models is in close proximity to our predicted accuracy. So, we can clearly state that our selected models might not overfit on our tested dataset.

7.2 Custom Multi-layer Perceptron Model

Our custom Multi-layer Perceptron model with 100 epochs, .30 validation, and batch size 50 has bought the most satisfactory result. It has gained 99.28% train accuracy and 98.00% test accuracy which is the highest among all the models.
Figure 7.4 and figure 7.5 demonstrate the model accuracy and model loss for Custom Multi-layer Perceptron.

The blue line shows accuracy and loss in test data and the yellow line shows accuracy and loss in test data.

In the case of model accuracy, the slope goes close to 1 as our model accuracy was .98 or 98%. On the other hand, the slope is close to .20 which indicates that our

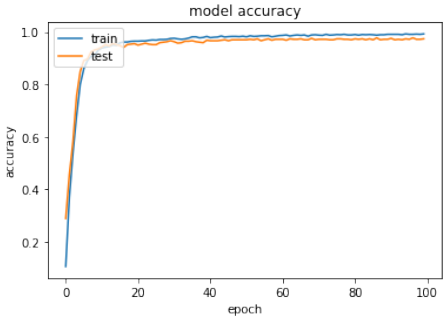


Figure 7.4: Model Accuracy of Custom Multi-layer Perceptron

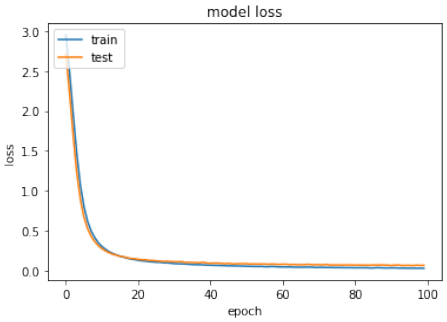


Figure 7.5: Model Loss of Custom Multi-layer Perceptron

model has 0.2 wrong predictions.

Moreover, in the case of both model accuracy and model loss, the test and train slope is in close proximity which indicates there is little chance of overfitting.

Crops	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	34
1	0.85	1.00	0.92	28
2	1.00	1.00	1.00	22
3	1.00	1.00	1.00	31
4	1.00	1.00	1.00	28
5	1.00	1.00	1.00	33
6	0.96	0.96	0.96	27
7	0.95	0.95	0.95	21
8	1.00	1.00	1.00	24
9	1.00	1.00	1.00	34
10	1.00	1.00	1.00	32
11	1.00	1.00	1.00	33
12	1.00	0.97	0.98	30
13	1.00	1.00	1.00	29
14	1.00	1.00	1.00	25
15	1.00	0.84	0.91	25
Accuracy			0.98	452
Macro avg	0.99	0.98	0.98	452
Weighted avg	0.99	0.98	0.98	452

Table 7.5: Evaluation Matrix of Multi-layer Multi-layer Perceptron

7.2.1 Confusion Matrix Of Multilayer Perceptron

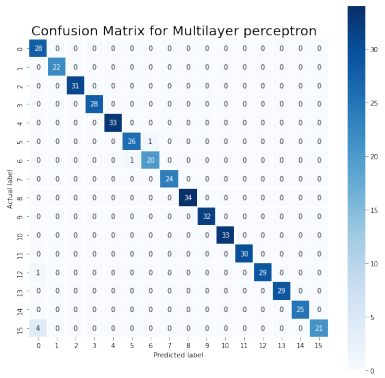


Figure 7.6: Confusion Matrix of Custom Multilayer Perceptron

In the confusion matrix, we can see the correct and wrong predictions of our labeled data from our Custom Multilayer Perceptron. For instance, out of 33 data samples on 01 or Banana our Custom Multi-layer Perceptron has predicted 28 to be correct 0, 4 samples were predicted to be data sample 15, and 1 sample to be 12.

7.3 Model Comparison

We found that our custom Multilayer perceptron brings out the best result of 98.00%. As it's a simple ANN model which works great with complex non-linear problems. Moreover, it's easy to train and test with both large and small datasets.

Secondly, the Decision Tree brings the most satisfactory test accuracy of 96.20%. Because it provides a highly effective structure within which we can lay out options and investigate the possible outcomes of choosing those options. As a result, it is one of the best fit for our multiclass classification problem.

Thirdly, we have found that logistic regression brings the second-best result. It has provided an accuracy of 94.10%. Because it brings out good performance with linearly separable classes. It is an extensively employed algorithm for multiclass classification.

Lastly, SVM brings out the most unsatisfactory result of 20.02%. SVM does not bring satisfactory results because in some instances our dataset has close or overlapping data points.

Model	Accuracy
Decision Tree	96.20%
Logistic Regression	94.10%
SVM	20.02%
Multilayer Perceptron	98.00%

Table 7.6: Accuracy Of Different Models

7.4 Final Result

As we have achieved the best result with Multilayer perception with our dataset. So, we will proceed with Multilayer perceptron for future research.

Because Multilayer perceptron can handle both large and small datasets. On the other hand, it is more precise than the tested machine learning classifiers. Moreover, Multilayer perceptron is less complex and requires less hardware capabilities to run.

Chapter 8

Conclusion and Future Work

8.1 Further Research

For finding the best-performing techniques for Bangladeshi crop yield prediction we have already tested 5 machine-learning classifiers and 3 deep-learning models. However, due to limited data avaiability, we could only bring satisfactory results from 3 machine learning models and 1 custom deep learning model.

So, our first future concern is enriching our dataset so that the usability and reliability of our models can increase. Moreover, we plan to release our dataset publicly in the future. We will also try to test more state-of-the-art models with our enriched dataset.

8.2 Final Words

In conclusion, building an effective prediction system for crop yields in Bangladesh using Machine Learning and deep learning approaches is one of the most difficult tasks. As our country has not seen notable association between agriculture and state of the art machine learning classifiers before. So, our big challenge is the successful integration of agriculture with Machine Learning and deep learning. However, we take these challenges as our inspiration to develop the most efficient prediction system for crop yields using different Machine Learning approaches. Though this study is still at its beginning level. But our step might be the start of something which will revolutionize the entire agricultural sector of Bangladesh.

Bibliography

[1] Sectoral Growth Rate of GDP at Current Prices. (2016, November 17) http://www.bbs.gov.bd

[2] Production of Major Crops 2012-16. (2017, January 18) http://data.gov.bd

[3] Asaduzzaman, M. (2021, March 25). Agriculture in Bangladesh: The last and the next fifty years. The Daily Star. Retrieved September 9, 2022, from https://www.thedailystar.net

[4] Reuters. (2022, April 30). Govt to buy rice from farmers for 40 Tk a kg. Prothomalo. Retrieved September 9, 2022, from https://en.prothomalo.com

[5] Wikipedia contributors. (2022, May 21). Agriculture in Bangladesh. Wikipedia. Retrieved September 9, 2022, from https://en.wikipedia.org/wiki/Agriculture$_in_Bangladesh$

[6] Satish Babu (2013), 'A Software Model for Precision Agriculture for Small and Marginal Farmers', at the International Centre forFree and Open Source Software (ICFOSS) Trivandrum, India

[7] Aymen E Khedr, Mona Kadry, Ghada Walid (2015), 'Proposed Framework for Implementing Data Mining Techniques to Enhance Decisions in Agriculture Sector Applied Case on Food Security Information Center Ministry of Agriculture, Egypt'

[8] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) , 'Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh' , (SNPD) IEEE/ACIS International Conference.

[9] Liying Yang (2011), 'Classifiers selection for ensemble learning based on accuracy and diversity' Published by Elsevier Ltd. Selection and/or peer-review under responsibility of [CEIS].

[10] Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh (2015), 'Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique', International Conference on Smart Technologies.

[11] Anshal Savla, Parul Dhawan, Himtanaya Bhadada, Nivedita Israni, Alisha Mandholia , Sanya Bhardwaj (2015), 'Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture'.

[12] Monali Paul, Santosh K. Vishwakarma, Ashok Verma (2015), 'Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach', International Conference on Computational Intelligence and Communication Networks.

[13] D Ramesh , B. Vishnu Vardhan, Jan, 2015, ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUE.

[14] Sudhanshu Sekhar Panda, Daniel P. Ames , and Suranjan Panigrahi, 2010, Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques.

[15] Eduard Cerny and Surrendra Dudani, 2009, SystemVerilog: Assertion-based Checker Libraries.

[16] Casey Casalnuovo, Prem Devanbu, Abilio Oliveira, Vladimir Filkov, Baishakhi Ray, 2007,Assert Use in GitHub Projects.

[17] Daniel Santana-Cedrés, Luis Gómez, Miguel Alemán-Flores, Agustín Salgado, Julio Esclarín, Luis Mazorra, Luis Álvarez, 2011, An Iterative Optimization Algorithm for Lens Distortion Correction Using Two-Parameter Models.

[18] S. Agatonovic-Kustrin, R Beresford,Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research,Journal of Pharmaceutical and Biomedical Analysis.

[19] Sytze de Bruin, Dick J. Brus, Gerard B.M. Heuvelink, Tom van Ebbenhorst Tengbergen, Alexandre M.J-C. Wadoux, Dealing with clustered samples for assessing map accuracy by cross-validation, Ecological Informatics, Volume 69, 2022.

[20] Yihu Zhang, Bo Yang, Haodong Liu, Dongsheng Li, A time-aware self-attention based neural network model for sequential recommendation, Applied Soft Computing, Volume 133, 2023