

# **Soybean Oil Price Prediction Using Machine Learning Approach**

**BY**

**Name: MD Mahadi Islam**

**ID: 171-35-1837**

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

Supervised By

**Mr. Khalid Been Badruzzaman Biplob**

Senior Lecturer

Department of SWE

Daffodil International University

Co-Supervised By

**Ms. Syeda Sumbul Hossain**

Senior Lecturer

Department of SWE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**20 JANUARY 2021**

## APPROVAL

This Thesis titled on “**Soybean Oil Price Prediction Using Machine Learning Approach**”, submitted by MD MAHADI ISLAM, ID: 171-35-1837 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

### BOARD OF EXAMINERS



Chairman

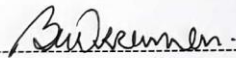
Dr. Imran Mahmud

Associate Professor and Head  
Department of Software Engineering  
Daffodil International University



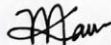
Internal Examiner 1

Nusrat Jahan  
Assistant Professor  
Department of Software Engineering  
Daffodil International University



Internal Examiner 2

Khalid Been Badruzzaman Biplob  
Senior Lecturer  
Department of Software Engineering  
Daffodil International University



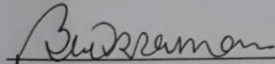
External Examiner

Professor Dr M Shamim Kaiser,  
Professor  
Institute of Information Technology  
Jahangirnagar University

## DECLARATION

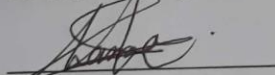
We hereby declare that this thesis has been done by us under the supervision of Mr. Khalid Been Badruzzaman Biplob, Senior Lecturer, Department of Software Engineering, and co-supervision of Ms. Syeda Sumbul Hossain, Senior Lecturer, and Department of Software Engineering Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

### Supervised by:



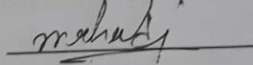
**Mr. Khalid Been Badruzzaman Biplob**  
Senior Lecturer  
Department of Software Engineering  
Daffodil International University

### Co-Supervised by:



**Ms. Syeda Sumbul Hossain**  
Senior Lecturer  
Department of Software Engineering  
Daffodil International University

### Submitted by:



**Md. Mahadi Islam**  
ID: 171-35-1837  
Department of Software Engineering  
Daffodil International University

## ACKNOWLEDGEMENT

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

We are really grateful and express our earnest indebtedness to Mr. Khalid Been Badruzzaman Biplob and Ms. Syeda Sumbul Hossain, Senior Lecturer, Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh. Profound Knowledge & intense interest of our supervisor in the field of “Machine Learning & Deep Learning” make our way very smooth to carry out this thesis. Her remarkable patience and dedication, scholarly guidance, continual encouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, great endurance during reading many inferior drafts and correcting the work to make it unique paves the way of work very smooth and ended with a great result.

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

We would like to express thankfulness to the fellow student of Daffodil International University, who took part in this discussion during the completion of this work.

We would like to express our immense thanks to the Ministry of Agriculture who provided us with the required raw data to make our work possible.

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

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

## **ABSTRACT**

In Bangladesh, market uncertainty is an ongoing issue. The pricing of our ordinary ingredients hence vary so often. It effects the component that we ingest every day considerably. In Bangladesh there are several types of oil. One of them is soybean oil. In Bangladesh, almost every meal includes soybean oil. The prices of the items that are used every day have to be recorded but manually organizing it is a difficult operation. It is quite handy to keep track of the pricing for persons living below the poverty line.

Now we have sophisticated devices in this era of artificial intelligence which can find information from the data. Data insight may be used with the use of machine learning algorithms for prediction purposes. Prediction can be a successful means of eliminating market volatility. We strive to identify approaches of machine learning to estimate the future price of soy bean oil in our study. Our analysis is based on raw data from the Ministry of Agriculture of Bangladesh (MOA).

Machine Learning has numerous prediction methods. For this, our solution was founded using Gradient Boosting, Decision Tree Regression, Lasso Regression, Linear Regression, MLP Regression, Random Forest algorithms. We compared the accuracy in performance to determine the best accuracy All algorithms perform about symmetrically. Our major objective was to find soybean oil future prices.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Acknowledgements	iv
Abstract	v
List of Figure	vii
List of Table	ix

### **CHAPTER**

#### **CHAPTER 1: INTRODUCTION**

**PAGE NO.**

**1-4**

1.1 Introduction	1
1.2 Motivation	2
1.3 Problem Definition	2
1.4 Research Questions	3
1.5 Research Methodology	3
1.6 Research Objective	3
1.7 Report Layout	4
1.8 Expected Outcome	4

#### **CHAPTER 2: BACKGROUND**

**5-8**

2.1 Introduction	5
2.2 Related Work	5
2.3 Bangladesh Perspective	8

#### **CHAPTER 3: RESEARCH METHODOLOGY**

**9-17**

3.1 Introduction	9
------------------	---

3.2 Data collection	9
3.3 Data Preprocessing	10
3.4 Data Analysis	10
3.5 Features	15
3.6 Algorithm Implementation	16
3.7 Evaluation	17
<b>CHAPTER 4: RESULT ANALYSIS</b>	<b>18-22</b>
4.1 Introduction	18
4.2 Experimental Result	18
<b>CHAPTER 5: SUMMARY, CONCLUSION AND FUTURE WORK</b>	
5.1 Summary of the Research	23
5.2 Conclusion	23
5.3 Recommendation	24
5.4 Future Work	24
<b>REFERENCES</b>	<b>25</b>
<b>APPENDIX</b>	<b>27</b>
<b>PLAGIARISM REPORT</b>	<b>28</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO.</b>
Figure 3.1: Methodology diagram	09
Figure 3.2 : price comparison	11
Figure 3.3: Seasonal price analysis.	12
Figure 3.4: Monthly price analysis	13
Figure 3.5 Box Plot of Dhaka price	14
Figure 3.6 KDE plot of dataset.	15
Figure 3.7: Comparison of real and predicted price	17



## LIST OF TABLE

<b>TABLE</b>	<b>PAGE NO.</b>
Table 3.1 Feature Descriptions	15
Table 3.2 Parameter usages	16
Table 4.1 R2 Score	18
Table 4.2 Mean absolute error	19
Table 4.3 Mean squared	19
Table 4.4: Mean squared	20

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

Soybean oil is an important part of our everyday food. Bangladesh's current yearly need for soybean oil is 1.3 million tons, while its demand for palm oil is 1.6 million tons, according to the United States Department of Agriculture's (USDA) April 2021 prediction. Bangladesh's demand is expected to increase by one million tons by 2025, according to the USDA. Bangladesh's current oilseed output is less than half a million tons, according to statistics from the Bangladesh Bureau of Statistics (2019-20). Bangladesh spends a whopping \$2 billion in imports each year to cover its edible oil demands. Unprecedented price volatility in Soybean oil have been a key source of concern in Bangladesh's economy.

Every year, the price of soybean oil in Bangladesh fluctuated drastically. According to the state-run TCB, the maximum retail price for a 5-litre soybean oil bottle in May 2020 was Tk520. Following the new authorization from the Commerce Ministry, the same volume will be sold for Tk728 in November 2021. According to TCB's market pricing report, palm oil prices have already increased by 75% year on year. We discovered that the price fluctuation range was quite wide. The poor people of Bangladesh cannot afford such a high price. Due to the unstructured nature of data and its unpredictability, financial forecasting is difficult. Aside from that, weather conditions, productivity, storage limitations, transportation, and the supply-demand ratio all impact forecasting, making it more complicated.

In this age of artificial intelligence, machines act like people. M. M. Hasan et al. [4] used several machine learning techniques to estimate future Soybean oil prices, which was highly helpful in removing soybean oil market volatility. In the finance industry, machine learning has the highest chance of success. To accomplish so, we use a Machine Learning (ML) approach to predict the price of soybean oil based on the information we have. For this research we used methods such as K-Nearest Neighbor (KNN), Nave Bayes, Decision

Tree, Neural Network (NN), and Vector Machine Support (SVM) to check which method performs the best in high accuracy.

## **1.2 Motivation**

Soybean oil are a significant feature of Bangladeshi food. Every year, we need about 1.3 million tons of Soybean oil in Bangladesh. The price of soybean oil in Bangladesh changed dramatically year after year. The maximum sale price for a 5-litre soybean oil bottle in May was Tk520. Now November 2021, the same quantity will be sold for Tk728 according to a new Commerce Ministry authorization. In terms of Soybean oil price fluctuation, the rate hit a high point. Instead of utilizing the country's production potential, the gap is filled through imports, which are frequently more than the gap, resulting in a market oversupply. It's not impossible that importers have a vested interest in playing games with growing costs, which hurts consumers from all walks of life. The poor are the ones who suffer the most from an increase in Soybean oil prices. As a result, imports may assist in price reductions, making customers pleased.

In this case, prediction can be beneficial. In the financial industry, prediction may be used to forecast future product prices. We now have a variety of powerful Machine Learning algorithms that can easily handle such a big workload. We can predict the future price of Soybean oil by making the best use of these algorithms. As a result, the appropriate authorities will be able to plan how to deal with the problem if it increases.

## **1.3 Problem Definition**

Machine learning is a very important term in the modern ICT field. However, before we can utilize it successfully, we need to understand where and how we can use it. To discover a suitable solution, we must first determine the nature of the problem. We need to understand everything we can about the circumstance if we want our plan to be as successful as possible. We must also effectively point out the investigation's criteria in order to arrive at the most satisfactory result.

We all know that demand and supply are the most important factors in achieving market equilibrium. We can determine the demand-supply balance in the market if we estimate future prices, and we can use this information to keep the market in equilibrium. We can also put our model on the web, which will be beneficial to the consumer.

## **1.4 Research Questions**

- How to get the data and prepare the dataset?
- Identifying recent usage of Machine Learning in pricing prediction applications?
- What is the best way to classify soybean oil prices?
- What is the benefit of this effort to the people?
- The limitations that occur when attempting to predict pricing?
- Is it possible for a machine learning algorithm to accurately predict Soybean oil prices?

## **1.5 Research Methodology**

This section describes our data refinement and procedure. We've spoken about data filtering and attribute selection. This part also goes over how to train models and how to use them. We've also discussed the results of the algorithms.

## **1.6 Research Objectives**

- To predict better model by creating data set of Soybean oil.

## 1.7 Research Layout

The contents of our report will have appeared as regards:

**Chapter 1** will give an overview of our study, including its introduction, motivation, problem definition, research question, research methodology, and predicted outcomes.

**Chapter 2** includes a background study as well as a brief discussion of relevant work in this topic. The following is a list of notable machine learning work, including prediction work.

**Chapter 3** is giving a clear description about methodology or the workflow. How the research has been done have been addressed in this sections?

**Chapter 4** is about the evaluation of the result. It contains the outcome of the research with the graph.

**Chapter 5** is the part of the ending of the research. This section presents the performance of the model. This section also shows the comparison in terms of accuracy. In this section the web implementation part of the model and output are also attached. The chapter ends with viewing the limitations of the work. It also encoded with the future work.

## 1.8 Expected Outcome

- We will predict the future price of Soybean oil.

## **CHAPTER 2**

### **BACKGROUND STUDY**

#### **2.1 Introduction**

For prediction, various machine learning techniques have been used. One of the most extensively utilized applications of Machine Learning is prediction. These research focused on specific problems and used a variety of machine learning techniques to resolve problems. This chapter summarizes the actions that were efficiently carried out by many professionals in the preceding region.

#### **2.2 Related Works**

In today's time machine learning is commonly used for prediction and classification. This section describes the significant exercises performed by a few experts in the recently referred field during the previous several years.

They [1] use machine learning approach to predict the price of gold in this research. Some of the indices they used to estimate gold prices were the stock market, crude oil price, rupee dollar exchange rate, inflation, and interest rate. Three distinct regression-based techniques were used. Regression Linear Random Forest Regression and Gradient Boosting Regression are two types of regression. They gathered monthly gold price data. The statistics and time period are from 2000 to 2018. The data was then separated into three time periods: 2000-2018, 2000-2011, and 2011-2012. 2018. For the years 2000 to 2018, random forest regression provides improved accuracy. For the spans 2000-2011 and 2011-2018, gradient boosting regression is used. They took measurements. The performance of algorithms is measured using MAE, MSE, RMSE, and the Root Mean Square Error value.

The purpose of their research is to use a machine learning technique to predict rice prices [2]. The price was predicted using data from Bangladesh's Ministry of Agriculture website. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Nave Bayes, Decision Tree, and Random Forest were among the machine learning techniques used to produce this prediction. All of these algorithms are compared to see which one delivers the greatest results. The random forest came out as the greatest performer. They can predict the potential future price of rice using their prediction model. It is possible to avoid the rice market from collapsing if such a prediction is made.

In this research [3], they predict the house price using regression techniques. The comparison of different machine learning a regression-based approach was the major emphasis of this work. They used six different regression methods to see which one was the best: Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting Regression, and Gradient Boosting Regression. The accuracy of algorithms is calculated using MSE, RMSE, and Root Mean Square Error. With a Root Mean Square score of 0.9177022, MSE 12037006 088.27804 and RMSE 10971390390[3]. In these strategies, Gradient Boosting Regression provides the greatest accuracy.

M. M. Hasan et al. [4] showed that onion market volatility may be handled by predicting onion price. For this, they used a machine learning system to forecast future pricing. They used daily data for two years. Information gathering The important four phases in finishing the project are the implementation and evaluation of Data Analysis Algorithms. The Neural Network technique, KNN, Nave Bayes, Decision Tree, SVM, and the Neural Network method were all used. The technique provides the maximum accuracy with a rate of 98.17 percent.

In this research, F. A. de Oliveira et al. [5] predicted the stock's current price rather than simply estimating the stock's future price. Leaning might be thought of as a configuration. They are capable of making both short-term (day or week-long) and long-term predictions. They discovered that the backmost produced superior outcomes, with a 79 percent accuracy rate. The recital assessment base of the network is another intriguing aspect of the article cerebrate. The manufacturing evaluation algorithm decides whether to purchase, sell, or retain the stock based on the predicted production.

They developed an Artificial Neural Network (ANN) model for gold price predictions in their paper. Their goal is to create a model that can accurately anticipate the gold price. With the varied layers, number of neurons, input shape, and activation function, they construct a Long Term Short Memory (LSTM) model. They measured the performance of their model using Root mean square error (RMSE) and Mean absolute error (MAE) [6]. The LSTM MODEL is a better model for predicting the gold stock value, according to a comparative study.

They use Support Vector Regression (SVM) and Adaptive Neural Fuzzy Inference System (ANFIS) models to forecast time-series gold price. They look at their Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Nash–Sutcliffe model efficiency coefficient (E), and Mean absolute percentage error (MAPE) scores to see how well they did. ANFIS-GP has the best performance in the ANFIS model [7].

In this research [8], they provide their models for forecasting gold prices on a daily basis . To do this, they used a machine learning method. We used Support Vector Regression (SVR), Random Forest Regressor (RFR), Decision Tree, Gradient Boosting, and XGBoost models to anticipate the daily gold price. All of the models they have created yield results that are very satisfactory. Random Forest Regressor (RFR) has produced the greatest



results in all phases out of all the models we designed. In all circumstances, the SVR algorithm achieves an accuracy of roughly 99 percent.

### **2.3 Bangladesh Perspective**

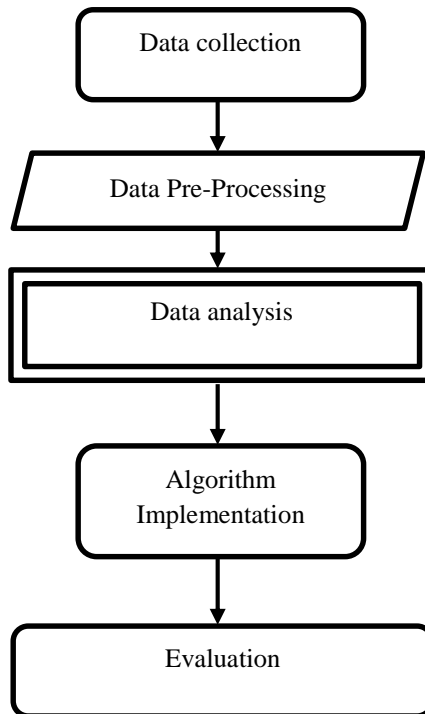
Bangladesh's economy is not very stable at the moment. The economic pace is slowing because of the COVID-19 epidemic. The major threat of reducing market uncertainty has become a serious threat for Bangladesh's lower-income people. Because of the rise in the food industry, those who are financially poor cannot afford many items. Our government usually tries to support everyone in our nation, but they don't have a strong understanding of market instability. The price of Soybean Oil is strongly linked to the state of the economy. [9] However, using Machine Learning, it may be simpler to identify the primary cause of the uncertainty. Furthermore, by appropriately analyzing the market, the government should take appropriate actions to combat it. People should also be aware of the significance of the agricultural industry.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

The strategy of our work comprises an add up to five steps which are the information collection, information investigation, calculation usage, assessment. Figure 3.1. appears



the chart of our work:

Figure 3.1: Methodology diagram

## **3.2 Data Collection**

POS information is expanding day by day. This sort of information is presently utilized for showcase examination. This examination moreover makes a difference the decision-maker to form choices. We collected the desired information from Bangladesh Agriculture Ministry (MOA). We assembled 1215 tests day to day of oil cost for calculation execution. The data was not arranged for advance investigation so we have taken after the another step for planning the data knowledge.

## **3.3 Data Pre-Processing**

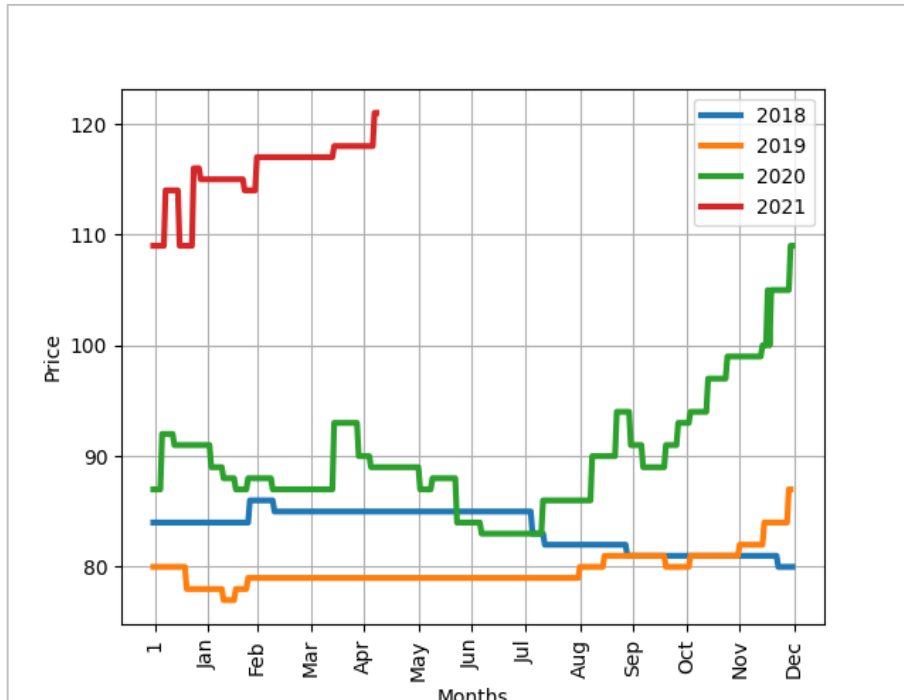
Our initial collected data was not in a clean organize. So we ought to do pre-processing for making the dataset prepared. The information was combined with undesirable data so we need to evacuate the undesirable information. For lost information, we have done some measurable investigation to discover the lost information. For expelling information excess, we connected panda library. ML devices offer assistance to form pre-processing simple.

## **3.4 Data Analysis**

Data in a structured format is required for algorithm implementation. In our research, we used year, month, day, season, location, price, and category as factors. We aim to draw insight from the data as we go over the parameters and produce the appropriate dataset. To aid us do this, we used the dataset graph that we constructed from the dataset. It paints a clear knowledge of the dataset for us.

### **3.4.1 Yearly Price Analysis**

Figure: 3.2 represents the monthly price range or year 2018,2019,2020, 2021. We collected Daily price of each month of year 2018-2020 for year 2021 we collected daily price of January, February, March April months from this figure we have got an important knowledge about our collected data. Firstly, year 2018 and year 2019 price of soybean remaining nearly stable. But for year 2020 and 2021 there are slight instability has been



displayed by our graph. Specifically, last of each year price are unstable. Blue yellow green red line graph represents monthly price of year 2018,2019,2020, 2021.

Figure 3.2: price comparison

### 3.4.2 Seasonally Price Analysis

This figure represents the scatter plot of six seasons. We have taken season as a parameter. Because there are six seasons in Bangladesh each season contains its own characteristics. We also analysed this and try to find out if a season actually affects the price of soybean. In this graph, the x-axis represents the price and the y-axis represents the seasons. Each season consists of two months. The seasons Summer, Rainy, Autumn, Late Autumn, Winter, Spring represented by 0-5. From this graph we can see that the price is quite high in the last 3 seasons.

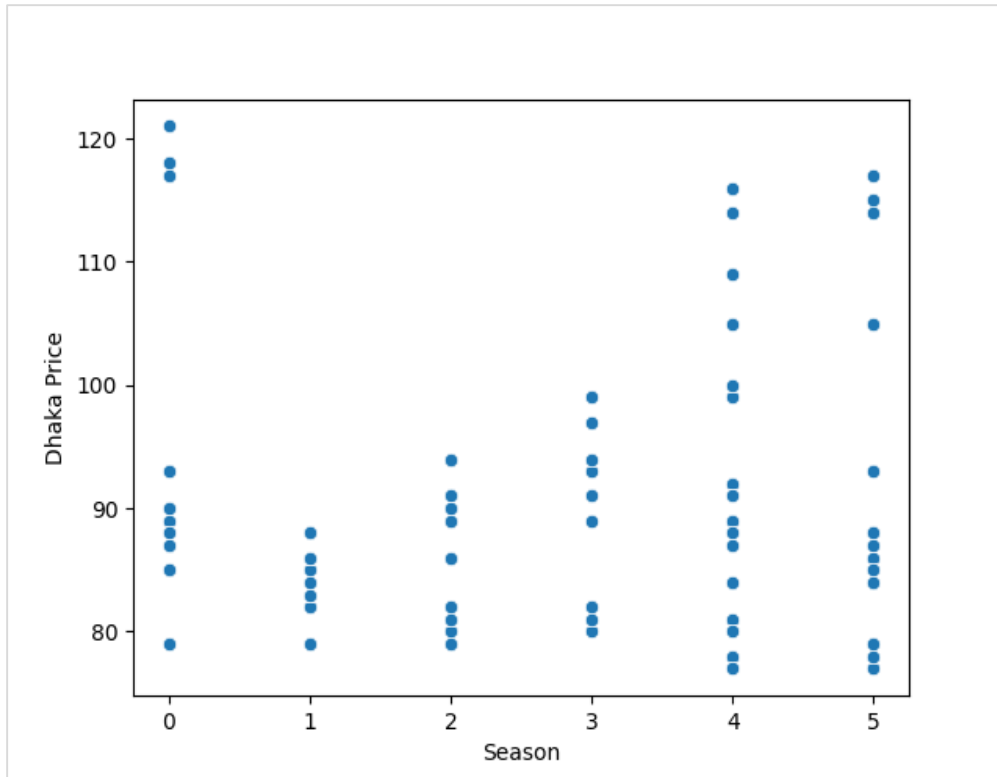


Figure 3.3: Seasonal price analysis

### 3.4.3 Monthly Price Analysis

This figure represents the scatter plot of monthly price. by this graph we tried to show price up down or price range of each month. This graph represents that average price of soybean remains high in the month of January February and March. Other seven months price is quite stable than first three months.

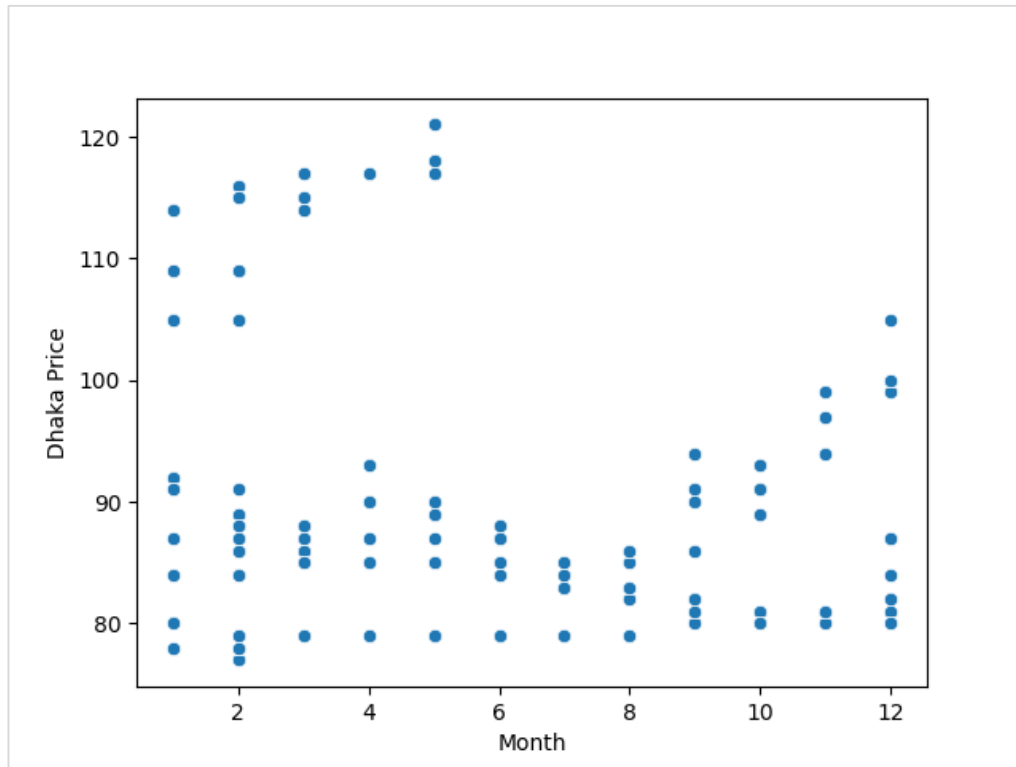


Figure 3.4: Monthly price analysis

### 3.4.4 Minimum maximum price range

A box and whisker plot, often known as a box plot, depicts a five-number summary of data. The five numbers that comprise the five-number summary are the lowest, first quartile, middle, lower quartile, and optimum. In a box plot, we draw a box from the first to the third quartile. This figure represents the box plot of Dhaka price. From this graph we can see that the minimum price and maximum price and the average price of Dhaka. From this graph we can see that the minimum price is less than 80 the maximum price is 100 and the average price rate that means the price are frequently come in between range 80-90.

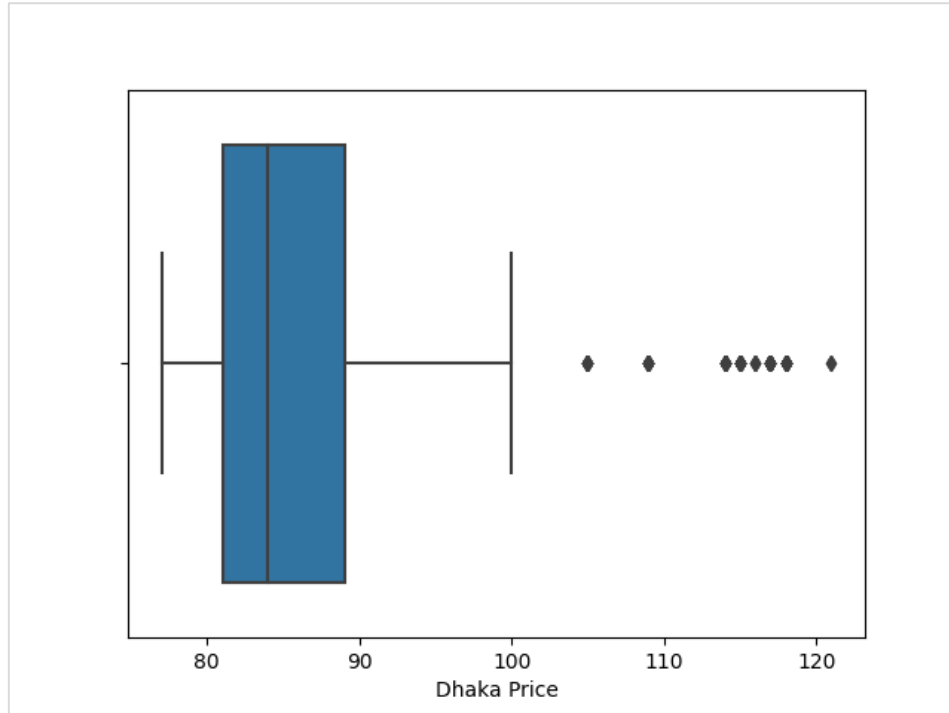


Figure 3.5: Box Plot of Dhaka price

### 3.4.5 KDE price analysis

A kernel density estimate (KDE) plot, like a histogram, appears the conveyance of perceptions in a dataset. KDE employs a ceaseless likelihood thickness bend to speak to information in one or more measurements. KDE Plot depicted as Part Thickness Appraise is utilized for visualizing the Likelihood Thickness of a nonstop variable. It portrays the likelihood thickness at distinctive values in a ceaseless variable. Figure 3.6 represents that the KDE plot of our dataset.

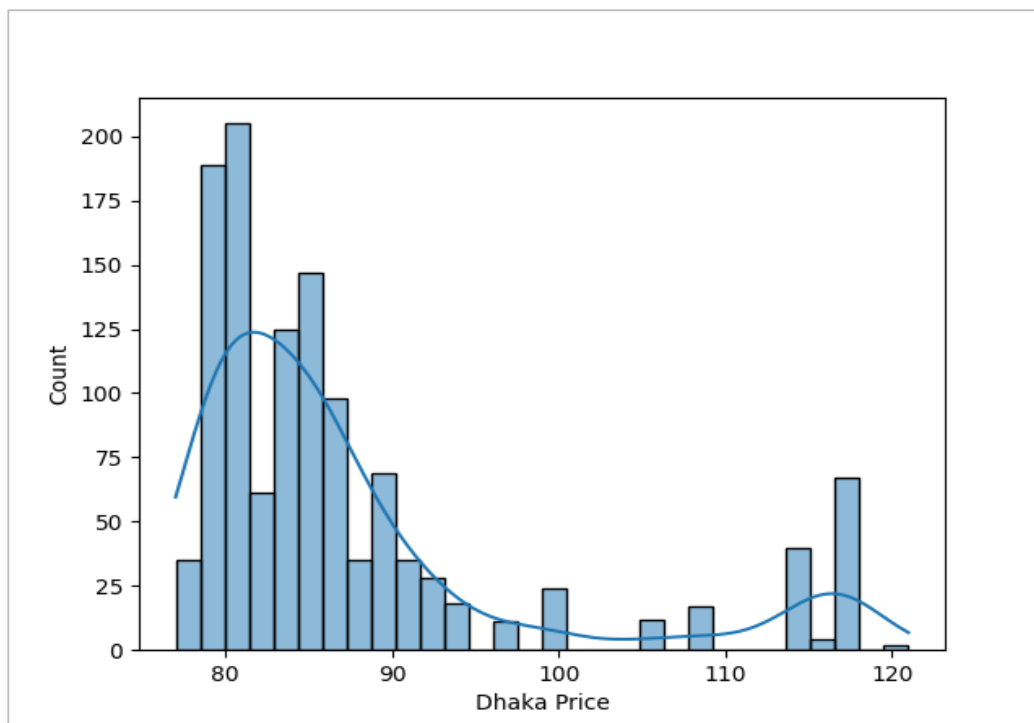


Figure 3.6: KDE plot of dataset.

### 3.5 Features

We create the work's characteristics after studying the data.

Table 3.1: Feature Descriptions

SL No.	Name of Attributes	Description
1	Year	The year characteristic is used to estimate pricing more precisely.
2	Day	We consider the current onion price to be the most important factor.
3	Month	Price changes throughout the course of a month are also considered a characteristic.
4	Season	Because price varies by season and cannot be avoided, we evaluated season as another primary attribute.
5	Location	The cost is also affected by where you live. The cost of an onion varies depending on where you live. As a result, we use location as an attribute.



6	Price	Our desired attribute is price. Our key focus is forecasting the price based on the year, day, and month.
---	-------	---

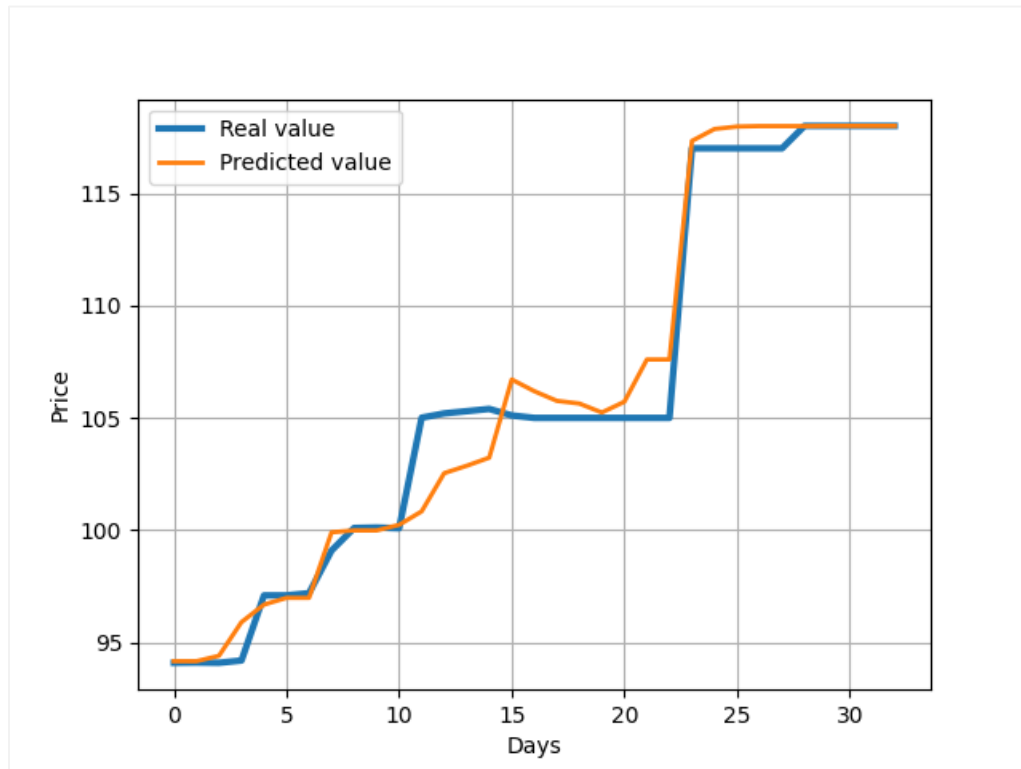
### 3.6 Algorithm Implementation

In this section, we discussed the algorithm implementation process. To finish this step, we must first complete the preceding process in order to create the appropriate dataset. When the dataset was ready, we divided the price into three categories: high, mid, and low. The algorithm is then put into action. Because our job is to anticipate the onion price, we aim to find the best strategies for doing so. We discovered that classification methods were the best fit for our needs. We gradually implemented five different ML algorithms named KNN, Nave Bayes, Decision Tree, SVM, and Neural Network and measured and compared the results to discover the ideal algorithm to forecast the price more efficiently. The results are shown in table 3.2, along with the accuracy rate. NN had the highest accuracy of all of them.

Table 3.2 Parameter usages

Algorithms	Details
Gradient Boosting	n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0
Decision Tree Regression	random_state= 42
Lasso Regression	Alpha = 0.1
Linear Regression	Karnel = 'linear'
MLP Regression	random_state=1, max_iter=500
Random Forest	max_depth=2, random_state=0

Table 3.3 depicts the parameters and various things that we used to implement the selected



algorithms.

### 3.7 Evaluation

Figure 3.7: Comparison of real and predicted price

Figure 4 represent the prediction evaluation of our work. For testing our model, we have collected 35 days of daily data this 35 data are never used in our model. Blue color represents the real daily price and orange color represents the predicted price. From this graph we can see that our predicted result is nearly overlapped our actual price. That means our model perform very good for real and unseen data.

# CHAPTER 4

## RESULT ANALYSIS

### 4.1 Introduction

This chapter 4 mainly focuses on the descriptive analysis of the data used in the research as well as the experimental results of our research. When we analysis it the question at first comes to our mind is what is result analysis? The consequence section should be programmed to narrate discoveries without having to understand or analyze them and should also provide guidelines for the study paper discussion section. The observations are documented and the analysis is revealed. This analysis portion examines what has been achieved in the results.

### 4.2 Experimental Result

Table 4.1: R2 Score

Data usage rate	Algorithms					
	Gradient Boosting	Lasso Regression	MLP Regression	Random Forest	Decision Tree	Linear Regression
30%	99.36%	54.79%	48.37%	<b>99.59%</b>	99.53%	57.25%
40%	98.62%	55.17%	49.31%	<b>98.76%</b>	98.34%	57.57%
50%	98.66%	55.18%	48.13%	<b>98.78%</b>	98.53%	56.67%
60%	98.71%	54.12%	47.14%	<b>98.89%</b>	98.61%	55.27%
70%	98.70%	53.96%	46.94%	<b>98.77%</b>	98.33%	54.72%

Table 4.1 represents the r2 sore of each of each used algorithms we used 30%-70% test data usage rate when test data is 30% training automatically set as 70% same way when test data is 70% then training will be 30%. We used this techniques to find out which percentage is better for our model. We found best r2 score 99.59% by random forest algorithm.

Table 4.2: Mean absolute error

Data usage rate	Algorithms					
	Gradient Boosting	Lasso Regression	MLP Regression	Random Forest	Decision Tree	Linear Regression
30%	0.52	6.33	6.23	<b>0.17</b>	0.23	6.46
40%	0.76	6.26	6.07	0.34	0.30	6.43
50%	0.71	6.17	5.91	0.34	0.25	6.39
60%	0.71	6.21	5.89	0.36	0.27	6.40
70%	0.75	6.21	5.92	0.45	0.37	6.41

Table 4.2 represents the mean absolute error of each algorithm. The less mean absolute error is achieved by random forest algorithm and the error rate is 0.23 only. And the highest error rate is achieved by linear regression algorithm.

Table 4.3: Mean squared

Data usage rate	Algorithms					
	Gradient Boosting	Lasso Regression	MLP Regression	Random Forest	Decision Tree	Linear Regression
30%	0.60	57.78	65.98	<b>0.52</b>	0.59	54.63
40%	1.75	57.40	64.91	1.57	2.12	54.33
50%	1.62	54.25	62.79	1.47	1.76	52.46
60%	1.50	53.57	61.72	1.29	1.62	52.22
70%	1.47	52.62	60.65	1.39	1.38	51.74

Table 4.3 represents the mean squared error. Form this table we can see that the less error is produced by random forest algorithm and the error rate is 0.52 using 30% test data usage rate.

Table 4.4: Root Mean squared

Data usage rate	Algorithms					
	Gradient Boosting	Lasso Regression	MLP Regression	Random Forest	Decision Tree	Linear Regression
30%	0.77	7.60	8.12	<b>0.72</b>	0.77	7.39
40%	1.32	7.57	8.05	1.25	1.45	7.37
50%	1.27	7.36	7.92	1.21	1.33	7.24
60%	1.22	7.31	7.78	1.13	1.27	7.22
70%	1.21	7.25	7.78	1.18	1.38	7.19

Table 4.4 represents the mean squared error. Form this table we can see that the less error is produced by random forest algorithm and the error rate is 0.72 using 30% test data usage rate.

### 4.2.1 Gradient Boosting

The gradient boosting technique is one of the most powerful strategies in machine learning. Machine learning algorithm errors, as we all know, are broadly classified into two types: bias errors and variance errors. Gradient boosting is one of the boosting procedures used to reduce the model's bias error. The gradient boosting procedure, unlike the Adaboosting approach, does not allow us to specify the base estimator. The base estimator of the Gradient Boost algorithm is fixed, i.e. Decision Stump. Using AdaBoost, for example, we can modify the n estimator of the gradient boosting algorithm. We get very expected result like r2 is 99.36% using 30% test data. By same data uses rate we found 0.60 mean squared error, 0.52 means absolute error. And 0.77 root mean squared error.

### **4.2.2 Lasso Regression**

Chris Hans et. al. [11] identified the lasso estimate corresponds to a posterior mode when distinct, double-exponential prior distributions are applied to the regression coefficient. Lasso regression is a sort of shrinkage-based linear regression. Data values are shrunk towards a central point, such as the mean, in shrinkage. Simple, sparse models are encouraged by the lasso approach. This form of regression is ideal for models with a lot of multi collinearity or when you wish to automate elements of the model selection process, such as variable selection and parameter removal. We get result like  $r^2$  is 48.37% using 30% test data. By same data uses rate we found 65.98 mean squared error, 6.23 means absolute error. And 8.12 root mean squared error.

### **4.2.3 MLP Regression**

J. Rynkiewicz et al. [12] describe Multilayer perceptron's (MLP) with one covered up layer have been utilized for a long time to bargain with non-linear relapse. In any case, in a few assignments, MLP's are as well capable models and a little cruel square blunder (MSE) may be more due to overfitting than to real modeling. Multilayer Perceptron is commonly utilized in straightforward relapse issues. In any case, MLPs are not perfect for preparing designs with consecutive and multidimensional information. We get result like  $r^2$  is 54.97% using 40% test data. By same data uses rate we found 57.78 mean squared error, 6.33 means absolute error. And 7.60 root mean squared error.

### **4.2.4 Random Forest**

F Livingston [11] A classical machine learner is created by collecting tests of information to speak to the whole populace. This information set is ordinarily subdivided into two or more datasets. Portion of the dataset set is commonly utilized for creating the machine learner, and the remaining information is utilized for assessment. Frequently this information set is imbalanced; the information comprises of as it were a very little minority of the information. We get result like  $r^2$  is 99.59% using 30% test data. By same data uses rate we found 0.52 mean squared error, 0.23 means absolute error. And 0.72 root mean squared error. This is the highest

rated algorithm among all 6 algorithm. Our highest r2 score is achieved by random forest algorithm. so we decided to use this algorithm for prediction.

#### **4.2.5 Decision Tree**

Freund et al. [13] explained a modern sort of classification run the show, the rotating choice tree, which may be a generalization of choice trees, voted choice trees and voted choice stumps. At the same time classier of this sort are generally simple to translate. They show a learning calculation for rotating choice trees that are based on boosting. Test comes about appear it is competitive with boosted choice tree calculations. Choice tree is the foremost capable and well known apparatus for classification and prediction. A Decision tree could be a flowchart like tree structure, where each inside hub indicates a test on a quality. We get result like r2 is 99.53% using 30% test data. By same data uses rate we found 0.59 mean squared error, 0.17 means absolute error. And 0.77 root mean squared error.

#### **4.2.6 Linear Regression algorithm**

Direct Relapse could be a machine learning calculation based on directed learning. It performs a relapse errand. Relapse models a target forecast esteem based on free factors. It is for the most part utilized for finding out the relationship between factors and estimating. Distinctive relapse models vary based on – the kind of relationship between subordinate and autonomous factors, they are considering and the number of autonomous factors being utilized. We get result like r2 is 5.57% using 40% test data. By same data uses rate we found 54.33 mean squared error, 6.43 means absolute error. And 7.37 root mean squared error.

## **CHAPTER 5**

### **SUMMARY, CONCLUSION AND FUTURE WORK**

#### **5.1 Summary of the Study**

Although there has been a large amount of study done in the field of machine learning, the volume of such research effort in Bangladesh is rather low. While predictive style work is a common term for machine learning, Bangladeshi goods are still unfamiliar with it. This type of research has recently been used when the outcome of such a task produces a significant change in our digital lives. As a result of this type of research, we receive some remarkable real-world applications. However, little research is being conducted in the realm of Bangladesh's economy. We trust, however, that a number of researchers from many countries have performed research in this topic. In our scientific capacities.

#### **5.2 Conclusion**

In this paper, we investigated the performance of various machine-learning algorithms for predicting soybean oil prices. We've been using 1357 data points to forecast soybean oil prices. We employed the Decision tree, gradient boosting, lass regression, linear regression, MLP regression, Random Forest Regression. techniques to build that model. Random Forest Regression (RFR) outperformed all other algorithms. In any event, it has a precision of roughly 99 percent. As a result, we used the Random Forest forecasting model to forecast soybean oil prices. Forecasting future pricing is a more effective way of reducing market imbalance. This endeavor will aid the government in taking actions to maintain market stability.



### **5.3 Recommendations**

Here are a few notable recommendations:

- ❖ To improve the study's outcomes by increasing the accuracy of data collecting.
- ❖ The use of Deep Learning would be beneficial.
- ❖ Improved data will also result in a better data gathering outcome.
- ❖ Use a more complicated algorithm, such as LSTM, to increase the precision of our dataset.

### **5.4 Future Work**

The future guidance on the development of this work is given bellow:

- In the future, we will explore at various optimization algorithms in our study.
- In the future, we will create an intelligence system based on deep learning techniques.
- We will utilize a huge dataset in the future to increase accuracy.
- In future, we will be working on a Web-based API to achieve this goal.

## REFERENCE

- [1] Manjula, K. A., & Karthikeyan, P. (2019, April). Gold Price Prediction using Ensemble based Machine Learning Techniques. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1360-1364). IEEE.
- [2] M. M. Hasan, M. T. Zahara, M. M. Sykot, A. U. Nur, M. Saifuzzaman and R. Hafiz, "Ascertaining the Fluctuation of Rice Price in Bangladesh Using Machine Learning Approach," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5
- [3] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: A comparative study. In 2019 International Conference on Smart Structures and Systems (ICSSS) (pp. 1-5). IEEE.
- [4] Hasan, M. M., Zahara, M. T., Sykot, M. M., Hafiz, R., & Saifuzzaman, M. (2020, July). Solving Onion Market Instability by Forecasting Onion Price Using Machine Learning Approach. In 2020 International Conference on Computational Performance Evaluation (ComPE) (pp. 777-780). IEEE.
- [5] F. A. de Oliveira, L. E. Zárata, M. de Azevedo Reis, and C. N. Nobre, "The use of artificial neural networks in the analysis and prediction of stock prices," in 2011 IEEE International Conference on Systems, Man, and Cybernetics, 2011, pp. 2151-2155: IEEE.
- [6] Salis, V. E., Kumari, A., & Singh, A. (2019). Prediction of gold stock market using hybrid approach. In Emerging Research in Electronics, Computer Science and Technology (pp. 803-812). Springer, Singapore.
- [7] Salis, V. E., Kumari, A., & Singh, A. (2019). Prediction of gold stock market using hybrid approach. In Emerging Research in Electronics, Computer Science and Technology (pp. 803-812). Springer, Singapore.
- [8] M. A. Mithu, K. M. Rahman, R. A. Razu, M. Riajuliislam, S. I. Momo and A. Sattar, "Gold Price Forecasting using Regression Techniques for Settling Economic and Stock Market Inconsistency," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-4, doi: 10.1109/ICCCNT51525.2021.9579755.
- [9] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016, pp. 1-5: IEEE
- [10] Jerome H. Friedman, Stochastic gradient boosting, Computational Statistics & Data Analysis, Volume 38, Issue 4, 2002, Pages 367-378, ISSN 0167-9473.
- [11] Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. ECE591Q Machine Learning Journal Paper, 1-13.
- [12] J. Rynkiewicz, General bound of overfitting for MLP regression models, Neurocomputing, Volume 90, 2012, Pages 106-110, ISSN 0925-2312.

- [13] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *icml* (Vol. 99, pp. 124-133).
- [14] Späth, H. Algorithm 39 Clusterwise linear regression. *Computing* 22, 367–373 (1979).

## **APPENDIX**

The primary was to characterize the methodological strategy for our study to carry out the inquire about which we confronted so numerous deterrents. Besides, not much work has been done sometime recently in this locale. This was not ordinary work. So from some place we couldn't get as well much back. The collection of information was another deterrent, and for us, this was a tremendous challenge. We too started physically gathering information. In comparison, categorizing the numerous posts is another issue. We may do it after a long time of difficult work.