# SOIL SUITABILITY
# FOR CROP CULTIVATION USING MACHINE LEARNING: A CASE STUDY OF BOGURA REGION

A thesis report submitted in partial fulfillment of the requirements
for the Degree of
Bachelor of Science [B.Sc]

Presented By

**Md. Abdul Alim**
**0322210105101132**
**$20^{th}$ Batch**

**Department of Computer Science and Engineering**
**Pundra University of Science & Technology**
**RangpurRoad, Gokul, Bogura, Bangladesh**

**January, 2026**

# SOIL SUITABILITY
# FOR CROP CULTIVATION USING MACHINE LEARNING: A CASE STUDY OF BOGURA REGION

Presented By

**Md. Abdul Alim**
**0322210105101132**
**20th Batch**

**Approved as to style and content by**

**........................................................**

**Mst. Rehana Khatun**

**Assistant Professor**
**Computer Science & Engineering**

(Supervisor)

**Department of Computer Science and Engineering**
**Pundra University of Science & Technology**
**RangpurRoad, Gokul, Bogura, Bangladesh**

**January, 2026**

# Declaration

I, Md. Abdul Alim, can declare and testify that the study presented in this thesis, i.e. soil suitability for crop cultivation using machine learning: a case study of bogura region, is the result of my personal original research, which was conducted under the supervision of Mst. Rehana Khatun. This thesis has not been submitted in full, or even in fractions, to any other university or institution to get a degree or diploma. Citations of all information and support materials utilized in this study have been given adequately. I testify that this is not a plagiarized piece of work in regard to the academic integrity rules.

Date : 01 January, 2026
ID : 0322210105101132
Computer Science & Engineering

…………………………….
[Signature]

# Acknowledgement

I would say that I have deep gratitude to my supervisor Mst. Rehana Khatun, Asst. Professor, Department of Computer Science and Engineering, due to her unshaken support, informed suggestions and valuable proposals, in the course of writing this thesis. It has been through their encouragement and their illuminating criticism that the direction of this study and its outcomes have been decided. I do appreciate the dedication and encouragement of Computer Science and engineering (CSE) faculty and staff in ensuring the provision of the required resources and creation of the conducive academic environment.. I would like to retain my friends and classmates who helped and supported me. My parents and my family remain the greatest support to me in this journey and whose love, support, and prayers have been unending.

Md. Abdul Alim

01 January, 2026

# ABSTRACT

The study of soil suitability is a very important aspect towards enhancing output in agricultural activities and sustainable crop production. Crop choice in Bangladesh is usually determined by traditional knowledge, and not by scientific consideration of the physical properties of soils, which can contribute to inefficient use of land and low productivity. This paper outlines a machine learning tool to evaluate the suitability of soil in the Bogura Bangladesh area where crops are going to be grown. The suggested system determines the soil suitability of the four major crops of rice, maize, banana and papaya based on some major parameters of soil such as the pH value, organic matter (OM), nitrogen (N), phosphorus (P), potassium (K), soil texture, drainage, rainfall, humidity, temperature and sunlight exposure etc. To classify soil suitability using these parameters, four machine learning algorithms, Decision Tree, Random Forest, Gaussian Naive Bayes, and Logistic Regression are supervised. Model performance is measured using standard metric such as accuracy, precision, recall and F1 score. Experimental results show that the proposed approach is suitable for classifying the soil suitability into several categories. Out of the tested models, the Decision Tree classifier proves to be more effective, and it is appreciated in its capacity to extract non-linear relationships between the soil characteristics and the crop requirements. The results indicate that machine learning approaches are efficient in the delivery of interpretable and precise soil suitability classification. The research will be helpful in making agriculture decisions based on data and also help to provide a practical framework to select the right crops to maximize productivity, resource use, and sustainable agriculture in the Bogura region. The suggested solution can assist farmers, researchers, and policymakers in embracing smart devices in the determination of suitability of crops based on the soil and could be applied to other farms that have environment-related features.

**Keywords**: Soil Suitability, Crop Cultivation, machine learning, Decision tree, Agriculture, Random Forest, Bogura region.

# CONTENTS

1.Chapter 1 : Introduction ........ 1
1.1 Background of the study ........ 1
1.2 Agriculture Scenario in Bangladesh ........ 2
1.3 Soil Characteristics of the Bogura Region ........ 2
1.4 Motivation of the Research ........ 5
1.5 Problem Statement ........ 6
1.6 Research Objectives ........ 6
1.7 Scope of the Study ........ 7
2.Chapter 2 : Literature Review ........ 8
2.1 Soil Suitability Analysis in Agriculture ........ 8
2.2 Application of Machine Learning in Agriculture ........ 9
2.3 Machine Learning-Based Crop Recommendation Systems ........ 10
2.4 Soil Parameter-Based Classification Using Machine Learning ........ 11
2.5 Limitations of Existing Studies and Research Gap ........ 13
3.Chapter 3 : Methodology ........ 14
3.1 Overall System Workflow ........ 14
3.2 Study Area Description ........ 14
3.2.1 General ........ 15
3.2.2 Geomorphology ........ 15
3.3 Dataset Description ........ 16
3.3.1 Data Source ........ 16
3.3.2 Input Features ........ 17
3.3.3 Target Variable and Class Labels ........ 17
3.3.4 Dataset Size and Discribution ........ 18
3.4 Dataset Preprocessing ........ 19
3.4.1 Handling Missing and Inconsistent Values ........ 19
3.4.2 Encoding of Categorical Features ........ 19
3.4.3 Feature Scaling and Normalizatio ........ 19
3.4.4 Outliers Detection ........ 20
3.4.5 Final Dataset Preparation ........ 20
3.5 Feature Selection ........ 20
3.6 Machine Learning Algorithm Used ........ 21
3.6.1 Decision Tree ........ 21

3.6.2 Random Forest ........ 21
3.6.3 Logistic Regression ........ 22
3.6.4 Gaussian Naïve Bayes ........ 22
3.7 Model Training and Testing ........ 22
3.8 Performance Metrics ........ 23
3.8.1 Accuracy ........ 23
3.8.2 Precision ........ 24
3.8.3 Recall ........ 24
3.8.4 F1-Score ........ 25
4.Chapter 4 : Results & Discussion ........ 26
4.1 Introduction ........ 26
4.2 Exploratory Data Analysis (EDA) ........ 26
4.2.1 Class Distribution Analysis ........ 27
4.2.2 Feature Distribution Analysis ........ 28
4.3 Model-Wise Comparison Performance ........ 33
4.3.1 Accuracy Comparison for Rice ........ 33
4.3.2 Accuracy Comparison for Banana ........ 34
4.3.3 Accuracy Comparison for Papaya ........ 34
4.3.4 Accuracy Comparison for Maize ........ 34
4.4 Graphs and Charts ........ 35
4.4.1 Accuracy Comparison Graph (Rice) ........ 35
4.4.2 Accuracy Comparison Graph (Banana) ........ 36
4.4.3 Accuracy Comparison Graph (Papaya) ........ 36
4.4.4 Accuracy Comparison Graph (Maize) ........ 37
4.5 Precision, Recall and F1-Score Comparison ........ 38
4.5.1 Precision, Recall and F1-Score (Rice) ........ 38
4.5.2 Precision, Recall and F1-Score (Banana) ........ 39
4.5.3 Precision, Recall and F1-Score (Papaya) ........ 39
4.5.3 Precision, Recall and F1-Score (Maize) ........ 40
4.6 Confusion Matrix Visualization ........ 40
4.6.1 Confusion Matrix Visualization (Rice) ........ 41
4.6.2 Confusion Matrix Visualization (Banana) ........ 41
4.6.3 Confusion Matrix Visualization (Papaya) ........ 42
4.6.4 Confusion Matrix Visualization (Maize) ........ 42
4.7 Discussion of Graphical Results ........ 43
4.8 Best Model Selection ........ 43
4.9 Interpretation of Results ........ 44
4.10 Comparison with Previous Studies ........ 44
4.11 Practical Implications of Farmers ........ 45

Chapter 5 : Conclusion & Future Work ........ 46
5.1 Summary of Finding ........ 46
5.2 Conclusion ........ 46
5.3 Limitations of my Works ........ 46
5.4 Future Research Directions ........ 47

# FIGURES

Figure 1.1: A map of the location showing the study area ..... 3
Figure 1.2: Composition of soil ..... 4
Figure 1.3 : Types of soil ..... 4
Figure 3.1 : Workflow of the system ..... 14
Figure 3.2 : Surface geology of Bogura created by FAO ..... 15
Figure 4.1 : Soil suitability classes for Banana ..... 27
Figure 4.2 : Soil suitability classes for Papaya ..... 27
Figure 4.3 : Soil suitability classes for Rice ..... 27
Figure 4.4 : Soil suitability classes for Maize ..... 27
Figure 4.5: Boxplot and histogram of the distribution of soil pH ..... 28
Figure 4.6: Boxplot and histogram of the distribution of soil nitrogen content ..... 29
Figure 4.7: Boxplot and histogram showing the distribution of phosphorus concentration in soil ..... 29
Figure 4.8 : Boxplot and histogram showing the distribution of potassium concentration in soil. ..... 30
Figure 4.9 : Boxplot and histogram showing the distribution of organic matter ..... 30
Figure 4.10 : Boxplot and histogram showing the distribution of humidity levels. ..... 31
Figure 4.11 : Boxplot and histogram showing the distribution of rainfall values. ..... 31
Figure 4.12 : Distribution of drainage in the dataset ..... 32
Figure 4.13 : Distribution of soil texture in the dataset ..... 33
Figure 4.14 : Accuracy comparison of machine learning models for Rice cultivation ..... 35
Figure 4.15 : Accuracy comparison of machine learning models for Banana cultivation ..... 36
Figure 4.16 : Accuracy comparison of machine learning models for Papaya cultivation ..... 37
Figure 4.17 : Accuracy comparison of machine learning models for Maize cultivation ..... 37
Figure 4.18 : Precision, recall, and F1-score comparison of machine learning models for rice ..... 38
Figure 4.19 : Precision, recall, and F1-score comparison of machine learning models for banana ..... 39
Figure 4.20 : Precision, recall, and F1-score comparison of machine learning models for papaya ..... 39
Figure 4.21 : Precision, recall, and F1-score comparison of machine learning models for maize ..... 40
Figure 4.22 : Confusion matrix of the best-performing machine learning model for soil suitability classification ..... 41
Figure 4.23 : Confusion matrix of the best-performing machine learning model for soil suitability classification ..... 41
Figure 4.24 : Confusion matrix of the best-performing machine learning model for soil suitability classification ..... 42
Figure 4.25 : Confusion matrix of the best-performing machine learning model for soil suitability classification ..... 42

# Tables

Table 3.1 : Description of soil and environment input features ............ 17
Table 3.2 : Selected crops for soil suitability classification ............ 17
Table 3.3 : showing the target variable (crop suitability) ............ 18
Table 4.1 : Showing the accuracy comparison for rice crop ............ 33
Table 4.2 : Showing the accuracy comparison for banana crop ............ 34
Table 4.3 : Showing the accuracy comparison for papaya crop ............ 34
Table 4.4 : Showing the accuracy comparison for maize crop ............ 34

# 1. Chapter 1 : Introduction

## 1.1 Background of the study

Bangladesh is one of the least developed countries in the South Asian region which is growing economically at a rapid pace. Agriculture and food security have advanced remarkably in the country despite the fact that it is a populated and prone to climate based disasters. Population is around 166.7 million with population density of 1278/km2, which hails it as the eighth most populated country in the world (WPR, 2021) [1]. One of the key aspects of sustainable agriculture is the soil health that is the key to the sustainability of agricultural productivity, ecological equilibrium, and climate. This is significant as the quality of soil directly affects food security, environmental sustainability and livelihoods of billions of individuals in the entire world. There are currently only a few arable land areas of approximately 1.4 billion hectares (or approximately 10 percent of the total land mass in the world) and approximately 33 percent that is already degraded by unsustainable land management practices, excessive use of chemical inputs, and climate change (Food and Agriculture Organization (FAO), 2023) [2]. The world largely depends on agriculture as the basic source of food. Farming is influenced by a number of factors such as pH value, nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, soil texture, etc [3]. The conventional soil analysis methods though useful when dealing with very small sets of data, can be laborious and involve lab tests and are thus time consuming and economically impractical when dealing with big agricultural decisions [4]. The recent technological advancements have provided novel approaches to soil classification that include remote sensing, machine learning and artificial intelligence. These new methods facilitate the speedy examination of bigger areas of land beyond improving the precision as well as effectiveness of classifying soils. Machine learning methods allow analyzing complex and large soil data quickly and have proven to be more efficient and predictive than traditional methods of soil classification [5].

## 1.2 Agriculture Scenario in Bangladesh

Agriculture is the biggest sector of employment in Bangladesh and contributes 14.2% of Bangladesh's Gross Domestic Product (GDP) in 2017 and 20% GDP in 2025, providing employment to approximately 42.7% of the workforce. As of the financial year "2022 to 2023" the agriculture sector accounted for more than 12% of GDP contribution to it. The performance of this sector has an overwhelming effect on major macroeconomic objectives such as employment generation, poverty alleviation, development of human resources, food security and other economic and social forces [6]. By the year 2025 AD, the population will grow to around 198 million (en.wikipedia.org/ wiki/world population). Total cultivable land of the country is about 8.44 million hectare. Demographic pressures and greater urbanization led to a decline in cultivated area at the rate of 1 percent per annum. Food requirement of the country is estimated to be doubled in the next 25 years (Islam and Haq 1999). The demand has to be met from our scarce and shrinking land resources [7]. Despite the rise in crop production, the farmers are finding it difficult to benefit from their work because of the shortage of agricultural labour, climate change, poor management of markets, and limited use of modern technology. Challenges such as lower productivity than neighbouring countries, the decline of farmland and lack of agricultural export capacity have further worsened the situation [8]. The barriers are high costs, low levels of digital literacy and gaps on the policy level. Hence, there is an urgent need to discover viable agri-technologies that are easy to replicate and obtain with respect to the localities. The proposed research will help in filling this research gap by looking at the latest development in the world and discussing the ways in which this technology can be applied to the context of Bangladesh agriculture. By harnessing the power of modern technology and combining it with classical farming knowledge, Bangladesh can develop a strong agricultural sector which will ensure food security, support rural livelihoods and ensure long term environmental sustainability [9].

## 1.3 Soil Characteristics of the Bogura Region

Bogura is one such agricultural district in the Northern region of Bangladesh and is most well known in terms of crop production. It is a crucial district in the provision of food security in Bangladesh because of the good agro-climatic factor. The region is blessed with sufficient rainfall, well developed irrigation systems, and availability of major rivers, which all contribute to boosting agricultural production.

The farmers of Bogura produce both old and new varieties of crops, which are useful in terms of regional and national food reserves. The country's agricultural regions differ in cropping pattern with the district of Bogura being a major agricultural district with a dual characteristic of economic dependence on agriculture as well as drought vulnerability. Cropping intensity, which is the ratio of gross cropped area and net sown area, is an important agricultural productivity indicator. While the national average cropping intensity is 191%, for Bogura, it is 234% indicating the efforts of this region in crop diversification and multiple cropping patterns [10].

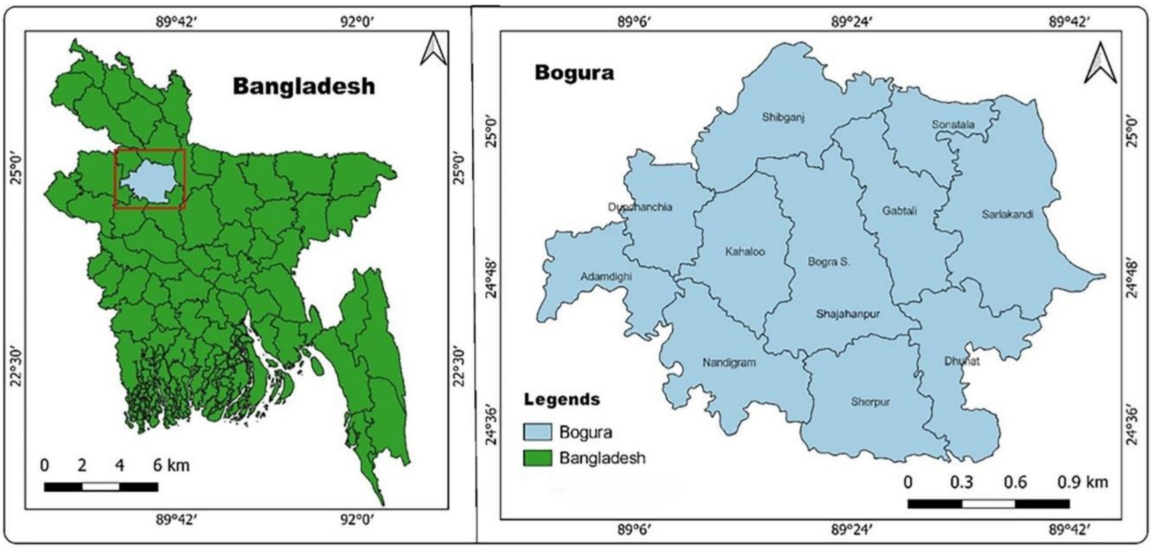


Figure 1.1: A map of the location showing the study area

**Properties of soil :**

Soil is a very complex system of a nature-based origin, consisting of air, water, mineral and organic matter. It is made up of abiotic components (minerals, air and water) and biotic components (living organisms). Generally, minerals constitute about 40-45% of the volume of soil, air and water 25% each and living organisms about 5%. Soil composition changes depending on location based on parent material, climate, vegetation, compaction, and organic matter (humus) [11].

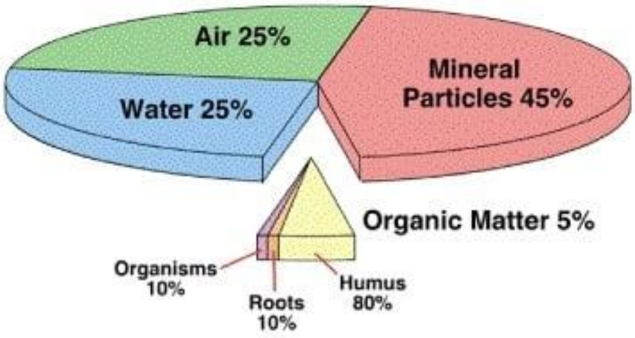


Figure 1.2: Composition of soil

The properties of soil depend on different amounts of biotic and abiotic components depending on the composition of the soil. Combinations of these components define the physical and chemical properties of soil. The major properties of soil such as pH level, nitrogen (N), phosphorus (P), potassium (K), organic matter (OM), soil texture, rainfall, humidity, temperature etc are important and highly used in deciding what crop to grow and the capacity of the soil to produce. The differences in these parameters may directly influence the availability of nutrients, development of roots, and total crop development.

**Types of soil :**

**Sandy soil** : It consists of large coarse particles that drain water very quickly, resulting in dry soil that has few nutrients. Although it is easy to work with and warms up easily in the spring time, lack of moisture and nutrients can be a limiting factor to plant growth.

**Clay soil** : It has very fine particles which are good at trapping water and nutrients but poorly drained and aerated. Although it helps in crops such as rice, wheat and cotton, its high density and compactness may limit the growth of roots and may cause cultivation to be problematic.


Figure 1.3 : Types of soil

**Silt soil** : The particles that constitute silt soil are fine and smooth and medium, which is more efficient in holding water and nutrients, unlike sandy soil, and very productive in agricultural activities. It however can shrink when wet resulting to poor aeration and drainage that could limit the growth of the plants.

**Loam soil** : Sandy, fine grains of soils with clay form a balanced soil, which is loam soil with good drainage and good moisture and nutrient storage. The well-structured and easy to work fertile nature of it has been ideal to most crops and has been popular among gardeners and farmers.

## 1.4 Motivation of the Research

In Bangladesh, the agricultural productivity is greatly reliant on the correct selection of crop based on the conditions of the soil. Nevertheless, in some places like Bogura agricultural choices are often supported by customary experience but not by scientific analysis of soil. Although these in the past have been known to sustain farming, they do not always suffice to deal with soil erosion, changes in climate as well as resource scarcity. The technology in agriculture has revolutionized agriculture to a high point of surpassing the current methods of agricultural practices in terms of productivity, efficiency and durability by the use of modern technology in agriculture. Traditional agricultural practices require physical labor, rudimentary machinery and natural cycle which is not as efficient and may not be easy to scale. The modern technology assists farmers in the adaption of climate change, including weather forecast and soil monitoring system which contributes in minimizing risk [12]. Bogura region possesses various types of soil and produces food including rice, maize, banana, and papaya hence it can be used in analyzing soil suitability using machine learning. But despite the significance of agriculture in this region, sophisticated equipment is not very common to assist farmers to select crops according to the soil characteristics. This disparity between the technology that has been made available and its practical application in agriculture is the driving force behind this study. Thus, this work will create a simple and stable machine learning-based tool to determine the suitability of soils and contribute to more successful crop choice. The study aims to enhance the agriculture decision making process, land utilization and sustainable agriculture in the Bogura region by comparing the various machine learning models.

## 1.5 Problem Statement

In Bangladesh, the success of agriculture is very reliant on improperly selected crops depending on the nature of the soil. But in most places, such as the Bogura district, the choice to cultivate a crop is still strongly influenced by traditional knowledge and experience of the farmers and not by a systematic study of the soil characteristics. Mismatches between soil properties and crop demands usually occur in the practice, which in turn causes decreased yield, unproductive application of fertilizers and water, as well as degraded soils in the long-term. The soil testing facilities and agricultural advisory services are not always available and affordable to the small-scale farmers even though they are available. In addition, traditional soil evaluation methodologies are time consuming and are also unable to analyze large and multi dimensional data. Consequently, the farmers do not have access to effective and timely decision-support tools to help them in selecting appropriate crops depending on the quality of the soil. Simultaneously, the developments in machine learning and data collection have enabled the opportunity to process sophisticated data in agriculture and make precise forecasts. Nevertheless, machine learning-based soil suitability assessment is not utilized at the regional level in Bangladesh. The modern computational techniques are evidently not well matched with their application in assisting more practical decisions in soil-based crop cultivation. Thus, the main issue that is going to be solved in the study is that the region has no efficient, data-driven and region-specific system to evaluate the soil suitability and suggest the crops to grow in the Bogura area.

## 1.6 Research Objectives

The main aim of the study is to come up with a machine learning-based model of soil suitability analysis to aid in the decision of crop planting in the Bogura district of Bangladesh.

The precise goals of the given research are as follows:

- To measure the major parameters of soil that can be used to cultivate crops such as the pH, nitrogen, phosphorus, potassium, organic matter (OM), soil texture, rainfall, humidity, temperature etc.
- To train machine learning to predict the suitability of soil to various crops (rice, maize, banana, and papaya).
- To apply and compare the performance of the Decision Tree, random forest, Gaussian Naive Bayes, and the Logistic regression algorithms.

- To assess the model performance in terms of normal metrics like accuracy, precision, recall, and F1-score.
- To determine the best machine learning model to use in the classification of soil suitability.
- To offer a decision-support methodology which can be used to help farmers and other agricultural stakeholders choose appropriate crops depending on the condition of the soil.

## 1.7 Scope of the Study

The paper is aimed at evaluating soil openness to crop farming in the Bogura, Bangladesh area through machine learning algorithms. The analysis of the chosen soil parameters, such as the pH value, nitrogen (N), phosphorus (P), potassium (K), and moisture content, is the scope of the research, as the given parameters can be regarded as the crucial indicators of the soil fertility and crop development. The paper takes into account four significant crops such as rice, maize, banana, and papaya, depending on their agricultural significance in the Bogura area. Decision Tree, Random Forest, Gaussian Naive Bayes, and Logistic Regression are supervised machine learning algorithms that classify the soil suitability. The assessment of the model is carried out in the conventional performance indicators: accuracy, precision, recall, and F1-score. This study is supposed to offer a decision-making system of the soil-based crop choice. The study however fails to consider the economic factors, pest and disease conditions and long term climatic projections. The results are region-specific to the Bogura area, and might need more modification before they are implemented in other areas that act differently due to the soil and environment.

# 2. Chapter 2 : Literature Review

## 2.1 Soil Suitability Analysis in Agriculture

Previous studies have shown that rapid population growth and inappropriate agricultural practices are among the main causes of soil degradation, putting greater pressure on the soil resources. To overcome this problem, Soil suitability assessment has been largely applied to estimating the agricultural capability of the land especially in drylands regions. A very interesting paper, carried out in the El-Fayoum depression of Northern Egypt, used the traditional Almagra model to evaluate soil suitability for 12 Mediterranean crops. The experiment involved contrasting the current and ideal scenarios of soil management as the study indicated that, there was a massive increment of highly suitable land areas under improved management. Major constraints were also noted to be soil salinity and sodium saturation that had a great effect on crop suitability. Nevertheless, it was based on the model that was made using rules and did not consider machine learning techniques, detailed analysis of soil properties and model validation which signifies a definite gap in the research [13]. The research problem dealt with the identification of appropriate agricultural land to alleviate sustainable land use planning and avert land degradation. The study was needed since the conventional land use decisions do not normally take into consideration integrated soil and environmental limitations. GIS layers were used to analyze spatial data of a chosen study area as opposed to a given numerical sample. Physical characteristics like soil texture, soil depth, drainage, slope and elevation were taken into consideration whereas the chemical characteristics like pH and NPK were not. The analysis utilized the traditional GIS-based Analytic Hierarchy Process (AHP) model rather than machine learning algorithms that include RF, SVM, or ANN. The method was an effective land classification method, but it was not integrated with machine learning, nor was it based on detailed soil chemistry and model validation, so it clearly has a research gap [14]. The research was aimed at understanding soil potential and crop appropriateness to dry land farming zones where degradation of the soil and food insecurity were a concern. This study was required to help in sustainable agricultural planning with synthesis of soil capability analysis coupled with land suitability analysis. The assessment was done using spatial data and soil profile of selected areas in the Northwest Nile Delta, Egypt.

The main soil parameters like depth of the soils, salinity, level of calcium carbonate, and the percentage of exchangeable sodium were taken into consideration with minimal detail given to the chemical properties. Instead of machine learning, the study used the Almagra (Micro-LEIS) system and multivariate statistical analysis to do the traditional GIS-based suitability modeling. Nevertheless, the lack of machine learning algorithms, few soil variables, and the validation of the models show that there is a definite gap in research to pursue data-driven soil suitability research [15]. The case was dealing with the issue of finding a viable land to produce food in arid and semi-arid areas where the inappropriate use of land lowers productivity. The study was required to underpin the sustainable land use planning approach through incorporating the soil and environmental indicators in the suitability evaluation. They were based on the use of spatial soil profile data, topography, climate data, and images of remote sensing of the western Nile Delta area. The soil characteristics were considered to be the texture, pH, electrical conductivity, exchangeable sodium percentage, organic matter, soil depth, slope and rainfall. A GIS-based multi-criteria analysis tool with AHP and weighted overlay analysis was used as a conventional method instead of machine learning models. The study involved no machine learning algorithms, did not validate the predictive accuracy, and assessed the weighting based on the experts that showed a definite gap in research [16].

## 2.2 Application of Machine Learning in Agriculture

The aim of the study was to forecast the soil-crop suitability patterns to enhance proper crop selection towards the establishment of sustainable agriculture. The study was necessary due to the subjectivity of the conventional methods of soil suitability that are inefficient in managing the multi-dimensional data of soil. The model was trained and tested on data of the soil samples in various parts of Negros Occidental, Philippines, which are a collection of several soil records. They were 14 soil properties such as the pH, organic matter, phosphorus, potassium, CEC, drainage, permeability and soil depth. Naive Bayes, Decision Tree, Random Forest, and Deep Learning models of machine learning algorithms were used. Though high accuracy was obtained (Random Forest ≈94.6%), the research lacked research diversity and external validation, which means that there is a research gap [17]. The chapter discusses the issue of the enhancement of agricultural yield and the decision making in a complex interaction of soil, crop, and climate. The study is required since the outdated farming systems and mere statistical analysis procedures are incapable of managing extensive and complex farm data. It is a review-based research that consolidates the results of a number of datasets, such as soil data, crop yield data, weather data and remote sensing data. Some of the major aspects that were covered are soil properties such as pH, NPK, moisture, climatic variables, crop characteristics and environmental indicators. Some of the machine learning techniques in the study include the Random Forest, Support Vector machine, Artificial Neural Networks, and ensemble models. It, however, points out the gaps, which included the problem of data quality, lack of interpretability of the models, and the un-standardized validation structures among studies [18]. The issue of traditional manual evaluation of agricultural land being time consuming, costly and error-prone was tackled by the study. This research was necessary, to build an automated, accurate, and efficient land suitability evaluation system using machine learning. Geospatial information of soil, climatic, and topographic data of the Abhanpur district in Chattisgarh, India was used for the analysis. Key factors comprised soil, climate and topographic parameters based on the FAO framework of land suitability. A hybrid approach was applied by combining MCDA (AHP) and machine learning, which Balanced Bagging classifier gave the best results. Although quite high balanced accuracy (Approximately 97%) was achieved, the research was confined to 2 crops and 1 region and therefore need to be more broadly validated [19]. The research was concerned with the issue of proper determination of suitability of the soil to support agricultural activities particularly during drought prone situations. This study was necessary since the conventional soil assessment techniques are ineffective and not able to capture complex interactions of soil. The suitability model was constructed with soil data in the various regions of the United States. The major soil parameters were the condition of the roots, nutrient status, soil toxicity, and access to oxygen. A prediction model in the form of a machine learning model was used to predict the suitability of soil. Even though high accuracy (approximately 98.81) was obtained, the research was restricted by data specific to the regions and other general validation was not done [20].

### 2.3 Machine Learning-Based Crop Recommendation Systems

The research will discuss the issue of appropriate crop choice to be grown to enhance the yield, resource utilization, and sustainability. The need to conduct this research is that it will be required to make informed decisions in the agricultural sector and also be able to guarantee food security at a time where population pressures are on the rise. Instead of an experimental dataset, the work is founded on a broad survey of the available work on crop recommendation systems. Such aspects as soil conditions, weather data, geospatial data, remote sensing data, and IoT-based data are mentioned as important. The paper deals with machine learning and data-driven solutions in precision agriculture. Nevertheless, it lacks a particular predictive model, crop-wise validation, and quantitative accuracy findings, and this data is a gap in the research [21]. The paper presents this issue of poor crop production and economic suffering of the marginal farmers caused by poor crop selection and yield forecasting. The study is required to enhance agricultural output, earnings of farmers, as well as decision making based on smart, user-friendly technology solutions. The system involves agricultural data that is supplied by the farmers via a smart phone application that comes with a GPS that determines the location and the soil. Major ones are the type of soil, area of land, crop type and other production based agricultural parameters. SVM, ANN, Random Forest, MLN and KNN regression machine learning models were used and the best results were achieved using the random forest. Despite the high accuracy (approximately 95%), lack of large-scale validation, and the absence of a discussion of environmental factors are limitations to the study, as the data set used is region-specific [22]. The paper solved the issue of lack of adequate and inaccurate soil suitability data on sustainable agriculture and food security in the arid and semi-arid areas. The necessity to conduct this research was that traditional assessment methods have been not able to effectively model the complex interactions between soil and the environment to assist with agricultural planning. The inventory of 238 soil suitability points was employed on the basis of soil data combined with remotely sensed phenological data. There were 14 physico-chemical soil factors and 4 phenological parameters among the key factors and phenological factors were determined as the most significant. Random Forest, Xgb Tree, ANN, KNN and SVM machine learning models were used to produce soil suitability maps. Even though it yielded high performance (best AUC $\approx$ 0.97), both the scope of the study (regional) and lack of multi-crop validation were limitations of the study [23].

### 2.4 Soil Parameter-Based Classification Using Machine Learning

The research countered the issue of proper estimation of the probability of soil liquefaction to be used in the risk-based geotechnical design, particularly under rare failure occurrences. The necessity of this research was due to the fact that the conventional Monte Carlo Simulation techniques are computationally ineffective in low-probability liquefaction evaluation. The model was developed and tested using geotechnical datasets that involved the state parameters of soil state, stress conditions, and liquefaction response parameters. The essential ones were state parameter ($\psi$), relative density, effective stress, cyclic resistance ratio, and other mechanical soil properties. A subset simulation structure along with machine learning models

(DNN, XGBoost, and CatBoost) were used, and DNN did the best. Despite the high accuracy (DNN: R 2 = 0.93, RMSE = 0.079), the paper is confined to the analysis of geotechnical liquefaction and does not cover the suitability of agricultural soils, as well as crop application [24]. The issue that was handled through the study was the need to predict biological activity in soils to promote sustainable land and soil management. The study was necessary since conventional soil evaluation techniques have limited abilities in the ability to measure the complicated biological functions in the soil. The data used were those that were obtained in the soil samples of agricultural fields with data subsets classified under training and testing subsets. The primary soil characteristics were that of carbon content, nitrogen content, and components of soil texture like sand, silt and clay. A regression model based on machine learning was utilized to forecast the activity of soil respiration through the use of the Random Forest model. Even though the results were moderate accuracy (approximately 70%), the analysis did not incorporate crop-specific and advanced climate variables [25]. The chosen issue in the study was the possibility of choosing appropriate crops depending on the soil and the environmental situation instead of the experience of farmers. This study was required to enhance the accuracy of crop recommendations and aid in the use of data to make decisions in agriculture. There was the use of soil and environmental datasets which were subdivided into training and testing sets but the sample size was not stated. The most important ones were the characteristics of soil and environmental variables that concerned the growth of crops. A number of machine learning classifiers were implemented, such as kNN, Naive Bayes, Decision Tree, SVM, Random Forest, and Bagging. Even though Bagging did the best, the analysis was not large scale and involved an in-depth and crop by crop performance analysis [26]. The research problem was that little information was available on suitability of the land to support sustainable agriculture in semi-arid areas of Iran. This study was necessary to enhance crop production planning and sustainability based on a scientific land suitability model. A sample of 100 soil profiles in the soil and topographic and climatic data of a 65 $km^2$ zone in the Kurdistan province of Iran were used. These were physical-chemical properties of soils, slope, rainfall, depth of soil, pH, and gravel content. The models of machine learning were tested and compared to conventional methods of FAO-based land suitability in rain-fed wheat and barley. Though the accuracy of ML was better than the traditional ones, the study was constrained by the regional conditions and environmental limitations like rainfall and slope [27].

## 2.5 Limitations of Existing Studies and Research Gap

A in-depth examination of existing literature reveals a number of shared limitations of existing studies on soil suitability and machine learning-based agricultural studies. Many previous works are based on traditional rule-based, GIS-based or expert driven land suitability evaluation methods such as Almagra, FAO frameworks and Analytic Hierarchy Process(AHP) which lack the flexibility to model complex interactions between the multiple soil variables [13]- [16]

Although recent research have presented machine learning approaches for soil and crop suitability study, a lot of them address limited geographic regions and region-specific data set limiting their generalized applicability (citations from [17], [19], [20], [22]).

Another important limitation that has been noted in the literature is the incomplete consideration of the soil chemical properties. Several studies focus on physical soil properties with little consideration of important chemical soil parameters such as pH and nutrient content that are important to provide an accurate assessment of crop suitability [14]-16].

Furthermore, many studies concerned with machine learning algorithms are interested in a limited number of crops or a single crop without conducting a thorough multi-crop suitability analysis (citations from [19], [21], [26], [27]).

In addition, a comparative evaluation of multiple machine learning algorithms is often missing and studies report their results for one best performing model without systematic comparison or interpretability analysis (citations from [17], [20], [25]).

Based on these limitations, there seems to be an apparent research gap in developing an area-specific and machine learning-based soil suitability framework that considers detailed soil chemical parameters, tests multiple crop species and compares the performance of various interpretable machine learning models. Moreover, lack of research work on Bogura region of Bangladesh is the motivation for the present research.

# 3. Chapter 3 : Methodology

## 3.1 Overall System Workflow

The generic methodology of this research consists of a structured workflow ending with the model evaluation, which starts with the collection of data. The major steps involved are the soil data collection, data preprocessing, feature selection, model training, model testing, and performance evaluation. Multiple machine learning classifiers are implemented and compared in order to find the most effective model to classify the soil suitability.

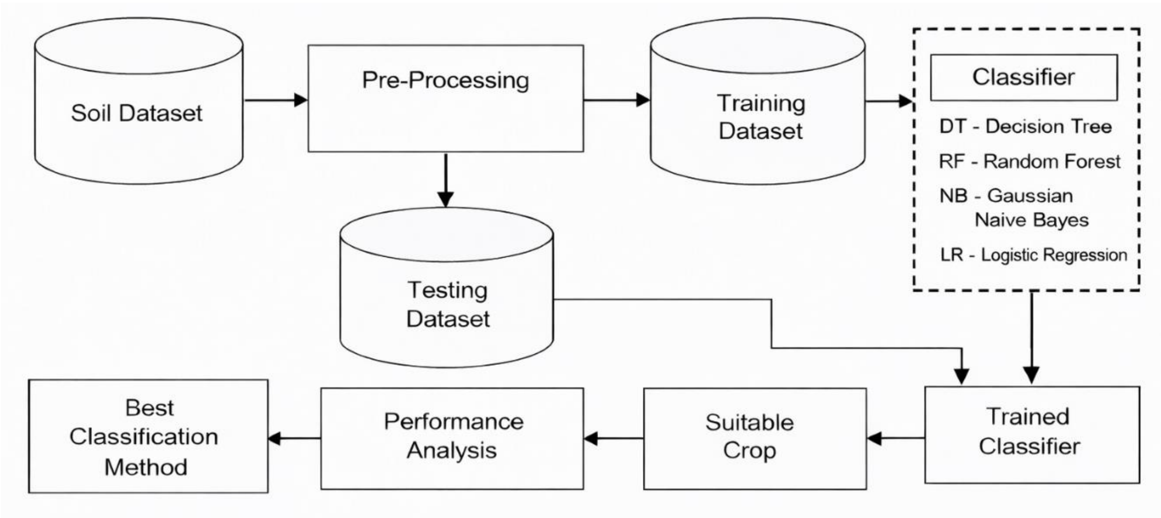


Figure 3.1 : Workflow of the system

## 3.2 Study Area Description

This study is devoted to the region of Bogura which is situated in the North of Bangladesh. The region has been chosen because of its agricultural significance and because it has a variety of crops that are grown like rice, maize, banana and papaya. The differences in the soil properties within the region make it an appropriate case to study soil suitability using the machine learning methods.

### 3.2.1 General

The Bogura district is located in the north western part of Bangladesh in the fertile area of Bengal Delta. The natural processes that occur along the river system, Ganges-Brahmaputra-Meghna (GBM), have created the geomorphology of the city of Bogura. Bogura is located in the geomorphic region of the Barind Tract due to its high elevation as compared to the surrounding floodplains.

### 3.2.2 Geomorphology

A Following are some of the detailed description of soil classification in Bogura District on the basis of digital soil map developed by Food and agriculture Organization (FAO): A soil classification map of Bogura district can be seen in Fig. 3.2.2. The map shows that areas of Bogura are covered mostly by 4 types of soils. These soil types are, alluvial silt, alluvial sand, alluvial silt and clay and clay residue. Alluvial sand occupies the area of 6% of the total area of Bogura district. About 32% of Bogura district is covered by the alluvial silt. These are fine particle which are smaller than sand but bigger than clay. Out of the total place that Bogura district covers, this type of soil has a coverage of 12%. In general, alluvial soils of silt and clay are cohesive, smooth and sometimes somewhat sticky when wet. Clay residuum is a type of soil which has been formed directly from the weathering and decomposition of parent rock material without much transport and deposition In the case of Bogura district 44% of the area is covered with this type of soil.

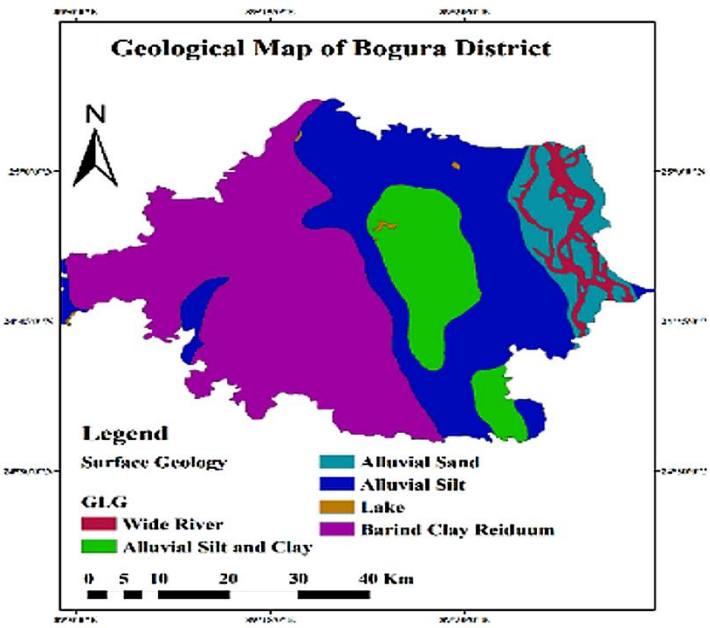


Figure 3.2 : Surface geology of Bogura created by FAO

## 3.3 Dataset Description

This study makes use of a soil dataset which has been specifically prepared for soil suitability analysis for crop cultivation in the Bogura region in Bangladesh. The dataset includes soil relevant attributes that are closely related to soil fertility and crop growth. These attributes are used as the input features for training and evaluating machine learning models for classifying soil suitability for selected crops. The data set is intended to reflect various soil conditions in different agricultural location in the Bogura district. This section contains a detailed description of the dataset with information of data source, selected soil parameters, targeted variables and the dataset attributes, which are fundamental to understand the data preparation process and further model development.

### 3.3.1 Data Source

The data of soil was utilized in this study in the Bogura district of Bangladesh in the Soil Resource Development Institute (SRDI). SRDI is a government agency that carries out soil survey, soil classification and soil fertility assessment in various parts of the country. The institute offers credible and standardized data on soils which are extensively utilized in planning and research of agriculture. The data consists of information on soil samples taken at the agricultural land of the Bogura area. These samples are used to represent different soil conditions in different locations within the district making them varied in soil properties. Field sampling and laboratory analysis were used to acquire the data with the help of SRDI under the regular conditions of soil testing. The gathered data will include the measurement of the most important soil parameters that can be applied to the crop cultivation such as pH value, organic matter (OM), nitrogen (N), phosphorus (P), potassium (K), temperature, rainfall, humidity, soil texture etc. These parameters have been chosen because they are important parameters that influence the fertility of soil and suitability of crop.

### 3.3.2 Input Features

Table 3.1 : Description of soil and environment input features

| Sl. No. | Soil Characteristics | Type | Description |
|---|---|---|---|
| 1 | pH | Numerical | Gives the acidity or alkalinity of soil which influences the availability of nutrients and growth of crops. |
| 2 | Organic_Matter (%) | Numerical | The proportion of organic matter that enhances soil fertility and soil moisture. |
| 3 | Nitrogen (%) | Numerical | Determines the nitrogen supply in soil that is critical in the growth and productivity of plants. |
| 4 | Phosphorus (mg_kg) | Numerical | Implicates the amount of phosphorus that is needed to build the root growth and energy conveyance in vegetation. |
| 5 | Potassium (mg_kg) | Numerical | Represents the level of potassium that promotes the strength and the resistance of plants against diseases. |
| 6 | Soil_Texture | Catagorial | Identifies the physical makeup of soil with the proportions of sand, silt and clay. |
| 7 | Rainfall(mm) | Numerical | Refers to the quantity of rain that affects the water content on the soil and water availability to crops. |
| 8 | Humidity (mm) | Numerical | Indicators of the moisture content of the atmosphere that influence the evapotranspiration and crop stress. |
| 9 | Temperature | Numerical | Means the surrounding temperature which has effects on growth of crops and the seasonal aptness. |
| 10 | Soil_Type | Catagorial | Categorizes soils in terms of physical and chemical properties of soils as far as crop suitability is concerned. |
| 11 | Drainage | Catagorial | Defines the capacity of the soil to take up surplus water and forego waterlogging. |
| 12 | Sunlight Exposure (hours) | Numerical | Represents the amount of solar radiation received by the soil and the crops and affects photosynthesis and the overall crop growth. |

### 3.3.3 Target Variable and Class Labels

The target variable used in this study is the crop suitability class assigned to each of the soil samples based on its soil and environmental characteristics.

Table 3.2 : Selected crops for soil suitability classification

| **Sl. No.** | **Crop Name** | **Scientific Name** | **Family** |
|---|---|---|---|
| 1 | Rice | Oryza sativa | Poaceae |
| 2 | Maize | Zea mays | Poaceae |
| 3 | Banana | Musa spp. | Musaceae |
| 4 | Papaya | Carica papaya | Caricaceae |

The goal of the machine learning models is to predict the best crop to grow based on the input features. In this research, four major crops generally cultivated in the Bogura region are considered as class labels i.e. Rice, Maize, Banana and Papaya. Each of the soil samples in the data set is identified with one of these crop classes based on suitability for that crop under specific soil conditions. The target variable is considered a categorical class label in which case the problem is a multi-class classification problem.

Table 3.3 : showing the target variable (crop suitability)

| **Label** | **Description** |
|---|---|
| S1 | Highly suitable |
| S2 | Moderately suitable |
| S3 | Marginally suitable |
| N1 | Currently not suitable |
| N2 | Permanently not suitable |

By learning the relationship between the soil parameters and crop suitability, the goal of the machine learning models is to be able to accurately classify soil samples into the correct crop category.

### 3.3.4 Dataset Size and Discribution

The data set used for this study is a soil sample collected in different agricultural locations in the Bogura region. Each piece of data is one unique set of attributes for a particular soil and environmental conditions with a corresponding label for crop suitability. The dataset has samples of different soil conditions to ensure that the variability and representativeness of the study area is met. The total number of samples in the dataset are 1699 and these samples are distributed between four classes of crops viz. Rice, Maize, Banana and Papaya. The distribution of samples in these classes is aimed at reflecting the general cultivation patterns of the region. Efforts were taken to have a relatively balanced distribution between the crop classes to minimize classification bias and thus increase model learning performance.For the model development and evaluation, the dataset is split into a training and testing subset. A significant portion of the data is used to train the machine learning models, while the remaining portion is used for testing and validation. This separation means that the models used for training are not used to evaluate the models, and hence we can get an unbiased measure of generalization capability of the trained models. Overall, the size of dataset and distribution of dataset gives enough ground for training and comparing various machine learning algorithms for classification of soil suitability.

## 3.4 Dataset Preprocessing

Data preprocessing is a very important step in machine learning as the quality of the input data directly influences the performance of the predictive models. Raw soil and environmental data usually have missing values, inconsistencies and scale variation, which need to be addressed before model training. In the present research work, several preprocessing steps are used to prepare the data set for the effective classification of soil suitability.

### 3.4.1 Handling Missing and Inconsistent Values

The dataset was checked to determine missing, incomplete or inconsistent records. Soil samples that had missing values in the critical parameters were addressed by the suitable data cleaning techniques. In the case of little missing values, statistical methods of imputation were used to replace the missing values with representative values such as the mean or median of the same feature. This method helped in preserving the overall data distribution whereas preserving the dataset integrity.

### 3.4.2 Encoding of Categorical Features

Some aspects of soil such as soil texture, soil type, and drainage condition are categorical by nature. Since machine learning algorithms require a numerical input, these categorical features were converted to numerical representation. Label encoding was used for effective processing of categorical classes into integer numbers so that the models can take both soil and environmental attributes well.

### 3.4.3 Feature Scaling and Normalizatio

The numerical features within the dataset, such as pH, concentrations of nutrients, rainfall, humidity, and temperature, will have a different range of values and unit. To make all features contribute equally while training the model, feature scaling was done. Normalization techniques were used to rescale numerical attributes to a similar range to avoid the fact that features with larger magnitudes are prioritized over others in the learning process, especially in distance-based algorithms.

### 3.4.4 Outliers Detection

Outliers for the soil data may be produced by measurement errors or extreme environmental conditions. The dataset was processed in order to identify abnormal values which were far away from the normal. Identified outliers were carefully reviewed and considered to downgrade the effect of outliers and leave meaningful variability in soil conditions.

### 3.4.5 Final Dataset Preparation

After data cleaning, encoding, scaling, and outlier handling techniques were applied to the data, the dataset was converted into a structured and machine learning-ready format. The final preprocessed dataset has the consist, normalized, and complete values for the features, so they can be used for training and testing of the machine learning models selected for the soil suitability classification.

## 3.5 Feature Selection

Feature selection is an important step in machine learning as it helps in identifying the most relevant input variables that contribute to the accurate prediction and also it reduces the model complexity. Selecting the right features helps to improve the performance of the model, lower computational cost and also reduce the risk of overfitting. In this study, feature selection is mainly directed by domain knowledge related to soil science as well as crop cultivation. Soil parameters such as pH, organic matter, nitrogen, phosphorus, potassium, moisture content and drainage condition are well known indicators of the soil fertility and crop growth. In addition, environmental factors such as rainfall, humidity, temperature, and sunlight are also taken into consideration because of their effects on plant development and water availability. An exploratory analysis was carried out to study the relationships between input features and crop suitability classes. Features that showed low variability or little influence in the differentiation of the crops were carefully reviewed. Redundant attributes were removed in order to prevent the unnecessary duplication of information and make the learning process easier. The last group of selected features gives a good balance of soil chemical properties, physical features and environment. This selection helps the machine learning models to learn about some meaningful patterns associated with soil suitability effectively without sacrificing the computational efficiency and interpretability.

## 3.6 Machine Learning Algorithm Used

In this study, supervised machine learning algorithms are used for soil suitability classification for crop cultivation depending on soil and environmental attributes. The goal of the multiple algorithms is to compare their performance and find the most effective algorithm to predict suitable crops. For this, four popular classifiers of machine learning are chosen: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Gaussian Naive Bayes (GNB). These algorithms are different learning approaches like tree-based model, probabilistic classifiers, and linear models, to comprehensively evaluate the soil suitability classification.

### 3.6.1 Decision Tree

Decision Tree (DT): Decision Tree is a rule-based classification algorithm which partitions the dataset into smaller subsets by using feature-based decision-imposed tree-like structure with class labels at the leaves. It offers an effective method of nonlinear correlation of soil characteristics and crop aptness. The high interpretability of the Decision Trees makes them suitable for agricultural decision making. However, they are prone to overfitting if the tree gets too deep and this can be controlled using proper training and pruning techniques. Based on rule-based learning, Decision Tree (DT) classifier determines the class label by a sequence of decision rules, from root node to leaf node and selects the best suitable crop from the testing samples. The performance evaluation of the DT classifier for the prediction of crop is done using the k-fold cross validation technique; here k-1 folds are used for training and the other fold is for validation.

### 3.6.2 Random Forest

Random Forest (RF) is one of the ensemble learning algorithms that build multiple decision trees with random subsets of the training data and feature space. The last classification outcome is established with the majority vote in all the trees of the ensemble. This approach helps to reduce the overfitting and improve the generalization of the model. Random Forest is suitable for soil suitability analysis for the following reasons: its robustness, ability to deal with high-dimensional and nonlinear data and its high resistance to noise and variability in soil properties. Based on the ensemble learning principles, the Random Forest (RF) classifier determines the class label based on majority vote by multiple decision trees and the most suitable crop is selected from the testing samples. The performance of the RF classifier for crop prediction is evaluated by using the k-fold cross-validation technique in which k-1 parts are used to train the model and the remaining part is used to validate the model.

### 3.6.3 Logistic Regression

Logistic Regression (LR) is a statistical classification model that estimates class probabilities by means of a logistic function and works fine when features and target classes are linearly related to each other. In this study, it is used as a baseline model for comparing the performance of more complicated classifiers. It's simplicity and computational efficiency and probabilistic output make it appropriate for the soil suitability classification. Based on the probability theory, Logistic Regression (LR) classifier is used to estimate the probability of class membership using logistic function and determines the class label with the highest estimated probability to select the suitable crop from testing samples. The performance of LR classifier for the prediction of crop is checked with the help of the k-fold cross validation technique, where k-1 folds are used for training and the remaining fold is used for validation.

### 3.6.4 Gaussian Naïve Bayes

The Gaussian Naive Bayes (GNB) technique is a probabilistic classifier that is modeled using the Bayes theorem and is a variant of the Naive Bayes family. The GNB is similar to NB in the sense that it also assumes that the input features are conditionally independent of the class label. However, GNB is different in the sense that it assumes that all the continuous features are described by a Gaussian (normal) distribution within the class. The probability of each feature is estimated with the help of the mean and variance calculated from the training data. Based on the probability theory, the GNB classifier assigns the class label to the testing samples with the maximum posterior probability and picks the most suitable crop based on the testing samples. The performance of the GNB classifier for crop prediction is evaluated using the k-fold cross validation technique in which k-1 folds is used for the training and the remaining fold is used for validation.

## 3.7 Model Training and Testing

Model training and testing is done to assess the effectiveness of the machine learning algorithms in soil suitability classification. After preprocessing, the dataset was split into training and testing sets with an 80:20 split, 80% of the data being used for training and 20%

for testing. To maintain the class distribution in both subsets, stratified split was used. A fixed random state (42) was used in order to make the experimental results reproducible. Four supervised machine learning models were trained and evaluated: Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes. For the case of Decision Tree and Random Forest, the models were trained with the original (non-scaled) feature space, whereas models for Logistic Regression and Gaussian Naive Bayes were trained with standardized features. Standardization was done by fitting the scaling on the training data and scaling the testing data in the same way to prevent data leakage. To ensure the reliability of the evaluation other than using a single split, 10-fold cross validation was performed with the training set for each model. The average accuracy on the cross validation samples was used as another measure of model generalization. After training, each model made predictions on the reserved test set, and performance was measured in terms of accuracy, precision, recall and F1-score. Confusion matrices were also generated for each classifier to get an idea of class-wise performance in prediction.

## 3.8 Performance Metrics

To assess the performance of the machine learning models for soil suitability classification, a number of standard metrics for classification are used. These metrics give an overall sense of effectiveness of the model, and not just checking the overall accuracy but the accuracy for each class. The selected performance metrics are Accuracy, Precision, Recall and F1-score.

### 3.8.1 Accuracy

Accuracy is a basic measure that is used to assess the performance of a classification model. It tells us what proportion of correct forecasts do the model make out of all forecasts. Accuracy is good but it gives False Positive impression of having high accuracy. The problem arises due to the possibility of misclassification of minor class sample to be very high.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

### 3.8.2 Precision

This is a measure of the number of positive predictions by the model are actually correct. It's useful when the cost of false positives is high like in medical diagnoses where it can have serious consequences to predict a disease while it's not there.

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

Where:

TP = True Positives and FP = False Positives

Precision assists in making sure that whenever the model makes a positive result, it will most probably be accurate.

### 3.8.3 Recall

Recall or Sensitivity is the measure of the number of actual positive cases that were correctly identified by the model. It is important when it is more costly to miss a positive case (false negative) than it is to get false positives.

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

Where:

TP = True Positives and FN = False Negatives

In cases where the detection of all positive cases is important (such as in the case of disease detection), recall is an important metric.

### 3.8.4 F1-Score

F1 Score is the harmonic mean of the precision and recall scores. It is useful in those cases when we want to find a balance between precision and recall as it gives both a number. A high F1 score indicates that the model has a good score on both metrics. Its range is [0,1]. Lower recall and higher precision give us great accuracy but then it misses a large number of instances. More the F1 score better will be performance. It can be represented by mathematical expression this way:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

# 4. Chapter 4 : Results & Discussion

## 4.1 Introduction

This chapter introduces the experimental results and analytical results achieved by the proposed machine learning-based soil suitability assessment framework. The main purpose of this chapter is to assess the performance of various machine learning models for classification of soil suitability for crop cultivation based on soil and environmental characteristics of Bogura region. The chapter starts with Exploratory Data Analysis (EDA) to gain insight on the underlying characteristics of the dataset such as class distribution, feature variability and relationships between soil parameters and suitability classes. EDA helps to get crucial insights about the data patterns and any possible issues like class imbalance or over-representation of one particular feature which is very important in understanding the model's behavior. Following the exploratory analysis, a model-wise performance comparison is performed in order to evaluate the effectiveness of the machine learning algorithms that have been selected. The performance of Decision Tree, Random Forest, Logistic Regression, and Gaussian Naive Bayes is evaluated by using the standard classification measures like accuracy, precision, recall, and F1-score. Tabular and graphical representations are used to clearly compare the model performance. Finally, the chapter is dedicated to selecting the best model based on the overall performance, consistency and robustness. The results are analyzed to determine the most reliable model for soil suitability classification and the implication of the results is discussed. The information learned from this chapter is the foundation for the conclusions and recommendations that follow in the next chapter.

## 4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to get an initial understanding of the data set and try to discover any patterns, trends and possible challenges that can be experienced before evaluating the model. EDA is helpful in studying the distribution of classes, variability of features, distribution of data, which are important in the interpretation of machine learning model performance.

4.2.1 Class Distribution Analysis

Class distribution analysis is performed to see the proportion of soil samples that belong to each class of soil suitability. In the present study, the suitability of the soil is classified into four groups namely; Highly Suitable, Moderately Suitable, Marginally Suitable, and Not Suitable. The understanding of this distribution of these classes is important because class imbalance can have a significant impact on model performance and may bias predictions toward certain classes, especially the dominant classes. To analyze class distribution normalized frequency counts were calculated based on percentage of each suitability class. This approach offers a good understanding of how the soil samples are spread in terms of their suitability levels. The distribution is shown using the graphical representation of percentage share of each suitability class.

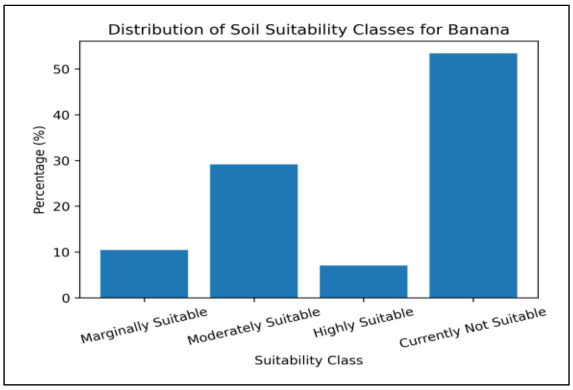


Figure 4.1 : Soil suitability classes for Banana

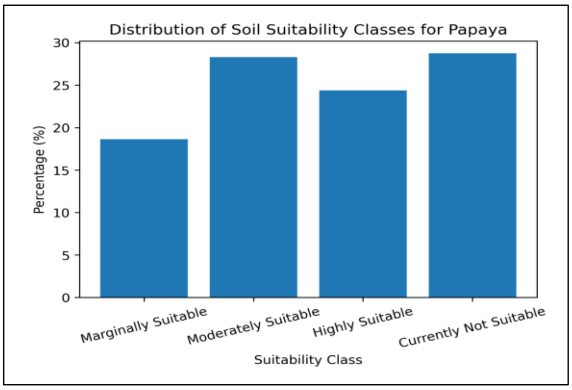


Figure 4.2 : Soil suitability classes for Papaya

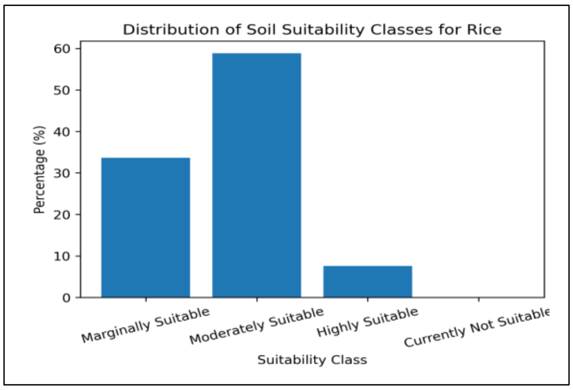


Figure 4.3 : Soil suitability classes for Rice

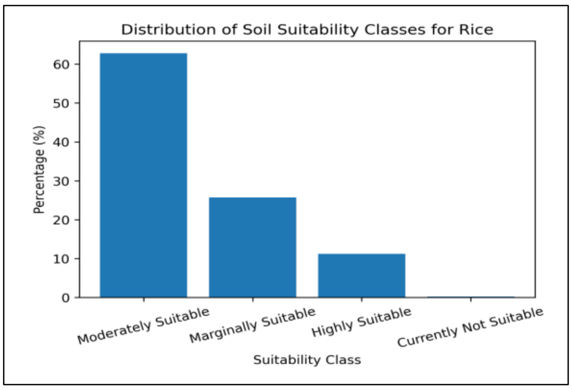


Figure 4.4 : Soil suitability classes for Maize

The distribution of soil samples into the various suitability classes is shown in the above figure in terms of percentage. The distribution shows that the dataset is balanced in terms of the samples from all the suitable categories, implying it is representative in terms of the classes. To make sure a fair evaluation of the model, stratified sampling was used in train-test split to maintain the same class distribution.

### 4.2.2 Feature Distribution Analysis

Feature distribution analysis is conducted to analyze the statistical characteristics and variation in the soil and environmental attributes used in this study. Understanding the distribution of individual features is useful to identify data spread, dominant ranges and potential outliers, important in understanding model behavior and prediction performance.

**Soil pH Distribution**

Soil pH is an important index of acidity or alkalinity of soil and it directly affects nutrient availability and crop growth. The distribution of pH values of the soil reveals that most of the samples lie in the slightly acidic to the neutral range. This shows good soil conditions for the growing of crops in several areas of the study area. However, the existence of extreme pH values indicates variability of soil quality which can impact on the classification of the crop suitability.

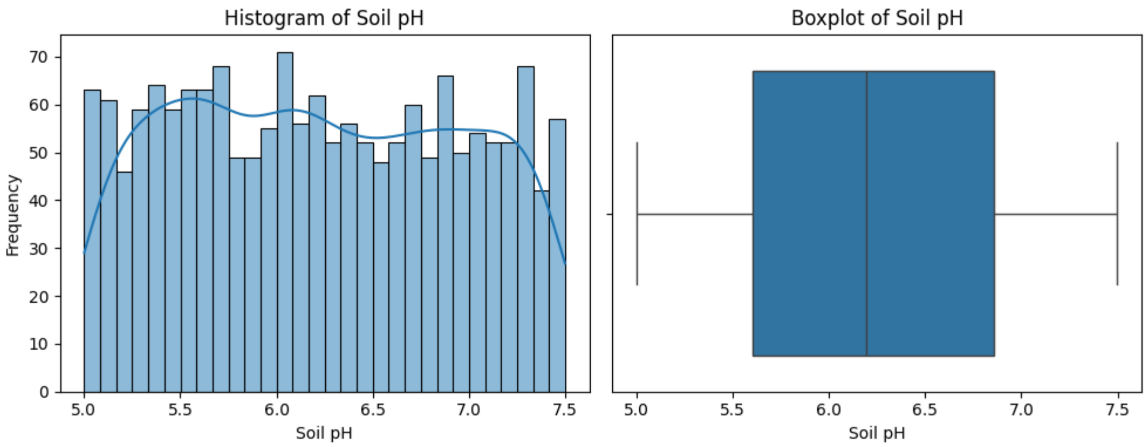


Figure 4.5: Boxplot and histogram of the distribution of soil pH

**Nitrogen (N) Distribution**

Nitrogen content has a major role to play in plant growth and yield. The distribution of the nitrogen values shows that there is some noticeable variation in the dataset, with a larger concentration of samples showing low to moderate levels of nitrogen. This variation points to the importance of nitrogen as a discriminative feature of variation in soil suitability classes.

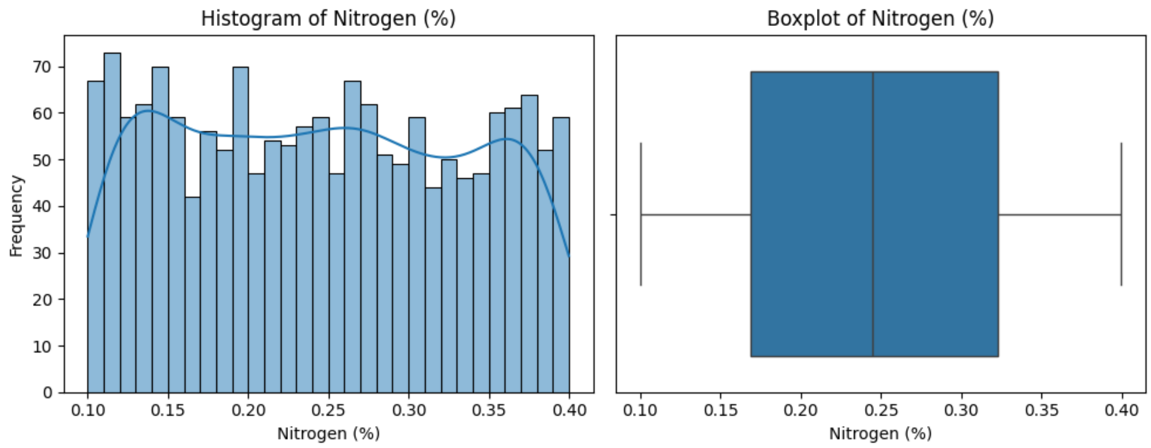


Figure 4.6: Boxplot and histogram of the distribution of soil nitrogen content

### Phosphorus (P) Distribution

Phosphorus is necessary for root development and the transfer of energy in plants. The phosphorus distribution shows a broader distribution than other nutrients indicating that there is a lot of variability in the availability of phosphorus across the different soil samples.

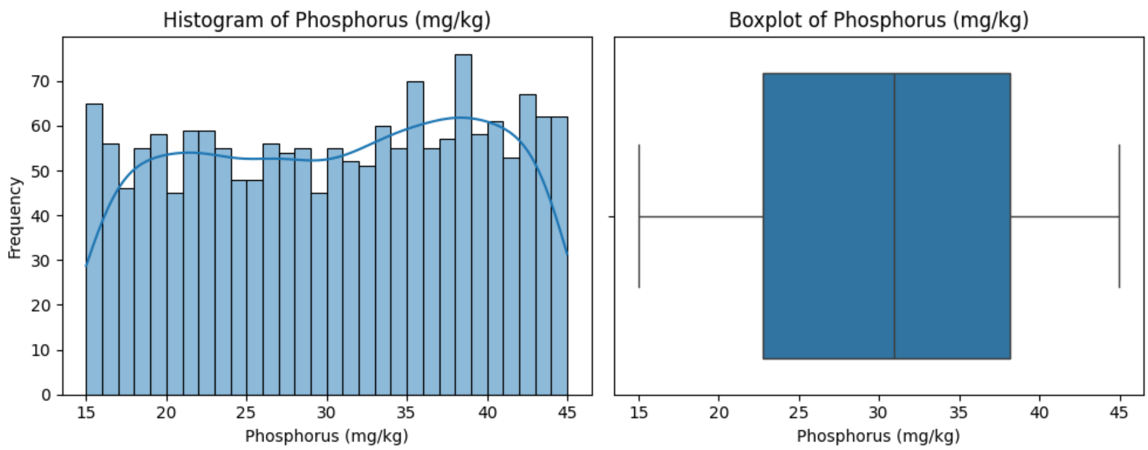


Figure 4.7: Boxplot and histogram showing the distribution of phosphorus concentration in soil

### Potassium (K) Distribution

Potassium is responsible for the strength and stress tolerance of vegetable plants. The distribution of the values of potassium shows it to be moderate with a majority of samples falling within a medium range. This raises the possibility that the availability of potassium may be adequate for some crops, but may limit others, and this is an important feature in suitability assessment.

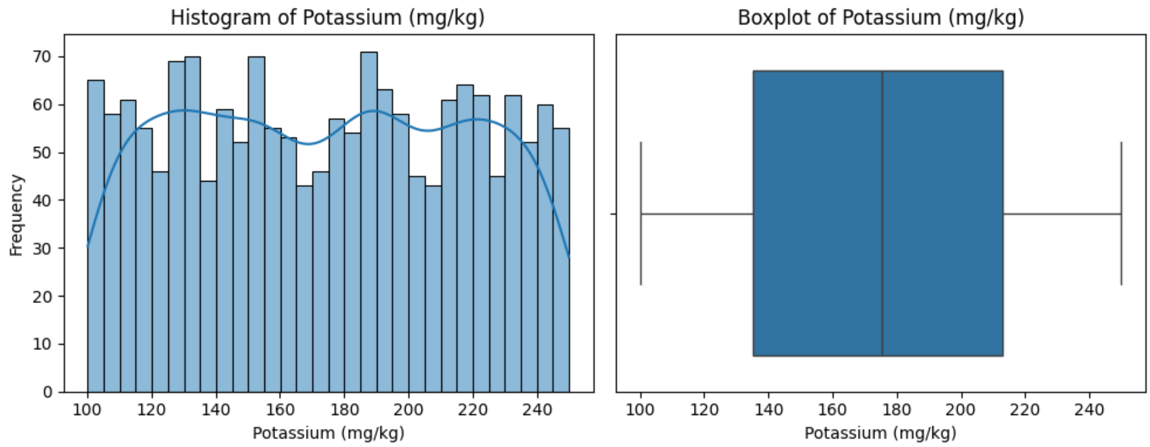


Figure 4.8 : Boxplot and histogram showing the distribution of potassium concentration in soil.

### Organic Matter Distribution

Organic matter content and soil moisture are important factors in the fertility of soils and the retention of water. The distribution of organic matter values indicates that most samples have low to moderate organic content of the samples.

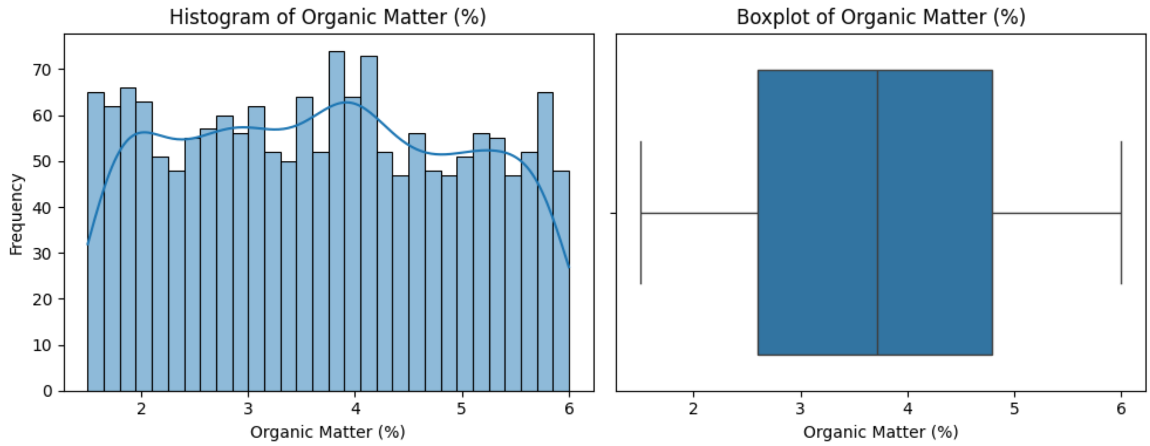


Figure 4.9 : Boxplot and histogram showing the distribution of organic matter

### Distribution of Environmental Features

Environmental attributes such as rainfall, temperature, humidity, and sunlight show a seasonal and regional change. Their distributions suggest that climatic conditions in the study are different in various zone, affecting the availability of soil moisture and the growth patterns of crop. Inclusion of these features provides the additional robustness to the machine learning models capturing the influence of the environment on soil suitability.

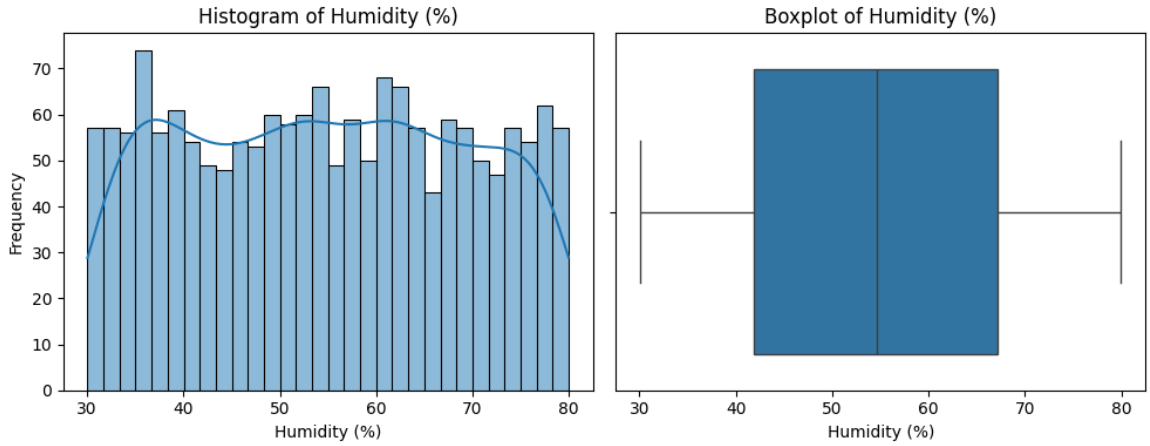


Figure 4.10 : Boxplot and histogram showing the distribution of humidity levels.

Overall, feature distribution analysis shows that the dataset has sufficient variability with respect to the soil and other environmental parameters. This heterogeneity allows the machine learning models to gain meaningful patterns and allows the effective classification of soil suitability.

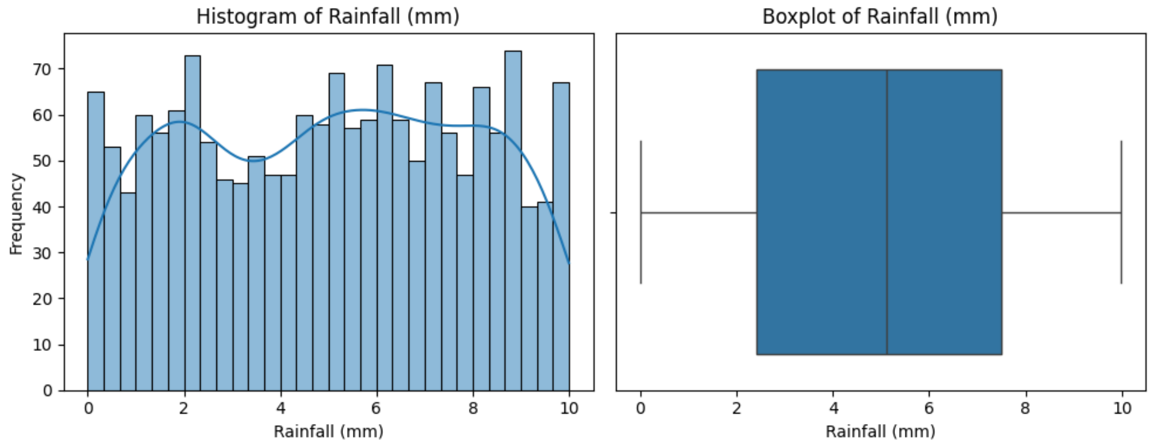


Figure 4.11 : Boxplot and histogram showing the distribution of rainfall values.

**Drainage Condition Distribution**

Drainage condition is one of the important properties of soil which determines the specific ability of soil to remove excess water after rainfall or irrigation. The distribution analysis of drainage condition conditions indicates that the soil samples are classified under different drainage classes such as good, fair, and poor. A higher proportion of samples fall under the fair and good drainage categories and this implies that a significant number of agricultural lands in the study area have acceptable water movement characteristics.

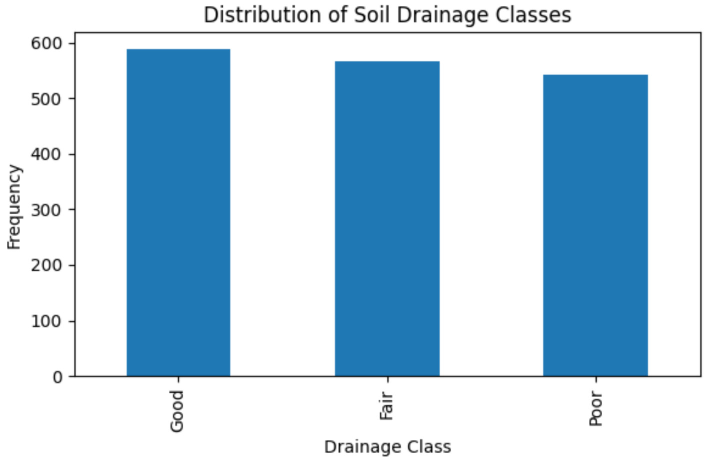

Figure 4.12 : Distribution of drainage in the dataset

However, the occurrence of poorly drained soils is the risk of water logging in some places which has negative impacts on root development and crop growth. This is the variability that makes a drainage an influential feature in the classification of soil into different soil suitability classes.

**Soil Texture Distribution**

Soil texture refers to the physical makeup of soil according to the proportionalities of sand, silt and clay. The soil texture distribution suggests that the dataset is a combination of the soil types, sandy, loamy, and clayey. Loamy soils occur more often, which is also good in terms of cultivating crops because they are balanced in their ability to retain water and aerate. Although well-drained, Sandy soils can have low nutrient retention but Clayey soils can have excessive moisture. The variation in the soil texture among the samples is a source of variability of the soil suitability and is of considerable importance in the classification process.

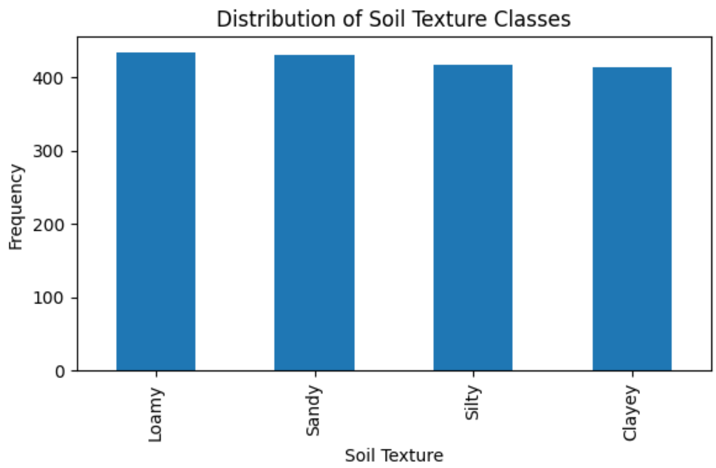


Figure 4.13 : Distribution of soil texture in the dataset

## 4.3 Model-Wise Comparison Performance

This section shows a comparative analysis of the machine learning models used for soil suitability classification. The purpose of this comparison is to test the relative efficacy of each model with regard to prediction accuracy, consistency and general classification performance. Four supervised machine learning algorithms, namely Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and the Gaussian Naive Bayes (GNB), were tested in the same experimental conditions to make the testing fair. The performance of the models was evaluated with the help of cross validation accuracy and testing accuracy, but also some other evaluation criteria such as precision, recall and F1 score. These metrics give a complete picture of the overall correctness and also class wise prediction behaviour.

### 4.3.1 Accuracy Comparison for Rice

Accuracy has been used as one of the main indicators to quantify the ratio of the number of correct soil classification. The results of the cross validation and test accuracy obtained for each model are summarized in the below table.

Table 4.1 : Showing the accuracy comparison for rice crop

| Model | CV Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| Decision Tree | 94.62% | 94.71% |
| Random forest | 97.05% | 98.24% |
| Logistic Regression | 72.24% | 72.94% |
| Gaussian Naïve Bayes | 77.39% | 76.18% |

The results suggest that there is noticeable variation in the accuracy of the evaluated models. Tree-based approaches tend to be more accurate than probabilistic and linear models because they can model nonlinear relationships between soil and environmental features.

### 4.3.2 Accuracy Comparison for Banana

Table 4.2 : Showing the accuracy comparison for banana crop

| Model | CV Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| Decision Tree | 95.95% | 95.59% |
| Random forest | 92.92% | 92.35% |
| Logistic Regression | 61.75% | 60.00% |
| Gaussian Naïve Bayes | 74.42% | 72.65% |

### 4.3.3 Accuracy Comparison for Papaya

Table 4.3 : Showing the accuracy comparison for papaya crop

| Model | CV Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| Decision Tree | 93.82% | 94.41% |
| Random forest | 93.67% | 96.76% |
| Logistic Regression | 67.16% | 72.06% |
| Gaussian Naïve Bayes | 59.36% | 62.94% |

### 4.3.4 Accuracy Comparison for Maize

Table 4.4 : Showing the accuracy comparison for maize  crop

| Model | CV Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| Decision Tree | | |
| Random forest | | |
| Logistic Regression | | |
| Gaussian Naïve Bayes | | |

On the contrary, the Logistic Regression and Gaussian Naive Bayes have fairly lower accuracy since they assume simplification. The fact that cross-validation and test accuracy values are consistent points to the assumption that the models have consistent generalization performance with regard to classifying rice suitability.

## 4.4 Graphs and Charts

Graphical representations are employed for visual analysis and comparison of performance of machine learning models for soil suitability classification. While tabular results give exact numerical values, graphs give an intuitive idea of performance trends as well as relative differences between models. In this paper, bar charts and comparison graphs are used to show the performance of models on different crops and evaluation metrics.

### 4.4.1 Accuracy Comparison Graph (Rice)

Accuracy comparison graphs are used to compare the accuracy of the evaluated models graphically. For each crop, namely rice, maize, banana and papaya, bar charts are plotted that are based on the accuracy values of the test set obtained from the trained models. These graphs clearly indicate the performance difference between the Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes classifiers.

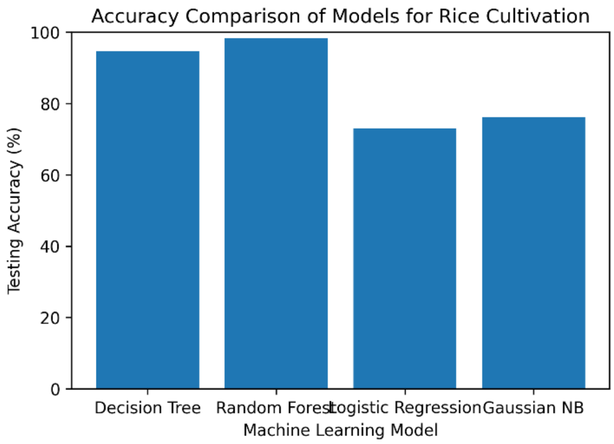


Figure 4.14 : Accuracy comparison of machine learning models for Rice cultivation

For the rice suitability classification, the accuracy of Random Forest model showed the best performance with testing accuracy of 98.24% and best cross validation accuracy demonstrating the good generalisation ability. The Decision Tree model also worked well with slightly inferior but similar accuracy. Logistic Regression performed moderately, indicating its low ability to capture complex non-linear relationships. Gaussian Naive Bayes gave a relatively lower accuracy and seems to be sensitive to assumptions of feature distribution and independence.

### 4.4.2 Accuracy Comparison Graph (Banana)

Accuracy comparison graphs are used to compare the accuracy of the evaluated models graphically. For each crop, namely rice, maize, banana and papaya, bar charts are plotted that are based on the accuracy values of the test set obtained from the trained models. These graphs clearly indicate the performance difference between the Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes classifiers. .

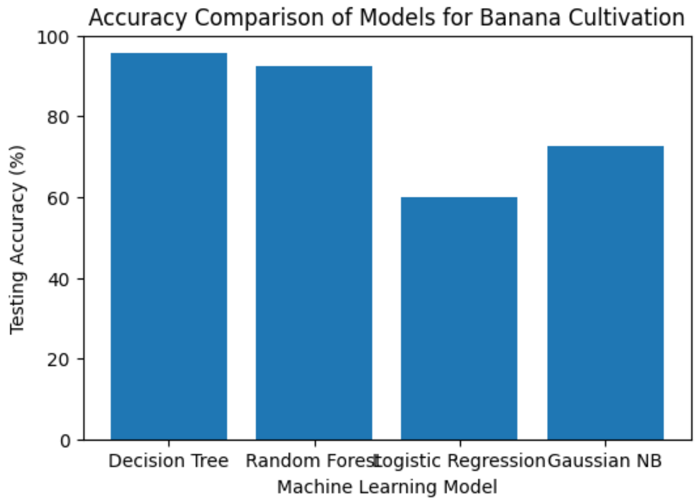


Figure 4.15 : Accuracy comparison of machine learning models for Banana cultivation

For banana suitability assessment, the classification accuracy of Decision Tree (DT) model was the highest 95.59%, which proved that the model is suitable to extract rule-based soil-climate relationship. The Random Forest (RF) model has also worked well, but slightly lesser than DT. Gaussian Naive Bayes (GNB) showed rather poor performance, which may be attributed to the fact that it is sensitive to the assumptions about the distribution of the features. Logistic Regression (LR) provided comparatively less performance, which indicates that it has certain limitations when it comes to modeling complex non-linear patterns.

### 4.4.3 Accuracy Comparison Graph (Papaya)

For each crop, namely rice, maize, banana and papaya, bar charts are also plotted that are based on the accuracy values of the test set obtained from the trained models. These graphs clearly indicate the performance difference between the Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes classifiers.

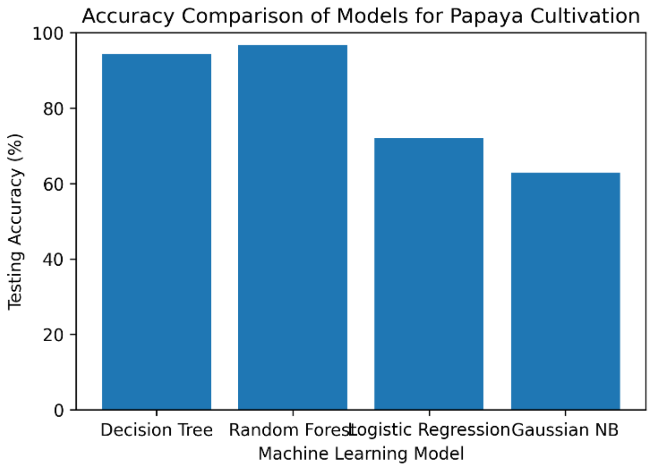


Figure 4.16 : Accuracy comparison of machine learning models for Papaya cultivation

For papaya suitability prediction, the Random Forest model gave the highest accuracy 96.76% of prediction on the testing data and overall performance. The Decision Tree model performed slightly worse than Random Forest but had pretty good classification ability. Logistic Regression has performed moderately with relatively less accuracy. Gaussian Naive Bayes had the worst performance out of the models evaluated.

### 4.4.4 Accuracy Comparison Graph (Maize)

For each crop, namely rice, maize, banana and papaya, bar charts are also plotted that are based on the accuracy values of the test set obtained from the trained models. These graphs clearly indicate the performance difference between the Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes classifiers.

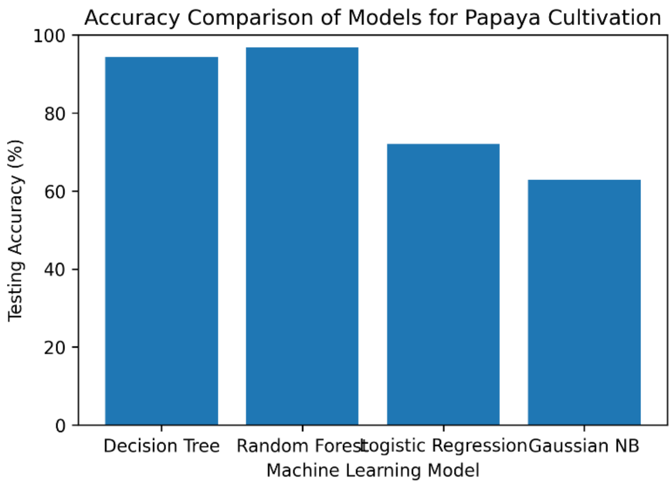


Figure 4.17 : Accuracy comparison of machine learning models for Maize cultivation

## 4.5 Precision, Recall and F1-Score Comparison

Accuracy alone does not fully represent the performance of multi-class classification model, especially in the case of soil suitability assessment where class imbalance and overlap of feature ranges are possible. Therefore, this study further evaluates the machine learning models with the help of precision, recall and F1- score so as to give comprehensive evaluation of performance. Precision describes the accuracy of predicted suitability classes, recall the capability to find all of the samples of a class which are relevant for our analysis and F1-score is a mean harmonic of precision and recall. is a balance of both metrics. The comparative analysis shows that tree-based models are better than linear and probabilistic models in these metrics for all comparisons. For rice and papaya crops, the Random Forest model gets higher and more stable values for precision, recall and F1 score, which shows the good generalization ability of the model. In contrast, in the case of banana cultivation, the Decision Tree model has a good performance, indicating that crop-specific soil requirements can affect the model performance. Logistic Regression and Gaussian Naive Bayes have relatively low scores because they have a limited capacity to model the relationships between soil and environmental features that may involve nonlinear relationships.

### 4.5.1 Precision, Recall and F1-Score (Rice)

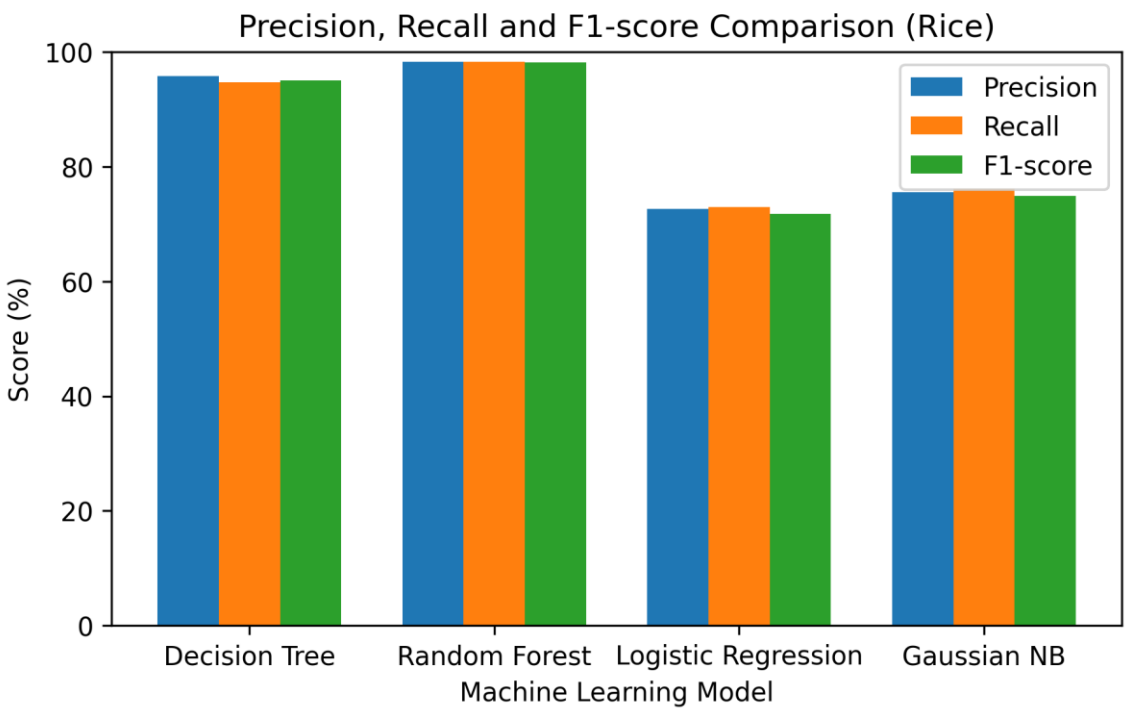


Figure 4.18 : Precision, recall, and F1-score comparison of machine learning models for rice

4.5.2 Precision, Recall and F1-Score (Banana)

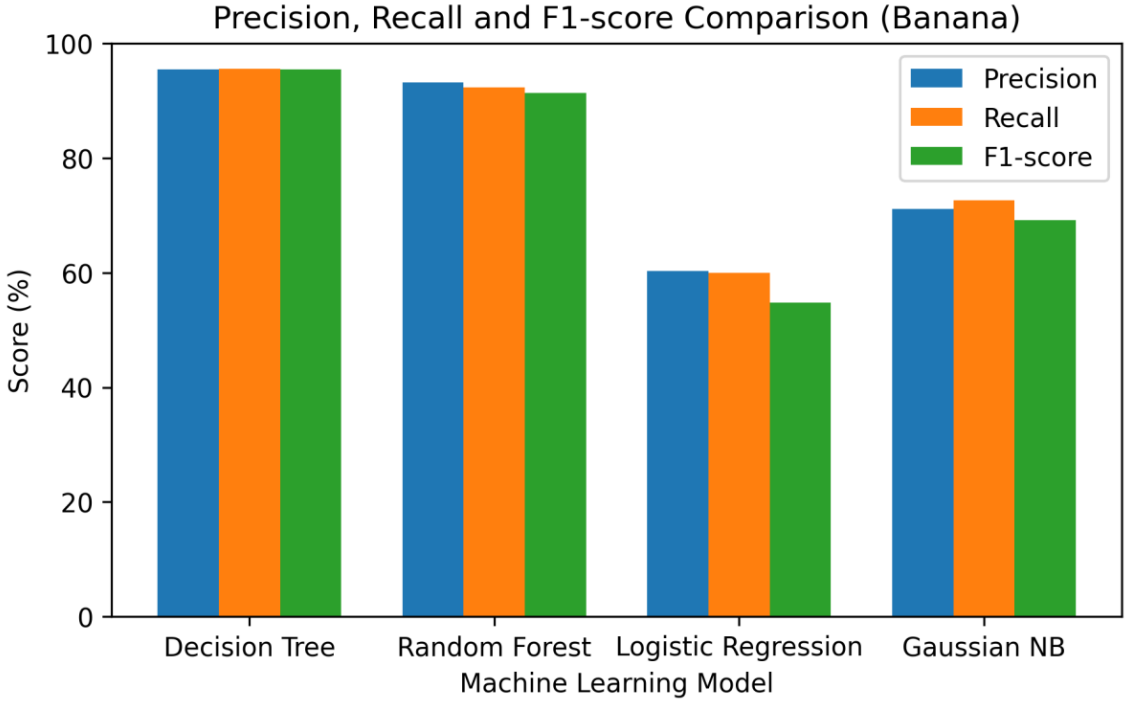


Figure 4.19 : Precision, recall, and F1-score comparison of machine learning models for banana

4.5.3 Precision, Recall and F1-Score (Papaya)

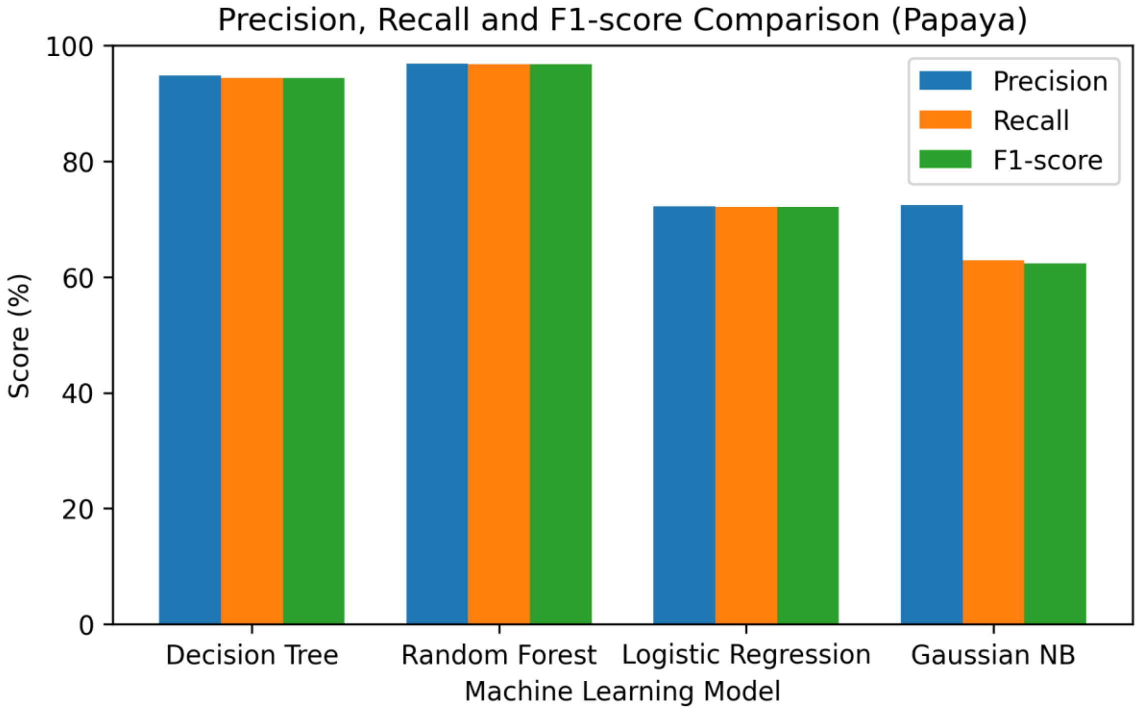


Figure 4.20 : Precision, recall, and F1-score comparison of machine learning models for papaya

### 4.5.3 Precision, Recall and F1-Score (Maize)

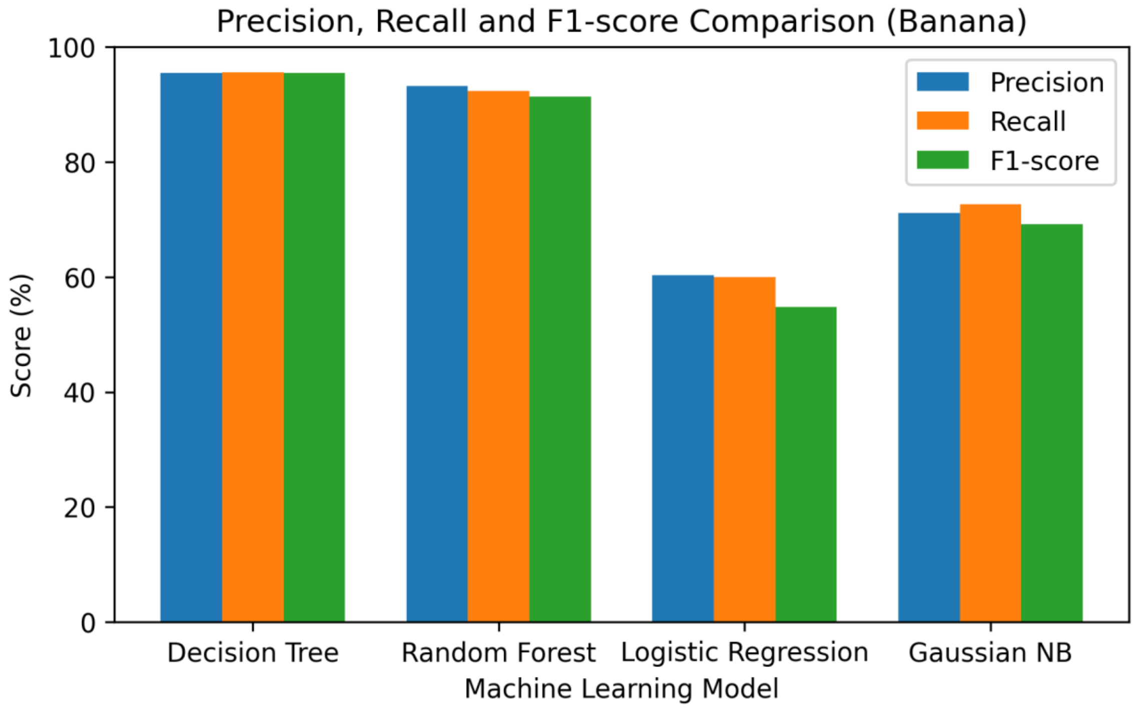


Figure 4.21 : Precision, recall, and F1-score comparison of machine learning models for maize

## 4.6 Confusion Matrix Visualization

To further analyze class wise prediction behavior, Confusion matrices were used to visualize the performance of a model: best performing model per crop. The confusion matrix provides detailed information on correct classifications and misclassifications in the different classes of soil suitability. Diagonal elements are correctly classified instances and off-diagonal elements misclassification occurs between classes. The results of the confusion matrix shows that Random Forest is effective in rice and papaya suitability classification as the number of correct predictions with less misclassifications are higher. For banana Decision tree model shows better class separation. Most of the misclassifications are between adjacent suitability classes such as Moderately Suitable and Marginally Suitable which is expected due to gradual variation of soil properties. These findings assist in the strength of tree-based models to model complex interactions between soils and crops.

### 4.6.1 Confusion Matrix Visualization (Rice)

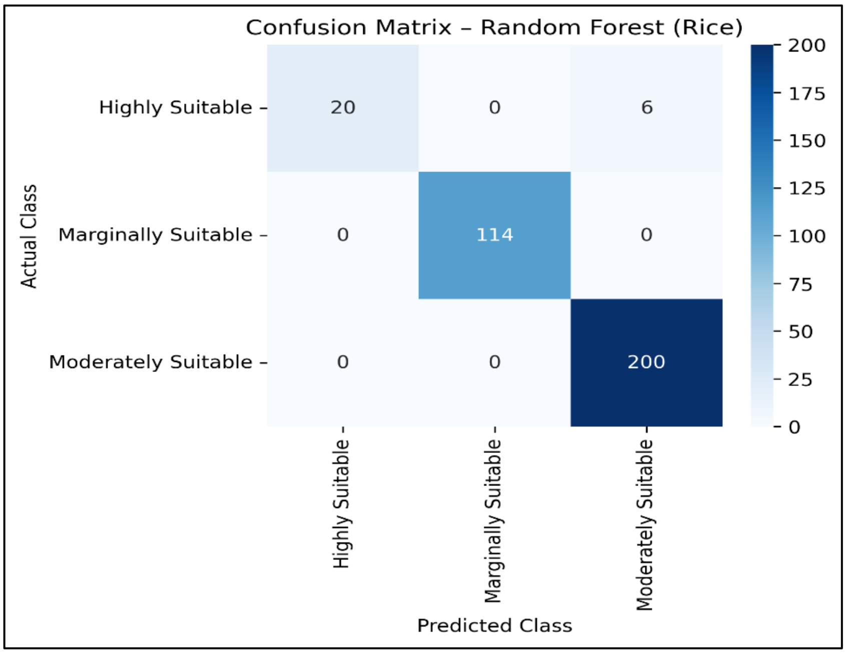


Figure 4.22 : Confusion matrix of the best-performing machine learning model for soil suitability classification

### 4.6.2 Confusion Matrix Visualization (Banana)

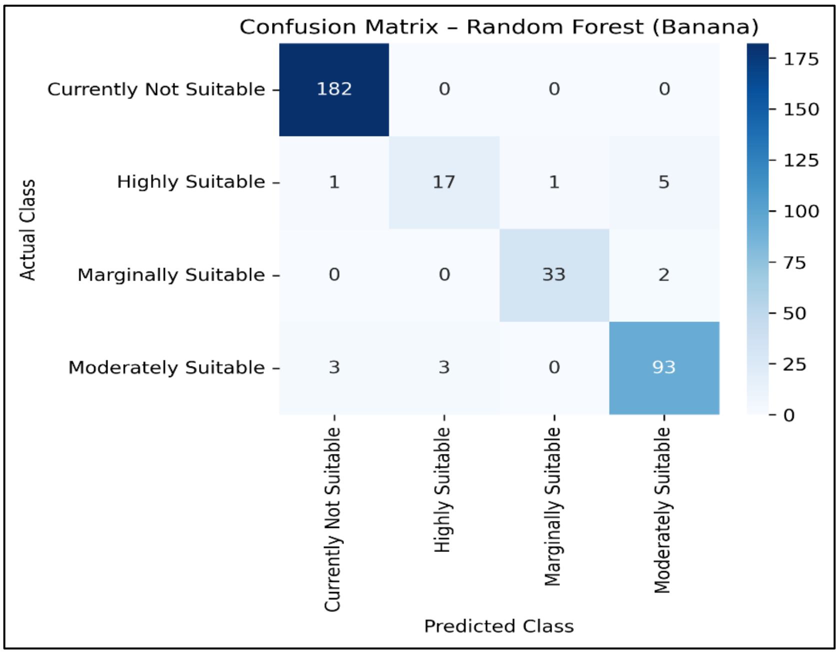


Figure 4.23 : Confusion matrix of the best-performing machine learning model for soil suitability classification

### 4.6.3 Confusion Matrix Visualization (Papaya)

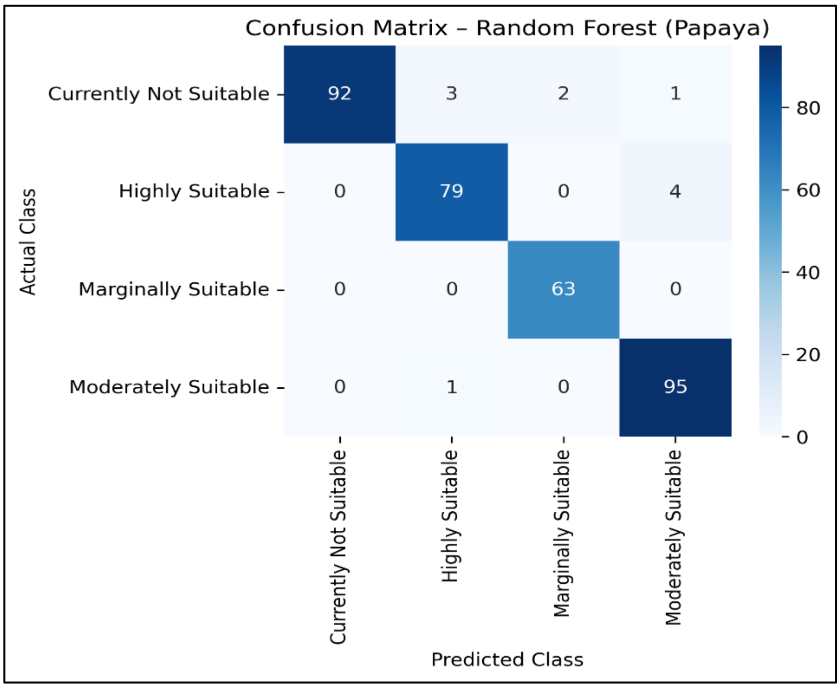


Figure 4.24 : Confusion matrix of the best-performing machine learning model for soil suitability classification

### 4.6.4 Confusion Matrix Visualization (Maize)

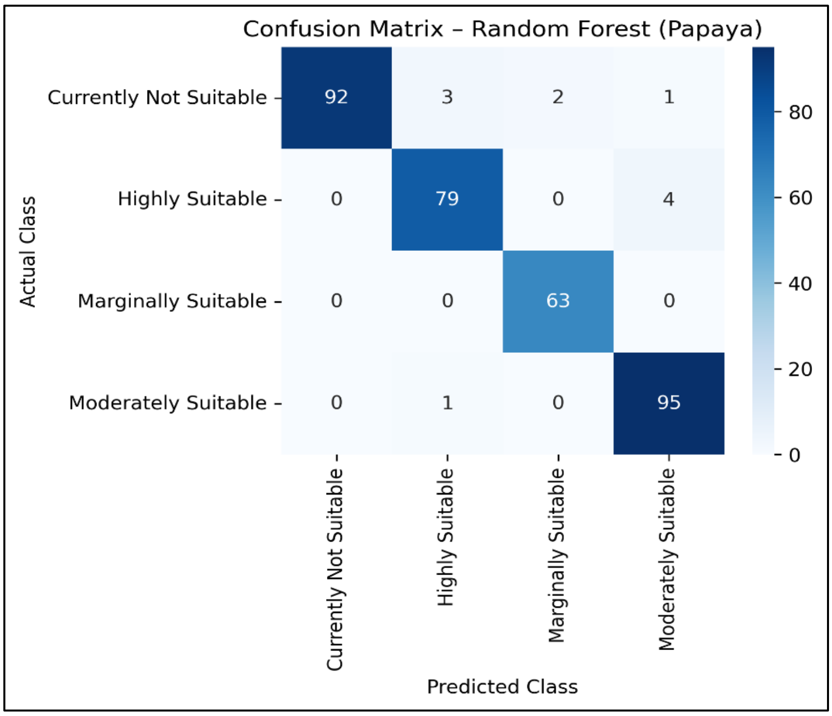


Figure 4.25 : Confusion matrix of the best-performing machine learning model for soil suitability classification

## 4.7 Discussion of Graphical Results

The graphical results offer a good way to compare the model performances for different crops and evaluation metrics. For rice and papaya the Random Forest model was able to achieve higher accuracy, precision, recall and F1-score consistently, which shows that it is able to capture complex and nonlinear relationships among the soil and environmental features. In contrast, for banana cultivation, the Decision Tree model proved to be better for predictions, which might indicate that simpler tree models are capable of modeling the soil requirements of specific crops in some cases. The confusion matrix visualizations further support such findings by displaying higher correct classification rates between selected best performing models. Most of the misclassifications were between adjacent suitability classes, indicating that soil properties changed gradually and therefore there were no sharp class boundaries. Overall, the graphical analysis shows the importance of the crop-specific model selection and validates the effectiveness of the tree-based machine learning models for soil suitability assessment.

## 4.8 Best Model Selection

Based on the comparative analysis of accuracy, precision, recall, F1-score and the results of the confusion matrix, the most appropriate machine learning model was chosen for each crop. The results of the evaluation show that tree-based models are generally superior to linear and probabilistic models, and can adequately deal with complex soil-crop relationships. For rice, and papaya, Random Forest model had the best overall performance based on several evaluation metrics. The ensemble based design allows it to generalize and be robust because of decisions of several trees, and it is suitable in capturing nonlinear interactions between soil and environmental features. In contrast, for banana farming, the Decision Tree model was found to have better performance suggesting that sometimes the soil requirements of crops can be well modeled with simpler trees. These results suggest that the best choice of a model can be crop-specific and that one model is not superior for a general field of crops. Overall, the results corroborate the fact that machine learning-based approaches, especially those based on a tree structure, are very suitable for assessing soil suitability and can serve for data-based decision making in agricultural planning.

## 4.9 Interpretation of Results

The experimental results show that machine learning models are effective in the classification of soil suitability for crop cultivation based on soil and environmental attributes. The performance analysis indicates that the tree-based models typically perform better than the linear and probabilistic models because of their capability of capturing non-linear interactions among multiple parameters of soil. For rice and maize cultivation, there was relatively better accuracy of tree-based approaches, meaning that soil suitability for these crops is related to complex interactions of nutrients, pH, moisture, and climatic factors.

In the case of banana cultivation, it was found that the Decision Tree model performed the best, which suggests that the suitability of banana cultivation is governed by relatively clear rule-based patterns that could be successfully captured through hierarchical decision rules. On the other hand, in case of papaya cultivation, the model performance of Random Forest was better than the other models, thereby showing the need for ensemble learning to model complex and overlapping feature interactions. Overall, the differences in the best-performing models for the various crops confirm that soil suitability assessment is crop-specific and affected by different soil and environmental requirements.

## 4.10 Comparison with Previous Studies

The results of the current study are in line with several published studies which reported good performances of tree-based and ensemble machine learning models in soil and land suitability assessment. Previous studies have found that Random Forest is especially useful in dealing with heterogeneous soil data and nonlinearity in the relationship between soil variables and the suitability of the crop [12], [15]. Similarly, past research has emphasized the use of Decision Tree models in the case where soil suitability is defined by explicit decision rules [18]. The observed crop-specific variation in model performance is consistent with results reported in other related studies in agricultural machine learning, in which different models performed optimally for different crops based on soil characteristics and environmental conditions [20]. These comparisons validate the proposed methodology is reliable and strengthens the use of machine learning methodologies in agricultural decision-making.

## 4.11 Practical Implications of Farmers

The findings of this study have some practical implications to the farmers and agricultural planners in Bogura region. The designed framework of soil suitability classification may assist farmers to select appropriate crops basing on the soil condition in the local area and thus minimize the risk of crop failure and increase the crop productivity. The farmers are able to make a good choice of crop, fertilisation and land management by understanding what class of soils are appropriate and inappropriate with various crops. It is also possible to make sure that the machine learning-based decision support systems can be used to guarantee the effective use of resources and their assistance in sustainable agriculture. Moreover, the choice of the model crop-specific in this work proves the need to offer technological solutions and other solutions applicable in the agricultural environment instead of a generalized model.

# Chapter 5 : Conclusion & Future Work

## 5.1 Summary of Finding

This research examined the soil suitability study for growing crops in the Bogura region using machine learning methods. A massive data set of soil chemical, physical and environmental attributes was analyzed on 4 main crops: rice, maize, banana and papaya. The exploratory data analysis gave information about the distribution of data, variability of features, and categorical information such as soil texture and drainage condition. Four supervised machine learning models including Decision Tree, Random Forest, Logistic Regression and Gaussian Naive Bayes were tested using Cross validation and testing accuracy. The results showed that tree-based models are always better than linear and probabilistic models and the best model depends on the crop.

## 5.2 Conclusion

The study concludes that machine learning techniques provide an effective approach for soil suitability classification and provide a reliable approach in soil suitability classification. The findings affirm that there is no universal model that is good across all crops hence the need to use a model that is specific to the crop. Decision Tree performed best for banana cultivation while the Random Forest algorithm did a better job for papaya cultivation. The proposed methodology represents a framework for data-driven approaches that can be used for precision agriculture and better decision-making in crop planning and land management. The results of this research show the potential of combining machine learning with conventional soil analysis in order to enhance sustainable agricultural development in Bangladesh.

## 5.3 Limitations of my Works

Despite the positive results obtained in this study, some limitations should be recognized. First, the dataset used in this research is relatively small in size, which might limit the generalization ability of the machine learning models. Although cross-validation was used to increase the reliability, a bigger dataset could increase the robustness of the model and accuracy of its predictions. Second, the geographic area for the study is concentrated in the Bogura region, so the study findings are region specific. While the Bogura is important agricultural area, the dataset does not represent soil samples from all sub-districts and agricultural zones in the district. As a result, the variability of the soils over the whole Bogura region may not be totally captured in the analysis. Additionally, the soil suitability labels were produced by predefined criteria for rules, which may not represent the full complexities and variations of experts' judgment about soils. Furthermore, the study is based on a selection of soil and environmental attributes, and other factors that might be influential such as micronutrients and long-term climate variability, as well as management practices, were not considered. These limitations mean that although the proposed approach works within the scope of the data available, care should be taken when generalizing to other data and region when drawing conclusions.

## 5.4 Future Research Directions

Future research can draw on the results of this research and further enhance the soil suitability assessment and generalization of the models. One direction that is important is using a bigger and more diverse data set that covers multiple regions and years. Incorporating the soil data of all the sub-districts in Bogura and adjacent areas would allow the models to incorporate a higher degree of the spatial variability and be robust. The other avenue of interest is the use of group or spatial validation methods (region-wise or block-wise cross-validation). Instead of random data splitting, group-based validation can test model performance on unseen geographical regions, which can provide a more realistic model performance evaluation of agricultural scenarios in the real world. Future research can also consider the concept of hybrid and ensemble modeling that can consider rule-based models, which are combined with machine learning models, or a combination of various classifiers to enhance the prediction accuracy. Further algorithms like the Gradient Boosting, XGBoost or the hybrid Decision Tree Neural Network models might be explored to answer a more intricate relationship between soils and crops. Moreover, it may be possible to incorporate Geographic Information Systems (GIS), remote sensing data, and real time environmental data that would allow spatial mapping of the suitability of the soil as well as to provide location-specific recommendations. The application of explainable AI (XAI) techniques could possibly further improve model transparency to enable farmers and agricultural experts to better understand the reason for suitability predictions. In general, the following future directions have a potential to make decision support systems that are more reliable, scalable and user-friendly to farmers in order to develop sustainable agricultural systems.

# References


[1] S. R. A. a. K. M. Rozario, "Insights on land use, agriculture and food security in Bangladesh: way forward with climate change and development.," 2021.

[2] R. K. S. R. D. J. P. a. S. T. Timilsina, "A comprehensive review of the soil health status for enhancing agricultural sustainability," Frontiers in Environmental Science, vol. 13, p. 1548095, 2025.

[3] A. A. S. R. R. S. B. A. D. A. a. A. D. S. S. S. B, "Machine Learning Techniques in Crop Recommendation based on Soil and Crop Yield Prediction System – Review," in 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), Karkala, India, 2022.

[4] M. S. M. Z. A. N. H. Z. L. L. W. Z. F. A. A. E. A. I. M. I. K. V. R. a. J. H. Awais, "AI and machine learning for soil analysis: An assessment of sustainable agricultural practices," Bioresources and Bioprocessing, vol. 10, no. 1, p. 90, 2023.

[5] ,. S. B. U. Chetana Shivanagi, "Machine Learning in Agricultural Soil Analysis: A Comprehensive Review on Classification and Fertility," International Journal of Engineering Research and Applications, vol. 15, no. 6, pp. 95-103, 2025.

[6] Wikipedia, "Agriculture in Bangladesh," 05 July 2025. [Online]. Available: https://en.wikipedia.org/.

[7] M. S. A. ,. M. Z. A. M. A. I. M. A. A. Md. Tanbir Hasan, "Four Crop-based Cropping Pattern Studies for Increasing Profitability and Productivity in Bogura Region of Bangladesh," Journal of Contemporary Agriculture and Bioscience, vol. 1, pp. 1-4, 2022.

[8] shaikh-abdullah, "Real challenges in agriculture: What's blocking Bangladesh's farmers from expected gains," 02 February 2025. [Online]. Available: http://www.tbsnews.net.

[9] M. N. M. a. F. T. a. A. A. O. a. Y. Ali, "Leveraging modern technologies in agriculture: Recent progress and prospects in Bangladesh," Plant Trends, vol. 3, no. 2, pp. 13-32, 2025.

[10] M. A. I. M. F. F. Z. S. M. M. A. M. Rakibul Alim, "Determinants of Cropping Intensity in Drought Prone Areas of Bogura," JSAT-International Journal on Science and Technology, no. 5, 2025.

[11] S. A. Anupama Sapkota, "Soil- Definition, Composition, Properties, Types and Uses," 25 March 2022. [Online]. Available: https://microbenotes.com/. [Accessed 03 January 2026].

[12] Y. a. M. M. Ali, "Leveraging Modern Technologies in Agriculture: Recent Progress and Prospects in Bangladesh," Bangladesh Society for Microbiology, Immunology, and Advanced Biotechnology, vol. 3, no. 2, pp. 13-32, 2025.

[13] S. K. B. N. M.-R. M. P. P. Z. Z. C. A. J. A. M. H. D. l. R. D. &. J. L. Abd-Elmabod, "Assessment of Soil Suitability for Improvement of Soil Factors and Agricultural Management. Sustainability," Sustainability (MDPI), vol. 11, no. 6, p. 1588, 2019.

[14] A. Y. Ö. B. T. Halil Akıncı, "Agricultural land use suitability analysis using GIS and AHP technique," Computers and Electronics in Agriculture, vol. 97, no. 0168-1699, pp. 71-82, 2013.

[15] R. A. A. A. E. B. M. M. I. E. S. M. D. E. K. a. M. S. S. El Behairy, "Assessment of soil capability and crop suitability using integrated multivariate and GIS approaches toward agricultural sustainability," Land, vol. 11, no. 7, p. 1027, 2022.

[16] M. S. A. &. A. S. AbdelRahman, "Assessment of land suitability using a soil-indicator-based approach in a geomatics environment," Scientific Reports, vol. 12, p. 18113, 2022.

[17] K. T. P. J. S. T. S. M. R. A. J. C. T. &. D. A. J. P. Soberano, "Predictive soil-crop suitability pattern extraction using machine learning algorithms," International Journal of ADVANCED AND APPLIED SCIENCES, vol. 10, no. 2, 2023.

[18] P. K. N. R. M. S. P. K. &. V. K. N. Tripathi, "Applications of Machine Learning in Agriculture.," IGI Global Scientific Publishing, pp. 99-118, 023.

[19] N. G. H. &. K. T. Agrawal, "Agricultural land suitability classification and crop suggestion using machine learning and spatial multicriteria decision analysis in semi-arid ecosystem," Journal of Environmental Health Science and Engineering, vol. 27, p. 13689–13726, 2025.

[20] K. K. A. K. A. B. E. V. Sergei Kurashkin, "Analysis of soil suitability for agricultural needs using machine learning methods," BIO Web of Conferences, vol. 149, p. 01053, 2024.

[21] N. P. R. K. P. P. a. R. Y. A. S. Arun, "Machine Learning-Based Crop Recommendation System," in Intelligent Systems and Machine Learning Conference (ISML), Hyderabad, India, 2024.

[22] R. B. P. Jadhav, "A Machine Learning Based Crop Recommendation System: A Survey.," Journal of Algebraic Statistics, vol. 13, no. 1, p. 426, 2022.

[23] S. K. M. N. A. H. M. Ismaili, "Assessment of Soil Suitability Using Machine Learning in Arid and Semi-Arid Regions.," Agronomy, vol. 13, no. 1, p. 165, 2023.

[24] K. S. P. C. S. e. a. Kumar, "State Parameter-Based Soil Liquefaction Assessment Using Subset Simulation and Machine Learning Approaches," Arabian Journal for Science and Engineering, 2025.

[25] D. E. V. O. A. K. A. P. Svetlana Kukartseva, "Machine learning in soil science for prediction and management of biological activity for sustainable land use," in E3S Web of Conferences, 2024.

[26] S. A. R. S. P. R. Mariammal Ganesan, "Prediction of Land Suitability for Crop Cultivation Using Classification Techniques," Agriculture, Agribusiness and Biotechnology, vol. 64, no. 1678-4324, 2021.

[27] K. N. L. R. R. Taghizadeh-Mehrjardi, "Land Suitability Assessment and Agricultural Production Sustainability Using Machine Learning Models," Agronomy, vol. 10, no. 4, p. 573, 2020.

[28] M. M. Rahman, "GIS based allowable bearing capacity thematic maps of shallow foundation for Bogura District, Bangladesh," International Journal of Engineering and Geosciences, vol. 10, no. 3, p. 329–338, 2025.

[29] M. H. a. M. Rahman, "Seismic microzonation and probability of ground failure assessment caused by liquefaction for Bogura District, Bangladesh," Journal of Rehabilitation in Civil Engineering, vol. 13, p. 218–242, 2025.