PREDICTION OF CROPS CULTIVATION CLASSIFYING SOIL THROUGH ARTIFICIAL INTELLIGENCE

FARDOWSI RAHMAN

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

PREDICTION OF CROPS CULTIVATION CLASSIFYING SOIL THROUGH ARTIFICIAL INTELLIGENCE

FARDOWSI RAHMAN

This Thesis Report submitted in partial fulfillment of the requirements
for the award of the degree of
Master of Science in Information and Communication Engineering

Department of Information and Communication Engineering
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

NOVEMBER 2022

SUPERVISOR'S DECLARATION

We hereby declare that we have checked this thesis and in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science in Department of Information and Communication Engineering of Noakhali Science and Technology University.

_____

Name of Supervisor: DR. MD. ASHIKUR RAHMAN KHAN
Position: PROFESSOR & CHAIRMAN
Date:

_____

Name of Co-Supervisor: ZAYED-US-SALEHIN
Position: ASSISTANT PROFESSOR
Date:

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged. This thesis report has not been accepted for any degree and is not concurrently submitted for the award of other degrees.

_____

Name: FARDOWSI RAHMAN

ID Number: BKH2011MSc102F

Date:

ACKNOWLEDGEMENTS

I am grateful and would like to express my sincere gratitude to my supervisor Professor Dr. Md. Ashikur Rahman Khan for his germinal ideas, invaluable guidance, continuous encouragement, and constant support in making this research possible. He has always impressed me with his outstanding professional conduct, his strong conviction for science, and his belief that an MSc program is only the start of a life-long learning experience. I appreciate his consistent support from the first day I applied to the graduate program to these concluding moments. I am truly grateful for his progressive vision about my training in science, his tolerance of my naive mistakes, and his commitment to my future carrier. I also would like to express very special thanks to my co-supervisor Zayed-Us-Salehin for his suggestions and co-operations throughout the study. I also sincerely thank you for the time spent proofreading and correcting my many mistakes.

My sincere thanks go to all my lab mates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways and made my stay at NSTU pleasant and unforgettable.

I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream, and sacrifice throughout my life. Special thanks should be given to my committee members. I would like to acknowledge their comments and suggestions which were crucial for the successful completion of this study.

ABSTRACT

The field of agriculture is one of the significant sources of people surviving in Bangladesh. Soil is one of the prominent components in agriculture for cultivating crops. There are different types of soil, and there are various features of each soil. Based on the soil types, several kinds of crops are grown. However, the agricultural land is decreasing day by day while the number of people is increased. On the other hand, farmers do not know what crops can be cultivated in their land. There are many lands in our country where the people think that cultivation is not possible at all. Due to these reasons, many lands are uncultivated in our country every year. Considering these issues, this research work aims to predict the crops cultivation for a particular soil so that the people can select and cultivates crops undoubtedly and correctly in their land. Consequently, all the agricultural land will be cultivated and we will be able to provide our own food. The soil types are determined first based on its distinct characteristics for a particular zone then the crops cultivation is ascertained according to the soil types. Artificial intelligence is an emerging and challenging research field in agriculture. Data is collected from different zone of Noakhali district. Then preprocessed data is splitted into training and testing. Several artificial intelligence techniques such as Support Vector Machine, Decision Tree, Multilayer Perceptron Neural Network, Random Forest, Logistic Regression and Naïve Bayes are used for prediction. For soil classification, Random Forest algorithm is selected on the basis of accuracy, precision, recall, f1-score and ROC. The accuracy of Random forest is 95.15% which is higher than others algorithm. And for crops prediction, Support vector machine is selected. The accuracy of support vector machine is 91.17% which is higher than other algorithms. Thus, this research work will become an effective and intelligent tool to enhance cultivation in our country and can play a significant role in economy development of Bangladesh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Meaning
μ	Mu
λ	Lambda
α	Alpha

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
LS- SVM	Least Square – Support Vector Machines
GIS	Geographic Information System
IDE	Integrated Development Environment
MLP	Multilayer Perceptron
MAPE	Mean Absolute Percentage Error
MFNN	Multilayer Feed-Forward Neural Networks
NNM	Neural Network Model
NLRM	Nonlinear Regression Model
SVR	Support Vector Regression
ADTree	Alternating Decision Tree
REP	Reduced-error Pruning
CART	Classification And Regression Tree
RS	Rank Search
KNN	K – Nearest Neighbor
MMRE	Mean Magnitude of Relative Error
MSE	Mean Squared Error
ROC	Receiver Operating Characteristic
LR	Logistic Regression
NB	Naïve Bayes
RF	Random Forest
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
TPR	True Positive Rate
FPR	False Positive Rate
DT	Decision Tree

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The introductory study of this thesis is carried out in this chapter. Since this research focuses on the prediction model for crop cultivation based on soil classification; thus, background analysis on these sectors is carried out first. The problem statement is then introduced, and the research objectives are described on this basis. At the end of this chapter, the outline of this thesis is presented.

1.2 BACKGROUND

Soil is a significant part of successful agriculture that we use to grow crops. It is a composition of organic materials, minerals, organisms, air, and water that provide the medium for plant growth. Soil type gives information about morphological, physical, and chemical characteristics and mineral content. There are different types of soil, and each has various features. Based on these multiple features, several kinds of crops are grown. The characteristics and behavior of different kinds of soil help to understand which crops grow in a particular soil type. Soil plays a significant role in crops ecosystem by providing nutrients essential for the growth of agricultural and horticultural crops. Fertile soil is rich in nutrients and highly suitable for agriculture. Existing water in the soil serves as the primary nutrient base for healthy crops. Rich soil contains pH and primary plant nutrients such as nitrogen, phosphorus, and potassium. Also, the soil has organic matter and minor nutrients

that help plant growth. Some of the functions associated with soil include nutrient cycling, water regulation, plant growth, recycling organic wastes and nutrients, water supply and purification, flood control, food production, and other normal processes in the ecosystem that improve the agricultural economy. Essential benefits of soil include the natural protection of seeds and plants, dispersal and germination of seeds within the soil ecosystem, the physical support system for plants, and retaining and delivering nutrients to crops.

Agriculture seeks to increase the crop yield and the quality of the crops to sustain human life. However, nowadays, people tend to take more immediately rewarding jobs. There are fewer and fewer people involved in crop cultivation. In addition, the continuous increase in human population makes the cultivation of crops at the right time and right place even more important. Because the climate is changing and the shifts from regular weather patterns are more frequent than before industrialization. Food insecurity is a problem that cannot be avoided, and humans must use new innovative technologies to make use of existing soil, water, and air conditions to obtain larger crops. The knowledge gap between traditional cultivating methods and new agricultural technologies is overcome if the system can be designed for the interactive impact of climate factors. Climate change definitely affects local and world food production, so crop cultivation requires new hypotheses for climate change studies, scenarios for climate change adaptation, and policymakers that can limit the devastating effects of weather on food supply. Farmers have certain expectations of how much crop will get and make financial decisions based on it. In the past, farmers used to predict crop cultivation based on their own experience and observed weather conditions. Weather, pests, and harvest operations may be kept as a reference for future years. Currently, the software can augment traditional knowledge. Maintaining accurate data is an essential aspect of agricultural risk management.

The physical and chemical properties of the soil always play an important role in precision and smart farming. Soil with essential nutrients is capable of supporting crop cultivation. But some nutrient level in the soil is declining because of the usage of more fertilizers. Due to this reason, crop production is falling [1]. Soil texture is the primary physical property of the soil and plays an essential role in farming. It defines the fraction of sand, silt, and clay present in a soil sample. Other related agriculture cultivating properties, such as soil water content, plant growth, and crop selection, relate to these three end members. Most rural farmers have

yet to learn about the texture and characteristics of the soil. Sandy soil is low in water-holding capacity and organic matter. The water-holding capacity of silt soil is more as compared to sandy soil. Clay soil shows a high water-holding capacity, high plasticity, and thickness, whereas sandy soils are conspicuous by the absence of these properties. The water holding capacity of sandy loam is low as compared to loam and silt clay soil. It indirectly affects the plant growth, the weight of the plant, and the percentage of chlorophyll. For example, chlorophyll contents of leaves are dropped in sandy loam as compared to silt clay loam. In the field study, it is found that rural farmers do not have any knowledge about soil classification. They are doing their farming without proper soil testing and are unaware of selecting soil and seed. It indirectly affects the overall growth of the plant.

The soil classification methods were developed step by step with different ideas. In the early days, the soil was classified depending on its composition and weight related to the total mass. Then soil was classified depending on texture which was finally developed into the triangular classification diagram method. This is shown in the following figure.

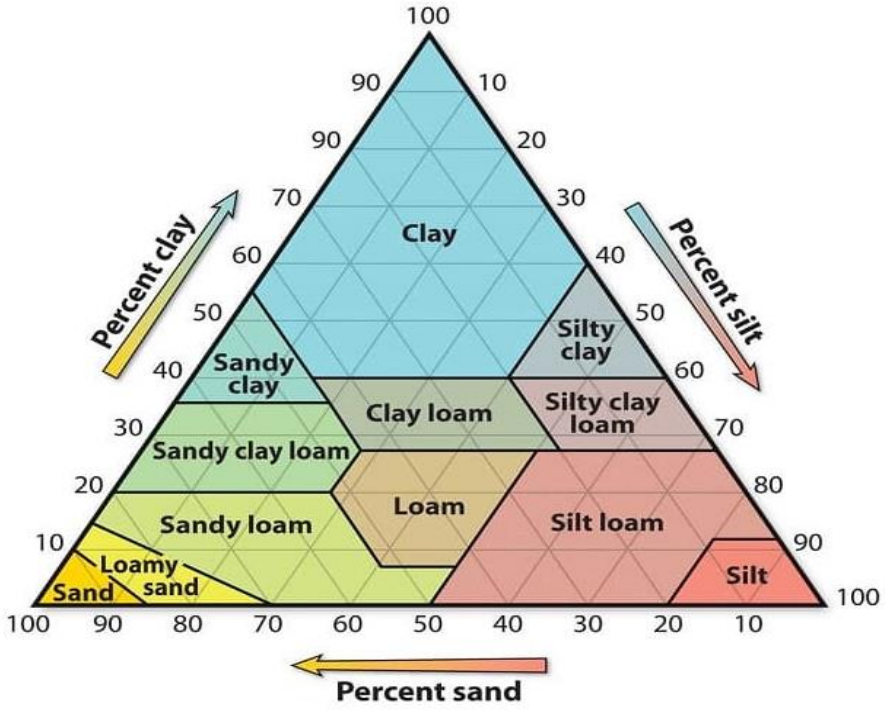


Figure 1.1: Triangular soil classification diagram

The properties of soil are determined by its composition. Here, the percentages of soil compositions such as sand, silt, and clay obtained from laboratory tests are used to determine the textural soil classification. In this process, a textural triangle chart is used to determine

the classification. In that chart, the corners of the triangle indicate the hundred percent of each composition – sand, silt, clay, and the twelve textural soil classes were noted within the triangle with the indication of thick line separation between those classes. It's simply clear that the summation of the percentage of sand, silt, and clay should give 100%. The percentage of sand, silt, and clay for the different types of soil are shown in the following table.

Table 1.1: Percentage of Sand, Silt, and Clay for different soil types.

Soil class	**Sand (%)**	**Silt (%)**	**Clay (%)**
Clay	0-45	0-40	40-100
Silt	0-20	80-100	0-12
Sand	85-100	0-15	0-10
Loam	23-52	27-50	7-27
Loamy Sand	70-85	0-30	0-15
Sandy Loam	43-85	0-50	0-20
Silt Loam	0-50	50-100	0-27
Sandy Clay Loam	45-80	0-27	20-35
Silty Clay Loam	0-20	40-71	27-40
Clay Loam	20-45	17-52	27-40
Sandy Clay	45-65	0-20	35-55
Silty Clay	70-88	0-30	10-14

Agriculture is the backbone of the Bangladesh economy, accounting for about half of the national income and slightly more than half of the merchandise exports. Most of the people of Bangladesh depend on agriculture as a source of food requirements. This implies that progress in reducing poverty, malnutrition, and food insecurity in Bangladesh depends

greatly on the performance of the agricultural sector. The main cash crops grown in the country include tea, jute, cotton, tobacco, sugarcane, coffee, silk, and rubber. At the same time, the main food crops include maize, rice, wheat, pulses, oil seeds, tomato, potato, fruit, and vegetables. There are many coastal lands of our country from which farmers do not get crops as expected. Most people think that cultivation is impossible in the coastal region. Due to these reasons, many lands are uncultivated in our country every year.

The main aim of agriculture is to grow crops. The agriculture factors such as rain, weather, soil type, pesticides, and fertilizers are the things to increase production. The population is increasing day by day, and with an increased population, the need for food is also increasing. Agriculture gives most people food and fabric. In the global economy, agriculture plays a vital role. Artificial neural networking techniques in agriculture are used to increase crop production. This research aims to classify the soil and predict the type of crops that can be cultivated in a particular soil type using artificial neural networking techniques. We provide a pathway for the farmers so that identification of soil and suggested crop production based on soil category will be easy for them.

Traditional methods used by farmers are insufficient to serve the huge demand, so they have to hamper the soil by using harmful pesticides in an intensified manner. This affects the agriculture practice a lot, and in the end, the land remains barren with no fertility. Machine learning in agriculture is used to improve the product quality of the crops in the agriculture sector. Machine Learning is the scientific field that gives the ability to the machine to learn without the intervention of human beings [2]. Farmers do not have the proper knowledge about the soil properties and usage of the optimum quantity of inputs in their crops which are essential for increasing productivity. The human race depends more on farm products for their existence than anything else since food and clothing – the prime necessities are products of farming. Even for industrial prosperity, farming forms the basic raw material. Most of the farmers are cultivating their lands just for the sake of cultivation without using full scientific and technological backup. They are satisfied with whatever they are getting finally [3].

Rapid and accurate crop type mapping is significant for agricultural management and sustainable development. Agriculture is one of the vital occupations practiced in Bangladesh. It is the most significant economic sector and plays the most important role in the overall development of the country. Most of the land in the country is used for agriculture in order

to suffice the needs of 16 crores of people. Thus adopting new agriculture technologies is very important. This will lead our country's farmers towards economic profit and a standardized life. Previous crop and yield predictions were based on the farmer's experience in a particular location. They will prefer the prior or neighborhood or more trend crops in the surrounding region only for their land, and they don't have enough knowledge about soil nutrient content such as nitrogen, phosphorus, and potassium in the land. This is the current situation without the rotation of the crop and applying an inadequate amount of nutrients to the soil, which leads to a reduction in yield and soil pollution (soil acidification). Considering all these problems takes into the account we designed the system using artificial neural networking techniques for the betterment of the farmer. Machine Learning (ML) is a game changer for the agriculture sector. Machine learning is a part of artificial intelligence and has emerged together with big data technologies and high-performance computing to create new opportunities for data-intensive science in the multi-disciplinary agri-technology domain. In the Agriculture field, machine learning, for instance, is not a mysterious trick or magic; it is a set of well-defined models that collect specific data and apply specific algorithms to achieve expected results.

Estimating agricultural crop cultivation prior to harvest is an important issue in agriculture, as the changes in crop production from year to year influence international business, food supply, and global market prices. Also, early prediction of crop cultivation provides useful information to policy planners. Appropriate prediction of crop productivity is required for efficient planning of land usage and economic policy. In recent times, forecasting of crop productivity at the within-field level has increased. The most influencing factor for crop productivity is weather conditions. If the weather-based prediction is made more precise, then farmers can be alerted well in advance so that the major loss can be mitigated and would be helpful for economic growth. The prediction will also aid the farmers in making decisions such as the choice of alternative crops or discarding a crop at an early stage in case of critical situations. Further, predicting crop production can facilitate the farmers to have a better vision of the cultivation of seasonal crops and their scheduling. Thus, it is necessary to simulate & predict crop cultivation for efficient crop management and expected outcome. Machine learning techniques might be efficient for predictions as there exists a non-linear relationship between crop cultivation and the factors influencing crop.

1.3 PROBLEM STATEMENT

Due to the lack of proper agricultural training and education, the failure of farmers in the field of agriculture is increasing day by day. Often farmers are unaware of the crop growth which suits the soil quality, nutrients, and composition—as a result, they are hampered in crop cultivation. On the other hand, the agricultural land is decreasing while the number of people is increasing. So this is a challenging task to provide vast amounts of food. But traditional methods used by the farmers are not working perfectly to produce vast quantities of food. Also, an unskilled farmer using harmful pesticides in the wrong manner results in hampering the soil. This affects agriculture a lot, and the land remains no fertility. Additionally, an inexpert farmer is unaware about the crop cultivation based on soil type.

There are many uncultivated lands in the coastal areas of Bangladesh. Most of the farmers do not cultivate the crop according to the quality of the soil, and due to this causes; they do not get the expected results. So, this thesis work will classify the soil and predict the cultivation of crops that can be cultivated in a particular soil type. And we will provide a pathway for farmers so that they can easily identify the soil and find out what crops are grown in certain soils.

The current population of Bangladesh is 168,426,046 based on Worldometer elaboration of the latest United Nations data. The population density in Bangladesh is 1265 per Km^2. The total land area is 130,170 Km2 (50,259 sq. miles). Agricultural land is decreasing day by day. On the other hand, the population is increasing day by day, and with an increase in population, the need for food is also increasing. Most of the people of Bangladesh depend on agriculture as a source of food requirements. Factors such as reducing poverty, malnutrition, and food insecurity in Bangladesh depend greatly on the performance of the agricultural sector.

Crops production in Bangladesh is constrained every year by challenges, such as a) Loss of Land, b) Population Growth, c) Climate Changes, d) Inadequate Management Practices, e) Unfair Prices of Products, f) Inadequate credit support to farmers, etc. The twin problem of land loss and population growth needs to be addressed simultaneously to ensure sustainable crop production. By producing more crops on a small area of land, the food supply for a large number of people will be ensured.

Noakhali is a coastal region of Bangladesh. There are many uncultivated lands in Noakhali. The farmers of this district are not very educated. They often do not understand what crops are grown in the coastal lands. That is why they do not know how to produce more crops from the land. Many farmers lose interest in cultivating crops as they do not get the expected results. Predicting the cultivation of a crop for a particular area is a very complex and challenging task as it is influenced by several parameters starting from the type of soil and different parameters. Again, the crops also depend on the type of method used by the farmers from field to field, so predicting the crop cultivation from its parametric point of view is a very critical task. With the increasing population, there is a considerable demand for crops throughout the globe; hence, farmers need to be aware of the type of crop that can be treatable for their geographical location and soil type. This is required to waive the economic pressure on the farmers. Therefore, it is essential to provide accurate, timely-based information to the farmers based on the different parameters of the soil, that's helping them to make the best decision for crop cultivation in the soil.

Plants require a specific kind of soil to grow properly. For example, a crop like rice requires a very high moisture content in the soil. By contrast, a crop like wheat prefers loamy soil rich in organic material. Improper soil analysis can lead to crop decline or outright failure. Due to improper soil analysis, farmers do not get crops as expected. A soil test is carried out to identify the nutrient content, composition, and other components contained in the soil. Soil tests are mainly conducted to measure the fertility and other defenses present in the soil. Soil classification techniques follow the existing knowledge and practical circumstances. On the land surfaces of the earth, the classification of soil creates a relation between soil samples and various kinds of natural entities.

The motto of this work is to overcome the difficulty in the agricultural sector and predict crop cultivation based on soil type using artificial neural networks.

1.4 RESEARCH OBJECTIVES

There are some objectives of this research. This research paper is worked on crop cultivation prediction based on soil types using artificial neural network techniques. The main purposes of this research are the following:

i. To analyze the chemical composition of different types of soil.
ii. To classify soil using artificial intelligence.
iii. To predict crop cultivation for specific areas based on soil classification.
iv. To make a comparison among distinct techniques.
v. To find an optimal technique for predicting crop cultivation through classifying soil.

1.5 MOTIVATION FOR THIS RESEARCH WORK

Soil properties play critical roles and have an influence on agricultural engineerings, such as soil improvement, land consolidation, management of drainage, soil erosion, and irrigation. Knowledge of the variations in the soil properties could provide valuable information to design a more rational use and management plan, especially in cultivated areas. Climate, biota, and geological history are the critical factors affecting the chemical and physical properties of soil at larger scales (e.g., regional and continental scales), while human activities and topography may be the dominant factors controlling the properties of soil at smaller scales. Topography has significant impacts on runoff, drainage, and soil erosion and hence on soil development [4].

Soil classification deals with the systematic categorization of soils based on distinguishing characteristics as well as criteria that dictate choices in use. Soil classification is a dynamic subject, from the structure of the system itself to the definitions of classes and, finally, the application in the field. Soil classification can be approached from the perspective of soil as a material and soil as a resource. The most common engineering classification system for soils is the Unified Soil Classification System (USCS) [5]. The USCS has three major classification groups: (1) coarse-grained soils (e.g., sands and gravels); (2) fine-grained soils (e.g., silts and clays); and (3) highly organic soils (referred to as "peat"). The USCS further subdivides the three major soil classes for clarification. A full geotechnical

engineering soil description will also include other properties of the soil, including color, in-situ moisture content, in-situ strength, and somewhat more detail about the material properties of the soil that is provided by the USCS code [5].

Soil formation is a long-term process, and diverse soils are formed in different localities due to various soil-forming factors over the landscape. Soil classification plays a critical role in various aspects of agricultural engineering. Physico-chemical parameters play an important role in soil classification. In this paper, K. Radhika et al. [6] present a comprehensive classification model for soil texture classification by using Linear Discriminant Analysis (LDA). They took the Physico-chemical properties of the soil, which include soil moisture, temperature, electrical conductivity, pH, organic carbon, available nitrogen, available phosphorus, and potassium, as independent variables, while the soil type was taken as the dependent variable. Feature selection is employed using the Boruta algorithm. The performance of the proposed classification model is evaluated and expressed in terms of overall accuracy and kappa coefficient. Results show that the average prediction accuracy and kappa coefficient of the proposed model is 80.3% and 0.814, respectively, indicating that the model can be used effectively for soil classification for a set of suitable dependent variables.

As a remark, most of the existing works consider systematic categorization of soils. Few researchers focus on chemical features based on soil classification. The main reason for the small amount of research work and accuracy is the less availability of existing datasets on the chemical features-based soil classification.

The purpose of this work is to categorize the soil and to predict the crops which can be cultivated in the particular soil.

1.6 SCOPE OF THE STUDY

In this paper, Artificial intelligence techniques are used to estimate the values of soil properties and soil type classification based on known values of particular chemical and physical properties in sampled profiles. The data used in this research work is collected from the different zone of the Noakhali District. Noakhali is a coastal region of Bangladesh. There are many uncultivated lands in Noakhali. The farmers of this district are not very

educated. They often do not understand what crops are grown in the coastal lands. That is why they do not know how to produce more crops from the land. Many farmers lose interest in cultivating crops as they do not get the expected results. This research work has been done to classify soil based on its distinct characteristics for a particular zone then the crop cultivation will be ascertained according to the soil types. The required tools carry out to work this work is:

- Python 3.7
- Anaconda Navigator
- Python Packages
- Strong CPU needed, and further processing needs GPU

1.7 THESIS OUTLINE

This thesis paper consists of 5 chapters, and these chapters are organized as follows:

Chapter 1 – Introduction: Performs the background of soil classification and crop cultivation. In addition problem statement, objectives, motivation, and the scope of the study are presented.

Chapter 2 – Literature Review: An illustrative review of previous related work of this thesis is described in this chapter. It gives information about soil classification, crop cultivation prediction, and techniques.

Chapter 3 – Methodology: Presents approaches that are used in the system. And also, the dataset details what function and algorithm are used are discussed.

Chapter 4 – Results Analysis: In this chapter, the results are described.

Chapter 5 – Conclusion and Future work: This chapter describes the conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Estimating agricultural crops cultivation prior to harvest is an important issue in agriculture. Early prediction of crop cultivation provides useful information to farmers. Appropriate prediction of crop productivity is required for efficient planning of land usage and economic policy. In this chapter, we take a few papers related to soil analysis and crop cultivation prediction using Artificial Intelligence techniques. Different past work are related with soil classification, crop prediction methods are discussed in literature review.

2.2 LITERATURE REVIEW

The literature review is performed on the works which relates on these sectors. In this section, we take some significant previous work related with crops cultivation prediction based on soil classification.

Cai Simina et al. [7] describes an effort to apply an improved support vector machine classifier to classify salt-affected soil. They used the support vector machine with texture features to extract thematic information for salt-affected soil. The SVM classification was conducted using a combination of multi-spectral features and texture features as the data source. They used mean, variance and homogeneity features, which were the best texture features, to improve the classification. In addition, they provided a contrast between the proposed SVM method and other SVM methods. SVM classification used here can effectively extract salinization soil thematic information for the Yinchuan Plain.

Rapid and accurate crop type mapping is of great significance for agricultural management and sustainable development. Han Gao et al. [8] proposed a pairwise proximity function support vector machine (ppfSVM) classification method with the time-weighted shape DTW (TWshapeDTW) alignment. With 42 Sentinel-1 dual-polarization SAR images located in Gansu Province, China, the crop classification maps in 2018 and 2019 are generated respectively. The proposed method obtains the overall accuracies of more than 90% in both two years, and the changes of crop types from 2018 to 2019 are also in line with the actual crop rotation. Compared with traditional classification methods (SVM and the NN method with the DTW alignment), it is found that the proposed method has higher overall accuracy (OA) and better robustness in the case of small number of samples.

In terms of soil mechanics and geotechnical engineering, the ground is basically uncertain and heterogeneous (inhomogeneous) because the inside cannot be visually inspected and easy predictions are not possible. Shinya Inazumi et al. [9] used image processing for the classification of soil. Deep learning was performed with a model using a neural network in this study. For three types of soil, namely, clay, sand, and gravel, an AI model was created that was conscious of the practical simplicity of the soil images. It was shown that artificial intelligence, along with deep learning, can be applied to soil classification determination by performing simple deep learning with a model using a neural network.

There is a works of Rahul Agarwal et al. [10] explain a new optimal bag-of-features based automated soil prediction method is used to categorize the soil images into a set of predefined categories. They used an enhanced Henry gas solubility optimization (HGSO) to generate optimal visual words. To improve exploration behavior of HGSO, chaotic map based Henry constant is introduced. The proposed variant of HGSO performs effectively on considered benchmark functions with better convergence precision. Moreover, the proposed classification method has been tested and validated on soil image dataset having seven classes and experimental results show that the proposed method returns 80.58% accuracy which is better than other meta-heuristic based classification methods. M.S. Sirsat et al. [11] used the chemical soil measurements to classify many relevant soil parameters: village-wise fertility indices of organic carbon (OC), phosphorus pentoxide (P2O5), manganese (Mn) and iron (Fe); soil pH and type; soil nutrients nitrous oxide (N2O), P2O5 and potassium oxide (K2O), in order to recommend suitable amounts of fertilizers; and preferable crop. To

classify these soil parameters allows to save time of specialized technicians developing expensive chemical analysis. These ten classification problems are solved using a collection of twenty very diverse classifiers, selected by their high performances, of families bagging, boosting, decision trees, nearest neighbors, neural networks, random forests (RF), rule based and support vector machines (SVM). The RF achieves the best performance for six of ten problems, overcoming 90% of the maximum performance in all the cases, followed by adaboost, SVM and Gaussian extreme learning machine.

Another works of Gayantha R.L. Kodikara et al. [12] proposed the Lunar Soil Characterization Consortium (LSCC) dataset to train and assess the predictive power of classification models based on their type (Mare soil and Highland soil), particle size, maturity, and the dominant type of pyroxene (High-Ca and Low-Ca). Nine Machine Learning (ML) algorithms including linear methods, non-linear methods, and rule-based methods (three from each) were selected, representing a range of characteristics such as simplicity, flexibility, computational complexity, and interpretability along with their ability to handle different types of data. Fifteen spectral parameters were initially introduced to the models as input features and a maximum of four features was selected as the best feature combinations to classify lunar soils based on their types, particle size, maturity, and the type of pyroxene. The Support Vector Machine with radial basis function (svmRadial) and the penalized logistic regression model (glmnet) performed well for all target variables with high accuracies. Band depths and Integrated band depths at 1 μm, 1.25 μm and 2 μm, band position of the 1 μm band, along with four band ratios (band tilt, band strength, band curvature, and olivine/pyroxene) are important features for classifying soil type, grain size, maturity, and type of pyroxene from reflectance spectra. This study shows that proper preprocessing and feature engineering techniques are crucial for high performance of the predictive models. Javed Mallick et al. [13] is to examine the efficacy of a novel and unique approach to automated complex wetland mapping in Bangladesh's Sunamganj District, using combined image fusion with machine learning techniques. Four advanced machine learning classifiers, i.e. random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), and decision tree (DT), were used to classify the complex wetland image. To validate the identified wetland maps, the Kappa coefficient and root mean square error (RMSE) have been used. The accuracy of all four machine learning classifiers was higher on the fused

images than on the MS image. The proposed fused ANN-based hybrid model outperformed all fused and MS classification models in terms of accuracy (kappa: 89.4% and RMSE: 0.13). Ritesh Dash et al. [14] focused on establishing and interrelating the micronutrients and the weather parameters and classifying the type of crop based on the micronutrients using a support vector machine and decision tree algorithm. Out of the different crop production types in this research, three major crops have been considered, such as rice, wheat, and sugarcane. The crop growth generally depends on the macronutrients and micronutrient content of the soil, which is again a parametric representation of different climatic conditions such as rain, humidity, temperature, sunlight, and pH content of the soil.

Utpal Barman et al. [15] proposed a conventional hydrometer method to determine the percentage of sand, silt, and clay present in a soil sample. The samples are photographed under a constant light condition using an Android mobile of 13 MP cameras. The fraction of sand, silt, and clay of the soil samples are determined using the hydrometer test. The feature vector is calculated from the Hue, Saturation, and Value (HSV) histogram, color moments, color auto Correlogram, Gabor wavelets, and discrete wavelet transform. Finally, Support Vector Machine classifier is used to classify the soil images using linear kernel. The proposed method gives an average of 91.37% accuracy for all the soil samples.

Liang Zhong et al. [16] explored the modeling potential of deep convolutional neural networks (DCNNs) for soil properties based on a large soil spectral library. The DCNN models produced accurate predictions for most soil properties, and were superior to a single-task shallow convolutional neural network and traditional machine learning methods. Soil organic carbon content, nitrogen content, cation exchange capacity, pH, and calcium carbonate content were well predicted. The overall classification accuracy of soil texture was 0.749 (four groups) and 0.566 (12 levels). Another works of Rushika Ghadge et al. [17] proposed a method to help farmers check the soil quality depending on the analysis done based on data mining approach. Thus the system focuses on checking the soil quality to predict the crop suitable for cultivation according to their soil type and maximize the crop yield with recommending appropriate fertilizer. It can be achieved using unsupervised and supervised learning algorithms, like Kohonen Self Organizing Map (Kohenon's SOM) and BPN (Back Propagation Network). Anastasia Angelaki et al. [18] used Support Vector Machine (SVM), artificial neural network (ANN) and adaptive Neuro-fuzzy inference

system (ANFIS) were employed to estimate the cumulative infiltration of soil. The models accuracy was depended upon the two performance evaluation parameter which is correlation coefficient (CC) and root mean square error (RMSE). . Results of performance evaluation parameters suggest that triangular membership function-based ANFIS model works well than SVM and ANN models. Srunitha.k et al. [19] explain support vector machine based classification of the soil types. Soil classification includes steps like image acquisition, image preprocessing, feature extraction and classification. The texture features of soil images are extracted using the low pass filter, Gabor filter and using color quantization technique. Mean amplitude, HSV histogram, Standard deviation are taken as the statistical parameters.

Kodimalar Palanivel et al. [20] proposed an approach for prediction of crop yield using machine learning techniques in big data computing paradigm. In this paper, an investigation has been performed on how various machine learning algorithms are useful in prediction of crop yield. The common input parameters are rainfall, temperature, humidity, solar radiation, crop population density, fertilizer application, irrigation, type of soil, depth, farm capacity, and soil organic matter.

Sheetal A. Jhare et al. [21] focused on how to improve the soil quality and predict the Crop selection. The survey of this paper related with Edaphic factor, Classification problem and prediction of village wise soil parameters. That is done by collecting number of soil testing samples for finding soil fertility indices and pH values which represent a detail overview on application of machine learning in agriculture base. Mostly above problem are solved using two advance classifier Xgboost and Logistical regression.

Rakesh Kumar et al. [22] developed a method named Crop Selection Method (CSM) to solve crop selection problem, and maximize net yield rate of crop over season and subsequently achieves maximum economic growth of the country. Selection of crop(s) is an important issue for agriculture planning. It depends on various parameters such as production rate, market price and government policies. The proposed method may improve net yield rate of crops.

B. Bhattacharya et al. [23] showed an approach for automating soil classification procedure. Firstly, a segmentation algorithm is developed and applied to segment the measured signals. Secondly, the salient features of these segments are extracted using boundary energy method. Based on the measured data and extracted features to assign classes to the segments

classifiers are built; they employ Decision Trees, ANN and Support Vector Machines. Another works of S. Aydemir et al. [24] introduced a new thin section method which provides reliable, automated classification of mineral, non-mineral constituents (e.g. organic matter), non-crystalline, or poorly crystalline components and voids. Five different classes were separated and quantified for each sample. Classified features were checked with 500 reference points under the petrographic microscope. Quantitative results of digital image processing based on pixel number of each class (aerial percentage) were compared with traditional point-counting method. Laura Almendra-Martín et al. [25] working with different methods were applied in the temporal and spatial domains and evaluated using the holdout cross-validation technique. A set of variables was introduced in the SVM model to estimate Soil Moisture(SM), land surface temperature, precipitation, normalized difference vegetation index (NDVI), potential evaporation, soil texture and geographical coordinates. In this work, the applicability of different gap-filling techniques is evaluated on the ESA Climate Change Initiative (CCI) SM combined product, which is the longest available satellite-based SM data record.

Nadia Ansari et al. [26] established a rapid inspection method to classify the paddy seed based on varietal purity using a machine vision technique with multivariate analysis methods. Three varieties of paddy seeds were taken, namely - BR 11, BRRI dhan 28 and BRRI dhan 29. An image processing algorithm was developed for extracting 20 important features from 375 paddy seed images. The accuracy of 83.8%, 93.9%, and 87.2% was achieved using combined selected features of color, morphological, and textural for the PLS-DA, SVM-C, and KNN model, respectively.

Xiyue Jia et al. [27] proposed a novel method to map soil contamination by combining high-resolution aerial imaging (HRAI) with machine learning algorithms. To support model establishment and validation, 1068 soil samples were collected from an arsenic (As) contaminated area in Zhongxiang, Hubei province, China. The average arsenic concentration was 39.88 mg/kg (SD = 23 213.70 mg/kg), with individual sample points determined as low risk (66.9%), medium 24 risk (29.4%), or high risk (3.7%), respectively. Srikanth Gorthi et al. [28] evaluated a novel smartphone-based soil image segmentation technique and subsequent machine learning (ML) optimization methodology with a set of soil images for rapidly predicting soil organic matter (SOM) with minimal soil processing. A Tree-based

Pipeline Optimization Tool was used to find an optimum ML stacking scheme using six different ML models. Model validation statistics indicated that reflectance image-extracted sub-color space could predict SOM with reasonable accuracy (R^2= 0.88, RMSE=0.28%) using original images in stack one. Hasan Muhammad Abdullah et al. [29] described the performance of different machine learning algorithms on three different spatial and multispectral satellite image classification in rural and urban extents. Random forest, Support Vector Machine (SVM), and their combined strength (stacked algorithms) were applied on Landsat-8, Sentinel-2, and Planet images separately to assess individual and overall class accuracy of the images. In both Bhola and Dhaka, the SVM on Sentinel image performed best with an overall accuracy of 0.969, 0.983, and overall kappa of 0.948, and 0.968, respectively. Huanxue Zhang et al. [30] summarized a look-up table for searching the optimum classification features and classifiers in different landscape, providing a reference for future crop classifications and preventing the consumption of computing time. The results revealed that (1) an optimal feature-subset (with reduced data volume by 65%) can achieve high-accuracy crop mapping in heterogeneous regions. (2) The optimum type and number of features and classifiers are landscape sensitive. When a specific accuracy was required, homogenous regions need a smaller number of features and a simple MLC could meet the requirement.

Zhice Fang et al. [31] is to assess landslide susceptibility by integrating convolutional neural network (CNN) with three conventional machine learning classifiers of support vector machine (SVM), random forest (RF) and logistic regression (LR) in the case of Yongxin Country, China. A total of 364 landslide historical locations were randomly divided into training (70%; 255) and verification (30%; 109) sets for modelling process and assessment. The experimental results demonstrated that the performance of the machine learning classifiers previously mentioned can be effectively improved by integrating the CNN technique. There is a works of Pedro Augusto de Oliveira Morais et al. [32] proposed a new analytical method to predict and classify soil texture using digital image processing of soil samples (image segmentation) and multivariate image analysis (MIA). The computer vision approach adopted for the recognition of soil textures based on soil images matched 100% of the classification predicted according to the standard method. The new method is low-cost, environment-friendly, nondestructive, and faster than the standard method. Another works

of T. Abimala et al. [33] is intended to support agriculture by classifying 7 different types of soils like Clay, Clayey Peat, Clayey Sand, Humus Clay, Peat, Sandy Clay and Silty Sand, and in suggesting suitable crops that could be grown in those particular soils using image processing. Pre-processing is done by using Low Pass filter. Matching of image features is achieved by training the Decision Tree classifier with statistical measurements like mean, standard deviation, skew etc.

P. Kanaga Priya et al. [34] used the technique of data analysis and Internet of Things (IoT) to improve the productivity of the crops by maintaining the finest crop health. The proposed system aims at helping the farmers by providing valuable insights to them. These insights can be driven with the help of parameters such as soil moisture level, humidity, temperature and pH collected from the sensors using IoT. Crop production may be a complicated development that's influenced by soil and environmental condition input parameters. Pavan Patil et al. [35] created a system where machine learning approach is used in agriculture for the improvement of crops production. Their proposed system is used to provide the information about the best time of sowing, plant growth and harvesting of plant. They showed their system is also useful to predict which crop can be grown in a particular region. In addition, Neha Rale et al. [36] develop a prediction model for crop yield production using machine learning techniques on the perspective of keeping accurate data to compromise the risk of agricultural risk management. They compare the performance of various linear and non-linear regression models using 5-fold cross validation. Another research work of Amitabha Chakrabarty et al. [37] has collected 6 different of crops data from various government organizations of Bangladesh that deal with agriculture. An attempt is made on using various machine learning algorithms to predict crop production based on the collected dataset. They developed a mobile apps available for doing small-scale management and accounting for crop production, but these apps cannot make a prediction of crop production based on historical data.

Senthil Kumar Swami Durai et al. [38] provides a precision farming techniques that is helps to increase the productivity by precisely determining the steps that needs to be practiced at its due season. Predicting the weather conditions, analyzing the soil, recommending the crops for cultivation, determine the amount of fertilizers, pesticides that need to be used are some elements of precision farming. There is a paper of Kiran M P et al. [39] uses data mining

techniques to improve the production of crops. They used parameters are soil, atmosphere, water thickness, pesticides, composts and Crop informational collection. The Classification calculations utilized in study were Adaptive boosting classification, Excess tree classification, neural based classification, Multiple Process classification, Decision making classification, K-closest neighbours, Bayesian theory classification, decision Forest classification, sup- port group machine, and Randomized Gradient Classification.

Crop cultivation prediction is an integral part of agriculture and is primarily based on factors such as soil, environmental features like rainfall and temperature, and the quantum of fertilizer used, particularly nitrogen and phosphorus. Another research works of G. Mariammal et al. [40] proposed a novel FS approach called modified recursive feature elimination (MRFE) to select appropriate features from a data set for crop prediction. The performance of proposed MRFE technique is evaluated by various metrics such as accuracy (ACC), precision, recall, specificity, F1 score, area under the curve, mean absolute error, and log loss. A. Suruliandia et al. [41] provides a comparative study of various wrapper feature selection methods are carried out for crop prediction using classification techniques that suggest the suitable crop/s for land. Their experimental results show that Boruta, SFFS, and RFE feature selection techniques with the bagging classifier performing best with 10 fold and 70% – 30% data splitting range and RFE with the bagging method, outperforms other methods. Another research work of Priyadharshini et al. [42] has proposed a system to assist the farmers in crop selection by considering all the factors like sowing season, soil, and geographical location. Vrushali C. Waikar et al. [43] have designed a web-based system to predict the crop based on soil classification. The main aim of them is to build a user-friendly system based on soil parameters. They used machine learning techniques to implement their system. The authors specify the proposed method to help to find out the crop selection problems and also help to raise the net yield rate of crop over seasons and help to get maximum economic growth.

There is a research work of Rakesh Kumar et al. [44] proposed a method named Crop Selection Method (CSM) to solve crop selection problem, and maximize net yield rate of crop over season and subsequently achieves maximum economic growth of the country. The proposed method may improve net yield rate of crops. Archana Gupta et al. [45] proposed a system for Smart Management of Crop Cultivation using IoT and Machine Learning – a

smart system that can assist farmers in crop management by considering sensed parameters (temperature, humidity) and other parameters (soil type, location of farm, rainfall) that predicts the most suitable crop to grow in that environment. Another research work of K. Aditya Shastry et al. [46] have built a cloud-based agricultural framework which enables the Indian farmers, agricultural departments, and agro industries to extract useful agricultural information. The designed agricultural cloud framework is providing two services, i.e., soil classification as a service and crop yield prediction as a service. For soil classification, hybrid support vector machine (M-SVM) and for wheat yield prediction, customized artificial neural network (M-ANN) was developed. To store the agricultural data, they are using Amazon S3 and for deployment of the services, we have used Heroku cloud. There is a works of Aishnavi S. et al. [47] proposed a crops recommendation system to help the farmers in crop cultivation using data mining. To implement such an approach, crops are recommended based on its climatic factors and quantity. Data Analytics paves a way to evolve useful extraction from agricultural database. Crop Dataset has been analyzed and recommendation of crops is done based on productivity and season.

Another research works of Sanjay H A. et al. [48] is focused on the development of a new soil classifier based on decision tree. They develop a soil classifier which would classify soils into acidic, alkali, saline and non-problematic soils with a higher accuracy. For this they select 2593 soil samples from National Bureau of Soil Sciences (NBSS), India as the data set. They choose two popular decision tree algorithms C4.5 and CART for soil classification. They split the total number of soil samples (2593) into 1555 for training set and 1038 for test set. They trained C4.5 and CART using WEKA tool on the training set. Varsha A. et al. [49] aims at classifying soils based on their characteristics and thus recommending the best crop that could be produced maximum in that soil. The project uses IoT and machine learning algorithm to implement the problem stated. The proposed system predict the crop suited for the given soil. There are different varieties of soil and only some specific crop grow on one particular soil.

Another research works of S.S.Baskar et al. [50] have developed an automated system for soil classification based on fertility. After obtaining the fertility class labels with the help of automated system, they carried out a comparative study of various classification techniques with the help of data mining tool known as WEKA. This research has implemented a very

sound practical application of linear regression technique by forecasting an obscure property of the soil test.

S. Rajeswari et al. [51] proposed a system to predict soil fertility label by analyzing the Virudhunagar District Soil information and recommendations are offered for crop selection and sowing by using C5.0: Advanced Decision Tree (ADT) classifier algorithm. . Using this technique, an Android based mobile phone applications named as Design of Smart Information System (DSIS) application has been developed. The proposed application activates the Global Positioning System (GPS) to identify the user location. The performance of proposed model is analyzed and it is compared with the existing classification model for agricultural data. There is a work of R. Reshma et al. [52] presents a system to predict the category of the analyzed soil datasets to indicate the yielding of the crop. Their proposed IoT system is composed of pH sensors, Humidity and temperature sensors, Soil moisture sensors, soil nutrient sensors (NPK) probes, microcontroller/microprocessor equipped with WiFi and Cloud storage. For the recommending system, the SVM and Decision Tree algorithm is proposed to get the crop suitable for the given soil data and helps to enhance the growth using an optimized farming process.

From the prior researcher, they have done classification of soil using some specific parameters. In our work, we will include additional parameters like pH, Salinity, Organic matter, Nitrogen, Sulphur and Boron for the soil classification. From the previous work of researcher, it is not clear to grow which crops cultivated in a particular soil type. In our research work, we are classifying soil according to soil attributes and predict which crops are cultivated in certain soil. There are many lands in our country where the people think that cultivation is not possible at all. Noakhali is a coastal region in which many lands are uncultivated. Most of the farmers think cultivation is not possible in coastal region. This thesis paper works with identification of soil and predict which crops grown in a particular soil type. Then it provides a great beneficial for the farmers.

The summary of past works is described in short in the table 2.1 which is shown below.

Table 2.1: Summary of literature

Sl No.	Author, Year	Implementation methods	Remark
[2]	Yadav, Jagdeep, Shalu Chopra, and M. Vijayalakshmi (2021)	Support Vector Machine(SVM), Naïve Bayes, Random Forest, Linear Regression, MLP, ANN	• Proposed a model that can find whether the soil is fertile or not. • Sowing crop seed on fertile soil. • Predicting the crop yield on different soil features.
[3]	Harlianto, Pramudyana Agus, Teguh Bharata Adji, and Noor Akhmad Setiawan (2017)	Support Vector Machine (SVM), Neural Network, Decision Tree, And Naïve Bayesian	• Simulation is run by using RapidMiner Studio. • The result shows that SVM, with the use of linear function kernel, outperforms the others algorithms. The SVM best accuracy is 82.35%.
[7]	Cai, Simin, Rongqun Zhang, Liming Liu, and De Zhou (2010)	Support Vector Machine	• Used mean, variance and homogeneity features as parameters. • Provided a contrast between the proposed SVM method and other SVM methods.
[8]	Han Gao , Changcheng Wang, Guanya Wang , Haiqiang Fu , Jianjun Zhu (2021)	Pairwise Proximity Function Support Vector Machine (ppfsvm)	• Proposed method has higher overall accuracy (OA) and better robustness. • The overall accuracies is more than 90%.
[9]	Inazumi, Shinya, Sutasinee Intui, Apiniti Jotisankasa, Susit Chaiprakaikeow, and Kazuhiko Kojima (2020)	Image Processing	• AI model was created that was conscious of the practical simplicity of the images used. • This AI model can be applied to make judgments on soil classification.
[10]	Agarwal, Rahul, Narpat Singh Shekhawat, and	Automation and image processing, Henry gas solubility optimization (HGSO)	• The proposed variant of HGSO has been tested and validated on soil image dataset having seven classes

	Ashish Kumar Luhach (2021)		• Proposed method returns 80.58% accuracy.
[11]	Sirsat, Manisha Sanjay, M. Fernández-Delgado (2017)	Decision Trees, Nearest Neighbors, Neural Networks, Random Forests (RF), Rule Based And Support Vector Machines (SVM).	• Soil parameters: village-wise fertility indices of organic carbon (OC), phosphorus pentoxide (P2O5), manganese (Mn) and iron (Fe); soil pH and type; soil nutrients nitrous oxide (N2O), P2O5 and potassium oxide (K2O). • RF achieves the best performance.
[12]	Kodikara, Gayantha RL, and Lindsay J. McHenry (2020)	Support Vector Machine with radial basis function (svmRadial) and the penalized logistic regression model (glmnet).	• To classify lunar soils based on their types, particle size, maturity, and the type of pyroxene. • Band depth and Integrated band depth at 1 μm and 2 μm absorption band provided the best results.
[13]	Javed Mallick, Swapan Talukdar, Shahfahad, Swades Pal, Atiqur Rahman (2021)	Random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), and decision tree (DT)	• The objective of this work is to automated complex wetland mapping in Bangladesh's Sunamganj District. • The Kappa coefficient and root mean square error (RMSE) have been used. • Proposed fused ANN-based hybrid model outperformed all fused and MS classification models in terms of accuracy (kappa: 89.4% and RMSE: 0.13).
[14]	Dash, Ritesh, Dillip Ku Dash, and G. C. Biswal (2021)	Support vector machine and decision tree algorithm.	• The parameter such as rain, humidity, temperature, sunlight, and pH content of the soil are used. • Based on the soil’s features and input parameters, the developed model forecasts a suitable crop

			with a 92% accuracy level (Fitness Score).
[15]	Barman, Utpal, and Ridip Dev Choudhury (2020)	Conventional hydrometer method, Support Vector Method.	• The fraction of sand, silt, and clay of the soil samples are determined using the hydrometer test. • Proposed method gives an average of 91.37% accuracy for all the soil samples.
[16]	Liang Zhong, Xi Guo, Zhe Xu, Meng Ding (2021)	Deep convolutional neural networks (DCNNs)	• The performance of a multi-task DCNN model built on the basis of LucasResNet-16 was improved compared with the performance of the single-task model. • The overall classification accuracy of soil texture was 0.749 (four groups) and 0.566 (12 levels).
[17]	Ghadge, Rushika, Juilee Kulkarni, Pooja More, Sachee Nene, and R. L. Priya (2018)	Kohonen's SOM (Self-Organizing Map), BPN (Back-Propagation Neural Networks), API (Application Programming Interface).	• The system focuses on checking the soil quality. • To maximize the crop yield with recommending appropriate fertilizer.
[18]	Angelaki, Anastasia, Somvir Singh Nain, Varun Singh, and Parveen Sihag (2021)	Artificial neural network (ANN) and Adaptive Neuro-fuzzy inference system (ANFIS).	• Two performance evaluation parameter which is correlation coefficient (CC) and root mean square error (RMSE) is used. • The ANFIS model works well than ANN model.
[19]	Srunitha, K., and S. Padmavathi (2016)	Support vector machine (SVM)	• Soil classification includes steps like image acquisition, image preprocessing, feature extraction and classification. • The classifications of non-sandy soils are better classified with SVM (through WEKA).

			• It is able to achieve a 95% accuracy rate for classifying.
[20]	Palanivel, Kodimalar, and Chellammal Surianarayanan (2019)	Linear and logistic regression, decision trees, Naïve Bayes.	• Various machine learning techniques such as prediction, classification, regression and clustering are utilized to forecast crop yield. • An approach has been proposed for prediction of crop yield.

2.3 SUMMARY

The methods used in previous studies did not have sufficient precision in predicting crops cultivation prediction. From the prior researcher, they have done classification of soil using some specific parameters. In our work, we will include additional parameters like pH, Salinity, organic matter, nitrogen, boron and sulfur etc for the soil classification. In Bangladesh, no work has been done to predict crops cultivation through classifying soil using ANN with so many attributes of real data. From the previous work of researcher, it is not clear to grow which crops cultivated in a particular soil type. In our research work, we will classify soil according to soil attributes and predict which crops are cultivated in certain soil. There are many lands in our country where the people think that cultivation is not possible at all. Noakhali is a coastal region in which many lands are uncultivated. Most of the farmers think cultivation is not possible in coastal region. This thesis paper works with identification of soil and predict which crops grown in a particular soil type. Then it provides a great beneficial for the farmers.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter presents an outline of research methods that were followed in the research work. It provides information on the data collection and preprocessing, workflow diagram for soil classification and crop prediction, artificial intelligence techniques, and performance analysis terminologies used in work. The requirements of this work and the procedures that were followed to carry out this study are included.

3.2 ATTRIBUTES SELECTION AND DATA COLLECTION

There are several parts included in this section of attributes selection and data collection.

3.2.1 Attributes of Soil Classification

For the proper classification of soil, it is necessary to select significant attributes of soil. Accurate classification of soil is a challenging task for farmers. Different classification techniques are used to classify soil. The description of these attributes is shown in Table 3.1. Attributes of soil are pH, Salinity, Organic matter, Potassium, Sulfur, Zinc, Boron, Nitrogen, Calcium, Phosphorus, Manganese, Magnesium, Copper, Iron, Sand, Silt, Clay, CaCo3, etc. From these attributes, seven parameters are selected for this thesis work. These attributes are pH, Salinity, Organic matter, Sulfur, Nitrogen, Phosphorus-B, and Boron.

Table 3.1: Attributes list and description

Attributes	Description
pH	pH value of soil
Salinity	D S\m
Organic matter	Percentage
Sulfur	Microgram/per gram soil
Nitrogen	Percentage
Phosphorus-B	Microgram/per gram soil
Boron	Microgram/per gram soil

3.2.2 Features of Soil in Crops Cultivation

For the proper prediction of crop cultivation, it is necessary to select significant attributes for crops. Several crop attributes are Soil, Temperature, Humidity, Rainfall, etc. For crop cultivation, soil is an essential attribute. Different types of crops grow on different types of soil. If the soil selection for the particular crops is not correct for that, farmers do not get the profit as they expected from the crop cultivation. There are different types of soils in Bangladesh such as clay, clay loam, loam, sandy clay, sandy clay loam, sandy loam, loamy sand, sand, silty clay, silty clay loam, silt loam, silt, etc. From these soil types, three types of soil classes are selected for this thesis work. These soil classes are clay, clay loam, and loam. Using soil attributes, we will classify soil and then predict which crops are suitable for a particular soil. There are 4 attributes are selected for the prediction of crop cultivation. These attributes are Location, Soil class, Season, and pH level.

3.2.3 Data Collection

The data used in this research work are collected from Begumganj, Sonaimuri, Hatiya, Senbag, and Subarnachar Upazilla of Noakhali District. Noakhali is a coastal region of Bangladesh. There are many uncultivated lands in Subarnachar of Noakhali. The farmers of Subarnachar are not very educated. They often do not understand what crops are grown in the coastal lands. That is why they do not know how to produce more crops from the land. Many farmers lose interest in cultivating crops as they do not get the expected results. If we provide them with some advanced technology, then they will be interested in cultivating crops and the uncultivated lands will play a leading role in the food production of the country. In this research work, we are providing a pathway for the farmers so that they can understand what crops are grown in which soil type. The total number of soil data used in this work is 1542, and the total number of crop data is 1162. In the entire data set, we will take 70% of the data used for training and 30% of the data for testing. Soil samples are collected from agricultural land to follow a proper way. Soil samples are collected at a depth of 6 inches. from the top level of the field. Soil samples were tested in the lab of Soil Resource Development Institute (SRDI) of Bangladesh. The chemical features of soil are listed in the soil dataset. Then crop datasets are built on the basis of Location, Soil type, pH, and Season.

3.2.4 Data Preprocessing

Before introducing the collected data into the classification algorithm for training, it should be noted that there are differences in the range and measure units of the variables. Using this data directly for the training of a classification algorithm could cause a considerable delay in convergence. A solution to this problem is to preprocess the training and testing data before using them. First of all, after taking all collected raw data, samples are included in the excel file and saved file with the 'csv' extension. There are several parameters whose values are a string. String data is converted into numeric via python. Preprocessed data is shown in the appendix.

3.2.5 Data Normalization

Many algorithms attempt to find trends in the data by comparing features of data points. However, there is an issue when the features are on drastically different scales. For this data, normalization is needed for the best performance of the classification algorithm. We use the Min-Max normalization formula for normalization. Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1. Here is the data normalization formula

$$Xnormalized = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.1)$$

Where $X_{normalized}$ is the new normalized value. All supervised learning methods start with an input data matrix called X here. Each tuple represents one observation. Each column represents one variable.

3.2.6 Data Analysis

This thesis works with what kind of chemical measurement exist in distinct soil category. It is important to find out the chemical measurement of soil. Soil class is used in crop selection. There are various features of the soil. Based on soil type different kinds of crops are cultivated. The measurements of pH, salinity, organic matter, nitrogen, sulfur, phosphorus-B, and boron are used to classify soil. For crop cultivation, soil type is a necessary tool for farmers.

3.3 WORKFLOW DIAGRAM FOR SOIL CLASSIFICATION

Soil samples are collected from different agricultural lands in Noakhali. Chemical features of soil like pH, salinity, organic matter, sulfur, nitrogen, phosphorus-B, and boron are selected as soil attributes for this work. The workflow diagram of soil classification of this thesis work is shown in figure 3.1.

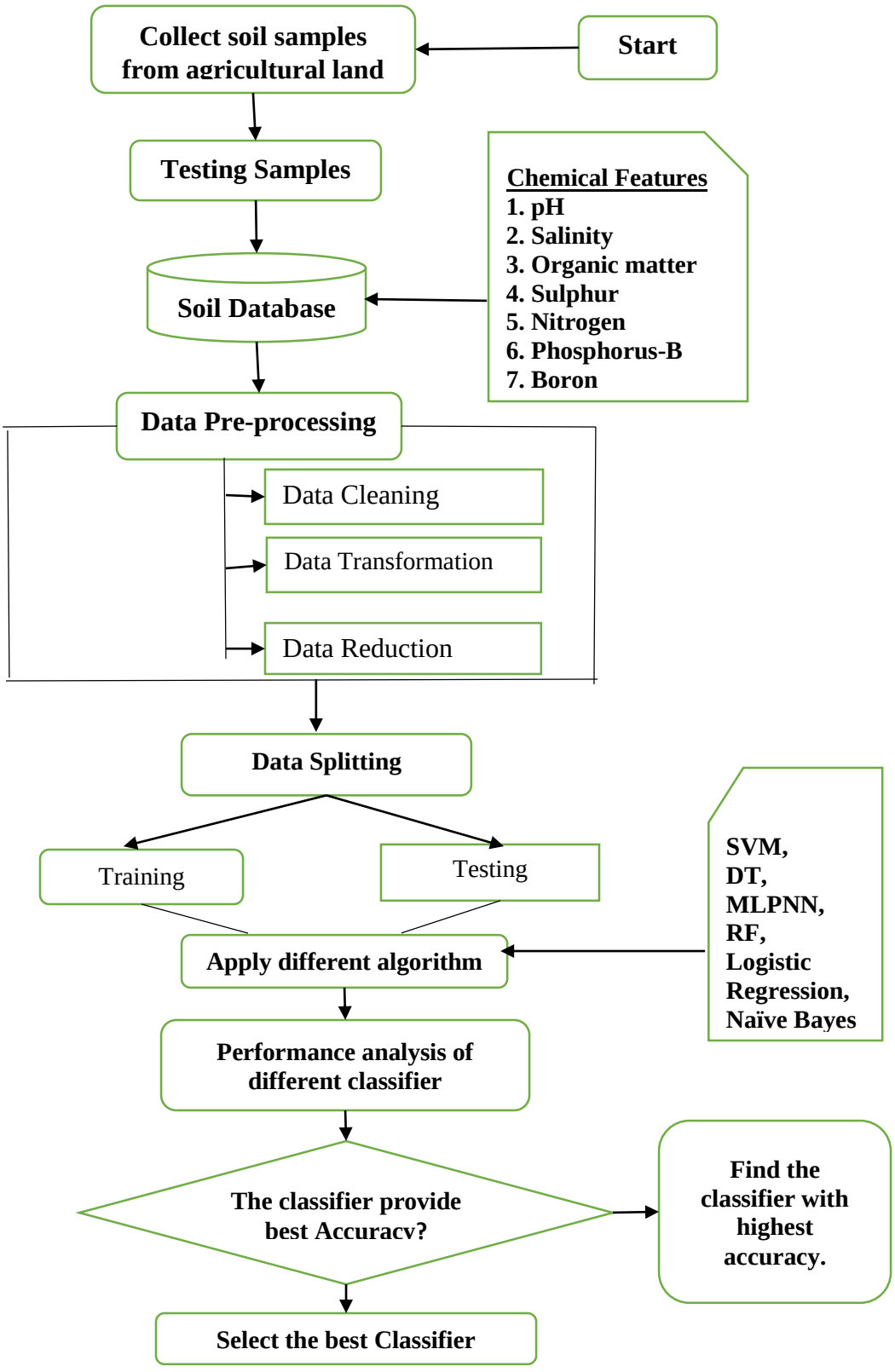


Figure 3.1: Workflow diagram for soil classification.

After collecting data, data is preprocessed for training the algorithm. Then the processed data is used in the training and testing phase. Different algorithms such as Support Vector Machine, Decision Tree, Multilayer Perceptron, Random Forest, Logistic Regression, and Naive Bayes are used to classify the soil.

3.4 WORKFLOW DIAGRAM FOR THE PREDICTION OF CROPS CULTIVATION

Estimating agricultural crop cultivation prior to harvest is an important issue in agriculture. Early prediction of crop cultivation provides useful information to farmers and policymakers. Appropriate prediction of crop productivity is required for efficient planning of land usage and economic policy. First, it is necessary to collect the coastal regional crops list. Crops are cultivated based on soil class. In this work, three soil classes, namely- clay, loam, and clay loam are used. The crop dataset consists of location, soil class, and other necessary features. After collecting data, data is preprocessed for training the algorithm. Then the processed data is used in the training and testing phase. Different algorithms such as Support Vector Machine, Decision Tree, Multilayer Perceptron, Logistic Regression, and KNN are used for crop cultivation prediction. The workflow diagram of crop cultivation prediction is shown in figure 3.2.

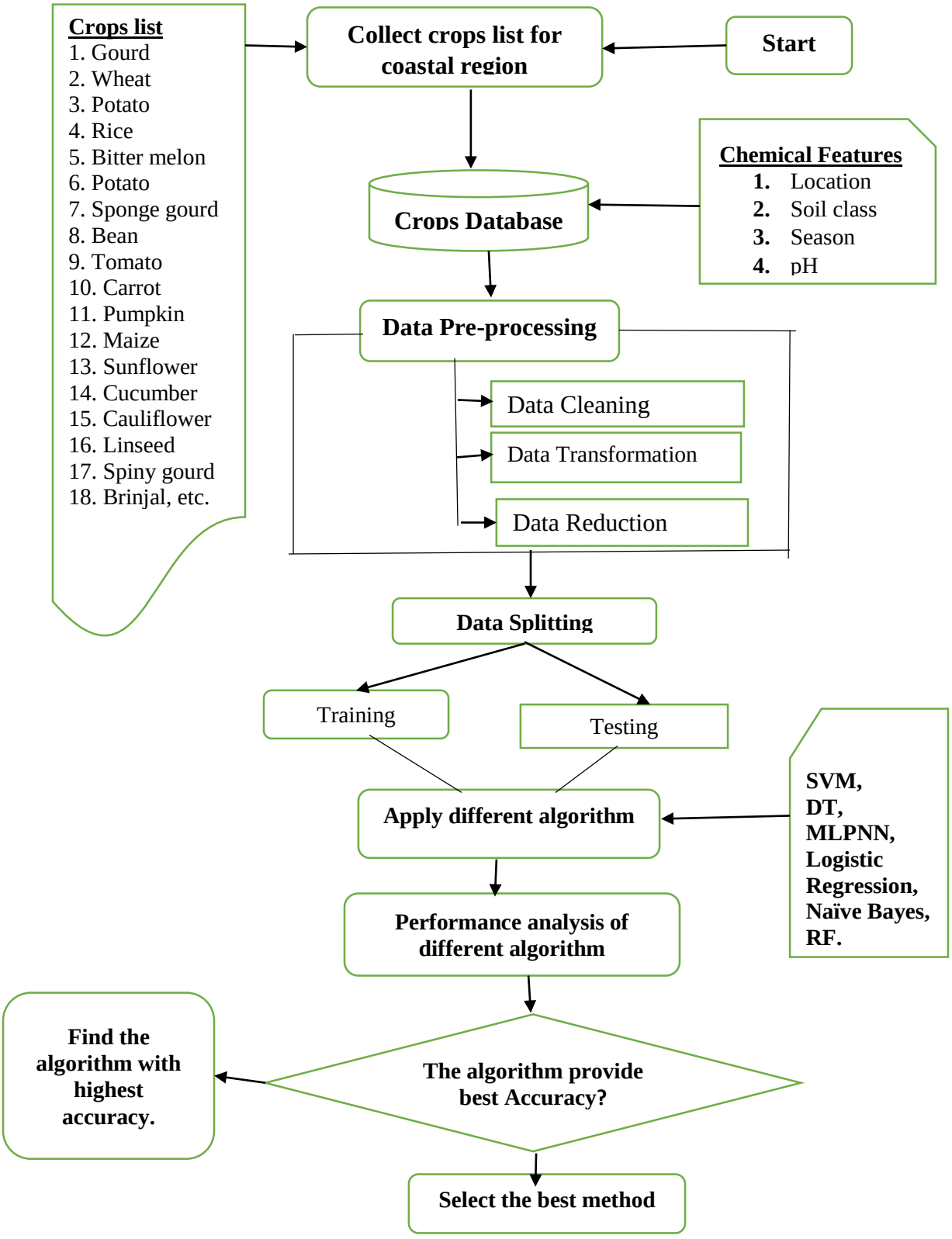


Figure 3.2: Workflow diagram for crop cultivation prediction.

3.5 ARTIFICIAL INTELLIGENT TECHNIQUES FOR SOIL CLASSIFICATION

There are several techniques used for soil classification. In this work, we will use Multilayer Perceptron Neural Network, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, and Logistic Regression.

3.5.1 Multilayer Perceptron Neural Network

The Multilayer Perceptron is composed of a set of sensorial units organized in three or more layers. The first layer is the input layer, which does not perform any computational task. Then there are one or more hidden (intermediate) layers and an output layer, all composed of computational nodes. In a typical MLP network, all the nodes from a layer are connected with every node from the previous and from the next layer. The non-computational nodes in the input layer use an identity function, while the computation nodes in the intermediate and the output layers use a sigmoid function [53].

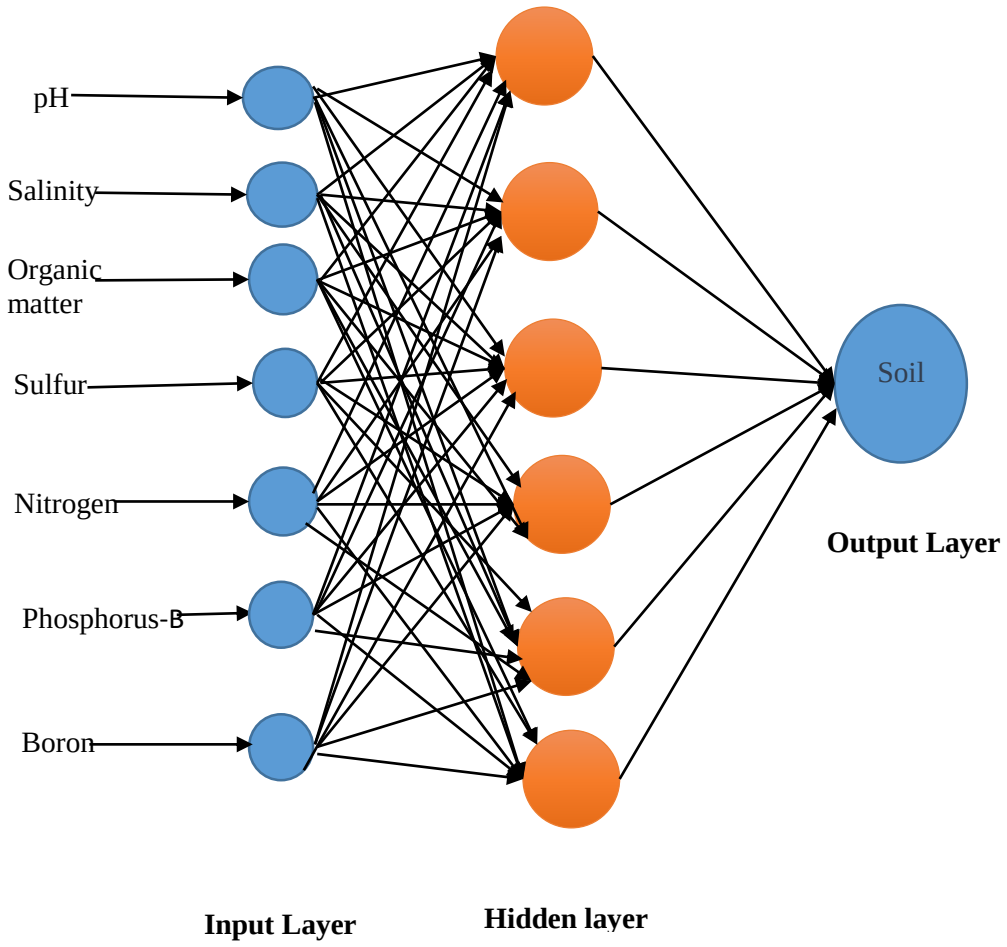


Figure 3.3: An example of a Multilayer perceptron neural network.

In this work, different soil attributes like pH, Salinity, Organic matter, Sulfur, Nitrogen, Phosphorus-B, and Boron work as the input layer. The hidden layer is designed with six neurons and is to be increased as the training process progresses on demand, as shown in Figure 3.3

Multilayer perceptron neural network has been used with success to solve difficult problems through its training by using the error backpropagation algorithm, which basically consists of two steps: a forward step where the signal propagates through the computational units until it gets to the output layer; and a backward step where all the synaptic weights are adjusted accordingly to an error correction rule [54].

3.5.2 Support Vector Machine

A support vector machine is a supervised machine learning model that is used for classification, regression, and outlier detection. After giving an SVM model set of labeled training data for either two categories, they are able to categorize new examples. It works based on the decision planes that define the boundaries. The decision plane separates one object from another object of a different class. The data points that are nearer to the hyperplane are called support vectors [55]. The process of training an SVM decision function amounts to identifying a reproducible hyperplane that maximizes the distance (i.e., the "margin") between the support vectors of both class labels. The geometrical interpretation of support vector classification (SVC) is that the algorithm searches for the optimal separating surface, i.e. the hyperplane that is, in a sense, equidistant from the two classes. SVC is outlined first for the linearly separable case. Kernel functions are then introduced in order to construct non-linear decision surfaces [56].

Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory. SVM has strong regularization properties. Regularization refers to the generalization of the model to new data [57]. Support vector machines, in particular, were designed as a tool to solve supervised learning classification problems. Supervised learning is an area of machine learning in which we are given some "training set" of data for which we know a priori the appropriate classifications.

- **Working Principles of SVM**

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

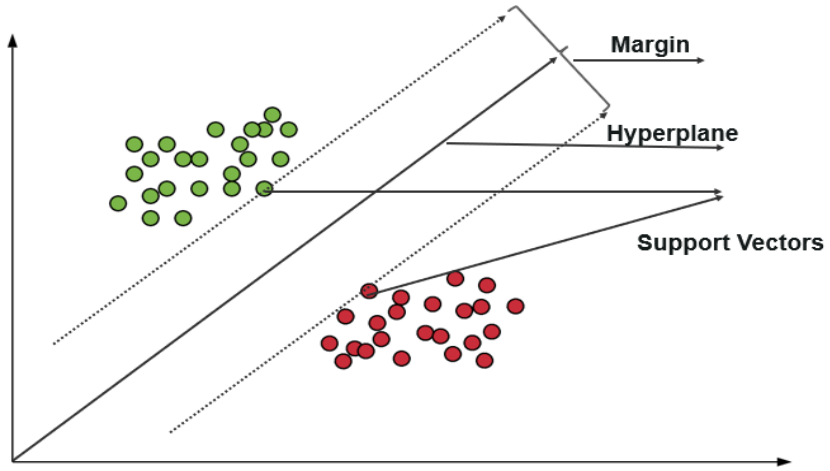


Figure 3.4: An example of a Support Vector Machine

The followings are important concepts of SVM –

Support Vectors: Data points that are closest to the hyperplane are called support vectors. Separating lines will be defined with the help of these data points.

Hyperplane: As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin: It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. A large margin is considered a good margin, and a small margin is considered a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps:

- First, SVM will generate hyperplanes iteratively that segregate the classes in the best way.
- Then, it will choose the hyperplane that separates the classes correctly.

❖ SVM Kernels

In practice, the SVM algorithm is implemented with the kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which the kernel takes a low dimensional input space and transforms it into a higher dimensional space. In simple words, the kernel converts non-separable problems into separable problems

by adding more dimensions to it. It makes SVM more powerful, flexible, and accurate. The following are some of the types of kernels used by SVM.

Linear Kernel: It can be used as a dot product between any two observations. The formula of the linear kernel is as below:

$$k(x_i x_j) = \sum x_i * x_j \qquad (3.2)$$

From the above formula, we can see that the product between two vectors, say x_i & x_j is the sum of the multiplication of each pair of input values.

Polynomial Kernel: It is a more generalized form of the linear kernel and distinguishes curved or nonlinear input space. Following is the formula for polynomial kernel –

$$k(x_i x_j) = (x_i . x_j + 1)^d \qquad (3.3)$$

Here d is the degree of the polynomial, which we need to specify manually in the learning algorithm.

Radial Basis Function (RBF) Kernel: RBF kernel, mostly used in SVM classification, maps input space in indefinite dimensional space. The following formula explains it mathematically –

$$k(x_i x_j) = \exp\left(-gamma * sum(x_i - x_j{}^2)\right) \qquad (3.4)$$

Here, *the gamma* ranges from 0 to 1. We need to specify it manually in the learning algorithm. A good default value of *gamma* is 0.1.

3.5.3 Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used mostly in the problems of Classification and Regression. It builds decision trees on different samples and takes their majority voting for classification, and takes their averaging in case of regression [58]. There is an important feature of the Random Forest Algorithm that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It provides better outcomes for classification problems. Random forest is developed on the basis of ensemble learning, that

is a process of combining multiple classifiers to solve a complex problem and to improve the performance model.

RF is a classifier that builds a number of decision trees on various samples of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and, based on the majority votes of predictions, predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting [59]. The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification or a "vote." The forest chooses the classification with the majority of the "votes." When using Random Forest for regression, the forest picks the average of the outputs of all trees. An important key here is that there is a low (or no) correlation between the individual decision trees that make up the larger Random Forest model. The process of the Random Forest algorithm is shown in following figure 3.5.

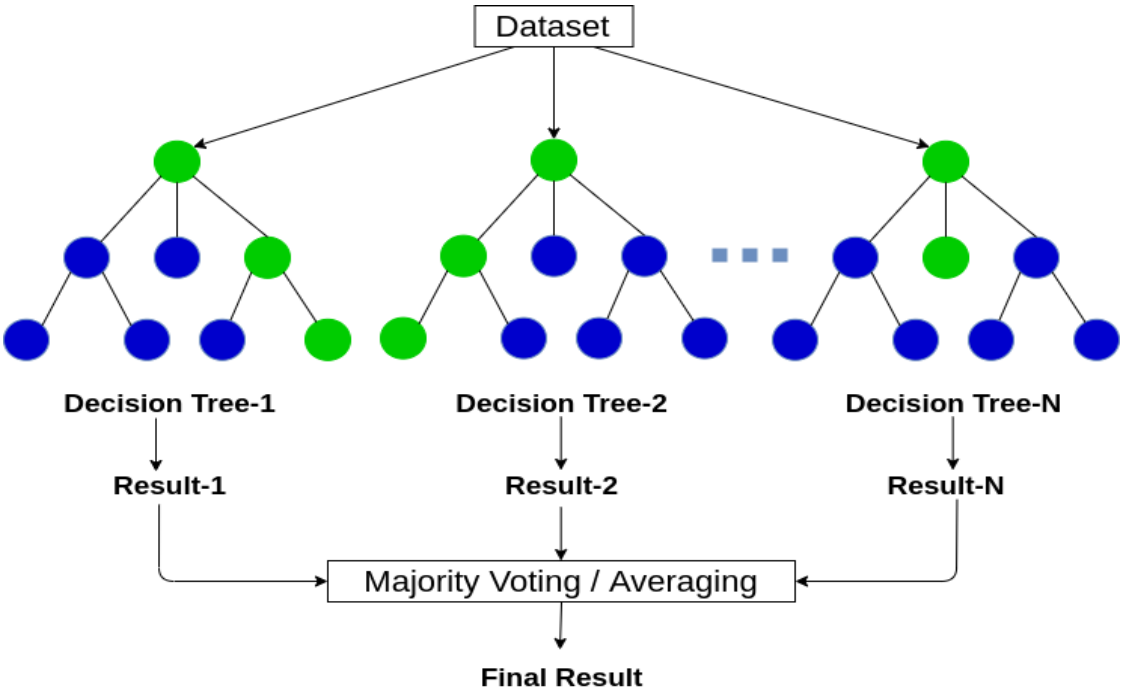


Figure 3.5: Process of Random Forest Algorithm

Decision trees in an ensemble, like the trees within a Random Forest, are usually trained using the "bagging" method. The "bagging" method is a type of ensemble machine learning algorithm called Bootstrap Aggregation. An ensemble method combines predictions from multiple machine learning algorithms together to make more accurate predictions than an individual model [60]. Bootstrap randomly performs row sampling and feature sampling from the dataset to form sample datasets for every model. Aggregation reduces these sample datasets into summary statistics based on the observation and combines them. Bootstrap Aggregation can be used to reduce the variance of high-variance algorithms such as decision trees.

Algorithm

Steps involved in the random forest algorithm:

Step 1: In a Random forest, N number of random records are taken from the data set having K number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

3.5.4 Decision Tree

Decision Tree is a supervised learning technique that can be used for both classification and regression problems, but it is mostly preferred for solving classification problems. It is a tree-structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. In a decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance [61].

The decision tree in figure 3.6 classifies a particular goal of a restaurant according to food category and returns the classification associated with the particular leaf (in this case, Yes or No).

For example, the instance (Patrons=Full, Hungry=Yes, Type=French) would be sorted down the leftmost branch of this decision tree and would therefore be classified as a positive instance.

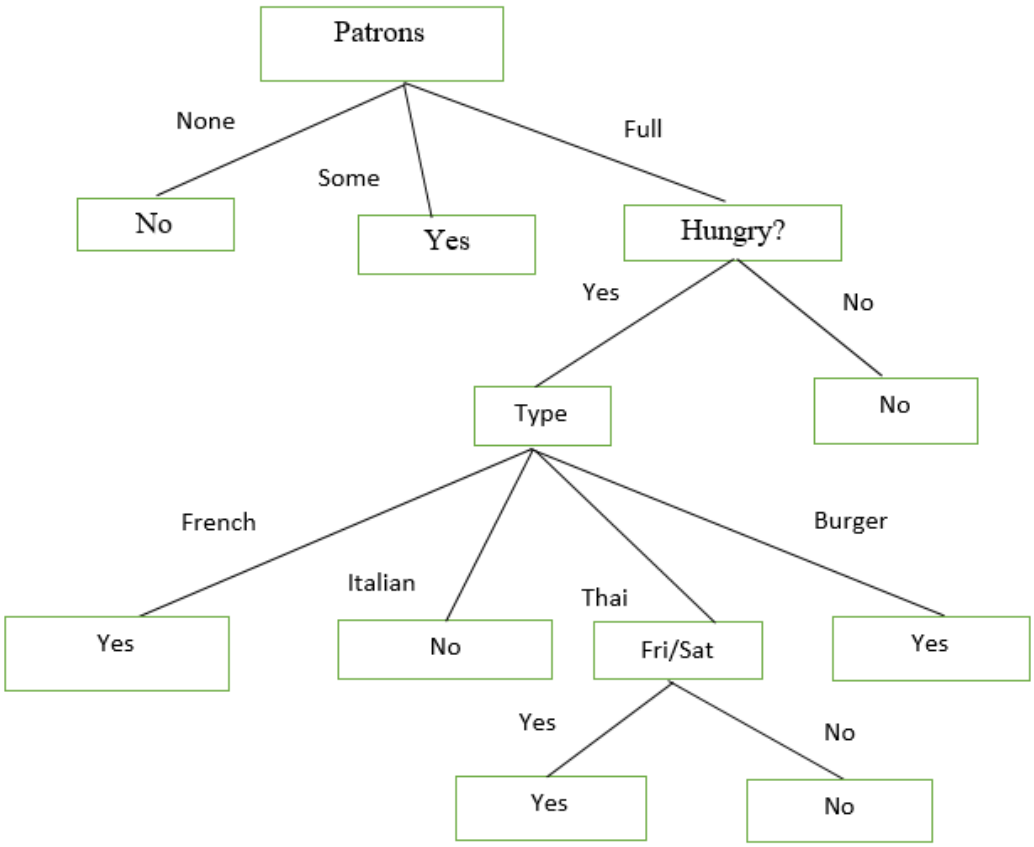


Figure 3.6: Example of decision tree for the concept of a restaurant.

❖ **Working Principles of Decision Tree**

Attribute selection measures are also called splitting rules to decide how the tuples are going to split. The splitting criteria are used to best partition the dataset. These measures provide a ranking of the attributes for partitioning the training tuples. The most popular methods of selecting the attribute are information gain and the Gini index.

1) Information Gain

This method is the main method that is used to build decision trees. It reduces the information that is required to classify the tuples. It reduces the number of tests that are needed to classify the given tuple. The attribute with the highest information gain will be selected. The original information needed for the classification of a tuple in dataset D is given by:

$$E(S) = \sum_{i=1}^{n} -p_i log_2 p_i \tag{3.5}$$

Where p is the probability that the tuple belongs to class C. The information is encoded in bits; therefore, log to base 2 is used. E(s) represents the average amount of information required to find out the class label of dataset D. This information gain is also called **Entropy**. Entropy is also represented by H as H(V) or H(S). The formula is:

$$H(V) = -\sum_{k} p(V_k)\, log_2\, P(V_k) \tag{3.6}$$

Where, V is the random variable, V_K is the value of each variable and P(V_k) is the probability of each variable.

If a training set contains p for positive values and n for negative values then the entropy of the whole set is-

$$H(Goal) = B\left(\frac{p}{p+n}\right) \tag{3.7}$$

An attribute A with d distinct values divides the training set E into subsets E_1,….E_d. Each subset E_k has P_k positive values and n_k negative values. So, the expected entropy remaining after testing attribute A is-

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n}\, B\left(\frac{p_k}{p_k + n_k}\right) \tag{3.8}$$

The information gain is-

$$Gain(A) = Entropy\ of\ the\ goal - Remainder(A) \tag{3.9}$$

$$or, \qquad Gain(A) = B\left(\frac{p}{p+n}\right) \tag{3.10}$$

$$or, \qquad Gain(A) = B\left(\frac{p}{p+n}\right) - \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right) \tag{3.11}$$

2) Gain Ratio

Information gain might sometimes result in portioning useless for classification. However, the Gain ratio splits the training data set into partitions and considers the number of tuples of the outcome with respect to the total tuples. The attribute with the max gain ratio is used as a splitting attribute.

3) Gini Index

Gini Index is calculated for binary variables only. It measures the impurity in training tuples of dataset D, as

$$Gini = 1 - \sum_{i} p(i/t)^2 \tag{3.12}$$

P is the probability that the tuple belongs to the class. The Gini index that is calculated for binary split dataset D by attribute A is given by:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \tag{3.13}$$

Where n is the nth partition of dataset D. The reduction in impurity is given by the difference of the Gini index of the original dataset D and Gini index after partition by attribute A. The maximum reduction in impurity or max Gini index is selected as the best attribute for splitting.

Algorithm

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

3.5.5 Logistic Regression

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e., binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type [62].

Logistic regression aims to solve classification problems. It does this by predicting categorical outcomes, unlike linear regression which predicts a continuous outcome. In the simplest case there are two outcomes, which is called binomial, an example of which is predicting if a tumor is malignant or benign. Other cases have more than two outcomes to classify; in this case it is called multinomial. A common example of multinomial logistic regression would be predicting the class of soil category between 3 classes.

The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets [63]. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The logistic function is shown in following figure 3.7.

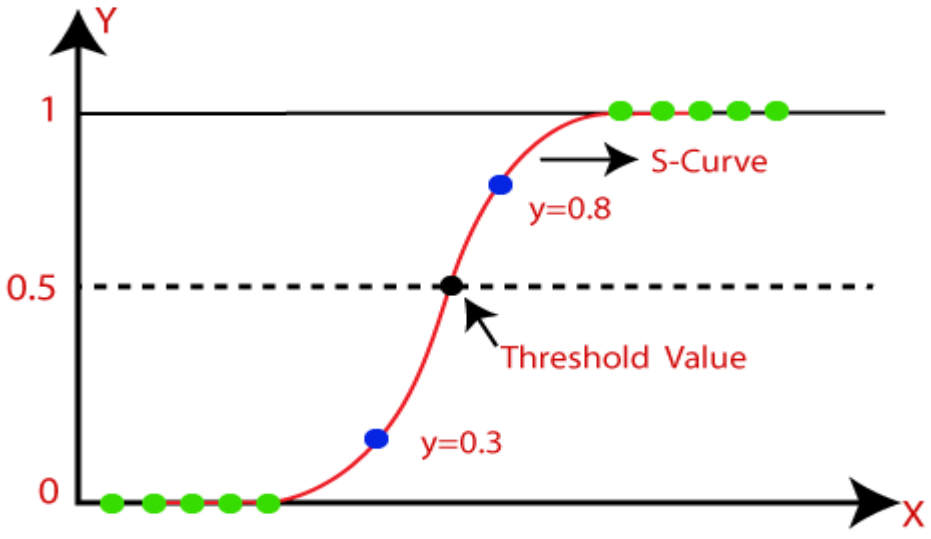


Figure 3.7: An example of a logistic function.

Probability always ranges between 0 (does not happen) and 1 (happens). The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

3.5.6 Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective classification algorithms which helps in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object [64].

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. Conditional probability is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred. The formula is:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{3.14}$$

Here, P(A|B) means how often A happens *given that B happens; P(B|A) means* how often B happens *given that A happens*; P(A) means how likely A is on its own; P(B) means how likely B is on its own.

3.6 ARTIFICIAL INTELLIGENT TECHNIQUES TO ASCERTAIN CROP CULTIVATION

There are several techniques used for crop cultivation prediction. In this work, we will use Multilayer Perceptron Neural Network, Decision Tree, Support Vector Machine, Random Forest, Logistic regression, and Naïve Bayes.

3.6.1 Multilayer Perceptron Neural Network

The general concept of MLPNN is discussed in section 3.5.1. In this work, we are predict crop cultivation using soil classification. There are different soil types, such as sandy soil, Loam soil, silt soil, clay soil, etc. In MLPNN techniques, the soil will play the role of input layer for crop prediction. There are four parameters, namely- location, soil class, season and soil pH is used as input. The hidden layer is designed with five neurons and is to be increased as the training process progresses on demand, as shown in Figure 3.8.

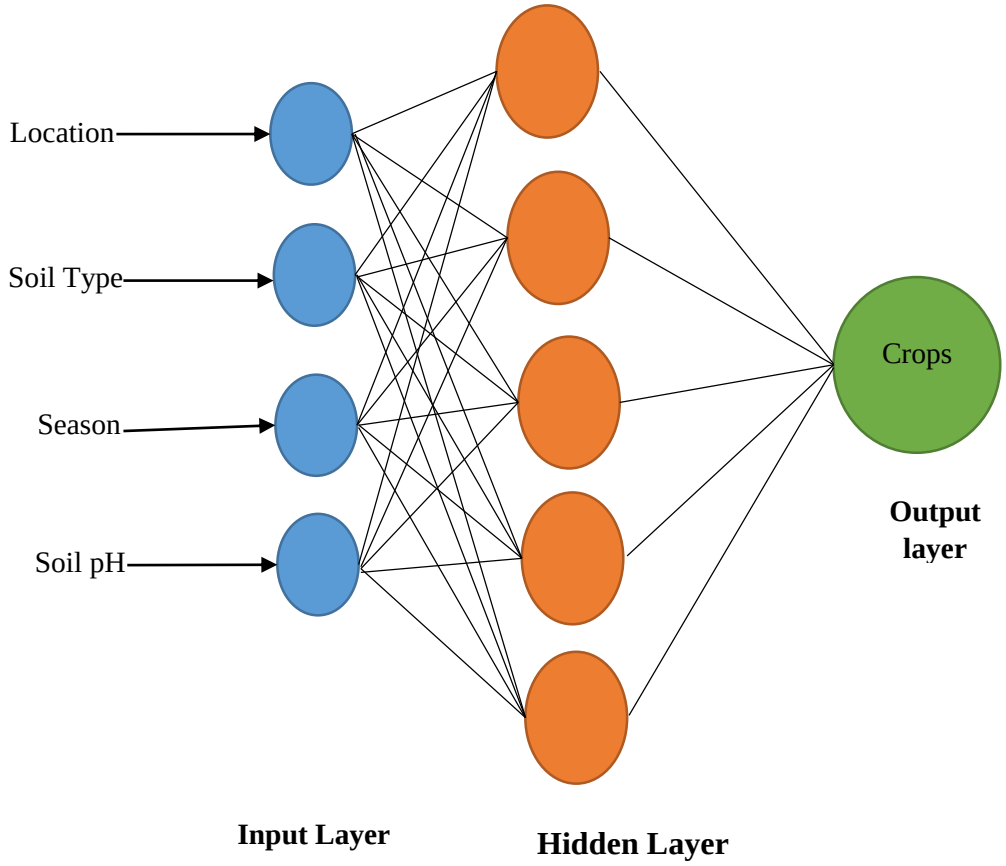


Figure 3.8: MLPNN structure for crop cultivation prediction

3.6.2 Decision Tree

The decision tree algorithm is based on recursive partitioning of input parameters to classify the data. It is an advanced algorithm that determines the importance and contribution value of input parameters. It learns on the dataset by calculating the Entropy and Information Gain of a decision. Different decisions over multiple iterations are made and the best fit is determined by that set of decisions that have the least entropy. It is also a part of the family of supervised learning algorithms where the dataset inputs have pre-hand knowledge of output. The decision tree algorithm has a recursive tree or graph network-like structure of decision calls. Each decision call contributes to a change in entropy. The successive decision calls leads to an output set of decision paths. On every decision call, the set of input parameters is divided into two or more subsets. It is called Splitting. There are different methods used to make the splitting decision like MARS, Chi-Squared, ID3, Cart, etc. The algorithm begins with the root node which contains all training data in the dataset. Then at each iteration of the algorithm calculates the Entropy and Information Gain. Using these values the attribute with the least entropy or highest information gain is selected as the splitting node and the rest attributes are split into subsets. The recursive splitting continues over each subset until the terminal nodes are reached. Entropy- Entropy of a decision is defined as the degree of randomness of the data. Higher value entropy means higher randomness in the data, which makes it harder to make logical analysis and conclusive study of data [65].

3.6.3 Support Vector Machine

The general concept of Support Vector Machine is discussed in the section of 3.5.2. We are adopting SVM techniques for crop cultivation prediction.

3.6.4 Random Forest

The general concept of Random Forest is discussed in the section of 3.5.3. We are adopting Random forest techniques for crop cultivation prediction.

3.6.5 Logistic Regression

The general concept of Logistic Regression is discussed in the section of 3.5.5. We are adopting Logistic Regression techniques for crop cultivation prediction.

3.6.6 Naïve Bayes

The general concept of Naïve Bayes is discussed in the section of 3.5.6. We are adopting Naïve Bayes techniques for crop cultivation prediction.

3.7 PERFORMANCE ANALYSIS

This research is compare performance by using various evaluation metrics. This paper carried out model development in order to assess the performance and usefulness of different classification algorithms for predicting crops cultivation through classifying soil. This work uses different artificial intelligence techniques, namely support vector machine, decision tree, Multilayer perceptron, Random Forest, naïve Bayes,and Logistic Regression. It is essential to review standard metrics to understand the performance of different classifier. Model evaluation or Performance Analysis is based on confusion matrix and all related measures like- True positives, True negatives, False positives, False negatives, Error rate (ERR), Accuracy, Precision and recall, F-measure, Sensitivity, Specificity, Precision, ROC curve, etc.

a) **Confusion Matrix**

Confusion matrix is one of the most common and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problems where the output can be of two or more types of classes. The confusion matrix is a table with two dimensions, "Actual class" and "Predicted class" in both dimensions. Actual classifications are Rows, and Predicted ones are columns. The classifiers for the two classes have the following confusion matrix, as depicted in Table 3.2.

Table 3.2: Confusion Matrix

		Predicted	
Actual		Positive(Class 1)	Negative(Class 2)
	Positive(Class 1)	TP	FN
	Negative(Class 2)	FP	TN

The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

True Positives (TP): True positives are the cases when the actual class of the data point was true and the predicted is also true.

True Negatives (TN): True negatives are the cases when the actual class of the data point was false and the predicted is also false.

False Positives (FP): False positives are the cases when the actual class of the data point was false and the predicted is true.

False Negatives (FN): False negatives are the cases when the actual class of the data point was true and the predicted is false.

b) **Precision**

Precision value is the proportion of correct positive classifications from cases that predict as positive. It is the accuracy of the positive prediction.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)} \tag{3.15}$$

For the prediction of crops cultivation, TP indicates the number of accurately predicted crops. And FP indicates that actual class of the data point was false and the predicted is true. For crops classification precision is as follows:

$$Precision = \frac{Total\ number\ of\ accurately\ predicted\ actual\ crops}{Total\ number\ of\ predicted\ crops} \tag{3.16}$$

c) **Recall or True Positive Rate**

Recall is the proportion of correct positive classification (True positive) from cases that are actually positive.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)} \tag{3.17}$$

d) **F-measure**

F-measure combines both precision and recall. It also known as F1 score. It is expressed as-

$$F_1 = 2 * \frac{Precision * recall}{Precision + recall} \quad (3.18)$$

e) **Accuracy**

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. Accuracy comparison is based on the performance among the four classification algorithms.

$$Accuracy = \frac{(TP + TN)}{(TN + FP + FN + TP)} \quad (3.19)$$

For crops cultivation prediction, accuracy is expressed as-

$$Accuracy = \frac{Total\ number\ of\ crops\ that\ are\ accuraty\ predicted}{Total\ number\ of\ crops} \quad (3.20)$$

f) **AUC and ROC curve**

Receiver operation characteristic (ROC) curve pots the true positive (recall) against false positive ratio. Area under the curve (AUC) metrics is used to calculate overall performance of a classification model based on area under the ROC curve.

3.7.1 Accuracy Testing For New Samples

After whole model development and performance analysis with different classification algorithm, the values of different models are investigated with the 42 performance on the classification methods. Then accuracy testing for new samples is done by using 20 numbers of new samples. These samples are completely new for the model. These will run on the model and then accuracy is testing by the values. New samples data is given in appendix.

3.8 SENSITIVITY

Sensitivity analysis is a strategy for modifying model input in a controlled manner and evaluating the impact of these changes on the model output. This method reveals the

model's sensitivity to these changes, as well as the impact of particular features on the model's performance. The objective of sensitivity analysis is to figure out how input and target factors interact.

In soil classification prediction, 7 attributes (pH, Salinity, Organic matter, Nitrogen, Phosphorus-B, Boron and Sulfur) of the dataset have been used as input variable. The output variable is categorized into three classes: clay (num = 0), clay loam (num = 1) and loam (num=2) of the soil classification.

Sensitivity is calculated while all parameters (input and output) are fixed in their reference condition. We have adjusted one variable at a time using this analysis technique to see how it affects the anticipated output. The change in the predicted output is typically calculated using the input variable's maximum and minimum values, with all other input variables held constant at their reference values. The percentage of change of an input variable has been determined using the following formulas if the maximum and minimum values of the input variable are Xmax and Xmin respectively:

((Xmax - Xmin)*100)/ Xmin) (from max to min value)

To determine sensitivity, the percentage change in input is divided by the percentage change in output. The output's sensitivity to each of the independent variables is then calculated by repeating those processes.

3.9 SUMMARY

This methodology chapter has discussed about the dataset and how to processed dataset. Also discuss several types of artificial intelligence techniques such as Multilayer Perceptron, Decision Tree, Support Vector Machine (SVM), Random Forest, Logistic Regression and Naïve Bayes. Performance analysis terminologies such as confusion matrix, roc curve, precision, recall and f-measure have included in this chapter.

CHAPTER 4

RESULTS ANALYSIS

4.1 INTRODUCTION

This chapter discussed the results of this research. In the first section, the results of soil classification from the perspective of different algorithms are shown. And at the second section, the results of crop cultivation prediction based on soil classification are shown. The training results of different classification algorithms for soil classification and crop cultivation prediction are discussed. Different algorithms such as Support Vector Machine, Multilayer Perceptron Neural Network, Decision Tree, Naive Bayes, Random Forest, and Logistic Regression are used for prediction. Results produced from all the selected algorithms are compared based on performance metrics such as precision, recall, F-measure, ROC, confusion metrics, accuracy, etc. Then comparing the performance of several algorithms, the testing results for some new samples are discussed.

4.2 SOIL CLASSIFICATION PREDICTION

Soil is a composition of organic materials, minerals, organisms, air, and water that provide the medium for plant growth. There are different types of soil, and there are various features of each soil. Based on these various features, several kinds of crops are grown. In this work, there are seven soil attributes, namely pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron, are used as input for soil classification. Several artificial intelligence techniques such as support vector machine, decision tree, multilayer perceptron, random forest, logistic regression, and Naïve Bayes are used for prediction.

4.2.1 Training Results

For classifying soil, the performance of different algorithms is necessary to disclose. The training results of the different classifier is described below-

a) Support Vector Machine

For train the dataset with Support Vector Machine classifier, load the dataset and initialize the input parameter and target value. There are seven parameters used as input. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.1 stands for the confusion matrix by support vector machine. This figure represents the matrix of 3 soil classes. The green cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

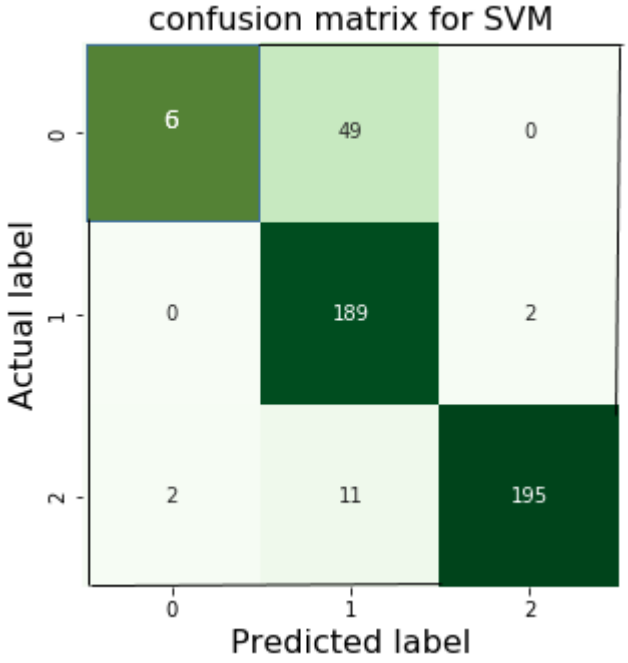


Figure 4.1: Confusion matrix for support vector machine.

In figure 4.1 above, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For classifying the soil, SVM provides an accuracy which is 85.90%. The following table 4.1 shows the precision, recall, f-measure, and accuracy score.

Table 4.1: Performance analysis of SVM

Performance metrics	Score
Precision	86.37%
F1-score	82.57%
Recall	85.90%
Accuracy	85.90%

Precision value is the proportion of correct positive classifications from cases that predict as positive. It is the accuracy of the positive prediction. SVM provides precision score is 86.37%. The recall is the proportion of correct positive classification (True positive) from cases that are actually positive. SVM provides recall is 85.90%. F-measure combines both precision and recall. For Soil classification, SVM provides F1-score is 82.57%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{6 + 189 + 195}{454} = \frac{390}{454} = 85.90\%$$

Here, SVM provides 85.90% accuracy for classifying soil.

Figure 4.2 is for the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.86; clay loam is 0.92, and loam is 1.00. The ROC score for SVM is 92.65%.

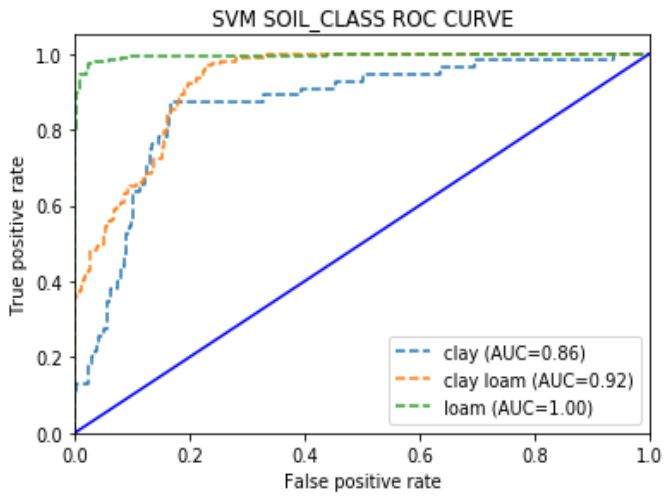


Figure 4.2: ROC Curve for measuring the performance of the SVM classifier.

b) Decision Tree

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. For train the dataset with Decision Tree classifier, load the dataset and initialize the input parameter and target value. There are seven parameters used as input for classifying soil. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.3 stands for the confusion matrix by Decision Tree. This figure represents the matrix of 3 soil classes. The green cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

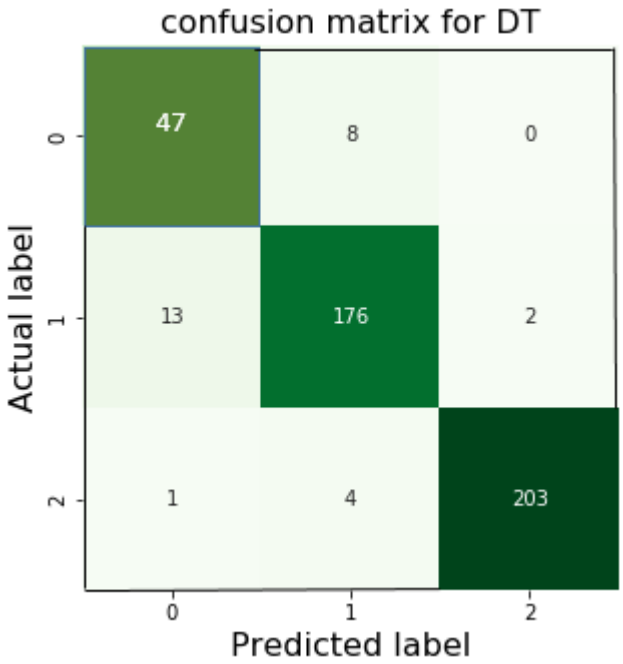


Figure 4.3: Confusion Matrix for Decision Tree.

In figure 4.3 above, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the classification of soil, DT provides an accuracy which is 93.83%. The following table shows the precision, recall, f-measure, and accuracy score:

Table 4.2: Performance analysis of DT

Performance metrics	Score
Precision	94.09%
Recall	93.83%
F1-score	93.93%
Accuracy	93.83%

Decision tree provides a precision score is 94.09%. The decision tree provides recall is 93.83%. F-measure combines both precision and recall. For Soil classification, DT provides F1-score is 93.93%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{47 + 176 + 203}{463} = \frac{426}{454} = 93.83\%$$

Here, the Decision tree provides 93.83% accuracy for classifying soil.

Figure 4.4 is the DT ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.89, clay loam is 0.93, and loam is 0.98. The ROC score for DT is 93.65%.

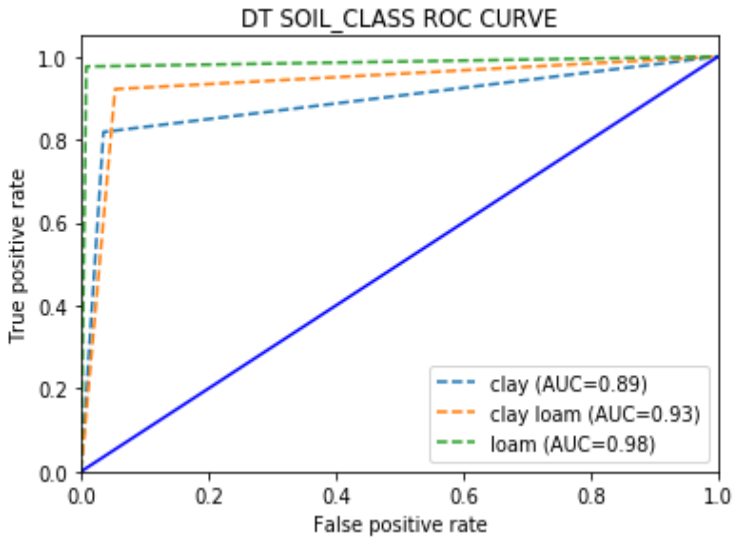


Figure 4.4: ROC Curve for measuring the performance of DT classifier.

c) Multilayer Perceptron Neural Network

For train the dataset with Multilayer Perceptron Neural Network, load the dataset, initialize input parameter and target value. There are seven parameters used as input for classifying soil. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.5 stands for the confusion matrix by MLPNN. This figure represents the matrix of 3 soil classes. The green cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

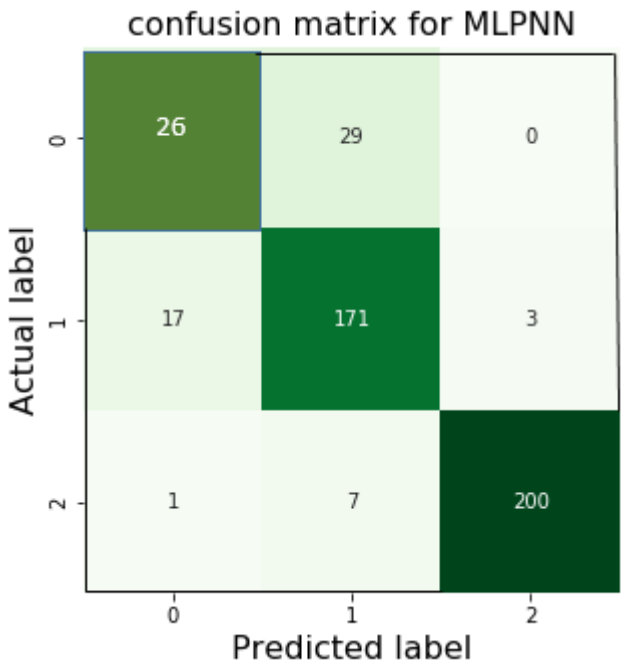


Figure 4.5: Confusion matrix for Multilayer Perceptron Neural Network

In the above figure, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For classifying the soil, MLPNN provides an accuracy which is 87.44%. The following table 4.3 shows the precision, recall, f-measure, and accuracy score.

Table 4.3: Performance analysis of MLPNN

Performance metrics	Score
Precision	87.05%
Recall	87.44%
F1-score	87.10%
Accuracy	87.44%

Multilayer perceptron provides the precision score for soil classification is 87.05%. The recall is the proportion of correct positive classification (True positive) from cases that are actually positive. MLPNN provides recall is 87.44%. F-measure combines both precision and recall. For Soil classification, MLPNN provides F1-score is 87.10%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{26 + 171 + 200}{454} = \frac{397}{454} = 87.44\%$$

Here, MLPNN provides 87.44% accuracy for classifying soil.

Figure 4.6 stands for the ROC curve to measure the performance of the MLPNN classifier. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.92, clay loam is 0.95, and loam is 0.99. The ROC score for MLPNN is 95.76%.

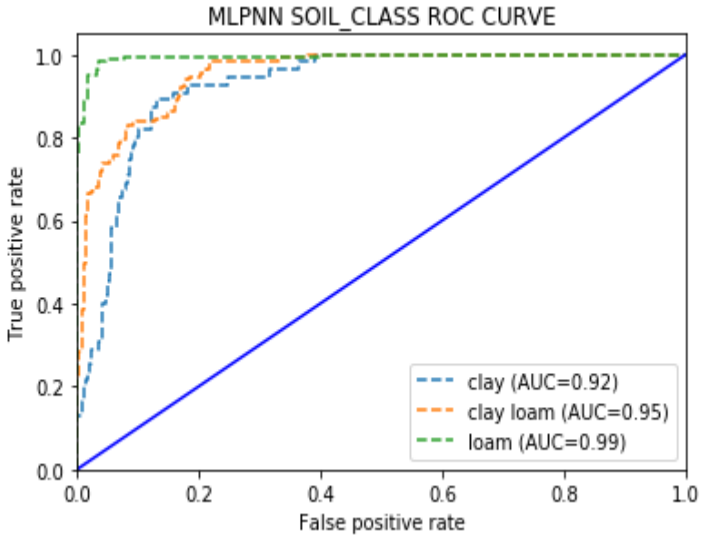


Figure 4.6: ROC Curve for measuring the performance of MLPNN classifier.

d) **Random Forest**

For train the dataset with the Random Forest classifier, load the dataset and initialize the input parameter and target value. There are seven parameters used as input for classifying soil. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.7 stands for the confusion matrix by the Random forest classifier. This figure represents the matrix of 3 soil classes. The green cell is shown where the output class is

matched with the target class. And the white cell is shown where the output class is not matched with the target class.

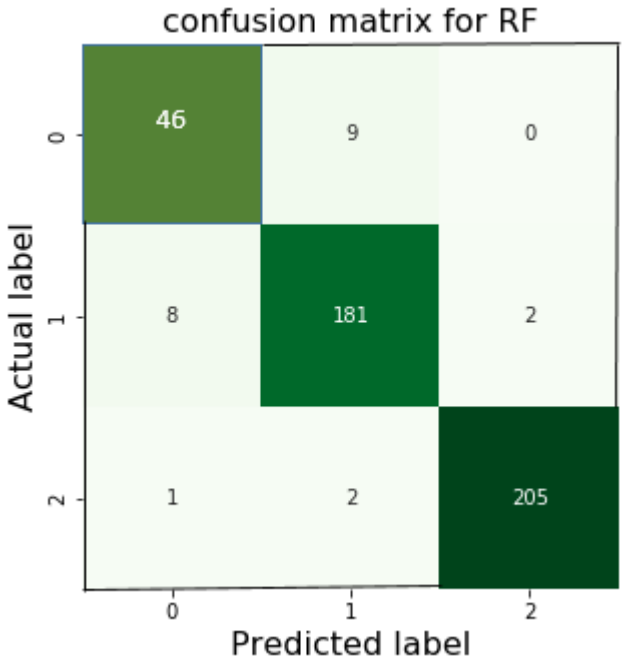


Figure 4.7: Confusion matrix for Random Forest.

In figure 4.7 above, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For classifying the soil, the RF classifier provides an accuracy which is 95.15%. The following table 4.4 shows the precision, recall, f-measure, and accuracy score.

Table 4.4: Performance analysis of Random Forest

Performance metrics	Score
Precision	95.16%
Recall	95.15%
F1-score	95.16%
Accuracy	95.15%

Random Forest provides the precision score for soil classification is 95.16%. The recall is the proportion of correct positive classification (True positive) from cases that are actually positive. Random forest classifier provides recall is 95.15%. F-measure combines both

precision and recall. For Soil classification, Random forest provides F1-score is 95.16%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{46 + 181 + 205}{454} = \frac{432}{454} = 95.15\%$$

Here, Random Forest provides 95.15% accuracy for classifying soil.

Figure 4.8 stands for the ROC curve to measure the performance of the Random forest classifier. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.99, clay loam is 0.99, and loam is 1.00. The ROC score for RF is 99.24%.

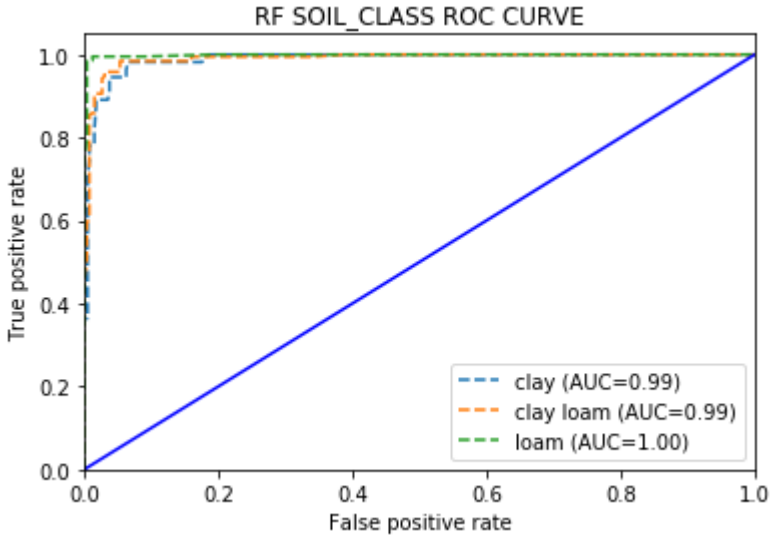


Figure 4.8: ROC Curve for measuring the performance of Random forest classifier.

e) Logistic Regression

For train the dataset with Logistic Regression, load the soil dataset and initialize the input parameter and target value. There are seven parameters used as input for classifying soil. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.9 stands for confusion matrix by Logistic Regression. This figure represents the matrix of 3 soil classes. The blue cell is shown where the output class is matched with the

target class. And the white cell is shown where the output class is not matched with the target class.

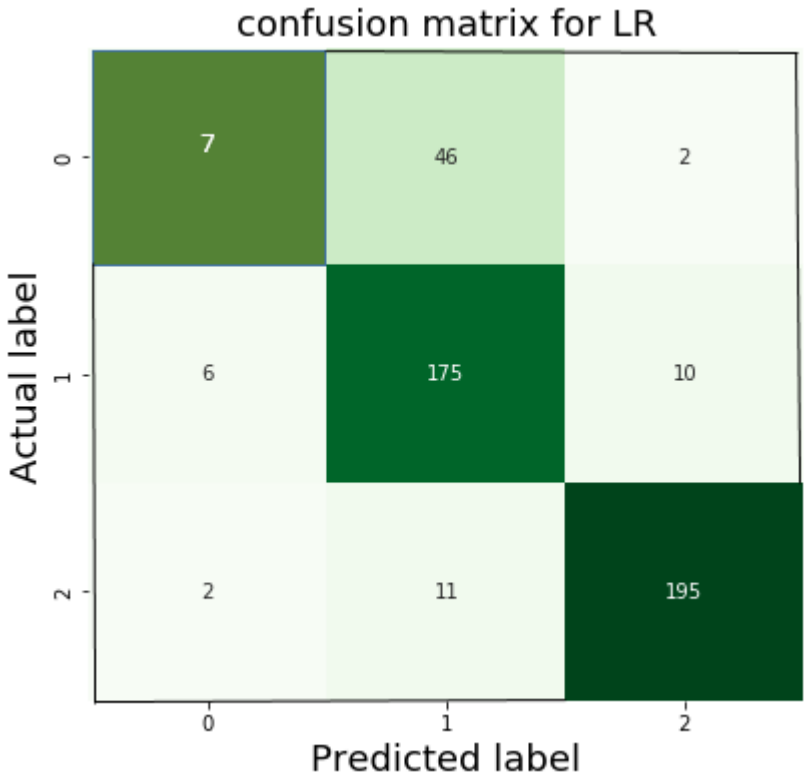


Figure 4.9: Confusion matrix for Logistic Regression.

In figure 4.9 above, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For classifying the soil, Logistic Regression provides an accuracy which is 83.04%. The following table 4.5 shows the precision, recall, f-measure, and accuracy score:

Table 4.5: Performance analysis of Logistic Regression

Performance metrics	Score
Precision	80.55%
Recall	83.04%
F1-score	80.29%
Accuracy	83.04%

Logistic Regression provides a precision score is 80.55%. The recall is the proportion of correct positive classification (True positive) from cases that are actually positive. Logistic Regression provides recall is 83.04%. F-measure combines both precision and recall. For

Soil classification, Logistic Regression provides F1-score is 80.29%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{7 + 175 + 195}{454} = \frac{377}{454} = 83.04\%$$

Here, Logistic Regression provides 83.04% accuracy for classifying soil.

Figure 4.10 stands for the ROC curve to measure the performance of Logistic Regression. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.83, clay loam is 0.93, and loam is 0.99. The ROC score for LR is 91.63%.

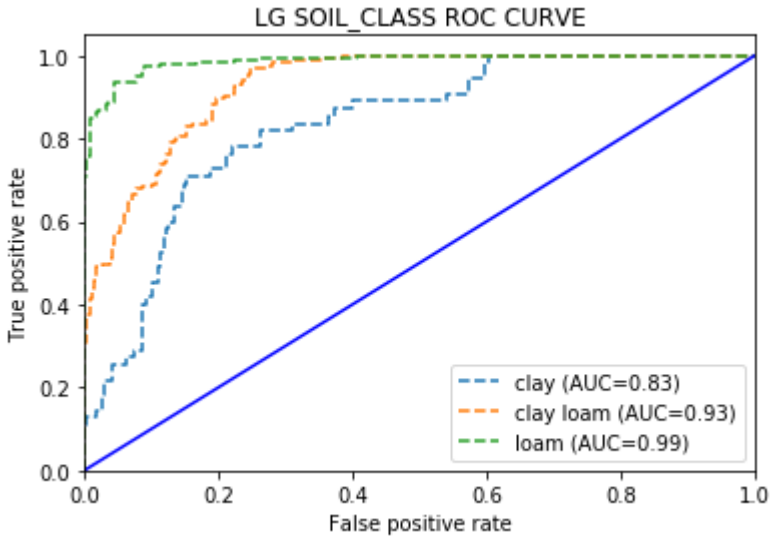


Figure 4.10: ROC Curve for measuring the performance of Logistic Regression.

f) Naïve Bayes Classifier

For train the dataset with the Naïve Bayes classifier, load the soil dataset and initialize the input parameter and target value. There are seven parameters used as input for classifying soil. These parameters are pH, salinity, organic matter, nitrogen, phosphorus-B, sulfur, and boron. The target column is "soil class" which consists of 3 soil classes. Three classes of soil are clay, loam, and clay loam.

Figure 4.11 stands for the confusion matrix by Naïve Bayes. This figure represents the matrix of 3 soil classes. The green cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

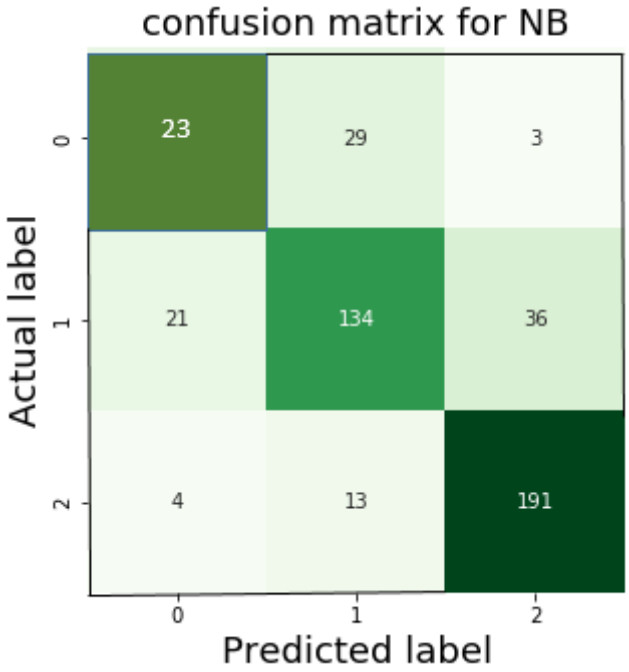


Figure 4.11: Confusion matrix for Naïve Bayes.

In figure 4.11 above, the model uses a dataset which is the soil dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For classifying the soil, Naïve Bayes provides an accuracy which is 76.65%. The following table shows the precision, recall, f-measure, and accuracy score:

Table 4.6: Performance analysis of Naïve Bayes

Performance metrics	Score
Precision	75.88%
Recall	76.65%
F1-score	76.09%
Accuracy	76.65%

Naïve Bayes provides precision score is 75.88%. The recall is the proportion of correct positive classification (True positive) from cases that are actually positive. Naïve Bayes provides recall 76.65%. F-measure combines both precision and recall. For Soil classification, Naïve Bayes provides F1-score is 76.09%. The calculation of accuracy using the value of the confusion matrix is-

$$Accuracy = \frac{23 + 134 + 191}{454} = \frac{348}{454} = 76.65\%$$

Here, Naïve Bayes provides 76.65% accuracy for classifying soil.

Figure 4.12 stands for the ROC curve to measure the performance of Naïve Bayes. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The area under for clay is 0.90, clay loam is 0.88, and loam is 0.95. The ROC score for NB is 91.12%.

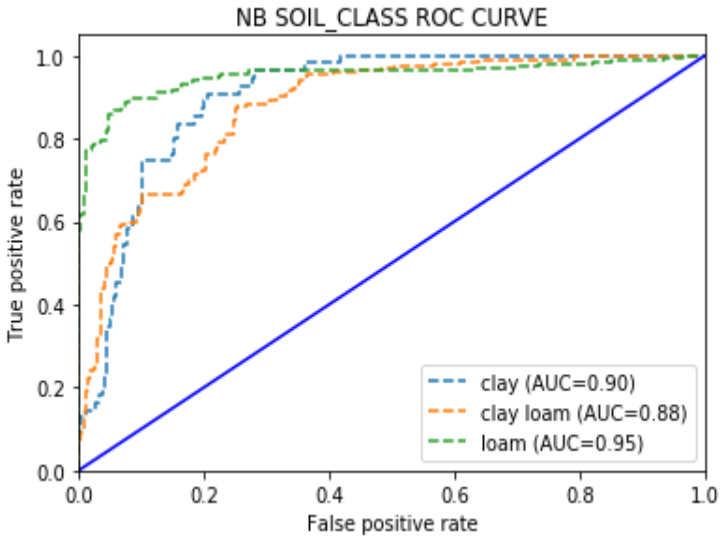
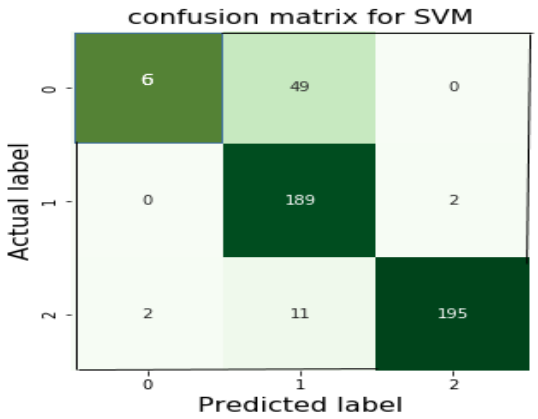


Figure 4.12: ROC Curve for measuring the performance of Naïve Bayes.

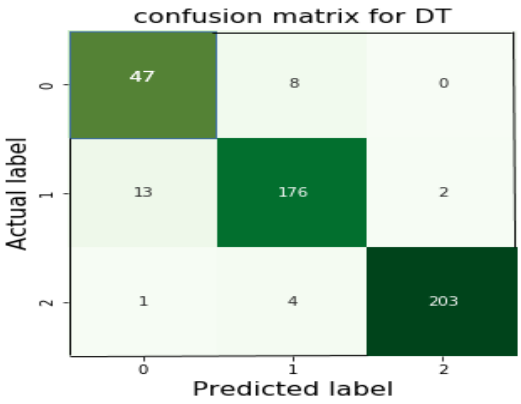
4.2.2 Comparison among Distinct Techniques for Soil Class Prediction

Comparison of different performance terminologies like confusion matrix, precision, recall, f-measure, accuracy, ROC curve, etc., of different artificial intelligence techniques for soil classification is shown below:

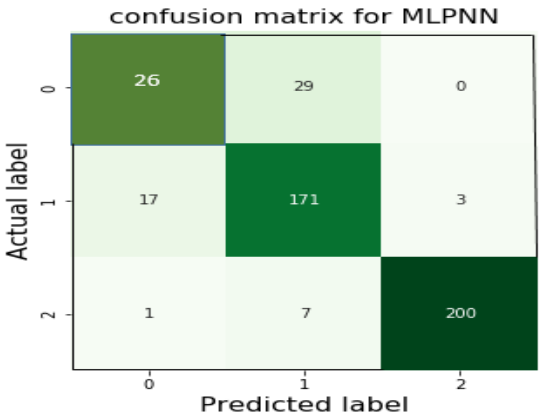
Comparison of the confusion matrix between different classifiers is shown in figure 4.13. In the figure 4.13(a) represents confusion matrix of the SVM algorithm, (b) represents confusion matrix of DT algorithm, (c) represents confusion matrix of MLPNN algorithm, (d) represents confusion matrix of RF algorithm, (e) represents confusion matrix of Logistic Regression and (f) represents confusion matrix of NB algorithm. This figure 4.13 shows that SVM provides accuracy is 85.9%, DT accuracy is 93.83%, MLPNN accuracy is 87.44%, RF accuracy is 95.15%, Logistic regression accuracy is 83.04%, and NB accuracy is 76.65%.



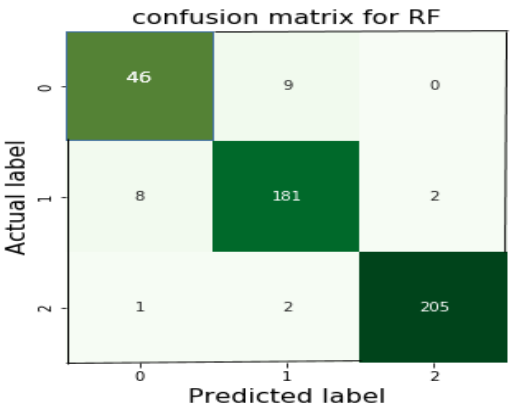
a) SVM



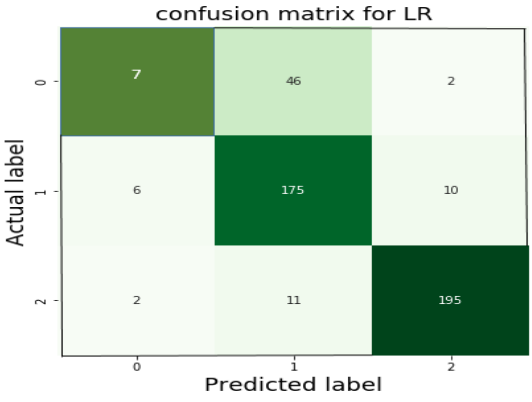
b) DT



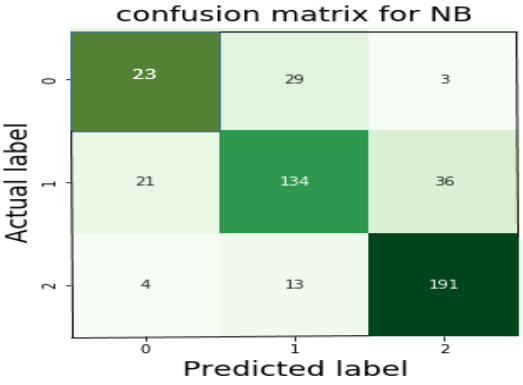
c) MLPNN



d) RF



e) LR

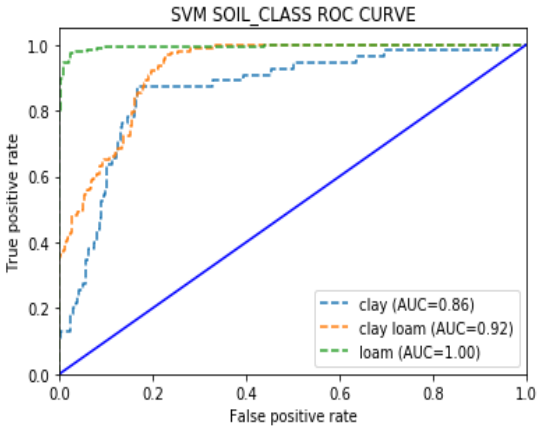


f) NB

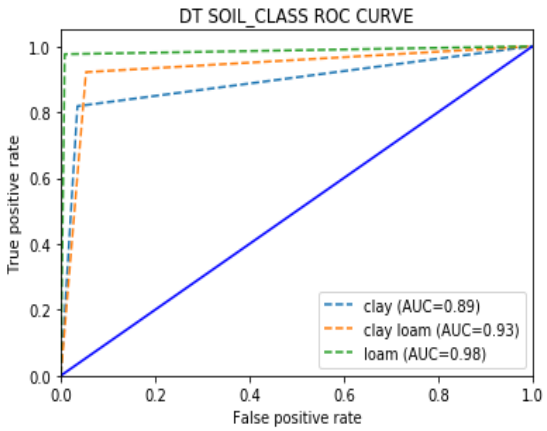
Figure 4.13: Comparison of confusion matrix between different classification algorithms- (a) SVM, (b) DT, (c) MLPNN, (d) RF, (e) LR, and (f) NB.

Comparison of ROC between different classification algorithms is shown in figure 4.14. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. In this figure 4.14, (a) represents ROC of SVM algorithm, (b) represents ROC of DT algorithm, (c) represents ROC of MLPNN algorithm, (d) represents ROC of RF algorithm, (e) represents ROC of LG algorithm and (f) represents ROC of NB algorithm.

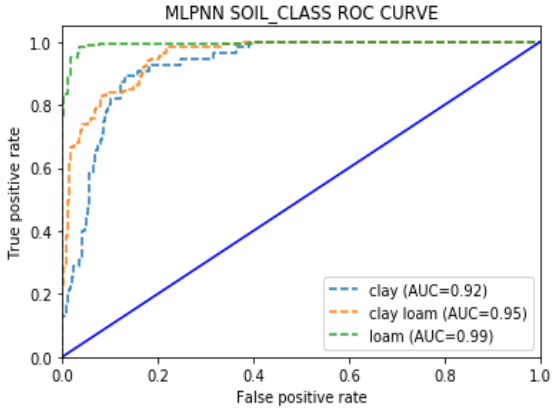
In figure 4.14(a) shows that area under ROC for clay is 0.83, clay loam is 0.91 and loam is 0.99, (b) shows that area under ROC for clay is 0.88, clay loam is 0.92 and loam is 0.98, (c) shows that area under ROC for clay is 0.90, clay loam is 0.95 and loam is 0.99, (d) shows that area under ROC for clay is 0.98, clay loam is 0.99 and loam is 1.00, (e) shows that area under ROC for clay is 0.85, clay loam is 0.92 and loam is 0.99 and (f) shows that area under ROC for clay is 0.92, clay loam is 0.90 and loam is 0.97. Here, ROC for the Random forest algorithm is higher than other algorithms.



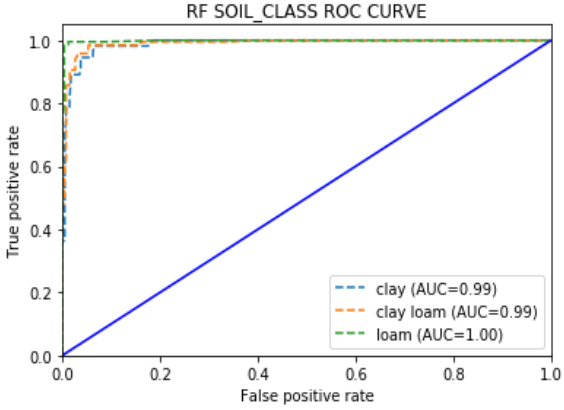
a) Support vector machine



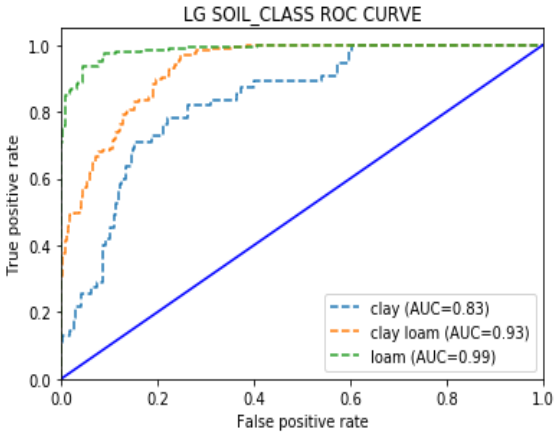
b) Decision Tree



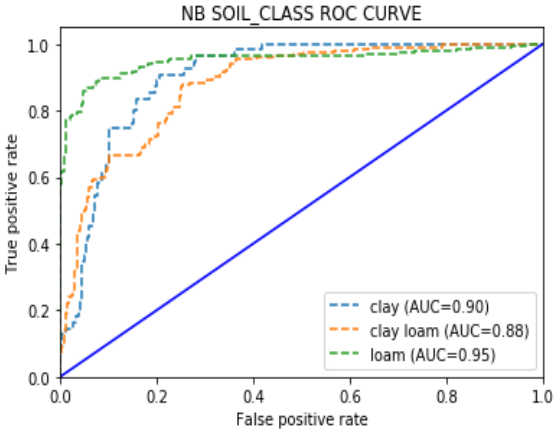
c) Multilayer perceptron



d) Random Forest



e) Logistic Regression



f) Naïve Bayes

Figure 4.14: Comparison of ROC between different classification algorithms- (a) SVM, b) DT, c) MLPNN, d) RF, e) LR, and f) NB.

Comparison table of prediction accuracy, error, and ROC between SVM, DT, MLPNN, RF, LG, and NB algorithms is shown in table 4.7.

Table 4.7: Comparison of accuracy, error, and roc between different algorithms

Name of algorithm	Accuracy (%)	Error (%)	ROC
SVM	85.90	14.1	92.65
DT	93.83	6.17	93.65
MLPNN	87.44	12.56	95.76
RF	95.15	4.85	99.24
LG	83.04	16.96	91.63
NB	76.65	23.35	91.12

In above table 4.7 shows that the RF algorithm's prediction accuracy is higher than other algorithms, which value is 95.15%. And ROC score for RF is also higher than other classifiers.

The following figure shows the comparison of different classification algorithms on the basis of prediction accuracy.

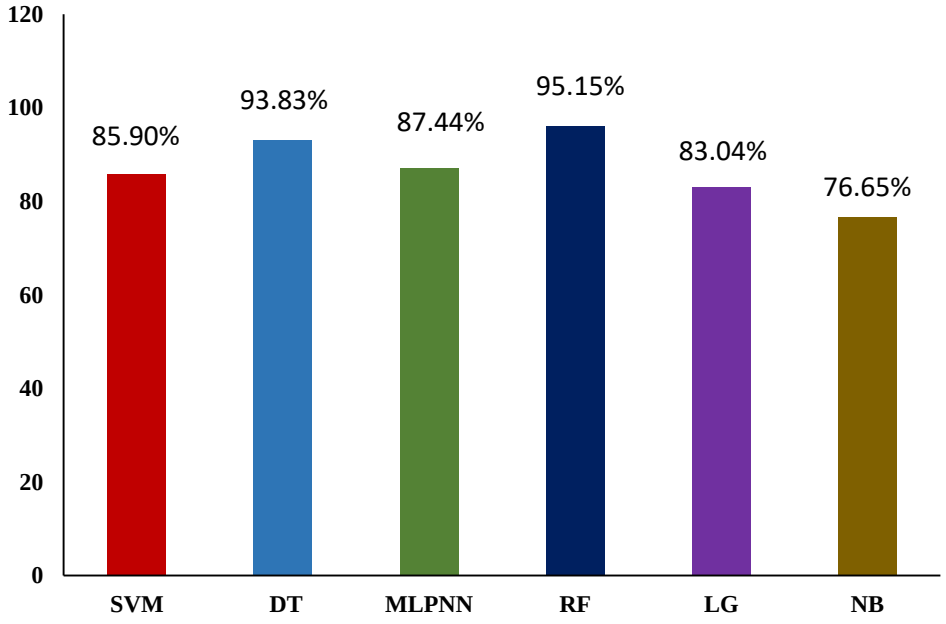


Figure 4.15: Comparison of accuracy score between different algorithms for soil classification.

Comparison of error measurement between different algorithms is shown in the following figure 4.16.

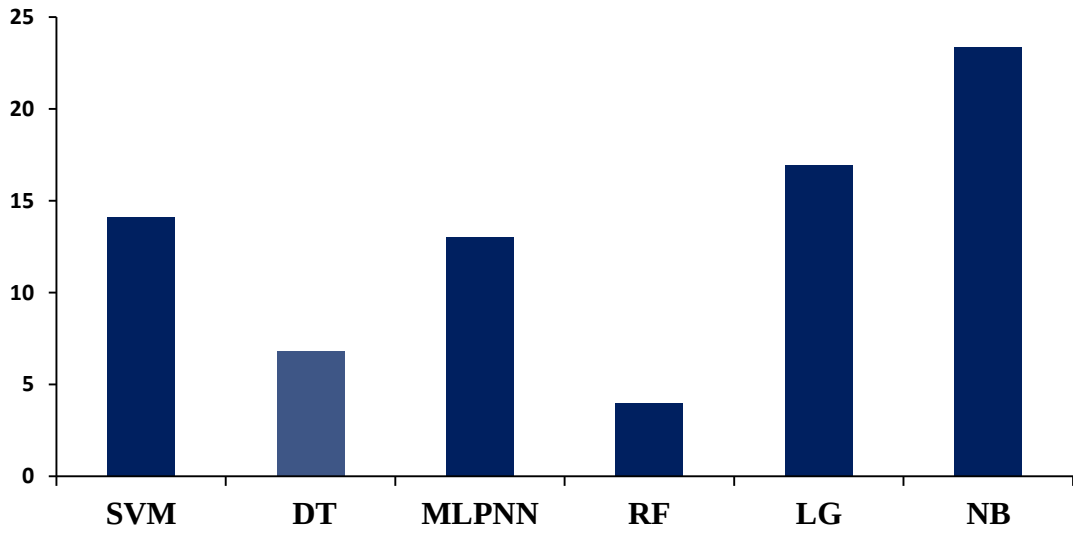


Figure 4.16: Error (%) of different algorithms for soil classification.

The above figure 4.15 and figure 4.16 shows that the RF algorithm has a higher accuracy rate which is 95.15%, and a lower error rate which is 4.85%.

Comparison table of prediction accuracy, precision, recall, and f1-score between SVM, DT, MLPNN, RF, LG, and NB algorithms is shown in table 4.8.

Table 4.8: Comparison of performance between different algorithms.

Name of algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Support Vector Machine	85.90	86.37	85.90	82.57
Decision Tree	93.83	94.09	93.83	93.93
Multilayer Perceptron NN.	87.44	87.05	87.44	87.10
Random Forest	95.15	95.16	95.15	95.16
Logistic Regression	83.04	80.55	83.04	80.29
Naïve Bayes	76.65	75.88	76.65	76.09

The following figure shows the comparison of performance evaluation for different algorithms on the basis of precision, recall, f-measure, and accuracy.

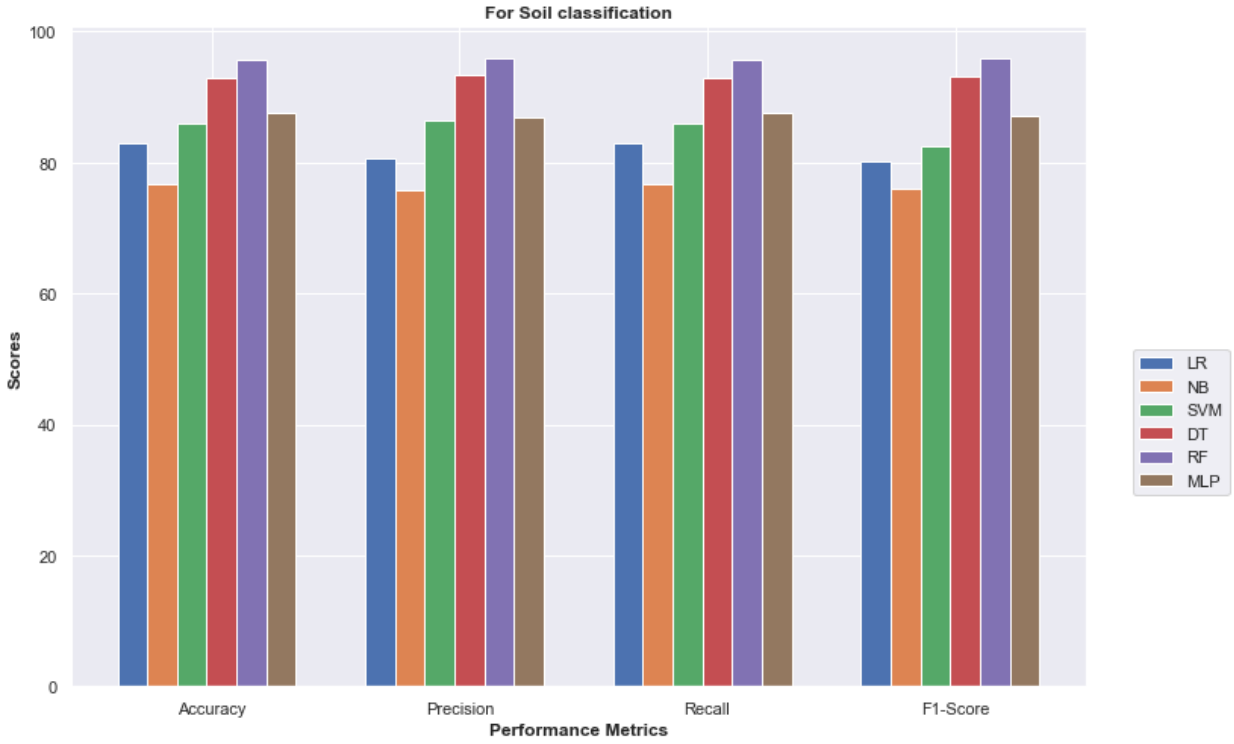


Figure 4.17: Comparison of performance metrics for different algorithms.

From the above figure 4.17 visibly shows that Random Forest (RF) algorithm has maximum accuracy is 95.15%, precision is 95.16%, recall is 95.15%, and f1-score is 95.16% which is larger than other algorithms.

From the discussion of this section, we selected the Random Forest algorithm for soil classification. Random forest is developed on the basis of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance model.

4.2.3 Testing

Now Random Forest (RF), Support Vector Machine, Decision Tree, Multilayer perceptron, logistic regression Naïve Bayes algorithm is used for testing with new soil samples which were unused before. So, test the classifier with the new 30 instances. The new

instance data set is in the main dataset. The last 30 samples of the total samples are used for prediction accuracy. These 30 samples are not used in training and testing before. All data samples are shown in the appendices section. One of the main tasks of the research is to show the prediction with new samples since the main goal of the research is to find out the best algorithm which will give better accuracy than the early classification system of soil classification. The following table 4.9 shows the results of new testing samples.

Table 4.9: Target and Predicted value of new samples for different algorithms

Sample No	**Target**	**Predicted**					
		Random Forest	**SVM**	**MLPNN**	**Decision Tree**	**Logistic Regress.**	**Naïve Bayes**
1	Clay loam	Clay	Clay loam	Clay loam	Clay	Clay loam	Clay
2	Loam	Loam	Loam	Loam	Loam	Loam	Loam
3	Clay	Clay	Clay loam	Clay loam	Clay	Clay loam	Loam
4	Loam	Loam	Loam	Loam	Loam	Loam	Loam
5	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
6	Loam	Loam	Loam	Loam	Loam	Loam	Loam
7	Loam	Loam	Loam	Loam	Loam	Loam	Loam
8	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
9	Clay loam	Clay loam	Clay loam	Clay	Clay loam	Clay loam	Clay loam
10	Loam	Loam	Loam	Loam	Loam	Loam	Loam
11	Loam	Loam	Loam	Loam	Loam	Loam	Loam
12	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay

13	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
14	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Loam
15	Loam	Loam	Loam	Loam	Loam	Loam	Loam
16	Loam	Loam	Loam	Loam	Loam	Loam	Loam
17	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
18	Loam	Loam	Loam	Clay loam	Loam	Clay loam	Clay loam
19	Loam	Loam	Loam	Loam	Loam	Loam	Loam
20	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
21	Clay loam	Clay loam	Clay loam	Clay	Clay loam	Clay loam	Clay loam
22	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
23	Loam	Loam	Loam	Loam	Loam	Loam	Loam
24	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
25	Clay	Clay	Clay loam	Clay loam	Clay	Clay	Clay loam
26	Loam	Loam	Loam	Loam	Loam	Clay loam	Loam
27	Loam	Loam	Loam	Loam	Loam	Loam	Loam
28	Loam	Loam	Loam	Loam	Loam	Loam	Loam
29	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam
30	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay loam	Clay
Accuracy		96.67%	93.33%	83.33%	96.67%	90%	76.66%

Here, the testing accuracy of random forest is 96.67%, SVM is 93.33%, MLPNN is 83.33%, Decision tree is 96.67%, Logistic Regression is 90%, and Naïve Bayes is 76.66%.

4.3 CROPS CULTIVATION PREDICTION

Rapid and accurate crop cultivation prediction is of great significance for agricultural management and sustainable development. In this work, there are four attributes, namely location, soil class, season, and soil pH level, are used as input for crop cultivation prediction. Several artificial intelligence techniques such as support vector machine, decision tree, multilayer perceptron, random forest, naïve bayes, and logistic regression are used for prediction.

4.3.1 Training Results

For the prediction of crop cultivation, the performance of different algorithms is necessary to disclose. The performance of different classifiers is described below-

a) Support Vector Machine

For train the crop dataset with the Support Vector Machine classifier, load the dataset and initialize the input parameter and target value. There are four parameters used as input. These parameters are location, soil class, season, and soil pH. The target is crops which consist of 29 crops.

Figure 4.18 stands for the confusion matrix by support vector machine. This figure represents the matrix of 29 crops. The red cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

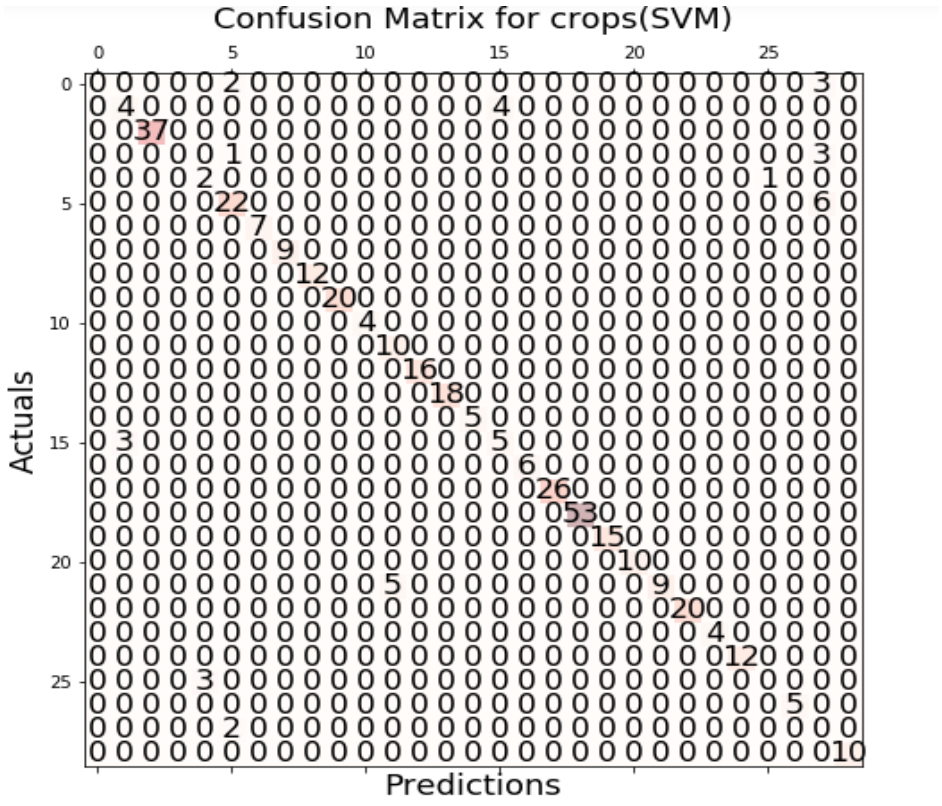


Figure 4.18: Confusion matrix for crop cultivation using SVM.

In the above figure, the model uses a dataset which is a crops dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, SVM provides an accuracy which is 91.18%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.10: Performance analysis of SVM for crops prediction

Performance metrics	Score
Precision	91.63%
Recall	91.18%
F1-score	91.13%
Accuracy	91.18%

In the above table, SVM provides a precision is 91.63%, recall is 91.18%, F1-score is 91.13%, and accuracy is 91.18%.

Figure 4.19 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for SVM is 99.53%.

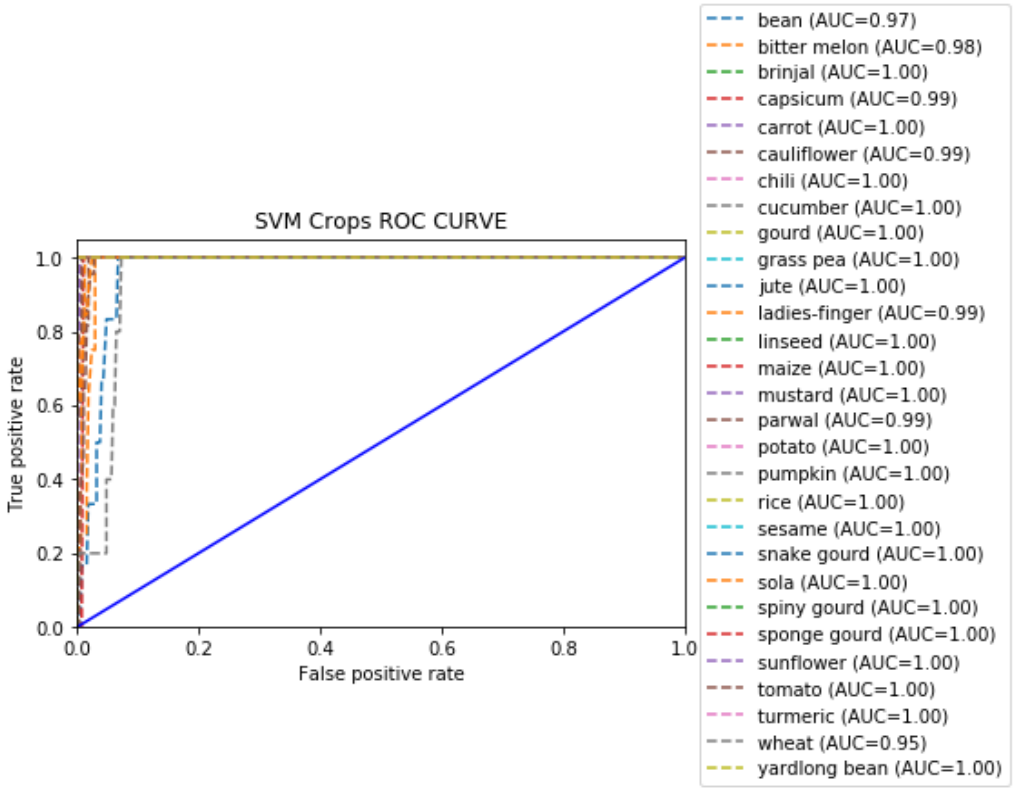


Figure 4.19: ROC Curve of SVM classifier for Crops prediction.

b) Decision Tree

For train the crops dataset with the Decision Tree algorithm, load the dataset and initialize the input parameter and target value. There are four parameters used as input. These parameters are location, soil class, season, and soil pH. The target is "crops" which consists of 29 crops.

Figure 4.20 stands for the confusion matrix by Decision Tree. This figure represents the matrix of 29 crops. The red cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

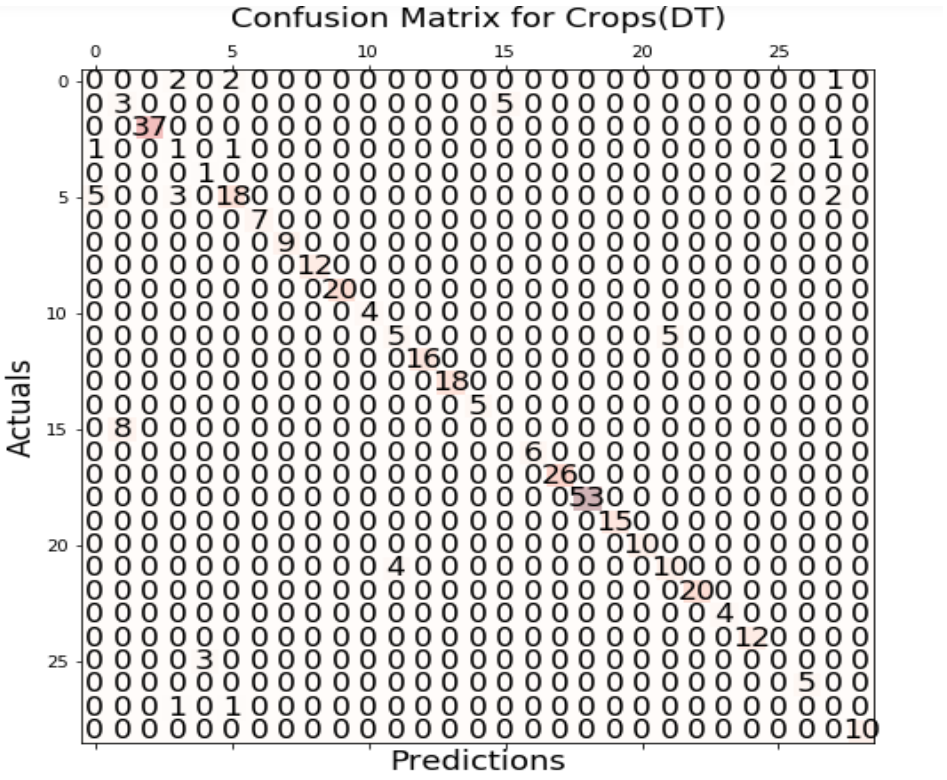


Figure 4.20: Confusion Matrix for crop cultivation using Decision Tree.

In the above figure, the model uses a dataset which is the crops dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, DT provides an accuracy which is 87.43%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.11: Performance analysis of DT for crops cultivation prediction

Performance metrics	Score
Precision	88.32%
Recall	87.43%
F1-score	87.75%
Accuracy	87.43%

In the above table, DT provides a precision is 88.32%, recall is 87.43%, F1-score is 87.75%, and accuracy is 87.43%.

Figure 4.21 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for DT is 88.79%.

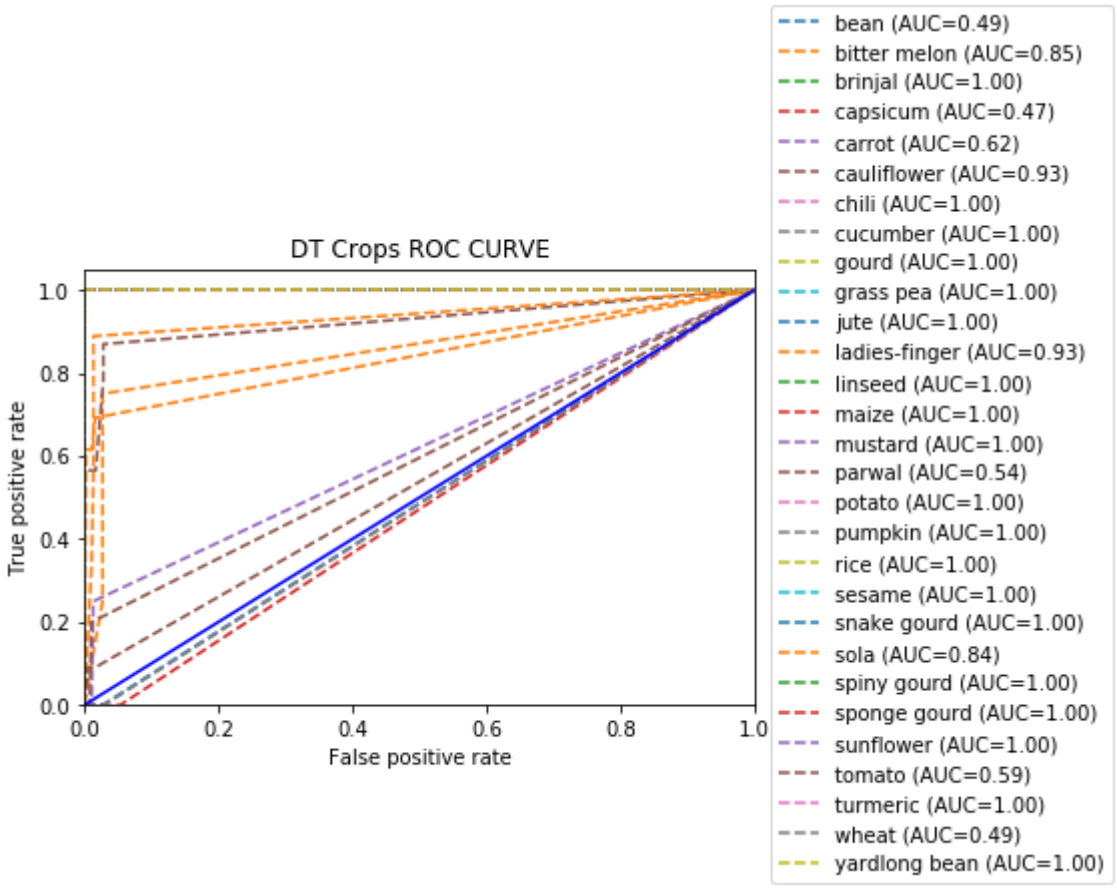


Figure 4.21: ROC Curve of DT classifier for crop prediction.

c) Multilayer Perceptron Neural Network

For train the crops dataset with Multilayer Perceptron Neural Network, load the dataset and initialize the input parameter and target value. There are 4 parameters are used as input. These parameters are location, soil class, season, and soil pH. The number of hidden layers is 13. The target column is "crops" which consists of 29 crops.

Figure 4.22 stands for the confusion matrix by Multilayer perceptron. This figure represents the matrix of 29 crops. The blue cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

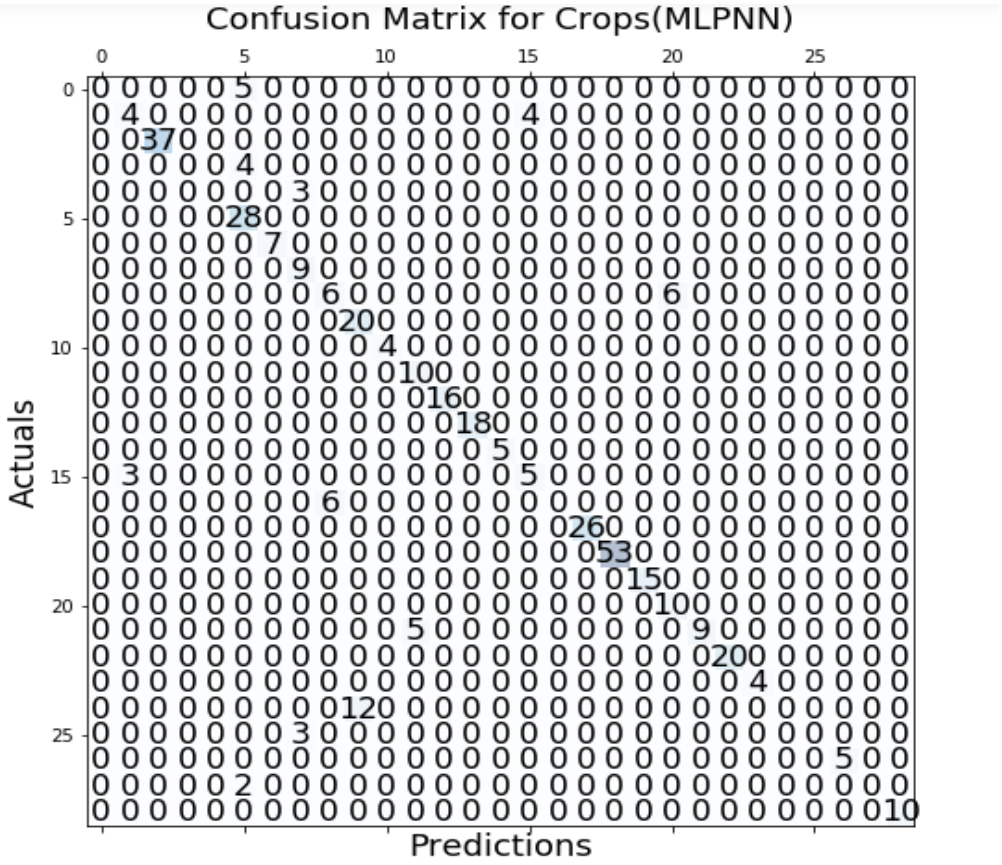


Figure 4.22: Confusion Matrix for Multilayer Perceptron Neural Network.

In the above figure, the model uses the dataset, which is the crops dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, MLPNN provides an accuracy which is 85.83%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.12: Performance analysis of MLPNN for crops cultivation prediction

Performance metrics	Score
Precision	80.20%
Recall	85.83%
F1-score	82.13%
Accuracy	85.83%

In the above table, MLPNN provides a precision is 80.20%, recall is 85.83%, F1-score is 82.13%, and accuracy is 85.83%.

Figure 4.23 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for MLPNN is 95.13%.

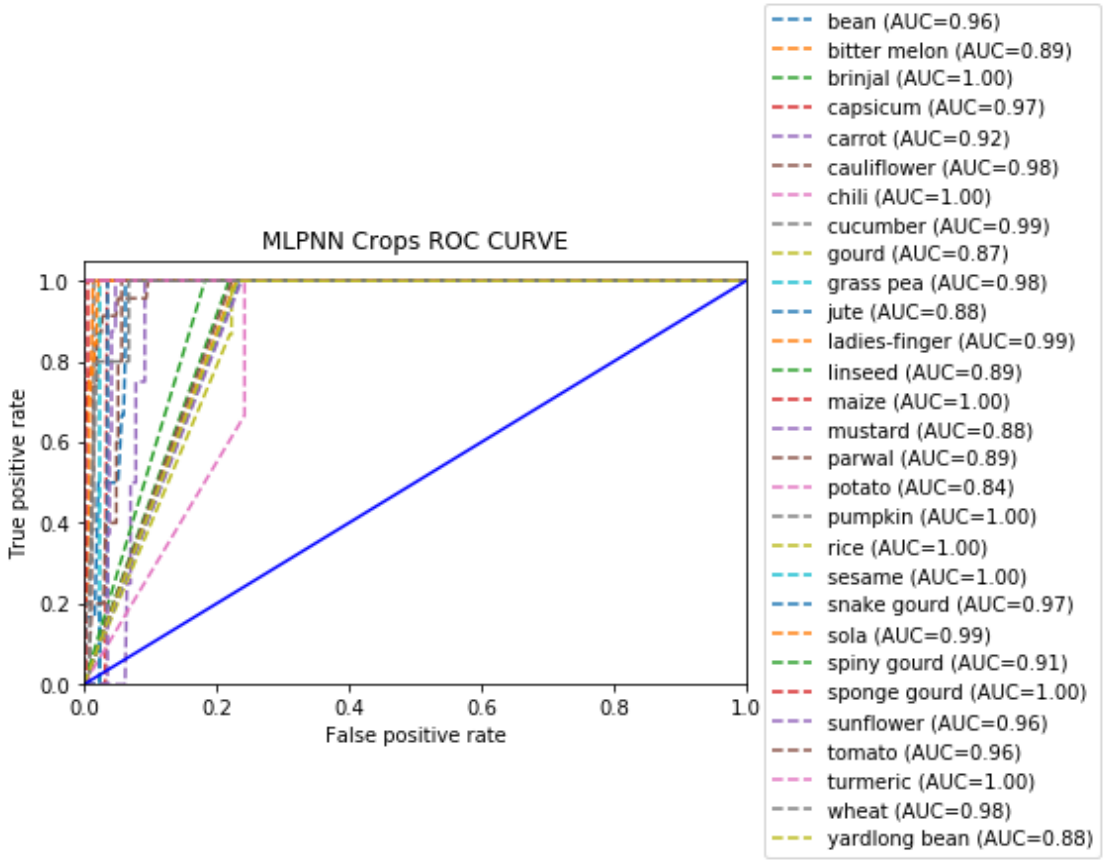


Figure 4.23: ROC Curve of MLPNN classifier for crop prediction.

d) Random Forest

For train the crops dataset with Random Forest classifier, load the dataset and initialize the input parameter and target value. There are four parameters used as input. These parameters are location, soil class, season, and soil pH. The target column is "crops" which consists of 29 crops.

Figure 4.24 stands for the confusion matrix by Multilayer perceptron. This figure represents the matrix of 29 crops. The blue cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

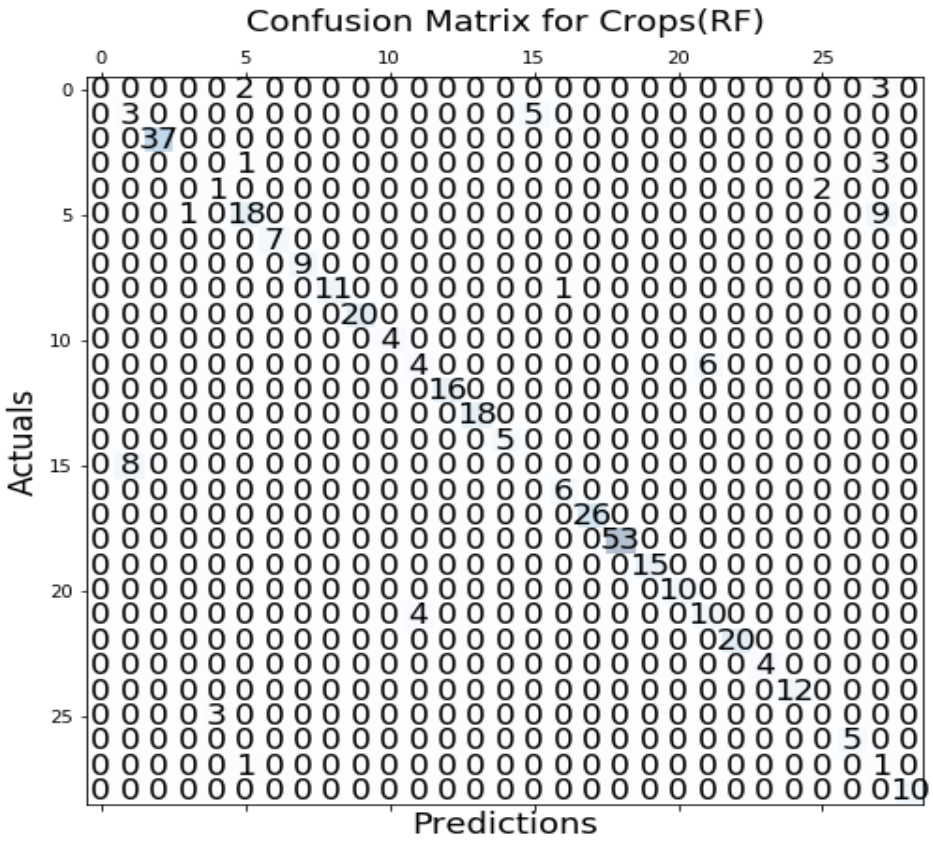


Figure 4.24: Confusion Matrix for Random Forest classifier.

In the above figure, the model uses a dataset, which is the crops dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, RF provides an accuracy which is 86.90%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.13: Performance analysis of RF for crops cultivation prediction

Performance metrics	Score
Precision	87.66%
Recall	86.90%
F1-score	87.05%
Accuracy	86.90%

In the above table, RF provides a precision is 87.66%, recall is 86.90%, F1-score is 87.05% and accuracy is 86.90%.

Figure 4.25 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for RF is 96.6%.

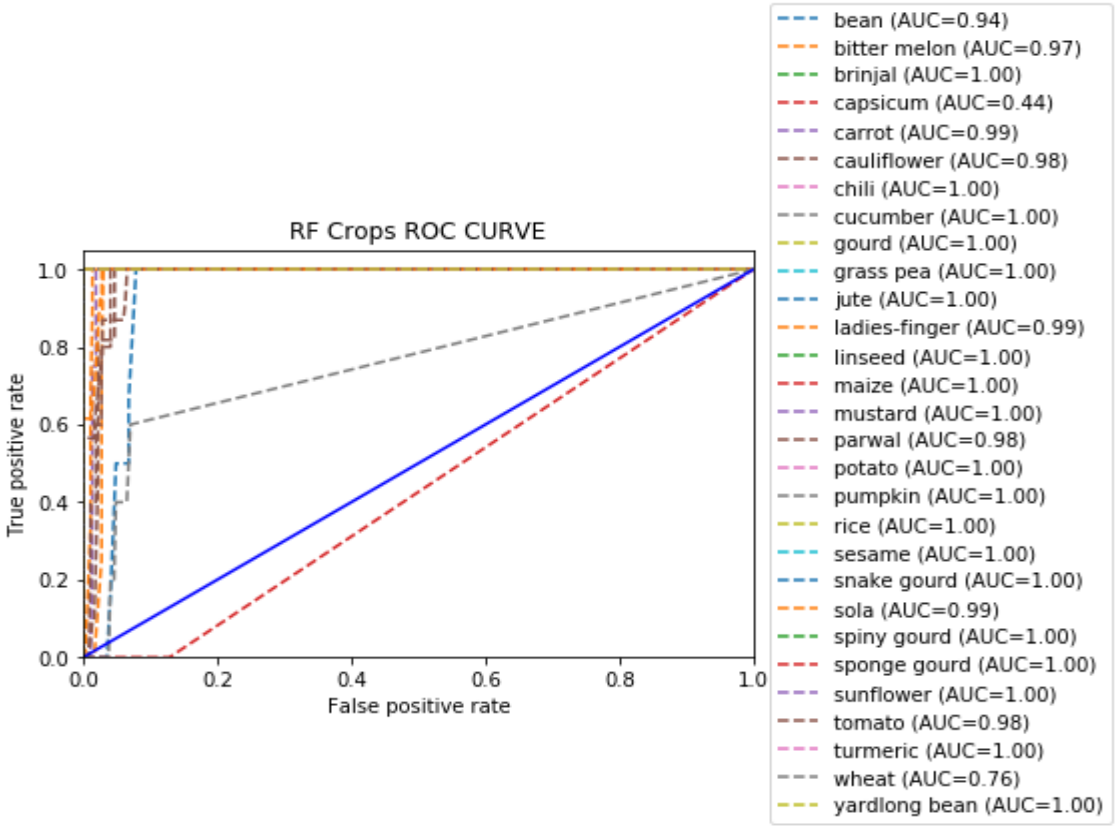


Figure 4.25: ROC Curve of RF classifier for crop prediction.

e) Logistic Regression

For train the crops dataset with Logistic Regression, load the dataset and initialize the input parameter and target value. There are four parameters used as input. These parameters are location, soil class, season, and soil pH. The target column is "crops" which consists of 29 crops.

Figure 4.26 stands for confusion matrix by Logistic Regression. This figure represents the matrix of 29 crops. The blue cell is shown where the output class is matched with the target class. And the white cell is shown where the output class is not matched with the target class.

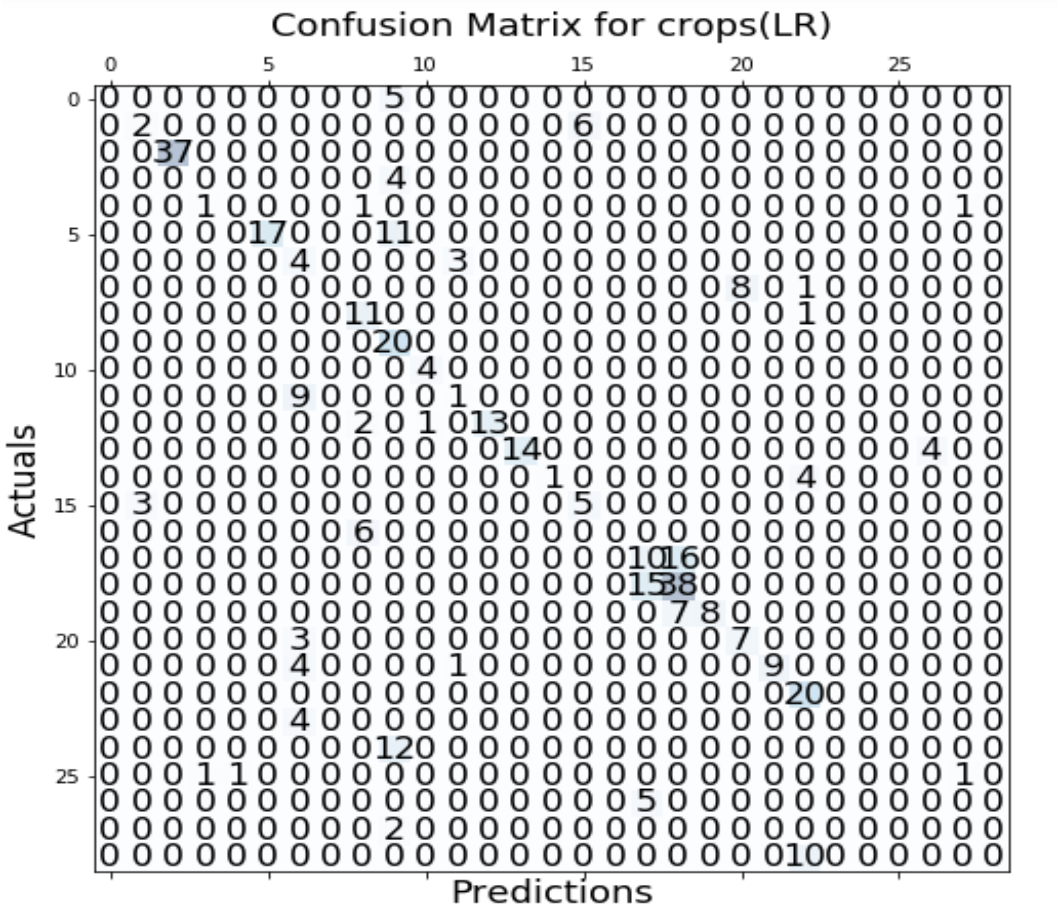


Figure 4.26: Confusion Matrix of LR for crop prediction.

In the above figure, the model uses a dataset which is the crops dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, Logistic Regression provides an accuracy which is 59.09%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.14: Performance analysis of LR for crops cultivation prediction

Performance metrics	Score
Precision	58.20%
Recall	59.09%
F1-score	55.68%
Accuracy	59.09%

In the above table, Logistic Regression provides a precision is 58.20%, recall is 59.09%, F1-score is 55.68%, and accuracy is 59.09%.

Figure 4.27 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for LR is 97.66%.

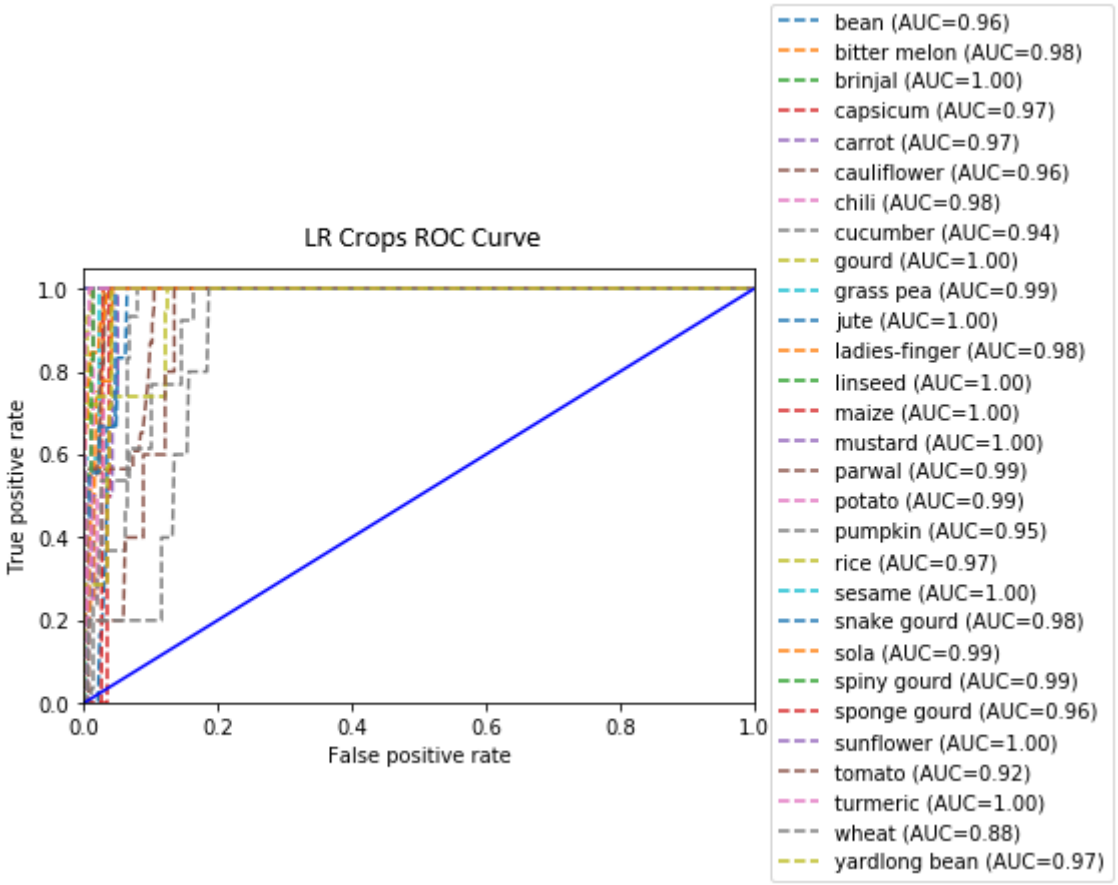


Figure 4.27: ROC curve of Logistic Regression for crops prediction.

f) Naïve Bayes Classifier

For train the crops dataset with the Naïve Bayes classifier, load the dataset and initialize the input parameter and target value. There are four parameters used as input. These parameters are location, soil class, season, and soil pH. The target column is "crops" which consists of 29 crops.

Figure 4.28 stands for the confusion matrix by the Naïve Bayes classifier. This figure represents the matrix of 29 crops. The blue cell is shown where the output class is matched

with the target class. And the white cell is shown where the output class is not matched with the target class.

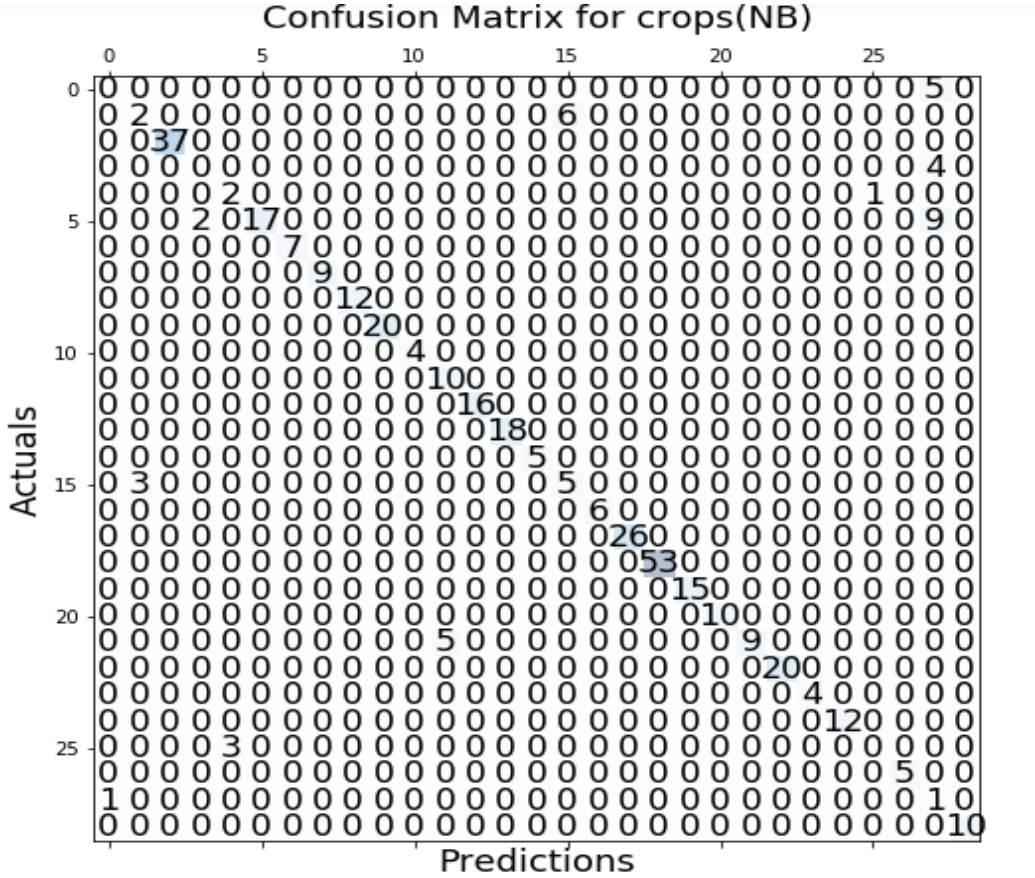


Figure 4.28: Confusion Matrix of Naïve Bayes for crop cultivation.

In the above figure, the model uses a dataset which is the crop dataset. Here, all rows show the actual classes, and all columns show the predicted classes. For the prediction of crop cultivation, Naïve Bayes provides an accuracy which is 89.57%. The following table shows the precision, recall, f-measure, and accuracy score.

Table 4.15: Performance analysis of NB for crops cultivation prediction

Performance metrics	Score
Precision	92.46%
Recall	89.57%
F1-score	90.23%
Accuracy	89.57%

In the above table, Logistic Regression provides precision is 92.46%, recall is 89.57%, F1-score is 90.23%, and accuracy is 89.57%.

Figure 4.29 shows the ROC curve for measuring performance. The Y axis stands for the true positive rate, and the X axis is for the false positive rate. The ROC score for NB is 99.42%.

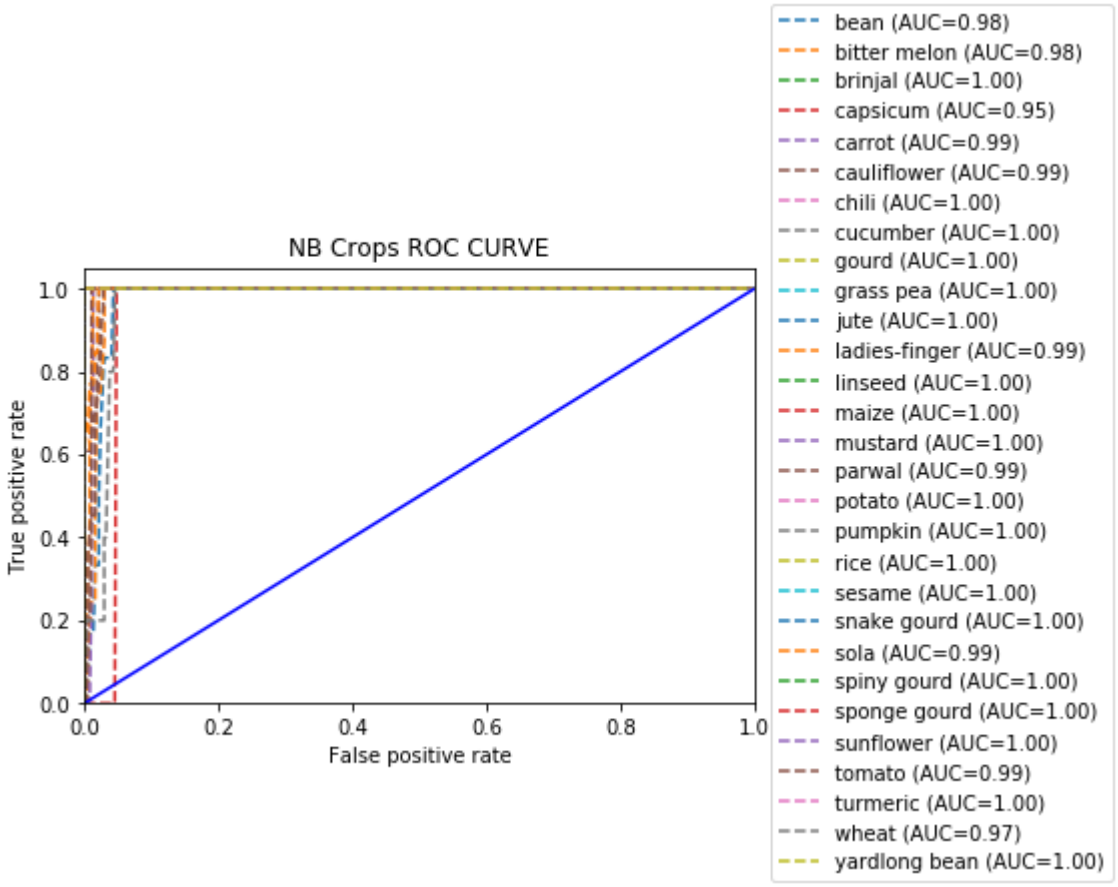


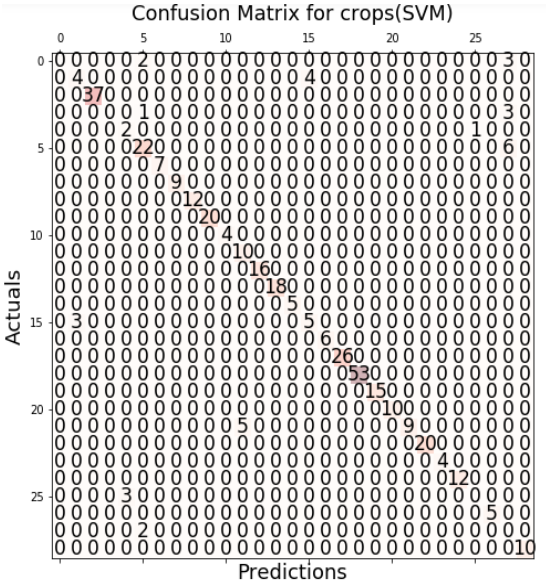
Figure 4.29: ROC curve of Naïve Bayes for crop prediction.

4.3.2 Comparative Analysis among Distinct Techniques

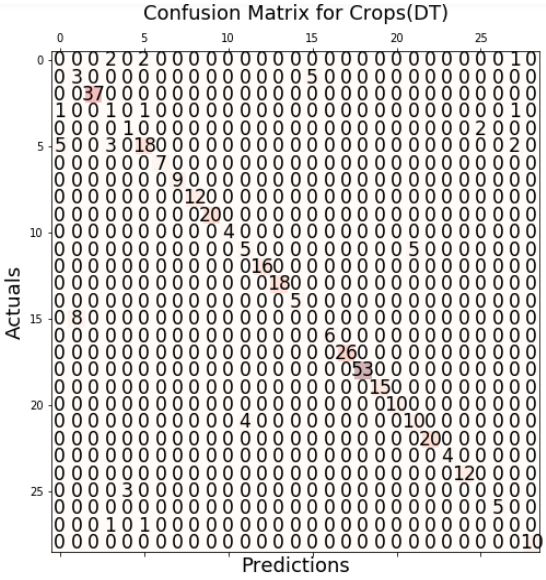
Comparison of different performance terminologies like confusion matrix, precision, recall, f-measure, accuracy, error, ROC curve, etc., of different artificial intelligence techniques for crops cultivation prediction is shown below:

Comparison of the confusion matrix between different techniques is shown in figure 4.30. In the figure of 4.30,(a) represents confusion matrix of SVM algorithm, (b) represents confusion matrix of DT algorithm, (c) represents confusion matrix of MLPNN algorithm, (d) represents confusion matrix of RF algorithm, (e) represents confusion matrix of Logistic Regression

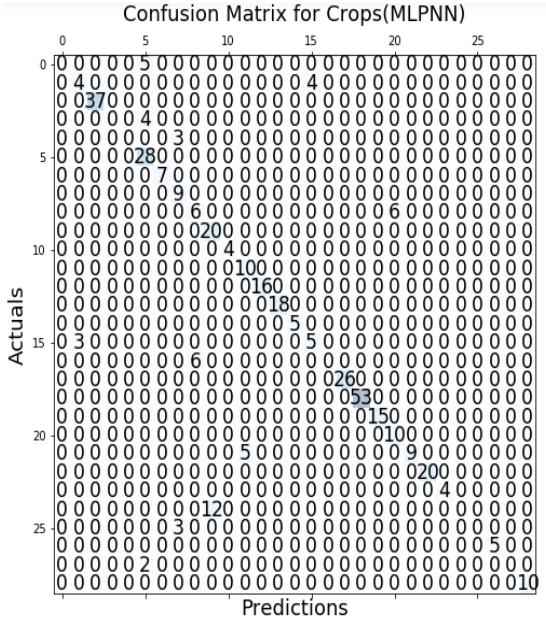
and (f) represents confusion matrix of NB algorithm. This figure shows that SVM provides accuracy is 91.18%, DT accuracy is 87.43%, MLPNN accuracy is 85.83%, RF accuracy is 86.90%, Logistic regression accuracy is 59.09%, and NB accuracy is 89.57%.



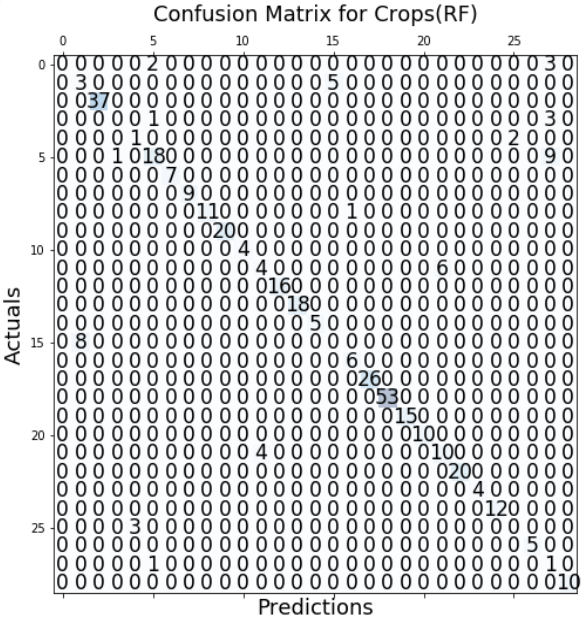
a) SVM



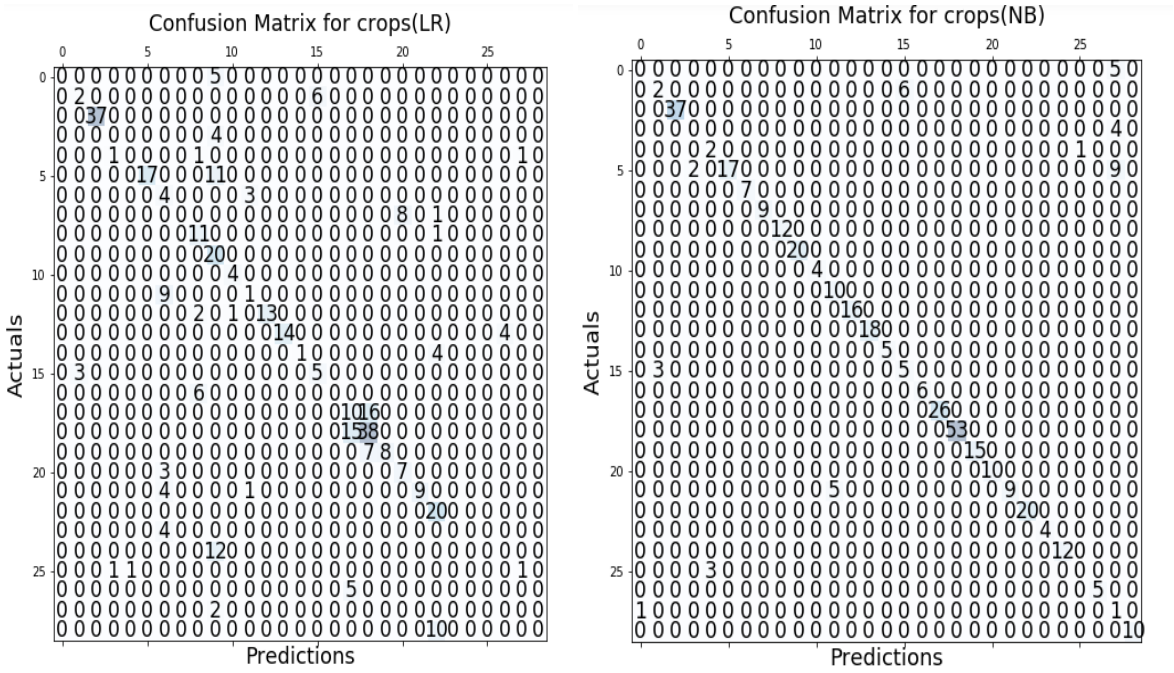
b) DT



c) MLPNN



d) RF



e) LR f) NB

Figure 4.30: Comparison of confusion matrix between different classification algorithms- (a) SVM, (b) DT, (c) MLPNN, (d) RF, (e) LR, and (f) NB.

Comparison table of prediction accuracy, error accuracy, and ROC between SVM, DT, MLPNN, RF, LG, and NB algorithms is shown in the following table 4.16.

Table 4.16: Comparison of accuracy, error, and roc between different algorithms

Name of algorithm	Accuracy (%)	Error (%)	ROC
Support Vector Machine	91.18	8.82	99.53
Decision Tree	87.43	12.57	88.79
Multilayer Perceptron NN	85.83	14.17	95.13
Random Forest	86.90	13.1	96.6
Logistic Regression	59.09	40.91	97.66
Naïve Bayes	89.57	10.43	99.42

The above table shows that the SVM algorithm's prediction accuracy is higher than other algorithms, which is 91.18%.

The following figure shows the comparison of different classification algorithms on the basis of prediction accuracy.

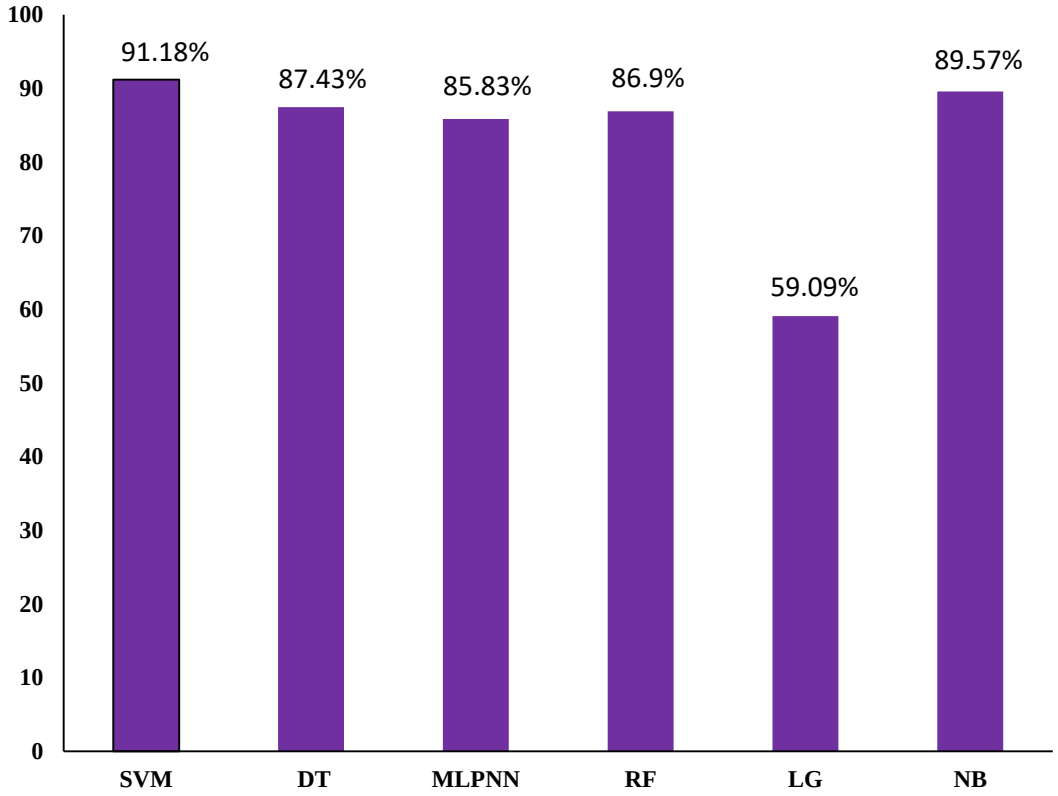


Figure 4.31: Accuracy score of different algorithms for crop prediction.

Comparison of error measurement between different algorithms is shown in the following figure 4.32.

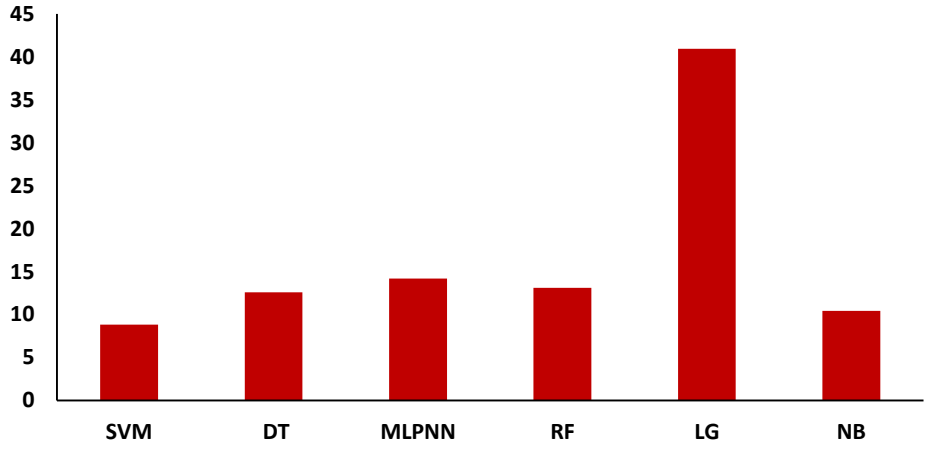


Figure 4.32: Error (%) of different algorithms for crop prediction.

The above figure 4.31 and figure 4.32 shows that the SVM algorithm has a higher accuracy rate which is 91.18%, and a lower error rate which is 8.82%.

Comparison table of prediction accuracy, precision, recall and f1-score between SVM, DT, MLPNN, RF, LG, and NB algorithms is shown in table 4.17.

Table 4.17: Comparison of performance between different algorithms.

Name of algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Support Vector Machine	91.18	91.63	91.18	91.13
Decision Tree	87.43	88.32	87.43	87.75
Multilayer Perceptron NN	85.83	80.20	85.83	82.13
Random Forest	86.90	87.66	86.90	87.05
Logistic Regression	59.09	58.20	59.09	55.68
Naïve Bayes	89.57	92.46	89.57	90.23

The following figure 4.33 shows the comparison of performance evaluation for different algorithms on the basis of precision, recall, f-measure, and accuracy.

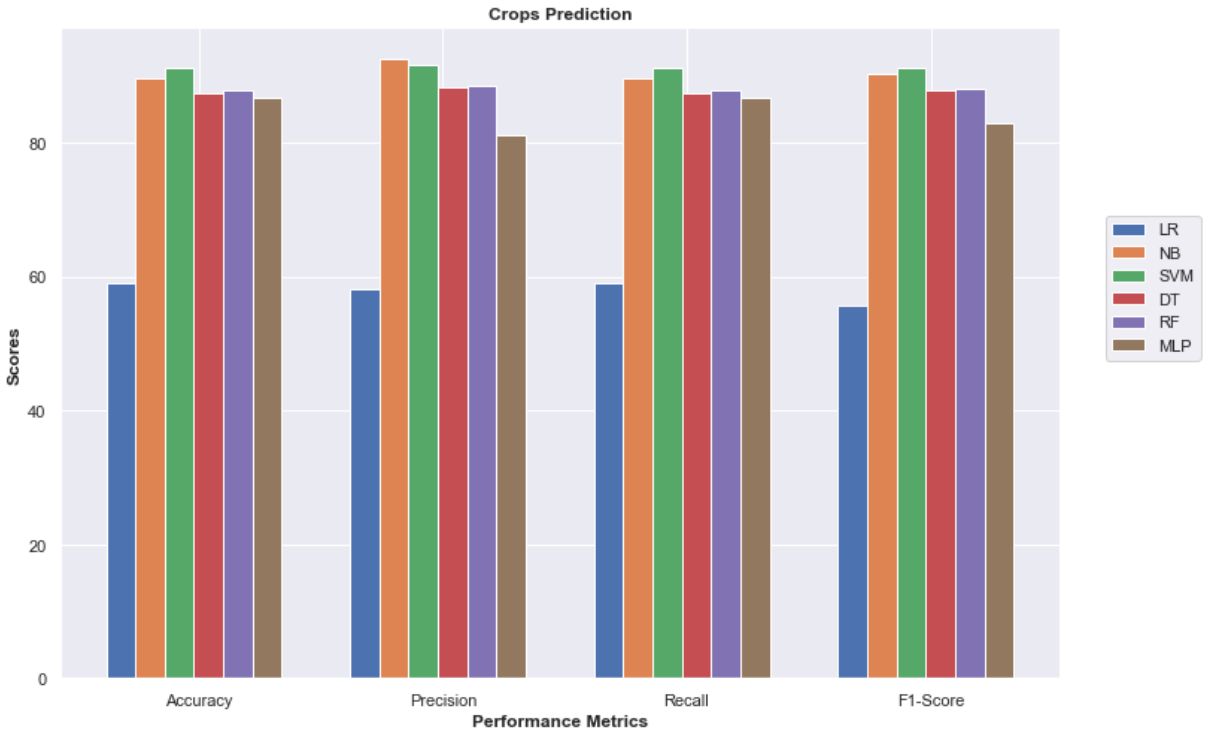


Figure 4.33: Comparison between different algorithms on the basis of performance metrics.

From the above figure 4.33, it visibly shows that the Support Vector Machine (SVM) algorithm has maximum accuracy, recall, and f1-score, which is larger than other algorithms. From the discussion of this section, we selected Support Vector Machine classifier for crops prediction. The attributes location, soil class, season, and pH is used as input, and crops are used as target.

4.3.3 Testing

Now Random Forest (RF), Support Vector Machine, Decision Tree, Multilayer perceptron, logistic regression Naïve Bayes algorithm is used for testing with new crops samples which were unused before. So, test the classifier with the new 30 instances. The new instance data set is in the main dataset. The last 30 samples of the total samples are used for prediction accuracy. These 30 samples are not used in training and testing before. All data samples are shown in the appendices section. One of the main tasks of the research is to show

the prediction with new samples. Since the main goal of the research is to find out the best algorithm which will give better accuracy than the early classification system of soil classification. The following table 4.18 shows the results of new testing samples:

Table 4.18: Target and Predicted value of new samples for different algorithms (crops)

Sample No	**Target**	**Predicted**			
		Random Forest	**SVM**	**MLPNN**	**Decision Tree**
1	Cucumber	Cucumber	Cucumber	Cucumber	Cucumber
2	Cucumber	Cucumber	Cucumber	Cucumber	Cucumber
3	Sola	Sola	Sola	Sola	Sola
4	Grass pea	Grass pea	Grass pea	Grass pea	Grass pea
5	Grass pea	Grass pea	Grass pea	Grass pea	Grass pea
6	Capsicum	Cauliflower	Cauliflower	Cauliflower	Bean
7	Cucumber	Cucumber	Cucumber	Cucumber	Cucumber
8	Cauliflower	Wheat	Cauliflower	Cauliflower	Bean
9	cauliflower	Cauliflower	Cauliflower	Cauliflower	cauliflower
10	sola	Ladies-finger	Ladies-finger	Ladies-finger	Ladies-finger
11	Cauliflower	Wheat	Cauliflower	Cauliflower	Bean
12	Gourd	Gourd	Gourd	Gourd	Gourd
13	Ladies-finger	Ladies-finger	Ladies finger	Ladies-finger	Ladies-finger
14	Ladies-finger	Ladies-finger	Ladies-finger	Ladies-finger	Ladies- finger
15	Cauliflower	Wheat	Cauliflower	Cauliflower	capsicum
16	Brinjal	Brinjal	Brinjal	Brinjal	Brinjal
17	Rice	Rice	Rice	Rice	Rice
18	Parwal	Parwal	Parwal	Parwal	Parwal
19	Rice	Rice	Rice	Rice	Rice

20	Grass pea	Grass pea	Grass pea	Grass pea	Grass pea
21	Spiny gourd	Spiny gourd	Spiny gourd	Spiny gourd	Spiny gourd
22	Ladies-finger	Sola	Ladies finger	Ladies-finger	Sola
23	Brinjal	Brinjal	Brinjal	Brinjal	Brinjal
24	Pumpkin	Pumpkin	Pumpkin	Pumpkin	Pumpkin
25	Maize	Maize	Maize	Maize	Maize
26	Bean	Bean	Parwal	Parwal	Bean
27	Pumpkin	Pumpkin	Pumpkin	Pumpkin	pumpkin
28	Maize	Maize	Maize	Maize	Maize
29	Linseed	Linseed	Linseed	Linseed	Linseed
30	Brinjal	Brinjal	Brinjal	Brinjal	Brinjal
Accuracy		80%	90%	90%	80%

Table 4.18: Continued.

Sample No	**Target**	**Predicted**	
		Naïve Bayes	**Logistic Regression**
1	Cucumber	Cucumber	Spiny gourd
2	Cucumber	Cucumber	Snake gourd
3	Sola	Sola	Sola
4	Grass pea	Grass pea	Grass pea
5	Grass pea	Grass pea	Grass pea
6	Capsicum	Capsicum	Grass pea
7	Cucumber	Cucumber	Snake gourd
8	Cauliflower	Bean	Grass pea
9	cauliflower	cauliflower	Cauliflower

10	sola	Ladies-finger	Chili
11	Cauliflower	Capsicum	Grass pea
12	Gourd	Gourd	Gourd
13	Ladies-finger	Ladies-finger	Ladies-finger
14	Ladies-finger	Ladies-finger	Ladies-finger
15	Cauliflower	Wheat	Grass pea
16	Brinjal	Brinjal	Brinjal
17	Rice	Rice	Rice
18	Parwal	Parwal	Parwal
19	Rice	Rice	Rice
20	Grass pea	Grass pea	Grass pea
21	Spiny gourd	Spiny gourd	Spiny gourd
22	Ladies-finger	Ladies-finger	Chili
23	Brinjal	Brinjal	Brinjal
24	Pumpkin	Pumpkin	Pumpkin
25	Maize	Maize	Maize
26	Bean	Parwal	Parwal
27	Pumpkin	Pumpkin	Pumpkin
28	Maize	Maize	Maize
29	Linseed	Linseed	Linseed
30	Brinjal	Brinjal	Brinjal
Accuracy		83.33%	66.67%

Here, the testing accuracy of random forest is 80%, SVM is 90%, MLPNN is 90%, Decision tree is 80%, Logistic Regression is 66.67%, and Naïve Bayes is 83.33%.

4.4 BETTER MODEL SELECTION

For soil classification, using the evaluation of experimental results of all algorithms, this work concludes that Random Forest (RF) algorithm can be considered as the best algorithm for soil classification because of its highest prediction accuracy 95.15%. As RF gives better accuracy, it can be used for Soil classification prediction.

For crops cultivation prediction, using the evaluation of experimental results of all algorithms, this work concludes that, the Support Vector Machine (SVM) algorithm can be considered as the best algorithm for crops cultivation prediction because of its highest prediction accuracy 91.18%. As SVM gives better accuracy it can be used for crops cultivation prediction.

4.5 COMPARISON WITH PREVIOUS RELATED WORK

Jagdeep Yadav, Shalu Chopra, and Vijayalakshmi M in their paper “Soil Analysis and Crop Fertility Prediction using Machine Learning” showed artificial neural networks gives 98% accuracy [2]. Ritesh Dash, Dillip Ku Dash, and G.C. Biswal in their paper “Classification of crop based on macronutrients and weather data using machine learning techniques” showed support vector machine gives 91% accuracy [14]. Utpal Barman and Ridip Dev Choudhury in their paper “Soil texture classification using multi class support vector machine” showed multi class SVM gives 91.37% accuracy [15]. In [23], they used decision tree for soil classification and showed 92.85% prediction accuracy. Pavan Patil, Virendra Panpatil and Prof. Shrikant Kokate used KNN classifier for crops prediction and showed 89.45% prediction accuracy [35]. The following table shows comparison with previous related work.

Table 4.19: Comparison with previous related works.

Research Reference	**Author**	**Methods/Techniques**	**Accuracy of previous work (%)**	**Accuracy of this work (%)**
[2]	Jagdeep Yadav, Shalu Chopra and Vijayalakshmi M	Artificial Neural Network (ANN)	Accuracy of 98%.	Accuracy of 95.15% for soil classification.
[14]	Ritesh Dash, Dillip Ku Dash and G.C. Biswal	Support Vector Machine (SVM)	F1 score is 91% for crops classification	F1 score is 91.13% for crops cultivation prediction.
[15]	Utpal Barman and Ridip Dev Choudhury	Support Vector Machine	Accuracy of 91.37% for soil classification.	Accuracy of 95.15%.
[23]	B. Bhattacharya and D.P. Solomatine	Decision tree	Accuracy of 92.85% for soil classification	Accuracy of 95.15%.
[35]	Pavan Patil, Virendra Panpatil and Prof. Shrikant Kokate	KNN classifier	Accuracy of 89.45% for crops prediction.	Accuracy of 91.18%.

[38]	Senthil Kumar Swami Durai and Mary Divya Shamili	Naïve Bayes	Accuracy of 94.72% for crops prediction.	Accuracy of 91.18%.
[40]	G. Mariammal , A. Suruliandi , S. P. Raja , and E. Poongotha	Bagging classifier	Accuracy of 95%.	Accuracy of 95.15% for soil classification.
[42]	Priyadharshini A, Swapneel Chakraborty, Aayush Kumar and Omen Rajendra.	Neural Network	Accuracy of 89.88% for crops prediction.	Accuracy of 91.18%.
[43]	Vrushali C. Waikar and Sheetal Y. Thorat	Artificial Neural Network	Accuracy of 92%.	Accuracy of 91.18% for crops prediction.

4.6 SENSITIVITY ANALYSIS

Sensitivity analysis is used to determine which attributes have the most influence on the soil classification prediction. This function calculates the sensitivity of each attribute to the class to estimate its worth. Sensitivity analysis has been performed on the seven attributes of the dataset using three different methods to determine the most significant attribute responsible for soil classification. Table 4.20 represents the result of the sensitivity analysis for the soil classification.

Table 20: Sensitivity Analysis of Soil classification

Attributes	Sensitivity (%)
pH	38.06
Salinity	52.66
Organic matter	68.41
Nitrogen	43.62
Phosphorus-B	19.24
Sulfur	40.55
Boron	73.15

The sensitivity analysis indices in this study show how important each parameter is to the prevalence of soil classification. Figure 4.34 demonstrates that Boron, Organic matter, and salinity are the most influential parameter in the sensitivity analysis method.

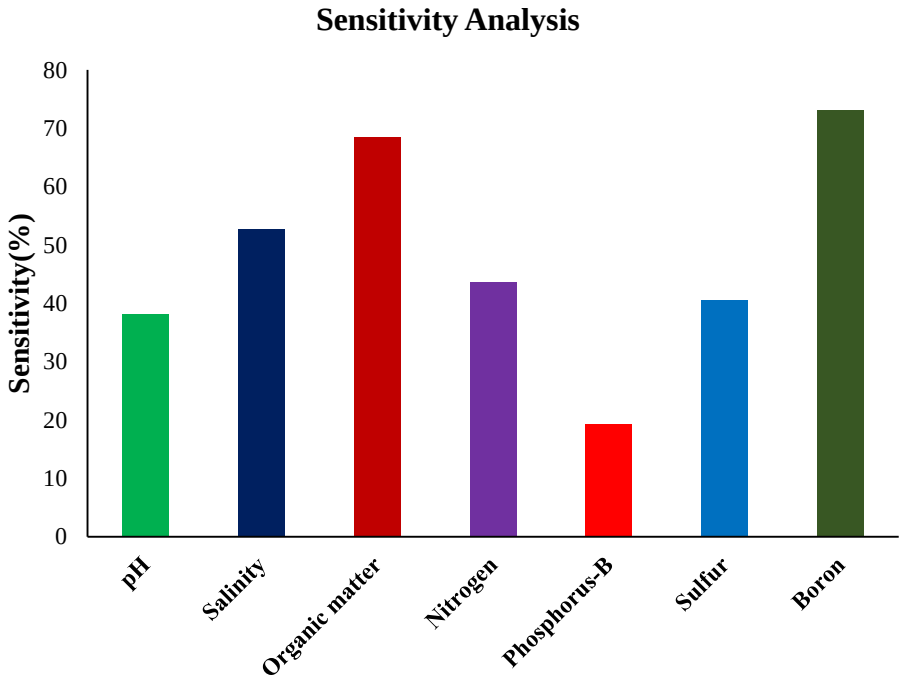


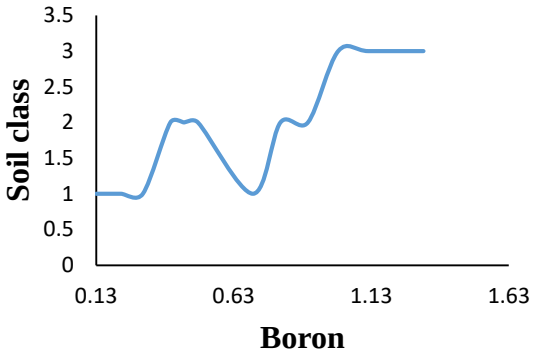
Figure 4.34: Graphical Representation of Sensitivity Analysis for soil classification.

- Boron is the most significant parameter in the sensitivity analysis of soil classification. Boron (B) is a micronutrient component of soil to the growth and health of all crops. It is a component of plant reproductive structures. Adequate boron

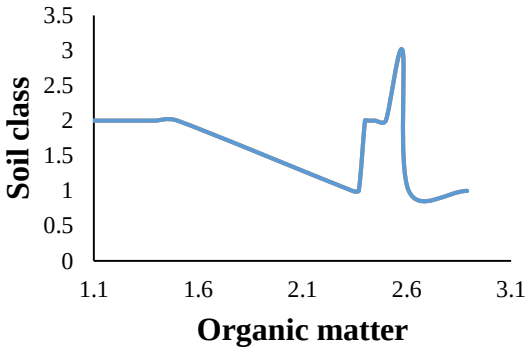
is required for effective nitrogen fixation in crops. Boron deficiency commonly results in empty pollen grains, poor pollen vitality, and a reduced number of flowers per plant. Low boron supply can stunt root growth.

- Organic matter is the second most important parameter in the sensitivity analysis of soil classification. Organic matter is a composition of any living or dead animal and plant material. It includes plant roots, leaves, and an animals dead body. The remains part of plant and animal is circulated at various stages of decomposition. The process of decomposition releases nutrients which can be taken up by plant roots as food.
- Salinity is the measurement of melted salts in the soil solution. The process of accumulating liquefiable salts in the soil is known as salinization. Salts in the soil have an important effect on the functions and management.
- Nitrogen is a naturally occurring element that is necessary for growth and reproduction in both plants and animals. The conversion of organic nitrogen in soil into mineral nitrogen is a significant source of nitrogen. Soil nitrogen is useful to estimate how much nitrogen from organic matter will become available to a crop.
- Sulfur is an essential plant nutrient. It's required for the production of amino acids, which make up the proteins critical to plant growth. Sulfur deficiency is more common in plants grown on cold and sandy soils as well as those that are low in organic matter. Sulfur deficiency is also more likely to occur in areas with high rainfall or pollution.
- pH is an indication of the acidity or alkalinity of the soil. Soil pH is defined as the negative logarithm of the hydrogen ion concentration. The pH scale goes from 0 to 14, with pH 7 as the neutral point. As the amount of hydrogen ions in the soil increases the soil pH decreases thus becoming more acidic. From pH 7 to 0 the soil is increasingly more acidic and from pH 7 to 14 the soil is increasingly more alkaline.
- Phosphorus-B is one of the major plant nutrients in the soil. It is a constituent of plant cells, essential for cell division and development of the growing tip of the plant. For this reason, it is vital for seedlings and young plants.

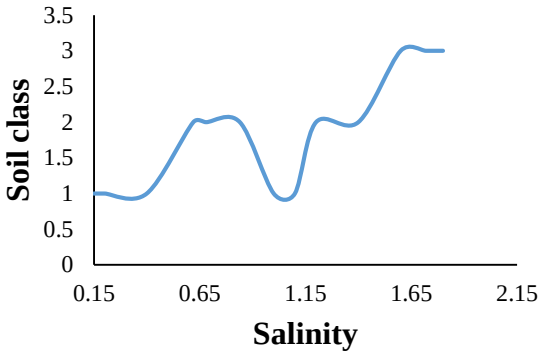
The following figure 4.35 shows the effect of distinct variables on Soil Classification:



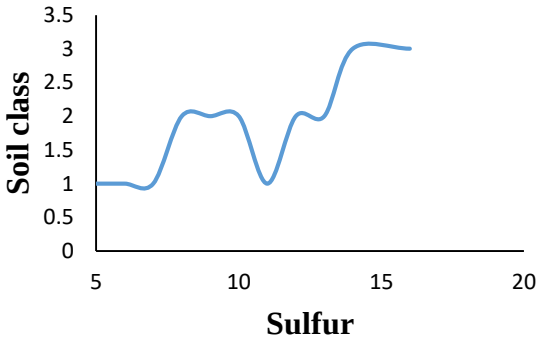
a) Boron Vs. Soil class



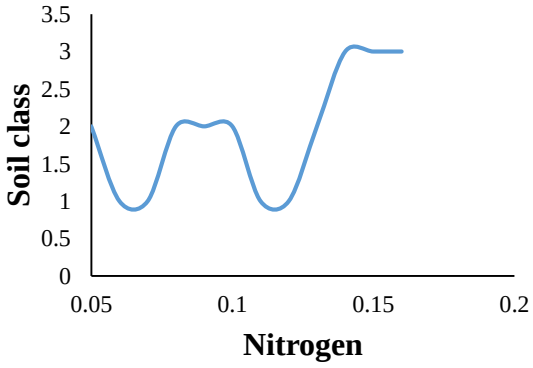
b) Organic matter Vs. Soil class



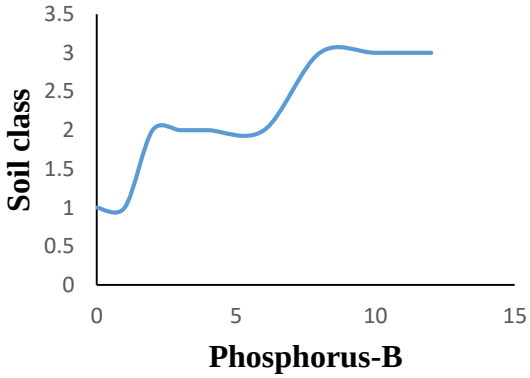
c) Salinity Vs. Soil class



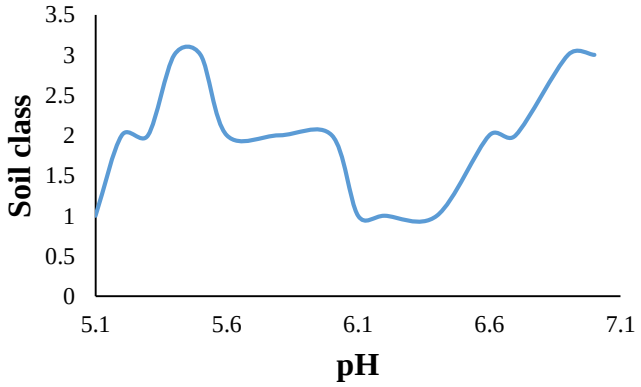
d) Sulfur Vs. Soil class



e) Nitrogen Vs. Soil class



f) Phosphorus-B Vs. Soil class



g) pH Vs. Soil class

Figure 4.35: Effect of distinct variables on Soil Classification (a) Boron (b) Organic matter (c) Salinity (d) Sulfur (e) Nitrogen (f) Phosphorus-B and (g) pH

The variation of the target (Soil Class) in relation to more useful input parameters is depicted in Figure 4.35. It is apparent from Figure 4.35 that the value of Boron, Organic matter, and salinity is more correlated with soil classification prediction. Compared to the findings of other studies conducted in this field, these results appear reasonable.

4.7 SUMMARY

In this chapter, the results of soil classification and crops cultivation prediction are discussed. All algorithms are discussed in detail and select the final model compared with other algorithms which is considered for this work. Random forest algorithm is selected for soil classification on the basis of training and testing results. And support Vector Machine is selected for crops cultivation prediction.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 INTRODUCTION

This chapter describes the conclusion and future research scope of this thesis work. First, the overall summary of this thesis work is presented in the conclusion section. Then the potential future research directions are provided in the future research scope section.

5.2 CONCLUSION

This research has been done to classify soil and predict crop cultivation based on soil type. This work focuses on the use of different artificial intelligence techniques to enhance crops cultivation. Soil is an important ingredient of agriculture. Here, used the real-time dataset of soil and crops which are collected from different zones of the Noakhali district. Soil chemical features are used as input, and soil class is used as target. In the crops dataset, crops are target based on soil type, location, and season. In this work, all experiment is based on the real dataset, and python programming is used for evaluation. The algorithm used for prediction is Support Vector Machine, Decision Tree, Multilayer Perceptron, Random forest, Logistic Regression, and Naive Bayes. Each algorithm's outcome is used to determine the better classifiers among them. Comparisons of these selected algorithms are measured based on the performance factors.

For soil classification, the Support Vector Machine (SVM) algorithm gives 85.90% accuracy, 86.37% precision, 85.90% recall, 82.57% f1-score, 14.1% error rate, and ROC 92.65. Decision Tree classifier gives 93.83% accuracy, 94.09% precision, 93.83% recall, 93.93% f1-score, 6.17% error rate, and ROC 93.65. Multilayer Perceptron Neural Network gives 87.44% accuracy, 87.05% precision, 87.44% recall, 87.10% f1-score, 12.56% error rate, and

ROC 95.76. Random forest classifier gives 95.15% accuracy, 95.16% precision, 95.15% recall, 95.16% f1-score, 4.85% error rate, and ROC 99.24. Logistic Regression gives 83.04% accuracy, 80.55% precision, 83.04% recall, 80.29% f1-score, 16.96% error rate, and ROC 91.63. Naïve Bayes algorithm gives 76.65% accuracy, 75.88% precision, 76.65% recall, 76.09% f1-score, 23.35% error rate, and ROC 91.12. And testing with the new unused samples, SVM gives 93.33% accuracy, Decision tree is 96.67%, Multilayer perceptron is 83.33%, Random forest is 96.67%, Logistic regression is 90% and Naïve Bayes is 76.66%.

By evaluating the experimental results of all algorithms, this work concludes that Random Forest (RF) algorithm can be considered the best algorithm because of its highest prediction accuracy 96.67%. As RF gives better accuracy, it can be used for Soil classification prediction.

For crop cultivation prediction, the Support Vector Machine (SVM) algorithm gives 91.18% accuracy, 91.63% precision, 91.18% recall, 91.13% f1-score, 8.82% error rate, and ROC 99.53. Decision Tree classifier gives 87.43% accuracy, 88.32% precision, 87.43% recall, 87.75% f1-score, 12.57% error rate, and ROC 88.79. Multilayer Perceptron Neural Network gives 85.83% accuracy, 80.20% precision, 85.83% recall, 82.13% f1-score, 14.17% error rate, and ROC 95.13. Random forest classifier gives 86.90% accuracy, 87.66% precision, 86.90% recall, 87.05% f1-score, 13.1% error rate, and ROC 96.6. Logistic Regression gives 59.09% accuracy, 58.20% precision, 59.09% recall, 55.68% f1-score, 40.91% error rate, and ROC 97.66. Naïve Bayes algorithm gives 89.57% accuracy, 92.46% precision, 89.57% recall, 90.23% f1-score, 10.43% error rate, and ROC 99.42. And testing with new unused samples, SVM gives 90% accuracy, Decision tree is 80%, Multilayer perceptron is 90%, Random forest is 80%, Logistic regression is 66.67%, and Naïve Bayes is 83.33%.

By evaluating the experimental results of all algorithms, this work concludes that Support Vector Machine (SVM) algorithm can be considered the best algorithm because of its highest prediction accuracy 90%. As SVM gives better accuracy, it can be used for crop cultivation prediction.

5.3 FUTURE RESEARCH SCOPE

There is a possibility to generate a new algorithm or a hybrid one that will provide better accuracy than the existing algorithms. Data from several districts of Bangladesh can be used for wide research on this topic in the future. Weather variables like temperature, rainfall, humidity, wind speed, geographic region, etc., can be used with the crops dataset as input parameters. Having a plan to develop software for the farmers so that they can easily cultivate crops based on their soil type.

REFERENCES

1. Rajeswari, A. M., A. Selva Anushiya, K. Seyad Ali Fathima, S. Shanmuga Priya, and N. Mathumithaa. "Fuzzy Decision Support System for Recommendation of Crop Cultivation based on Soil Type." In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pp. 768-773. IEEE, 2020.
2. Yadav, Jagdeep, Shalu Chopra, and M. Vijayalakshmi. "Soil analysis and crop fertility prediction using machine learning." *Machine Learning* 8, no. 03 (2021).
3. Harlianto, Pramudyana Agus, Teguh Bharata Adji, and Noor Akhmad Setiawan. "Comparison of machine learning algorithms for soil type classification." In *2017 3rd International Conference on Science and Technology-Computer (ICST)*, pp. 7-10. IEEE, 2017.
4. Wu, Wei, Ai-Di Li, Xin-Hua He, Ran Ma, Hong-Bin Liu, and Jia-Ke Lv. "A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China." *Computers and Electronics in Agriculture* 144 (2018): 86-93.
5. Bhargavi, P., and S. Jyothi. "Soil classification using data mining techniques: a comparative study." *International Journal of Engineering Trends and Technology* 2, no. 1 (2011): 55-59.
6. Radhika, K., and D. Madhavi Latha. "Machine learning model for automation of soil texture classification." *Indian Journal of Agricultural Research* 53, no. 1 (2019).
7. Cai, Simin, Rongqun Zhang, Liming Liu, and De Zhou. "A method of salt-affected soil information extraction based on a support vector machine with texture features." *Mathematical and Computer Modelling* 51, no. 11-12 (2010): 1319-1325.
8. Gao, Han, Changcheng Wang, Guanya Wang, Haiqiang Fu , Jianjun Zhu. "A novel crop classification method based on ppfSVM classifier with time-series alignment kernel from dual-polarization SAR datasets." *Remote Sensing of Environment* 264 (2021): 112628.
9. Inazumi, Shinya, Sutasinee Intui, Apiniti Jotisankasa, Susit Chaiprakaikeow, and Kazuhiko Kojima. "Artificial intelligence system for supporting soil classification." *Results in Engineering* 8 (2020): 100188.

10. Agarwal, Rahul, Narpat Singh Shekhawat, and Ashish Kumar Luhach. "Automated classification of soil images using chaotic Henry's gas solubility optimization: Smart agricultural system." *Microprocessors and Microsystems* (2021): 103854.
11. Sirsat, Manisha Sanjay, M. Fernández-Delgado. "Classification of agricultural soil parameters in India." *Computers and electronics in agriculture* 135 (2017): 269-279.
12. Kodikara, Gayantha RL, and Lindsay J. McHenry. "Machine learning approaches for classifying lunar soils." *Icarus* 345 (2020): 113719.
13. Javed Mallick, Swapan Talukdar, Shahfahad, Swades Pal, Atiqur Rahman. "A novel classifier for improving wetland mapping by integrating image fusion techniques and ensemble machine learning classifiers." *Ecological Informatics* 65 (2021): 101426.
14. Dash, Ritesh, Dillip Ku Dash, and G. C. Biswal. "Classification of crop based on macronutrients and weather data using machine learning techniques." *Results in Engineering* 9 (2021): 100203.
15. Barman, Utpal, and Ridip Dev Choudhury. "Soil texture classification using multi class support vector machine." *Information processing in agriculture* 7.2 (2020): 318-332.
16. Zhong, Liang, Xi Guo, Zhe Xu, Meng Ding. "Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks." *Geoderma* 402 (2021): 115366.
17. Ghadge, Rushika, Juilee Kulkarni, Pooja More, Sachee Nene, and R. L. Priya. "Prediction of crop yield using machine learning." *Int. Res. J. Eng. Technol.(IRJET)* 5 (2018).
18. Angelaki, Anastasia, Somvir Singh Nain, Varun Singh, and Parveen Sihag. "Estimation of models for cumulative infiltration of soil using machine learning methods." *ISH Journal of Hydraulic Engineering* 27, no. 2 (2021): 162-169.
19. Srunitha, K., and S. Padmavathi. "Performance of SVM classifier for image based soil classification." *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, 2016.
20. Palanivel, Kodimalar, and Chellammal Surianarayanan. "An approach for prediction of crop yield using machine learning and big data techniques." *International Journal of Computer Engineering and Technology* 10.3 (2019): 110-118.

21. Farooqui, Safia. "A Study on the Status of Women Empowerment in Modern India." *Celebrating International Women's Day: Achieving an Equal Future in a COVID-19 World* 59 (2021): 15.
22. Kumar, Rakesh, M. P. Singh, Prabhat Kumar, and J. P. Singh. "Crop Selection Method to maximize crop yield rate using machine learning technique." In *2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, pp. 138-145. IEEE, 2015.
23. Bhattacharya, Biswanath, and Dimitri P. Solomatine. "Machine learning in soil classification." *Neural networks* 19.2 (2006): 186-195.
24. Aydemir, S., S. Keskin, and L. R. Drees. "Quantification of soil features using digital image processing (DIP) techniques." *Geoderma* 119.1-2 (2004): 1-8.
25. Almendra-Martín, Laura, José Martínez-Fernández, María Piles, and Ángel González-Zamora. "Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe." *Remote Sensing of Environment* 258 (2021): 112377.
26. Ansari, Nadia, Sharmin Sultana Ratri, Afroz Jahan, Muhammad Ashik-E-Rabbani, and Anisur Rahman. "Inspection of paddy seed varietal purity using machine vision and multivariate analysis." *Journal of Agriculture and Food Research* 3 (2021): 100109.
27. Jia, Xiyue, Yining Cao, David O'Connor, Jin Zhu, Daniel CW Tsang, Bin Zou, and Deyi Hou. "Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field." *Environmental Pollution* 270 (2021): 116281.
28. Gorthi, Srikanth. "Soil organic matter prediction using smartphone-captured digital images: Use of reflectance image and image perturbation." *Biosystems Engineering* 209 (2021): 154-169.
29. Rahman, Ashikur, Hasan Muhammad Abdullah, Md Tousif Tanzir, Md Jakir Hossain, Bhoktear M. Khan, Md Giashuddin Miah, and Imranul Islam. "Performance of different machine learning algorithms on satellite image classification in rural and urban setup." *Remote Sensing Applications: Society and Environment* 20 (2020): 100410.
30. Zhang, Huanxue, Yuji Wang, Jiali Shang, Mingxu Liu, and Qiangzi Li. "Investigating the impact of classification features and classifiers on crop mapping performance in

heterogeneous agricultural landscapes." *International Journal of Applied Earth Observation and Geoinformation* 102 (2021): 102388.

31. Fang, Zhice, Yi Wang, Ling Peng, and Haoyuan Hong. "Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping." *Computers & Geosciences* 139 (2020): 104470.
32. Bayad, Mohamed, Henry Wai Chau, Stephen Trolove, Karin Müller, Leo Condron, Jim Moir, and Li Yi. "Time series of remote sensing and water deficit to predict the occurrence of soil water repellency in New Zealand pastures." *ISPRS Journal of Photogrammetry and Remote Sensing* 169 (2020): 292-300.
33. Abimala, T., S. Flora Sashya, and K. Sripriya. "Soil Classification & Crop Suggestion based on HSV, GLCM, Gabor Wavelet Techniques and Decision Tree Classifier in Image Processing." (2020).
34. Priya, P. Kanaga, and N. Yuvaraj. "An IoT based gradient descent approach for precision crop suggestion using MLP." *Journal of Physics: Conference Series*. Vol. 1362. No. 1. IOP Publishing, 2019.
35. Patil, Pavan, Virendra Panpatil, and Shrikant Kokate. "Crop prediction system using machine learning algorithms." *Int. Res. J. Eng. Technol.(IRJET)* 7, no. 02 (2020).
36. Rale, Neha, Raxitkumar Solanki, Doina Bein, James Andro-Vasko, and Wolfgang Bein. "Prediction of crop cultivation." In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0227-0232. IEEE, 2019.
37. Chakrabarty, Amitabha, Nafees Mansoor, Muhammad Irfan Uddin, Mosleh Hmoud Al-Adaileh, Nizar Alsharif, and Fawaz Waselallah Alsaade. "Prediction approaches for smart cultivation: a comparative study." *Complexity* 2021 (2021).
38. Durai, Senthil Kumar Swami, and Mary Divya Shamili. "Smart farming using Machine Learning and Deep Learning techniques." *Decision Analytics Journal* 3 (2022): 100041.
39. Kiran, M. P., and N. R. Deepak. "Crop prediction based on influencing parameters for different states in india-the data mining approach." In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1785-1791. IEEE, 2021.
40. Mariammal, G., A. Suruliandi, S. P. Raja, and E. Poongothai. "Prediction of land suitability for crop cultivation based on soil and environmental characteristics using

modified recursive feature elimination technique with various classifiers." *IEEE Transactions on Computational Social Systems* 8, no. 5 (2021): 1132-1142.

41. Suruliandi, A., G. Mariammal, and S. P. Raja. "Crop prediction based on soil and environmental characteristics using feature selection techniques." *Mathematical and Computer Modelling of Dynamical Systems* 27, no. 1 (2021): 117-140.
42. Priyadharshini, A., Swapneel Chakraborty, Aayush Kumar, and Omen Rajendra Pooniwala. "Intelligent Crop Recommendation System using Machine Learning." In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 843-848. IEEE, 2021.
43. Waikar, Vrushali C., Sheetal Y. Thorat, Ashlesha A. Ghute, Priya P. Rajput, and Mahesh S. Shinde. "Crop prediction based on soil classification using machine learning with classifier ensembling." *Int. Res. J. Eng. Technol* 7, no. 5 (2020).
44. Kumar, Rakesh, M. P. Singh, Prabhat Kumar, and J. P. Singh. "Crop Selection Method to maximize crop yield rate using machine learning technique." In *2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, pp. 138-145. IEEE, 2015.
45. Gupta, Archana, Dharmil Nagda, Pratiksha Nikhare, and Atharva Sandbhor. "Smart crop prediction using IoT and machine learning." *International Journal of Engineering Research & Technology (IJERT)* (2021): 18-21.
46. Aditya Shastry, K., and H. A. Sanjay. "Cloud-Based Agricultural Framework for Soil Classification and Crop Yield Prediction as a Service." In *Emerging Research in Computing, Information, Communication and Applications*, pp. 685-696. Springer, Singapore, 2019.
47. Vaishnavi, S., M. Shobana, R. Sabitha, and S. Karthik. "Agricultural crop recommendations based on productivity and season." In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 883-886. IEEE, 2021.
48. Shastry, K. Aditya, H. A. Sanjay, and H. Kavya. "A novel data mining approach for soil classification." In *2014 9th International Conference on Computer Science & Education*, pp. 93-98. IEEE, 2014.

49. Varsha, A., V. M. Midhuna, and R. Divya. "Soil classification and crop recommendation using IoT and machine learning." *International Journal of Scientific Research & Engineering Trends* 6, no. 3 (2020).
50. Baskar, S. S., L. Arockiam, and S. Charles. "Applying data mining techniques on soil fertility prediction." *International Journal of Computer Applications Technology and Research* 2, no. 6 (2013): 660-662.
51. Rajeswari, S., and K. Suthendran. "C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud." *Computers and Electronics in Agriculture* 156 (2019): 530-539.
52. Reshma, R., V. Sathiyavathi, T. Sindhu, K. Selvakumar, and L. SaiRamesh. "IoT based classification techniques for soil content analysis and crop yield prediction." In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 156-160. IEEE, 2020.
53. Breve, Fabricio A., Moacir P. Ponti-Junior, and Nelson DA Mascarenhas. "Multilayer perceptron classifier combination for identification of materials on noisy soil science multispectral images." In *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pp. 239-244. IEEE, 2007.
54. Haykin, Simon. *Redes neurais: princípios e prática*. Bookman Editora, 2001.
55. Liu, Yong, Huifeng Wang, Hong Zhang, and Karsten Liber. "A comprehensive support vector machine-based classification model for soil quality assessment." *Soil and Tillage Research* 155 (2016): 19-26.
56. Bhavsar, Himani, and Mahesh H. Panchal. "A review on support vector machine for data classification." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1, no. 10 (2012): 185-189.
57. Cristianini, Nello, and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
58. Lin, Weiwei, Ziming Wu, Longxin Lin, Angzhan Wen, and Jin Li. "An ensemble random forest algorithm for insurance big data analysis." *Ieee access* 5 (2017): 16568-16575.
59. Liu, Yanli, Yourong Wang, and Jian Zhang. "New machine learning algorithm: Random forest." In *International Conference on Information Computing and Applications*, pp. 246-252. Springer, Berlin, Heidelberg, 2012.

60. Ren, Qiong, Hui Cheng, and Hai Han. "Research on machine learning framework based on random forest algorithm." In *AIP conference proceedings*, vol. 1820, no. 1, p. 080020. AIP Publishing LLC, 2017.
61. Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 20-28.
62. Kleinbaum, David G., and Mitchel Klein. "Introduction to logistic regression." In *Logistic regression*, pp. 1-39. Springer, New York, NY, 2010.
63. Nick, Todd G., and Kathleen M. Campbell. "Logistic regression." *Topics in biostatistics* (2007): 273-301.
64. Yang, Feng-Jen. "An implementation of naive bayes classifier." In *2018 International conference on computational science and computational intelligence (CSCI)*, pp. 301-306. IEEE, 2018.
65. Rakhra, Manik, Priyansh Soniya, Dishant Tanwar, Piyush Singh, Dorothy Bordoloi, Prerit Agarwal, Sakshi Takkar, Kapil Jairath, and Neha Verma. "Crop price prediction using random forest and decision tree regression:-a review." *Materials Today: Proceedings* (2021).

APPENDICES

A. SOIL DATASET

A-1 Soil Raw Dataset

Sample No	pH	Salinity	Nitrogen	Phosphorus-B	Sulfur	Boron	Organic Matter	Soil class	Location
1	6.5	0.56	0.131	15	7	0.74	2.61	clay loam	purbo char jabbar
2	6.5	0.44	0.113	15	12	0.87	2.87	clay loam	purbo char jabbar
3	6.8	1.88	0.11	25	31	1.43	2.51	clay loam	purbo char jabbar
4	7	1.45	0.103	12	60	1.77	2.06	loam	purbo char jabbar
5	6.9	2.75	0.12	12	38	1.68	2.41	loam	purbo char jabbar
6	7.2	1.84	0.155	12	67	1.64	3.1	clay loam	purbo char jabbar
7	7.4	0.66	0.093	20	28	0.8	2.48	clay loam	purbo char jabbar
8	6.8	0.54	0.11	24	16	0.75	2.55	clay loam	purbo char jabbar
9	6.6	0.75	0.072	19	6	0.58	3.44	CLAY	purbo char jabbar
10	6.7	0.78	0.117	18	12	0.94	3.11	CLAY	pashim char jabbar
11	6.6	0.56	0.072	17	24	0.52	1.44	loam	pashim char jabbar
12	6.5	0.69	0.117	33	13	0.86	3.34	CLAY	pashim char jabbar
13	6.5	0.49	0.11	27	18	0.67	2.8	clay loam	pashim char jabbar
14	5.8	3.95	0.089	15	73	0.85	2.79	clay loam	pashim char jabbar

15	6.3	2.27	0.14	13	76	1.06	2.4	loam	pashim char jabbar
16	6.7	0.78	0.11	25	27	0.84	2.87	clay loam	char jabbar
17	6.6	0.91	0.13	35	13	0.79	1.61	loam	char jabbar
18	6	0.51	0.13	29	9	0.62	2.61	clay loam	char jabbar
19	6.2	1.88	0.14	36	57	1.02	2.75	clay loam	char jabbar
20	6.2	2.36	0.18	21	78	1.06	3.58	CLAY	char jabbar
21	6.2	2.4	0.13	23	100	0.94	2.61	clay loam	char jabbar
22	6.4	1.44	0.1	18	59	0.92	2.99	clay loam	char jabbar
23	6.6	1.19	0.09	26	49	0.75	2.79	CLAY	char jabbar
24	6.3	5.53	0.12	45	71	1.52	2.51	clay loam	char jabbar
25	6.6	2.21	0.15	29	31	1.36	2.96	clay loam	char panaullah
26	6.4	1.33	0.18	21	18	1.35	2.48	loam	char panaullah
27	6.7	0.68	0.09	18	39	0.93	1.72	loam	char panaullah
28	6.5	1.53	0.1	13	14	1.04	3.5	clay loam	char panaullah
29	6.6	1.05	0.12	86	32	0.81	2.34	loam	char panaullah
30	6.7	1.56	0.12	29	34	1	2.54	clay loam	char panaullah
31	6.5	0.76	0.12	13	36	0.93	2.6	clay loam	char panaullah
32	6.5	0.53	0.12	19	14	0.84	2.6	CLAY	char panaullah
33	6.6	0.63	0.12	15	26	0.8	3.34	CLAY	char panaullah
34	6.3	0.5	0.09	22	18	0.68	2.79	clay loam	uttar char vagga
35	6.2	0.53	0.09	22	10	0.74	1.86	loam	uttar char vagga
36	5.8	0.45	0.07	31	17	0.54	1.44	loam	uttar char vagga

37	5.9	0.41	0.12	13	18	0.91	3.41	CLAY	uttar char vagga
38	6.1	0.81	0.13	12	32	1.08	2.55	CLAY	uttar char vagga
39	6.1	1.52	0.14	11	29	1.05	2.4	loam	uttar char vagga
40	6.3	1.39	0.13	17	51	1.56	2.61	clay loam	uttar char vagga
41	6.4	0.91	0.13	14	29	1.23	2.61	CLAY	uttar char vagga
42	6.3	1.37	0.17	17	47	0.91	3.37	clay loam	uttar char vagga
43	6.4	1.53	0.09	23	36	0.99	2.78	clay loam	madhya char vagga
44	6.4	0.45	0.08	12	13	0.53	1.65	loam	madhya char vagga
45	6.5	0.63	0.12	11	24	1.12	3.41	CLAY	madhya char vagga
46	6.1	0.98	0.16	13	60	1.01	2.1	loam	madhya char vagga
47	6.1	1.73	0.19	15	106	1.95	3.85	clay loam	madhya char vagga
48	6.4	1.01	0.2	12	82	1.34	1.99	loam	madhya char vagga
49	6.2	1.57	0.16	7	103	1	3.23	clay loam	madhya char vagga
50	6.1	4.03	0.17	30	131	1.56	3.44	CLAY	dakshin char vagga
51	6	5.63	0.17	19	135	1.76	3.44	clay loam	dakshin char vagga
52	6.3	4.44	0.17	20	127	1.77	3.37	clay loam	dakshin char vagga
53	6.4	4.05	0.14	16	106	2.03	2.89	clay loam	dakshin char vagga

54	6.3	6.17	0.16	22	121	2.03	3.16	CLAY	dakshin char vagga
55	6.4	6.25	0.13	20	117	1.02	2.68	CLAY	dakshin char vagga
56	6.3	10.26	0.12	13	120	0.86	3.41	CLAY	dakshin char vagga
57	6.4	8.48	0.11	10	143	0.7	2.77	clay loam	dakshin char vagga
58	6.6	5.91	0.13	9	119	0.85	2.61	clay loam	dakshin char vagga
59	7.1	0.81	0.12	9	29	0.56	3.4	CLAY	char vagga
60	7	1.63	0.1	11	78	0.72	3.4	CLAY	char vagga
61	6.9	4.65	0.14	12	118	0.67	2.75	CLAY	char vagga
62	6.9	8.56	0.1	16	48	0.82	3.06	clay loam	char zia uddin
63	7.3	5.54	0.08	15	75	0.76	3.58	CLAY	char zia uddin
64	7.8	1.59	0.13	14	38	0.67	2.55	CLAY	char zia uddin
65	7.8	4.43	0.21	17	78	0.63	4.13	clay loam	char zia uddin
66	7.8	1.15	0.2	12	14	0.48	3.99	CLAY	char zia uddin
67	7.4	2.3	0.15	17	87	0.73	2.96	clay loam	char zia uddin
68	7.3	4.16	0.21	12	43	0.87	4.26	CLAY	char zia uddin
69	7.4	2.49	0.15	11	100	0.81	3.03	clay loam	char zia uddin
70	7.5	1.27	0.15	13	8	1.09	3.03	CLAY	char mohiuddin
71	7.4	1.64	0.18	17	12	1.15	3.58	CLAY	char mohiuddin
72	7.7	0.75	0.12	11	18	2.1	2.34	clay loam	char mohiuddin
73	7.5	1.05	0.12	13	35	2	2.8	CLAY	char mohiuddin

74	7.1	4.43	0.2	12	69	1.8	3.99	clay loam	char mohiuddin
75	7.1	4.98	0.18	13	92	1.55	3.58	CLAY	char mohiuddin
76	7.3	2.1	0.21	12	21	1.32	4.2	CLAY	char mohiuddin
77	7.3	1.26	0.16	9	93	0.98	3.1	CLAY	char mohiuddin
78	7.5	0.96	0.08	12	66	0.81	2.58	clay loam	char mohiuddin
79	7.4	0.5	0.19	12	24	0.78	3.78	CLAY	dakshin kachapia
80	7.2	0.88	0.12	110	52	0.88	3.41	CLAY	dakshin kachapia
81	7.4	0.51	0.15	19	25	1.05	3.03	clay loam	dakshin kachapia
82	7.1	0.58	0.17	13	28	1.25	3.3	clay loam	dakshin kachapia
83	6.6	1.39	0.18	14	96	1.41	3.58	CLAY	dakshin kachapia
84	7.4	0.56	0.09	14	29	1.5	3.72	CLAY	dakshin kachapia
85	6.3	1.05	0.17	24	101	1.22	3.44	CLAY	dakshin kachapia
86	5.3	1	0.17	12	98	1.75	3.44	CLAY	dakshin kachapia
87	5.3	2.34	0.17	12	87	2.02	3.3	clay loam	dakshin kachapia
88	6.1	1.74	0.14	46	134	2.1	2.75	CLAY	uttar kachapia
89	6.7	0.57	0.12	34	28	1.95	3.48	CLAY	uttar kachapia
90	6.7	0.4	0.15	17	26	1.82	3.03	CLAY	uttar kachapia
91	6.4	1.38	0.17	74	62	1.67	3.44	CLAY	uttar kachapia
92	6.7	0.76	0.12	22	18	0.71	2.85	clay loam	uttar kachapia
93	6	0.6	0.16	21	16	0.68	3.16	CLAY	uttar kachapia
94	5.9	0.72	0.13	15	12	0.57	2.61	clay loam	uttar kachapia
95	5.4	2.87	0.09	26	58	0.88	2.79	clay loam	uttar kachapia
96	6	0.96	0.1	105	20	0.85	2.93	CLAY	uttar kachapia

97	5.8	4.59	0.08	24	78	0.71	2.65	CLAY	pashim char jubille
98	6.1	2.56	0.11	16	92	0.47	2.85	clay loam	pashim char jubille
99	5.9	11.32	0.14	14	98	0.56	2.75	clay loam	pashim char jubille
100	6	9.21	0.13	23	80	0.46	2.61	Clay loam	pashim char jubille
….	….	….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….	…..	….
1529	5.3	0.36	0.02	6	31	0.42	2.51	clay loam	Noakhali
1530	6.3	0.51	0.09	3	20	0.32	2.52	clay loam	Noakhali
1531	6.5	0.41	0.02	2	22	0.41	2.53	loam	Noakhali
1532	6.1	0.2	0.07	3	36	0.43	2.54	clay loam	Noakhali
1534	5.4	0.45	0.12	2	30	0.36	3.4	clay loam	Noakhali
1534	5.2	0.94	0.02	3	50	0.87	2.61	clay loam	Noakhali
1535	6	0.16	0.02	2	34	0.89	2.63	clay loam	Noakhali
1536	6.1	2	0.09	2	43	0.3	2.53	clay loam	Noakhali
1537	5.7	1.6	0.06	3	45	0.31	1.02	loam	Noakhali
1538	6	0.48	0.01	3	10	0.32	0.17	loam	Noakhali
1539	5.3	0.43	0.01	4	11	0.22	2.12	loam	Noakhali
1540	5.3	0.49	0.02	2	9	0.14	2.33	loam	Noakhali
1541	6.2	0.48	0.02	3	32	0.57	2.34	loam	Noakhali
1542	5.8	0.39	0.02	3	14	0.34	2.4	loam	Noakhali

A-2 Processed Soil Data

Sample No	pH	Salinity	Nitrogen	Phosphorus-B	Sulfur	Boron	Organic Matter	Soil class
1	6.5	0.56	0.131	15	7	0.74	2.61	1
2	6.5	0.44	0.113	15	12	0.87	2.87	1

3	6.8	1.88	0.11	25	31	1.43	2.51	1
4	7	1.45	0.103	12	60	1.77	2.06	2
5	6.9	2.75	0.12	12	38	1.68	2.41	2
6	7.2	1.84	0.155	12	67	1.64	3.1	1
7	7.4	0.66	0.093	20	28	0.8	2.48	1
8	6.8	0.54	0.11	24	16	0.75	2.55	1
9	6.6	0.75	0.072	19	6	0.58	3.44	0
10	6.7	0.78	0.117	18	12	0.94	3.11	0
11	6.6	0.56	0.072	17	24	0.52	1.44	2
12	6.5	0.69	0.117	33	13	0.86	3.34	0
13	6.5	0.49	0.11	27	18	0.67	2.8	1
14	5.8	3.95	0.089	15	73	0.85	2.79	1
15	6.3	2.27	0.14	13	76	1.06	2.4	2
16	6.7	0.78	0.11	25	27	0.84	2.87	1
17	6.6	0.91	0.13	35	13	0.79	1.61	2
18	6	0.51	0.13	29	9	0.62	2.61	1
19	6.2	1.88	0.14	36	57	1.02	2.75	1
20	6.2	2.36	0.18	21	78	1.06	3.58	0
21	6.2	2.4	0.13	23	100	0.94	2.61	1
22	6.4	1.44	0.1	18	59	0.92	2.99	1
23	6.6	1.19	0.09	26	49	0.75	2.79	0
24	6.3	5.53	0.12	45	71	1.52	2.51	1
25	6.6	2.21	0.15	29	31	1.36	2.96	1
26	6.4	1.33	0.18	21	18	1.35	2.48	2
27	6.7	0.68	0.09	18	39	0.93	1.72	2
28	6.5	1.53	0.1	13	14	1.04	3.5	1
29	6.6	1.05	0.12	86	32	0.81	2.34	2
30	6.7	1.56	0.12	29	34	1	2.54	1
31	6.5	0.76	0.12	13	36	0.93	2.6	1
32	6.5	0.53	0.12	19	14	0.84	2.6	0
33	6.6	0.63	0.12	15	26	0.8	3.34	0
34	6.3	0.5	0.09	22	18	0.68	2.79	1
35	6.2	0.53	0.09	22	10	0.74	1.86	2
36	5.8	0.45	0.07	31	17	0.54	1.44	2
37	5.9	0.41	0.12	13	18	0.91	3.41	0
38	6.1	0.81	0.13	12	32	1.08	2.55	0
39	6.1	1.52	0.14	11	29	1.05	2.4	2
40	6.3	1.39	0.13	17	51	1.56	2.61	1
41	6.4	0.91	0.13	14	29	1.23	2.61	0
42	6.3	1.37	0.17	17	47	0.91	3.37	1
43	6.4	1.53	0.09	23	36	0.99	2.78	1

44	6.4	0.45	0.08	12	13	0.53	1.65	2
45	6.5	0.63	0.12	11	24	1.12	3.41	0
46	6.1	0.98	0.16	13	60	1.01	2.1	2
47	6.1	1.73	0.19	15	106	1.95	3.85	1
48	6.4	1.01	0.2	12	82	1.34	1.99	2
49	6.2	1.57	0.16	7	103	1	3.23	1
50	6.1	4.03	0.17	30	131	1.56	3.44	0
51	6	5.63	0.17	19	135	1.76	3.44	1
52	6.3	4.44	0.17	20	127	1.77	3.37	1
53	6.4	4.05	0.14	16	106	2.03	2.89	1
54	6.3	6.17	0.16	22	121	2.03	3.16	0
55	6.4	6.25	0.13	20	117	1.02	2.68	0
56	6.3	10.26	0.12	13	120	0.86	3.41	0
57	6.4	8.48	0.11	10	143	0.7	2.77	1
58	6.6	5.91	0.13	9	119	0.85	2.61	1
59	7.1	0.81	0.12	9	29	0.56	3.4	0
60	7	1.63	0.1	11	78	0.72	3.4	0
61	6.9	4.65	0.14	12	118	0.67	2.75	0
62	6.9	8.56	0.1	16	48	0.82	3.06	1
63	7.3	5.54	0.08	15	75	0.76	3.58	0
64	7.8	1.59	0.13	14	38	0.67	2.55	0
65	7.8	4.43	0.21	17	78	0.63	4.13	1
66	7.8	1.15	0.2	12	14	0.48	3.99	0
67	7.4	2.3	0.15	17	87	0.73	2.96	1
68	7.3	4.16	0.21	12	43	0.87	4.26	0
69	7.4	2.49	0.15	11	100	0.81	3.03	1
70	7.5	1.27	0.15	13	8	1.09	3.03	0
71	7.4	1.64	0.18	17	12	1.15	3.58	0
72	7.7	0.75	0.12	11	18	2.1	2.34	1
73	7.5	1.05	0.12	13	35	2	2.8	0
74	7.1	4.43	0.2	12	69	1.8	3.99	1
75	7.1	4.98	0.18	13	92	1.55	3.58	0
76	7.3	2.1	0.21	12	21	1.32	4.2	0
77	7.3	1.26	0.16	9	93	0.98	3.1	0
78	7.5	0.96	0.08	12	66	0.81	2.58	1
79	7.4	0.5	0.19	12	24	0.78	3.78	0
80	7.2	0.88	0.12	110	52	0.88	3.41	0
81	7.4	0.51	0.15	19	25	1.05	3.03	1
82	7.1	0.58	0.17	13	28	1.25	3.3	1
83	6.6	1.39	0.18	14	96	1.41	3.58	0
84	7.4	0.56	0.09	14	29	1.5	3.72	0

85	6.3	1.05	0.17	24	101	1.22	3.44	0
86	5.3	1	0.17	12	98	1.75	3.44	0
87	5.3	2.34	0.17	12	87	2.02	3.3	1
88	6.1	1.74	0.14	46	134	2.1	2.75	0
89	6.7	0.57	0.12	34	28	1.95	3.48	0
90	6.7	0.4	0.15	17	26	1.82	3.03	0
91	6.4	1.38	0.17	74	62	1.67	3.44	0
92	6.7	0.76	0.12	22	18	0.71	2.85	1
93	6	0.6	0.16	21	16	0.68	3.16	0
94	5.9	0.72	0.13	15	12	0.57	2.61	1
95	5.4	2.87	0.09	26	58	0.88	2.79	1
96	6	0.96	0.1	105	20	0.85	2.93	0
97	5.8	4.59	0.08	24	78	0.71	2.65	0
98	6.1	2.56	0.11	16	92	0.47	2.85	1
99	5.9	11.32	0.14	14	98	0.56	2.75	1
100	6	9.21	0.13	23	80	0.46	2.61	1
….	….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….	…..
1529	5.3	0.36	0.02	6	31	0.42	2.51	1
1530	6.3	0.51	0.09	3	20	0.32	2.52	1
1531	6.5	0.41	0.02	2	22	0.41	2.53	2
1532	6.1	0.2	0.07	3	36	0.43	2.54	1
1534	5.4	0.45	0.12	2	30	0.36	3.4	1
1534	5.2	0.94	0.02	3	50	0.87	2.61	1
1535	6	0.16	0.02	2	34	0.89	2.63	1
1536	6.1	2	0.09	2	43	0.3	2.53	1
1537	5.7	1.6	0.06	3	45	0.31	1.02	2
1538	6	0.48	0.01	3	10	0.32	0.17	2
1539	5.3	0.43	0.01	4	11	0.22	2.12	2
1540	5.3	0.49	0.02	2	9	0.14	2.33	2
1541	6.2	0.48	0.02	3	32	0.57	2.34	2
1542	5.8	0.39	0.02	3	14	0.34	2.4	2

A-3 Unused 30 Soil Samples

pH	Salinity	Nitrogen	Phosphorus-B	Sulfur	Boron	Organic matter	Soil class
5.4	6.09	0.14	19	108	0.89	2.89	1
6.8	0.22	0.03	3	11	0.1	0.51	2
7.3	5.54	0.08	15	75	0.76	3.58	0
5.8	2.6	0.07	3	12	0.51	1.19	2

5.3	2.38	0.14	1	30	0.35	3.91	1
6	0.24	0.07	3	41	0.32	1.45	2
5	2.38	0.14	2	15	0.61	0.14	2
5.8	1.53	0.09	2	24	0.59	3.04	1
5.7	0.48	0.09	21	53	0.65	3.92	1
6	1.36	0.08	2	6	1.11	0.04	2
5.3	2.72	0.16	2	19	0.56	0.24	2
5	5.77	0.17	21	77	1.01	3.44	1
5.6	0.43	0.04	8	73	0.51	3.12	1
7.1	2.5	0.14	2	205	0.99	3.09	1
6.8	0.7	0.08	0	70	0.63	1.69	2
7	0.3	0.07	4	122	1.1	1.52	2
5.7	3.57	0.21	2	84	0.71	2.71	1
5.3	0.49	0.02	2	9	0.14	2.33	2
8.1	13.9	0.08	4	145	1.4	1.93	2
6.4	0.13	0.1	11	49	0.5	3.92	1
6.1	1.73	0.19	15	106	1.95	3.85	1
5.3	2.34	0.17	12	87	2.02	3.3	1
6.6	2	0.12	2	121	0.99	2.32	2
5.4	3.06	0.18	5	44	0.38	3.91	1
5.4	2.06	0.14	14	104	1.38	2.75	0
7.1	2.3	0.07	2	111	0.46	1.25	2
5.5	0.85	0.05	3	7	1.05	0.24	2
7.7	0.9	0.05	0	80	0.77	1.01	2
5.8	1.53	0.09	2	24	0.59	4.02	1
6.9	8.56	0.1	16	48	0.82	3.07	1

B. CROPS DATASET

B-1 Crops Raw Data

Sample no	Location	Soil_class	Season	Soil_ph_for_crops	Crops
1	purbo char jabbar	clay loam	all_season	5.5-6.6	brinjal
2	purbo char jabbar	clay loam	all_season	6.00-6.8	pumpkin
3	purbo char jabbar	clay loam	kharif	5.5-7.00	spiny gourd
4	purbo char jabbar	clay loam	all_season	6.00-7.00	rice
5	purbo char jabbar	clay loam	kharif	6.5-7.00	sola
6	purbo char jabbar	clay loam	rabi	6.5-7.00	grass pea
7	purbo char jabbar	clay loam	rabi	5.8-6.2	gourd
8	purbo char jabbar	clay loam	rabi	5.00-7.00	linseed
9	purbo char jabbar	Loam	kharif	6.00-6.8	sponge gourd

10	purbo char jabbar	Loam	kharif	6.5-7.00	ladies-finger
11	purbo char jabbar	Loam	kharif	5.5-7.5	yardlong bean
12	purbo char jabbar	Loam	kharif	5.5-6.7	bitter melon
13	purbo char jabbar	Loam	all_season	6.00-6.8	pumpkin
14	purbo char jabbar	Loam	kharif	5.5-6.7	parwal
15	purbo char jabbar	Loam	kharif	6.00-6.5	cucumber
16	purbo char jabbar	Loam	kharif	5.5-7.00	spiny gourd
17	purbo char jabbar	loam	kharif	6.00-6.7	snake gourd
18	purbo char jabbar	loam	all_season	5.8-6.00	maize
19	purbo char jabbar	loam	all_season	6.00-7.00	rice
20	purbo char jabbar	loam	rabi	6.5-7.5	sunflower
21	purbo char jabbar	loam	all_season	7.00-7.00	sesame
22	purbo char jabbar	loam	all_season	6.00-6.5	turmeric
23	purbo char jabbar	loam	kharif	6.3-6.5	chili
24	purbo char jabbar	loam	kharif	6.5-7.00	sola
25	purbo char jabbar	loam	rabi	6.5-7.00	grass pea
26	purbo char jabbar	loam	kharif	5.00-7.4	jute
27	purbo char jabbar	loam	kharif	6.5-7.00	ladies-finger
28	purbo char jabbar	loam	kharif	5.5-7.5	yardlong bean
29	purbo char jabbar	loam	kharif	5.5-6.7	bitter melon
….					
….					
…					
1152	char amanullah	loam	rabi	6.00-7.00	capsicum
1153	char amanullah	loam	rabi	6.00-7.00	cauliflower
1154	char amanullah	loam	rabi	6.00-7.00	wheat
1155	char amanullah	loam	all_season	5.8-6.00	maize
1156	char amanullah	loam	all_season	6.00-7.00	rice
1157	char amanullah	loam	rabi	6.5-7.5	sunflower
1158	char amanullah	loam	rabi	5.5-6.8	mustard
1159	char amanullah	loam	all_season	7.00-7.00	sesame
1160	char amanullah	loam	kharif	6.3-6.5	chili
1161	char amanullah	loam	rabi	6.00-6.5	potato
1162	char amanullah	loam	rabi	6.5-7.00	grass pea

B-2 Processed Crops Data

Sample no	Location	Soil_class	Season	Soil_ph_for_crops	Crops
1	19	1	0	2	2
2	19	1	0	11	17
3	19	1	1	5	22

4	19	1	0	12	18
5	19	1	1	14	21
6	19	1	2	14	9
7	19	1	2	8	8
8	19	1	2	0	12
9	19	2	1	11	23
10	19	2	1	14	11
11	19	2	1	6	28
12	19	2	1	3	1
13	19	2	0	11	17
14	19	2	1	3	15
15	19	2	1	9	7
16	19	2	1	5	22
17	19	2	1	10	20
18	19	2	0	7	13
19	19	2	0	12	18
20	19	2	2	15	24
21	19	2	0	16	19
22	19	2	0	9	26
23	19	2	1	13	6
24	19	2	1	14	21
25	19	2	2	14	9
26	19	2	1	1	10
27	19	2	1	14	11
28	19	2	1	6	28
29	19	2	1	3	1
….	….	….	…..	…..	….
….	…..	….	…..	…..	…
…	…	….	…..	…	….
1152	0	2	2	12	3
1153	0	2	2	12	5
1154	0	2	2	12	27
1155	0	2	0	7	13
1156	0	2	0	12	18
1157	0	2	2	15	24
1158	0	2	2	4	14
1159	0	2	0	16	19
1160	0	2	1	13	6
1161	0	2	2	9	16
1162	0	2	2	14	9

B-3 Unused 30 Samples

Location	Soil_class	Season	Soil_ph_for_crops	Crops
20	2	1	9	7
15	2	1	9	7
7	1	1	14	21
13	1	2	14	9
2	2	2	14	9
4	2	2	12	3
7	2	1	9	7
1	2	2	12	5
19	0	2	12	5
10	2	1	14	21
3	2	2	12	5
5	1	2	8	8
19	2	1	14	11
21	2	1	14	11
7	2	2	12	5
4	0	0	2	2
18	0	0	12	18
10	2	1	3	15
16	1	0	12	18
18	1	2	14	9
14	1	1	5	22
3	2	1	14	11
10	1	0	2	2
4	2	0	11	17
20	2	0	7	13
17	2	1	3	1
1	2	0	11	17
19	2	0	7	13
11	1	2	0	12
7	0	0	2	2

C. CODE FOR SOIL CLASSIFICATION

```
# libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as splt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.tree import plot_tree
from sklearn.metrics import roc_curve
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import f1_score
from sklearn.metrics import auc
from sklearn import tree
import matplotlib.pyplot as plt
import sklearn.metrics as metric
#load dataset
df=pd.read_csv('F:\\Soil31.csv')
df.isnull().sum()
df=df.drop('Location', axis=1)
df['Soil class'].unique()
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df['Soil class']= label_encoder.fit_transform(df['Soil class'])
df['Soil class'].unique()
target=df['Soil class']
df1=df.copy()
df1=df1.drop('Soil class',axis=1)
x=df1
y=target
```

```
# split the dataset
from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest= train_test_split(x,y, test_size= .3, random_state=1)
# Other libraries
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import roc_auc_score
from sklearn.metrics import classification_report
# Machine Learning
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
#Support Vector Machine
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
sv=SVC(kernel='linear', random_state=1)
sv.fit(xtrain, ytrain)
y_pred_svm = sv.predict(xtest)
accuracy_svm = round(accuracy_score(ytest, y_pred_svm)*100, 2)
precision_svm = round(precision_score(ytest, y_pred_svm, average='weighted')*100, 2)
```

```
recall_svm = round(recall_score(ytest, y_pred_svm, average='weighted')*100, 2)
f1_score_svm = round(f1_score(ytest, y_pred_svm, average='weighted')*100, 2)
print("Accuracy score of SVM: "+str(accuracy_svm)+" %")
print("Precision score of SVM: "+str(precision_svm)+" %")
print("Recall score of  SVM: "+str(recall_svm)+" %")
print("F1 score of SVM: "+str(f1_score_svm)+" %")
import seaborn as sns
conf_matrix = confusion_matrix(ytest, y_pred_svm)
fig, ax = plt.subplots(figsize=(6,6))
ax.matshow(conf_matrix, cmap=plt.cm.Blues, alpha=0.3)
for i in range(conf_matrix.shape[0]):
    for j in range(conf_matrix.shape[1]):
        ax.text(x=j, y=i,s=conf_matrix[i, j], va='center', ha='center', size='xx-large')
plt.xlabel('Predictions', fontsize=18)
plt.ylabel('Actuals', fontsize=18)
plt.title('Confusion Matrix(SVM)', fontsize=18)
plt.show()
print(classification_report(ytest, y_pred_svm))
#Decision Tree
from sklearn.tree import DecisionTreeClassifier
dt=DecisionTreeClassifier()
dt.fit(xtrain, ytrain)
y_pred_dt = dt.predict(xtest)
accuracy_dt = round(accuracy_score(ytest, y_pred_dt)*100, 2)
precision_dt = round(precision_score(ytest, y_pred_dt, average='weighted')*100, 2)
recall_dt = round(recall_score(ytest, y_pred_dt, average='weighted')*100, 2)
f1_score_dt = round(f1_score(ytest, y_pred_dt, average='weighted')*100, 2)
print("Accuracy score of Decision Tree: "+str(accuracy_dt)+" %")
print("Precision score of Decision Tree: "+str(precision_dt)+" %")
print("Recall score of Decision Tree: "+str(recall_dt)+" %")
print("F1 score of Decision Tree: "+str(f1_score_dt)+" %")
```

```
conf_matrix = confusion_matrix(ytest, y_pred_dt)
fig, ax = plt.subplots(figsize=(6,6))
ax.matshow(conf_matrix, cmap=plt.cm.Blues, alpha=0.3)
for i in range(conf_matrix.shape[0]):
for j in range(conf_matrix.shape[1]):
ax.text(x=j, y=i,s=conf_matrix[i, j], va='center', ha='center', size='xx-large')
plt.xlabel('Predictions', fontsize=18)
plt.ylabel('Actuals', fontsize=18)
plt.title('Confusion Matrix(DT)', fontsize=18)
plt.show()
print(classification_report(ytest, y_pred_dt))
# Multilayer perceptron
mlp = MLPClassifier(solver='lbfgs', alpha=1e-5,
hidden_layer_sizes=(13,13,13),max_iter=5000)
mlp.fit(xtrain,ytrain)
y_pred_mlp = mlp.predict(xtest)
accuracy_mlp = round(accuracy_score(ytest, y_pred_mlp)*100, 2)
precision_mlp = round(precision_score(ytest, y_pred_mlp, average='weighted')*100, 2)
recall_mlp = round(recall_score(ytest, y_pred_mlp, average='weighted')*100, 2)
f1_score_mlp = round(f1_score(ytest, y_pred_mlp, average='weighted')*100, 2)
print("Accuracy score of Multi-Layer Perceptron: "+str(accuracy_mlp)+" %")
print("Precision score of Multi-Layer Perceptron: "+str(precision_mlp)+" %")
print("Recall score of Multi-Layer Perceptron: "+str(recall_mlp)+" %")
print("F1 score of Multi-Layer Perceptron: "+str(f1_score_mlp)+" %")
#Random Forest
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(xtrain,ytrain)
y_pred_rf = rf.predict(xtest)
accuracy_rf = round(accuracy_score(ytest, y_pred_rf)*100, 2)
precision_rf = round(precision_score(ytest, y_pred_rf,average='weighted')*100, 2)
```

```
recall_rf = round(recall_score(ytest, y_pred_rf,average='weighted')*100, 2)
f1_score_rf = round(f1_score(ytest, y_pred_rf,average='weighted')*100, 2)
print("Accuracy score of Random Forest: "+str(accuracy_rf)+" %")
print("Precision score of Random Forest: "+str(precision_rf)+" %")
print("Recall score of Random Forest: "+str(recall_rf)+" %")
print("F1 score of Random Forest: "+str(f1_score_rf)+" %")
#Logistic Regression
lr = LogisticRegression()
lr.fit(xtrain, ytrain)
y_pred_lr = lr.predict(xtest)
accuracy_lr = round(accuracy_score(ytest, y_pred_lr)*100, 2)
precision_lr = round(precision_score(ytest, y_pred_lr,average='weighted' )*100, 2)
recall_lr = round(recall_score(ytest, y_pred_lr, average='weighted')*100, 2)
f1_score_lr = round(f1_score(ytest, y_pred_lr, average='weighted')*100, 2)
print("Accuracy score of Logistic Regression: "+str(accuracy_lr)+" %")
print("Precision score of Logistic Regression: "+str(precision_lr)+" %")
print("Recall score of Logistic Regression: "+str(recall_lr)+" %")
print("F1 score of Logistic Regression: "+str(f1_score_lr)+" %")
# Naïve Bayes
nb = GaussianNB()
nb.fit(xtrain,ytrain)
y_pred_nb = nb.predict(xtest)
accuracy_nb = round(accuracy_score(ytest, y_pred_nb)*100, 2)
precision_nb = round(precision_score(ytest, y_pred_nb, average='weighted')*100, 2)
recall_nb = round(recall_score(ytest, y_pred_nb, average='weighted')*100, 2)
f1_score_nb = round(f1_score(ytest, y_pred_nb, average='weighted')*100, 2)
print("Accuracy score of Naive Bayes: "+str(accuracy_nb)+" %")
print("Precision score of Naive Bayes: "+str(precision_nb)+" %")
print("Recall score of Naive Bayes: "+str(recall_nb)+" %")
print("F1 score of Naive Bayes: "+str(f1_score_nb)+" %")
```

D. CODE FOR CROPS CULTIVATION PREDICTION

```
# load the crops data set
df=pd.read_csv('F:\\randomCrop.csv')
df=df.drop('Crops',axis=1)
df['Crops'].unique()
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df['Crops']= label_encoder.fit_transform(df['Crops'])
df['Crops'].unique()
df['Location'].unique()
df['Location']= label_encoder.fit_transform(df['Location'])
df['Location'].unique()
df['Season'].unique()
df['Season']= label_encoder.fit_transform(df['Season'])
df['Season'].unique()
df['Soil_pH_for_crops']= label_encoder.fit_transform(df['Soil_pH_for_crops'])
df['Soil_pH_for_crops'].unique()
df['Soil_pH_for_crops'].unique()
df.isnull().sum()
df
target=df['Crops']
df1=df.copy()
df1=df1.drop('Crops',axis=1)
x=df1
y=target
#splitting the dataset
from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest= train_test_split(x,y, test_size= .33, random_state=1)
#Support Vector Machine
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
```

```
sv=SVC(kernel='linear', random_state=1)
sv.fit(xtrain, ytrain)
y_pred_svm = sv.predict(xtest)
accuracy_svm = round(accuracy_score(ytest, y_pred_svm)*100, 2)
precision_svm = round(precision_score(ytest, y_pred_svm, average='weighted')*100, 2)
recall_svm = round(recall_score(ytest, y_pred_svm, average='weighted')*100, 2)
f1_score_svm = round(f1_score(ytest, y_pred_svm, average='weighted')*100, 2)
print("Accuracy score of SVM: "+str(accuracy_svm)+" %")
print("Precision score of SVM: "+str(precision_svm)+" %")
print("Recall score of  SVM: "+str(recall_svm)+" %")
print("F1 score of SVM: "+str(f1_score_svm)+" %")
#Decision Tree
from sklearn.tree import DecisionTreeClassifier
dt=DecisionTreeClassifier()
dt.fit(xtrain, ytrain)
y_pred_dt = dt.predict(xtest)
accuracy_dt = round(accuracy_score(ytest, y_pred_dt)*100, 2)
precision_dt = round(precision_score(ytest, y_pred_dt, average='weighted')*100, 2)
recall_dt = round(recall_score(ytest, y_pred_dt, average='weighted')*100, 2)
f1_score_dt = round(f1_score(ytest, y_pred_dt, average='weighted')*100, 2)
print("Accuracy score of Decision Tree: "+str(accuracy_dt)+" %")
print("Precision score of Decision Tree: "+str(precision_dt)+" %")
print("Recall score of Decision Tree: "+str(recall_dt)+" %")
print("F1 score of Decision Tree: "+str(f1_score_dt)+" %")
#MLPNN
mlp = MLPClassifier(solver='lbfgs', alpha=1e-5,
hidden_layer_sizes=(13,13,13),max_iter=5000)
mlp.fit(xtrain,ytrain)
y_pred_mlp = mlp.predict(xtest)
accuracy_mlp = round(accuracy_score(ytest, y_pred_mlp)*100, 2)
precision_mlp = round(precision_score(ytest, y_pred_mlp, average='weighted')*100, 2)
```

```
recall_mlp = round(recall_score(ytest, y_pred_mlp, average='weighted')*100, 2)
f1_score_mlp = round(f1_score(ytest, y_pred_mlp, average='weighted')*100, 2)
print("Accuracy score of Multi-Layer Perceptron: "+str(accuracy_mlp)+" %")
print("Precision score of Multi-Layer Perceptron: "+str(precision_mlp)+" %")
print("Recall score of Multi-Layer Perceptron: "+str(recall_mlp)+" %")
print("F1 score of Multi-Layer Perceptron: "+str(f1_score_mlp)+" %")
#Random Forest
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(xtrain,ytrain)
y_pred_rf = rf.predict(xtest)
accuracy_rf = round(accuracy_score(ytest, y_pred_rf)*100, 2)
precision_rf = round(precision_score(ytest, y_pred_rf,average='weighted')*100, 2)
recall_rf = round(recall_score(ytest, y_pred_rf,average='weighted')*100, 2)
f1_score_rf = round(f1_score(ytest, y_pred_rf,average='weighted')*100, 2)
print("Accuracy score of Random Forest: "+str(accuracy_rf)+" %")
print("Precision score of Random Forest: "+str(precision_rf)+" %")
print("Recall score of Random Forest: "+str(recall_rf)+" %")
print("F1 score of Random Forest: "+str(f1_score_rf)+" %")
#Naive Bayes
nb = GaussianNB()
nb.fit(xtrain,ytrain)
y_pred_nb = nb.predict(xtest)
accuracy_nb = round(accuracy_score(ytest, y_pred_nb)*100, 2)
precision_nb = round(precision_score(ytest, y_pred_nb, average='weighted')*100, 2)
recall_nb = round(recall_score(ytest, y_pred_nb, average='weighted')*100, 2)
f1_score_nb = round(f1_score(ytest, y_pred_nb, average='weighted')*100, 2)
print("Accuracy score of Naive Bayes: "+str(accuracy_nb)+" %")
print("Precision score of Naive Bayes: "+str(precision_nb)+" %")
print("Recall score of Naive Bayes: "+str(recall_nb)+" %")
print("F1 score of Naive Bayes: "+str(f1_score_nb)+" %")
```