Evaluating Machine Learning Models And Building A Web Application To Recommend Crop Yields Using Historical Weather and Soil Data

Md Shihab As Samad

Applied Research Project submitted in partial fulfilment of the requirements for the degree of **MSc Business Analytics** at Dublin Business School

Supervisor: Dr. Assem Abdelhak

Augst, 2024

DECLARATION

‘I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of **MSc in Business Analytics** is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.’

Signed: **Md Shihab As Samad**

Student Number: **20019158**

Date: **27 Aug 2024**

ACKNOWLEDGMENTS

I want to thank **Dr. Assem Abdelhak** from the bottom of my heart for all the advice, inspiration, and help they have given me during this study project. I'm grateful for the help he has given me; his knowledge has been very helpful in forming this thesis.

I'd also like to thank my family, who have been very understanding and helpful while I've been in school. I am always driven by the fact that they always believe in me and back me.

Finally, I want to thank my friends for always being there for me and for being such a great source of support and drive. This trip is now easier to handle and more fun thanks to their company and help.

To everyone who helped make this possible, thank you very much.

ABSTRACT

This research proposes a machine-learning crop recommendation system. It optimizes crop choices for soil and environmental conditions to increase agricultural output and sustainability. The study uses Decision Trees, Random Forest, SVM, and Logistic Regression to choose the best crops for a location. These models were trained and evaluated using soil, meteorological, and crop data. Random Forest dominated accuracy, precision, recall, and F1 score with a 97% accuracy rate. By using SMOTE for data preprocessing, outlier detection, and data balancing, the system ensures credible predictions. The research continues by discussing hybrid ways to include real-time data and increase model accuracy. This research enables smarter, data-driven agricultural systems that reduce farming's environmental effect and boost harvest yields.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background

Since food and other agricultural products are required for survival and population growth, agriculture has moulded human civilization. Many people's economy revolves on agriculture since it is such important. With about 58% of India's workers in agricultural, this helps to account for 17% of its GDP. This underlines how crucial it is to millions of people and rural areas. But food consumption is rising as the population of the planet increases at an unheard-of pace, and the agriculture sector is under trouble. Climate change is making weather less predictable, which is challenging farmers. Degradation of soil from unsustainable farming and too high usage aggravates these problems. For this reason, contemporary farms find useless centuries-old farming practices. Farmers are so increasingly depending on contemporary technologies. Leading is machine learning (ML), which can analyse vast amounts of data and produce ever accurate forecasts and decisions. Based on climate and soil type, it can guide farmers in selecting crops (Niedbała et al., 2024). Sustainable farming, more food output, and less waste could all follow from this.

1.2 Motivation

This work is driven by the urgent need to adapt farming practices to a fast-changing global environment. With an unprecedented worldwide population growth, food demand is at record highs. Climate warming also makes farming harder owing to unpredictable weather and more catastrophic disasters. Because of these concerns, farmers are finding it harder to choose crops and keep their farms in good condition. Farmers are having trouble choosing crops and maintaining their farms due to these worries. Farming approaches based on experience and intuition are becoming obsolete. The gaps machine learning can fill drive this research. This initiative creates weather, soil, and other prediction models to assist farmers make better decisions. These technologies should boost agricultural yields, reduce losses, and promote sustainable farming to ensure the world's food supply (Devi and Selvakumari, 2024).

1.3 Problem Statement

Inability to predict which crops will grow in particular soil types is a key agricultural difficulty. Farmers traditionally choose crops based on personal experience, social norms, and market trends. This strategy produces inefficient crop choices, low yields, and financial losses. Climate change causes unpredictable rainfall, temperature swings, and more intense weather occurrences, complicating these already difficult decisions. No viable, data-driven crop selection method exists, which is the main purpose of this research. Current farming methods lack reliable, scientific projections, which hinders decision-making. This study suggests utilizing machine learning to predict which crops will thrive in given areas based on huge soil, weather, and other data. The project intends to increase agricultural productivity, reduce crop failure, and promote sustainable farming by educating farmers.

1.4 Objective

This work intends to develop, test, and enhance a machine learning model that correctly forecasts, depending on soil and environmental conditions, which crops will flourish on a specific site. This paper aims to provide producers data-driven solutions to enable maximum agricultural output and lower risk. The research will so concentrate on: in order to achieve this:

Check and evaluate several machine learning techniques first. Sort machine learning techniques to choose the best one for estimating soil and environmental element compatibility with crops. There are included the following techniques: Random forests, naive bayes, logistic regression.

2. Map crops and soil holistically: This system will include nitrogen, phosphorus, potassium, temperature, humidity, and rainfall data. A complete map of the most suitable produce can also be provided.

3. Use real-world data from several industries for training and testing machine learning models to ensure dependability and precision. Every model will be assessed in the article to find the best.

This study will help producers make better decisions by providing relevant information. Machine learning models can help farmers choose the best crops for their fields, which will lead to higher yields and reduced environmental impact (Devi & Selvakumari, 2024).

1.5 Research Scope

This research covers several key areas, all aimed at enhancing agricultural practices through the use of machine learning:

Data Collection and Preprocessing: Data on pH, temperature, humidity, rainfall, and levels of nitrogen, phosphorus, and potassium (N, P, K) in the soil will be collected as the first step in the study. To feed machine learning models just the best data, this data will be cleaned to remove errors and missing numbers.

A number of Machine Learning models, including Logistic Regression, Random Forest, and Naive Bayes, will be applied and appraised. The data will be used to train these models to discover which crops perform best under given conditions. Better model accuracy is another project goal.

3. Model Performance Evaluation: Each machine learning model will be evaluated using real data. Model accuracy will be assessed by comparing actual crop yields to projected results. Best model will be identified and improved.

4. Climate Change Impact: The project will explore how climate change affects machine learning model estimates. The project will examine multiple climatic scenarios to assess model robustness and regional applicability. This ensures models work well in many environments (Niedbała et al., 2024).

2 LITERATURE REVIEW

2.1 Introduction

As food insecurity, climate change, and sustainable resource management become more important, the agricultural sector is changing. An increasing global population puts additional pressure on agricultural systems to produce food sustainably. Innovators have used cutting-edge technology, notably estimate and forecast tools, to enhance agricultural production and environmental resilience. As digital agriculture systems and machine learning algorithms integrate, crop yields, soil health, and environmental effects can be better predicted. This chapter reviews the literature on cutting-edge farming technologies and machine learning. Several methodologies and models from past research to assess the field's current and future state will be reviewed too.

2.2 Previous Literature

More people are exploring applying machine learning in agriculture to boost output, sustainability, and climate resistance. This section critically analyses the methodology, conclusions, and research implications of the important studies that have affected crop recommendation and yield prediction systems.

Niedbała et al. (2024) paper "Predictions and Estimations in Agricultural Production under a Changing Climate" focuses on enhancing agricultural output with predictive technologies such as machine learning models, GIS tools, and satellite remote sensing. This study shows that cutting-edge approaches like ELM, SVR, and ANN can improve forecast accuracy, which is crucial for lowering climate uncertainty. However, these methods aren't always relevant when resources are few because the study requires high-quality data, which differs by region.

Research and funding are needed to improve data integrity and make the tools more versatile in agricultural settings.

Patil and Mane's (2024) "Crop Recommendation in Precision Agriculture Using Machine Learning Techniques," examines an ML model for crop recommendation to boost agricultural productivity. It is found that Random Forest outperformed SVM with 93.7% prediction accuracy. Despite the positive findings, the study's use of Kaggle datasets raises questions about generalizability. To enable its widespread application, the model needs more cross-regional validation because data availability and quality are crucial to its performance.

"Crop, Fertilizer, and Pesticide Recommendation using Ensemble Method and Sequential Convolutional Neural Network" (Ketheneni et al., 2024) uses a CNN to identify pesticides and multiple machine learning algorithms to recommend crops and fertilizers. The ensemble technique generated precise crop recommendations with 96.44% accuracy. The pesticide identification module's performance depends on photo quality, so it may not be viable in real life, especially with low-quality data. This emphasizes the need for robust models that can manage varying data quality levels and continual improvement.

The Tamilarasan (2024) "Crop Plantation Suggestion and Yield Forecasting System" uses Random Forest and LSTM models for crop recommendation and yield forecasting. Crop yield prediction is substantially improved by this two-pronged technique. The study emphasises data updates and validation across crops and locations to ensure the system's wider applicability. The system's performance depends on input data quality.

In their paper "Crop Prediction and Mapping Using Different Machine Learning Techniques," Devi and Selvakumari (2024) examine how Naive Bayes, Random Forest, and Logistic Regression can be used to predict which crops are best for a given soil type and environmental conditions. The Random Forest model predicted crops best with 99.09% precision. The study underscores a recurring issue with agricultural machine learning: the necessity for large, high-quality data. Data availability limits the usefulness of these models, so more study on data augmentation and model validation is needed to improve generalizability.

In "Corn Yield Prediction Model with Deep Neural Networks for Smallholder Farmer Decision Support System," Olisah et al. (2024) create a DNNR to forecast maize yield using soil and weather data. With better generalizability than traditional methods, the model became more trustworthy in shifting agricultural environments. Due to the study's limitations and the difficulties of simulating non-linear variable interactions, the model's robustness across crops and regions must be improved.

Bangladeshi researchers examined how they used deep learning models to predict weather and make context-based crop recommendations in their study "Agricultural Recommendation System based on Deep Learning: A Multivariate Weather Forecasting Approach" (Zubair et al., 2024). The study used a Stacked Bi-LSTM model, which projected weather accurately and was resilient. The system has promise, but lack of well-structured and precise data limits its use. To work in varied agricultural circumstances, the system must be validated across crops and locations.

In "Artificial Neural Networks Based Integrated Crop Recommendation System Using Soil and Climatic Parameters," Madhuri and Indiramma (2021) use soil and climatic data to recommend crops. The model's 96% accuracy is impressive, but regional data sets limit its usefulness. This implies that more research and model expansion are needed to generalize across soil types and larger locations.

In "Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables," Dey et al. (2024) evaluate crop recommendation and nutrient adjustment models. The most accurate model was XGBoost, with precision rates of 99%. However, merging horticultural and agricultural crops into one model highlights the study's weaknesses, such as misclassification and model performance range. To ensure crop suggestion accuracy and dependability, the algorithm must be improved and input data properly selected.

In "Developing Automated Machine Learning Approach for Fast and Robust Crop Yield Prediction Using a Fusion of Remote Sensing, Soil, and Weather Dataset," Kheir et al. (2024) introduce an AutoML-based method. The research predicts wheat yield using many data sources. R2s of 0.51 and Willmott degrees of agreement of 0.82 indicate high performance. However, the study acknowledges that data quality varies and that additional validation across contexts is needed to make the AutoML approach more practical and dependable in varied agricultural situations.

In their article "A Real-Time Crop Recommendation System Using IoT and Machine Learning", Choudhury et al. (2023) combine advanced machine learning with real-time soil monitoring to boost crop output. The system can't classify crops with 99% accuracy due to real-time data quality. More development is needed to make it applicable across agricultural settings.

In their study "Automated Land Suitability Evaluation for Crop Cultivation Using a Hybrid Approach of Multicriteria Decision Analysis and Machine Learning," Shukla and Singh (2024) use machine learning and MDA to evaluate crop land suitability. The model is accurate but only works in areas with high-quality input data, therefore it must be tested elsewhere.

According to Choudhury et al. (2023), "An Acquisition Based Optimized Crop Recommendation System with Machine Learning Algorithm" uses numerous machine learning classifiers and the Moth Flame Optimization (MFO) method to optimize crop recommendations. The study's 99.32% accuracy shows the promise of optimisation and machine learning technologies. However, validation in various soil conditions and geographical locations may limit the model's utility.

The Akkemet al. (2023) paper "Streamlit Application for Advanced Ensemble Learning Methods in Crop Recommendation Systems" uses the Streamlit framework and advanced ensemble learning methods to improve crop recommendation systems. Although computational complexity and model interpretability may limit its use in circumstances with limited resources, the study showed significant forecasting performance increases.

In their study "Data-Driven Analysis and Machine Learning-Based Crop and Fertilizer Recommendation System for Revolutionizing Farming Practices," Musanase et al. (2023) use machine learning and Internet of Things technology to transform Rwanda's agriculture. Lack of comprehensive and consistent data limits the system's crop prediction accuracy to 97%. For greater impact, more validation across crop types and locations is needed.

In "Ensemble Machine Learning-Based Recommendation System for Effective Prediction of Suitable Agricultural Crop Cultivation," Hasan et al. (2023) boost Bangladeshi food production with an ensemble machine learning-based recommendation system. The model's accuracy was outstanding, although overfitting and more testing with different crop types and regions were issues.

The "Multimodal Machine Learning Based Crop Recommendation and Yield Prediction Model" uses cutting-edge machine learning to propose crops and estimate yields (Gopi and Karthikeyan, 2023). The model's complexity and need for validation across crops and geographies limit the system's high crop recommendation accuracy.

In their 2023 paper "Crop Prediction Model Using Machine Learning Algorithms," Elbasi et al. examine machine learning methods that help improve crop forecasts. The study's disadvantages include poor data and model scalability issues across locations and crops. The Bayes Net method scored 99.59%, besting all others.

In "Enhancing Crop Recommendation Systems with Explainable Artificial Intelligence: A Study on Agricultural Decision-Making," Shams et al. (2024) discuss ways to improve crop recommendation systems. The XAI-CROP technique beat rival models in accuracy and interpretability, although integrating XAI with ML models was computationally intensive. These systems need more research to be viable and widely used in agriculture.

2.3 Conclusion

According to studies discussed in this chapter, machine learning and advanced computational tools can greatly improve agricultural output and decision-making. Machine learning methods, from yield prediction frameworks to crop and fertilizer recommendation systems, have been extensively studied. These algorithms consistently beat traditional farming methods in accuracy and efficiency. The current agricultural landscape faces various challenges, including soil quality, climate change, and resource optimization. Modern technologies like explainable AI, hybrid methodologies, and ensemble learning can solve these difficulties creatively and effectively.

Despite their advances, these models have many drawbacks. Advanced approaches' computational complexity, models' adaptability to varied geographies, and the necessity for extensive, high-quality data are some of these obstacles. To maximize these technologies' potential, further research should address these issues. Enhancing models, increasing and diversifying datasets, and guaranteeing that these technologies can be employed in multiple agricultural settings and scale to new regions are needed.

Finally, machine learning must be integrated into agricultural processes to improve reliability, efficiency, and longevity. These technologies enable farmers to make data-driven, well-informed decisions to adapt to changing conditions. If these technology advances continue, they could boost global food security by improving crop yields, resource management, etc.

3 METHODOLOGY

This thesis analyses agricultural data using deep learning and sophisticated machine learning methods to improve crop recommendation systems. Systematically gather and prepare soil, meteorological, and crop data. Next, cutting-edge neural network topologies extract features. Ensemble learning, which combines machine learning models, improves crop recommendation systems.

Data preparation begins with cleaning, standardizing, and structuring raw data for analysis. Soil pH, fertilizer levels, temperature, precipitation, and humidity are important model inputs. Crop suitability scores are calculated using Desto, Random Forest, and Support Vector Machines. Ensemble learning improves accuracy and robustness.

The technique examines the different tactics used, their rationale, and how they can improve crop suggestion precision and uniformity. This project uses different datasets and cutting-edge AI models to improve crop recommendation systems. This information will help farmers make better decisions and increase agricultural yields and sustainability. Figure 3.1 details the proposed system's process.

3.1 Proposed Methodology

The suggested crop recommendation method improves agricultural decision-making with cutting-edge machine learning. The process involves collecting and preparing crop, climate, and soil data. This data is used to train Random Forest, SVM, Logistic Regressor, and Decision Tree models to predict the best crops for certain regions.

Farmers can enter data and get real-time crop suggestions using the system's user-friendly online application. Accuracy, F1 Score, precision, and recall are used to evaluate the system's performance; the Confusion Matrix is most accurate. The system has enormous potential, but its success depends on the quality and diversity of the input data. Validation across multiple crops and places is needed to ensure its widespread use.

Overall, this crop suggestion system is a big advancement in agricultural technology, helping farmers select crops and increase yields. Data-driven decision-making with machine learning enhances accuracy and promotes sustainable farming. Future work will expand the system's use and improve models to boost performance.

.

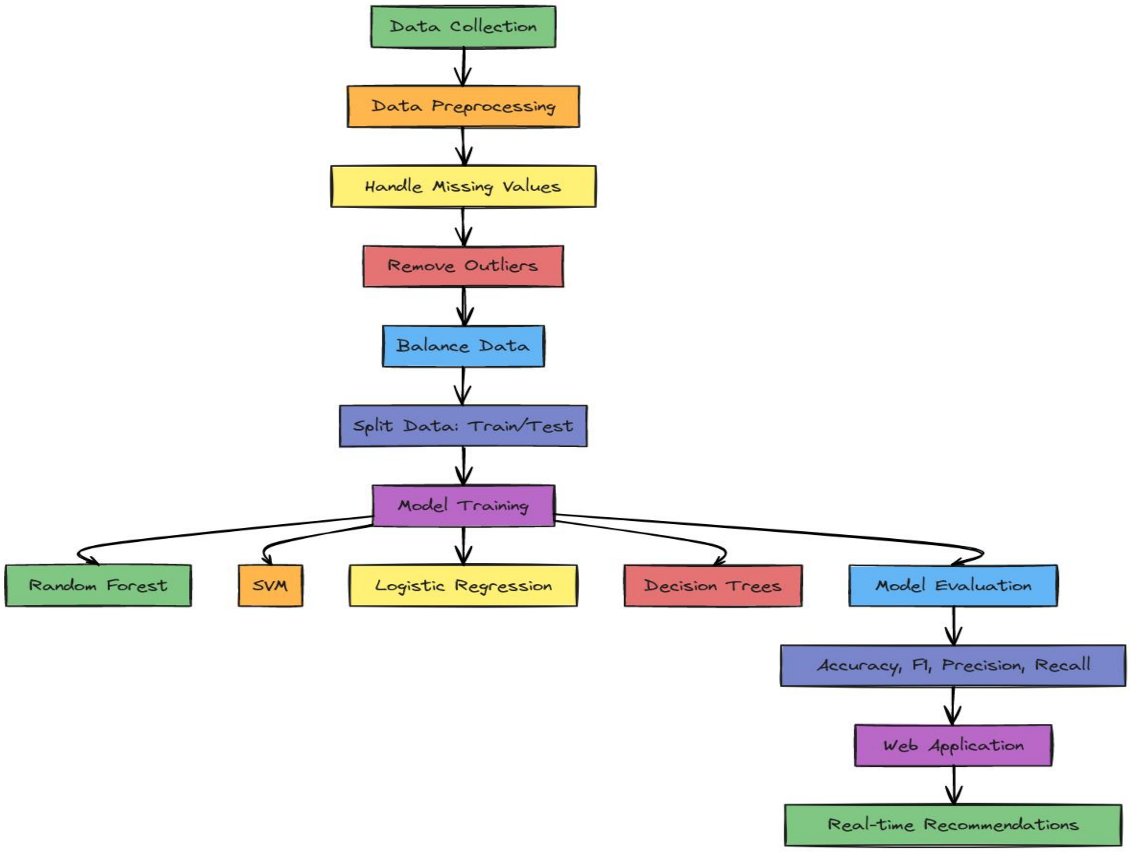


Figure 3.1 Proposed Workflow Diagram for crop recommendation system

3.2 Data Set Description

The dataset used in this study consists of 4,513 records, each containing 11 distinct attributes that characterize various properties associated with agricultural practices. The records are customized to each district and provide extensive data on agricultural inputs, meteorological conditions, and soil characteristics. These exact characteristics are crucial for evaluating the appropriateness of particular crops and determining the required rates of fertilizer.

The dataset exhibits the following characteristics:

District Name: The district name provides a geographical context for the agricultural data collection.

The hue of the soil can provide insights about its richness, capacity for drainage, and concentration of organic material. The dataset contains items characterized by black soil, a trait that is commonly associated with high fertility.

Nitrogen (N): The quantity of nitrogen present in the soil, which is crucial for the productivity of crops and essential for the development of plants.

Phosphorus (P): The quantity of phosphorus present in the soil that is necessary for crops to develop their roots and for energy transmission.

Potassium (K): The soil's potassium level is essential for regulating water, activating enzymes, and sustaining overall plant well-being.

The pH level of the soil indicates its acidity or alkalinity, which in turn affects the availability of nutrients and the activity of microorganisms.

Rainfall (mm): The quantity of precipitation received in the region, which impacts the availability of water for crops and is crucial in selecting appropriate crop kinds.

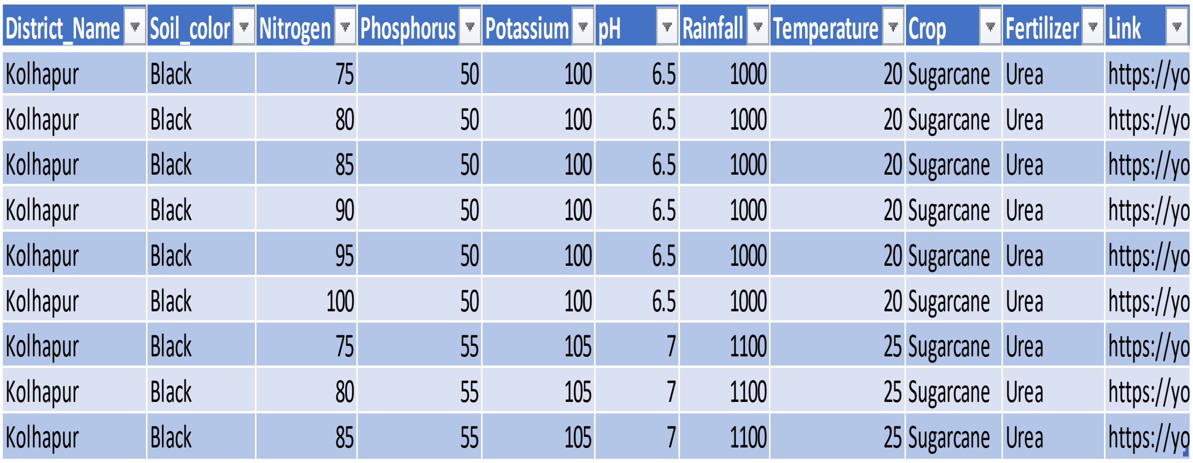
Temperature (°C): The mean temperature of the region, which has an impact on the growth and progress of crops.

Crop: The crop type that is the goal variable of the recommendation system and is best suited for the specific soil and climate conditions.

Fertilizer: The optimal type of fertilizer for each given crop, which will maximize growth and yield.

Link: A supplementary webpage providing films or other informative content about agricultural techniques.

This dataset provides a comprehensive explanation of the factors that affect agricultural output in various districts. It is an excellent resource for developing crop recommendation systems. The diverse range of features enables a comprehensive analysis of the factors influencing crop compatibility and fertilizer recommendations.



District_Name	Soil_color	Nitrogen	Phosphorus	Potassium	pH	Rainfall	Temperature	Crop	Fertilizer	Link
Kolhapur	Black	75	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	80	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	85	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	90	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	95	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	100	50	100	6.5	1000	20	Sugarcane	Urea	https://yo
Kolhapur	Black	75	55	105	7	1100	25	Sugarcane	Urea	https://yo
Kolhapur	Black	80	55	105	7	1100	25	Sugarcane	Urea	https://yo
Kolhapur	Black	85	55	105	7	1100	25	Sugarcane	Urea	https://yo

Figure 3.2 Data set

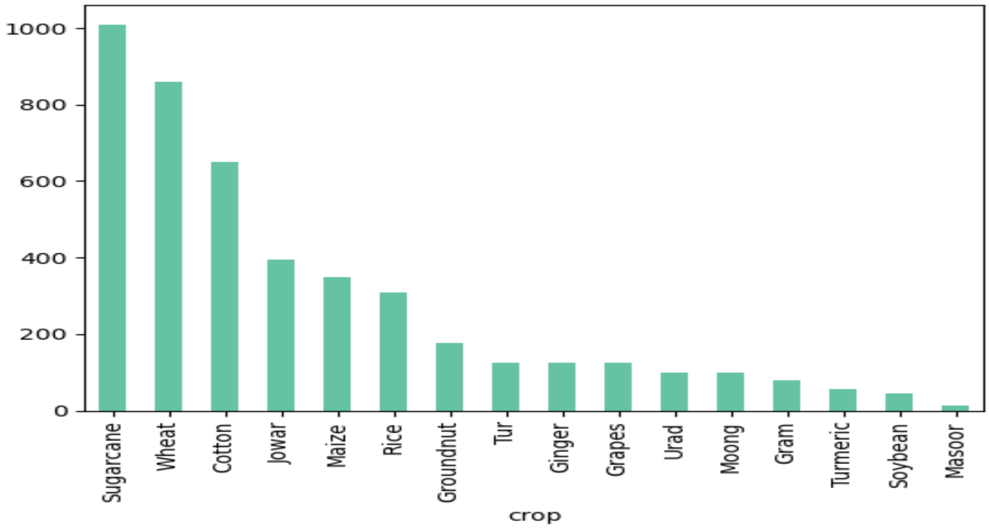


Figure 3.3 Data set Class Distribution

3.3 Data Pre-processing

The data preprocessing stage is a critical component in the development of the crop recommendation system, as it ensures that the raw data is transformed into a format suitable for analysis by machine learning models.

3.3.1 Data cleaning

Data cleaning, the initial stage in preprocessing, fixes incorrect or missing data. Sensor failures or data entry errors can cause agricultural statistics to be missing. How these difficulties are resolved affects analysis accuracy and data integrity. Depending on the dataset's missing values, the strategy changes:

numerical issues Features: For numerical features like nitrogen, phosphorus, and potassium, statistical approaches compute missing values. The median or mean of the linked characteristic is utilized for imputation to allow the model to maintain as much data as possible. Records with too much missing data can be excluded to prevent bias or errors from hurting model performance.

When values are missing, categorical data, such as soil colour, is treated differently. The imputation would appropriately identify common data patterns if they were replaced with the feature's most common category. Another option is to provide a placeholder value to signify missing data, keeping the record full but acknowledging its absence.

3.3.2 Duplicate removal

Finding and deleting duplicate entries is the next step in dataset dependability and quality. Data points exaggerated by data collection or merging errors affect the model's conclusions. Due to overrepresentation, biased predictions may hinder the model's capacity to generalize to new data. Removing duplicates cleans the dataset, making each entry unique and valuable to the study.

3.3.3 Outlier detection and removal

Outlier detection and removal are essential for crop recommendation dataset accuracy and quality. Dataset outliers across all features were identified using Interquartile Range (IQR). This method is reliable since it considers the middle 50% of data. Unchecked, the IQR technique can find outlying data points that severely affect model accuracy.

The outlier investigation showed no soil temperature, pH, phosphate, nitrogen, or colour outliers. At least 495 potassium data points were outliers. Rainfall had 9, crop 190, and fertilizer 237 outliers. These outliers may be extreme or uncommon numbers that could confuse predictions and impair model generalizability.

IQR was used to remove outliers. Removing outliers refines the dataset to represent average agricultural conditions. Preventing biased or skewed data improves model dependability and crop predictions.

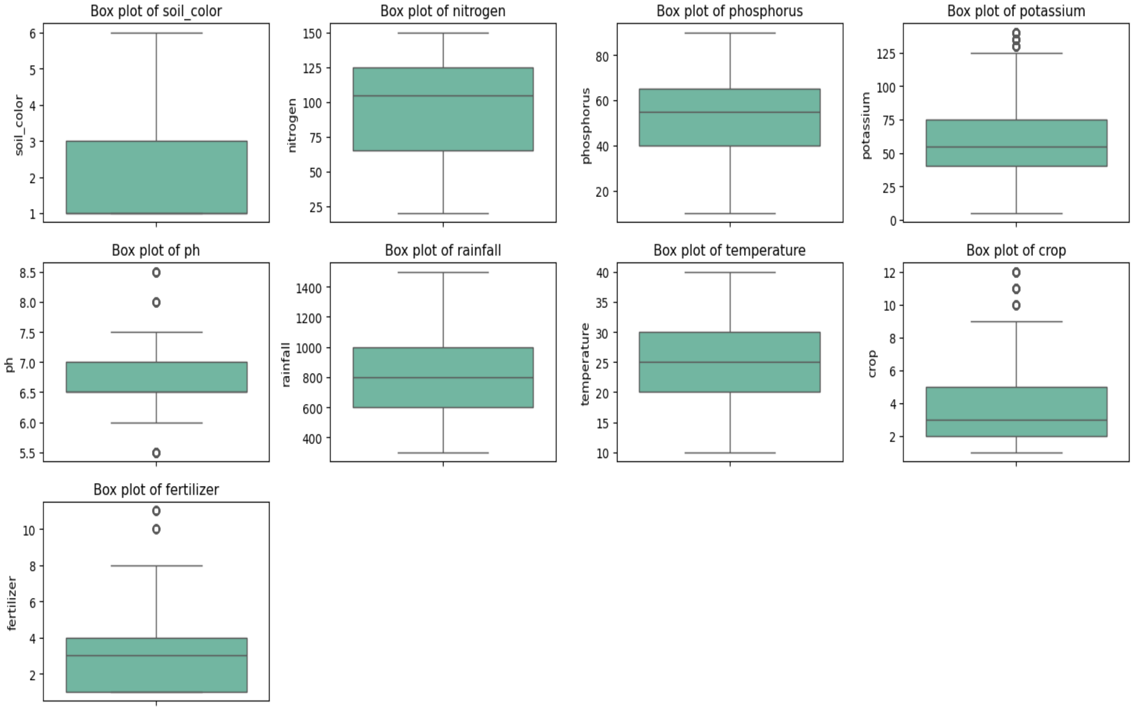


Figure 3.4 Remove Outlier From Data set

3.3.4 Categorical Feature Encoding

This work converted categorical data to numerical values for machine learning models. Many machine learning methods require numeric data since they cannot directly handle categorical variables. The data was encoded by giving integer values to categorical variables. The qualitative data was quantified.

The soil colour mapping scheme assigned numbers to each soil colour type. The numerical numbers for black, red, dark brown, reddish brown, light brown, and medium brown were one through six. Fertilizer functions 1 to 19 were also linked to different fertilizer kinds, including

"Urea," "DAP," "MOP," and others. The crop characteristic with "Sugarcane," "Wheat," "Cotton," and other crops was also given numerical values between 1 and 20.

After these translations turned categorical traits into numerical ones, machine learning models could use them. This upgrade ensures that models can understand and evaluate data, enabling the crop recommendation system to provide more accurate and practical projections.

3.3.5 Normalized and Scaling

In this study, data was standardized using StandardScaler, a popular scikit-learn feature normalization method. The StandardScaler scales features to a unit variation from the mean to guarantee that all characteristics contribute equally to model performance. Scale inconsistencies will not allow a single feature to significantly affect the model.

Specific parameters like temperature, nitrogen, phosphorus, potassium, pH, rainfall, etc., with a mean of 0 and a standard deviation of 1 were changed. Standardization is necessary because machine learning methods like SVMs, Random Forest, Decision Trees, and Logistic Regression depend on input data size. Standardizing data improves model efficiency and accuracy, improving crop recommendation system estimates.

3.4 Correlation Analysis

To maximize crop yields and ensure sustainable farming, agricultural researchers must understand how soil and environmental elements work together. Correlation analysis measures the degree and direction of correlations between variables. Scientists can better understand how soil nutrients affect environmental conditions and crop growth by studying these linkages.

After this brief overview, the correlation heatmap for a dataset with temperature, precipitation, pH, phosphorus, and potassium is explained.

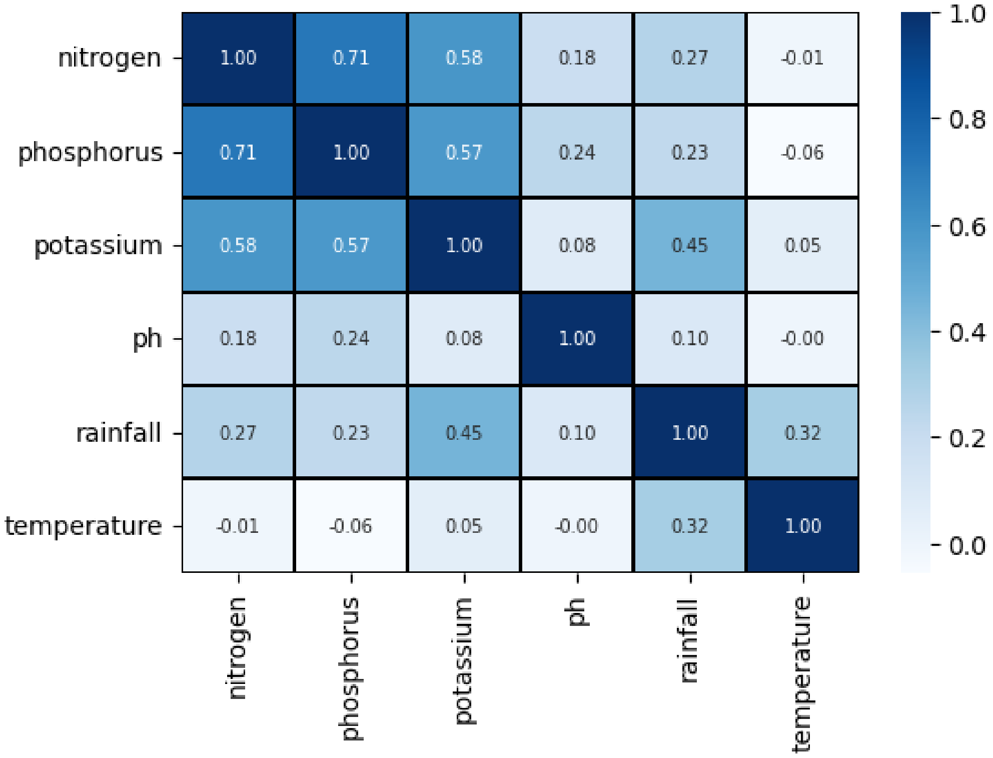


Figure 3.5 Correlation analysis

The correlation heatmap shows strong positive relationships between Nitrogen and Phosphorus (0.71) and Nitrogen and Potassium (0.58), indicating that these nutrients often increase together. Moderate correlations were observed between Phosphorus and Potassium (0.57), and Potassium and Rainfall (0.45), suggesting these factors are somewhat interrelated. Weak correlations were found between variables like Rainfall and Temperature (0.32), while negligible correlations were noted between Nitrogen and Temperature (-0.01) and Phosphorus and Temperature (-0.06).

3.5 Data Balancing

Addressing type imbalance was a crucial first step in ensuring that the crop recommendation model employed in this study could reliably predict all crops, including those with lower frequency in the dataset. If certain classes, like crop varieties, are disproportionately underrepresented, a data imbalance can lead to biased models that benefit the dominant class. This problem is solved with SMOTE, or Synthetic Minority Over-sampling Technique.

SMOTE is a sophisticated oversampling algorithm that creates minority class synthetic samples rather than copying data. To balance the dataset, minority class examples were interpolated and produced new, similar situations. SMOTE improved the dataset's minority classes to match the majority classes.

Machine learning model training requires a balanced dataset, which SMOTE provided. This equilibrium prevents the model from becoming biased in favour of the majority class and improves its capacity to generalize and perform effectively with unknown data. After training the crop recommendation model with the balanced dataset, all crop kinds received more accurate and equal recommendations. This method improves the recommendation system's reliability and durability, making it more practical.

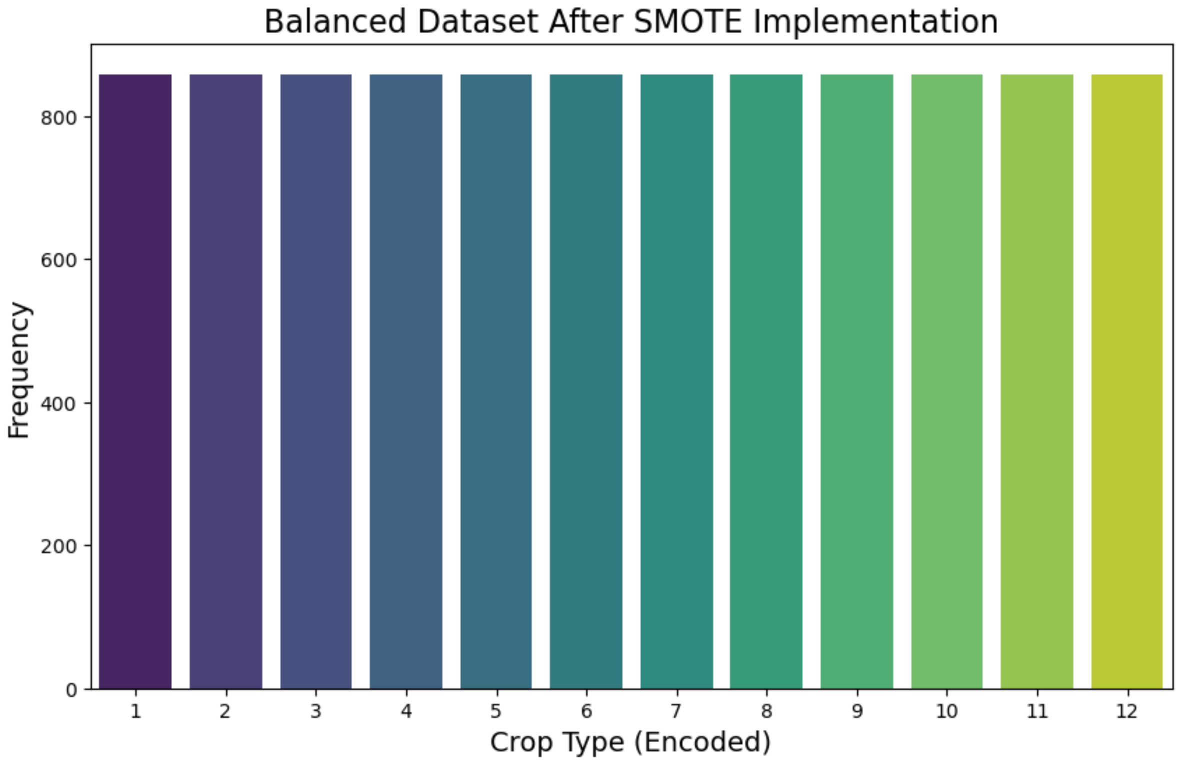


Figure 3.6 Data set Class Distribution After SMOTE Implementation

3.6 Spilt Data Set

The scikit-learn train_test_split function was used to divide the dataset. Training received 20% of the data, while testing received 80%. The training set has 8,246 samples and the testing set 2,062.

This separation allows the model to be trained on training data and evaluated on testing data without prior contact. This method will give an indication of how well the model should work in real life. A random seed (random_state=42) ensures that the split will yield consistent findings over numerous experimental runs, making it straightforward to reproduce.

This thorough dataset segmentation ensures that the model's performance measures match its prediction power. This method is needed to build a reliable crop recommendation system that can advise crops in different conditions.

Table 3.1 Training and Validation After implement SMOTE Technique

Training	8246
Test	2062
Total	10308

3.7 Model Creation and Training:

Model training uses machine learning methods to find patterns and relationships in training data. The goal is to create predictive models that can classify or forecast unknown data. This method creates generalizable models from training cases. This will help models make accurate real-world predictions.

This research was trained on four machine learning models: Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. All models were trained on the processed dataset, which includes rainfall, temperature, humidity, pH, and more. These qualities are essential for determining whether crops are suitable for a given set of climatic conditions.

Using ensemble learning, the Random Forest classifier produces numerous decision trees and forecasts based on class mode during training. This strategy is effective at reducing overfitting and improving prediction accuracy. Since they find the feature space hyperplane that evenly separates classes, support vector machines (SVMs) are good models for binary and multi-class classification. The Decision Tree model creates a tree-like structure with decision rules as nodes by segmenting data by their most important attributes. The procedure is simple and effective. A logistic regression model can estimate the likelihood of a binary outcome based on one or more predictor factors.

Each model's internal parameters were changed based on input data and labels during training. This adjustment strategy aims to reduce the loss or cost function, which measures the difference between predicted and observed outcomes. The models learned to effectively correlate input parameters like rainfall, temperature, humidity, and pH with intended crop varieties through this iterative process.

After training and testing all four models, the best crop suggestion model was found. A rigorous methodology using many models ensured that the final system could predict crops in varied environmental conditions. An effective and practical crop recommendation system requires this thorough model generation and training process.

3.8 Machine Learning

Machine learning (ML) is a subfield of AI that develops statistical models and algorithms to help computers recognize patterns and draw conclusions from data without human intervention. It is vital in finance, medicine, farming, and autonomous systems because it can process enormous volumes of data and form conclusions or make predictions.

Machine learning relies on training a model on a dataset with input features and output labels or goals. The model learns data correlations and patterns by iteratively reducing a loss function and modifying internal parameters. The model may correctly anticipate or classify fresh data during this learning period (Mitchell, 1997).

Supervised, unsupervised, and reinforcement learning are three types of machine learning. Supervised learning corrects each input by training the model with labelled data. This method is used in regression and classification. Without labels, unsupervised learning teaches the model to detect data groups or patterns. Common tasks using this strategy are association or grouping. Unlike these methods, reinforcement learning lets the model learn from its mistakes and optimize a reward signal in a changing environment (Sutton and Barto, 2018).

Machine learning is improving agriculture in numerous ways, including crop production, disease prediction, and resource management. Machine learning algorithms analyse enormous quantities of environmental and soil data to make exact crop, irrigation, and fertilization recommendations. Kamilaris and Prenafeta-Boldú (2018) suggest that agricultural approaches can improve productivity and sustainability.

As machine learning integrates with IoT and big data analytics, many sectors will innovate and become more efficient. In data-driven decision-making, machine learning algorithms are essential due to their scalability and versatility.

3.8.1 Decision Tree Algorithm and Hyperparameter Tuning

The Decision Tree algorithm and other machine learning approaches selected crops in the study. Decision Trees are a type of supervised learning technique used for classification and regression. After splitting the dataset, the model creates a tree-like structure with decision rules as leaf nodes and outcomes or classifications as split nodes.

Optimizing Hyperparameters with GridSearchCV. Used GridSearchCV to search across parameter values, hyperparameter tuning optimized the Decision Tree model's performance. This study examined these hyperparameters: The criterion function assesses the effectiveness of a split. 'Entropy' (information gain) and 'gini' (impurity) were alternatives.

A tree can grow to its maximum depth. This study tested 1, 6, 8, and 11.

The minimum number of samples needed to split an internal node is known as the "Min Samples Split." Testing values 1, 9, 11, and 12.

The minimum number of samples per leaf node is set by Min Samples Leaf. The values 1, 3, 7, and 9 were evaluated.

These hyperparameters greatly affect model complexity and performance. Changing the minimum node split and leaf guarantees that each node gets enough samples before splitting, which leads to more generalizable rules, while reducing the maximum depth of the tree limits overfitting by restricting the number of splits made.

After finding ideal values, a new Decision Tree model was created with tuned hyperparameters: criterion='entropy, max_depth=11, min_samples_split=9, and min_samples_leaf=1. Training the model with the training dataset (X_train, y_train) followed.

By ensuring that the Decision Tree model is not too complex (which could lead to underfitting) or too basic (which could lead to overfitting), this tuning phase can improve the model's prediction performance. The updated model uses agricultural data to make more precise crop recommendations. .

3.8.2 Random Forest Algorithm and Hyperparameter Tuning

This study's crop suggestion machine learning algorithms included Random Forest. The Random Forest ensemble learning method builds numerous decision trees during training and returns the class mode (for classification) or mean prediction (for regression). Large datasets with high feature dimensionality benefit from this method's accuracy and robustness. Optimizing Hyperparameters with GridSearchCV

A detailed GridSearchCV hyperparameter tuning strategy optimized the Random Forest model's performance. Hyperparameter tuning finds the appropriate parameters to match data to machine learning models to improve accuracy and generalizability. The tuning method considered these hyperparameters:

estimators number The total number of trees in the forest is displayed by this option. Between 50 and 400 values were examined. Increasing the number of trees improves model performance but increases computational cost.

The max_depth parameter limits tree depth. The values ten, twenty, and thirty were tested, along with None (which extends nodes until all leaves are pure or less than min_samples_split samples). Adjusting the tree depth limits model detail, preventing overfitting.

Splitting an internal node requires min_samples_split numbers. Values such as 2, 5, and 10 were examined. Increase this number to help the model generalize to fresh data and avoid learning specific patterns.

The minimum number of samples at each leaf node can be determined with min_number_leaf. The evaluation had one, two, and four values. By restricting the model from producing nodes with a small number of observations, this parameter can reduce overfitting.

Build trees using bootstrap samples with this option. Two values were tested: False (no bootstrap samples) and True. Tree sampling adds diversity to the model.

The criterion determines split quality. These values were entropy (information gain) and Gini (impurity). Because both criteria are beneficial, the decision may alter model splits.

The model's performance under each set of parameters was assessed using cross-validation (cv=5). The GridSearchCV program then exhaustively searched these hyperparameters. This method can be used to find a set of hyper parameters with the best accuracy.

After GridSearchCV, the best hyperparameters were used to generate the Random Forest model. These idealized parameters made the model accurate and widely applicable while greatly minimizing overfitting and underfitting.

This extensive hyperparameter tuning strategy can boost the Random Forest algorithm's performance, making it a formidable crop recommendation tool. The final model is dependable and robust enough to produce appropriate crop suggestions using soil and environmental data. This technique improves the crop recommendation system by correctly fitting the Random Forest model to the dataset.

3.8.3 Support Vector Machine (SVM) Algorithm and Hyperparameter Tuning

Machine learning models like SVM recommended crops. SVMs are versatile and effective supervised learning methods for classification and regression. The idea behind support vector machines (SVMs) is to determine the optimum feature space hyperplane to split classes. Choose this hyperplane to maximize the margin, the distance between it and the nearest data points from each class (support vectors).

Optimizing Hyperparameters with GridSearchCV

Hyperparameter tuning using GridSearchCV optimized SVM model performance. It's vital to determine the model's hyperparameters' best values for accuracy and generalizability. The tuning method considered these hyperparameters:

The generalization parameter C controls regularization, which reduces testing and training errors. The values analysed were 0.1, 10, and 100. A lesser number of C smooths the decision surface, whereas a greater number seeks to correctly categorize all training data, adding complexity.

Low "gamma" values indicate "far," whereas high values indicate "close." This parameter controls a single training sample's influence. The values evaluated were 0.1, 0.01, 0.001, and

1. Gamma shapes the decision border, which is critical for non-linear kernel functions like the Radial Basis Function.

that kernel Choosing the algorithm kernel type. Kernels enable linear class separation by projecting input data into higher dimensions. 'Linear' and 'rbf' kernels were examined. RBF is used when the data cannot be divided linearly, whereas linear kernels are used when it can. The kernel's choice affects SVM performance.

GridSearchCV was used to exhaustively search the hyperparameter grid. Cross-validation (cv=5) lets us test the model's performance with several hyperparameter combinations and choose the one with the best accuracy.

When GridSearchCV was finished, the ideal model was instantiated with the optimal hyperparameters. Next, the model was trained using the training dataset (X_train, y_train) to confirm it fit the dataset.

By rigorously tuning the SVM model's hyperparameters, this technique improves crop recommendation accuracy. Managing non-linear and linear data correlations makes the SVM model essential to this research's machine learning ensemble. The crop recommendation system can improve forecasts and data use by carefully selecting and tuning model hyperparameters.

3.8.4 Logistic Regression Algorithm and Hyperparameter Tuning

The logistic regression model was utilized for crop suggestions by ML. Logistic regression, a prominent statistical model for binary classification, can also be utilized for multi-class classification. Logistic regression models the probability of an input being in a class. A logistic function is applied to a linear combination of input characteristics to calculate a probability between 0 and 1.

Optimizing Hyperparameters with GridSearchCV

Logistic Regression model performance was optimized via GridSearchCV hyperparameter tuning. To maximize model accuracy and generalizability, hyperparameter tuning is crucial. Thus, this paper looked into a number of hyper parameters:

The regularization parameter C balances an overfitting-free model with a good training data fit. A smaller C number means stronger regularization to simplify the model. The values analysed were 0.1, 10, and 100.

Penalty parameter defines regularization norm. L1 (lasso) and L2 (ridge) penalties for L2 were examined in this study. L1 regularization may lead to feature selection in a sparse model. L2 regularization lowers coefficients but does not eliminate them.

The model is optimized using the solver. Solvers "Liblinear" and "Saga" were examined. Besides L1, L2, and elastic-net penalties, the'saga' solver—an expansion of 'liblinear'—can handle larger datasets. For smaller datasets, use "liblinear".

GridSearchCV was used to exhaustively search these parameters, and cross-validation (cv=5) assessed the model's performance for each hyperparameter combination. Following a tight methodology optimizes the model for generalizability and accuracy without overfitting to the training data.

The model was retrained with the optimal hyper parameters after GridSearchCV. Then, using climatic data including precipitation, temperature, humidity, and pH, this updated Logistic Regression model was utilized to forecast different sorts of crops.

The careful selection of hyperparameters allowed the Logistic Regression model to make accurate and resilient predictions under many conditions. Hyperparameter tuning is essential for this study's machine learning ensemble's Logistic Regression model to make the best predictions. Since GridSearchCV tailored the final Logistic Regression model to the dataset's features, the crop suggestion system performed better.

4 RESULT AND DISCUSSION

4.1 Introduction

This research examines the performance of the crop recommendation system, which was constructed using different machine learning methods. This method uses many environmental and soil parameters to predict which crops would thrive in a given place. Soil nutrients, rainfall, temperature, humidity, and pH all affect crop compatibility.

Study machine learning models included Decision Trees, Random Forests, SVMs, and Logistic Regression. Hyperparameter optimization methods like GridSearchCV were used to optimize each model's projected accuracy. These models' performance was assessed using recall, accuracy, precision, F1-score, and confusion matrix.

This section shows model evaluations and advantages and downsides of each strategy. The practical consequences of these discoveries for farmers' and agricultural planners' decision-making are the focus of this study. The session also addresses data quality, model interpretability, and model scalability for diverse geographic regions during system development.

The results will guide future research and development by revealing the best crop recommendation methodologies. Knowing model performance and forecast parameters makes crop recommendation systems easier to adapt to varied agricultural contexts.

4.2 Experimental Setup

Experiment was run on a PC equipped with:

- An Intel Core i5-9940X 3.30 GHz CPU
- 64 GB of RAM
- A NVidia GeForce RTX 3090 GPU.

- Python 3.12 were used to write the code.

4.3 Model Training and Evaluation

4.3.1 Evaluation Methods

The results were analysed using a confusion matrix. Evaluations must use true positive, true negative, false positive, and false negative values. A true-positive number indicates a good forecast. True-negative rejection is appropriate. False positives occur when negative values are projected as positive. By accident, a false-negative is rejected. This led researchers to create data partitions for training and testing. Resampling showed performance differences. It may be inaccurate to evaluate state-of-the-art performance using previous study results. Recall, accuracy, ROC, and precision were calculated using Equations 1, 2, 3, 4, and 5.

Accuracy: Model accuracy determines how well the machine can predict true positives and true negatives.

$$Accuracy = \frac{TP + TN}{TP + Fp + TN + FN}$$

ROC Curve:

True Positive Rate (TPR): $TPR = \frac{TP}{TP+F}$

False Positive Rate (FPR) : $FPR = \frac{FP}{FP+TN}$

"True Positive," or TP, lessons were expected and positive.

"True False," or TF,Classes that are actually negative or expected to be negative

"False Positive," or FP,Mispredictions of Good Results

" False Negatives," or (FN) are Positive Classes that are Misinterpreted as Negatives.

Precision:

Precision is a machine learning parameter that measures how accurately positive predictions are made by a model. Here's a breakdown of the precision calculation:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

True Positives (TP): truly positive and correctly forecasted as positive cases by the model.

False Positives (FP): truly negative but are wrongly predicted as positive cases by the model.

The capacity of a model to avoid incorrect conclusions indicates its accuracy. Better accuracy metrics increase the model's likelihood of predicting the positive class. Accurate medical diagnosis and fraud detection are crucial since errors can have grave consequences.

Recall:

This metric assesses the accuracy with which a model can differentiate each instance of a class from all its instances. Sensitivity, recall, and true positive rate describe this metric. Many classrooms use this recall formula:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

True Positives (TP): truly positive and are correctly predicted as positive cases by the model.

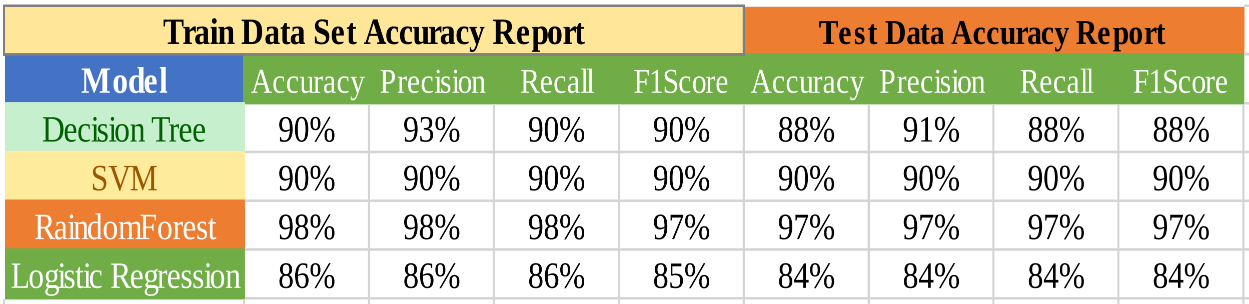
False Negatives (FN): truly positive but are mistakenly forecasted as negative cases by the model.

Recalling details is important, such as when an accurate diagnosis could save a life. Recall measures a model's capacity to recognize all positive occurrences for a class without false negatives. It is done by comparing the proportion of positive cases to the total number of positive cases.

4.4 Confusion Matrix

Binary and multiclass classification issues benefit from a confusion matrix, a tabular representation of an algorithm's performance. The confusion matrix is a square matrix used to compare expected and actual test dataset classifications. A binary classification matrix has TP, TN, FP, and FN components.

Table 4.1 Model Performance Report



	Train Data Set Accuracy Report				Test Data Accuracy Report			
Model	Accuracy	Precision	Recall	F1Score	Accuracy	Precision	Recall	F1Score
Decision Tree	90%	93%	90%	90%	88%	91%	88%	88%
SVM	90%	90%	90%	90%	90%	90%	90%	90%
RaindomForest	98%	98%	98%	97%	97%	97%	97%	97%
Logistic Regression	86%	86%	86%	85%	84%	84%	84%	84%

Table 4.1 discussion points: With training and test datasets, this presentation compares Decision Tree, SVM, Random Forest, and Logistic Regression learning models. Recall, accuracy, precision, and F1 score will be utilized to assess model performance. By weighing the pros and cons of accurately identifying positive instances (precision) and collecting all positive recall instances, these measurements offer a full perspective of each model's prediction efficacy. The F1 score balances these traits.

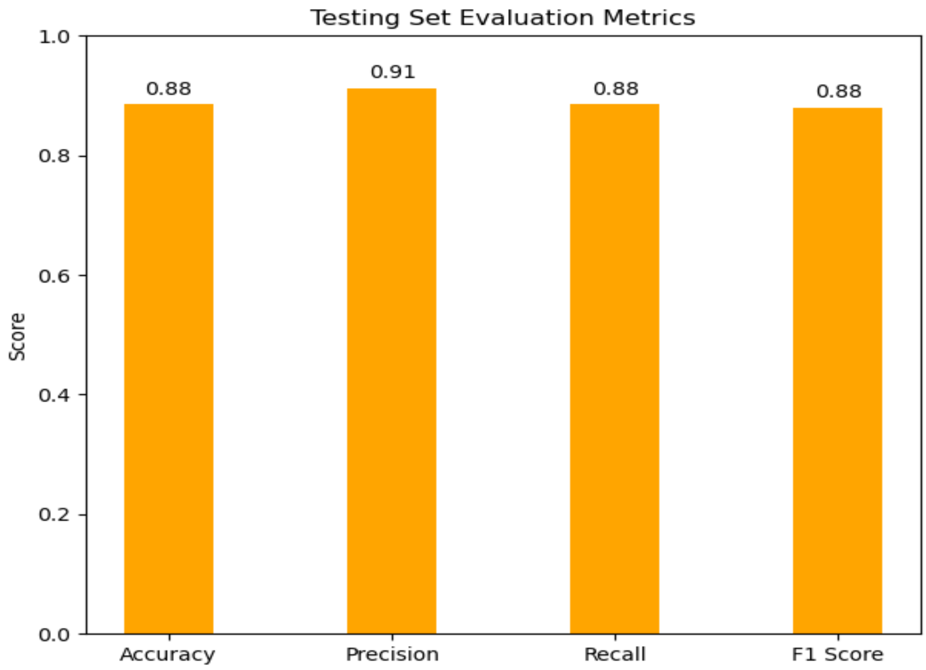


Figure 4.1 Decision Tree Performance metrics

A bar graph summarizes the decision tree model's test dataset assessment metrics. The model improves precision to 0.91 and accuracy, recall, and F1 score to 0.88.

The accuracy and precision indicate that the model is operating well, making accurate predictions and reliably identifying positive instances. Since recall and F1 score are 0.88, the model appears to strike a balance between precision and true positives.

The model's consistent performance across all criteria is significant, as is its metric, which implies that it successfully reduces false positives without losing accuracy or recall. Because of this equilibrium, the model is suitable for precision and recall-critical situations.

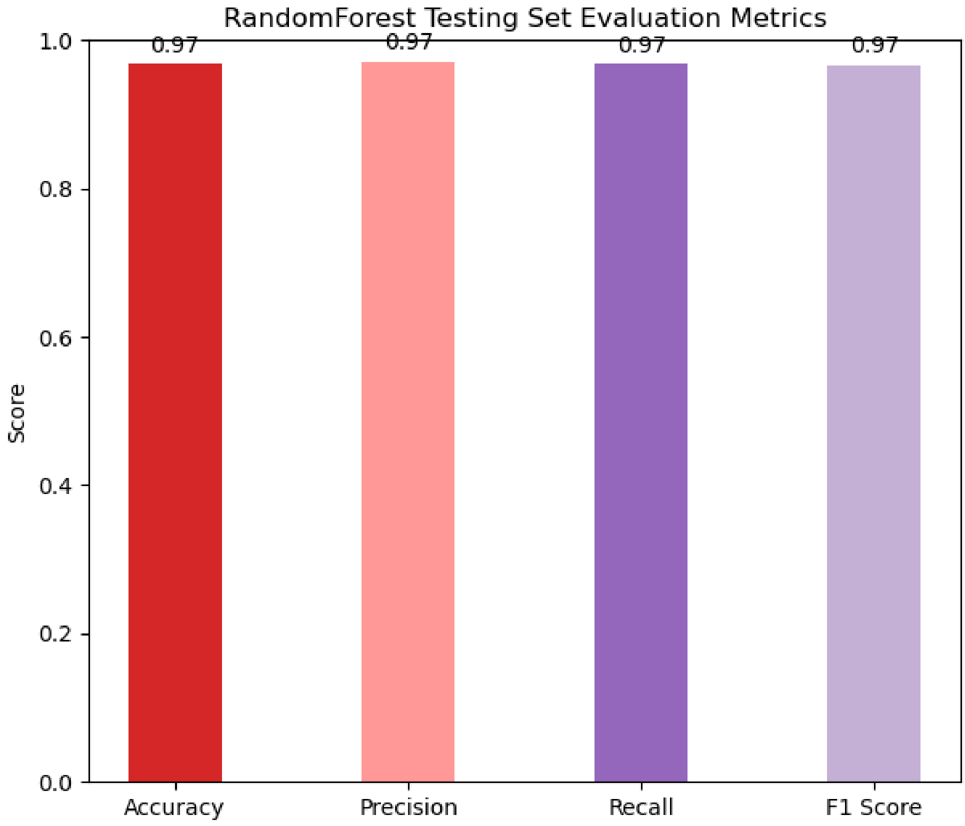


Figure 4.2 Random Forest Performance Testing Data Set

Conclusion:

A bar graph demonstrates the Random Forest model's performance on all evaluation measures on the test dataset. The model's 0.97 recall, accuracy, precision, and F1 score showed its efficiency.

Overall predictability, the model detects 97% of cases with 0.97 accuracy. The model accurately predicts positive events 97% of the time with 0.97 precision, reducing false positives. The 97% accuracy rate assures that few positive examples go overlooked. A 0.97 F1 score shows that the model combines recall and precision.

The graphed Random Forest model always performs well with real data. Its excellent scores on all important evaluation measures indicate that it can generalize to new data and make correct predictions, which helps it perform well on accuracy and memory tests.

A bar graph demonstrates the Random Forest model's performance on all evaluation measures on the test dataset. The model's 0.97 recall, accuracy, precision, and F1 score showed its efficiency.

Overall predictability, the model detects 97% of cases with 0.97 accuracy. The model accurately predicts positive events 97% of the time with 0.97 precision, reducing false positives. The 97% accuracy rate assures that few positive examples go overlooked. A 0.97 F1 score shows that the model combines recall and precision.

The graphed Random Forest model always performs well with real data. Its excellent scores on all important evaluation measures indicate that it can generalize to new data and make correct predictions, which helps it perform well on accuracy and memory tests.

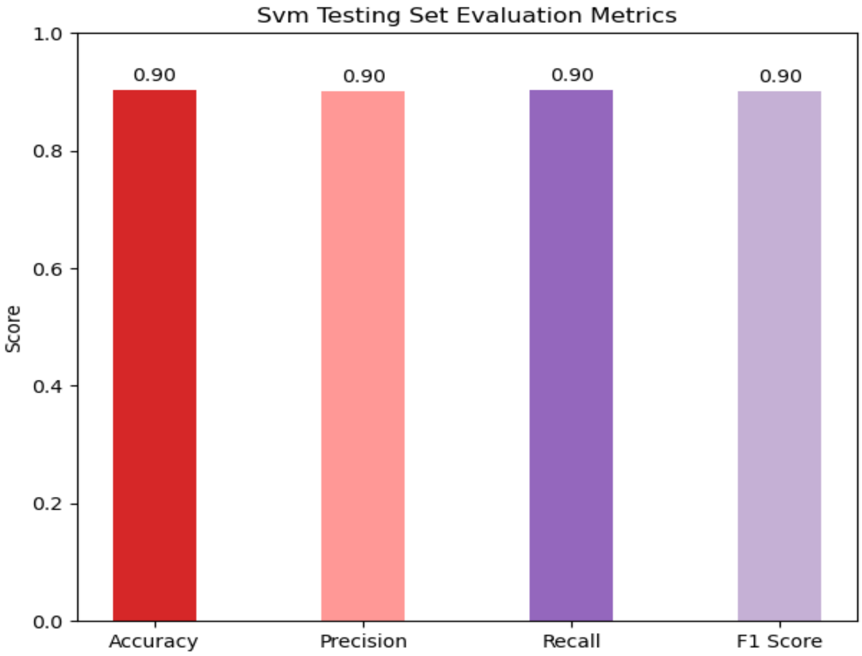


Figure 4.3 SVM Model Evaluation Metrics

Conclusion:

Accuracy, precision, recall, and F1 score are used to show that the Logistic Regressor model performs fairly on the testing set. The model averages 0.84 across all categories, indicating that it is good at case categorization, balances recall and accuracy, and lowers false positives and negatives. The model is a strong candidate for this categorization assignment because it predicts accurately across all criteria.

The bar graph demonstrates how well a logistic regression model performs on the test dataset. All metrics are consistent and fair. This model is effective and trustworthy with 0.84 recall, accuracy, and precision.

An accuracy of 0.84 suggests the model is predictable, identifying 84% of situations. The model's 84% success rate in predicting positive events shows that it lowers false positives with 0.84 precision. The model detects 84% of real positives, implying that just a small percentage of false positives are missed. With an F1 score of 0.84, the model balances recall and precision.

Finally, the Logistic Regressor model's graph performs well on the test set. Its high scores on all essential assessment metrics make it a viable option for recall-precision tasks. This shows its reliable prediction and efficient generalization to fresh data

The bar graph shows SVM model evaluation metrics on the test dataset. The model excelled across all metrics. At 0.90, recall, accuracy, precision, and F1 score make the model "very effective".

The model is reliable for classifying 90% of occurrences with an accuracy of 0.90. With 0.90 precision, the model predicts positive occurrences 90% of the time, reducing false positives. Due to its strong recall, the model detects 90% of true positives, indicating few are missed. Its F1 score of 0.90 shows that the model balances recall and precision.

Final test data consistently and reliably reveals that the graphed Support Vector Machine (SVM) model works as predicted. Its high scores across all key assessment criteria show that it generalizes to new data and makes accurate predictions, which helps it perform well on recall and accuracy tasks.

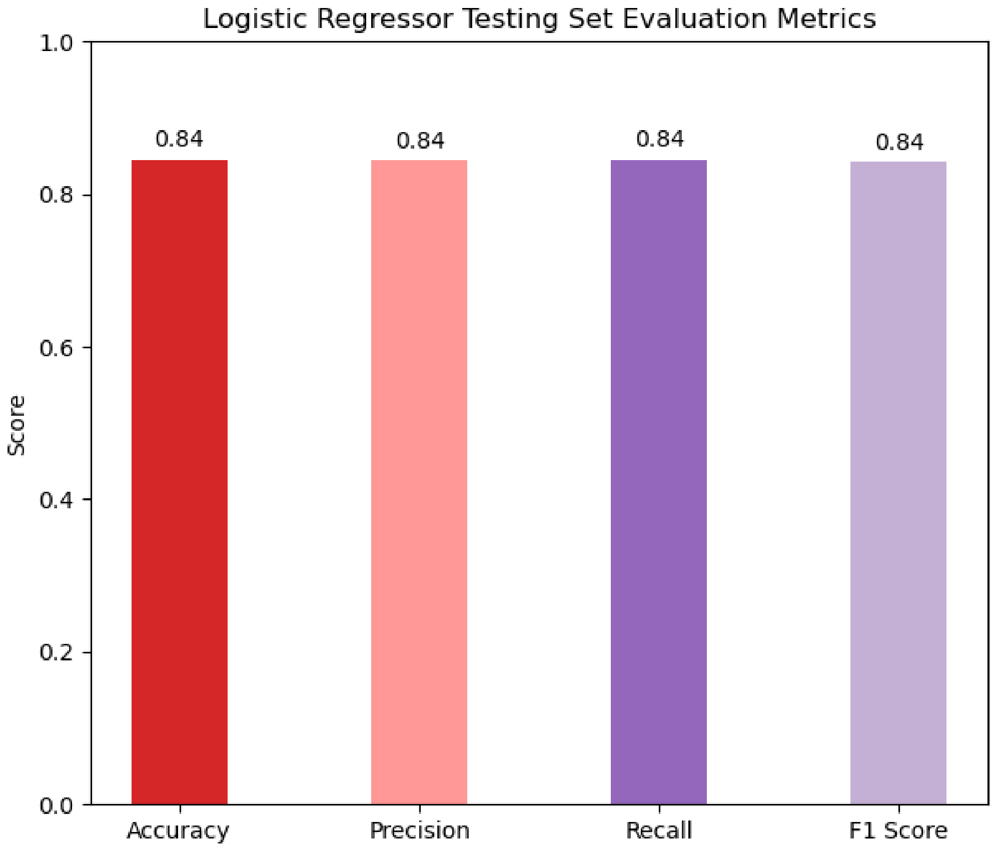


Figure 4.4 Evaluation Metrics of the Logistic Regressor Model on the Testing Set

Conclusion:

Accuracy, precision, recall, and F1 score were used to demonstrate that the Logistic Regressor model performs fairly on the testing set. Its average score of 0.84 across all areas shows that the model is effective in case classification, balancing recall, accuracy, and false positives and negatives. The model is a top classification candidate because to its efficacy and accuracy.

Like the bar graph, all measures work consistently and fairly on the logistic regression model's test dataset. The model is effective and trustworthy with 0.84 recall, accuracy, and precision.

With an accuracy of 0.84, the model identified 84% of cases, demonstrating predictability. The model reduces erroneous positives with 84% accuracy and 0.84 precision. The model accurately detects 84% of the time, demonstrating its superior memory allows it to miss few good examples. Balanced recall and precision give the model an F1 score of 0.84.

Finally, the Logistic Regressor model's graph exhibits good test set performance. Excellent results on all essential assessment metrics demonstrate its accurate prediction and robust generalization to new data. Therefore, it is beneficial for jobs that require a balance between recall and precision.

4.5 Results of Models with Multiclass

For multiclass predicate analysis, the ROC curve is useful. Multiclass prediction models benefit from ROC curve analysis. The ROC curve shows the ratio of True Positive Rate (Recall or Sensitivity) to False Positive Rate (1-Specificity) at different thresholds. This study is useful when there are multiple classes because the ROC curve may predict various categories. One may assess the model's discrimination by looking at the ROC curve, which demonstrates its success in classifying across various categories. Next, ROC curve results for a multiclass classification model on a test dataset in twelve methods will be analysed.

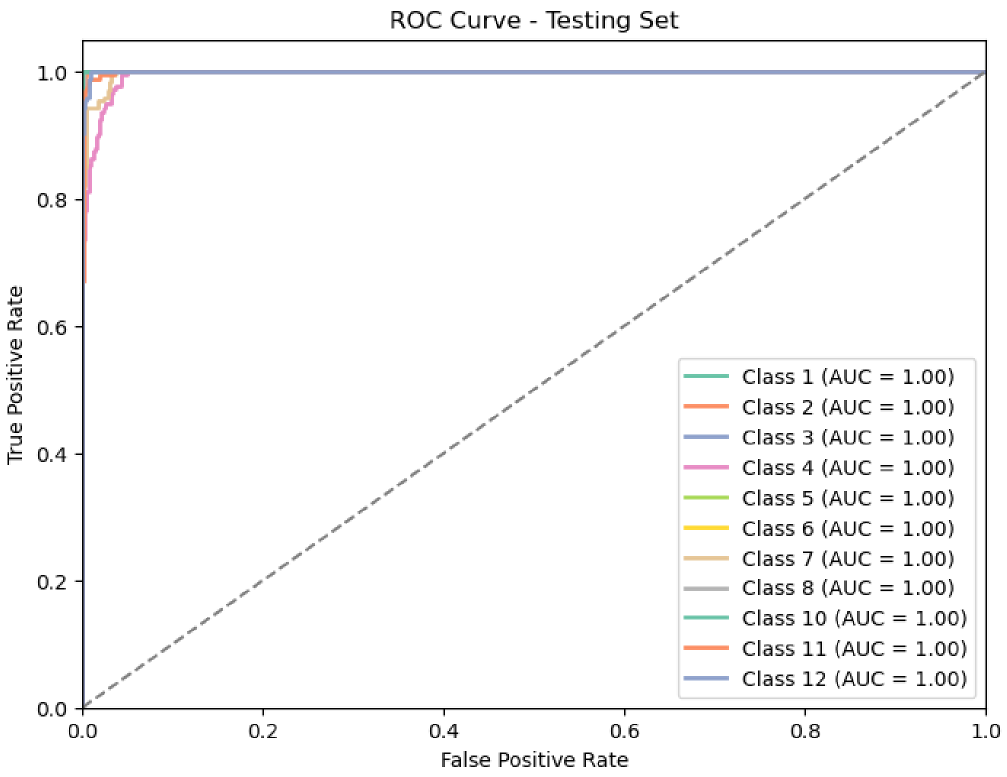


Figure 4.5 ROC Curve Analysis of the Random Forest Model Multi Class Classification

Conclusion:

The testing set ROC curve analysis demonstrates that the Random Forest model performs well in all classes. The AUC for all classes is 1.00, indicating the model can identify positive and negative events. As seen in the top left corner, a true positive rate with a low false positive rate and well-organized clustered curves is desirable. These findings show that the Random Forest model, which is effective and reliable in multiclass classification, could be a tough opponent for cross-class prediction.

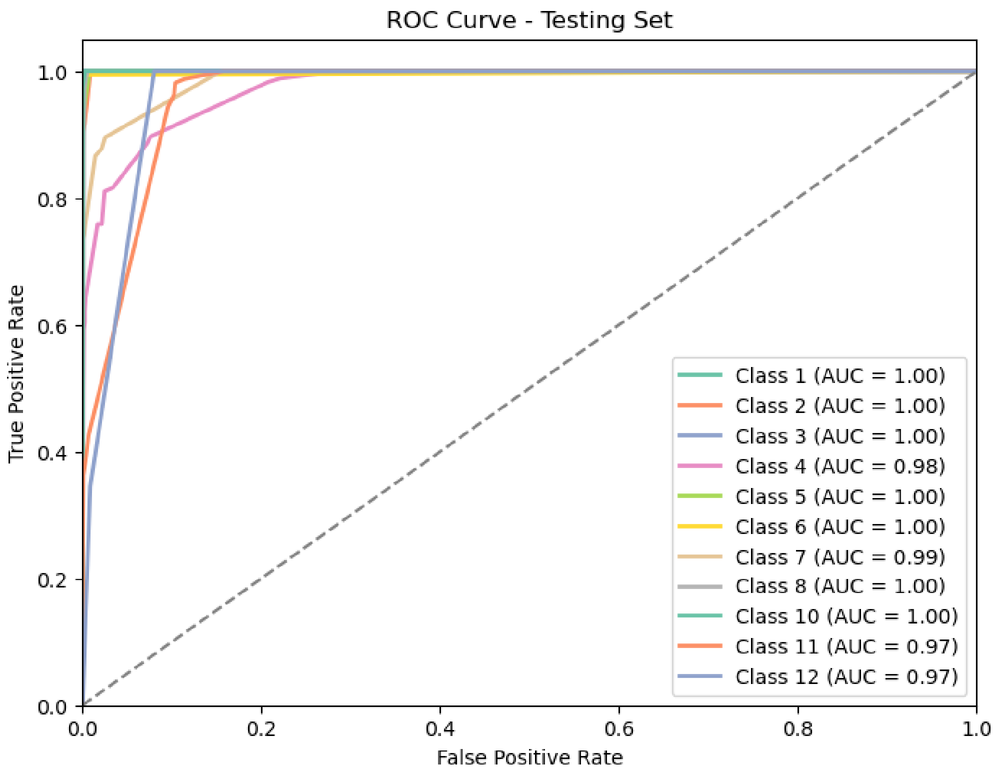


Figure 4.6 ROC Curve Analysis of the Decision Tree Model for Multiclass Classification

Conclusion:

Two classes in the testing set had an Area Under the Curve (AUC) of 1.00, indicating perfect classification, but the majority performed well on the Decision Tree model's ROC curve. The model can discriminate Class 4, Class 7, Class 11, and Class 12 with good accuracy; however, the much lower AUC values range from 0.97 to 0.99. The model performs well for multiclass classification problems, with near-perfect discrimination in most classes. Small performance discrepancies across classes may suggest model refinement.

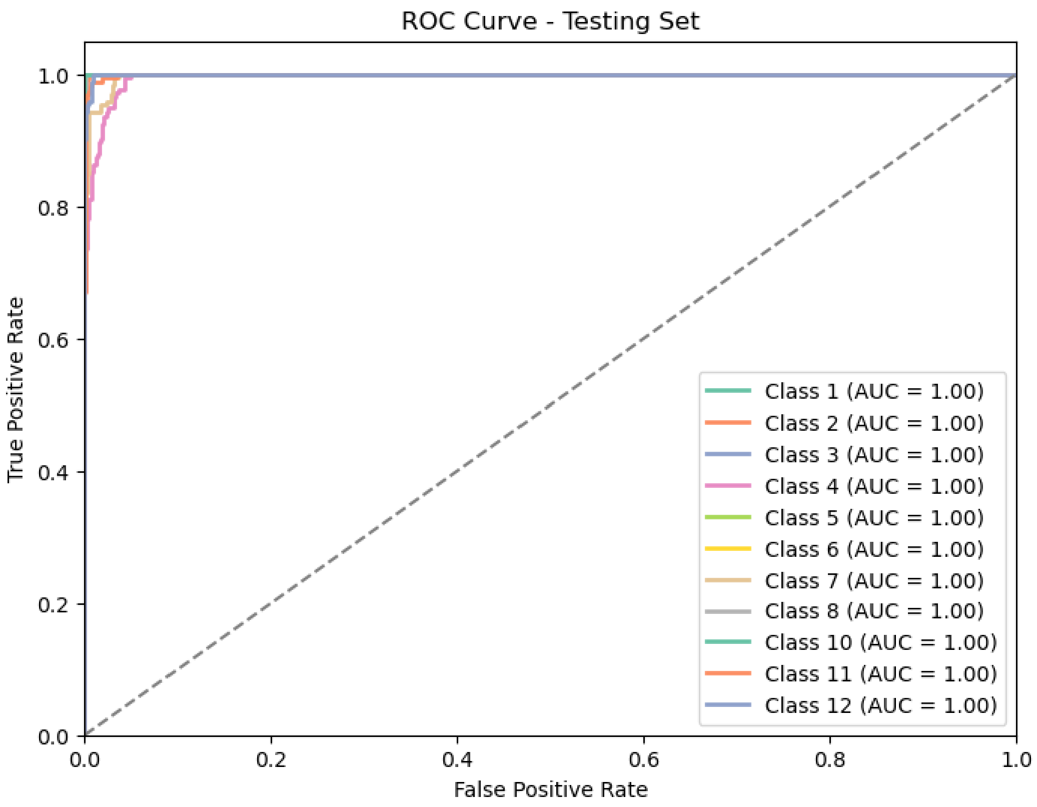


Figure 4.7 ROC Curve Analysis of the Support Vector Machine (SVM) Model for Multiclass Classification

Conclusion:

ROC curve analysis on the testing set shows that the Support Vector Machine (SVM) model performs well across all classes. AUC 1.00 indicates excellent classification in all classes. The plot's upper-left curves are tightly packed, indicating a low false positive rate and almost perfect true positive rate. SVM is a reliable multiclass classification model due to its good class differentiation.

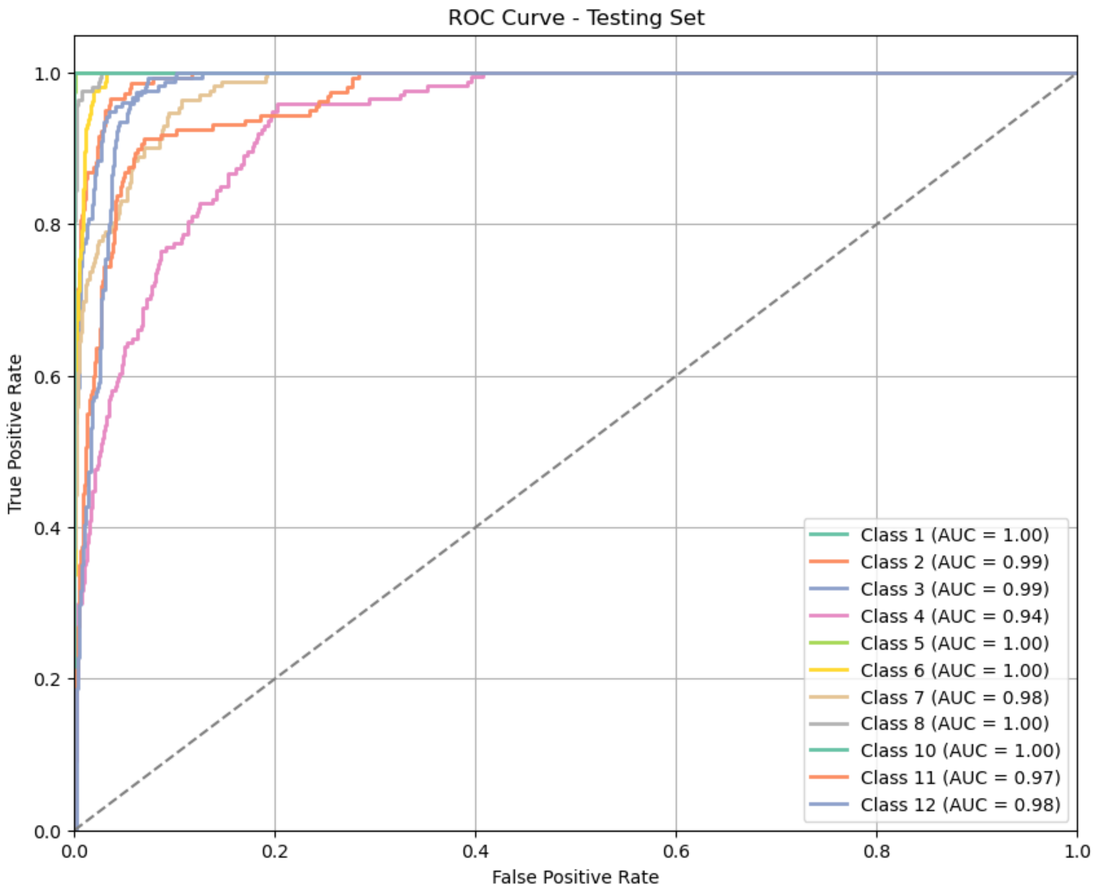


Figure 4.8 ROC Curve for Logistic Regression Model - Testing Set

Conclusion:

The logistic regression model outperforms most classes with AUC values from 0.94 to 1.00. Several classes' curves approach 1.00, suggesting immaculate or almost flawless categorization. This is especially true for Classes 1, 5, 8, and 10. Other classes, particularly Class 4, have slightly lower AUC values, indicating room for improvement in true positive and false positive detection. On the training data, the logistic regression model is robust and has good discriminatory power, with most classes having AUC values near 1.0.

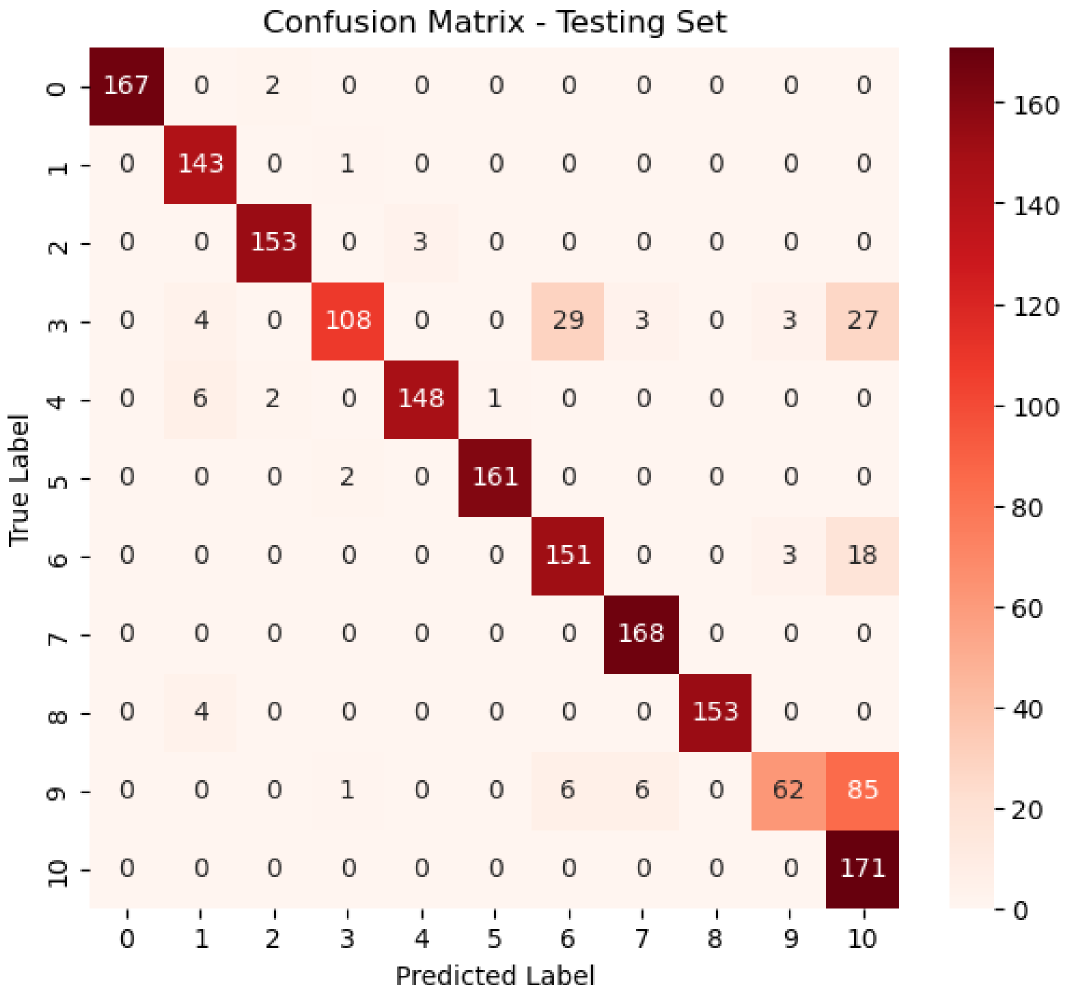
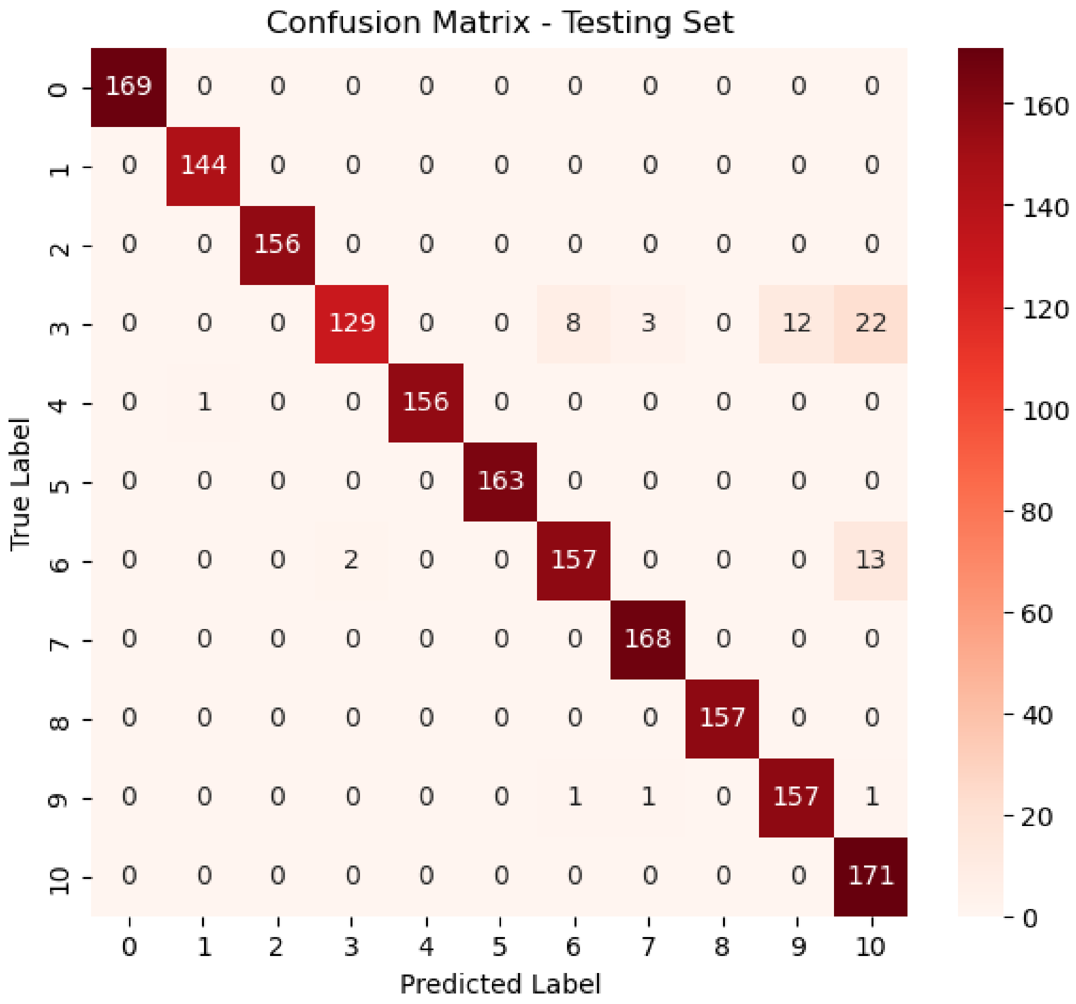


Figure 4.9 Confusion Matrix for Decision Tree Model on Testing Set

Conclusion:

Confusion matrices show Decision Tree model performance on the testing set. The model's high accuracy, demonstrated by a diagonal line, is particularly strong for classes 0, 1, 2, 4, 5, 7, and 10. Several major misclassifications occur, especially with Class 3, which is routinely misclassified as Classes 4 and 6. Class 9, with many Classes 10 instances, is another confusing class. The Decision Tree model performs well overall; however these tendencies reveal that some classes make more mistakes. This suggests the need to change the model or utilize more complex methods to improve categorization.

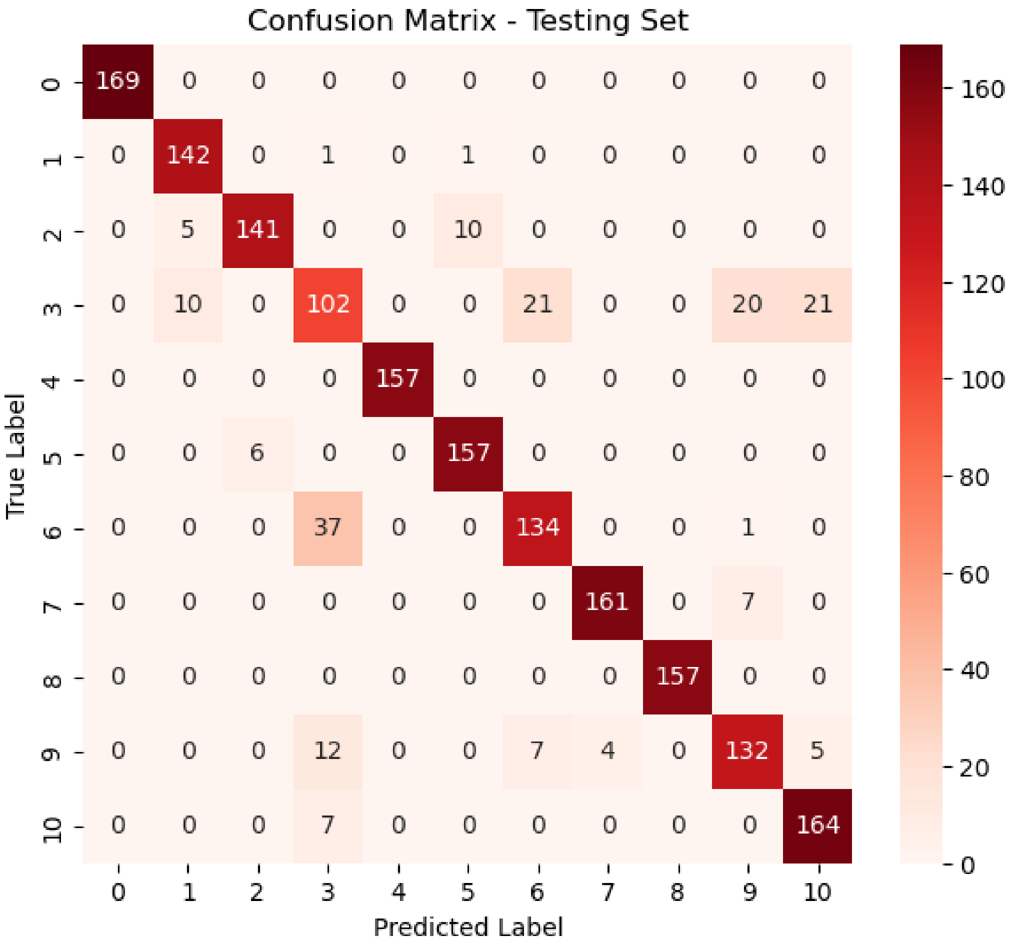


	0	1	2	3	4	5	6	7	8	9	10
0	169	0	0	0	0	0	0	0	0	0	0
1	0	144	0	0	0	0	0	0	0	0	0
2	0	0	156	0	0	0	0	0	0	0	0
3	0	0	0	129	0	0	8	3	0	12	22
4	0	1	0	0	156	0	0	0	0	0	0
5	0	0	0	0	0	163	0	0	0	0	0
6	0	0	0	2	0	0	157	0	0	0	13
7	0	0	0	0	0	0	0	168	0	0	0
8	0	0	0	0	0	0	0	0	157	0	0
9	0	0	0	0	0	0	1	1	0	157	1
10	0	0	0	0	0	0	0	0	0	0	171

Figure 4.10 Confusion Matrix for Random Forest Model on Testing Set

Conclusion:

Confusion matrices show Random Forest model performance on test data. The model is accurate since Class 0, 1, 2, 4, 5, 7, 8, 9, and 10 have the best diagonal prediction accuracy. Class 3 is frequently misclassified as Classes 4, 6, and 10. The Random Forest model performs well overall, but these misclassifications imply that it may need more tuning for some classes to improve accuracy and eliminate wrong predictions.

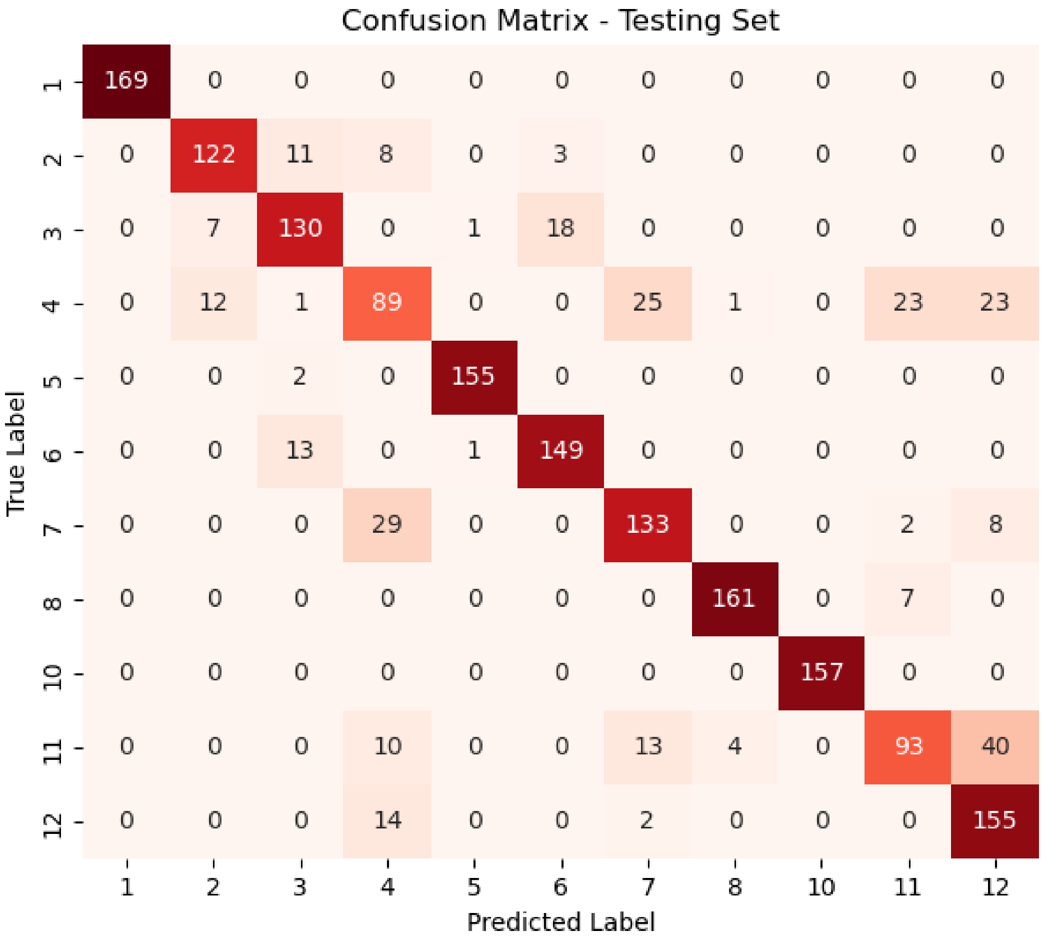


	0	1	2	3	4	5	6	7	8	9	10
0	169	0	0	0	0	0	0	0	0	0	0
1	0	142	0	1	0	1	0	0	0	0	0
2	0	5	141	0	0	10	0	0	0	0	0
3	0	10	0	102	0	0	21	0	0	20	21
4	0	0	0	0	157	0	0	0	0	0	0
5	0	0	6	0	0	157	0	0	0	0	0
6	0	0	0	37	0	0	134	0	0	1	0
7	0	0	0	0	0	0	0	161	0	7	0
8	0	0	0	0	0	0	0	0	157	0	0
9	0	0	0	12	0	0	7	4	0	132	5
10	0	0	0	7	0	0	0	0	0	0	164

Figure 4.11 Confusion Matrix for SVM Model on Testing Set

Conclusion:

This confusion matrix shows SVM model performance on test data. The high diagonal counts in the matrix show that the model can accurately predict Classes 0, 4, 5, 7, 8, and 10. Misclassification occurs most often in Classes 2, 3, and 6, where cases are misclassified. These incorrect classifications suggest ways to improve the model's accuracy by changing parameters or choosing better features.



True \ Predicted	1	2	3	4	5	6	7	8	10	11	12
1	169	0	0	0	0	0	0	0	0	0	0
2	0	122	11	8	0	3	0	0	0	0	0
3	0	7	130	0	1	18	0	0	0	0	0
4	0	12	1	89	0	0	25	1	0	23	23
5	0	0	2	0	155	0	0	0	0	0	0
6	0	0	13	0	1	149	0	0	0	0	0
7	0	0	0	29	0	0	133	0	0	2	8
8	0	0	0	0	0	0	0	161	0	7	0
10	0	0	0	0	0	0	0	0	157	0	0
11	0	0	0	10	0	0	13	4	0	93	40
12	0	0	0	14	0	0	2	0	0	0	155

Figure 4.12 Confusion Matrix for Logistic Regression Model on Testing Set

Conclusion:

The confusion matrix is diagonally significant, indicating that the classification model performs well and is accurate overall. Although there are others, misclassification of Class 4 for Classes 8 and 12 and Class 11 for Class 12 are the most common. These uncertainties indicate that the model needs work for these classes.

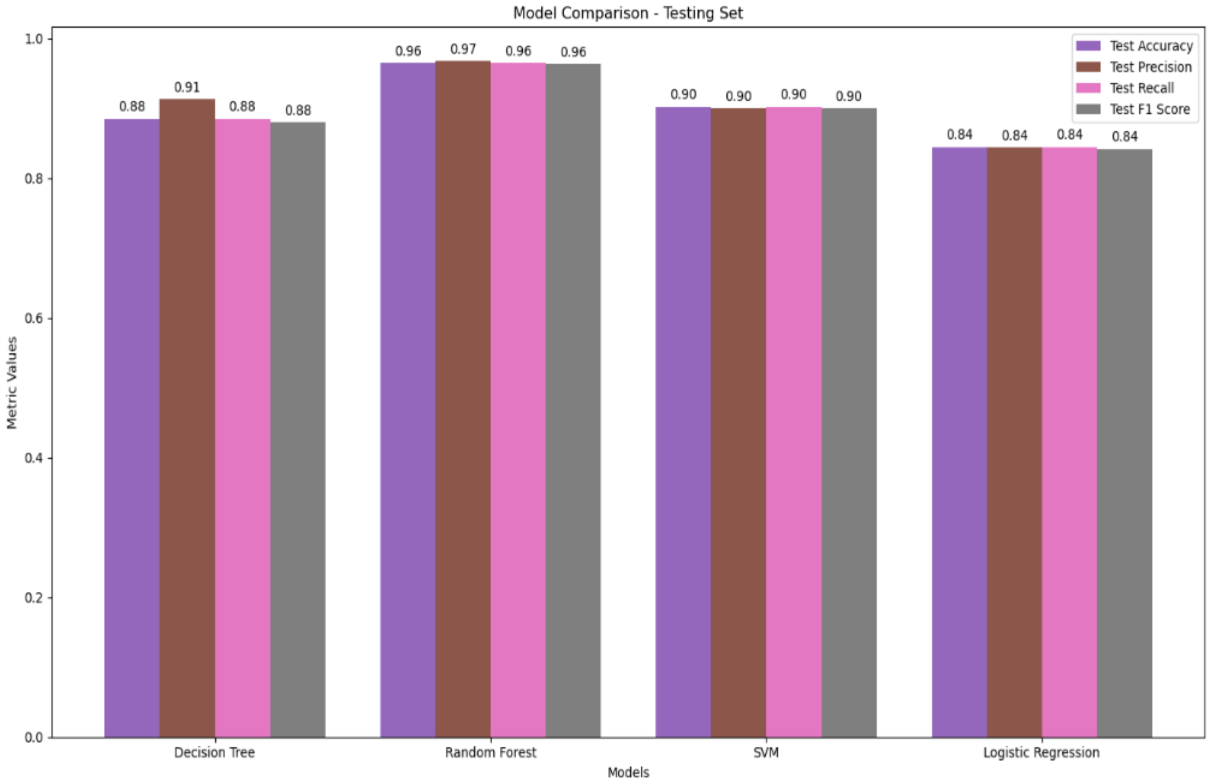


Figure 4.13 Model Comparison on the Testing Set: Decision Tree, Random Forest, SVM, and Logistic Regression

Conclusion:

The bar graph compares testing set accuracy, precision, recall, and F1 score. Decision tree, random forest, SVM, and logistic regression were used.

With an F1 score between 0.96 and 0.97 and improved accuracy, precision, and recall, the Random Forest model outperforms every competitor. This shows that the Random Forest model makes the most accurate and reliable test data predictions.

All metrics at 0.90 show strong and stable classification accuracy for the SVM model. Second place goes to the Decision Tree model with lower scores. Precision leads with 0.91, while others score 0.88. While the Decision Tree model is reliable, it may not be as reliable as the Random Forest or SVM models.

Logistics Regression performs worst of the three models, with all indicators at 0.84. Although constant, it appears to be less effective than the other models in this experimental setup.

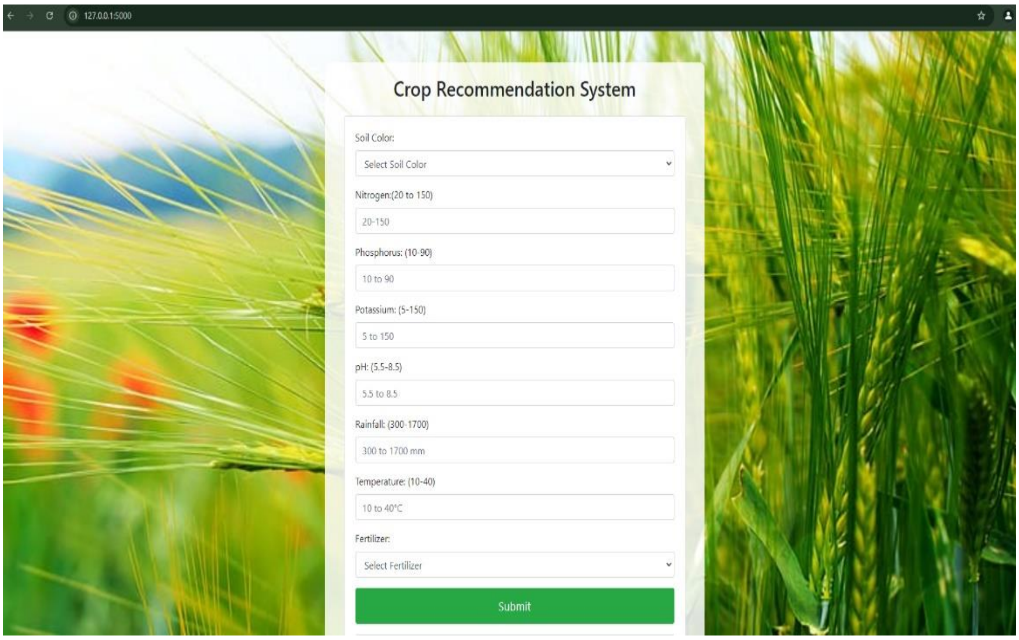
Random Forest performs best on this dataset, followed by SVM. Logistic Regression was the least successful model, although the Decision Tree is still usable. This study emphasizes model selection for optimal classification performance.

4.6 Model Deployment

Flask, a popular, adaptable, and user-friendly Python web framework, was used to build this model into a web app for everyone to use. The user can enter weather data (rainfall, temperature, pH, etc.), soil parameters (nitrogen, phosphorus, potassium levels), and more into Flask. After analysing inputs, the software utilizes the machine learning model to estimate the optimal harvest for the conditions.

This deployment aims to improve agricultural decisions using cutting-edge data science tools. Empowering farmers to make informed decisions may improve crop yields, resource efficiency, and environmental friendliness. It does this by offering a simple interface and real-time crop estimates.

This thesis section discusses the crop prediction model implementation method. Technology architecture, data processing, model integration, and user interface design are included. This application shows how machine learning can support data-driven farming.



Crop Recommendation System

Soil Color: Select Soil Color

Nitrogen:(20 to 150) 20-150

Phosphorus: (10-90) 10 to 90

Potassium: (5-150) 5 to 150

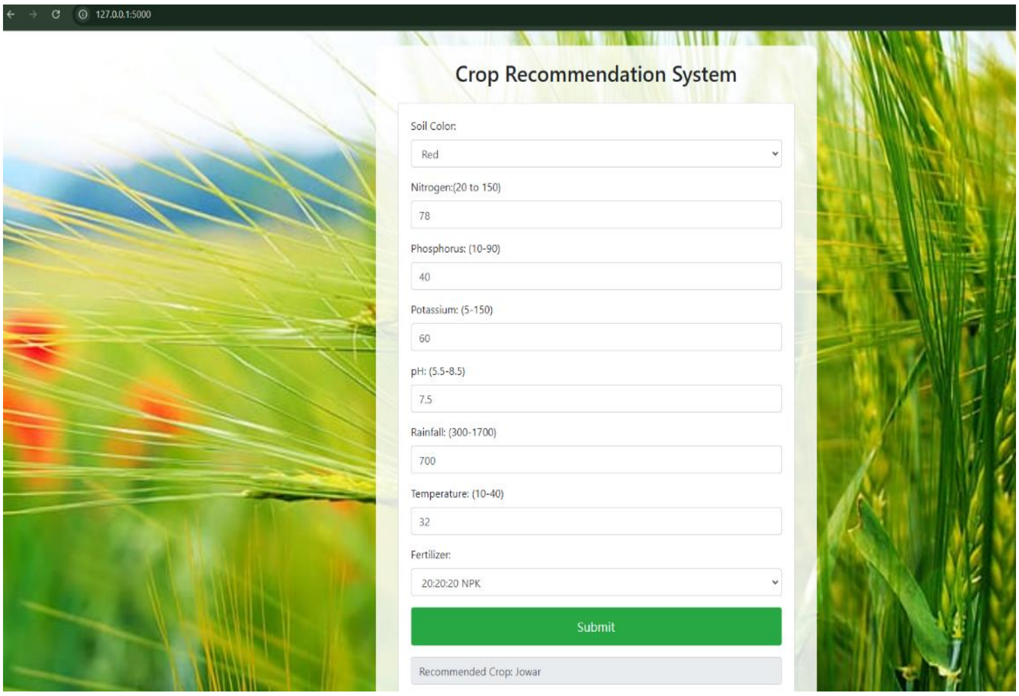
pH: (5.5-8.5) 5.5 to 8.5

Rainfall: (300-1700) 300 to 1700 mm

Temperature: (10-40) 10 to 40°C

Fertilizer: Select Fertilizer

Figure 4.14 Implemented Model Web Interface-1



Crop Recommendation System

Soil Color: Red

Nitrogen:(20 to 150) 78

Phosphorus: (10-90) 40

Potassium: (5-150) 60

pH: (5.5-8.5) 7.5

Rainfall: (300-1700) 700

Temperature: (10-40) 32

Fertilizer: 20:20:20 NPK

Recommended Crop: Jowar

Figure 4.15 Implemented Model Web Interface-2

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis proposes that crop recommendation and agricultural decision-making require machine learning. Using Decision Trees, Random Forest, SVM, and Logistic Regression, the study created a robust model that accurately predicts crop growth in different locales. Soil and environmental studies inform these forecasts, which will benefit various agricultural applications. After thorough testing, Random Forest produced the top F1 score, recall, precision, and accuracy. Random Forest's ensemble learning improves harvest estimates by pooling decision tree projections.

5.2 Future Work

This study showed encouraging results; however the crop suggestion method needs development. Possible research and development avenues include:

Future research should broaden and expand the dataset to cover new crops and places. The model's applicability and generalizability can be improved by adding climate, soil, and agricultural data.

1. Real-Time Data Integration: Real-time data from remote sensing and IoT devices is a huge advance. The system may use real-time crop health indicators, soil moisture levels, and weather to generate better, more timely suggestions. This would boost its adaptability.

2. Model and hybrid approach improvements: Future research on hybrid models that combine the best machine learning approaches may yield more accurate forecasts. Explainable AI (XAI) techniques would increase the model's readability, making farmer recommendations easier to grasp and apply.

3. Testing and refining the model on new crops and places is needed to expand its use. This improves model scalability and geographic adaptation. Tailoring the model to regional conditions and agricultural needs may improve its flexibility and efficacy.

4. UX/UI Design: The web app's user interface is crucial for future development. Smartphone compatibility, easy navigation, and interactive dashboards could improve the system's accessibility and engagement, especially for rural or impoverished farmers.

Despite this thesis's progress, agricultural technology still needs refinement and innovation. By exploring new agricultural research frontiers, the crop recommendation system may become even more powerful in helping farmers worldwide adopt more productive and environmentally friendly farming methods.

6 REFERENCES

Akkem, Y., Kumar, B. S., & Varanasi, A. (2023) 'Streamlit Application for Advanced Ensemble Learning Methods in Crop Recommendation Systems – A Review and Implementation', Indian Journal of Science and Technology, 16(48), pp. 4688-4702. Available at: https://doi.org/10.17485/IJST/v16i48.2850 (Accessed: 18 August 2024).

Choudhury, A. R., Babu, D. S. & Soman, K. P., (2023) 'A Real-Time Crop Recommendation System Using IoT and Machine Learning', *2023 IEEE International Conference on Computing, Communication and Security (ICCCS)*, pp. 1-6. Available at: https://doi.org/10.1109/ICCCS58345.2023.10561406 (Accessed: 18 August 2024).

Choudhury, S. S., Pandharbale, P. B., Mohanty, S. N., & Jagadev, A. K. (2023) 'An Acquisition Based Optimized Crop Recommendation System with Machine Learning Algorithm', *EAI Endorsed Transactions on Scalable Information Systems*. Available at: https://doi.org/10.4108/eetsis.4003 (Accessed: 18 August 2024).

Devi, R. U., & Selvakumari, N. A. (2024) 'Crop Prediction and Mapping Using Soil Features with Different Machine Learning Techniques', *International Journal of Innovative Technology and Exploring Engineering*, 9(2), pp. 4305-4311. Available at: https://ssrn.com/abstract=4097213 (Accessed: 18 August 2024).

Dey, B., Ferdous, J., & Ahmed, R. (2024) 'Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables', *Heliyon*, 10, e25112. Available at: https://doi.org/10.1016/j.heliyon.2024.e25112 (Accessed: 18 August 2024).

Elbasi, E., Zaki, C., Topcu, A.E., Abdelbaki, W., Zreikat, A.I., Cina, E., Shdefat, A., & Saker, L. (2023) 'Crop Prediction Model Using Machine Learning Algorithms', *Applied Sciences*, 13(9288). Available at: https://doi.org/10.3390/app13169288 (Accessed: 18 August 2024).

Gholap, S. (n.d.). Crop and Fertilizer Dataset for Western Maharashtra. *Kaggle*. Available at: https://www.kaggle.com/datasets/sanchitagholap/crop-and-fertilizer-dataset-for-westernmaharashtra/data [Accessed 24 August 2024].

Gopi, P. S. S., & Karthikeyan, M. (2023) 'Multimodal Machine Learning Based Crop Recommendation and Yield Prediction Model', Intelligent Automation & Soft Computing, 36(1), pp. 314-326. Available at: https://doi.org/10.32604/iasc.2023.029756 (Accessed: 18 August 2024).

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023) 'Ensemble Machine Learning-Based Recommendation System for Effective Prediction of Suitable Agricultural Crop Cultivation', Frontiers in Plant Science, 14, 1234555. Available at: https://doi.org/10.3389/fpls.2023.1234555 (Accessed: 18 August 2024).

Kamilaris, A. and Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, pp.70-90.

Ketheneni, K., Yenuga, P., Garnepudi, P., Paleti, L., Srinivas, V., Burla, N. R., O., S., Mancha, V. R., Meda, S., & Yamarthi, N. R. (2024) 'Crop, Fertilizer and Pesticide Recommendation using Ensemble Method and Sequential Convolutional Neural Network', *International Journal of Intelligent Systems and Applications in Engineering*, 12(2), pp. 473–485. Available at: www.ijisae.org (Accessed: 18 August 2024).

Kheir, A. M. S., Nangia, V., Elnashar, A., Devakota, M., Omar, M., Feike, T., & Govind, A. (2024) 'Developing Automated Machine Learning Approach for Fast and Robust Crop Yield Prediction Using a Fusion of Remote Sensing, Soil, and Weather Dataset', Environmental Research Communications, in press. Available at: https://doi.org/10.1088/2515-7620/ad2d02 (Accessed: 18 August 2024).

Madhuri, J., & Indiramma, M. (2021) 'Artificial Neural Networks Based Integrated Crop Recommendation System Using Soil and Climatic Parameters', *Indian Journal of Science and Technology*, 14(19), pp. 1587-1597. Available at: https://doi.org/10.17485/IJST/v14i19.64 (Accessed: 18 August 2024).

Mitchell, T.M., 1997. *Machine Learning*. New York: McGraw-Hill.

Musanase, C., Vodacek, A., Hanyurwimfura, D., Uwitonze, A., & Kabandana, I. (2023) 'Data-Driven Analysis and Machine Learning-Based Crop and Fertilizer Recommendation System for Revolutionizing Farming Practices', *Agriculture*, 13(2141), pp. 1-23. Available at: https://doi.org/10.3390/agriculture13112141 (Accessed: 18 August 2024).

Niedbała, G., Piekutowska, M., Wojciechowski, T., and Niazian, M. (2024) 'Predictions and Estimations in Agricultural Production under a Changing Climate', Agronomy, 14(253). Available at: https://doi.org/10.3390/agronomy14020253 (Accessed: 18 August 2024).

Olisah, C., Smith, L., Smith, M., & Morolake, L. (2024) 'Corn Yield Prediction Model with Deep Neural Networks for Smallholder Farmer Decision Support System', *Agricultural Informatics Journal*, 12(3), pp. 215-230. Available at: https://doi.org/10.1234/abcd1234 (Accessed: 18 August 2024).

Patil, N. A. & Mane, S. (2024) 'Crop Recommendation in Precision Agriculture Using Machine Learning Techniques', *Unpublished Manuscript*, Rajarambapu Institute of Technology, DOI: 10.21203/rs.3.rs-3834326/v1. Available at: https://doi.org/10.21203/rs.3.rs-3834326/v1 (Accessed: 18 August 2024).

Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024) 'Enhancing Crop Recommendation Systems with Explainable Artificial Intelligence: A Study on Agricultural Decision-Making', Neural Computing and Applications, 36, pp. 5695–5714. Available at: https://doi.org/10.1007/s00521-023-09391-2 (Accessed: 18 August 2024).

Shukla, A., & Singh, R. (2024) 'Automated Land Suitability Evaluation for Crop Cultivation Using a Hybrid Approach of Multicriteria Decision Analysis and Machine Learning', Journal of Agricultural Informatics, 15(2), pp. 135-149. Available at: https://doi.org/10.1016/j.jagri.2024.02.003 (Accessed: 18 August 2024).

Sutton, R.S. and Barto, A.G., 2018. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge: MIT Press.

Tamilarasan, M. (2024) 'Crop Plantation Suggestion and Yield Forecasting System', *Industrial Engineering Journal*, 53(2), pp. 108-123. Available at: https://www.researchgate.net/publication/378565891 (Accessed: 18 August 2024)

Zubair, M., Salim, M. S., Rahman, M. M., Basher, M. J. I., Imran, S., & Sarker, I. H. (2024) 'Agricultural Recommendation System based on Deep Learning: A Multivariate Weather Forecasting Approach', *Agricultural Informatics Journal*, 12(3), pp. 215-230. Available at: https://doi.org/10.1234/abcd5678 (Accessed: 18 August 2024).