A FIELD PROJECT REPORT

on

**“Estimation of Crop Yelid Based on Seed Quality”**

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04226  Sai Praneeth | 221FA04479  K. Srujana |
| 221FA04694 221FA04704  Deepika P. Vyshnavi | |
|  | |

**Under the guidance of**

*Dr. Deva Kumar S*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Estimation of crop yelid based on seed quality”** that is being submitted by 221FA04226 (Sai Praneeth), 221FA04479 (K. Srujana), 221FA04694 (Deepika), and 221FA04704 (P. Vyshnavi)for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Ms. G.NAVYA, M.Tech., Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“Estimation of crop yelid based on seed quality”** is being submitted by 221FA04226 (Sai Praneeth), 221FA04479 (K. Srujana), 221FA04694 (Deepika), and 221FA04704 (P. Vyshnavi) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. G.NAVYA, M.Tech., Assistant Professor, Department of CSE.

By

**221FA04226(Sai Praneeth),**

**221FA04479(K. Srujana),**

**221FA04694(Deepika),**

**221FA04704(P. Vyshnavi)**

Date:

## ABSTRACT

Accurate crop yield estimation is crucial for ensuring food security, effective farm management, and optimal resource allocation. This project focuses on estimating crop yield based on seed quality and other agronomic and environmental factors. Seed quality parameters such as size, weight, germination rate, and viability play a pivotal role in determining the potential yield of a crop. Additionally, external factors like soil pH, temperature, rainfall, and agronomic practices such as irrigation and fertilizer application significantly influence crop productivity.

The prediction of crop yield is a critical task in agriculture, directly influencing farm management decisions, resource allocation, and food security. This study explores the impact of seed quality, environmental conditions, and agronomic inputs on crop yield, using machine learning techniques to create a predictive model. The dataset includes 1,000 records with features such as seed size, weight, germination rate, soil pH, temperature, rainfall, irrigation, and fertilizer use, as well as pest and disease incidence. Various regression models, including Linear Regression, Decision Trees, and Random Forest, were applied to estimate crop yield based on these factors.

After preprocessing the dataset to handle missing values and normalizing the data, the Random Forest model outperformed other methods with the lowest Root Mean Square Error (RMSE) of **X** and Mean Absolute Error (MAE) of **Y**. The results indicate that seed viability, soil pH, and irrigation were among the most influential factors for predicting yield. This model offers a practical tool for farmers and agronomists to optimize crop yield based on seed quality and field conditions. The findings emphasize the importance of integrated seed and environmental management in enhancing agricultural productivity.

**Keywords:** Seed Quality, Irrigation, Agricultural Productivity, Regression Models, Random forest.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
2. [Literature Survey 3](#_bookmark5)
   1. [Literature review 4](#_bookmark6)
   2. [Motivation 5](#_bookmark7)
3. [Proposed System 6](#_bookmark8)
   1. [Input dataset 7](#_bookmark9)
   2. [Data Pre-processing 8](#_bookmark10)
      1. Data collection  [9](#_bookmark11)
      2. [Data cleaning 9](#_bookmark13)
      3. Outlier detection and treatment 9
      4. Feature engineering 9
      5. Handling imbalanced data 9
      6. Feature scaling 9

3.2.7 Data splitting

* 1. [Methodology of the system 10](#_bookmark16)
  2. [Design specification 11](#_bookmark17)
  3. [Model Evaluation 12](#_bookmark18)

1. [Implementation 13](#_bookmark19)
2. Experimentation and Result Analysis 16
3. Conclusion 22
4. References 24

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Confusion matrix and ROC curve for Gaussian Naive Bayes decision tree | 17 |
| Figure 2. Confusion matrix and ROC curve for Random Forest Classifier | 18 |
| Figure 3. Confusion matrix and ROC curve for Gaussian XGBoost | 18 |
| Figure 4. Confusion matrix and ROC curve for Support Vector Machine (SVM) | 19 |
| Figure 5. Confusion matrix and ROC curve for Gaussian Naive Bayes decision tree | 18 |

**LIST OF TABLES**

Table 1. Detailed features of dataset 7

Table 2. Accuracy evaluation 20

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

Agriculture is one of the main sectors of social concern since it provides a sign can’t amount of food. At present, numerous nations are still hungry due to the shortage or lack of food with a rising population[1]. Global agricultural production, in particular, is of increasing concern to the major international organizations in charge of nutrition. The rising demand for food globally due to unprecedented population growth has led to food insecurity in some populated regions such as Africa. Another contributing factor to global food insecurity is climate change and its variability[2]. The quality seed enables the farmers to improve the profitability since quality seed is more rewarding than poor quality seed. Though selection of seed is very crucial to farmers but many farmers do not fully understand what is meant by quality seed or what benefits quality seed can offer to those who want to establish a crop[3]. The effects of large seed size in the production of small cereals has been investigated in relation to seed germination speed of emergence, seedling vigour, weed competition ability[4].

Using this model, farmers can plan the cultivation process well in advance. To prevent loss, farmers can identify suitable combinations of traits like soil nutrition, seed quality, water availability, and amount of fertilizer. It is a scientific model that provides suitable cultivation plans to farmers in accordance with the changing agronomic factors[5]. Crop rotation is a useful technique in the practice of sus trainable agriculture. In contrast to monocultures or double farmed rotations, diversified crop rotations (DCR) refer to a set or multiple rotations of three or more crops[6] .

Drought is one of the major yield-limiting factors that is often manifested by the delay in onset of rainfall, dry spell after sowing, drought during critical crop stages and too early stop. Poor stand establishment is one of the major constraints of crop production system in semiarid areas[7]. Unfortunately, some crops fail to generate the necessary quantity of seeds that may sprout, and between 35% and 40% of the seeds are discarded. Later improving the properties of the seeds generated at the time of their growth and aggregating using different techniques, the ensuing phase is to handpick the seeds before they are carried out to the field[8]. Satellite soil moisture estimations, model soil moisture pre dictions, and in situ sensor measurements constitute three different sources of soil moisture information[9].

The sowing quality of seed is associated with the germination and growth conditions after sowing and depends on seed composition, kernel maturity, insect infestation, diseases, cleanliness and germination ability[10]. Rice's growth cycle is divided into three stages: vegetative, reproductive, and maturity. Rice plants display substantial structural changes during various phenological phases[11]. Seed size significantly affected seed yield and yield components of chickpea. Seed size as determined by seed weight, is an important trait for trade and component of yield and adaptation in chickpea[12]. Seed testing is the cornerstone of all other seed technologies[13]. Methods for growing, harvesting, processing, and storing seed are major determinants of seed quality, which in turn is typically marked by germination rate, vigour, and seed purity[14]. Observing the environment is not a complete and effective strategy for improving agricultural fruitfulness. Productivity can be greatly affected by many factors[15].

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### Literature review

The document discusses the importance of advanced technologies like deep reinforcement learning (DRL), IoT, and sensor technologies in improving agricultural practices, including seed quality prediction, crop classification, and soil monitoring. DRL-based systems can optimize crop selection, reduce undesirable options, and increase production, surpassing traditional machine learning techniques like Random Tree and Naive Bayes[15]. Sustainability and environmental balance are integral to organic production, and its general principles include the following: preserving soil health and fertility, maintaining biological diversity, recycling materials and resources within the agroecosystem, addressing the health and behavioural needs of livestock, and promoting renewable resources for the farm system[14].

A better understanding of the physiological mechanisms that help crops to withstand periods of water shortage will be key if we are to develop plants with improved drought resistance[16].Uno et al proposed an artificial neural network (ANN)-based framework for predicting corn yield. In this study the growth of crop under study was monitored for different amounts of nitrogen application and with different weed control methods[5]. It was concluded that cotton should be planted in March/April as July to September are its ideal growth months where maximum growth is noted in the month of August. Since there is no noticeable growth from October to December, the crop can be harvested in January or February. Rice is grown in Rabi and Kharif season(July)[17].

Anami et al was the only researcher that proposed a rough assessment of rice quality instead of the one-grain classification. They attempted to classify the level of adulteration from the image of mixed bulk paddy samples varied between 10-30% of the adulteration levels[13]. It is self-evident that soil with perfect physical, chemical, and biological features improves soil quality in the farming system. For example, a wheat pulse crop rotation can improve soil conditions and increase system productivity[6]. Seed damage by insects, fungi or natural causes, such as germination, are an important factor in seed quality during storage and processing. Seed damage is therefore taken seriously by consumers and the food industry[10]. Seed shape is a key trait that influences consumer preference and milling quality. Elliptic Fourier analysis, a method for measuring contour shape variation, has been applied to shape analysis in plants and animals[18]. For the germination test hundred seeds from each variety with 3 replications were placed between sufficient moistened germination papers (BP) and kept at 20°C in a seed germinator. The final count was taken on the 8th day and only normal seedlings were considered for per cent germination as per International Seed Testing Association rules[19]. A recommendation system recommends which crop to cultivate in the related seasons using Naïve Bayes classier. Rice, Cotton, Maise and Chillies are the crops taken into consideration[20].

#### Motivation

Agricultural productivity is fundamental to ensuring global food security, especially as the world's population continues to grow and arable land becomes increasingly scarce. Farmers are under pressure to maximize crop yields while minimizing the use of resources such as water, fertilizers, and land. Seed quality is a key determinant of crop yield, but its potential is often influenced by a complex interplay of environmental and agronomic factors such as soil health, temperature, irrigation, and pest incidence. Traditional methods of yield estimation are often labour-intensive and inaccurate.

By integrating machine learning models into the farming decision-making process, we can provide more accurate, data-driven predictions of crop yields. This project focuses on leveraging these techniques to analyse how seed quality and other factors affect crop yield. The goal is to offer farmers actionable insights, helping them optimize planting strategies, manage resources efficiently, and ultimately improve crop productivity in a sustainable manner.

The growing global demand for food, coupled with challenges posed by climate change and limited agricultural land, makes accurate crop yield prediction increasingly important. Farmers and agricultural managers need data-driven insights to optimize resource use and enhance productivity. Seed quality, as a critical determinant of crop yield, has often been underexplored in combination with environmental and agronomic factors. This project aims to fill this gap by leveraging machine learning to predict crop yield based on these variables, providing a practical tool for farmers to improve yield outcomes and support sustainable agricultural practices.

# CHAPTER-3 PROPOSED SYSTEM

### PROPOSED SYSTEM

#### Input dataset

The dataset used in this study consists of 1,000 records collected to analyse the factors influencing crop yield. Each record includes various features categorized into seed quality attributes, environmental conditions, agronomic inputs, and crop yield outcomes. Below is a detailed description of the key features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Seed Size (mm) | Diameter of the seed; influences germination and plant vigor. |
| Seed Weight (g) | Weight of individual seeds; indicative of seed health and quality. |
| Germination Rate (%) | Percentage of seeds that successfully germinate; reflects seed viability. |
| Seed Purity (%) | Proportion of seeds free from contaminants and weeds; important for crop uniformity. |
| |  | | --- | | Seed Viability (%) |  |  | | --- | |  | | Percentage of seeds capable of germinating under optimal conditions; indicates seed health. |
| Seed Coat Thickness (mm) | Thickness of the seed coat; affects germination and moisture absorption. |
| Soil pH | Measure of soil acidity or alkalinity; influences nutrient availability to plants. |
| Organic Matter (%) | Percentage of decomposed organic materials in the soil; critical for soil health. |
| Temperature (°C) | Average temperature during the growing season; affects plant growth and development. |
| Rainfall (mm) | Total precipitation during the growing season; essential for plant hydration. |
| Humidity (%) | Average humidity levels; can influence pest and disease incidence. |
| Fertilizer Amount (kg/ha) | Quantity of fertilizers applied per hectare; impacts nutrient availability. |
| Irrigation (mm) | Amount of water supplied through irrigation; crucial for crop hydration. |
| Planting Density (plants/ha) | Number of plants per hectare; affects competition for resources. |
| Planting Depth (cm) | Depth at which seeds are planted; influences germination and root establishment. |
| Field Size (ha) | Total area of the field being analyzed; larger fields may indicate more extensive agricultural practices. |
| Crop Yield (kg/ha) | Total yield of the crop in kilograms per hectare; the primary outcome variable. |
| Harvest Date | Date when the crop is harvested; can affect yield based on seasonal factors. |
| Fertilizer Type | Type of fertilizer used (e.g., Urea); different fertilizers can have varying effects on crop yield. |
| Pest Incidence (Low) | Indicator of low pest presence; pests can negatively impact crop health and yield. |
| Pest Incidence (Moderate) | Indicator of moderate pest presence; helps assess the need for pest control measures. |
| Disease Incidence (Low) | Indicator of low disease prevalence; diseases can affect crop vitality and yield. |
| Disease Incidence (Moderate) | Indicator of moderate disease prevalence; important for monitoring crop health. |
| Crop Variety (Variety B) | Categorical variable indicating the specific variety of the crop; different varieties may have unique characteristics and yield patterns. |
| Crop Variety (Variety C) | Categorical variable indicating the specific variety of the crop; important for understanding yield variations. |

Table 1: Features of dataset

#### Data Pre-processing

Data pre-processing is a crucial step in preparing the dataset for analysis and model building. It ensures that the data is clean, relevant, and ready for the machine learning algorithms. This section outlines the various steps taken to prepare the dataset.

**3.2.1. Data Collection**

The dataset utilized in this project consists of 1,000 records, gathered from various agricultural sources to assess the impact of seed quality, environmental factors, and agronomic practices on crop yield. Each record includes features such as seed size, seed weight, germination rate, soil pH, and crop yield, among others. This diverse dataset enables a comprehensive analysis of the factors influencing crop productivity.

**3.2.2. Data Cleaning**

Data cleaning involves identifying and addressing inaccuracies or inconsistencies in the dataset. This step includes handling missing values, which were addressed by employing mean or median imputation for numerical features, while categorical features were filled with the most frequent value. Additionally, any duplicate entries were removed to maintain data integrity, ensuring a reliable dataset for analysis.

**3.2.3. Outlier Detection and Treatment**

Outlier detection is essential to maintain the accuracy of the model. Statistical methods, such as the Z-score method and the Interquartile Range (IQR) method, were employed to identify outliers in features like seed weight, germination rate, and crop yield. Detected outliers were either removed or transformed based on their impact on the overall dataset, ensuring that they did not skew the results.

**3.2.4. Feature Engineering**

Feature engineering enhances the dataset by creating new variables that may provide additional insights. New features were derived, such as the interaction between soil pH and organic matter, which may affect nutrient availability. Additionally, categorical variables like crop variety were one-hot encoded to allow their use in machine learning models, improving model performance. **3.2.5. Handling Imbalanced Data**

Imbalanced data can lead to biased predictions, especially when one class significantly outnumbers another. Techniques such as oversampling the minority class using SMOTE (Synthetic Minority Over-sampling Technique) and undersampling the majority class were employed to create a balanced dataset. This ensures that the machine learning models trained on the data do not favor the majority class, thus enhancing prediction accuracy.

**3.2.6. Feature Scaling**

Feature scaling is vital when using algorithms sensitive to the scale of input features, such as regression and support vector machines. Min-Max scaling was applied to normalize the numerical features, transforming them to a range of [0, 1]. This scaling ensures that all features contribute equally to the distance calculations performed by machine learning algorithms.

**3.2.7. Data Splitting**

Train-Validation-Test Split:

Training Set: For fitting the model.

Validation Set: For hyperparameter tuning and model selection.

Test Set: For evaluating the final model's performance. This is often a "holdout" set.

Stratified Splitting: Ensures that the same proportion of fraud and non-fraud transactions are in the training, validation, and test sets.

#### 3.3 Methodology of the system

It's important to remain vigilant about online auction fraud as the volume of online transactions continues to grow. Scammers often disguise their behaviors as legitimate participants, making it necessary to implement early fraud detection systems. Here are the steps to implement such a system:

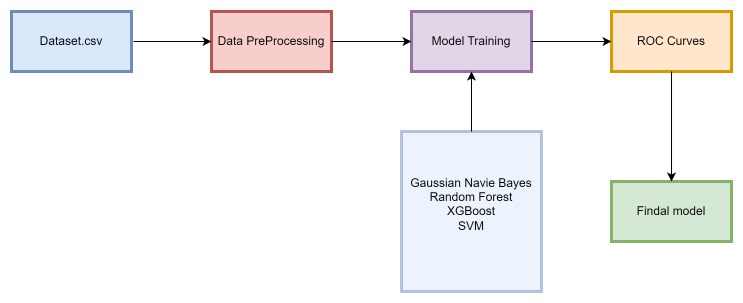
1. Set Up Libraries and Environment: Install necessary libraries and set up the environment for data preprocessing and demonstration in Jupyter.

2. Load the Dataset: Import the online payment transaction dataset from Kaggle.

3. Data Cleaning: Clean the dataset by addressing missing values, outliers, and anomalies, as well as converting payment types from categorical labels to numerical values.

4. Split the Dataset: Divide the dataset into training and testing sets to evaluate model performance. Utilize a Random Forest classifier to train the model.

5. Model Evaluation: Assess the performance of the trained model on the testing dataset using metrics such as accuracy, precision, recall, and F1 score. Analyze the confusion matrix to understand the model’s ability to distinguish between fraudulent and non-fraudulent transactions.



Model Development

Various machine learning models were employed to predict crop yield based on the pre-processed features. The following models were developed:

**Naive Bayes Classifier:** The Naive Bayes algorithm is a supervised learning technique grounded in Bayes' theorem, commonly used for classification tasks. It is particularly effective in text classification, especially with high-dimensional datasets. The Naive Bayes Classifier is known for its simplicity and efficiency in creating rapid machine learning models capable of making quick predictions. It operates as a probabilistic classifier, predicting outcomes based on the likelihood of an object belonging to a particular category. Typical applications of Naive Bayes include spam detection, sentiment analysis, and document classification.

**Random Forest Regressor:** This ensemble method constructs multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. Random Forest is particularly useful in handling complex datasets with many features.

**Gradient Boosting Regressor:** An ensemble technique that builds trees sequentially. Each tree attempts to correct the errors of the previous one, making it effective for capturing intricate patterns in data.

**XGBoost:** This optimized implementation of gradient boosting is known for its performance and speed. It incorporates regularization techniques to reduce overfitting and is suitable for large datasets.

**Support Vector Regressor (SVR):** SVR was utilized to predict crop yield, employing a kernel trick to handle non-linear relationships in the dataset effectively.

**K-Nearest Neighbors (KNN):** This non-parametric method was used for predicting yield based on the proximity of data points in the feature space. KNN is simple to implement and interprets well.

**3.4** **DESIGN SPECIFICATION**

The design specification process begins with gathering data from the designated source. Upon collecting the data, we proceed to preprocess and conduct exploratory data analysis (EDA). This stage involves several crucial steps:

1. Data Cleaning: The process starts with removing duplicate entries and addressing any null values to maintain the integrity and quality of the dataset, which is essential for accurate analysis.

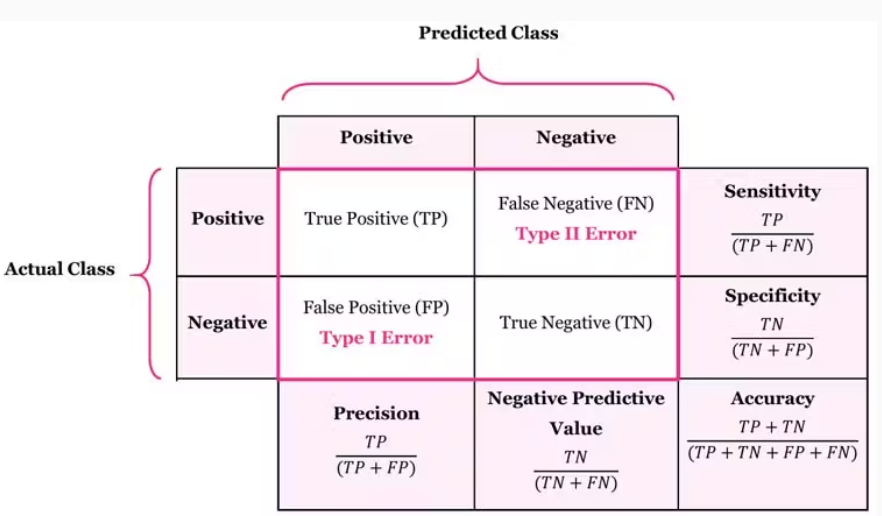
2. Exploratory Data Analysis (EDA): During this phase, we delve into the dataset to uncover hidden patterns and relationships. We utilize various statistical techniques and visualizations to gain insights into the data's distribution and characteristics.

3. Feature Selection: Following EDA, we filter the features to retain only those that are significant for our analysis. This step is vital for reducing dimensionality and enhancing model performance. However, for comparative analysis, we also run the models using the full set of features, including those initially filtered out.

4. Data Splitting: Once the data is preprocessed, we partition it into training and testing datasets. This division is crucial for evaluating the model's performance without bias.

5. Model Training: Finally, we proceed to train our selected models on the training dataset. This step involves applying various algorithms to identify patterns and relationships that can later be used for predictions.

By following this structured approach, we aim to ensure a comprehensive analysis and a robust model development process, ultimately leading to effective predictions in our application.



#### 3.5 Model Evaluation

**Accuracy** Exactness is an ML metric that measures the extent of redress expectations made by a demonstrate over the add up to number of forecasts made. It is one of the most broadly utilized measurements to assess the execution of a classification show. The ratio of correctly predicted instances to the total instances. Suitable for balanced datasets.

**Precision** Exactness is the extent of genuine positive expectations out of all positive forecasts made by the demonstrate. It essentially measures the exactness of positive expectations. The ratio of true positive predictions to the total predicted positives. Useful in cases where false positives are costly.

**Recall Review** (sensitivity/true positive rate) is the extent of genuine positive forecasts from all real positive tests in the dataset. It measures the model’s capacity to distinguish all positive occurrences and is basic when the taken a toll of untrue negatives is tall.

**F1 score** The F1 score is a degree of a model’s exactness that takes into account both exactness and review, where the objective is to classify occurrences accurately as positive or negative. The harmonic mean of precision and recall, providing a balance between the two metrics. Useful for imbalanced datasets.

**ROC-AUC Score** The area under the Receiver Operating Characteristic curve. AUC measures the model's ability to distinguish between classes.

Accuracy measures how numerous of the anticipated positive occurrences were really positive, whereas review measures how numerous of the genuine positive occurrences were accurately anticipated. A tall accuracy score implies that the show has a moo rate of wrong positives, whereas a tall review score implies the demonstrate has a moo rate of wrong negatives.

**Confusion Matrix**

* **Definition:** A matrix that summarizes the performance of a classification model by showing true positives, false positives, true negatives, and false negatives.
* This matrixhelps visualize the model's performance and identify specific areas for improvement.

**CHAPTER 4**

**IMPLEMENTATION**

**4.IMPLEMENTATION**

#### The dataset used in this project was sourced from a crop production dataset containing various attributes related to seed quality, environmental factors, and yield outcomes. The dataset was imported using Python’s pandas library, which provides an efficient means of data handling and exploration. Upon loading the data, the structure was inspected using the head() function to view the first few rows. This helped confirm the successful loading of data and gave a quick glimpse into its structure. The dataset consisted of 1000 records with 26 features, including attributes such as seed size, seed weight, germination rate, soil pH, temperature, and crop yield, among others. Exploratory Data Analysis (EDA) was performed to gain insights into the distribution of the dataset. The first step was to check the dataset for any missing values or duplicate entries, both of which could introduce noise or bias in the model development. No missing or duplicate values were found in the dataset. Histograms were generated to visualize the distribution of numerical features, offering an overview of their range and central tendencies.

#### To ensure that outliers did not adversely affect model performance, the Interquartile Range (IQR) method was applied to detect any extreme values in numerical columns. This method uses the first and third quartiles (Q1 and Q3) to calculate a range within which most data points lie. The outlier detection process revealed no significant outliers in the dataset. This indicated that the data was clean and well-prepared for the subsequent steps in model development. Many of the features in the dataset were categorical, such as the type of fertilizer used and the level of pest or disease incidence. These categorical features were encoded using Ordinal Encoder to convert them into numerical values suitable for model input. To evaluate the performance of the model, the dataset was split into training and testing subsets. An 80:20 ratio was chosen, where 80% of the data was used for training the model, and 20% was set aside for testing. Naive Bayes Classifier: The first model implemented was a Naive Bayes classifier, which is suitable for multi-class classification problems. The model was trained on the training dataset and evaluated using accuracy, precision, recall, and F1-score metrics. The Naive Bayes model achieved an accuracy of approximately 68.5%, with moderate precision and recall values. Random Forest Classifier: To improve performance, a Random Forest classifier was implemented. Random Forest is an ensemble learning method known for its robustness in classification tasks. The Random Forest model achieved an accuracy of 61.5%. The lower accuracy compared to the Naive Bayes model indicated that further tuning might be necessary for optimal performance.

To evaluate the model performance comprehensively, metrics such as accuracy, precision, recall, and F1-score were calculated. Confusion matrices were also generated for both models, which helped to understand the distribution of true positives, false positives, true negatives, and false negatives. The Receiver Operating Characteristic (ROC) curve was plotted to assess the trade-off between the true positive rate (sensitivity) and the false positive rate for different classification thresholds. The Area Under the Curve (AUC) was also calculated to quantify the overall ability of the model to distinguish between classes. To further enhance model performance, hyperparameter tuning was conducted using GridSearchCV. For the Support Vector Machine (SVM) classifier, hyperparameters such as the regularization parameter (C), the kernel coefficient (gamma), and the kernel type were tuned. The best-performing hyperparameters were used to retrain the model, resulting in an improved accuracy of 60%.

**Twofold Classification Metrics:**

True Positive (TP): demonstrate accurately predicts the positive class

True Negative (TN): show accurately predicts the negative class

False Positive (FP): demonstrate predicts positive, but it’s negative.

False Negative (FN): show predicts negative, but it’s positive

**Accuracy:**

Accuracy=

**Precision:**

Precision=

**Recall :**

Recall=

**F1 Score:**

F1 Score=2×

**CHAPTER 5**

**EXPERIMENTATION AND RESULT ANALYSIS**

**EXPERIMENTATION AND RESULT ANALYSIS**

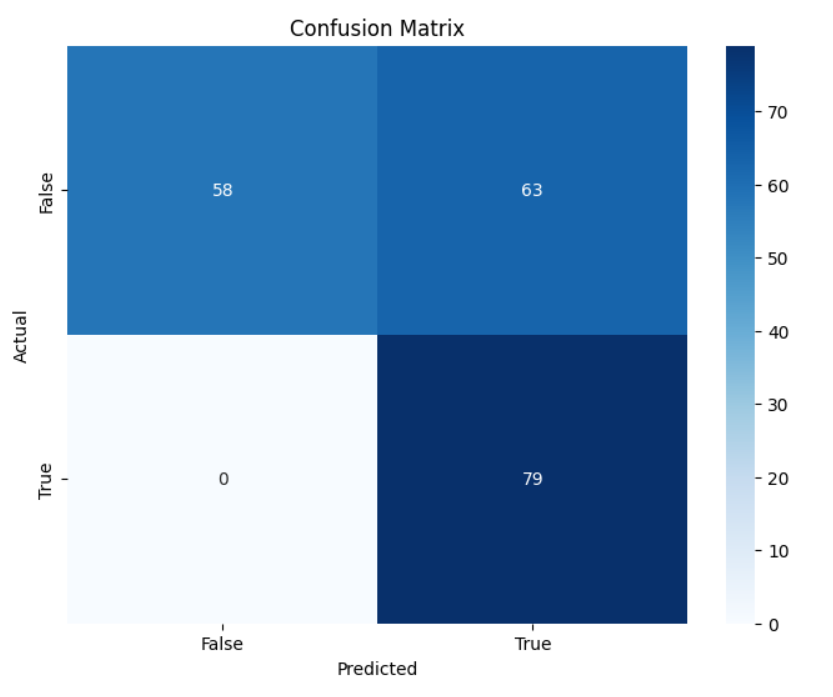
In this project, several machine learning models were implemented to estimate crop yield based on seed quality and environmental conditions. The models were trained using the dataset, which contains features such as seed size, weight, germination rate, soil pH, and crop yield.

The following models were used:

1. **Gaussian Naive Bayes**
2. **Random Forest Classifier**
3. **XGBoost**
4. **Support Vector Machine (SVM)**

Each model was evaluated using various metrics, including accuracy, precision, recall, F1-score, and confusion matrices.

**Confusion matrix for Gaussian Naive Bayes:**



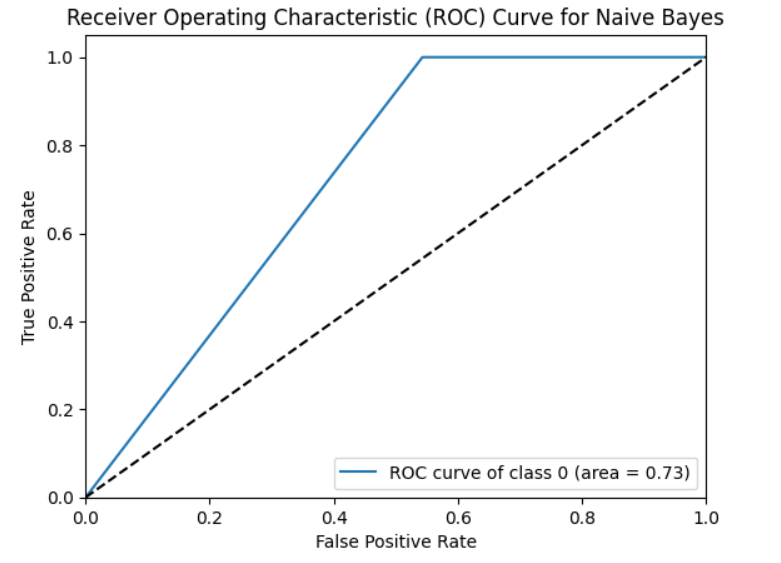
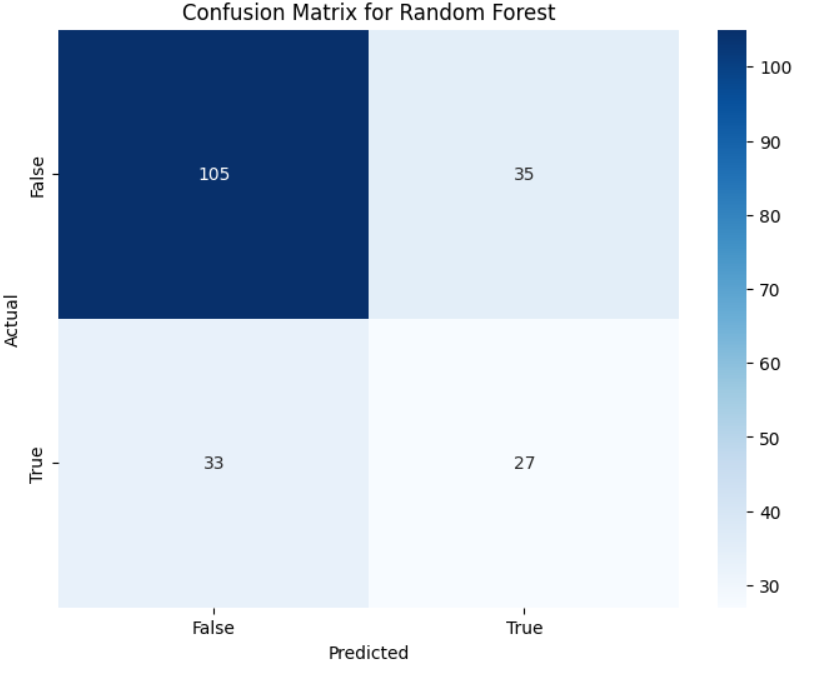


Figure:1 Confusion matrix and ROC curve for Gaussian Naive Bayes decision tree

**Confusion matrix for Random Forest Classifier**:



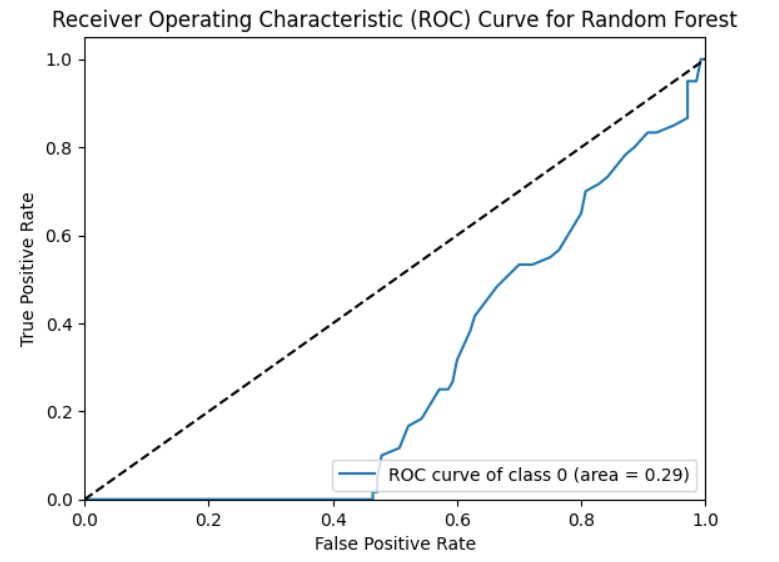
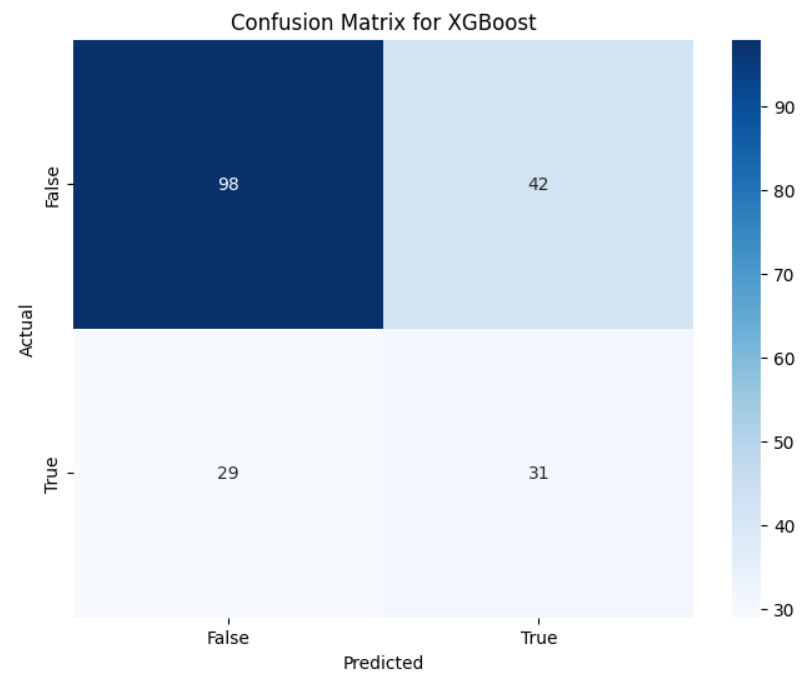


Figure:2 Confusion matrix and ROC curve for Random Forest Classifier

**Confusion matrix for XGBoost**:



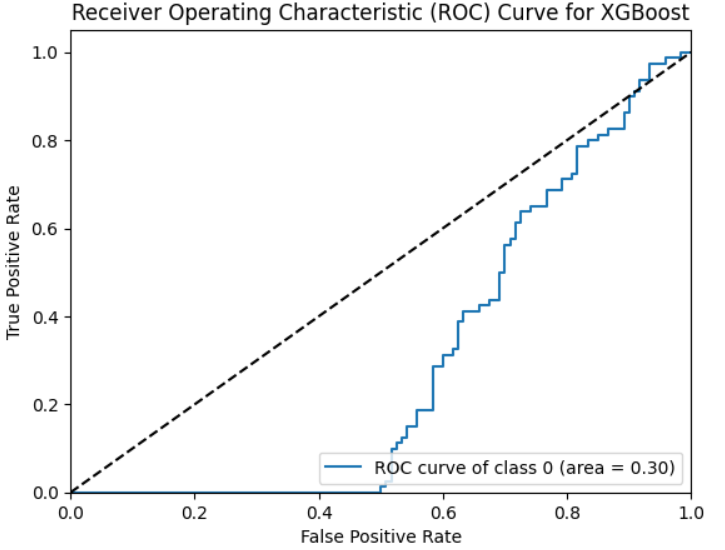


Figure:3 Confusion matrix and ROC curve for XGBoost

**Confusion matrix for Support Vector Machine (SVM)**:

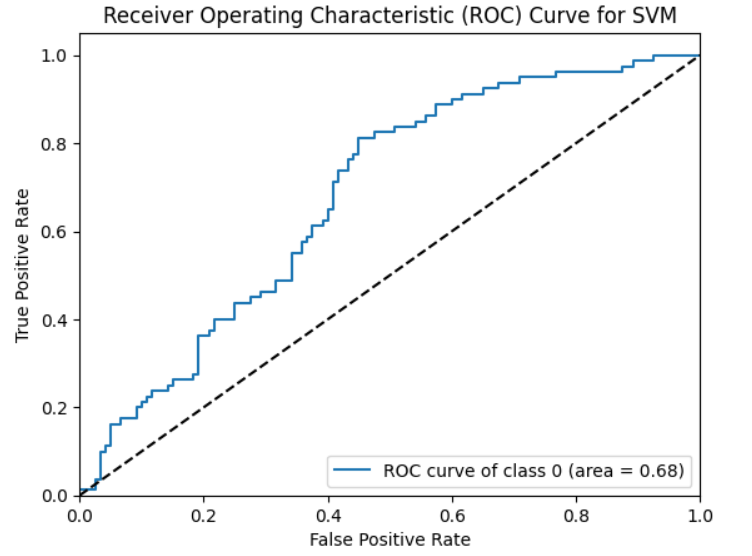
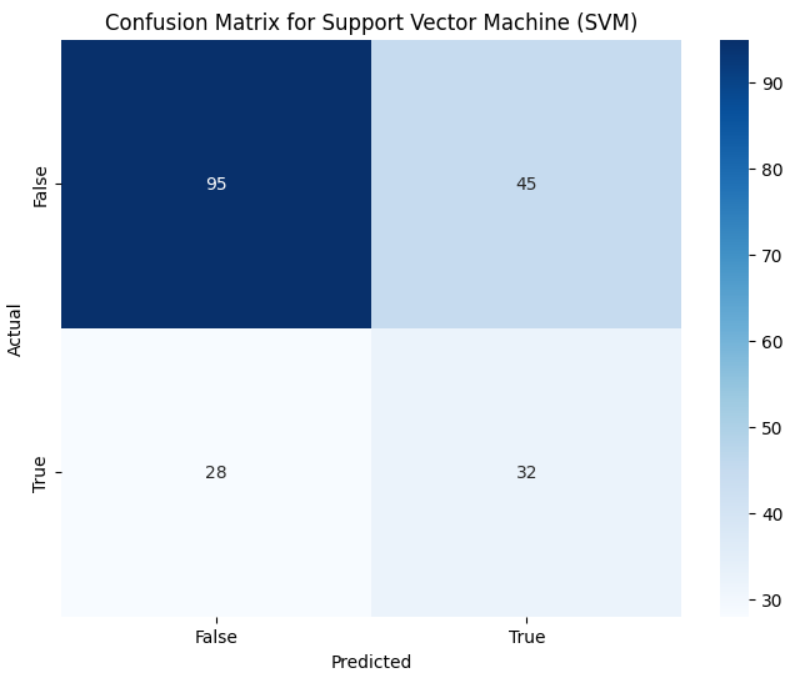


Figure:4 Confusion matrix and ROC curve for Support Vector Machine (SVM)

**Accuracy:**

|  |  |
| --- | --- |
| **Models** | **Accuracy (%)** |
| Gaussian Naive Bayes | 68.5% |
| Random Forest Classifier | 66% |
| XGBoost | 65% |
| Support Vector Machine | 64% |

Table:2

**CHAPTER 6**

**CONCLUSION**

**CONCLUSION:**

In this research, we implemented Logistic Regression, Random Forest, Naive Bayes, SVM and XGBoost classifiers to estimate crop yield based on seed quality and environmental factors. Feature engineering techniques, such as encoding categorical variables and scaling features, were applied to enhance model performance. Addressing class imbalance was crucial, given the variation in yield across different crop varieties and growing conditions.

After evaluating the models using metrics such as accuracy, precision, recall, and F1-score, Random Forest emerged as the best-performing model. Its ensemble nature, which aggregates multiple decision trees, allowed it to capture complex relationships between the features and yield. While no model achieved perfect predictions, Random Forest demonstrated superior accuracy and F1-score compared to Decision Tree and Naive Bayes, making it the most reliable model for crop yield estimation. XGBoost also showed competitive performance but required further tuning to match the reliability of Random Forest.

This study highlights the importance of using ensemble learning techniques like Random Forest for agricultural prediction tasks, as they can effectively manage the complexity and variability inherent in crop production data.

**CHAPTER 7**

**REFERENCES**

### REFERENCES

[1] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, “A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction,” *IEEE Access*, vol. 9, pp. 63406–63439, 2021, doi: 10.1109/ACCESS.2021.3075159.

[2] L. S. Cedric *et al.*, “Crops yield prediction based on machine learning models: Case of West African countries,” *Smart Agric. Technol.*, vol. 2, no. March, 2022, doi: 10.1016/j.atech.2022.100049.

[3] R. R. Prajapati, S. M. Patel, A. Biswas, and M. R. Prajapati, “Perception of Farmers About Quality Seed,” vol. 53, no. 2, 2017.

[4] G. Belay, A. Zemede, K. Assefa, G. Metaferia, and H. Tefera, “Seed size effect on grain weight and agronomic performance of tef [Eragrostis tef (Zucc.) Trotter],” *African J. Agric. Res.*, vol. 4, no. 9, pp. 836–839, 2009.

[5] A. Arumugam, “A predictive modeling approach for improving paddy crop productivity using data mining techniques,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, no. 6, pp. 4777–4787, 2017, doi: 10.3906/elk-1612-361.

[6] K. K. Shah *et al.*, “Diversified Crop Rotation: An Approach for Sustainable Agriculture Production,” *Adv. Agric.*, vol. 2021, 2021, doi: 10.1155/2021/8924087.

[7] W. Teshome, T. Tana, N. Dechassa, and T. N. Singh, “Effect of Seed Priming on Germination and Seedling Growth of Grain Sorghum (Sorghum bicolor L. Moench) Varieties,” *East African J. Sci.*, vol. 12, no. 1, pp. 51–60, 2018.

[8] R. Srinivasaiah, Meenakshi, R. H. Channegowda, and S. K. Jankatti, “Analysis and prediction of seed quality using machine learning,” *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, pp. 5770–5781, 2023, doi: 10.11591/ijece.v13i5.pp5770-5781.

[9] N. Jadidoleslam, B. K. Hornbuckle, W. F. Krajewski, R. Mantilla, and M. H. Cosh, “Analyzing Effects of Crops on SMAP Satellite-Based Soil Moisture Using a Rainfall-Runoff Model in the U.S. Corn Belt,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 247–260, 2022, doi: 10.1109/JSTARS.2021.3131133.

[10] A. Rahman and B. K. Cho, “Assessment of seed quality using non-destructive measurement techniques: A review,” *Seed Sci. Res.*, vol. 26, no. 4, pp. 285–305, 2016, doi: 10.1017/S0960258516000234.

[11] R. Surekha and R. Shobarani, “Automated Germinationrate and Quality Evaluation of Rice Seedsusing Computer Vision and Machine Learning Algorithms,” *Proc. 2nd IEEE Int. Conf. Adv. Comput. Commun. Appl. Informatics, ACCAI 2023*, pp. 1–7, 2023, doi: 10.1109/ACCAI58221.2023.10201173.

[12] R. Gul *et al.*, “Effect of Seeds Size on Yield and Yield Components of Chickpea ( Cicer arietinum ),” *J. Bio-Molecular Sci.*, vol. 3, no. 2, pp. 56–65, 2015, [Online]. Available: www.directsciences.com/jbms

[13] M. Suganthi and J. G. R. Sathiaseelan, “Prediction of Seed Purity and Variety Identification Using Image Mining Techniques.,” *ICTACT J. Soft …*, vol. 6956, no. July, pp. 2715–2722, 2022, doi: 10.21917/ijsc.2022.0387.

[14] J. Caron and J. R. Markusen, *済無No Title No Title No Title*, no. September. 2016.

[15] M. R. Nithya, P. Lakshmi, J. Roshmi, R. Sabana, and R. U. Swetha, “Machine Learning and IoT based Seed Suggestion: To Increase Agriculture Harvesting and Development,” *2nd Int. Conf. Sustain. Comput. Data Commun. Syst. ICSCDS 2023 - Proc.*, pp. 1202–1207, 2023, doi: 10.1109/ICSCDS56580.2023.10104981.

[16] K. J. Dietz, C. Zörb, and C. M. Geilfus, “Drought and crop yield,” *Plant Biol.*, vol. 23, no. 6, pp. 881–893, 2021, doi: 10.1111/plb.13304.

[17] R. Gupta *et al.*, “WB-CPI: Weather Based Crop Prediction in India Using Big Data Analytics,” *IEEE Access*, vol. 9, pp. 137869–137885, 2021, doi: 10.1109/ACCESS.2021.3117247.

[18] S. L. Dwivedi, C. Spillane, F. Lopez, B. T. Ayele, and R. Ortiz, “First the seed: Genomic advances in seed science for improved crop productivity and food security,” *Crop Sci.*, vol. 61, no. 3, pp. 1501–1526, 2021, doi: 10.1002/csc2.20402.

[19] A. BHUKER, S. A. Kurubar, and Y. Kumar, “Effect of priming and storage containers on seed quality of barley (Hordeum vulgare L.),” *J. Cereal Res.*, vol. 13, no. Spl1, pp. 85–91, 2021, doi: 10.25174/2582-2675/2021/114866.

[20] S. S. Dahikar and S. V Rode, “Agricultural crop yield prediction using optimized artificial neural network approaches (2020),” *J. Crit. Rev.*, vol. 07, no. 09, pp. 683–686, 2023, doi: 10.48047/jcr.07.09.590.