# Crop Prediction using Gaussian Naive Bayes Algorithm

Author 1: Mallikarjun
*Department: Information Science and Engineering*
*Institution: M.S. Ramaiah Institute of Technology*
Email: mallikarjundm9591@gmail.com

Author 2: Mandar Desurkar
*Department: Information Science and Engineering*
*Institution: M.S. Ramaiah Institute of Technology*
Email: mandardesurkar3@gmail.com

***Abstract*—India's agricultural sector is crucial to its economy, but many farmers continue to grow the same crops and apply fertilizers without adequate knowledge, leading to soil degradation and reduced crop yields. This study proposes a novel system that utilizes machine learning, specifically the Gaussian Naive Bayes algorithm, to recommend optimal crops based on soil and weather parameters. By analyzing factors like soil pH, temperature, humidity, rainfall, crop data, and NPK values (nitrogen, phosphorus, potassium), the system can provide data-driven recommendations to farmers, helping them select the most suitable crops for their specific conditions. We achieved an accuracy of 99.3**

**This approach aims to enhance crop yield and profitability and seeks to reduce soil pollution and promote sustainable farming practices. The implementation of this system could lead to more efficient resource utilization and improved agricultural productivity, ultimately benefiting both farmers and the environment.**

***Index Terms*—Machine Learning, Crop Prediction, Gaussian Naive Bayes, Precision Agriculture, Crop Recommendation.**

## I. Introduction

The agriculture sector, a cornerstone of global sustenance and economic stability, faces unprecedented challenges due to rapid population growth, climate change, and resource constraints. To meet the increasing demand for food while maintaining sustainability, innovative technological solutions are essential. One promising avenue is the application of machine learning (ML) to enhance crop prediction and management. Among the various ML techniques, the Naive Bayes algorithm stands out for its simplicity, efficiency, and robustness in handling agricultural data.

India's agricultural sector, crucial to its economy, exemplifies these challenges. Indian farmers face numerous problems, including soil degradation, unpredictable weather patterns, inefficient use of resources, and a lack of access to modern farming techniques. Continuous use of chemical fertilizers without adequate knowledge has led to significant soil degradation, with over 120 million hectares of land in India affected according to the Indian Council of Agricultural Research (ICAR). Climate change has resulted in erratic weather patterns, affecting crop yields; for instance, Maharashtra experienced a 40% deficit in rainfall in 2019, severely impacting agricultural output. Many farmers lack access to data-driven insights for optimal resource utilization, leading to inefficient use of water, fertilizers, and pesticides, which increases costs and causes environmental harm. Furthermore, many Indian farmers are not equipped with modern farming techniques and rely on traditional methods, resulting in lower productivity and profitability.

The emergence of big data and advances in artificial intelligence herald a transformative change for the agriculture industry. Precision agriculture, made possible by machine learning algorithms, can uncover patterns and insights in vast datasets that are beyond human comprehension. Crop prediction, a critical component of precision agriculture, allows for informed decisions about planting, irrigation, fertilization, and harvesting. These decisions can lead to increased yields, reduced waste, and more efficient use of resources[1].

This study proposes a novel system that utilizes the Gaussian Naive Bayes algorithm to recommend optimal crops based on soil and weather parameters. By analyzing factors like soil pH, temperature, humidity, rainfall, crop data, and NPK values (nitrogen, phosphorus, potassium), the system can provide data-driven recommendations to farmers, helping them select the most suitable crops for their specific conditions. We achieved an accuracy of 99.3% with this method.

This research is important because it has the potential to transform agricultural practices by offering insights based on data. With the use of the Naive Bayes algorithm, this research attempts to create an efficient and reasonably priced crop prediction tool[7]. With the use of such a tool, farmers will be able to make more informed decisions, which will ultimately boost agricultural sustainability and productivity. Moreover, this research contributes to the broader field of precision agriculture, showcasing how ML can be applied to solve complex problems in farming. The insights gained from this study can inform future research and development efforts, paving the way for more advanced and integrated agricultural technologies[2].

By addressing the challenges faced by Indian farmers and leveraging the power of machine learning, this approach aims to enhance crop yield and profitability while reducing soil pollution and promoting sustainable farming practices. The implementation of this system could lead to more efficient resource utilization and improved agricultural productivity, ultimately benefiting both farmers and the environment[6].

## II. Literature Review

Crop prediction using machine learning techniques has become increasingly important for improving agricultural productivity and supporting farmers in decision-making[7]. Researchers have explored various methods to predict crop outcomes effectively.

In the study titled "Crop Prediction using Machine Learning Approaches" by Nischitha K et al. (2020)[11], the authors highlight the critical role of agriculture in India and propose a system that uses machine learning to recommend suitable crops based on soil conditions and weather data. For instance, Ashwani Kumar Kushwaha et al. (2020)[[11] utilized big data technologies like Hadoop to analyze extensive soil and weather data, showing potential in enhancing agricultural decisions with an accuracy rate of 85%.

Different machine learning algorithms have been studied for their effectiveness in predicting crop outcomes. Girish L et al. (2020) [11] found Support Vector Machine (SVM) to be highly efficient for rainfall prediction, achieving an accuracy of 86.5%, emphasizing the need for selecting appropriate algorithms for different agricultural tasks. Rahul Katarya et al. (2020) explored artificial intelligence techniques such as K-Nearest Neighbors (KNN) and neural networks, showing significant improvements in crop yield predictions with reported accuracy rates of up to 95

The proposed systems generally involve several stages. Data collection includes gathering information from government sources and databases, focusing on factors like soil pH, temperature, humidity, and rainfall. Data preprocessing ensures accuracy by cleaning data and handling missing values. Machine learning models like SVM and Decision Trees are then applied to predict crop suitability and recommend appropriate crops, fertilizers, and seeds. Nischitha K et al. (2020) achieved an accuracy of 98.2% using Decision Trees.

Experimental results have demonstrated the effectiveness of these systems in providing accurate crop recommendations and agricultural advice. For example, some models have achieved accuracy rates of over 90%, helping farmers make informed decisions to maximize profits and reduce environmental impact. Future enhancements could include integrating GPS for automated data collection and developing models to anticipate food shortages or surpluses.

Another relevant study, "Machine Learning Techniques Based Prediction for Crops in Agriculture" by Dr. D. ManendraSai et al. (2023) [5], focuses on precision agriculture. It emphasizes the importance of accurate crop yield predictions for planning and decision-making in agriculture. Machine learning, categorized under artificial intelligence, proves beneficial for analyzing historical data to forecast future crop outcomes based on environmental factors, with models reaching accuracy levels of 92

The study distinguishes between descriptive and predictive models within machine learning, with descriptive models used to understand past data and predictive models forecasting future events. Related work has explored various techniques like data mining and computer vision to optimize crop yield rates and enhance agricultural efficiency.

In conclusion, these studies underscore the potential of machine learning in revolutionizing agricultural practices by providing precise crop recommendations and supporting sustainable farming. The findings highlight the importance of ongoing research to improve prediction accuracy and enhance agricultural productivity globally.

## III. Comparative Analysis

| Authors / Year | Technique | Accuracy |
|---|---|---|
| Ashwani Kumar Kushwaha et al. (2020) | Hadoop | 85% |
| Girish L et al. (2020) | SVM | 86.5% |
| Rahul Katarya et al. (2020) | KNN | 95% |
| Nischitha K et al. (2020) | DT | 98.2% |
| Proposed method | NB | 98.3% |
| Dr. D. ManendraSai et al. (2023) | ML | 92% |

TABLE I
Comparative Analysis of Different Techniques

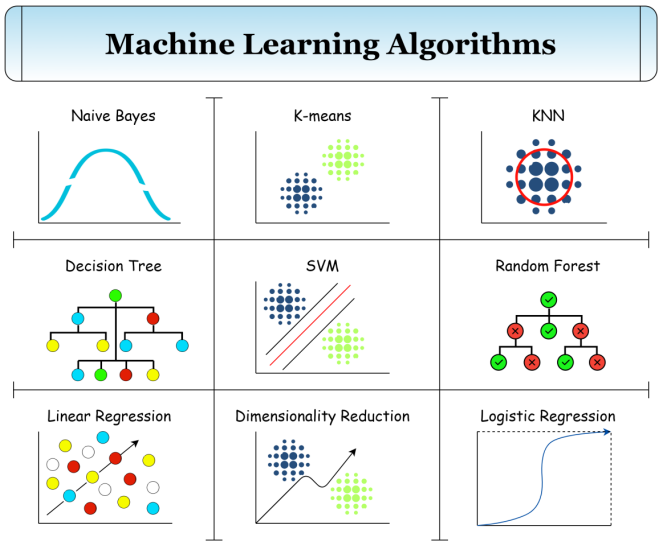


Fig. 1. Types of ML Algorithm

## IV. Proposed Methodology

### A. System Architecture

The proposed system architecture for crop prediction and recommendation using the Gaussian Naive Bayes algorithm is designed to enhance agricultural practices and outcomes by leveraging advanced machine learning techniques. This system encompasses several critical stages: data collection, data preprocessing, machine learning implementation, and crop recommendation[1].

*1) Data Collection:* Data collection forms the foundation of the system, gathering relevant information from various credible sources. Primary sources include government agricultural websites and meteorological departments, which provide comprehensive datasets on essential parameters such as soil pH,

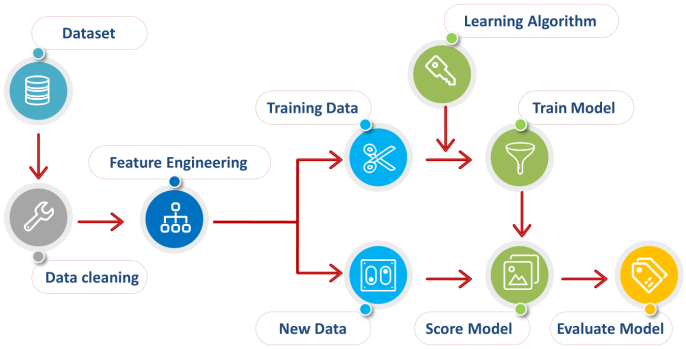


Fig. 2. Model Flow

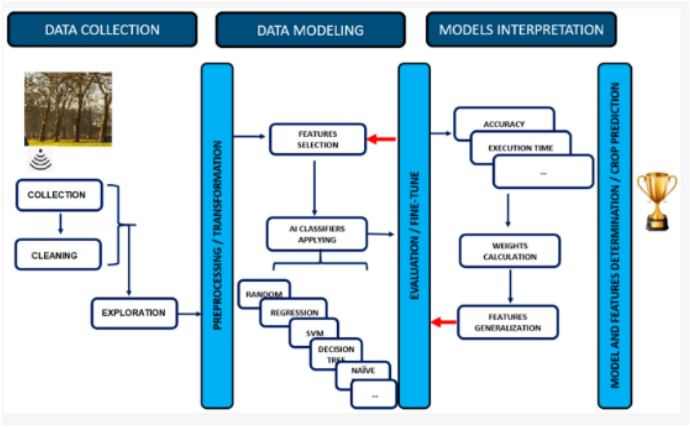


Fig. 3. Feature Determination

temperature, humidity, rainfall, crop-specific data, and NPK (Nitrogen, Phosphorous, Potassium) values. These datasets are critical as they offer a detailed view of the environmental and soil conditions necessary for accurate crop prediction and recommendation. Collecting high-quality, diverse data ensures the system can make well-informed and precise recommendations to farmers.

*2) Data Preprocessing:* Data preprocessing is a crucial step to ensure the quality and usability of the collected data. This stage involves several key techniques:

- **Data Cleaning:** Removing inconsistencies and handling missing values through imputation or deletion. Techniques such as statistical imputation are used to fill in missing values appropriately, ensuring data integrity.
- **Feature Engineering:** Enhancing the dataset by creating new features derived from existing ones can significantly improve the predictive power of the model. For example, combining soil quality indicators with weather patterns to create indices that reflect optimal crop conditions provides more nuanced inputs for the model.
- **Normalization:** Scaling numerical features to a standard range to prevent any single feature from disproportionately influencing the model during training. This step ensures that all features contribute equally to the model's learning process, thereby enhancing its stability and performance.
- **Data Splitting:** Dividing the dataset into training and testing subsets is essential to evaluate the model's performance accurately. The training set is used to train the model, while the testing set serves to assess how well the model generalizes to new, unseen data. This step helps in detecting overfitting and provides an estimate of the model's predictive accuracy in real-world scenarios.

*3) Machine Learning Algorithm:* The Gaussian Naive Bayes (GNB) algorithm is chosen for its efficacy in handling continuous data and its straightforward implementation in crop prediction models. The steps include:

- **Training Process:** During the training phase, the GNB algorithm utilizes the preprocessed dataset to establish correlations between input features and target labels. This involves learning how various factors such as soil attributes, weather conditions (like temperature, humidity, and rainfall), and potentially historical crop yields relate to determining the most suitable crop types for specific agricultural contexts. By analyzing these relationships, the algorithm builds a probabilistic model that forms the basis for subsequent predictions.
- **Prediction Mechanism:** Once trained, the GNB model applies probabilistic principles to predict the optimal crop for given input parameters. For instance, based on inputs such as soil composition, forecasted weather conditions, and historical agricultural performance data, the model calculates probabilities for each crop type. The crop with the highest probability is then recommended as the most suitable choice for planting on the specified land.
- **Crop Recommendation:** The ultimate goal of the GNB-based system is to provide actionable recommendations tailored to local agricultural conditions. By leveraging its predictive capabilities, the system advises farmers on crops likely to thrive based on comprehensive data analysis. This empowers decision-makers to make informed choices that can potentially optimize yield outcomes and resource allocation.

The Gaussian Naive Bayes algorithm works by assuming that the features follow a Gaussian (normal) distribution and calculates the probability of each crop being suitable based on the input parameters. The probability $P(C|X)$ of a class $C$ given an input vector $X$ is computed as follows:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{1}$$

Where:

- $P(X|C)$ is the likelihood of $X$ given $C$
- $P(C)$ is the prior probability of class $C$
- $P(X)$ is the probability of $X$

For continuous features, the likelihood $P(X_i|C)$ is given by the Gaussian distribution:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(X_i - \mu_C)^2}{2\sigma_C^2}\right) \tag{2}$$

Where:

- $\mu_C$ is the mean of feature $X_i$ for class $C$
- $\sigma_C$ is the standard deviation of feature $X_i$ for class $C$

Inputs to the algorithm include critical factors such as temperature, humidity, soil pH, and NPK values. By analyzing these inputs, the algorithm can predict which crops are most likely to thrive under the given environmental conditions. The model's simplicity allows for quick computation, making it practical for real-time applications in farming.

*4) Crop Recommendation:* The final component of the system is the crop recommendation module, which provides actionable insights to farmers. Based on the predictions generated by the Gaussian Naive Bayes algorithm, the system recommends the most suitable crops for the given land conditions. This recommendation considers all the input parameters, ensuring that the suggested crops are optimal for the specific soil and weather conditions. The goal is to help farmers make informed decisions that enhance crop yield and profitability while promoting sustainable farming practices. By offering precise crop recommendations, the system aims to reduce soil degradation and improve the overall efficiency of agricultural practices.

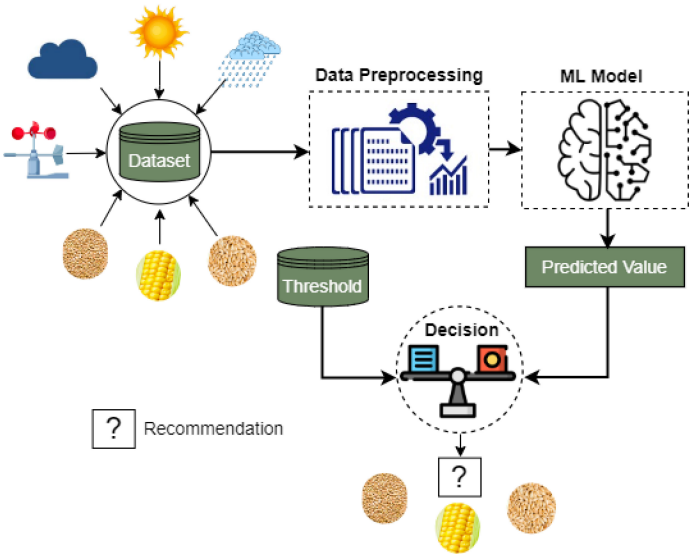


Fig. 4. Naive bias Prediction Model

### B. Performance Metrics for Model Evaluation

To ensure the effectiveness and accuracy of the Gaussian Naive Bayes algorithm in crop prediction, several performance metrics are used:

These formulas and tables provide a comprehensive foundation for understanding and implementing the Gaussian Naive Bayes algorithm in the proposed crop prediction and recommendation system. By integrating these metrics into the evaluation process, the model's accuracy, precision, recall, and overall performance can be rigorously assessed, ensuring reliable and actionable crop recommendations for farmers.

TABLE II
PERFORMANCE METRICS FOR MODEL EVALUATION

| Metric | Formula |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F1 Score | $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$ |
| Confusion Matrix | True vs. predicted classifications |

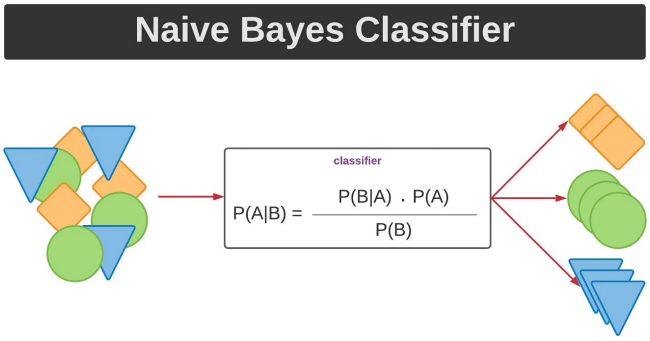


Fig. 5. Naive Baye Formula

## V. IMPLEMENTATION

### A. System Architecture

The proposed system for crop prediction and recommendation is built using the Gaussian Naive Bayes algorithm. The system encompasses several stages: data collection, data preprocessing, machine learning implementation, and crop recommendation. This design aims to enhance agricultural practices by leveraging advanced machine learning techniques.

### B. Data Collection

Data collection forms the foundation of the system, gathering relevant information from credible sources such as government agricultural websites and meteorological departments. These sources provide comprehensive datasets on essential parameters including soil pH, temperature, humidity, rainfall, crop-specific data, and NPK (Nitrogen, Phosphorous, Potassium) values. High-quality and diverse data ensure that the system can make well-informed and precise recommendations to farmers.

### C. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and usability of the collected data. This stage involves several key processes:

1) **Loading the Data**: The dataset is loaded into a Pandas DataFrame for easy manipulation and analysis. The initial structure, shape, and size of the dataset are examined to understand its dimensions and contents.
2) **Handling Missing Values**: Missing values within the dataset are identified and handled appropriately, using

techniques such as imputation with mean or median values or by dropping rows/columns if necessary.

3) **Feature Engineering**: Enhancing the dataset by creating new features derived from existing ones can significantly improve the model's predictive power. For example, combining soil quality indicators with weather patterns to create indices that reflect optimal crop conditions.
4) **Normalization**: Scaling numerical features to a standard range is crucial to prevent any single feature from disproportionately influencing the model during training. This step ensures that all features contribute equally to the model's learning process.
5) **Feature and Target Variables**: The dataset is divided into features (X) and the target variable (y). This division is essential for training machine learning models.

### *D. Exploratory Data Analysis (EDA)*

EDA helps in understanding the data and uncovering patterns. Various visualizations and analyses were performed:

1) **Pairplot Visualizations**: Pair plots were generated to observe the relationships and interactions between different features, helping in identifying trends and correlations.
2) **Correlation Matrix and Heatmap**: A correlation matrix was computed to quantify the relationships between numerical features, visualized using a heatmap to easily identify strong positive or negative correlations.
3) **Histograms**: Histograms for each numerical feature were plotted to understand their distributions, helping in identifying any skewness or outliers in the data.
4) **Scatter Plots**: Scatter plots were used to investigate potential correlations between pairs of features, such as temperature and humidity, which are critical in understanding their impact on the target variable.

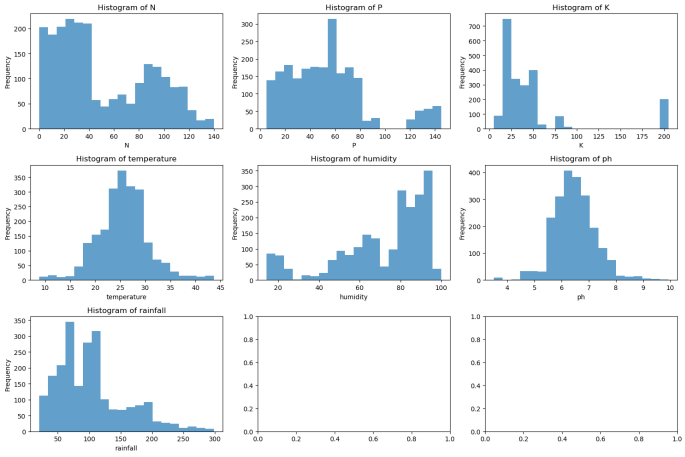


Fig. 6. Histogram

### *E. Data Splitting*

The dataset was split into training and testing sets to evaluate the model's performance on unseen data. Typically, an 80-20 split was used, where 80% of the data was used for training and 20% for testing.

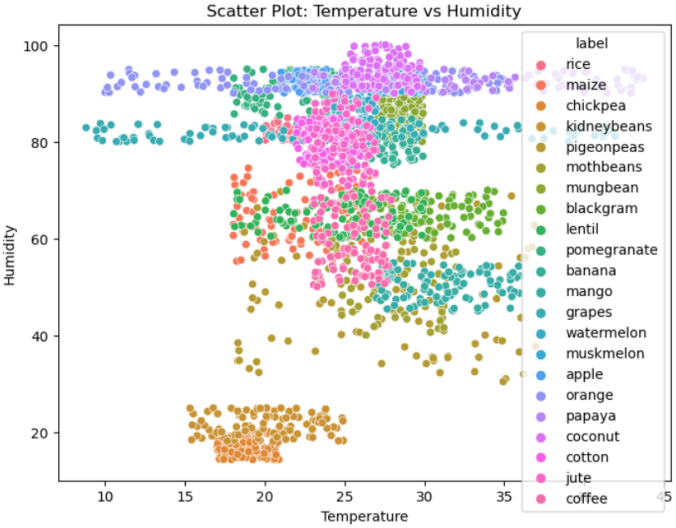


Fig. 7. Scatter Plot

### *F. Machine Learning Algorithm*

The core of the system is the implementation of the Gaussian Naive Bayes (GNB) algorithm for crop prediction. This algorithm is chosen for its simplicity, efficiency, and effectiveness in handling probabilistic data[4].

1) **Training Process**: During the training phase, the GNB algorithm utilizes the preprocessed dataset to establish correlations between input features and target labels. This involves learning how various factors such as soil attributes, weather conditions, and historical crop yields relate to determining the most suitable crop types.
2) **Prediction Mechanism**: Once trained, the GNB model applies probabilistic principles to predict the optimal crop for given input parameters. For instance, based on inputs such as soil composition, forecasted weather conditions, and historical agricultural performance data, the model calculates probabilities for each crop type.
3) **Crop Recommendation**: The system recommends the most suitable crops for the given land conditions based on the predictions generated by the GNB algorithm. This recommendation considers all input parameters, ensuring that the suggested crops are optimal for the specific soil and weather conditions.

### *G. Model Training and Evaluation*

To validate the model's performance, additional machine learning algorithms were also tested:

1) **Logistic Regression**:
   - **Training**: The Logistic Regression model was trained on the training dataset.
   - **Evaluation**: Model performance was evaluated using accuracy scores and validated through cross-validation techniques.

2) **Random Forest Classifier**:

- **Training**: The Random Forest Classifier was trained with a specified number of estimators to ensure robustness.
- **Evaluation**: Model performance was evaluated using accuracy scores and validated through cross-validation techniques.

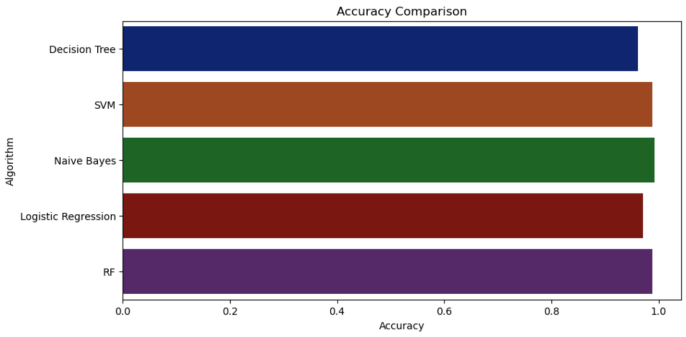


Fig. 8. Accuracy Comparison of Various Models

### H. Model Saving

To allow for future use without retraining, the trained models were saved using pickle.

```
import pickle

# Save the model to disk
filename = 'finalized_model.sav'
pickle.dump(model, open(filename, 'wb'))
```

### I. Comparison of Models

The performance of different models was compared using accuracy scores. A bar plot was created to visualize the accuracy of each algorithm, aiding in the identification of the most effective model for our dataset.

### J. Predictions

Trained models were used to make predictions on new data samples, demonstrating the practical applicability of the models in real-world scenarios.

### K. Conclusion

Our implementation effectively utilized various data preprocessing techniques, exploratory data analysis methods, and machine learning algorithms to build robust predictive models[10]. The Gaussian Naive Bayes algorithm emerged as a highly accurate model, showcasing its strength in handling the given dataset and providing reliable crop recommendations. This system aims to help farmers make informed decisions that enhance crop yield and profitability while promoting sustainable farming practices.[8]

## VI. Experimental Analysis

The Gaussian Naive Bayes model demonstrated impressive performance on both the training and test datasets, reflected in the following metrics:

- Training Accuracy: 0.9943
- Test Accuracy: 0.9932

### A. Understanding the Metrics

The accuracy metric is a measure of how well the model performs in correctly predicting the target variable. It is calculated as the ratio of the number of correct predictions to the total number of predictions made. The formula to compute accuracy is as follows:

$$Accuracy = \frac{Number of Correct Predictions}{Total Number of Predictions} \tag{3}$$

*1) Calculating Accuracy:* For both the training and test datasets, the accuracy can be computed using the above formula. Here's a detailed explanation of how you can calculate it:

1) **Determine Correct Predictions:** Identify the number of instances where the model's predicted value matches the actual value.
2) **Total Predictions:** Count the total number of instances in the dataset.
3) **Apply the Formula:** Divide the number of correct predictions by the total number of predictions.

### B. Classification Report

The classification report provides a detailed analysis of the performance of the classification model for each class (crop in this case). The key metrics included in the report are Precision, Recall, F1-Score, and Support. Here's a breakdown of what each metric represents:

- **Precision**: This is the ratio of correctly predicted positive observations to the total predicted positives. Precision is a measure of the accuracy of the positive predictions. It is calculated as:

$$Precision = \frac{True Positives}{True Positives + False Positives}$$

- **Recall (Sensitivity)**: This is the ratio of correctly predicted positive observations to the all observations in the actual class. Recall is a measure of the model's ability to capture all relevant instances. It is calculated as:

$$Recall = \frac{True Positives}{True Positives + False Negatives}$$

- **F1-Score**: This is the weighted average of Precision and Recall. The F1-Score is especially useful when you need a balance between Precision and Recall. It is calculated as:

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

- **Support**: This indicates the number of actual occurrences of the class in the dataset. It provides context for the other

metrics, as some classes might have more instances than others.

Detailed Metrics for Each Crop

Below is the classification report for the model's performance on different crops:

| Crop | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| apple | 1.00 | 1.00 | 1.00 | 22 |
| banana | 1.00 | 1.00 | 1.00 | 18 |
| blackgram | 1.00 | 1.00 | 1.00 | 21 |
| chickpea | 1.00 | 1.00 | 1.00 | 15 |
| coconut | 1.00 | 1.00 | 1.00 | 18 |
| coffee | 1.00 | 0.96 | 0.98 | 27 |
| cotton | 1.00 | 1.00 | 1.00 | 24 |
| grapes | 1.00 | 1.00 | 1.00 | 17 |
| jute | 0.94 | 0.89 | 0.92 | 19 |
| kidneybeans | 1.00 | 1.00 | 1.00 | 21 |
| lentil | 1.00 | 1.00 | 1.00 | 23 |
| maize | 1.00 | 1.00 | 1.00 | 20 |
| mango | 1.00 | 1.00 | 1.00 | 16 |
| mothbeans | 1.00 | 1.00 | 1.00 | 15 |
| mungbean | 1.00 | 1.00 | 1.00 | 25 |
| muskmelon | 1.00 | 1.00 | 1.00 | 20 |
| orange | 1.00 | 1.00 | 1.00 | 21 |
| papaya | 1.00 | 1.00 | 1.00 | 22 |
| pigeonpeas | 1.00 | 1.00 | 1.00 | 16 |
| pomegranate | 1.00 | 1.00 | 1.00 | 21 |

TABLE III
CLASSIFICATION REPORT

## C. Analysis of the Confusion Matrix

The provided confusion matrix is a detailed visualization used to evaluate the performance of a classification algorithm on a multi-class problem. In this specific case, the matrix represents the classification results for various agricultural crops.

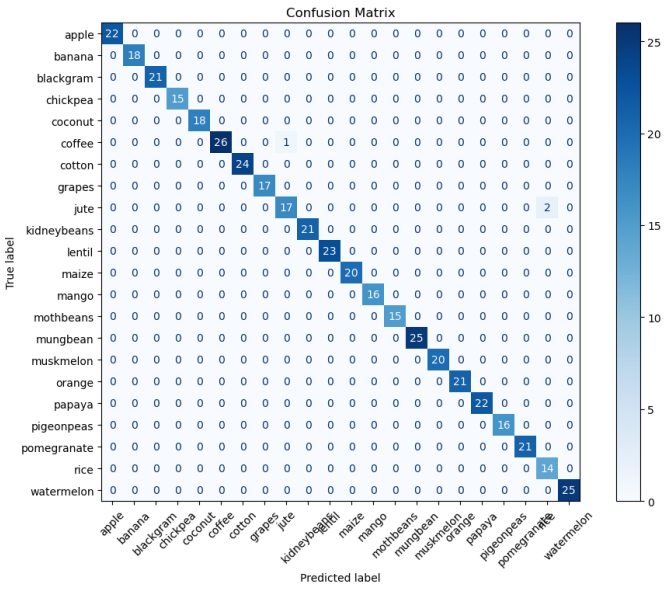


Fig. 9. Confusion Matrix

*1) Structure of the Confusion Matrix:*

*a) Axes::*

- The **vertical axis (True Label)** represents the actual classes of the crops.
- The **horizontal axis (Predicted Label)** represents the predicted classes assigned by the classification model.

*b) Diagonal Elements::*

- The values on the diagonal (from top-left to bottom-right) represent the number of correct predictions for each class. For instance, the model correctly predicted 22 instances of "apple," 18 instances of "banana," and so on.

*c) Off-Diagonal Elements::*

- Values not on the diagonal indicate misclassifications. For example, the model misclassified one instance of "coffee" as "mothbeans" and two instances of "jute" as "mungbean."

*d) Color Intensity::*

- The intensity of the color in each cell reflects the number of instances. Darker shades indicate a higher count of instances, providing a quick visual indication of where the model performs well or poorly.

*2) Insights from the Confusion Matrix:*

*a) High Accuracy Classes::*

- Classes like "coffee" (26), "mungbean" (25), "watermelon" (25), and "lentil" (23) have high numbers of correct predictions, suggesting that the model performs particularly well in identifying these crops.

*b) Misclassification Observations::*

- There are minimal misclassifications, indicating a generally robust model. Notable misclassifications include:
  - One instance of "coffee" was predicted as "mothbeans."
  - Two instances of "jute" were predicted as "mungbean."

*c) Perfect Predictions::*

- Several classes such as "blackgram," "cotton," "grapes," "kidneybeans," "maize," and "rice" have no misclassifications, indicating perfect predictions for these classes.

## D. Analysis of the Correlation Heatmap

The provided correlation heatmap is a graphical representation that displays the correlation coefficients between pairs of variables. The heatmap uses colors to indicate the strength and direction of the correlations, making it easier to identify patterns and relationships in the data.

*1) Structure of the Correlation Heatmap:*

*a) Axes::*

- Both the **horizontal** and **vertical axes** represent the variables in the dataset. In this heatmap, the variables are: N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall.

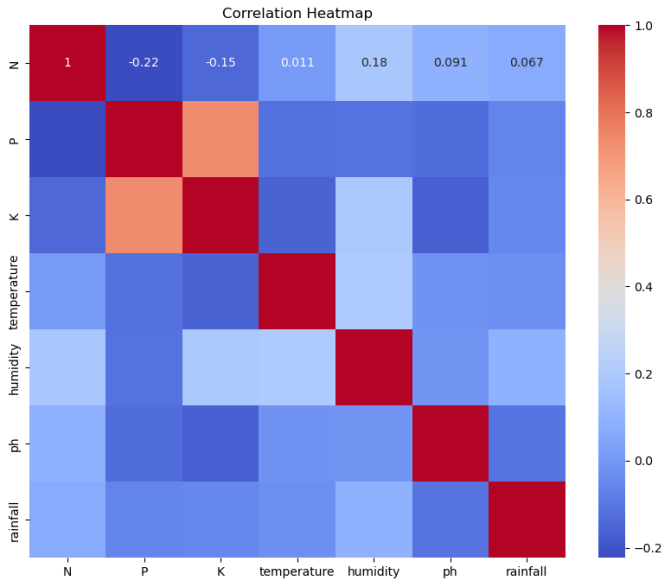


Fig. 10. Correlation Heatmap

*b) Color Coding::*

- The color intensity in each cell represents the correlation coefficient between the corresponding pair of variables.
- The scale on the right ranges from -1 to 1, where:
  - **Red** indicates a strong positive correlation (+1).
  - **Blue** indicates a strong negative correlation (-1).
  - **White** represents no correlation (0).

*c) Correlation Coefficients::*

- Each cell contains a numerical value that specifies the exact correlation coefficient between the two variables.

*2) Insights from the Correlation Heatmap:*

*a) High Positive Correlations::*

- The variables show varying degrees of positive correlation. For instance:
  - **temperature** and **humidity** have a correlation of 0.18.
  - **N** and **temperature** have a correlation of 0.18.
- These positive correlations suggest that as one variable increases, the other tends to increase as well.

*b) High Negative Correlations::*

- Some pairs of variables exhibit negative correlations, indicating an inverse relationship:
  - **P** and **N** have a correlation of -0.22.
  - **K** and **N** have a correlation of -0.15.
- These negative correlations suggest that as one variable increases, the other tends to decrease.

*c) Notable Observations::*

- There are no perfect correlations (1 or -1), indicating that while there are relationships, they are not perfect linear relationships.
- The correlation between **P** and **K** is 0.59, which is moderately high and indicates a considerable positive relationship.
- **temperature** and **ph** show a significant negative correlation (-0.20), suggesting that higher temperatures might be associated with lower pH values.

*3) Summary of the Correlation Analysis:* The correlation heatmap provides a visual and quantitative overview of the relationships between different variables in the dataset. Understanding these correlations is crucial for:

- Identifying which variables are related.
- Making informed decisions in data preprocessing.
- Feature selection for modeling.

The heatmap effectively highlights the interdependencies between the variables, which can guide further analysis and interpretation in your research.

By analyzing the correlation heatmap, researchers can gain valuable insights into the data's structure and relationships, aiding in the development of more accurate and interpretable models.

### *E. Additional Metrics Explained*

- **Precision: 0.9936**
  - Precision measures the accuracy of positive predictions. A precision of 0.9936 means that 99.36% of the predictions classified as positive are actually correct.
- **Recall: 0.9932**
  - Recall, also known as sensitivity, measures the ability to identify all actual positive cases. A recall of 0.9932 means that 99.32% of the actual positive cases were correctly identified by the model.
- **F1 Score: 0.9932**
  - The F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. An F1 score of 0.9932 indicates a high level of overall accuracy, balancing both precision and recall.

| Authors / Year | Technique | Accuracy |
|---|---|---|
| Ashwani Kumar Kushwaha et al. (2020) | Hadoop | 85% |
| Girish L et al. (2020) | SVM | 86.5% |
| Rahul Katarya et al. (2020) | KNN | 95% |
| Nischitha K et al. (2020) | DT | 98.2% |
| Proposed method | NB | 98.3% |
| Dr. D. ManendraSai et al. (2023) | ML | 92% |
| Proposed Model (2024) | NB | 99.3% |

TABLE IV
COMPARATIVE ANALYSIS OF PROPOSED MODEL WITH EXISTING MODELS

## VII. Front-End

The model is deployed using HTML, CSS, and Flask

## VIII. Conclusion

The Gaussian Naive Bayes model has demonstrated remarkable accuracy in predicting suitable crops, achieving an accuracy of 98.3%. This high level of accuracy suggests that the model is highly effective in making correct positive predictions

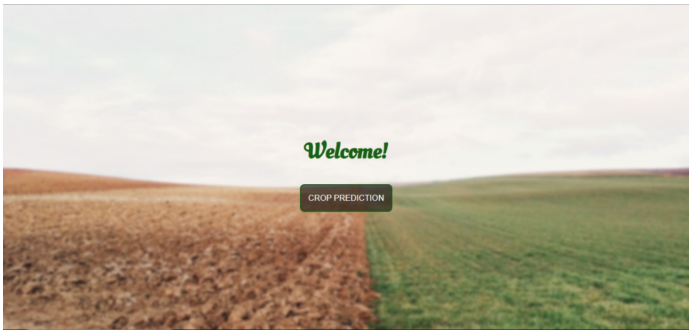


Fig. 11. Home Page

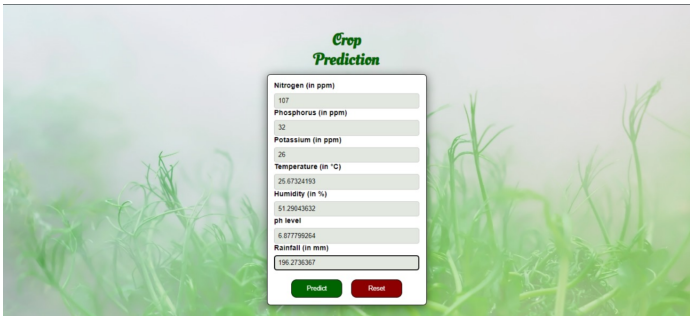


Fig. 12. Input Page


Fig. 13. Input Result Page

and identifying nearly all actual positive cases. The simplicity and efficiency of the Gaussian Naive Bayes model make it a practical choice for agricultural applications, particularly in handling continuous data common in agricultural datasets.

This predictive capability can greatly benefit farmers and the agro-industry by providing reliable recommendations on suitable crops based on various environmental and agricultural factors. By leveraging this model, farmers can optimize their planting decisions, potentially increasing yield and reducing resource waste. This, in turn, can lead to more sustainable and efficient agricultural practices[9].

However, there's room for improvement. Future research could explore more complex algorithms such as neural networks, Random Forest, and Support Vector Machines to uncover deeper patterns in agricultural data. Continuous updates to the model with real-time data could further enhance its accuracy and relevance[4]. Additionally, enhancing the model's robustness and generalizability through advanced feature engineering and broader geographic and temporal scope would be beneficial.

Collaboration with agricultural experts and farmers is crucial for refining the model and ensuring its practical applicability in real-world farming scenarios[6]. Their qualitative insights can complement quantitative data, providing a more comprehensive understanding of the factors influencing crop suitability.

By continuously refining these predictive models, we can offer more reliable and actionable insights, supporting sustainable and efficient agricultural practices, ultimately benefiting farmers, the agro-industry, and global food security.

## REFERENCES

[1] S. Pudumalar et al., "Crop recommendation system for precision agriculture," in Proceedings of the Eighth International Conference on Advanced Computing (ICoAC), IEEE, June 2017.

[2] M. Mohammed et al., "Machine Learning: Algorithms and Applications," CRC press.

[3] J. Zhang et al., "Evolutionary Computation Meets Machine Learning: A Survey," IEEE Computational Intelligence Magazine, Vol. 6, Issue 4, pp. 68-75, Nov. 2011.

[4] A. Singh et al., "A review of supervised machine learning algorithms," in Proceedings of the International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, October 2016.

[5] Dr. D. ManendraSai et al., "Machine Learning Techniques Based Prediction for Crops in Agriculture," Journal of Survey in Fisheries Sciences, Vol. 10, Issue 1S, March 2023.

[6] Li, L.; Wang, B.; Feng, P.; Liu, D.L.; He, Q.; Zhang, Y.; Wang, Y.; Li, S.; Lu, X.; Yue, C.; et al. Developing machine learning models with multi-source environmental data to predict wheat yield in China. Comput. Electron. Agric. 2022, 194, 106790.

[7] van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 2020, 177, 105709

[8] Kuradusenge, M.; Hitimana, E.; Hanyurwimfura, D.; Rukundo, P.; Mtonga, K.; Mukasine, A.; Uwitonze, C.; Ngabonziza, J.; Uwamahoro, A. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. Agriculture 2023, 13, 225.

[9] Xu, W.; Kaili, Z.; Tianlei, W. Smart Farm Based on Six-Domain Model. In Proceedings of the IEEE 4th International Conference on Electronics Technology (ICET), Chengdu, China, 7–10 May 2021; pp. 417–421.

[10] Moysiadis, V.; Tsakos, K.; Sarigiannidis, P.; Petrakis, E.G.M.; Boursianis, A.D.; Goudos, S.K. A Cloud Computing web-based application for Smart Farming based on microservices architecture. In Proceedings of the 11th International Conference on Modern Circuits and Systems Technologies (MOCAST), Bremen, Germany, 8–10 June 2022; pp. 1–5.

[11] Nischitha K, Vishwakarma D, Ashwini, Mahendra N, Manjuraju M.R. "Crop Prediction using Machine Learning Approaches." International Journal of Engineering Research Technology (IJERT), vol. 9, no. 08, August 2020, pp. 23-26.