A Smart Recommendation System for Crop Seed Selection Using Gradient Boosting Based on Environmental and Geospatial Data

Aryanti Aryanti [1]*, Nanda Iryani [2]*, Khairunnisa' [3]*
* Teknik Telekomunikasi, Politeknik Negeri Sriwijaya
aryanti@polsri.ac.id [1], nandairyani@polsri.ac.id [2], wahyudikhairunnisa@gmail.com [3]

Article Info

Article history:

Received 2025-07-15
Revised 2025-11-11
Accepted 2025-11-22

Keyword:

Crop seed recommendation,
Gradient Boosting,
Environmental data,
Geographical data,
Precision agriculture.

ABSTRACT

The selection of appropriate crop seeds is a critical factor in enhancing agricultural productivity. Nevertheless, farmers frequently face challenges when trying to determine which crop seeds match the unique features of their surrounding environment and geographic location. To address this, the study introduces a smart recommendation model that leverages real-time environmental measurements alongside vital geographical characteristics to support informed seed selection. The environmental features include temperature, humidity, and rainfall, while the geographical attributes encompass nitrogen, phosphorus, and potassium content. A Gradient Boosting classification algorithm is employed to model the relationships between these features and the optimal crop seed types, based on a labeled dataset. Experimental results demonstrate that the model achieves strong classification performance, indicating its effectiveness in delivering accurate and context-specific seed recommendations. The proposed system highlights the potential of data-driven approaches in supporting agricultural decision-making and can be further integrated into smart farming platforms to optimize crop planning and seed selection, ultimately contributing to improved agricultural outcomes.

I. Introduction

The agricultural sector is fundamental to sustaining food availability and driving economic progress, particularly in nations where a significant share of the population relies on farming-related occupations as a primary source of income [1]. In such settings, increasing agricultural productivity is not only an economic objective but also a vital component of national resilience and food sovereignty [2]. One of the most fundamental aspects influencing agricultural productivity is the selection of appropriate crop varieties. The choice of crop seeds greatly determines whether the farming process will yield optimal results or be hindered by environmental incompatibility and resource inefficiency [3].

Selecting the most suitable crop variety for a specific area is not a trivial taskThis process takes into account a combination of climatic and locational attributes, including temperature, humidity, precipitation levels, soil nutrient profiles, as well as the physical contours of the terrain [4]. Farmers, particularly those operating in resource-limited settings, often rely on personal experience, intuition, or generalized recommendations that may not reflect the unique conditions of their specific farming plots [5]. These limitations can lead to a mismatch between crop requirements and environmental conditions, resulting in reduced yields, increased susceptibility to pests and diseases, and considerable economic loss.

The problem addressed in this study centers on the lack of intelligent tools that can provide accurate, data-based recommendations for seed selection tailored to the specific environmental and geographical conditions of individual farmlands. While various agricultural extension services and platforms have been developed, many of them still offer static advice that does not consider dynamic real-time changes or detailed local variations. This condition underscores a critical opportunity that could be resolved by leveraging contemporary computational methods, especially those stemming from advancements in machine learning.

Machine learning has emerged as a powerful approach in many fields including agriculture [6]. Its ability to learn complex patterns from data, adapt to new information, and generate predictions makes it an ideal candidate for applications that require precision and personalization [7]. In agriculture, machine learning has been used in areas such as yield forecasting, pest detection, disease diagnosis, and irrigation management [8]. The integration of machine learning into agricultural decision-making supports the broader transition towards precision farming, where data is used to guide every aspect of crop management [9].

One illustration of how machine learning has been utilized within the agricultural domain can be found in research carried out through the work of [10]. In their approach, they utilized the C4.5 algorithm a type of decision tree to assess the appropriateness of various seed and fertilizer pairings. Their classification model delivered a notable performance, reaching 86.4% in predictive success. Nevertheless, the reported error rate of 13.6% pointed to inherent limitations, stemming from both the algorithm's simplicity and the restricted scope of the input variables. Additionally, the intricate branching structure led to overfitting issues, diminishing the model's effectiveness when applied to unfamiliar datasets.

In contrast, the research by [11] focused on the use of the Gradient Boosting algorithm in environmental classification tasks. Specifically, they implemented a hybrid approach combining Principal Component Regression with Gradient Boosting to classify water quality levels based on multiple environmental factors. Their model achieved a perfect classification accuracy of 100 percent, outperforming other commonly used classifiers such as Random Forest and Support Vector Machine. This result demonstrated the robustness of Gradient Boosting in handling complex, multivariate data and its capacity to model nonlinear relationships effectively.

Inspired by these developments, this study explores the application of the Gradient Boosting classification algorithm for the recommendation of crop seed types. The system presented in this study integrates various climatic conditions including ambient temperature, moisture levels, and precipitation along with key soil nutrient metrics such as nitrogen, phosphorus, and potassium concentrations. By processing this input through a Gradient Boosting model trained on labeled data, the system aims to support farmers in identifying crop varieties that best match the current conditions of their land.

This research aims to develop and assess a recommendation framework powered by machine learning, designed to support seed selection tailored to specific environmental and geographical contexts. By using this system, farmers can benefit from a decision-support tool that enhances accuracy in planning and optimizes the use of agricultural resources potentially resulting in better harvest outcomes and minimized losses. The work adds value to the expanding literature on precision agriculture through the implementation of a functional machine learning solution in the domain of crop selection. Unlike prior approaches that may have used fewer variables or simpler models, the use of multiple input factors and a robust ensemble learning method seeks to produce more reliable outcomes under various environmental conditions. Furthermore, the system can serve as a foundation for integration with broader smart farming platforms, offering potential scalability and long-term usability.

II. Research Method

A, Research Canvas

This study employed the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, a structured framework developed by SAS Institute to support systematic data mining workflows. SEMMA is widely adopted in data science for its clear, sequential approach in transforming raw data into insightful models [12]. in Figure 1. The research flow begins with the Sample phase, where environmental and geographical data are collected and prepared. The process continues with Explore, which involves examining data patterns and correlations through exploratory analysis. Next, the Modify phase is carried out to preprocess the data and engineer relevant features. This model is evaluated using accuracy, precision, recall, and F1-score.

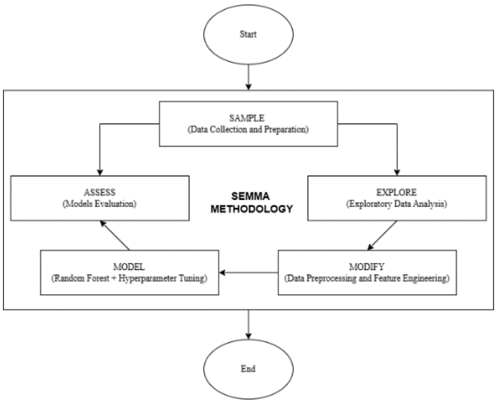

Start
SAMPLE
(Data Collection and Preparation)
ASSESS
(Models Evaluation)
SEMMA
METHODOLOGY
EXPLORE
(Exploratory Data Analysis)
MODEL
(Random Forest + Hyperparameter Tuning)
MODIFY
(Data Preprocessing and Feature Engineering)
End

Figure 1. Research flow diagram

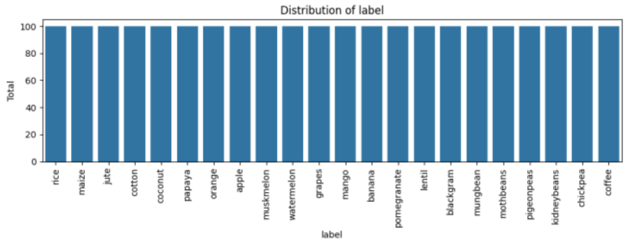
B. Sample

This study uses the Crop Recommendation Dataset, which is openly accessible via Kaggle and widely utilized in agricultural machine learning research [13]. The dataset comprises 2,200 records with 7 numerical features and 1 categorical target label, where each record corresponds to a specific agricultural condition associated with a crop type. The attributes include key environmental and soil-related parameters: Nitrogen (N), Phosphorus (P), Potassium (K) contents in the soil, temperature (°C), humidity (%), pH level, and rainfall (mm). These features collectively represent the essential conditions influencing crop growth. The target variable is the crop label, a categorical feature indicating the most suitable crop for the given environmental inputs. This dataset provides a solid foundation for training predictive models to recommend optimal crop types based on environmental and soil characteristics, and is highly relevant for smart farming and precision agriculture applications.

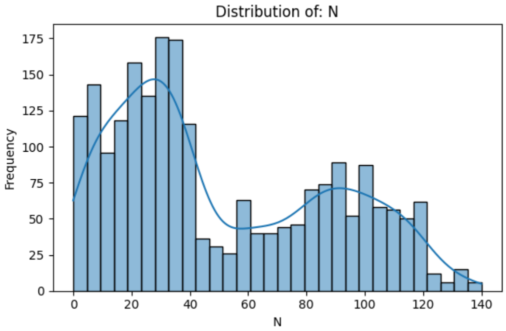
TABEL I
DESCIPTION DATA

Feature	Data Type
Nitrogen (N)	Numeral (Integer)
Phosphorus (P)	Numeral (Integer)
Potassium (K)	Numeral (Integer)
Temperature	Numeral (Integer)
Humidity	Numeral (Float)
ph	Numeral (Float)
Rainfall	Numeral (Float)

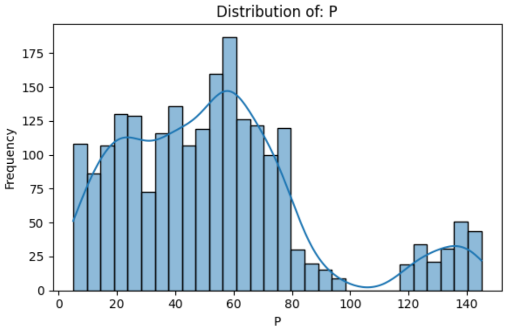
The feature Potassium (K) contains the highest number of outliers, accounting for 9.09% of the data in that attribute, followed by Phosphorus (P) and Rainfall. In contrast, Nitrogen (N) shows no outliers, indicating a consistent distribution of values within the expected range. Despite the presence of outliers, they are not removed at this stage. This decision is based on the robustness of the Gradient Boosting algorithm used in this study, which is known to perform well even in the presence of extreme values. However, the impact of these outliers will be closely examined during the model evaluation phase to ensure reliability and generalization.



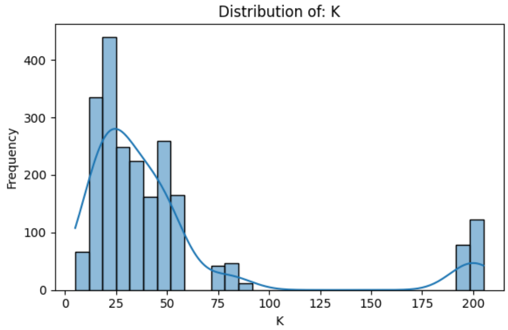
(a)



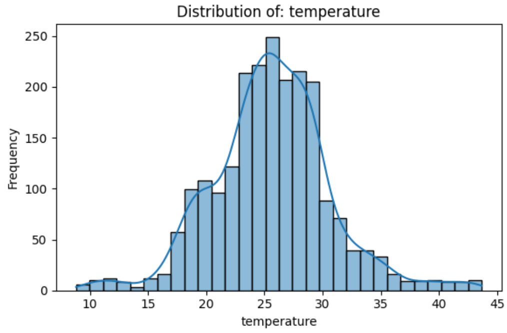
(b)



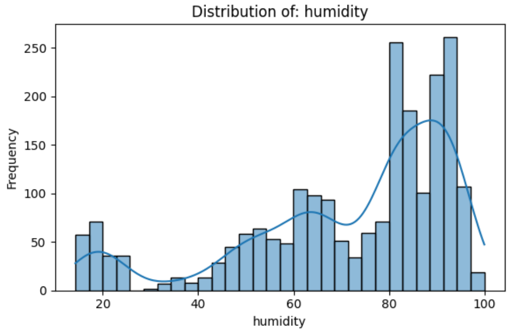
(c)



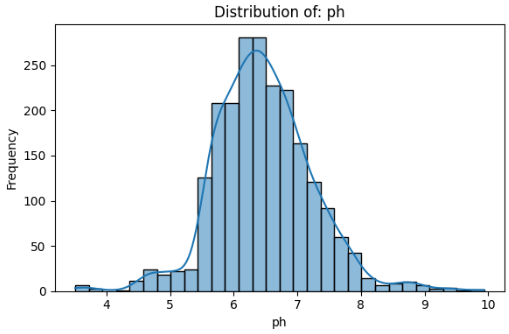
(d)



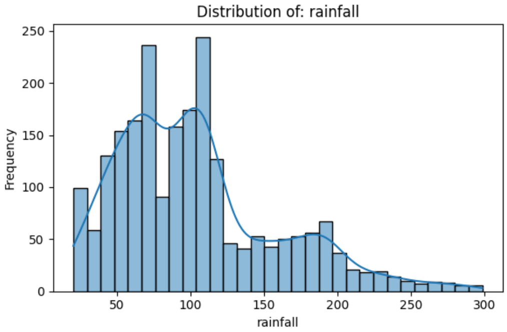
(e)



(f)



(g)



(h)

Figure 2. Dataset Explore

The relationships between the numerical features were further examined using Pearson correlation analysis. The results are visualized in the correlation heatmap presented in Figure 9. The correlation matrix serves to reveal how strongly and in which direction features are linearly associated. Based on the heatmap, it is evident that Phosphorus (P) and Potassium (K) are closely linked, demonstrating a high positive association with a correlation value of 0.74 suggesting that these two soil elements tend to increase or decrease in tandem. This strong association is consistent with their role as complementary macronutrients essential for crop growth. In contrast, most of the other feature pairs display weak or negligible correlations, such as temperature and pH (correlation ≈ -0.018), or rainfall and Potassium (correlation ≈ -0.053). This suggests low multicollinearity among the features, which is advantageous for building reliable machine learning models as it reduces redundancy in the input data.

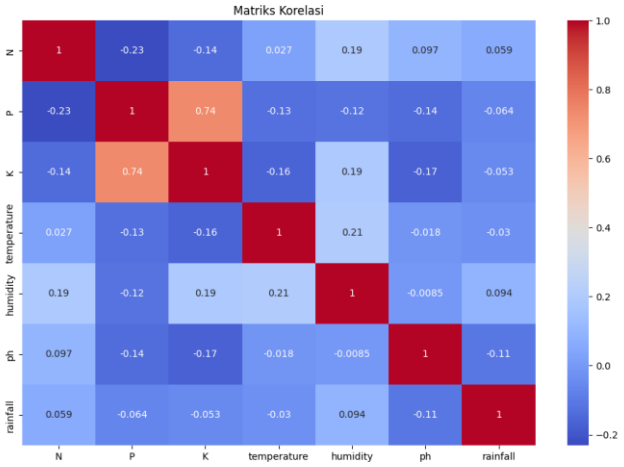


Figure 3. Correlation Heatmap

Outliers in the collected dataset were identified using the Interquartile Range (IQR) method [14]. A data point is classified as an outlier if it lies beyond the lower or upper bounds, which are calculated using the following equations:

$$UpperBound = Q3 + (1{,}5 \times IQR) \tag{1}$$

$$LowerBound = Q1 - (1{,}5 \times IQR) \tag{2}$$

Q1 and *Q3* denote the lower and upper quartiles, while the interquartile range (I*QR*) is calculated as the difference between *Q3* and *Q1*. A summary of the detected outliers both in terms of their count and proportion for each numerical attribute is presented in Table 2.

TABEL II
FEATURE OUTLIER

Feature	Outlier	Percentase (%)
Nitrogen (N)	0	0.00%
Phosphorus (P)	138	6.27%
Potassium (K)	200	9.09%
Temperature	86	3.91%
Humidity	30	1.36%
ph	57	2.59%
Rainfall	100	4.55%

The feature Potassium (K) contains the highest number of outliers, accounting for 9.09% of the data in that attribute, followed by Phosphorus (P) and Rainfall. In contrast, Nitrogen (N) shows no outliers, indicating a consistent distribution of values within the expected range. Despite the presence of outliers, they are not removed at this stage. This decision is based on the robustness of the Gradient Boosting algorithm used in this study, which is known to perform well even in the presence of extreme values. However, the impact of these outliers will be closely examined during the model evaluation phase to ensure reliability and generalization.

C. *Explore*

The Explore phase in the SEMMA methodology focuses on examining the structure, distribution, and relationships within the dataset to detect potential anomalies or irregularities before proceeding to the next stages. This phase includes identifying feature types, assessing data distribution, evaluating the presence of missing values, and visualizing data patterns through exploratory data analysis (EDA) [15]. The Crop Recommendation Dataset used in this study comprises 2,200 rows and 8 columns, which include 7 numerical input features namely *Nitrogen (N)*, *Phosphorus (P)*, *Potassium (K)*, *temperature*, *humidity*, *pH*, and *rainfall* and 1 categorical target label, *crop*. Notably, there are no missing values in any of the features, ensuring the integrity of the dataset for modeling. To better understand the distribution of each feature, histogram plots with kernel density estimations were generated, as shown in Figures 2 through 9. These visualizations reveal that several features, such as temperature and pH, exhibit a near-normal distribution, whereas others, like Nitrogen (N), Phosphorus (P), and Potassium (K), demonstrate a multimodal or skewed distribution, indicating potential variations in soil nutrient conditions. Additionally, the target variable distribution is presented in Figure 1. It shows a balanced class distribution, where each crop type has a relatively equal number of samples, approximately 100 instances per class. This balance simplifies model training by reducing the need for rebalancing techniques during the Modify phase.

D. Modify

Modify is the third stage in the SEMMA methodology. It encompasses several data preparation techniques, including the removal of outliers, data partitioning, categorical encoding, normalization, and feature selection. These steps are essential to improve model performance and generalization. The following procedures were applied during this phase:

1. Outlier Removal
   Outliers in the dataset were detected and removed using the Interquartile Range (IQR) method [14]. Any data point falling outside these bounds is considered an outlier. The number and percentage of outliers for each numerical feature are summarized in Table 2. Although several features exhibit significant outliers such as Potassium (K) and Phosphorus (P) these values are retained in the dataset, considering that Gradient Boosting is inherently robust to extreme values. Their influence will be examined during model evaluation.
2. Data Splitting
   After completing the data preprocessing stage, the dataset was partitioned into two portionsone for training the model and the other for evaluation using an 80:20 distribution ratio through the train_test_split utility from scikit-learn [16]. Stratification was not applied since the target variable represents multiple crop types with relatively balanced class distributions. This split ensures that both subsets adequately represent the full range of environmental conditions.
3. Feature Scaling (Standardization)
   All continuous variables underwent normalization through the application of the *StandardScaler* method, which transformed their distributions to a common scale. This process restructured the values so that every variable had a zero-centered mean and a unit standard deviation [17]. To avoid data leakage and maintain unbiased evaluation, the scaler was trained solely on the training subset and subsequently applied to both training and testing partitions.
4. Feature Selection
   To reduce dimensionality and emphasize relevant attributes, feature selection was performed using SelectKBest with the ANOVA F-test scoring function [18]. This method ranks features based on their statistical relevance to the target variable. Although the original dataset contains only seven numerical input features, the technique helps validate feature importance and supports future model scalability.
5. Label Encoding
   The categorical target label (*crop*) was encoded into numerical form using LabelEncoder from scikit-learn [19]. This transformation is necessary for model compatibility, as machine learning algorithms typically require numerical target values.

E. Model

The Model phase in the SEMMA methodology involves applying machine learning algorithms to the prepared dataset in order to build a predictive model. In this study, the chosen algorithm is Gradient Boosting, a powerful ensemble method known for its high accuracy, robustness to outliers, and ability to handle complex, non-linear relationships [20]. Gradient Boosting was selected due to its demonstrated effectiveness in agricultural prediction problems [21]. For instance, it has been successfully applied in precision agriculture applications.

$$-log\ L1 = -\sum_{i=1}^{N} y_i log(odds) + log(1 + e^{odds}) \quad (3)$$

Gradient Boosting represents a stepwise ensemble learning strategy where predictive models are constructed in a progressive manner [22]**.** In this approach, each new decision tree aims to minimize the shortcomings of the previous one by aligning with the direction opposite to the gradient of the loss function. This approach gradually improves the model's overall predictive accuracy. For binary classification problems, one of the most commonly used loss functions is the logistic loss, which can be mathematically expressed as follows.

F. Assess

Within the SEMMA methodology, the evaluation stage holds significant importance in determining the validity of classification models. This stage entails measuring the degree to which the predicted outputs correspond to the true class values. A commonly used performance indicator is Overall Accuracy (OA) [23], which reflects the probability that a randomly selected instance is accurately labeled by the model. Nevertheless, depending exclusively on accuracy can lead to an incomplete interpretation of the model's performance, especially when dealing with imbalanced class distributions. To achieve a more thorough performance assessment, the current research adopts supplementary metrics such as Precision, Recall, and F1-score. These indicators provide a clearer picture of how effectively the model distinguishes between classes, while also reducing the occurrence of both false positives and false negatives.

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (4)$$

$$F1 - Score = \frac{2\times TP}{2\times TP+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FN} \quad (6)$$

$$Recall = \frac{TN}{TN+FP} \quad (7)$$

III. RESULT AND DISCUSSION

A. Result

Before training the model, the dataset was preprocessed to improve its quality and reliability. The dataset was split into training and testing sets using an 80:20 ratio, where 80 percent of the data was used for training and 20 percent for evaluation.

To ensure data quality and model reliability, outlier removal was conducted during the preprocessing phase using the Interquartile Range (IQR) method. This step eliminated extreme values that could negatively affect model training. The summary of remaining outliers after the cleaning process is shown in Table 3.

TABEL III
OUTLIER REMOVER

Column Name	Number of Outliers	Percentage (%)
N	0	0.00%
P	0	0.00%
K	26	1.41%
Temperature	4	0.21%
Humidity	0	0.00%
pH	10	0.54%
Rainfall	18	0.95%

The results show that most outliers were effectively removed, with no remaining outliers in the nitrogen (N), phosphorus (P), and humidity attributes. The highest remaining outlier percentage was found in potassium (K), at 1.41%, which is still within an acceptable threshold. These outcomes confirm that the dataset is now cleaner and more reliable for robust model training.

In this study, the Gradient Boosting algorithm was employed as the primary classification model to predict the most appropriate crop seed types based on environmental and geographical data. Gradient Boosting was chosen due to its strong performance in managing structured datasets and its capability to reduce prediction errors through sequential learning. The model was trained using an 80 to 20 percent train-test split, and the hyperparameters were manually tuned to achieve an optimal balance between predictive performance and generalization. The main hyperparameters used in the model are summarized in Table 4.

TABEL IV
MODEL HYPERPARAMETER

Hyperparameter	Value	Description
n_estimators	150	Number of boosting stages (trees) to perform. More trees increase model complexity.
learning_rate	0.1	Step size shrinkage used in updating predictions. Lower values slow down learning.
max_depth	5	Maximum depth of individual trees. Controls model complexity to prevent overfitting.
subsample	0.9	Fraction of samples used for fitting each base learner. Adds randomness to reduce overfitting.
random_state	42	Seed used by the random number generator to ensure reproducible results.

The classification table in Table 5. shows that the Gradient Boosting model performs exceptionally well in classifying various crop types. Most crops, such as apple, banana, chickpea, coffee, grapes, maize, mango, musk melon, and watermelon, achieved precision, recall, and F1-score values of 1.00. A few crops, including cotton, jute, lentil, pomegranate, and blackgram, recorded slightly lower values but still above 0.90. These results indicate that the model can accurately distinguish between crop types and demonstrates strong generalization on the test data.

TABEL V
CLASSIFICATION TABLE

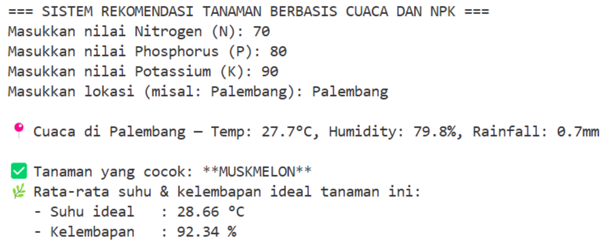
Crop Types	Precision	Recall	F1-Score
Apple	1.00	1.00	1.00
Banana	1.00	1.00	1.00
Blackgram	0.95	1.00	0.98
Chick_Pea	1.00	1.00	1.00
Coconut	0.96	0.96	0.96
Coffee	1.00	1.00	1.00
Cotton	0.94	1.00	0.97
Grapes	1.00	1.00	1.00
Jute	0.92	0.96	0.94
Kidneys_Beans	1.00	1.00	1.00
Lentil	0.92	1.00	0.96
Maize	1.00	0.95	0.98
Mango	1.00	1.00	1.00
Moth_Beans	1.00	0.96	0.98
Mung_Bean	1.00	1.00	1.00
Musk_Melon	1.00	1.00	1.00
Orange	1.00	1.00	1.00
Papaya	0.96	1,00	0.98
Pigeon_Peas	1.00	0.96	0.98
Pomegranate	1.00	0.91	0.95
Rice	0.95	0.95	0.95
Watermelon	1.00	1.00	1.00

The overall performance metrics indicate that the Gradient Boosting model achieves high reliability in predicting crop seed types. The model reached an accuracy of 0.98, with macro and weighted averages for precision, recall, and F1-score also consistently at 0.98. This demonstrates that the model performs well across all classes, including those with smaller representation, and maintains balanced performance without favoring any specific crop type. These results further confirm the model's effectiveness and suitability for real-world implementation in precision agriculture.

TABEL VI
PERFORMANCE MATRIX

	Precision	Recall	F1-Score
Accuracy			0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

The developed system generates different outputs depending on the combination of NPK values and real-time weather conditions. Three representative scenarios are presented below:



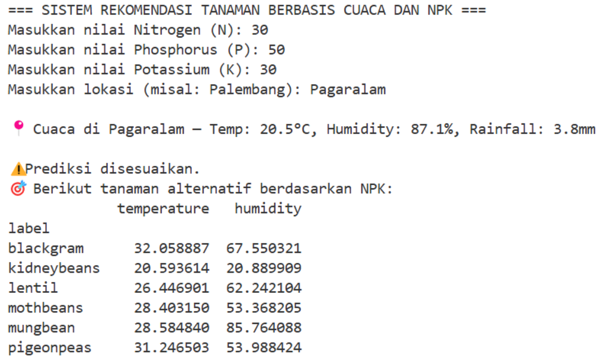
```
=== SISTEM REKOMENDASI TANAMAN BERBASIS CUACA DAN NPK ===
Masukkan nilai Nitrogen (N): 70
Masukkan nilai Phosphorus (P): 80
Masukkan nilai Potassium (K): 90
Masukkan lokasi (misal: Palembang): Palembang

📍 Cuaca di Palembang – Temp: 27.7°C, Humidity: 79.8%, Rainfall: 0.7mm

✅ Tanaman yang cocok: **MUSKMELON**
🌿 Rata-rata suhu & kelembapan ideal tanaman ini:
   - Suhu ideal    : 28.66 °C
   - Kelembapan    : 92.34 %
```

Figure 4. Result of Scenario 1

In Figure 1, the system receives input values of nitrogen, phosphorus, and potassium (70, 80, 90) along with the selected location of Palembang. The real-time weather data fetched shows a temperature of 27.6°C and humidity of 85.5%. Under these conditions, the model confidently recommends Musk Melon as the most suitable crop. The predicted crop closely aligns with its known ideal conditions (temperature of 28.66°C and humidity of 92.34%), indicating that the model performs well when both nutrient and climate factors fall within familiar patterns learned during training. This scenario illustrates the system’s strength in producing direct and reliable predictions when conditions are favorable.



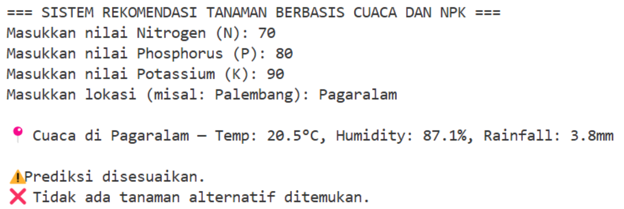
```
=== SISTEM REKOMENDASI TANAMAN BERBASIS CUACA DAN NPK ===
Masukkan nilai Nitrogen (N): 30
Masukkan nilai Phosphorus (P): 50
Masukkan nilai Potassium (K): 30
Masukkan lokasi (misal: Palembang): Pagaralam

📍 Cuaca di Pagaralam – Temp: 20.5°C, Humidity: 87.1%, Rainfall: 3.8mm

⚠️Prediksi disesuaikan.
🎯 Berikut tanaman alternatif berdasarkan NPK:
             temperature   humidity
label
blackgram      32.058887  67.550321
kidneybeans    20.593614  20.889909
lentil         26.446901  62.242104
mothbeans      28.403150  53.368205
mungbean       28.584840  85.764088
pigeonpeas     31.246503  53.988424
```

Figure 5. Result of Scenario 2

In Figure 2, the system is tested under a different setting where the input NPK values are relatively low (30, 50, 30) and the location is set to Pagaralam. The retrieved weather shows a cooler temperature of 20.7°C and high humidity of 92.7%. Under these conditions, the system is unable to find a direct match from the model, and instead adapts by suggesting several alternative crops based on NPK similarity, such as blackgram, kidneybeans, lentil, and mungbean. Although these alternatives may not exactly fit the current climate, this output demonstrates the system’s flexibility in adjusting recommendations by prioritizing nutrient compatibility. This capability ensures that users are still guided with meaningful suggestions even when full match conditions are not met.



```
=== SISTEM REKOMENDASI TANAMAN BERBASIS CUACA DAN NPK ===
Masukkan nilai Nitrogen (N): 70
Masukkan nilai Phosphorus (P): 80
Masukkan nilai Potassium (K): 90
Masukkan lokasi (misal: Palembang): Pagaralam

📍 Cuaca di Pagaralam – Temp: 20.5°C, Humidity: 87.1%, Rainfall: 3.8mm

⚠️Prediksi disesuaikan.
❌ Tidak ada tanaman alternatif ditemukan.
```

Figure 6. Result of Scenario 3

In Figure 3, the same location (Pagaralam) is used but with higher nutrient input values (70, 80, 90). Despite the soil being nutrient-rich, the environmental conditions remain unfavorable, particularly the low temperature and excessive humidity, which results in no suitable crop being found. The system displays a warning and refrains from forcing an inaccurate prediction. This behavior underscores the robustness of the model and its ability to reject unreliable outputs, ensuring that recommendations are only provided when the model has sufficient confidence based on both climate and nutrient features.

B. Analysis

The results demonstrate that the Gradient Boosting model is capable of delivering highly accurate crop seed recommendations based on a combination of environmental and geographical parameters. The high accuracy of 98%, along with strong precision, recall, and F1-scores across almost all crop types, indicates that the model successfully captures complex relationships in the dataset.

The system's performance across three output scenarios highlights its practical adaptability. In the first scenario, the model confidently provides a direct crop recommendation that aligns with both NPK values and current weather conditions, showcasing the effectiveness of the integrated real-time environmental input. In contrast, the second scenario illustrates the model's ability to adjust and provide alternatives when the environmental conditions deviate from optimal thresholds. This flexibility is crucial in real-world agricultural applications, where ideal conditions are not always present. The third scenario further reinforces the model's reliability by rejecting unsuitable predictions rather than producing inaccurate suggestions, thereby ensuring responsible decision-making.

The analysis also reveals that weather data plays a pivotal role in refining predictions. Even when NPK values match those in the training data, the model avoids making recommendations if environmental parameters fall outside the learned tolerances. This suggests that incorporating dynamic climate data enhances the model’s robustness and makes it more context-aware.

The recommendation results have also been validated using Feature Importance, which describes the relative contribution of each input variable to the overall model decision. This technique calculates the increase in accuracy (gain) when a feature is used in the decision tree. In the XGBoost model, feature importance can be obtained using Python:

xgb.plot_importance(model, importance_type='gain')
Results:

TABEL VII
FEATURE IMPORTANCE

Feature	Gain Importance (%)
Annual rainfall	34.2
Average temperature	26.5
Location altitude	18.9
Soil pH	12.7
Soil texture	7.7

From table VII, it can be concluded that rainfall and temperature are the main determinants in plant seed recommendations.

Nevertheless, some limitations are observed. Certain crops with lower representation in the dataset may slightly underperform, as shown by precision or recall values below 1.00. Additionally, the system relies on the accuracy of external weather data sources, which could affect performance in areas with limited or inconsistent data availability.

However, analysis bias can occur if the training data is dominated by a particular region (e.g., lowlands with high rainfall), so that the feature distribution does not represent all geographic conditions. As a result, the model may be overfitted to local environmental patterns and less accurate in areas with different conditions.

Overall, the model's ability to make direct predictions, propose alternatives, or reject unsuitable inputs shows strong potential for implementation in smart farming systems. Its decision-making logic, grounded in both soil nutrient levels and real-time environmental factors, aligns well with the principles of precision agriculture.

IV. CONCLUSION

This study successfully developed a smart recommendation system for crop seed selection using the Gradient Boosting algorithm, leveraging both environmental and geospatial data to provide recommendations for farmers. The research addressed the critical challenge of matching crop varieties to specific local conditions, a task that is often hindered by reliance on generalized advice or personal intuition. By utilizing a comprehensive dataset containing key soil nutrients and environmental parameters, the system was able to process and analyze complex, multivariate data relevant to crop growth.

Exploratory data analysis confirmed the dataset's integrity and revealed meaningful relationships among features, while robust preprocessing steps, including outlier detection, feature scaling, and label encoding, further enhanced model reliability. The Gradient Boosting algorithm proved effective, given its resilience to outliers and ability to model non-linear relationships, making it well-suited for the complexities of agricultural data. A set of performance indicators including accuracy, precision, recall, and the F1-score was employed to thoroughly evaluate the model's effectiveness, confirming its relevance and applicability in real-life seed selection contexts. The balanced class distribution in the dataset simplified training and minimized the risk of bias toward any particular crop type. The system demonstrates significant potential to support precision agriculture by delivering accurate, context-specific crop recommendations. This method enhances the efficiency of resource utilization and promotes better crop productivity, while simultaneously lowering the likelihood of failure caused by incompatibility with environmental conditions. Moreover, the study supports the progression of intelligent agricultural technologies and establishes a foundation for future incorporation into more extensive decision-support systems in farming. The technical initial steps are to design using components in the form of sensors (soil moisture, temperature, pH, light intensity, rainfall), actuator modules (irrigation pumps), and multispectral drone devices to collect agricultural environmental data periodically and in real-time. Mobile app development for real-time monitoring, early warnings (notifications), and irrigation/fertilizer recommendations for field officers and farmers.

REFERENCES

[1] K. Pawlak and M. Kołodziejczak, "The Role of Agriculture in Ensuring Food Security in Developing Countries: Considerations in the Context of the Problem of Sustainable Food Production," *Sustainability 2020, Vol. 12, Page 5488*, vol. 12, no. 13, p. 5488, Jul. 2020, doi: 10.3390/SU12135488.

[2] R. Byaruhanga and E. Isgren, "Rethinking the Alternatives: Food Sovereignty as a Prerequisite for Sustainable Food Security," *Food Ethics*, vol. 8, no. 2, pp. 1–20, Oct. 2023, doi: 10.1007/S41055-023-00126-6/METRICS.

[3] D. M. K. S. Hemathilake and D. M. C. C. Gunathilake, "Agricultural productivity and food supply to meet increased demands," *Future Foods: Global Trends, Opportunities, and Sustainability Challenges*, pp. 539–553, Jan. 2022, doi: 10.1016/B978-0-323-91001-9.00016-5.

[4] A. Cravero, S. Pardo, P. Galeas, J. López Fenner, and M. Caniupán, "Data Type and Data Sources for Agricultural Big Data and Machine Learning," *Sustainability 2022, Vol. 14, Page 16131*, vol. 14, no. 23, p. 16131, Dec. 2022, doi: 10.3390/SU142316131.

[5] M. F. Kanazoe, A. Keïta, D. Yamegueu, Y. Konate, B. Sawadogo, and B. Boube, "Integrating Fish Farming into Runoff Water Harvesting Ponds (RWHP) for Sustainable Agriculture and Food Security: Farmers' Perceptions and Opportunities in Burkina Faso," *Sustainability 2025, Vol. 17, Page 880*, vol. 17, no. 3, p. 880, Jan. 2025, doi: 10.3390/SU17030880.

[6] S. O. Araújo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata, "Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives," *Agronomy 2023, Vol. 13, Page 2976*, vol. 13, no. 12, p. 2976, Dec. 2023, doi: 10.3390/AGRONOMY13122976.

[7] A. Haleem, M. Javaid, M. Asim Qadri, R. Pratap Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literature-based study," *International Journal of Intelligent Networks*, vol. 3, pp. 119–132, Jan. 2022, doi: 10.1016/J.IJIN.2022.08.005.

[8] I. Attri, L. K. Awasthi, and T. P. Sharma, "Machine learning in agriculture: a review of crop management applications," *Multimed Tools Appl*, vol. 83, no. 5, pp. 12875–12915, Feb. 2024, doi: 10.1007/S11042-023-16105-2/METRICS.

[9] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone, "Towards leveraging the role of machine learning and artificial intelligence

in precision agriculture and smart farming," *Comput Electron Agric*, vol. 198, p. 107119, Jul. 2022, doi: 10.1016/J.COMPAG.2022.107119.

[10] M. P. Sulanggalih, S. R. Nudin, and R. A. J. Firdaus, "Sistem Klasifikasi Tingkat Kesesuaian Bibit Dan Pupuk Dengan Algoritma C4.5 Berbasis Website (Studi Kasus : Kecamatan Megaluh)," *Inovate : Jurnal Ilmiah Inovasi Teknologi Informasi*, vol. 6, no. 1, pp. 28–37, Sep. 2021, doi: 10.33752/INOVATE.V6I1.3159.

[11] M. S. Islam Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4773–4781, Sep. 2022, doi: 10.1016/J.JKSUCI.2021.06.003.

[12] O. Firas, "A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach ," *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, 2023.

[13] S. Sharma, "Crop Recommendation Dataset." Accessed: Jul. 02, 2025. [Online]. Available: https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset

[14] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/J.DAJOUR.2023.100164.

[15] J. Peng *et al.*, "DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 2271–2280, 2021.

[16] A. Testas, "Support Vector Machine Classification with Pandas, Scikit-Learn, and PySpark," *Distributed Machine Learning with PySpark*, pp. 259–280, 2023, doi: 10.1007/978-1-4842-9751-3_10.

[17] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics 2022, Vol. 10, Page 1942*, vol. 10, no. 11, p. 1942, Jun. 2022, doi: 10.3390/MATH10111942.

[18] S. Abdumalikov, J. Kim, and Y. Yoon, "Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification," *Applied Sciences 2024, Vol. 14, Page 10511*, vol. 14, no. 22, p. 10511, Nov. 2024, doi: 10.3390/APP142210511.

[19] A. Zollanvari, "Supervised Learning in Practice: the First Application Using Scikit-Learn," *Machine Learning with Python*, pp. 111–131, 2023, doi: 10.1007/978-3-031-33342-2_4.

[20] P. Trizoglou, X. Liu, and Z. Lin, "Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines," *Renew Energy*, vol. 179, pp. 945–962, Dec. 2021, doi: 10.1016/J.RENENE.2021.07.085.

[21] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, Apr. 2023, doi: 10.1002/eng2.12599.

[22] A. Thakur *et al.*, "Product Length Predictions with Machine Learning: An Integrated Approach Using Extreme Gradient Boosting," *SN Comput Sci*, vol. 5, Aug. 2024, doi: 10.1007/s42979-024-02999-8.

[23] C. Y. Chang and D. H. S. Silverman, "Accuracy of early diagnosis and its impact on the management and course of Alzheimer's disease," in *Expert Review of Molecular Diagnostics*, Jan. 2004, pp. 63–69. doi: 10.1586/14737159.4.1.63.