

Optimizing Seed Selection Using Random Forest for Sustainable Agriculture: Using Weather and Soil Data

Bambang Pratama*, Khairunnisa, Aqilla Khairunnisya

Sriwijaya State Polytechnic, Palembang

*Corresponding author: bprtmma@gmail.com, nisaxica@gmail.com, aqillakhrnsy@gmail.com

Abstract

Optimizing seed selection is essential for managing issues with unpredictable weather patterns and a variety of soil types, both of which are crucial for sustainable agriculture especially in Indonesia. Machine learning techniques, particularly ensemble methods, have gained prominence in agricultural solutions for their predictive reliability. This study introduces a web-based seed selection tool designed for Indonesian farmers, using information data from Visual Crossing Weather Data and manually inputted soil type. A predictive model was developed using the Random Forest method, a method of ensemble learning that is well-known for improving accuracy and reducing the danger of overfitting. The system achieved a notable accuracy of 87.05%, offering reliable recommendations for seed selection based on environmental and soil types. The platform's web-based design ensures accessibility and eliminating the need for large storage capacities. This research highlights the effectiveness of integrating Random Forest with diverse datasets to address agricultural challenges. The system encourages sustainable farming methods and helps to improve agricultural productivity and resource management in an economical and user-friendly way by equipping farmers with machine learning.

Keywords: Machine learning, Random forest, Seed selection, Sustainable agriculture.

INTRODUCTION

According to a recent analysis by UNESCO's World Water Assessment Program (WWAP), the global population is projected to grow by 33% by 2050, which will significantly increase the demand for food and water, effectively doubling their current needs (Roser, 2020). Moreover, according to the Food and Agriculture Organization of the United Nations (FAO), meeting the needs of this growing population will require a nearly 70% increase in food production rates. This challenge is further exacerbated by mounting competition and intensified exploitation of land, water, and other critical natural resources. These pressing concerns underscore the urgent need for innovative, data-driven approaches to enhance agricultural efficiency and sustainability (Gupta *et al.*, 2020). With the growing global demand for food, there is increasing pressure to achieve both higher crop yields and better quality. To address this, strategies are being implemented to reduce post-harvest losses, enhance supply chain efficiency, and minimize food waste across different stages of production and distribution. Smart farming technologies and precision agriculture are gaining significant attention due to their potential to address the growing global demand for food and resources (Stafford, 2019).

Studies have shown that machine learning techniques like Random Forest are increasingly applied in agricultural settings, enhancing productivity and sustainability. However, despite these advancements, research gaps remain in how effectively these technologies can be tailored to specific agricultural contexts, such as different soil types and unpredictable weather patterns. Some of the key research gaps identified include the lack of real-time data integration for crop recommendations, limited exploration of additional soil parameters for predictions, insufficient

validation across diverse geographical regions, the need for user-friendly interfaces for farmers, and limited focus on economic impact analysis of recommendations (Varshitha & Choudhary, 2022). This paper aims to address these research gaps by presenting a machine learning-based seed selection model that integrates real-time weather data, incorporates additional soil types for more precise predictions, and utilizes a user-friendly web interface for farmers. These innovations are intended to offer practical solutions to local agricultural challenges while promoting sustainable farming practices. By leveraging real-time data, this study proposes a more adaptive and effective model for seed selection and crop recommendations, ultimately contributing to the enhancement of agricultural productivity and sustainability.

LITERATURE REVIEW

a. Machine Learning in Agriculture

Creating computing algorithms that can recognize patterns and produce data-driven predictions or judgments is the main goal of the specialist field of machine learning within artificial intelligence (AI). By using structured or unstructured data to train models, this method enables systems to iteratively increase their accuracy and generalization skills. We offer our knowledge in implementing machine learning algorithms, but in order to properly and efficiently direct analysis, the data providers must supply domain information. Since machine learning is not a "one-shot" process, we typically need to work closely with the people who collected the data (Amini & Rahmani, 2023). Machine learning, enabled by big data technologies and high-performance computers, has opened up new avenues for data-intensive study in the interdisciplinary field of agriculture technology (Liokas *et al.*, 2018).

b. Precision Agriculture

A lot of accurate information about a particular region is necessary for precision agriculture (PA), which enables the utilization of that information in line with agronomic principles. It allows farmers more control over existing practices, including the use of chemicals, crop placement, and soil conditions. However, integrating precision livestock solutions into these growing systems can save waste and expenses while bringing animals closer to producers (Žetrović *et al.*, 2024). In essence, precision agriculture uses a focused and strategic approach to increase total agricultural productivity, reduce waste, and maximize efficiency (Zang, 2016).

As the field progressed, the concept expanded to incorporate more advanced technologies, enabling farmers to make data-driven decisions that improved efficiency and sustainability. In the context of the paper at hand, the optimization of seed selection is an evolution of this precision-driven approach. By integrating weather data and soil types, the aim is to enhance the precision of agricultural practices, providing farmers with tailored recommendations for seed selection based on real-time environmental factors. This shift towards using diverse datasets for decision-making is in line with the ongoing transformation of Precision Agriculture into Smart Farming, where the focus is on real-time data and collaborative decision-making across multiple stakeholders.

c. Random Forest

Breiman developed the Random Forest ensemble learning method to solve problems with classification and regression. A machine learning technique called "ensemble learning" aims to increase accuracy by integrating several models to address a single issue. Specifically, "multiple

classifiers" use ensemble classification to produce more accurate results than using a single classifier. This integration can produce more dependable results by lowering variance, especially for unstable classifiers. Next, unlabeled samples are given labels using a "voting scenario" (Liokas *et al.*, 2018; Fawagreh, 2014; Mohammed & Kora).

The trees in the ensemble have different overfitting cases and outliers as a result of the different tree architectures and splitting variables that RFs produce. In order to obtain the "global optimum," about two-thirds of the data are used to train the trees. The remaining samples, known as "out-of-bag (OOB)" samples, independently validate the RF model. The "OOB error" is utilized to evaluate "variable importance measures (VIMs)" and compute prediction error (Ogregon & Jung, 2022; akar & Güngör, 2015). "Variable importance metrics (VIM)" (Song *et al.*, 2021; Stafford, 2019; Varshitha & Choudhary, 2022; Wei & Song, 2015) are a prominent characteristic of RFs. Predictor variables are assessed and ranked by VIM according to their relative importance. It looks at sample proximities to find outliers in training samples or employs internal proximity matrix measurements. "Mean decrease in accuracy (MDA)" and "mean decrease in Gini (MDG)" are the two main VIMs (Behnamian *et al.*, 2017; Belgiu & Drăgu, 2016; Berhane *et al.*, 2016; Fawagreh 2014; Gronski *et al.*, 2014).

Two parameters must be provided for RF models: "Ntree" and "Mtry." In contrast to Mtry, RFs are less responsive to Ntree (Berhane *et al.*, 2016). Studies often find 500 trees to be optimal for Ntree, as higher values rarely improve accuracy (Belgiu & Drăgu, 2016). However, the dataset determines the ideal value for Mtry. Mtry is advised to be one-third of the input features for regression and the square root of the number of input features for classification tasks (Zang, 2016). However, when dealing with noisy predictors, smaller Mtry values may lead to inaccurate trees (Belgiu & Drăgu, 2016).

In this study, the Random Forest method is implemented following the concepts and configurations outlined above. A model consisting of 100 trees is utilized to achieve an optimal balance between computational efficiency and predictive accuracy, adhering to the recommended parameter settings to maximize performance.

RESEARCH METHOD

The research methodology for developing a crop recommendation system based on machine learning consisted of five interconnected stage.

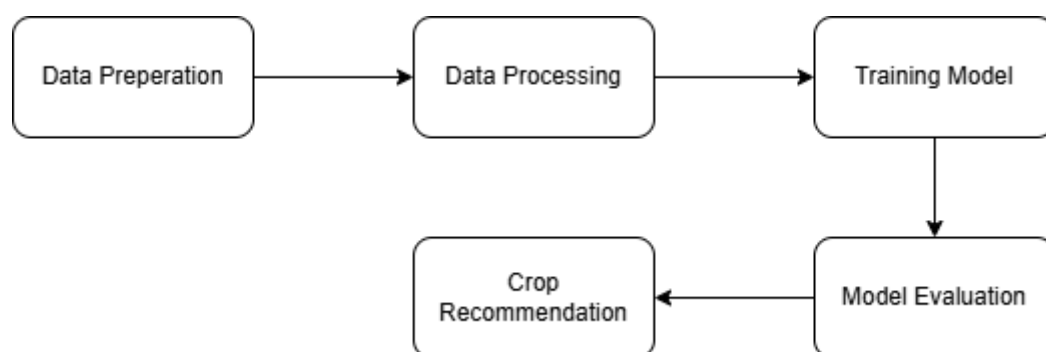


Figure 1. Stage of research method

The process began with data preparation and data processing. The dataset was imported from a CSV file using the following code:

```
dataset = pd.read_csv('Dataset.csv')
```

Key predictor variables such as temperature, humidity, and soil type were identified and selected for further analysis:

```
X = dataset[['Temperature', 'Humidity', 'Jenis_Tanah']].
```

To handle categorical data effectively, the soil type variable (Jenis_Tanah) was converted into a numerical representation using the one-hot encoding technique:

```
X = pd.get_dummies(X, columns=['Jenis_Tanah'], drop_first=True)
```

This transformation ensured that the machine learning algorithm could process categorical data efficiently. Subsequently, The dataset was separated into both testing and training subgroups using the `train_test_split()` method in an 20:80 ratio. This division ensured independent validation of the model during the evaluation phase. In the data modeling stage, a Random Forest Classifier was chosen and implemented with optimized hyperparameters. The configuration of the model is as follows:

Table 4. Hyperparameters

Model	Parameters
RandomForestClassifier	n_estimators=100, max_depth=15, max_features='log2', min_samples_leaf=4, min_samples_split=4, criterion='entropy', n_jobs=-1, random_state=42

Each parameter was strategically chosen to optimize the model's performance. For instance, the `n_estimators` parameter specified the number of decision trees, enhancing robustness. The `max_depth` parameter limited tree complexity to prevent overfitting, while the `criterion='entropy'` parameter ensured that splits maximized information gain. Additionally, setting `n_jobs=-1` enabled parallel processing, significantly reducing computation time.

A number of indicators were used to assess the model's performance. The model's overall accuracy was assessed using the `accuracy_score()` function:

```
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
```

Further insights were gained using the `classification_report()` function, which provided detailed precision, recall, and F1 scores for each crop type:

```
classification_report(y_test, y_pred)
```

These metrics facilitated a comprehensive evaluation of the model's performance, highlighting its strengths and identifying areas for improvement. An innovative feature of this research was the integration of real-time weather data using the Visual Crossing Weather API. The `get_weather_data()` function fetched weather conditions based on geographical coordinates, converted temperatures from Fahrenheit to Celsius, and returned actual weather data for real-time crop recommendations.

The core recommendation logic was implemented through two functions: `rekomendasi_tanaman` and `get_alternative_crops`. These functions intelligently filtered data based on soil type, matched crops to temperature and humidity conditions, and provided alternative crop suggestions when exact matches were unavailable. The system was deployed as a web application using Flask. The `/rekomendasi` route allowed users to input latitude, longitude, and soil type to receive dynamic crop recommendations based on their environmental conditions. This methodological approach resulted in an adaptive machine learning system capable of delivering crop recommendations based on a comprehensive analysis of specific environmental parameters. The system not only predicted the most suitable crops but also provided valuable insights into optimal growth conditions.

The system's main advantages include:

- a. Leveraging machine learning for accurate crop predictions.
- b. Integrating real-time weather data for dynamic recommendations.
- c. Flexibility in handling various soil types.
- d. The ability to suggest alternative crops when ideal conditions are unmet.

This study makes a substantial contribution to the advancement of intelligent agricultural technology, enabling farmers to make informed choices that are suited to their unique environmental circumstances. This methodological framework emphasizes the integration of machine learning with real-world agricultural challenges, ensuring that the developed system is practical, interpretable, and user-friendly for farmers.

RESULTS AND DISCUSSION

This study successfully implemented the stages of preparation, modeling, and evaluation using the Random Forest algorithm. In the preparation phase, data cleaning was conducted to remove rows with null values, and data formatting was performed to convert numeric data types into float format, including changing the format of numbers from commas to dots. These processes ensured that the dataset was structured and ready for modeling.

The modeling phase involved dividing the dataset into 80% training data and 20% testing data, following the formula below:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The `train_test_split` function from the scikit-learn library was used to randomly divide the dataset into two subsets: testing data for accuracy evaluation and training data for model creation. According to the `test_size=0.2` parameter, 80% of the dataset is used for training and 20% is kept aside for testing. Deterministic splitting is guaranteed with the `random_state=42` argument, making the experiment repeatable.

Table 2. Classification report

Crop Types	Precision	Recall	F1-Score
Apple	1.00	1.00	1.00
Banana	0.90	0.86	0.88
Blackgram	0.86	0.95	0.90
Chick_Pea	1.00	1.00	1.00
Coconut	0.84	1.00	1.00
Coffee	0.65	0.76	0.62
Cotton	0.52	0.76	0.70
Grapes	0.91	0.71	0.80
Jute	0.81	0.57	0.67
Kidneys_Beans	1.00	1.00	1.00
Lentil	0.67	0.73	0.70
Maize	1.00	1.00	1.00
Mango	0.90	0.95	0.92
Moth_Beans	0.88	0.62	0.73
Mung_Bean	1.00	1.00	1.00
Musk_Melon	0.94	1.00	0.97
Orange	0.91	0.71	0.80
Papaya	1.00	0.96	0.98
Pigeon_Peas	1.00	1.00	1.00
Pomegranate	0.95	0.87	0.91
Rice	0.82	0.95	0.88
Watermelon	0.89	0.89	0.89

The model's accuracy was 88.64% after it was trained using the Random Forest technique. Accuracy was assessed using metrics such as precision, recall, and F1-score for each class. The model's performance across many classes is highlighted in the classification report, which is displayed in the image above. For instance, the "Apple" class had perfect predictions with an F1-score of 1.00, precision, and recall. Nevertheless, several classifications, including "Cotton" and "Coffee," performed somewhat worse, indicating that the model could use some additional work to better handle these categories.

Table 3. Report average

	Precision	Recall	F1-Score
Accuracy			0.89
Macro Avg	0.88	0.88	0.88
Weighted Avg	0.90	0.89	0.89

Additionally, the graphic displays averaged metrics, such as the weighted average and macro average, which have respective precision, recall, and F1-scores of 0.88 and 0.89. These values indicate the model's overall robust predictive performance.

The results demonstrate that the Random Forest algorithm's capacity to handle complicated data and lower the danger of overfitting makes it an appropriate option for managing datasets with numerous categories. With an accuracy of 88.64%, the model proves reliable for data-driven analysis

and decision-making in this study's context. The modeling process also yielded results showcased in the website interface, which demonstrates the system's functionality in two scenarios:

Rekomendasi Tanaman

Gunakan Lokasi GPS Saya

Latitude: -2.983404

Longitude: 104.733532

Jenis Tanah: Andisol

Dapatkan Rekomendasi

Hasil Rekomendasi

Suhu: 27.133°C

Kelembapan: 79%

Jenis Tanah: Andisol

Rekomendasi Tanaman: Banana

Figure 2. Crop recommendation for matching conditions

Figure 2 illustrates the website's display when the input data, including soil type, temperature, and humidity, match the available dataset. In this scenario, the system provides precise crop recommendations based on the matching environmental conditions.

Rekomendasi Tanaman

Gunakan Lokasi GPS Saya

Latitude: -2.983404

Longitude: 104.733532

Jenis Tanah: Lempung

Dapatkan Rekomendasi

Tidak Ada Rekomendasi Tepat

Suhu: 27.133°C

Kelembapan: 79%

Jenis Tanah: Lempung

Tanaman Alternatif

Maize

Suhu: 18.283622°C - 26.074701°C

Kelembapan: 55.647099% - 74.829137%

Figure 3. Alternative crop recommendations for non-matching conditions

Figure 3 presents the website's display when the input data, such as soil type, temperature, and humidity, do not align with the dataset. In this case, the system intelligently suggests alternative crops that are suitable for the given soil type while recommending optimal temperature and humidity conditions for planting those crops. These visual results emphasize the system's adaptability in handling both exact matches and non-matching environmental conditions, ensuring that users receive meaningful crop recommendations even under suboptimal scenarios.

CONCLUSION

This study demonstrates that implementing machine learning for crop recommendations using classification algorithms can be effectively achieved through the normalization of agricultural data and the application of Random Forest modeling techniques. The classification approach in this research yielded highly accurate results which 88.64%, enabling the model to predict suitable crop types based on specific weather conditions and soil types.

The system provides a valuable tool for farmers and agricultural decision-makers, allowing them to simulate and strategize planting decisions for specific periods. By leveraging data-driven insights into weather, temperature, humidity, and soil type, this approach aligns with ongoing efforts toward digital transformation in agriculture. The system's capability extends beyond primary crop recommendations by offering alternative crop suggestions when environmental conditions are suboptimal. This flexibility underscores the depth of machine learning analysis, ensuring comprehensive solutions tailored to specific parameters for the benefit of farmers.

Future research can expand on this work by incorporating additional environmental variables, socio-economic factors, and the integration of satellite or sensor data. Further exploration of regression algorithms or forecasting techniques could also facilitate predictions of crop yields and plant growth conditions, supporting sustainable agricultural practices. This research is important because it has the potential to replace conventional techniques of agricultural decision-making with data-driven, intelligent methods. This transition promises enhanced productivity and efficiency in the agricultural sector, paving the way for smarter, more adaptive farming practices.

ACKNOWLEDGEMENT

This research was made possible through the support and guidance of Mr. Mohammad Fadhli, S. Pd., M.T., Head of the D4 Telecommunication Engineering program at Politeknik Negeri Sriwijaya (Polsri). We sincerely thank him for his invaluable insights and encouragement throughout this study. Our gratitude also extends to Mrs. Lindawati, S.T., M.T.I., for her constructive feedback and support during the research process.

REFERENCES

- Akar, Ö., & Güngör, O. (2015). Integrating multiple texture methods and NDVI to the random forest classification algorithm to detect tea and hazelnut plantation areas in northeast Turkey. *International Journal of Remote Sensing*, 36(2), 442–464.
- Amini, M., & Rahmani, A. (2023). Agricultural databases evaluation with machine learning procedure. *Australian Journal of Engineering and Applied Science*, 8(6).
- Banks, S., *et al.* (2019). Wetland classification with multi-angle/temporal SAR using random forests. *Remote Sensing*, 11(6), Art. no. 670.
- Bassa, Z., Bob, U., Szantoi, Z., & Ismail, R. (2016). Land cover and land use mapping of the iSimangaliso Wetland Park, South Africa: Comparison of oblique and orthogonal random forest algorithms. *Journal of Applied Remote Sensing*, 10(1), Art. no. 015017.
- Behnamian, A., Millard, K., Banks, S. N., White, L., Richardson, M., & Pasher, J. (2017). A systematic approach for variable selection with random forests: Achieving stable variable importance values. *IEEE Geoscience and Remote Sensing Letters*, 14(11), 1988–1992.
- Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.

- Berhane, T., *et al.* (2018). Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. *Remote Sensing*, 10(4), Art. no. 580.
- Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1), 602–609
- Gromski, P. S., Xu, Y., Correa, E., Ellis, D. I., Turner, M. L., & Goodacre, R. (2014). A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Analytica Chimica Acta*, 829, 1–8.
- Gupta, M., Abdelsalam, M., Khorsandroo, S., & Mittal, S. (2020). Security and privacy in smart farming: Challenges and opportunities. *IEEE Access*, 8, 34564–34584.
- Liakos, K., Busato, P., Moshu, D., Pearso, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *MDPI Agriculture*, 18.
- Mohammed, A., & Kora, R. (n.d.). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*. Advance online publication.
- Obregon, J., & Jung, J.-Y. (2022). Explanation of ensemble models. In *Human-Centered Artificial Intelligence: Research and Applications* (pp. 51–72). Elsevier.
- Roser, M. (2020). Future population growth. *Our World in Data*. Retrieved from <https://ourworldindata.org/future-population-growth>
- Song, S., He, R., Shi, Z., & Zhang, W. (2021). Variable importance measure system based on advanced random forest. *CMES - Computer Modeling in Engineering and Sciences*, 128(1), 65–85.
- Stafford, J. V. (2019). *Precision Agriculture 19*. Wageningen, The Netherlands: Academic Press.
- Varshitha, D. N., & Choudhary, S. (2022). An artificial intelligence solution for crop recommendation. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(3), 1688–1695.
- Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 142, 399–432.
- Zhang, Q. (2016). *Precision Agriculture Technology for Crop Farming* (p. 374). Taylor & Francis.
- Žetrović, B., Bumbálek, R., Zoubek, T., Kuneš, R., Smutný, L., & Bartoš, P. (2024). Application of precision agriculture technologies in Central Europe: Review. *Journal of Agriculture and Food Research*, 15, 101048.