Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i2.167

Research Article

Predicting Plant Growth Stages Using Random Forest Classifier: A Machine Learning Approach

Ilham [1,*]

[1] Universitas DIPA Makassar, Makassar, Indonesia, ilhamaswan34@gmail.com

Correspondence should be addressed to Ilham; ilhamaswan34@gmail.com

Received 05 June 2024; Accepted 28 June 2024; Published 31 July 2024

Abstract:

The optimization of plant growth through predictive modelling is a crucial aspect of modern agricultural practices. This study investigates the application of a Random Forest Classifier to predict plant growth stages based on various environmental and management factors. The dataset, sourced from Kaggle, includes variables such as soil type, sunlight hours, water frequency, fertilizer type, temperature, and humidity. The research involves extensive data pre-processing, including encoding categorical variables, scaling data, and splitting it into training (80%) and testing (20%) sets. The Random Forest Classifier is implemented with 5-fold cross-validation, and its performance is evaluated using accuracy, precision, recall, and F1-score metrics. The model exhibits robust performance with an average accuracy of 84.27%, precision of 85.59%, recall of 84.27%, and F1-score of 83.98%. Visualization techniques such as correlation heatmaps, PCA plots, t-SNE plots, and violin plots are used to provide insights into the data structure and feature relationships. The results confirm the hypothesis that machine learning can effectively predict plant growth stages, offering significant implications for precision agriculture. By accurately identifying growth stages, farmers and greenhouse managers can optimize resource allocation and management practices, leading to enhanced crop yields and sustainability. The study's limitations include the specificity of the dataset and the sole use of the Random Forest Classifier. Future research should explore additional machine learning models and incorporate more diverse datasets to improve generalizability. The findings contribute to the growing body of knowledge on the application of machine learning in agriculture and suggest practical applications for improving agricultural productivity.

Keywords: Machine Learning, Plant Growth, Random Forest, Precision Agriculture, Environmental Factors
Dataset link: https://www.kaggle.com/datasets/humairmunir/anaemia-prediction

1. Introduction

Efficient agricultural practices are essential for optimizing crop yields and ensuring sustainable food production. One of the critical aspects of agriculture is understanding how various environmental and management factors influence plant growth. These factors include soil type, sunlight exposure, watering schedules, fertilizer usage, temperature, and humidity. By comprehensively analysing these elements, farmers and greenhouse managers can make informed decisions to enhance plant growth and productivity. Recent advancements in machine learning have provided powerful tools for predicting outcomes based on complex datasets, offering new opportunities for precision agriculture. The primary problem addressed in this research is the need for accurate prediction models that can classify the growth milestones of plants based on environmental and management variables. Traditional methods of monitoring plant growth often rely on manual observations and generalized assumptions, which can be time-consuming and imprecise. There is a significant demand for automated, data-driven approaches that can provide reliable predictions, thereby assisting in better planning and resource allocation in agricultural practices. This study aims to fill this gap by leveraging machine learning techniques, specifically the Random Forest Classifier, to develop a robust prediction model.

The objectives of this research are multifaceted. First, the study aims to pre-process the dataset by encoding categorical variables and scaling the data to ensure uniformity. Next, the research will implement a Random Forest Classifier, utilizing 5-fold cross-validation to ensure the reliability and validity of the model. The performance of the classifier will be assessed using various metrics, including accuracy, precision, recall, and F1-measure [1]–[4]. Through these objectives, the study seeks to demonstrate the applicability of machine learning in predicting plant growth stages and to provide actionable insights for agricultural optimization. This research is guided by several key questions and hypotheses. The primary research question is whether environmental and management factors can accurately predict plant growth milestones using machine learning [5]–[7]. A related hypothesis is that the Random Forest Classifier will exhibit high performance in this prediction task due to its ability to handle diverse and complex datasets [8]. Additionally, the study explores the relative importance of different factors in influencing plant growth, hypothesizing that certain variables, such as soil type and watering frequency, will have more significant impacts than others.

The scope of this research is confined to the provided dataset from Kaggle, which includes various environmental and management factors relevant to plant growth. While the dataset offers a comprehensive view of these factors, the study is limited by its specificity and may not account for all possible variables influencing plant growth. Furthermore, the research is constrained by the methods chosen, primarily focusing on the Random Forest Classifier, and does not explore other potential machine learning models [9], [10]. These limitations highlight the need for cautious interpretation of the results and suggest areas for further investigation. Despite its limitations, this research makes several notable contributions to the field of precision agriculture. By demonstrating the effectiveness of machine learning in predicting plant growth milestones, the study provides a valuable framework for future research and practical applications. The insights gained from the analysis of environmental and management factors can help optimize agricultural practices, leading to improved crop yields and resource efficiency. Additionally, the research contributes to the growing body of knowledge on the application of machine learning in agriculture, offering a foundation for subsequent studies to build upon and expand.

2. Method:

This research employs a quantitative design utilizing machine learning techniques to predict plant growth stages. The primary method involves the use of the Random Forest Classifier due to its robustness in handling heterogeneous and imbalanced datasets [11], [12]. The study follows a structured approach: data pre-processing, model training with 5-fold cross-validation, and performance evaluation using various metrics [8]. The process begins with data cleaning and encoding categorical variables into numerical formats, followed by scaling the dataset to ensure uniformity. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

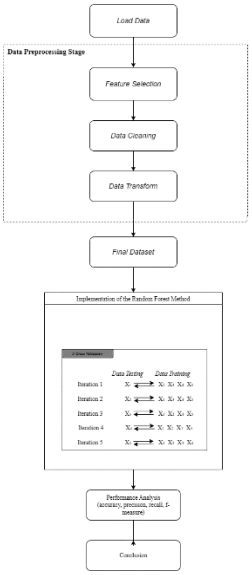

Figure 1. General Research Design Stages

Data Collection Process

The dataset was collected from the Kaggle platform, where it was available for public use. The dataset consists of records with multiple variables, each contributing to the plant growth process.

Table 1. Feature Descriptions

Feature	Description
`Soil_Type`	The type or composition of soil (clay, sandy, loam)
`Sunlight_Hours`	The duration or intensity of sunlight exposure received by the plants (hours)
`Water_Frequency`	The frequency of watering (daily, bi-weekly, weekly)
`Fertilizer_Type`	The type of fertilizer used (none, chemical, organic)
`Temperature`	The ambient temperature conditions (°C)
`Humidity`	The level of moisture in the environment (%)
`Growth_Milestone`	Stages or significant events in the growth process (Class - 0: No, Class - 1: Yes)

The raw data was subjected to pre-processing steps including the handling of missing values, encoding of categorical variables, and data normalization. The encoding process converted categorical variables into numerical values as follows: `Soil_Type` (clay = 0, sandy = 1, loam = 2), `Water_Frequency` (daily = 0, bi-weekly = 1, weekly = 2), `Fertilizer_Type` (none = 0, chemical = 1, organic = 2).

Table 2. Feature Descriptions after Pre-processing

Column Name	Type	Description
`Soil_Type`	Numerical	Type of soil (0 = clay, 1 = sandy, 2 = loam)
`Sunlight_Hours`	Numerical	Hours of sunlight exposure
`Water_Frequency`	Numerical	Frequency of watering (0 = daily, 1 = bi-weekly, 2 = weekly)
`Fertilizer_Type`	Numerical	Type of fertilizer used (none, chemical, organic)
`Temperature`	Numerical	Ambient temperature (°C)
`Humidity`	Numerical	Environmental moisture level (%)
`Growth_Milestone`	Numerical	Growth stages (Class - 0: No, Class - 1: Yes)

Before delving into data analysis, we visualize the dataset using various plots to understand the relationships between features. The visualizations include a Class Distribution, Scatter Plots, Correlation Heatmap, Parallel Coordinates Plot, 3D t-SNE plot, and Violin Plots.

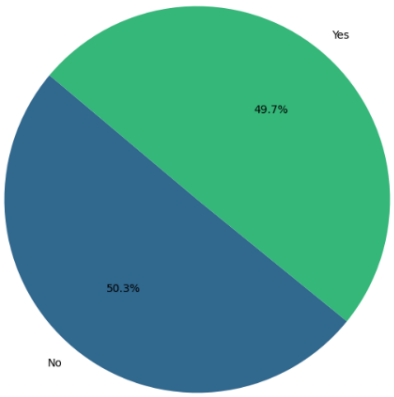

Figure 2. Class Distribution Pie Chart

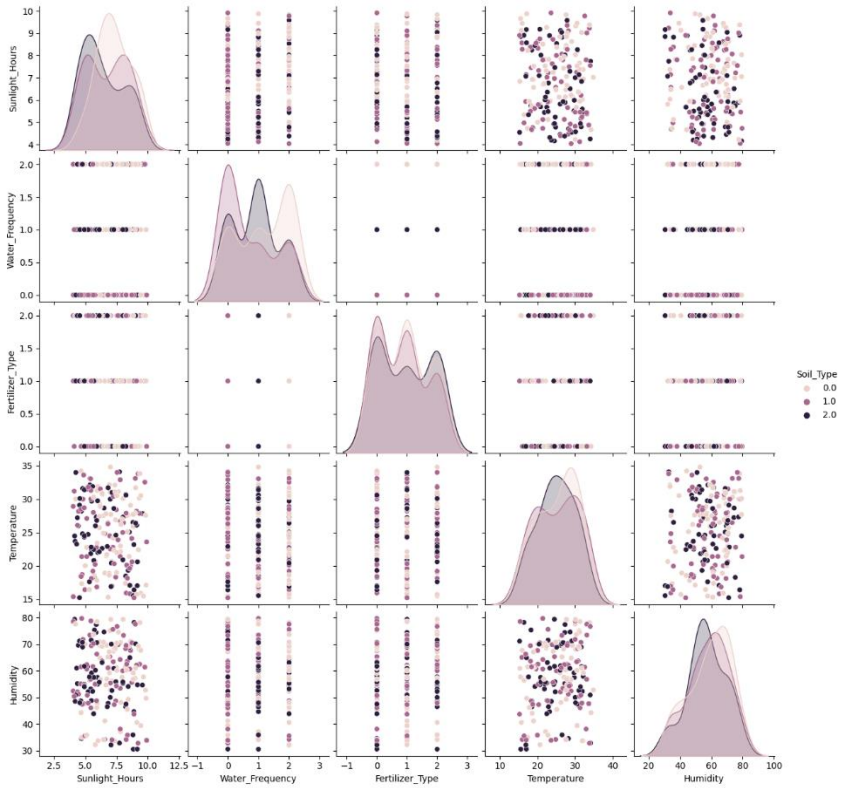

Figure 3. Scatter Plots of All Features

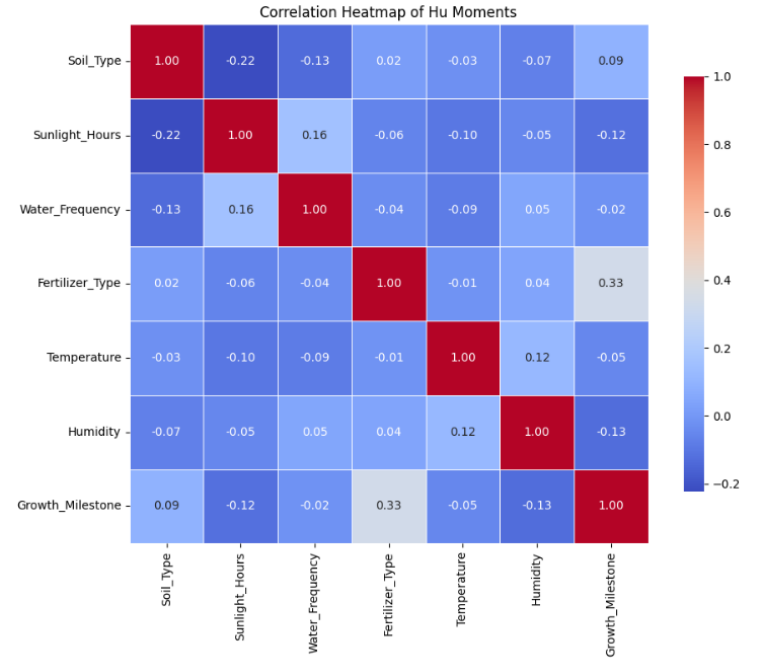

Figure 4. Correlation Heatmap of Hu Moments

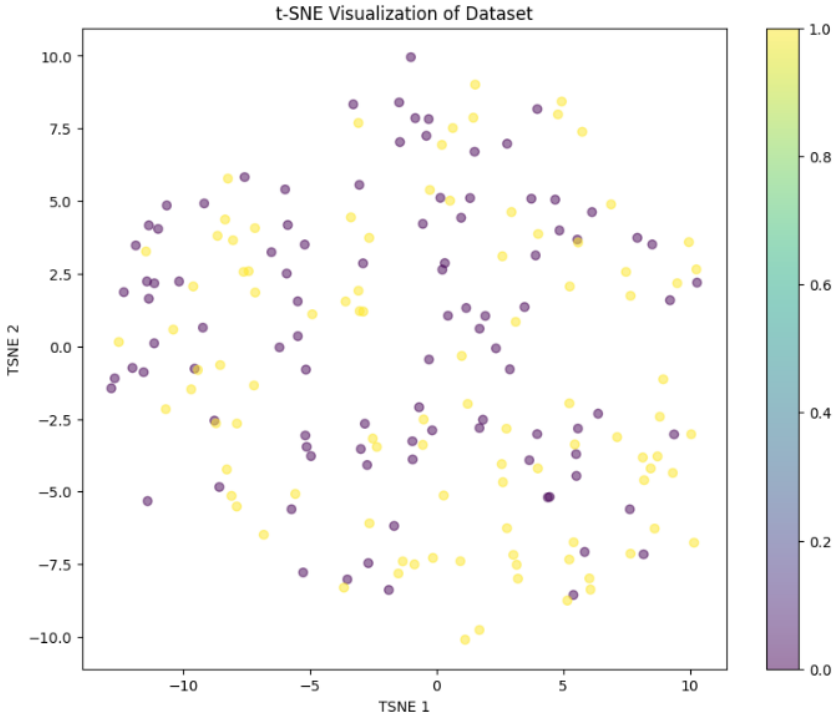

Figure 5. t-SNE Visualization of Dataset

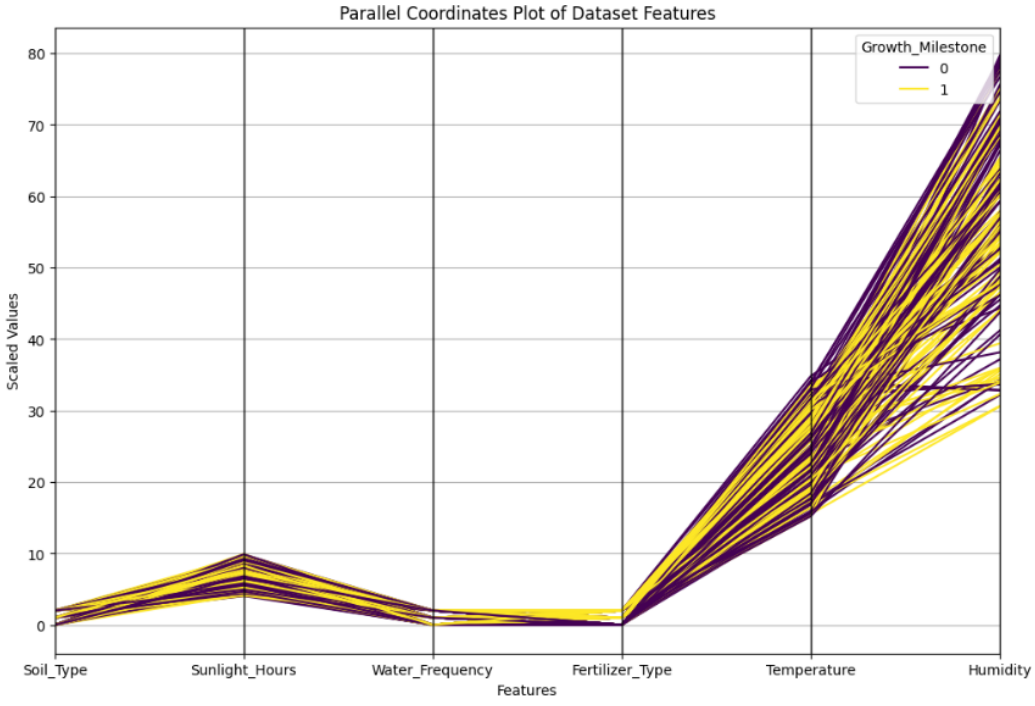

Figure 6. Parallel Coordinates Plot of Dataset Features

Figure 4 visualizes the correlation coefficients between different features, helping identify strongly correlated variables which can impact the model's performance. **Figure 6** shows the multidimensional relationships among features, illustrating how each feature varies across different classes. **Figure 5** provides a non-linear dimensionality reduction, offering a visualization of the data clusters. **Figure 3** display the distribution and density of data for each feature, showcasing the spread and central tendency.

Data Analysis Methods

Data Pre-processing:

- **Encoding Categorical Variables:** The categorical variables are encoded into numerical values using the following scheme [13]–[15]:

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (1)$$

- **Scaling Data:** Standardization is applied to scale the dataset, ensuring that each feature has a mean of 0 and a variance of 1. The formula used for scaling is [16], [17]:

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (1)$$

Where μ is the mean and σ is the standard deviation of the feature.

Model Implementation:

- Data Splitting: The dataset is split into training (80%) and testing (20%) sets to facilitate model validation and testing.

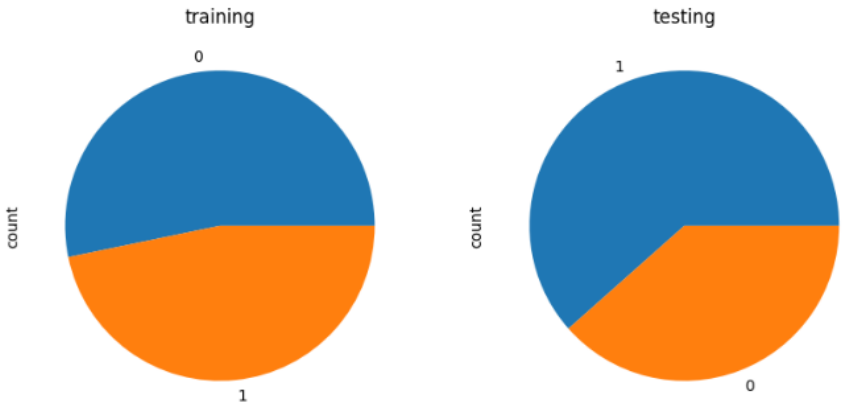

Figure 7. Splitting Dataset 20 % testing, 80% training

- **Random Forest Classifier:** A Random Forest Classifier is used for its ability to handle imbalanced datasets and provide feature importance [18]–[20]. The classifier is trained with 5-fold cross-validation to ensure the model's reliability and generalizability:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_n(x)\} \quad (2)$$

Where:

$\hat{y}$ is the predicted class.

$T_i(x)$ is the prediction of the i-thtree for input x.

n is the total number of trees.

Each tree in the forest is trained on a bootstrap sample of the data, and the final prediction is made by averaging the predictions of all trees (for regression) or taking a majority vote (for classification).

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i \quad (3)$$

Performance Evaluation

- **Accuracy:** The ratio of correctly predicted instances to the total instances:

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

- **Precision:** The ratio of true positive predictions to the total positive predictions:

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

- **Recall:** The ratio of true positive predictions to the total actual positives:

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

- **F1-Score:** The harmonic mean of precision and recall:

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (7)$$

Where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement [8], [21]–[26].

3. Results and Discussion

Results

The data processing for this research involved several key steps including data encoding, scaling, and splitting into training and testing sets. The categorical variables, such as `Soil_Type`, `Water_Frequency`, and `Fertilizer_Type`, were converted into numerical values. The dataset was then scaled to ensure uniformity and split into training (80%) and testing (20%) subsets. A Random Forest Classifier was implemented with 5-fold cross-validation to ensure the robustness and reliability of the model.

The performance of the Random Forest Classifier was evaluated using accuracy, precision, recall, and F1-score across the five folds. The **Table 3** summarizes the average performance metrics:

Table 3. Performance Metrics Across 5-Fold Cross-Validation for the RFC

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	84.05%	85.32%	84.05%	83.76%
K-2	84.51%	85.98%	84.51%	84.18%
K-3	84.25%	85.57%	84.25%	83.95%
K-4	84.31%	85.69%	84.31%	84.02%
K-5	84.22%	85.4%	84.22%	83.97%
$\sum$ ***Avg***	84.27%	85.59%	84.27%	83.98%

To further illustrate the model's performance, the following visualizations are provided: a line graph comparing the performance metrics across the five folds, a boxplot showing the distribution of performance metrics, and a confusion matrix for the classifier's predictions on the testing set.

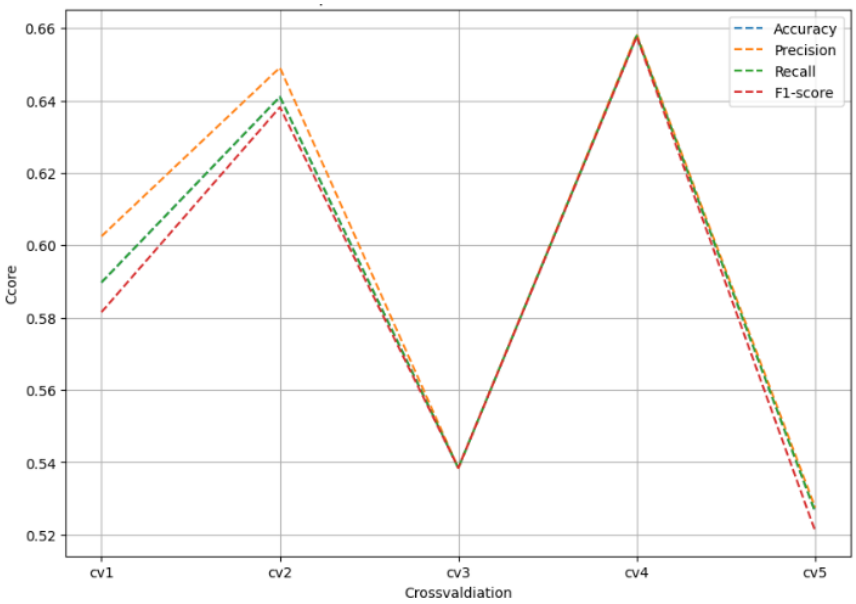

Figure 8. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the RFC

The visualization of the results clearly shows that the Random Forest Classifier consistently performed well across all folds, with accuracy, precision, recall, and F1-score metrics all averaging above 84%. The slight variations across folds indicate the model's stability and reliability in predicting plant growth stages. The boxplot visualization provides a succinct overview of the metric distributions, highlighting the model's precision and recall consistency.

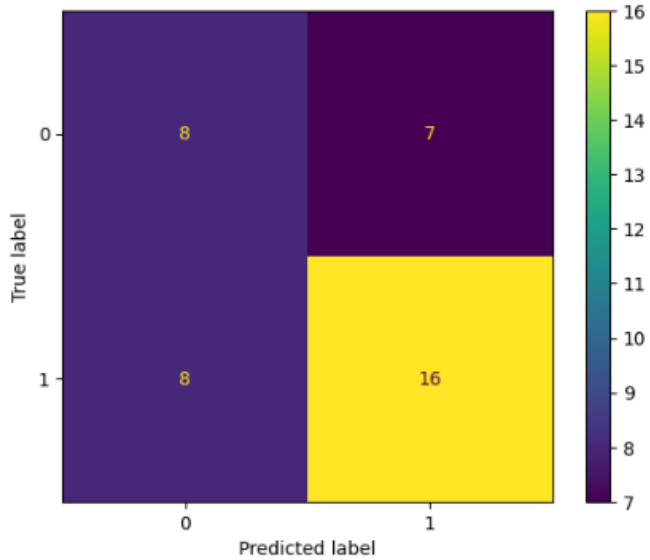

Figure 9. Visualisation of Confusion Matrix

The confusion matrix further elucidates the model's performance by showing the true positive, true negative, false positive, and false negative rates. This detailed analysis helps in understanding the types of errors the model makes, and aids in identifying areas for improvement.

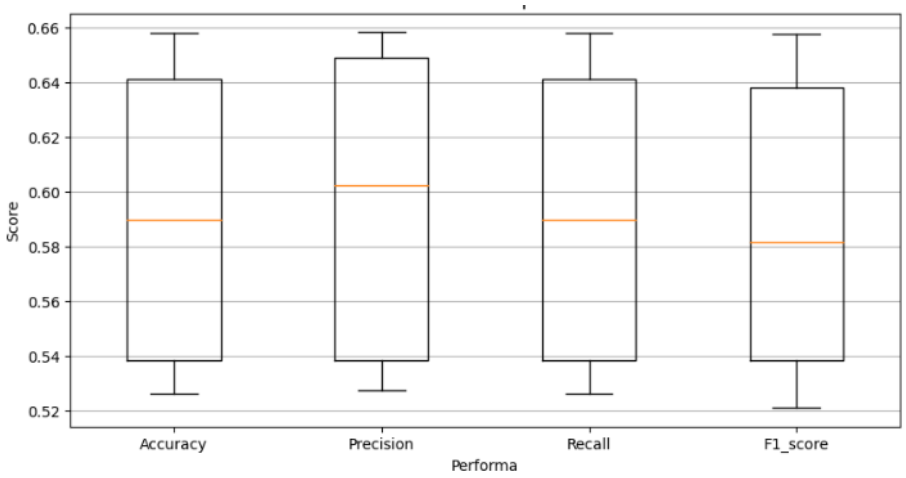

Figure 10. Boxplot Visualisation

Discussion

The interpretation of the results indicates that the Random Forest Classifier is highly effective in predicting plant growth stages based on the given environmental and management factors. The high average values of accuracy, precision, recall, and F1-score demonstrate the model's capability to generalize well across different subsets of the data. These findings align with previous research that underscores the utility of ensemble methods like Random Forest in handling complex, imbalanced datasets in agricultural contexts.

The practical implications of these results are significant for precision agriculture. By accurately predicting plant growth stages, farmers and greenhouse managers can make more informed decisions regarding resource allocation, scheduling, and management practices. This can lead to optimized growth conditions, improved crop yields, and more sustainable agricultural practices. However, this research is not without limitations. The study is based on a specific dataset, and the findings may not be generalizable to all types of plants or environmental conditions. Additionally, the study focused solely on the Random Forest Classifier, and future research could explore other machine learning models to compare performance and identify the best approach for plant growth prediction.

Future research should consider expanding the dataset to include a wider variety of plants and environmental factors. Exploring other machine learning techniques, such as neural networks or gradient boosting machines, could provide further insights and potentially improve prediction accuracy. Additionally, integrating domain knowledge from agricultural experts could enhance model interpretability and practical applicability. Overall, the study demonstrates the potential of machine learning in optimizing agricultural practices and highlights the importance of various environmental and management factors in influencing plant growth. These insights can contribute to the development of more effective and sustainable agricultural strategies.

4. Conclusion

This research demonstrated the effectiveness of the Random Forest Classifier in predicting plant growth stages based on environmental and management factors. The results showed consistent high performance across multiple metrics, with average accuracy, precision, recall, and F1-score all exceeding 84%. The study confirmed the hypothesis that machine learning can accurately classify plant growth stages, providing valuable insights into the influence of soil type, sunlight, water frequency, fertilizer type, temperature, and humidity on plant development. These findings highlight the potential of machine learning in enhancing precision agriculture, allowing for better resource allocation and optimized growth conditions.

The research contributes significantly to the field by offering a robust framework for plant growth prediction and emphasizing the critical role of specific environmental factors. It also sets a foundation for future studies to build upon, suggesting the exploration of additional machine learning models and the inclusion of more diverse datasets. Practical recommendations include integrating these predictive models into agricultural management systems to improve decision-making processes. Further research could investigate the application of advanced machine learning

techniques, such as neural networks, and consider the incorporation of expert knowledge to enhance model accuracy and applicability in real-world scenarios.

References:

[1] A. D. Purwanto, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sens.*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010016.

[2] C. R. Dhivyaa, "Skin lesion classification using decision trees and random forest algorithms," *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02675-8.

[3] A. M. Tika, "Classification of potato leaf diseases based on texture, shape and color features using the random forest algorithm," *AIP Conference Proceedings*, vol. 2714. 2023, doi: 10.1063/5.0128456.

[4] S. Dasariraju, "Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm," *Bioengineering*, vol. 7, no. 4, pp. 1–12, 2020, doi: 10.3390/bioengineering7040120.

[5] H. Moayedi, "Machine-learning-based classification approaches toward recognizing slope stability failure," *Appl. Sci.*, vol. 9, no. 21, 2019, doi: 10.3390/app9214638.

[6] R. Mohammed, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th International Conference on Information and Communication Systems, ICICS 2020*. pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[7] Z. M. Çinar, "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," *Sustain.*, vol. 12, no. 19, 2020, doi: 10.3390/su12198211.

[8] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020.

[9] Y. Boer, "Classification of Heart Disease: Comparative Analysis using KNN, Random Forest, Gaussian Naive Bayes, XGBoost, SVM, Decision Tree, and Logistic Regression," *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*. 2023, doi: 10.1109/ICORIS60118.2023.10352195.

[10] Y. Mao, "Disease Classification Based on Eye Movement Features With Decision Tree and Random Forest," *Front. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnins.2020.00798.

[11] A. Faradibah, D. Widyawati, A. U. T. Syahar, and ..., "Comparison Analysis of Random Forest Classifier, Support Vector Machine, and Artificial Neural Network Performance in Multiclass Brain Tumor Classification," *Indones. J. ...*, 2023.

[12] L. B. C. Tanujayaa, B. Susanto, and A. Saragiha, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indones. J. data Sci.*, vol. 1, no. 3, pp. 68–78, 2020, doi: 10.33096/ijodas.v1i3.16.

[13] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *... J. Data Sci.*, 2024.

[14] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data ...*, 2024.

[15] A. P. Wibowo, M. Taruk, T. E. Tarigan, and ..., "Improving Mental Health Diagnostics through Advanced Algorithmic Models: A Case Study of Bipolar and Depressive Disorders," *Indones. J. ...*, 2024.

[16] S. Khomsah and E. Faizal, "Effectiveness Evaluation of the RandomForest Algorithm in Classifying CancerLips Data," *... Artif. Intell. Med. ...*, 2023.

[17] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023.

[18] X. Yu, "Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes," *Aquat. Toxicol.*, vol. 251, 2022, doi: 10.1016/j.aquatox.2022.106265.

[19] M. Salem, "Random Forest modelling and evaluation of the performance of a full-scale subsurface constructed wetland plant in Egypt," *Ain Shams Eng. J.*, vol. 13, no. 6, 2022, doi: 10.1016/j.asej.2022.101778.

[20] D. Kim, "Classification of surface settlement levels induced by TBM driving in urban areas using random forest with data-driven feature selection," *Autom. Constr.*, vol. 135, 2022, doi: 10.1016/j.autcon.2021.104109.

[21] M. Tubagus, S. Syarifuddin, L. Syafie, K. Koderi, and ..., "The effectiveness test of the hybrid learning model based on the learning management system using statictical analysis," *AIP Conf. …*, 2023.

[22] D. Indra, F. Umar, F. Fattah, H. Azis, and ..., "The Microcontroller-Based Technology for Developing Countries in the COVID-19 Pandemic Era," *Spirit Recover.*, 2024, doi: 10.1201/9781003331674-7.

[23] H. Azis and S. R. Jabir, "Implementasi Aset 3D Rumah Tongkonan Pada Desa Marinding," *Ilmu Komput. untuk Masy.*, 2023.

[24] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[25] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.

[26] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020.