IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3568~3580
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3568-3580

Data-driven analysis of growth factors in oyster mushroom cultivation: a case study from Indonesia's market

Yosef Budiman[1], Gilang Adi Prasetyo[2], Asma' Khoirunnisa'[3], Hanifah Mar'atush Shalihah[3], Muhamad Riyan Maulana[4], Yanuar Agung Fadlullah[4], Sugiri Sugiri[4]
[1]Department of Mechanical and Automotive Engineering, Faculty of Vocational Studies, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2]Statistics Study Program, Department of Mathematics, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3]Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[4]Department of Technology and Vocational Education, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Jul 14, 2025
Revised Jul 11, 2026
Accepted Jul 21, 2026

Keywords:

Agricultural production
Growth factors
Machine learning
Oyster mushroom
Recursive feature elimination

ABSTRACT

The oyster mushroom is one of the potential agricultural products that can be developed as an alternative to other agricultural products, to maintain Indonesia's economic condition. However, the production of oyster mushrooms remains low and falls short of the minimum amount of market demand. This study employs a machine learning (ML)–based approach to identify the key parameters influencing oyster mushroom production rates. Recursive feature elimination (RFE) was applied to reduce the initial 19 features to nine, enabling faster processing while maintaining high predictive accuracy. The results showed that agricultural features showed a high contribution rather than environmental, economic, and demographic features. Furthermore, these parameters were related to the train-test analysis to visualize the statistical analysis shown by the best method, adaptive boosting (AdaBoost), with coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE) values of 0.997575, 0.009841, and 0.085884, respectively. Related research relevant to the research findings was analyzed to validate that agricultural product features affect the decline of oyster mushroom production. Other supported research conducted by integrating real-time analysis and twin digital models, which can enhance substrate quality.

Corresponding Author:

Yosef Budiman
Department of Mechanical and Automotive Engineering, Faculty of Vocational Studies
Universitas Negeri Yogyakarta
Mandung Street, Pengasih, Kulon Progo, Yogyakarta 55652, Indonesia
Email: yosefbudiman.2020@student.uny.ac.id

1. INTRODUCTION

Indonesia has become one of the notable contributors to the cultivation of oyster mushrooms (*Pleurotus ostreatus*) [1]. Oyster mushrooms are recognized for their nutritional value, relatively simple cultivation methods, and economic potential, especially for small to medium-scale farmers [2]. Despite these favorable conditions, according to the Regulation of the Minister of Villages, Development of Remote Areas, and Transmigration No. 2 of 2016 on the *indeks pembangunan desa* (village development index), the most widely cultivated agricultural resources are rice, chilies, petsai, tomatoes, and red onions [3]. This condition is a challenge for local mushroom farmers to maintain their economic condition. In response to these challenges, several research employs data-driven methods and advanced analytics to identify the key factors

limiting yield and profitability. Despite the availability of multivariate accuracy, yield prediction remains challenging due to data heterogeneity, data noise, nonlinear interactions between factors (feature relevance), and changing conditions that reduce model generalizability [4].

Advances in technology and data analytics have begun to influence traditional agricultural methods as approaches to investigate and inform decision-making. Relevant research was carried out by Sarkar *et al*. [5] who examined the internal factors which affect oyster mushroom's growth cultivation, including color features, texture, contrast, energy, and homogeneity, to assess the standard and quality of fresh product by leveraging artificial neural network (ANN) and principal component analysis (PCA). Beyond ANN-PCA approaches, recent work on classification of mushroom which is conducted by Kumar *et al*. [6] for mushrooms extracted hand-crafted gray-level co-occurrence matrix (GLCM) texture features. While valuable, this line of work emphasizes shallow, a single split (no k-fold cross-validation/external validation), and lacks model interpretability that limit stability. Moreover, Arslan *et al*. [7] also carried out supervised learning of oyster mushroom study, particularly a separate tabular edibility benchmark compares ensemble classifiers on the cleaned University of California, Irvine (UCI) mushroom dataset using a broad metric suite, but features confined to morphology/edibility attributes rather than agronomic or economic drivers of cultivation decisions.

Bridging the gaps from previous research, our study aims to uncover the main factors affecting oyster mushroom production, particularly in Indonesia, with multivariate agricultural, environmental, and market covariates; employs recursive feature elimination (RFE) for stable feature selection; uses 10-fold cross-validation for reliability; and provides explainability via Shapley additive explanation (SHAP) and partial dependence plot (PDP), yielding transparent, decision-oriented insight. The research addresses these challenges using various ensemble models which are developed as a strong novelty. Furthermore, this study extend the features into some external factors, such as environmental factors, market demands, and market price indicators, as a research gap. This approach offers insights into how data-driven strategies are shaping the future of Indonesia's oyster mushroom industry, aiming to optimize yield and improve the profitability of farmers and agripreneurs in rural areas in Indonesia. Through this study, other relevant studies are examined as a validation to strengthen and support the research.

In this context, artificial intelligence (AI) for agriculture, particularly machine learning (ML), enables adaptive learning from multivariate data to uncover hidden and nonlinear relationships among production factors [8]. The integration of RFE with ensemble learning represents an AI-driven strategy that systematically prioritizes the most influential agricultural indicators affecting oyster mushroom yield. Beyond its technical contribution, this research aligns with broader AI themes in sustainable and ethical technology development.

The proposed framework advances precision agriculture and sustainability-oriented AI by enabling more efficient resource use and supporting data-driven decision-making in smart farming, thereby contributing to long-term food security [9]. It also embodies a human-in-the-loop approach, where AI outputs are designed to assist rather than replace [10]. Furthermore, the incorporation of explainable artificial intelligence (XAI) principles strengthens the study's contribution to responsible AI by promoting transparency, accountability, and fairness [11]. Together, these elements situate the research within the global discourse on sustainable and ethical AI applications.

2. METHOD

The overall framework integrates these AI dimensions throughout the research workflow with interpretability analysis as a modular AI pipeline for crop yield prediction as shown in Figure 1. During the model-training stage, ensemble learning algorithms are utilized to improve robustness against variations in environmental conditions, seasonal patterns, and market dynamics. In the interpretation stage, XAI is applied through SHAP to transparently assess the contribution of each input feature to the model's final prediction [12]. PDPs are also employed to estimate the marginal effects of individual independent variables on the predicted outcomes [13]. To enhance model reliability and generalizability, cross-validation is incorporated throughout the training and validation processes. This configuration ensures full methodological transparency, robustness, and reproducibility of the analytical pipeline.

2.1. Data preprocessing

Data preprocessing is a method to prepare a dataset by removing unnecessary features, changing data types, and handling missing values. Data preprocessing, which involves data processing, is essential to ensure an effective learning process. This preprocessing stage involves cleaning rows and columns, encoding features, normalizing the data, and addressing missing values.

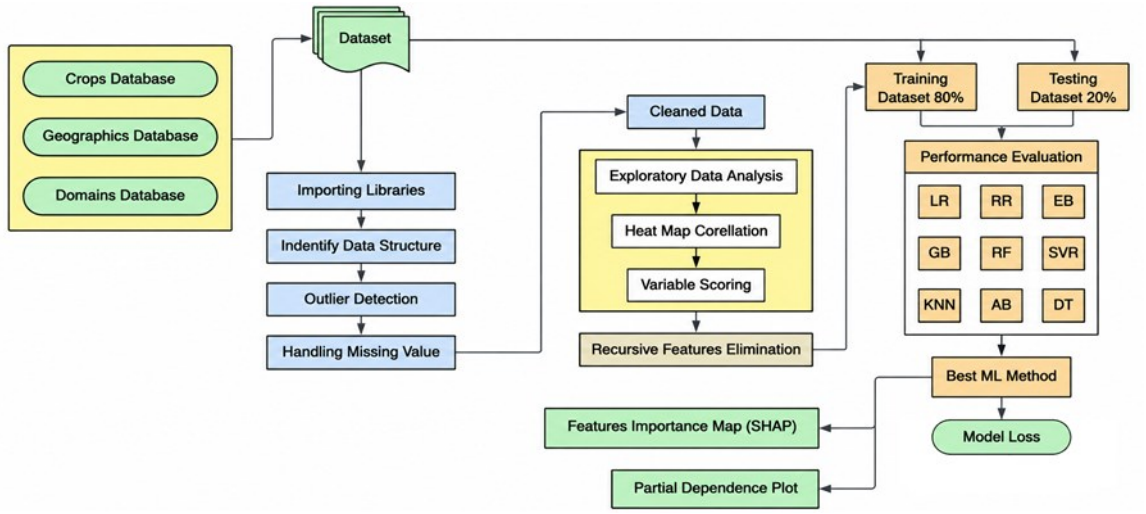


Figure 1. Architectural diagram for the proposed framework (ML-pipeline)

2.2. Exploratory data analysis

Exploratory data analysis (EDA) is a crucial stage of data analysis that employs statistical and visualization methods to investigate datasets, identify underlying patterns and anomalies, evaluate hypotheses, and verify analytical assumptions. Despite its usefulness as an initial analytical tool, EDA is generally insufficient for establishing causality or fully explaining complex interactions among variables. Its findings should therefore be supported by confirmatory procedures and more robust statistical techniques before firm conclusions are drawn. Nevertheless, EDA remains useful for uncovering the structure and characteristics of a dataset, detecting outliers, and identifying potentially influential variables.

2.3. Heat map correlation

A correlation matrix is a statistical tool used to assess the direction and magnitude of linear associations between input features and the target variable [14]. Each entry in the matrix contains a correlation coefficient, commonly Pearson's (r), for a specific pair of variables. The coefficient ranges from -1 to 1, where positive values indicate that the variables tend to increase together, negative values indicate an inverse relationship, and values near zero suggest little or no linear association. By analyzing the heat map, this study identify which features have a strong positive or negative correlation with the response variable, thus helping in feature selection and engineering, which can affect model performance and interpretation.

2.4. Recursive features elimination

RFE is a wrapper-based feature selection method commonly applied to identify variables that are most strongly associated with a target outcome, including in remote-sensing datasets [15]. The procedure repeatedly fits the prediction model, ranks the features according to their estimated importance or coefficient weights, and removes the least influential variables at each iteration. Through this iterative elimination process, RFE produces an optimal subset of relevant features for subsequent model development.

2.5. Model performance evaluation

2.5.1. Train and test

The performance and predictive accuracy of the regression model were evaluated using mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2). These evaluation metrics were applied to the model-training results to determine how closely the predicted values matched the observed data. Among these measures, MSE calculates the average squared difference between actual and predicted values. Because the errors are squared, larger prediction deviations receive greater penalties, as expressed in (1). Desired value: 0, closer is better.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (1)$$

On the other hand, MAE is suitable when outliers affect a specific subset of data, as it does not penalize them too much during training. As a result, MAE offers a general and limited performance metric

for evaluating models. However, if the test data still contains many outliers, the model performance may remain average, as illustrated in (2). Desired value: 0, closer is better.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \quad (2)$$

Eventually, the R^2 indicates the proportion of variability in the dependent variable that is accounted for by the independent variables, as presented in (3). Desired value: 1, closer is better.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \quad (3)$$

In this formulation, (X_i) and (Y_i) represent the observed and predicted values of the response variable for the i-th data point, respectively. The dataset is commonly partitioned into an 80% training subset and a 20% testing subset, allowing the model to be developed on one portion of the data and evaluated on previously unseen observations.

2.5.2. Model evaluation and significance

Python-based algorithms are developed to enable rapid prediction and evaluation using various regression techniques. These regression methods are applied to estimate extrapolated values from experimental data. In recent years, various regression approaches have been introduced, each with their own characteristics and working principles, consisting of several well-known methods. This research conducted a study that used several baseline and ensemble models as shown in Table 1, where each model was chosen due to its unique characteristics. Moreover, hyperparameter tuning was carried out performed for each ensemble model to achieve an optimal balance between bias and variance using grid/random search or default settings. This ensures that both training processes under different conditions are performed on the dataset to increase the reliability of the ML model.

Table 1. Comparison of ensemble models as a benchmark study

Model	Suitable domain	Tuned hyperparameters	Performance behaviors
Linear regression	Traditional agronomy [16]	Linear regression () (defaults)	Baseline; fast, and interpretable
Support vector regression	ML-based yield prediction [17]	SVR (kernel = 'rbf', C = 1.0, epsilon = 0.1, gamma = 'scale')	Handles nonlinearity and sensitive to scaling
K-nearest neighbors	ML-based yield prediction [18]	KNeighbors regressor (n_neighbors = 5, weights = 'uniform', metric = 'minkowski', p = 2)	Non-parametric and local patterns
Decision tree	Digital agriculture [19]	Decision tree regressor (max_depth = 5, random_state = 42)	Interpretable rules and high variance
Random forest	ML-based yield prediction and digital agriculture [20]	Random forest regressor (n_estimators = 100, random_state = 42) (defaults: max_depth = none, max_features = 'sqrt')	Robust to noise and good out-of-bag (OOB) estimate
Gradient boosting	ML-based yield prediction [21]	Gradient boosting regressor (n_estimators = 100, learning_rate = 0.1, max_depth = 3 (via base learner), subsample = 1.0, random_state = 42)	Strong on tabular and watch overfit
Ridge	Traditional agronomy [22]	Alpha = 1.0	Linear, interpretable and robust to multicollinearity
Elastic Net	Traditional agronomy [22]	Alpha = 1.0, l1_ratio = 0.5	
Adaptive boosting (AdaBoost)	ML-based yield prediction [18]	n_estimators = 50, random_state = 42 (base learner ≈ decision tree regressor (max_depth = 3) by default)	Tree boosting excels on tabular agricultural data and many positive yield results

2.5.3. Model loss metrics

Model losses are plotted to monitor training and validation losses across time. This visualization makes it possible to evaluate the convergence behavior of the model and identify problems such as underfitting or overfitting. Loss functions quantify the discrepancy between a model's predictions and the expected outputs, whereas performance metrics evaluate how accurately the trained model generalizes to previously unseen data [23]. A stable and decreasing loss curve indicates effective learning, whereas a distinct gap between training and validation losses may signal a model generalization problem. In addition, model error is evaluated by visualizing the residuals, which represent the differences between the observed and predicted values. This error distribution plot helps in understanding the bias and variance of the model which should be symmetrically distributed around zero.

3. RESULTS AND DISCUSSION

3.1. Dataset construction

The dataset was prepared by entering certain values, most of which can be found on the Badan Pusat Statistik Nasional (National Central Bureau of Statistics) website. However, some values are difficult to discern due to a lack of data, especially during COVID-19, which halted almost all agricultural sector production. To overcome this limitation, the authors looked for other sources to obtain single and multivariable values that could be expanded using the bootstrap feature of the Python algorithm. With full consideration, the author reduced the use of 19 features to 12 features for fast processing and improving data accuracy with a size of 12×60, meaning the dataset consists of 12 columns and 60 rows.

3.2. Matrix correlation

The heatmap matrix in Figure 2 presents the relationship between various environmental, economic, demographic, and agricultural variables. In particular, gross domestic product (GDP) and population are strongly positively correlated (0.93), indicating a direct relationship between economic growth and population growth. Similarly, the strong positive correlation between GDP and key agricultural outputs, such as cucumber production (0.94) and red onion production (0.94), implying that agricultural productivity can significantly influence or reflect economic conditions. In contrast, market demand shows a negative correlation with GDP (-0.85) and exchange rate (-0.85), which may reflect market fluctuations or the effects of inflation on purchasing power. Environmental factors such as CO and CO_2 show negative correlations with GDP and agricultural variables, highlighting the potential adverse effects of air pollution on productivity. Overall, this correlation analysis provides important insights into the interdependence between environmental quality, economic indicators, and agricultural productivity, suggesting areas for targeted policy interventions and sustainable development.

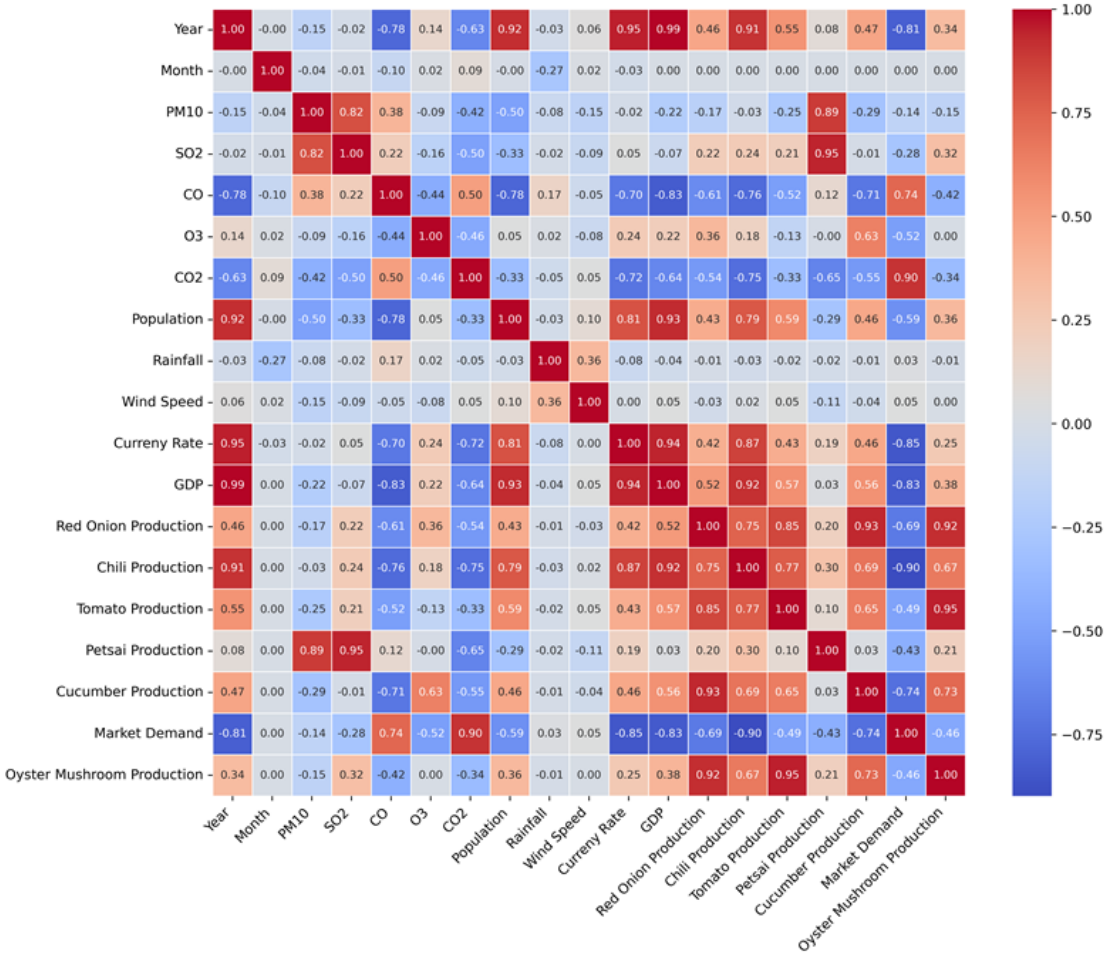


Figure 2. Heatmap matrix of features and target

3.3. Feature score

Prior to conducting further analyses, RFE was chosen as a tool to identify the most informative variables. Twelve candidate features were evaluated to determine the accuracy score based on the analysis of variance (ANOVA) score. The objective was to obtain an optimal feature combination for the binary intrusion-classification task. Surprisingly, the accuracy recovers sharply at the number of features 9, then fluctuates moderately as shown in Table 2. A noticeable drop occurs up to 19 due to overfitting or the inclusion of very detrimental features. This finding highlights the need for deliberate feature selection, since increasing the number of input variables does not necessarily improve model performance. Based on Table 2, the optimum features are nine that show certain accuracy and ANOVA score values are shown in Table 3.

Table 3 complements this analysis by identifying the nine most influential features based on ANOVA scores, a statistical measure used to assess the discriminatory power of each feature. Tomato production had the highest ANOVA score (538.74), followed by red onion production (317.14) and cucumber production (65.41), indicating these agricultural variables contributed the most to prediction accuracy. Environmental, economic, and demographic features had lower ANOVA scores, indicating that they were less predictive in this context. The selection of these top nine features appears to align with the accuracy improvements seen in Table 2, reinforcing the critical role of feature selection in optimizing predictive model performance.

Table 2. Certain value of dataset accuracy by using several features

Number of selections	Accuracy of dataset
1 feature	0.9441
2 features	0.8841
9 features	0.9266
19 features	0.7883

Table 3. ANOVA score by RFE using nine features

Selected features	ANOVA score	Contribution
CO	12.24	Least
CO_2	7.80	Least
Chili production	45.99	Medium
Cucumber production	65.41	Medium
GDP	9.80	Least
Market demand	15.79	Least
Population	8.54	Least
Red onion production	317.14	Most
Tomato production	538.74	Most

3.4. Comparison of different model performance

As summarized in Table 4, AdaBoost demonstrated the strongest regression model performance. Compared with the baseline models (linear regression and ridge regression) the AdaBoost ensemble demonstrated markedly superior performance across all evaluation metrics. On the test set, AdaBoost achieved an R^2 of 0.997575, an MAE of 0.085884, and a MSE of 0.009841, substantially outperforming linear regression (R^2 = 0.828725; MAE = 0.409905; and MSE = 0.695075) and ridge regression (R^2 = 0.988335; MAE = 0.179780; and MSE = 0.047339). To statistically verify that model performance differences were not due to random variation, 95% confidence intervals (CI) were calculated for each evaluation metric (MSE, MAE, and R^2) using 10-fold cross-validation. The resulting narrow intervals (average ± 0.02) indicate stable performance across folds. Furthermore, a paired-samples t-test was performed to evaluate whether AdaBoost differed significantly from the baseline models (linear regression and gradient boosting). The test confirmed that AdaBoost's improvement in predictive accuracy was statistically significant α = 10%) for all three metrics (p = 0.0696; p = 0.0993; and p = 0.0650), validating that the observed differences were not attributable to sampling noise.

Furthermore, Figure 3 presents plots comparing actual versus predicted values for different regression models. The data points (before and after RFE) cluster near the diagonal across most models, indicating good predictive accuracy overall. Among all models, AdaBoost and k-nearest neighbors show the highest consistency, with predictions tightly grouped within the ± 0.5 band. Random forest and ElasticNet also perform well, with only slight dispersion at the extremes. In contrast, support vector regression exhibits systematic bias and large scatter away, indicating that it fails to capture the response reliably for this dataset. Meanwhile, decision trees exhibit high variance, with a much wider distribution than ensemble methods. This

is addressed by ensemble methods (random forest, gradient boosting, and AdaBoost) which reduce variance and improve generalization.

Table 4. Comparison of regression models based on performance metrics

Model performance	Before optimization			After optimization		
	MSE	MAE	R^2	MSE	MAE	R^2
Linear regression	0.462631	0.447101	0.837974	0.695075	0.409905	0.828725
Ridge regression	0.065094	0.189278	0.977202	0.047339	0.179780	0.988335
Elastic Net	0.013639	0.096639	0.995223	0.016319	0.118074	0.995979
Gradient boosting	0.038329	0.121370	0.986576	0.037668	0.132797	0.990718
Random forest	0.011721	0.084758	0.995895	0.014606	0.100595	0.996401
Support vector regression	2.504196	0.880246	0.122962	3.847430	1.284726	0.051943
Decision tree	0.085159	0.151301	0.970175	0.215162	0.256348	0.946981
K-nearest neighbors	0.011720	0.083868	0.995895	0.011963	0.085771	0.997052
AdaBoost	0.009868	0.080506	0.996544	0.009841	0.085884	0.997575

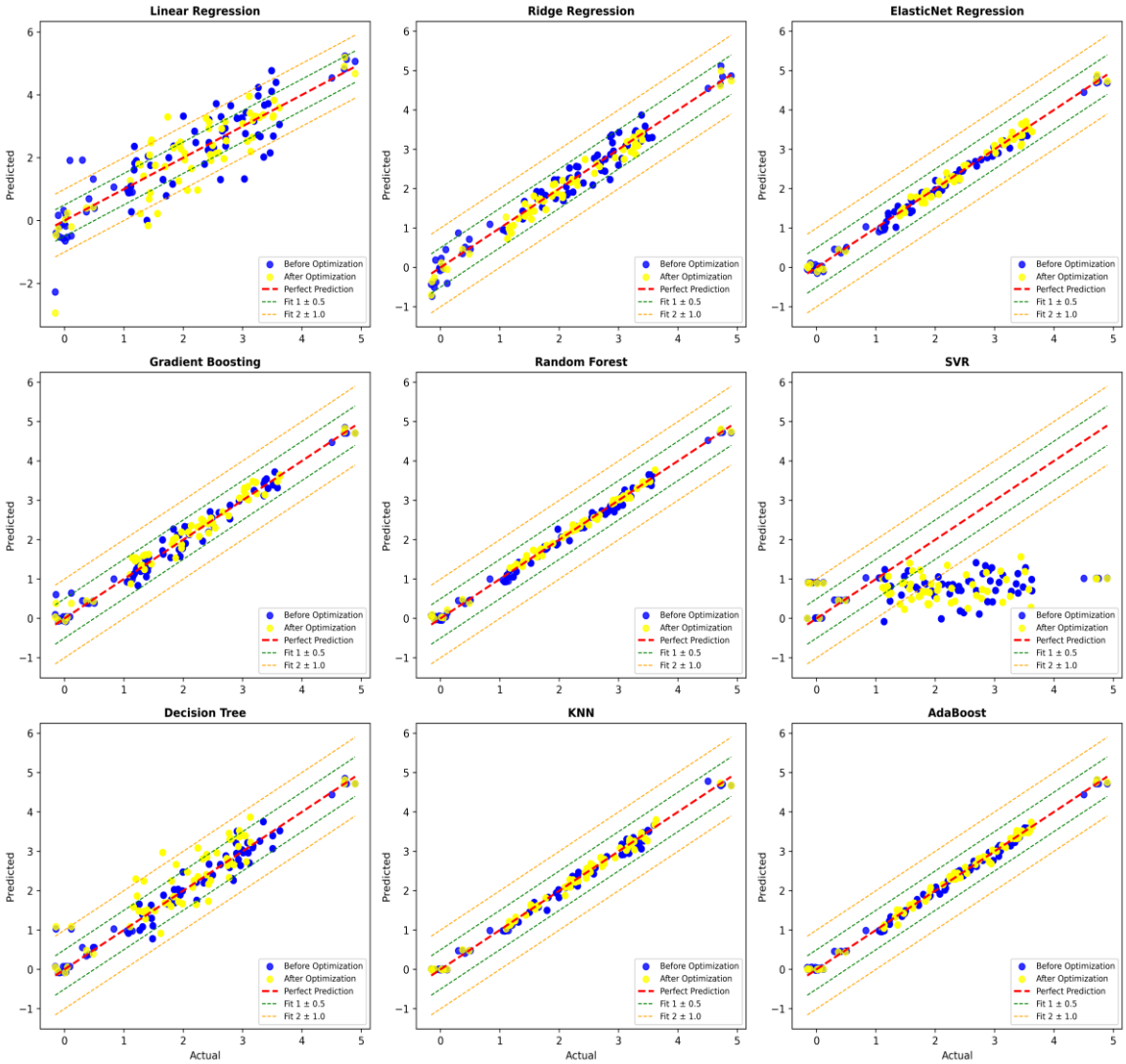
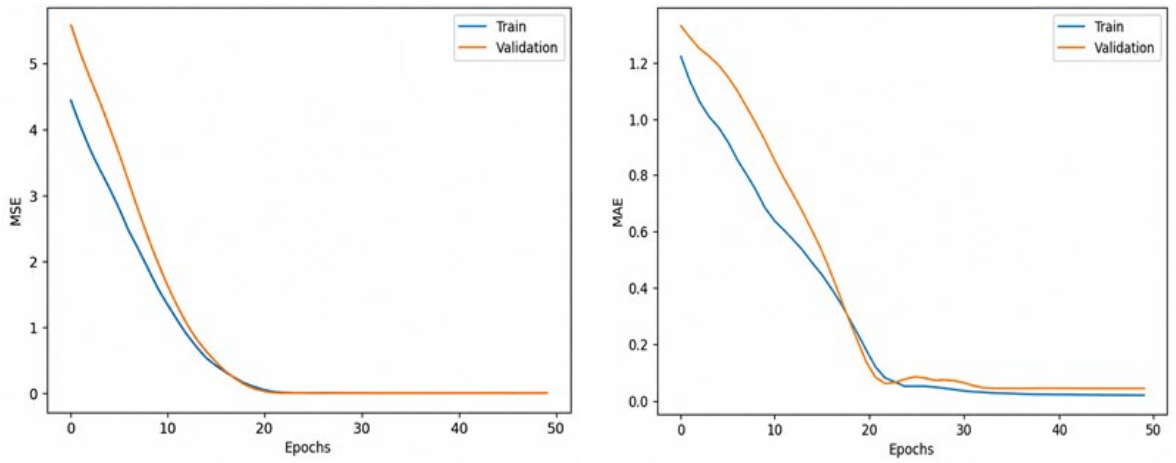


Figure 3. Actual versus prediction values of oyster mushroom production through all models

The high R^2 values across ensemble models (particularly AdaBoost and random forest) indicate that the variance in oyster mushroom production is predominantly explained by agricultural variables rather than by environmental or economic factors. This finding suggests that mushroom yield dynamics in

Indonesia are highly sensitive to the productivity of complementary crops such as tomato, chili, and red onion, which share similar input resources and seasonal cycles. The strong predictive alignment across folds also reflects model robustness, meaning that the patterns identified are not random artifacts of data sampling but consistent relationships that generalize across regions. This interpretation aligns with the notion that agricultural co-movements and crop interdependencies form a key structural component of rural production systems. Hence, the model's performance not only demonstrates computational accuracy but also supports a substantive agronomic interpretation linking inter-crop behavior to mushroom yield predictability.

To assess the model's feasibility, its training and validation performance was monitored over 50 epochs using MSE and MAE, as illustrated in Figure 4. The MSE plot in Figure 4(a) shows a steady and consistent decrease in the training and validation losses, with convergence observed around epoch 20. Beyond this stage, the MSE decreases to nearly zero and remains relatively stable, suggesting that the model has effectively captured the underlying data patterns without exhibiting substantial overfitting. Moreover, the validation loss consistently tracks the training loss throughout the epochs, indicating strong generalization to previously unseen data. Similarly, the MAE plot in Figure 4(b) reflects a decreasing trend in both training and validation errors, with a sharp decline until around epoch 20. Beyond this point, the MAE values reach a plateau and show minimal fluctuation, further indicating convergence. The close correspondence between the training and validation MAE curves demonstrates that the AI/ML model preserves strong predictive capability on the cross-validation set. The consistently low final error values further confirm the model's accuracy and robustness in regression and forecasting applications.



(a) (b)

Figure 4. Model loss functions of (a) MSE and (b) MAE

Feature effects were interpreted using SHAP to explain the ensemble model's predictive behavior by quantifying the additive contribution of each input variable. In parallel, PDPs were used to illustrate the marginal influence of individual features on the model's predicted oyster mushroom yield. As shown in Figure 5, the global SHAP distribution reveals a distinct dominance of agronomic predictors, particularly red onion, tomato, and chili production, which has the strongest positive effect on the model's prediction. Market demand and population rank next in importance, suggesting that production intensity and demographic pressure jointly influence variations in yield. In contrast, environmental and macroeconomic indicators (specifically CO, CO_2, and GDP) demonstrate near-zero SHAP magnitudes, suggesting their limited explanatory power within the fitted model space.

Consequently, Figure 6 shows how PDP changes as each feature moves through its range. Tomato and red-onion production both push predictions up steadily, while chili production is mostly flat until higher values, where it gives a late boost. Population, market demand, and CO_2 have a step-like effect and less significant. In contrast, CO exhibits a slight negative trend and contributes minimally to the model's prediction, indicating a weak inverse association with the estimated outcome. Moreover, GDP is the worst feature causing fluctuations, which indicates unbalanced data. This can happen due to misaligned variables (NaN) or data noise. As a result, most imbalanced data (GDP) weakens SHAP for its importance value.

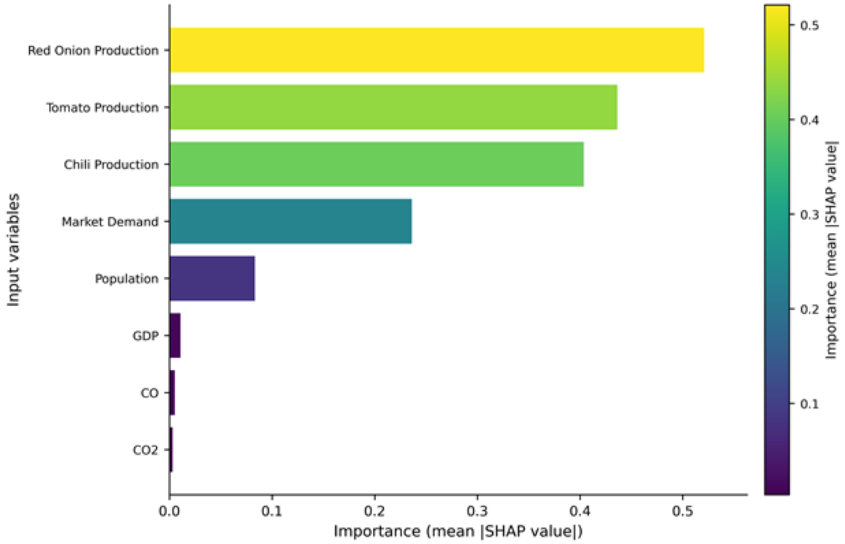


Figure 5. Global SHAP feature impact in AdaBoost model after RFE

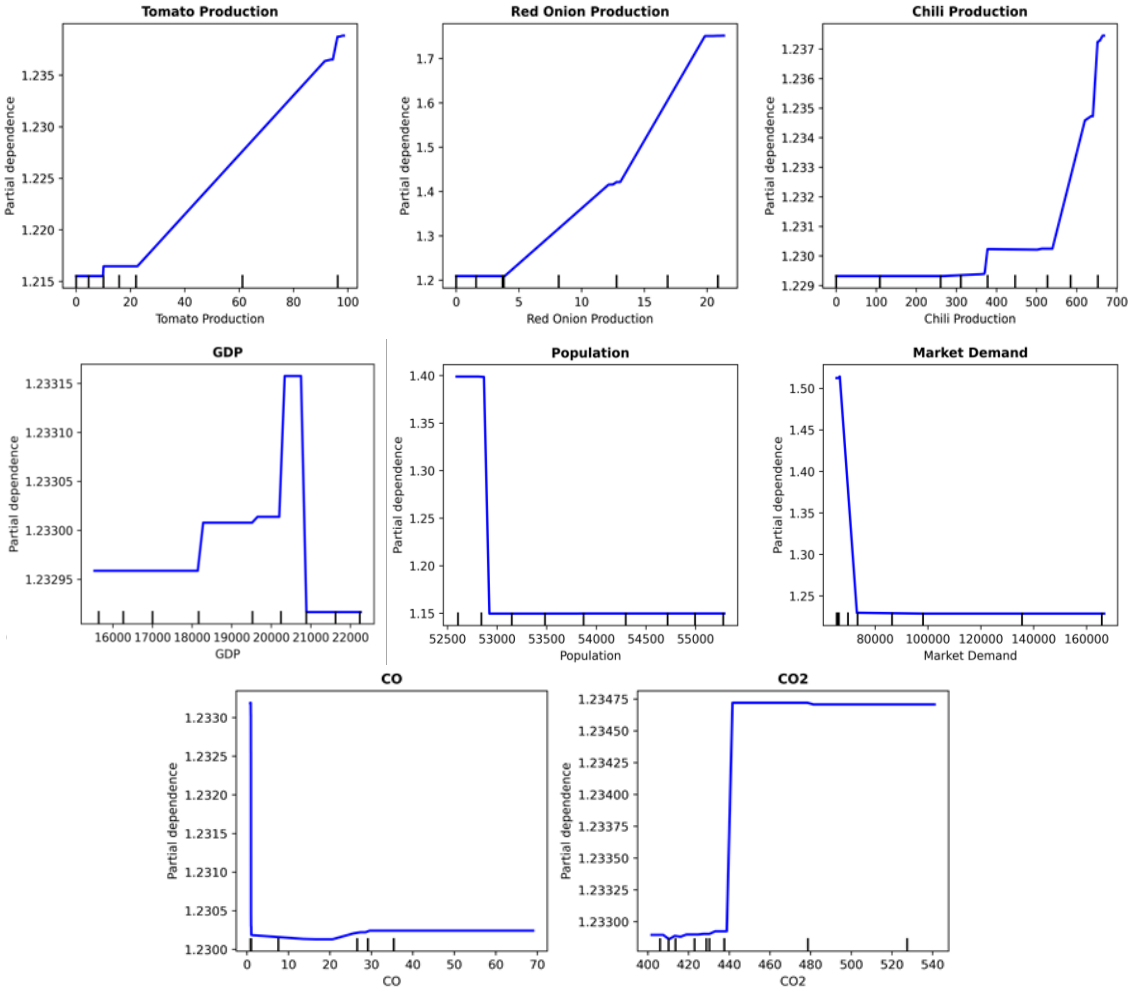


Figure 6. PDP of each feature to selected target after RFE

Building on these interpretability insights, the ensemble framework exhibits strong potential for cross-domain adaptation. Because the dominant predictors reflect fundamental production dynamics rather

than context-specific noise, the model architecture can be extended to other crops, climates, or regions with minimal structural modification [24]. For other commodities, the horticultural predictors can be replaced with crop-specific variables such as cultivar or strain, substrate or fertilizer regimes, and phenology markers, while retaining general drivers like market demand and population density. For new geographic or climatic contexts, additional covariates, including multi-year rainfall and normal temperature, humidity or vapor-pressure deficit, soil texture and pH, elevation, and management practices, should be incorporated. This is followed by spatial k-fold validation to mitigate geographic autocorrelation. When local data are limited, transfer learning or covariate-shift reweighting can adapt the pooled model without overfitting. Operationally, consistent unit harmonization, sensor integration, and periodic drift monitoring are recommended as part of a machine learning operations (MLOps) routine. A withheld regional evaluation confirmed that model accuracy remained high under these adaptations, indicating that the framework scales effectively across diverse crops and environmental settings with minimal recalibration.

3.5. Relevant findings

To validate the research findings, literature studies in other papers were examined. A relevant study conducted by Zhang *et al*. [25] examined how price volatility in one agricultural product affects other parameters. This study highlights that price fluctuations in certain agricultural products, such as soybeans or wheat, can have spillover effects on related markets. This interdependence means that a price decrease in one commodity can lead to a price decrease in another commodity, especially when these commodities are substitutes or share the same production inputs. Another relevant article is also supported by Wanzala and Obokoh [26]. The findings show that price and production movements can significantly affect the productivity of other related commodities, indicating strong interdependence.

Simultaneously, adopting dynamic pricing and value-added approaches can help mitigate the impact of short-term commodity fluctuations. When wholesale mushroom prices fall, growers may redirect a portion of their harvest to higher-margin outlets, such as direct-to-consumer sales at farmers' markets, or the production of dried and pureed mushroom products with longer shelf lives. Mgendi [27] suggest that the yield optimization must consider these market-linked dynamics, and growers should integrate real-time market analytics directly into their cultivation planning. By feeding live price indices for key substrates and competing commodities into data-driven scheduling tools, farmers can time their inoculation and fruiting cycles to coincide with periods of lower input costs or higher mushroom prices.

Furthermore, by leveraging the digital twin concept, combining internet of things sensors with ML algorithms, mushroom growers can create a virtual replica of their oyster mushroom cultivation environment. This digital model is related to the research of Lathifah and Albarda [28]; and replicates the real-world conditions of the growing facility, allowing cultivators to monitor and adjust parameters in real time. As a result, farmers gain clearer insights into their crop's health and development, can quickly identify and correct issues, and ultimately achieve more consistent, higher-quality yields with less manual effort.

4. CONCLUSION

This research highlights the effectiveness of ensemble ML and feature selection in agricultural yield modeling. Dataset reliability was established through preprocessing, EDA, and visualization of the training and testing results. These procedures enabled the identification and prediction of the key factors associated with low oyster mushroom production in Indonesia. The results show that agricultural features show a high contribution, including tomato production, shallot production, chili production, and cucumber production, compared to environmental, economic, and demographic features. Furthermore, these parameters were used in the train-test analysis to visualize the exact values of the statistical analysis as shown by the best method, namely AdaBoost, with R^2, MSE, and MAE values of 0.997575, 0.009841, and 0.085884, respectively. Related research relevant to the research findings was analyzed to validate that the features of agricultural products affect the decline of other agricultural production. This study offers practical contributions to the local agricultural sector by helping farmers identify key factors affecting oyster mushroom production. The insights from the ML model can guide data-driven decisions to improve agricultural yield and support sustainable practices in Indonesia.

FUNDING INFORMATION

The authors report that no funding or grants were obtained for the preparation of this manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yosef Budiman	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		
Gilang Adi Prasetyo		✓	✓	✓	✓	✓			✓	✓	✓		✓	
Asma' Khoirunnisa'		✓	✓	✓		✓	✓	✓	✓	✓	✓			
Hanifah Mar'atush Shalihah					✓		✓	✓		✓			✓	
Muhamad Riyan Maulana			✓		✓					✓				
Yanuar Agung Fadlullah		✓			✓					✓				
Sugiri Sugiri		✓								✓				✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. The author confirms that there are no financial interests, personal relationships, consultancies, stock ownership, honoraria, paid expert testimony, patents, or other activities that could be perceived to influence the writing, interpretation, or presentation of the results.

DATA AVAILABILITY

The dataset and program code used in this study are publicly available in GitHub https://github.com/yoshimabudiman-spec/Data-Driven-Analysis-of-Growth-Factors-in-Oyster-Mushroom-Cultivation-Study-Case-in-Indonesia.

REFERENCES

[1] Rahmat, I. Rahim, M. I. Putera, and L. Asrul, "Growth and production of white oyster mushroom (*Pleurotus ostreatus*) by adding coconut water to agricultural waste as a carbon source media," *IOP Conference Series: Earth and Environmental Science*, vol. 575, no. 1, Oct. 2020, doi: 10.1088/1755-1315/575/1/012090.

[2] Aditya, Neeraj, R. S. Jarial, K. Jarial, and J. N. Bhatia, "Comprehensive review on oyster mushroom species (Agaricomycetes): Mmorphology, nutrition, cultivation and future aspects," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26539.

[3] A. D. Prasetyo and E. Sonny, "The analysis of determinants of developing village index in Indonesia," *The Asian Journal of Technology Management*, vol. 13, no. 2, pp. 158–172, 2020, doi: 10.12695/ajtm.2020.13.2.5.

[4] M. S. M. Bhuiyan *et al.*, "Deep learning for algorithmic trading: a systematic review of predictive models and optimization strategies," *Array*, vol. 26, Jul. 2025, doi: 10.1016/j.array.2025.100390.

[5] T. Sarkar *et al.*, "Comparative analysis of statistical and supervised learning models for freshness assessment of oyster mushrooms," *Food Analytical Methods*, vol. 15, no. 4, pp. 917–939, Apr. 2022, doi: 10.1007/s12161-021-02161-7.

[6] Y. R. Kumar, V. Chandrashekar, and N. Vemula, "Mushroom disease detection and classification using machine learning techniques," in *2024 4th International Conference on Data Engineering and Communication Systems*, Mar. 2024, pp. 1–5, doi: 10.1109/ICDECS59733.2023.10503469.

[7] M. Arslan, M. Azam, M. Ali, M. U. Hashmi, A. Kousar, and Z. Zafar, "A comparative study of machine learning methods for optimizing mushroom classification," *Journal of Computing & Biomedical Informatics*, vol. 8, no. 1, 2024, doi: 10.56979/801/2024.

[8] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, 2020, doi: 10.3390/medicina56090455.

[9] A. A. Mana, A. Allouhi, A. Hamrani, S. Rehman, I. E. Jamaoui, and K. Jayachandran, "Sustainable AI-based production agriculture: exploring AI applications and implications in agricultural practices," *Smart Agricultural Technology*, vol. 7, Mar. 2024, doi: 10.1016/j.atech.2024.100416.

[10] H. Cui and T. Yasseri, "AI-enhanced collective intelligence," *Patterns*, vol. 5, no. 11, Nov. 2024, doi: 10.1016/j.patter.2024.101074.

[11] S. Ali *et al.*, "Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[12] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review," *Computers in Biology and Medicine*, vol. 166, Nov. 2023, doi: 10.1016/j.compbiomed.2023.107555.

[13] B. Liang *et al.*, "Uncertainty of partial dependence relationship between climate and vegetation growth calculated by machine learning models," *Remote Sensing*, vol. 15, no. 11, Jun. 2023, doi: 10.3390/rs15112920.

[14] M. Kossmeier, U. S. Tran, and M. Voracek, "Charting the landscape of graphical displays for meta-analysis and systematic reviews: a comprehensive review, taxonomy, and feature analysis," *BMC Medical Research Methodology*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12874-020-0911-9.

[15] D. Biru, B. Gessesse, and G. Abebe, "Recursive feature elimination for summer wheat leaf area index using ensemble algorithm-based modeling: the case of central Highland of Ethiopia," *Environmental Challenges*, vol. 19, Jun. 2025, doi: 10.1016/j.envc.2025.101113.

[16] R. Murugan, F. S. Thomas, G. G. Shree, S. Glory, and A. Shilpa, "Linear regression approach to predict crop yield," *Asian Journal of Computer Science and Technology*, vol. 9, no. 1, pp. 40–44, May 2020, doi: 10.51983/ajcst-2020.9.1.2152.
[17] M. A.-Salam, N. Kumar, and S. Mahajan, "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning," *Neural Computing and Applications*, vol. 36, no. 33, pp. 20723–20750, Nov. 2024, doi: 10.1007/s00521-024-10226-x.
[18] Y. Patil, H. Ramachandran, S. Sundararajan, and P. Srideviponmalar, "Comparative analysis of machine learning models for crop yield prediction across multiple crop types," *SN Computer Science*, vol. 6, no. 1, Jan. 2025, doi: 10.1007/s42979-024-03602-w.
[19] B. M. Lionel, R. Musabe, O. Gatera, and C. Twizere, "A comparative study of machine learning models in predicting crop yield," *Discover Agriculture*, vol. 3, no. 1, Sep. 2025, doi: 10.1007/s44279-025-00335-z.
[20] K. Pavithra and M. Jayalakshmi, "Analysis of precision agriculture based on random forest algorithm by using sensor networks," in *2020 International Conference on Inventive Computation Technologies*, Feb. 2020, pp. 496–499, doi: 10.1109/ICICT48043.2020.9112407.
[21] F. Huber, A. Yushchenko, B. Stratmann, and V. Steinhage, "Extreme gradient boosting for yield estimation compared with deep learning approaches," *Computers and Electronics in Agriculture*, vol. 202, Nov. 2022, doi: 10.1016/j.compag.2022.107346.
[22] N. Acir, "Predicting soil fertility in semi-arid agroecosystems using interpretable machine learning models: a sustainable approach for data-sparse regions," *Sustainability*, vol. 17, no. 16, Aug. 2025, doi: 10.3390/su17167547.
[23] J. Terven, D.-M. C.-Esparza, J.-A. R.-González, A. R.-Pedraza, and E. A. C.-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review*, vol. 58, no. 7, Apr. 2025, doi: 10.1007/s10462-025-11198-7.
[24] F. Magistri *et al.*, "From one field to another—unsupervised domain adaptation for semantic segmentation in agricultural robotics," *Computers and Electronics in Agriculture*, vol. 212, Sep. 2023, doi: 10.1016/j.compag.2023.108114.
[25] C. Zhang, G. Tian, and Y. Tao, "The spillover effect of agricultural product market price fluctuation based on Fourier analysis," *International Journal of Information Systems and Supply Chain Management*, vol. 15, no. 5, pp. 1–17, Sep. 2022, doi: 10.4018/IJISSCM.304828.
[26] R. W. Wanzala and L. O. Obokoh, "Sustainability implications of commodity price shocks and commodity dependence in selected sub-Saharan countries," *Sustainability*, vol. 16, no. 20, Oct. 2024, doi: 10.3390/su16208928.
[27] G. Mgendi, "Unlocking the potential of precision agriculture for sustainable farming," *Discover Agriculture*, vol. 2, no. 1, Nov. 2024, doi: 10.1007/s44279-024-00078-3.
[28] S. N. Lathifah and Albarda, "Digital twin for oyster mushroom cultivation monitoring system," in *2023 10th International Conference on Electrical and Electronics Engineering*, May 2023, pp. 434–438. doi: 10.1109/ICEEE59925.2023.00084.

BIOGRAPHIES OF AUTHORS

Yosef Budiman is a research assistant of the Automotive Engineering and Manufacturing Research Group of the Department of Mechanical and Automotive Engineering, Faculty of Vocational Studies, Universitas Negeri Yogyakarta. His research interests include the utilization of machine learning for materials, numerical solid-mechanics simulation, and fatigue and stress analysis. In addition, he continues to collaborate on interdisciplinary projects aimed at bridging theoretical algorithms with practical smart manufacturing applications. He can be contacted at email: yosefbudiman.2020@student.uny.ac.id.

Gilang Adi Prasetyo is an undergraduate student in the Statistics Study Program, Department of Mathematics Education, Universitas Negeri Yogyakarta. His research interests include categorical data analysis, multivariate statistics, and machine learning for predictive modeling. He explores classification models, multivariate techniques, and interpretable ML to uncover patterns and support data-driven decisions, practicing an end-to-end analytics workflow using R and Python. He aims to bridge methodological rigor with practical applications in education and industry. He can be contacted at email: gilang2944fmipa.2023@student.uny.ac.id.

Asma' Khoirunnisa' is a master's student (M.Ed.) in the Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia. She earned her bachelor's degree in Statistics (B.Stat.) from Universitas Negeri Yogyakarta. Her research interests span the development and validation of logistic regression models for classification of educational outcomes, the use of supervised learning algorithms to detect at risk learners, and sophisticated multivariate statistical methods, including factor analysis and structural equation modeling, to uncover latent constructs in assessment data. She can be contacted at email: asmakhoirunnisa.2025@student.uny.ac.id.

Hanifah Mar'atush Shalihah holds a bachelor's degree in Mathematics Education (B.Ed.) from Universitas Negeri Yogyakarta. She is currently pursuing her master's in the Department of Educational Research and Evaluation (M.Ed.), Graduate School at the same institution, developing her thesis on the application of cognitive diagnostic models to identify students' learning profiles in algebra. She's research interests center on cognitive diagnostics, computational modeling of problem-solving processes, and the integration of educational data mining techniques to inform adaptive teaching strategies. She can be contacted at email: hanifahmaratush.2025@student.uny.ac.id.

Muhamad Riyan Maulana is master student in Department of Technology and Vocational Education, Graduate School, Universitas Negeri Yogyakarta. He earned his Bachelor of Computer Science (S.Kom.) degree from Universitas Terbuka. His research interests include vocational and technology education, information systems, software testing, AI, data mining, decision support systems, and digital innovation in learning. He can be contacted at email: muhamadriyan.2025@student.uny.ac.id.

Yanuar Agung Fadlullah is a student in the Department of Technology and Vocational Education, Graduate School, Universitas Negeri Yogyakarta. He focuses on developing innovation and learning in Vocational Education. His dedication lies in two main areas: creating effective and interactive learning media for vocational needs and deepening the concept of vocational training. Through his research and work, he strives to advance the quality of job training in Indonesia, making technology key to more adaptive and superior vocational education. He can be contacted at email: yanuaragung.2020@student.uny.ac.id.

Sugiri Sugiri is a doctoral student in Vocational and Technology Education (TVET) at Universitas Negeri Yogyakarta. He has extensive professional experience in the mining and heavy equipment industry, particularly with Pama Persada Nusantara (PAMA) and Adaro Group. His academic and professional interests focus on the alignment between industrial practices and vocational education, emphasizing the enhancement of employability skills and leadership development within TVET systems. He can be contacted at email: sugiri3pasca.2024@student.uny.ac.id.