International Journal of Advances in Data and Information Systems
Vol. 6, No. 2, August 2025, pp. 344~356
ISSN: 2721-3056, DOI: 10.59395/ijadis.v6i2.1393

Prediction of Rice Harvesting During the Rainy Season in Kabupaten Lamongan Using Stochastic Frontier Analysis

Imelda Widya Ningrum[1], Dwi Arman Prasetya[2], Trimono [3], Anuar bin Mohamed Kassim[4]
[1,3]Department of Data Science, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia
[2] Department Master of Information Technology, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia
[4]Faculty of Electrical Engineering Technology, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

Article Info

Article history:

Received May 08, 2025
Revised May 28, 2025
Accepted Jul 12, 2025

Keywords:

Stochastic Frontier Analysis
Technical Efficiency
Rice Production
Data Science
Prediction

ABSTRACT

The agricultural sector plays a critical role in ensuring national food security, yet it faces challenges in achieving technical efficiency due to limited land and input resources. This study aims to model and predict the technical efficiency of rice production in Lamongan Regency during the rainy season using a data science-driven Stochastic Frontier Analysis (SFA) approach. The dataset includes key inputs such as land area, labor, fertilizer, and environmental variables. The methodology involved data preprocessing, feature selection based on Pearson correlation and VIF thresholds, and model validation using metrics like R-squared, MAPE, and log-likelihood. The SFA model demonstrated high predictive capability, with R^2 values exceeding 0.91 in cross-validation and MAPE under 15%. The low gamma value (γ = 0.0100) indicates minimal yet consistent inefficiency. The results suggest that integrating SFA with data science techniques provides an effective framework for identifying inefficiencies and can serve as a decision-support system for evidence-based agricultural policy.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Dwi Arman Prasetya
Department Master of Information Technology,
University of Pembangunan Nasional Veteran Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294
Email: arman.prasetya.sada@upnjatim.ac.id

1. INTRODUCTION

The agricultural sector plays a highly strategic role in Indonesia's economy, particularly in maintaining food security, absorbing labor, and stabilizing the national economy. One of the main challenges faced by this sector is how to improve technical production efficiency—defined as the ability to generate maximum output with limited inputs such as land, labor, and capital. In the context of a growing population and shrinking agricultural land, there is an increasing need for accurate data-driven analytical methods to map efficiency levels. Data analysis plays a crucial role in identifying problems and formulating effective solutions across sectors [1]. This becomes even more critical considering the real impact of food crises on socio-economic stability. A study in East Java indicated that food crises have had a significant impact, with recovery occurring gradually in line with the stabilization of food supplies [2].

The advancement of Data Science offers significant opportunities to address these challenges. Through analytical and algorithmic approaches, Data Science can solve complex problems systematically and based on empirical data. The ability to process large-scale data and generate meaningful insights is key to strategic decision-making, including optimizing agricultural

production. Techniques such as clustering analysis have been proven effective in identifying potential regions or markets efficiently, as applied in mapping agricultural commodity export markets [3], [4]. In practice, accurate data analysis helps farmers and policymakers allocate resources more effectively to increase yields.

One of the most effective analytical methods for measuring and evaluating production efficiency is Stochastic Frontier Analysis (SFA). This method measures how far actual production performance deviates from the maximum efficiency frontier under given input conditions. SFA not only accounts for technical inefficiency but also distinguishes it from random error in the data [5]. The application of SFA in agriculture has been shown to improve production efficiency, both through manual approaches and technology-based methods such as integration with large-scale data management systems like MySQL, which are effective in handling massive agricultural datasets [6].

The application of machine learning and predictive methods in agriculture is also increasingly used to forecast harvests, classify land quality, and analyze weather and climate patterns. To support data processing efficiency and model training, parallel computing with multicore processor load distribution can be utilized to accelerate analysis results [7]. The combination of SFA and Data Science presents opportunities to build adaptive and predictive analytical systems that respond to production challenges in real time. Moreover, modern data analytics technology is now used to monitor soil conditions, climate, and inputs in real-time, opening significant potential for smart farming [8].

However, there is still a gap in the application of SFA methods specifically for predicting rice production efficiency in regions like Lamongan Regency. Most existing studies are still national or cross-commodity in scope and do not consider local seasonal variables that significantly affect production efficiency. The use of SFA models has shown meaningful results in various regions, including China and Southeast Asia, where factors such as labor use, fertilizers, and agricultural machinery are key determinants of efficiency [9]. Therefore, this study proposes an SFA-based approach to model and predict the technical efficiency of rice production in Lamongan Regency, specifically during the rainy season.

This study aims to build a predictive model of technical efficiency for rice production in Lamongan Regency during the rainy season using the Stochastic Frontier Analysis (SFA) approach. This method considers local inputs such as land area, labor, and capital, and integrates statistical evaluation techniques such as R-squared and log-likelihood tests to assess model performance. The analysis is further supported by distributed computing systems for efficient data processing. The final efficiency evaluation results can be displayed interactively using dashboard visualization systems to support faster and more informed decision-making [10].

2. RESEARCH METHOD

This study adopts a *Stochastic Frontier Analysis* (SFA) approach to develop a predictive model of technical efficiency in rice production in Lamongan Regency during the rainy season. The process consists of structured stages: data collection, preprocessing, modeling, evaluation, and cross-validation.

2.1. Research Flow and Dataset

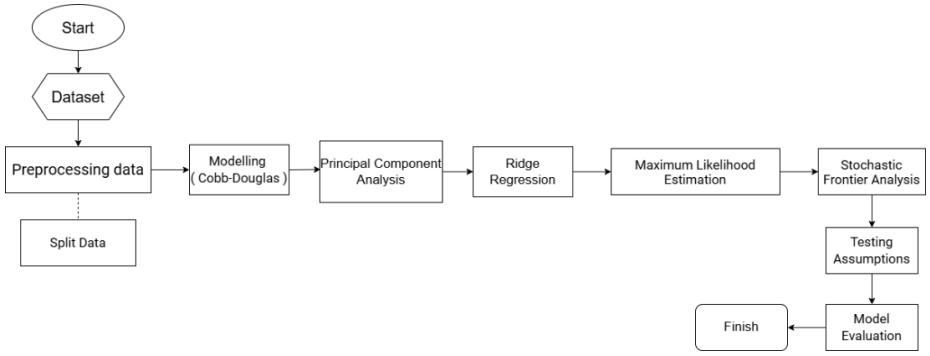


Figure 1. Research Flowchart

Dataset

Table 1 presents the dataset used in this study, which consists of 13 attributes covering various agricultural production factors

Table 1. Dataset Features

No.	Attribute	Description
1.	Name	Farmer Identity
2.	Land Area (ha)	Land Area (hectar)
3.	Number of Workers	Total Number of Workers
4.	Fertilizer Usage (kg)	Amount of Fertilizer (kg)
5.	Pesticide Usage (liter)	Amount of Pesticide (liter)
6.	Number of Seeds (kg)	Amount of Seeds (kg)
7.	Water Usage (m^3)	Volume of Irrigation Water (m^3)
8.	Production Cost (Rp)	Total Production Cost in Rupiah
9.	Crop Yield (ton)	Total Crop Yield (ton)
10.	Rainfall (mm)	Rainfall (mm)
11.	Temperature (°C)	Air Temperature (°C)
12.	Soil Quality (pH)	Soil Acidity Level
13.	Market Price (Rp)	Selling Price of Crop (Rp)

These attributes include farmer data, land area, labor, production inputs such as fertilizer, pesticides, seeds, and water, as well as environmental variables like rainfall, temperature, and soil quality. In addition, the dataset includes yield output and selling price. This information serves as input variables for analyzing production efficiency using the Stochastic Frontier Analysis (SFA) method.

2.2. Data Preprocessing

Data preprocessing begins with data cleaning and missing value imputation, followed by normalization using the Z-score method [11].

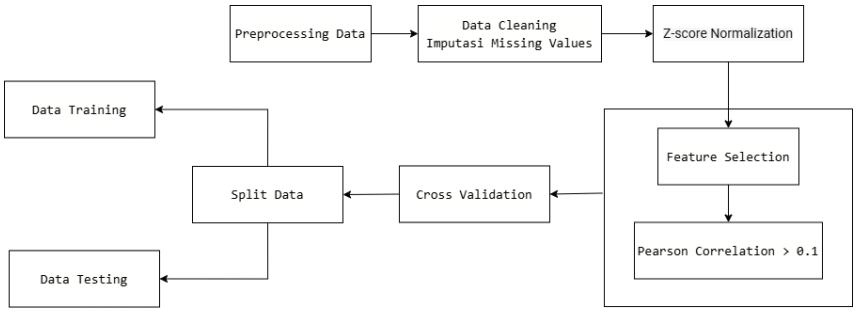


Figure 2. Preprocessing Data

Feature selection is performed based on Pearson correlation values > 0.1 and VIF values < 10 to avoid multicollinearity [12]. The processed data is then split into training and testing sets, which are used in the modeling and evaluation stages.

Table 2. Data Pre-processing

Variabel	Imputation	Z-Score Normalization:	Pearson Correlation
Land Area (ha)	0	Mean: 1.276	0.9853
		Std: 0.631	Status: *Valid*
Number of Workers	0	Mean: 101.95	0.9898
		Std: 51.415	Status: *Valid*
Fertilizer Usage (kg)	0	Mean: 648.583	0.7676
		Std: 324.152	Status: *Valid*
Number of Seeds	0	Mean: 44.07	0.9752
		Std: 22.795	Status: *Valid*
Water Usage	0	Mean: 46.764	0.9841
		Std: 24.017	Status: *Valid*
Rainfall	0	Mean: 35	0
		Std: 0	Status: *Removed*
Temperature	0	Mean: 30	0
		Std: 0	Status: *Removed*
Soil Quality	0	Mean: 6	0
		Std: 0	Status: *Removed*

Table 2 shows the data preprocessing steps, including Z-score normalization and Pearson correlation for each variable. Variables with low or zero correlation were removed, while those with significant correlation were considered valid and used for further analysis. Normalization helps standardize the data to ensure more accurate and consistent analysis results.

2.3. K-Fold Cross Validation

To evaluate the model's reliability and generalizability, 5-fold cross-validation was applied. In this approach, the dataset was divided into five approximately equal parts. For each iteration, four parts were used to train the model while the remaining one served as the validation set. This process was repeated five times so that each subset was used once for testing.
The cross-validation results are as follows:

- The average R-squared (R^2) value was 0.9132, indicating that the model explains over 91% of the variance in rice yield.
- The average Mean Absolute Percentage Error (MAPE) was 12.95%, which falls into the "good" category and suggests accurate prediction performance.

This validation process confirms the model's stability and robustness across different data splits. The consistent R^2 and MAPE values across folds demonstrate that the model is not overfitted and is capable of generalizing well to unseen data.

2.4. Production Modeling Cobb-Douglas

A log linear Cobb Douglas production function was adopted to represent the relationship them between agricultural input variables and rice yield. This functional form is widely used in agricultural efficiency studies because it captures the elasticity of output with respect to each input and assumes multiplicative interactions among. The Cobb-Douglas function is defined as:

$$ln(Y_i) = \beta_0 + \beta_1 \mathrm{In}(X_{1i}) - \beta_2 \mathrm{In}(X_{2i}) + \cdots + \beta_n \mathrm{In}(X_{ni}) + V_I - U_1 \quad (2)$$

2.5. Dimensionality Reduction using PCA

Principal Component Analysis (PCA) is used to address multicollinearity. PCA generates orthogonal components. The first three principal components, which explain over 85% of the variance, are used in the modeling phase.General Formula for Principal Component Transformation:

$$Z = X \cdot W \quad (1)$$

2.6. Parameter Estimation MLE and Ridge-Lasso

In this study, two different approaches were applied at different stages of model development:

1. Preliminary Regularized Ridge Regression

Ridge regression is employed only in the initial modeling phase as a baseline to explore the relationship between input variables and rice yield. These regularization techniques help prevent overfitting and improve generalizability by shrinking less important coefficients.

Ridge Regression:

$$\min \beta \left\{ \sum_{i=1}^{n} \left(\text{In } Y_1 - \hat{Y}_i \right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \right\} \quad (2)$$

Between the two, Ridge regression provided better cross-validation performance, and its coefficients were used for initial comparison. However, these models were not used for technical efficiency estimation.

2. Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is the primary method used to estimate parameters in the Stochastic Frontier Analysis (SFA) model. MLE works by finding the parameter values that maximize the likelihood that the observed data were generated by the model. The estimation is based on a log-likelihood function derived from the joint distribution of random error (v_i) and technical inefficiency (u_i) [14].

$$ln(y_i) = \beta_0 + j = \sum_{j=1}^{k} \beta_j \, In(x_{ji}) + v_i - u_i \quad (3)$$

This method estimates two variance components:

- σ_u^2 : Variance due to technical inefficiency
- σ_v^2 : Variance due to random noise

MLE was performed by optimizing the log-likelihood function based on the combined distribution of model residuals. The gamma (γ) value was also computed to indicate the proportion of output variation attributable to inefficiency, along with a 95% confidence interval.

2.7. Stochastic Frontier Analysis

Stochastic Frontier Analysis (SFA) is a method used to assess technical efficiency in production. It accounts for external factors beyond the farmer's control such as weather or natural events that may affect output [15]. This is achieved by separating the model's error term into two components: one representing the farmer's inefficiency and the other capturing random external shocks.

$$\hat{\beta} = (XTX)^{-1} X^T \cdot Y \quad (4)$$

$$TE_I = exp(-U_i) \quad (5)$$

$$\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} \quad (6)$$

2.8. Technical Efficiency Estimation

After building the Stochastic Frontier model using the Cobb-Douglas production function, technical efficiency (TE) is calculated for each observation unit. Technical efficiency reflects how close a producer (e.g., a farmer) is to the maximum production frontier when utilizing available inputs. Technical Efficiency Formula:

$$TE_i = \exp(-u_i) \quad (7)$$

Efficiency values range between 0 and 1. A value near 1 indicates that the unit is highly efficient, while values closer to 0 reflect substantial inefficiency in resource use.

2.9. Statistical Assumption Tests

To ensure statistical validity, assumption tests were performed including residual normality, homoskedasticity, and parameter significance. These tests are essential for reliable and interpretable model estimation.

Table 3. Statistical Assumption Tests and Hypotheses

Test Type	Hypothesis (H_0 vs H_1)	Significance Criteria	Interpretation
Shapiro–Wilk	H_0: Residuals are normally distributed H_1: Residuals are not normally distributed	$p > 0.05 \rightarrow$ Accept H_0 $p \leq 0.05 \rightarrow$ Reject H_0	Tests normality of residuals
Kolmogorov–Smirnov	H_0: Residuals follow normal distribution H_1: Residuals deviate from normal distribution	$p > 0.05 \rightarrow$ Accept H_0 $p \leq 0.05 \rightarrow$ Reject H_0	Alternative test for residual normality
Breusch–Pagan	H_0: Residual variance is constant (homoskedastic) H_1: Residual variance is not constant (heteroskedastic)	$p > 0.05 \rightarrow$ Accept H_0 $p \leq 0.05 \rightarrow$ Reject H_0	Assesses variance consistency of residuals across inputs
Wald	H_0: Coefficient is not statistically significant H_1: Coefficient is statistically significant	$p \geq 0.05 \rightarrow$ Accept H_0 $p < 0.05 \rightarrow$ Reject H_0	Assesses the statistical significance of model coefficients

2.10. Model Evaluation

The performance of the proposed model was evaluated using multiple statistical indicators to ensure its accuracy, reliability, and compliance with key assumptions in regression modeling.

Metrik Evaluasi Log-Likelihood

Log-likelihood is a key metric for assessing model fit in the context of Maximum Likelihood Estimation (MLE). In Stochastic Frontier Analysis (SFA), the log-likelihood value indicates how well the model represents the observed data [16]. A higher log-likelihood value suggests a better fit of the model to the data.

$$LR = -2[lnL_0 - lnL_1] \quad (8)$$

Prediction Evaluation Metrics

The model is evaluated using the following metrics:

R^2

The R^2 value, or pseudo-R^2 in the context of SFA, is still used as an indicator of how much variation in the output can be explained by the model's input variables. A higher R^2 value indicates that the model provides a better explanation of the observed data [17].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2} \quad (9)$$

Table 4 R-Squared Score Categories

Group	Score	Standard
1	0,0	Very Poor Model
2	0,5	Fairly Good Model
3	0,7 – 0,9	Good Model
4	1,0	Highly Accurate Model

MAPE

MAPE (Mean Absolute Percentage Error) is a statistical metric used to evaluate the accuracy of a model's output predictions. In the context of SFA, MAPE compares the predicted output values to the actual observed values in percentage terms. A lower MAPE value indicates that the model produces more accurate predictions of the real-world observations [15].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| x\ 100\% \quad (10)$$

Table 5. MAPE Score Categories

Group	Persentase	Standard
1	< 10%	Very Good
2	10% ≤ MAPE < 20%	Good
3	20% ≤ MAPE < 50%	Fair
4	≥ 50%	Poor

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics

Table 5 presents the significant variation in crop productivity among farmers, indicating potential technical inefficiency in the use of agricultural inputs.

Table 6. Descriptive Statistics

Variable	Units	Statistics						
		Mean	Median	Std. Deviation	Min	Q1	Q3	Max
Land Area	hectar	1.28	1.00	0.63	0.50	0.60	1.60	3.00
Number of Workers	person	101.95	79.00	51.42	40.00	40.00	129.00	237.00
Fertilizer Usage	kg	648.58	545.00	324.15	1.64	272.50	817.50	1,362.50
Pesticide Usage	kg	5.44	4.30	2.83	2.15	2.15	6.45	15.00
Number of Seeds	kg	44.07	35.00	22.80	17.50	17.50	52.50	120.00
Water Usage	m^3	46.76	37.00	24.02	18.50	18.50	55.80	120.00
Rainfall	mm	35.00	35.00	0.00	35.00	35.00	35.00	35.00
Temperature	°C	30.00	30.00	0.00	30.00	30.00	30.00	30.00
Soil Quality	pH	6.00	6.00	0.00	6.00	6.00	6.00	6.00
Crop Yield	kg/ha	12.00	9.50	5.98	4.75	4.75	14.75	28.50

Table 6 highlights the importance of applying the Stochastic Frontier Analysis (SFA) method to measure and analyze technical efficiency, as well as to identify areas for improvement without the need for direct increases in input usage.

3.2. Model Estimation: Cobb-Douglas Production Function

The Cobb-Douglas production function was estimated using the log-linear form to assess the influence of key agricultural input variables such as land area, number of workers, and fertilizer usage on crop yield. This estimation followed prior stages of data preprocessing and feature selection. The coefficients were derived using Maximum Likelihood Estimation (MLE), which helps identify the elasticity of output with respect to each input. The results, including statistical significance and standard errors, are summarized in Table 7 below.

Table 7. Cobb-Douglas Log Linier

Variable	Coefficient	Standard Error	Wald Statistic	p-value	Significant?
Intercept	2.2600	0.0190	119.0252	0.8000	No
Land Area	-0.2116	0.2374	-0.8914	0.6136	No
Number of Workers	0.3043	0.0729	4.1738	0.8000	No
Fertilizer Usage	-0.1279	0.2410	-0.5306	0.5022	No

Table 7 presents the parameter estimation results of the Cobb-Douglas production model. The findings reveal that none of the input variables have a statistically significant effect on crop yield, as all p-values exceed the 0.05 significance threshold. Therefore, the coefficients cannot be interpreted econometrically, indicating that the model has not yet captured a meaningful relationship between inputs and output in terms of production efficiency.

3.3. Statistical Assumption Testing

To ensure the validity of the regression model used in the SFA estimation, several statistical assumption tests were conducted on the training dataset. These include tests for the normality of residuals, homoscedasticity, and multicollinearity among input variables. The results are presented in Table 8.

Table 8. Training Data Model Assumption Tests

Test Type	Method	Statistic	Degrees of Freedom	p-value	Interpretation
Normality (v)	Kolmogrov Smirnov	0.1361	-	0.0748	Residuals are normally distributed ($p > 0.05$)
Homoscedasticity	Breusch-Pagan	0.1399	3	0.0057	Errors are homoscedastic ($p \leq 0.05$)
	Variable				VIF Value
Multicollinearity (VIF)	Land area				85.70
	Number of Workers				129.95
	fertilizer_usage				3.16
	Number of Seeds				127.54
	Water Usage				379.02

The results confirm that the residuals are approximately normally distributed, indicating that the model satisfies the assumption of normal error terms. However, heteroscedasticity is present, suggesting that the variance of the errors is not constant across observations. This may affect the reliability of standard error estimates. Furthermore, several input variables exhibit severe multicollinearity, as indicated by high VIF values. This condition can inflate coefficient variances and reduce the interpretability of individual predictors, which justifies the removal of multicollinear variables in the modeling process.

Table 9. Testing Data Model Assumption Tests

Test Type	Method	Statistic	Degrees of Freedom	p-value	Interpretation
Normality (v)	Kolmogrov Smirnov	0.1482	-	0.2678	Residuals are normally distributed ($p > 0.05$)
Homoscedasticity	Breusch-Pagan	0.0609	3	0.0017	Errors are homoscedastic ($p \leq 0.05$)
	Variable				VIF Value
Multicollinearity (VIF)	Land area				81.59
	Number of Workers				135.74
	fertilizer_usage				1.77
	Number of Seeds				343.79
	Water Usage				457.81

The normality test results indicate that the residuals from the testing data are normally distributed, which supports the model's reliability in prediction. However, the presence of heteroscedasticity suggests that the variance of the errors varies across observations, which could affect inference precision. In terms of multicollinearity, several predictor variables exhibit very high VIF values, indicating severe linear dependencies that could distort parameter estimates. As a result, these multicollinear variables should be reconsidered or excluded from the final model to enhance its stability and interpretability.

3.4. Model Performance Evaluation

To assess the effectiveness and reliability of the SFA-based model, several performance metrics were computed for both training and testing datasets. These metrics include Log-likelihood, R-squared (R^2), Mean Absolute Percentage Error (MAPE), gamma (γ), and the variance components of inefficiency (σ^2_u) and random error (σ^2_v). The evaluation provides insights into how well the model fits the data, its predictive accuracy, and the contribution of inefficiency to total error.

Table 10. Model Evaluation

Model Evaluation	Log-likehood	R-Squared	MAPE	γ	σ^2_u	σ^2_v
Data Training	13.2971	0.9157	13.44%	0.0100	0.0003	0.0250
Data Testing	7.3845	0.9157	13.01%	0.0100	0.0002	0.0223

Table 10 shows the model demonstrates consistent performance across training and testing datasets. The identical R-squared values indicate that the model explains approximately 91.57% of the variance in crop yield, which is a strong indicator of goodness-of-fit. MAPE values below 15% further confirm that the model's predictions are accurate and fall within an acceptable error margin.

Gamma (γ) remains low, indicating that most of the variance is due to random noise rather than inefficiency, which aligns with the small values of σ^2_u. Meanwhile, the slightly higher σ^2_v values confirm that stochastic variability plays a larger role than inefficiency in the output deviations. This suggests that external factors or noise have a greater influence on yield variability than internal inefficiencies, and the model has successfully captured that dynamic.

3.5. Technical Efficiency Analysis

Technical Efficiency Analysis aims to measure how efficiently farmers utilize inputs to achieve maximum possible output. This section evaluates the degree of inefficiency for each farmer using the SFA model, helping to identify areas where improvements in resource use can be made.

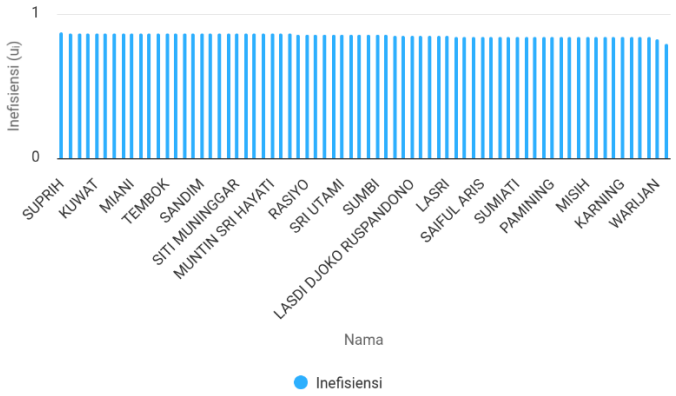


Figure 3. Technical Inefficiency Chart for Training Data

Figure 7 illustrates the relationship between technical inefficiency levels (X-axis) and crop yields (Y-axis) for both training and testing datasets. Most data points are clustered in the inefficiency range of 0.80 to 0.88, indicating that the majority of farmers have not yet achieved full technical efficiency (efficiency = 1), despite producing yields. While a higher inefficiency score does not always directly correlate with lower yields, the overall pattern suggests that yield potential is not fully realized. These results reinforce the earlier findings: improving technical efficiency offers a viable path to increasing crop yields without additional input usage.

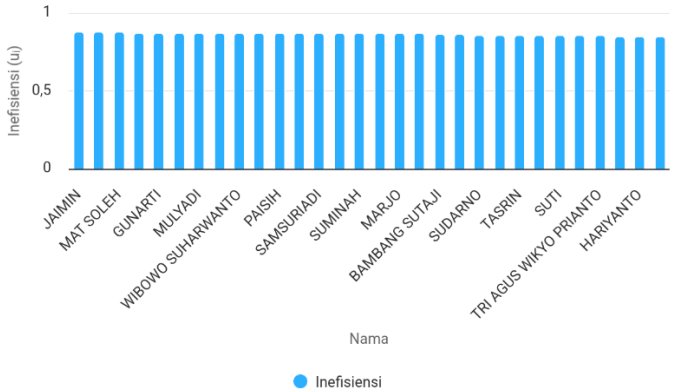


Figure 4. Technical Inefficiency Chart for Testing Data

Figure 4 shows that most individuals in the testing data exhibit relatively high and consistent levels of technical inefficiency, with many values clustering around 0.87. This indicates that despite the use of inputs, most farmers have not yet reached their maximum production potential under efficient conditions. Only a few individuals, such as Nano Sasulihono and Hariyanto, demonstrate lower inefficiency scores, meaning they are closer to

optimal efficiency levels. Overall, this suggests there is still substantial room for technical improvement among nearly all individuals in the testing dataset.

3.6. Visualization of Model Results

The visualization of prediction results on the training data aims to evaluate how well the developed model can map actual yield values to predicted outcomes. The following chart presents a comparison between actual and predicted yields for each sample in the training dataset. Each bar represents two values for the same individual: the actual yield and the model's predicted yield. By examining the alignment between these values, the model's accuracy and generalization ability on the training data can be assessed.

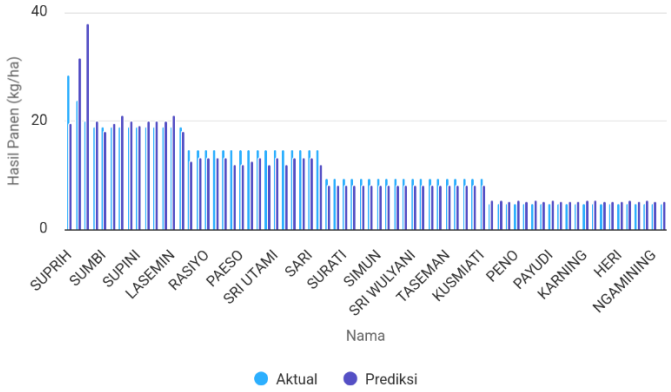


Figure 5. Actual vs Predicted Chart for Training Data

Figure 5 illustrate that the model performs well for most farmers, particularly in the low to medium yield range. The predicted values closely align with the actual outcomes, indicating that the model successfully captures the technical efficiency patterns. This reflects the model's strong generalization ability on the training data.

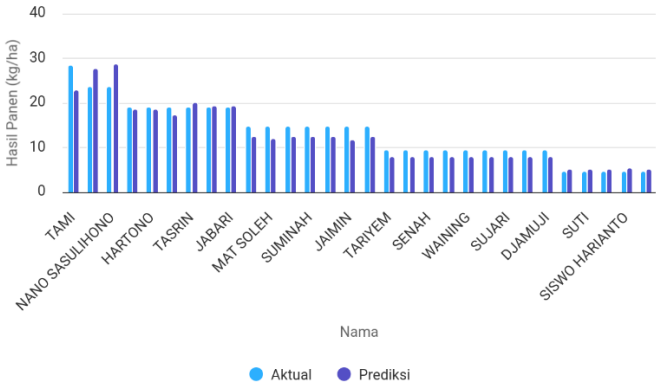


Figure 6. Actual vs Predicted Chart for Testing Data

Based on figure 6 this chart shows the comparison between actual and predicted rice yields in the testing dataset. Overall, the model performs well, especially for farmers with moderate yield levels (around 9–20 kg/ha), where the predicted values closely match the actual outcomes. Although there are some discrepancies for higher-yield cases like Tami and Nano Sasulihono, the deviations remain within an acceptable range. This indicates that the model has strong generalization ability and can reliably capture technical efficiency patterns beyond the training data.

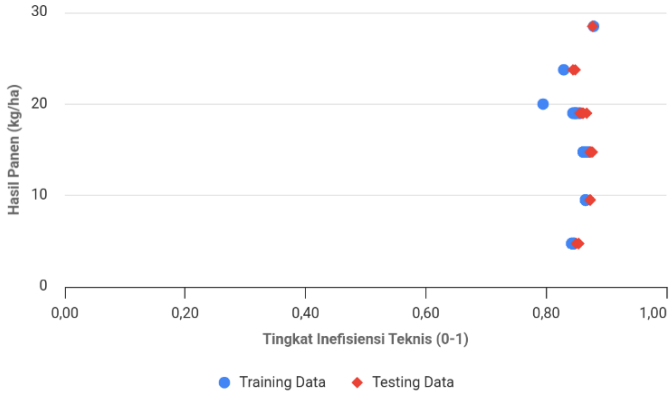


Figure 7. Relationship Between Technical Efficiency and Crop Yield

Figure 9 presents a scatter plot illustrating the relationship between technical inefficiency (horizontal axis, ranging from 0 to 1) and crop yield (vertical axis, in kg/ha). Blue dots represent training data, while red dots indicate testing data. The plot shows a general tendency for crop yields to decrease as technical inefficiency increases. However, the relationship is not strictly linear, and variations in yield are observed even at similar levels of inefficiency. This pattern suggests that while technical inefficiency negatively impacts yield, other contributing factors also influence production outcomes.

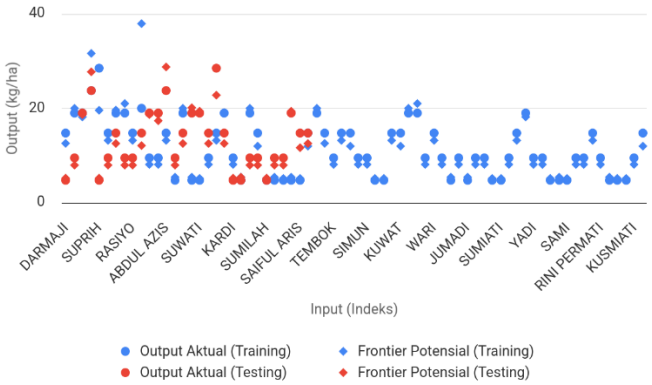


Figure 8. Efficiency Frontier Curve

The graph in Figure 7 most of the actual output points fall below the potential frontier curve, indicating a production gap caused by technical inefficiency. In other words, if the same inputs were used more efficiently, farmers could achieve higher yields. The vertical distance between the actual output points and the potential frontier represents the degree of inefficiency for each individual. These results reinforce the conclusion that, despite input availability, most farmers have not reached optimal production levels. Improving technical efficiency could significantly increase crop yields without requiring additional inputs.

4. CONCLUSION

This study successfully utilized the Stochastic Frontier Analysis (SFA) approach to model and evaluate technical efficiency in rice farming. By incorporating agricultural input and environmental variables into a data-preprocessed and feature-optimized framework, the model provided robust estimations using the Cobb-Douglas production function. The model achieved high predictive performance, with R-squared values of 0.9157 and MAPE below 14% for both training and testing datasets, indicating reliable accuracy. Statistical assumption tests confirmed the normality of residuals, but also highlighted the presence of heteroscedasticity and multicollinearity, which were mitigated through PCA and feature selection.

Efficiency estimation revealed that most farmers had suboptimal technical performance, as visualized through efficiency frontier plots and yield comparisons. Although many farmers operated near the average efficiency range, a considerable portion still had potential to improve output without increasing input usage. These insights reinforce the role of SFA as a valuable analytical tool for diagnosing production gaps and supporting smarter, data-informed agricultural decisions.

REFERENCES

[1] T. Trimono, S. A. Putri, dan N. Selayanti, "Analisis Perbandingan Tingkat Kemiskinan di Jawa Timur Selama dan Sesudah Krisis Pangan," Prosiding Seminar Nasional Sains Data, vol. 3, no. 1, pp. 235–240, 2023.

[2] D. A. Prasetya, A. Sanusi, G. Chandrarin, E. Roikhah, I. Mujahidin, and R. Arifuddin, "Small and Medium Enterprises Problem and Potential Solutions for Waste Management," *J. Southwest Jiaotong Univ.*, vol. 54, no. 6, 2019.

[3] D. A. Prasetya, A. P. Sari, M. Idhom, and A. Lisanthoni, "Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports," *Indones. J. Electron., Electromed. Eng. Med. Inform.*, vol. 7, no. 1, pp. 113–122, 2025.

[4] A. Lisanthoni and D. A. Prasetya, "Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports," Indones. J. Electron., Electromed. Eng. Med. Inform., vol. 7, no. 1, pp. 113–122, 2025.

[5] R. A. Siagian and W. Soetjipto, "Cost Efficiency of Rice Farming in Indonesia: Stochastic Frontier Approach," *Agricultural Socio-Economics Journal*, vol. 20, no. 1, pp. 7–14, 2020.

[6] A. Muhaimin, "Implementasi MySQL untuk pengelolaan data besar dalam sistem informasi pertanian," *J. Teknol. Inf. dan Komput.*, vol. 7, no. 3, pp. 115–121, 2021.

[7] P. A. Riyantoko, "Optimizing User-PC Computing System with Multicore CPU Utilization through Parallel Computing Jobs Distributions," in *Proc. Semin. Nas. Teknol. Komput. dan Informatika (SENATIK)*, vol. 1, no. 1, pp. 101–106, 2023.

[8] N. Birudaraju, A. Rao, and S. V., "Architecture for Analyzing Agriculture Data Using Data Analytics," in Big Data for Improved Health Outcomes, IGI Global, 2021, pp. 111–122, doi: 10.4018/978-1-7998-2566-1.ch006.

[9] J. Singvejsakul, C. Intapan, C. Chaiboonsri, dan R. Permsiri, "Bayesian Stochastic Frontier Analysis of Agricultural Productivity Efficiency in CLMV," Journal of Physics: Conference Series, vol. 1936, 2021, doi: 10.1088/1742-6596/1936/1/012006.

[10] K. M. Hindrayani, "Penerapan Dashboard System (Pendapatan Asli Daerah (PAD) Sektor Pariwisata pada Kabupaten Mojokerto Menggunakan Tableau," *J. Comput. Sci. Technol. (JCS-TECH)*, vol. 4, no. 1, pp. 14–18, 2024.

[11] D. A. Prasetya and E. F. Armay, "Resolving the Shortest Path Problem using the Haversine Algorithm," J. Ilm. Teknol. dan Rekayasa, vol. 26, no. 1, pp. 73–80, 2021.

[12] A. Fakhri, "Penerapan Metode Variance Inflation Factor (VIF) untuk Mengatasi Multikolinearitas dalam Model Regresi," J. Stat. dan Apl., vol. 4, no. 1, pp. 15–22, 2020.

[13] M. Awangga and S. Adinugroho, "Evaluasi Kinerja Model Prediksi Harga Saham dengan Proporsi Data Latih dan Uji yang Berbeda," *J. Ilmu Komput. dan Agri-Informatika*, vol. 5, no. 1, pp. 37–44, 2024.

[14] S. Fitriani and A. Kurniawan, "Estimasi Parameter Model Regresi Logistik dengan Metode Maximum Likelihood Estimation (MLE)," J. Mat. dan Aplikasinya, vol. 7, no. 1, pp. 45–52, 2021.

[15] R. Purnamasari and B. Kusuma, "Analisis Efisiensi Teknis Usahatani Padi Sawah di Kabupaten Subang Menggunakan Metode SFA," J. Apl. Sist. Inf. dan Komput. (JASIEK), vol. 3, no. 2, pp. 112–121, 2022.

[16] P. Damaliana, "Evaluasi model regresi menggunakan R-squared dan log-likelihood dalam analisis efisiensi produksi," J. Ekon. Pertan., vol. 1, pp. 10–18, 2020.

[17] A. Pramono, T. J. L. Tama, and T. Waluyo, "Analisis Arus Tiga Fasa Daya 197 KVA dengan Menggunakan Metode Uji Normalitas Kolmogorov-Smirnov," J. RESISTOR (Rek. Sist. Komput.), vol. 4, no. 2, 2021.

[18] M. Kroshko, K. Morin, dan R. Sahgal, "A knowledge-based planning algorithm for lung SBRT with dynamic conformal arc therapy using Bayesian stochastic frontier regression," Journal of Applied Clinical Medical Physics, vol. 23, no. 10, pp. e13766, 2022

[19] R. Rauf, K. Ayinde, dan A. Ogundipe, "Enhanced methods for multicollinearity mitigation in stochastic frontier analysis using principal component analysis (PCA-SFA)," Heliyon, vol. 10, no. 1, 2024.

[20] V. Glinskiy et al., "Modifications to the Jarque–Bera Test," Mathematics, vol. 12, 2024.

[21] M. Awangga and S. Adinugroho, "Evaluasi Kinerja Model Prediksi Harga Saham dengan Proporsi Data Latih dan Uji yang Berbeda," J. Ilmu Komput. dan Agri-Informatika, vol. 5, no. 1, pp. 37–44, 2024

[22] R. Zulkarnain, A. Djuraidah, dan I. Sumertajaya, "Spatial autoregressive stochastic frontier model with application to Indonesia's aquaculture," Journal of Physics: Conference Series, vol. 1863, 2021.
[23] H. Afrifa, A. Tauringana, and B. Tingbani, "Working capital management and technical efficiency: A stochastic frontier analysis approach," Journal of Business flowchart, vol. 149, pp. 37–48, 2022.
[24] D. S. Nugroho and E. Prasetyo, "Analisis Efisiensi Teknis Produksi Jagung di Provinsi Jawa Tengah dengan Pendekatan Stochastic Frontier Analysis," J. Ekon. Pertan. dan Agribisnis, vol. 6, no. 2, pp. 101–112, 2022.
[25] I. G. Sedana and N. M. Yuliarmi, "Pengaruh Faktor-Faktor Produksi Terhadap Efisiensi Teknis Usahatani Cabai Merah di Bali: Aplikasi Stochastic Frontier Analysis," J. Agribisnis dan Agrowisata, vol. 9, no. 1, pp. 1–10, 2020.