Zero : Jurnal Sains. Matematika. dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10 Number 1. 2026
DOI: 10.30829/zero.v10i1.28721
Page: 202-212

Comparative Modeling of Pineapple Production Using Gaussian GLM and Random Forest Regression

[1]Radot MH Siahaan
Department of Actuarial Science, Institut Teknologi Sumatera, 35365. Indonesia

[2]Indah Gumala Andirasdini
Department of Actuarial Science, Institut Teknologi Sumatera, 35365, Indonesia

[3]Fuji Lestari
Department of Actuarial Science, Institut Teknologi Sumatera, 35365, Indonesia

[4]Dwi Mahrani
Department of Actuarial Science, Institut Teknologi Sumatera, 35365, Indonesia

[5]Amalia Listiani
Department of Actuarial Science, Institut Teknologi Sumatera, 35365, Indonesia

Article Info

Article history:

Accepted 25 March 2026

Keywords:

Climatic Factor;
Gaussian GLM;
MAPE;
Pineapple Production;
Random Forest Regression;

ABSTRACT

This study aims to conduct a comparative modelling of pineapple production at PT Great Giant Pineapple (GGP) using Gaussian GLM as parametric statistical approach and Random Forest Regression method as machine learning based on monthly data from 2014 to 2022. Multicollinearity testing and distribution fitting were conducted to validate the Gaussian assumption. For the Random Forest Regression, hyperparameters were optimized by tuning the number of trees (ntree) and the number of predictors at each split (mtry) with model stability evaluated using Out-of-Bag (OOB) error. The Gaussian GLM achieved a MAPE of 8.41% (R^2 = 0.106) for the GP3 clone and 11.27% (R^2 = 0.149) for the F180 clone. Random Forest Regression produced a testing MAPE of 9.28% (R^2 = 0.144) for GP3 and 12.11% (R^2 = 0.105) for F180. While both models achieved low prediction error based on MAPE, they differed in identifying influential variables and showed limited explanatory power as indicated by low R^2 values. The Gaussian GLM identifies air pressure as significant for both clones and rainfall for F180 clone, while Random Forest consistently identifies rainfall as the most influential predictor. These findings confirm the complementary strengths of parametric and machine learning approaches in supporting climate-based production planning and risk mitigation.

Corresponding Author:

Indah Gumala Andirasdini
Department of Actuarial Science
Institut Teknologi Sumatera
Email: indah.andirasdini@at.itera.ac.id

1. INTRODUCTION

This study conducts a comparative analysis of Gaussian GLM and Random Forest Regression using monthly production data from PT Great Giant Pineapple (GGP) for the period 2014–2022 to evaluate predictive accuracy and identify influential climatic variables. Unlike smallholder agricultural datasets, which are often incomplete, the industrial data used in this study are systematically recorded, enabling a more reliable evaluation of climate–yield relationships. This approach aims to bridge the methodological gap between parametric inference and machine learning prediction while providing practical insights for climate-based production planning and risk mitigation in industrial pineapple management.

The agricultural sector plays a crucial role in national economic development by supporting food security and improving farmers' welfare [1], [2]. Climate variability has been widely recognized as a major determinant of agricultural productivity and yield stability worldwide, increasing production uncertainty and operational risk [3], [4]. Among horticultural commodities, pineapple represents an important industrial crop, particularly for large-scale producers such as PT Great Giant Pineapple (GGP), where production stability is essential for maintaining supply chain efficiency and market competitiveness [5]. Pineapple production is highly influenced by climatic factors including rainfall, temperature, humidity, air pressure, and solar radiation.

In agricultural production modeling, statistical approaches such as the Generalized Linear Model (GLM) remain widely used due to their clear inferential interpretation and hypothesis testing capability under explicit distributional assumptions [6], [7]. Previous studies have widely applied the Generalized Linear Model (GLM) to analyze climate–yield relationships in agricultural systems. For instance, GLM-based approaches have identified rainfall and temperature as significant determinants of crop productivity in commodities such as coffee and staple crops, providing interpretable parameter estimates and statistical inference [8], [9]. In terms of predictive performance, GLM also demonstrates strong accuracy, achieving low MAPE values of 9.05% for soybean production and 8.84% for peanut productivity, indicating reliable model performance.

With advances in computational methods, machine learning techniques such as Random Forest Regression have increasingly been adopted for agricultural prediction. Random Forest is capable of capturing nonlinear relationships and complex interactions among variables without strict distributional assumptions [10], [11]. Several studies report that Random Forest often improves predictive accuracy in agricultural and environmental modelling such as maize and other plantation commodities, particularly under heterogeneous climatic conditions [11], [12], [13], [14], [15]. For example, the application of the random forest algorithm to palm oil using the Mtry parameter of 1 and the Ntree parameter of 500 produces a percentage accuracy rate of 100%. Furthermore, Random Forest modeling successfully explained 81% of maize yield variability, identified soil variables as the most important predictors over climatic factors [16].

Despite the growing use of both statistical and machine learning approaches, most existing studies apply these methods independently rather than conducting structured comparisons using the same dataset. Without evaluating models under identical data conditions, it is difficult to determine whether differences in predictive performance arise from model capability or from variations in data context. This creates a methodological gap in understanding the relative strengths of parametric and machine learning approaches for agricultural production modeling.

Therefore, this study evaluates predictive accuracy using MAPE and R-squared and identifies influential climatic variables under both Gaussian GLM and Random Forest approaches, providing methodological insights and practical implications for climate-based production planning and risk mitigation in industrial pineapple systems.

2. RESEARCH METHOD

This study employs two modeling approaches, the Gaussian GLM and Random Forest Regression, to analyze factors affecting pineapple production. The GLM assumes a Gaussian (Normal) distribution with an identity link function, while Random Forest does not rely on distributional or linearity assumptions and evaluates variable importance based on reductions in prediction error measured by Mean Squared Error (MSE). By combining these approaches, the study provides both inferential interpretation through GLM and predictive evaluation through Random Forest, enabling a comprehensive assessment of the determinants of pineapple production for GP3 and F180 clones.

The dataset consists of 108 observation (monthly data from 2014–2022). Climatic and pineapple production data were collected from the company's internal meteorological station managed by the Production Planning and Inventory Control (PPIC) division, PT Great Giant Pineapple (GGP), while production data were systematically recorded as part of routine industrial monitoring activities. The stages of analysis in this study were as follows.

2.1 Data Preprocessing

Data was verified using summary statistics and visualization techniques, and no missing values were detected, therefore, imputation procedures were not required. Outliers with interquartile range (IQR) criterion were assessed using boxplot visualization. However, these observations were not removed because they represent

genuine extreme weather conditions rather than measurement errors. Retaining such extreme climatic events is important to avoid bias in modeling climate–yield relationships. To improve numerical stability and comparability across predictors, variables were standardized using z-score prior to modeling GLM [15] [17] [16]-[19]. Random Forest is relatively robust to outliers due to its recursive partitioning and prediction aggregation mechanisms and generally does not require variable standardization, as tree-based algorithms are insensitive to differences in predictor scales [10], [11], [15]. The research variables are summarized in Table 1. Although the data consist of monthly observations, the Gaussian GLM implemented in this study does not incorporate lagged terms or autoregressive structures; therefore, it should not be interpreted as a time-series model. The primary objective of this study is a comparative methodological evaluation rather than modeling temporal dependence, while incorporating explicit serial correlation structures remains a direction for future research.

Table 1. Research Variable

Variable	Description	Unit
Y_1	Pineapple Yield (GP3 Clone)	ton/ha
Y_2	Pineapple Yield (F180 Clone)	ton/ha
X_1	Rainfall	Millimeter (mm)
X_2	High Humidity	Millimeter (mm)
X_3	Average temperature	Celsius (℃)
X_4	Sunshine Duration	Hours
X_5	Solar Radiation Intensity	Percentage (%)
X_6	Relative Humidity	Percentage (%)
X_7	Wind Speed	Kilometers per hour (Km/h)
X_8	Air Pressure	hectoPascal (hPa)
X_9	Evaporation	Millimeter (mm)

2.2 Multicollinearity Test

Multicollinearity among predictors was examined using the Pearson correlation coefficient and Variance Inflation Factor (VIF) [15], [18], [19], [20]. Correlation coefficients exceeding 0.80 may indicate strong linear relationships among predictors. The VIF measures the inflation of coefficient variance due to multicollinearity. In this study, multicollinearity is considered absent when VIF values are below 10.

2.3 Fitting Distribution

Distribution fitting was conducted using histogram visualization, Q–Q plots, and information criteria. Candidate distributions included Gaussian, Gamma, and Lognormal. The best distribution was selected based on log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC)[21], [22].

2.4 Gaussian Generalized Linear Model

This study uses a Gaussian GLM which is equivalent to the Normal distribution. The model with an identitiy link function expresses the conditional mean of pineapple yield given the predictor variables (μ_i), vector of predictor variables ($\boldsymbol{X_i}$), vector of regression parameters (β), and variance of errors (σ^2) is expressed in Equation (1).

$$Y_i \sim N(\mu_i, \sigma^2)$$

with

$$\mu_i = X_i\beta \quad (1)$$

Parameter estimation is conducted using the Maximum Likelihood Estimation (MLE) approach, which under the Gaussian assumption is equivalent to the Ordinary Least Squares (OLS) estimator [17], [18]. The significance of parameters in the Gaussian GLM is evaluated using a partial t test at the 5% significance level [17], [18], [19], [20].

Model performance is evaluated using the Mean Absolute Percentage Error (MAPE) and R-Squared (R^2) to facilitate comparison with the Random Forest Regression model [15], [20], [23]. MAPE values below 10% indicate low prediction error, while values between 10% and 20% are generally considered acceptable [24], [25], [26]. The coefficient of determination R^2 represents the proportion of variance in pineapple yield explained by the predictor variables, with values closer to 1 indicating stronger explanatory power [23].

2.5 Random Forest Regression

The Random Forest regression procedure begins by dividing the dataset into training and testing subsets. Two key parameters are used: ntree, representing the number of trees, and mtry, representing the number of predictor variables randomly selected at each split. At each node, the algorithm selects mtry variables from the

total p predictors and determines the best split among them to reduce correlation among trees and improve model generalization A common rule is mtry = p/3 where p denotes the number of predictors [10], [11]. Final hyperparameter values were chosen based on OOB performance [10]. For each training dataset $\{(x_i, y_i)\}_{i=i}^{n}$ with K ntree, prediction from the k-th tree ($T_b(x)$), The last prediction of Random Forest model ($\hat{f}_{rf}^{K}(x)$) for a new observation is given by Equation (2). [11][25]

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(x) \tag{2}$$

The dataset was divided into 80:20 training and testing subsets. Within the training set, each tree was constructed using bootstrap sampling, leaving approximately 37% of observations as Out-of-Bag (OOB) samples [27]. These OOB samples provide an internal validation mechanism for evaluating model performance without requiring additional validation data. Sensitivity analysis was conducted by progressively increasing the number of trees and examining the stabilization of the OOB error, with the optimal ntree selected when the prediction error converged and additional trees no longer substantially improved model performance [23], [27].

In addition, variable importance was evaluated based on the reduction in prediction error measured by Mean Squared Error (MSE). Variables that produce larger reductions in MSE across tree splits are considered more influential in predicting pineapple production [10], [23], [27].

3. RESULT AND ANALYSIS

3.1 Descriptive Statistics

Descriptive statistics for each research variable were presented in Table 2.

Table 2. Descriptive Statistics of Research Variable for the Period 2014-2022

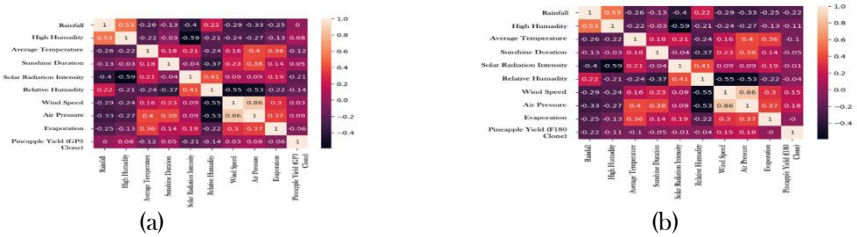
Variable	Mean	Min	Max	Standard Deviation
Pineapple Yield (GP3 Clone)	76.132	25.110	109.720	21.005
Pineapple Yield (F180 Clone)	93.50	47.60	117.04	13.15
Rainfall	203.178	0	746.000	157.014
High Humidity	13.42	0	31.00	7.98
Average Temperature	27.814	24.490	30.200	0.655
Sunshine Duration	6.29	1.45	22.83	3.58
Solar Radiation Intensity	56.221	20.500	88.340	16.174
Relative Humidity	86.904	73.840	92.700	5.827
Wind Speed	4.799	2.300	8.770	1.202
Air Pressure	7.281	3.800	11.600	1.732
Evaporation	4.470	1.780	10.800	1.128

Based on Table 2, the average pineapple yield for the GP3 clone is 76.132 tons, ranging from 25.110 to 109.720 with a standard deviation of 21.005, indicating considerable production variability. Among climatic variables, rainfall shows the largest variability with a mean of 203.178 mm and a standard deviation of 157.014 mm, suggesting substantial fluctuations in precipitation levels. In contrast, temperature remains relatively stable with a mean of 27.814°C and a low standard deviation of 0.655°C. Solar radiation intensity and relative humidity exhibit moderate variability, while wind speed, air pressure, and evaporation show comparatively lower dispersion. Overall, these statistics suggest that rainfall variability may play an important role in influencing pineapple production.

3.2 Generalized Linear Model

3.2.1 Multicollinearity Test

Figure 1 ilustrate that several correlations ranging from weak to strong are observed among the climatic variables.



(a) (b)

Figure 1. Heatmap of Correlation Matrix from Research Variable a) Pineapple Yield (GP3 Clone); b) Pineapple Yield (F180 Clone)

Rainfall shows a moderate positive correlation with high humidity (0.53) and relative humidity (0.22), but a moderate negative correlation with solar radiation intensity (−0.40) and wind speed (−0.29). Solar radiation intensity is moderate negatively correlated with high humidity (−0.59), while wind speed and air pressure show a strong positive correlation (0.86), indicating a close linear relationship between these variables. In contrast, correlations between pineapple yield (GP3 and F180 clones) and climatic variables are generally weak, with coefficients close to zero. For example, rainfall shows almost no linear relationship with GP3 yield (0.00), and only small correlations are observed with air pressure (0.08), wind speed (0.03), and temperature (−0.12). A similar pattern is observed for the F180 clone, where most climatic variables exhibit weak associations with production. These results suggest that the relationship between climatic factors and pineapple production cannot be fully explained by simple linear correlations, indicating the need for modeling approaches capable of capturing more complex relationships. Since pineapple production is continuous and strictly positive, subsequent analysis employs a Generalized Linear Model (GLM) to obtain consistent parametric estimates and a structured interpretation.

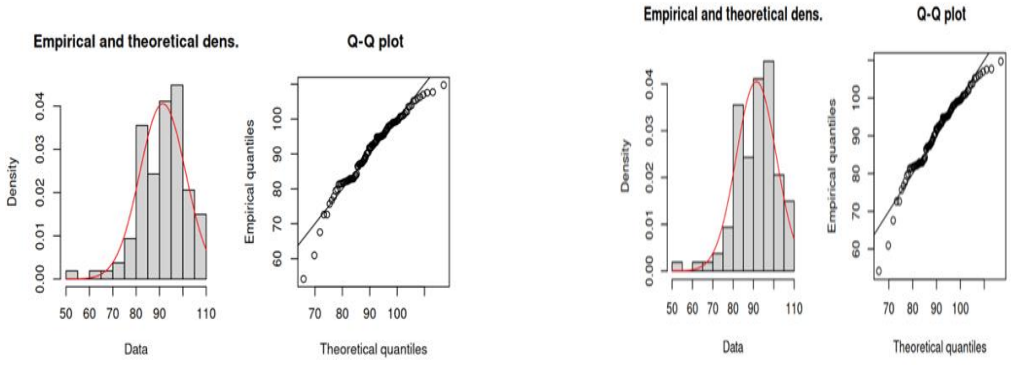
All VIF values from Table 3 are below the commonly accepted threshold of 10. These indicating that there is no severe multicollinearity among the explanatory variables. The highest VIF values are observed for Wind Speed (6.183563) and Air Pressure (6.337869), these finding is consistent with the correlation matrix shown Figure 1 where Wind Speed and Air Pressure exhibit a strong positive correlation. Thus, it can be concluded that there is no indication of multicollinearity among the predictor variables.

Table 3. VIF Value

Variable	VIF	
Rainfall	1.894961	No multicollinearity
High Humidity	2.144229	No multicollinearity
Average Temperature	1.723380	No multicollinearity
Sunshine Duration	1.363127	No multicollinearity
Solar Radiation Intensity	2.342883	No multicollinearity
Relative Humidity	3.296729	No multicollinearity
Wind Speed	6.183563	No multicollinearity
Air Pressure	6.337869	No multicollinearity
Evaporation	1.285334	No multicollinearity

3.2.2Fitting Distribution

Fitting distribution to check asumstion distribution in GLM show at Figure 2



(a) (b)

Figure 2. Empirical and theoretical density comparison and Q–Q plots illustrating the fitting of Pineapple Yield a) GP3 Clone) and b) F180 Clone to Gaussian Distribution

The empirical density plot and Q–Q plot at Figure 2 indicate that pineapple production data approximately follow a Gaussian distribution. Most observations lie close to the theoretical normal line, with only minor deviations at the lower tail. These results suggest that the normality assumption is reasonably satisfied, supporting the use of a Gaussian GLM in subsequent modeling. To ensure this distribution, subsequently, distribution fitting was performed on several candidate distributions, namely Gaussian (Normal), Gamma, and Lognormal. Parameter estimation was carried out using the Maximum Likelihood Estimation (MLE) method.

Table 4. Godness of fit Criteria Distribution

Variable	Distribution	Estimated parameter	Log-Likelihood	AIC	BIC
Pineapple Yield (GP3 Clone)	Gaussian	$\hat{\mu} = 91.3989, \hat{\sigma} = 9.8393$	-396.4706	796.9413	802.2869
	Gamma	$\hat{\alpha} = 79.2068, \hat{\beta} = 0.8665$	-400	805.2118	810.5574
	Log Normal	$\hat{\mu}_{log} = 4.5089,$ $\hat{\sigma}_{log} = 0.1153,$	-403.1746	810.3491	815.6948
Pineapple Yield (F180 Clone)	Gaussian	$\hat{\mu} = 93.5, \hat{\sigma} = 13.0839$	-426.9644	857.9288	863.2744
	Gamma	$\hat{\alpha} = 44.2266, \hat{\beta} = 0.473$	-433.8311	871.662	877.0078
	Log Normal	$\hat{\mu}_{log} = 4.5266,$ $\hat{\sigma}_{log} = 0.1569$	-438.0503	880.1006	885.4463

Based on the log-likelihood, AIC, and BIC values presented in Table 4, for both pineapple clones, the Gamma and Lognormal distributions give AIC and BIC values larger than Gaussian, indicating a poorer model fit. Therefore, it can be concluded that the Gaussian distribution provides the best goodness-of-fit criteria for modeling pineapple production in both GP3 and F180 clones.

3.2.3 Gaussian Generalized Linear Model

The parameter estimates of the Gaussian Generalized Linear Model, obtained through MLE are presented in Table 5.

Table 5. Gaussian Generalized Linear Model Regression Results for Pineapple Yield (GP3 Clone)

	Estimated parameter	t-value	p-value	Lower 95%	Upper 95%	Decision
Intercept	213.8686	3.2780	0.0014	85.997	341.74016	Significant
Rainfall	-0.0034	-0.4101	0.6826	-0.0196	0.0128	No significant
High Humidity	-0.0364	-0.2092	0.8346	-0.3778	0.3049	No significant
Average Temperature	-3.5962	-1.8917	0.0615	-7.3223	0.1297	No significant
Sunshine Duration	-0.1520	-0.4915	0.6241	-0.7585	0.4543	No significant
Solar Radiation Intensity	-0.0969	-1.0792	0.2831	-0.2729	0.0790	No significant
Relative Humidity	-0.2007	-0.6800	0.4981	-0.7794	0.3778	No significant
Wind Speed	-3.3580	-1.7100	0.0904	-7.2068	0.4906	No significant
Air Pressure	2.7973	2.0296	0.0451	0.0961	5.4985	Significant
Evaporation	-0.3714	-0.3903	0.6971	-2.2364	1.4935	No significant

Gaussian Generalized Linear Model Regression for Pineapple Yield (GP3 Clone) can be expressed by

$$y = 213.8686 - 0.0034x_1 - 0.0364x_2 - 3.5962x_3 - 0.1520x_4 - 0.0969x_5 - 0.2007x_6 - 3.3580x_7 + 2.7973x_8 - 0.3714x_9$$

Table 5 shows that only the Air Pressure variable significantly affects the production of the GP3 pineapple clone at the 5% significance level (p-value = 0.0451<0.05). The positive coefficient (2.7973) indicates that for every 1 hPa increase in air pressure, pineapple production is expected to increase by 2.7973 tons/ha, holding other variables constant. Meanwhile, rainfall, high humidity, average temperature, sunshine duration, solar radiation intensity, relative humidity, wind speed, and evaporation do not show statistically significant effects (p-value > 0.05). The intercept value of 213.8686 is statistically significant (p-value = 0.0014 < 0.05) and represents the average production level when all predictor variables are equal to zero. Overall, these results suggest that under the parametric Gaussian GLM, air pressure is the only variable that significantly contributes to the variation in GP3 pineapple production.

Table 6. Gaussian Generalized Linear Model Regression Results for Pineapple Yield (F180 Clone)

	Estimated parameter	t-value	p-value	Lower 95%	Upper 95%	Decision
Intercept	199.8423	2.3616	0.0202	33.9841	365.7004	Significant
Rainfall	-0.0238	-2.2093	0.0295	-0.0449	-0.0027	Significant
High Humidity	0.0007	0.0029	0.9977	-0.4422	0.4435	No significant
Average Temperature	-4.6627	-1.8910	0.0616	-9.4956	0.1702	No significant
Sunshine Duration	-0.5462	-1.3608	0.1767	-1.3328	0.2405	No significant
Solar Radiation Intensity	-0.1242	-1.0662	0.2890	-0.3525	0.1041	No significant
Relative Humidity	0.2699	0.7048	0.4826	-0.4806	1.0205	No significant
Wind Speed	-2.1123	-0.8293	0.4090	-7.1044	2.8798	No significant
Air Pressure	3.7365	2.0902	0.0392	0.2328	7.2401	Significant
Evaporation	-0.4334	-0.3512	0.7262	-2.8525	1.9856	No significant

Gaussian Generalized Linear Model Regression for Pineapple Yield (F180 Clone) can be expressed by

$$y = 199.8423 - 0.0238x_1 + 0.0007x_2 - 4.6627x_3 - 0.5462x_4 - 0.1242x_5 + 0.2699x_6 - 2.1123x_7 + 3.7365x_8 - 0.4334x_9 + e$$

Table 6 shows that Rainfall and Air Pressure variable significantly affects the production of the F180 pineapple clone at the 5% significance level. The positive coefficient of 3.7365 indicates that for every 1 hPa increase in air pressure, pineapple production is expected to increase by 3.7365 tons/ha, assuming other variables remain constant. Similar with the negative coefficient of 0.0238 indicates that for every 1 mm increase in rainfall, pineapple production is expected to decrease by 0.0238 tons/ha, assuming other variables remain constant. Meanwhile, high humidity, average temperature, sunshine duration, solar radiation intensity, relative humidity, wind speed, and evaporation do not show statistically significant effects (p-value > 0.05).

Although several climatic variables were not statistically significant in the Gaussian GLM, this does not necessarily imply that these variables lack predictive relevance. The Gaussian GLM assumes linear relationships and does not explicitly model nonlinear interactions unless specified. Therefore, variables such as rainfall in the GP3 clone (p-value = 0.6826 > 0.05) may not exhibit significant linear effects but may contribute through nonlinear or interaction structures captured by Random Forest Regression.

Table 7. Model Performance Evaluation for Gaussian GLM

	MAPE (%)	R^2
Pineapple Yield (GP3 Clone)	8.4149	0.1058
Pineapple Yield (F180 Clone)	11.2679	0.1492

According to the MAPE evaluation criteria, Table 7 shows that the Gaussian GLM achieves low prediction error for the GP3 clone and acceptable performance for the F180 clone. However, the relatively low R^2 values indicate limited explanatory power, suggesting that additional factors beyond climatic variables may influence yield.

3.3 Random Forest Regression

The dataset was divided into training and testing sets using an 80:20 ratio. The optimal number of trees in the Random Forest Regression model was determined through several experimental scenarios using mtry = p/3 = 9/3 = 3 [11][12]. The accuracy obtained from each tree is presented in Table 8.

Table 8. OOB Performance Evaluation of Random Forest Regression Across Different Numbers of Trees

	mtry	ntree	OOB- MAPE	OOB- MSE
Pineapple Yield (GP3 Clone)	3	50	8.9921	97.0093
		100	8.7351	96.42054
		200	8.8595	95.50618
		500	8.7667	92.88579
		1000	8.8203	93.73952
Pineapple Yield (F180 Clone)	3	50	11.5677	165.9474
		100	11.9940	173.3059

200	11.8125	167.7794
500	11.8617	165.7466
1000	11.7183	164.7236

Table 8 indicates that increasing the number of trees generally improves model performance, with the most not iceable improvement occurring between 50 and 200 trees.

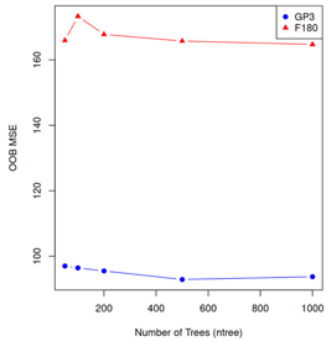


Figure 3. Stability of OOB MSE for GP3 and F180 Clones

Figure 3 shows the convergence of OOB-MSE as the number of trees increases for both clones. The error decreases rapidly at lower ntree values and stabilizes after approximately 500 trees, indicating ensemble convergenc. Therefore, 500 trees were selected as the final model configuration and subsequently evaluated using testing data.

Table 9. Evaluation Performance of Testing Data

	MAPE(%)	MSE	R^2
Pineapple Yield (GP3 Clone)	9.276	103.74	0.1435
Pineapple Yield (F180 Clone)	12.111	169.42	0.1053

This result is consistent with the findings in Table 7. Random Forest Regression also provides low prediction error for the GP3 clone and acceptable performance for the F180 clone. Subsequently, the important variables influencing pineapple production for the both clone will be identified based on the reduction in error, measured by Mean Squared Error (MSE).

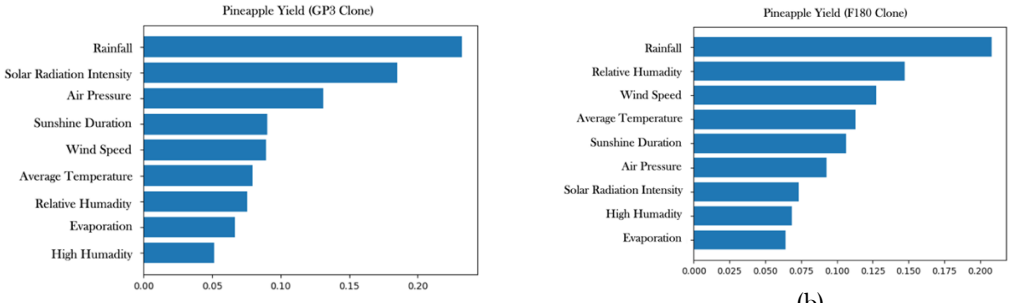


Figure 4. Variable importance a) Pineapple Yield (GP3 Clone); b) Pineapple Yield (F180 Clone)

Figure 4 shows that rainfall is the most important variable influencing pineapple production for both clones under the Random Forest model. For the GP3 clone, rainfall has the highest importance score (≈0.23), followed by solar radiation intensity (≈0.19) and air pressure (≈0.14). A similar pattern is observed for the F180 clone, where rainfall ranks first (≈0.20), followed by relative humidity (≈0.15) and wind speed (≈0.13). The consistently highest importance scores indicate that rainfall contributes the largest reduction in prediction error across trees, highlighting its dominant role in explaining production variability for both pineapple clones.

3.4 Discussion

As shown in Table 10, the Gaussian GLM shows slightly lower MAPE values than the Random Forest model for both clones, although both models exhibit low explanatory power as indicated by the relatively low R^2 values (0.10–0.15). This suggests that additional agronomic factors may influence pineapple production but are not included in the current model. From an inferential perspective, the Gaussian GLM provides statistical significance testing and confidence intervals for parameter estimates (Tables 5 and 6), allowing formal evaluation of linear relationships. The results show that air pressure is the only significant predictor for the GP3 clone, while rainfall and air pressure significantly affect the F180 clone.

Table 10. Comparative Summary of Model Performance

Model	Pineapple Yield	MAPE (%)	R^2	Important Variables
Gaussian GLM	GP3 Clone	8.4149	0.1058	Air Pressure
	F180 Clone	11.2679	0.1492	Rainfall, Air Pressure
Random Forest Regression	GP3 Clone	9.276	0.1435	Rainfall, Solar Radiation
	F180 Clone	12.111	0.1053	Rainfall, Relative Humidity

The differences in identifying influential variables presented in Table 10 indicate that the effect of rainfall does not always increase or decrease proportionally with pineapple yield, suggesting the presence of nonlinear relationships that cannot be fully captured by the linear structure of the Gaussian GLM. In contrast, Random Forest evaluates variable importance based on reductions in prediction error across nonlinear decision trees, allowing rainfall and radiation-related variables to emerge as dominant predictors for both clones. This difference suggests that rainfall variability may influence pineapple production through threshold effects such as excessive soil moisture, waterlogging, or increased disease incidence, which are difficult to represent under linear model assumptions. These results indicate that pineapple yield responses to climatic variability likely involve nonlinear interactions among precipitation, humidity, and radiation conditions. These findings indicate that pineapple yield responses to climatic variability involve complex interactions among precipitation, humidity, and radiation. While GLM offers clear statistical interpretability, Random Forest provides greater flexibility in capturing nonlinear patterns, highlighting the complementary strengths of both approaches. Although additional techniques such as cross-validation or statistical comparison tests could further strengthen the evaluation, this study focuses on MAPE and R^2 to ensure a consistent and interpretable comparison.

4. CONCLUSION

Based on MAPE values, both methods achieve relatively low prediction error. However, they differ in identifying the variables influencing pineapple production for both clones. Under the parametric Gaussian GLM approach, air pressure is a significant variable for both clones, while rainfall is significant only for the F180 clone. In contrast, Random Forest Regression identifies rainfall as the most influential variable for both clones and highlights the contribution of other climatic factors, indicating the presence of nonlinear relationships among variables. These findings confirm that parametric and nonparametric approaches are complementary. The GLM is more stable in inferential interpretation and significance testing, whereas Random Forest is more flexible in capturing complex relationship patterns, consistent with recent studies. This study is limited to climatic variables and production data from a single company within a specific period. Therefore, the findings may not be broadly generalizable to other conditions or production systems. Practically, the results can support harvest planning and risk mitigation in pineapple production management. Future research may incorporate additional agronomic variables or develop hybrid models that combine statistical and machine learning approaches, including the incorporation of temporal correlation structures. For instance, Random Forest can first identify the most influential climatic variables, which are then incorporated into models such as GLM or Generalized Additive Models (GAM) for statistical interpretation. This approach is expected to improve predictive accuracy while maintaining interpretability, potentially reducing prediction error beyond the current MAPE range. Additionally, including agronomic factors such as soil moisture, fertilization schedules, planting density, pest incidence, and lagged yield effects may improve model explanatory power and support more accurate climate-based production optimization.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support provided by the Institute for Research and Community Service (LPPM), Institut Teknologi Sumatera (ITERA), under the 2022 research funding scheme GBU45/Kolaborasi. We also thank to PT Great Giant Pineapple (GGP) for support and agreeing to be a collaborative partner in this research. The support has significantly contributed to the completion of this research.

5. REFERENCES

[1] Kementerian Pertanian Republik Indonesia, "Statistik Pertanian 2023," Jakarta, 2023.

[2] Badan Pusat Statistik, "Hasil Sensus Pertanian 2023," Jakarta, Indonesia, 2023.

[3] Food and Agriculture Organization, "World Programme for the Census of Agriculture 2020," Rome, Italy, 2020.

[4] World Bank, "Agriculture, forestry, and fishing, value added (% of GDP) – Indonesia," https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=ID.

[5] Kementerian Pertanian Republik Indonesia, "Outlook Nanas 2024," Jakarta, Indonesia, 2024.

[6] T. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression: Models, Methods and Applications*, 2nd ed. Berlin, Germany: Springer, 2021.

[7] P. K. Dunn and G. K. Smyth, *Generalized Linear Models with Examples in R Edition: 2*. New York, NY, USA: Springer, 2022.

[8] F. Ceballos-Sierra and S. Dall'Erba, "The effect of climate variability on Colombian coffee productivity: A dynamic panel model approach," *Agric. Syst.*, vol. 190, p. 103126, 2021, [Online]. Available: 10.1016/j.agsy.2021.103126

[9] M. J. Wellington, R. Lawes, and P. Kuhnert, "A framework for modelling spatio-temporal trends in crop production using generalised additive models," *Comput. Electron. Agric.*, vol. 212, Sep. 2023, doi: 10.1016/j.compag.2023.108111.

[10] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. Springer, 2021.

[12] A. Bagus Prakoso, M. A. Putra, M. H. Hilmi, S. Yoma, P. Risky, and J. Maulindar, "Penerapan Algoritma Regresi Random Forest Untuk Prediksi Produksi Jagung Menggunakan Data Statistik Sistem Pertanian Cerdas Smart City," *Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, no. Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) 2025, pp. 28–33, 2025, doi: https://doi.org/10.47701/19h5ny78.

[13] M. Anang Pratama, E. Hermawan, and S. Agustian Hudjimartsu, "Identification Of Potential Forest Fires Using The Random Forest Method In Kubu Raya Regency," *e-Jurnal Penyelidikan dan Inovasi*, vol. 12, no. 5, pp. 1–18, Dec. 2025, doi: 10.53840/ejpi.v12i5.309.

[14] Z. Wirda, A. Ramadhani, and P. Studi Agroekoteknologi, "Model Prediksi Produksi Pertanian Berbasis Machine learning dan Data Lapangan," *SISFO : Jurnal Ilmiah Sistem Informasi*, vol. 9, no. 2, 2025.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[16] Q. Zhang *et al.*, "Maize yield prediction using federated random forest," *Comput. Electron. Agric.*, vol. 210, p. 107930, 2023, doi: https://doi.org/10.1016/j.compag.2023.107930.

[17] A. J. Dobson and A. G. Barnett, *An Introduction to Generalized Linear Models*, 4th ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2018.

[18] A. Agresti, *Foundations of Linear and Generalized Linear Models*. Hoboken, NJ, USA: Wiley, 2015.

[19] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed. Boston, MA, USA: Cengage, 2019.

[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 6th ed. Hoboken, NJ, USA: Wiley, 2021.

[21] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York, NY, USA: Springer, 2002.

[22] I. Gumala Andirasdini, M. Aliem, and A. Sofia, "Regression Models with Arma Errors For Predicting Tabarru Fund In Islamic Insurance: A Normally Distributed Simulation Approach," *PARAMETER: Jurnal Matematika, Statistika, dan Terapannya*, vol. 04, no. 2, pp. 239–248, 2025, doi: 10.30598/parameterv4i1pp239-248.

[23] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[24] I. Gumala Andirasdini, D. Saputra, M. Rivai, S. Eka, and M. Putra, "Analysis of the Health Social Security Administration (BPJS Kesehatan) Claim Amount using Random Forest Regression," 2025.

[25] Syakirah Fachid and Agung Triayudi, "Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19," *MIB Jurnal Media Informatika Budidarma*, vol. 6, no. 1, pp. 68–73, 2022, doi: https://doi.org/10.30865/mib.v6i1.3492.

[26] E. Omotoye and B. Rotimi, "Stationarity in Prophet Model Forecast: Performance Evaluation Approach," *American Journal of Theoretical and Applied Statistics*, vol. 14, no. 3, pp. 109–117, Jun. 2025, doi: 10.11648/j.ajtas.20251403.12.

[27] L. Chen, P. W. Gamage, and J. Ryan, "Debias Random Forest Regression Predictors," *J. Stat. Res.*, vol. 56, no. 2, pp. 115–131, 2022, doi: 10.3329/jsr.v56i2.67466.