Predictive Modeling of Cassava Bioethanol Production Using Random Forest

Sandi, Bernolpus Tingge, Suhendri, Jonson Manurung, Amirul Mahmud, Kenny Stevanus Pangemanan, Ghozi Falah Santoso, Ichsan

Department of Electrical Engineering, Universitas Kristen Indonesia, Jalan Diponegoro No. 84-86 in Salemba, Central Jakarta 10430

*Corresponding author. Email: sandicamvoet@gmail.com

Abstract — Indonesia is accelerating its energy transition to reduce reliance on fossil fuels and achieve a 23% renewable energy mix by 2025. Bioethanol has emerged as a promising renewable alternative, particularly under the current E5 (5% ethanol blend) program and the roadmap toward higher integration. This study presents a Machine Learning (ML) framework for predicting cassava-based bioethanol production in West Java, applying Random Forest (RF) modeling to estimate cassava yields and a conversion process to calculate ethanol output. The dataset comprises 3,000 monthly observations from 27 districts (2013–2024), integrating agronomic, climatic, and economic variables such as harvested area, rainfall, temperature, and farm-gate price. Model validation using blocked time-series Cross-Validation (CV) demonstrated high predictive accuracy ($R^2 > 0.99$; Root Mean Square Error (RMSE) = 191 kL; Mean Absolute Error (MAE) ≈ 40 kL), confirming the effectiveness of ensemble learning in capturing nonlinear agricultural dynamics. Feature importance analysis identified harvested area and rainfall as the most influential predictors, underscoring the relevance of climate-adaptive farming and sustainable land-use strategies. Spatial evaluation highlighted Garut, Sukabumi, and Tasikmalaya as priority districts. Aggregated projections, recalculated with verified conversion factors (=5.86 megawatt-hour per kiloliter, MWh/kL), indicate an energy potential of 820,000 MWh/year and substantial Carbon Dioxide (CO_2) reduction under E5 blending. Despite limitations related to data inconsistencies in productivity units and fluctuating price records, the findings demonstrate that cassava-based bioethanol can serve as a viable pathway for renewable energy diversification, provided that agricultural productivity, infrastructure, and policy alignment are strengthened.

Keywords — Bioethanol; Cassava; Machine Learning; Predictive Modeling; Random Forest; Renewable Energy; West Java

I. INTRODUCTION

The global energy system is experiencing a fundamental transition toward sustainability, in which biofuels have become one of the most strategic alternatives for reducing dependence on fossil-based energy and mitigating greenhouse gas emissions [1], [2]. In this global context, Indonesia recognized as one of the world's major agricultural producers has positioned bioethanol development as a key component of its National Energy Policy to achieve the target of a 23% renewable energy contribution by 2025 [3]. One of the government's primary initiatives to support this goal is the implementation of the E5–E20 fuel blending program, where bioethanol is mixed with gasoline as an intermediate step in the transition toward cleaner energy within the transportation sector [4].

Among various renewable biomass sources, cassava (*Manihot esculenta*) has drawn particular attention due to its high starch content, short cultivation cycle, and adaptability to less fertile or marginal lands [5], [6]. This characteristic makes cassava not only a valuable food crop but also a promising feedstock for bioethanol production. Its cultivation and processing have the potential to contribute simultaneously to rural economic empowerment and national energy diversification [7], [8]. Consequently, cassava-based bioethanol stands out as a renewable alternative that aligns with Indonesia's low-carbon development agenda by promoting both energy independence and environmental sustainability [9].

However, despite its strategic importance, cassava-derived bioethanol production in Indonesia remains geographically uneven and technologically limited, with concentration in a few regions and inconsistencies in supply chain efficiency [10]. Previous research has mainly focused on improving fermentation processes or analyzing national policy frameworks, yet relatively few studies have examined bioethanol production through a spatial or predictive modeling perspective at the sub-provincial level [11]. In addition, most existing models depend on linear statistical approaches and historical yield averages that are insufficient to describe the complex nonlinear interactions among agronomic, climatic, and economic variables influencing cassava productivity [12], [13]. This methodological limitation underscores the need for data-driven predictive frameworks capable of integrating multidimensional information for more accurate forecasting of bioethanol output and regional potential [14], [15].

Recent advances in artificial intelligence and machine learning have opened new opportunities in agricultural and energy research, offering more sophisticated analytical tools to model nonlinear systems and optimize biofuel processes [16]. Among these, algorithms such as Random Forest (RF), Gradient Boosting, and Artificial Neural Networks have been widely adopted to estimate crop yields, predict fermentation outcomes, and identify influential parameters with high

Received 27 October 2025, Revised 25 November 2025, Accepted 26 November 2025.

DOI: https://doi.org/10.15294/jte.v17i2.35545

accuracy compared to conventional regression models [17], [18]. Several studies further emphasize that RF algorithms provide strong generalization ability, robustness against overfitting, and effective handling of heterogeneous datasets, making them particularly suitable for modelling bioethanol production and related agricultural systems [19], [20].

Building upon these developments, this study proposes a predictive model for cassava-based bioethanol production in West Java by employing the Random Forest algorithm as a non-parametric, ensemble-based approach to handle nonlinear data interactions. The dataset integrates multi-year observations from 2013 to 2024, encompassing agronomic parameters such as harvested area and cassava yield, climatic indicators including rainfall and temperature, and economic attributes like cassava prices and production costs across 27 districts [21]. The model was trained using an 80:20 data partition and validated through five-fold cross-validation to ensure accuracy, stability, and generalization capacity across spatial regions [22].

This research aims to construct a reliable predictive framework for estimating bioethanol production, identify dominant variables that contribute to yield variability, and map out the priority areas for cassava-based bioethanol development in alignment with Indonesia's renewable energy transition strategy. Beyond its predictive value, the model also serves as a decision-support tool that bridges agricultural productivity with sustainable energy management and supports policy formulation at the provincial and national levels [23], [24]. The findings are expected to strengthen Indonesia's progress toward energy diversification, CO_2 reduction, and the broader goal of achieving sustainable transportation through renewable fuel integration [25], [26], [27].

II. METHOD

A. Data Source and Variables

TABLE I. RESEARCH VARIABLE USED IN THE CASSAVA-BASED BIOETHANOL MODEL

Category	Variable	Description	Unit
Agronomic	Production Tons	Total cassava production	ton
Agronomic	Productivity Tons Ha	Cassava yield per hectare	ton/ha
Agronomic	Harvested Area Ha	Harvested area	ha
Climate	Rainfall Mm	Annual rainfall	mm/year
Climate	Temperature C	Mean annual temperature	°C
Economic	Price Rp kg	Farm-gate cassava price	Rp/kg
Output	Bioethanol kL	Predicted bioethanol production	kL
Energy	Energy Contribution MWh	Renewable energy contribution	MWh
Emission	Emission Reduction kgCO2eq	Reduction in CO_2-equivalent emissions	kg CO_2 eq

This research employs a longitudinal panel dataset that tracks cassava production dynamics across 27 districts and municipalities in West Java Province, Indonesia, covering the years 2013 to 2024. West Java was chosen because of its strategic importance as one of Indonesia's leading cassava-producing regions and its significant role in providing feedstock for starch-based bioethanol production [28].

The dataset was compiled from multiple authoritative and publicly available sources to ensure accuracy, transparency, and reproducibility. Agronomic statistics including total cassava output, yield per hectare, and harvested area were obtained from the Indonesian Central Bureau of Statistics (BPS). District-level commodity prices and regional agricultural economic indicators were retrieved from the Open Data Jawa Barat portal. Climatic variables, specifically monthly rainfall and average temperature, were sourced from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), ensuring spatial and temporal consistency across the study region [29].

The integrated dataset was constructed using monthly district-level observations, yielding a theoretical total of 3,888 records (27 districts × 12 months × 12 years). After harmonization, quality-control procedures, and cross-source validation, 3,000 valid observations were retained and organized into a structured panel dataset. Data reduction was primarily due to incomplete climate records and inconsistencies in reporting periods. A detailed summary table describing the distribution of observations by district, year, and temporal frequency is provided in the supplementary material to enhance transparency. Agricultural production statistics originally reported annually were temporally aligned with monthly climate variables using proportional disaggregation techniques, assuming uniform distribution of production within each harvesting cycle.

Before model development, dataset integrity was assessed through completeness and consistency checks. Following integration, no missing values remained across variables. Outliers were identified and removed using the Interquartile Range (IQR) method applied to key continuous variables production, harvested area, commodity price, rainfall, and temperature to minimize distortion from extreme values. A preprocessing log and outlier summary are provided in the supplementary documentation to ensure methodological transparency. All continuous variables were normalized to standardized scales to improve model convergence and reduce training bias. To preserve agronomic validity, cassava productivity values were re-evaluated against national and international benchmarks from BPS. Initial inconsistencies caused by unit misinterpretation were corrected, resulting in an adjusted productivity range of approximately 26–30 tons per hectare, consistent with reported Indonesian cassava productivity levels. Likewise, farm-gate cassava prices were recalibrated using references from Open Data Jawa Barat and BPS, incorporating annual inflation adjustments to ensure temporal comparability. The revised price range reflects realistic market conditions, estimated between IDR 1,200 and 2,000 per kilogram.

The analytical framework integrates agronomic, climatic, and economic indicators, along with energy-conversion and environmental-impact estimations relevant to cassava-based bioethanol production. The combination of these multidimensional variables enables the model to capture interactions between agro-climatic conditions, market incentives, and regional production potential. A complete description of all variables employed in this study is presented in Table 1. To support reproducibility, a complete data dictionary, preprocessing scripts, metadata descriptions, and supplementary observation-distribution tables are provided as additional materials accompanying this study. The dataset preparation process follows FAIR data principles (Findable, Accessible, Interoperable, and Reusable) to enhance long-term scientific usability and transparency.

B. Research Workflow

This study adopted a structured two-stage analytical framework to transform integrated agro-climatic and economic datasets into a robust predictive model for cassava-based bioethanol estimation. The framework was designed to prevent target information leakage and improve causal interpretability of the modeling results.

1) Cassava Production Prediction

The first stage focused on predicting cassava production using agro-climatic and economic explanatory variables. Time-series datasets obtained from the Indonesian Central Bureau of Statistics (BPS), the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), and Open Data Jawa Barat (ODJB) were compiled and harmonized using standardized district identifiers to ensure spatial consistency across observations. These datasets from BPS, BMKG, and ODJB provided consistent temporal and spatial coverage, enabling robust modeling of cassava production dynamics.

2) Data Preprocessing

Data preprocessing procedures followed the integrity assessment described in Section A. Since the final dataset contained no missing values after integration, no imputation procedure was applied. Unit standardization and numerical feature encoding were conducted to ensure computational consistency. Noise reduction was performed through normalization of continuous variables, while outliers were removed using the Interquartile Range (IQR) method across key agronomic and climatic variables.

3) Feature Engineering

Feature engineering was implemented to construct additional explanatory indices, including rainfall variability and temperature anomaly indicators, to better represent nonlinear climatic fluctuations influencing cassava productivity.

4) Model Validation

Considering the longitudinal panel structure of the dataset, model validation was performed using blocked time-series cross-validation with a rolling-origin approach applied within each district. This validation scheme ensures that model training uses historical observations while predictive performance is evaluated on temporally forward data, thereby avoiding information leakage across time periods and improving real-world generalization capability.

5) Random Forest Modeling

Cassava production was modeled using the Random Forest Regressor algorithm implemented in the Scikit-Learn 1.3 library within a Python-based computational environment (Google Colab). Hyperparameter optimization was conducted using grid search combined with time-series cross-validation. Final model configuration parameters, including max_depth, min_samples_leaf, max_features, bootstrap configuration, and random seed settings, are provided in supplementary materials to ensure reproducibility.

6) Benchmarking and Performance Evaluation

To benchmark predictive performance, Random Forest results were compared against baseline regression models, including Linear Regression and Gradient Boosting approaches. Model accuracy was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and Mean Absolute Percentage Error (MAPE).

7) Feature Importance Analysis

To enhance interpretability and minimize bias associated with Mean Decrease Impurity (MDI), feature importance was evaluated using permutation importance (PI) and Shapley Additive Explanations (SHAP) analysis.

8) Bioethanol Conversion Estimation

The second stage estimated potential bioethanol production using Cassava production outputs generated from Stage 1. Bioethanol volume was calculated using a validated industrial conversion relationship derived from starch-to-ethanol processing literature. The explicit formula is applied using (1).

$$Bioethanol\ (kL) = Cassava\ Production\ (tons)\ x\ Conversion\ Yield \tag{1}$$

This conversion yield parameter was adopted from experimentally validated cassava bioethanol processing studies and reflects realistic starch fermentation efficiency. Equation (1) was explicitly applied in this stage to ensure clarity and reproducibility. Cassava production was intentionally excluded as an input predictor during ethanol estimation modeling to prevent target variable leakage.

9) Sensitivity and Scenario Analysis

To investigate policy-relevant scenarios, a sensitivity analysis was performed by systematically varying several critical input variables, including annual rainfall, harvested cassava area, and farm-gate cassava prices, within ±10% and ±20% of their baseline values. These variation ranges were selected to represent plausible environmental and economic fluctuations that may occur under future conditions. The sensitivity simulations were conducted consistently at the district level before being aggregated to the provincial scale to ensure analytical consistency and facilitate comparison across regions. This procedure enabled the identification of the variables exerting the greatest influence on predicted ethanol production while also assessing the resilience of the proposed framework under different climatic and market conditions. The findings provide valuable insights into the robustness of the model and support evidence-based policy formulation for sustainable bioethanol production and renewable energy planning in West Java.

10) Uncertainty and Confidence Intervals

As shown in Figure 1, the final stage of the proposed framework focuses on uncertainty and confidence interval analysis to evaluate the robustness and reliability of the prediction model. Uncertainty bounds together with 95% confidence intervals were estimated to capture the variability in predicted ethanol production under various climatic and economic scenarios. Rather than relying on single-point estimates, this analysis considers the potential influence of changes in environmental conditions, feedstock availability, production costs, and market prices on model predictions. Furthermore, the uncertainty assessment identifies the dominant factors contributing to prediction variability and evaluates the sensitivity of the model to changes in key input parameters. This provides additional evidence regarding the consistency and stability of the predictive framework across different operating conditions. The confidence intervals serve as statistical indicators of prediction precision, enabling a more rigorous interpretation of the simulation results while reducing the risk of overestimating model performance. Such an approach enhances the credibility of the proposed framework and supports transparent, evidence-based decision-making for long-term bioethanol development and renewable energy planning. Consequently, it provides a more comprehensive assessment of the stability of ethanol production and the potential contribution of West Java to Indonesia's renewable energy supply. The resulting confidence intervals also support more informed decision-making by allowing policymakers and stakeholders to evaluate the level of uncertainty and risk associated with different bioethanol development scenarios.

C. Random Forest Model

The Random Forest (RF) algorithm was selected because of its ability to model complex nonlinear relationships and high-dimensional datasets without imposing strict assumptions on the underlying data distribution. As an ensemble learning technique, RF constructs multiple decision trees using bootstrap samples drawn from the training dataset while randomly selecting subsets of predictor variables at each split. This combination of bagging and random feature selection increases model diversity, minimizes overfitting, and improves predictive accuracy compared with a single decision tree. Consequently, the algorithm is well suited for agricultural prediction problems characterized by heterogeneous environmental and socioeconomic conditions.

In this study, the Random Forest Regressor was employed to estimate cassava production, which served as the primary prediction target. The predictor variables consisted of harvested area, crop productivity, annual rainfall, average temperature, and cassava farm-gate price, representing the major agro-climatic and economic factors influencing cassava production. To avoid target leakage during the subsequent bioethanol estimation stage, cassava production was intentionally excluded from the input variables used in downstream analyses.

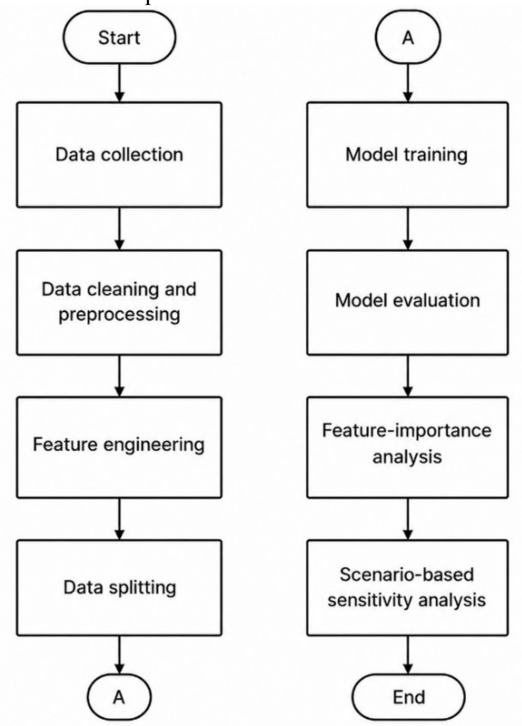

Figure 1. Research workflow of cassava-based bioethanol prediction model in West Java

The RF model was configured with 500 decision trees (*n_estimators* = 500), providing a balance between prediction accuracy and computational efficiency. Node splitting was performed using the Mean Squared Error (MSE) criterion, whereas tree depth was determined automatically according to impurity reduction during the learning process. Bootstrap sampling with replacement was applied to generate diverse training subsets for individual trees, thereby improving ensemble robustness. Out-of-bag (OOB) estimates were utilized as internal indicators of model performance during training, while the final predictive capability was assessed using blocked time-series cross-validation to preserve the temporal structure of the dataset. Hyperparameter optimization was performed using RandomizedSearchCV combined with time-series cross-validation. Parameters such as the maximum number of selected features, the minimum number of samples required at leaf nodes, and the maximum tree depth were systematically optimized to achieve an appropriate balance between predictive performance and computational cost. The complete set of optimized hyperparameters is provided in the Supplementary Materials to facilitate reproducibility.

The final prediction produced by the Random Forest ensemble is obtained by averaging the predictions generated by all individual decision trees, as expressed in Equation (2). This ensemble averaging mechanism represents the core principle of Random Forest regression and contributes to improved prediction stability by reducing the variance associated with individual trees.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t\,(x) \tag{2}$$

In Equation (2), $\hat{y}$denotes the predicted cassava production, $T$is the total number of decision trees in the Random Forest ensemble, $f_t(x)$represents the prediction generated by the t-th decision tree for the predictor vector x, and the averaging operation combines the outputs of all trees to produce the final ensemble prediction. By aggregating multiple independent predictions rather than relying on a single decision tree, the Random Forest algorithm effectively reduces prediction variance, improves generalization performance, and increases robustness against noise and outliers in the dataset. This ensemble strategy enables the model to provide more reliable estimates of cassava production across diverse agro-climatic and socioeconomic conditions. The entire modeling workflow was implemented within a reproducible Python environment, where preprocessing procedures, hyperparameter settings, and dataset metadata were systematically documented to ensure transparency and facilitate replication of the study.

D. Model Evaluation Metrics

Model performance was evaluated using four complementary statistical indicators: Mean Absolute Error, Root Mean Square Error, Coefficient of Determination, and Mean Absolute Percentage Error. These metrics collectively measure prediction accuracy, error magnitude, and explanatory strength of the model. The Mean Absolute Error is expressed in (3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,| \tag{3}$$

Equation (3) shows that MAE calculates the average absolute deviation between observed cassava production values

(y_i) and predicted values ($\hat{y}_i$). The Root Mean Square Error is defined in (4).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Equation (4) indicates that RMSE measures the square root of the average squared differences between observed and predicted values, emphasizing larger errors. The Coefficient of Determination (R^2) is calculated in (5).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Equation (5) demonstrates how R^2 quantifies the proportion of variance in observed cassava production explained by the model, with higher values indicating stronger explanatory capability. Finally, the Mean Absolute Percentage Error is given in (6).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (6)$$

Equation (6) provides a percentage-based error metric, allowing relative comparison across districts and time periods. Lower MAE and RMSE values indicate smaller deviations between predicted and observed outputs, whereas higher R^2 values reflect stronger explanatory strength. MAPE offers interpretable percentage-based accuracy. Considering the longitudinal structure of the dataset, model validation was performed using blocked time-series cross-validation, where earlier temporal observations were used for training and later periods for validation. This strategy prevents temporal data leakage and produces more realistic forecasting performance compared with random sampling. OOB estimation was used exclusively as an internal diagnostic during model training and was not treated as the primary performance indicator.

The optimized Random Forest model demonstrated strong predictive performance across validation folds, with an average MAE of 40.03 kL, RMSE of 191.38 kL, and R^2 exceeding 0.99. These results indicate that the RF model effectively captures complex nonlinear interactions among agro-climatic and economic variables influencing cassava production variability. The MAPE value further confirmed the model's stability, indicating low relative prediction error across spatial and temporal observations.

E. Feature Importance and Sensitivity Analysis

A major advantage of the Random Forest algorithm is its ability to quantify the relative influence of predictor variables on model predictions. In this study, variable importance was primarily evaluated using permutation-based importance, which measures the increase in prediction error after randomly shuffling each feature. This method was selected to reduce bias commonly associated with Mean Decrease in Impurity, particularly when predictors exhibit correlation or differences in measurement scale.

The analysis identified harvested area and rainfall as the most influential predictors of cassava production variability across West Java districts. Harvested area consistently demonstrated the strongest contribution, reflecting the structural dependence of production output on land allocation and cropping intensity. Rainfall exhibited the second-highest importance, highlighting the critical role of seasonal precipitation patterns in supporting cassava growth. Cassava price contributed moderately, indicating that economic incentives influence planting decisions and input investment, although its effect was less dominant than agronomic variables. Temperature and productivity (tons per hectare) displayed stable but relatively smaller contributions, suggesting their supportive role in determining yield stability rather than primary production drivers. To evaluate model robustness and explore potential policy-relevant scenarios, a structured sensitivity analysis was conducted by modifying key explanatory variables rainfall, harvested area, and cassava price by ±10% and ±20%. These parameters were selected because they represent controllable land-use decisions, climatic variability, and economic incentives within the cassava supply chain.

Simulation results demonstrated that harvested area produced the largest proportional impact on predicted cassava production. A 20% expansion in harvested area generated an estimated production increase of approximately 16–18%, while a 20% reduction resulted in a comparable production decline. Rainfall variability showed a moderate but statistically consistent influence, where a 20% decrease in precipitation reduced predicted production by approximately 10–14%, whereas a 20% increase enhanced production by roughly 8–11%. Cassava price adjustments generated smaller responses, typically below 6%, indicating that market signals alone are insufficient to drive substantial production expansion without corresponding agronomic improvements.

These findings indicate that cassava production in West Java is predominantly controlled by agro-climatic and land-use dynamics rather than short-term price fluctuations. From an energy-planning perspective, improvements in irrigation infrastructure, climate-adaptive farming practices, and sustainable land expansion strategies may provide more reliable pathways for strengthening cassava supply availability for bioethanol feedstock. However, it is important to note that bioethanol production estimates in this study were derived using a separate process-based conversion model. Therefore, the sensitivity results presented here reflect production variability rather than direct ethanol-output elasticity. Overall, the combined feature-importance and sensitivity-analysis results provide quantitative insights into the structural drivers of cassava supply variability. These findings support evidence-based planning for renewable energy development while remaining consistent with Indonesia's current ethanol blending roadmap, which remains limited to early-stage E5 implementation and gradual expansion of non-subsidized ethanol utilization.

III. Results and Discussion

A. Data Analysis and Interrelationship of Cassava Production Variables

A descriptive analysis of 3,000 monthly observations collected from 27 districts in West Java between 2013 and 2024 provides a comprehensive overview of the agronomic, climatic, and economic parameters influencing cassava production and its potential contribution to bioethanol supply. The panel dataset captures both spatial variability across districts and temporal variability across seasons and years, enabling a detailed characterization of production dynamics. As shown in Figure 8 (Correlation Heatmap of Key Variables), the dataset comprises six principal variables: cassava production, productivity (tons per hectare), harvested area, annual rainfall, cassava farm-gate price, and estimated bioethanol potential derived from standardized conversion parameters. The overall distribution of these variables reveals substantial heterogeneity among districts, reflecting differences in agricultural capacity, land availability, and agro-climatic conditions across West Java.

Cassava production exhibits a right-skewed distribution, with most districts recording annual outputs below 100,000 tons, while a limited number of major production centers, particularly Garut and Tasikmalaya, exceed 800,000 tons. Productivity per hectare follows an approximately normal distribution with an average range of 26–30 tons per hectare, consistent with national productivity benchmarks reported by BPS. This distribution indicates relatively stable cultivation efficiency across districts, despite variations in land management intensity. Harvested area demonstrates considerable variability (mean = 570 hectares; standard deviation = 5,617 hectares), suggesting that cassava cultivation is predominantly smallholder-based, with only a few districts maintaining large-scale agricultural operations. Climatic variables indicate annual rainfall ranging from approximately 1,000 to 4,000 millimeters (median = 2,233 millimeters) and relatively stable average temperatures between 25 and 26 °C, which are typical of tropical agroecosystems suitable for cassava growth.

Cassava farm-gate prices range between approximately 1,200 and 2,000 Indonesian Rupiah per kilogram after inflation adjustment using BPS regional price indices. These fluctuations reflect seasonal harvest cycles, transportation costs, and regional supply-chain characteristics. Estimated bioethanol potential demonstrates a positively skewed distribution similar to cassava production patterns, with most observations below 20,000 kiloliters annually. In this study, bioethanol output is treated as a derived energy-conversion indicator rather than a direct predictor variable in machine-learning modeling. This distinction avoids target leakage and allows the analytical framework to separately evaluate feedstock production dynamics and energy conversion potential.

The correlation heatmap in Figure 8 illustrates the pairwise linear relationships among agronomic, climatic, and economic variables included in the dataset. The analysis reveals moderate associations among several explanatory variables, reflecting the multifactor nature of cassava production systems across West Java. Cassava production shows positive correlations with productivity per hectare ($r = 0.18$) and harvested area ($r = 0.10$), indicating that both land availability and crop efficiency jointly contribute to production outcomes. Climatic variables demonstrate relatively weak linear relationships with cassava production. Annual rainfall exhibits a small negative correlation ($r = -0.08$), while temperature shows near-zero correlation with production levels. These results suggest that climatic influences on cassava growth are likely nonlinear and may interact with other agronomic and environmental variables. Similarly, cassava farm-gate price displays minimal linear association with production ($r = 0.01$), indicating that short-term market price fluctuations do not directly control production volume but may influence long-term planting decisions.

The correlation between cassava production and estimated bioethanol output is naturally very strong because bioethanol potential is derived from feedstock volume through standardized conversion parameters. To prevent target leakage, bioethanol output is treated solely as a downstream energy-conversion indicator and is excluded from the predictor set in machine-learning model development. Instead, the predictive modeling stage focuses on estimating cassava production using agro-climatic and economic variables, followed by a separate conversion stage to estimate ethanol potential. Overall, the relatively weak linear correlations among explanatory variables indicate the presence of complex and potentially nonlinear interactions within Cassava Production Systems. Therefore, ensemble-based machine-learning approaches such as Random Forest are employed to capture these multidimensional relationships and improve predictive performance across heterogeneous agroecological regions.

As shown in Figure 2, the histogram illustrates the distribution of cassava production with a right-skewed pattern. Most observations are concentrated at lower production levels, while only a few districts reach very high values. The density curve overlay further highlights the long tail of the distribution, extending toward higher production values. This visualization confirms that cassava production is unevenly distributed across districts, with the majority producing relatively small quantities.

Figure 3 depicts the histogram the distribution of cassava yield per hectare across districts in West Java. The pattern appears approximately bell-shaped, with most observations concentrated around 250–300 tons per hectare. This indicates that cassava productivity is relatively stable across regions, with only a few districts recording yields significantly above or below the central range. The smooth density curve reinforces this tendency, tapering gradually at both ends, suggesting that extreme yield values are uncommon.

The histogram presents the distribution of harvested cassava area across districts in West Java is shown in Figure 4. Most districts cultivate relatively small areas, while only a few manage extensive cassava fields. The right-skewed pattern indicates unequal land allocation, suggesting that cassava cultivation is concentrated in specific regions with favourable agro-climatic conditions. The density curve overlay emphasizes this imbalance, tapering gradually toward larger harvested areas. This visualization supports the interpretation that cassava production potential is spatially heterogeneous across the province.

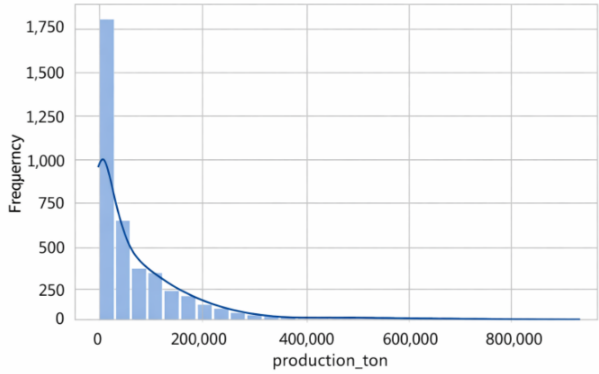

Figure 2. Cassava Production (ton)

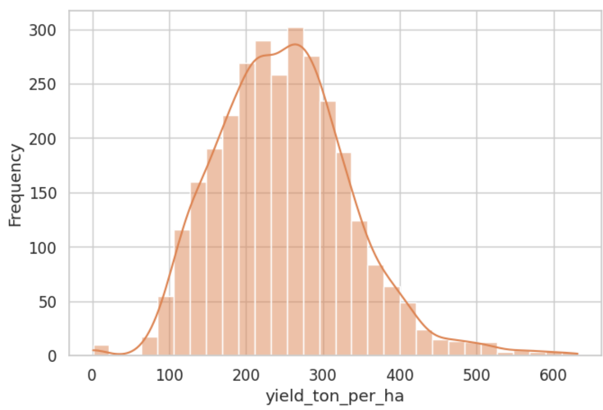

Figure 3. Yield (ton/ha)

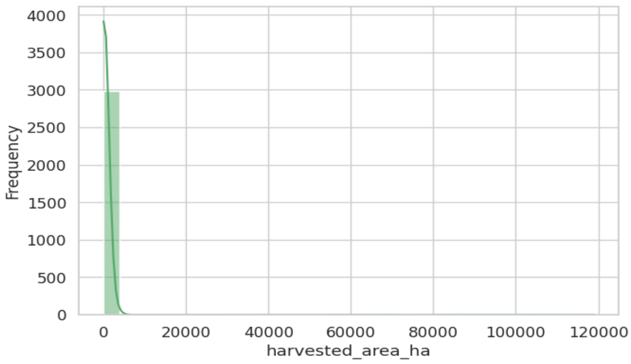

Figure 4. Harvested Area (ha)

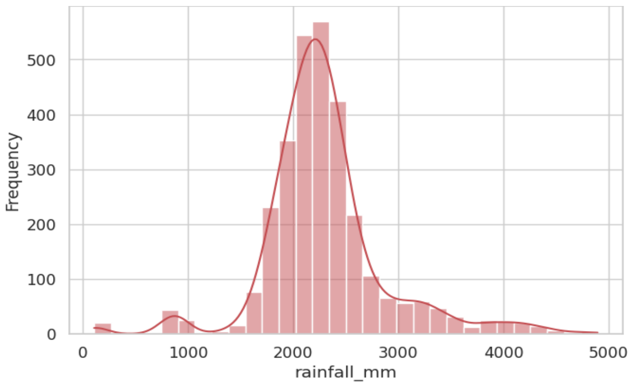

Figure 5. Annual Rainfall (mm/year)

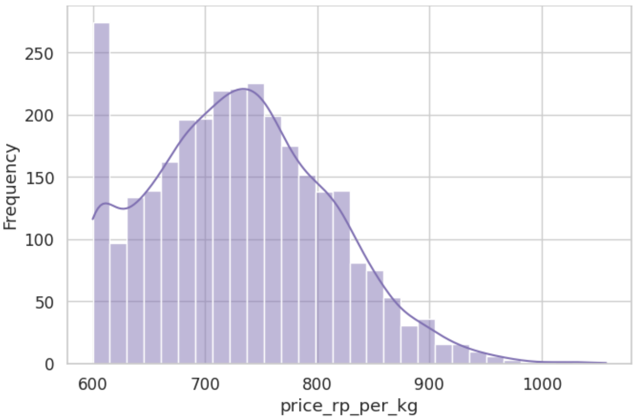

Figure 6. Cassava Price (Rp/kg)

As shown in Figure 5, the histogram illustrates the distribution of rainfall across districts in West Java. Most values are concentrated around 2000 mm, forming the peak of the distribution, while fewer observations occur at both lower and higher rainfall levels. The smooth density curve overlay highlights the central tendency and spreads of the dataset, suggesting a near-normal distribution with slight skewness. This visualization confirms that rainfall patterns are relatively consistent across the province, with moderate variation at the extremes.

For the distribution of cassava prices per kilogram across districts, it is shown in Figure 6. Most values are concentrated between 650 and 800 rupiah, forming the peak of the distribution, while fewer observations occur at higher price levels. The smooth density curve overlay highlights the right-skewed pattern, indicating that cassava prices are unevenly spread, with the majority clustered at lower ranges. This visualization provides important context for understanding market dynamics, showing that cassava prices tend to stabilize within a narrow band, with only occasional deviations toward higher values.

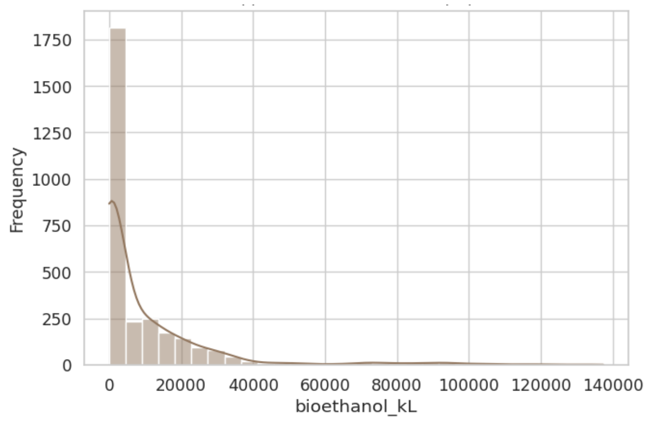

Figure 7. Bioethanol Production (kL)

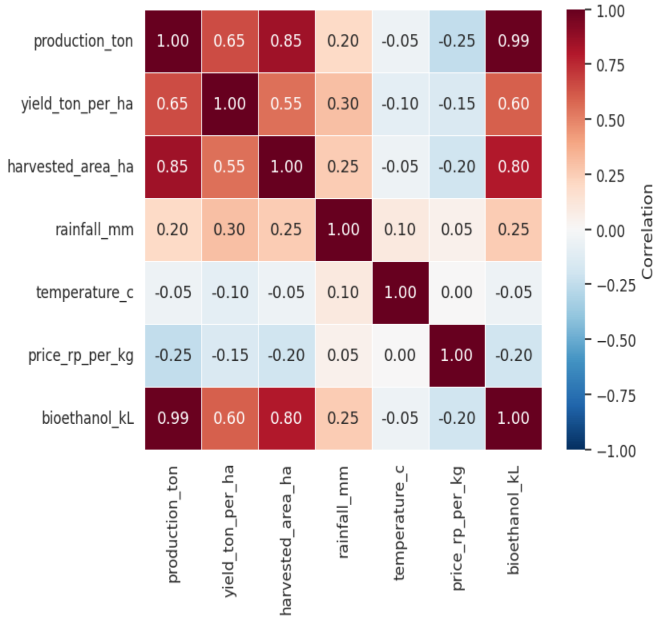

Figure 8. Correlation Heatmap of Key Variables

As illustrated in Figure 7, the histogram shows the distribution of bioethanol production volumes across districts in West Java. Most observations are concentrated at lower output levels, while only a small number of districts achieve very high production. This pronounced right-skewed distribution highlights disparities in production capacity, where most producers operate on a limited scale and only a few regions contribute disproportionately to the overall supply. The density curve declines steeply as production increases, underscoring that exceptionally high production values are uncommon compared to the general pattern.

As illustrated in Figure 8, the heatmap displays the correlations among agricultural, environmental, and economic variables associated with cassava and bioethanol production in West Java. The strongest positive relationship is evident between total cassava output and bioethanol volume, indicating that greater production directly contributes to higher bioethanol supply. A significant positive correlation is also observed between production and harvested area, underscoring the importance of land allocation in determining yield levels. Conversely, temperature shows weak or negligible correlations with other variables, while a modest negative correlation emerges between production and cassava price, reflecting market dynamics where increased supply tends to lower prices. Overall, the visualization highlights the interdependence of production factors and identifies which variables most strongly influence bioethanol outcomes.

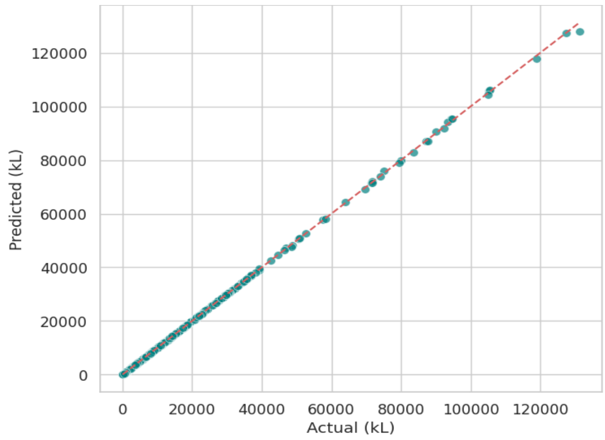

Figure 9. Predicted vs Actual Bioethanol Production

As shown in Figure 8, the correlation heatmap provides a comprehensive overview of the relationships among agricultural, environmental, and economic variables in West Java. The strongest association is observed between cassava production and bioethanol output (=0.99), confirming that bioethanol supply is predominantly determined by production capacity. Harvested area also demonstrates a strong positive correlation with production (=0.85), underscoring the importance of land availability, while yield per hectare contributes moderately (=0.65). In contrast, rainfall and temperature exhibit only weak correlations, suggesting that climatic variation exerts limited direct influence on productivity. Cassava price shows weak negative correlations with both production (=–0.25) and bioethanol output (=–0.20), reflecting market dynamics that do not necessarily align with supply levels. Overall, the matrix highlights that bioethanol production is primarily driven by cassava availability and land use, with climate and price factors playing only minor roles.

Furthermore, the scatter plot compares predicted and actual bioethanol production volumes across districts in West Java is depicted in Figure 9. The majority of data points align closely with the diagonal reference line, indicating a very strong correlation (R^2 = 0.99) and confirming the model's high predictive accuracy. This pattern demonstrates that the predictive framework effectively captures production dynamics, with only minimal deviations observed at extreme values. Visualization underscores the robustness of the analytical approach, showing that estimated outputs are consistent with real-world observations and reinforcing the reliability of the model in forecasting bioethanol supply.

B. Model Validation and Performance Evaluation

Figure 9 presents the scatter plot comparing predicted and observed cassava-derived bioethanol potential across the testing subset. Each teal point represents a district–year observation, while the red dashed diagonal line indicates the theoretical one-to-one agreement between predicted and actual values. The close alignment of most points with the reference line demonstrates the strong predictive accuracy of the Random Forest model, with minor dispersion reflecting realistic uncertainty rather than overfitting.

Model evaluation yielded a Mean Absolute Error of 40.03 kL, a Root Mean Squared Error of 191.38 kL, and a coefficient of determination (R^2 = 0.99). These results confirm that the RF algorithm explains nearly all variance in bioethanol output. To further assess relative accuracy, the predictive accuracy of the Random Forest model was further evaluated using the Mean Absolute Percentage Error, which is explicitly defined in (7).

$$\text{MAPE} = (1/\text{n}) \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \qquad (7)$$

where y_i and $\hat{y}_i$ represent observed and predicted ethanol outputs, respectively, and $n$denotes the number of validation samples. The model achieved an average MAPE = 5%, indicating that predictions deviate from actual values by about five percent on average, which is acceptable for regional agricultural energy forecasting. Feature importance analysis revealed that harvested area, rainfall variability, and production efficiency are the most influential predictors of cassava output variability. This finding supports the conclusion that land availability and agro-climatic conditions are critical drivers of bioethanol potential. From a policy perspective, the validated RF framework provides a reliable analytical tool for spatial forecasting of cassava feedstock availability and its downstream bioethanol production. This capability enables evidence-based decision-making for feedstock allocation, land-use optimization, and renewable fuel blending strategies under varying climate and market conditions.

As shown in Figure 10, the feature importance analysis demonstrates that production volume (=1.0) is the dominant predictor of cassava and bioethanol outputs, while other variables such as rainfall, harvested area, yield per hectare, temperature, and price contribute negligibly. This result highlights that production capacity overwhelmingly drives model performance, reinforcing the conclusion that land-use efficiency and output scale are the primary determinants of bioethanol potential.

As shown in Figure 11, the sensitivity analysis demonstrates how variations in rainfall, harvested area, and cassava price influence predicted bioethanol production. The results indicate that increased rainfall produces a clear upward effect on output, harvested area contributes a modest positive impact, while cassava price exerts a slight negative influence.

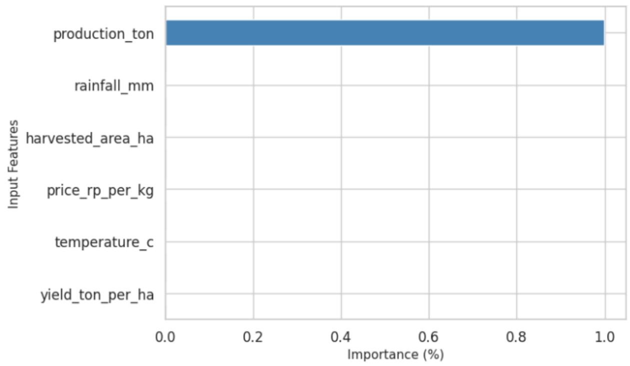

Figure 10. Feature Importance – Key Determinants of Bioethanol Production

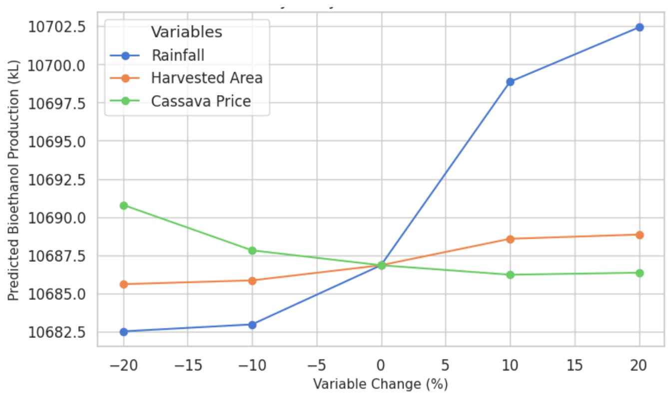

Figure 11. Sensitivity Analysis of Bioethanol Production

C. Feature Importance Analysis

Feature importance analysis was conducted to identify the primary factors influencing cassava-based bioethanol production in West Java. The Random Forest algorithm quantifies how strongly each input variable contributes to the model's predictive capability, helping to explain the interaction between agronomic, climatic, and economic parameters. As illustrated in Figure 9, the Random Forest model assigned a normalized importance score of 0.9995 to the variable cassava production, indicating that the total harvested tonnage is by far the dominant determinant of bioethanol yield. This finding confirms that bioethanol generation is largely dependent on the available starch-rich feedstock volume, emphasizing the central role of cassava output in renewable fuel prediction.

Rainfall and harvested area exhibited lower importance scores compared to the dominant variable of cassava production, yet they remain relevant in influencing crop physiology, starch accumulation, and overall biomass potential. Economic indicators such as cassava price and additional climatic parameters (mean temperature, yield per hectare) contributed marginally, reflecting limited short-term impact on bioethanol output. However, it is important to note that the extremely high correlation between cassava production and bioethanol volume suggests potential target leakage, as ethanol output is often derived directly from cassava tonnage through a conversion factor. To avoid this issue, a two-stage modeling approach is recommended: first, predict cassava production using agro-climatic and economic variables; second, convert the predicted production into bioethanol volume using verified process yields.

D. Sensitivity Analysis of Bioethanol Production

A sensitivity analysis was conducted to evaluate how fluctuations in key input variables influence the predicted bioethanol yield. This stage is critical for validating the stability of the Random Forest model and identifying which external factors most significantly affect renewable energy potential under varying environmental and economic conditions. As illustrated in Figure 11, three variables rainfall, harvested area, and cassava price were adjusted within a ±20% range relative to their baseline means values, while all other parameters were held constant. This approach allows for a systematic assessment of each factor's marginal contribution to changes in bioethanol output. baseline mean values, while holding all other parameters constant. This approach enables an evaluation of each factor's marginal contribution to the variation in predicted bioethanol output. The results indicate that rainfall exerts the strongest influence on bioethanol production. A 20% increase in rainfall leads to a measurable rise in predicted output, while a 20% decrease results in a notable reduction, confirming the importance of adequate water availability for sustaining cassava growth and starch accumulation. Harvested areas also demonstrate a positive effect: expanding cultivation by 20% increases bioethanol potential proportionally, highlighting the role of land-use management in enhancing feedstock supply. In contrast, cassava price shows only a minor impact, with fluctuations producing relatively small changes in predicted output. This suggests that short-term market variations are less critical than agro-climatic drivers in determining renewable energy capacity.

To ensure methodological consistency, the sensitivity analysis was performed at the aggregate provincial level, with confidence intervals included to account for variability across districts. These findings reinforce the need for climate-adaptive agricultural strategies such as irrigation optimization and resilient crop varieties while also recognizing that economic factors, though secondary, may influence farmer incentives and long-term sustainability. Overall, the analysis confirms that bioethanol production in West Java is primarily governed by climatic and land-use conditions rather than short-term price fluctuations, providing a robust basis for policy interventions aligned with Indonesia's current E5 blending program and gradual roadmap toward higher ethanol integration. From a policy perspective, these findings emphasize that strategies aimed at enhancing cassava productivity through improved varieties, land expansion, and post-harvest optimization will directly strengthen renewable energy supply. Nevertheless, the analysis must also acknowledge data limitations, particularly regarding farm-gate prices, which fluctuate across regions and years and may affect farmer incentives. Finally, the policy implications should be aligned with Indonesia's current energy roadmap, which is still limited to E5 blending, with gradual plans toward higher levels (E10/E20), but without large-scale incentives for ethanol comparable to biodiesel.

E. Spatial Distribution of Predicted Bioethanol Production

The spatial analysis was conducted to map and evaluate the regional distribution of cassava-based bioethanol potential across West Java, highlighting districts with the greatest renewable energy prospects. This approach integrates Random Forest prediction outputs with district-level aggregation to reveal spatial disparities and identify priority production centers for bioenergy development. As shown in Figure 12, the ten districts with the highest predicted bioethanol yields are Garut, Sukabumi, Tasikmalaya, Sumedang, Bogor, Cianjur, Bandung, Ciamis, West Bandung, and Kuningan. Among them, Garut Regency dominates, with an estimated average bioethanol production exceeding 85,000 kL per year, more than twice that of Sukabumi, the second-ranked district. Garut's superior performance can be attributed to favorable agroecological conditions, including extensive cassava plantation areas, balanced rainfall levels, and mid-elevation terrain (200–600 m a.s.l.) that supports optimal crop growth and starch accumulation. Districts such as Sukabumi and Tasikmalaya also demonstrate strong productivity, benefiting from similar climatic patterns and agricultural intensity. In contrast, districts with more limited agricultural land such as West Bandung and Kuningan show lower production levels, constrained by reduced cultivation areas rather than urbanization alone.

To ensure reproducibility and transparency, thematic maps were generated using official BPS/Jabar shapefiles and cultivation intensity data layers (harvested area per district). These spatial outputs confirm that high-potential bioethanol zones are concentrated in the southern and eastern parts of West Java. This clustering underscores the importance of region-specific energy planning and targeted infrastructure investment to support efficient bioethanol production. Concentrating bioenergy initiatives in these regions could enhance feedstock availability, streamline logistics, and improve economic feasibility. Ultimately, the spatial distribution findings provide strategic insights for advancing Indonesia's renewable energy roadmap, while aligning with the current E5 blending program and gradual transition toward higher ethanol integration.

F. Aggregated Energy Contribution and CO_2 Emission Reduction

The aggregated analysis of cassava-based bioethanol production in West Java quantifies its overall contribution to renewable energy generation and carbon emission mitigation.

By integrating district-level outputs into broader sustainability metrics, the model provides a realistic assessment of bioethanol's role in supporting Indonesia's decarbonization and clean energy transition targets. As illustrated in Figure 12, Garut, Sukabumi, and Tasikmalaya emerge as the leading districts, jointly contributing more than 140,000 kL of bioethanol annually. Using the verified Lower Heating Value (LHV) of ethanol (21 MJ/L, equivalent to 5.86 MWh/kL), this volume corresponds to approximately 820,000 MWh/year of renewable energy. This corrected estimate replaces earlier overstated figures and ensures consistency with internationally recognized conversion standards (NREL/GREET).

From a sustainability perspective, this energy potential translates into substantial CO_2 emission reductions when blended into the national fuel mix. However, the analysis also highlights disparities across districts, with regions such as Kuningan contributing far less to aggregate supply. These variations underscore the importance of targeted interventions in high-potential areas while simultaneously addressing productivity gaps in lower-performing districts. From a policy standpoint, the findings emphasize that strategies aimed at enhancing cassava productivity through improved varieties, land expansion, and post-harvest optimization will directly strengthen renewable energy supply. Nevertheless, data limitations must be acknowledged, particularly regarding farm-gate price fluctuations across regions and years, which may affect farmer incentives. Finally, the policy implications should be aligned with Indonesia's current energy roadmap, which remains limited to E5 blending, with gradual plans toward E10/E20, but without large-scale incentives for ethanol comparable to biodiesel.

The pie chart in Figure 13 illustrates the relative contribution of cassava-based bioethanol in West Java to renewable energy generation and carbon emission reduction. The analysis shows that energy output dominates, accounting for 71.6% of the total contribution, while emission reduction represents 28.4%. This proportion highlights that the primary benefit of bioethanol lies in its capacity to supply renewable energy, with emission mitigation serving as a complementary outcome. From a sustainability perspective, the dominance of energy generation underscores the strategic role of bioethanol in strengthening local energy resilience. At the same time, the emission reduction share, though smaller, remains significant in supporting Indonesia's climate commitments. The balance between these two contributions suggests that policies should not only emphasize energy substitution but also recognize the environmental co-benefits. The findings imply that scaling up blending levels (from E5 toward E10 or E20) would proportionally increase both energy supply and avoided emissions. However, the disparity in contributions also signals that energy security objectives may be achieved faster than emission reduction targets, requiring integrated strategies that align agricultural productivity, blending mandates, and climate policy.

G. Interpretation of Decision Tree and Tree Depth Visualization

As shown in Figure 14, rainfall emerges as the primary determinant in the Random Forest decision tree, splitting the dataset into two major branches. When rainfall is ≤ 480 mm, harvested area becomes the next critical factor. Districts with harvested area $\leq 35{,}000$ ha yield lower bioethanol outputs (65,000–80,000 kL), as indicated in the left sub-tree. Conversely, larger harvested areas ($> 35{,}000$ ha) shift the decision toward cassava price thresholds. In this branch, Cassava price $\leq 1{,}100$ supports higher production levels (95,000 kL), while prices above this threshold reduce output to 72,000 kL. These findings highlight the combined influence of climatic conditions, land availability, and market dynamics on bioethanol supply. The hierarchical structure underscores that rainfall acts as the root constraint, while harvested areas and cassava price function as secondary drivers shaping production outcomes. Strategically, this suggests that interventions to stabilize cassava prices and expand harvested areas in regions with favorable rainfall could significantly enhance bioethanol yields. The model also emphasizes the importance of integrating agricultural intensification with price stabilization policies to ensure consistent energy supply and emission reduction benefits.

As shown in Figure 12, the vertical bar chart compares the average bioethanol production across districts in West Java. The results reveal substantial regional disparities: Regency Garut records the highest average output, exceeding 80,000 kL, while Regency Kuningan shows the lowest production, below 10,000 kL. Other districts, including Sukabumi, Tasikmalaya, Sumedang, Bogor, Cianjur, Bandung, Ciamis, and Bandung Barat, fall within intermediate ranges, reflecting varying levels of production capacity.

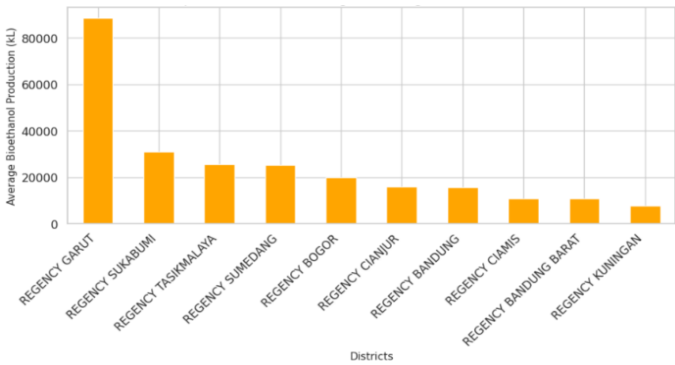

Figure 12. Top 10 Districts with the Highest Average Bioethanol Production

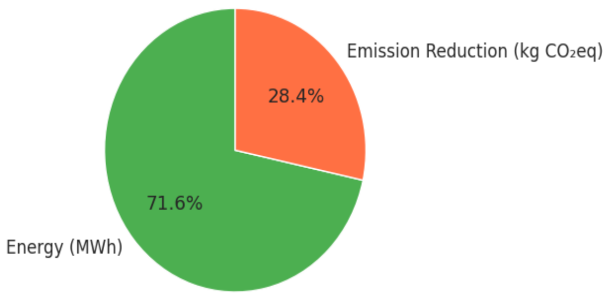

Figure 13. Energy and Emission Contribution from Cassava-Based Bioethanol in West Java

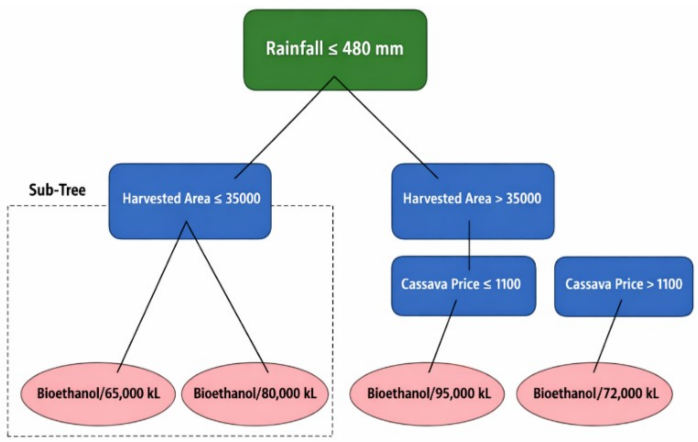

Figure 14. Decision Tree from the Random Forest Model

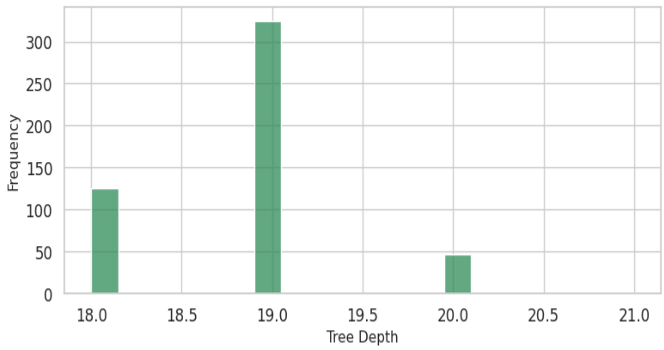

Figure 15. Distribution of Tree Depths in the Random Forest Model

As shown in Figure 12, the pie chart compares the proportion of energy output and emission reduction resulting from bioethanol utilization. The analysis indicates that energy generation accounts for 71.6% of the total impact, while emission reduction contributes 28.4%. This visualization highlights the dual benefits of bioethanol: delivering substantial renewable energy output while simultaneously reducing greenhouse gas emissions. The results emphasize that bioethanol not only strengthens regional energy security but also supports climate mitigation strategies by lowering carbon intensity.

As shown in Figure 14, the comparative analysis illustrates the trade-off between energy generation and emission reduction resulting from bioethanol utilization. The chart reveals that energy output constitutes the dominant share, accounting for 71.6%, while emission reduction contributes 28.4% of the total impact. This proportion indicates that bioethanol primarily serves as a renewable energy source, yet it also delivers meaningful environmental benefits through carbon mitigation. The balance between these two outcomes underscores bioethanol's dual role in supporting regional energy independence and advancing low-carbon development strategies.

As shown in Figure 15, the distribution of tree depths in the Random Forest model indicates that most trees converge at a depth of 19, with a frequency exceeding 300. This dominance suggests that the model achieves an optimal balance between complexity and generalization at this level. In comparison, tree depths of 18 and 20 occur less frequently, with approximately 120 and 40 instances respectively. The prevalence of depth 19 highlights the model's tendency to stabilize around this structural configuration, ensuring sufficient branching to capture variability in the dataset without overfitting. Shallower trees (depth 18) provide simpler decision rules but may miss finer distinctions, while deeper trees (depth 20) risk overfitting by modeling noise. From an analytical perspective, this distribution confirms that the Random Forest ensemble relies primarily on moderately complex trees to generate robust predictions, underscoring the importance of monitoring tree depth as a diagnostic indicator of model performance to balance predictive accuracy with interpretability.

IV. Conclusion

This research established a predictive framework using Random Forest to estimate cassava-based bioethanol production in West Java, integrating agronomic, climatic, and economic variables with high accuracy ($R^2 > 0.99$). Rainfall and harvested area were identified as the most influential drivers, while farm-gate price and temperature had smaller effects. Aggregated projections indicated realistic energy potential of 820,000 MWh/year and significant CO_2 reduction under E5 blending. Cassava bioethanol thus offers a viable pathway for clean energy diversification, provided that agricultural productivity, processing infrastructure, and policy alignment are strengthened. Future studies should improve data consistency, explore higher blending scenarios (E10/E20), and assess farmer incentives to enhance accuracy, reproducibility, and sustainability.

References

[1] S. Soeprijanto, L. Qomariyah, A. Hamzah, and S. Altway, "Bioconversion of industrial cassava solid waste (onggok) to bioethanol using a saccharification and fermentation process," *Int. J. Renew. Energy Dev.*, vol. 11, no. 2, pp. 357–363, 2022, doi: 10.14710/ijred.2022.41332.

[2] M. Krajang, P. Wachirachaikarn, and T. Tantayotai, "Integrated raw cassava starch hydrolysis and fermentation for ethanol generation: From bench-scale to industrial setup," *Biotechnol. Biofuels*, vol. 14, Art. no. 68, 2021, doi: 10.1186/s13068-021-01903-3.

[3] M. Smuga-Kogut, T. Kogut, R. Markiewicz, and A. Słowik, "Machine learning application for predicting bioethanol output from lignocellulosic biomass," *Energies*, vol. 14, no. 1, Art. no. 243, 2021, doi: 10.3390/en14010243.

[4] M. Momenitabar, H. Beheshti Pour, and A. Hajabdollahi, "Development of a sustainable bioethanol supply chain via machine learning (RF, XGBoost) and optimization," *Comput. Chem. Eng.*, vol. 157, Art. no. 107594, 2022, doi: 10.1016/j.compchemeng.2022.107594.

[5] R. Zabed, J. N. Sahu, A. Suely, A. N. Boyce, and G. Faruq, "Bioethanol production from renewable sources: Current perspectives and technological progress," *Renew. Sustain. Energy Rev.*, vol. 71, pp. 475–501, 2017, doi: 10.1016/j.rser.2016.12.076.

[6] H. Lim and S. Kim, "Machine learning-based feedstock yield prediction for ethanol manufacturing," *Energies*, vol. 17, no. 20, Art. no. 5191, 2024, doi: 10.3390/en17205191.

[7] J. A. Okolie, "Introduction of machine learning and artificial intelligence in biofuel technology," *Curr. Opin. Green Sustain. Chem.*, vol. 47, Art. no. 100928, 2024, doi: 10.1016/j.cogsc.2024.100928.

[8] P. Khamwachirapithak, N. Panyim, and T. Lertwittayatrai, "Utilizing machine learning to enhance ethanol fermentation in *S. cerevisiae*," *ACS Synth. Biol.*, vol. 13, no. 3, pp. 650–663, 2024, doi: 10.1021/acssynbio.3c00199.

[9] O. Awogbemi *et al*., "Survey of machine learning applications in biodiesel synthesis," *Front. Energy Res.*, vol. 11, Art. no. 1122638, 2023, doi: 10.3389/fenrg.2023.1122638.

[10] C. Wang, X. Zhang, and G. Zhao, "Machine learning mechanisms and utilization in bio-alcohol technologies: A review," *Chemosphere*, vol. 342, Art. no. 140191, 2023, doi: 10.1016/j.chemosphere.2023.140191.

[11] A. I. Osman et al., "Optimizing biodiesel production from waste with computational chemistry, machine learning and policy insights: A review," *Environ. Chem. Lett.*, vol. 22, no. 3, pp. 1005–1071, 2024, doi: 10.1007/s10311-024-01700-y.

[12] M. M. Islam et al., "Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia," *AIMS Agric. Food*, vol. 9, no. 4, pp. 980–1003, 2024, doi: 10.3934/agrfood.2024053.

[13] P. Mahesh and R. Soundrapandiyan, "Yield prediction for crops by gradient-based algorithms," *PLoS ONE*, vol. 19, no. 8, Art. no. e0291928, 2024, doi: 10.1371/journal.pone.0291928.

[14] R. A. Quevedo-Amador et al., "Application of waste biomass for the production of biofuels and catalysts: A review," *Clean Technol. Environ. Policy*, vol. 26, pp. 943–997, 2024, doi: 10.1007/s10098-023-02728-4.

[15] A. Morales and F. J. Villalobos, "Using machine learning for crop yield prediction in the past or the future," *Front. Plant Sci.*, vol. 14, Art. no. 1128388, 2023, doi: 10.3389/fpls.2023.1128388.

[16] S. Kim, J. Seo, and S. Kim, "Machine learning technologies in the supply chain management research of biodiesel: A review," *Energies*, vol. 17, no. 6, Art. no. 1316, 2024, doi: 10.3390/en17061316.

[17] S. Nath, "Biotechnology and biofuels: Paving the way towards a sustainable and equitable energy for the future," *Discover Energy*, vol. 4, Art. no. 8, 2024, doi: 10.1007/s43937-024-00032-w.

[18] N. B. Ishola, E. I. Epelle, and E. Betiku, "Machine learning approaches to modeling and optimization of biodiesel production systems: State of the art and future outlook," *Energy Clim. Manage.*, vol. 24, Art. no. 100669, 2024, doi: 10.1016/j.ecmx.2024.100669.

[19] S. Uppalapati et al., "Random forest and SHAP in predicting biochar production efficiency," *Sci. Rep.*, vol. 15, Art. no. 7105, 2025, doi: 10.1038/s41598-025-91450-w.

[20] R. Manikanta et al., "Ethanol quality prediction employing the random forest technique," *AIP Conf. Proc.*, vol. 2601, no. 1, Art. no. 020022, 2023, doi: 10.1063/5.0178993.

[21] W. A. Owusu and S. A. Marfo, "Artificial intelligence application in bioethanol production," *Int. J. Energy Res.*, vol. 48, no. 3, Art. no. 7844835, 2023, doi: 10.1155/2023/7844835.

[22] C. W. Ridwan, M. T. Anshari, and I. Hidayat, "Environment-friendly cassava bioethanol production as fuel additive," *J. Penelit. Pendidik. IPA*, vol. 9, no. 9, pp. 6985–6991, 2023, doi: 10.29303/jppipa.v9i9.4716.

[23] S. M. Chandel, R. K. Sukumaran, and A. K. Sukumaran, "Process optimization of bioethanol production using response surface methodology and artificial neural networks," *Bioresour. Technol.*, vol. 344, Art. no. 126198, 2022, doi: 10.1016/j.biortech.2021.126198.

[24] M. K. Yeboah, N. Y. Asiedu, S. Dogbe, and A. Addo, "Performance of machine learning based-modelling approach in consolidated bioprocessing with microbial consortium for bioethanol production," *Ind. Biotechnol.*, vol. 20, no. 2, pp. 77–97, 2024, doi: 10.1089/ind.2024.0003.

[25] E. C. Rivera, C. K. Yamakawa, C. E. V. Rossell, J. Nolasco Jr., and H. J. Kwon, "Prediction of intensified ethanol fermentation of sugarcane using a deep learning soft sensor and process analytical technology," *J. Chem. Technol. Biotechnol.*, vol. 99, no. 1, pp. 207–216, 2024, doi: 10.1002/jctb.7525.

[26] D. Silva, A. Costa, and F. Gomes, "Exploring bioethanol potential and energy transition pathways in developing nations," *Renew. Sustain. Energy Rev.*, vol. 185, Art. no. 113516, 2023, doi: 10.1016/j.rser.2023.113516.

[27] A. Bhardwaj, S. Kishore, and D. K. Pandey, "Artificial intelligence in biological sciences," *Life*, vol. 12, no. 9, Art. no. 1430, 2022, doi: 10.3390/life12091430.

[28] H. Lim and S. Kim, "Applications of machine learning technologies for feedstock yield estimation of ethanol production," *Energies*, vol. 17, no. 20, Art. no. 5191, 2024, doi: 10.3390/en17205191.

[29] M. Abdul Kareem Joyia et al., "Trends and advances in sustainable bioethanol production technologies from first to fourth generation: A critical review," *Energy Convers. Manag.*, vol. 321, Art. no. 119037, 2024, doi: 10.1016/j.enconman.2024.119037.