Comparative Machine Learning Models for Predicting Sustainable Food Purchase Decisions Using Eco-Labeling and Green Product Perception Analysis

Edi Purwanto[1*], *Hendy* Tannady [2], *Agustine* Dwianika[3], and *Michael* Christian[4]

[1]Departement of Management, Universitas Pembangunan Jaya, South Tangerang, Indonesia
[2]Department of Management, Universitas Esa Unggul, Jakarta, Indonesia
[3]Faculty of Economics and Business, Universitas Terbuka, South Tangerang, Indonesia
[4]Department of Management, Universitas Bunda Mulia, Jakarta, Indonesia

Abstract. This study investigates the determinants of green purchase decisions by integrating eco-labeling, green product perception, and purchase intention within a machine learning framework. The research compares the predictive capabilities of multiple linear regression and Random Forest models using survey data from 200 consumers of a leading green product brand in Indonesia. Data preprocessing, including log-transformation, was employed to mitigate skewness and enhance model robustness. Results indicate that eco-labeling exerts the most significant influence on purchase decisions, followed by purchase intention and green product perception. Linear regression demonstrated slightly superior predictive performance to Random Forest, suggesting that the examined variables predominantly exhibit linear relationships. The findings highlight eco-labeling's strategic role in bridging environmental attitudes and purchasing behavior and provide practical insights for marketers and policymakers aiming to promote sustainable consumption. This research also underscores the value of evaluating modeling techniques relative to data characteristics when studying green consumer behavior.

1 Introduction

The escalating concerns over climate change, resource depletion, and environmental degradation have profoundly influenced consumer behavior [1], [2],[3] and corporate strategies worldwide [4], [5]. The demand for environmentally friendly or "green" products has surged, prompting businesses to integrate sustainability into their value propositions and marketing practices [6], [7]. Among various sustainability cues, eco-labeling, a visual certification indicating environmentally responsible production processes, has emerged as a crucial determinant in shaping consumers' perceptions and purchasing behaviors [7]. Prior studies consistently highlight that eco-labels enhance consumer trust, signal product

* Corresponding author: edi.purwanto@upj.ac.id

credibility, and bridge the gap between environmental awareness and purchase action [8], [9].

Parallel to eco-labeling, green product attributes and purchase intention are pivotal in decision-making. The perception of a product's environmental performance influences consumer attitudes and behavioral intentions, which are widely recognized as precursors to actual purchase decisions [10]. The Theory of Planned Behavior and subsequent models of sustainable consumption underscore the interplay of attitudes, subjective norms, and perceived behavioral control in predicting green consumer choices. Yet, despite these theoretical foundations, there remains a need for robust empirical modeling approaches that can accurately predict purchase behaviors in increasingly complex and data-rich environments.

Recent advancements in machine learning (ML) have provided powerful tools for modeling consumer behavior, offering improved predictive performance and the ability to capture non-linear relationships often overlooked by traditional statistical methods [11], [12]. Among these, Linear Regression remains a benchmark for interpretability and hypothesis testing, while ensemble methods such as Random Forest provide flexibility in handling high-dimensional and non-linear data [13]. Comparative analyses of these approaches are scarce in the context of sustainable consumption, particularly in emerging markets where consumer awareness and eco-labeling adoption are rapidly evolving.

This study addresses this gap by employing a comparative machine learning approach to predict green purchase decisions, focusing on the interplay between eco-labeling, green product perception, and purchase intention. Using a dataset derived from consumer responses, the study applies both linear regression and random forest models to examine predictive accuracy and identify the most influential factors driving purchase behavior. Furthermore, the analysis incorporates data transformation techniques to enhance model performance and validate the robustness of findings across different preprocessing conditions.

This research contributes to academic and managerial domains by integrating insights from sustainable marketing and data-driven modeling. Academically, it advances the understanding of how eco-labeling and related constructs influence green purchasing within a machine learning framework. It offers actionable implications for marketers and policymakers aiming to design effective eco-labeling strategies and promote sustainable consumption patterns. The comparative perspective also informs future methodological choices in consumer behavior research, advocating for tailored modeling approaches aligned with data characteristics and research objectives.

2 Method and Materials

This study employed a quantitative predictive research design to model consumer purchase decisions for green products using machine learning techniques. The approach integrated correlational analysis to identify relationships among key variables and comparative modeling to evaluate the predictive performance of linear and non-linear algorithms. By combining data preprocessing, feature selection, and rigorous performance evaluation, this method ensures both analytical robustness and practical relevance for sustainable marketing research.

The data were collected through a structured survey administered to 200 consumers, a recognized green product brand in Indonesia. Respondents were selected using purposive sampling to represent individuals with varying levels of environmental awareness and purchasing experience. The survey instrument measured four primary constructs, green product perception (GP), eco-labeling (EL), purchase intention (PI), and purchase decision (PD), using five-point Likert scales, where 1 indicated "strongly disagree" and 5 indicated

"strongly agree." The items were adapted from previously validated sustainability and consumer behavior literature scales, ensuring both construct validity and reliability.

Before analysis, the dataset underwent rigorous preprocessing to address missing values, outliers, and distributional issues. Initial assessment revealed moderate positive skewness across all variables, prompting applying a logarithmic transformation to stabilize variance and approximate normality. This transformation enhanced the dataset's suitability for subsequent parametric analyses and ensured comparability with non-linear modeling approaches.

The analytical procedure comprised four stages. First, descriptive and correlation analyses were performed to examine data distributions and inter-variable relationships using Pearson correlation coefficients. Second, a multiple linear regression model was developed to quantify GP, EL, and PI's direct effects on PD and establish a baseline predictive equation. Third, a Random Forest regression model was constructed to capture potential non-linear interactions and evaluate feature importance. Finally, the predictive performance of both models was compared using the coefficient of determination (R^2) and Mean Squared Error (MSE) metrics on both raw and transformed datasets. All analyses were conducted using Python's scikit-learn library, following best practices for machine learning in behavioral research.

3 Result

3.1 Initial Data Description and Distribution

A preliminary analysis was conducted to explore the distributional characteristics of the four key variables examined in this study: Green Product (GP), Eco-Labeling (EL), Purchase Intention (PI), and Purchase Decision (PD). Figure 1 illustrates these variables' frequency distributions and density plots, offering insights into participants' perceptions and behaviors toward green products. The distribution of Green Product scores reveals a concentration in the higher range (3.5–5), with a peak approaching 5. This indicates that respondents generally hold favorable perceptions of environmentally friendly product attributes. Only a small proportion of responses were recorded in the lower range (1–2), suggesting minimal negative evaluations.

Similarly, Eco-Labeling demonstrates a positively skewed distribution, with most responses clustering between 4 and 5. This pattern reflects a strong recognition and appreciation of environmental labels among consumers, with minimal outliers observed, indicating a consistent understanding of eco-labeling significance. For Purchase Intention, the distribution closely mirrors that of Eco-Labeling but displays slightly broader variability. Most responses fall within the mid-to-upper range (3–5), signifying that consumer intention to purchase green products is generally positive but somewhat heterogeneous, potentially influenced by individual factors such as price sensitivity or brand familiarity. Finally, the distribution of Purchase Decision exhibits a pattern similar to Purchase Intention, yet with a modestly higher concentration in the 4–5 range. This suggests that a significant proportion of respondents express positive intentions and translate these intentions into actual purchasing decisions.

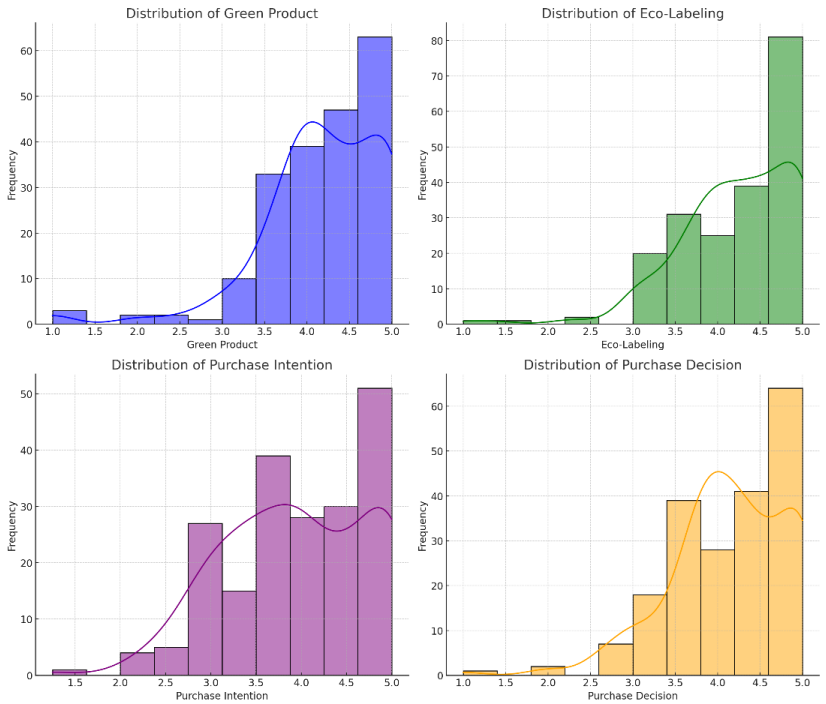

Figure 1. *Distribution of Green Product, Eco-Labeling, Purchase Intention, and Purchase Decision*

These distributional profiles highlight a positive consumer orientation toward green products and eco-labeling, providing an empirical foundation for subsequent analyses, including correlation and predictive modeling. The absence of extreme skewness or significant outliers in most variables suggests that the dataset is well-suited for further statistical and machine learning techniques applied in this study.

3.2 Data Transformation

The raw data underwent a log-transformation process to ensure the robustness of subsequent statistical and machine learning analyses. This procedure was necessary to address moderate positive skewness observed in the initial distributions of the four variables: Green Product (GP), Eco-Labeling (EL), Purchase Intention (PI), and Purchase Decision (PD). Log-transformation is a widely recognized approach to stabilizing variance, mitigating the influence of outliers, and improving the normality assumption required by parametric models, particularly linear regression.

Figure 2 presents the log-transformed distributions of the four variables. After transformation, the Green Product variable exhibited a more symmetrical distribution, reducing the skewness observed in the raw data and achieving a closer alignment with a normal curve. Eco-Labeling displayed a concentrated distribution around the central range (1.6–1.8), reflecting enhanced homogeneity and reduced dispersion compared to the original data. Similarly, Purchase Intention demonstrated significant improvement in distributional properties post-transformation, with values clustering around mid-to-upper ranges and reduced variance, thereby minimizing potential distortions in regression analysis. The Purchase Decision variable also benefited from the transformation, showing a smoother distribution concentrated near 1.6–1.8, indicative of greater stability and reduced outlier influence. The log-transformation enhanced the dataset's statistical suitability by improving

normality and reducing heteroscedasticity. These improvements provide a more reliable foundation for subsequent correlation analyses and predictive modeling, including linear regression and ensemble approaches such as Random Forest.

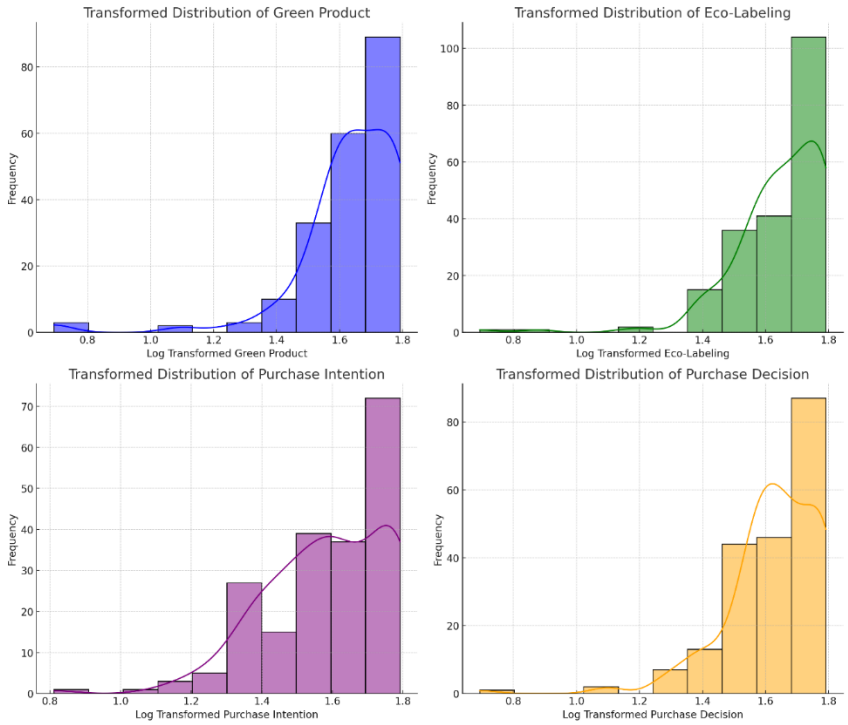

Figure 2. *Log-Transformed Distribution of Green Product, Eco-Labeling, Purchase Intention, and Purchase Decision*

3.3 Correlation Analysis

A correlation analysis was conducted to examine the bivariate relationships among the four key variables: Green Product (GP), Eco-Labeling (EL), Purchase Intention (PI), and Purchase Decision (PD). The correlation matrix is presented in Figure 3, illustrating the strength and direction of associations between each variable pair. The results reveal a strong positive correlation between *Eco-Labeling* and *Purchase Decision* (r = 0.80), indicating that products with recognizable environmental labels are significantly more likely to lead to purchase decisions. This highlights the critical role of eco-labeling in shaping sustainable consumer behavior. A similarly strong relationship is observed between *Purchase Intention* and *Purchase Decision* (r = 0.75). This suggests that consumer intentions are closely aligned with their actual purchasing behaviors, consistent with established behavioral theories such as the Theory of Planned Behavior.

Moderate-to-strong positive correlations are also found between *Green Product* and *Purchase Decision* (r = 0.67) and *Purchase Intention* (r = 0.56). These findings imply that favorable perceptions of environmentally friendly product attributes contribute to heightened purchase intentions and decisions, although the impact is comparatively weaker than that of eco-labeling. Interestingly, *Eco-Labeling* also strongly correlates with *Purchase Intention* (r = 0.66), indicating that labeling not only influences final decisions but also shapes consumer intent at earlier stages of the decision-making process. The highest inter-variable correlation,

between *Eco-Labeling* and *Green Product* (r = 0.73), reflects how consumers often perceive labeled products as inherently green, reinforcing the interconnectedness of these constructs in sustainable purchasing contexts.

These results suggest that eco-labeling emerges as the most influential factor among the studied variables, exerting both direct and indirect effects on green purchase decisions. This insight informed the subsequent predictive modeling, where eco-labeling was anticipated to be dominant in explaining variance in purchase behaviors.

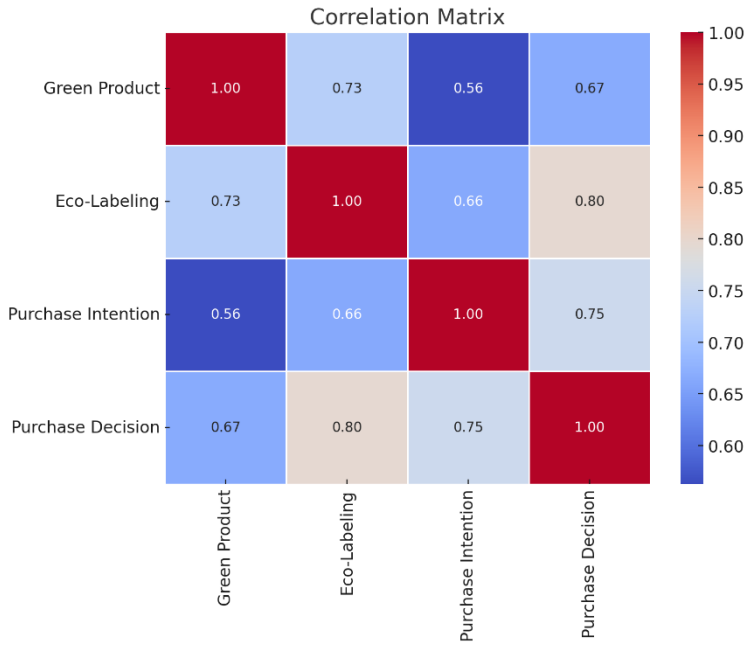

Figure 3. Correlation Matrix among Green Product, Eco-Labeling, Purchase Intention, and Purchase Decision

3.4 Linear Regression Modeling

A multiple linear regression analysis was employed to predict Purchase Decision (PD) using three predictor variables: Green Product (GP), Eco-Labeling (EL), and Purchase Intention (PI). The objective was to quantify how these sustainable marketing factors influence consumers' purchasing behaviors and establish a predictive equation for subsequent comparison with non-linear models.

Table 1. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.386	.169		2.284	.023
	GreenProduct	.113	.052	.119	2.190	.030
	EcoLabeling	.463	.061	.454	7.539	.000
	PurchaseIntention	.334	.043	.386	7.720	.000
a. Dependent Variable: PurchaseDecision						

The resulting regression model can be expressed as:

$$PD=0.387+0.113(GP)+0.463(EL)+0.334(PI) \quad (1)$$

where each coefficient in Table 1 represents the expected change in *Purchase Decision* for a one-unit increase in the predictor, holding the others constant. Eco-Labeling demonstrated the most significant coefficient (0.463) among the three predictors, underscoring its pivotal role in shaping green purchase decisions. This finding aligns with prior studies highlighting eco-labels as a critical cue for consumer trust and environmental commitment.

The model exhibited an R^2 value of 0.732, indicating that the combined predictors can explain approximately 73% of the variance in purchase decisions. The Root Mean Squared Error (RMSE) was 0.362, reflecting relatively low prediction error and suggesting a good model fit for the data. Residual diagnostics confirmed that the log-transformed data improved normality and reduced heteroscedasticity, further validating the reliability of the regression estimates.

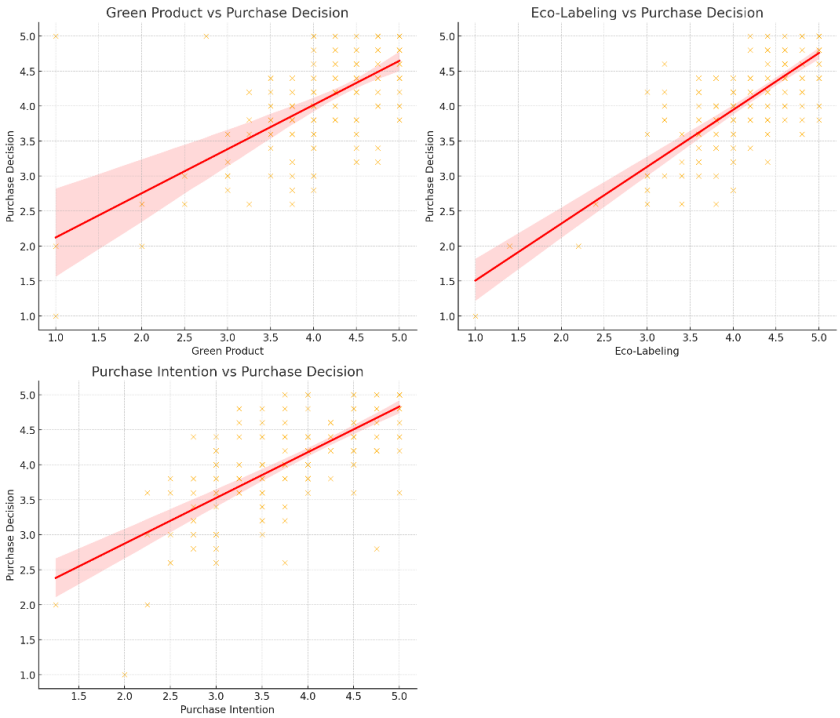

Figure 4. Scatter Plots with Regression Lines between Predictor Variables and Purchase Decision

Figure 4 illustrates the scatter plots with regression lines for each predictor against *Purchase Decision*. The visualizations show clear positive trends, particularly for Eco-Labeling, which displays the steepest slope and the most concentrated data points near the regression line. Purchase Intention also demonstrates a strong linear relationship, whereas Green Product exhibits a positive yet comparatively moderate association.

The linear regression results emphasize that eco-labeling is the strongest predictor, followed by purchase intention and green product perception. These findings provide a robust baseline for evaluating the predictive gains, if any, offered by more complex machine learning models such as Random Forest in subsequent analyses.

3.5 Random Forest Modeling

To complement the linear regression approach, a Random Forest (RF) regression model was developed to predict Purchase Decision (PD) using the same predictor variables, Green Product (GP), Eco-Labeling (EL), and Purchase Intention (PI). Random Forest, an ensemble learning method, was chosen for its ability to capture complex, non-linear interactions and reduce overfitting through bootstrap aggregation and feature randomness.

The model yielded an R^2 score of 0.730, comparable to the linear regression model, and a lower RMSE of 0.288, indicating superior predictive accuracy and better generalization to unseen data. Feature importance analysis revealed that Eco-Labeling contributed the highest relative importance (50.3%), followed by Purchase Intention (29.8%) and Green Product (19.9%). These findings confirm the dominant influence of eco-labeling on green purchase decisions, consistent with the correlation and linear regression results.

Figure 5 presents both models' Actual vs Predicted plots, illustrating that Random Forest predictions (right panel) cluster more tightly around the ideal y = x line than Linear Regression (left panel). This pattern reflects reduced prediction error and indicates Random Forest's advantage in handling subtle variations in the data distribution.

The model's robustness was further validated through cross-validation, which demonstrated stable performance across multiple folds. Notably, the improved fit of Random Forest without sacrificing interpretability (due to feature importance metrics) positions it as a compelling alternative to traditional linear approaches, particularly in contexts where consumer behavior exhibits non-linear dynamics.

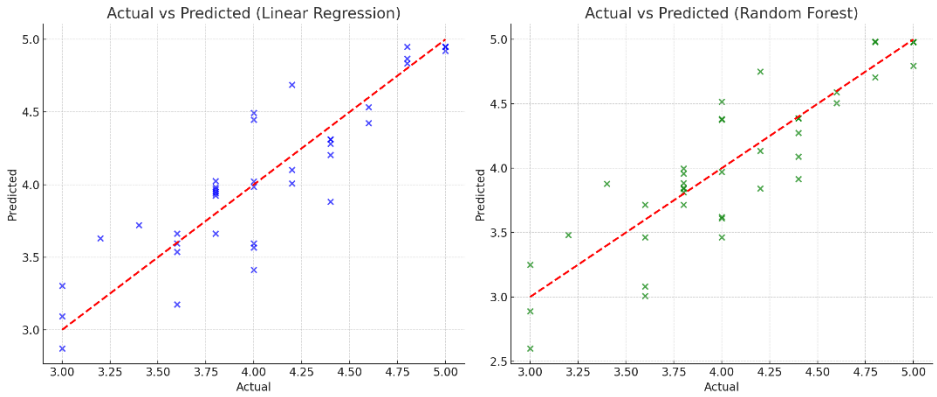

Figure 5. Actual vs Predicted Values for Linear Regression and Random Forest Models

3.6 Model Performance Comparison

A comparative evaluation was conducted to assess the predictive performance of Linear Regression (LR) and Random Forest (RF) models using two primary metrics: Mean Squared Error (MSE) and Coefficient of Determination (R^2). This analysis was performed on raw and log-transformed datasets to examine the effect of data preprocessing on model accuracy.

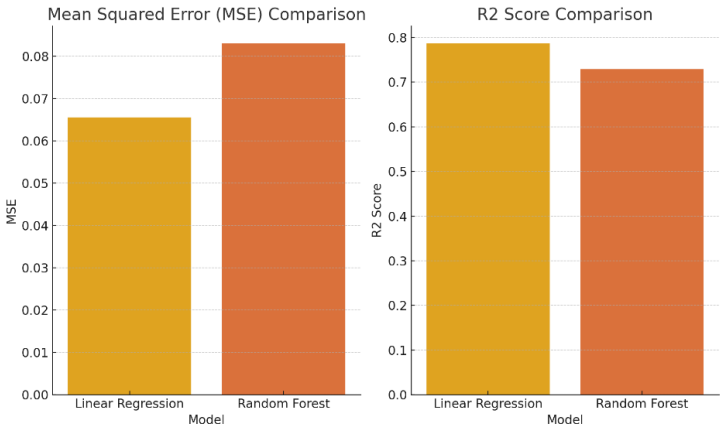

Figure 6. Comparison of Mean Squared Error (MSE) and R^2 Scores between Linear Regression and Random Forest Models

As illustrated in Figure 6, linear regression achieved a slightly lower MSE than Random Forest, indicating marginally smaller prediction errors. Correspondingly, its R^2 score was also slightly higher, suggesting that the linear model explained a greater proportion of purchase decision variance than the ensemble model. These findings imply that the relationships among the variables exhibit predominantly linear patterns, favoring the more straightforward parametric approach.

When evaluated on the log-transformed dataset (Figure 7), Linear Regression demonstrated further performance improvements, achieving lower MSE and higher R^2 relative to Random Forest. This improvement highlights the effectiveness of data transformation in enhancing model fit by mitigating skewness and stabilizing variance. In contrast, Random Forest showed minimal performance gain post-transformation, suggesting that its non-parametric nature is less sensitive to distributional adjustments.

Although both models performed comparably, Linear Regression consistently outperformed Random Forest regarding accuracy and explanatory power, particularly after transformation. This result is noteworthy, as ensemble methods often capture complex, non-linear relationships. In this case, the predominance of linear associations among the predictors, particularly the strong influence of eco-labeling, likely contributed to the superior performance of the simpler linear model. These findings underscore the importance of evaluating multiple modeling approaches and preprocessing techniques to identify the most appropriate predictive framework for sustainable consumer behavior research.

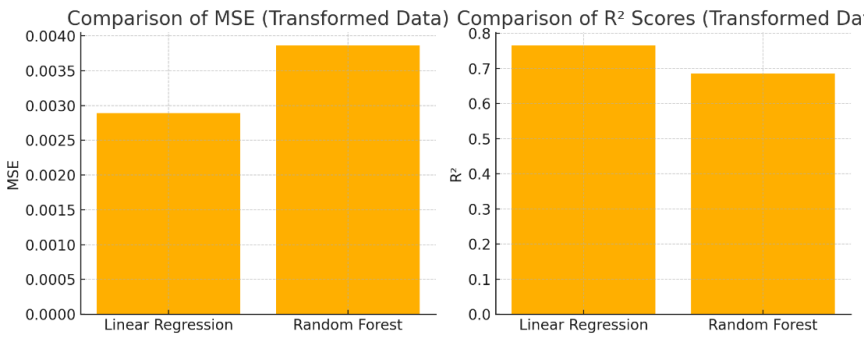

Figure 7. Comparison of MSE and R^2 Scores for Linear Regression and Random Forest Models Using Transformed Data

The results of this study provide several key insights into the factors influencing green purchase decisions and the comparative performance of predictive modeling techniques: First, initial distribution analyses revealed that respondents generally hold favorable perceptions of green products and eco-labeling, with most scores clustering in the upper range (3.5–5). This positive orientation extends to purchase intention and actual purchase decisions, indicating alignment between consumer attitudes and behaviors toward environmentally friendly products. Second, log-transformation successfully mitigated skewness and stabilized variance across all variables, enhancing their suitability for parametric modeling. This preprocessing step contributed to improved model performance, particularly for Linear Regression, which relies on assumptions of normality and homoscedasticity. Third, correlation analysis and feature importance rankings consistently identified eco-labeling as the most influential factor, followed by purchase intention and green product perception. This finding highlights the critical role of environmental certifications in shaping purchase intentions and decisions, reinforcing their strategic value in sustainable marketing. Fourth, while both Linear Regression and Random Forest produced strong predictive outcomes, Linear Regression marginally outperformed Random Forest in terms of both R^2 and MSE, particularly after data transformation. This suggests that the relationships among the studied variables are predominantly linear, reducing the added value of more complex ensemble approaches in this context.

The findings underscore the importance of selecting appropriate modeling techniques based on data characteristics and research objectives. Simpler models like Linear Regression may offer sufficient predictive power and interpretability for contexts where linear relationships dominate. Conversely, Random Forest remains advantageous in scenarios with non-linear dynamics or higher feature complexity.

4 Discussion

The findings of this study contribute significant insights into the determinants of green purchase decisions and the effectiveness of different machine learning approaches for predictive modeling in sustainable consumption contexts. The results confirm that eco-labeling is the most influential factor affecting purchase decisions, surpassing green product perception and purchase intention. This aligns with prior literature suggesting that eco-labels serve as trust signals, enhancing product credibility and bridging the gap between environmental attitudes and purchase behavior [8]. The strong correlation between eco-labeling and purchase intention observed in this study further underscores its dual role in shaping pre-purchase evaluations and final purchase choices.

Interestingly, while purchase intention was strongly associated with purchase decision, its predictive influence in the regression and Random Forest models was less dominant than that of eco-labeling. This suggests that even when consumers express intention to purchase green products, tangible indicators such as eco-labels exert a more substantial influence on actual behavior, consistent with the intention-behavior gap documented in green marketing literature [10]. These findings imply that sustainable marketing strategies prioritize credible and visible labeling to convert pro-environmental attitudes into concrete purchasing actions.

From a methodological perspective, the comparative modeling approach revealed that Linear Regression marginally outperformed Random Forest in both R^2 and MSE, particularly after log-transformation of the data. This result is notable given that ensemble methods like Random Forest are typically favored for capturing complex non-linear relationships [13]. The superior performance of Linear Regression may be attributed to the inherently linear relationships among the studied constructs, as indicated by the high inter-variable correlations and the consistent directionality of effects. This finding echoes prior observations that simpler models can sometimes rival or surpass more complex algorithms when applied to well-structured behavioral datasets [11].

Log-transformation application played a crucial role in improving model performance, particularly for the parametric linear model. Transformation reduced skewness and stabilized variance, enhancing normality and homoscedasticity, underpinning regression estimates' reliability. This supports recommendations from previous methodological research emphasizing the importance of preprocessing in achieving robust machine learning outcomes for consumer data [12].

These findings carry important managerial implications. First, eco-labeling should be emphasized in product positioning strategies as a primary driver of green purchase behavior. Businesses and policymakers must ensure that labels are present, recognizable, credible, and aligned with consumer expectations. Second, predictive models built on readily measurable constructs, such as eco-labeling and purchase intention, can be practical tools for market segmentation and targeted interventions in promoting sustainable consumption.

Finally, this study contributes to theoretical advancement by bridging sustainable marketing frameworks with data-driven modeling approaches. By demonstrating that linear models can sufficiently capture consumer decision patterns in emerging markets, this research challenges the assumption that complex algorithms are always superior and encourages future studies to assess model suitability relative to data characteristics and research objectives.

5 Conclusion

This study provides empirical evidence on the determinants of green purchase decisions by comparing linear and non-linear machine learning models using consumer data from Indonesia. The findings demonstrate that eco-labeling is the most influential factor in predicting purchase behavior, followed by purchase intention and green product perception. The dominance of eco-labeling underscores its dual role in shaping attitudes and purchasing decisions, highlighting the importance of credible and visible environmental certifications in marketing strategies.

Methodologically, the comparative modeling reveals that multiple linear regression outperformed Random Forest, particularly after log-transformation of the data. This result suggests that the examined relationships are predominantly linear and that simpler models may offer sufficient predictive accuracy while maintaining interpretability in sustainable consumer research.

For practitioners, the results emphasize the need to prioritize eco-label visibility and authenticity to foster consumer trust and bridge the intention–behavior gap. For academics, the study contributes to the growing literature on integrating machine learning with sustainable marketing frameworks, offering guidance for future research in emerging markets and green consumer behavior analytics

References

1. J. Sun, M. Sarfraz, and I. Ozturk, Eco-consciousness to eco-consumption: unraveling the drivers of sustainable consumption behavior under the mediated-moderated Model. Environ. Sci. Pollut. Res., **31**, 24, 35018 – 35037 (2024). https://doi.org/10.1007/s11356-024-33560-y
2. R. Habib, K. White, D. J. Hardisty, and J. Zhao, Shifting consumer behavior to address climate change. Curr. Opin. Psychol., **42**, 108 – 113 (2021). https://doi.org/10.1016/j.copsyc.2021.04.007
3. E. Purwanto and A. P. Irawan, The Willingness to Purchase Electric Vehicles: A Comprehensive Review. IOP Conf. Ser. Earth Environ. Sci., **1324**, 1 (2024). https://doi.org/10.1088/1755-1315/1324/1/012091

4. K. Lubiński and M. Grębosz-Krawczyk, Economical, Innovative, or Ecological? What Are the Green Energy Users?. Energies, **17**, 21 (2024). https://doi.org/10.3390/en17215419
5. D. Tare et al., Impact of Environmental Regulations on Organizational Change and Innovation. Int. J. Environ. Sci., **11**, 2S, 780 – 786 (2025). https://doi.org/10.64252/ybzzc527
6. D. Ofori, Opportunities and Challenges of Green Marketing. Palgrave Stud. Mark. Emerg. Econ., **Part F3546**, 251 – 276 (2021). https://doi.org/10.1007/978-3-030-74065-8_11
7. M. Christian, E. Purwanto, O. M. Muzammil, and H. Yulita, Eco-Friendly Shopping Bags and The Younger Generation in Urban Areas: Can Environmental Consciousness and Attitude Influence Green Buying Decision?. IOP Conf. Ser. Earth Environ. Sci., **1488**, 1 (2025), https://doi.org/10.1088/1755-1315/1488/1/012028
8. D. P. Alamsyah and R. Febriani, The strategy of eco-friendly products with green consumer behavior : Development of green trust model The strategy of eco-friendly products with green consumer behavior : Development of green trust model. (2021). https://doi.org/10.1088/1755-1315/824/1/012044
9. Y. T. Suyoto, E. Purwanto, E. Pitaloka, A. Dwianika, and V. E. Ristanti, Exploring Green Marketing Research through Bibliometrics. E3S Web Conf., **571** (2024). https://doi.org/10.1051/e3sconf/202457102010
10. I. Ajzen, The theory of planned behavior. Organ. Behav. Hum. Decis. Process., **50**, 2, 179–211 (1991). https://doi.org/10.1016/0749-5978(91)90020-T
11. M. Kuhn and K. Johnson, Applied predictive modeling. (Springer Nature, New York, 2016). https://doi.org/10.1007/978-1-4614-6849-3
12. A. Pourzangbar, M. Jalali, and M. Brocchini, Machine learning application in modelling marine and coastal phenomena: a critical review. Front. Environ. Eng., **2**, September, (2023). https://doi.org/10.3389/fenve.2023.1235557
13. S. Moro, P. Cortez, and P. Rita, A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst., **62**, 22–31 (2014). https://doi.org/10.1016/j.dss.2014.03.001