Griya Journal of Mathematics Education and Application
Volume 6 Nomor 2, Juni 2026
e-ISSN 2776-124X| | p-ISSN 2776-1258
https://mathjournal.unram.ac.id/index.php/Griya/index

Comparative mathematical modeling of coffee shelf-life using linear regression and ensemble learning under simulated storage conditions

Lingga Gita Dwikasari[1*], Dilla Afriansyah[1]

[1] *Ilmu dan Teknologi Pangan, Universitas Mataram*

linggadwikasari@unram.ac.id

Diterima: 16-06-2026; Direvisi: 29-06-2026; Dipublikasi: 29-06-2026

Abstract
Accurate prediction of coffee shelf-life is essential for maintaining product quality and supporting storage management. However, comparative studies evaluating interpretable statistical models and ensemble learning algorithms for coffee shelf-life prediction remain limited, particularly using simulation-based datasets. This study compared the predictive performance of Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR) using a simulation-based dataset of 400 observations representing realistic storage conditions. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). MLR achieved the best performance with the lowest MAE (10.01 days), the lowest RMSE (12.55 days), and the highest R^2 (0.8649), outperforming both ensemble learning models. Feature importance analysis consistently identified storage temperature as the most influential predictor of coffee shelf-life. These findings demonstrate that increased model complexity does not necessarily improve predictive accuracy and support the use of simulation-based datasets for developing predictive models prior to validation with experimental data.

Keywords: Coffee shelf-life; mathematical modeling; multiple linear regression; random forest regression; gradient boosting regression

Abstrak
Prediksi umur simpan kopi yang akurat penting untuk menjaga mutu produk dan mendukung pengelolaan penyimpanan. Namun, penelitian yang membandingkan model statistik yang mudah diinterpretasikan dengan algoritma ensemble learning untuk prediksi umur simpan kopi masih terbatas, terutama menggunakan dataset berbasis simulasi. Penelitian ini membandingkan kinerja Multiple Linear Regression (MLR), Random Forest Regression (RFR), dan Gradient Boosting Regression (GBR) menggunakan dataset simulasi sebanyak 400 observasi yang merepresentasikan kondisi penyimpanan kopi yang realistis. Kinerja model dievaluasi menggunakan Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan koefisien determinasi (R^2). Hasil penelitian menunjukkan bahwa MLR memberikan performa terbaik dengan MAE terendah (10,01 hari), RMSE terendah (12,55 hari), dan R^2 tertinggi (0,8649), melampaui kedua model ensemble learning. Analisis feature importance menunjukkan bahwa suhu penyimpanan merupakan faktor paling berpengaruh terhadap umur simpan kopi. Temuan ini menunjukkan bahwa kompleksitas model yang lebih tinggi tidak selalu meningkatkan akurasi prediksi serta mendukung penggunaan dataset simulasi sebagai tahap awal pengembangan model sebelum divalidasi menggunakan data eksperimen.

Kata Kunci: Gradient boosting regression; multiple linear regression; pemodelan matematika; random forest regression; umur simpan kopi

1. INTRODUCTION

Coffee is one of the most widely consumed beverages in the world and represents an important commodity in the global food industry (Angeloni et al., 2023). In addition to contributing to economic development and international trade, the coffee sector supports the livelihoods of millions of farmers and plays a significant role in food processing activities (Sachs et al., 2019). As consumer expectations continue to increase, maintaining product quality throughout storage and distribution has become essential. Consumers increasingly demand coffee products that retain their characteristic aroma, flavor, freshness, and overall acceptability until the end of their shelf-life.

After processing and packaging, coffee products continue to undergo physical and chemical changes that gradually reduce their quality (Smrke et al., 2022). Oxidation of volatile compounds, moisture migration, and degradation of flavor constituents may alter sensory characteristics and eventually decrease consumer acceptance (Kam et al., 2026). Therefore, accurate estimation of shelf-life is crucial for determining expiration dates, optimizing inventory management, minimizing product losses, and supporting decision-making throughout the coffee supply chain.

Coffee shelf-life is influenced by several interacting factors. Storage temperature can accelerate deterioration reactions, while relative humidity affects moisture exchange between the product and its surrounding environment (Lopriore et al., 2025). Product characteristics such as moisture content and water activity also influence physicochemical stability and susceptibility to quality changes (Ibrahim Garba, 2023). Furthermore, storage duration reflects the cumulative effects of these conditions. Because these variables interact simultaneously, predicting shelf-life is often challenging and may not be adequately represented by simple relationships.

Conventional approaches to shelf-life estimation in food products have largely relied on experimental procedures and kinetic models (Cui et al., 2023). Although these methods provide valuable insights into deterioration mechanisms, they often require extensive experimentation, considerable time, and substantial resources (X. Guo et al., 2026). Moreover, the assumptions underlying traditional models may not fully capture the complexity of food deterioration processes involving multiple interacting factors.

Recent advances in computational methods have introduced data-driven approaches that offer greater flexibility in predictive modeling (X. Guo et al., 2026). Machine learning techniques have been increasingly applied in food science because of their ability to identify complex patterns and model nonlinear relationships without strict assumptions regarding data structure (Lin et al., 2023). Among statistical approaches, Multiple Linear Regression remains widely used due to its simplicity and interpretability, allowing researchers to explicitly quantify the influence of predictor variables on the response (Miller et al., 2022). However, its predictive performance may be limited when the underlying relationships are nonlinear.

Ensemble learning algorithms provide promising alternatives for addressing these limitations. Random Forest Regression improves prediction robustness by aggregating multiple decision trees, whereas Gradient Boosting Regression sequentially refines predictions by minimizing errors from previous learners (Kalule et al., 2023). Both methods have demonstrated strong predictive capabilities across various applications involving complex datasets (Jain et al., 2023).

Despite the growing application of predictive analytics in food research, studies specifically focusing on coffee shelf-life prediction remain limited. Existing investigations predominantly emphasize experimental shelf-life determination, kinetic analyses, or the use of a single predictive approach (Rosillo et al., 2023). Comparative studies evaluating interpretable statistical models alongside ensemble learning techniques in the context of coffee storage are still scarce. To the best of our knowledge, no previous study has comparatively evaluated MLR, RFR, and GBR for coffee shelf-life prediction using simulation-based datasets representing diverse storage conditions. In addition, generating extensive experimental datasets under multiple storage scenarios is often costly and time-consuming, limiting opportunities for methodological exploration (Aung Moon et al., 2024).

Although several studies have applied statistical models or machine learning techniques to food quality prediction, most investigations have focused on evaluating a single predictive approach or optimizing model accuracy within specific experimental datasets (X. Wang et al., 2022). Comparative analyses involving interpretable statistical models and ensemble learning algorithms under identical storage scenarios remain limited, particularly for coffee shelf-life prediction (Collazos-Escobar, Gutiérrez-Guzmán, et al., 2025). Moreover, the scarcity of large experimental datasets covering diverse storage conditions has constrained systematic evaluation of predictive methodologies and limited the reproducibility of comparative studies (Bahamón-Monje et al., 2026). These limitations highlight the need for a standardized framework that enables fair comparison of predictive models under controlled yet realistic storage scenarios.

Simulation based approaches offer an efficient alternative for addressing these challenges (Crowell et al., 2023). By generating datasets that represent realistic storage conditions and natural variability in deterioration processes, simulations enable researchers to evaluate predictive methodologies in a controlled environment before applying them to empirical observations (W. Guo et al., 2022). Such approaches are particularly useful for comparing model performance and investigating the influence of storage variables under diverse scenarios.

Therefore, this study proposes a comparative mathematical modeling framework for predicting coffee shelf-life using a simulation-based dataset representing realistic storage conditions. Unlike previous studies that primarily focused on a single predictive approach or relied solely on experimental observations, this study systematically compares Multiple Linear Regression, Random Forest Regression, and Gradient Boosting

Regression within the same simulation framework. In addition to evaluating predictive performance using multiple statistical metrics, this study incorporates feature importance analysis to identify the relative contribution of storage variables to shelf-life prediction. The proposed framework provides a reproducible and computationally efficient approach for benchmarking predictive models prior to validation with experimental shelf-life data, thereby contributing to the development of data-driven methodologies for coffee storage management.

2. MATERIALS AND METHODS

Research Design

This study employed a comparative mathematical modeling approach to evaluate the performance of different predictive methods in estimating coffee shelf-life under simulated storage conditions. The proposed framework combined a classical statistical approach and ensemble learning techniques to investigate their predictive capabilities. Multiple Linear Regression (MLR) was used as the baseline model because of its interpretability, whereas Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) were selected to represent ensemble learning approaches capable of capturing nonlinear relationships among variables (Kutty et al., 2022; Zhou et al., 2023).

The three models were selected to represent different levels of predictive complexity while maintaining methodological interpretability. Multiple Linear Regression was chosen as a conventional statistical baseline because of its transparent mathematical formulation and ease of interpretation (Kurt, 2024). Random Forest Regression and Gradient Boosting Regression were selected as representative ensemble learning algorithms due to their proven ability to capture nonlinear relationships and complex interactions among predictor variables (Kalule et al., 2023). Comparing these models enables evaluation of whether increased algorithmic complexity provides meaningful improvements over a simpler and more interpretable mathematical model under identical storage conditions.

The overall procedure consisted of dataset generation, data preprocessing, model development, model evaluation, and feature importance analysis. The workflow of the study is illustrated in Figure 1.

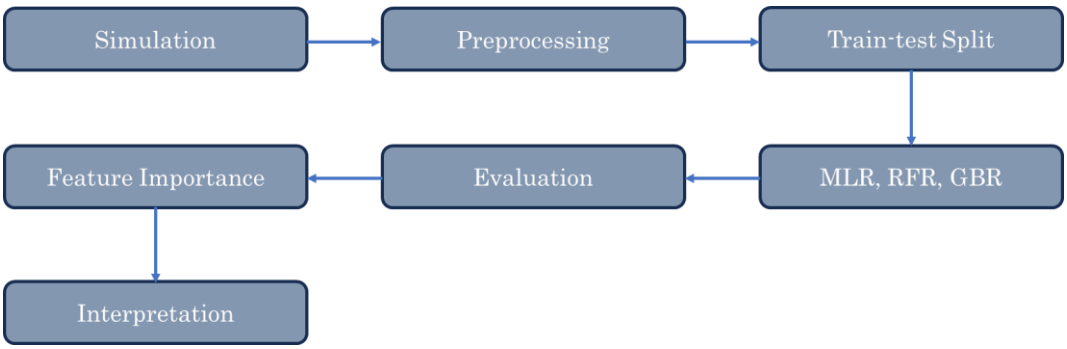

Figure 1. Research Workflow

Simulated Dataset Generation

Experimental datasets describing coffee shelf-life under various storage environments are often limited because they require prolonged observation periods and substantial resources. (Cueva Ríos et al., 2023; Rosillo et al., 2023) Therefore, this study utilized a simulation-based dataset designed to represent realistic storage conditions commonly encountered during coffee storage and distribution.

The simulation ranges adopted in this study were derived from values commonly reported in previous coffee storage and food shelf-life studies. Storage temperature, relative humidity, storage duration, moisture content, and water activity were selected because they are consistently recognized as key factors influencing physicochemical deterioration during storage (Collazos-Escobar, Gutiérrez-Guzmán, et al., 2025). Rather than representing a specific experimental dataset, these ranges were designed to capture realistic variations that may occur during coffee storage, transportation, and distribution. This simulation strategy provides a controlled environment for systematically comparing predictive models while preserving plausible storage conditions reported in the literature.

A total of 400 observations were generated by combining plausible ranges of storage-related variables associated with coffee deterioration. The predictor variables included storage temperature, relative humidity, storage duration, moisture content, and water activity. These variables were selected because they are known to influence physicochemical stability and the rate of quality deterioration during storage (Alessandro Del Nobile & Conte, 2023; Collazos-Escobar, Hurtado-Cortés, et al., 2025). Storage temperature ranged from 20–40 °C, representing typical environmental conditions during storage and transportation. Relative humidity varied between 50–90%, storage duration ranged from 1–180 days, moisture content ranged from 2–8%, and water activity ranged from 0.30–0.70. The response variable was coffee shelf-life, expressed in days.

Each predictor variable was generated independently using continuous uniform distributions within the predefined ranges. This approach ensured that all plausible storage conditions had equal probabilities of being represented while avoiding bias toward specific environmental scenarios. The simulated shelf-life values were subsequently generated according to predefined deterioration assumptions with an added stochastic error component to reproduce natural variability commonly observed in food quality studies.

Shelf-life values were simulated based on realistic deterioration patterns under varying storage conditions, with stochastic variability incorporated to mimic experimental observations. The simulation assumptions were established based on general deterioration principles reported in food shelf-life studies, in which unfavorable storage conditions tend to accelerate quality degradation (Herron et al., 2022). In general, the simulation framework assumed that unfavorable storage conditions, such as higher temperature, elevated relative humidity, prolonged storage duration, increased moisture content, and greater water activity, were associated with shorter shelf-life values.

Random variation was introduced to reproduce the natural fluctuations commonly encountered in food quality studies and to avoid deterministic relationships among variables.

The resulting dataset represented diverse storage scenarios that allowed the predictive performance of different modeling approaches to be evaluated under realistic conditions. This simulation-based strategy provided an efficient alternative to lengthy experimental studies and facilitated methodological exploration prior to the application of empirical observations obtained from actual shelf-life experiments. The simulation framework was intended to represent general coffee storage conditions rather than a specific coffee variety or product type, thereby emphasizing methodological comparison over product-specific prediction.

Data Preprocessing

The generated dataset was examined to ensure completeness and consistency before model development. Since the observations were produced programmatically, no missing values were present and no imputation procedures were required. The dataset was then divided into training and testing subsets using an 80:20 ratio. The training data were used to construct predictive models, while the testing data were used to evaluate the predictive performance of each model on unseen observations.

An 80:20 train-test split was selected because it provides a balanced compromise between model training and unbiased performance evaluation and is widely adopted in predictive modeling studies (Ngusie et al., 2024). The larger training subset allows robust parameter estimation, whereas the independent testing subset provides an objective assessment of model generalization to previously unseen data.

Prior to model development, exploratory analyses were performed to ensure that the simulated dataset was suitable for regression modeling. Because the predictor variables were generated independently within predefined ranges and no missing values were present, the dataset did not exhibit structural inconsistencies that could adversely affect model estimation. The primary objective of this study was comparative predictive performance rather than statistical inference; therefore, emphasis was placed on evaluating model accuracy across different predictive approaches using the same dataset.

Model Development

Three predictive models were developed and compared in this study.

Multiple Linear Regression

Multiple Linear Regression was employed as the conventional mathematical model for predicting coffee shelf-life (Alessandro Del Nobile & Conte, 2023). The relationship between predictor variables and shelf-life was expressed as

$$SL = \beta_0 + \beta_1 T + \beta_2 RH + \beta_3 D + \beta_4 MC + \beta_5 a_w + \varepsilon$$

where SL represents the predicted shelf-life (days), T denotes storage temperature ($°C$), RH denotes relative humidity (%), D represents storage duration (days), MC denotes moisture content (%), a_w represents water activity, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_5$ are regression coefficients, and ε is the random error term.

MLR was selected because it provides explicit mathematical relationships that facilitate interpretation of the effects of predictor variables on shelf-life (Alessandro Del Nobile & Conte, 2023; Cui et al., 2023).

Random Forest Regression

Random Forest Regression is an ensemble learning method that combines predictions from multiple decision trees constructed using bootstrap samples of the training data (Rosati et al., 2023; Thelen et al., 2024). The predicted shelf-life can be expressed as

$$\widehat{SL} = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

where $\widehat{SL}$ is the predicted shelf-life, B denotes the total number of decision trees, and $f_b(x)$ represents the prediction produced by the b-th tree.

By aggregating multiple trees, Random Forest can improve prediction stability, reduce overfitting, and accommodate nonlinear interactions among predictor variables (Camur et al., 2024). This aggregation strategy reduces variance and improves generalization performance (Che et al., 2023).

For model implementation, the Random Forest Regression model was constructed using 100 decision trees with the default Scikit-learn settings. Bootstrap sampling was enabled, and the random state was fixed to ensure reproducibility of the simulation results.

Gradient Boosting Regression

Gradient Boosting Regression constructs predictive models sequentially by minimizing the prediction errors generated by previous learners (Mandal et al., 2021). The model can be formulated as

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $F_m(x)$ is the prediction function at iteration m, $F_{m-1}(x)$ is the prediction obtained from the previous iteration, $h_m(x)$ represents the weak learner fitted at iteration m, and γ_m is the shrinkage parameter (learning rate).

This iterative approach enables the model to capture complex patterns and nonlinear relationships within the data (Ayinde et al., 2023).

Gradient Boosting Regression was implemented using 100 estimators, a learning rate of 0.1, and the default maximum tree depth provided by Scikit-learn. A fixed random state was also specified to ensure reproducibility throughout the model development process.

Model Evaluation

The predictive performances of the developed models were evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

Mean Absolute Error was calculated as

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

where y_i and $\hat{y}_i$ denote the observed and predicted shelf-life values, respectively.

Root Mean Square Error was computed as

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

which gives greater penalties to larger prediction errors (Terven et al., 2025).

The coefficient of determination was calculated as

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

where $\bar{y}$ is the mean of the observed shelf-life values.

Lower MAE and RMSE values indicate better predictive performance, whereas higher R^2 values indicate that a larger proportion of the variability in shelf-life can be explained by the model (S. Wang et al., 2022).

The three evaluation metrics were selected because they provide complementary perspectives on predictive performance. MAE measures the average magnitude of prediction errors and is easily interpretable in the original measurement unit. RMSE assigns greater penalties to larger prediction errors, making it more sensitive to substantial deviations between observed and predicted values. Meanwhile, the coefficient of determination (R^2) quantifies the proportion of variability explained by the model, allowing comprehensive comparison of predictive capability across different modeling approaches (X. Wang et al., 2022).

Feature Importance Analysis

Feature importance analysis was conducted for the ensemble learning models to quantify the relative contribution of each predictor variable to shelf-life estimation. The resulting importance scores were used to identify which storage factors exerted the greatest influence on coffee shelf-life predictions. Higher importance scores indicated stronger contributions of predictor variables to the predictive process. Understanding the

contribution of individual variables may provide practical insights for quality control and support decision-making related to coffee storage management.

Computational Environment

All simulations, model development procedures, and statistical analyses were performed using Python 3 within the Google Colab environment. Several open-source libraries were utilized, including NumPy for numerical computation, Pandas for data management, Scikit-learn for predictive modeling, and Matplotlib for data visualization.

3. RESULTS AND DISCUSSIONS

Descriptive Statistics of the Simulated Dataset

The descriptive statistics of the simulated dataset are presented in Table 1. A total of 400 observations representing diverse coffee storage scenarios were successfully generated. The predictor variables covered a wide range of storage conditions that may realistically occur during coffee storage and distribution. The average storage temperature was 29.88 ± 5.88 °C, ranging from 20.10 to 39.81 °C. Relative humidity exhibited a mean value of 69.99 ± 11.75%, with observations varying between 50.43% and 89.99%. Storage duration ranged from 1.83 to 179.63 days, with an average of 91.14 ± 53.13 days. Moisture content and water activity showed mean values of 4.98 ± 1.74% and 0.50 ± 0.12, respectively. The simulated coffee shelf-life values ranged from 15.99 to 166.55 days, with an average shelf-life of 83.26 ± 34.43 days. These values fall within realistic storage periods reported for packaged coffee products under varying environmental conditions (Rosillo et al., 2023). The broad distribution of shelf-life observations indicates that the generated dataset adequately captured variability in storage environments, thereby providing an appropriate basis for evaluating predictive models.

Table 1. Descriptive Statistics of the Simulated Dataset

Variable	Mean ± SD	Minimum	Maximum
Storage Temperature (°C)	29.88 ± 5.88	20.10	39.81
Relative Humidity (%)	69.99 ± 11.75	50.43	89.99
Storage Duration (days)	91.14 ± 53.13	1.83	179.63
Moisture Content (%)	4.98 ± 1.74	2.03	7.99
Water Activity	0.50 ± 0.12	0.30	0.70
Shelf-Life (days)	83.26 ± 34.43	15.99	166.55

Comparative Performance of Predictive Models

The predictive performances of Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR) are summarized in Table 2. Contrary to the common assumption that more sophisticated machine learning approaches necessarily provide superior predictive capabilities, the present study demonstrated that the conventional Multiple Linear Regression model achieved the highest predictive accuracy (Landschaft et al., 2025). MLR produced the lowest Mean Absolute Error (MAE) of 10.01 days and the lowest Root Mean Square Error (RMSE) of

12.55 days. In addition, MLR attained the highest coefficient of determination (R^2 = 0.8649), indicating that approximately 86.49% of the variability in coffee shelf-life could be explained by the linear model.

In comparison, Random Forest Regression yielded an MAE of 11.49 days, an RMSE of 14.78 days, and an R^2 value of 0.8127. Similarly, Gradient Boosting Regression produced an MAE of 11.50 days, an RMSE of 14.39 days, and an R^2 value of 0.8225. Although both ensemble learning approaches demonstrated satisfactory predictive performance, neither surpassed the simpler MLR model under the simulated storage conditions considered in this study. These findings suggest that the relationships between storage variables and coffee shelf-life within the generated scenarios were sufficiently approximated by linear associations. Consequently, the increased complexity introduced by ensemble learning algorithms did not necessarily translate into improved prediction accuracy.

The present findings are consistent with previous studies reporting that simpler statistical models may achieve predictive performance comparable to or even exceeding that of more sophisticated machine learning algorithms when the underlying relationships among variables are predominantly linear. Ensemble learning methods generally provide greater advantages for highly nonlinear datasets with complex feature interactions. In the present simulation framework, however, the storage variables were generated within realistic yet controlled ranges that preserved predominantly linear deterioration patterns. Consequently, the additional flexibility offered by Random Forest Regression and Gradient Boosting Regression did not produce substantial improvements in prediction accuracy over Multiple Linear Regression.

The superior performance of Multiple Linear Regression also reflects the characteristics of the simulated dataset. Because the response variable was generated according to predefined deterioration assumptions with moderate stochastic variability, the dominant relationships between storage conditions and shelf-life remained largely additive and monotonic. Under such circumstances, the linear model was sufficiently flexible to describe the underlying data structure while avoiding the unnecessary complexity introduced by ensemble learning algorithms. This finding emphasizes that model selection should be guided by data characteristics rather than by algorithmic sophistication alone.

These findings complement previous studies demonstrating the successful application of machine learning techniques for food quality prediction while highlighting that predictive superiority is not universal across all datasets. Several studies have shown that ensemble learning algorithms outperform conventional regression models when nonlinear relationships dominate the data. In contrast, the present study demonstrates that under controlled simulation-based storage scenarios characterized by relatively stable deterioration patterns, Multiple Linear Regression remained highly competitive and provided superior interpretability without compromising predictive performance.

Table 2. Performance Comparison of Predictive Models

Model	MAE	RMSE	R^2
Multiple Linear Regression	10.0147	12.5522	0.8649
Random Forest Regression	11.4913	14.7775	0.8127
Gradient Boosting Regression	11.4990	14.3875	0.8225

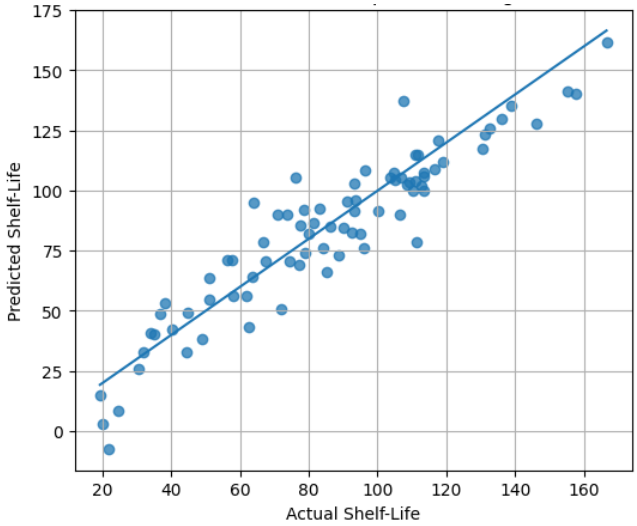

Figure 2. Actual vs Predicted: Multiple Linear Regression

Figure 2 demonstrates that the predictions generated by Multiple Linear Regression closely followed the observed shelf-life values. Most observations clustered around the 1:1 reference line, suggesting that the linear model adequately captured the relationships between storage conditions and coffee shelf-life. Only minor deviations were observed at the lower and upper ranges of shelf-life values.

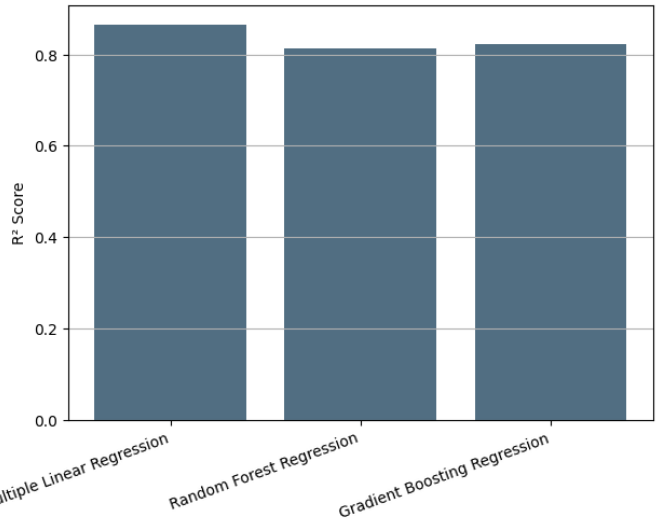

Figure 3. Comparison of Model Performance Based on R^2

Figure 3 further illustrates the comparative predictive performances of the evaluated models. Multiple Linear Regression exhibited the highest coefficient of determination, indicating that the linear model explained approximately 86.49% of the variability in coffee shelf-life. In contrast, Random Forest Regression and Gradient Boosting Regression achieved R^2 values of 81.27% and 82.25%, respectively. These findings suggest that increasing model complexity did not necessarily improve prediction accuracy within the simulated storage scenarios considered in this study.

Feature Importance Analysis

Feature importance analyses derived from Random Forest Regression and Gradient Boosting Regression revealed highly consistent patterns regarding the relative influence of storage variables on coffee shelf-life prediction. For Random Forest Regression, storage temperature emerged as the most influential predictor, accounting for 57.88% of the total importance score. Relative humidity ranked second with an importance value of 21.88%, followed by storage duration at 11.94%. Water activity and moisture content contributed 6.36% and 1.94%, respectively. Similarly, Gradient Boosting Regression identified storage temperature as the dominant predictor, contributing 55.92% of the overall importance. Relative humidity accounted for 21.49%, while storage duration represented 14.04% of the importance score. Water activity and moisture content contributed 6.29% and 2.26%, respectively. The consistent ranking across both ensemble models strengthens the reliability of these findings. Temperature appears to be the primary determinant of coffee shelf-life, highlighting its critical role in accelerating physicochemical deterioration processes (Alessandro Del Nobile & Conte, 2023; Rosillo et al., 2023). Elevated temperatures are generally associated with faster degradation of volatile compounds, increased oxidation rates, and reduced preservation of desirable sensory attributes . Relative humidity and storage duration also exert substantial influences because they affect moisture exchange and cumulative deterioration over time (Feitosa et al., 2020).

In contrast, moisture content exhibited the smallest contribution to predictive performance. This observation may indicate that, within the simulated ranges considered in this study, moisture content had a relatively limited independent effect compared with broader environmental conditions.

Table 3. Feature Importance Obtained from Random Forest Regression

Variable	Importance
Storage Temperature	0.5788
Relative Humidity	0.2188
Storage Duration	0.1194
Water Activity	0.0636
Moisture Content	0.0194

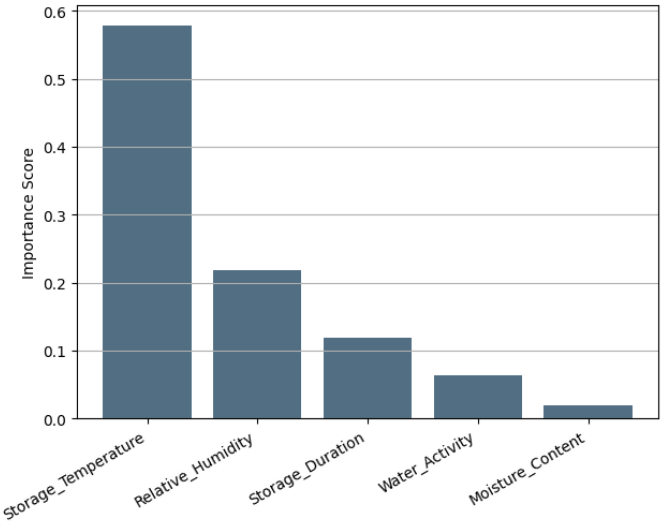

Figure 4. Feature Importance: Random Forest Regression

Figure 4 illustrates the relative importance of predictor variables obtained from Random Forest Regression. The results emphasize the dominant influence of storage temperature, followed by relative humidity and storage duration, indicating that environmental storage conditions exert stronger effects on shelf-life deterioration than intrinsic moisture characteristics within the simulated ranges considered.

Table 4. Feature Importance Obtained from Gradient Boosting Regression

Variable	Importance
Storage Temperature	0.5592
Relative Humidity	0.2149
Storage Duration	0.1404
Water Activity	0.0629
Moisture Content	0.0226

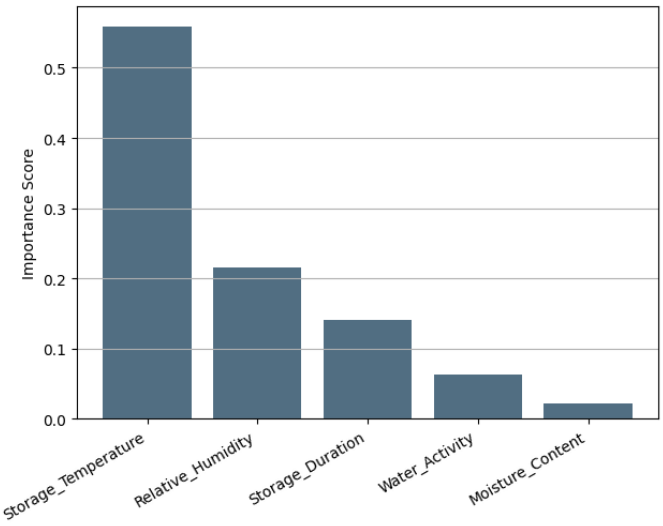

Figure 5. Feature Importance: Gradient Boosting Regression

Figure 5 presents the feature importance scores obtained from Gradient Boosting Regression. The ranking pattern was consistent with the Random Forest model, confirming that storage temperature was the dominant predictor of coffee shelf-life, followed by relative humidity and storage duration. This consistency strengthens the finding that environmental storage conditions played a greater role than moisture-related intrinsic factors within the simulated dataset.

Implications of the Finding

An important implication of this study is that model complexity should not automatically be equated with superior predictive capability. While ensemble learning techniques have gained widespread popularity because of their ability to capture nonlinear relationships, the present findings demonstrate that simpler and more interpretable models may perform equally well or even better when the underlying relationships among variables are predominantly linear.

From a practical perspective, the superior performance of Multiple Linear Regression offers advantages for shelf-life assessment because the model provides explicit mathematical relationships that facilitate interpretation and implementation. This characteristic may be particularly beneficial in industrial settings where transparency and ease of application are essential considerations.

From an industrial perspective, the proposed Multiple Linear Regression model may serve as a practical decision-support tool for estimating coffee shelf-life using routinely measurable storage variables. Because the model is mathematically transparent and computationally inexpensive, it can be readily implemented by coffee producers, storage facilities, and supply chain managers without requiring advanced computational resources. This practicality may facilitate inventory planning, storage optimization, and expiration date estimation in small- and medium-scale coffee industries.

Overall, this study highlights that the selection of predictive methodologies should be guided by the characteristics of the data rather than by assumptions regarding methodological sophistication. Under the simulated storage conditions evaluated herein, conventional linear modeling remained a competitive and effective approach for predicting coffee shelf-life.

Study Limitations

This study employed a simulation-based dataset to enable systematic comparison of predictive models under controlled storage scenarios. Although the simulated data were generated using realistic parameter ranges derived from the literature, they cannot fully represent the complexity and variability of experimental coffee shelf-life observations. Consequently, the predictive performance reported in this study should be interpreted as a methodological comparison rather than definitive evidence of real-world model superiority. Future research should validate the proposed framework using experimental

datasets collected under actual storage conditions and incorporate additional physicochemical and sensory quality indicators to further improve prediction reliability.

4. CONCLUSION

This study comparatively evaluated the performance of Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR) for predicting coffee shelf-life using a simulation-based dataset representing realistic storage conditions. Among the evaluated models, MLR achieved the highest predictive accuracy, demonstrating that a simple and interpretable statistical model can outperform more complex ensemble learning algorithms when the underlying relationships between predictor variables and shelf-life are predominantly linear. Beyond model comparison, this study provides a reproducible simulation-based framework for benchmarking predictive models prior to experimental validation. The consistent identification of storage temperature as the most influential predictor further highlights the importance of environmental storage conditions in maintaining coffee quality and supports its consideration in future predictive modeling studies.

The findings should be interpreted within the scope of the simulated dataset employed in this research. Although the simulation ranges were derived from realistic storage conditions reported in the literature, experimental validation using real coffee shelf-life data remains necessary to confirm the generalizability of the proposed framework. Future studies are therefore encouraged to integrate experimental observations, additional physicochemical and sensory quality indicators, and more diverse predictive algorithms to further enhance the robustness and practical applicability of coffee shelf-life prediction models.

5. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the University of Mataram for its institutional support in facilitating this research and publication. The authors also gratefully acknowledge all co-authors for their valuable contributions, constructive discussions, and collaboration throughout this study. Special appreciation is extended to colleagues and all individuals who provided technical, academic, and administrative assistance during the research process. Their support has been invaluable in the completion of this work.

6. REFERENCES

Alessandro Del Nobile, M., & Conte, A. (2023). Secondary Shelf Life of Foods: State of the Art and Future Perspective. *Food Engineering Reviews*, *15*(4), 748–762. https://doi.org/10.1007/s12393-023-09360-4

Angeloni, S., Giacomini, J., Maponi, P., Perticarini, A., Vittori, S., Cognigni, L., & Fioretti, L. (2023). Computer Percolation Models for Espresso Coffee: State of the Art, Results and Future Perspectives. *Applied Sciences*, *13*(4), 2688. https://doi.org/10.3390/app13042688

Aung Moon, S., Wongsakul, S., Kitazawa, H., Kittiwachana, S., & Saengrayap, R. (2024). Application of ATR-FTIR for Green Arabica Bean Shelf-Life Determination in Accelerated Storage. *Foods*, *13*(15), 2331. https://doi.org/10.3390/foods13152331

Ayinde, A. S., Yu, H., & Wu, K. (2023). Sea level variability and modeling in the Gulf of Guinea using supervised machine learning. *Scientific Reports*, *13*(1), 21318. https://doi.org/10.1038/s41598-023-48624-1

Bahamón-Monje, A. F., Collazos-Escobar, G. A., & Gutiérrez-Guzmán, N. (2026). Data-driven modeling of water adsorption isotherms in cocoa beans: Dataset and Python-based Machine Learning tools for multivariate analysis and storage management. *Data in Brief*, *65*, 112616. https://doi.org/10.1016/j.dib.2026.112616

Camur, M. C., Ravi, S. K., & Saleh, S. (2024). Enhancing supply chain resilience: A machine learning approach for predicting product availability dates under disruption. *Expert Systems with Applications*, *247*, 123226. https://doi.org/10.1016/j.eswa.2024.123226

Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. *Academic Journal of Science and Technology*, *8*(1), 278–280. https://doi.org/10.54097/ajst.v8i1.14333

Collazos-Escobar, G. A., Gutiérrez-Guzmán, N., Váquiro, H. A., García-Pérez, J. V., & Cárcel, J. A. (2025). Analysis of Machine Learning Algorithms for the Computer Simulation of Moisture Sorption Isotherms of Coffee Beans. *Food and Bioprocess Technology*, *18*(6), 5419–5430. https://doi.org/10.1007/s11947-025-03785-x

Collazos-Escobar, G. A., Hurtado-Cortés, V., Bahamón-Monje, A. F., & Gutiérrez-Guzmán, N. (2025). Mathematical modeling of water sorption isotherms in specialty coffee beans processed by wet and semidry postharvest methods. *Scientific Reports*, *15*(1), 3898. https://doi.org/10.1038/s41598-024-83702-y

Crowell, H. L., Morillo Leonardo, S. X., Soneson, C., & Robinson, M. D. (2023). The shaky foundations of simulating single-cell RNA sequencing data. *Genome Biology*, *24*(1), 62. https://doi.org/10.1186/s13059-023-02904-1

Cueva Ríos, M. A., Fernández Rosillo, F., Quiñones Huatangari, L., & Milagros Cabrejos Barrios, E. (2023). Estimation of coffee shelf life under accelerated storage conditions using mathematical models - Systematic review. *Czech Journal of Food Sciences*, *41*(2), 92–102. https://doi.org/10.17221/163/2022-CJFS

Cui, F., Zheng, S., Wang, D., Tan, X., Li, Q., Li, J., & Li, T. (2023). Recent advances in shelf life prediction models for monitoring food quality. *Comprehensive Reviews in Food Science and Food Safety*, *22*(2), 1257–1284. https://doi.org/10.1111/1541-4337.13110

Feitosa, R. M., Figueirêdo, R. M. F. de, Queiroz, A. J. de M., Silva, R. C. da, Oliveira, E. N. A. de, & Moreira, I. D. S. (2020). Evaluation of Lyophilized Myrtle Pulp Powder Stability During Storage. *Journal of Agricultural Studies*, *8*(1), 138. https://doi.org/10.5296/jas.v8i1.15447

Guo, W., Sun, Z., Vilsen, S. B., Meng, J., & Stroe, D. I. (2022). Review of "grey box" lifetime modeling for lithium-ion battery: Combining physics and data-driven methods. *Journal of Energy Storage*, *56*, 105992. https://doi.org/10.1016/j.est.2022.105992

Guo, X., Harbers, C. J., Heuer, H. E. J. M., Min, Q., Liu, Y., Woltering, E. J., Marcelis, L. F. M., & van der Sman, R. G. M. (2026). A generic scientific machine learning framework for fruit and vegetable quality prediction. *Postharvest Biology and Technology*, *234*, 114103. https://doi.org/10.1016/j.postharvbio.2025.114103

Herron, C. B., Garner, L. J., Siddique, A., Huang, T.-S., Campbell, J. C., Rao, S., & Morey, A. (2022). Building "First Expire, First Out" models to predict food losses at retail due to cold

chain disruption in the last mile. *Frontiers in Sustainable Food Systems*, *6*. https://doi.org/10.3389/fsufs.2022.1018807

Ibrahim Garba, A. (2023). Food Preservation Packaging. In *Food Processing and Packaging Technologies - Recent Advances*. IntechOpen. https://doi.org/10.5772/intechopen.110043

Jain, K., Kaushik, K., Gupta, S. K., Mahajan, S., & Kadry, S. (2023). Machine learning-based predictive modelling for the enhancement of wine quality. *Scientific Reports*, *13*(1), 17042. https://doi.org/10.1038/s41598-023-44111-9

Kalule, R., Abderrahmane, H. A., Alameri, W., & Sassi, M. (2023). Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. *Scientific Reports*, *13*(1), 9855. https://doi.org/10.1038/s41598-023-36096-2

Kam, Y. F., Noor Hasnan, N. Z., Abd Ghani @ Hashim, N. H., Jahari, M., Mohd Basri, M. S., & Kamarudin, K. (2026). Analysis of Quality Loss in The Roasted Coffee Bean Storage and Improvement Based on Internet of Things (IoT). *Advances in Agricultural and Food Research Journal*, *7*(1). https://doi.org/10.36877/aafrj.a0000547

Kurt, O. (2024). Model-based prediction of water levels for the Great Lakes: a comparative analysis. *Earth Science Informatics*, *17*(4), 3333–3349. https://doi.org/10.1007/s12145-024-01341-3

Kutty, A. A., Wakjira, T. G., Kucukvar, M., Abdella, G. M., & Onat, N. C. (2022). Urban resilience and livability performance of European smart cities: A novel machine learning approach. *Journal of Cleaner Production*, *378*, 134203. https://doi.org/10.1016/j.jclepro.2022.134203

Landschaft, A., Abdelmoity, L., Chafjiri, F. M. A., Puckett, M. A., Gettings, J., & Loddenkemper, T. (2025). Clinician-in-the-loop screening saturation: predicting annotation yield for efficient EHR review. *BMC Medical Informatics and Decision Making*, *25*(1), 404. https://doi.org/10.1186/s12911-025-03241-y

Lin, Y., Ma, J., Wang, Q., & Sun, D.-W. (2023). Applications of machine learning techniques for enhancing nondestructive food quality and safety detection. *Critical Reviews in Food Science and Nutrition*, *63*(12), 1649–1669. https://doi.org/10.1080/10408398.2022.2131725

Lopriore, M., Alongi, M., Valentino, M., Anese, M., & Nicoli, M. C. (2025). Impact of Environmental Humidity on Instant Coffee Stability: Defining Moisture Thresholds for Quality Degradation and Shelf Life Prediction. *Foods*, *14*(10), 1826. https://doi.org/10.3390/foods14101826

Mandal, P. P., Rezaee, R., & Emelyanova, I. (2021). Ensemble Learning for Predicting TOC from Well-Logs of the Unconventional Goldwyer Shale. *Energies*, *15*(1), 216. https://doi.org/10.3390/en15010216

Miller, A., Panneerselvam, J., & Liu, L. (2022). A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors. *Neurocomputing*, *489*, 466–485. https://doi.org/10.1016/j.neucom.2021.08.150

Ngusie, H. S., Mengiste, S. A., Zemariam, A. B., Molla, B., Tesfa, G. A., Seboka, B. T., Alene, T. D., & Sun, J. (2024). Predicting adverse birth outcome among childbearing women in Sub-Saharan Africa: employing innovative machine learning techniques. *BMC Public Health*, *24*(1), 2029. https://doi.org/10.1186/s12889-024-19566-8

Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Viti, P., Mancini, A., & Frontoni, E. (2023). From knowledge-based to big data analytic model: a novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0. *Journal of Intelligent Manufacturing*, *34*(1), 107–121. https://doi.org/10.1007/s10845-022-01960-x

Rosillo, F. F., Ríos, M. A. C., Huatangari, L. Q., & Lalangui, C. G. S. (2023). Shelf-Life Prediction of Specialty Coffees using the Arrhenius Model. *OnLine Journal of Biological Sciences*, *23*(1), 17–24. https://doi.org/10.3844/ojbsci.2023.17.24

Sachs, J. D., Cordes, K., Rising, J., Toledano, P., & Maennling, N. (2019). Ensuring Economic Viability and Sustainability of Coffee Production. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3660936

Smrke, S., Adam, J., Mühlemann, S., Lantz, I., & Yeretzian, C. (2022). Effects of different coffee storage methods on coffee freshness after opening of packages. *Food Packaging and Shelf Life*, *33*, 100893. https://doi.org/10.1016/j.fpsl.2022.100893

Terven, J., Cordova-Esparza, D.-M., Romero-González, J.-A., Ramírez-Pedraza, A., & Chávez-Urbiola, E. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, *58*(7), 195. https://doi.org/10.1007/s10462-025-11198-7

Thelen, A., Huan, X., Paulson, N., Onori, S., Hu, Z., & Hu, C. (2024). Probabilistic machine learning for battery health diagnostics and prognostics—review and perspectives. *Npj Materials Sustainability*, *2*(1), 14. https://doi.org/10.1038/s44296-024-00011-1

Wang, S., Takyi-Aninakwa, P., Jin, S., Yu, C., Fernandez, C., & Stroe, D.-I. (2022). An improved feedforward-long short-term memory modeling method for the whole-life-cycle state of charge prediction of lithium-ion batteries considering current-voltage-temperature variation. *Energy*, *254*, 124224. https://doi.org/10.1016/j.energy.2022.124224

Wang, X., Bouzembrak, Y., Lansink, A. O., & van der Fels-Klerx, H. J. (2022). Application of machine learning to the monitoring and prediction of food safety: A review. *Comprehensive Reviews in Food Science and Food Safety*, *21*(1), 416–434. https://doi.org/10.1111/1541-4337.12868

Zhou, Z., Qiu, C., & Zhang, Y. (2023). A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. *Scientific Reports*, *13*(1), 22420. https://doi.org/10.1038/s41598-023-49899-0