IJASST International Journal of Applied Sciences and Smart Technologies
Volume 08
Issue 01
p-ISSN 2655-8564, e-ISSN 2685-943, pp. 183-202, 2026

Principal Component Analysis-Driven Feature Reduction for Predicting Coffee Quality Using a Machine Learning Approach

Siti Yuliyanti[1 *], Heni Sulastri [2], Sakifah [3]

[1]*Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, West Java, Indonesia*
[2]*Department of System Information, Faculty of Engineering, Siliwangi University, Tasikmalaya, West Java, Indonesia*
[3]*Department of Digital Banking and Finance, Faculty of Economics and Business, Siliwangi University, Tasikmalaya, West Java, Indonesia*
[*]*Corresponding Author: sitiyuliyanti@unsil.id*

(Received 29-09-2025; Revised 13-01-2026; Accepted 09-04-2026)

Abstract

Coffee quality assessment using a machine learning approach faces major challenges, including high data dimensionality and redundancy between features. Therefore, PCA is proposed as a feature reduction technique to improve the efficiency and accuracy of coffee quality prediction models. The research phase began with data acquisition, data cleaning, feature engineering, explanatory data analysis, testing the normalization of coffee parameter profiles, implementing PCA on Random Forest and XGBoost models, and then evaluating model performance. Model evaluation using MAE and MAPE showed that Random Forest provided more precise predictions than XGBoost, particularly when applying PCA. This resulted in a 39% performance increase for Random Forest from 0.11903 to 0.08542 and an 8% increase for XGBoost, shifting the score from 0.12511 to 0.11570. Prediction visualization reinforced the consistency and precision of the Random Forest model, regardless of whether PCA was used. The findings of this study highlight the importance of feature cleaning and engineering, and the role of PCA in improving the precision of coffee quality predictions. The use of the Random Forest model with PCA is recommended as an efficient method for modeling the quality of Arabica coffee, taking into account sensory and environmental factors.

Keywords: Coffee Quality, Feature Reduction, Pearson Correlation, PCA, Principal Component

1 Introduction

Arabica coffee (Coffea arabica) is the most highly valued coffee variety globally, accounting for over 60% of global coffee production. It possesses high economic value, with a complex flavor and distinctive aroma influenced by environmental, genetic, and post-harvest factors. Traditionally, coffee quality evaluation is conducted through a

cupping process by trained panelists, but this method is subjective, time-consuming, and inconsistent across evaluators[1], [2]. Therefore, a data-driven approach using artificial intelligence is increasingly needed to improve efficiency and objectivity in coffee quality analysis. One challenge in coffee quality data analysis is the high dimensionality of the data, encompassing sensory, chemical, and physical features. To address this, Principal Component Analysis (PCA) is used as a dimensionality reduction technique to remove redundancy and retain the most significant information in the form of principal components [3]. PCA can help simplify input data while improving the performance of classification models.

Machine learning approaches have been widely used to automatically classify coffee quality based on sensory, physical, or chemical data[4], [5]. Support Vector Machines and Random Forests are capable of classifying coffee based on geographic origin with high accuracy. However, one of the main challenges in coffee data processing is the large number of features, which can lead to overfitting and decreased predictive performance[6]. The Random Forest and Extreme Gradient Boosting (XGBoost) algorithms have proven effective for classifying complex and heterogeneous data, including in the agriculture and food sectors. Random Forest is a decision tree-based ensemble algorithm that works by randomly constructing multiple trees and combining their results through voting. Its advantages lie in its resistance to overfitting and its ability to handle nonlinear data[7], [8]. Meanwhile, XGBoost is one of the most advanced and efficient boosting algorithms designed for high performance, large scalability, and high accuracy. Compared to traditional boosting algorithms, XGBoost offers optimizations in terms of regularization, execution speed, and overfitting control[9], [10]. In previous studies, XGBoost demonstrated superiority in the quality classification of agricultural products, including coffee, with higher accuracy than other models[11], [12].

Previous studies have only used coffee datasets from a single country or region [13], [14], so the prediction models have not been tested on global Arabica coffee data, which has high genetic, environmental, and post-harvest diversity. Previous studies have rarely conducted direct comparisons between the Random Forest and XGBoost algorithms in the context of sensory data-based coffee quality prediction, especially after dimensionality reduction using PCA[3], [15], [16]. Most studies focus solely on

prediction accuracy, without considering model efficiency and its potential application in automated quality assessment systems in the coffee industry. No research has yet tested the combination of PCA as the primary feature extraction with Random Forest and XGBoost, and applied it to global Arabica data from various countries (with differences in flavor, aroma, and body).

This study aims to develop a quality classification model for Arabica coffee from various countries by applying PCA as a feature extraction stage, which is then used as input to the Random Forest and XGBoost models. The global dataset used includes sensory data from cupping results such as aroma, flavor, acidity, body, and aftertaste. Evaluation was conducted to measure the performance of both models in predicting coffee quality based on the results of dimensionality reduction. This approach is expected to produce an accurate, efficient, and reliable coffee quality evaluation system, as well as serve as a basis for the development of an automated coffee grading system on a global industrial scale.

2 Material and Methods

This research methodology outlines the research stages, into several parts as illustrated in Fig. 1.

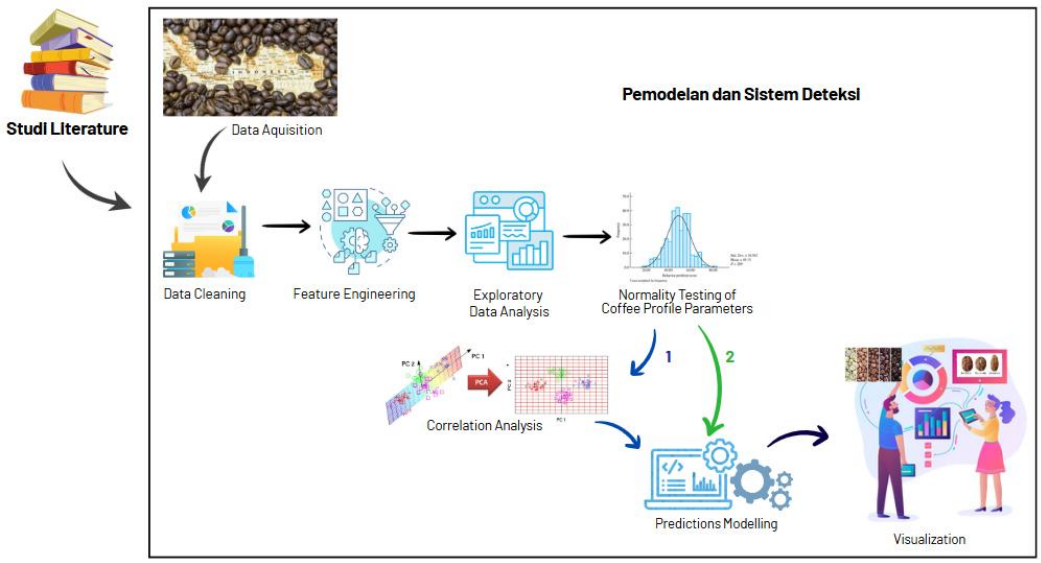


Figure 1. Framework Research

Data Acquisition

The dataset used in this study was obtained from the Coffee Quality Institute (CQI), which provides publicly accessible Arabica coffee cupping data collected under standardized evaluation protocols. The dataset is available for academic research purposes under open access terms. After data cleaning and filtering, the final dataset consisted of 207 samples with 41 original features, including sensory attributes (e.g., aroma, flavor, aftertaste), environmental factors (e.g., altitude), and processing information. PCA was applied only at the modeling stage, reducing the feature space to 2–3 principal components while preserving the original dataset dimensionality for exploratory and correlation analysis.

A limitation of this dataset is its moderate sample size relative to the number of sensory features, which motivates the use of dimensionality reduction to prevent overfitting.

Data Cleansing

Data cleansing is essential for accurate statistical calculations and machine learning models. The steps include:

1. Feature scaling and feature selection: filling all empty cells with mode data, standardizing the scale for a value, and eliminating redundant columns to speed up the data analysis process[17], [18].
2. Altitude: cleaning numeric data defined as strings. The altitude feature is used in models to predict flavor profiles or quality scores. For example, coffee from Garut, at an altitude of 1,400 m above sea level, likely has higher acidity and aroma than coffee from lower areas. Based on the values listed in all cells in the "Altitude" column, we can see that the values are a minimum value. This is not good for data analysis and model building. The middle value of the range is used to replace it.

Feature Engineering

This research adds new features that may be useful for the model creation process, based on existing data. For the harvest year data, we will use the first year[19], [20]. This is so that when calculating the coffee bean age, we can consider the possibility of rotting beans first. The data does not specify the harvest date; therefore, we will assume that all

harvests occurred simultaneously on January 1, 2021, or 2022 and 2023. This simplifies the calculation of coffee age. Next, perform the calculation as below. Next, process the coffee plantation height data, categorizing the height of the farm based on the dataset. After the data cleaning and feature engineering process, the dataset, consisting of 270 rows and 41 columns, became 207 rows and 2 columns.

Exploratory Data Analysis

Well-organized data makes the EDA process more effective. Initially, summary statistics are shown for the complete dataset, indicating the type of data and the quantity of missing (null) entries in the dataframe. The examination proceeds with visual representations of coffee variety information from each nation to identify which country boasts the highest diversity of coffee varieties.

Normality Test for Coffee Profile Parameters

Normality assessments are conducted on each coffee profile characteristic, including acidity, sweetness, balance, and so forth. This step is vital for every data point, confirming a suitable distribution with the help of the Shapiro-Wilk normality test, especially since a vast amount of data is sourced from worldwide datasets[4], [12].

Correlation Analysis of Factors Influencing Coffee Quality

A Pearson correlation analysis was carried out to assess how strongly the sensory attributes observed during cupping are related to the overall coffee quality rating (Total Cup Points)[21]. The Pearson correlation coefficient (r) for every combination of parameters was determined using Equation 1. Visualization is done using heatmaps to facilitate identification of relationship patterns between features.

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})}.\sqrt{\sum_{i=1}^{n}(x_i-\bar{y})^2}.} \quad (1)$$

r_{xy} : correlation coefficient between variables x and y

$\bar{x}, \bar{y}$: average of each variable

n : number of data, the r value ranges from −1 to +1

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is applied as a key dimensionality reduction technique to extract the most informative components from high-dimensional coffee sensory data[3], [15]. The original dataset consists of several correlated sensory attributes—such as aroma, flavor, aftertaste, acidity, body, balance, and overall score—which, while individually informative, tend to exhibit multicollinearity and redundancy when used collectively as input to a machine learning model. PCA transforms these original variables into a new set of uncorrelated variables known as principal components (PCs), which are ordered by the amount of variance they capture from the original data. The first step is to standardize the data so that all features have the same scale (mean = 0, std = 1) using Equation 1.

$$Z_{ij} = \frac{Xij - \mu j}{\sigma_j} \quad (1)$$

X_{ij} : value of the j-th feature from the i-th observation

μ_j : mean of the j-th feature

σ_j : sigma_j

σj : standard deviation of the j-th feature

Next, determine the Covariance matrix using Equation 2, in which z represents the normalized data matrix (mean = 0), C denotes the covariance matrix, n indicates the number of samples, and p refers to the number of features.

$$C = \frac{1}{n-1} Z^T Z \quad (2)$$

After obtaining the covariance value, calculate the eigenvalues (λ) and eigenvectors (*e*) using Equation 3. Then, project the principal component data using the projection matrix or principal components using Equation 4. Where Z is the standardized data matrix, E_k: the p × k eigenvector matrix, and PCS is the result of transforming the data into a new *k*-dimensional space.

$$C_e = \lambda_e \quad (3)$$

$$PCS = Z.E_k \quad (4)$$

Random Forest (RF) Modeling

Random Forest is a group learning method that creates several decision trees during the training phase, producing class predictions for classification purposes. This approach enhances prediction accuracy by combining the outcomes of numerous weak learners, each trained on a randomly selected subset of the data and characteristics (bagging)[8]. The RF model utilizes principal components derived from PCA as input variables, which helps minimize redundancy and boosts computational effectiveness[22]. Tuning of hyperparameters is conducted to refine key settings, including the number of trees (n_estimators), maximum allowable depth (max_depth), and the number of features evaluated at each split (max_features). This model showcases strong generalization skills and is less likely to face overfitting due to its smoothing properties.

Extreme Gradient Boosting (XGBoost) Modeling

XGBoost is a sophisticated version of gradient boosting that constructs trees one after another, with each new tree focusing on correcting the mistakes of the preceding one. This model incorporates a regularized learning approach to limit overfitting and supports parallel processing along with effective management of missing data[22]. The features transformed by PCA were employed to train the XGBoost model, and a grid search was carried out to fine-tune important parameters like the learning rate (eta), maximum depth of the trees (max_depth), and the number of boosting iterations (n_estimators). XGBoost showed remarkable performance in both training and validation phases, excelling in recognizing intricate feature interactions and non-linearities in the dataset.

Model Evaluation

Performance assessment in this research aims to analyze the use of Random Forest and XGBoost, both with and without PCA, utilizing MAE to quantify the average absolute deviation between actual and predicted figures, and MAPE (Mean Absolute Percentage Error) serves as a measure of the average absolute error expressed as a percentage of the actual value[23], [24].

3 Results and Discussions

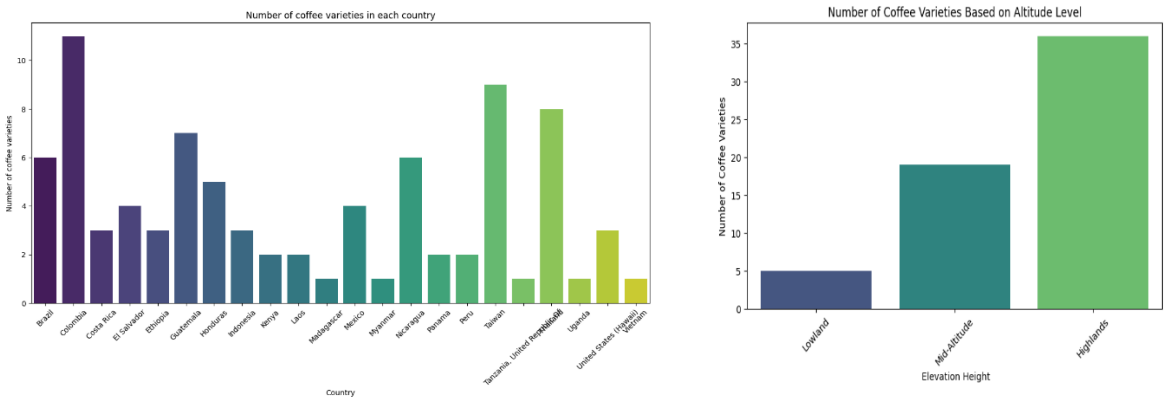
The process of cleaning data involves an altitude characteristic that is utilized in the model to forecast taste characteristics or quality ratings. For instance, coffee sourced from Garut, situated at 1,400 meters above sea level, typically possesses greater acidity and fragrance compared to coffee harvested from lower altitudes. By examining the figures in the "Altitude" column, we can observe a variation of values present. Such discrepancies are not optimal for analysis and model creation. Therefore, we will replace the varied values with the middle value of the range. Additionally, due to the lengthy nature of the task, we will eliminate certain columns deemed unnecessary and subsequently review for any parameters that contain null values or vacant cells. The empty entries will be filled using the mode of the data. For instance, in this analysis, the "Processing Method" column contains missing entries. We will use mode imputation to address these gaps. At this stage, the aim is to display the unique values, maximum values, and minimum values from the "Altitude" column. To facilitate this, we assign each altitude a category, namely Highlands, Mid-Altitude, and Lowland, as shown in Table 1.

Exploratory data analysis in this study visualizes how many coffee varieties there are in each country, as shown in Fig. 2. Fig. 2 (a) shows the number of coffee varieties in each country, and (b) displays a bar graph illustrating the variety of coffee types found across three elevation groups: lowland, mid-altitude, and highland.

Table 1. Altitude category labeling table

ID	Altitude	Altitude_Category
0	1815	Highlands
1	1200	Mid-Altitude
2	1300	Highlands
...	...	...
206	1300	Highlands
207	1200	Mid-Altitude

The horizontal axis (X) denotes the elevation groups, while the vertical axis (Y) indicates how many coffee varieties have been documented within each group. The depicted data suggests that the total number of coffee varieties tends to rise as altitude increases. The lowland group contains the fewest varieties, approximately 5. In the mid-altitude range, this amount jumps significantly to around 18 varieties. The highland group shows the greatest diversity, with about 35 varieties recorded. This observation implies that regions with higher altitudes tend to exhibit a richer diversity of coffee types. This trend can be linked to the more advantageous agroclimatic conditions found in highland areas, such as cooler temperatures, reduced temperature variations, and improved drainage in the soil. These factors may create a more consistent and suitable environment for a variety of coffee plants, especially the Coffea arabica species, which is recognized for its ability to thrive at specific elevations. Overall, this chart provides evidence supporting the idea that altitude is a significant ecological element affecting the distribution and variety of coffee species within a particular area. These findings hold valuable significance for strategies aimed at conserving coffee genetic resources and for the planning of cultivation practices that foster the growth of elite varieties suited to particular geographic conditions.



(b)

Figure 2. Visualisasi exploratory data analysis: (a) Number of coffee varieties in each country (b) Number of coffee varieties based on altitude level

The stages of the coffee profile parameter normality test using the Kolmogorov-Smirnov test are to check whether each column in the dataset is normally distributed or not. Then, a correlation analysis is performed on the numerical parameters and total cup points (coffee value), consisting of sweetness, acidity, aroma, flavor, aftertaste, body, balance, uniformity, and total cup points. The Pearson correlation is then calculated and visualized in a heatmap in Fig. 3.

Fig. 3 displays a heatmap that illustrates the correlations among various coffee quality evaluation factors, which consist of sweetness, acidity, aroma, flavor, aftertaste, body, balance, uniformity, and total cup points. The correlation coefficients were derived using the Pearson correlation technique, with values ranging from -1 to 1, signifying both the intensity and direction of the linear relationships among the factors. The findings reveal that flavor has the strongest correlation with total cup points, scoring 0.93, closely followed by aftertaste (0.92), balance (0.92), acidity (0.89), and sweetness (0.89). This suggests that flavor, aftertaste, and balance are key elements in assessing the overall quality of coffee flavor. Additionally, uniformity demonstrated a very weak correlation with total cup points (0.00) as well as with other factors, implying that it has little impact on the overall cup rating.

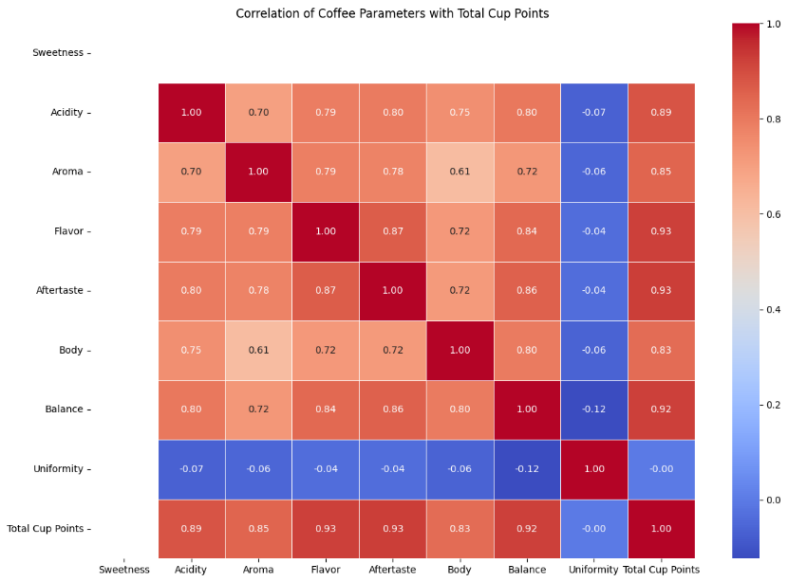


Figure 3. Correlation of coffee parameters with total cup points

At the same time, a reasonably strong positive correlation was observed between flavor and aftertaste (0.87), along with a notable correlation between balance and aftertaste (0.86), underscoring the interconnections in sensory attributes related to flavor and balance during coffee assessment.

Cumulative Variance Analysis Based on Principal Component Analysis (PCA) in Fig. 4 presents a plot of the cumulative variance explained by each principal component based on the results of a PCA analysis of coffee quality feature data.

The purpose of this visualization is to determine the optimal number of principal components capable of representing the information in the initial data with minimal information loss. The analysis results show that the first principal component explains approximately 74% of the total variance in the dataset. The addition of the second component increases the cumulative explained variance to 87%, while the first three components account for over 90% of the total variance. Furthermore, six principal components successfully explain almost all of the variance in the data (approaching 100%). The curve shown shows a decreasing pattern in the rate of variance increase, with an elbow visible at the third to fourth components. This indicates that most of the information in the dataset can be effectively represented using only three to four principal components without significant information loss.

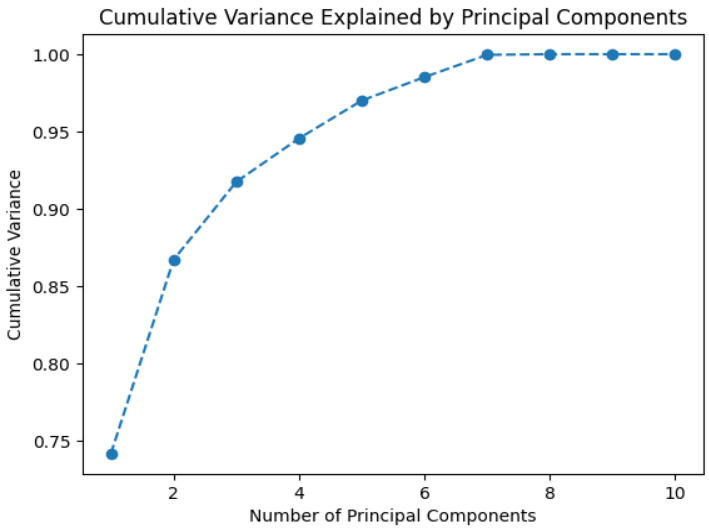


Figure 4. Result: cumulative variance explained by PCA

Thus, these results can be used as a basis in the dimensionality reduction process to simplify data, speed up the computational process, and reduce the risk of overfitting in subsequent modeling, such as classification or regression using machine learning algorithms.

The following step involves choosing the number of components for PCA. In this research, two primary components were utilized since the cumulative variance chart indicates that the initial two components account for roughly 87% of the total variance, whereas including the first three components surpasses 90%. As a result, opting for three components is deemed better to preserve adequate information from the original dataset, particularly when using the Random Forest and XGBoost methods. The PCA results for the explained variance ratio reveal that PC1, PC2, and PC3 have values of 0.741, 0. 26, and 0.050, respectively. Explained Variance Ratio is used in dimensional analysis in Principal Component Analysis (PCA).

Based on Table 2, the PCA analysis results show that the first principal component (PC1) explains 74.1% of the total variance in the data, while the second component (PC2) accounts for 26.0%. The combination of these two components accounts for approximately 100% of the information in the dataset. The third component (PC3) only explains an additional 5.0% of the variance and is therefore considered less significant. Therefore, the first two principal components are considered representative enough for use in further analysis and dimensionality reduction. The modeling phase begins with dividing the training and test data and scaling based on the PCA-reduced dataset.

Principal Component Analysis (PCA) is theoretically justified in this study due to the strong multicollinearity observed among sensory attributes such as flavor, aftertaste, balance, and acidity, as evidenced by Pearson correlation coefficients exceeding 0.85. High feature correlation can degrade the performance of tree-based ensemble models by introducing redundant splits and unstable decision boundaries.

PCA addresses this issue by projecting the original correlated feature space into a lower-dimensional orthogonal space defined by eigenvectors of the covariance matrix. Each principal component corresponds to an eigenvector associated with an eigenvalue that quantifies the amount of variance captured. According to Kaiser’s criterion,

components with eigenvalues greater than 1 should be retained, while the scree plot elbow method helps identify the point of diminishing returns in variance contribution. Based on both the scree plot elbow observed between PC2 and PC3 and the Kaiser criterion, retaining two principal components is statistically sufficient, capturing over 87% of the total variance without introducing noise from lower-variance components.

X is the principal feature, and y is the target prediction, which is then evaluated using MAE and MAPE for both the Random Forest and XGBoost models. The Random Forest model uses n_estimators = 300 and random_state = 42, where n_estimators is the number of decision trees in the Random Forest. The more trees, the more stable and accurate the prediction results, up to a certain point. The optimal random_state in previous machine learning research approaches averages 42. At the same time, the settings implemented in XGBoost include learning_rate set at 0.1, n_estimators at 300, random_state as 42, reg_lambda valued at 2, and verbosity level of 2.

Table 3 displays model performance using MAPE and MAE, which aim to assess, compare, and communicate the accuracy of regression models quantitatively and interpretably. The combination of the two provides a comprehensive picture of how well the model performs in a real-world context.

Table 2. Eigenvalues and Cumulative Explained Variance

PC	Eigenvalue	Explained Variance (%)	Cumulative (%)
PC1	5.93	74.1	74.1
PC2	2.08	26.0	100
PC3	0.41	5.0	105

Table 3. Evaluation of model performance

Evaluation	non-PCA		PCA	
	Random Forest	XGBoost	Random Forest	XGBoost
MAPE	0.11903	0.12511	0.08542	0.11570
MAE	0.09974	0.10464	0.07186	0.09735

Based on the findings under conditions without PCA, the Random Forest model exhibited a MAPE of 0. 11903 and an MAE of 0. 09974, while XGBoost had a MAPE of 0. 12511 and an MAE of 0. 10464. This suggests that, when dimensionality reduction isn't applied, Random Forest outperforms XGBoost in both absolute and relative error metrics. After implementing PCA, there was a noticeable enhancement in the performance of both models, especially for Random Forest. The Random Forest model utilizing PCA achieved the lowest MAPE of 0. 08542 and an MAE of 0. 07186, demonstrating a substantial improvement in accuracy compared to its performance without PCA. On the other hand, the XGBoost model also showed a minor improvement in accuracy with PCA, as reflected by a drop in MAPE to 0. 11570 and a decrease in MAE to 0. 09735, although it remained less accurate than Random Forest in both scenarios. In summary, these outcomes suggest that employing PCA as a method for dimensionality reduction enhances the performance of the Random Forest model by 39% and the XGBoost model by 8%. Additionally, Random Forest consistently achieves better results than XGBoost across all evaluation metrics in every scenario. These results indicate that, for the dataset evaluated, Random Forest demonstrates greater adaptability to the data structure both prior to and following the PCA transformation, making it a more dependable option for the predicted task at hand.

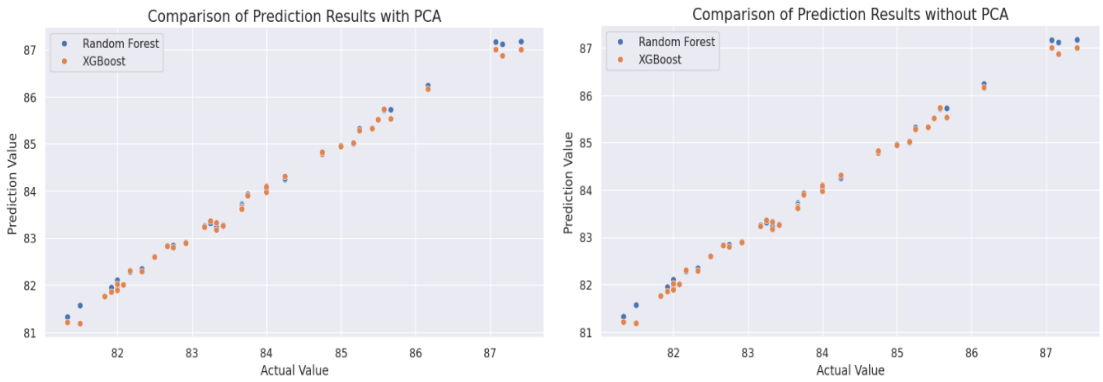


Figure 5. Visualisasi comparison of prediction: (a) model using PCA (b) model without PCA

Fig. 5 (a) illustrates a comparison of the results obtained from two regression techniques, Random Forest and XGBoost, about the real values after utilizing Principal Component Analysis (PCA) for reducing dimensions. The graph shows the correspondence between the actual values on the X-axis and the predicted values on the Y-axis, with each dot representing an individual observation. The blue dots signify the forecasts made by the Random Forest model, whereas the orange dots indicate predictions from XGBoost. Overall, both models display a linear pattern and move closer to the identity line (which signifies a perfect fit), suggesting that their predictions tend to be fairly accurate when compared to the actual data. Nonetheless, it appears that the forecasting points from the Random Forest model generally fall nearer to the real values than those from XGBoost, which sometimes shows slightly greater discrepancies. This graph reinforces prior quantitative results shown in Table 3, which indicated that Random Forest yielded lower MAE and MAPE figures after PCA was applied. Consequently, it can be deduced that the Random Forest model not only has superior numerical effectiveness but also offers greater consistency in prediction stability against the real values represented in this graph. Additionally, this visualization verifies that implementing PCA does not significantly weaken the model's predictive power; rather, it enhances prediction accuracy, particularly for the Random Forest model.

Fig. 5(b) presents a visualization of the comparison of the prediction results between the Random Forest and XGBoost models against the actual values, without Principal Component Analysis (PCA). This graph displays the relationship between the actual values (X-axis) and the predicted values generated by both models (Y-axis), where each point represents a prediction from a single observation. The blue color represents the prediction results from the Random Forest model, while the orange color represents the prediction results from XGBoost. In general, both models show a fairly accurate and linear trend in their predictions relative to the actual values. Most data points are quite close to the diagonal line (an imaginary line representing perfect prediction), indicating that both Random Forest and XGBoost have good predictive capabilities without dimensionality reduction. However, a visual comparison shows that Random Forest slightly outperforms in terms of prediction consistency, especially in the mid-to-high range. This aligns with the quantitative evaluation results in Table 3, where Random

Forest recorded lower MAPE (0.11903) and MAE (0.09974) values compared to XGBoost (MAPE 0.12511, MAE 0.10464) in the scenario without PCA. This visualization reinforces the interpretation that although both models perform reasonably well without feature transformation, Random Forest provides more stable and accurate performance than XGBoost in representing the relationship between input features and target outputs.

4 Conclusions

The data cleaning phase of this research incorporated altitude features that affect the flavor characteristics and quality of coffee. Variations in altitude measurements that were inconsistent were substituted with the median of their respective altitude groups to assist in analysis. Furthermore, irrelevant columns were eliminated, and any gaps in the data were filled in using the most frequent value to ensure uniformity in the dataset. The preliminary data analysis indicated that as elevation increased, there was a rise in the variety of coffee types, with the highest diversity found in elevated areas. This trend likely stems from the favorable climate and agricultural conditions in highlands, which promote the cultivation of high-quality Coffea arabica types. Normality assessments and Pearson correlation analyses revealed a strong positive relationship between flavor characteristics such as taste, aftertaste, and balance with total cup points, underscoring the significance of these factors in assessing overall coffee quality. The Principal Component Analysis (PCA) demonstrated that three primary components accounted for more than 90% of the data's variance, showing that PCA is effective for reducing the complexity of the data without significant information loss. During the modeling phase, both Random Forest and XGBoost regression models were evaluated, with and without PCA.

The performance analysis using MAPE and MAE indicated that Random Forest consistently performed better than XGBoost. The introduction of PCA enhanced the accuracy of the Random Forest model by 39%, while XGBoost's accuracy increased by 8%. Visual representations of the predictions affirmed that the Random Forest model offered more precise and reliable forecasts compared to XGBoost, whether PCA was applied or not. These results suggest that Random Forest is a more trustworthy method

for forecasting coffee quality in this dataset, with PCA proving to be an effective approach for enhancing model accuracy and computational speed.

Acknowledgements

The research team would like to express its gratitude to LPPM Siliwangi University for the assistance and funding for the 2025 budget year so that this research can be completed and published.

References

[1] V. Sari, F. Firdausi, and Y. Azhar, "Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.

[2] R. Rusinek *et al.*, "How to Identify Roast Defects in Coffee Beans Based on the Volatile Compound Profile," *Molecules*, vol. 27, no. 23, pp. 1–13, 2022, doi: 10.3390/molecules27238530.

[3] F. H. Alkallas, A. M. Mostafa, E. A. Rashed, A. B. G. Trabelsi, M. A. I. Essawy, and R. A. Rezk, "Authentication of Roasted Coffee Beans via LIBS: Statistical Principal Component Analysis," *Coatings*, vol. 13, no. 10, 2023, doi: 10.3390/coatings13101790.

[4] Kamil Fadli, "Pengolahan Citra Digital Menggunakan Metode Yolo Untuk Mendeteksi Kualitas Dari Biji Kopi Berbasis Android," *J. Artif. Intel. dan Sist. Penunjang Keputusan*, vol. 1, no. 1, pp. 120–125, 2023, [Online]. Available: https://jurnalmahasiswa.com/index.php/aidanspk

[5] S. Asghar, J. Choi, D. Yoon, and J. Byun, "Spatial pseudo-labeling for semi-supervised facies classification," *J. Pet. Sci. Eng.*, vol. 195, Dec. 2020, doi:

10.1016/j.petrol.2020.107834.

[6] V. No, D. Gusmaliza, and S. Aminah, "Edumatic : Jurnal Pendidikan Informatika Sistem Identifikasi Kualitas Biji Kopi Robusta berbasis Image Processing dengan Support Vector Machine," vol. 8, no. 2, pp. 744–753, 2024, doi: 10.29408/edumatic.v8i2.28008.

[7] Rizkya Nur Amalda, "Decision Tree: Algoritma C5.0." [Online]. Available: https://rizkyanuramalda.medium.com/decision-tree-algoritma-c5-0-52e76661a31d

[8] K. Ciptady, M. Harahap, J. Jonvin, Y. Ndruru, and I. Ibadurrahman, "Prediksi Kualitas Kopi Dengan Algoritma Random Forest Melalui Pendekatan Data Science," *Data Sci. Indones.*, vol. 2, no. 1, 2022, doi: 10.47709/dsi.v2i1.1708.

[9] G. Idoje, C. Mouroutoglou, T. Dagiuklas, A. Kotsiras, I. Muddesar, and P. Alefragkis, "Comparative analysis of data using machine learning algorithms: A hydroponics system use case," *Smart Agric. Technol.*, vol. 4, Aug. 2023, doi: 10.1016/j.atech.2023.100207.

[10] T. Firmansyah, R. Kurniawan, and A. T. Hidayat, "Klasfikasi Tingkat Kematangan Roasting Biji Kopi Berbasis Deep Learning dengan Arsitektur MobileNet," vol. 6, no. 2, pp. 1432–1442, 2025, doi: 10.47065/josh.v6i2.6811.

[11] D. Herdhiansyah, S. Sudarmi, S. Sakir, and A. Asriani, "Analisis Faktor Prioritas Pengembangan Komoditas Perkebunan Unggulan Dengan Metode Ahp (Analtycal Hierarchy Process," *J. Tek. Pertan. Lampung (Journal Agric. Eng.*, vol. 10, no. 2, p. 239, 2021, doi: 10.23960/jtep-l.v10i2.239-251.

[12] S. Wahyudi, "Aplikasi Deteksi Kualitas Biji Kopi Menggunakan Metode Histogram Equalization Berbasis Android," vol. 2, no. 1, pp. 50–56, 2018.

[13] K. Przybył *et al.*, "Application of Machine Learning to Assess the Quality of Food Products—Case Study: Coffee Bean," *Appl. Sci.*, vol. 13, no. 19, 2023, doi: 10.3390/app131910786.

[14] J. A. Suarez-Peña, H. F. Lobaton-García, J. I. Rodríguez-Molano, and W. C. Rodriguez-Vazquez, *Machine Learning for Cup Coffee Quality Prediction from Green and Roasted Coffee Beans Features*, vol. 1274 CCIS, no. October. Springer International Publishing, 2020. doi: 10.1007/978-3-030-61834-6_5.

[15] N. Neighbor *et al.*, "Jurnal Darma Agung Implementasi Algoritma Principal Component Analysis (PCA) dan K -," pp. 1–10, 2023.

[16] R. P. Tanjung and D. W. Utomo, "Implementasi Principal Component Analysis (PCA) pada Pengenalan Wajah Resolusi Rendah," vol. 15, no. 01, pp. 109–116, 2024, doi: 10.35970/infotekmesin.v15i1.2148.

[17] A. Z. Putra, C. Chalvin, A. Nurhadi, A. E. Tambun, and S. Defha, "Coffee Quality Prediction with Light Gradient Boosting Machine Algorithm Through Data Science Approach," *Sinkron*, vol. 8, no. 1, pp. 563–573, 2023, doi: 10.33395/sinkron.v8i1.12169.

[18] A. Astaraja, B. S. Syamsudin, M. Diaz, and M. Dhafin, "Optimizing Coffee Ripeness Classification Using Yolov5 for Automated Detection and Sorting," vol. 4, 2025.

[19] A. E. Minarno, M. Fadhlan, Y. Munarko, and R. Chandranegara, "International Journal On Informatics Visualization journal homepage : www.joiv.org/index.php/joiv International Journal On Informatics Visualization Classification of Dermoscopic Images Using CNN-SVM." [Online]. Available: www.joiv.org/index.php/joiv

[20] X. Mu, L. He, P. Heinemann, J. Schupp, and M. Karkee, "Mask R-CNN based apple flower detection and king flower identification for precision pollination," *Smart Agric. Technol.*, vol. 4, Aug. 2023, doi: 10.1016/j.atech.2022.100151.

[21] "Non-destructive method for identification and classific.pdf."

[22] C. H. de Freitas, R. D. Coelho, J. de O. Costa, and P. C. Sentelhas, "Smart Coffee: Machine Learning Techniques for Estimating Arabica Coffee Yield," *AgriEngineering*, vol. 6, no. 4, pp. 4925–4942, 2024, doi: 10.3390/agriengineering6040281.

[23] U. P. Ganesha, "Klasifikasi Kualitas Biji Kopi Robusta Dengan Metode Naive Bayes," vol. 10, pp. 280–289, 2023.

[24] E. Aghdamifar, V. Rasooli Sharabiani, E. Taghinezhad, A. Rezvanivand Fanaei, and M. Szymanek, "Non-destructive method for identification and classification of varieties and quality of coffee beans based on soft computing models using VIS/NIR spectroscopy," *Eur. Food Res. Technol.*, vol. 249, no. 6, pp. 1599–1612, 2023, doi: 10.1007/s00217-023-04240-x.