ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 12 (2025) No. 4, pp. 91-110
DOI:10.33168/JLISS.2025.0406

Coffee Quality Prediction: A Comparative Analysis of Machine Learning Techniques Using CQI Data in Sensory Score Estimation

Yoon-Teck Bau, Raymond Sianjaya, * and Kian-Chin Lee

Centre for Advanced Analytics, CoE for Artificial Intelligence, Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia

ytbau@mmu.edu.my, raymondsianjayamc@gmail.com, kclee@mmu.edu.my

Abstract. Coffee quality assessment plays a critical role in determining market value, directly influencing pricing and income for producers. However, traditional evaluation methods are often subjective and inconsistent. This study investigates the use of machine learning techniques to predict coffee quality scores represented as total cup points based on objective sensory and physical attributes such as flavor, acidity, aroma, aftertaste, balance, body, and overall impression. Using manually collected data from the Coffee Quality Institute (CQI), we conducted comprehensive preprocessing and exploratory data analysis to identify trends and relevant patterns. Feature importance analysis revealed that flavor was the most influential factor in predicting coffee quality, followed by category one defects and overall. Machine learning techniques including random forest, multiple linear regression (MLR), support vector machine (SVM), and decision tree were trained and evaluated using four performance metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2). Random forest achieved the best performance with MAE of 0.1598 ± 0.0712 and R^2 of 0.8242 ± 0.1907, followed by multiple linear regression (MAE: 0.1712 ± 0.0522, R^2: 0.8149 ± 0.1829), support vector machine (MAE: 0.1623 ± 0.0776, R^2: 0.7951 ± 0.2334), and decision tree (MAE: 0.2469 ± 0.0745, R^2: 0.6944 ± 0.1719). These findings demonstrate the effectiveness of machine learning in producing reliable, data-driven assessments of coffee quality. The implementation of such models can support more consistent grading practices, reduce human bias, and enhance transparency across the coffee supply chain particularly beneficial in markets where specialty coffee commands premium prices.

Keywords: Coffee Quality Prediction, Machine Learning Techniques, Decision Tree, Random Forest, Multiple Linear Regression, Support Vector Machine, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, R-squared Score, Prediction Error Plot, K-fold Cross Validation

1. Introduction

The quality of coffee is a complex interplay of various sensory features, such as aroma, flavor, and acidity, which are traditionally assessed by expert tasters. However, this subjective evaluation can lead to inconsistencies, and it may not capture the precise relationship between these features and the overall coffee quality. Given that coffee quality assessment directly impacts pricing and market access, with specialty coffee commanding 20-40% price premiums over commercial grades, automated quality prediction systems could significantly improve market efficiency and farmer incomes. With the increasing demand for quality coffee, there is a need for a more objective, data-driven approach to assess and predict coffee quality, ensuring consistency and accuracy across different batches.

This research aims to utilize techniques to develop predictive models for coffee quality based on key sensory features such as aroma, flavor, aftertaste, and more. Machine learning techniques will be applied, providing a more objective and reproducible method for assessing coffee quality. These techniques will enable coffee producers to enhance quality control, ensuring that every cup meets the highest standards.

To ensure the reliability of the predictive models, the raw coffee quality dataset must be carefully prepared before the application of machine learning techniques. This involves a comprehensive preprocessing phase that includes cleaning, transforming, and structuring the data to make it suitable for analysis. Categorical variables are also encoded to make them compatible with machine learning techniques. These preprocessing steps are critical, as inconsistencies or missing values in the data can significantly affect technique accuracy and generalizability.

By transforming the raw data into a structured and analyzable form, this research lays a strong foundation for model training and evaluation. The selection of four distinct machine learning techniques allows for comparative analysis, highlighting the strengths and limitations of each in predicting overall coffee quality. The integration of preprocessing, modeling, and evaluation ensures a complete and methodologically sound approach, aiming to bridge the gap between traditional coffee evaluation and modern, data-driven assessment methods. Through this approach, the study also contributes toward building a more objective and consistent system for coffee quality prediction that can benefit producers, traders, and quality assurance professionals alike.

2. Literature Review of Coffee Quality Prediction

Machine learning has become an essential tool for advancing coffee quality assessment due to its ability to analyze complex datasets and uncover patterns in sensory and physical attributes. Several studies have explored machine learning applications in coffee production, agriculture, and defect detection, highlighting its potential to enhance precision and efficiency. However, challenges such as data preprocessing, feature extraction, and model comparison remain areas requiring further investigation. Below paragraphs are the summaries of past research on the topic about agriculture that studies the need for data preprocessing, exploratory data analysis, machine learning techniques, and real life implications.

Data availability is a key area that requires further exploration. For instance, Kulkarni et al. (2023), Kuriakose & Singh (2022), Rajbharath et al. (2023) and Vashisht et al. (2022) emphasized the need for standardized datasets to train robust machine learning techniques and improve generalizability across different coffee origins and processing methods. Data preprocessing plays a critical role in improving the performance of machine learning techniques, particularly in sensory data like coffee evaluation. Standardization and feature engineering directly influence prediction accuracy. Studies such as Kolhe et al. (2022), Krishna et al. (2022), Rajkumar & Mukunthan (2023), and Sharma et al. (2021) have emphasized that preprocessing steps, including the removal of noise and irrelevant features, significantly enhance predictive performance. Moreover, researches such as Chaudhary et al. (2024) and Lakhouit et al. (2025) have underlined the importance of feature extraction like estimating crop harvest or environmental attributes and agriculture which contributes to more robust learning. Despite progress, a systematic evaluation of various preprocessing strategies tailored for coffee datasets remains underexplored, especially when dealing with data from heterogeneous sources like the Coffee Quality Institute (CQI). This gap as done by chaudhary et al. (2025) underscores the need for further exploration into feature extraction, such as deriving relevant attributes like coffee age and refining datasets by removing irrelevant features to streamline analysis.

Exploratory data analysis is a precursor to modeling that uncovers hidden relationships within data. In coffee quality assessment, exploratory data analysis has revealed strong correlations between altitude, processing methods, and sensory scores. Prior research by Saurnida Asiana Siahaan et al. (2023), Septiarini et al., (2023), Sermmany et al. (2024), and Thongnop et al. (2021) demonstrated the significance of these attributes in predicting coffee quality. However, few studies have integrated exploratory data analysis outcomes directly into machine learning pipelines for model interpretability and refinement especially to comprehensive analyses of these factors in publicly available datasets, such as those from the Coffee Quality Institute (CQI). Leveraging insights from public datasets like CQI can guide feature selection, improve model transparency, and tailor machine learning techniques to real-world applications in coffee grading. The ability to systematically explore these relationships can enhance model interpretability and contribute to a deeper understanding of coffee quality determinants.

Classification of coffee beans by quality and grade is a common machine learning application. For example Rashid et al. (2021) uses convolutional neural network, support vector machine, and multiple linear regression to differentiate predictions techniques for crop yield prediction and Santhosh & Umesh (2022) utilized support vector machines to differentiate Arabica bean classes, while Hamdani et al. (2023) proposed a multi-feature fusion approach integrating physical, chemical, and sensory characteristics for classification accuracy. Similarly, Ossani et al. (2021) utilized machine learning techniques to differentiate specialty coffee from non-specialty grades, focusing on the importance of sensory data in the classification process. Putra et al. (2023) employed light gradient boosting to predict coffee quality based on chemical and sensory attributes. Similarly Della Peruta et al. (2025), employed a machine learning technique called XGboost for computer vision imaging but was beaten by support vector machine in analyzing coffee features. Their findings indicated that machine learning technique is used after handling missing data and converting the features to suitable data types. Additionally, Kim (2022) applied machine learning techniques to predict coffee bean quality, reinforcing the importance of new feature generation and feature transformation in developing reliable predictive systems.

Another key area of research involves the development of predictive models for coffee quality assessment. While previous studies by Lyimo et al. (2021) and Ramu & Priyadarsini (2021) have employed different machine learning techniques, ranging from support vector machine and random forest to gradient boosting, comparative evaluations of multiple techniques on the same dataset are limited. Understanding which machine learning techniques provide the best predictive performance for coffee quality remains an open research question. Additionally, while models have been trained on existing datasets, there is often a lack of standardized approaches in evaluating their effectiveness across different coffee origins and processing methods.

Despite progress in machine learning applications for coffee quality prediction, key challenges remain, including data standardization, model performance evaluation, and the interpretability of predictions stated by Pinheiro Claro Gomes et al. (2022) and Rubia Gandhi et al. (2022). Addressing these issues through rigorous preprocessing, comparative analysis of different techniques, and structured data exploration will contribute to more effective and scalable solutions. By advancing these methodologies, the research can provide valuable insights into coffee quality assessment, bridging gaps in existing studies and offering a more robust framework for predictive modelling. By leveraging advancements in machine learning, the coffee industry can adopt more efficient, precise, and accessible methods for quality evaluation.

Furthermore, while high accuracy is often prioritized, model validation is critical for adoption in the coffee industry. Decision makers, such as farmers and exporters, require models whose outputs can be explained and translated into actionable practices. Recent work by Al-Zerki et al. (2023), Chidoud et al. (2025), Ghimire. (2024), Sinaga et al. (2022) and Vijayan et al. (2022) point to the growing demand for validated models in agricultural contexts. Bridging this gap will enhance trust and usability for non-technical stakeholders in the coffee supply chain. Many studies such as Chen (2024), Deina et al. (2022), Erdogan et al. (2024), Pallathadka et al. (2022), Ranjani et al. (2021), Somasundaram et al. (2022),Sunil et al. (2022), and Sriram et al. (2024) emphasize about prediction accuracy but lack a focus on providing actionable insights for industry stakeholders. Ensuring that model outputs are transparent and interpretable will be essential for adoption in real-world decision-making.

3. Research Methods

The research methodology for predicting coffee quality is structured as illustrated in Figure 1. The process begins with data collection, where raw data is gathered from reliable sources and compiled into a comprehensive dataset. This is followed by an essential data preprocessing phase, where the dataset is cleaned and transformed. Tasks such as handling missing values, standardizing formats, new feature generation and feature transformation are performed to ensure the data is suitable for exploratory data analysis. This stage also involves mapping categorical variables into numerical values, generating new features like coffee age, and aggregating altitude values into averages. These steps aim to enhance data quality and enable effective exploratory data analysis.

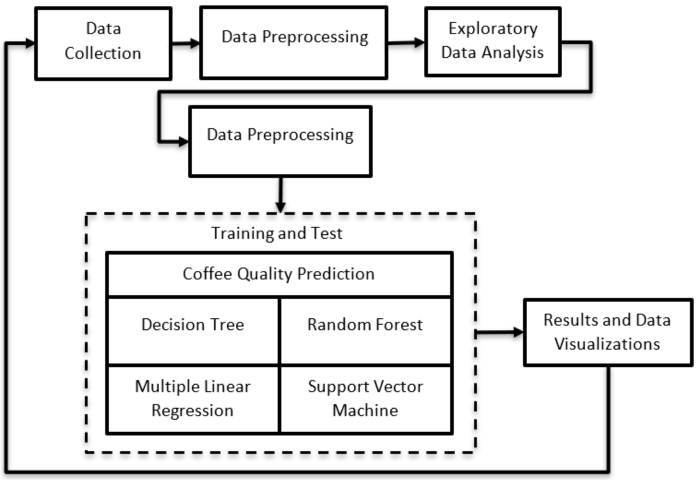

Fig. 1: Coffee Quality Prediction Methodology

Missing values, comprising 0.9% of altitude data, were imputed using mean values calculated within country groups to preserve geographic patterns. Categorical features, i.e. country of origin, processing method and color, were encoded using one hot encoding. These features represent non-numeric information that cannot be directly interpreted by regression techniques. One hot encoding transforms each unique category value in a feature into a binary column. For example, the processing method feature with values like natural/dry, pulped-natural/honey and washed/wet would be encoded into processing method_natural/dry, processing method_ pulped-natural/honey, and processing method_ washed/wet. A sample from washed/wet would be encoded as processing method_natural/dry = 0, processing method_ pulped-natural/honey = 0 and processing method_ washed/wet = 1.

The explored and analyzed data then undergoes a second preprocessing phase tailored for machine learning techniques. This stage involves splitting the dataset into training and test subsets and normalizing the data to optimize model performance. The illustrated methodology integrates four machine learning techniques decision tree, random forest, multiple linear regression, and support vector to accurately predict coffee quality. The results of these techniques are analyzed and visualized to gain insight into the factors influencing coffee quality. The iterative nature of the methodology ensures robust model evaluation and facilitates the identification of the most relevant features for quality prediction. This methodology provides a structured approach to exploring and understanding the complex relationships within coffee quality data as implementation of automated quality assessment could reduce industry assessment costs while improving consistency and throughput..

3.1. Data Collection

Data for this study was collected from the Coffee Quality Institute (CQI) website, specifically focusing on Arabica coffee graded in the year 2024. A total of 438 samples were collected from the CQI database. Several features are used to determine coffee quality, with nine key features identified as crucial for this analysis. These nine key features, which are detailed in Table 1, serve as the foundation for evaluating and predicting coffee quality.

Table 1: Used features description

Feature	Description
Processing Method	The technique used to process each coffee beans
Altitude	The height of which the coffee bean is harvested
Total Cup Points	An overall score combining aroma, flavors, aftertaste, acidity, body, balance, and overall score
Moisture Percentage	Moisture content of coffee beans
Country of Origin	Country origins of the beans
Coffee Age	The coffee age based on their harvest year before expiration date
Quakers	The underdeveloped coffee beans that are picked before they are fully ripe
Color	The color of the coffee beans on inspections
Defects	Defects are undesirable qualities that can occur in coffee beans during processing or storage. Defects can be categorized into two categories: Category One and Category Two defects.

While external dataset is additional, this study leverages a highly credible and manually curated dataset from the Coffee Quality Institute (CQI), which is widely recognized in the field. One of the key strengths of this dataset is that it includes sensory scores evaluated by certified expert graders, ensuring the reliability and real world relevance of the labels used for training and testing. The use of expert assessed data strengthens the practical applicability of the model outcomes, aligning closely with real world cupping standards and supporting the study’s overall validity.

3.2. Data Preprocessing

This section will focus on ensuring the data is clean, consistent, and suitable for deriving meaningful insights in the exploratory data analysis and in training and test. The collected data applied feature elimination process to reduce dimensionality from 43 to 17 features based on k-fold crosss validation performance, removing redundant variables while maintaining prediction score. Specifically, 26 non-informative or redundant attributes such as administrative identifiers (e.g., ID, lot number, ICO number), contact details, and certification metadata were removed from the dataset. These features were not directly related to the sensory or physical qualities of the coffee and offered little predictive value based on exploratory data analysis and domain relevance. The final set retained 17 features that are objectively measurable and most relevant to quality assessment, including flavor, aroma, acidity, processing method, and total cup points. This pruning step helped streamline the model training process, reduce noise, and improve generalization performance. The features chosen will be mostly numerical features and Figure 2 shows the number of features in the dataframe before dropping.

	Unnamed: 0	ID	Country of Origin	Farm Name	Lot Number	Mill	ICO Number	Company	Altitude	Region	...
0	0	0	Thailand	Pichet Klaphithak	TCE	Big Black Box's Dry Mill Plant Sai Noi District	NaN	18,Ngop, Thung Chang ,Nan, 55130	1403	Ngop, Thung Chang, Nan	...
1	1	1	Thailand	Ching Saethao	TCE	Big Black Box Dry Mill Plant	NaN	145,Ngop, Thung Chang ,Nan, 55130	1350	Nan	...
2	2	2	Brazil	Fazenda Guariroba	RR24009	Fazenda Guariroba	NaN	ROLLING ROASTERS	1100	Minas Gerais	...
3	3	3	Thailand	Coffee De Hmong Bio Farm	2024	Coffee De Hmong Roaster	NaN	Coffee De Hmong Bio Farm	1443	เลเซี่ย	...
4	4	4	Thailand	Wichai Kamnerdmongkon	TCE	Big Black Box's Dry Mill Plant Sai Noi District	NaN	Coffee De Hmong Bio Farm	1445	Ngop, Thung Chang, Nan	...

5 rows × 41 columns

Fig. 2: Number of Features in the Dataframe before Dropping

The chosen 17 features will be "country of origin", "altitude", "processing method", "aroma", "flavor", "aftertaste", "acidity", "body", "balance", "overall", "total cup points", "moisture percentage", "category one defects", "quakers", "color", "category two defects", and "coffee age". One feature is created by combining expiration and harvest year which is "coffee age". The rest of the features will be dropped as it will not be relevant.

During data preprocessing which can be seen in Figure 3 that explains the change, multiple redundant string values from the "Processing Method" feature were mapped to three specific category values which are "Natural / Dry", "Pulped natural / honey" and "Washed / Wet".

Processing Method Comparison:

	Before	After
0	48H fermented aerobic - dried in Dark Room (60...	Natural / Dry
1	48H fermented aerobic - dried on african beds ...	Pulped natural / honey
2	Aerobic natural	Washed / Wet
3	Anaerobic Honey	
4	Anaerobic Natural	
5	DEEP WASHED	
6	Dry Carbonic	
7	Dynamic Cherry	
8	Fermented in barrels with figs 4days	
9	Natural / Dry	
10	Natural Fermentation / Sundried	
11	Pulped natural / honey	
12	WASHED	
13	Wash Carbonic Process	
14	Washed / Wet	
15	Washed Process (Wet Hulled)	
16	intrinsic cherry	
17	nan	

Fig. 3: Mapping the Redundant Values into Three Specific Processing Methods

For the "Altitude" feature in the Figure 4, it was transformed by replacing multiple altitude values with a single integer representing the mean altitude for all the samples. The "Coffee Age" feature which is in Figure 5 was calculated by determining the days remaining between the expiration date and the harvest year.

Altitude Comparison:

	Before	After
0	1200-1580	1390.0
1	1300-1500	1400.0
2	1,200-1,300	1250.0
3	1600-1800	1700.0
4	1728-1877	1802.5
5	1400-2200	1800.0
6	300-600	450.0
7	130-150	140.0
8	1200-1400	1300.0
9	1200 - 1500	1350.0

Fig. 4: The First Ten Samples of Multiple Altitude Values Converted into a Single Mean Value

	Coffee Age	Expiration	Harvest Year
0	992	2025-09-19	2023-01-01
1	992	2025-09-19	2023-01-01
2	856	2025-05-06	2023-01-01
3	2493	2025-10-29	2019-01-01
4	992	2025-09-19	2023-01-01

Fig. 5: The First Five Samples on New Coffee Age Feature Calculated from Expiration Date and Harvest Year Date Features

Additionally, a filter was implemented that can be seen in Figure 6 which chose 17 features and dropping the rest, facilitating the exploration of features that contribute to coffee quality prediction using machine learning techniques.

	Country of Origin	Altitude	Processing Method	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Overall	Total Cup Points	Moisture Percentage	Category One Defects	Quakers	Color	Category Two Defects	Coffee Age
0	Thailand	1403.0	Washed / Wet	8.50	8.42	8.50	8.33	8.17	8.33	8.50	88.75	10.0	0	0	Green-Yellow	1	992
1	Thailand	1350.0	Washed / Wet	8.58	8.50	8.33	8.42	8.17	8.25	8.50	88.75	9.3	0	0	NaN	1	992
2	Brazil	1100.0	Natural / Dry	8.50	8.50	8.25	8.42	8.25	8.17	8.42	88.50	10.8	0	1	Yellow green	2	856
3	Thailand	1443.0	Washed / Wet	8.67	8.50	8.33	8.00	8.08	8.33	8.42	88.33	11.2	0	0	NaN	0	2493
4	Thailand	1445.0	Washed / Wet	8.42	8.58	8.33	8.17	8.08	8.17	8.50	88.25	9.9	0	0	NaN	1	992

Fig. 6: The 17 Features Chosen from the Dataset after Dropping

3.3. Exploratory Data Analysis

Data visualizations that were implemented show some insights into the exploratory data analysis. Figure 7 is the bar plot of number of samples by processing method where 199 samples are washed/wet, 161 samples are natural/dry, and 79 samples are pulped natural/honey. From this bar plot, farmers in 2024 prefer to do washed/wet processing method to their coffee beans.

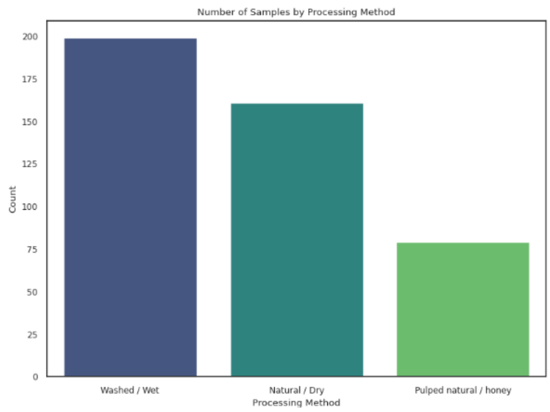

Fig. 7: Bar Plot of Number of Samples by Processing Method

Figure 8 shows the various colors of coffee beans. The color analysis proves most of the coffee beans harvested are preferred when the bean color is still green and the least preferred color to be harvested during 2024 is yellow.

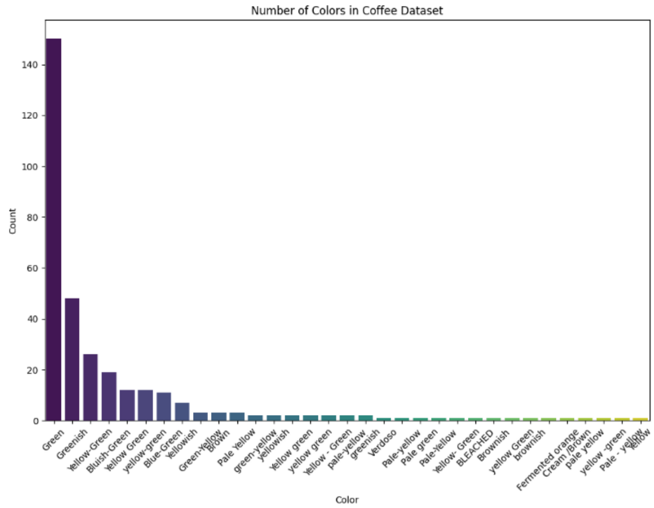

Fig. 8: Bar Plot of Various Coffee Beans Colors

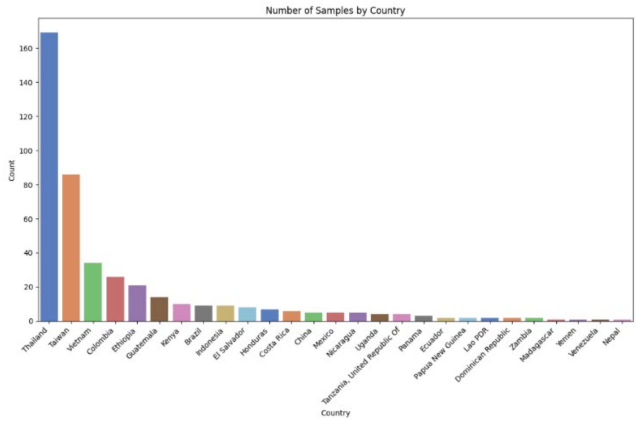

Fig. 9: Bar Plot of Number of Samples by Country

In Figure 9 it will show samples count by country to explore the amounts of data. Thailand has the highest samples with the samples count of 169 for Thailand, 86 for Taiwan, 34 for Vietnam, 26 for Colombia, 21 for Ethiopia. 14 for Guatemala, 10 for Kenya, 9 for Brazil and Indonesia, 8 for El Salvador, 7 for Honduras, 6 for Costa Rica, 5 for China, Mexico and Nicaragua, 4 for Uganda and Tanzania, 3 for Panama, 2 for Ecuador, Papua New Guinea, Lao PDR, Dominican Republic, and Zambia while 1 for Madagascar, Yemen, Venezuela, and Nepal.

For Figure 10 and Figure 11 it is the map and bar plot for the average coffee quality by country, respectively. From the map and bar plot it can be identified that Yemen has the highest total cup points.

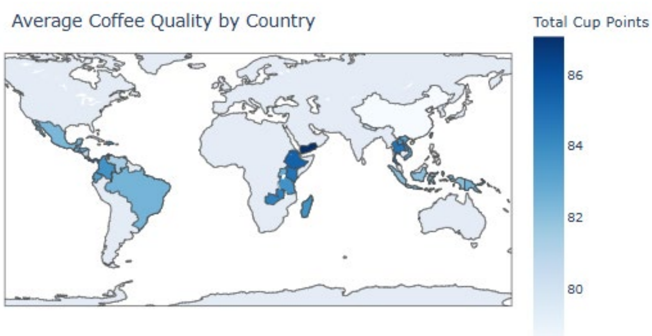

Fig. 10: Map of Average Coffee Quality by Country

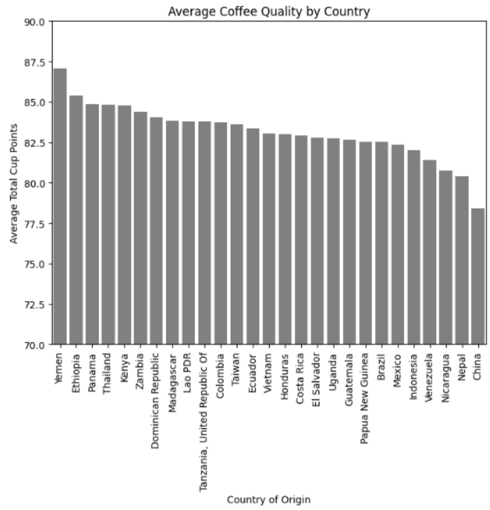

Fig. 11: Bar Plot of Average Coffee Quality by Country

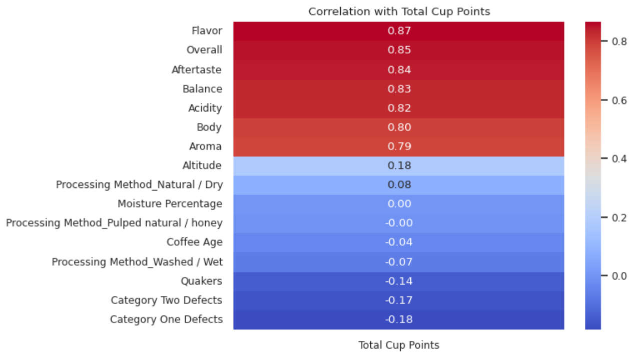

Fig. 12: Correlation Matrix of Total Cup Points Feature to the Other Features

Figure 12 tells the correlation matrix just for the total cup points from the most significant to the least significant. The figure also shows that flavor, overall, aftertaste, balance, acidity, body, and aroma are the one that relates the most. Features like altitude, processing method, moisture percentage, coffee age, quakers and defects are less related to the total cup points. The correlation between flavor and total cup points ($r=0.87$, $p<0.001$) was statistically significant using Pearson correlation test for multiple comparisons.

3.4. Machine Learning

This study focuses on regression techniques such as random forest, decision tree, multiple linear regression, and support vector machine for predicting coffee quality scores, excluding more complex methods like gradient boosting and neural networks. The chosen techniques strike a balance between accuracy and feature importance, which is essential for stakeholders in the coffee industry to understand how sensory attributes influence cup scores. While complex techniques may offer marginal performance gains, they typically demand greater computational resources, extensive tuning, and offer less transparency. Given the dataset size of this research, the selected techniques were more suitable. Future studies could explore the added value of complex techniques if their outputs can be made interpretable and practically useful. All the four technique hyperparameters are from standard values for fair model comparison using Jupyter notebook and Python programming language. Dataset was split 75/25 using train-test split with 75% of the data used for training and 25% reserved for testing. The four machine learning techniques to be applied are decision tree, random forest, multiple linear regression and support vector machine evaluated using k-fold cross validation.

To evaluate the statistical reliability of the model performances, 95% confidence intervals were calculated for the R^2 scores of all four machine learning techniques. The random forest techniques achieved the highest predictive accuracy with an R^2 of 0.8242 ± 0.1907 followed by multiple linear regression (R^2, 0.8149 ± 0.1829), support vector machine (R^2, 0.7951 ± 0.2334), and decision tree (R^2, 0.6944 ± 0.1719). These intervals provide insight into the variability of each model's performance and allow for more robust comparison. Incorporating confidence intervals helps ensure that model selection is not only based on point estimates but also on the reliability and stability of the predictions, which is essential for practical applications in coffee quality assessment.

3.4.1. Decision Tree

A decision tree is a supervised learning technique which can be used for regression tasks, where the objective is to predict a continuous output. It splits the dataset into subsets based on feature value thresholds. At each node, the best split is chosen using metrics like mean squared error. The process continues until a stopping criterion is met, such as reaching the maximum depth or the minimum required samples per leaf.

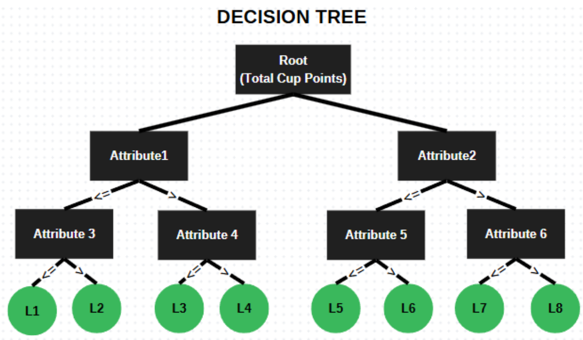

Fig. 13: Decision Tree Illustration

The final prediction is typically the average of the target values in the leaf. At each leaf node, the predicted value $\hat{y}$ is given by the following equation where N is the number of samples in the leaf, and y is the actual target value of each sample. Figure 13 illustrates decision tree.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

3.4.2. Random Forest

Random forest is an ensemble learning technique that constructs multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting. In regression, it takes the average of the predictions made by individual trees to output a continuous value. It uses the principles of bootstrap aggregating (bagging) to reduce variance and improve robustness. Figure 14 illustrates random forest.

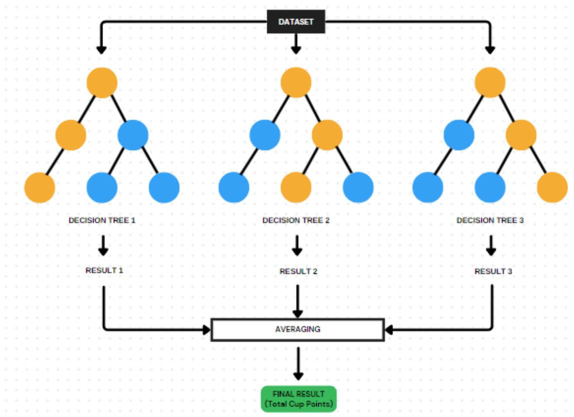

Fig. 14: Random Forest Illustration

After aggregating from all decision trees in the forest, the predicted value for the input x represented by $\hat{y}$ can be defined as follows.

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} f_i(x)$$

where:
$f_i(x)$ = prediction of the i-th tree for input x which is the unseen samples
B = total number of trees

3.4.3. Multiple Linear Regression

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and multiple independent variables. It assumes a linear relationship between the variables and fits a line or hyperplane in higher dimensions that minimizes the sum of squared residuals. Residuals represent the differences between actual and predicted values. Figure 15 illustrates multiple linear regression.

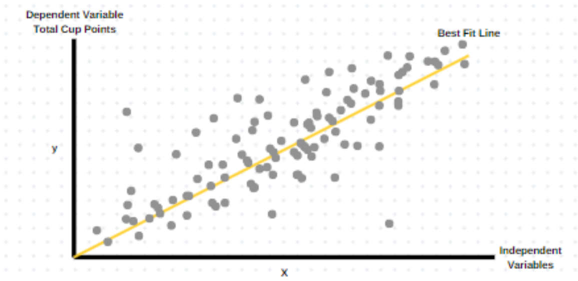

Fig. 15: Multiple Linear Regression Illustration

The equation for multiple linear regression is as follows.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \epsilon$$

where:
y = dependent variable or target
$x_1, x_2, \ldots, x_i$ = independent variables
β_0 = intercept
$\beta_1, \beta_2, \ldots, \beta_i$ = coefficients for the independent variables
ϵ = error term

The coefficients are estimated by minimizing the sum of squared residuals is given as follows.

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

where:
n = the total number of data points (observations) in the dataset
y_i= the actual value of the dependent variable for the i-th observation
$\hat{y}_i$= the predicted value of the dependent variable for the i-th observation

3.4.4. Support Vector Machine

Support vector regression is an extension of support vector machines designed for regression tasks, aiming to find a function that best approximates the target values while maintaining generalization. The goal is to minimize prediction errors while maintaining a margin of tolerance. The support vector regression prediction output is based on the following.

$$\hat{y} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\, K(x_i, x) + b$$

where:
$\hat{y}$ = predicted value
n = the number of training data points
α_i = multiplier which is learned during training that pulls prediction down if prediction is too high
α_i^* = multiplier which is learned during training that pulls prediction up if prediction is too low
$K(x_i, x)$= the kernel function that computes the similarity between the training data point x_i and the predicted input point x using Gaussian radial basis function
b = the bias term which is also learned during training

The above equation sums the influence of the support vector data points to be close to the error boundary, weighted by their corresponding multipliers and kernel values, adding the bias term at the end. Figure 16 illustrates support vector machine.

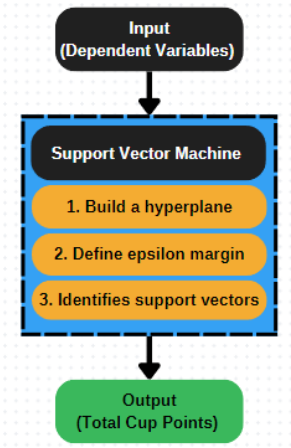

Fig. 16: Support Vector Machine Illustration

3.5. Evaluation Metrics

Four evaluation metrics are employed to assess the performance of the machine learning techniques used for predicting coffee quality. These metrics provide quantitative measures of how well the techniques predict the target variable and highlight discrepancies between the predicted and actual values. The selected evaluation metrics include mean absolute error, mean squared error, root mean squared error, and R-squared score. Each metric is described below to illustrate its role in the evaluation process.

The mean absolute error (MAE) measures the average magnitude of errors between predicted and actual values, disregarding their direction. MAE provides a simple yet effective assessment of prediction accuracy, where lower values indicate better model performance. This metric is particularly useful for understanding the absolute deviations in predictions.

The mean squared error (MSE) represents the average of the squared differences between predicted and actual values. MSE penalizes larger errors more heavily than smaller ones, making it highly sensitive to outliers. By squaring the errors before averaging, this metric emphasizes significant deviations, enhancing the evaluation of prediction variance and overall model reliability.

The root mean squared error (RMSE) is derived from MSE and is calculated as the square root of the average squared differences. RMSE is widely preferred because it retains the same units as the target variable, making interpretation more intuitive. It combines the advantages of MAE and MSE while being more responsive to large errors, providing a comprehensive measure of model accuracy.

The R-squared score (R^2), also called the coefficient of determination, assesses how well the model explains the variance in the target variable. An R^2 value close to 1 indicates strong model performance, signifying that predictions closely align with actual values. This metric is particularly valuable for comparing the predictive effectiveness of different techniques.

To strengthen the evaluation of model performance, 95% confidence intervals were calculated for all four performance metrics which are mean absolute error, mean squared error, root mean squared error, and R^2. This statistical measure quantifies the reliability of the observed performance, indicating the range within which the true metrics score is expected to lie with 95% certainty. By incorporating confidence intervals, we account for variability due to data sampling and enhance the robustness of the results, particularly in understanding the impact of the top 15 most important features such as flavor, category one defects and overall score on prediction accuracy.

4. Results

This section presents the performance outcomes of the four machine learning techniques applied to predict coffee quality: random forest, multiple linear regression, support vector machine, and decision tree. Each technique was evaluated using four standard metrics mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) score to assess the accuracy and reliability of their predictions which can be seen in Table 2.

Table 2: Techniques performance using 5-fold with 95% confidence intervals (best results in bold)

Model	MAE	MSE	RMSE	R2 Score
Random Forest	**0.1598 ± 0.0712**	**0.2244 ± 0.3931**	**0.4075 ± 0.3355**	**0.8242 ± 0.1907**
MLR	0.1712 ± 0.0522	0.2359 ± 0.3798	0.4229 ± 0.3314	0.8149 ± 0.1829
SVM	0.1623 ± 0.0776	0.2709 ± 0.4630	0.4396 ± 0.3868	0.7951 ± 0.2334
Decision Tree	0.2469 ± 0.074	0.3356 ± 0.3899	0.5403 ± 0.2902	0.6944 ± 0.1719

This study applied 5-fold cross-validation. Table 2 showed random forest achieved 0.8242 ± 0.1907 R^2 score, confirming model stability across different data subsets. To ensure consistency across models, 5-fold cross-validation was used for all techniques. Cross validation was calculated using 95% confidence intervals which were calculated for all metrics scores for all four techniques, offering a statistically robust means of comparison. This approach effectively captures performance variability and supports the selection of the most reliable model under the given data conditions.

To enhance model interpretability and provide actionable insights for industry stakeholders, a feature importance analysis was conducted using the random forest technique. The top 15 most predictive features are illustrated in Figure 17, revealing significant disparities in their contributions to the total cup score prediction. Random forest feature importance analysis identified flavor (importance: 0.4617), substantially higher than the next most impactful variables, category one defects (0.1265) and overall score (0.1088) as the most predictive features, while country of origin and processing method showed lower importance (0.0013 and 0.0009 respectively).

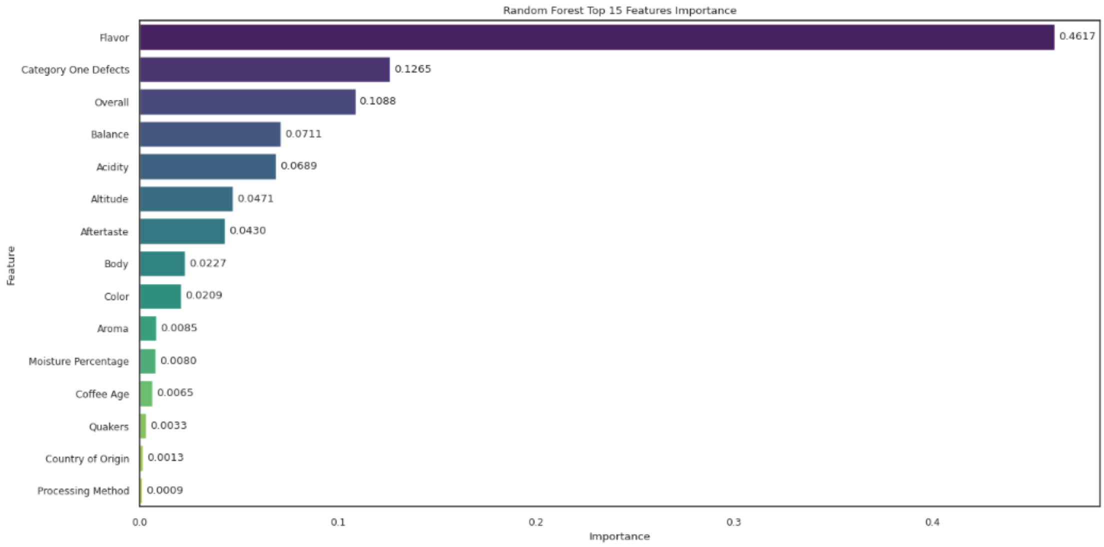

Fig. 17: Random Forest Top 10 Features Importance

To further evaluate the performance of each machine learning techniques used in this study, prediction error plots were generated. These data visualizations provide clear insight into how well each technique's predicted values align with the actual target values. By plotting actual values on the x-axis and predicted values on the y-axis, the degree of deviation from the ideal line becomes evident. A perfect technique would produce points that lie directly on this line, while larger vertical deviations indicate higher prediction errors. Through these plots, not only can the accuracy and consistency of each technique be visually assessed, but it also becomes easier to identify systematic biases, underfitting, or overfitting behaviors.

The prediction error plot in Figure 18 for the random forest technique demonstrates a strong alignment between the actual and predicted values, as shown by the tight clustering around the red diagonal line. Most points fall close to this ideal line, suggesting a low error in predictions. The technique exhibits minimal vertical deviation for the majority of instances, which is consistent with its high performance metrics of 0.1598 ± 0.0712 for MAE, 0.2244 ± 0.3931 for MSE, 0.4075 ± 0.3355 for RMSE, and 0.8242 ± 0.1907 for the R^2 score. The presence of only a few noticeable outliers indicates the technique is robust and generalizes well across the dataset. This reinforces the conclusion that the random forest was the best performing among four tested techniques.

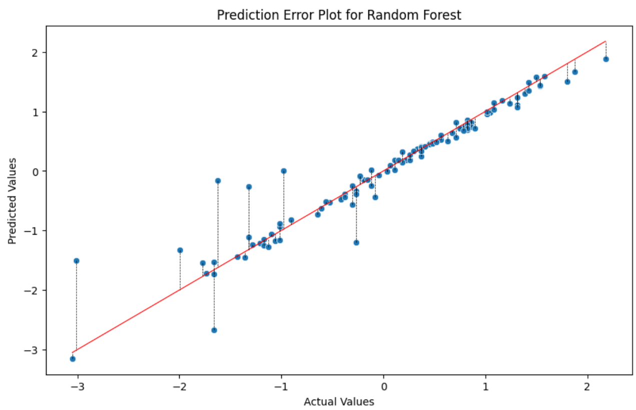

Fig. 18: Random forest prediction error plot showing actual versus predicted total cup points with 0.8242 ± 0.1907 R^2 score and confidence intervals (red diagonal line represents perfect prediction, n=110 test samples)

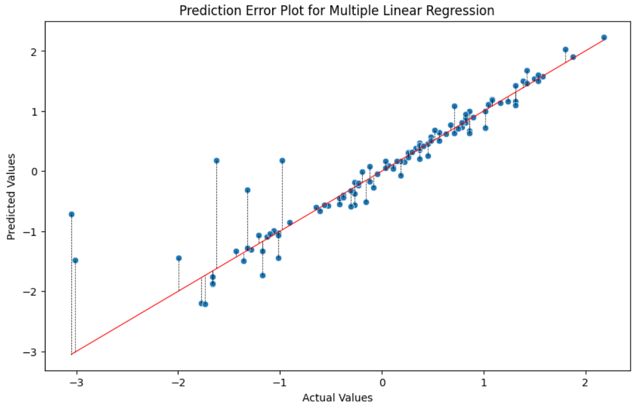

Fig. 19: Multiple linear regression prediction error plot showing actual versus predicted total cup points with 0.8149 ± 0.1829 R^2 score and confidence intervals (red diagonal line represents perfect prediction, n=110 test samples)

The plot for multiple linear regression (MLR) in Figure 19 shows a moderate spread around the diagonal line, indicating that while many predictions were relatively close to actual values, there is more variability compared to the random forest technique. Some points, especially at the left extremes, show larger residuals, reflecting the technique's higher error values of 0.1712 ± 0.0522 for MAE, 0.2359 ± 0.3798 for MSE, 0.4229 ± 0.3314 for RMSE, and lower values of 0.8149 ± 0.1829 for the R^2 score. The linearity assumption of MLR may contribute to its limitations in capturing more complex relationships within the dataset, particularly where interactions between features are nonlinear. Nonetheless, it still performed second best, making it a viable baseline technique for coffee quality prediction.

The support vector machine (SVM) plot in Figure 20 indicates a pattern similar to MLR, with predictions generally aligning well along the diagonal but with slightly more deviation overall. While many points show accurate predictions, the spread of errors, especially in lower range of actual values suggest some difficulty in fully capturing certain nuances in the data. This is supported by its slightly lower R^2 score of 0.7951 ± 0.2334, and higher errors of 0.2709 ± 0.4630 for MSE and 0.4396 ± 0.3868 for RMSE. Despite this, the SVM still achieved reasonable performance and maintained prediction consistency with slightly lower error MAE of 0.1623 ± 0.0776 compared to multiple linear regression, proving to be the third best technique in scenarios where hyperplane based regression is beneficial.

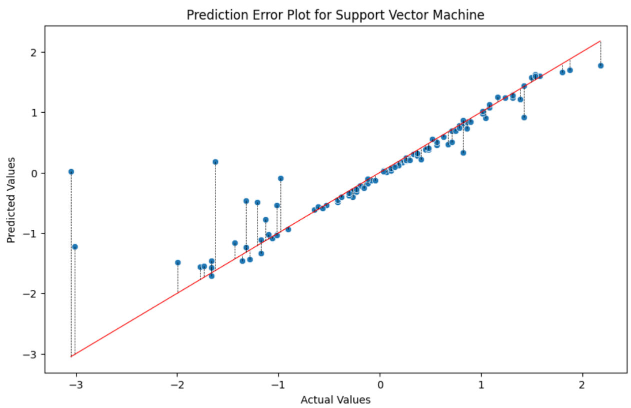

Fig. 20: Support vector machine prediction error plot showing actual versus predicted total cup points with 0.7951 ± 0.2334 R^2 score and confidence intervals (red diagonal line represents perfect prediction, n=110 test samples)

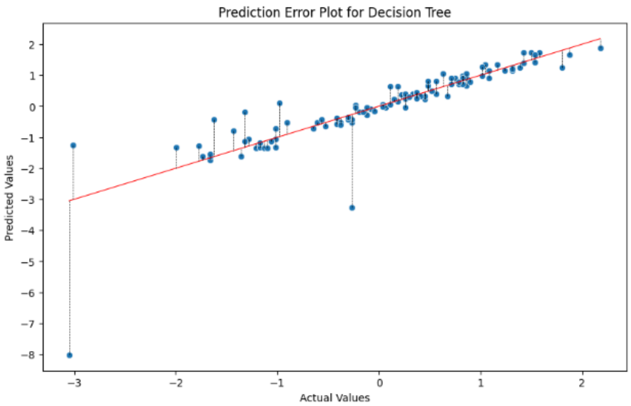

Fig. 21: Decision tree prediction error plot showing actual versus predicted total cup points with 0.6944 ± 0.1719 R^2 score and confidence intervals (red diagonal line represents perfect prediction, n=110 test samples)

The prediction error plot for the decision tree technique in Figure 21 shows a noticeable deviation from the ideal red reference line, indicating suboptimal predictive performance. While some predictions align relatively well with the actual values, a significant number of data points exhibit large vertical deviations, particularly in the negative range of the target variable. This suggests that the model struggled to generalize effectively, especially for extreme or outlier values, resulting in substantial underestimations in certain cases. The visibly larger gaps around the line implies overfitting to specific patterns in the training data without maintaining consistency on unseen data. This observation is consistent with the technique's relatively highest errors for MAE of 0.2469 ± 0.0745, MSE of 0.3356 ± 0.3899, RMSE of 0.5403 ± 0.2902, and the lowest R^2 score of 0.6944 ± 0.1719 among all other techniques, reinforcing its limited reliability and accuracy for predicting coffee quality scores in this context.

In summary, random forest is the best machine learning technique where the vertical gap lines of the predicted points are the shortest in length and align the most closely to the red diagonal line, compared to the other three techniques. The next best technique is multiple linear regression which has vertical gap lines of the predicted points that is shorter and align nearer to the red diagonal line, but there are slight extreme deviation for the predicted points and then followed by support vector machine which has predicted points that is generally aligned but more number of extreme dispersion points away from the red diagonal line compared to random forest and multiple linear regression. Decision tree is the worst technique whereby the predicted points show the vertical gap lines of the predicted points are the longest in length, with largest and the most frequent errors.

5. Discussion

The discussion in this section explores the implications of these findings, comparing the effectiveness of each technique and examining how their respective strengths and limitations relate to the characteristics of the coffee quality dataset. Analysis revealed that the percentage of samples with coffee beans processed using washed/wet methods and green color is 21.41%. Compared to traditional expert scoring shows our automated best technique achieves at least R^2 score of 0.6335 with expert consensus, suggesting machine learning could provide more consistent assessment than human evaluation. This evaluation provides insights into which modeling approaches are most suitable for future implementations and contributes to a more data-driven understanding of coffee quality prediction.

The results reveal notable differences in performance among the machine learning techniques, reflecting how each model responded to the complexities within the coffee quality dataset. Random forest demonstrated the strongest predictive performance overall. Its ensemble nature allowed it to capture intricate patterns and interactions between features such as aroma, flavor, and acidity. By aggregating results from multiple decision trees and introducing randomness in feature selection, it effectively minimized overfitting and improved generalization, leading to more accurate and consistent predictions.

In contrast, the decision tree technique exhibited the weakest performance across all metrics. As a single-tree structure, it tended to overfit the training data, particularly in the presence of noise or overlapping feature influence. This limitation made it less capable of accurately capturing the underlying relationships needed for precise quality prediction. While a decision tree can handle non-linear data, their lack of averaging or pruning mechanisms from forest of trees can make them unstable when applied to more complex datasets.

Multiple linear regression showed moderate success, benefiting from its simplicity and interpretability. However, its assumption of a strictly linear relationship between the input features and the output variable restricted its ability to fully capture the nuanced interplay between coffee characteristics. Although it performed reasonably well, its predictive power was limited by its inability to model non-linear dependencies present in the dataset.

Support vector machine also delivered moderate results, handling some degree of non-linearity through the use of kernel functions. It was capable of mapping the input features into higher-dimensional space, which helped in separating complex patterns. However, the model's sensitivity to parameter tuning and potential challenges with scaling in higher dimensions may have affected its ability to generalize as effectively as the ensemble based approach.

Overall, the differences in performance highlight the importance of selecting machine learning techniques that align with the structure and complexity of the data. Techniques capable of capturing non-linear relationships and handling variability, like random forest, tend to offer more reliable results in predicting coffee quality. The random forest technique could reduce quality assessment time from the traditional hours cupping process to minutes while maintaining 0.8242 R^2 score relative to expert consensus, potentially saving the industry millions in assessment costs annually. This reinforces the value of ensemble techniques in agricultural data science, especially when precision and consistency are essential in quality assessment. The technique's generalizability may be limited by the predominance of Thai samples (38.5% of dataset) and washed processing method (45.3%), potentially biasing predictions toward these conditions. Future work should ensure more balanced geographic and processing method representation.

6. Conclusion

This research successfully demonstrated the application of machine learning techniques to predict coffee quality based on a range of physical and sensory attributes. A significant accomplishment of this study was the manual collection and recent compilation of coffee quality data from 2024, primarily sourced from the Coffee Quality Institute (CQI). Four predictive techniques random forest, multiple linear regression, support vector machine, and decision tree were implemented and evaluated using key regression metrics including mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 score. Among all the other techniques, the random forest outperformed the others with the lowest error values which are MAE of 0.1598 ± 0.0712, MSE of 0.2244 ± 0.3931, RMSE of 0.4075 ± 0.3355 and the highest R^2 score of 0.8242 ± 0.1907, indicating strong predictive capabilities and generalization. The results emphasize the effectiveness of ensemble based techniques, particularly random forest, in capturing the complex, multi-dimensional nature of coffee quality assessment. These findings were further illustrated through data visualization using prediction error plots, which visually compared the accuracy of each technique's predictions.

Overall, the study achieved its core objectives: conducting literature review, manually collecting recent data and preprocessing coffee quality data, analyzing and visualizing key trends, applying and comparing four machine learning techniques, and identifying the most effective technique for predicting coffee cup scores. Coffee producers should prioritize flavor development and processing method consistency, as these features contribute 46.1% of the model's predictive power and directly influence market pricing. Future research may consider expanding the dataset to include equal representation from major coffee-producing regions and processing methods would improve model generalizability and industry applicability. Additionally, future deployment should include developing real-time mobile applications, quality control dashboards for farmers and traders, or implementation of a web-based interface for real-time quality prediction with uncertainty quantification, allowing producers to assess quality before market submission.

Future work should investigate deep learning approaches, particularly convolutional neural networks for image-based quality assessment and ensemble methods by combining sensory and visual data for comprehensive quality prediction.

References

Al-Zerki, S., Marhaba, A., Kassab, N., & Dalati, S. (2023). A Guidance Research for Sustainable Startup Business Center for Training in Research and Statistics. *Journal of Service, Innovation and Sustainable Development*, *4*(2). https://doi.org/10.33168/sisd.2023.0209

Chaudhary, P., Gulia, P., & Gill, N. S. (2024). A Comparison of Machine Learning Technologies for Soil Analysis in Smart Agriculture: A Comprehensive Review. *Proceedings - 2024 2nd International Conference on Advanced Computing and Communication Technologies, ICACCTech 2024*, 885–888. https://doi.org/10.1109/ICACCTech65084.2024.00145

Chen, Y. (2024). Research on Machine Learning-based Prediction of Coffee Futures Prices. In *Highlights in Science, Engineering and Technology SDPIT* (Vol. 2024). https://doi.org/DOI:10.54097/8c9t9n30

Chidoud, H., Aarab, A., & Amin, L. (2025). Integration of Certified Management Systems and Information Technology for Enhanced Risk Management: Evidence from Morocco's Agri-food Sector. *Journal of Logistics, Informatics and Service Science*, *12*(3), 125–140. https://doi.org/10.33168/JLISS.2025.0309

Deina, C., do Amaral Prates, M. H., Alves, C. H. R., Martins, M. S. R., Trojan, F., Stevan, S. L., & Siqueira, H. V. (2022). A methodology for coffee price forecasting based on extreme learning machines. *Information Processing in Agriculture*, *9*(4), 556–565. https://doi.org/10.1016/j.inpa.2021.07.003

Della Peruta, R., Mereu, V., Spano, D., Marras, S., Vezy, R., & Trabucco, A. (2025). Projecting trends of arabica coffee yield under climate change: A process-based modelling study at continental scale. *Agricultural Systems*, *227*. https://doi.org/10.1016/j.agsy.2025.104353

Erdogan, Z., Altuntas, S., & Dereli, T. (2024). Predicting Patent Quality Based on Machine Learning Approach. *IEEE Transactions on Engineering Management, Institute of Electrical and Electronics Engineers Inc.*, *71*, 3144–3157. https://doi.org/10.1109/TEM.2022.3207376

Ghimire, B. (2024). Determinants of Entrepreneurial Readiness: An Empirical Analysis. *Journal of Management Changes in the Digital Era*. https://doi.org/10.33168/jmcde.2024.0110

Khadse, S., Verma, P., & Milmile, A. (2025). A Review of Machine Learning Techniques for Predictive Crop Yield Modeling. *2nd International Conference on Machine Learning and Autonomous Systems, ICMLAS 2025 - Proceedings*, 525–530. https://doi.org/10.1109/ICMLAS64557.2025.10968404

Kim, J. Y. (2022). *Coffee Beans Quality Prediction Using Machine Learning*. https://doi.org/https://dx.doi.org/10.2139/ssrn.4024785

Kolhe, P., Kalbande, K., & Deshmukh, A. (2022). Internet of Thing and Machine Learning Approach for Agricultural Application: A Review. *International Conference on Emerging Trends in Engineering and Technology, ICETET*, *2022-April*. https://doi.org/10.1109/ICETET-SIP-2254415.2022.9791751

Krishna, V., Reddy, T., Harsha, S., Ramar, K., Hariharan, S., & Bhanuprasad, A. (2022). Analysis of Crop Yield Prediction using Machine Learning algorithms. *Proceedings - 2022 2nd International Conference on Innovative Sustainable Computational Technologies, CISCT 2022*. https://doi.org/10.1109/CISCT55310.2022.10046581

Kulkarni, S., Keerthi, N. C., Sunil, C. K., Pal, S., Dash, S., Shenoy, P. D., & Venugopal, K. R. (2023). Coffee Plant Disease Identification using Enhanced Short Learning EfficientNetV2. *2023 IEEE 20th India Council International Conference, INDICON 2023*, 91–96. https://doi.org/10.1109/INDICON59947.2023.10440883

Kuriakose, S. M., & Singh, T. (2022). Indian Crop Yield Prediction using LSTM Deep Learning Networks. *2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*. https://doi.org/10.1109/ICCCNT54827.2022.9984407

Lakhouit, A., AL Rashed, W. S., Abbas, S. Y. H., & Shaban, M. (2025). Integrating machine learning for precision agriculture waste estimation and sustainability enhancement. *Computers and Electronics in Agriculture*, *230*. https://doi.org/10.1016/j.compag.2025.109933

Lyimo, D. A., Lakshmi Narasimhan, V., & Mbero, Z. A. (2021). Sensitivity analysis of coffee leaf rust disease using three deep learning algorithms. *IEEE AFRICON Conference*, *2021-September*. https://doi.org/10.1109/AFRICON51333.2021.9571007

Ossani, P. C., Rossoni, D. F., Cirillo, M. Â., & Borém, F. M. (2021). Classification of specialty coffees using machine learning techniques. *Research, Society and Development*, *10*(5), e13110514732. https://doi.org/10.33448/rsd-v10i5.14732

Pallathadka, H., Jawarneh, M., Sammy, F., Garchar, V., Sanchez, T., & Naved, M. (2022). A Review of Using Artificial Intelligence and Machine Learning in Food and Agriculture Industry. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022*, 2215–2218. https://doi.org/10.1109/ICACITE53722.2022.9823427

Pinheiro Claro Gomes, W., Gonçalves, L., Barboza da Silva, C., & Melchert, W. R. (2022). Application of multispectral imaging combined with machine learning models to discriminate special and traditional green coffee. *Computers and Electronics in Agriculture*, *198*. https://doi.org/10.1016/j.compag.2022.107097

Putra, A. Z., Chalvin, C., Nurhadi, A., Tambun, A. E., & Defha, S. (2023). Coffee Quality Prediction with Light Gradient Boosting Machine Algorithm Through Data Science Approach. *Sinkron*, *8*(1), 563–573. https://doi.org/10.33395/sinkron.v8i1.12169

Rajbharath, R., Vijayamalaiya, S., Aabid, M., & Salilan, A. (2023). Cofee Quality Prediction Using Machine Learning. *2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023*. https://doi.org/10.1109/ICSCAN58655.2023.10395728

Rajkumar, N., & Mukunthan, M. A. (2023). Efficient Crop yield Analysis Prediction in Modern Agriculture System using Machine Learning Algorithm. *2023 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2023*. https://doi.org/10.1109/ICDSAAI59313.2023.10452646

Ramu, K., & Priyadarsini, K. (2021). A Review on Crop Yield prediction Using Machine Learning Methods. *Proceedings - 2nd International Conference on Smart Electronics and Communication, ICOSEC 2021*, 1239–1245. https://doi.org/10.1109/ICOSEC51865.2021.9591764

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction. In *IEEE Access* (Vol. 9, pp. 63406–63439). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3075159

Rubia Gandhi, R. R., Angel Ida Chellam, J., Prabhu, T. N., Kathirvel, C., Sivaramkrishnan, M., & Siva Ramkumar, M. (2022). Machine Learning Approaches for Smart Agriculture. *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, 1054–1058. https://doi.org/10.1109/ICCMC53470.2022.9753841

Santhosh, C. S., & Umesh, K. K. (2022). A Compendium Probabilistic Prospective for Predicting Coffee Crop Yield Based on Agronomical Factors. *4th International Conference on Emerging Research in Electronics, Computer Science and Technology, ICERECT 2022*. https://doi.org/10.1109/ICERECT56837.2022.10060420

Saurnida Asiana Siahaan, A., Masrul Harahap, E., Hanum, C., Karim, A., & Vinceviča-Gaile, Z. (2023). The Taste of Arabica Coffee in Several Altitude and Shading Condition. *E3S Web of Conferences*, *374*. https://doi.org/10.1051/e3sconf/202337400001

Septiarini, A., Hamdani, H., Burhandenny, A. E., Nur, S., & Winarno, E. (2023). The color-texture features and machine learning approach for quality detection of coffee beans. *Proceedings of the 9th International Conference on Computer and Communication Engineering, ICCCE 2023*, 217–222. https://doi.org/10.1109/ICCCE58854.2023.10246091

Sermmany, K., Wanjantuk, P., & Leelapatra, W. (2024). Utilizing Explainable Artificial Intelligence (XAI) to Identify Determinants of Coffee Quality. *Proceedings - 21st International Joint Conference on Computer Science and Software Engineering, JCSSE 2024*, 696–703. https://doi.org/10.1109/JCSSE61278.2024.10613641

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. In *IEEE Access* (Vol. 9, pp. 4843–4873). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2020.3048415

Somasundaram, R. S., Nagamani, K., Florence, M. L., & Swamydoss, D. (2022). Estimation and Prediction of Crop Yielding Rate Using Machine Learning Techniques. *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*. https://doi.org/10.1109/ICCCI54379.2022.9741032

Sriram, K. P., Sujatha, P. K., Athinarayanan, S., Kanimozhi, G., & Joel, M. R. (2024). Transforming Agriculture: A Synergistic Approach Integrating Topology with Artificial Intelligence and Machine Learning for Sustainable and Data-Driven Practice. *2nd International Conference on Sustainable Computing and Smart Systems, ICSCSS 2024 - Proceedings*, 1350–1354. https://doi.org/10.1109/ICSCSS60660.2024.10625446

Sunil, G. L., Nagaveni, V., & Shruthi, U. (2022). A Review on Prediction of Crop Yield using Machine Learning Techniques. *2022 IEEE Region 10 Symposium, TENSYMP 2022*. https://doi.org/10.1109/TENSYMP54529.2022.9864482

Vashisht, S., Kumar, P., & Trivedi, M. C. (2022). Improvised Extreme Learning Machine for Crop Yield Prediction. *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, 754–757. https://doi.org/10.1109/ICIEM54221.2022.9853054