Specialty Coffees Classification Utilizes Feature Selection and Machine Learning

Nelly Oktavia Adiwijaya[1, a)] and Riyanarto Sarno[2, b)]

[1] *Informatics Department, Faculty of Computer Science, University of Jember, Indonesia*
[2] *Informatics Department, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS), Indonesia*

[a)] *nelly.oa@unej.ac.id*
[b)] *Corresponding author: riyanarto@if.its.ac.id*

Abstract. An important factor that influences the price of coffee bean commodities is their quality. Specialty coffee beans are the quality of coffee beans with the highest price. Determining the quality of specialty coffee beans is determined through a long and complicated series of physical tests and cupping test by an expert called Qgrader. This research proposes classifying specialty coffee beans using several machine learning methods. The first step taken was to label the data in accordance with the Specialty Coffee Association of America standard rules. The coffee classes used in this research are Grade 1, Grade 2 and Grade 3. Next, feature selection was carried out using correlation analysis and important features which resulted in 6 features out of 11 features. This study compares the results of classification using 3 different models, namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Random Forest (RF) with accuracy results of 78%, 100% and 100% respectively.

Keywords: Specialty coffee classification, machine learning models, feature selection, cupping quality assesment

INTRODUCTION

Coffee is one of the highest traded commodities in the world [1]. Coffee is also included in the list of superior products in Indonesian trade as a commodity exported to dozens of countries in the world [2]. Coffee beans are an important component in the coffee industry. The quality of coffee beans has a significant influence on taste and consumer satisfaction [3]. Classifying the quality of coffee beans is an important key in this industry because it influences the price of the coffee commodity.

Coffee quality is influenced by several things, including origin which is related to ground level [4], taste and aroma [5], acidity [6], [7], shape and size of beans, the color of coffee beans [8]and value of coffee defects. The taste and aroma of coffee are influenced by the chemical composition contained in coffee beans [9]. By having clear information about the quality of coffee beans, producers and retailers can choose the best coffee beans for further processing or offering to consumers [3]. International standards for the quality of specialty coffee beans are set by the SCAA (Specialty Coffee Association of America) organization [10]. The method for ranking the quality of coffee beans is based on the relationship between bean defects and the taste of the coffee beans [11]. The process of assessing coffee quality is carried out by experts using a long and complicated procedure [12]. Therefore, research on the classification of coffee beans in terms of quality has become important to provide more efficient and accurate solutions in this process.

Recent research has discussed coffee classification using several type of coffee bean. The classification includes evaluation of the high quality and low quality of fresh coffee bean [13], estimation of caffeine content of Arabica and Robusta samples [14], assessing the roasting level of Arabica Gayo coffee [15], identifying pure Arabica coffee and pure Robusta coffee using several machine learning method such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN) [16]. The research that has been carried out discusses more about differentiating coffee beans based on their cultivars.

The aim of this research is to be able to differentiate coffee beans based on their quality which can influence the price of coffee bean commodities. The data used in this research specifically uses data with parameters that are relevant to assessing the quality of coffee beans based on International standard of coffee quality. The parameters used have different characteristics to identify the aroma and taste of coffee beans. This data is then carried out in the preprocessing stage by cleaning the data. Feature selection is also applied to obtain the most important features to differentiate the quality of coffee beans. After the preprocessing stage, the calculation continues with the classification

I. H. Agustin (ed.), *Proceedings of the 2nd International Conference on Neural Networks and Machine Learning 2023 (ICNNML 2023)*, Advances in Intelligent Systems Research 183,
https://doi.org/10.2991/978-94-6463-445-7_11

stage using several machine learning methods. This study has four grades as quality output, namely Grade 1, grade 2, grade 3 and grade 4. The accuracy [17] is used in this study to evaluate the classification method.

Related Work

Previous research on the classification of coffee beans has been carried out with various types of coffee beans and used various methods for data analysis. Detection and classification of Indonesian luwak and nonluwak coffee. The objects used are Aceh coffee, Arjuno Malang coffee and Bengkulu coffee, so there are 6 different types of luwak and nonluwak coffee in these three coffee shops. Feature extraction is performed statistically in this study, which are mean, standard deviation, maximum and minimum. Then, the classification process is performed using four machine learning methods namely Logistic Regression, Support Vector Machine, Decision Tree, and Naive Bayes. The best accuracy result obtained was 97% in the decision tree method with average statistical parameters and standard deviation [17].

Further research was carried out to determine the origin of the coffee. The coffees used as subjects were coffees from Yigacheffe and Kona. Of the 850 volatile compounds identified in roasted coffee beans, the study used only 30 compounds related to coffee flavor and aroma. The advantage of this study is that humidity conditions are adjusted and maintained to maintain accuracy. Because the environmental conditions of humidity, temperature and their changes greatly affect the moisture content of the coffee beans and the sensor response and ultimately affect the aroma concentration results obtained. Increased humidity and temperature will cause the sensor's sensitivity to decrease, which will affect the output [18].

In research that predicts the acidity level of roasted coffee beans using electronic nose, 8 gas sensors were choosen. Data collection here is done by cupping, namely brewing water at a temperature of 9397 degrees for 3 minutes. The method used is an Artificial Neural Network and produces values that are close to the same as the assessments made by cupping experts with an accuracy of up to 95% [19]. Another approach used for level classification of Gayo Arabica coffee is using stepwise Linear Discriminant and K-Nearest Neighbor. Coffee is roasted using 3 different tempera tures and different roasting times. The accuracy obtained was 91.67% [15].

International Standard of Coffee Quality

The international standard for determining the quality of coffee beans is arranged by Specialty Coffee Association of America (SCAA). The SCAA sets international standards for the quality of specialty coffee beans. Coffee quality measurements are carried out through samples per 300 grams [20]. For specialty quality coffee, no more than 5 pieces of coffee are full of defects. No major defects are permitted. Maximum tolerance ±5% above or below the tolerated screen size. This coffee must have a special characteristic in body, taste, aroma, or acidity. This coffee must be defect free and there must be no quakers. For premium quality coffee, no more than 8 copies of coffee are full of defects and may have major defects. Maximum tolerance ±5% above or below the tolerated screen size. This coffee must have a special characteristic in body, taste, aroma, or acidity and must only contain 3 Quaker coffee beans. The complete characteristics of coffee bean quality regulation according to the SCAA are as in Table 1.

TABLE 1. SCAA grading criteria

Grade / Parameter	Specialty Grade	Premium Grade	Exchange Grade	Below Standard Grade	Off Grade
Screen size	≤ 5%	≤ 5%	50% sz.15 ≤ 5% Sz.14	> 5%	> 5%
Aroma					
Flavor					
Acidity					
Fault & Taints	+	+	+		
Primary defect	+				
Full Defect	≤ 5	≤ 8	923	2486	> 86
Quaker	+	≤ 3	≤ 5	>5	>5

Grade / Parameter	Specialty Grade	Premium Grade	Exchange Grade	Below Standard Grade	Off Grade
Moisture 913%					

METHOD

This research aims to analyze coffee sample data that has been rated by a Q grader. Technical Reports include an array of information, including cupping notes and scores from three certified Q Graders and a breakdown of green coffee defects. To obtain the best classification accuracy, a comparison is carried out between the performance of various combinations of classification methods. The classification system scheme used in this research is shown in Figure 1.

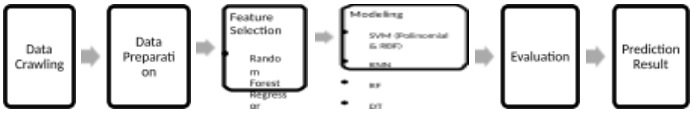

FIGURE 1. Classification scheme

Data Description

In this study, the data used came from https://database.coffeeinstitute.org/. The data obtained is coffee data from cupping test results in 2022 and 2023. The coffee samples contained in this database come from various countries and certified assessors come from various countries too. The results of data collection amounted to 193 data for Arabica coffee types. The number of parameters from the data crawling results was 51. Of all the parameters, only 16 parameters were taken which were related to the aroma, taste and visual profile of the coffee beans, altitude and coffee bean processing method after harvest. The other parameter are not included about the coffee identity like the country, company, owner, etc. Data related to coffee owner company, country and address are not included. The list of parameters used in this research can be seen in Table 2.

TABLE 2. Data description.

Parameter	type	description
Aroma	Float64	scent or fragrance of the coffee
Flavor	Float64	Value taste of coffee
Acidity	Float64	the brightness or liveliness of the taste
Body	Float64	the thickness or viscosity of the coffee in the mouth
Moisture	Float64	Humidity of coffee beans
primary defect	Int64	Value of primary defect
Secondary defect	Int64	Value of secondary defect
Quaker	Int64	Value of quaker bean
Aftertaste	Float64	the lingering taste that remains in the mouth after swallowing the coffee
Balance	Float64	how well the different flavor components of the coffee work together
Uniformity	Int64	the consistency of the coffee from cup to cup
Clean cup	Int64	a coffee that is free of any offflavors or defects, such as sourness, mustiness, or staleness
Sweetness	Float64	caramellike, fruity, or floral, and is a desirable quality in coffee
Total cup points	Float64	Total score from the cupping form
Processing method	object	The process method of postharvest
Altitude	Float64	The altitude origin of the coffee

Data Preparation

In data preprocessing, data cleaning is carried out to ensure that there is no data that is inconsistent and has null values. For the parameter ‘uniformity’, ‘clean cup’, and ‘sweetness’ are ignored because they have the same value so they are not considered to have a different influence on the results. Apart from that, the altitude parameter which contains the range value is divided into the lowest value and the highest value and then an average value is created for the altitude value. What is used then is ‘altitude_mean’ as the average altitude value. Feature engineering is also carried out to detect outliers. Seven outliers data are found in the altitude parameter that shows outside the reasonable height limit. Handling of outliers is done by checking regional identity and actual height via the internet.

Data labeling is carried out by calculating parameters related to SCAA categories. The parameters used for labeling here are the top 8 variables in table 2. The grading result form the dataset are 4 classess (Grade 14) as seen in table 3.

TABLE 3. Amount of labelling result

Grade	Sum of the data
Grade 1	37
Grade 2	19
Grade 3	116
Grade 4	21

The processing method parameter in the dataset is object data type that consist of categorical data, so its needs to be encode before doing the classification. Encoding categorical data is a crucial step in data preparation for machine learning. Categorical data, which consists of finite possible values divided into groups, needs to be converted into numerical format before being provided to machine learning algorithms for better results [21]. The encode method used here is OneHot Encoding. This method creates binary columns for each category of a categorical variable. If a category is present in a row, the corresponding binary column is marked as 1; otherwise, it is marked as 0. The results of the encoding process get 19 features.

Feature Selection

In this step, feature selection is perform to identified and selected the most relevant features from a dataset to improve model performance and interpretability. The method we used here is random forest regressor. Random Forests are a popular machine learning algorithm that can provide good predictive accuracy by creating an ensemble of decision trees. The feature selection process in a Random Forest Regressor is an example of an embedded method, which combines the qualities of filter and wrapper methods and is implemented by algorithms that have their own builtin feature selection methods [22].

Random Forests can calculate the importance of a feature based on its ability to increase the purity of the leaves in the decision trees. Features that are selected at the top of the trees are generally more important than features that are selected at the end nodes of the trees, as the top splits lead to bigger information gains [11]. Feature selection in a Random Forest Regressor can be beneficial in situations where the dataset contains a large number of features, some of which may be irrelevant or noisy. By selecting only the most informative features, the model's performance can be improved, and the interpretability of the model can be enhanced. However, it is important to note that Random Forests are often said to perform well "out-of-thebox," with no tuning or feature selection needed, even with highdimensional datasets [23].

In this study, the most important features selected were 6 out of 31 features using random forest regressor. The six highest feature are ‘total_cup_points’, ‘quakers’, ‘moisture’, ‘sec-def’, ‘acidity’, and ‘aroma’ as seen in Figure 2.

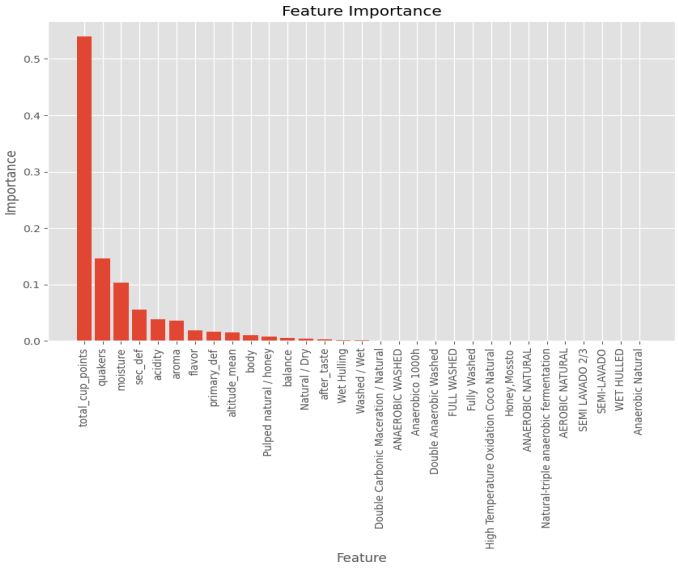

FIGURE 2. Feature selection using random forest regressor

Modelling

The modeling method we used here are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB). KNN is Non-generalizing model that "remembers" all of its training data. Supports multiclass classification. Classification is computed from a simple majority vote of the nearest neighbors. It does not require any training phase, fast for large datasets and can handle imbalanced datasets [24]. The SVM method can handle both linear and nonlinear relationships, good for high dimensional data, and can handle imbalanced datasets [22]. Random Forest is an ensemble method that constructs multiple decision trees, and can identifies important features for prediction. It can handle a large number of features. The Decision tree method can handle both categorical and numerical data, flexible and can be used for regression and classification tasks, also with pruning can lead to better accuracy [24]. Naïve bayes method is a generative model that assumes prior probabilities of class labels are known, Requires less training time compared to other algorithms [22]. This study focuses on employing five different classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), and Naive Bayes (NB). The dataset is divided into training and testing sets with an 80:20 ratio, ensuring that the models are trained on a substantial portion of the data and evaluated on a separate, unseen portion. The dataset comprises 31 features, and a feature selection process is implemented to narrow down the relevant variables for modeling. Six features, selected through this process, are included in the modeling phase. To enhance the performance of each algorithm, hyperparameter tuning is applied, optimizing the parameters of the models for better predictive accuracy. This comprehensive approach aims to create robust and welloptimized predictive models capable of making accurate predictions on new, unseen data.

Evaluation

Accuracy measurement stages are carried out for each method. The aim of accuracy evaluation is to measure how well a machine learning model correctly identifies and predicts the true labels of the data. Accuracy is calculated as the ratio of correctly predicted instances to the total number of instances [25].

RESULT

This research was conducted with the aim of classifying the quality of coffee based on parameters related to the aroma, taste and physical properties of the coffee beans. The data obtained was 193 samples from various countries by various certified Q grader assessors. The data is divided into four classifications based on the SCAA calculation values, namely grade 1, grade 2, grade 3 and grade 4. First, we try to obtain calculation accuracy by selecting features using the Random Forest Reressor method which produces 6 highest important parameters. Then training are made using KNN, SVM, RF, DT and NB methods. The splitting data scenario is carried out in a ratio of 80:20. The training process is carried out in both schemes, namely by involving all features and using only 6 features resulting from feature selection. The results are then compared as shown in table 4. The analysis results show that the application of the feature selection method has an impact on increasing prediction accuracy on testing data for all classification methods used. The experience have shown that feature selection has the potential to improve classification accuracies, especially when applied in the context of specific methods and datasets. a feature selection approach was found to improve the accuracy up to 22% for KNN.

TABLE 4. Comparison of the accuracy with and without feature selection

Classifier	Accuration	
	without feature	With feature selection
KNN	71,79	87,18
SVM	79,49	84,62
RF	76,92	87,18
DT	84,62	87,2
NB	28,21	51,28

The results of the predictive modeling indicate that the Decision Trees (DT) method achieved the highest accuracy among the considered algorithms. Specifically, the DT model exhibited an impressive accuracy of 87.2% when feature selection was applied. This outcome underscores the effectiveness of the Decision Trees algorithm in capturing the underlying patterns within the dataset and making accurate predictions. The utilization of feature selection also played a crucial role in enhancing the model's performance by focusing on the most relevant variables. The notable accuracy obtained with the DT method suggests its suitability for the given dataset and reinforces its potential as a robust classification algorithm for similar predictive tasks. The findings underscore the importance of not only selecting appropriate algorithms but also employing feature selection techniques to optimize model performance in predictive modeling scenarios. Further discussions on the interpretability and generalizability of the model, as well as potential areas for improvement, can contribute to a comprehensive understanding of the results.

CONCLUSION

This research aims to differentiate coffee of different qualities using several machine learning methods. Feature selection at the preprocessing stage is used to obtain the most important parameters related to coffee quality. The highest classification accuracy results were obtained from the KNN, Random Forest and DT algorithms. The accuracy can still be improved by doing data augmentation for the imbalanced dataset.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to Prof. Drs. Ec. Ir. Riyanarto Sarno, M.Sc., Ph.D for his invaluable guidance and support throughout my doctoral research. His unwavering commitment to academic excellence, insightful feedback, and encouragement have been instrumental in shaping the trajectory of my doctoral journey. I am truly fortunate to have had Him as my doctoral advisor, and his expertise has significantly enriched the quality of my research. I look forward to continuing this research journey under his supervision and anticipate the insightful contributions that will further enrich the subsequent phases of my doctoral investigation.

REFERENCES

1. economicshelp. economicshelp.org, “Most traded commodities”, (2023)
2. Kemendag. http://ppejp.kemendag.go.id/, ”Produk Unggulan Indonesia”, (2022)
3. LannaCoffeeCo. www.lannacoffeeco.com, “The Importance of Choosing HighQuality Coffee Beans for Your Business”, (2023)
4. de Assis Silva S, de Queiroz DM, Ferreira WPM, Corrêa PC, dos Santos Rufino JL. Mapping the potential beverage quality of coffee produced in the Zona da Mata, Minas Gerais, Brazil. *J Sci Food Agric*, (2016 Jul 1);96(9):3098–108.
5. M.R.N. Alcantara G, Dresch D, R. Melchert W, “Use of nonvolatile compounds for the classification of specialty and traditional Brazilian coffees using principal component analysis”, *Food Chem*, (2021 Oct 30), 360.
6. Abreu MB, Marcheafave GG, Bruns RE, Scarminio IS, Zeraik ML., “Spectroscopic and Chromatographic Fingerprints for Discrimination of Specialty and Traditional Coffees by Integrated Chemometric Methods”, *Food Anal Methods*, (2020 Dec 1), 13(12):2204–12.
7. Perrone D, Farah A, Donangelo CM, de Paulis T, Martin PR., “Comprehensive analysis of major and minor chlorogenic acids and lactones in economically relevant Brazilian coffee cultivars”, *Food Chem,* (2008 Jan 15), 106(2):859–67.
8. Adiwijaya NO, Romadhon HI, Putra JA, Kuswanto DP., “The quality of coffee bean classification system based on color by using K-Nearest neighbor method”, *Journal of Physics: Conference Series*, (2022)
9. Tsai JJ, Chang CC, Huang DY, Lin TS, Chen YC., “Analysis and classification of coffee beans using single coffee bean mass spectrometry with machine learning strategy”, *Food Chem*, (2023 Nov 15), 426.
10. sca.coffee. https://sca.coffee/research/protocolsbestpractices, (2023), Protocols & Best Practices.
11. Malato G. www.yourdatateacher.com, “Feature selection with Random Forest”, (2021)
12. BareaRamos JD, Cascos G, Mesías M, Lozano J, MartínVertedor D., “Evaluation of the Olfactory Quality of Roasted Coffee Beans Using a Digital Nose”, *Sensors*, (2022 Nov 1), 22(22).
13. Cascos G, Lozano J, Arroyo P, RuizCanales A, Oates MJ, MartinVertedor D., “Fusion data of digital olfation devises for the evaluation of the quality of fresh coffee beans”, *Res Sq*, (2023)
14. Aghdamifar E, Sharabiani VR, Taghinezhad E, Szymanek M, DziwulskaHunek A., “Enose as a nondestructive and fast method for identification and classification of coffee beans based on soft computing models”, *Sens Actuators B Chem*, (2023 Oct 15), 393.
15. Nasution IS, Delima DP, Zaidiyah Z, Fadhil R., “A Low Cost Electronic Nose System for Classification of Gayo Arabica Coffee Roasting Levels Using Stepwise Linear Discriminant and K-Nearest Neighbor”, *Mathematical Modelling of Engineering Problems*, (2022 Oct 1), 9(5):1271–6.
16. Harsono W, Sarno R, Izza Sabilla S., “Recognition of Original Arabica Civet Coffee based on Odor using Electronic Nose and Machine Learning”, *In 2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 333339, IEEE, (2020, September)
17. Wakhid S, Sarno R, Sabilla SI, Maghfira DB., ”Detection and classification of indonesian civet and noncivet coffee based on statistical analysis comparison using ENose”, *International Journal of Intelligent Engineering and Systems*, 13(4):56–65, (2020)
18. Lee CH, Chen I Te, Yang HC, Chen YJ., ”An AIpowered Electronic Nose System with Fingerprint Extraction for Aroma Recognition of Coffee Beans”, *Micromachines (Basel)*, 13(8), (2022 Aug 1)
19. coffeeresearch.org. http://www.coffeeresearch.org/coffee/scaaclass.htm., SCAA Coffee Beans Classification, (2023)
20. Rohith S. M. devgenius.io, “Encoding Methods to encode Categorical data in Machine Learning”, (2022)
21. Untoro MC, Praseptiawan M, Widianingsih M, Ashari IF, Afriansyah A, Oktafianto, “Evaluation of Decision Tree, KNN, Naive Bayes and SVM with MWMOTE on UCI Dataset”, *In: Journal of Physics: Conference Series. Institute of Physics Publishing*, (2020)
22. Varghese D., “Towards Data Science”, *Comparative Study on Classic Machine learning Algorithms*, (2018)
23. Mansur HS, Adiwijaya NO, Dharmawan T., “Optimization of Machine Learning Algorithms with Bagging and AdaBoost Methods for Stroke Disease Prediction”, *Applied Medical Informatics Original Research*, Vol. 45, (2023)
24. Varghese V, Krishnan V, Kumar GS., ”Evaluating PedicleScrew Instrumentation Using DecisionTree Analysis Based on Pullout Strength”, *Asian Spine J*., 12:611–21, (2018), doi: 10.31616/asj.2018.12.4.611
25. Mansour, S., Jalali, A., Ashjaee, M., Houshfar, E., “Multiobjective optimization of a sandwich rectangular channel liquid cooling plate battery thermal management system: A deeplearning approach”, *Energy Conversion and Management*, 290, 117200, (2023)