Advances in Artificial Intelligence and Machine Learning; Research 2 (1) 273-287 Received 28-12-21; Accepted 05-02-22; Published 15-02-22

Machine learning to support geographical origin traceability of Coffea Arabica

Elisabete A. De Nadai Fernandes[1*], LIS@CENA.USP.BR
Gabriel A. Sarriés[2],
Yuniel T. Mazola[1], **Robson C. de Lima**[1],
Gustavo N. Furlan[1], **Márcio A. Bacchi**[1]
[1]*Nuclear Energy Center for Agriculture, University of São Paulo, Avenida Centenário 303, 13416-000 Piracicaba, SP, Brazil*
[2]*College of Agriculture Luiz de Queiroz, University of São Paulo, Avenida Pádua Dias 11, 13418-900 Piracicaba, SP, Brazil.*

[*]**Corresponding Author:** Elisabete A. De Nadai Fernandes.

Abstract

The species, variety and geographic origin of coffee directly influence the characteristics of the coffee beans and, consequently, the quality of the beverage. The added economic value that these features bring to the product has boosted the use of non-designative tools for authentication purposes. In this work, the feasibility of implementing a traceability system for Arabica coffee by country of origin was investigated using quality attributes and supervised machine learning algorithms: Multilayer Perceptron (MLP), Random Forest (RF), Random Tree (RT) and Sequential Minimal Optimization (SMO). We used an available database containing quality parameters for coffee beans produced in 15 countries, including the largest exporters and importers. Overall, Ethiopia, Kenya and Uganda had the highest coffee quality index (Total Cup Points). Differences between countries were found with 99% confidence using Robust Multivariate Data Science with original data and 98% accuracy using Bootstrapping resampling method and Supervised Machine Learning algorithms. The model obtained by RF provided the best classification accuracy. The most important attributes to discriminate Arabica coffee by country of origin, in descending order, were body, moisture, total cup points, cupper points, acidity, aftertaste, flavor, aroma, balance, sweetness and uniformity. The coffee variety proved to be a promising variable to increase accuracy and can be incorporated among the quality attributes for classification and grading of coffee beans.

Keywords: coffee producing countries; coffee quality; coffee varieties; data science

1. INTRODUCTION

Coffee is one of the most popular beverages worldwide and also one of the most profitable international commodities. Coffee is the name of the seed of the coffee plant that belongs to the botanical family of Rubiaceae and the genus Coffea. Among the more than hundred species of this genus, the

Citation: Elisabete A. De Nadai Fernandes et al. Machine learning to support geographical origin traceability of Coffea Arabica. Advances in Artificial Intelligence and Machine Learning. 2022;2(1):18.

two most economically cultivated are Coffea arabica and Coffea canephora, which supply almost all the world´s coffee consumption. The Arabica type is traditionally known for giving the best coffee drink, mainly attributed to a large number of different varieties and hybrids that bring a myriad of aroma, sweetness, acidity, body and flavor, depending on the region of origin and other factors from cultivation to roasting. The most common currently cultivated are Bourbon, Mundo Novo, Blue Mountain, Catuaí, Caturra, Kona, Typica, among others.

It is believed that the consumption of coffee as a beverage first emerged in Yemen in the mid-15th century in shrines, brewed as current methods from roasted coffee seeds originating in the Ethiopian Highlands. From the 16th century onwards, coffee spread to the Middle East, North Africa, Europe, Indonesia and the Americas. Although introduced in Brazil in 1727, coffee cultivation only gained prominence after the country's independence in 1822, when plantations took place in tropical forest areas in the states of Rio de Janeiro and São Paulo. Brazil already became the world's largest producer in 1852 and has been ever since. In the 1910s, the country exported around 70% of the world's coffee, followed by Colombia, Guatemala and Venezuela with 15% [1]. Nowadays, coffee is grown in more than 70 countries, mainly in the equatorial regions of the Americas, Southeast Asia, the Indian subcontinent and Africa. The ten countries ranked by the highest coffee production rates are Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, Mexico and Guatemala [2].

Climatic conditions are the most critical in coffee production, as they affect plant growth and development in different ways. For example, when frost occurs, the dew resulting from the extreme drop in temperature can form an ice deposition on plant surface, being responsible for coffee crop failure. This type of frost hits the plants by the action of very cold winds, which cause the tissue temperature to fall below the limit corresponding to the internal freezing point of 2°C [3, 4]. Brazil, being the largest coffee producer in the world, plays a decisive role in the global coffee market. Severe frosts hit Brazilian coffee producing areas in the southeast of the country in July 2021, with temperatures as low as -1.2 °C, irreparably damaging coffee plants. Earlier this year, there was a prolonged drought in the country, which means about double the damage caused by frosts for crops in the coming years. Estimated production losses are increasing and market analysts and coffee traders predict repercussions for the global coffee market and rising prices for many years to come.

The world coffee market is a billion-dollar trade, but there is a great disparity in the production chain, between the profit from sales and that of the people who actually produce the coffee. Coffee production has been criticized for the environmental impact caused by land clearing for insertion of crops and large use of water. These gaps are leading to the emergence of environmentally and socially differentiated markets like organic coffee and fair trade. Global market trends for coffee supply chain sustainability emphasize certification schemes such as Fairtrade, Organic, Rainforest Alliance, UTZ, considering the influence of these standards on coffee producers and the industry [5]. The relevance of the ESG (Environmental, Social and Governance) policy on coffee agribusiness is currently on the agenda of sustainable coffee growing, bringing together innovations, technologies and efficient practices in the environmental, managerial and social areas, especially among small rural producers who are the weakest link in the production chain. Alternative production methods, reducing the use of herbicides, insecticides and fungicides, provide environmental, technical and economic benefits, such as increased soil organic matter, soil protection, nutrient cycling, market opportunities and cost savings [6].

Coffee grading and classification system for green coffee is primarily aiming at producing homogenous commercial lots that meet defined quality criteria, thus facilitating a fair price system. Producing countries have developed their own classification and grade charts, used mainly to establish minimum standards for coffee trade, based on characteristics such as botanical variety, growing altitude and region, processing method, bean size, shape, color, density, number of defects, and cup quality. Coffee industry is making great efforts in scientific research for sensory development with the aim of having a universal sensory language that adapts to all grading situations [7]. The Specialty Coffee Association (SCA) recommends standards for cupping coffee to enable more accurate assessment of coffee quality. Still, there are some limitations as two coffees could get exactly the same score despite having different profiles. The World Coffee Research Sensory Lexicon, developed at Kansas State University's Sensory Analysis Center, uses sensory science tools and technologies to understand the primary sensory qualities of coffee, in a descriptive, quantifiable and replicable way, containing coffee flavors, aromas and textures as determined by a panel of sensory experts and coffee industry leaders [8].

The influence of production conditions on coffee quality has led several authors to combine analytical techniques to unequivocally authenticate different species [9-11], categories [12-15], geographical origin [16-18], processing methods [19, 20], production systems [21, 22], among others. However, obtaining a database with reliable chemical analysis data from representative coffee samples from the world's largest producers entails high costs and logistical difficulties. In this paper, we have focused on a database available from the Coffee Quality Institute (CQI), which is a non-profit organization working internationally to improve the quality of coffee and the lives of people who produce it, supporting them to achieve high quality standards and rewarding its marketing. Supervised machine learning algorithms were applied as a means of investigating geographical origin traceability of Arabica coffee based on internationally adopted quality attributes.

2. MATERIALS AND METHODS

Data contained in a database available on the Kaggle platform [23], collected in January 2018 from the Coffee Quality Institute (CQI) [24], have been used for this study. The database contains assessments made by CQI experts on quality parameters of Arabica coffee produced in several countries. Data referring to quality parameters of 15 coffee producing countries having at least 20 records were analyzed: Brazil (n = 132), Colombia (n = 183), Costa Rica (n = 51), El Salvador (n = 21), Ethiopia (n = 44), Guatemala (n = 181), Honduras (n = 52), Kenya (n = 25), Mexico (n = 236), Nicaragua (n = 26), Taiwan (n = 75), Tanzania (n = 40), Thailand (n = 32), Uganda (n = 26) and United States (n = 81). Then, from a total of 1312 records available, 1191 were used in the study encompassing 11 attributes related to coffee quality: Acidity, Aftertaste, Aroma, Balance, Body, Cup Cleanliness (Clear Cup), Cupper Points, Flavor, Moisture, Sweetness and Total Cup Points.

In order to provide asymmetry correction, eliminate outliers and reduce the variance of the results, 200 records were generated for each country using bootstrapping branching (10% sample size) [25, 26]. For this, all the values of a column (X) belonging to a country (Y) were transformed into an array, from which N values were drawn, where N = array size/10, and the arithmetic mean was calculated. This process was repeated 200 times, thus obtaining the number of observations for each country.

Linear Regression (LR), Quadratic Discriminant Analysis (QDA) and unsupervised machine learning (Cluster Analysis) were implemented as tasks of prediction, generative model and exploratory data analysis, respectively. Supervised Machine Learning (SML) techniques such as Multilayer Perceptron (MLP), Random Forest (RF), Random Tree (RT) and Sequential Minimal Optimization (SMO) were implemented for classification. The k-cross-validation (k=10) was used to divide the data into two groups: one used to learn (train) the model and the other to validate the model (test) [26, 27]. Each algorithm was repeated 20 times to reduce the estimated error of the model´s classification performance [28].

Artificial Neural Network is a classification model based on the interconnection of nodes, also referred to as perceptrons. They were inspired by the cognitive system and the neurological functions of the human brain, simulating the system of transmission of nerve impulses by neurons and their ligaments [29]. A perceptron has three basic elements (input connections, linear combiner and activation function). The input connections are weighted by a synaptic weight. Each node Xi is multiplied by the synaptic weight Wi (calculated by the algorithm), and is subsequently connected to the neuron. The entry has a fixed value, nonzero. The linear combiner is responsible for the sum of the input values (X1*W1 + X2*W2 + ... + Xi*Wi), generating the activation potential U. The activation function evaluates the activation potential U and through a function f(U) calculates the output signal of the neuron to identify the classes (Brazil, Colombia, Costa Rica, El Salvador, Ethiopia, Guatemala, Honduras, Kenya, Mexico, Nicaragua, Taiwan, Tanzania, Thailand, Uganda and United States). The Multilayer Perceptron (MLP) is a neural network similar to the simple perceptron, with hidden layers between input and output layers. It works with the backward propagation of errors. Prediction errors obtained during the training set analysis are propagated from the output layer to the previous layers. This error value is used to adjust the weight values on each edge [30, 31]. FIGURE 1 shows the model used in the Multilayer Perceptron network.

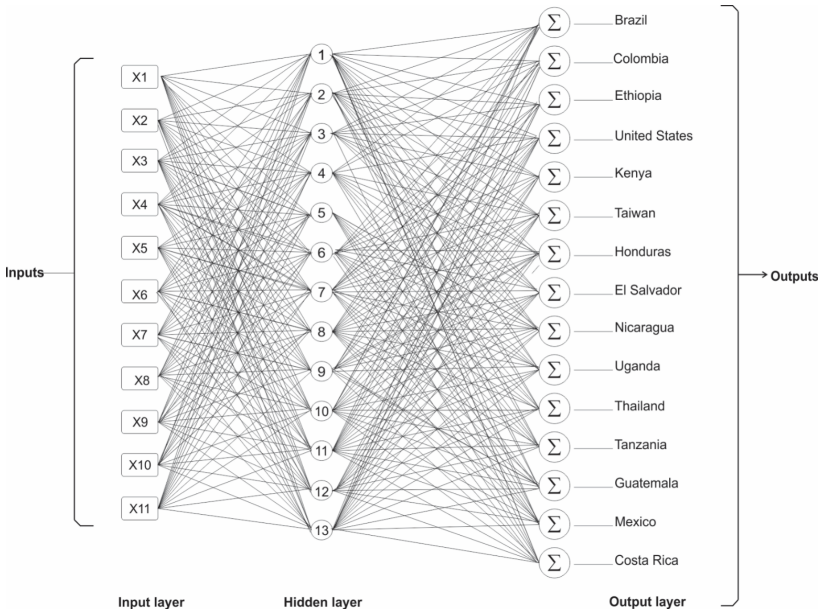


Figure 1: Network architecture used in the Multilayer Perceptron.

Decision Tree (DT) algorithms can be used for classification or regression predictive modelling. Each attribute of the training data set is analyzed individually, and these questions and their answers form classification rules [32]. The DT expresses the classification rules in a tree structure. Each node in the tree corresponds to a quality attribute and each edge that exits a node represents a value or a range of values.. The classification of an unknown example occurs by checking their values, starting with the root node, and following the edge representing the value obtained. This iterative process will generate a path through the decision tree until the associated class label is identified with the training class label. Random Forest (RF) adds an additional layer of randomness to bagging and builds multiple decision trees using bootstrapped samples from the original training data. This type of resampling, in which a large number of smaller samples of the same size are repeatedly drawn, changes the way the tree is constructed. In a random forest, each node is split using the best among a subset of predictors (attributes related to coffee quality) chosen randomly at that node. It achieves high levels of accuracy, generally much higher than those obtained with a single decision tree [33]. RF has only two parameters, the number of variables in the random subset at each node and the number of trees in the forest [34]. The set of trees is then used for classification of an example based on the most frequent classification among them, each tree can be considered as an individual classifier.

Sequential Minimal Optimization (SMO) implements the support vector machines (SVMs) algorithm. The purpose of the SVMs is to find a hyperplane that plays a decision boundary role and shows two parallel lines of them. The greatest distance between those parallel lines touching class boundaries is defined as the maximum margin hyperplane. When working with non-separable linear data, SVM projects the data from its original space to a new coordinate space, where the examples are separable by a linear decision limit. This process is done by applying a transformation function to the dataset´s attributes. An alternative to this transformation is the use of nucleon's functions, the so-called kernel functions. These functions make it possible to deal with the training phase as well as the classification of new examples when explicit mapping calculations need to be performed [35, 36]. This criterion is used to find the hyperplane that provides the most robust separation by obtaining the greatest possible discrimination between classes. A quadratic optimization model is used to find such a hyperplane. The main parameter to be adjusted by the algorithm in the SVM classifier refers to the type of kernel functions used. While SMO is one of the fastest techniques for learning SVMs, it is often slow to converge to a solution, particularly with noisy data [37].

The results obtained by classification models were compared to the results of Non-Parametric Multivariate Analysis Of Variance (NPMANOVA), Kruskal Wallis (NPANOVA) and Robusta Regression [28, 38]. The analyzes were performed using Python (Version 3.10.1), SAS (Version 8.2.0.1201) and default setting for Weka (Waikato Environment for Knowledge Analysis, Version 3.8) programs. TABLE 1 shows the main parameters used for the SML implemented in the Weka GUI. Optimal sample size for machine learning was estimated using Monte Carlo simulation [39] and nonlinear regression to obtain 99% classification accuracy [40].

Table 1: Main configuration from each Supervised Machine Learning implemented in the Weka GUI

MLP	
Activation function	Sigmoid
Optimization function	Stochastic gradient descent
Learning rate	0.3
Momentum	0.2
Epochs	500
Batch size	100
Hidden layer	1
Nodes in hidden layer	13
RF	
Maximum depth	None
Tree in forest	100
Batch size	100
RT	
Batch size	100
Maximum depth	None
Minimum variance proportion	0.001
Minimum variance proportion	0.001
K value	2
Minimum instances in a leaf	1
SMO	
Kernel exponent	1
Kernel C	25007
Batch size	100
Calibrator	Logistic

3. RESULTS AND DISCUSSION

The contribution of the analyzed attributes for the separation of coffee from different countries was based on the Chi-square value obtained by the Kruskal Wallis test. All attributes had a statistically significant importance to differentiate between classes (p value < 0.0001).

The attributes that most contributed to the discrimination of country of origin were Body (Chi-square = 308.1) and Total Cup Points (Chi-square = 262.2), followed by Cupper Points (Chi-square = 231.3), Acidity (Chi-square = 230.7), Aftertaste (Chi-square = 208.4), Flavor (Chi-square = 198.7), Aroma (Chi-square = 191.7), Balance (Chi-square = 112.9), Sweetness (Chi-square = 105.1) and Uniformity (Chi-square = 82.2), in that order. When the 1191 original records of Arabica coffee samples from 15 producing countries were analyzed using Cluster Analysis by Hierarchical Agglomerative clustering by the Average method [41], the similarity between the characteristics of coffee from different countries was observed (FIGURE 2).

Based on the quality attributes analyzed, the producing countries were separated into two groups, the first comprising Ethiopia, Kenya and Uganda, which have the best quality scores (FIGURE 2)

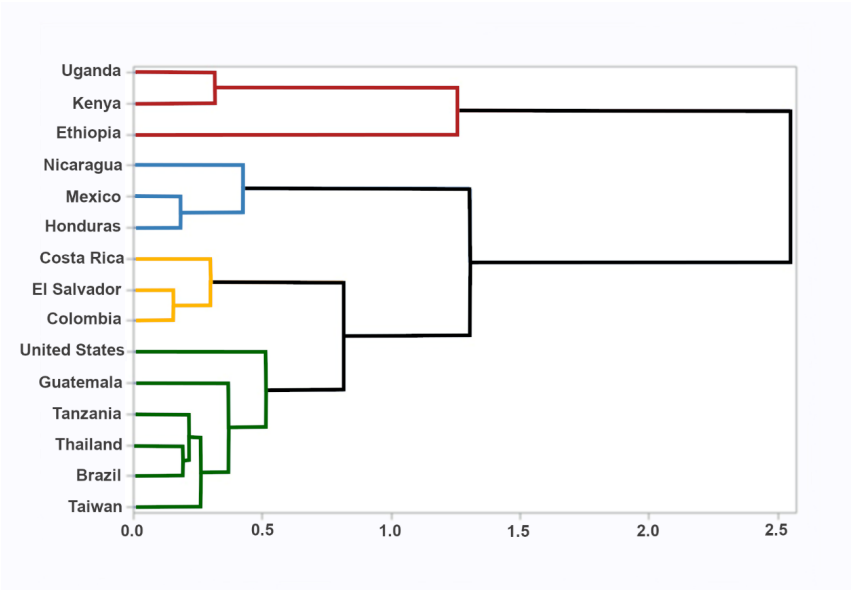


Figure 2: Cluster analysis of coffee quality attributes from 15 producing countries.

and the second with the 12 remaining countries. In this second group, there is a tendency to have three subgroups, the first formed by Nicaragua, Mexico and Honduras, the second by Costa Rica, Salvador and Colombia, and the third by Brazil, Guatemala, United States, Taiwan, Tanzania and Thailand. This trend was confirmed through the Elbow Method, from which it was obtained that the ideal for our database would be the formation of 4 distinct clusters.

The result of the non-parametric multivariate analysis of variance (NPMANOVA) was similar to that obtained from the univariate approach (NPANOVA), for the variable Total Cup Points (p value < 0.0001). FIGURE 3 shows the Box Plot of Total Cup Point attributes for the 15 producing countries. The best performance was seen for countries of the African continent, Ethiopia and Kenya, which had more than 3/4 of the records in quartile 1 (above 80% of the overall data) and Uganda more than 1/2 (quartile 2). The worst performances were observed for Nicaragua and Mexico, with the first quartile below the 20^{th} percentile.

Although a statistically significant difference was observed in the discrimination of producing countries, when implementing MLP, RF, RT and SMO algorithms the classification models obtained were not more accurate than 50%. This low accuracy may be due to the coffee variety, an attribute not included in the original model, as the primary objective was focused on classification of countries. In addition, the database had unfilled records of this important variable. Using factorial ANOVA test, statistically significant interactions were obtained between the country of origin and the variety for all combinations whose degrees of freedom allowed the test to be carried out. FIGURE 4 shows the number of observations for the most representative variety for each country. The more quality-oriented varieties, such as Typica, Bourbon and Caturra, are grown in all the main coffee producing regions in the world. Typica is one of the most iconic coffee varieties in the world, originating from the birthplace of Arabica coffee in Ethiopia. Bourbon and Typica are an integral part of the coffee variety's family tree. They gave rise to popular varieties of coffee, such as Mundo Novo (natural hybrid of Typica and Bourbon), Caturra (natural muta-

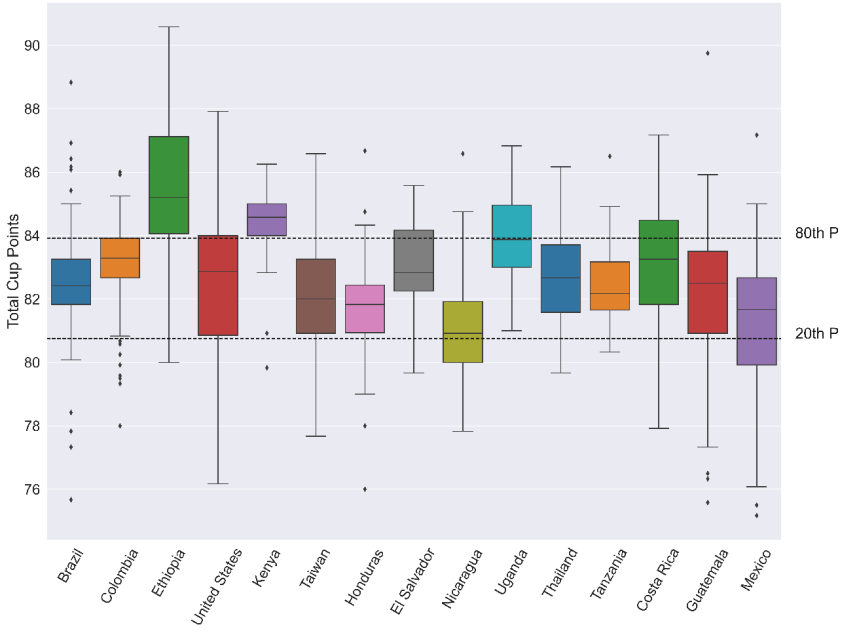
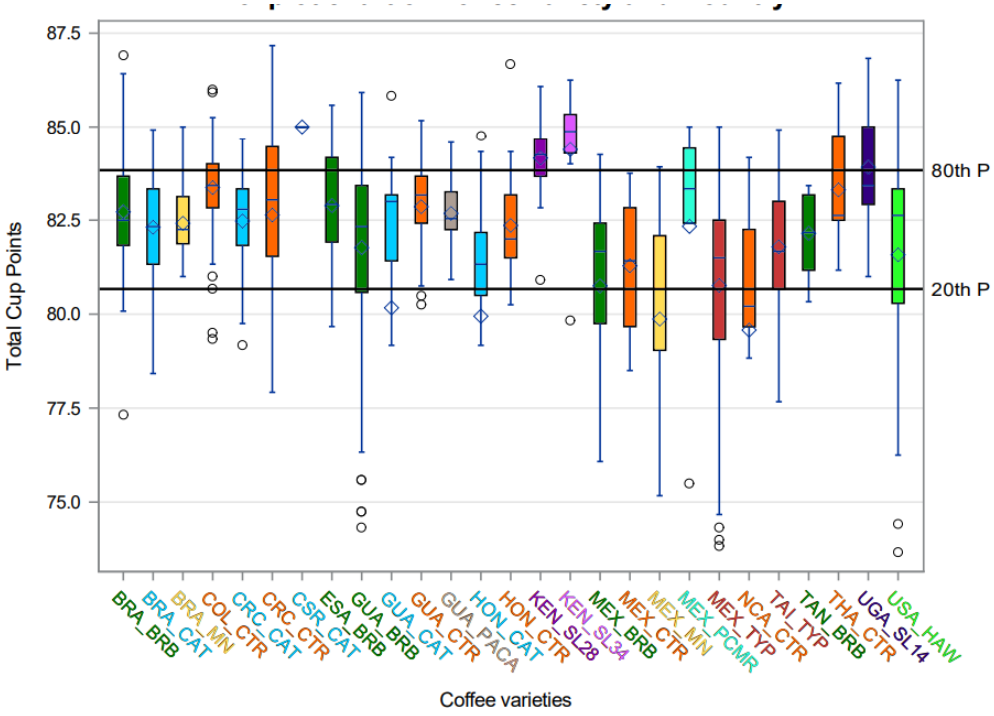


Figure 3: Boxplots of total cup points with percentiles for 15 producing countries.

tion of Bourbon), Catuaí (hybrid of Caturra and Mundo Novo), Maragogipe (natural mutation of Typica), Pacas (natural mutation of Bourbon), Pacamara (hybrid of Pacas and Maragogipe). When considering the varieties for each country, Kenya and Uganda maintained the highest quality values (FIGURE 2).

Bourbon is found in 5 countries (Brazil, El Salvador, Guatemala, Mexico and Tanzania), Caturra in 7 countries (Colombia, Costa Rica, Guatemala, Honduras, Mexico, Nicaragua and Thailand) and Typica in 2 countries (Mexico and Thailand). The importance of the variety as a quality attribute was evidenced for Mexico, which presented samples for 5 different varieties: Bourbon (n = 35), Caturra (n = 19), Mundo Novo (n = 12), Pacamara (n = 6) and Typica (n = 137). The Pacamara variety stood out from the others, achieving high scores in the total cup points (above 80^{th} percentile) while the other four had lower values (below 20^{th} percentile). The importance of variety as a coffee quality attribute reinforces the need to be included in future databases for specialty coffee discrimination.

Seven of the fifteen countries were selected in an attempt to minimize the effect of varieties as well as to consider the commercial interest of unequivocal discrimination of the country of origin: Brazil, the largest producer of Arabica coffee; Colombia, the second largest producer of Arabica coffee; United States, the biggest importer of coffee exported by Brazil and Colombia; Ethiopia, third largest producer of Arabica coffee; Honduras, fourth largest producer of Arabica coffee; Kenya, a producer of high-quality Arabica coffee sought by third parties to blend with other varieties; and Taiwan, a producer of Arabica coffee and one of the main export markets for USA coffee. Statistically significant differences (p value < 0.001) for pairwise comparisons were observed except for Ethiopia



	Country	Variety	No.
BRA_BRB	Brazil	Bourbon	73
BRA_CAT	Brazil	Catuai	19
BRA_MN	Brazil	Mundo Novo	20
COL_CTR	Colombia	Caturra	129
CRC_CAT	Costa Rica	Catuai	15
CRC_CTR	Costa Rica	Caturra	28
ESA_BRB	El Salvador	Bourbon	13
GUA_BRB	Guatemala	Bourbon	119
GUA_CAT	Guatemala	Catuai	9
GUA_CTR	Guatemala	Caturra	29
GUA_PACA	Guatemala	Pacas	6
HON_CAT	Honduras	Catuai	21
HON_CTR	Honduras	Caturra	22
KEN_SL28	Kenya	SL28	14
KEN_SL34	Kenya	SL34	8
MEX_BRB	Mexico	Bourbon	35
MEX_CTR	Mexico	Caturra	19
MEX_MN	Mexico	Mundo Novo	12
MEX_PCMR	Mexico	Pacamara	6
MEX_TYP	Mexico	Typica	137
NCA_CTR	Nicaragua	Caturra	12
TAI_TYP	Taiwan	Typica	59
TAN_BRB	Tanzania	Bourbon	9
THA_CTR	Thailand	Caturra	6
UGA_SLl4	Uganda	SL14	17
USA_HAW	USA	Hawailian	44

Figure 4: Boxplots of total cup points with percentiles considering coffee varieties in each of 15 producing countries.

and Kenya (p value = 0.0989). Cluster Analysis showed that the distance between them was greater than the distance between Brazil, Colombia, United States and Taiwan.

A bootstrap resampling technique was applied and the data used to generate canonical functions of quadratic discriminant analysis. The obtained ellipsoids indicated the multivariate distribution with no overlapping and different volume, depending on the variance and covariance matrix of each country. In the Linear Discriminant Analysis, the ellipsoids have a consensus distribution, reinforcing the need for use of quadratic discriminant analysis.

Classification algorithms of Supervised Machine Learning and Bootstrapping resampling method provided results with higher accuracy than those obtained with the original dataset. TABLE 2 shows a summary of the classification performance of each algorithm. For all the models, a Kappa statistic was greater than 0.95, considered almost perfect according to Landis and Koch [42]. The lower the mean absolute error (MSE), the root mean squared error (RMSE), the relative absolute error (RSE) and the root relative squared error (RRSE) prediction indicators, the better and more accurate the predictor will be [43].

The general accuracy for classification models implementing RF, SMO, MLP and RT were 98.4%, 97.8%, 95.7% and 94.9%, respectively. Considering the general classification accuracy and the performance indicators, the best classification algorithm was RF, followed by MLP, SMO and RT. TABLEs 3 and 4 show the confusion matrix obtained for RF and SMO, respectively.

Table 2: Summary of the indicators of the prediction performance of each classification algorithm.

	MLP	**RF**	**RT[a]**	**SMO[b]**
Correctly Classified Instances	2928	2953	2846	2934
Incorrectly Classified Instances	72	47	154	66
Kappa Statistic	0.97	0.98	0.95	0.98
Mean Absolute Error	0.0052	0.0093	0.0068	0.1156
Root Mean Squared Error	0.0513	0.0484	0.0827	0.2346
Relative Absolute Error	4.20 %	7.45 %	5.50 %	92.88 %
Root Relative Squared Error	20.58 %	19.39 %	33.17 %	94.06 %
Total Number of Instances	3000			

[a] Size of the tree: 247; [b] Number of kernel evaluations: 1420 (67.786% cached).

Table 3: Confusion matrix obtained implementing RF algorithm.

a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***	***n***	***o***	classified as
195	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*4*	*0*	*1*	*0*	*0*	***a = Brazil***
0	*198*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2*	*0*	*0*	***b = Colombia***
0	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***c = Ethiopia***
0	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***d = United States***
0	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***e = Kenya***
0	*0*	*0*	*0*	*0*	*199*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	***f = Taiwan***
0	*0*	*0*	*0*	*0*	*0*	*188*	*0*	*11*	*0*	*0*	*0*	*0*	*0*	*1*	***g = Honduras***
0	*0*	*0*	*0*	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***h = El Salvador***
0	*0*	*0*	*0*	*0*	*0*	*12*	*0*	*188*	*0*	*0*	*0*	*0*	*0*	*0*	***i = Nicaragua***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	***j = Uganda***
5	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*193*	*0*	*2*	*0*	*0*	***k = Thailand***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*199*	*0*	*1*	*0*	***l = Tanzania***
0	*3*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*2*	*0*	*194*	*0*	*0*	***m = Costa Rica***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*199*	*0*	***n = Guatemala***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	***o = Mexico***

Table 4: Confusion matrix obtained implementing SMO algorithm.

a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***	***n***	***o***	classified as
200	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***a = Brazil***
0	*199*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	***b = Colombia***
0	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***c = Ethiopia***
0	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***d = United States***
0	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***e = Kenya***
0	*0*	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***f = Taiwan***
0	*0*	*0*	*0*	*0*	*0*	*185*	*0*	*5*	*0*	*0*	*0*	*0*	*0*	*10*	***g = Honduras***
0	*15*	*0*	*0*	*0*	*0*	*0*	*184*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	***h = El Salvador***
0	*0*	*0*	*0*	*0*	*0*	*19*	*0*	*177*	*0*	*0*	*0*	*0*	*0*	*4*	***i = Nicaragua***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	*0*	*0*	*0*	*0*	*0*	***j = Uganda***
6	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*191*	*1*	*2*	*0*	*0*	***k = Thailand***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	*0*	*0*	*0*	***l = Tanzania***
0	*2*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*198*	*0*	*0*	***m = Costa Rica***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	*0*	***n = Guatemala***
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*200*	***o = Mexico***

The models based on Random Forest (RF) and Sequential Minimal Optimization (SMO) achieved classification performance with 100% accuracy, precision, sensitivity and specificity, respectively, for coffees produced in Ethiopia, United States, Kenya and Uganda, and for coffees produced in Ethiopia, United States, Kenya, Taiwan, Uganda and Guatemala. The major misclassification by implementing RF and SMO were between samples of Honduras and Nicaragua, with samples of Nicaragua classified as Honduras (RF=12 and SMO=19) and samples of Honduras classified as Nicaragua (RF=11 and SMO=5). According to SMO, 15 samples of El Salvador were misclassified as Colombia. The sample size should have approximately 25,500 observations to discriminate with 99% classification accuracy the Arabica coffee geographical origins.

Supervised Machine Learning tools for Regression were used to understand the relationship between each of the attributes and Total Cup Points. Due to the effect induced by the Country Factor, they were not considered in the model. Brazil, the largest producer of Arabica coffee of distinct varieties, was chosen to model these relationships. The Linear Regression model achieved a correlation coefficient of 0.979 and a relative absolute error of 15%. The model described in Equation 1 explains 97.9% of the Total Cup Points variability when Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Clean Cup, Sweetness and Cupper Points parameters vary between the limits of the dataset.

$$\textit{Total Cup Points} = 0.8727*\textit{Aroma} + 1.1988*\textit{Flavor} + 0.8029*\textit{Aftertaste} + 1.2585*\textit{Acidity} + 1.0366*\textit{Body} + 1.283*\textit{Balance} + 1.5263*\textit{Clean Cup} + 1.1746*\textit{Sweetness} + 1.1107*\textit{Cupper Points} - 1.3016 \quad (1)$$

Moisture was the only attribute not selected by the model, indicating that its variability is not significantly affecting the quality of Arabica coffee (Total Cup Points). The parameters that most influence the quality of the coffee, with variable coefficients greater than 1, were Clean Cup, followed by Balance, Acidity, Flavor, Sweetness, Cupper Points and Body. The results obtained by Robust Regression are shown in Equation 2.

$$\textit{Total Cup Points} = 0.9901*\textit{Aroma} + 10.060*\textit{Flavor} + 10.028*\textit{Aftertaste} + 10.056*\textit{Acidity} + 10.010*\textit{Body} + 0.9976*\textit{Balance} + 19.976*\textit{Clean Cup} + 0.9868*\textit{Sweetness} + 10.100*\textit{Cupper Points} + 0.0276*\textit{Moisture} - 0.1290 \quad (2)$$

R^2 value for Robust Multiple Regression model was 0.8244, lower than the R^2 for Linear Regression, nevertheless, a low residual scale (0.0088) was obtained. The only statistically non-significant attribute was moisture (p value < 0.052). All other attributes were highly significant (p value < 0.0001). As for the Linear Regression, the Robust Multiple Regression indicated the Clean Cup as the most important variable for Total Cup Points. The results of Linear Regression and Robust Multiple Regression obtained by bootstrapping resampling method were corroborated by multivariate analysis, estimating the residual correlation and removing the influence of Country Factor for the original dataset. Residual correlations were highly significant (p value < 0.0001) and positive for Total Cup Points, except for the variable Moisture, which showed a negative (r = - 0.058) and significant (p value < 0.047) correlation.

4. CONCLUSION

Traceability of Arabica coffee by country of origin is feasible using quality attributes, machine learning, robust multivariate and univariate data science. The number of records in the dataset (1191) was insufficient to classify the country of origin with high accuracy, and should be about 25 times higher. The bootstrapping resampling and classification models provided discrimination of producing countries with up to 98% accuracy. The coffee variety attribute showed its importance for the discrimination of coffee quality and can be considered a predictor variable to increase the classification accuracy. The best classification algorithm was Random Forest, with the highest classification accuracy (98 %), lowest root mean squared error (0.0484) and smallest root relative squared error (19.39 %).

5. CONFLICT OF INTEREST

We have no conflicts of interest to disclose.

6. ETHICAL GUIDELINE

We made not any examinations on a human or living creatures.

7. ACKNOWLEDGEMENT

This research was supported by the Research Center in Technology and Innovation for Agriculture Sustainability – USP 2012.1.17654.1.0 and the Nucleus for Technological Innovation in Metrology and Quality in Agriculture - CENA/USP, CNPq 420799/2013-3.

References

[1] Peters EJ. ”A rich and tantalizing brew: a history of how coffee connected the world, by Jeanette M. Fregulia, Fayetteville, University of Arkansas, 2019, 193 pp., ISBN 978-1682260876. Uncommon grounds: the history of coffee and how it transformed our world, by Mark Pendergrast, New York, Basic Books, 2019, 480 pp., ISBN 978-1541699380.” Food, Culture & Society. 2020;23:455-457.

[2] https://apps.fas.usda.gov/psdonline/circulars/coffee.pdf.

[3] Ahmed S, Brinkley S, Smith E, Sela A, Theisen M, et. Al. ”Climate Change and Coffee Quality: Systematic Review on the Effects of Environmental and Management Variation on Secondary Metabolites and Sensory Attributes of Coffea arabica and Coffea canephora.” Frontiers in Plant Science. 2021;12.

[4] Dos Santos DG, Coelho CCDS, Ferreira ABR, Freitas-Silva O. ”Brazilian Coffee Production and the Future Microbiome and Mycotoxin Profile Considering the Climate Change Scenario.” Microorganisms. 2021;9:858.

[5] Pierrot J, Giovannucci D, Kasterine A. ”Trends in the trade of certified coffees.” International Trade Centre Technical Paper, 2010.

[6] De Queiroz VT, Azevedo MM, Da Silva Quadros IP, Costa AV, Do Amaral Aa, Dos Santos, Gmada, Juvanhol, Rs, De Almeida Telles, La, Dos Santos, Ar. ”Environmental risk assessment for sustainable pesticide use in coffee production.” Journal of Contaminant Hydrology. 2018;219:18-27.

[7] Gutiérrez-Guzmán N, Cortés-Cabezas A, Chambers Iv E. ”A novel tasting platform for sensory analysis of specialty coffee.” Coffee Science. 2018;13:401 - 409.

[8] https://worldcoffeeresearch.org/resources/sensory-lexicon.

[9] Adnan A, Naumann M, Mörlein D, Pawelzik E. ”Reliable Discrimination of Green Coffee Beans Species: A Comparison of UV-Vis-Based Determination of Caffeine and Chlorogenic Acid with Non-Targeted Near-Infrared Spectroscopy.Foods.2020;9:788.

[10] Bertrand B, Villarreal D, Laffargue A, Posada H, Lashermes P, et. al. ”Comparison of the Effectiveness of Fatty Acids, Chlorogenic Acids, and Elements for the Chemometric Discrimination of Coffee (Coffea arabica L.) Varieties and Growing Origins.” Journal of Agricultural and Food Chemistry 2008;56:2273-2280.

[11] Núñez N, Saurina J, Núñez O. ”Authenticity Assessment and Fraud Quantitation of Coffee Adulterated with Chicory, Barley, and Flours by Untargeted HPLC-UV-FLD Fingerprinting and Chemometrics.” Foods. 2021;10:840.

[12] Alcantara G, Dresch D, Melchert W R. ”Use of non-volatile compounds for the classification of specialty and traditional Brazilian coffees using principal component analysis.” Food Chemistry. 2021; 360:130088.

[13] De Araújo TKL, Nóbrega RO, Fernandes DDDS, De Araújo MCU, Diniz PHGD, et. al. ”Non-destructive authentication of Gourmet ground roasted coffees using NIR spectroscopy and digital images.” Food Chemistry. 2021;364:130452.

[14] Agnoletti BZ, Folli GS, Pereira LL, Pinheiro PF, Guarçoni RC, Da Silva Oliveira EC, Filgueiras PR. ”Multivariate calibration applied to study of volatile predictors of arabica coffee quality.” Food Chemistry. 2022;367:130679.

[15] Cheserek JJ, Ngugi K, Muthomi JW, Omondi CO. ”Assessment of Arabusta coffee hybrids [Coffea arabica L. X Tetraploid Robusta (Coffea canephora)] for green bean physical properties and cup quality.” African Journal of Food Science. 2020;14:119-127.

[16] Robert JV, De Gois JS, Rocha RB, Luna AS. ”Direct solid sample analysis using synchronous fluorescence spectroscopy coupled with chemometric tools for the geographical discrimination of coffee samples.” Food Chemistry. 2022;371:131063.

[17] Arana Va, Medina J, Esseiva P, Pazos D, Wist J. ”Classification of Coffee Beans by GC-C-IRMS, GC-MS, and 1H-NMR.” Journal of Analytical Methods in Chemistry. 2016;2016:8564584.

[18] Yang SI, Li C, Mei Y, Liu W, Liu R, Chen W, Han D, Xu K. ”Determination of the Geographical Origin of Coffee Beans Using Terahertz Spectroscopy Combined With Machine Learning Methods.” Frontiers in Nutrition. 2021;8.

[19] Filho VRA, Polito WL, Gomes Neto JA. ”Comparative studies of the sample decomposition of green and roasted coffee for determination of nutrients and data exploratory analysis.” Journal of the Brazilian Chemical Society. 2007;18:47-53.

[20] Pires FDC, Pereira RGFA, Baqueta MR, Valderrama P, Alves Da Rocha R. ”Near-infrared spectroscopy and multivariate calibration as an alternative to the Agtron to predict roasting degrees in coffee beans and ground coffees.” Food Chemistry. 2021;365:130471.

[21] Barbosa RM, Batista BL, Varrique RM, Coelho VA, Campiglia AD, et al. ”The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee.” Food Research International. 2014;61:246-251.

[22] De Nadai Fernandes EA, Tagliaferro FS, Azevedo-Filho A, Bode P. ”Organic coffee discrimination with INAA and data mining/KDD techniques: new perspectives for coffee trade.” Accreditation and Quality Assurance. 2002;7:378-387.

[23] https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi.

[24] https://database.coffeeinstitute.org/.

[25] https://arxiv.org/pdf/2108.06624.pdf

[26] Costa WGD, Barbosa IDP, De Souza JE, Cruz CD, Nascimento M, et al. ”Machine learning and statistics to qualify environments through multi-traits in Coffea arabica.” PLOS ONE. 2021;16:e0245298.

[27] Refaeilzadeh P, Tang L, Liu H. Cross-Validation, in Encyclopedia of Database Systems, L. Liu and M.T. Özsu, Editors. Springer New York: New York, NY. 2016:1-7.

[28] De Nadai Fernandes EA., Sarriés G.A., Bacchi MA, Mazola YT, Gonzaga CL, et. al. ”Trace elements and machine learning for Brazilian beef traceability.” Food Chemistry. 2020;333:127462.

[29] https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf

[30] Gardner MW, Dorling SR. ”Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences.” Atmospheric Environment. 1998;32:2627-2636.

[31] DE Mello RF, Ponti MA. Machine Learning: A Practical Approach on the Statistical Learning Theory. 1 ed. 2018: Springer International Publishing.

[32] Bel L, Allard D, Laurent JM, Cheddadi R, Bar-Hen A. CART algorithm for spatial data: Application to environmental and ecological data. Computational Statistics & Data Analysis. 2009;53:3082-3093.

[33] Breiman L. ”Random Forests.” Machine Learning. 2001;45:5-32.

[34] Liaw A. Wiener M. ”Classification and regression by Random Forest.” R news. 2002;2:18-22.

[35] XU Y, Zomer S, Brereton RG. ”Support vector machines: a recent method for classification in chemometrics.” Critical Reviews in Analytical Chemistry. 2006;36:177-188.

[36] Maione C. ”Mineração de dados para o reconhecimento da origem e do tipo de alimentos e outras substâncias com base em sua composição química”. Mestrado em Computação, Universidade Federal de Goiás, Goiânia. 2016

[37] KeerthI SS, Shevade SK, Bhattacharyya C, Murthy KRK. ”Improvements to Platt’s SMO algorithm for SVM classifier design.” Neural Computation. 2001;13:637-649.

[38] Fernandes EADN, Mazola YT, Sarriés GA, Bacchi MA, Bode P, Gonzaga CL, Sarriés SR. ”Discriminating Beef Producing Countries By Multi-Element Analysis And Machine Learning.” Advances in Artificial Intelligence and Machine Learning.2021;1:1-11.

[39] Byrd RH, Chin GM, Nocedal J, WU Y. ”Sample size selection in optimization methods for machine learning.” Mathematical Programming. 2012;134:127-155.

[40] https://arxiv.org/pdf/1109.2378.pdf

[41] Syakur M, Khotimah B., Rochman E, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. in IOP Conference Series: Materials Science and Engineering. 2018. IOP Publishing.

[42] Landis JR, Koch GG. ”The Measurement of Observer Agreement for Categorical Data.” Biometrics. 1977;33:159-174.

[43] Guo H, Yin J, Zhao J, Yao L, Xia X, Luo H. ”An Ensemble Learning for Predicting Breakdown Field Strength of Polyimide Nanocomposite Films.” Journal of Nanomaterials. 2015;2015:950943.