Research, Society and Development, v. 10, n. 5, e13110514732, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i5.14732

# Classification of specialty coffees using machine learning techniques

## Classificação de cafés especiais usando técnicas de aprendizado de máquina

## Clasificación de cafés especiales utilizando técnicas de aprendizaje automático

Received: 04/06/2021 | Reviewed: 04/12/2021 | Accept: 04/16/2021 | Published: 01/05/2021

**Paulo César Ossani**
ORCID: https://orcid.org/0000-0002-6617-8085
State University of Maringá, Brazil
E-mail: ossanipc@hotmail.com
**Diogo Francisco Rossoni**
ORCID: https://orcid.org/0000-0001-6337-6628
State University of Maringá, Brazil
E-mail: diogo.rossoni@gmail.com
**Marcelo Ângelo Cirillo**
ORCID: https://orcid.org/0000-0003-2026-6802
Federal University of Lavras, Brazil
E-mail: macufla@dex.ufla.br
**Flávio Meira Borém**
ORCID: https://orcid.org/0000-0002-6560-8792
Federal University of Lavras, Brazil
E-mail: flavioborem@deg.ufla.br

**Abstract**
Specialty coffees have a big importance in the economic scenario, and its sensory quality is appreciated by the productive sector and by the market. Researches have been constantly carried out in the search for better blends in order to add value and differentiate prices according to the product quality. To accomplish that, new methodologies must be explored, taking into consideration factors that might differentiate the particularities of each consumer and/or product. Thus, this article suggests the use of the machine learning technique in the construction of supervised classification and identification models. In a sensory evaluation test for consumer acceptance using four classes of specialty coffees, applied to four groups of trained and untrained consumers, features such as flavor, body, sweetness and general grade were evaluated. The use of machine learning is viable because it allows the classification and identification of specialty coffees produced in different altitudes and different processing methods.
**Keywords:** Supervised classification; Classification models; Sensory analysis.

**Resumo**
Os cafés especiais têm grande importância no cenário econômico, e sua qualidade sensorial é apreciada pelo setor produtivo e pelo mercado. Pesquisas têm sido constantemente realizadas na busca por melhores misturas a fim de agregar valor e diferenciar preços de acordo com a qualidade do produto. Para isso, novas metodologias devem ser exploradas, levando em consideração fatores que possam diferenciar as particularidades de cada consumidor e/ou produto. Assim, este artigo sugere o uso da técnica de machine learning na construção de modelos de classificação e identificação supervisionados. Em um teste de avaliação sensorial para aceitação do consumidor utilizando quatro classes de cafés especiais, aplicados a quatro grupos de consumidores treinados e não treinados, foram avaliadas características como sabor, corpo, doçura e grau geral. O uso de machine learning é viável porque permite a classificação e identificação de cafés especiais produzidos em diferentes altitudes e diferentes métodos de processamento.
**Palavras-chave:** Classificação supervisionada; Modelos de classificação; Análise sensorial.

**Resumen**
Los cafés especiales son de gran importancia en el escenario económico, y su calidad sensorial es apreciada por el sector productivo y el mercado. La investigación se ha llevado a cabo constantemente en la búsqueda de mejores mezclas con el fin de añadir valor y diferenciar los precios de acuerdo con la calidad del producto. Para ello, deben explorarse nuevas metodologías, teniendo en cuenta factores que puedan diferenciar las particularidades de cada consumidor y/o producto. Por lo tanto, este artículo sugiere el uso de la técnica de aprendizaje automático en la construcción de modelos supervisados de clasificación e identificación. En una prueba de evaluación sensorial para la aceptación del consumidor utilizando cuatro clases de cafés especiales, aplicadas a cuatro grupos de consumidores capacitados y no entrenados, se evaluaron características como sabor, cuerpo, dulzura y grado general. El uso del aprendizaje automático es factible porque permite la clasificación e identificación de cafés especiales producidos a diferentes altitudes y diferentes métodos de procesamiento.
**Palabras clave:** Clasificación supervisada; Modelos de clasificación; Análisis sensorial.

## 1. Introduction

The coffee business holds a big social importance due to its capacity of creating jobs and acting in the Brazilian socioeconomic development. Also, the domestic consumption is growing higher and higher (Fehr, et al., 2012).

The coffee consumer in Brazil changed is consumption habits and acquired new perceptions regarding the beverage. Thus, new strategies aim at the appreciation of the product with remarkable attributes with tangible and intangible aspects. Therefore, competition happens not only through prices but also through products with innovative features (Nicoleli & Moller, 2006).

The differentiated coffee segment is the one with the highest growth, and the purchase of the product is connected to the attributes brand and flavor which are connected to former experiences inherent to the sensory memory that characterizes each consumer associating loyalty to brand to favorite flavor. Thus, there are evidences that Brazilian consumers are ready to acquire quality coffees once there are differences between their preferences and the coffee segments. These characteristics should be attended through marketing strategies that involve differentiation standards which would increase the quality adding value to the consumer's satisfaction (Spers, et al., 2004). Thus, the sensory study of the consumers is an important tool for the identification of the motivation in the processes of coffee purchase in the different segments of this market.

The preferences and the acceptance tests must be considered in a sensory test focusing on the evaluation of the taster in order to differentiate the sensory quality of a product when compared to others. Some outside factors are inherent to the formation of the sensory panel such as individual preferences, panel training, and taster experience which might cause statistical problems coming from errors in measurement when filling the sensory evaluation sheet, or even during data analysis (Ossani, et al., 2017).

According to Figueiredo et al. (2018) there is a growing participation and appreciation of specialty coffees in the international market. A study was carried out with the Bourbon genotypes in different environments relating to the chemical composition of the grains with their sensory profile. It was observed that the genotypes Bourbon Amarelo IAC J9 and Bourbon Amarelo / SSP were the most suitable for the production of specialty coffees. In which the caffeine content made it possible to differentiate the coffee in relation to the quality of the drink, with coffees with higher quality having the lowest caffeine content.

Consequently, there is a large field of researches with the aim at and the use of new approaches that add more precise results to the acceptance analyzes and discrimination of sensory quality in coffees.

In the work carried out by Borem et al. (2009), it was used techniques of logistic regression and correspondence analysis in the environmental aspects such as latitude, longitude, altitude and slope as well as coffee varieties and processing methods in consecutive harvests with the objective of setting the sensory quality of the cultivated coffee. The results suggest that the quality did not correspond to the sample discrimination between the direction of the slope face and the sensory profile of the coffee.

In the work presented by Liska et al. (2015), it was used Fisher's conventional linear discriminant analysis (LDA) and the discriminant analysis via boosting algorithm (Adaboost) as a proposal for a classification rule to discriminate trained and untrained tasters. The authors concluded that the boosting method applied to the discriminant analysis show a higher sensibility rate in the trained panel.

The Multiple Factor Analysis for Contingency Tables (MFACT) was used by Ossani et al. (2017) using categorized data obtained from sensory experiments carried out with different consumer groups investigating similarities among four specialty coffees. The use of the technique was viable since it allowed the discrimination of specialty coffees produced in different environments (altitudes) and processing taking into consideration the heterogeneity of the consumers involved in the sensory analysis.

Unsupervised classification techniques were used by Ossani *et at.* (2020) in specialty coffees, obtaining groupings that were in line with the original groups, having very satisfactory results in the algorithms used in the process.

The supervised classification and the data identification represented in classes deserve special attention since the factors, generally unknown, but related to the sensory quality, can be identified. Among the many techniques proposed for data classification, machine learning is characterized for allowing the classification of groups of variables with different sizes and distinct nature.

According to Amaral (2016), machine learning is the application of computational techniques in the attempt of trying to find hidden patterns in data in order to produce algorithms capable of making the computers learn and, not only, run algorithms. This technique is closely connected to the statistics and the artificial intelligence and is directly related to data mining.

Classification techniques have been employed in other situations, for example, in Ossani et al. (2020) worked on the supervised classification process of unconventional vegetables obtaining excellent results with few attributes.

In Zamora et al. (2020) used machine learning techniques in the supervised classification process in riparian forests in the areas of influence of the Goitá and Tapacurá reservoirs.

Classification techniques have been employed in other situations, for example, in Ossani et al. (2020) worked on the supervised classification process of unconventional vegetables obtaining excellent results with few attributes. In Zamora et al. (2020) used machine learning techniques in the supervised classification process in riparian forests in the areas of influence of the Goitá and Tapacurá reservoirs. In the work of Neve et al. (2019) used neural networks in the recognition of body gestures, resulting in an accuracy of 87.6%.

There are many supervised data classification techniques covered by machine learning with each one having their own specificities. They can generate different results depending on the inherent structure of the analyzed database. This is easily verified given the previous data classification in order to choose the one that better model the data.

The classification and identification of coffees take an important role in the consumer's choice form a product of better quality that attends to their economic and taste requirements besides allowing a better marketing targeting to the specific segments. Thus, the choice of a good classifier that may model the sensory data accurately to the consumers characteristics becomes an excellent quality tool for the products guaranteeing better results to consumer satisfaction.

The current work was carried out with the objective of proposing the use of the machine learning technique in the construction of algorithms for supervised classification and identification of specialty coffees produced in different processing and altitudes taking into consideration trained and untrained in a sensory analysis experiment.

## 2. Material and Methods

### 2.1 Data description

According to the proposed objectives, it was considered the data referring to a sensory experiment (Ossani, et al., 2017) relating to the acceptance of specialty coffees produced in Serra da Mantiqueira characterized according to the specifications give in Table 1.

**Table 1 -** Description of specialty coffees evaluated in sensory analysis.

| Classes | Genotype | Altitude | Processing |
|---|---|---|---|
| A | Yellow Bourbon | Above 1.200m | Natural |
| B | Acaiá | Below 1.100m | Peeled Cherry |
| C | Acaiá | Below 1.100m | Natural |
| D | Yellow Bourbon | Above 1.200m | Peeled Cherry |

Source: Authors.

In Ossani et al. (2017) is described all methodology used in the process of sensory experiment performed with tasters, and the way coffees were treated for sensory analyses.

The structure of the juxtaposed table considering the sensory grades obtained in the classed for the attributes body, acidity, sweetness and general grade is presented in the layout in Table 2 added by other attributes.

**Table 2 -** Table layout with results of the sensory analysis of the classes of coffees.

| Instance | Quantitative variables | | | | | | Category Variables | | | Cls |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acidity | Body | Sweetness | General Grade | Altitude | Age | Sex | Proc | Groups | |
| 1 | | | | | | | | | | |
| 2 | | | | $x_{iGC}$ | | | | | | |
| ⋮ | | | | | | | | | | |
| 696 | | | | | | | | | | |

Proc = Processing, Groups = 1, 2, 3, 4 (groups of individuals) and Cls = A, B, C, D (Classes of coffees). $x_{iGC}$ = i-th observation (instance) in the group G and coffee C. Source: Authors.

The groups $G = 1$ and 2 were formed by consumers that were trained for the sensory evaluations. They were constituted, respectively, by 52 and 47 individuals while the other groups ($G = 3$ e $4$) were not trained. However, these last individuals were technicians or researchers in the field of coffee researches with 32 and 43 individuals, respectively, in a total of 696 instances (observations) with 174 instances for each coffee class.

The individuals pointed at all the attributes of all the coffee classes with values belonging to the range $[0; 10]$ with 10 as the maximum grade.

### 2.2 Projection pursuit

It is a technique for exploratory analysis of multivariate data that searches low dimension linear projection in high dimension data. Such projections are hit through the optimization of an objective function called projection index. In this work, we will use this technique to find groupings in the analyzed data indicating the existence of separation among the analyzed coffees in the process of supervised classification.

Therefore, according to the number of variables (Table 2), it was applied the indices Legendre and PDA with the purpose of researching the formation of groupings to be detected. The Legendre index is based in the distance $L^2$between the density of the projected data and the standard bivariate normal density. It is built by the inversion of the density through a normal cumulative distribution function with the transformations $y^{\alpha} = 2\phi(z^{\alpha}) - 1$and $y^{\beta} = 2\phi(z^{\beta}) - 1$ in which $\phi$ is the standard normal distribution and using $J$ polynomial terms of Legendre for the expansion (Martinez and Martinez 2007). Meanwhile, the PDA index is based in the penalty of the LDA index being applied in situations with many predictors highly correlated when the classification is necessary. However, the LDA index is obtained through the linear discriminant analysis with the objective of searching for linear projections with the highest separation among classes and the lowest intraclass dispersion (Espezua, et al., 2015).

### 2.3 Classification methods

In order to compare the procedure of reduction of dimension carried out with the projection pursuit (section 2.2), it was taken into consideration the classification methods applied in the data of Table 2.

**Bayes models**

There are many algorithms based on the Bayes rule

$$P(C_i|x) = \frac{P(x|C_i) \cdot P(C_i)}{P(x)} = \frac{P(x|C_i) \cdot P(C_i)}{\sum_{k=1}^{K} P(x|C_k) \cdot P(C_k)} \quad (1)$$

with $C_j$ the j-th class and $x$the instance. The Bayes classifier chooses the class with the highest probability; thus, it chooses $C_j$ if $P(C_j \vee x) = \underset{k}{max} P(C_k \vee x)$; therefore, the class value with the highest index (Alpaydin, 2010). In this work, it is not used the supervised discretization to convert numeric attributes into nominal ones.

The Naive Bayes algorithm considers the non-dependence among attributes used in the construction of the models. It evaluates how much the attribute helps in the classification of the instance building a probability table and based on the training data, the precision values of the numeric estimator are chosen (Alpaydin, 2010).

The *Bayes Net* algorithm is based in graphs to represent the conditional probability relations allowing the data classification (Alpaydin, 2010). In this work, a simple estimator is used to estimate the conditional probability tables once the structure was learned.

There are also algorithms based on bayesian networks that provide excellent results in data classification here characterized by *Naive Bayes Multinomial* and *Naive Bayes Multinomial Updateable* (Alpaydin, 2010).

In this work the *Naive Bayes Multinomial* algorithm ignored the words that do not occur at least in the minimum frequency in the training data. The lemmatization algorithm is also used in in the words.

In the *Naive Bayes Multinomial Updateable* algorithm in this work, it is not used the supervised discretization to convert numeric attributes into nominal ones.

**Function models**

The logistic regression can be used as a very efficient classification algorithm. Thus, it there are $k$classes for $n$instances with $m$attributes, with $\beta$ as the order parameter matrix $m \times (k-1)$, then, the probability for the $j$class, with the exception of the last class, is given by (Landwehr, et al., 2006)

$$P_j(X_i) = \frac{exp(X_i\beta_j)}{1 + \sum_{j=1}^{k-1} exp(X_i\beta_j)}, \quad (2)$$

and the last class will have probability of

$$1 - \sum_{j=1}^{k-1} P_j(X_i) = \frac{1}{1 + \sum_{j=1}^{k-1} exp(X_i\beta_j)}. \quad (3)$$

Support Vector Machine (SVM) is a classification method that creates an optimized vector maximizing the margin between the nearest instances allowing the classification. It minimizes the overfits and supports many attributes. In a high dimension space, the input vector is mapped non-linearly and, in this space, a linear decision surface is built (Keerthi & Shevade, 2001). In this work, it was used the logistic function as calibrator and normalized data.

The multilayered artificial neural network perceptron is used as classification algorithm working as a non-parametric estimator in which perceptron is given by

$$y = \sum_{j=1}^{d} w_j x_j + w_0, \quad (4)$$

with $x_j \in R, j = 1, \cdots, d$ the network input, $w_j$ the synaptic weight; $y$the output and $w_0$ the intercept value to generalize the model (Alpaydin, 2010). In this work, it was used a neural network with three layers in the classification process.

**Lazy models**

According to Nicoletti (2005), the algorithms of the Instance Based Leaning (IBL) family are considered as an extension of the Nearest Neighbor (NN) algorithm, outlining some limitations associated to the NN model.

In these algorithms, the instances are represented as points in the n-dimensional space defined by r-attributes that describe them with the training instance stored in the memory.

There are many algorithms based on instances that represent the IBL family. Mitchell (1997) points out that the k-Nearest Neighbor (KNN) algorithm is the most basic method based on instances. This algorithm assumes that all instances correspond to points in the n-dimensional space $R^n$.

The nearest neighbors of an instance are defined based on the euclidean distance in which, given two instances $x_i$ and $x_j$ with $i \neq j$, there is

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{m} (x_{ir} - x_{jr})^2}, \quad (1)$$

with $r$ the r-th attribute of the $x$instance. Other metrics can be used (Nicoletti, 2005). In this work, it was used $k = 1$ nearest neighbors with the euclidean distance in the classification process.

The LWL algorithm is based on instances, but it considers the instances locally in order to classify them using the Naive Bayes algorithm or the linear regression (Frank, et al., 2003).

The Kstar algorithm is, also, a classifier based on instances, but it differs itself because it used a distance function based on entropy (Cleary & Trigg, 1995).

**Rules models**

The Rules models are characterized by the use of rules in the classification of instances, and there are many algorithms based on rules.

The Jrip algorithm, proposed by Cohen (1995), makes use of the propositional rules in the classification process. It was used the value 2 as minimum total weight of instances in a rule.

The Decision Table algorithm used decision tables as hypothesis space in the classification process (Kohavi, 1995). In this work, the research method applied to find good combinations of attributes for the decision table was BestFirst, in which it is possible to research the space of attributes subsets by augmented scale with a setback facility.

According to Frank and Witten (1998), the PART algorithm creates a partial decision tree in each iteration and turns the best leaf into a classification rule. In this work, it was used the minimum description length (MDL) correction when locating divisions in numeric attributes.

The OneR algorithm discretizes the numeric attributes and used the minimum-error attribute for prevision (Holte 1993). The minimum interval sized employed in this work to discretize attributes was 6.

**Tree models**

There are many algorithms based on decision trees that generate excellent classifiers.

The REPTree algorithm generates multiple decision trees in changed iterations based on the information gain with the entropy, and minimizes the error resulting from the variation. Then, it chooses the best out of all the trees generated (Lakshmi, 2015). In this work, it was used the value 2 as the minimum total weight of the instances in a leaf, and the minimum proportion of variance in all the data that must be present in a knot for the division to be carried out in regression trees was 0.001.

The Hoeffding Tree algorithm explores the fact that a small sample might be sufficient in the choice of an attribute, and it also assumes that the distribution of generation of examples do not change over time. The Hoeffding limit quantifies the number of examples necessary to estimate how good an attribute is (Hulten, et al., 2001). In this work, it was employed Naive Bayes adaptive as strategy of prevision of leaf to be used. The number of instances (or total weight of instances) that a leaf must attend between division trials was 200. The limit under which a division will be forced to a tie break was 0.05.

The J48 algorithm creates a binary tree using the decision tree C4.5. Next, the algorithm is applied to each tuple in the database resulting in their classification (Quinlan, 1993). In this work, it was used the MDL correction when locating divisions in numeric attributes.

The Decision Stump algorithm consists in one-level decision trees. The prevision was made based on the value of a single input resource (Oliver & Hand, 1994). In this work, it was used the regression based on the mean squared error or the classification based on entropy.

The Random Forest used a mixture of decision tree predictors in a way that each tree depends on the values of a random vector autonomously and with the same distribution to every tree (Breiman, 2001). In this work, it was used 100 trees in the random forest.

The LMT algorithm uses logistic regression functions in the leaves of the decision trees (Landwehr, et al., 2005). It was considered, in this work, the value of 15 as the minimum number of instances in which a knot is considered for the division.

The Random Tree algorithm considers k attributes randomly chosen in each knot in the construction of the decision tree (Hall, et al., 2009). It was employed the value 1 as the minimum total weight of instances in a leaf, and the minimum proportion of variance in all the data that must be present in a knot for the division to be carried out in regression trees was 0.001.

**Meta models**

The Meta models are characterized by algorithms made of multiple learners that complement themselves so, when combined, they may obtain higher precision since, according to Alpaydin (2010), no algorithm is always the most precise. Therefore, the Meta models improve the performance of the classification algorithms.

The Bagging model uses a voting method to differentiate the classifiers employing training sets slightly different in the training process of the classifiers (Breiman, 1996). In this work, it was used to improve the performance of the REPTree algorithm with 10 iterations carried out.

The AdaBoost model is based in a training set to build a set of classifiers. Since it is a metaheuristic algorithm, it is used to improve the performance of other classifiers (Freund & Schapire, 1996). In this work, it was used to improve the performance of the Decision Stump algorithm carrying out 10 iterations.

As cited by Alpaydin (2010), Stacking is a technique proposed by Wolpert (1992) in which it used a voting methods in which the outputs of the classifiers are combined. In this work, it was used to improve the performance of the Naive Bayes algorithm.

The Random SubSpace algorithm is based in decision trees to build a classifier improving the precision with the increase of complexity (Ho, 1998). In this work, it was used to increase the performance of the REPTree algorithm, and 10 iterations were carried out.

The CV Parameter Selection algorithm used cross validation in parameter selection for any classifier (Kohavi, 1995). In this work it was used to increase the performance of the J48 algorithm.

The Logit Boost algorithm uses logistic regression in the classification process when dealing with multiple classes (Friedman, et al., 2000). In this work, it was used to increase the performance of the Decision Stump algorithm, and 10 iterations were carried out.

In the Classification Via Regression algorithm, the classed are binarized, and a regression algorithm is built for each class value. Then, it used the decision tree in the classification process (Frank, et al., 1998).

**2.4 Procedure for validation of the proposed model**

With the objective of validating the proposed model, and taking into consideration the coffee classes in Table 2, it was adopted the procedure describe in the following steps:

1) it was used the cross-validation method k-fold in which the set of original data was subdivided in $k$subsets. Next, the $k-1$ subsets were used for training and the remaining subset for test. This procedure was repeated $k$ times with each instance using the same number of times for training and test. In this work, it was used $k = 10$. Using bootstrap, the instances for training and test are random samples with substitution.
2) after adjusting the $k$ machine learning models in (1), the validation error rate of the proposed model was given by

$$VE = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_j - \hat{y}_j\right),$$

in which $y_j$ and $\hat{y}_j$ denote, respectively, the values observed and predict of classes for the j-th instance. There is, also, $n_i$ as the number of observations of the k-th test set. It was established that, if the classes were equal, $y_j - \hat{y}_j = 0$. Otherwise, $y_j - \hat{y}_j = 1$.

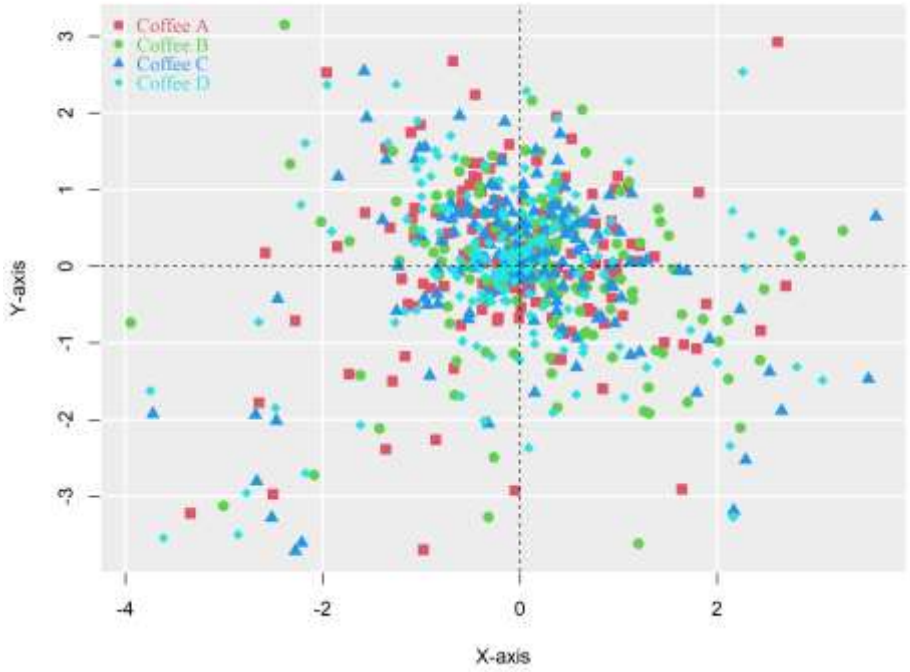
3) next, it was verified it there was a good adjustment considering the validation error rate under 30%.

The supervised classification analyzes were carried out using the software Waikato Environment for Knowledge Analysis (Weka) version 3.9.4 (Hall, et al., 2009).

## 3. Results and Discussions

When using the projection pursuit technique in the quantitative variables (Table 2), using the Legendre index in spherical, and with the grant tour simulated annealing optimization algorithm, through the pack MVar 2.1.4 (Ossani & Cirillo, 2020) of the R software (R Development Core Team 2020), the objective was to verify the presence of grouping. It is possible to observe, in Figure 1, that the data in each class are very disperse. Also, there is not a separation among the coffee classes, that being, there are no grouping formation in the analyzed sample. It is important to highlight that other indexes were used in the search of grouping formations without success.

**Figure 1** - Graph of the results of the projection pursuit in the quantitative data using the Legendre index.




Source: Authors.

When applying the supervised classification techniques (section 2.3) in the quantitative variable (Table 2), it is obtained the Table 3 with the results of the classification error rates.

**Table 3 -** Results of the classification techniques used in the quantitative variables.

| Nº | Classifier | Model | Validation error rate % |
|---|---|---|---|
| 1 | Naive Bayes | Bayes | 48,56 |
| 2 | Bayes Net | Bayes | 45,83 |
| 3 | Naive Bayes Updateable | Bayes | 48,56 |
| 4 | Naive Bayes Multinomial | Bayes | 70,97 |
| 5 | Naive Bayes M. Updateable | Bayes | 70,98 |
| 6 | Logistic | Function | 45,55 |
| 7 | SVM [1] | Function | 47,13 |
| 8 | Multilayer Perceptron | Function | 42,67 |
| 9 | KNN [1] | Lazy | 49,86 |
| 10 | LWL | Lazy | 46,41 |
| 11 | KStar | Lazy | 52,01 |
| 12 | JRip | Rules | 47,27 |
| 13 | Decision Table | Rules | 43,10 |
| 14 | PART | Rules | 44,25 |
| 15 | OneR | Rules | 51,15 |
| 16 | REPTree | Trees | 43,68 |
| 17 | Hoeffding Tree | Trees | 46,55 |
| 18 | Decision Stump | Trees | 51,15 |
| 18 | J48 | Trees | 44,11 |
| 20 | Random Forest | Trees | 46,98 |
| 21 | LMT | Trees | 45,83 |
| 22 | Random Tree | Trees | 49,57 |
| 23 | Bagging | Meta | 43,10 |
| 24 | AdaBoostM1 | Meta | 51,15 |
| 25 | Stacking | Meta | 74,85 |
| 26 | Random SubSpace | Meta | 41,81 |
| 27 | CV Parameter Selection | Meta | 44,11 |
| 28 | Logit Boost | Meta | 45,11 |
| 29 | Classification Via Regression | Meta | 48,56 |

[1]SVM and KNN correspond, respectively, in the WEKA software, to the SMO and IBK classifiers. Source: Authors.

Overall, the classifiers are guided by the inherent characteristics of each class in order to obtain differeces in the moment of classification. If the point cloud of the classes present high degree of intersection (Figure 1), that being, high degree of homogeneity, these differences become hard since the algorithms are mixed in the moment of the classification. This occurred with the quantitative data analyzed, in Table 3, in which it is possible to observe that the cross-validation error rates were high.

Since the validation error rates in Table 3 were high, it is justifiable the search for new forms of classification for data with structures of this nature in which the groups present little variability among the classes, what can be observed in Figure 1.

With the quantitative data of the consumer classes having high homogeneity among themselves, the use of categorical variables might help in the differentiation among classes allowing the classification and identification of different coffee classes, what can be seen in the following section.

## 4. Classification Proposal

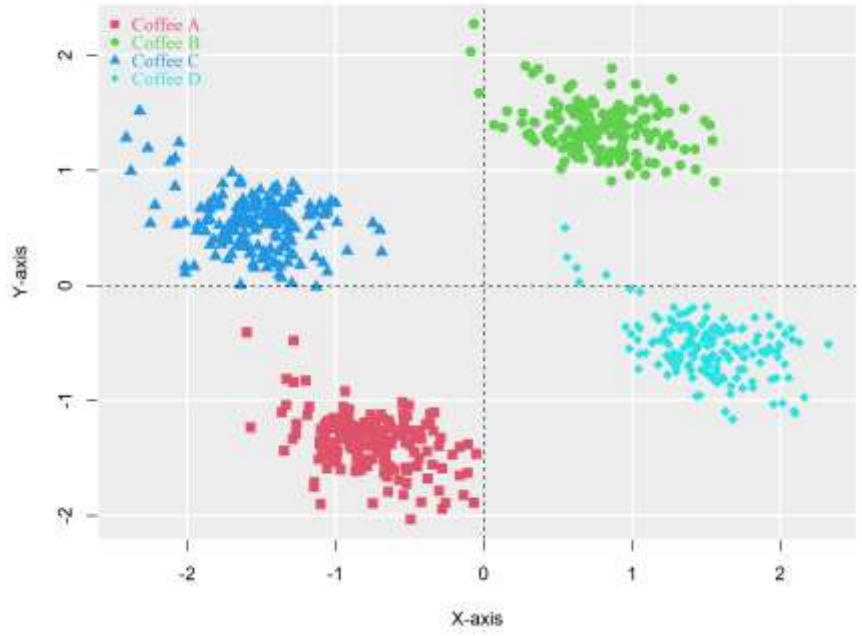
With the objective of outlining the high degree of homogeneity in the classes present in Table 2, and minimizing the cross-validation error rates (Table 3), regarding quantitative variables, an alternative would be the addition of other attributes that allow the differentiation of the classes in the moment of the classification. Here, it is proposed the use of the variable

categories (Table 2) turning into dummy variables in order to outline the high degree of homogeneity and differentiate the referred classes.

Although the sensory attributes are relevant in the discrimination of a coffee, and, since the specialty coffees are of higher quality than the commercial coffees, the grades given by the tasters are not different in the composed groups. Thus, since the coffees are indeed different, there are non-numeric attributes that characterize them in the moment of the sensory research that might be used in the classification. Therefore, the variables "Sex", "Processing", and "Groups are added in the classification.

Based on the use of the dummy variables, it was generated Figure 2 through the projection pursuit technique now using the PDA index, in spherical data, and with the optimization algorithm grant tour simulated annealing through the pack MVar 2.1.4 (Ossani & Cirillo, 2020) of the R software (R Development Core Team, 2020). Other indexes were used in the search of groupings formations. This was the one that better represented the groupings.

**Figure 2** - Graph of the results of projection pursuit in the data using the PDA index.




Source: Authors.

It is possible to observe through Figure 2 that, unlike the data presented in Figure 1, there was a distinction among classes suggesting the existence of algorithms in machine learning capable of classifying and identifying the specialty coffees studies based on the quantitative and qualitative attributes.

With the objective of creating parcimonial models, that being, models with few variables (attributes) capable of explaining the entire variability contained in the model with all the variables, it was applied the variable selection method ReliefFAttributeEval (Kira & Rendell, 1992) implemented in the WEKA software version 3.9.4 (Hall, et al., 2009). The variables general grade, altitude and processing were enough in the classification process of the specialty coffees with the results being presented in Table 4.

**Table 4 -** Results of the classification techniques using quantitative and qualitative techniques.

| Nº | Classifier | Model | Validation error rate % |
|---|---|---|---|
| 1 | Naive Bayes | Bayes | 0,00 |
| 2 | Bayes Net | Bayes | 0,00 |
| 3 | Naive Bayes Updateable | Bayes | 0,00 |
| 4 | Naive Bayes Multinomial | Bayes | 75,57 |
| 5 | Naive Bayes M. Updateable | Bayes | 46,55 |
| 6 | Logistic | Functio | 0,00 |
| 7 | SVM [1] | Functio | 0,00 |
| 8 | Multilayer Perceptron | Functio | 0,00 |
| 9 | KNN [1] | Lazy | 0,00 |
| 10 | LWL | Lazy | 0,00 |
| 11 | KStar | Lazy | 0,00 |
| 12 | JRip | Rules | 0,00 |
| 13 | Decision Table | Rules | 0,14 |
| 14 | PART | Rules | 0,00 |
| 15 | OneR | Rules | 51,15 |
| 16 | REPTree | Trees | 0,00 |
| 17 | Hoeffding Tree | Trees | 0,00 |
| 18 | Decision Stump | Trees | 51,15 |
| 18 | J48 | Trees | 0,00 |
| 20 | Random Forest | Trees | 0,00 |
| 21 | LMT | Trees | 0,00 |
| 22 | Random Tree | Trees | 0,00 |
| 23 | Bagging | Meta | 0,00 |
| 24 | AdaBoostM1 | Meta | 41,38 |
| 25 | Stacking | Meta | 74,85 |
| 26 | Random SubSpace | Meta | 0,29 |
| 27 | CV Parameter Selection | Meta | 0,00 |
| 28 | Logit Boost | Meta | 0,00 |
| 29 | Classification Via Regression | Meta | 0,00 |

[1]SVM and KNN correspond, respectively, in the WEKA software, to the SMO and IBK classifiers. Source: Authors.

In Table 4, it is possible to observe that the use of the qualitative variable processing alongside the quantitative variables general grade and altitude, was efficient in the supervised classification process although some algorithms did not reach a good adjustment what it justified by the inherent specificities. This shows that only quantitative information regarding the data set were not enough in the differentiation of classes with high degree of similarity in its structure in the n-dimensional space in which they are inserted.

Although in Table 4 there are classifiers that obtained validation error rates of 0%, their use is not advisable since, even though they were analyzed via k-fold cross-validation, there is the possibility of overfitting, that being, the model might have an excellent precision in the development environment and a terrible performance in new data.

The classifiers suggested are the ones under 30% and over 0% in the validation error rate. Thus, the Decision Table and Random SubSpace classifiers are the ones best adjusted to the task of classifying these data.

It is important to highlight that there were improvements and declines in the classifiers of the parcimonial models presented in Table 4 when compared to the models presented in Table 3. The results of the algorithms OneR, Decision Stump and Stacking stayed the same while the Naive Bayes Multinomial algorithm worsened the result. With the exclusion of possible overfitting, the rest of the classifiers showed high improvement in the classification error rates.

Most classifiers shown in Table 4 obtained validation error rates under 1% which are excellent results. This shows that classifying and identifying specialty coffees are actions viable through machine learning techniques using only the general grade given by the consumers, trained or untrained, the altitude where they were produced and the processing methods.

The high number of classifiers with excellent results makes clear the intrinsic differentiation of each specialty, adding to the results observer by Borem et al. (2019), Silveira and Pinheiro (2016) and Taveira et al. (2011) which allowed the separation among classes in order to attend the specificities of many classifiers what greatly characterizes these specialty coffees.

According to Silveira and Pinheiro (2016), the factors altitude, slope exposure and fruit color influences the sensory quality of the coffee when analyzed separately or in its interactions (altitude x slope exposure, altitude x fruit color, and slope exposure x fruit color) with the altitude being the major factor to influence the sensory quality of the coffee. In higher altitudes, the coffee producers take long to complete the cycle making the period for grain filling longer allowing higher accumulation of starch in the coffee fruits. Thus, the period for carbohydrate productions becomes sufficient to accumulate substances such as sugars, some acids, and amino-acids, adding to a more pleasant flavor,

In Taveira et al. (2011), there are reports that the altitude and the slope face are empirically known as factors that favor the quality of the coffee. These factors allow the formation of a mild micro-climate, and lower temperatures are pointed as responsible for slowing the speed of fruits maturation allowing higher accumulation of precursors of flavor and aroma.

According to Borem et al. (2019). the sensory properties of the coffees directly depend on the cultivation environment, on the genetic characteristics inherent to the varieties, and on the technology used for post-harvest processing. Besides environmental factors, genetic factors, and factor associated with the handling of the coffee culture, the differences in the quality of coffee beverages are directly associated with the changes in the coffee grains during the different processing stages.

In a study by Benedito et al. (2020) which evaluated the acceptance of coffee by consumers using olfactory sensory analysis, observed that the samples most accepted by consumers are associated with coffees classified as hard and soft.

The peeled coffees present a more desirable acidity when compared to the natural coffees. However, its is important to point out that the time of exposition to the drying conditions of the grains produced via drought is higher when compared to the ones produced via wet what produces irreversible damage to the grains decreasing their physiological quality and changing the beverage (Taveira, et al., 2010).

## 5. Conclusions

In accordance to the proposed objectives and methodology, it is possible to conclude that the machine learning technique is viable to be applied in the supervised classification and identification of specialty coffees.

When working with quantitative data, it was not possible to find good classification models, nor show inherent distinction in each specialty coffee studies. However, the addition of the qualitative variable processing allowed the classification and identification of the coffees studies.

Excellent results were obtained using only the attributes general grade, altitude, and processing with validation error rates under 1% in the classifiers used.

As future research suggests the use only of qualitative variables in the validation process, since its impact in this study was relevant. Another theme of research interest would be the development of other classification methods derived from data dimension reduction techniques such as projection pursuit.

## References

Alpaydin, E. (2010). *Introduction to machine learning.* Adaptive Computation and machine learning Series. MIT Press.

Amaral, F. (2016). *Introdução a ciência de dados: mineração de dados e Big Data*. Rio de Janeiro: Alta Books. 320 p.

Benedito, L. Z., Lima, C. M. G., Silva, J. F. da, Cardoso, D. C., Verruck, S., & Pereira, R. G. F. A. (2020). Acceptance of coffee by different consumer profiles using multivariate statistics. *Research, Society and Development*, 9(6), e102963592. 10.33448/rsd-v9i6.3592.

Borém, F. M., Cirillo, M. A., Alves, A. P. C., Santos, C. M., Liska, G. R., Ramos, M. F., & Lima, R. R. (2019). Coffee sensory quality study based on spatial distribution in the Mantiqueira mountain region of Brazil, *Journal of Sensory Studies*. e12552. 10.1111/joss.12552

Breiman, L. (1996). Bagging predictors. *Machine learning*. 24(2):123-140, 10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Machine learning*. 45(1):5-32.

Cleary, J. G., &Trigg L. E. (1995) K*: An Instance-based Learner Using an Entropic Distance Measure. *In: 12th International Conference on Machine Learning*, 108-114.

Cohen, W. W. (1995). Fast Effective Rule Induction. *In: Twelfth International Conference on machine learning*, 115-123.

Espezua, S., Villanueva, E., Maciel, C. D., & Carvalho, A. (2015). A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets. *Neurocomputing* 149, 767–776, 10.1016/j.neucom.2014.07.057

Fehr, L. C. F., Duarte, A. S. L., Tavares, M., & Reis, E. A. (2012). Análise temporal das variáveis de custos da cultura do café arábica nas principais regiões produtoras do Brasil *Custos e Agronegócio Online*, v. 8, n. 1 – Jan/Mar.

Figueiredo, L. P., Borém, F. M.; Ribeiro, F. C., Giomo, G. S., Malta, M. R., & Taveira, J. H. S. (2018). Sensory analysis and chemical composition of `bourbon' coffees cultivated in different environments. *COFFEE SCIENCE*, 13, 122.

Frank, E., Hall, M., & Pfahringer, B. (2003). Locally Weighted Naive Bayes. *In: 19th Conference in Uncertainty in Artificial Intelligence*, 249-256.

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine learning*. 32(1):63-76, 10.1023/A:1007421302149

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *In: Thirteenth International Conference on machine learning*, San Francisco, 148-156.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). *Additive Logistic Regression: a Statistical View of Boosting*. Stanford University. The Annals of Statistics 2000, 28(2), 337-407.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(8):832-844, 10.1109/34.709601

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning*. 11:63-91, 10.1023/A:1022631118932

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *In: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 97-106, 10.1145/502512.502529

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 13(3):637-649, 10.1162/089976601300014493

Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In: *Ninth International Workshop on Machine Learning*, 249-256, 10.1016/B978-1-55860-247-2.50037-1

Kohavi, R. (1995). The Power of Decision Tables. *In: 8th European Conference on machine learning*, 174-189, 10.1007/3-540-59286-5_57

Kohavi, R. (1995). Wrappers for Performance Enhancement and Oblivious Decision Graphs. *Department of Computer Science*, Stanford University.

Lakshmi, D., C. (2015). Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction. *IJCA Proceedings on International Conference on Communication, Computing and Information Technology*. ICCCMIT 2014(3):30-36.

Landwehr, N., Hall, M., & Frank, E. (2006). *Logistic Model Trees. Kluwer Academic Publishers*. Printed in the Netherlands.

Liska, G. R., Menezes, F. S., Cirillo, M. A., Borem, F. M., Cortez, R. M., & Ribeiro, D. E. (2015). Evaluation of sensory panels of consumers of specialty coffee beverages using the boosting method in discriminant analysis. Semina. *Ciências Agrárias (Online)*, 36, 3671-3679, 10.5433/1679-0359.2015v36n6p3671

Martinez, W. L., & Martinez, A. R. (2007). *Computational Statistics Handbook with MATLAB*, (2th. ed.), Chapman & Hall/CRC, 794 p.

Mitchell, T. M. (1997). *Machine learning*, Mc-Graw Hill, 421p.

Neves, A. das, Okada, H., & Shitsuka, R. (2019). Recognition in Images Using Neural Networks. *Research, Society and Development*, 8(11), e278111470. 10.33448/rsd-v8i11.1470.

Nicoleli, M., & Moller, H. D. (2006). Análise da competitividade dos custos do café orgânico sombreado irrigado. *Custos e Agronegócio Online*, 2(1).

Nicoletti, M. C. (2005). *O modelo de aprendizado de máquina baseado em exemplares: principais características e algoritmos.* EdUFSCar, 61 p.

Oliver, J. J., & Hand, D. (1994). Averaging over decision stumps. *Lecture Notes in Computer Science*, 231–241, 10.1007/3-540-57868-4_61

Ossani, P. C., & Cirillo, M. A. (2020). *MVar: Multivariate Analysis*. URL <https://cran.r-project.org/web/packages/MVar/index.html>. R package version 2.1.4.

Ossani, P. C., de Souza, D. C., Rossoni, D. F., & Resende, L. V. (2020). Machine learning in classification and identification of nonconventional vegetables. *Journal of Food Science*, 85: 4194-4200. 10.1111/1750-3841.15514

Ossani, P. C., Rossoni, D. F., Cirillo, M. Â., & Borém, F. M. (2020). Unsupervised classification of specialty coffees in Homogeneous sensory attributes through machine learning. *Coffee Science*, 15, e151780. 10.25186/cs.v15i.1780

Ossani, P. C., Cirillo, M. A., Borém, F. M., Ribeiro, D. E., & Cortez, R. M. (2017). Qualidade de cafés especiais: uma avaliação sensorial feita com consumidores utilizando a técnica MFACT. *Revista Ciência Agronômica*, 48(1), 92-100. 10.5935/1806-6690.20170010

Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Morgan Kaufmann Publishers*, San Mateo, CA.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. *Vienna: Vienna University of Economics and Business*, 2020. <http:// www.R-project.org/>.

Silveira, A. S., Pinheiro, A. C. T., Ferreira, W. P. M., Silva, L. J., Rufino, J. L. S., & Sakiyama, N. S. (2016). Sensory analysis of specialty coffee from different environmental conditions in the region of Matas de Minas, Minas Gerais, Brazil. *Revista Ceres*, *63*(4), 436-443, 10.1590/0034-737X201663040002

Spers, E. E., Saes, M. S. M., & Souza, M. C. M. (2004). Análise das preferências do consumidor brasileiro de café: um estudo exploratório dos mercados de São Paulo e Belo Horizonte. *RAUSP - Revista de Administração da Universidade de São Paulo*, 39(1), 53-61.

Taveira, J. H., Borém, F. M., Rosa, S. D. V. F., Ribeiro, D. E., Chaves, A. R. C. S., Ferreira, D. A., Ferreira, I. T., & Ribeiro, R. C. (2011). Aspectos fisiológicos de grãos de café produzidos em ambientes variados da micro região da Serra da Mantiqueira. *In: 7º Simpósio de Pesquisa dos Cafés do Brasil*, Araxá. Anais, Epamig.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*. 5:241-259, 10.1016/S0893-6080(05)80023-1

Zamora, V. R. O., Cruz, A. F. da S., Andrade, A. R. S. de, Silva, E. G. da, Andrade, E. K. P. de, Silva, J. D. De S., & Silva, E. T. da. (2020). Supervised classification of riparian forest areas of influence in the Goitá and Tapacurá dams through Spring. *Research, Society and Development*, 9(11), e4829119947. 10.33448/rsd-v9i11.9947.