eISSN: 1984-3909
Coffee Science, e151780, 2020

Doi: https://doi.org/10.25186/cs.v15i.1780

UNSUPERVISED CLASSIFICATION OF SPECIALTY COFFEES IN HOMOGENEOUS SENSORY ATTRIBUTES THROUGH MACHINE LEARNING

Paulo César Ossani[1*], Diogo Francisco Rossoni[1], Marcelo Ângelo Cirillo[2], Flávio Meira Borém[3]

[1]Universidade Estadual de Maringá, Departamento de Estatística, Maringá, Paraná, Brazil
[2]Universidade Federal de Lavras/UFLA, Departemento de Estatística, Lavras, MG, Brasil
[3]Universidade Federal de Lavras/UFLA, Departamento de Engenharia Agrícola Engineering, Lavras, MG, Brasil, 37200-000,
Contact authors: ossanipc@hotmail.com, diogo.rossoni@gmail.com, macufla@dex.ufla.br, flavioborem@deg.ufla.br
Received in July 15, 2020 and approved September 10, 2020

ABSTRACT

Brazil is the largest exporter of coffee beans, 29% world exports, 15% this volume in specialty coffees. Thereby researches are done, so that identify different segments in the market, in order to direct the end consumer to a better quality product. New technologies are explored to meet an increasing demand for high quality coffees. Therefore, in this article has an objective to propose the use of machine learning techniques combined with projection pursuit in the construction of unsupervised classification models, in a sensory acceptance experiment, applied to four groups of trained and untrained consumers, in four classes of specialty coffees in which they were evaluated sensory characteristics: aroma, body coffee, sweetness and general note. For evaluating classifier performance, in the data with reduced dimension, all instances were used, and considering four groupings, the models were adjusted. The results obtained from the groupings formed were compared with pre-established classes to confirm the model. Success and error rates were obtained, considering the rate of false positives and false negatives, sensitivity and classification methods accuracy. It was concluded that, machine learning use in data with reduced dimensions is feasible, as it allows unsupervised classification of specialty coffees, produced at different altitudes and processes, considering the heterogeneity among consumers involved in sensory analysis, and the high homogeneity of sensory attributes among the analyzed classes, obtaining good hit rates in some classifiers.

Key words: Classification models; Data dimension reduction; Groupings identification; Projection pursuit.

1 INTRODUCTION

There is an exponential growth in global coffee consumption, and with the capacity to produce in large volumes. Brazil has become the largest exporter of coffee beans, accounting for about 29% of world coffee exports, equivalent to more than 34 thousand bags, with 15% of that volume in specialty coffees (Boaventura et al., 2018).

Still according to Boaventura et al. (2018) there is a revolution in the specialty coffees consumption through changes in product differentiation and consumption experience.

As there are differences among consumer preferences and the coffee segments, they should be served through marketing strategies that involve differentiation standards, which would increase quality, in order to add value to consumer satisfaction (Spers; Saes; Souza, 2004).

Thus, when considering acceptance or preference tests, in a sensory analysis experiment, focusing on the evaluation of the taster's intention, and discerning the sensorial quality of a product in relation to the others. In addition to the statistical problem, associated with the measurement error in filling out the sensory form, or in the resulting data analysis from the sensory notes distribution, some external factors that are relevant to the sensory panel formation, such as the taster experience, training of the panel and individual preferences, can be contemplated (Ossani et al., 2017).

Given this fact, research involving different statistical methodologies is proposed to validate more accurate results, with the purpose of refining results that discriminate the acceptance and discrimination of coffees sensory quality and derived products.

Ossani et al. (2017) proposed to use multiple factor analysis method for contingency tables (MFACT), which uses categorized data obtained from a sensory experiment carried out with different consumers groups. Seeking to identify similarities among four specialty coffees. In conclusion, the technical use is feasible, as it allows discriminating the specialty coffees produced in different environments (altitudes) and processing when considering the heterogeneity of consumers involved in sensory analysis.

In the study by Ferreira et al. (2016), a model was proposed using the extreme values distribution, to identify outliers presence. According to probabilities obtained, they concluded that untrained consumers are not able to differentiate specialty coffees.

In the case of specialty coffees sensory analysis, Ramos et al. (2016) used decision trees built using the CHAID technique, in sensory analysis of specialty coffees. According to the decision trees obtained, it was possible discriminating samples of coffees with sensory scores higher than 88 points, thus strongly associated with ambient coffee growing at altitudes higher than 1200m above sea level.

Liska et al. (2015) proposed a classification rule to discriminate trained and untrained tasters, using the conventional Fisher's discriminant analysis (LDA) and the discriminant analysis via the boosting algorithm (Adaboost). They concluded that, boosting method applied to the discriminant analysis showed a higher sensitivity rate in relation to the trained panel, approximately 80.63%, and there was a reduction in the false negative rate approximately 19.37%.

Among the numerous techniques proposed to classify data, machine learning is characterized by allowing to classify groups of variables with different sizes and different nature, wherein the unsupervised data classification deserves special attention, since the factors, in general, unknown, but related to sensory quality can be identified.

Machine learning techniques are closely linked to statistics and artificial intelligence, directly related to data mining. It can be defined as the computational techniques application that seeks to find hidden patterns in the data, producing algorithms capable of making computers learn and not just execute the algorithms.

There are many unsupervised data classification techniques, understood by machine learning, with particular characteristics. Making that different results are obtained, according to the structure inherent to the database analyzed used in the classification, thus the classifier choice is made by the one that best fits the data. In this context, it can mention the projection pursuit, which can be seen as a type of statistical technique used to find linear projections in multidimensional data, since, by reducing dimensions, making it possible to identify groupings. In this way, the projection pursuit can constitute an unsupervised classification method.

Unsupervised classification applied to coffee discrimination can be used as a consumer market segmentation tool, since by discovering characteristics inherent to specific groupings, research and marketing can be efficiently directed. In this way, the consumer will have greater reliability as that products found in the gondolas will be those that best suit their tastes, resulting in greater satisfaction, in addition to greater added value to the final product.

Therefore, the present work has as main objective to propose the use of several techniques of machine learning technique use combined with projection pursuit in the construction of unsupervised classification models, to be validated in the sensory attributes analysis, referring to specialty coffees, produced with different varieties, processing and altitudes. It will also be considered groups of consumers characterized by trained and untrained individuals in relation to experiences in sensory analysis of coffees.

2 MATERIAL AND METHODS

2.1 Data description

In order to attend the objective proposed, the data relative to a sensorial experiment was considered as the case study (Ossani et al., 2017), in which the samples of naturals and peeled cherry coffees from two genotypes of Coffea Arabica (Yellow Bourbon and Acaia) form the base of the experiment.

The samples were processed naturally, referring to the coffee drying process in which all the anatomical components are kept intact, and processed in the peeled cherry method, referring to the coffee drying process in which the exocarp and the mesocarp of the fruits are removed.

In the altitudes under 1.100m and over 1.200m, samples representative of the combination between genotype and processing were harvested. The removal of dirtiness and strange material was performed after the harvest of selected mature and healthy fruits. Later, the fruits were processed and sun-dried. The final level of water in the samples was of 11%. After the 30-day rest period, the samples were improved aiming at the obtaining of grains destined to the coffee roasting.

After the preparation of the coffee samples for sensorial evaluation, the defective grains were removed. Then, complying with the maximum period of 24 hours for tasting, the coffee was roasted, in accordance with the protocol of the Specialty Coffee Association of America (Specialty Coffee Association of America, SCAA, 2009).

In addition, in obedience to the rules of Re 466/12, the preventive measures proper to the preparation of food were taken, according to the opinion consolidated by the Ethics and Research Council, registered with the CAAE: 14959413.1.0000.5148.

Using the color classification system by means of standardized discs (SCAA, 2009), the roasting point was determined visually. In preparing the drink, the concentration of 7% m/v was maintained using filtered drinking water and without the addition of sugar. With these specifications, four classes of specialty coffees were coded in the samples by A, B, C and D, as described in Table 1.

The sensorial characteristics: acidity, body, sweetness and general grade were evaluated in the acceptance test for each type of coffee. The test was performed in four sessions with voluntary consumers with knowledge related to their personal experiences in relation to the sensorial analysis of coffees.

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

The groups $G = 1$ and 2 were formed by consumers who were trained for the sensorial evaluations. These groups were constituted, respectively, of 52 and 47 individuals. The members of the other groups ($G = 3$ and 4), were not trained; however, they were technical professionals or researchers in the area of interest, being 32 and 43 individuals, respectively.

Table 1: Specialty coffees description evaluated in the 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

Table 2: Table layout with sensory analysis results from classes of coffee.

Sample	Quantitative variables					Categorical Variables				Cls
	Acidity	Body coffee	Sweetness	General Note	Altitude	Age	Gender	Proc	Groups	
1										
2				x_{iGT}						
⋮										
696										

Proc = Processing, Groups = 1, 2, 3, 4 (groups of individuals)
Cls = A, B, C, D (Classes of coffees).
xiGT = i-th observation (instance) in group G and coffee classes T.

The individuals pointed out all the attributes from every coffee class with values in the range [0;10], with 10 as the highest grade adding up to 696 instances (observations), with 174 instances of each coffee type.

2.2 Projection pursuit

It is a technique for exploratory analysis of multivariate data, which looks for low-dimensional linear projections in high-dimensional data. Such projections are achieved through objective function optimization, called of projection pursuit (Friedman; Tukey, 1974). Thus, depending on the number of variables (Table 2), the Moment index was applied, with the purpose of researching linear projections that maximize the samples discrimination, among the different groupings to be detected. The index is defined by Equation (1), for two dimensions (Martinez; Martinez; Solka, 2010; Posse, 1995).

$$PI_M(A) = \frac{1}{12}\left\{k_{30}^2 + 3k_{21}^2 + 3k_{12}^2 + k_{03}^2 + \frac{1}{4}\left(k_{40}^2 + 4k_{31}^2 + 6k_{22}^2 + 4k_{13}^2 + k_{04}^2\right)\right\}, \quad (1)$$

The terms involved in the index composition correspond to the third and fourth bivariate moments. According to Martinez, Martinez and Solka (2010) the expressions for these moments are give in Equation 2-10.

$$k_{30} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\alpha})^3 \quad (2)$$

$$k_{03} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\beta})^3 \quad (3)$$

$$k_{12} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\beta})^2 z_i^{\alpha} \quad (4)$$

$$k_{21} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\alpha})^2 z_i^{\beta} \quad (5)$$

so, *nn* in the sample size, z_i the *i*-th observation of spherical data, and (z^{α},z^{β}) the spherical instances projected on the vectors α and β, that is, $z^{\alpha} = z^T \alpha$ and $z^{\beta} = z^{\beta} \beta$.

$$k_{40} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}\left(z_i^{\alpha}\right)^4 - \frac{3(n-1)^3}{n}\right\} \quad (6)$$

$$k_{04} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}\left(z_i^{\beta}\right)^4 - \frac{3(n-1)^3}{n}\right\} \quad (7)$$

$$k_{22} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}\left(z_i^{\alpha}\right)^2\left(z_i^{\beta}\right)^2 - \frac{(n-1)^3}{n}\right\} \quad (8)$$

$$k_{31} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(z_i^{\alpha}\right)^3 z_i^{\beta} \quad (9)$$

$$k_{13} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(z_i^{\beta}\right)^3 z_i^{\alpha} \quad (10)$$

The Holes and Central Mass indexes were also applied, defined by Equation (11) and (12), respectively, both formalized by Cook, Buja and Cabrera (1993), and are derived from the Normal density function, the first is a sensitive index to projections with few points in the center, and the second sensitive to projections with many points in the center, as suggested by Cook and Swayne (2007).

$$PI_{Holes}(A) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} exp\left(-\frac{1}{2}\mathbf{y}_i^T\mathbf{y}_i\right)}{1 - exp\left(-\frac{d}{2}\right)}, \quad (11)$$

$$PI_{MC}(A) = 1 - PI_{Holes}(A) \quad (12)$$

so, d reduction dimension of the original data dimension, and y_i the i-th data projected observation in the new dimension.

2.3 Classification methods

For comparison purposes of dimension reduction procedure performed with the projection pursuit (section 2.2), the following methods were considered**:**

Farthest First Model

The method consists of a complete graph G = (V, E) with weights at the edges $w_e \geq 0$, $e \in E$ and $w_{(v,v)} = 0$, $v \in V$.. The problem is to find a subset $S \subseteq V$ of k maximum such that $w(S) = max_{i \in V}\, min_{j \in S}\, w_{(i,j)}$ is minimized (Hochbaum; Shmoys, 1985). In general terms, given a X set of n points in a d-dimensional space and a k integer, when choosing a k set points in X, so that $\{c_1, c_2, \ldots c_k\}$ are the centers of the clusters $\{c_1, c_2, \ldots c_k\}$ and have

$$w(C) = \max_j \min_{x \in Cj} d(x, C_j), \quad (13)$$

minimized, wherein $d(x, C_j)$ is the distance $x \in X$ in cluster Cj.

K-means Model

K-means method is a widely used clustering technique that seeks to minimize the mean quadratic distance between points in the same cluster, aiming to partition n observations among k groups wherein each observation belongs to the group closest to the mean (Rencher; Christensen, 2002).

According to Lattin, Carroll and Green (2003), the method is prone to finding locally optimal solutions, as it is based on a heuristic that makes only local improvements for an initial partition, but its simplicity and speed are very attractive in the practice.

It can be used the Euclidean distance or Manhattan distance. If the Manhattan distance is used, then the centroids are calculated as the component median instead of the mean.

Computationally, different ways of implementing k-means method can be used. In this work, the algorithm presented by Arthur and Vassilvitskii (2007) was used.

Hierarchical Clustering Model

In this method, the instances grouped are in hierarchical manner. The criterion used to decide to what extent two instances of the data set can be considered similar or not, it uses several different measures of similarity and dissimilarity, and producing each one a certain grouping type.

Given two instances x_i and x_j, with $i \neq j$, wherein r is the r-th attribute of instance x, we have the Minkowski distance, which is defined by

$$d\left(x_i, x_j\right) = \left(\sum_{r=1}^{m} \left| x_{ir} - x_{jr} \right|^{\lambda} \right)^{\frac{1}{\lambda}} \quad (14)$$

Setting $\lambda = 1$ the distance is known as city-block or Manhattan, and $\lambda = 2$ has the Euclidean distance (Johnson; Wichern, 2007).

The Canberra distance (Johnson; Wichern, 2007), is defined by

$$d\left(x_i, x_j\right) = \sum_{r=1}^{m} \frac{\left| x_{ir} - x_{jr} \right|}{x_{ir} + x_{jr}} \quad (15)$$

The Maximum distance, known as Chebyshev distance, is defined by

$$d\left(x_i, x_j\right) = \max\left(\left| x_{ir} - x_{jr} \right| \right) \quad (16)$$

The choice of groups number in which a data set should be divided is subjective. Although, there are techniques, that help in determining number of groups.

There are several hierarchical groupings methods, based on the distances established among the instances.

The methods used in this work are presented below. Thus, given two clusters C_l and C_k, with $l \neq k$, with $x_i \epsilon C_k$, we have that:

Single Linkage method, also called the nearest neighbor method (Rencher; Christensen, 2012) is defined as the minimum distance between a point in C_l and a point in C_k, that is,

$$d\left(C_l, C_k\right) = \min\left(d\left(x_i, x_j\right) \right) \quad (17)$$

Complete Linkage method, also called the most distant neighbor method (Rencher; Christensen, 2012) is defined as the maximum distance between a point in C_l and a point in C_k, that is,

$$d\left(C_l, C_k\right) = \max\left(d\left(x_i, x_j\right) \right) \quad (18)$$

Average Linkage method (Rencher; Christensen, 2012) is defined as the average of $n_{C_l}\, n_{C_k}$ distance between n_{C_l} points in C_l and n_{C_k} points in C_k, that is,

$$d\left(C_l, C_k\right) = \frac{1}{n_{C_l} n_{C_k}} \sum_{i=1}^{n_{C_l}} \sum_{j=1}^{n_{C_k}} d\left(x_i, x_j\right), \text{para } x_i \in C_l \text{ e } x_j \in C_k. \quad (19)$$

Centroid method (Rencher; Christensen, 2012), is defined as the Euclidean distance among the average vectors, called centroids, between C_l and C_k, that is,

$$d\left(C_l, C_k\right) = d\left(\bar{x}_{C_l}, \bar{x}_{C_k}\right) = \left(\bar{x}_{C_l} - \bar{x}_{C_k}\right)^T \left(\bar{x}_{C_l} - \bar{x}_{C_k}\right) \quad (20)$$

wherein, $\bar{x}_{C_l}$ and $\bar{x}_{C_k}$ are the vectors of average of each observation in C_l and C_k, respectively.

Median method (Rencher; Christensen, 2012), is defined as the median between C_l and C_k, that is,

$$m_{C_l C_k} = \frac{1}{2}\left(\bar{x}_{C_l} + \bar{x}_{C_k}\right) \quad (21)$$

Ward method, also called the incremental sum of squares method, uses the square distances within the clusters and the square distances among clusters (Ward, 1963; Wishart, 1969). The combination that gives the least sum of squares is chosen. Be very sensitive to outliers, and produces clusters of approximately equal size (Rencher; Christensen, 2012).

McQuitty method (Mcquitty, 1966), is defined as the simple average of the distances among the clusters, that is,

$$d\left(C_l, C_k \cup C_w\right) = \frac{d\left(C_l, C_k\right) + d\left(C_l, C_w\right)}{2} \quad (22)$$

2.4 Procedures for proposed model validation

Aiming at the validation of the model proposed, the procedure described below was adopted:

1. The coffee classes established in Table 2 were ignored.
2. Using all the instances and taking into consideration four groupings, the following models were adjusted: Farthest First, K-means using the Euclidean distance, and Hierarchical Clustering using the Euclidean distances, Maximum, Manhattan, Canberra and Minkowski with Linkages Single, Average, Complete, McQuitty e Ward, for the non-supervised classifications.
3. After the adjustment of each machine learning model cited in (2), the results obtained from the formed groupings were compared to the coffee classes in Table 2 allowing the verification of the adjustment of the model. The validation error rate was given by:

$$VE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)$$

in which y_i, $\hat{y}_i$ e n denote, respectively the values observed e predicted of the classes for the i^{th} observation and the number of observations in the sample, here n = 696 according to Table 2. It was established that if the classes are equal $y_i - \hat{y}_i = 0$, otherwise $y_i - \hat{y}_i = 1$.

4. It was verified if there was a good adjustment considering the validation error rate under 30%.

The unsupervised classification analyzes were performed using the Weka software (Waikato Environment for Knowledge Analysis) version 3.8.3 (Hall et al., 2009) and the MVar 2.1.2 pack (Ossani; Cirillo, 2020) of the R software (R Development Core Team, 2020).

The classifiers Farthest First and K-means were done in Weka, while the classifier Hierarchical Clustering was done in R.

3. RESULTS

When using the projection pursuit technique, using the Moment index, in spherical data and with the optimization algorithm grant tour simulated annealing, through the MVar pack 2.1.2 (Ossani; Cirillo, 2020) of the software R (R Development Core Team , 2020), aiming to verify the groupings presence. It can be seen from Figure 1 that data in each class are very dispersed, also not showing a separation among classes of coffees, that is, there is no groupings formation in the analyzed sample.

As the classes are known, n = 4 was assigned, corresponding to four classes already established in Table 2, for the clusters analysis.

When applying the procedures in section 2.4, Table 3 shows the results of the classification error rates, using some of the known unsupervised classification techniques in the quantitative data (section 2.3).

The validation error rate presented in Table three was high, what justifies the search for new classification methods for data with structures of this nature in which the groups present little intraclass variability, what can be observed in Figure 1.

3.1 Classification proposal

In order to circumvent the high degree of homogeneity presented by the clusters shown in Table 2, and to minimize the error rates in the validation (Table 3), regarding quantitative variables, other attributes were added to enable the clusters differentiation at the classification time. It is proposed to use the category variables (Table 2), transforming them into dummy variables, in order to circumvent the high degree of homogeneity, and differentiate the referred clusters.

By the fact these are specialty coffees, whose quality is superior to that commercial coffees, which makes the notes of sensory attributes not differentiable in the clusters that are composed. In fact, different coffees, there are non-numeric attributes inherent to each cluster, and this can be used in an attempt to classify, so the use of the variables “Gender”, “Processing” and “Groups” were added in the classification.

Based on the use of dummy variables, and by the projection pursuit technique, using the Holes index, on spherical data and with the optimization algorithm grant tour simulated annealing. The data dimension was reduced in three dimensions, wherein presented the best results, and Figure 2

was generated using the scatterplot3d 0.3-41 pack (Ligges; Mächler, 2003). With the dimension two, or greater than three, the results were not satisfactory, therefore are omitted.

The results illustrated in Figure 2, unlike the data presented in the Figure 1, suggest that there is statistical evidence to point out that there was an intraclass distinction, with the addition of categorical variables. Suggesting that, it could favor the data classification. But, when applying the classifiers to all attributes in Table 2, it happened that the data were not classified, although the clusters were noticed in the data with reduced dimensions.

When applying the procedure suggested in this work (section 2.4) in every attribute from Table 2, which included quantitative and qualitative variables, the data was not classified although it was possible to notice the clusters in the data with reduced dimensions (Figure 2). The results were

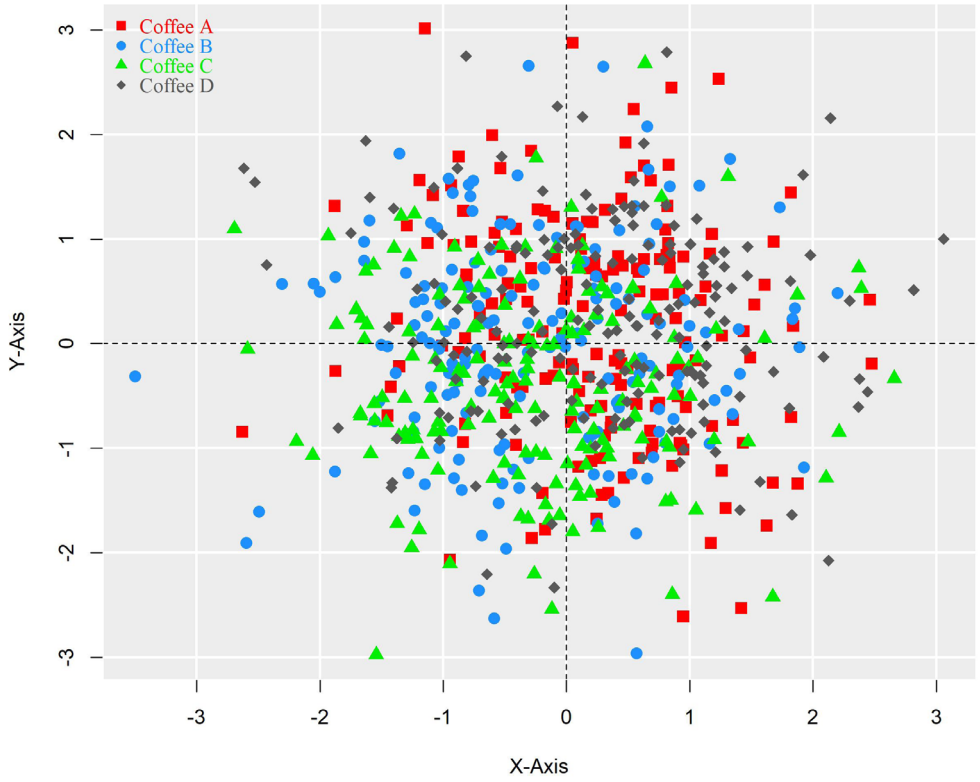

Figure 1: Graph of projection pursuit results in the data using Moment index.

Table 3: Machine learning techniques results used in the quantitative variables.

Nº	Classifier	Distance	Linkage	Validation error rate %
1	Farthest First[1]	-	-	50.28
2	K-means[1]	Euclidean	-	48.85
3	Hierarchical Clustering[2]	Euclidean	Single	50.14
4	Hierarchical Clustering[2]	Euclidean	Average	48.42
5	Hierarchical Clustering[2]	Maximum	Ward	65.52
6	Hierarchical Clustering[2]	Maximum	Average	69.83
7	Hierarchical Clustering[2]	Maximum	McQuitty	65.95
8	Hierarchical Clustering[2]	Manhattan	Single	50.14
9	Hierarchical Clustering[2]	Manhattan	Average	72.70
10	Hierarchical Clustering[2]	Canberra	Complete	72.13
11	Hierarchical Clustering[2]	Minkowski[3]	Ward	65.52
12	Hierarchical Clustering[2]	Minkowski[3]	Average	65.23

[1] Weka software [2] Software R.
[3] The best result with the parameter p ranging from 3 to 25 integers.

omitted here because they presented validation error rates as high as the ones presented in Table 3.

However, when the procedure present in section 2.4 was applied again directly in the data with the reduced dimensions obtained through the projection pursuit technique using the qualitative and quantitative variables in the data projected in three dimensions, the results described in Table 4 were obtained.

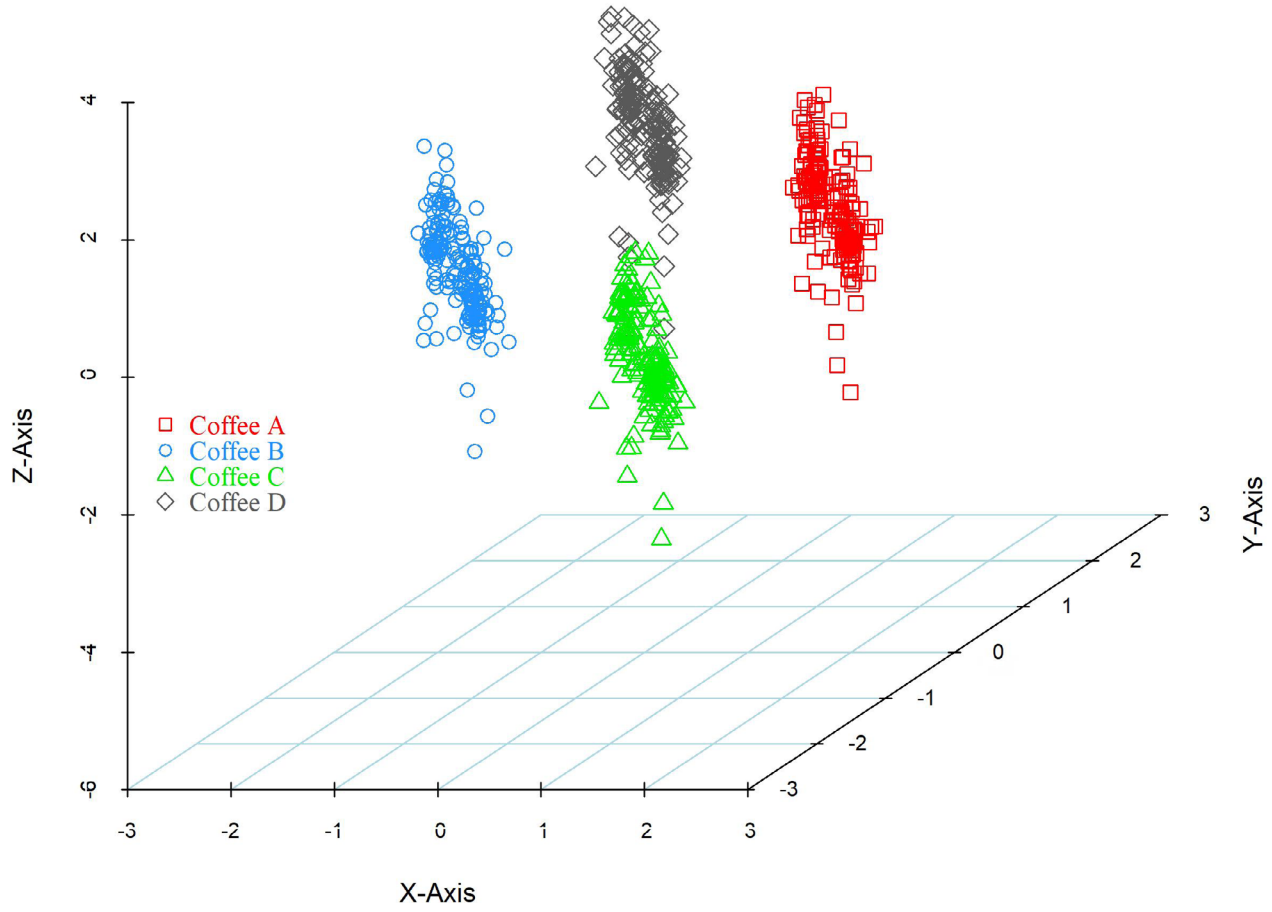

Figure 2: Graph of projection pursuit results in the data using the Holes index in three dimensions.

Table 4: Results of classification techniques, showing better results in unsupervised classification, in data with reduced dimensions.

Nº	Classifier	Distance	Linkage	Validation error rate %
1	Farthest First[1]	-	-	50.43
2	K-means[1]	Euclidean	-	36.49
3	Hierarchical Clustering[2]	Euclidean	Single	24.86*
4	Hierarchical Clustering[2]	Euclidean	Average	25.57*
5	Hierarchical Clustering[2]	Maximum	Ward	1.44*
6	Hierarchical Clustering[2]	Maximum	Average	25.57*
7	Hierarchical Clustering[2]	Maximum	McQuitty	32.04
8	Hierarchical Clustering[2]	Manhattan	Single	24.86*
9	Hierarchical Clustering[2]	Manhattan	Average	25.29*
10	Hierarchical Clustering[2]	Canberra	Complete	24.28*
11	Hierarchical Clustering[2]	Minkowski[3]	Ward	1.44*
12	Hierarchical Clustering[2]	Minkowski[3]	Average	25.57*

[1] Weka software [2] Software R.
[3] The best result with the parameter p ranging from 3 to 25 integers.
* Results with a rate below 30%.

Unlike the results presented in Table 3, which refers to the data in the original dimensions, the results presented in Table 4 for the data in reduced dimensions projected in three dimensions have proven themselves efficient in the non-supervised classification with validation error rates under 26% in most of the tested algorithms. In some algorithms, the validation error rates were under 2%, what is considered a good result.

It should be noted that the results presented in Table 4, using the Holes index, were the same as those obtained with the Central Mass index, with the results of the latter omitted because they are equal.

4 DISCUSSION

The classifiers in the clusters formation based on characteristics adjacent to the elements that make up each group, which allows the clusters separation at the classification time. Now, if the global point cloud does not present a well-defined intraclass separation (Figure 1), that is, since the groups are highly homogeneous with each other, this clusters differentiation becomes difficult, given the difficulty of the algorithms to distinguish groupings due to the lack relevant classificatory elements. This fact occurred with the quantitative data analyzed, presented in Table 3, with high error rates in the validation, and the best value given by Hierarchical Clustering, using the Euclidean distance and Average method, result in 48.42%, that means, each 100 instances, 48 were classified incorrectly, which is a bad result.

The addition of the dummy variables was not enough for the non-supervised classification method although the projection pursuit method presented a structure in reduced dimension that indicated the presence of distinct groupings in the data in original dimension (Figure 2). However, the classification algorithms were not efficient in catching the differences present in these groupings.

When working directly with the data in reduced dimension projected in three dimensions, the data classification was possible. Such was captured by the classifiers used in the analyzes with significant improvement in the validation error rates presented in Table 4 in relation to those in Table 3. This result suggests that structures in high dimensions, represented by matrices of order n x m, hide information that can be picked up in their projections in smaller dimensions. Thus, the best results were reached through the Hierarchical Clustering classifier, using the distance Maximum and the Ward method, and, for the same classifier, using the Minkowski distance and the Ward method. In both cases, a validation error rate of 1,44% was reached. Other inferior results with classification error rate of 26% were reached by the Hierarchical Clustering classifier with other settings beings presented in Table 4.

Among the main results, stands out Holes and Central Mass indexes they were originally created with the purpose of finding data structures that have few points in the center and accumulations, respectively, but in this work an application for them was in the joint use with unsupervised classification techniques in the search of groupings that corresponded to a previous classification, in which they were successful. This result can indicate that other indexes used in projection pursuit can be used for other purposes different from the ones for which it was originally created, therefore expanding the use of projection pursuit in other research situations.

5 CONCLUSIONS

In line with the objectives and the proposed methodology, in this application it is concluded that the machine learning technique is feasible to be applied in the unsupervised classification of sensory data of consumer classes with different skills in relation to the specialty coffees quality, as long as the groupings data have a well-defined intraclass structure, otherwise it is necessary to ally with other techniques in order to highlight the groupings structures. In this work, machine learning technique combined with the projection pursuit was efficient in differentiating among different classes of coffees. Suggesting that, structures in high dimensions hide information that can be captured in their projections in smaller dimensions.

6 REFERENCES

ARTHUR D.; VASSILVITSKII, S. k-means++: the advantages of carefull seeding. **Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms**, 1027-1035, 2007.

BOAVENTURA, P. S. M. et al. Value co-creation in the specialty coffee value chain: The third-wave coffee movement. **Revista de Administração de Empresas,** 58(3):254-266, 2018.

COOK, D.; BUJA, A.; CABRERA, J. Projection pursuit indexes based on orthonormal function expansions. **Journal of Computational and Graphical Statistics**, 2(3):225-250, 1993.

COOK, D.; SWAYNE, D. F. **Interactive and dynamic graphics for data analysis:** With R and GGobi. New York: Springer, 2007, 202p.

FRIEDMAN, J. H.; TUKEY, J. W. A projection pursuit algorithm for exploratory data analysis. **IEEE Transaction on Computers**, 23(9):881-890, 1974.

HALL, M. et al. The WEKA data mining software: An update. **ACM SIGKDD Explorations Newsletter**, 11(1):10-18, 2009.

HOCHBAUM, D. S.; SHMOYS, D. B. A best possible heuristic for the k-center problem. **Mathematics of Operations Research**. 10(2):180-184, 1985.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**, 6th ed. New Jersey: Pearson Prentice Hall, 2007, 794p.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. **Analyzing multivariate data**. Pacific Grove, CA: Thomson Brooks/ Cole, 2003, 455p.

LIGGES, U.; MÄCHLER, M. Scatterplot3d - An R package for visualizing multivariate data. **Journal of Statistical Software**, 8(11):1-20, 2003.

LISKA, G. R. et al. Evaluation of sensory panels of consumers of specialty coffee beverages using the boosting method in discriminant analysis. **Semina: Ciências Agrárias**, 36(6):3671-3679, 2015.

MARTINEZ, W. L.; MARTINEZ, A. R.; SOLKA, J. **Exploratory data analysis with MATLAB**. 2nd ed. New York: Chapman & Hall/CRC, 2010, 499p.

MCQUITTY, L. L. Similarity analysis by reciprocal pairs for discrete and continuous data. **Educational and Psychological Measurement**, 26(4):825-831, 1966.

FERREIRA, H. A. et al. Selecting a probabilistic model applied to the sensory analysis of specialty coffees performed with consumer. **IEEE Latin America Transactions**, 14(3):1507-1512, 2016.

POSSE, C. Tools for two-dimensional exploratory projection pursuit. **Journal of Computational and Graphical Statistics**, 4(2):83-100, 1995.

OSSANI, P. C.; CIRILLO, M. A. **MVar:** Multivariate analysis. 2020. R package version 2.1.2. Available in: <https://cran.r-project.org/web/packages/MVar/index.html>. Access in: September, 10, 2020.

OSSANI, P. C. et al. Quality of specialty coffees: A sensory evaluation by consumers using the MFACT technique. **Revista Ciência Agronômica**, 48(1):92-100, 2017.

RAMOS, M. F. et al. Discrimination of the sensory quality of the *Coffea arabica* L. (cv. Yellow Bourbon) produced in different altitudes using decision trees obtained by the CHAID method. **Journal Of The Science Of Food And Agriculture**, 96(10):3543-3551, 2016.

RENCHER, A. C.; CHRISTENSEN, W. F. **Methods of Multivariate Analysis**. 3th. ed. New York: J. Wiley, 2012. 758p.

R DEVELOPMENT CORE TEAM. **R:** A language and environment for statistical computing. R foundation for statistical computing. 2020. Vienna: Vienna University of Economics and Business. Available in: <http://www.R-project.org/>. Access in: September, 10, 2020.

SPECIALITY COFFEE ASSOCIATION OF AMERICA. SCAA Protocols. **Cupping Specialty Coffee**. Long Beach: SCAA, 2009, 7p.

SPERS, E. E.; SAES, M. S. M.; SOUZA, M. C. M. 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, 2004

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, 58(301):236-244. 1963.

WISHART, D. An algorithm for hierarchical classifications. **Biometrics**, 25:165-170, 1969.