[Applied Computing and Informatics 15 (2019) 153–162](https://doi.org/10.1016/j.aci.2018.02.003)



Contents lists available at [ScienceDirect](http://www.sciencedirect.com/science/journal/22108327)

Applied Computing and Informatics

journal homepage: [www.sciencedirect.com](http://www.sciencedirect.com/)

Original Article

[](http://crossmark.crossref.org/dialog/?doi=10.1016/j.aci.2018.02.003&domain=pdf)Research on social data by means of cluster analysis

Camila Maione [a](#_bookmark0), Donald R. Nelson [b](#_bookmark1), Rommel Melgaço Barbosa [a](#_bookmark0),[⇑](#_bookmark2)

a *Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás, Brazil*

b *Department of Anthropology, University of Georgia, Athens, USA*

a r t i c l e i n f o

*Article history:*

Received 13 December 2017

Revised 20 February 2018

Accepted 20 February 2018

Available online 21 February 2018

*Keywords:* Clustering Social data Classification Pam

Data mining

a b s t r a c t

This paper presents a data mining study and cluster analysis of social data obtained on small producers and family farmers from six country cities in Ceará state, northeast Brazil. The analyzed data involve demographic, economic, agriculture and food insecurity information. The goal of the study is to establish profiles for the small producer families that reside in the region and to identify relevant features which differentiate these profiles. Moreover, we provide an efficient data mining methodology for analysis of social data sets which is capable of handling its natural challenges, such as mixed variables and abun- dance of null values. We use the Silhouette method for the estimation of the best number of natural groups within the data, along with the Partitioning Around Medoids clustering algorithm in order to com- pute the profiles. The Correlation-Based Feature Selection method is used to identify which social criteria are the most important to differentiate the families from each profile. Classification models based on sup- port vector machines, multilayer perceptron and decision trees were developed aiming to predict in which of the identified clusters an arbitrary family would be best fit. We obtained a good separation of the families into two clusters, and a multilayer perceptron model with approximately 93.5% prediction accuracy.

© 2018 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an

open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The Brazilian law (Lei 11.326, 2006) defines the family farmer as the family which practices agricultural activities in a rural area that is not larger than four fiscal modules and uses predominantly the family’s own labor in their business activities. Moreover, according to the law, the income of a family farmer must be originated from activities related to their own business, which must be ran by the farmer and his family. According to Guanziroli [[15]](#_bookmark17), in 2001, family-based agriculture in Brazil was so far responsible for almost 40% of the total agricultural production, and accounts for 76.8% of agricultural employment.

Most of the family-based agriculture in Brazil is consolidated in the Northeast region. However, the family farmers which reside in this region face various challenges related to the soil and climate.

\* Corresponding author address: Alameda Palmeiras, Quadra D, Câmpus Samambaia, Instituto de Informática, CEP 74690-900 Goiânia, Goiás, Brazil.

*E-mail address:* [rommel@inf.ufg.br](mailto:rommel@inf.ufg.br) (R.M. Barbosa).

Peer review under responsibility of King Saud University.

**Production and hosting by Elsevier**

Approximately 70% of the semiarid region is above the crystalline basement where soils are generally shallow and with low water infiltration capacity, which limit the development of agriculture [[16]](#_bookmark17). Furthermore, the Northeast region frequently suffers from drought, resulting in insufficient or irregular distribution of rain- fall, which cannot keep the soil moist enough during periods longer than the harvest cycle [[8]](#_bookmark17). In addition, the income distribution among the farmer families in the region appears to be deficient. In 2006, family-based agriculture in the Northeast region was com- posed of more than 4,500,000 farmers; only around 450,000 of them comprised the most capitalized group, with a annual income of more than 53,000 reais, while more than 2,500,000 farmers sur- vived only on subsistence activities, with an annual monetary income of approximately 255 reais [[16]](#_bookmark17).

One possible way to identify categories among these families and to gain insight on the most important factors for this division is the analysis of social data regarding these families by cluster analysis and data mining techniques. Data mining is a process of discovery of useful information and hidden patterns in databases, widely used in recent research in various fields such as food science [[3,4,26,27]](#_bookmark25), agriculture [[28,29]](#_bookmark24), social media [[2,17]](#_bookmark26), business and customer management [[30]](#_bookmark27), sports [[6]](#_bookmark17) and others. However, the analysis of social data sets present some challenges derived from the manner in which they are gathered and stored. Social data sets

<https://doi.org/10.1016/j.aci.2018.02.003>

2210-8327/© 2018 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

are composed of many mixed variables, which restrain the direct application of one the most popular clustering algorithms, the K- means. There is also a large quantity of null values distributed among the data that must be handled. Due to the exploratory nat- ure of the analysis that we wish to conduct, the number of natural groups existing within the data is unknown a priori, and many clus- tering techniques require this value as an input parameter.

This study presents a data mining study which aims to establish profiles for the family farmers residing in Northeastern Brazil. The proposed methodology include data mining, cluster analysis and variable selection techniques in order to identify profiles among the analyzed families based on demographic, economic, agricul- tural production and food insecurity information. We focus our study on family farmers from six municipalities from the state of Ceará. We identified homogeneous groups within the available data, aiming to find which families are relatively similar in terms of the gathered information and which of these social criteria are more relevant for the differentiation of the family groups. We built classification models capable of predicting in which of these groups an arbitrary family would be best fit. Decision rules were employed to aid the interpretation of the clustering results.

Finally, this paper presents the following contributions:

* We conducted a data analysis of social information in order to identify profiles of small farmer families from the state of Ceará,

using statistical and machine learning techniques from data mining and cluster analysis;

* We provide a data mining methodology which is capable of

handling challenges inherent to social data sets, such as mixed

variables, abundant null values and the absence of information about the number of natural groups within the data.

The content of this paper is organized as follows: Section [2](#_bookmark3) describes the methodology employed, the data set, concepts and algorithms used; Section [3](#_bookmark7) details the results obtained; Section [4](#_bookmark16) brings the conclusion.

1. Methodology

The analyzed data was obtained from research conducted on

476 households distributed around six country cities from the Ceará state: Barbalha, Guaraciaba do Norte, Boa Viagem, Limoeiro do Norte, Itarema and Parambu ([Fig. 1](#_bookmark4)). Each family was evaluated by a questionnaire composed of 79 questions regarding demo- graphic, economic, agricultural production and food security infor- mation. The answers for the questions were defined as the features for the data set. Some important questions (features) are listed in [Table 1](#_bookmark5). The number of evaluated families from each city is similar: 80 from Barbalha, 80 from Boa Viagem, 80 from Guaraciaba do Norte, 79 from Itarema, 79 from Limoeiro do Norte and 78 from Parambu. Data were collected in 2012 over a 1-month period, as part of a research project focused on understanding household drought vulnerability. The six municipalities selected each repre- sent one of the agro-ecological zones in the state. Questions were developed based on a sustainable livelihoods approach [[34]](#_bookmark28), which categorizes five types of capital – social, natural, financial, human and physical – and provides a framework for understanding their

related to the nature and domain of the social questions involved. For this matter, we employed cluster analysis concepts and techniques.

Data mining refers to the process of automatic discovery of use- ful information in large databases [[35]](#_bookmark29). Data mining techniques merge machine learning, statistical calculation, linear algebra and mathematic optimization concepts in order to uncover hidden pat- terns in data. Machine learning is a fundamental procedure for data mining processes as it generates smart algorithms capable of dis- covering patterns and information in data bases automatically, which aid decision making [[5]](#_bookmark17).

Cluster analysis is a data mining process which consists in dividing the samples into groups (clusters) based on information found within the data which describes these samples and its rela- tionships [[35]](#_bookmark29). Samples belonging to the same cluster must show a similarity pattern among them while being as dissimilar as possi- ble from samples associated to other clusters.

Due to the nature of social data and the way they are usually obtained and stored, the application of clustering techniques on this kind of data proves challenging. Research on social data is usually conducted through the use of structured forms which bring a num- ber of questions to be answered by the members of the observed population. Once collected, these data are stored in spreadsheets or plain text. The conversion of these files to data matrixes usually generates columns of mixed types, including numeric, categoric and ordinal variables. Moreover, fields that are not filled are mapped into null values, which represent obstacles for data analysis.

Some recent research brings methodologies and techniques for handling the problem of data sets with variables of mixed types during cluster analysis [[1,21,32]](#_bookmark30). Another point to be observed is that, in this study, we conduct an exploratory analysis where we have no pre-established models or hypotheses regarding the data [[22]](#_bookmark18). Most clustering techniques require the number of clusters as an input parameter, and in our case this number is not known a priori. We employed the Silhouette method which aids in the estimation of the best number of clusters within the data set.

* + 1. *Estimating the best number of clusters*

In this study, we used the Silhouette method [[33]](#_bookmark31) to estimate the best number of natural clusters within the data set. Silhouette is a graphical display method for clustering algorithms. Each clus- ter is represented by a silhouette which is designed based on sim- ilarity and dissimilarity between the samples. The silhouette shows which samples are well placed in the clusters and which samples are floating between two or more clusters. The entire par- tition is shown in a single plot as a combination of silhouettes which allows visualization of the quality of the clusters. The best number of clusters can be estimated by computing the average weigth of the silhouettes designed for a range of partitions with different numbers of clusters.

Consider a(i) as the average distance between the sample *i* and

*k* being the number of clusters, d(i; C) is the average distance the other samples in the same cluster as *i*. For all other clusters C*k*, d(i; C), given by b(i) := min d(i; C), is the distance between *i* and between *i* and the other samples in C*k*. The minimum value of

c

the nearest cluster. Hence, s(i) is computed as:

relationships. The respondents were randomly selected from a list

# s

b(i)— a(i)

of farmers that were provided by the local Farm Workers Syndi-

cate, which contains an almost exhaustive list of farmers [Table 2](#_bookmark6).

* 1. *Knowledge discovery using data mining and cluster analysis*

The first step of the analytical procedure was to identify rele- vant groups of the interviewed families based on a similarity factor

# (i) := max(a(i); b(i))

Samples with show high values for s(i) are well clustered, otherwise ples with negative values for s(i) are placed in wrong clusters. The they are probably placed between two or more clusters, and sam- value of s(i) for all samples associated to it. quality of a cluster can be estimated by computing the average

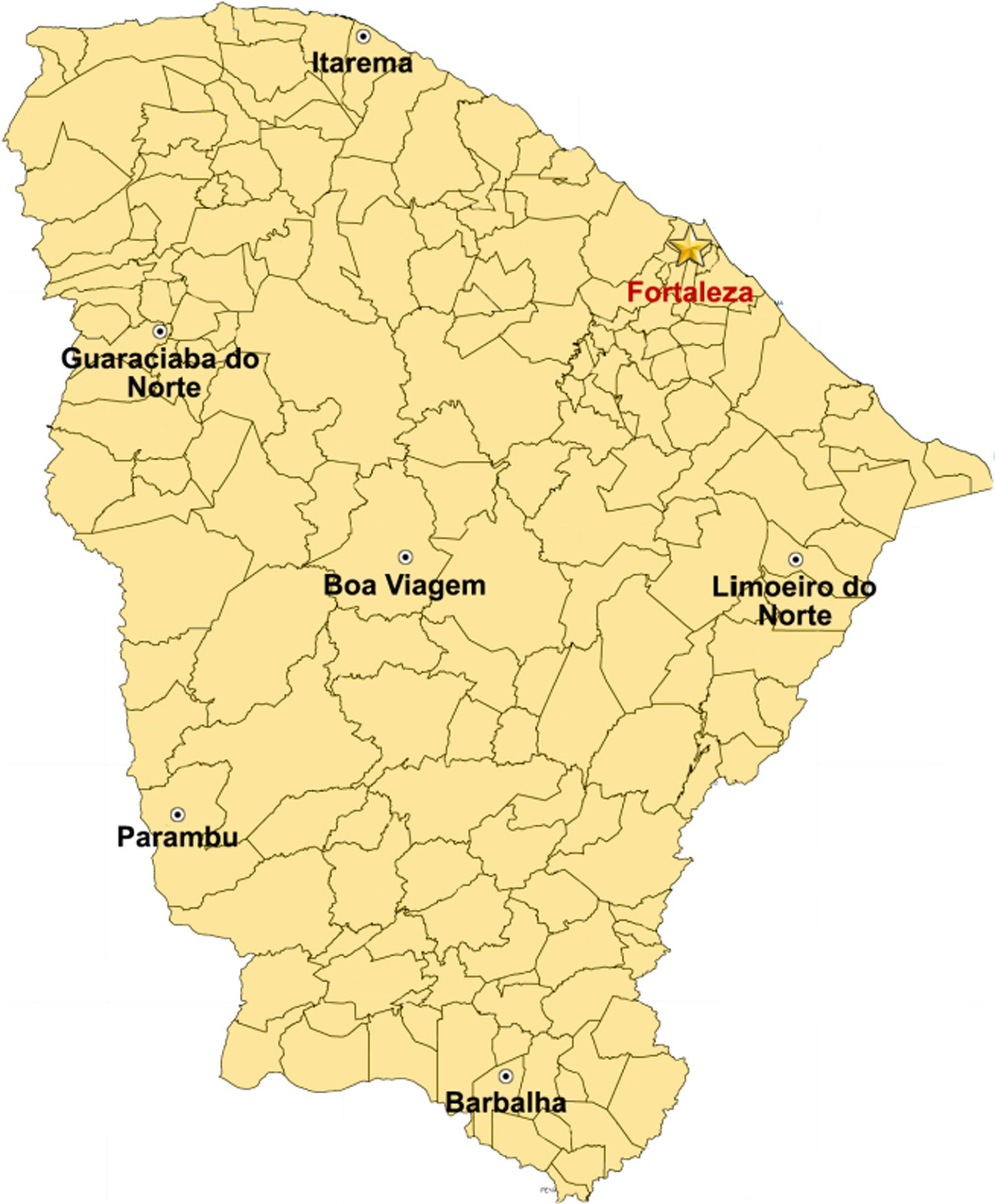


Fig. 1. Map of Ceará state. The six cities of origin of the family farmers analyzed (Barbalha, Guaraciaba do Norte, Boa Viagem, Limoeiro do Norte, Itarema and Parambu) and the capital, Fortaleza, are highlighted.

range of possible values, F = {2; ... ; n}. For each k ∈ F, a clustering In order to choose the best number of clusters, we selected a ette avg(s(i))k is computed for all *k* clusters. To conclude, the final algorithm divides the data set in *k* clusters, and the average silhou-

P

average n media(s(i))k is computed and selected as the best

k=2

number of clusters for the data set.

* + 1. *Partitioning around medoids (PAM)*

The clustering algorithm selected was the PAM [[24]](#_bookmark19), an exten- sion of the K-means algorithm. K-means is an antique clustering algorithm which is still very popular due to its speed, efficiency and simplicity [[22]](#_bookmark18). This method searches for *k* centers within the data set which minimizes the total sum of the squared dis- tances between each sample and its nearest center. The K-means is executed in the following steps:

1. Randomly choose *k* samples from the data set as the initial centers *C* = {*c*1, *c*2,.. .*ck*}.
2. For each *i* ∈ {1,.. .,*k*}, set the cluster *Ci* as the subset of samples which are nearest to *ci* than to any other center *cj* for all *j* – *i*.
3. For each *i* ∈ {1,.. .,*k*}, update *ci* to be the center of mass of all samples in *Ci*.
4. Repeat Steps 2 and 3 until no changes are observed in *C* or a maximum number of iterations is reached.

Although K-means is currently the best-known and most used clustering algorithm, it has the key limitation of working with numerical values only, which makes its application on our mixed data set unfeasible. We opted for the use of the PAM algorithm, which implements the K-medoids [[23]](#_bookmark20)clustering algorithm.

K-medoids is similar to K-means, but more robust and less sensitive to outliers. The k-medoids algorithm is based on finding *k* representative samples within the data set, called medoids, and the *k* clusters will be constructed through the association of each sample to its nearest representative object. K-medoids can also receive a dissimilarity matrix instead of the data, a matrix which

Table 1

Type and description of the descriptive variables, which correspond to information obtained on 476 family farmers from six country rural towns in Ceará state.

|  |  |  |
| --- | --- | --- |
| Identificação | Tipo | Descrição |
| Municipality | Categoric | City where the family resides in |
| F\_1 | Binary | Household has saving account |
| F\_2 | Binary | Household received credit in the past |
| F\_4 | Binary | Household received remittances |
| F\_6 | Binary | Household never received Bolsa Família |
| H\_1 | Binary | Household has access to health services |
| H\_5 | Binary | School enrollment |
| H\_5a | Binary | Household has kids in school age |
| N\_1 | Binary | Household own their own land |
| N\_3 | Binary | Household own more land than they cultivate |
| N\_4 | Nominal | Dependability of irrigation source |
| N\_5 | Nominal | Household have made land improvements |
| N\_7 | Binary | Household has running water |
| M\_3 | Binary | Household owns a car or motorcycle |
| M\_4 | Binary | Household uses a tractor |
| S\_2c | Binay | Household provided support to other families |
| S\_2d | Binary | Household received support from other families |
| AS\_1 | Ratio | Number of plots of agricultural land |
| AS\_5 | Binary | Household has total subsistence crop loss |
| P\_1 | Binary | Household participated in ‘‘Hora de Plantar” government seed distribution program |
| P\_2 | Binary | Household has knowledge of PAA (government food purchasing program) |
| P\_4 | Binary | Household has knowledge of crop insurance program |
| P\_5 | Binary | Household participated in crop insurance |
| P\_6 | Binary | Household used loaned farm equipment |
| P\_7 | Binary | Household participated in PAA |
| FS\_3 | Binary | Household has insufficient quantity of food |
| FS\_4 | Binary | Household has food of insufficient quality |
| MG\_1 | Binary | Household has an individual that migrated in the past and has returned |
| MG\_4 | Binary | Household has an individual that is a current migrant |
| MG\_6 | Nominal | Types of migrants in the household |
| AG\_1 | Binary | Household planted corn or beans for subsistence |
| AG\_2 | Binary | Household planted other subsistence crops |
| AG\_3 | Binary | Household planted horticulture |
| AG\_4 | Binary | Household planted cash crops |
| AG\_5 | Binary | Household planted fruit crops |
| F\_3 | Numeric | Per capita annual income |
| F\_3a | Numeric | Per capita annual income, inflation adjusted |
| F\_5 | Numeric | Per capita livestock assets |
| H\_2 | Ratio | Dependency ratio |
| H\_3 | Ratio | Percentage of adults older than 17 that have finished high school |
| N\_2 | Numeric | Quantity of land cultivated in hectares |
| M\_1 | Ordinal | Number of pieces of farm equipment owned |
| M\_2 | Ratio | Asset index |
| M\_2a | Ratio | Household asset index |
| M\_2b | Ratio | Age productivity asset index |
| M\_5 | Numeric | Number of agricultural technologies used |
| S\_1 | Numeric | Number of social groups that the household belongs to |
| S\_2 | Ratio | Frequency of given and received support |
| S\_2a | Ratio | Frequency of support given |
| S\_2b | Ratio | Frequency of support received |
| LS\_1 | Numeric | Number of sources of income per capita |
| LS\_2a | Ratio | Percentage of total income from climate sensitive sources |
| LS\_2b | Ratio | Percentage of total income from climate neutral sources |
| LS\_2c | Ratio | Percentage of total income from Bolsa Família cash transfer program |
| LS\_2d | Ratio | Percentage of total income from social security |
| LS\_3 | Ordinal | Drought impacts on household relative to other households in the area |
| AS\_2 | Ordinal | Soil quality index |
| AS\_3 | Ratio | Number of crop types |
| FS\_1 | Ordinal | Food security index |
| FS\_2 | Ratio | Acess index |
| MG\_2 | Numeric | Number of migration episodes in the past per household |
| MG\_3 | Numeric | Maximum number of individual migration episodes |
| MG\_5 | Numeric | Number of current migrants |
| MG\_7a | Numeric | Sum of individuals in household that migrated to city seat |
| MG\_7b | Numeric | Sum of individuals in household that migrated to another city |
| MG\_7c | Numeric | Sum of individuals in household that migrated to Fortaleza (capital of Ceará) |
| MG\_7d | Numeric | Sum of individuals in household that migrated to another state |
| MG\_7e | Numeric | Sum of individuals in household that migrated to another country |
| R\_1 | Numeric | Annual household income from climate sensitive sources |
| R\_2 | Numeric | Annual household income from social security |
| R\_3 | Numeric | Annual household income from Bolsa Família |
| R\_4 | Numeric | Annual household income from climate neutral sources |
| R\_1a | Numeric | Annual household income from climate sensitive sources, inflation adjusted |
| R\_2a | Numeric | Annual household income from social security, inflation adjusted |
| R\_3a | Numeric | Annual household income from Bolsa Família, inflation adjusted |
| R\_4a | Numeric | Annual household income from climate neutral sources, inflation adjusted |

Table 2

Best features determined by the CFS method.

iii. If *j* is more distant from *i* than from at least one of the other representative samples, but nearest to *h* than from any other representative sample, the contribution of *j* to the swap is

Cjih = d(j; h)— Dj.

tions Tih = Pj Cjih. 2. Compute the final result of the swap by adding the contribu-

3. Select the pair (*i*,*h*) which satisfies the function mini;h

Tih

.

4. If the minimum value of *Tih* is negative, swap and return to step

|  |  |  |
| --- | --- | --- |
| Feature | Type | Description |
| F\_2 | Categorical | Household received credit in the past |
| F\_3 | Numeric | Per capita annual income |
| F\_6 | Categorical | Household never received Bolsa Família |
| H\_1 | Categorical | Household has health access |
| S\_2a | Numeric | Frequency of support given |
| S\_2c | Categorical | Household gave support to other families |
| LS\_2C | Numeric | Percentage of total income from Bolsa Família cash  transfer program |
| LS\_2D | Numeric | Percentage of total income from social security |
| P\_5 | Categorical | Household participated in crop insurance |
| FS\_1 | Numeric | Food security index |
| FS\_2 | Numeric | Access index |
| FS\_4 | Categorical | Household has food of insufficient quality |
| R\_2 | Numeric | Annual household income from social security |
| R\_2A | Numeric | Annual household income from social security, 2012 |
| R\_3A | Numeric | inflation adjusted  Annual household income from Bolsa Família, 2012 |
| AG\_1 | Categorical | inflation adjusted  Household planted corn or beans for subsistence |

1. Otherwise, the swap is considered undesirable and the algo- rithm ends.

stores pairwise distances between the samples. The medoid of a cluster will be the sample from this cluster such that the average dissimilarity between this sample and the other samples in this

same cluster is as small as possible, instead of the squared sum

* + 1. *Dissimilarity matrix calculation*

The dissimilarity matrix which PAM algorithm receives as input parameter can be computed using the Daisy function detailed by [[24]](#_bookmark19). This function accepts mixed data types, including numeric, categoric, ordinal, symmetric and asymmetric binary values, which is useful for the data set we analyze in this study. In order to handle mixed variables, Daisy uses the Gower dissimilarity coefficient [[14]](#_bookmark17). Each variable is normalized by dividing each value by the range of values of the corresponding variable, after the subtraction of the minimum value. Thereafter, the variable will be scaled to [0,1]. Null values are discarded from the calculation.

The Gower coefficient computes the distance between two sam- ples *i* and *j* as the weighted average of the contributions of each variable, given by the equation:

P*p wk*d*kdk*

P

of Euclidian distances computed by the K-means. The algorithm

*d*

= *d*(*i*; *j*)= *k*=1 *ij ij*

*ij*

is executed in two steps, BUILD and SWAP. In BUILD phase, an

*ij*

initial clustering is performed by the successive selection of *k* rep-

*k*=1 *p*

*ij*

*ij*

*wk*d*k*

resentative samples. The first sample selected must present the

where *dij* is the weighted average of *dk* , *wk*d*k* are the corresponding

least average dissimilarity possible from the other samples and,

weights, d*k* is 0 or 1, and *dk* is the *k*th contribution of the variable for

*ij ij*

set. The next *k* — 1 samples are selected in order to decrease therefore, should be located at the center of mass of the data the objective function value, following the steps:

1. Consider *i* as an unselected sample.
2. Consider *j* an unselected sample. Compute the difference between *Dj* (dissimilarity between *j* and the nearest selected sample) and *d*(*j*,*i*) (dissimilarity between *j* and *i*).
3. If the computed difference is positive, compute the contribution of *j* to the decision if *i* should be selected, given by

Cji = max(Dj — d(j; i); 0).

Pj

1. Compute the total gain obtained if *i* is selected given by Cji.
2. Select the unselected sample *i* which satisfies maxiPj Cji.

The SWAP phase tries to optimize the set of selected represen- tative samples. Consider all pairs (*i*,*h*) of selected sample *i* and uns- elected sample *h*. In order to decide if *i* should be swaped with *h*, the following equations and steps are performed:

1. Consider an unselected sample *j* and compute its contribution Cjih to the swap:

a. If *j* is more distant from *i* and *h* than from the other represen- tative samples, Cjih is zero.

tative samples,d(j; i)= Dj; b. If *j* is not more distant from *i* than from the other represen-

sample, (d(j; i) < Ej, where Ej is the dissimilarity between *j* i. If *j* is nearest to *h* than to the second nearest representative and the second nearest representative sample), the contribu-

Cjih = d(j; h)— d(j; i). tion of *j* to the swap between *i* and *h* is given by

representative sample, (d(j; i) P Ej), then the contribution ii. If *j* is equally distant from *h* than from the second nearest of *i* to the swap is Cjih = Ej — Dj.

the total distance. The weight d*k* is 0 when variable *k* presents a null

value for *i* or *j*, or 1 otherwise. The contribution *dk* of a categoric variable for the total distance is 0 if both values are equal 1, or 0 otherwise. For the other types of variable, the contribution is the difference between the two values divided by the range of values

*ij*

*ij*

of the variable.

* 1. *Data classification*

In addition to cluster analysis, data mining also offers techniques to perform predictive analyses called classification. Classification models implement supervised learning, in which a certain information represented by a class label of an unknown sample can be predicted based on previous observation of labeled samples. These labeled samples are also called training set. The final product of supervised learning is a classifier that is

mathematically described by a function ^*f* : *X* → {*true*; *false*} which uses a subset of *D* and ^*f* (*x*)ﬃ *f* (*x*), being D a set of labeled samples and *D* = {(*x*; *y*)|*x* ∈ *Sey* = *f* (*x*)} for a labeling function *f* : *X* → {*true*; *false*} [[20]](#_bookmark21).

After clustering the data, we developed classification models

capable of predicting in which cluster an arbitrary family would be associated to. We selected three of the most popular classifica- tion models in the data classification literature: artificial neural networks, support vector machines and decision trees.

* + 1. *Decision trees*

Decision trees [[31]](#_bookmark32) refers to one of the oldest classification mod- els. Each variable of the data set is individually questioned, and all the questions and answers can be arranged in a hierarchical struc- ture called a tree. Each node of the tree refers to a variable, and each edge originated in a node represents a value, or a range of

values, for the variable represented by the node. Leaf nodes store the class labels, the final point of the classification. Whenever a new test sample is analyzed, each one of the variables are ques- tioned, following a preexisting path along the tree until a leaf node is reached and its corresponding class label is set as the class label predicted for the analyzed sample. The recursive Hunt algorithm is one of the most commonly used algorithms for building decision

trees. The idea behing it is that the nodes of the tree should be

study, we use a classification model of this family called Multilayer Perceptron (MLP).

A perceptron is a structure composed of *n* input nodes that receive the values from the feature variables which describe the samples. Each node from this initial layer is linked to the output node through an edge that is associated to a weight value *w*. The output value is determined by the equation:

arranged in order to maximize the information gain, i.e., the ques- tioned variables should be capable of dividing the data set in pure partitions, with larger frequencies of a determined class.

*n*

*f* =

X

*k*=1

*wkxk*! — *t*

Another advantage of decision trees is the ease of visualization and interpretation. The decision rules exposed on the paths allow us to generate hypotheses regarding the individual influence of each variable in the classification process. Therefore, although they are used primarily as classification models, in this study we employ decision trees in order to interpret the clustering results and visualize which variables are the most relevant to the separa- tion of the clusters [[25,35]](#_bookmark22).

* + 1. *Support vector machines*

Support vector machine (SVM) is a popular classification model in the recent data mining literature due to its efficiency and empir- ical success. Data sets usually have more than one decision bound- ary capable of dividing the data in the n-dimensional space, and the goal of SVM is to compute the hyperplane with the largest mar- gin possible to act as a decision boundary [[35]](#_bookmark29). When applied to data that is linearly separable, the decision boundary has a linear representation given by the equation:

*w* · *x* + *b* = 0

where *x* refers to the variable set of an arbitrary sample *e,* and *w* is a

set of weights whose linear combination gives the class label *y* for *e*. The width of the decision boundary margin is given by the distance between the two parallel hyperplanes (support vectors) which touch the samples from each class that is nearest to the decision boundary, and is defined as [[7,10]](#_bookmark17):

*d* = 2

# ||*w*||

Finally, the goal of SVM is to find the decision boundary with the

largest margin possible, which means minimizing the following objective function:

# ||*w*||2

The class label is 1 if *f* > 0 and 0 otherwise. The model’s perfor-

mance can be improved by systematic adjustments in the weigth values *w*1; *w*2; ... ; *wn*. The MLP has hidden layers between the initial layer and output layer which works with backpropagation: the pre- diction errors obtained during the training phase are backpropa-

gated to the previous layers, and this error value is used to adjust the weights in each edge. The training phase with the implementa- tion of the backpropagation is summarized in the following steps [[13]](#_bookmark17):

1. Initialize the neural network with the weight values in each edge
2. Read the first input sample
3. Propagate the feature values of the given samples through the neural network until an output value is obtained
4. Compute the error value by comparing the output value obtained with the expected output value
5. Propagate the error back through the network
6. Adjust the weigth values in order to minimize the overall clas- sification error
7. Repeat steps 2–7 for new training samples, until the overall error is minimized

*2.2.4. Evaluating the performance of classification models*

The evaluation of classification models is performed by using a set of training samples to train the model and a set of test samples formed by unknown samples to test it. In the k-fold cross valida- tion method, the data set D is randomly divided into *k* mutually

exclusive subsets *D*1; *D*2; ... ; *Dk* (folds) of equal or similar size.

iteration *t* ∈ {1; 2; ... ; *k*}, the folds *D* — *Dt* are used to train the The classification model is trained and tested *k* times, and in each model and the remaining fold *Dt* is used to test it [[12]](#_bookmark17). Given the

classification results obtained during each test phase, the accuracy of the model is computed as:

min*w* 2 *c*

When handling data that is not linearly separable, SVM projects the data from its original n-dimensional space into a new space where the samples can be separated by a linear decision boundary. This process is achieved with the aid of kernel functions *K*(*x*,*y*). Kernel functions express the similarity between two samples in the new transformed dimensional space according to their dot product. In this study, we use the radial basis kernel function (RBF) to trans- form the data, given by the equation:

*K*(*x*; *y*)= exp(—c||*x* — *y*||2)

* + 1. *Artificial neural networks*

Artificial neural networks are inspired in the cognitive system and neurological functions of the human brain, simulating its neu- ron and links. A neuron has a filament called axon which connects to another neuron through dendrites, and the connection point is called synapse. Similarly, a classification model based on artificial neural networks is composed of interconnected nodes [[35]](#_bookmark29). In this

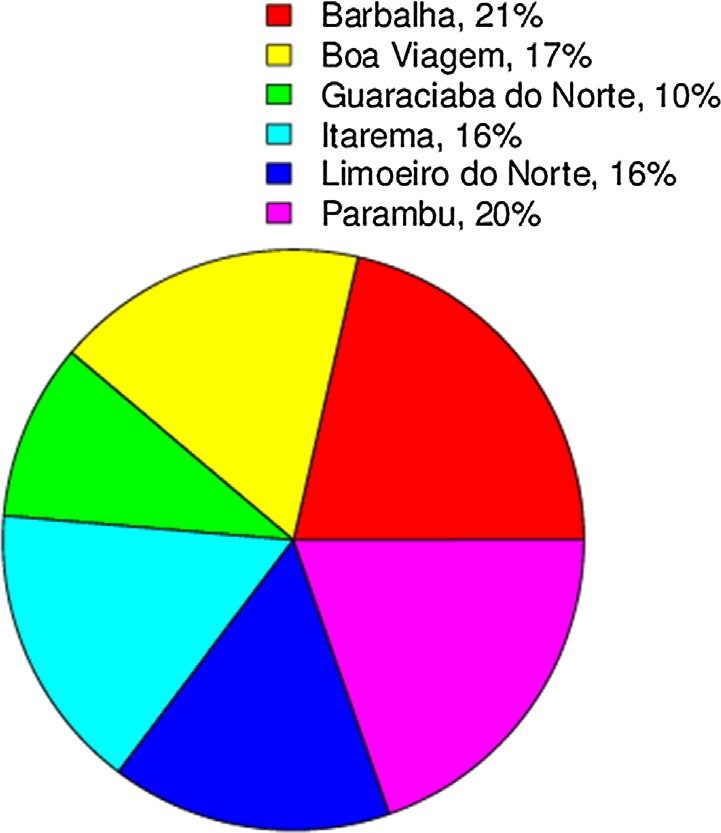
*Accuracy*(%)= *n* × 100%

where *c* is the number of test samples which were correctly classi-

fied, and *n* is the total number of test samples. The final accuracy of the model estimated by the k-fold cross validation is the average of all accuracies obtained in each iteration.

* 1. *Feature selection*

Feature selection is a data mining process which aims for the removal of feature variables which are considered unimportant of the classification. In general, we say that a feature is considered important for the analysis if it is relevant and irredundant [[9]](#_bookmark17). A variable is considered redundant when it presents high depen- dence to the other variables and the information contained in it can be expressed by a fewer number of these variables. A variable is considered irrelevant when the information it contains does not contribute to generate hypotheses regarding the samples with relation to their class labels. The removal of such variables can

improve the accuracy of classification models, reduce outliers and simplify the analysis [[11,18]](#_bookmark17).

* + 1. *Correlation Based Feature Selection*

CFS (Correlation Based Feature Selection) is a feature selection method proposed by Hall [[19]](#_bookmark23) which employs a heuristic based on correlations to estimate the importance of a subset of features. This method generates subsets with different combinations of fea- tures, and each subset is evaluated by the following Pearson corre- lation equation:

*kr*

*M cf*

# *s* = p*k*ﬃﬃﬃﬃ+ﬃﬃﬃﬃﬃ*k*ﬃﬃ(ﬃﬃ*k*ﬃﬃﬃﬃ—ﬃﬃﬃﬃﬃ1ﬃﬃ)ﬃﬃ*r*ﬃﬃﬃﬃﬃ

being *k* the total number of features, *Ms* is the merit of a feature

*ff*

subset *S*, *r**cf* is the average correlation between the features and the class label, and *rff* is the average intercorrelation between the features. Irrelevant features present low value for *rcf* and redundant features present high value for *rff* , and both cases lead to a mini- mization of the *Ms* value. Finally, the feature subset which achieved the highest merit is selected as the best feature subset.

1. Results
   1. *Groups identified*

The first step of the analysis was estimating the empirically best number of clusters within the data set. We determined the average width of the silhouette for 2–50 clusters, and we found that the analyzed families could be divided into a best number of 2 clusters. The PAM clustering algorithm was applied, and the data was divided into 2 clusters of 224 and 252 samples, respectively. A t- SNE (t-Distributed Stochastic Neighbor Embedding) graph for the partition is presented in [Fig. 2](#_bookmark10), allowing the visualization of the formed clusters in a bidimensional plot. The frequency of families from each city in the clusters is shown in [Figs. 3 and 4](#_bookmark8).

In [Fig. 3](#_bookmark8), we can see that families from Barbalha and Parambu cities are more frequent in Cluster 1, corresponding to 21% and 20% respectively of the total families associated to this cluster. This percentage varies between 10% and 17% for the other cities. Fami- lies from Guaraciaba do Norte are less frequent in this cluster, accounting for 10% of the families. This scenario is reversed for Cluster 2, presented in [Fig. 4](#_bookmark9). This cluster has more families from Guaraciaba do Norte (23%) and fewer families from Barbalha and Parambu (13% both). [Figs. 5–10](#_bookmark11) bring another visualization of this result, presenting the frequency of the two clusters in each city. As a matter of fact, we can see that families from Boa Viagem ([Fig. 6](#_bookmark12))

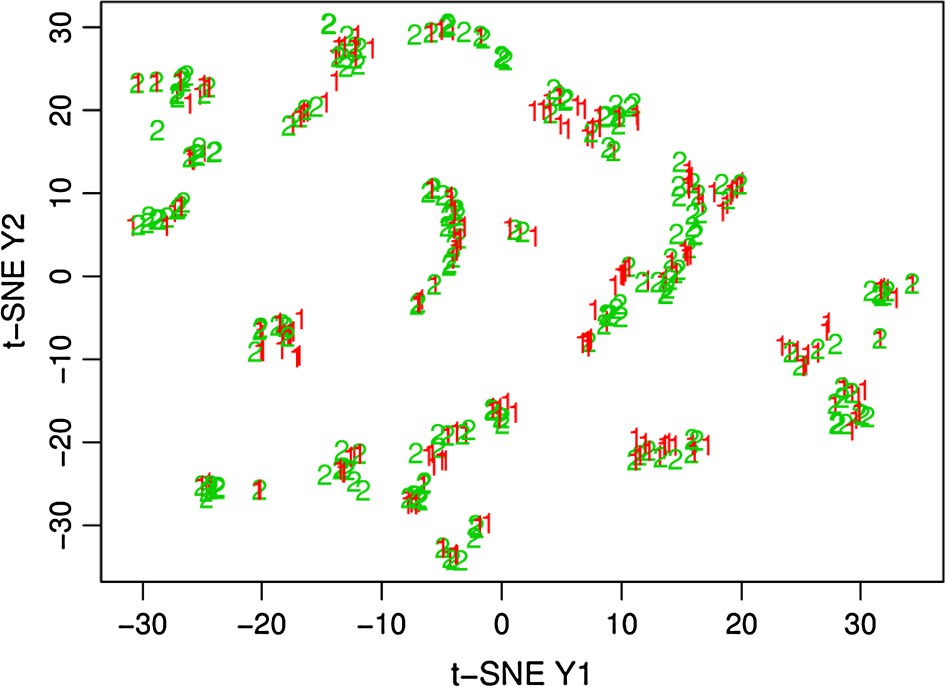


Fig. 2. The t-SNE bidimensional graph for the found clusters.

Fig. 3. Frequency of cities in Cluster 1. 21% Barbalha, 17% Boa Viagem, 10% Guaraciaba do Norte, 16% Itarema, 16% Limoeiro do Norte, 20% Parambu.

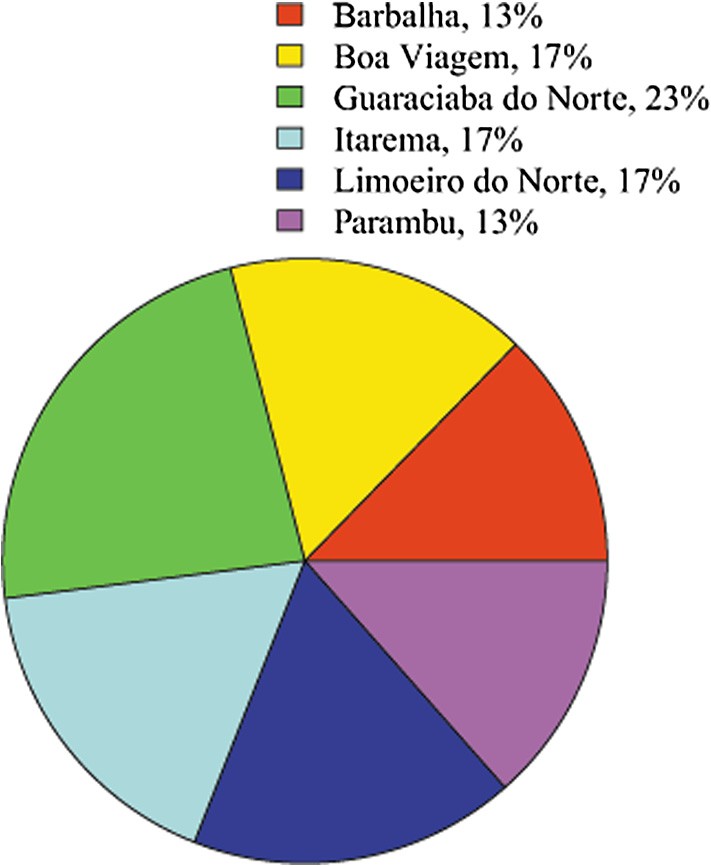


Fig. 4. Frequency of cities in Cluster 2. 13% Barbalha, 16% Boa Viagem, 23% Guaraciaba do Norte, 17% Itarema, 17% Limoeiro do Norte, 13% Parambu.

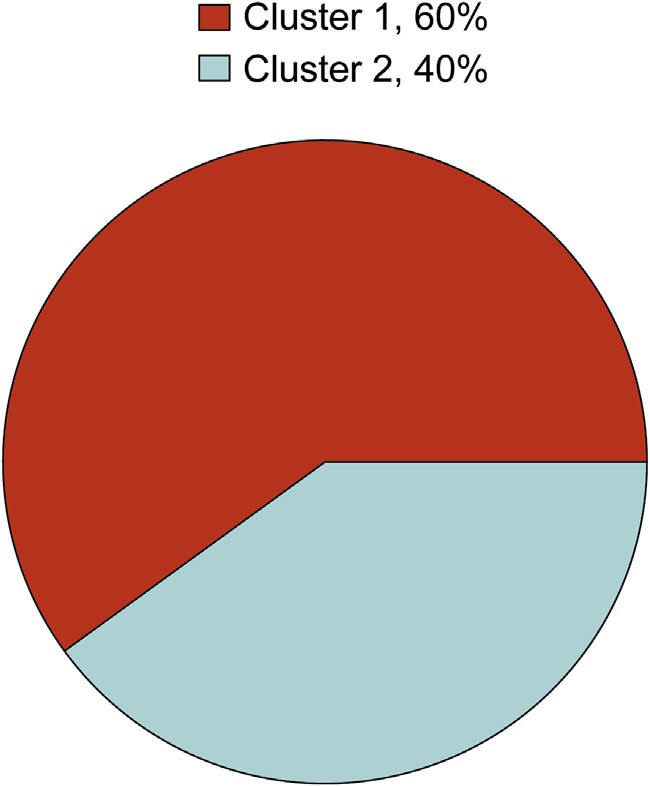


Fig. 5. Barbalha. 60% Cluster 1, 40% Cluster 2.

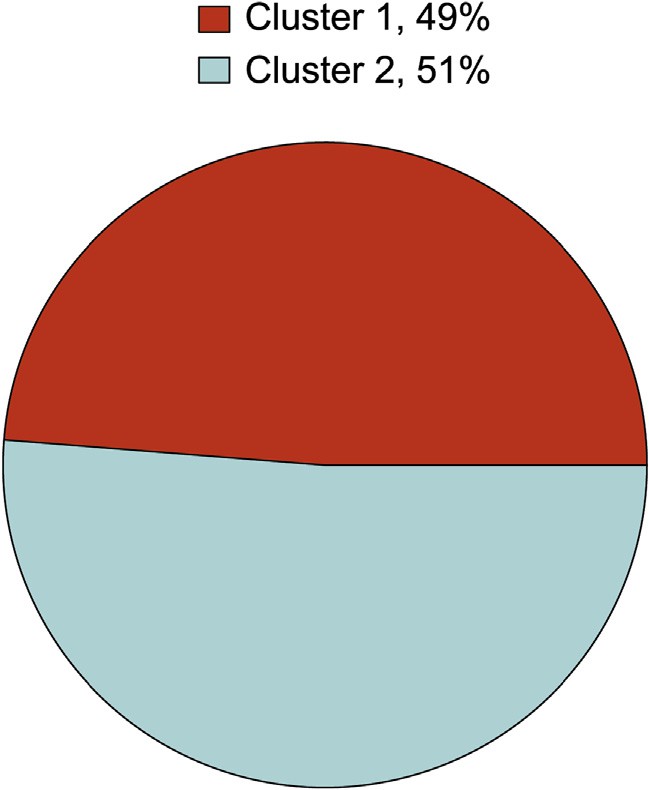
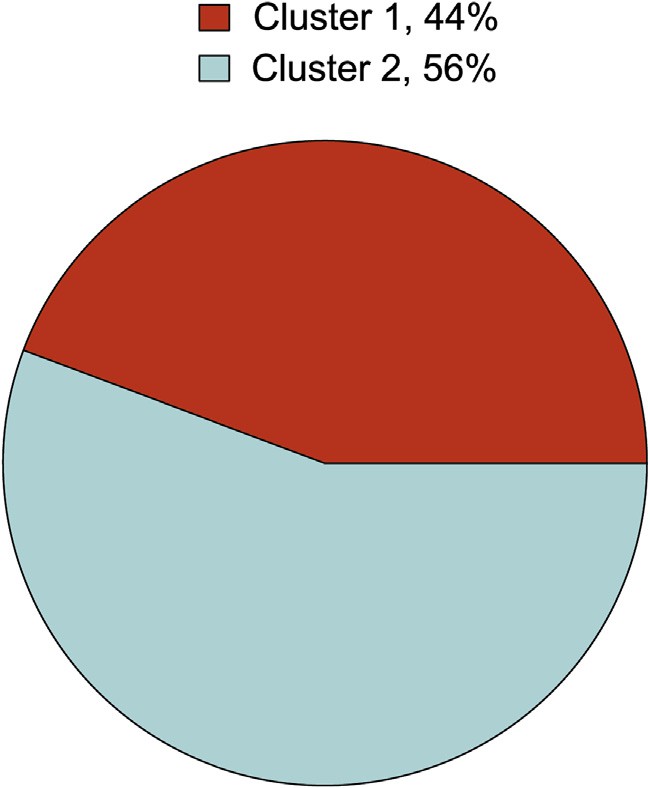
 

Fig. 6. Boa Viagem. 49% Cluster 1, 51% Cluster 2. Fig. 9. Limoeiro do Norte. 44% Cluster 1, 56% Cluster 2.

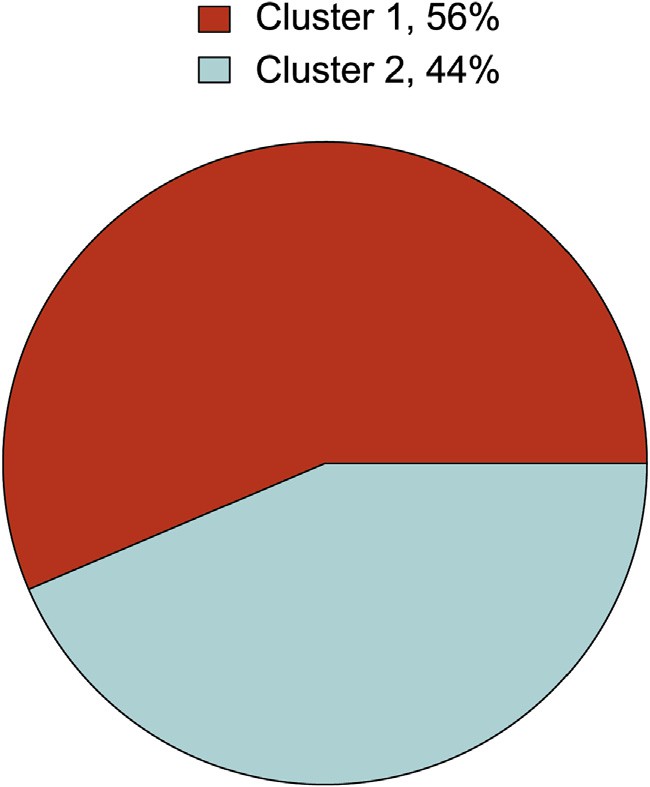
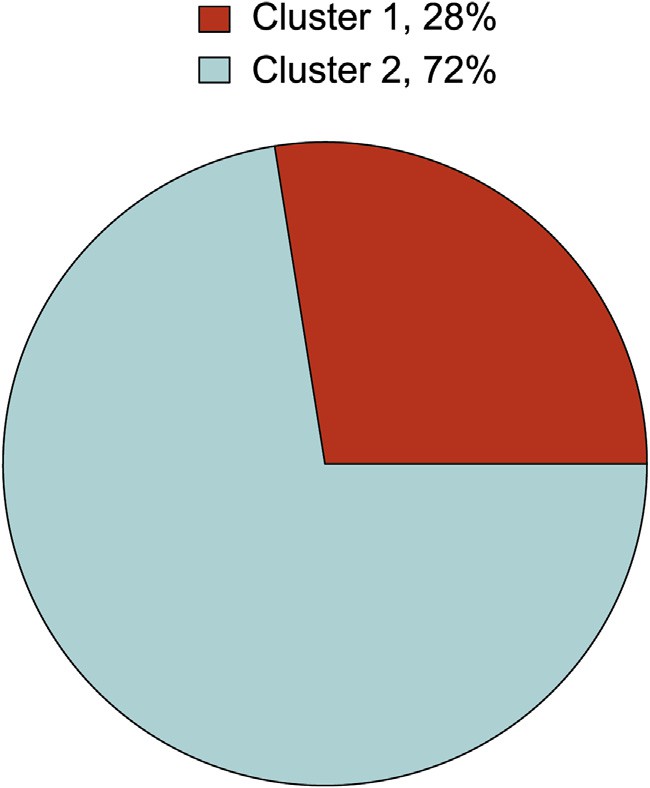


Fig. 7. Guaraciaba do Norte. 28% Cluster 1, 72% Cluster 2.

Fig. 10. Parambu. 56% Cluster 1, 44% Cluster 2.

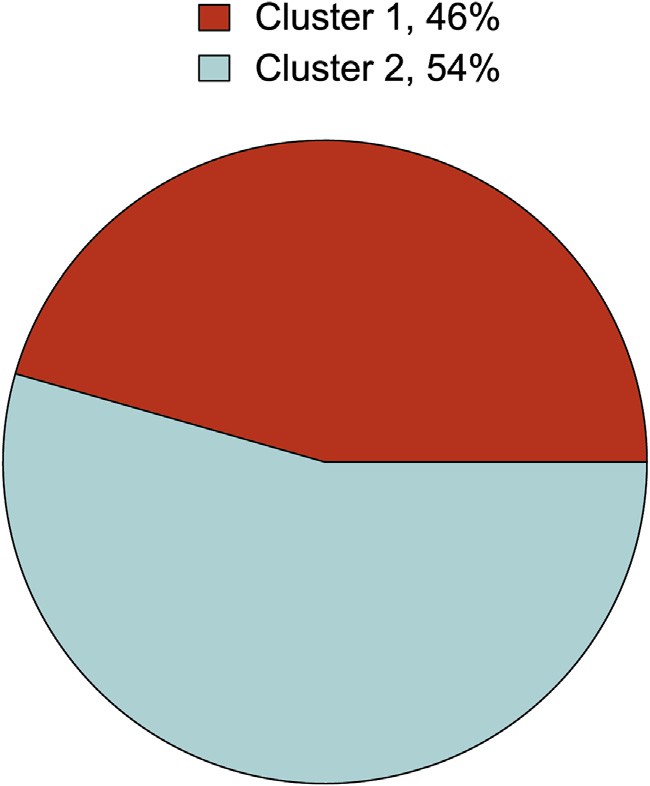


Fig. 8. Itarema. 46% Cluster 1, 54% Cluster 2.

and Itarema ([Fig. 7](#_bookmark13)) are equally distributed in both clusters. Fami- lies from Guaraciaba do Norte were mostly associated to Cluster 2 (72%), and families from Barbalha were mostly associated to Clus- ter 1 (60%).

* 1. *Feature selection and classification*

After determination of the clusters, we considered the cluster value of each sample as class label and proceeded to develop clas- sification models capable of predicting in which cluster a new, unknown family, would be associated to. We also applied the CFS algorithm to determine the most relevant features which differen- tiate the families of the two clusters. The best feature subset esti- mated by the CFS is shown in [Table 1](#_bookmark5).

The prediction results obtained by the classification model before and after feature selection are presented in [Table 3](#_bookmark14). It is vis- ible that both SVM and MLP models achieved best accuracy when all feature variables are used for training. The classification model which achieved the best performance was the MLP with 93.48% prediction accuracy when all variables are used.

Table 3

Accuracy values achieved by the classification models before and after feature selection is performed.

Model Before CFS After CFS

SVM 92.22% 86.76%

MLP 93.48% 85.50%

Decision trees 81.51% 85.50%

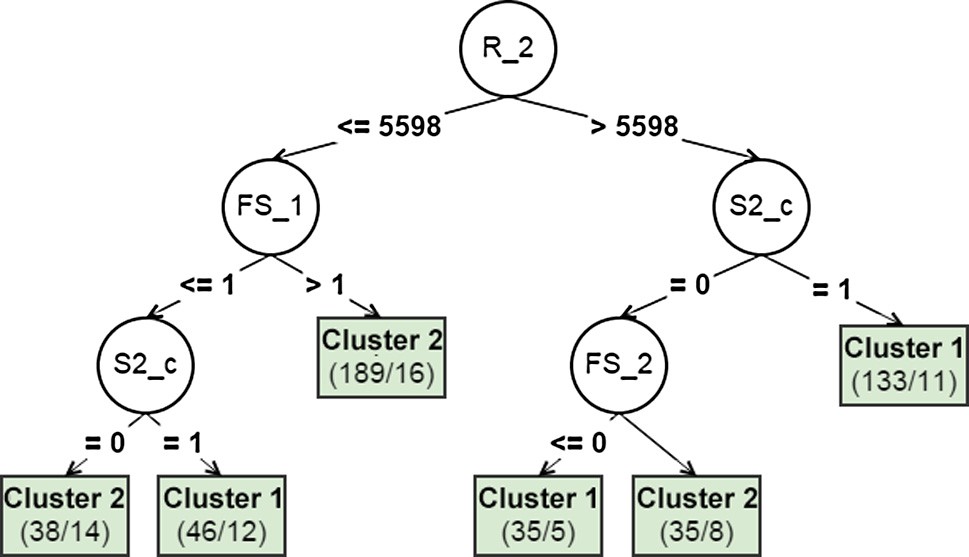


Fig. 11. Decision tree built for the analyzed data set, considering only the most important features selected by the CFS method and the cluster value as class label.

* 1. *Interpretation and discussion*

We used Weka software (version 3.6) to build a decision tree considering the cluster value of the samples as class label and only the most important features determined by the CFS as input vari- ables. The algorithm employed was the C4.5, which yielded 85.5% prediction accuracy. The structure of the decision tree achieved is shown in [Fig. 11](#_bookmark15).

The first feature selected by the C4.5 to divide the samples and which present the highest information gain is R\_2 (annual house- hold income from social security), which means that we can orga- nize the results in two cases: when the families have annual household income from social security relatively higher than the others (> R$5598), and otherwise.

Considering the families which have annual income from social security higher than R$5598, we observe the following patterns:

* Cluster 1: in this cluster, the families with higher income from social security give support to others more frequently, other-

wise they are characterized by a minimum access index.

* Cluster 2: in this cluster, the families with higher income do not give support to others frequently, and have higher access index

than the families with higher income associated to Cluster 1.

Considering the families which have annual income from social security less than R$5598, we observe the following patterns:

* Cluster 1: in this cluster, families with lower income from social security have the minimum value of food security index (1), but

they give support to other families more frequently.

* Cluster 2: in this cluster, families with lower income from social security have a higher food security index, and families with

low food security index do not give support to other families frequently.

An individual analysis of the variables and decision rules pre- sented in the tree’s structure lead us to the following conclusions:

* Annual household income from social security (R\_2): Consider- ing the prediction error rate, the tree shows that a total of 160

families (71%) associated to Cluster 1 have income from social

security higher than R$5598, while only 43 families (17%) of the families in Cluster 2 have income from social security also higher than this value. Therefore, the majority of families asso- ciated to Cluster 1 have income from social security higher than families associated to Cluster 2.

* Food security index (FS\_1): This variable considers whenever

the family has income from social security lower than R$5598.

Given the 64 families from Cluster 1 which have this income, 48 families (21% of the total of families of Cluster 1) are charac- terized by minimum food security index. On the other hand, in Cluster 2, 173 families (69% of the total of families associated to Cluster 2) have income from social security lower than R$5598 and food security index higher than the minimum. Therefore, in the few cases where a family from Cluster 1 has relatively low income from social security, it is likely that its food security index is also minimum. Moreover, the majority of families from Cluster 2 have income from social security lower than the fam- ilies from Cluster 1, however, they present higher food security index overall.

* Household gave support to other families (S2\_c): This variable is

always considered to determine if a family belongs to Cluster 1.

Around 70% of the families associated to Cluster 1 do give sup- port to other families. Considering the remaining 30% families that do not give support, 13% have relatively high income from social security, but are characterized by the minimum access index (FS\_2), and 11% have relatively low income from social security and minimum food security index. For Cluster 2, we can see that, considering the 43 families who have income from social security higher than R$5598, 32 of them did not give sup- port the other families. Therefore, we can conclude that families associated to Cluster 1 usually gave support to other families.

Overall, the majority of families from Cluster 1 receive income from social security higher than the families in Cluster 2, and gave support to other families even in the rare cases where they show lower income from social security and minimum food security index. When families of this cluster did not gave support, they pre- sent a minimum access index. As discussed previously, the major- ity of families from Barbalha (60% of the analyzed families) and Parambu (56%) belong to this cluster.

On the other hand, families from Cluster 2 received lower income from social security, but they had a higher access index when they received lower income, and present higher food secu- rity index when the income from social security was low. As dis- cussed previously, most families from Boa Viagem, Guaraciaba do Norte, Itarema and Limoeiro do Norte are associated to this cluster – approximately 51%, 72%, 54% and 56%, respectively.

1. Conclusion

We presented a data mining study and cluster analysis of social data obtained from small producers and family farmers from six municipalities in Ceará state, Northeastern Brazil. The social data were obtained through research conducted directly with the fam- ilies, which involved personal questions regarding demography, economics, agricultural production, and food security. The answers obtained were used as feature variables for our analysis. In order to work with this challenging data, we employed a methodology cap- able of handling variables of mixed types and null values. We esti- mated the best number of clusters within the data with the Silhouette technique, and the PAM clustering algorithm was employed in order to partition the data in the clusters. With the aid of the C4.5 decision tree, we observed that, overall, families from Cluster 1 received higher income from social security than the families associated to Cluster 2, and gave more support to other

families, even in the rare cases where the families presented lower income from social security and a in minimum value of food secu- rity. When families of the cluster did not give support to others, they present a minimum access index. Most of the analyzed fami- lies from Barbalha and Parambu belong to this cluster. On the other hand, families from Cluster 2 receive less income from social secu- rity, but they have a higher access index when the income from social security is higher. These families also have a higher food security index, even when the income from social security is low. Most of the analyzed families from Boa Viagem, Guaraciaba do Norte, Itarema and Limoeiro do Norte are associated with this cluster.

The proposed methodology presented promising results, how- ever, there are a few issues that could be addressed in future works.

First, we worked with a relatively small data set, and the anal- ysis of a larger number of families could help us uncover more trends regarding their profile. Moreover, we investigated families from Ceará state only, and it would be interesting to extend our analysis to families from other states as well. Piauí, Rio Grande do Norte, Paraíba and Pernambuco are states bordering Ceará and located in the Northeast region as well, sharing very similar geological, climatic and demographic characteristics, and families from these states could be inserted into the analysis in order for us to investigate demographic, economic and social profiles of small producers and family farmers of the Northeast region as a whole instead of one state only.

Also, working with categorical values is one of the major chal- lenges for data mining research~~es~~. We expect, for future works, to improve the processability of our data set, either by standardiz- ing our variables to numeric values only, changing the data collec- tion method (i.e., replace the common surveys) or by employing forthcoming clustering and classification techniques in the data mining literature that can handle categorical variables.

Declarations of interest

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. [A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0005) [categorical data, Data Knowl. Eng. 63 (2007) 503–527](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0005).
2. [G. Barbier, H. Liu, Data Mining in Social Media, in: Social Network Data](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0010) [Analytics,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0010) [Springer, US, Boston, MA, 2011, pp. 327–352](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0010).
3. [R.M. Barbosa, B.L. Batista, C.V. Barião, R.M. Varrique, V.A. Coelho, A.D.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0015) [Campiglia, F. Barbosa, A simple and practical control of the authenticity of](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0015) [organic sugarcane samples based on the use of machine-learning algorithms](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0015) [and trace elements determination by inductively coupled plasma mass](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0015) [spectrometry, Food Chem. 184 (2015) 154–159](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0015).
4. [R.M. Barbosa, B.L. Batista, R.M. Varrique, V.A. Coelho, A.D. Campiglia, F.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0020) [Barbosa, The use of advanced chemometric techniques and trace element](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0020) [levels for controlling the authenticity of organic coffee, Food Res. Int. 61 (2014)](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0020) [246–251](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0020).
5. C.M. Bishop, Pattern Recognition and Machine Learning, Springer – Verlag New York, 2006.
6. [R.P. Bunker, F. Thabtah, A Machine Learning Framework for Sport Result](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0030) [Prediction, Appl. Comput, Informatics. 15 (2019) 27–33](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0030).
7. [C.J.C. Burges, A tutorial on support vector machines for pattern recognition,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0035) [Data Min. Knowl. Discov. 2 (1998) 121–167](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0035).
8. [J.N.B. Campos, Paradigms and public policies on drought in Northeast Brazil: a](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0040) [historical perspective, Environ. Manage. 55 (2015) 1052–1063](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0040).
9. [G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0045) [Electr. Eng. 40 (2014) 16–28](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0045).
10. [C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0050)

[297](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0050).

1. [M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997)](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0055) [131–156](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0055).
2. R.O. Duda, P.E. Hart (E. Peter, D.G. Stork). Pattern classification. Wiley, 2001.
3. [M. Gardner, S. Dorling, Artificial neural networks (the multilayer perceptron)—](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0065) [a review of applications in the atmospheric sciences, Atmos. Environ. 32](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0065) [(1998) 2627–2636](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0065).
4. [J.C. Gower, A general coefficient of similarity and some of its properties,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0070) [Biometrics 27 (1971) 857](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0070).
5. Guanziroli, C.E., 2001. Agricultura familiar e reforma agrária no século XXI. Garamond.
6. [C.E. Guanziroli, A.M. Buainain, A. Di Sabbato, Dez anos de evolução da](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0080) [agricultura familiar no Brasil: (1996 e 2006), Rev. Econ. e Sociol. Rural 50](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0080) [(2012) 351–370](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0080).
7. P. Gundecha, H. Liu, Mining social media: a brief introduction, in: Tutorials in Operations Research. INFORMS, 2012, pp. 1–17.
8. [I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0090) [Learn. Res. 3 (2003) 1157–1182](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0090).
9. M.A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, in: Pat Langley (Ed.), Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000). Morgan Kaufmann Publishers Inc., Stanford, CA, USA, 2000, pp. 359–366.
10. [L. Hamel, Knowledge Discovery with Support Vector Machines,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0100) [Wiley-](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0100) [Interscience, 2009](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0100).
11. [Z. Huang, Extensions to the k-means algorithm for clustering large data sets](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0105) [with categorical values, Data Min. Knowl. Discov. 2 (283–304) (1998) 1](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0105).
12. [A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0110) [(2010)](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0110).
13. L. Kaufman, P. Rousseeuw, Clustering by means of medoids, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L1-Norm and Related Methods, 1987, pp. 405–416.
14. [L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0120) [Analysis,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0120) [Wiley, 2005](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0120).
15. [D.M. Khan, N. Mohamudally, An integration of K-means and decision tree (ID3)](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0125) [towards a more efficient data mining algorithm, J. Comput. 3 (2011) 76–82](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0125).
16. [C. Maione, E.S. De Paula, M. Gallimberti, B.L. Batista, A.D. Campiglia, F. Barbosa,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0130)

[R.M. Barbosa, Comparative study of data mining techniques for the](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0130) [authentication of organic grape juice based on ICP-MS analysis, Expert Syst.](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0130) [Appl. 49 (2016) 60–73](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0130).

1. [C. Maione, B. Lemos, A. Dobal, F. Barbosa, R. Melgaço, Classification of](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0135) [geographic origin of rice by data mining and inductively coupled plasma mass](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0135) [spectrometry, Comput. Electron. Agric. 121 (2016) 101–107](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0135).
2. [A. Mucherino, P. Papajorgji, P.M. Pardalos, A survey of data mining techniques](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0140) [applied to agriculture, Oper. Res. 9 (2009) 121–140](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0140).
3. [M.R. Naroui Rad, S. Koohkan, H.R. Fanaei, M.R. Pahlavan Rad, Application of](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0145) [Artificial Neural Networks to predict the final fruit weight and random forest](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0145) [to select important variables in native population of melon (*Cucumis melo* L.),](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0145) [Sci. Hortic. (Amsterdam) 181 (2015) 108–112](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0145).
4. [E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0150) [customer relationship management: a literature review and classification,](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0150) [Expert Syst. Appl. 36 (2009) 2592–2602](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0150).
5. [J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0155).
6. M.V.J. Reddy, B. Kavitha, Clustering the mixed numerical and categorical dataset using similarity weight and filter method, Int. J. Database Theory Appl. (2012) 5.
7. [P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0165) [validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0165).
8. [I. Scoones, Livelihoods perspectives and rural development, J. Peasant Stud. 36](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0170) [(2009) 171–196](http://refhub.elsevier.com/S2210-8327(17)30352-6/h0170).
9. P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining. Pearson Addison Wesley, 2005.