Classification Methods for SARLAFT Segmentation

Yuly Andrea Lizarralde Bejarano

Department of Mathematics
School of Sciences and Engineering
Sergio Arboleda University
Bogotá, Colombia
2016

Classification Methods for SARLAFT Segmentation

Yuly Andrea Lizarralde Bejarano

Undergraduate Thesis Work

Advisor: Phd(c) Martha Corrales

Department of Mathematics
School of Sciences and Engineering
Sergio Arboleda University
Bogotá, Colombia
2016

A los michis.

Acknowledgements

...mmmm

Abstract

Here, It is supposed to be a great abstract.

Keywords clustering, classification, business risk, SARLAFT

Contents

Chapter 1

Introduction

The purpose of this work is to make use of the classification methods theory to create a model which will be applied to risk management issues and allows effective customer segmentation linked to a financial institution. This taking into account that Colombian Financial Superintendence considers that customers or users are agents of money laundering and terrorist financing sources (hereinafter ML/TF). This segmentation, as defined in the external circular letter 022 of 2007, is the process by which the separation of elements takes place in homogeneous groups, within them and heterogeneous among them. The separation is based on the recognition of major differences in their characteristics (segmentation variables). The segmentation variables that will be considered for customers are: economic activity, volume or frequency of their transactions, amount of income and assets. As starting point, we will see the definitions of money laundering and terrorist financing as they are stipulated in Colombian law, what is SARLAFT [1] and what is requiered by supervised entities to implement this program. After that, will be given the relevant definitions of the main classification methods and

[1] Certification on risk administration system to prevent money laundry and terrorism financing, SARLAFT for its abbreviation in Spanish (Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo)

there will be a try to show how can the theory of a given classification method be applied to customer segmentation, performed by Financial Superintendence´s supervised entities in order to identify the risk of ML/TF.

Chapter 2

Preliminaries

2.1 Money Laundering

Money laundering is a try to give the legality appearance to illegal origin resources and it is a serious problem in our country, mainly because it is related to criminal activities such as drug trafficking, migrant traffic, human trafficking, extortion, illicit enrichment, kidnapping, arms trafficking, corruption and crimes against the financial system.

Article 323 of the Criminal Code defines the crime of money laundering as follows:

whoever acquires, safeguards, investments, transports, transforms, safekeeps or manages assets that have remote or immediate origin from activities of migrant smuggling, human trafficking, extortion, illicit enrichment, kidnapping, rebellion, arms trafficking, financing terrorism and activities related to terrorist management, traffic of drugs, narcotics or psychotropic substances, crimes against the financial system, crimes against public administration, or related to the proceeds of crime executed under conspiracy, or give to these resources or goods from such activities the appearance of legality or legalizes them, hide them or disguises their

true nature, origin, location, destination, movement or rights to such property or perform any other act to conceal or disguise their illicit origin. (Art.323, Law 599 of 2000)

2.2 Terrorist Financing Sources

In Colombia it is a crime financing terrorism or managing any resources related to terrorist activities.

It is meant by financing of terrorism: *Whoever directly or indirectly provides, collects, delivers, receives, manages, contributes, custodies or saves funds, assets or resources, or perform any other act promoting, organizing, supporting, maintaining, or economically sustaining illegal armed groups or any of their members, or national or foreign terrorist groups, or terrorist activities. (Art. 345, Law 599 of 2000).*

It is meant by the financing of terrorism: *Whoever directly or indirectly provides, collects, delivers, receives, manages, contributes, custodies or saves funds, assets or resources, or perform any other act promoting, organizing, supporting, maintaining, or economically sustaining illegal armed groups or any of their members, or national or foreign terrorist groups, or terrorist activities. (Art. 345, Law 599 of 2000).*

2.3 SARLAFT

It is the management system that must implement the supervised entities by the Colombian Financial Superintendence to protect against the risk of ML/FT and it

is implemented through the stages and elements that compose it. (Regulation 022 Colombian Superintendence of Finance).

According to Colombian Superintendence of Finance to identify the risk of ML/FT controlled entities as a minimum they must comply:

1. To establish methodologies for segmentation of risk factors.
2. Based on established methodologies, segment the risk factors.
3. To establish methodologies for risk identification ML/TF and the associated risks for each one of the risk factors.
4. Based on the established methodologies described in the developing of the preceding paragraph, identify ways through, which may arise a risk of ML/TF.

Customers or users represent one of the risk factors that for SARLAFT effects should be taken into account by supervised entities. According to the Colombian Financial Superintendence, the customer is any natural or legal person with whom the entity establishes and maintains a contractual or legal relationship for supplying any product of its own activity.

Taking this into consideration, the purpose of this document is to present a methodology for customer segmentation by using the classification methods theory.

2.4 Warning Signs

(From Regulation 022) Are the set of qualitative and quantitative indicators that ensure timely and/or prospectively identify atypical behaviors of relevant variables, previously determined by the entity.

2.5 Segmentation

(From Regulation 022) Is the process by which takes place the separation of elements in homogeneous groups, within them and heterogeneous between them. The separation is based on the recognition of significant differences in their characteristics (segmentation variables).

Chapter 3

Theory

3.1 Clustering

Clustering is defined as the process of classifying a large group of data items into smaller groups that share the same or similar properties. A cluster is a basic unit of classification of initial unclassified data based on common properties. There are various definitions of a cluster as follows:

* A cluster is a set of entities that are alike or a set of entities from different clusters that are not alike.

* A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point within the cluster and any other point outside the cluster.

* A cluster is a connected region of a multidimensional space containing a relatively high density of points.

* A cluster is a group of contiguous elements of a statistical population.

From these definitions, we can see that whether clusters consist of entities, points, or regions, the components within a cluster are more similar in some respects to each other than to other components outside of the cluster or to entities classified into other clusters. This description covers two important points. One is similar, which can be reflected with distance measures, and the other is classification, which suggests the objective of clustering. Therefore, clustering can be defined as a process of identifying groups of data that are similar in a certain aspect and building a classification among them. A clustering process usually involves the following steps:

* **Object Selection:** The entities to be clustered are selected in a manner such that the entities are representations of the cluster structures that are inherent in the data.

* **Variable Selection:** The variable that will represent the measurements of the entities must be selected. Correct selection of the variable will result in a meaningful cluster structure

* **Variable Standardization:** Since variables may be measured with different systems, they may initially be incomparable. To solve this problem, the variables are usually standardized, although this step is optional.

* **Similarity Measurement:** Similarity or dissimilarity between a pair of data items or among many items must be calculated. This will usually be the basis of a similarity matrix. Sometimes more than one attribute can be considered and analyzed. This is called multivariate analysis.

* **Clustering Entities:** Based on the similarity or dissimilarity measurement a pair of items can be compared and classified in the same group or different

groups. This process is applied to all items in one data record until the items can be classified into two clusters. Recursively following this step will result in a classification with various clusters.

3.1.1 *Distance Measures*

The most commonly used proximity measure, is the Minkowsky metric, which is a generalization of the normal distance between points in the Euclidian space. It is defined as

$$p_{ij} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^r \right)^{1/r}$$

where r is a parameter, d is the dimensionality of the data object, and x_{ik} and x_{jk} are, respectively, the k^{th} components of the i^{th} and j^{th} objects, x_i and x_j. The following is a list of the common Minkowski distances for specific values of r.

1. $r = 1$. City block (Manhattan, taxicap, L_1 norm) distance. A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors.

2. $r = 2$. Euclidean distance. The most common measure of the distance between two points.

3. $r = \infty$. "supremum" (L_{max} norm, L_∞ norm) distance. This is the maximum difference between any component of the vectors.

3.2 K-means

The *K-means* algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity measures in regard to the mean value of the objects in a cluster, which can be viewed as the cluster centroid or center of gravity. The *k-means* algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and m_i is the mean of cluster C_i (both p and m_i are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is square, and the distances are summed. This criterion tries to make the resulting k clusters, as compact and as separate as possible. **Algorithm:** *k-means*. The *k-means* algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.
Input:

* k: the number of clusters,

* D: a data set containing n objects.

Output: A set of k clusters. **Method:**

1) arbitrarily choose k objects from D as the initial cluster centers;

2) repeat

3) re-assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

5) until no change.

3.3 Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, an each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. A typical decision tree is shown in Figure. It represents the concept *buys cell phone*, that is, it predicts whether a customer at AllElectronics is likely to purchase a cell phone. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees, whereas others can produce nonbinary trees. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. The strategy for make a decision tree is as follows:

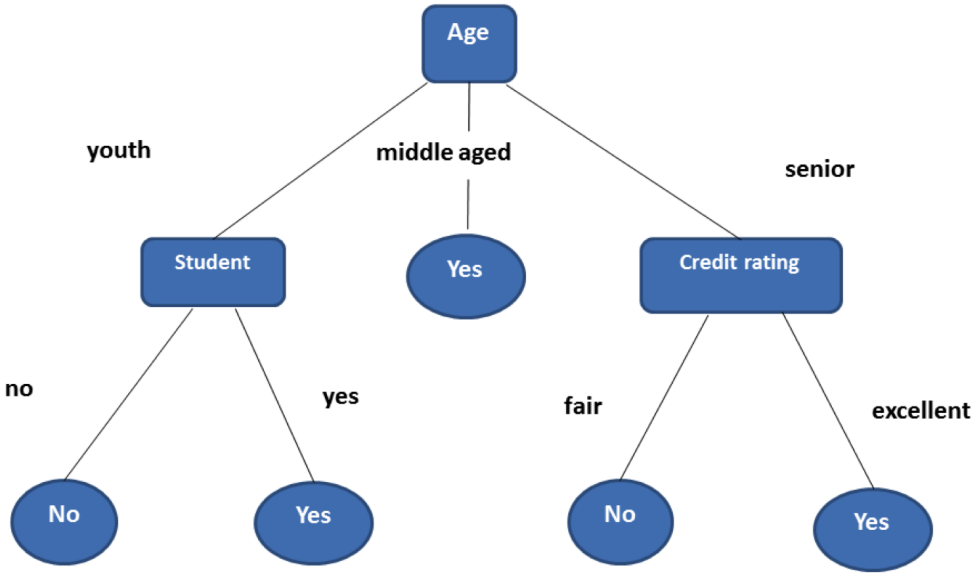

Figure 3.1: caption

* The algorithm is called with three parameters: *D*, *attribute list*, and *attribute selection method*. We refer to *D*as a data partition. The parameter *attribute list* is a list of attributes describing the tuples. *attribute selection method* specifies a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class.

* The tree starts as a single node, *N*, representing the training tuples in *D*.

* If the tuples in *D* are all of the same class, then the node *N* becomes a leaf and is labeled with that class.

* Otherwise, the algorithm calls *attribute selection method* to determine the splitting criterion. The splitting criterion tell us which attribute to test at node *N* by determining the best way to separate or partition the tuples in *D* into individual classes. The splitting criterion is determined so that, ideally,

the resulting partitions at each branch are as pure as posible. A partition is pure if all of the tuples in it belong to the same class.

* The node N is labeled with the splitting criterion, which serves as a test at the node. A branch is grown from node N for each of the outcomes of the splitting criterion. The tuples in D are partitioned accordingly. Let A be the splitting attribute. A has v distinct values, $\{a_1, a_2, \cdots, a_v\}$, based on the training data.

    1. A is *discrete valued:* In this case, the outcomes of the test at node N correspond directly to the known values of A. A branch is created for each known value, a_j, of A and labeled with that value.

    2. A is *continuous valued:* In this case, the test at node N has two possible outcomes, corresponding to the conditions $A \leq$ *split point* and $A >$ *split point*, respectively, where *split point* is the split-point returned by *attribute selection method* as part of the splitting criterion. Two branches are grown from N and labeled according to the above outcomes.

    3. A is *discrete valued* and a *binary tree* must be produced: The test at node N is of the form $A \in S_A$. S_A is the splitting subset for A, returned by *attribute selection method* as part of the splitting criterion. It is a subset of the known values of A. If a given tuple has value a_j of A and if $a_j \in S_A$, then the test at node N is satisfied. Two branches are grown from N.

* The algorithm uses the same process recursively to form a decision tree for the tuples at each resulting partition, D_j, of D.

* The recursive partitioning stops only when any one of the following terminating conditions is true:

    1. All of the tuples in partition D (represented as node N) belong to the same class, or
    2. There are no remaining attributes on which the tuples may be further partitioned.
    3. There are no tuples for a given branch, that is, a partition D_j is empty. In this case, a leaf is created with the majority class in D.

* The resulting decision tree is returned.

When a decision tree is built, many of the branches will reflect anomalies in the training data to noise or outliers. Tree pruning methods typically use statistical measures to remove the least reliable branches. In the pre-pruning approach, a tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples. There are difficulties, however, in choosing an appropriate threshold. High thresholds could result in oversimplified trees, whereas low thresholds could result in very little simplification. The second and more common approach is post pruning, which removes subtrees from a fully grown tree. A subtree of a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced.

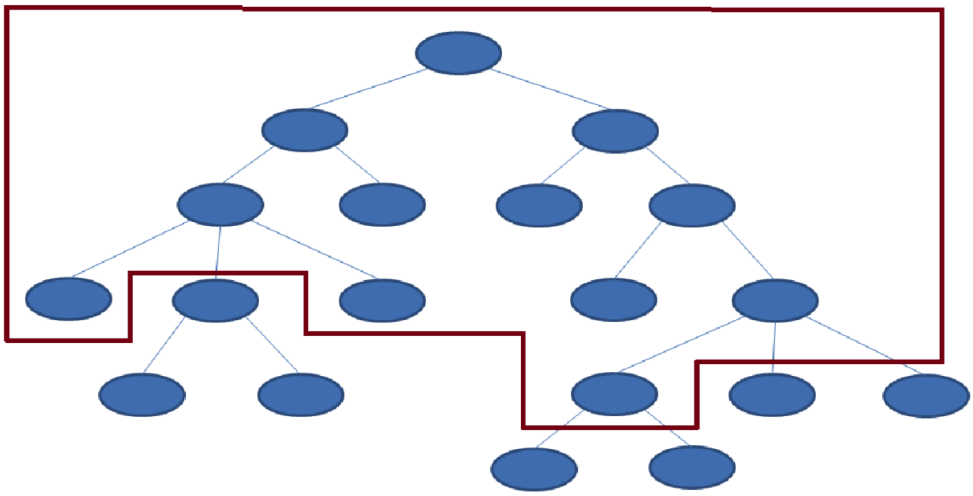

Figure 3.2: caption

Chapter 4

Problem Statement

4.1 Problem

The 11th circular letter of the Financial Superintendence of Colombia stipulates that all entities governed by it, should set a money laundering and terrorist financing risk management system (quote). Two methodologies are proposed to try to comply with the standard. The first comes from building clusters by using the K-Means method, and the second by building decision trees, methodology used in companies (citing training). It is intended to show whether through clusters can achieve an acceptable and applicable classification in the financial field.

These methodologies will be applied on the same database and the obtained results will be compared, on the assumption that valid results are obtained from decision trees.

4.2 Implementation

We will now explain the way how decision tree implementation and K-means were developed using the program SPSS.

4.2.1 Decision trees

We will use the SPSS statistical program to elaborate the customer segmentation through classification, decision trees.

The decision tree procedure makes a classification model based on trees and classifies cases into groups or predicts values of a dependent variable (standard) based on independent variable values (predictors).

The first thing to do is clean the information which is required to elaborate the model. We have a database that provides information such as Income, Assets and six months financial movements. As in most of the cases we find that there are some missing data, in this opportunity some those cases that did not have the necessary information that could affect segmentation were excluded (Asobancaria training).

Following the Financial Superintendence suggestions, we will take into account for customer segmentation the following variables:

1. Incomes
2. Assets
3. Credits volume

From the variables, asset, income and credit volume we calculate percentiles to round numbers to the nearest unit and be able to establish better groups. Then we get the following classification for the volume of credits:

* Lower than 3.000.000
* Between 3.000.000 and 15.000.000

* Between 15.000.000 and 30.000.000
* Between 30.000.000 and 60.000.000
* Between 60.000.000 and 120.000.000
* Between 120.000.000 and 210.000.000
* More than 210.000.000

for the variable incomes:

* Lower than 5.000.000
* Between 5.000.000 and 10.000.000
* Between 10.000.000 and 20.000.000
* Between 20.000.000 and 50.000.000
* More than 50.000.000

for the variable assets:

* Lower than 50.000.000
* Between 50.000.000 and 150.000.000
* Between 150.000.000 and 300.000.000
* Between 300.000.000 and 500.000.000
* Between 500.000.000 and 1.000.000.000
* More than 1.000.000.000

Next step consists on running the program. SPSS asks for a dependent variable that in this case is *Credits volume* (deposits made by every customer) and the independent variables that will be incomes and assets.

SPSS offers the following growing methods for the trees:

1. CHAID
2. Exhaustive CHAID
3. CRT
4. QUEST

We will select CRT due to it divides the data into the most possible homogeneous segments in relation to the dependent variable. Next graphic will show us the first division of the decision tree made taking into account incomes variable:

We analyze the tree's results and go about the subsequent pruning that corresponds to the CRT algorithm

Alerts determination

For determining the alerts of the obtained segments by mean of the decision trees, we will use the percentages showed by every node of the tree with customer movement's information:

The node shows us how many customers made some kind of movements and indicates the related percentage inside the segment, as well. In such a way we can say that if a lower percentage of customers make bigger movements than the rest of them, it is a warning sign because it is outside of the movement's expected range.

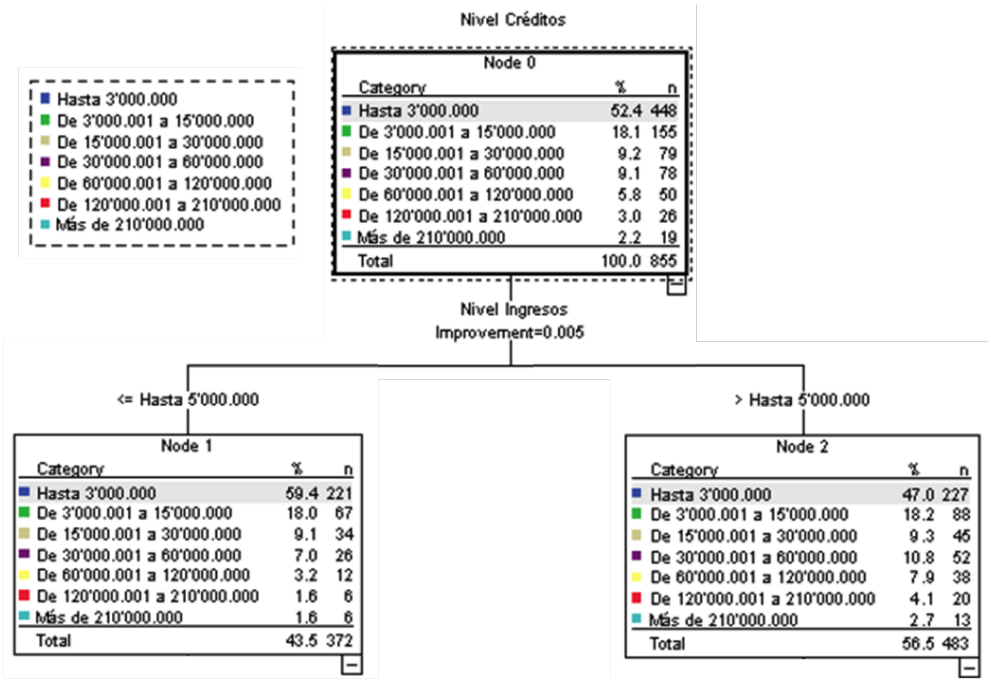

Figure 4.1: caption

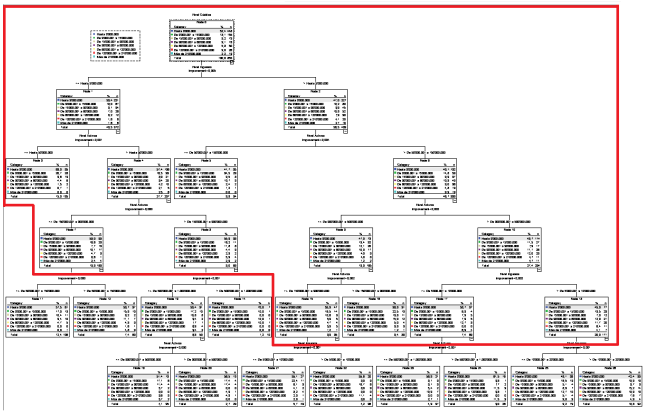

Figure 4.2: caption

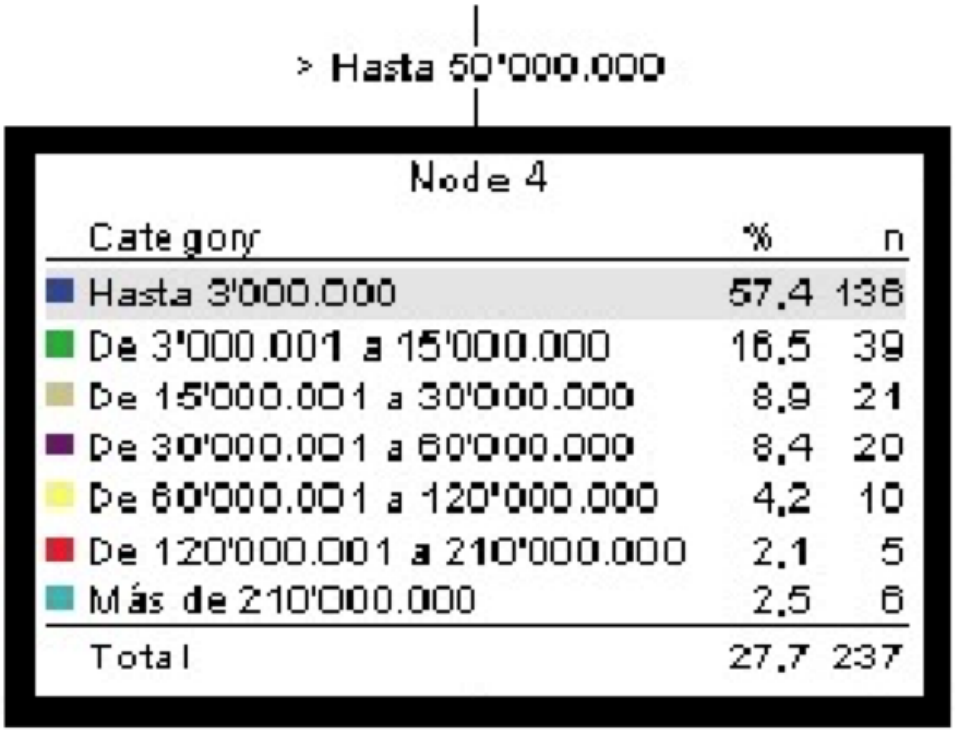

Figure 4.3: caption

4.2.2 K-Means Algorithm

We will use the SPSS statistical program to elaborate the customer's classification through K - means.

This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that can handle large numbers of cases. However, the algorithm requires you to specify the number of clusters. We can specify initial cluster centers if we know this information. In our case, we are going to determine cluster centers iteratively.

To use K - means, we have to keep in mind that the variables should be quantitative at the interval or ratio level. Also, we assume that the distances are computed using simple Euclidean distance.

Similarly to what was done with decision trees, we clean the information which is required to elaborate the model. We have a database that provides information such as Income, Assets and six months financial movements. As in most of the cases we find that there is some missing data, in this opportunity some those cases that did not have the necessary information that could affect segmentation were excluded (Asobancaria training). Following the Financial Superintendence suggestions, we will take into account for customer segmentation the following variables:

1. Incomes
2. Assets
3. Credits volume

Similarly to what was done with decision trees, we clean the information which is required to elaborate the model. We have a database that provides information

such as Income, Assets and six months financial movements. As in most of the cases we find that there are some missing data, in this opportunity some those cases that did not have the necessary information that could affect segmentation were excluded (Asobancaria training). for the variable incomes:

* Lower than 5.000.000
* Between 5.000.000 and 10.000.000
* Between 10.000.000 and 20.000.000
* Between 20.000.000 and 50.000.000
* More than 50.000.000

for the variable assets:

* Lower than 50.000.000
* Between 50.000.000 and 150.000.000
* Between 150.000.000 and 300.000.000
* Between 300.000.000 and 500.000.000
* Between 500.000.000 and 1.000.000.000
* More than 1.000.000.000

The variable Credits volume will be use to determine the alerts.

Alerts determination

For determining the alerts of the obtained segments by mean of the K - means, we will use the standard deviation of the variable credits volume, will be doing this for each segment.

Chapter 5

Results

5.1 Decision trees

From the implementation of a decision tree, we get the following segments and the respective warning signs for each segment.

* **Segment 1:** Income lower or equal than 5.000.000 and assets, lower or equal than 50.000.000.

  The warning sign for this segment is the transactions made by the costumers above 2.500.000 per month. 16.2% of the customers in the segment 1, made transactions over 2.500.000.

* **Segment 2:** Income lower or equal than 5.000.000 and assets between 50.000.000 and 300.000.000.

  The warning sign for this segment is the transactions made by the customers above 5.000.000 per month. 17.8% of the customers in the segment two, made transactions over 5.000.000.

* **Segment 3:** Income lower or equal than 5.000.000 and assets more than

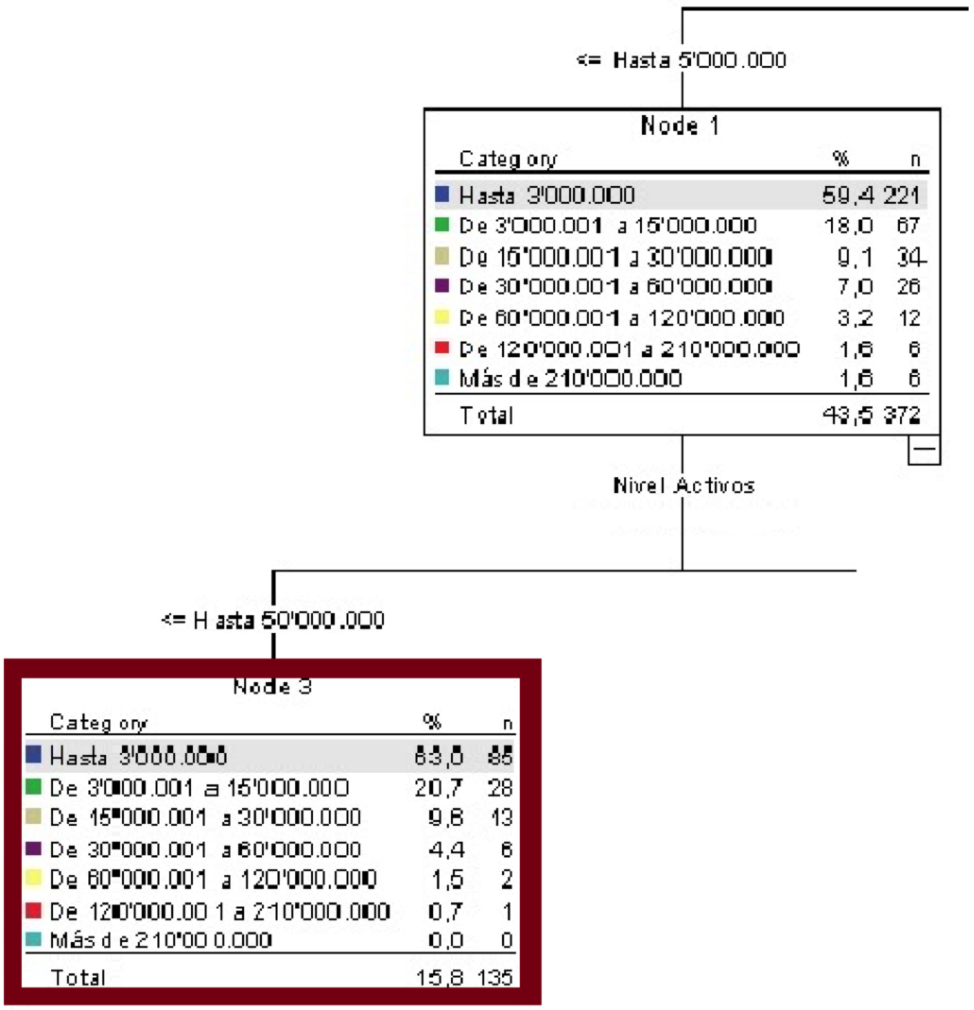

Figure 5.1: caption

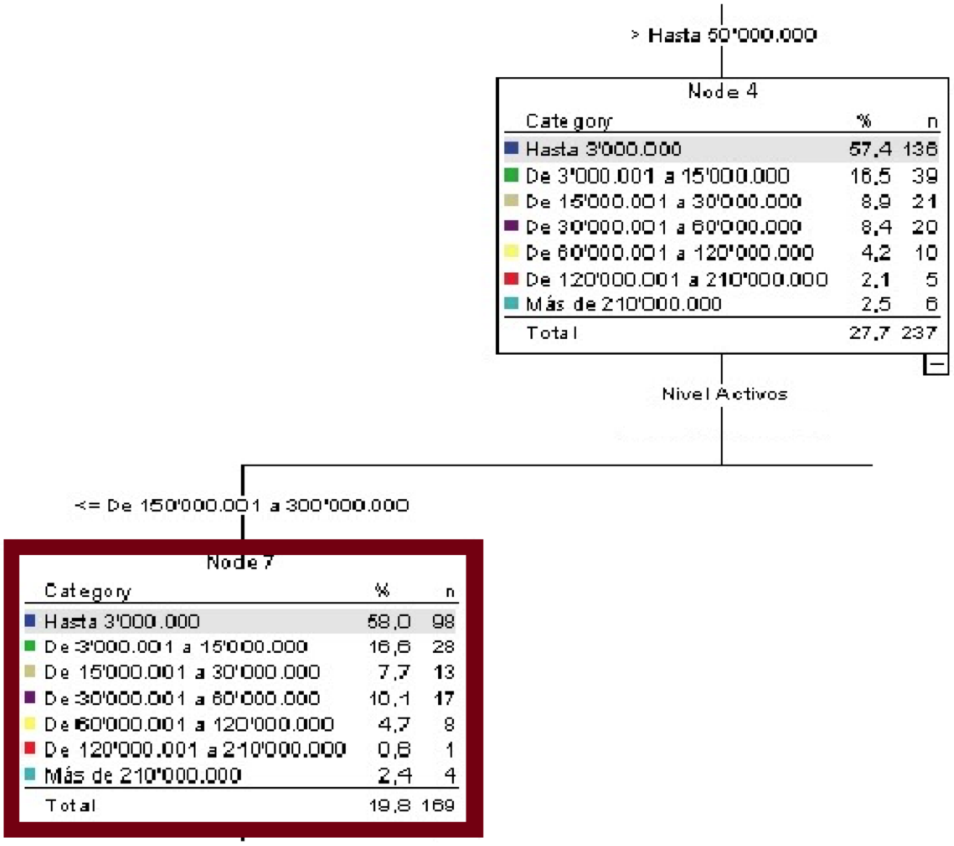

Figure 5.2: caption

300.000.000.

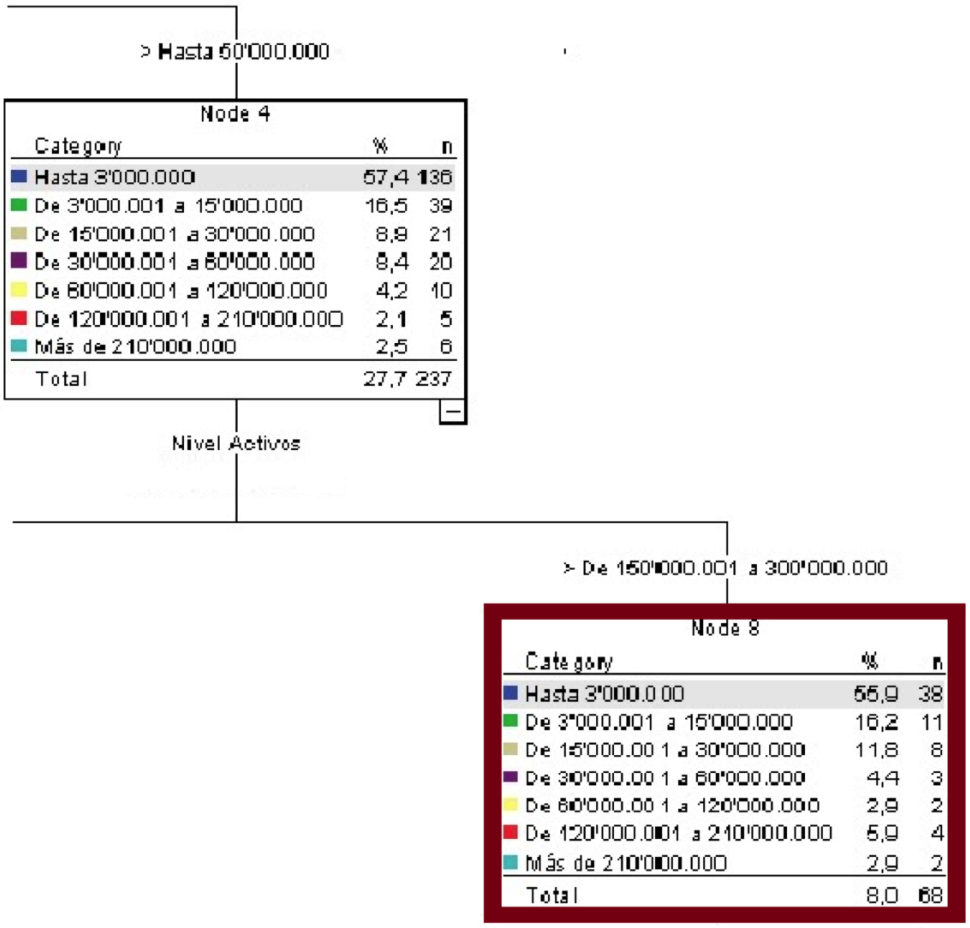

Figure 5.3: caption

The warning signs for this segment are:

* Transactions made by the customers above 10.000.000 per month, or
* customers that they have assets more than 1.000.000.000.

11.7% of the customers in the segment 3, made transactions over 10.000.000 and 14.7% of the customers in segment 3, have assets more than 1.000.000.000.

* **Segment 4:** Income more than 5.000.000 and assets, lower or equal than 150.000.000. The warning sign for this segment is the transactions made

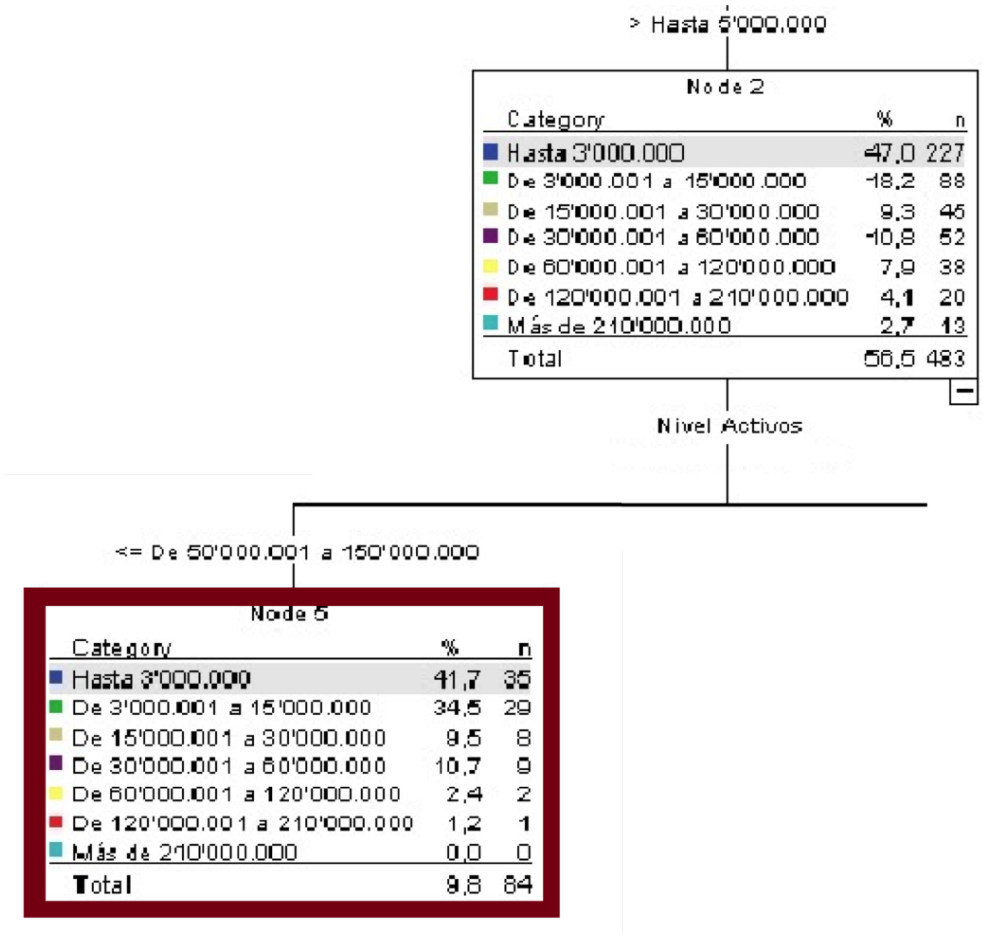

Figure 5.4: caption

by the customers above 10.000.000. 3.6% of the customers in segment 4, made transactions over 10.000.000 per month.

* **Segment 5:** Income more than 5.000.000 and assets between 150.000.000 and 300.000.000. The warning sign for this segment is the transactions

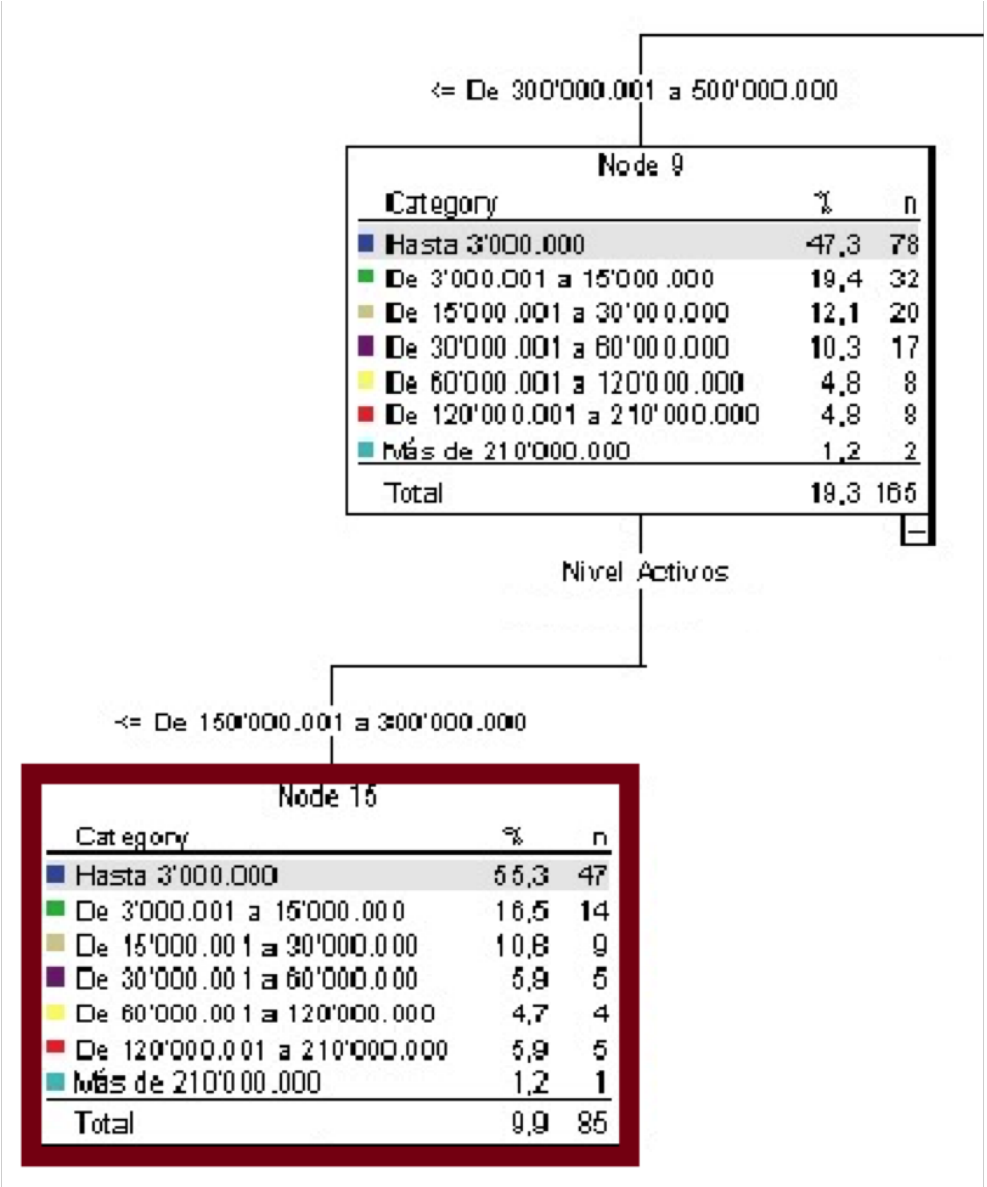

Figure 5.5: caption

made by the customers above 20.000.000 per month. 7.1% of the customers in the segment 5, made transactions over 20.000.000.

* **Segment 6:** Income more than 5.000.000 and assets between 300.000.000 and 500.000.000. The warning sign for this segment is the transactions made by the customers above 20.000.000 per month. 4.8% of the customers in the segment 6, made transactions over 20.000.000. Another warning sign for this segment is the customers, which economic activity is housewives, 2,5% of the segment 6, and business associate, 1,2% of the segment 6.

* **Segment 7:** Income between 5.000.000 and 10.000.000 and assets more than 500.000.000 The warning sign for this segment is the transactions made by the customers above 35.000.000 per month. 6.35% of the customers in the segment 7, made transactions over 35.000.000. Another warning sign for this segment is the customers, which economic activity is a business associate, 4.76% of the segment 7.

* **Segmento 8:** Income more than 10.000.000 and assets more than 500.000.000. The warning sign for this segment is the transactions made by the customers above 35.000.000 per month. 4.1% of the customers in the segment 8, made transactions over 35.000.000. Another warning sign for this segment is the customers, which economic activity is a housewife, 6.43% of the segment 8.

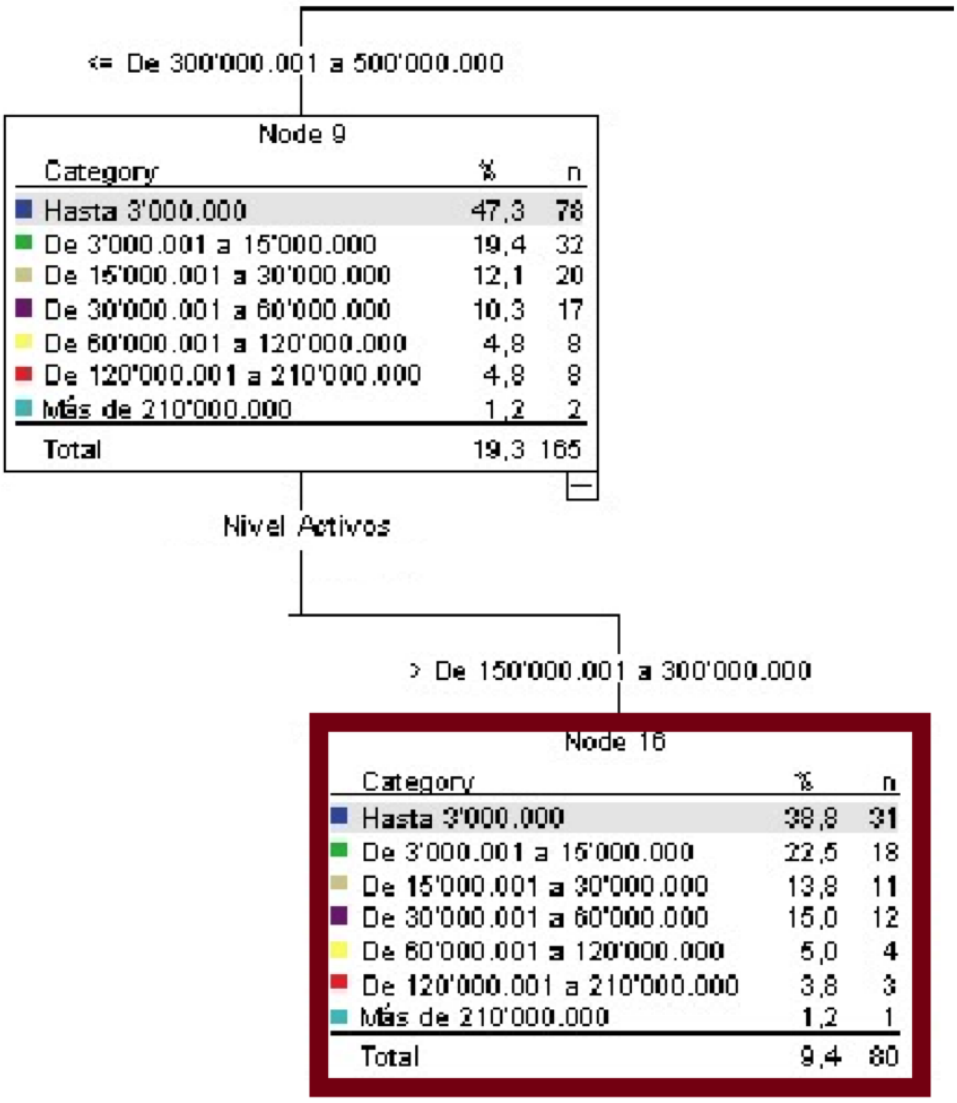

Figure 5.6: caption

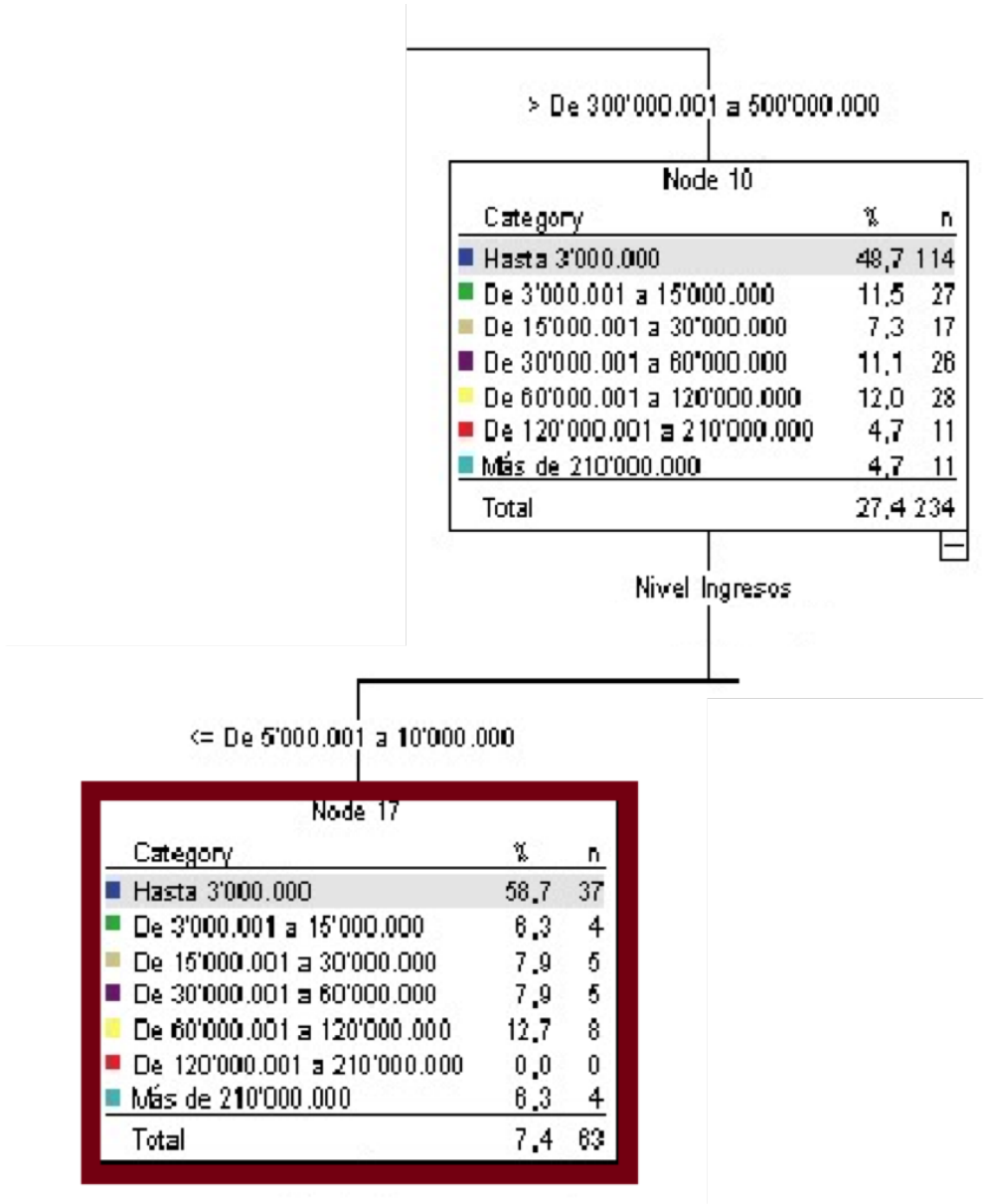

Figure 5.7: caption

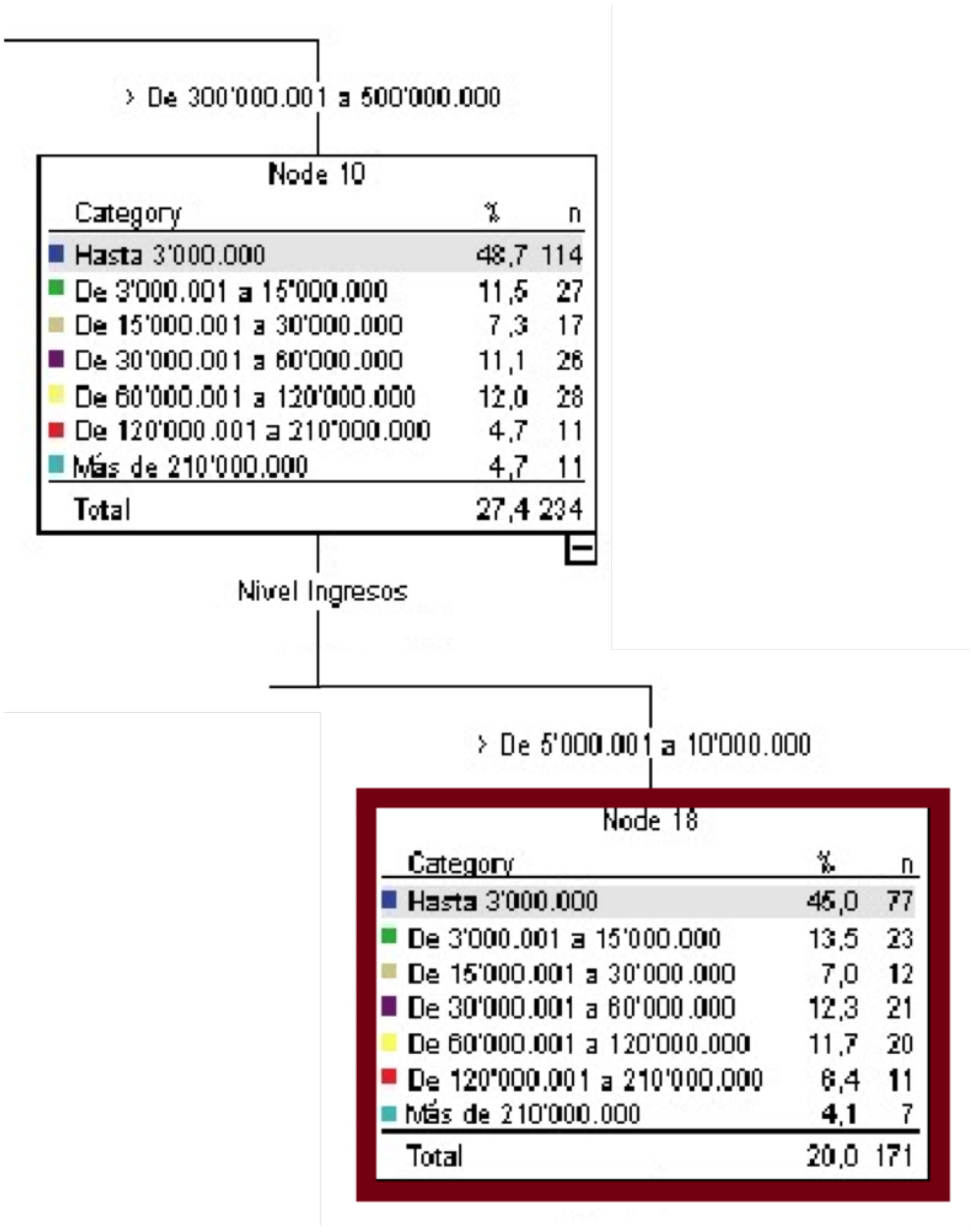

Figure 5.8: caption

5.2 K-Means Algorithm

Since the implementation of a K-means, we get the following segments and the respective warning signs for each segment.

* **Segment 1:** Assets more than 300.000.000 and income more than 20.000.000. The warning sign for this segment is the transactions made by the customers above 34.000.000 per month. 5.12% of the customers in this segment made transactions over 34.000.000.

* **Segment 2:** Assets between 150.000.000 and 500.000.000, and income between 10.000.000 and 50.000.000. The warning sign for this segment is the transactions made by the customers above 10.000.000 per month. 11.62% of the customers in this segment made transactions over 10.000.000.

* **Segment 3:** Assets lower or equal than 150.000.000 and income lower or equal than 10.000.000. The warning sign for this segment is the transactions made by the customers above 18.000.000 per month. 1.31% of the customers in this segment, made transactions over 18.000.000.

Chapter 6

Conclusions