DIGIT.D1 – Big Data

Community of Madrid – Social Insertion

D04.01.Data Modeling

everis Spain S.L.U

Table of contents

Table of figures

1 INTRODUCTION

1.1 Context of the project

According to statistic data, more than 120 million people in the EU are at risk of poverty or social exclusion. The fight against poverty and social exclusion is at the top of the Europe 2020 strategy for intelligent, sustainable and inclusive growth of the society.

A strategic objective of the European administration (national, regional and local) is to help the social insertion of these groups using several instruments and mechanisms.

During the last years the steps that have been carried out to reduce this number have been failed because the number has continued to increase and the Community of Madrid has not been an exception.

Public administrations must understand better the different groups of single units and family units with homogeneous themes that lead to exclusion and, therefore, need particular policies and follow-up of inclusion.

Measuring the results of these policies applied to each group and analysing their chronological evolution is a critical issue to understand the factors that affect the possible social insertion of an individual or his / her family unit.

1.2 Objetive

The main purpose of the project that will be carried out in the Community of Madrid are:

- Identify the different groups of single units and family units existing in the Community of Madrid with homogeneous themes that lead to social exclusion.
- Detect chronological evolution and discover its causes.
- Establish an insertion index to measure the potential inclusion of an single unit or family unit identifying, if possible, strengths and weaknesses in applied social policies.

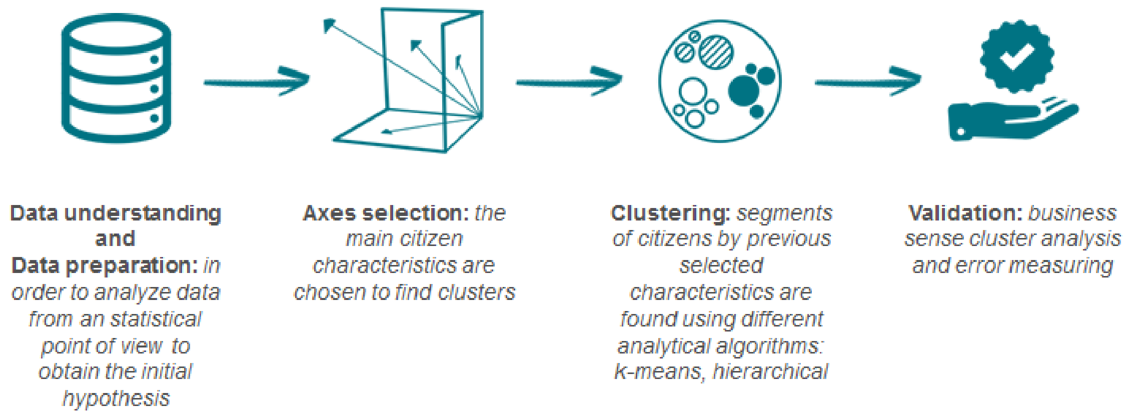
To meet these purposes, two lines of work will be carried out:

- On the one hand, an analytical process will be carried out, which will process and analyse the raw information received using analytical algorithms and information segmentation.
- On the other hand, a tool will be developed that visually presents the previously described information in a way that facilitates its understanding, reflecting the segmentation obtained from the different groups at risk of social exclusion in the Community of Madrid, its chronological evolution and identifying causes and factors that help make social insertion possible for each group.

The objective of this document is to reflect the analyses and processes which have taken place once the data understanding phase under the CRISP-DM methodology, which explores, analyzes and validates the quality of the information for the development of the project, has been carried out. These phases are: **Axes selection**, **Clustering** and **Validation**

2 DATA ANALYSIS

Once anonymous data are received, the phase of data understanding will start using the CRISP-DM methodology which explores, analyses and validates the quality of the information for the development of the project



Data understanding and Data preparation: *in order to analyze data from an statistical point of view to obtain the initial hypothesis*

Axes selection: *the main citizen characteristics are chosen to find clusters*

Clustering: *segments of citizens by previous selected characteristics are found using different analytical algorithms: k-means, hierarchical*

Validation: *business sense cluster analysis and error measuring*

Figure 1- Data analysis methodology

2.1 Data understanding and data preparation

The process of understanding data is divided into 5 sequential phases:

- Business understanding (Validation of business data, validation and table description, correct number of records).
- Integrity analysis (Table integrity, referential integrity).
- Descriptive analysis (Descriptive statistics, distribution analysis, temporal series).
- Specific analysis (Ad-hoc analysis focused on target variable and business, Bi-variants analysis and correlations).
- Data refinement (Detected bugs correction, request corrective tables).

The details of this analysis can be found in the deliverable D02.01 Data Source Analysis.

3 AXES SELECTION

To select the axes that will mark the segmentation, a main analysis of variables of interest for the project will be carried out. It will be considered:

- Files:
  - Years from issue.
  - File ratio over town/district.

- Petitioner:
  - Age.
  - Sex.
  - Nationality.
  - Country of birth.
  - Employment situation.
  - Education level.
  - Housing type.

- Family unit:
  - Number of members.
  - Members age.
  - Members employment situation.
  - Members education level.
  - Income or patrimony of the family unit.

- Individual Insertion Plans (IIP):
  - Areas of action.
  - Plan valuation.
  - Social worker who supervises.
  - Insertion causes.

Once all these variables have been studied, those that will be most useful as possible axes for segmentation will be selected.

4 CLUSTERING

Secondly, it is necessary to apply analytical techniques for the segmentation of data. In this case the following techniques have been used

4.1 Cluster analysis (K-means)

K-means clustering is a method of vector quantification, that is popular for cluster analysis in data mining.

The k-means algorithm is a statistical model technique which consists in the creation of homogeneous groups (segments) with similar features, being the resulting segments as much heterogeneous as possible among them.

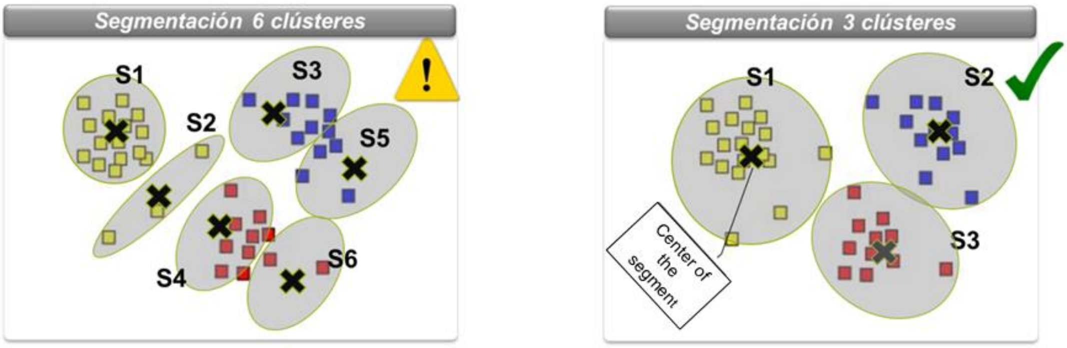


Figure 2 - K means algorithm

4.2 Hierarchical Algorithms

This methodology of cluster analysis seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

The figure bellow shows an example of agglomerative and divisive clustering

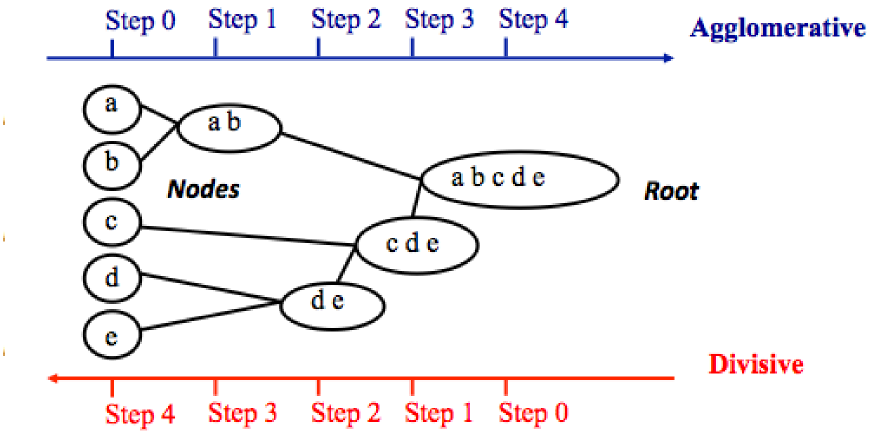


Figure 3 - Hierarchical Algorithms

5 VALIDATION

Once the business is understood, the information and data analysis actions are carried out and the data is validated in order to detect possible errors and validate the data clustering process.