

**Data Mining Project**

**MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**CLUSTERING ANALYSYS ON INSURANCE**

GROUP BB

Miguel Lince, number: M20211300

<Second student name>, number: <student number>

<Third student name>, number: <student number>

January, 2021

INDEX

[1. Introduction iii](#_Toc91586272)

[2. Data Exploration iv](#_Toc91586273)

[3. Data preprocessing v](#_Toc91586274)

[3.1. Missing imputation v](#_Toc91586275)

[3.2. Creating new features v](#_Toc91586276)

[3.3. Correlation v](#_Toc91586277)

[4. Clustering vi](#_Toc91586278)

[4.1. Density-Based clustering vi](#_Toc91586279)

[4.2. Product segmentation vi](#_Toc91586280)

[4.3. Value segmentation vi](#_Toc91586281)

[4.4. Demographic segmentation vi](#_Toc91586282)

[4.5. Cluster merging vi](#_Toc91586283)

[5. References vii](#_Toc91586284)

[6. Appendix viii](#_Toc91586285)

# Introduction

This report aims to document the development of a clustering analysis for a Portuguese insurance company. The analysis was based on data regarding the company’s portfolio with the aim of client segmentation and the proposition of targeted actions based on the characteristics of the segments found that will help the company improving profitability, costumer experience and potentially attract new customers.

High and fearless competition in a digital era where information is so easily accessible makes it paramount for the companies to find new ways of keeping themselves relevant in the market. Understanding their client base and being able of putting the customer in the center of their operation reveals to be essential when, now more than ever, each client wants their needs to be unique and are less prone to compromising.

In the following chapters we will start by explaining our tough process during data exploration and validation and then move to outlier filtering, missing values imputation, the computation of new features and standardization of the metric features we will use for the segmentation.

The segmentation is based on three different segments: Product, Value and Demographic. We relayed on K-Means and K-Medoids/K-Prototypes to define the clusters on each of the segmentations in conjunction with the Elbow method to define the optimal number of clusters. Before defining the clusters using the distance based algorithms we ran a density-based algorithm called DCSCAN to identified new outliers that emerge by the fact that we are now in a manyfold space.

After combining the 3 segments we proceed with the analysis of the dimension of each combination and the assignation of the clients filtered to one of the clusters using a Decision Tree.

In the Product and Demographic segments, we also applied a dimensionality reduction algorithm from Scikit-learn called MDS to be able to visualize the clusters in a 2-dimensional space.

In the last chapter we propose some marketing and underwriting actions to be taken by the company in order to target these segments.

# Data Exploration

# Data preprocessing

## Missing imputation

## Creating new features

## Correlation

# Clustering

## Density-Based clustering

## Product segmentation

## Value segmentation

## Demographic segmentation

## Cluster merging

# References

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix