

Métodos Descritivos de Data Mining

Fall Semester 2020/2021

**Group XXX**

M20170723– Afonso Diniz

MXXX– Bruno Santos

MXXX – Nickolas Lago

**Table of Contents**

[1. Business understanding 4](#_Toc58079811)

[2. Data understanding 5](#_Toc58079812)

[2.1 Shape 5](#_Toc58079813)

[2.2 Info 5](#_Toc58079814)

[2.3 Head 5](#_Toc58079815)

[3. Data preparation 5](#_Toc58079816)

[3.1 Columns dropping and creation 5](#_Toc58079817)

[3.2 Outliers 5](#_Toc58079818)

[3.2.1 Interval Variables 5](#_Toc58079819)

[3.2.2 Categorical Variables 5](#_Toc58079820)

[3.3 Data Exploration 5](#_Toc58079821)

[3.4 NaN Values 5](#_Toc58079822)

[4. Modeling 6](#_Toc58079823)

[4.1 Variable Binning and dummying 6](#_Toc58079824)

[4.2 Data Scaling 6](#_Toc58079825)

[4.3 Building the Model 6](#_Toc58079826)

[4.4 Plotting Model results 6](#_Toc58079827)

[5. Evaluation 6](#_Toc58079828)

[5.1 Optimal number of clusters 6](#_Toc58079829)

[5.2 Top features 6](#_Toc58079830)

[5.3 Cluster results visualization 6](#_Toc58079831)

[5.4 Recommendations for future Marketing Campaigns 6](#_Toc58079832)

[6. Conclusions 6](#_Toc58079833)

**Table of Figures**

[Figure 1 – Crisp DM 4](#_Toc58075248)

# Business understanding

This project is related to a fictional insurance company in Portugal and has as objective the creation of a Customer Segmentation that will enable the Marketing Department to better understand the customer profiles. Better understand of the customer profiles can lead to specific/direct/targeted Marketing campaigns which can lead to more revenue and company growth.

In order to accomplish this goal, we developed all the logic of the segmentation in Python and we ran it on Jupyter Notebook.

We followed CRISP-DM’s methodology, which stands for Cross-industry standard process for Data Mining. It starts on the Business/data understanding, passing through the data preparation stage, which is always where data related projects take more time, and finally the cluster modelling, using the K-means algorithm. After the modeling we assessed the cluster results, and now the Marketing Department can target specific customer that they think will use their product, according to the customer’s characteristics.

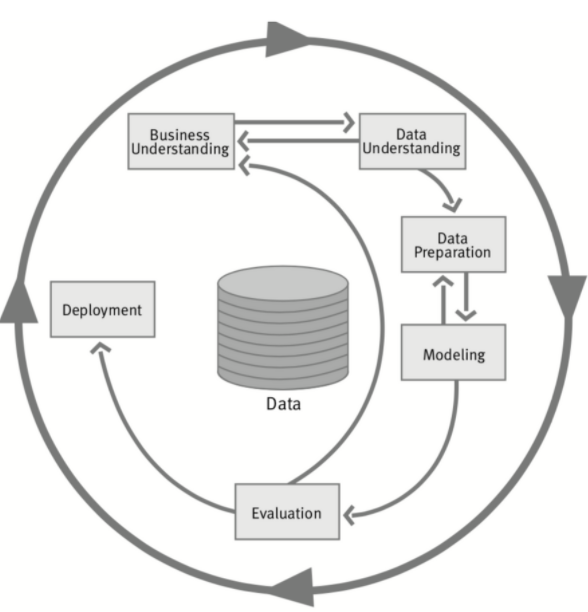
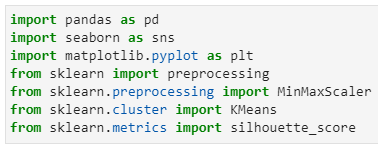


Figure 1 – CRISP DM

# Data understanding

To help us through the realization of this segmentation project we were assisted by several python libraries, being those:

* pandas – library used for data manipulation and analysis
* seaborn & matplotLib – library used for plotting and visualizations
* MinMaxScaler – library used for scaling the data before applying our clustering model
* KMeans – library that represents our model to be applied in our clustering
* silhouette\_score – library that was used to assess the optimal number of ‘k’ of our clustering algorithm



## Import dataset

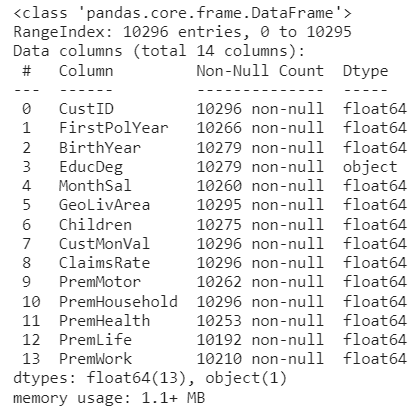
The first thing we had to do was import the given dataset that represented the customers of the Portuguese fictional insurance company.



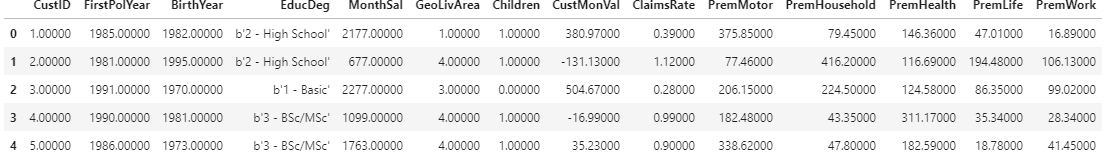
## Shape()



## Info()



## Head()



# Data preparation

## Columns dropping and creation

## Outliers

### Interval Variables

### Categorical Variables

## Data Exploration

## NaN Values

# Modeling

## Variable Binning and dummying

## Data Scaling

## Building the Model

## Plotting Model results

# Evaluation

## Optimal number of clusters

## Top features

## Cluster results visualization

## Recommendations for future Marketing Campaigns

## 

# Conclusions