1.Business Understanding

**1.1 Determine Business Goals:**

Background:WWW is a company focused on selling Wines and related products from all over the world through several channels (telephone, online or physical shops).

Business objectives:WWW pretends to know the different types of customer based on several variables in order to design more efficient marketing tools . Those marketing tools include everything from pricing to marketing campaigns, which creates more loyal (decrease of the churn rate) and valuable customers.

Business Success Criteria:1.(Subjective) Provide useful information about each type of customer; 2.(objective) Decrease churn rate by being able to identify and target potential leavers.

**1.2 Assess the situation**

Inventory of resources: Data (excel file) with 10 000 observations and 29 variables with a time span of 4 years.

Requirement, Assumption and constraints:

Constraints:The dataset includes only customers that have purchased something during the last 18 months.

Risks and Contigency:

|  |  |
| --- | --- |
| **Risks** | **Contigency** |
| Missing Data | Fill missing with median,mode,mean or KNN Imputer. |
| Outliers | Use IQR, Z-score to detect outliers and delete them . |
| Redundant Features | Use correlation matrix to detect redundant features and not use them in the clustering process. |

Terminology:

* Market segmentation: The process of identfying groups of customers based on their purchase behavior.
* Pricing: The process of determining the value of the product.
* Churn rate: Rate of customers that have stoped buying the product.

Costs and Benefits

Benefits:

1. No cost of Mass-Marketing

2. Decrease of the churn Rate

3. Increase of the revenue

Costs

1. Cost of a Differentiated-Marketing

**1.3 Determining Data Mining goals**

Data Mining goals: Clusters that represent the different market segments

**1.4 Produce Project Plan**

Project Plan

1.Explore the data: The team starts off with the analysis of the different types of data in each variable the statistical information of each variable and the relation between.

2.Acquire insights from the data visualization.

3.Coherence test: Check if there was any information that did not respect the rules, for example, check if all the clients bought something from the shop during the last 18 months (active clients).

4.Data cleanup: The variables “Custid” and “Rand” are deleted as they are useless for the analysis.

5.Outliers: Since the clustering analysis is based on the distance between points, outliers can distort the analysis, as such they had to be eliminated.

6.Normalization: Besides outliers, the difference scales of the variables also distort the distance-based clustering process, therefore the data has to be put in the same scale.

7.Results: In this step, it is shown several vizualizations(pairwise,boxplot,correlation matrix) of the processed data.

8. Separation of the feature into several perpectives.

Initial Assessment of tools and Techniques

1. Explore the data: info(),describe(),Pairwise,Boxplot and Correlation matrix

5.Outliers: IQR, Z-score method and the combination of both.

6.Normalization: StandardScaler() and then normalize()

7.Results:Pairwise,Boxplot and Correlation matrix.

O que falta fazer:

1.Acabar Project Plan