Multivariate Analysis Final Project
Heart Disease Data Set

Javier Ferrando Monsonís Marcel Porta Vallés Mehmet Fatih Cagil

June 2019

1 Description of the problem and available data

The problem we face is the identification of the presence of heart disease in a patient. For this, we use the public data set *UCI Machine Learning Repository*[3] where 14 features can be found together with the target variable:

- **age**: age in years
- **sex**: 1 = male, 0 = female
- **cp**: chest pain type
  - 0 = typical angina
  - 1 = atypical angina
  - 2 = non-anginal pain
  - 3 = asymptomatic
- **trestbps**: resting blood pressure (in mm Hg on admission to the hospital)
- **chol**: serum cholestoral in mg/dl
- **fbs**: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- **restecg**: resting electrocardiographic results
  - 0 = normal
  - 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
- **thalach**: thallium stress test maximum heart rate achieved
- **exang**: exercise induced angina (1 = yes; 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest
- **slope**: the slope of the peak exercise ST segment
  - 0: upsloping
  - 1: flat
  - 2: downsloping

- **ca**: number of major vessels (0-3) colored by flourosopy
- **thal**: exercise thallium scintigraphic defects
  - 1: normal
  - 2: fixed defect
  - 3: reversable defect
- **target**: 1 = disease or 0 = no disease ¡

Below can be found a subset of individuals of the Heart Disease data set

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

2 Pre-process of data

2.1 Outliers detection

2.1.1 Mahalanobis

By means of the *Maout R* function, Mahalanobis distance is used for outlier detection. In the following plot it is shown the combination of the Classic Mahalanobis distance and the resuls of the iterative algorithm that gives the Robust Mahalanobis distance. Mahalanobis metric computes the differences between the data points with respect to the centroid G taking into account the covariance matrix V^-1.

$$D_M(i, G) = \sqrt{(x_i - G)^T V^{-1}(x_i - G)}$$

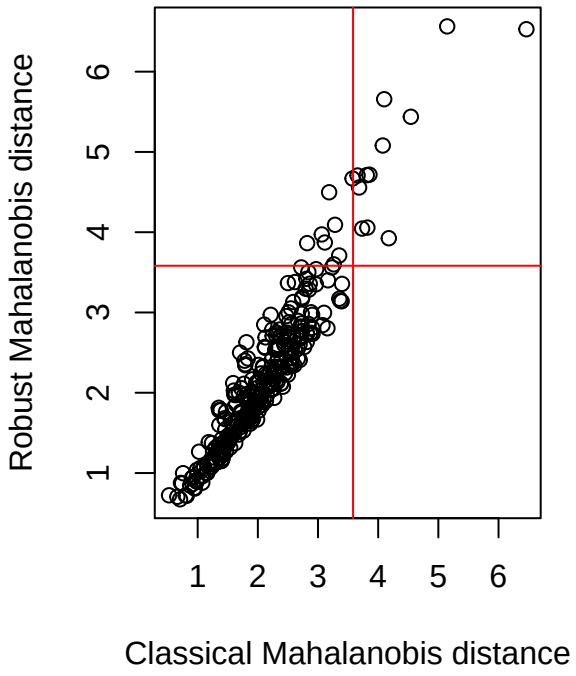

Figure 1: Classic vs Robusts Mahalanobis distances

Data points in Figure 1 placed at the top right quadrant suggest potential outliers. These points are above the quantile cutoff value of 0.975.

2.1.2 Local Outlier Factor

Local Outlier Factor is an algorithm that detects outliers by measuring the local deviation of a given data point with respect to its neighbours [1].

$$LOF_k(x) = \frac{\sum_{nei_x} \frac{maxdist_x(x)}{maxdist_k(nei_x^k)}}{k}$$

Where k denotes the number of individuals taken as neighbors of x, nei_x.

LOF(k) $\approx$ 1 means Similar density as neighbors
LOF(k) < 1 means Higher density than neighbors (Inlier)
LOF(k) > 1 means Lower density than neighbors (Outlier)

From the distribution of Local Outliers Factor scores it can be seen that the great majority of individuals have scores larger than 1, which generally propose outliers. The nature of this dataset could make us increase the cutoff.

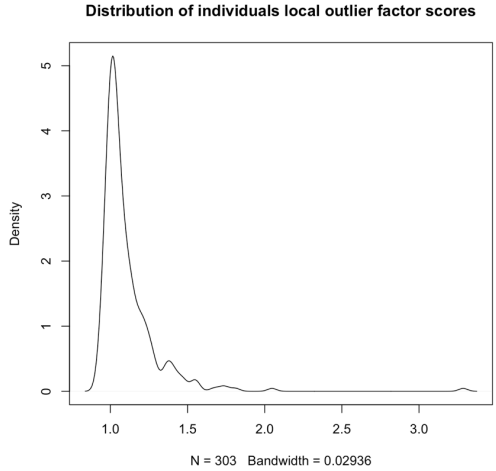

Figure 2: Local Outlier Factor scores

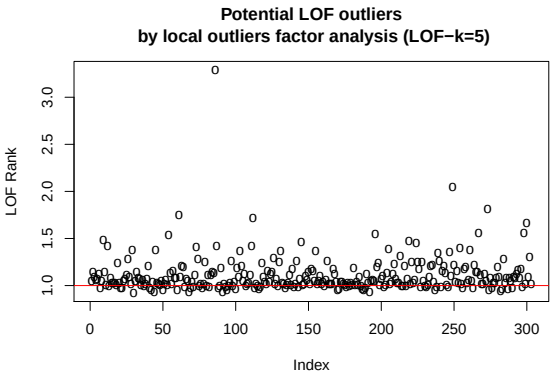

Figure 3: Indicator matrix

3 Visualisation and interpretation of latent factors

3.1 Exploratory Data Analysis

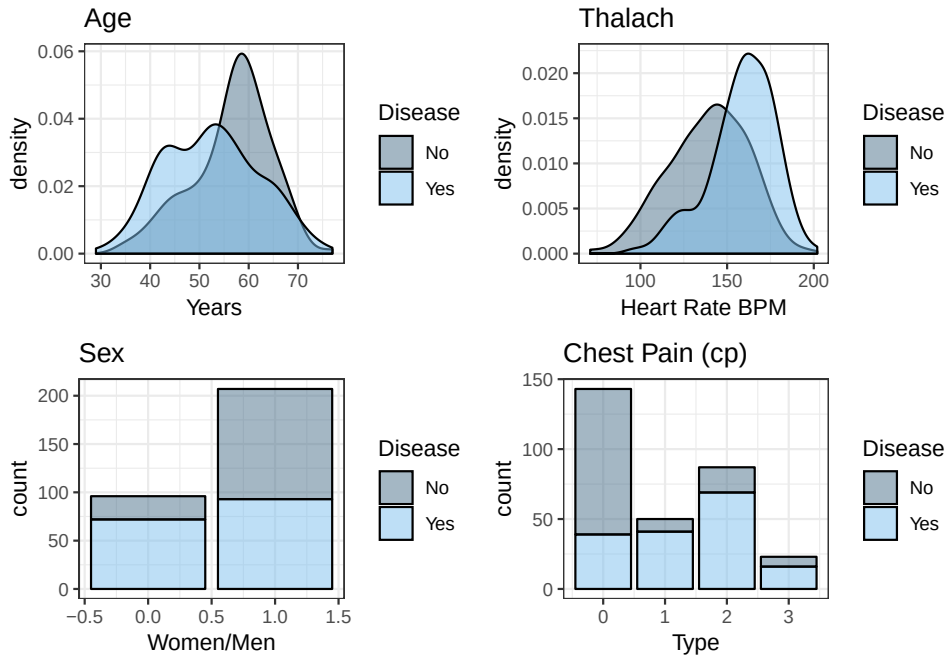

Figure 4: Density plots of explanatory variables vs target variable

From the density plots and histograms we can observe the differences between individuals respect to the heart disease result. The distribution of diseased people age achieves a peak around 50 while the majority of healthy patients are around older ages. Although in the data set sample there are almost double male patients, female proportion of diseased cases is much higher than in males. Chest Pain of type 1 shows much less diseases patients than in categories. Finally, people with diseases tend to achieve higher maximum heart rate in the thallium stress test. This could be partly due to the fact that diseased patients are younger and their maximum predicted heart rate, calculated as 220 (210 for women) minus the patient's age, are higher.

It has to be taken into account that this is a biased sample, not a random population sample. This means that individuals in our data set are people who have already experienced any kind of symptom and have visit a doctor.

3.2 Principal Component Analysis

In order to perform Principal Component Analysis, every categorical variable has to be deleted. Only 5 explanatory variables are left.

We consider $\mathbf{X}$ as the standardized data set matrix of n observations and p features. $\mathbf{N}$ is a diagonal matrix containing weights (importance) for each of the observations in the data, in this case every individual has same weights (unweighted PCA).

$\mathbf{u}_1 \in R^p$ is considered a unitary vector defining a direction in R^p.

Ψ_{1i} represents the projection of the observation i on $\mathbf{u}_1$. When projecting all the individuals on $\mathbf{u}_1$, we get

$$\Psi_1 = \mathbf{X} \cdot \mathbf{u}_1$$

The goal is to obtain orthogonal vectors $\mathbf{u}$ in the directions which maximizes the variance (or inertia I_n) of their Ψ, maximizing the sum of the individual's projections on $\mathbf{u}$. So, in the case of the First Principal Component the objective function we will try to maximize is

$$\max_{\mathbf{u}_1} I_{total} = \max_{\mathbf{u}_1} \sum_{i=1}^{n} w_i \Psi_{1i}^2 = \Psi_1^\intercal \mathbf{N} \Psi_1 = \mathbf{u}_1^\intercal \mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1$$

Subject to $\mathbf{u}_1 \mathbf{u}_1^\intercal = \|\mathbf{u}_1\|_2^2 = 1$

Following an optimization procedure we get

$$\mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

Since we are using a standardized matrix, $\mathbf{X}^\intercal \mathbf{N} \mathbf{X} = Cor(\mathbf{X})$, $\mathbf{u}_1$ represents an eigenvector of $Cor(\mathbf{X})$ and λ_1 its associated eigenvalue. Taking the largest λ_1 will give the eigenvector with maximum variance (First Principal Component).

$\Psi_\alpha \in R^n$ where each component represent the projection of each individual i on the Principal Component u_α.

Since u_1 is a unitary vector we deduce from previous formulas that $\Psi_1^\intercal \mathbf{N} \Psi_1 = \lambda_1 = var(\Psi_1)$

$$I_{total} = \sum_{j=1}^{p} \sum_{i=1}^{n} w_i (x_{ij} - \overline{x}_j)^2 = \sum_{j=1}^{p} var(x_j) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

Projected inertia on the first axis

$$I_1 = \sum_{i=1}^{n} \frac{1}{n} \Psi_{1i}^2 = \lambda_1$$

When plotting the projected inertia on every Principal Component we get

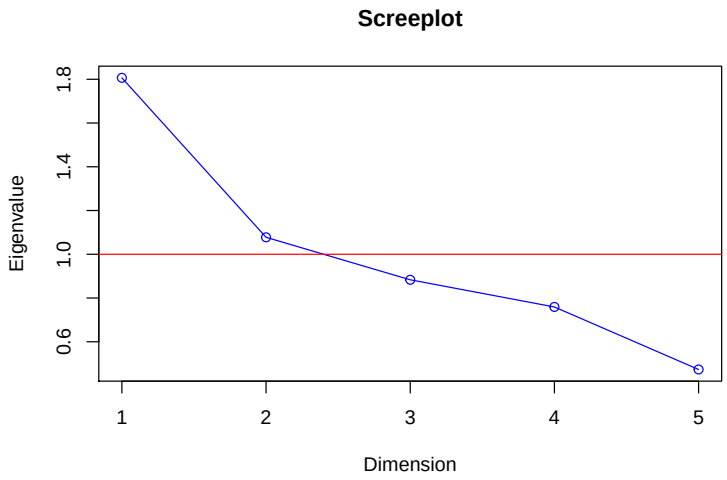

Figure 5: Eigenvalues associated with each Principal Component

To stay with the significant Principal Components, we use the method of subtracting the average value of the eigenvalues and getting the ones staying above. Since we are using a standardized $\mathbf{X}$ data set, the mean of the eigenvalues is equal to 1. Only the first two Principal Components laid above that margin.

The result of projecting our $\mathbf{X}$ matrix onto the first two eigenvectors of $Cor(\mathbf{X})$ gives the plot of Figure 4

$$\Psi_\alpha = \mathbf{X} \cdot \mathbf{u}_\alpha$$

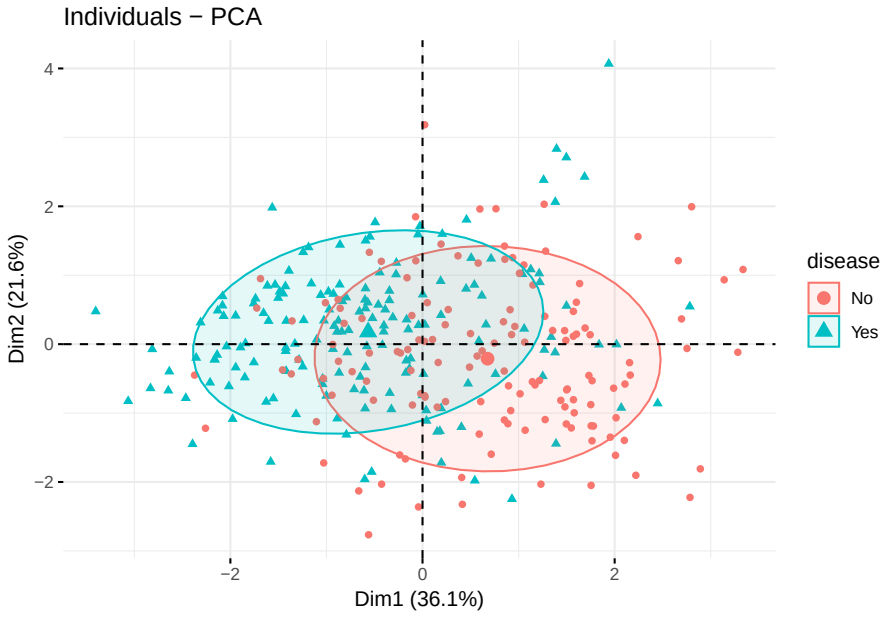

Figure 6: Individuals projection in First Factorial Plane

Figure 6 shows a small difference between disease groups along the First Principal Component. Individuals without heart disease are more likely to appear in the positive coordinates of the First Factor while diseased patients tend to appear in the negative side.

Now, we are going to analyze the association between variables and their correlation between the Principal Components.

$\mathbf{v}_1 \in R^n$ is considered a unitary vector defining a direction in R^n. φ_{1j} denotes the projections of variable j onto $\mathbf{v}_1$, $\mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1$, when using a standardized matrix, $\varphi_1 = cor(\mathbf{X}, \Psi_1)$. The function to maximize is

$$\max_{\mathbf{v}_1} I_{total} = \max_{\mathbf{v}_1} \sum_{j=1}^{p} \varphi_{1j}^2 = \varphi_1^\intercal \varphi_1 = \mathbf{v}_1^\intercal \mathbf{N}^{1/2} \mathbf{X}\mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1$$

Subject to $\mathbf{v}_1 \mathbf{v}_1^\intercal = \|\mathbf{v}_1\|_2^2 = 1$

Following the same optimization procedure that in R^p we get

$$\mathbf{N}^{1/2} \mathbf{X}\mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

The result of projecting $\mathbf{X}^\intercal \mathbf{N}^{1/2}$ matrix onto the first two eigenvectors with highest eigenvalues associated gives the plot of Figure 5

$$\varphi_\alpha = \mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_\alpha$$

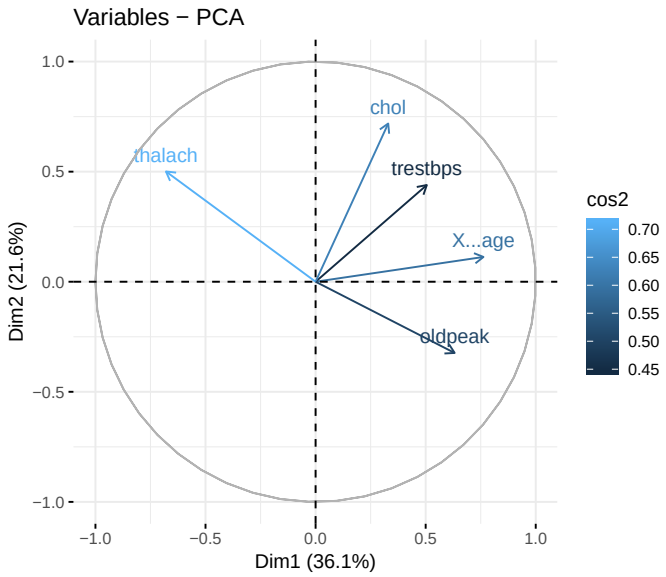

Figure 7: Projection of variables onto First Factorial plane

The arrows are coloured by each variable $cos^2(j, \alpha)$, which is the contribution (importance) of a component to the squared distance of the observation to the origin (G) in the original cloud of points. Can be interpreted as the quality of the representations of each variable.

$$cos^2(j, \alpha) = \frac{\varphi_{\alpha j}^2}{s_j^2}$$

Figure 7 shows correlation between variables and Principal Components. Variable *thalach* appears inverse correlated with *oldpeak* in both components. This means that high maximum heart rates achieved appear with low ST depression [4] produced by exercise and vice versa. Hearts that beat at high rates (potentially unhealthy (Section 4.1)) will present low values of ST Depression (ST Segment length in [Figure 9]), which is considered as a normal behaviour of the heart rate signal detected in the electrocardiogram [2]. This means that ST Depression metric doesn't present too much information about the type of disease presented in this data set (vessel diameter narrowing disease).

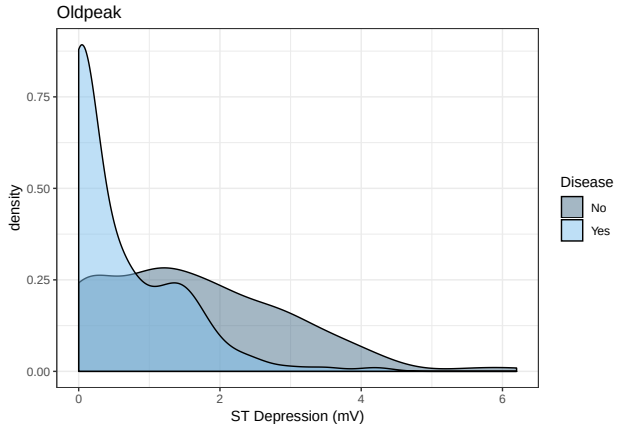

Figure 8: Density plots of explanatory variables vs target variable

Moreover, *chol* (cholesterol), *trestbps* (blood preasure at rest) and *X..age* (age) are closely correlated with the First Factorial Plane. It's natural to think that each of these variables would increase or decrease at the same time.

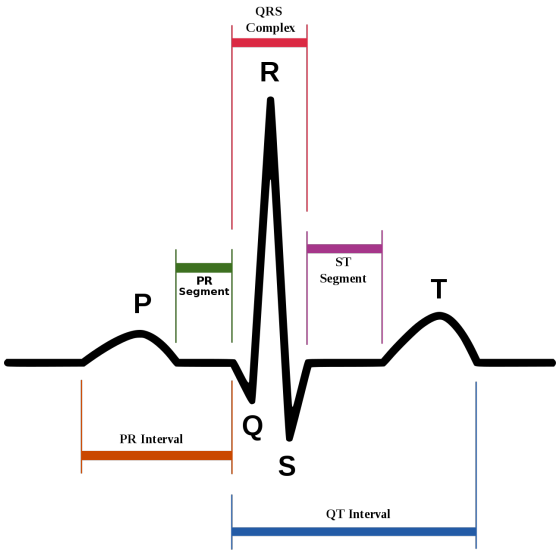

Figure 9: Metrics used in an electrocardiogram

3.3 Multiple Correspondence Analysis

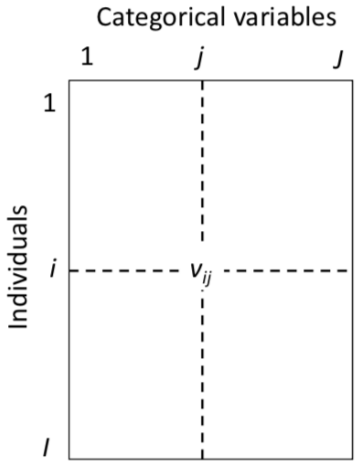

Figure 10: MCA table

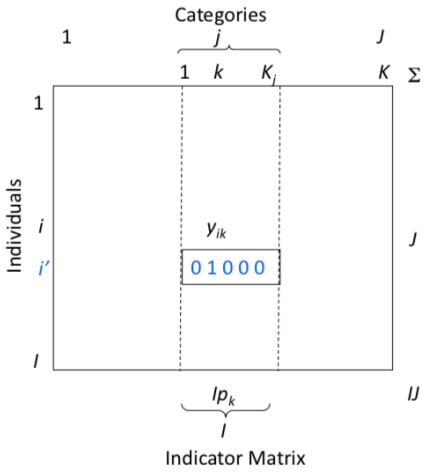

Figure 11: Indicator matrix

Where I refers to the number of individuals, J to the qualitative variables, v_{ij} the category of the j-th variable possessed by the i-th individual.

In order to make use of Multiple Correspondence techniques, a modification in the values of our predictor variables is needed to be done, quantitative variables are transformed into categorical ones by grouping its values in different intervals.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	disease
[60,70)	1	3	[140,160)	[200,300)	1	0	[150,170)	0	[2,3)	0	0	1	No
[30,40)	1	2	[120,140)	[200,300)	0	1	[170,190)	0	[3,4)	0	0	2	No
[40,50)	0	1	[120,140)	[200,300)	0	0	[170,190)	0	[1,2)	2	0	2	No

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	disease
[50,60)	1	1	[120,140)	[200,300)	0	1	[170,190)	0	[0,1)	2	0	2	No
[50,60)	0	0	[120,140)	[300,400)	0	1	[150,170)	1	[0,1)	2	0	2	No
[50,60)	1	0	[140,160)	[100,200)	0	1	[130,150)	0	[0,1)	1	0	1	No

FactoMineR package builds the complete disjunctive in Figure 20 table from Figure 19.

When studying the cloud of individuals, every row (individual) is represented by a vector y_{ik} of zeros and ones indicating whether *i* belongs to each of the categories of each of the variables. y_{ik} equals 1 if the i-th individual is in the k-th category of on category of the J variables, 0 otherwise. p_k refers to the proportion of individuals that share category *k*.
Distance between two individuals, between centroid of the cloud of points and the total Inertia are calculated with the following formulas

$$d^2_{i,i'} = \frac{1}{J}\sum_{k=1}^{K}\frac{1}{p_k}(y_{ik} - y_{i'k})$$

$$d^2_{i,G^2_I} = \frac{1}{J}\sum_{k=1}^{K}\frac{y_{ij}}{p_k} - 1$$

$$Inertia(N_I) = \frac{K}{J}$$

In Figure 12 is represented the projections of the cloud of individuals onto the First Factorial Plane.

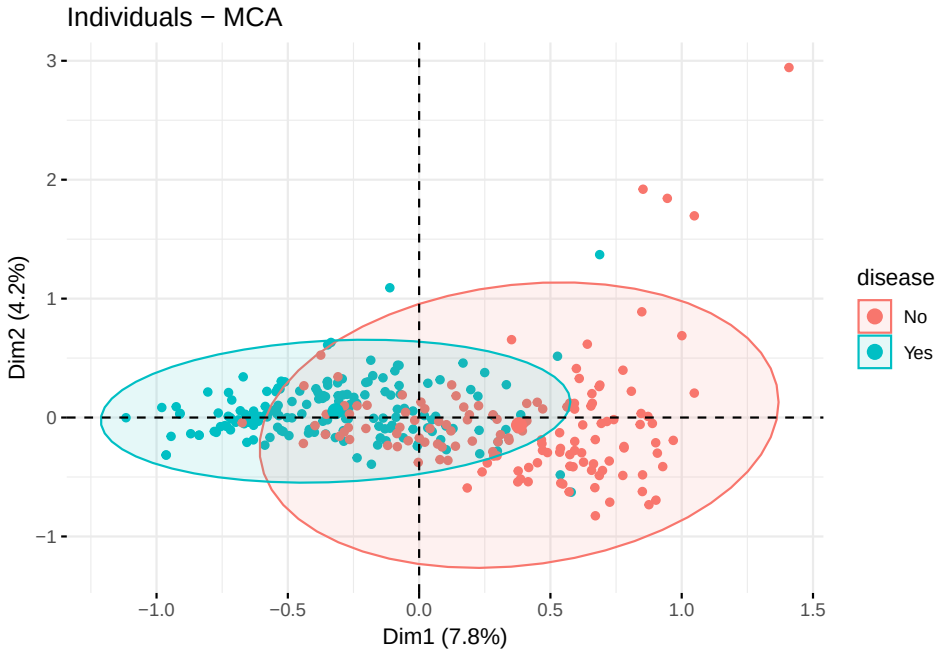

Figure 12: Individuals projection onto First MCA Factorial Plane

The distance between two categories in the cloud of variables is represented as

$$d^2_{k,k'} = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

The Inertial of a category k and of variable j as

$$Inertia(k) = \frac{1 - p_k}{J}$$

$$Inertia(j) = \frac{K_j - 1}{J}$$

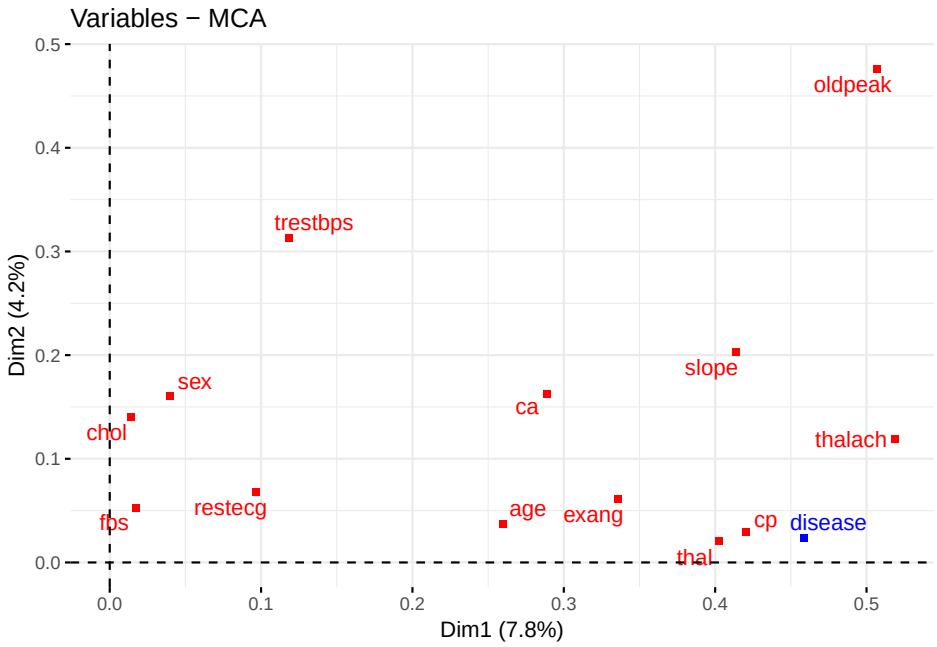

Figure 13: Variables projection onto First MCA Factorial Plane

Figure 13 shows the variables that are better represented in the Factorial Plane far from the origin like *oldpeak*, *slope* and *thalach* .On the contrary, variables such as *chol*, *fbs* or *retecg* are worse represented and explain less of the total inertia.

In the following figure, a representation of the Top 15 categories in terms of the quality (cos^2) of their representation in the plane together with the supplementary variable (disease) is showed. Category *slope=0*, *cp=0*, *exang=1*, *thal=3* come up in healthy individuals. On the other hand, *slope=2*, *thalach* $\in [170, 190)$, *thal = 2*, *ca=0*, *exang=0* and *Oldp* $\in [0, 1)$ are measured by diseased individuals.

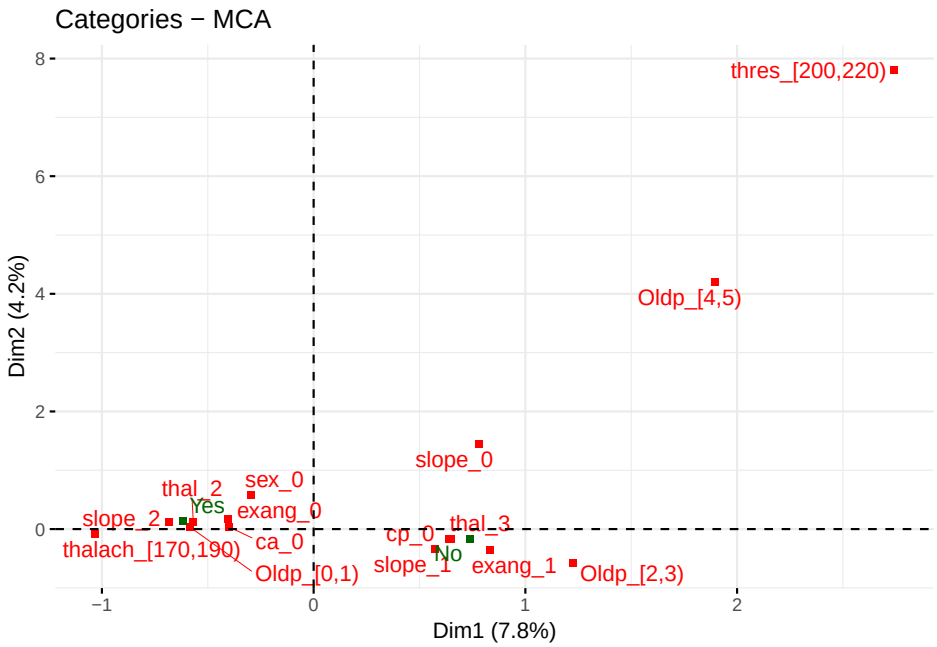

Figure 14: Top 15 categories by quality of representation

The scree plot shows the percentage of explained variance by each of the components of the performed MCA.

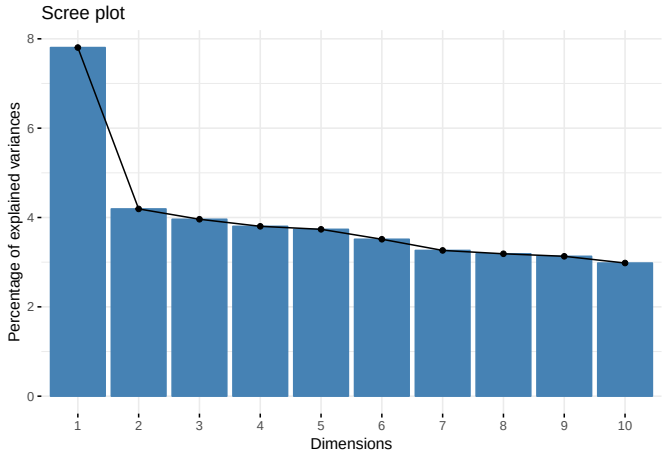

Figure 15: Percentage of the total Inertia explained by each component

4 Clustering performed

4.1 Hierarchical Clustering

When dealing with Hierarchical clustering, we have at our disposal different agglomeration methods that will change our results depending on the one chosen. In this case, we make use of Ward's minimum variance method, that minimizes the total within-cluster variance. At each step of the iterative algorithm, the pair of clusters with minimum between-cluster distance are merged.

Each leaf in the dendogram (Figure 16) represents one individual of our data. Hierarchical Clustering takes as input Euclidean distances between data points in and performs the iterative algorithm by joining leafs into branches by means of the agglomeration specified. The height of the fusion, provided on the vertical axis, indicates the dissimilarity between two observations, The higher the height, the less similar the individuals are. The height of the cut of the dendogram controls the number of clusters obtained.

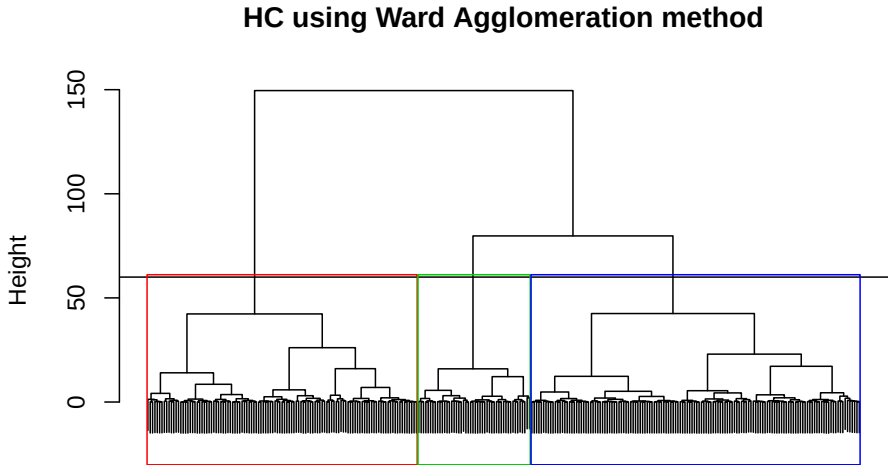

Figure 16: Dendogram showing Hierarchical Clustering results

Once the Hierarchical Clustering has been performed, the tree is cut at a height of 60, which gives us 3 different clusters. In the next figure it is showed individual projections labelled with their associated cluster.

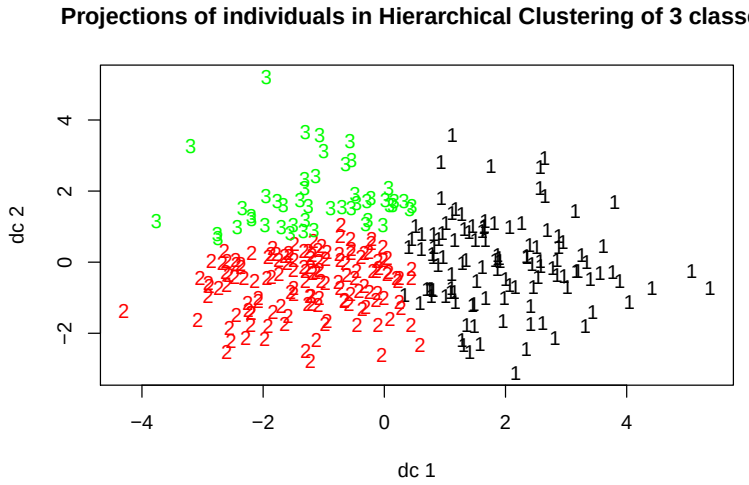

Figure 17: Individuals with its associated cluster after Hierarchical Clustering

4.2 Consolidation clustering

Hierarchical Clustering can be used as input to the K-Means algorithm, instead of the random initialization of standard K-Means. K-Means Clustering iterates until each cluster assignment stops changing. In each iteration, each cluster centroid is calculated and every observation is assigned to the closest cluster (using Euclidean distance). The results of Hierarchical + K-means algorithms are showed below.

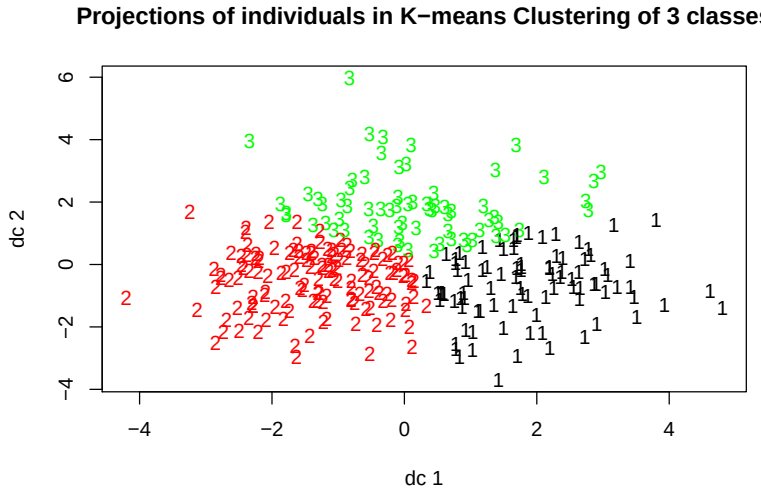

Figure 18: Individuals with its associated cluster after Consolidation Process

In order to decide the optimal number of clusters Calinski-Harabasz index is used as a measure of the quality of the clustering.

$$CH_k = \frac{SSB}{SSW}\frac{N-k}{k-1}$$

Where SSB represents the overall between-cluster variance, SSW the overall within-cluster variance, N the total number of observations and k the number of clusters. In the following figures, CH_k values are plotted. It can be seen a overall increment in the index values after the consolidation clustering. 3 clusters seems to be a fairly good choice based on it index values.

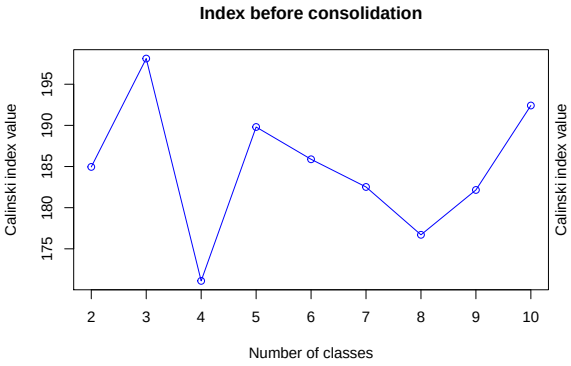

Figure 19: CH_k before consolidation

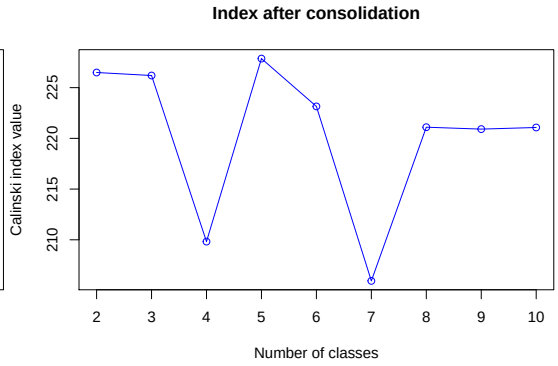

Figure 20: CH_k after consolidation

5 Interpretation of the found clusters

The same procedure as in section 4.1 and 4.2 has been done with the results obtained from the Multiple Correspondence Analysis (not showed to avoid repetition). After the consolidation process with the combination of Hierarchical Cluster and K-Means and choosing the optimum number of clusters following the Calinski-Harabasz index, the following clusters are obtained

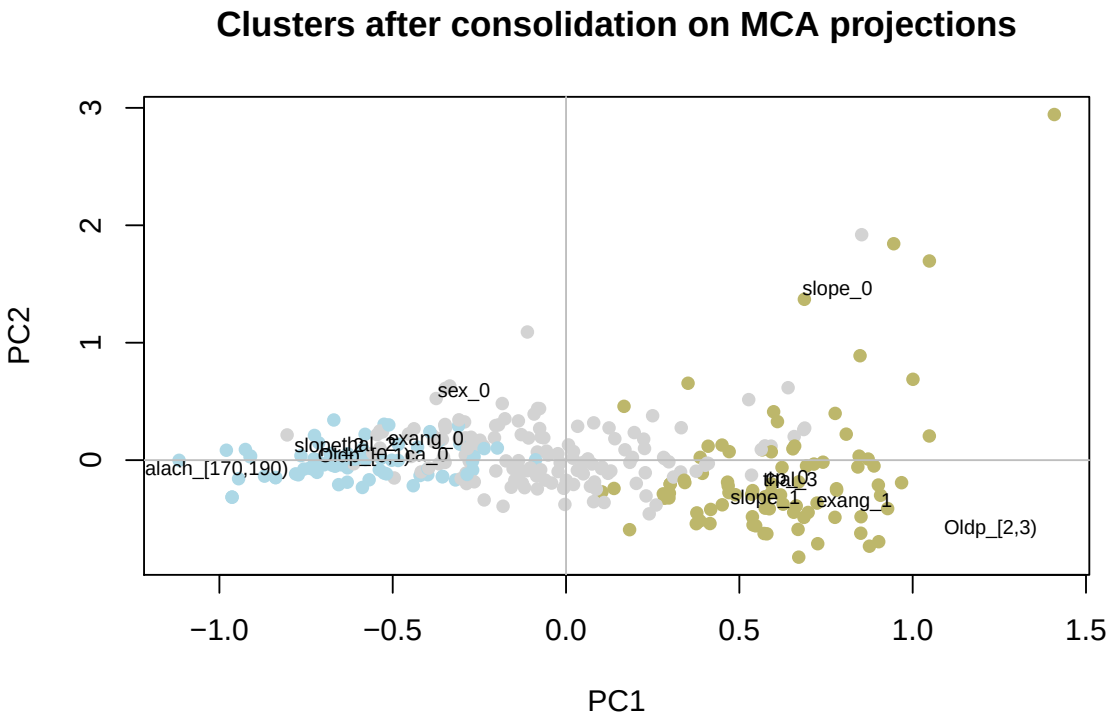

Figure 21: Clusters and Top categories by cos^2

For each of the categorical variables in each cluster a v-test is performed. In this case, the null hypothesis H_0 for categorical variables is that the average of each variable in any given group is the same to the mean of every variable in the whole data set. The results of the test are shown by the p-value. In the following tables, the first five categories in ascendant p-value order of each of 2 cluster is showed:

Cluster 1: women with medium-low maximum heart rate achieved in the stress test, fixed defect and non induced angina.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
thalach=thalach_[150,170)	73.148148	71.1711712	35.643564	0.0000000	9.847403
sex=0	56.250000	48.6486486	31.683168	0.0000020	4.750784
thal=2	45.783133	68.4684685	54.785479	0.0002701	3.642439
exang=0	42.647059	78.3783784	67.326733	0.0016694	3.143502
cp=2	50.574713	39.6396396	28.712871	0.0016771	3.142156

Cluster 2: individuals with medium-high maximum heart rate achieved in the stress test, normal ST Depression values and medium age.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
thalach=thalach_[170,190)	84.482759	55.056180	19.141914	0.0000000	9.841015
oldpeak=Oldp_[0,1)	47.590361	88.764045	54.785479	0.0000000	8.015739
age=Age_[40,50)	68.055556	55.056180	23.762376	0.0000000	7.920538
slope=2	50.000000	79.775281	46.864686	0.0000000	7.504068
exang=0	40.686275	93.258427	67.326733	0.0000000	6.682068

Cluster 3: patients with induced angina, chest pain the first type and a flat slope in the ST segment. This cluster is the most closely associated with healthy individuals.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
exang=1	69.696970	66.9902913	32.673267	0.0000000	9.052541
cp=0	59.440559	82.5242718	47.194719	0.0000000	9.028255
slope=1	57.142857	77.6699029	46.204620	0.0000000	7.974188
oldpeak=Oldp_[2,3)	94.117647	31.0679612	11.221122	0.0000000	7.768497
thal=3	57.264957	65.0485437	38.613861	0.0000000	6.730015

6 Differences between the test sample and the training one

The whole data set is splitted randomly into training set (2/3) and test set (1/3). 108 individuals make up the test set and 195 the training set. Histograms of target and explanatory continuous variables show similiar distributions between both sets.

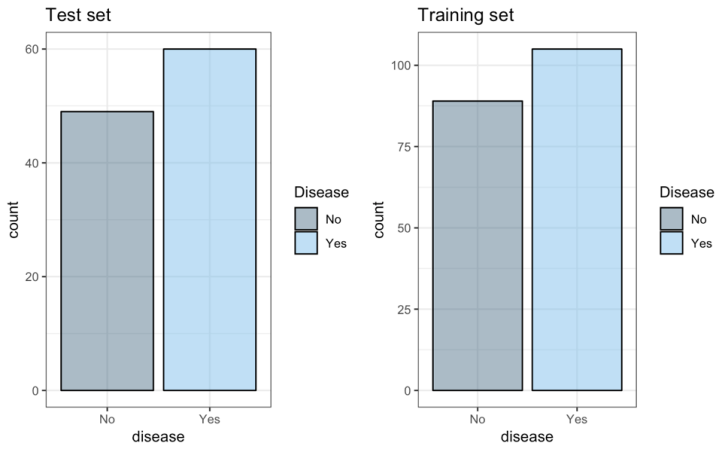

Figure 22: target variable distribution test vs training data sets

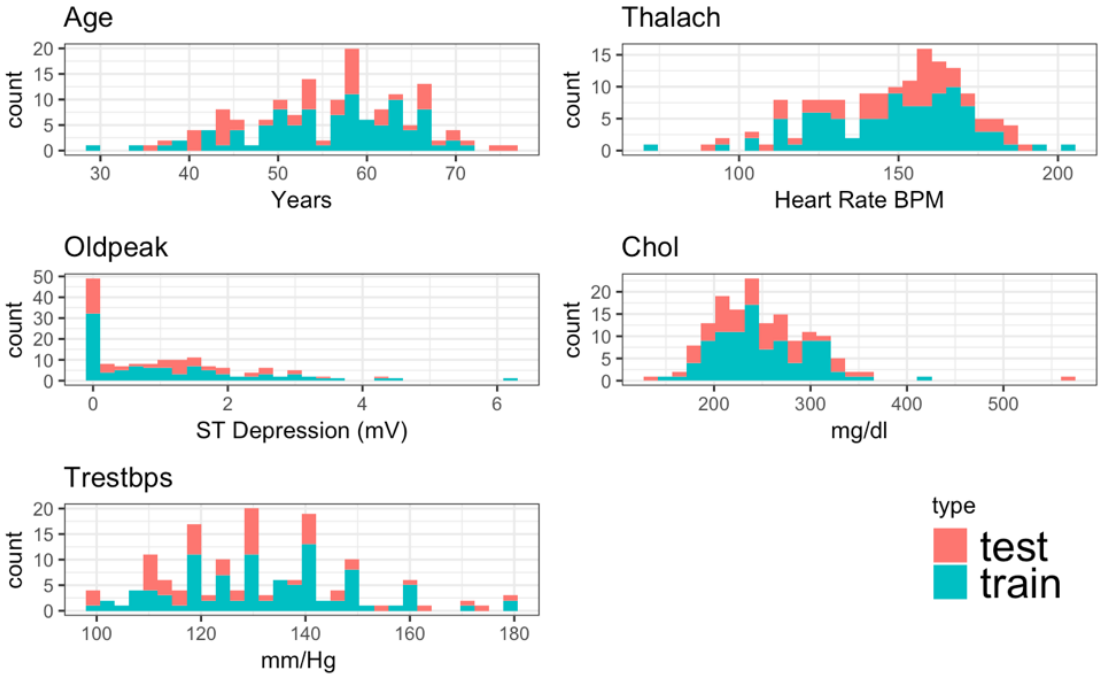

Figure 23: Explanatory variables distributions test vs training data sets

7 Predictive model

7.1 Decision Trees

In order to make predictions model, we make use of the power of decision trees. Cost of the tree R(T) is tried to be minimized by selecting the appropriate variable and value to split the node into two children. The cost to minimize is

$$R(t) = \frac{\sum_{t \in T} p(t) r(t)}{r(root)} \cdot 100$$

Where $r(t)$ is the cost of a node t and $r(root)$ the cost of the root. The cost of a node in a classification tree is calculated based on $max_j p(j|t)$, where $p(j|t) = \frac{n_{tj}}{n_t}$

$$r(t) = 1 - max_j p(j|t)$$

Since the previous objective function would lead to large trees, we use α as a complexity parameter to control its size.

The new objective function becomes

$$Min(R(t) + \alpha |T|)$$

In the next plot it is shown the cost R(T) of the tree using the training data (red) and the test data (blue) as a function of the size of the tree. Test data results has been obtained by means of 10 fold Cross-Validation. The complexity parameter used to penalize the length of the trees has been $\alpha = 0.01$.

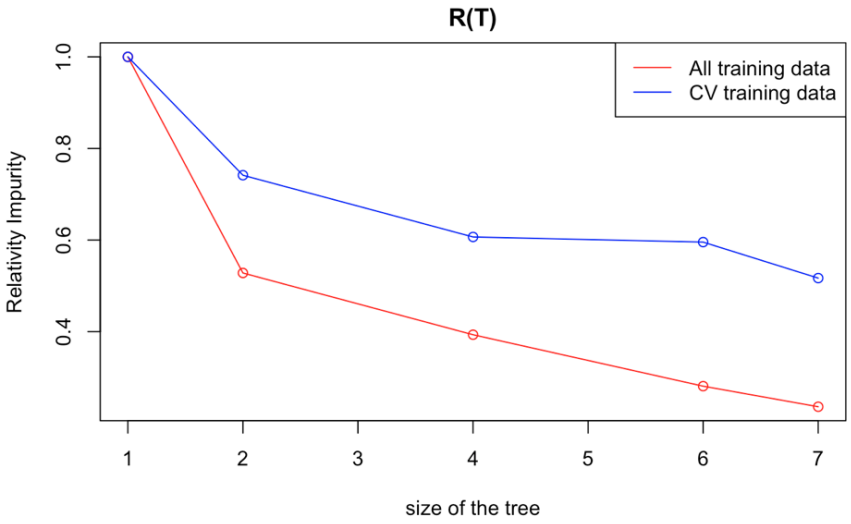

Figure 24: Training vs Cross-Validation error

7 leafs leads to the optimal tree, expanding the splits in each node we obtain

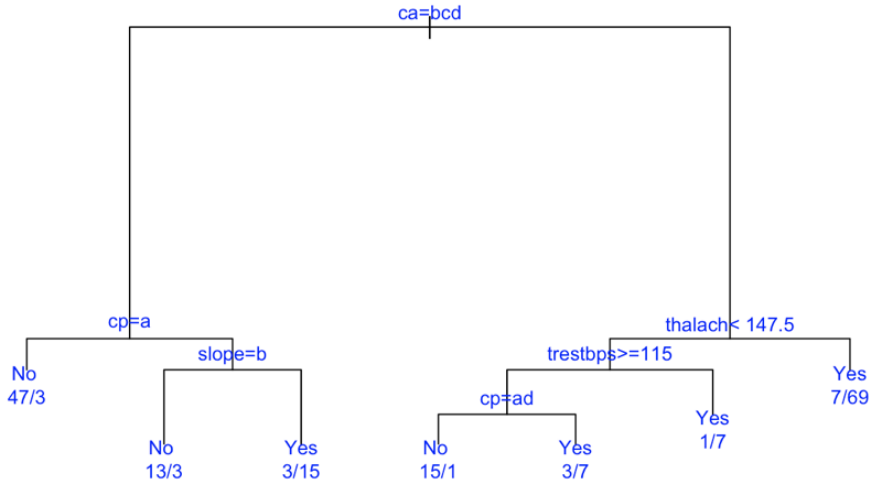

Figure 25: Optimal tree

Absence of major vessels (0) colored by flourosopy *ca = 0* and typical angina *cp = 0* is predicted as no disease by the classification tree. It finds 47 cases in the test set where patients with this characteristics doesn't show a disease and only 3 with heart disease. On the contrary, presence of major vessels, together with high heart beat rates in the stress test are predicted as diseased people (69 out of 76 in these categories in the test set).
The most important variables, in terms of influence in the decision tree making are *ca* and *thalach*.

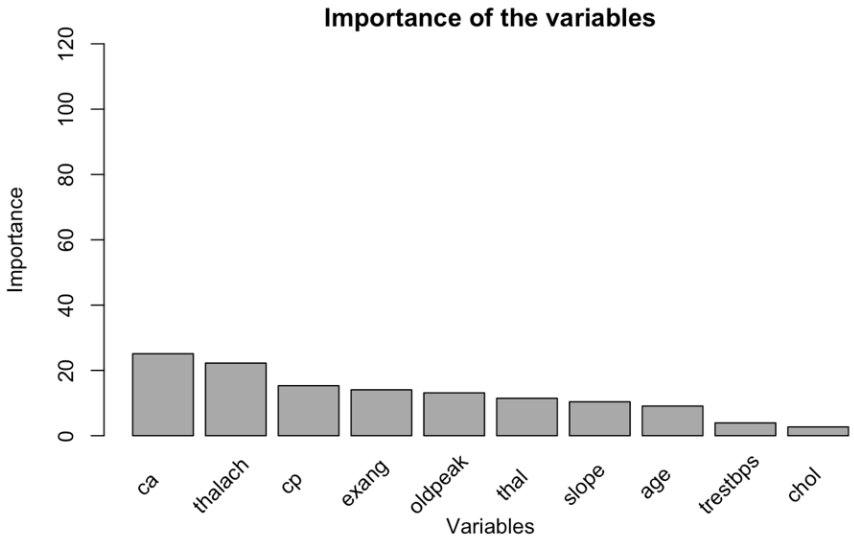

Figure 26: Variables importance

Following we display the confusion matrix output by the decision tree and its corresponding classification metrics.

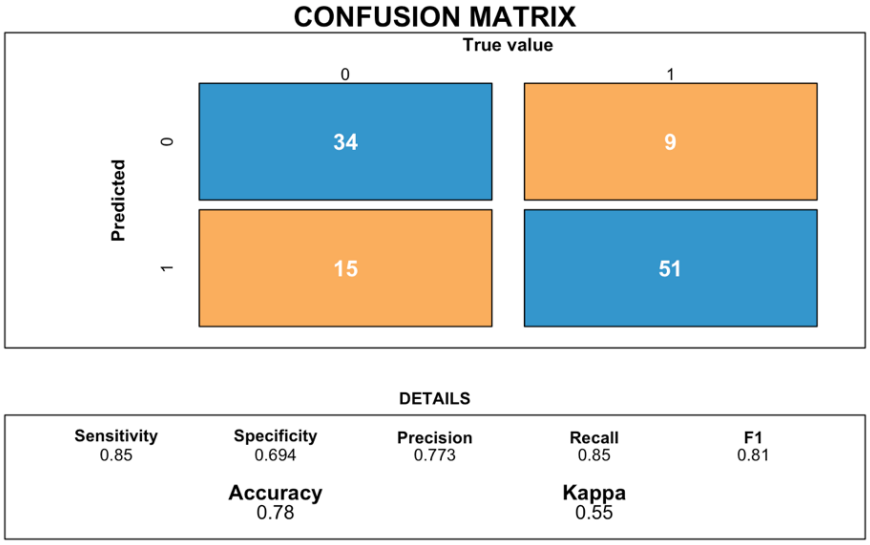

Figure 27: Decision tree confusion matrix

$$Precision = \frac{1}{2}\left[\frac{TP}{TP+FP} + \frac{TN}{TN+FN}\right]$$

$$Accuracy = 1 - \frac{FN + FP}{n}$$

$$Recall = Sensitivity = \frac{TP}{P}$$

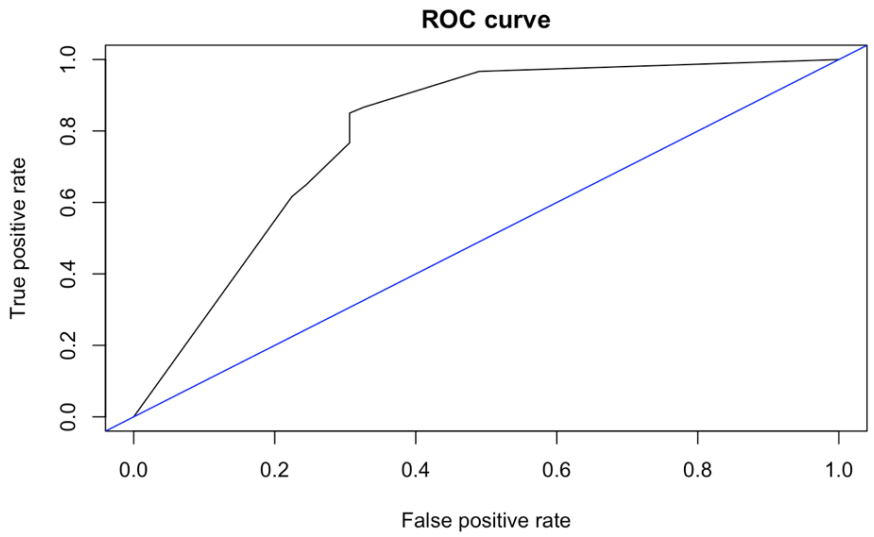

Figure 28: ROC curve

8 Scientific and personal conclusions

In this work, it has been presented an analysis of the Heart Disease data set *UCI Machine Learning Repository*[3] using the techniques learned in Multivariate Analysis course. An exploratory analysis has made us interpret better the available data and come up with first hypotheses that have been tested during PCA and MCA analysis. Clustering procedures such as K-Means and Hierarchical Clustering have been used to find groups between individuals with similar characteristics. Finally, Decision Trees and Random Forest have been utilised as predictive models showing up to 80% of accuracy.

Every approach followed during this work have made us increase our knowledge, not only Multivariate Analysis wise but also in terms of heart diseases subject. Moreover, we have been able to get more comfortable with R programming language and with the writing of scientific work, skills that for sure will be useful in our future.

References

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. 2000.

[2] A. Goldberger, Goldberger L., Z. A., and A. Shvilkin. How to make basic ecg measurements. 2018.

[3] UCI. Uci repository of machine learning databases, 1998.

[4] Wikipedia. St depression wikipedia, 2019.