UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Master in Innovation and Research in Informatics

Multivariate Analysis

Heart Disease Prediction

Final Project

Carol Jhasmin Azparrent, Mohana Fathollahi,
Joan Gaztelu, Paula Iborra

January 17, 2022

Contents

1 Heart Disease Data Set **1**

2 Project summary **1**

3 Data source **1**

3.1 Data description 1

4 Project plan **2**

4.1 Aims 2

4.2 Work plan and assignment of tasks 3

4.3 Potential risks 3

5 Data Cleaning **4**

6 Data Explanatory Analysis **4**

6.1 Visualization of Variables 6

6.2 Detection of Missing Data 8

6.3 Detection of Outliers 10

6.3.1 Multivariate Outlier Detection 10

6.3.2 Univariate Outlier Detection 11

7 Prepare data for analysis **13**

7.1 Validation Protocol 13

7.2 Pre-processing 13

7.2.1 Imputation of missings - MICE 14

7.2.2 Imputation of missings - Miss Forest 15

8 Dimensionality reduction **15**

8.1 Principal Component Analysis (PCA) 15

8.2 Multiple Correspondence Analysis (MCA) 17

9 Clustering **18**

10 Predictive Models **21**

10.1 Trees-Based Models 21

10.1.1 Classification Trees 21

10.1.2 Random Forests 27

10.2 LDA 29
10.2.1 Normality assumption 29
10.2.2 Multivariate Gaussian Conditions 30
10.2.3 Variance and covariance in groups 31
10.3 QDA 34
10.4 Final Results 34

A Additional Tables **iii**

B Additional Figures **iv**

List of Figures

1 Pairwise scaterplot and correlations of continuous variables colored according to absence(0) or presence(1) of CVD. . . . 7
2 Barplot of categorical variables colored according to absence(0) or presence(1) of CVD. . . . 8
3 Overall missing values per variables across all the observations. . . . 9
4 Missing values per variables and origin source. . . . 10
5 Squared Mahalanobis Distances for our data set observations against the empirical Chi-Squared distribution function of these values . . . 11
6 Boxplot of numeric variables . . . 12
7 Comparing observed and imputed data (train partition) . . . 14
8 Biplot . . . 16
9 Number of cluster obtained with PAM . . . 19
10 Cluster 1 and Cluster 2 . . . 19
11 Head of the final dataset resultados . . . 20
12 Number of individuals in each cluster . . . 20
13 Classification trees based on Gini Index (A) and Information Entropy (B), with a complexity parameter of 0.00001. . . . 23
14 Cross-validation error plots: Representation of the relative errors of the cross-validation. Horizontal line is drawn 1SE above the minimum of the curve. . . . 25
15 ROC curves for all tree models and its computed Area Under the Curve (AUC). . . . 26
16 Tune randomForest for the optimal `mtry` parameter with respect to the Out-of-Bag error estimates. . . . 28
17 Important predictors of RF model based on the mean decrease in node impurity (Gini index). . . . 28
18 qqplot for patients with heart disease . . . 29
19 qqplot for patients without heart disease . . . 30
20 Royston-test for two group of patients . . . 31
21 Boxplot of different variables in two groups . . . 31
22 Covariance in each pair of variables . . . 32
23 confusion matrix of lda on real data . . . 32
24 Result of lda on real data . . . 33
25 Confusion matrix of lda on scaled dataset . . . 33
26 Result of lda on scaled dataset . . . 34

27 Final performance of RF model in test partition data that approximate the true error of this model. . . . 35
A.1 Project work plan and assignment of tasks . . . iii
B.1 Pearson correlation between all continuous variables. . . . iv
B.2 Stats of Tretbps variable before the changes . . . iv
B.3 Stats of Tretbps variable after the changes . . . iv
B.4 Stats of Chol variable before the changes . . . v
B.5 Stats of Chol variable after the changes . . . v
B.6 Boxplot of numeric variables after changing zeros . . . v
B.7 Missing values with *va* source removed . . . vi
B.8 Missing values with *ca* and *thal* variables removed. . . . vii
B.9 plot for *chol* and *tretbps*. . . . vii
B.10 plot for *thalach* and *oldpeak*. . . . viii
B.11 Density plot for *chol* and *Oldpeak*. . . . viii
B.12 Correlation values. . . . ix
B.13 Dimensions summary. . . . ix
B.14 Eigenvalues plot. . . . ix
B.15 Dimension 1 test . . . x
B.16 Dimension 2 test . . . x
B.17 Circular plot by quality . . . xi
B.18 Contribution of variables to first and second dimensions. . . . xii
B.19 Number of tree vs error. . . . xiii
B.20 Correlations categorical data. . . . xiii
B.21 Variance retained by each dimension. . . . xiv
B.22 Biplot distances. . . . xv
B.23 Individuals correlation. . . . xvi
B.24 Categories correlation. . . . xvii
B.25 Cos2 values. . . . xviii
B.26 Cos2 of variables in dimensions 1. . . . xix
B.27 Cos2 of variables in dimension 2. . . . xx
B.28 Scatter plot of most important individuals. . . . I

List of Tables

1 Variables description and its specified values. . . . 2
2 Centroids of each cluster . . . 20

3 Predictive power of tree based models when predicting validation data partition. Accuracy refers to the number of correct predictions made divided by the total number of predictions. Sensitivity (True Positive Rate, TPR) refers to the proportion of those who have risk of CVD that received a positive result (1) on this test. Specificity (True Negative Rate, TNR) refers to the proportion of those who do not have the condition that received a negative result (0) on this test. 26

1 Heart Disease Data Set

2 Project summary

We have decided to base our course project on the Heart Disease dataset [2] due to its important role in cardiovascular diseases (CVDs) prediction. CVDs are one of the main cause of death globally every year, representing in 2019 30% of all deaths worldwide. Moreover, 85% of these deaths were caused by heart attacks and strokes.

For this purpose, a large dataset containing the more significant physiological variables have been be created in order to be able to predict a possible heart disease in patients. Observations will be classified in two groups; heart disease (1) or healthy (0).

Our goal with this project is to assess through different techniques what are the factors that affect heart disease the most and produce models that using those factors would be able to predict the presence of risk for CVD. By detecting patients with risk of CVD based on their physiological risk factors such as cholesterol, hypertension, diabetes, etc., an early diagnosis of the patient could be made.

3 Data source

The chosen data belongs to the UCI Machine Learning Repository. The linked directory contains 4 databases concerning heart disease diagnosis. The data was collected from the four following locations:

- Cleveland Clinic Foundation (303 observations)
- Hungarian Institute of Cardiology, Budapest (294 observations)
- V.A. Medical Center, Long Beach, CA (200 observations)
- University Hospital, Zurich, Switzerland (123 observations)

3.1 Data description

Each database has the same instance format. While the databases have 76 raw attributes, only 14 of them will be actually used. Hence, our aim is to create a large dataset by combining the different datasets already available independently over 14 common attributes.

Variable information

Heart Disease dataset with 14 physiological variables associated with CVD described in Table 1.

Variable	Description [value]
age	age of the patient [years]
sex	sex of the patient [1 = male; 0 = female]
cp	chest pain type [1= typical angina, 2= atypical angina, 3= non-anginal pain, 4= asymptomatic]
trestbps	resting blood pressure [mm Hg on admission to the hospital]
chol	serum cholestoral [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, 2= showing probable or definite left ventricular hypertrophy by Estes' criteria]
thalach	maximum heart rate achieved [value between 60 and 202]
exang	exercise induced angina [1 = yes; 0 = no]
oldpeak	ST depression induced by exercise relative to rest [value measured in depression]
slope	the slope of the peak exercise ST segment [1= upsloping, 2= flat, 3= downsloping]
ca	number of major vessels colored by flourosopy [0-3]
thal	[3 = normal; 6 = fixed defect; 7 = reversable defect]
num	diagnosis of heart disease, angiographic disease status [0 (no presence) to 4]

Table 1: Variables description and its specified values.

4 Project plan

4.1 Aims

- To determine the most significant associated physiological risk factors with heart failure and CVDs in adult patients (aged 28-77 years).
- To create a predictive model for adult patients with potential CV physiological risks to make and early diagnosis of possible heart attacks and strokes, and therefore, gain time for a preventive treatment.

4.2 Work plan and assignment of tasks

We have divided the project in 6 main tasks with different sub-tasks, for which we have assigned a task leader. All the members of the project will work collaboratively in all of the project tasks. However, each member will be responsible of at least one sub-tasks. The task leader will be in charge of the organisation of the work among the other members and will lead the discussions related to that task. Specified details for work plan organization in Table A.1

4.3 Potential risks

- Imbalance classification: Many real-world classification problems have an imbalanced class distribution. In heart failure prediction, distribution across classes in the dataset is more likely to be skewed. We expect that many of the observations could belong to healthy patients. Most of the algorithms used for classification problems are designed under the assumption of equal distribution of classes. Therefore, imbalance pose a challenge for predictive modeling and will require specialized techniques in data preparation, where we will need to make stratified partitions of the dataset.

- MultiCollinearity: It occurs when two or more independent variables are highly correlated with one another in a predictive model. We expect that some of the physiological variables (predictors) of the dataset will change similarly depending on the patient health condition. Thus, the predictive model will result unstable. Even though multicollinearity may not affect the accuracy of the model as much, we might lose reliability in determining the effects of individual independent variables on the target variable in our model leading to problems when interpreting it. To reduce risk of these problems, correlation matrices should be calculated and we should pay attention to situations where there is a significant correlation between independent variables.

- Missing data: In a first brief exploration of the dataset we have notice the presence of missing data in some of the databases that are used to create the whole dataset. By dropping many rows or columns with missing values, we will lose valuable information that might have a significant impact on the target variable, which will lead to model overfitting in the training set and worse prediction performance on the test partition. Therefore, we will need to explore which type of missing data we have and deal with it using appropriate techniques for its imputation (k-nearest neighbors, mean input, etc.) or deletion. Normally, classification is the best when a model-based technique (like knn method) is used when imputing missing data. This is due to the fact that the original variance of the data is better approached when using knn as a replacement of missing data.

- Miscommunication: In another hand, we have potential risk regarding people. In this case, as it's a project group, there are different people working together so sometimes it's difficult to speak each other and have a smooth communication.
- Lack of commitment: This is a critical problem for the group because if somebody is committed to doing a part of the project and suddenly he/she doesn't do it, the group is affected by this person.

5 Data Cleaning

In order to have a unique data set as we desire containing all the information the first step is combine all datasets and add data `source` column to keep track of the origin of each observation. An `ID` column with the row number is added to ensure we can always identify properly all the observations. Therefore, we obtain a initial data set with dimensions a 920×16.

Since in our dataset description it inidicates that missing values are identified with -9, we replace those values with `NA`, so that we identify them as missing.

Due to the fact that with a brief first exploration of our data we detected that very few samples represent high degrees of risk (`num` $= 3$ or 4), our study will focus the goal in to determine the presence/absence of risk of suffering CVDs rather than the degree of risk. Therefore, we have transform our response variable into 0-1 binary variable, grouping together levels 1 to 4.

Finally, variables classes are properly converted into `factor` and `numeric` types.

6 Data Explanatory Analysis

The very first step in any multivariate analysis project is an explanatory analysis of the data. This first look at our data is focused on finding any irregularities on our data such as outliers, missing values, skewed categorical variables and unbalanced classes. Our dataset is composed with mixed types of variables stated below:

- Categorical variables: sex, cp, fbs, restecg, exang, slope, ca, thal, num, (id, source)
- Continuous variables: age, tretbps, chol, thalach, oldpeak

A first insight of our data is obtained with the built-int function in R `summary()`:

```
      age         sex      cp          tretbps            chol          fbs
 Min.   :28.00   0:194   1: 46   Min.   :  0.0   Min.   :  0.0   0   :692
```

```
1st Qu.:47.00   1:726   2:174   1st Qu.:120.0   1st Qu.:175.0   1   :138
Median :54.00           3:204   Median :130.0   Median :223.0   NA's: 90
Mean   :53.51           4:496   Mean   :132.1   Mean   :199.1
3rd Qu.:60.00                   3rd Qu.:140.0   3rd Qu.:268.0
Max.   :77.00                   Max.   :200.0   Max.   :603.0
                                NA's   :59      NA's   :30
restecg       thalach       exang        oldpeak         slope
0   :551   Min.   : 60.0   0   :528   Min.   :-2.6000   1   :203
1   :179   1st Qu.:120.0   1   :337   1st Qu.: 0.0000   2   :345
2   :188   Median :140.0   NA's: 55   Median : 0.5000   3   : 63
NA's:  2   Mean   :137.5              Mean   : 0.8788   NA's:309
           3rd Qu.:157.0              3rd Qu.: 1.5000
           Max.   :202.0              Max.   : 6.2000
           NA's   :55                 NA's   :62
      ca        thal     num             source          id
0.0    :176   3    : 30   0:411   cleveland  :303   Min.   :  1.0
1.0    : 65   3.0  :166   1:509   hungarian  :294   1st Qu.:230.8
2.0    : 38   6    : 28           switzerland:123   Median :460.5
3.0    : 20   6.0  : 18           va         :200   Mean   :460.5
0      :  5   7    : 75                             3rd Qu.:690.2
(Other):  6   7.0  :117                             Max.   :920.0
NA's   :610   NA's:486
```

Firstly, it can be observed from our response categorical variable `num` that the number of observation belonging to each category is well balanced. However, other qualitative variables are not as well balanced across their categories. We can observe from `sex` that our data set is composed of much more males (726) than females (194) patients. Therefore, we will need to take this into account in our validation protocol to obtain stratified partitions of the data.

Regarding the continuous variables, `tretbps`, `chol` and `thalach` present a slightly high range of magnitudes. Therefore, due to its high variability when performing exploratory visualizations it might be better to try to plot them in log-scale and consider a log10() transformation in our predictive analysis. For cholesterol (`chol`) and resting blood pressure (`trestbps`), it is noticed that a minimum of 0 are non-sense values that may be detected as outliers and treat them consequently.

It can also be observed that some variables `tretbps, chol, fbs, restecg, thalach, exang` and `oldpeak` contain few missing values, whereas `slope, ca` and `thal` contain high percentage

of missingness. Therefore, appropriate techniques for the deletion or appropriate imputation of missing values will be discussed in later steps.

The `source` variable will not be considered in our analysis and further predictive models, since our aim is to built a model based on patients physiological characteristics. Nevertheless, we will keep this variable for informative purposes.

6.1 Visualization of Variables

For a better understanding of the variables we used different visualization plots for both qualitative and quantitative variables.

To study the continuous numerical variables we have perform a pairwise plot with the `ggpairs()` built in function in `GGally` library. It consists of two-dimensional scatter-plots for each variable-combination along with its distributions. In the upper panel the Pearson correlation coefficients for each combination are displayed, as an overall and studied per CVD condition too, along with the significance levels (Figure 1).

As we have previously commented, due to the high variability of some variables, variables on logarithmic scale could show better the linear tends between the variables compared with the non-scaled data. Therefore, we have tried to transform `chol` and `thal` into logarithmic scale. However, since our dataset contain many missing values, a log-scale can not be directly applied. If we omit missing values for visualization purposes only, log-plots show no more information than normal scaled variables. Consequently, in this report we considered normal scale for all variables in pairwise plots.

As expected, it is observed in Figure 1 that age distribution of patients with risk of CVD present is slightly higher than for healthy patients. Similarly, those patients also present higher values of `tretbps, chol, thalch` and `oldpeak`. All variables follows an approximately *Gaussian* shape distribution.

As previously stated, in these plots we can be observed the abnormal values in 0 detected for both `chol` and `thalach`.

The *Pearson* correlation between all continuous variables have been showed in Figure B.1 too with its Pearson correlation coefficients where we can visualize them better.

To study whether significant differences in presence or absence of CVD (`num`) occur between different levels our categorical variables we used bar-plots (Figure 2).

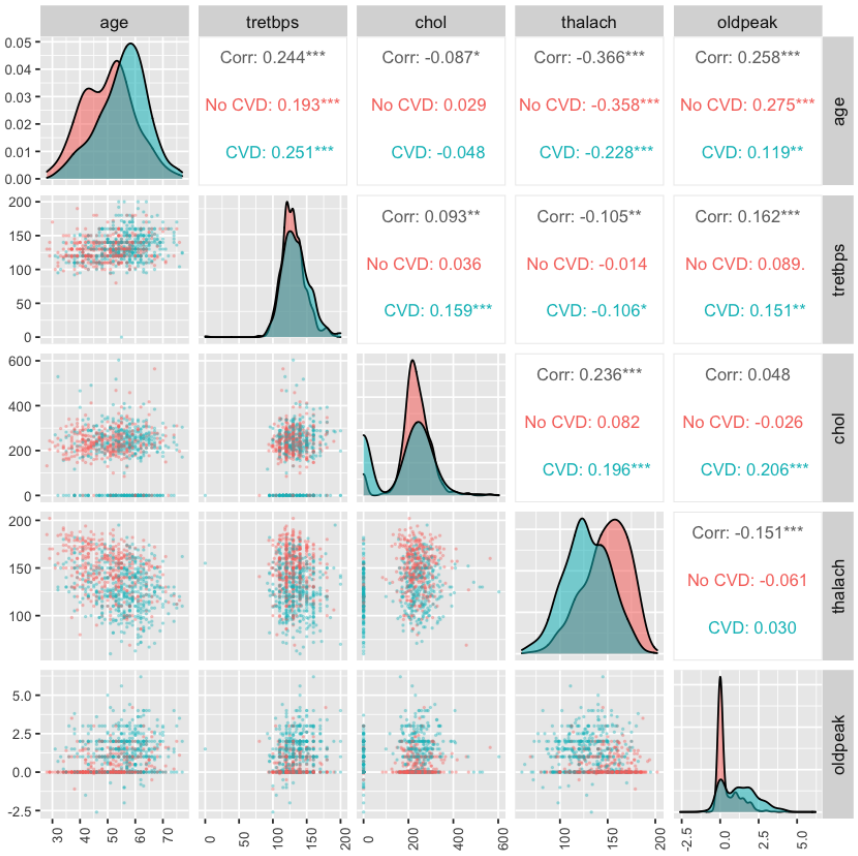

Figure 1: Pairwise scaterplot and correlations of continuous variables colored according to absence(0) or presence(1) of CVD.

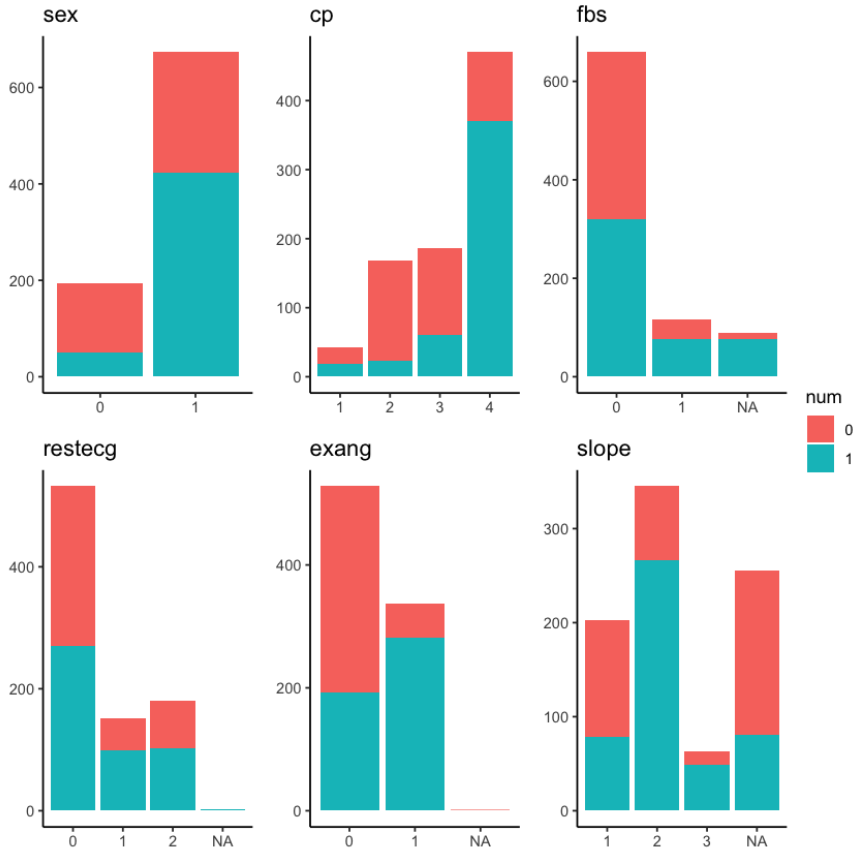

Figure 2: Barplot of categorical variables colored according to absence(0) or presence(1) of CVD.

Diseased patients are more present in level 0 in exgang, fbs and restecg. The other variables categories are approximately stratified among patient disease condition.

6.2 Detection of Missing Data

With respect to the missigness of the data set, as previously mentioned, for variables `age`, `sex`, `cp`, `restec` and `num` no missing values are reported. However, we have three variables with high percentage of missingness. `Slope` (34%) and two that exceeds the 50%: `ca` (66%) and `thal` (53%). The rest of the variables contain a percentage of missgness lower than 10%. Overall, the amount of missingnes present in our dataset it 12.7% (see Figure 3).

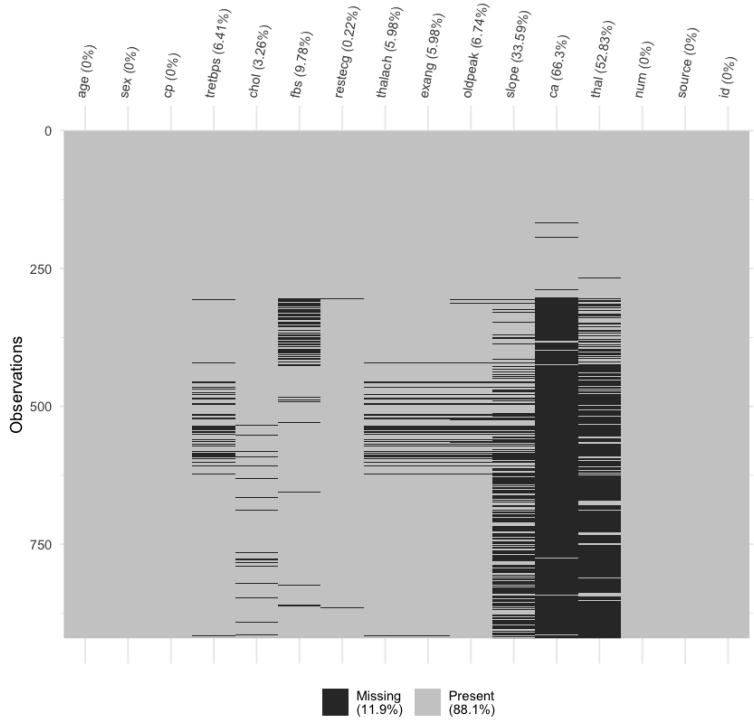

Figure 3: Overall missing values per variables across all the observations.

We also wanted to observed how missing values were distributed across different dataset sources. Observing Figure 4 it can be noticed that `va` source contain a higher level of missinged compared with the others.

Thereby, we have tried two different strategies to reduce the amout of missingnes to be able to apply an appropiate method to impute missing values in fuerther steps.

1. Drop `va` hospital from our study since it contains many missings (23.3%).
2. Remove both `ca` and `thal` variables that exceeds 50% missigness. We consider that the imputation of those many missing could introduce too much biased data in these variables making them uninformative for the analysis. We could be biasing our analysis more than gaining information.

Aditionally, we have removed patients with more than 4 values missing en each of the strategies.

Dropping both `ca` and `thal` qualitative variables reduces the overall percentage of missingness in our data set up to 3.5%, which is better than the 9.7% missingnes retained after removing `va` source dataset (see Appendix B.7 and B.8). Therefore, we will continue our analysis removing

ca and thal variables but keeping observations from va hospital. In later steps at Section 7.2 we will impute our missing with appropriate techniques.

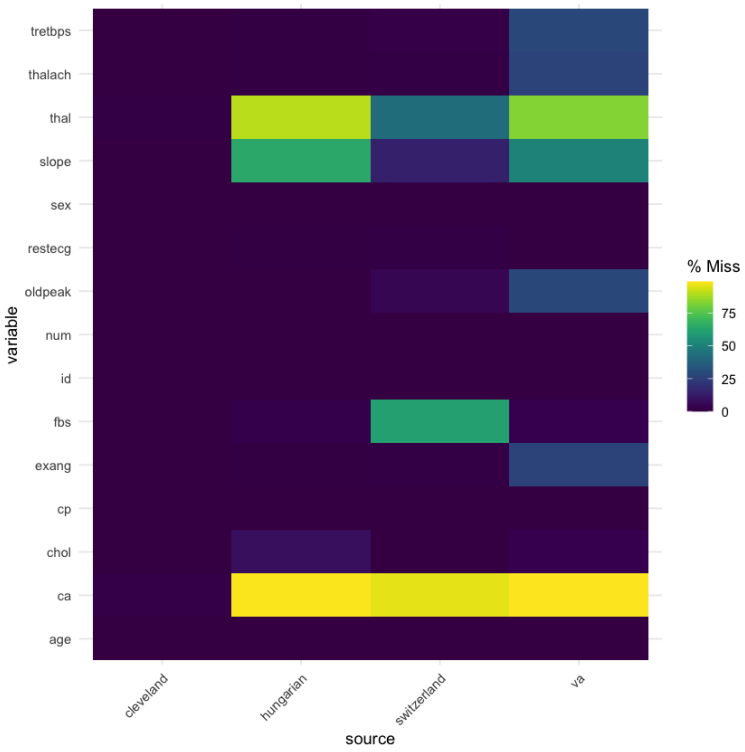

Figure 4: Missing values per variables and origin source.

6.3 Detection of Outliers

6.3.1 Multivariate Outlier Detection

The classic Mahalanobis Distance is a method that measures the distance between a point and a distribution to which that point belongs. This techniquee acts as a statistical measure for the classification of a point as an outlier based on a chi-square distribution.

For many of its multivariate functions, this package uses the Robust Mahalanobis distance based the Minimum Co-variance Determination. While classic calculation of the distances is widely accepted, the robust calculation is preferred.

Outlier detection techniques will normalize all of the data, so the mismatch in scaling is of no consequence.

```
Projection to the first and second robust principal components.
Proportion of total variation (explained variance):  0.7481015
```

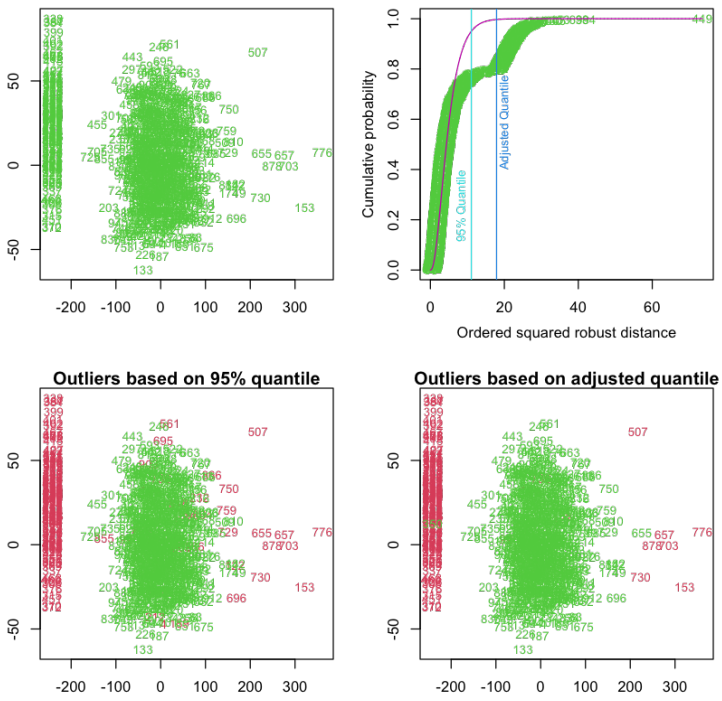

Figure 5: Squared Mahalanobis Distances for our data set observations against the empirical Chi-Squared distribution function of these values

The graphics above (Figure 5) provide an intuitive visualization of the outliers contained in this dataset. All observations are plotted against the 1st and 2nd principal components, and every observation outside of the Chi-Square quantile is colored in red. The function displays outliers outside of the adjusted Chi-square quantile as well.

Outliers showed at left side of the 4th plot (negative side) in Figure 5 outliers observed correspond to observations with 0 values in `chol` and/or `tretbps` variables. In Univariate section it is specified their treatment. At right side (positive) the observation in the 4th plot the detected outliers correspond with observation with high values of `chol` specially. However this is crucial information in our data for unhealthy patients en thereby we will keep these observations.

6.3.2 Univariate Outlier Detection

As we want to detect which numeric variables have the outliers, we visualize our qualitative data in boxplots (see Figure 6).

From the boxplot it can be observed that the `chol` and `tretbps` variables have many outliers specially the first one (0 detected previously in Figure 1). So we will treat them separately.

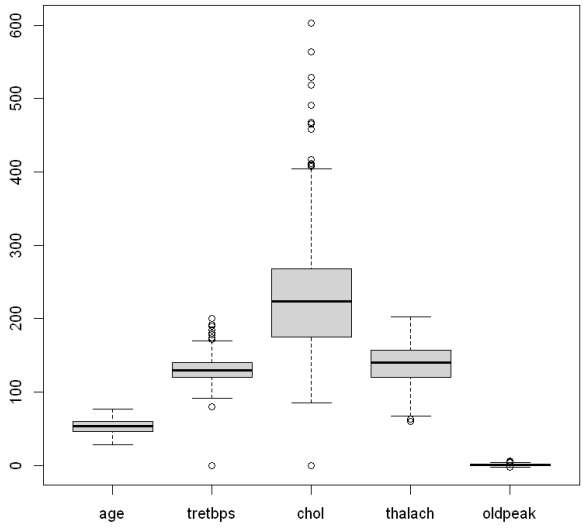

Figure 6: Boxplot of numeric variables

Tretbps variable:

In the appendix B.2, we have obtained tretbs outliers. Since the acceptable ratio of resting blood pressure(trestbps) starts from 80-90 mmHg, values below this range are meaningless such as 0 mmHg. So taking this into account, we will change these with NA values.

According to the American Heart Association, it is important to emphasize that less from 80 to 100 mmHg is low blood pressure, from 100 to 129 mmHg is normal, high than 130 mmHg is high blood pressure. If the patient has more than 180 mmHg, he/she has a hypertesive crisis who has to go to the doctor inmediately. The outliers which have a high value(more than 170 mmHg) are crucial because it give us an significant information.

Consequently, we have changed the values less than 80 for NA values because these value are zeros as it shows in appendix B.3.

It seems that we only have a 0 value. These are new outliers which are relevant for our research so we will keep them.

Chol variable:

In the appendix B.4, we have obtained chol outliers. According to the National Institutes of Health, it's possible to have values more than 240 mg/dL which are consider high cholesterol

level, inclusive we could have values above 404 mg/dL. However, it is inconceivable to have 0 mg/dL. As consequence, 0 values will be replaced by NA values and consequently will be imputed with missing techniques in later steps.

As you can see in appendix B.5, we don't have 0 values anymore. Here we have the new outliers with NA values which don't influence in our results. Also, these new outliers are considered as explained above.

Finally, we obtained the ideal boxplot (appendix B.6) after changing the zeros value to NA values of chol and tretbps variable.

7 Prepare data for analysis

7.1 Validation Protocol

Since our data set contains 866 patients, we will split the data into 60% training data, 20% validation and 20% test, stratified by the response variable `num`. For this prupose we use built-in function form the `splitTools` R library.

1. Train: the training set is the largest split of our dataset that we reserve for training the predictive models that will be further developed in next reports.

2. Validation: separate section of the data that we will use during model training to get sense of how well models perform (validation error) and to tune possible hyper-parameters.

3. Test: due to the fact that the validation set is heavily used in model creation, to detect if overfited during validation has been introduced, it is important to hold back a unobserved separate set of data. Evaluation metrics (i.e. generalized error or predictive error) will be based on the test set at the very end of the project, to get unbiased evaluation of a final model to accurately asses its performance.

7.2 Pre-processing

In order to consider the data ready to be fed into a predictive model, different pre-processings need to be done depending on the technique applied.

It is important that the pre-processing is applied separately in each partition of the data (train, validation and test splits) in order to not contaminate the data contained in each one.

Thereby, in this report we will describe a first brief pre-processing applied in `train` partition that

will be the base for further steps.

For `validation` and `test` partitions the same preprocessing will be applied separately.

7.2.1 Imputation of missings - MICE

This method use random forests to get their imputations, and it is a parametric model and it assumes linearity in data set. Mice method gives multiple imputations and we should find which imputation is better. Different values for "m" were used and we found that $m = 3$ has better performance, additionally for different type of variables different method have been used, *pmm* for continuous variables, *logreg* for binary categorical variables and *polyreg* for categorical variables. Based on Figure 9 we can see that Mice imputation does not change distribution for features with missing values.First row belong to continuous features *fbs*, *restecg* and *slope* and second row is for categorical features with missing values. More plots that show performance of Mice are provided in (Appendix B.9, B.10 and B.11).

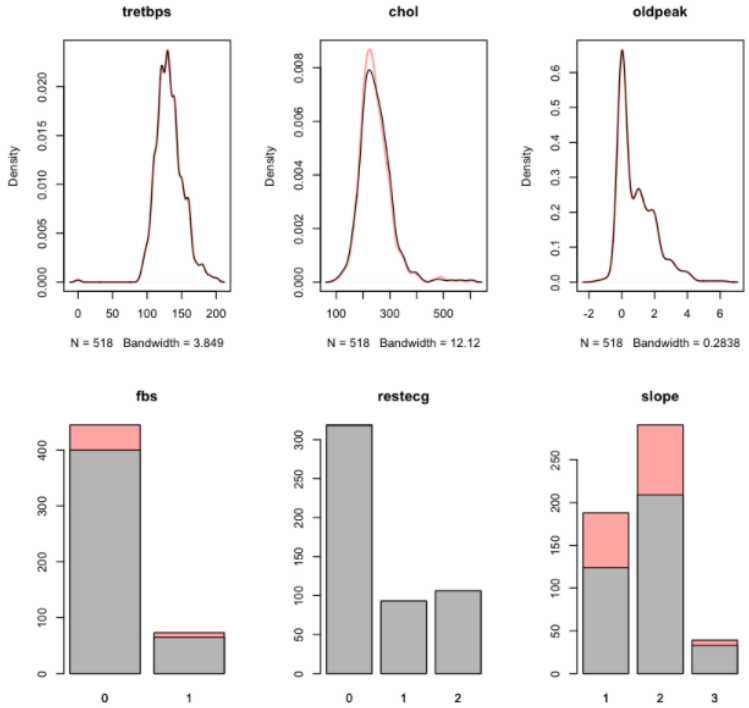

Figure 7: Comparing observed and imputed data (train partition)

Results obtained for missing imputation of `validation` and `test` partitions are equal to the ones obtained in the train set. Missing values imputation works also well in these two sets showing no difference in variables distribution before and after imputation.

7.2.2 Imputation of missings - Miss Forest

Missforest uses random forest like Mice, but it can be used for mixed data types (numeric and categorical variables) and it is a non-parametric model that makes no assumptions about the relationship between the features. To find performance of this method, estimated error for each variable is calculated and based on it's result we can conclude that Miss forest does not behave well and the assumption of linearity in Mice helps us to have a better estimation.

8 Dimensionality reduction

8.1 Principal Component Analysis (PCA)

It is well known that PCA models work with numerical variables. As seen previously, our model not only contains the desired variable type. Before starting the dimensionality reduction, we create a new dataframe containing only numerical variables and then we check the correlations within all the preprocessed data. (See figure B.12)

There are positive and negative correlations. Negative correlations inform us that if one variable increases another decreases so we will see how they behave when doing the dimensionality reduction. (See figure B.13).

Once we perform the model, we are ready to explore the PC. Firstly we summarize the model. Dimension 1 explains 33 % of the overall variance. Dimensions 2 and 3 explain 22% and 16% respectively. The variance explained by Dimensions 3, 4 and 5 is close to be the same. This can be a first check to know which dimensions to focus on as we will not work with all of them.

Another good practice to start would be to check the Eigenvalues, for dimension 1 the value is 1.6971 and for dimension 2 1.1929. They inform us about the magnitude of the principal components. They also help us to decide how many dimensions we are interested in. The other 3 dimensions have values under 0.9.

A scree plot can help us understand the importance of each dimension by plotting its eigenvalues. The first two dimensions are the ones that capture most of the variance/information so we can ignore the rest without loosing too much information. So we could say that we place the "elbow" (cutting point) after dimension 2. (See figure B.14)

Next we plot a correlation circle, this way we will see the correlations between the original dataset and the principal components. Variables that are uncorrelated are orthogonal to each other. Variables with same direction have a higher correlation. B.17

Orange/red variables/arrows are those whose contibution is of the highest quality. Can be noticed that the lenght of tretbps and oldpeak arrows is less than the other 3. So we can plot a barplot to see the contribution of the variables to the dimensions. (See Figure B.18)

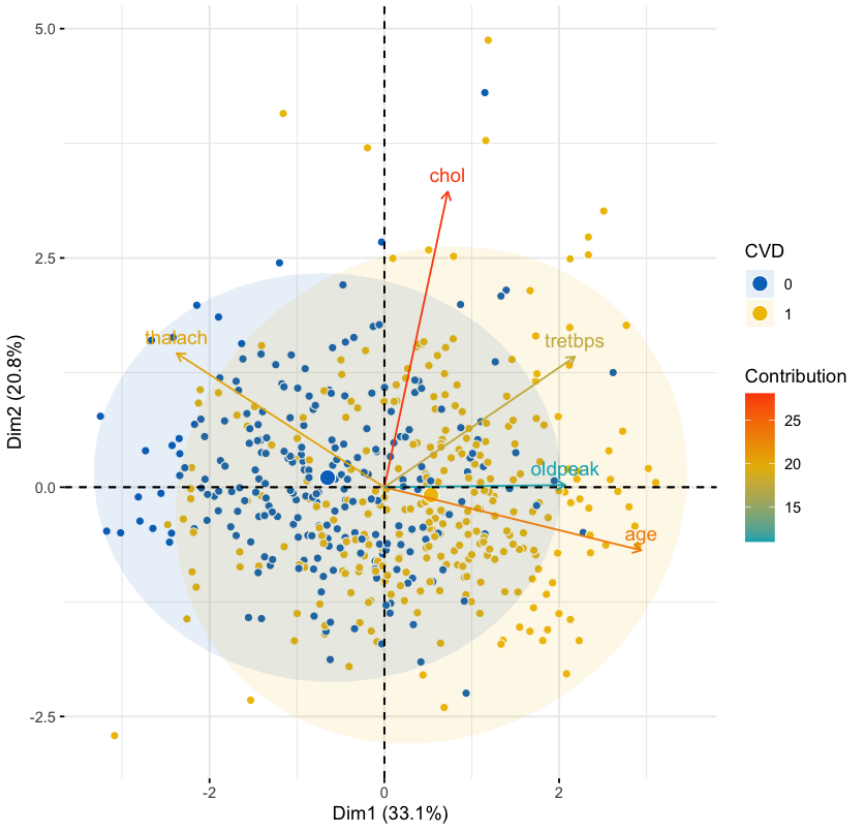

Figure 8: Biplot

We added to the information we have a biplot (Figure 8) to see if one of the two dimensions can separate individuals coordinates across disease condition. Individuals labelled in blue (no CVD) seem to be shifted to the left side of Dim1 (associated with thalach variable), while individuals that present the disease labelled in yellow are more located at the right side. These are more positively associated then with chol, tretbps oldpeak and age. Thereby, we expect that older individuals along with mild/high values in variables positively correlated with dimension 1 will more likely present CVD. Considering this, Dimension 1 is the one that explains if the individuals will have a cardiovascular disease, as we can not very well differentiate disease presence using only dimension 2. All this analysis have a visual character. So we do some statistical tests to check our assumptions and gain in robustness. (See Figure B.15 and B.16) Variables negatively correlated in dimension 1 are the ones with higher correlation in dimension 2. This means that

dimension 2 better explains chol and tretbps. P-values are significant so the choice of just getting two dimensions was right.

8.2 Multiple Correspondence Analysis (MCA)

We first start our analysis by selecting all categorical columns in all our preprocessed data. Correlations of the data can be seen in B.20).

Once the model is run the proportion of variances retained by the different dimensions can be seen in B.21). Same as PCA, eigenvalues are sorted and used as they are informative about the variation that occurs. BY checking them visually, we see that first dimension has higher percentage of explained variation. We will only focus on the first two dimensions as the variation explained by further dimension is not significant.

By plotting B.22) a global pattern within the data is shown where rows/individuals are represented by blue points and columns/categories by red triangles.

A measure of dis/similarity is show represented by the distance between column and row points. Positive values of dimension 1 seem to be closer on the factor map. This means that they share similar profile. Values in dimension 2 are grouped around the 0 axis across positive and negative values of dimension 1.

Next we focus on studying the variables, specially the correlation between variables and principal dimensions. To visualize the correlation check B.23).

In this graph we can see which variables are more correlated with each dimension. Sex, slope, exang, num and even cp are the most correlated categories with dimension 1. Fbs and restecg in the other hand are more correlated with dimension 2.

Following we checked the coordinates of categories, the plot B.24). shows the relationships between variable categories.

Variable categories with a similar profile are grouped together such as exang1, slope1,cp3, sex1,num1. Negatively correlated variable categories are positioned on opposed quadrants like category sex or restecg. The distance between category points and the origin measures the quality of the variable category on the factor map. Category points that are away from the origin are well represented on the factor map.

Also the quality of the representations can be seen using squared cosines. Follwing figure B.25) help us measure the degree of association between variable categories and a particular axis. All

variables well represented by both dimensions will have a vulue close to 1. First plot shows us the importance of each category. Red categories are the ones with highest degree of association. Therefore values are much closer to 1 than if variables are coloured by blue, representing the lowest degree of relation.

The plot does not show many variables with red meaning that the association degree between variables and axis is medium.

The barplot B.26) helps us identify which variables are explained by first dimension. Num, exang, cp, slope and sex are the variables with more contribution to dimension 1 while restecg source and fbs contribute more in second dimension B.27) . If contributions were uniform, values would be around the red line.

The most important variables are summarized in the scatter plot B.28). Values away from the 0 axis have a higher degree of association with axis. Individuals that are more far of the 0 axis have a higher degree of cos2, having a greater contribution.

9 Clustering

In this project, we are going to apply Clustering with Gower due to the fact that our dataset has continuous and categorical variables. So, in order to do that, we must use the preprocess data without id variable neither source variable. Firstly, we have to keep in mind that the continuous variables must be scaled. As a consequence, we have obtained `datclus` dataset which is ready to work in Gower. For calculating the Gower distance, Daisy function is applied for clustering.

In order to calculate silhouette width for many k, Partitioning Around Medoids algorithm (PAM) has employed which is less sensitive to outliers that are important for our project.

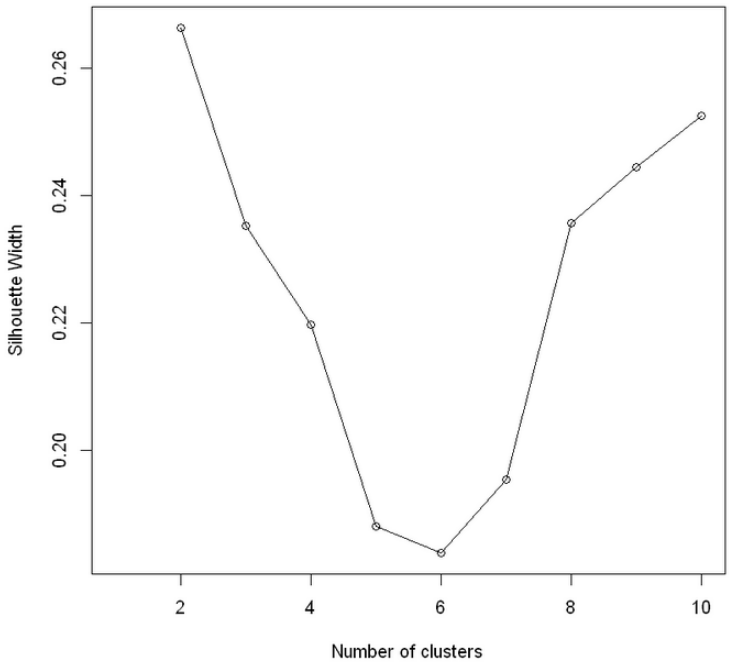

Figure 9: Number of cluster obtained with PAM

As it shows in the figure 9, we have obtained two clusters which is the optimal.

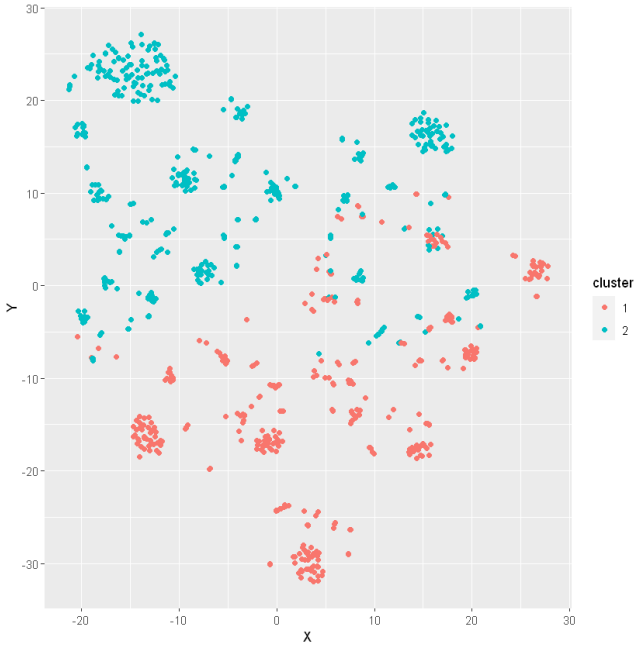

Figure 10: Cluster 1 and Cluster 2

The previous graph (figure 10) shows the two cluster widely separated.

Due to the cluster giving us two cluster, we have obtained two centroids as it shows in the next table. The cluster 1 has a centroid of 755 which means that all individual who is in cluster 1 is similar to the individual 755. The cluster 2 has a centroid of 647 which means that all the individuals who are in the cluster 2 are pretty similar to the individual 647. This happens because with PAM, each centroid corresponds to an individual of the dataset.

	sex	cp	fbs	restecg	exang	slope	age	tretbps	chol	thalach	oldpeak
	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
755	2	2	1	1	1	1	-0.2267856	-0.1269202	-0.3761502	0.4796299	-0.8081812
647	2	4	1	1	2	2	0.1992368	0.4113408	0.3799203	-0.3691448	0.5628876

Table 2: Centroids of each cluster

Finally, the `resultados` is created to constrast the values of the given output `num` and the result of the clustering `pam_fit.cluster`. So, this new dataset `resultados` shows the real values of all the variables and the result of the clustering (`pam_fit.cluster` variable).

	age	sex	cp	tretbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num	source	id	pam_fit.cluster
	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<fct>	<fct>	<fct>	<dbl>	<fct>
1	63	2	1	145	233	2	3	150	1	2.3	3	0	cleveland	1	1
2	67	2	4	160	286	1	3	108	2	1.5	2	1	cleveland	2	2
3	53	2	4	140	203	2	3	155	2	3.1	3	1	cleveland	10	2
4	56	2	3	130	256	2	3	142	2	0.6	2	1	cleveland	13	1

Figure 11: Head of the final dataset resultados

In order to verify the effective of the clustering, we are going to compare the output given at the beginning `num` with the obtained result `pam_fit.cluster`. As a reminder, `num` = 0 means that the patient is healthy(no heart break) and if `num`= 1, the patient is sick. On the other hand, `pam_fit.cluster` = 1 means that the patient is healthy and if `pam_fit.cluster` = 2 means that the patient is sick.

```
num      pam_fit.cluster
0:392    1:400
1:474    2:466
```

Figure 12: Number of individuals in each cluster

`pam_fit.cluster` and `num` have nearly the same among of value of number of healthy individuals with a difference of 8 individuals. Also, when it's about of sick patient, there is a difference of 8 individuals. As a consequence, Gower method works very well in our dataset.

10 Predictive Models

10.1 Trees-Based Models

Tree based Models are relatively fast to construct and they produce simple and useful results for interpretation. They naturally incorporate mixtures of numeric and categorical predictor variables and missing values, and they are invariant under transformations of individual predictors. Hence, scaling and other more general transformations are not a problem, and they are not affected by outliers in the predicted value. A basic aspect of tree models is the performance of internal feature selection as an integral part of the process.

In this section, different Tree Based methods for classification will be performed upon the Heart Disease dataset, which contains several continuous and categorical variables that can explain whether or not a patient has a heart disease.

10.1.1 Classification Trees

Classification trees are single trees characterized by the easy interpretation and intuition of the decisions they make in order to do predictions.

Good classification models are the ones for which the *validation error rate* is low. Therefore, to properly evaluate the performance of our classification trees, we must estimate the *validation error* or the *validation accuracy* rather than simply computing the training statistics. We would then compare the *validation accuracy* for the competing models. Thus, we will use the training and validation preprocessed partitions of data that we obtained in previous section 7.2.1. Then the trees are trained using the training set and its performance is evaluated on the validation data.

The first two models that are built corresponds to classification trees using `rpart` function from `rpart R` library.

Each of these two first models have been based on different splitting criterion that search for the splits that produce the minimum impurity[1]. The impurity of a node can be calculated either by the Gini index (Gi) or by the Information Entropy (Hi), which are the two measurements that we used for our models.

The first classification tree is based on Gi and the second on Hi. The tree models (Figure 13) shows the variables that are actually used to construct each tree. The R implementation of

[1]Measure of the homogeneity of the labels on the node. The purest nodes are the ones that contain only one class after the split.

the CART algorithm has determined that the other variables did not contribute to the predictive power of the models. The tree model based on Gi has 17 rules (splits), whereas the tree based on Hi is more complex with 25 rules, both with a minimum complexity parameter (cp) of 0.00001. We observe that both models have achieve equal results on the top splits and variable importance. For both models the most important indicator of CVD appears to be chest pain with a 31.3% importance, followed by exang and thalach with a 15% and 13.4% importance, respectively. Further more, the first split based on the presence of chest pain (1= typical angina, 2= atypical angina, 3= non-anginal pain) or not (4 = asymptomatic), classifies the latest as having an increased risk of CVD. We were expecting opposite results, however we do know how these measurements where taken and we are not experts in the field.

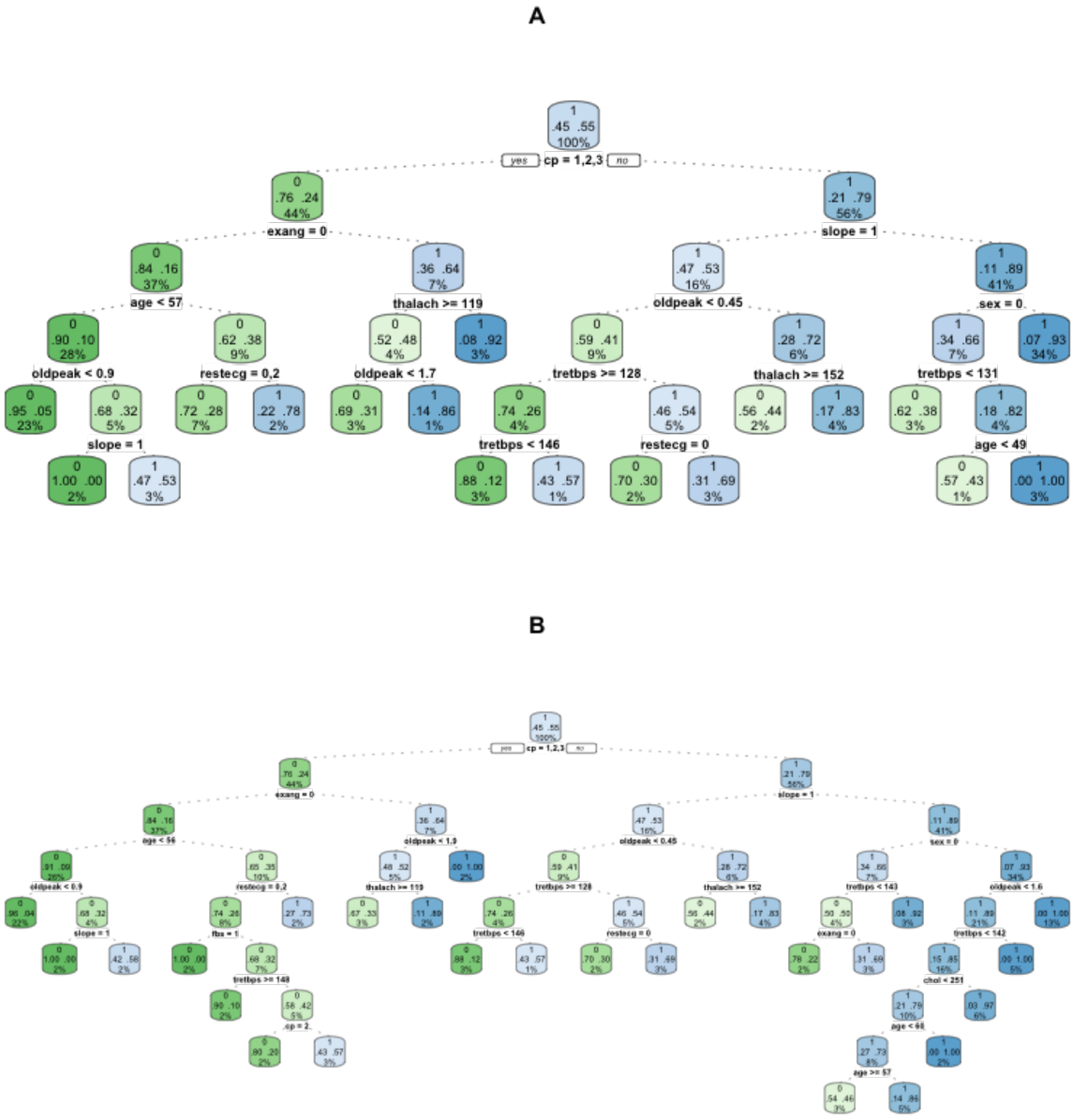

Figure 13: Classification trees based on Gini Index (A) and Information Entropy (B), with a complexity parameter of 0.00001.

To evaluate the models built, the response variable (num) is predicted using the validation data set. Here we use the `predict ()` function. It is important to note that while these models may produce good predictions on the training set, it is likely they overfit the data, leading to poor validation set performance. For this purpose, we have compute for each model the prediction error rate (or miss classification) based on training set and we have estimated this metric also

with the cross validation, using 10-fold. After, we have predict the results on the validation data set and computed the statistics (accuracy, sensitivity and specificity) from the confusion matrix.

For *Gi* tree the prediction error rate in training is 14% while the estimation of miss classification rate given with the cross validation (using 10-fold CV) is 24.7%. This estimation approximates accurately the predictive power achieved when predicting the response variable in the validation set as it can be observed in Table 3. For *Hi* tree the prediction error rate in training is 13% with an estimation in cross validation of 24%. Again these results meet the ones obtained when predicting the presence of risk of CVD in the validation set (see Table 3).

However, the results obtained are far from perfect and can be improved. Smaller trees with fewer splits might lead to lower variance and better interpretation at the cost of a little bias. Therefore, to avoid our models to overfit the data we consider whether pruning these trees can improve performance. We base the pruning on the cp or complexity parameter that determines how deep the trees will grow. In order to select an optimal value of the hyperparameter of complexity we will apply the power of cross-validation on the build complete decision trees to prune them. To obtain the final models, we selected the optimal value of complexity following the criterion of *one standard error* of Breiman et al. (1984) [1]. They found that in the case of selecting optimal tree size for classification tree models, the tree size with minimal cross-validation error generates a model which generally overfits. Therefore, we will select the most parsimonious model whose error is no more than one standard error above the CV error of the best model, hence choosing the simplest model whose accuracy is comparable with the best model.

For both models we have obtain equal pruned trees with a cp=0.0427, which is represented in the cross-validation error plot in Figure 14. Both trees have been pruned up to the first split based on chest pain, which from our point of view, are over simply tree models.

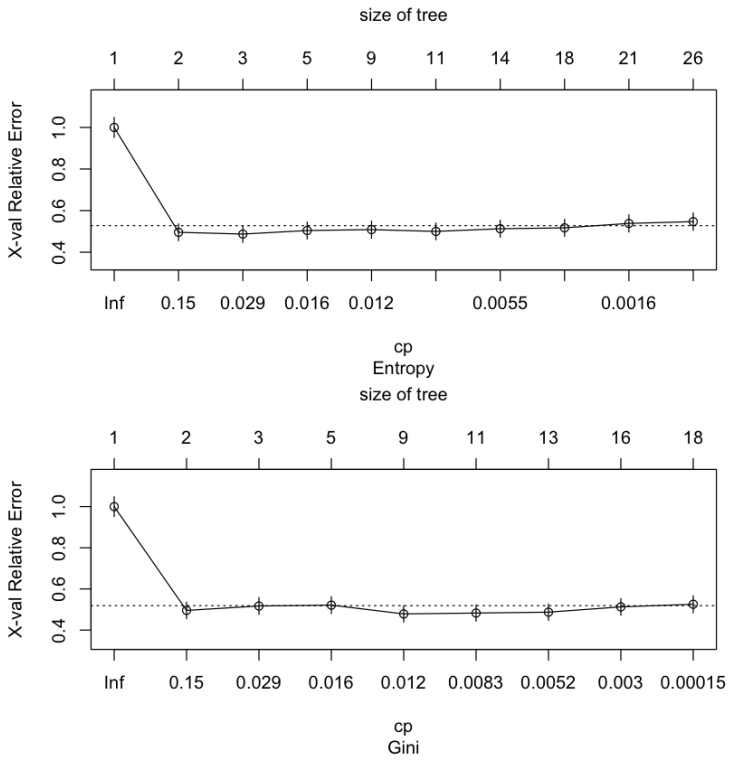

Figure 14: Cross-validation error plots: Representation of the relative errors of the cross-validation. Horizontal line is drawn 1SE above the minimum of the curve.

For all models we have computed the Area Under the Curve (AUC), that can be used as an additional quality metric for comparing classifiers, and plotted its ROC curves to visualize and compare the performance among all tree models (Figure 15).

From Table 3 we can compare all the results obtained for the classification decision trees when predicting the validation set. Based on AUC metric, unpruned decision tree based on Entropy index is the best (0.83) followed by the unpruned *Gi* tree (AUC=0.80). Regarding the accuracy results, both unpruned trees are the models with the highest accuracy 0.75 followed by pruned trees models which have both achieved the same decision tree after punning them with equal results (accuracy of 0.747). We observed that pruning is not always effective in reducing bias. From the obtained results we can observe that pruning has decreased sensitivity in both models compared with the best sensitivity archived by an unpruned tree model, miss-detecting more CV diseased patients. It is important to note that we could choose the pruned trees, which have lower number of splits, if interpretability would be more important than a lower bias since results are close. Nonetheless, our most important aim is to correctly detect vulnerable patients with increased risk of CVD. Therefore, since both unpruned trees have obtain equal highest accuracy, we will choose the model that achieved the highest sensitivity (0.75) which correspond to the

unpruned tree based on Entropy. Overall, this model is the one that better approach our final aim of correctly detecting more vulnerable patients that could be in risk of suffering from CVD.

	Accuracy	Sensitivity	Specificity	AUC
Dtree_entropy	0.7528736	0.7578947	0.7468354	0.8342438
Dtree_gini	0.7528736	0.7157895	0.7974684	0.7997335
Dtree_entropy_prunned	0.7471264	0.7157895	0.7848101	0.7502998
Dtree_gini_prunned	0.7471264	0.7157895	0.7848101	0.7502998
RF	0.8448276	0.8421053	0.8481013	0.8864757

Table 3: Predictive power of tree based models when predicting validation data partition. Accuracy refers to the number of correct predictions made divided by the total number of predictions. Sensitivity (True Positive Rate, TPR) refers to the proportion of those who have risk of CVD that received a positive result (1) on this test. Specificity (True Negative Rate, TNR) refers to the proportion of those who do not have the condition that received a negative result (0) on this test.

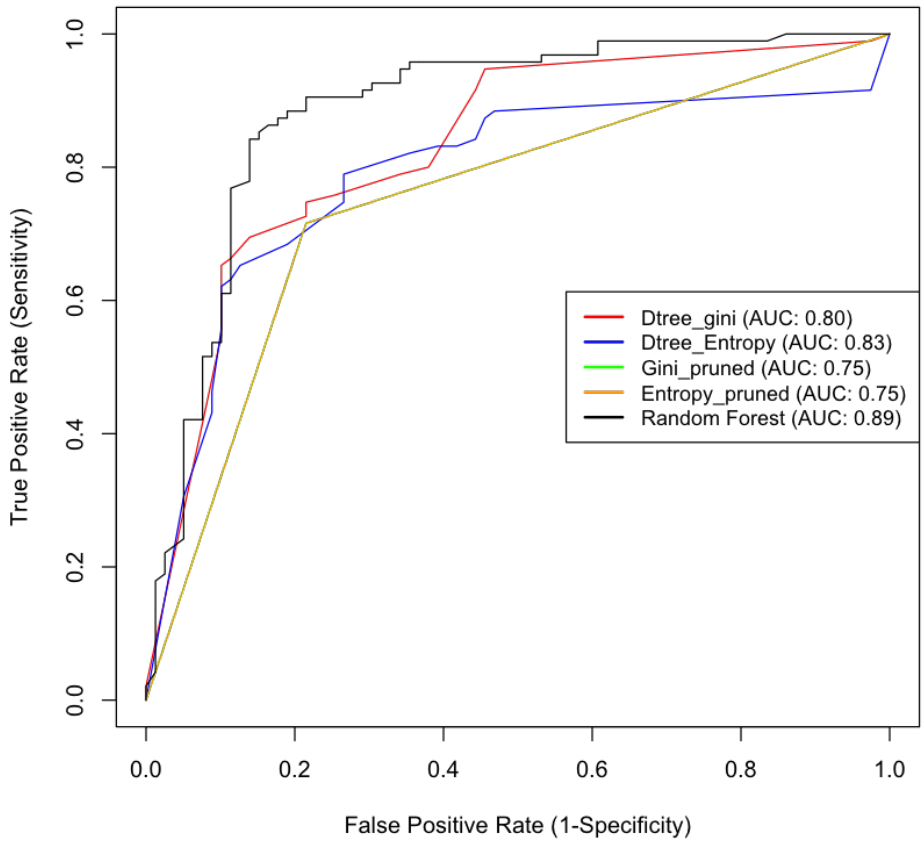

Figure 15: ROC curves for all tree models and its computed Area Under the Curve (AUC).

10.1.2 Random Forests

Random forest (RF) is a tree-based algorithm composed of many decision trees where their outputs are combined to enhance the performance of a given model. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. The idea in RF is to improve the variance reduction of bagging by reducing the correlation between the trees. This is achieved by randomly selecting input variables during the tree-growing process. Thereby, when growing a tree on bootstrapped training samples, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of predictors.

Here, a RF classifier is going to be implemented using `randomForest` from `randomForest R` library. We have used the `tuneRF` function that search for the optimal value (with respect to Out-of-Bag, OOB, error estimate) of `mtry` for randomForest. The minimum OOB error is achieved setting $mtry = 1$, as shown in Figure 27. The random forest model it fitted with the training partition of the data specifying $mtry = 1$ and the number of trees is set to 500. We have tried to run the model with up to 1000 trees, however error stabilize at $\sim$350 (see Figure B.19), so we keep the number of tree at 500 which give us good results without too much computationally cost. With these adjustments, we obtain a RF model with an OOB estimate of error rate of 18.34%. In Figure 17 are represented the important predictors of RF model based on the mean decrease in node impurity.

To evaluate the RF model more accurately the response variable is predicted using the validation data set. As it is summarized in Table 3, the RF achieved an accuracy >84% with a sensitivity and specificity of $\sim$84% too, improving significantly the results obtained with single classification trees.

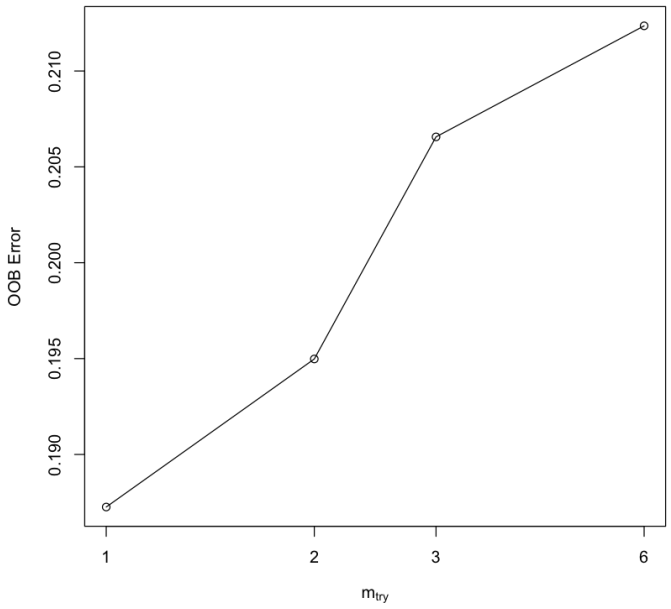

Figure 16: Tune randomForest for the optimal `mtry` parameter with respect to the Out-of-Bag error estimates.

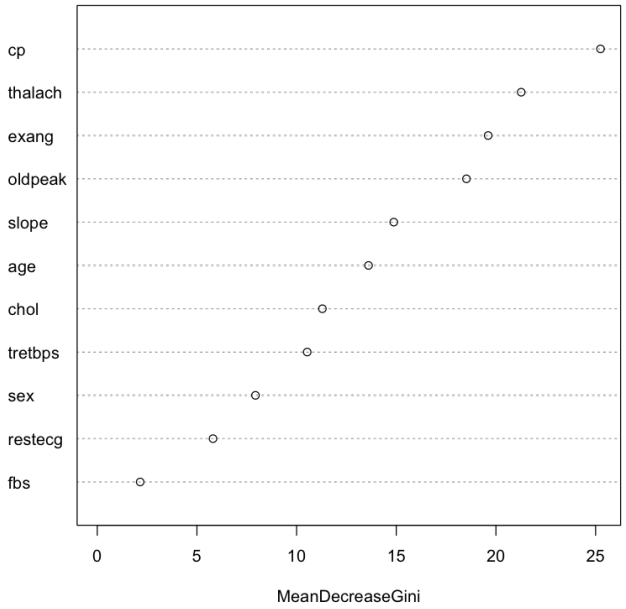

Figure 17: Important predictors of RF model based on the mean decrease in node impurity (Gini index).

10.2 LDA

Linear Discriminant Analysis (LDA) is a basic classification method from parametric statistics. The main idea in LDA is calculating the posterior probability of a sample that gives us with which probability samples belong to a class. As we mentioned, LDA is a parametric model, therefore there is some assumption in this model, that we bring these assumptions in below:

- Each variable should have normal distribution
- Multivariate Gaussian Conditions
- Same variance and covariance between groups

First, we check these assumptions in two group of patients, with heart disease and without heart disease.

10.2.1 Normality assumption

In fisrt assumption we should check that variables has normal distribution. To check normal distribution in continuous variables we used qqplot and Shapiro test.In table below we provided result of qqplot test for these two groups:

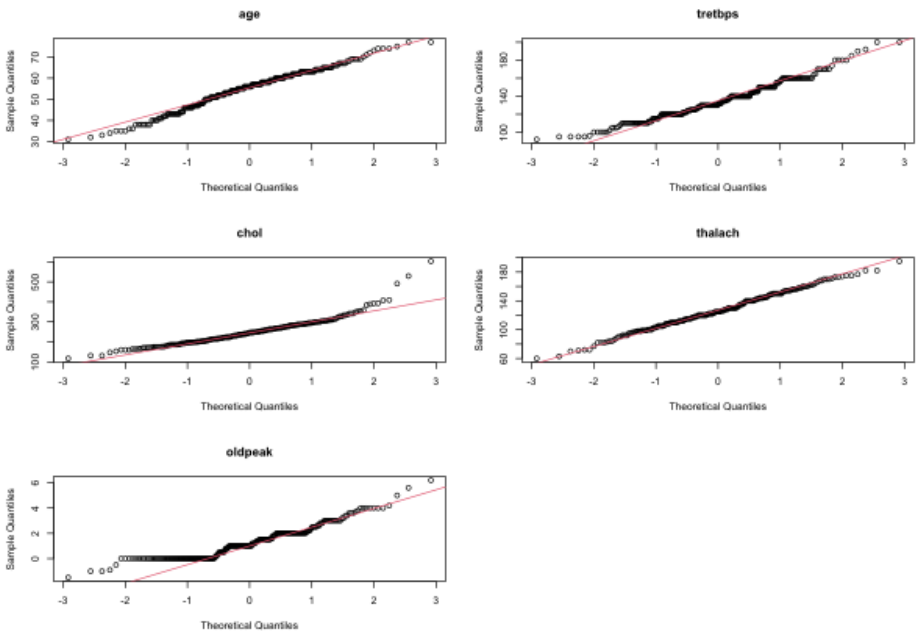

Figure 18: qqplot for patients with heart disease

Based on plot 18 we can see that "thalach" is normally distributed because the dots fall on the red

line. On the other side, distribution of the other variables are not normally distributed because the points deviate from the line. Result of Shapiro test gave us this result and just "thalach" has p-value larger than 0.05, that means it has normal distribution.

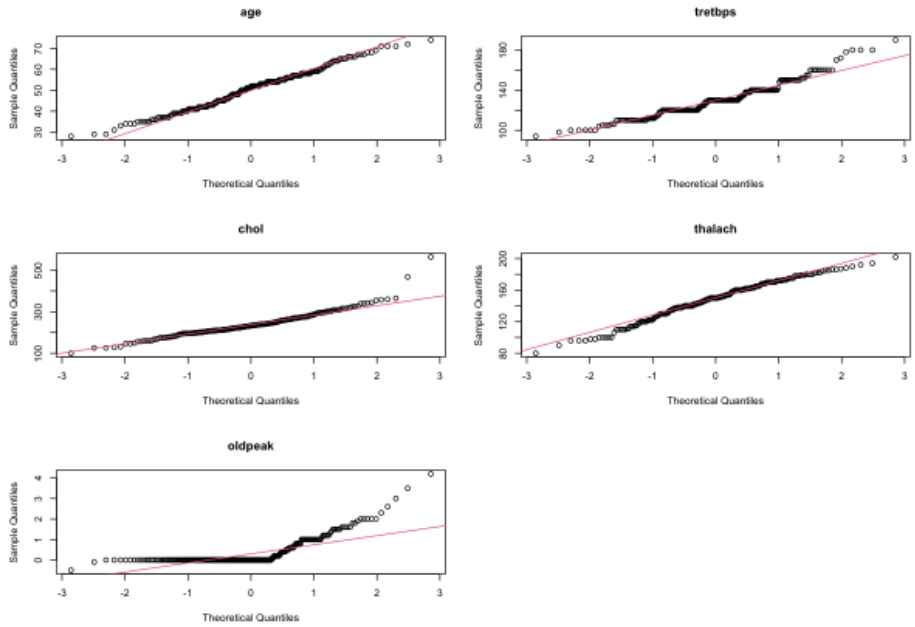

Figure 19: qqplot for patients without heart disease

Based on plot 19 we can see that "age" is normally distributed and has p-value larger than 0.05 in Shapiro test.

To normalize data set we used Box Cox because we have skewed dataset and this package has a better performance in skewed data set. Unfortunately we could not get a normal data after applying basicPower function.

10.2.2 Multivariate Gaussian Conditions

To check this assumption we used Royston test that gives us p-value and chi-square qqplot. Based on plot 20 there are some deviation from the line in both group of patients, additionally result of p-value gave us same result. Therefore, this assumption is violated.

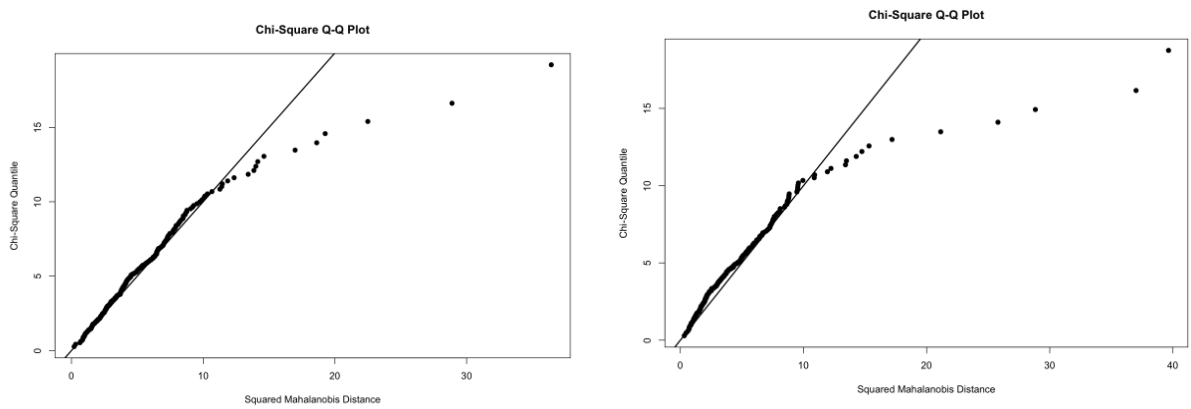

Figure 20: Royston-test for two group of patients

10.2.3 Variance and covariance in groups

In first part of this assumption, we should check that two group of patients have equal variance. First, we used box plot to clearly see changes in variance in our variables and then we use leventest test.

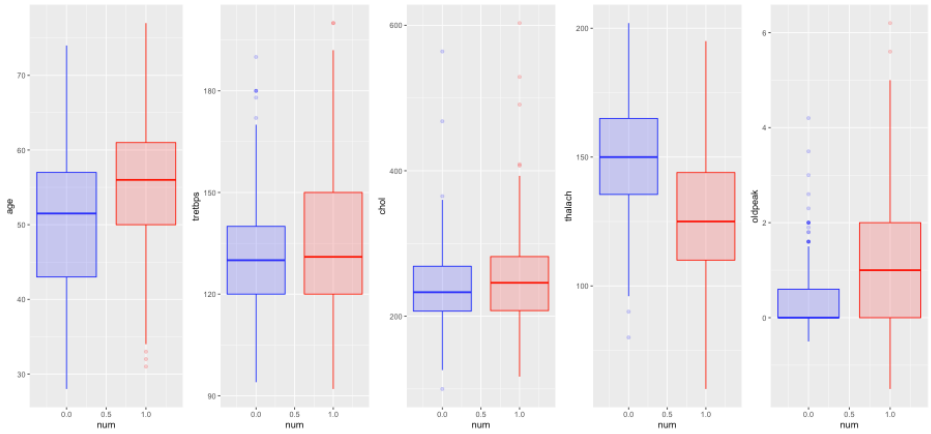

Figure 21: Boxplot of different variables in two groups

As we can see in plot 21 length of plots for "oldpeak" and "tretbps" in 2 groups are different. This is an indication of non-equal variances. To make sure about these 2 variables we used leventest test too. P-value in these two variables are less than 0.05. In second part we used ellipses to see covariance in 2 groups for pair of variables. As we know covariance refers to the measure of how two random variables will change when they are compared to each other. For example in oldpeak and thalach, changes in patients with heart disease is so different with changes in patients without heart diseas. We can see clear relation between variables in plot 22.

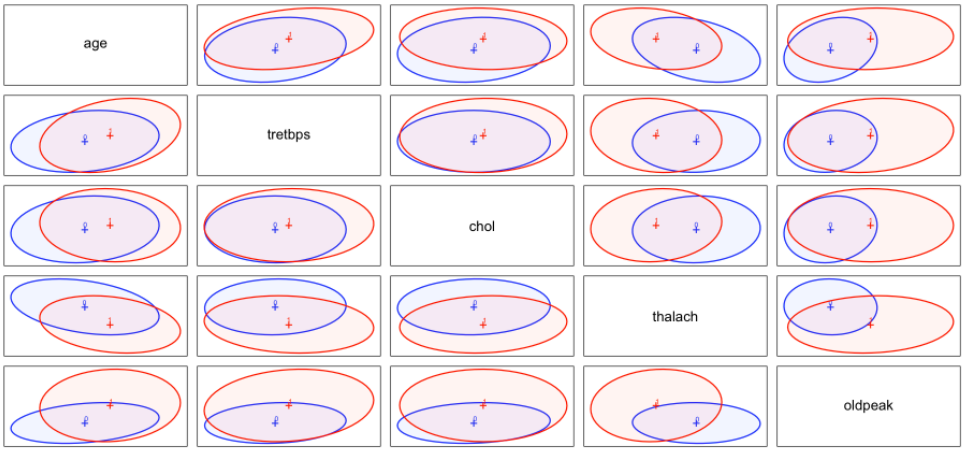

Figure 22: Covariance in each pair of variables

Additionally, we used Box M test because it is more sensitive and can detect even small departures from homogeneity of variance-covariance matrices. Based on result of this test, p-value = 2.2e-16, we can conclude that we have a problem of heterogeneity of variance-covariance matrices in this data set.

Even though, all assumptions are violated, we will applied lda and qda. We train model with real and scaled train data set and make prediction on validation data set. We used these two scenario to see effect of having scaled data in train and validation. In below, we can see confusion matrix of the model that train with real data set:

```
            Reference
Prediction  0  1
         0 76 59
         1  3 36
```

Figure 23: confusion matrix of lda on real data

Based on results in Figure 23, we can say that 59 of patients predicted as patients without heart disease while they have heart disease, that is too risky. On the other side 3 patients predicted as patients with heart disease while they do not have a heart disease. To be more precise in this model in figure 24 we can see accuracy and other metrics such as specificity and sensitivity and our goal is to classify correctly those patients that are at risk of suffering from heart disease.

```
               Accuracy : 0.6437
                 95% CI : (0.5677, 0.7147)
    No Information Rate : 0.546
    P-Value [Acc > NIR] : 0.005675

                  Kappa : 0.3218

 Mcnemar's Test P-Value : 2.848e-12

            Sensitivity : 0.9620
            Specificity : 0.3789
         Pos Pred Value : 0.5630
         Neg Pred Value : 0.9231
             Prevalence : 0.4540
         Detection Rate : 0.4368
   Detection Prevalence : 0.7759
      Balanced Accuracy : 0.6705
```

Figure 24: Result of lda on real data

In the next step, we train lda model with scaled train data set and make prediction on scaled validation data set. For continuous variables we used min max scaling and for categorical variables we used hot encoding.

The main reason of scaling data set before applying some Machine Learning models is that range of some features are so different and large values in these features may affect on training model.

Based on confusion matrix in figure 25 , we can see that number of miss classification in group of patients without heart disease decreased to 13 patients.

```
          Reference
Prediction  0  1
         0 59 13
         1 20 82
```

Figure 25: Confusion matrix of lda on scaled dataset

To compare result of lda in both scenarios, it is better to compare metrics that we described before. In plot 26 we can see that accuracy improves to 81% and although our Sensitivity decreased from 96% to 74%, but Specificity increased from 38% to 86%, that is very good improvement.

```
               Accuracy : 0.8103
                 95% CI : (0.7441, 0.8657)
    No Information Rate : 0.546
    P-Value [Acc > NIR] : 2.27e-13

                  Kappa : 0.6146

 Mcnemar's Test P-Value : 0.2963

            Sensitivity : 0.7468
            Specificity : 0.8632
         Pos Pred Value : 0.8194
         Neg Pred Value : 0.8039
             Prevalence : 0.4540
         Detection Rate : 0.3391
   Detection Prevalence : 0.4138
      Balanced Accuracy : 0.8050
```

Figure 26: Result of lda on scaled dataset

10.3 QDA

Quadratic Discriminant Analysis (QDA)is a variant of LDA in which an individual covariance matrix is estimated for every class of observations. QDA is particularly useful if there is prior knowledge that individual classes exhibit distinct covariances. Therefore, we check first two assumptions and we do not check third assumption in this model. We should keep in mind that LDA is better for small data set and QDA is better for large data set.
Unfortunately, because our data set is not big we could not run this model in our data set.

10.4 Final Results

Regarding the results of the accuracy obtained by all of our predictive models, Random Forest is the algorithm that obtained a better performance beating decision trees and LDA.

We have kept apart a test partition of data that have not been used through the models building nor the validation process. Therefore, we want to asses which is the performance of our best model in this new data to get a generalized error that could approximate the true error of the model.

Using the test partition on the RF model to predict the risk of suffering CVD disease we achieve slightly worse than expected results.

In Figure **??** we can observe that accuracy on new data is $\sim$76% with very similar sensitivity

(∼76%) and specificity (∼77%). This decrease on the performance can be due to the variables used on building the RF model and its distributions in each data partition (train/validation/test). The most important variable in this model is *chest pain*, which is a categorical variable of with 4 categories (1= typical angina, 2= atypical angina, 3= non-anginal pain,4= asymptomatic) where a lot of the patients (mostly diseased) fall in the 4 category being asymptomatic, making the predictions more difficult.

Unfortunately, these results indicate that our RF model is not so good on predicting completely new data.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 61 23
         1 18 72

               Accuracy : 0.7644
                 95% CI : (0.6942, 0.8253)
    No Information Rate : 0.546
    P-Value [Acc > NIR] : 1.951e-09

                  Kappa : 0.5272

 Mcnemar's Test P-Value : 0.5322

            Sensitivity : 0.7579
            Specificity : 0.7722
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.7262
             Prevalence : 0.5460
         Detection Rate : 0.4138
   Detection Prevalence : 0.5172
      Balanced Accuracy : 0.7650

       'Positive' Class : 1
```

Figure 27: Final performance of RF model in test partition data that approximate the true error of this model.

References

[1] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[2] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[3] Peter Filzmoser and Moritz Gschwandtner. *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*, 2021. URL https://CRAN.R-project.org/package=mvoutlier. R package version 2.1.1.

[4] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.4.0.

[5] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020. URL https://CRAN.R-project.org/package=factoextra. R package version 1.0.7.

[6] Max Kuhn. *caret: Classification and Regression Training*, 2021. URL https://CRAN.R-project.org/package=caret. R package version 6.0-90.

[7] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008. doi: 10.18637/jss.v025.i01.

[8] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

[9] Michael Mayer. *splitTools: Tools for Data Splitting*, 2020. URL https://CRAN.R-project.org/package=splitTools. R package version 0.3.1.

[10] Stephen Milborrow. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*, 2021. URL https://CRAN.R-project.org/package=rpart.plot. R package version 3.1.0.

[11] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. URL https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-2.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[13] Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. URL https://CRAN.R-project.org/package=GGally. R package version 2.1.2.

[14] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005. URL http://rocr.bioinf.mpi-sb.mpg.de.

[15] Daniel J. Stekhoven and Peter Buehlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[16] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-15.

[17] Nicholas Tierney, Di Cook, Miles McBain, and Colin Fay. *naniar: Data Structures, Summaries, and Visualisations for Missing Data*, 2021. URL https://CRAN.R-project.org/package=naniar. R package version 0.6.1.

[18] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

[19] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[20] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

A Additional Tables

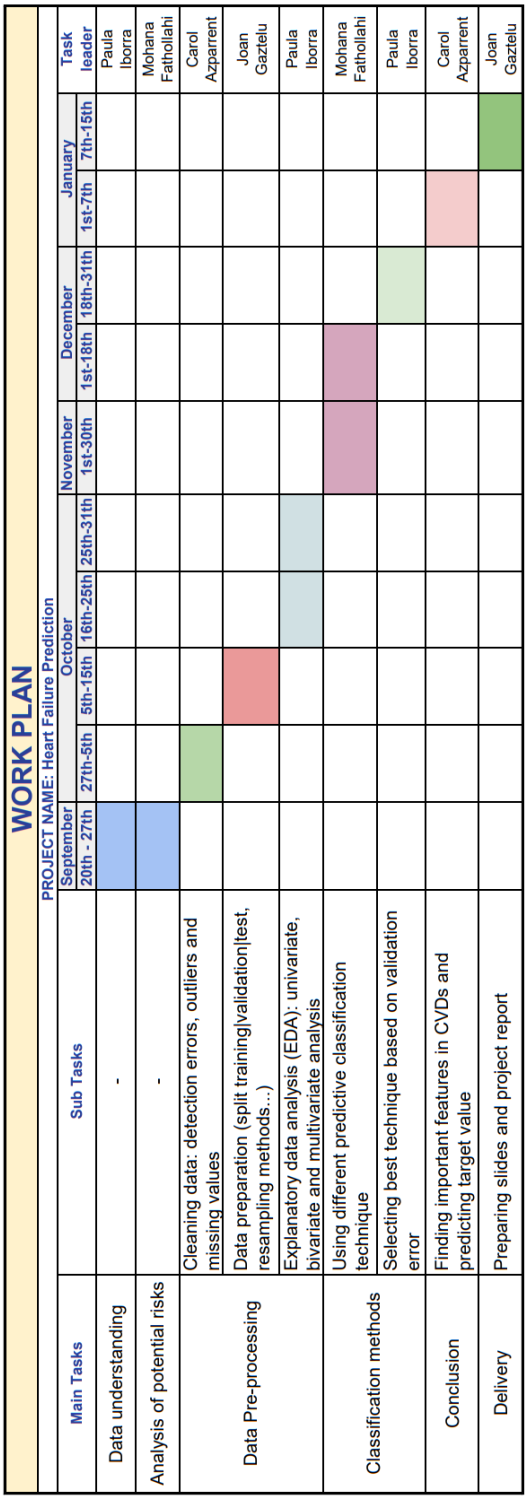

Figure A.1: Project work plan and assignment of tasks

B Additional Figures

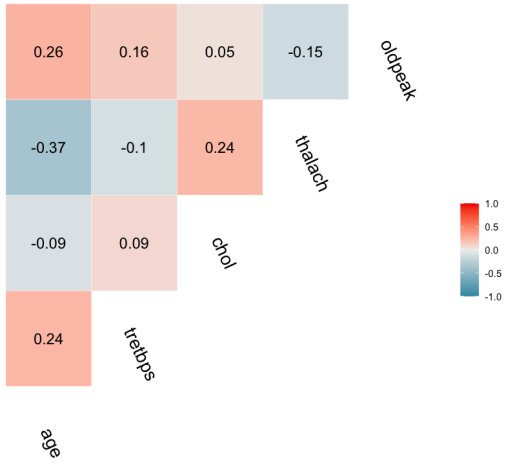

Figure B.1: Pearson correlation between all continuous variables.

$stats
92 120 130 140 170

$n
861

$conf
128.923074880986 131.076925119014

$out
172 180 200 174 178 192 180 178 180 80 180 200 185 200 180 0 178 172 180
190 190 180 180 180 200 180 180 180

Figure B.2: Stats of Tretbps variable before the changes

$stats
92 120 130 140 170

$n
860

$conf
128.922448943637 131.077551056363

$out
172 180 200 174 178 192 180 178 180 80 180 200 185 200 180 178 172 180
190 190 180 180 180 200 180 180 180

Figure B.3: Stats of Tretbps variable after the changes

$stats
85 175.5 224 268 404

$n
839
$conf
218.954341662656 229.045658337344

$out
417 407 564 409 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 458 0 0 0
0 468 518 412 529 466 603 491

Figure B.4: Stats of Chol variable before the changes

$stats
117 210 240 276 365

$n
677
$conf
235.992194027286 244.007805972714

$out
417 407 564 394 409 100 458 384 468 518 412 529 100 85 392 466 393 388
603 404 491 394

Figure B.5: Stats of Chol variable after the changes

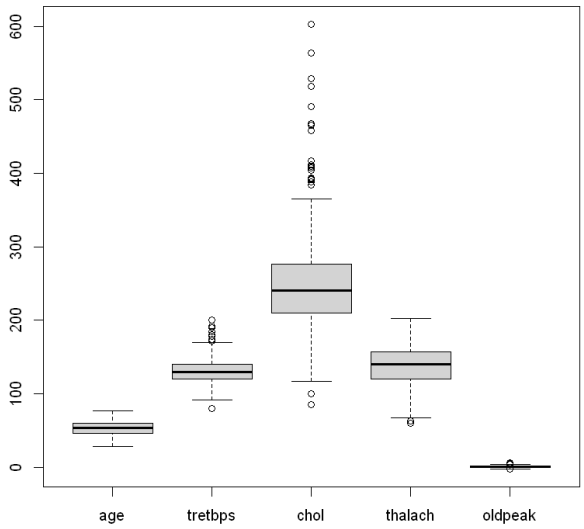

Figure B.6: Boxplot of numeric variables after changing zeros

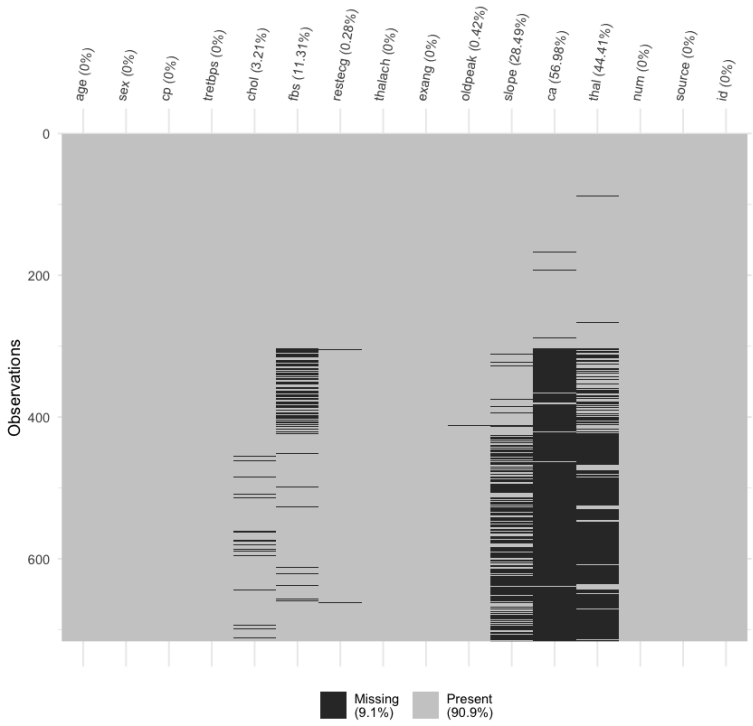

Figure B.7: Missing values with *va* source removed

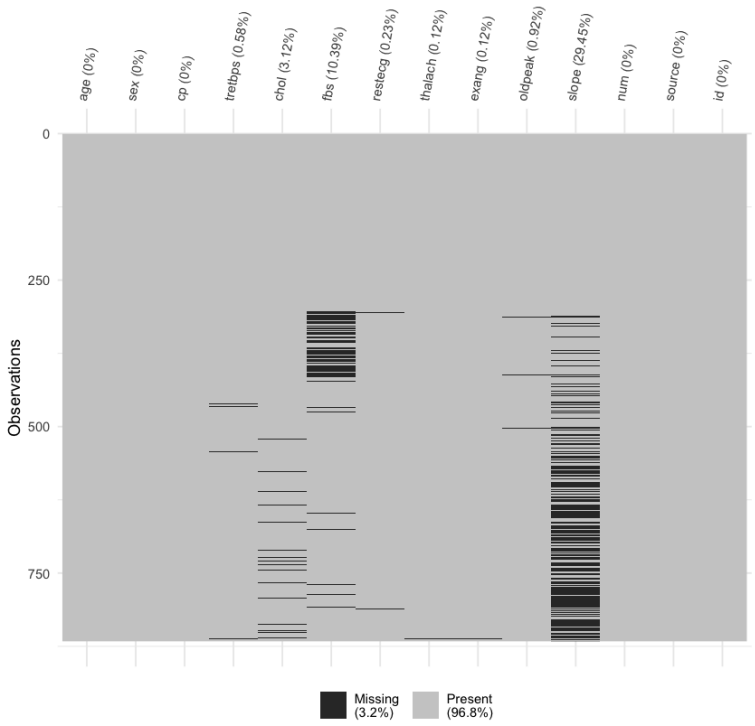

Figure B.8: Missing values with *ca* and *thal* variables removed.

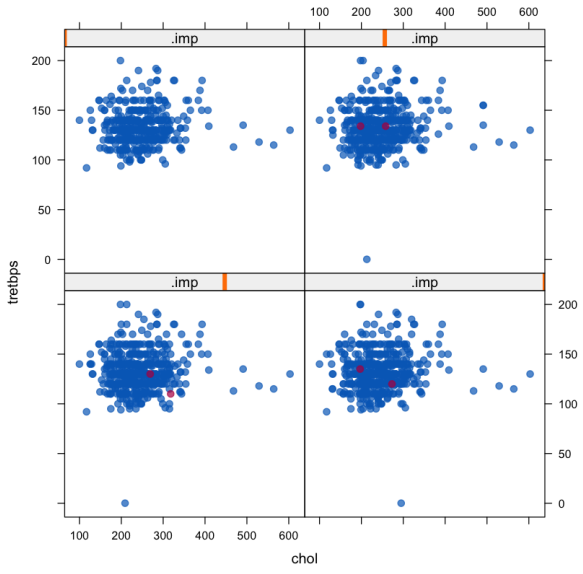

Figure B.9: plot for *chol* and *tretbps*.

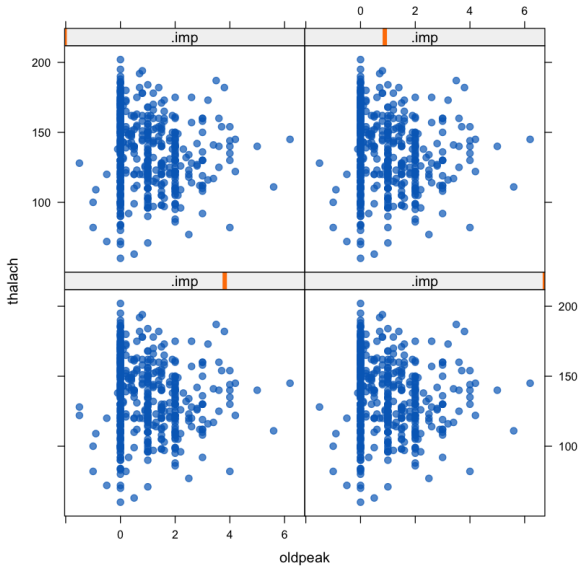

Figure B.10: plot for *thalach* and *oldpeak*.

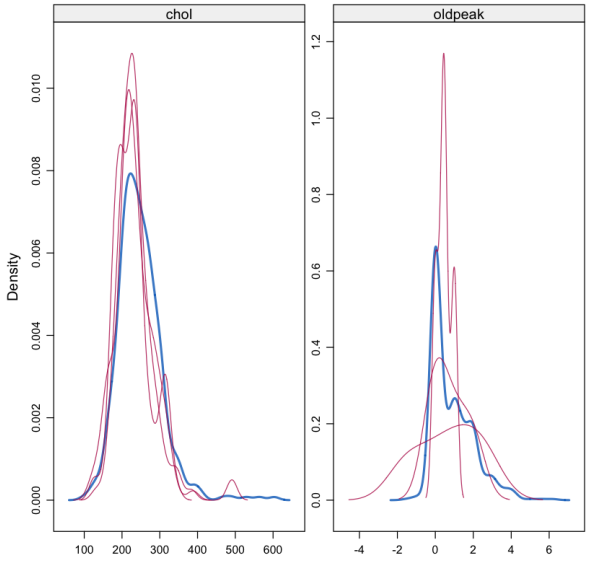

Figure B.11: Density plot for *chol* and *Oldpeak*.

	age	tretbps	chol	thalach	oldpeak
age	1.0000000	0.24270354	-0.11586479	-0.36861008	0.26626098
tretbps	0.2427035	1.00000000	0.08107652	-0.09280759	0.17681540
chol	-0.1158648	0.08107652	1.00000000	0.23341656	0.05527817
thalach	-0.3686101	-0.09280759	0.23341656	1.00000000	-0.15560212
oldpeak	0.2662610	0.17681540	0.05527817	-0.15560212	1.00000000

Figure B.12: Correlation values.

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Variance	1.697	1.193	0.822	0.700	0.588
% of var.	33.942	23.858	16.435	14.006	11.758
Cumulative % of var.	33.942	57.800	74.236	88.242	100.000

Figure B.13: Dimensions summary.

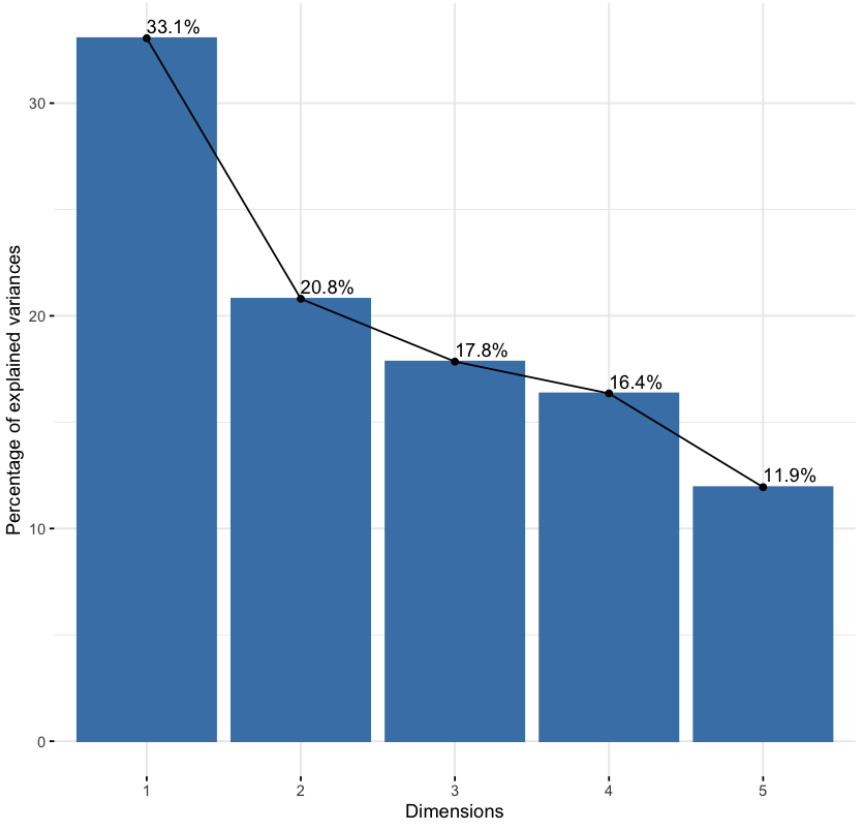

Figure B.14: Eigenvalues plot.

	correlation	p.value
age	0.7860630	1.867266e-109
oldpeak	0.5541881	7.064833e-43
tretbps	0.4756264	1.740898e-30
chol	-0.2626163	1.381041e-09
thalach	-0.6905893	2.395933e-74

Figure B.15: Dimension 1 test

	correlation	p.value
chol	0.7889592	8.494924e-111
tretbps	0.5176465	1.036849e-36
oldpeak	0.3983590	4.522052e-21
thalach	0.3791981	4.320040e-19

Figure B.16: Dimension 2 test

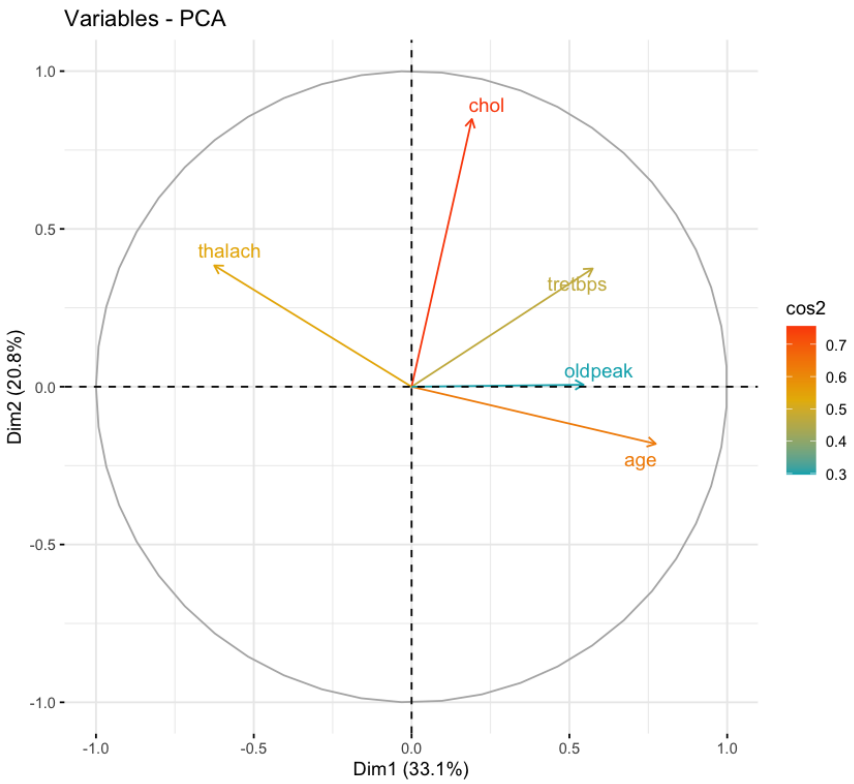

Figure B.17: Circular plot by quality

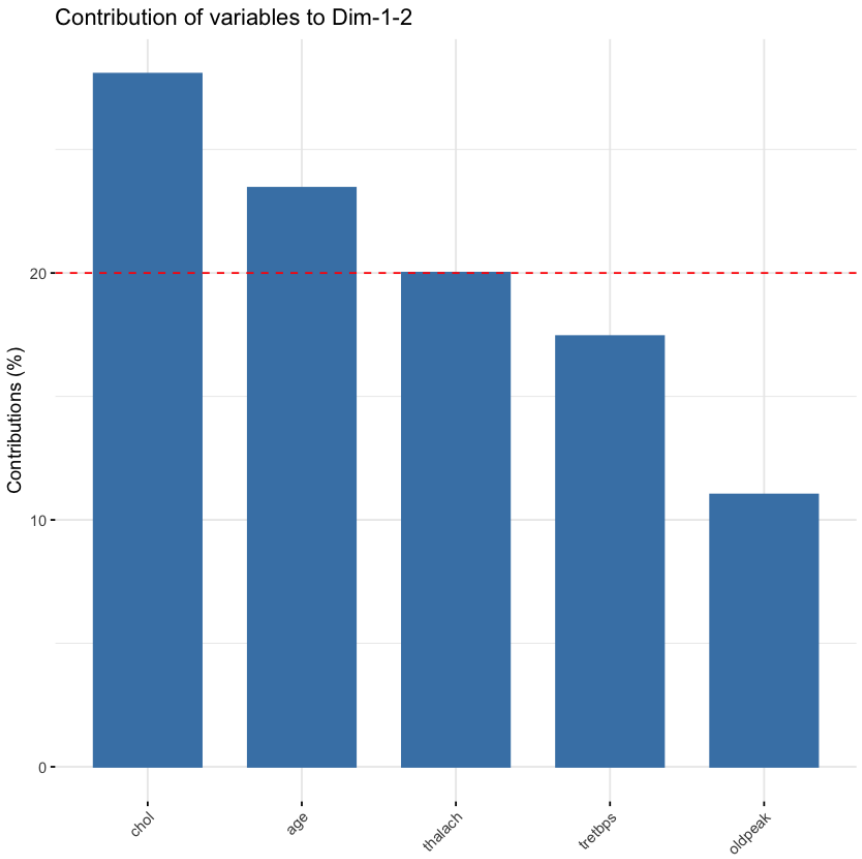

Figure B.18: Contribution of variables to first and second dimensions.

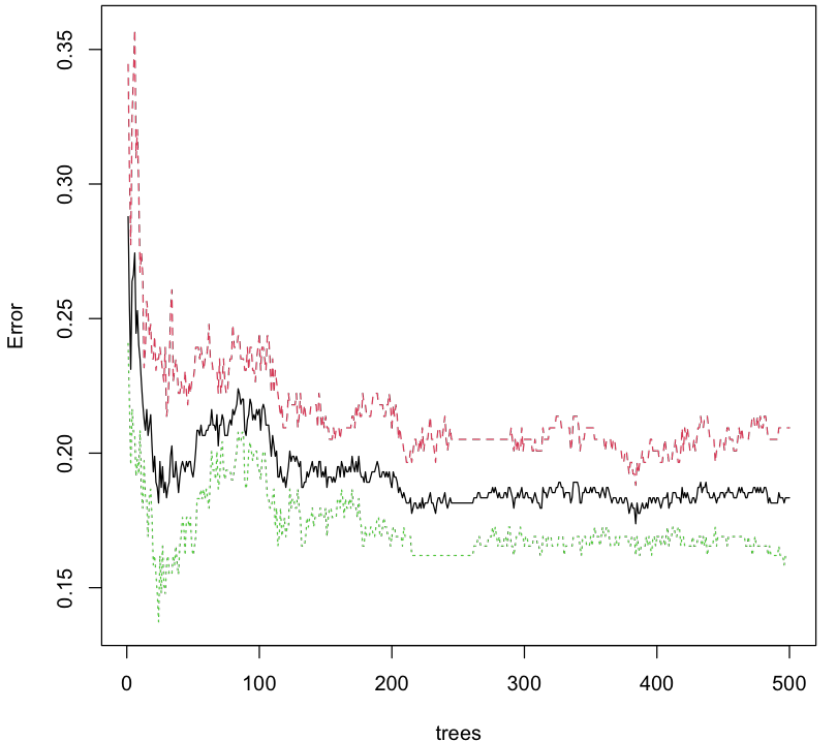

Figure B.19: Number of tree vs error.

	sex	cp	fbs	restecg	exang	slope	num	source
sex	1.00000000	0.13841894	0.06615162	0.01308489	0.15716633	0.09479835	0.2394171	0.26776070
cp	0.13841894	1.00000000	-0.02000949	0.05742827	0.41994810	0.19691010	0.4141348	0.20020414
fbs	0.06615162	-0.02000949	1.00000000	0.16593917	0.02240523	0.08417098	0.1544816	0.13947265
restecg	0.01308489	0.05742827	0.16593917	1.00000000	0.07883434	0.07929706	0.1180162	-0.09616149
exang	0.15716633	0.41994810	0.02240523	0.07883434	1.00000000	0.30010243	0.4394460	0.24152427
slope	0.09479835	0.19691010	0.08417098	0.07929706	0.30010243	1.00000000	0.3679893	0.16638256
num	0.23941710	0.41413478	0.15448158	0.11801621	0.43944604	0.36798931	1.0000000	0.25658716
source	0.26776070	0.20020414	0.13947265	-0.09616149	0.24152427	0.16638256	0.2565872	1.00000000

Figure B.20: Correlations categorical data.

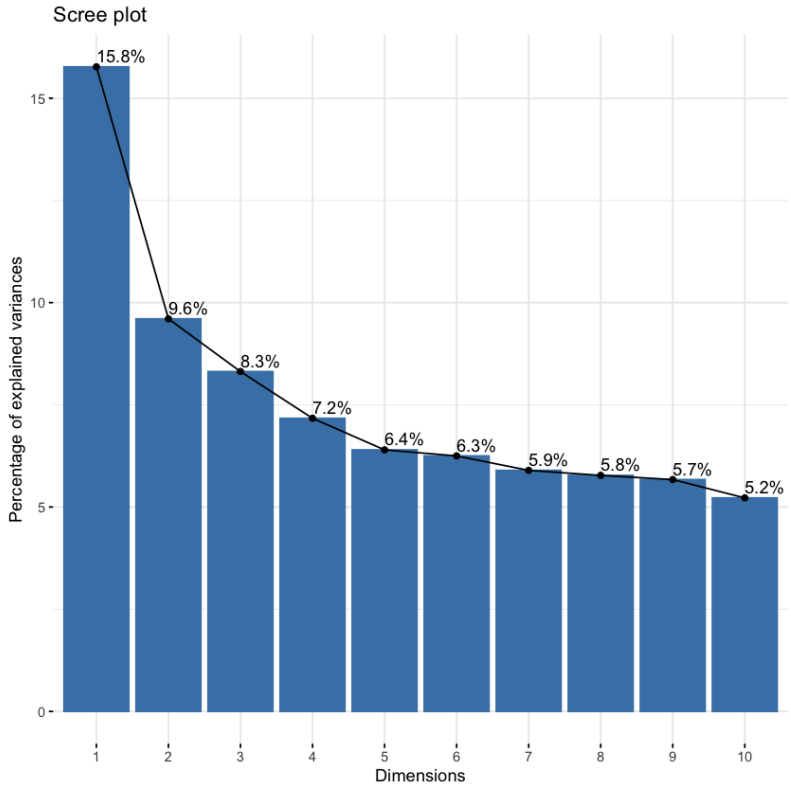

Figure B.21: Variance retained by each dimension.

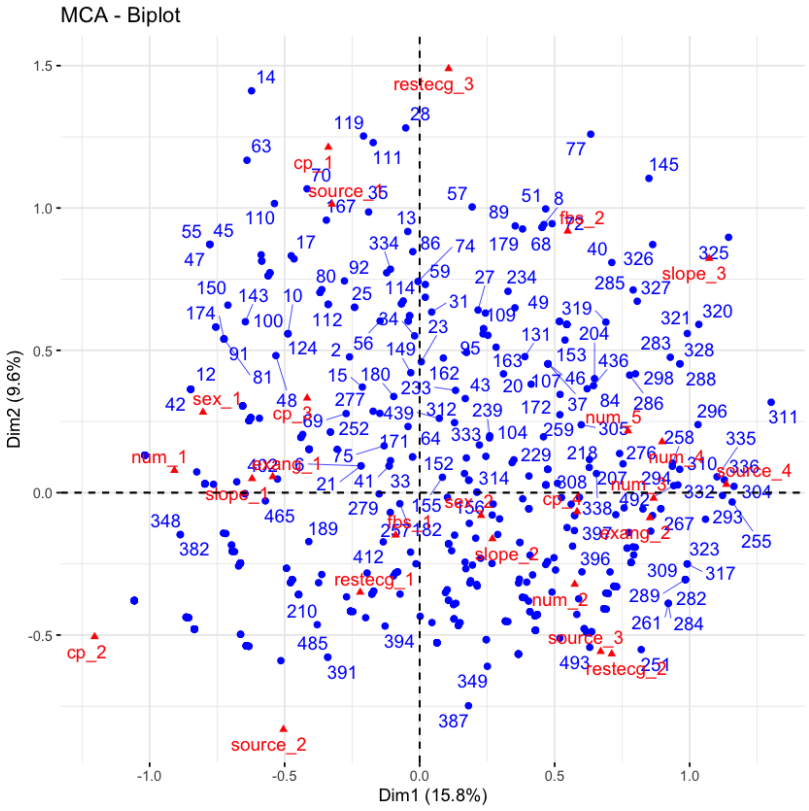

Figure B.22: Biplot distances.

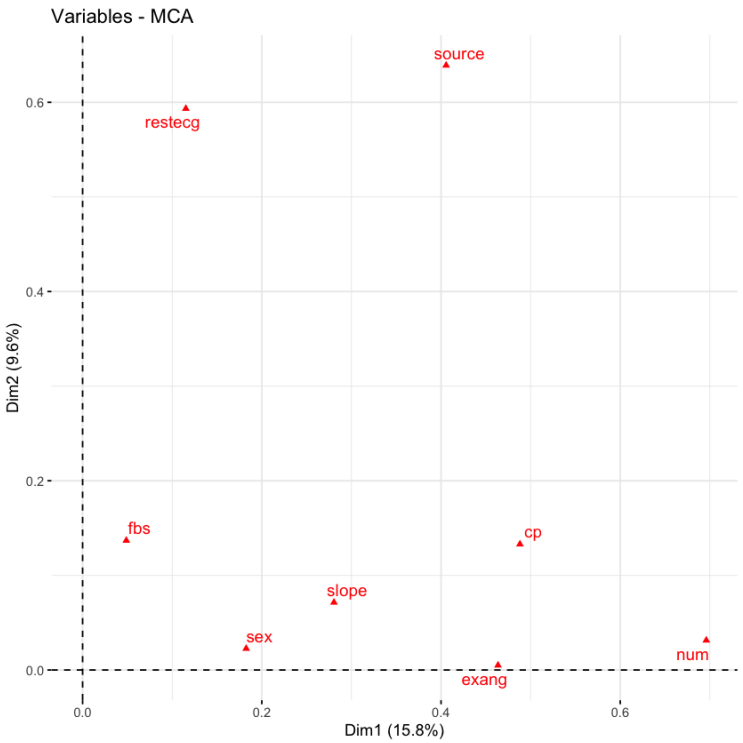

Figure B.23: Individuals correlation.

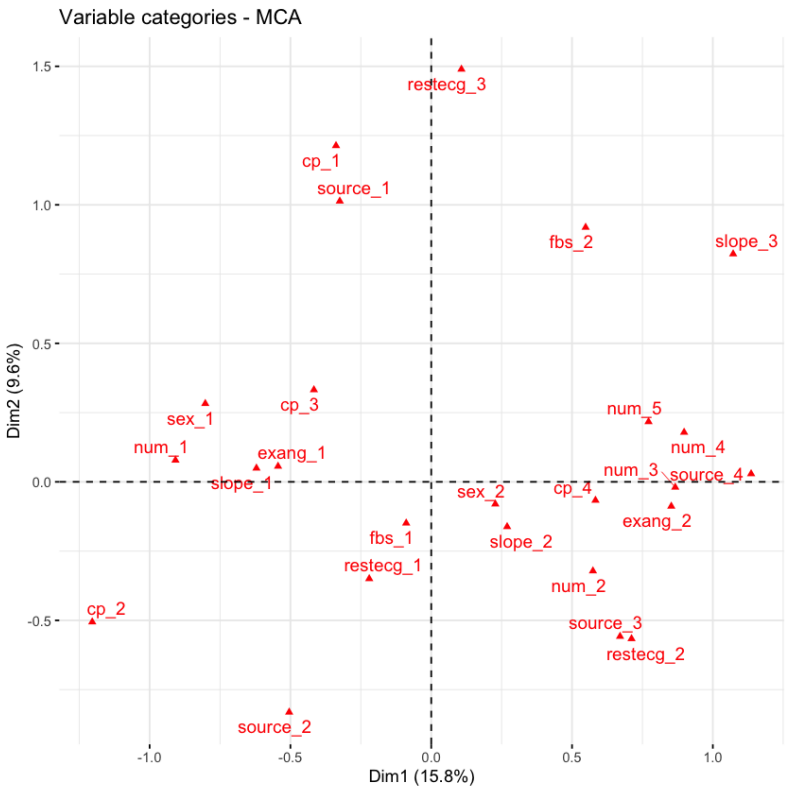

Figure B.24: Categories correlation.

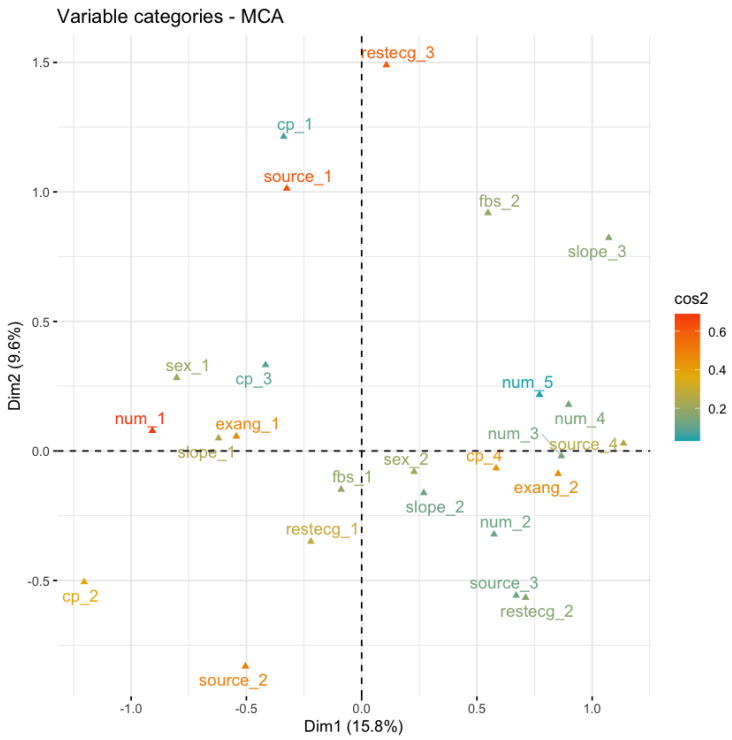

Figure B.25: Cos2 values.

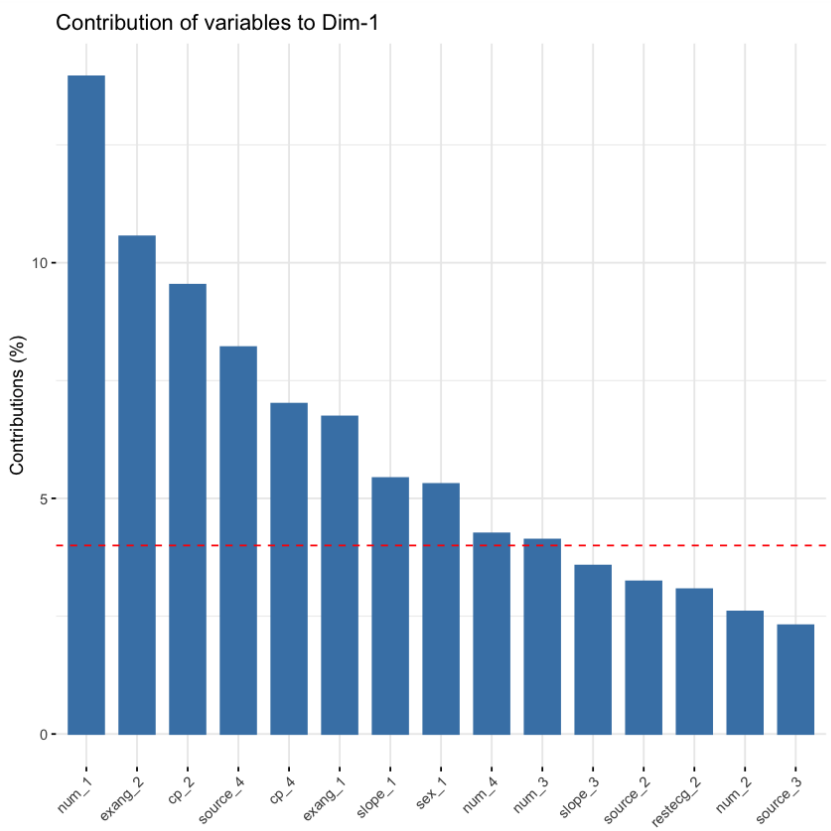

Figure B.26: Cos2 of variables in dimensions 1.

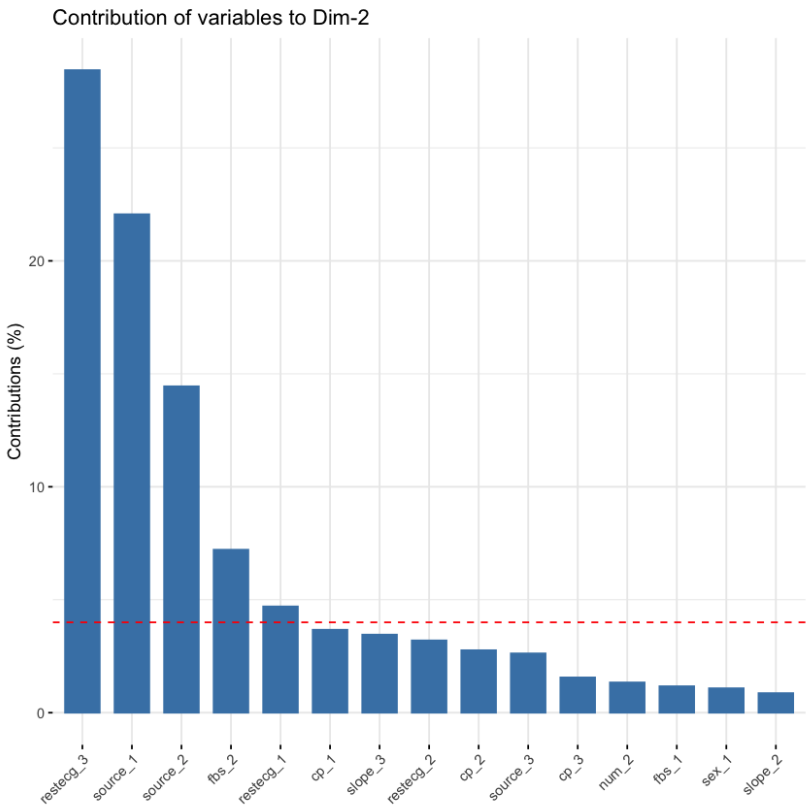

Figure B.27: Cos2 of variables in dimension 2.

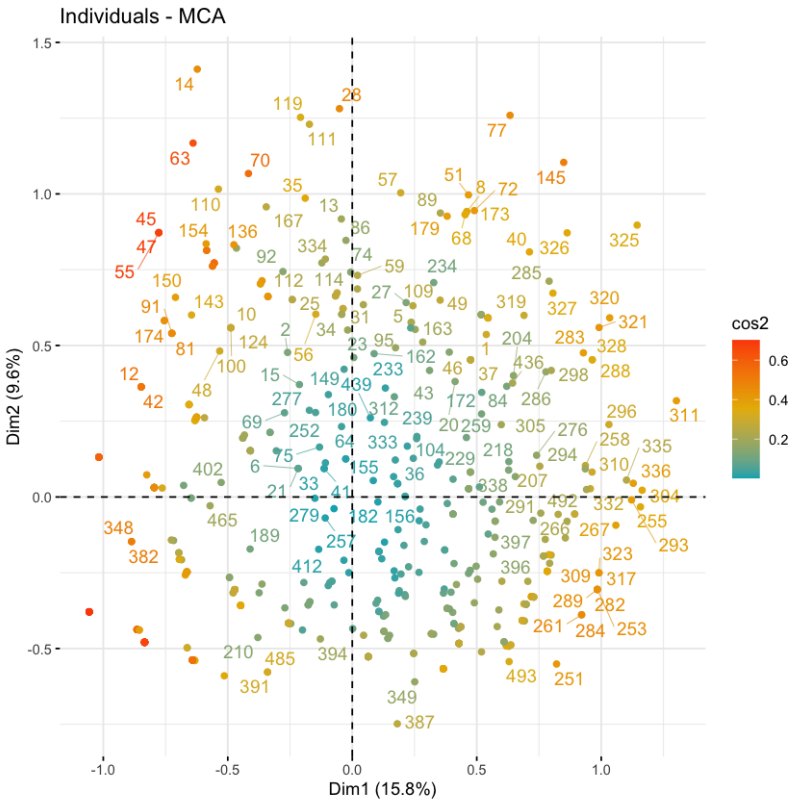

Figure B.28: Scatter plot of most important individuals.