Heart Disease
Machine Learning 1

Jose Pérez Cano and Álvaro Ribot Barrado

GCED – FIB – UPC

April 26, 2020

En aquesta entrega requerim un informe del que heu fet amb el dataset que heu escollit, això inclou: preprocés, exploració inicial de dades (visualitzacions senzilles) i modelització inicial amb models vists a classe. Expliqueu bé la metodologia de selecció de models (resampling etc.) per tal d'evitar disgustos de cara a la entrega final. Gràcies.

1 Preprocessing

1.1 Format

The data provided by the UCI repository is not in a nice format to be read directly using R. So we have done a previous preprocessing step converting .data, where each patient was split into several lines, to .csv format, in which the 75 attributes of each patient were altogether in one line, using C++. The repository consists of different datasets, corresponding to different locations but all of them with the same variables (75). This databases are called Cleveland, Hungarian, Switzerland and Long Beach VA. We have begun our study using only the Cleveland database, since it is the one other researchers used. The target variable in this project is $V58$, corresponding to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4 and we will deal with it as a factor.

1.2 Missing values

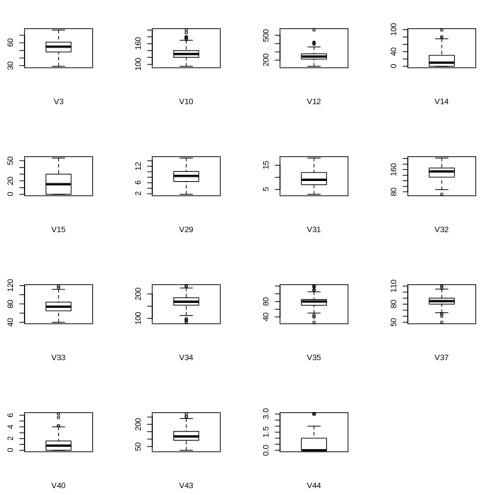
There are 20 variables with all missing values, so we have deleted them from our Data Set. We have deleted two more variables, one with more than 90% missing values that we also have deleted, and another one with 69 missing values, corresponding to the time where ST measure depression was noted.

Missing values are encoded with -9. Before encoding them with NA we have imputed the remaining missing values using the k nearest neighbour technique, using $k = 7$.

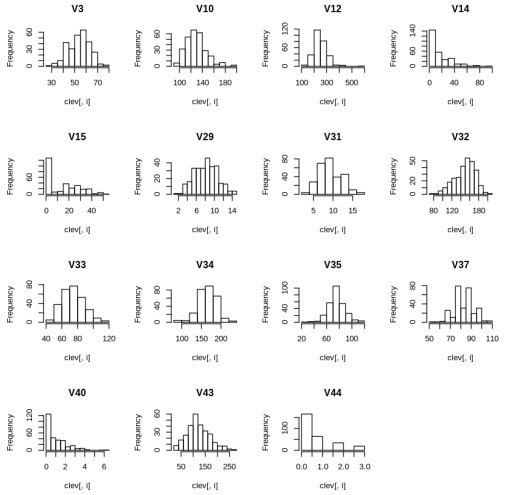
After doing this, we have changed several variables types from integer to factor, according to the dataset description. Also, we've removed some dummy variables which had only one value for all. The number of remaining variables is 45, from the initial 75.

1.3 Data visualization

Initially, we start by doing boxplots and histograms of the different variables to see possible outliers and if the distributions are more or less symmetric. This is the result, it contains symmetric and skewed variables.



(a) Boxplots of numeric variables



(b) Histograms of numeric variables

Figure 1: Visualization

1.4 Correlation of variables

The dataset has many variables which are related between them. This insight will be helpful in a possible reduction of variables later on. After looking their description every correlation found seems reasonable. The more correlated pair was the duration of a test exercise and a MET score in a test, with a correlation of 0,93. MET means Metabolic Equivalent of Task, which is simply a measure of how much exercise is done, as is the duration, thus the relation. There are other pairs of moderately correlated variables ($0,5 < cor < 0,7$), like cigarettes per day and years as a smoker. Also peak exercise blood pressure and resting blood pressure, and maximum heart rate achieved and METs achieved.

1.5 Modification of values

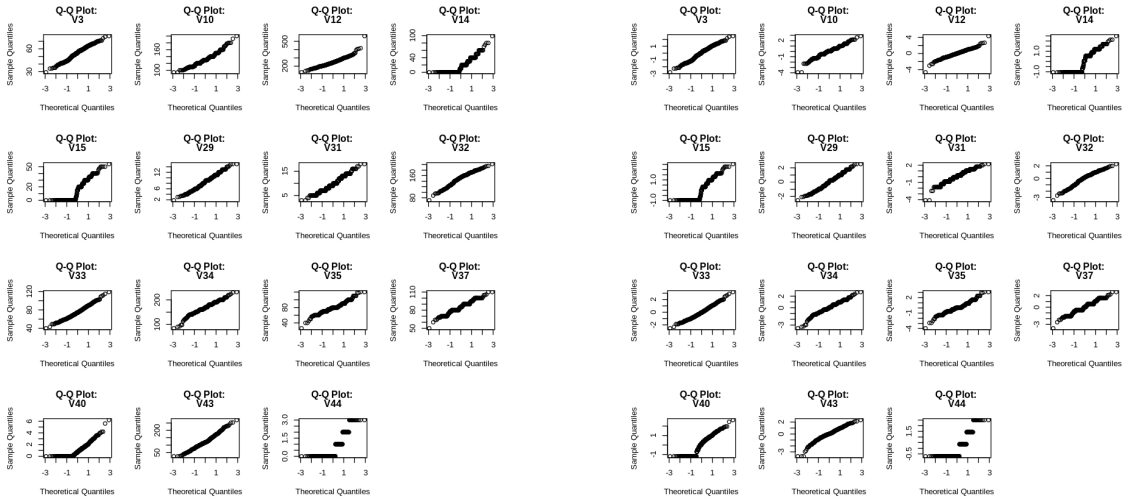
As it was seen previously many variables aren't normal, and most of our models will rely partly on this assumption. Therefore we are going to apply a Box-Cox transformation to approximate the distribution to a Gaussian as much as possible. Furthermore, data is standardised in order to obtain better results, since the ranges of the variables vary substantially across variables.

Some of the variables are already normal and in the Box-Cox the estimated lambda was statistically 1, or the Q-Q plot was a straight line. These were the age of the patient, the maximum heart rate, the peak blood pressure exercise, the resting blood pressure and the duration of the exercise. The estimated transformation for the other variables is a square root transformation.

However, there are three special variables which are years as smoker, cigarettes per day and ST depression induced by exercise relative to rest. They have two groups, one with all zeros, and the other with a variable quantity. We have decided to transform this variables considering the zeros

apart, this way, the estimated transformation to cigarettes per day and ST depression is a square root, and years as a smoker is normal if we ignore the zeros.

Since there are non-positive values we decided to subtract the minimum plus some epsilon to be able to use Box-Cox. Another approach could have been to use the Manly transformation which is exponential and it allows to use non-positive values.



(a) Before (b) After

Figure 2: Q-Q plots of numeric variables

1.6 Feature extraction

1.6.1 PCA

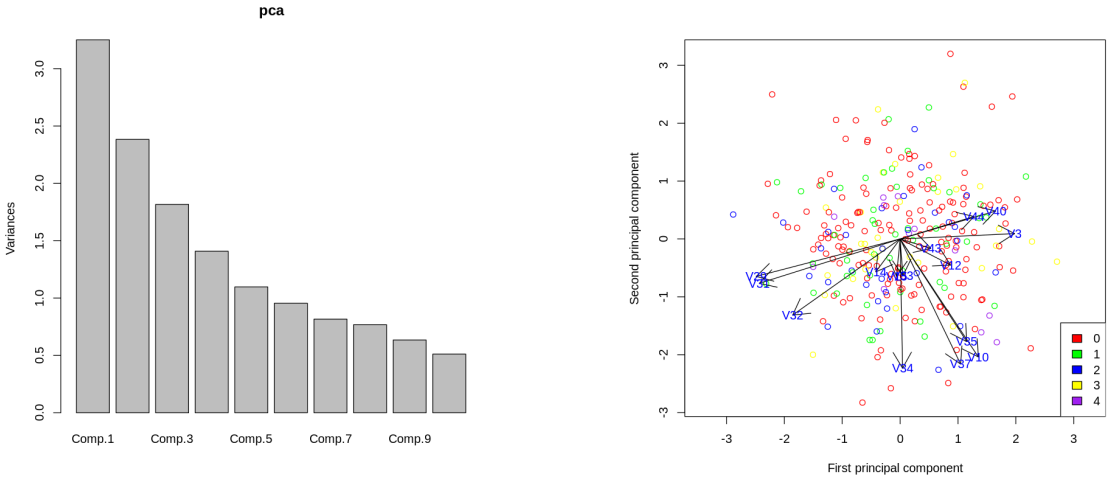
After doing the PCA we have obtained the following screeplot of the variance explained by the principal components. With 2 components we can explain 38% of the total variance, and if we want to explain more than 80% we would need 8 components. The Figure 3 represents the screeplot, where we can see the variance explained by each principal components. We have also computed a biplot to visualize the data in two dimensions. As we expected, it is quite difficult the interpretation of this biplot and we cannot distinguish the different levels of the target variable. Because of that we will not use the PCA in our models.

1.6.2 FDA

To compute the Linear Discriminant Analysis we have removed three problematic variables from the dataset. This variables are

- $V1$: the patient ID,
- $V57$: year of cardiac (which can be expressed as a linear combination of the other variables),
- $V59$: left main coronary trunk. It takes only two values. 1 is taken by 270 patients and 2 by only 12 patients. So it cannot be used to discriminate the data propoerly.

In Figure 4 we can see the plot obtained after FDA. We can clearly distinguish different levels of the target variables (each one is represented by one colour). We also can note that the variables obtained seem to follow a multivariate normal distribution.



(a) Screeplot (b) Biplot

Figure 3: PCA plots

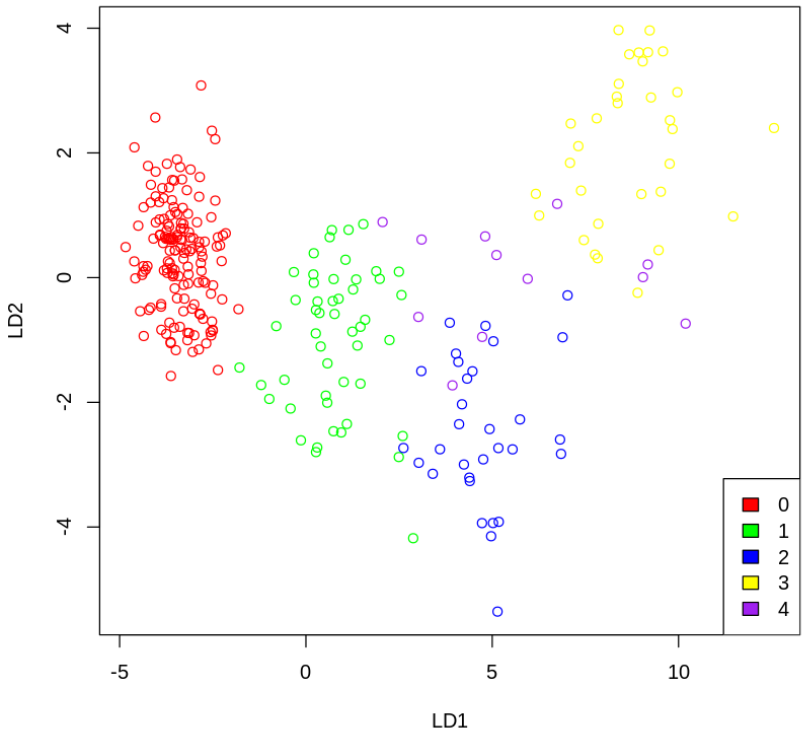


Figure 4: FDA plot

1.6.3 MCA

We have studied the relation between factor variables performing a Multivariate Corresponding Analyis. But as we can see in Figure 5, it is not very helpful.

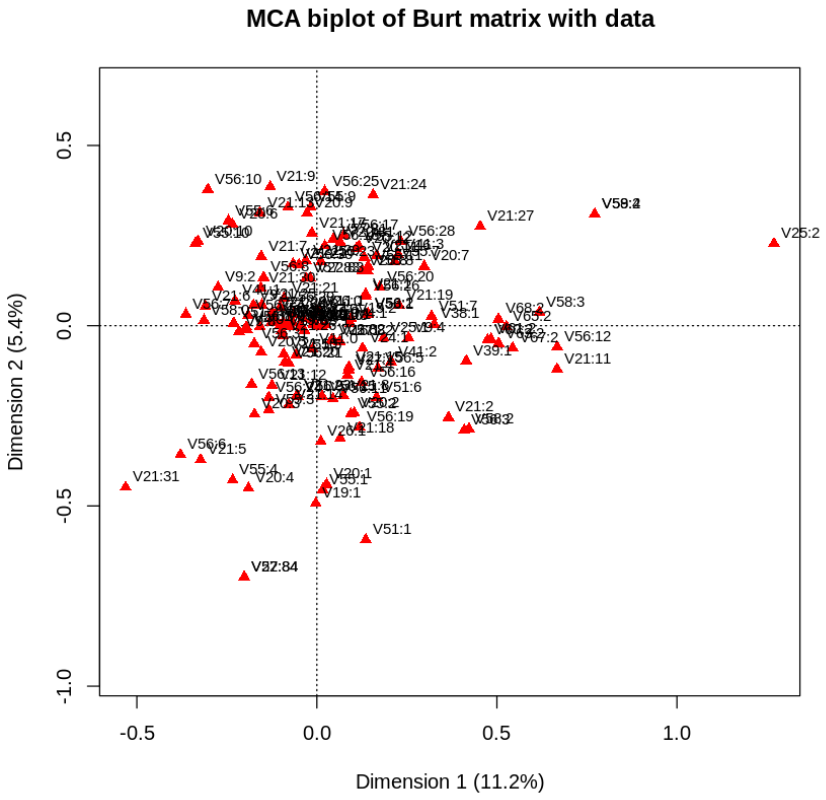


Figure 5: MCA biplot

2 Initial model