Statistics and Data Analysis first project report

Maria Lavoura (up201908426)
Nuno Gomes (up199300242)

April 22, 2020

1 Introduction

The aim of this project was to apply all the regression and classification methods studied in the classes of Statistics and Data Analysis to a data set composed of quantitative and qualitative variables, and several observations. This report accounts for the analysis performed on the data, and on the results obtained from it.

In section 2, an exploratory data analysis (EDA) is made, highlighting the main characteristics of the data set. In section 3, we describe the application of the regression models to the data set, and the performance of all models is compared. Finally, in section 4, we present our conclusions and prospects for future work.

2 Exploratory Data Analysis

2.1 Data set description

The goal of this project was to determine the presence or absence of cardiovascular disease (CVD) in patients from eleven features, five quantitative—*age*, *height*, *weight*, *systolic pressure (aphi)*, and *diastolic pressure* (*aplo*)—and six qualitative—*gender*, *cholesterol* (*choles*), *glucose* (*gluc*), *smoking* (*smoke*), *alcohol intake* (*alco*), and *physical activity* (*active*)— and 70 000 instances. All data set values were collected at the moment of a medical examination. There are three types of input features: (i) *objective*, with factual information; (ii) *examination*, containing results of a medical examination; and (iii) *subjective*, corresponding to information given by the patient. Table 1 presents the features description.

2.2 Data analysis

Figure 10 illustrates a matrix of distributions, correlations, and graphical relations between all variables of the original data set. This type of representation was useful to get a glimpse on the full data set and the relations between the variables within, on the distributions of all the features, and to identify potential outliers. While the skewnesses can be spotted from the representations of the distributions of experimental values (either by densities or histograms) in the main diagonal of the matrix, the graphical interrelations between variables (such as

Table 1: Description of the features. The variables *age*, *height*, *weight*, *aphi*, and *aplo* are quantitative, while *gender*, *choles*, *gluc*, *smoke*, *alco*, and *active* are qualitative and treated as *factors* in the data processing. The target variable is *cardio*.

Description	**Input type**	**Name**	**Type**
Age	Objective	age	int (days, converted to years)
Gender	Objective	gender	1: women 2: men
Height	Objective	height	int (cm, converted to m)
Weight	Objective	weight	float (kg)
Systolic blood pressure	Examination	aphi	int
Diastolic blood pressure	Examination	aplo	int
Cholesterol	Examination	choles	1: normal 2: above normal 3: well above normal
Glucose	Examination	gluc	1: normal 2: above normal 3: well above normal
Smoking	Subjective	smoke	binary
Alcohol intake	Subjective	alco	binary
Physical activity	Subjective	active	binary
Cardiovascular disease	Target	cardio	binary

boxplots and scatter plots) allowed us to identify possible outliers. A careful analysis of each variable was carried in order to infer the correlation between each of them and the target variable, and to identify and remove possible outliers.

2.2.1 Quantitative variables

The *ages* of the patients vary between 29.56 and 64.92 years, with a mean value of 53.30 and a median equal to 53.95. The variance is equal to 45.63, and the skewness −0.3070 (*i.e.*, the distribution is *left-skewed*, with a longer tail on the left-hand side). The histogram and the boxplot of the sample distribution of the age are illustrated in fig. 1. The outliers identified in the boxplot were not disregarded on a first examination, since the corresponding ages are reasonable in the context of the experiment.

Regarding the *height* and the *weight*, both variables presented several outliers in their sample distributions. In the case of the former, there were too many values below 1.5 m and some around 2.5 m. Given the age range of the patients, those values were considered unlikely and were removed. For the weight, having in mind the age range of the patients, values below 50 kg and above 150 kg were also considered as outliers (*conf.* figs. 2 and 3).

The systolic and diastolic variables presented several outliers, with values of some thousands or a few units of mmHg, respectively well above and below normal values for human beings. Those values were removed before the application of the regression models (*conf.* figs. 4 and 5).

When removing the outliers of the variables, their skew were reduced and their distributions approximated better the normal distribution.

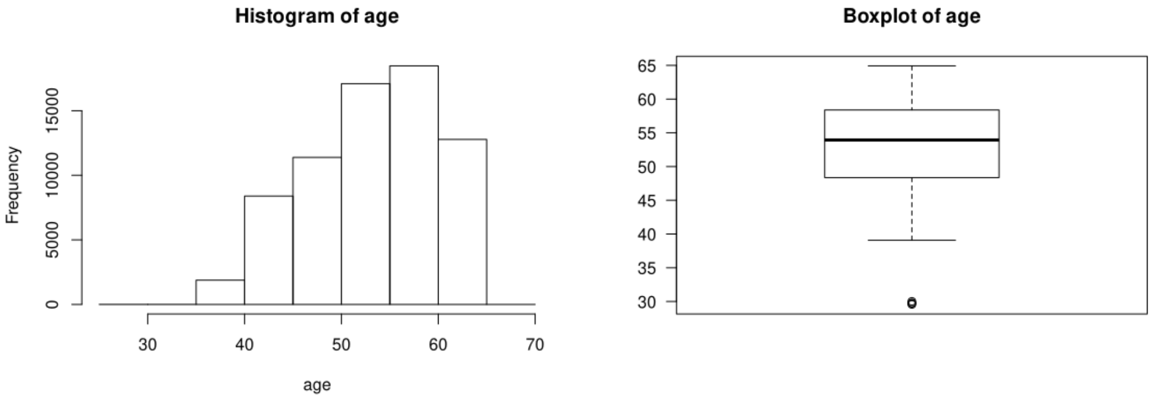


Figure 1: Histogram and boxplot of the sample distribution of the *age* variable.

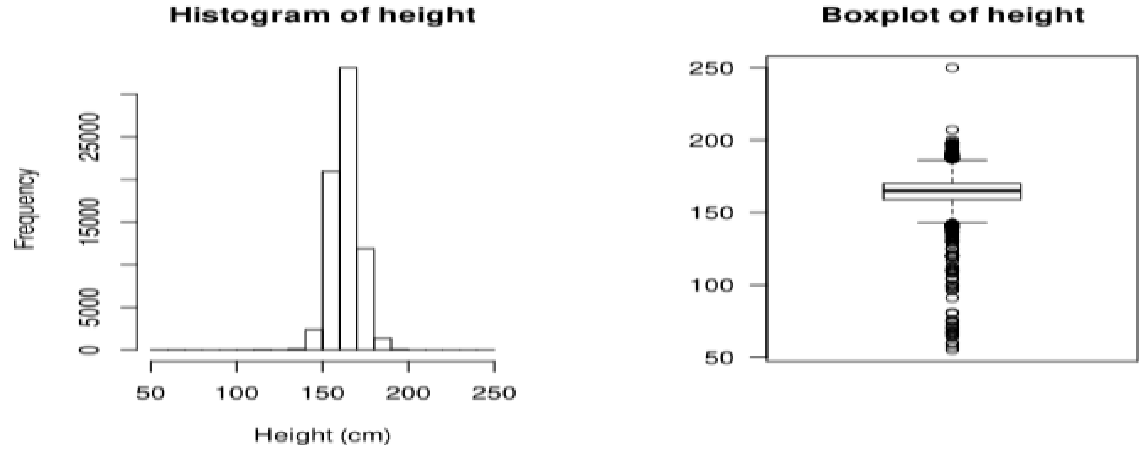


Figure 2: Histogram and boxplot of the sample distribution of the *height*.

2.2.2 Qualitative variables

Regarding the gender, the sample contains about 65 % of women and 35 % of men (fig. 6a). Care was taken to verify the correlation between the gender and the existence/absence of cardio-vascular disease (CVD). No apparent correlation exists between the variables *gender* and *cardio* (see fig. 6b).

Similar analysis was made for the variables *cholesterol*, *glucose*, *smoking*, *alcohol intake*, and *physical activity*. Concerning the former, there is a strong evidence for the increase of the cholesterol levels and the incidence of CVD (fig. 7); there is also a noticeable correlation between the levels of glucose and the existence or absence of CVD (fig. 8); the correlation between CVD and the other three qualitative variables is less notorious, being practically absent in the alcohol intake (see figs. 9a to 9c).

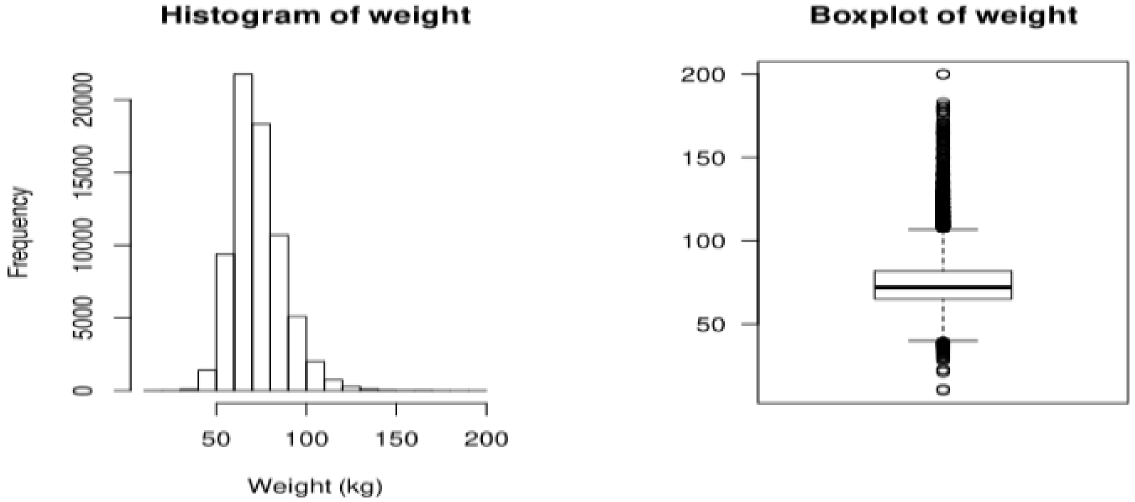


Figure 3: Histogram and boxplot of the sample distribution of the *weight*.

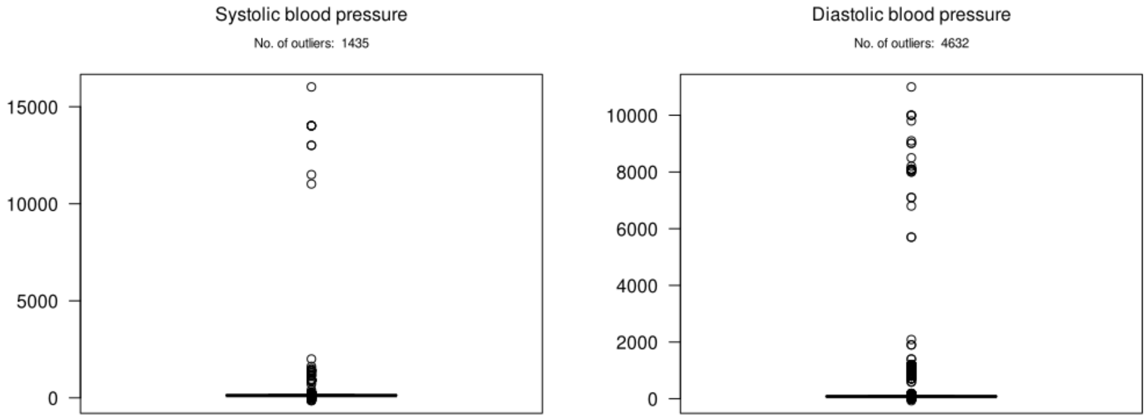


Figure 4: Histogram and boxplot of the *systolic blood pressure*, with outliers.

2.2.3 Correlations

In fig. 11 is presented the correlations between all variables of the data set. The analysis of the correlations between the predictors showed that all predictors but *aphi* and *aplo* were slightly correlated. Taking into account that high correlated predictors might lower the model performance, one of *aphi* and *aplo* predictors had a higher chance of being eliminated from the final dataset. On the other side, *aphi* and *aplo* were also the most correlated ones to the target, in other words, the ones that most contribute to the prediction of presence of cardiovascular disease. The variables *active*, *alco*, *smoke*, *height* and *gender*, being not correlated at all to the target, were prone to being eliminated.

The discussion and conclusion of the final feature selection of the data set is covered in the next subsection.

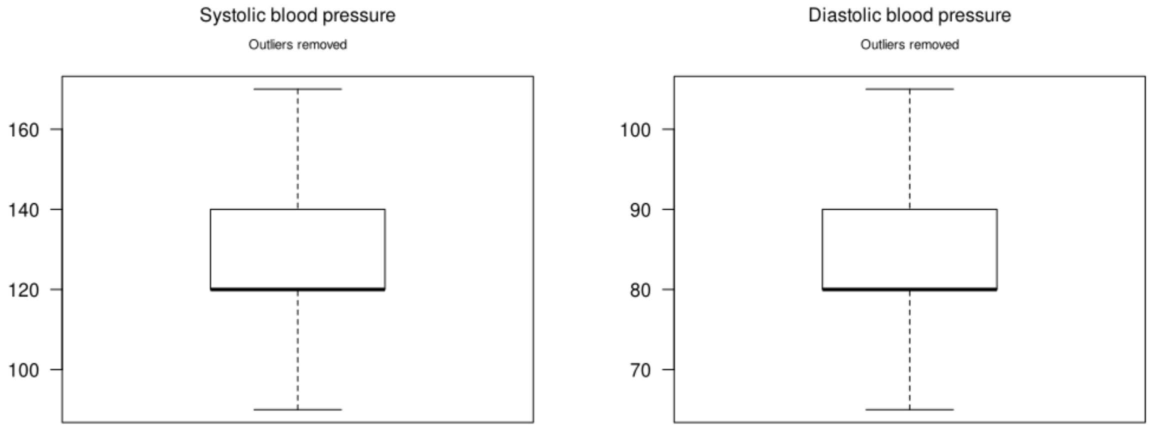
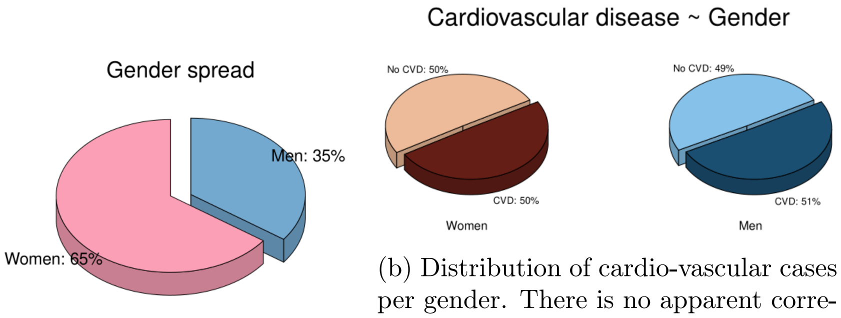


Figure 5: Histogram and boxplot of the *systolic blood pressure* without outliers.



(a) Gender spread in the sample.

(b) Distribution of cardio-vascular cases per gender. There is no apparent correlation between the variables *gender* and *cardio*.

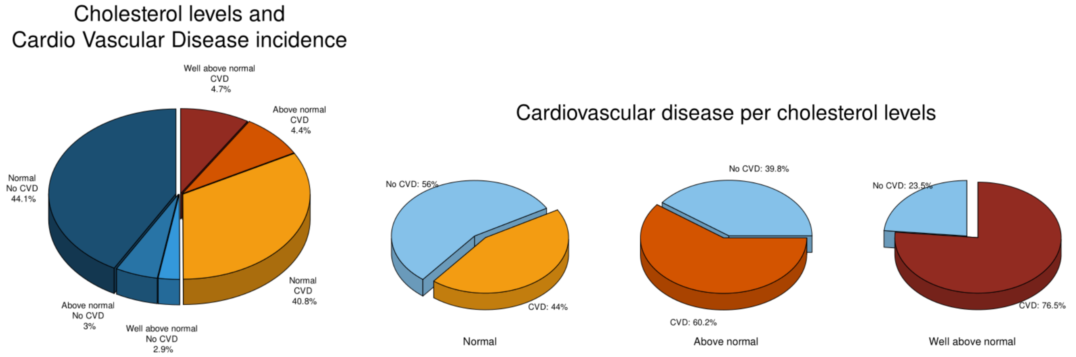
Figure 6: Gender spread of the sample (*left*), and distribution of cardio-vascular cases per gender (*right*). CVD stands for *cardio-vascular disease*.

2.3 Final data set

All outliers identified in section 2.2 were excluded from the original data set, obtaining a new one with 62 505 observations. This set was the basis for the feature selection, which is going to be explained in the following.

In order to decide which features were going to be kept in the final data set, a quick study was made to the variables importance and the residuals by fitting a Linear Regression model.

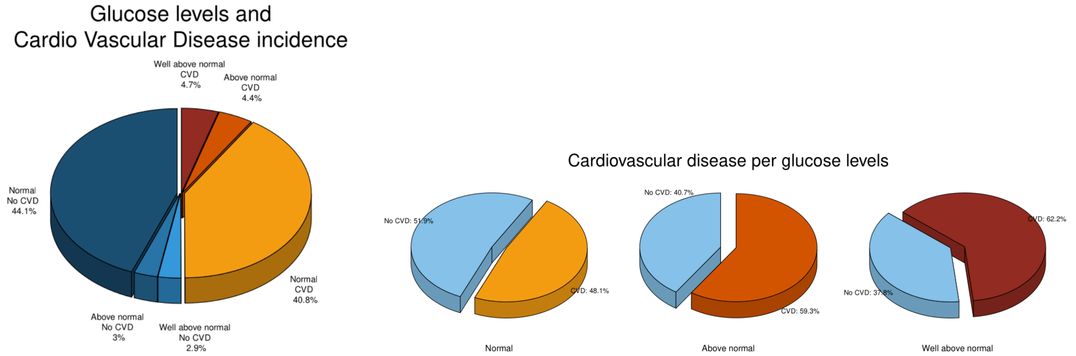
Based on the variance importance returned by the complete model, all but height and gender were statistically significant. Adding to this the conclusions from the analysis referred in section 2, a total of five new linear regression models were fitted, namely, with all minus the *gender*, with all minus the *height*, with all minus the *aphi*, with all minus the *aplo* and with all minus the *gender* and the *height*. Using the adjusted R-Squared and the Anova test as a selecting criteria, the model with all predictors minus the gender and height was the best one.



(a) Overview of the cardio-vascular disease incidence per level of cholesterol.

(b) Distribution of cardio-vascular disease cases per cholesterol level. There is evidence for the linear correlation between the variables *choles* and *cardio*.

Figure 7: Overview of cardio-vascular disease (CVD) per level of cholesterol (*left*), and distribution of CVD cases per cholesterol level (*right*).

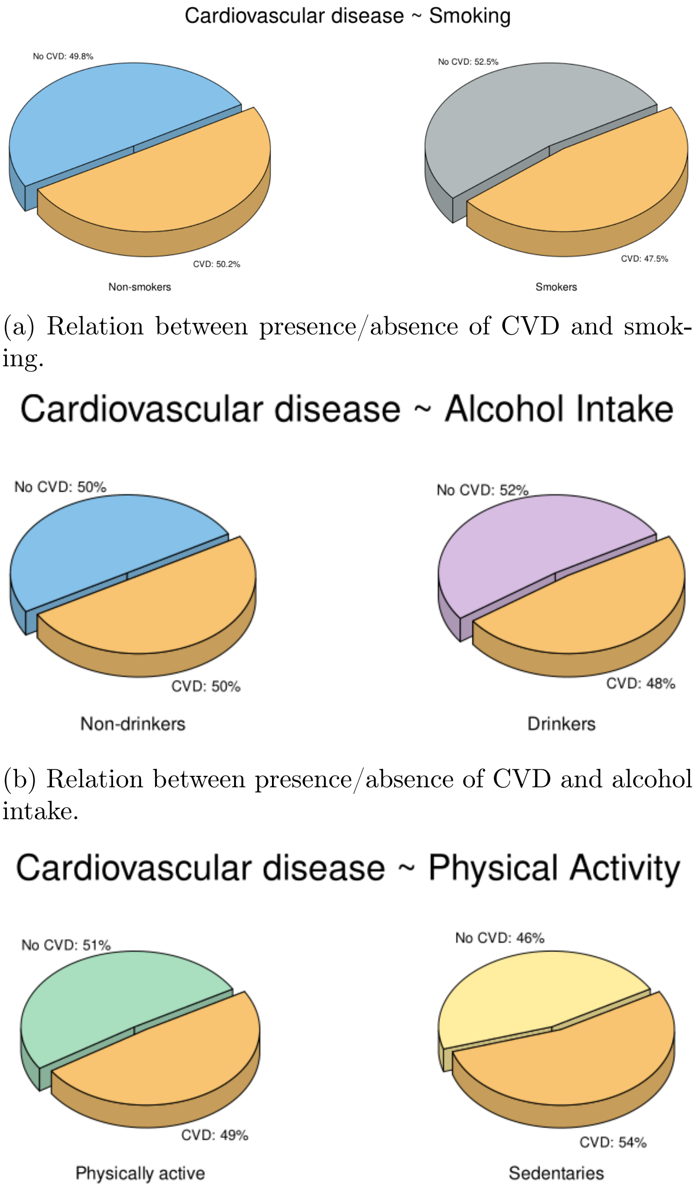


(a) Overview of the cardio-vascular disease incidence per glucose level.

(b) Distribution of cardio-vascular disease cases per glucose level. There is evidence for the linear correlation between the variables *gluc* and *cardio*.

Figure 8: Overview of cardio-vascular disease (CVD) per level of glucose (*left*), and distribution of CVD cases per glucose level (*right*).

Figure 12 presents the linear model graphics analysis. The Cook's distance were all bellow 1, reveling an nonexistence of influential points that could possibly be interfering with the fit of the model. The analysis of residuals showed that they were not normal, as can be seen from the histogram and the QQ-plot of the residuals illustrated in fig. 12. Even though a linear regression model is not the best fitting model when the target is categorical, the previous conclusions are not to discard completely. Therefore, this study compared the results between the complete model and the model with all features minus the gender and height.



(a) Relation between presence/absence of CVD and smoking.

(b) Relation between presence/absence of CVD and alcohol intake.

(c) Relation between presence/absence of CVD and the practice of physical activity.

Figure 9: Overview of presence or absence of cardio-vascular disease (CVD) and smoking (*top*), alcohol intake (*middle*), and the practice of physical activity (*bottom*).

3 Modeling

The models covered in this study were Logistic Regression (LogR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Lasso Regression (Lasso), Ridge

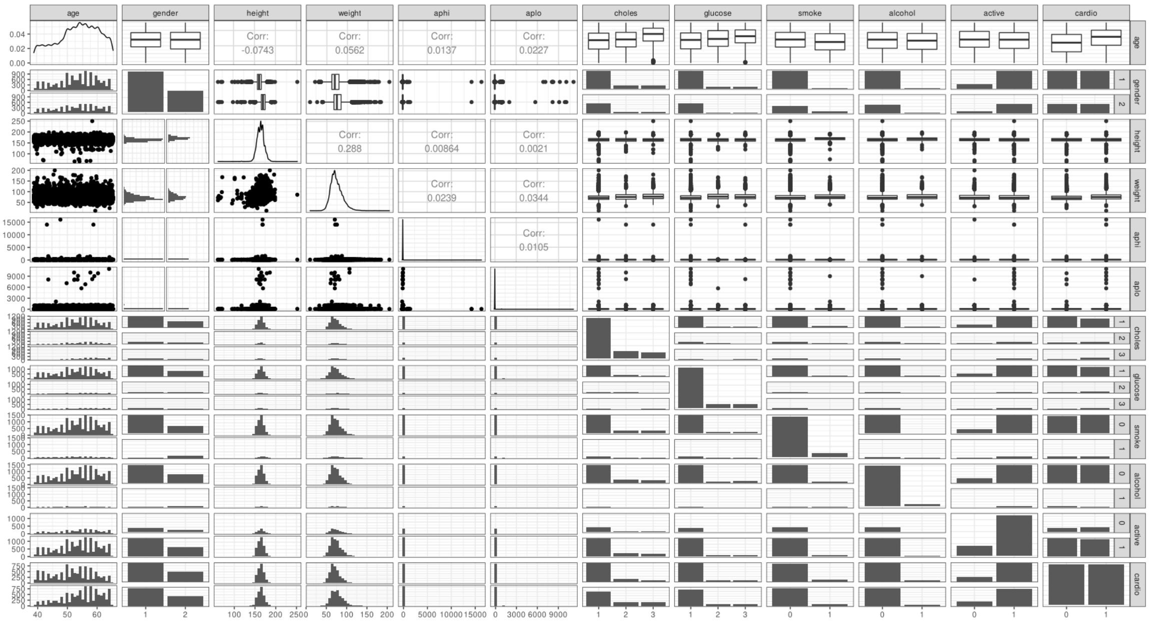


Figure 10: Graphics description of data

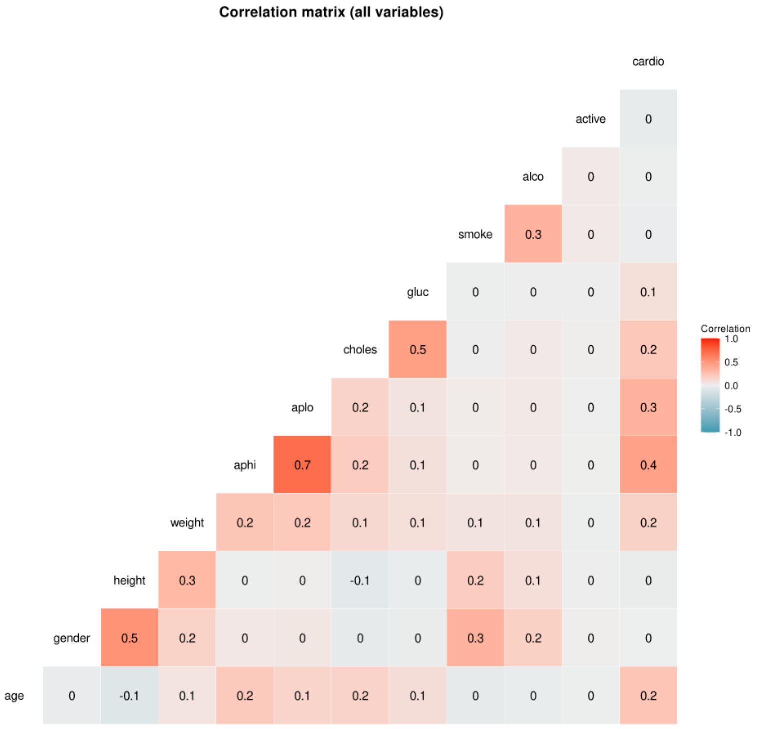


Figure 11: Correlations of all variables

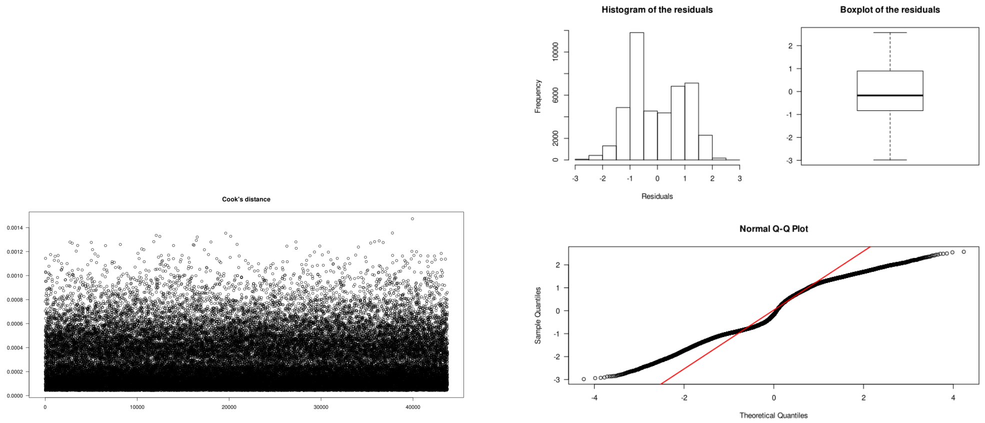


Figure 12: Linear Model graphics analysis. The Cook's distances (*left*) are all below 0.1 while the histogram and the QQ-plot (*right*) clearly show that the distribution of the residuals is not normal.

Regression (Ridge), K-Means Clustering (KM), and Hierarchical Clustering (HC). The analysis of these methods and their application to the data set is going to be explained in the following sections.

3.1 Logistic Regression, LDA and QDA

The Logistic Regression, Linear Discriminant Analysis and Quadratic Discriminant Analysis have no hyperparameters to be tuned. These models can only be fine tuned by means of the transformation of the data. The closer the sample distributions to the theoretical ones, presumed by the models, the better the performance by the latter. For example, LDA assumes a Gaussian distribution of the data. Therefore, the closer the sample distribution to a normal one, the better the performance of LDA.

In our case, as was identified in fig. 11 and section 2.2, none of the sample distributions of the variables were normal. Hence, we did not expect a good performance of the models under study.

3.2 Lasso and Ridge Regression

The Lasso and Ridge Regression have an α parameter to be tuned. The alpha is a weight that determines the importance of each predictor. To find the best α, 10-fold cross validation was used.

3.3 K-Means Clustering

The `kmeans` function in **R** was used for the K-Means clustering. In order to identify the best clustering, the number of clusters to identify was varied from 1 to 20, and the sum of the variations within each cluster was tracked on. In each iteration, 25 random sets were chosen for the identification of the clusters. Then, the reduction in variance per value of k was plotted. Figure 13 corresponds to the "elbow" point illustrating the variation within each cluster as a function of the number of clusters. From inspection of the graph, the optimal number of clusters is found to be approximately equal to 4 or 5.

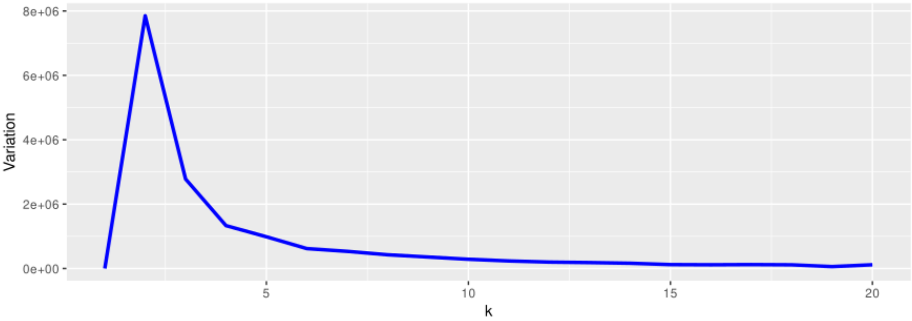


Figure 13: "Elbow" plot of the total variation within each cluster as a function of the number of clusters, computed using the `kmeans` function in **R**. The optimal number of clusters to split the data is approximately 4 or 5.

3.4 Hierarchical Clustering

In the Hierarchical Clustering method, two distances were used to fit the model: the Euclidean's distance, and the Gower's distance. The Gower's distance is used when the dataset has a mixture of categorical and numerical features. "In short, Gower's distance (or similarity) first computes distances between pairs of variables over two data sets and then combines those distances to a single value per record-pair." [1].

After testing which is the best clustering method for this train set, fig. 14, the chosen method was "Ward.D", or the *Ward's minimum variance method*, which "aims at finding compact, spherical clusters." [2]. Since the data set was large, *i.e.*, it was composed of many observations, it was necessary to reduce it to 500 instances, keeping the proportion of presence and absence of cardio diseases occurrences, in order to get a reliable output from the hierarchical clustering. To perform the *HC* with Euclidean distance all predictors were previously standardized.

The dendograms obtained from hierarchical clustering using the Ward method and Euclidean and the Gower distances are represented in figs. 15a and 15b, respectively. The heatmaps obtained from the `heatmap` function in **R** and the heatmap ordered by patient and predictors using Euclidean distance are represented in figs. 16a and 16b, respectively.

3.5 Results

The accuracies obtained for each model are presented in table 2. In general, the accuracies from the complete model and from the one without *gender* and *height* predictors are identical. Even between different machine learning models the results are very close. The model

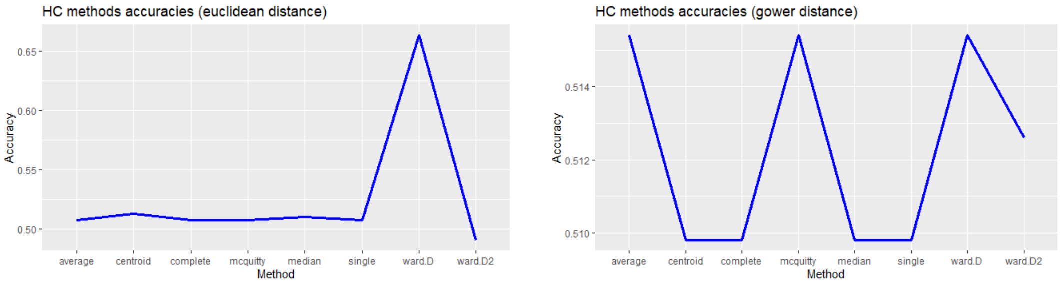
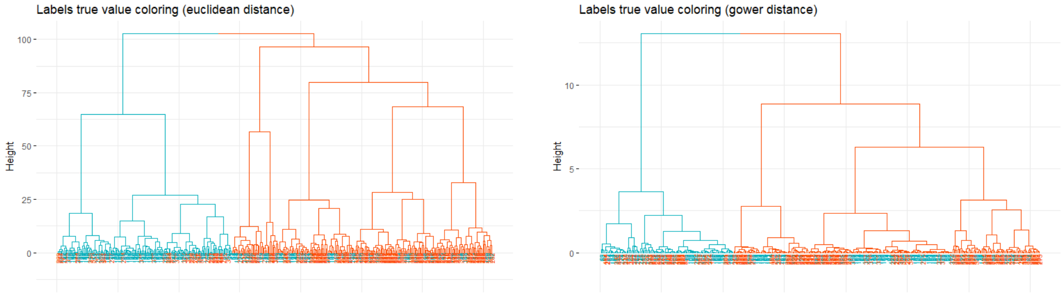


Figure 14: Hirarchical Clustering training set accuracies for different clustering methods for both Euclidean and Gower's distance



(a) Dendogram obtained from hierarchically clustering the data with Ward method and Euclidean distance.

(b) Dendogram obtained from hierarchically clustering the data with Ward method and Gower distance.

Figure 15: Dendograms obtained from the Hierarchical Cluster model.

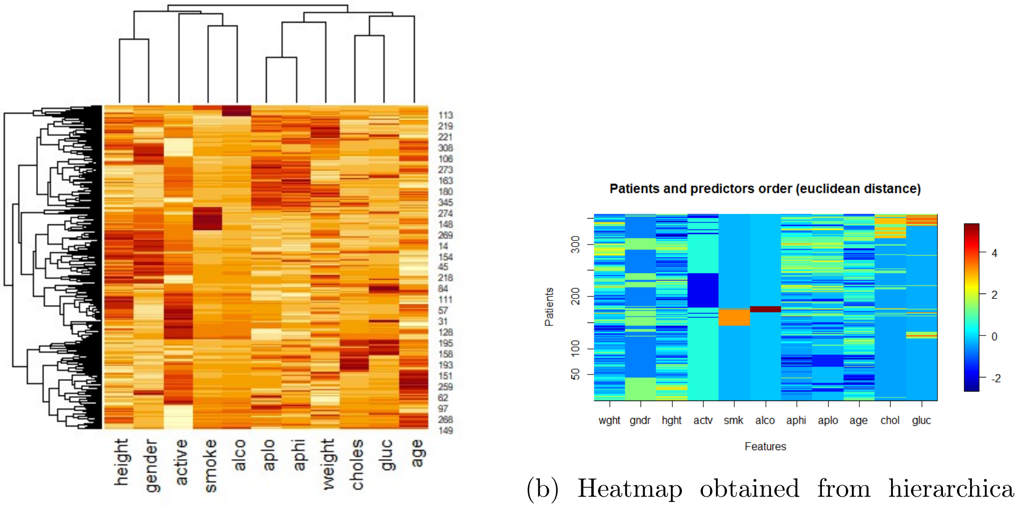
with the best performance was *LDA*. Based on the ROC curves of the models, fig. 17, we have a draw between all except the QDA model, which performed the worst. The comparison of models is discussed in more detail in the next subsection.

Table 2: The accuracies obtained for the different models studied in this project.

Model	**Complete model**		**- gender and height**	
	Train Acc.	**Test Acc.**	**Train Acc.**	**Test Acc.**
LogR	0.705	0.703	0.704	0.702
LDA	0.723	0.724	0.723	0.726
QDA	0.687	0.684	0.688	0.687
Lasso	0.724	0.723	0.724	0.723
Ridge	0.721	0.723	0.722	0.725

3.6 Models comparison

In Ridge Regression, the main idea is to introduce a small amount of bias in the training model in order to reduce the variance in the testing set, so it can provide better long term



(a) Heatmap obtained from hierarchically clustering the data with Ward method.

(b) Heatmap obtained from hierarchically clustering the data with Ward method and Euclidean distance ordered by patients and predictors.

Figure 16: Hierarchical Clustering model's heatmap

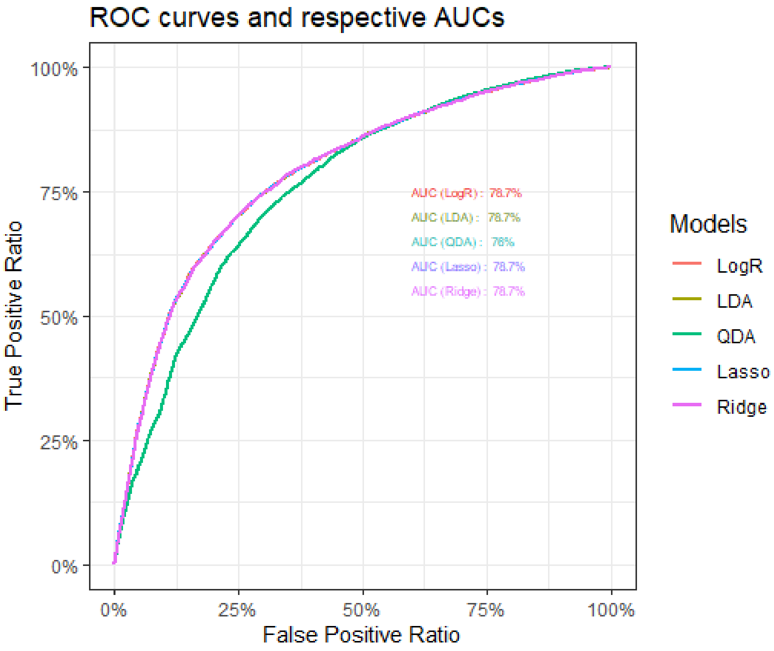


Figure 17: ROC curves

predictions. The major difference between Ridge and Lasso Regression is that Ridge can only reduce the slope asymptotically close to zero, while Lasso can shrink the slope all the way to zero. Hence, as the λ parameter increases, Lasso can remove some predictors (because the slope became zero) and, because of this, Lasso is a little better than Ridge at reducing the variance in models that contain a lot of useless variables.

The main difference between the clustering methods is that *K-Means* specifically tries to put the data into the number of clusters pre-defined, while *HC* outputs, pairwise, what two things are most similar. However, by means of an "elbow" plot, it is possible to estimate the optimal number of clusters to search for with *K-means*.

4 Conclusion

In this project, several supervised and non-supervised models were applied to a data set of 70 000 observations over eleven features, quantitative and qualitative in nature, in order to perform regression and clustering. Outliers removal and feature selection were performed during the exploratory data analysis, although both the full and reduced data set were used throughout the application of the models to the data, being the results yielded by both sets compared.

Using the accuracies on the test set as comparison metric, all the models performed identically, with exception for QDA, which behaved worse than the others. LDA was the model which performed the best, followed closely by Ridge and Lasso.

Hierarchical clustering indicates three or four clusters in the data, depending if the Ward method is used with the Euclidean or the Gower distance, respectively. By creating an "elbow" plot of the variations within each cluster as a function of the number of clusters, K-means yielded four to five clusters. Therefore, both methods indicated approximately the same number of clusters in the data.

This project reveled to be challenging, specially at the level of the data analysis. Extra care was needed in preparing the data set prior to the application of the models. In fact, during the first attempt, without the removal of the outliers, the results were poor in terms of accuracy and identification of clusters. The analysis of this data set allowed the deepening of the contents studied during the classes and fostered a will to pursue further research on other methods. An example is the *Elastic-net* method, very similar to Ridge and Lasso in nature, and which could be applied and compared to those models.

References

[1] Mark van der Loo, CRAN, [0nline], Accessed at `https://cran.r-project.org/web/packages/gower/vignettes/intro.pdf`, 21/04/2020

[2] RDocs, [0nline], Accessed at `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html`, 21/04/2020