Università di Catania

# Statistical Learning Report - Heart Attack Data

Academic year 2022-2023

Giovanni Spadaro

May 25, 2023

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction and data exploration

This report aim at exploring some statistical learning techniques applied to the field of medical care.
In this area the decision making process and the diagnostics are yet mostly heuristic and require a certain amount of time, so the development and the application of these kind of tools can be very useful in this terms.
The setting of interest for this analysis is the prediction of heart diseases. Thanks to a dataset choosen from Kaggle [1], the aim is to classify patients based on whether they have an heart disease or not.

The source code is available on GitHub [2].

## 1.1 Data description

By looking at the dataset documentation [1][3], the followings are the variables with their description:

- *age*: age in years
- *sex*: sex. Values:
  - Value 0: female
  - Value 1: male
- *cp*: chest pain type. Values:
  - Value 0: typical angina
  - Value 1: atypical angina
  - Value 2: non-anginal pain
  - Value 3: asymptomatic
- *trestbps*: value of resting blood pressure on admission to the hospital (measured in $mmHg$). The documentation doesn't report whether this variable refers to systolic or diastolic blood pressure, but probably it refers to the systolic one since the distribution of high values suggests this.

- *chol*: serum cholesterol (measured in $mg/dl$)
- *fbs*: fasting blood sugar. Values:
  - Value 0: $fbs \leq 120\ mg/dl$
  - Value 1: $fbs < 120\ mg/dl$
- *restecg*: resting electrocardiographic results. Values:
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $> 0.05\ mV$)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- *thalach*: maximum heart rate achieved (measured in BPM).
- *exang*: exercise induced angina. Values:
  - Value 0: no
  - Value 1: yes
- *oldpeak*: ST depression induced by exercise relative to rest
- *slope*: the slope of the peak exercise ST segment. Values:
  - Value 0: upsloping
  - Value 1: flat
  - Value 2: downsloping
- *ca*: number of major vessels (0-3) colored by fluoroscopy
- *thal*: thalassemia. Values:
  - Value 1: normal (no blood flow in some part of the heart)
  - Value 2: fixed defect
  - Value 3: reversable defect (a blood flow is observed but it isn't normal)
- *target*: diagnosis of heart disease (angiographic disease status). Values:
  - Value 0: absence of heart disease
  - Value 1: presence of heart disease

The following tables shows the first rows of the data.

The first 13 variables represent the health status of the patients and will be used as predictors $\boldsymbol{X}$ in the statistical models. The last column in the set is related to the health status of the patient, which is the response $y$.

The training data is composed of 242 rows.

|  | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 68 | 45 | 0 | 1 | 130 | 234 | 0 | 0 | 175 | 0 | 0.60 | 1 | 0 | 2 | 1 |
| 43 | 45 | 1 | 0 | 104 | 208 | 0 | 0 | 148 | 1 | 3.00 | 1 | 0 | 2 | 1 |
| 51 | 51 | 0 | 2 | 130 | 256 | 0 | 0 | 149 | 0 | 0.50 | 2 | 0 | 2 | 1 |
| 88 | 46 | 1 | 1 | 101 | 197 | 1 | 1 | 156 | 0 | 0.00 | 2 | 0 | 3 | 1 |
| 29 | 65 | 0 | 2 | 140 | 417 | 1 | 0 | 157 | 0 | 0.80 | 2 | 1 | 2 | 1 |
| 99 | 43 | 1 | 2 | 130 | 315 | 0 | 1 | 162 | 0 | 1.90 | 2 | 1 | 2 | 1 |

Table 1.1: Train dataset head.

### 1.1.1 Aim of the analysis and report structure

So, after a quick look at the data, some questions arises:

- How accurately can we predict if a patient is affected by any heart disease?
- Which characteristic of the patient has the most impact on his heart condition?

The aim of this analysis is to give an answer to these questions.

The analysis is divided in four main parts: an exploratory analysis of the training set which will provide a deep understanding of the data, a training phase of three statistical models (logistic regression, random forest and neural networks), a validation phase of the models and a final comparison to choose the best model according to some metrics.

## 1.2 Data exploration

The starting point of this project is the Exploratory Data Analysis (EDA) of the training set.

First of all some questions are considered:

1. Is the dataset balanced according to target and sex?
2. Is there a relationship between the age and the target? And what about the sex and the target?
3. Is there correlation between the numerical variables?
4. Which variables influence the most the target?

### 1.2.1 Univariate analysis

The univariate analysis starts by having a look at the overall statistics for each variable according to the type.

It has been noted that there are no missing values in the dataset, all of the categorical variables are unordered and, by comparing the number of rows with the number of unique rows of the dataset, it has been found a duplicate row, which has been removed.

| type | variable | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|
| numeric | age | 54.07 | 8.77 | 29.00 | 47.00 | 56.00 | 60.00 | 76.00 |
| numeric | trestbps | 132.60 | 17.04 | 100.00 | 120.00 | 130.00 | 140.00 | 200.00 |
| numeric | chol | 246.81 | 53.58 | 126.00 | 211.00 | 243.00 | 274.00 | 564.00 |
| numeric | thalach | 150.10 | 22.94 | 88.00 | 136.00 | 154.00 | 167.75 | 202.00 |
| numeric | oldpeak | 1.05 | 1.18 | 0.00 | 0.00 | 0.80 | 1.60 | 6.20 |

Table 1.2: Summary table of numerical variables

| type | variable | n_unique | top_counts |
|---|---|---|---|
| factor | sex | 2 | 1: 161, 0: 81 |
| factor | cp | 4 | 0: 118, 2: 65, 1: 42, 3: 17 |
| factor | fbs | 2 | 0: 208, 1: 34 |
| factor | restecg | 3 | 0: 121, 1: 119, 2: 2 |
| factor | exang | 2 | 0: 163, 1: 79 |
| factor | slope | 3 | 2: 114, 1: 109, 0: 19 |
| factor | ca | 5 | 0: 141, 1: 53, 2: 30, 3: 14, 4: 3 |
| factor | thal | 4 | 2: 132, 3: 92, 1: 16, 0: 2 |
| factor | target | 2 | 1: 132, 0: 110 |

Table 1.3: Summary table of categorical variables

Moreover, from the Table 1.3 some strange values can be observed: the variables *ca* and *thal* presents the levels 4 and 0 respectively which are not specified in the documentation. Further investigation will follow during the EDA.

Let's now have a look at the most important variables of the dataset.

### Age

Looking at the variable's left histogram it seems that there are two peeks and whis could be a symptom of two data generating populations. This observation can be more or less confirmed by the right histogram which highlights the patients' age distribution according to their diagnosis.

### Resting blood pressure

According to [4], high blood pressure can cause many heart problems, including: coronary artery disease, enlarged left heart, heart failure. Usually blood pressure's values higher 130 $mmHg$ are considered high.

Despite this premise, from the Figure 1.2 it doesn't looks like this variable has an impact on the target variable.

In regards with the overall variable distribution, it is positively skewed with a coefficient of skewness of 0.82.

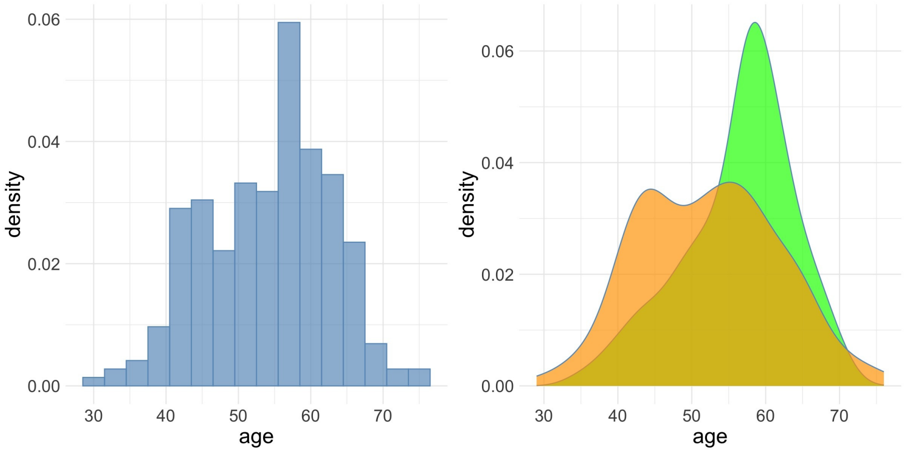

Figure 1.1: Left: Histogram of *age*. Right: Histogram of *age* with *target* hue (green for healthy patients and orange for sick patients).

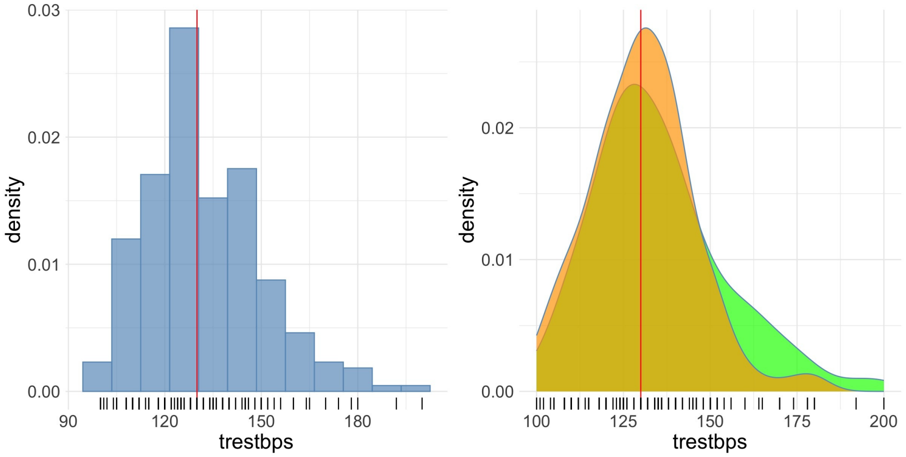

Figure 1.2: Left: Histogram of *trestbps*. Right: Histogram of *trestbps* with *target* hue (green for healthy patients and orange for sick patients). In both graphs the vertical redline in located at 130 $mmHg$.

### Cholesterol

High levels of cholesterol could be a cause of developing an heart disease. According to [5], values from 240 $mg/dl$ are considered high.

Although, looking at the right histogram in Figure 1.3, the cholesterol levels doesn't help to discriminate a sick patient from an healthy one.

The overall distribution is simmetric and there seems to be some outliers. The boxplot (Figure 1.4) highlights this more this aspect and it shows that there 5 outliers with values over 400 $mg/dl$.

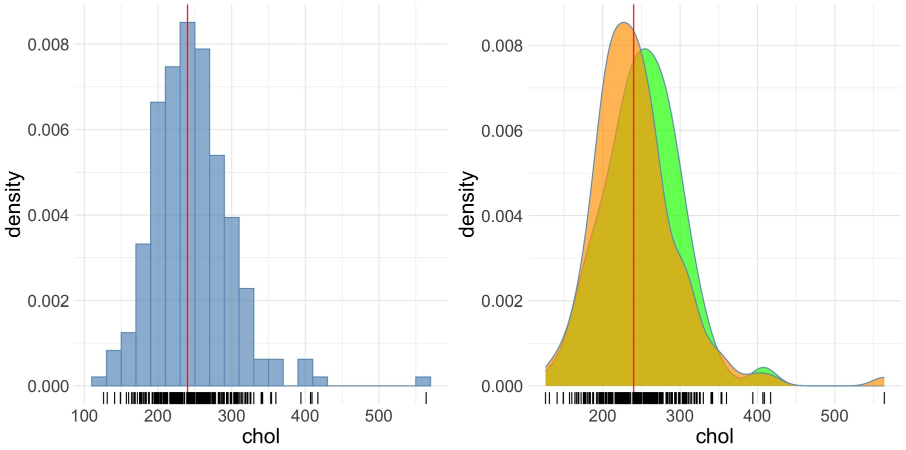

Figure 1.3: Left: Histogram of *chol*. Right: Histogram of *chol* with *target* hue (green for healthy patients and orange for sick patients). In both the graph the vertical redline in located at 240 $mg/dl$.

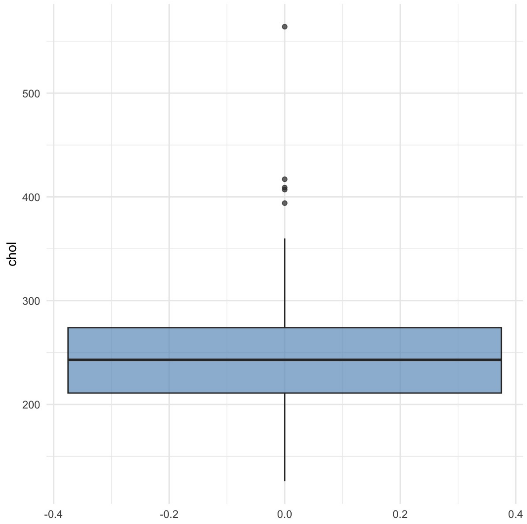

Figure 1.4: Boxplot of *chol*.

### Maximum heart rate achieved

The variable distribution looks slightly negatively skewed from the Figure 1.5 and its coefficient of skewness of $-0.47$ confirms what is shown in the left graph.

Considering the right graph, it seems that the variable is generated from two overlapped distribution: the first more or less simmetric and the second negatively skewed and with a higher mean. Calculating the mean and the skewness of the two distributions are $140.56 BPM$ and $-0.21$ for the healthy patients and $157.92 BPM$ and $-0.68$ for the sick patients.

So the maximum heart rate achieved by the patient has an impact on his probability to develop an heart disease.

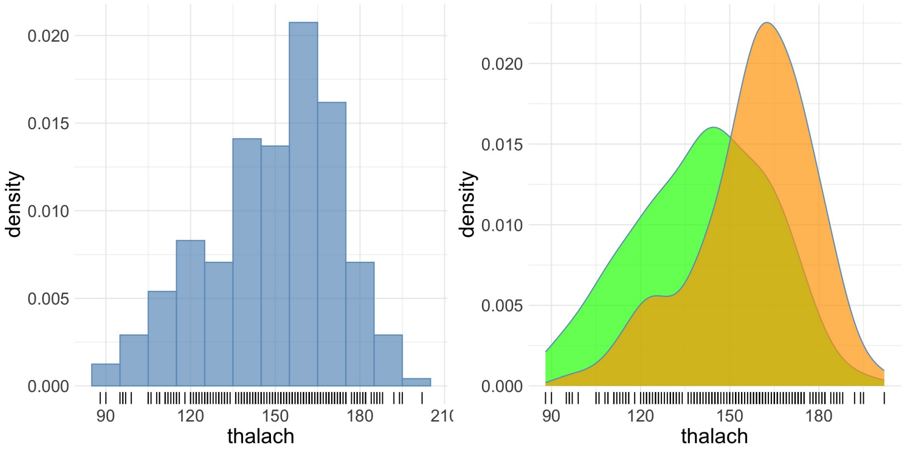

Figure 1.5: Left: Histogram of *thalach*. Right: Histogram of *thalach* with *target* hue (green for healthy patients and orange for sick patients).

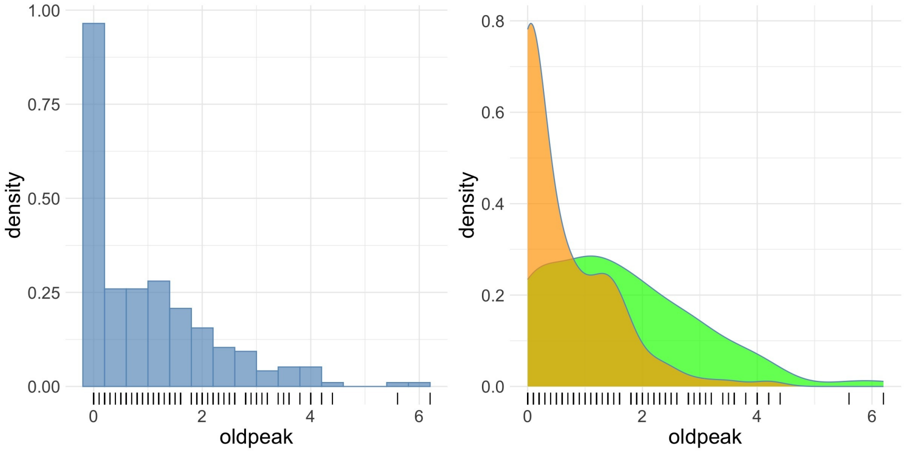

Figure 1.6: Left: Histogram of *oldpeak*. Right: Histogram of *oldpeak* with *target* hue (green for healthy patients and orange for sick patients).

### ST depression induced by exercise relative to rest

According to Figure 1.6, ST depression refers to a finding on the electrocardiogram wherein the trace in the ST segment is abnormally low below the baseline. An abnormal depression of the ST segment can be a symptom of myocardial ischaemia or infarction.

Despite this premise, the right graph doesn't show this trend, in fact for higher values of depression there isn't an high frequency of sick patients.

The overall distribution is positively skewed with a coefficient equal to 1.33. The 0 value seems to have an abnomal frequency, in fact by checking the table frequencies there are 79 rows (out of 241) with *oldpeak* $= 0$. This is supposedly due to a standard association of wrong or insecure readings to the 0 value.

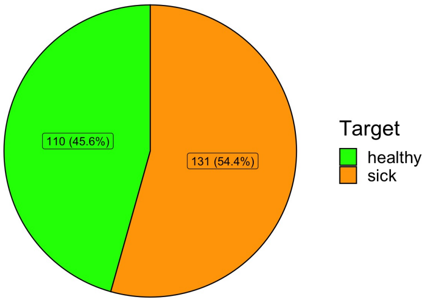

Figure 1.7: Pie chart of *target*

### Target

When performing a classification task the proportion of each class in the dataset has to be checked. As the Figure 1.7 shows, the dataset is balanced according the response variable since the proportion of each class is more or less the same. This ensures that the models won't have a bias in classifying patients according to the presence of an heart disease.

### Sex

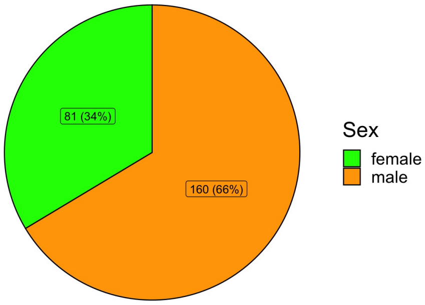

Figure 1.8: Pie chart of *sex*.

The same reasoning can be applied to the other categorical variables, such as the sex. Although the impact in the classifiers is not so relevant with respect to the response variable, it has to be noted that the dataset is unbalanced according to the sex. In particular, the Figure 1.8 shows that 2/3 of the dataset is made of male patients, so maybe the classifiers could develop a bias in better distinguishin the health status of a male patient rather the health status of a female one.

### Number of major vessels

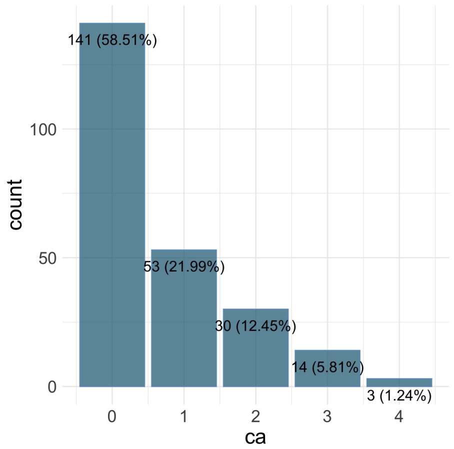

Figure 1.9: Barplot of *ca*.

As shown by Figure 1.9, there are 3 observations with registered value of *ca* equal to 4.

By looking at these observation from Table 1.4 the values for the other variables looks like normal.

So, since this value is not specified in the documentation, it is replaced with 3.

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 159 | 58 | 1 | 1 | 125 | 220 | 0 | 1 | 144 | 0 | 0.40 | 1 | 4 | 3 | 1 |
| 164 | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0.00 | 2 | 4 | 2 | 1 |
| 93 | 52 | 1 | 2 | 138 | 223 | 0 | 1 | 169 | 0 | 0.00 | 2 | 4 | 2 | 1 |

Table 1.4: Rows of the dataset with $ca = 4$

### Thalassemia

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | 53 | 0 | 2 | 128 | 216 | 0 | 0 | 115 | 0 | 0.00 | 2 | 0.00 | 0 | 1 |
| 282 | 52 | 1 | 0 | 128 | 204 | 1 | 1 | 156 | 1 | 1.00 | 1 | 0.00 | 0 | 0 |

Table 1.5: Rows of the dataset with $thal = 0$

As in the previous case, the variable presents a few observation of undocumented values. So these values will be replaced the value 1.

## 1.2.2 Multivariate analysis

Let's now explore the relationship between the numerical variables.

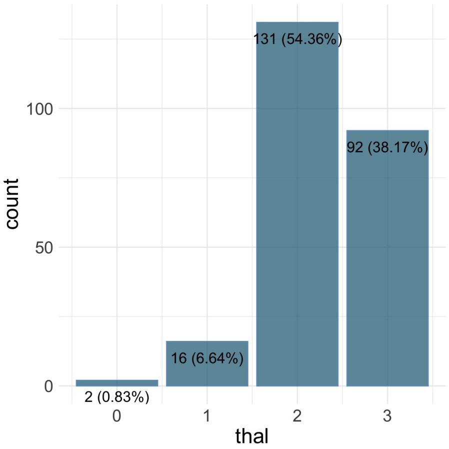

Figure 1.10: Barplot of *thal*.

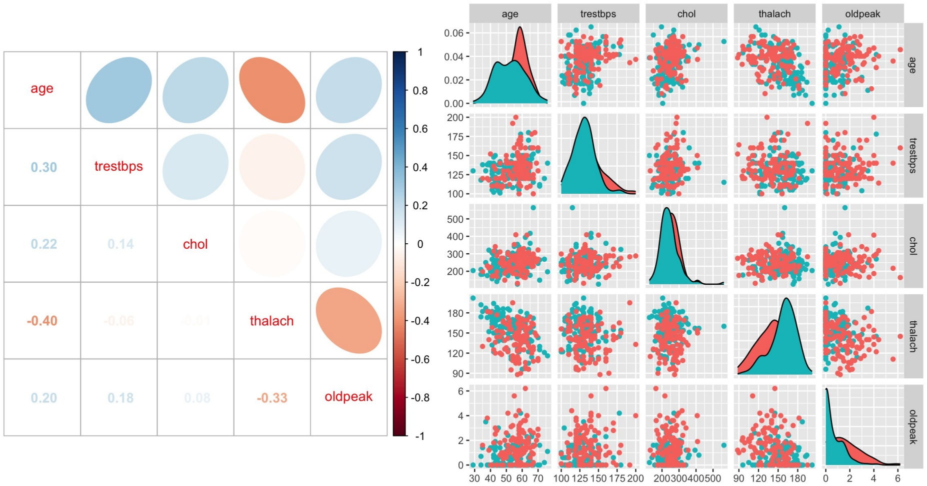

Figure 1.11: Left: Correlation matrix of numerical variables. Right: Matrix of pairwise scatterplots of numerical variables. Red dots represent healthy patients and blue dots represent sick patients.

The correlation matrix (Figure 1.11 Left) shows that the variables are mostly uncorrelated between them, but there are a few expections. *age* has a slight negative correlation with the maximum heart rate achieved (*thalach*), which is expected since the heart usually has more difficulties pumping blood as the patient gets older, and it has also a slight positive correlation with the resting blood pressure. The maximum heart rate achieved is slightly negatively correlated with the ST depression (*oldpeak*).

The Figure 1.11 Right shows the matrix of the scatterplot considering all possible pairs of the numerical variables. The diagonal is dedicated to the densities of the variables and the off-diagonal elements are the pairwise scatterplots.

From the graphs it can be noted that there are almost no visible groups in the dataset and also the response variable doesn't help in any other kind of classification. The only exception is in the scatterplot of *age*-*thalach*, from which it can be noted a slight higher concentration of blue points in the upper left corner with respect to an higher concentration of red points in the bottom right corner, but still with a very noticeable overlap.

## 1.3 Comments and conclusions of EDA

At the end of this analysis some conclusions can be made:

- The dataset presents a very limited number of observations. Although this is a common issue among the medical dataset, the models will be affected from this aspect.
- The dataset is balanced, so the models the be trained without any worry for possible classifications biases.
- The univariate and bivariate analysis don't show good results in group separation between healthy and sick patients, with the only weak expections of the variables *thalach* and *age*.

# Chapter 2

# Data modeling

In this chapter it will be analized three models in order to predict the response variable. Since *target* is a categorical variable this task falls into the classification setting, more specifically it's a binary balanced classification since the variable has two possible values and it's balanced.

The assigned data comes divided into two sets: a training set and a test set. The first one will be divided: one for training the models (80%) and one for validating them (20%).

The models that will be tested are the logistic regression (section 2.1), the random forest (section 2.2) and the feedforward neural networks (2.3).

## 2.1 Logistic regression

The logistic regression is a linear model based on the sigmoidal function described in equation 2.1.

$$\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \tag{2.1}$$

Thanks to the shape of the sigmoidal function, the probability of an observation to belong to class 1 can be modeled. Then by choosing a threshold under which the observation falls in class 0 and over which the observation falls in class 1, the classification can be made.

Before training the model the train set is partitioned in two parts: the first part (80%) will be used to effectively train it and the second part (20%) will be used to validate its performance.

The first logistic regression model has been fitted using all variables available.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---:|---:|---:|---:|---:|---|
| (Intercept) | 1.72e+00 | 3.87 | 0.44 | 0.66 | |
| age | 2.27e-03 | 0.04 | 0.06 | 0.95 | |
| sex1 | -1.46e+00 | 0.72 | -2.02 | 0.04 | * |
| cp1 | 7.06e-01 | 0.74 | 0.96 | 0.34 | |
| cp2 | 2.34e+00 | 0.73 | 3.19 | 0.00 | ** |
| cp3 | 1.47e+00 | 0.93 | 1.59 | 0.11 | |
| trestbps | -1.22e-02 | 0.02 | -0.80 | 0.42 | |
| chol | -4.16e-03 | 0.01 | -0.83 | 0.41 | |
| fbs1 | -1.49e-01 | 0.72 | -0.21 | 0.84 | |
| restecg1 | 2.61e-01 | 0.52 | 0.50 | 0.62 | |
| restecg2 | -1.26e+01 | 1455.40 | -0.01 | 0.99 | |
| thalach | 1.90e-02 | 0.02 | 1.17 | 0.24 | |
| exang1 | -1.43e+00 | 0.64 | -2.25 | 0.02 | * |
| oldpeak | -2.65e-02 | 0.29 | -0.09 | 0.93 | |
| slope1 | -1.26e-01 | 0.94 | -0.13 | 0.89 | |
| slope2 | 1.07e+00 | 1.01 | 1.06 | 0.29 | |
| ca1 | -2.90e+00 | 0.67 | -4.33 | 0.00 | *** |
| ca2 | -3.40e+00 | 0.93 | -3.65 | 0.00 | *** |
| ca3 | -3.23e+00 | 1.25 | -2.58 | 0.01 | ** |
| thal2 | 1.61e-01 | 1.01 | 0.16 | 0.87 | |
| thal3 | -1.68e+00 | 0.94 | -1.79 | 0.07 | . |

Table 2.1: Coefficient estimates of logistic regression model using all variables.

The table 2.1 shows the result of the fit. Almost all predictors are not statistically significant because of a p-value higher than 0.05. The only variables that are statistically important are *sex*, *cp*, *exang* and *ca*. It can be noted that all of these predictors are categorical variables.

The AIC of the model is equal to 158.26.

|  | AIC | Variables selection |
|---|---|---|
| Model 1 | 146.26 | sex + cp + exang + slope + ca + thal |
| Model 2 | 146.82 | sex + cp + thalach + exang + slope + ca + thal |
| Model 3 | 147.49 | sex + cp + chol + thalach + exang + slope + ca + thal |

Table 2.2: Ranking of logistic regression models according to AIC.

A stepwise logistic regression is now performed in order to a see which variables should be included in the model according to the AIC.

The table 2.2 shows the 3 best models. Although there is a ranking due to the AIC, the differences of the coefficients are very small, so the domain knowledge can help understand which of the should be selected. Since the maximum heart rate achieved and the cholesterol level can be important symptoms of the presence of an heart disease they should be included in the model. For what concerns the *thalach* variable, it's worth noting that in the EDA it showed an (small) impact on the response variable.

Let's now validate the models using the validation set.

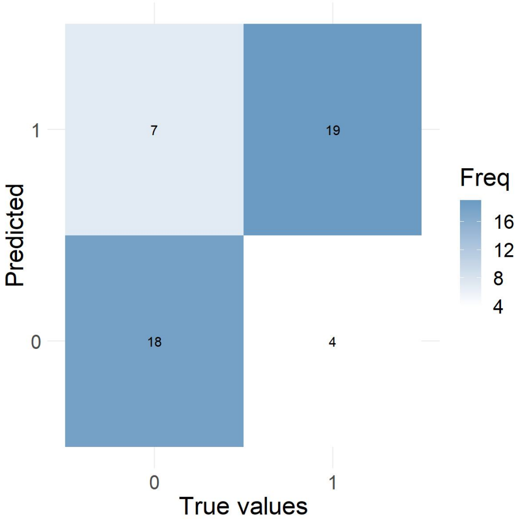

Figure 2.1: Correlation matrix of logistic regression model using all predictors.

Figures 2.1 and 2.2 and Table 2.3 show the performance of the model fitted using all variables.

|  | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| all_predictors | 0.771 | 0.229 | 0.818 | 0.826 | 0.774 |
| third_best_AIC | 0.771 | 0.229 | 0.773 | 0.800 | 0.771 |

Table 2.3: Performance metrics of logistic regression models

Figures 2.3 and 2.4 and Table 2.3 show the performance of the third best model according to AIC, which uses the following variables: *sex*, *cp*, *chol*, *thalach*, *exang*, *slope*, *ca*, *thal*.

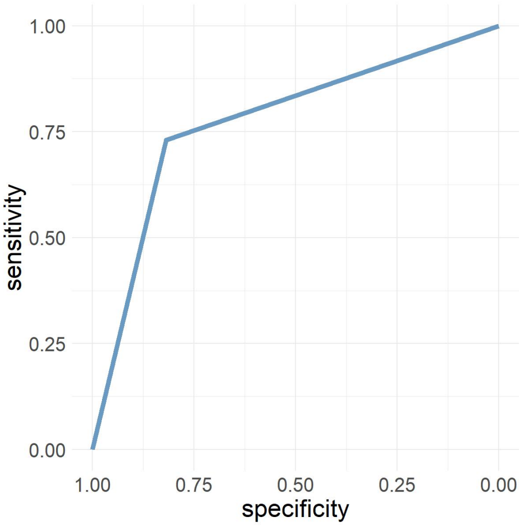

Figure 2.2: ROC curve of logistic regression model using all predictors.

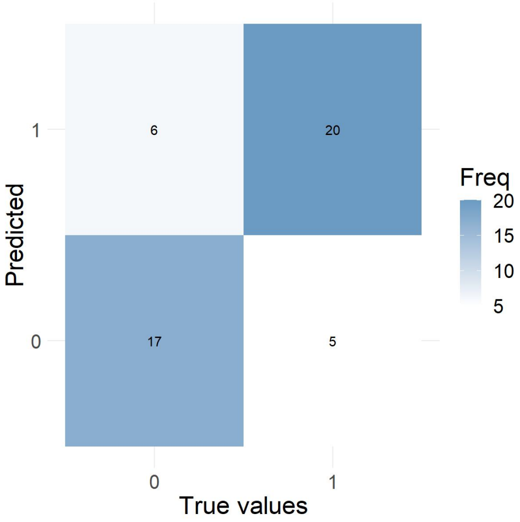

Figure 2.3: Correlation matrix of the third best logistic regression model according to AIC.

This model performes quite similarly to the previous one, in fact by looking at the confusion matrix there are the same number of miss classified patients. Despite this sensitivity and specificity have decresed, but this can be also due to the very small number of units.

It can be then concluded that the third best model choosed by the AIC is the best among the two models validated.

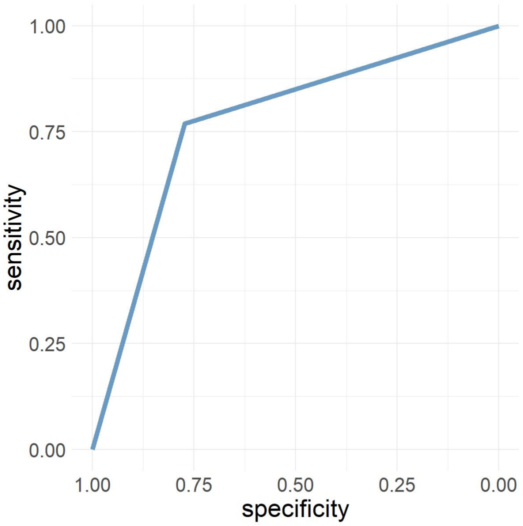

Figure 2.4: ROC curve of the third best logistic regression model according to AIC.

## 2.2 Random forests

The random forest is a statistical ensemble method based on decision tree. It is a modified version of bagging (bootstrap aggregation) in which the trees are grown considering a different subset of features for each split.

Usually the number of features choosen for each split is $\sqrt{p}$, being $p$ the number of features available. Since the dataset has $p = 13$ features, 4 has been choosen as the dimension of the feature subset each tree has access to for each split.

As before, the training set has been divided in two parts, one for training and one for validation. Though, due to how bagging works (each tree works with a bootstrapped version of the training data, which is statistically large 2/3 of the available data), there is another set of validation metrics which takes into account the OOB (Out-Of-Bag) observations.

| | 0 | 1 | class.error |
|---|---|---|---|
| 0 | 71 | 17 | 0.19 |
| 1 | 14 | 91 | 0.13 |

Table 2.4: Confusion matrix of the random forest model with OOB observations.

Looking at the confusion matrix made of the predicted OOB observations (Table 2.4), it can be observed that there's a 19% of error rate for the first class (which correspond to an healthy condition) and an 13% of error rate for the second class (which correspond to a having an heart disease). The total OOB estimate of error rate is 16%.

Considering now the validation metrics, Figures 2.5 and 2.6 and Table 2.5 shows the confusion matrix, the roc curve and the classification metrics.

It can be observed that the error rate is higher with respect to the OOB estimate error rate.

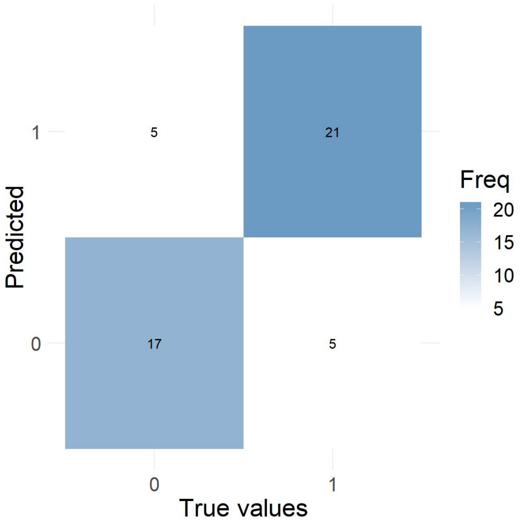

Figure 2.5: Confusion matrix of the random forest model.

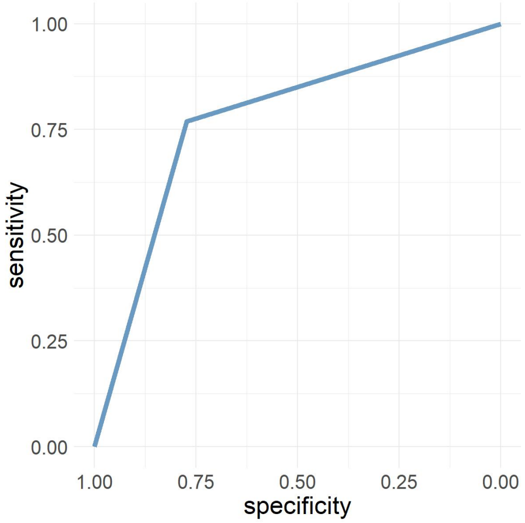

Figure 2.6: ROC curve of the random forest model.

By comparing it to the logistic regression, there is one more observation correctly classified, so the random forest performs a little bit better.

A relevant characteristic of the random forest, being a tree-based method, is the possibility to compute a coefficient of importance for every feature used during training. This is done by looking at how much the Gini index reduces by splitting the feature space using a certain predictor.

Figure 2.7 shows the feature importance of the predictors according to 2 metrics: the mean decrease accuracy and the mean decrease Gini.

- The mean decrease accuracy represents how much the accuracy in the OOB set decreases on average by excluding that predictor.

| Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---:|---:|---:|---:|---:|
| 0.792 | 0.208 | 0.773 | 0.808 | 0.790 |

Table 2.5: Performance metrics of the random forest model.

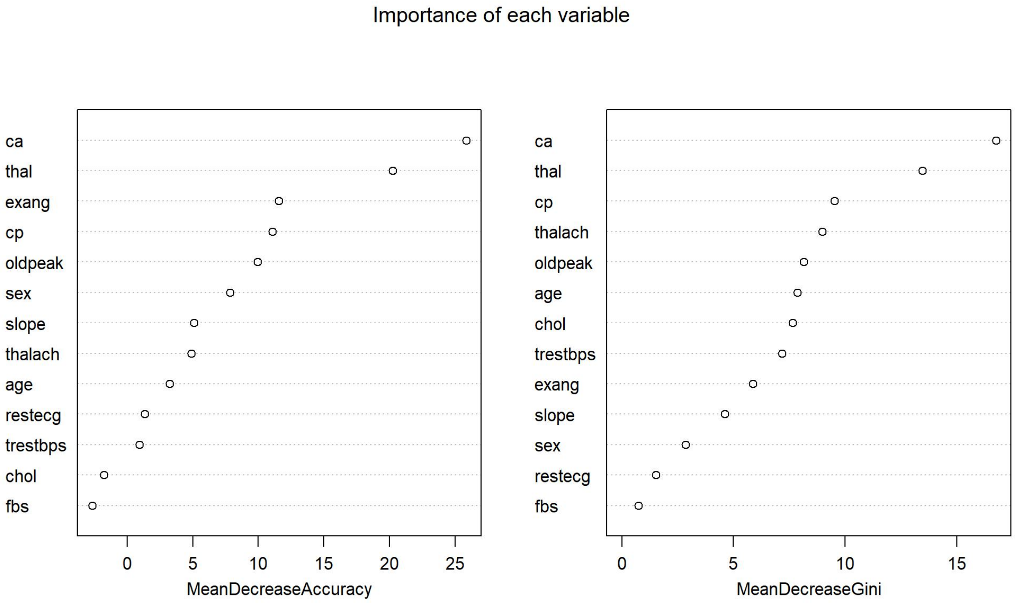

Figure 2.7: Variable importance graph.

- The mean decrease Gini represents how much a feature decreases the node impurity, that is determined by the Gini index.

So, the two best features in predicting the target variable are *ca* and *thal*.

## 2.3 Neural networks

The last model used for this classification task is the feedforward neural network. Neural networks are machine learning models which consist in an architecture of interconnected layers of neurons. Each neuron can be seen as a process node for its input values based on its activation function, usually the rectified linear unit (ReLU) is the most common choice (equation 2.2).

$$g(z) = max(0, z) \quad where\ z = w^T x \tag{2.2}$$

The architecture choosen for the model is the feedforward one: there are $n$ layers in which each neuron is connected to all the neurons in the next layer and retrieves information from all neurons in the previous layer.

The training process is focused on estimating the weights of each neuron-to-neuron connection in the network and it's performed thanks to the errors' backpropagation algorithm. Before executing this algorithm a loss function and an optimization method are defined: in the case of binary classification problems the binary crossentropy (BCE) is the most common loss function (equation 2.3); for what concerns the optimization method the gradient descent is the most used algorithm, although some more sophisticated methods are available, such as the adam optimizer.

$$L(\theta) = -\sum_{i=1}^{n} y_i\ log(f_0(x_i)) + (1 - y_i)\ log(f_1(x_i)) \tag{2.3}$$

To build the feedforward neural network model several attempts were made to find the best network configuration in terms of *number of hidden layers*, *number of neurons for each layer*, *learning rate* of the optimizer (which governs how aggressively the method will update the weights at each step), *batch size* (that is the number of observations to work through before updating the weights) and *number of training epochs* (that is the number of times the network will see the entire dataset during the training process).

The final network has 2 hidden layers, with 7 neurons in the first layer and 4 neurons in the second one (both layers uses ReLU as activation function), and an output layers with 2 neurons, one for each possible outcome, governed by the logistic activate function (equation 2.1).

For the learning process the BCE and the adam optimizer were choosen as loss function and optimization method respectively. The best observed lerning hyper-parameters are: 32 samples as *batch size*, 70 *epochs* of learning, and *learning rate* equal to 0.001.

Figures 2.8 and 2.9 shows losses, accuracies and their respective trends (given by polinomial fitting estimation [6]) over training epochs. The behaviour is stable and the training and validation curves follow the same trend, but the final results are quite poor. Table 2.6 shows a final loss value of 0.508 (for the validation set) which is well above the ideal value of 0.

This behaviour can be a symptom of underfitting.

The final validation metrics are reported in Table 2.7 and Figures 2.10 and 2.11.

This model obtains slightly better results than the previous ones.

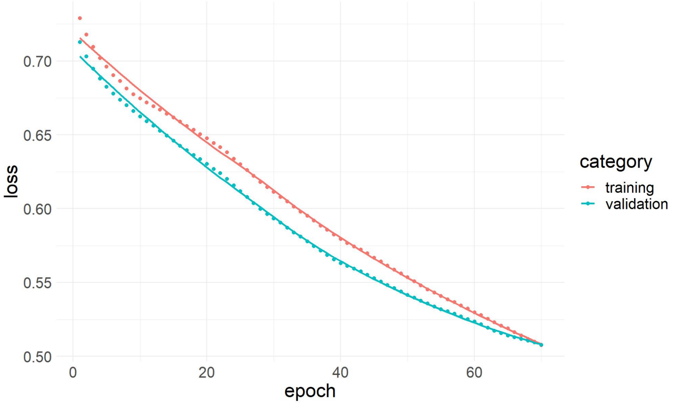

Figure 2.8: Loss trend over epochs.

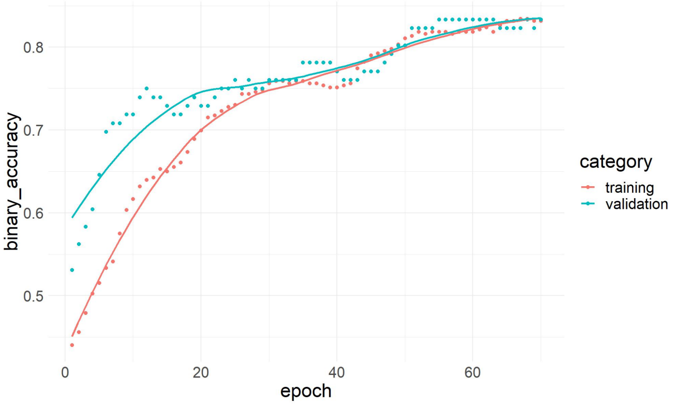

Figure 2.9: Accuracy trend over epochs.

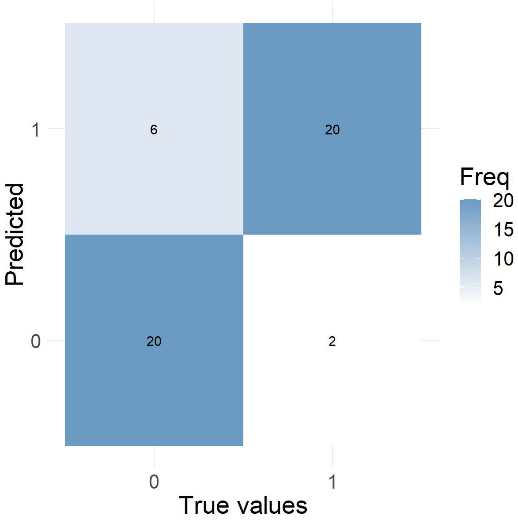

Figure 2.10: Confusion matrix of the neural network model.

|  | Loss | Accuracy |
|---|---|---|
| Training | 0.506 | 0.834 |
| Validation | 0.508 | 0.833 |

Table 2.6: Loss and accuracy on last epoch.

| Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|
| 0.833 | 0.167 | 0.909 | 0.909 | 0.839 |

Table 2.7: Performance metrics of the neural network model.

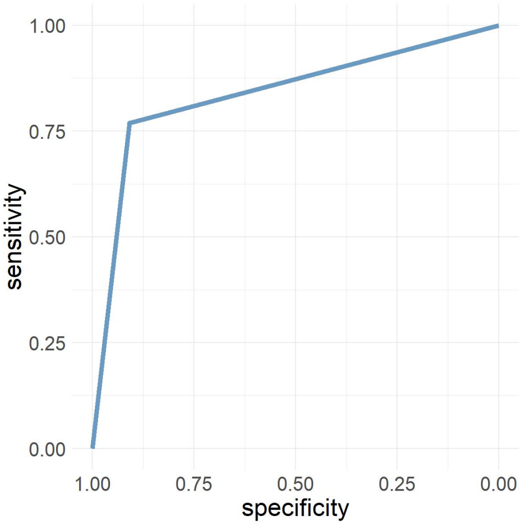

Figure 2.11: ROC curve of the neural network model.

# Chapter 3

# Results and conclusions

To evaluate the tested models the Table 3.1 can be considered.

| | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| Logistic regression | 0.771 | 0.229 | 0.818 | 0.826 | 0.774 |
| Random forest | 0.792 | 0.208 | 0.773 | 0.808 | 0.790 |
| Neural network | 0.833 | 0.167 | 0.909 | 0.909 | 0.839 |

Table 3.1: Performance metrics of all models.

As it can be seen the best model according to all metrics is the feedforward neural network.

So the choosen model was applied to the provided test set in order to predict the target variable. The results are available in the *data* folder of the GitHub repo [2].

### Some conclusions

In section 1.1.1 two questions arised to guide the analysis made. These questions can now be answered:

- How accurately can we predict if a patient is affected by any heart disease?

  Not so accurately. The best accuracy score is achieved by the neural network model with a value of 0.833. In this type of applications the 83.3% of accuracy is considered a bad result since only a very small error rate is tolerated when dealing with patients' health.

  This result can be due to the very small number of observations available in the dataset. This is a common issue in the medical framework since most of the existing data isn't publicly available due to privacy concerns.

  Another reason explaining this result can be the empirical diagnosis process at the base of the patients' labeling.

- Which characteristic of the patient has the most impact on his heart condition?

  Thanks to models like random forest and logistic regression the variables' importance can be analyzed.

Figure 2.7 (from random forest model) shows that the number of major vessels and the thalassemia are the two most important aspects that helps to distinguish healthy patients from sick ones.

Table 2.1 (from logistic regression model) confirms the importance of the number of major vessels and highlights a significant importance for the chest pain. In particular when a patient presents non-anginal pain the logistic regression model seems to perform better.

In the framework of extracting the most important features these two models can be used even though the feedforward neural network was choosen as the best model. This is due to the very similar scores obtained by the three models.

# Bibliography

[1] *Heart disease data set*, Kaggle. [Online]. Available: `https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility`.

[2] G. Spadaro, *Heart-attack-analysis*, GitHub, 2023. [Online]. Available: `https://github.com/Giovo17/heart-attack-analysis`.

[3] *Uci machine learning repository: Heart disease data set*, Uci.edu, 2019. [Online]. Available: `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`.

[4] M. Clinic, *How high blood pressure can affect your body*, Mayo Clinic, 2022. [Online]. Available: `https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868`.

[5] MedlinePlus, *Cholesterol levels: Medlineplus lab test information*, Medlineplus.gov, 2017. [Online]. Available: `https://medlineplus.gov/lab-tests/cholesterol-levels/`.

[6] *Loess: Local polynomial regression fitting*, rdrr.io. [Online]. Available: `https://rdrr.io/r/stats/loess.html`.