SICILIAE STVDIVM GENERALE

1434

# Project Report for Statistical Learning

## *Heart Attack dataset*

Authors Giuseppe Lombardia

Statistical Learning course

Prof. Salvatore Ingrassia

MSc in Data Science for Management

# Contents

**1 Introduction and Data exploration 2**

1.1 Data description . . . 2

1.2 Data Exploration . . . 3

1.2.1 Age . . . 4

1.2.2 Trestbps . . . 5

1.2.3 Cholesterol . . . 6

1.2.4 Thalach . . . 7

1.2.5 Oldpeak . . . 7

1.2.6 Target . . . 8

1.2.7 Sex . . . 8

1.2.8 Number of major vessels . . . 9

1.2.9 Thalassemia . . . 9

1.3 Multivariate analysis . . . 10

**2 Data modeling 10**

2.1 Logistic Regression . . . 10

2.2 Random Forest . . . 14

2.3 Neural Networks . . . 15

**3 Results and conclusions 18**

3.1 Applying the model to the test set . . . 19

**4 Conclusion 20**

# 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 disease. This dataset is of a multivariate type, meaning it involves a variety of distinct mathematical or statistical variables, suitable for multivariate numerical data analysis. It comprises 14 attributes, which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels, and Thalassemia. Although the database includes 76 attributes, published studies typically focus on a subset of 14. The Cleveland database is the only one that has been used by machine learning researchers to date. One of the primary tasks with this dataset is to predict whether a given patient has heart disease based on their attributes. Additionally, the dataset is used for experimental tasks to diagnose and uncover various insights that could enhance understanding of the problem.

## 1.1 Data description

By looking at the dataset on kaggle, the followings are the variables description:

1. *age*: age in years
2. *sex*: sex values: 0 for female and 1 for male
   (a) Value 0: female
   (b) Value 1: male
3. *cp* : chest pain type value:
   (a) Value 0: typical angina
   (b) Value 1: atypical angina
   (c) Value 2: non-anginal pain
   (d) Value 3: asymptomatic
4. *trestbps*: The person's resting blood pressure (mm Hg on admission to the hospital)
5. *chol*: The person's cholesterol measurement in mg/dl
6. *fbs*: fasting blood sugar: Values:
   (a) Value 0: fbs $\leq$ 120 mg/dl
   (b) Value 1: fbs $>$ 120 mg/dl
7. *thalach*: maximum heart rate achieved (measured in BPM)
8. *exang*: excercise induced angina. Values:
   (a) Value 0: no
   (b) Value 1: yes
9. *oldpeak*: ST depression induced by exercise relative to rest
10. *slope*: the slope of the peak exercise ST segment. Values:
    (a) Value 0: unsloping

(b) Value 1: flat

(c) Value 2: downsloping

11. *ca*: number of major vessels (0-3) colored by fluoroscopy.

12. *thal*: thalassemia. Values:

    (a) Value 1: normal (no blood flow in some part of the heart)

    (b) Value 2: fixed defect

    (c) Value 3: reversible defect (a blood flow is observed but it isn't normal).

13. *target*:

    (a) Value 0: absence of heart disease

    (b) Value 1: presence of heart disease

The following tables shows the first rows of the data. The dataset consists of 13 variables representing the health status of certain patients. These 13 variables will be used as X predictors in statistical models. The last column in question is the patient's health status, which is the response variable. The training data is composed of 180 observations.

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 68 | 1 | 2 | 118 | 277 | 0 | 1 | 151 | 0 | 1.00 | 2 | 1 | 3 | 1 |
| 2 | 54 | 0 | 1 | 132 | 288 | 1 | 0 | 159 | 1 | 0.00 | 2 | 1 | 2 | 1 |
| 3 | 42 | 0 | 2 | 120 | 209 | 0 | 1 | 173 | 0 | 0.00 | 1 | 0 | 2 | 1 |
| 4 | 63 | 0 | 1 | 140 | 195 | 0 | 1 | 179 | 0 | 0.00 | 2 | 2 | 2 | 1 |
| 5 | 62 | 1 | 1 | 128 | 208 | 1 | 0 | 140 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 6 | 59 | 1 | 0 | 135 | 234 | 0 | 1 | 161 | 0 | 0.50 | 1 | 0 | 3 | 1 |

Table 1: Training dataset head

## 1.2 Data Exploration

The starting point of the project is the **EDA** (**Exploratory Data Analysis**). We can start 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 the rows with the number of unique rows of the dataset, it has been found a duplicate row, which has been removed.

| | type | variable | factor.n_unique | factor.top_counts |
|---|---|---|---|---|
| 1 | factor | sex | 2 | 1: 125, 0: 55 |
| 2 | factor | cp | 4 | 0: 79, 2: 55, 1: 34, 3: 12 |
| 3 | factor | fbs | 2 | 0: 158, 1: 22 |
| 4 | factor | restecg | 3 | 1: 92, 0: 85, 2: 3 |
| 5 | factor | exang | 2 | 0: 122, 1: 58 |
| 6 | factor | slope | 3 | 2: 84, 1: 83, 0: 13 |
| 7 | factor | ca | 5 | 0: 104, 1: 42, 2: 21, 3: 11 |
| 8 | factor | thal | 3 | 2: 97, 3: 72, 1: 11 |
| 9 | factor | target | 2 | 1: 98, 0: 82 |

Table 2: Training dataset head (Categorical Variables)

|  | variable | complete_rate | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 |
|---|---|---|---|---|---|---|---|---|---|
| 10 | age | 1.00 | 54.04 | 8.79 | 29.00 | 48.00 | 55.00 | 60.00 | 76.00 |
| 11 | trestbps | 1.00 | 131.39 | 17.94 | 94.00 | 120.00 | 130.00 | 140.00 | 200.00 |
| 12 | chol | 1.00 | 241.47 | 43.69 | 131.00 | 207.00 | 239.00 | 270.25 | 342.00 |
| 13 | thalach | 1.00 | 149.43 | 23.14 | 71.00 | 132.00 | 153.00 | 165.00 | 202.00 |
| 14 | oldpeak | 1.00 | 1.02 | 1.14 | 0.00 | 0.00 | 0.60 | 1.80 | 5.60 |

Table 3: Training dataset head (Numerical Variables)

### 1.2.1 Age

Examining the histogram on the left, it appears there are two peaks, which might indicate the data comes from two distinct populations. This idea is further supported by the histogram on the right, which shows the age distribution of patients according to their diagnosis.

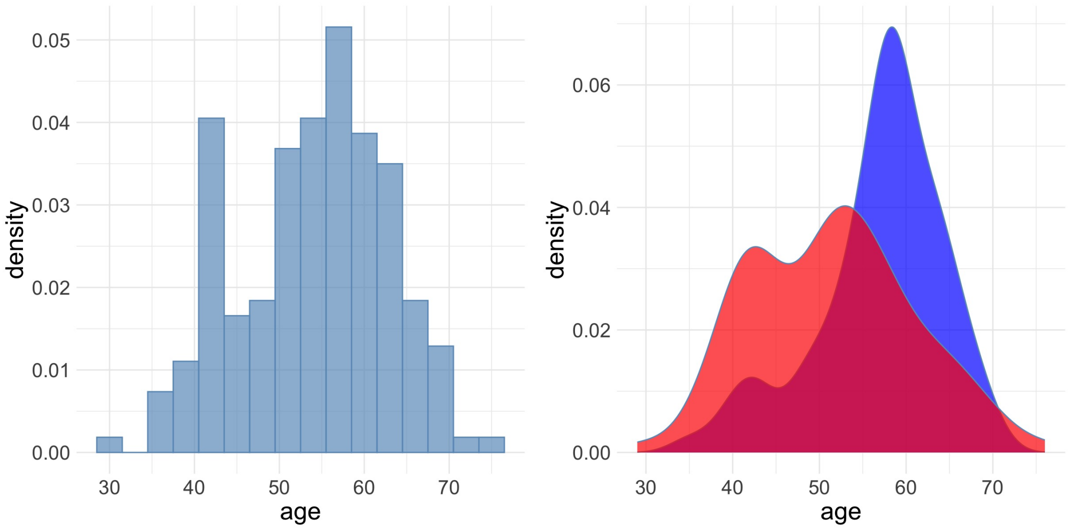

Figure 1: Left: histogram of age. Right: Histogram of age with target hue. Blue for healthy patient and red for sick ones

### 1.2.2 Trestbps

Trestbps stands for ”resting blood pressure” and is a common parameter used in medical studies and health assessments. It refers to the blood pressure (both systolic and diastolic) measured while a person is at rest, typically in a seated or lying position. This measurement is an important indicator of cardiovascular health. The skewness value of 0.9633282 indicates a positive skew, meaning that the distribution of resting blood pressure values is not symmetric and has a longer right tail.

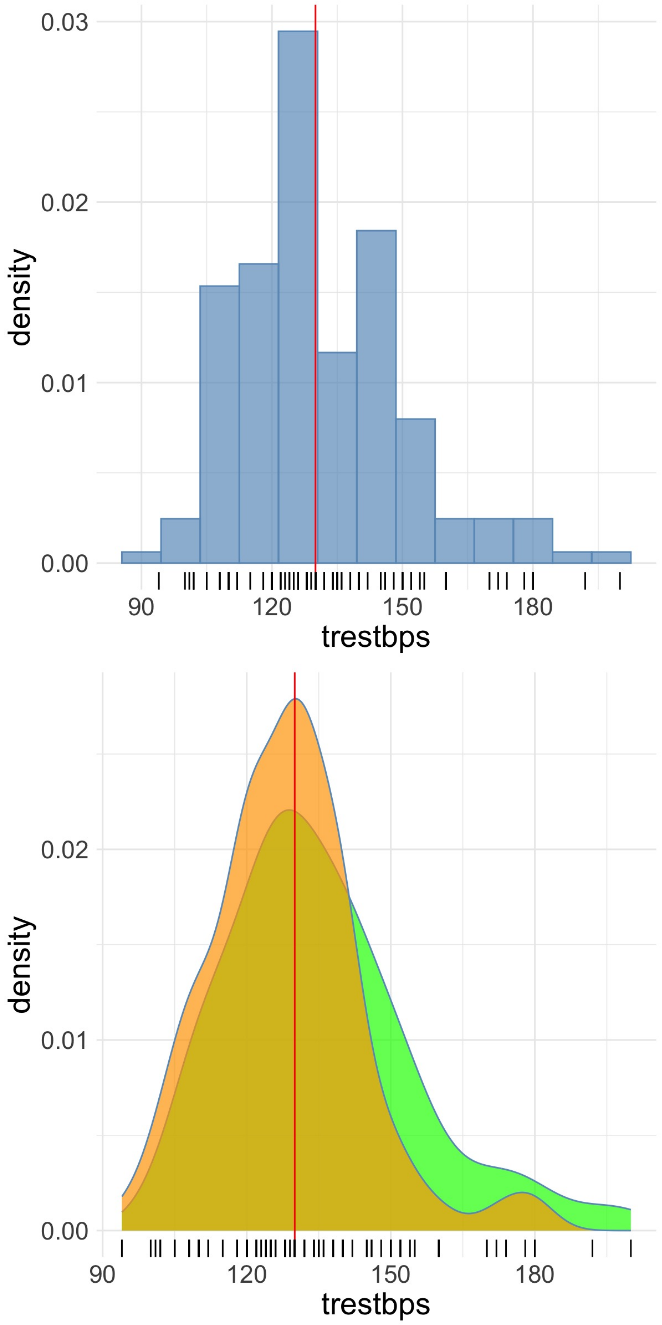

Figure 2: Left: histogram of trestbps. Right: Histogram of trestbps with target hue. Green for healthy patient and Orange for sick ones

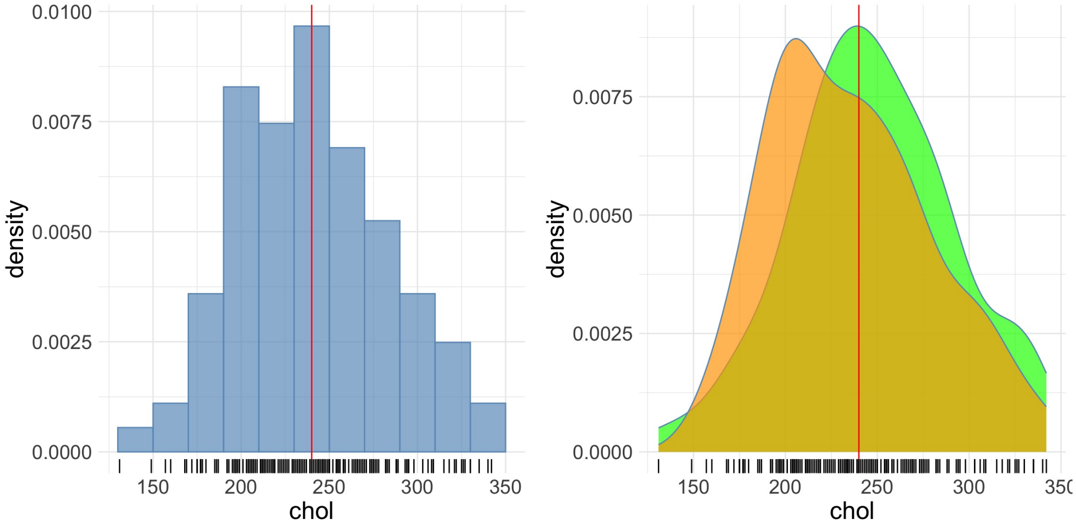

Figure 3: Left: histogram of chol. Right: Histogram of chol with target hue. Green for healthy patient and Orange for sick ones

### 1.2.3 Cholesterol

High levels of cholesterol could be the cause of developing heart disease. Values of 240 mg / dl are considered high. The overall distribution is simmetric.

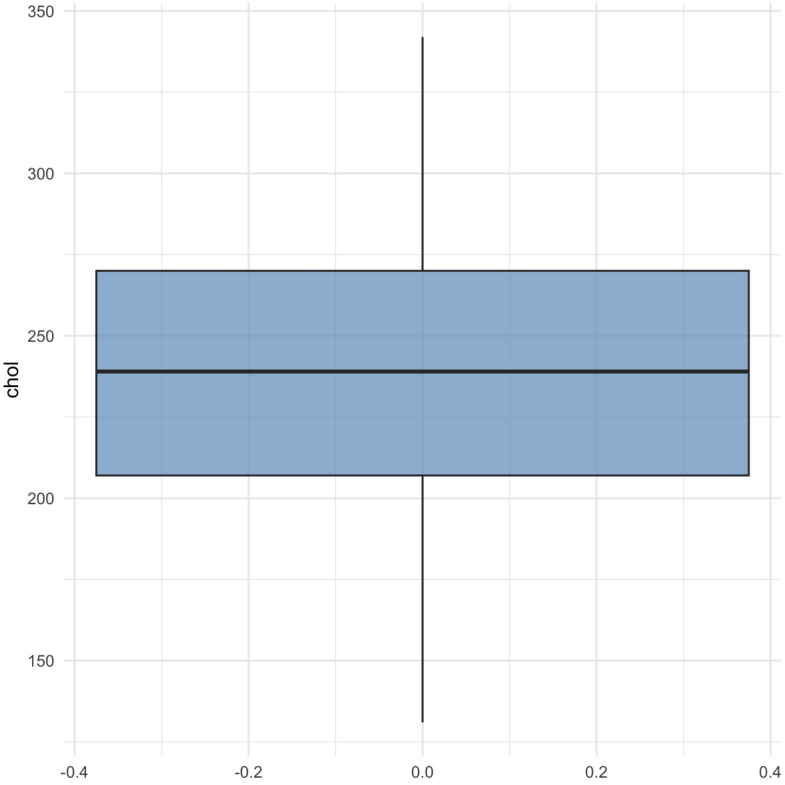

Figure 4: box plot of cholesterol

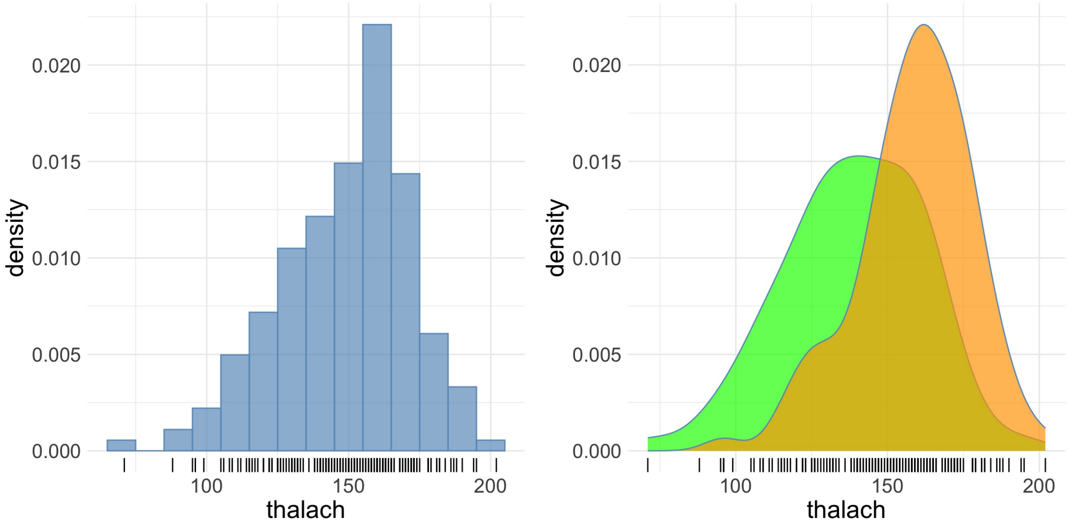

Figure 5: Left: histogram of thalach . Right: Histogram of thalach with target hue. Green for healthy patient and Orange for sick ones

### 1.2.4 Thalach

Thalach stands for ***maximum heart rate achieved*** during a stress test or exercise test. It is a common parameter used in cardiology to assess cardiovascular fitness and the heart's response to physical stress.

### 1.2.5 Oldpeak

"Oldpeak" refers to the ST depression induced by exercise relative to rest. It is a measure obtained from an electrocardiogram (ECG or EKG) and is used to assess the severity of heart disease. Specifically, it quantifies the depression in the ST segment of the ECG during peak exercise compared to rest. The overall distribution is significantly skewed with a coefficient equal to 1.129694.

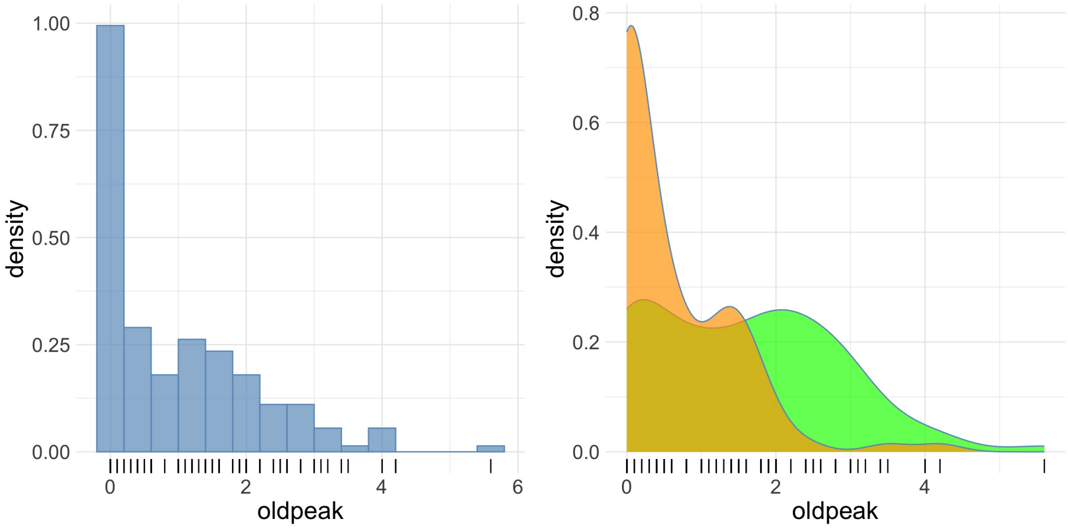

Figure 6: Left: histogram of oldpeak . Right: Histogram of oldpeak with target hue. Green for healthy patient and Orange for sick ones

### 1.2.6 Target

When performing a classification task, you need to check the percentage of each class in the data set. As shown in Figure 7, the ratio of each class is more or less the same,so the data set is balanced according to the response variables. This allows the model to be bias-free in classifying patients according to the presence of heart disease.

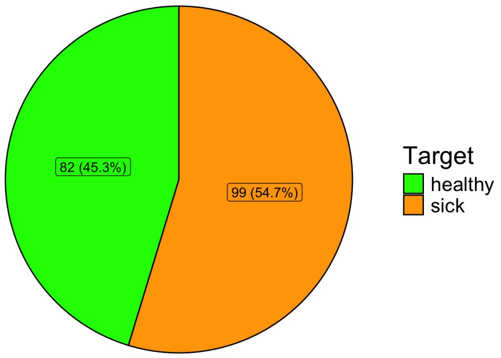

Figure 7: Pie chart of target

### 1.2.7 Sex

I apply the same reasoning to another categorical variable. It should be said that the data set is unbalanced in this categorical variable since about 2/3 are male patients. It could be the case that classifiers may develop a bias in distinguishing a male patient's health status better than a female patient's health status.

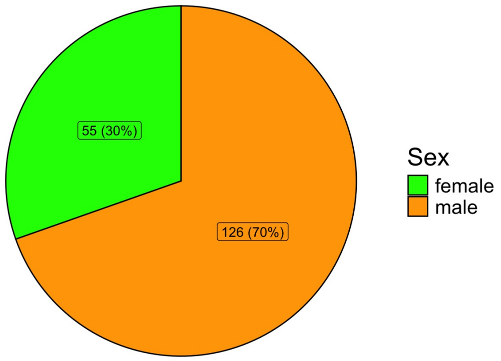

Figure 8: Pie chart of sex

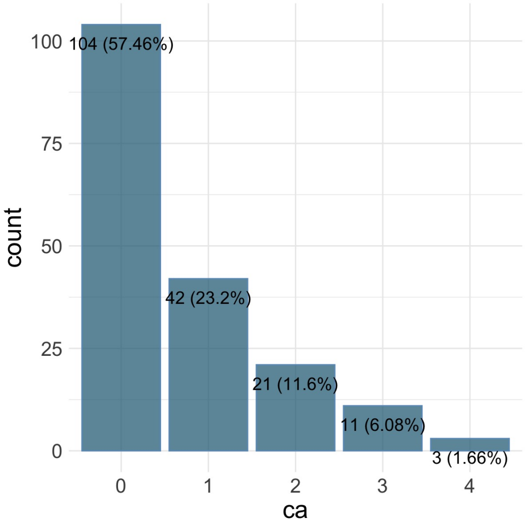

Figure 9: Barplot of ca

### 1.2.8 Number of major vessels

As shown by Figure 9, there are only 3 observations with registered value of ca equal to 4. So, 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 42 | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0.00 | 2 | 4 | 2 | 1 |
| 64 | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0.00 | 2 | 4 | 2 | 1 |
| 120 | 43 | 1 | 0 | 132 | 247 | 1 | 0 | 143 | 1 | 0.10 | 1 | 4 | 3 | 0 |

Table 4: Rows of dataset with ca = 4

### 1.2.9 Thalassemia

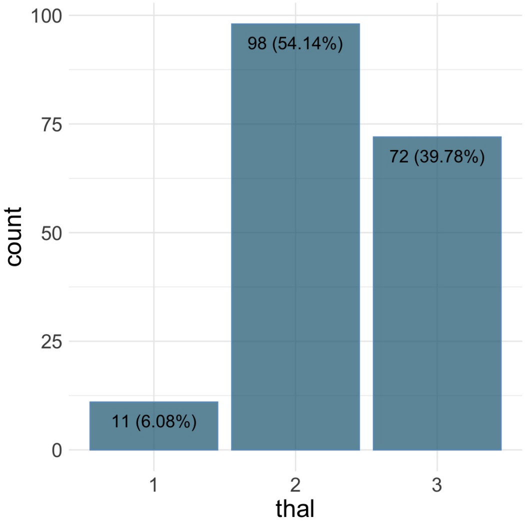

Figure 10: Barplot of thalassemia

### 1.3 Multivariate analysis

Now, the focus will be on analyzing the numerical variables between them. Therefore, we proceed with the exploration of possible relationships between them.

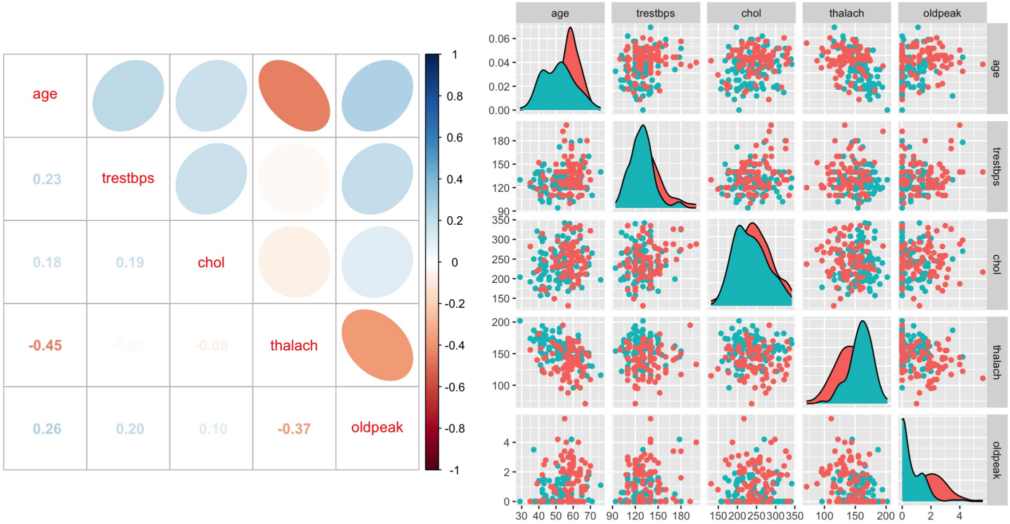

Figure 11: Left: correlation matrix of numerical variables. Right: Matrix of pairwise scatterplots of numerical variables. Red dots represent healthy patients and blue dots represent sick patients

The correlation matrix shows the variables that are mostly uncorrelated between them. **Age** has a slight negative correlation with **thalach**, which is normal, because the heart usually has more difficulty pumping blood as the patient gets older and also has a slight positive correlation with **oldpeak**. In the Figure 11 Right shows the matrix of the scatterplot considering all possible pairs of the numerical variables. In the diagonal we have the densities of the variables and in the off-diagonal we have the pairwise scatterplots. In the graphs it can be noted that there are almost no visible groups in the dataset and also the response variable does not help in any other kind of classification. There is only one exception in the scatterplot of *age-thalach*, where there is a high concentration of blue dots in the upper left corner of the graph and an high concentration of red dots in the bottom right corner. So, the data set is balanced so that the models that are trained without any worries possible classification biases. The univariate and bivariate analysis don't show good results in group separation between sick and healthy patients.

## 2 Data modeling

In this section, three models will be analyzed in order to predict the response variable. The assigned data are divided into three sets: a training set (60%), a validation set (20%), and a test set (20%). The models that will be tested are logistic regression, the random forest and the neural networks.

### 2.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks. This approach utilizes the logistic (or sigmoid) function to transform a linear combination of input features into a probability value ranging between 0 and 1. The sigmoid function is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

where $\sigma(x)$ represents the sigmoid function applied to the input $x$.

This probability indicates the likelihood that a given input corresponds to one of two predefined categories. The formula for logistic regression is as follows:

$$P(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}$$

where $P(Y = 1|\mathbf{X})$ represents the probability that the output $Y$ is 1 given the input vector $\mathbf{X}$, $\beta_0$ is the intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the explanatory variables $X_1, X_2, \ldots, X_p$. The first logistic regression model has been fitted using all available variables.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | * |
|---:|---:|---:|---:|---:|---:|
| (Intercept) | 4.29 | 4.13 | 1.04 | 0.30 | |
| age | -0.02 | 0.04 | -0.60 | 0.55 | |
| sex1 | -2.43 | 0.87 | -2.81 | 0.01 | ** |
| cp1 | 0.85 | 0.73 | 1.17 | 0.24 | |
| cp2 | 2.46 | 0.74 | 3.34 | 0.00 | *** |
| cp3 | 3.95 | 1.16 | 3.40 | 0.00 | *** |
| trestbps | -0.04 | 0.02 | -2.39 | 0.02 | * |
| chol | -0.00 | 0.01 | -0.47 | 0.64 | |
| fbs1 | 1.35 | 0.90 | 1.50 | 0.13 | |
| restecg1 | 1.18 | 0.59 | 2.00 | 0.05 | * |
| restecg2 | 0.63 | 2.52 | 0.25 | 0.80 | |
| thalach | 0.03 | 0.02 | 1.62 | 0.11 | |
| exang1 | -1.04 | 0.62 | -1.69 | 0.09 | . |
| oldpeak | -0.12 | 0.33 | -0.38 | 0.71 | |
| slope1 | -0.42 | 1.13 | -0.37 | 0.71 | |
| slope2 | 0.82 | 1.22 | 0.67 | 0.50 | |
| ca1 | -1.84 | 0.67 | -2.76 | 0.01 | ** |
| ca2 | -3.12 | 1.22 | -2.56 | 0.01 | * |
| ca3 | -1.66 | 1.15 | -1.44 | 0.15 | |
| thal2 | -0.47 | 1.19 | -0.39 | 0.69 | |
| thal3 | -1.55 | 1.13 | -1.37 | 0.17 | |

Table 5: Coefficient estimates of logistic regression model using all variables

Table 5 shows the result of the fit. Almost all predictors are not statistically significant because they have a p-value higher than 0.05. The only variables that are statistically significant are sex, cp, trestbps, restecg1 and ca. The **AIC** of the model is equal to 147.44.

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

| | AIC | Variables selection |
|---|---:|---:|
| Model 1 | 147.44 | all variables |
| Model 2 | 145.59 | age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + slope + ca + thal |
| Model 3 | 143.84 | age + sex + cp + trestbps + fbs + restecg + thalach + exang + slope + ca + thal |
| Model 4 | 142.24 | sex + cp + trestbps + fbs + restecg + thalach + exang + slope + ca + thal |

Table 6: Ranking of logistic regression according to AIC

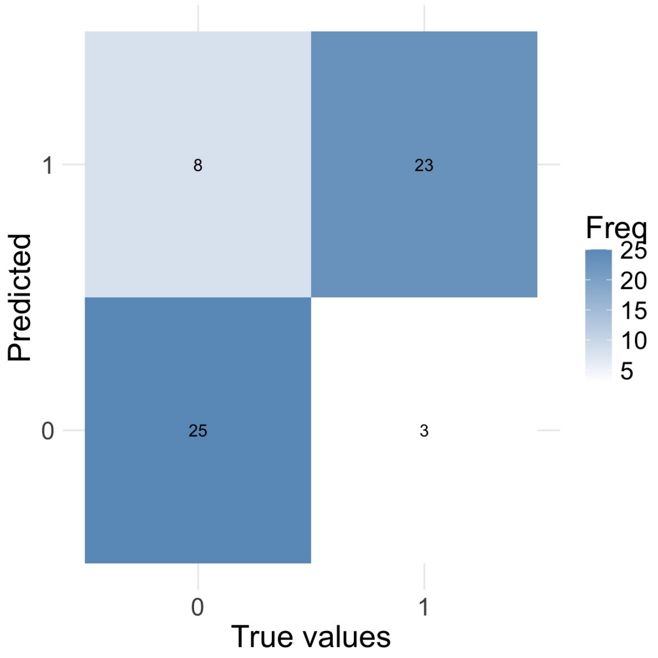

Figure 12: Correlation matrix of logistic regression model using all predictors

Let us now validate the model using the validation set.

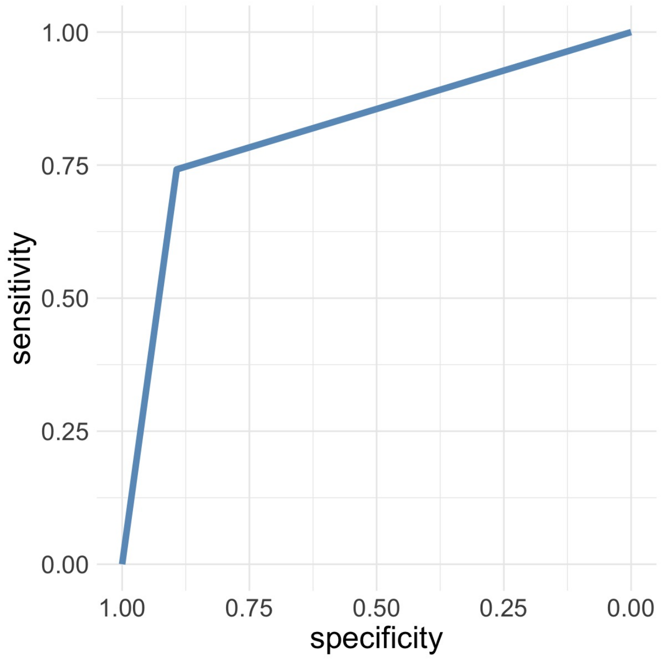

Figure 13: Roc Curve of logistic regression model using all predictors

Table 7 shows the performance of the regression model with best predictors according to AIC, which uses the following variables: *sex, cp, trestbps, fbs, restecg, thalach, exang, slope, ca, thal.* This model performers quite

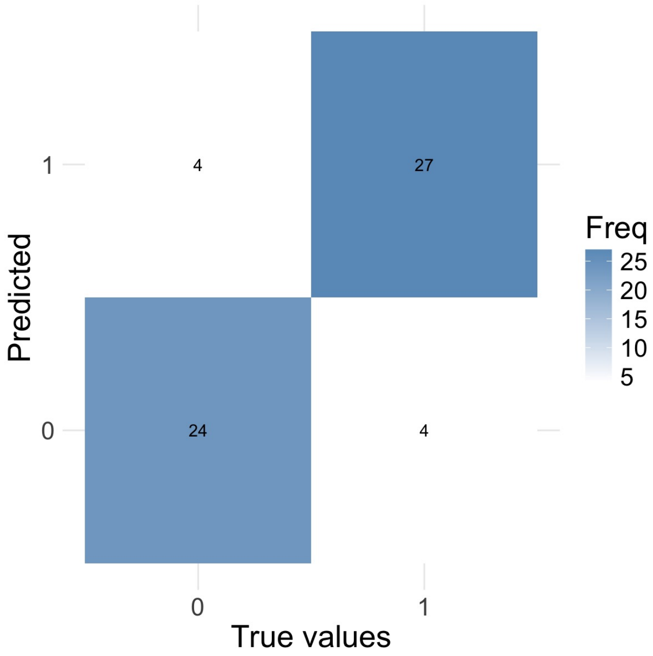

Figure 14: Correlation matrix of the model with best predictors

similarly to the previous one.

| | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| all predictors | 0.814 | 0.186 | 0.893 | 0.885 | 0.817 |
| best predictors | 0.864 | 0.136 | 0.857 | 0.871 | 0.864 |

Table 7: Performance metrics of logistic regression models

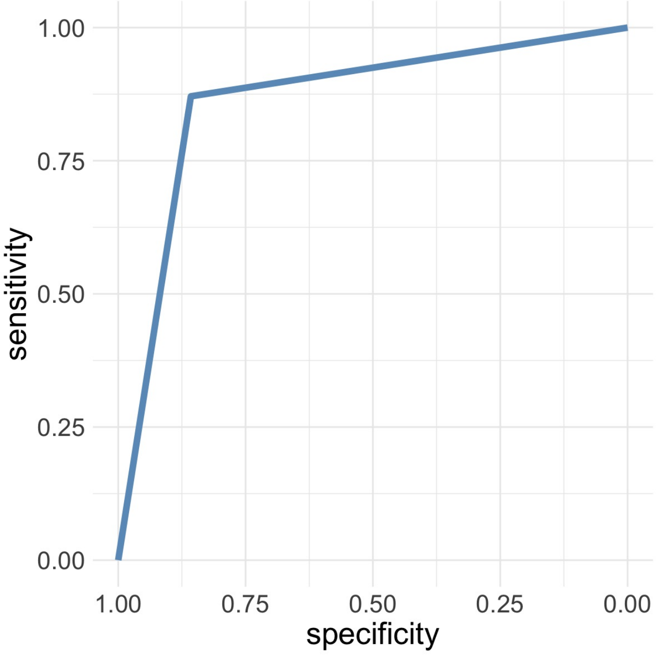

Figure 15: Roc curves of the model with best predictors

## 2.2 Random Forest

The random forest is a statistical ensemble method based on a decision tree. Usually, the number of features chosen for each split is $\sqrt{p}$, p being the number of available features. This data set has $p = 13$ features. In this case, 4 features has been chosen as the dimension of the feature subset each tree has access to for each split. The data set is divided into three parts, one for training, one for validation, and the last for testing. Due to how bagging works (each tree works with a bootstrapped version of the training data), there is another set of validation metrics which takes into account the OOB (Out-Of-Bag) observations.

| | 0 | 1 | class.error |
|---|---|---|---|
| 0 | 60 | 22 | 0.27 |
| 1 | 14 | 85 | 0.14 |

Table 8: Confusion matrix of the random forest model with OOB observations

Looking at the Table 8, the confusion matrix made of the predicted OOB observations has a 27% of error rate for the first class and 14% of error rate for the second class. The total OOB estimate rate of error is about 20%. Now we can observe the validation metrics,

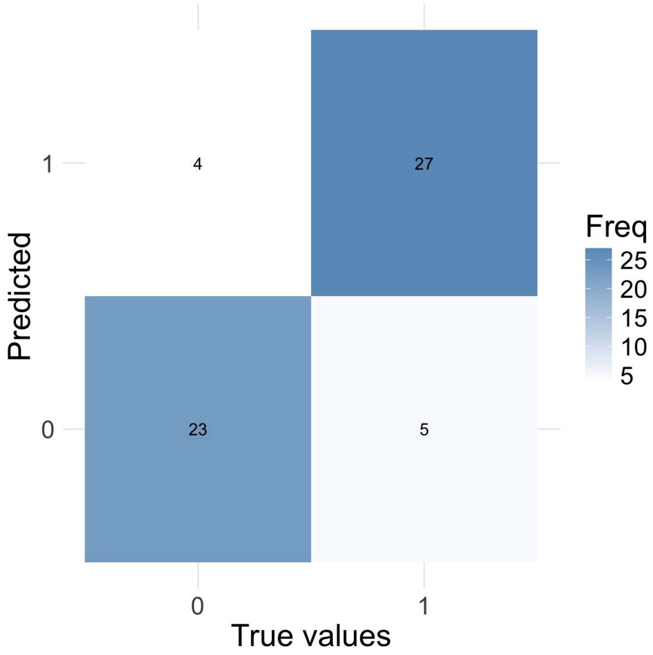

Figure 16: Confusion matrix of the random forest model

By comparing it with logistic regression with all predictors, the random forest performs a little better. An important characteristic of the random forest, being a tree-based method, is the possibility to compute a coefficient of importance for every feature used during the training. This is possible thanks to looking at how much the GINI index reduces by splitting the feature space using a certain predictor. Figure 18 shows the importance of the predictors' characteristics according to 2 metrics: the mean decreases the accuracy and the mean decreases GINI. The mean decrease in accuracy represents how much the accuracy in the OOB set decreases on average by excluding that predictor. The mean decrease Gini represents how much a feature decreases the node impurity, that is determined by the Gini index.

| | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| 1 | 0.839 | 0.161 | 0.828 | 0.848 | 0.838 |

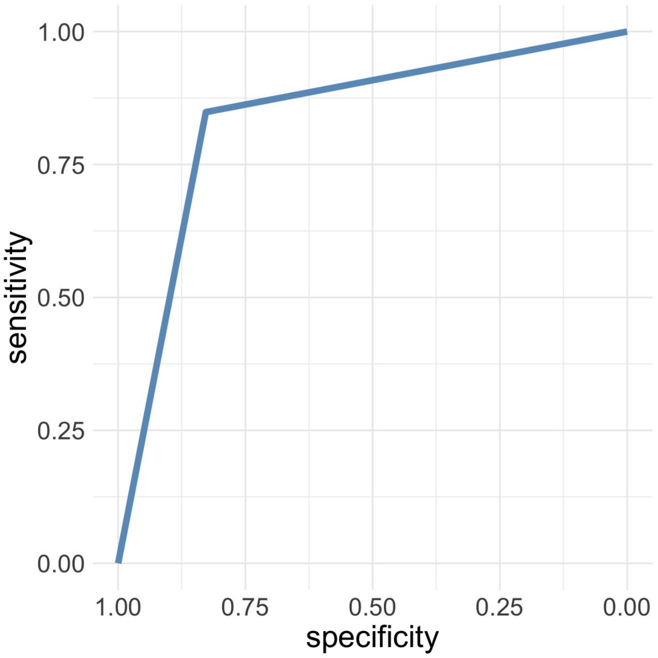

Figure 17: ROC curve of the random forest model

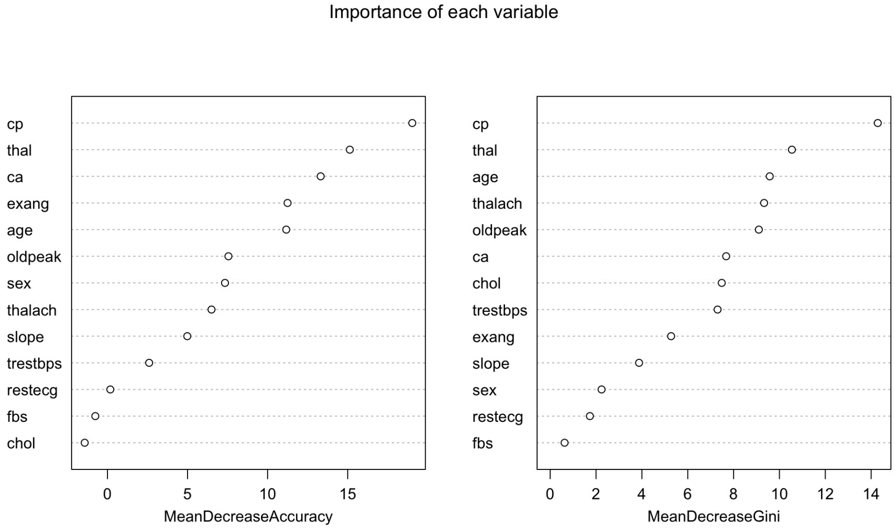

Figure 18: Variable importance graph

## 2.3 Neural Networks

The final model applied to this classification task is the neural network. Neural networks are sophisticated machine learning models composed of multiple layers of interconnected nodes, known as neurons. Each neuron functions as a processing unit that computes an output based on its input values through an activation function.

A typical neural network architecture includes the following.

1. **Input Layer**: This layer consists of neurons representing the input features of the dataset.

2. **Hidden Layers**: One or more layers of neurons between the input and output layers. Each neuron in a hidden layer receives input from the neurons of the previous layer, processes it using an activation function, and passes the output to the neurons of the next layer.

3. **Output Layer**: This layer consists of neurons that provide the final prediction or classification result.

Each neuron computes a weighted sum of its inputs, adds a bias term, and applies an activation function to produce its output. Mathematically, the output $z_i$ of neuron $i$ in a layer can be expressed as:

$$z_i = \phi\left(\sum_{j=1}^{n} w_{ij}x_j + b_i\right)$$

where:

- $\phi$ is the activation function.
- $w_{ij}$ represents the weight connecting the $j$-th input to the $i$-th neuron.
- $x_j$ is the $j$-th input.
- $b_i$ is the bias term for the $i$-th neuron.

The Rectified Linear Unit (ReLU) is a common choice for the activation function, defined as:

$$\phi(z) = \max(0, z)$$

In the context of neural networks, the output of a single layer can be represented as a vectorized operation:

$$\mathbf{z} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$$

where:

- $\mathbf{z}$ is the vector of outputs for the layer.
- $\mathbf{W}$ is the matrix of weights.
- $\mathbf{x}$ is the input vector.
- $\mathbf{b}$ is the bias vector.

The network learns by adjusting the weights and biases to minimize a loss function, often through an optimization algorithm such as stochastic gradient descent (SGD). The loss function quantifies the difference between the predicted outputs and the actual target values. For classification tasks, the cross-entropy loss is frequently used, defined as:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

where:

- $y_i$ is the actual label for the $i$-th sample.

- $\hat{y}_i$ is the predicted probability for the $i$-th sample.
- $n$ is the number of samples.

By iterative adjustment of the weights and biases to minimize this loss, the neural network improves its performance in the classification task. The final network has 2 hidden layers, with 7 neurons in the first layer and 4 neurons in the second one (both layers use ReLU as activation function), and an output layer with 2 neurons, one for each possible outcome, governed by the logistic activate function. 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.

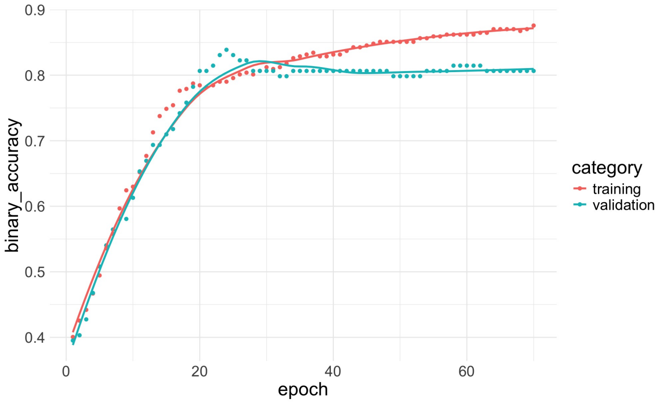

Figure 19: Accuracy over epochs

Figure 19 and 20 show loss, accuracy and their respective trends. The behavior is stable and the training and validation curves follow the same trend, but the final results are quite poor. The table

| | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| 1 | 0.806 | 0.194 | 0.828 | 0.839 | 0.808 |

Table 9: Performance metrics of neural network model

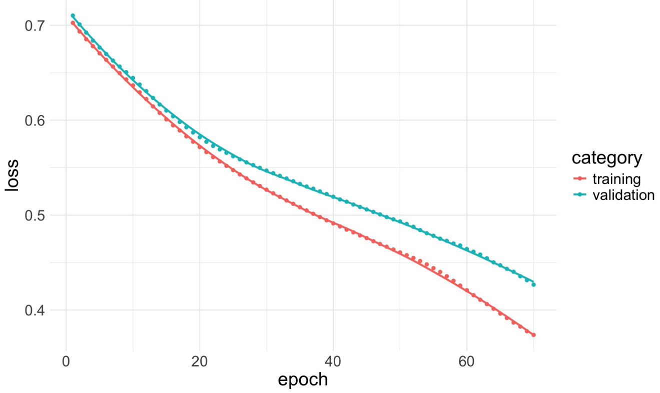

Figure 20: Loss over epochs

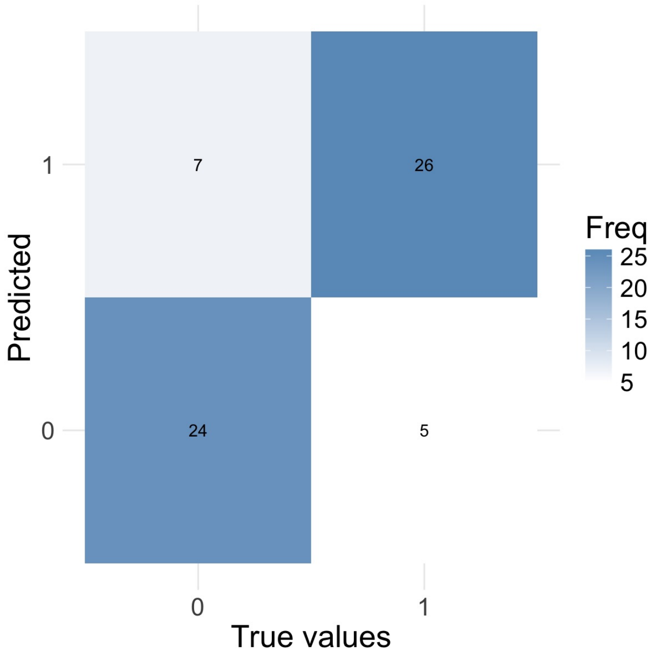

Figure 21: Confusion matrix of neural network model

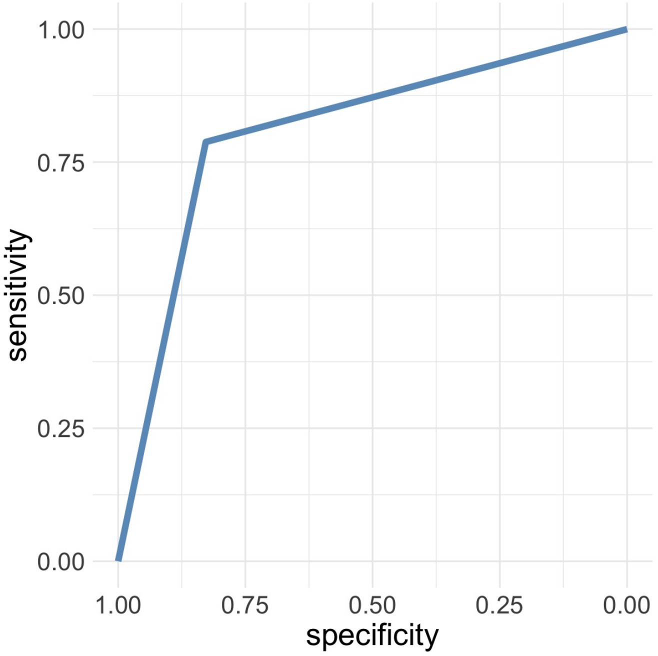

Figure 22: ROC curve of neural network model

# 3 Results and conclusions

To evaluate the tested models we can consider the following table.

To determine the best model among Logistic Regression, Random Forest, and Neural Network, we compare their performance metrics: Accuracy, Error Rate, Specificity, Sensitivity, and AUC. Below is a detailed explanation of each metric and how it contributes to evaluating the model:

| | Accuracy | Error rate | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.814 | 0.186 | 0.893 | 0.885 | 0.817 |
| Random Forest | 0.839 | 0.161 | 0.828 | 0.848 | 0.838 |
| Neural Network | 0.806 | 0.194 | 0.828 | 0.839 | 0.808 |

Table 10: Performance metrics of all models

- **Accuracy**: The proportion of correctly classified instances.
- **Error Rate**: The proportion of incorrectly classified instances.
- **Specificity**: The proportion of true negatives out of all negatives (ability to correctly identify negative classes).
- **Sensitivity**: The proportion of true positives out of all positives (ability to correctly identify positive classes).
- **AUC (Area Under the Curve)**: A comprehensive measure of model performance, considering both sensitivity and specificity.

1. **Accuracy**:

- **Random Forest** has the highest accuracy (0.839), followed by Logistic Regression (0.814) and Neural Network (0.806).

2. **Error Rate**:

- **Random Forest** has the lowest error rate (0.161), followed by Logistic Regression (0.186) and Neural Network (0.194).

3. **Specificity**:

- **Logistic Regression** has the highest specificity (0.893), followed by both Random Forest and Neural Network (0.828).

4. **Sensitivity**:

- **Logistic Regression** has the highest sensitivity (0.885), followed by Random Forest (0.848) and Neural Network (0.839).

5. **AUC**:

- **Random Forest** has the highest AUC value (0.838), followed by Logistic Regression (0.817) and Neural Network (0.808).

The **Random Forest** model emerges as the best model based on most of the metrics considered, particularly in terms of precision, error rate, and AUC. Although Logistic Regression shows slightly higher values in Specificity and Sensitivity, the Random Forest model offers a better overall balance and superior performance in terms of AUC, which is a key metric for evaluating the overall performance of the model.

## 3.1 Applying the model to the test set

Here it is shown some results on the test set are shown.

|  | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | Predicted |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 64 | 0 | 2 | 140 | 313 | 0 | 1 | 133 | 0 | 0.20 | 2 | 0 | 3 | 1 |
| 2 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.60 | 2 | 0 | 2 | 1 |
| 3 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.30 | 0 | 0 | 1 | 1 |
| 4 | 44 | 1 | 2 | 140 | 235 | 0 | 0 | 180 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 5 | 47 | 1 | 2 | 138 | 257 | 0 | 0 | 156 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 6 | 67 | 0 | 2 | 152 | 277 | 0 | 1 | 172 | 0 | 0.00 | 2 | 1 | 2 | 1 |
| 7 | 53 | 1 | 2 | 130 | 246 | 1 | 0 | 173 | 0 | 0.00 | 2 | 3 | 2 | 1 |
| 8 | 57 | 1 | 2 | 150 | 126 | 1 | 1 | 173 | 0 | 0.20 | 2 | 1 | 3 | 0 |
| 9 | 45 | 1 | 1 | 128 | 308 | 0 | 0 | 170 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 10 | 43 | 1 | 0 | 110 | 211 | 0 | 1 | 161 | 0 | 0.00 | 2 | 0 | 3 | 1 |
| 11 | 44 | 1 | 2 | 120 | 226 | 0 | 1 | 169 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 12 | 74 | 0 | 1 | 120 | 269 | 0 | 0 | 121 | 1 | 0.20 | 2 | 1 | 2 | 1 |
| 13 | 60 | 0 | 3 | 150 | 240 | 0 | 1 | 171 | 0 | 0.90 | 2 | 0 | 2 | 1 |
| 14 | 34 | 1 | 3 | 118 | 182 | 0 | 0 | 174 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 15 | 39 | 0 | 2 | 94 | 199 | 0 | 1 | 179 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 16 | 66 | 1 | 0 | 160 | 228 | 0 | 0 | 138 | 0 | 2.30 | 2 | 0 | 1 | 0 |
| 17 | 65 | 0 | 2 | 140 | 417 | 1 | 0 | 157 | 0 | 0.80 | 2 | 1 | 2 | 1 |
| 18 | 44 | 1 | 1 | 120 | 220 | 0 | 1 | 170 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 19 | 54 | 0 | 2 | 135 | 304 | 1 | 1 | 170 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 20 | 58 | 0 | 3 | 150 | 283 | 1 | 0 | 162 | 0 | 1.00 | 2 | 0 | 2 | 1 |
| 21 | 34 | 0 | 1 | 118 | 210 | 0 | 1 | 192 | 0 | 0.70 | 2 | 0 | 2 | 1 |
| 22 | 54 | 0 | 2 | 108 | 267 | 0 | 0 | 167 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 23 | 46 | 0 | 0 | 138 | 243 | 0 | 0 | 152 | 1 | 0.00 | 1 | 0 | 2 | 0 |
| 24 | 55 | 1 | 1 | 130 | 262 | 0 | 1 | 155 | 0 | 0.00 | 2 | 0 | 2 | 1 |
| 25 | 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1.00 | 1 | 0 | 2 | 1 |
| 26 | 52 | 1 | 2 | 138 | 223 | 0 | 1 | 169 | 0 | 0.00 | 2 | 4 | 2 | 1 |
| 27 | 48 | 1 | 1 | 130 | 245 | 0 | 0 | 180 | 0 | 0.20 | 1 | 0 | 2 | 1 |
| 28 | 71 | 0 | 0 | 112 | 149 | 0 | 1 | 125 | 0 | 1.60 | 1 | 0 | 2 | 1 |
| 29 | 66 | 0 | 3 | 150 | 226 | 0 | 1 | 114 | 0 | 2.60 | 0 | 0 | 2 | 1 |
| 30 | 56 | 1 | 1 | 120 | 240 | 0 | 1 | 169 | 0 | 0.00 | 0 | 0 | 2 | 1 |

## 4 Conclusion

Is it possible to predict if a patient is affected by heart disease? It is very difficult because the best accuracy is 84%. In this type of applications the accuracy of 84% is a bad result. This result may be due to the very small number of observations available in the dataset. This is a common problem in the medical setting, as most existing data is not publicly available due to privacy concerns. But another question is which characteristics of the patient have the greatest impact on his heart condition? Thanks to models such as random forest and logistic regression, the importance of variables can be analyzed. Figure 18 (from random forest). The plots show the importance of the variables for the Random Forest model used. There are two plots, each showing different measures of variable importance. MeanDecreaseAccuracy and MeanDecreaseGini. The MeanDecreaseAccuracy plot (left) indicates the reduction in model accuracy when each variable is excluded, with higher values representing more important variables. The MeanDecreaseGini plot (right) shows the decrease in Gini impurity when a variable is used for splitting, with higher values indicating greater importance for node purity.

From the MeanDecreaseAccuracy plot, the most important variables are ”cp” (chest pain type), followed by ”thal” (thalassemia) and ”ca” (number of major vessels colored by fluoroscopy). Other significant variables include ”exang” (exercise induced angina), ”age”, and ”oldpeak” (ST depression induced by exercise relative to rest). Similarly, the MeanDecreaseGini plot highlights ”cp” as the most important variable, followed by ”thal”, ”age”, and ”thalach” (maximum heart rate achieved). Additional important variables are ”oldpeak”, ”ca”, and ”chol” (serum cholesterol).

In conclusion, the most important variables to consider, based on both metrics, are ”cp”, ”thal”, ”ca”, ”age”, and ”oldpeak”, as they have a significant impact on both model accuracy and node purity. These variables should be prioritized for model interpretation and further analysis.