HarvardX Data Science Part 9: Own Project - Heart Disease Prediction

Joerg Schors

Contents

Final version **10.12.2024**

1 Executive Summary

This report describes the development of a Machine Learning (ML) model for a data set freely available with 1190 observations containing 11 different variables and one "result" called *target*. The *target* states, if the observation (i.e. a person) suffers from heart disease or not. The variables are a combination of continuous and nominal (factorial) values most probably taken during a medical examination with exercise and in combination with the record of a Electrocardiogram (ECG).

The data set was thoroughly inspected with respect to missing or non plausible values, which have been corrected to a certain extent. Data analysis has been performed to find correlations or patterns in the combination of different variables. Stratification has been used to reduce complexity of one variable (*age*).

To find patterns, a large set of graphical representations of the data in different combination has been created, but has not lead to a definite conclusion which approach for the ML model would be convenient.

At least five different models, one simply by guessing, k-Nearest Neighbors, two generalized linear models (glm & GAM loess) and a Random Forest algorithm (two different models) have been evaluated on the basis of the accuracy as criterium for the best model.

The conclusion is, that a Random Forest model (*rf*) with optimized parameters gives the highest accuracy of about 0.91, while *cforest* gives an accuracy of only 0.86.

All steps will be described in detail in the following sections.

2 Introduction

This is the own ML project to complete the HarvardX Data Science Course Series. After a long research (also on different governmental data platforms) which data could I use as a final project I choose a dataset from *www.kaggle.com*, because I can not use data from my own business.

The data analysis and model building was done using R [1] with RStudio [2] as IDE using techniques and approaches as described in the course script [3]. Reference [4] was used for additional information.

3 Description of the Data

3.1 Data set information

The data set is available for download from *www.kaggle.com* (see: *https://www.kaggle.com/datasets/mexwell/heart-disease-dataset*). It consists of eleven different attributes that may possibly have influence on a heart disease and the health status (i.e. heart disease or not, labeled as *target*).

The provider of the data stated:

This heart disease dataset is curated by combining five popular heart disease datasets already available independently but not combined before. In this dataset, the five heart datasets are combined which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- *Cleveland*
- *Hungarian*
- *Switzerland*
- *Long Beach VA*
- *Statlog (Heart) Data Set*

3.2 Detailed data description

The detailed description of the variables is given below:

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

A more detailed description of the "nominal" values, i.e. categorial values is given in the table below. The meaning, i.e. background and importance with respect to heart disease can be found in literature (see for example references [5] to [7]). Some of these aspects will be described below with respect to the data analysis and model building.

Attribute	Description
Sex	"1 = male / 0= female"
Chest Pain Type	– Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic
Fasting Blood sugar	> 120 mg/dl (1 = true; 0 = false)
Resting electrocardiogram results	– 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
Exercise induced angina	"1 = yes / 0 = no"
the slope of the peak exercise ST segment	– Value 1: upsloping – Value 2: flat – Value 3: downsloping
class	1 = heart disease, 0 = Normal

4 Project Goal

The goal of this project is to set up a Machine Learning (ML) model convenient and accurate to predict heart disease based on the attributes available in this data set.

5 General Remarks

The essential steps consist in general of the following tasks, which may vary in extent based on the problem and data available:

1. Download / read the data (may be organized in several files and different formats)
2. Reformat the data in a way to have a structured data set, which then can be transformed into tidy or similar data format
3. Explanatory Data Analysis to get an impression of the accuracy, limits, content and faulty entries using statistics and graphical representation
4. Data wrangling, i.e. filter, remove or replace content not usable or not reliable, reshape data or reduce dimension
5. Split up the data into one ore more training and test sets to check the validity of the selected model(s) or combination of models
6. Depending on the problem to solve and the type of data (numerical, factors) select at least one ore more ML models to fit the data
7. Train the model(s) and test on the appropriate data sets
8. Present the final result (i.e. accuracy, confusion matrix etc.)
9. Conclusion

It is essential to understand the meaning of the different parameters and their correlations not only on a statistical basis, but in the sense of causality and inference. A good approach has to incorporate reasoning and knowledge of the data. It has to be kept in mind, that the data is not simply a set of numbers to deal with.

6 Data Analysis, preprocessing and data wrangling

6.1 Loading the local file

Downloading the file directly from *kaggle* needs the use of user credentials to log in before access to the data is possible. To solve this, the data file is provided with all other files (code, report etc.) required in a .csv format and will be read from the actual directory. The file is named *heart_statlog_cleveland_hungary_final.csv*.

6.2 General information about the data

A first display during the process of inspection of the data is the printout of the header of the data file. This gives a good overview over the content of the data set with variable (column) names and type of data. Typically a set of six rows is displayed.

```
##   age sex chest.pain.type resting.bp.s cholesterol fasting.blood.sugar
## 1  40   1               2          140         289                   0
## 2  49   0               3          160         180                   0
## 3  37   1               2          130         283                   0
## 4  48   0               4          138         214                   0
## 5  54   1               3          150         195                   0
## 6  39   1               3          120         339                   0
##   resting.ecg max.heart.rate exercise.angina oldpeak ST.slope target
## 1           0            172               0     0.0        1      0
## 2           0            156               0     1.0        2      1
## 3           1             98               0     0.0        1      0
## 4           0            108               1     1.5        2      1
## 5           0            122               0     0.0        1      0
## 6           0            170               0     0.0        1      0
```

It can be seen, that all data is numeric (integer, double). Another fast overview can be achieved by using the function *str* (see table below).

```
## 'data.frame':    1190 obs. of  12 variables:
##  $ age                : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ sex                : int  1 0 1 0 1 1 0 1 1 0 ...
##  $ chest.pain.type    : int  2 3 2 4 3 3 2 2 4 2 ...
##  $ resting.bp.s       : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ cholesterol        : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ fasting.blood.sugar: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ resting.ecg        : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ max.heart.rate     : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ exercise.angina    : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ oldpeak            : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST.slope           : int  1 2 1 2 1 1 1 1 2 1 ...
##  $ target             : int  0 1 0 1 0 0 0 0 1 0 ...
```

We can see, that all the data is numeric (*int* or *num*), which may not be convenient to establish a ML model, because some of the variables consist of a few distinctive values, that is categorical or factor values (i.e. no continuous values like *age* or *cholesterol*).

6.3 Detailed data exploration

A detailed exploration of the data available shows missing data, possible inconsistent data and the general data quality, which is essential for all further steps to build a ML model and achieve a good prediction.

This task is usually named as Explanatory Data Analysis (EDA) and is used to show general features of the different variables and also correlations between variables using graphics or statistical values directly. It is also helpful to analyse the distribution of the different variables.

6.3.1 Missing data

The following table shows the number of missing (non existent = NA) values in each column:

Table 3: Missing values

Variable	N
age	0
sex	0
chest.pain.type	0
resting.bp.s	0
cholesterol	0
fasting.blood.sugar	0
resting.ecg	0
max.heart.rate	0
exercise.angina	0
oldpeak	0
ST.slope	0
target	0

There are no missing values in the data set.

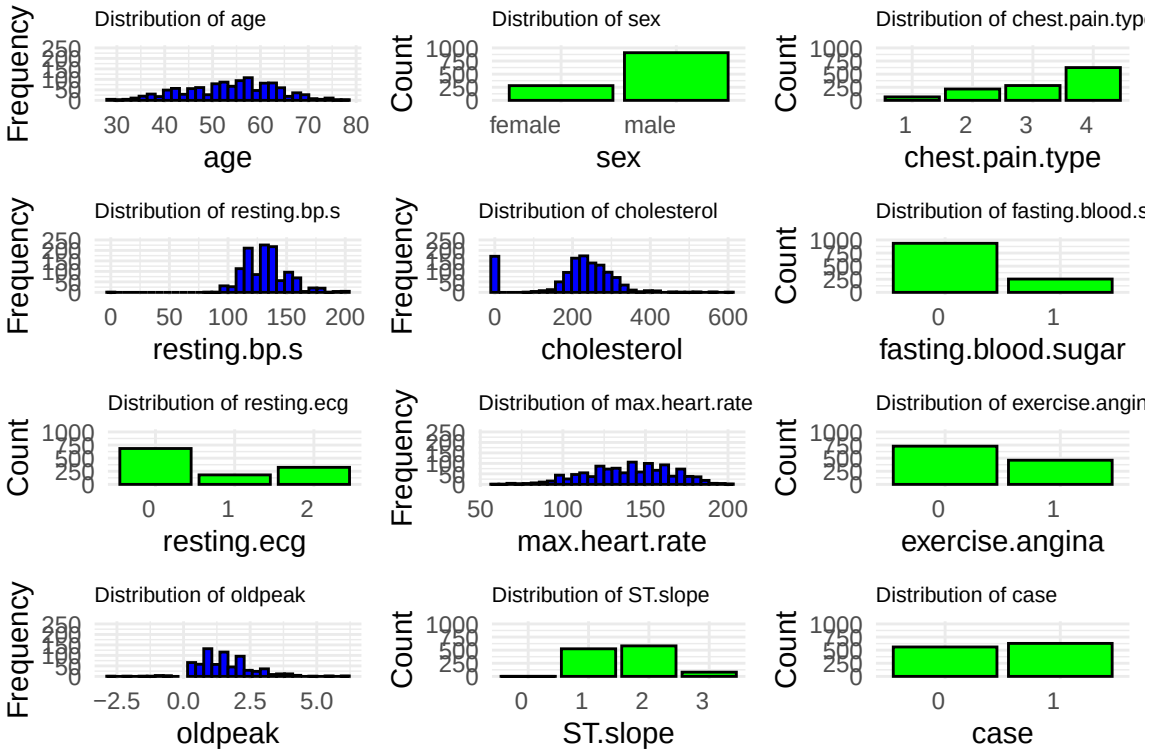
6.3.2 Reorganization of the data

From the description of the data (see chapter 3), we know, that some of the variables (Columns) contain nominal values, i.e. have to be handled as factors. Before we proceed with the data analysis, the data type will be set appropriate. This affects the following columns (attributes / variables): **sex, chest.pain.type, fasting.blood.sugar, resting.ecg, exercise.angina, ST.slope and target**

The *target* remains as a numerical value, but the new column *case* will be a factor. A simple code below converts the columns mentioned to categorical variables in a new data frame.

6.3.3 Data spread and distribution

I choose to use *GridExtra* to arrange the graphics just to have a fast overview with no detailed presentation of the data. The following graphic shows the distributions of the variables to get an impression of the spread and the values of the different variables without distinction whether the variable is continuous or categorical.



The graphic displays all categorical data in *green*, all continuous (numerical) data in *blue*. The frequency of the value *0* in the graphic of *oldpeak* can not be displayed, because of the overall y axis limit due to an acceptable representation of the data.

A summary of min, max, mean, standard deviation and median values is presented in a table below the graphics for variables with continuous values only.

It can be seen, that *ST.slope* has four levels, while the data description explains only three levels. Because there is only one *zero* in the data set, I assume this a wrong entry. Because it is unclear, this data will removed from the dataset.

Table 4: Continuous Variables: spread, average, standard deviation and median

Variable	min	max	average	sd	median
age	28.0	77.0	53.71	9.35	54.0
resting.bp.s	0.0	200.0	132.14	18.37	130.0
cholesterol	0.0	603.0	210.38	101.46	229.0
max.heart.rate	60.0	202.0	139.74	25.53	141.0
oldpeak	-2.6	6.2	0.92	1.09	0.6

Several of the variables contain also *zero* values (*resting.bp.s, cholesterol*) or negative values (*oldpeak*). This has to be evaluated in detail and - if not plausible or realistic - replaced or filtered out.

6.3.4 Inconsistent data and data cleaning/correction

The variable *cholesterol* consists of an amount of *zero* values, which is not explained in the data description. We must assume, that in these cases no information is available, i.e. the value has not been measured. These zero values distort the distribution of cholesterol measurements and have also an influence on any ML model.

The average and median of the *cholesterol* values with and without *zeros* are presented in the table below. They show a significant difference.

Table 5: Cholesterol average and median

Data	Average mg/dl	Median mg/dl
Zeros included	210.4	229
Zeros removed	246	240

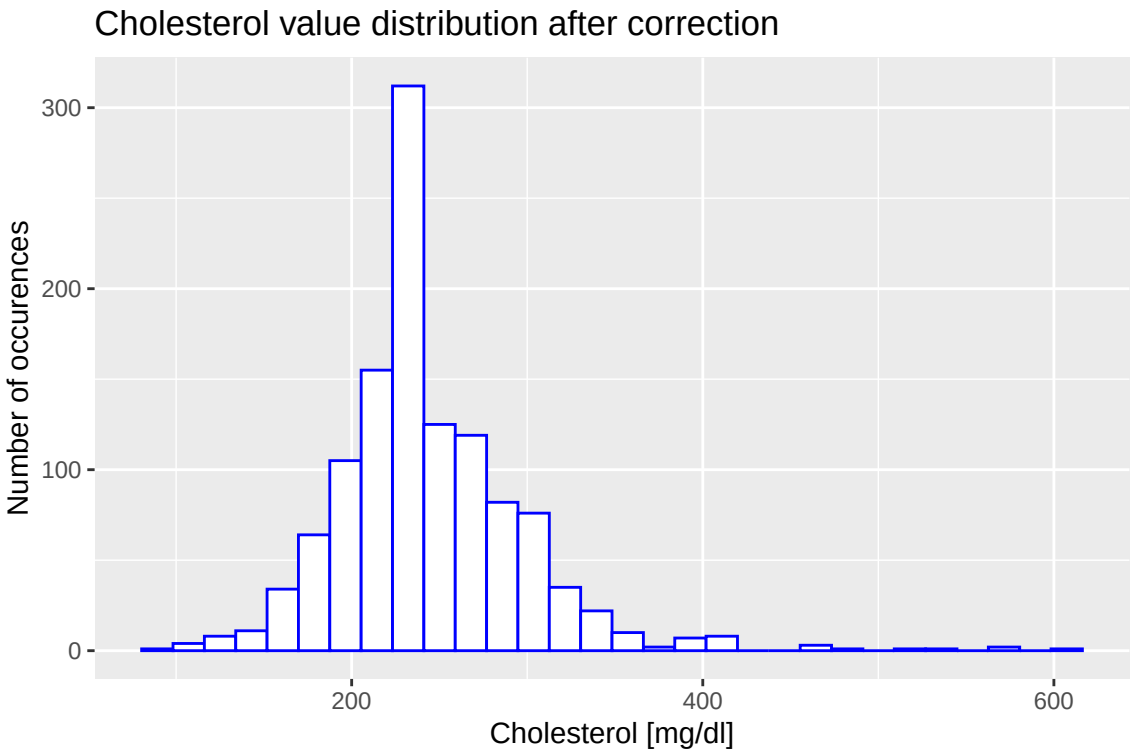
In total 172 *zero* values exist in the data set, which represents a proportion of 14.4% of all values. Four possible solutions exist to correct the data:

1. Remove the data sets to avoid a negative influence on the model
2. Replace all zeros by the average value calculated from non-zero values only
3. Replace all zeros by the median value calculated from non-zero values only
4. Replace all zeros by generated values based on a normal distribution calculated from the non-zero values by use of mean and standard deviation

In a first step , option 3 is used, i.e. the zero values are replaced by the median of the cholesterol value calculated without the zeros. It will influence the distribution of the cholesterol values in a way, that the distribution will get narrower as about 15% of the values are set to the average. This is displayed in the next graphic. A replacement with the average (option 2) instead should have the same effect, because average and median differ only by -2.5%.

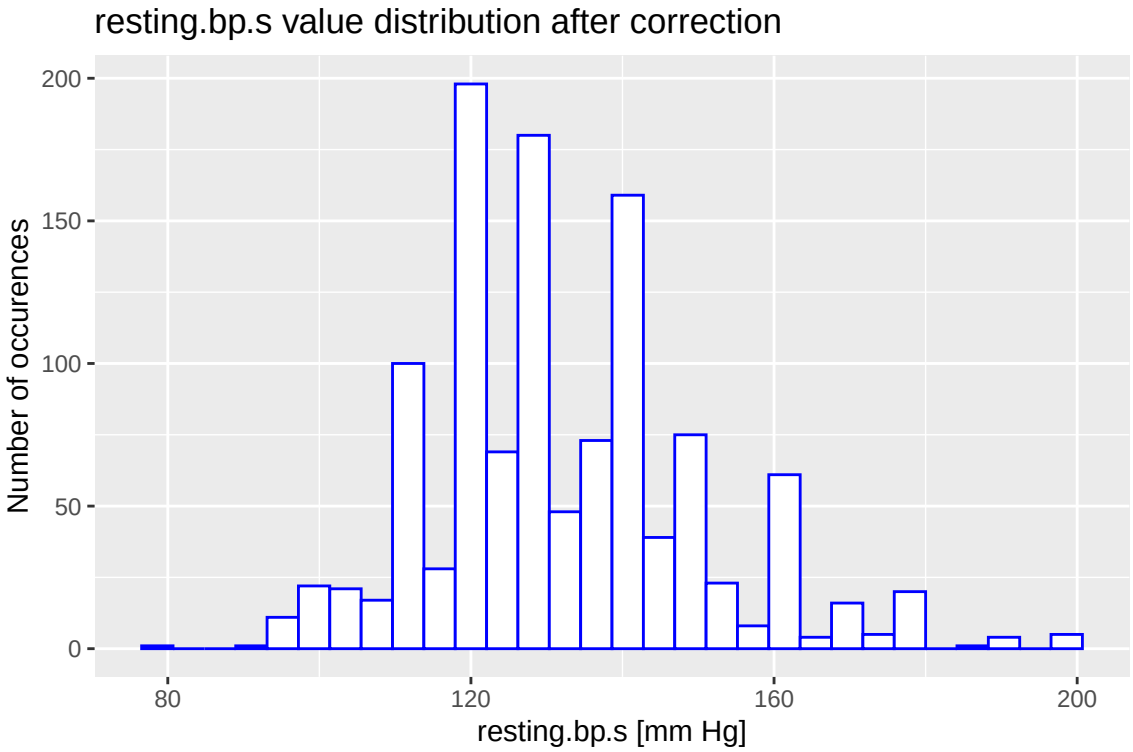
Option 4 was tested also, but causes a distortion of the distribution of the *colesterol* values, depending on the sample created to replace the values. Option 3 seems to be an acceptable approach.

It can also be seen, that some of the values are far beyond 400 ml/dL, which is far too high according to actual knowledge (see for example [8]). A value above 240 mg/dL gives rise to a high risk of a heart disease, values above 450 mg/dL could be interpreted as outliers and removed. We do not take into account a drop or correction of these very large cholesterol values, because the total number is very small compared to the data set (number of values above 450 mg/dL: 9).



After the correction, we find the average with 245.09504, and the median as 240.

There is only one *zero* value in the variable *resting.bp.s*. The content of this variable is the blood pressure at rest in mm Hg and has to be a positive value. This value is also set to the median of the data set. In this case we can use the median calculated for the full data set, because the influence on the average and median can neglected.



The distribution of the values gives rise to questions. Certain blood pressure values seem to be far more prominent than others. We calculate the frequencies of the top ten occurrences and as shown in the table below.

Table 6: Top ten blood pressure values

BP in mm Hg	Frequency
120	166
130	150
140	137
110	76
150	72
160	61
125	39
128	27
135	26
138	26

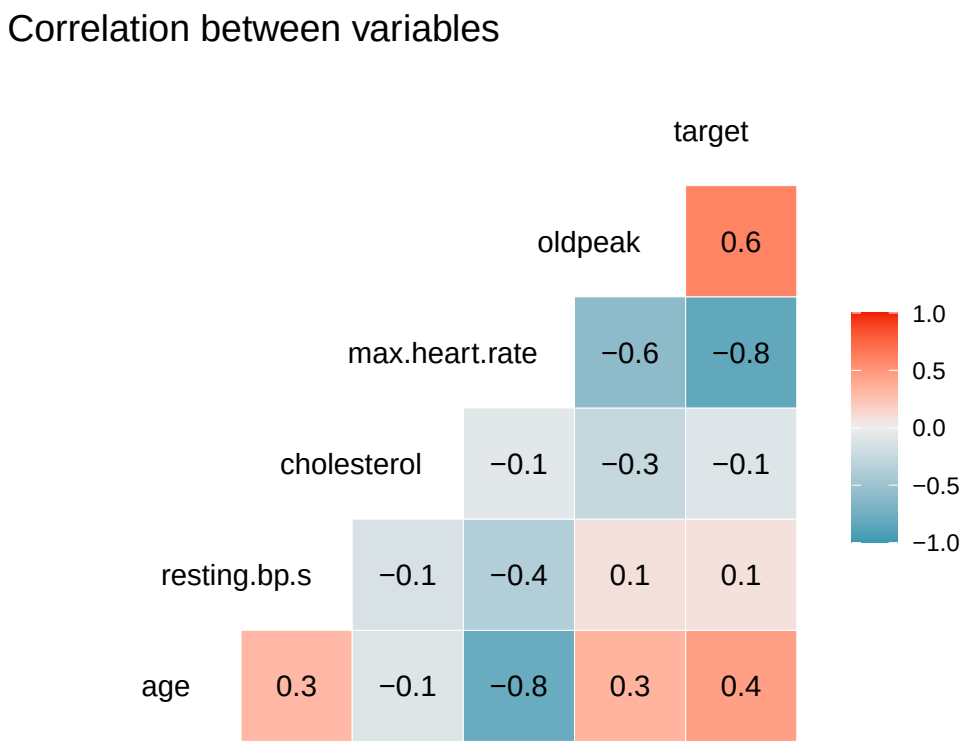
Nearly 40% of the data belong to a blood pressure of 120, 130 and 140 mm Hg (top three frequencies), followed by 110, 150 and 160. This sums up in total to 663 (56%) entries. It is not clear, if the values are really measured or if the persons tested had to give the result of their private measurement. The trust in the accuracy of the blood pressure data is quite low.

Since the data has a large spread and most of the values are multiples of ten, the variable was as stratified into classes as multiples of 10 to test, if the models improve, but the accuracy was slightly lower. Therefore this approach for *resting.bp.s* is not implemented in this final report.

At this point the data is ready for the application of ML models.

6.3.5 Correlation of data

To get more insight we check, which variables are correlated. This can easily be done using the appropriate function and visualize the data as shown below assuming the data is numerical and consists of continuous values, i.e. the data as read from the file is used.



The matrix shows some positive and negative correlations between the *target* and *oldpeak* as well as between *age* and *max.heart.rate*, but most of the data seems to have nearly no correlation.

But the numerical data is only a subgroup of the data set. This has to be taken into account for the choice of the ML model.

6.3.6 Grouping of variables

As explained in the *Introduction*, the data is composed of sets from different countries, but the information about the origin (i.e. country/region) is omitted and not available. It is not possible to use the origin for grouping. Since the data comes from the US and Europe, this could have an influence due to different cultural behavior.

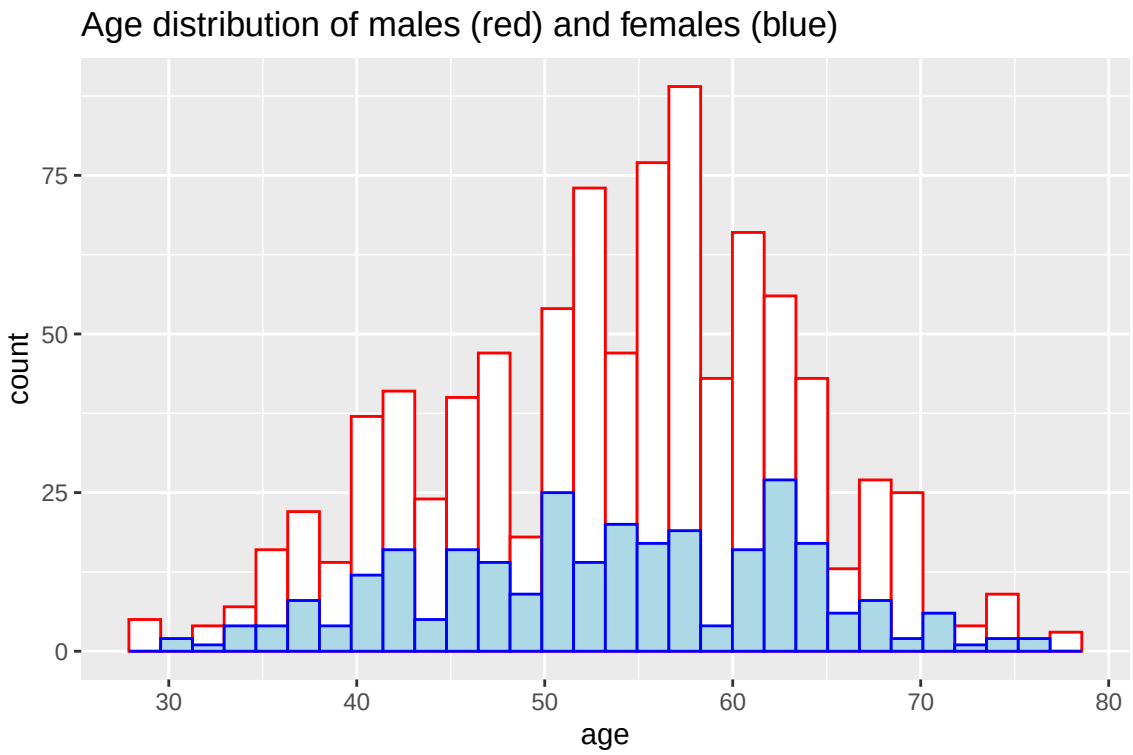
Thus a first analysis is focused on the influence of gender (i.e. male vs. female), because it is known, that symptoms of men and women are different with respect to heart disease (see reference [5]).

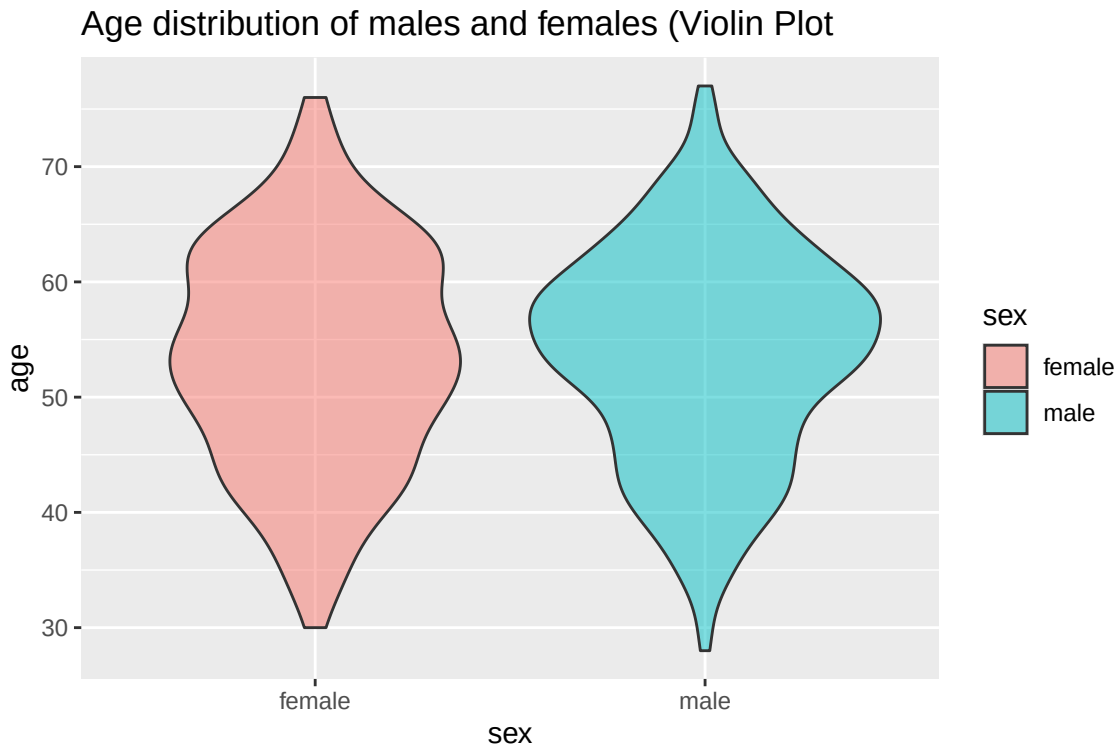
The dataset consists of 76% males and 24% females.

According to [5] which states "*Sex-specific differences grounded in biology may play a role. The overall prevalence of coronary artery disease is lower in women, and they tend to develop heart problems at older ages (the average age for a first heart attack in men is 65, compared with 72 in women).*", a significant difference in the data for men and women can be expected and also an influence of the age.

Table 7: Number of datsets for males/females

Gender	N
female	281
male	908





From the histograms, it is difficult to judge, if the age distributions of males and females differ. An additional plot (*Violin plot*) helps th see the differences. The portion of males between 50 and 65 years is much higher relative to the total as for females. A direct use of this information is not possible. So it is necessary to calculate the relative numbers (percentage) of males and females.

Because the age distribution of males and females consists of different values over a large scale, the age data is stratified such, that every person is put into a class (called *age_strat*) within a five year interval.

The resulting data is shown below graphically as well as in tabular form.

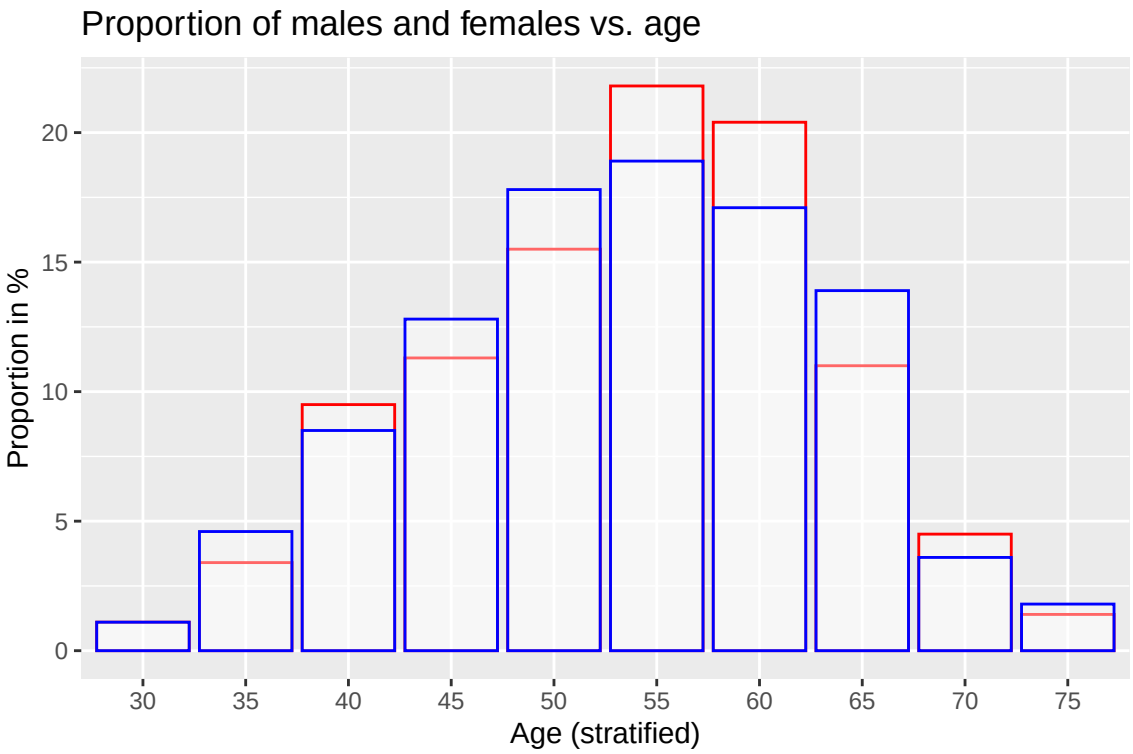
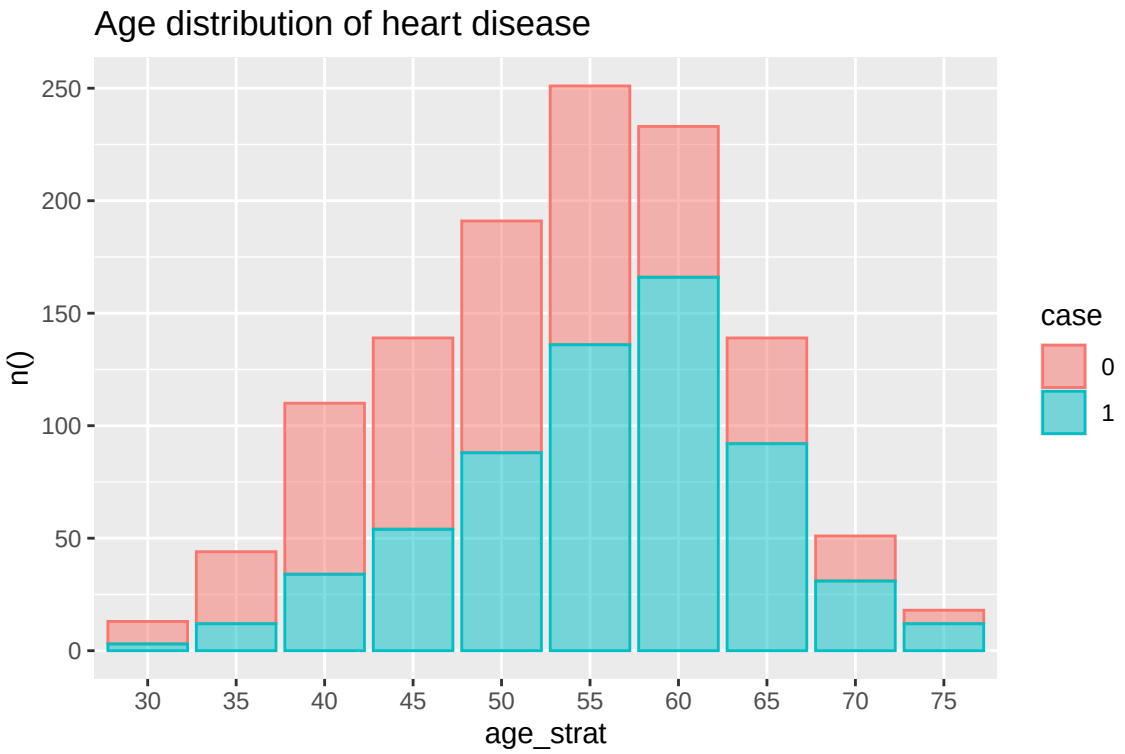


Table 8: Age distribution of males/females

Age (class)	females	males	males %	females %
30	3	10	1.1	1.1
35	13	31	3.4	4.6
40	24	86	9.5	8.5
45	36	103	11.3	12.8
50	50	141	15.5	17.8
55	53	198	21.8	18.9
60	48	185	20.4	17.1
65	39	100	11.0	13.9
70	10	41	4.5	3.6
75	5	13	1.4	1.8

Now it can be seen, that the age distribution of males and females differs only slightly by a few percent. We can also assume, that significant differences in the data for males and females is not simply an outcome of completely different samples with respect to the age (e.g. only people above 60 etc.).



This histogram reveals, that the number of heart disease cases is highest around the age of 60 years (+/- 5 years), that means age seems to be a significant factor for heart disease. On the other hand is the number of persons of that age the highest. The number of disease cases must be weighted with the number of persons of the same age class.

It can be calculated similar to the proportions of males and females in each age class. The outcome of this calculation is shown below.

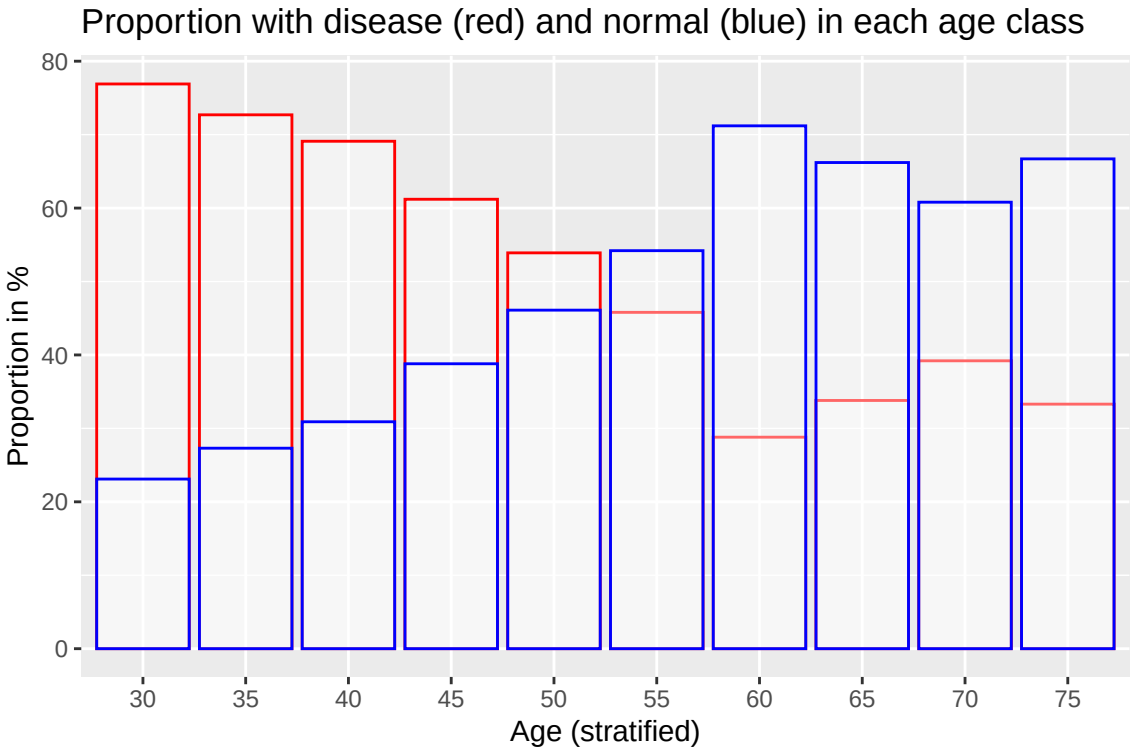


Table 9: Age distribution of Disease

Age (class)	Disease	Normal	Disease %	Normal %
30	10	3	76.9	23.1
35	32	12	72.7	27.3
40	76	34	69.1	30.9
45	85	54	61.2	38.8
50	103	88	53.9	46.1
55	115	136	45.8	54.2
60	67	166	28.8	71.2
65	47	92	33.8	66.2
70	20	31	39.2	60.8
75	6	12	33.3	66.7

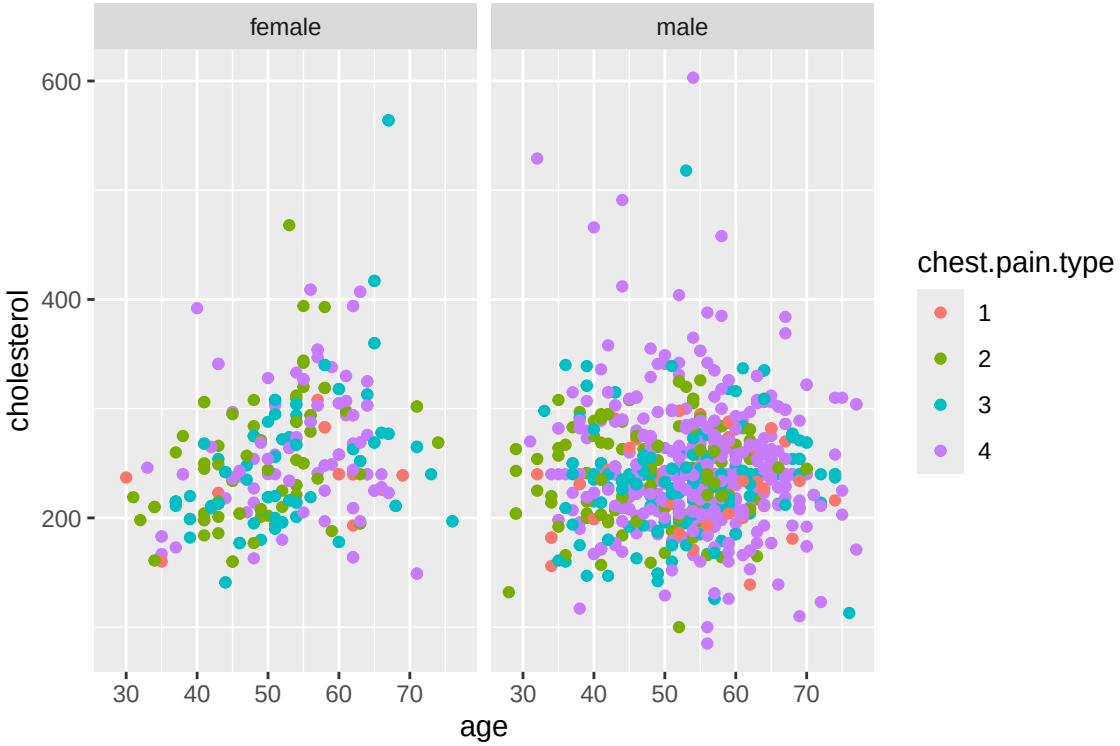
This result is surprising. The relative proportion of disease cases for the age classes of 55 years and below is higher than for people with or above 60 years. The total number of cases in each age class is also shown in the table above and corresponds with the proportions. It is not possible to explain this based on the data set. Additional information would be necessary to interpret this. For very small numbers of cases (e.g. 13 cases in the 30-year class), the values will not be reliable, since a small number of cases has a high influence on the proportion. No significant influence is seen with respect to a ML model.

6.3.7 Significant patterns

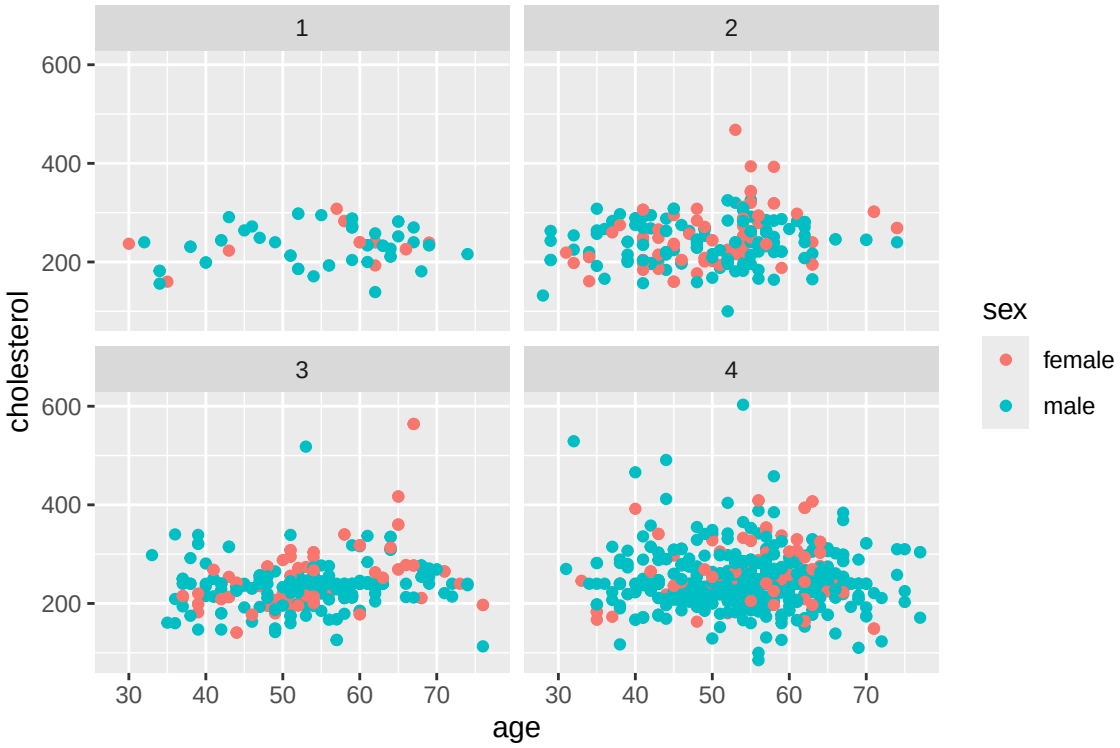
Can we find significant patterns in the data (i.e. clustering etc.)? The first choice is to separate males and females and check relevant attributes. This is done using graphics, because this easily reveals, if data is spread or concentrated in regions of interest. The graphics combine up to four different variables to check for possible clustering or dependencies.

The set of graphics is only presented for combinations of data, which reveal possibly some interesting features/structure and also described in short.

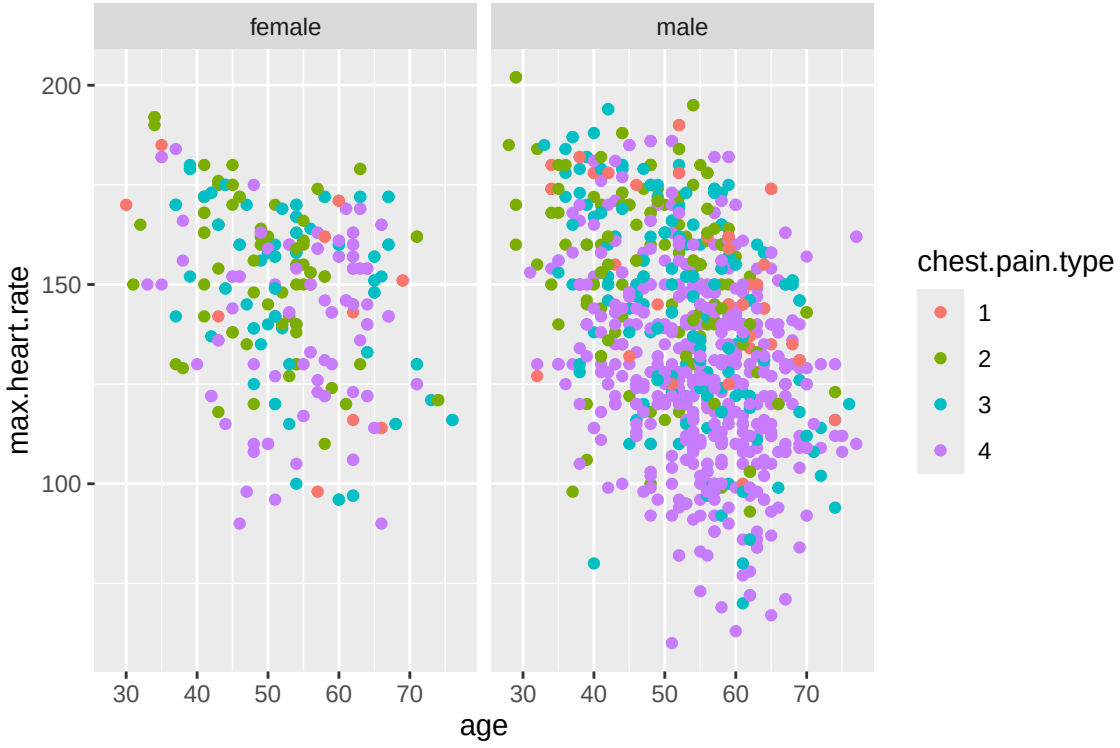
Since cholesterol is an important factor for cardiovascular disease, it is interesting, how the cholesterol values are distributed with respect to additional grouping.



The graphic above shows the cholesterol values vs. age, grouped by sex and colored by chest.pain.type. Only one significant pattern is visible: the corrected cholesterol values (0 replaced by the median 240) are prominent in the data for males.



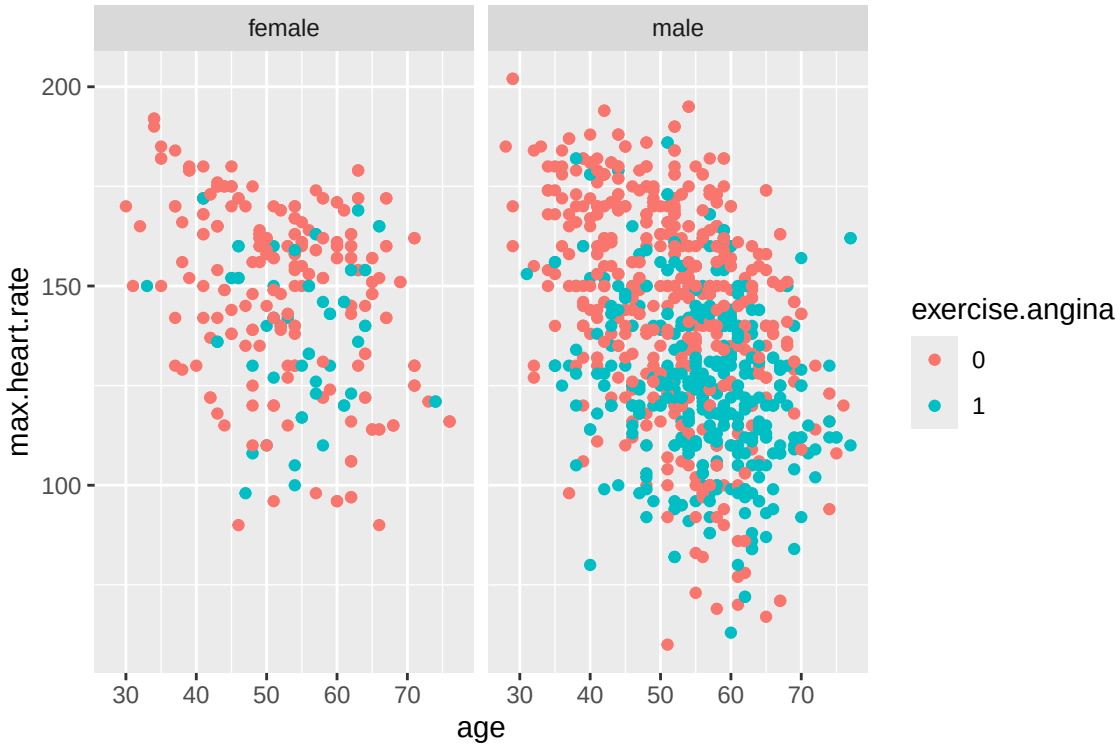
The graphic above shows the cholesterol values vs. age, grouped by chest.pain.type and colored by sex. No significant pattern or correlation is visible.



From the graphic above (max.heart.rate vs. age, grouped by sex and colored by chest.pain.type), it can be seen, that the type of chest pain is related to the maximum heart beat rate. Type 4 (asymptomatic) seems

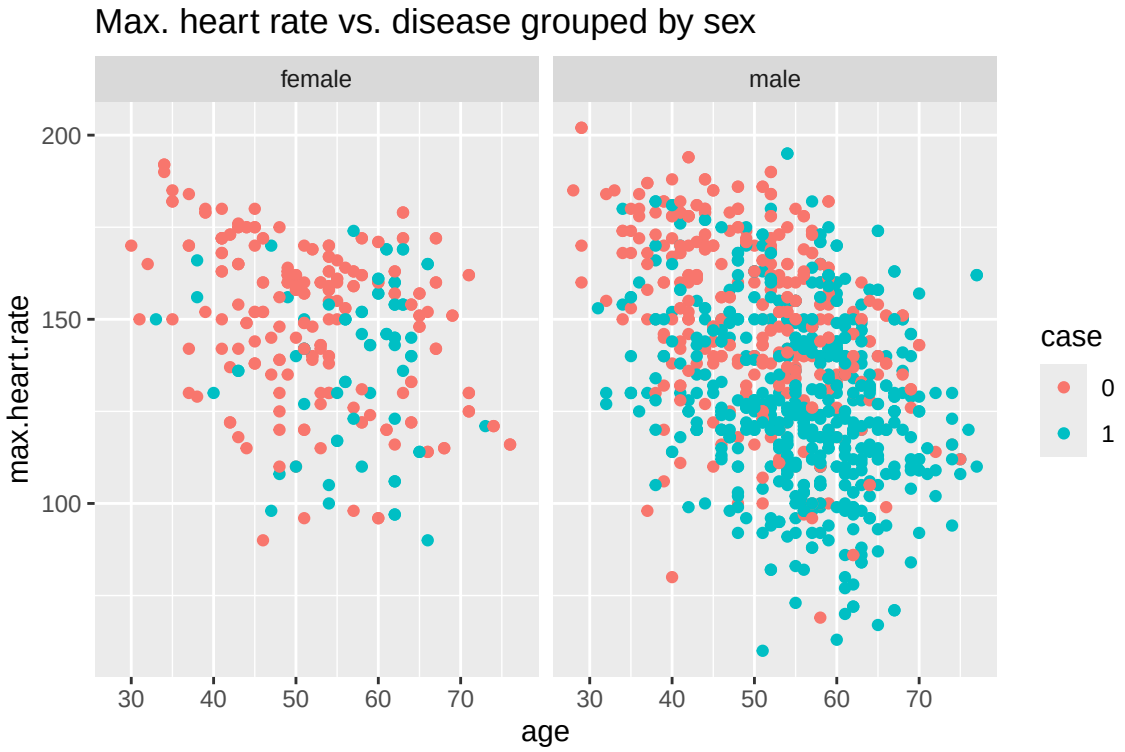
to be more often, when the max. heart beat rate is lower than 150 per minute. A similar picture can be generated with the variable *ST.slope* instead of *chest.pain.type*. In that case, the lower maximum heart beat rate is connected with *ST.slope = 2*, which is not a negative sign (see below). It can also be seen, that *ST.slope = 1* and *ST.slope = 2* dominate.

In case of a heart disease, not only chest pain, but also a feeling of heavy breathing, called angina often occurs. In data set consists also of a value called *exercise.angina*, which means, that the person encounters angina symptoms induced by exercise. It is surely of interest, if there is some pattern in the data, which is shown in the following graphic.



In the graphic above the *max.heart.rate* is displayed vs. *age*, grouped by *sex* and the value of *exercise.angina*. A lower *max.heart.rate* corresponds to *exercise.angina* symptoms.

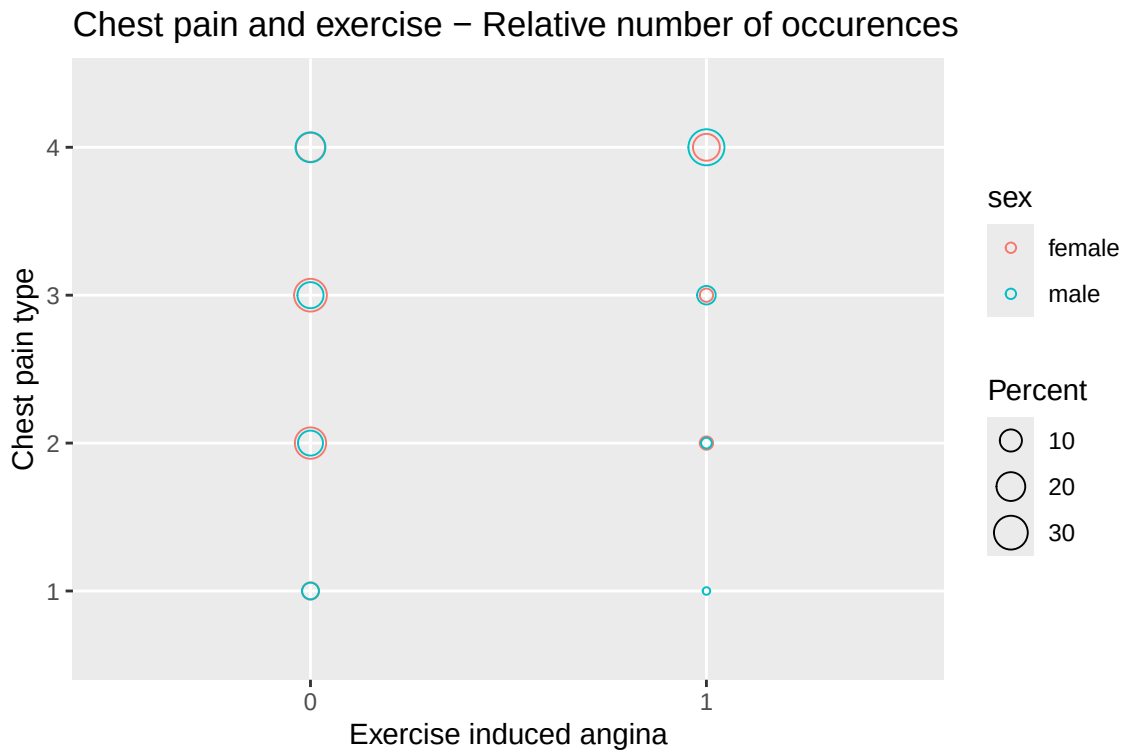
The next graphic shows the *max.heart.rate* vs. *age* grouped by *sex* and *case*. For males, the two cases reveal the same characteristics like the graphics above with *exercise.angina*, *ST.slope* and *chest.pain.type*. Two slightly separated regions with a "center of mass" at lower *age* (about 45) and higher *max.heart.rate* (about 150), while the other group is centered at an *age* of about 55 and *max.heart.rate* of about 120. This behavior can not bee seen in the data for females, but this could be due to the fact, that less data is available.

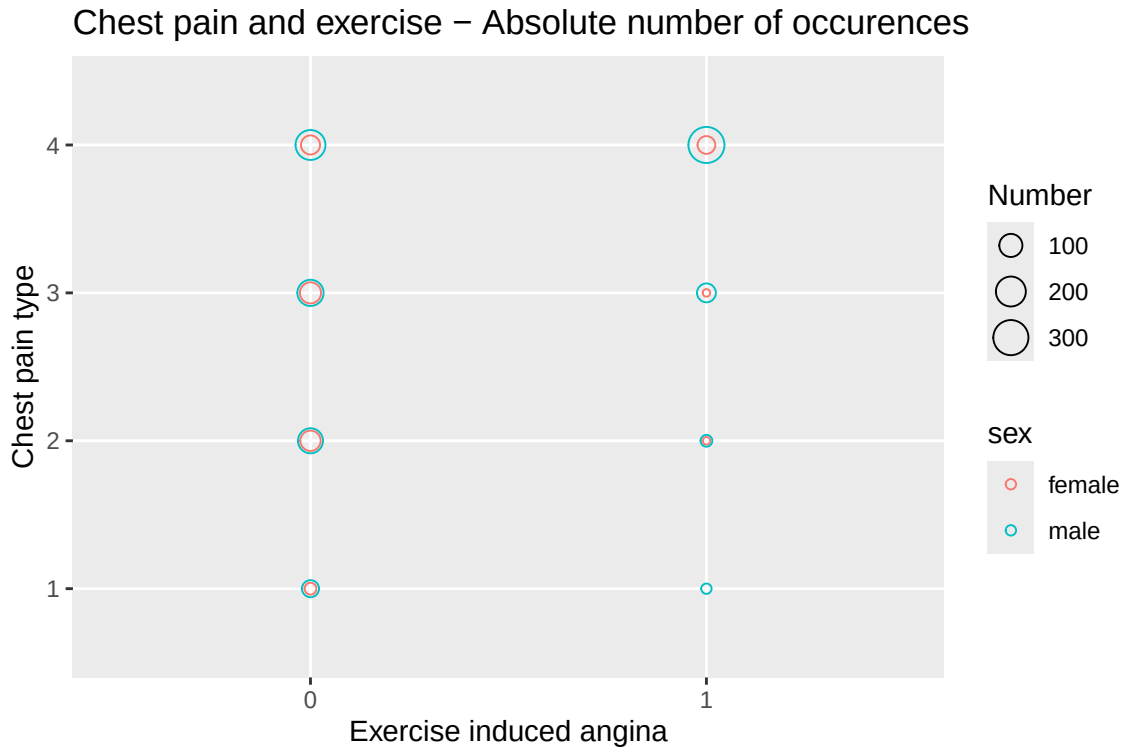


The next plot shows the distribution of cases for males and females together with the *cholesterol* value and the *age*. The plot reveals no new insights. Heart disease occurs more often at higher ages, while the cholesterol shows no significant change with age.

The next two plots show the occurrences of *exercise.angina* in combination with *chest.pain.type* with absolute

values (number of persons) and the relative amount, separated for males and females.





From the graphics above we can deduce the following information:

1. Most of the persons (men and women) had no *exercise.angina*, but had also different angina symptoms (i.e. *chest.pain.type = 1 or 2*)

2. The most *chest.pain.type = 4* occures in combination *exercise.angina = 1*
3. The graphic with the relative number of different symptoms suggests, that no female in the data set encounters a *chest.pain.type = 4* while *exercise.angina = 0*, but this is not true as the graphic with the absolute values shows. In this case, the percentage is nearly identical (see table), which may result in a complete overlay of the two values (male / female).

Conclusion: Graphics should also be checked for consistency and possibly misleading depiction.

In summary, the data shows some characteristics based on variables with nominal values. The continuous variables reveal no strong correlations as suggested by the correlation matrix in chapter 6.3.5.

7 Methods applied

The first step here is to generate the data sets for training and testing. This is done using the *caret* library. AS a starting point the training set consists of 80% of the data, while the test set consists of the remaining 20%.

The columns *target* and *age* will be removed

7.1 Preparation of train and test sets

7.2 Model building

As described above the data set consists of continuous and categorical variables. The question is, what model could be used to establish an acceptable prediction of heart disease (i.e. the value of target / case). In the following sections, this will be described and tested with various approaches.

The *caret* library provides the function *train* to use different algorithms in a simple and efficient way and will be used as described below to test several models for the prediction of heart disease. But in a first step we will check the power of a prediction by random sampling.

The performance of the models will be compared using the accuracy.

7.2.1 Simple statistical model (guessing)

From the data exploration it can be seen, that the number of heart disease cases is nearly 50% of all cases. What accuracy can a guess with 50% chance get? The result is given below and as expected a nearly 50% accuracy is reached.

In this very simple (and stupid) approach, the accuracy gained for the train set is 0.51263 and 0.4477 for the test set. This result is of course not acceptable for a ML model.

Table 10: Accuracy of the model

Method	Accuracy
Guessing	0.448

7.2.2 Linear combination of variables (glm)

Since no strong correlations between the continuous variables exist, it is not expected, that a generalized linear model will give a high accuracy.

Table 11: Accuracy of the model

Method	Accuracy
Linear Combination (glm)	0.841

The result is a little bit surprising, but gives at least an accuracy of 84 percent.

7.2.3 k-Nearest Neighbors

Based on the fact, that most of the attributes are categorical, we could assume that an algorithm based on the neighborhood of the data in a multi-dimensional space would lead to a high accuracy. This expectation has not come to real. The facts show, that this approach leads to a relatively low accuracy in data prediction, as shown in the table below compiled of the data of the prediction models used.

The *train* function can in this case be used with a tuning parameter k. The influence of the tuning parameter and its optimal value is shown in the graphic below.

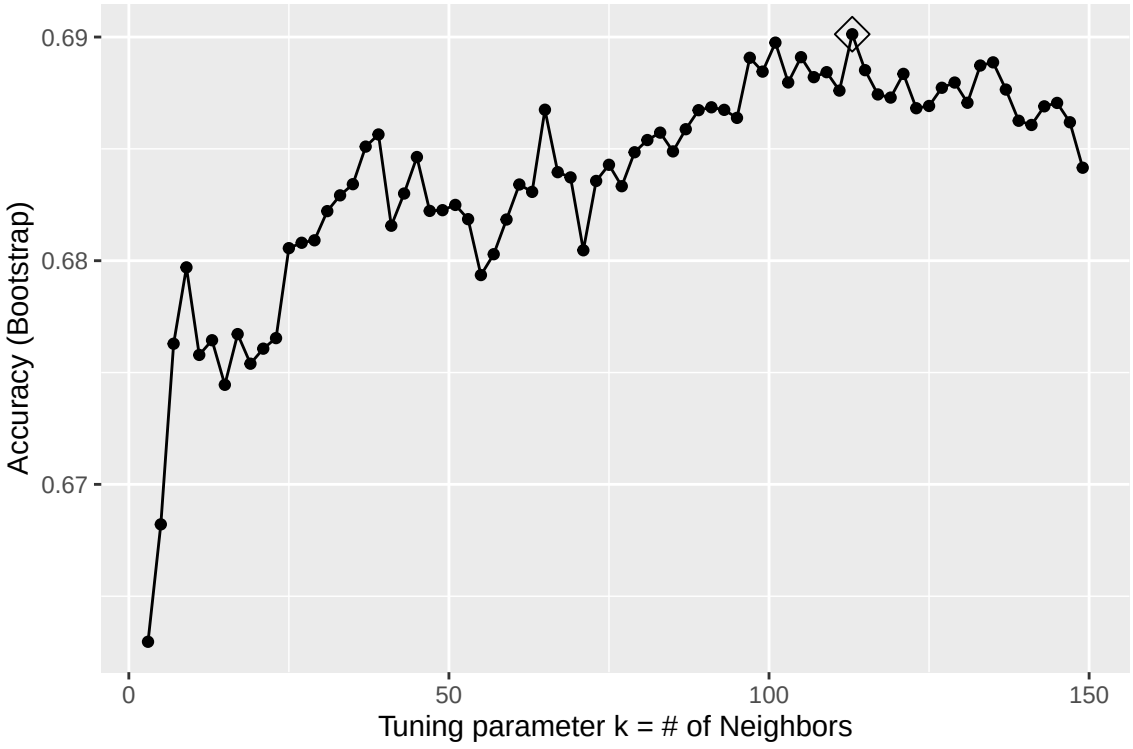


Table 12: Accuracy of the model

Method	Accuracy
Linear Combination	0.665

The resulting accuracy is very low and even below the *GLM* model.

7.2.4 GAM loess

According to the course material [3], this algorithm should perform similar to the kNN model. Here, also a tuning parameter could set, but in this case we just let the function give the best value.

This algorithm is a *General Additive Model (GAM)* similar to the *Generalized Linear Models (GLM)* using the *LOESS* algorithm for regression.

Table 13: Accuracy of the model

Method	Accuracy
GAM Loess	0.837

7.2.5 Random Forest

Because the results so far are not satisfying, a Random Forest model (with algorithm *cforest*) will be used. This is a generalized approach based on the principles of a decision tree.

Here, we do not use the tuning parameter, but relay in the first step on the optimization strategy of the algorithm.

Table 14: Accuracy of the model

Method	Accuracy
Random Forest (cforest)	0.862

7.2.6 Random Forest with Tuning Parameters

It is also possible to use the library *randomForest* directly to train a model instead of using the *caret* function *train*. This is done also and the result using the tuning parameter *mtry* is given below. An additional improvement of the model was achieved by adapting the parameter *ntree* and using the algorithm *rf*.

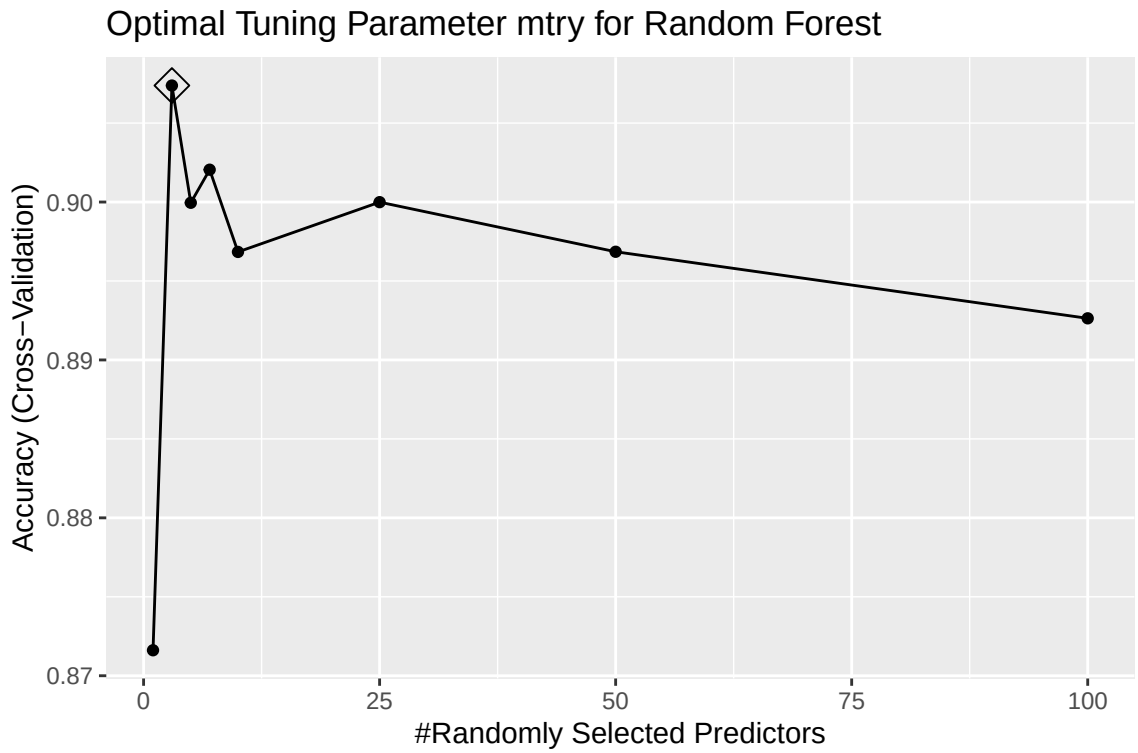


Table 15: Accuracy of the model

Method	Accuracy
Random Forest Optimized	0.912

8 Final Result

The table below lists all models used at this point with their accuracy. We can see, that the best model so far is generated using *Random Forest Optimized*, which seems to be reasonable due to a large amount of categorical data in the data set. I tested also possible improvements by combining models, but none of the combinations reached the accuracy of the last *Random Forest Optimized* model.

Table 16: Comparison of the models

Method	Accuracy
Guessing	0.448
k-Nearest neighbors	0.665
GAM Loess	0.837
Linear Combination (glm)	0.841
Random Forest (cforest)	0.862
Random Forest Optimized	0.912

9 Conclusion

The prediction of heart disease based on the data available in this data set is investigated by use of different models. The accuracy for prediction is best approached by a Random Forest model with accuracy of 0.91 and 0.86 respectively, while a generalized linear model reaches an accuracy of 0.84, which is about 8% less than the best model.

Since a lot of questions arised with respect to the data quality, the result seems quite acceptable.

10 References and Acknowledgements

1. R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
2. RStudio 2024.09.1+394 "Cranberry Hibiscus" Release (a1fe401fc08c232d470278d1bc362d05d79753d9, 2024-11-03) for windows, Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) RStudio/2024.09.1+394 Chrome/124.0.6367.243 Electron/30.4.0 Safari/537.36, Quarto 1.5.57
3. Introduction to Data Science, Rafael A. Irizarry https://rafalab.github.io/dsbook/
4. Statistics for Data Scientists, Maurits Kaptein & Edwin van den Heuvel Verlag Springer, 2022 https://doi.org/10.1007/978-3-030-10531-0
5. The heart disease gender gap, September 1, 2022 https://www.health.harvard.edu/heart-health/the-heart-disease-gender-gap
6. Angina pectoris (in German) https://de.wikipedia.org/wiki/Angina_pectoris
7. Angina: Symptoms, diagnosis and treatments, September 21, 2021 https://www.health.harvard.edu/heart-health/angina-symptoms-diagnosis-and-treatments

8. Cholesterol Levels, Cleveland Clinic (July 2022) https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean