Heart Disease Prediction

COMP3354 Statistical Learning

# MOTIVATION

What problem are we trying to solve?

According to the World Health Organization heart diseases, also known as Cardiovascular diseases (CVDs) represent the number one cause of death globally. With an estimated 17.9 million cases in 2016 they make up a staggering 31% of all global deaths. As deaths of CVDs can be prevented through appropriate treatment, we are trying to find a way to ensure, that high-risk patients can be identified.

Why do you choose this topic?

We chose this subject because we believe that a lot of the deaths would have been easily preventable, if patients would have received the necessary care. Using a statistical approach, we hope to ensure that patients with a high risk of CVDs can get identified, even before the first symptoms show. This is especially important, since the first warning of a cardiovascular disease is often already a heart attacks or a stroke.

What will be the potential impact of your project?

Our project aims to ensure, that people with a high risk of cardiovascular disease can get diagnosed and subsequentially receive the necessary treatment. As three quarters of the world’s deaths from CVDs occur in low- and middle-income countries we are also aiming to bridge a global equality gap and reduce the burden on the countries’ economies.

Current Practices

Currently there are two types of interventions to reduce the likelihood of cardiovascular diseases: population-wide and individual.  
Population wide interventions include measures, such as taxation on fatty and salty foods, tobacco and alcohol, building walking and cycle paths to encourage physical activity and providing healthy school meals to children. As these are general measures this paper will not cover them more in depth.

There have also been previous efforts to use a data-driven approach[[1]](#footnote-1)1, who have used a Naïve Base Classifier with an accuracy of up to 87.98% and a true positive recognition rate between 82.06% and 100% depending on the risk level of the patient.

What do you want to improve?

Our aim is to contribute to the research in the field of cardiovascular diseases and to contribute to the World Health Organisations “Global action plan for the prevention and control of NCDs 2013-2020”.

# METHODS

Data Source[[2]](#footnote-2)2

For our project we are using the Heart Disease dataset from the UCI Machine Learning Repository. It consists of 303 individuals, with 165 cases of cardiovascular diseases and 136 members of the control group.

Variables of Interest

The dataset originally contained a total of 76 variables, many of which were deemed statistically irrelevant by previous researchers. As previous research has only focused on a smaller, complete subset of data, we planned on following the precedents and chose the same data-subset. It’s compromised of the following 14 variables:

1. age

“age” is a quantitative variable indicating the individuals age.

1. sex

“sex” is a qualitative variable describing the gender of the individual. Its value is set to 0 if the individual is female and 1 if the individual is male.

1. cp

“cp” is a qualitative variable describing four different types of chest pain:

1. typical angina

2. atypical angina

3. non-anginal pain

4. asymptomatic

Angina is a type of chest pain or discomfort, caused by the heart not getting enough blood and oxygen[[3]](#footnote-3)3. It’s caused by plaque build-up in coronary arteries, which reduces blood flow to the heart.

1. trestbps

“trestbps” is a quantitative variable, describing resting blood pressure measured in mmHg on admission to the hospital. Optimal values are below 120, high blood pressure starts at 140.[[4]](#footnote-4)4

1. chol

“chol” is a quantitative variable, describing serum cholesterol levels in mg/dL. Normal ranges are from 125-200mg/dL.

Cholesterol is a lipid, a type of body fat. It’s a waxy, fatty substance that some cells in the liver produce and release into the bloodstream. High amounts can clog up arteries, which will prevent blood flow and oxygen from reaching organs and tissues. [[5]](#footnote-5)5

1. fbs

“fbs” is a qualitative variable, indicating if the Fasting blood sugar (fbs) levels are above 120 mg/dL. fbs levels are measured by making the individual not consume anything but water for eight hours. Subsequentially, the levels of glucose in the body are measured.

Normal values range from 70 to 99mg/dL, Prediabetes or Impaired Glucose Tolerance ranges from 100 to 125mg/dL and diabetes is normally indicated by levels of 126 or above.

1. restecg

“restecg” is a quantitative valuable indicating if the ECG is normal (Value = 0), if there is a ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) (Value = 1) or if it is probable or definite that the individual has left ventricular hypertrophy as defined by Estes’ criteria (Value = 2).

1. thalach

“thalach” is a quantitative variable showing the maximum heart rate achieved by the individual.

1. exang

“exang” is a quantitative variable showing if the individual has suffered of an angina through exercising.

1. oldpeak

“oldpeak is a” quantitative value describing ST depression induced by exercise relative to rest. ST depression refers to a downward slope in the ST section of an ECG. (QUOTE)

1. slope

“slope” is a quantitative value referring to the slope of the peak exercise ST segment. It can be upsloping (value = 0), flat (value = 1), or downsloping (value = 1).

1. ca

“ca” a quantitative value ranging from 0 to 3, indicating the number of major vessels coloured by fluoroscopy.

1. thal

thal is a quantitative value indicating the hereditary disease Thalassemia. People with Thalassemia have an abnormal form or inadequate amount of haemoglobin. It is caused by large amounts of red blood cells getting destroyed. Since they are the ones carrying oxygen, Thalassemia can lead to anaemia.   
If the individual shows no signs, then the value will be set to 3. If the defect has been fixed, the value will be 6, otherwise it’s set to 7. The effects of the disease will still be reversible.

1. target

The target value determines if the individual does have a cardiovascular disease or not. Its qualitative values are set to 0 or 1 accordingly.

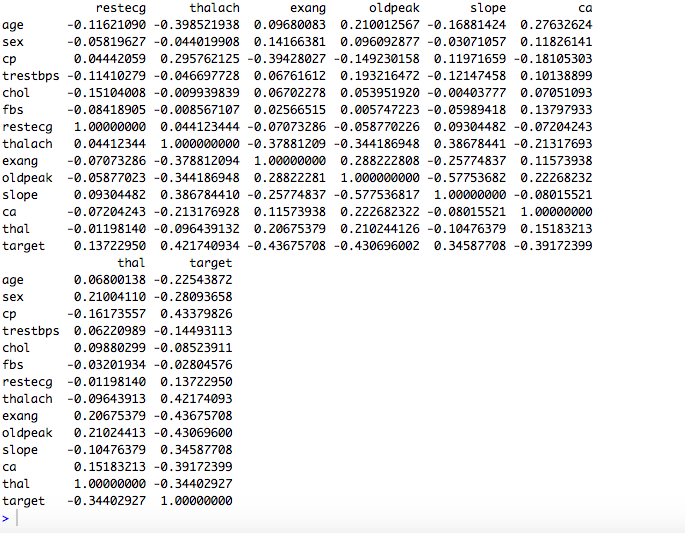
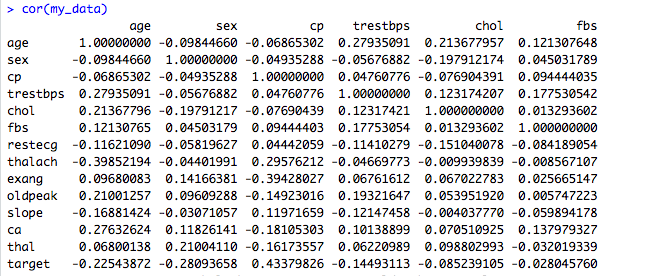
The descriptions were taken from the “heart-disease.names” file from the UCI Machine learning repositories website (ADD LINK). Since they were only rudimentary descriptions, often just consisting of a single word we had to do some additional research on the variables.

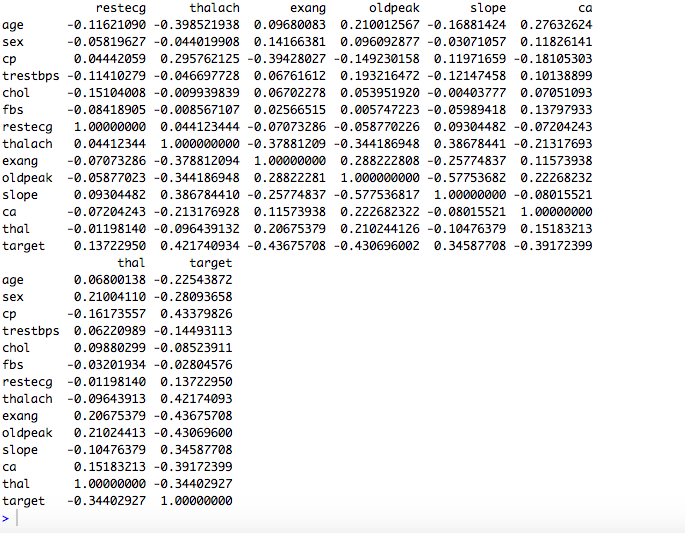
Analytical Procedure

1) Looking for Important Correlations

The correlation matrix shows how strongly correlated the respective variables are to each other. As we define a strong correlation of being > |0.5|, there is only one pair of variables, which satisfies this criteria: “oldpeak” with “slope”.  
This was to be expected, since both of the variables deal with the ST-T segment of the ECG.  
Since “oldpeak” is a quantitatitive variable describing how much the effects of exercise effect the segment and “slope” is a qualitative variable describing how the segment is affected, we decided to keep both variables.

In the last row you can see how much “target” correlates with the other variables. Worth mentioning is that “age” and “sex” aren’t strong predictors for cardiovascular disease. It’s already measurable symptoms such as occurrence of an angina, ECG results and having a family history of thalassemia which are much more statistically relevant.



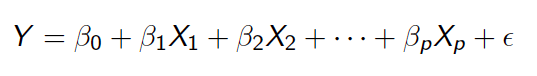


2) Methods

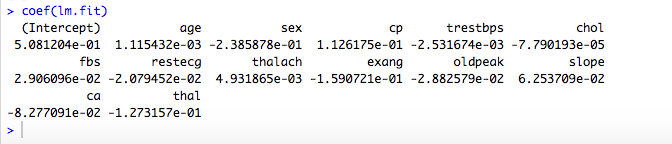
For the following methods we have chosen to apply all the predictors.

Multiple Linear Regression:

Linear regression is very simple approach but interesting to have as a base and for later comparisons. We based our approach to multiple linear regression on the formula:

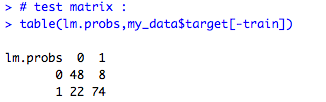


Where the Xi’s are the different predictors (age, sex…) and the βᵢ’s are the different coefficients:



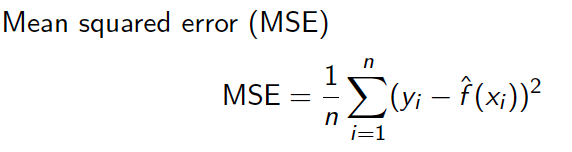
We used half of our data as training set, and the other half as test set. Then we defined the target to be predicted as 1 if y 0.5, and as 0 otherwise.

On the test set, we obtained these results :

This amounts to 122 correct predictions out of 152 in total.

To test the accuracy of the model we also computed the MSE as 0.1973684.

*Recall*:



where yi = correct value

f̂ (xi) = estimated value

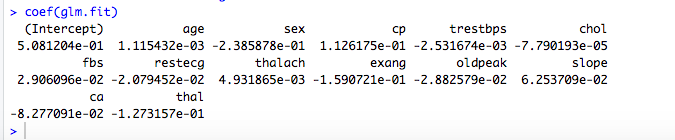
Logistic Regression

Trying to predict a heart disease means that our response is qualitative instead of quantitative.

This means that we have to classify our « observations ».

Logistic regression is known to be an effective method in this case.

Here are the coefficients founds computing a logistic regression :



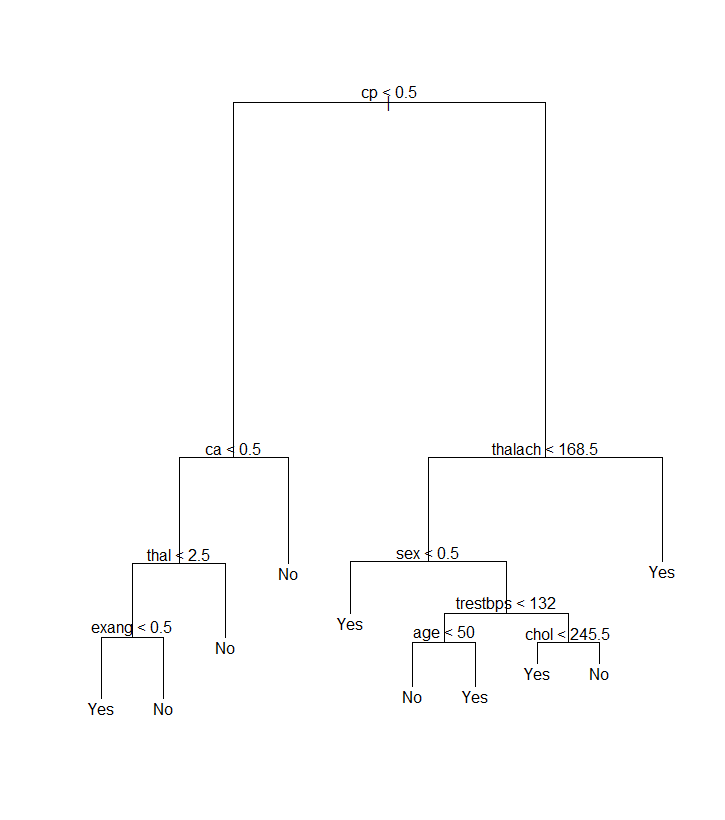
We used two ways to estimate the error. At first we computed the mean square error to compare it to the one linear regression:

Mean square error MSE = 0.1973684

As the both the MSE and the coefficients are exactly the same, we conclude that computing the logistic regression model leads to the exact same model as the linear one.  
We also computed the cross validation error to estimate the test error and compare it with our further models:

Leave-one-out cross validation error: CV = 0.1319878

Classification Tree:



We are aware that “trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen”. [[6]](#footnote-6)4

On the other hand, we also noticed that they provide a global view of the situation and can easily be interpreted.

This can be useful, for example in the case where a doctor wants to explain the reason of the different tests and conclusions to a patient.

This is the reason why we provided one model.

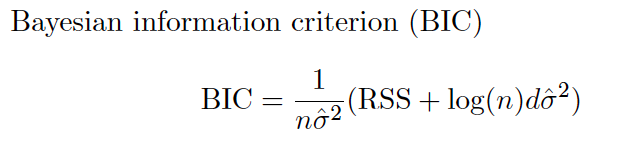
After computing the classification tree, we constructed the pruning tree of size 9, as our main goal is to have a general overview, and the tree above.

At this point of the analysis, even if the data set has already been pre-processed (selecting 14 variables instead of 76), we tried to find if we still couldn’t find an accurate subset, which would reduce the number of variables.

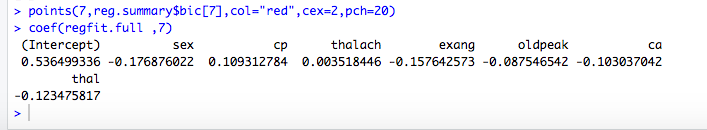
The Best Subset Selection.

We tried to find the optimal subset according to criterions:

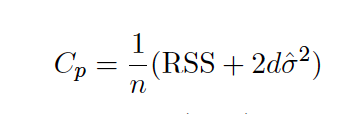
**BIC**:



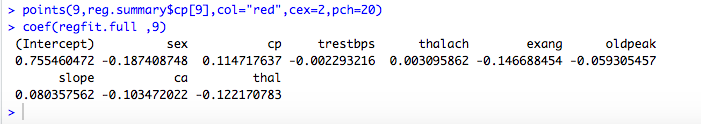
Subset of 7 selected predictors:



**Cp**:

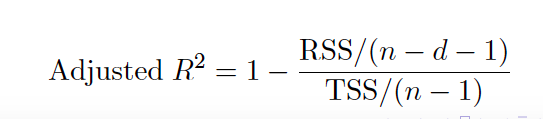


Subset of 9 predictors selected :

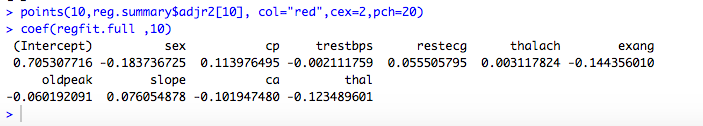


Cp and BIC are estimates of test MSE.

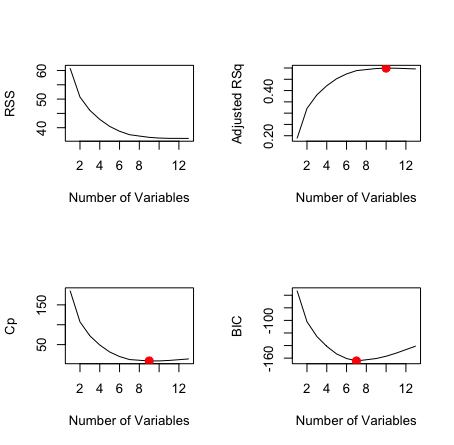
**Adjusted R square**:



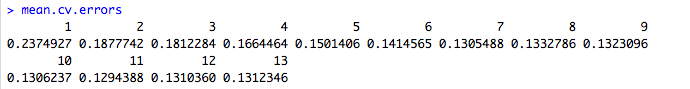
Subset of 10 predictors selected :



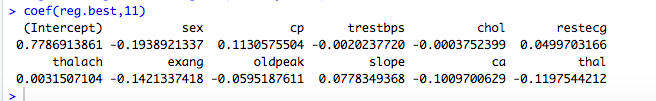
Plots of the results :



**Applying the Cross** **Validation**:



We find the best subset has 11 variables, with the following coefficients :



We concluded, that there is no clear subset selection, which is not surprising because the selection has already been done previously.

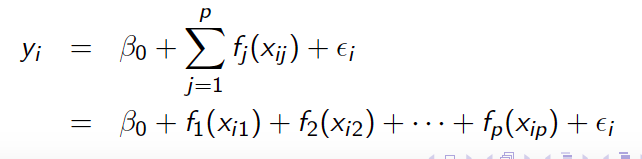
After this subset selection, we started again with different methods, using the last 11 variables found.

Logistic Regression:

We got results, but logistic regression is not hard to compute, so it didn’t make sense to remove variable in this general model.

Generalize Additive Model (GAM) :

Instead of allowing coefficients to the variables, we provide non-linear functions of these variables:

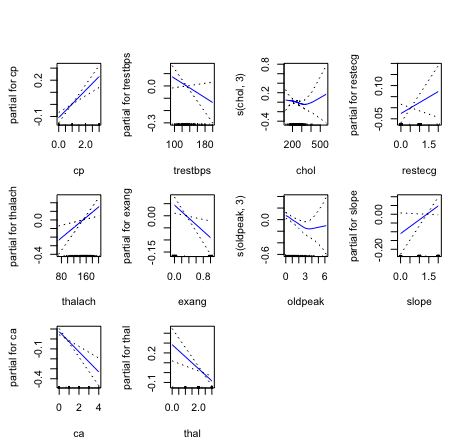


We compared 5 different models.

When we computed the best subset selection, we noticed that even if the 11-variables (cv = 0.1294388) was the best one, the 7-variable was also very close (cv = 0.1305488). So we’ll use this difference between the 7 and others 3 variables of the 11-variables model.

We started giving cubic functions to 7 variables of the best subset selection and quadratic functions to the 3 left (the ones in the 11-variables best subset but not in the 7-variable).

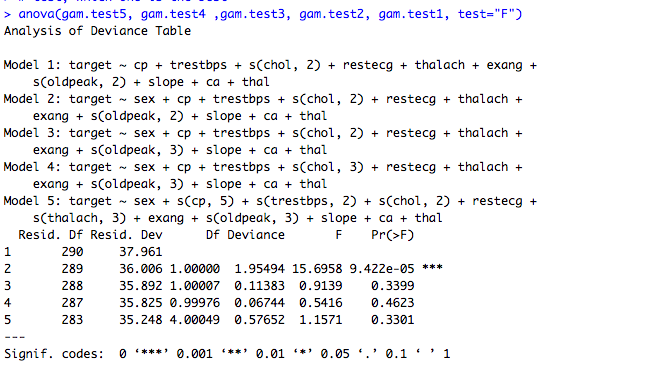
To approximate the functions for each of the variables we made a plot of this first try:



We notice that except for chol and oldpeak, all function can actually be approximated by using a linear function.

The next moove was then to identify wich were the best functions for these two variables left.

We tried cubique and quadratic functions and compared the models :



We find that the second model is the best, using cubic functions for our variables chol and oldpeak.

We finish and obtain our best model, with MSE = 0.1121472

# RESULTS

We have found, that there was no one distinct way, for which to select the predictors to use in the statistical model. Interesting here is that the 7 variables found through the BIC Criterion are a subset of the 9 predictors of the Cp Criterion, which in turn is a subset of the predictors we found trough the adjusted R2 and eventually through Cross Validation.

Common to all four subsets are the predictors “sex”, “cp”, “thalach”, “exang”, “oldpeak”, “ca” and “thal”.

Worth noting here is, that “cp”, “thalach”, “exang”, “oldpeak” can all be directly linked to the heart and it measurably not functioning correctly. “thal” is similar in that it indirectly affects how much oxygen reaches the heart. We didn’t expect the gender of the person to be included in all the four subsets. It’s especially surprising considering it only has an approximate -0.28 coefficient, whereas slope had an approximate 0.35 coefficient and was only included in three of the subsets. Our guess is that the data is generally slightly screwed towards male subjects, since they made up 68.3% of the population.

Also interesting is the fact, that age and fasting blood sugar were included in none of the data subsets. As the first quartile of the population starts at 47.5 years of age and the third at 61 years, most of the there isn’t that much of an age gap in between most of the individuals. This shows, that being healthy in general is more relevant to preventing heart disease, than just being young.

The fbs value is typically used for predicting diabetes and its pre-stage. Since the rest of the dataset’s variables are purely heart focussed or general variables such as sex and age, it comes to no surprise, that fbs isn’t as relevant as the others.

We still decided to go with the full set of 11 variables, since the population of our dataset is only 303 entries, which wasn’t too much for our computer to handle. For a considerably bigger dataset we would have chosen a variation with only the 7 variables or would have tried to reduce the dimensions with Principle Component Analysis.

In order to predict heart disease, the most accurate model found, comparing tests errors, is the generalize additive model. It uses non-linear functions as coefficients for the predictors.

A shown in the method part, cubic functions must be assigned to the variables chol and oldpeak, while all the other variables can sufficiently have linear function.

This project wasn’t meant to be ground-breaking research in and on its own. We hope to support current existing research in the field and ideally supplement it with our own insights and perspectives. At the very least we’d like to confirm past results and use the opportunity to learn about public health.  
It is our goal to raise awareness on heart diseases as they are the number one cause of death worldwide and as can be seen from our data can’t just easily be predicted by just sex and age. They can often only be predicted when the first symptoms start to show, which can fatal.

We are happy to have had the opportunity to explore the dataset and learn something about heart disease and how it can be prevented and are looking forward to receiving feedback.

1. 1 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981580/> [↑](#footnote-ref-1)
2. 2 https://www.kaggle.com/ronitf/heart-disease-uci [↑](#footnote-ref-2)
3. 3 <https://www.heartfoundation.org.au/your-heart/heart-conditions/angina> [↑](#footnote-ref-3)
4. 4 <https://www.heartfoundation.org.au/your-heart/know-your-risks/blood-pressure/is-my-blood-pressure-normal> [↑](#footnote-ref-4)
5. 5 <https://www.medicalnewstoday.com/articles/321519.php#health-impact-of-serum-cholesterol> [↑](#footnote-ref-5)
6. 4 slide 21, lecture 8, COMP33654 Fall semester 2019 [↑](#footnote-ref-6)