Johanna Kasenurm
Elisabet Hein

HEART DISEASE PREDICTION

BACKGROUND

Heart diseases, also known as cardiovascular diseases (CVDs) are causing the most deaths around the world. It is argued that CVDs are among the deadliest diseases in both developed and developing countries. CVDs take an estimated 17.9 million lives each year which is about 32% of all deaths globally. However, the chance of survival is higher if the diagnosis is made early enough. Therefore machine learning has become one of the most important tools in solving this problem by helping to predict which people are at risk of heart diseases.

GOALS

- *Explore diverse machine learning techniques*
  - *Random forest*
  - *Decision tree*
  - *K Nearest Neighbors*
  - *Neural Network*
- *Optimizing all those models*
- *Look deeper into features and see which factors influence or boost the development of the diseases the most.*
- *Write a programm to predict the presence of heart disease based on user input*

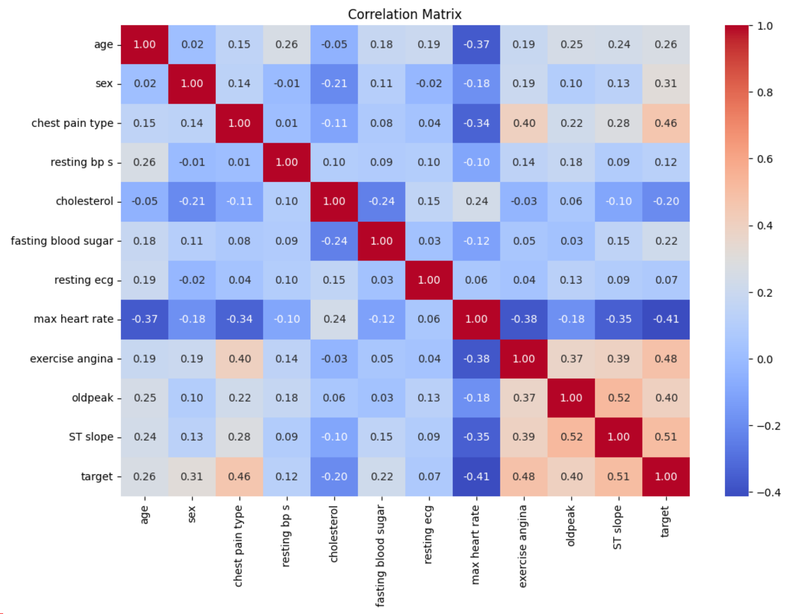
DATA

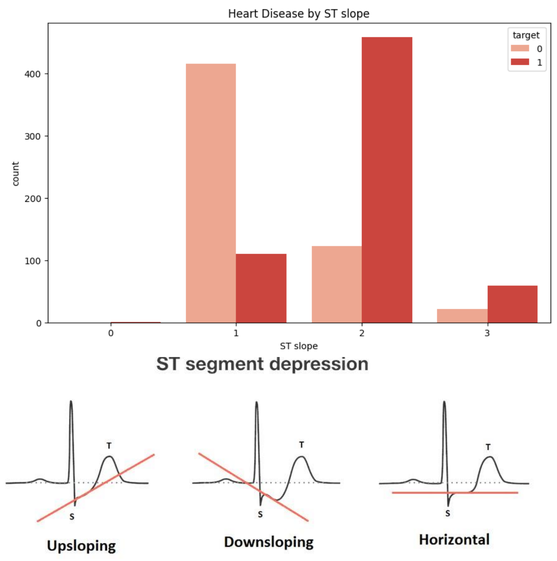
Our database has data from different datasets, which have been combined. There is data from Cleveland with 303 instances, from Hungary with 294 instances, from Switzerland with 123 instances, from Long Beach VA with 200 instances and from another general dataset with 270 instances. The features are:

- *age (numeric)* - Patient's age in years.
- *sex (nominaal)* - Patient's gender. Male is 1 and Female is 0.
- *chest pain type (categorical)* - 1 - typical; 2 - typical angina; 3 - non-angina; 4 - asymptomatic.
- *resting bs s (numeric)* - Level of blood pressure at resting mode in mm/GG.
- *cholesterol (numeric)* - Serum cholesterol in mg/dl.
- *fasting blood sugar (nominal)* - Blood sugar levels on fasting > 120 mg/dl represent 1 in case of true and 0 as false.
- *resting ecg. (categorical)* - 0: Normal, 1: Abnormality in ST-T wave, 2: Left ventricular hypertrophy
- *max heart rate (numeric)* - Maximum heart rate achieved
- *exercise angina (nominal)* - Angina induced by exercise 0: no and 1: yes
- *oldpeak (numeric)* - Exercise induced ST-depression in comparison with the state of rest.
- *ST slope (categorical)* - 0: Normal 1: Upsloping 2: Flat 3: Downsloping
- *target (nominal)* - Heart risk, 1: heart disease 0: normal

CORRELATIONS

The negative correlation with max heart rate supports the idea that lower maximum heart rates may be associated with heart disease, which is a common belief that ageing could cause risk for different diseases. The positive correlation between chest pain type, exercise angina, oldpeak and ST slope suggests that individuals who experience atypical pain (not typical chest pain) in the chest or angina during exercise have more likely heart diseases and individuals who have abnormalities in ST slope too. The matrix shows that the biggest influence on heart-related diseases may be caused by abnormal ST slope.





MOST IMPORTANT FEATURES

From the correlation matrix, we have found the biggest association between the ST segment slope and the risk of heart disease. The graphical representation on the left further accentuates this finding, suggesting that individuals with a downsloping ST segment (value 3) are approximately three times more likely to have a heart disease. This observation aligns with expectations, considering that, in many cases, a normal ST segment slope trends slightly upwards—a reflection of the critical mechanics required for optimal heart function.

RESULTS

As a result, we were able to create models with an accuracy of over 92%. Random Forest has the highest accuracy among the models we created, with a score of 0.9454.
There isn't a significant difference between all the models. If there was it could suggest overfitting where a certain model learns the training data too well and fails to generalize to new, unseen data. We see that some optimizations we tried didn't make much difference in the accuracy. For example, feature importance with the random forest model underperformed and also the regulations for Neural networks didn't perform better.
So in conclusion the random forest model was our best model with an accuracy of 0.9454, which meets our goal to train a model with an accuracy over 0.92.