# 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.  
Individual level practices … (see who print-out)

There have also been previous efforts to use a data-driven approach such as <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981580/> (look up how to quote correctly), 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

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

Shows the age of the individual

1. sex

Shows 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 (chest-pain type)

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 (<https://www.heartfoundation.org.au/your-heart/heart-conditions/angina>) Over time a fatty material called plaque builds up in your coronary arteries, making them become narrow. This reduces the blood flow to your heart, and sometimes it may not get as much blood as it needs. (1:1 quote, adjust)

1. trestbps

Resting blood pressure measured in mmHg on admission to the hospital

Optimal values are below 120, high blood pressure starts at 140

(<https://www.heartfoundation.org.au/your-heart/know-your-risks/blood-pressure/is-my-blood-pressure-normal>)

1. chol

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 bloodflow and oxygen from reaching organs and tissues. (<https://www.medicalnewstoday.com/articles/321519.php#health-impact-of-serum-cholesterol>)

1. fbs

Fasting blood sugar (fbs) levels are measured by making the individual not consume anything but water for eight hours. 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. As the goal of the dataset isn’t to diagnose diabetes but risk of heart disease, the criterion is if the value is greater than 120mg/dl. If so then 1 is returned, otherwise 0.

1. restecg

Resting Electrocardiographic Results are 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.

1. thalach

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

1. exang

Exercise induced angina (exang) is a quantitative variable showing if the individual has suffered of an angina through exercising.

1. oldpeak

oldpeak is a value, which is either an integer or float 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

The slope value 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

The ca value is also 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 anemia.   
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. num

The num value determines if the individual does have a cardiovascular disease or not. It’s qualitative values are set to 0 or 1 accordingly.

####MAKE UNIFORM, SENTENCES, ADD IF QUALITATIVE OR QUANTITATIVE ######

Analytical Procedure

Our main objective was to find the model with the highest accuracy for predicting heart disease.

# RESULTS

The Heart Disease Dataset from the UCI Machine Learning Repository consists of 303 entries out of which 165 are cases of cardiovascular diseases and 136 belong to the control group. This amounts an approximate 54.46% of cases.  
The age of the participants ranges from 29 to 77, with the average person being 54.37 years old. Out of the 303 participants 207 were male and 96 were female.

Due to the dataset being used in previous research, it has already been pre-processed, making it easier to work with.

####ADD PLOTS, Descriptions###  
Say something about chol outlier?

Sources

<https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)> (WHO on Heart disease (statistics, etc.)

<https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-15/prevention-of-cardiovascular-disease-recent-achievements-and-remaining-challeng> (not used yet)

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981580/> (predicting CVDs with a Bayesian classifier)

<https://apps.who.int/iris/bitstream/handle/10665/94384/9789241506236_eng.pdf;jsessionid=DFB5ACC1C246C34F2879CEAA2551504A?sequence=1> (WHO noncommunicable diseases)