# Heart Disease analysis with classification

**Authors:** Lola Sõukand, Robin Mürk

**Data:** <https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/discussion>

Aim is to build a binary classifier that predicts a person's risk of a heart attack and to analyse individual attributes to determine what lowers and what increases the risk of heart attack using regression analysis.

**Goal 1:** train a model to predict whether or not a patient is at high risk or at low risk of a heart attack.  
**Goal 2:** find factors, that lower the risk of heart attack  
**Goal 3:** find factors, that increase the risk of heart attack

## Gathering data

* Outline data requirements
* Verify data availability
* Define selection criteria

## Describing data

* HeartDiseaseorAttack
  + Label
  + Binary
  + 1=yes
  + 0= no
* HighBP
  + Indicates if the person has been told by a health professional that they have High Blood Pressure
  + Binary
  + 1= yes
  + 0= no
* HighChol
  + Indicates if the person has been told by a health professional that they have High Blood Cholesterol
  + Binary
  + 1= yes
  + 0= no
* CholCheck
  + Cholesterol Check, if the person has their cholesterol levels checked within the last 5 years
  + Binary
  + 1= yes
  + 0= no
* BMI
  + Body Mass Index, calculated by dividing the persons weight (in kilogram) by the square of their height (in meters).
* Smoker
  + Indicates if the person has smoked at least 100 cigarettes.
  + Binary
  + 1= yes
  + 0= no
* Stroke
  + Indicates if the person has a history of stroke.
  + Binary
  + 1= yes
  + 0= no
* Diabetes
  + Indicates if the person has a history of diabetes, or currently in pre-diabetes, or suffers from either type of diabetes.
  + Non-Binary
  + 2= type 2
  + 1= type 1
  + 0= no
* PhysActivity
  + Indicates if the person has some form of physical activity in their day-to-day routine.
  + Binary
  + 1= yes
  + 0= no
* Fruits
  + Indicates if the person consumes 1 or more fruit(s) daily.
  + Binary
  + 1= yes
  + 0= no
* Veggies
  + Indicates if the person consumes 1 or more vegetable(s) daily.
  + Binary
  + 1= yes
  + 0= no
* HvyAlcoholConsump
  + Indicates if the person has more than 14 drinks per week.
  + Binary
  + 1= yes
  + 0= no
* AnyHealthcare
  + Indicates if the person has any form of health insurance.
  + Binary
  + 1= yes
  + 0= no
* NoDocbcCost
  + Indicates if the person wanted to visit a doctor within the past 1 year but couldn’t, due to cost.
  + Binary
  + 1= yes
  + 0= no
* GenHlth
  + Indicates the persons response to how well is their general health, ranging from 1 (excellent) to 5 (poor).
  + ?
  + 1= excellent
  + 2=
  + 3=
  + 4=
  + 5= poor
* Menthlth
  + Indicates the number of days, within the past 30 days that the person had bad mental health.
* PhysHlth
  + Indicates the number of days, within the past 30 days that the person had bad physical health.
* DiffWalk
  + Indicates if the person has difficulty while walking or climbing stairs.
* Sex
  + Indicates the gender of the person, where 0 is female and 1 is male.
  + Binary
  + 1= male
  + 0= female
* Age
  + Indicates the age class of the person, where 1 is 18 years to 24 years up till 13 which is 80 years or older, each interval between has a 5-year increment.
* Education
  + Indicates the highest year of school completed, with 0 being never attended or kindergarten only and 6 being, having attended 4 years of college or more.
* Income
  + Indicates the total household income, ranging from 1 (at least $10,000) to 6 ($75,000+)

## Exploring data

## Verifying data quality

Consult the above-given book chapter to understand what is expected under all these deliverables. Take inspiration from when describing and exploring the data. As a result of this exercise, you should have gathered and understood the data. You should have decided which parts of the data you will use and understood the meaning of all fields within these parts. Note that data cleaning is part of the data preparation step in CRISP-DM, but you might choose to do some of it already during this task.

**The report of task 3 should be 400-800 words.**