Project Proposal

Cardiovascular Disease dataset

The system will use 13 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The system will predicts the likelihood of patients getting Cardiovascular Disease. It will enables significant knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established. I will employ logistic regression, and a bunch of other models .

The obtained results will hopefully illustrate that the designed diagnostic system can effectively predict the risk level of Cardiovascular diseases.

Question/need:

here are some framing questions that hopefully will be answered in my analysis :

- What is the most powerful predictor of cardiovascular disease?
- can smoking be a reason for getting a cardiovascular disease?
- What are three known predictors of cardiovascular disease?
- What is the main cause of cardiovascular disease?

Who will benefits from exploring this question or building this model/system are researchers, and doctors, or health professionals

Data Description:

The dataset consists of 70 000 records of patients data, 11 features + target.

There are 3 types of input features:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.

Tools:

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting and visualization .