

1 Assignment 1

- Deadline 27.10.2019, 23:59
- All files need to be available through your GIT repository, in the directory “Project 1”
- You can work in teams up to 3 people. Please state in the report the group member names.

5.1 The Data

We will use the Heart Disease Data Set

(„The Cleveland Database“, see <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>)

This database contains clinical and bio-medical attributes for more than 300 patients. The "goal" field refers to the presence of a heart disease in a patient. It is integer valued from 0 (no presence) to 4.

5.2 The Task

- You are supposed to develop (or: train) five classifiers to diagnose a heart disease, based on the available data.
- You can start with this example code:
 - <https://www.kaggle.com/heyricl/clinical-data-explanation-and-standard-eda-95>
 - <https://medium.com/@wanjirumaggie45/data-science-for-good-machine-learning-for-heart-disease-prediction-289234651fed>
- The main difference from this task and the example code is:
 - You need to train five classifiers and compare them.
 - The data used in the example code is different to the original one (which can be found in the UCI machine learning repository). Your code needs to work with the original data.

5.3 Deliverables

Your need to upload all source codes and a report to your GIT repository.

If your code is in python, I must be able to run your code within a Google Colab notebook (<https://colab.research.google.com> – so you can test it yourself). If your code is not in Python or R, you must provide a manual how to compile and run it on a Linux machine.

- The report should be about 600-1200 words in length (this is roughly 1-2 pages, depending on your layout).
- The report must be delivered in PDF format.
- The report must contain a screenshot of the final classifier results.
- The following sections must be present (you can add more if needed):
 - Scientific background of the project
 - Goal of the project
 - Description of the data and the data-preprocessing steps
 - Description and comparison of the used methods
 - Results
 - Discussion: why is this a typical project for a data-scientist? (Or why not?)