Introduction to Focus Areas in Bioinformatics – WS21/22
Lecturers: Alexander Bockmayr, Michael Grünstäudl, Tim Conrad, Robert Preissner, Torsten Semmler, Heike Siebert, Martin Vingron, Jana Wolf

Assignment 1

- Deadline 30.10.2021, 18:00
- All files need to be available through your GIT repository, in the directory "Project 1"
- You should work in teams of 4 people. Please state in the report the group member names.

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.

The Task

- You are supposed to develop (or: train) three classifiers to diagnose a heart disease, based on the available data.
- You can start with this example code:
  - PYTHON: https://www.kaggle.com/heyrict/clinical-data-explanation-and-standard-eda-95
  - R: 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 at least three 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.

Deliverables

You need to upload (1) all source codes and (2) the report (PDF) 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.
- The report must be delivered in PDF format and written in Latex as described in the lecture.
- The report must contain a screenshot of the final classifier results.
- The following sections **must** be present (you can add more if needed):
  - Abstract (as defined in the lecture)
  - Scientific Background ("Scientific background of the project")
  - Goal ("Goal of the project")
  - Data and Preprocessing ("Description of the data and the data-preprocessing steps")
  - Methods ("Description of the used method(s)")
  - Results
  - Discussion ("Why is this a typical project for a data-scientist? (Or why not?)")
  - Appendix ("Who of the team did what? (Each person one sentence.)")