Introduction to Focus Areas in Bioinformatics – WS19/20

Prof. Dr. Tim Conrad

2 Project 3

- Deadline: 10.11.2019, 23:59
- All files need to be available through your GIT repository, in the directory “Project 3”.
- You can work in teams up to 3 people. Please state in the report the group member names.
- If your code is in python, I must be able to run your code within a Google Colab notebook. If your code is not in Python or R, you must provide a manual how to compile and run it on a Linux machine.

2.1 Data & Goal

We will use the McKinsey Stroke Dataset (see https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data). This dataset contains patient data for more than 60.000 patients. The “stroke" field refers to the diagnosis. The goal is to build a tool that could help medical doctors to diagnose.

2.2 Task 1: Statistics & Data Cleaning

- Gather statistical information about the data, such as summary statistics, attribute distributions, outliers, information about missing data etc.
- Visualize some of the findings.
- Preprocess the data (e.g. for data cleaning) if needed.
- Compare the effect of imputation for the regression model, i.e. how do the results differ if (a) imputation is used and (2) no imputation is used

2.3 Task 2: Prediction

- Implement and compare three approaches to reach the goal:
    - A clustering approach, to analyze the underlying structure of the data.
    - A predictor to predict the probability of a stroke happening to a patient.
    - A classifier to classify a patient as “no danger of stroke” vs. “danger of stroke”.
- Evaluate the predictor and classifier, e.g. by calculating their accuracy using an independent test-set.
- These example codes might help you:
    - https://www.kaggle.com/njalan/healthcare-dataset-stroke-data-pyspark
    - https://github.com/aman1002/McKinseyOnlineHackathon-Healthcare-
    - https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data/kernels

2.4 Deliverables

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

- The report should be about 600-1200 words in length (this is roughly 1-2 pages, depending on your layout) and must be delivered in PDF format.
- The report must contain a screenshot of the evaluation of the predictor and classifier.
- The following sections must be present (you can add more if needed):
    - Background and goal of the project
    - Description of the data (including summary of the data statistics)
    - Preprocessing steps (e.g. cleaning)
    - Results (including description and comparison of the three approaches)
    - Effect of imputation (comparison of the predictor trained w/ and w/o imputed data)
    - Discussion: why is this a typical project for a data-scientist? (Or why not?)