Introduction to Focus Areas in Bioinformatics – WS21/22

Lecturer: Priyanka Banerjee, Project Idea: Tim Conrad

4 Project 4

- Deadlines: For the **REPORT**: 20.11.2021, 18:00.

4.1 The Data

We will use the McKinsey Stroke Dataset (available from the course Whiteboard page). This dataset contains patient data for 60.001 patients. The “stroke" field refers to the diagnosis. You have two different datasets: Training and test dataset.

4.2 Task 1: Data Exploration

- USE SPARK / SPARK SQL methods for the following sub-tasks, if possible.
- Read-in the data and gather statistical information about the data, such as summary statistics, attribute distributions, correlation between attributes, outliers, information about missing data etc.
- Present and discuss your findings in the report (at least 3 things) and use appropriate visualization if possible.
- Preprocess the data (imputation, sub-sampling, etc.). Describe in the report in detail what you did.
- Use at least two imputation methods. Analyze and compare the results.

4.3 Task 2: Prediction

- USE SPARK methods for the following sub-tasks, if possible.
- Develop, evaluate and compare two predictors to predict the probability (!) of a stroke happening to a patient. (If predicting the probability does not work, do classification.)
- Do this once with the unbalanced and once with a balanced data-set. For the balanced data-set: use and compare both sampling methods discussed in the lecture.
- Some example code that you might find helpful can be found here:
    - https://www.kaggle.com/njalan/healthcare-dataset-stroke-data-pyspark
    - https://github.com/aman1002/McKinseyOnlineHackathon-Healthcare-
    - https://dataxboost.wordpress.com/2018/04/17/mckinsey-online-hackathon-on-healthcare/

4.4 Deliverables

Please send your files too **priyanka.charite@gmail.com** and I will send the feedback about the reports (including your team members name)

- The report should be about 900-1200 words in length.
- The report must be delivered in PDF format using the usual template.
- The report must contain at least one screenshot of the final classifier results.
- The following sections must be present (you can add more if needed):
    - Scientific background
    - Goal
    - Data
    - Results

      For each task: describe what you did (e.g. which methods you used) and what the results are. Make sure to describe the used methods in detail.

- Discussion 1: Discuss your results.
- Discussion 2: Discuss, why this is a typical project for a data-scientist? (Or why not?)