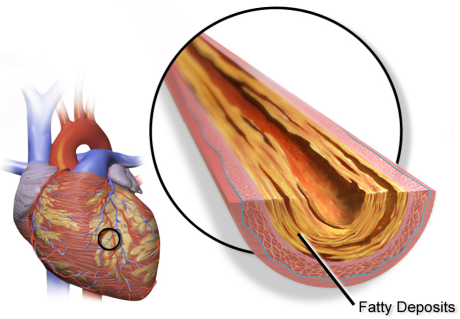
IML'22 Project: Heart Disease Dataset



Background. Coronary artery disease (CAD), the most common cardiovascular disease, is a reduction in blood flow to the heart. Symptoms include chest pain or discomfort that may extend into the shoulder, arm, back, neck, or jaw. Heart attacks are often the first sign, although heart failure and abnormal heartbeat can also occur. There are many risk factors such as high blood pressure, smoking, diabetes, obesity, high blood cholesterol, poor diet, depression, and excessive drinking. Electrocardiograms and cardiac stress tests are among the tests that can help diagnose. The primary task of this project is to determine whether a patient has heart disease or not based on their given attributes, while the experimental task is to diagnose and gain more insights from this dataset that may help in a better understanding of coronary artery disease (CAD).

Dataset. The Heart Disease Dataset is a publicly available UCI Machine Learning Repository dataset. Four databases are included in the dataset. Nevertheless, we will focus on the Cleveland database (Detrando et al. 1989). The database, which can be downloaded from here, contains 303 patients with 14 numerical value attributes. such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic, and maximum heart rate, among others. You are flexible to use the other databases if needed. However, this has to be agreed on at the group level.

Requirements.

1. Formulate the machine learning problem by identifying the task (classification or regression), data, and challenges.
2. Perform Exploratory Data Analysis. This includes:
    - Plot relationship between variables
    - Identify correlated features or the ones that highly correlated with the label/outcome, if any
    - Perform data imputation for missing variables
    - Encode your outcome to a one-hot vector, if needed
    - Remove redundant variables by dimensionality reduction techniques
3. Consider splitting the data into Training, Validation, and Testing sets. The suggested split is 70%, 10%, and 20% held-out testing set, respectively. (same split for all tasks)
4. Build and develop the following models (tasks):
    - **Task#1:** Clustering in an unsupervised fashion
    - **Task#2:** Logistic/Linear Regression according to your formulation in step 1
    - **Task#3:** Neural Network for Classification/Regression according to your formulation in step 1

5. For each task, run the model *with* and *without* processing the data, e.g., without normalization or dimensionality reduction, and compare the model's performance after you normalize and/or reduce the dimensionality of the data.
6. For each task, show how you performed the *model selection*. For example, demonstrate the performance of variants of your model with different hyper-parameters, e.g., number of clusters and initialization when it comes to clustering methods.
7. For each task, perform a 5-fold Cross Validation.
8. For each task, run the corresponding evaluation metrics on each fold to demonstrate the performance of your model on the held-out testing set

Submission.

- **Prepare** a well-documented Jupyter notebook demonstrating the following sections:
    - Your name along with your Uni-ID and task assignment within the group
    - The background of the given task
    - Sample of your dataset
    - The rationale behind your ML formulation
    - The exploratory data analysis and data split
    - The developed models (Task #1-#3) including the model selection (hyper-parameters) and their results
    - A brief comparison between different models + concluding remarks
    - Lessons learned from this project
- **Consider** the following naming convention when you name your Jupyter notebook, `HeartDiseaseDataset_G##`, and consider replacing ## with your group number, e.g, if your group number is 5, then the name of your file should be `HeartDiseaseDataset_G05.ipynb`
- **Submit** the Jupyter notebook to the following folder https://uni-bonn.sciebo.de/s/tO0R2lXI1RamUWB which allows you to only upload files. You can upload the same file multiple times in case updates/modifications were required.
- **The submission deadline** is 18th Jan. 2023 at 11:59 PM

Grading.

- **Individual-level grading (65%) + Group-level grading (35%)**
    - The problem addressed (clearly stated and understood) – *Group-level*
    - ML problem formulation (rationale) – *Group-level*
    - Exploratory Data Analysis – *Group-level*
    - Task Completion and Findings – *Individual-level*
    - Interpretation and conclusion – *Individual-level*
    - Limitations – *Group-level*
    - Documentation, Submission, and Presentation – *Group-level*