INFO411 Assignment 2: Data Analysis and Modelling (40 marks)

J Deng, jddeng@ieee.org

September 11, 2023

This assignment is a group project. You are given the datasets and some rough directions for data processing, analysis and modelling. It is up to you to decide on the exact paths and solutions of your project. You may work as a group; each group may contain up to three members.

Datasets

There are two data sources you can use:

- The Statlog Heart dataset (let's call it "DS1"), available from the UCI Machine Learning Repository[1].
- The Heart Disease Data Set ("DS2"), also available from the UCI Machine Learning Repository[2].

DS1 is actually a subset of the much larger DS2, but contains no missing values. It has been used widely as a benchmark. This should be the dataset you start with.

However, it is expected that you should use DS2 for the main body of your work.

It contains a very clean "cleveland" dataset, as well as datasets from other locations that contain some significant amounts of missing values on various attributes.

The `location.data` files contain the full data for each *location*. Many attributes have missing values, some are *not used*.

The `processed.`*location*`.data` files contain 14 attributes only (and less missing values), the same as those in DS1.

Tasks

A number of perspectives can be considered.

Exploratory data analysis (10 marks)

You may use DS1 to gain some insight on the statistical properties of the data. This EDA part may look at the probability distributions of various attributes, their correlations, visualization, and/or find potential clusters (if any).

Some further exploration of a larger scope of attributes in DS2 may be attempted.

Report on insights gained from your EDA experiments.

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[2] https://archive.ics.uci.edu/ml/datasets/heart+disease

Data imputation (5 marks)

Impute at least three datasets (the `processed.location.data` files) from DS2. Using the imputed datasets, contrast their statistical profiles. Use a subset of 14 features if necessary.

You may use the `Impute.jl` package [3] to carry out imputation, or search [4] on MLJ interfaces for more options.

Modelling (10 marks)

You can build predictive models for classification (binary prediction on the `heart_disease` attribute) or regression (predicting on the 0 - 4 scale of the same attribute). Carry out performance validation and tuning. Use full or partial sets of features where applicable.

You may experiment with the clean DS1 data, or the imputed `processed.location.data` files, or both.

Find out whether your prediction models can generalize well on data obtained from other locations. Explore potential improvement of the model performances.

Deliverables (15 marks)

You may build separate Pluto notebooks for different experiments, depending on their purposes, length or relevance. Include them in your submission.

Dashboards (10) In addition, you should build a few Pluto *dashboards* as highlights of your EDA or modelling work.

For a simple demo of a Pluto dashboard, see the demo code on Piazza's Resources page (Resources/Demo notebooks). You may refer to a detailed JuliaCon 2022 talk by G. Haetinger [5].

Report (5) In addition to the Pluto notebooks, you should submit a short report where you summarize your work and reflect on your main findings. List each group member's main contributions.

Submission

Each group must prepare a private github repository that contains *all* used data files (original, processed, compiled etc.), your Julia notebook(s) with the full source code (commented where necessary), the final Report, and a README.md file that outlines repo content.

Share your repo with "jddeng@ieee.org" by **Friday 6/10 11:59pm**. No further commits to the repo are allowed after the deadline.

There will be an interview that *all* group members should attend (to be scheduled on Wednesday 11/10 and Thursday 12/10) . It is expected that *all* members make significant technical contributions to various parts of this group project.

[3] `https://invenia.github.io/Impute.jl/stable/`

[4] `https://alan-turing-institute.github.io/MLJ.jl/dev/search/?q=imputer`

[5] `https://www.youtube.com/watch?v=dP9UuEL00iM`