

Requirement:

Using Rapidminer, you are required to:

- a) Read in the heart-Training data from the heart-Training.csv file. (10 Marks)
- b) Clean the data in the CSV file. (30 Marks)
- c) Develop four separate models that can classify the target label using the cross validation technique. (30 Marks)
- d) Identify the best model using suitable techniques. (15 Marks)
- e) Apply the best model to the unseen data file (heart-Unseen.xlsx) and predict which of the patients could possibly have heart disease. (15 Marks)

Notes:

1. The relevant data files can be found in the Work Package 2 folder on the Moodle webpage. There are two files saved there, they are :
 - a. Heart-Training.csv
 - b. Heart_Unseen.xlsx
2. The data set is a small data set, there are only 1,000 samples in the training data set.
3. All pre-processing of data and model generation / comparison must be carried out using the Rapidminer software. Wherever possible and relevant, you should document/annotate all of the operators.
4. Marks will be deducted (up to 10%) for cluttered presentation of Rapidminer processes.
5. Please ensure that you submit your .rmp file (containing all of the relevant operators) through the Moodle link setup for Workpackage 2. Failure to do so will result in a zero grade.
6. The due date for this assignment 29th March 2021 @ 09:00

Context

The "goal" of this dataset is to identify the presence of heart disease in the patient. The Target column is integer valued from 0 (no presence) to 1 (Heart Disease present).

Content

Attribute Information:

- | | |
|---------|---|
| 01. Age | Patient's Age in years (Numeric) |
| 02. sex | Patient's Gender Male as 1 Female as 0 (Nominal) |
| 03. cp | Type of chest pain categorized into 1 typical, 2 typical angina, 3 non-anginal pain, 4 asymptomatic |

04. resting bp s	Level of blood pressure at resting mode in mm/HG (Numerical)
05. cholesterol	serum cholestorol in mg/dl
06. fbs	fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
07. restecg	electrocardiographic results (values 0,1,2)
08. Max heart rate	maximum heart rate achieved
09. Exercise angina	exercise induced angina
10. oldpeak	ST depression induced by exercise relative to rest
11. slope	the slope of the peak exercise ST segment
12. target	heart disease detected (1 = true, 0 = false)