229352 Statistical Learning for Data Science 2 **Fall 2021**

Lab 3: due Sunday July 25

Instructor: Donlapark Ponnoprat

Note: The report must be turned in as a PDF file. If you are using the Python Notebook, go to: `File>Download as>pdf(.tex)`.

In this lab, we are going to create a decision tree model in order to predict whether a patient has a heart disease or not.

The data that we will be using is stored in `heart_disease.csv`. The description of this data is appended at the end of this file.

Do the following tasks.

1. Use any method to deal with the missing data. Then split the data into a training set and a test set.
2. Apply a grid search or a random search via cross-validation on the training set to find the best criterion for node's purity (gini index or entropy) and the optimal value of pruning hyperparameter α. You can also include some other hyperparameters (tree's depth, minimum number of samples in each leaf etc.) if you want.
   Note: Try small values of α—it should be somewhere between 10^{-4} to 10^{-1}.
3. Plot the tree model that you just obtained. Report all the features used in the classification. What is the most important feature that indicates that a patient has a heart disease?
4. Report the accuracy (and possibly any other classification scores) of your predictions on the test set.

Data description

The data consists of following 13 features and 1 label.

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

- restecg: resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: downsloping
- ca: number of major vessels (0-3) colored by flourosopy
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
- label: diagnosis of heart disease (1 = positive; 0 = negative)