CENG 484 - Data Mining
Exercise 2

Try to implement rule-based classification on heart diseases data. It will be detect whether a person has heart diseases according to some features. The data contains significant information about the diagnosis of the disease such as age, chest pain.

This data was uploaded as “heart_data.csv”. You can choose Python or R for coding. Attributes of data are as follows:

- age (numeric)
- sex: 1 = male; 0 = female
- cp: chest pain type
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholesterol in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results
- 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
- ca: number of major vessels (0-3) colored by flourosopy
- thal: normal; 6 = fixed defect; 7 = reversable defect
- target: person with heart disease = 1 or healthy = 0

Rules will be extracted from the data by selecting a few of the above features. The following will be done in this exercise:

a) Read data and select only “age”, “cp” (chest pain), “trestbps” (blood pressure), “thalach” (maximum heart rate) and “chol” (cholesterol) features.

b) Write a code to replace the age values: if age>55 new value of age “older person” else new value of age “younger person”.

c) You should convert all categorical values to **numerical values**. You can use label encoder in Python.

d) Split the data into two subsets: training data (80%) and testing data (20%).

e) Apply **RIPPER** algorithm to create rules from the data. You can use available library such as wittgenstein in Python.

f) Create a **decision tree** classifier.

g) Compare algorithm **running times (milliseconds)** and classification performances according to area under the curve (**AUC**) value.

Note: This exercise will be graded as +10 bonus points for the assignment, you can improve your practice and prepare yourself for the assignments. Answers will be shared on May 6.