CENG 484 - Data Mining
Exercise 1

Try to implement a decision tree model that predicts if the client will subscribe a term deposit at the bank. The data contains marketing phone calls of a real Portuguese banking institution. Often, more than one call to the same client is required, in order to convince the term deposit. Term deposit is a product of a bank. It is type of deposit account held at a bank where money is locked. You will predict whether the customer accept (yes answer) or not accept (no answer) the term deposit based on some features such as age, job, education.

This data was uploaded as bank_customer.csv. You need to construct a decision tree classification model to make predictions in Python or R. Attributes of data are as follows:

- age (numeric)
- job : type of job (categorical: such as 'admin.','blue-collar')
- marital : marital status (categorical: 'divorced','married','single')
- education (categorical: such as 'basic.4y','basic.9y','high.school')
- default: has credit in default? (categorical: 'no','yes','unknown')
- balance: the amount of money in a deposit account (numeric)
- housing: has housing loan? (categorical: 'no','yes','unknown')
- loan: has personal loan? (categorical: 'no','yes','unknown')
- contact: contact communication type (categorical: 'cellular','telephone')

- day: last contact day of the month (numeric)
- month: last contact month of year (categorical: such as 'jan', 'feb', 'mar')
- duration: last contact duration, in seconds (numeric)
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: such as 'failure','success')
- deposit: has the client subscribed a term deposit? (binary: 'yes','no')

The following will be done in this exercise:

a) Combine similar “job” as white-collar, pink-collar, other and print all jobs with counts after combination.

    - management + admin -> white-collar
    - services + housemaid -> pink-collar
    - retired + student + unemployed + unknown -> other

    Combine “poutcome” values as other + unknown -> unknown, so obtain 3 types unknown, failure, success. Print all types with counts.

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

c) Split the data into two subsets: training data (70%) and testing data (30%).

d) Create data by selecting 8 attributes: “age”, “job”, “marital”, “education”, “balance”, “housing”, “duration”, “poutcome” attributes and name it “**data_1**”.

As an alternative, create the second data by selecting only 4 attributes: “job”, “marital”, “education”, “housing”, name it “**data_2**”.

e) Train a decision tree (with data_1 and data_2) using **entropy**, make predictions on the test data and calculate training and testing accuracy.

f) Train a decision tree (with data_1 and data_2) using **gini index**, make predictions on the test data and calculate training and testing accuracy.

g) Try to make **pruning** by changing **depth limits** parameter, observe changes in accuracy.

h) Calculate 95% **confidence interval** (z = 1.96) for accuracy values, find lower and upper p values.

i) Display the decision trees obtained in steps e and f.

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 April 29.