# Programming for Data Analytics
# Project-II

Lecturer: Dr Mohammed Hasanuzzaman

Saturday 7th December, 2019

Saturday 7th December, 2019

## 1 Third Assessment. Project-II

The bank dataset ('bank.csv') is related with direct marketing campaigns of a Portuguese banking institution. The dataset has 19 features with different types of values.

1. age (numeric)
2. job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self employed','services','student','technician','unemployed','unknown')
3. marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','univers
5. default: has credit in default? (categorical: 'no','yes','unknown')
6. balance: balance in the account
7. housing: has housing loan? (categorical: 'no','yes','unknown')
8. loan: has personal loan? (categorical: 'no','yes','unknown')
9. contact: contact communication type (categorical: 'cellular','telephone')
10. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
12. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed.

13. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

15. previous: number of contacts performed before this campaign and for this client (numeric)

16. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') social and economic context attributes

17. bank_arg1: An argument that is internally used by bank

18. bank_arg2: An argument that is internally used by bank

19. y - has the client subscribed a term deposit? (binary: 'yes','no')

## 1.1 Project Specification

The objective of this project is to provide an insight into the underlying pattern of the dataset such as relationship between features, feature prediction and etc. Please perform the following tasks:

1. Create two data-sets as follows:

   (a) Dataset 1: **age**,**job**,**poutcome**, **balance**, **default** and **y**, where **y** is the class attribute.

   (b) Dataset 2: **age**,**job**,**poutcome**, **balance**, **default** and **loan**, where **loan** is the class attribute.

   Apply a classification algorithm on each dataset and report the error. Which dataset has a higher accuracy (lower error)? Note, all the data should be used for training the model from both dataset

2. Create a dataset by using: *age* and **marital**. Apply an unsupervised learning algorithm and cluster all the individuals in the dataset. Determine the optimal number of clusters/groups for your dataset using an appropriate visualization technique.

3. *bank_arg1* attribute is a continuous valued attribute. Convert this attribute to discreet valued attribute using $n$ buckets. Create a dataset using *y*, *loan* and *bank_arg1*. Consider *bank_arg1* as the class attribute (target class). Apply a supervised learning algorithm to train a model for predicting the class attribute. Test the algorithm with different values for **n** and report which one has the better accuracy.

4. Create a dataset using **age**,**job**,**marital**, **education**,**loan** and **y**. Consider **y** as your class attribute (target attribute).

   Apply the following learning algorithms on your dataset:

   - KNeighborsClassifier
   - DecisionTreeClassifier
   - GaussianNB
   - SVM
   - RandomForestClassifier

   Apply cross-validation technique, when test data is %25 of the whole data, and conclude which technique is the best one and why?

5. Use **bank_arg1** and **bank_arg2** attributes and apply an unsupervised learning algorithm to do the following:

   - What is the best number of groups of individuals? Depict your answer visually.
   - Visualize your data individuals and assign different color for each group. The centroid of each group should also be visualized.

6. Create a dataset using *housing*, *balance* and *y* Consider **y** as your class attribute (target attribute). Apply decision tree classifier on your data. Depict when the problem of over-fitting happens and show how the problem of over-fitting can be resolved. (Hint: A high accuracy for training data and low accuracy for test data is a sign of over-fitting).

7. Create a dataset using **loan**, **balance**, **y** and *bank_arg1*. Split your data into training and test sets. Apply two supervised learning algorithms: *Decision Tree* and *Random Forest*, to create a model for each algorithm. Calculate the score when predicting the test data. Discuss which of them generates a better score. Note that the *bank_arg1* should be descritized.

Note that visualization plots need to have proper labels and annotations.

## 1.2 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function. Please make sure you fully comment your code. You should clearly be explaining the operation of important lines of code.

Please write your name and student ID as a comment in the designated area in the provided python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the the python file should be named: s1234567.py

Please Go to the Assignment Project -II in Canvas to upload your file.

Once you have submitted your files you should verify that you have correctly uploaded them. It is your responsibility to make sure you upload the correct files.

The deadline for this project is 21st of December at 23:59.

Any question about project should be communicated with Mohammed Hasanuzzaman mohammed.hasanuzzaman@cit.ie or via Canvas.

## 1.3 Rubric

This rubric is subject to change.

1. Correct task implementation (model training, error reporting, visualization if needed etc). (100%)
2. Relatively correct task implementation (model training, error reporting, visualization if needed etc). (70%)
3. Partly correct task implementation (model training, error reporting, visualization if needed etc). (40%)
4. Wrong task implementation. (0%)