**Final-EENG 860**

**Due date: May 11, 2023 @ 11pm**

The attached data base “bank-database.csv” with 41188 samples and 20 attributes (features) is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The data ordered by date (from May 2008 to November 2010).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

**Attribute Information:**

Input variables:

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', 'university.degree', 'unknown')
5. default: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. housing: has a housing loan? (categorical: 'no', 'yes', 'unknown')
7. loan: has a personal loan? (categorical: 'no', 'yes', 'unknown') # related with the last contact of the current campaign:
8. contact: contact communication type (categorical: 'cellular', 'telephone')
9. month: last contact month of the year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day\_ of\_ week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu',' fri')
11. 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. Also, after the end of the call, y is obviously known. Thus, this input should be discarded.

# other attributes:

1. campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact)
2. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)
3. previous: number of contacts performed before this campaign and for this client (numeric)
4. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success') # social and economic context attributes
5. emp.var.rate: employment variation rate - quarterly indicator (numeric)
6. cons.price.idx: consumer price index - monthly indicator (numeric)
7. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
8. euribor3m: euribor 3 month rate - daily indicator (numeric)
9. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21- has the client subscribed to a term deposit? (binary: 'yes' and 'no')

Compare the classification performance (sensitivity, specificity, F1- score, AUC, and total accuracy) for predicting the client subscription using supervised classifiers SVM, Adaboost, Gradient boost, XGboost, and HMM and unsupervised classifiers K-mean clustering and FCM clustering approaches when 20% of the data in each class is used for testing.

Apply sequential forward selection, Minimum Redundancy Maximum Relevance, and Relief approaches to order the features to the 80% evaluation data, and then apply the features based on the order provided by feature selection methods to the classifiers. Start from the first feature from each feature selection method and increase the number of features until you get the best performance for each classifier. Indicate this number as the best number of features and calculate the sensitivity, specificity, F1- score, AUC, and total accuracy of each classifier on 20% test data using the best features. Compare the performance of the classifiers. Plot the total accuracy versus the number of features for each feature selection and classification method using 80% evaluation data. Provide three tables, each showing the features in the order provided with each of the three feature selection methods.

Submit all your results in a PowerPoint presentation along with all codes. Indicate the parameters you used for each classifier in the presentation.

Name the codes with classification method\_Last name\_first name.