Computer Science Department
CS675 – Introduction to Data Science (CRN: 73405)
Fall 2024

Project #2 / Due 12-Nov-2024

The goal of this assignment is to understand the 'power' of various Machine Learning Classification algorithms applied into a dataset. By contrasting these very well-diverse and widely used models within Machine Learning space. The end goal is to find the 'best' algorithm to do the job in quest.

Write up **Python/R code** snippets that will device **6 different classification algorithms** on the same dataset. Namely, apply the following ML models:

1- **Logistic Regression** (LR)

2- **Naive Bayes** (NB)

3- **K-Nearest Neighbors** (KNN)

4- **Decision Tree** (DT)

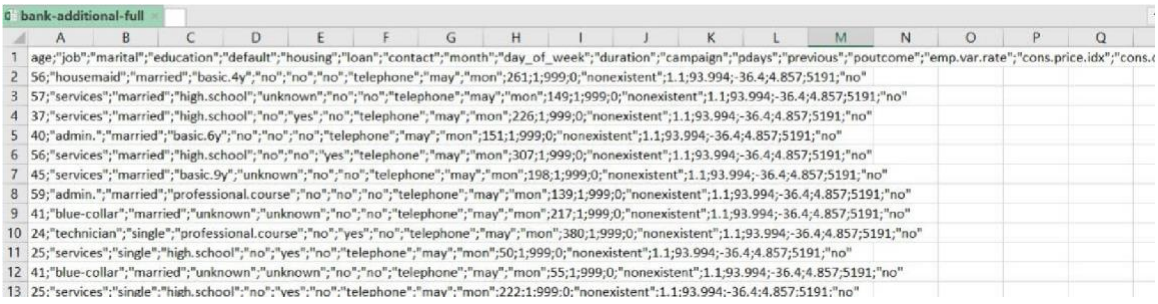
5- **Random Forest** (RF)

6- **XGBoost Algorithm** (XGB)

You should download the following Bank dataset: **Bank Marketing Data Set**
The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

You should get only one (1) dataset, the 'bank-additonal-full.csv' with 45 211 records.
https://archive.ics.uci.edu/dataset/222/bank+marketing
Download the 'bank-additional.zip' file and extract the 'bank-additonal-full.csv' file. Read details of what the variable/features mean.

Here is what the file looks like:



Perform various Machine Learning activities in order to complete the following tasks along with their output. All work should be done and submitted in a single **Jupyter/Colab Notebook.**

1- **Prep the data** in order to be ready to be fed to ML models.

2- Split the source dataset into **training** and **test** datasets at a 70%/30% ratio.

3- Run all algorithms with default values and report their **model performance** on the following metrics: -

- Accuracy
- Precision
- Recall
- F1 Harmonic Mean

4- Generate **Classification Report** (for each model) including Confusion Matrices, ROC Curves, and AUCs.

5- Extra points, rerun some of the models by **tuning** some **hyperparameters**.