Computer Science Department CS675 – Introduction to Data Science (CRN: 22548)

Spring 2025 Final Project / Due 09-May-2025

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 'bank-additional-full.csv' with 41,188 records.

Write Python scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single Jupyter Notebook.

1- Prep the data in order to be ready to be fed to a model. Look for missing, null, NaN records. Find outliers. Transform data – all entries should be numeric.
   List all types of data, numeric, categorical,…
   Perform EDA on data. Present dependencies and correlations among the various features in the data. List the most variables (Feature Importance) that will affect the target label.
   State limitations/issues (if any) with the given dataset

2- Prep the data in order to be ready to be fed to ML models. Split the source dataset into training and test datasets at a 70%/30% ratio. Run all algorithms with default values and report their model performance on the following metrics: - Accuracy - Precision - Recall - F1 Harmonic Mean Generate Classification Report (for each model) including: Confusion Matrices, ROC Curves, and AUCs. Extra points, rerun some of the models by tuning some hyperparameters