CSCI - 4146/6409 - The Process of Data Science - Fall 2022
Assignment 3

The submission must be done through Brightspace.
Due date and time as shown on Brightspace under Assignments.

- To prepare your assignment solution use the assignment template notebook available on Brightspace.
- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- Assignments can be done by a pair of students, or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
- We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). File names should be:
  - **A3-<your_name1>-<your_name2>.ipynb**
  - **A3-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 markings for both students.**

In this Assignment, you will experiment with the ensemble methods from Sklearn API on the Credit card fraud detection

- Use this link or the command below to download the dataset -
  `sklearn.datasets.fetch_openml(data_id: Optional[int] = 1597)`

1. **[0.2]** Data Preparation
   a. Build the data quality report and the data quality plan.
   b. Preprocess your data according to the data quality plan.
2. **[0.4]** Select three models: a strong learner and a bagging model and a boosting model. For each of the models:
   a. Motivate the choice of the model. Explain how your data satisfies the model's requirements.
   b. Perform hyperparameter tuning if applicable.
   c. Train and evaluate your model
   d. Plot the learning curve and analyze it
3. **[0.4]** Model comparison
   a. Analyze the bias and variance of the models

Variance: Captures how much your classifier changes if you train on a different training set.
Bias: What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution.
More information about bias and variance can be found here:

- http://scott.fortmann-roe.com/docs/BiasVariance.html
- https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

Answer the following question:

- What model has the best (worst) bias?
- What model has the best (worst) variance?
- Given this information, which model is preferable for fraud detection and why?

b. **Analyze the impact of the models on imbalance classification**
Imbalance classification is a classification where some classes have more data points than others.
Answer the following question:
   - What model can handle the imbalance classification better?

c. **Analyze the running time for inference and training of each model.**
Using the code provided in the case-study tutorial measure and visualize the running time of training and testing of the models
   - Which model is the fastest with one worker thread?
   - Which model is the fastest with 4 worker threads (use the **n_jobs** parameter of an estimator)?