Computer Science 3202/6915
Assignment 5 – Classification – Do your best!

Goal
Apply classification methods and select the best model.

Due date
Sunday March 28th by 11:30pm.

Specifications
Your program should be called A5_Class.py. You program should take two command-line arguments: A filename specifying a tab-delimited plain-text file containing the training data, and a filename specifying a tab-delimited plain-text file containing the test data.

The data you should use is provided in Brigthspace. For generating your model, you will train with the data given for Assignment 2 in the file A2_training_dataset.tsv.

Once you have your final model, you will predict the likelihood of belonging to class 1 for the observations given in A5_test_dataset.tsv. This is a tab-delimited text file containing 3009 instances. Instances in this file are not included in the training data. The order of the features is the same as that in the training data.

Functionality
Your program should do the following:

1. Read the input data. You can assume the input files are in the working directory. **DONE IN ASSIGNMENT 2**
2. Perform grid-search CV to find the optimal hyper-parameter setting for at least two machine learning methods for classification (you can choose any method available in sklearn).
3. Perform K-fold CV to compare the best models (one model per ML method).
4. Generate a graphical representation of performance of the best models (one model per ML method).
5. Select your best model and create your final model using all of the training data.
6. Using your final model predict the likelihood to belong to class 1 for the test instances provided in the file A5_test_dataset.tsv. The larger the number the more likely the instance is to belong to class 1.
7. Write in a file called A5_predictions_groupXX.txt (where XX is your group number) the predicted values for the instances in the test data. This file must be a text file with a single column containing the likelihood (or confidence value) for the test instances to belong to class 1 in exactly the same order as given in the file A5_test_dataset.tsv

Submission
Submit through Brightspace the following (one submission per team):

a) Your python code in a single file called A5_Class.py
b) A PDF file containing:
    1. short and clear description and justification of any data pre-processing done,
    2. short and clear description and justification of the ML methods evaluated and the grid used for grid search CV,

3. short and clear justification of the performance metric used to select the best model,
4. the best hyper-parameter setting found per method,
5. a table with the CV results for the best model per ML method with mean and standard deviation,
6. a figure showing a graphical representation of the cross-validation performance of the best models (one model per ML method),
7. an acknowledgement section listing your collaborations and online sources, and
8. a program specification section listing the Python version and libraries you used.

c) A text file called A5_predictions_groupXX.txt (where XX is your group number) with your predicted values for the instances in the test data. This file must be a text file with a single column containing the likelihood to belong to class 1 for the test instances provided in the file A5_test_dataset.tsv. The larger the number the more likely the instance is to belong to class 1.

Performance assessment:
25% of the evaluation for this assignment will be graded based on predictive performance. The team(s) with the best performing model (referred below as 1st ranked model) in terms of AUPRC (rounded to 2 decimal places) on the test dataset will receive full marks for the performance of their model (2.5). All other teams will receive a mark for the performance of their model proportional to their decrease in performance with respect to the 1st ranked model. For example, if the AUPRC of the best model of a given team is 10% lower than the AUPRC of the 1st ranked model then their mark for performance of their model will be 2.25.

Common pitfalls to avoid (i.e., DO NOT do the following or points will be deducted):
1. Submit files in a compressed file.
2. Hard coded file names in your code.
3. Fail to include some of the sections in the PDF file.
4. Fail to have the predicted values **exactly in the same order** as given in the file A5_test_dataset.tsv.
5. Forget to acknowledge a collaboration or source.
6. Miss some functionality.
7. Use different data sets.

Online examples/tutorials:
- https://scikit-learn.org/stable/tutorial/statistical_inference/model_selection.html