Computer Science 6915 - Winter 2019
Assignment 3
Do your best!

Following best practices for model selection and research reproducibility, choose a good ML model for classifying the dataset provided in D2L (A3_training_dataset.tsv). This is the same dataset used in Assignment 2 but with some features removed. You need to:

1. Apply at least three different machine learning methods using their implementation available in scikit learn (note that ML methods are not restricted to those seen in class).
2. Select the best parameters and features for each of the ML methods you are evaluating.
3. Predict the likelihood to belong to class 1 for those instances provided in D2L (A3_test_dataset.tsv). The larger the number the more likely the instance is to belong to class 1. 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.

For this task, submit through D2L the following (one submission per team):

a) Your python code to select your best ML model and to generate the predictions in a single file called A3.py. Your program should take as command-line arguments the training and test data, and generate as output a text file with the predictions.

b) A three-page description of how you selected the best model for this task including:
   - a short description of the ML methods evaluated,
   - a short justification of why these methods were selected,
   - an explanation of any data pre-processing step,
   - the range of parameters considered for each ML method,
   - how the best model was selected with a clear definition of the specific measure or statistic used to select the model,
   - a figure showing a graphical representation of the cross-validation performance of the best models (one model per classifier evaluated) with mean and standard deviation.

   This description has to be submitted as a PDF file.

c) A text file with the predicted probability (or confidence score) for each instance in the test set. Make sure the predictions are given for the instances in exactly the same order as in the A3_test_dataset.tsv file.
   The first few lines of a sample file with the predictions might look like this:

```
1.7681105
0.91168957
0.23909187
0.94302765
0.32095104
0.99913759
```

This task will be graded based on whether correct practices for model selection and research reproducibility were followed (40%), quality of the description (35%), and classification performance of the model selected (25%).

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 the model selected (25% as above). All other teams will receive a mark for the performance of the model selected 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 the model selected will be 22.5%.