CS 422/622
INTRODUCTION TO MACHINE LEARNING

FALL 2022
Assignment #6

Due Date/Time: **12/10/2022 @ 11:59PM**
Total Points: **100**

CS 422 students may complete this assignment individually or in teams of two; CS 622 students are to complete it individually.

Description:

- For this assignment, you will implement the ***k nearest neighbors (KNN)*** algorithm in Python or MATLAB to solve a ***classification problem*** for a dataset of your choosing with continuous input and categorical output.
- You may use all built-in functions. The built-in KNN models in Python and MATLAB that can be used for this assignment are:
  - Python: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
  - MATLAB: https://www.mathworks.com/help/stats/classificationknn.html
- Train and test your models with a dataset of your choosing that meets the following criteria:

○ Number of input features:	3+
○ Input characteristics:	Continuous real-valued
○ Output characteristics:	Categorical

- Use 80% of the dataset for training and 20% for testing your model.
- Experiment with ***three (3)*** different ***k*** values.
- If you'd like, you may use one of the following sources to find a dataset:
  - University of California, Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php
  - Kaggle https://www.kaggle.com/
  - Awesome Public Datasets https://github.com/awesomedata/awesome-public-datasets
  - Google Dataset Search Engine https://datasetsearch.research.google.com/
  - Microsoft Research Open Data https://msropendata.com/
  - U.S. Government's Open Data https://www.data.gov/
  - Registry of Research Data Repositories https://www.re3data.org/
  - CMU Libraries https://guides.library.cmu.edu/machine-learning/datasets

- Summarize your approach and results in a report that includes at least the following:
  - The dataset you used, its source and characteristics.
  - The data preprocessing steps you took (if any).
  - A table showing relevant evaluation metrics (accuracy, sensitivity, specificity, f1 score, log loss) for the training dataset for the different ***k*** values.
  - A table showing relevant evaluation metrics (accuracy, sensitivity, specificity, f1 score, log loss) for the test dataset for the different ***k*** values.
  - Any additional details you would like to include.
- Submit your report along with your dataset and source code. Feel free to include your code in the report, but you also need to submit your source code files (.py or .m) and your dataset separately, so that your results can be replicated for scoring.

Submission Instructions:

Compress all files and submit it through Canvas as a **.zip** file.

If you are completing the assignment as a team of two, only one of the team members needs to submit the assignment.

I will set Canvas to allow unlimited number of submissions and will only grade the last submission. So, please do not wait until the last minute to submit as you can always submit an updated version.