CSE391: Programming for data Science
Cancer Classification

The goal of this project is to perform different data analyses on RNAseq gene expression data collected from 500 (COAD) Colon adenocarcinoma tumors (one of prevalent types of cancer) from TCGA data Portal (https://tcga-data.nci.nih.gov/docs/publications/tcga/?)
The project has two parts:

Part I

Perform the following steps:

1- Go to Blackboard , in Assignment, download the file named TCGA_COAD. This file contains the COAD cancer data.
2- Read this file in Python. The first column of the file contains the gene IDs and the first row contains the patient IDs. The numerical values are the gene expression measured using RNA-seq
3- Select randomly 2,000 genes out of 18,026 and 50 patients out of 500 patients in this dataset (so now you have a matrix of 2,000 by 50) and perform the following clustering approach: Use Kmeans clustering and cluster these genes into 20 clusters--- note you cluster genes and not patients. Plot the centers of each cluster and report the number of genes in each cluster. Do you think you need to increase or decrease the number of cluster in order to have a better clustering for these data.
4- Perform the hierarchical clustering on patients and display the dendrogram. How many dominant clusters you can observe.
5- Perform the PCA analysis on patients. Compute how much of variation is explained by each principle component. If you want to compress these data, how many PCs you would choose.

Part II
The goal of this project is to classify the mRNA data pertaining two different subtypes of breast cancer. We measured the expression of 200 genes across 500 patients diagnosed with subtype 1 (250 patients) and subtype 2 (250 patients) breast cancer.

Perform the following steps:

1- Go to Blackboard , in Assignment, download the folder named "cancer classification project. This is in form of a MAT file.
2- Design a SVM classifier, to classify these two groups of subtype cancer.
3- When you upload the data, you have two matrices, X and Y whose dimensions are 200 by 250. Divide the data into the training and test sets. Assign 200 patients for training

and 50 for the test for each class. Train the SVM classifier on the training dataset and report the performance results on the test set.

4- When training the classifier, you have two classes, subtype I (X) and subtype II (Y). You can label all data belonging class I as 0 and all data belonging to class II as 1. Note that you have 250 observation vectors (patients), each has a dimension of 200 for each class.

5- Report the classification results in form of confusion matrix.