**Outbreak investigation using Machine Learning.**

**Due date Oct 3, 2022**

Problem 1: You are given a data set consisting of DNA sequences (the file is available [here](https://drive.google.com/file/d/1AWopOscYefdNsXmEmcDzqrsX8bHzC6KR/view?usp=sharing)) of the same length. Each DNA sequence is a sequence of characters from the alphabet ‘A’,’C’,’T’,’G’, and it represents a particular viral strain sampled from an infected individual. Your goal is to write a code that helps to identify so-called transmission clusters that correspond to ongoing viral outbreaks.

The sequences should be considered as feature vectors and characters - as features. The data set is stored as a fasta file, which is essentially a text file that has the following form:

>Name of Sequence1

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Name of Sequence2

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Name of Sequence3

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

…..

Here each line starting with ‘>’ symbol contains the name of a sequence followed by the sequence itself in the next line.

You may proceed as follows:

1) Read sequences from the file.

2) Calculate pairwise distances between sequences. Use Hamming distance: it is the number of positions at which the sequences are different (see https://en.wikipedia.org/wiki/Hamming\_distance)

3) Project the sequences in 2-D space using Multidimensional Scaling (MDS) based on Hamming distance matrix.

4) Plot the obtained 2-D data points. Estimate the number of clusters *K* by visual inspection.

5) Use k-means algorithm to cluster the 2-D data points.

You may use library functions to read data from the file and perform MDS. For parsing sequence files, see e.g. <http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec11> (it is a big tutorial, but you need only a part about opening fasta files). For multidimensional scaling in python, see e.g. <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html>. For

You can use library functions for k-means clustering see e.g. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> However, if you implement k-means clustering yourself, you will get an extra credit

Your submission should contain:

* The code of your script
* Visualization plots for MDS with different clusters highlighted in different colors.

Please do not hesitate to ask questions.

Problem 2: Use linear regression to predict COVID-19 mortality in the countries worldwide. For each country, let us use the following two features:

1. Percentage of persons over the age of 65 in the population. This information can be found at <https://en.wikipedia.org/wiki/List_of_countries_by_age_structure>
2. The number of hospital beds per 1,000 people in the most recent reported year. This information can be found at <https://data.worldbank.org/indicator/SH.MED.BEDS.ZS>
3. Use any statistics you think it may be related to the number of deaths. You can add more features, but you need to find statistics for all the countries.

The response variable for each country will be the number of COVID-19 deaths per 100K population. This information can be found at https://coronavirus.jhu.edu/data/mortality

To train the model, use 5-fold cross-validation to validate the model. It means that you should split the countries into 5 groups. For each group, train the model on the remaining groups, apply the trained model to predict the number of death for countries in that group and calculate the prediction error for them.

An average error over all iterations will be an assessment of the quality of linear regression model.

You can use Linear Regression functions from sklearn package (<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>) Please submit

1. Your source code written in Python.
2. All external files with the data used in your python script. Please make sure that the paths in your script are set correctly. If the TA will not be able to run your code, you will be penalized.
3. A text file that contains (a) the list the countries that you used to train the model; (b) the results of 5-fold cross-validation.

If you cannot submit the files to iCollege directly, please, archive them, and then add an extension .txt to your archive. For example, you have archived your files of the source code into SuperStudent.zip. Then, you have to rename it as SuperStudent.zip.txt and upload to iCollege.