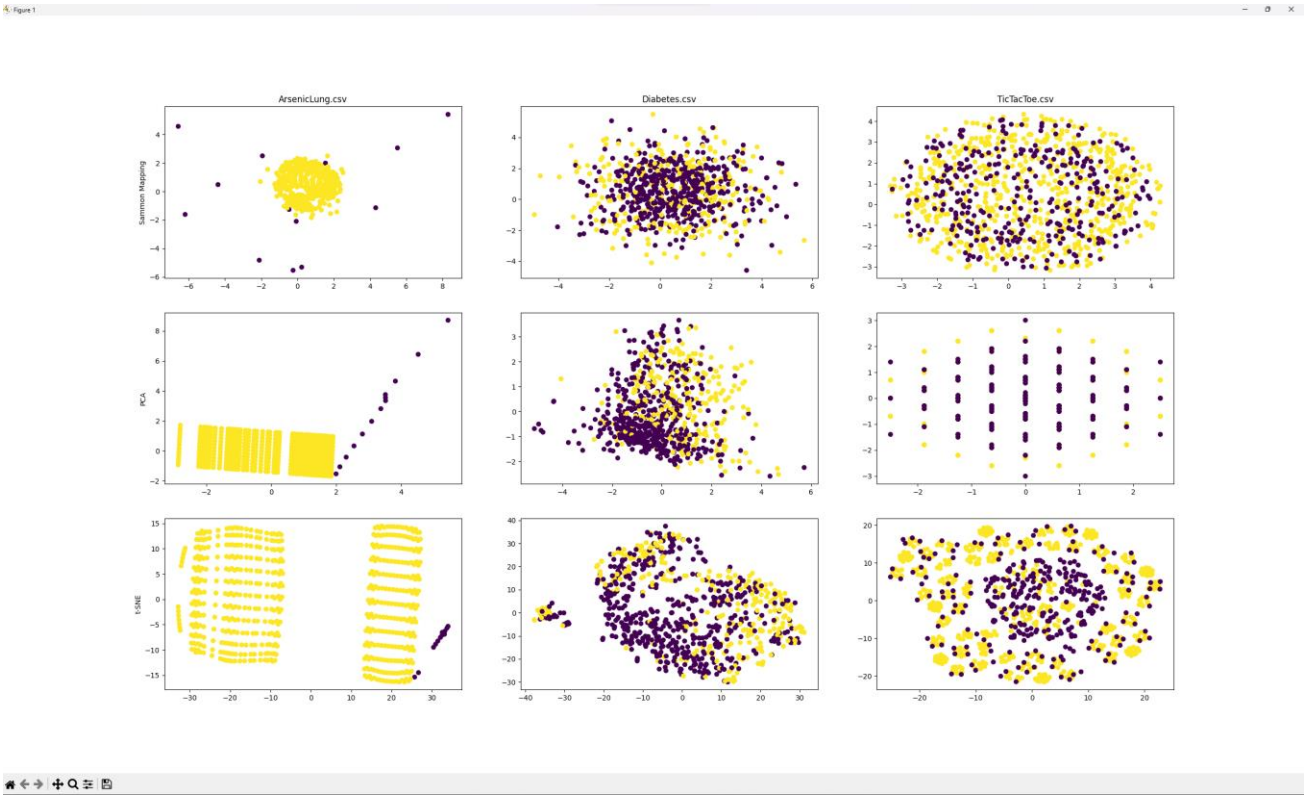
Assignment 4

By Nedko Nedkov

Task 1 & 2:

I have implemented the methods in the Common.py file to be able to use them in all 3 tasks. Task1.py and Task2.py just make calls to those methods.

Task 3.1:



To assess the performance of each technique, we can look at the boundaries and overlap of the target classes, as well as the distance between points.

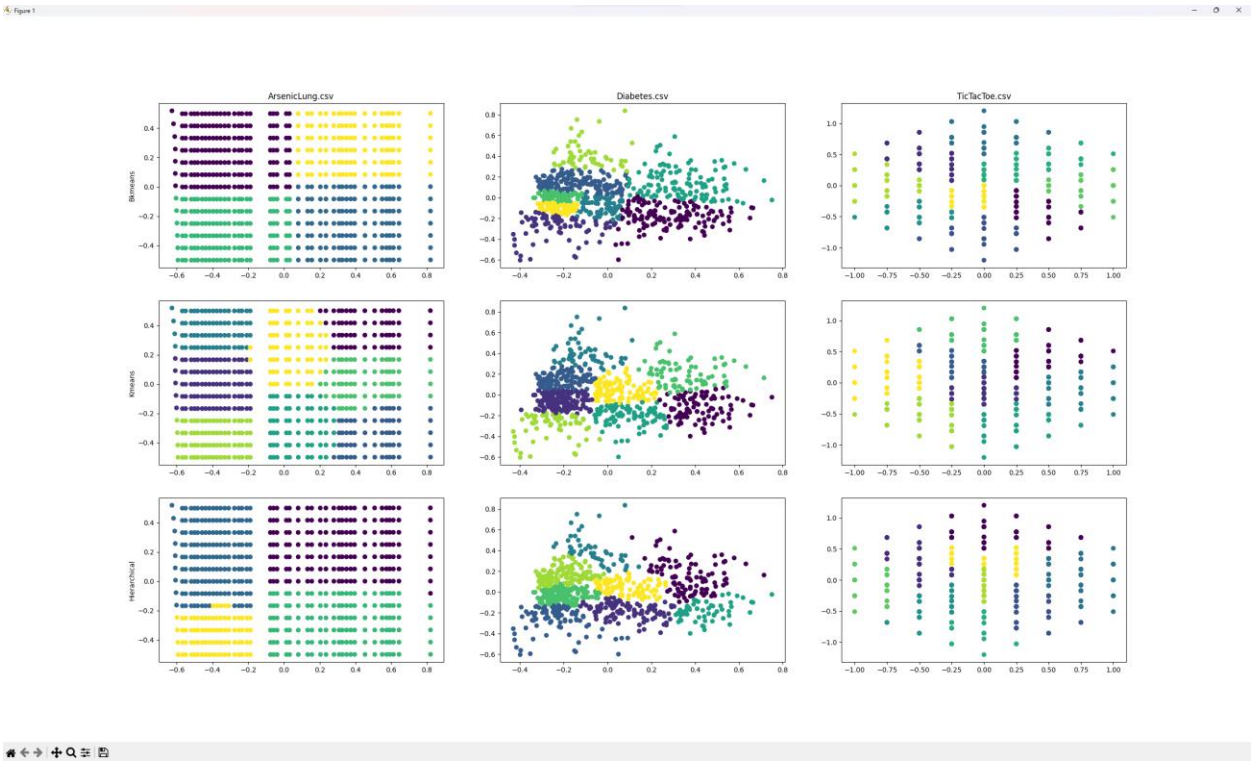
For the first dataset, ArsenicLung.csv, we can see that the Sammon Mapping results in a bit of overlap of the two classes, hence is not our preferred technique. t-SNE also has an issue – great distance between the points of the yellow class. That leaves PCA, which has a greater distance between the points of the purple class but compensates for it with the compactness of the yellow class, so it is the technique that would be the most fitting to use in this case.

For the second dataset, Diabetes.csv, the points are more chaotically spread out, so with each technique giving great class overlap. In this case, we would look at the compactness of the class points and choose the PCA again, which gives the most compact classes.

Lastly, for the TicTacToe.csv, we will go with PCA again, which gives the minimum overlap and clearest boundaries, even if the distance between the points is greater than the other two techniques.

Some classes are easier to separate than others if they have clear boundaries and no overlap with other classes. If there is no clear boundary, it can be more difficult to separate the classes but still possible if the correct technique is chosen, such as t-SNE.

Task 3.2:



To assess the performance of each clustering technique, we consider the separation (distance between clusters, with higher distance being preferred) and compactness (distance between each point in the cluster) of different clusters.
For the first dataset, ArsenicLung.csv, we observe that Bkmeans creates the simplest clusters with no anomalies (for example the yellow-colored cluster in the other two techniques), with the least distancing between points. Therefore, we can conclude that Bkmeans is the most fitting technique to use for the first dataset.
For the second dataset, Diabetes.csv, it is a bit more difficult to determine which is the preferred technique to use, since it is scattered more randomly. There, we can note that the bkmeans and kmeans techniques result in clusters that are overlapping with each other, hence are not good techniques to use. Rather, we should pick the hierarchical technique to use, since it gives us well defined clusters with points close to each other.
Lastly, the TicTacToe.csv dataset is similar to our first set in the sense that it is more organized. There, the kmeans seems to be giving the most compact clusters with highest distance between them, so we should choose that technique.

Dataset clusters are separated depending on their characteristics and dimensionality. Some of them have clear boundaries with little-to-no overlap and are easy to separate, while others are

the opposite. An example of clusters that are hard to separate are high-dimensional dataset clusters, since they seem to be closer to each other.