Task - 4

K-MEANS

The *K-Means* is a technique for partitioning N observations into K disjoint clusters C_1, C_2, ..., C_k in which each observation belongs to the cluster with the nearest mean known as *centroid*. The algorithm tries to minimize the within cluster sum of a certain distance function of each point in the cluster to the cluster centroid:

$$min\ j = \sum_{j=1}^{K} \sum_{n \in C_j} \mid x_n - \mu_j \mid^2$$

where x_n is a vector representing the nth observation and μ_j is the *centroid* of observations in C_j. The K-Means algorithm, that implement the technique, is simple and intuitive. It works as follows (there is plenty information about this technique on the internet):

1. Decide how many clusters you want. Call this K.
2. Create K random "centroids". Each cluster centroid will have the same number of dimensions as the input observations. You can choose random values for each dimension for each of the K centroids or you can choose a random data point to represent each initial cluster centroid.
3. For each observation use Euclidean distance (which works in any number of dimensions) to determine which cluster's centroid is closest to the observation. Assign this observation to that cluster. (Note it may have already been assigned to that cluster.)
4. Now that all observation have been assigned (or reassigned) to clusters, recalculate the cluster means. This simply involves summing all data vectors in the cluster and dividing by the number of members in the cluster.
5. Go back to step 3 until no cluster assignments change.

Write a templated version of the K-Means algorithm using STL Iterators and containers. Use the algorithm to cluster the following datesets:

1. the well know Iris dataset (`https://archive.ics.uci.edu/ml/datasets/Iris` for more documentation about it): this dataset contains contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Use the K-Means templated implementation to try to classify the observations into the 4 types of plants, considering the *sepal length*, *sepal width*, *petal length* and *petal width* attributes.
2. a set of points located in São Paulo state: the points have an *ID*, its *latitude*, *longitude* and the *municipality* where they are located. Use the K-Means templated implementation to cluster the points using its latitude and longitude attributes.

Basically you have to:

1. read the dataset files using the facilities of `<iostream>` and store them in memory using STL containers, classe, structs, etc.
2. run you K-Means implementation to classify the two datasets:
   - the iris dataset into 4 clusters. How close you to the right classification (attribute *type*)?
   - the points try more classify in different number of clusters (10 to 15 for example).