MACHINE INTELLIGENCE AND EXPERT SYSTEMS
AUTUMN SEMESTER - 2021
PROGRAMMING ASSIGNMENT-2 (CLUSTERING)

Problem Statement:

Write a program to cluster the given set of points using K-means. Consider, K=3, clusters. Consider Euclidean distance as the distance measure. Randomly initialize a cluster mean as one of the data points. Iterate for 10 iterations.

a) After iterations are over, print the final cluster means for each of the clusters. Use the ground truth cluster label present in the data set to compute and print the Jacquard distance of the obtained clusters with the ground truth clusters for each of the three clusters.
b) Plot the final clusters using a scatter plot and any 2 features at a time.
c) **Bonus:** Modify the code to initialize the cluster centres using the concept of KMeans++, i.e choose the first centre randomly then the second centre will be the point farthest away from the 1st centre and similarly initialize the 3rd centre keeping into account its distance from 1st and the 2nd centre.

Note:

1) Don’t use any inbuilt library for performing K-Means clustering. Ex scikit-learn. However, you are allowed to use Numpy and pandas for data handling.
2) Avoid Hardcoding wherever possible, use global variables for changeable parameters.
3) The code should be well documented/commented along with an explanation of each function used.
4) Code will be evaluated on the basis of readability, result, and approach. Any kind of plagiarism will be highly penalized.

Data Set Description:

Data Filename: iris_plant.csv

The data set contains 150 data points, there are three clusters where each cluster refers to a type of iris plant. The first four columns represent the attributes listed below. The last column is the ground truth cluster name and is to be used for evaluating the cluster quality.

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. Ground truth cluster name:
    - -- Iris Setosa
    - -- Iris Versicolour
    - -- Iris Virginica

Submission Instruction:

Use Python for this assignment. Submit the final code along with a ReadMe file specifying each instruction to run the code. Submit a single zip file containing your code, ReadMe, and Results, and name the zip file as <your_name>_<rollnumber>.zip.