Assignment #02 – Machine Learning

Instructor	**Asif Ameer**
Session	**Fall 2022**

Instructions:

- Questions requiring paperwork must be done on A4 size blank pages and then the images may be scanned through any application like CamScanner. Upload a **PDF** file containing all your scanned images in the proper format.
- The paperwork must contain your name and roll number on every single page. **Any page without the name and roll number will not be considered.**
- PDF file format should be "Roll-Number_Section_AssignmentNo", for example *19F-0111_A_Assignment #04*. **Marks will be deducted** for not following the correct format.
- Plagiarism will not be tolerated, either done from the internet or from any fellow classmate (of same/different section) and **will lead to zero or negative marks in the assignment.**
- **No late submissions will be accepted.**

Question 1

Load the dataset available in the folder with name “data”. The data contains the basic information (ID, age, gender, income, spending score) about the customers of a mall. Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

A Notebook is provided to you where the dataset has already been normalized.

1. Label encoding is done of the Gender Column to convert its values to numerical values.
2. Feature Extraction is already performed to extract the features required for the problem.

Your task is to build up the **K-Means Clustering** function from scratch and use that to train this normalized data.

Apply **Elbow Method** on your problem to find out the optimal number of clusters with the K-Means method.

Visualize your clusters with the help of scatter plot on your dataset by assigning colors based on the clusters computed through K-Means Algorithm. (Don’t use any two random values, instead use Principal Component Analysis for Reduction of Columns)

Question 2

Consider the same normalized dataset used in the above problem and apply the following two techniques on the dataset,

1. K-Medoid Algorithm (for k = 2, find the cluster with least total cost)

2. Hierarchical Clustering (Technique = Agglomerative Clustering, Linkage = Single-Link)

Note: Using Built-in libraries (scikit-learn, etc.) is not allowed for algorithm implementation and will result in zero marks. You may use Numpy, Pandas, Seaborn or Matplotlib etc. however.

Both the Questions are coding questions. You are required to follow the instructions as mentioned above.