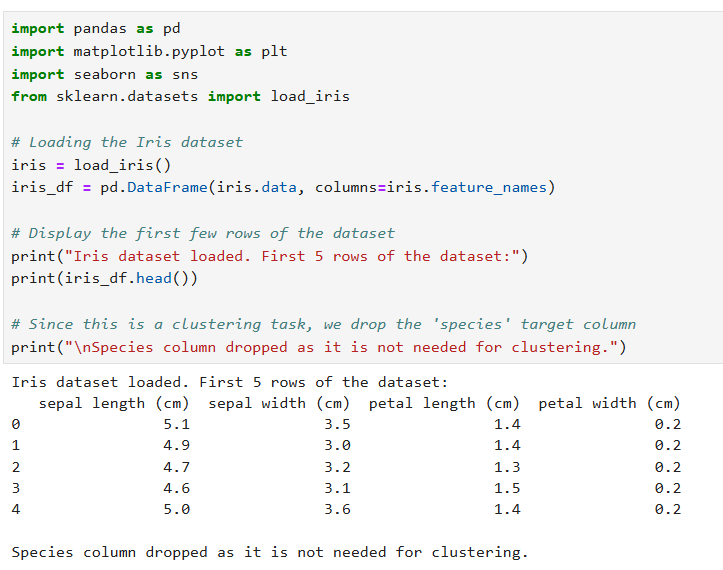
**Formative Assessment: Unsupervised Learning**

**Due on 11th October 24**

**Submitted by Aiswarya Jayaprakash**

**Objective:**  
The objective of this assessment is to evaluate your understanding and ability to apply clustering techniques to a real-world dataset.  
  
**Dataset**  
Use the Iris dataset available in the sklearn library.  
  
**Key components to be fulfilled :**  
  
**1. Loading and Preprocessing (1 marks)**

* Load the Iris dataset from sklearn.
* Drop the species column since this is a clustering problem.



**Iris Dataset c**ontains of a total of 150 samples of iris flowers with 4 features: sepal length, sepal width, petal length, and petal width. In the p**reprocessing step t**he species column is dropped because clustering is unsupervised, and the data is grouped without using labels.

**2.Clustering Algorithm Implementation (8 marks)**

* Implement the following two clustering algorithms:

**A) KMeans Clustering (4 marks)**

* Provide a brief description of how KMeans clustering works.

KMeans clustering is an iterative clustering algorithm that assigns each data point to one of k clusters. The centroids of the clusters are recalculated after each iteration to minimize the within-cluster sum of squares (WCSS). It is a distance-based algorithm and uses Euclidean distance to assign points to clusters.

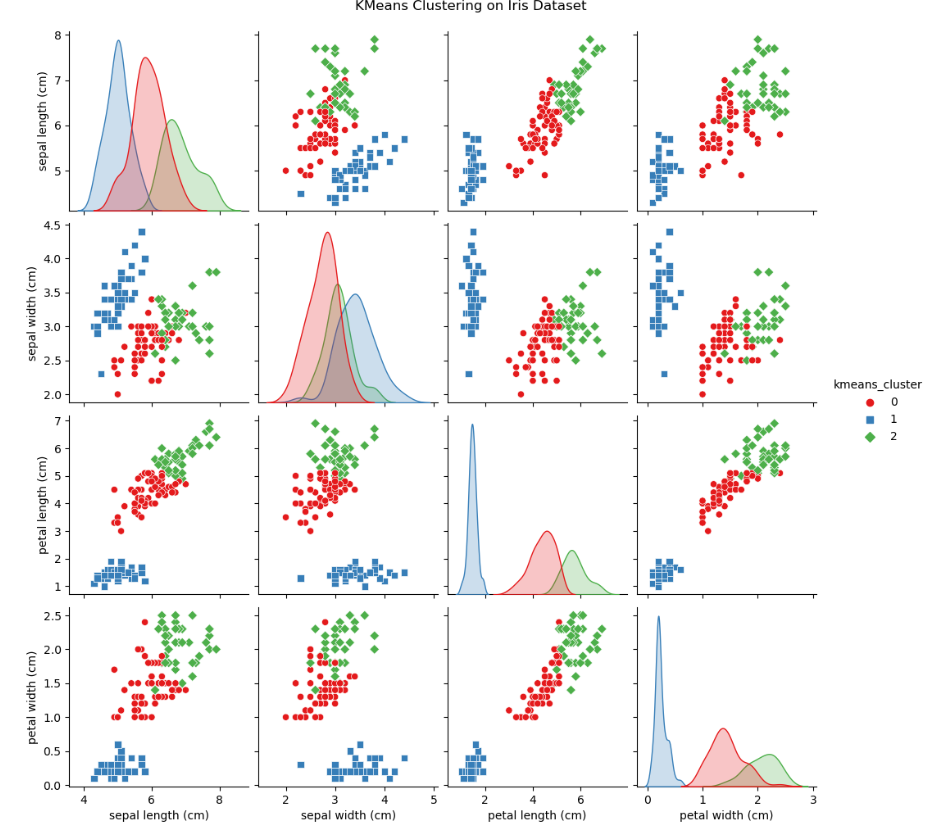
The clustering works by randomly initializing k centroids. It assigns each point to the nearest centroid and recalculates the centroid of each cluster. This process is repeated until the centroids do not change.

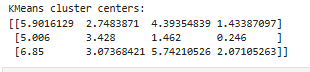
* Explain why KMeans clustering might be suitable for the Iris dataset.

The Iris dataset has clearly defined classes and features that are easy to separate in terms of distance. KMeans performs well when clusters are globular and linearly separable, which is likely the case for the Iris dataset.

* Apply KMeans clustering to the preprocessed Iris dataset and visualize the clusters.







**B) Hierarchical Clustering (4 marks)**

* Provide a brief description of how Hierarchical clustering works.

Hierarchical clustering builds a hierarchy of clusters. The main types are **agglomerative clustering and divisive clustering.** Agglomerative clustering is more commonly used. It works by calculating the distance between clusters using a linkage criterion (e.g., single-linkage, complete-linkage). Agglomerative clustering starts with each point as its own cluster and merges clusters iteratively whereas **divisive clustering starts with all points in one cluster and splits them iteratively.**

* Explain why Hierarchical clustering might be suitable for the Iris dataset.

Hierarchical clustering is particularly useful when the dataset has a natural hierarchical structure, which is often the case with biological datasets. It does not require specifying the number of clusters in advance and can be visualized using a dendrogram

* Apply Hierarchical clustering to the preprocessed Iris dataset and visualize the clusters.

