In-Class Activity: K-Means and Hierarchical Clustering

# Objective

In this activity, you will explore two unsupervised learning techniques—K-means clustering and hierarchical clustering—using real-world data. You'll also use principal components (PCs) to visualize clusters in a simplified 2D space.

**🌸 Introduction to the Iris Dataset**

The **Iris dataset** is one of the most famous datasets in data science and statistics. It was introduced by the British biologist and statistician **Ronald Fisher** in 1936 as an example for **discriminant analysis**.

**📦 What’s in the dataset?**

The dataset contains measurements of **150 flowers** from three different species of the *Iris* flower:

* *Iris setosa*
* *Iris versicolor*
* *Iris virginica*

Four features describe each **flower**:

| **Feature** | **Description** |
| --- | --- |
| Sepal Length (cm) | Length of the outer part of the flower |
| Sepal Width (cm) | Width of the outer part |
| Petal Length (cm) | Length of the inner petal |
| Petal Width (cm) | Width of the inner petal |

There are **50 samples from each species**, for a total of **150 rows** in the dataset.

**🎯 Why is this dataset useful?**

* It’s a **clean and small** dataset—great for learning.
* It allows us to practice both **supervised** and **unsupervised learning**.
* The differences between the species are subtle but present, making it ideal for:
  + **Clustering** (unsupervised)
  + **Classification** (supervised)
  + **Dimensionality reduction** (e.g., PCA)

## Part 1: Visualize the DataA graph showing different colored dots AI-generated content may be incorrect.

## Part 2: K-Means Clustering

A graph with different colored dots

AI-generated content may be incorrect.

## Part 3: Hierarchical Clustering

A diagram of a clustering structure

AI-generated content may be incorrect.A graph with green and red dots

AI-generated content may be incorrect.

## Reflection Questions

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hierarchical | |  |
| True | 1 | 2 | 3 |
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 50 | 0 |
| virginica | 0 | 47 | 3 |
|  | KMeans |  |  |
| True | 1 | 2 | 3 |
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 39 | 11 |
| virginica | 0 | 14 | 36 |

## PCA Analysis

A graph with dots and numbers

AI-generated content may be incorrect.

A diagram of a graph

AI-generated content may be incorrect.

K-means using the first 3 PCA Scores (equal result versus before)

|  |  |  |  |
| --- | --- | --- | --- |
|  | K-Means |  |  |
| TRUE | 1 | 2 | 3 |
| Setosa | 50 | 0 | 0 |
| Versicolor | 0 | 39 | 11 |
| Virginica | 0 | 14 | 36 |