**DATA WAREHOUSING AND DATA MINING LAB**

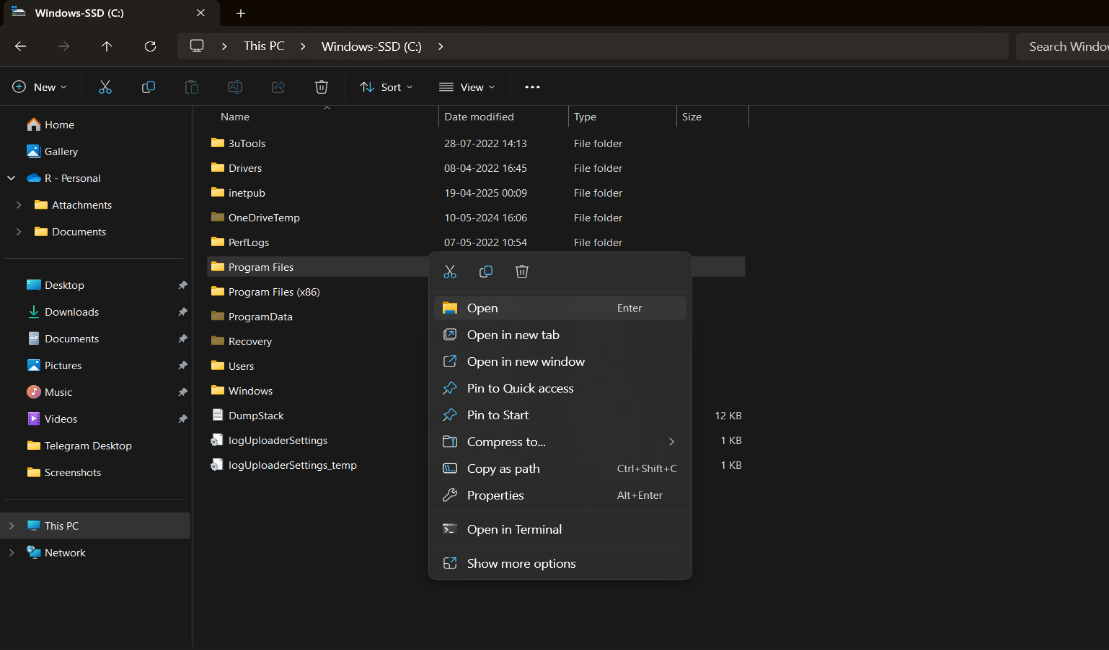
**Experiment 10: Use Hierarchical clustering and dendrogram visualization**

**Aim:**  
 To apply hierarchical clustering on a dataset, visualize the resulting dendrogram, and analyze the cluster formation at different levels.

**Description:**

Hierarchical clustering creates a tree-like structure (dendrogram) that shows the order in which data points are merged or split. It can be **agglomerative** (start with individual points and merge) or **divisive** (start with one cluster and split). The linkage method determines how distances between clusters are calculated.

**Steps to Run in WEKA :  
Step 1:**  
Open **File Explorer**  
Select **This PC** 🡪 Open the **C drive** 🡪 Open **Program Files**

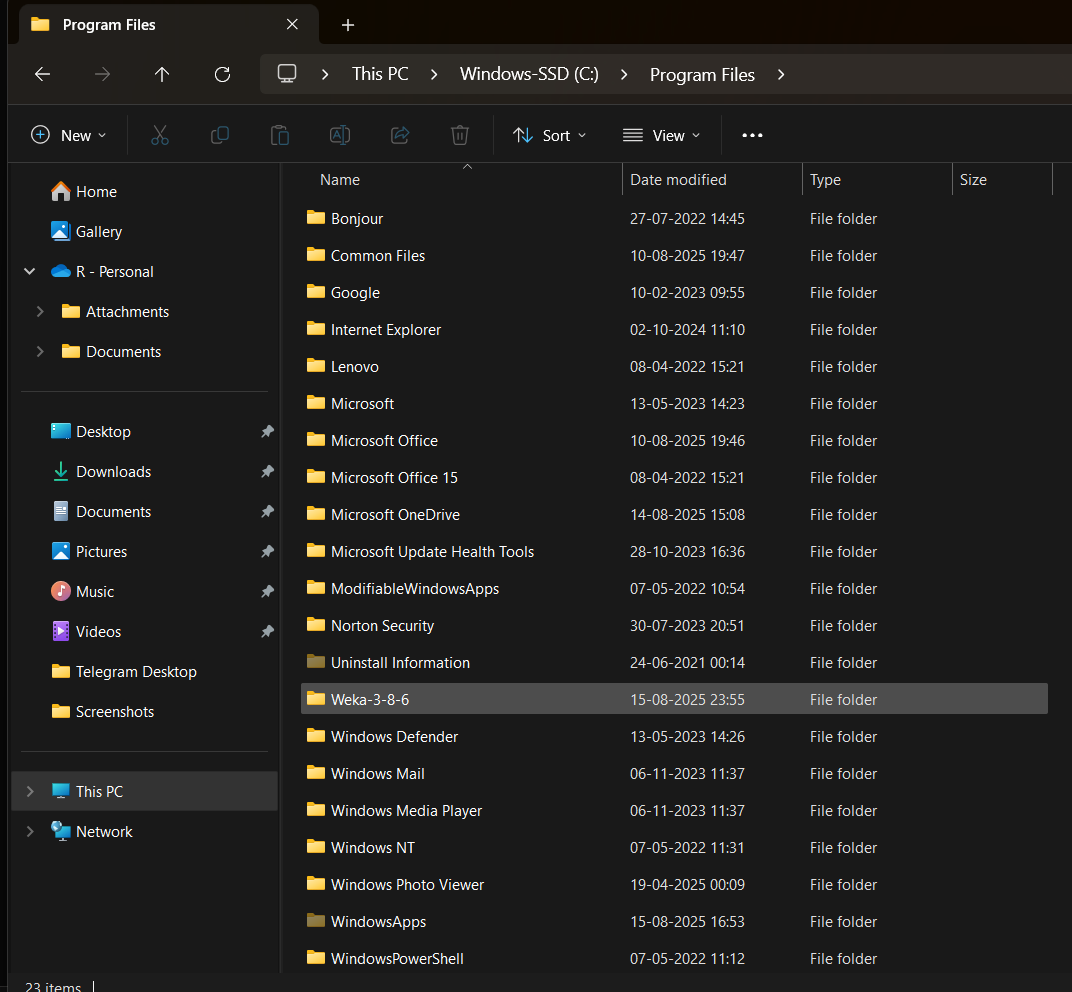


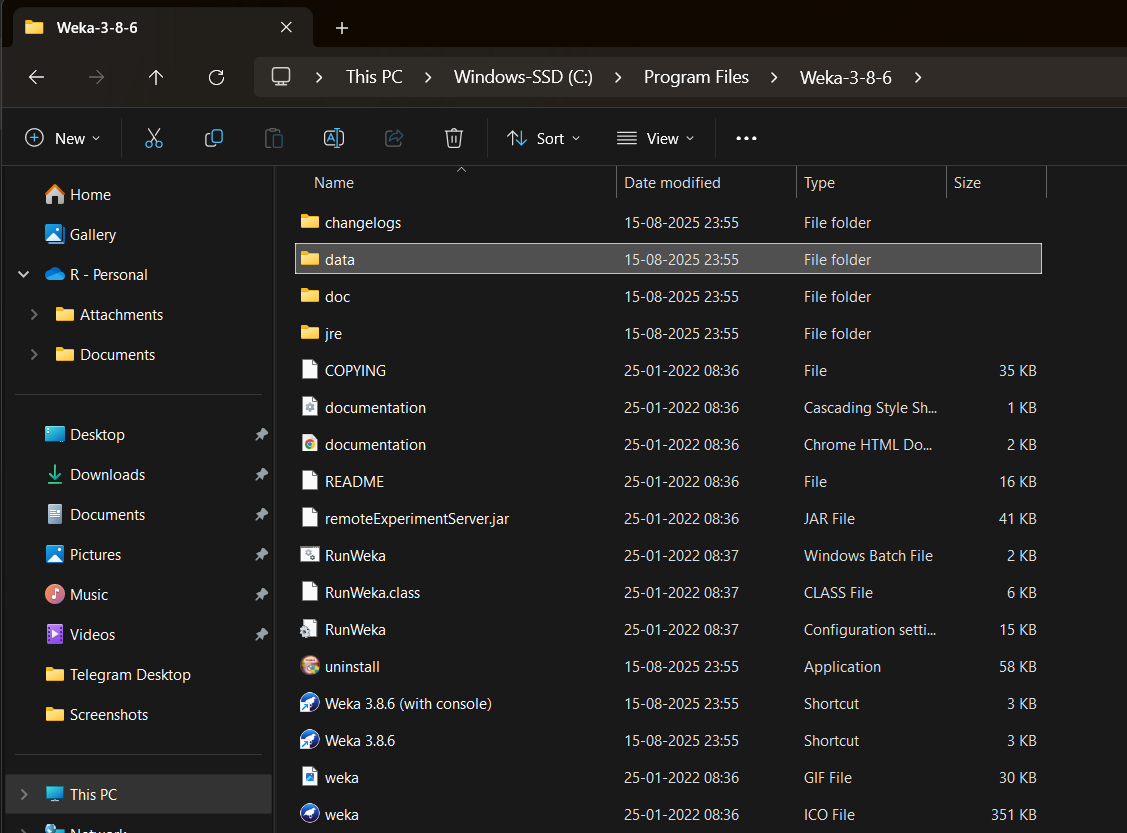
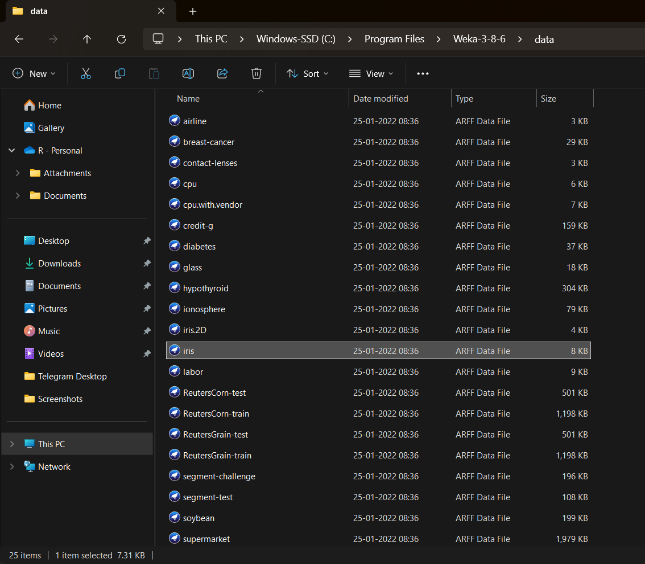
**Step 2 :**

Open the **WEKA 3.8.6** folder 🡪 Open the **Data** folder 🡪 Select the **Iris** dataset 🡪 Double-click on it

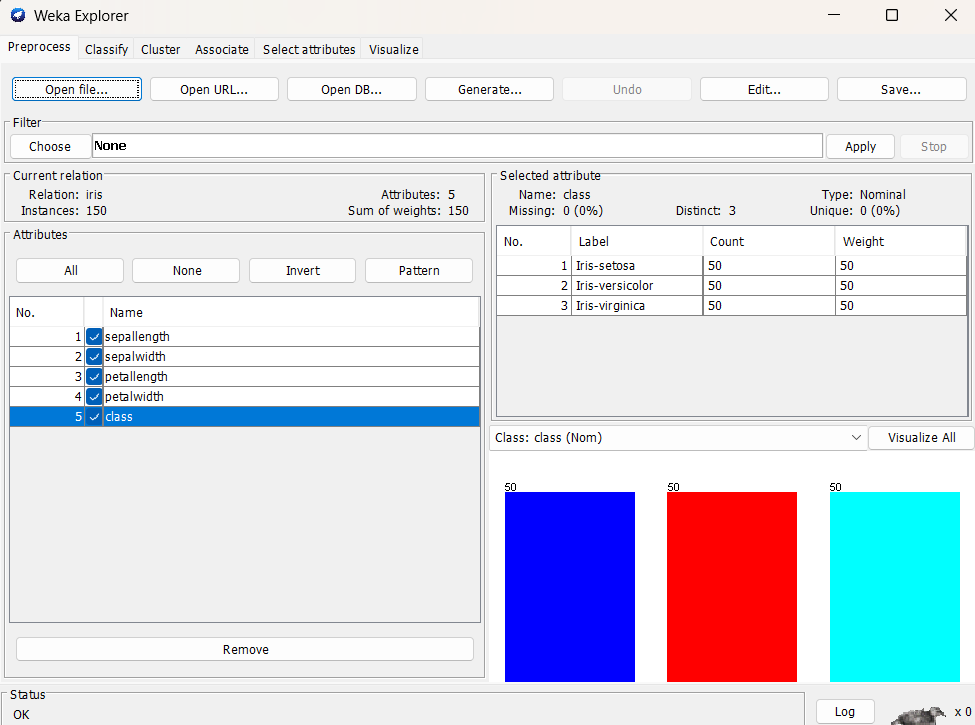
### **Iris Dataset – Information**

* **Introduced by**: Ronald A. Fisher in 1936 (in his paper on discriminant analysis).
* **Purpose**: Used for pattern recognition and classification tasks; one of the most famous benchmark datasets in machine learning.
* **Instances**: 150 flower samples.
* **Attributes**: 4 numerical features (in centimeters):
  1. Sepal length
  2. Sepal width
  3. Petal length
  4. Petal width
* **Class Labels (Species)**: 3 classes, each with **50 samples**:
  1. Iris-setosa
  2. Iris-versicolor
  3. Iris-virginica



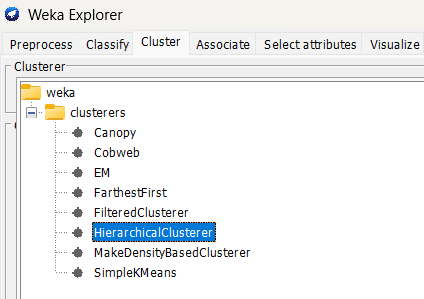
 

**Step 3:**  
 After WEKA opens, select all the parameters displayed there, such as **Sepallength, Sepalwidth, Petallength, Petalwidth,** and **Class**

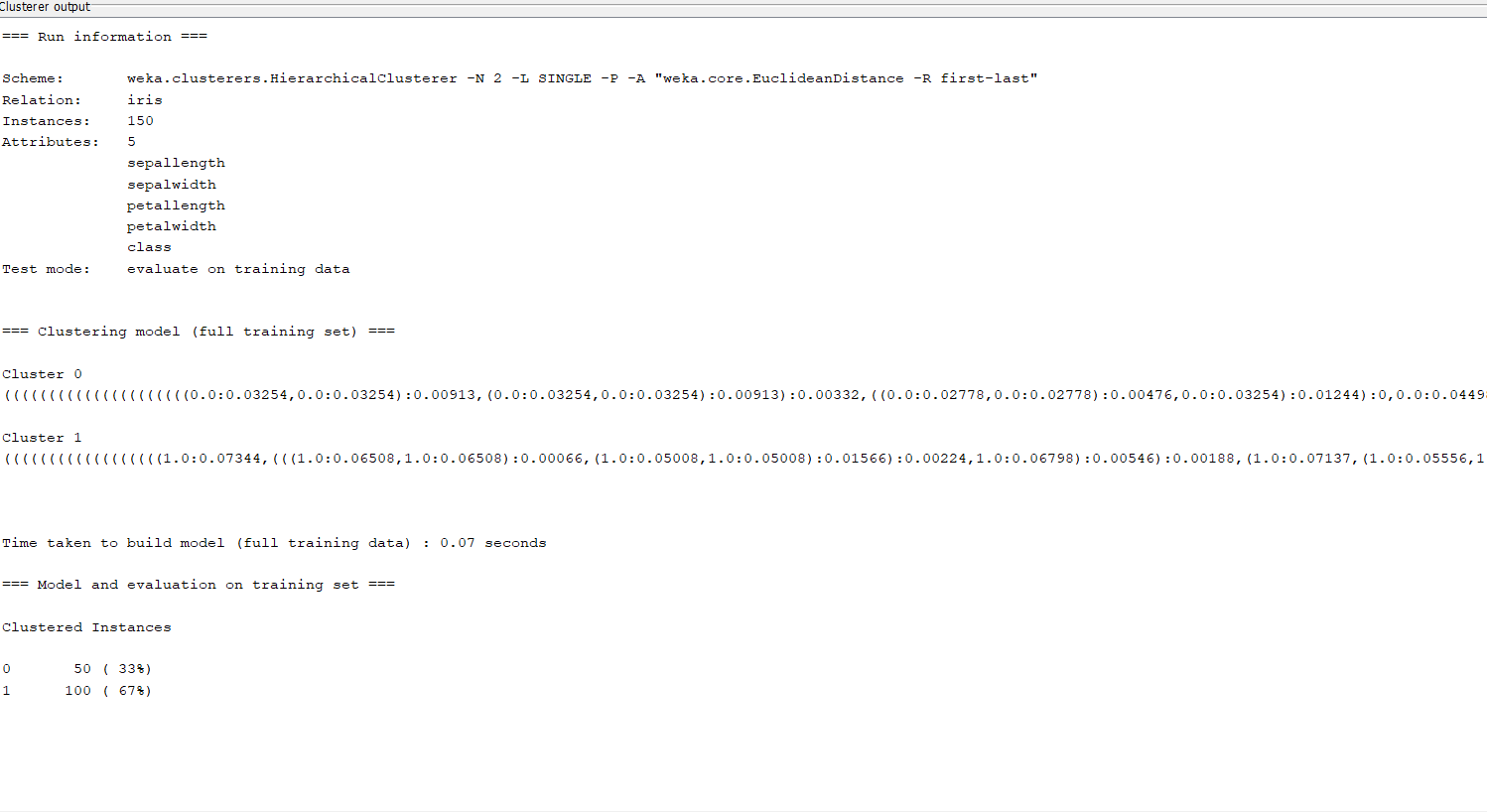


**Step 4:**

Select **Cluster** 🡪 Click the **Choose** option 🡪 In the **Clusterer** section, select **HierarchicalClusterer**



**Step 5:** Click on **Start**, and the output will be as follows:



### **Hierarchical Clustering – Summary**

* **Dataset**: Iris (150 instances, 5 attributes) (The data used for clustering)
* **Evaluation**: Trained and evaluated on the dataset itself (unsupervised learning, no predefined labels used)
* **Clusters formed**: 2 clusters (The algorithm grouped the data into 2 main clusters)
* **Build time**: 0.07 sec (Time taken to construct the clustering model)

**Clustered Instances:**

* **Cluster 0** → 50 instances (33%) → mostly Iris-setosa
* **Cluster 1** → 100 instances (67%) → mix of Iris-versicolor and Iris-virginica

**Result:**

* Setosa forms a clear, separate cluster.
* Versicolor and Virginica end up together in one cluster due to their similarity.

**Conclusion:** Hierarchical Clustering grouped the Iris dataset into 2 clusters: one for Setosa and one combining Versicolor and Virginica, similar to K-Means results.