**DATA WAREHOUSING AND DATA MINING LAB**

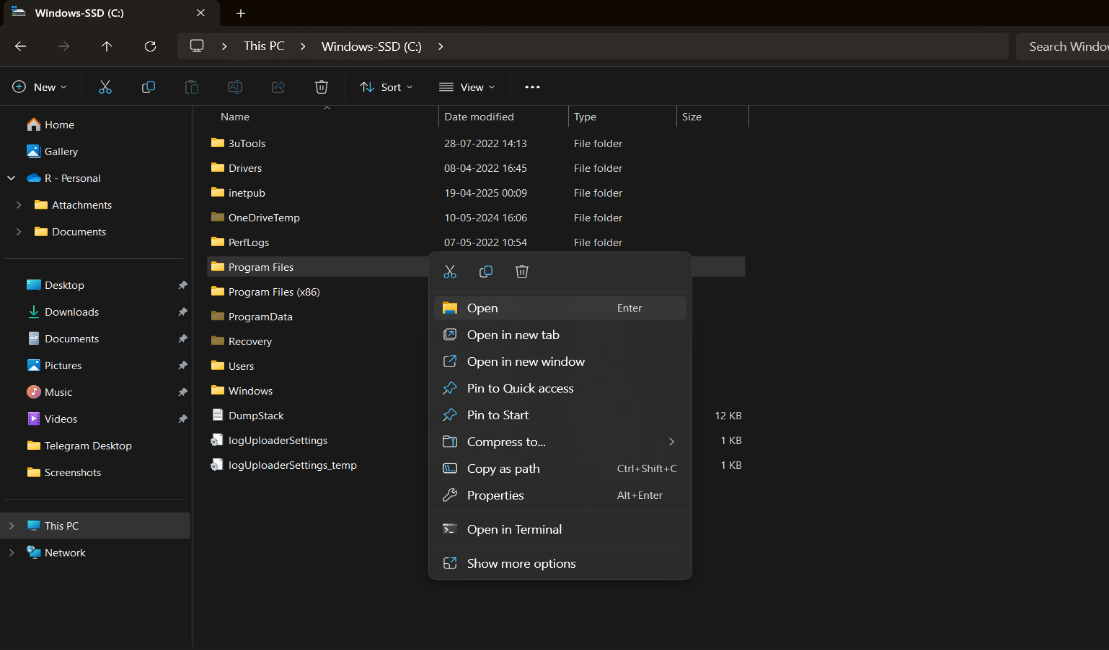
**Experiment 9: Apply K-Means clustering on a dataset and visualize cluster separation**

**Aim:**  
 To perform clustering on a dataset using the K-Means algorithm and visualize cluster separation to understand unsupervised grouping.

**Description:**

K-Means clustering partitions the dataset into k clusters where each point belongs to the cluster with the nearest mean (centroid). It iteratively updates cluster centroids until assignments no longer change. It is widely used in market segmentation, image compression, and document clustering.

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

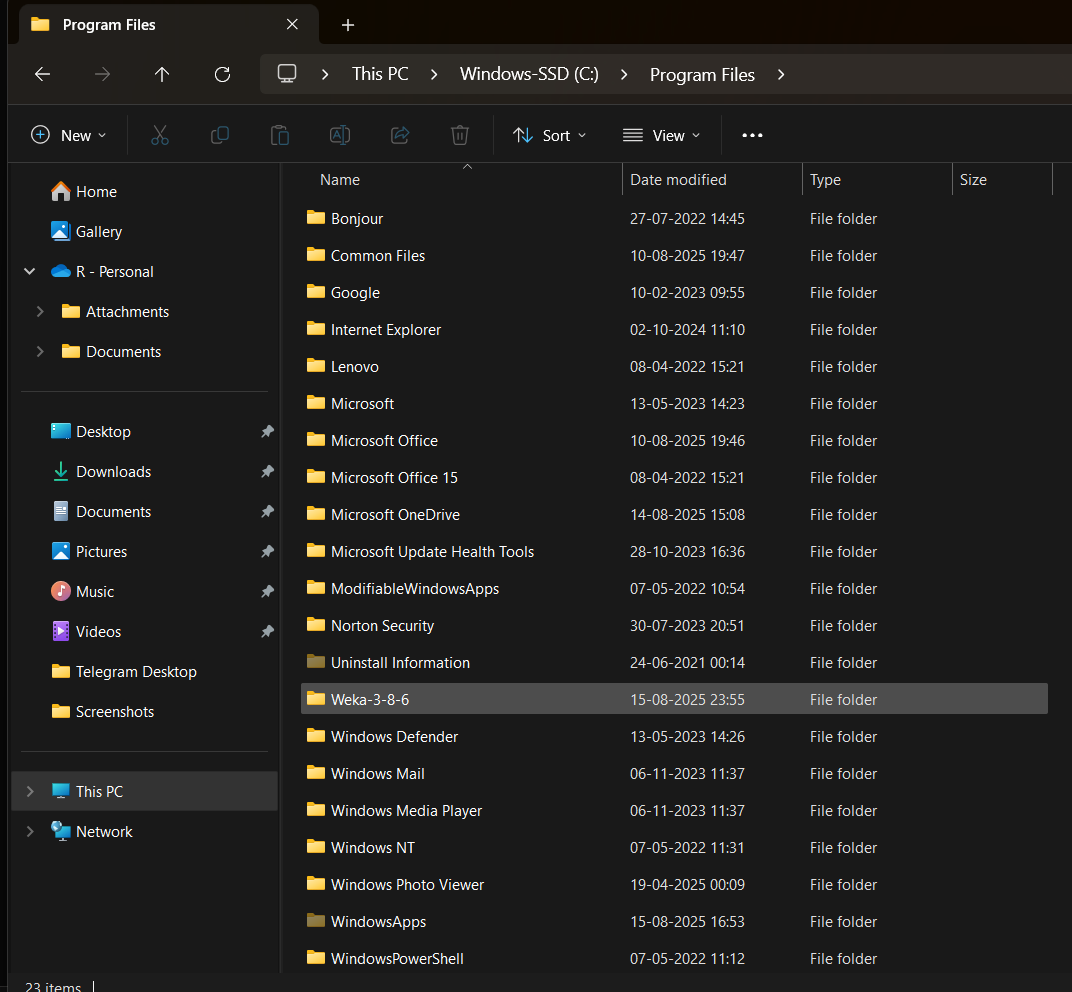


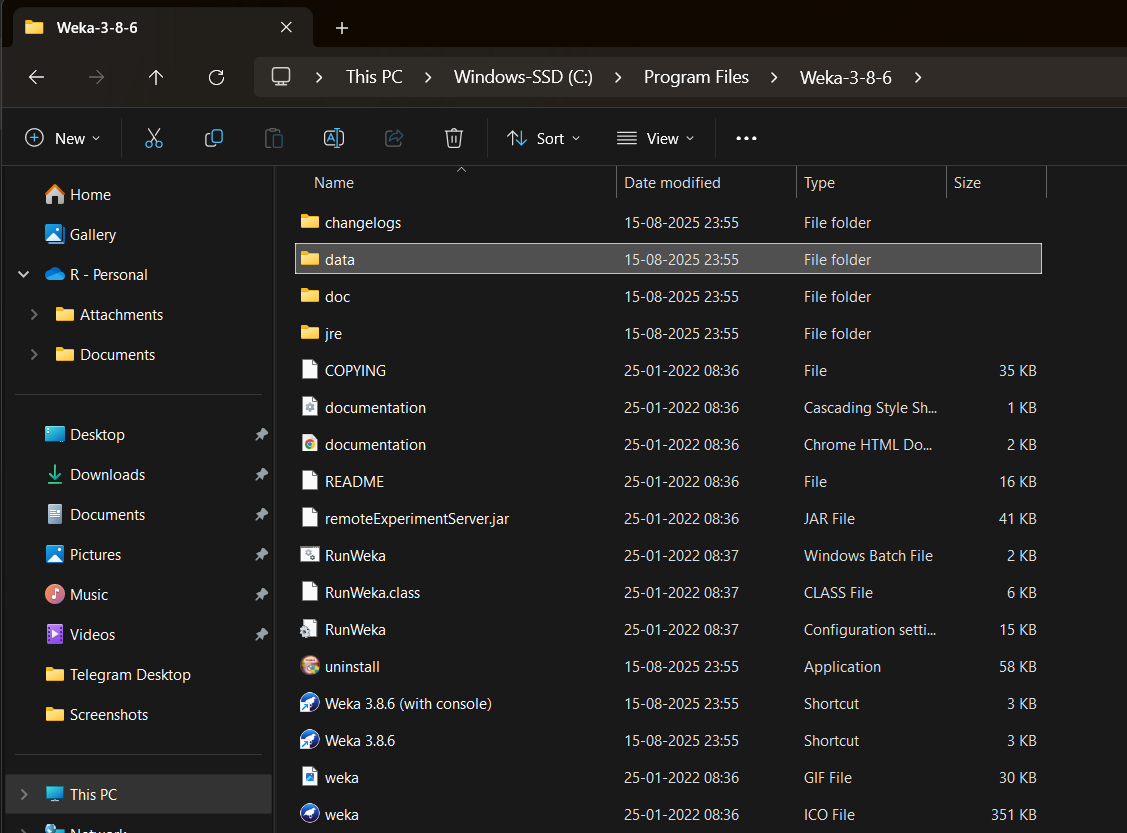
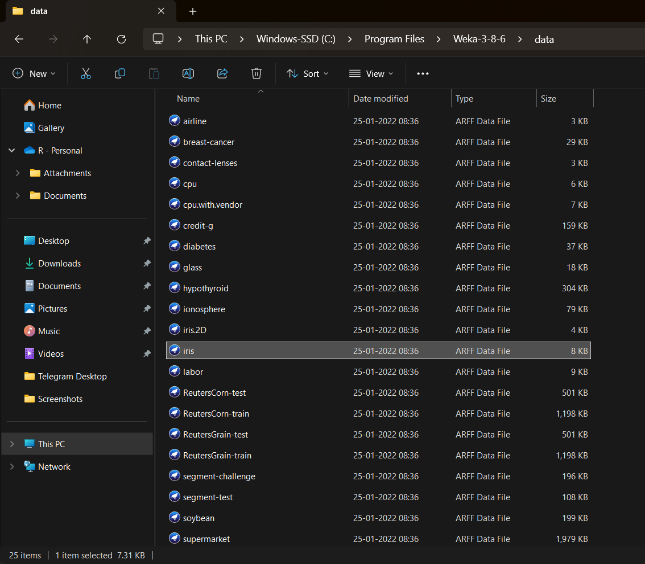
**Step 2 :**

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

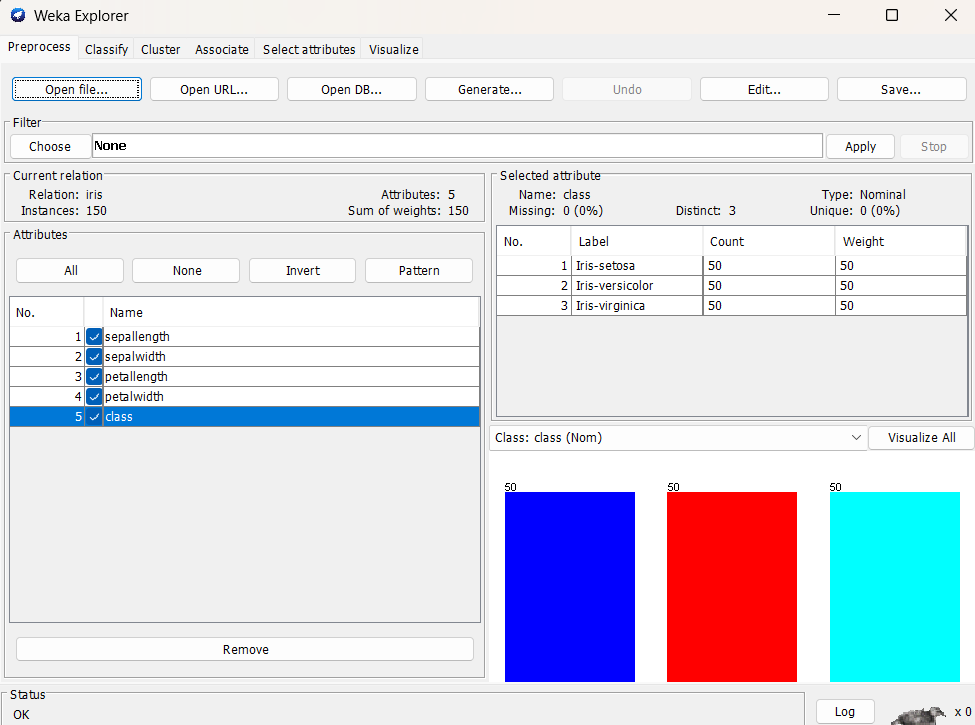
### **Iris Dataset – Information**

* **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



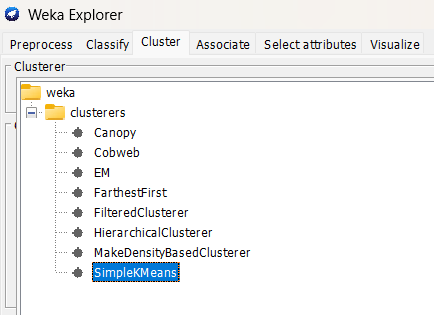
 

**Step 3:**  
 After WEKA opens, select all the parameters displayed there, such as **Sepallength, Sepalwidth, Petallength, Petalwidth,** and **Class**

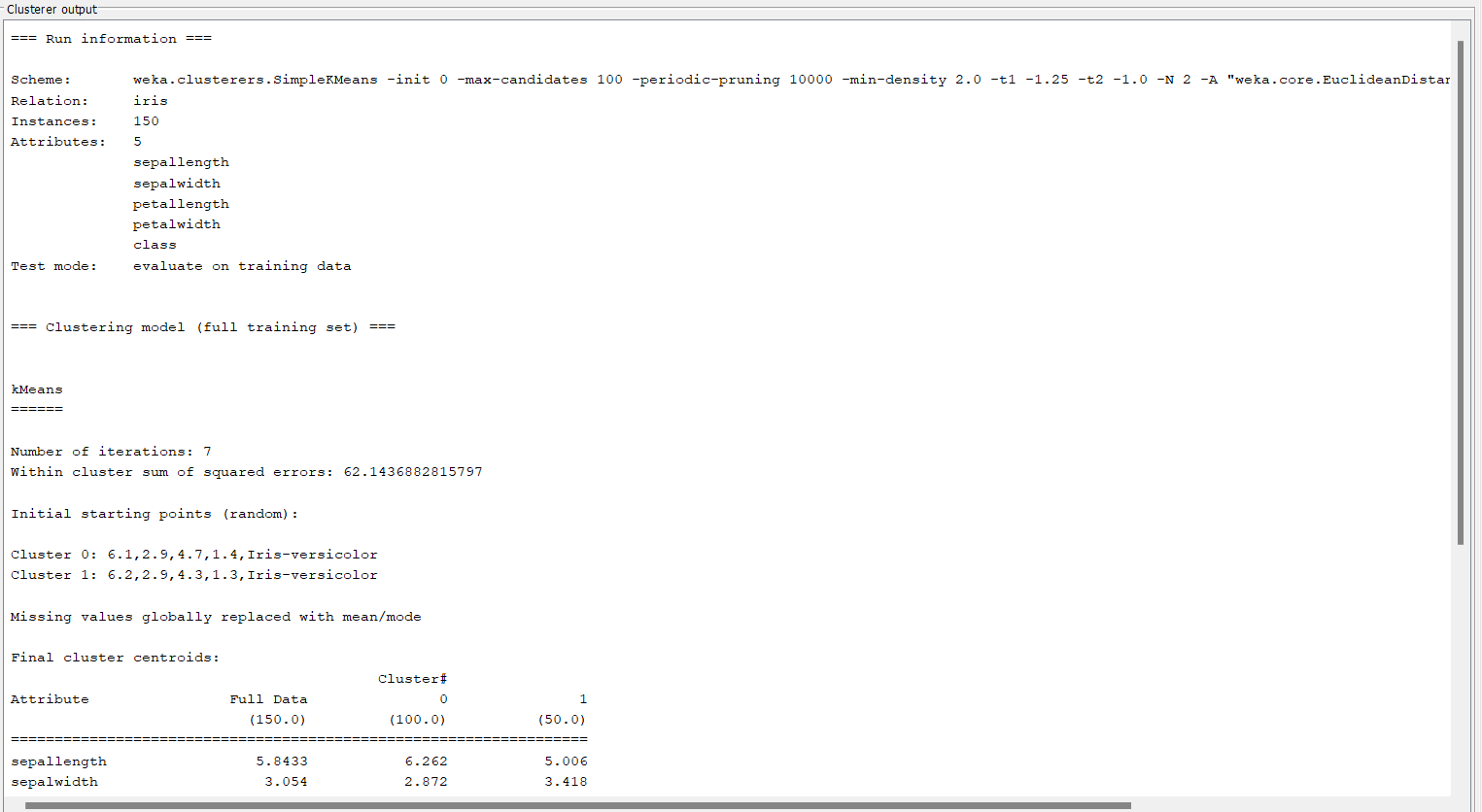


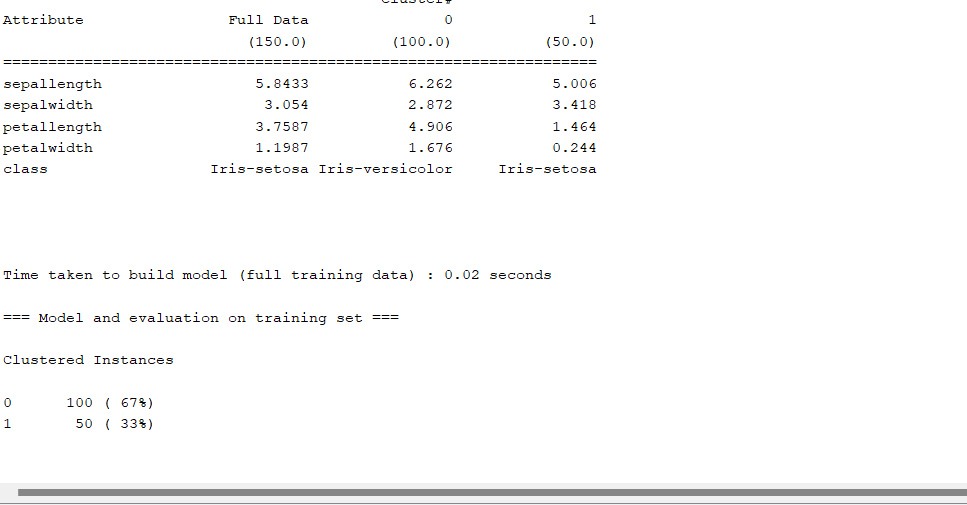
**Step 4:**

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



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





### **K-Means 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 for training)
* **Clusters formed**: 2 clusters (Instead of 3 Iris species, the algorithm grouped data into 2 clusters)
* **Iterations**: 7 (Number of times K-Means updated centroids until convergence)
* **Build time**: 0.02 sec (Time taken to form clusters)

**Cluster Centroids (Simplified):**

* **Cluster 0 (100 instances, 67%)** → Higher petal length & petal width (mostly versicolor + virginica)
* **Cluster 1 (50 instances, 33%)** → Smaller petal length & petal width (mostly setosa)

**Result:**

* Setosa is well-separated into its own cluster.
* Versicolor and Virginica are grouped together into one cluster (as they are more similar).

**Conclusion:** K-Means clustered the Iris dataset into 2 groups: one cluster perfectly represents Setosa, while the other mixes Versicolor and Virginica due to feature overlap.