**DATA WAREHOUSING AND DATA MINING LAB**

**Experiment 8: Perform classification using K-Nearest Neighbors (KNN) and analyze the results**

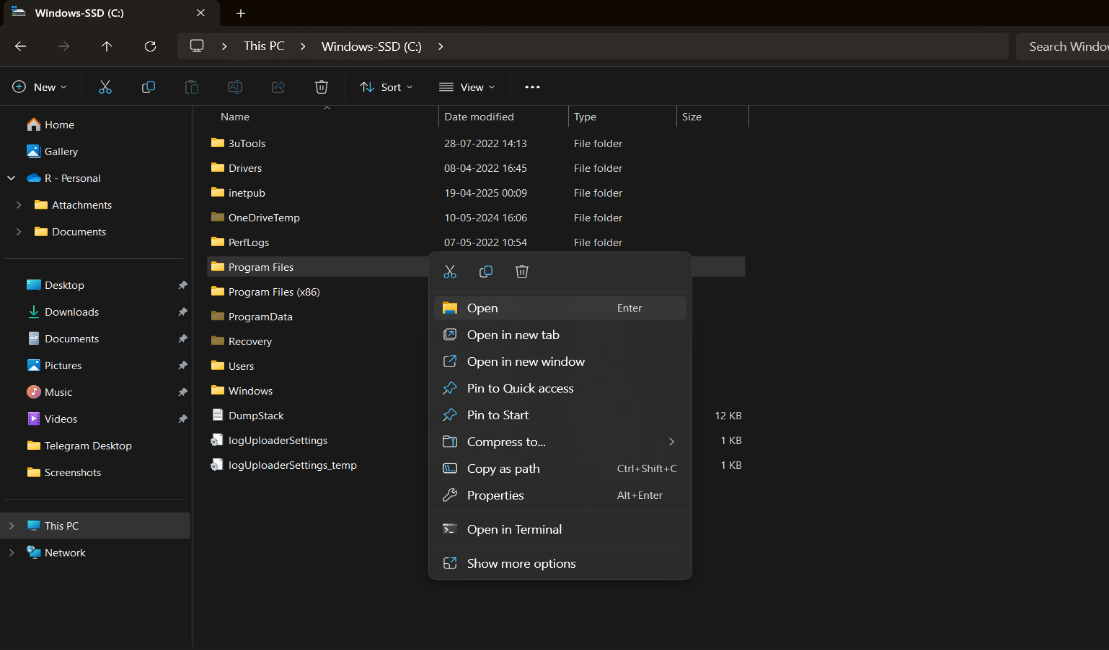
**Aim:**

To classify instances based on the majority vote of their k nearest neighbors using the KNN algorithm and compare performance for different k values.

**Description**

KNN is a non-parametric and instance-based learning algorithm. It does not build a model in advance; instead, it memorizes the dataset and classifies new instances based on similarity. The similarity is usually measured using Euclidean distance or other distance metrics.

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

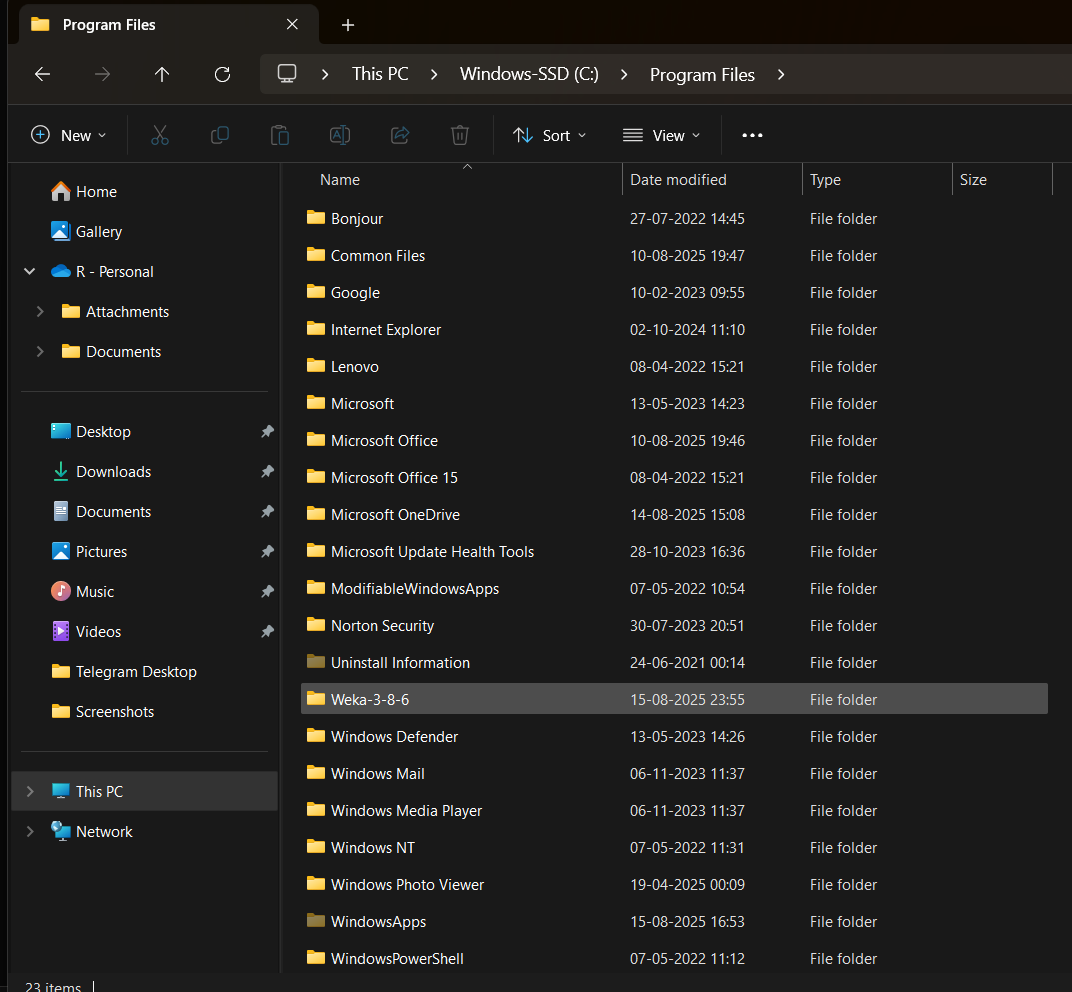


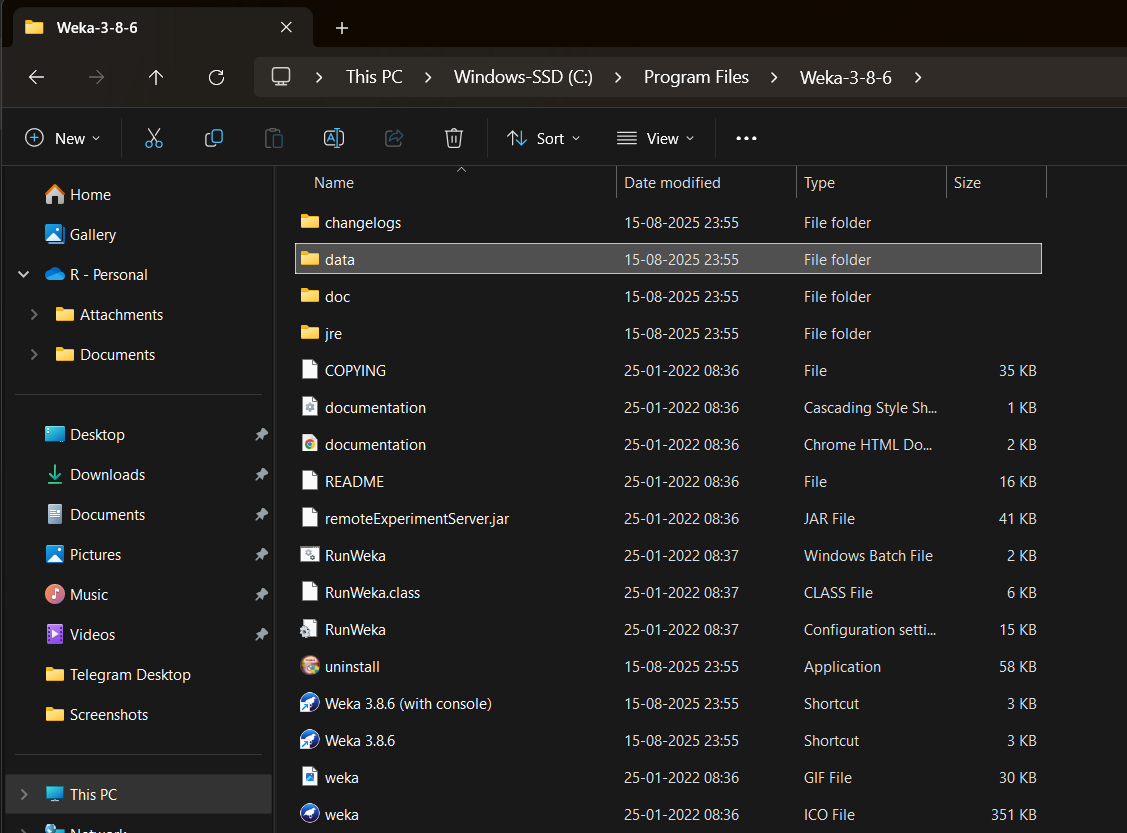
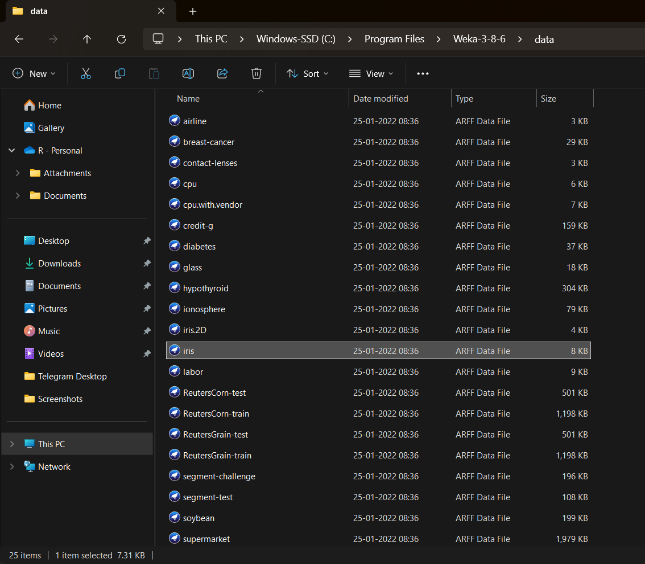
**Step 2 :**

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

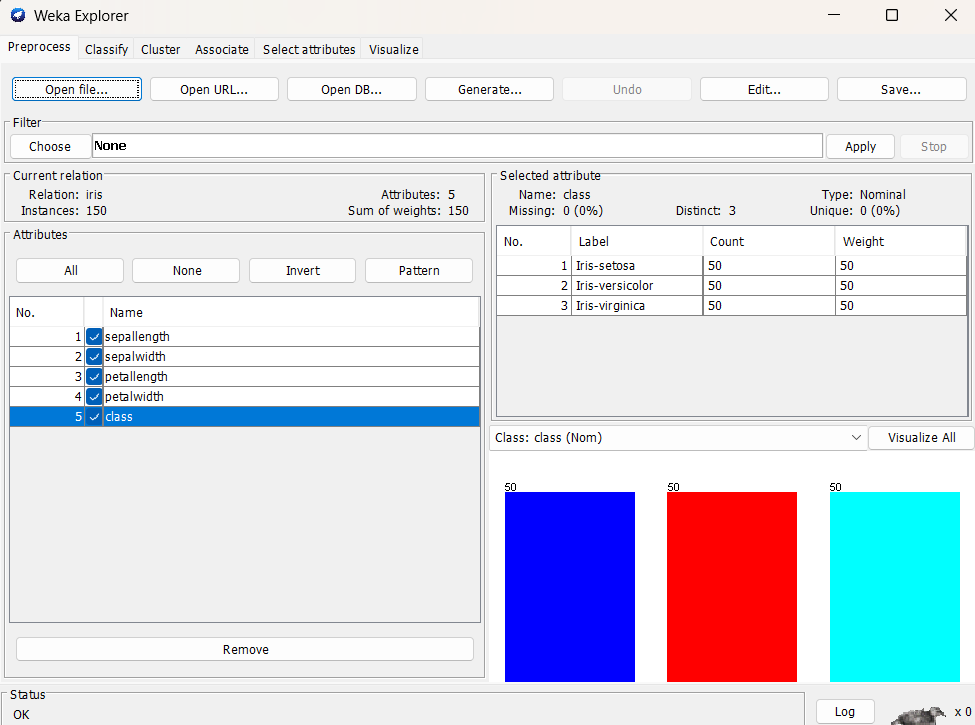
### **Iris Dataset – Information**

* **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



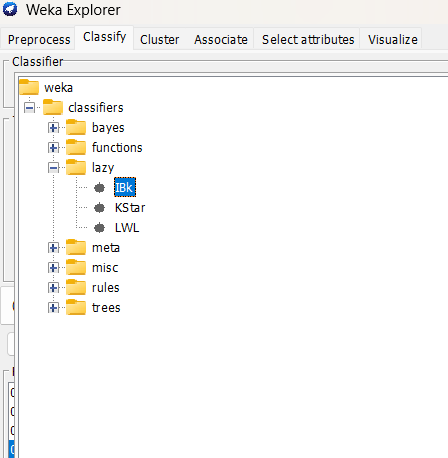
 

**Step 3:**  
 After WEKA opens, select all the parameters displayed there, such as **Sepallength, Sepalwidth, Petallength, Petalwidth,** and **Class**

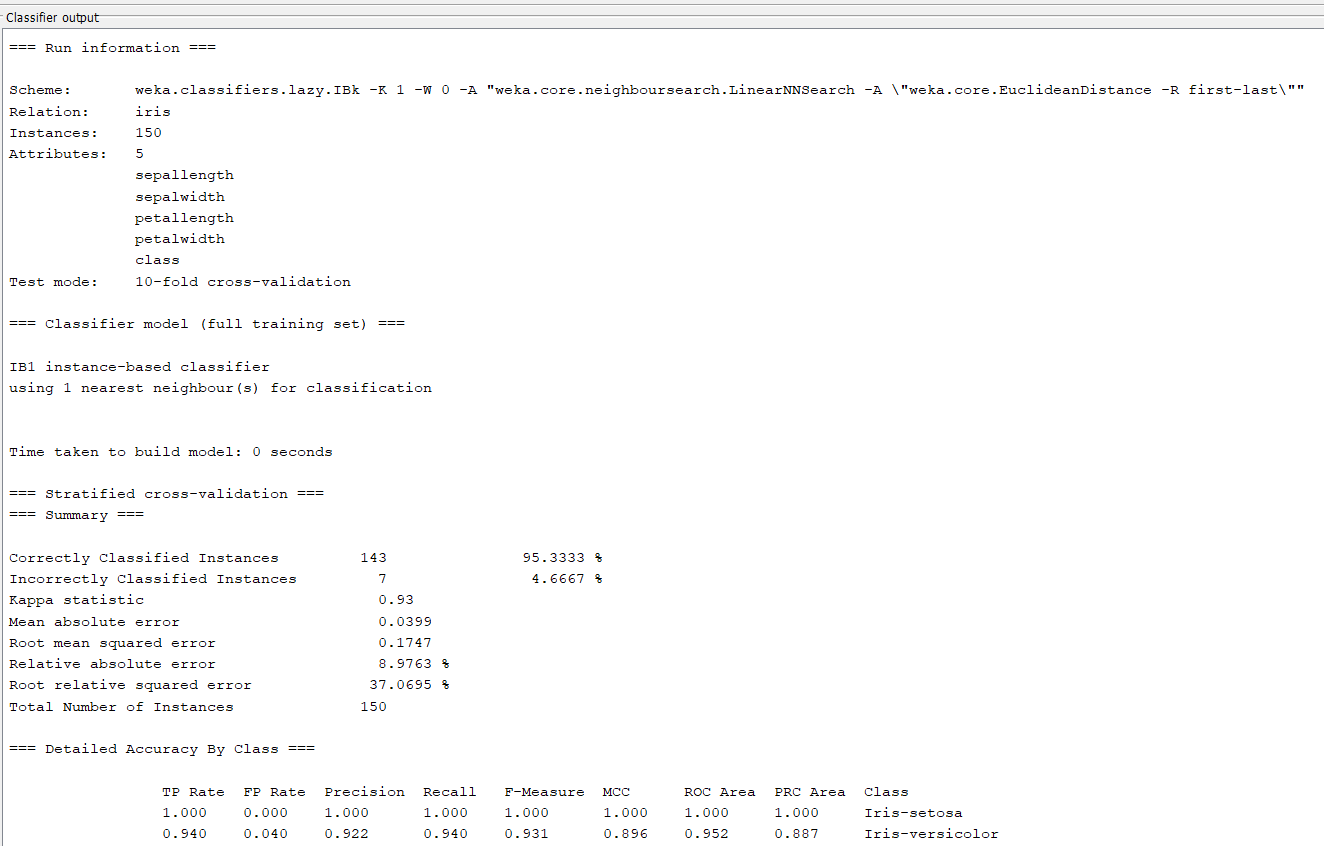


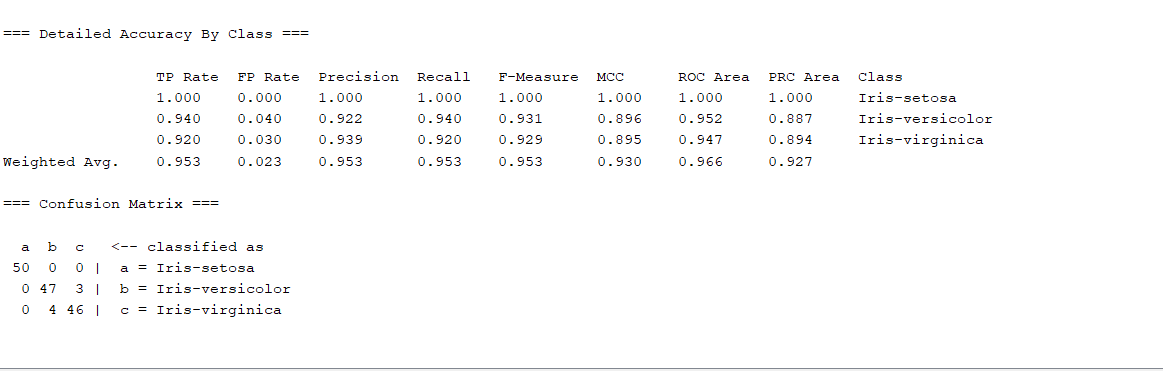
**Step 4:**

Select **Classify** 🡪 Click the **Choose** option 🡪 In the **Classifiers** section, select **Lazy** 🡪 Then, select **IBK** from the options



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





### **IBk (k-NN) – Summary**

* **Dataset**: Iris (150 instances, 5 attributes) (The data used to train and test the model)
* **Evaluation**: 10-fold cross-validation (The method used to test the model’s performance)
* **Accuracy**: 95.33% (143/150 correct) (The percentage of correctly classified instances)
* **Build time**: 0 sec (Lazy learner, computes only at query time)

**Confusion Matrix (Simplified):**

* Setosa → 50/50 correct
* Versicolor → 47/50 correct (3 misclassified as virginica)
* Virginica → 46/50 correct (4 misclassified as versicolor)

**Result:**

* Setosa is classified perfectly.
* Few errors occur between versicolor and virginica.
* Accuracy is slightly lower than Naive Bayes and J48.

**Conclusion:** IBk (k=1) achieved 95.33% accuracy; it performs very well on Iris but, like other models, struggles a bit to separate versicolor and virginica.