

**NAME :- Dev Parekh**

**ROLL NO :-42**

**DIV:-TY09/B**

**Aim:** Perform data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool WEKA.

**Introduction:** Data mining is the process of extracting useful patterns from large datasets. WEKA is a powerful open- source tool that supports various data mining techniques through an easy-to-use interface. In this experiment, we use WEKA to demonstrate three key tasks:

- **Classification:** Predicting predefined class labels (e.g., spam detection).
- **Clustering:** Grouping similar data without prior labels.
- **Association:** Finding relationships between items (e.g., market basket analysis).

Before applying these algorithms, data preprocessing is done to clean and prepare the data for better accuracy.

### **Procedure:**

#### **1. Open Weka Knowledge Flow:**

- Go to **Program Files** on your **PC** and launch **Weka 3.6**.
- Choose the **Knowledge Flow** environment from the initial menu (Explorer, Experimenter, Knowledge Flow, etc.).

#### **2. Load Dataset Using Arff Loader:**

- Drag the **ArffLoader** from the "Data Sources" section into the canvas.

- Right-click → **Configure**, then click **Browse** and select a dataset (e.g., from the **Data** folder like
- `iris.arff` ). This loads your data into the flow.

### 3. Configure Evaluation Component:

- Add the **Evaluation** component to evaluate the clustering model.
- Set the evaluation type to **Static** for using the dataset as-is.

### 4. Prepare the Training Format:

- Add a **TrainingSetMaker** component.
- This prepares your data in a format suitable for training.
- Connect it to the output of the ArffLoader.

### 5. Add and Configure Clusterer:

- Drag the **Clusterer** component into the workspace.
- Choose **SimpleKMeans** as the clustering algorithm.
- Configure it (e.g., set number of clusters, distance function, etc.).

### 6. Analyze Clustering Performance:

- Add the **ClustererPerformanceEvaluator** component.
- Connect it to the output of the Clusterer to measure model effectiveness.

### 7. Add Output Viewers:

- Drag in a **TextViewer** to view textual output (e.g., cluster assignments, summary).
- Add a **Visualization** component for graphical display of cluster distribution.

### 8. Connect Components and Run Flow:

- Right-click on each component to **Connect** them in order: ArffLoader → TrainingSetMaker → Clusterer → ClustererPerformanceEvaluator → TextViewer/Visualization
- Finally, right-click the **last component** and choose **Start Execution** to run the workflow.

## Implementation/Outputs:

Weka KnowledgeFlow Environment

Program File Edit Insert View

Data mining processes Attribute summary Scatter plot matrix SQL Viewer Simple CLI

Design

- SerializedInstancesLoader
- SVMLightLoader
- TextDirectoryLoader
- XRFFLoader
- DataGrid
- DataSinks
- DataGenerators
- Filters
- Classifiers
- Clusters
  - Canopy
  - Cobweb
  - EM
  - FarthestFirst
  - FilteredClusterer
  - HierarchicalClusterer
  - MakeDensityBasedClust
  - SimpleKMeans
- Associations
- AttSelection
- Evaluation
  - TrainingSetMaker
  - TestSetMaker
  - TrainTestSplitMaker
  - ClassAssigner
  - ClassValuePicker
  - ClassifierPerformanceEv
  - ClustererPerformanceEv
  - CrossValidationFoldMak
  - PredictionAppender
  - IncrementalClassifierEva

Untitled1 x

ArffLoader

data Set

TrainingSetMaker

training Set

SimpleKMeans

batchClusterer

Clusterer Performance Evaluator

text

TextViewer

Status Log

Component	Parameters	Time	Status
[KnowledgeFlow]		-	OK.
ArffLoader		-	Finished.
TrainingSetMaker		-	Finished.
SimpleKMeans	-init 0 -max-candidates 100 -peri...	-	Finished.
ClustererPerformanceE...		-	Finished.
TextViewer		-	Finished.

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...Generate...UndoEdit...Save...

FilterChooseNoneApplyStop

Current relation  
Relation: iris  
Instances: 150

Attributes: 5  
Sum of weights: 150

AttributesAllNoneInvertPattern

No.	Name
1	<input checked="" type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

StatusOK

Selected attribute  
Name: sepallength  
Missing: 0 (0%)

Distinct: 35

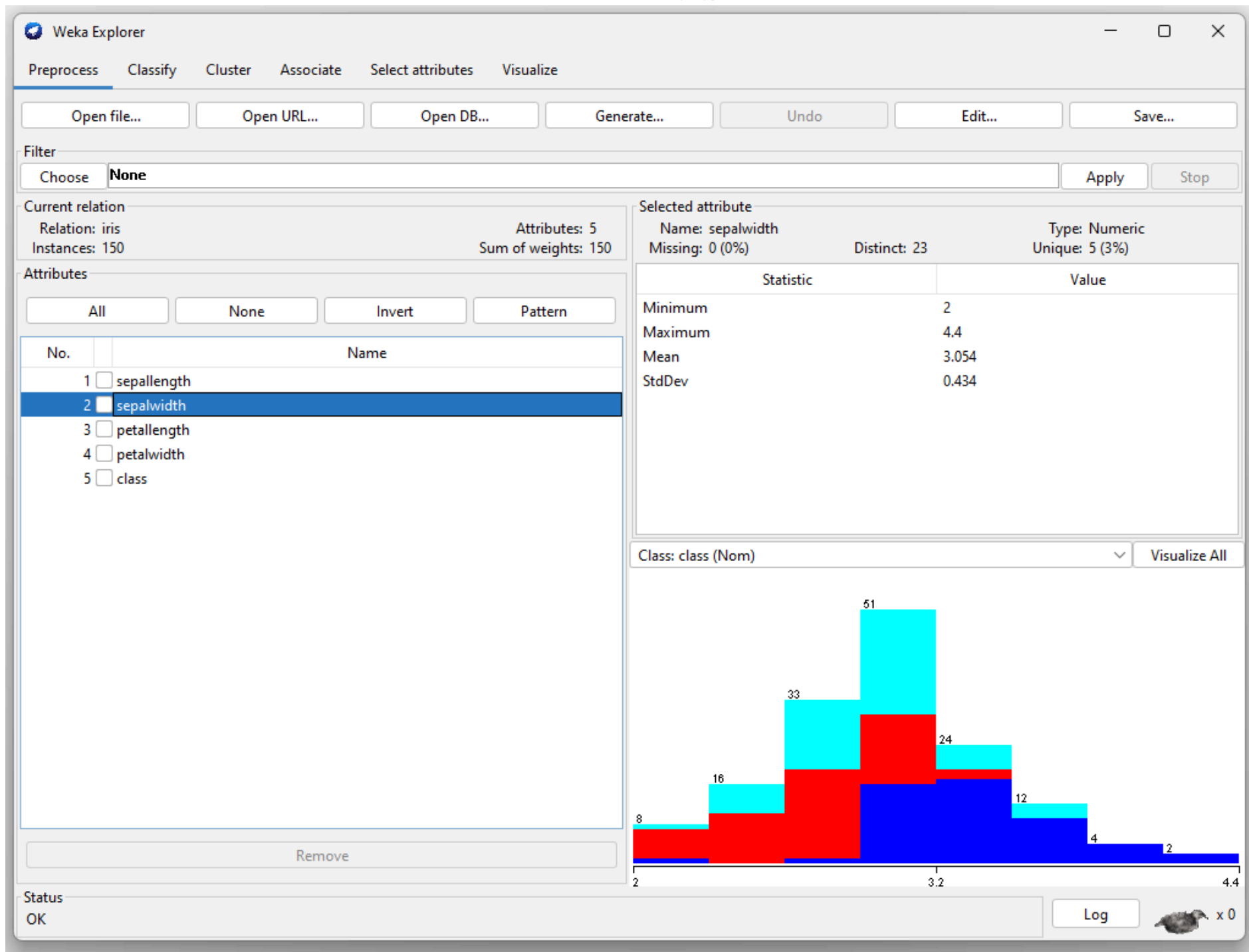
Type: Numeric  
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)Visualize All

Bin Range	Count	Color
4.3 - 5.0	16	Blue
5.0 - 5.7	30	Red
5.7 - 6.4	34	Cyan
6.4 - 7.1	28	Red
7.1 - 7.9	25	Cyan

Log x 0



Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...Generate...UndoEdit...Save...

FilterChooseNoneApplyStop

Current relation  
Relation: iris  
Instances: 150  
Attributes: 5  
Sum of weights: 150

Attributes  
AllNoneInvertPattern  

No.	Name
1	<input type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input checked="" type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

Selected attribute  
Name: petallength  
Missing: 0 (0%)  
Distinct: 43  
Type: Numeric  
Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Class: class (Nom)Visualize All

petallength Range	Count	Color
1.0 - 2.5	50	Blue
2.5 - 3.0	3	Red
3.0 - 5.5	34	Red
5.5 - 6.9	47	Cyan

StatusOKLog x 0



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply Stop

Current relation  
Relation: iris  
Instances: 150  
Attributes: 5  
Sum of weights: 150

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input checked="" type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

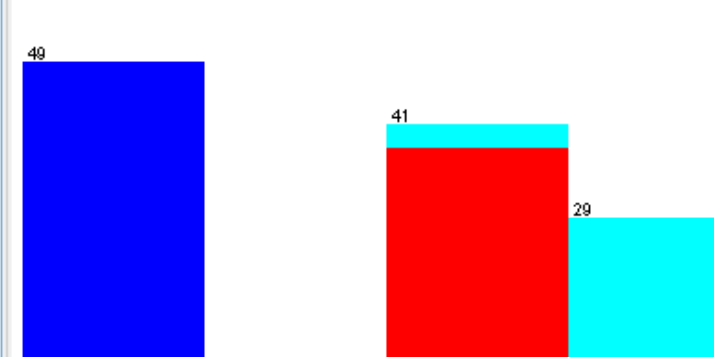
Remove

Status: OK

Selected attribute  
Name: petalwidth  
Missing: 0 (0%)  
Distinct: 22  
Type: Numeric  
Unique: 2 (1%)

Statistic	Value
Minimum	0.1
Maximum	2.5
Mean	1.199
StdDev	0.763

Class: class (Nom) Visualize All



A histogram showing the distribution of petalwidth for three classes: setosa (blue), versicolour (red), and virginica (cyan). The x-axis represents petalwidth from 0.1 to 2.5, and the y-axis represents the count of instances. The setosa class has a single peak at 0.1 with a count of 49. The versicolour class has a peak at 1.3 with a count of 41. The virginica class has a peak at 2.0 with a count of 29. There are also smaller counts of 8 for versicolour at 0.1 and 23 for virginica at 2.5.

Class	Petalwidth	Count
setosa	0.1	49
versicolour	0.1	8
versicolour	1.3	41
virginica	2.0	29
virginica	2.5	23

Log x 0

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active. The 'iris' dataset is loaded, with 150 instances and 5 attributes. The 'Attributes' list on the left shows 'class' selected. The 'Selected attribute' panel on the right displays statistics for 'class' (Nominal, 3 distinct values, 50 instances each). The 'Visualize All' button is visible at the bottom right.

No.	Label	Count	Weight
1	Iris-setosa	50	50
2	Iris-versicolor	50	50
3	Iris-virginica	50	50

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Clusterer

Choose **EM** -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set
 ☐ Supplied test set 
☐ Percentage split % 
☐ Classes to clusters evaluation (Nom) class 
☒ Store clusters for visualization

Start

Stop

Result list (right-click for options)

11:28:21 - EM

Clusterer output

```

=== Run information ===

Scheme:      weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6
Relation:    iris
Instances:   150
Attributes:  5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:   evaluate on training data

=== Clustering model (full training set) ===

EM
==

Number of clusters selected by cross validation: 4
Number of iterations performed: 16

Attribute      Cluster
                0      1      2      3
                (0.32) (0.33) (0.2) (0.14)
=====
sepallength
  mean          5.897  5.006  6.9426  6.1304
  std. dev.     0.5279  0.3489  0.498  0.2943


sepalwidth

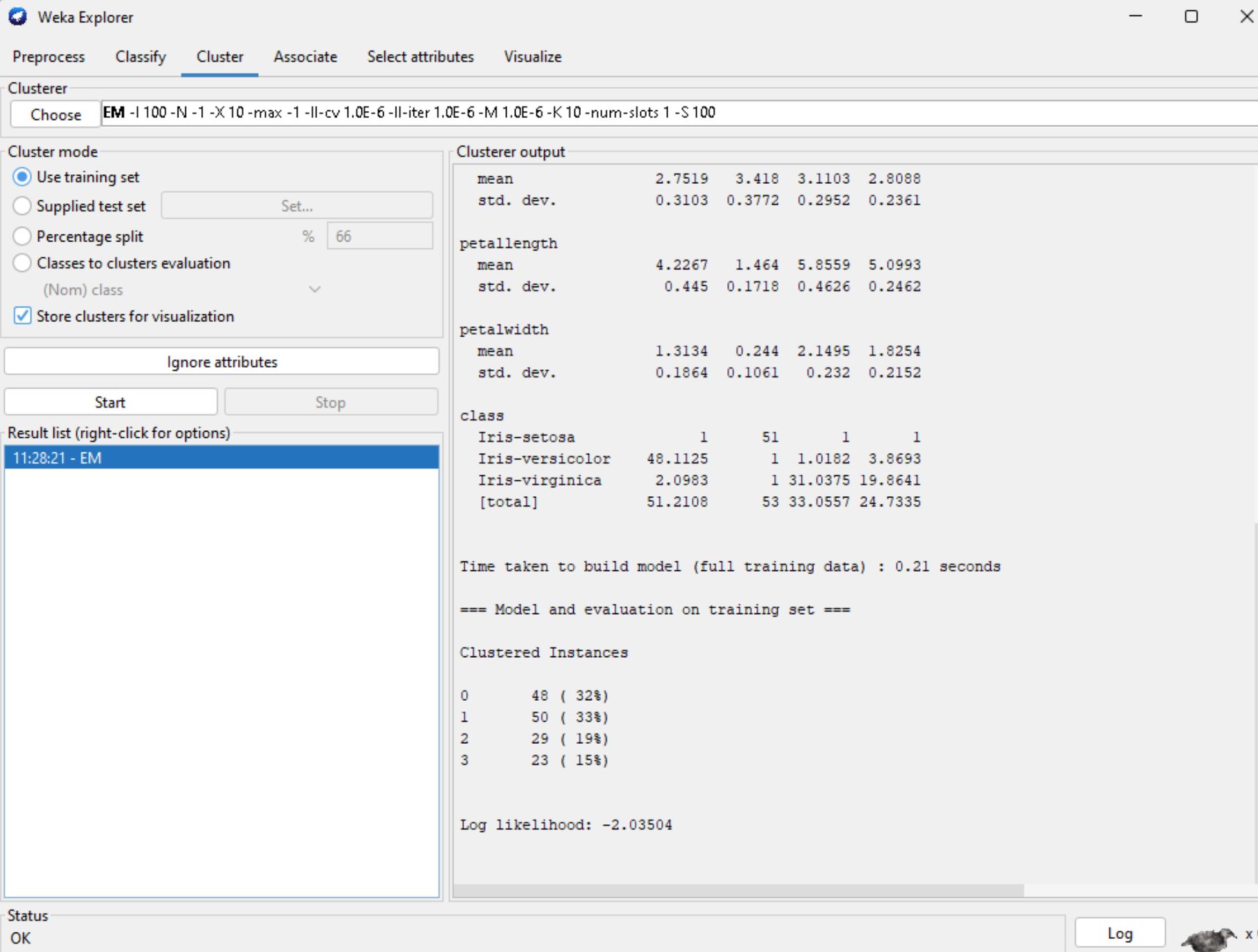
```

Status

OK

Log

 x 0



**Conclusion:** We successfully demonstrated data preprocessing and applied key data mining techniques—Classification, Clustering, and Association—using the WEKA tool. WEKA's intuitive interface and built-in algorithms made it easy to load datasets, configure models, and visualize results. Through this practical approach, we understood how to classify data, group it into clusters, and discover hidden associations, all of which are essential in real-world data analysis and decision-making.