EXP 8:-Demonstrate Classification, Clustering, Association using weka

Aim:

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

Introduction:

WEKA (Waikato Environment for Knowledge Analysis) is a popular open-source software developed at the University of Waikato, New Zealand. It provides a collection of machine learning algorithms for data mining tasks such as data preprocessing, classification, regression, clustering, association rules, and visualization. WEKA supports a user-friendly graphical interface that simplifies applying various algorithms on datasets without writing code.

The tool operates on datasets in the ARFF (Attribute-Relation File Format) or CSV format and allows users to easily load, process, and analyze data. The strength of WEKA lies in its wide range of implemented algorithms and built-in support for cross-validation, visualization, and model evaluation.

In this experiment, we demonstrate:

- Classification: Predicting categorical labels based on input features using algorithms like J48, Naive Bayes, etc.
- Clustering: Grouping similar data points together without pre-defined labels using algorithms like k-Means.
- Association: Discovering interesting relationships or associations among attributes in large datasets using Apriori algorithm.

Procedure

Step 1: Load Dataset

- Open WEKA GUI Chooser.
- Choose "Explorer".
- Click on "Open file" and load a dataset

Step 2: Preprocess Data

- View attribute summary.
- Remove or transform attributes if required.
- Ensure no missing values or irrelevant fields.

Step 3: Apply Classification

- Go to "Classify" tab.
- Choose classifier (e.g., J48 decision tree, Naive Bayes).
- Select class attribute.

- Click "Start" to train and test the model.
- Analyze results: accuracy, confusion matrix, ROC curve, etc.

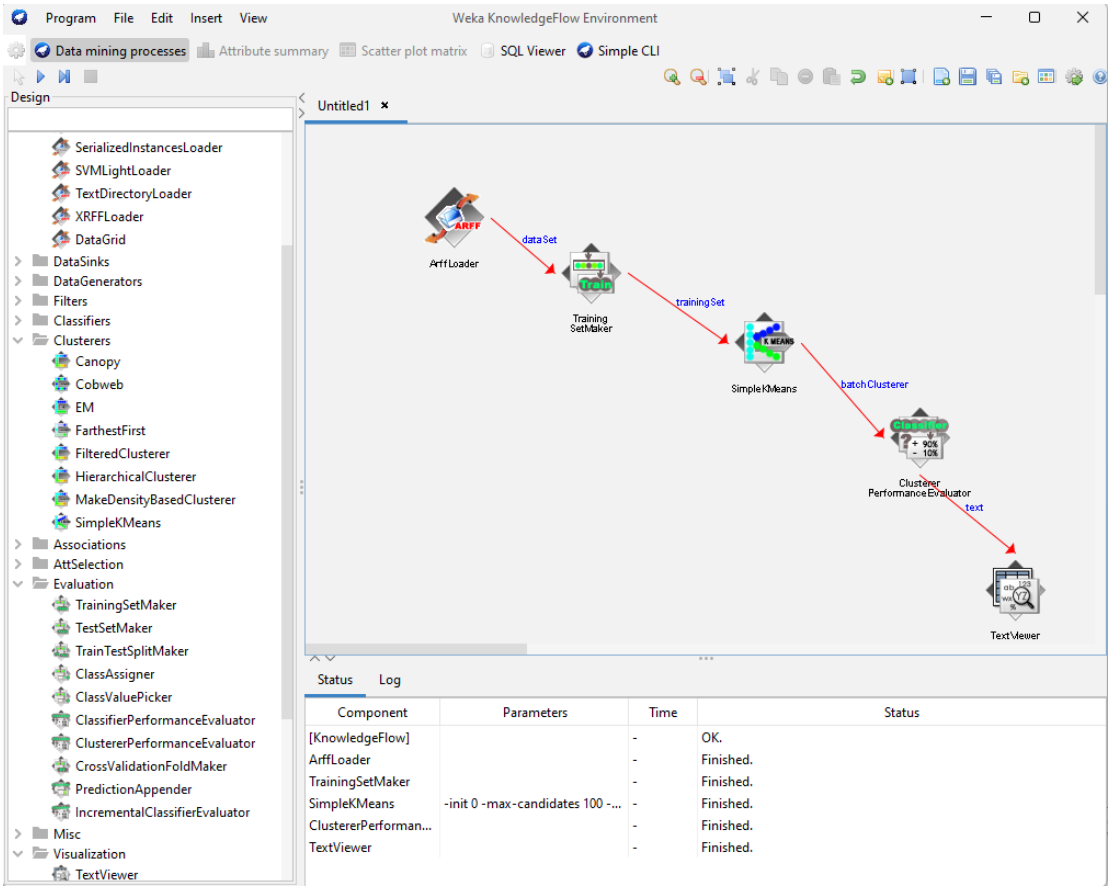
Step 4: Apply Clustering

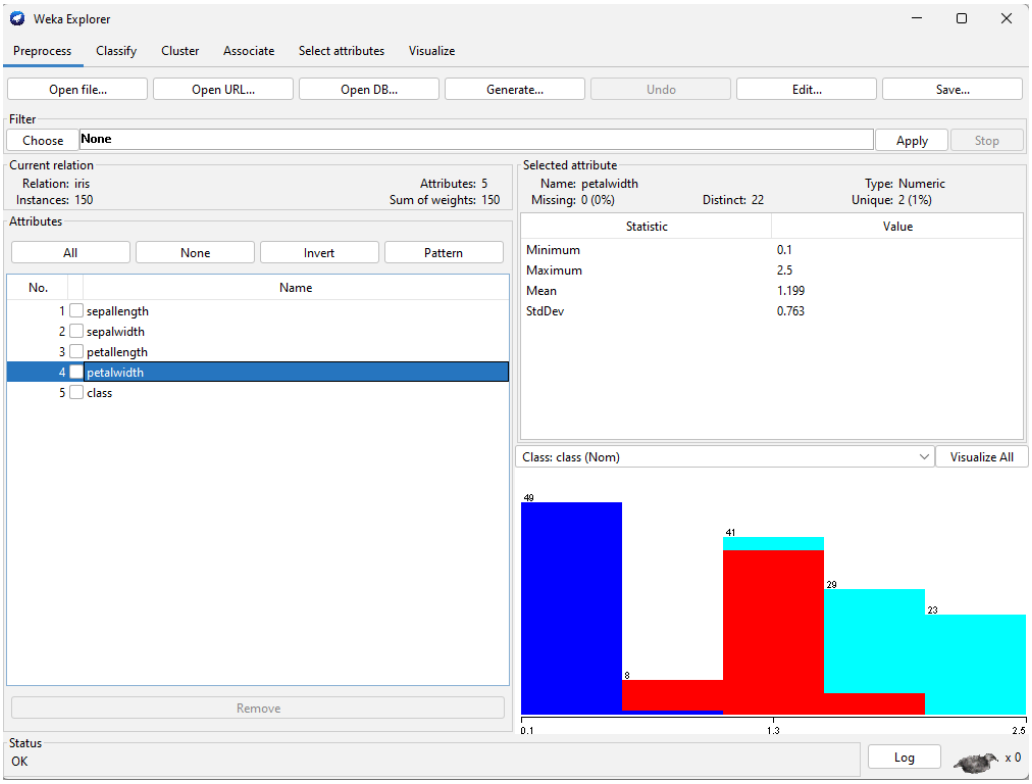
- Go to "Cluster" tab.
- Choose clustering algorithm (e.g., SimpleKMeans).
- Configure number of clusters.
- Click "Start" and view cluster assignments.

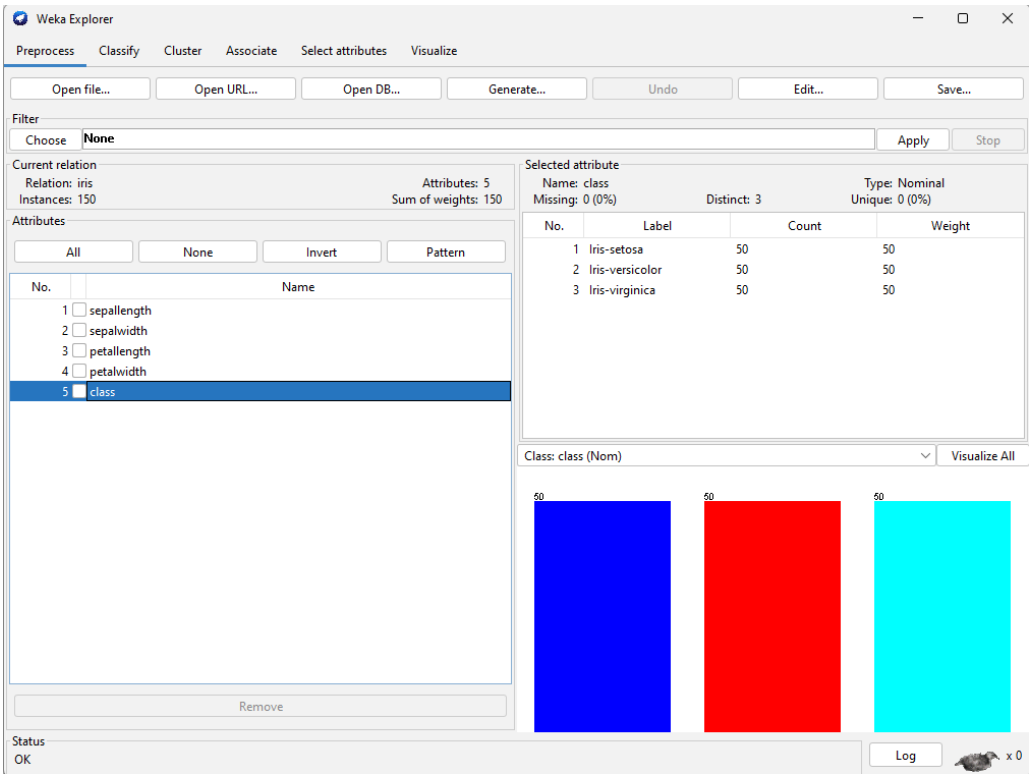
Step 5: Apply Association

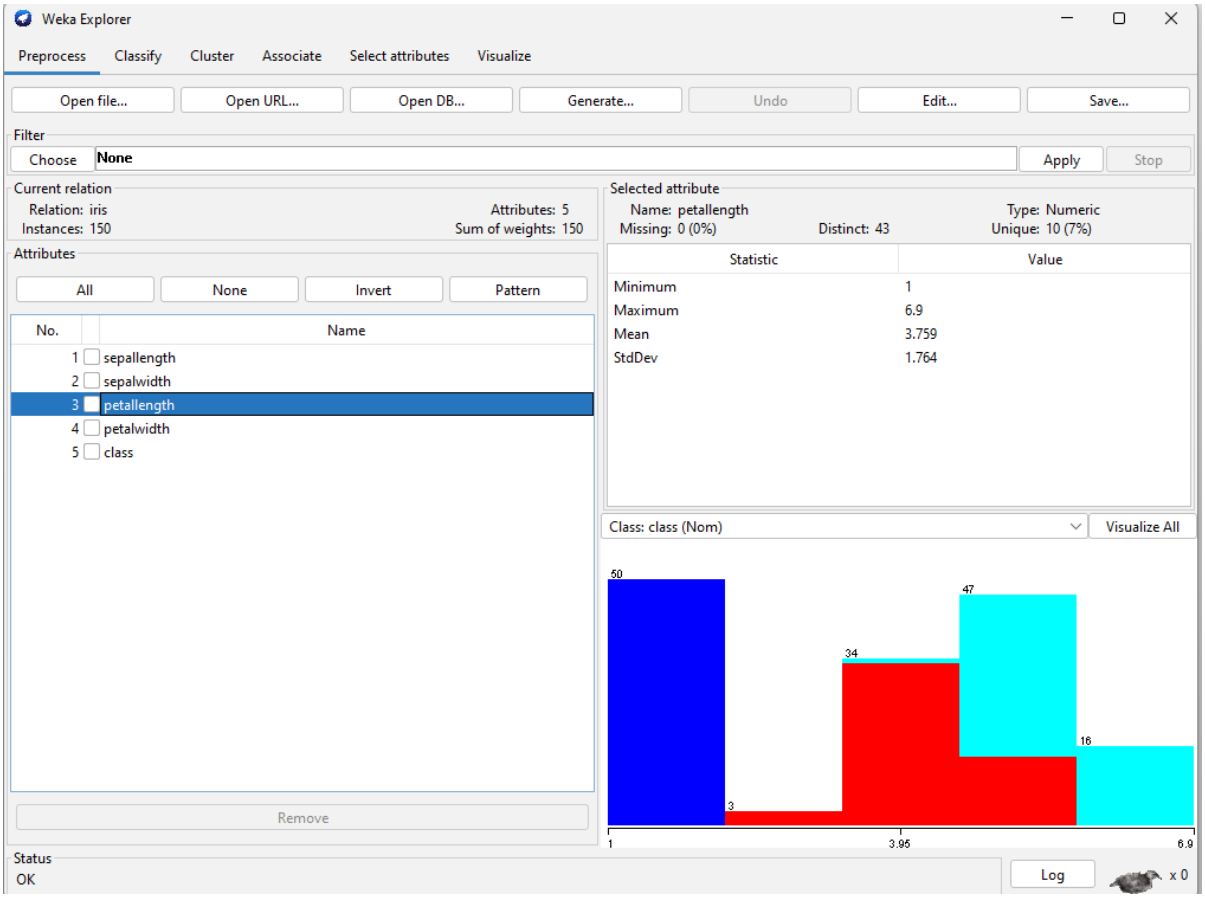
- Go to "Associate" tab.
- Choose algorithm (e.g., Apriori).
- Set minimum support and confidence thresholds.
- Click "Start" and view generated rules.

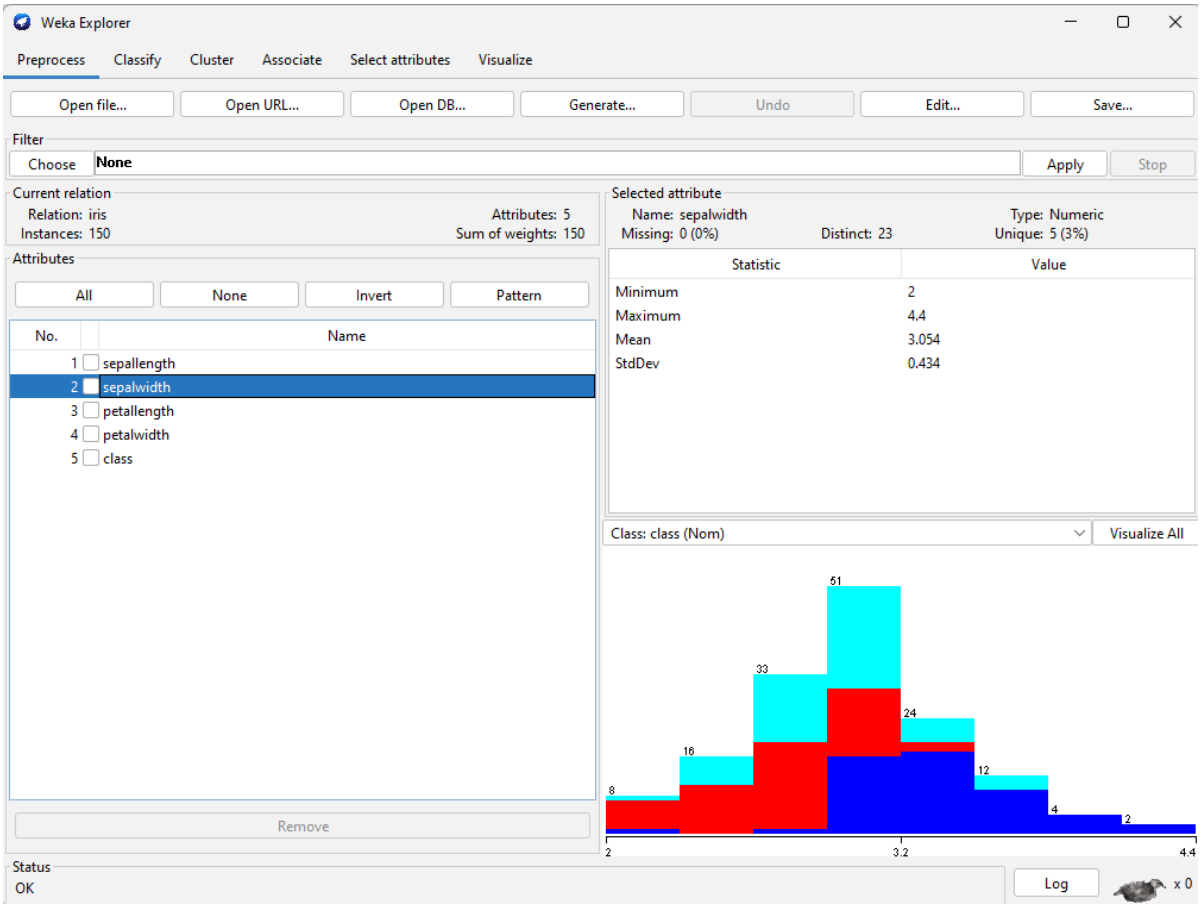
Output

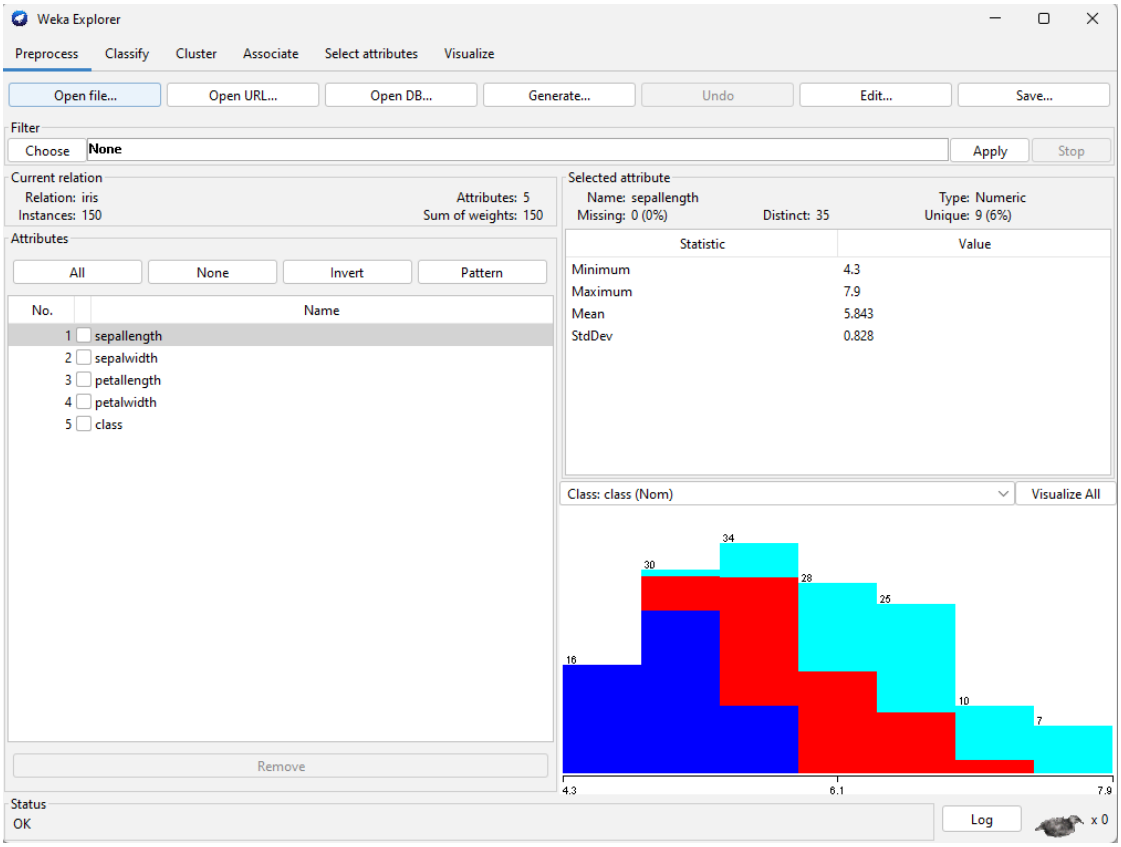


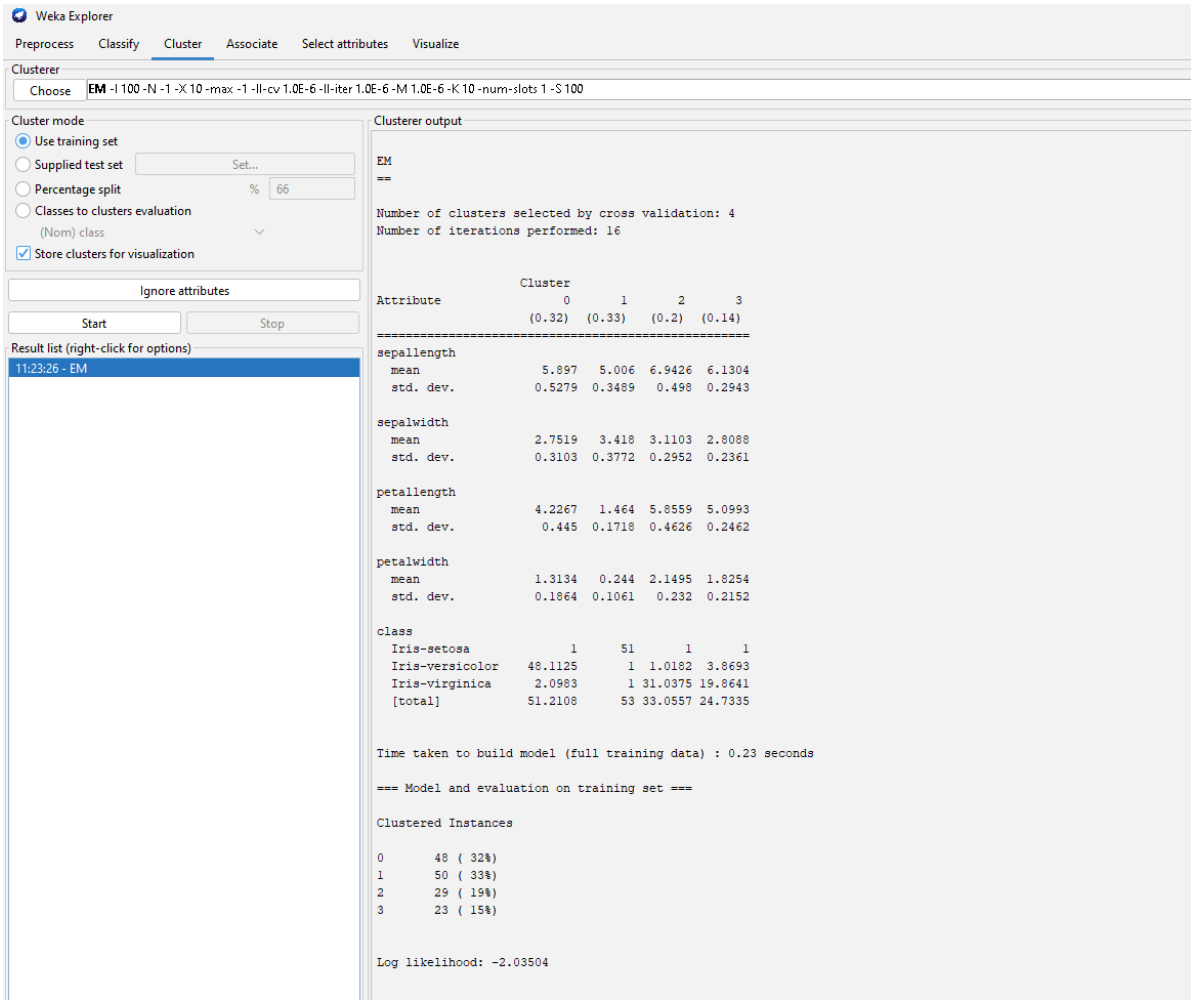

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: petalwidth
Missing: 0 (0%)
Distinct: 22
Type: Numeric
Unique: 2 (1%)
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum 0.1
Maximum 2.5
Mean 1.199
StdDev 0.763
Class: class (Nom)
Visualize All
49
8
41
29
23
0.1
1.3
2.5
Remove
Status
OK
Log
x 0

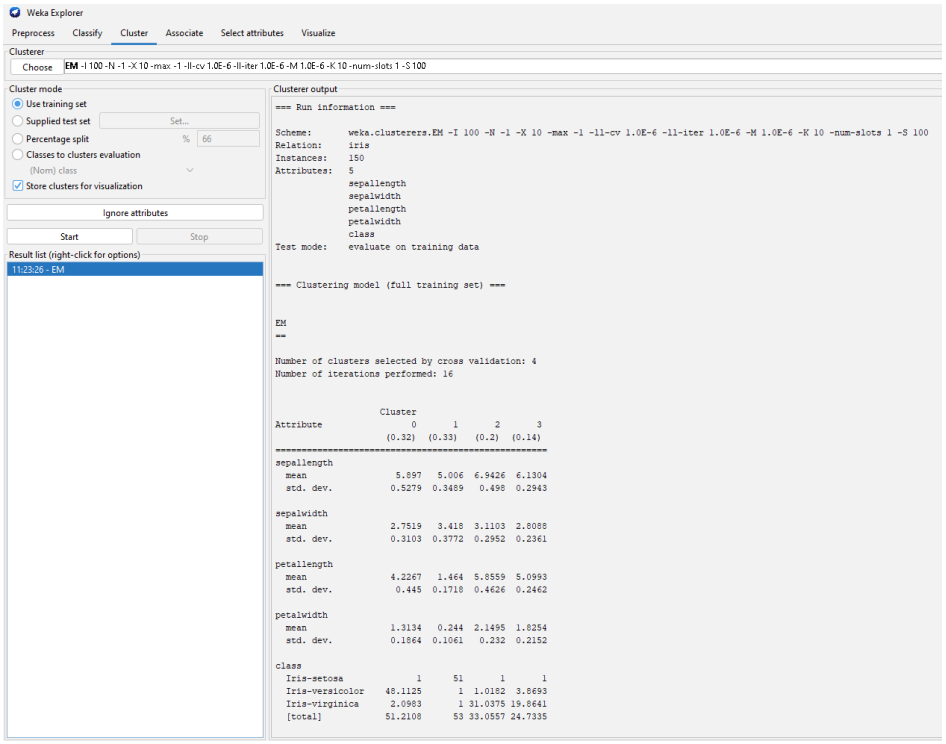

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
No.
Label
Count
Weight
1 Iris-setosa 50 50
2 Iris-versicolor 50 50
3 Iris-virginica 50 50
Class: class (Nom)
Visualize All
50
50
50
Remove
Status
OK
Log
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 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 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
50
3
34
47
16
1
3.95
6.9
Status
OK
Log
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 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Selected attribute
Name: sepalwidth
Missing: 0 (0%)
Distinct: 23
Type: Numeric
Unique: 5 (3%)
Statistic
Value
Minimum 2
Maximum 4.4
Mean 3.054
StdDev 0.434
Class: class (Nom)
Visualize All
8
16
33
51
24
12
4
2
2
3.2
4.4
Status
OK
Log
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
Selected attribute
Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Class: class (Nom)
Visualize All
34
30
28
26
16
10
7
4.3
6.1
7.9
Remove
Status
OK
Log
x 0


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
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
11:23:26 - EM
Clusterer output
EM
==
Number of clusters selected by cross validation: 4
Number of iterations performed: 16
Cluster
Attribute 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
mean 2.7519 3.418 3.1103 2.8088
std. dev. 0.3103 0.3772 0.2952 0.2361
petallength
mean 4.2267 1.464 5.8559 5.0993
std. dev. 0.445 0.1718 0.4626 0.2462
petalwidth
mean 1.3134 0.244 2.1495 1.8254
std. dev. 0.1864 0.1061 0.232 0.2152
class
Iris-setosa 1 51 1 1
Iris-versicolor 48.1125 1 1.0182 3.8693
Iris-virginica 2.0983 1 31.0375 19.8641
[total] 51.2108 53 33.0557 24.7335
Time taken to build model (full training data) : 0.23 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 48 (32%)
1 50 (33%)
2 29 (19%)
3 23 (15%)
Log likelihood: -2.03504



Conclusion

In this experiment, we explored how to use the WEKA tool to apply various data mining techniques, including classification, clustering, and association rule mining. We learned how to load and preprocess data, apply different algorithms, and interpret the results effectively. WEKA provides a simple yet powerful interface for applying machine learning models and is a great educational tool for understanding core concepts in data mining.

Github link:-https://github.com/omk279/DWM