Name: Sonika Kulkarni

Class: TY3-B-34

Subject: DWM

Experiment No. 8

Title: 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: 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 **Clusterer Performance Evaluator** 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 → TrainingSet Maker → Clusterer → ClustererPerformance Evaluator → TextViewer/Visualization
- Finally, right-click the **last component** and choose **Start Execution** to run the workflow.

Implementation/Outputs :

DataSinks

DataGenerators

Filters

Classifiers

Clusterers

Cano
py

Cobweb

EM

FarthestFirst

Filtered Clusterer

Hierarchical Clusterer

MakeDensityBasedClu
st

SimpleKMean
s

Associations

AttSelection

Evaluation

TrainingSetMake
r

Untitled1 x

ARFF

data Set

Arff Loader

Train

training
Set

Training
SetMaker

KMEAN
S

batchClusterer

Simple
KMeans

Clake Rat

TestSetMaker

Train
TestSplitMaker

Status Log

ClassAssig
ner

ClassValuePicker

Compone
nt

[KnowledgeFlow]

Parameters

Time

OK

.

ClassifierPerformanceEv

ArffLoader

Finished.

ClustererPerformanceEv

CrossValidationFoldMak

Training
SetMaker
SimpleKMeans

Finished.

-init 0-max-candidates 100-peri...

Finishe
d.

PredictionAppender

ClustererPerformanceE...

Finished.

IncrementalClassifierEva TextViewer

Finished.

90
%
10%

Clusterer Performance
Evaluator

test

Status

ok

Text
Mewer

Weka
Explorer

Preproces
s

Classify Cluster Associate

Select attributes Visualize

Open
file...

Open
URL...

Open
DB...

Generate...

Undo

Edit...

Filter

Choose None

Current relation

Relation: iris

Instances: 150

Attributes

All

No.

1
sepallengt
h

2
sepalwidt
h

3
petalleng
th

4
petalwid
th

5
class

Statu
s

OK

None

Invert

Name

Save...

X

Appl
y

Stop

Attributes: 5 Sum of weights: 150

Selected attribute

Name: sepallength Missing: 0 (0%)

Distinct: 35

Statistic

Type: Numeric Unique: 9 (6%)

Valuc

Pattern

Minimum

Maximum

Mean

StdDev

Class: class (Nom)

16

30

34

4.3

7.9

5.843

0.828

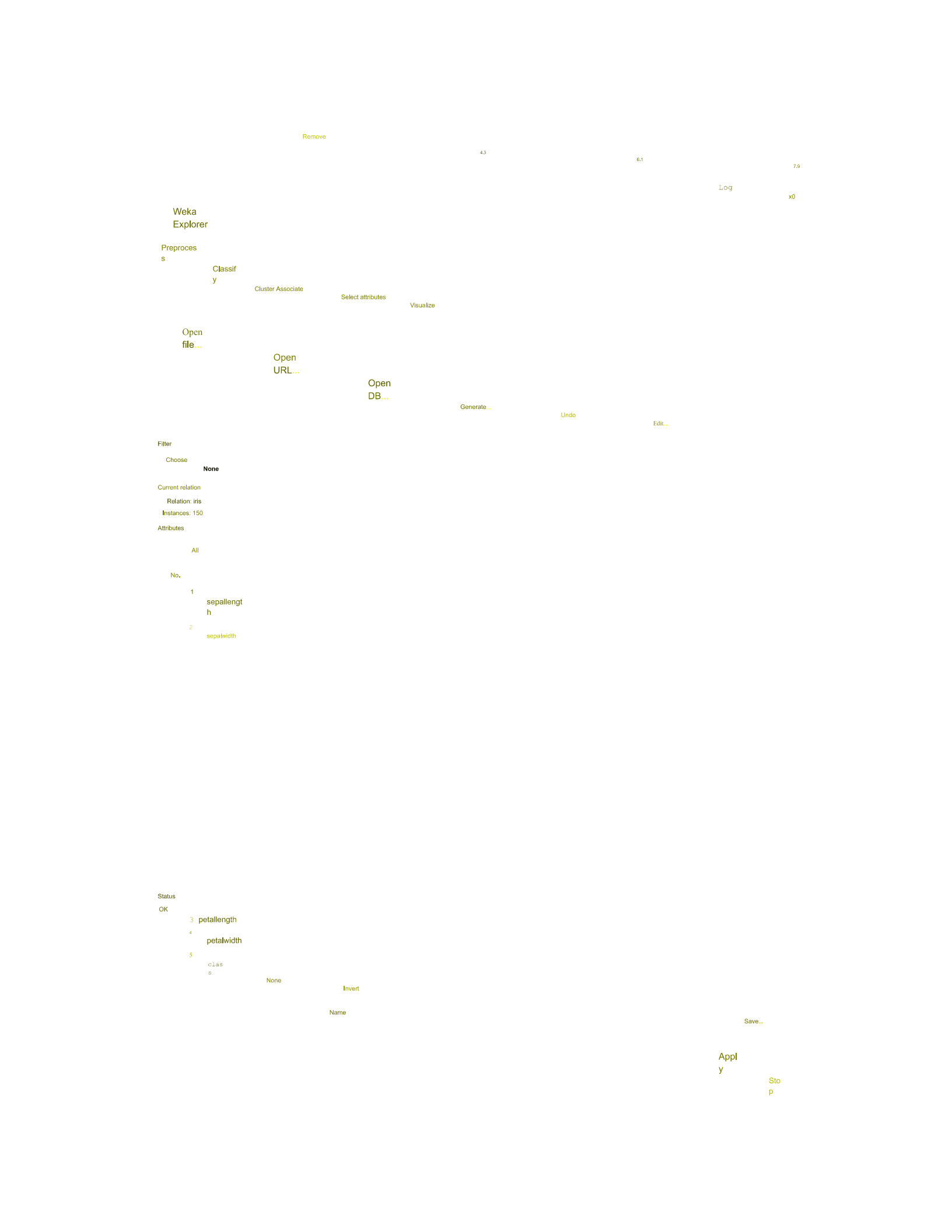
28

25

10

Visualize All

7


Remove
4.3
6.1
7.9
Log
x0
Weka
Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Filter
Choose
None
Current relation
Relation: iris
Instances: 150
Attributes
All
No.
1
sepallength
2
sepalwidth
Status
OK
3
petallength
4
petalwidth
5
class
None
Invert
Name
Save...
Apply
Stop

Attributes: 5 Sum of weights: 150

Selected attribute

Name: sepalwidth Missing: 0 (0%)

Distinct: 23

Statistic

Type: Numeric Unique: 5 (3%)

Valuc

Pattern

Remove

2

Minimum

Maximum

Mean

StdDev

Class: class (Nom)

8

16

33

51

2

4.4

3.054

0.434

24


12
X
Visualize All
4
2
4.4
3.
2
Log
x0
Weka Explorer
Preproces
s
Classify Cluster Associate
Select attributes Visualize
Open
file...
Open
URL...
Open
DB...
Generate...
Undo
Edit...
Filter
Choose
None
Current relation
Relation: iris
Instances: 150
Attributes

Status

OK

All

No.

1 sepallength

2 sepalwidth

3 petallength

4 petalwidth

5 class

None

Invert

Name

Save...

Apply

Stop

Attributes: 5 Sum of weights: 150

Selected attribute

Name: petallength Missing: 0 (0%)

Distinct: 43

Statistic

Type: Numeric Unique: 10 (7%)

Value

Pattern

Minimum

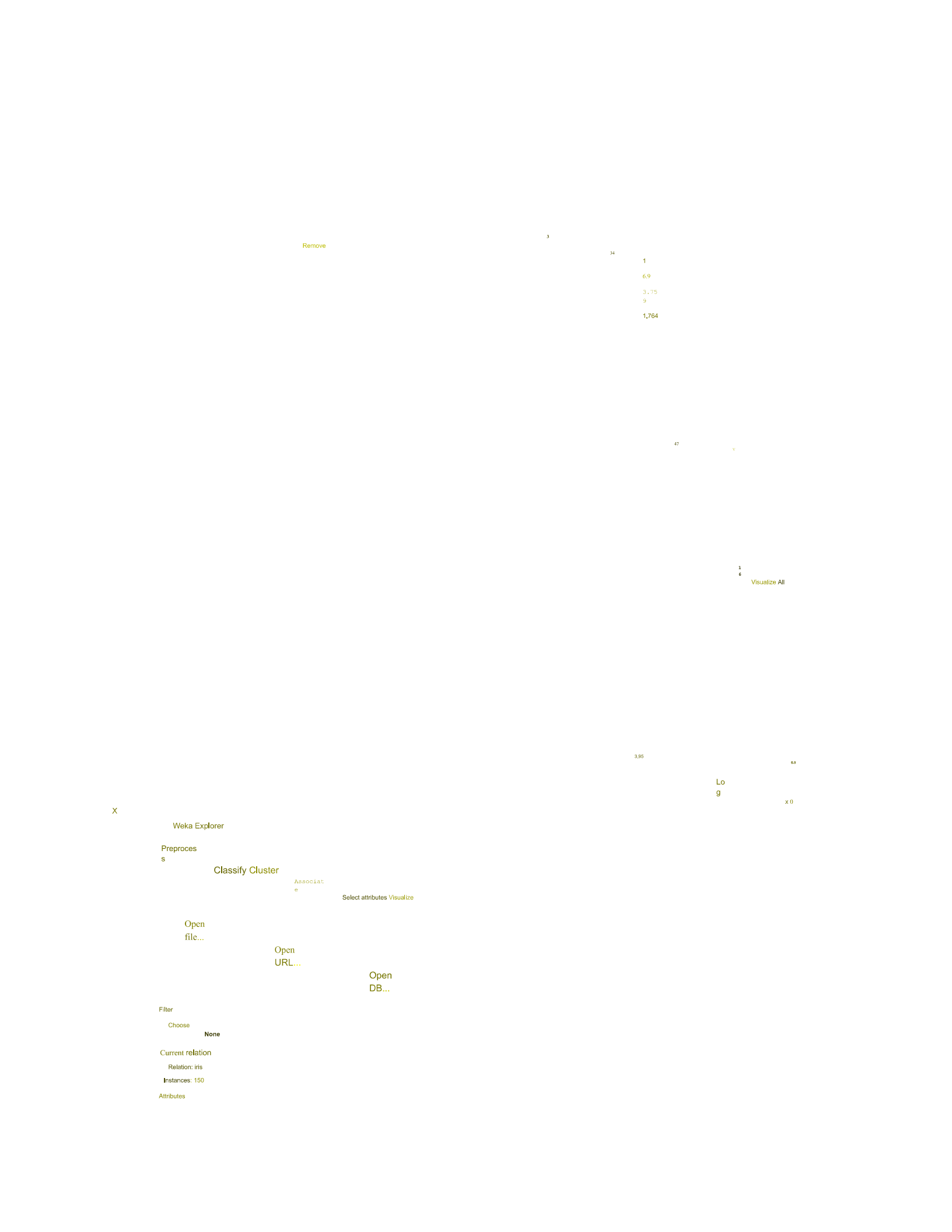
Maximum

Mean

StdDev

Class: class (Nom)

50


3
Remove
34
1
6.9
3.75
9
1.764
47
v
1
6
Visualize All
3.95
6.3
Lo
g
x 0
X
Weka Explorer
Preproces
s
Classify Cluster
Associat
e
Select attributes Visualize
Open
file...
Open
URL...
Open
DB...
Filter
Choose
None
Current relation
Relation: iris
Instances: 150
Attributes

All

No.

None

Invert

Name

1 sepallength

2 sepalwidth

Status

OK

3 petallength

4 petalwidth

5 class

Generate...

Undo

Edit...

Save...

Appl
y

Stop

Attributes: 5 Sum of weights: 150

Selected attribute

Name: petalwidth Missing: 0 (0%)

Distinct: 22

Type: Numeric Unique: 2 (1%)

Statistic

Value

Pattern

Minimum

Maximum

Mean

StdDev

Class: class
(Nom)

49


Remove
0.1
8
41
0.1
2.5
1.19
9
0.763
29
23
X
Visualize All
1.
3
2.5
Log
x 0
Weka
Explorer
Preproces
s
Classify Cluster
Associate

Select attributes Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Filter

Choose None

Current relation

Relation: iris

Instances: 150

Attributes

All

No.

1 sepallength

2 sepalwidth

Status

OK

3 petallength

4 petalwidth

5 class

None

Invert

Name

Remove

Save...

Appl
y

Stop

Selected attribute

Attributes: 5 Sum of weights: 150

Name: class

Missing: 0 (0%)

Distinct: 3

Type: Nominal Unique: 0 (0%)

No.

Label

Count

Weight

Pattern

1 Iris-setosa

50

2 Iris-versicolor

3 Iris-virginica

5
0

50

5555

50

50

50

Class: class
(Nom)

50

50

50

X

Visualize
All

Lo

g

x 0

Weka
Explorer

Preproces
s

Clusterer

Classify Cluster Associate

Select attributes Visualize

Choose

Cluster mode

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

Clusterer output

X

Use training
set

Supplied test
set

Set...

Percentage
split

%

66

Classes to clusters evaluation

(Nom)
class

Scheme :

Relation:

Instances:

Attributes:

=== Run information ===

weka.clusterers.EM -I 100 -N 1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6

iris

150

5

Store clusters for visualization

sepallengt
h

sepalwidth

Ignore
attributes

petallength

petalwidt
h

Start

Stop

class

Test mode:

evaluate on training
data

Result list (right-click for options)

11:28:21 - EM

Status

OK

=== Clustering model (full training set)

EM

==

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

Attribute	Cluster 0 (0.32)	1	2 (0.33)	3 (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				

Log

x 0

Weka Explorer

Preprocess

Clusterer

Classify Cluster Associate Select attributes Visualize

Choose **EM**-I 100-N-1-X 10-max **-1** -II-cv 1.0E-6-**11**-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

Clusterer output

mean				
std dev.				
petallength				
mean	2.7519	3.418	3.1103	2.8688
std. dev.	0.3103	0.3772	0.2952	0.2361
	4.2267	1.464	5.8559	5.0993
	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

Store clusters for visualization

Ignore attributes

Start

Stop

Result list (right-click for options)

11:28:21 - EM

class

Iris-setosa

Iris-versicolor

Iris-virginica [total]

1

51

1

1

48.1125

2.0983

51.2108

1 1.0182 3.8693

1 31.0375 19.8641

53 33.0557 24.7335

Status

OK

Time taken to build model (full training data) : 0.21 seconds

=== Model and evaluation on training set ===

Clustered Instances

O

H

N

m

0

48 (32%)

1

50 (33%)

2

29 (19%)

3

23 (15%)

Log likelihood: -2.03504

0

X

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.

GitHub Link: https://github.com/sonikak19/DWM_Exp8.git