

A
Practical File
On
Data ware house and data mining

Paper Code: ITD08



Bachelor of Engineering
(2016-2020)
Computer Science Engineering

Submitted By:

Kavita Maurya

Batch-3

Group-1

2016UCO1579

1.

AIM:

Create an Employee Table with the help of Data Mining Tool WEKA.

DESCRIPTION:

We need to create an Employee Table with training data set which includes attributes like name, id, salary, experience, gender, phone number.

PROCEDURE:

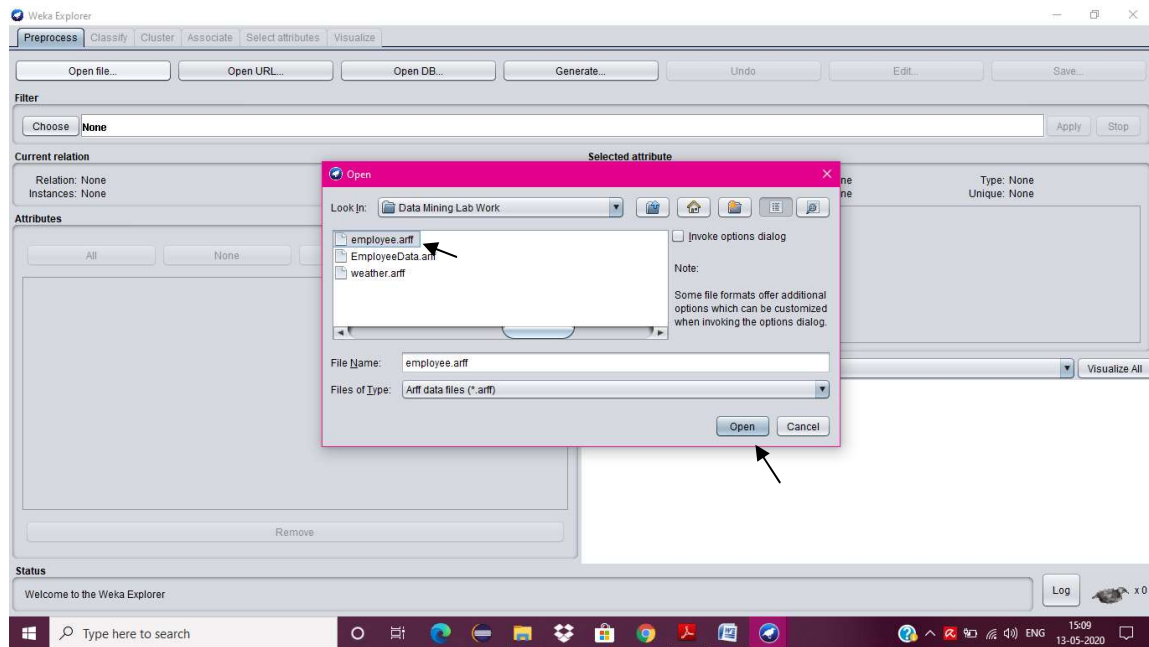
1. Open Start -> Programs -> Accessories -> Notepad.
2. Type the following training data set with the help of Notepad for Employee Table.

```
@relation employee
@attribute name {x,y,z,a,b}
@attribute id numeric
@attribute salary {low,medium,high}
@attribute exp numeric
@attribute gender {male,female}
@attribute phone numeric
@data
x,101,low,2,male,250311
y,102,high,3,female,251665
z,103,medium,1,male,240238
a,104,low,5,female,200200
b,105,high,2,male,240240
```

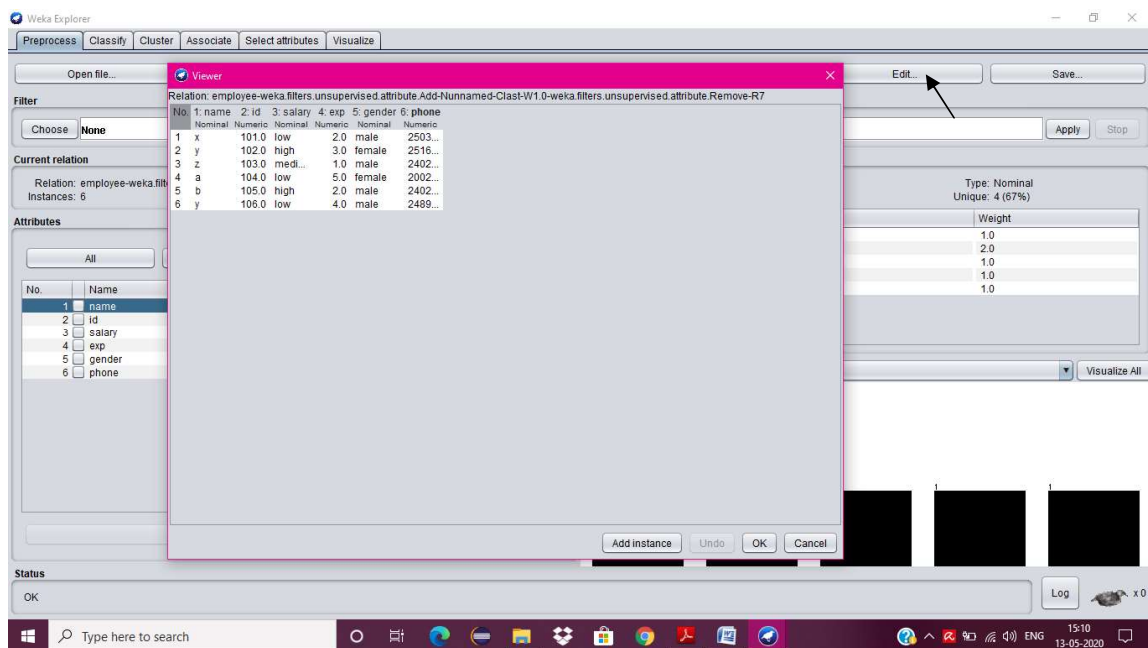
3. After that, save the file in **.arff** format.
4. Click on **weka-3-4**, then Weka dialog box is displayed on the screen.
5. In that dialog box there are four modes, click on **explorer**.



6. Explorer shows many options. In that click on '**open file**' and select the arff file.



7. Click on **edit** button which shows employee table on weka .



2.

AIM:

Apply Pre-Processing techniques to the training data set of Employee Table.

DESCRIPTION:

Real world databases are highly influenced to noise, missing and inconsistency due to their queue size so the data can be pre-processed to improve the quality of data and missing results and it also improves the efficiency.

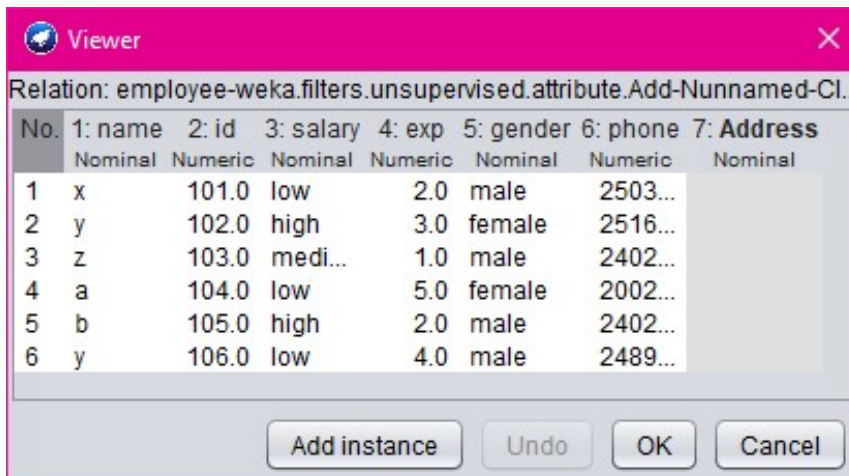
There are 3 pre-processing techniques, they are:

- Add
- Remove
- Normalization

PROCEDURE:

ADD:

1. Start -> Programs -> Weka-3-4 ->Weka-3-4
2. Click on **explorer**.
3. Click on **open file**.
4. Select **Employee.arff** file and click on open.
5. Click on **Choose button** and select the **Filters option**.
6. In Filters, we have **Supervised** and **Unsupervised data**.
7. Click on **Unsupervised data**.
8. Select the attribute **Add**.
9. A new window is opened.
10. In that we enter attribute index, type, data format, nominal label values for **Address**.
11. Click on **OK**.
12. Press the **Apply button**, then a new attribute is added to the Employee Table.
13. **Save** the file.
14. Click on the **Edit button**, it shows a new Employee Table on Weka.



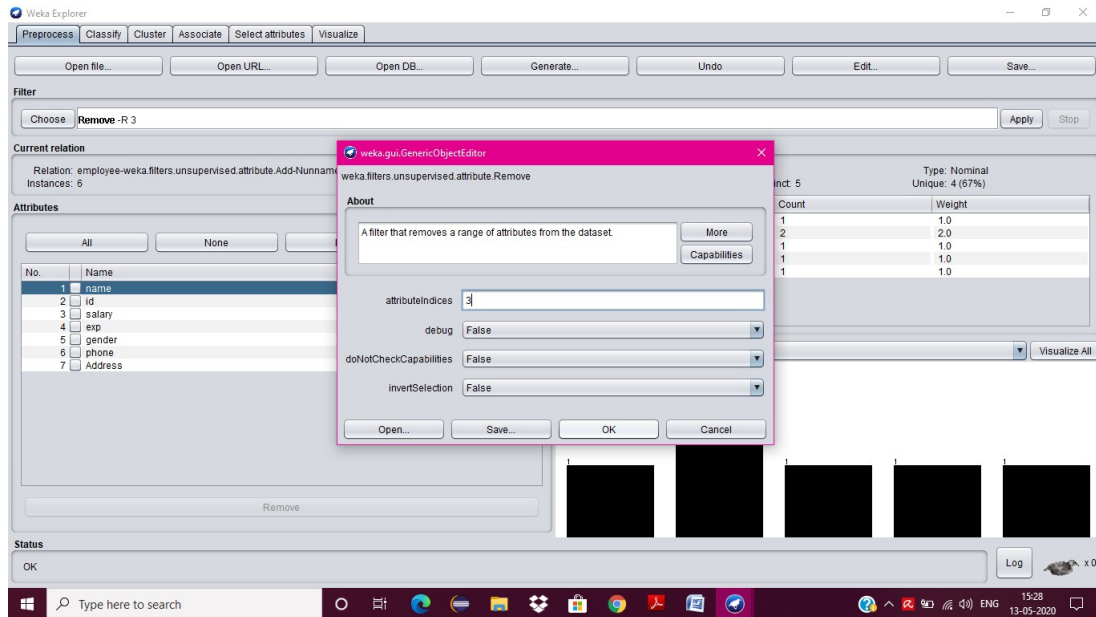
Relation: employee-weka.filters.unsupervised.attribute.Add-Nunamed-Cl...

No.	1: name	2: id	3: salary	4: exp	5: gender	6: phone	7: Address
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal
1	x	101.0	low	2.0	male	2503...	
2	y	102.0	high	3.0	female	2516...	
3	z	103.0	medi...	1.0	male	2402...	
4	a	104.0	low	5.0	female	2002...	
5	b	105.0	high	2.0	male	2402...	
6	y	106.0	low	4.0	male	2489...	

Add instance Undo OK Cancel

REMOVE:

1. Click on **Choose** button and select the **Filters** option.
2. In Filters, we have **Supervised** and **Unsupervised** data.
3. Click on **Unsupervised** data.
4. Select the attribute **Remove**.
5. Select the attributes **salary** to Remove.



6. Click **Remove** button and then **Save**.
7. Click on the **Edit** button, it shows a new Employee Table on Weka.

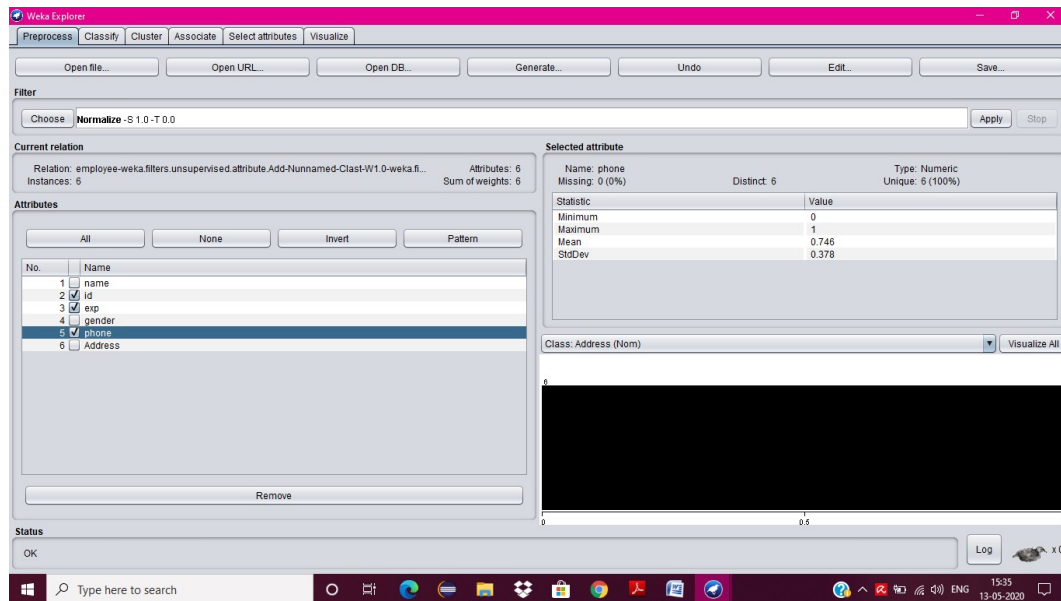
The screenshot shows the Weka Viewer window displaying the resulting dataset. The relation is 'employee-weka.filters.unsupervised.attribute.Add-Nu...'. The table has 6 columns: No., 1: name, 2: id, 3: exp, 4: gender, 5: phone, and 6: Address. The data is as follows:

No.	1: name	2: id	3: exp	4: gender	5: phone	6: Address
	Nominal	Numeric	Numeric	Nominal	Numeric	Nominal
1	x	101.0	2.0	male	2503...	
2	y	102.0	3.0	female	2516...	
3	z	103.0	1.0	male	2402...	
4	a	104.0	5.0	female	2002...	
5	b	105.0	2.0	male	2402...	
6	y	106.0	4.0	male	2489...	

At the bottom of the window, there are buttons for 'Add instance', 'Undo', 'OK', and 'Cancel'.

NORMALIZE:

1. Click on **Choose** button and select the **Filters** option.
2. In Filters, we have **Supervised** and **Unsupervised data**.
3. Click on **Unsupervised data**.
4. Select the attribute **Normalize**.
5. Select the attributes **id**, **experience**, **phone** to Normalize.



6. Click on **Apply** button and then **Save**.
7. Click on the **Edit** button, it shows a new Employee Table with normalized values on Weka.

Relation: employee-weka.filters.unsupervised.attribute.Add-Nunna...

No.	1: name	2: id	3: exp	4: gender	5: phone	6: Address
	Nominal	Numeric	Numeric	Nominal	Numeric	Nominal
1	x	0.0	0.25	male	0.973...	
2	y	0.2	0.5	female	1.0	
3	z	0.4	0.0	male	0.777...	
4	a	0.6	1.0	female	0.0	
5	b	0.8	0.25	male	0.778	
6	y	1.0	0.75	male	0.947	

Right click (or left+alt)

Add instance Undo OK Cancel

3.

AIM:

Normalize Weather Table data using Knowledge Flow.

DESCRIPTION:

The knowledge flow provides an alternative way to the explorer as a graphical front end to WEKA's algorithm. Knowledge flow is a working progress. So, some of the functionality from explorer is not yet available. So, on the other hand there are the things that can be done in knowledge flow, but not in explorer. Knowledge flow presents a dataflow interface to WEKA. The user can select WEKA components from a toolbar placed them on a layout canvas and connect them together in order to form a knowledge flow for processing and analyzing the data.

PROCEDURE:

1. Open Start -> Programs -> Weka-3-4 -> Weka-3-4
2. Open the Knowledge Flow.
3. Select the Data Source component and add Arff Loader into the knowledge layout canvas.
4. Select the Filters component and add Attribute Selection and Normalize into the knowledge layout canvas.
5. Select the Data Sinks component and add Arff Saver into the knowledge layout canvas.
6. Right click on Arff Loader and select Configure option then the new window will be opened and select Weather.arff
7. Right click on Arff Loader and select Dataset option then establish a link between Arff Loader and Attribute Selection.
8. Right click on Attribute Selection and select Dataset option then establish a link between Attribute Selection and Normalize.
9. Right click on Attribute Selection and select Configure option and choose the best attribute for Weather data.
10. Right click on Normalize and select Dataset option then establish a link between Normalize and Arff Saver.
11. Right click on Arff Saver and select Configure option then new window will be opened and set the path, enter .arff in look in dialog box to save normalize data.
12. Right click on Arff Loader and click on Start Loading option then everything will be executed one by one.
13. Check whether output is created or not by selecting the preferred path.
14. Rename the data name as a.arff
15. Double click on a.arff then automatically the output will be opened in MS-Excel.

Weka KnowledgeFlow Environment

Program File Edit Insert View

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

Design

DataSourcees
DataSinks
DataGenerators
Filters
Classifiers
Clusterers
Associations
AttributeSelection
Evaluation
Misc
Visualization
Flow
Tools

weatherJ4

```
graph LR; ArffLoader -- "data Set" --> ClassAssigner; ClassAssigner -- "data Set" --> CrossValidationFoldMaker; CrossValidationFoldMaker -- "test Set" --> J48; J48 -- "training Set" --> benchClassifier; benchClassifier -- "test" --> ClassifierPerformanceEvaluator; ClassifierPerformanceEvaluator -- "test" --> TextViewer;
```

Status Log

Component	Parameters	Time	Status
[KnowledgeFlow]		-	OK.
ArffLoader		-	Finished.
ClassAssigner		-	Finished.
CrossValidationFoldMaker		-	Finished.
J48	-C 0.25 -M 2	-	Finished.
ClassifierPerformanceEvaluator		-	Finished.
TextViewer		-	Finished.

Type here to search

15:44 13-05-2020

4.

AIM:

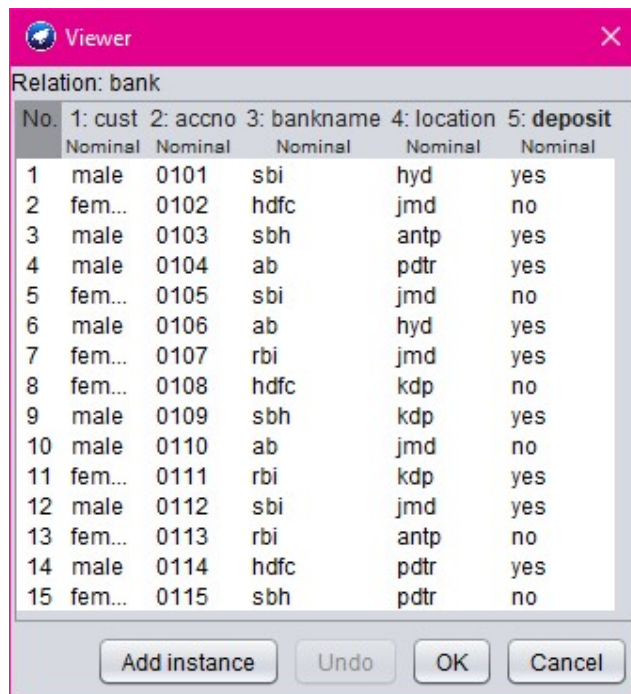
Finding Association Rules for Banking data.

DESCRIPTION:

In data mining, **association rule learning** is a popular and well researched method for discovering interesting relations between variables in large databases. It can be described as analyzing and presenting strong rules discovered in databases using different measures of interestingness. In market basket analysis association rules are used and they are also employed in many application areas including Web usage mining, intrusion detection and bioinformatics.

PROCEDURE:

1. Open Start ▢ Programs ▢ Weka-3-4 ▢ Weka-3-4
2. Open **explorer**.
3. Click on **open file** and select **banking.arff**



The screenshot shows the 'Viewer' window in Weka Explorer. The title bar is pink with a close button. The window displays the 'bank' relation data. The table has 5 columns: 'No.', '1: cust', '2: accno', '3: bankname', '4: location', and '5: deposit'. The data is as follows:

No.	1: cust	2: accno	3: bankname	4: location	5: deposit
	Nominal	Nominal	Nominal	Nominal	Nominal
1	male	0101	sbi	hyd	yes
2	fem...	0102	hdfc	jmd	no
3	male	0103	sbh	antp	yes
4	male	0104	ab	pdtr	yes
5	fem...	0105	sbi	jmd	no
6	male	0106	ab	hyd	yes
7	fem...	0107	rbi	jmd	yes
8	fem...	0108	hdfc	kdp	no
9	male	0109	sbh	kdp	yes
10	male	0110	ab	jmd	no
11	fem...	0111	rbi	kdp	yes
12	male	0112	sbi	jmd	yes
13	fem...	0113	rbi	antp	no
14	male	0114	hdfc	pdtr	yes
15	fem...	0115	sbh	pdtr	no

At the bottom of the window, there are four buttons: 'Add instance', 'Undo', 'OK', and 'Cancel'.

4. Select **Associate option** on the top of the Menu bar.
5. Select **Choose button** and then click on **Apriori Algorithm**.
6. Click on **Start button** and output will be displayed on the **right side** of the window.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click...)

15:51:20 - Apriori

Associator output

```
Apriori
=====

Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14
Size of set of large itemsets L(2): 24
Size of set of large itemsets L(3): 8

Best rules found:

1. bankname=ab 3 ==> cust=male 3    <conf:(1)> lift:(1.88) lev:(0.09) [1] conv:(1.4)
2. bankname=rbi 3 ==> cust=female 3  <conf:(1)> lift:(2.14) lev:(0.11) [1] conv:(1.6)
3. location=hyd 2 ==> cust=male 2    <conf:(1)> lift:(1.88) lev:(0.06) [0] conv:(0.93)
4. location=hyd 2 ==> deposit=yes 2  <conf:(1)> lift:(1.67) lev:(0.05) [0] conv:(0.8)
5. bankname=sbi deposit=yes 2 ==> cust=male 2  <conf:(1)> lift:(1.88) lev:(0.06) [0] conv:(0.93)
6. cust=male bankname=sbi 2 ==> deposit=yes 2  <conf:(1)> lift:(1.67) lev:(0.05) [0] conv:(0.8)
7. bankname=sbh deposit=yes 2 ==> cust=male 2  <conf:(1)> lift:(1.88) lev:(0.06) [0] conv:(0.93)
8. cust=male bankname=sbh 2 ==> deposit=yes 2  <conf:(1)> lift:(1.67) lev:(0.05) [0] conv:(0.8)
9. bankname=ab deposit=yes 2 ==> cust=male 2  <conf:(1)> lift:(1.88) lev:(0.06) [0] conv:(0.93)
10. location=hyd deposit=yes 2 ==> cust=male 2  <conf:(1)> lift:(1.88) lev:(0.06) [0] conv:(0.93)
```

Status

Type here to search

5.

AIM:

To Construct Decision Tree for Weather data and classify it.

DESCRIPTION:

Classification & Prediction:

Classification is the process for finding a model that describes the data values and concepts for the purpose of Prediction.

Decision Tree:

A decision Tree is a classification scheme to generate a tree consisting of root node, internal nodes and external nodes.

Root nodes representing the attributes. Internal nodes are also the attributes. External nodes are the classes and each branch represents the values of the attributes

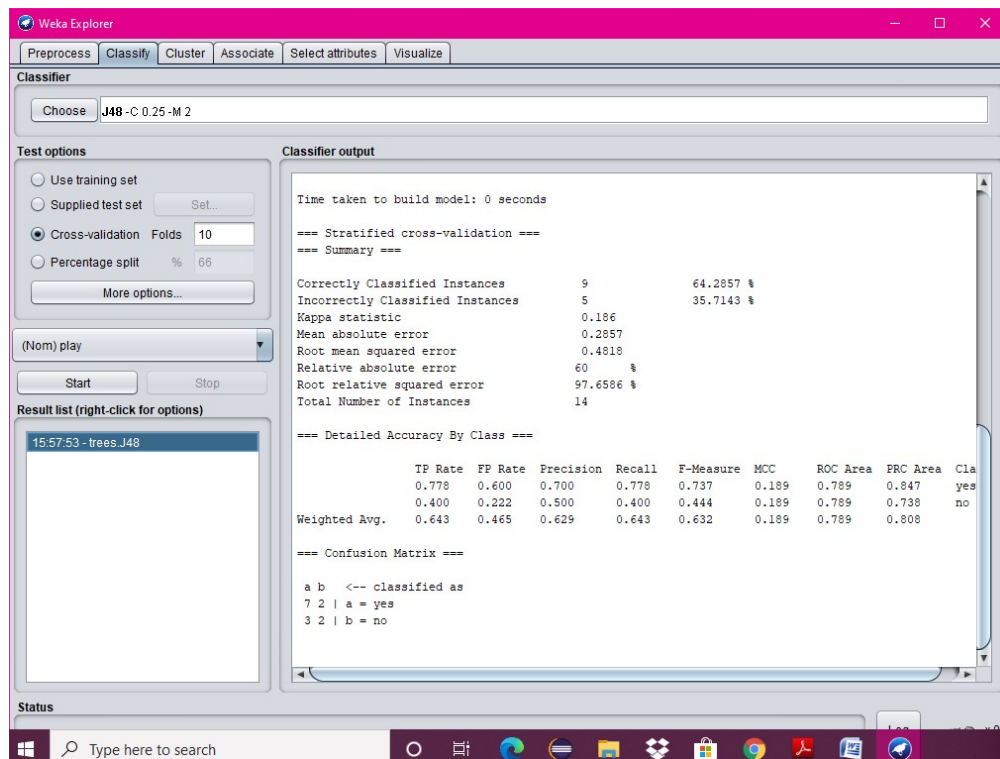
Decision Tree also contains set of rules for a given data set; there are two subsets in Decision Tree. One is a Training data set and second one is a Testing data set. Training data set is previously classified data. Testing data set is newly generated data

PROCEDURE:

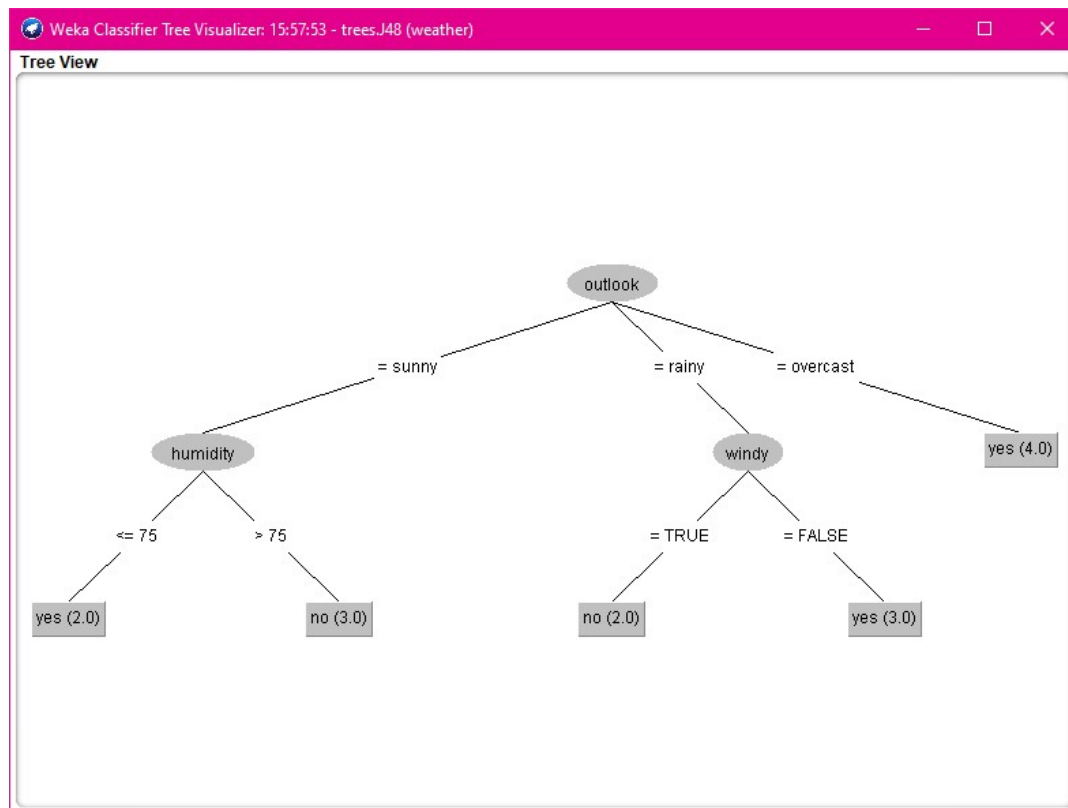
1. Open Start -> Programs -> Weka-3-4 -> Weka-3-4
2. Open **explorer**.
3. Click on **open file** and select **weather.arff**

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

4. Select **Classifier** option on the top of the Menu bar.
5. Select **Choose** button and click on **Tree** option.
6. Click on **J48**.
7. Click on **Start** button and output will be displayed on the **right side** of the window.



8. Select the **result list** and **right click** on result list and select **Visualize Tree** option.
9. Then **Decision Tree** will be displayed on **new window**.



6.

AIM:

Write a procedure for cross-validation using J48 Algorithm for weather table.

DESCRIPTION:

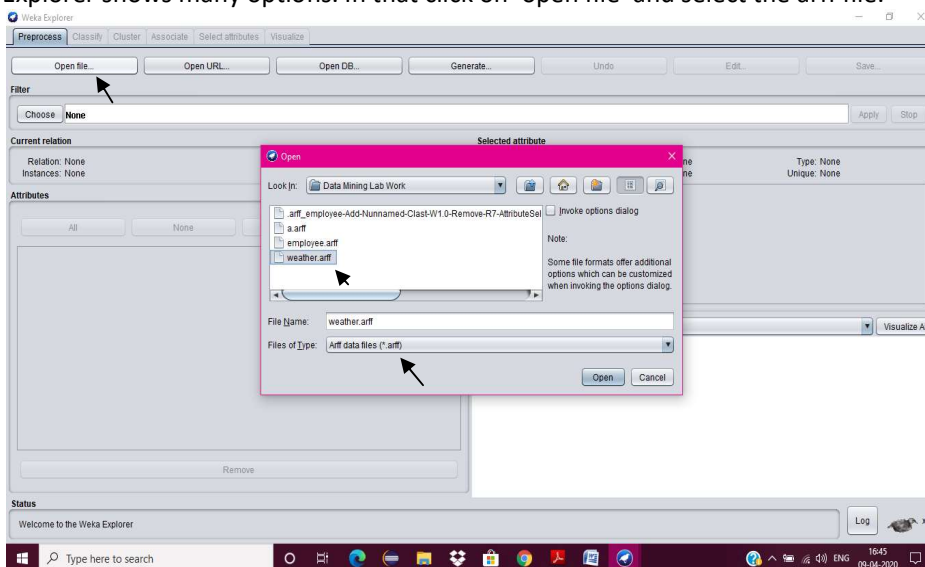
Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

PROCEDURE:

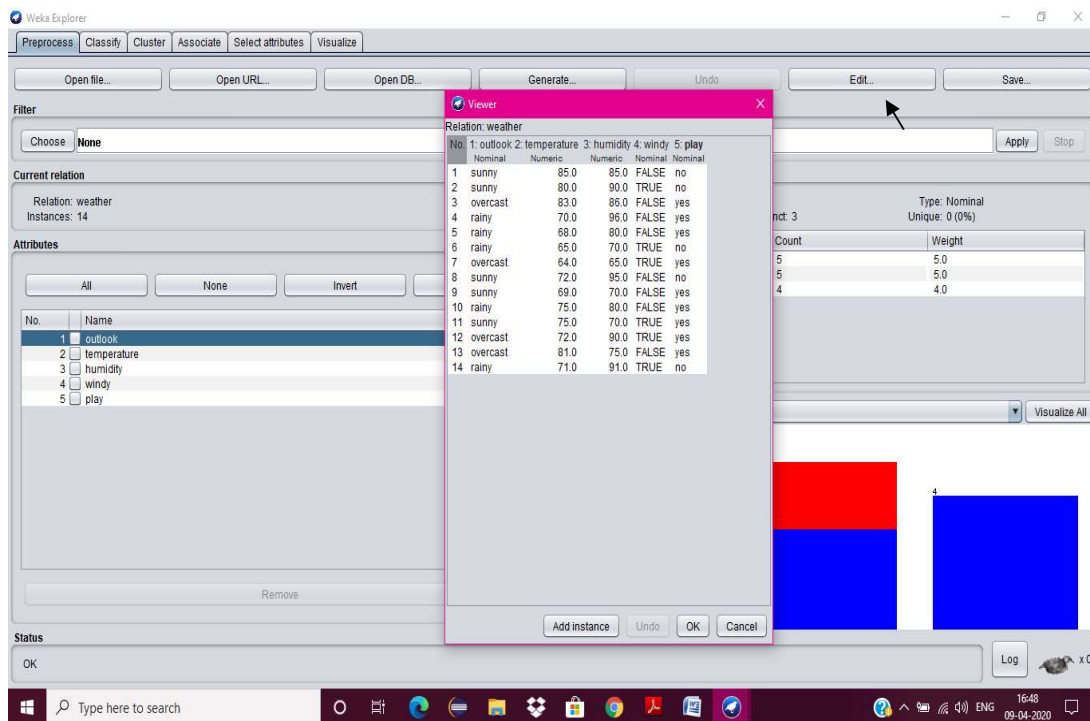
1. Create or download **weather.arff** file having weather data.
2. Click on weka and then click on explorer.



3. Explorer shows many options. In that click on 'open file' and select the arff file.



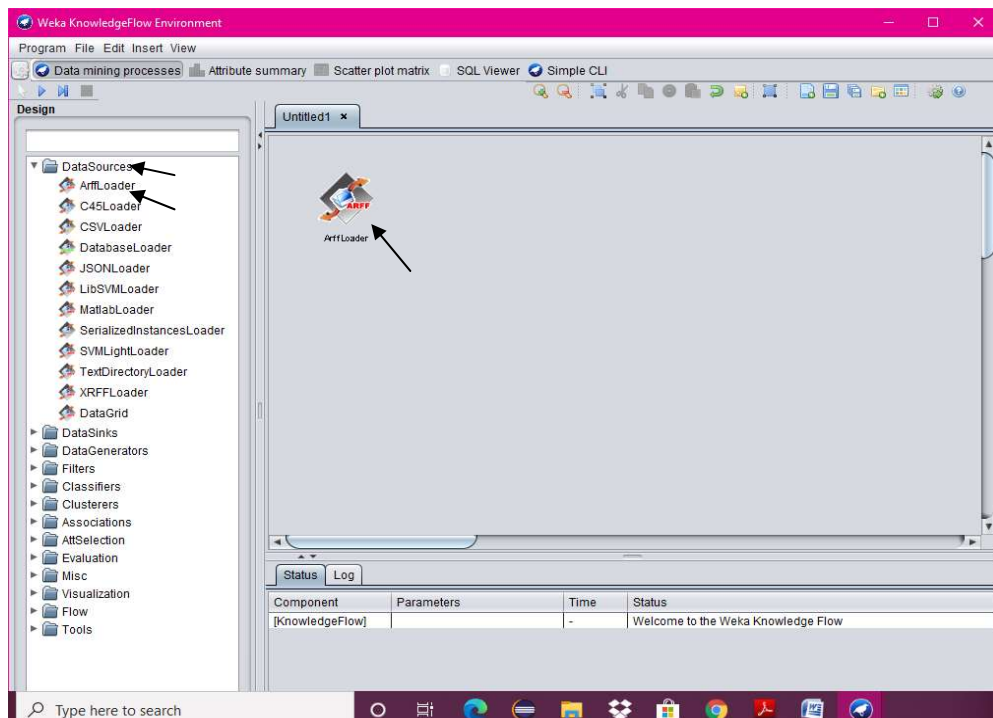
4. Click on edit button which shows weather table on weka.



5. Close Weka Explorer and then select KnowledgeFlow from GUI.

6. Select Data Source tab & choose Arff Loader.

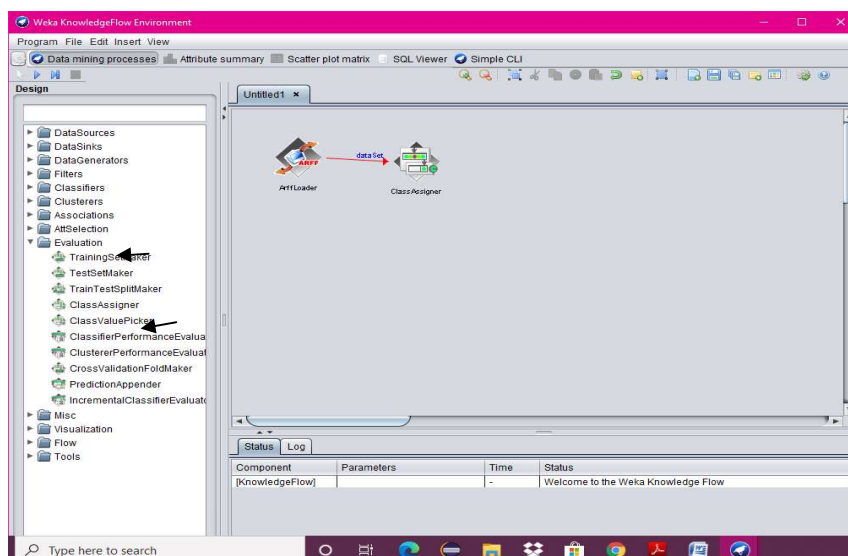
7. Place Arff Loader component on the layout area by clicking on that component.



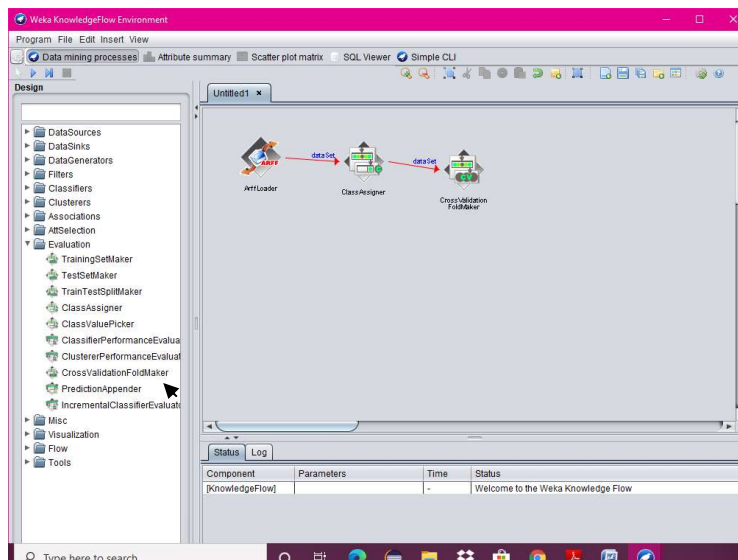
8. Specify an Arff file to load by right clicking on Arff Loader icon, and then a pop-up menu will appear. In that select Configure & browse to the location of weather.arff



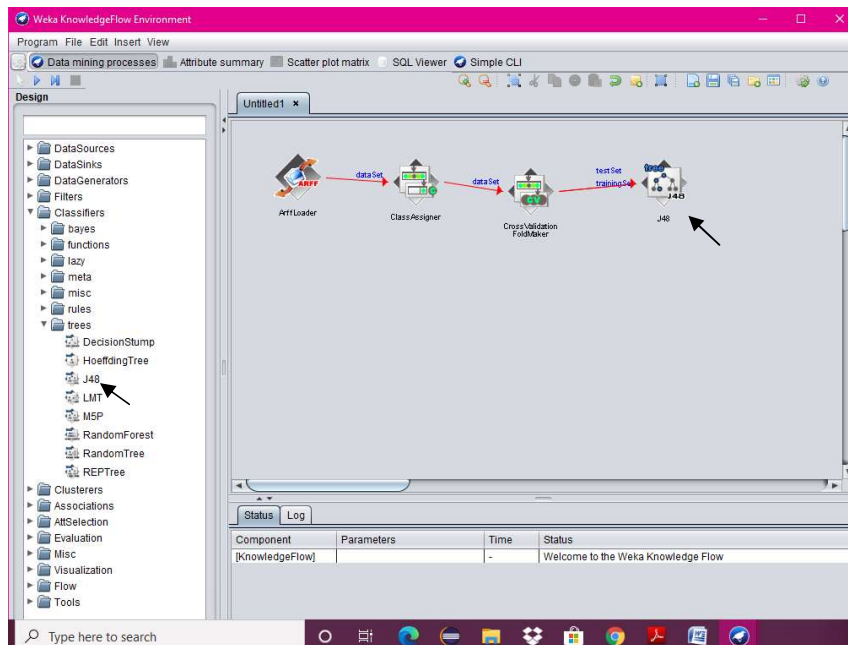
9. Click on the Evaluation tab & choose Class Assigner & place it on the layout.
10. Now connect the Arff Loader to the Class Assigner by right clicking on Arff Loader, and then select Data Set option, now a link will be established.



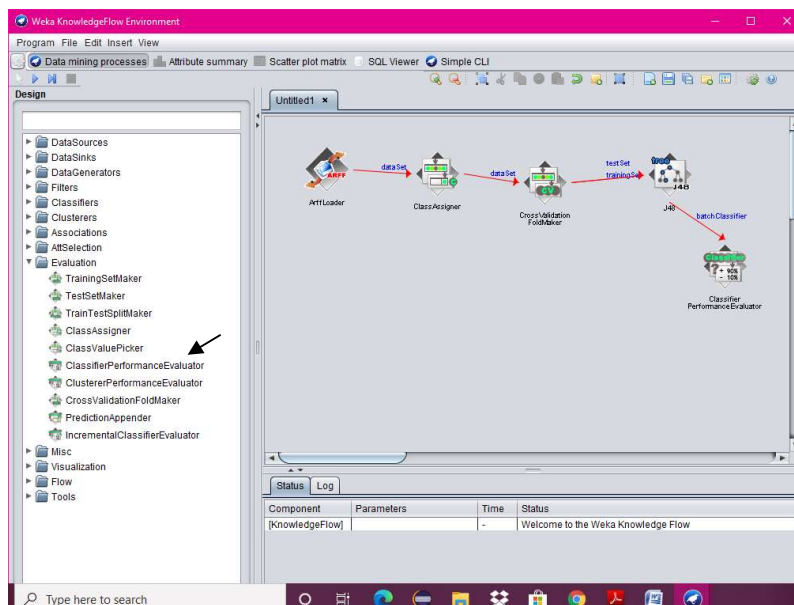
11. Right click on Class Assigner & choose Configure option, and then a new window will appear & specify a class to our data.
12. Select Evaluation tab & select Cross-Validation Fold Maker & place it on the layout.
13. Now connect the Class Assigner to the Cross-Validation Fold Maker.



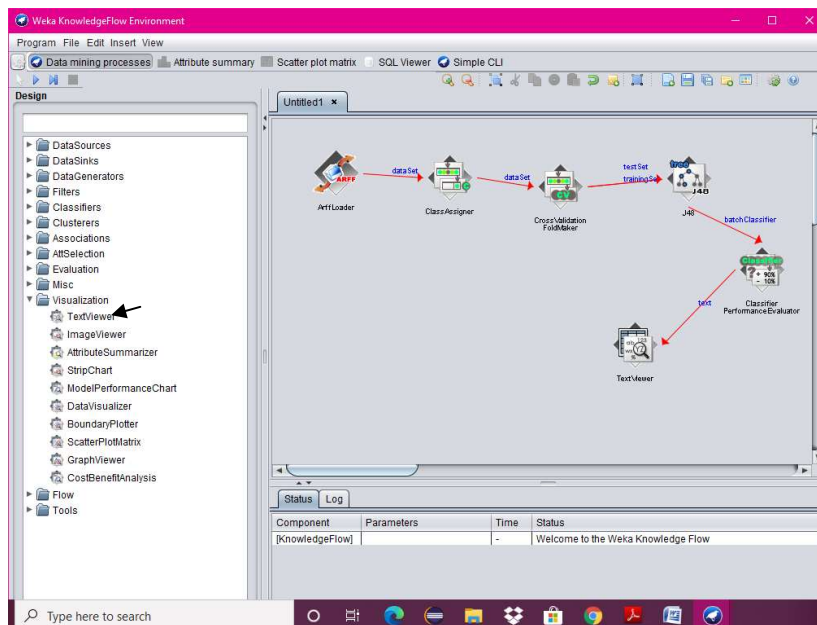
14. Select Classifiers tab & select J48 component & place it on the layout.
15. Now connect Cross-Validation Fold Maker to J48 twice; first choose Training Data Set option and then Test Data Set option.



16. Select Evaluation Tab & select Classifier Performance Evaluator component & place it on the layout.
17. Connect J48 to Classifier Performance Evaluator component by right clicking on J48 & selecting Batch Classifier.

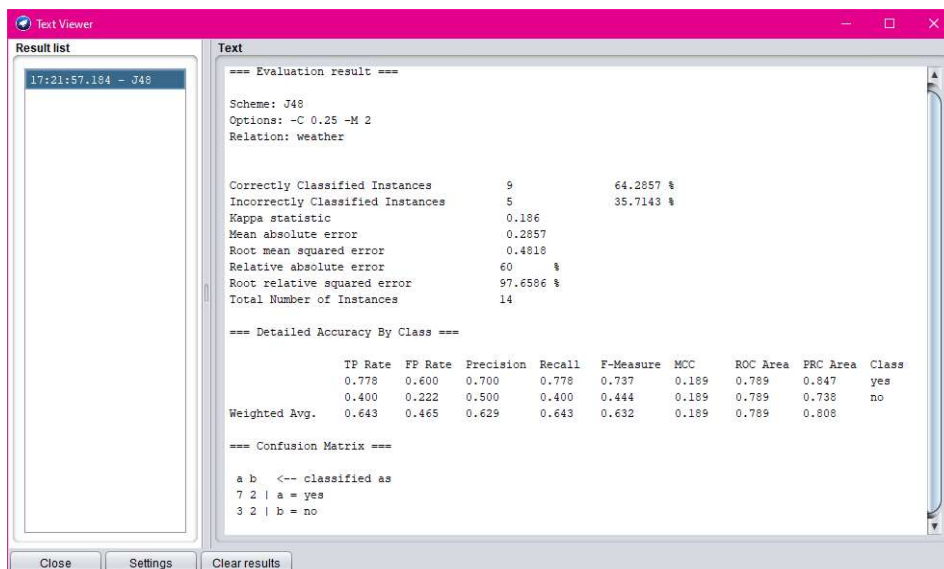


18. Select Visualization tab & select Text Viewer component & place it on the layout.
19. Connect Classifier Performance Evaluator to Text Viewer by right clicking on Classifier Performance Evaluator & by selecting Text option.



20. Start the flow of execution by selecting Start Loading from Arff Loader.

21. For viewing result, right click on Text Viewer & select the Show Results, and then the result will be displayed on the new window.



7.

AIM:

Write a procedure for Clustering Buying data using Cobweb Algorithm.

DESCRIPTION:

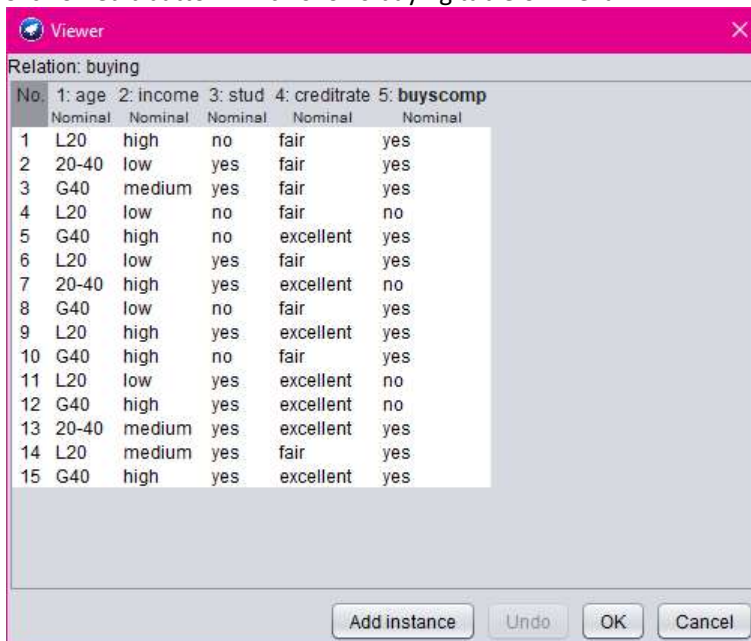
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

PROCEDURE:

8. Create or download **buying.arff** file having buying data.
9. Click on weka and then click on explorer.

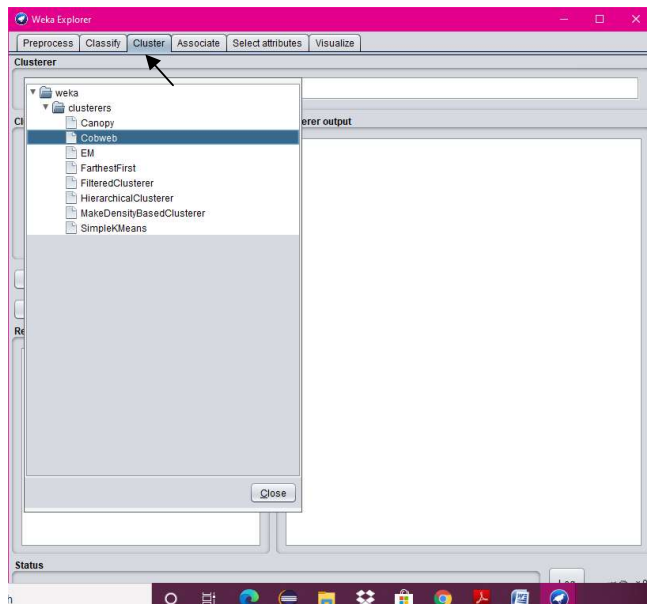


10. Explorer shows many options. In that click on 'open file' and select the arff file.
11. Click on edit button which shows buying table on weka.

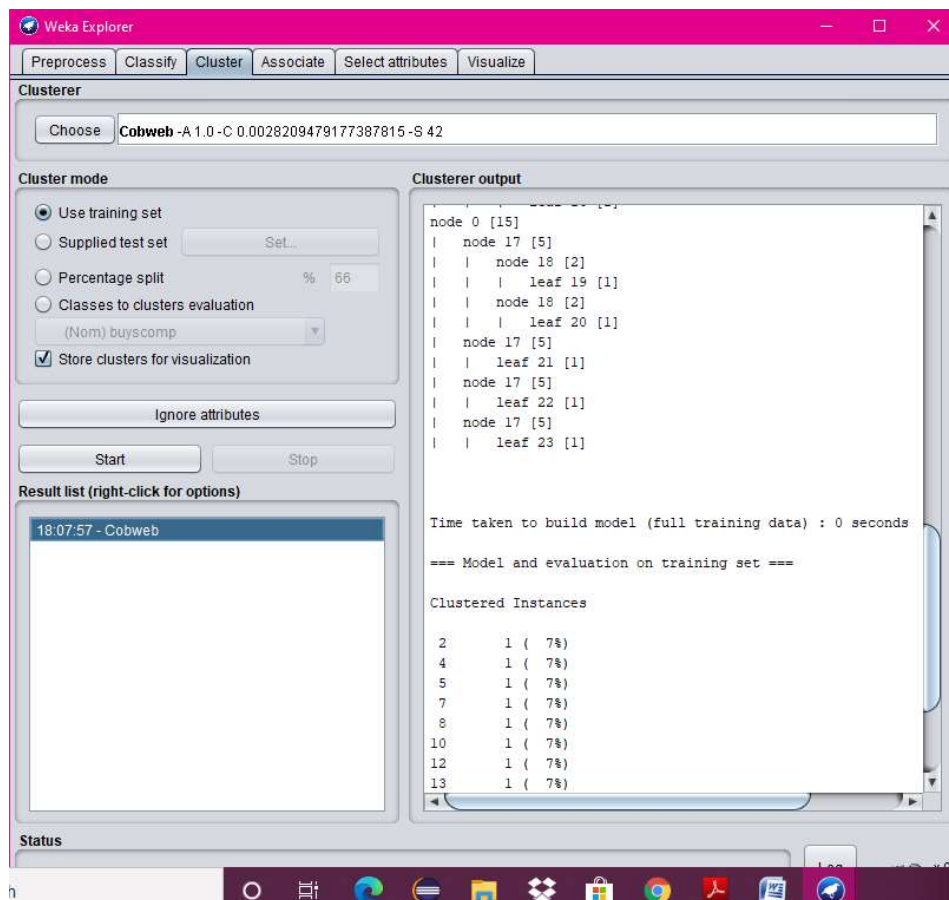


No.	1: age	2: income	3: stud	4: creditrate	5: buyscomp
	Nominal	Nominal	Nominal	Nominal	Nominal
1	L20	high	no	fair	yes
2	20-40	low	yes	fair	yes
3	G40	medium	yes	fair	yes
4	L20	low	no	fair	no
5	G40	high	no	excellent	yes
6	L20	low	yes	fair	yes
7	20-40	high	yes	excellent	no
8	G40	low	no	fair	yes
9	L20	high	yes	excellent	yes
10	G40	high	no	fair	yes
11	L20	low	yes	excellent	no
12	G40	high	yes	excellent	no
13	20-40	medium	yes	excellent	yes
14	L20	medium	yes	fair	yes
15	G40	high	yes	excellent	yes

12. Close the file.
13. Click on Cluster menu. In this there are different algorithms are there.
14. Click on Choose button and then select cobweb algorithm.



15. Click on Start button and then output will be displayed on the screen.



8.

AIM:

Write a procedure for Clustering Weather data using EM Algorithm.

DESCRIPTION:

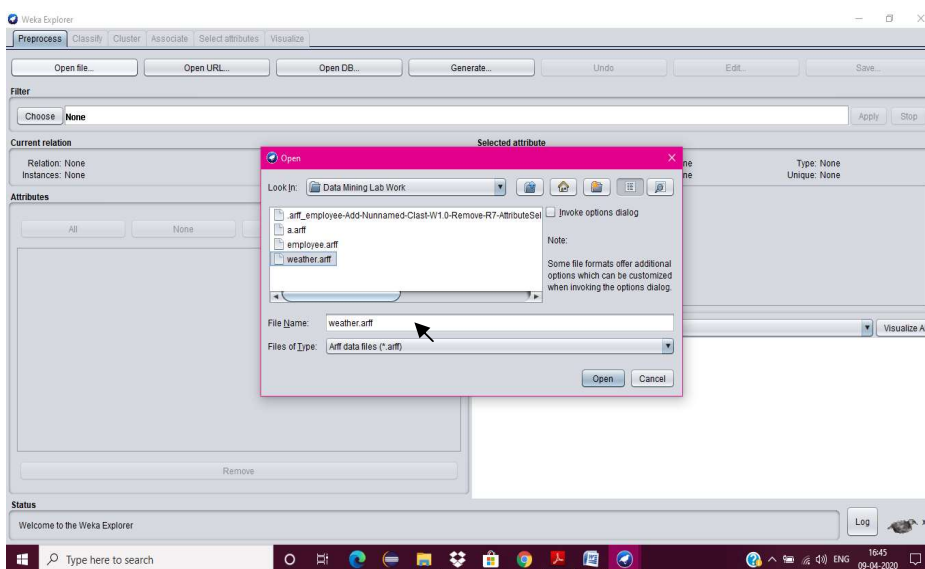
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

PROCEDURE:

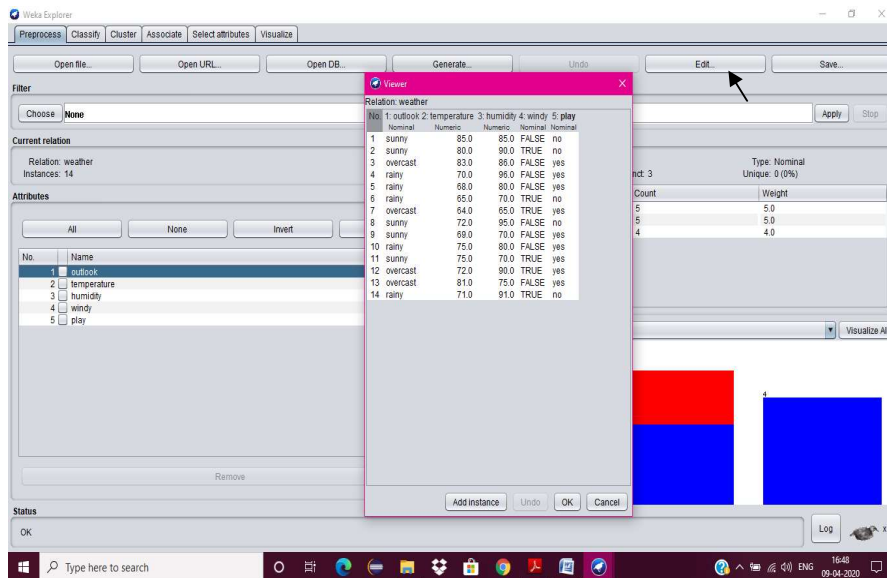
16. Create or download **weather.arff** file having weather data.
17. Click on weka and then click on explorer.



18. Explorer shows many options. In that click on 'open file' and select the arff file.



19. Click on edit button which shows weather table on weka.

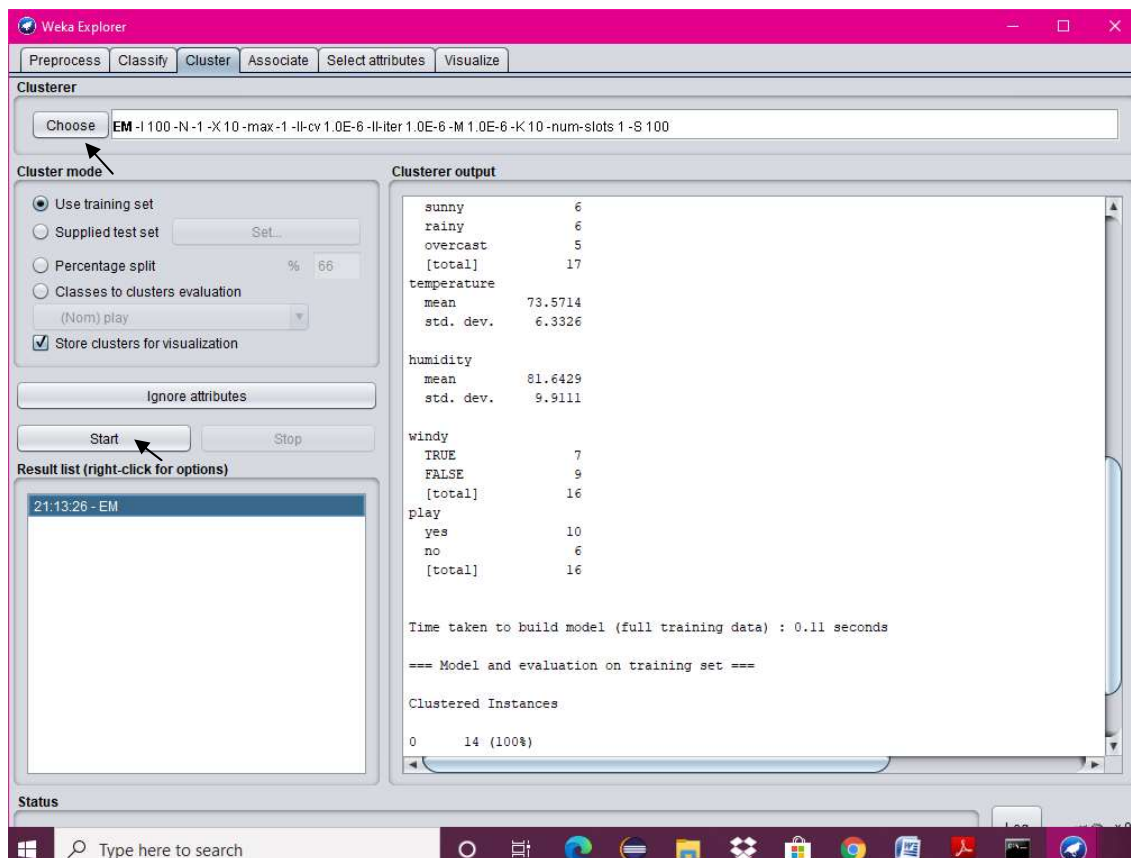


20. Close the file.

21. Click on Cluster menu. In this there are different algorithms are there.

22. Click on Choose button and then select EM algorithm.

23. Click on Start button and then output will be displayed on the screen.



9.

AIM:

Write a procedure for Clustering Banking data using the farthest first Algorithm.

DESCRIPTION:

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

PROCEDURE:

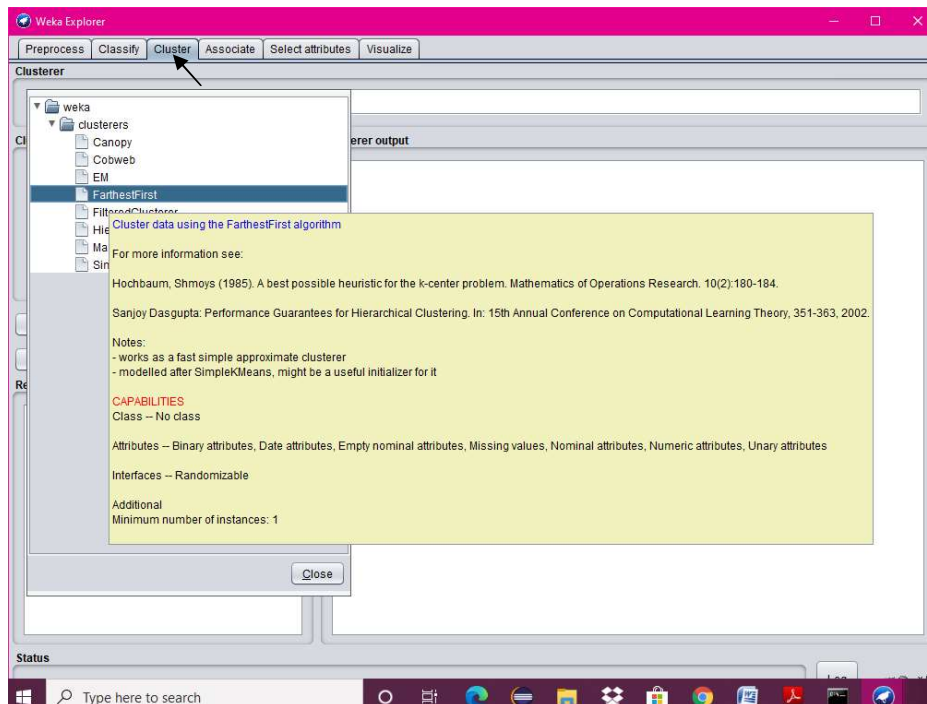
1. Create or download **banking.arff** file having banking data.
2. Click on weka and then click on explorer.



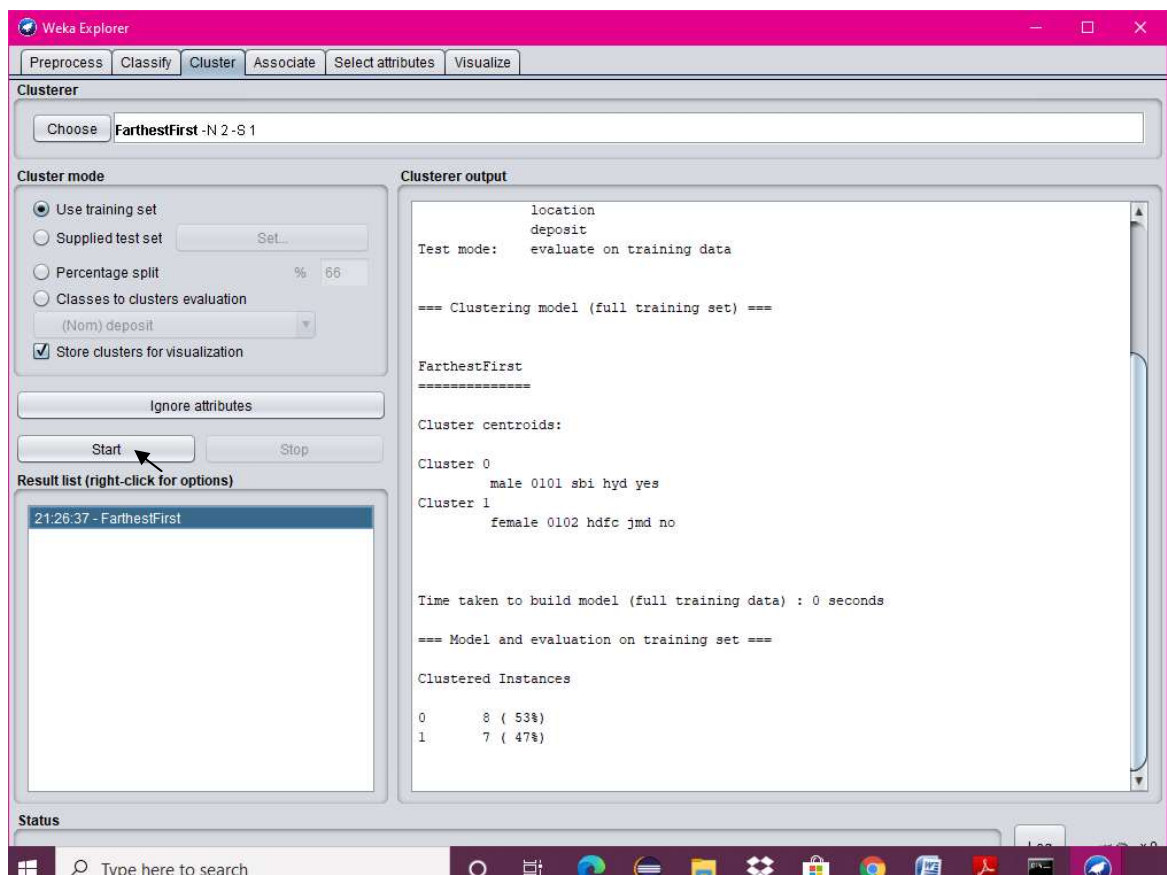
3. Explorer shows many options. In that click on 'open file' and select the arff file.
4. Click on edit button which shows banking table on weka.

No.	1: cust	2: accno	3: bankname	4: location	5: deposit
	Nominal	Nominal	Nominal	Nominal	Nominal
1	male	0101	sbi	hyd	yes
2	fem...	0102	hdfc	jmd	no
3	male	0103	sbh	antp	yes
4	male	0104	ab	pdtr	yes
5	fem...	0105	sbi	jmd	no
6	male	0106	ab	hyd	yes
7	fem...	0107	rbi	jmd	yes
8	fem...	0108	hdfc	kdp	no
9	male	0109	sbh	kdp	yes
10	male	0110	ab	jmd	no
11	fem...	0111	rbi	kdp	yes
12	male	0112	sbi	jmd	yes
13	fem...	0113	rbi	antp	no
14	male	0114	hdfc	pdtr	yes
15	fem...	0115	sbh	pdtr	no

5. Close the file.
6. Click on Cluster menu. In this there are different algorithms are there.
7. Click on Choose button and then select **FarthestFirst** algorithm.



8. Click on Start button and then output will be displayed on the screen.



10.

AIM:

Write a procedure for Employee data using Make Density Based Cluster Algorithm.

DESCRIPTION:

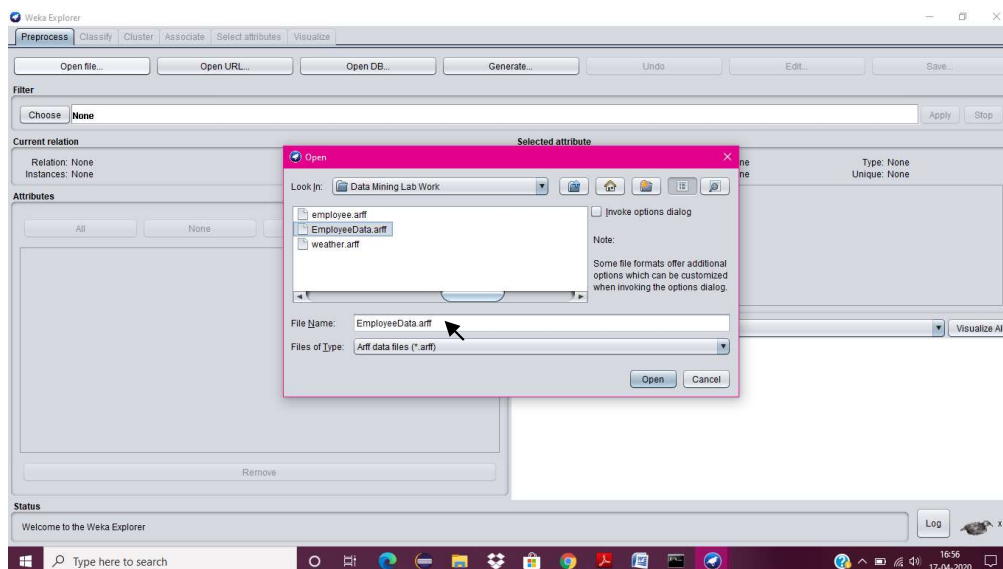
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

PROCEDURE:

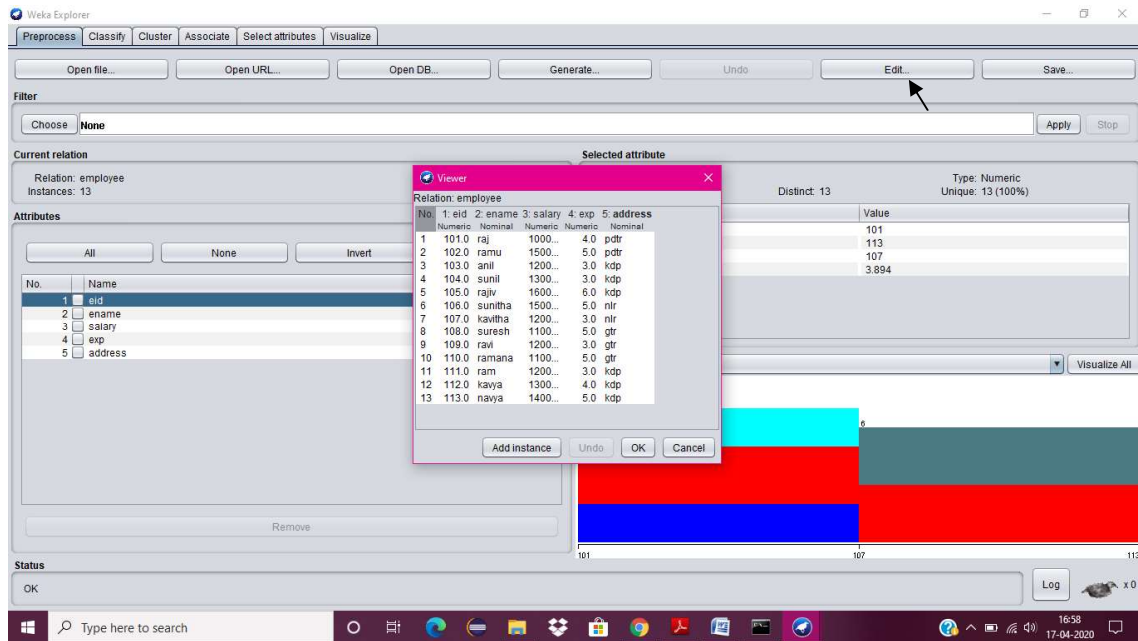
24. Create or download **EmployeeData.arff** file having weather data.
25. Click on weka and then click on explorer.



26. Explorer shows many options. In that click on 'open file' and select the arff file.



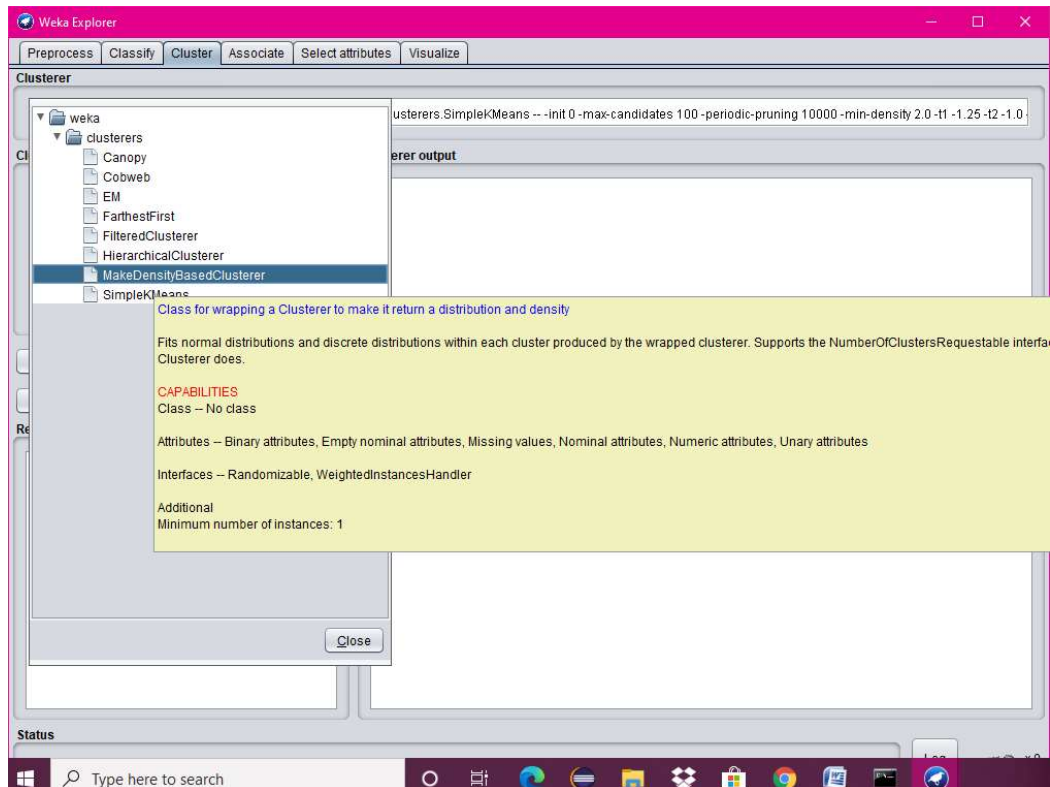
27. Click on edit button which shows employee table on weka.



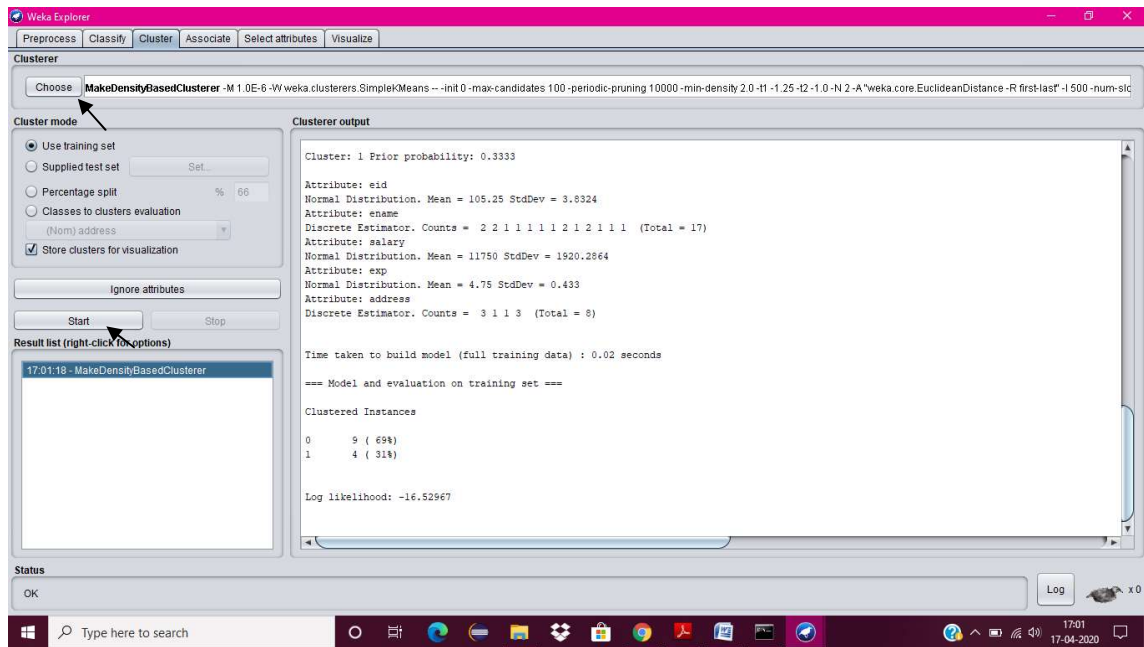
28. Close the file.

29. Click on Cluster menu. In this there are different algorithms are there.

30. Click on Choose button and then select Make Density Based Cluster Algorithm.



31. Click on Start button and then output will be displayed on the screen.



11.

AIM:

Write a procedure for Clustering Customer data using Simple KMeans Algorithm.

DESCRIPTION:

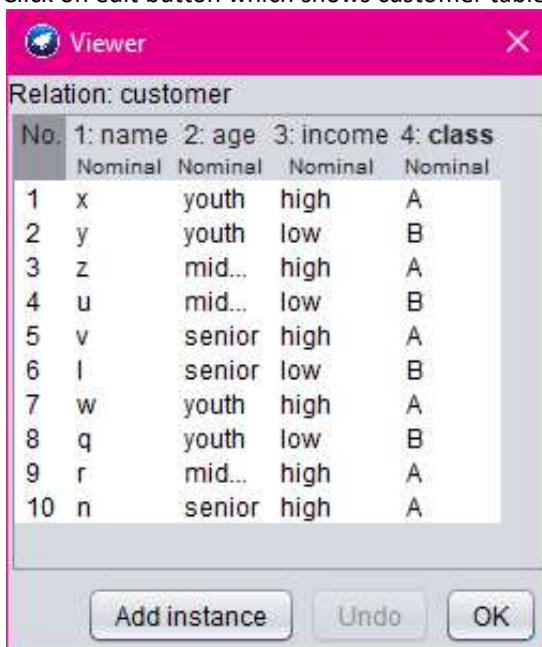
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

PROCEDURE:

9. Create or download **customer.arff** file having banking data.
10. Click on weka and then click on explorer.

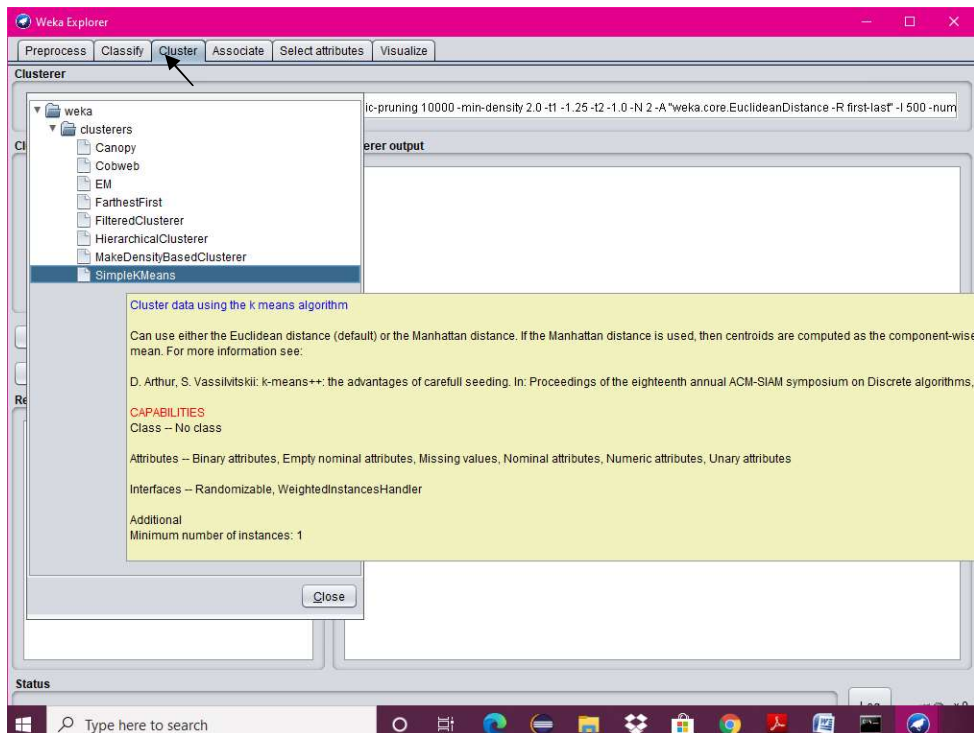


11. Explorer shows many options. In that click on 'open file' and select the arff file.
12. Click on edit button which shows customer table on weka.

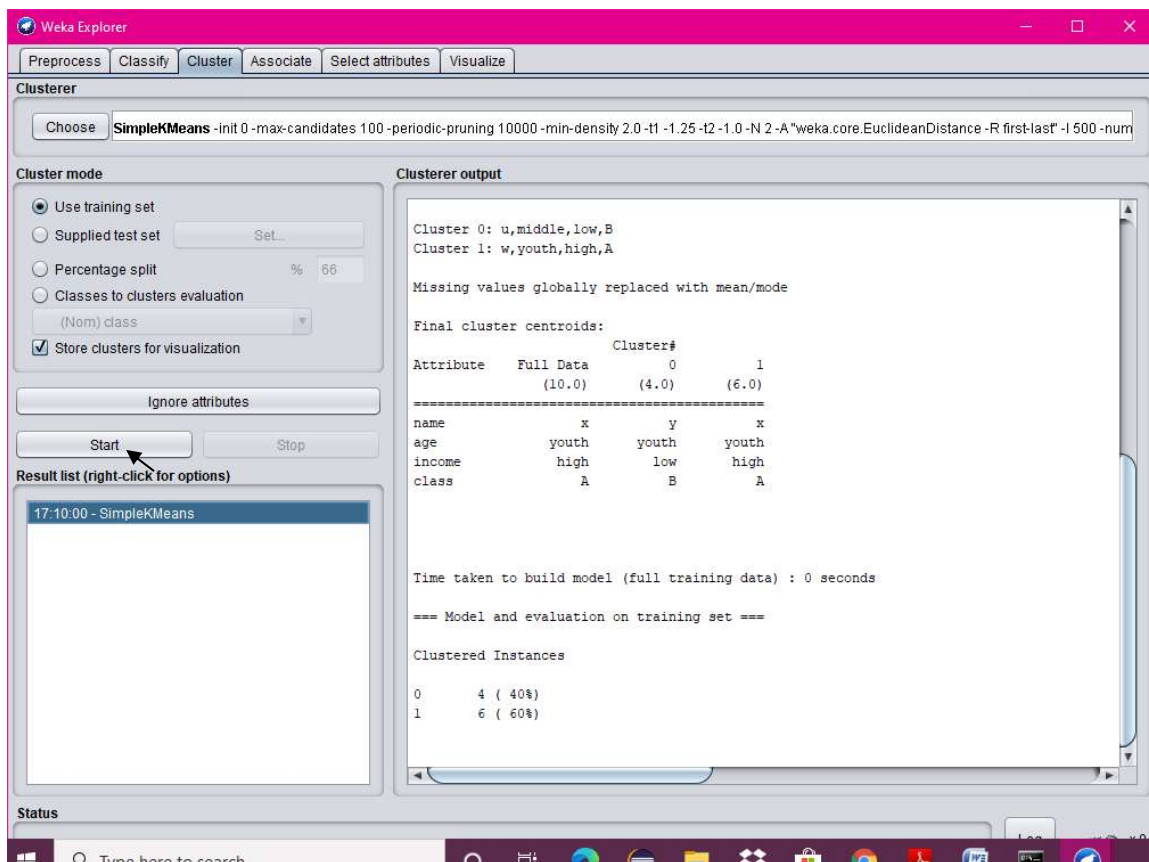


No.	1: name	2: age	3: income	4: class
	Nominal	Nominal	Nominal	Nominal
1	x	youth	high	A
2	y	youth	low	B
3	z	mid...	high	A
4	u	mid...	low	B
5	v	senior	high	A
6	l	senior	low	B
7	w	youth	high	A
8	q	youth	low	B
9	r	mid...	high	A
10	n	senior	high	A

13. Close the file.
14. Click on Cluster menu. In this there are different algorithms are there.
15. Click on Choose button and then select **SimpleKMeans** algorithm.



16. Click on Start button and then output will be displayed on the screen.



12.

AIM:

Write a procedure for Visualization for Weather Table.

DESCRIPTION:

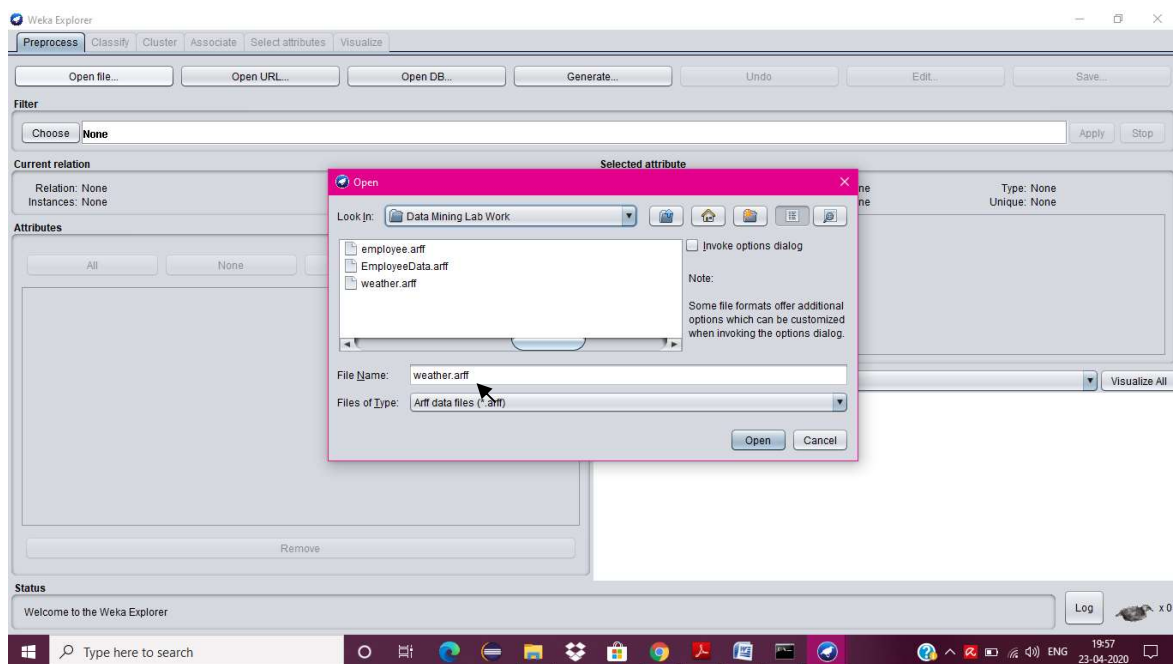
This program calculates and has comparisons on the data set selection of attributes and methods of manipulations have been chosen. The Visualization can be shown in a 2-D representation of the information.

PROCEDURE:

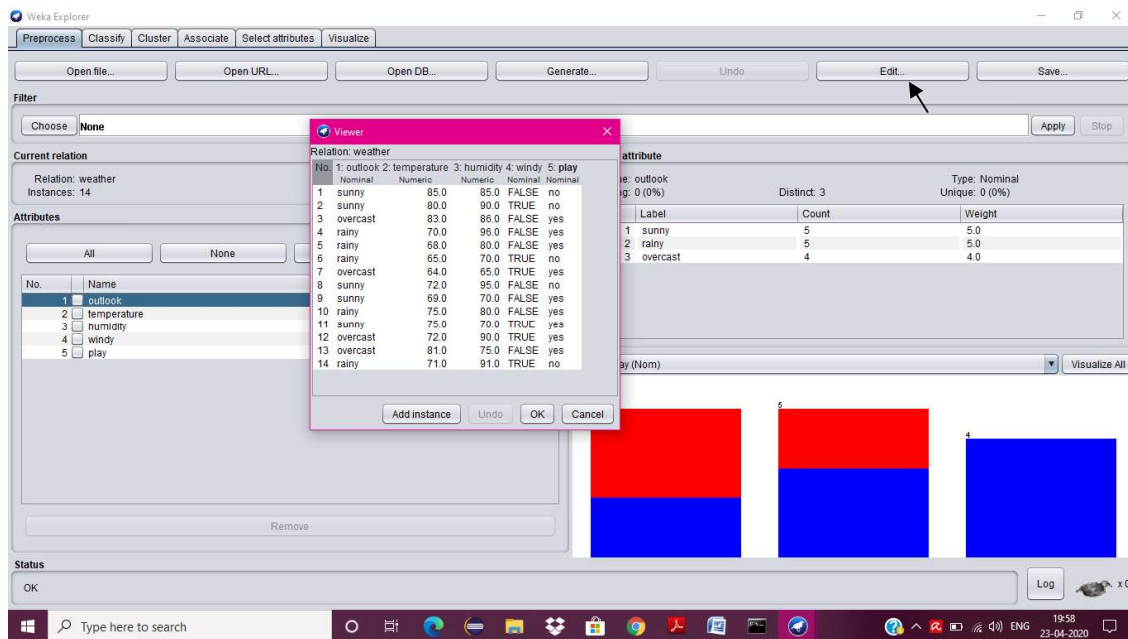
1. Create or download **Weather.arff** file having weather data.
2. Click on weka and then click on explorer.



3. Explorer shows many options. In that click on 'open file' and select the arff file.

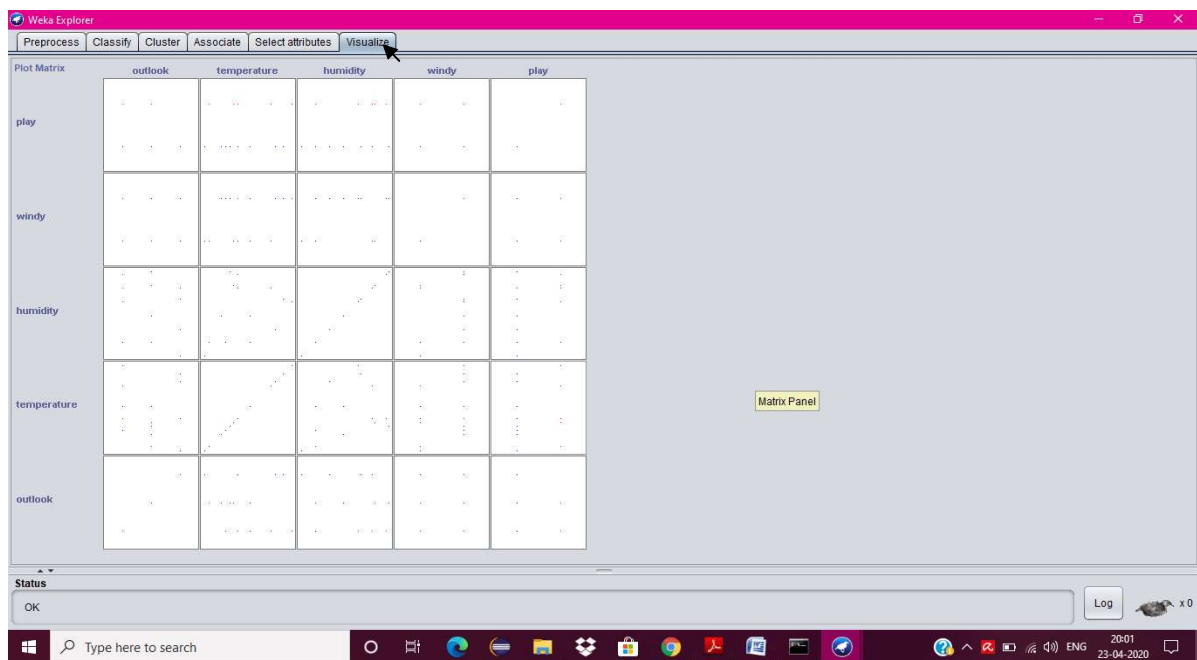


- Click on edit button which shows weather table on weka.



- Close the file.

- Now click on Visualize tab.

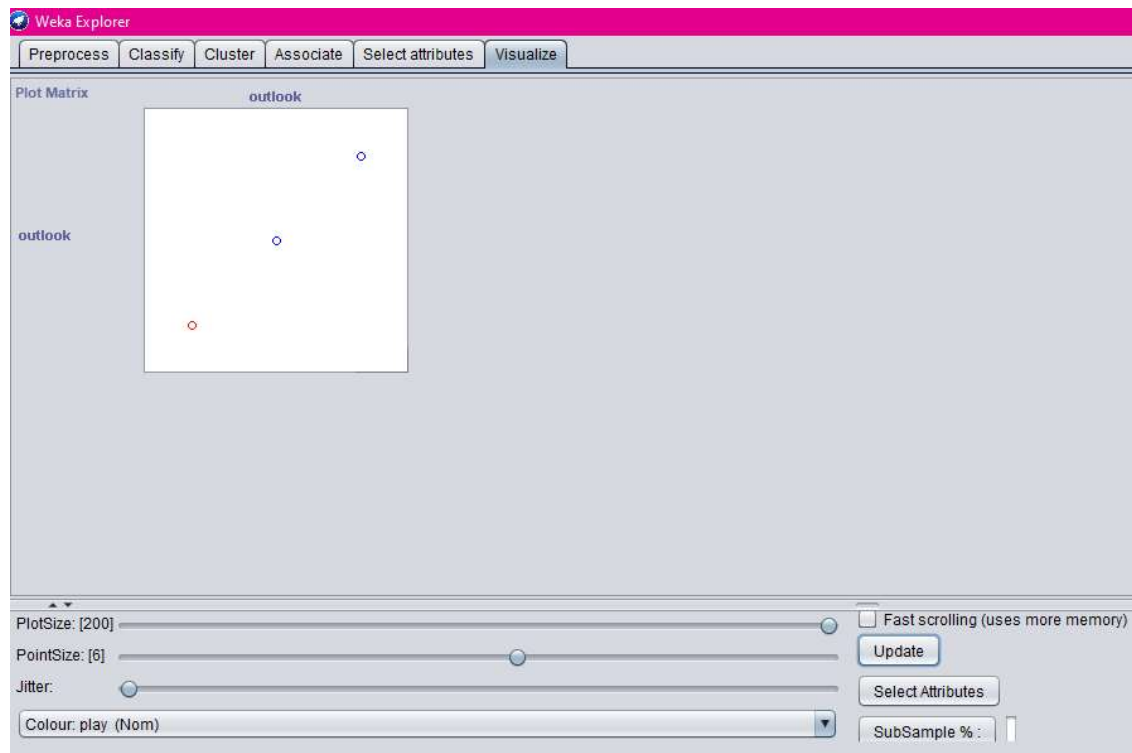


- Now visualize each attribute one by one.

- Select the **Select Attribute** button, then select **Outlook attribute** and click OK.

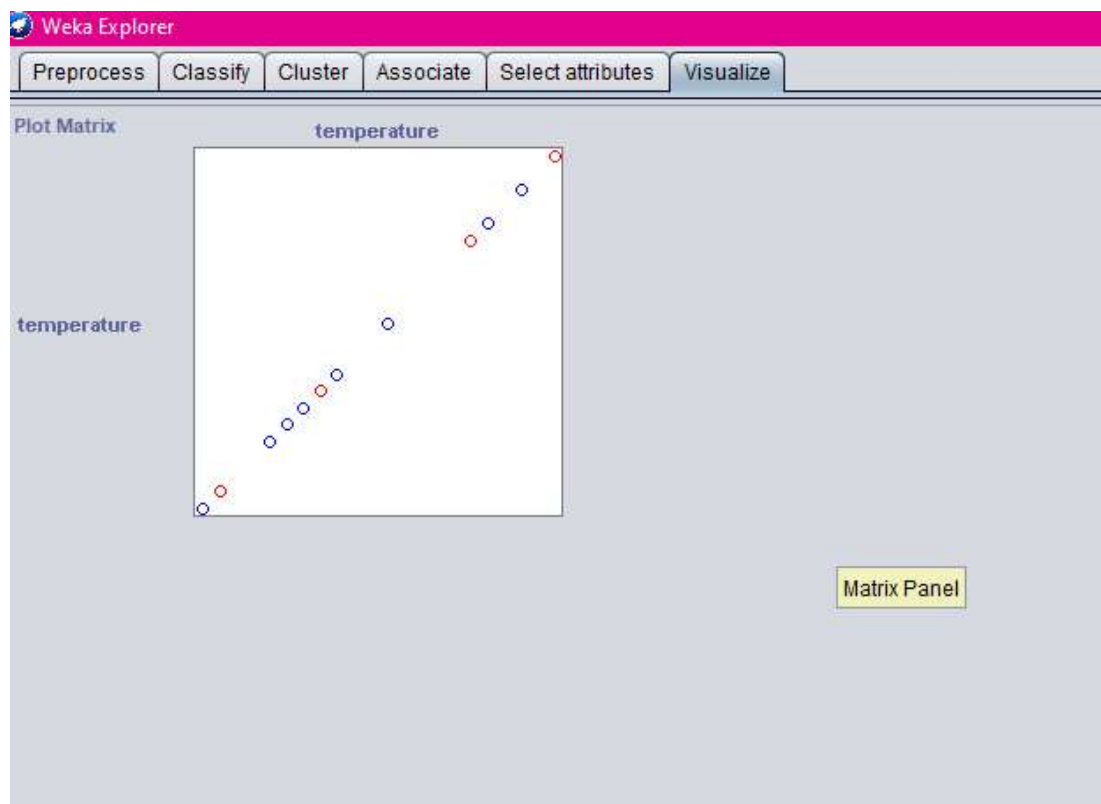
- Click on the update button to display the output.

- Increase the point size and plot size.

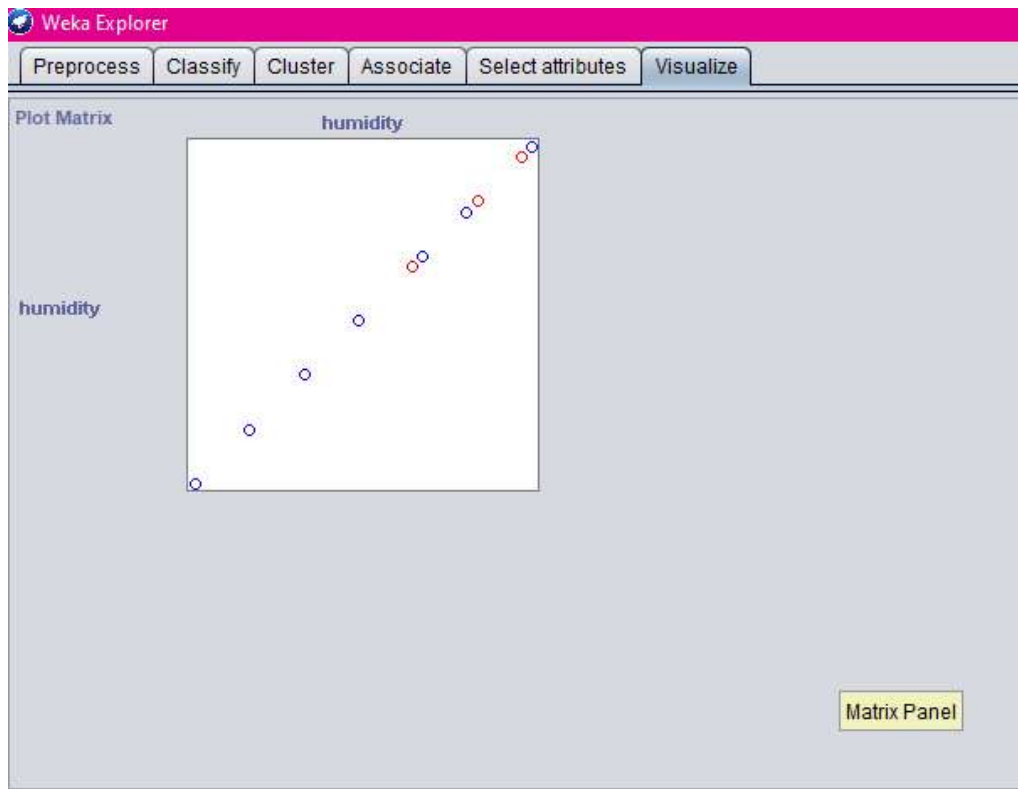


11. Repeat step 8 and 9 for other attributes also.

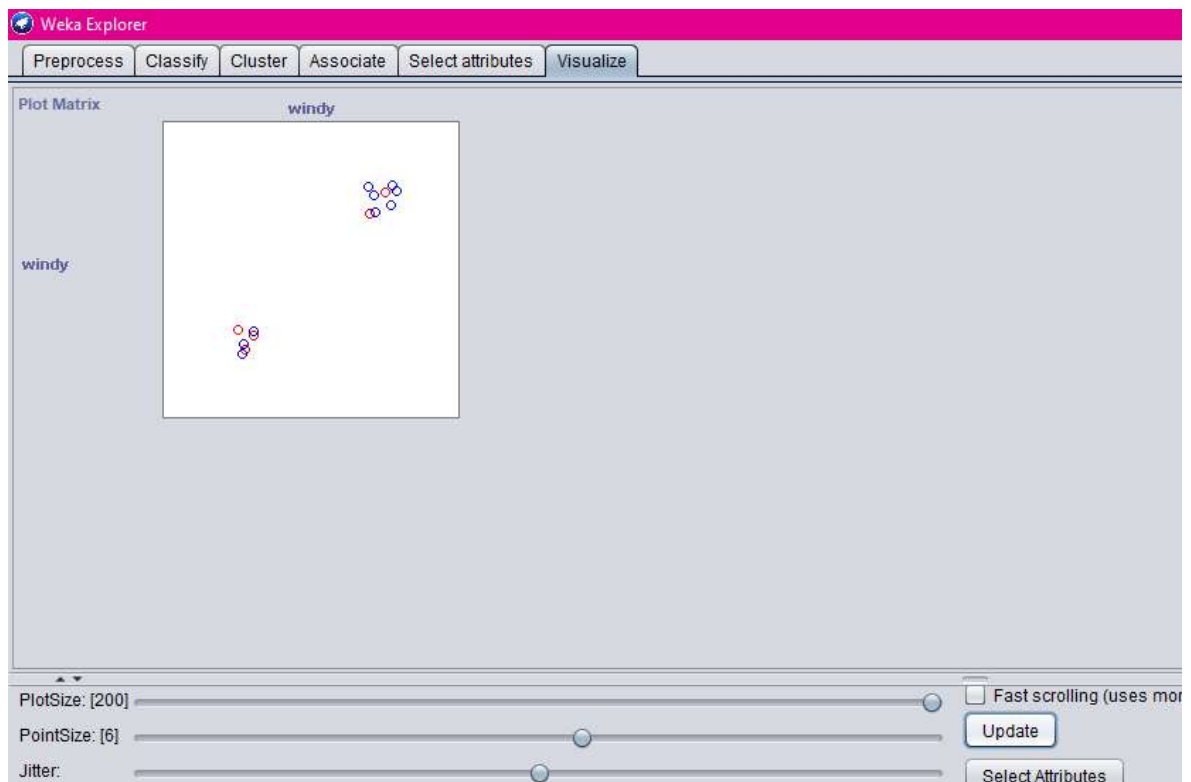
12. Temperature



13. Humidity



14. Windy



15. Play

