

COMSATS UNIVERSITY ISLAMABAD

LAHORE CAMPUS.

**WEKA MANUAL**

DATA MINING

Supervisor:

**Usman Shahzaib**

**C block.**

|  |  |
| --- | --- |
| Aasim Bashir | FA16-BCS-036 |
| Muneeb Ahmed | **FA16-BCS-042** |
| Zarrar Ahmad | **FA16-BCS-030** |

Table of Contents

Title page1

The patient arff file 3

Basic load file into WEKA5

Data Pre-Processing5-12

Knowledge Flow13-14

Association Rule15

Decision tree16-17

Visualization18-22

Clustering 22-31

**The representation of arff file.**

**@relation patient**

**@attribute name {1,2,3,4,5,6,7}**

**@attribute** Scan Advised String

**@attribute** City String

**@attribute** Age/Gender numeric/String

**@attribute** Patient Type String {Inpatient, Emergency, Outpatient}

**@attribute** History String

**@attribute** Contrast/Plain String {Plain, Contrast}

**@attribute** Report Findings String {Abnormal, Normal}

**@data**

Head, SHEIKHUPURA,70Y/M, Inpatient, CVA symptoms,Plain,Abnormal

Head, LAYYAH,50Y/M,Emergency,,CVA symptoms,Plain,Abnormal

Head, KHUSHAB,51Y/M,Emergency,,CVA symptoms,Plain,Normal

Head, KHUSHAB,57Y/F,Emergency,,CVA symptoms,Plain,Normal

Head, VEHARI,0Y/M,Emergency,,CVA symptoms,Plain,Abnormal

Head, SHEIKHUPURA,65Y/F,Emergency,,CVA symptoms,Plain,Abnormal

Head, SHEIKHUPURA,45Y/M,Emergency,,CVA symptoms,Plain,Abnormal

Head, LAYYAH,25Y/F,Emergency,,CVA symptoms,Plain,Abnormal

Head, MUZAFARGARH,75Y/M,Emergency,,SOB/ Dyspnea,Plain,Abnormal

Head, MUZAFARGARH,48Y/M,Emergency,,Trauma,Plain,Abnormal

Cervical Spine, SHEIKHUPURA,55Y/M,Inpatient,,Trauma,Plain,Abnormal

Chest, OKARA,90Y/M,Outpatient,,Chest Pain,Plain,Abnormal

Head, MUZAFARGARH,55Y/M,Emergency,,Trauma,Plain,Abnormal

Head, JHELUM,10Y/F,Emergency,,Trauma,Plain,Normal

Head, MUZAFARGARH,70Y/M,Emergency,,CVA symptoms,Plain,Abnormal

Head, MUZAFARGARH,60Y/M,Emergency,,Seizure,Plain,Abnormal

Head, MUZAFARGARH,40Y/M,Emergency,,Trauma,Plain,Abnormal

**EXPERIMENT NO:1**

**Aim:**

Create a Patient Table with the help of Data Mining Tool WEKA

**Description:**

We need to create a Patient Table with training data set which includes attributes like

Scan Advised, City, Age/Gender, Patient Type, History, Contrast/Plain Report, Findings.

* Start > Programs > weka3-9.
* Click on weka3-9, then Weka dialog box is displayed on the screen.
* In that dialog box there are four modes, click on explorer.
* Explorer shows many options. In that click on ‘open file’ and select the csv from file of type and select the csv file from there.
* The attributes will appear and select the desire attributes.

A screenshot of a social media post

Description automatically generated

* And then click the edit button on right top

A screenshot of a cell phone

Description automatically generated

Result: This program has been successfully executed.

**EXPERIMENT NO:2**

**Aim:**

Apply Pre-Processing techniques to the training data set of Patient 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 and also increase the speed of computation.

**Do the same steps of experiment1**

There are 3 pre-processing techniques they are:

* Add
* Remove
* Normalization
* Binning
* Replace missing value
* Nominal to binary
* Ordinal to numeric
* Principle component

Add > Pre-Processing Technique:

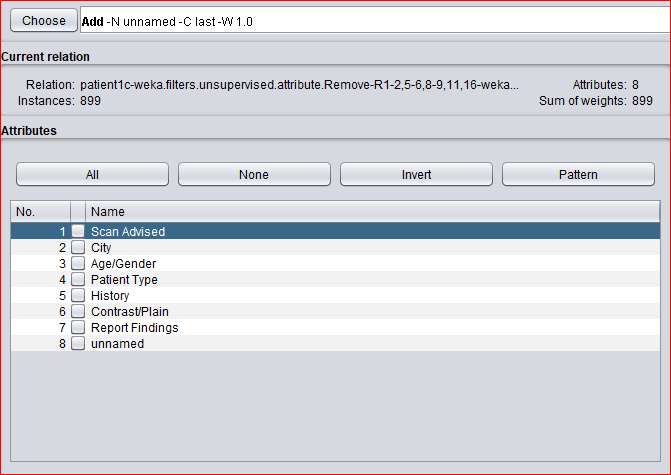
* Click on Choose button and select the Filters option.

A screenshot of a cell phone

Description automatically generated

* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the attribute Add.
* A new window is opened.
* In that we enter attribute index, type, data format, nominal label values like unnamed
* Click on OK.
* Press the Apply button then a new attribute is added to the Table.
* Save the file.
* Click on the Edit button, it shows a new Patient Table on Weka.

New attribute is created.



Remove >Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the attribute Remove.

A screenshot of a cell phone

Description automatically generated

* Select the attributes City to Remove.

A screenshot of a cell phone

Description automatically generated

* Click Remove button and then Save.
* Click on the Edit button, it shows a new Patient Table is appear without city on Weka.

A screenshot of a cell phone

Description automatically generated

Normalize> Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the attribute Normalize.
* Select the attributes which is numeric to Normalize.
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with normalized values on Weka.

A screenshot of a cell phone

Description automatically generated

Binning> Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the attribute Discretize.
* Select the attributes which is numeric to binning.
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with binning values on Weka.

A screenshot of a cell phone

Description automatically generated

Missing value> Pre-Processing Technique:

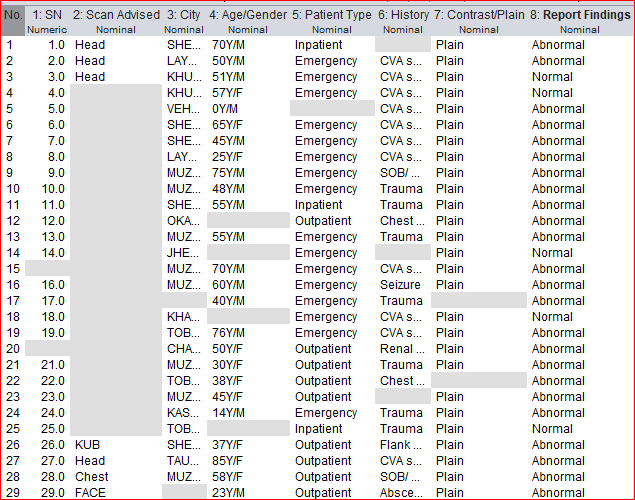
* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the attribute replace missing values/replace missing values with constant/ replace with missing values.
* Select the all attributes.
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with replace the missing value values on Weka.

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a cell phone

Description automatically generated

Nominal to binary> Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the all attributes.
* Select the attributes which is nominal to binary
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with nominal to binary values on Weka.

A screenshot of a cell phone

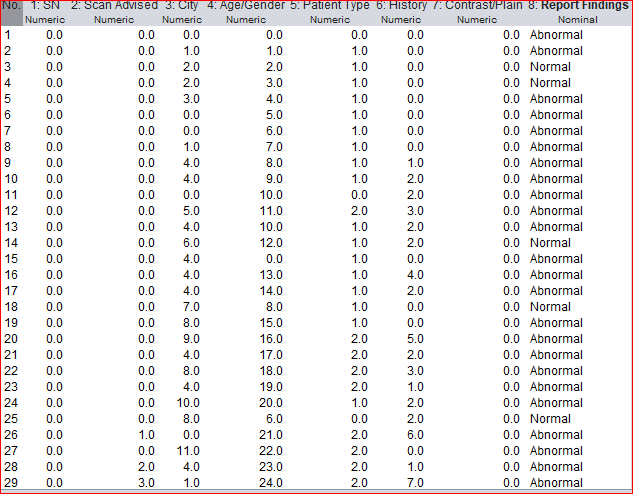
Description automatically generated

A screenshot of a social media post

Description automatically generated

Ordinal to numeric> Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the all attributes.
* Select the attributes which is ordinal to numeric.
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with ordinal to numeric values on Weka.



Principle component> Pre-Processing Technique:

* Click on Choose button and select the Filters option.
* In Filters, we have Supervised and Unsupervised data.
* Click on Unsupervised data.
* Select the all attribute.
* Select the attributes which principle component.
* Click on Apply button and then Save.
* Click on the Edit button, it shows a new Patient Table with principle component values on Weka.

A screenshot of a cell phone

Description automatically generated

Result: This program has been successfully executed.

**EXPERIMENT NO:3**

**Aim:**

Normalize Patient 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 campus and connect them together in order to form a knowledge flow for processing and analyzing the data.

**Do the same steps of experiment1**

Procedure for Knowledge Flow:

**You can also take the csv file as well.**

* Open Start >Programs> weka3-9
* Open the Knowledge Flow.

**A screenshot of a cell phone

Description automatically generated**

* Select the **Data Source** component and add Arff/Csv Loader into the knowledge layout canvas.

A screenshot of a cell phone

Description automatically generated

* Select the **Filters** component and go to supervised and add **Attribute Selection** and Normalize into the knowledge layout canvas.

A screenshot of a cell phone

Description automatically generated

* Select the **Data Sinks** component and add Arff/Csv Saver into the knowledge layout canvas.
* Right click on Arff/ Csv Loader and select Configure option then the new window will be opened and select Patient.arff/csv.
* Right click on Arff/ Csv Loader and select **Dataset** option then establish a link between Arff/ Csv Loader and Attribute Selection.
* Right click on Attribute Selection and select **Dataset** option then establish a link between Attribute Selection and Normalize.
* Right click on Attribute Selection and select **Configure** option and choose the best attribute for Weather data.
* Right click on Normalize and select Dataset option then establish a link between Normalize and Arff/ Csv Saver.
* Right click on Arff/ Csv Saver and select Configure option then new window will be opened and set the path, enter. arff /csv in look in dialog box to save normalize data.
* Right click on Arff/ Csv Loader and click on Start Loading option then everything will be executed one by one.
* Check whether output is created or not by selecting the preferred path.
* Rename the data name as a. arff/ a. Csv.
* Double click on a.arff/a.csv then automatically the output will be opened in MS-Excel

A screenshot of a map

Description automatically generated

Result: This program has been successfully executed.

**EXPERIMENT NO:4**

**Aim:**

Finding Association Rules for Patient 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.

* Open Start>Programs> weka3-9
* Open explorer.
* Click on open file and select Patient.csv/arff.
* Select Associate option on the top of the Menu bar.

Or

* Do the same steps of experiment1

A screenshot of a cell phone

Description automatically generated

* Select Choose button and then click on Apriori Algorithm.
* Click on Start button and output will be displayed on the right side of the window

A screenshot of a computer

Description automatically generated

Result: This program has been successfully executed.

**EXPERIMENT NO:5**

**Aim:**

To Construct Decision Tree for Patient 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 represent 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:**

**Do experiment no. 1**

**Procedure for Decision Trees:**

**1)** Open Start >Programs > Weka-3-9

**2)** Open **explorer**.

**3)** Click on **open file** and select **patient.csv**

**4)** Select **Classifier option** on the top of the Menu bar.

**5)** Select **Choose button** and click on **Tree option**.

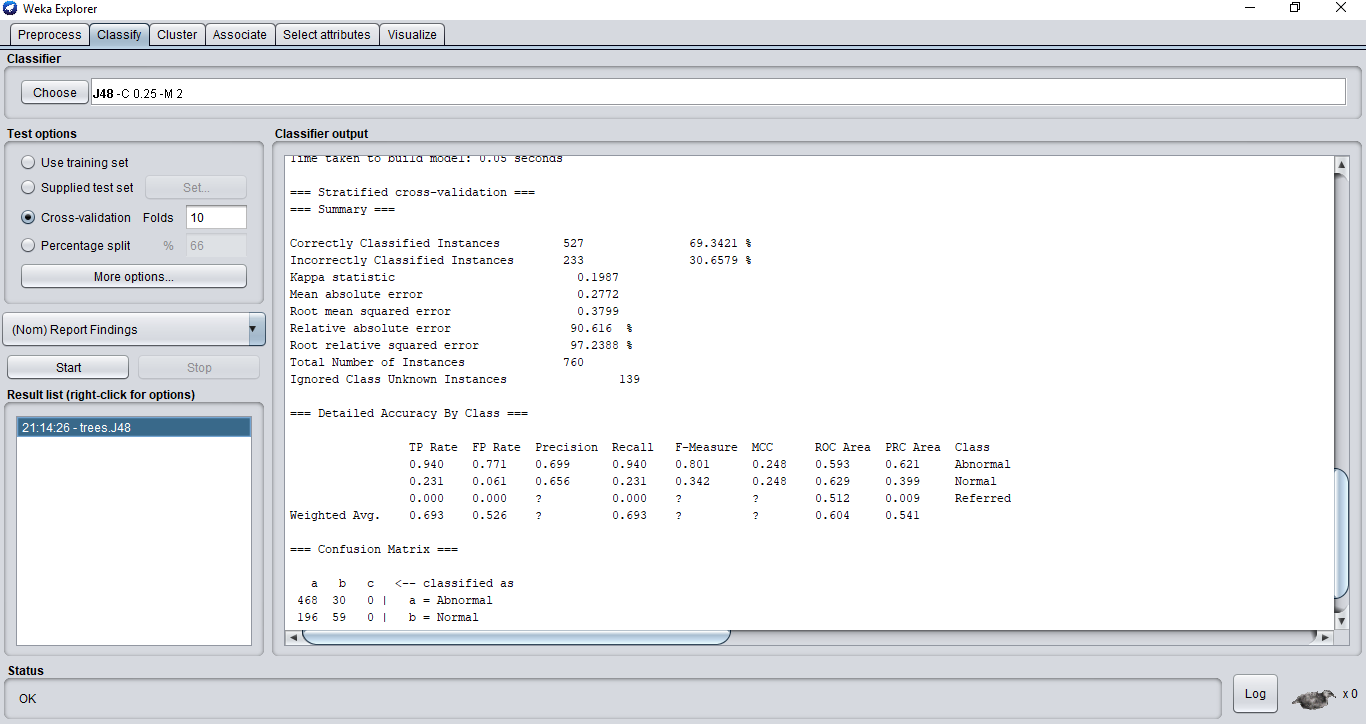
**6)** Click on **J48.**

**7)** Click on **Start button** and output will be displayed on the **right side** of the window.

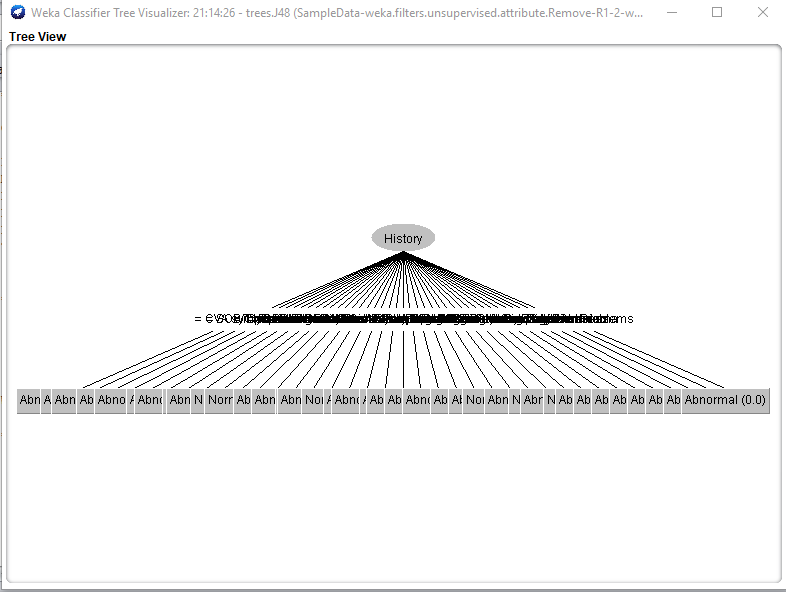
**8)** Select the **result list** and **right click** on result list and select **Visualize Tree option**.

**9)** Then **Decision Tree** will be displayed on **new window**.

**OUTPUT:**

****

**Decision Tree:**

****

**Result:**

This program has been successfully executed.

**EXPERIMENT NO:6**

**Aim:**

Write a procedure for Visualization for Patient Table.

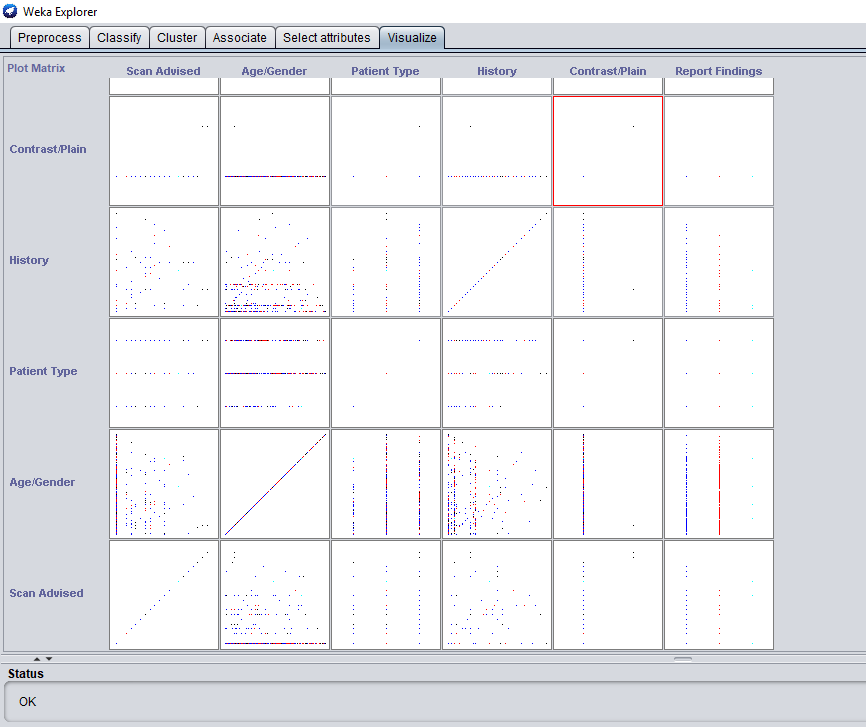
**Description:**

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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no.1**

**2-D Plot Matrix:**

****

**Procedure:**

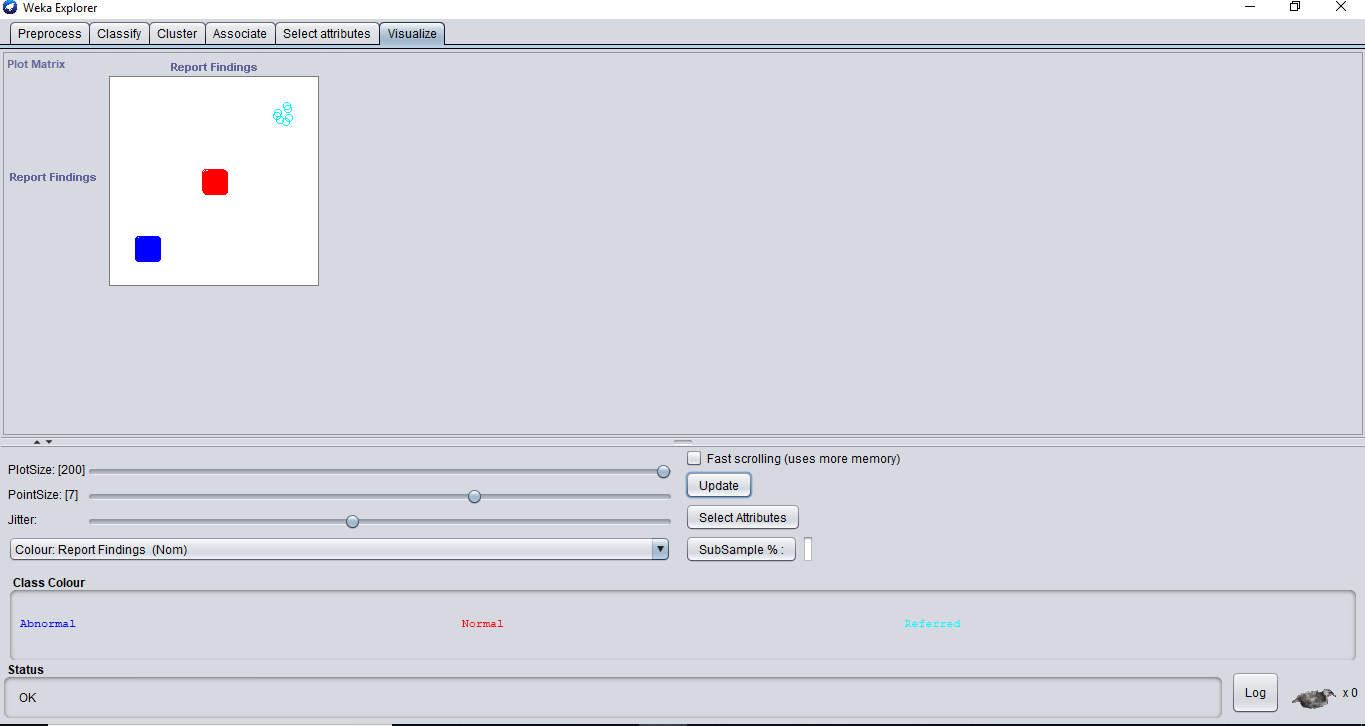
**1)** Open Start >programs> weka3-9

**2)** Open the explorer and click on **Preprocess**, then a new window will appear. In that window select **patient.csv** file then the data will be displayed.

**3)** After that click on the **Visualize tab** on the top of the Menu bar.

**4)** When we select **Visualize tab** then **Plot Matrix** is displayed on the screen.

**Output:**

****

**5)** After that we select the **Select Attribute button**, then select **Outlook attribute** and click OK.

**6)** Click on the **Update button** to display the output.

**7)** After that select the **Select Attribute button** and select **Temperature attribute** and then click OK.

**8) Increase** the **Plot Size** and **Point Size**.

**9)** Click on the **Update button** to display the output.

**10)** After that we select the **Select Attribute button**, then select **Humidity attribute** and click OK.

**11)** Click on the **Update button** to display the output.

**12)** After that select the **Select Attribute button** and select **Windy attribute** and then click OK.

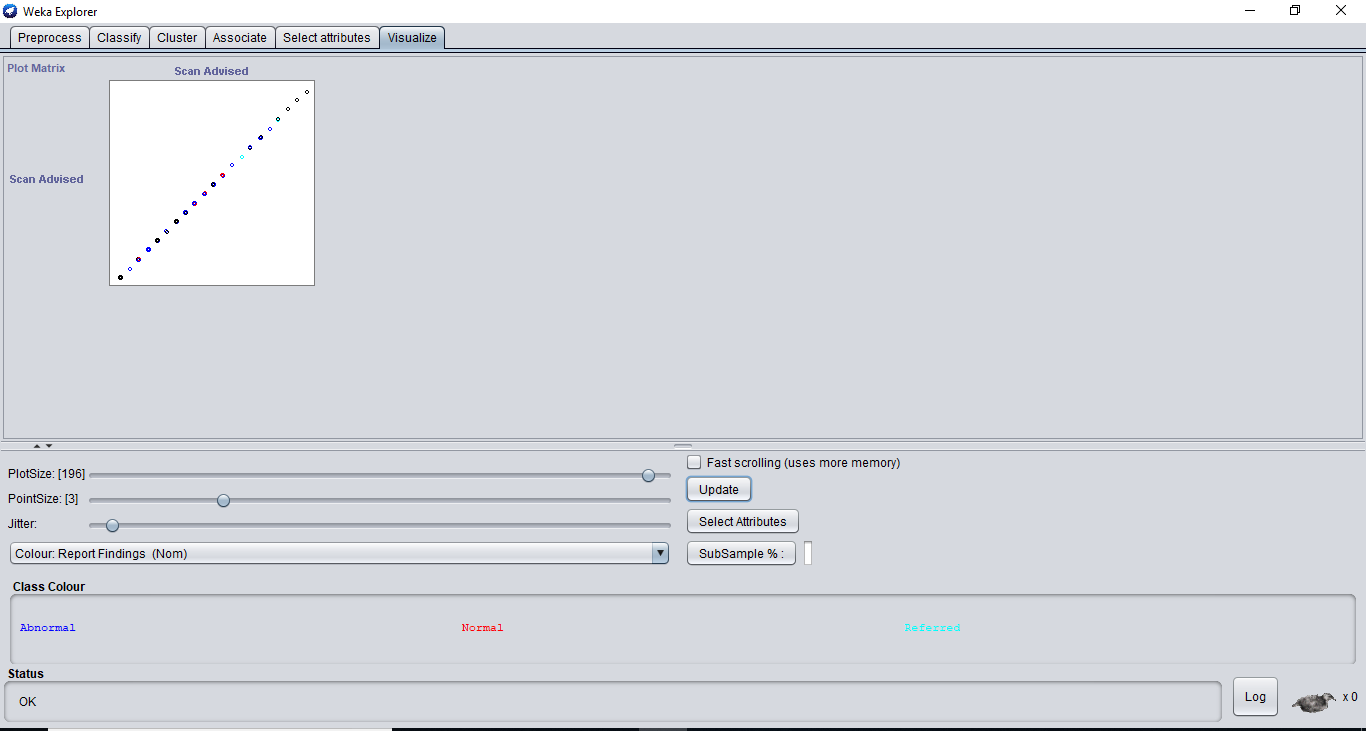
**13) Increase** the **Jitter Size**.

**14)** Click on the **Update button** to display the output.

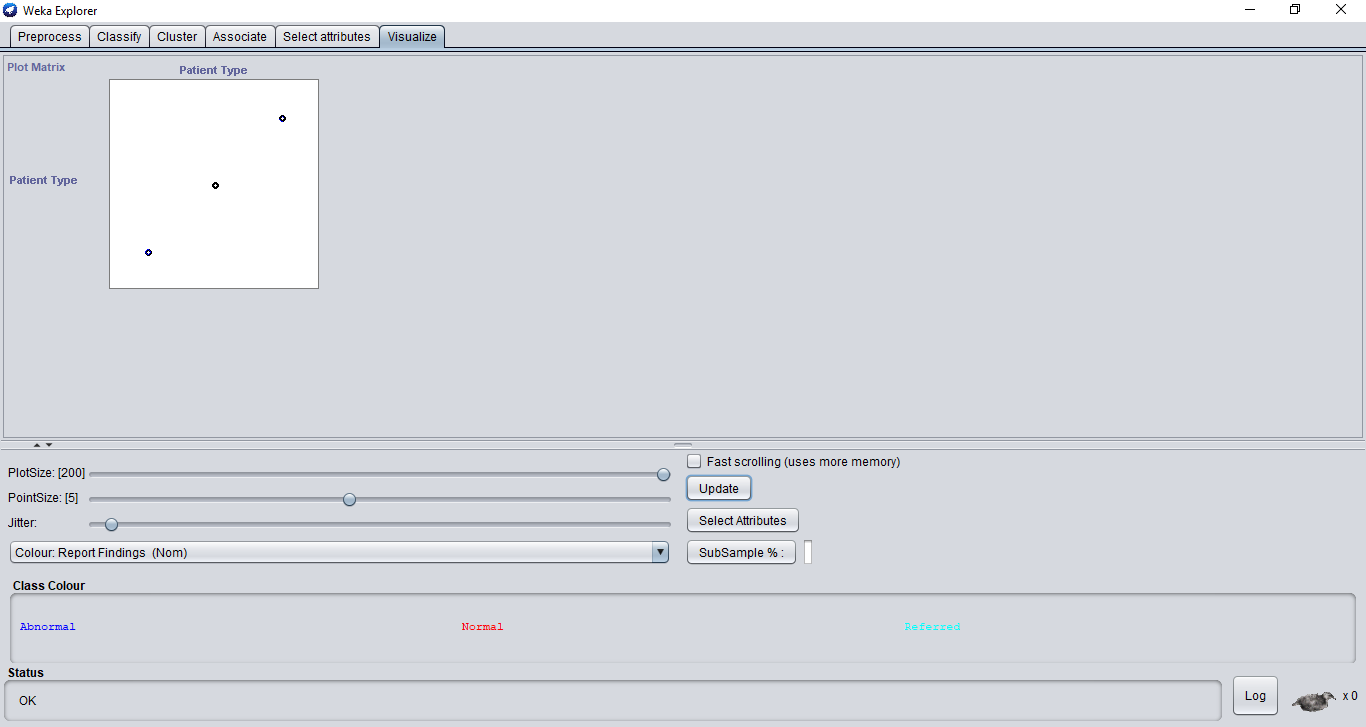
**15)** After that we select the **Select Attribute button**, then select **Play attribute** and click OK.

**16)** Click on the **Update button** to display the output.

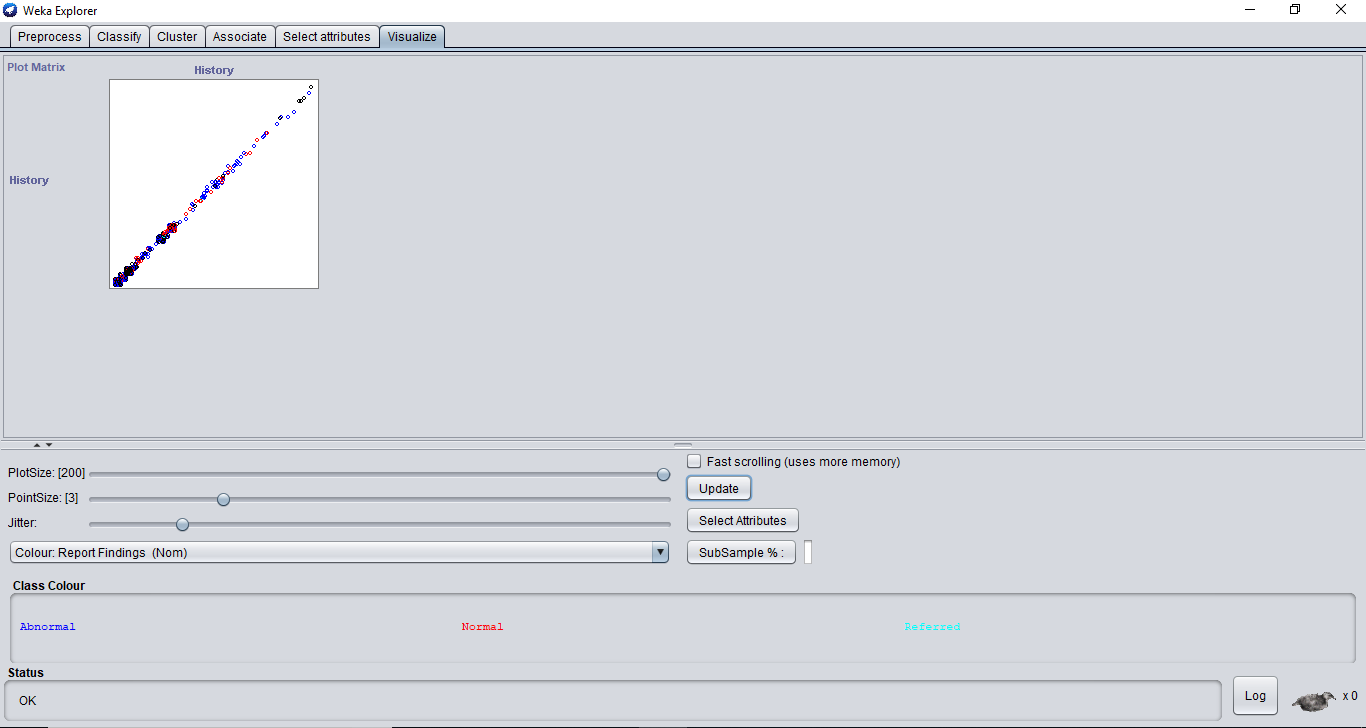
**Output:**

****

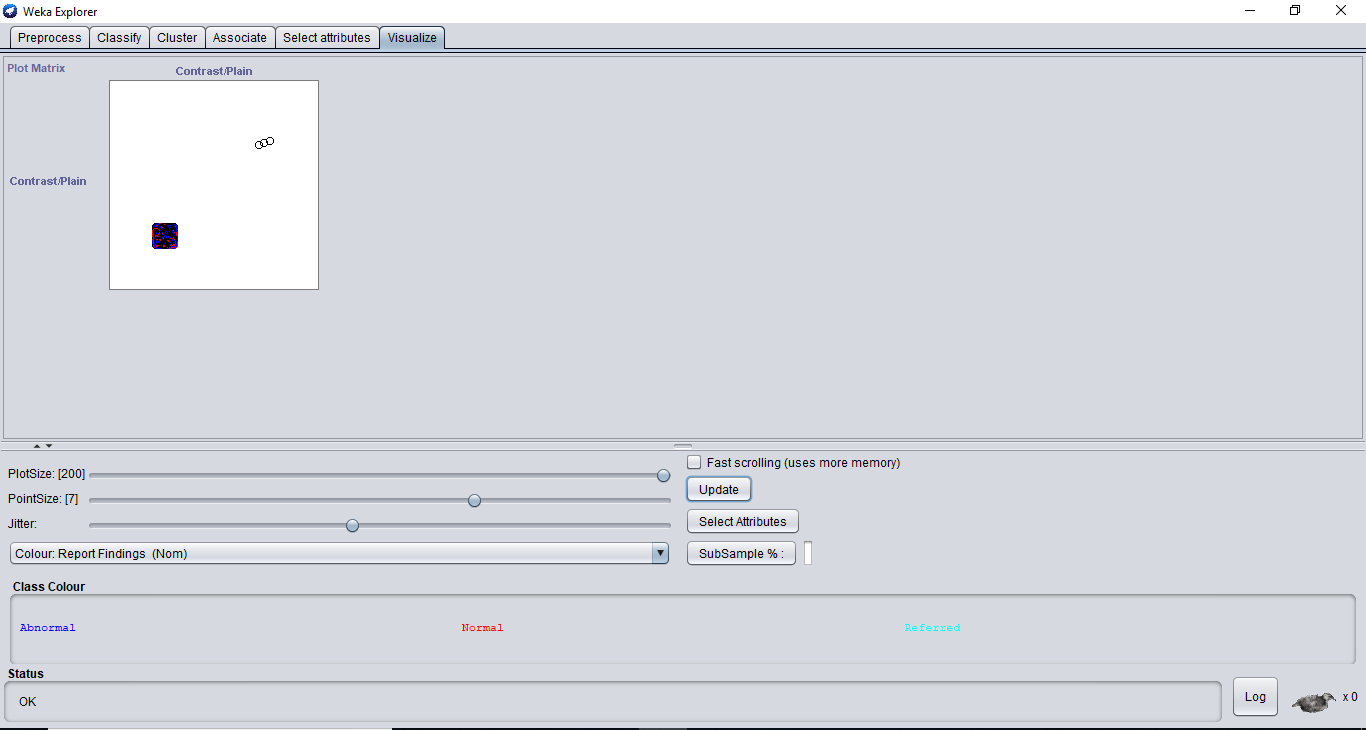
**Output:**

****

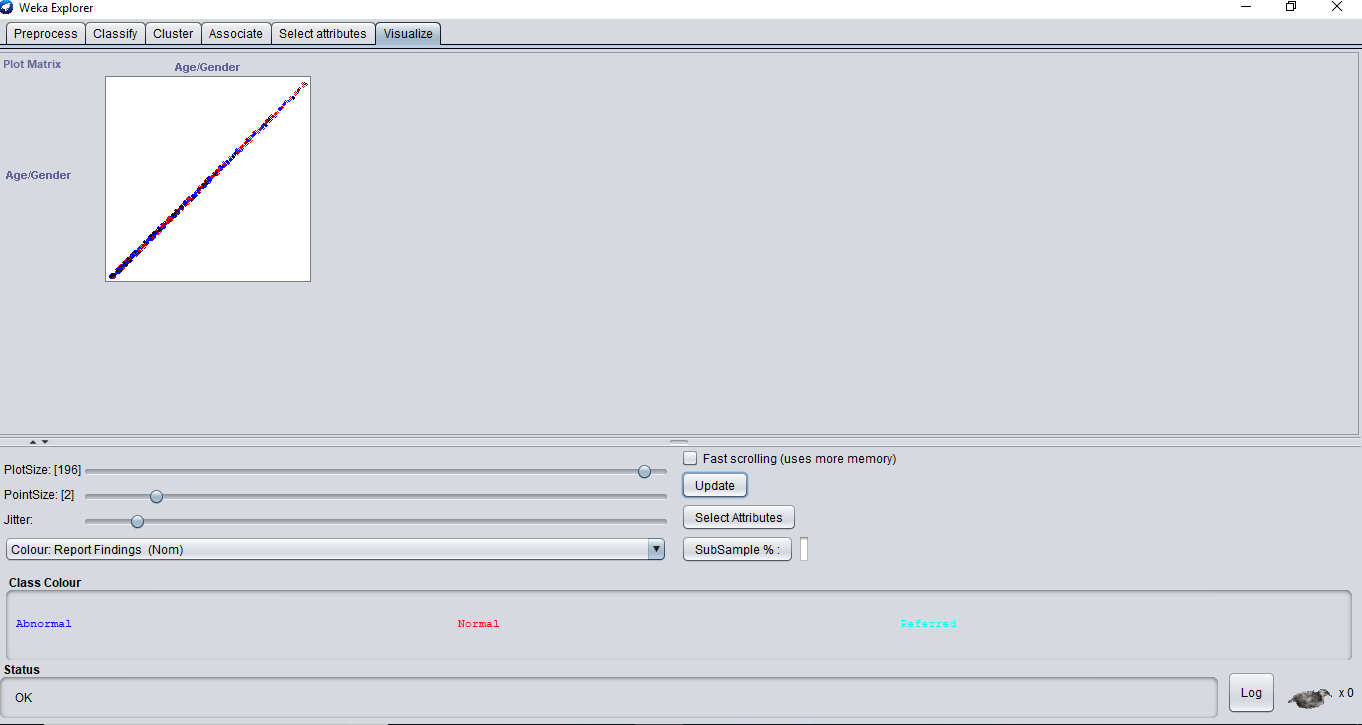
**Output:**

****

**Output:**

****

**Output:**

****

**Result:**

This program has been successfully executed.

**EXPERIMENT NO:7**

**Aim:**

Write a procedure for Clustering Patient 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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no. 1**

**Procedure:**

**1)** Click **Start** > **Programs** > **Weka 3.9**

**2)** Click on **Explorer**.

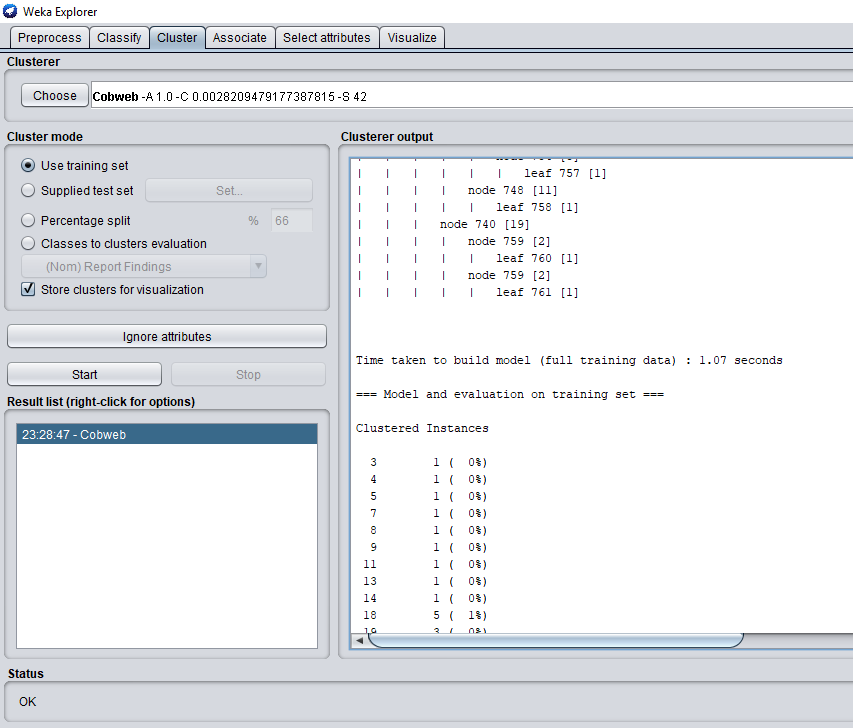
**3)** Click on **open file** & then select **Patient.csv** file.

**4)** Click on **Cluster menu**. In this there are different algorithms are there.

**5)** Click on **Choose button** and then select **cobweb** algorithm.

**6)** Click on **Start button** and then **output** will be displayed on the screen.

**Output:**



**EXPERIMENT NO:8**

**Aim:**

Write a procedure for Clustering Patient 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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no. 1**

**Procedure:**

**1)** Click **Start** -> **Programs** -> **Weka 3.9**

**2)** Click on **Explorer**.

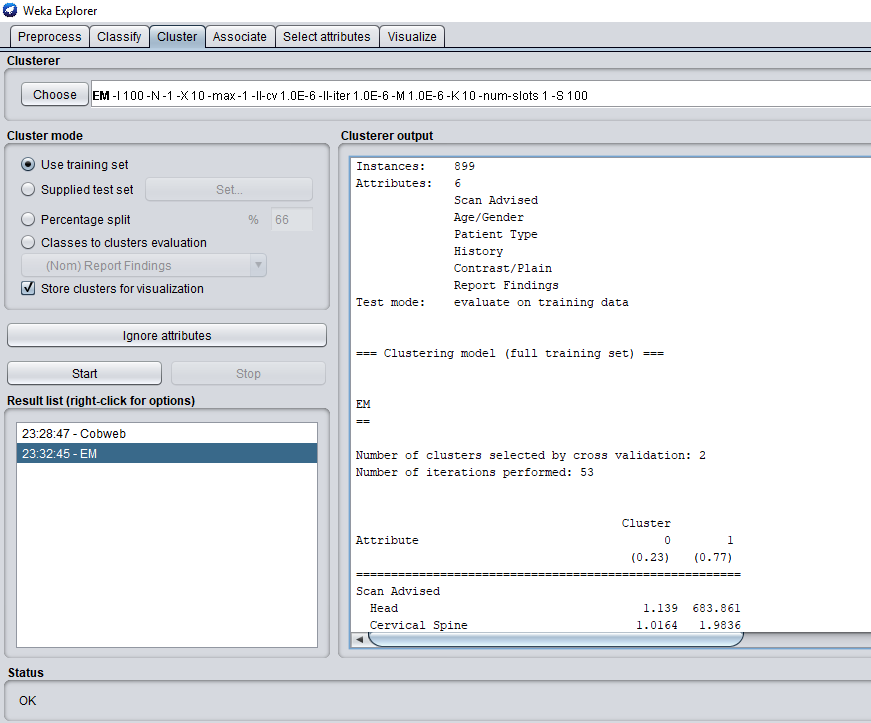
**3)** Click on **open file** & then select **Patient.csv** file.

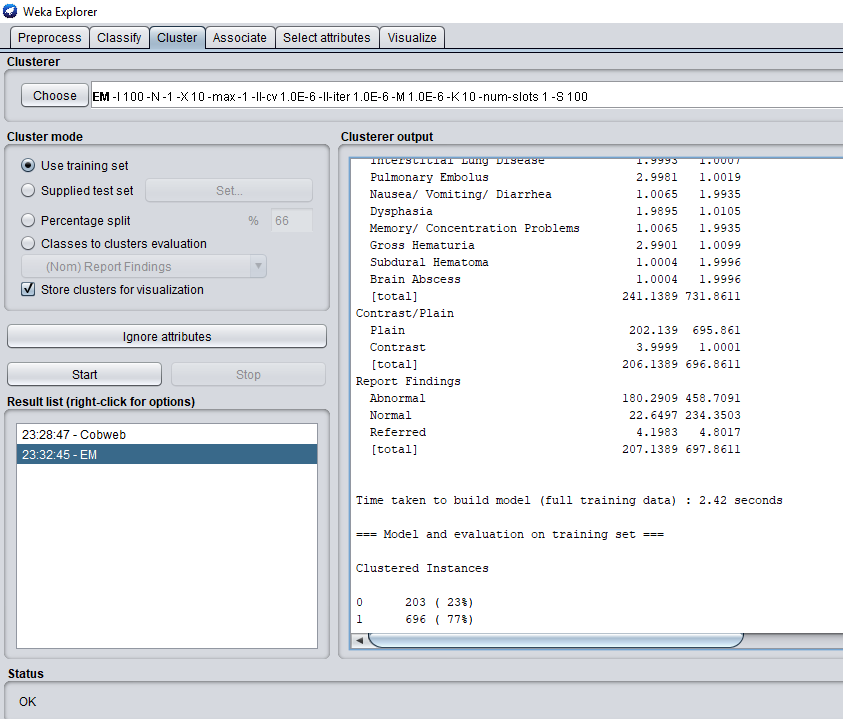
**4)** Click on **Cluster menu**. In this there are different algorithms are there.

**5)** Click on **Choose button** and then select **cobweb** algorithm.

**6)** Click on **Start button** and then **output** will be displayed on the screen.

**Output:**

****

****

**Result:**

The program has been successfully executed.

**EXPERIMENT NO:9**

**Aim:**

Write a procedure for Clustering Patient data using 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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no. 1**

**Procedure:**

**1)** Click **Start** > **Programs** > **Weka 3.9**

**2)** Click on **Explorer**.

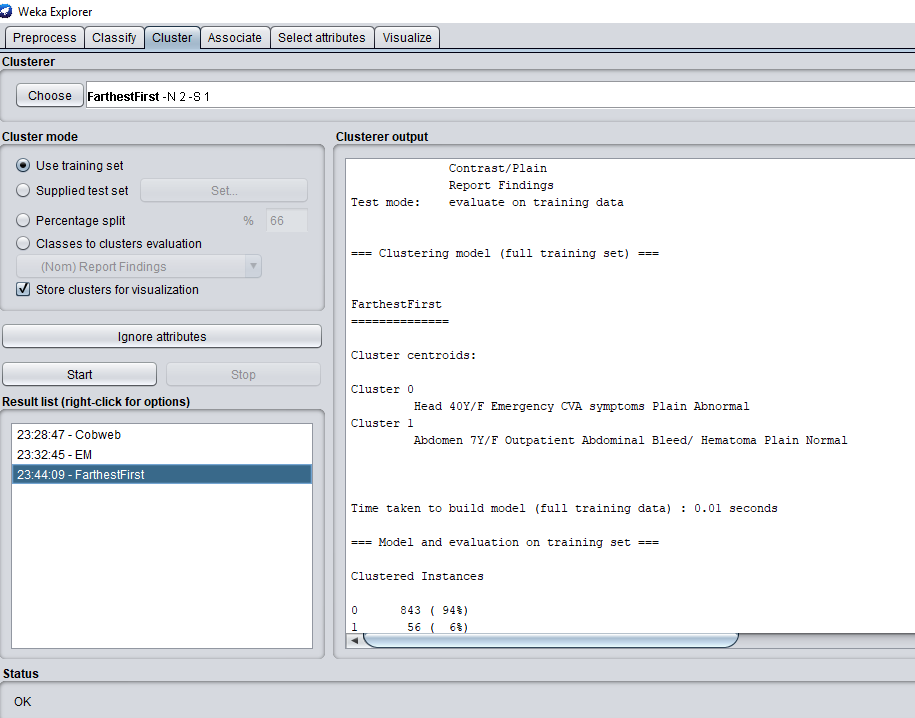
**3)** Click on **open file** & then select **Patient.csv** file.

**4)** Click on **Cluster menu**. In this there are different algorithms are there.

**5)** Click on **Choose button** and then select **Farthest First** algorithm.

**6)** Click on **Start button** and then **output** will be displayed on the screen.

**Output:**

****

**Result:**

This program has been successfully executed.

**EXPERIMENT NO:10**

**Aim:**

Write a procedure for Patient 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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no. 1**

**Procedure:**

**1)** Click **Start** > **Programs** > **Weka 3.9**

**2)** Click on **Explorer**.

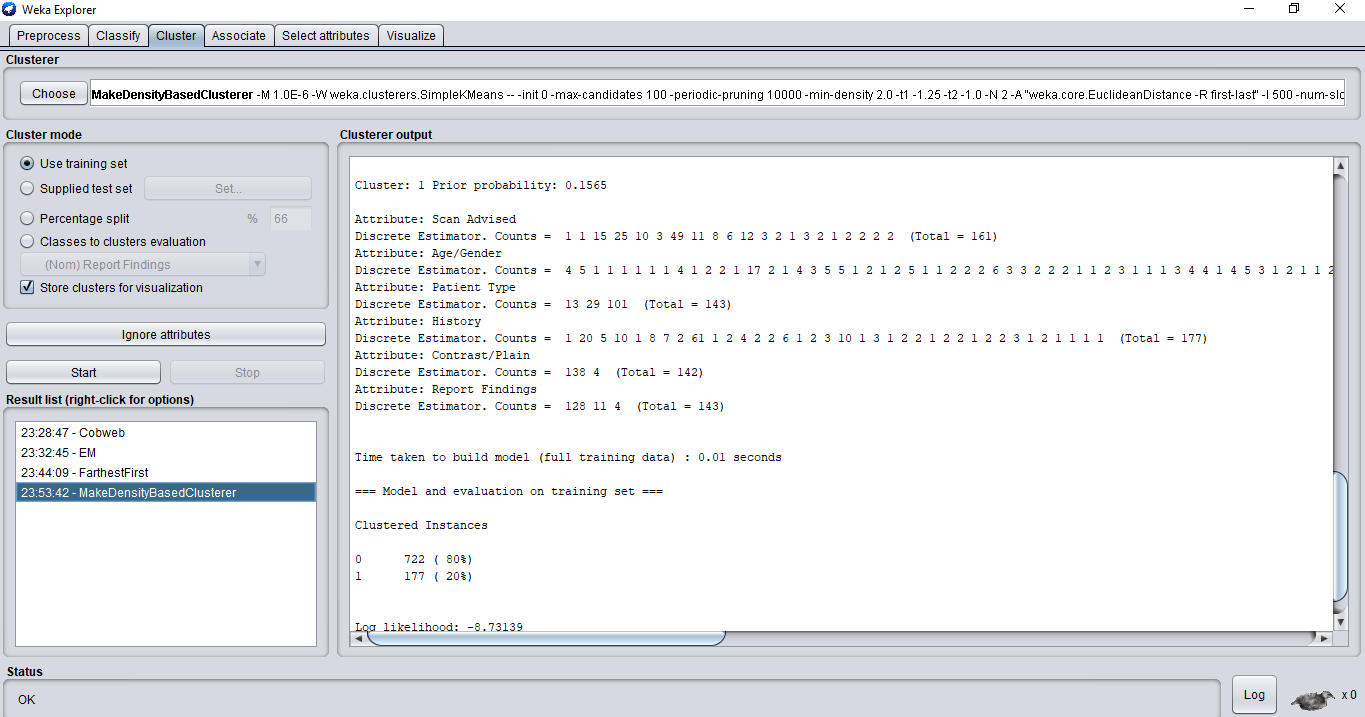
**3)** Click on **open file** & then select **Patient.csv** file.

**4)** Click on **Cluster menu**. In this there are different algorithms are there.

**5)** Click on **Choose button** and then select **MakeDensityBasedClusterer** algorithm.

**6)** Click on **Start button** and then **output** will be displayed on the screen.

**Output:**

****

**Result:**

The program has been successfully executed.

**EXPERIMENT NO:11**

**Aim:**

Write a procedure for Clustering Patient 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.

**Creation of Patient Table:**

**Procedure: Repeat experiment no. 1**

**Procedure:**

**1)** Click **Start** > **Programs** > **Weka 3.9**

**2)** Click on **Explorer**.

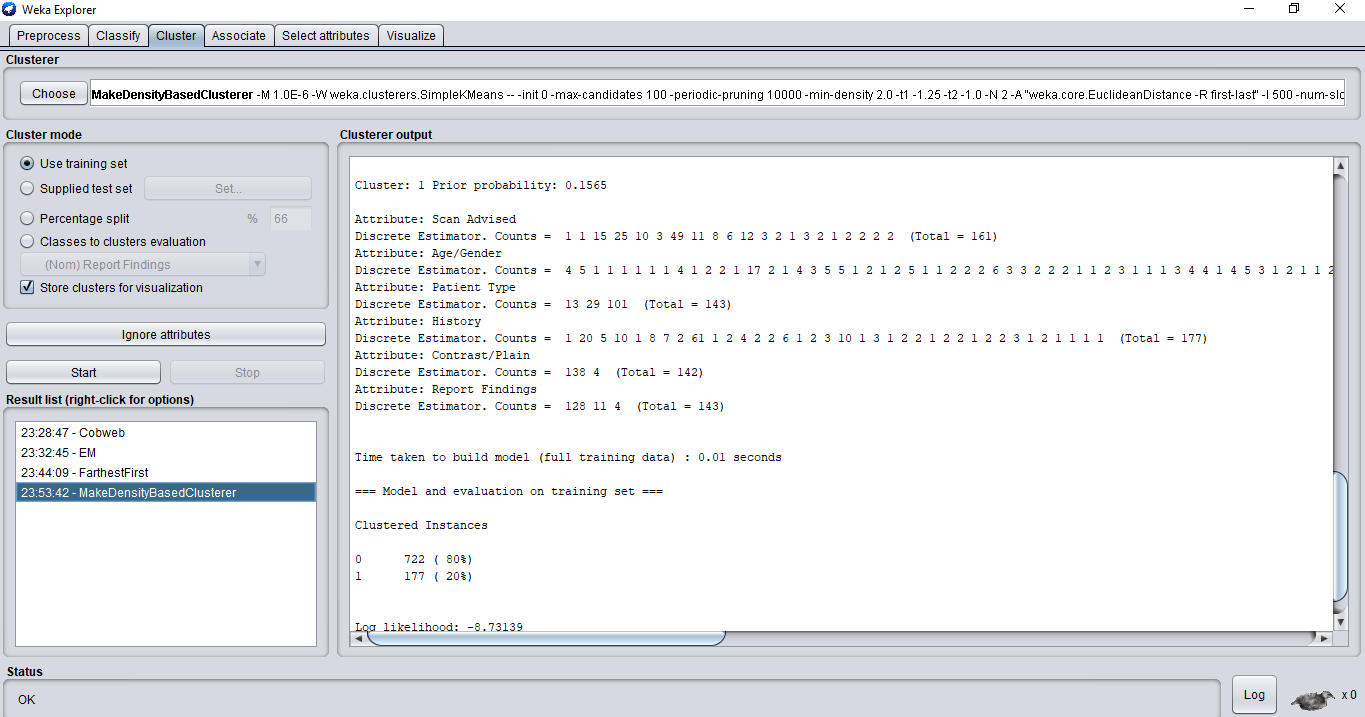
**3)** Click on **open file** & then select **Patient.csv** file.

**4)** Click on **Cluster menu**. In this there are different algorithms are there.

**5)** Click on **Choose button** and then select **Simple Means** algorithm.

**6)** Click on **Start button** and then **output** will be displayed on the screen.

**Output:**

****

**Result:**

The program has been successfully executed.

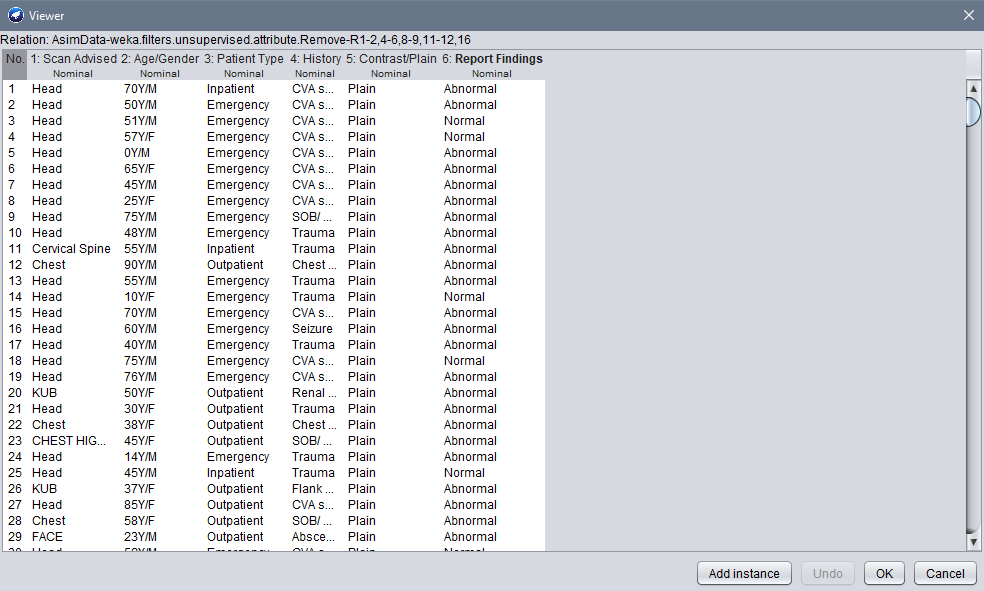
**EXPERIMENT NO: 12**

**Aim:**

Write a procedure for cross-validation using J48 Algorithm for Patient 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:**

**You can also take csv file as well.**

**1) Start** > **Programs** > **Weka 3.9**

**2)** Open **Knowledge Flow**.

**3)** Select **Data Source** tab & choose **Arff Loader**.

**4)** Place **Arff Loader** component on the **layout area** by clicking on that component.

**5)** 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 **ReportFinding.arff**

**6)** Click on the **Evaluation** tab & choose **Class Assigner** & place it on the layout.

**7)** 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.

**8)** Right click on **Class Assigner** & choose **Configure** option, and then a new window will appear & specify a class to our data.

**9)** Select **Evaluation** tab & select **Cross-Validation Fold Maker** & place it on the layout.

**10)** Now **connect** the **Class Assigner** to the **Cross-Validation Fold Maker**.

**11)** Select **Classifiers** tab & select **J48** component & place it on the layout.

**12)** Now **connect Cross-Validation Fold Maker** to **J48 twice**; **first** choose **Training Data Set** option and **then Test Data Set** option.

**13)** Select **Evaluation Tab** & select **Classifier Performance Evaluator** component & place it on the layout.

**14)** Connect **J48** to **Classifier Performance Evaluator** component by right clicking on J48 & selecting **Batch Classifier**.

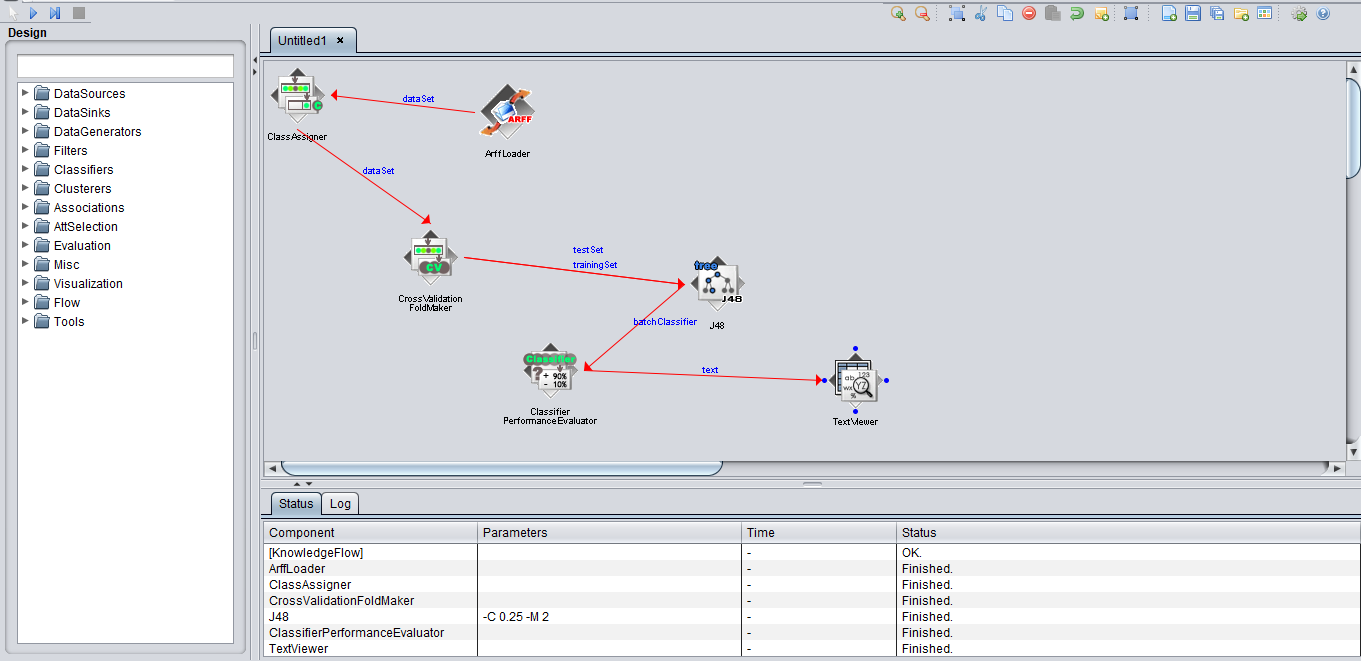
**15)** Select **Visualization** tab & select **Text Viewer** component & place it on the layout.

**16)** Connect **Text Viewer** to **Classifier Performance Evaluator** by right clicking on Text Viewer & by selecting **Text** option.

**17)** Start the flow of execution by selecting **Start Loading** from **Arff Loader**.

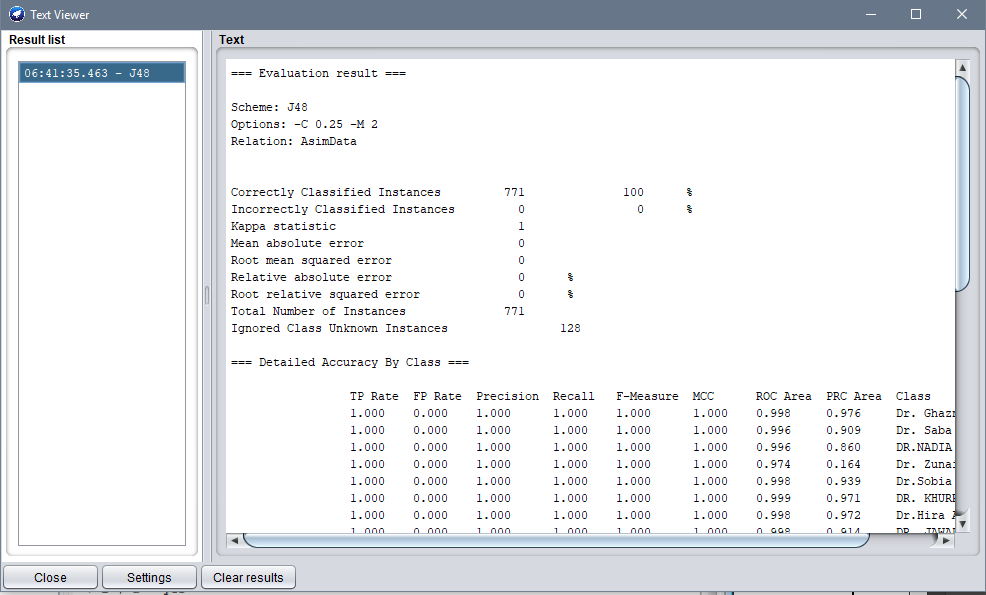
54

After **17 step** we get something like that:



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

**Results are:**

****

**Result:**

This program has been successfully executed.