ABESIT

COLLEGE CODE – 290

Lab File

NAME	SANDEEP KUMAR SHUKLA
BRANCH	CSE
UNIVERSITY ROLL NO.	1729010140
SESSION	2019-20
NAME OF LAB	Data Warehousing & Data Mining Lab (RCS-654)

INDEX

Experiment No. 1

Title: Create data-set in .arff file format.

S/w Requirement: Excel, Text Editor(Notepad)

Objective: To create data-set in .arff file format for WEKA using Excel and Notepad.

Steps for creating a ARFF File in Weka :

Step1: You have a XLSX file then you need to convert it into a CSV(Comma Separated Values)File.

Step2: Then Open the CSV File with a text editor eg .Notepad

Step3: Append header relation eg.@relation student

Step4: After that append the file with headers equal to the number of instances in your XLSX file eg. @attribute age {<30,30-40,>40} @attribute income {low, medium, high} @attribute student {yes, no} @attribute credit-rating {fair, excellent}. This means the file has four columns excluding the class label.

Step5: Add the class label relation eg. @attribute buyspc {yes, no}.This has 2 classes mainly yes and no.

Step 6: After that append the header with @data and then save the file as .Arff

Dataset (.arff)

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
```

overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

Conclusion: The data is converted from excel to csv format and then to arff format.

*****

Experiment No. 2

Title: Demonstration of preprocessing on WEKA data-set.

S/w Requirement: WEKA

Objective: To demonstrate some of the basic data preprocessing operations that can be performed using WEKA-Explorer. The sample dataset used for this example is the student data available in .arff format.

Steps to implement preprocessing operations on WEKA-Explorer:

Step1: Loading the data. We can load the dataset into weka by clicking on open button in preprocessing interface and selecting the appropriate file.

Step2: Once the data is loaded, weka will recognize the attributes and during the scan of the data weka will compute some basic strategies on each attribute. The left panel in the above figure shows the list of recognized attributes while the top panel indicates the names of the base relation or table and the current working relation (which are same initially).

Step3: Clicking on an attribute in the left panel will show the basic statistics on the attributes for the categorical attributes the frequency of each attribute value is shown, while for continuous attributes we can obtain min, max, mean, standard deviation and deviation etc.

Step4: The visualization in the right button panel in the form of cross-tabulation across two attributes.
Note:we can select another attribute using the dropdown list.

Step5: Selecting or filtering attributes
Removing an attribute-When we need to remove an attribute,we can do this by using the attribute filters in weka.In the filter model panel,click on choose button,This will show a popup window with a list of available filters.
Scroll down the list and select the "weka.filters.unsupervised.attribute.remove" filters.

Step 6:
a)Next click the textbox immediately to the right of the choose button.In the resulting dialog box enter the index of the attribute to be filtered out.
b)Make sure that invert selection option is set to false.The click OK now in the filter box.you will see "Remove-R-7".
c)Click the apply button to apply filter to this data.This will remove the attribute and create new working relation.
d)Save the new working relation as an arff file by clicking save button on the top(button)panel.(student.arff)

Discretization
Sometimes association rule mining can only be performed on categorical data.This requires performing discretization on numeric or continuous attributes.In the following example let us discretize age attribute.

i)Let us divide the values of age attribute into three bins(intervals).

ii)First load the dataset into weka(student.arff)

iii)Select the age attribute.

iv)Activate filter-dialog box and select “WEKA.filters.unsupervised.attribute.discretize”from the list.

v)To change the defaults for the filters,click on the box immediately to the right of the choose button.

vi)We enter the index for the attribute to be discretized.In this case the attribute is age.So we must enter ‘1’ corresponding to the age attribute.

vii)Enter ‘3’ as the number of bins.Leave the remaining field values as they are.

viii)Click OK button.

ix)Click apply in the filter panel.This will result in a new working relation with the selected attribute partition into 3 bins.

x)Save the new working relation in a file called student-data-discretized.arff

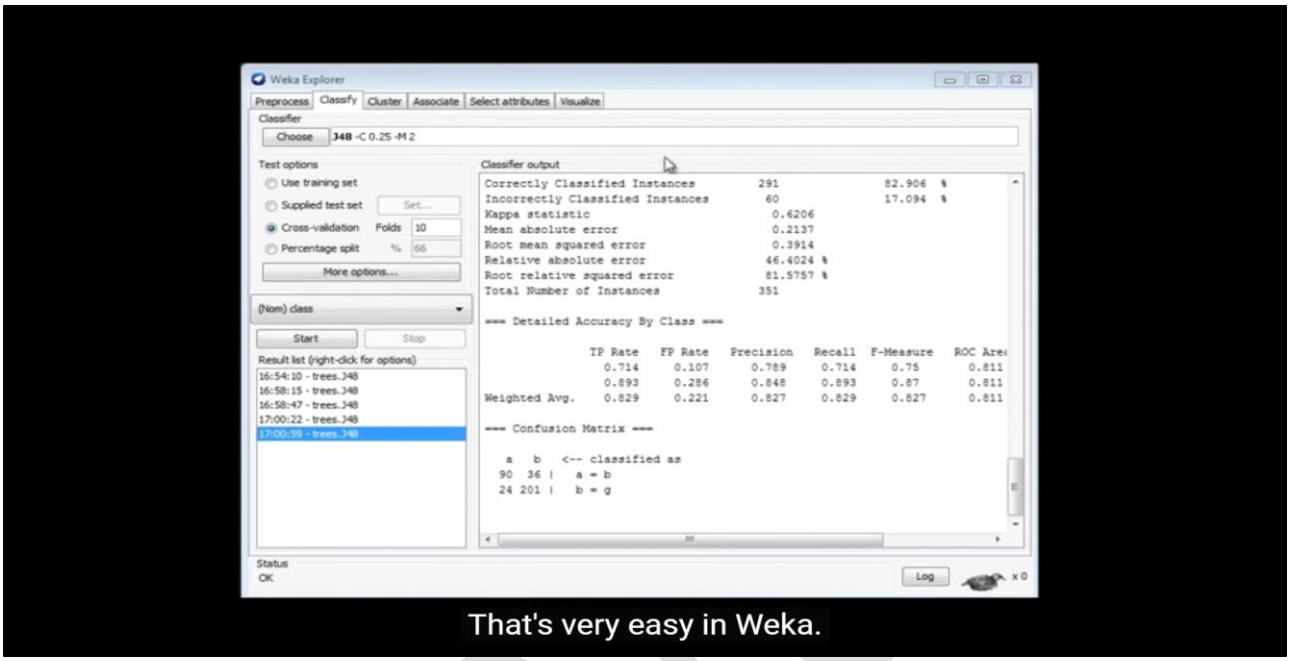
Dataset (.arff)

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

The following screenshot shows the effect of discretization :



Conclusion: We have performed the discretization of data set into a new attribute.

*****

Experiment No. 3

Title: Demonstration of Association rule process on WEKA data-set using Apriori Algorithm.

S/w Requirement: WEKA

Objectives: To demonstrate some of the basic elements of asscociation rule mining using WEKA. The sample dataset used for this example is contactlenses.arff.

Steps to implement Association rule process using Apriori Algorithm:

Step1: Open the data file in Weka Explorer. It is presumed that the required data fields have been discretized. In this example it is age attribute.

Step2: Clicking on the associate tab will bring up the interface for association rule algorithm.

Step3: We will use apriori algorithm. This is the default algorithm.

Step4: Inorder to change the parameters for the run (example support, confidence etc) we click on the text box immediately to the right of the choose button.

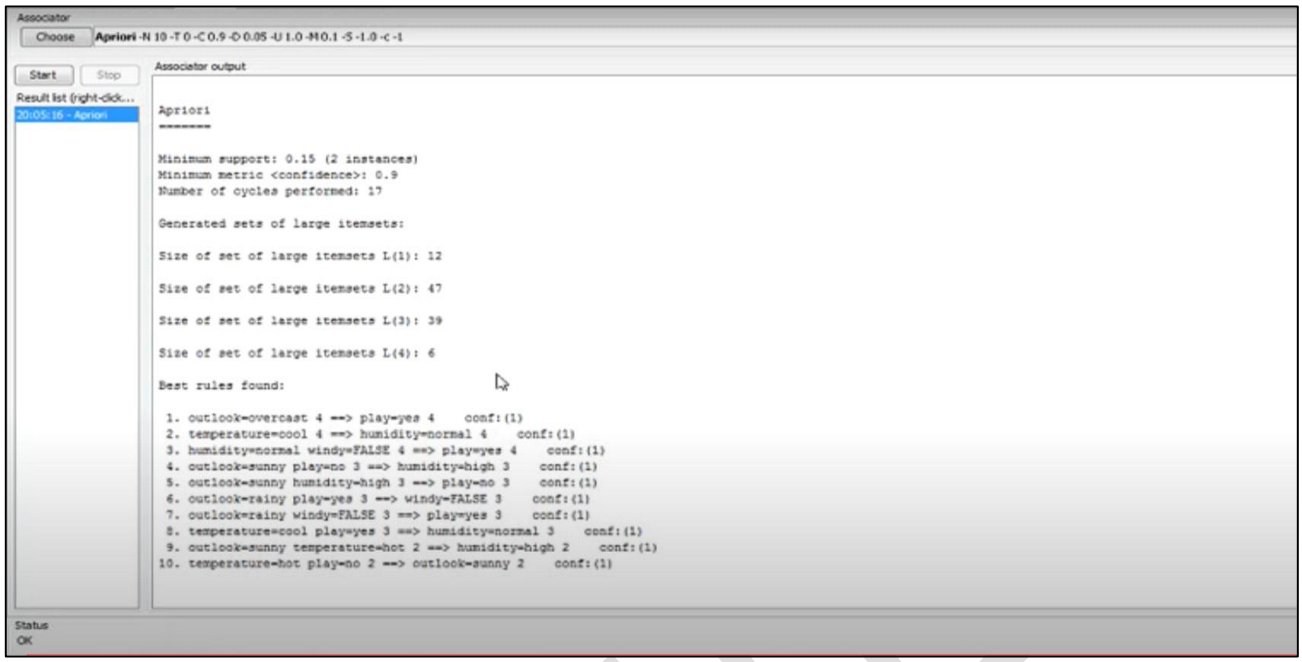
Dataset (.arff)

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

The following screenshot shows the association rules that were generated when Apriori algorithm is applied on the given dataset.



Conclusion: The experiment displays Set of large itemsets, best rule found for the given support and the confidence values. We get the results faster using the toolkits.

*****

Experiment No. 4

Title: Demonstration of classification rule process on WEKA data-set using Naive Bayes algorithm.

S/w Requirement: WEKA

Objective: To learn the use of naïve bayes classifier in weka. The sample data set used in this experiment is "employee"data available at .arff format.

Execution steps :

Step1: We begin the experiment by loading the data (employee.arff) into weka.

Step2: Next we select the "classify" tab and click "choose" button to select the "id3"classifier.

Step3: Now we specify the various parameters. These can be specified by clicking in the text box to the right of the chose button. In this example, we accept the default values his default version does perform some pruning but does not perform error pruning.

Step4: Under the "text "options in the main panel. We select the 10-fold cross validation as our evaluation approach. Since we don't have separate evaluation data set, this is necessary to get a reasonable idea of accuracy of generated model.

Step-5: we now click"start"to generate the model .the ASCII version of the tree as well as evaluation statistic will appear in the right panel when the model construction is complete.

Step-6: note that the classification accuracy of model is about 69%.this indicates that we may find more work. (Either in preprocessing or in selecting current parameters for the classification)

Step-7: now weka also lets us a view a graphical version of the classification tree. This can be done by right clicking the last result set and selecting "visualize tree" from the pop-up menu.

Step-8: we will use our model to classify the new instances.

Step-9: In the main panel under "text "options click the "supplied test set" radio button and then click the "set" button. This will show pop-up window which will allow you to open the file containing test instances.

Dataset (.arff)

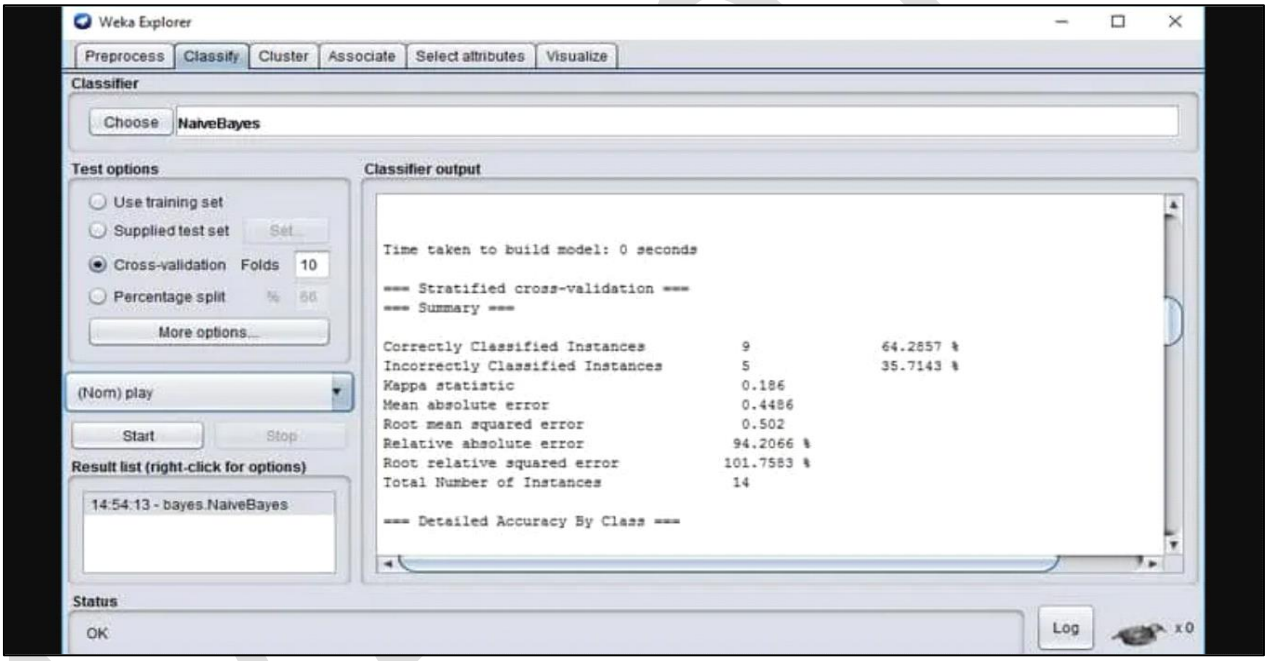
```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
```

sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

The following screenshot shows the classification rules that were generated when naive bayes algorithm is applied on the given dataset.



Conclusion: The naive bayes algorithm is able to classify the data in the employee database.

*****

Experiment No. 5

Title: Demonstration of classification rule process on WEKA data-set using j48 algorithm.

S/w Requirement: WEKA

Objective: To learn to use the Weka machine learning toolkit for j48, decision tree classifier

Execution steps :

Step-1: We begin the experiment by loading the data (student.arff)into weka.

Step2: Next we select the "classify" tab and click "choose" button t o select the "j48"classifier.

Step3: Now we specify the various parameters. These can be specified by clicking in the text box to the right of the chose button. In this example, we accept the default values. The default version does perform some pruning but does not perform error pruning.

Step4: Under the "text" options in the main panel. We select the 10-fold cross validation as our evaluation approach. Since we don't have separate evaluation data set, this is necessary to get a reasonable idea of accuracy of generated model.

Step-5: We now click "start" to generate the model .the Ascii version of the tree as well as evaluation statistic will appear in the right panel when the model construction is complete.

Step-6: Note that the classification accuracy of model is about 69%.this indicates that we may find more work. (Either in preprocessing or in selecting current parameters for the classification)

Step-7: Now weka also lets us a view a graphical version of the classification tree. This can be done by right clicking the last result set and selecting "visualize tree" from the pop-up menu.

Step-8: We will use our model to classify the new instances.

Step-9: In the main panel under "text" options click the "supplied test set" radio button and then click the "set" button. This wills pop-up a window which will allow you to open the file containing test instances.

Dataset (.arff)

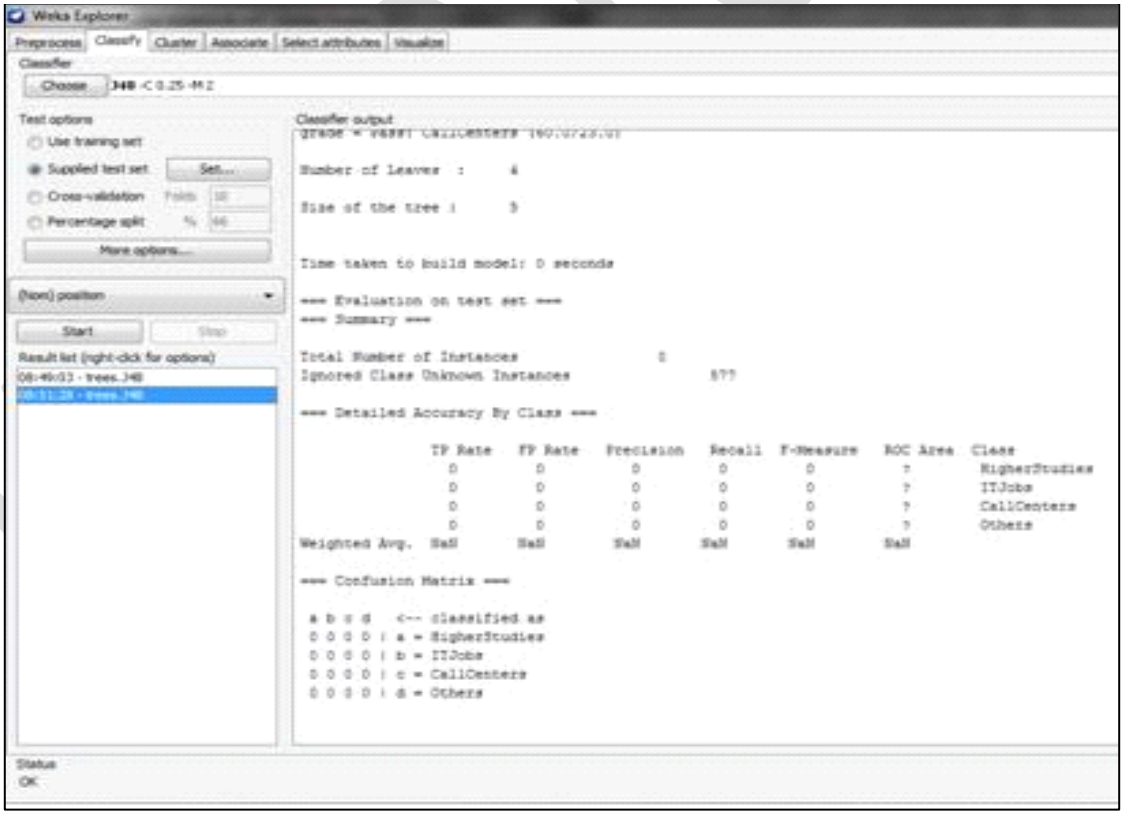
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real

@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

The following screenshot shows the classification rules that were generated when j48 algorithm is applied on the given dataset.



Conclusion: The experiment displays decision tree, which is annotated (labeled). It also gives the time taken to build the tree and the confusion matrix.

*****

Experiment No. 6

Title: Demonstration of clustering rule process on WEKA data-set using k-means clustering algorithm.

S/w Requirement: WEKA

Objective: To learn to use the WEKA machine learning toolkit for simple k-means clustering.

Execution steps :

Step 1: Run the Weka explorer and load the data file iris.arff in preprocessing interface.

Step 2: Inorder to perform clustering select the 'cluster' tab in the explorer and click on the choose button. This step results in a dropdown list of available clustering algorithms.

Step 3: In this case we select 'SimpleKMeans'.

Step 4: Next click in text button to the right of the choose button to get popup window shown in the screenshots. In this window we enter six on the number of clusters and we leave the value of the seed on as it is. The seed value is used in generating a random number which is used for making the internal assignments of instances of clusters.

Step 5: Once of the option have been specified. We run the clustering algorithm there we must make sure that they are in the 'cluster mode' panel. The use of training set option is selected and then we click 'start' button. This process and resulting window are shown in the following screenshots.

Dataset (.arff)

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
```

sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

Conclusion: The k means clustering is able the cluster the data in the iris database.

*****

Experiment No. 7

Title: Demonstration of clustering rule process on WEKA data-set using hierarchical clustering .

S/w Requirement: WEKA

Objective: To learn to use the WEKA machine learning toolkit for hierarchical clustering.

Execution steps :

Step 1: Run the Weka explorer and load the data file iris.arff in preprocessing interface.

Step 2: Inorder to perform clustering select the 'cluster' tab in the explorer and click on the choose button. This step results in a dropdown list of available clustering algorithms.

Step 3: In this case we select 'HierarchicalClusterer'.

Step 4: Next click in text button to the right of the choose button to get popup window shown in the screenshots. In this window we enter six on the number of clusters and we leave the value of the seed on as it is. The seed value is used in generating a random number which is used for making the internal assignments of instances of clusters.

Step 5: Once of the option have been specified. We run the clustering algorithm there we must make sure that they are in the 'cluster mode' panel. The use of training set option is selected and then we click 'start' button. This process and resulting window are shown in the following screenshots.

Dataset (.arff)

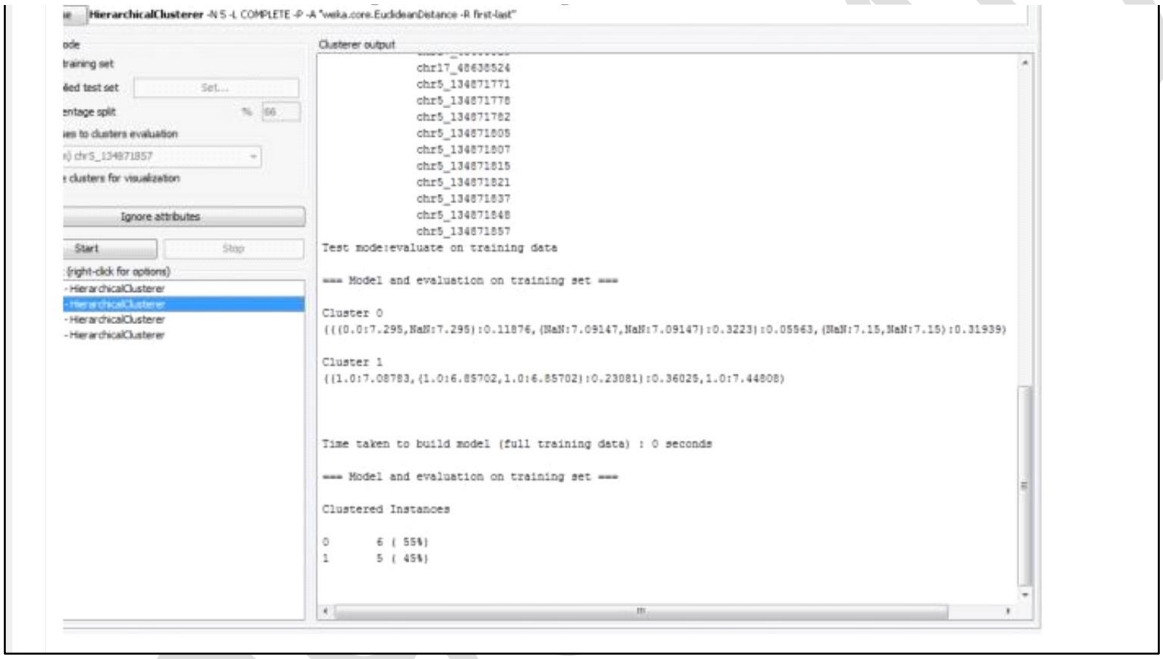
```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
```

sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

The following screenshot shows the clustering rules that were generated when hierarchical clustering algorithm is applied on the given dataset:



Conclusion: The hierarchical clustering is able the cluster the data in the iris database.

*****

Experiment No. 8

Title: Demonstration of clustering rule process on WEKA data-set using density based clustering .

S/w Requirement: WEKA

Objective: To learn to use the WEKA machine learning toolkit for density based clustering.

Execution steps :

Step 1: Run the Weka explorer and load the data file iris.arff in preprocessing interface.

Step 2: Inorder to perform clustering select the 'cluster' tab in the explorer and click on the choose button. This step results in a dropdown list of available clustering algorithms.

Step 3: In this case we select 'MakeDensityBasedClusterer'.

Step 4: Next click in text button to the right of the choose button to get popup window shown in the screenshots. In this window we enter six on the number of clusters and we leave the value of the seed on as it is. The seed value is used in generating a random number which is used for making the internal assignments of instances of clusters.

Step 5: Once of the option have been specified. We run the clustering algorithm there we must make sure that they are in the 'cluster mode' panel. The use of training set option is selected and then we click 'start' button. This process and resulting window are shown in the following screenshots.

Dataset (.arff)

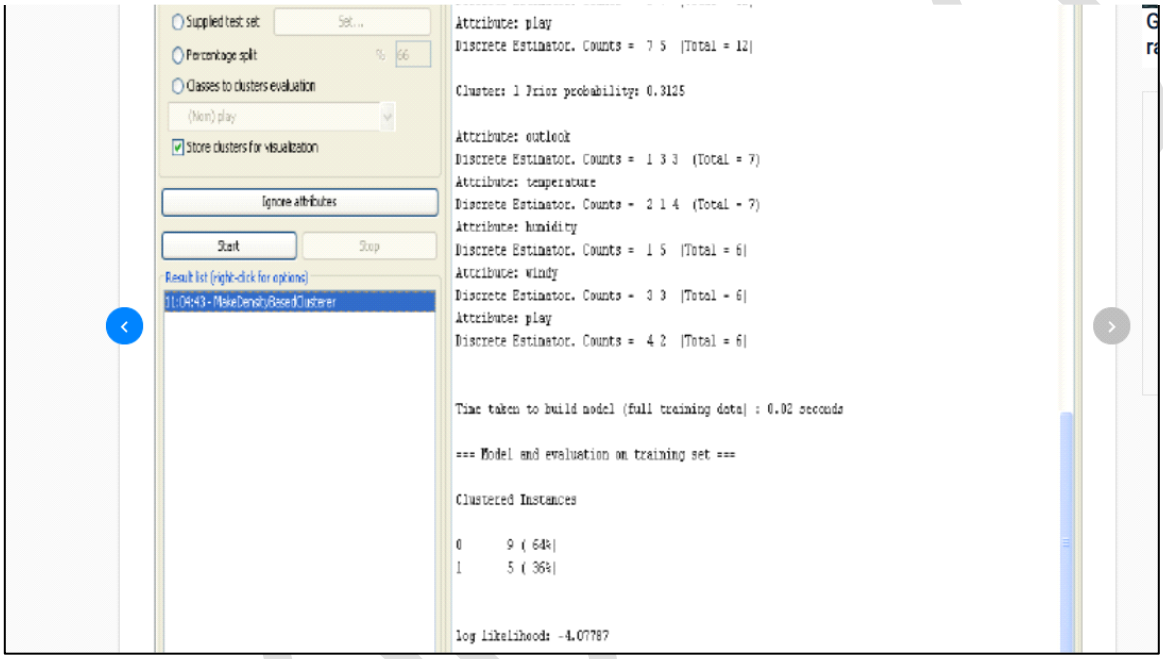
```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
```

sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

The following screenshot shows the clustering rules that were generated when density based clustering algorithm is applied on the given dataset:



Conclusion: The density based clustering is able the cluster the data in the iris database.

*****

Experiment No. 9

Title: Demonstration of any ETL tool

S/w Requirement: (like INFORMATICA, WEKA etc.)

Objective: To understand the concept of ETL(Extract, Transform & Load)

Steps of ETL Process in WEKA are as follows-:

1st Step – Extraction

The first step before you can begin organizing your data is pulling or extracting the data from all the relevant sources and compiling it. These sources may include on-premise databases, CRM systems, marketing automation platforms, unstructured and structured files, cloud applications, and any other data sources you wish to draw insights from via analytical processing.

ETL Extraction Steps

Compile data from relevant sources.

Organize data to make it consistent.

2nd Step – Transformation

Data Transformation is the second step of the ETL process in data warehouse. Here the compiled data is converted, reformatted, and cleansed in the staging area to be fed into the target database in the next step. The transformation step involves executing a series of functions and applying sets of rules to the extracted data, to convert it into a standard format to meet the schema requirements of the target database.

ETL Transformation Steps

Convert data according to the business requirements.

Reformat converted data to a standard format for compatibility.

Cleanse irrelevant data from the datasets.

Sort & Filter data.

Clear duplicate information.

Translate where necessary.

<u>3rd Step – Loading</u>

The concluding step is the act of loading the datasets that've been extracted and transformed earlier, into the target database. There are two ways to go about it; first is a SQL insert routine that involves the manual insertion of each record in every row of your target database table. While, the other loading approach uses a process called bulk load of data, reserved for massive loading of data.

ETL Loading Steps

Load well transformed datasets through bulk loading.

Load questionable datasets through SQL Inserts.

Conclusion: The ETL process is demonstated above.

*****

Experiment No. 10

Title: Case study/ Create a mini project on data set using WEKA software.

S/w Requirement: WEKA

Case Study:

Take data set related to prediction for sales, health diagnosis(heart disease prediction, cancer disease prediction) weather forecasting, credit card fraudulent ….many more.

Perform different operations using WEKA software and observe the results shown in the output window.

Newsgroups Data Set

§ Content and structure:

• approximately 20,000 newsgroup documents

» 19,997 originally

» 18,828 without duplicates

• partitioned evenly across 20 different newsgroups

• we are only using a subset

WEKA can import data from:

• files: ARFF, CSV, binary

• URL

• SQL database

§ Pre-processing tools (filters) are used for:

• Discretization, normalization,

resampling, attribute selection,

transforming and combining attributes,

etc.

§ Classifiers in WEKA are models for:

• classification (predict a nominal class)

• regression (predict a numerical quantity)

§ Learning algorithms:

• Naïve Bayes, decision trees, kNN, support vector machines,

multi-layer perceptron, logistic regression, etc.

§ Meta-classifiers:

• cannot be used alone

• always combined with a learning algorithm

• examples: boosting, bagging etc.

*****