**TRIBHUVAN UNIVERSITY**

**Institute of Engineering**

Kathmandu Engineering College

Department of Computer Engineering



Report On

**Classification of dataset**

By

Sujit Kumar Shestha - KAT075BCT087

Saurav Khakurel - KAT075BCT079

Kathmandu, Nepal

Shrawan 16, 2079

**INTRODUCTION**

**CLASSIFICATION**

Classification is the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories, a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

1. Naive Bayesian Classifier

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is easy and fast to predict class of test data set. It also perform well in multi class prediction. The algorithm for the Naive Bayesian classifier is

Step 1: Convert the data set into a frequency table.

Step 2: Create Likelihood table by finding the probabilities

Step 3: Use Naive Bayesian equation to calculate the posterior probability for each class.

The class with the highest posterior probability is the outcome of prediction.

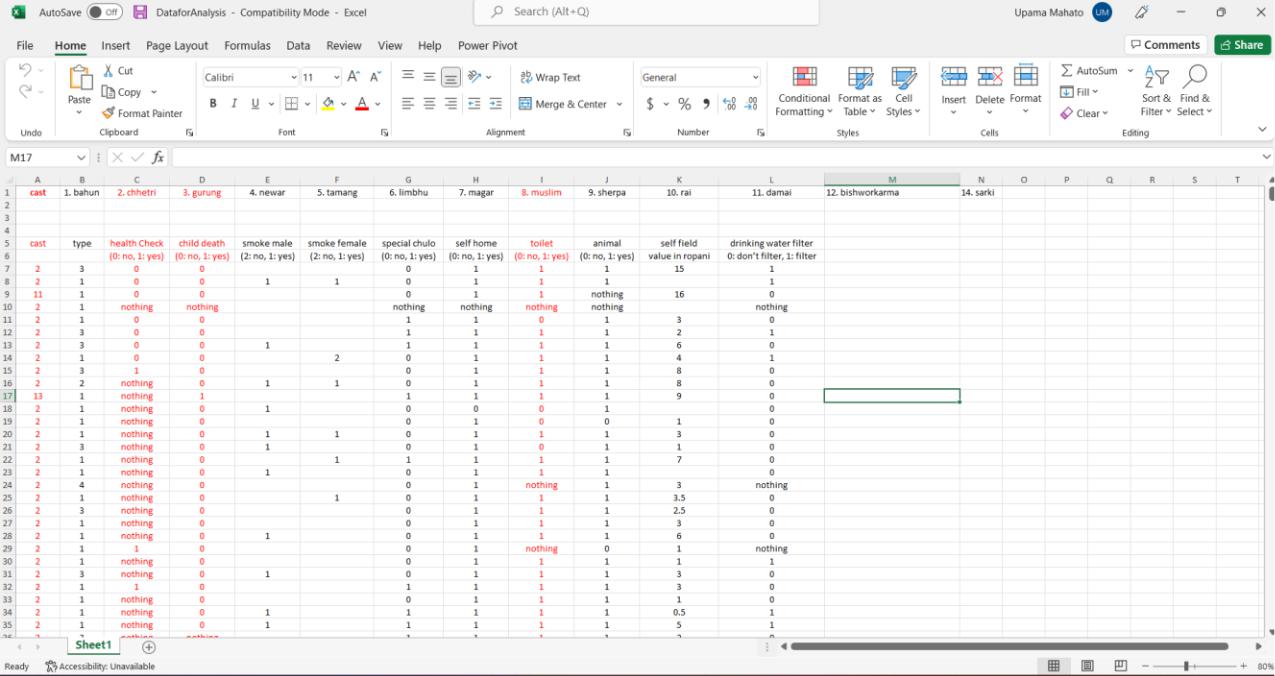
2. Multilayer Perceptron

A multilayer perceptron is a feed-forward artificial neural network that generates a set of outputs from a set of inputs. It is characterized by several layers of input nodes connected as a directed graph between the input and output layers. It uses back-propagation for training the network.

4

**DESCRIPTION OF DATABASE**

The database consists of 1385 data with 12 attributes. Every attribute has its own values. There are some data with either null value or value written as nothing. Also each of the values of attributes are represented by numbers 0 or 1, 0 stands for no and 1 stands for yes. During ARFF file generation, we either use their nominal values or real values.



*Figure 1. Screenshot of Rough database*

5

**KDD STEPS FOLLOWED IN DATABASE**

1. Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.

* Cleaning in case of Missing values.
* Cleaning noisy data, where noise is a random or variance error.
* Cleaning with Data discrepancy detection and Data transformation tools.

1. Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).

* Data integration using Data Migration tools.
* Data integration using Data Synchronization tools.
* Data integration using ETL (Extract-Load-Transformation) process.

1. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. Data selection using neural network, Decision Trees, Clustering, Regression and Naive Bayes.
2. Data Transformation: Data Transformation is defined as the process of transforming data into the appropriate form required by mining procedure.

Data Transformation is a two-step process Data Mapping and Code generation:

1. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

* Transforms task relevant data into patterns.
* Decides purpose of model using classification or characterization.

6

1. Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

* Find interestingness score of each pattern.
* Uses summarization and Visualization to make data understandable by user.

1. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

* Generate reports.
* Generate tables.
* Generate discriminant rules, classification rules, characterization rules, etc.

7

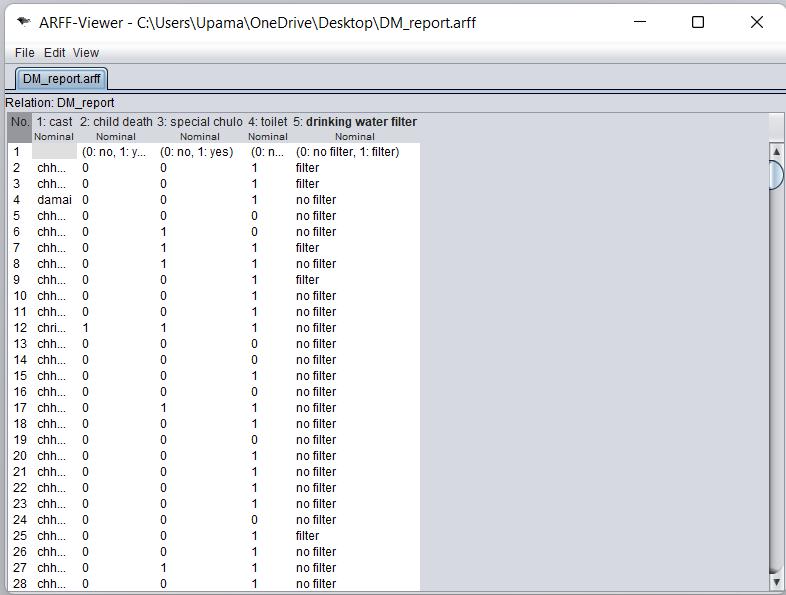
**ARFF FILE DESCRIPTION**

The header section defines the relation name, attribute name and the type of data.

The @relation tag defines the name of the database.

The @attribute tag defines the attributes. It may have nominal values or real values.

The @data tag starts the list of data rows each containing the comma separated fields.



*Figure 2. Screenshot of ARFF file*

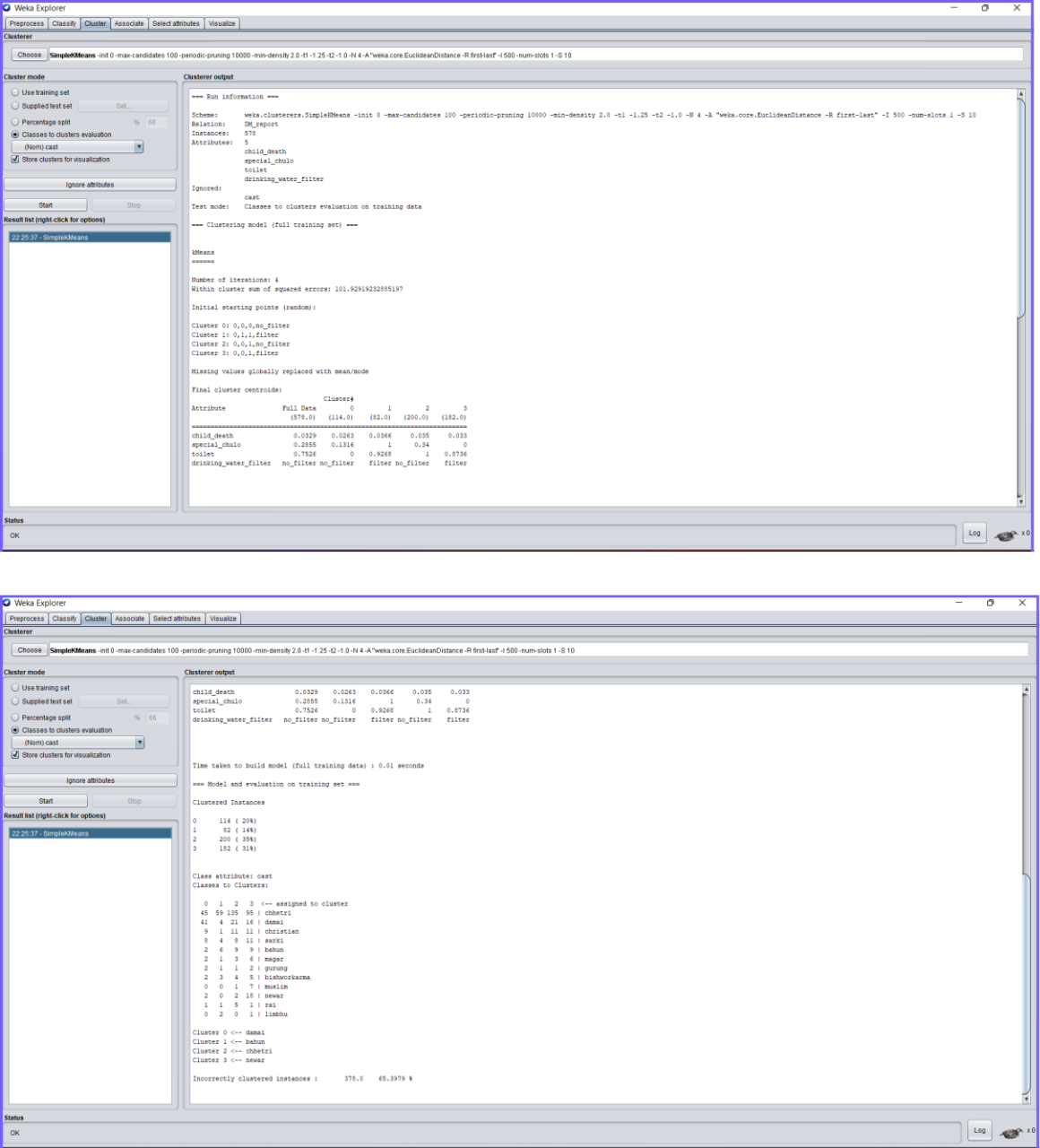
8

**OUTPUT FROM THE ALGORITHM**

**For clustering:**

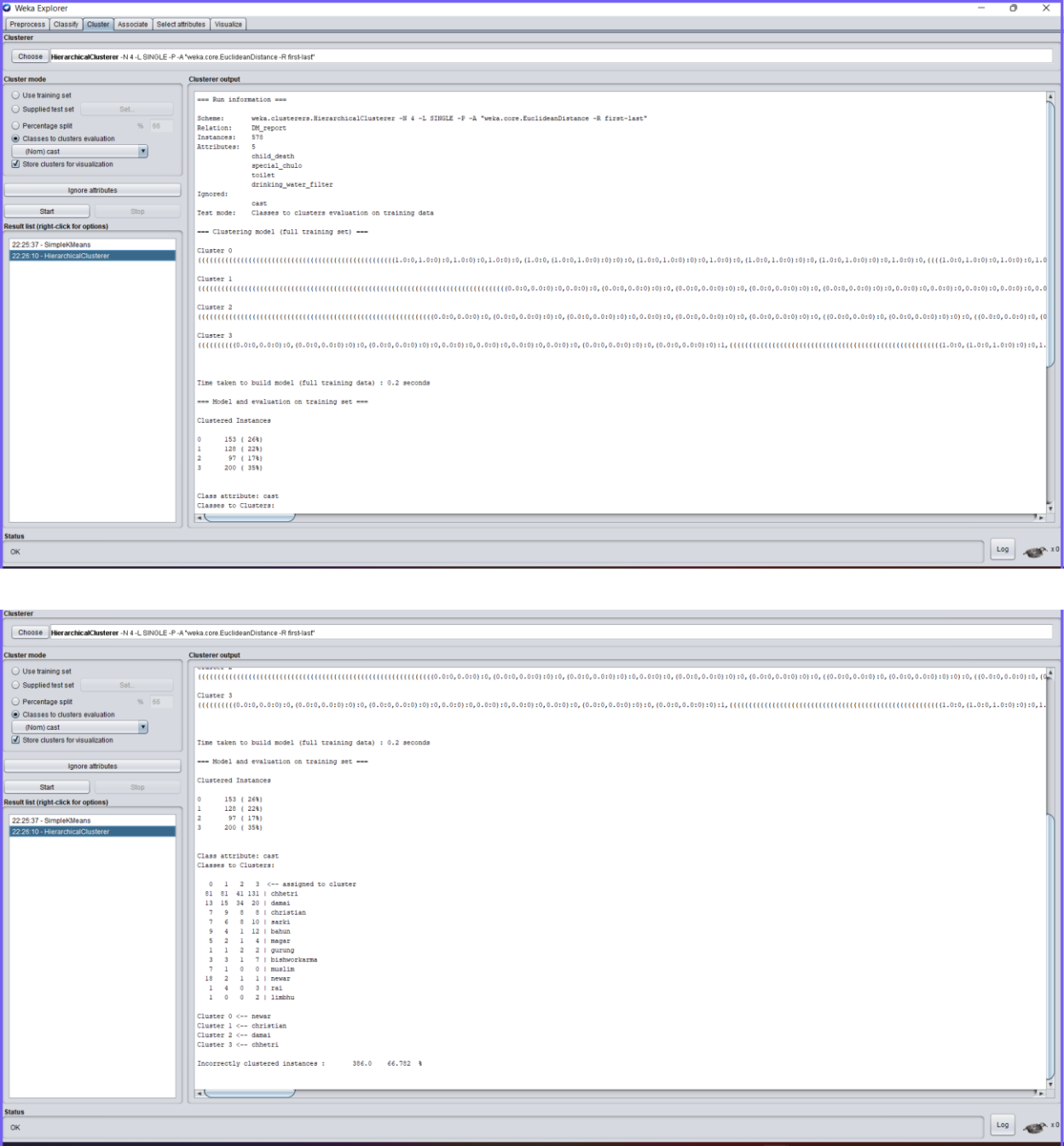
We have used the k-means algorithm and hierarchical method for clustering. In both methods, we have created 4 clusters to analyze the output.

For a simple k-means algorithm with Euclidean Distance for cast:



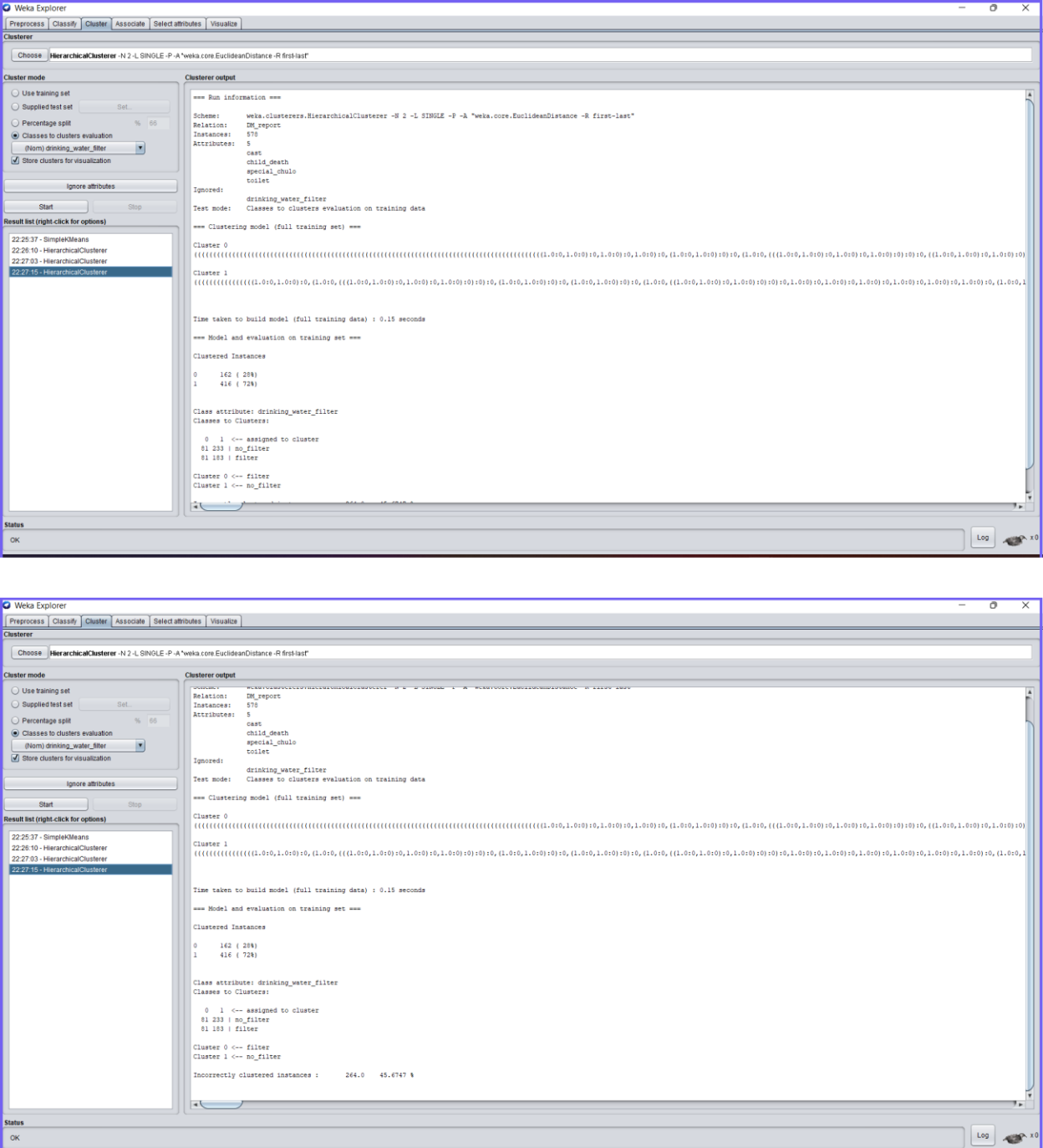
9

For a simple Hierarchical algorithm with Euclidean Distance for cast:



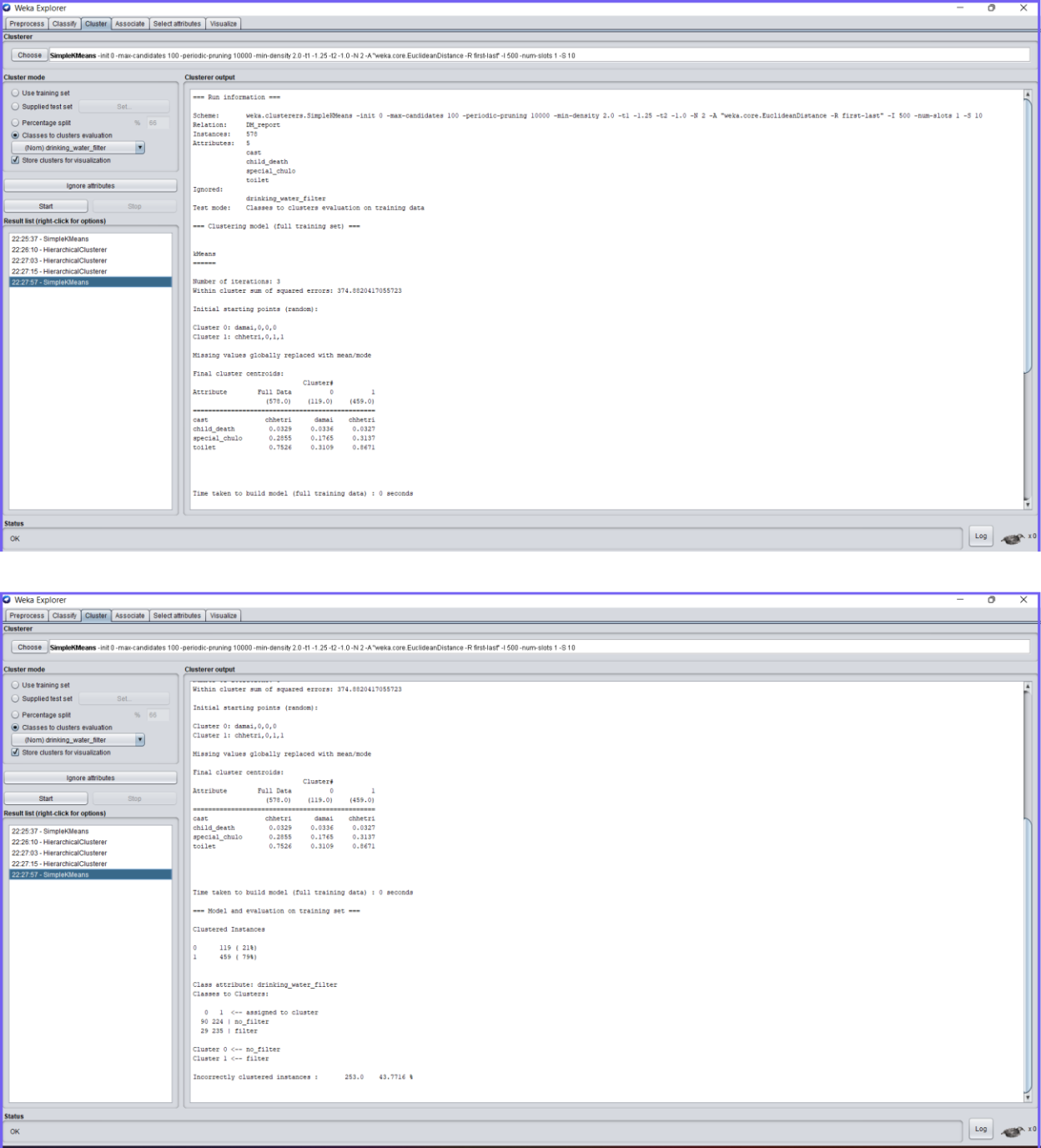
10

For a simple Hierarchical algorithm with Euclidean Distance for drinking\_water\_filter:



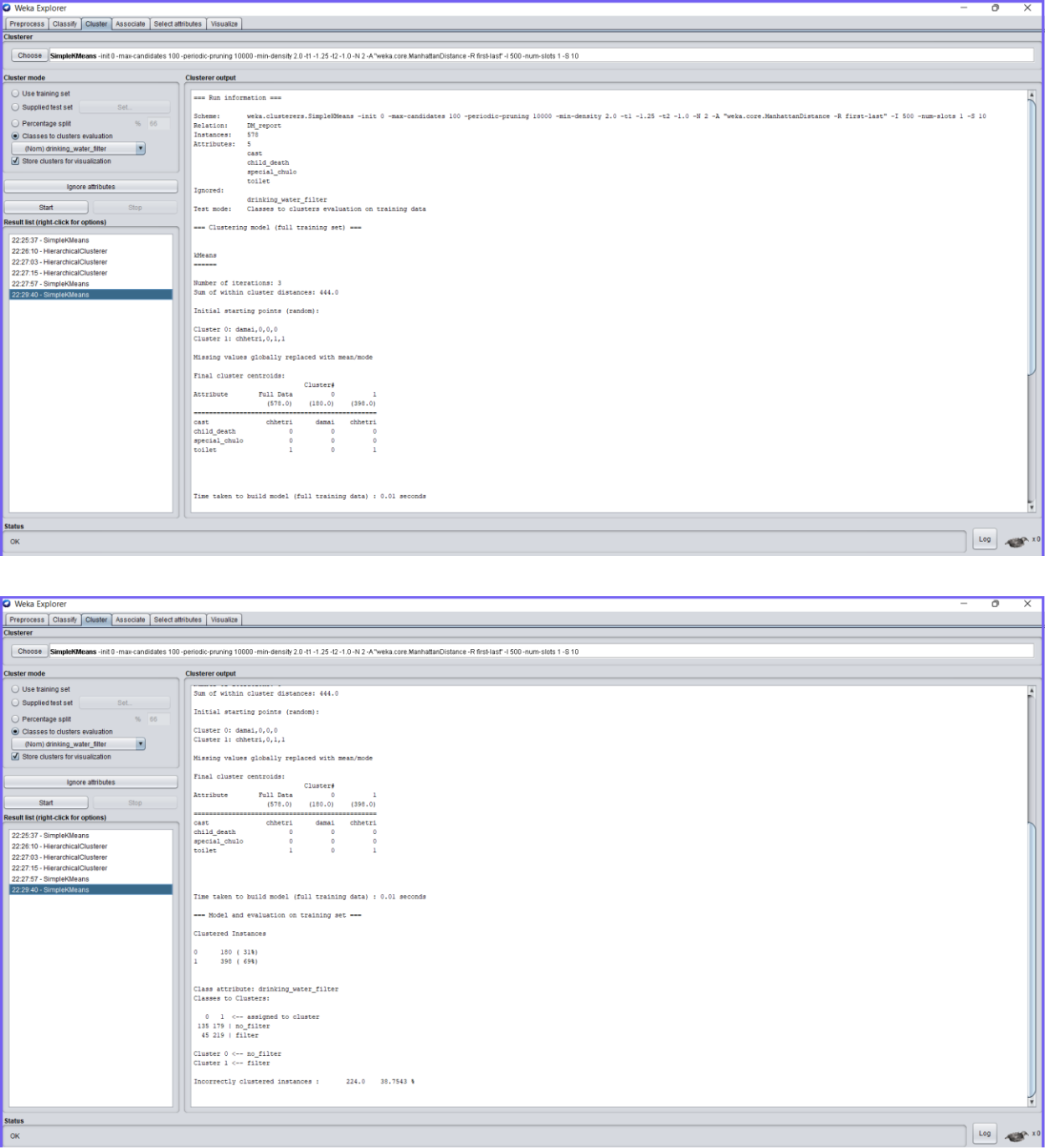
11

For a simple k-means algorithm with Euclidean Distance for drinking\_water\_filter:



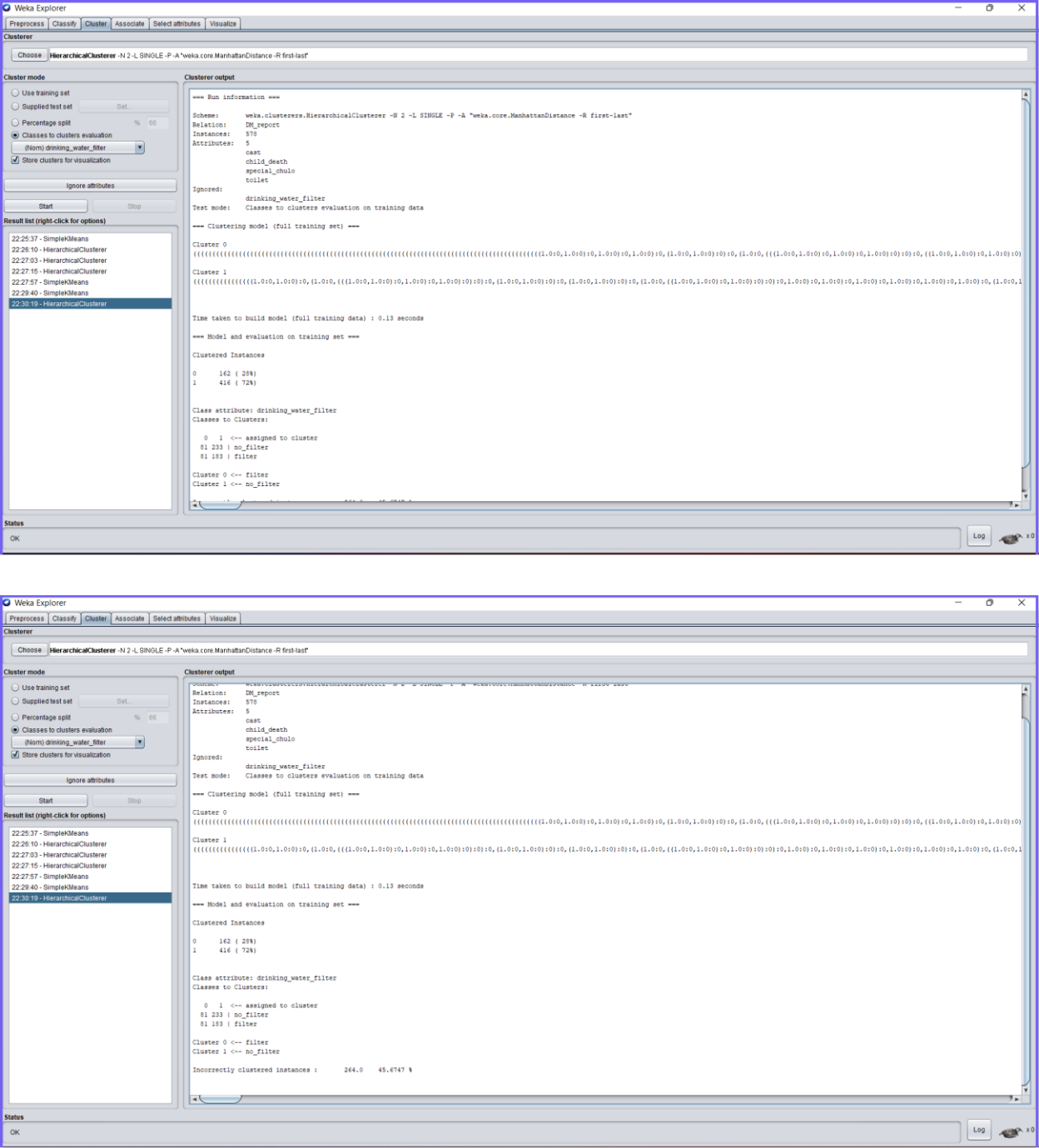
12

For a simple k-means algorithm with Manhattan Distance for the attribute drinking water filter:



13

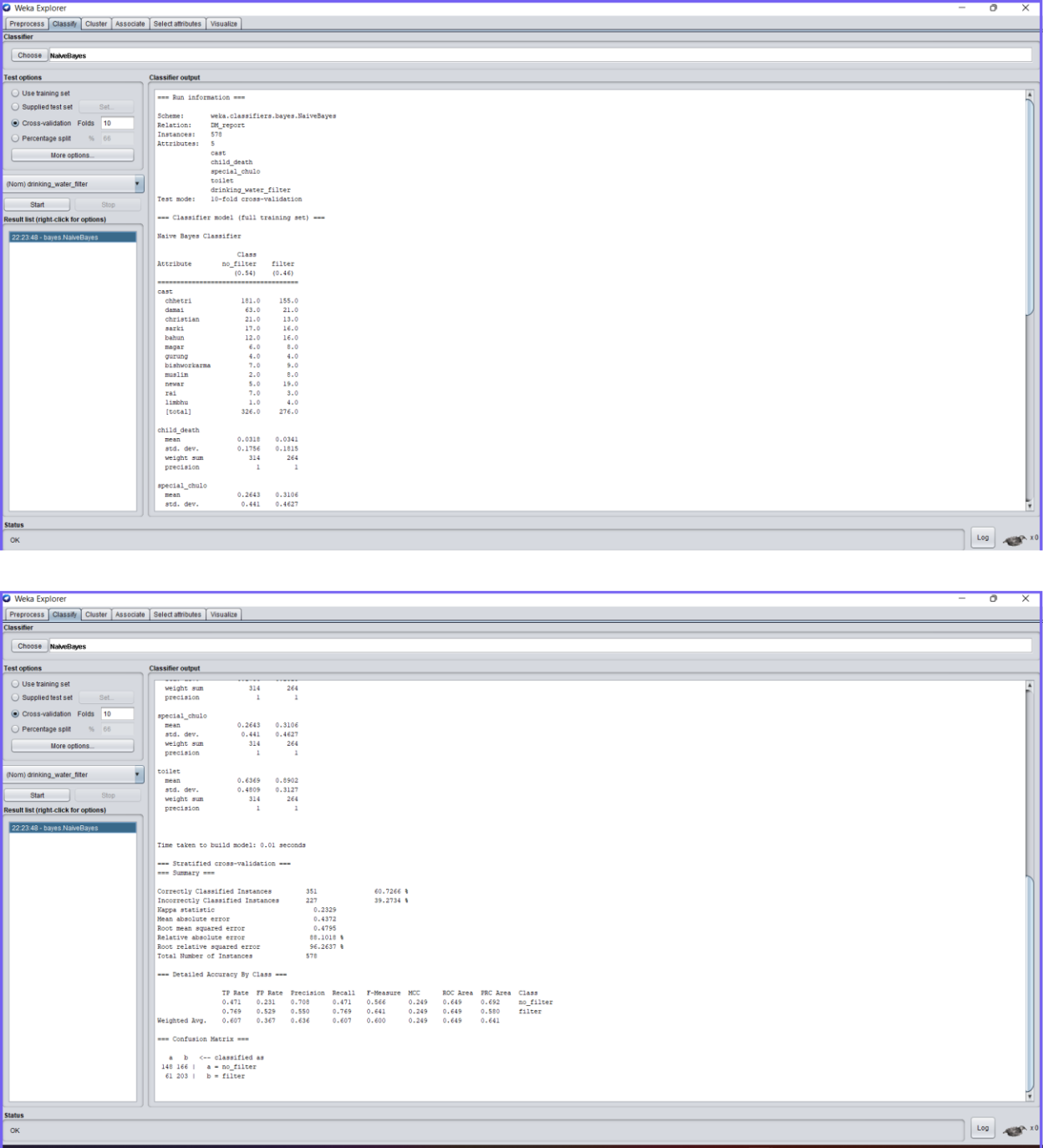
For a Hierarchical algorithm with Manhattan Distance for the attribute drinking water filter:



14

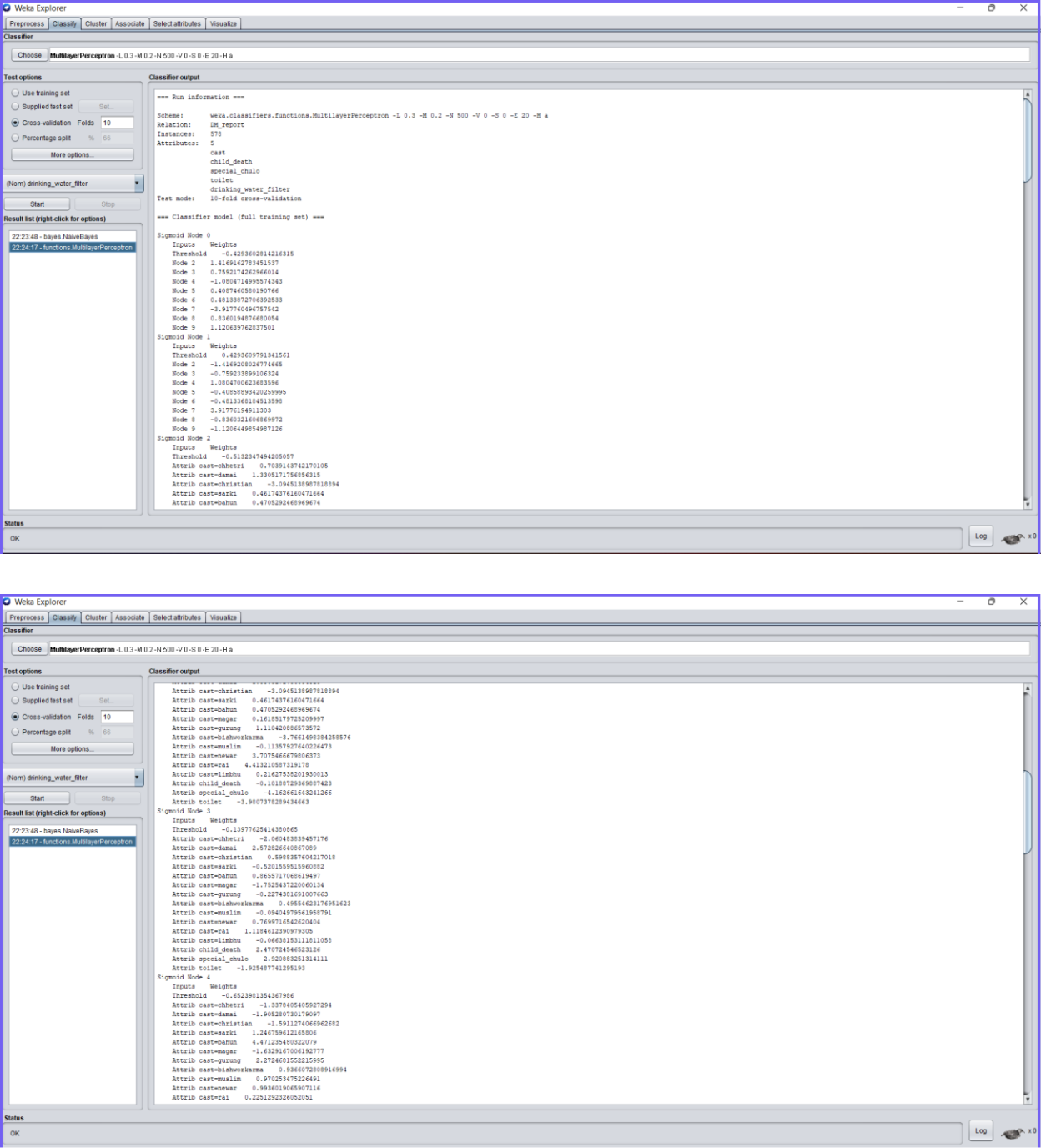
**For classifier:**

We have used the Naive Bayesian algorithm classification.

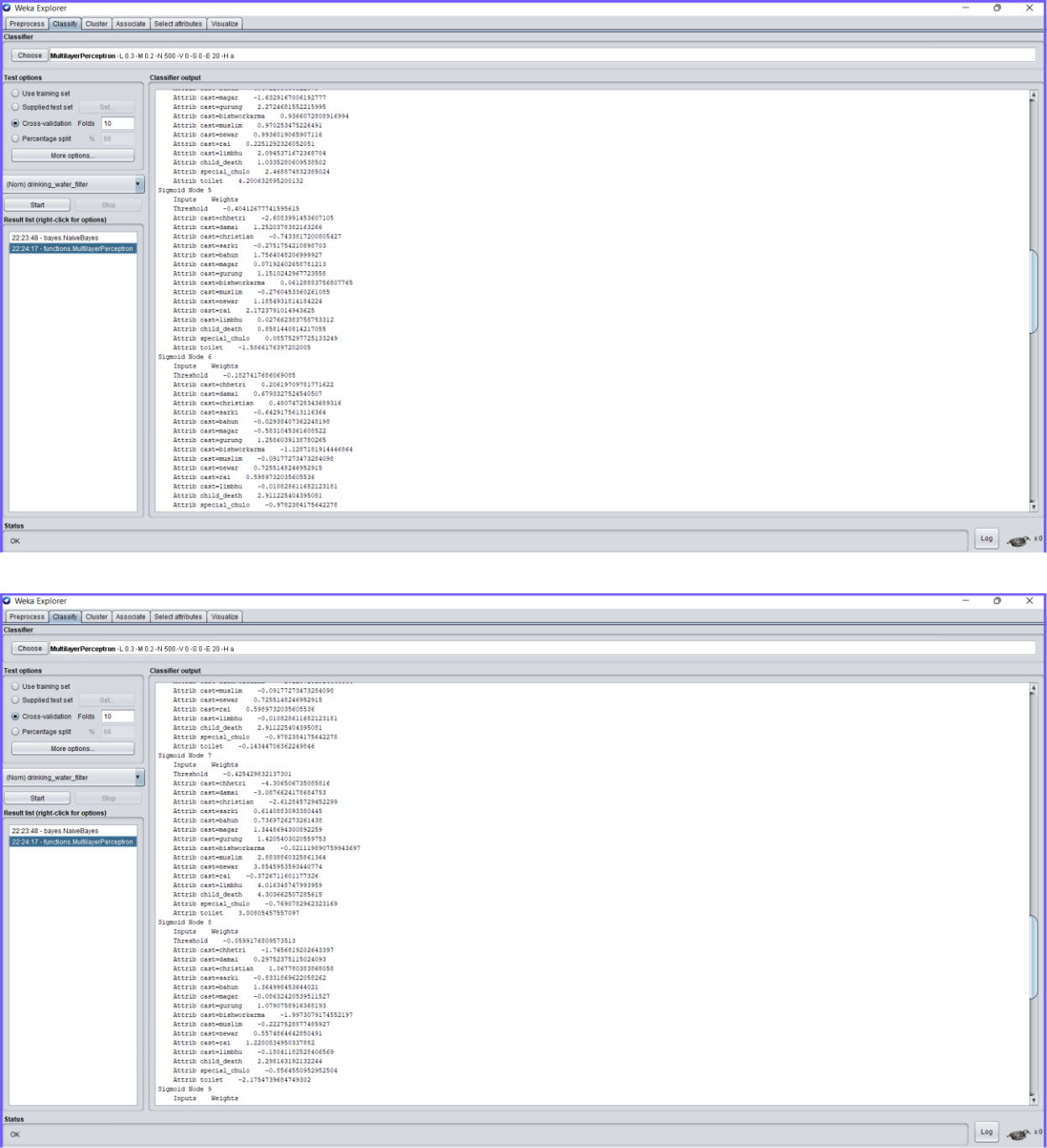


15

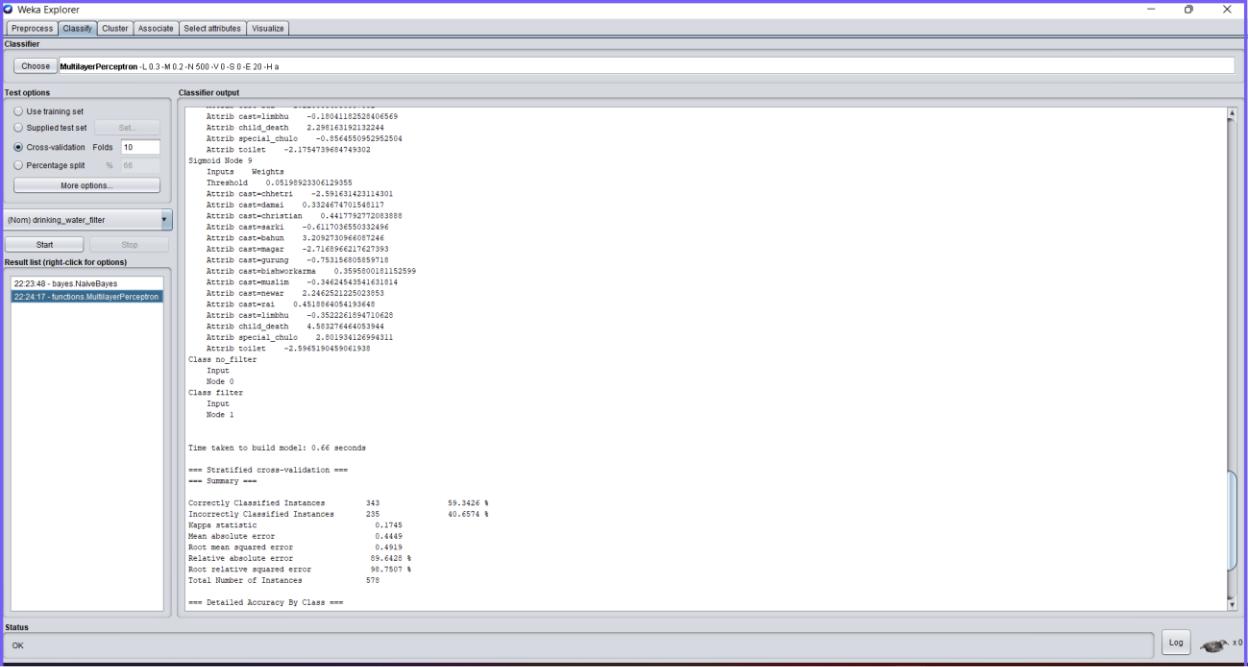
We have used multilayer perceptron method for classification.



16



17



18

**DECISION FROM THE OUTPUT**

**For clustering**

1. Using simple k-means algorithm with Euclidean Distance function:

Through the analysis of the given data using the above algorithm for cast attribute, we

created 4 clusters such that the results are:

Cluster Instances:

Cluster 0 114(20%)

Cluster1 82(14%)

Cluster2 200(35%)

Cluster3 182(31%)

Then,

Cluster 0 => damai

Cluster1 => bahun

Cluster2 => chhetri

Cluster3=> newar

2. Using Hierarchical algorithm with Euclidean Distance function:

Through the analysis of the given data using the algorithm for cast attribute, we created 4

clusters such that the results are:

Cluster Instances:

Cluster 0 153(26%)

Cluster1 128(22%)

Cluster2 97(17%)

Cluster3 200(35%)

Then,

Cluster 0 => newar

Cluster1 => christian

Cluster2 => damai

Cluster3=> chhetri

19

3. Using simple k-means algorithm with Euclidean Distance function:

Through the analysis of the given data using the above algorithm for the drinking\_water\_filter attribute, we created 2 clusters such that the results are: Cluster Instances:

Cluster 0 119(21%)

Cluster1 459(79%)

Then,

Cluster 0 => no\_filter

Cluster1 => filter

4. Using Hierarchical algorithm with Euclidean Distance function:

Through the analysis of the given data using the above algorithm for the drinking\_water\_filter attribute, we created 2 clusters such that the results are: Cluster Instances:

Cluster 0 162(28%)

Cluster1 416(72%)

Then,

Cluster 0 => filter

Cluster1 => no\_filter

5. Using simple k-means algorithm with Manhattan Distance function:

Through the analysis of the given data using the above algorithm for the drinking\_water\_filter attribute, we created 2 clusters such that the results are: Cluster Instances:

Cluster 0 180(31%)

Cluster1 398(69%)

Then,

Cluster 0 => no\_filter

Cluster1 => filter

20

6. Using Hierarchical algorithm with Manhattan Distance function:

Through the analysis of the given data using the above algorithm for the drinking\_water\_filter attribute, we created 2 clusters such that the results are: Cluster Instances:

|  |  |
| --- | --- |
| Cluster 0 | 162(28%) |
| Cluster1 | 416(72%) |

Then,

Cluster 0 => filter

Cluster1 => no\_filter

Now, by comparing a simple k-means algorithm for the attribute drinking\_water\_filter using Euclidean and Manhattan Function we came to the following decision that Manhattan Distance function gave better clustering of data.

Also, by comparing a hierarchical algorithm for attribute drinking\_water\_filter using Euclidean and Manhattan Function we came to the following decision that both distance functions gave us the same result for clustering of data.

Now, comparing simple k-means algorithm and hierarchical algorithm, we came to the decision that using hierarchical algorithm gave us better clustering of data for a given dataset.

**For classification**

1. Using the Naive Bayesian algorithm for the attribute drinking\_water\_filter we got the following confusion matrix:

a

a

148

b

166

b 61 203

a : no\_filter

b : filter

21

1. Using the Multilayer Perceptron algorithm for the attribute drinking\_water\_filter we got the following confusion matrix:

a

a

209

b

105

b 130 134

a : no\_filter

b : filter

In the case of the Naive Bayesian algorithm,

For actual data which uses no\_filter, the prediction result for no\_filter gave us 148.

For the actual data which uses filter but predicted result is for a no\_filter which gave us 61.

For the actual data which uses a no\_filter but prediction result is for filter which gave us 166.

For the actual data which uses the filter, the predicted result is for the filter which gave us 203.

Its accuracy is:

(148+203)/578

* 0.6072

=60.72%

In the case of the Multilayer Perceptron algorithm,

For actual data which uses no\_filter, the prediction result for no\_filter gave us 209.

For the actual data which uses filter but predicted result is for a no\_filter which gave us 130.

For the actual data which uses a no\_filter but prediction result is for filter which gave us 105.

22

For the actual data which uses the filter, the predicted result for the filter which gave us 134.

Its accuracy is:

(209+134)/578

=0.5934

=59.34%

From the accuracy calculated above, we came to decision that Naive Bayesian algorithm gave better result compare to Multilayer Perceptron algorithm for the given dataset.

**CONCLUSION**

Hence, from this lab, we learned about different ways of clustering and classifying a dataset. Depending upon the type of algorithm used, the distance function selected and various other parameters, the result obtained will be different. From our work, by contrasting a straightforward k-means approach for the attribute drinking\_water\_filter utilizing the Manhattan Function and the Euclidean Function, we arrived at the conclusion that the Manhattan Distance function provided better data clustering. Additionally, by contrasting a hierarchical approach for the attribute drinking\_water\_filter utilizing the Manhattan and Euclidean functions, we determined that both distance functions produced the same outcome. Now that we have compared the hierarchical method to the standard k-means approach, we have determined that the hierarchical algorithm provided us with better data clustering for a specific dataset. We also determined that, for the provided dataset, the Naive Bayesian method performed better than the Multilayer Perceptron technique.

23