Name:	Prerna Sunil Jadhav
Sap Id:	60004220127
Class:	T. Y. B.Tech (Computer Engineering)
Course:	Data Mining and Warehouse Laboratory
Course Code:	DJ19CEL501
Experiment No.:	01

AIM: Perform data Pre-processing task using Weka data mining tool

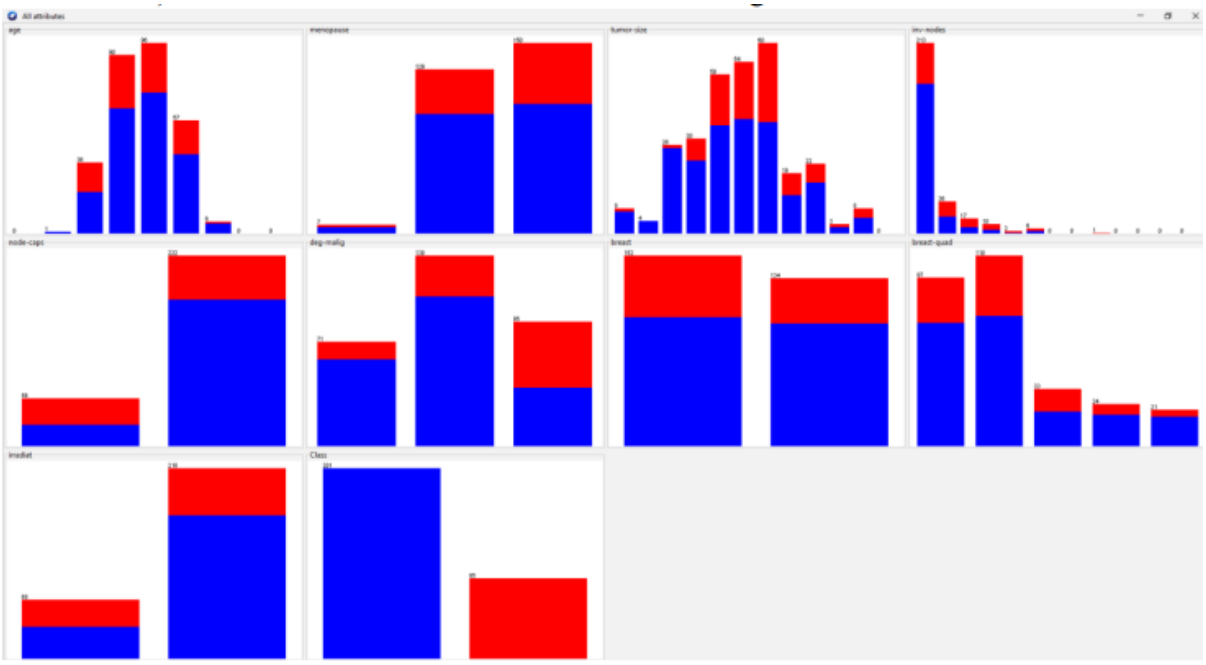
THEORY:
WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems

TASKS PERFORMED THROUGH WEKA:

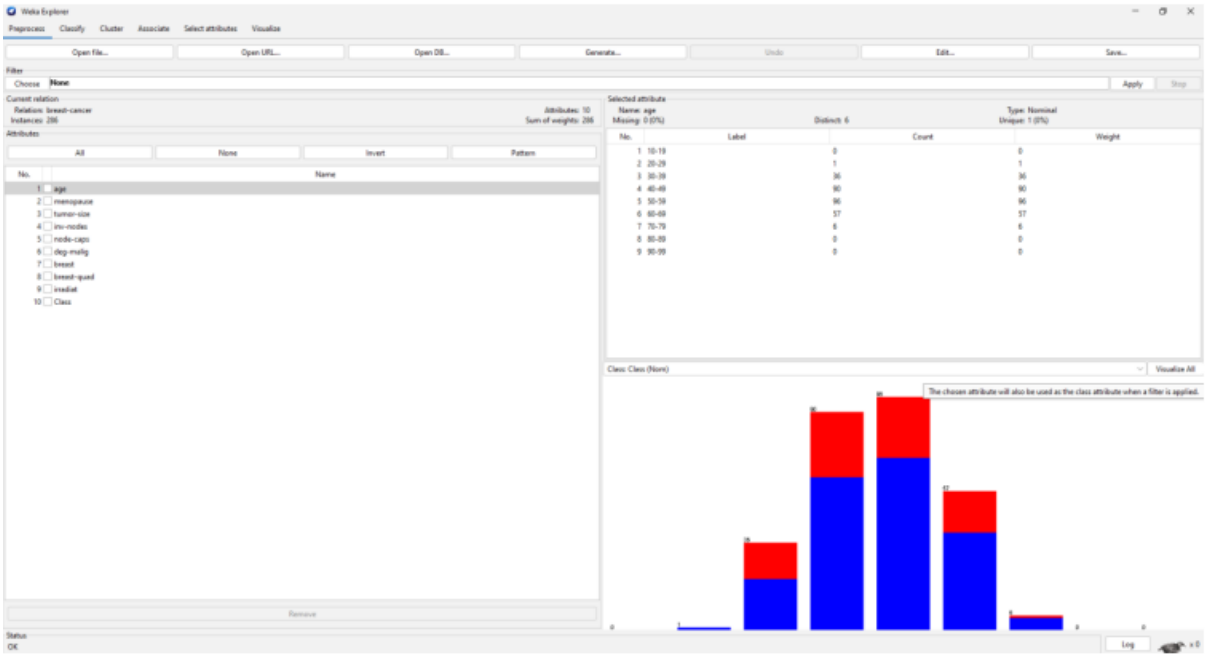
- PREPROCESSING:

Procedure:

a. Visualize All: Select this button to visualize histograms of all attributes.

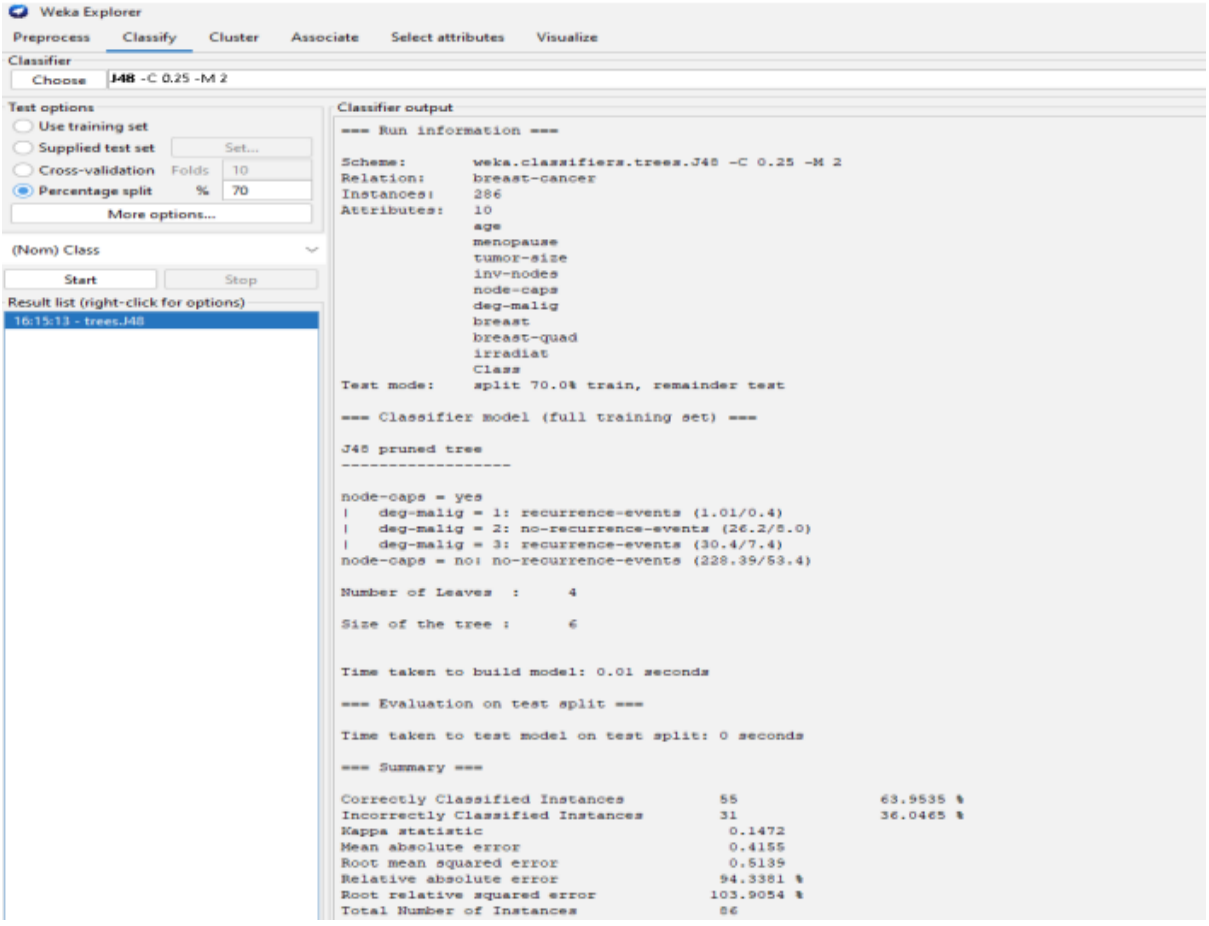


b. Filter: Choose Discretization under Unsupervised and Supervised methods. Observe the discretization and the outliers.
c. IQR: Observe the IQR values for a selected attribute. Observe the outlier and extreme values
d. Remove the value: Remove instances with outlier values and show the screenshots of dataset before and after the removal.



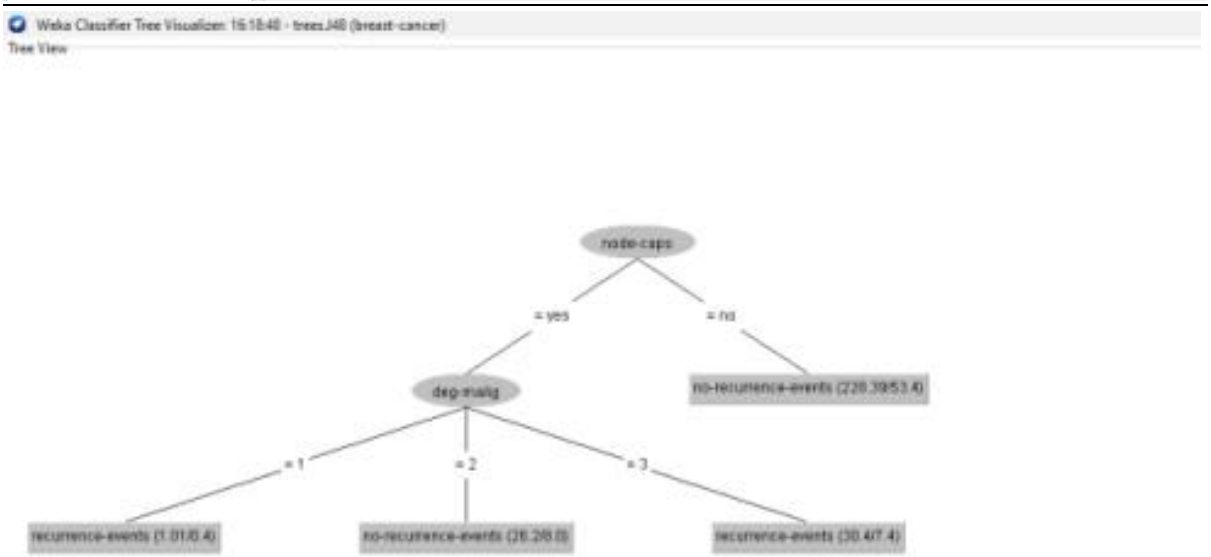
- CLASSIFICATION:

PROCEDURE: Perform NB, kNN and DT/rule based classification



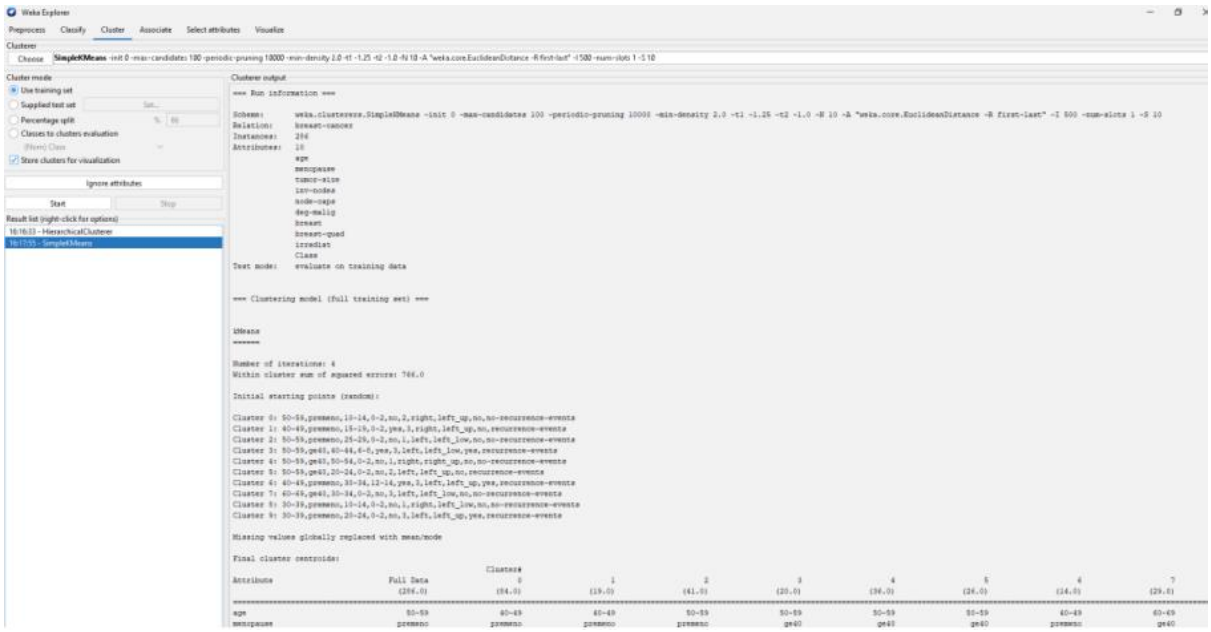
```
Weka Explorer
Preprocess  Classify  Cluster  Associate  Select attributes  Visualize

Classifier
Choose  J48 -C 0.25 -M 2

Test options                          Classifier output
Use training set                      === Run information ===
Supplied test set   Set...
Cross-validation    Folds  10         Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Percentage split    %      70         Relation:     breast-cancer
More options...                       Instances:    286
                                      Attributes:   10
(Nom) Class                                         age
                                                    menopause
Start          Stop                                 tumor-size
                                                    inv-nodes
Result list (right-click for options)               node-caps
16:15:13 - trees.J48                                deg-malig
                                                    breast
                                                    breast-quad
                                                    irradiat
                                                    Class
                                      Test mode:    split 70.0% train, remainder test

                                      === Classifier model (full training set) ===

                                      J48 pruned tree
                                      ------------------

                                      node-caps = yes
                                      |   deg-malig = 1: recurrence-events (1.01/0.4)
                                      |   deg-malig = 2: no-recurrence-events (26.2/8.0)
                                      |   deg-malig = 3: recurrence-events (30.4/7.4)
                                      node-caps = no: no-recurrence-events (228.39/53.4)

                                      Number of Leaves  :     4

                                      Size of the tree :     6


                                      Time taken to build model: 0.01 seconds

                                      === Evaluation on test split ===

                                      Time taken to test model on test split: 0 seconds

                                      === Summary ===

                                      Correctly Classified Instances          55               63.9535 %
                                      Incorrectly Classified Instances        31               36.0465 %
                                      Kappa statistic                          0.1472
                                      Mean absolute error                      0.4155
                                      Root mean squared error                  0.5139
                                      Relative absolute error                 94.3381 %
                                      Root relative squared error            103.9054 %
                                      Total Number of Instances               86
```

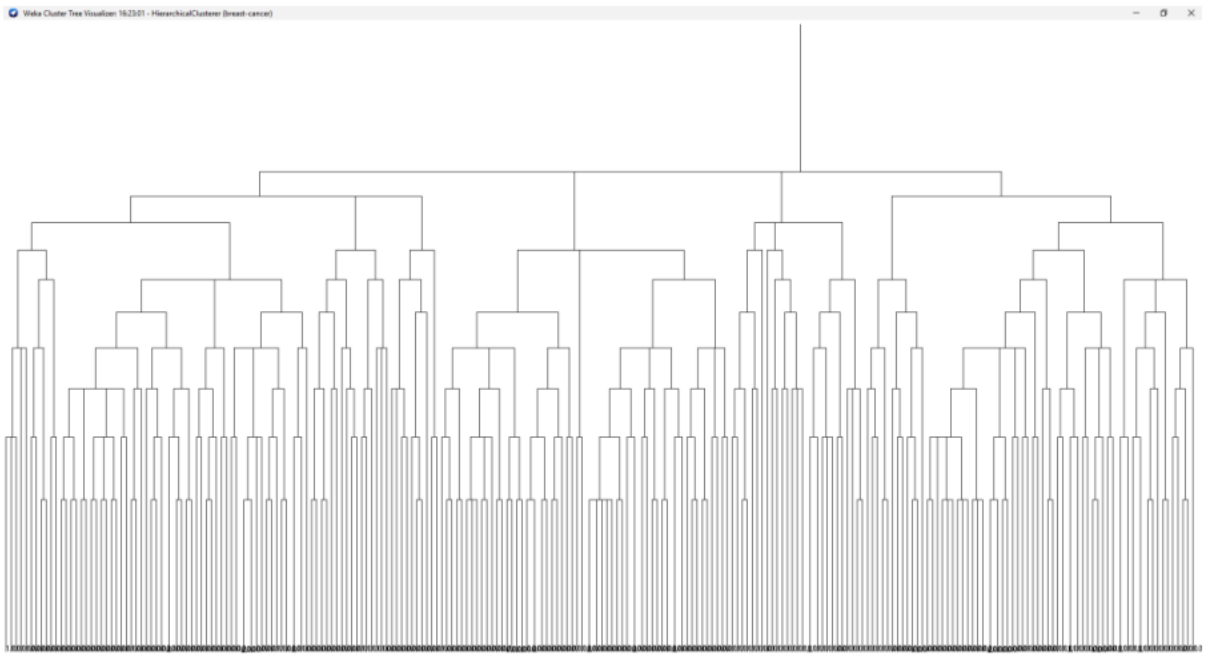
```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.852    0.719      0.667      0.852    0.748      0.162    0.573     0.680     no-recurrence-events
                 0.281    0.148      0.529      0.281    0.367      0.162    0.573     0.429     recurrence-events
Weighted Avg.    0.640    0.506      0.616      0.640    0.606      0.162    0.573     0.587

=== Confusion Matrix ===

  a  b   <-- classified as
 46  8 |  a = no-recurrence-events
 23  9 |  b = recurrence-events
```



- CLUSTERING:

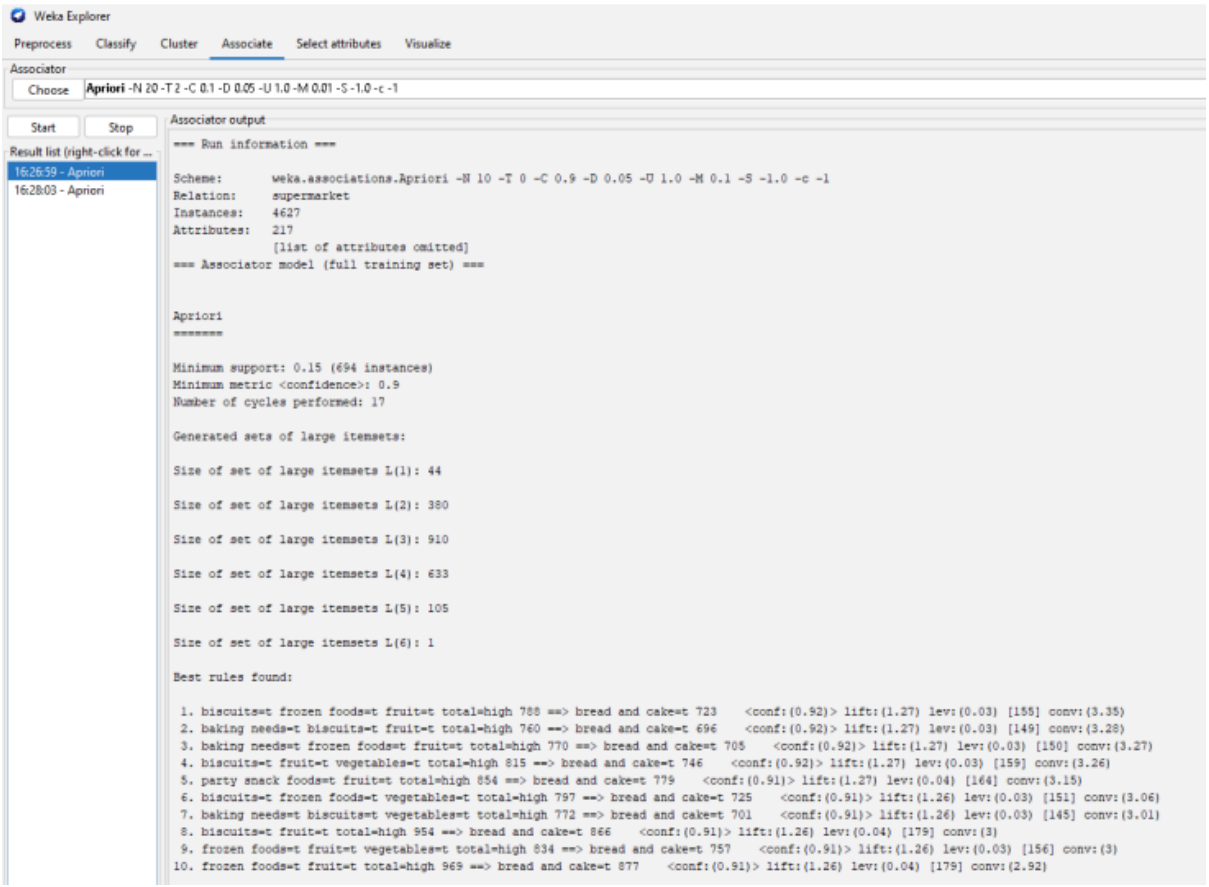
PROCEDURE: Perform kmeans, hierarchical clustering and explain the output


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
16:16:33 - HierarchicalClusterer
16:17:55 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: breast-cancer
Instances: 286
Attributes: 10
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Class
Test mode: evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 4
Within cluster sum of squared errors: 766.0
Initial starting points (random):
Cluster 0: 50-59,premeno,10-14,0-2,no,2,right,left_up,no,no-recurrence-events
Cluster 1: 40-49,premeno,15-19,0-2,yes,3,right,left_up,no,recurrence-events
Cluster 2: 50-59,premeno,25-29,0-2,no,1,left,left_low,no,no-recurrence-events
Cluster 3: 50-59,ge40,40-44,0-2,yes,3,left,left_low,yes,recurrence-events
Cluster 4: 50-59,ge40,50-54,0-2,no,1,right,right_up,no,no-recurrence-events
Cluster 5: 50-59,ge40,20-24,0-2,no,2,left,left_up,no,recurrence-events
Cluster 6: 40-49,premeno,30-34,12-14,yes,3,left,left_up,yes,recurrence-events
Cluster 7: 40-49,ge40,30-34,0-2,no,3,left,left_low,no,no-recurrence-events
Cluster 8: 30-39,premeno,10-14,0-2,no,1,right,left_low,no,no-recurrence-events
Cluster 9: 30-39,premeno,20-24,0-2,no,3,left,left_up,yes,recurrence-events
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1 2 3 4 5 6 7
(286.0) (84.0) (19.0) (41.0) (20.0) (36.0) (26.0) (14.0) (29.0)
age 50-59 40-49 40-49 50-59 50-59 50-59 50-59 40-49 40-49
menopause premeno premeno premeno premeno ge40 ge40 ge40 premeno ge40


Weka Cluster Tree Visualizer: 16:23:01 - HierarchicalClusterer (breast-cancer)

ASSOCIATION RULE:

PROCEDURE: Perform apriori algorithm and show the rules created



SELECT ATTRIBUTES:

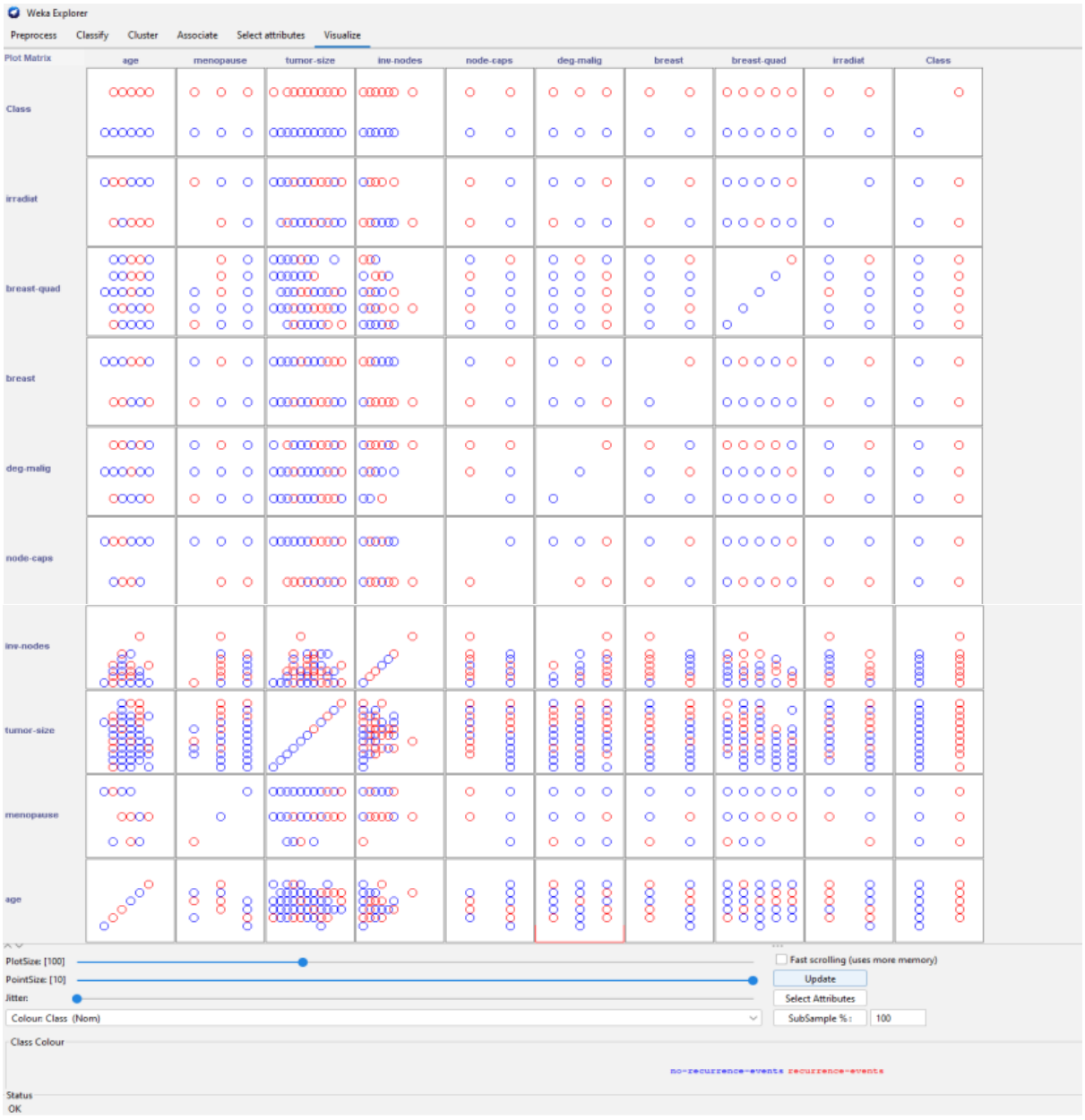
PROCEDURE:

a. Apply suitable feature selection filter like GainRatio etc. to choose relevant attributes from the list of attributes.
b. Observe the ranks / priority provided by the filter.

- VISUALIZATION:

PROCEDURE:

a. Visualize scatter plot for all the attributes from a dataset selected from Weka.
b. Determine correlation if any using these plots for different datasets



CONCLUSION:

- the experiment involving data pre-processing using the Weka data mining tool has been a valuable and essential step in preparing data for subsequent analysis.
- Weka provides a wide range of functionalities that aid in cleaning, transforming, and organizing data, making it more suitable for data mining and machine learning tasks.