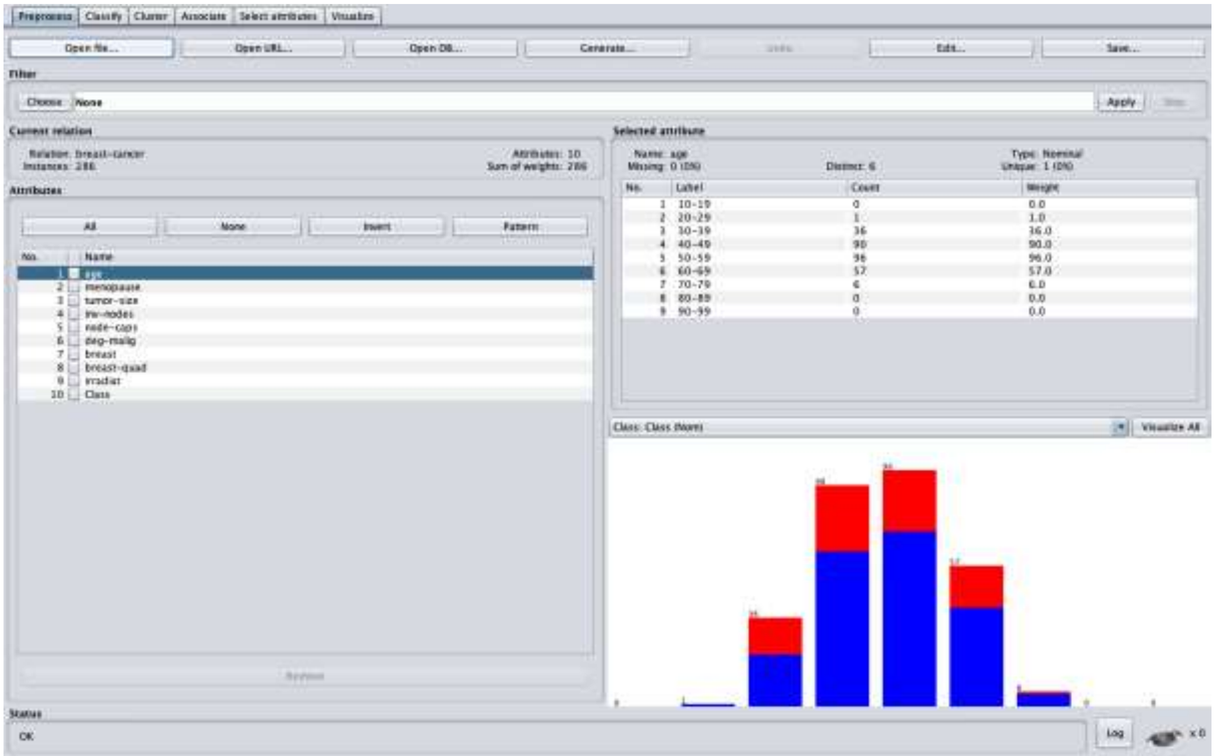
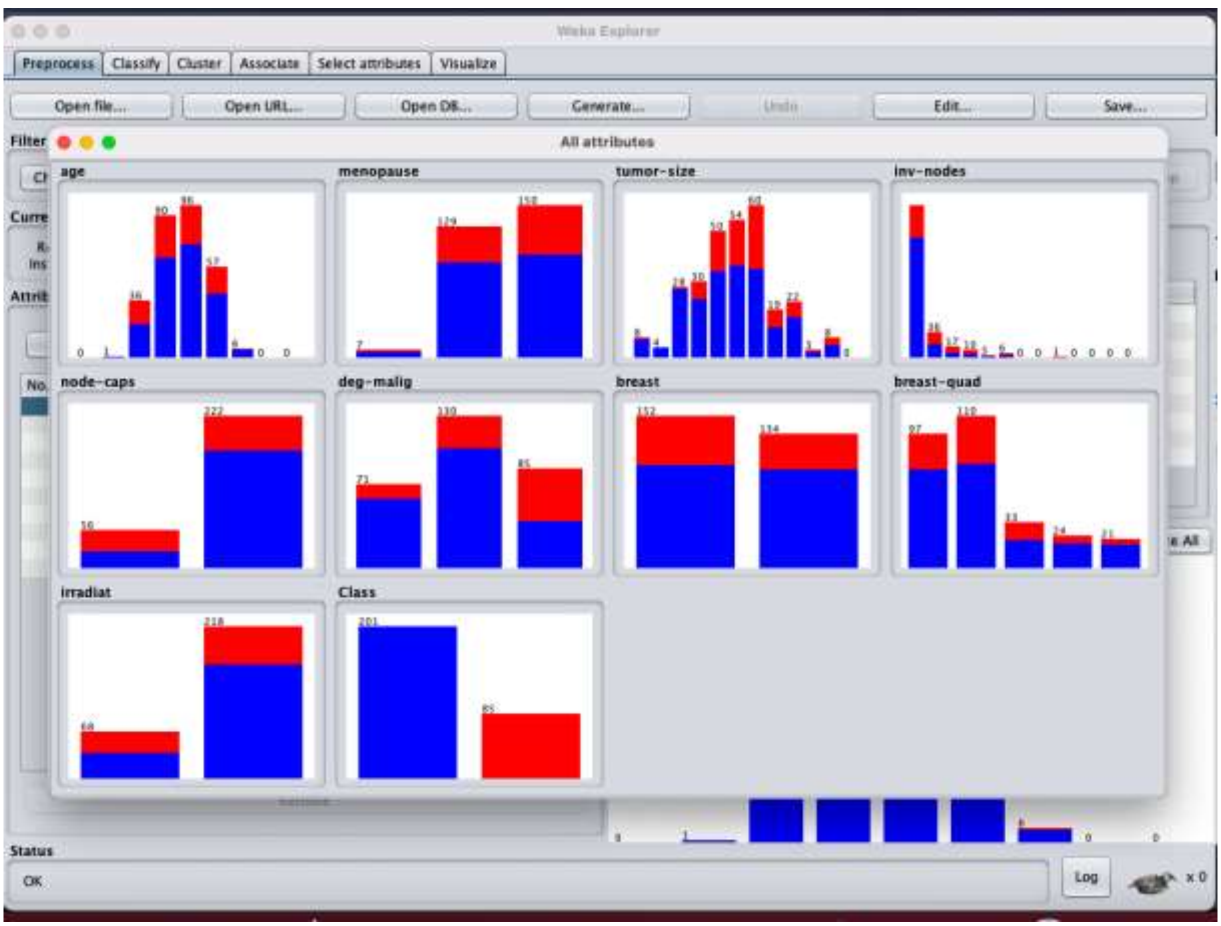
Name: Mareena Fernnades

TE IT	Roll number : 8669
Expt. number : 8	Date of implementation:10/05/2021
Aim : To implement Apriori algorithm on large dataset using WEKA (data mining tool)	
Related Course outcome : CO3 Upon completion of this course students will be able to evaluate the performance of different data mining algorithms using latest tools	
Theory : WEKA contains an implementation of the Apriori algorithm. The algorithm works only with discrete data. It can identify statistical dependencies between groups of attributes. Apriori algorithm can compute all rules that have a given minimum support and exceed a given confidence. Clicking on the "Associate" tab will bring up the interface for the association rule algorithms. The Apriori algorithm which we will use is the default algorithm selected. However, in order to change the parameters for this run (e.g., support, confidence, etc.) we click on the text box immediately to the right of the "Choose" button. Note that this box, at any given time, shows the specific command line arguments that are to be used for the algorithm. WEKA allows the resulting rules to be sorted according to different metrics such as confidence, leverage, and lift. We can also change the default value of rules (10) to be 20; this indicates that the program will report no more than the top 20 rules. The upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%). Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments which by default is set to 0.05 or 5%). The algorithm halts when either the specified number of rules are generated, or the lower bound for min. support is reached. Once the parameters have been set, the command line text box will show the new command line. We now click on start to run the program. This results in a set of rules. The panel on the left ("Result list") now shows an item indicating the algorithm that was run and the time of the run. You can perform multiple runs in the same session each time with different parameters. Each run will appear as an item in the Result list panel. Clicking on one of the results in this list will bring up the details of the run, including the discovered rules in the right panel. In addition, right-clicking on the result set allows us to save the result buffer into a separate file. Note that the rules were discovered based on the specified threshold values for support and lift. For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction. In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the "quality" of the rule. However, the real value of a rule, in terms of usefulness and action ability is subjective and depends heavily of the particular domain and business objectives.	

Dataset:





Output using Apriori:

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: breast-cancer
Instances: 286
Attributes: 10
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Class
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.5 (143 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 4

Size of set of large itemsets L(4): 1

Best rules found:

1. inv-nodes=0-2 irradiat=no Class=no-recurrence-events 147 ==> node-caps=no 145 <conf:(0.99)> lift:(1.27) lev:(0.11) [30] conv:(10.97)
2. inv-nodes=0-2 irradiat=no 183 ==> node-caps=no 177 <conf:(0.97)> lift:(1.25) lev:(0.12) [34] conv:(5.85)

3. node-caps=no irradiat=no Class=no-recurrence-events 151 ==> inv-nodes=0-2 145 <conf:(0.96)> lift:(1.29) lev:(0.11) [32] conv:(5.51)
4. inv-nodes=0-2 Class=no-recurrence-events 167 ==> node-caps=no 160 <conf:(0.96)> lift:(1.23) lev:(0.11) [30] conv:(4.67)
5. inv-nodes=0-2 213 ==> node-caps=no 201 <conf:(0.94)> lift:(1.22) lev:(0.12) [35] conv:(3.67)
6. node-caps=no irradiat=no 188 ==> inv-nodes=0-2 177 <conf:(0.94)> lift:(1.26) lev:(0.13) [36] conv:(4)
7. node-caps=no Class=no-recurrence-events 171 ==> inv-nodes=0-2 160 <conf:(0.94)> lift:(1.26) lev:(0.11) [32] conv:(3.64)
8. irradiat=no Class=no-recurrence-events 164 ==> node-caps=no 151 <conf:(0.92)> lift:(1.19) lev:(0.08) [23] conv:(2.62)
9. inv-nodes=0-2 node-caps=no Class=no-recurrence-events 160 ==> irradiat=no 145 <conf:(0.91)> lift:(1.19) lev:(0.08) [23] conv:(2.38)
10. node-caps=no 222 ==> inv-nodes=0-2 201 <conf:(0.91)> lift:(1.22) lev:(0.12) [35] conv:(2.58)

Output using FilteredAssociator:

```
=== Run information ===
Scheme:       weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F
\"weka.filters.unsupervised.attribute.ReplaceMissingValues \"" -c -1 -W
weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
=== Associator model (full training set) ===
FilteredAssociator using weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
-c -1 on data filtered through weka.filters.MultiFilter -F
"weka.filters.unsupervised.attribute.ReplaceMissingValues "
Filtered Header
@relation breast-cancer-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.MultiFilter-Fweka.filters.unsupervised.attribute.ReplaceMissingValues

@attribute age {10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99}
@attribute menopause {lt40,ge40,premeno}
@attribute tumor-size {0-4,5-9,10-14,15-19,20-24,25-29,30-34,35-39,40-44,45-49,50-54,55-59}
@attribute inv-nodes {0-2,3-5,6-8,9-11,12-14,15-17,18-20,21-23,24-26,27-29,30-32,33-35,36-
39}
@attribute node-caps {yes,no}
@attribute deg-malig {1,2,3}
@attribute breast {left,right}
@attribute breast-quad {left_up,left_low,right_up,right_low,central}
@attribute irradiat {yes,no}
@attribute Class {no-recurrence-events,recurrence-events}

@data
Associator Model
Apriori
=======

Minimum support: 0.35 (100 instances)
```

Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:
Size of set of large itemsets L(1): 10
Size of set of large itemsets L(2): 20

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 2

Best rules found:

1. inv-nodes=0-2 breast=left irradiat=no 101 ==> node-caps=no 100 <conf:(0.99)> lift:(1.23) lev:(0.07) [18] conv:(9.89)
2. inv-nodes=0-2 irradiat=no Class=no-recurrence-events 147 ==> node-caps=no 145 <conf:(0.99)> lift:(1.23) lev:(0.09) [26] conv:(9.59)
3. inv-nodes=0-2 breast=left 115 ==> node-caps=no 113 <conf:(0.98)> lift:(1.22) lev:(0.07) [20] conv:(7.51)
4. inv-nodes=0-2 irradiat=no 183 ==> node-caps=no 179 <conf:(0.98)> lift:(1.22) lev:(0.11) [31] conv:(7.17)
5. inv-nodes=0-2 Class=no-recurrence-events 167 ==> node-caps=no 161 <conf:(0.96)> lift:(1.2) lev:(0.09) [26] conv:(4.67)
6. node-caps=no breast=left irradiat=no 104 ==> inv-nodes=0-2 100 <conf:(0.96)> lift:(1.29) lev:(0.08) [22] conv:(5.31)
7. node-caps=no irradiat=no Class=no-recurrence-events 151 ==> inv-nodes=0-2 145 <conf:(0.96)> lift:(1.29) lev:(0.11) [32] conv:(5.51)
8. inv-nodes=0-2 213 ==> node-caps=no 204 <conf:(0.96)> lift:(1.19) lev:(0.11) [32] conv:(4.17)
9. menopause=premeno inv-nodes=0-2 112 ==> node-caps=no 106 <conf:(0.95)> lift:(1.18) lev:(0.06) [15] conv:(3.13)
10. node-caps=no irradiat=no 190 ==> inv-nodes=0-2 179 <conf:(0.94)> lift:(1.26) lev:(0.13) [37] conv:(4.04)

Post-lab Questions:

1. Give the difference between maximal frequent item set and closed frequent item set

Ans:

Maximal Frequent Itemsets	Closed Frequent Itemsets
It’s a frequent itemset for which none of its intermediate supersets are frequent.	It’s a frequent itemset that is a superset of maximal frequent itemsets.
In this, we identify all the immediate supersets.	In this we identify all the frequent itemsets.
They’re valuable because they provide a compact representation of frequent itemsets.	They’re useful in removing in redundant association rules.
They form the smallest representation of frequent itemsets and hence they are the most practical to use when space is an issue.	They are a subset of frequent itemsets and so they provide a compact representation.

2. What is multi-level association.

Ans:

Multilevel Association Rule:

a) Association rules created from mining information at different degrees of reflection are called various level or staggered association rules.
b) Multilevel association rules can be mined effectively utilizing idea progressions under a help certainty system.
c) Rules at a high idea level may add to good judgment while rules at a low idea level may not be valuable consistently.
d) Sometimes at the low data level, data does not show any significant pattern but there is useful information hiding behind it.
e) The aim is to find the hidden information in or between levels of abstraction.

3. Approaches to multilevel association rule mining:

Ans:

a) Uniform Support (Using uniform minimum support for all level)
b) Reduced Support (Using reduced minimum support at lower levels)
c) Group-based Support (Using item or group-based support)