



**A.Y. 2022-2023**

Name : Siddharth Unny

SAP-ID : 60004200080

Batch : A4

LAB 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

Classification

Clustering

Association Rule

Select Attributes

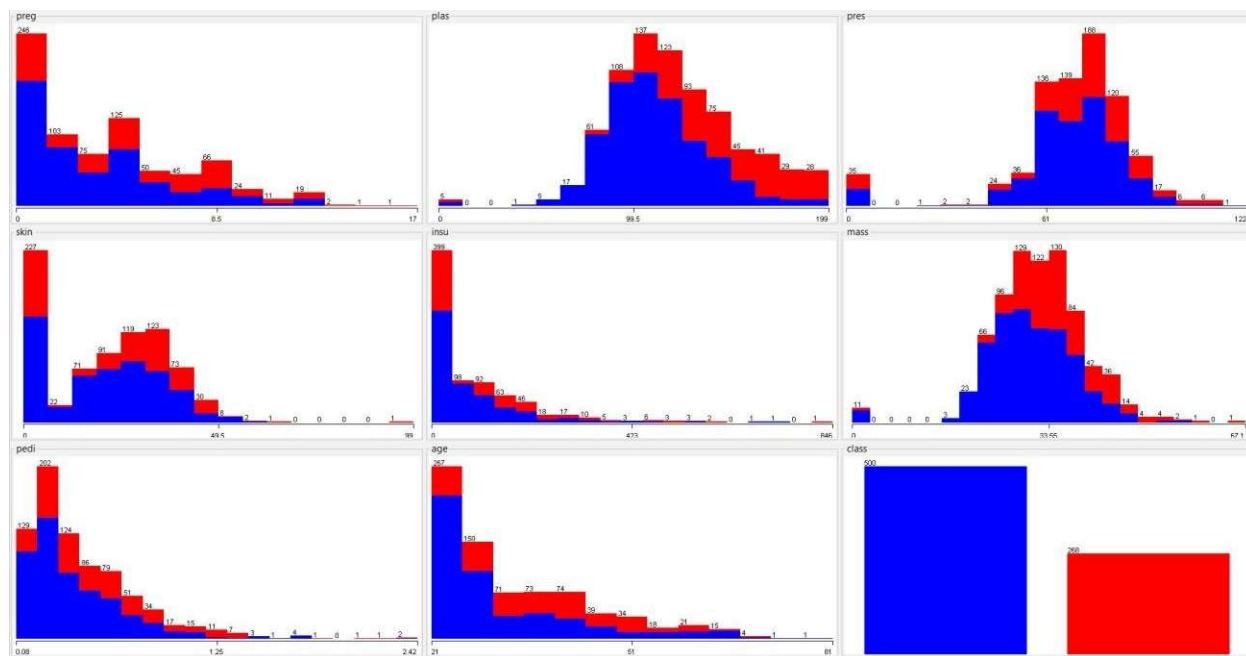
Visualization



A.Y. 2022-2023

Output :

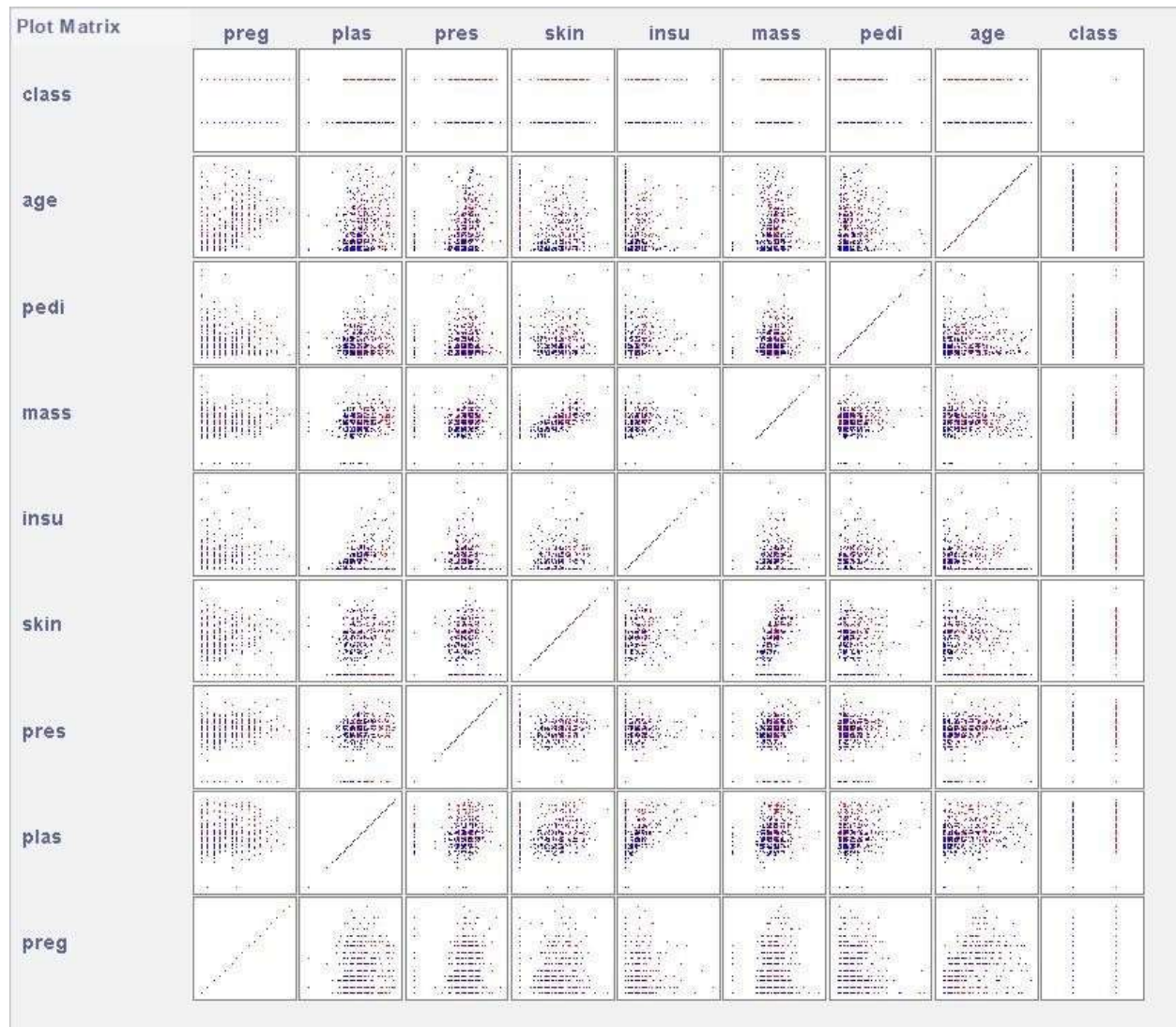
Visualization of all attributes





A.Y. 2022-2023

Scatter plot for all attributes



From the plot, we can see that attribute insu and skin are highly correlated along with minor correlation in plas, mass, pedi, age and class attributes.



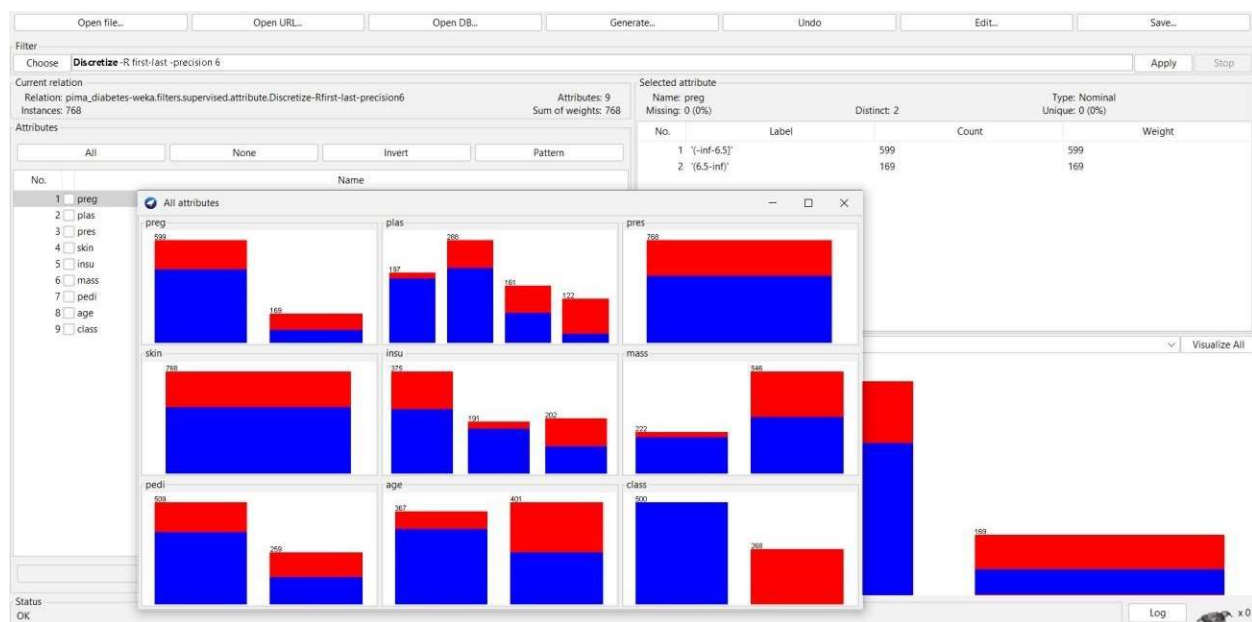
A.Y. 2022-2023





A.Y. 2022-2023

## Supervised Filter – Discretization



## Unsupervised Filter – Discretization





A.Y. 2022-2023

## Supervised Filter – AttributeSelection

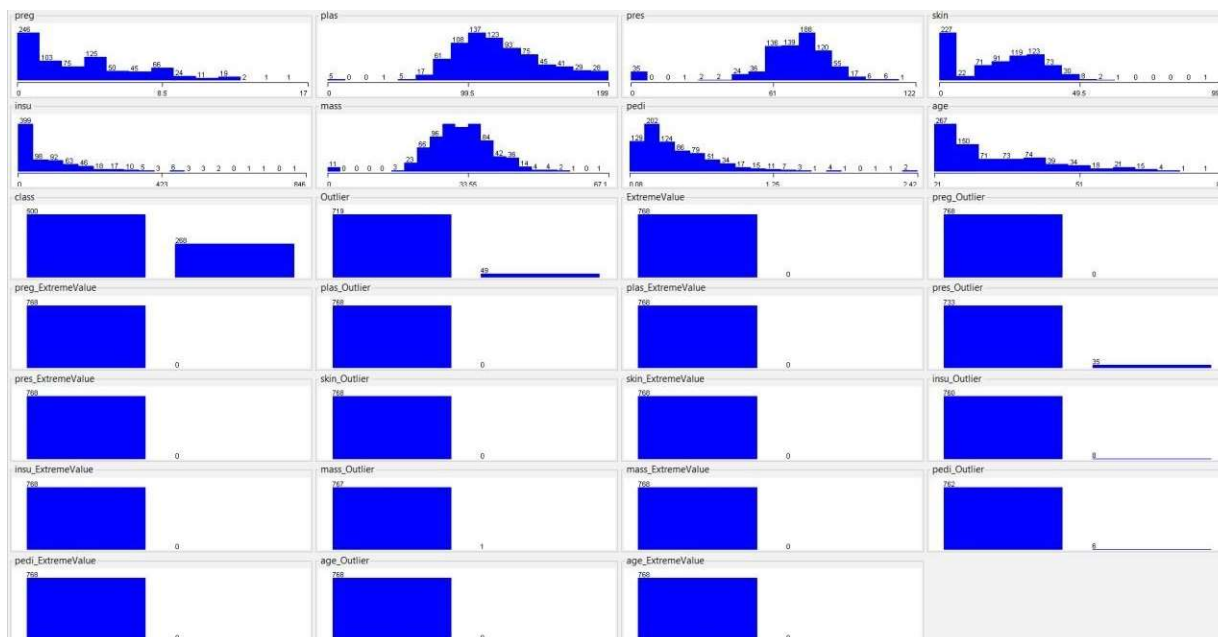






A.Y. 2022-2023

## Unsupervised Filter – InterQuartileRange



Classification :

Naïve Bayes

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier  
Choose **NaiveBayes**

Test options  
☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds: 10  
☐ Percentage split %: 66  
More options...

(Nom) class  
Start Stop  
Result list (right-click for options)  
200535 - bayes.NaiveBayes

Classifier output

precision 0.0045 0.0045

age  
mean 31.2494 37.0808  
std. dev. 11.6059 10.9146  
weight sum 500 268  
precision 1.1765 1.1765

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===  
=== Summary ===  
Correctly Classified Instances 586 76.3021 %  
Incorrectly Classified Instances 182 23.6979 %  
Kappa statistic 0.4664  
Mean absolute error 0.2841  
Root mean squared error 0.4168  
Relative absolute error 62.5028 %  
Root relative squared error 87.4349 %  
Total Number of Instances 768

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
Weighted Avg.	0.612	0.136	0.678	0.612	0.643	0.468	0.819	tested_negative
	0.763	0.307	0.759	0.763	0.760	0.468	0.819	tested_positive

=== Confusion Matrix ===  
a b <-- classified as  
422 78 | a = tested\_negative  
104 164 | b = tested\_positive

Status  
OK Log x0



A.Y. 2022-2023

Decision Table:

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **DecisionTable** -X1-5 "weka.attributeSelection.BestFirst-D1-N5"

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds: 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

200535 - bayes.NaiveBayes

200711 - rules.DecisionTable

Classifier output

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 43

Merit of best subset found: 77.604

Evaluation (for feature selection): CV (leave one out)

Feature set: 1,2,4,9

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	547	71.224 %
Incorrectly Classified Instances	221	28.776 %
Kappa statistic	0.3492	
Mean absolute error	0.3448	
Root mean squared error	0.4277	
Relative absolute error	75.8525 %	
Root relative squared error	89.7294 %	
Total Number of Instances	768	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.610	0.470	0.763	0.810	0.786	0.351	0.773	0.855	tested_negative
	0.530	0.190	0.599	0.530	0.562	0.351	0.773	0.628	tested_positive
	0.712	0.372	0.706	0.712	0.708	0.351	0.773	0.778	

=== Confusion Matrix ===

a b <-- classified as

405 95 | a = tested\_negative

126 142 | b = tested\_positive

Status OK

Log x0

Clustering :

Simple K-Means

Preprocess Classify Cluster Associate Select attributes Visualize

Cluster

Choose **SimpleKMeans** -init0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

☒ Use training set

☐ Supplied test set

☐ Percentage split % 66

☐ Classes to clusters evaluation

(Nom) class

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

200814 - SimpleKMeans

Cluster output

Within cluster sum of squared errors: 149.5177664501119

Initial starting points (random):

Cluster 0: 1,126,56,29,152,20,7,0.801,21,tested\_negative

Cluster 1: 8,95,72,0,0,36,8,0.485,57,tested\_negative

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data	Cluster#	
	(768.0)	(500.0)	(268.0)
preg	3.8451	3.298	4.8657
plas	120.8945	105.58	141.2575
pres	69.1055	68.184	70.0246
skin	20.5365	19.664	22.1642
insu	79.7995	68.792	100.3358
mass	31.9926	30.3042	35.1425
pedi	0.4719	0.4297	0.5505
age	33.2409	31.19	37.0672
class			tested_negative tested_negative tested_positive

Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

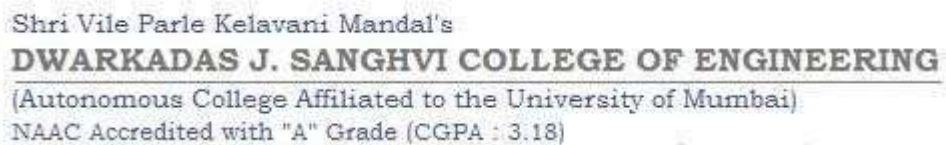
Clustered Instances

0	500 ( 65%)
1	268 ( 35%)

Status OK

Log x0





**Preprocess**   **Classify**   **Cluster**   Associate   Select attributes   Visualize

---

Cluster  
Choose **HierarchicalCluster -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"**

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)

2008.14 - SimpleKMeans
2008.47 - HierarchicalCluster

Cluster output

```

Scheme:      weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:    pima_diabetes
Instances:   768
Attributes:  9

preg
plas
pres
skin
insu
mass
pedi
age
class

Test mode:   evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
(((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((1.0;0.16243,1.0;0.16243);0.01512,1.0;0.17755);0.01402,1.0;0.19157);0.00239,(1.0;0.

Cluster 1
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((

Time taken to build model (full training data) : 5.96 seconds

=== Model and evaluation on training set ===

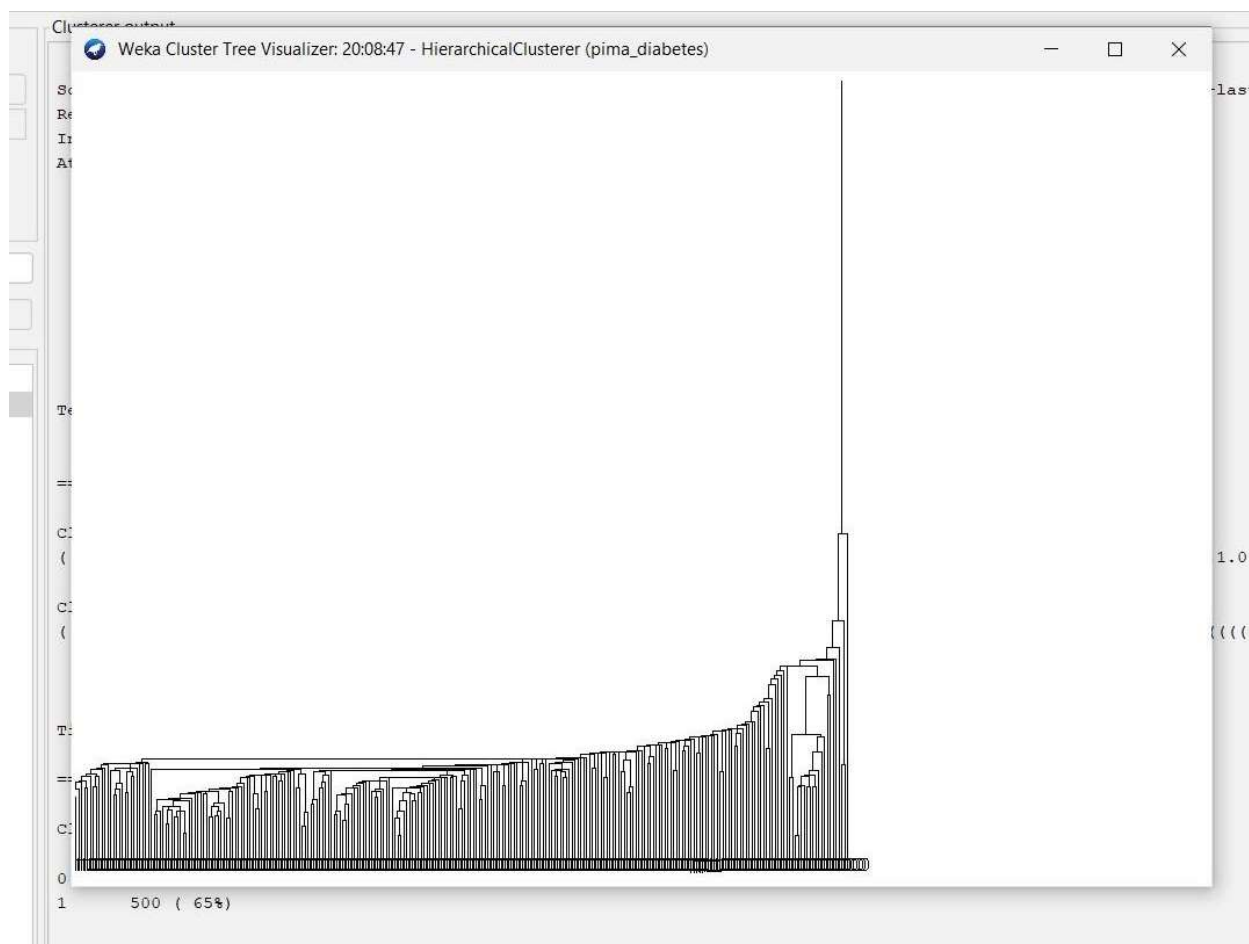
Clustered Instances

0       268   ( 35%)
1       500   ( 65%)
        
```

Status OK
Log



A.Y. 2022-2023



In clustering based on the attribute needed to be predicted random amount of clusters are formed, in this case there are two clusters 0 and 1 which denote the tested\_negative and tested\_positive values of attribute 'class' based on which samples from each cluster are passed in the model for training and the corresponding output displays the clusters with count of instances present in each cluster.

A.Y. 2022-2023

Since there is no association between the attributes, associate Apriori can't be used here.

Dataset – Supermarket – Apriori

The screenshot shows a software interface with a menu bar (Preprocess, Classify, Cluster, Associate, Select attributes, Visualize) and a toolbar (Start, Stop). The 'Associate' menu is active, and the 'Apriori' option is selected. The 'Apriori' window displays the following output:

```
==== Apriori model (full training set) ====

Apriori
=====

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1

Best rules found:

1. biscuits== frozen foods== fruit== total=high 708 ==> bread and cake== 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs== biscuits== fruit== total=high 760 ==> bread and cake== 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
3. baking needs== frozen foods== fruit== total=high 770 ==> bread and cake== 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
4. biscuits== fruit== vegetables== total=high 815 ==> bread and cake== 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
5. party snack foods== fruit== total=high 854 ==> bread and cake== 775 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
6. biscuits== frozen foods== vegetables== total=high 797 ==> bread and cake== 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
7. baking needs== biscuits== vegetables== total=high 772 ==> bread and cake== 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
8. biscuits== fruit== total=high 954 ==> bread and cake== 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
9. frozen foods== fruit== vegetables== total=high 834 ==> bread and cake== 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods== fruit== total=high 969 ==> bread and cake== 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)
```

The status bar at the bottom shows 'Status OK' and a 'Log' button.

Here we can see that attributes which associate with attributes bread and cake allowing the owner to understand which items should be placed together and such.