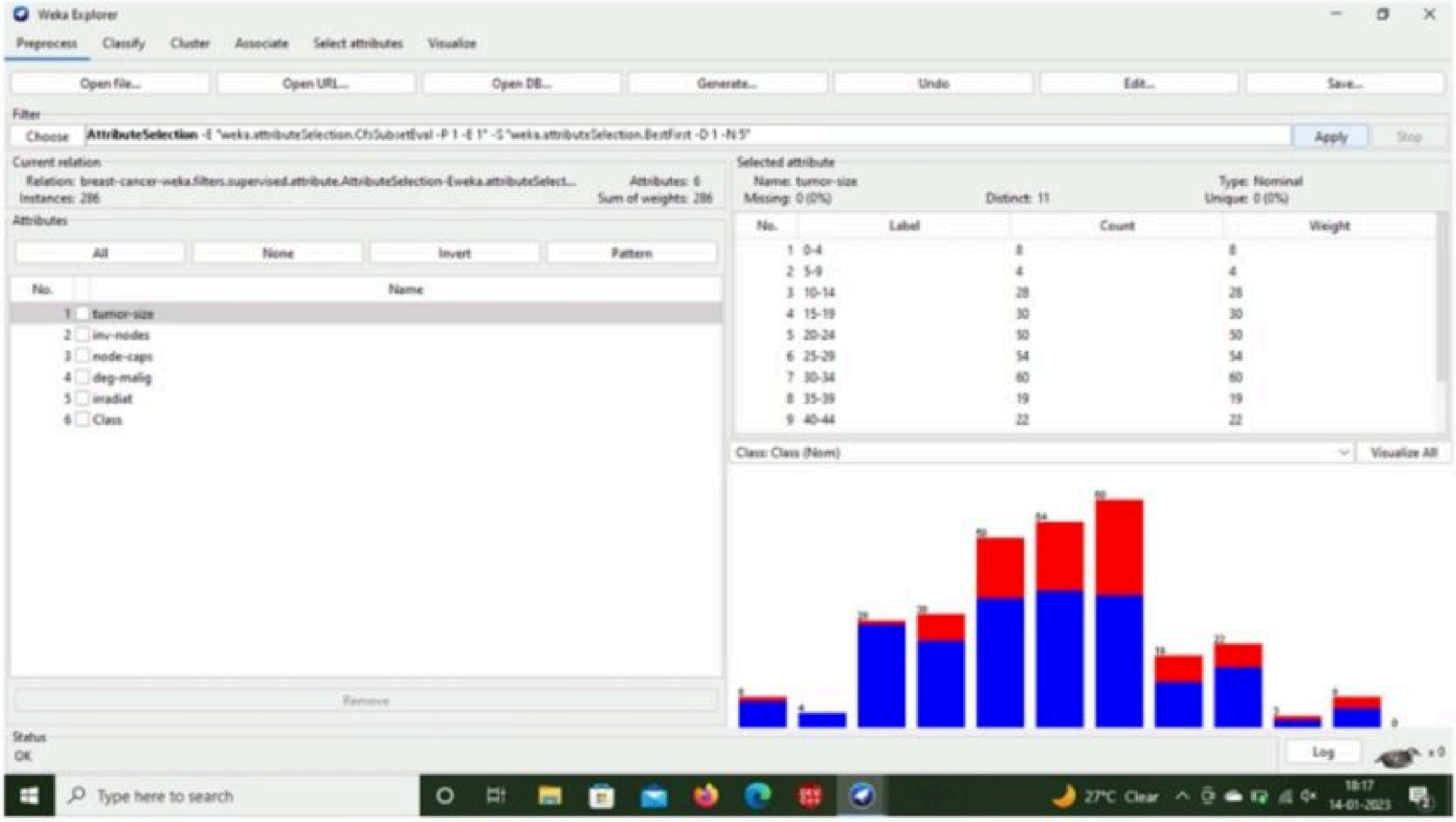
CSA1668-DATA WAREHOUSING AND DATA MINING

T, MANOHAR
192110219

DATA PRE PROCESSING



Filter
Choose **AttributeSelection** -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Current relation
Relation: breast-cancer-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelect... Attributes: 6
Instances: 286 Sum of weights: 286

Attributes

No.		Name
1	□	tumor-size
2	□	inv-nodes
3	□	node-caps
4	□	deg-malig
5	□	irradiat
6	□	Class

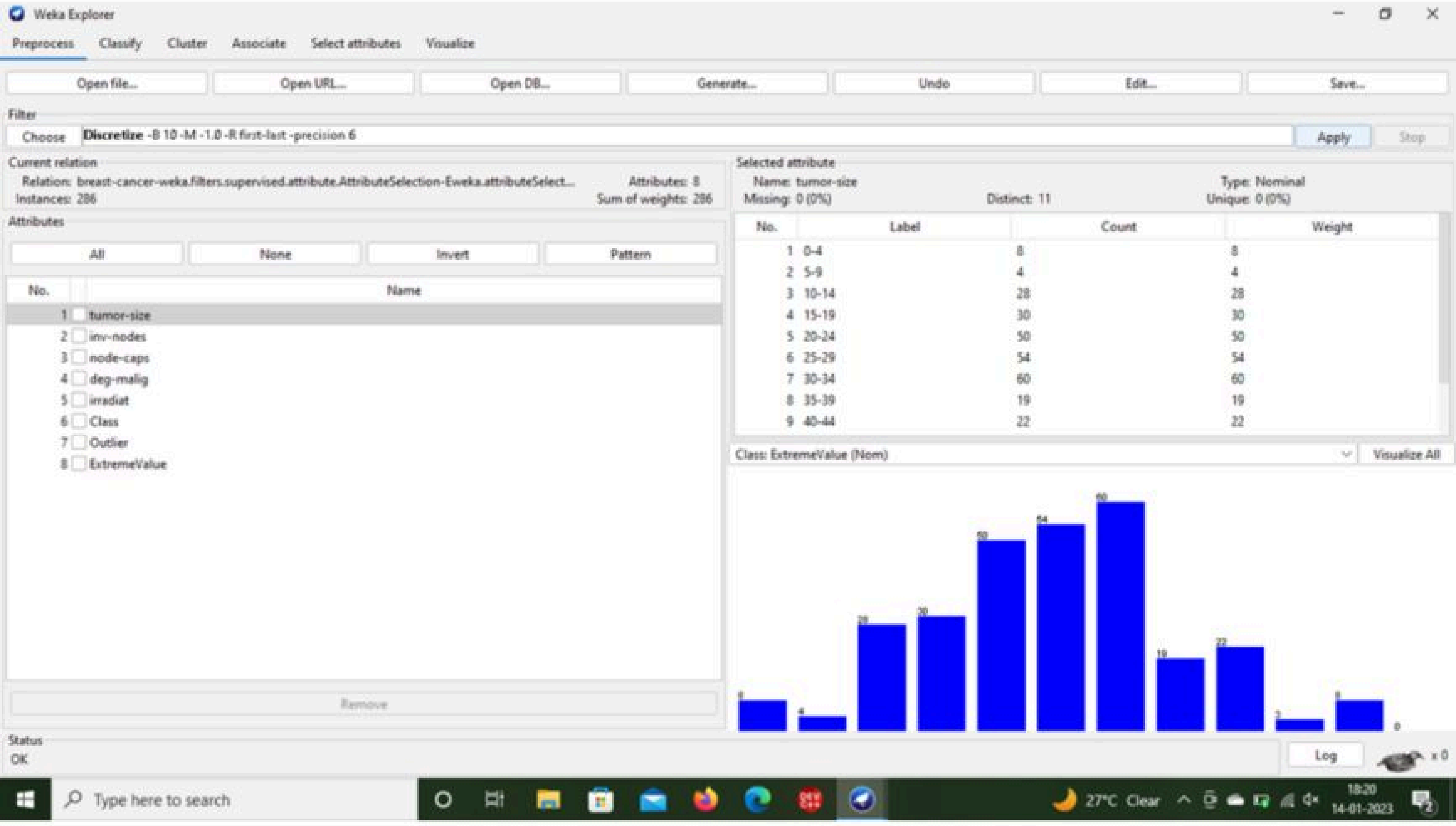
Selected attribute
Name: tumor-size Type: Nominal
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
1	0-4	8	8
2	5-9	4	4
3	10-14	28	28
4	15-19	30	30
5	20-24	50	50
6	25-29	54	54
7	30-34	60	60
8	35-39	19	19
9	40-44	22	22

Class: Class (Nom)

Status
OK

DISCRETIZE



Filter

Discretize -B 10 -M -1.0 -R first-last -precision 6

Current relation

Relation: breast-cancer-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelect...

Attributes: 8

Instances: 286

Sum of weights: 286

Attributes

No.		Name
1	□	tumor-size
2	□	inv-nodes
3	□	node-caps
4	□	deg-malig
5	□	irradiat
6	□	Class
7	□	Outlier
8	□	ExtremeValue

Selected attribute

Name: tumor-size

Type: Nominal

Missing: 0 (0%)

Distinct: 11

Unique: 0 (0%)

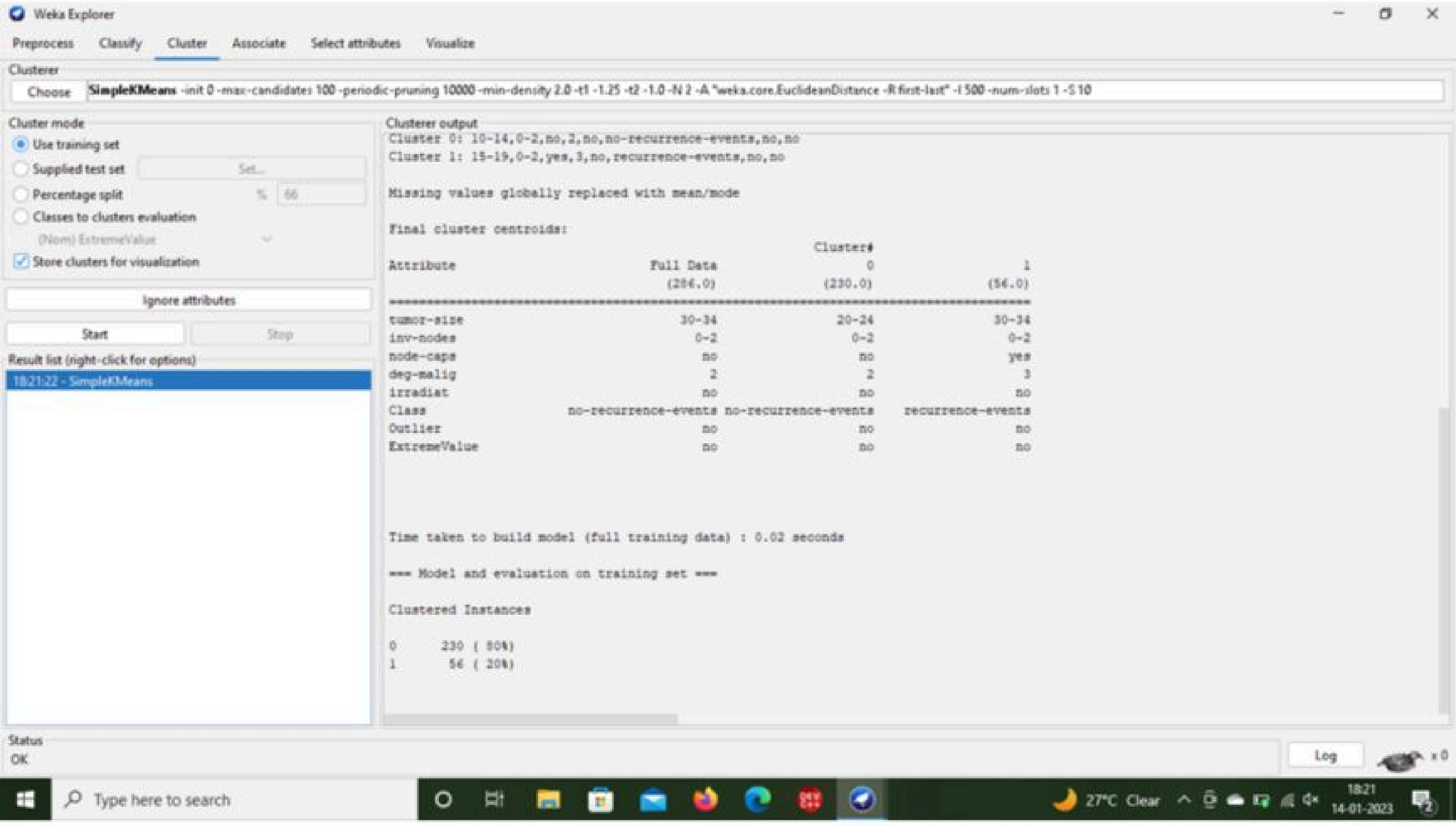
No.	Label	Count	Weight
1	0-4	8	8
2	5-9	4	4
3	10-14	28	28
4	15-19	30	30
5	20-24	50	50
6	25-29	54	54
7	30-34	60	60
8	35-39	19	19
9	40-44	22	22

Class: ExtremeValue (Nom)

Status

OK

CLUSTERING



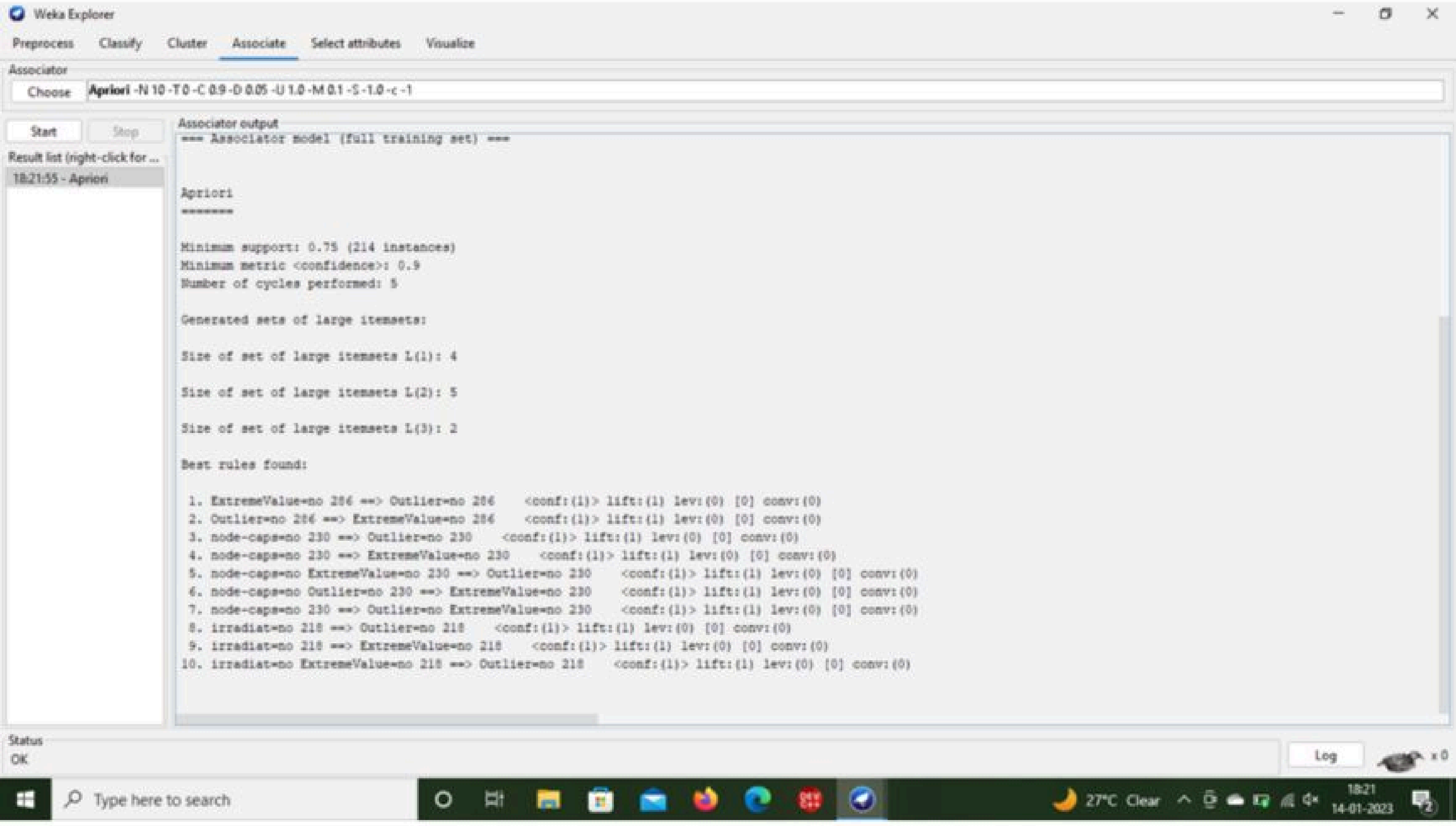
Clusterer: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode:

- ● Use training set
- ○ Supplied test set
- ○ Percentage split
- ○ Classes to clusters evaluation
- ☑ Store clusters for visualization

```
Cluster 0: 10-14,0-2,no,2,no,no-recurrence-events,no,no
Cluster 1: 15-19,0-2,yes,3,no,recurrence-events,no,no

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                  Cluster#
Attribute                        Full Data               0                   1
                                   (286.0)         (230.0)              (56.0)
==============================================================================
tumor-size                           30-34           20-24               30-34
inv-nodes                              0-2             0-2                 0-2
node-caps                               no              no                 yes
deg-malig                                2               2                   3
irradiat                                no              no                  no
Class                 no-recurrence-events no-recurrence-events  recurrence-events
Outlier                                 no              no                  no
ExtremeValue                            no              no                  no




Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0      230 ( 80%)
1       56 ( 20%)
```

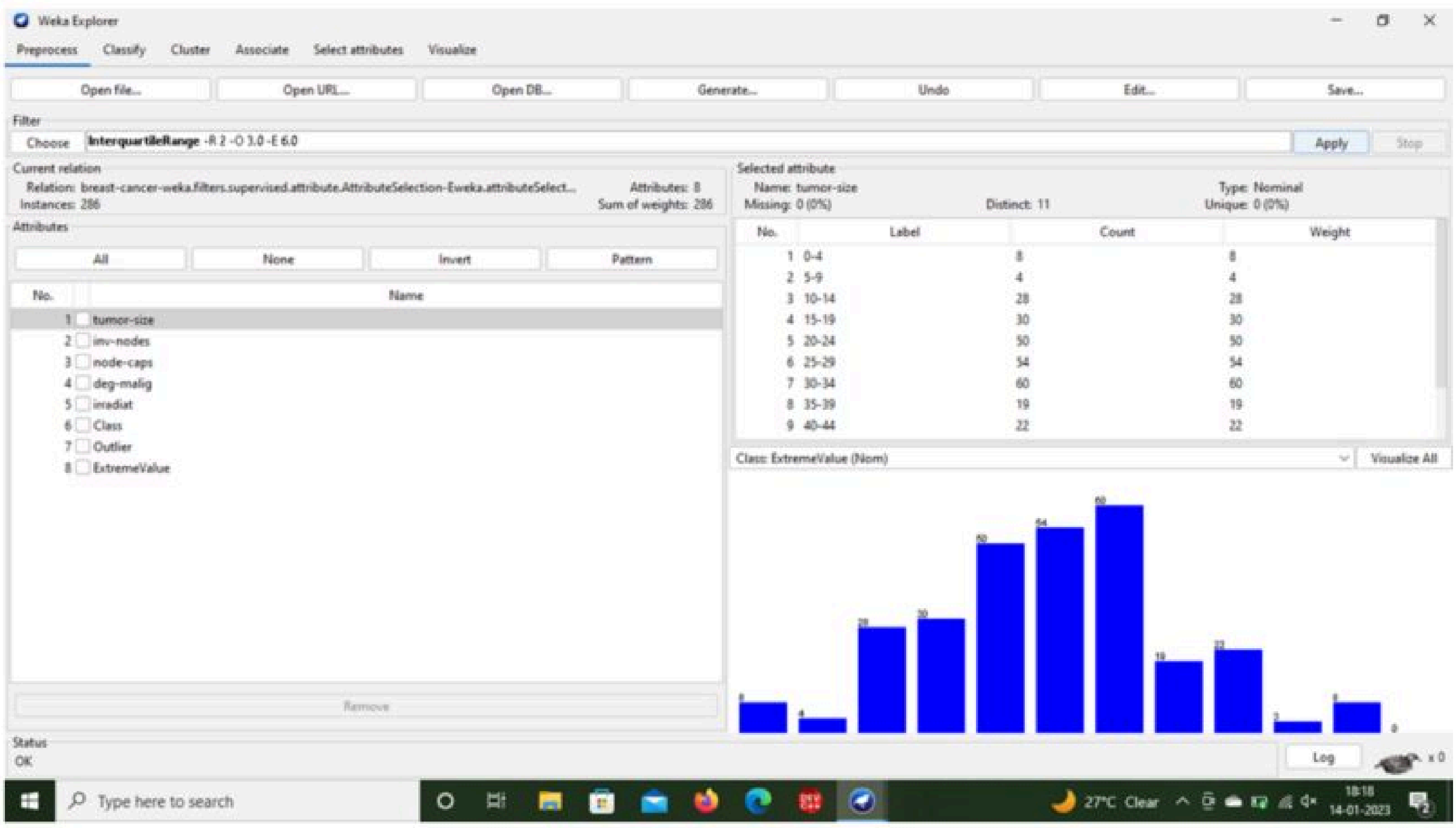
ASSOCIATE



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

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


Apriori
=======

Minimum support: 0.75 (214 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 5

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

 1. ExtremeValue=no 286 ==> Outlier=no 286    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. Outlier=no 286 ==> ExtremeValue=no 286    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. node-caps=no 230 ==> Outlier=no 230    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. node-caps=no 230 ==> ExtremeValue=no 230    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. node-caps=no ExtremeValue=no 230 ==> Outlier=no 230    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. node-caps=no Outlier=no 230 ==> ExtremeValue=no 230    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. node-caps=no 230 ==> Outlier=no ExtremeValue=no 230    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. irradiat=no 218 ==> Outlier=no 218    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. irradiat=no 218 ==> ExtremeValue=no 218    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. irradiat=no ExtremeValue=no 218 ==> Outlier=no 218    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

INTERQUARTILE RANGE



Filter

InterquartileRange -R 2 -O 3.0 -E 6.0

Current relation

Relation: breast-cancer-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelect... Attributes: 8

Instances: 286 Sum of weights: 286

Attributes

No.		Name
1	□	tumor-size
2	□	inv-nodes
3	□	node-caps
4	□	deg-malig
5	□	irradiat
6	□	Class
7	□	Outlier
8	□	ExtremeValue

Selected attribute

Name: tumor-size Type: Nominal

Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

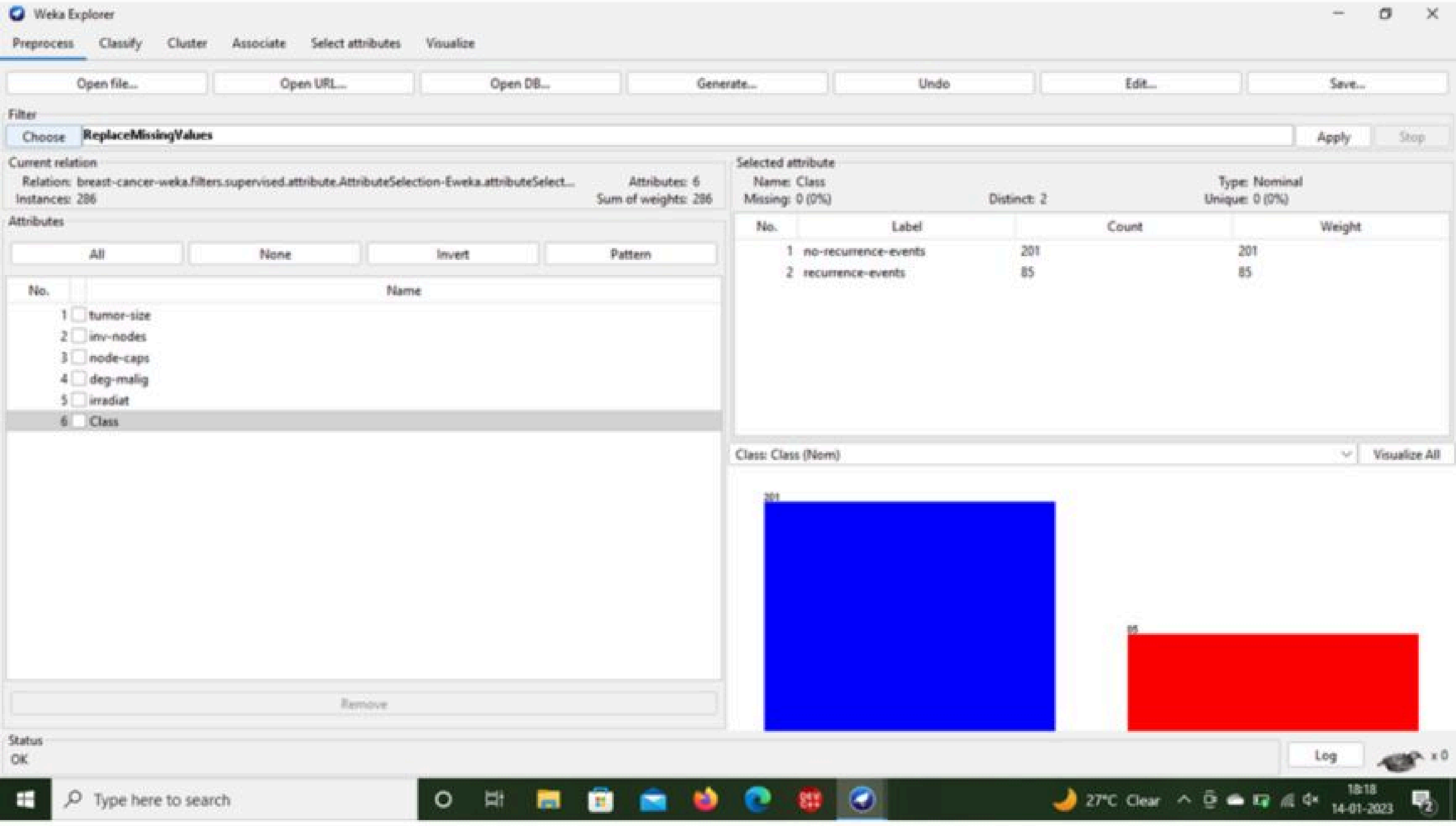
No.	Label	Count	Weight
1	0-4	8	8
2	5-9	4	4
3	10-14	28	28
4	15-19	30	30
5	20-24	50	50
6	25-29	54	54
7	30-34	60	60
8	35-39	19	19
9	40-44	22	22

Class: ExtremeValue (Nom)

Status

OK

REPLACE MISSING VALUES



Filter

ReplaceMissingValues

Current relation

Relation: breast-cancer-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelect... Attributes: 6

Instances: 286 Sum of weights: 286

Attributes

No.		Name
1	□	tumor-size
2	□	inv-nodes
3	□	node-caps
4	□	deg-malig
5	□	irradiat
6	□	Class

Selected attribute

Name: Class Type: Nominal

Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

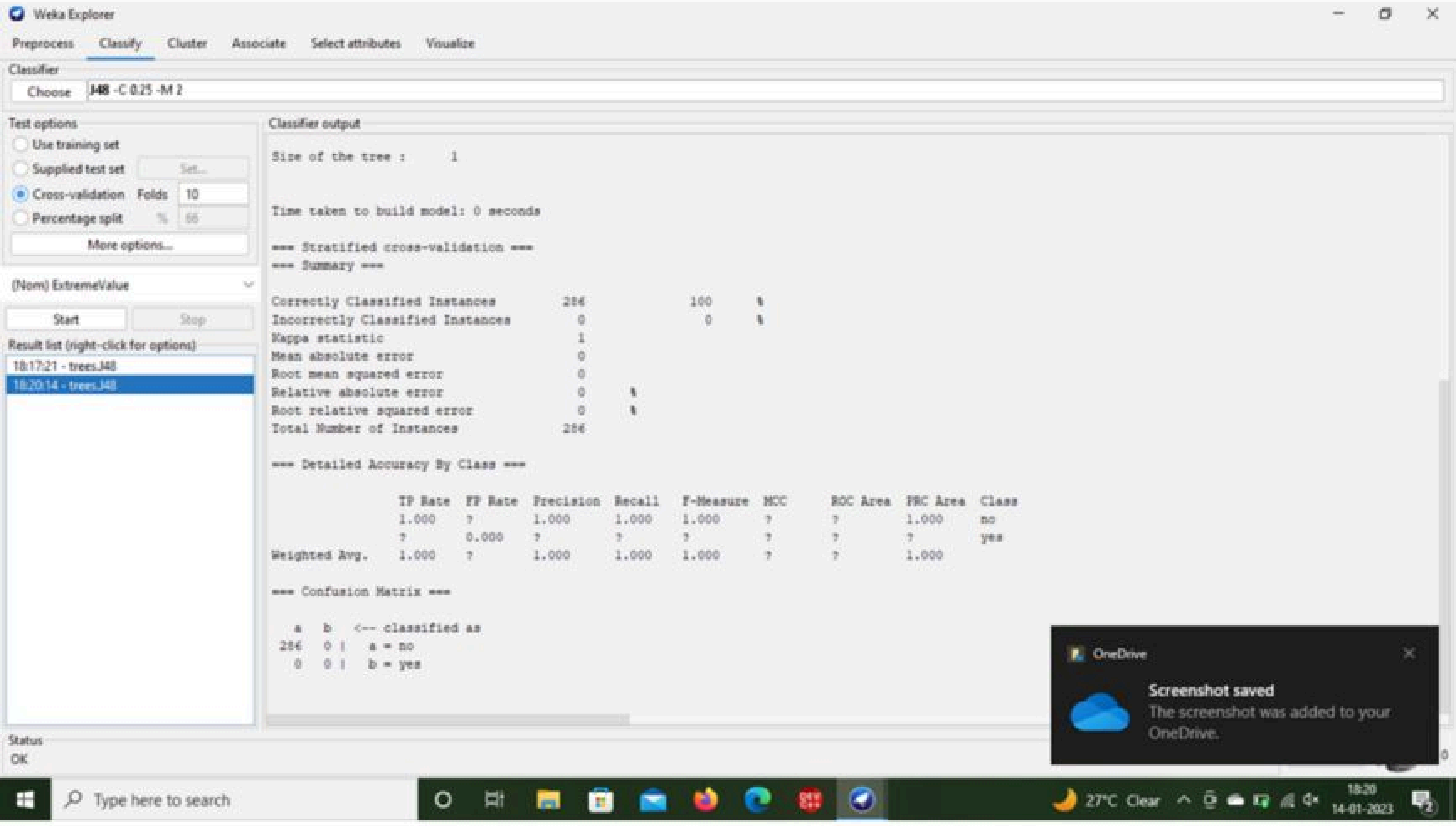
No.	Label	Count	Weight
1	no-recurrence-events	201	201
2	recurrence-events	85	85

Class: Class (Nom)

Status

OK

CLASSIFICATION



Classifier: J48 -C 0.25 -M 2

Test options:
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) ExtremeValue

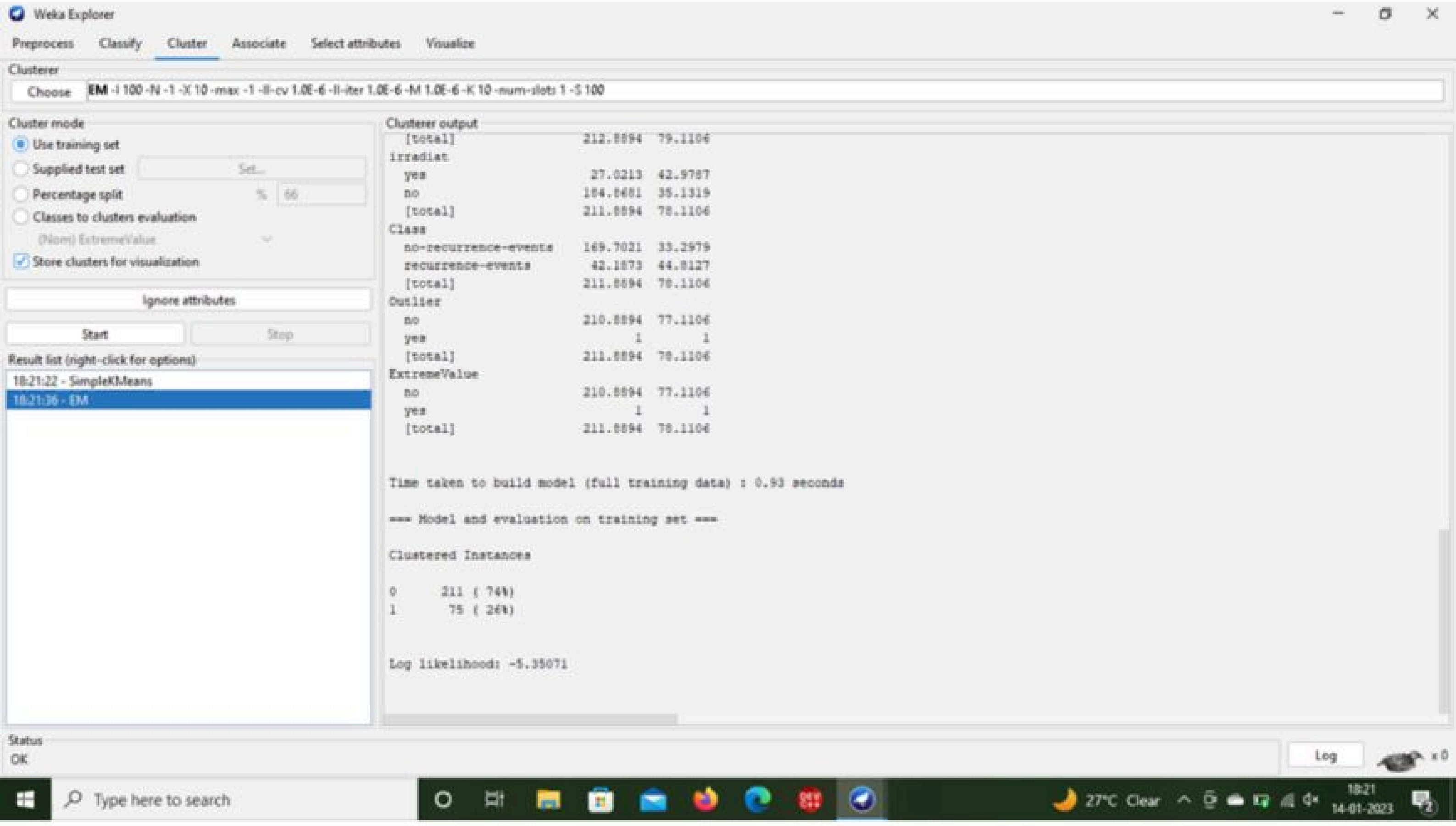
Result list (right-click for options)
- 18:17:21 - trees.J48
- 18:20:14 - trees.J48

Status: OK

Classifier output

```
Size of the tree :      1


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         286              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    ?        1.000      1.000    1.000      ?        ?         1.000     no
                 ?        0.000    ?          ?        ?          ?        ?         ?         yes
Weighted Avg.    1.000    ?        1.000      1.000    1.000      ?        ?         1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 286   0 |   a = no
   0   0 |   b = yes
```

EM



Clusterer

```
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
```

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation (Nom) ExtremeValue
- ☑ Store clusters for visualization

Result list (right-click for options)

- 18:21:22 - SimpleKMeans
- 18:21:36 - EM

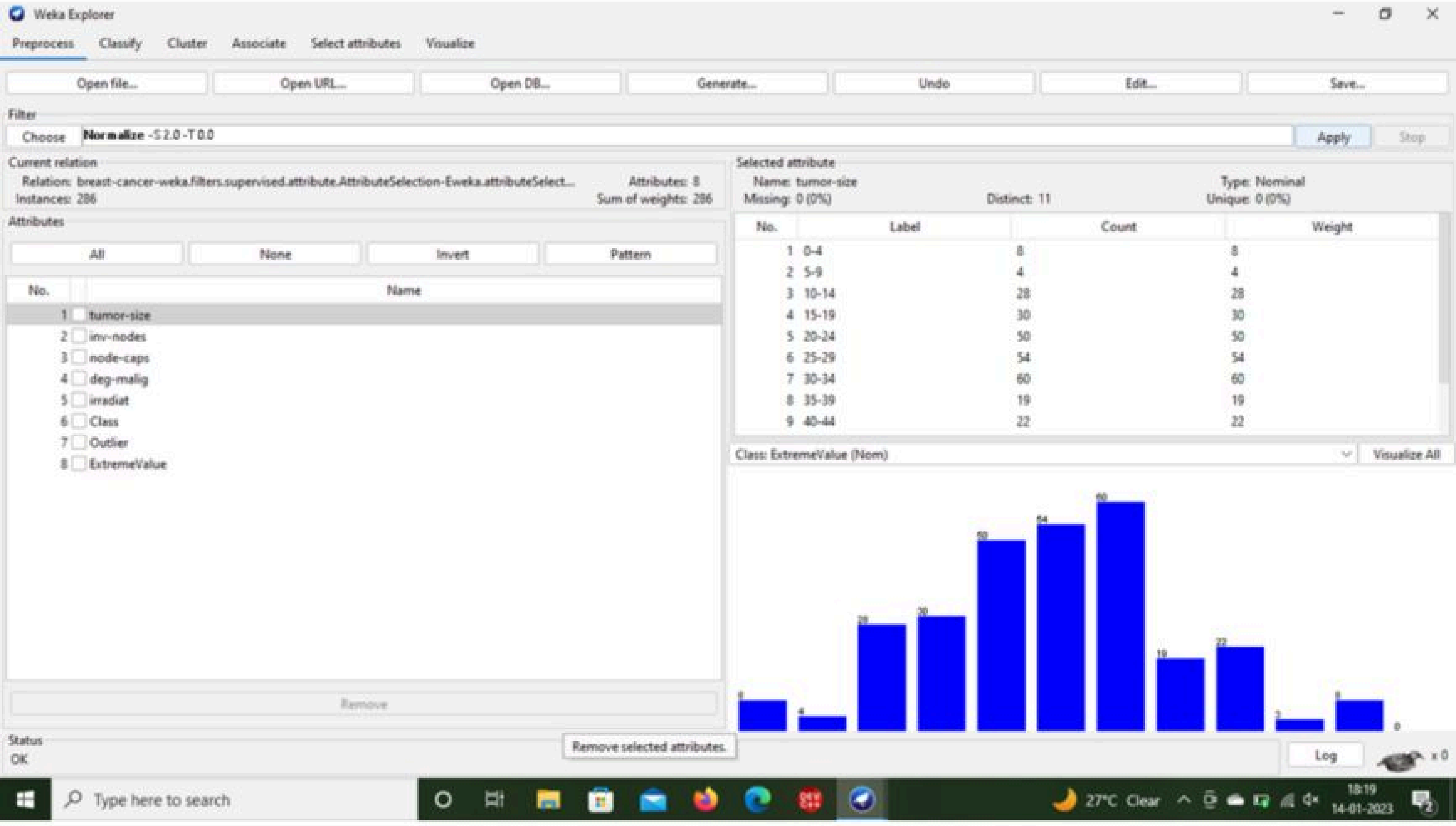
Clusterer output

```
  [total]                212.8894  79.1106
irradiat
  yes                     27.0213  42.9787
  no                     184.8681  35.1319
  [total]                211.8894  78.1106
Class
  no-recurrence-events   169.7021  33.2979
  recurrence-events       42.1873  44.8127
  [total]                211.8894  78.1106
Outlier
  no                     210.8894  77.1106
  yes                           1        1
  [total]                211.8894  78.1106
ExtremeValue
  no                     210.8894  77.1106
  yes                           1        1
  [total]                211.8894  78.1106


Time taken to build model (full training data) : 0.93 seconds

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

Clustered Instances

0      211 ( 74%)
1       75 ( 26%)


Log likelihood: -5.35071
```

Status

OK

NORMALIZE



Filter

Normalize -S 2.0 -T 0.0

Current relation

Relation: breast-cancer-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelect... Attributes: 8

Instances: 286 Sum of weights: 286

Attributes

No.		Name
1	□	tumor-size
2	□	inv-nodes
3	□	node-caps
4	□	deg-malig
5	□	irradiat
6	□	Class
7	□	Outlier
8	□	ExtremeValue

Selected attribute

Name: tumor-size Type: Nominal

Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
1	0-4	8	8
2	5-9	4	4
3	10-14	28	28
4	15-19	30	30
5	20-24	50	50
6	25-29	54	54
7	30-34	60	60
8	35-39	19	19
9	40-44	22	22

Class: ExtremeValue (Nom)

Status

OK