Data Mining Assignment 1 using Weka

Student Name and ID of the member submitting the assignment: Harshith Pashikanti, 1001974588

Student Name and ID of the remaining members: Venkata Sainath Reddy Palavala, 1001949223

Introduction to Weka:

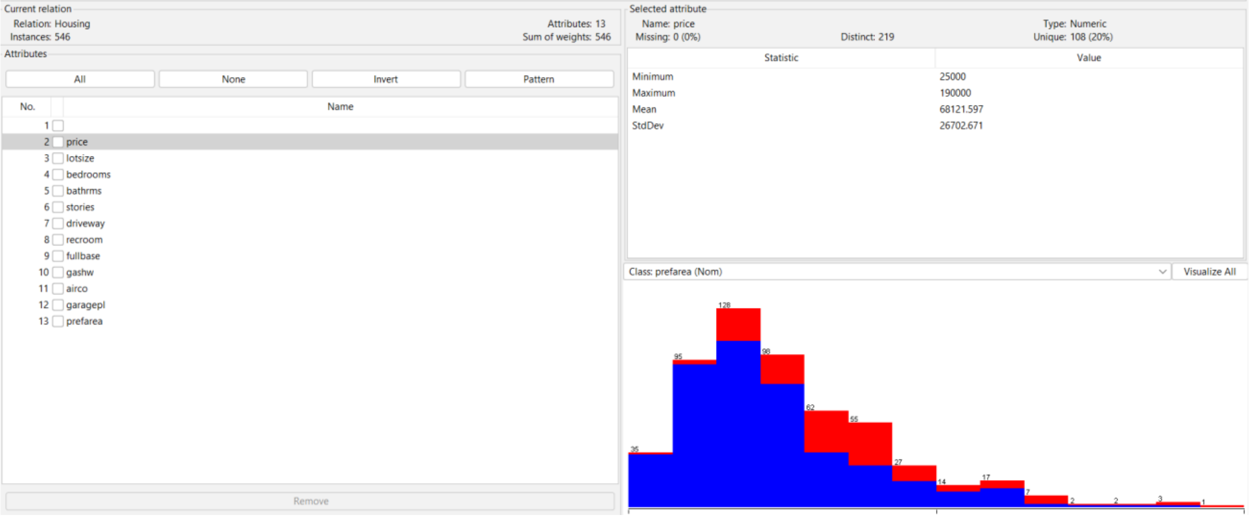
Weka is a collection of machine learning algorithms for data mining tasks. Weeks has the following tool that can be used in understanding the data and visualizing the data with utmost accuracy.

- Data pre-processing
- Regression
- Classification
- Clustering
- Association
- Visualization

To visualize the data in Weka, we need to the import the Housing.csv file which is provided. The dataset has Relation: Housing, Attributes: 13, Instances: 546, Sum of weights: 546.

For the selected attributes in the Weka, it will show the Type of the attribute, Missing values, Unique values, Maximum, Minimum, Mean, and StdDev.

List of Attributes:



Current relation

Relation: Housing
Instances: 546

Attributes: 13
Sum of weights: 546

Attributes

No.		Name
1	□	
2	□	price
3	□	lotsize
4	□	bedrooms
5	□	bathrms
6	□	stories
7	□	driveway
8	□	recroom
9	□	fullbase
10	□	gashw
11	□	airco
12	□	garagepl
13	□	prefarea

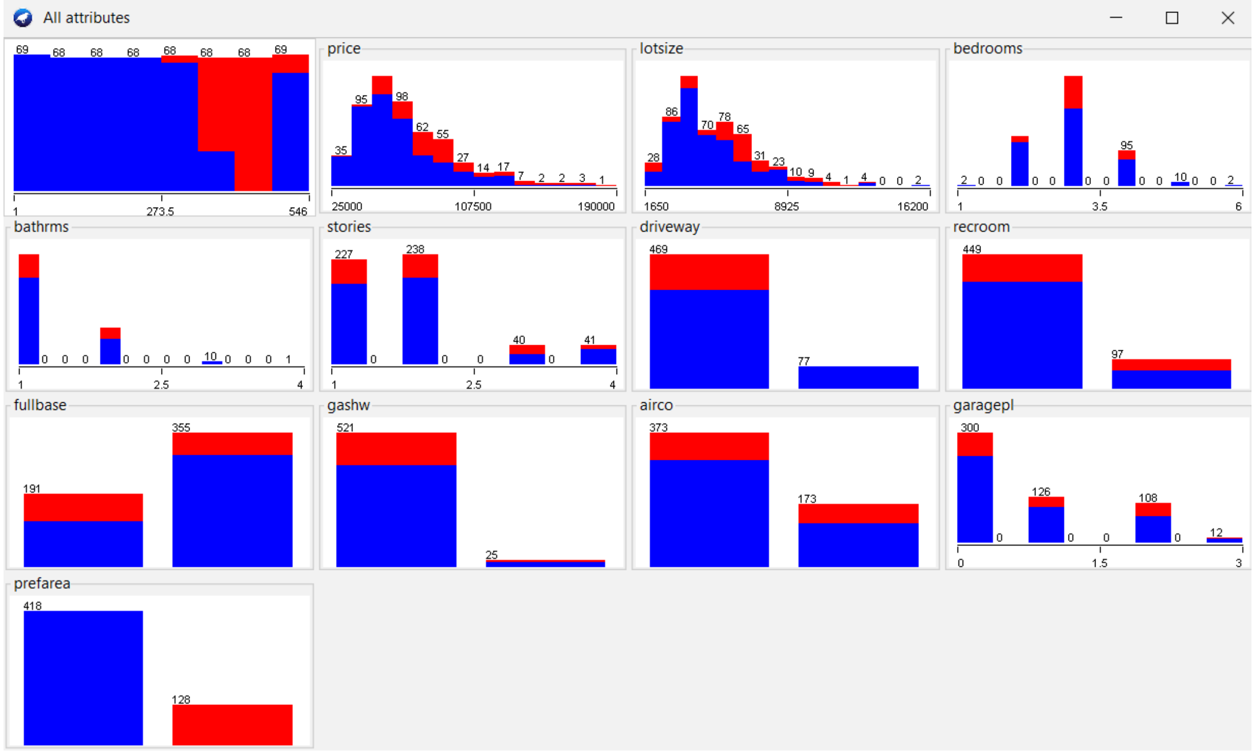
Selected attribute

Name: price
Missing: 0 (0%)
Distinct: 219
Type: Numeric
Unique: 108 (20%)

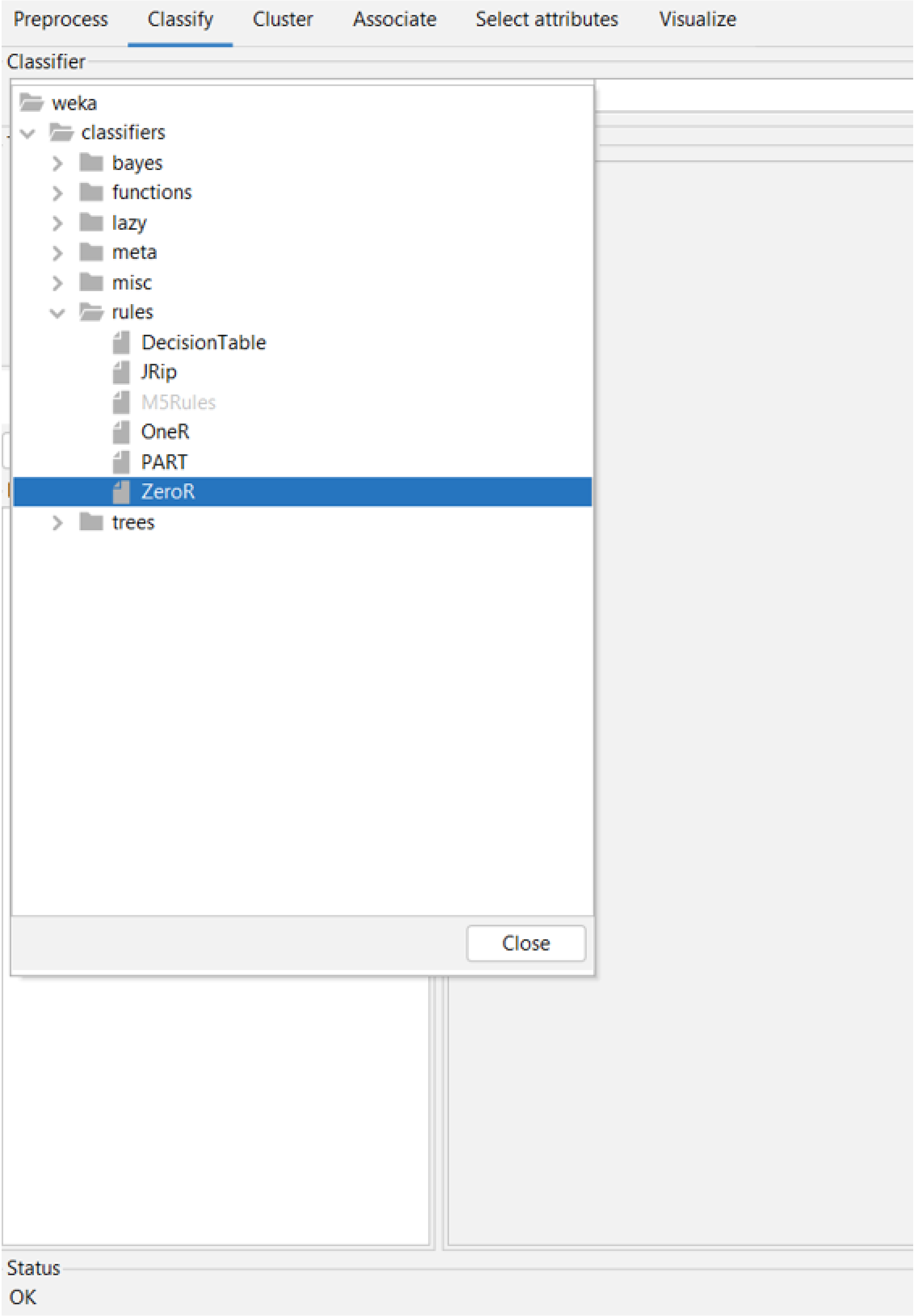
Statistic	Value
Minimum	25000
Maximum	190000
Mean	68121.597
StdDev	26702.671

Class: prefarea (Nom)

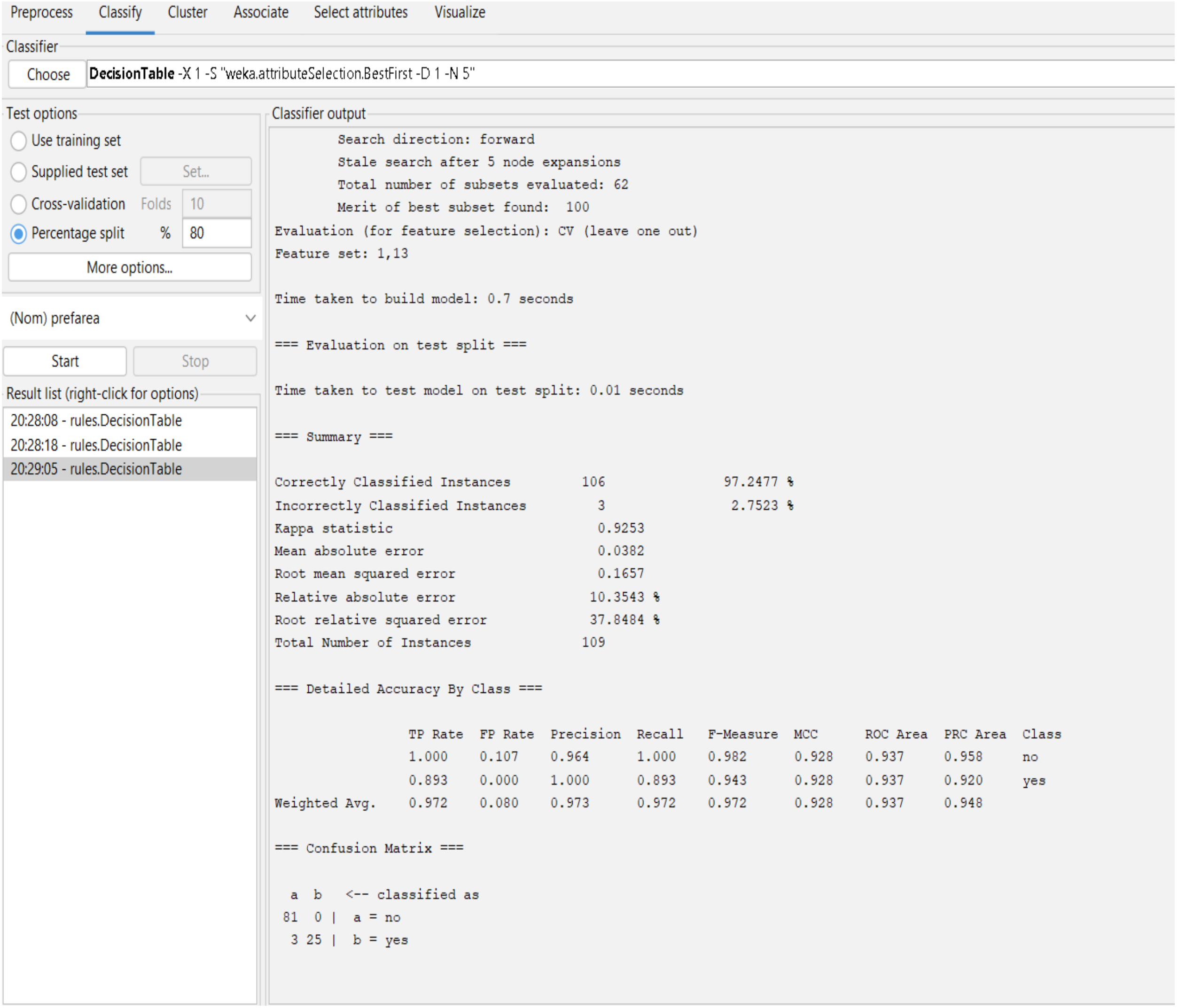
Visualize All for all attributes:



Classify: Decision table



:



Classifier

DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 80

(Nom) prefarea

Result list (right-click for options)

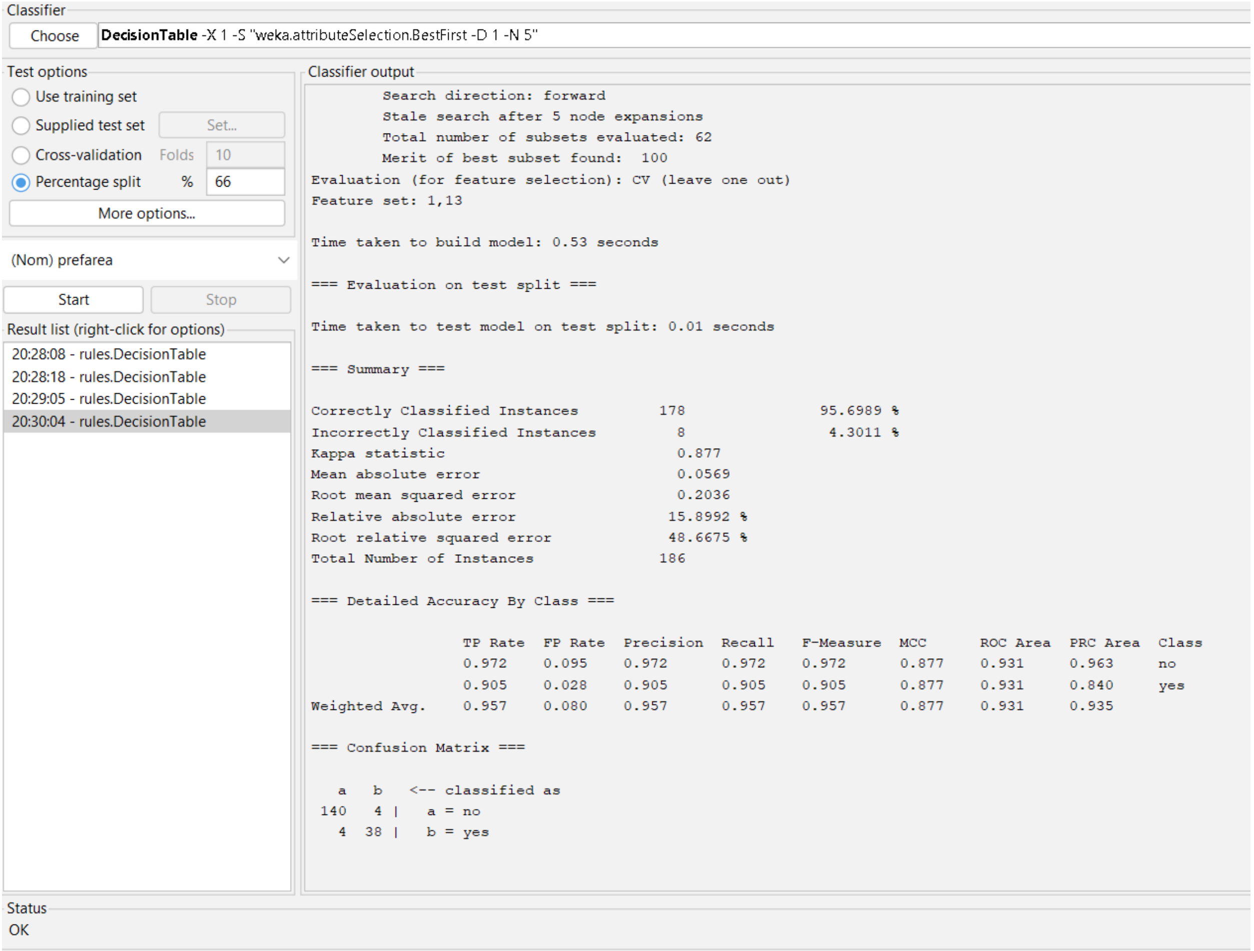
20:28:08 - rules.DecisionTable

20:28:18 - rules.DecisionTable

20:29:05 - rules.DecisionTable

Classifier output

```
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 62
        Merit of best subset found:  100
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,13

Time taken to build model: 0.7 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         106               97.2477 %
Incorrectly Classified Instances         3                2.7523 %
Kappa statistic                          0.9253
Mean absolute error                      0.0382
Root mean squared error                  0.1657
Relative absolute error                 10.3543 %
Root relative squared error             37.8484 %
Total Number of Instances              109

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.107    0.964      1.000    0.982      0.928    0.937     0.958     no
                 0.893    0.000    1.000      0.893    0.943      0.928    0.937     0.920     yes
Weighted Avg.    0.972    0.080    0.973      0.972    0.972      0.928    0.937     0.948

=== Confusion Matrix ===

  a  b   <-- classified as
 81  0 |  a = no
  3 25 |  b = yes
```



Classifier

Choose **DecisionTable** -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Test options

○ Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
● Percentage split % 66

More options...

(Nom) prefarea

Start Stop

Result list (right-click for options)

20:28:08 - rules.DecisionTable
20:28:18 - rules.DecisionTable
20:29:05 - rules.DecisionTable
20:30:04 - rules.DecisionTable

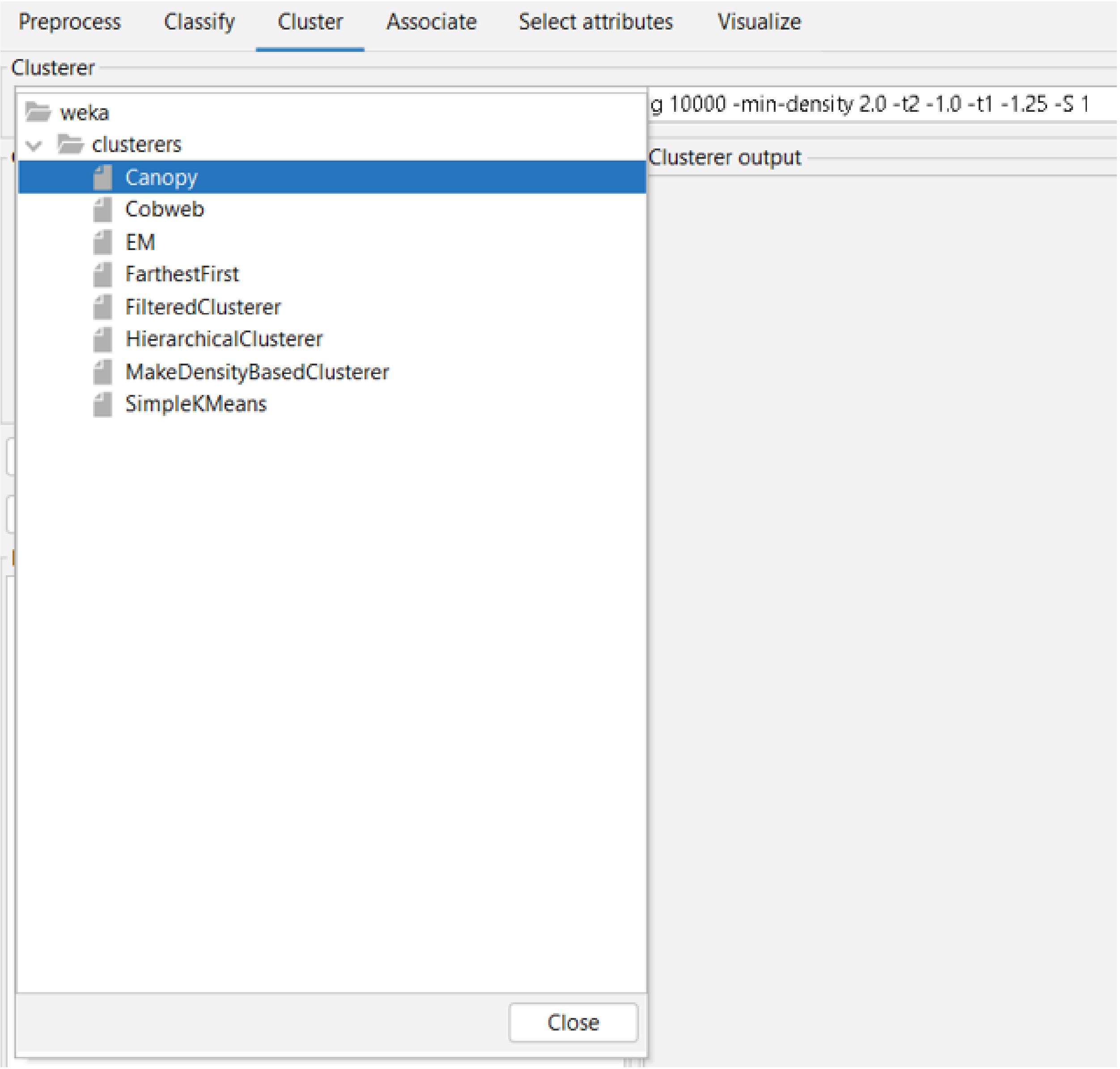
Classifier output

```
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 62
        Merit of best subset found:  100
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,13

Time taken to build model: 0.53 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         178               95.6989 %
Incorrectly Classified Instances         8                4.3011 %
Kappa statistic                          0.877
Mean absolute error                      0.0569
Root mean squared error                  0.2036
Relative absolute error                 15.8992 %
Root relative squared error             48.6675 %
Total Number of Instances              186

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.095    0.972      0.972    0.972      0.877    0.931     0.963     no
                 0.905    0.028    0.905      0.905    0.905      0.877    0.931     0.840     yes
Weighted Avg.    0.957    0.080    0.957      0.957    0.957      0.877    0.931     0.935

=== Confusion Matrix ===

   a   b   <-- classified as
 140   4 |   a = no
   4  38 |   b = yes
```

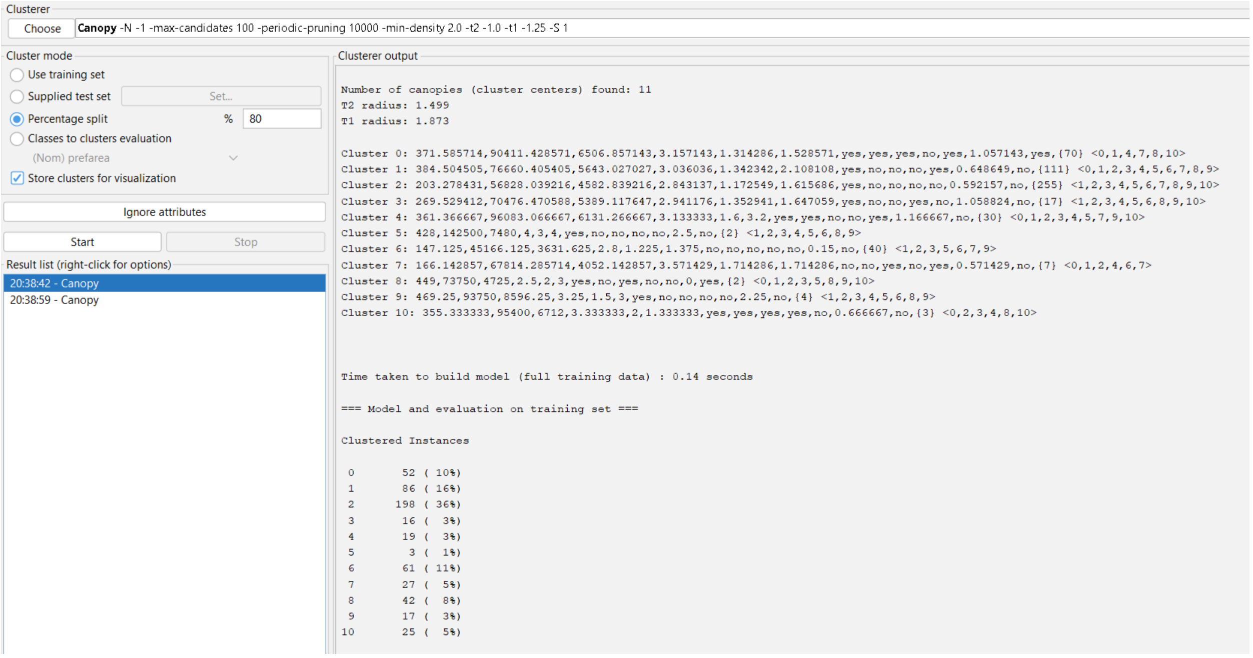
Status
OK

By comparing both the snipps, we can observe that as the percentage split increases the plotting percentage “**Correctly Classified Instances**” increases with using decision tree.

Cluster:



```
g 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1
```

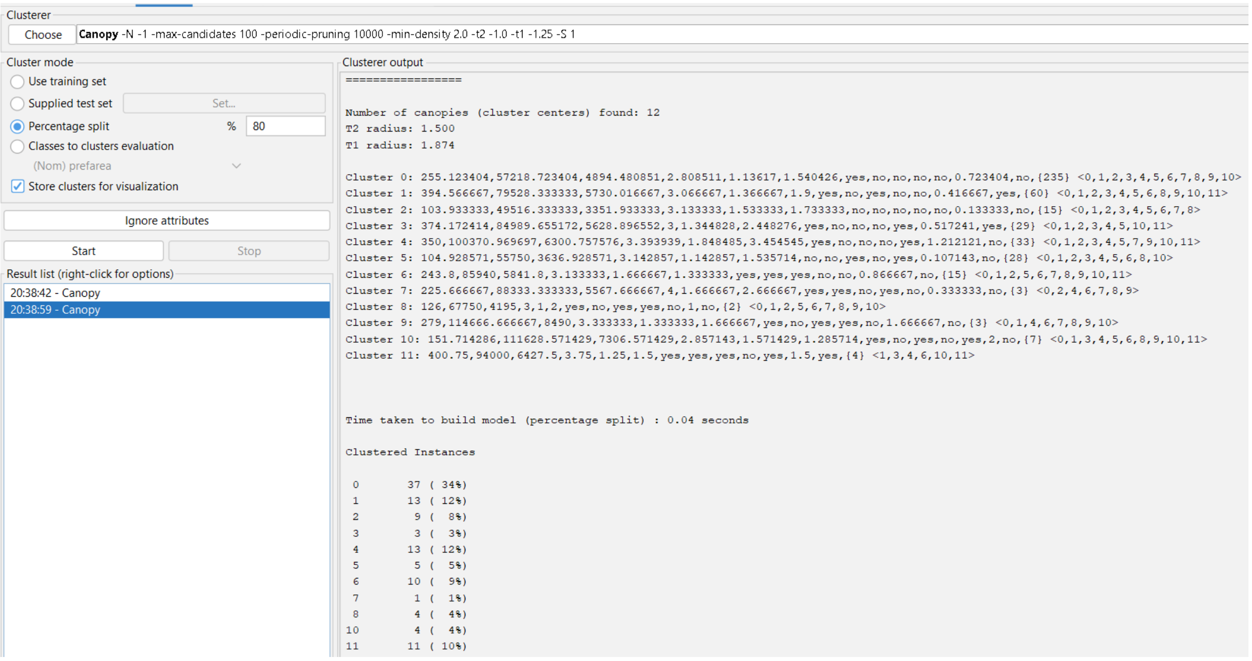


Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

○ Use training set
○ Supplied test set
● Percentage split % 80
○ Classes to clusters evaluation
(Nom) prefarea
☑ Store clusters for visualization

Result list (right-click for options)
20:38:42 - Canopy
20:38:59 - Canopy

```
Number of canopies (cluster centers) found: 11
T2 radius: 1.499
T1 radius: 1.873

Cluster 0: 371.585714,90411.428571,6506.857143,3.157143,1.314286,1.528571,yes,yes,yes,no,yes,1.057143,yes,{70} <0,1,4,7,8,10>
Cluster 1: 384.504505,76660.405405,5643.027027,3.036036,1.342342,2.108108,yes,no,no,no,yes,0.648649,no,{111} <0,1,2,3,4,5,6,7,8,9>
Cluster 2: 203.278431,56828.039216,4582.839216,2.843137,1.172549,1.615686,yes,no,no,no,no,0.592157,no,{255} <1,2,3,4,5,6,7,8,9,10>
Cluster 3: 269.529412,70476.470588,5389.117647,2.941176,1.352941,1.647059,yes,no,no,yes,no,1.058824,no,{17} <1,2,3,4,5,6,8,9,10>
Cluster 4: 361.366667,96083.066667,6131.266667,3.133333,1.6,3.2,yes,yes,no,no,yes,1.166667,no,{30} <0,1,2,3,4,5,7,9,10>
Cluster 5: 428,142500,7480,4,3,4,yes,no,no,no,no,2.5,no,{2} <1,2,3,4,5,6,8,9>
Cluster 6: 147.125,45166.125,3631.625,2.8,1.225,1.375,no,no,no,no,no,0.15,no,{40} <1,2,3,5,6,7,9>
Cluster 7: 166.142857,67814.285714,4052.142857,3.571429,1.714286,1.714286,no,no,yes,no,yes,0.571429,no,{7} <0,1,2,4,6,7>
Cluster 8: 449,73750,4725,2.5,2,3,yes,no,yes,no,no,0,yes,{2} <0,1,2,3,5,8,9,10>
Cluster 9: 469.25,93750,8596.25,3.25,1.5,3,yes,no,no,no,no,2.25,no,{4} <1,2,3,4,5,6,8,9>
Cluster 10: 355.333333,95400,6712,3.333333,2,1.333333,yes,yes,yes,yes,no,0.666667,no,{3} <0,2,3,4,8,10>



Time taken to build model (full training data) : 0.14 seconds

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

Clustered Instances

 0       52 ( 10%)
 1       86 ( 16%)
 2      198 ( 36%)
 3       16 (  3%)
 4       19 (  3%)
 5        3 (  1%)
 6       61 ( 11%)
 7       27 (  5%)
 8       42 (  8%)
 9       17 (  3%)
10       25 (  5%)
```

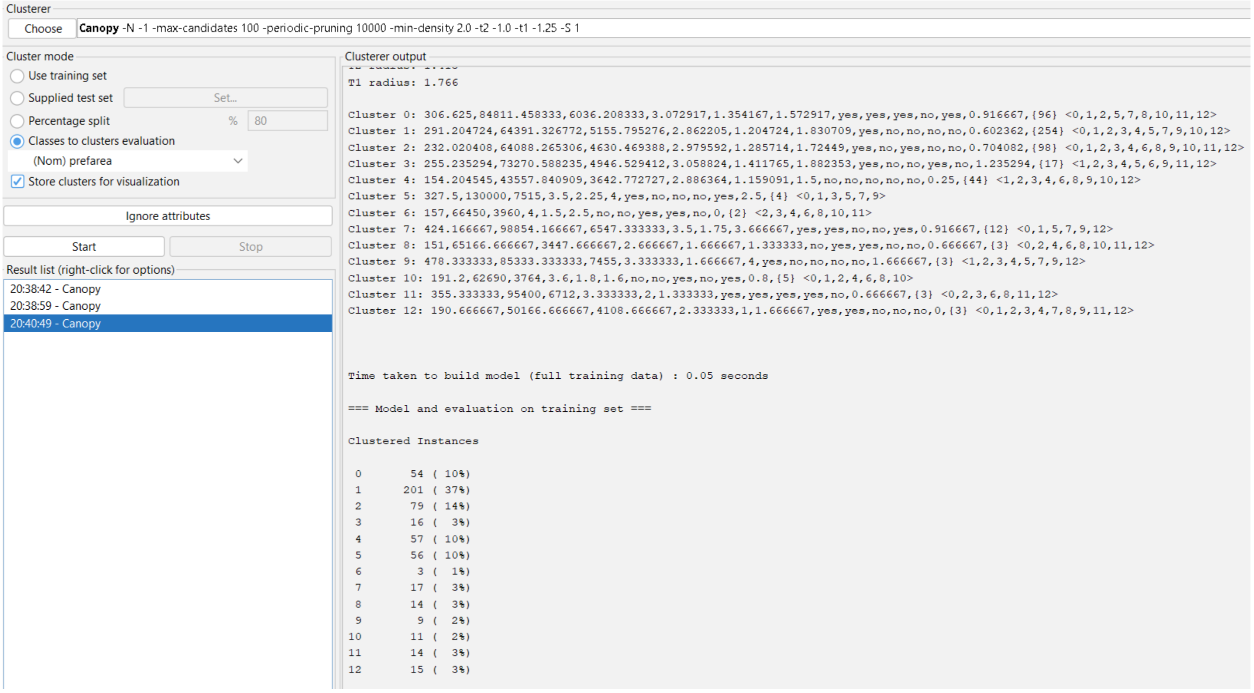


Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

○ Use training set
○ Supplied test set
● Percentage split % 80
○ Classes to clusters evaluation
(Nom) prefarea
☑ Store clusters for visualization

Result list (right-click for options)
20:38:42 - Canopy
20:38:59 - Canopy

```
=================

Number of canopies (cluster centers) found: 12
T2 radius: 1.500
T1 radius: 1.874

Cluster 0: 255.123404,57218.723404,4894.480851,2.808511,1.13617,1.540426,yes,no,no,no,no,0.723404,no,{235} <0,1,2,3,4,5,6,7,8,9,10>
Cluster 1: 394.566667,79528.333333,5730.016667,3.066667,1.366667,1.9,yes,no,yes,no,no,0.416667,yes,{60} <0,1,2,3,4,5,6,8,9,10,11>
Cluster 2: 103.933333,49516.333333,3351.933333,3.133333,1.533333,1.733333,no,no,no,no,no,0.133333,no,{15} <0,1,2,3,4,5,6,7,8>
Cluster 3: 374.172414,84989.655172,5628.896552,3,1.344828,2.448276,yes,no,no,no,yes,0.517241,yes,{29} <0,1,2,3,4,5,10,11>
Cluster 4: 350,100370.969697,6300.757576,3.393939,1.848485,3.454545,yes,no,no,no,yes,1.212121,no,{33} <0,1,2,3,4,5,7,9,10,11>
Cluster 5: 104.928571,55750,3636.928571,3.142857,1.142857,1.535714,no,no,yes,no,yes,0.107143,no,{28} <0,1,2,3,4,5,6,8,10>
Cluster 6: 243.8,85940,5841.8,3.133333,1.666667,1.333333,yes,yes,yes,no,no,0.866667,no,{15} <0,1,2,5,6,7,8,9,10,11>
Cluster 7: 225.666667,88333.333333,5567.666667,4,1.666667,2.666667,yes,yes,no,yes,no,0.333333,no,{3} <0,2,4,6,7,8,9>
Cluster 8: 126,67750,4195,3,1,2,yes,no,yes,yes,no,1,no,{2} <0,1,2,5,6,7,8,9,10>
Cluster 9: 279,114666.666667,8490,3.333333,1.333333,1.666667,yes,no,yes,yes,no,1.666667,no,{3} <0,1,4,6,7,8,9,10>
Cluster 10: 151.714286,111628.571429,7306.571429,2.857143,1.571429,1.285714,yes,no,yes,no,yes,2,no,{7} <0,1,3,4,5,6,8,9,10,11>
Cluster 11: 400.75,94000,6427.5,3.75,1.25,1.5,yes,yes,yes,no,yes,1.5,yes,{4} <1,3,4,6,10,11>



Time taken to build model (percentage split) : 0.04 seconds

Clustered Instances

 0       37 ( 34%)
 1       13 ( 12%)
 2        9 (  8%)
 3        3 (  3%)
 4       13 ( 12%)
 5        5 (  5%)
 6       10 (  9%)
 7        1 (  1%)
 8        4 (  4%)
10        4 (  4%)
11       11 ( 10%)
```



Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Percentage split: 80

(Nom) prefarea

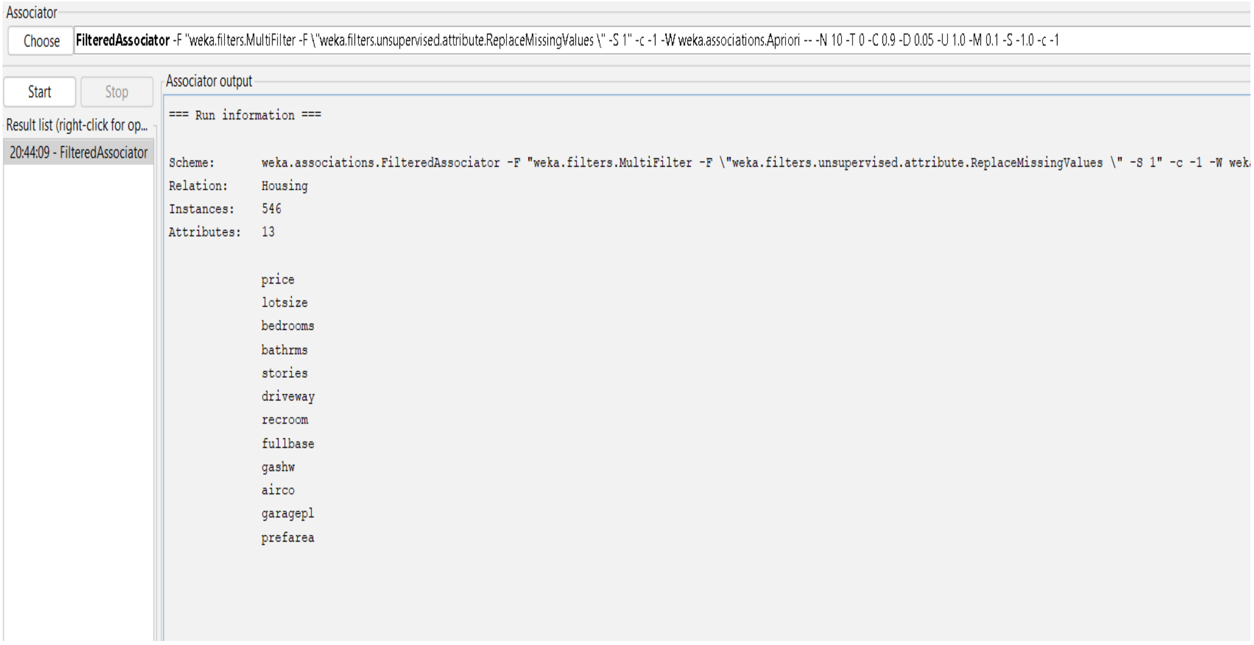
☑ Store clusters for visualization

20:38:42 - Canopy
20:38:59 - Canopy
20:40:49 - Canopy

```
T1 radius: 1.766

Cluster 0: 306.625,84811.458333,6036.208333,3.072917,1.354167,1.572917,yes,yes,yes,no,yes,0.916667,{96} <0,1,2,5,7,8,10,11,12>
Cluster 1: 291.204724,64391.326772,5155.795276,2.862205,1.204724,1.830709,yes,no,no,no,no,0.602362,{254} <0,1,2,3,4,5,7,9,10,12>
Cluster 2: 232.020408,64088.265306,4630.469388,2.979592,1.285714,1.72449,yes,no,yes,no,no,0.704082,{98} <0,1,2,3,4,6,8,9,10,11,12>
Cluster 3: 255.235294,73270.588235,4946.529412,3.058824,1.411765,1.882353,yes,no,no,yes,no,1.235294,{17} <1,2,3,4,5,6,9,11,12>
Cluster 4: 154.204545,43557.840909,3642.772727,2.886364,1.159091,1.5,no,no,no,no,no,0.25,{44} <1,2,3,4,6,8,9,10,12>
Cluster 5: 327.5,130000,7515,3.5,2.25,4,yes,no,no,no,yes,2.5,{4} <0,1,3,5,7,9>
Cluster 6: 157,66450,3960,4,1.5,2.5,no,no,yes,yes,no,0,{2} <2,3,4,6,8,10,11>
Cluster 7: 424.166667,98854.166667,6547.333333,3.5,1.75,3.666667,yes,yes,no,no,yes,0.916667,{12} <0,1,5,7,9,12>
Cluster 8: 151,65166.666667,3447.666667,2.666667,1.666667,1.333333,no,yes,yes,no,no,0.666667,{3} <0,2,4,6,8,10,11,12>
Cluster 9: 478.333333,85333.333333,7455,3.333333,1.666667,4,yes,no,no,no,no,1.666667,{3} <1,2,3,4,5,7,9,12>
Cluster 10: 191.2,62690,3764,3.6,1.8,1.6,no,no,yes,no,yes,0.8,{5} <0,1,2,4,6,8,10>
Cluster 11: 355.333333,95400,6712,3.333333,2,1.333333,yes,yes,yes,yes,no,0.666667,{3} <0,2,3,6,8,11,12>
Cluster 12: 190.666667,50166.666667,4108.666667,2.333333,1,1.666667,yes,yes,no,no,no,0,{3} <0,1,2,3,4,7,8,9,11,12>



Time taken to build model (full training data) : 0.05 seconds

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

Clustered Instances

 0       54 ( 10%)
 1      201 ( 37%)
 2       79 ( 14%)
 3       16 (  3%)
 4       57 ( 10%)
 5       56 ( 10%)
 6        3 (  1%)
 7       17 (  3%)
 8       14 (  3%)
 9        9 (  2%)
10       11 (  2%)
11       14 (  3%)
12       15 (  3%)
```

By seeing the above snipps, we can observe that based on the 10 epocs done by the clusters, based on the pooling layers (clusters) we can determine the creation and training of model its time.

Associator:



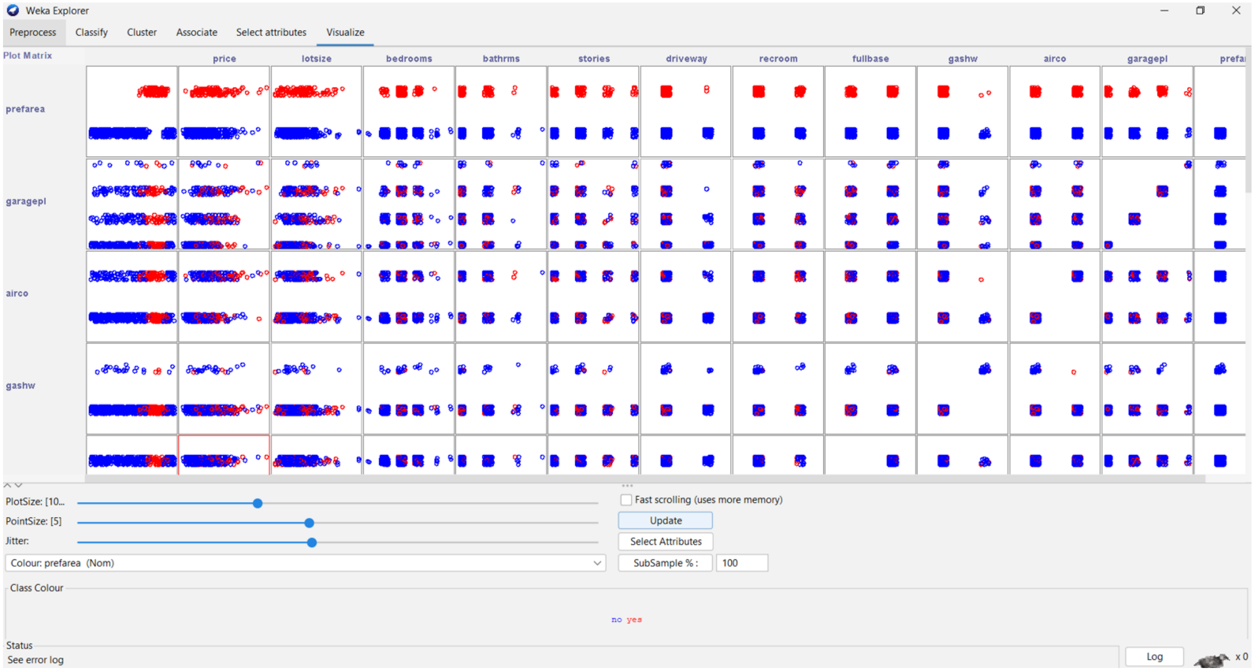
FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \" -S 1" -c -1 -W weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

20:44:09 - FilteredAssociator

```
=== Run information ===

Scheme:       weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \" -S 1" -c -1 -W wek
Relation:     Housing
Instances:    546
Attributes:   13

              price
              lotsize
              bedrooms
              bathrms
              stories
              driveway
              recroom
              fullbase
              gashw
              airco
              garagepl
              prefarea
```

Associators are used to determine Scheme, Relations, Instances and Attributes.

Visualize:



PlotSize: [10...

PointSize: [5]

Jitter:

Colour: prefarea (Nom)

□ Fast scrolling (uses more memory)

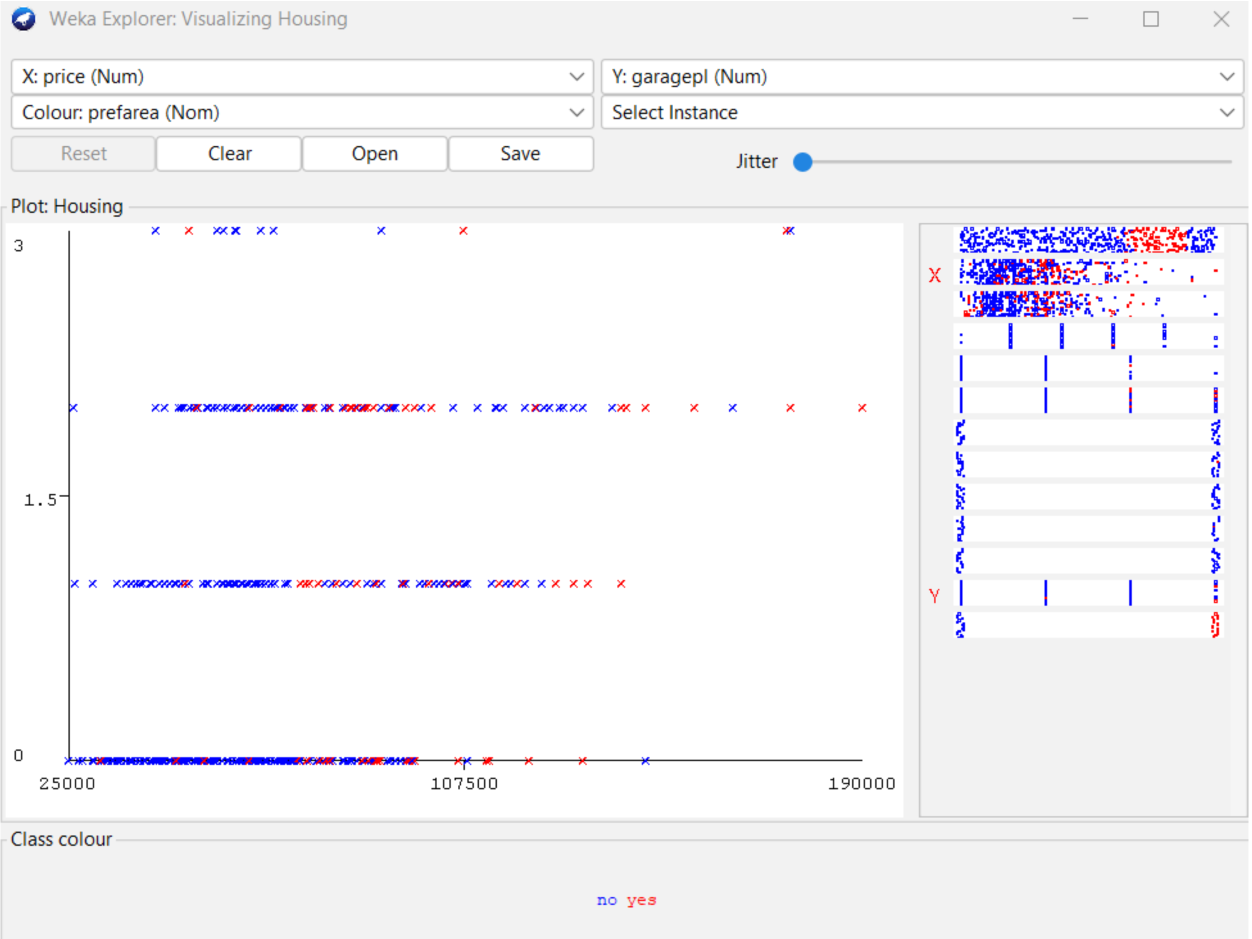
SubSample % : 100

Status

See error log

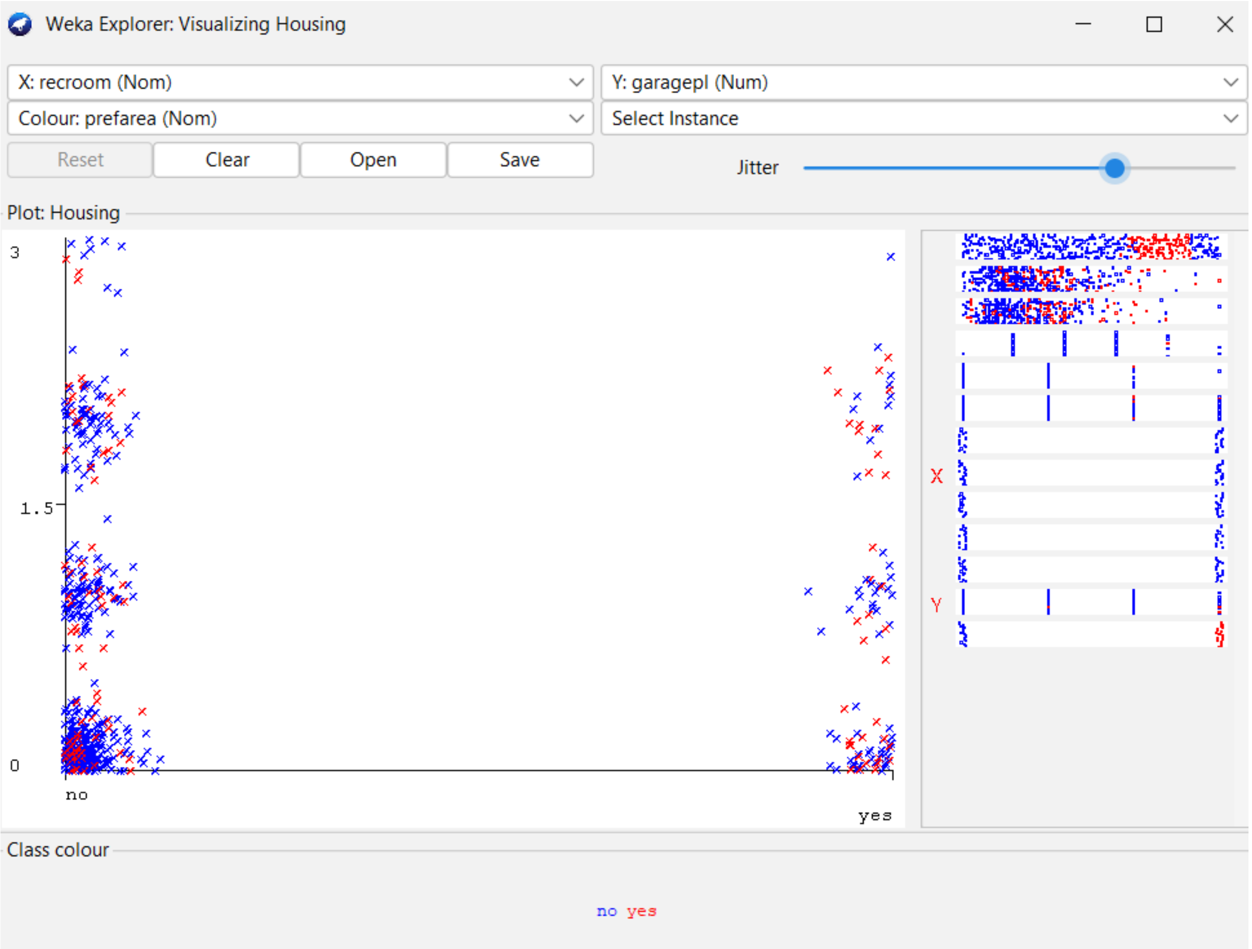
Visualization of each to each other column by adjusting Plotsize, Pointsize and Jitter.

Where plotting can be adjusted based on the values.



X-axis: Price vs Y-axis: garagepl

Colour: prefarea



Weka Explorer: Visualizing Housing

X: recroom (Nom)

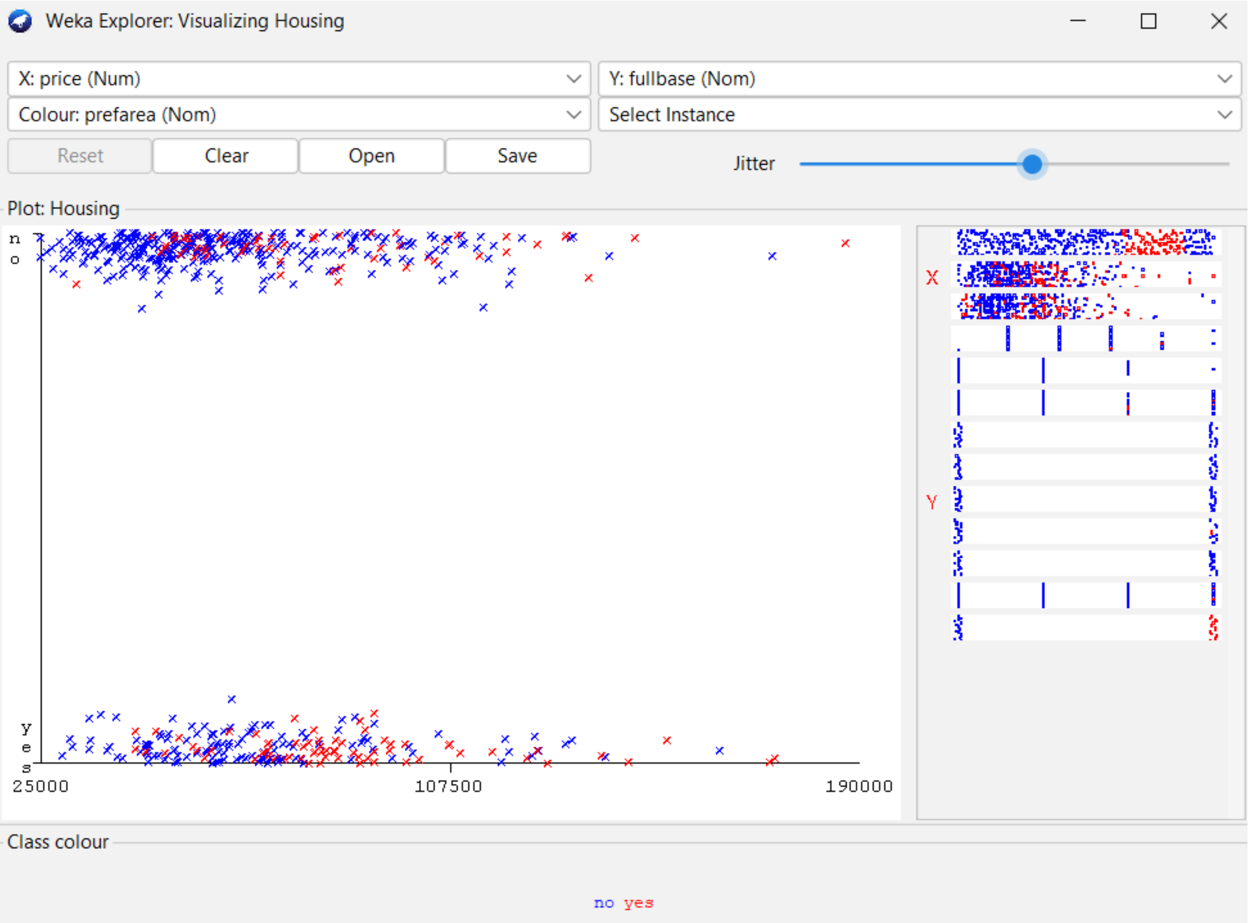
Y: garagepl (Num)

Colour: prefarea (Nom)

Select Instance

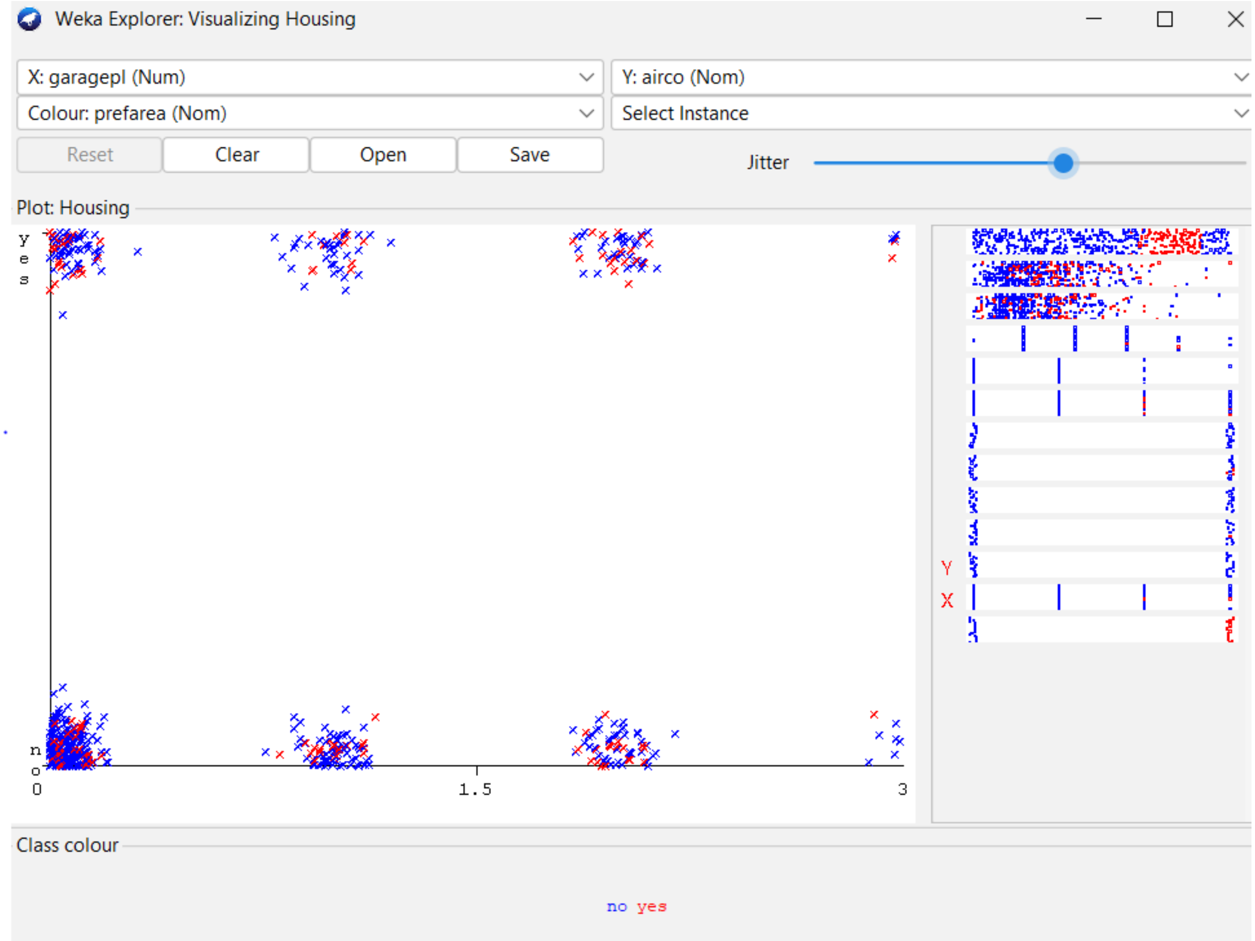
X-axis: recroom vs Y-axis: garagepl

Colour: prefarea



X-axis: Price vs Y-axis: fullbase

Colour: prefarea



X-axis: garagepl vs Y-axis: airco

Colour: prefarea