Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

DATA WAREHOUSING AND DATA MINING

NAME : MOTHISHWARAN C.

REG No : 19MID0017

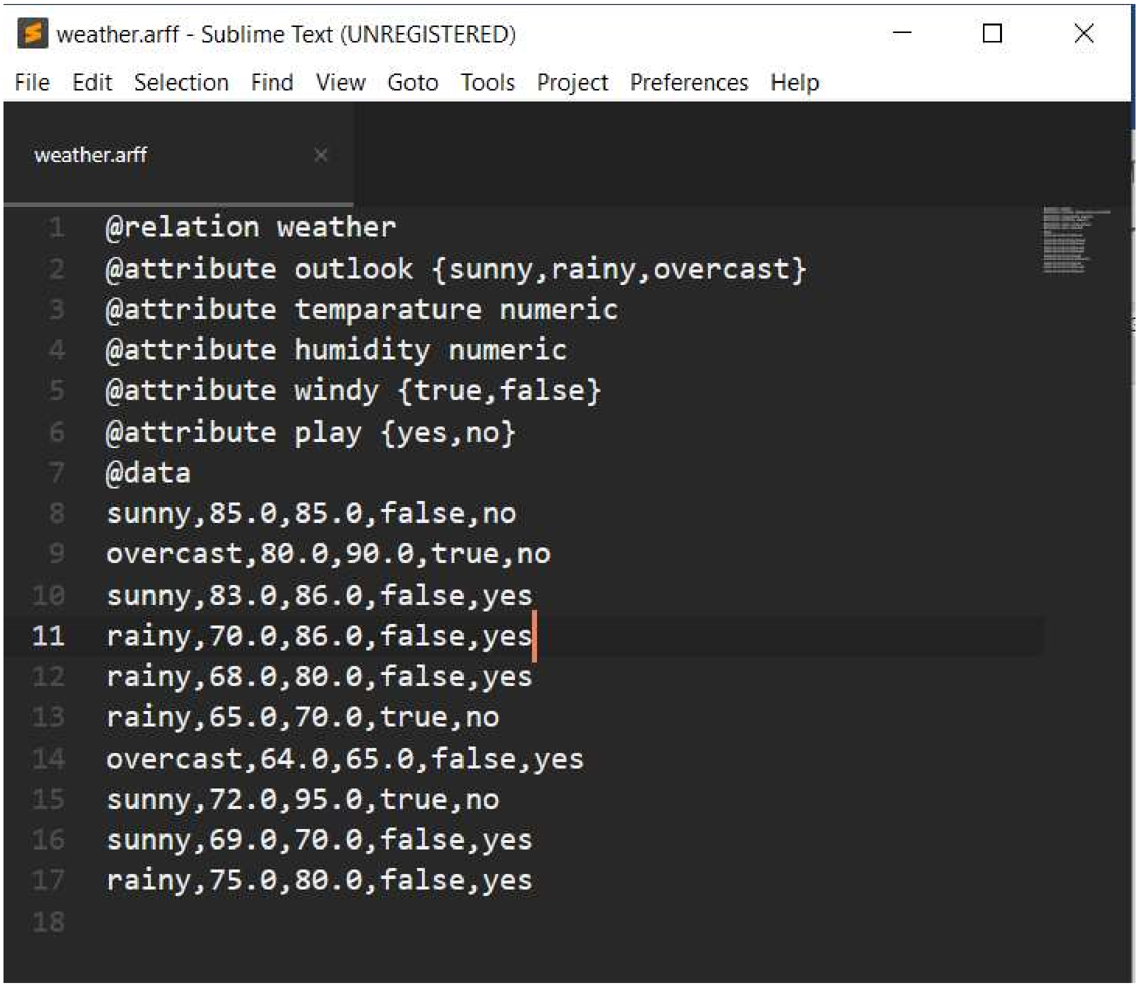
COURSE CODE : CSI3010

SLOT : L39+L40

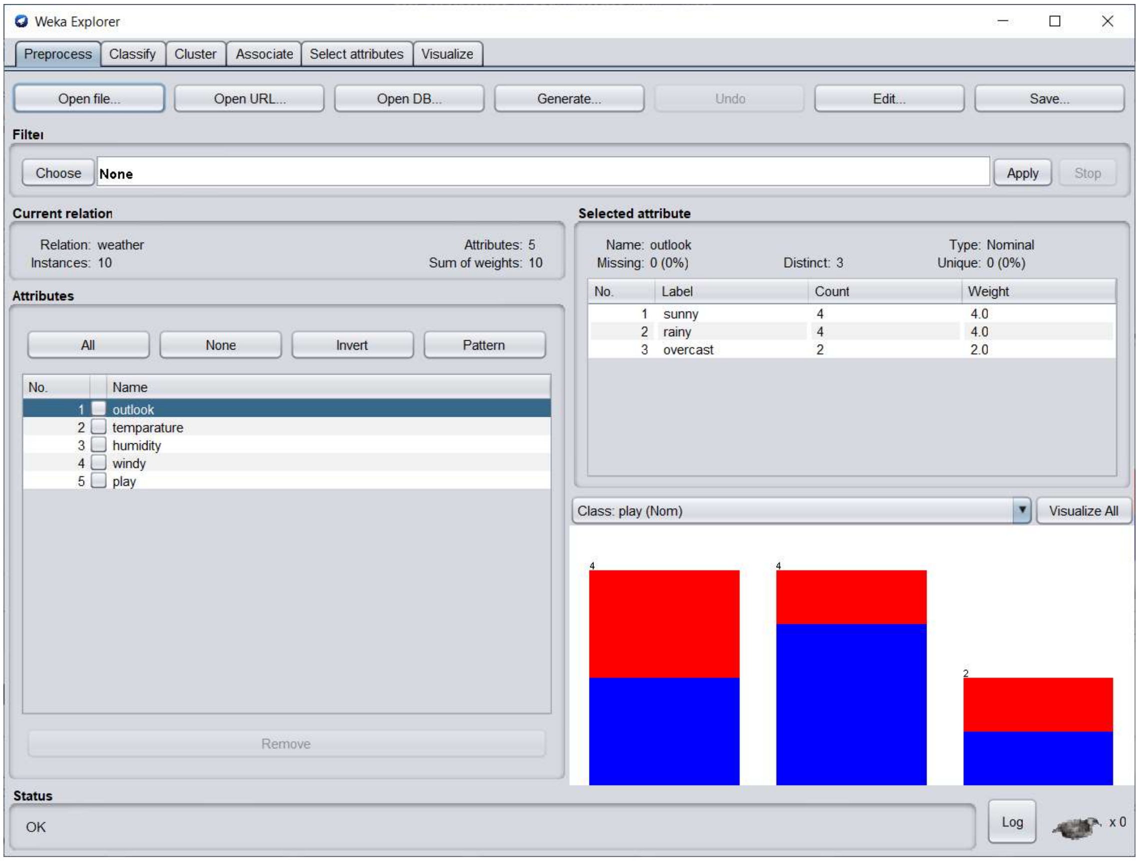
FACULTY : DR. CHELLATHAMILAN T.

DISCRETIZATION OF DATA USING WEKA

1) Weather dataset:



```
@relation weather
@attribute outlook {sunny,rainy,overcast}
@attribute temparature numeric
@attribute humidity numeric
@attribute windy {true,false}
@attribute play {yes,no}
@data
sunny,85.0,85.0,false,no
overcast,80.0,90.0,true,no
sunny,83.0,86.0,false,yes
rainy,70.0,86.0,false,yes
rainy,68.0,80.0,false,yes
rainy,65.0,70.0,true,no
overcast,64.0,65.0,false,yes
sunny,72.0,95.0,true,no
sunny,69.0,70.0,false,yes
rainy,75.0,80.0,false,yes

```



Filter: None

Current relation

Relation: weather
Instances: 10
Attributes: 5
Sum of weights: 10

Attributes

No.		Name
1	□	outlook
2	□	temparature
3	□	humidity
4	□	windy
5	□	play

Selected attribute

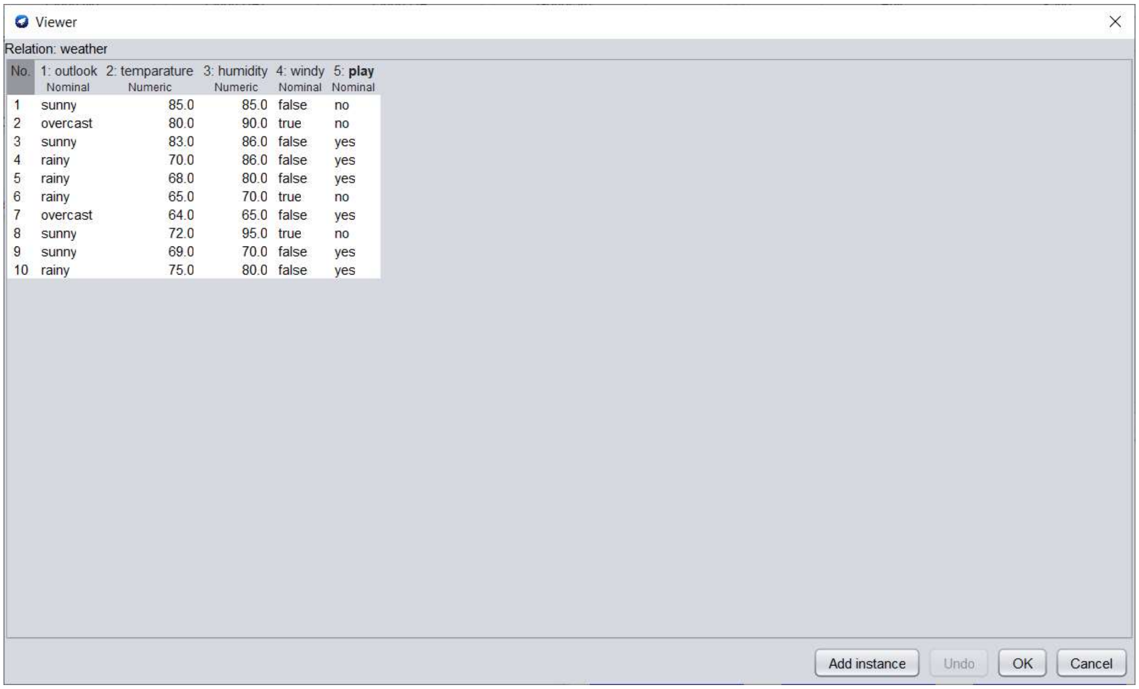
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	4	4.0
2	rainy	4	4.0
3	overcast	2	2.0

Class: play (Nom)

Status

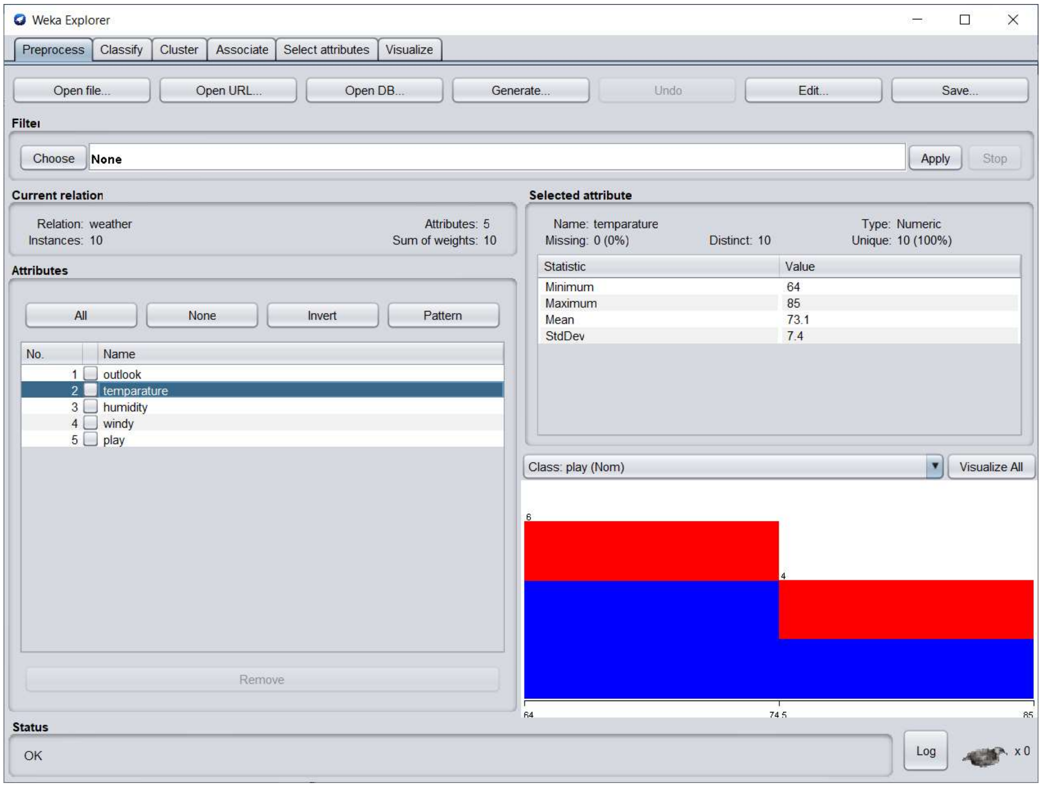
OK



Relation: weather

No.	1: outlook Nominal	2: temparature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	85.0	85.0	false	no
2	overcast	80.0	90.0	true	no
3	sunny	83.0	86.0	false	yes
4	rainy	70.0	86.0	false	yes
5	rainy	68.0	80.0	false	yes
6	rainy	65.0	70.0	true	no
7	overcast	64.0	65.0	false	yes
8	sunny	72.0	95.0	true	no
9	sunny	69.0	70.0	false	yes
10	rainy	75.0	80.0	false	yes

Before Discretization (no bins added):



Filter: None

Current relation

Relation: weather Attributes: 5
Instances: 10 Sum of weights: 10

Selected attribute

Name: temparature Type: Numeric
Missing: 0 (0%) Distinct: 10 Unique: 10 (100%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.1
StdDev	7.4

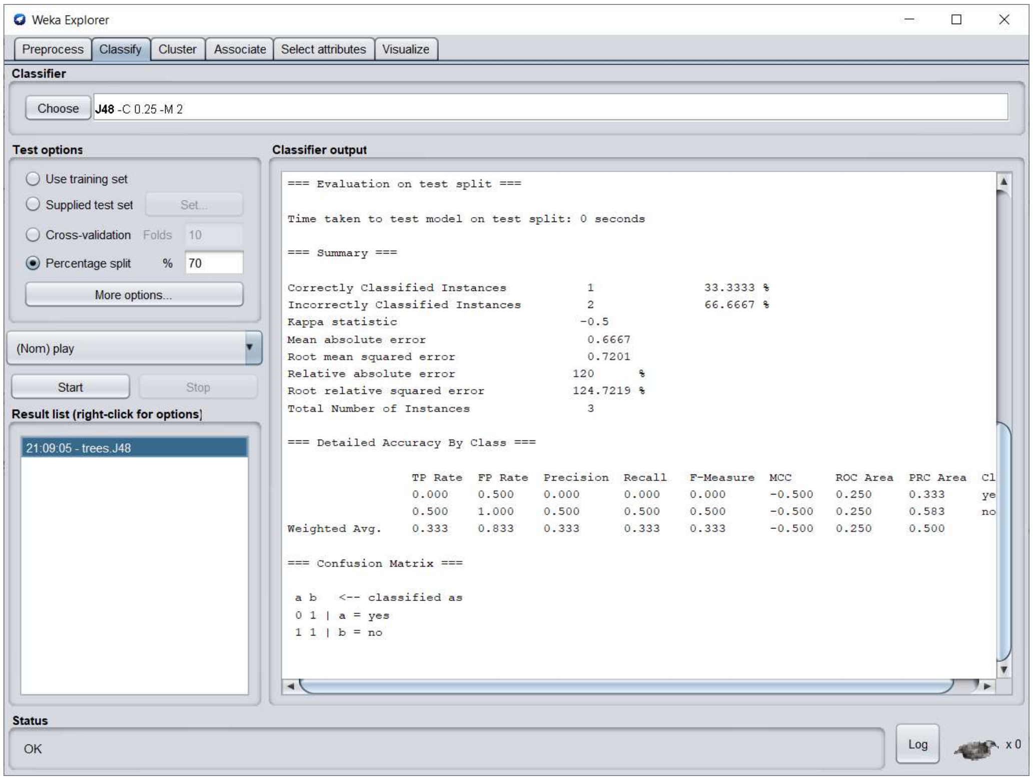
Attributes

No.		Name
1	□	outlook
2	□	temparature
3	□	humidity
4	□	windy
5	□	play

Class: play (Nom)

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) play

Result list (right-click for options)

21:09:05 - trees.J48

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           1               33.3333 %
Incorrectly Classified Instances         2               66.6667 %
Kappa statistic                         -0.5
Mean absolute error                      0.6667
Root mean squared error                  0.7201
Relative absolute error                120      %
Root relative squared error            124.7219 %
Total Number of Instances                3

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.000    0.500    0.000      0.000    0.000      -0.500   0.250     0.333     ye
                 0.500    1.000    0.500      0.500    0.500      -0.500   0.250     0.583     no
Weighted Avg.    0.333    0.833    0.333      0.333    0.333      -0.500   0.250     0.500

=== Confusion Matrix ===

 a b   <-- classified as
 0 1 | a = yes
 1 1 | b = no
```

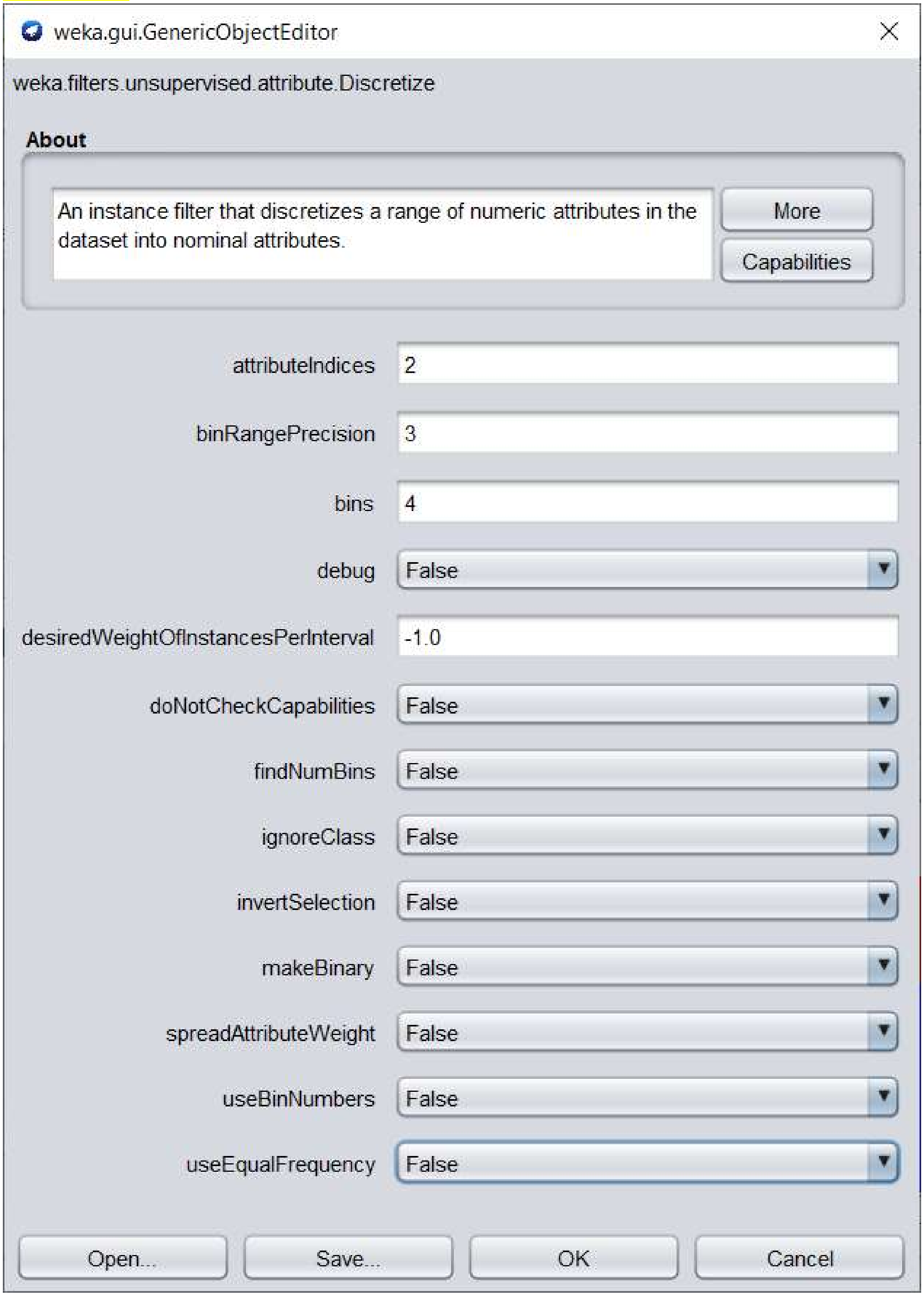
Status

OK

Accuracy: 33.333%

After Discretization (4 bins added):

No. of. Bins: 4



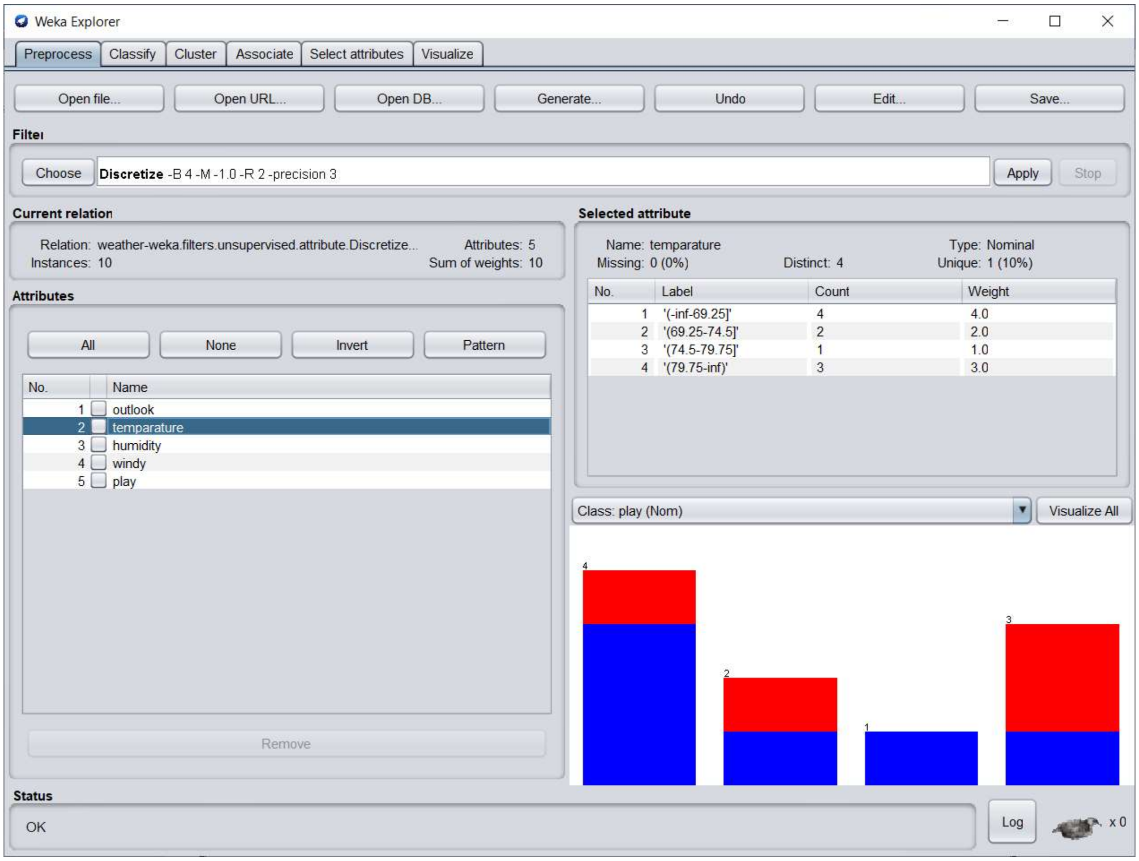
weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	2
binRangePrecision	3
bins	4
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False



Filter

Discretize -B 4 -M -1.0 -R 2 -precision 3

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize... Attributes: 5
Instances: 10 Sum of weights: 10

Selected attribute

Name: temparature Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 1 (10%)

No.	Label	Count	Weight
1	'(-inf-69.25]'	4	4.0
2	'(69.25-74.5]'	2	2.0
3	'(74.5-79.75]'	1	1.0
4	'(79.75-inf)'	3	3.0

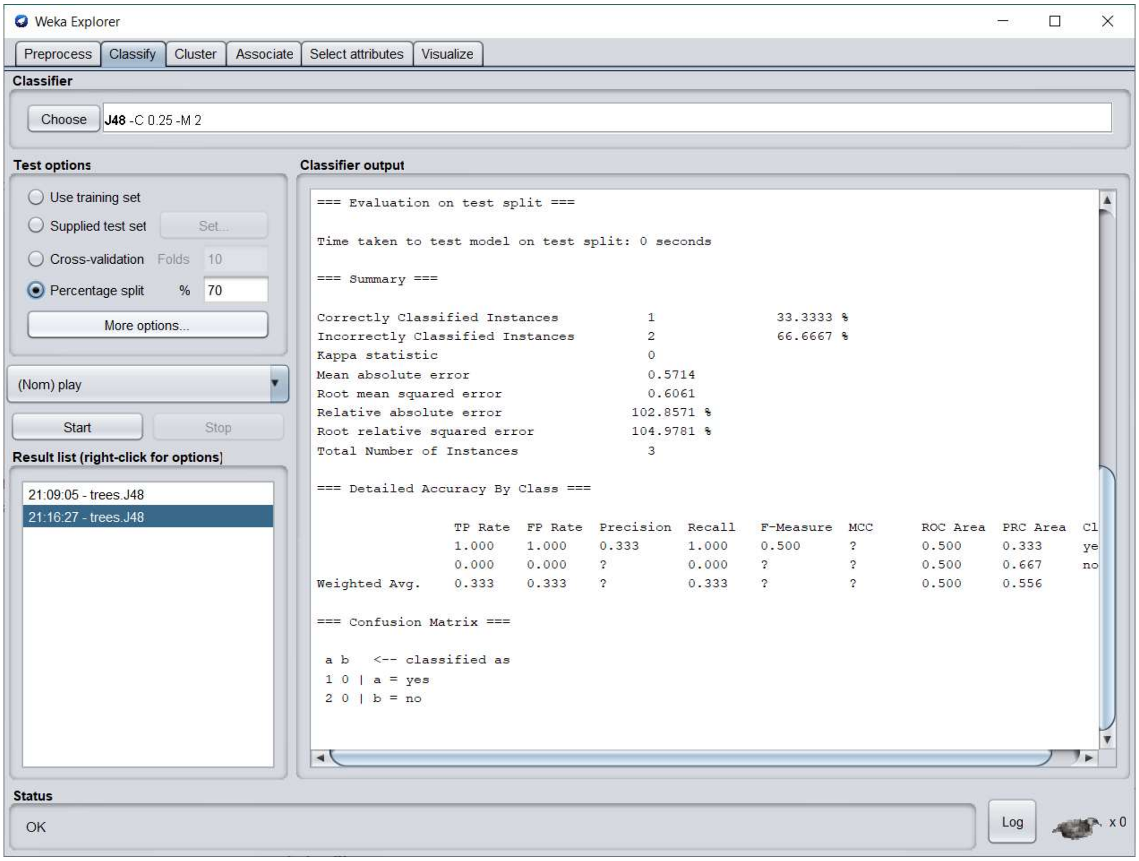
Attributes

No.	Name
1	□ outlook
2	□ temparature
3	□ humidity
4	□ windy
5	□ play

Class: play (Nom)

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) play

Result list (right-click for options)

21:09:05 - trees.J48
21:16:27 - trees.J48

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           1               33.3333 %
Incorrectly Classified Instances         2               66.6667 %
Kappa statistic                          0
Mean absolute error                      0.5714
Root mean squared error                  0.6061
Relative absolute error                102.8571 %
Root relative squared error            104.9781 %
Total Number of Instances                3

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 1.000    1.000    0.333      1.000    0.500      ?        0.500     0.333     ye
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.667     no
Weighted Avg.    0.333    0.333    ?          0.333    ?          ?        0.500     0.556

=== Confusion Matrix ===

 a b   <-- classified as
 1 0 | a = yes
 2 0 | b = no
```

Status

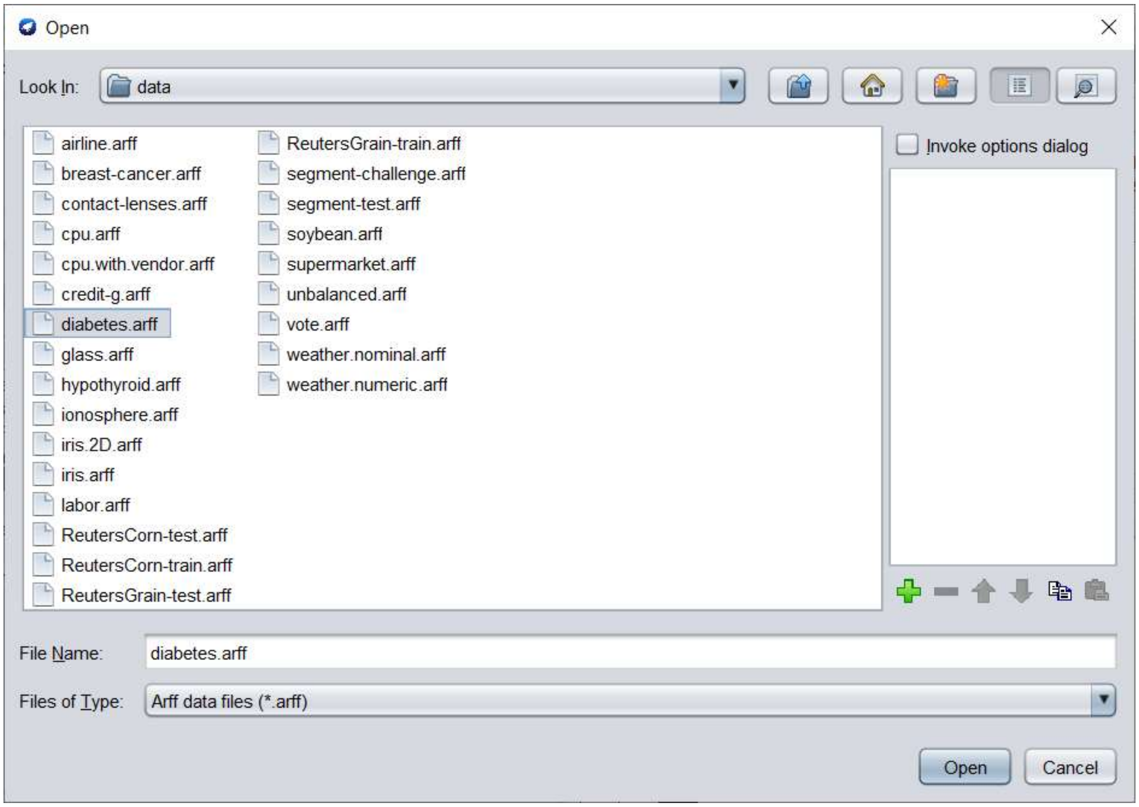
OK

Accuracy: 33.333%

Accuracy varies when we change the split percentages.

Split %	Before Discretization	After Discretization
70%	33.33%	33.33%
50%	40%	40%

1) Pima Diabetes Dataset:

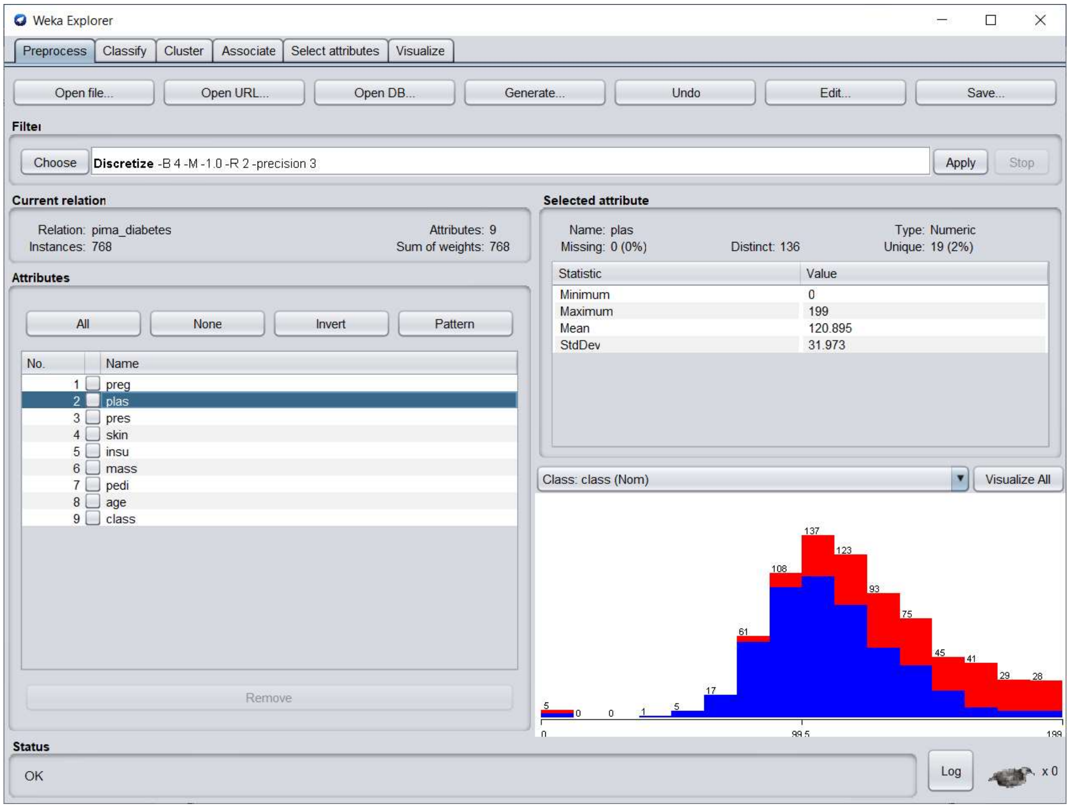


Look In: data

File Name: diabetes.arff

Files of Type: Arff data files (*.arff)

Before Discretization (no bins added):



Filter

Discretize -B 4 -M -1.0 -R 2 -precision 3

Current relation

Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

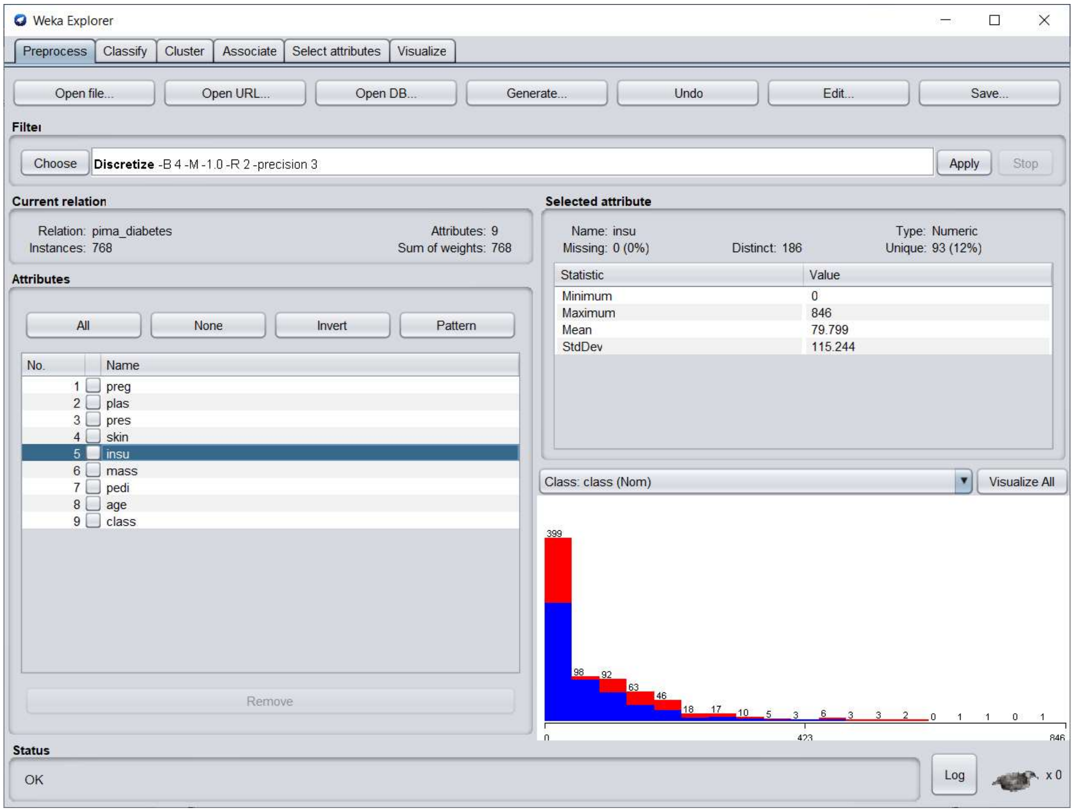
Selected attribute

Name: plas
Missing: 0 (0%)
Distinct: 136
Type: Numeric
Unique: 19 (2%)

Statistic	Value
Minimum	0
Maximum	199
Mean	120.895
StdDev	31.973

Status

OK



Filter

Discretize -B 4 -M -1.0 -R 2 -precision 3

Current relation

Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

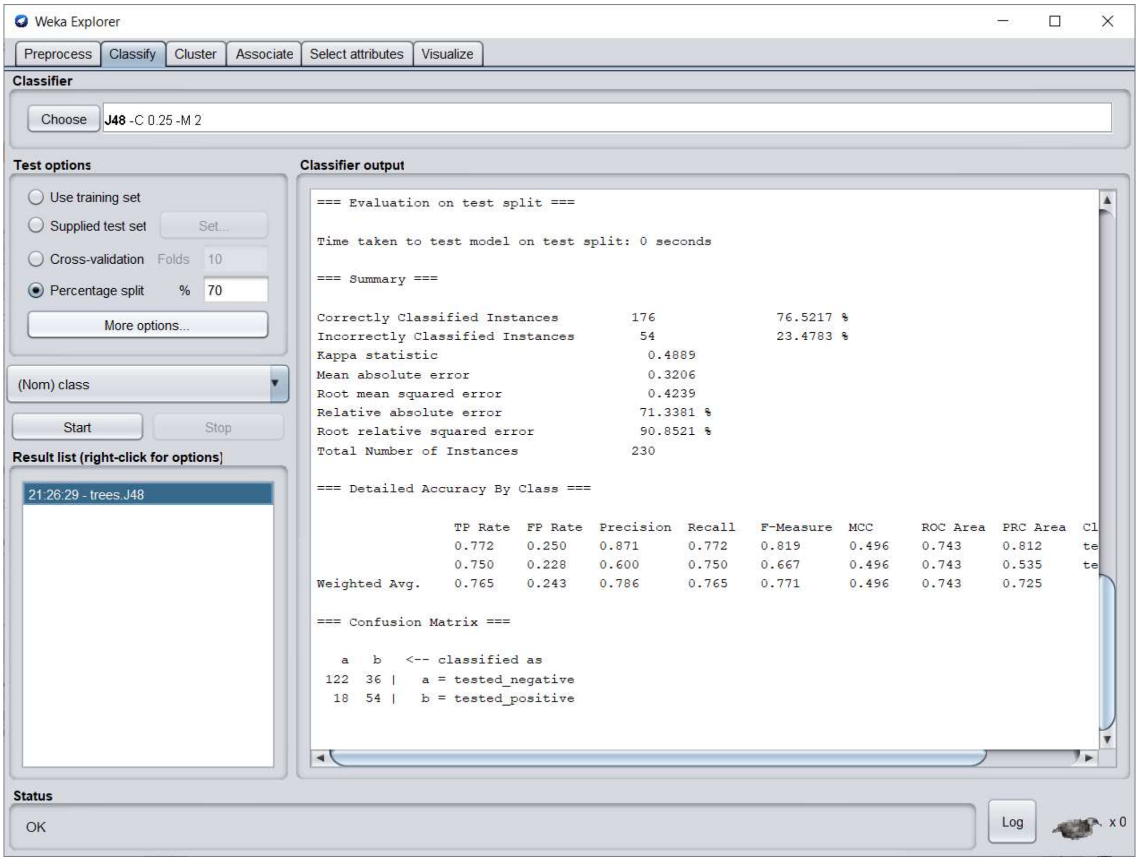
Selected attribute

Name: insu
Missing: 0 (0%)
Distinct: 186
Type: Numeric
Unique: 93 (12%)

Statistic	Value
Minimum	0
Maximum	846
Mean	79.799
StdDev	115.244

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 70

(Nom) class

Result list (right-click for options)

21:26:29 - trees.J48

Classifier output

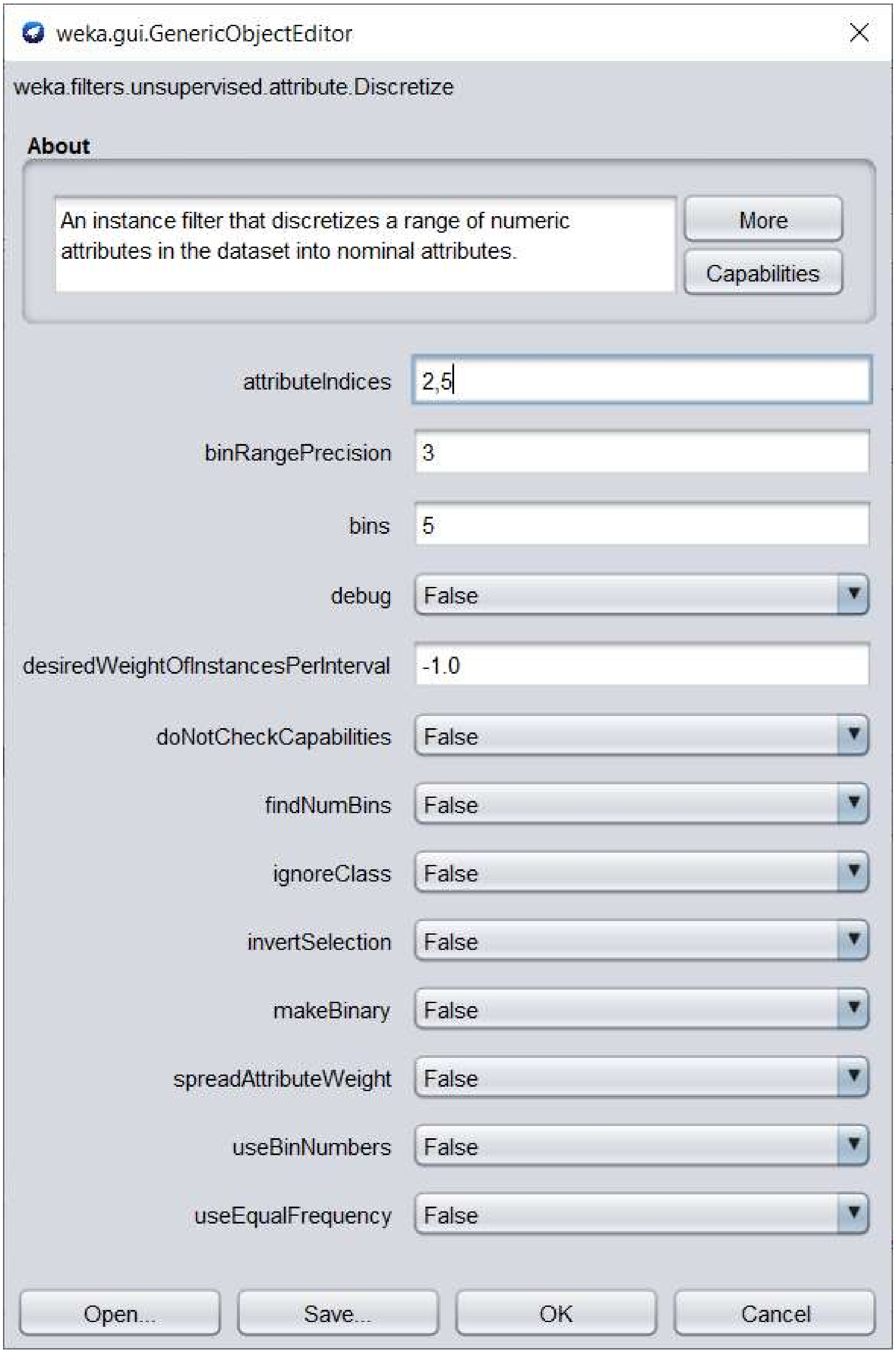
```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         176               76.5217 %
Incorrectly Classified Instances        54               23.4783 %
Kappa statistic                          0.4889
Mean absolute error                      0.3206
Root mean squared error                  0.4239
Relative absolute error                 71.3381 %
Root relative squared error             90.8521 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.772    0.250    0.871      0.772    0.819      0.496    0.743     0.812     te
                 0.750    0.228    0.600      0.750    0.667      0.496    0.743     0.535     te
Weighted Avg.    0.765    0.243    0.786      0.765    0.771      0.496    0.743     0.725

=== Confusion Matrix ===

   a   b   <-- classified as
 122  36 |   a = tested_negative
  18  54 |   b = tested_positive
```

Status

OK

Accuracy: 76.52%

After Discretization (5 bins added):



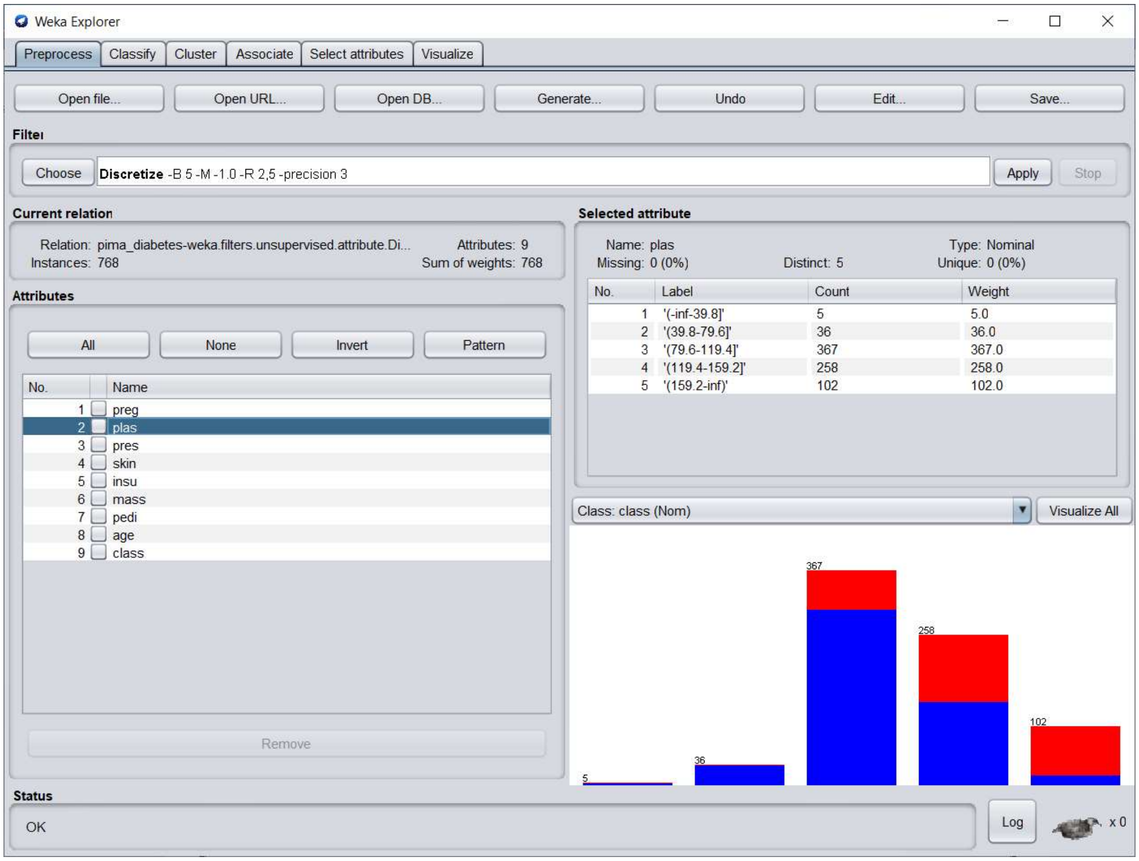
weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	2,5
binRangePrecision	3
bins	5
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False



Filter: **Discretize** -B 5 -M -1.0 -R 2,5 -precision 3

Relation: pima_diabetes-weka.filters.unsupervised.attribute.Di...
Attributes: 9
Instances: 768
Sum of weights: 768

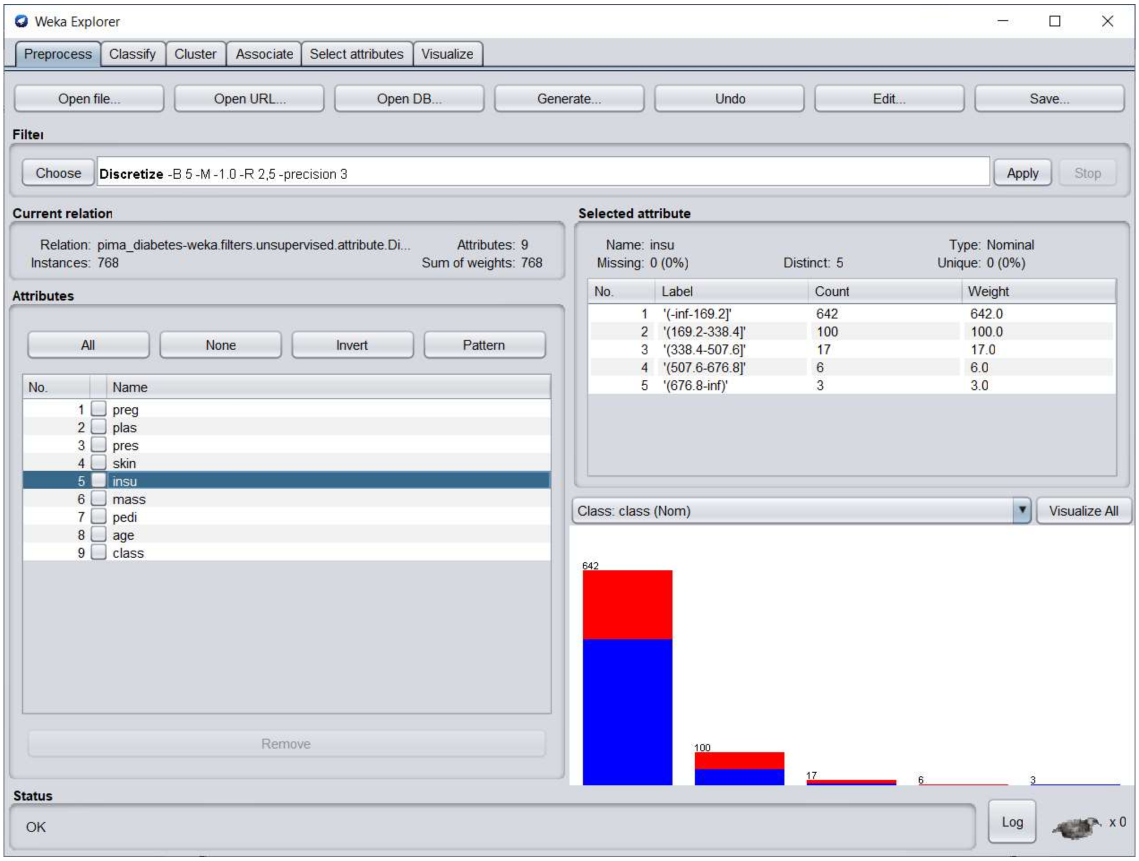
No.	Name
1	☐ preg
2	☐ plas
3	☐ pres
4	☐ skin
5	☐ insu
6	☐ mass
7	☐ pedi
8	☐ age
9	☐ class

Name: plas
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-39.8]'	5	5.0
2	'(39.8-79.6]'	36	36.0
3	'(79.6-119.4]'	367	367.0
4	'(119.4-159.2]'	258	258.0
5	'(159.2-inf)'	102	102.0

Class: class (Nom)

Status: OK



Filter: **Discretize** -B 5 -M -1.0 -R 2,5 -precision 3

Relation: pima_diabetes-weka.filters.unsupervised.attribute.Di...
Attributes: 9
Instances: 768
Sum of weights: 768

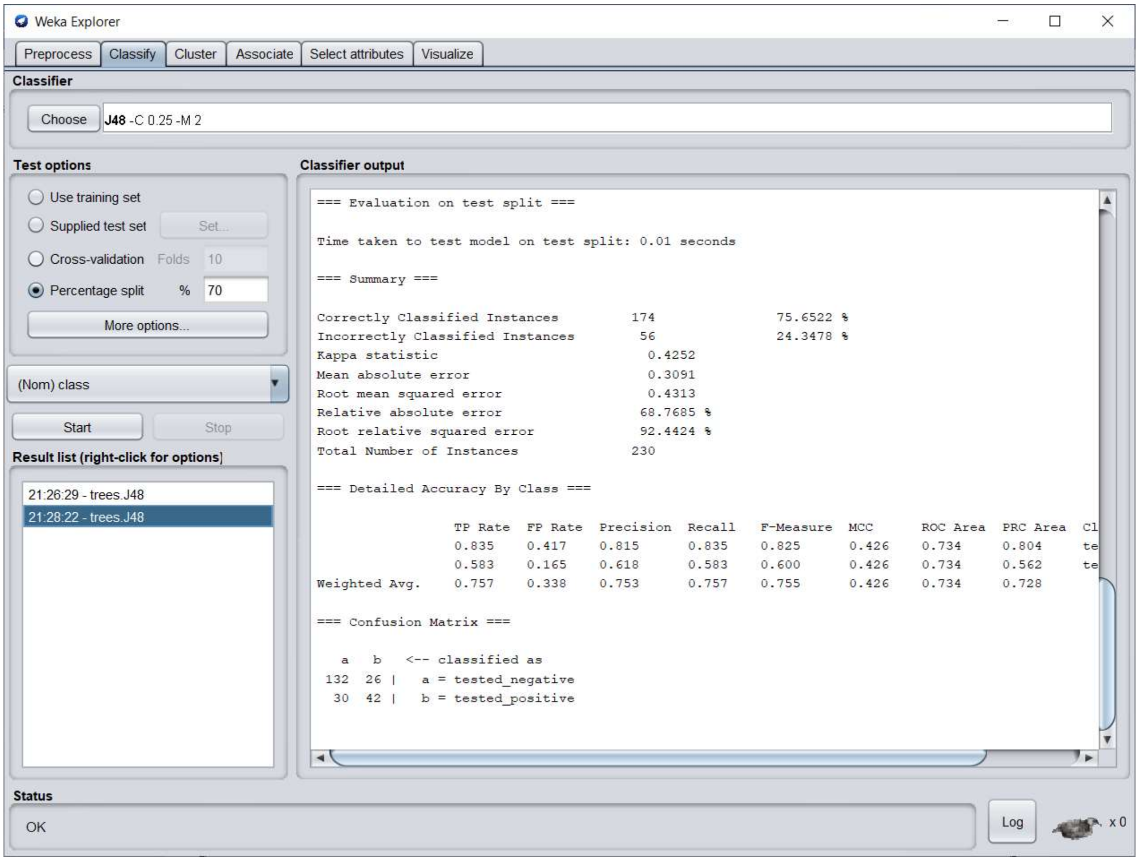
No.	Name
1	☐ preg
2	☐ plas
3	☐ pres
4	☐ skin
5	☐ insu
6	☐ mass
7	☐ pedi
8	☐ age
9	☐ class

Name: insu
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-169.2]'	642	642.0
2	'(169.2-338.4]'	100	100.0
3	'(338.4-507.6]'	17	17.0
4	'(507.6-676.8]'	6	6.0
5	'(676.8-inf)'	3	3.0

Class: class (Nom)

Status: OK



Classifier: J48 -C 0.25 -M 2

Test options:

○ Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) class

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         174               75.6522 %
Incorrectly Classified Instances        56               24.3478 %
Kappa statistic                          0.4252
Mean absolute error                      0.3091
Root mean squared error                  0.4313
Relative absolute error                 68.7685 %
Root relative squared error             92.4424 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.835    0.417    0.815      0.835    0.825      0.426    0.734     0.804     te
                 0.583    0.165    0.618      0.583    0.600      0.426    0.734     0.562     te
Weighted Avg.    0.757    0.338    0.753      0.757    0.755      0.426    0.734     0.728

=== Confusion Matrix ===

   a   b   <-- classified as
 132  26 |   a = tested_negative
  30  42 |   b = tested_positive
```

Status: OK

Accuracy: 75.65%

Accuracy varies when we change the split percentages.

Split %	Before Discretization	After Discretization
70%	76.52%	75.65%
50%	74.21%	71.35%