EXPERIMENT NO. 5

Aim: To implement Data Discretization and Visualization.

Requirement: Windows OS and Weka Tool.

Theory:

Data Discretization: Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

Table before Discretization

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

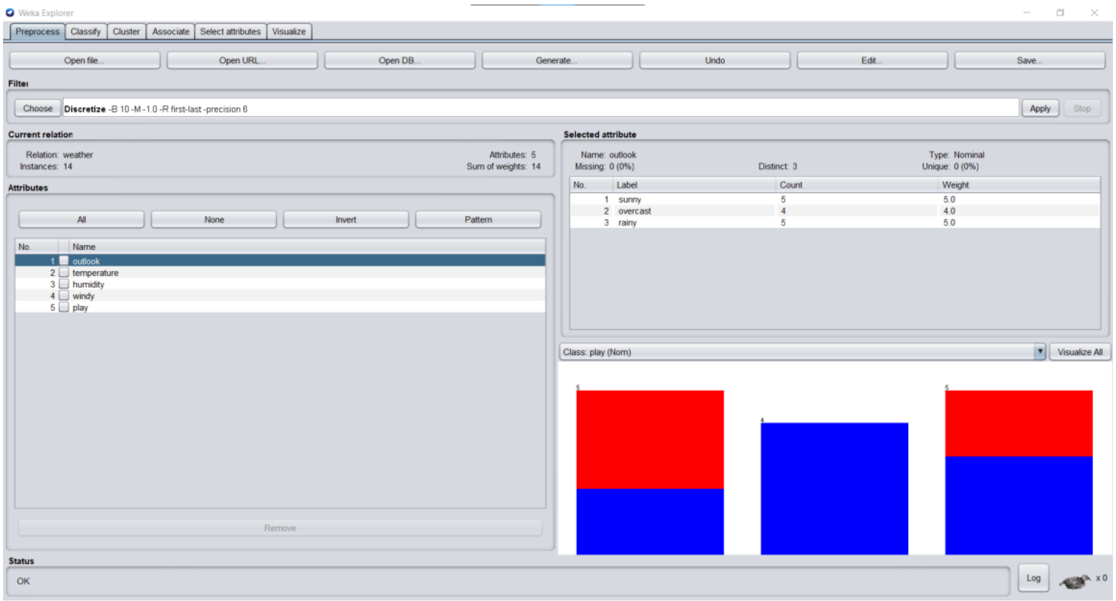
Table after Discretization

Data Visualization: Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

A dashboard is an information visualization tool. It helps you monitor events or activities at a glance by providing insights on one or more pages or screens. Unlike an infographic, which presents a static graphical representation, a dashboard conveys real-time information by pulling complex data points directly from large data sets. An interactive dashboard makes it easy to sort, filter, or drill into different types of data as needed. Data science techniques can be used to identify what is happening, why it's happening, and what will happen next at speed.

As the amount of big data increases, more people are using data visualization tools to access insights on their computer and on mobile devices. Dashboards are used by business people, data analysts, and data scientists to make data-driven business decisions.

Weather Data before Discretization:



Filter

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

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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

Selected attribute

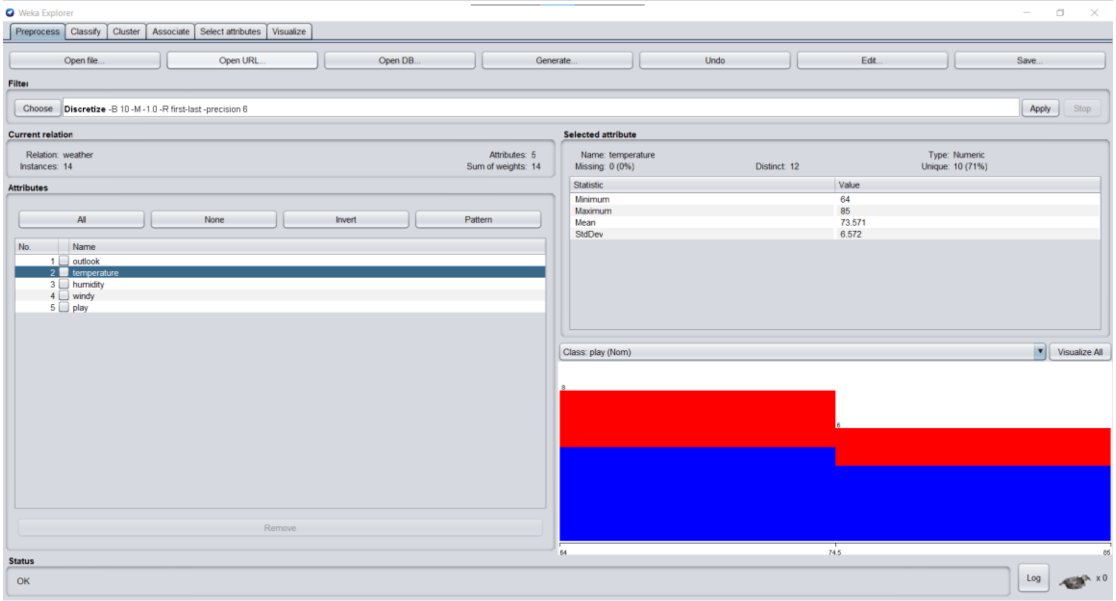
Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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

Selected attribute

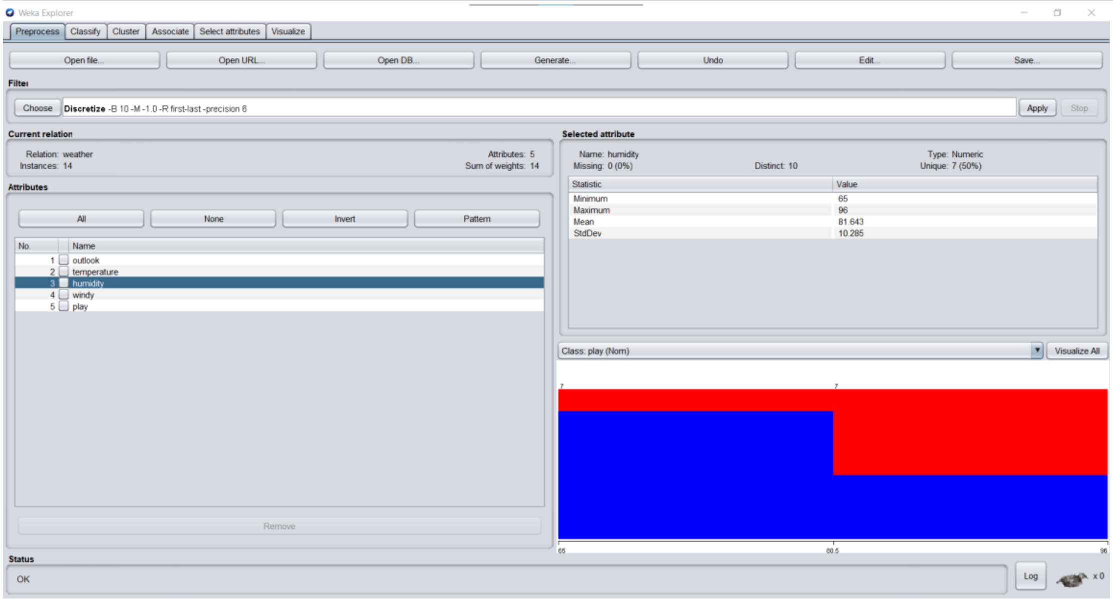
Name: temperature
Missing: 0 (0%)
Distinct: 12
Type: Numeric
Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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

Selected attribute

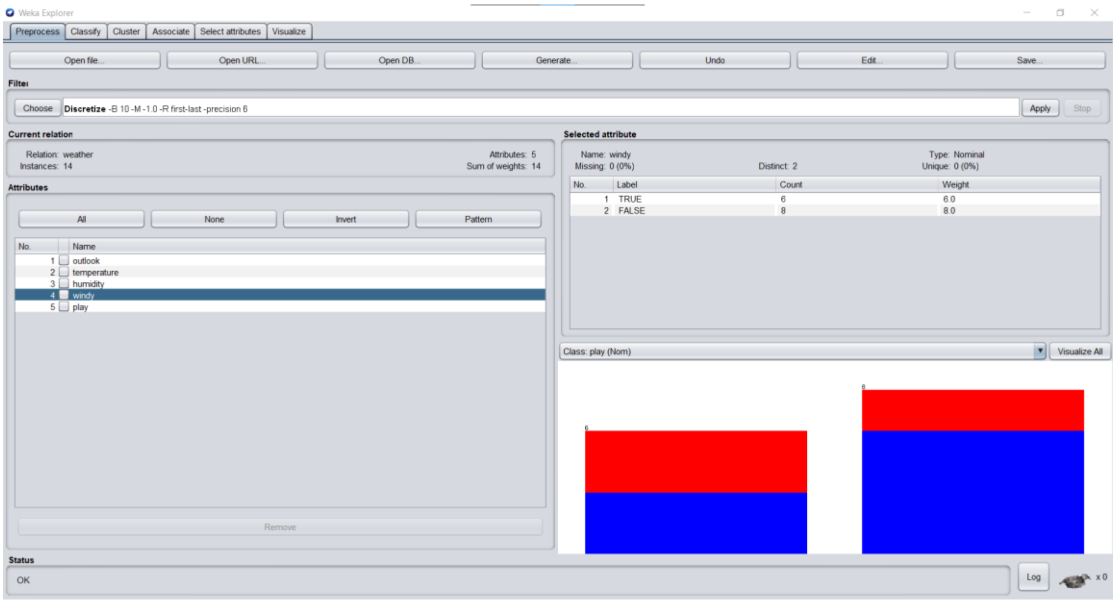
Name: humidity
Missing: 0 (0%)
Distinct: 10
Type: Numeric
Unique: 7 (50%)

Statistic	Value
Minimum	65
Maximum	96
Mean	81.643
StdDev	10.285

Class: play (Nom)

Status

OK



Filter: **Discretize** -B 10 -M -1.0 -R first-last -precision 6

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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

Selected attribute

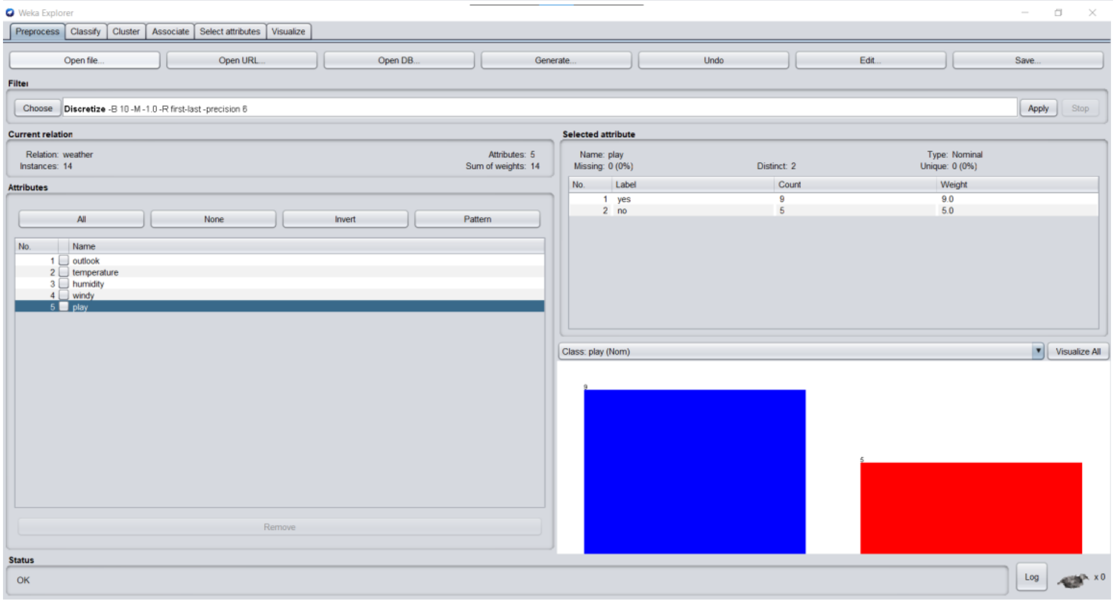
Name: windy
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	TRUE	6	6.0
2	FALSE	8	8.0

Class: play (Nom)

Status

OK



Filter: **Discretize** -B 10 -M -1.0 -R first-last -precision 6

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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

Selected attribute

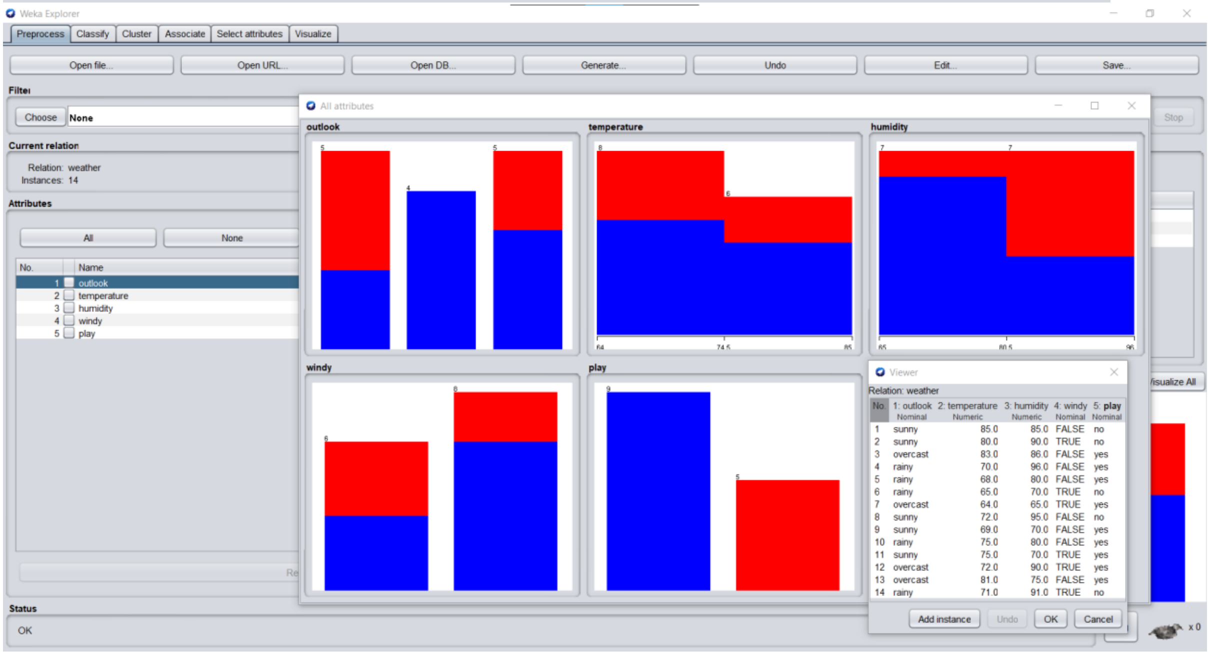
Name: play
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	9	9.0
2	no	5	5.0

Class: play (Nom)

Status

OK



Filter: **None**

Current relation

Relation: weather
Instances: 14

Attributes

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

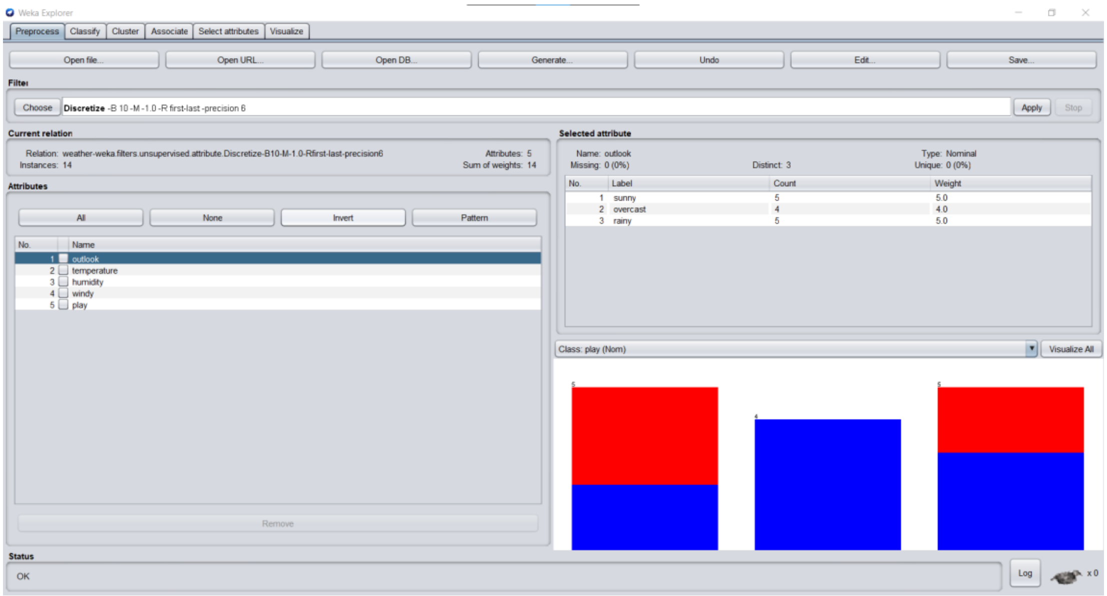
Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Status

OK

Weather Data after Discretization:



Filter

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

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6 Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

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

Selected attribute

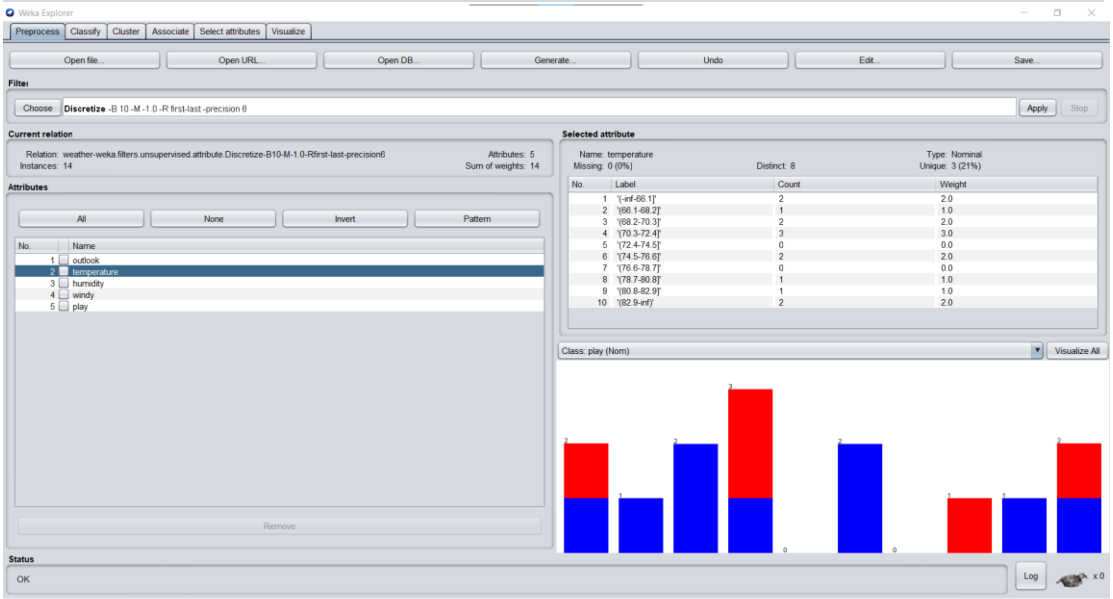
Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6 Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

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

Selected attribute

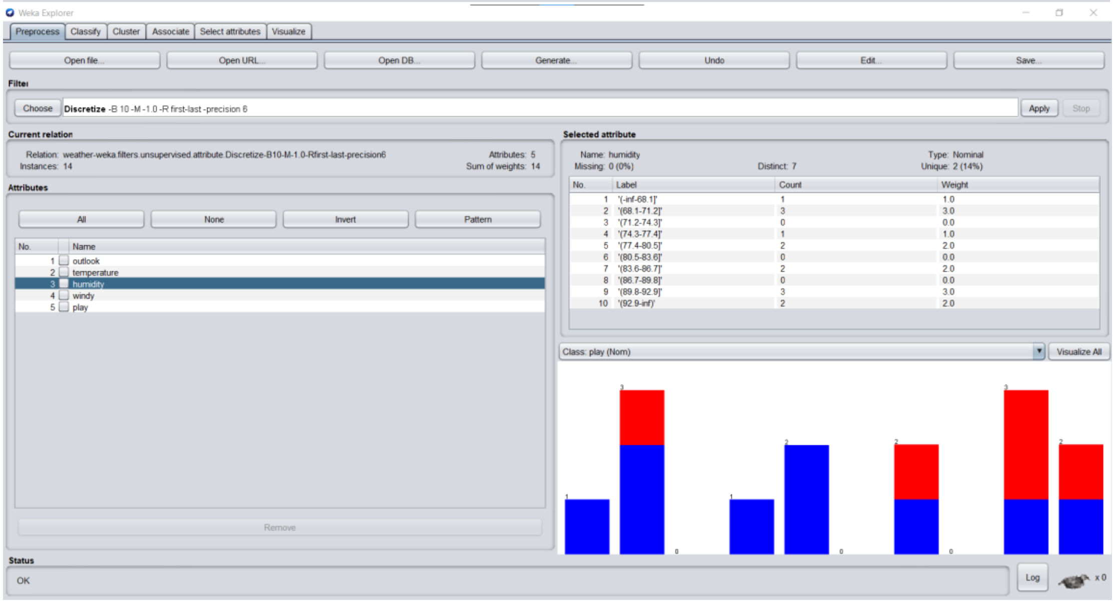
Name: temperature Type: Nominal
Missing: 0 (0%) Distinct: 8 Unique: 3 (21%)

No.	Label	Count	Weight
1	'(-inf-66.1]'	2	2.0
2	'(66.1-68.2]'	1	1.0
3	'(68.2-70.3]'	2	2.0
4	'(70.3-72.4]'	3	3.0
5	'(72.4-74.5]'	0	0.0
6	'(74.5-76.6]'	2	2.0
7	'(76.6-78.7]'	0	0.0
8	'(78.7-80.8]'	1	1.0
9	'(80.8-82.9]'	1	1.0
10	'(82.9-inf)'	2	2.0

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6 Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

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

Selected attribute

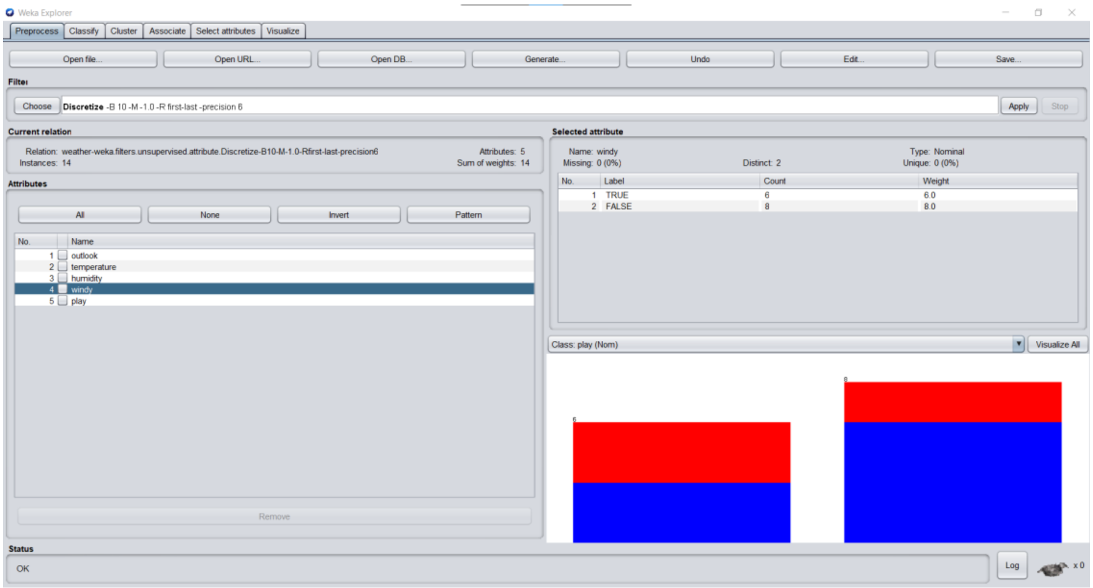
Name: humidity Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 2 (14%)

No.	Label	Count	Weight
1	'(-inf-68.1]'	1	1.0
2	'(68.1-71.2]'	3	3.0
3	'(71.2-74.3]'	0	0.0
4	'(74.3-77.4]'	1	1.0
5	'(77.4-80.5]'	2	2.0
6	'(80.5-83.6]'	0	0.0
7	'(83.6-86.7]'	2	2.0
8	'(86.7-89.8]'	0	0.0
9	'(89.8-92.9]'	3	3.0
10	'(92.9-inf)'	2	2.0

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6 Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

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

Selected attribute

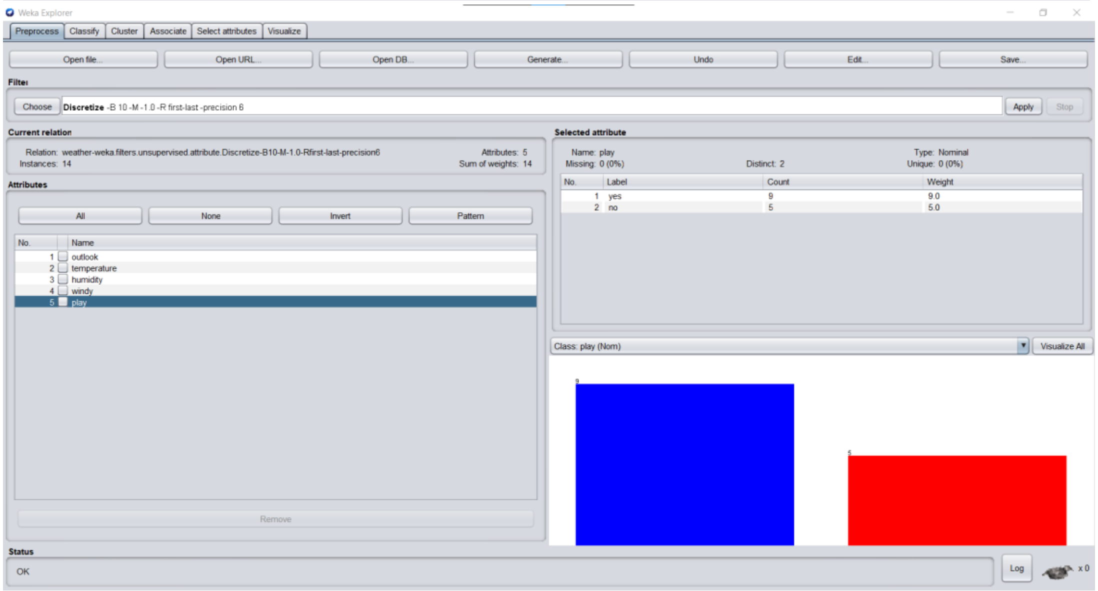
Name: windy Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	TRUE	6	6.0
2	FALSE	8	8.0

Class: play (Nom)

Status

OK



Filter

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

Current relation

Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6 Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

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

Selected attribute

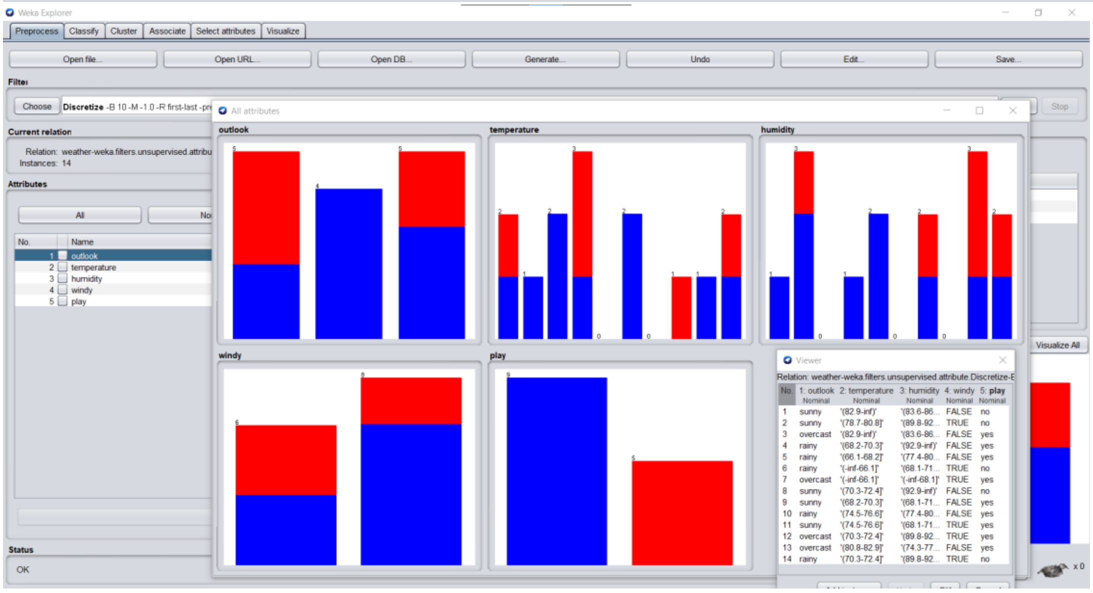
Name: play Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	9	9.0
2	no	5	5.0

Class: play (Nom)

Status

OK



Filter

Discretize -B 10 -M -1.0 -R first-last -pr

Current relation

Relation: weather-weka.filters.unsupervised.attribu
Instances: 14

Attributes

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

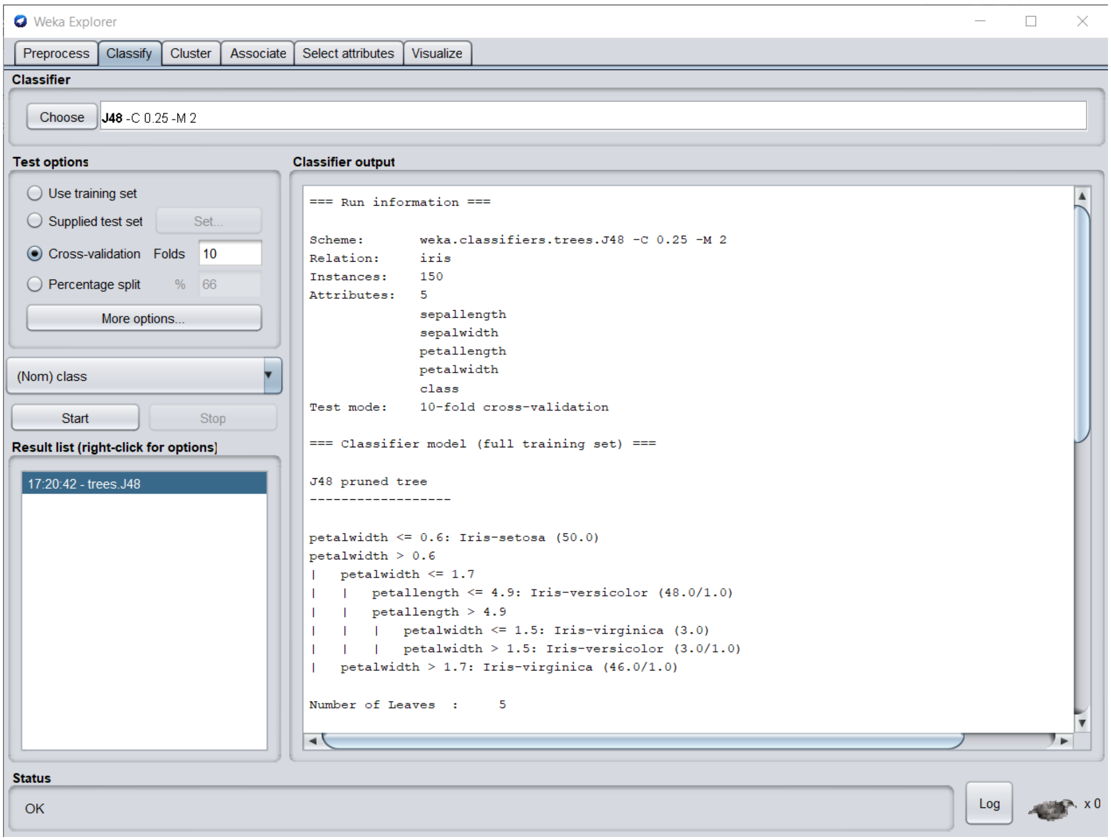
Status

OK

Relation: weather-weka.filters.unsupervised.attribute.Discretize-E

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	'(82.9-inf)'	'(83.6-86...	FALSE	no
2	sunny	'(78.7-80.8]'	'(89.8-92...	TRUE	no
3	overcast	'(82.9-inf)'	'(83.6-86...	FALSE	yes
4	rainy	'(68.2-70.3]'	'(92.9-inf)'	FALSE	yes
5	rainy	'(66.1-68.2]'	'(77.4-80...	FALSE	yes
6	rainy	'(-inf-66.1]'	'(68.1-71...	TRUE	no
7	overcast	'(-inf-66.1]'	'(-inf-68.1]'	TRUE	yes
8	sunny	'(70.3-72.4]'	'(92.9-inf)'	FALSE	no
9	sunny	'(68.2-70.3]'	'(68.1-71...	FALSE	yes
10	rainy	'(74.5-76.6]'	'(77.4-80...	FALSE	yes
11	sunny	'(74.5-76.6]'	'(68.1-71...	TRUE	yes
12	overcast	'(70.3-72.4]'	'(89.8-92...	TRUE	yes
13	overcast	'(80.8-82.9]'	'(74.3-77...	FALSE	yes
14	rainy	'(70.3-72.4]'	'(89.8-92...	TRUE	no

J48 Tree classification on iris data set before discretization:



J48 -C 0.25 -M 2

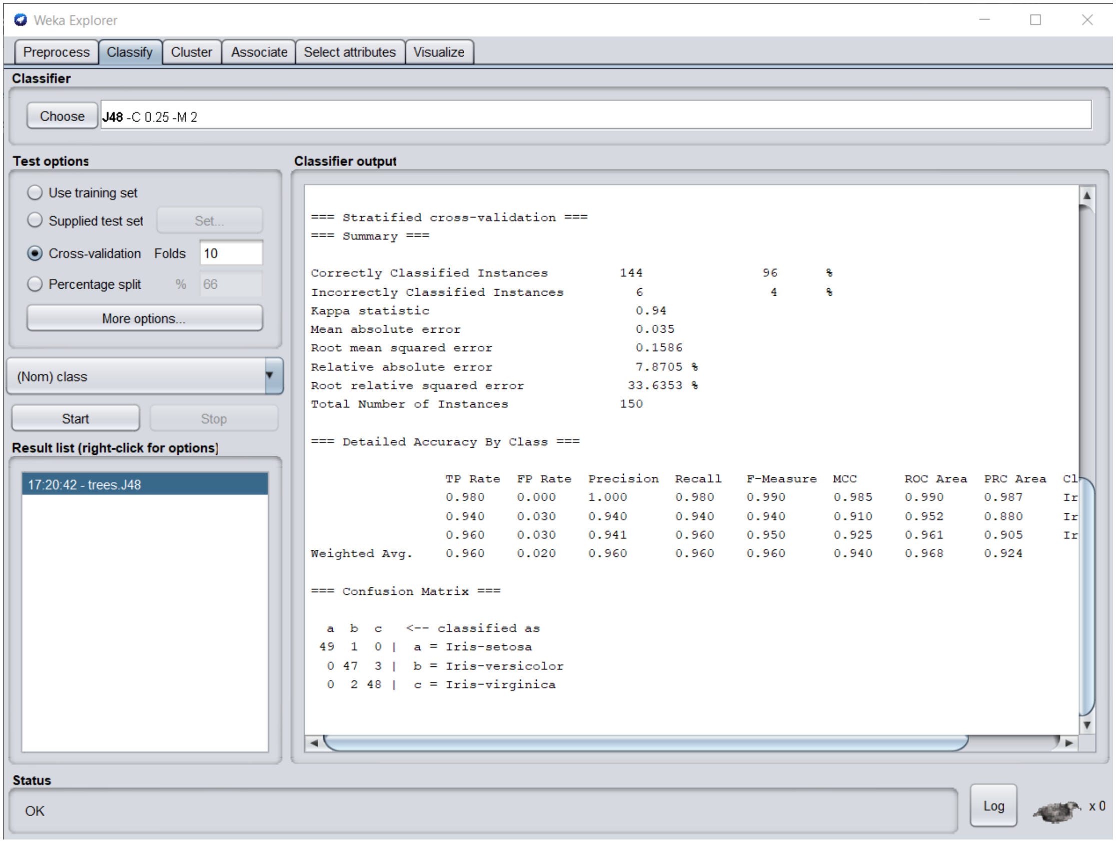
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

17:20:42 - trees.J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```

OK



J48 -C 0.25 -M 2

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

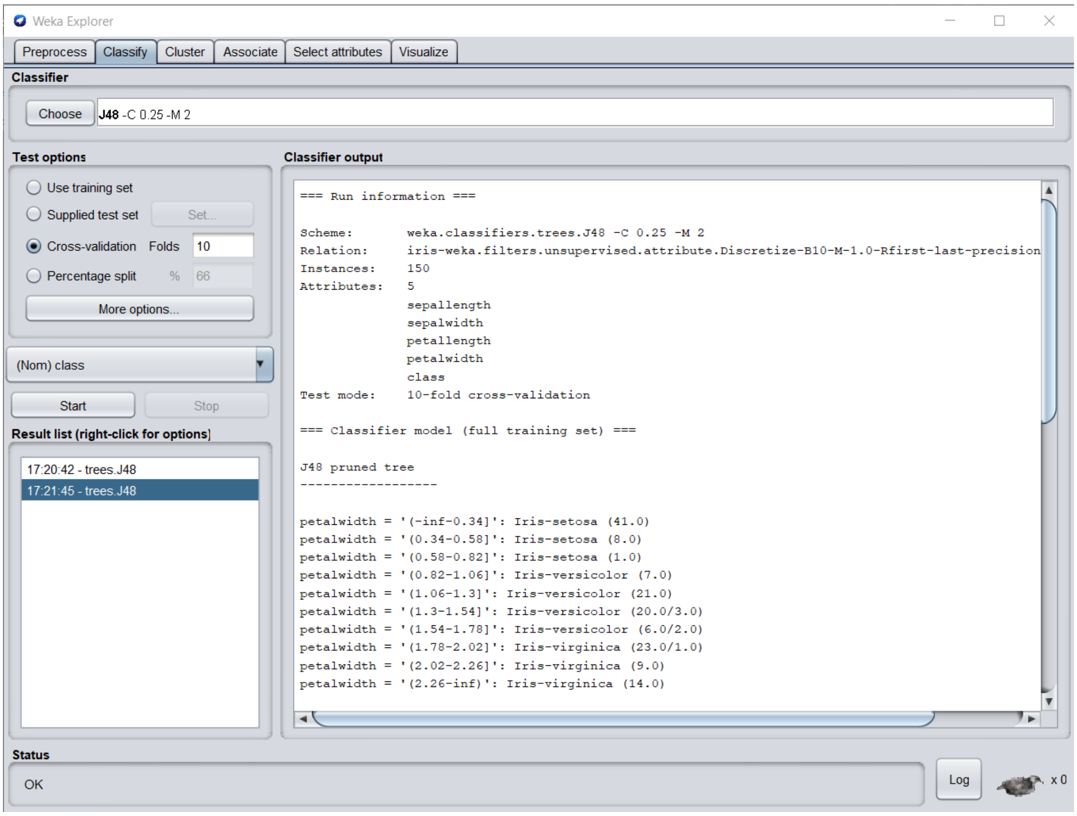
(Nom) class

17:20:42 - trees.J48

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

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.980    0.000    1.000      0.980    0.990      0.985    0.990     0.987     Ir
                 0.940    0.030    0.940      0.940    0.940      0.910    0.952     0.880     Ir
                 0.960    0.030    0.941      0.960    0.950      0.925    0.961     0.905     Ir
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.968     0.924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

OK

J48 Tree classification on iris data set after discretization:



J48 -C 0.25 -M 2

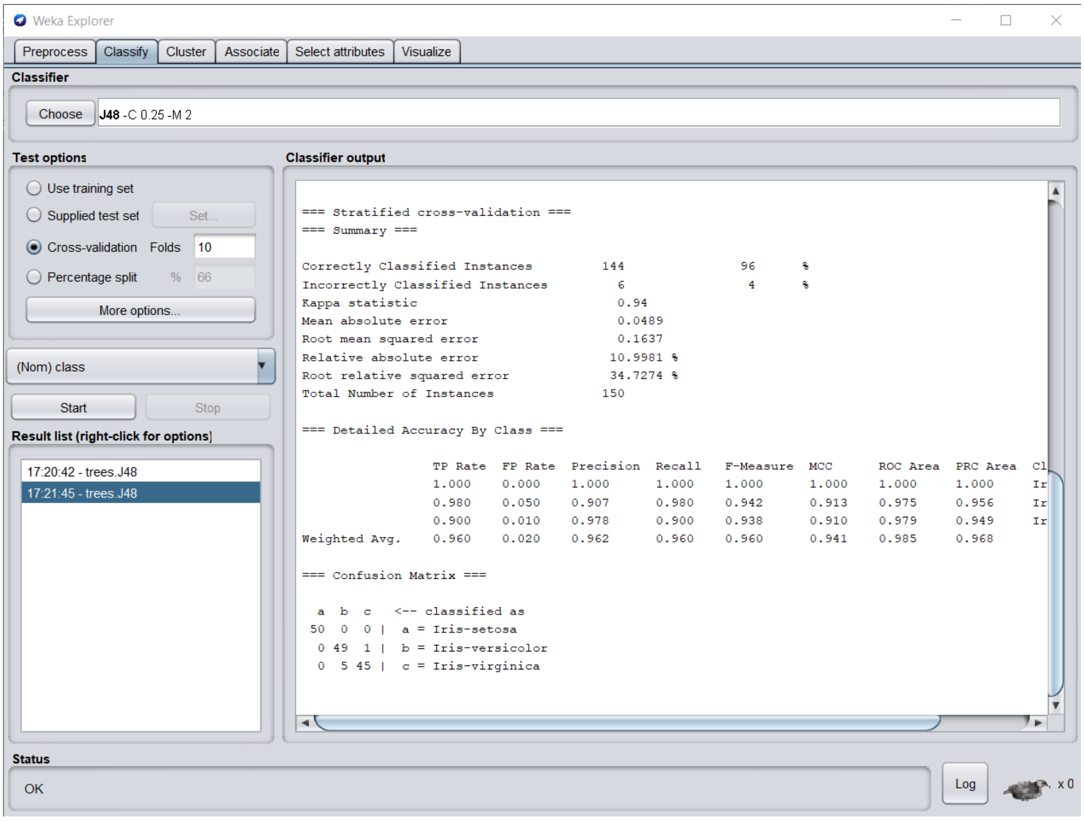
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

17:20:42 - trees.J48
17:21:45 - trees.J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth = '(-inf-0.34]': Iris-setosa (41.0)
petalwidth = '(0.34-0.58]': Iris-setosa (8.0)
petalwidth = '(0.58-0.82]': Iris-setosa (1.0)
petalwidth = '(0.82-1.06]': Iris-versicolor (7.0)
petalwidth = '(1.06-1.3]': Iris-versicolor (21.0)
petalwidth = '(1.3-1.54]': Iris-versicolor (20.0/3.0)
petalwidth = '(1.54-1.78]': Iris-versicolor (6.0/2.0)
petalwidth = '(1.78-2.02]': Iris-virginica (23.0/1.0)
petalwidth = '(2.02-2.26]': Iris-virginica (9.0)
petalwidth = '(2.26-inf)': Iris-virginica (14.0)
```

OK



J48 -C 0.25 -M 2

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

(Nom) class

17:20:42 - trees.J48
17:21:45 - trees.J48

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

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0489
Root mean squared error                  0.1637
Relative absolute error                 10.9981 %
Root relative squared error             34.7274 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Ir
                 0.980    0.050    0.907      0.980    0.942      0.913    0.975     0.956     Ir
                 0.900    0.010    0.978      0.900    0.938      0.910    0.979     0.949     Ir
Weighted Avg.    0.960    0.020    0.962      0.960    0.960      0.941    0.985     0.968

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

OK

Conclusion: We have successfully implemented Data Discretization and Visualization on weather Data Set using Weka Tool.