

## 2) Discretization of Data using Weka

19MID0020

### Data-Set Creation using Editor

```
C:\github\winter_semester-2022\CSI3010 Data Warehousing Data Mining\lab exercises\11_1_21\weather.arff - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help

weather.arff x
1 @relation weather
2 @attribute outlook {sunny,rainy,overcast}
3 @attribute temperature numeric
4 @attribute humidity numeric
5 @attribute windy {true,false}
6 @attribute play {yes,no}
7
8 @data
9 sunny,85.0,85.0,false,no
10 overcast,80.0,90.0,true,no
11 sunny,83.0,86.0,false,yes
12 rainy,70.0,86.0,false,yes
13 rainy,68.0,80.0,false,yes
14 rainy,65.0,70.0,true,no
15 overcast,64.0,65.0,false,yes
16 sunny,72.0,95.0,true,no
17 sunny,69.0,70.0,false,yes
18 rainy,75.0,80.0,false,yes
```

### Opening the created Data-Set using Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply Stop

Current relation: Relation: weather, Instances: 10, Attributes: 5, Sum of weights: 10

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Status: OK Log x 0

Selected attribute: Name: outlook, Missing: 0 (0%), Distinct: 3, Type: Nominal, 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) Visualize All

## Before Discretization (No Bins Added)

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply Stop

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

Selected attribute: Name: temperature Missing: 0 (0%) Distinct: 10 Type: Numeric Unique: 10 (100%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.1
StdDev	7.4

Class: play (Nom) Visualize All

Status: OK Log x 0

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options:
 

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

 More options...

(Nom) play Start Stop

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

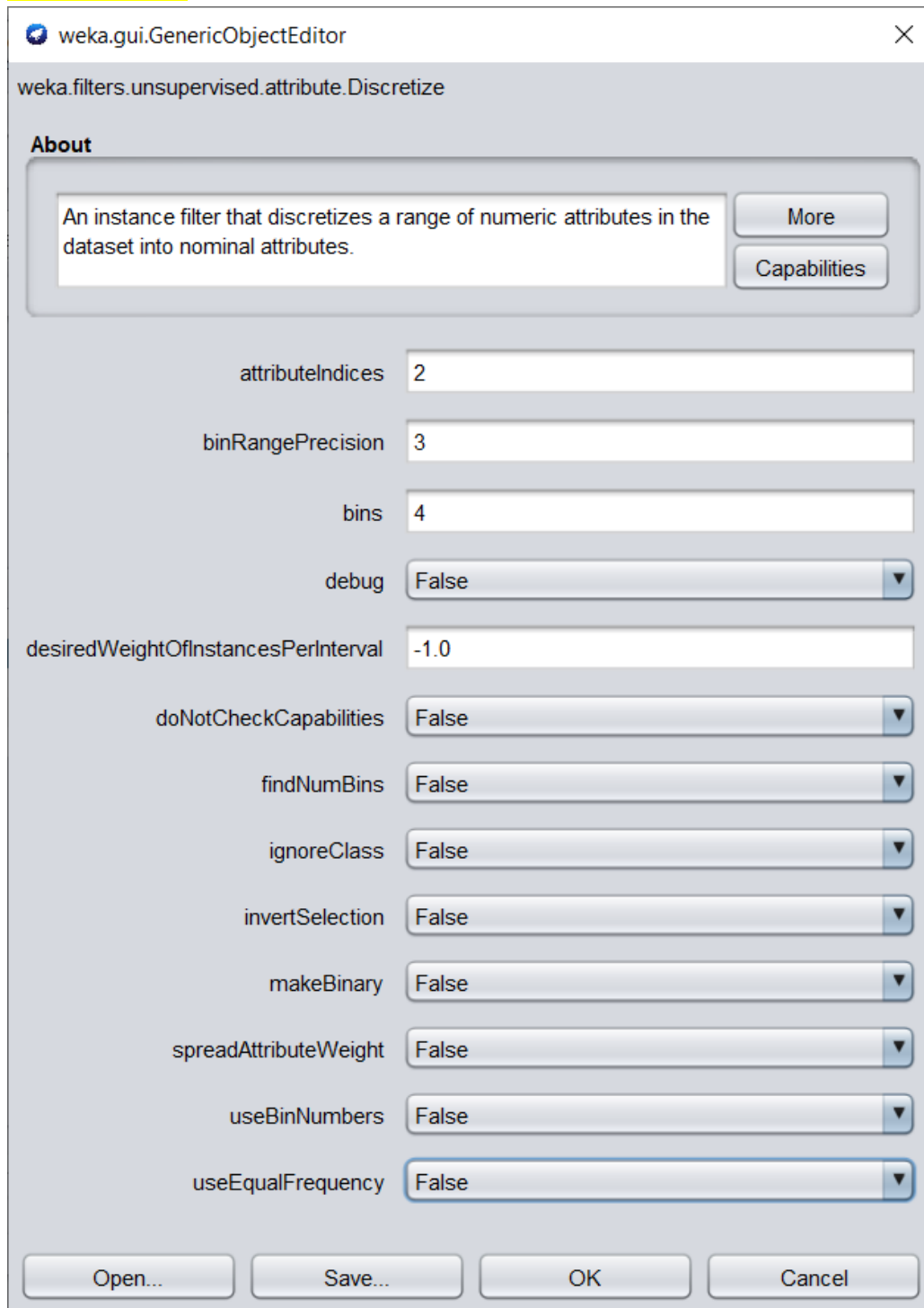
=== Detailed Accuracy By Class ===
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Cl
      0.000    0.500    0.000     0.000    0.000   -0.500  0.250    0.333    ye
      0.500    1.000    0.500     0.500    0.500   -0.500  0.250    0.583    no
Weighted Avg.   0.333    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 Log x 0

## After Discretization (4 Bins Added)

No. of. Bins: 4



The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.filters.unsupervised.attribute.Discretize' filter. The 'About' section describes it as an instance filter that discretizes a range of numeric attributes into nominal attributes. The settings are as follows:

Property	Value
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

Buttons at the bottom: Open..., Save..., OK, Cancel.

# ISCRETIZATION OF DATA USING WEKA

19MID0020

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose Discretize -B 4 -M -1.0 -R 2 -precision 3 Apply Stop

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

Attributes  
All None Invert Pattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input checked="" type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Selected attribute  
Name: temperature  
Missing: 0 (0%) Distinct: 4 Type: Nominal  
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

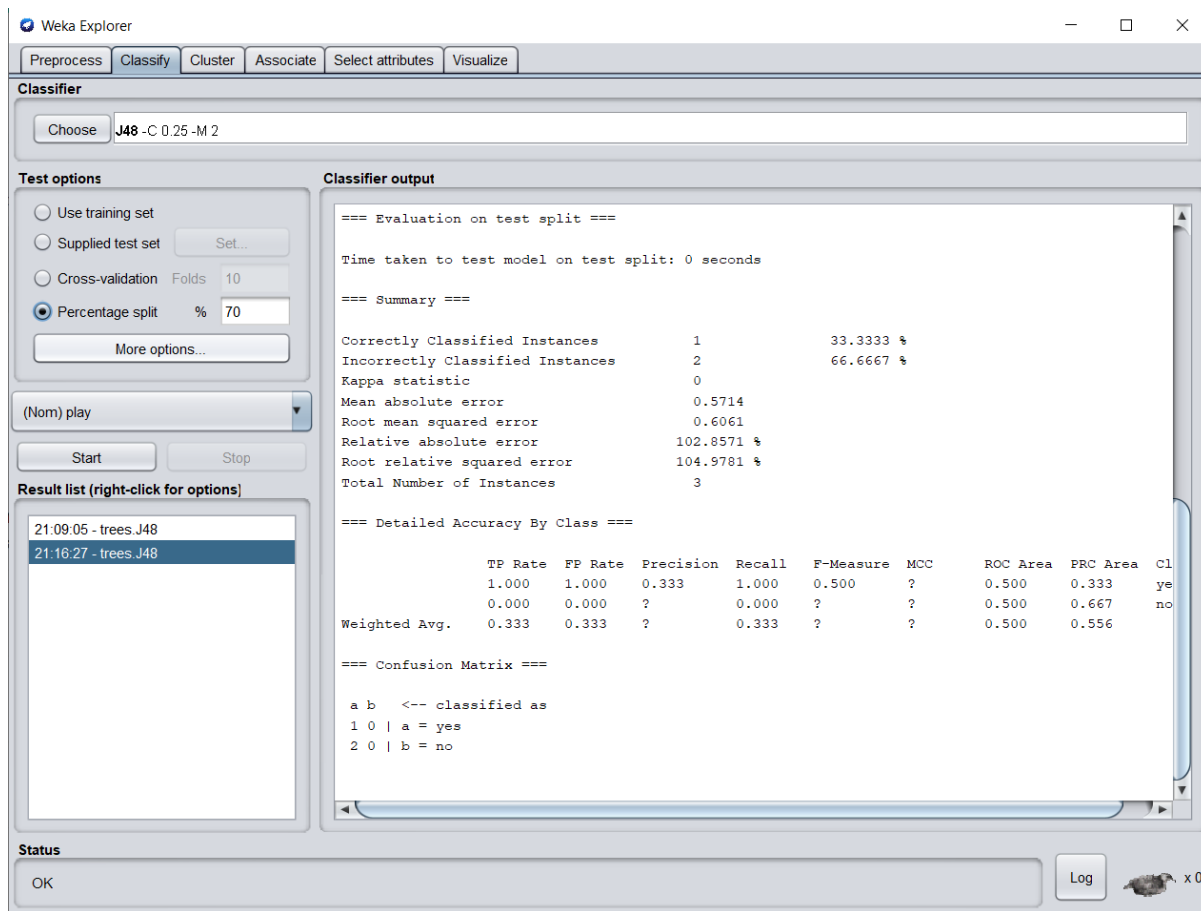
Class: play (Nom) Visualize All

Temperature Bin	Blue Count	Red Count	Total Count
'(-inf-69.25]'	4	0	4
'(69.25-74.5]'	2	0	2
'(74.5-79.75]'	1	0	1
'(79.75-inf]'	3	0	3

Status  
OK Log x 0

# ISCRETIZATION OF DATA USING WEKA

19MID0020

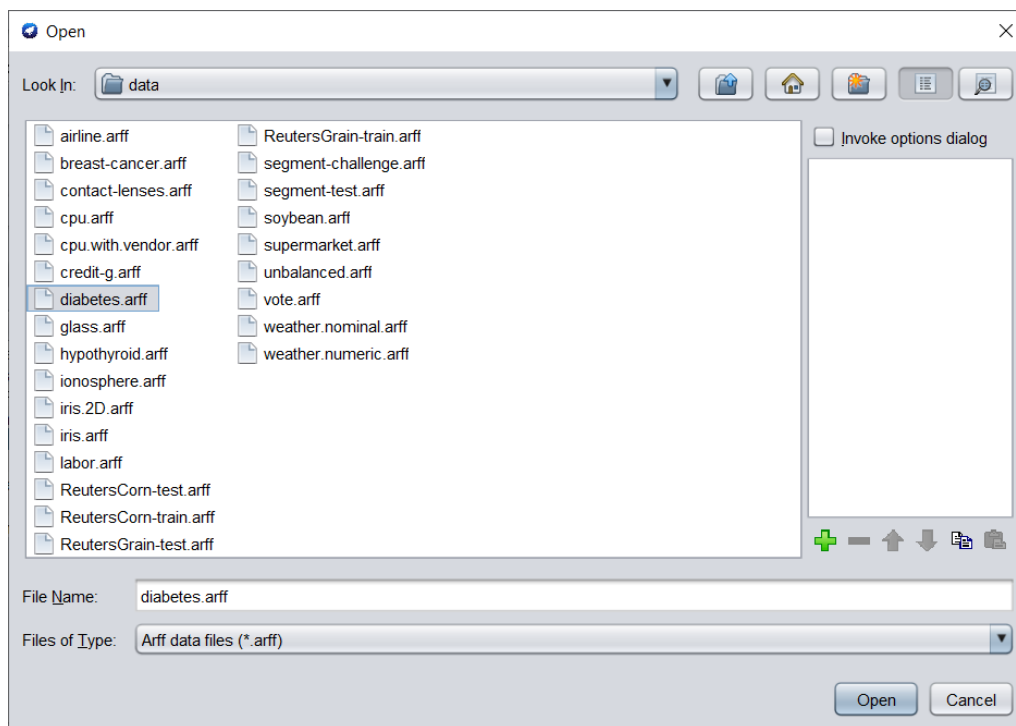


Accuracy: 33.333%

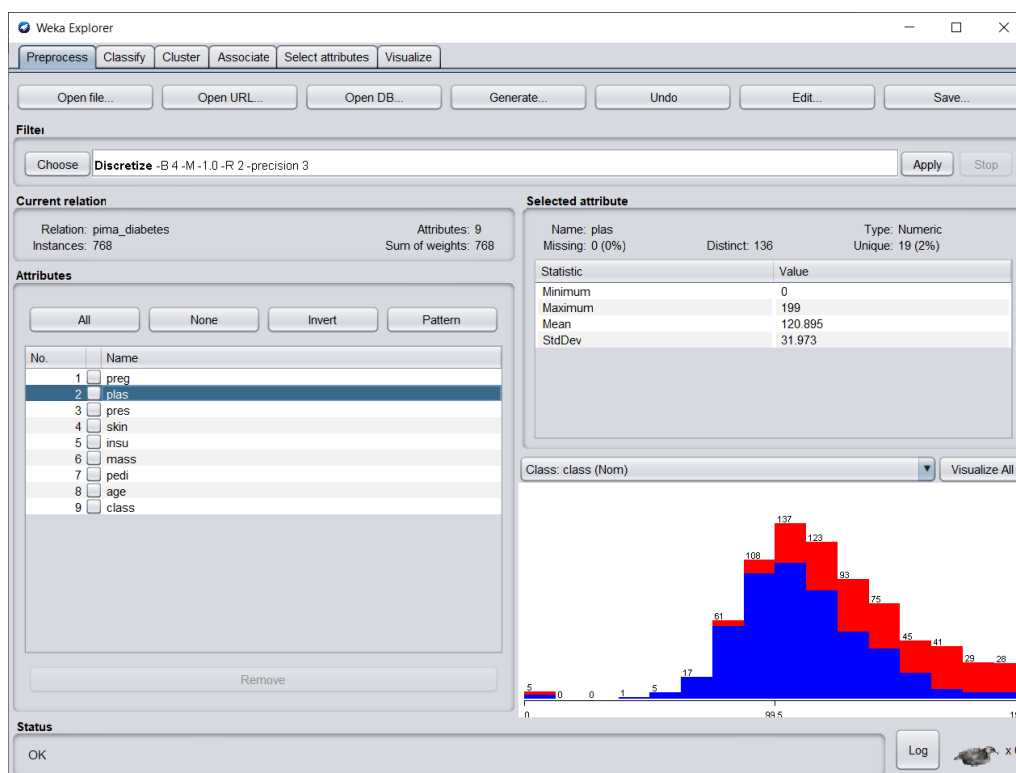
Accuracy varies when we change the split percentages.

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

## Prima Diabetes Data-Set



## Before Discretization (No bins added)



# ISCRETIZATION OF DATA USING WEKA

19MID0020

**Weka Explorer**

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

**Filter:** Choose **Discretize -B 4 -M 1.0 -R 2 -precision 3** Apply Stop

**Current relation**  
Relation: pima\_diabetes  
Instances: 768  
Attributes: 9  
Sum of weights: 768

**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

**Attributes**  
All None Invert Pattern

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

Remove

**Class: class (Nom)** Visualize All

**Status**  
OK Log x 0

**Weka Explorer**

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

**Classifier**  
Choose **J48 -C 0.25 -M 2**

**Test options**  
☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 70  
More options...

(Nom) class

Start Stop

**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  MOC     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 Log x 0

Accuracy: 76.52%

## After Before 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.

More

Capabilities

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

Open... Save... OK Cancel



# ISCRETIZATION OF DATA USING WEKA

19MID0020

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

**Filter**

Choose **Discretize -B 5 -M -1.0 -R 2,5 -precision 3** Apply Stop

**Current relation**

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

**Attributes**

All None Invert Pattern

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

Remove

**Selected attribute**

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) Visualize All

Status: OK Log x 0

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

**Filter**

Choose **Discretize -B 5 -M -1.0 -R 2,5 -precision 3** Apply Stop

**Current relation**

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

**Attributes**

All None Invert Pattern

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

Remove

**Selected attribute**

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) Visualize All

Status: OK Log x 0

# ISCRETIZATION OF DATA USING WEKA

19MID0020

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is selected, and 'J48 -C 0.25 -M 2' is chosen. Under 'Test options', 'Percentage split' is set to 70%. The 'Classifier output' pane displays the following results:

```
=== 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
```

The 'Result list' on the left shows two entries: '21:26:29 - trees.J48' and '21:28:22 - trees.J48', with the latter selected.

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%