VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
Vellore-632 014, Tamil Nadu, India.
www.vit.ac.in

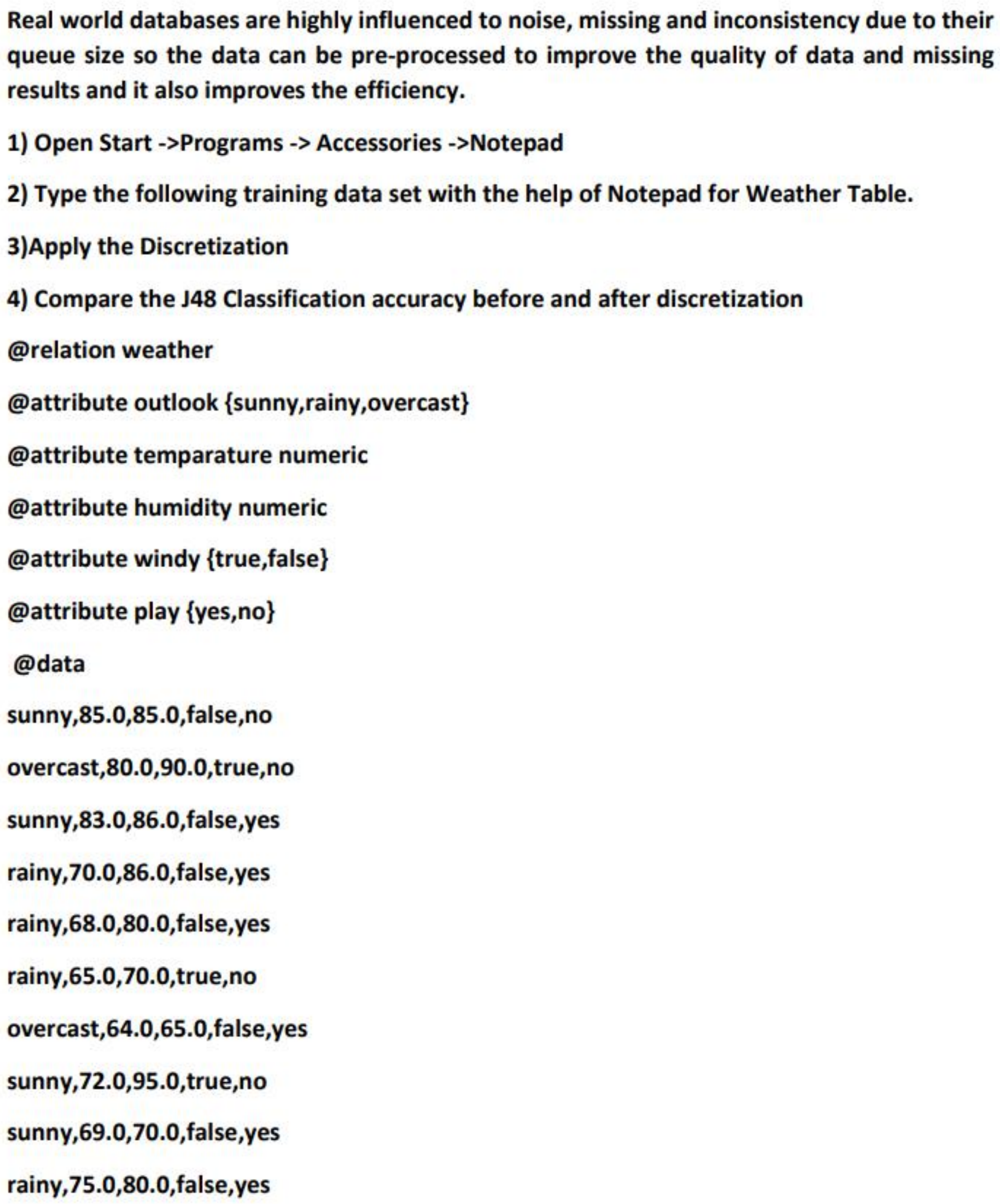
S.THARUN

19MID0031

CSI3010 – DATA WAREHOUSING AND DATA MINING

FACULTY : CHELLATAMILAN T

PRE-PROCESSING TECHNIQUES USING WEKA- DISCRETIZATION



Real world databases are highly influenced to noise, missing and inconsistency due to their queue size so the data can be pre-processed to improve the quality of data and missing results and it also improves the efficiency.

1) Open Start ->Programs -> Accessories ->Notepad

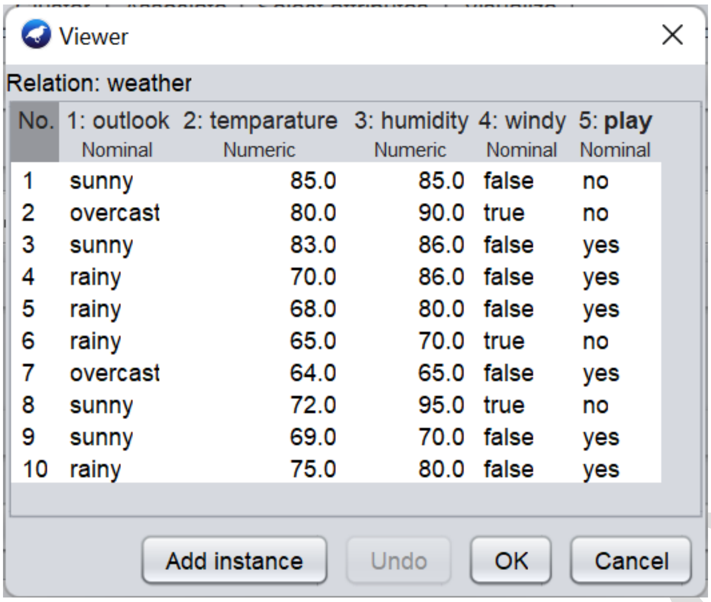
2) Type the following training data set with the help of Notepad for Weather Table.

3)Apply the Discretization

4) Compare the J48 Classification accuracy before and after discretization

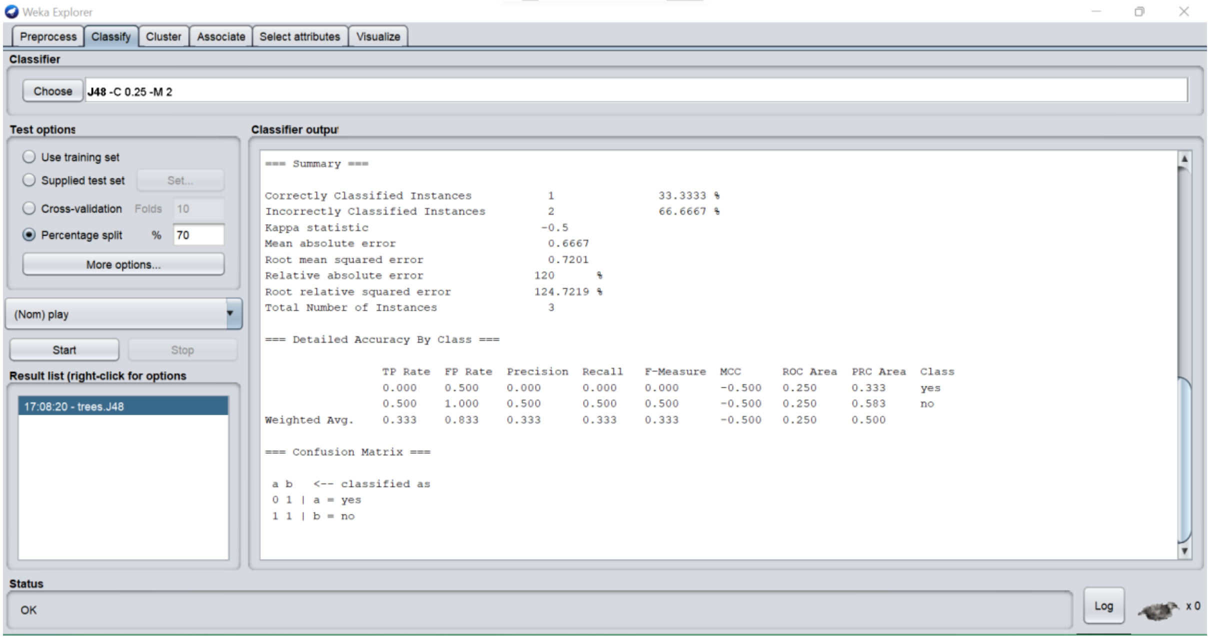
```
@relation weather

@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
```

BEFORE DISCRETIZATION



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



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

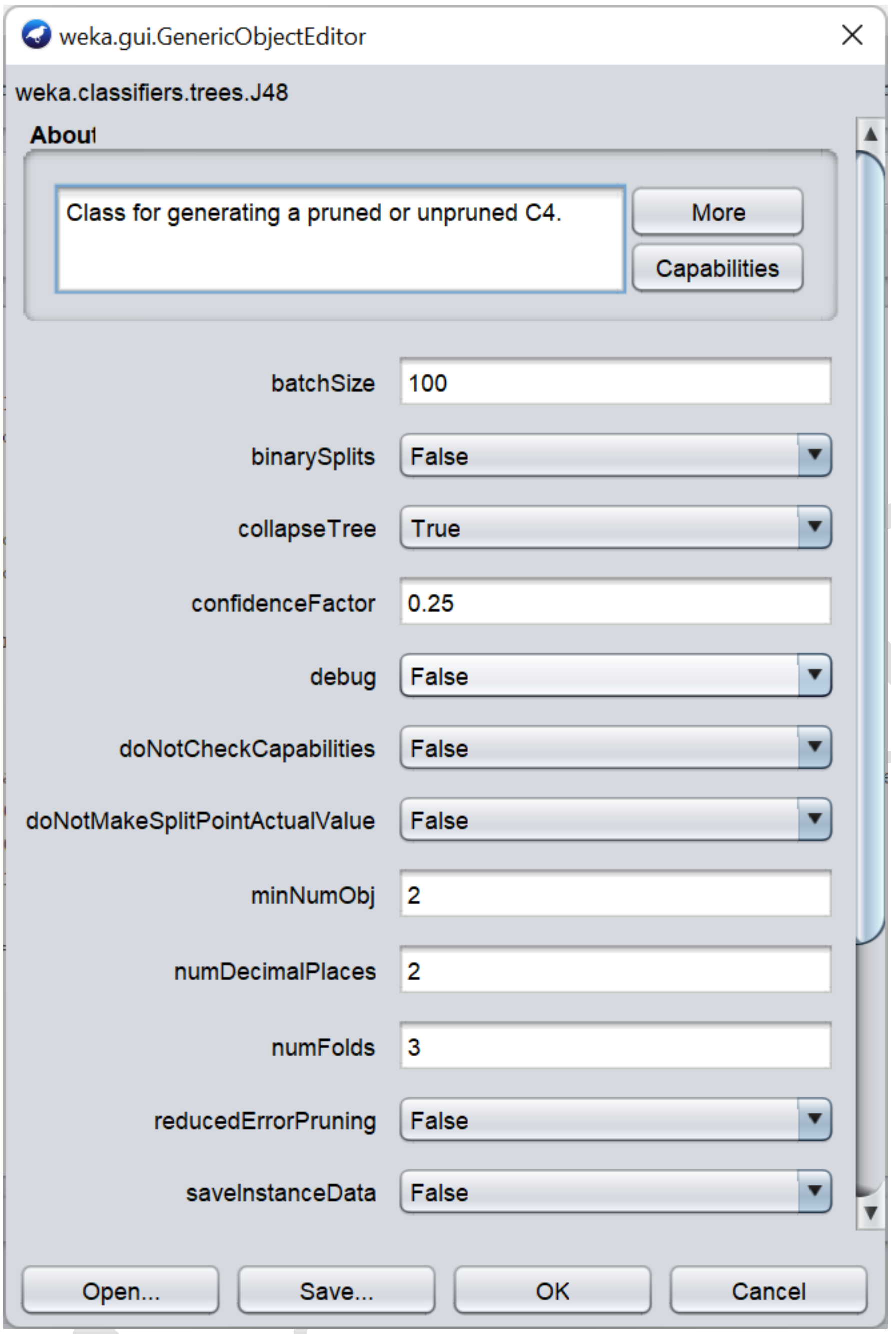
17:08:20 - trees.J48

Classifier outpu

```
=== 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  Class
                 0.000    0.500    0.000      0.000    0.000      -0.500   0.250     0.333     yes
                 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



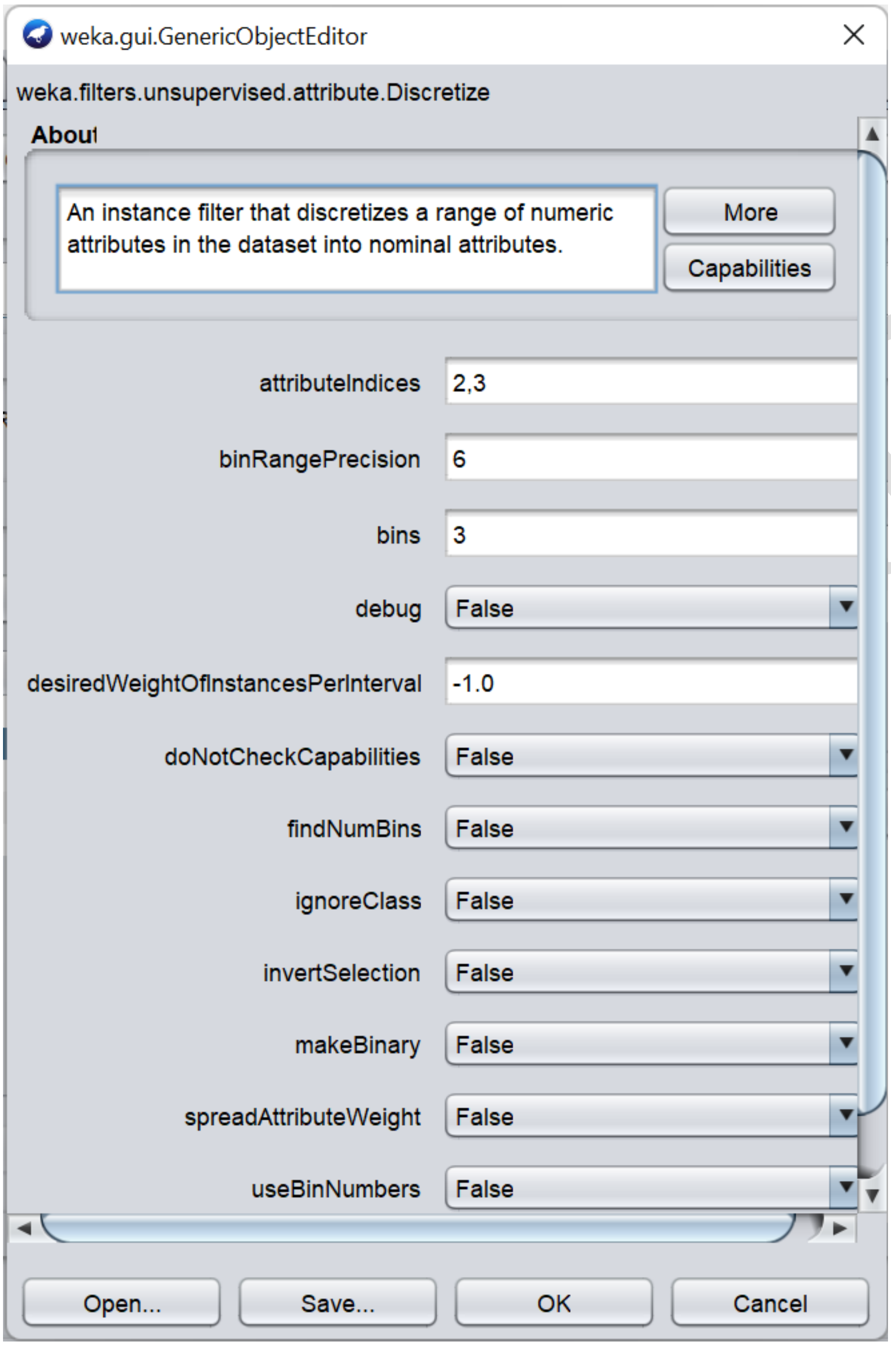
weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

batchSize	100
binarySplits	False
collapseTree	True
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	2
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False

AFTER DISCRETIZATION

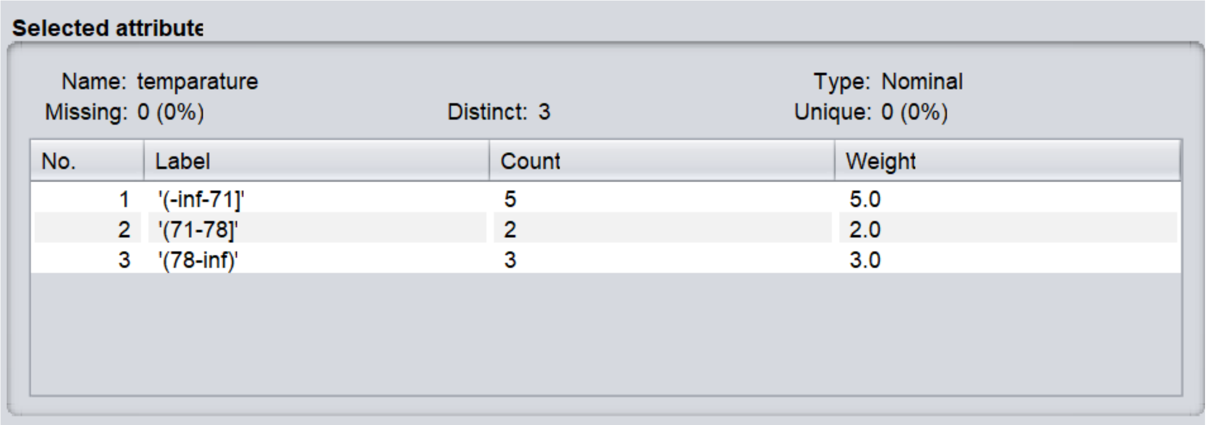


weka.filters.unsupervised.attribute.Discretize

About

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

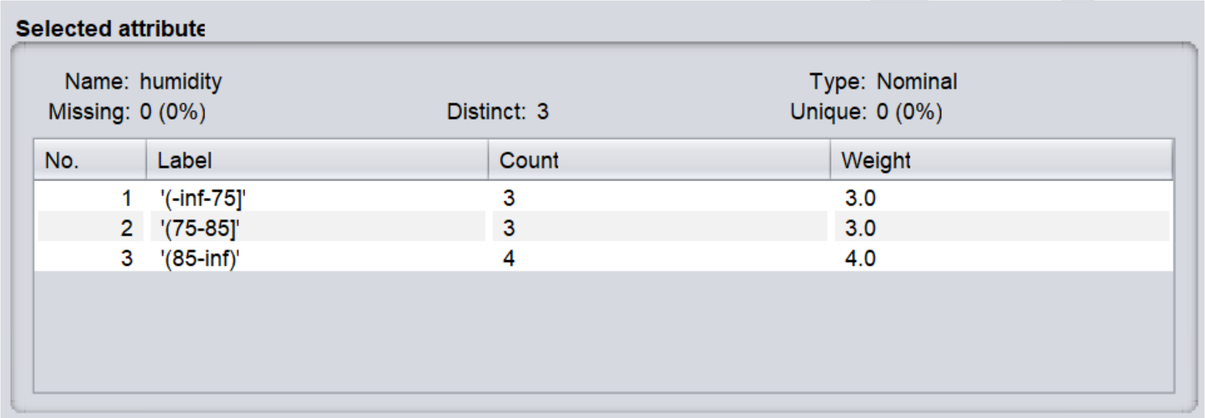
attributeIndices: 2,3
binRangePrecision: 6
bins: 3
debug: False
desiredWeightOfInstancesPerInterval: -1.0
doNotCheckCapabilities: False
findNumBins: False
ignoreClass: False
invertSelection: False
makeBinary: False
spreadAttributeWeight: False
useBinNumbers: False



Selected attribute

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

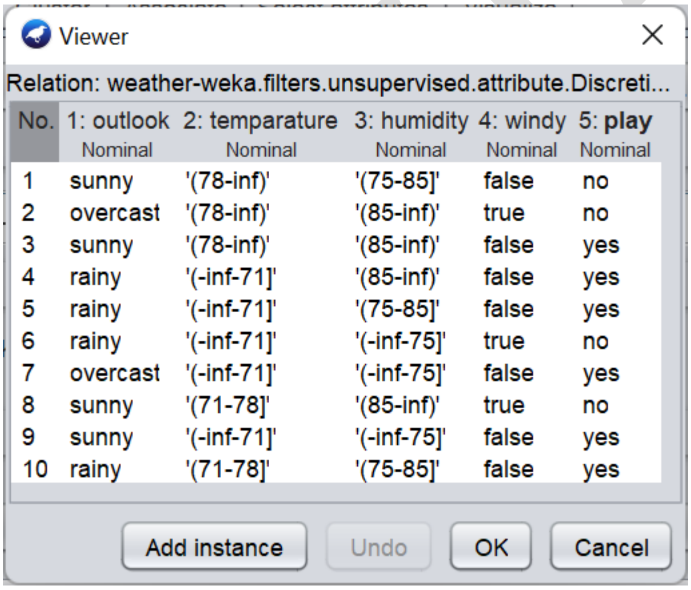
No.	Label	Count	Weight
1	'(-inf-71]'	5	5.0
2	'(71-78]'	2	2.0
3	'(78-inf)'	3	3.0



Selected attribute

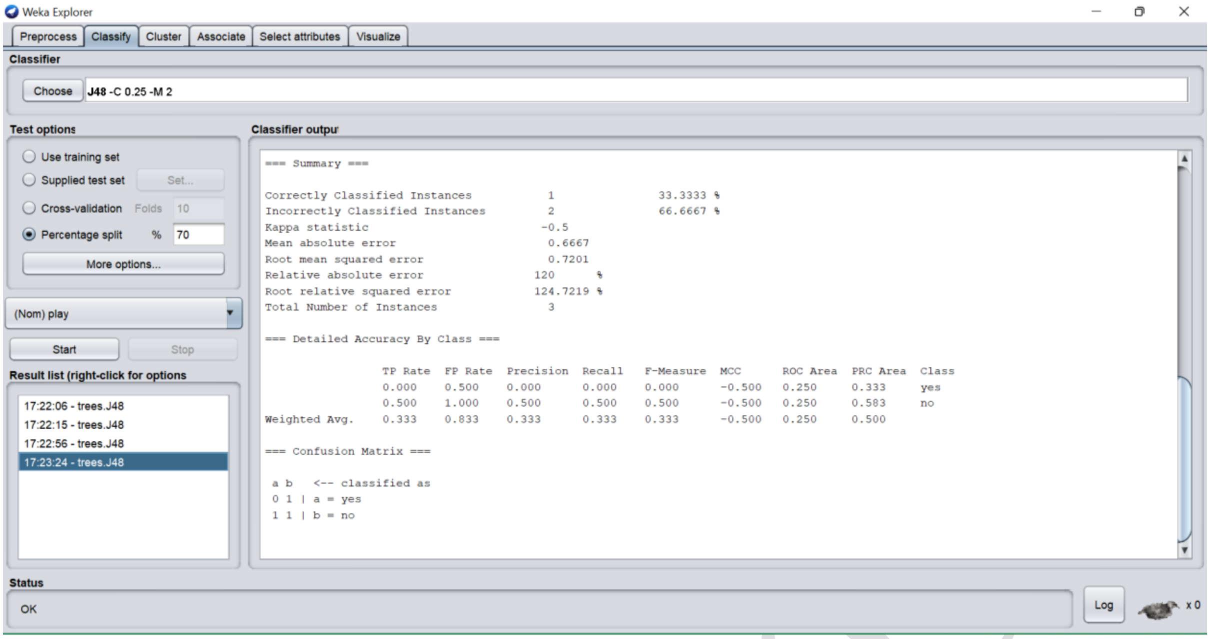
Name: humidity Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-75]'	3	3.0
2	'(75-85]'	3	3.0
3	'(85-inf)'	4	4.0



Relation: weather-weka.filters.unsupervised.attribute.Discreti...

No.	1: outlook Nominal	2: temparature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
1	sunny	'(78-inf)'	'(75-85]'	false	no
2	overcast	'(78-inf)'	'(85-inf)'	true	no
3	sunny	'(78-inf)'	'(85-inf)'	false	yes
4	rainy	'(-inf-71]'	'(85-inf)'	false	yes
5	rainy	'(-inf-71]'	'(75-85]'	false	yes
6	rainy	'(-inf-71]'	'(-inf-75]'	true	no
7	overcast	'(-inf-71]'	'(-inf-75]'	false	yes
8	sunny	'(71-78]'	'(85-inf)'	true	no
9	sunny	'(-inf-71]'	'(-inf-75]'	false	yes
10	rainy	'(71-78]'	'(75-85]'	false	yes



J48 -C 0.25 -M 2

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

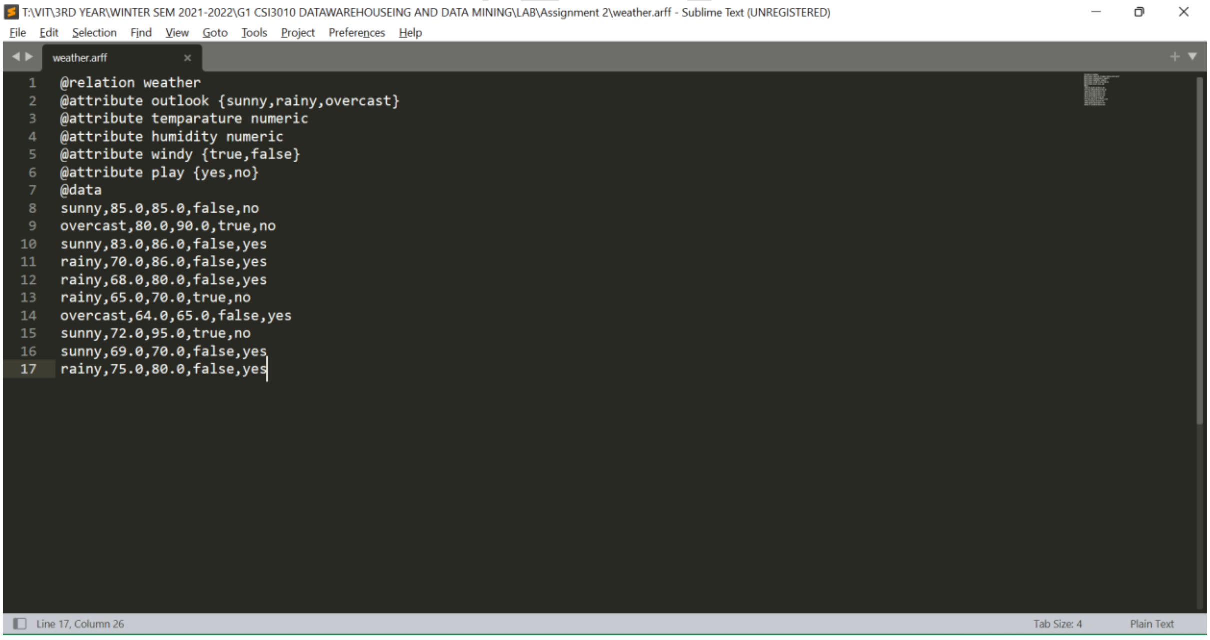
(Nom) play

17:22:06 - trees.J48
17:22:15 - trees.J48
17:22:56 - trees.J48
17:23:24 - trees.J48

```
=== 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  Class
                 0.000    0.500    0.000      0.000    0.000      -0.500   0.250     0.333     yes
                 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
```

OK

MANUALLY CREATED DATASET FILE :



T:\VIT\3RD YEAR\WINTER SEM 2021-2022\G1 CSI3010 DATAWAREHOUSEING AND DATA MINING\LAB\Assignment 2\weather.arff - Sublime Text (UNREGISTERED)

```
@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
```

Line 17, Column 26