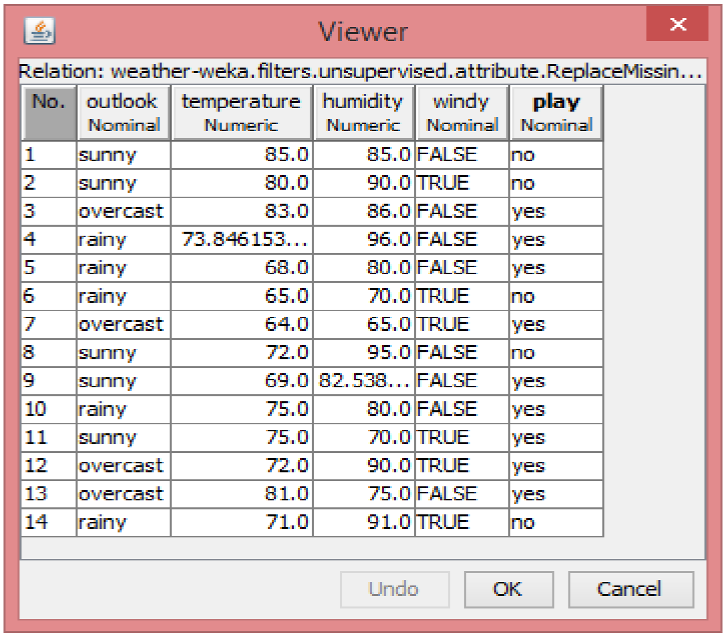
Lab Exercise One
Data Preprocessing with WEKA Explorer

Binning With Filter

Unsupervised Attribute Filter – Discretize: This filter converts numeric attributes to nominal use equal-width (default) or equal-depth (frequency) binning.

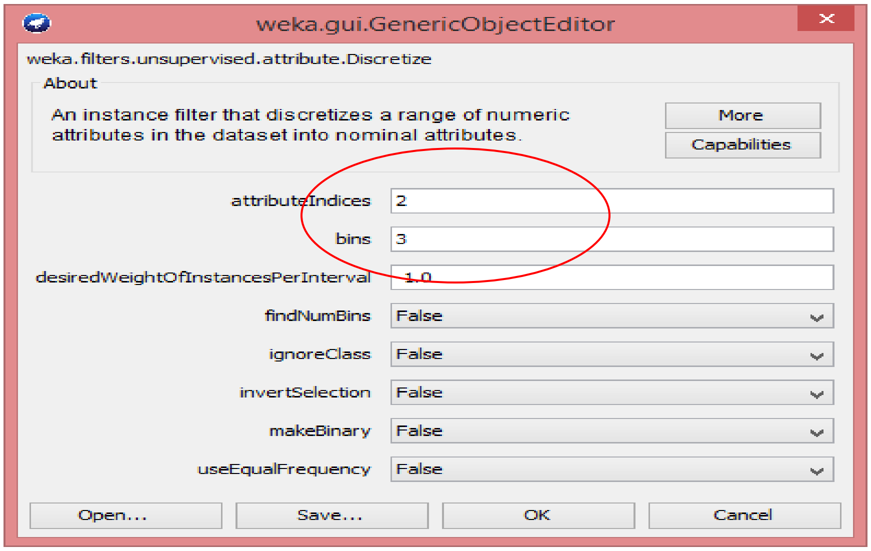
1. Open the dataset **weather.numeric**. First, replace missing values with the filter.



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

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	**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	73.846153...	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	82.538...	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

2. Choose filter **Discretize** from the drop-down list of **unsupervised attribute** filters and then left-click to open its properties window. We want to perform **equal-width** binning on 2nd attribute – **temperature** with three bins.



weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndices: 2

bins: 3

desiredWeightOfInstancesPerInterval: -1.0

findNumBins: False

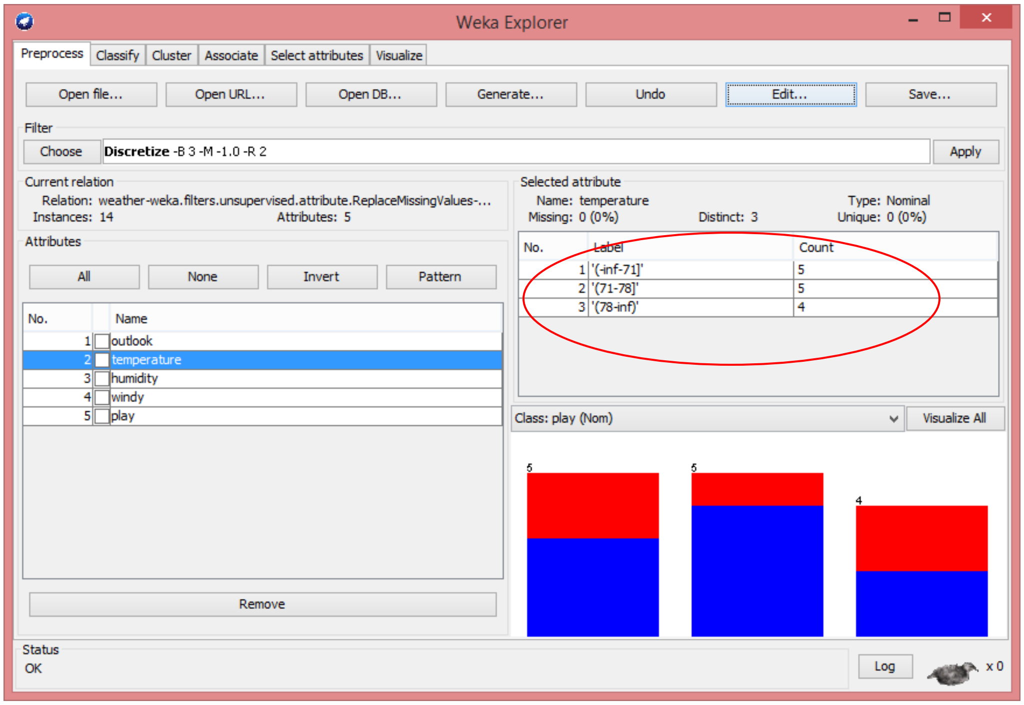
ignoreClass: False

invertSelection: False

makeBinary: False

useEqualFrequency: False

3. Click **Apply** button. Then select **temperature** attribute to check the results.



Filter

Discretize -B 3 -M -1.0 -R 2

Current relation

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues-...
Instances: 14 Attributes: 5

Selected attribute

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

No.	Label	Count
1	'(-inf-71]'	5
2	'(71-78]'	5
3	'(78-inf)'	4

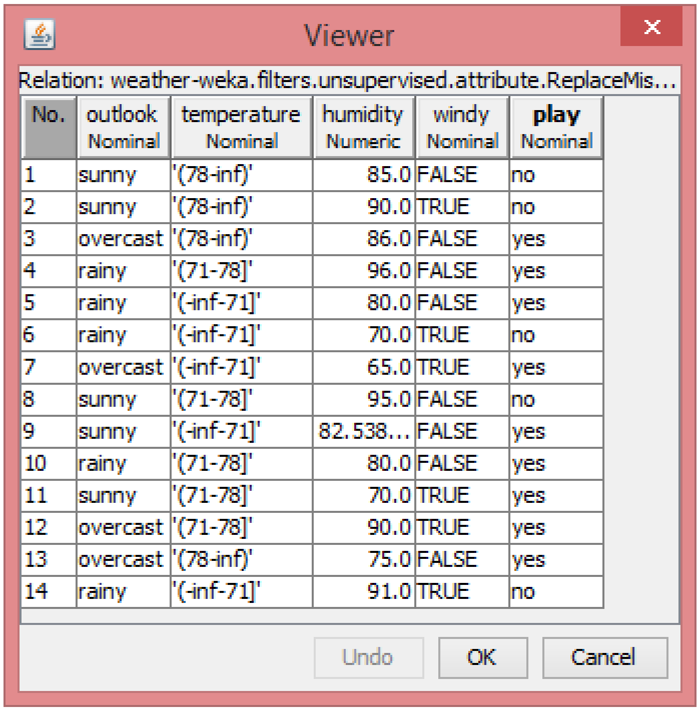
Attributes

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

Class: play (Nom)

Status

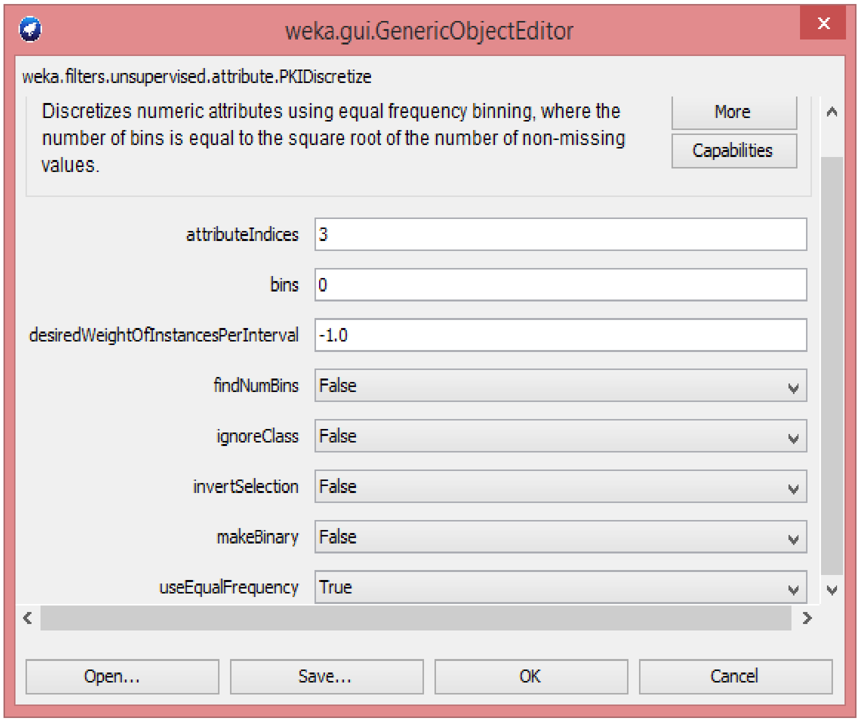
OK



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

No.	outlook Nominal	temperature Nominal	humidity Numeric	windy Nominal	**play** Nominal
1	sunny	'(78-inf)'	85.0	FALSE	no
2	sunny	'(78-inf)'	90.0	TRUE	no
3	overcast	'(78-inf)'	86.0	FALSE	yes
4	rainy	'(71-78]'	96.0	FALSE	yes
5	rainy	'(-inf-71]'	80.0	FALSE	yes
6	rainy	'(-inf-71]'	70.0	TRUE	no
7	overcast	'(-inf-71]'	65.0	TRUE	yes
8	sunny	'(71-78]'	95.0	FALSE	no
9	sunny	'(-inf-71]'	82.538...	FALSE	yes
10	rainy	'(71-78]'	80.0	FALSE	yes
11	sunny	'(71-78]'	70.0	TRUE	yes
12	overcast	'(71-78]'	90.0	TRUE	yes
13	overcast	'(78-inf)'	75.0	FALSE	yes
14	rainy	'(-inf-71]'	91.0	TRUE	no

4. To perform equal-depth (frequency) binning on the 3rd attribute **humidity**, we choose filter **PKIDiscretize** from the drop-down list of **unsupervised attribute** filters and then left-click to open its properties window. This filter use the square root of the number of values as the number of bins.

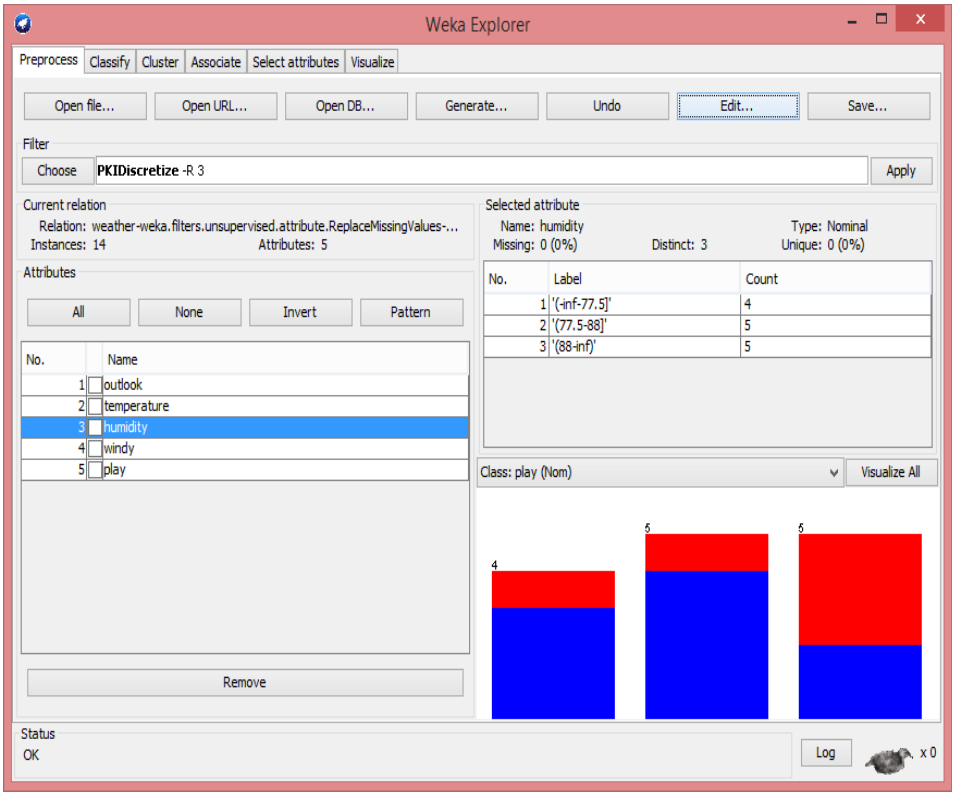


weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.PKIDiscretize

Discretizes numeric attributes using equal frequency binning, where the number of bins is equal to the square root of the number of non-missing values.

attributeIndices	3
bins	0
desiredWeightOfInstancesPerInterval	-1.0
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
useEqualFrequency	True



Weka Explorer

Filter: PKIDiscretize -R 3

Current relation
Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues-...
Instances: 14   Attributes: 5

Attributes

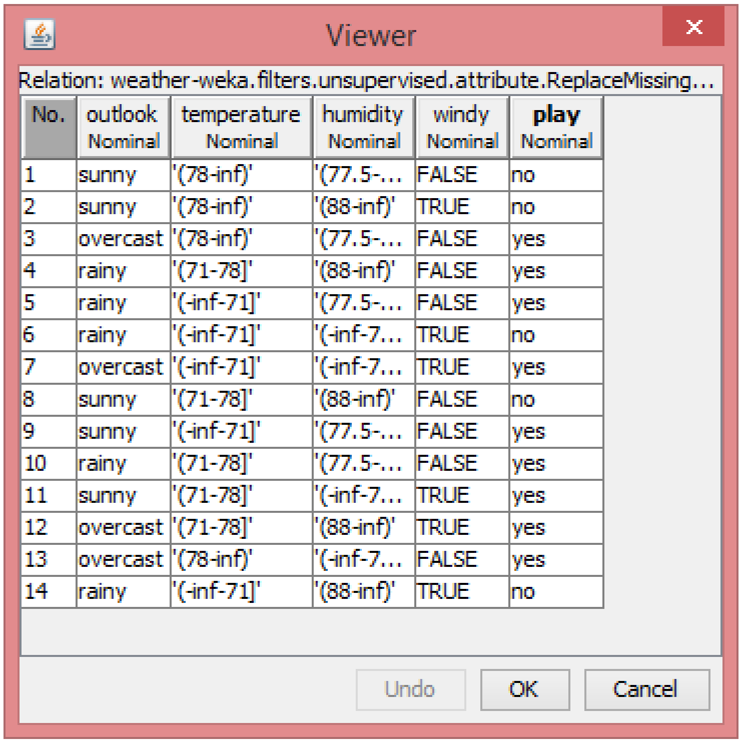
No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

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

No.	Label	Count
1	'(-inf-77.5]'	4
2	'(77.5-88]'	5
3	'(88-inf)'	5

Class: play (Nom)

Status
OK



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

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	'(78-inf)'	'(77.5-...	FALSE	no
2	sunny	'(78-inf)'	'(88-inf)'	TRUE	no
3	overcast	'(78-inf)'	'(77.5-...	FALSE	yes
4	rainy	'(71-78]'	'(88-inf)'	FALSE	yes
5	rainy	'(-inf-71]'	'(77.5-...	FALSE	yes
6	rainy	'(-inf-71]'	'(-inf-7...	TRUE	no
7	overcast	'(-inf-71]'	'(-inf-7...	TRUE	yes
8	sunny	'(71-78]'	'(88-inf)'	FALSE	no
9	sunny	'(-inf-71]'	'(77.5-...	FALSE	yes
10	rainy	'(71-78]'	'(77.5-...	FALSE	yes
11	sunny	'(71-78]'	'(-inf-7...	TRUE	yes
12	overcast	'(71-78]'	'(88-inf)'	TRUE	yes
13	overcast	'(78-inf)'	'(-inf-7...	FALSE	yes
14	rainy	'(-inf-71]'	'(88-inf)'	TRUE	no