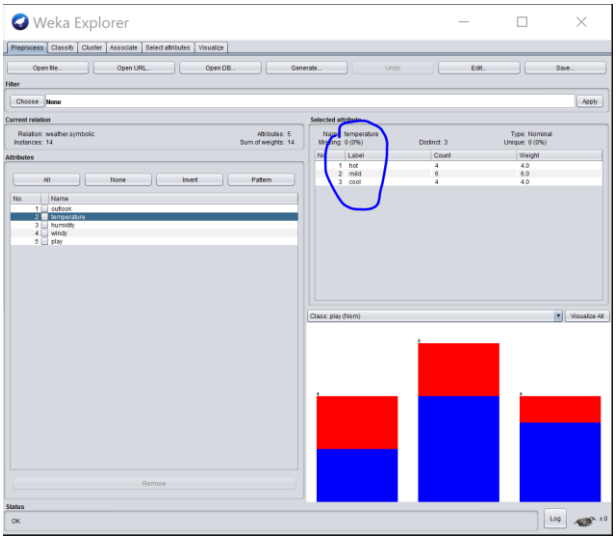
LIS590DT Data Mining Assignment #1

Jialu Wang (jwang282)

Exercise 17.1.1. What are the values that the attribute temperature can have?

The values of the attribute “temperature” are: hot, mid, cool.

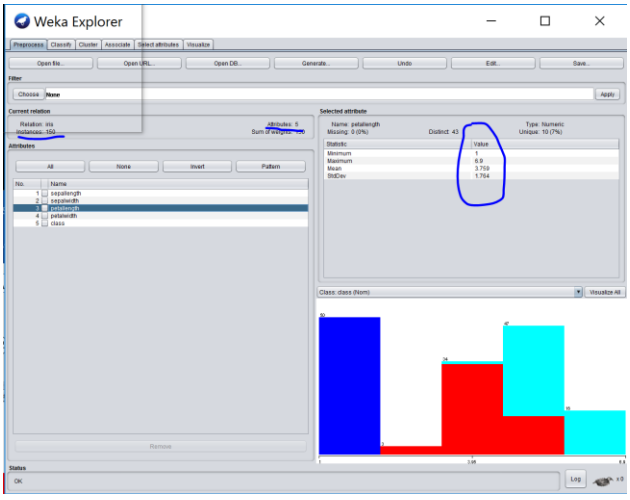


Exercise 17.1.2. Load a new dataset. Click the Open file button and select the file iris.arff, which corresponds to the iris dataset in Table 1.4. How many instances does this dataset have? How many attributes? What is the range of possible values of the attribute petal length?

Instances: 150

Attributes:5

Petallength: 6.9-1 = 5.9



Exercise 17.1.3. What is the function of the first column in the Viewer window?

It serves as the index of all 14 instances.

Exercise 17.1.4. What is the class value of instance number 8 in the weather data?

Instance#8: outlook:sunny;temperature:mild;humidity:high;windy:FALSE; play: no.

Relation: weather.symbolic

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

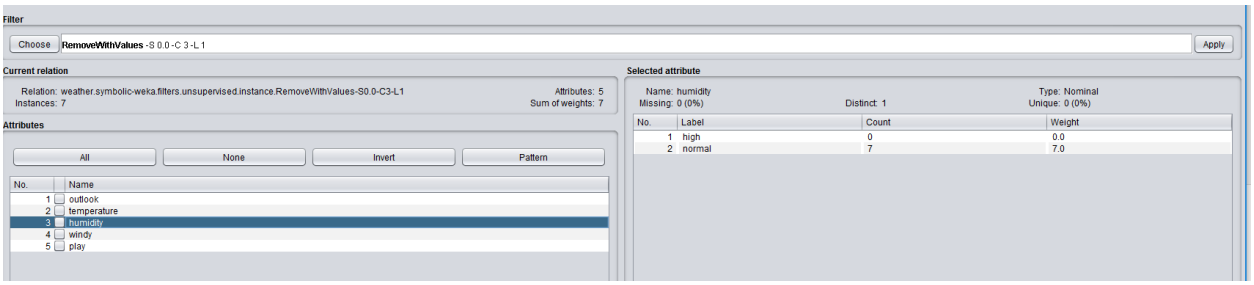
Exercise 17.1.5. Load the iris data and open it in the editor. How many numeric and how many nominal attributes does this dataset have?

4 numerical attributes and 1 nominal attribute.

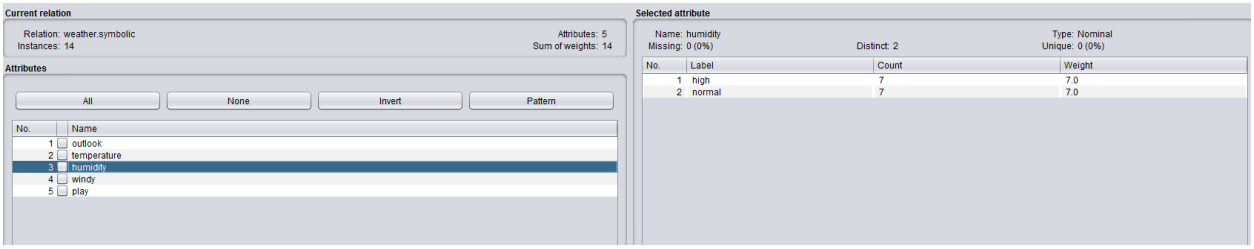
Relation: iris

No.	1: sepallength	2: sepalwidth	3: petallength	4: petalwidth	5: **class**
	Numeric	Numeric	Numeric	Numeric	Nominal
1	[illegible]	[illegible]	[illegible]	0.2	Iris-s...
2	4.9	3.0	1.4	0.2	Iris-s...
3	4.7	3.2	1.3	0.2	Iris-s...
4	4.6	3.1	1.5	0.2	Iris-s...
5	5.0	3.6	1.4	0.2	Iris-s...
6	5.4	3.9	1.7	0.4	Iris-s...

Exercise 17.1.6. Load the weather.nominal dataset. Use the filter weka.unsupervised.instance. RemoveWithValues to remove all instances in which the humidity attribute has the value high. To do this, first make the field next to the Choose button show the text RemoveWithValues. Then click on it to get the Generic Object Editor window, and figure out how to change the filter settings appropriately. [illustrated with image]

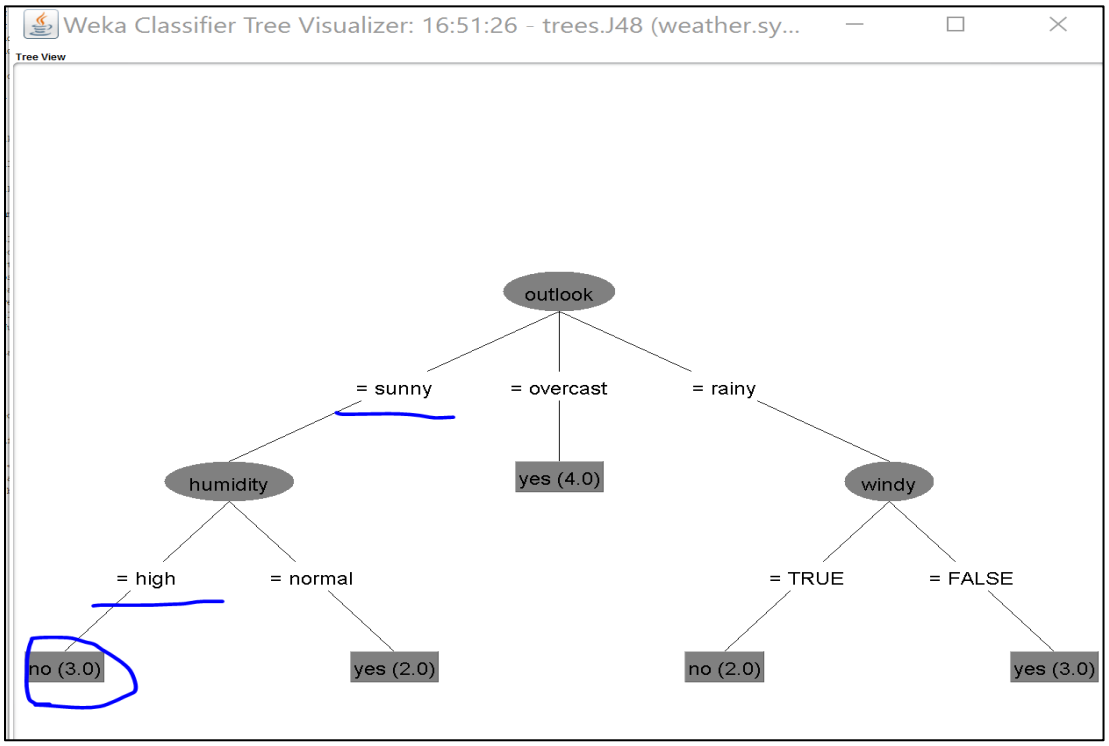


Exercise 17.1.7. Undo the change to the dataset that you just performed, and verify that the data has reverted to its original state. [illustrated with image]



Exercise 17.1.8. How would this instance be classified using the decision tree? outlook = sunny, temperature = cool, humidity = high, windy = TRUE

The instance would be classified as no (3.0). [illustrated with image]



Exercise 17.1.9. Load the iris data using the Preprocess panel. Evaluate C4.5 on this data using (a) the training set and (b) cross-validation. What is the estimated percentage of correct classifications for (a) and (b)? Which estimate is more realistic?

a) Training set. Estimated percentage of correct classifications: 98%

```
=== Summary ===

Correctly Classified Instances         147               98       %
Incorrectly Classified Instances         3                2       %
Kappa statistic                          0.97
Mean absolute error                      0.0233
Root mean squared error                  0.108
Relative absolute error                  5.2482 %
Root relative squared error             22.9089 %
Total Number of Instances              150
```

b) Cross validation. Estimated percentage of correct classifications: 96%

```
=== 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
```

The former is more realistic with smaller errors.

Exercise 17.1.10. Use the Visualize classifier errors function to find the wrongly classified test instances for the cross-validation performed in Exercise 17.1.9. What can you say about the location of the errors?

The 6 errors are outliers of the data set which are located at the middle of the border of the graph and significantly away from other 144 entries of data. They are within the right value range but of the wrong values.