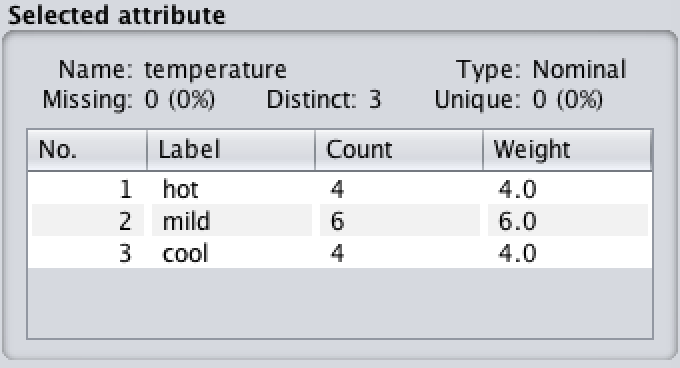
**IS590DT Data Mining Application**

**HW1 2017/8/29**

**yuweic3**

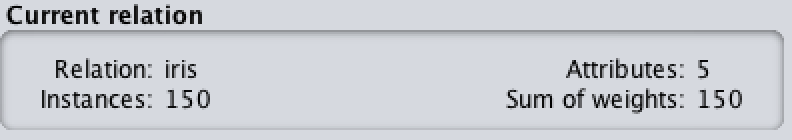
* **Ex 17.1.1**

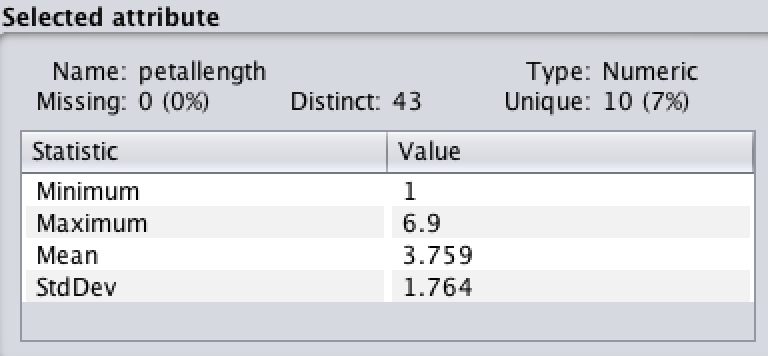
The values of *temperature* attribute include “hot”, mild” and “cool”.



* **Ex 17.1.2**

There are 150 instances and 5 attributes in the dataset. The range of “petallength” is from 1 to 6.9.





* **Ex 17.1.3**

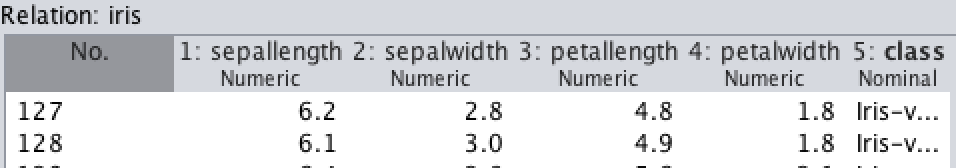
It is row number.

* **Ex 17.1.4**

The class value of instance number 8 is “no”.

* **Ex 17.1.5**

There are 4 numeric attributes and 1 nominal attributes.

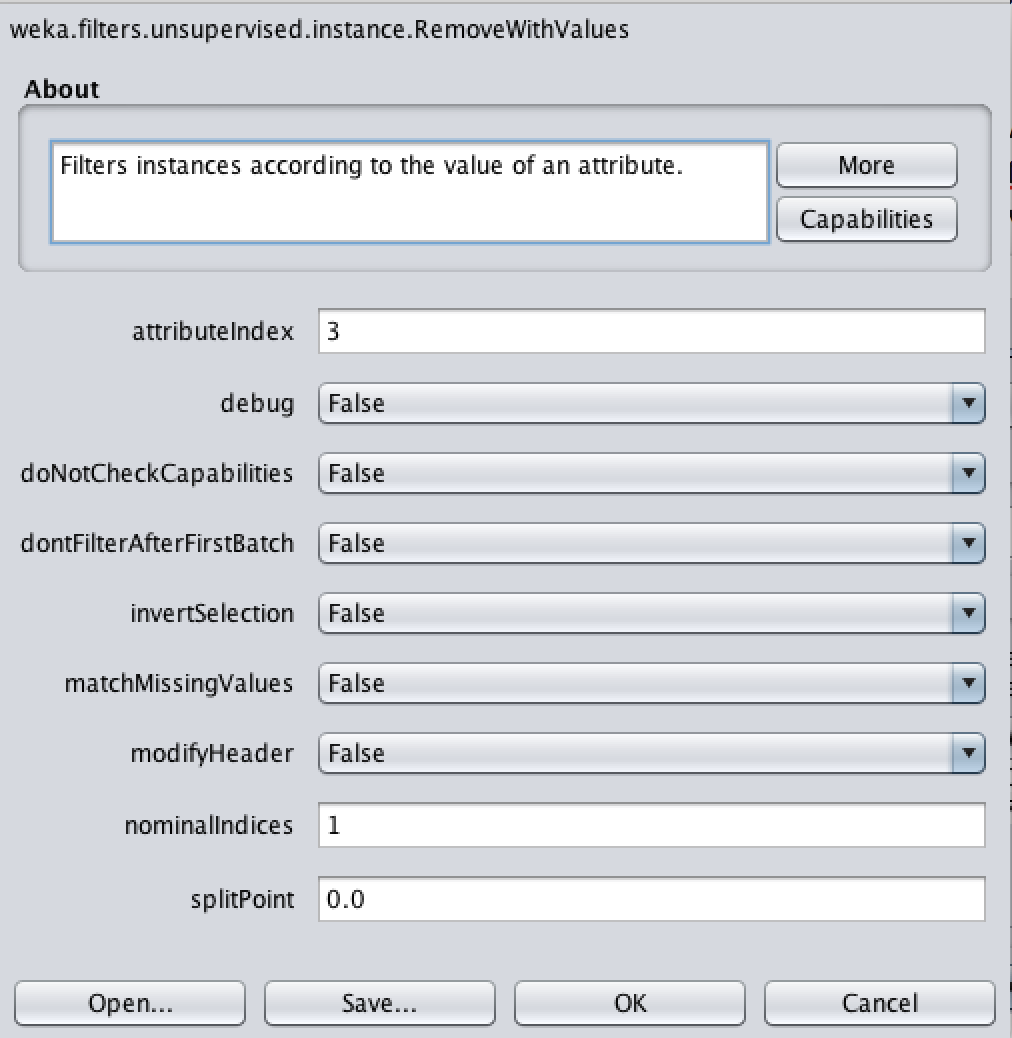


* **Ex 17.1.6**

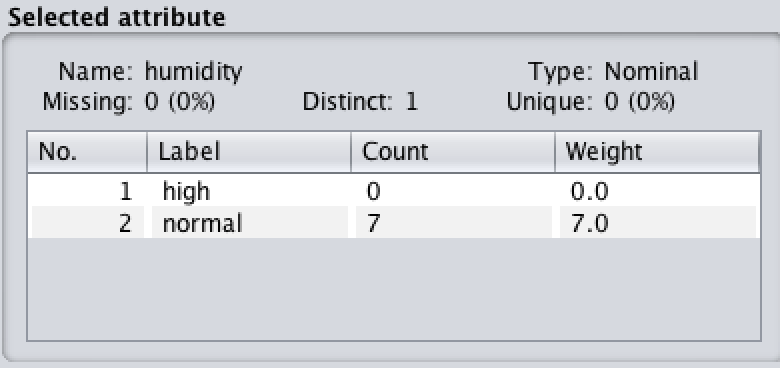
Before filtering, we check the *humidity* attribute to find that there are 7 instances labeled “high”.

To remove instances with “humidity-high” label, there are two things to be set in *Generic Object Editor* window.

1. attributeIndex = 3
2. nominalIndices = 1



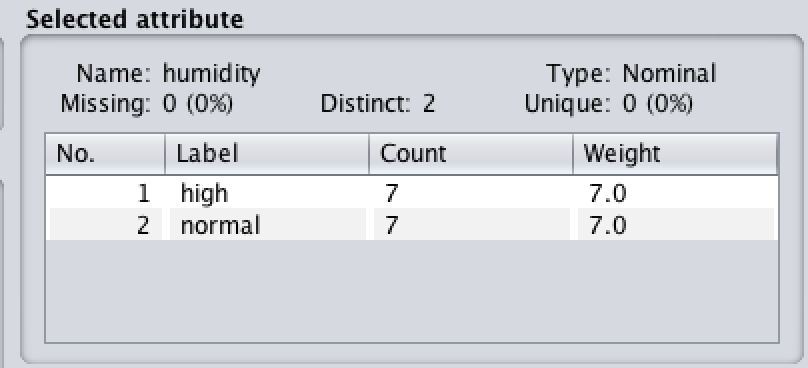
The filtered result is showed below:



Now there is no instance with “humidity-high” label.

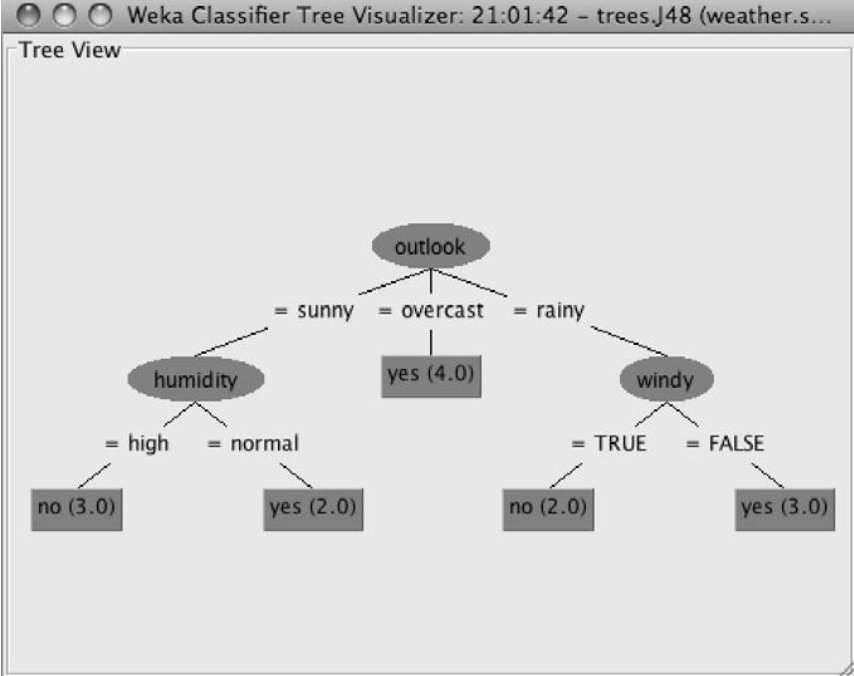
* **Ex 17.1.17**

Just click on “Undo” button.



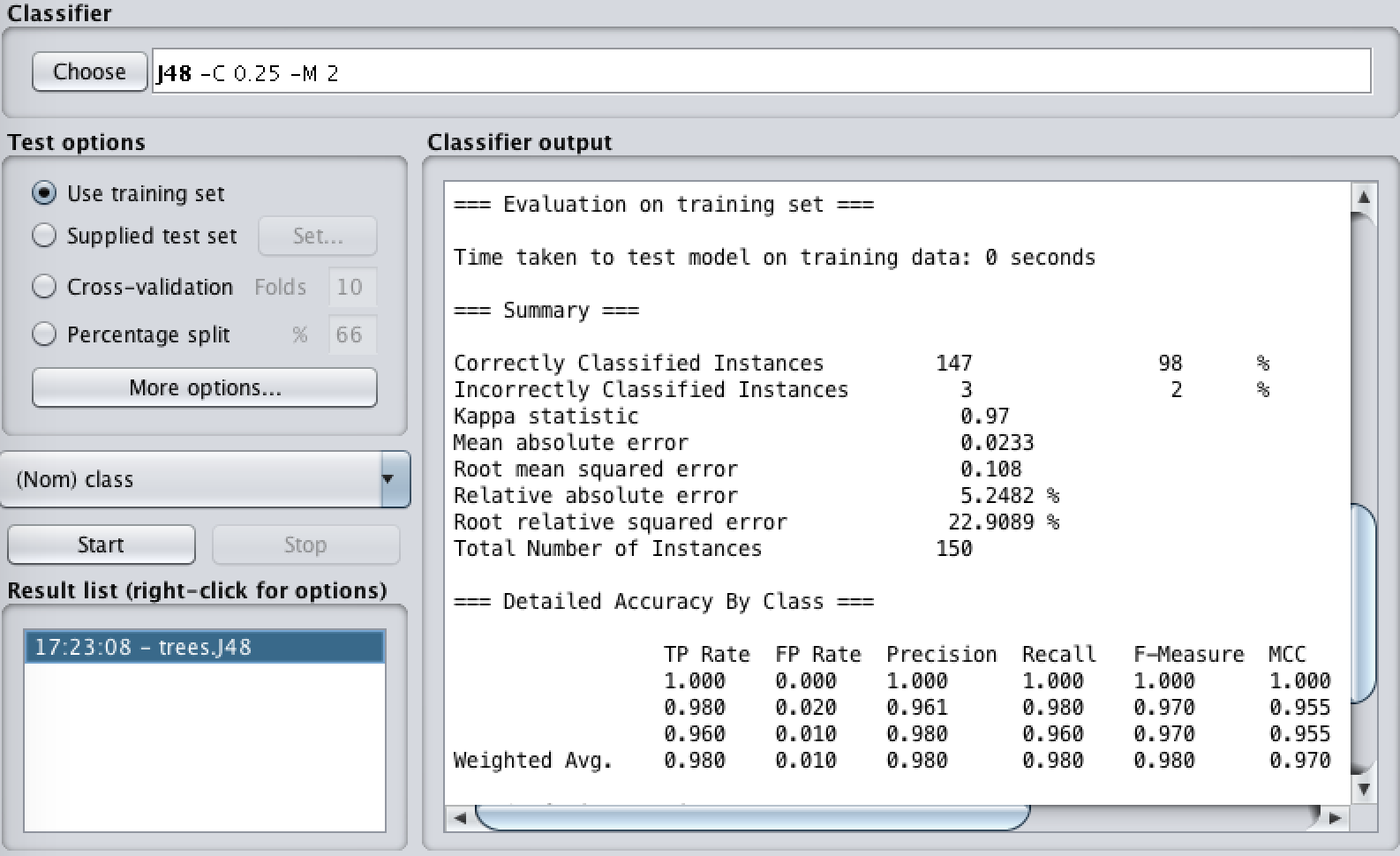
* **Ex 17.1.8**

It would be classified as “no”. From the decision tree, we apply the criterion of “outlook = sunny” and “humidity = high” and it leads us to “no”.

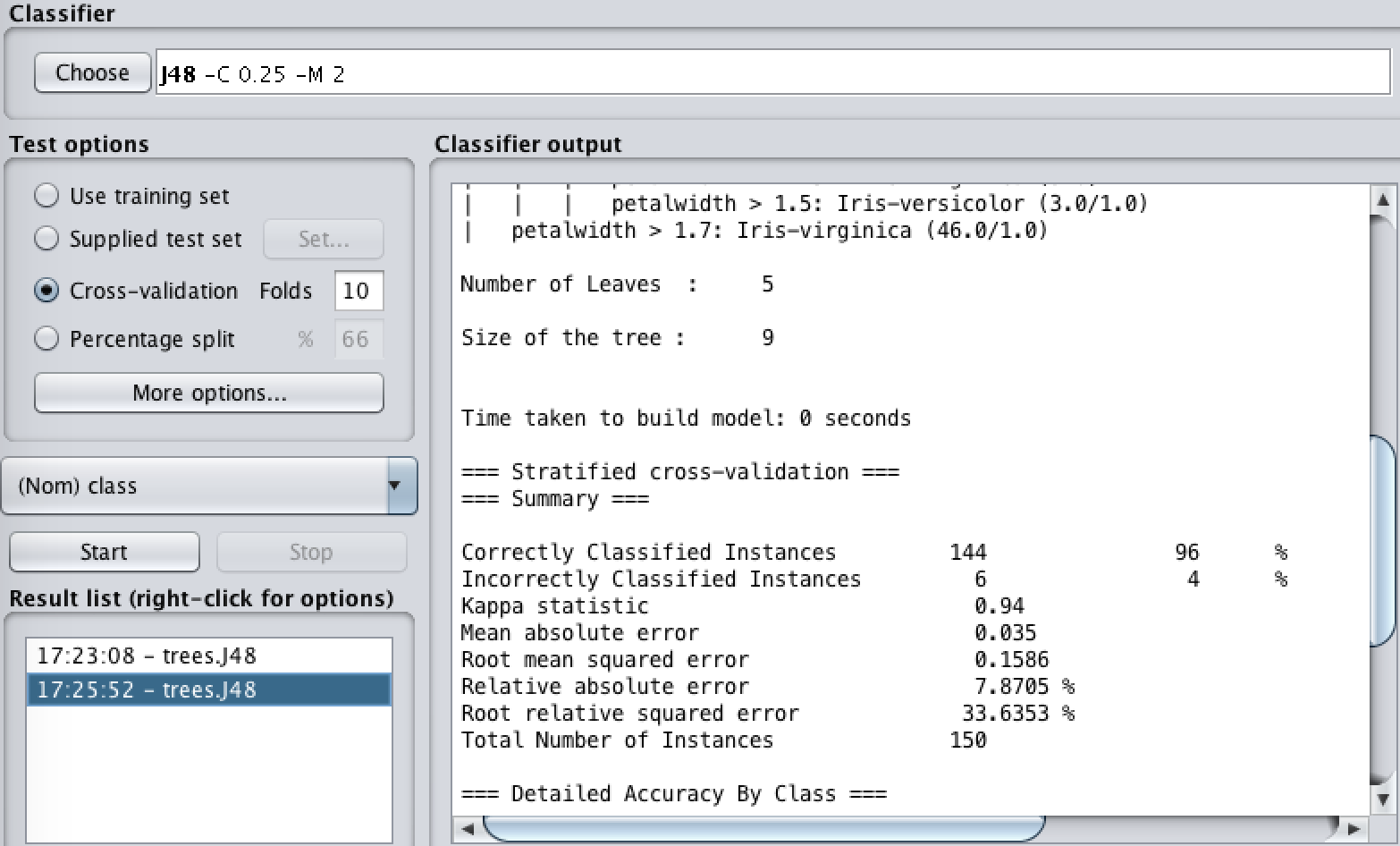


* **Ex 17.1.9**

The estimated percentage of correct classifications for “a) the training set” is 98%.



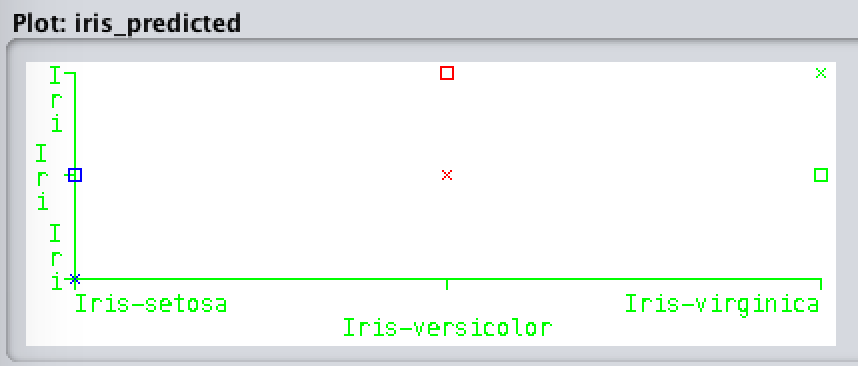
The estimated percentage of correct classifications for “b) cross-validation” is 96%.



The cross validation is more realistic.

Although a) outperforms b) with higher corrected classification, it is unrealistically optimistic performance estimate, because in the real world the classifier derived from training data set are almost evaluated on different test set instead of training set itself.

* **Ex 17.1.10**



Clearly, those correct classifications are located on the diagonal line () of the Coordinate.

Also from the location of squares, we know that

1. “Iris-Setosa” is more likely incorrectly classified into “Iris-Versicolor”;
2. “Iris-Versicolor” is more likely incorrectly classified into “Iris-Virginica”;
3. “Iris-Virginica” is more likely incorrectly classified into “Iris-Versicolor”.