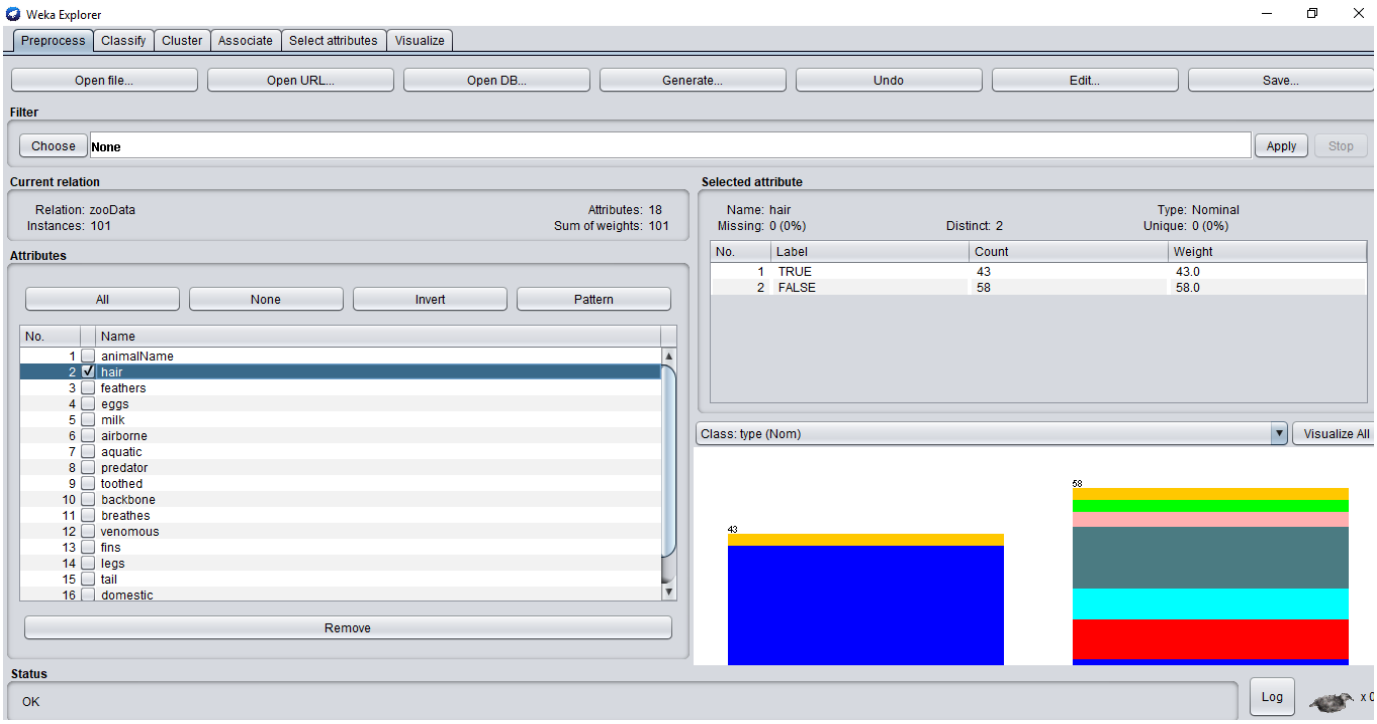
E/14/049

Chamara k.k.v

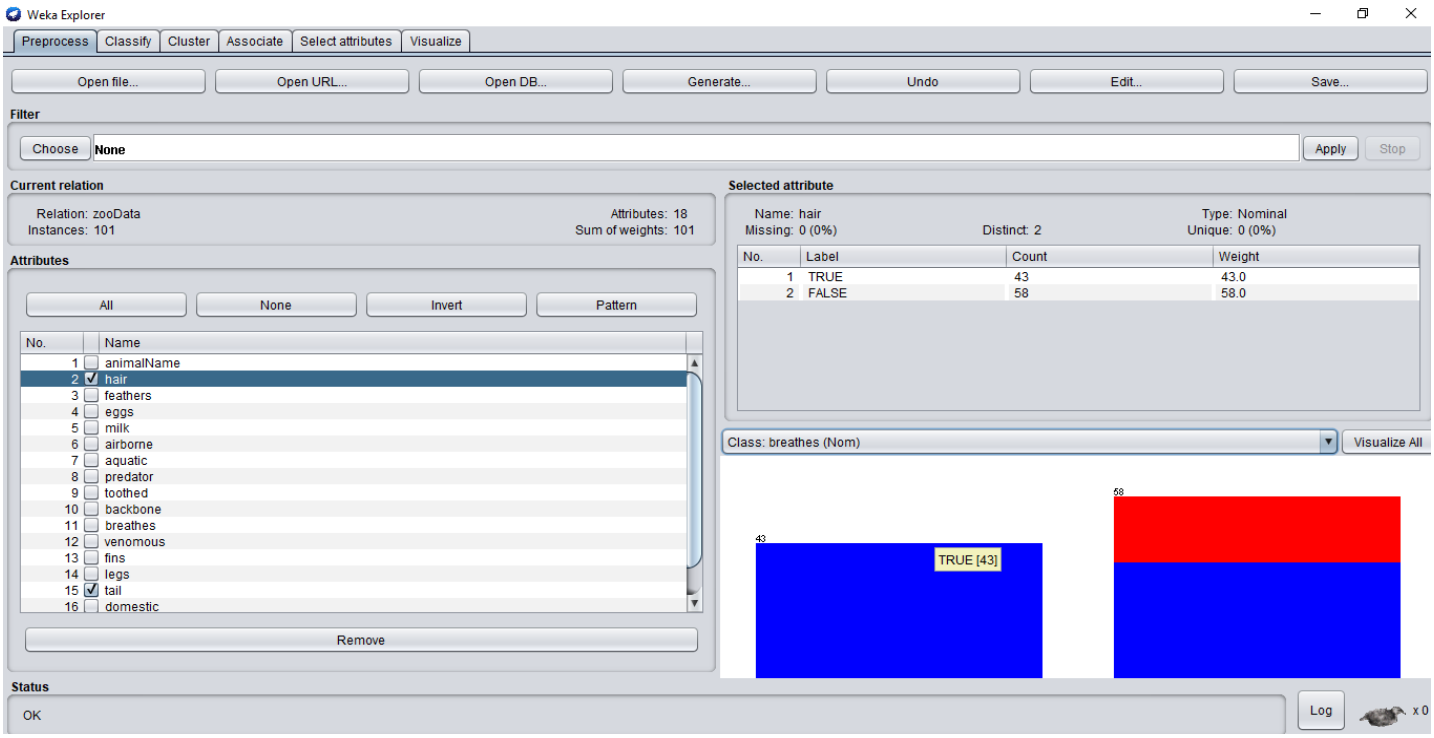
CO 544 Machine Learning and Data Mining – Lab 01 Report

Data Preprocessing in Weka

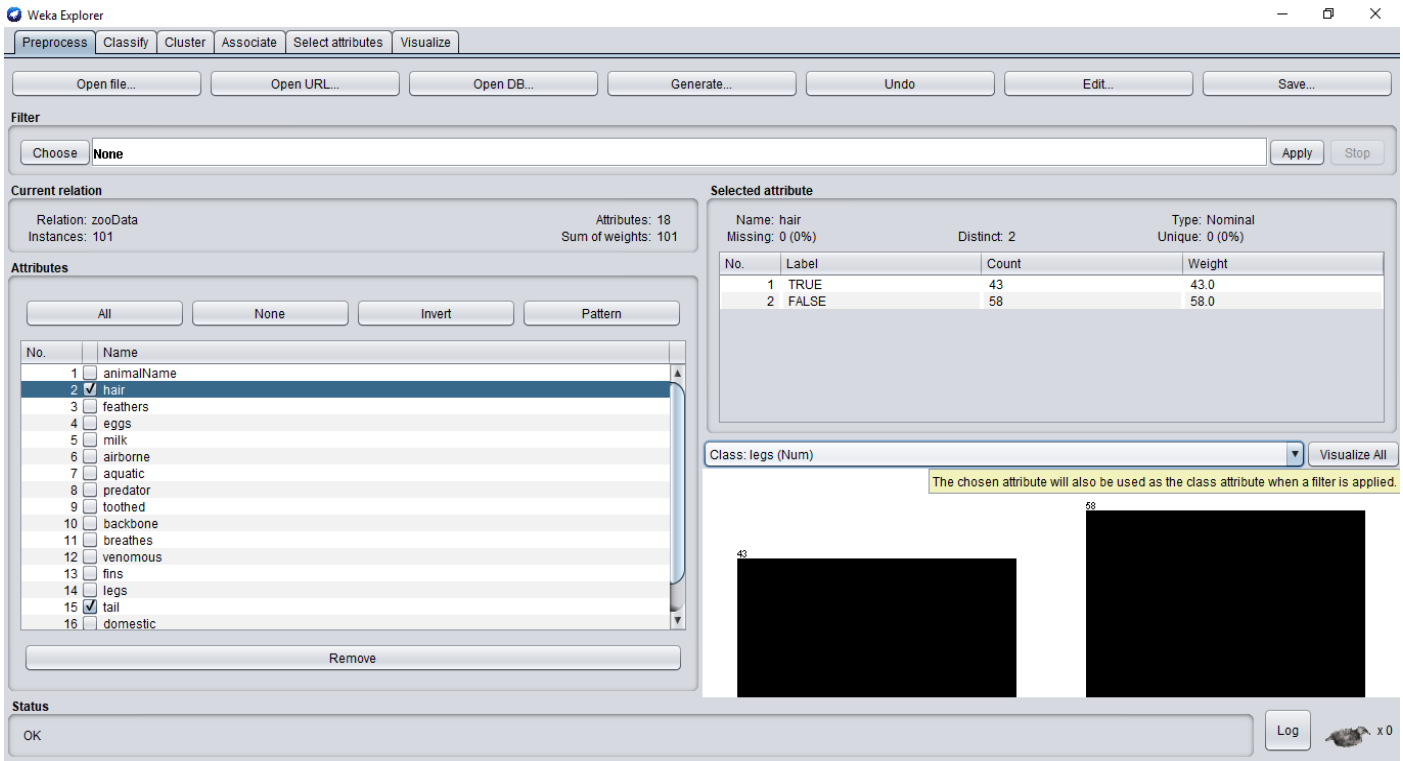
- In our example "zooData.arff " file we have categories such as animal name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, fins, legs, tail, domestic, capsize and type for each of seven different types of animals (mammal, bird, reptile, fish, amphibian, insect, invertebrate).
- After explore that arff file this window will pop up. In here we can select one attribute and see the selected attribute labels which may be nominal value, string or numeric value. Then in the selected attribute window we can get the results of selected attributes. In the below figure can see that I have selected the "hair" attribute and the results are either "true" and "false" which is 43 and 58 respectively. And we can get the weight or we can call it as a probability percentage of getting that label.



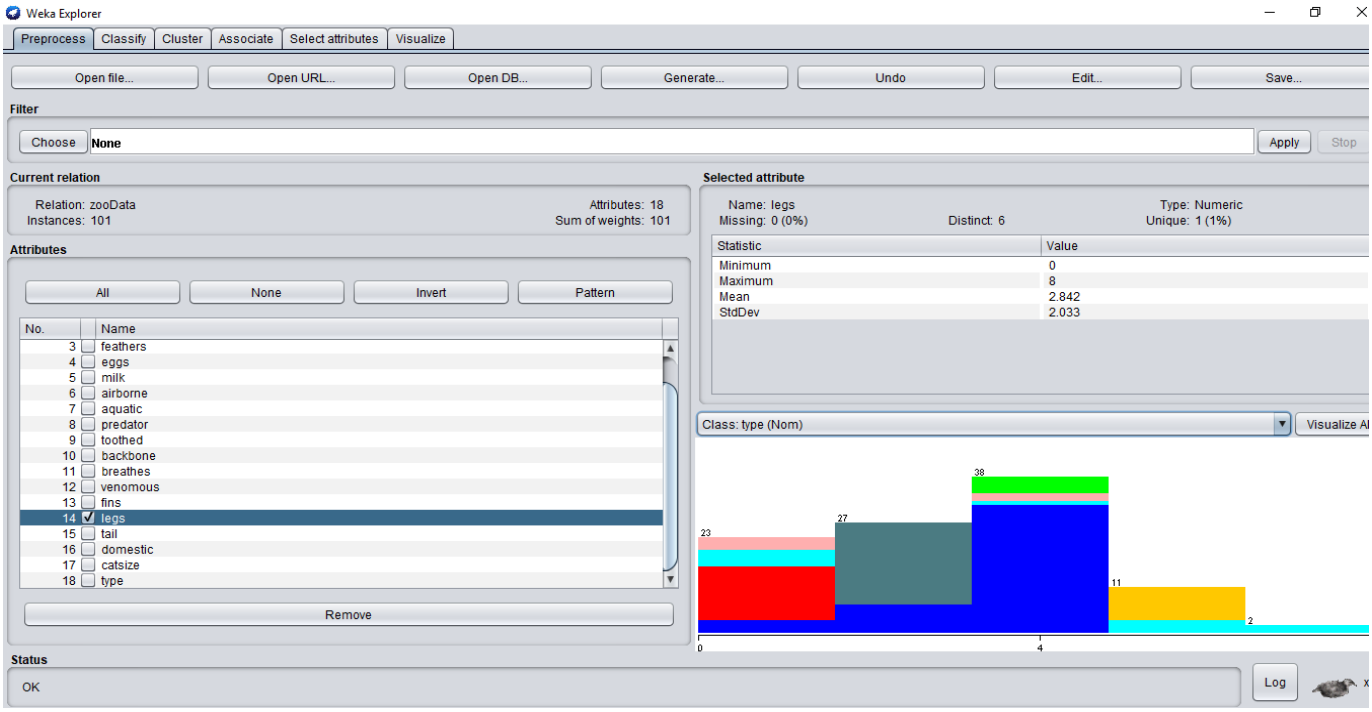
- And in the below bar graph shows the selected attribute results label wise. Which is also can be analysis as class wise too. In above figure show the Class "type". The below one show the result of class attribute of "breathes". Which shows that the animals who's having hair are in breathes animals. These kind of decision we can get by looking at these graphs.



As in the below figure if we select a irrelevant class then we will get the graph showing the results of the selected attribute only.

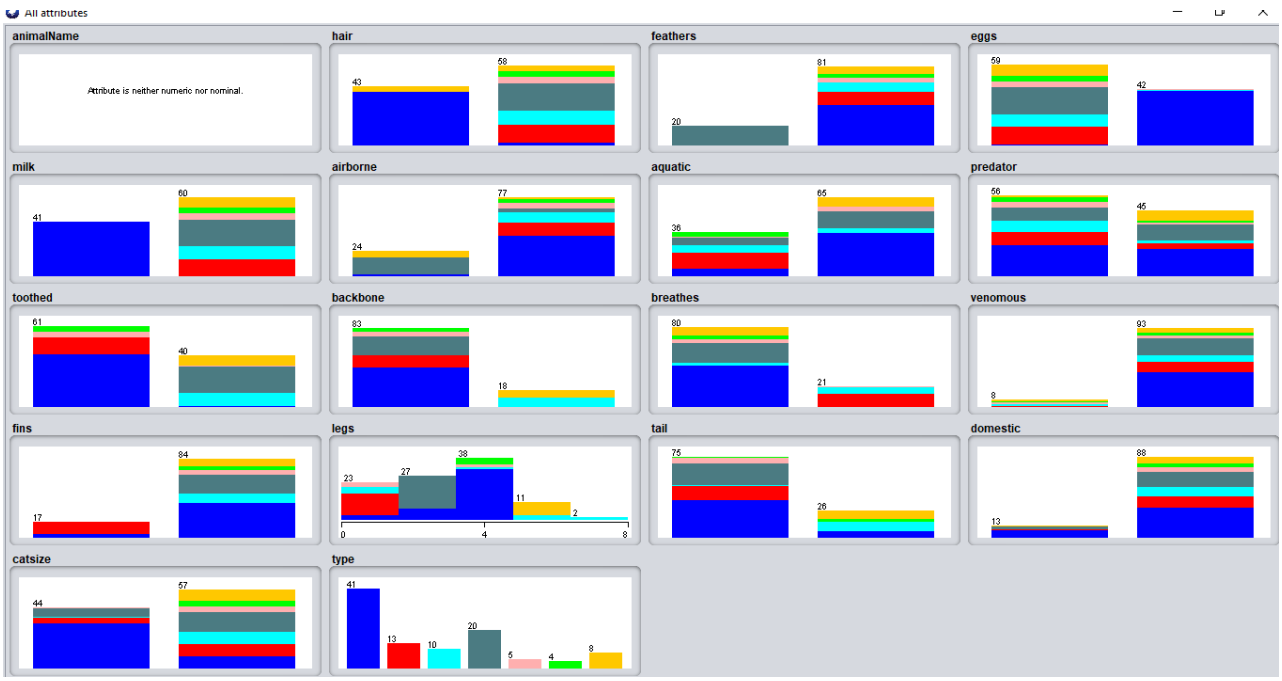


- If we visualize result of numerical value or sting value we can see the results will be show for each attributes. In the field Missing we can see how many percentage of data losses in the selected attribute. And also type is also mentioned. In numeric attributes we can have addition statistical analysis relevant to the data and their variations by looking at the average and



standard deviation.

Finally we can see the results of all attributes of by “visualizing all” like in the below figure.



Additional Useful Weka Capabilities:

- In Weka GUI chooser we can found ARFF file viewer. In there we can see the ARFF file and we can edit it.

ARFF-Viewer - F:\hard\sem 6\data mining\weka\Weka-3-8\data\lab.arff

File Edit View

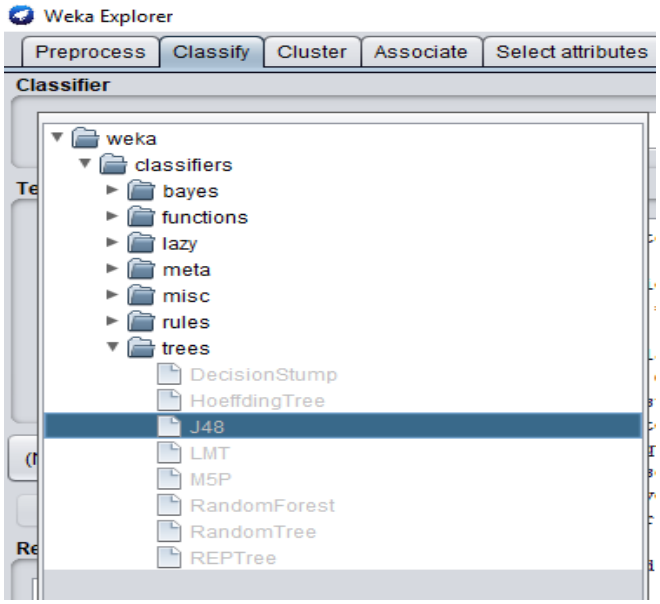
lab.arff

Relation: zooData

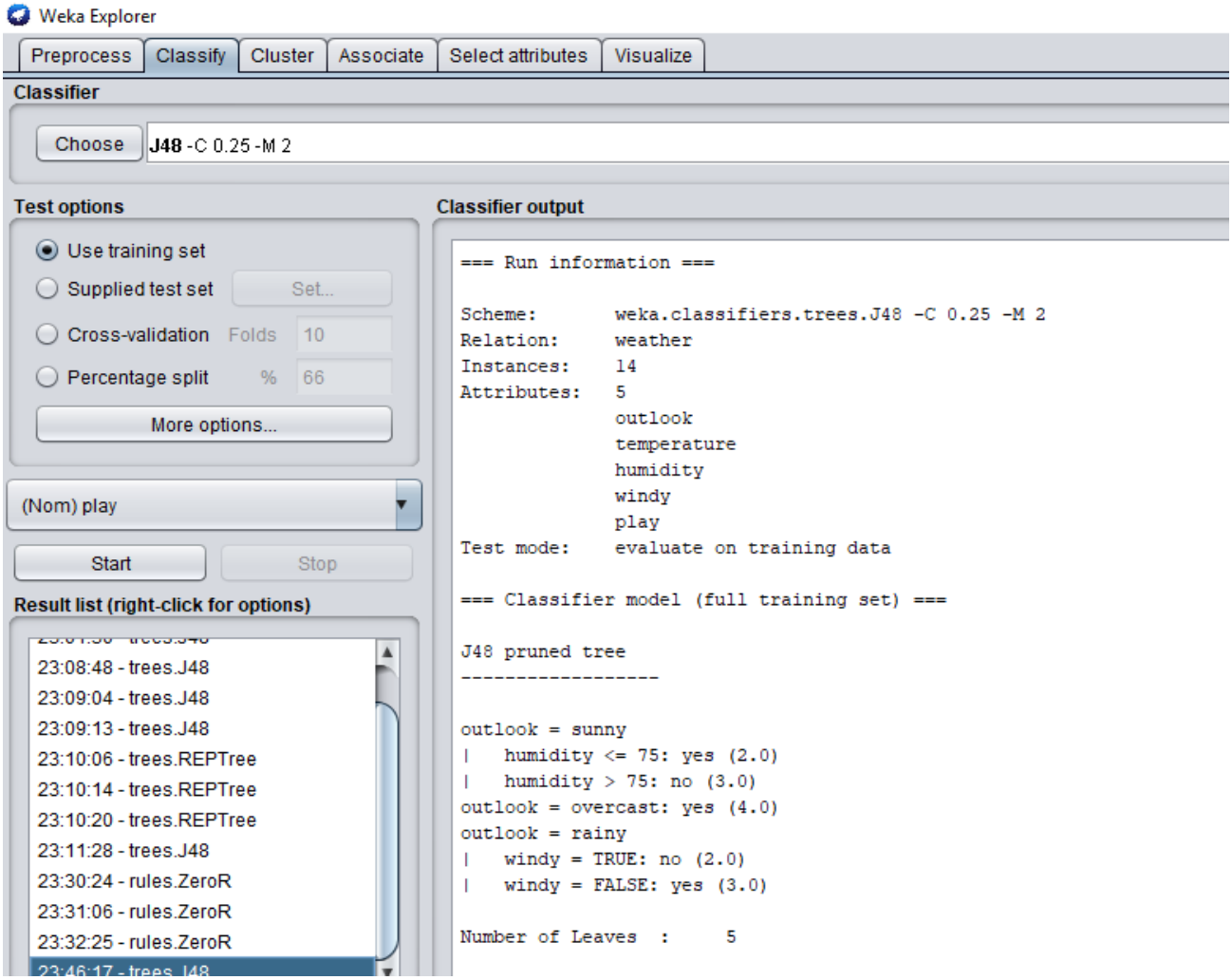
No.	1: animalName String	2: hair Nominal	3: feathers Nominal	4: eggs Nominal	5: milk Nominal	6: airborne Nominal	7: aquatic Nominal	8: predator Nominal	9: toothed Nominal	10: backbone Nominal	11: breathes Nominal	12: venomous Nominal	13: fins Nominal	14: legs Numeric	15: tail Nominal	16: domestic Nominal	17: catsize Nominal	18: type Nominal
1	aardvark	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	FAL...	FALSE	TRUE	mam...
2	antelope	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
3	bass	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
4	bear	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	FAL...	FALSE	TRUE	mam...
5	boar	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
6	buffalo	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
7	calf	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	TRUE	TRUE	mam...
8	carp	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	TRUE	FALSE	fish
9	catfish	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
10	cavy	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	FAL...	TRUE	FALSE	mam...
11	cheetah	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
12	chicken	FAL...	TRUE	TRUE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	TRUE	FALSE	bird
13	chub	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
14	clam	FAL...	FALSE	TRUE	FAL...	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FAL...	0.0	FAL...	FALSE	FALSE	invert...
15	crab	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FAL...	4.0	FAL...	FALSE	FALSE	invert...
16	crayfish	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FAL...	6.0	FAL...	FALSE	FALSE	invert...
17	crow	FAL...	TRUE	TRUE	FAL...	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	FALSE	FALSE	bird
18	deer	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
19	dogfish	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	TRUE	fish
20	dolphin	FAL...	FALSE	FAL...	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	0.0	TRUE	FALSE	TRUE	mam...
21	dove	FAL...	TRUE	TRUE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	TRUE	FALSE	bird
22	duck	FAL...	TRUE	TRUE	FAL...	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	FALSE	FALSE	bird
23	elephant	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
24	flamingo	FAL...	TRUE	TRUE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	FALSE	TRUE	bird
25	flea	FAL...	FALSE	TRUE	FAL...	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAL...	6.0	FAL...	FALSE	FALSE	insect
26	frog	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	FAL...	FALSE	FALSE	amp...
27	frog	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FAL...	4.0	FAL...	FALSE	FALSE	amp...
28	fruitbat	TRUE	FALSE	FAL...	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	FALSE	FALSE	mam...
29	giraffe	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	TRUE	mam...
30	gorilla	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FAL...	2.0	FAL...	TRUE	TRUE	mam...
31	gnat	FAL...	FALSE	TRUE	FAL...	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAL...	6.0	FAL...	FALSE	FALSE	insect
32	goat	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	TRUE	TRUE	mam...
33	gorilla	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	2.0	FAL...	FALSE	TRUE	mam...
34	gull	FAL...	TRUE	TRUE	FAL...	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FAL...	2.0	TRUE	FALSE	FALSE	bird
35	haddock	FAL...	FALSE	TRUE	FAL...	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
36	hamster	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	TRUE	FALSE	mam...
37	hare	TRUE	FALSE	FAL...	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FAL...	4.0	TRUE	FALSE	FALSE	mam...

Classifiers & Clustering:

- There are several classification methods available in Weka as in the below figure. The method that we are considering is J48 tree.

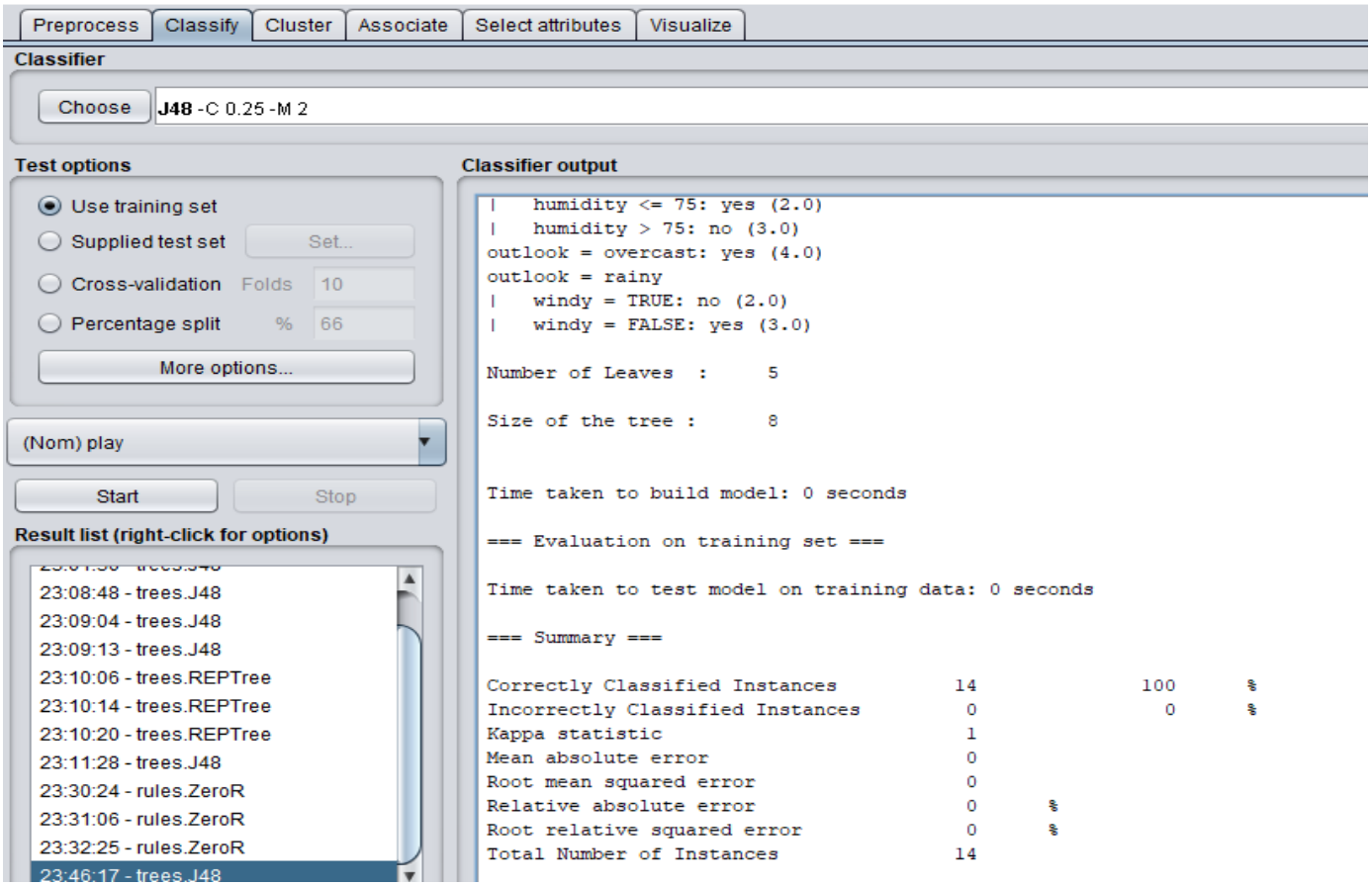


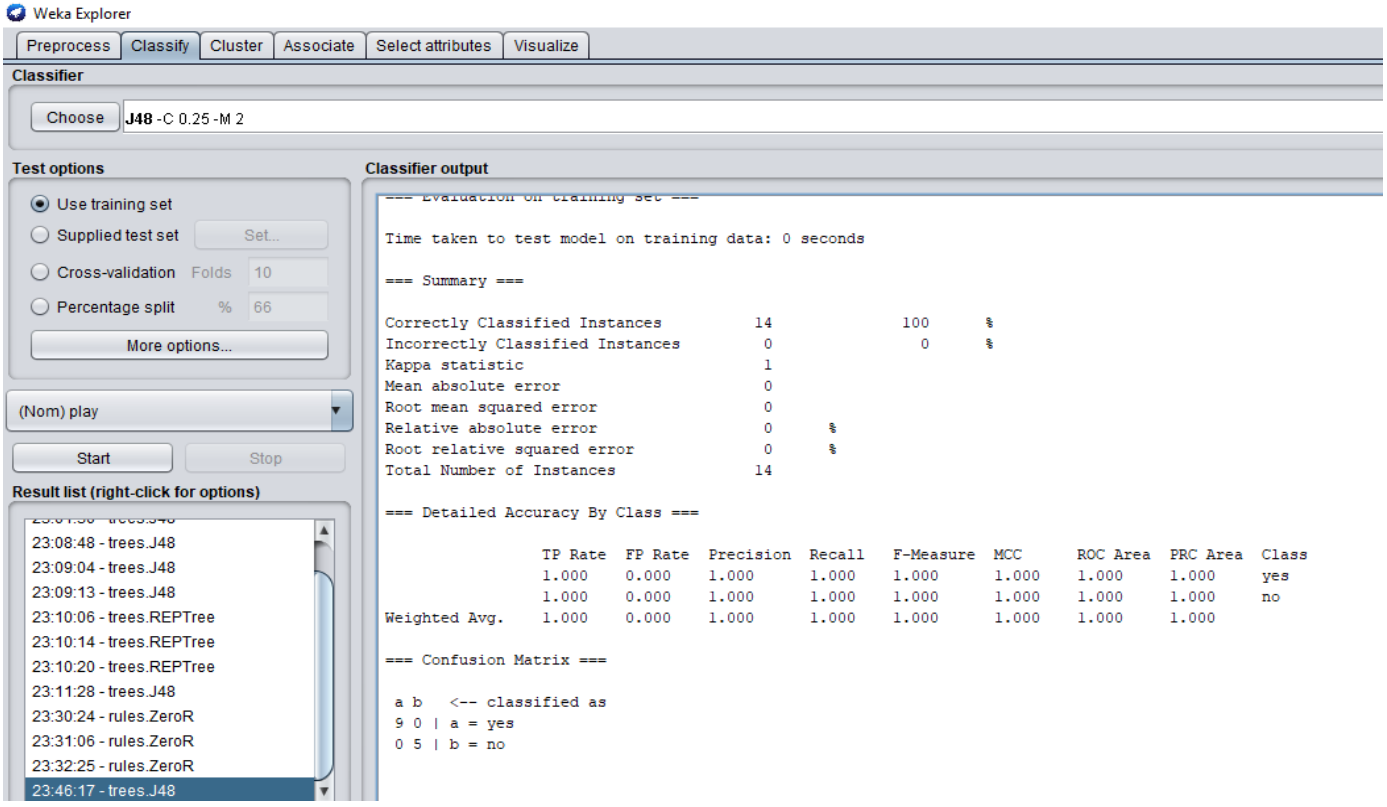
- At that point we are ready to create the model. Selecting the “Use Selecting Set” so that it uses the data set that have just loaded to the model. Then hit on the button “Start”.



- The important numbers to focus on here are the numbers next to the "Correctly Classified Instances" (100 percent) and the "Incorrectly Classified Instances" (0 percent). Other important numbers are in the "ROC Area" column, in the first row (the 1). Finally, in the "Confusion Matrix," it shows you the number of false positives and false negatives. Therefore no false positives and false negatives in this data set.

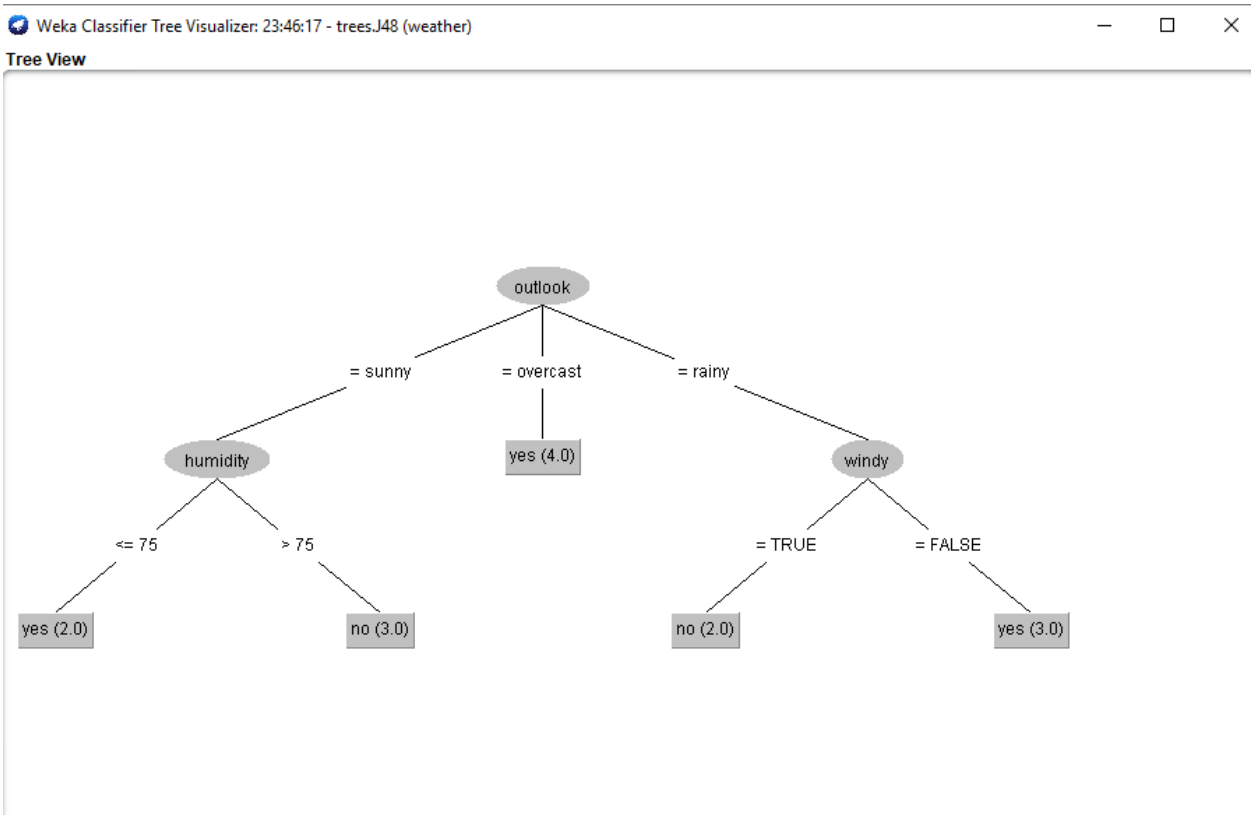
- We can say that this model is good model for this data set because the accuracy is 100%.


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
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
23:08:48 - trees.J48
23:09:04 - trees.J48
23:09:13 - trees.J48
23:10:06 - trees.REPTree
23:10:14 - trees.REPTree
23:10:20 - trees.REPTree
23:11:28 - trees.J48
23:30:24 - rules.ZeroR
23:31:06 - rules.ZeroR
23:32:25 - rules.ZeroR
23:46:17 - trees.J48
Classifier output
| humidity <= 75: yes (2.0)
| humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 14 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0
Root mean squared error 0
Relative absolute error 0 %
Root relative squared error 0 %
Total Number of Instances 14


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
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
23:08:48 - trees.J48
23:09:04 - trees.J48
23:09:13 - trees.J48
23:10:06 - trees.REPTree
23:10:14 - trees.REPTree
23:10:20 - trees.REPTree
23:11:28 - trees.J48
23:30:24 - rules.ZeroR
23:31:06 - rules.ZeroR
23:32:25 - rules.ZeroR
23:46:17 - trees.J48
Classifier output
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 14 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0
Root mean squared error 0
Relative absolute error 0 %
Root relative squared error 0 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 yes
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 no
Weighted Avg. 1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000
=== Confusion Matrix ===
a b <-- classified as
9 0 | a = yes
0 5 | b = no

- Confusion matrix explain that number of correct and incorrect predictions that an instance is negative or positive. (http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html)

- We can see the tree by right-clicking on the model created, in the result list. On the pop-up menu, select Visualize tree. We will see the classification tree we just created.



- If the data set is giving low accuracy then we can test using another test data as giving them “Supplied test set” radio button. Then testing result may be better than getting data which are loaded to the model.