

Viewer

Relation: Capitals

No.	1: Name Nominal	2: Has_subway Nominal	3: major_government_office Nominal	4: Has_airport Nominal	5: Is_Capital Nominal
1	New York	true	true	true	false
2	Washington DC	true	true	true	true
3	London	true	true	true	true
4	Toronto	true	true	true	true
5	Cape Town	true	true	false	true
6	Paris	true	true	true	true
7	Amsterdam	false	true	true	true
8	Munich	false	true	true	false
9	Milan	false	true	true	false

Imagine a fictional dataset which we are using to build a machine learning model that predicts whether a city is the capital of a country or not. **The data is completely fictional and does not represent the actual cities in any way.** So after the dataset is loaded we can then use the classification task in Weka. We choose classification because the class variable “Is_Capital” is discrete.



Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set
 Supplied test set
 Cross-validation Folds
 Percentage split %

(Nom) Is_Capital

Result list (right-click for options)

20:48:12 - trees.J48

Classifier output

NUMBER OF LEAVES : 2

Size of the tree : 3

Time taken to build model: 0 seconds

== Stratified cross-validation ==

== Summary ==

Correctly Classified Instances	5	55.5556 %
Incorrectly Classified Instances	4	44.4444 %
Kappa statistic	-0.2	
Mean absolute error	0.5	
Root mean squared error	0.6078	
Relative absolute error	100	%
Root relative squared error	116.9749	%
Total Number of Instances	9	

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.167	0.000	0.000	0.000	0.000	-0.250	0.556	0.556	false
0.833	1.000	0.625	0.833	0.833	0.714	-0.250	0.556	0.806	true
Weighted Avg.	0.556	0.722	0.417	0.556	0.476	-0.250	0.556	0.722	

== Confusion Matrix ==

a b <-- classified as
0 3 | a = false
1 5 | b = true

Using J48 and cross validation, we didn't do too well. We were only allowed to use up to 9 folds, as Weka does not allow more folds than the number of instances.

Status

OK

Log



Scenario: Suppose now that we have created this J48 classifier with the data, we want to see whether it really works or not. So, side by side a friend of ours develops a test set independently to test it with the classifier that we have built.

```
@RELATION Capitals

@ATTRIBUTE Name {'New York', 'Washington DC', 'London', 'Toronto', 'Cape Town', 'Paris', 'Amsterdam', 'Munich', 'Milan'}
@ATTRIBUTE Has_subway {false, true}
@ATTRIBUTE major_government_office {false, true}
@ATTRIBUTE Has_airport {false, true}
@ATTRIBUTE Is_Capital {false, true}

@DATA
|
'New York',true,true,true,false
'Washington DC',true,true,true,true
'London',true,true,true,true
'Toronto',true,true,true,true
'Cape Town',true,true,false,true
'Paris',true,true,true,true
'Amsterdam',false,true,true,true
'Munich',false,true,true,false
'Milan',false,true,true,false
```

This is the original arff file with the attributes, you see how the nominal attributes must be predefined such as the names of the cities.

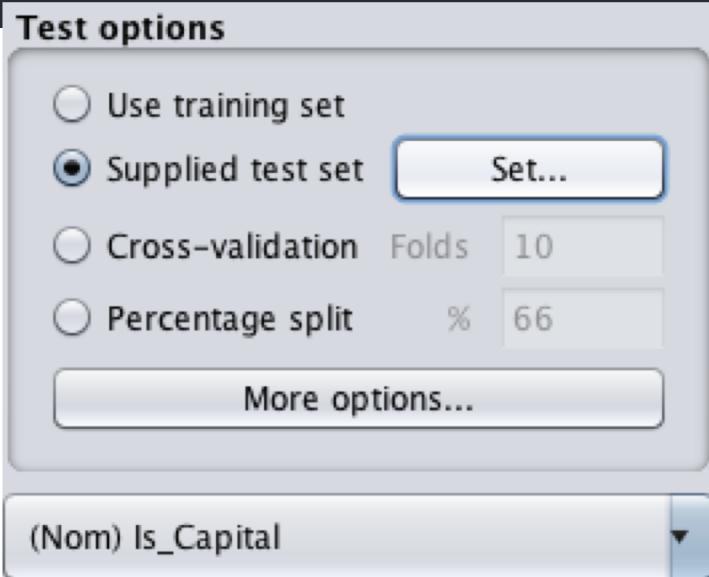
@RELATION Capitals

```
@ATTRIBUTE Name {'New York', 'Washington DC', 'London', 'Toronto', 'Cape Town', 'Paris', 'Amsterdam', 'Munich', 'Milan', 'Stockholm'  
@ATTRIBUTE Has_subway {false, true}  
@ATTRIBUTE major_government_office {false, true}  
@ATTRIBUTE Has_airport {false, true}  
@ATTRIBUTE Is_Capital {false, true}
```

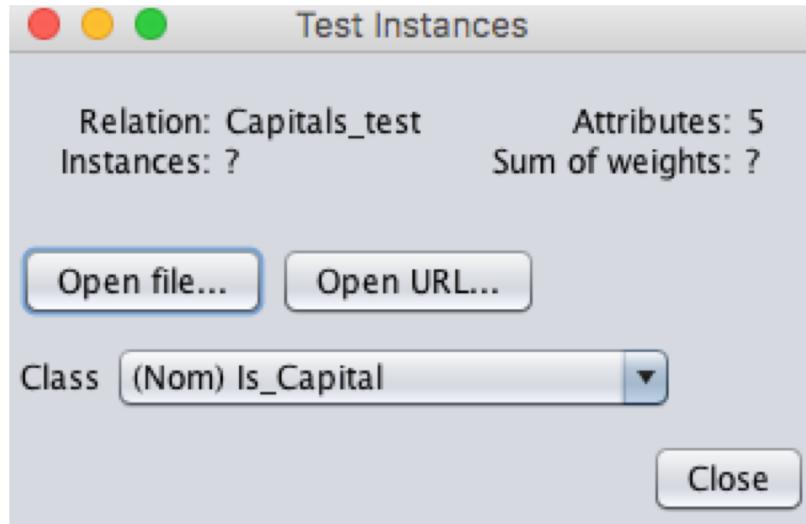
@DATA

```
'Stockholm',true,true,true,true  
'Barcelona',true,true,true,false
```

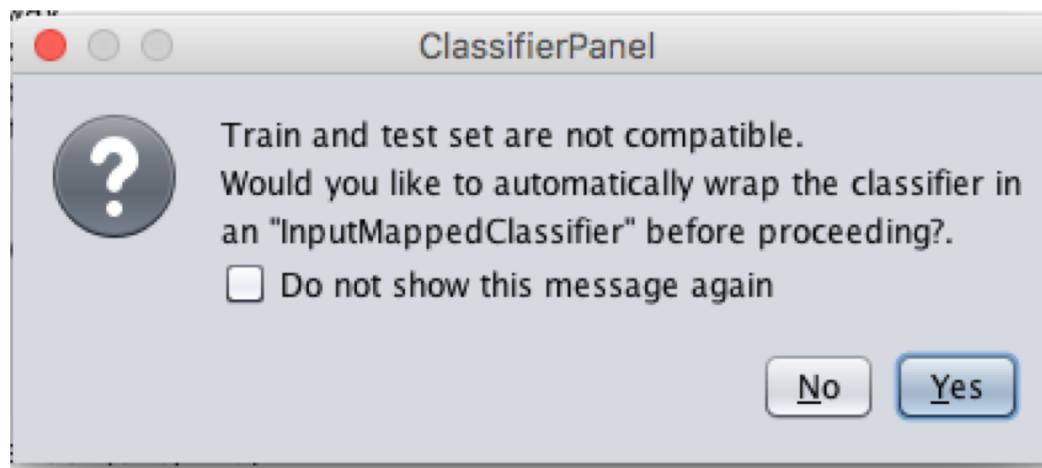
This is the new test set that you want to put into Weka, we think of two new cities to test Stockholm and Barcelona, we know if they have subways, have airports, have government offices or whether or not they are capitals. So we can fill in the information for each attributes for our test cities. The one thing we have to remember is because we have new cities, this new information should be updated in the attribute above as well.



After that we want to supply the test set into Weka



After loading the test set we might be interested in declaring the class, maybe we want to find certain metrics.



Weka detects some changes between the test and train set, which is true but Weka is able to adjust the changes.