Data mining and classification

Dra. Dª. Macarena Espinilla Estévez

Universidad de Jaén

Data mining and classification

UJA.es

- Dr. Macarena Espinilla
  - University of Jaén
  - Spain
- Research group
  - Intelligent Systems Based on Fuzzy Decision Analysis
- Research topics
  - Soft computing techniques
  - Decision making models
    - Uncertainty context
    - Linguistic information
  - Applications
    - Sensory evaluation
    - Sustainable energy evaluation
    - Performance appraisal

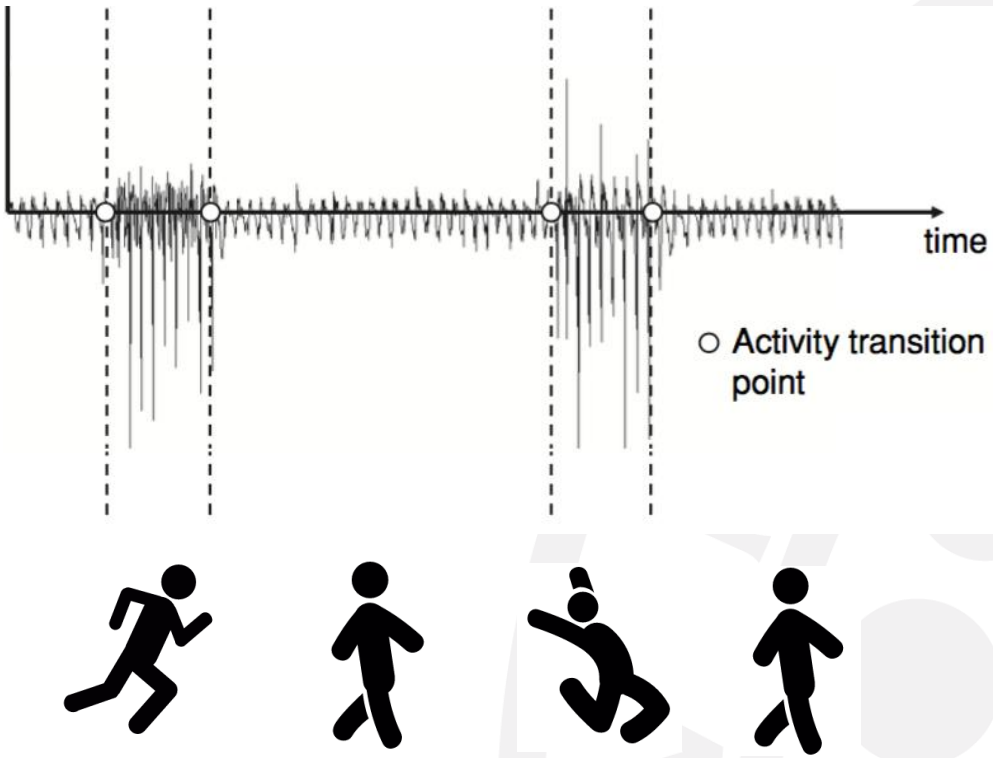
SINBAD²

http://sinbad2.ujaen.es

MacarenaUja

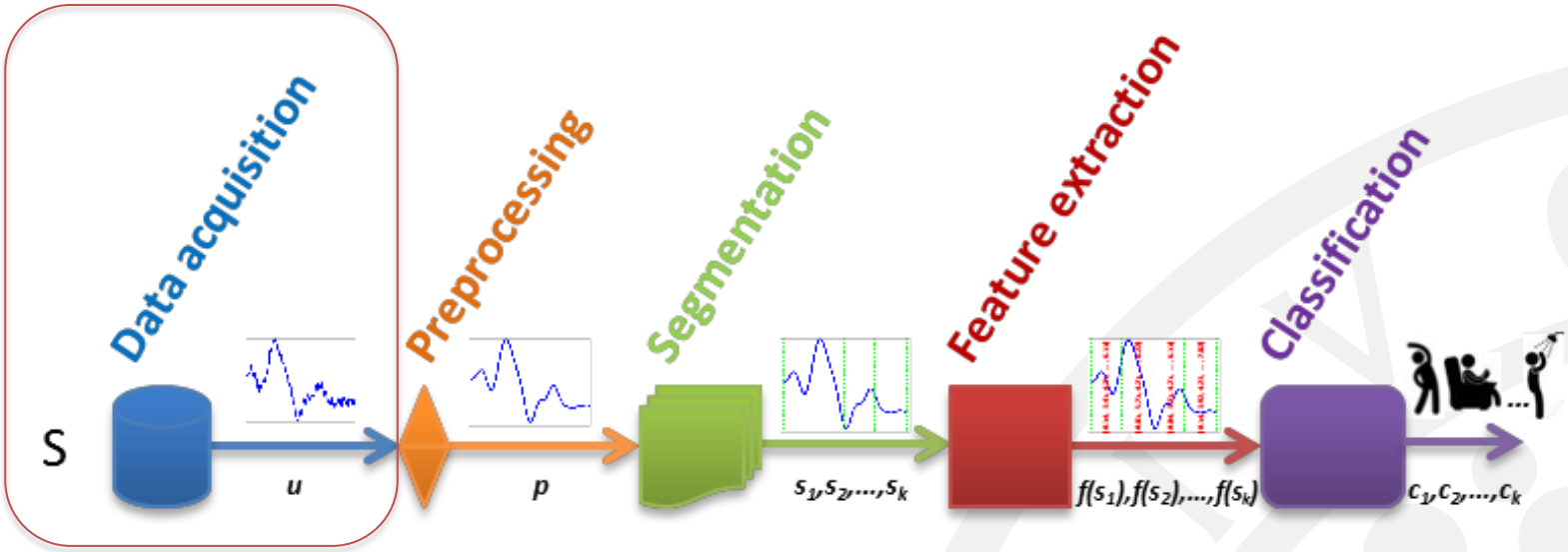
Tutorial Context

- The final aim is to build an activity model in order to classify activities based on the data generated by a sensor



Tutorial Context

- Session 1: Data collection. Shimmer sensor



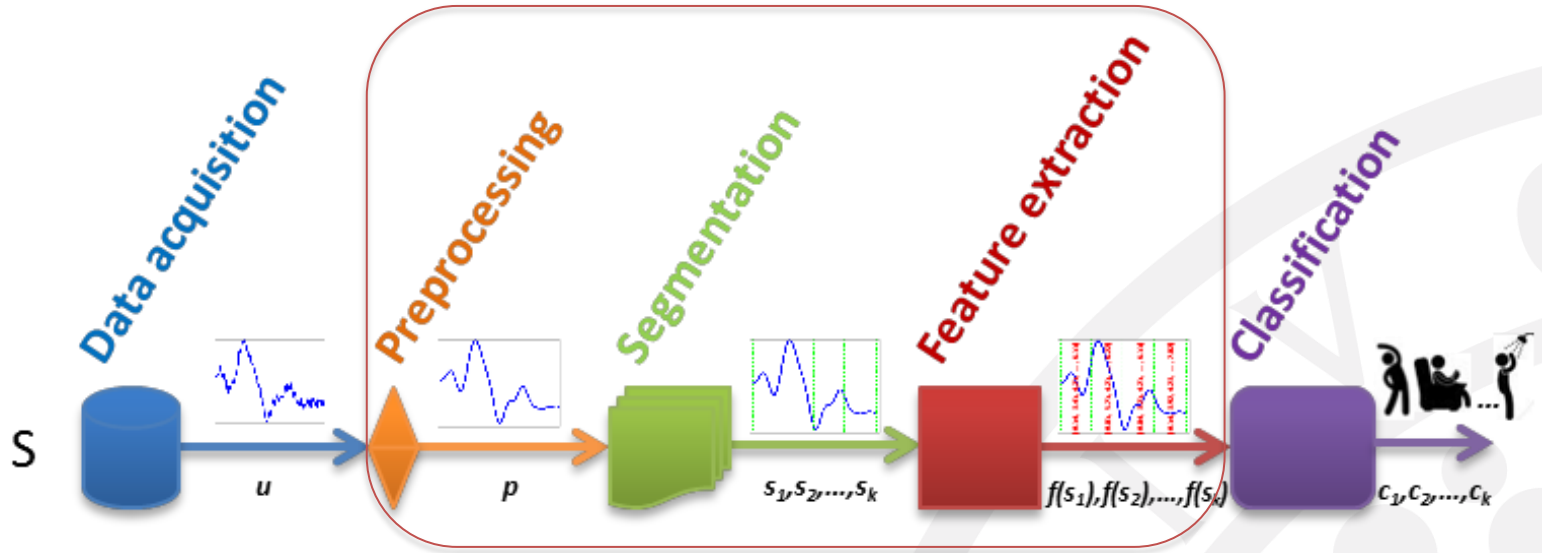
S = data source (sensor)
s_i = segment of data

u = raw/unprocessed data
$f(s_i)$ = feature vector

p = preprocessed data
c_i = class/label

Tutorial Context

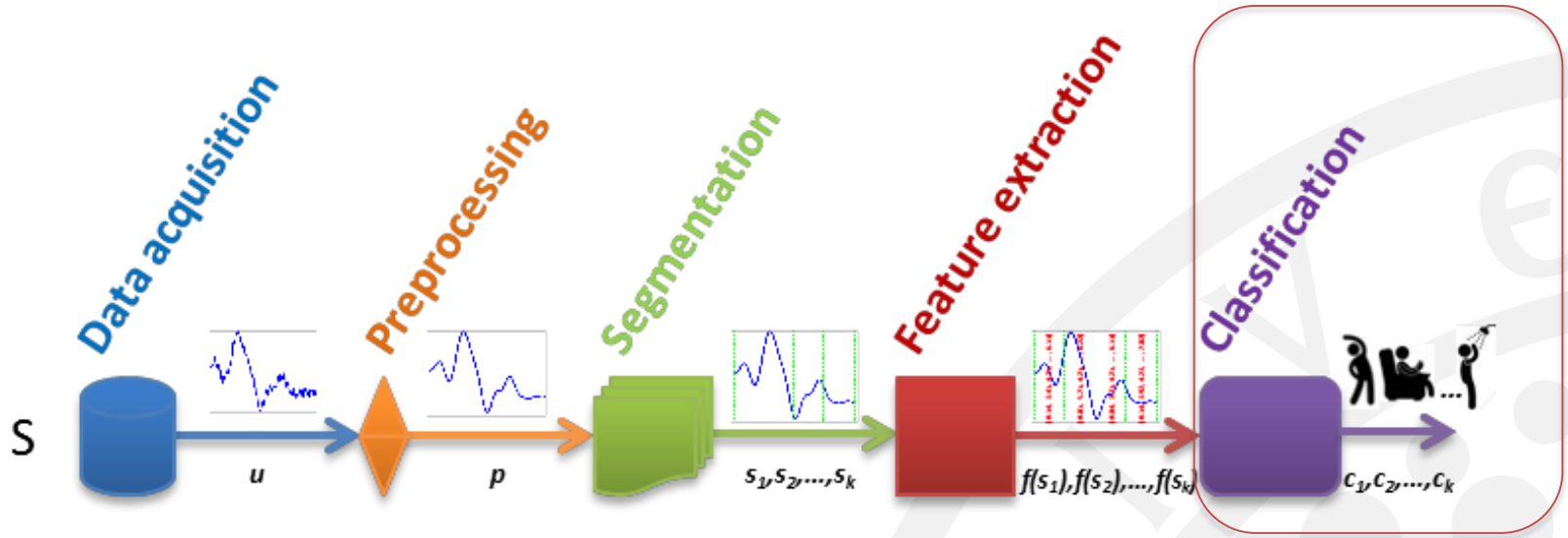
- Session 2: Processing data sensor. Feature matrix



S = data source (sensor) *u = raw/unprocessed data* *p = preprocessed data*
s_i = segment of data *$f(s_i)$ = feature vector* *c_i = class/label*

Tutorial Context

- Session 3: Build a classification model



S = data source (sensor)
s_i = segment of data

u = raw/unprocessed data
$f(s_i)$ = feature vector

p = preprocessed data
c_i = class/label

Classification context

- Classification
  - Data Mining technique
- The classification task
  - Build a classifier model to classify new objects.
  - **Given a dataset**
    - that contains a set examples with their classes
- A type of supervised learning
  - The real class of each example is used to build the classifier model
- Multiple classifiers
  - Set of rules
  - Decision tree
  - Neuronal network
  - Etc.

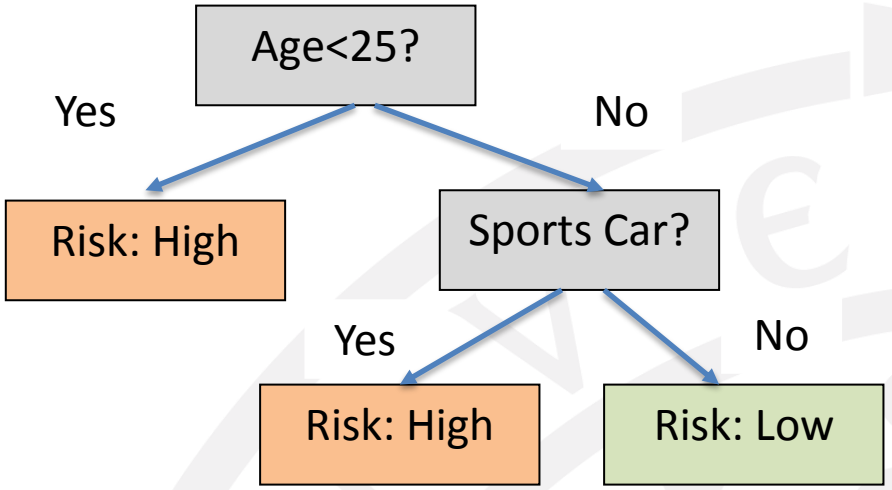
Example of classifier

Dataset

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Insurance Risk Evaluation

Decision Tree



Age	Car Type	Risk
34	Family	?

Unseen class

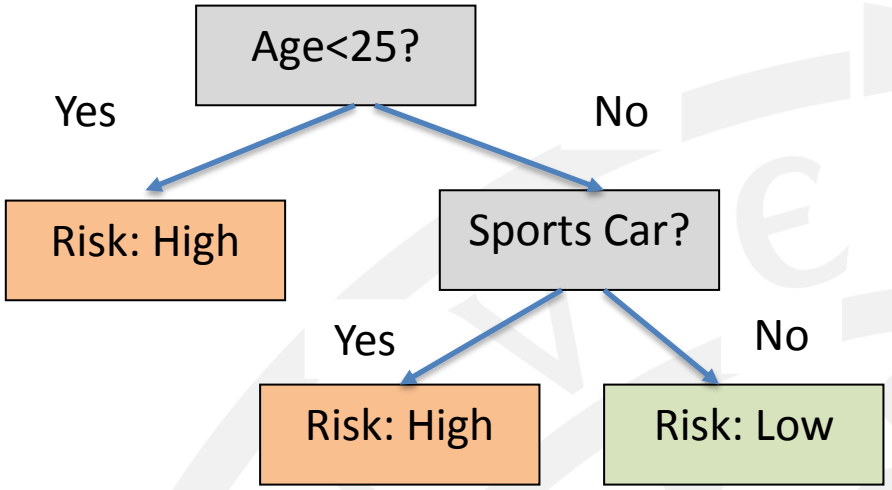
Example of classifier

Dataset

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Insurance Risk Evaluation

Decision Tree



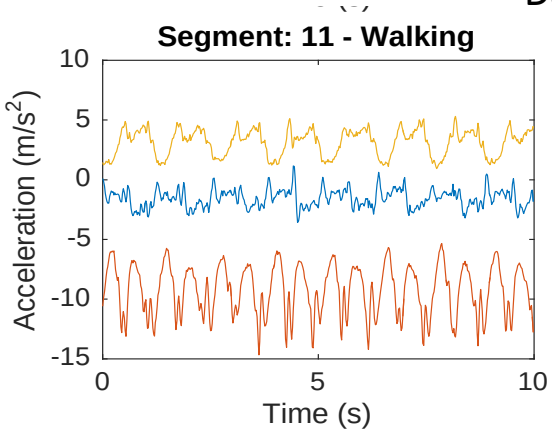
Age	Car Type	Risk
34	Family	Low

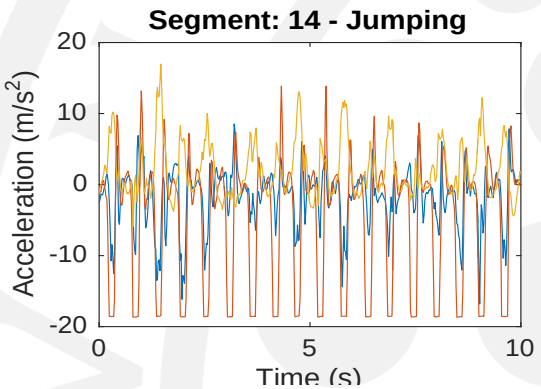
Activity Recognition

- In our case, feature matrix is the dataset

Mean X	St. Des. -X	Max X	Min X	…	Activity
-6.2041	-4.1083	1.9135	9.9164	…	Jumping
-6.1358	-3.854,	1.9593	9.797	…	Jumping
-4.0384	-8.1126	1.5507	1.7009	…	Walking
-5.8076	-7.0869	2.8533	11.415	…	Running

Dataset





How is the activity model evaluated?

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Dataset

Classification

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low

Training Set

Build the classifier model

Age	Car Type	Risk
32	Truck	? – Low
20	Family	? – High

Test Set

Compute the precision

Evaluation measures

- Different evaluation measures
  - Precision - Corrected classified instance percentage
    - The number of correct positive predictions divided by the total number of positive predictions.

Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

Precision= 7/15=0,46%

Evaluation measures

- Different evaluation measures
  - Precision - Corrected classified instance percentage
    - The number of correct positive predictions divided by the total number of positive predictions.

Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

		Predicted		
		Short	Medium	Tall
Real	Short	0	4	0
	Medium	0	5	3
	Tall	0	1	2

Confusion matrix

Evaluation measures

- Different evaluation measures
  - Precision - Corrected classified instance percentage
    - The number of correct positive predictions divided by the total number of positive predictions.

Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	**Short**	**Medium**
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	**Short**	**Medium**
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	**Short**	**Medium**
Dave	M	1.7m	**Short**	**Medium**
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

		Predicted		
		Short	Medium	Tall
Real	Short	0	**4**	0
	Medium	0	5	3
	Tall	0	1	2

Evaluation measure

- Different evaluation measure
  - Precision - Corrected classified instance percentage
    - The number of correct positive predictions divided by the total number of positive predictions.

Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	**Medium**	**Tall**
Martha	F	1.88m	**Medium**	**Tall**
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	**Medium**	**Tall**
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

		Predicted		
		Short	Medium	Tall
Real	Short	0	4	0
	Medium	0	5	**3**
	Tall	0	1	2

Evaluation measure

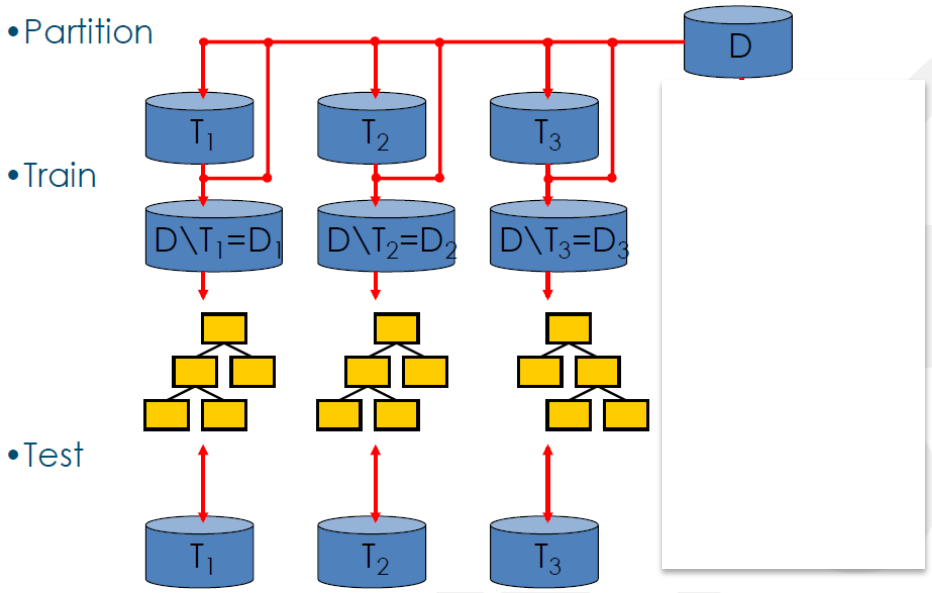
- Different evaluation measure
  - Precision - Corrected classified instance percentage
    - The number of correct positive predictions divided by the total number of positive predictions.

Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	**Medium**	**Medium**
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	**Medium**	**Medium**
Todd	M	1.95m	**Medium**	**Medium**
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

		Predicted		
		Short	Medium	Tall
Real	Short	0	4	0
	Medium	0	5	3
	Tall	0	1	2

Evaluation measure

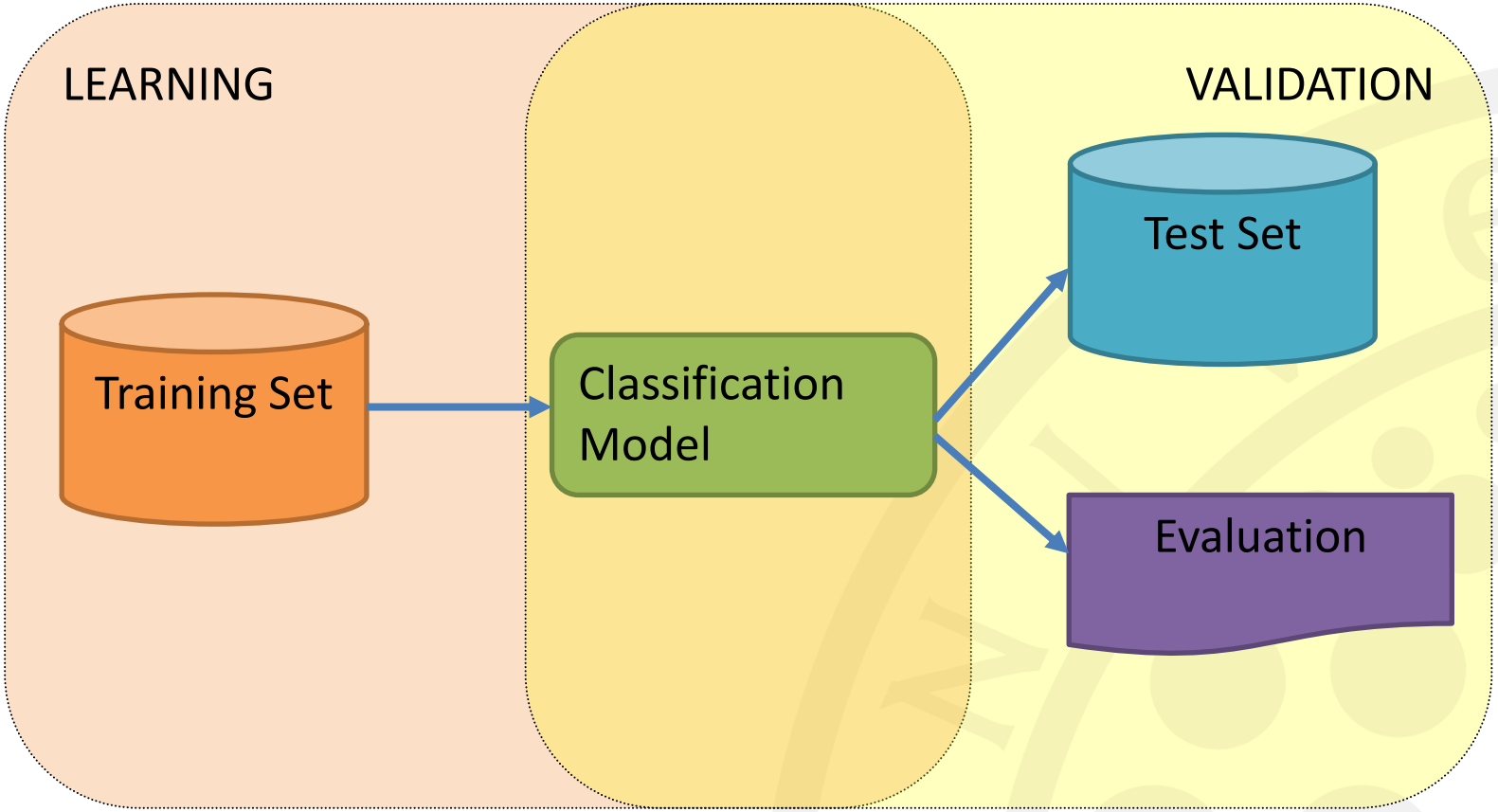
- Different evaluation methods
  - 10 Cross Validation
    - Dataset is split into 10 parts
    - Iterative process:
      - From 1st part to 9th part are used to train the classification model
      - The 10th part is used to assess the algorithm



3 Cross Validation

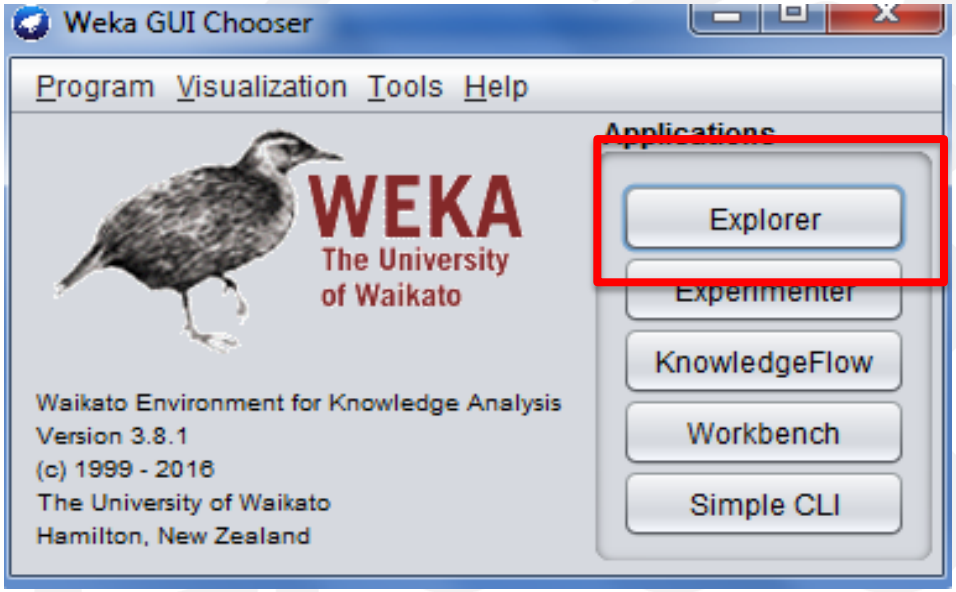
Summary

- Two stages



Weka

- Weka – Data mining software
  - Machine learning algorithms
  - Given a dataset
    - Build a classifier model
    - Evaluate the classifier model



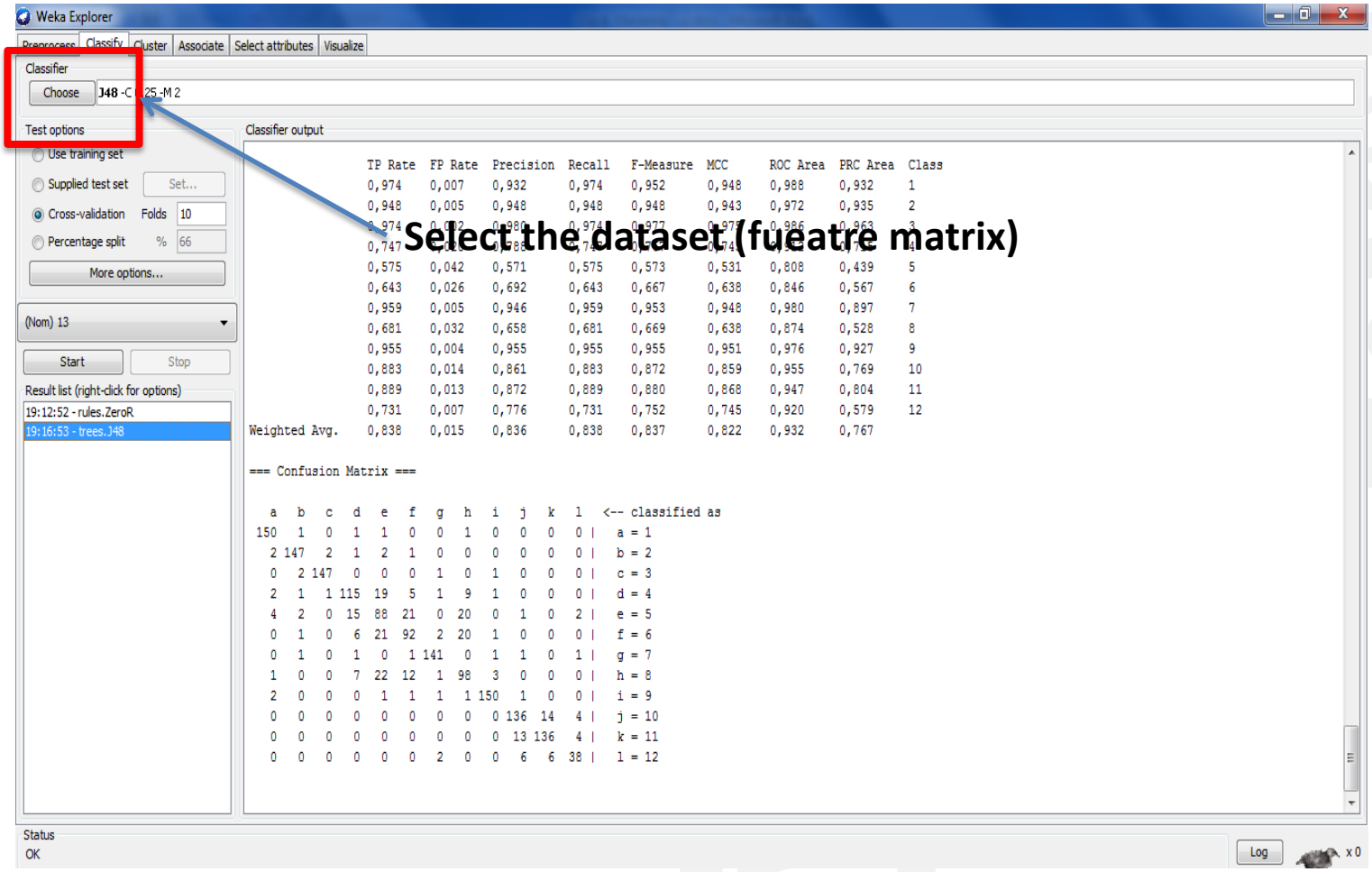
Weka

- Weka has a specific format: arff

```
@relation feature_matrix_FS3

@attribute 1 numeric
@attribute 2 numeric
@attribute 3 numeric
@attribute 4 numeric
@attribute 5 numeric
@attribute 6 numeric
@attribute 7 numeric
@attribute 8 numeric
@attribute 9 numeric
@attribute 10 numeric
@attribute 11 numeric
@attribute 12 numeric
@attribute 13 {1,2,3,4,5,6,7,8,9,10,11,12}

@data
-2.8735,-9.1443,1.747,0.10948,0.10133,0.10736,-2.5823,-8.8917,2.0668,-
3.178,-9.4535,1.4767,1

-2.8971,-9.1307,1.7251,0.13095,0.10742,0.10986,-2.5602,-8.842,2.046,-
3.4133,-9.4541,1.3498,1

.
.
.
```

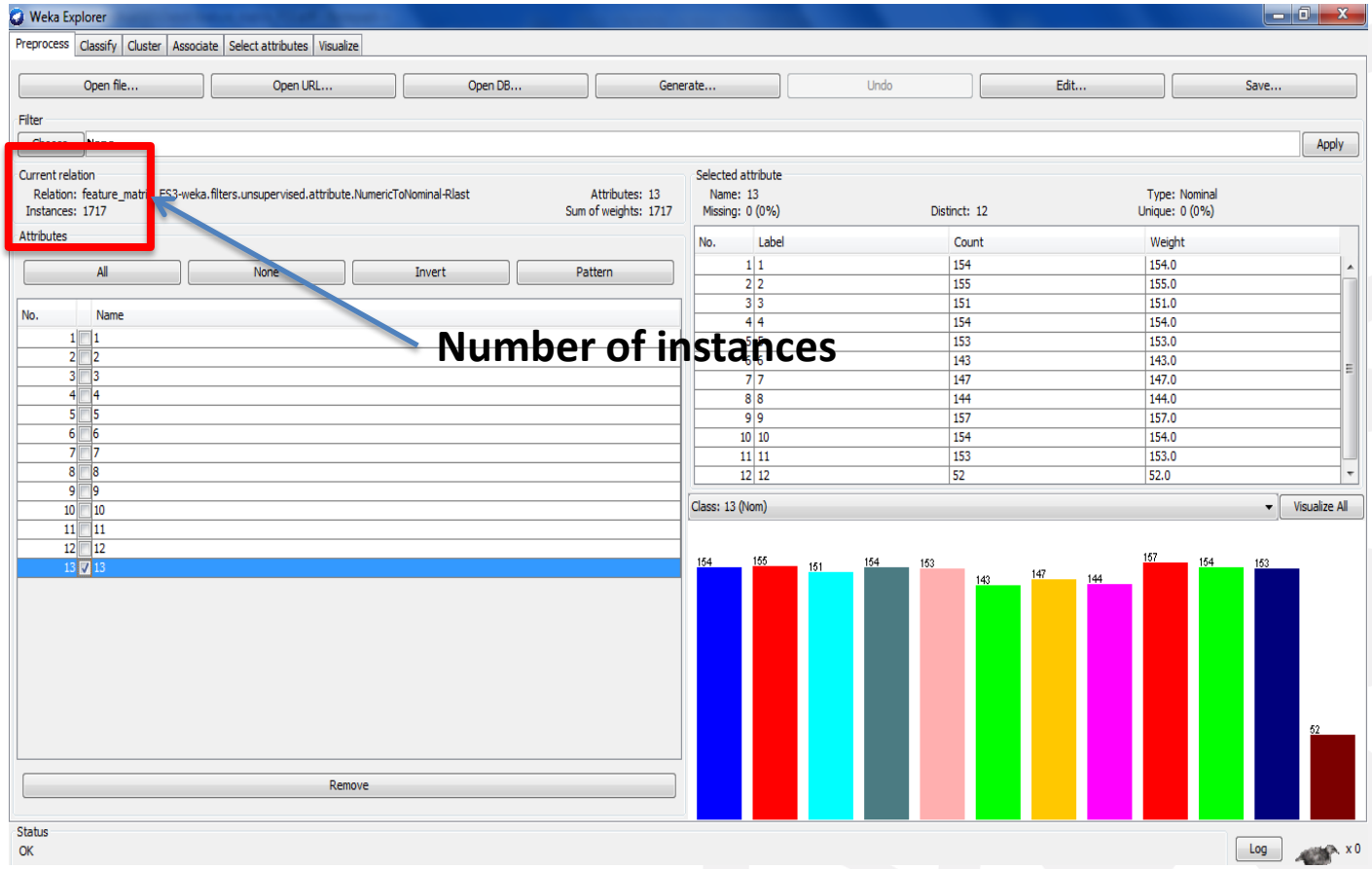
Weka

- Classification
  - Open the dataset (feature matrix)



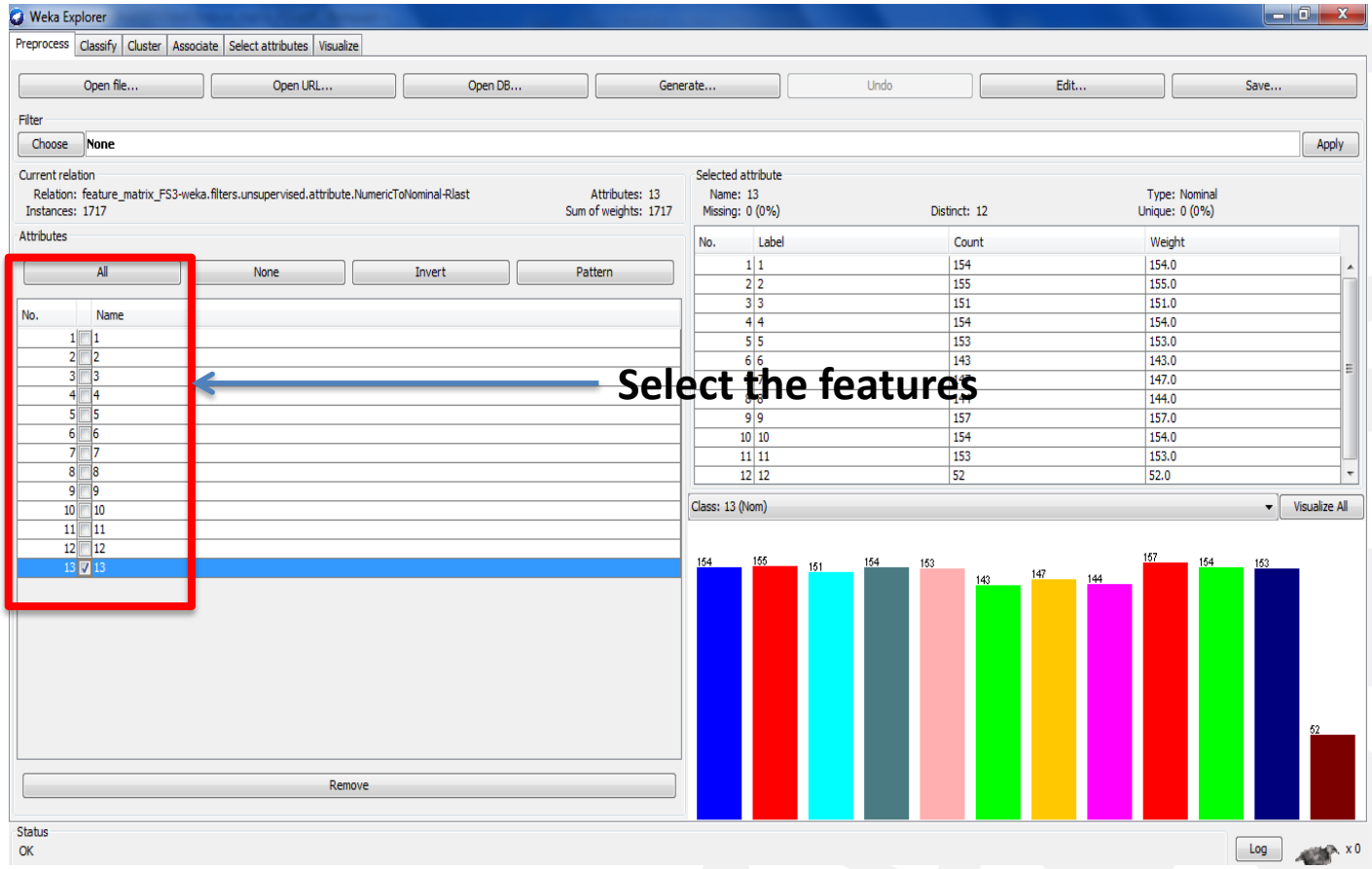
Weka

- Classification
  - Dataset (feature matrix)



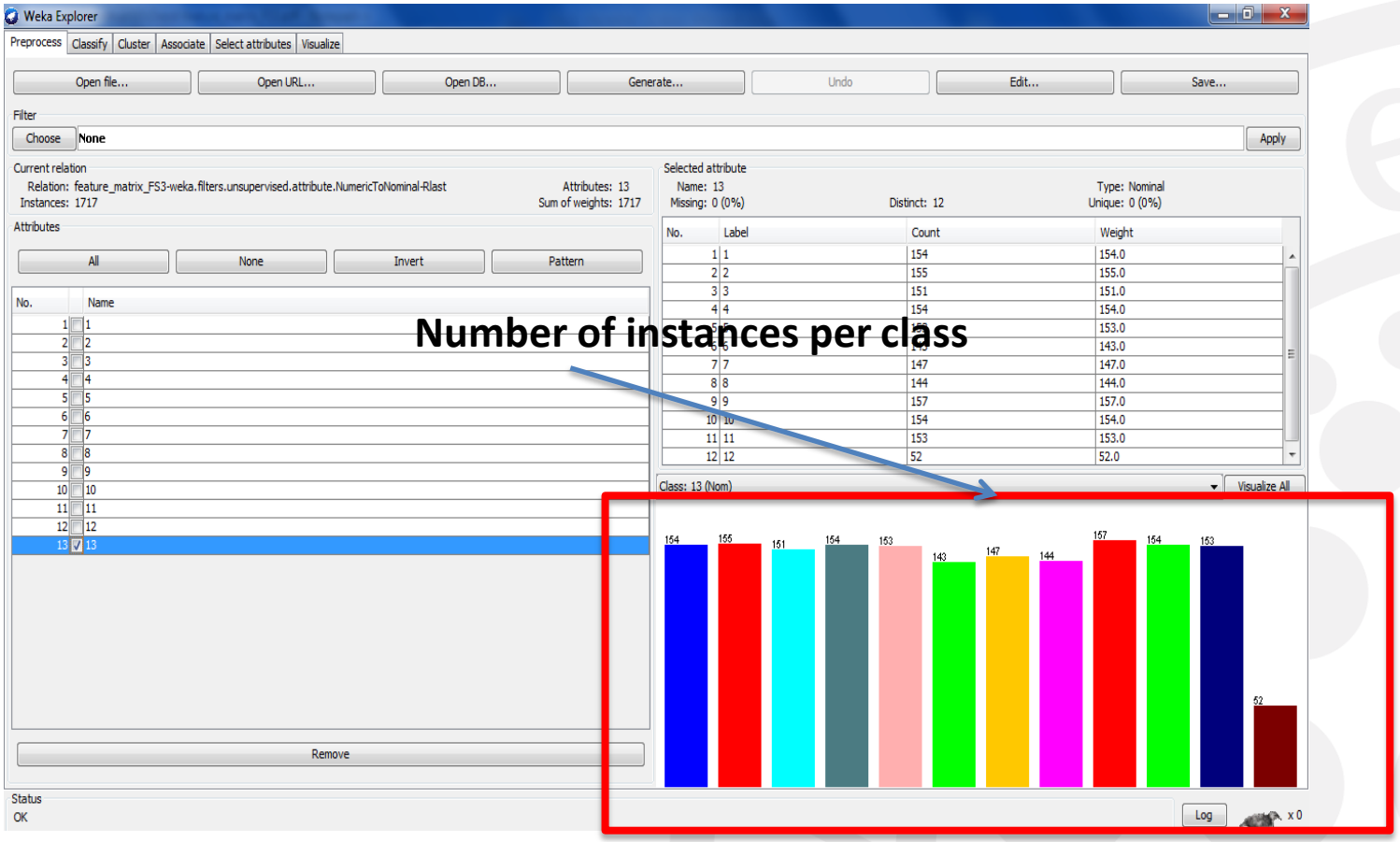
Weka

- Classification
  - Dataset (feature matrix)



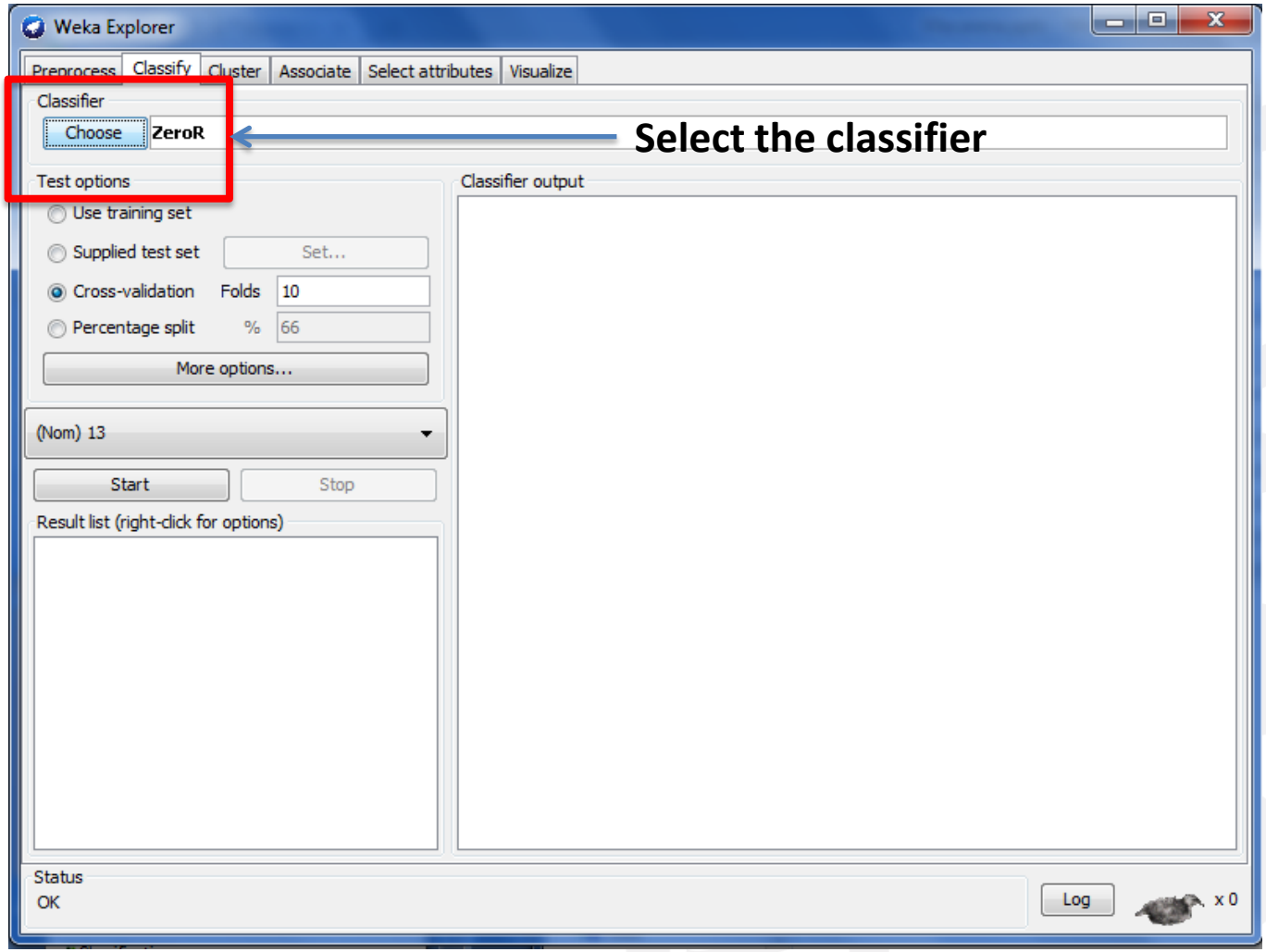
Weka

- Classification
  - Dataset (feature matrix)



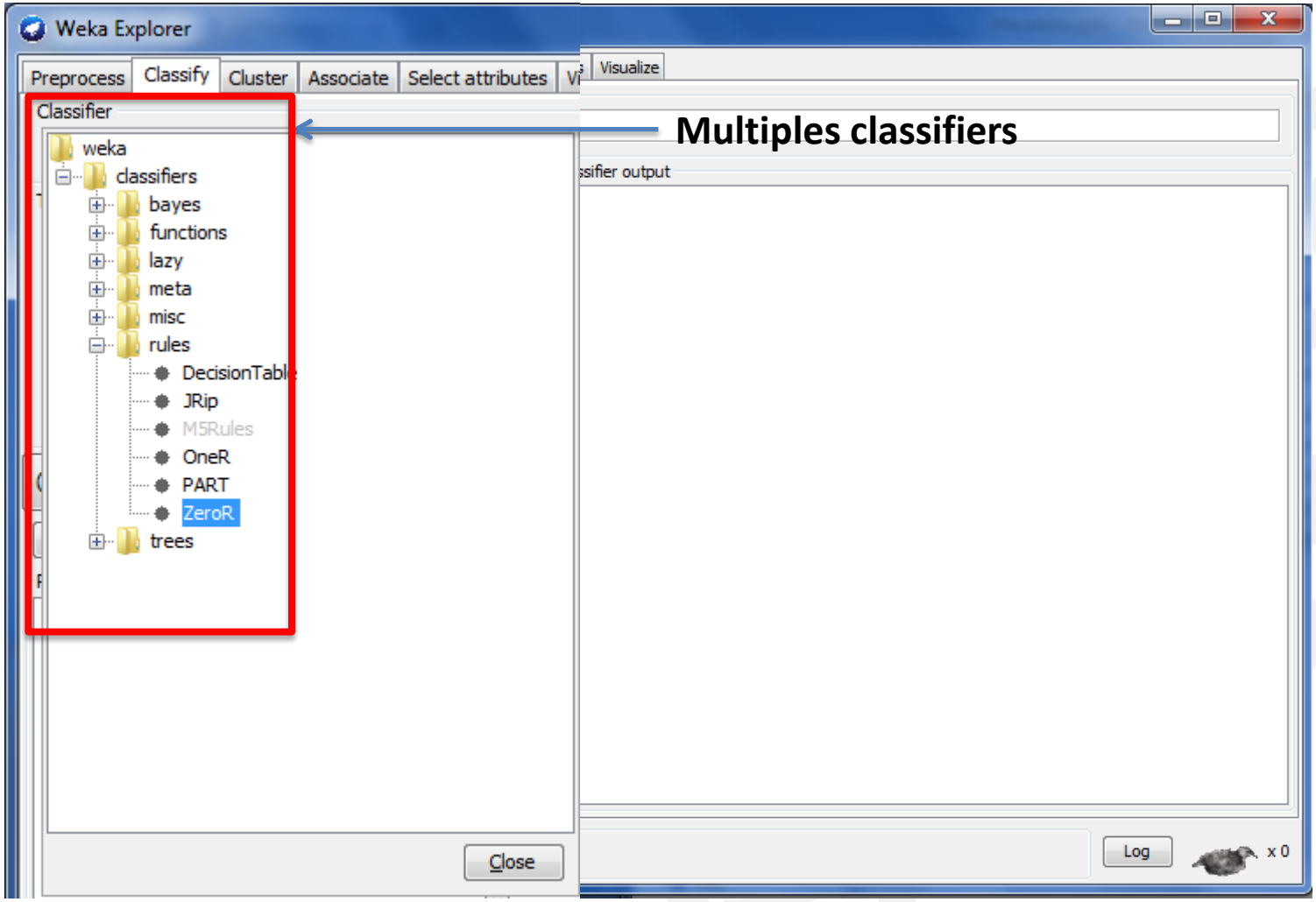
Weka

- Classification → Activity Model



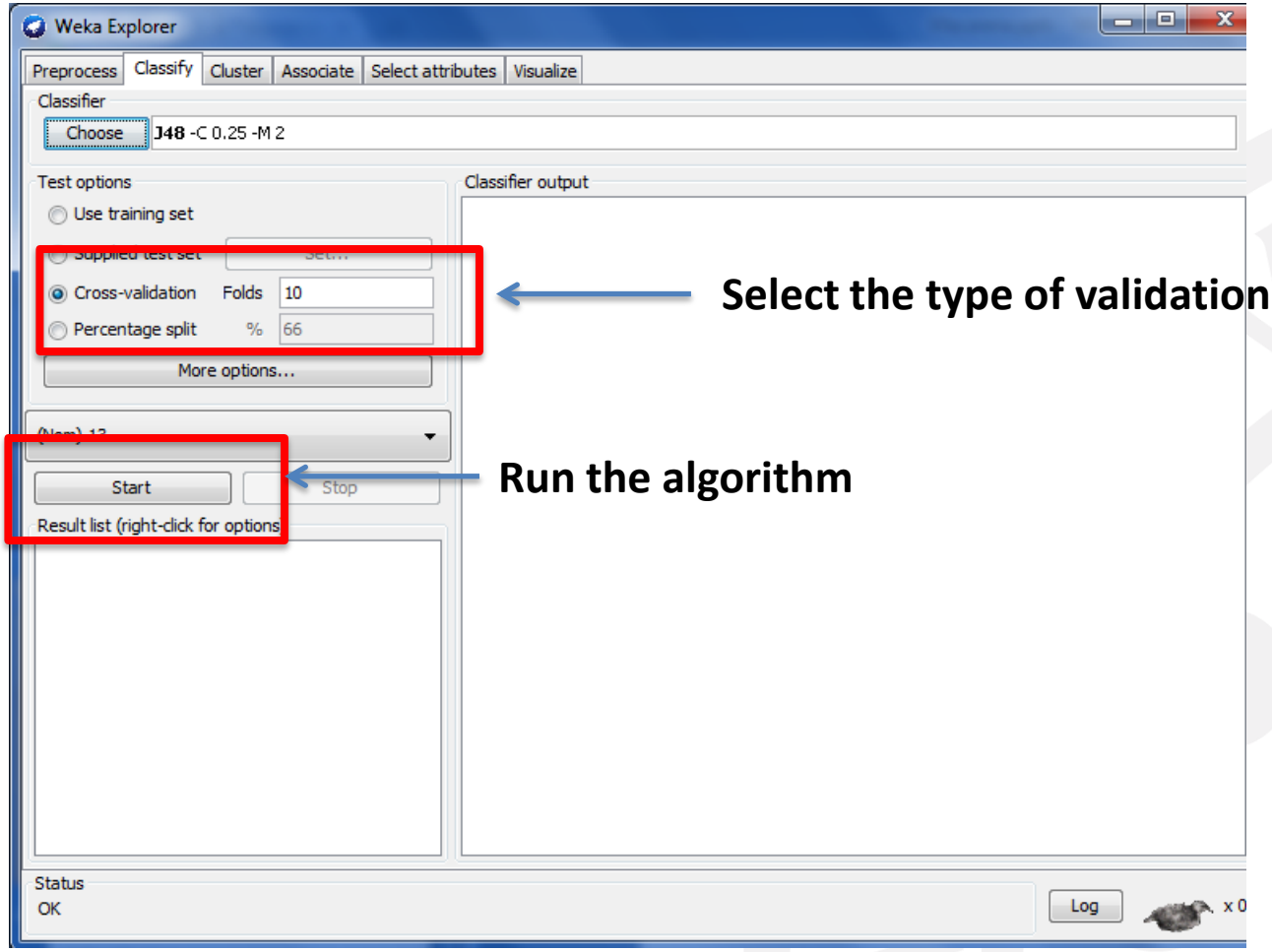
Weka

- Classification → Activity Model



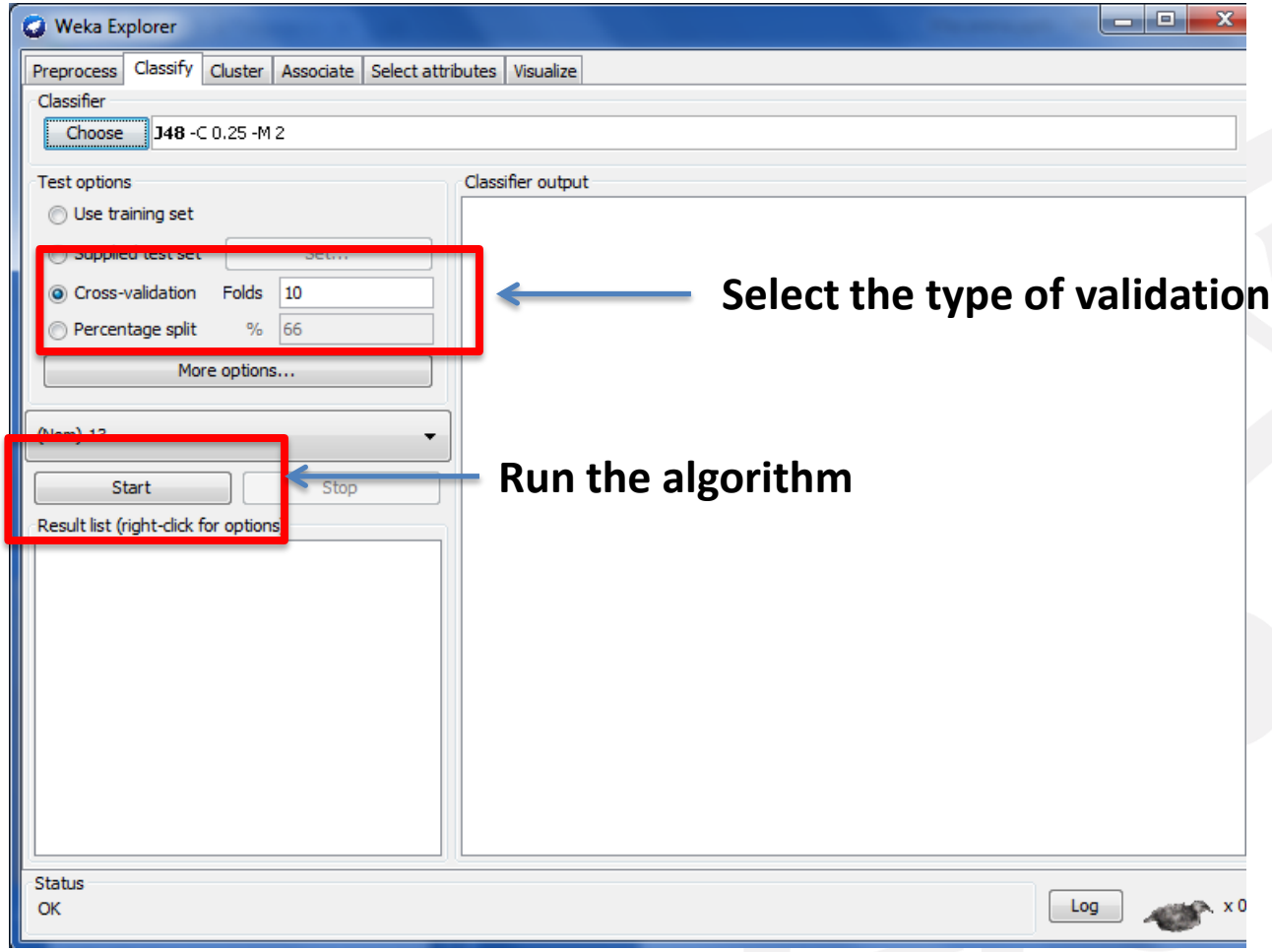
Weka

- Classification → Build the activity model



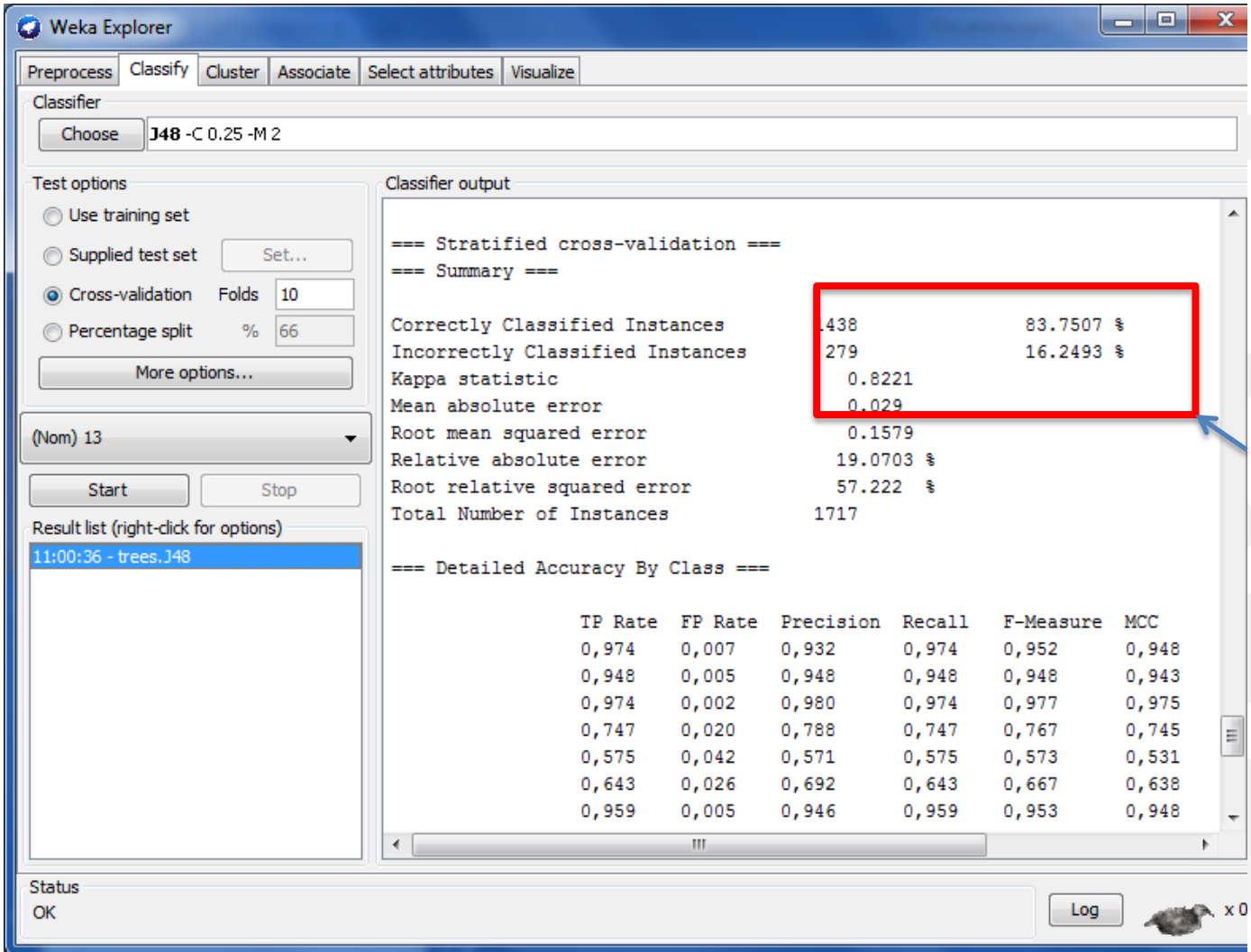
Weka

- Classification → Build the activity model



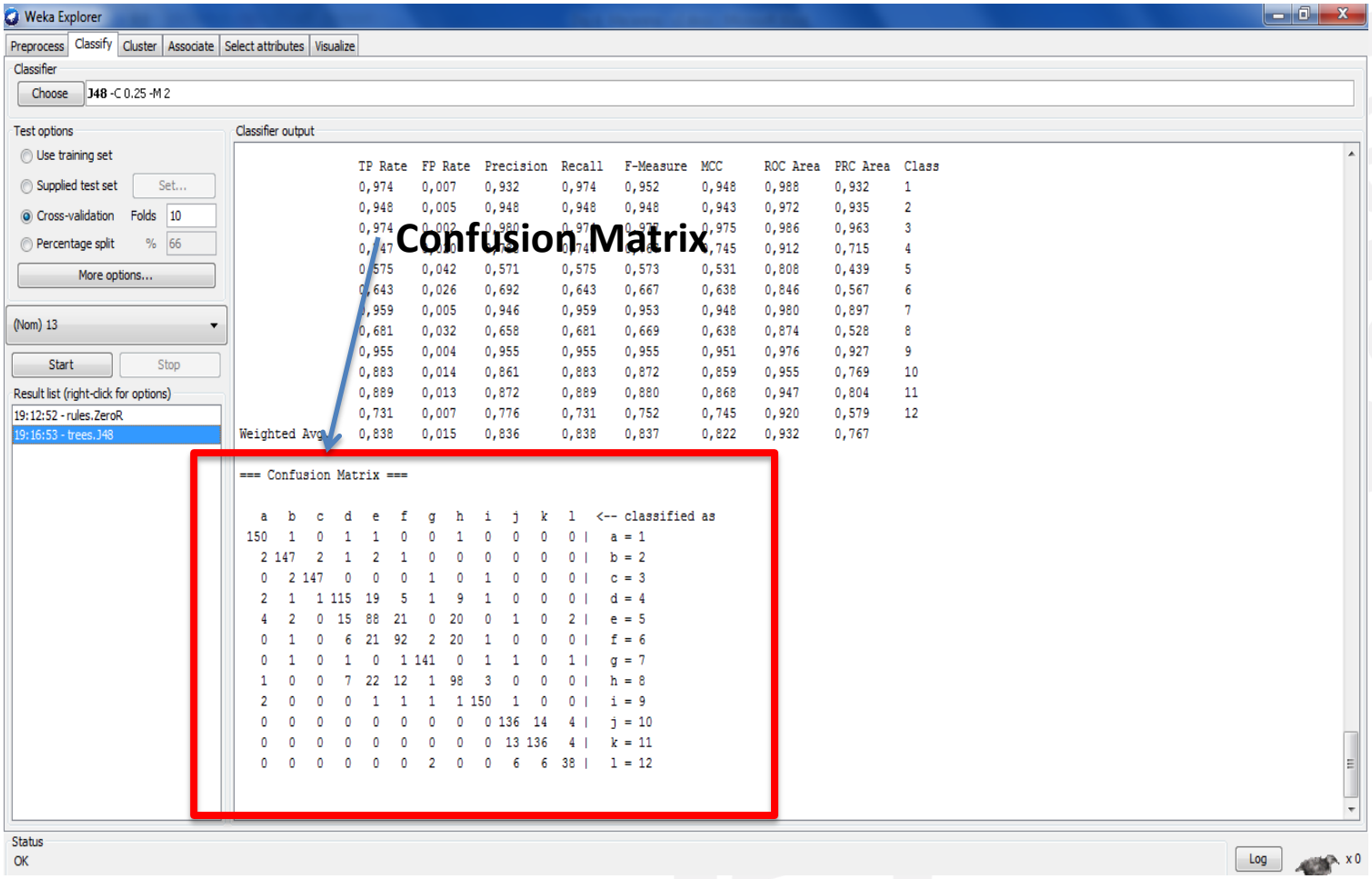
Weka

- Classification → Evaluation of the activity model



Weka

- Classification → Evaluation of the activity model



Weka

- Tutorial time!
  - Let's go!!
  - Material

goo.gl/JY7cN1