



Data mining and classification

Dra. D^a. Macarena Espinilla Estévez



Universidad de Jaén

- 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

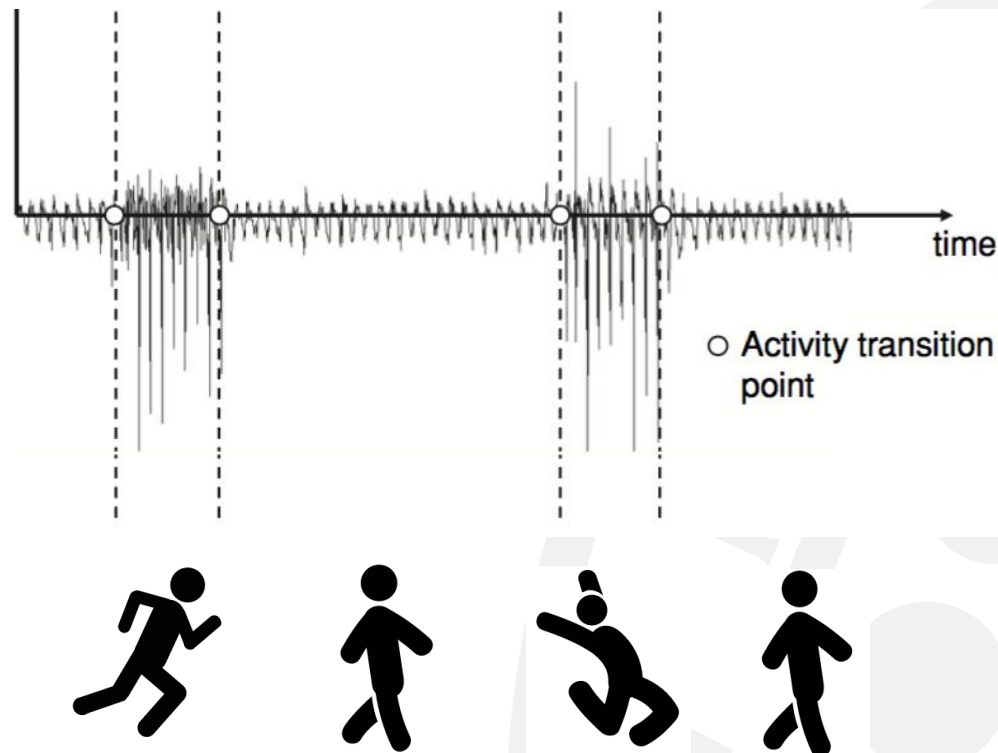
<http://sinbad2.ujaen.es>



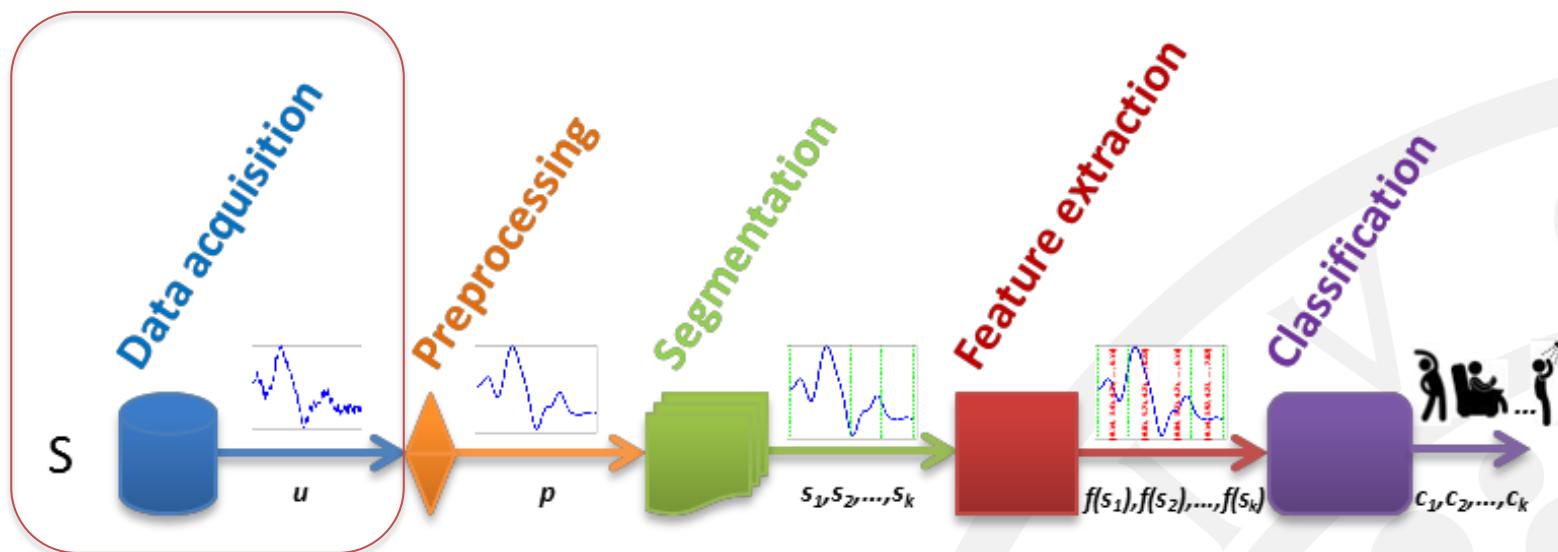
MacarenaUja

Tutorial Context

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



- Session 1: Data collection. Shimmer sensor

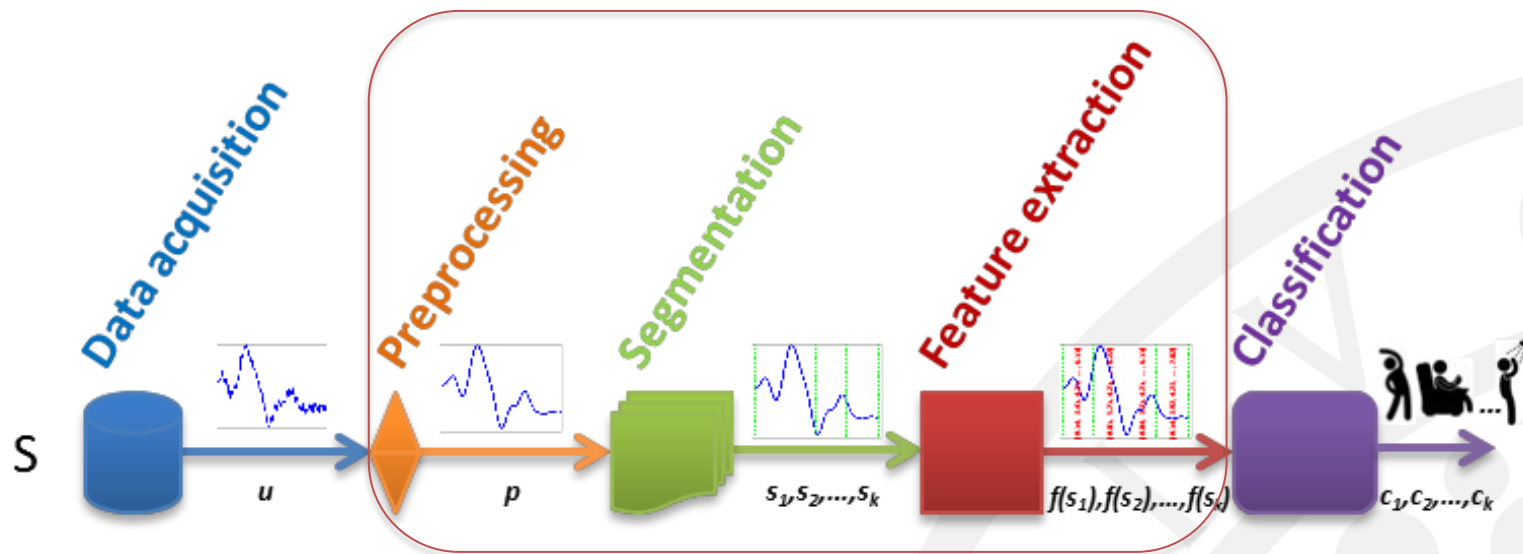


S = data source (sensor)
 s_i = segment of data

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

p = preprocessed data
 c_i = class/label

■ Session 2: Processing data sensor. Feature matrix

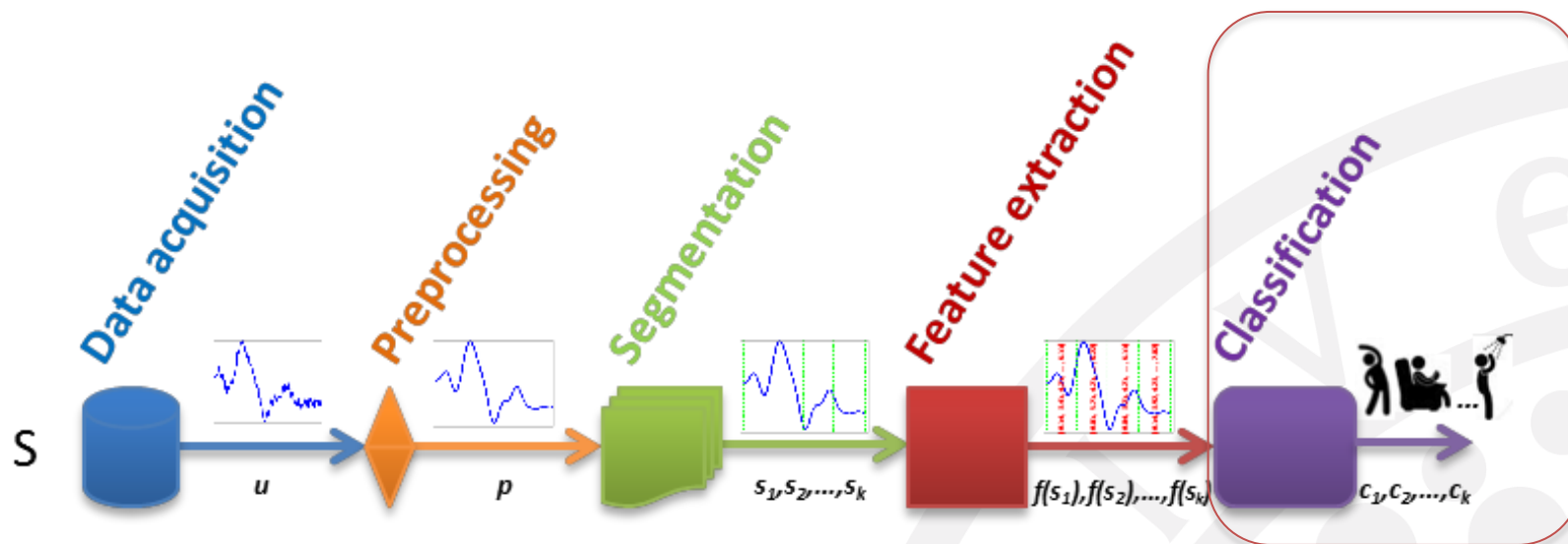


S = data source (sensor)
 s_i = segment of data

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

p = preprocessed data
 c_i = class/label

■ 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
 - 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.

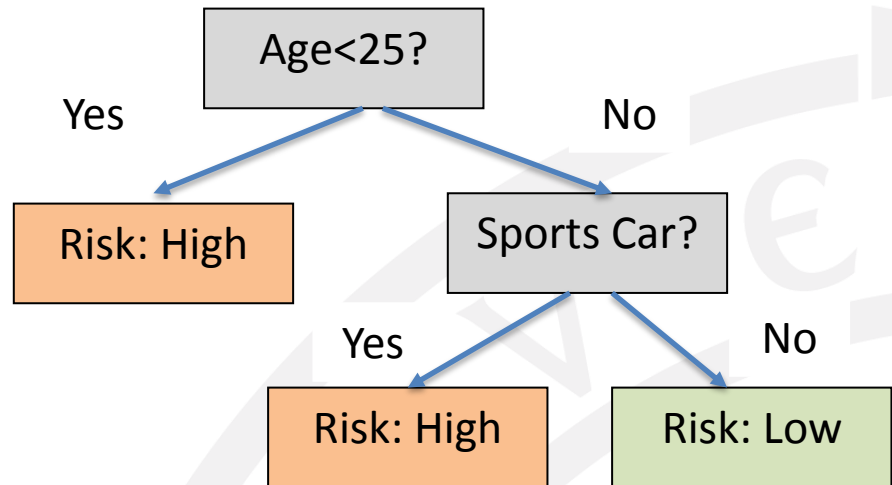
Example of classifier

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

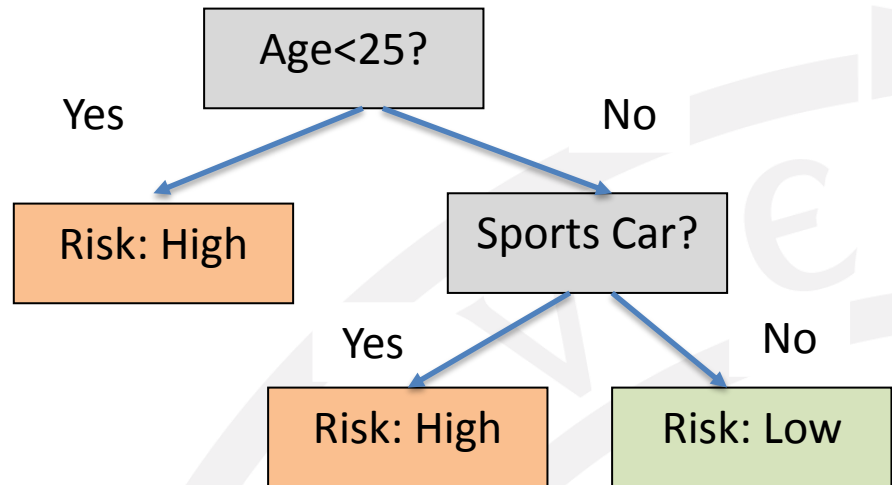
Example of classifier

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



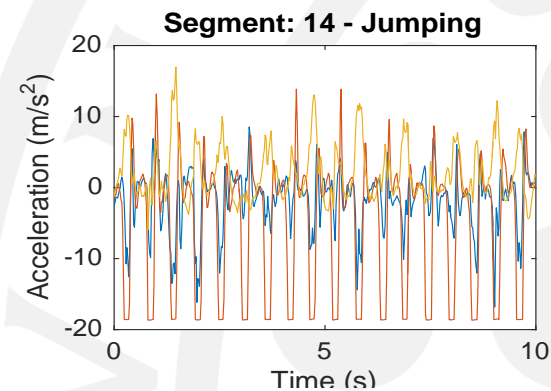
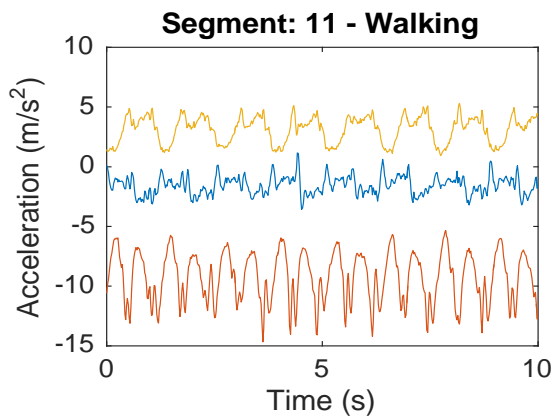
Age	Car Type	Risk
34	Family	Low

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

■ 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 = $7/15 = 0,46\%$

■ 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

■ 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

■ 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

■ Different evaluation measure

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

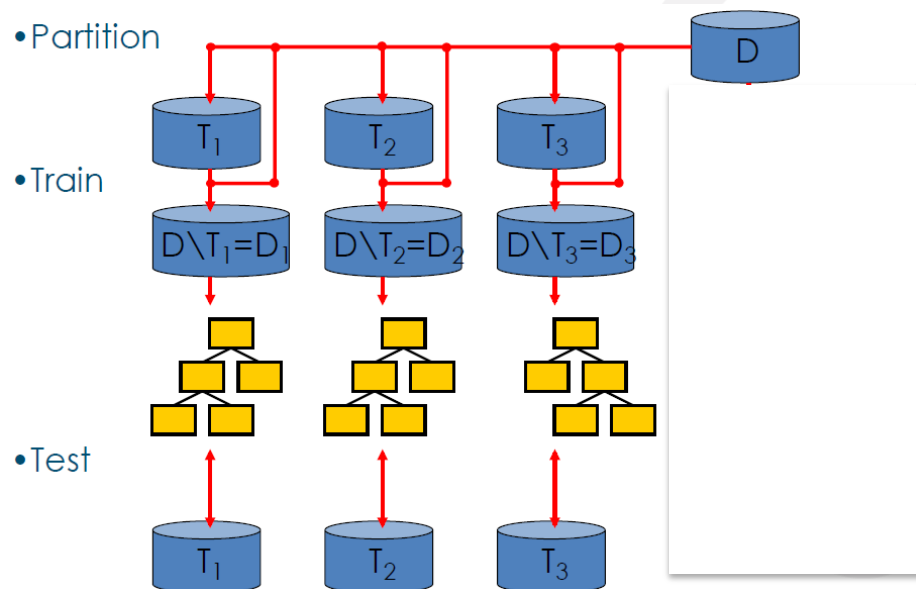
Name	Gender	Height	Output1 Real	Output2 Predicted
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	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

■ Different evaluation methods

– 10 Cross Validation

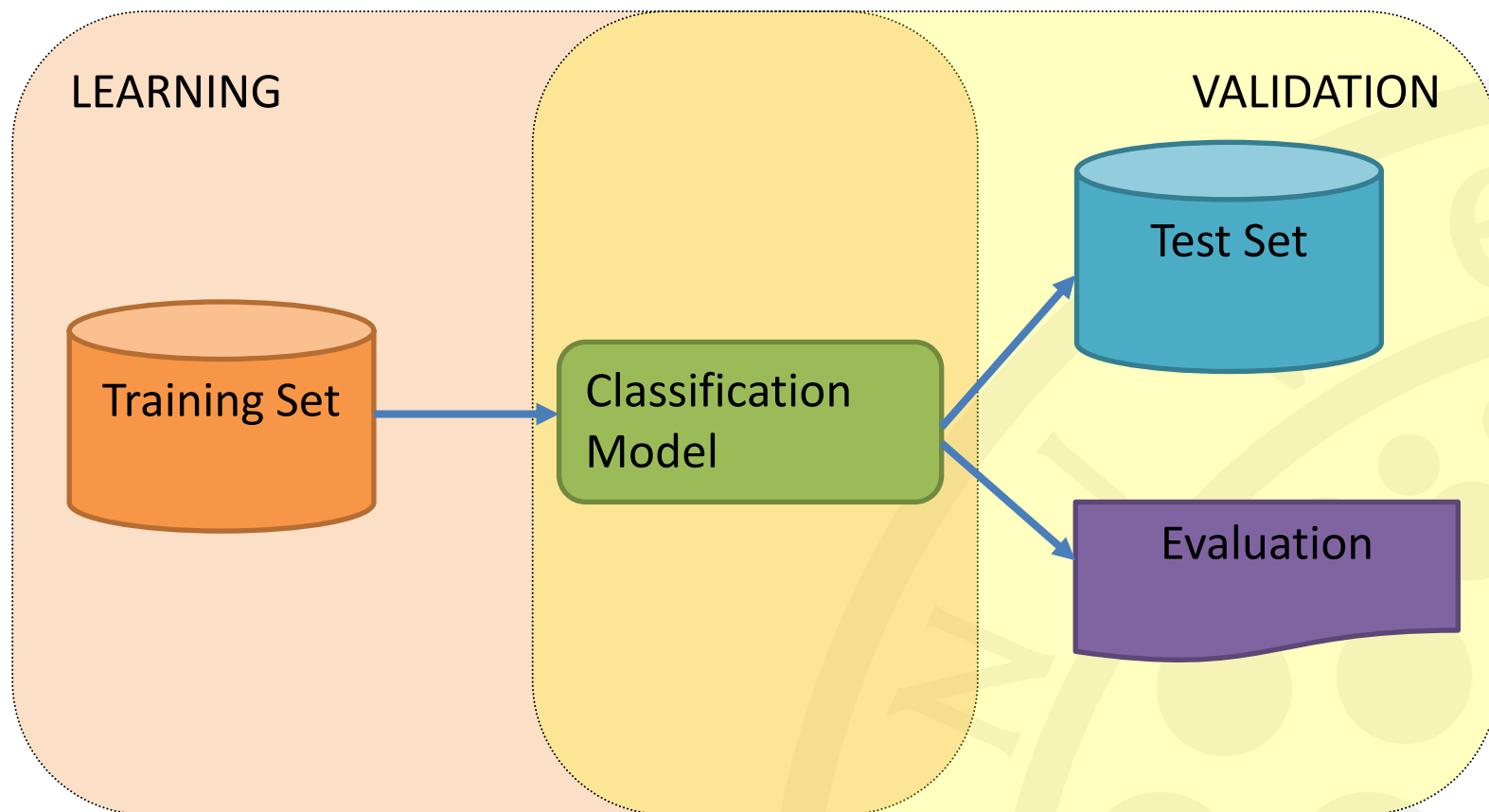
- Dataset is split into 10 parts
- Iterative process:
 - From 1st part to 9th part are used to train the classification model
 - The 10th part is used to assess the algorithm



3 Cross Validation

Summary

■ Two stages



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



- 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

-
-
-
```

■ Classification

- Open the dataset (feature matrix)

Select the dataset (feature matrix)

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,974	0,007	0,932	0,974	0,952	0,948	0,988	0,932	1
0,948	0,005	0,948	0,948	0,948	0,943	0,972	0,935	2
0,974	0,002	0,988	0,974	0,977	0,977	0,986	0,963	3
0,747	0,012	0,883	0,747	0,777	0,744	0,744	0,713	4
0,575	0,042	0,571	0,575	0,573	0,531	0,808	0,439	5
0,643	0,026	0,692	0,643	0,667	0,638	0,846	0,567	6
0,959	0,005	0,946	0,959	0,953	0,948	0,980	0,897	7
0,681	0,032	0,658	0,681	0,669	0,638	0,874	0,528	8
0,955	0,004	0,955	0,955	0,955	0,951	0,976	0,927	9
0,883	0,014	0,861	0,883	0,872	0,859	0,955	0,769	10
0,889	0,013	0,872	0,889	0,880	0,868	0,947	0,804	11
0,731	0,007	0,776	0,731	0,752	0,745	0,920	0,579	12
Weighted Avg.	0,838	0,015	0,836	0,838	0,837	0,822	0,767	


```
=== Confusion Matrix ===
 a b c d e f g h i j k l <-- classified as
150 1 0 1 1 0 0 1 0 0 0 0 | a = 1
 2 147 2 1 2 1 0 0 0 0 0 0 | b = 2
 0 2 147 0 0 0 1 0 1 0 0 0 | c = 3
 2 1 1 115 19 5 1 9 1 0 0 0 | d = 4
 4 2 0 15 88 21 0 20 0 1 0 2 | e = 5
 0 1 0 6 21 92 2 20 1 0 0 0 | f = 6
 0 1 0 1 0 1 141 0 1 1 0 1 | g = 7
 1 0 0 7 22 12 1 98 3 0 0 0 | h = 8
 2 0 0 0 1 1 1 1 150 1 0 0 | i = 9
 0 0 0 0 0 0 0 0 0 136 14 4 | j = 10
 0 0 0 0 0 0 0 0 0 13 136 4 | k = 11
 0 0 0 0 0 0 0 2 0 0 6 6 38 | l = 12
```

■ Classification

– Dataset (feature matrix)

The screenshot shows the Weka Explorer interface. The 'Current relation' section is highlighted with a red box and contains the following information:

- Relation: feature_matrix
- Instances: 1717
- Attributes: 13
- Sum of weights: 1717

A blue arrow points from the text 'Number of instances' to the 'Instances: 1717' value.

The 'Selected attribute' section shows a table with 13 attributes:

No.	Label	Count	Weight
1	1	154	154.0
2	2	155	155.0
3	3	151	151.0
4	4	154	154.0
5	5	153	153.0
6	6	143	143.0
7	7	147	147.0
8	8	144	144.0
9	9	157	157.0
10	10	154	154.0
11	11	153	153.0
12	12	52	52.0

Below the table is a bar chart showing the count for each attribute. The bars are colored and labeled with their respective counts: 154, 155, 151, 154, 153, 143, 147, 144, 157, 154, 153, 52.

■ Classification

– Dataset (feature matrix)

The screenshot shows the Weka Explorer interface. The 'Attributes' list on the left is highlighted with a red box, and a blue arrow points to it with the text 'Select the features'. The 'Selected attribute' table on the right shows the following data:

No.	Label	Count	Weight
1	1	154	154.0
2	2	155	155.0
3	3	151	151.0
4	4	154	154.0
5	5	153	153.0
6	6	143	143.0
7	7	147	147.0
8	8	144	144.0
9	9	157	157.0
10	10	154	154.0
11	11	153	153.0
12	12	52	52.0

Below the table is a bar chart showing the distribution of the 13 classes. The bars are colored and labeled with their respective counts: 154, 155, 151, 154, 153, 143, 147, 144, 157, 154, 153, 154, and 52.

■ Classification

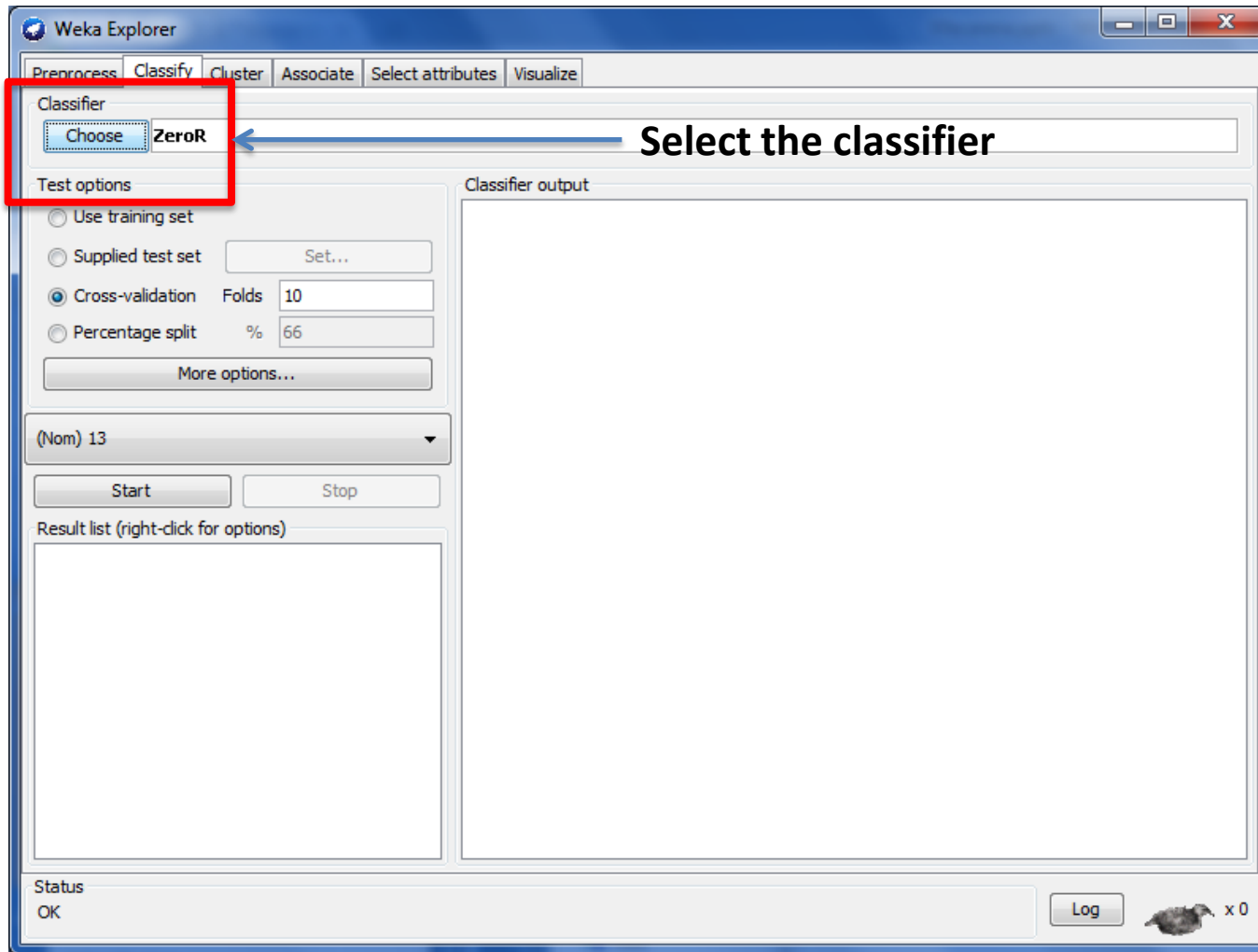
– Dataset (feature matrix)

The screenshot shows the Weka Explorer interface. The 'Selected attribute' table displays the following data:

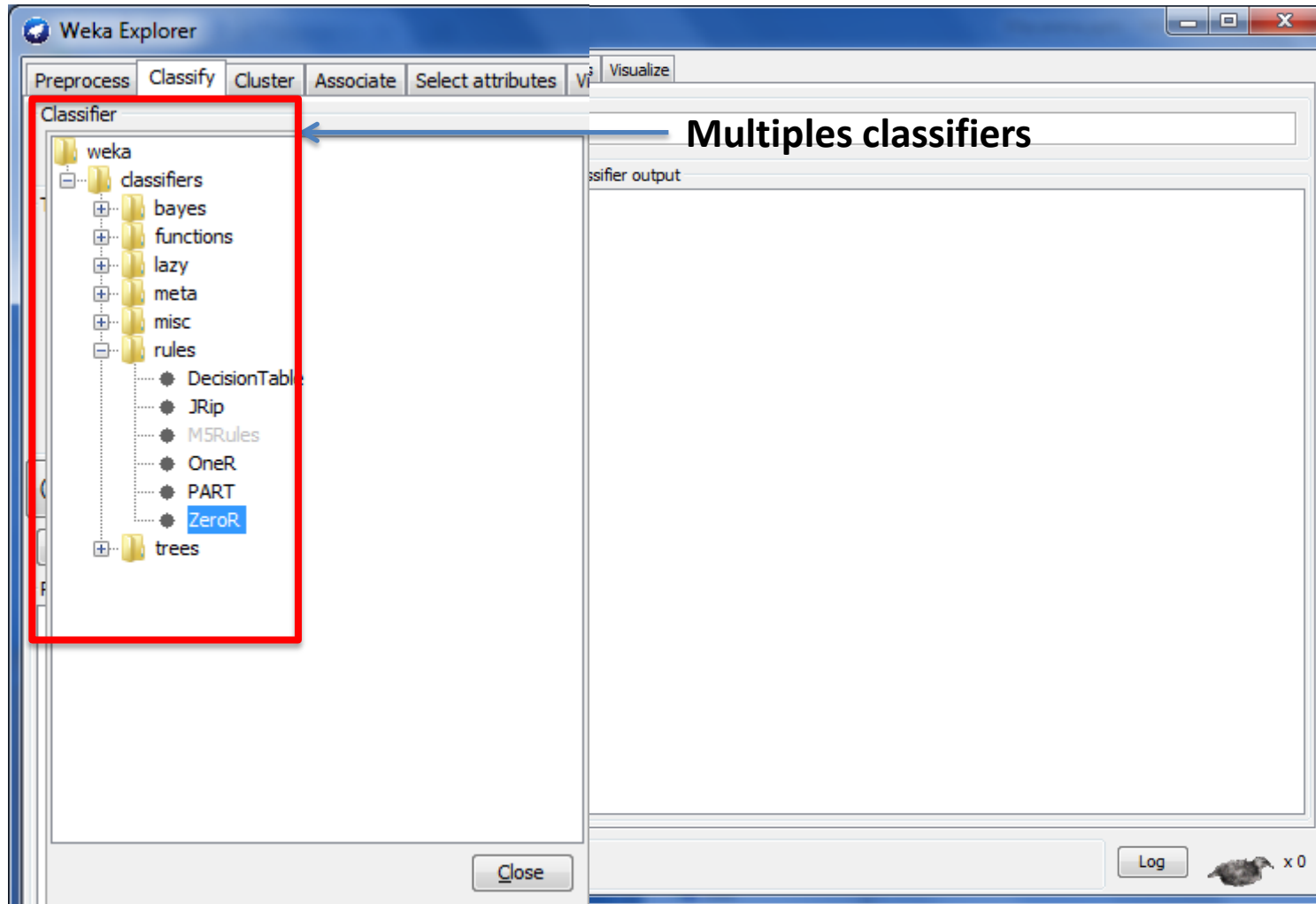
No.	Label	Count	Weight
1	1	154	154.0
2	2	155	155.0
3	3	151	151.0
4	4	154	154.0
5	5	153	153.0
6	6	143	143.0
7	7	147	147.0
8	8	144	144.0
9	9	157	157.0
10	10	154	154.0
11	11	153	153.0
12	12	52	52.0

The bar chart below the table visualizes the 'Count' column. The bars are colored and labeled with their respective values: 154 (blue), 155 (red), 151 (cyan), 154 (grey), 153 (pink), 143 (green), 147 (yellow), 144 (magenta), 157 (red), 154 (green), 153 (blue), and 52 (dark red). A red box highlights the bar chart, and an arrow points from the text 'Number of instances per class' to it.

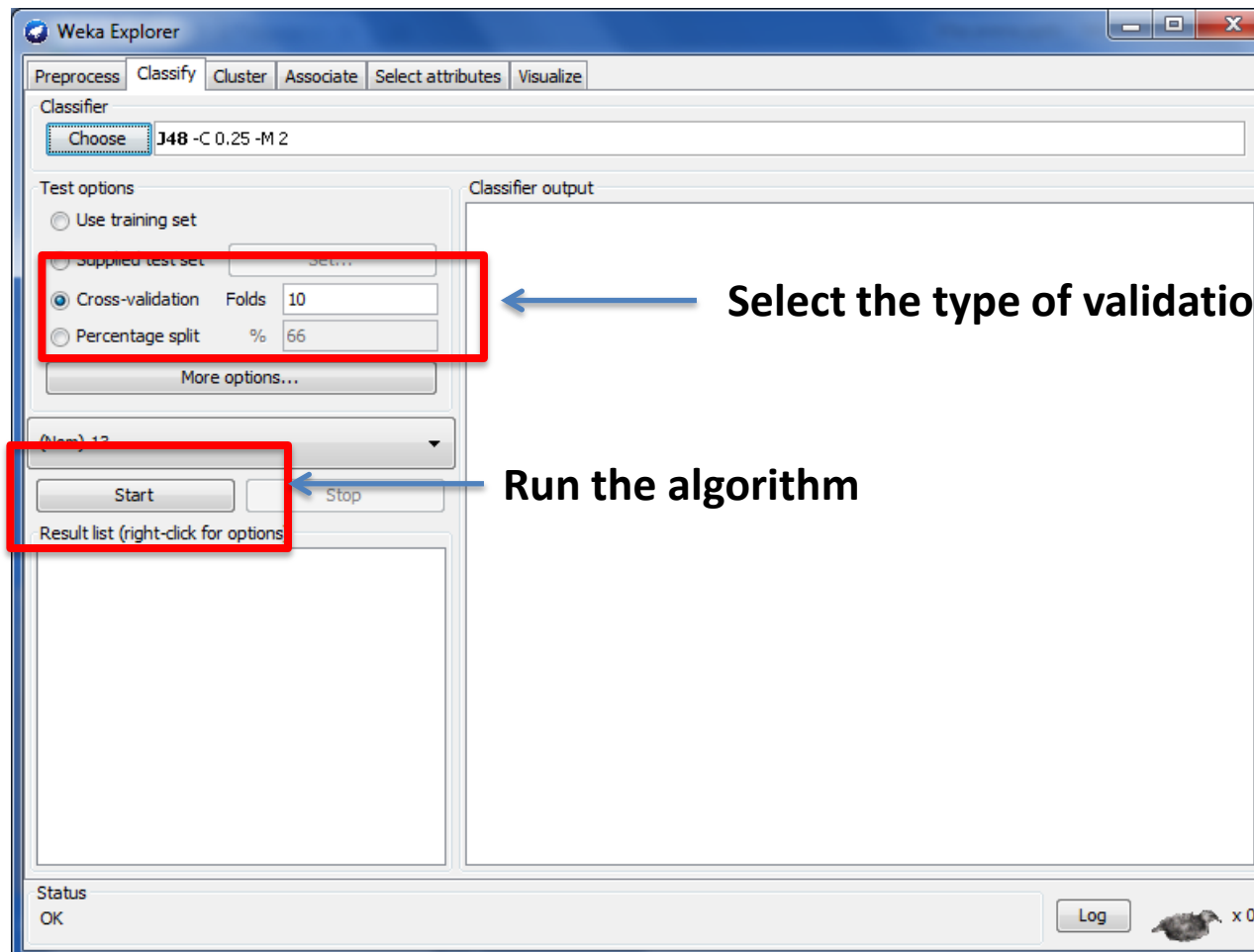
■ Classification → Activity Model



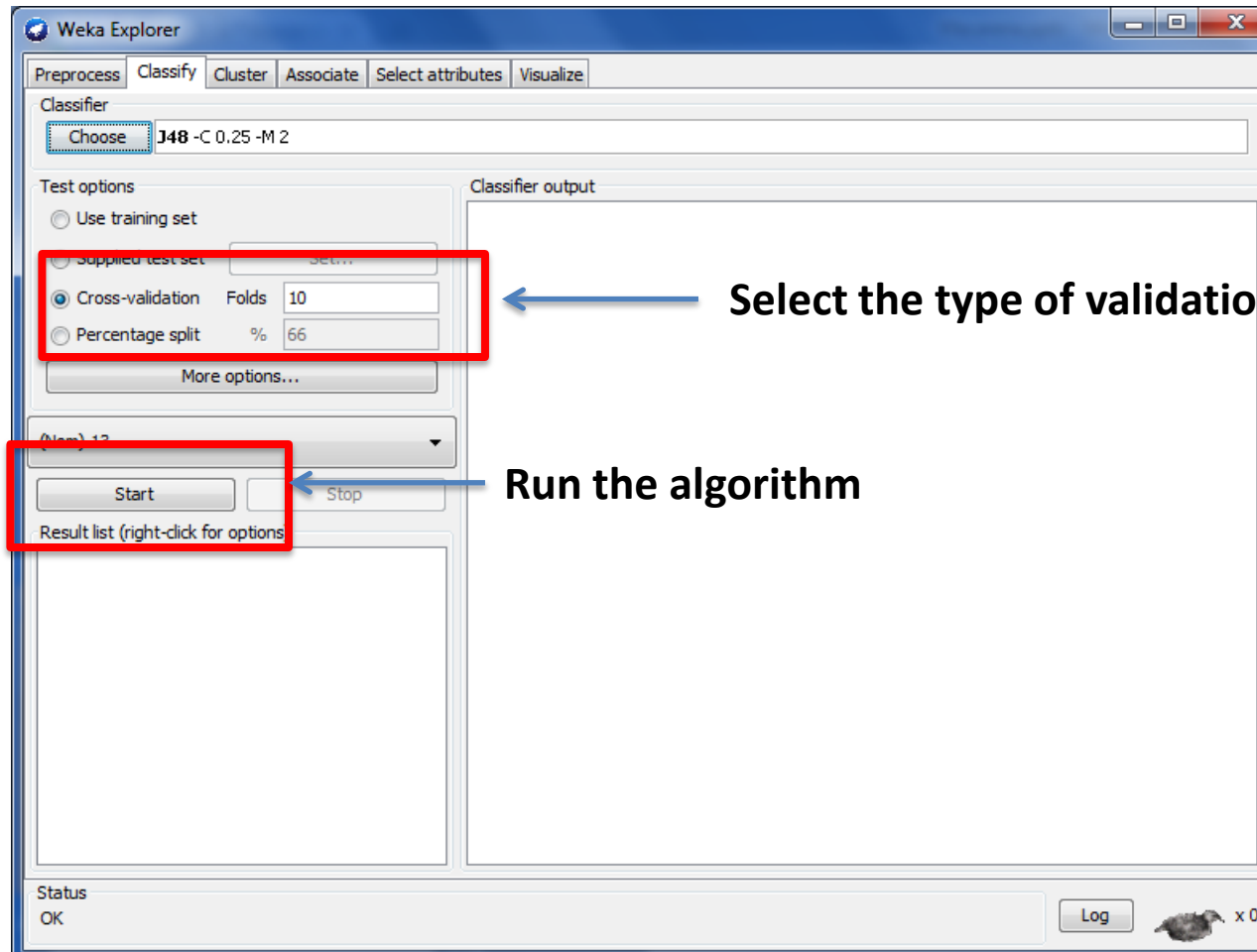
■ Classification → Activity Model



- Classification → Build the activity model



- Classification → Build the activity model



Classification → Evaluation of the activity model

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) 13

Start Stop

Result list (right-click for options)

11:00:36 - trees.J48

Classifier output:

```

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      438           83.7507 %
Incorrectly Classified Instances    279           16.2493 %
Kappa statistic                    0.8221
Mean absolute error                 0.029
Root mean squared error             0.1579
Relative absolute error             19.0703 %
Root relative squared error         57.222 %
Total Number of Instances          1717

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
	0,974	0,007	0,932	0,974	0,952	0,948
	0,948	0,005	0,948	0,948	0,948	0,943
	0,974	0,002	0,980	0,974	0,977	0,975
	0,747	0,020	0,788	0,747	0,767	0,745
	0,575	0,042	0,571	0,575	0,573	0,531
	0,643	0,026	0,692	0,643	0,667	0,638
	0,959	0,005	0,946	0,959	0,953	0,948

Status: OK

Log x 0

Precision

Classification → Evaluation of the activity model

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier used is 'J48 - C 0.25 - M 2'. The 'Test options' are set to 'Cross-validation' with 10 folds and a 66% split. The 'Classifier output' table displays performance metrics for 12 classes. A blue arrow points from the 'Confusion Matrix' text to the 'Weighted Avg' row in the table. Below the table, a red box highlights the 'Confusion Matrix' output, which is a table with 12 columns (a-l) and 12 rows (a-l), showing the number of instances classified correctly and incorrectly for each class.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,974	0,007	0,932	0,974	0,952	0,948	0,988	0,932	1	
0,948	0,005	0,948	0,948	0,948	0,943	0,972	0,935	2	
0,974	0,002	0,980	0,974	0,977	0,975	0,986	0,963	3	
0,47	0,000	0,000	0,000	0,000	0,000	0,000	0,000	4	
0,575	0,042	0,571	0,575	0,573	0,531	0,808	0,439	5	
0,643	0,026	0,692	0,643	0,667	0,638	0,846	0,567	6	
0,959	0,005	0,946	0,959	0,953	0,948	0,980	0,897	7	
0,681	0,032	0,658	0,681	0,669	0,638	0,874	0,528	8	
0,955	0,004	0,955	0,955	0,955	0,951	0,976	0,927	9	
0,883	0,014	0,861	0,883	0,872	0,859	0,955	0,769	10	
0,889	0,013	0,872	0,889	0,880	0,868	0,947	0,804	11	
0,731	0,007	0,776	0,731	0,752	0,745	0,920	0,579	12	
Weighted Avg	0,838	0,015	0,836	0,838	0,837	0,822	0,932	0,767	

```

=== Confusion Matrix ===
 a  b  c  d  e  f  g  h  i  j  k  l  <-- classified as
150 1  0  1  1  0  0  1  0  0  0  0  | a = 1
 2 147 2  1  2  1  0  0  0  0  0  0  | b = 2
 0  2 147 0  0  0  1  0  1  0  0  0  | c = 3
 2  1  1 115 19  5  1  9  1  0  0  0  | d = 4
 4  2  0  15  88 21  0 20  0  1  0  2  | e = 5
 0  1  0  6  21  92  2 20  1  0  0  0  | f = 6
 0  1  0  1  0  1 141  0  1  1  0  1  | g = 7
 1  0  0  7  22  12  1  98  3  0  0  0  | h = 8
 2  0  0  0  1  1  1  1 150  1  0  0  | i = 9
 0  0  0  0  0  0  0  0  0 136 14  4  | j = 10
 0  0  0  0  0  0  0  0  0 13 136  4  | k = 11
 0  0  0  0  0  0  0  2  0  0  6  6 38  | l = 12
    
```

■ Tutorial time!

- Let's go!!
- Material

goo.gl/JY7cN1

