



# **P**HÂN CỤM DỮ LIỆU

---

Giảng viên: Nguyễn Tu Trung  
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2021

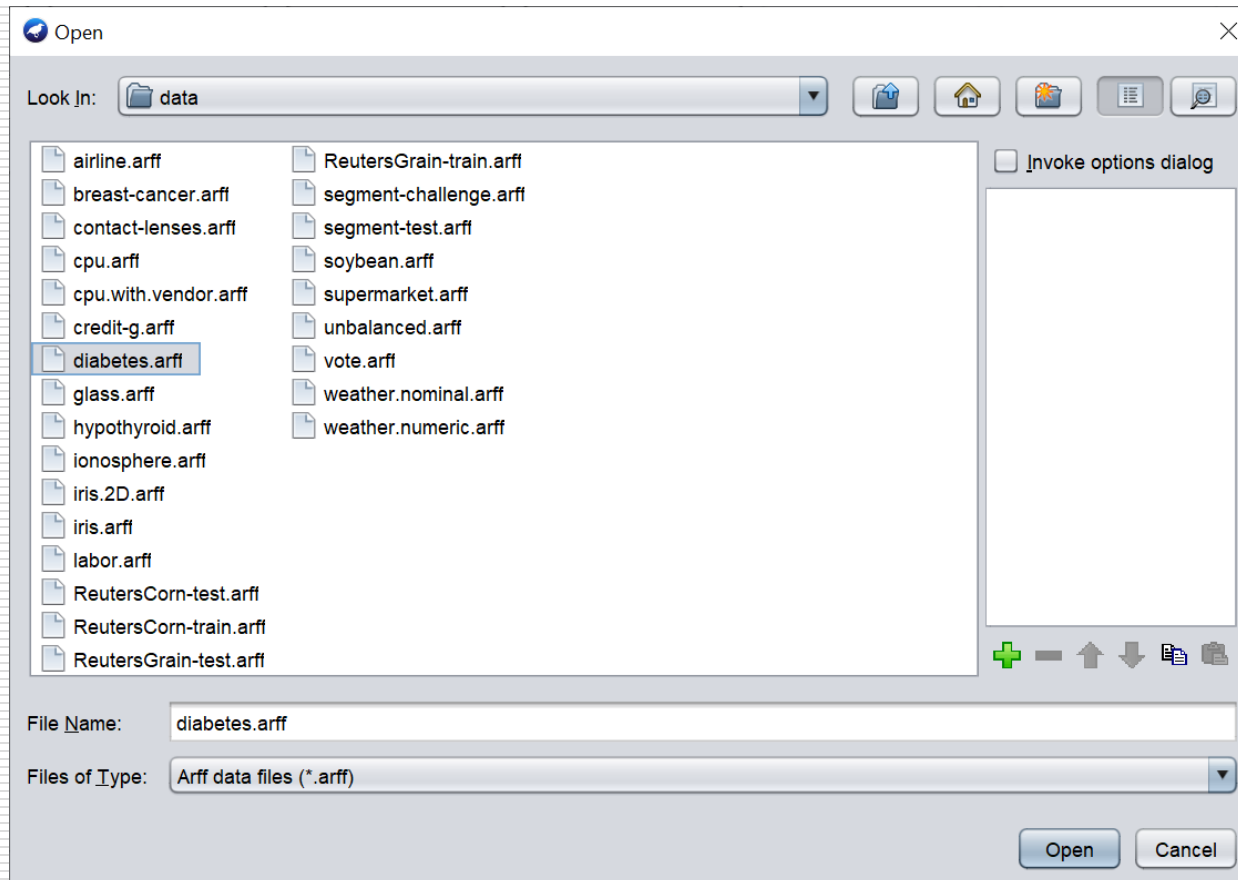
# Nội dung

---

- ❖ Mở file diabetes.arff
- ❖ Loại bỏ thuộc tính class (phân lớp)
- ❖ Phân cụm k-Means
- ❖ So sánh kết quả phân cụm và trường class
- ❖ Phân cụm với thuật toán EM

# Mở file diabetes.arff

- ❖ Trong Tab Preprocessing, chọn Open File
- ❖ Chọn file diabetes.arff



# Loại bỏ thuộc tính class (phân lớp)

- ❖ Tick chọn các trường class và nhấn Remove để loại bỏ:

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active. In the 'Attributes' list, the 'class' attribute is selected (checked). The 'Selected attribute' summary shows the 'class' attribute with 2 distinct values: 'tested\_negative' (500 instances) and 'tested\_positive' (268 instances). A bar chart at the bottom right visualizes these counts with a blue bar for 500 and a red bar for 268.

**Current relation**  
Relation: pima\_diabetes  
Instances: 768  
Attributes: 9  
Sum of weights: 768

**Attributes**

No.	Name
1	<input type="checkbox"/> preg
2	<input type="checkbox"/> plas
3	<input type="checkbox"/> pres
4	<input type="checkbox"/> skin
5	<input type="checkbox"/> insu
6	<input type="checkbox"/> mass
7	<input type="checkbox"/> pedi
8	<input type="checkbox"/> age
9	<input checked="" type="checkbox"/> class

**Selected attribute**

Name: class  
Missing: 0 (0%)  
Distinct: 2  
Type: Nominal  
Unique: 0 (0%)

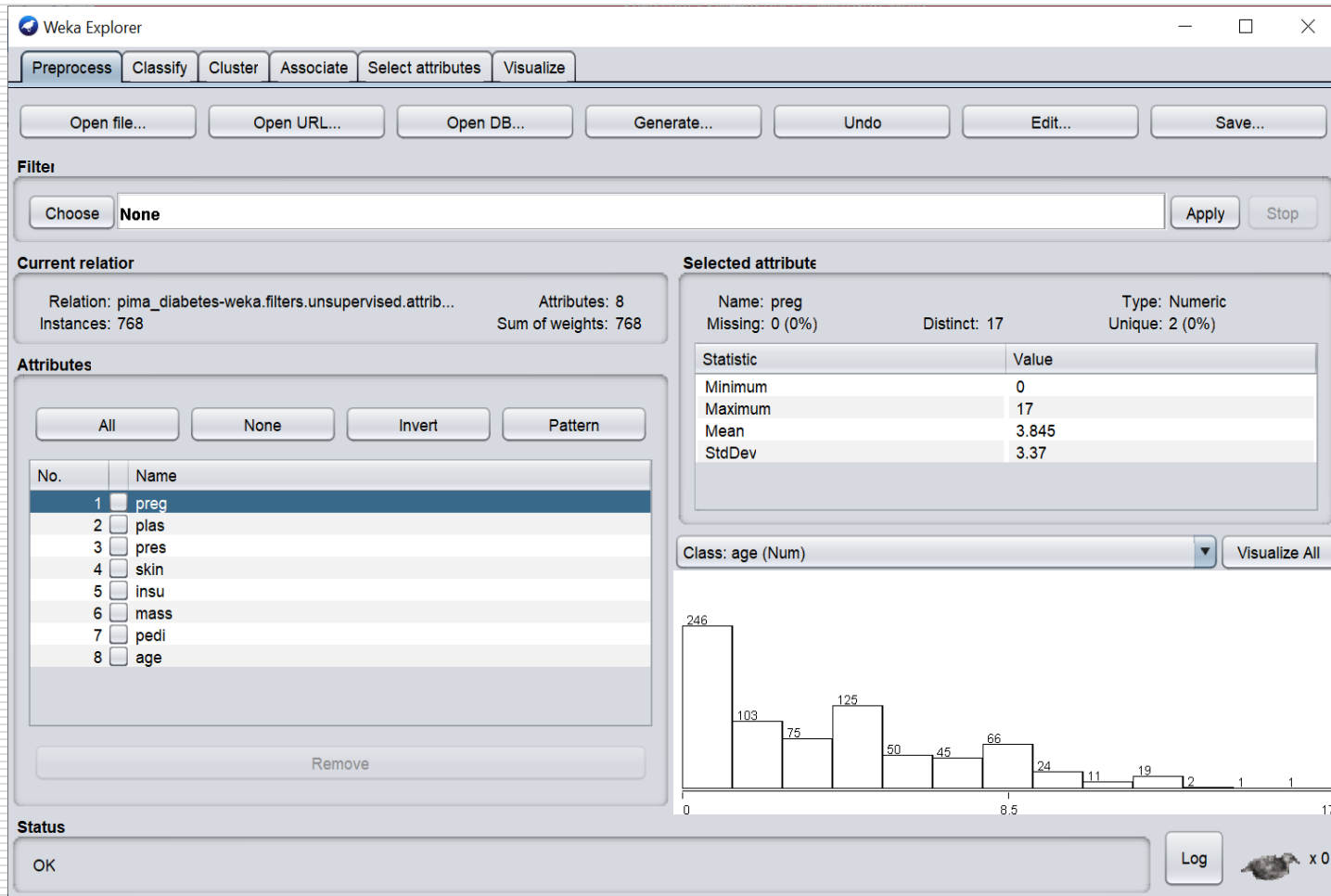
No.	Label	Count	Weight
1	tested_negative	500	500.0
2	tested_positive	268	268.0

Class: class (Nom) [Visualize All]

Bar chart showing counts: 500 (blue bar) and 268 (red bar).

# Loại bỏ thuộc tính class (phân lớp)

## ❖ Kết quả:



# Phân cụm k-Means

- ❖ Chuyển tab Cluster
- ❖ Nhấn nút Choose để chọn thuật toán
- ❖ Chọn “Use training set”

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' list on the right shows 'SimpleKMeans' selected. The 'Choose' button is highlighted with a red box. Below, the 'Cluster mode' section has 'Use training set' selected with a red box. The 'Clusterer output' pane displays the following statistics:

	mean	std. dev.	age
mean	0.3491	0.5238	0.5432
std. dev.	0.2085	0.3844	0.3345

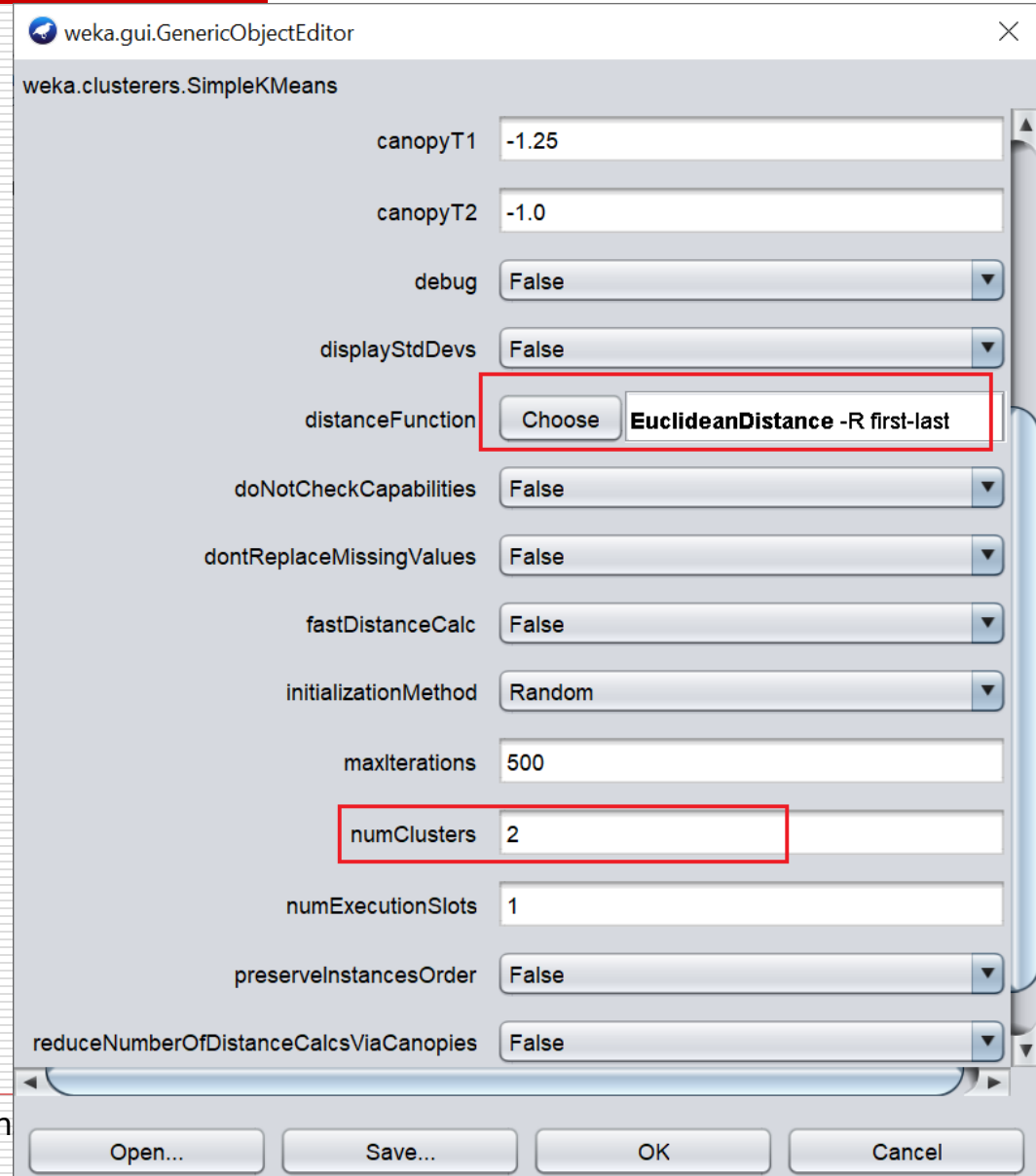
	mean	std. dev.	age
mean	32.3749	24.5802	44.2618
std. dev.	9.7378	3.0226	11.1922

Time taken to build model (full training data) : 2.28 seconds  
=== Model and evaluation on training set ===  
Clustered Instances  
0 369 ( 48%)  
1 250 ( 33%)  
2 149 ( 19%)  
Log likelihood: -23.10594

The 'Result list' shows '19:40:06 - EM'.

# Phân cụm k-Means

- ❖ Nhấn chuột trái vào Hộp thông tin thuật toán bên cạnh để mở cửa sổ cấu hình thuật toán
- ❖ Cấu hình tham số cho kMeans
- ❖ Nhấn OK để đóng cửa sổ



# Phân cụm k-Means

❖ Nhấn nút Start và xem kết quả:

The screenshot shows the Weka Explorer application window. The 'Clusterer' tab is selected. The 'Choose' button is pressed, and the 'SimpleKMeans' algorithm is selected with the following options: `-init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -num-slot:`

**Cluster mode**

- ☒ Use training set
- ☐ Supplied test set (Set...)
- ☐ Percentage split % 66
- ☐ Classes to clusters evaluation (Num) age
- ☒ Store clusters for visualization

**Clusterer output**

Attribute	Full Data	0	1
	(768.0)	(515.0)	(253.0)
preg	3.8451	2.0835	7.4308
plas	120.8945	115.3282	132.2253
pres	69.1055	65.9903	75.4466
skin	20.5365	21.8194	17.9249
insu	79.7995	85.0194	69.1739
mass	31.9926	31.7751	32.4352
pedi	0.4719	0.4708	0.4741
age	33.2409	26.7728	46.4071

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	515 ( 67%)
1	253 ( 33%)

**Status**

OK



# So sánh kết quả phân cụm và trường class

## ❖ Giữ nguyên trường class:

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply Stop

**Current relation**  
Relation: pima\_diabetes  
Instances: 768  
Attributes: 9  
Sum of weights: 768

**Attributes**  
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> preg
2	<input type="checkbox"/> plas
3	<input type="checkbox"/> pres
4	<input type="checkbox"/> skin
5	<input type="checkbox"/> insu
6	<input type="checkbox"/> mass
7	<input type="checkbox"/> pedi
8	<input type="checkbox"/> age
9	<input type="checkbox"/> class

Remove

**Selected attribute**  
Name: preg  
Missing: 0 (0%)  
Distinct: 17  
Type: Numeric  
Unique: 2 (0%)

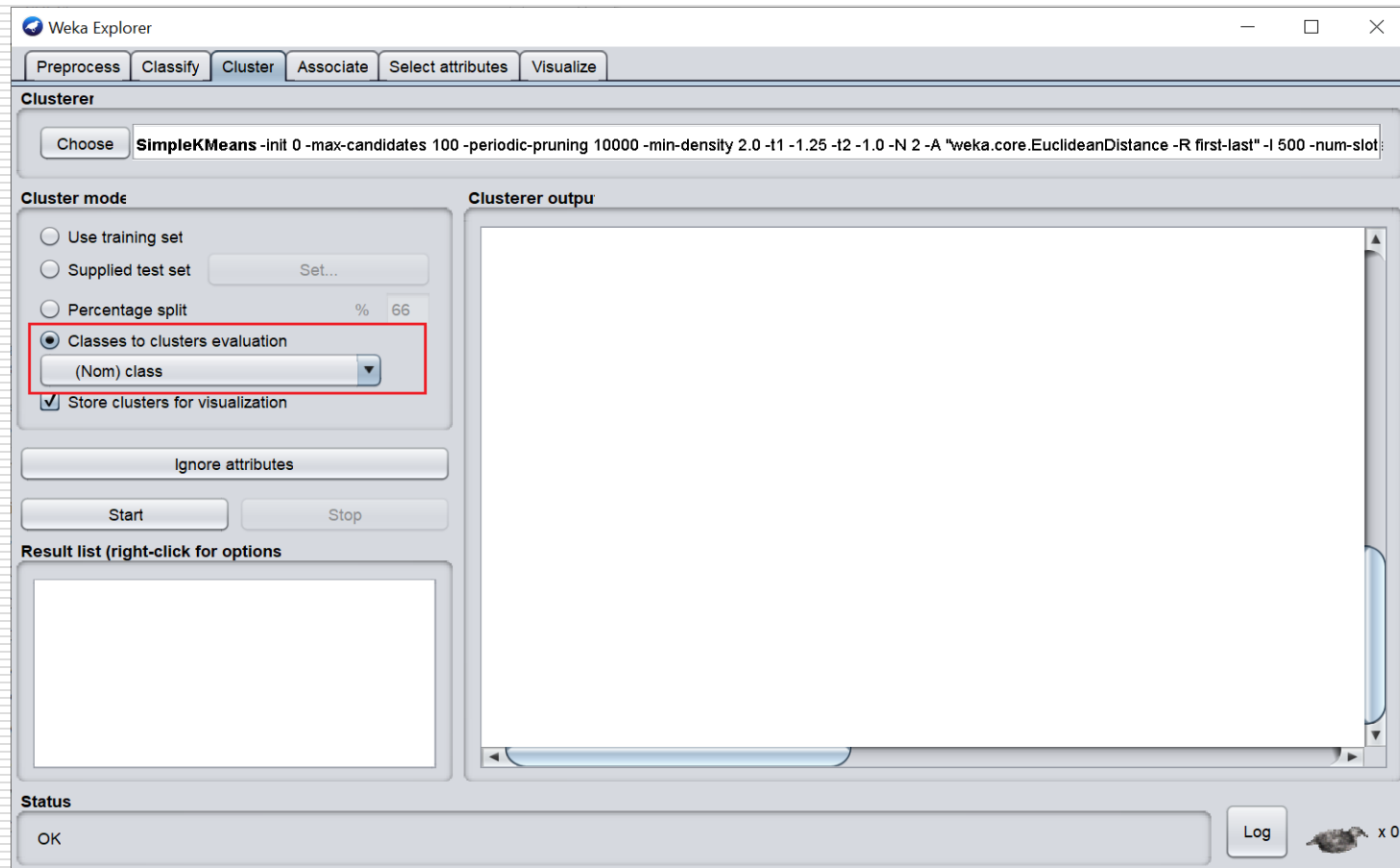
Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Class: class (Nom) Visualize All

Log x 0

# So sánh kết quả phân cụm và trường class

- ❖ Trong tab cluster, ko chọn training set mà chọn trường để so khớp:



# So sánh kết quả phân cụm và trường class

- ❖ Nhấn Start và xem kết quả:
  - ❖ Trong 515 về cụm 0 thì có 135 đúng
  - ❖ Trong 253 về cụm 1 thì có 133 đúng

The screenshot shows the Weka Explorer application window. The 'Cluster' tab is selected. The 'Clusterer' section shows 'SimpleKMeans' with various parameters. The 'Cluster mode' section has 'Classes to clusters evaluation' selected. The 'Clusterer output' section displays the following text:

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      515 ( 67%)
1      253 ( 33%)

Class attribute: class
Classes to Clusters:

  0   1  <-- assigned to cluster
380 120 | tested_negative
135 133 | tested_positive

Cluster 0 <-- tested_negative
Cluster 1 <-- tested_positive

Incorrectly clustered instances :      255.0      33.2031 %
```

The 'Result list' section shows '19:47:23 - SimpleKMeans'. The 'Status' section shows 'OK'.

# Phân cụm với thuật toán EM

- ❖ Cách làm giống kMeans, chỉ khác về tham số mô hình

The image shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'EM'. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' pane displays the following data:

Clusterer output			
age			
mean	39.7217	24.4448	
std. dev.	11.5994	2.9329	
class			
tested_negative	211.0899	290.9101	
tested_positive	233.1071	36.8929	
[total]	444.197	327.803	

Time taken to build model (full training data):

=== Model and evaluation on training set ===

Clustered Instances

Cluster	Count	Percentage
0	432	( 56%)
1	336	( 44%)

Log likelihood: -29.68781

The 'GenericObjectEditor' dialog is open, showing the 'weka.clusterers.EM' class. The 'numClusters' parameter is highlighted with a red box and set to 2.

Status: OK

Log x 0

# Phân cụm với thuật toán EM

- ❖ Có thể chọn số cụm là -1 và EM tự động chọn số cụm tốt

The image shows two windows from the Weka software. The left window is 'Weka Explorer' and the right is 'weka.gui.GenericObjectEditor'.

**Weka Explorer:**

- Buttons: Preprocess, Classify, Cluster, Associate, Select attributes, Visualize.
- Clusterer: Choose, EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
- Cluster mode:
  - ☒ Use training set
  - ☐ Supplied test set (Set...)
  - ☐ Percentage split (% 66)
  - ☐ Classes to clusters evaluation (Nom) class
  - ☒ Store clusters for visualization
- Buttons: Ignore attributes, Start, Stop
- Result list (right-click for options):
  - 15:29:23 - EM
  - 15:32:01 - EM (selected)
- Status: OK

**weka.gui.GenericObjectEditor (weka.clusterers.EM):**

- About: Simple EM (expectation maximisation) class. (More, Capabilities)
- debug: False
- displayModelInOldFormat: False
- doNotCheckCapabilities: False
- maxIterations: 100
- maximumNumberOfClusters: -1
- minLogLikelihoodImprovementCV: 1.0E-6
- minLogLikelihoodImprovementIterating: 1.0E-6
- minStdDev: 1.0E-6
- numClusters: -1 (highlighted with a red box)
- numExecutionSlots: 1
- numFolds: 10
- numKMeansRuns: 10
- seed: 100
- Buttons: Open..., Save..., OK, Cancel

**Clusterer output:**

	mean	std. dev.	
36.6656	23.9583	38.9112	
10.7088	2.5897	12.6593	

class

	tested_negative	tested_positive	[total]
108.1931	236.1879	158.619	
142.6434	19.676	108.6807	
250.8365	255.8639	267.2997	

Time taken to build model (full training data) :

=== Model and evaluation on training set ===

Clustered Instances

0	228	( 30%)
1	203	( 26%)
2	337	( 44%)

Log likelihood: -24.97229