Ejercicio 2

PASO 2 – OPERADOR Y PARÁMETROS

Las medidas disponibles son mútliples, tanto para atributos nominales como numéricos.

Entre los nominales está Jaccard, Simple Matching

Y entre los numéricos Distancia Euclideana, Chebychev, Correlación, Similaridad del coseno.

También se encuentra la divergencia de Bregman que ofrece varias medidas de divergencia.

PASO 3 – EVALUACION

El vector de performance de cluster no tiene parámetros configurables.

Data to similarity pide solicita configurar la medida de distancia a ser utilizada.

PASO 4 – EJECUCION E INTERPRETACIÓN

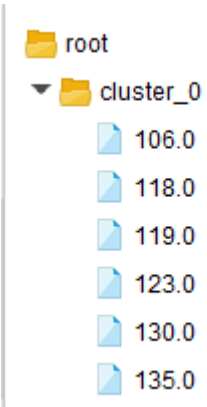
La salida de model tiene información sobre la cantidad de casos en cada cluster:
Con un épsilon de 0.25 se obtiene la siguiente clasificación:

Cluster Model

```
Cluster 0: 6 items
Cluster 1: 50 items
Cluster 2: 94 items
Total number of items: 150
```

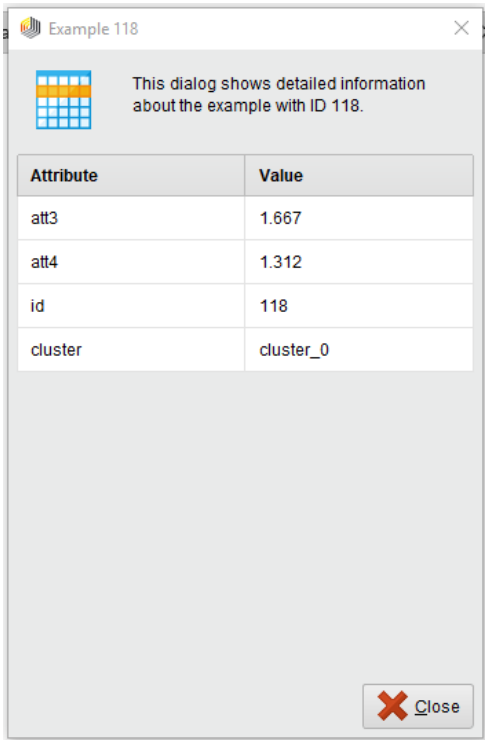
Identifica como ruido 6 observaciones, mientras que 50 son clasificadas en el cluster_1 y 94 en el cluster_2.

En folder view se pueden observar por ejemplo los 6 casos identificados como ruido:



Análogo para los demás cluster.

Si hacemos click en cualquiera de los casos obtenemos información de sus atributos:

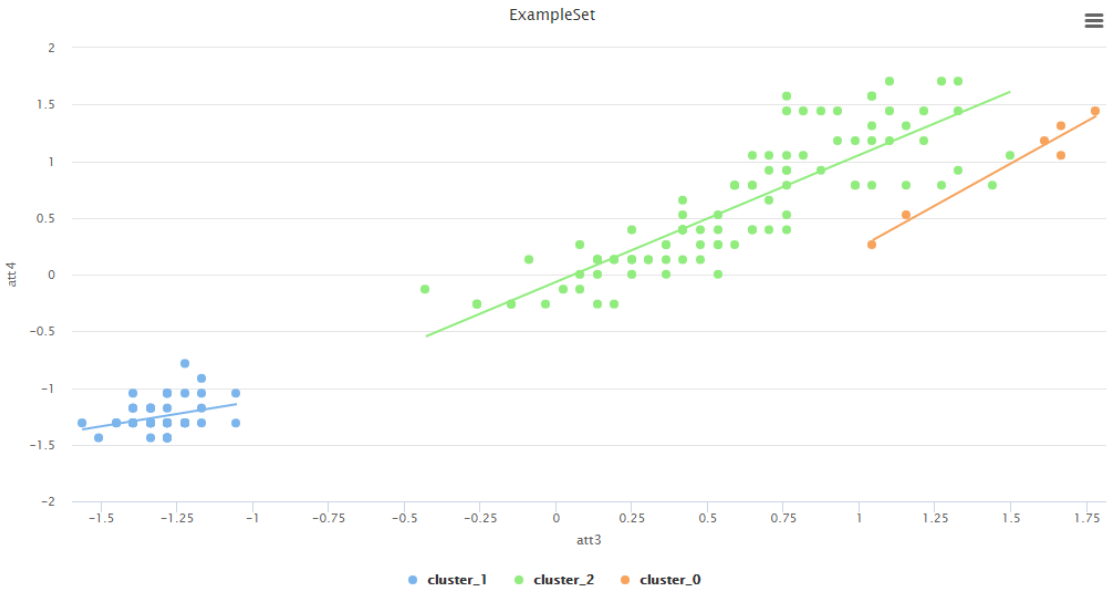


Clustered example set:

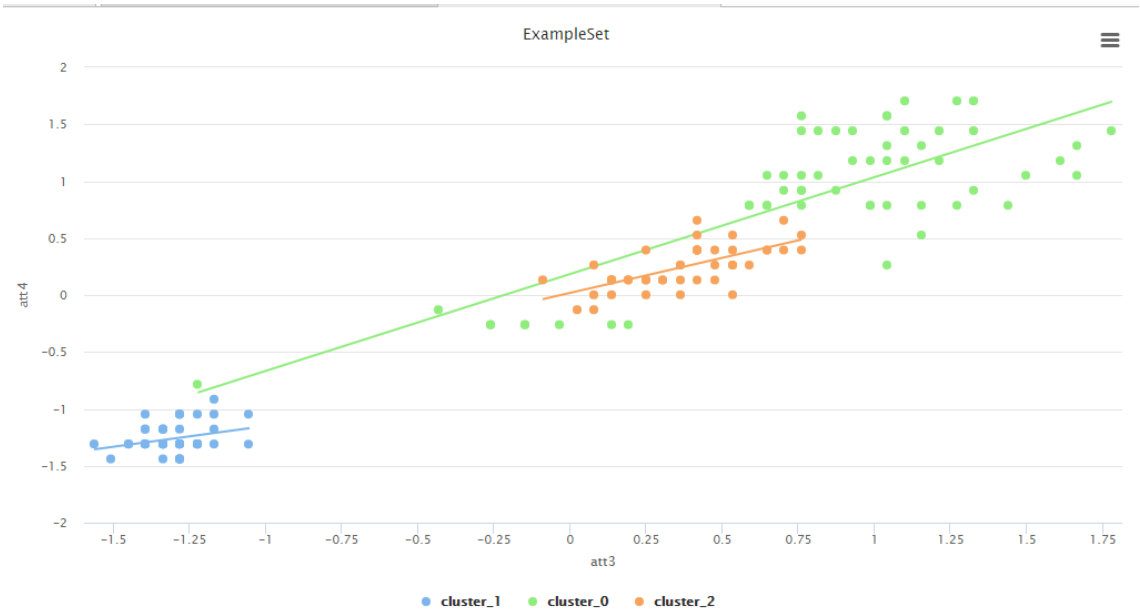
Row No.	id	cluster	att3	att4
1	1	cluster_1	-1.337	-1.309
2	2	cluster_1	-1.337	-1.309
3	3	cluster_1	-1.393	-1.309
4	4	cluster_1	-1.280	-1.309
5	5	cluster_1	-1.337	-1.309
6	6	cluster_1	-1.167	-1.047
7	7	cluster_1	-1.337	-1.178
8	8	cluster_1	-1.280	-1.309
9	9	cluster_1	-1.337	-1.309
10	10	cluster_1	-1.280	-1.440
11	11	cluster_1	-1.280	-1.309
12	12	cluster_1	-1.223	-1.309
13	13	cluster_1	-1.337	-1.440
14	14	cluster_1	-1.507	-1.440
15	15	cluster_1	-1.450	-1.309

Se observa la incorporación de la etiqueta del número de cluster al modelo.

En la siguiente imagen se observan los cluster identificados:



Aumentando el número de puntos mínimos a 15 se obtienen la misma cantidad de clusters



Pero con una transversalidad mayor del cluster_1

<u>Vector de performance.</u>

La distancia promedio entre los cluster se muestra en el vector de performance con mínima cantidad de puntos de 15:

Avg. within cluster distance

```
Avg. within cluster distance: -27.447
```

Que tiene un desempeño mejor que al haber seteado mínima cantidad de puntos en 5:

Avg. within cluster distance

```
Avg. within cluster distance: -76.595
```

También puede observar la distancia promedio de los puntos de cada cluster

> # Avg. within cluster distance for cluster 0
>
> ```
> Avg. within cluster distance for cluster 0: -52.717
> ```

El cluster_0 por ser el ruido tiene una distancia promedio alta, mientras que en los demás desciende a 9-14.

Dist promedio	Cluster_0	Cluster_1	Cluster_2	Épsilon	Min num ptos
-27.447	-52.717	-9.233	-14.993	0.25	15
-61.074	-	-10.013	-86.604	0.5	7
-50.486	-6.301	-10.013	-75.572	0.3	7