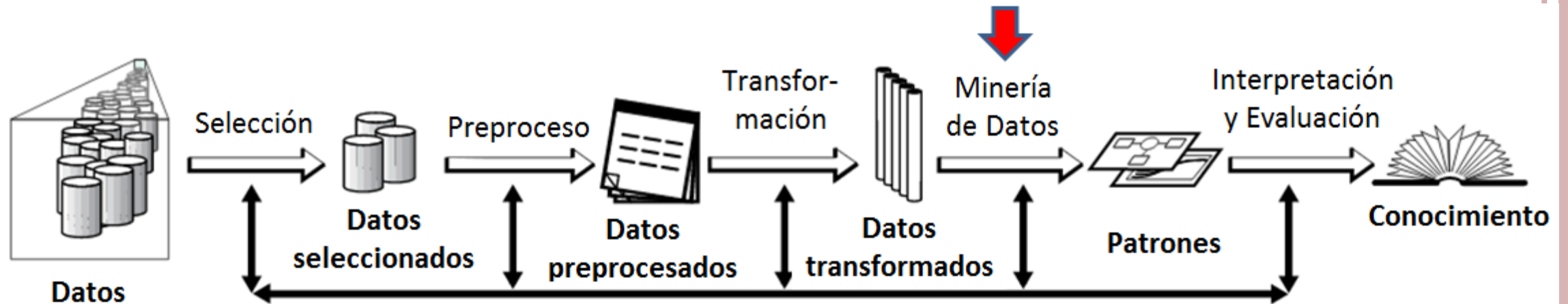
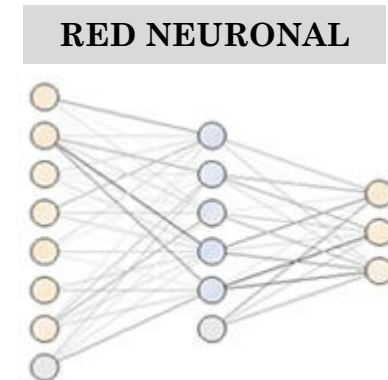
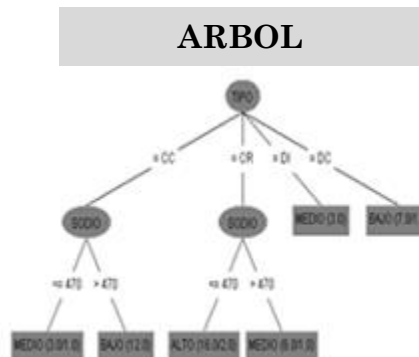


MINERÍA DE DATOS Y EL PROCESO DE KDD



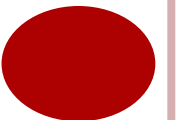
Fayyad (1996)

□ Técnicas de Minería de Datos

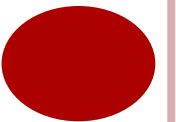


AGRUPAMIENTO O CLUSTERING

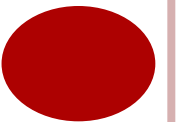
- El *clustering* es uno de los métodos de aprendizaje **no supervisado** más importantes y busca caracterizar conceptos desconocidos a partir de los ejemplos disponibles.
- Generalmente, en un problema real **se desconoce la clase** y es allí donde el agrupamiento puede ayudar a identificar las características comunes entre instancias.
- Al no disponer de la clase utiliza una **medida de similitud** para determinar el parecido entre instancias.



¿CÓMO LOS AGRUPARÍAS?

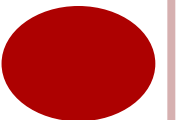


¿CÓMO LOS AGRUPARÍAS?



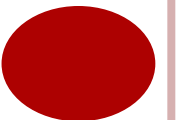
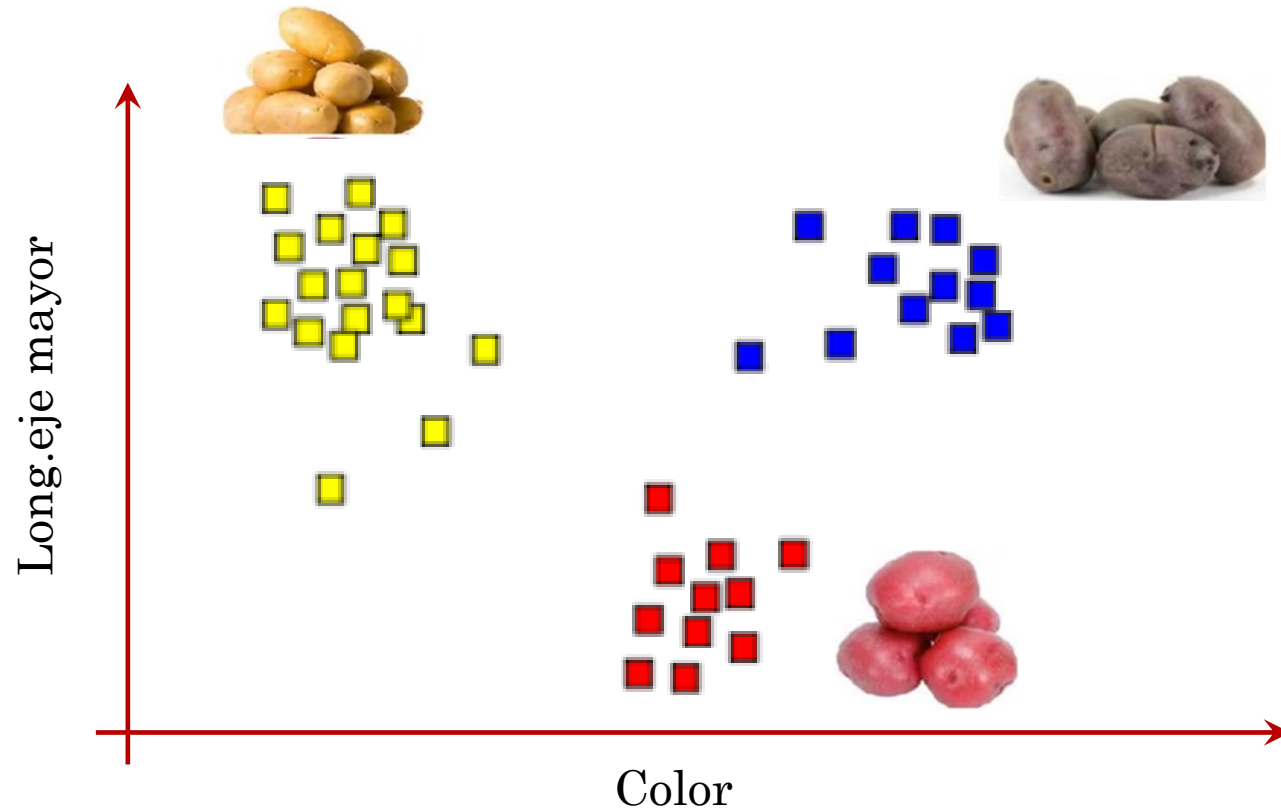
ELEMENTOS BÁSICOS DEL AGRUPAMIENTO

- Identificar las características relevantes de cada tipo de elemento.
- Indicar la manera en que se realizará la comparación (**DISTANCIA**)

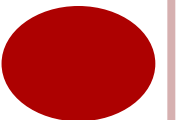
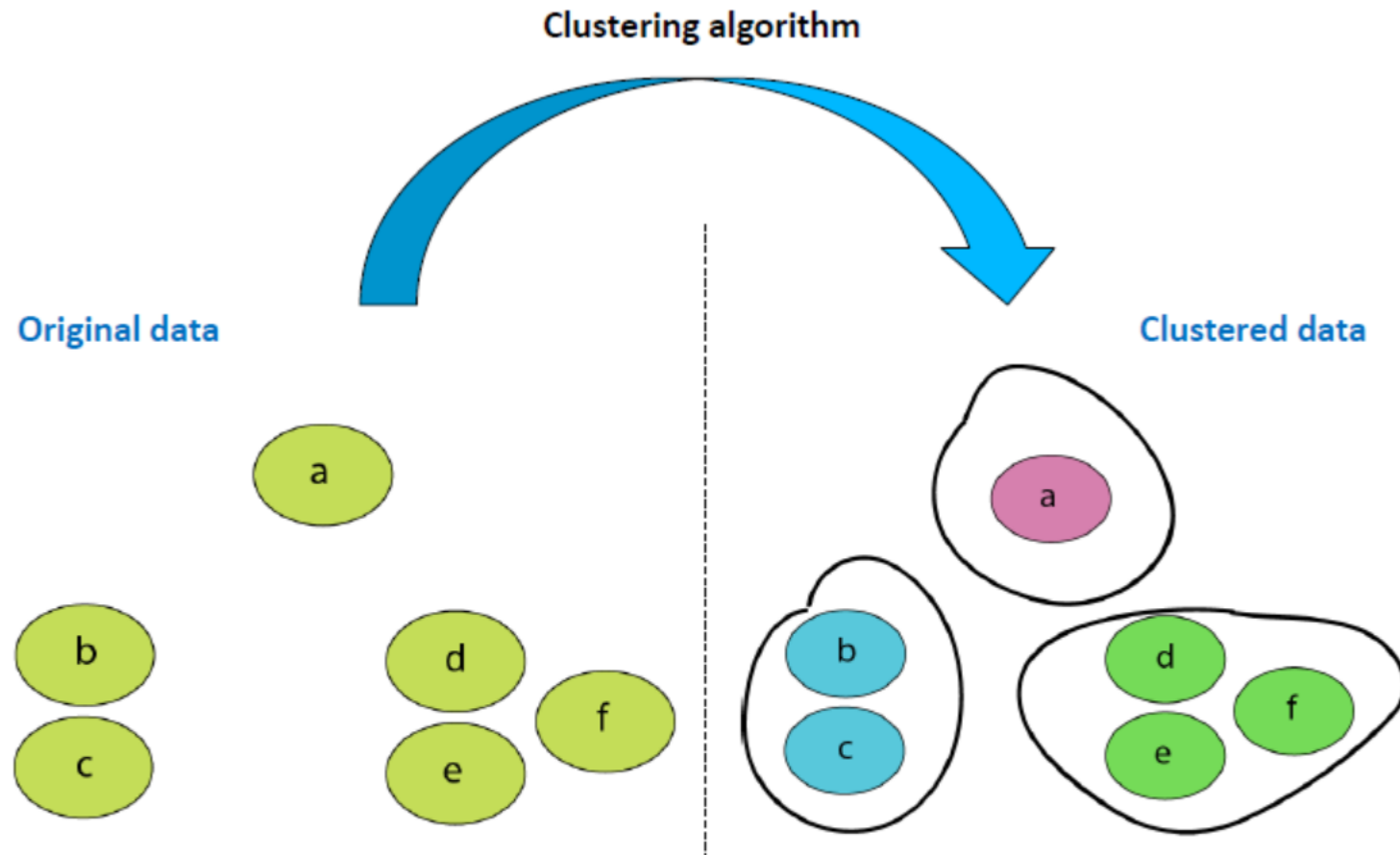


AGRUPAMIENTO O CLUSTERING

- El resultado de aplicar una técnica de *clustering* es una serie de agrupamientos o *clusters* formados al particionar las instancias.

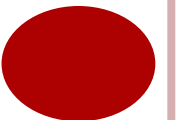


AGRUPAMIENTO - OBJETIVO



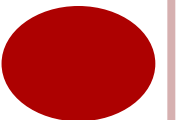
AGRUPAMIENTO O CLUSTERING

- Permite encontrar grupos de instancias con características similares.
- Aplicaciones
 - Identificar grupos y describirlos
 - Detectar clientes con características similares para ofrecer servicios adecuados.
 - Identificar alumnos con rendimientos académicos similares con el objetivo de reducir la deserción escolar.
 - Detección de casos anómalos
 - Detección de fraudes.



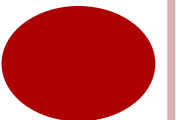
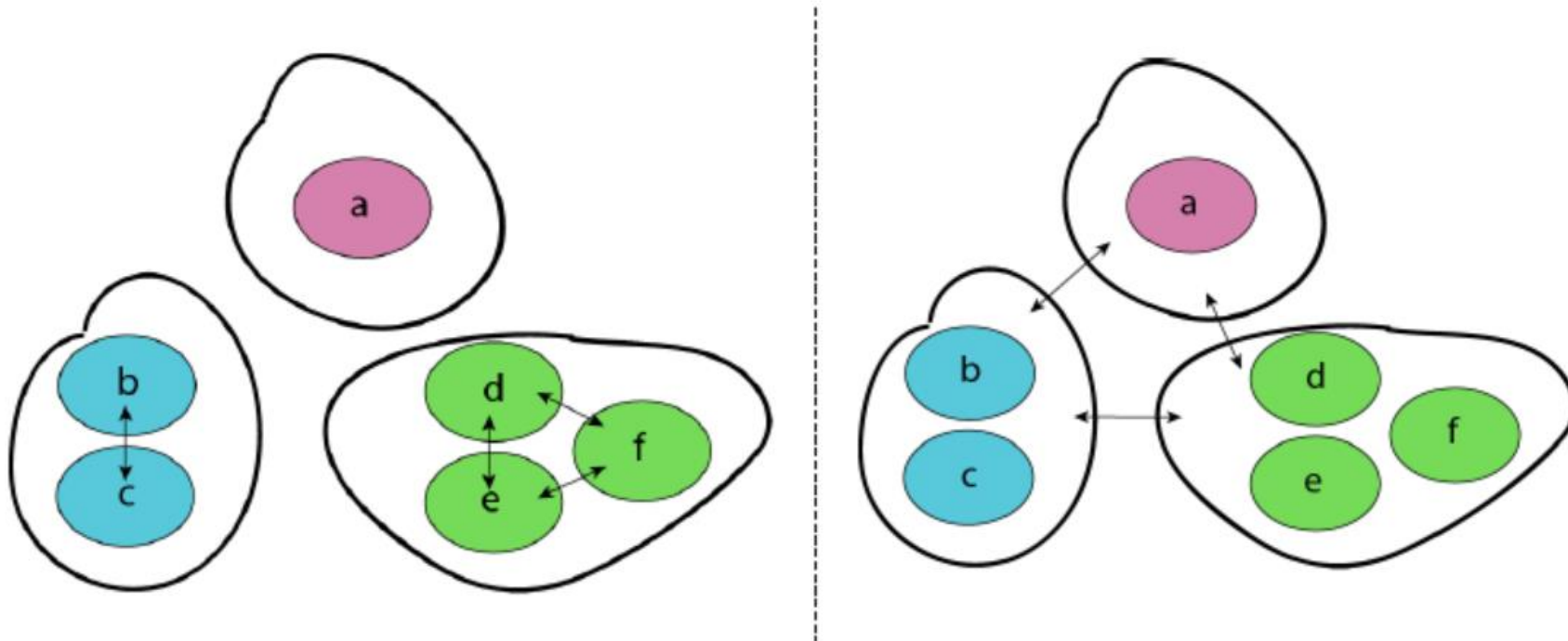
CALIDAD DEL AGRUPAMIENTO OBTENIDO

- Un buen método de agrupamiento producirá grupos de alta calidad en los cuales
 - El parecido entre los elementos que componen un mismo grupo es alto (intra-cluster).
 - El parecido entre los elementos de grupos distintos es bajo (inter-cluster).



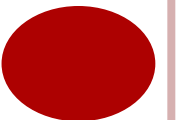
CALIDAD DEL AGRUPAMIENTO OBTENIDO

- Minimizar la distancia entre los elementos de un mismo cluster (intra-cluster)
- Maximizar la distancia entre clusters (inter-cluster)

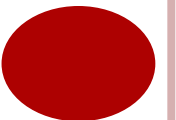
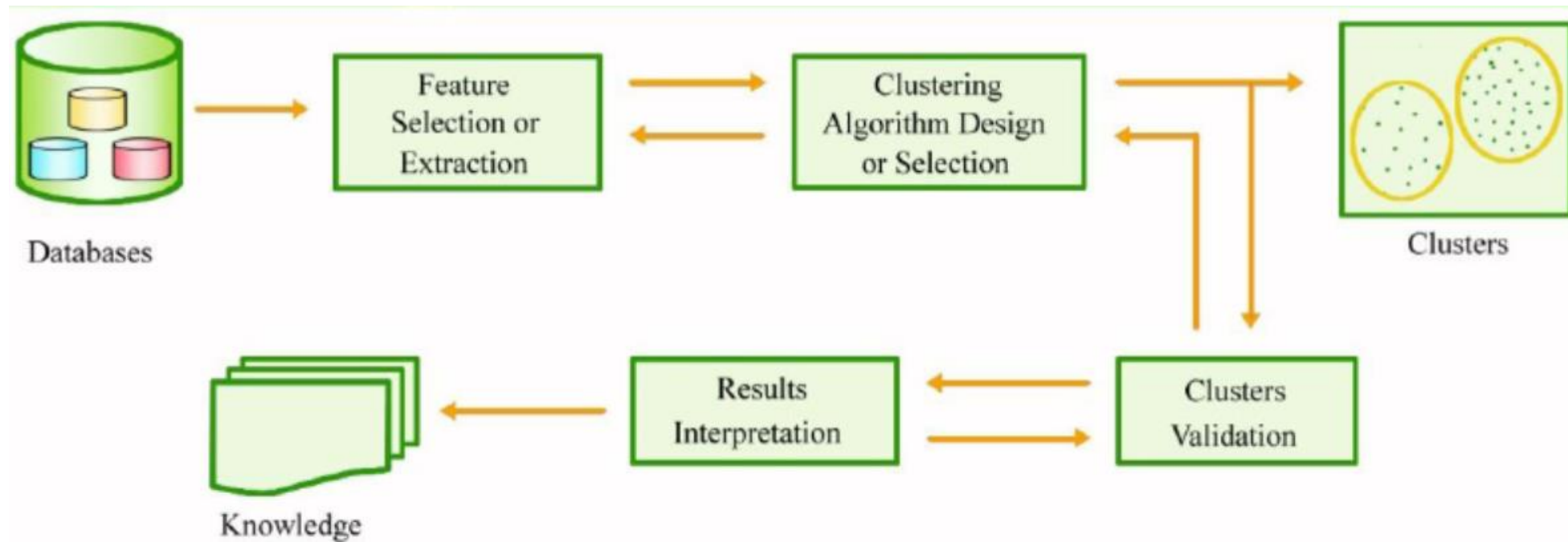


PROCESO DE AGRUPAMIENTO

- Seleccionar las características relevantes
- Definir una representación adecuada.
- Definir la medida de similitud a utilidad (medida de distancia).
Depende del problema.
- Aplicar un algoritmo de agrupamiento
- Validar los grupos obtenidos y de ser necesario volver a repetir el proceso.



PROCESO DE AGRUPAMIENTO



TIPOS DE ALGORITMOS DE AGRUPAMIENTO

○ Algoritmo Partitivo

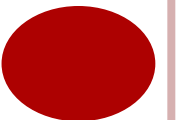
- Particionan los datos creando un número K de clusters.
- Una instancia pertenece a un único grupo.

○ Algoritmo Jerárquico

- Generan una estructura jerárquica de clusters que permiten ver las particiones de las instancias con distinta granularidad.
- Una instancia pertenece a un único grupo.

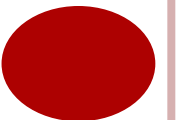
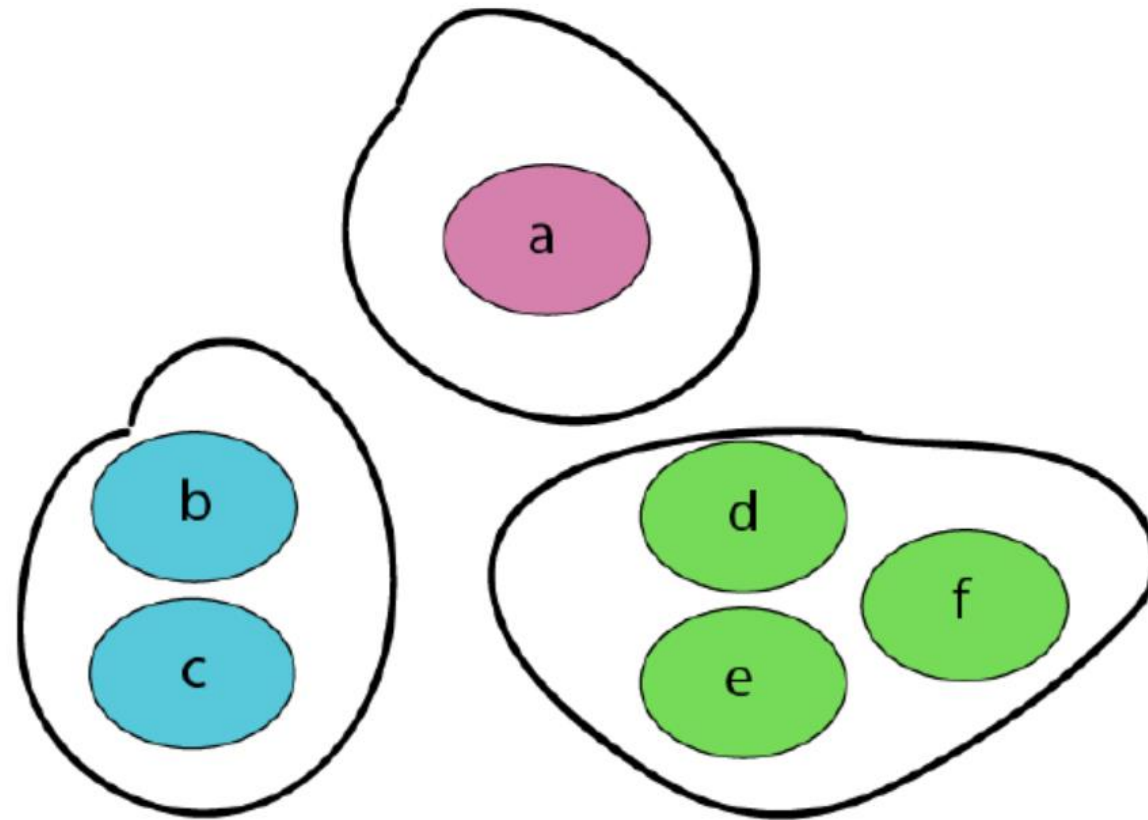
○ Algoritmo probabilista

- Los clusters se generan con un método probabilístico



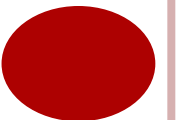
ALGORITMOS DE CLUSTERING PARTITIVOS

- Obtiene una única partición de los datos



K-MEDIAS

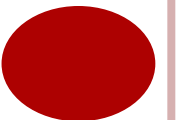
- El algoritmo K-Medias fue propuesto por MacQueen, en 1967.
- Requiere conocer a priori el número K de grupos a formar.
- El algoritmo está basado en la minimización de la distancia interna (la suma de las distancias de los ejemplos asignados a un agrupamiento al centroide de dicho agrupamiento).
- De hecho, este algoritmo minimiza la suma de las distancias al cuadrado de cada ejemplo al centroide de su agrupamiento.



K-MEDIAS

○ Características

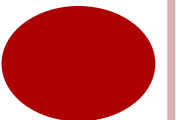
- El algoritmo es sencillo y eficiente.
- Procesa los ejemplos secuencialmente (por lo que requiere un almacenamiento mínimo).
- Está sesgado por el orden de presentación de los ejemplos (los primeros ejemplos determinan la configuración inicial de los agrupamientos)
- Su comportamiento depende enormemente del parámetro K .



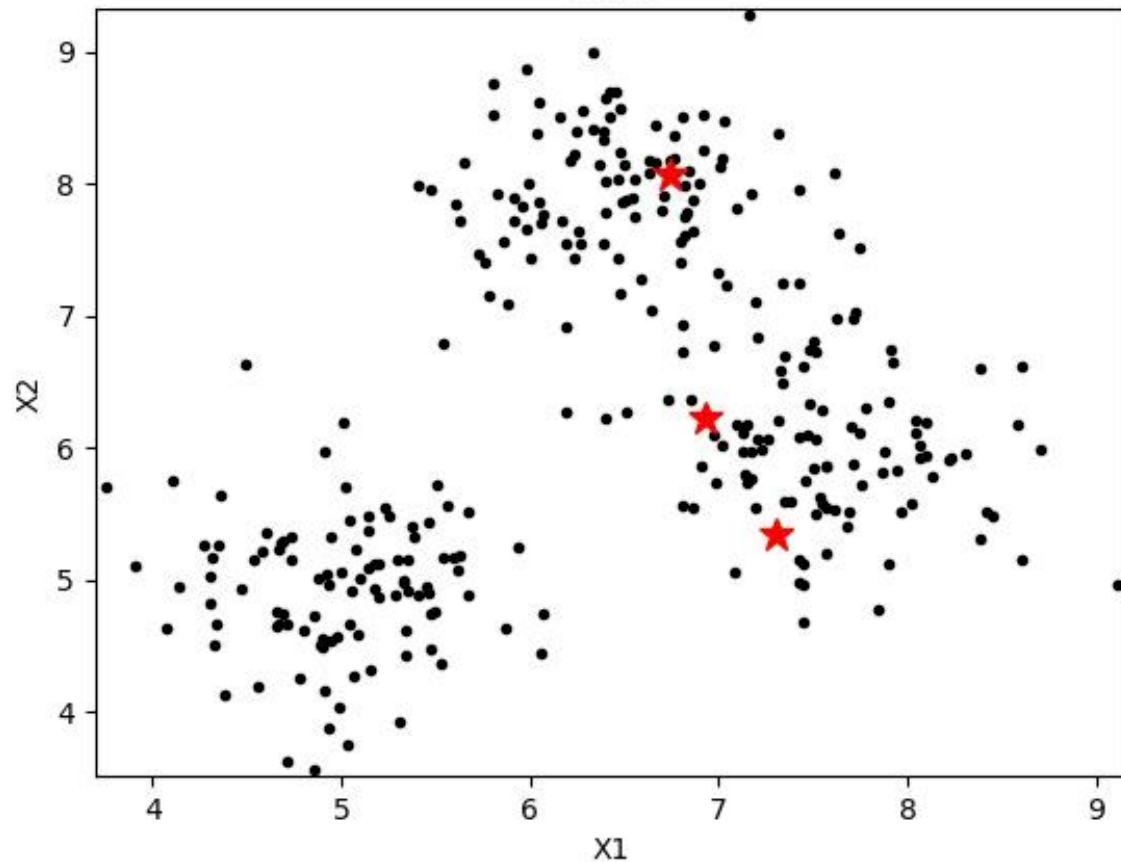
ALGORITMO K-MEDIAS

- Elegir aleatoriamente K ejemplos de entrada como centros iniciales.
- Repetir
 - Redistribuir los ejemplos entre los clusters utilizando la mínima distancia euclídea al cuadrado como clasificador.
 - Calcular los centros de los K clusters.

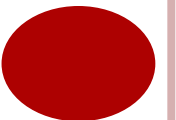
hasta que no cambien los centros de los clusters



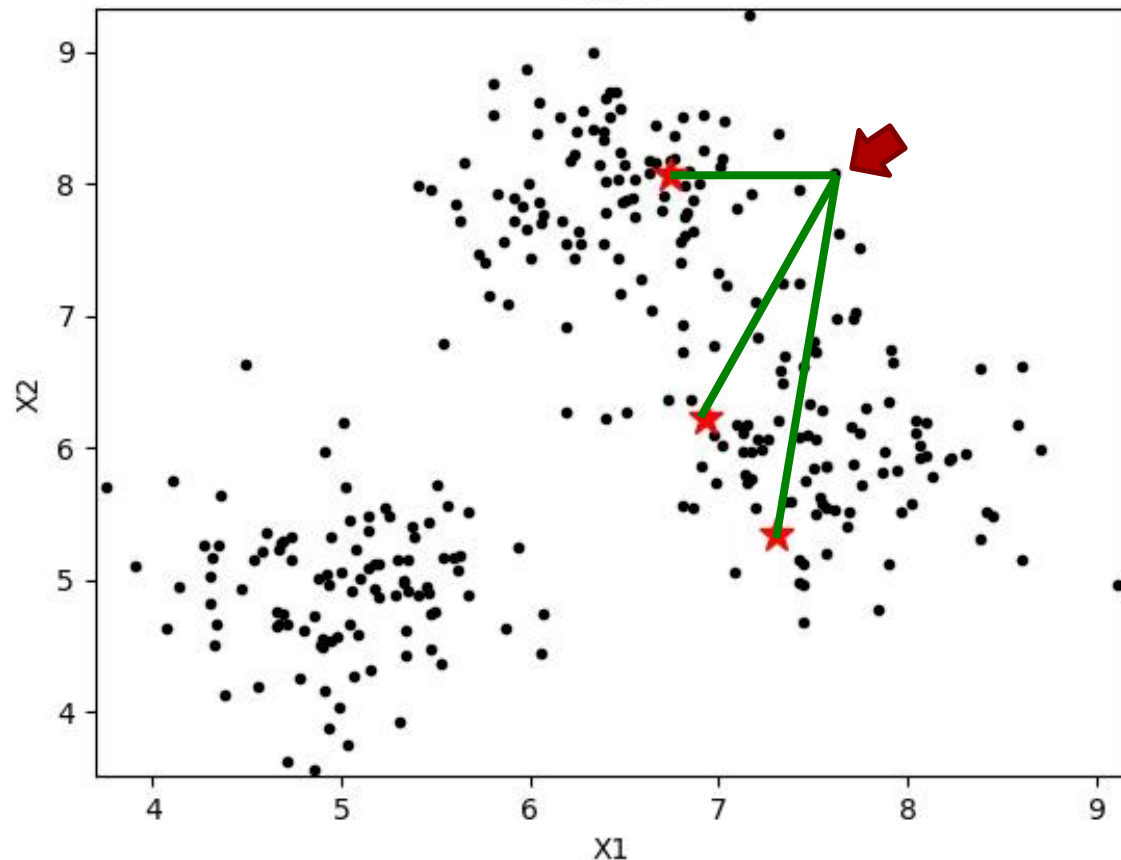
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



El proceso inicia
tomando **$k=3$**
ejemplos como
centros iniciales



Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters

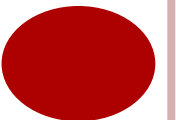


Calcular la distancia de cada ejemplo a cada centro y tomar la menor

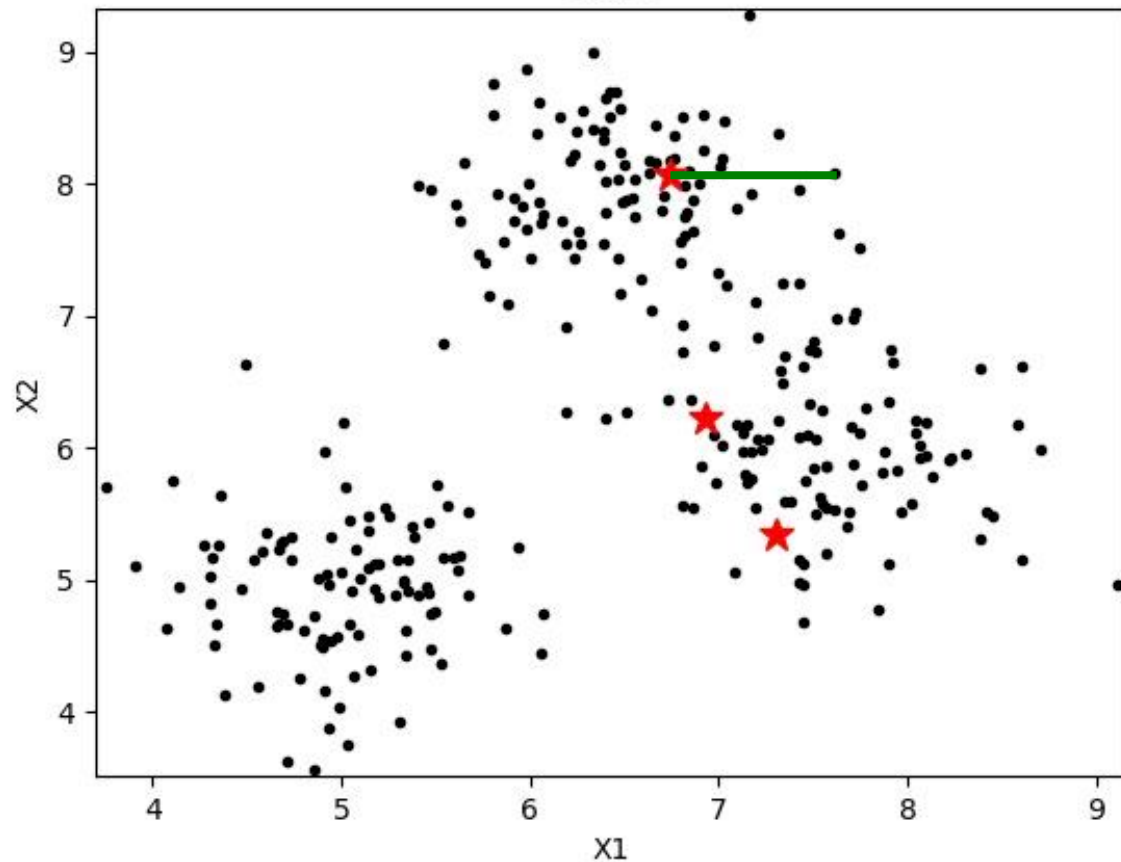
$$X = (x_1, x_2)$$

$$C = (c_1, c_2)$$

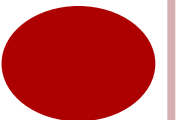
$$\text{dist}^2(C, X) = (c_1 - x_1)^2 + (c_2 - x_2)^2$$



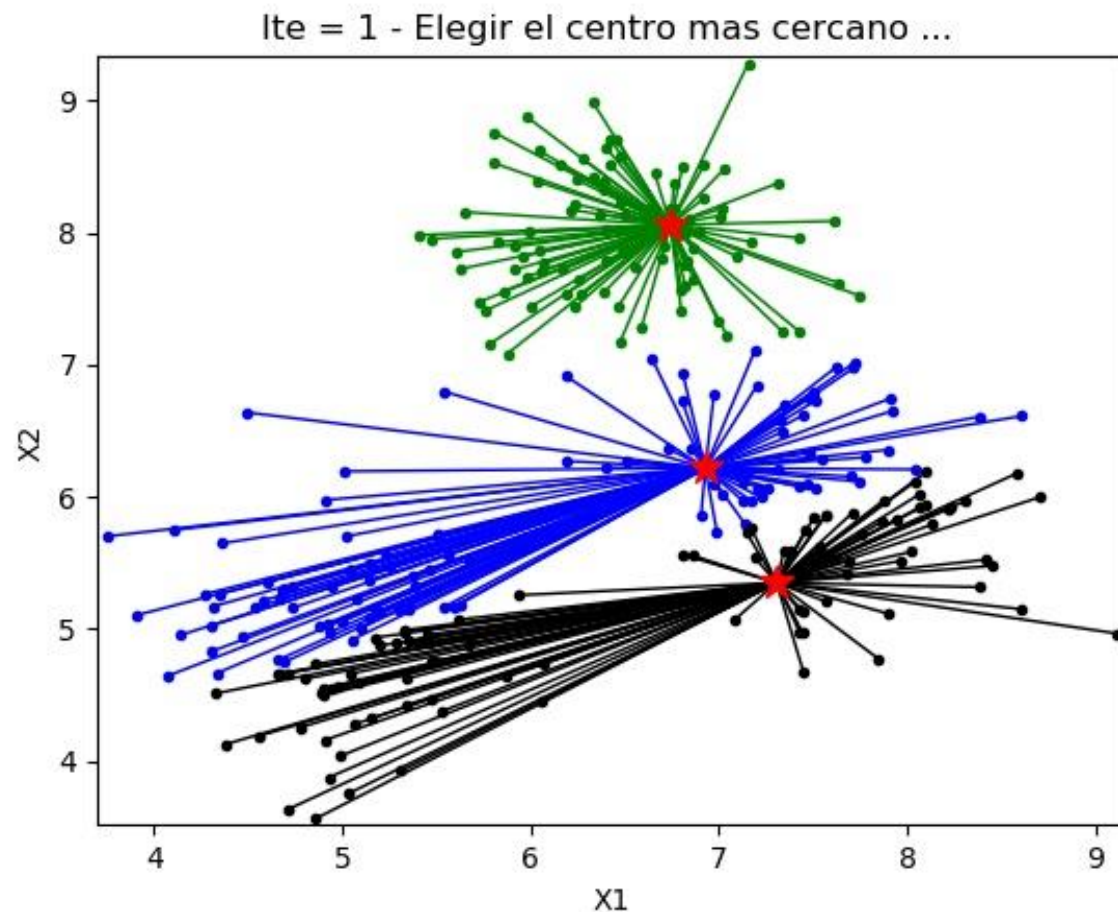
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



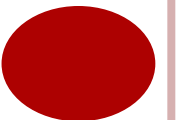
Calcular la distancia de cada ejemplo a cada centro y tomar la menor



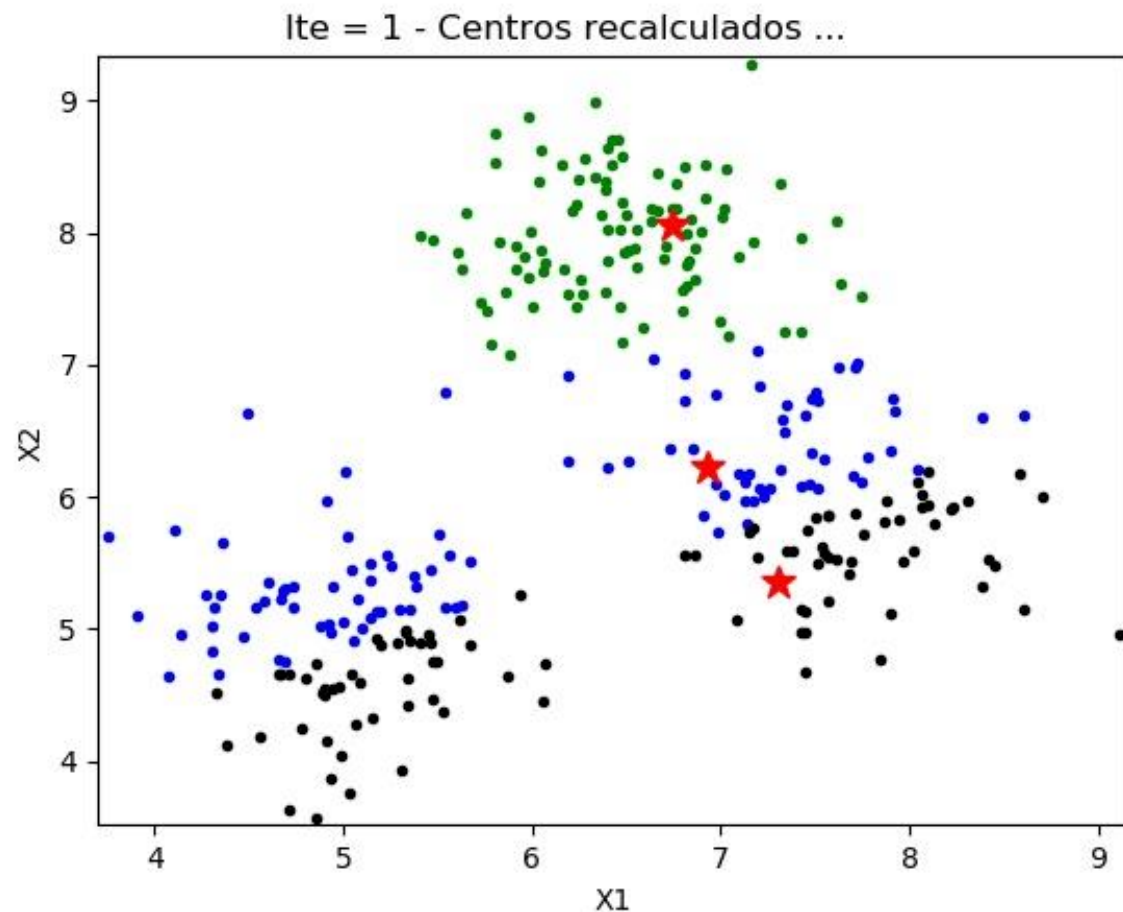
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



Asignar cada ejemplo al centro más cercano

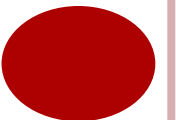


Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters

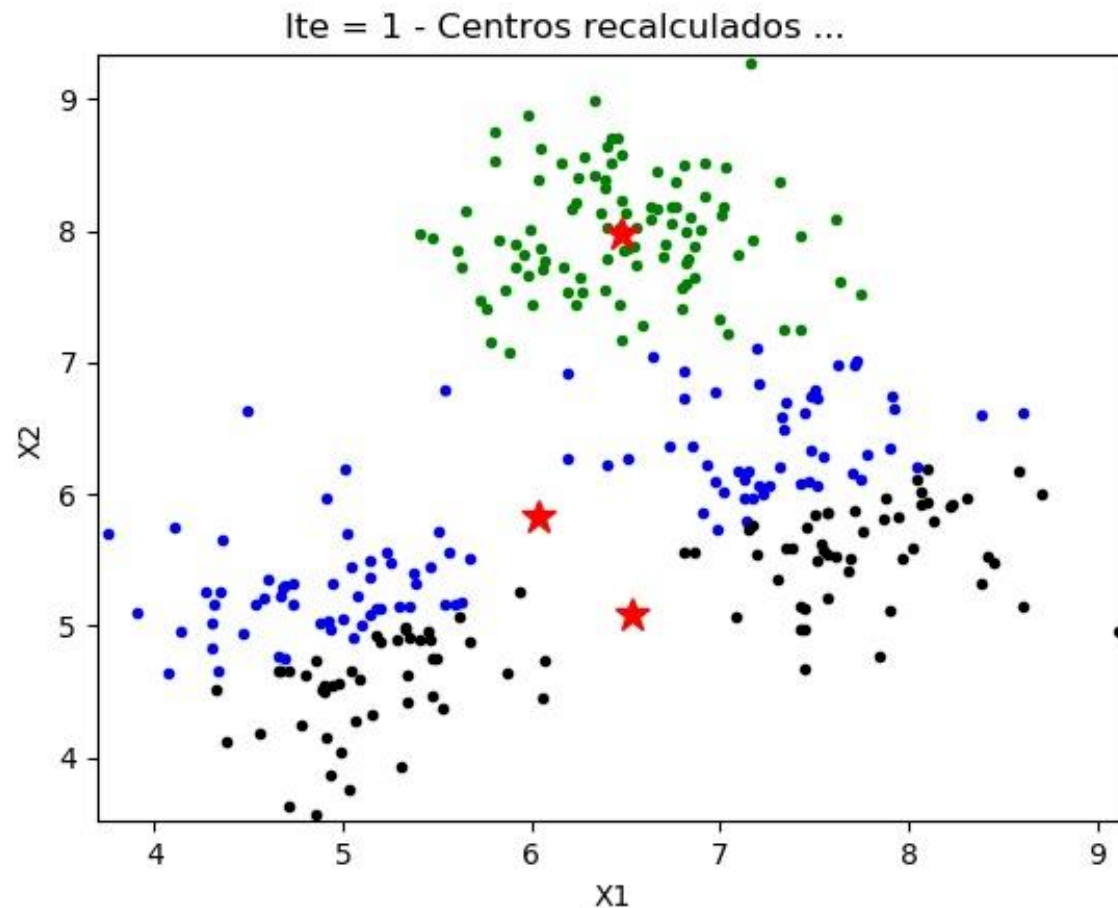


Recalcular la posición de los centros.

- Cada centro se reubica promediando los valores de los atributos de los ejemplos que los conforman.

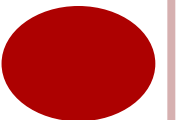


Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters

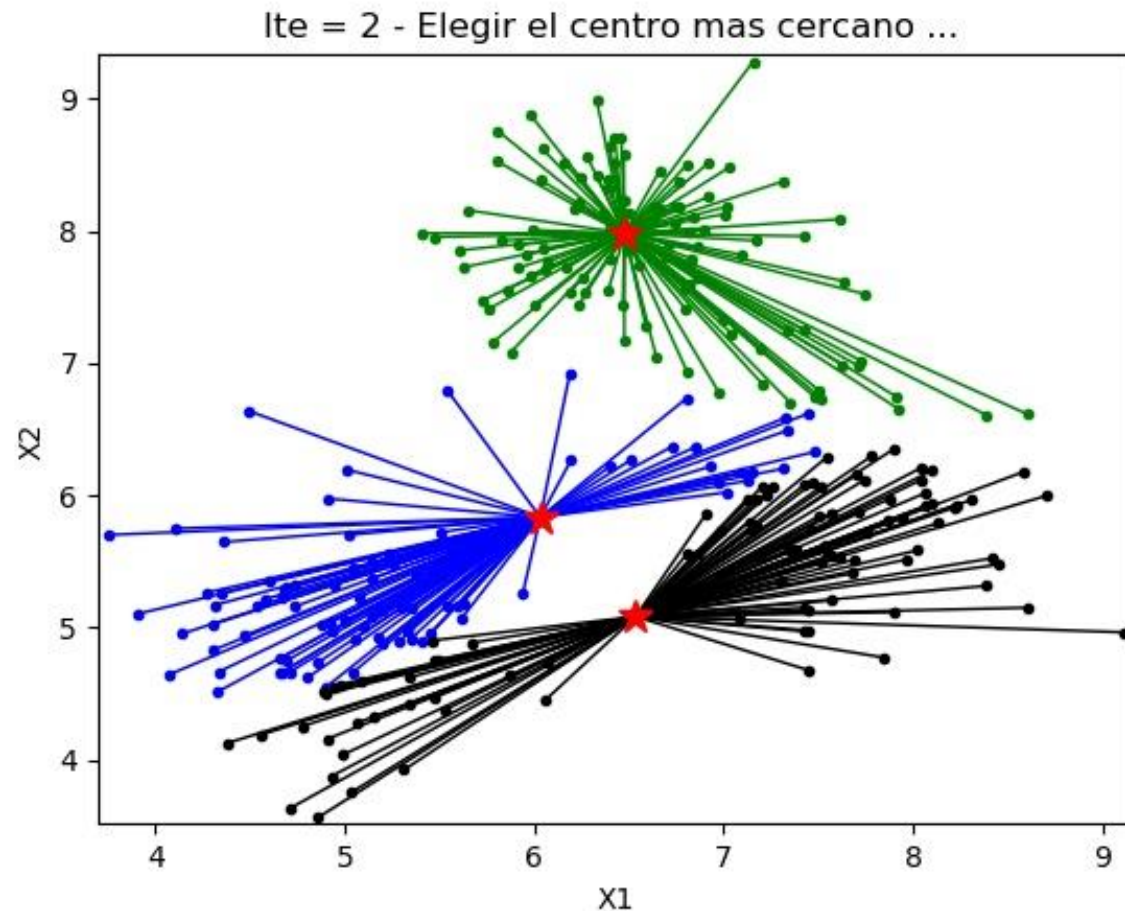


Recalcular la posición de los centros.

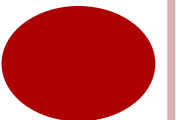
- Cada centro se reubica promediando los valores de los atributos de los ejemplos que los conforman.



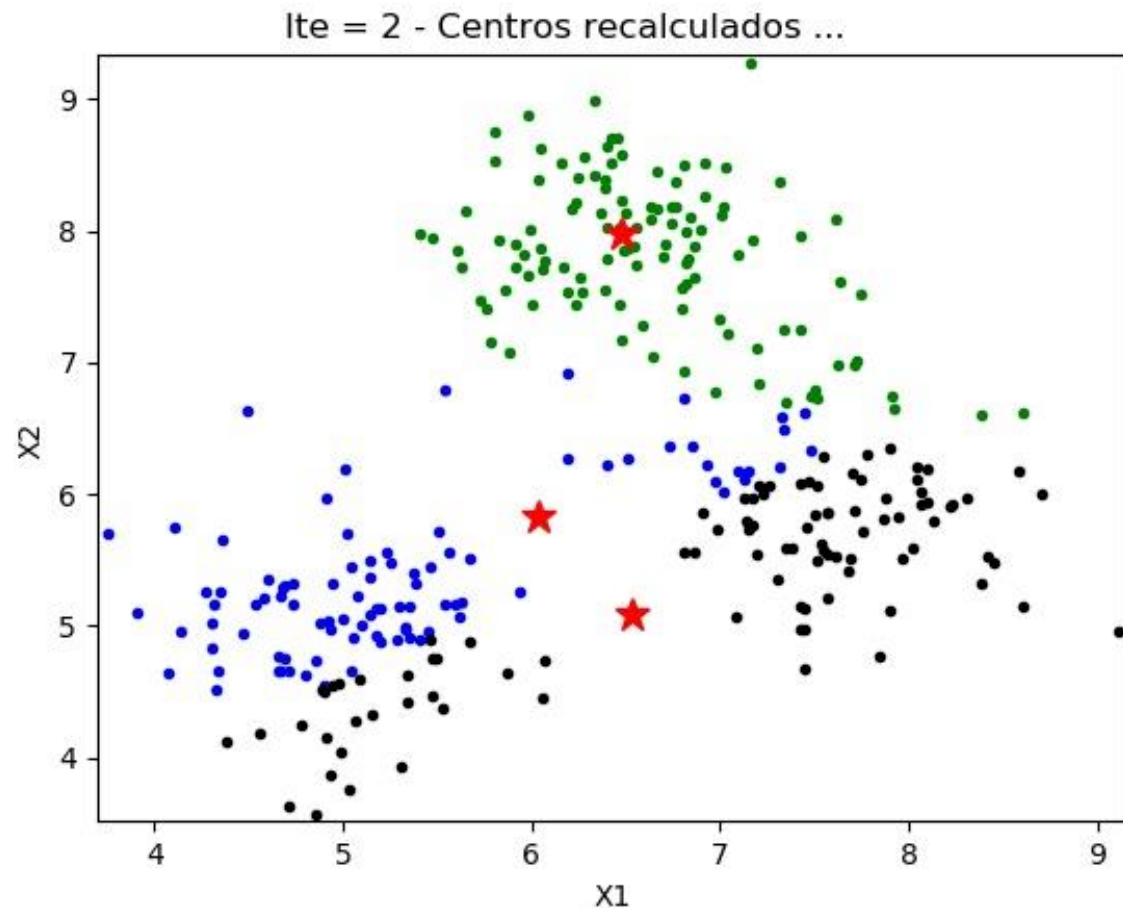
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



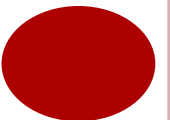
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



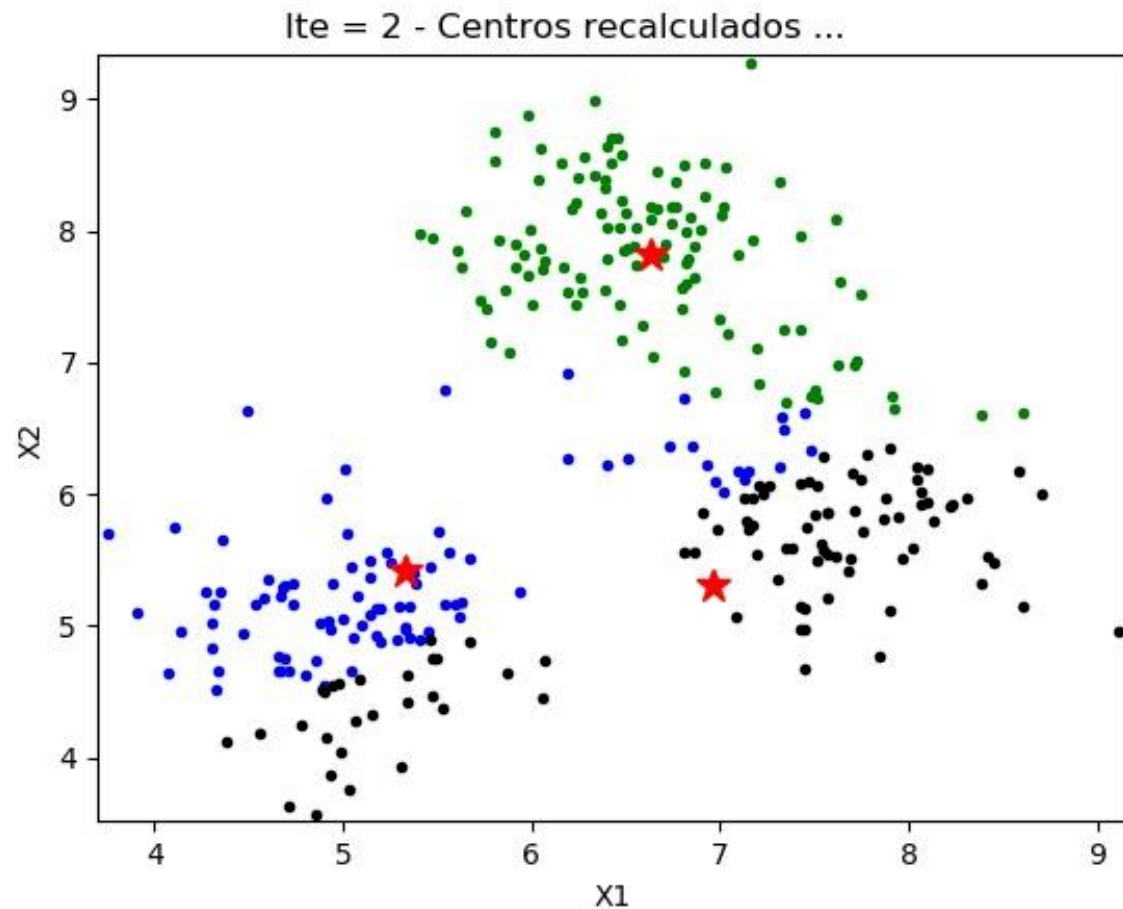
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



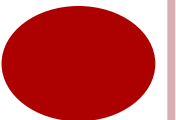
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



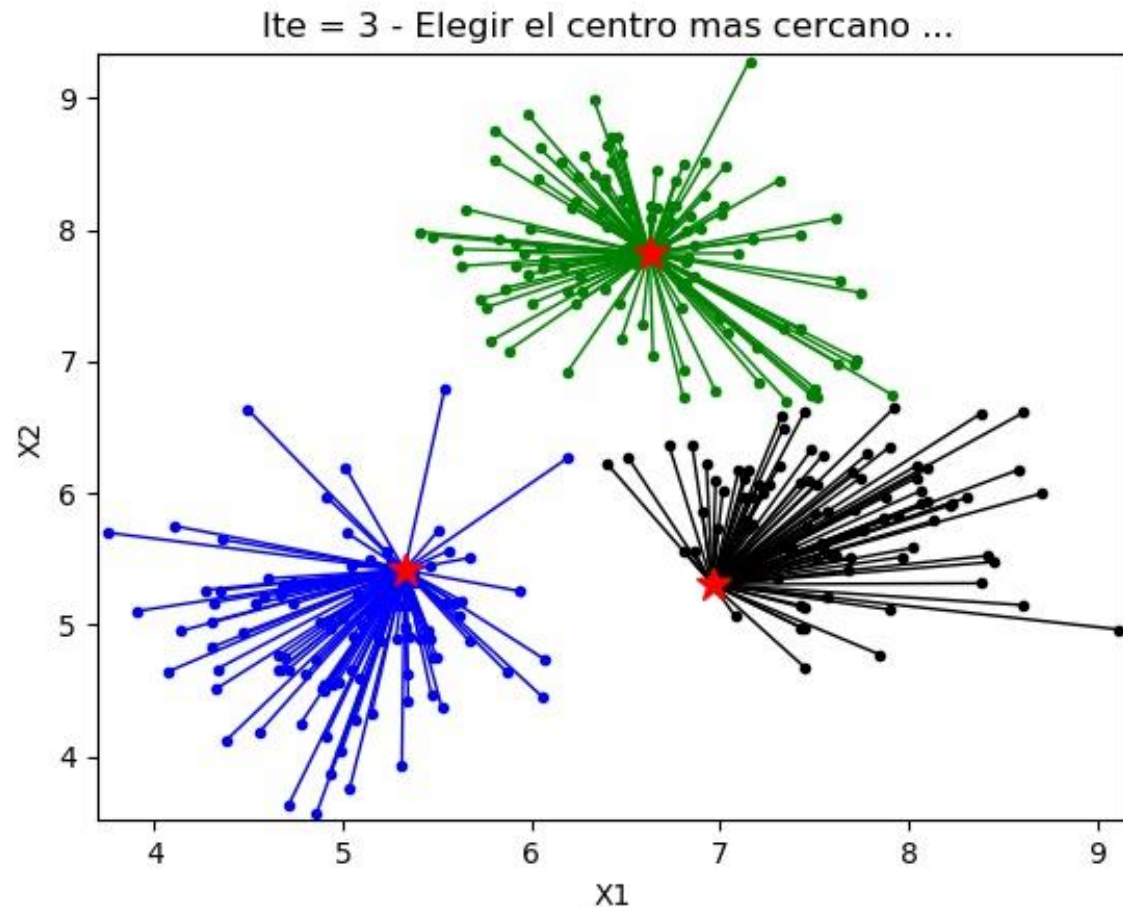
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



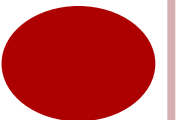
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



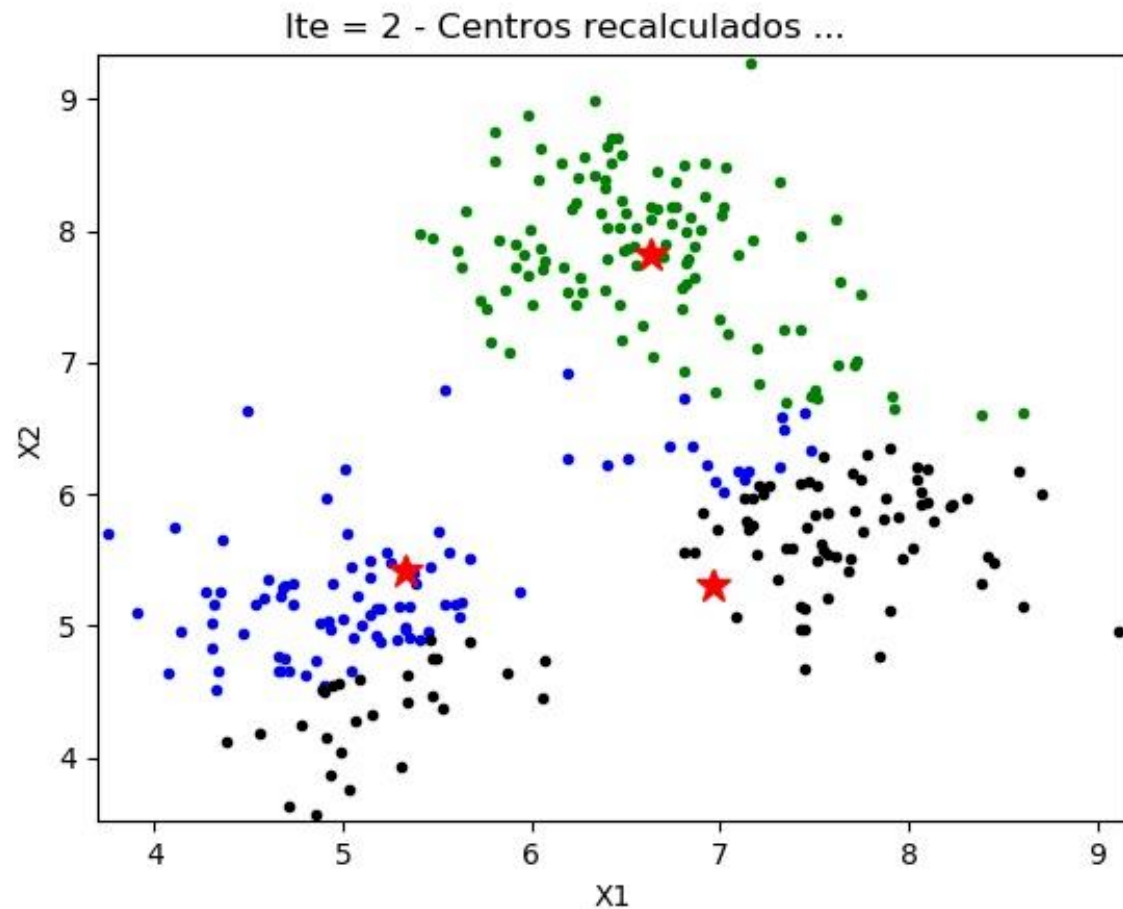
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



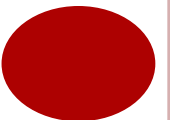
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



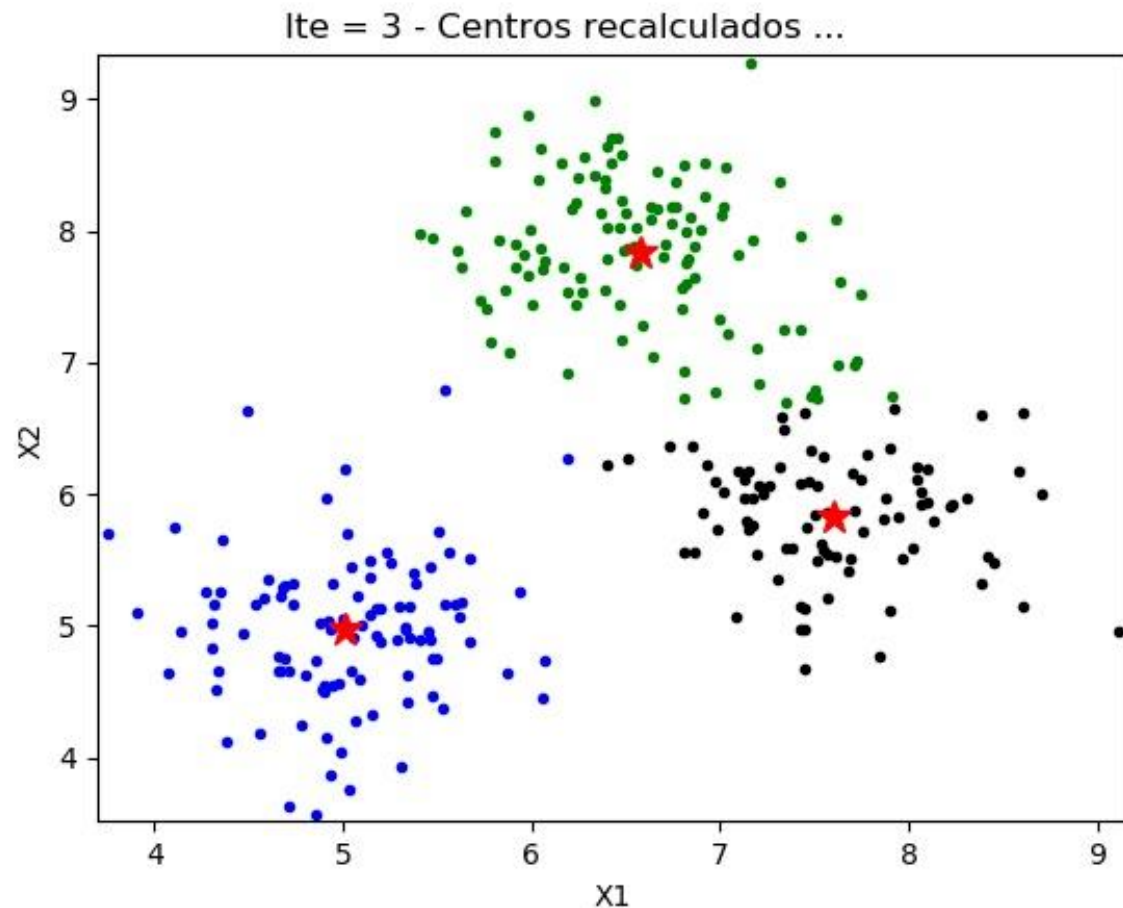
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



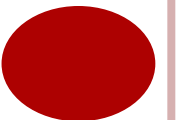
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



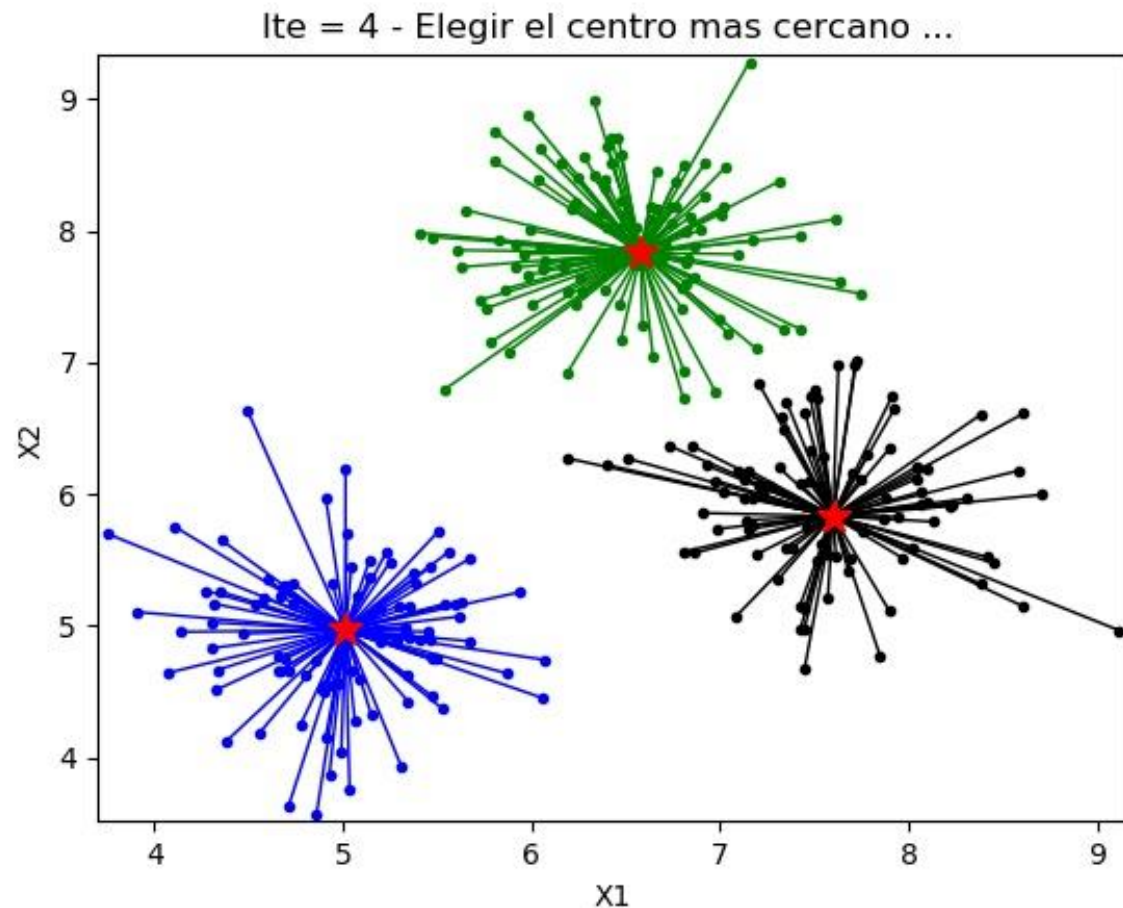
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



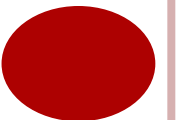
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



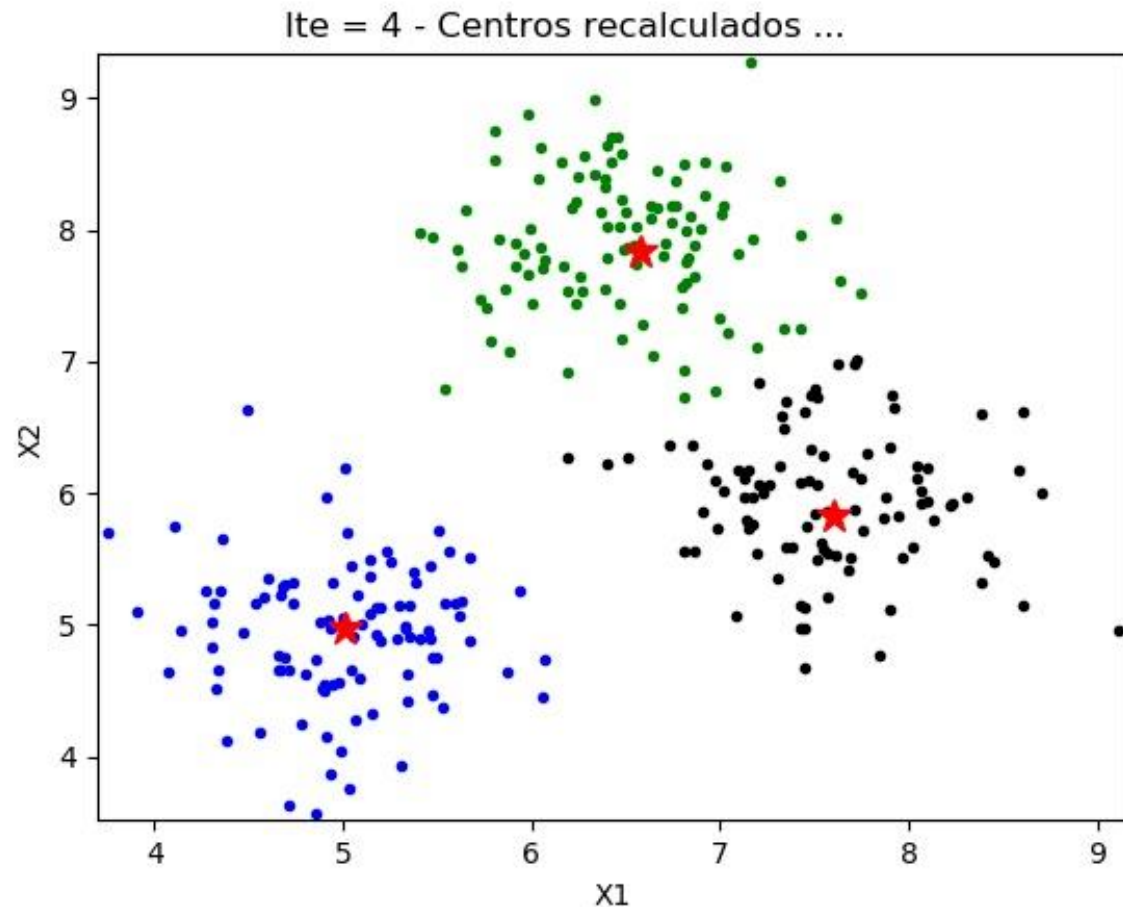
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



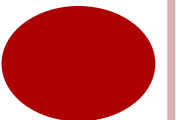
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



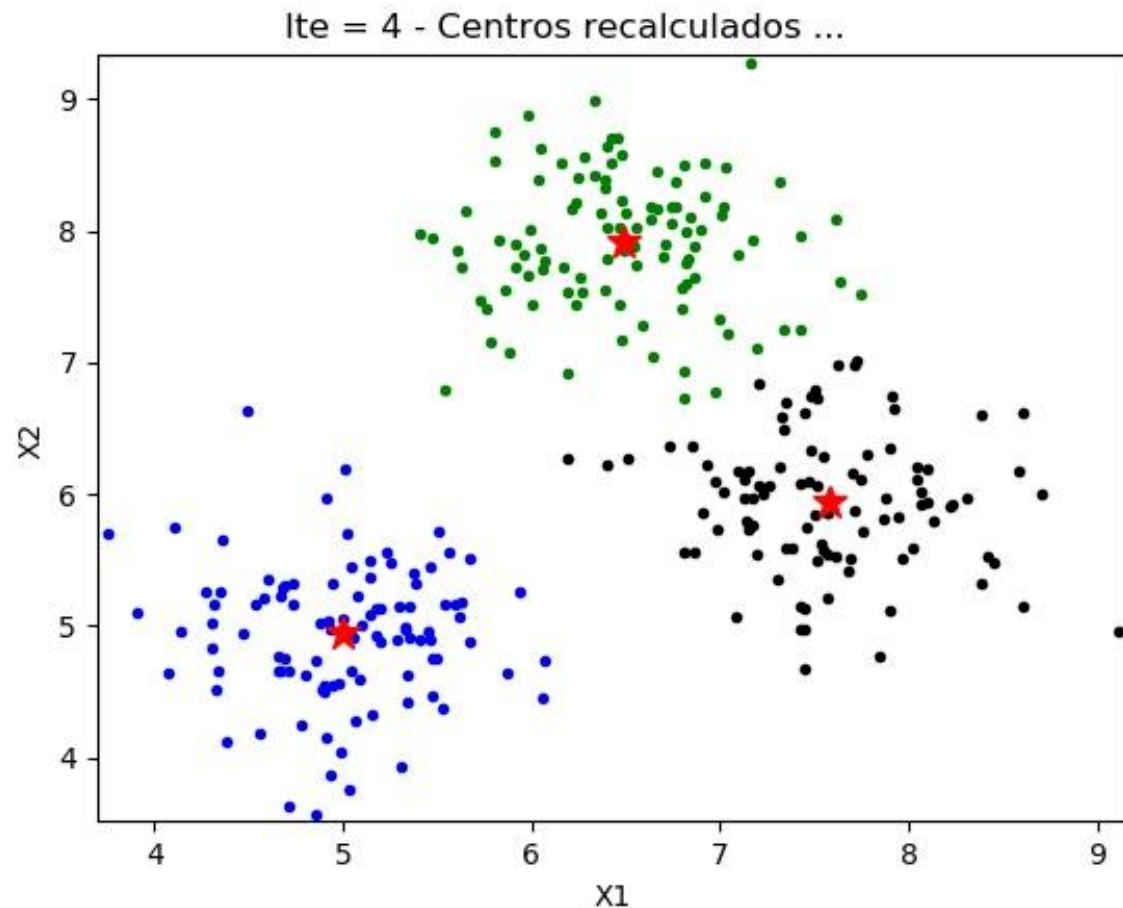
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



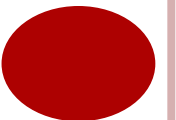
- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



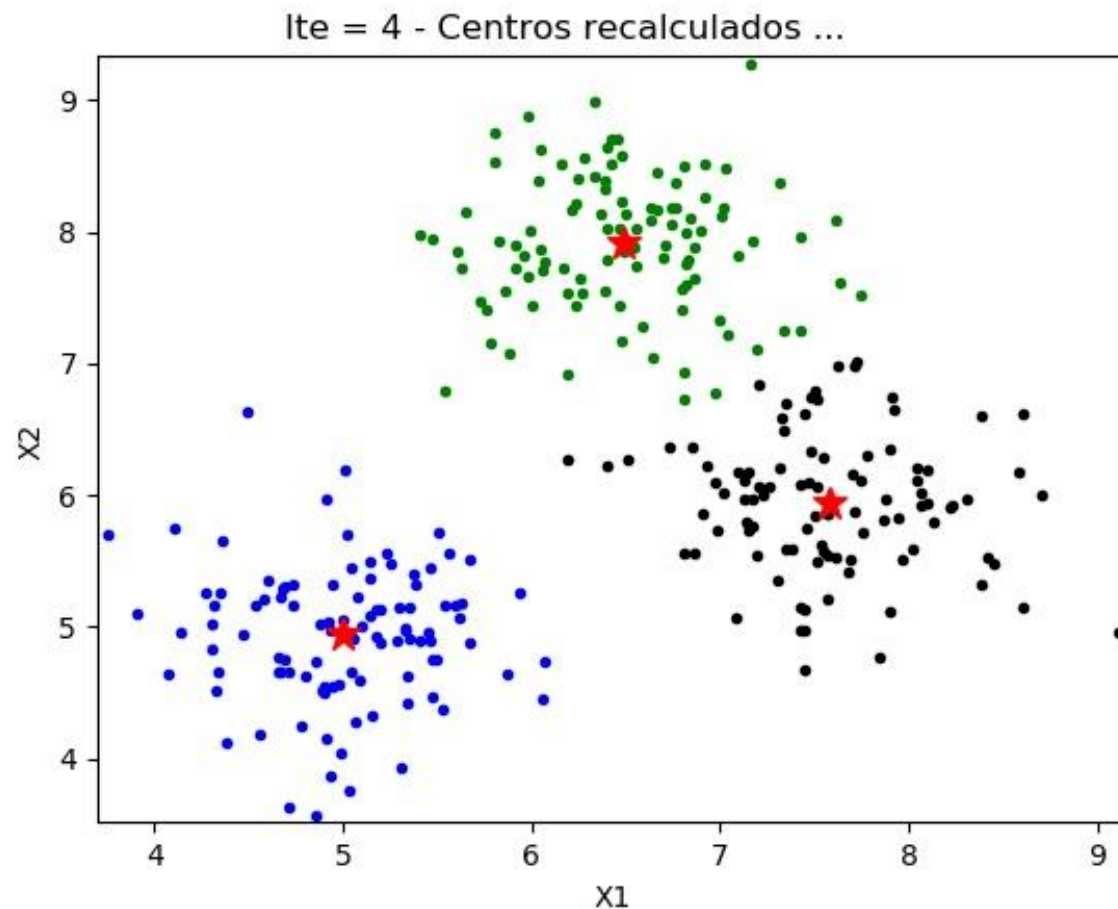
Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters



- Repetir ambos pasos
 - Asignar cada ejemplo al centro más cercano
 - Recalcular los centros hasta que los centros no cambien



Agrupando los ejemplos de **PuntosClusters.csv** usando k-medias con $k = 3$ clusters

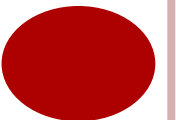


El modelo resultante
estará definido por los
centros (individuos
promedio) de cada grupo.

$$C_1 = (4.9972, 4.9380)$$

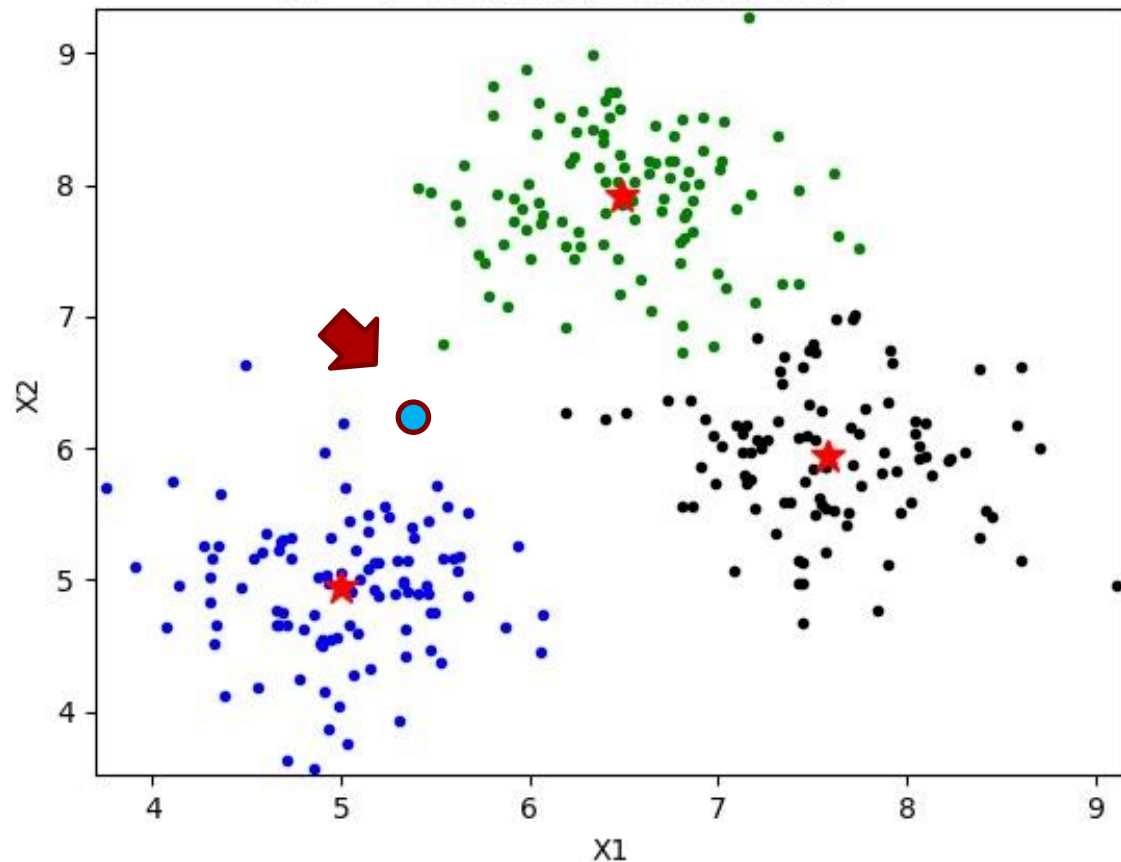
$$C_2 = (7.5672, 5.9504)$$

$$C_3 = (6.4782, 7.9324)$$

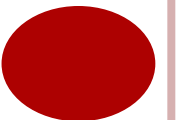


Utilizando el modelo obtenido con k-medias

- A qué grupo asignaría el nuevo ejemplo? A cuál de los 3 se parece más?



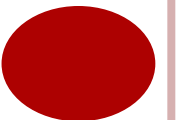
Calcular la distancia del nuevo ejemplo a cada centro y asignarlo al cluster más cercano



EJEMPLO DE K-MEANS USANDO CARTAS DE PÓKER



<https://www.youtube.com/watch?v=zHbxb2ye3E>

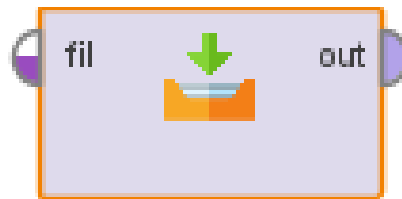


K-MEDIAS DESDE RAPIDMINER

Utilice los datos del
archivo
PuntosClusters.csv




PtosClusters.csv



res

Parameters

 PtosClusters.csv (Read CSV)

data set meta data information

 Edit List (3)...

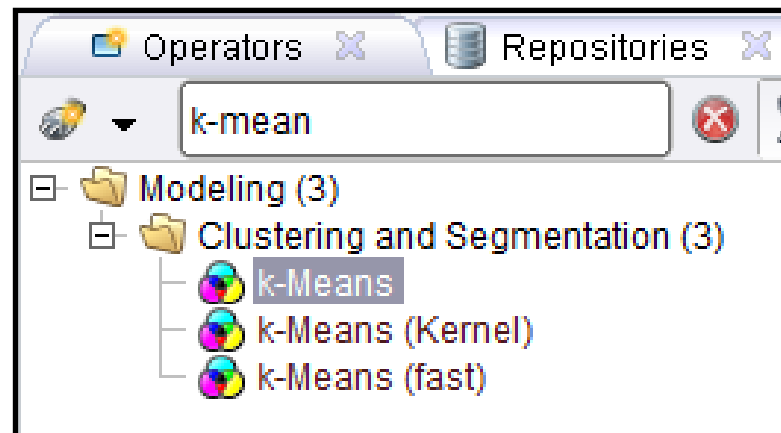
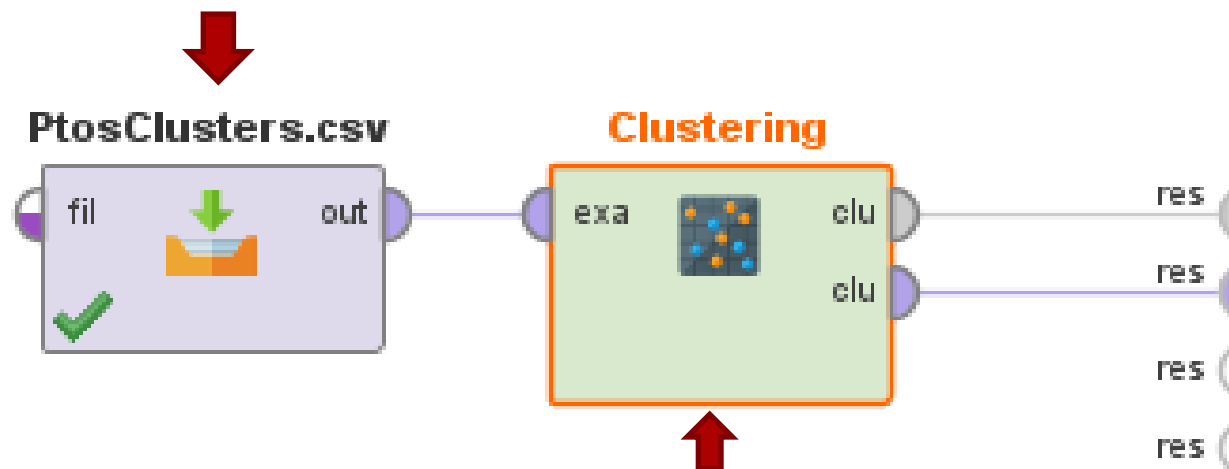


Edit Parameter List: **data set meta data information**
The meta data information

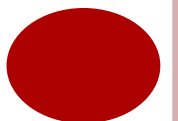
column index	attribute meta data information			
0	X1	<input checked="" type="checkbox"/> <i>column selected</i>	real ▼	attribute
1	X2	<input checked="" type="checkbox"/> <i>column selected</i>	real ▼	attribute
2	Clase	<input checked="" type="checkbox"/> <i>column selected</i>	polynomial ▼	label

K-MEDIAS DESDE RAPIDMINER

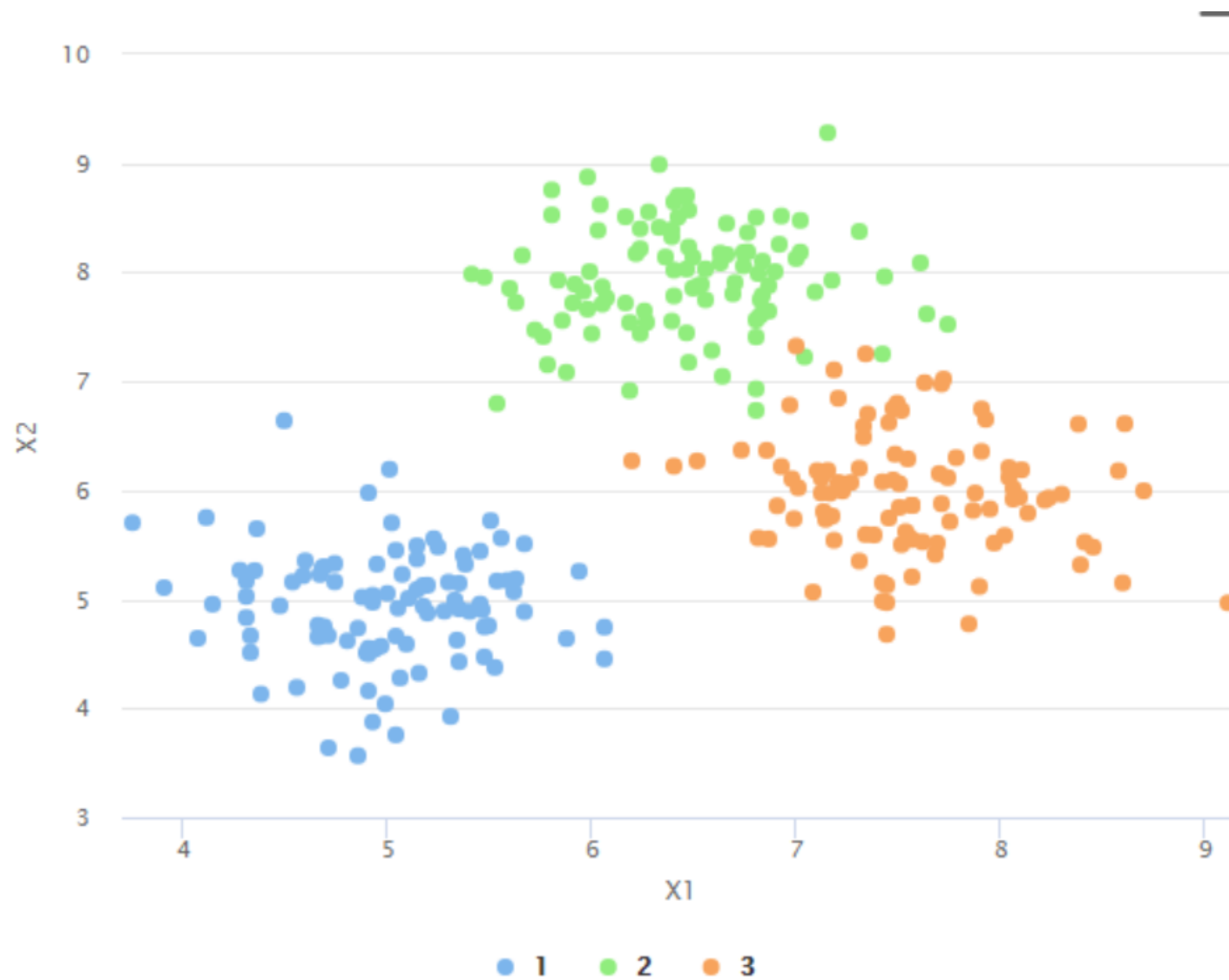
Utilice los datos del
archivo
PuntosClusters.csv



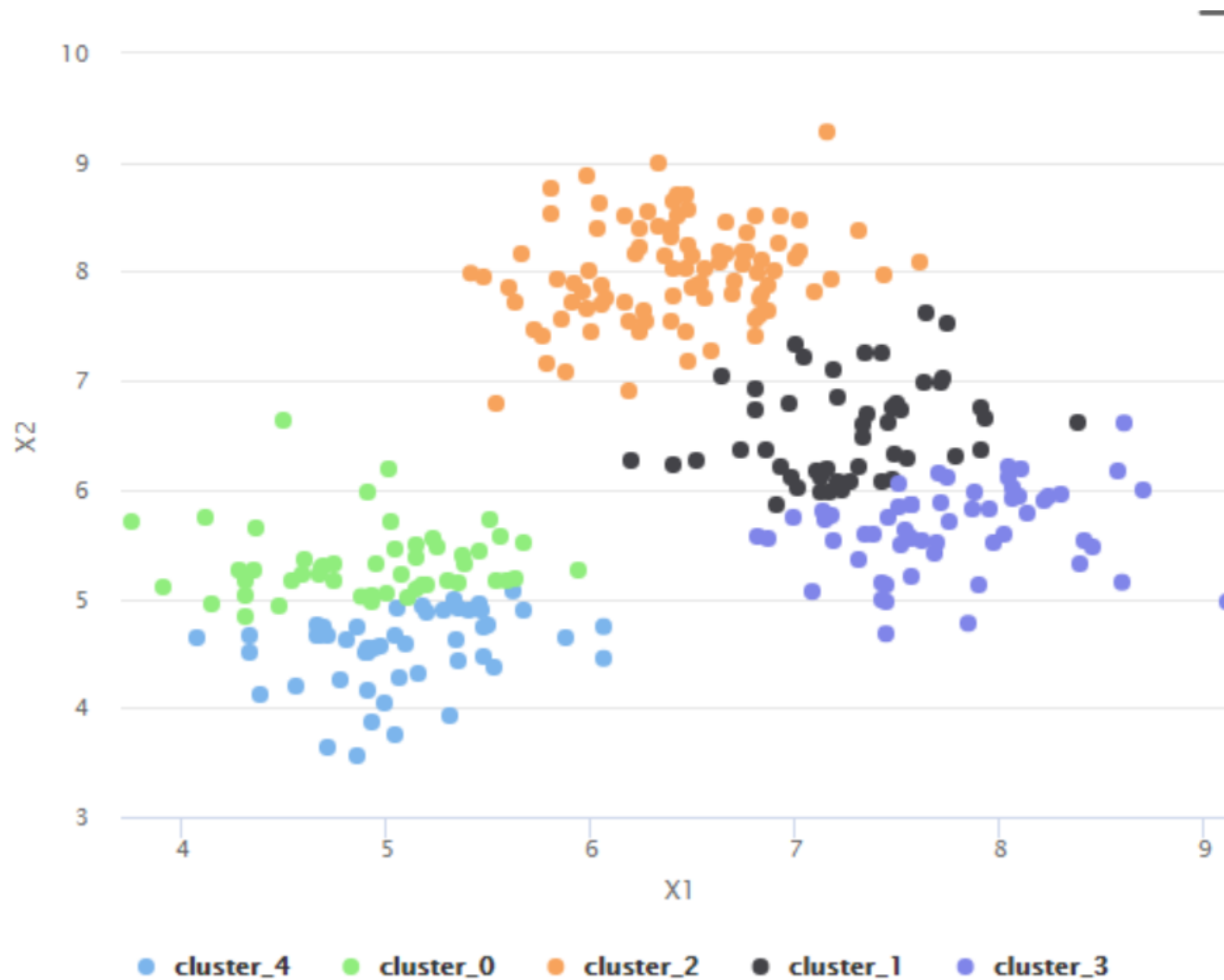
The screenshot shows the configuration panel for the 'Clustering (k-Means)' operator. The panel has a title bar with a cluster icon and the text 'Clustering (k-Means)'. Below the title bar, there are four checkboxes: 'add cluster attribute' (checked), 'add as label' (unchecked), 'remove unlabeled' (unchecked), and 'determine good start values' (checked). Below the checkboxes, there are two input fields: 'k' with the value '3' and 'max runs' with the value '100'.



K-MEDIAS DESDE RAPIDMINER



K-MEDIAS DESDE RAPIDMINER





K-MEDIAS DESDE RAPIDMINER

- Modelo obtenido como resultado del agrupamiento con $k=3$

Result History

Cluster Model (Clustering)


Description




Cluster Model

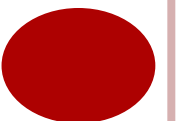
Cluster 0: 100 items
Cluster 1: 98 items
Cluster 2: 102 items
Total number of items: 300

Result History

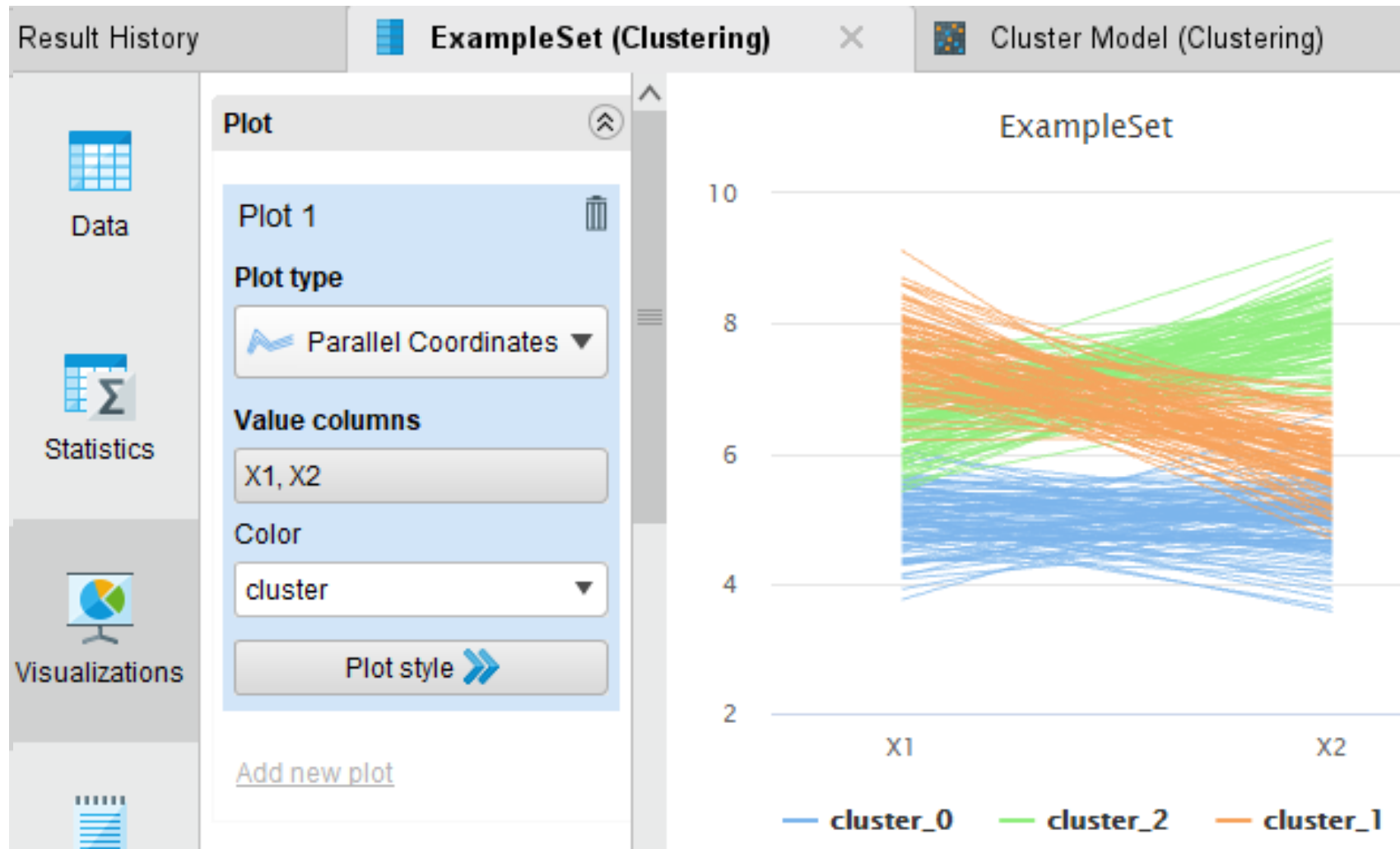
Cluster Model (Clustering) X


Centroid Table

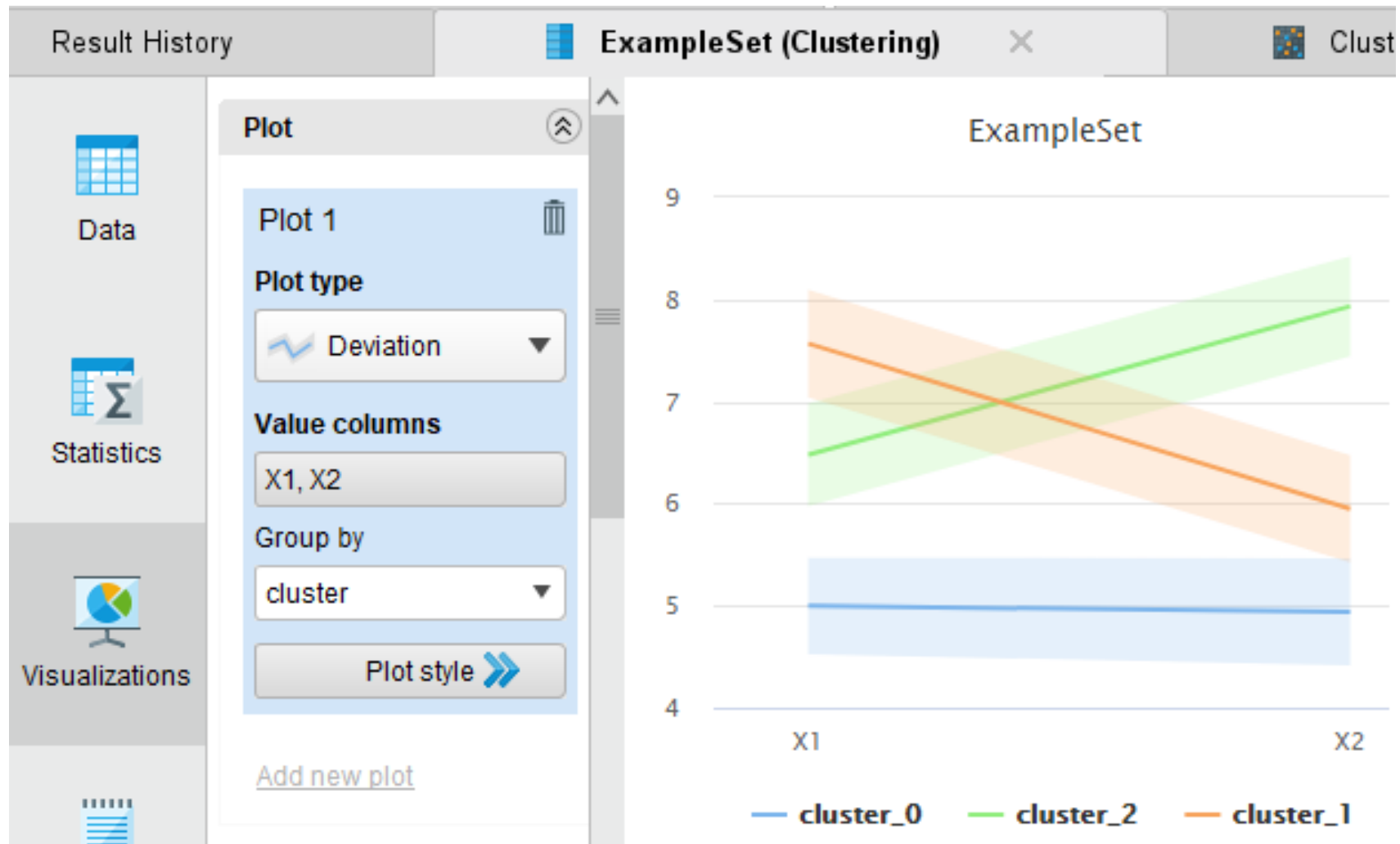
Attribute	cluster_0	cluster_1	cluster_2
X1	4.997	7.567	6.478
X2	4.938	5.950	7.932



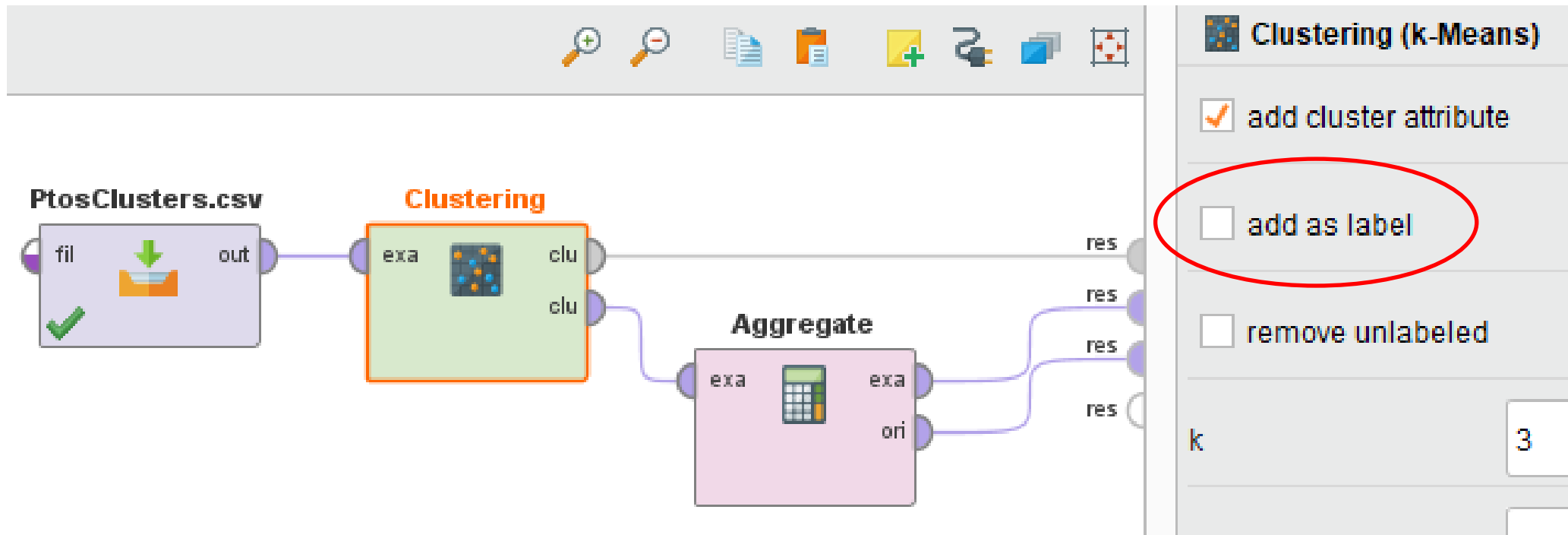
VISUALIZACIÓN DE LOS GRUPOS



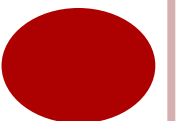
VISUALIZACIÓN DE LOS GRUPOS



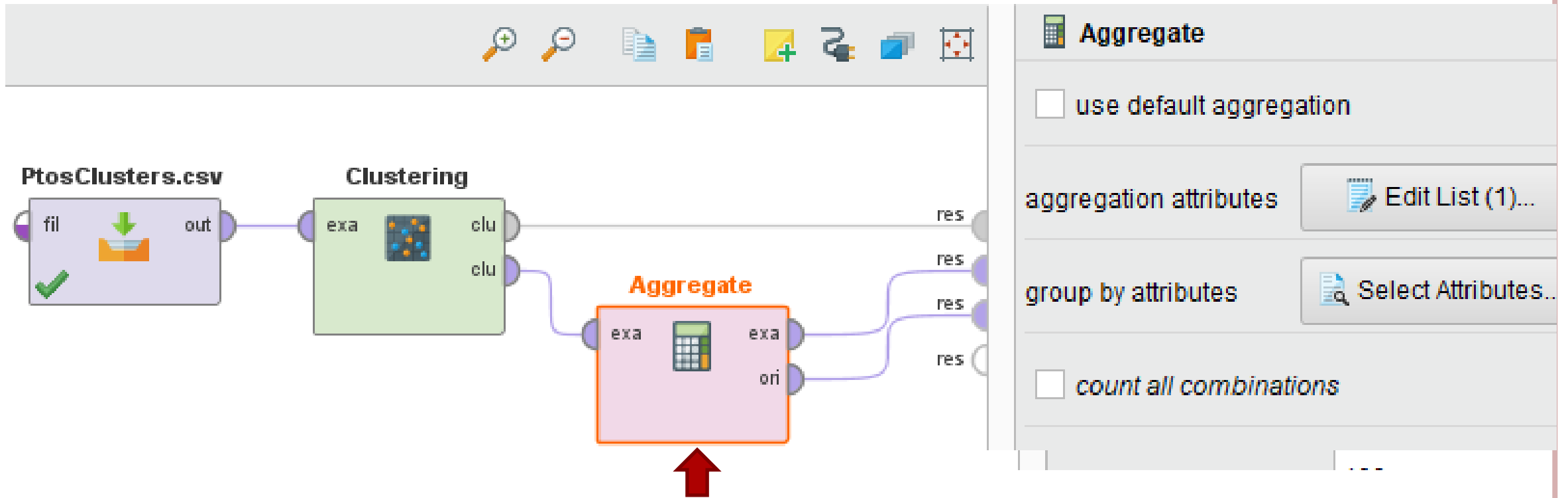
VISUALIZANDO LOS VALORES DE CLASE EN CADA GRUPO



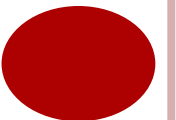
Es preciso disponer del label original y de la etiqueta de cluster por lo que no debe estar tildada esta opción



VISUALIZANDO LOS VALORES DE CLASE EN CADA GRUPO



Falta configurarlo



VISUALIZANDO LOS VALORES DE CLASE EN CADA GRUPO

Edit Parameter List: aggregation attributes
The attributes which should be aggregated.

aggregation attribute	aggregation functions
id	count

Add Entry

Remove Entry

Apply

Cancel

Aggregate

☐ use default aggregation

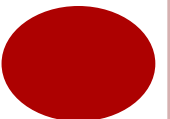
aggregation attributes

Edit List (1)...

group by attributes

Select Attributes..

☐ count all combinations



VISUALIZANDO LOS VALORES DE CLASE EN CADA GRUPO

Select Attributes: group by attributes

Select Attributes: **group by attributes**
Performs a grouping by the values of the attributes by the selected attributes.

Attributes

Search

id
X1
X2

Selected Attributes

Search

Clase
cluster

Apply Cancel

Aggregate




☐ use default aggregation

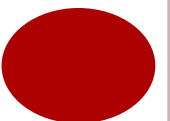
aggregation attributes [Edit List \(1\)...](#)

group by attributes [Select Attributes..](#)

☐ count all combinations

RESULTADO DE LA AGREGACIÓN

Result History				ExampleSet (Aggregate) X				ExampleSet			
 Data				Open in  Turbo Prep				 Auto Model			
Row No.	Clase	cluster ↑	count(id)								
1	1	cluster_0	100								
2	2	cluster_1	1								
4	3	cluster_1	97								
3	2	cluster_2	99								
5	3	cluster_2	3								



EJEMPLO 2 - AGRUPANDO FLORES DE IRIS

- Se dispone de información 3 tipos de flores Iris



Setosa

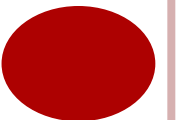


Versicolor



Virginica

<https://archive.ics.uci.edu/ml/datasets/Iris>

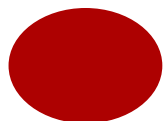
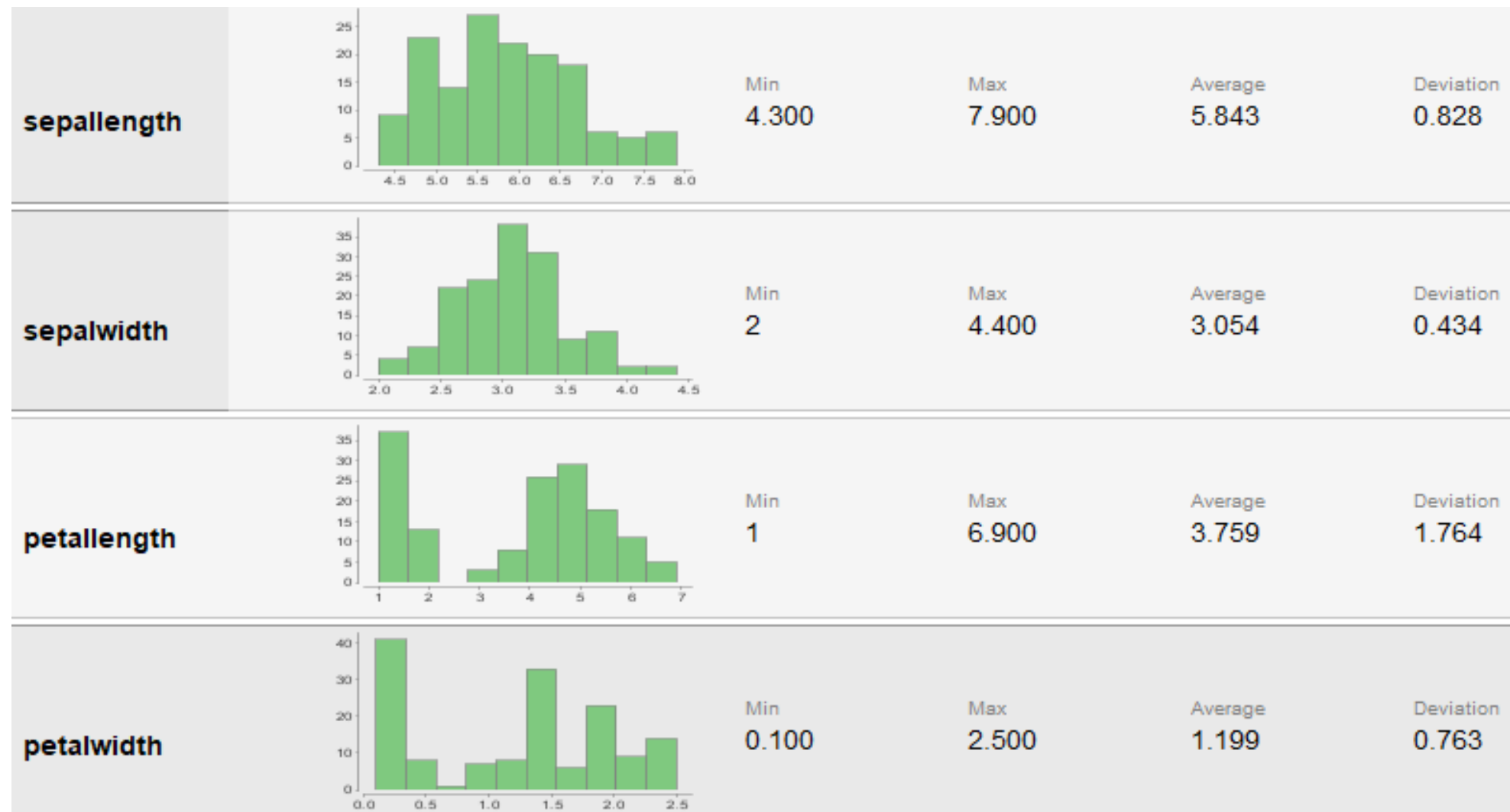


EJEMPLO 2 - AGRUPANDO FLORES DE IRIS

Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica

<https://archive.ics.uci.edu/ml/datasets/Iris>

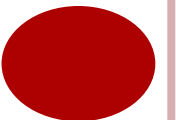
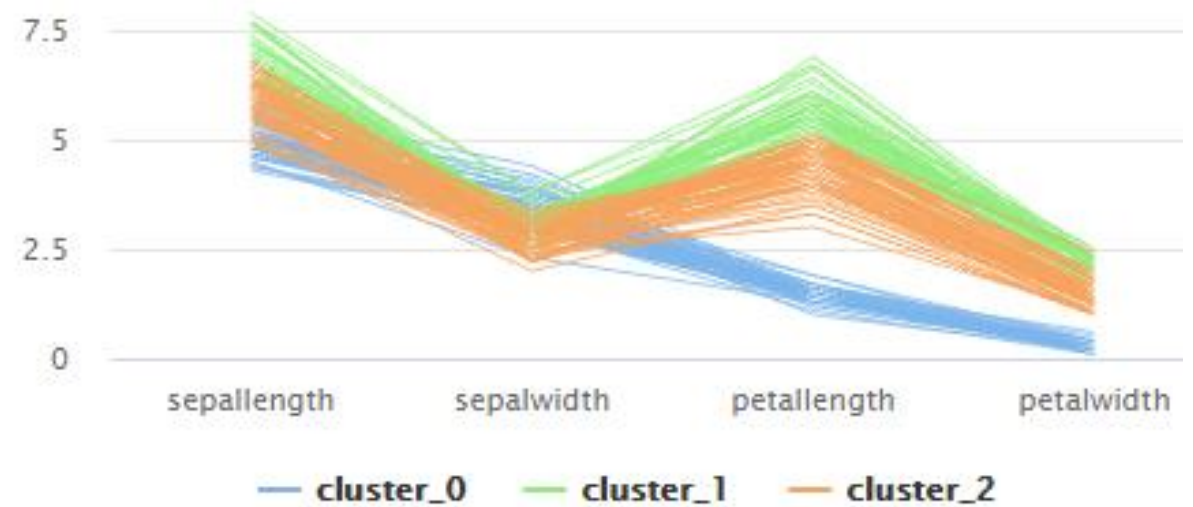
EJEMPLO 2 – CARACTERÍSTICAS DE FLORES DE IRIS



EJEMPLO 2 - AGRUPANDO FLORES DE IRIS

- Se busca identificar atributos con valores distintos entre grupos

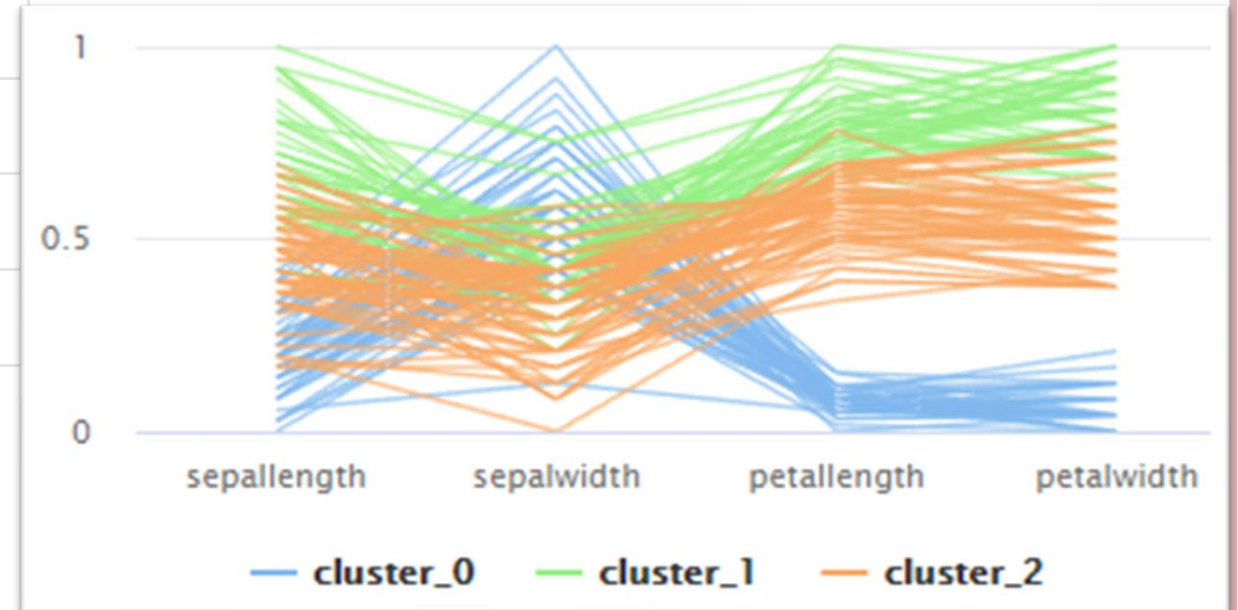
Attribute	cluster_0	cluster_1	cluster_2
sepalength	5.006	6.854	5.884
sepalwidth	3.418	3.077	2.741
petallength	1.464	5.715	4.389
petalwidth	0.244	2.054	1.434



EJEMPLO 2 - AGRUPANDO CON K-MEDIAS (K=3)

- Normalizando linealmente entre 0 y 1 los valores de cada atributo ANTES de agrupar se obtiene lo siguiente:

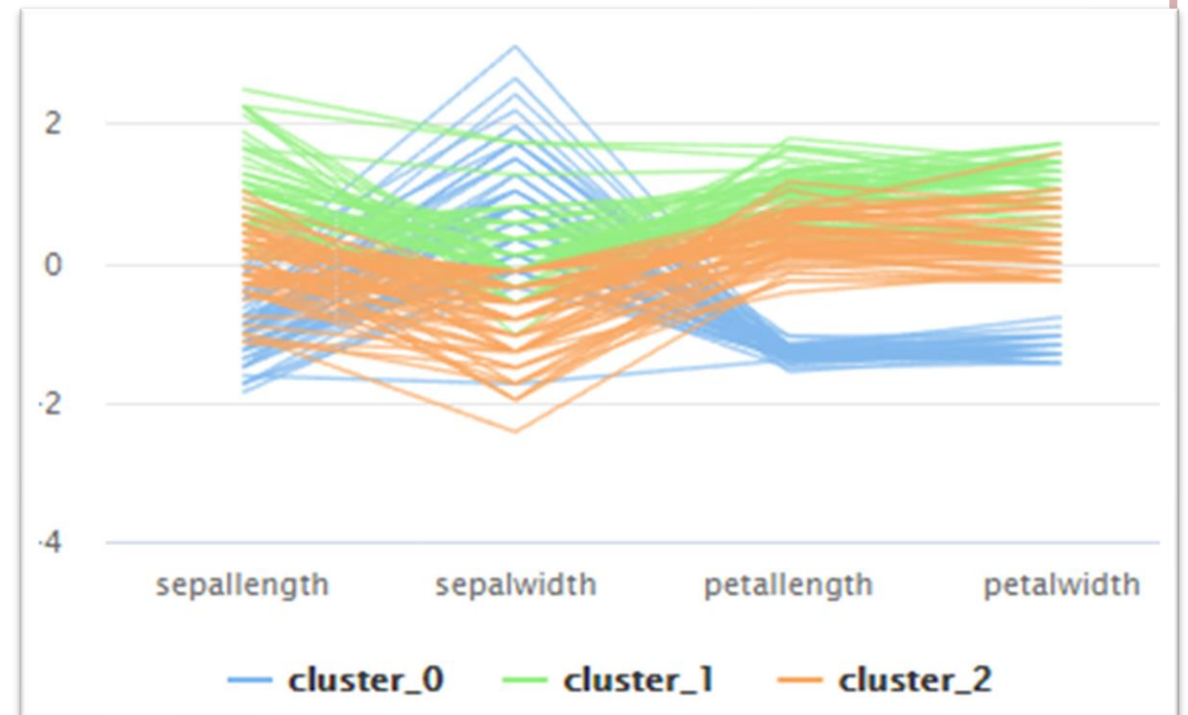
Attribute	cluster_0	cluster_1	cluster_2
sepallength	0.196	0.707	0.441
sepalwidth	0.591	0.451	0.307
petallength	0.079	0.797	0.576
petalwidth	0.060	0.825	0.549



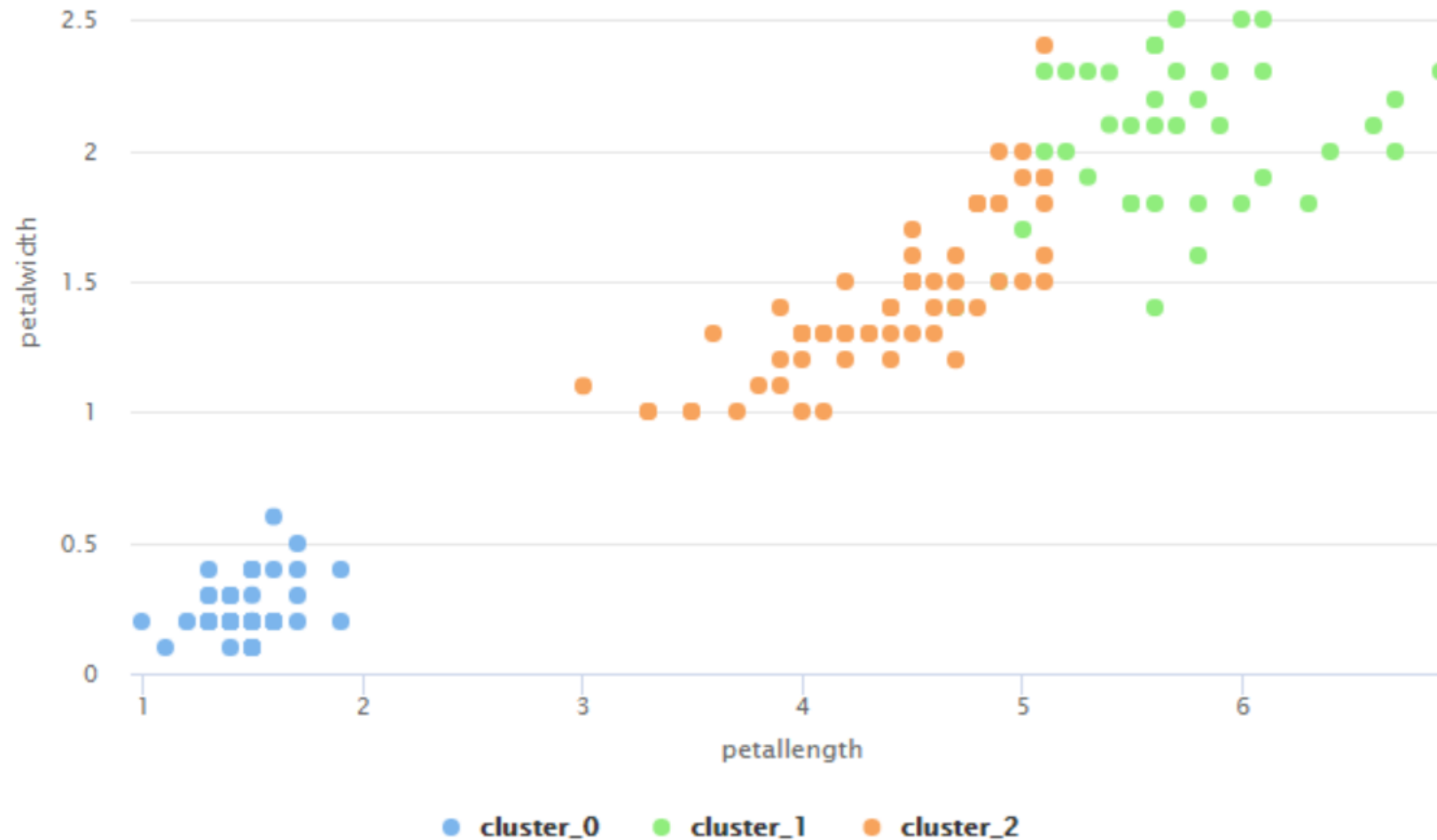
EJEMPLO 2 - AGRUPANDO CON K-MEDIAS (K=3)

- Normalizando los atributos ANTES de agrupar (transformación Z de Rapid Miner) se obtiene lo siguiente:

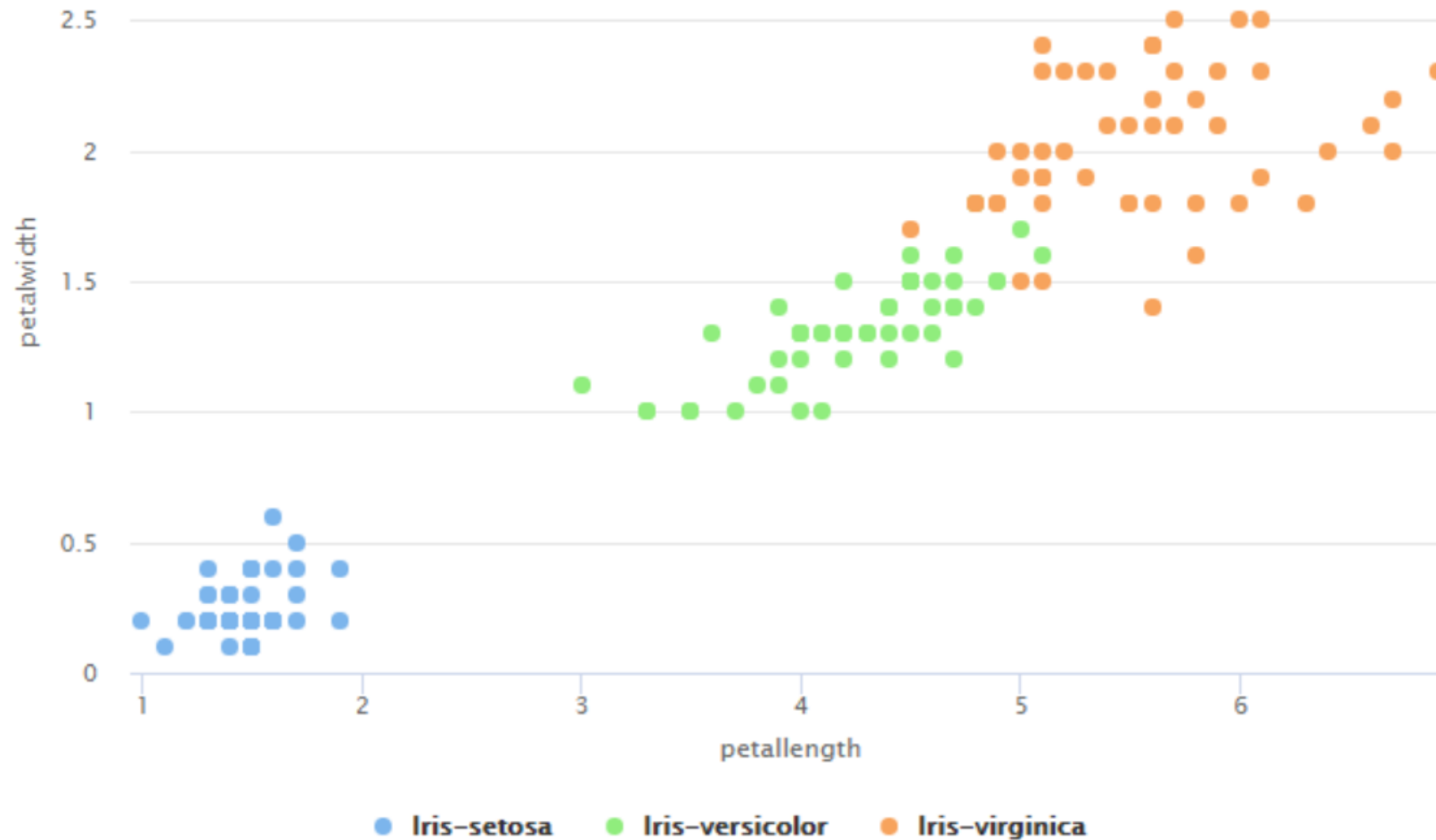
Attribute	cluster_0	cluster_1	cluster_2
sepallength	-1.011	1.164	-0.011
sepalwidth	0.839	0.153	-0.870
petallength	-1.301	1.000	0.376
petalwidth	-1.251	1.026	0.311



EJEMPLO 2 - AGRUPANDO FLORES DE IRIS

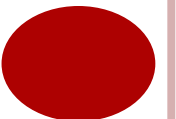


EJEMPLO 2 - AGRUPANDO FLORES DE IRIS



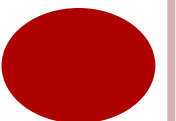
EJERCICIO

- El archivo **SEMILLAS.csv** contiene información de granos que pertenecen a tres variedades de trigo: Kama, Rosa y Canadiense.
- Para cada grano se midieron las siguientes características:
 - área A,
 - perímetro P,
 - compacidad $C = 4 * \pi * A / P^2$,
 - longitud del núcleo,
 - ancho del núcleo,
 - coeficiente de asimetría
 - longitud del surco del núcleo
- Describa los tipos de semillas inspeccionados utilizando el algoritmo K-medias



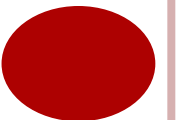
EJERCICIO

- El archivo **MachineCPU.csv** contiene datos de rendimiento relativo de la CPU.
- Los atributos relevados fueron los siguientes:
 - tiempo de ciclo de máquina en nanosegundos (MYCT)
 - memoria principal mínima en kilobytes (MMIN)
 - memoria principal máxima en kilobytes (MMAX)
 - memoria caché en kilobytes (CACH)
 - canales mínimos en unidades (CHMIN) y canales máximos en unidades (CHMAX).
 - rendimiento relativo publicado (PRP) de la CPU.
- Agrupe utilizando k-means con $K=2, 3$ y 4 . Observe la cantidad de ejemplos por grupo y analice los valores de los centroides obtenidos.



EJERCICIO

- El archivo **Vinos.csv** contiene los resultados de un análisis químico de vinos cultivados en la misma región de Italia pero procedentes de tres cultivares diferentes.
- El análisis determinó las cantidades de 13 constituyentes que se encuentran en cada uno de los tres tipos de vinos.
- Agrupe las muestras usando k-medias con $K=3$. Relacione los grupos obtenidos con los tres tipos de vinos.



TIPOS DE ALGORITMOS DE AGRUPAMIENTO

○ Algoritmo Partitivo

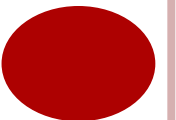
- Particionan los datos creando un número K de clusters.
- Una instancia pertenece a un único grupo.

○ Algoritmo Jerárquico

- Generan una estructura jerárquica de clusters que permiten ver las particiones de las instancias con distinta granularidad.
- Una instancia pertenece a un único grupo.

○ Algoritmo probabilista

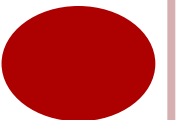
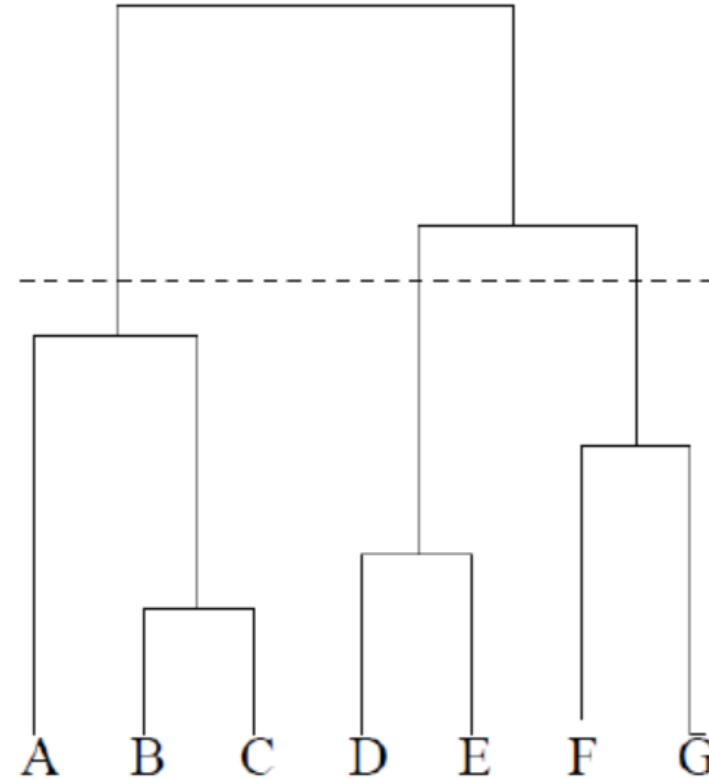
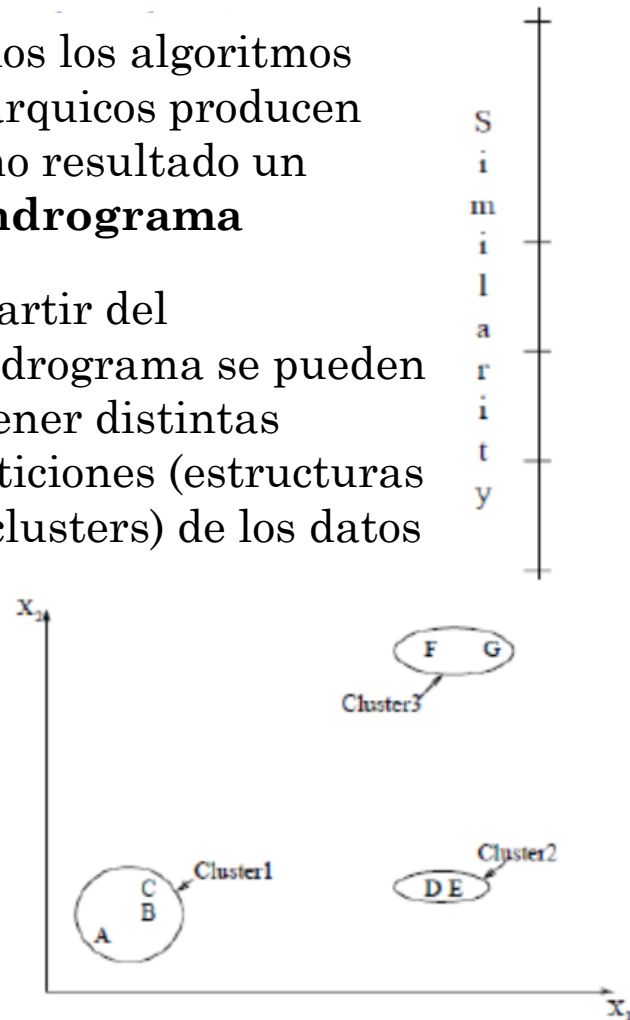
- Los clusters se generan con un método probabilístico



ALGORITMO DE CLUSTERING JERÁRQUICOS

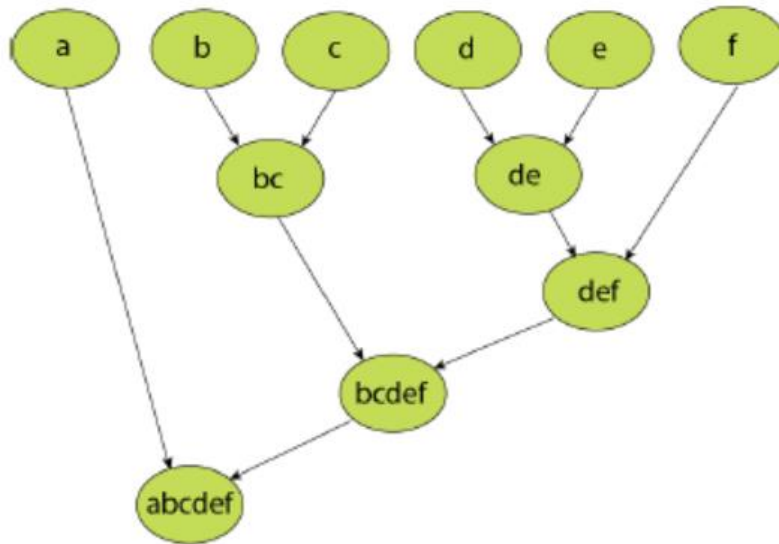
Todos los algoritmos jerárquicos producen como resultado un **dendrograma**

A partir del dendrograma se pueden obtener distintas particiones (estructuras de clusters) de los datos

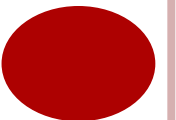
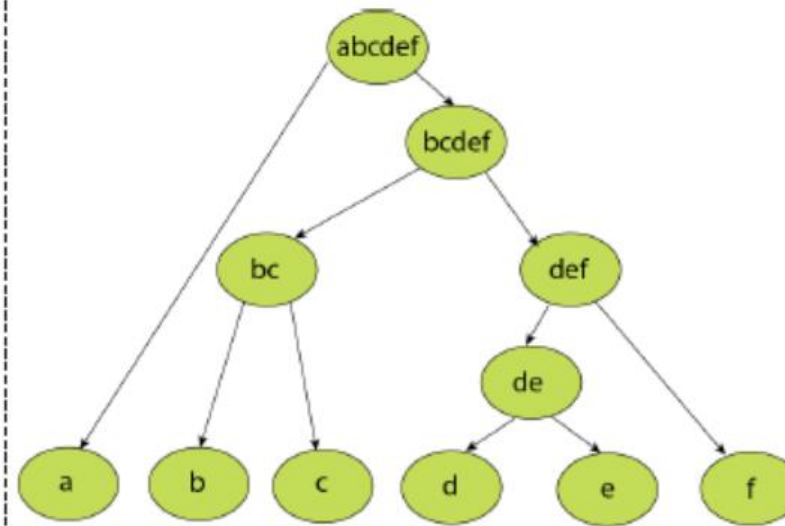


ALGORITMO DE CLUSTERING JERÁRQUICOS

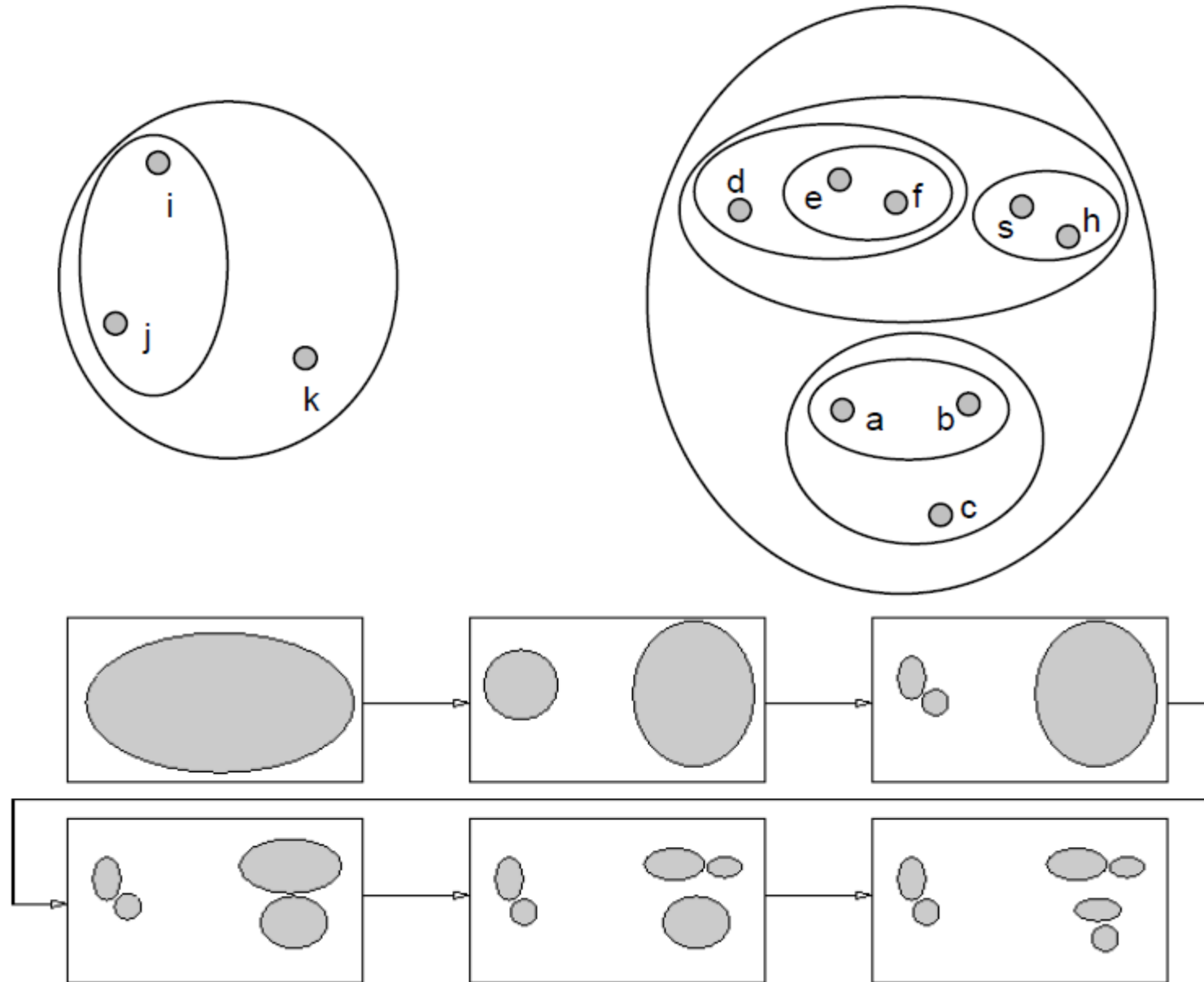
Aglomerativo



Divisible

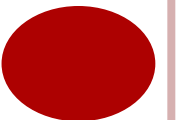


AGRUPAMIENTO JERARQUICO



ALGORITMO JERÁRQUICO AGLOMERATIVO

- **Paso 1:** A cada instancia se le asigna un cluster, de modo que inicialmente si hay N instancias se tienen N clusters de 1 elemento cada uno.
- **Paso 2:** Calcular la **distancia entre clusters** y unir en uno solo a los dos más cercanos.
- **Paso 3:** Calcular la distancia del nuevo cluster a los restantes.
- **Paso 4:** Repetir los pasos 2 y 3 hasta que todas las instancias pertenezcan al mismo cluster

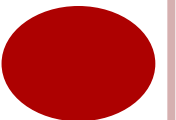
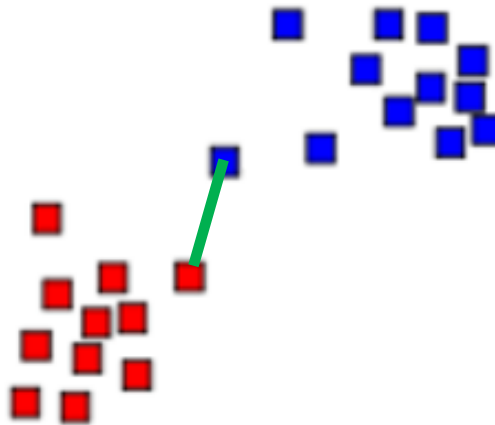


MEDIDAS DE CONECTIVIDAD (*LINKAGE MEASURES*)

○ Enlace simple (*single-linkage*)

- La similitud entre dos clusters se calcula como la similitud de los **dos puntos más cercanos** pertenecientes a los diferentes clusters.

$$\min\{dist(a, b): a \in A, b \in B\}$$

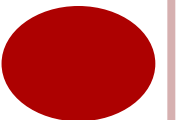
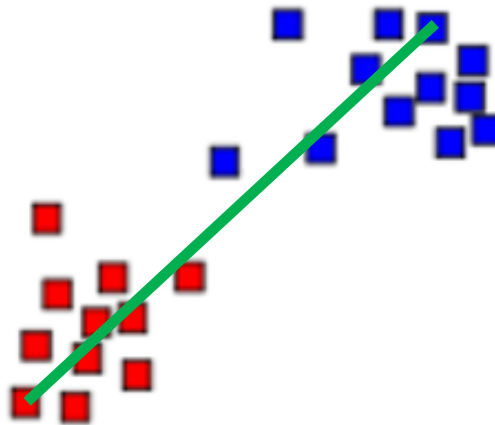


MEDIDAS DE CONECTIVIDAD (*LINKAGE MEASURES*)

○ Enlace completo (*complete-linkage*)

- La similitud entre dos clusters se calcula como la similitud de los **dos puntos más lejanos** pertenecientes a los diferentes clusters.

$$\max\{dist(a, b) : a \in A, b \in B\}$$



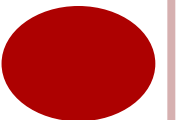
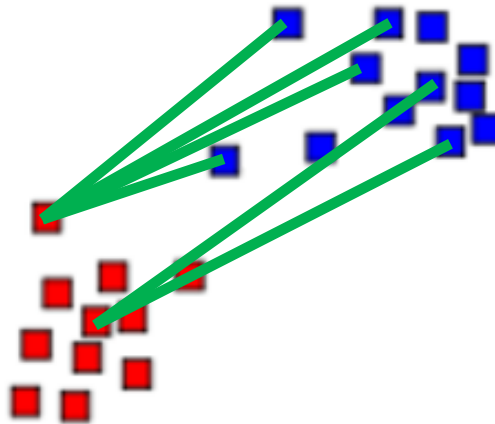
MEDIDAS DE CONECTIVIDAD

(*LINKAGE MEASURES*)

○ Enlace promedio (*average-linkage*)

- La distancia entre dos grupos se calcula promediando las distancias entre todos los pares que se puedan formar tomando una instancia de cada cluster.

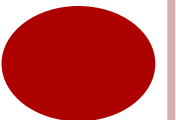
$$\frac{1}{|A| |B|} \sum_{a \in A} \sum_{b \in B} \text{dist}(a, b)$$



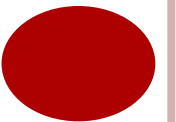
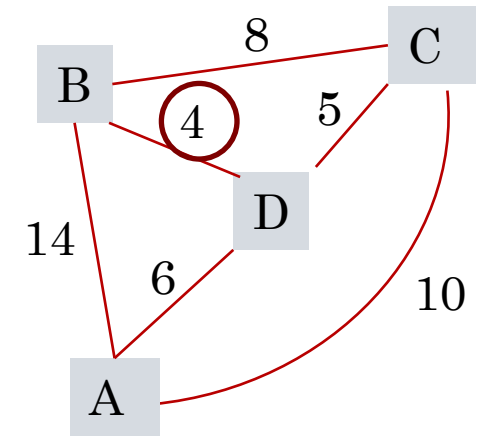
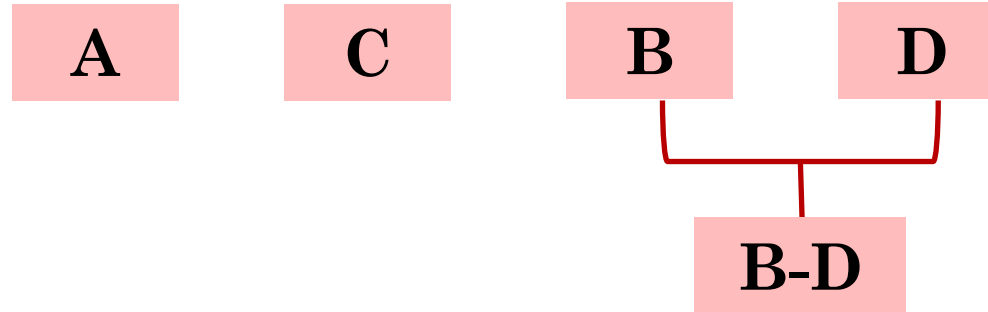
EJEMPLO

- Aplique un algoritmo jerárquico aglomerativo para agrupar las instancias A, B, C y D cuya matriz de distancias se indica a continuación

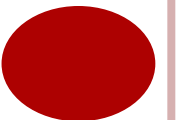
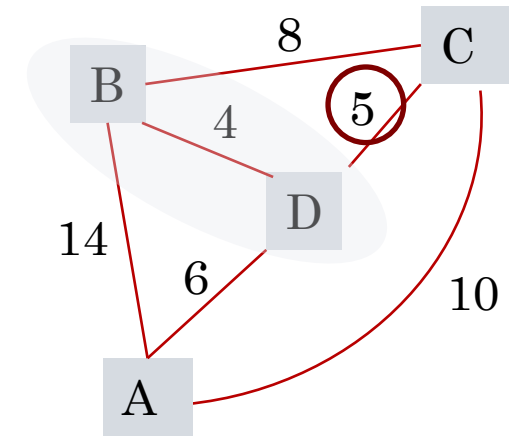
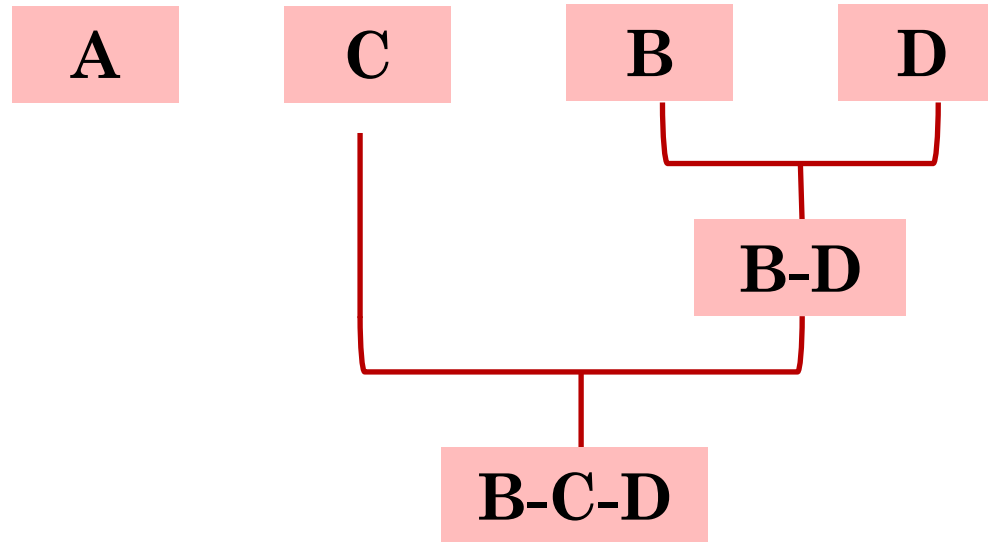
	A	B	C	D
A	0	14	10	6
B	14	0	8	4
C	10	8	0	5
D	6	4	5	0



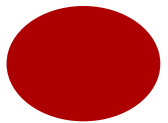
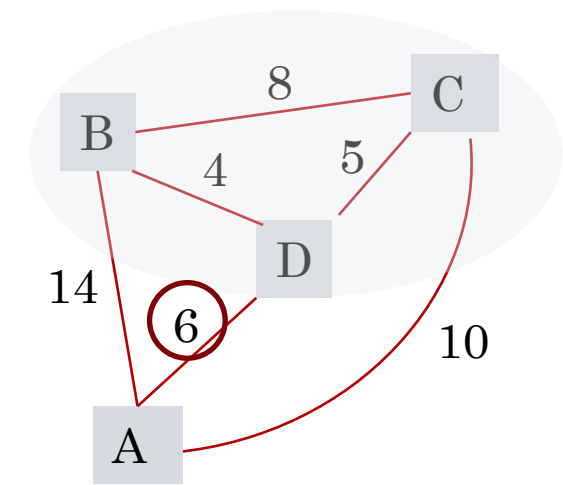
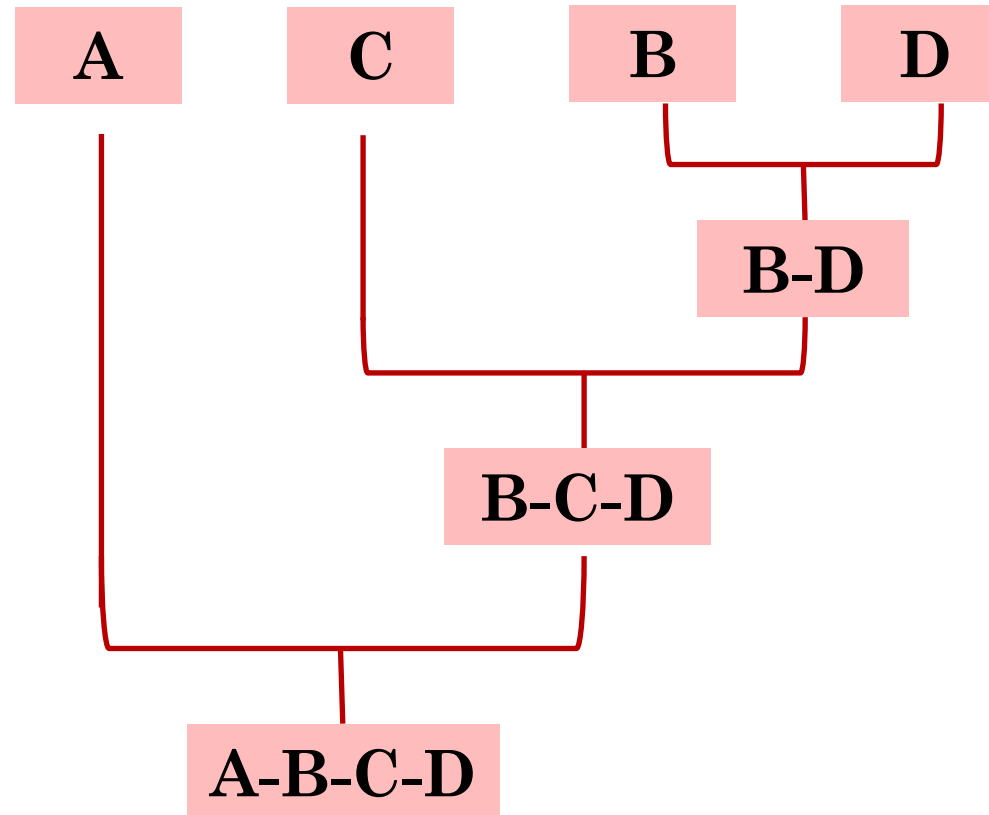
EJEMPLO USANDO ENLACE SIMPLE



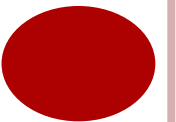
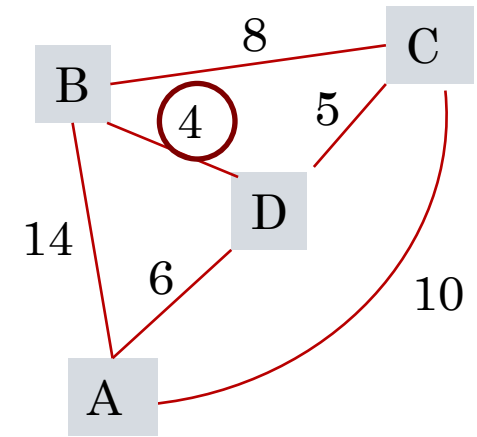
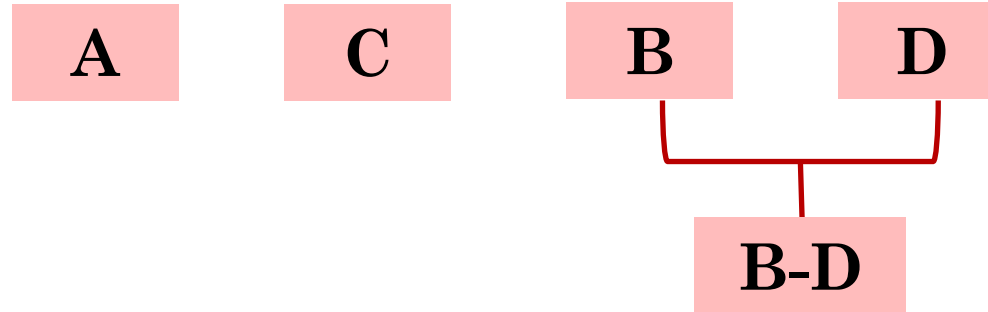
EJEMPLO USANDO ENLACE SIMPLE



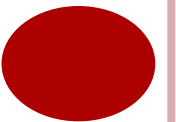
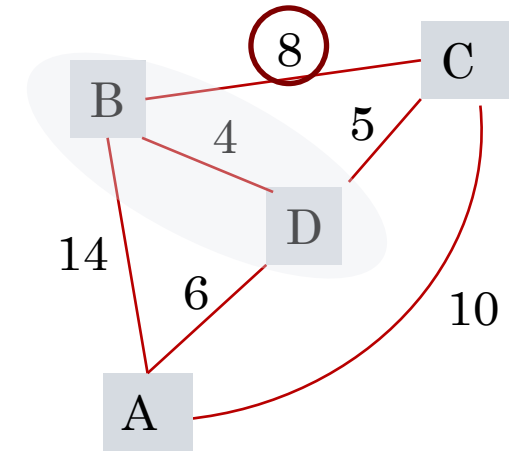
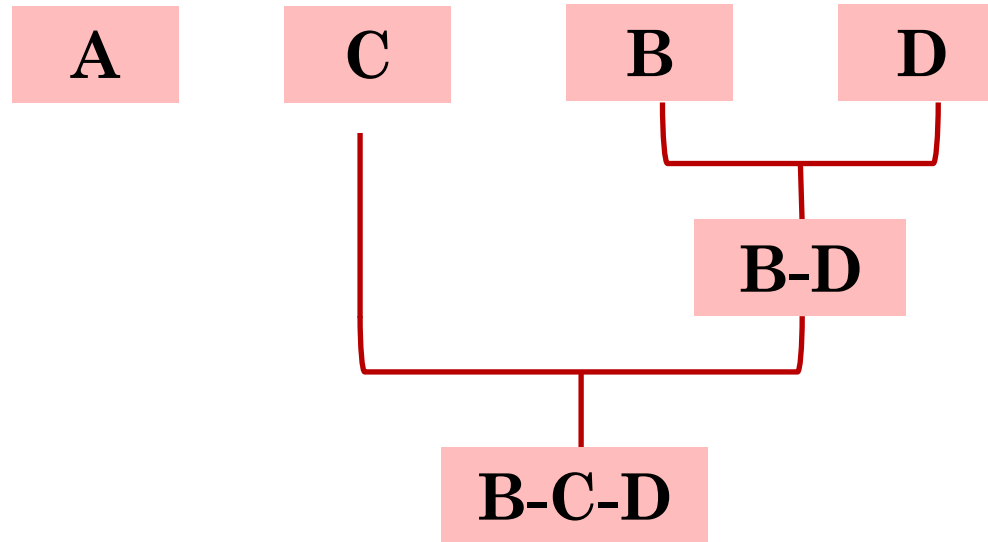
EJEMPLO USANDO ENLACE SIMPLE



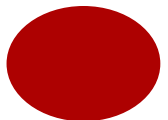
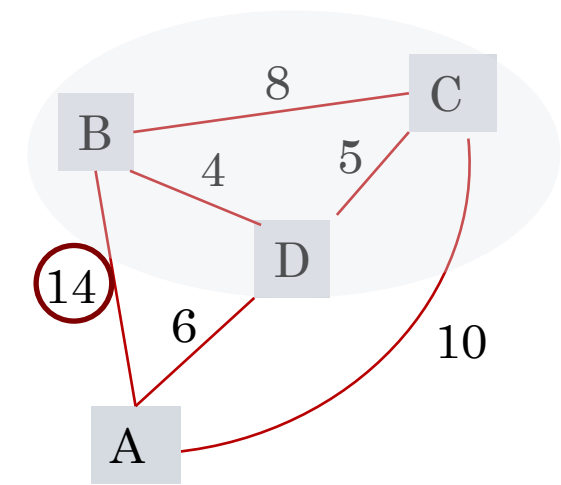
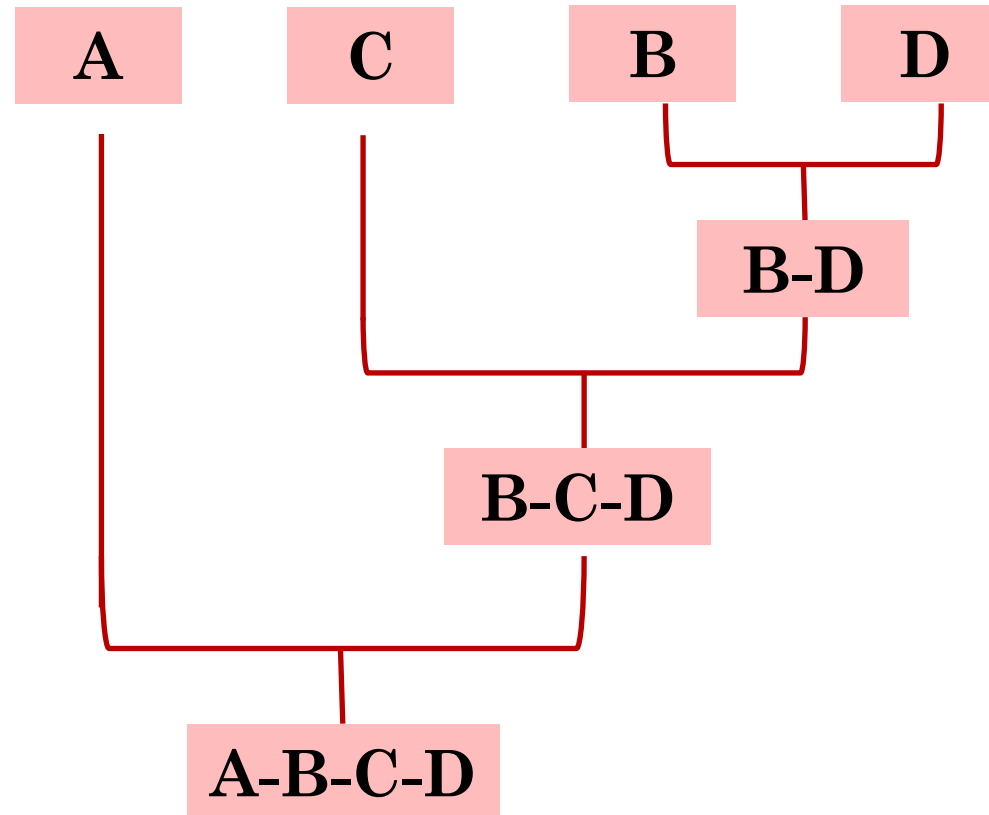
EJEMPLO USANDO ENLACE COMPLETO



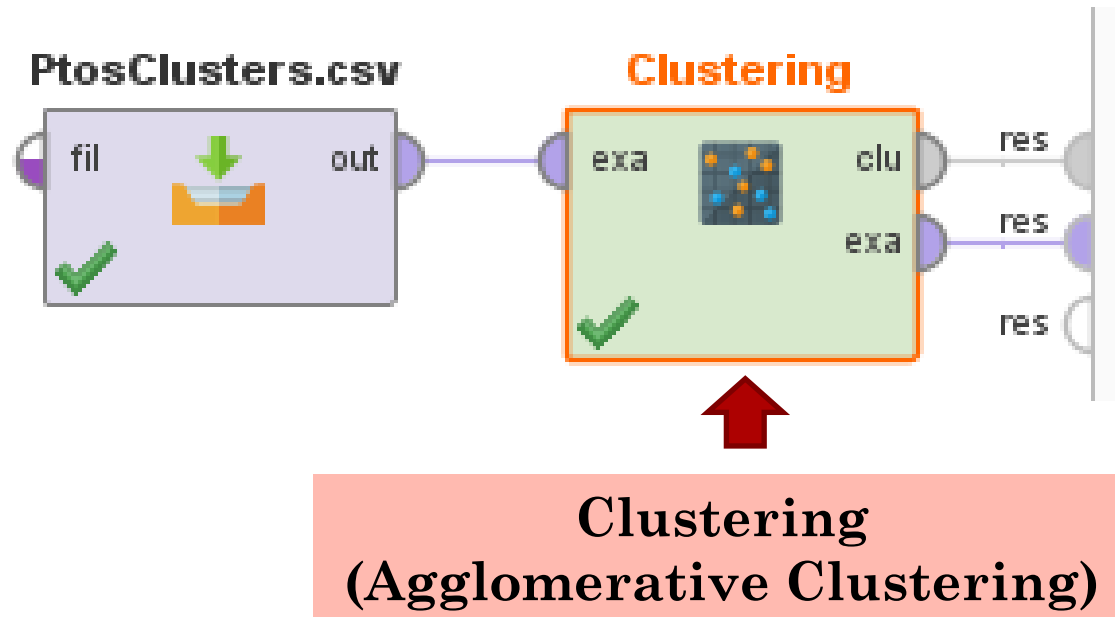
EJEMPLO USANDO ENLACE COMPLETO



EJEMPLO USANDO ENLACE COMPLETO



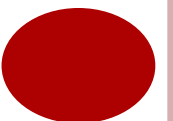
CLUSTERING JERÁRQUICO CON RAPID MINER



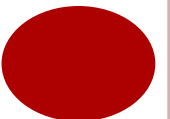
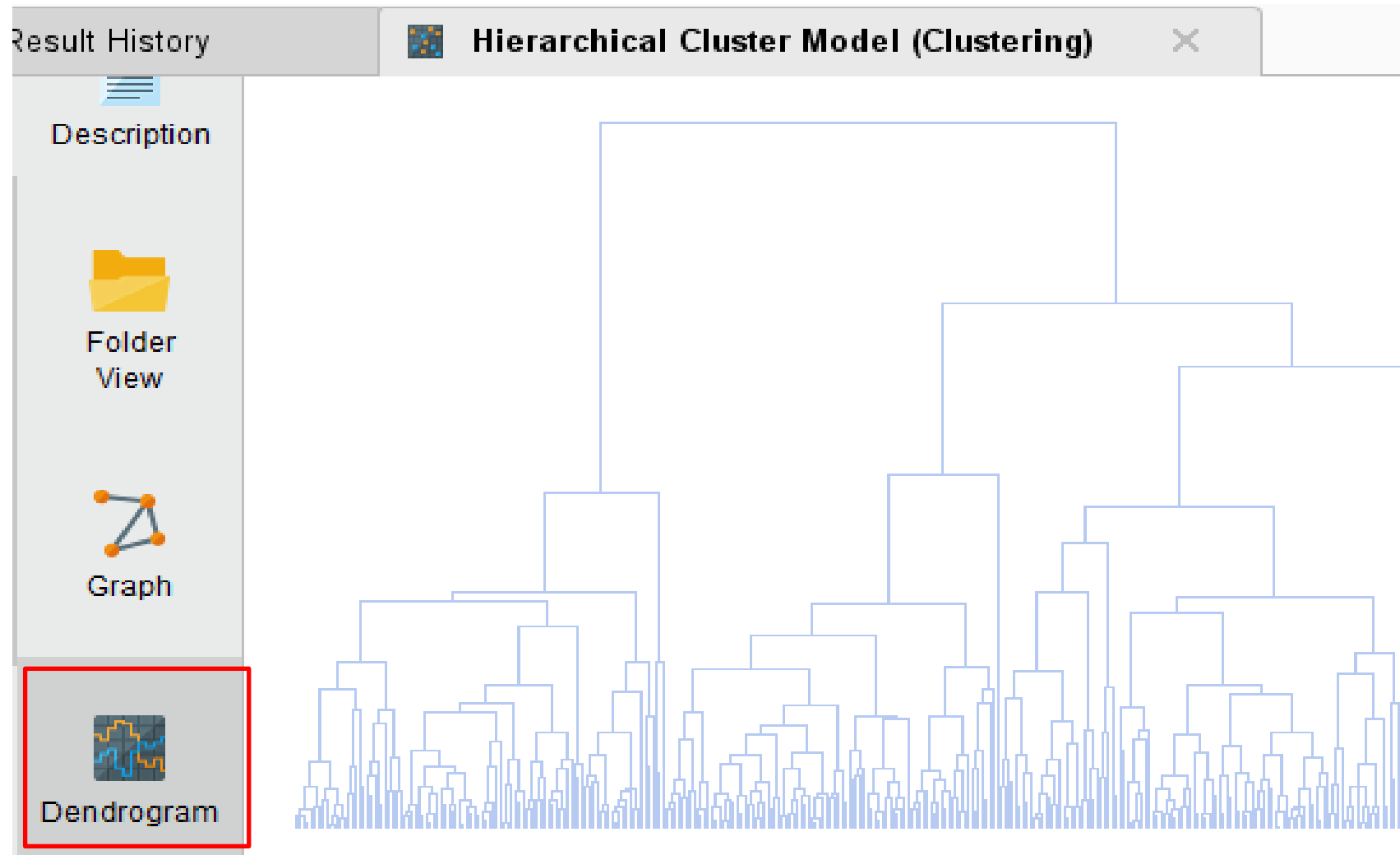
Parameters [X]

Clustering (Agglomerative Clustering)

mode	AverageLink ▼
measure types	SingleLink CompleteLink AverageLink
mixed measure	MixedEuclidean... ▼

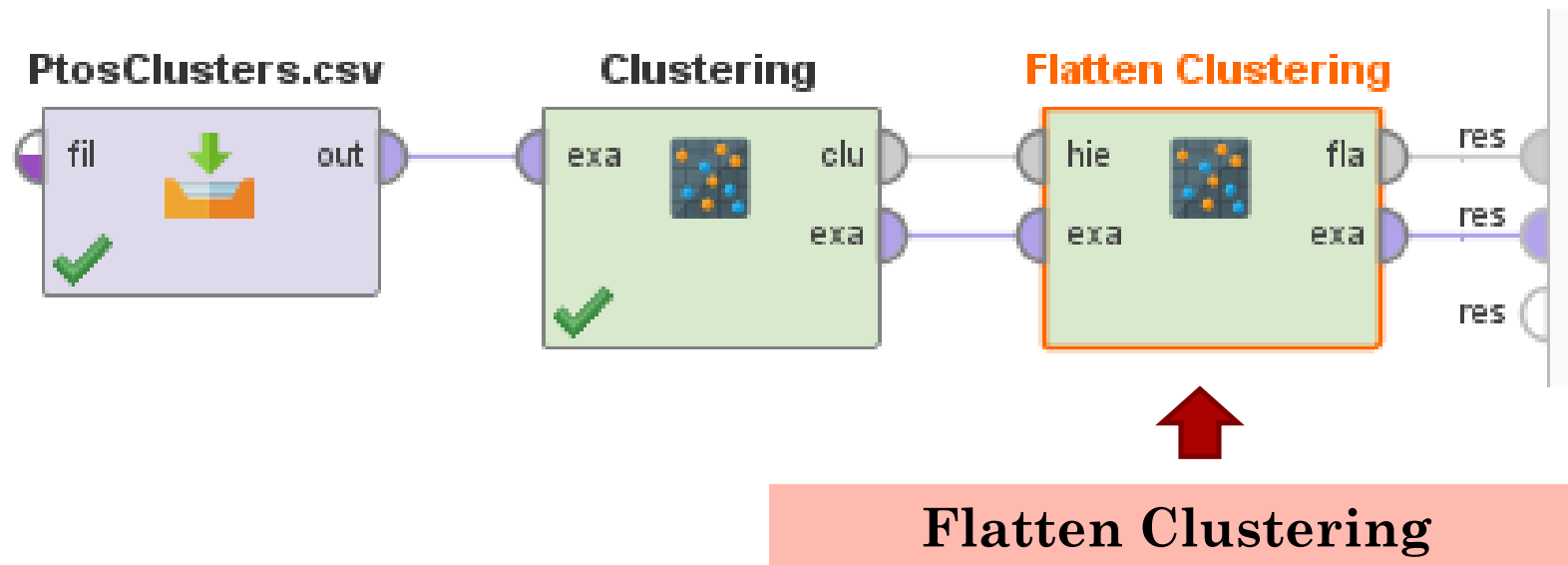


CLUSTERING JERÁRQUICO CON RAPID MINER

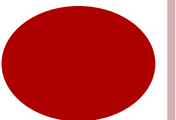


CLUSTERING JERÁRQUICO CON RAPID MINER

- Puede seleccionarse la cantidad de grupos para cortar el dendrograma

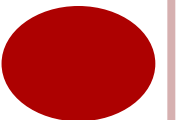


Probar con distintos números de clusters y observar los resultados



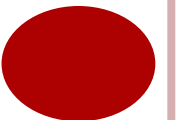
AGRUPAMIENTO PROBABILISTA

- Los algoritmos basados en probabilidad, en lugar de asociar una instancia a un único cluster, utilizan la probabilidad de **que las instancias pertenezcan a cada uno de los clusters**.
- Utilizan un conjunto de k distribuciones de probabilidad para representar k clusters.
- *El problema es determinar los parámetros que modelan las distribuciones*



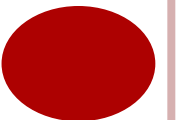
ALGORITMO EM (EXPECTATION - MAXIMIZATION)

- Estimar los valores de los parámetros que modelan los clustes.
- Repetir hasta que no mejoren los agrupamientos
 - **Esperanza (*expectation*):** Utiliza los valores de los parámetros, iniciales o proporcionados por el paso *Maximization*, obteniendo diferentes formas de la *fdp* (función de densidad de probabilidad) buscada.
 - **Maximización (*maximization*):** Recalcular los parámetros de la distribución haciendo uso de los datos proporcionados por el paso anterior.

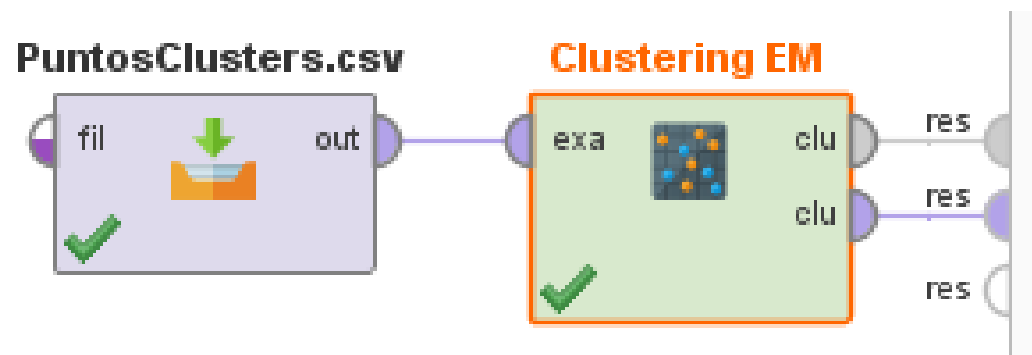


ALGORITMO EM (EXPECTATION - MAXIMIZATION)

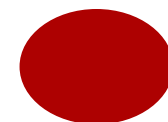
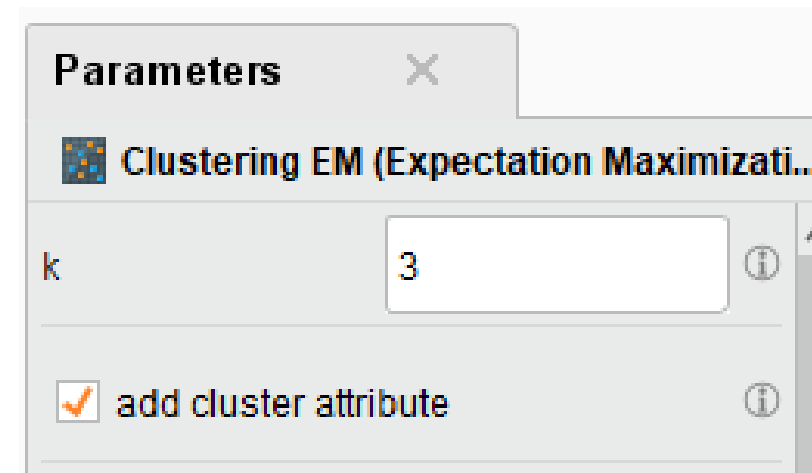
- Se puede usar siempre que se conozca la forma de la distribución (ej : gaussiana)
- Ventajas
 - Suele dar buenos resultados por ser más general que otros métodos de agrupamiento.
 - Es simple de implementar.
- Desventajas
 - Puede converger a un óptimo local.
 - Tiene un costo computacional alto en especial si se utilizan muchas distribuciones de probabilidad.



EJEMPLO – PUNTOSCLUSTERS.CSV



↑
Expectation Maximization
Clustering



MODELO OBTENIDO

- Para cada cluster se indica
 - La cantidad de ejemplos
 - La probabilidad de que un ejemplo pertenezca a cada cluster
 - La media (centroide)
 - La matriz de covarianza que define la forma de la gaussiana

Cluster Model

```
Cluster 0: 100 items
Cluster 1: 97 items
Cluster 2: 103 items
Total number of items: 300
```

```
-----

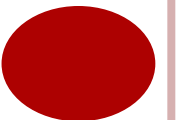
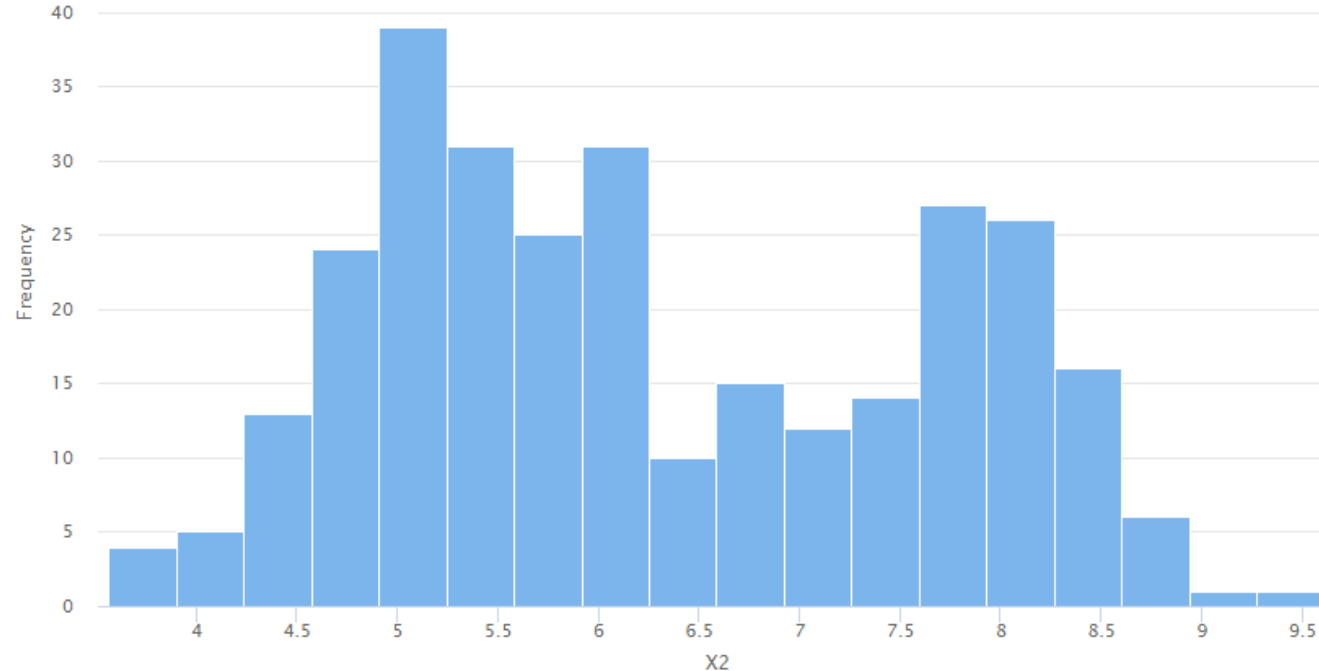
cluster probabilities:
Cluster 0: 0.3335837422817361
Cluster 1: 0.319724737525268
Cluster 2: 0.3466915201929961

cluster means:
Cluster 0: 4.997910966696282; 4.939540833319785
Cluster 1: 6.437133148970626; 7.973591615428699
Cluster 2: 7.5425466310502; 6.027543336040732

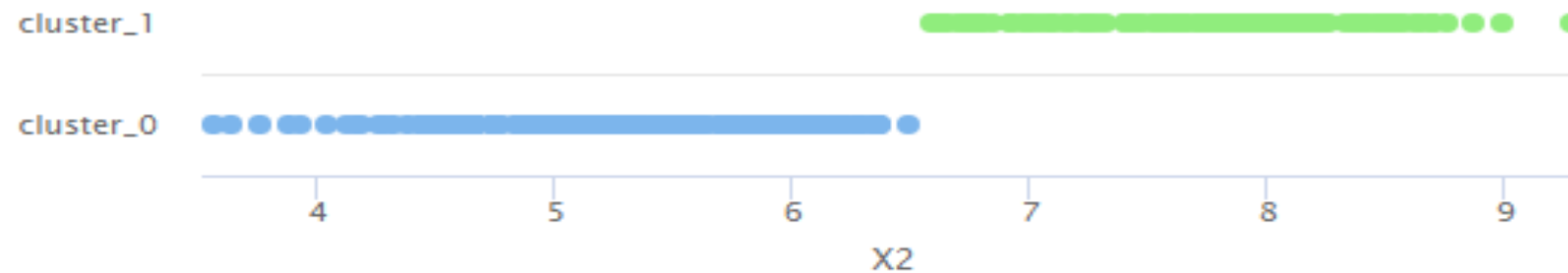
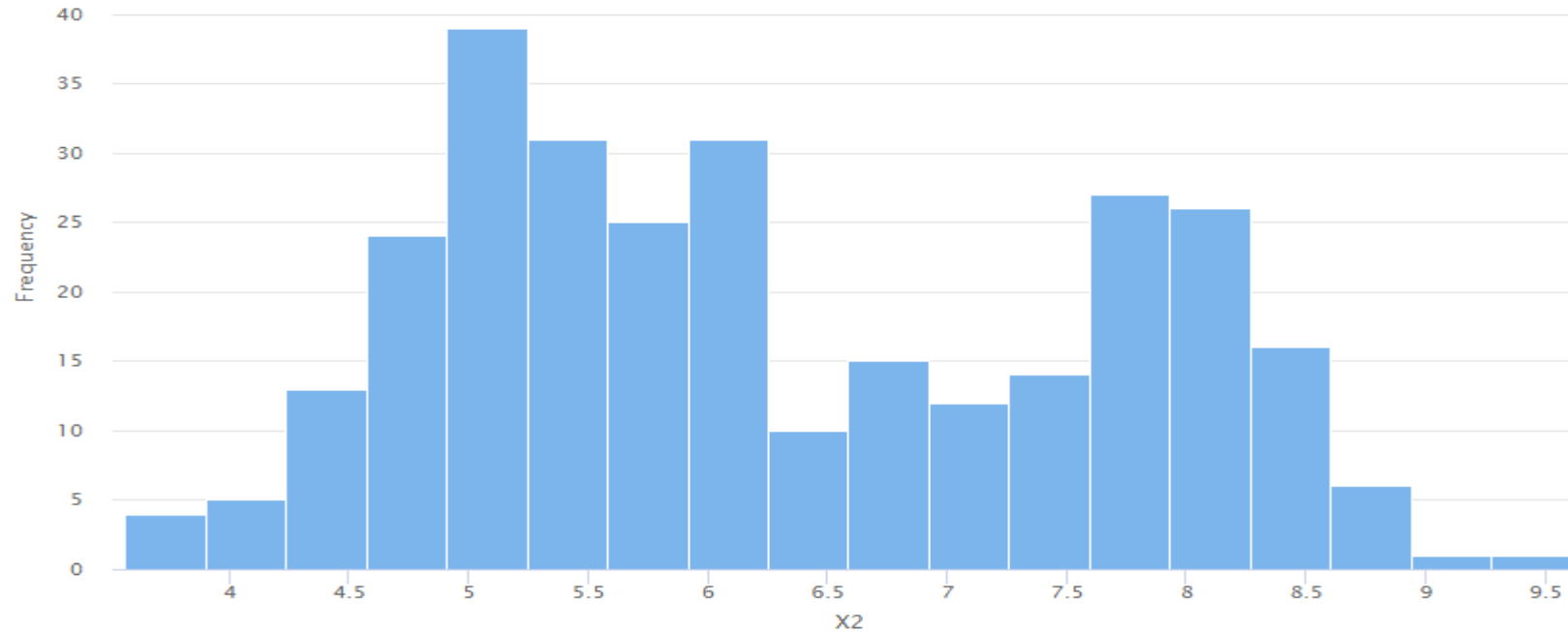
cluster covariance matrices:
Cluster 0:
0.21991011183395798      -0.009384755850033549
-0.009384755850033549    0.27550579702420064
Cluster 1:
0.22684606151038703      0.026317339496083773
0.026317339496083773    0.22119055361492618
Cluster 2:
0.277464592267925        -0.06738634016284242
-0.06738634016284242    0.35526376793989023
```

EJERCICIO

- Analice los valores del atributo X2 del archivo PuntosClusters.csv utilizando:
 - K-medias con $K=2$
 - Agrupamiento jerárquico visualizando sólo 2 grupos
 - EM con $K=2$



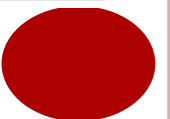
ALGORITMO K-MEDIAS CON K=2



Cluster Model

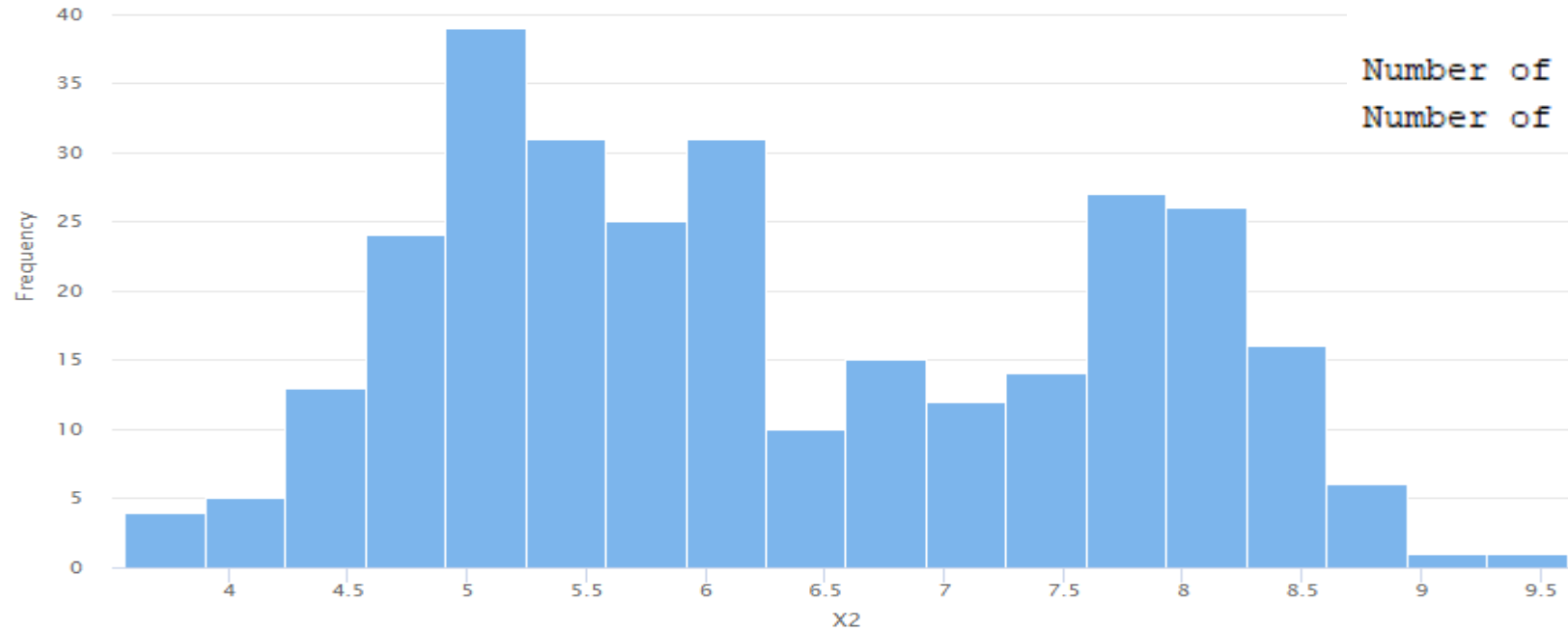
Cluster 0: 181 items
Cluster 1: 119 items
Total number of items: 300

cluster_0	cluster_1
5.316	7.764



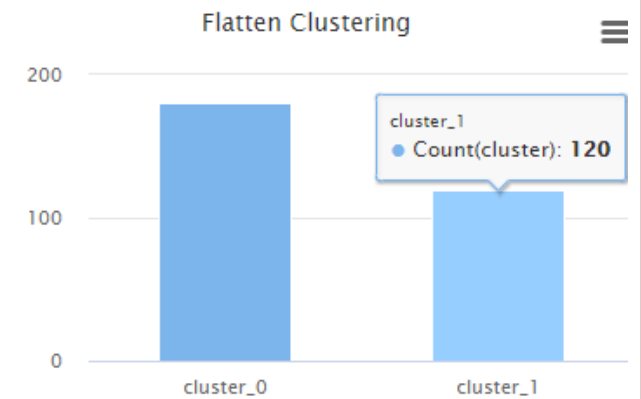
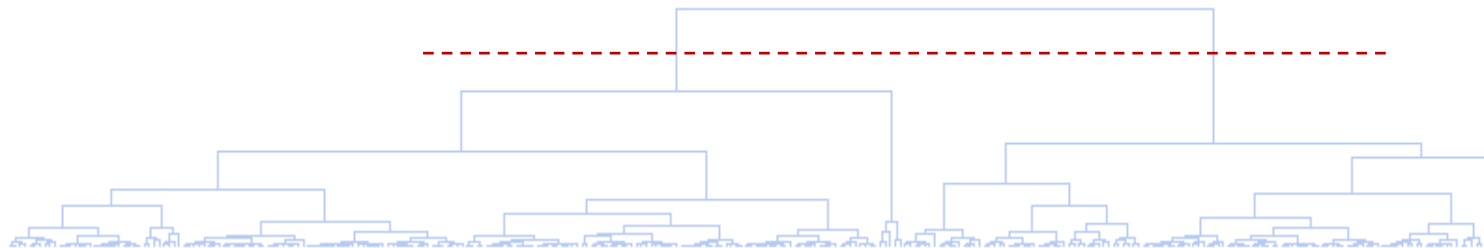
AGRUPAMIENTO JERÁRQUICO

Hierarchical Cluster Model

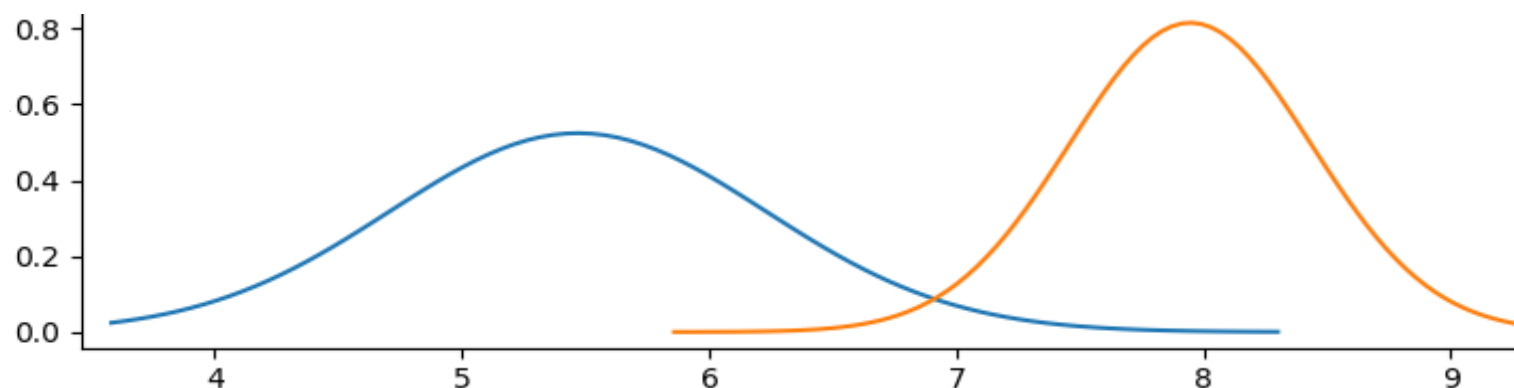
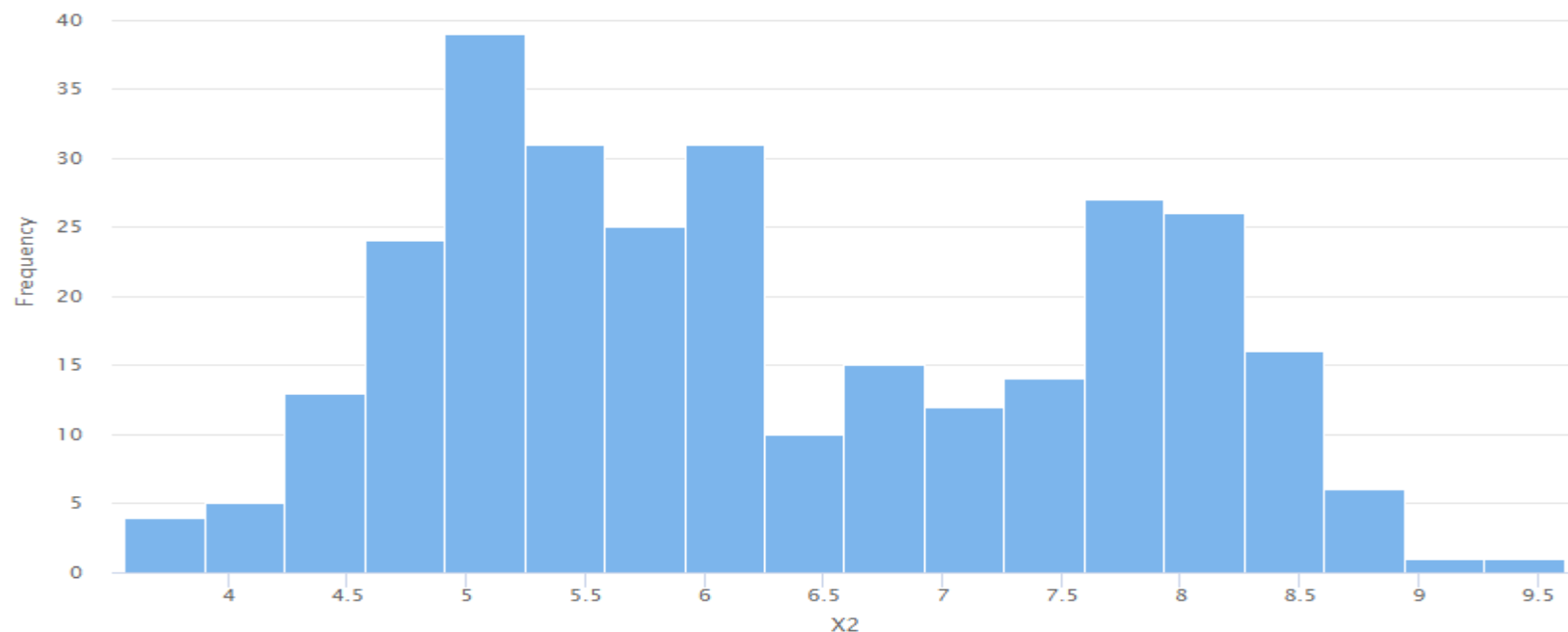


Number of clusters :599

Number of items :300



ALGORITMO EM



Cluster Model

Cluster 0: 200 items

Cluster 1: 100 items

Total number of items: 300

cluster probabilities:

Cluster 0: 0.6697867226226227

Cluster 1: 0.33021327737737705

cluster means:

Cluster 0: 5.468834450684805

Cluster 1: 7.945926591623768

cluster covariance matrices:

Cluster 0:

0.579384540264987

Cluster 1:

0.23957506465667355