

# Aprendizaje Automático

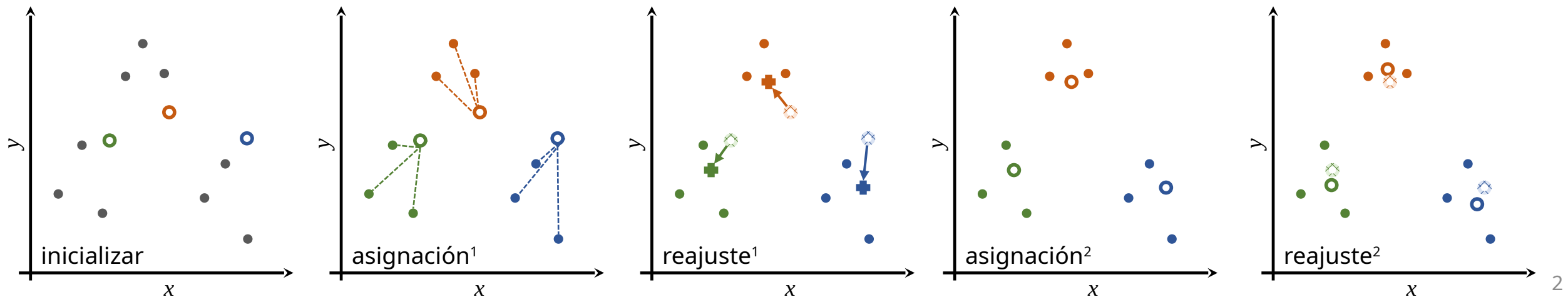
Departamento de Informática – UC3M

TUTORIAL 8 – Clustering

# Tutorial 8

## Recordando teoría. KMeans

- Inicializar
  - Situar aleatoriamente los centros de las agrupaciones
- Repetir
  - Asignación  
Cada instancia se asocia a la agrupación del centro más cercano
  - Reajuste  
Desplazar los centros de las agrupaciones al centro de gravedad de las instancias asignadas



# Tutorial 8

## Kmeans. Sklearn

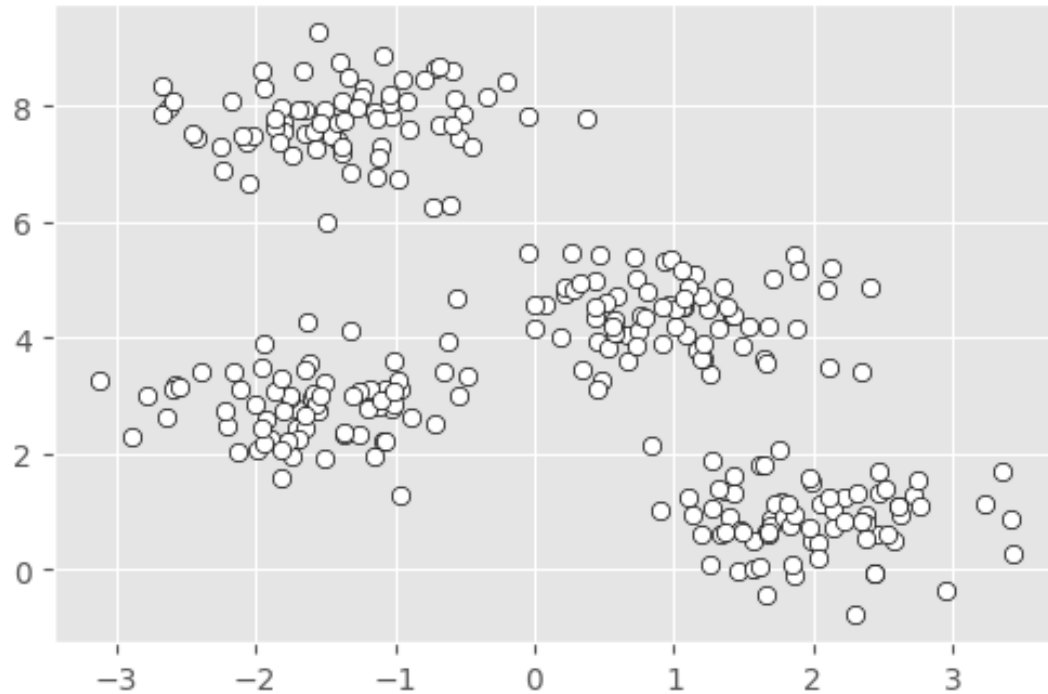
- **n\_clusters**: determina el número  $K$  de clusters que se van a generar.
- **init**: estrategia para asignar los centroides iniciales. Por defecto se emplea 'k-means++', una estrategia que trata de alejar los centroides lo máximo posible facilitando la convergencia. Sin embargo, esta estrategia puede ralentizar el proceso cuando hay muchos datos, si esto ocurre, es mejor utilizar 'random'.
- **n\_init**: determina el número de veces que se va a repetir el proceso, cada vez con una asignación aleatoria inicial distinta. Es recomendable que este último valor sea alto, entre 10-25, para no obtener resultados subóptimos debido a una iniciación poco afortunada del proceso.
- **max\_iter**: número máximo de iteraciones permitidas.
- **random\_state**: semilla para garantizar la reproducibilidad de los resultados.



# Tutorial 8

## Número de clusters. Método del codo (Elbow method)

Datos simulados



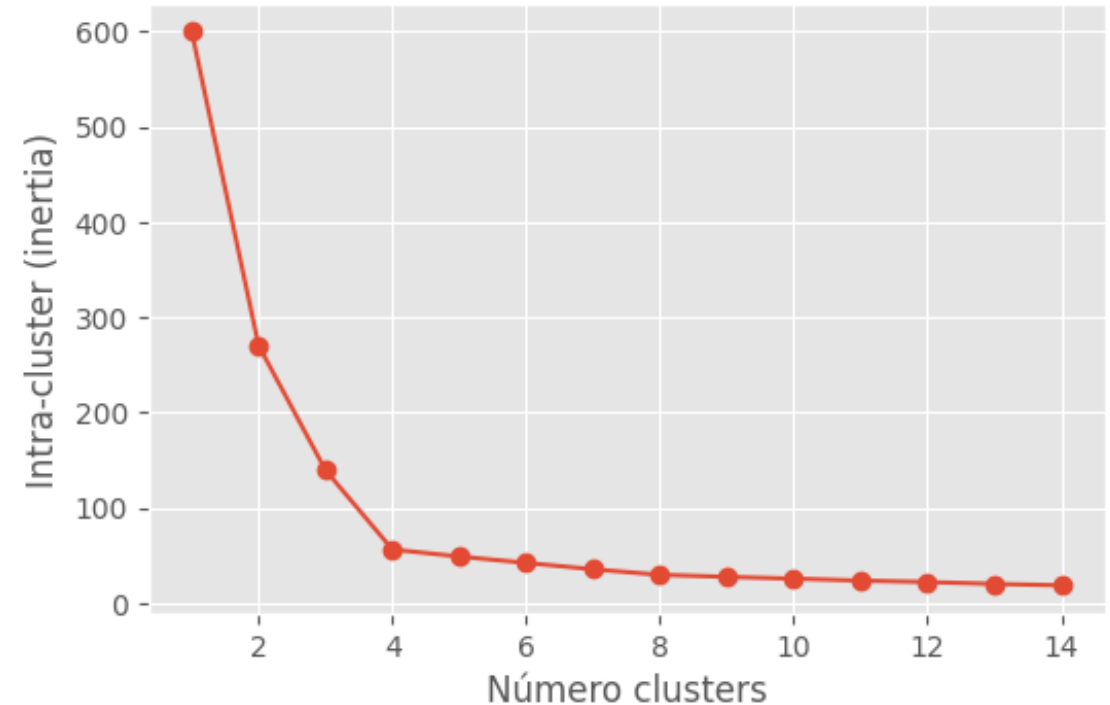
**Fórmula:**

$$\text{Inercia} = \sum_{i=1}^n \min_{\mu_j \in C} \|x_i - \mu_j\|^2$$

**Donde:**

- $x_i$  es cada punto de datos.
- $\mu_j$  es el centroide del cluster más cercano.
- $C$  es el conjunto de centroides.

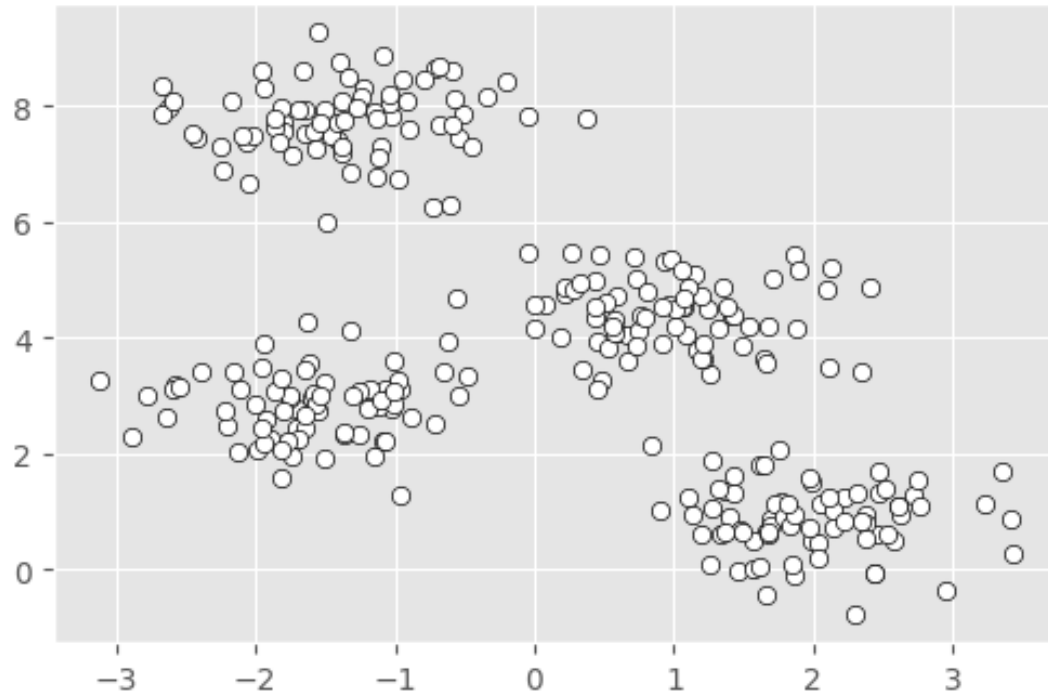
Evolución de la varianza intra-cluster total



# Tutorial 8

## Número de clusters. Método Silhouette

Datos simulados

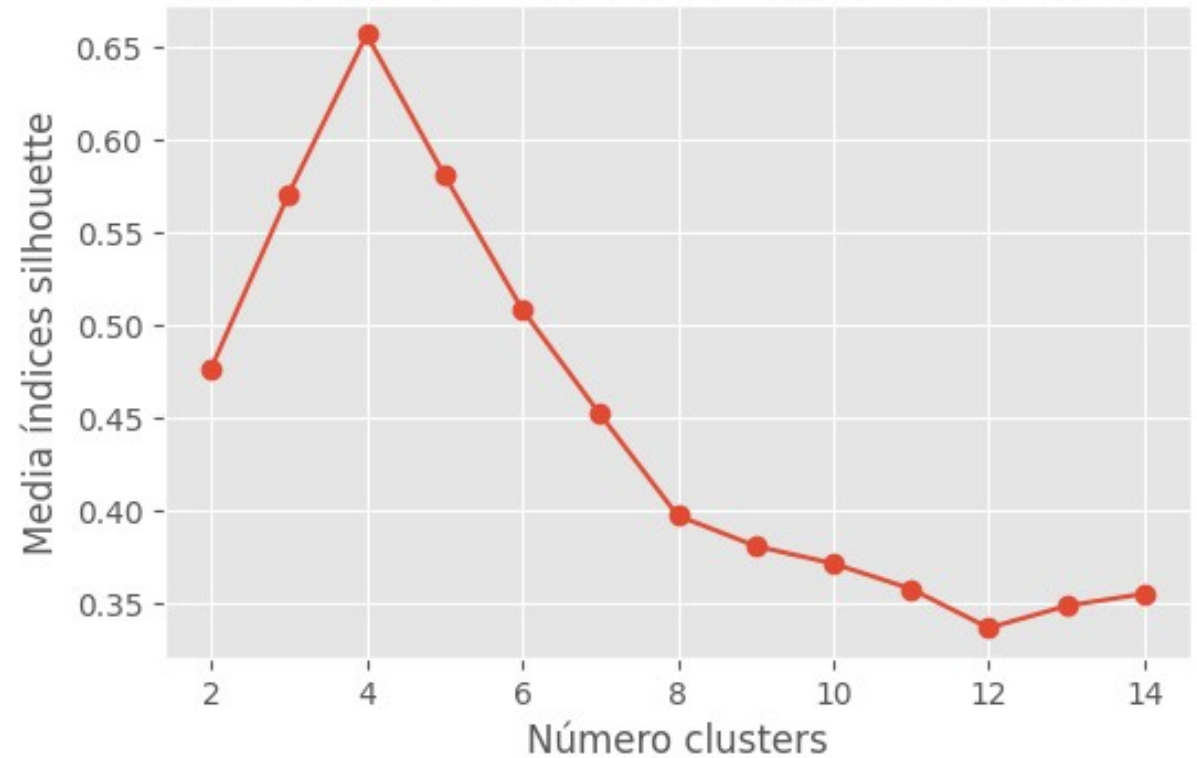


$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Donde:

- $a(i)$  es la **distancia media** entre el punto  $i$  y todos los demás puntos **de su mismo cluster**.
- $b(i)$  es la **distancia media** entre el punto  $i$  y todos los puntos del **cluster más cercano diferente**.

Evolución de media de los índices silhouette



### Interpretación:

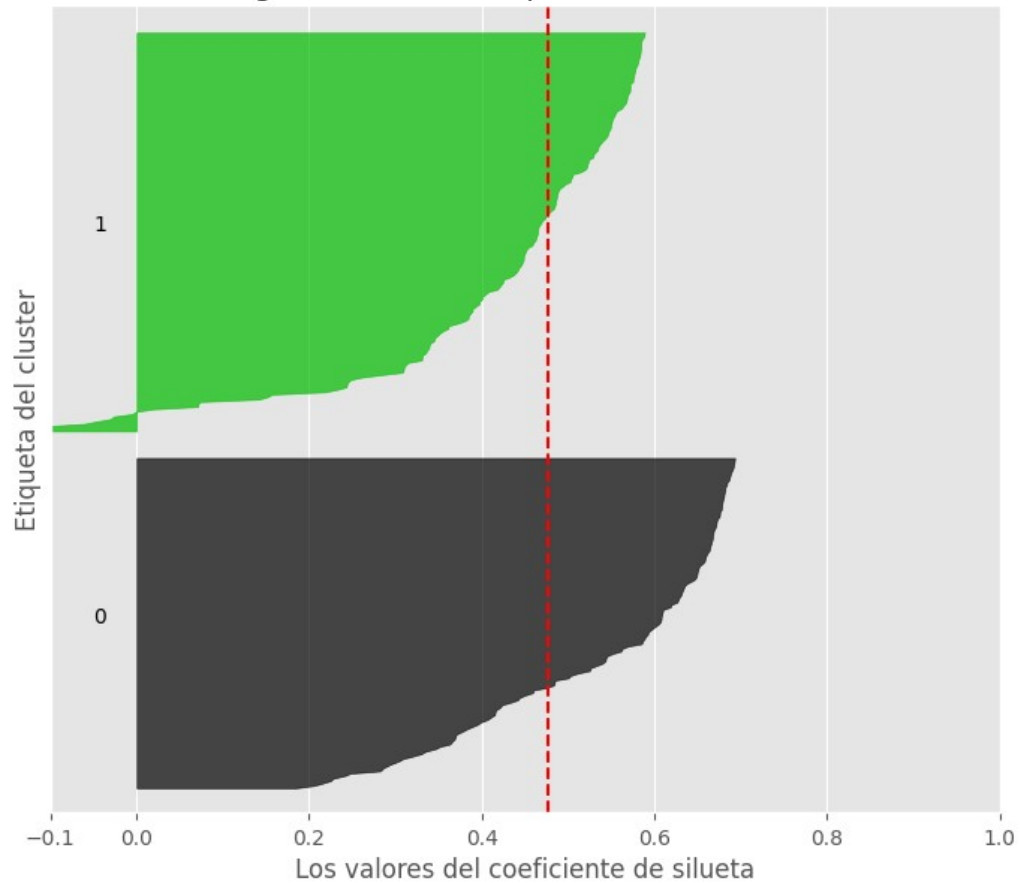
- $s(i)$  cercano a 1. Bien asignado
- $s(i)$  negativo. Mal asignado

# Tutorial 8

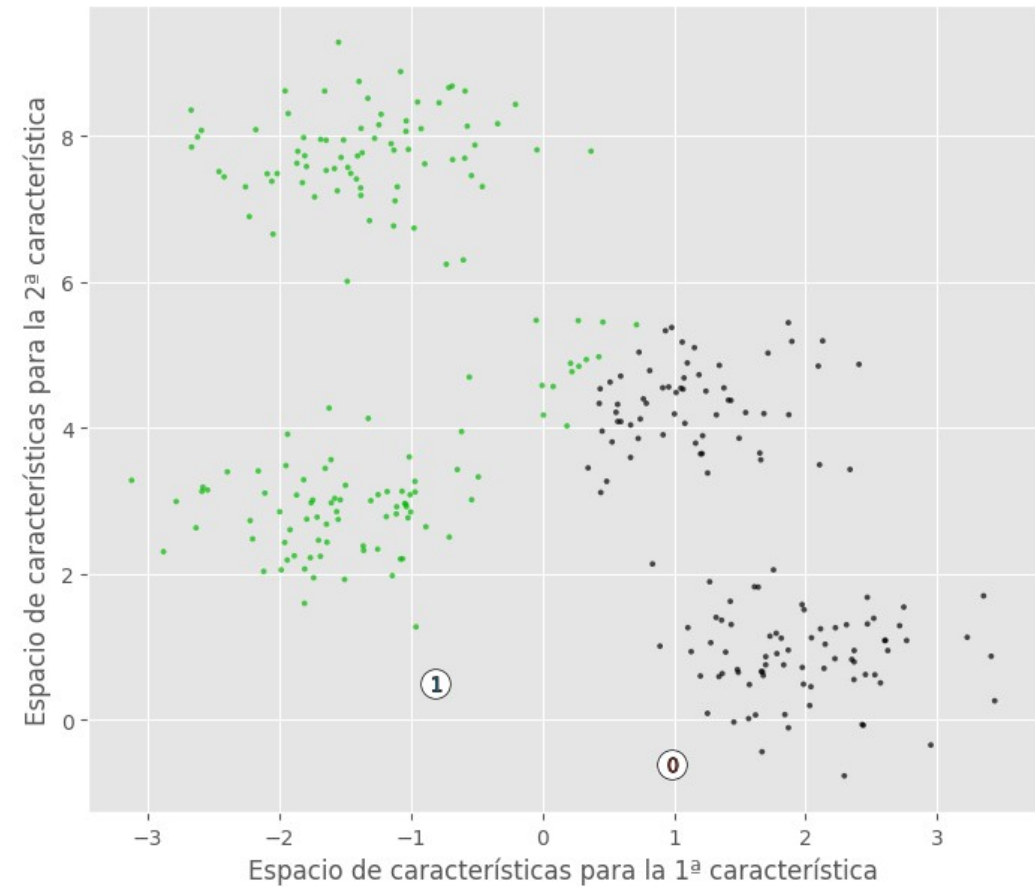
## Número de clusters. Método Silhouette

**Análisis silhouette para la agrupación KMeans en datos de muestra con  $n\_clusters = 2$**

El gráfico silhouette para varios clusters.



Visualización de los datos clusterizados.

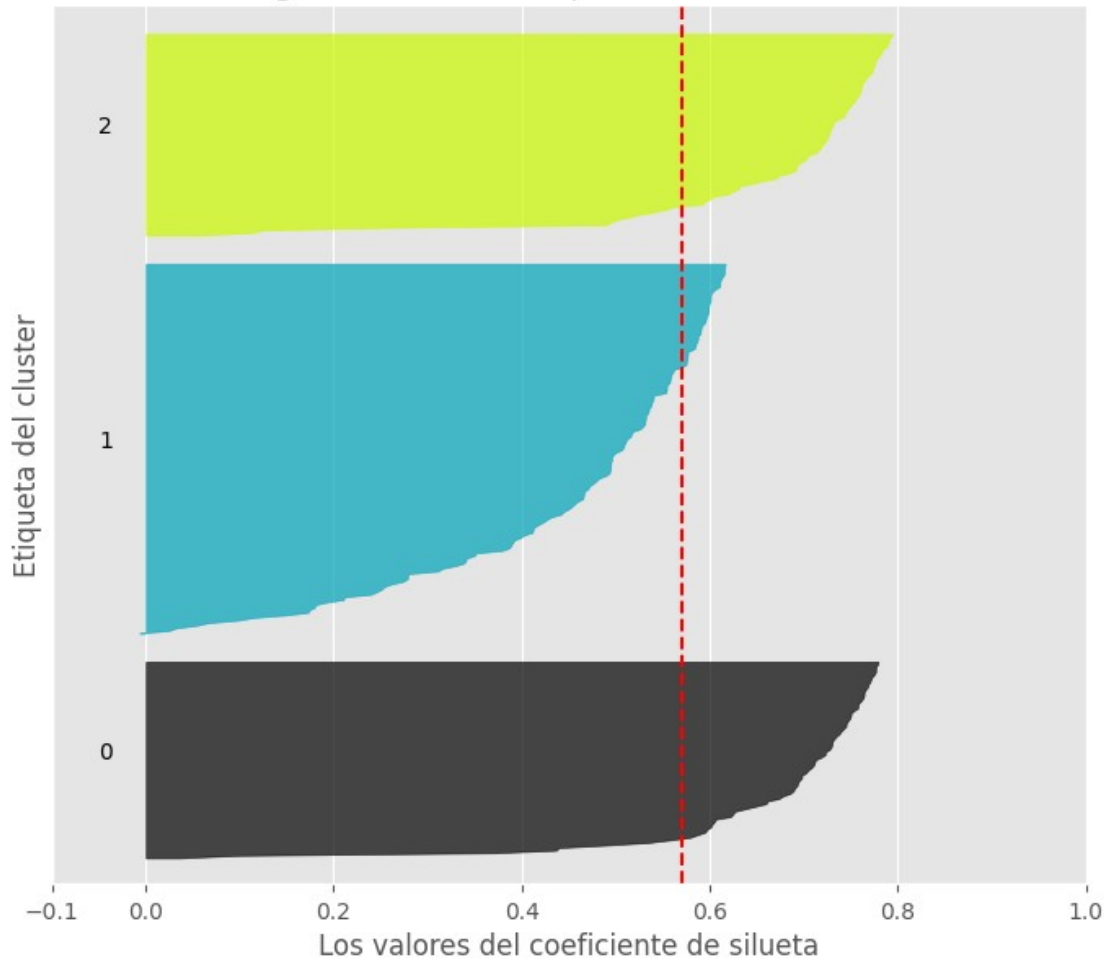


# Tutorial 8

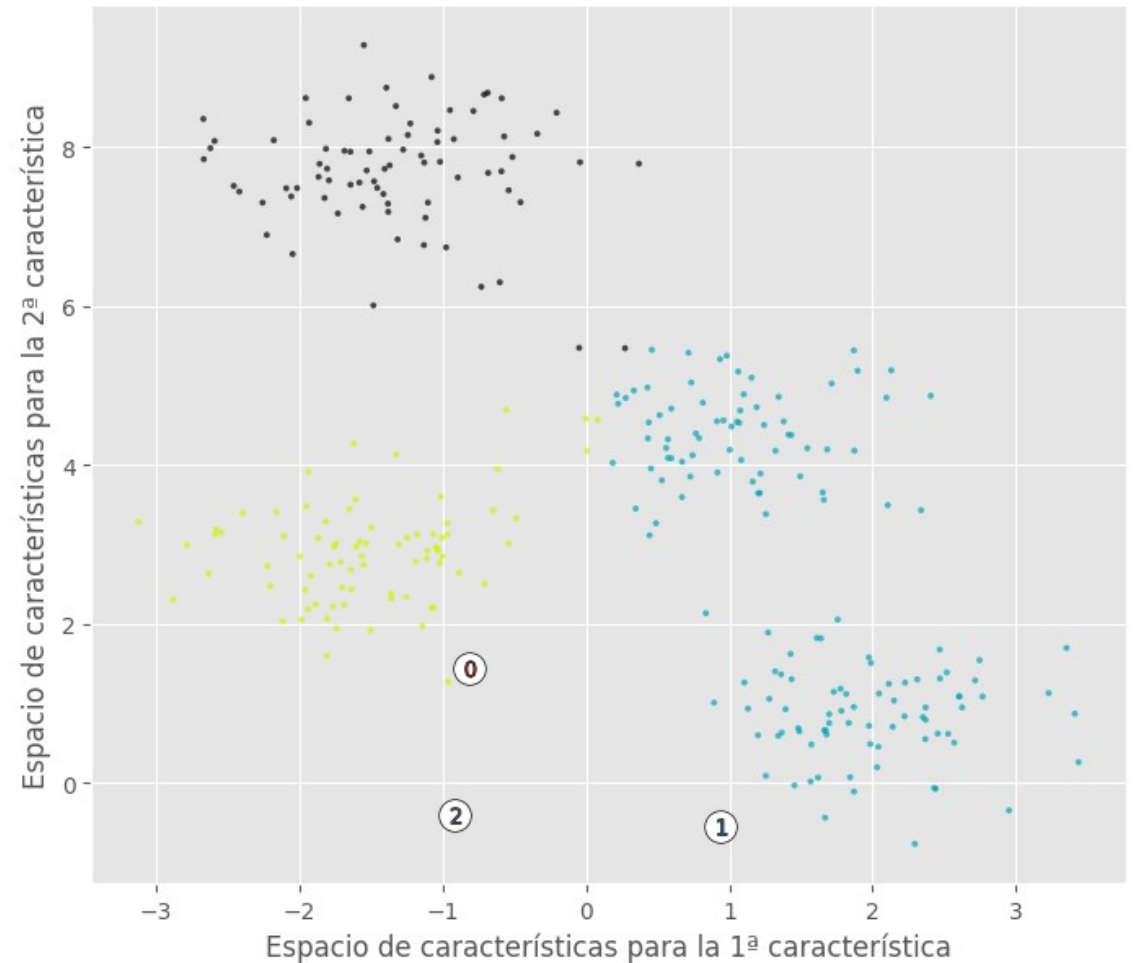
## Número de clusters. Método Silhouette

**Análisis silhouette para la agrupación KMeans en datos de muestra con  $n\_clusters = 3$**

El gráfico silhouette para varios clusters.



Visualización de los datos clusterizados.

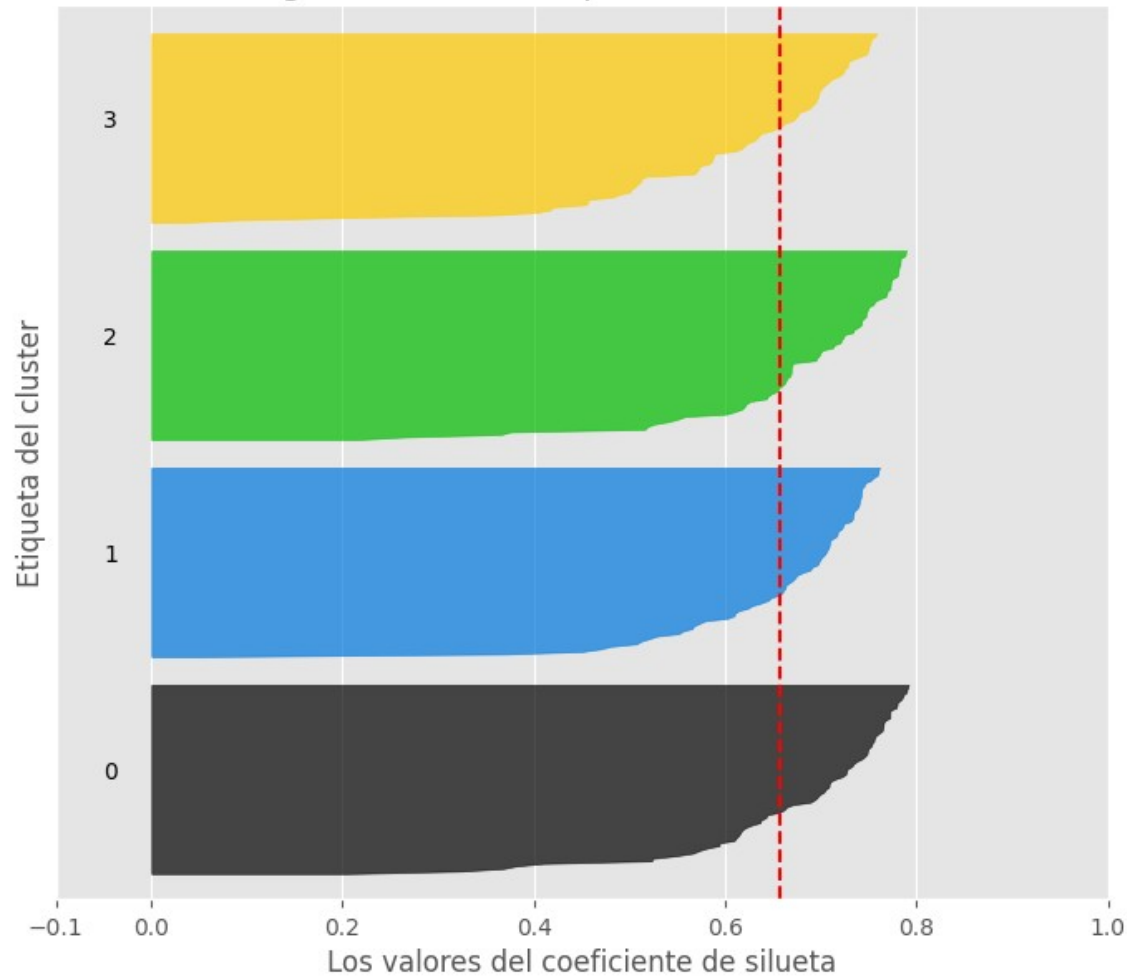


# Tutorial 8

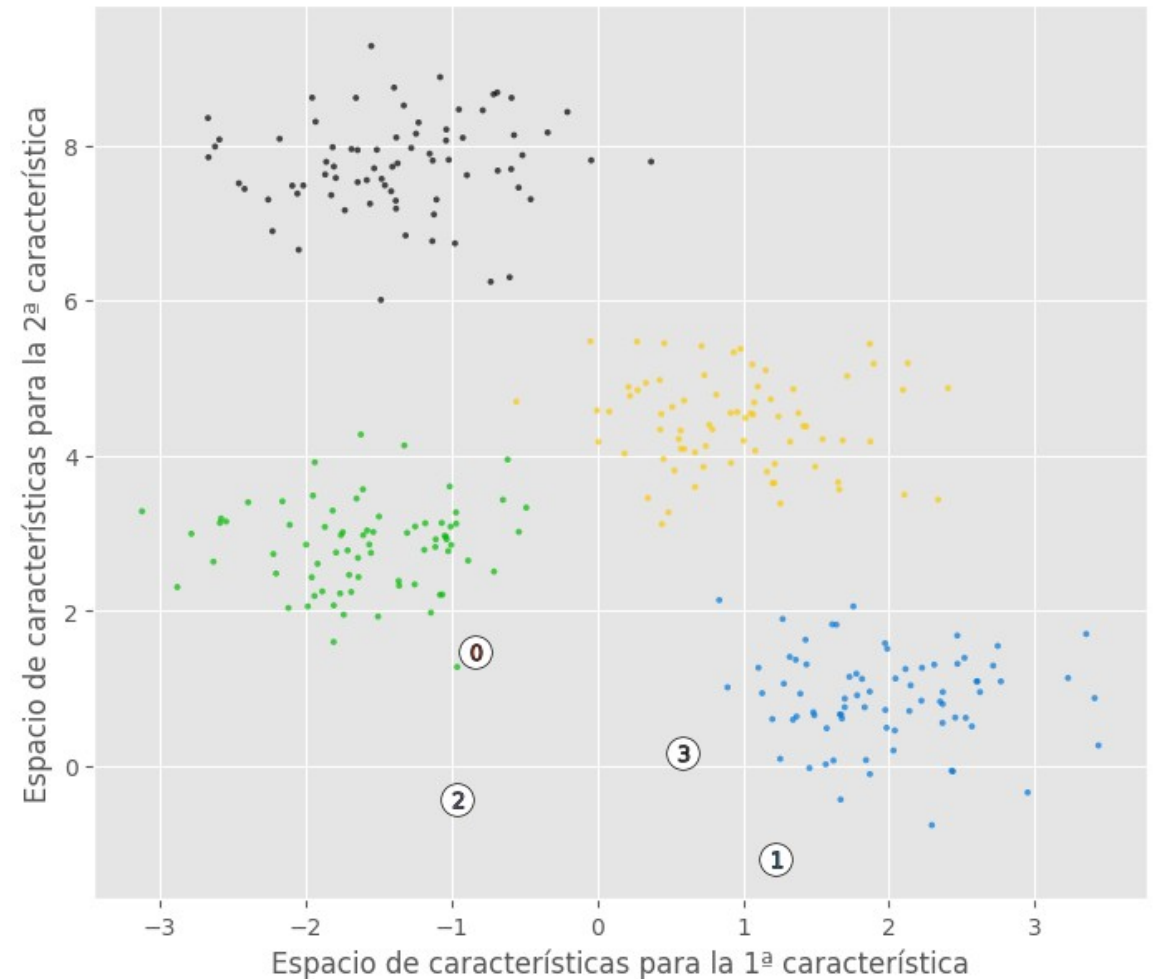
## Número de clusters. Método Silhouette

**Análisis silhouette para la agrupación KMeans en datos de muestra con  $n\_clusters = 4$**

El gráfico silhouette para varios clusters.



Visualización de los datos clusterizados.



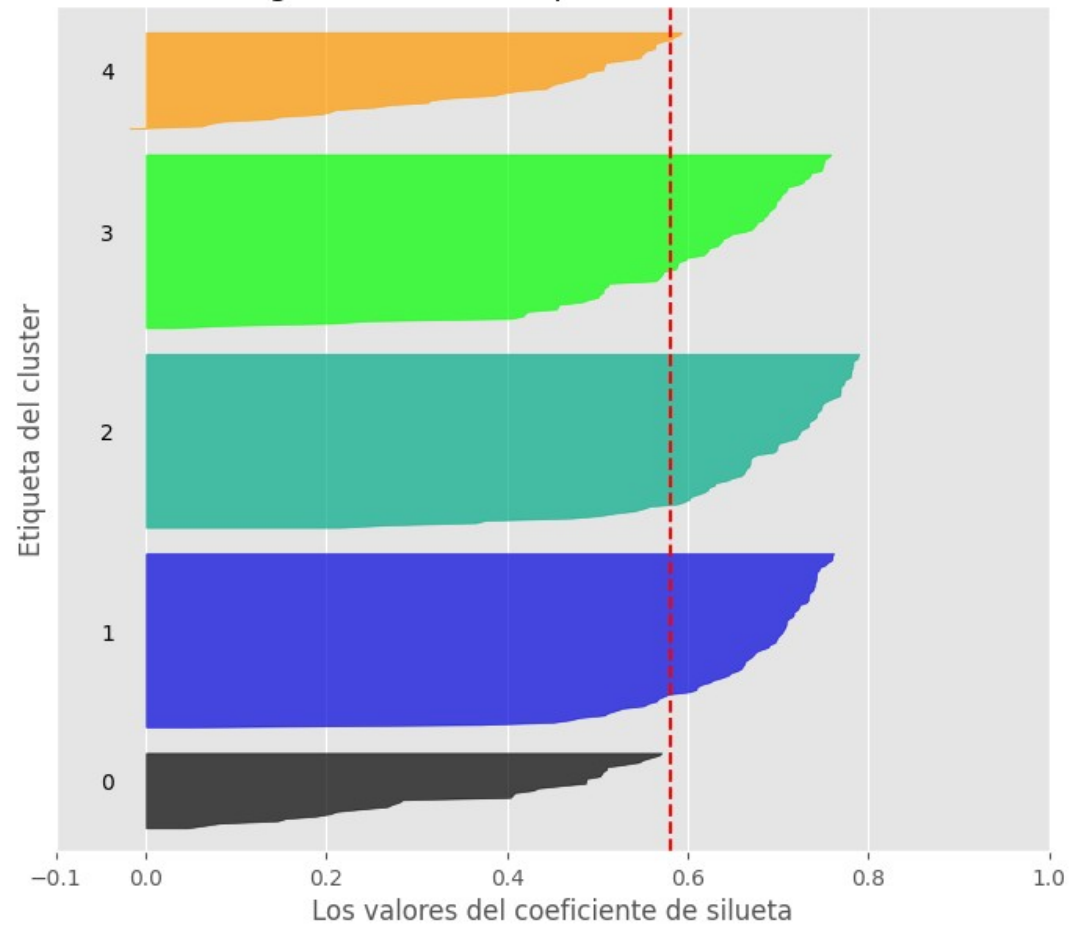


# Tutorial 8

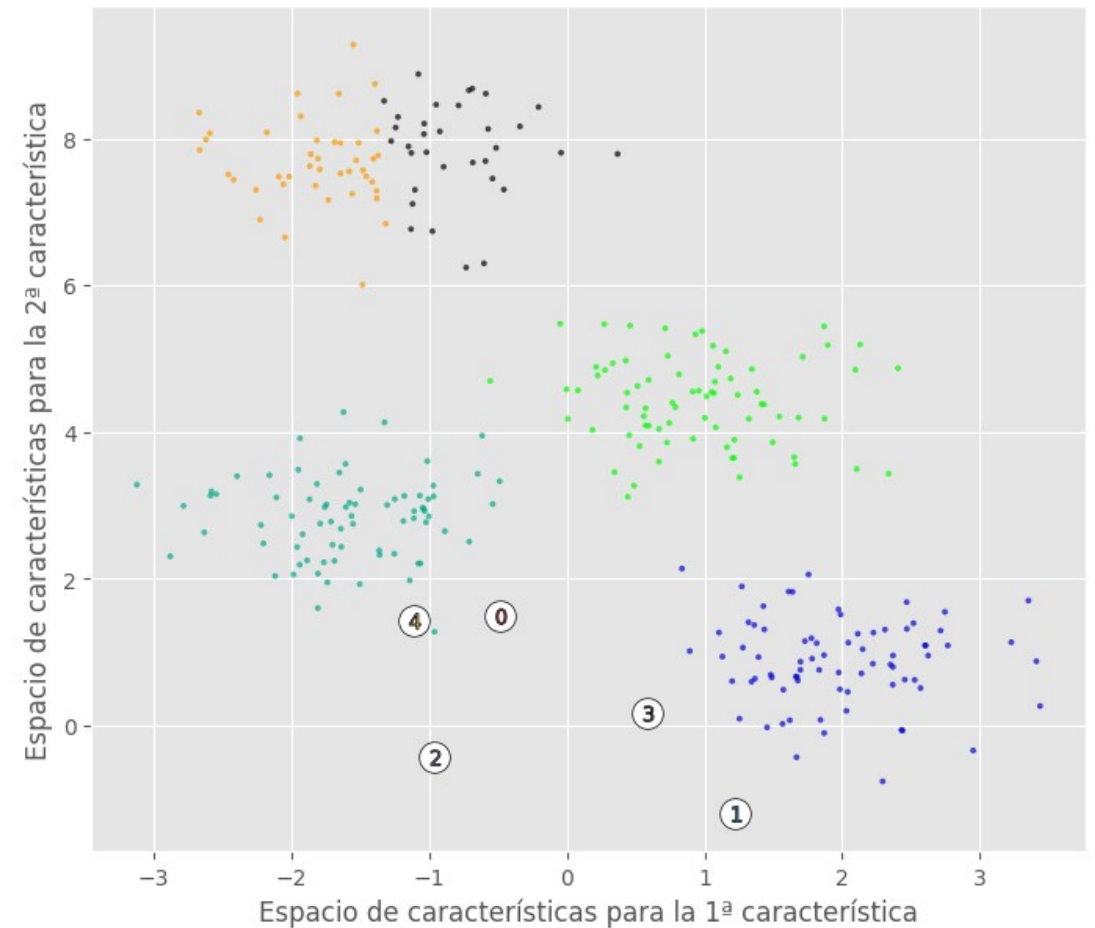
## Número de clusters. Método Silhouette

**Análisis silhouette para la agrupación KMeans en datos de muestra con  $n\_clusters = 5$**

El gráfico silhouette para varios clusters.



Visualización de los datos clusterizados.

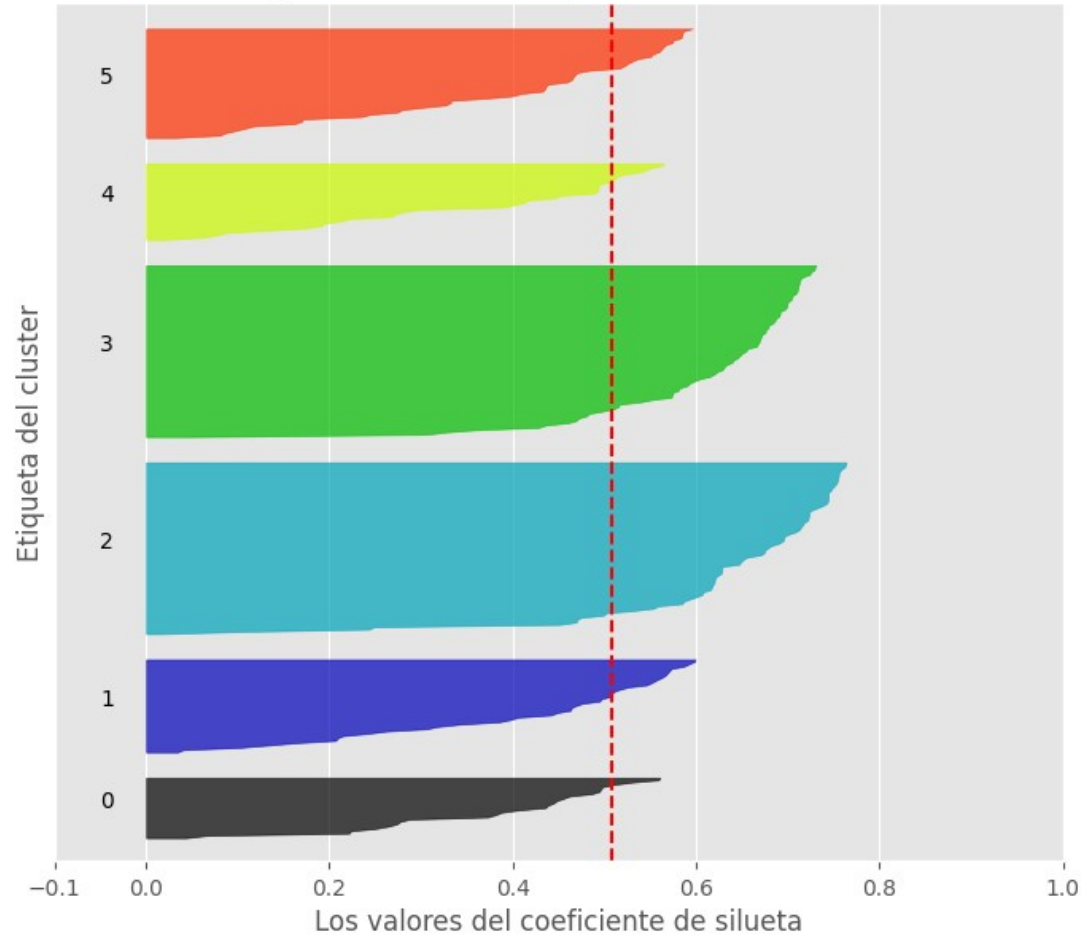


# Tutorial 8

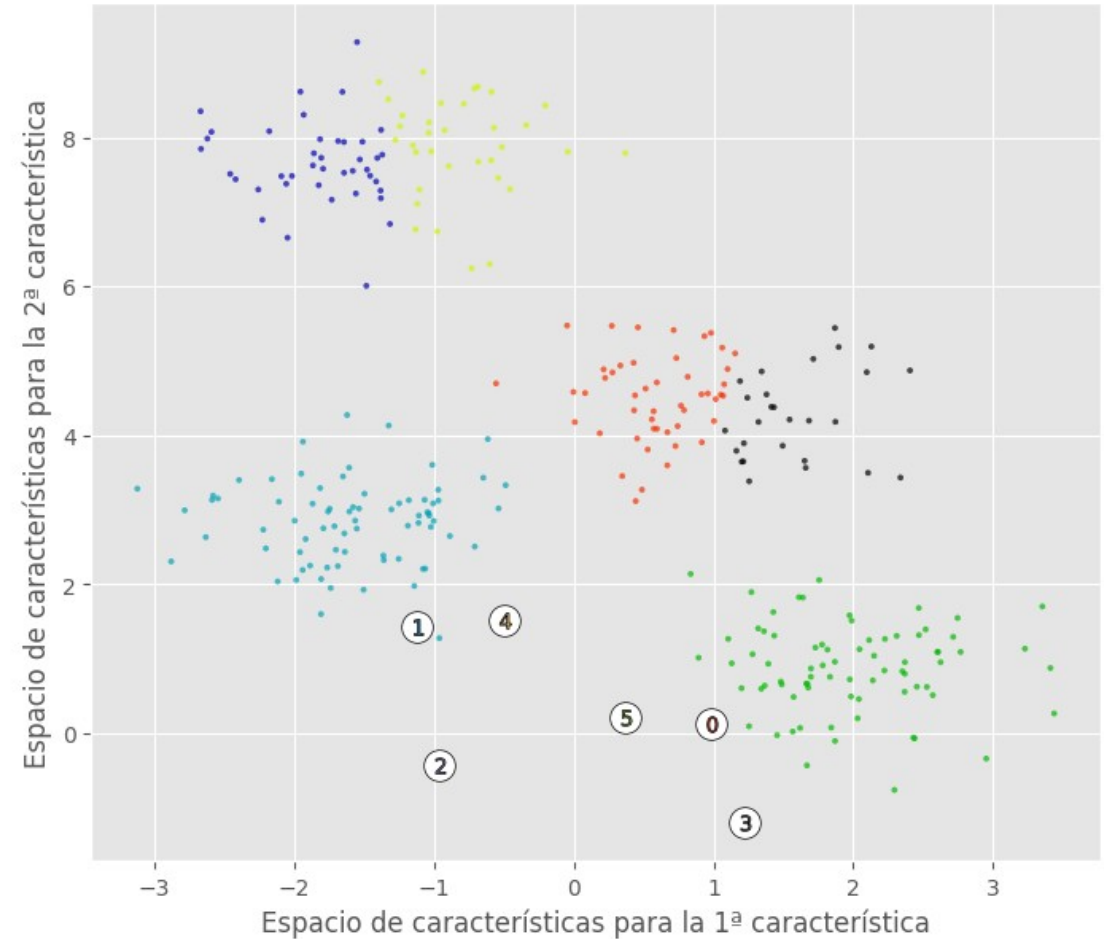
## Número de clusters. Método Silhouette

**Análisis silhouette para la agrupación KMeans en datos de muestra con  $n\_clusters = 6$**

El gráfico silhouette para varios clusters.



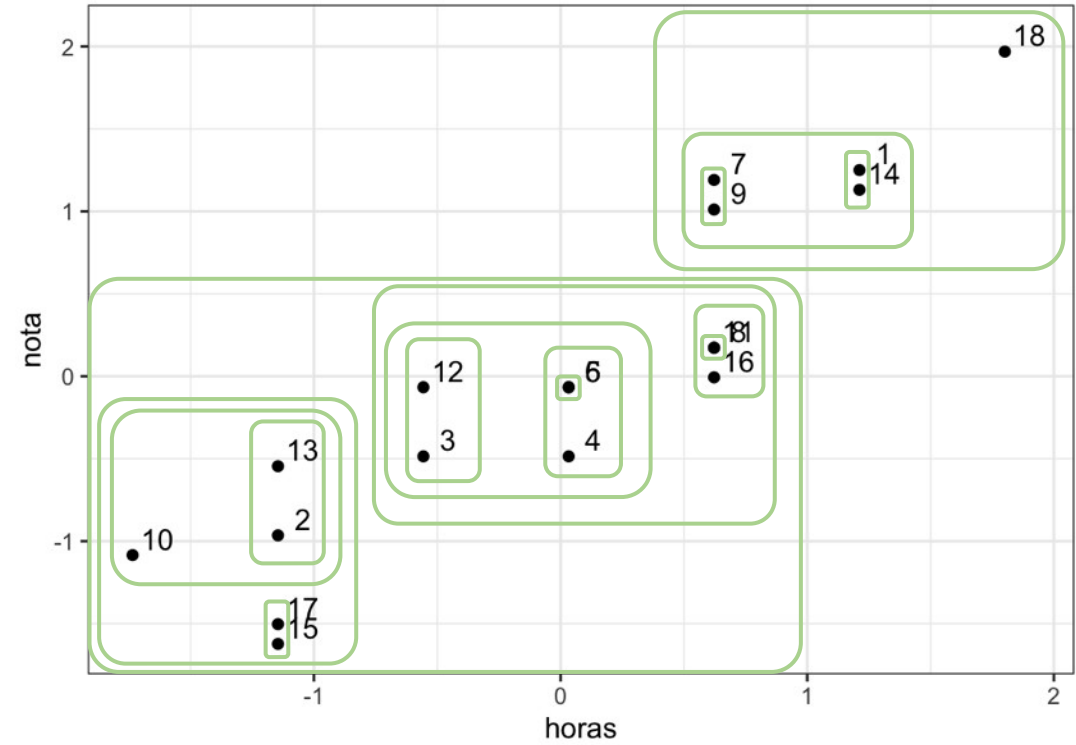
Visualización de los datos clusterizados.



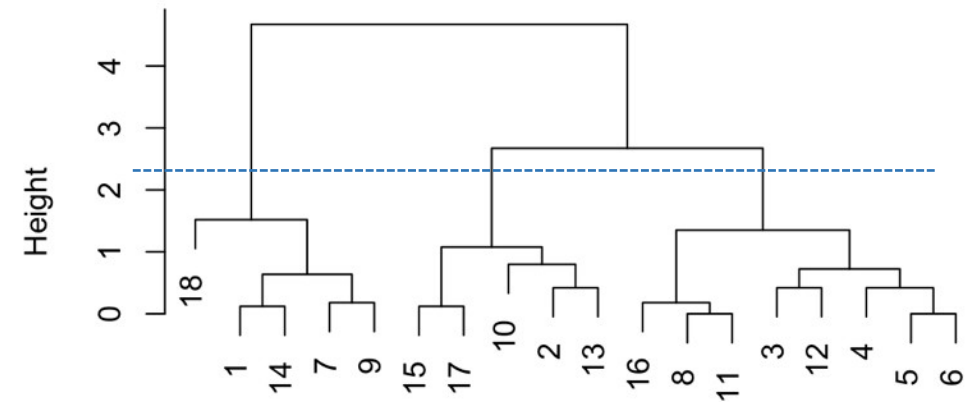
# Tutorial 8

## Cluster jerárquico

Horas estudio al día	Nota media t	Horas estudio al día	Nota media t	Horas estudio al día	Nota media t
5	8,8	5	8,7	5	8,7
1	5,1	1	5,1	1	5,1
2	5,9	2	5,9	2	5,9
3	5,9	3	5,9	3	5,9
3	6,6	3	6,6	3	6,6
4	8,7	4	8,7	4	8,7
4	7,0	4	7,0	4	7,0
4	8,4	4	8,4	4	8,4
0	4,9	0	4,9	0	4,9
2	6,6	2	6,6	2	6,6
1	5,8	1	5,8	1	5,8
5	8,6	1	4,0	1	4,1
1	4,0	4	6,7	4	6,7
4	6,7	1	4,2	6	10,0
1	4,2	6	10,0		
6	10,0				



Cluster Dendrogram

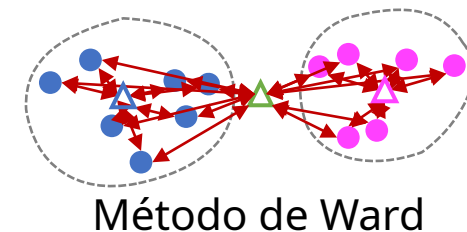
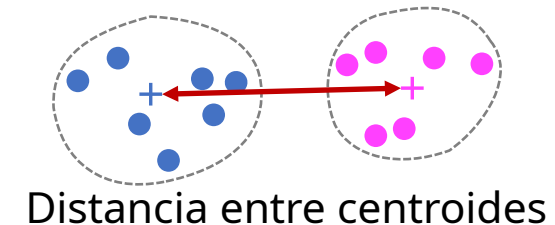
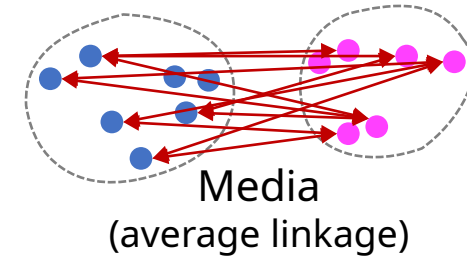
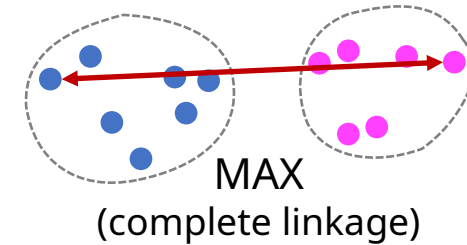
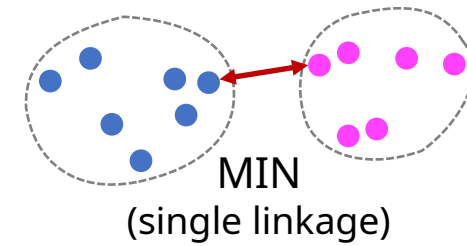


# Tutorial 8

## Cluster jerárquico. "linkage"

### Cálculo de la distancia

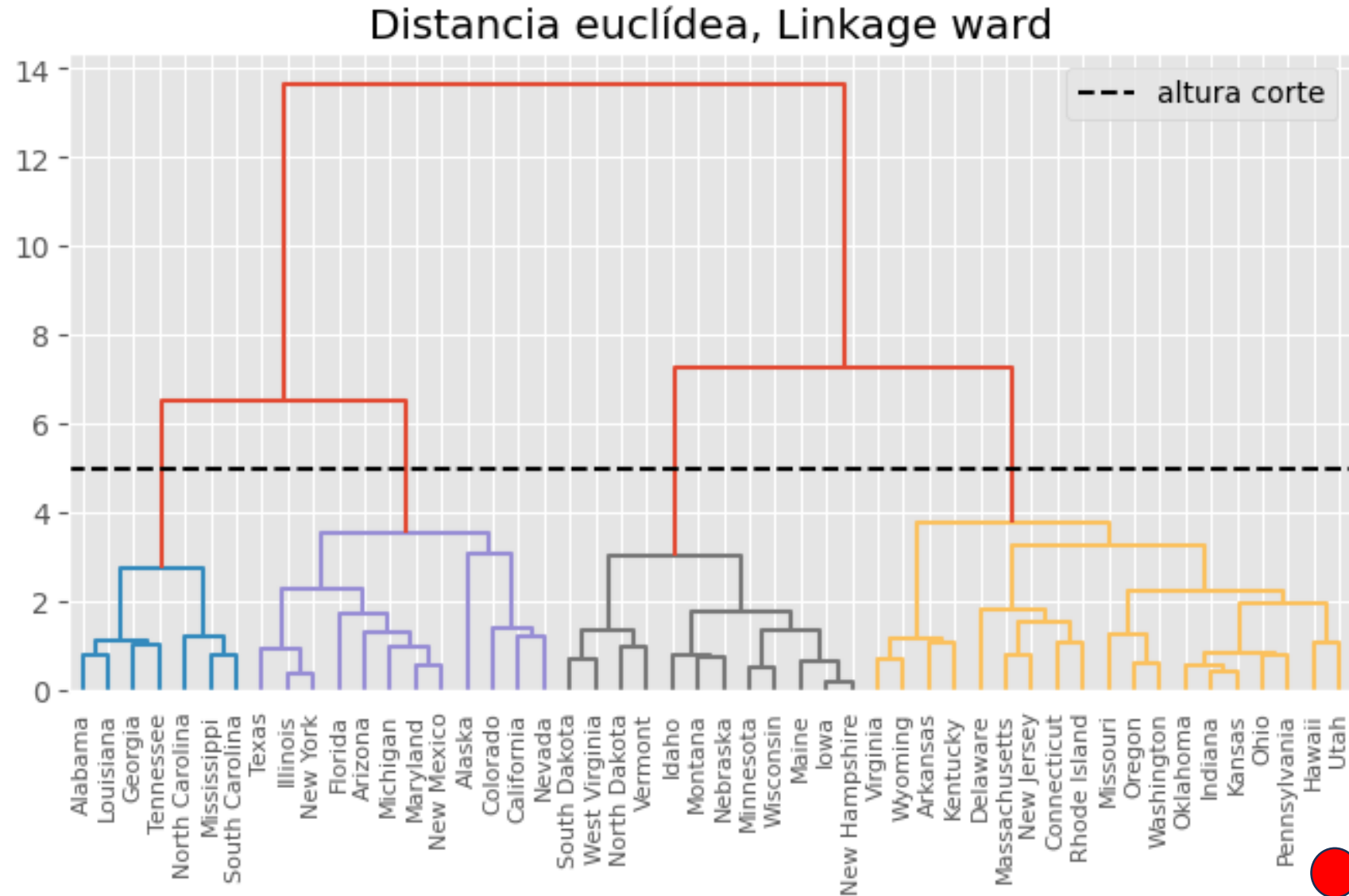
- MIN ("Single")
  - Distancia menor entre instancias de ambas agrupaciones
- MAX ("Complete")
  - Mayor distancia entre instancias de ambas agrupaciones
- Media ("Average")
  - Distancia media entre las instancias de una y otra agrupación
- Distancia entre centroides
  - Distancia entre los centroides de las agrupaciones
- Ward (función objetivo)
  - **Minimiza la varianza total en las agrupaciones**
  - Más robusto al ruido



# Tutorial 8

## Cluster jerárquico. USArrests

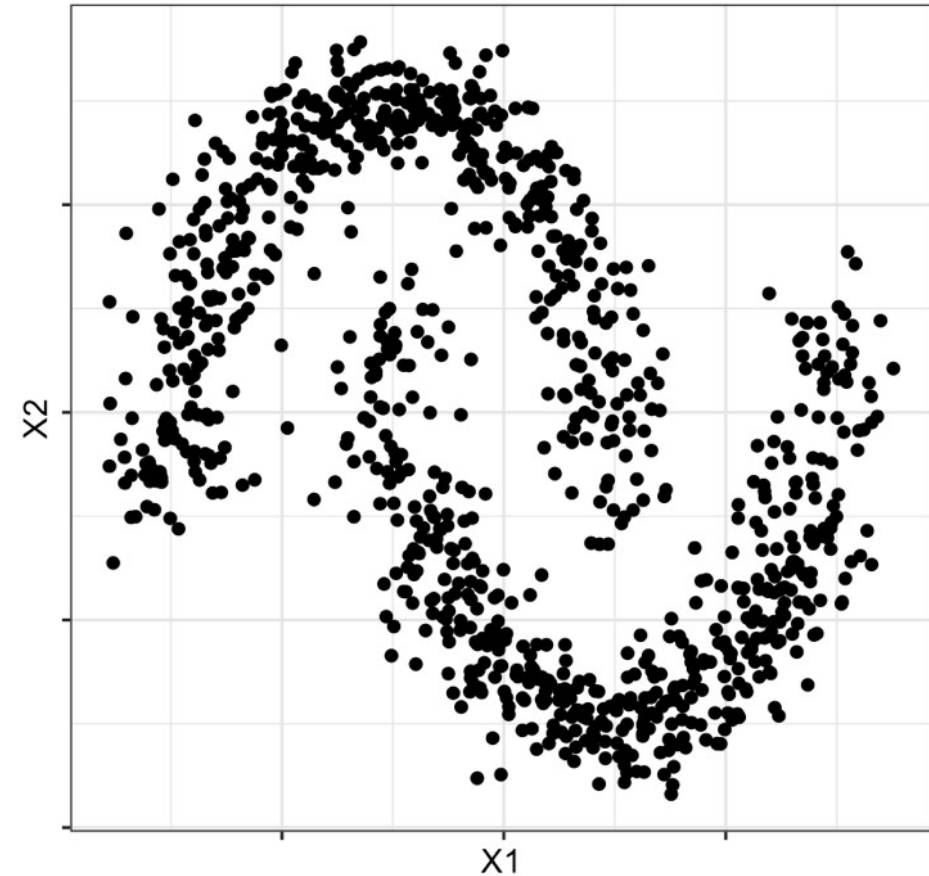
	Murder	Assault	UrbanPop	Rape
rownames				
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5



# Tutorial 8

## Recordando teoría. Cluster por densidad.

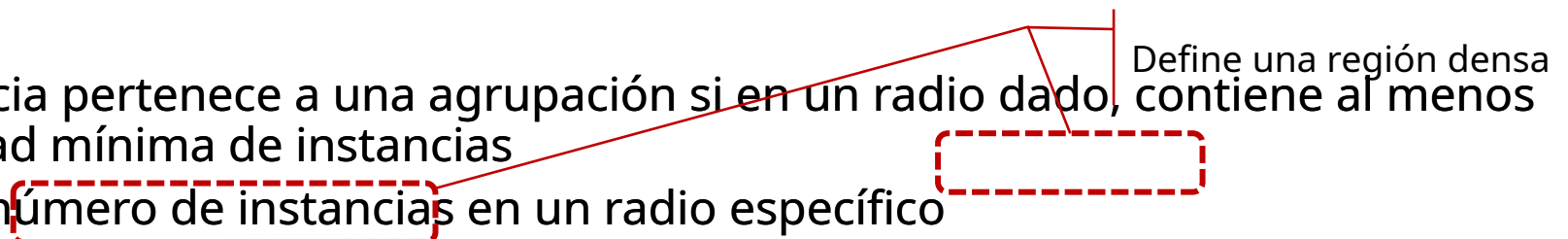
- Las agrupaciones son regiones de **alta densidad** de instancias separadas por regiones de baja densidad o definidas mediante una función explícita
  - Utiliza un **criterio local** (como el agrupamiento jerárquico) a diferencia de k-means que usa un criterio global
- Características principales
  - Permite obtener agrupaciones con cualquier forma
  - Pueden manejar instancias con ruido
  - Requieren de definir/ajustar parámetros de densidad



# Tutorial 8

## Recordando teoría. DBSCAN.

### DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

- Conceptos:
    - Una instancia pertenece a una agrupación si en un radio dado, contiene al menos una cantidad mínima de instancias
    - Densidad: número de instancias en un radio específico
    - Radio (épsilon, *eps*): dos instancias a igual o menor distancia que *eps* se consideran que son vecinas
    - Cantidad mínima (Puntos mínimos, *minPts*): número mínimo de instancias vecinas que formarían una región densa
- 

# Tutorial 8

## DBSCAN - sklearn.

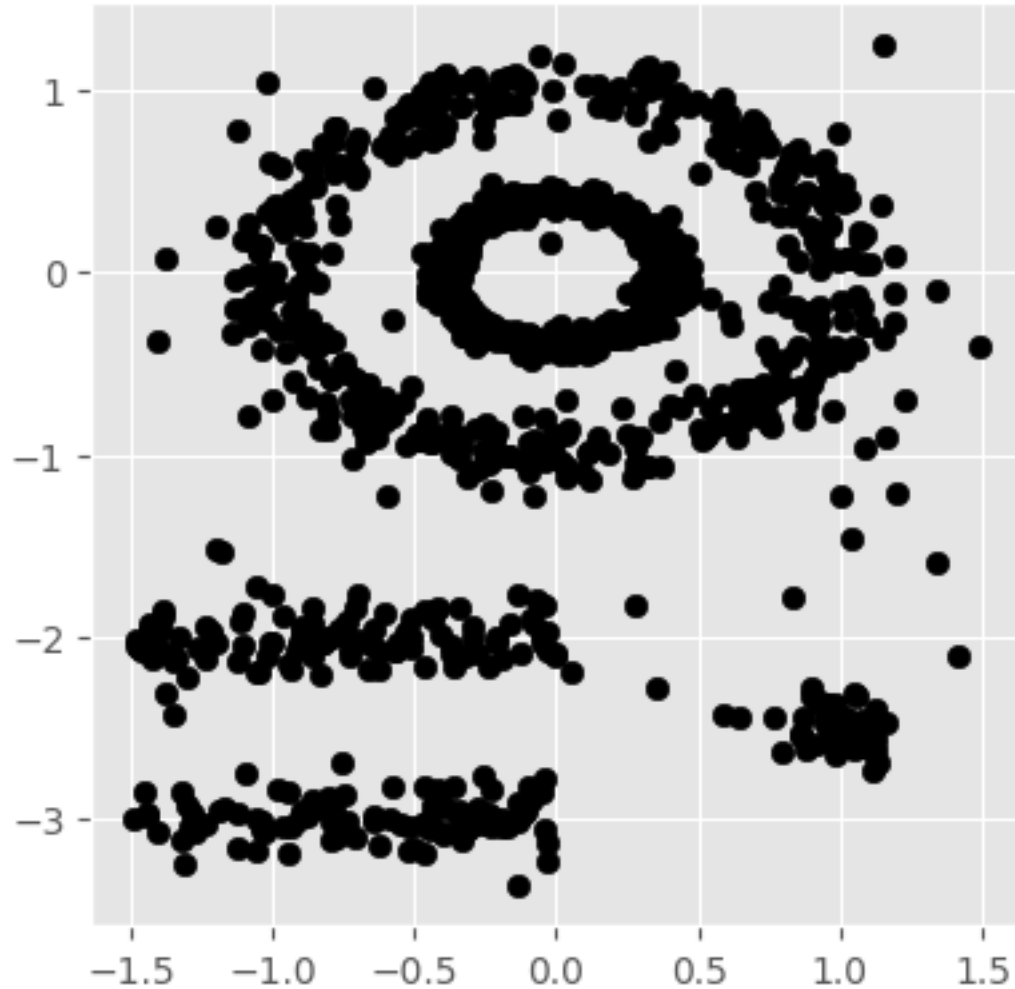
- **eps**: Distancia máxima entre dos muestras para que una se considere vecina de la otra. Define el  $\epsilon$ -neighborhood
- **min\_samples**: El número de muestras (o peso total) en un vecindario para que un punto se considere un core point. Esto incluye el propio punto.
- **metric**: métrica utilizada como distancia. Puede ser: "euclidean", "l1", "l2", "manhattan", "cosine", or "precomputed". Por defecto es "euclidean".



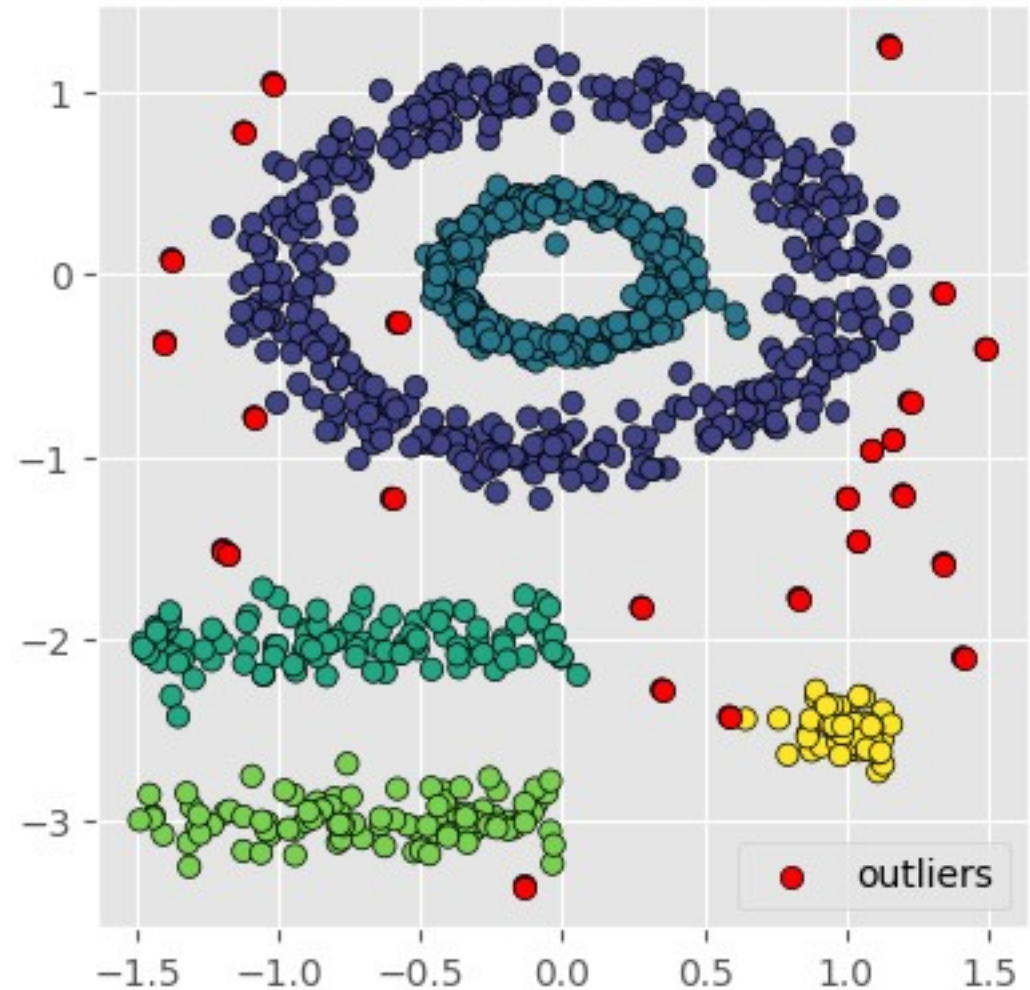
# Tutorial 8

## DBSCAN - sklearn.

Nube de puntos iniciales

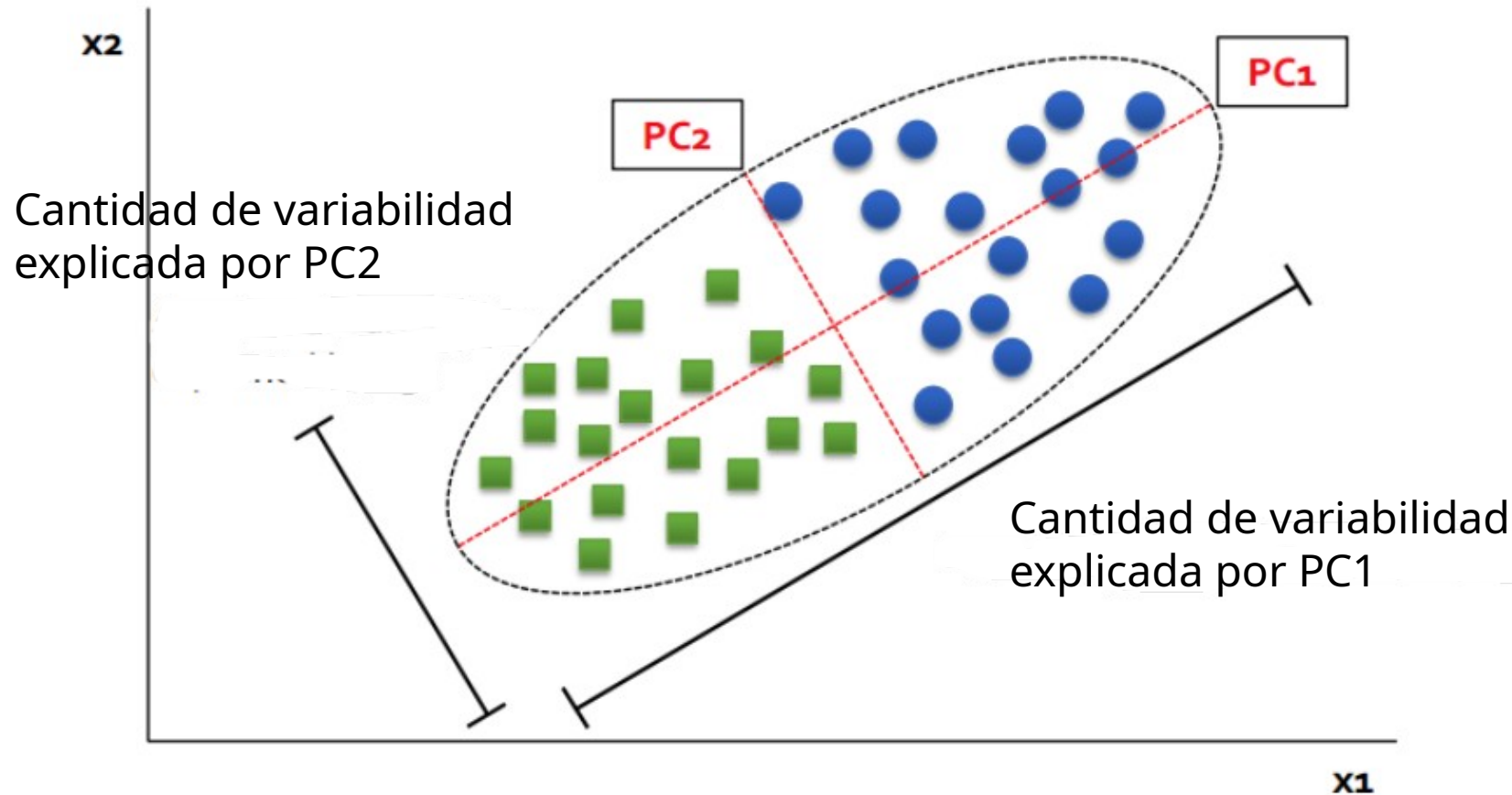


Clusterings generados por DBSCAN



# Tutorial 8

## Análisis de Componentes Principales y Clustering



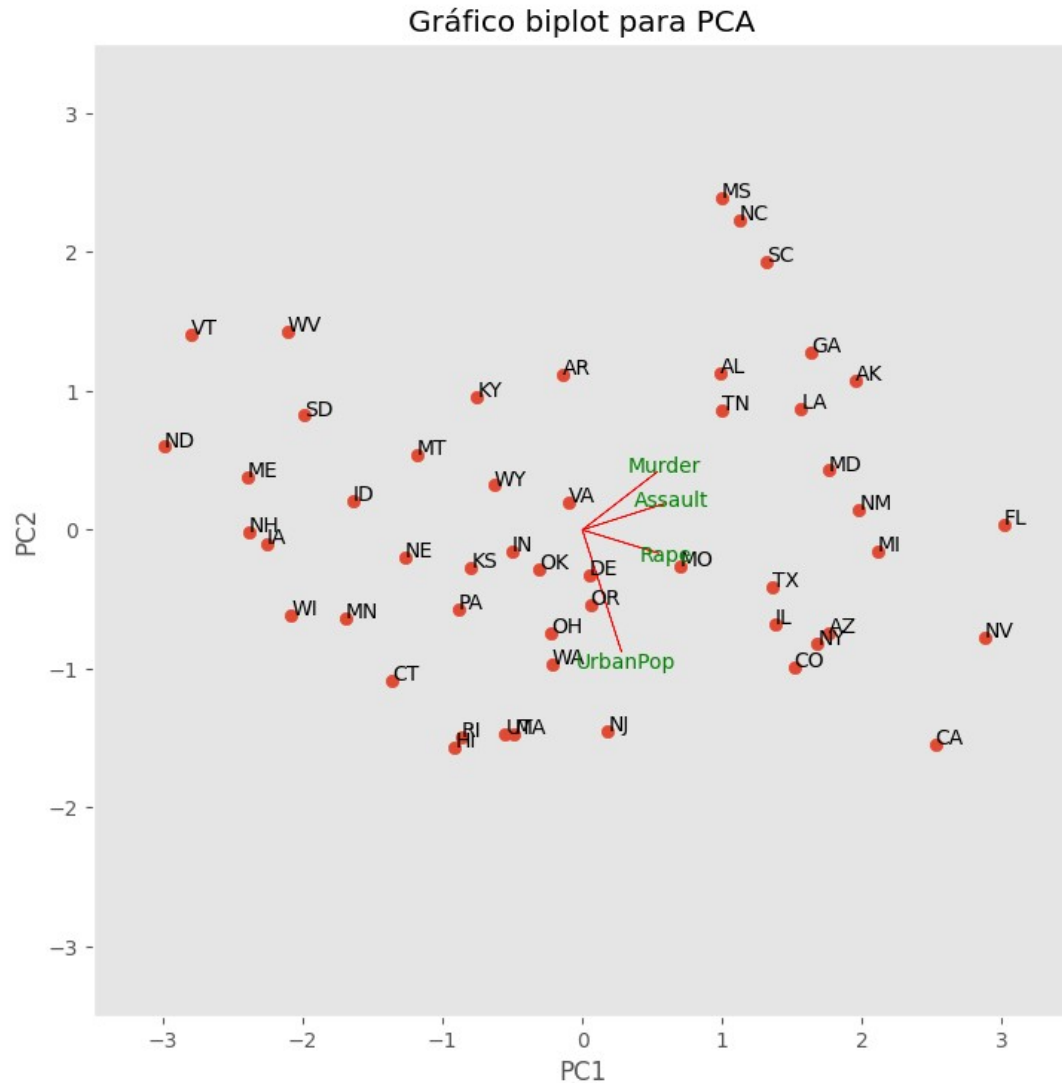
$$PC1 = a_{11}x_1 + a_{12}x_2$$
$$PC2 = a_{21}x_1 + a_{22}x_2$$

# Tutorial 8

## Teoría. PCA

- Es el algoritmo más popular de la segunda clases de algoritmos no supervisados, los métodos de proyección
  - Es válido para atributos numéricos continuos
  - Encuentra las proyecciones, combinaciones lineales, de los datos que explican la variabilidad
  - Por tanto, **permiten reducir con pérdida las dimensiones de los datos**
  - Aplicaciones:
    - Visualización
    - Preprocesamiento
    - Compresión

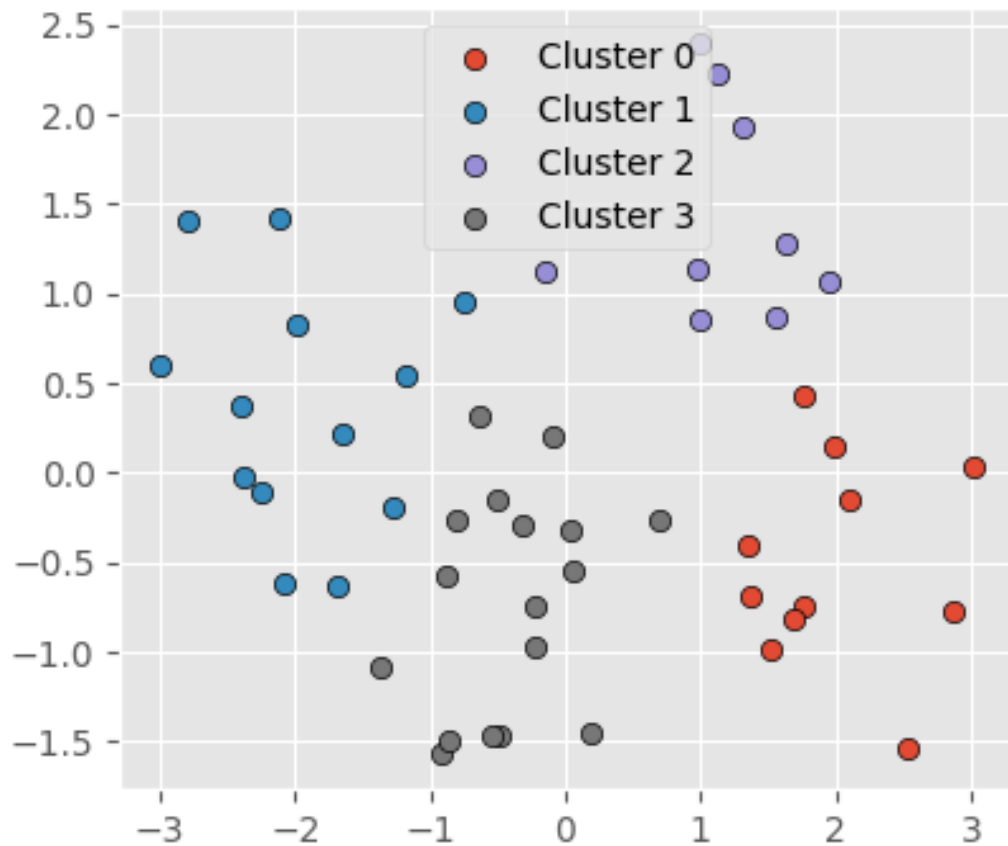
# Análisis de Componentes Principales y Clustering. Biplot



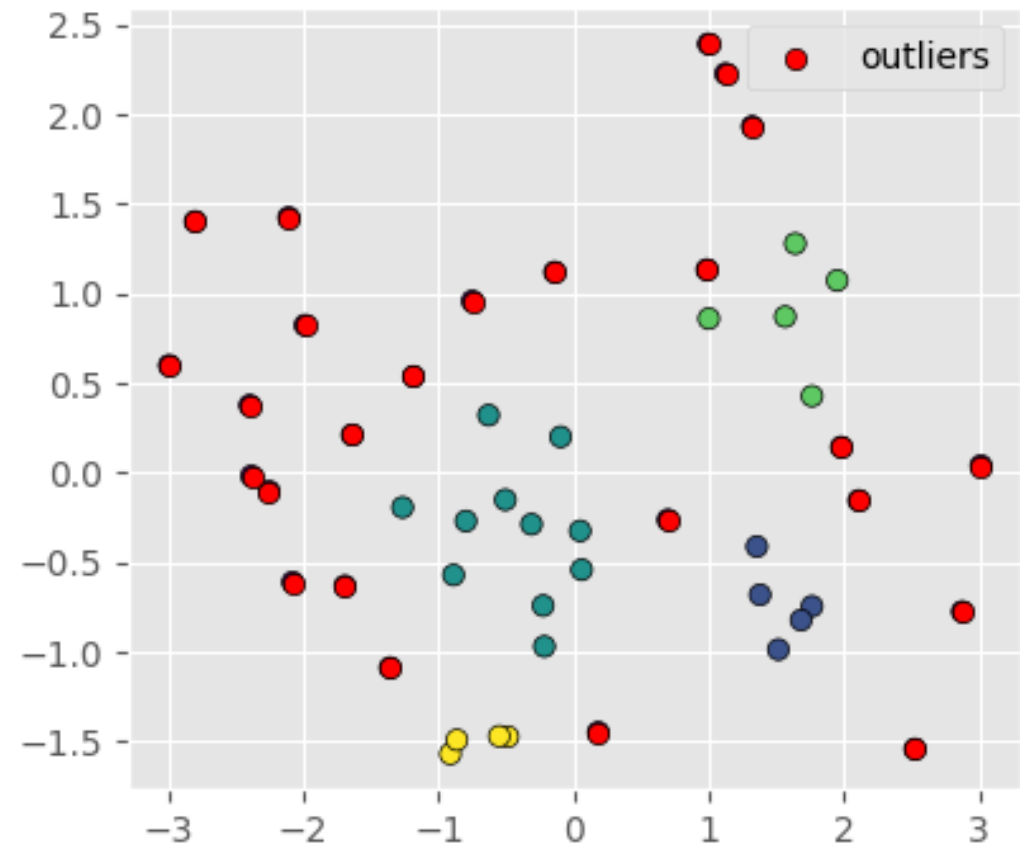
# Tutorial 8

## Análisis de Componentes Principales y Clustering.

Clusters encontrados con K-Means

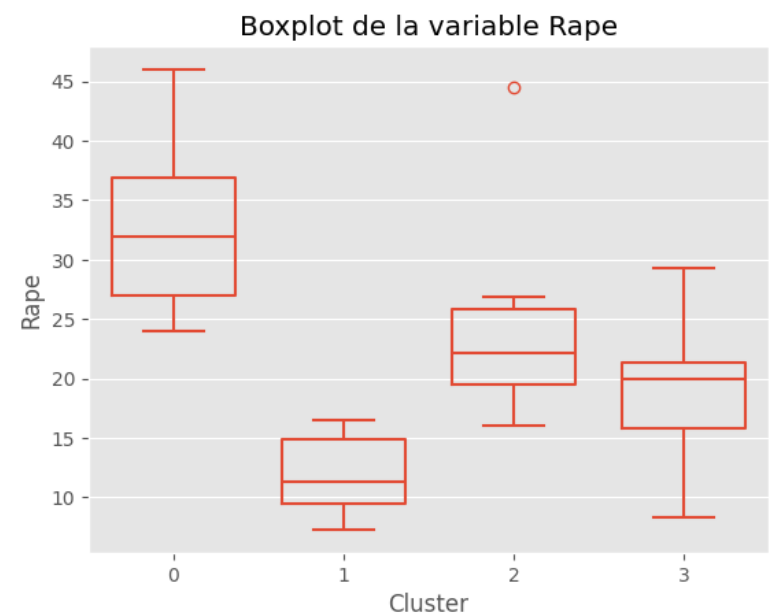
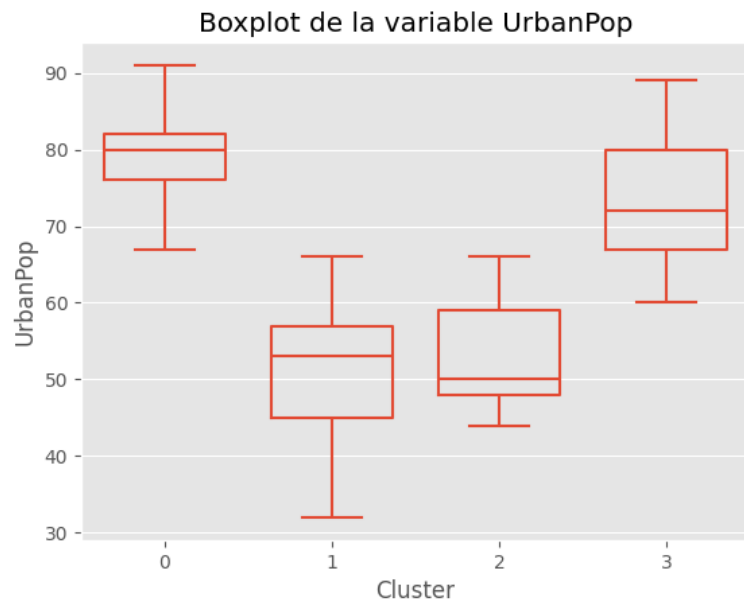
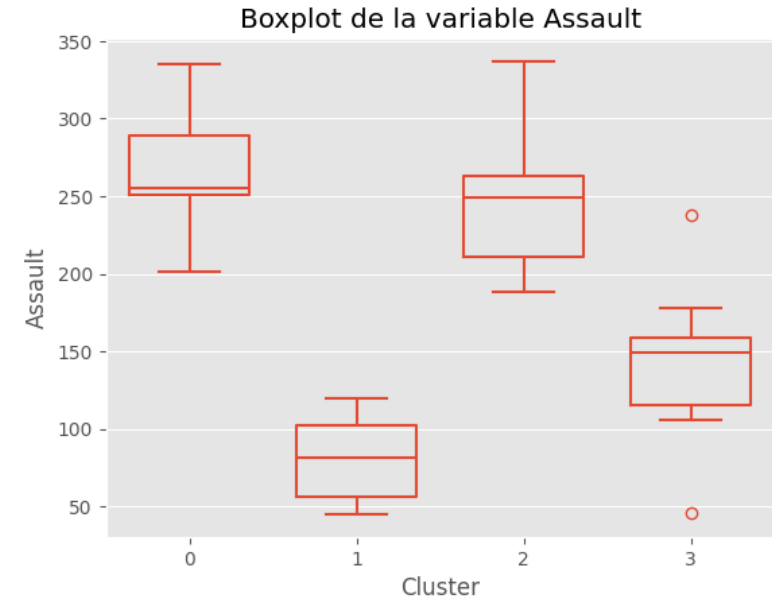
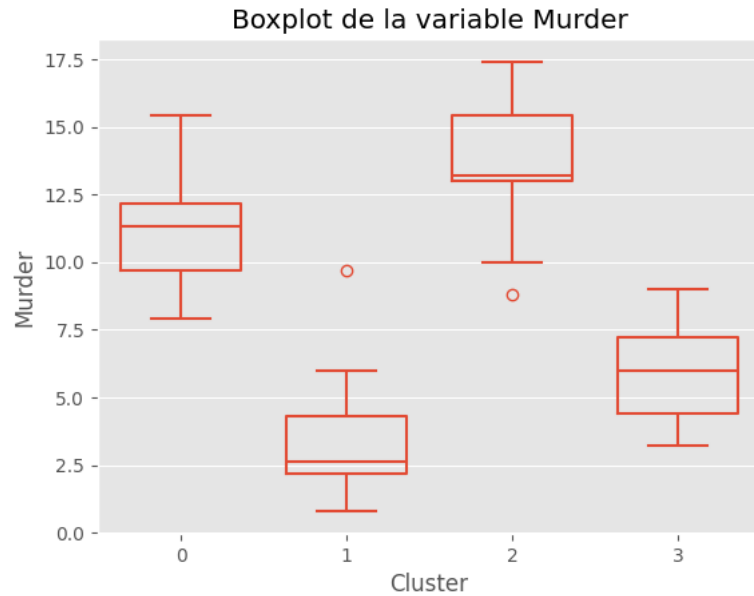


DBSCAN (clusters = 4)



# Tutorial 8

## Análisis de Componentes Principales y Clustering. Caracterización

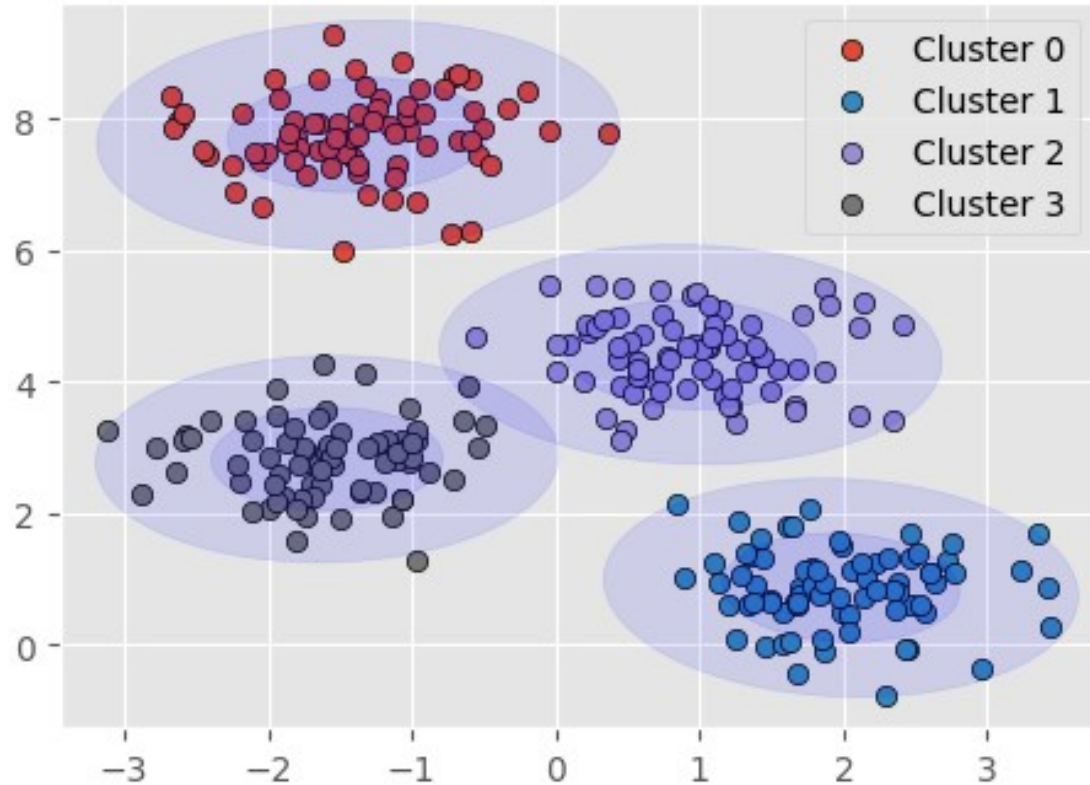




# Tutorial 8

## GMMs: Gaussian mixture models. Expectation-Maximization (EM)

Distribución de prob. de cada componente



Distribución de prob. del modelo completo

