



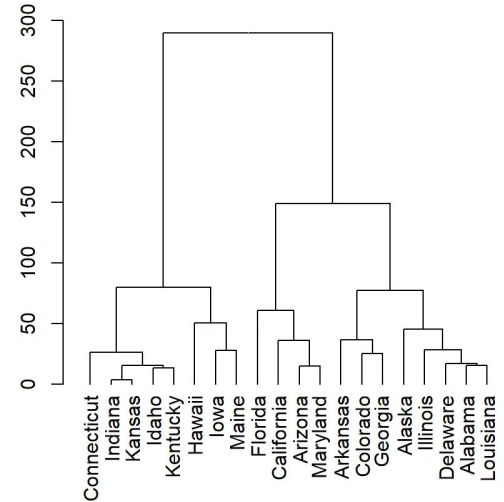
Aprendizaje Automático

**Métodos de Aprendizaje No Supervisado,
Agrupamiento Jerárquico**

2023-2Q

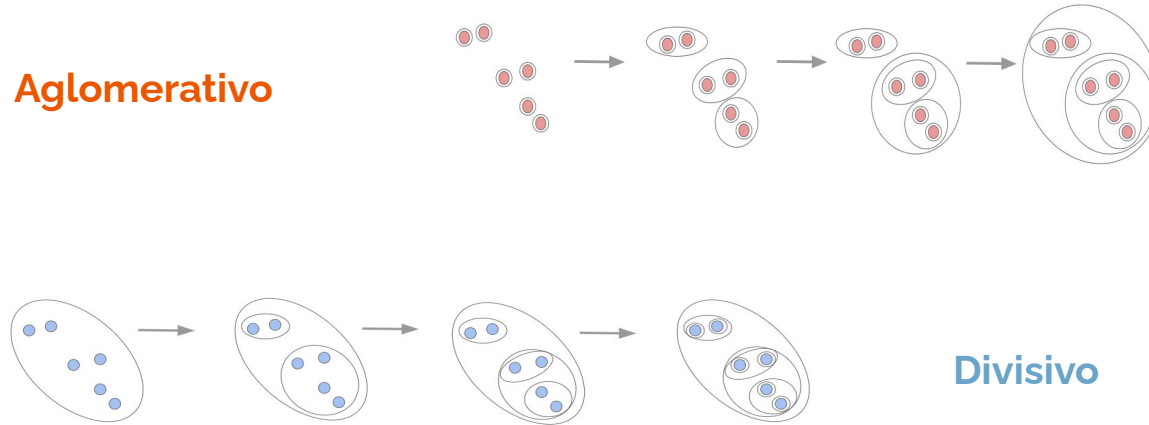
Agrupamiento Jerárquico

- Algoritmo **no supervisado** que organiza puntos de datos en una **jerarquía de clústeres** basados en su **similitud o distancia**.
- Se lo representa mediante un **dendrograma**

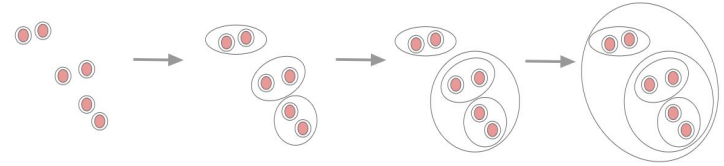


Agrupamiento Jerárquico

- El agrupamiento jerárquico tiene dos variantes: **aglomerativo** y **divisivo**.

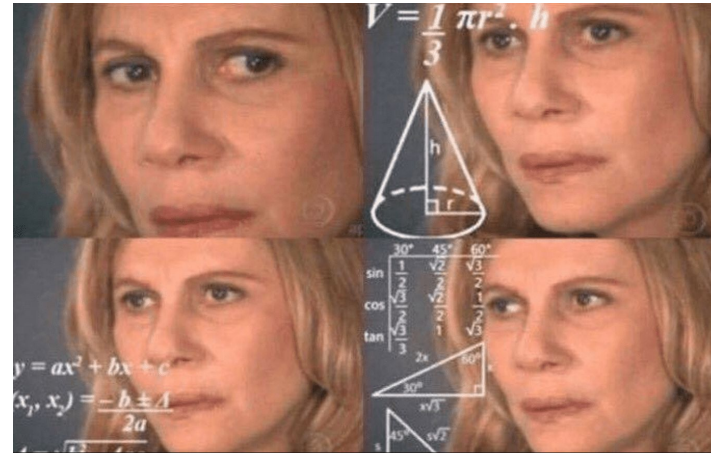
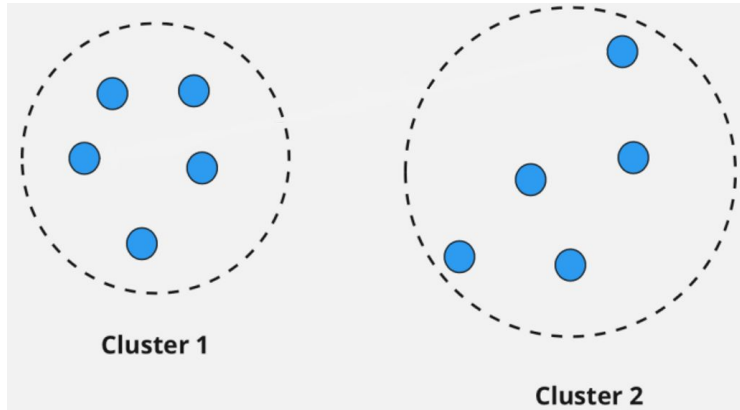


Algoritmo - Aglomerativo



1. Cada una de las n observaciones son un grupo (**grupos = n**)
2. **Tomar la distancia entre** cada uno de los **clusters**, en este paso son $\frac{n(n-1)}{2}$ **distancias**.
3. **Tomar la menor** medida entre grupos **y unir** esos dos grupos.
4. Seguir hasta que quede **un solo grupo**.

¿Cómo medimos la similitud entre grupos?

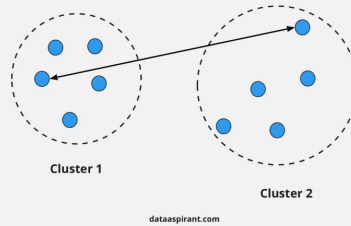


Medidas de similitud

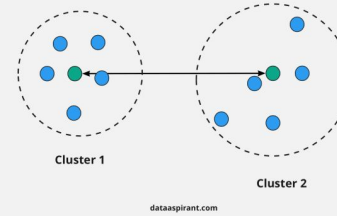
Máxima

$$d_{12} = \max_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$$

Complete Linkage Method



Centroid Linkage Method



Centroid Point

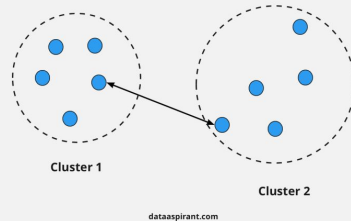
Centroide

$$d_{12} = d(\bar{\mathbf{x}}, \bar{\mathbf{y}})$$

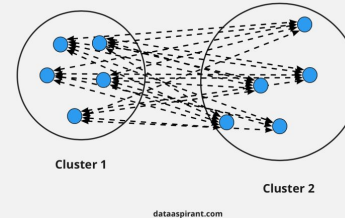
Mínima

$$d_{12} = \min_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$$

Simple Linkage Method



Average Linkage Method

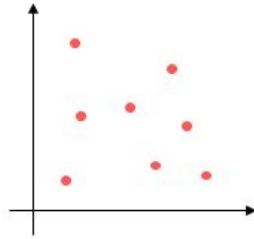


Promedio

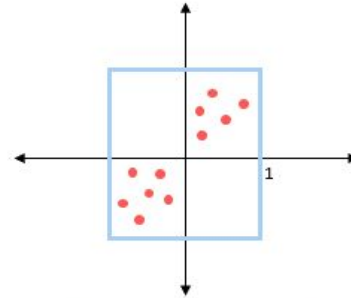
$$d_{12} = \frac{1}{kl} \sum_{i=1}^k \sum_{j=1}^l d(\mathbf{X}_i, \mathbf{Y}_j)$$

Estandarizar las variables

Cuando las variables están en diferentes escalas, es conveniente estandarizarlas para que sean comparables.



Actual Data



After standardization

Distancia Correlación

x, y observaciones d dimensión n

$$d_{cor}(x, y) = 1 - \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{(\sum_{i=1}^n (x_i - \bar{x})^2 * \sum_{i=1}^n (y_i - \bar{y})^2)^{\frac{1}{2}}}$$

Mide la similitud en términos de correlación lineal. **Si es 0, entonces son l.i.**

Ejemplo: Similitud con la Distancia Euclídea

Puntos en el plano

$p1 = (0.40, 0.53)$

$p2 = (0.22, 0.38)$

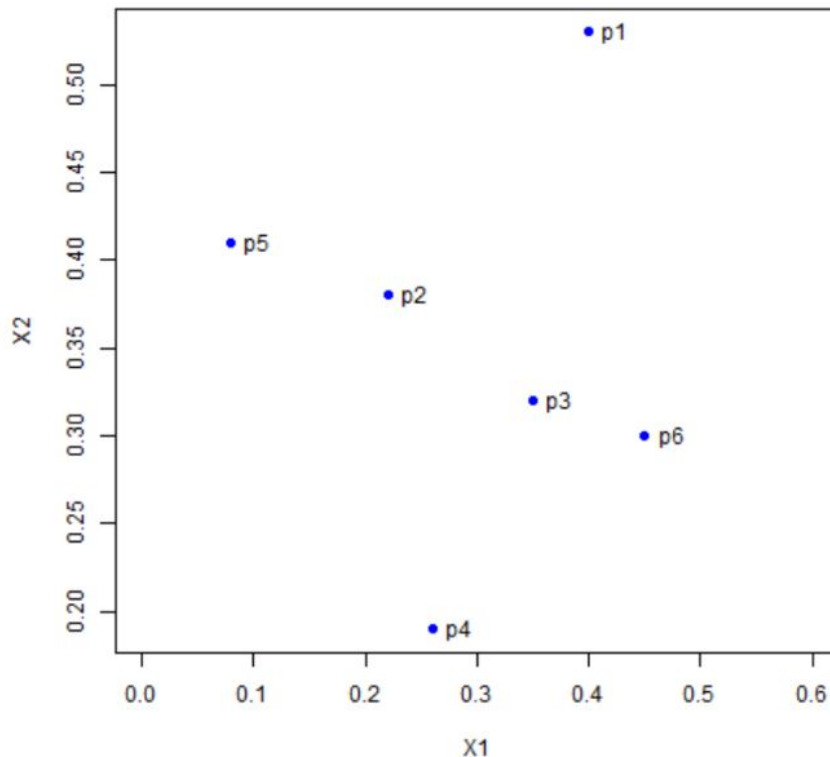
$p3 = (0.35, 0.32)$

$p4 = (0.26, 0.19)$

$p5 = (0.08, 0.41)$

$p6 = (0.45, 0.30)$

¿Qué par de puntos están
más cerca?



Ejemplo: Similitud con la Distancia Euclídea

Matriz de distancias

	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$
$p1$	0	0,23	0,22	0,37	0,34	0,24
$p2$	0,23	0	0,14	0,19	0,14	0,24
$p3$	0,22	0,14	0	0,16	0,28	0,10
$p4$	0,37	0,19	0,16	0	0,28	0,22
$p5$	0,34	0,14	0,28	0,28	0	0,39
$p6$	0,24	0,24	0,10	0,22	0,39	0

Matriz simétrica

Agrupamos



- Ahora tenemos los grupos: **{p1, p2, {p3, p6}, p4, p5}**
- Tomar la distancia entre todos los grupos, por ejemplo: grupo 1 = p1 con todos los otros grupos:

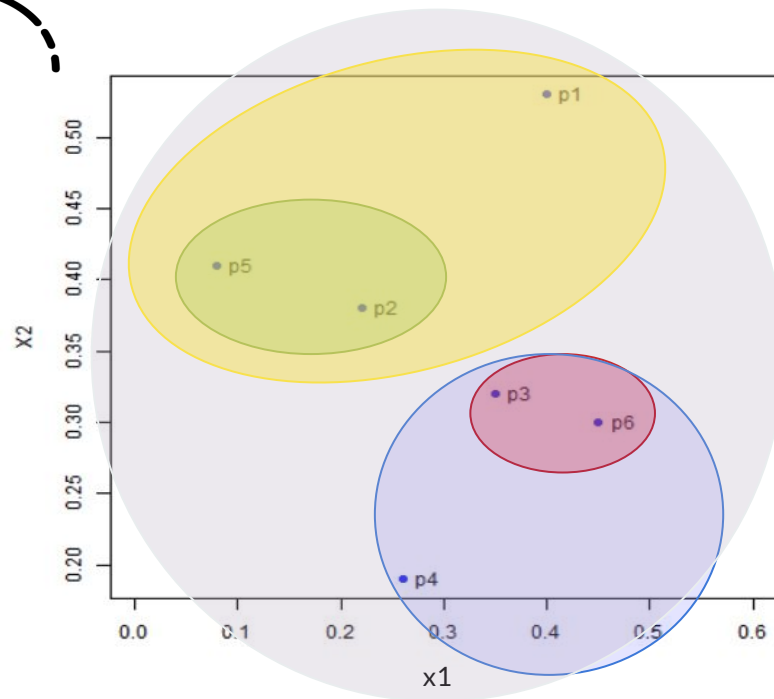
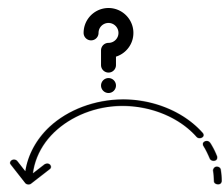
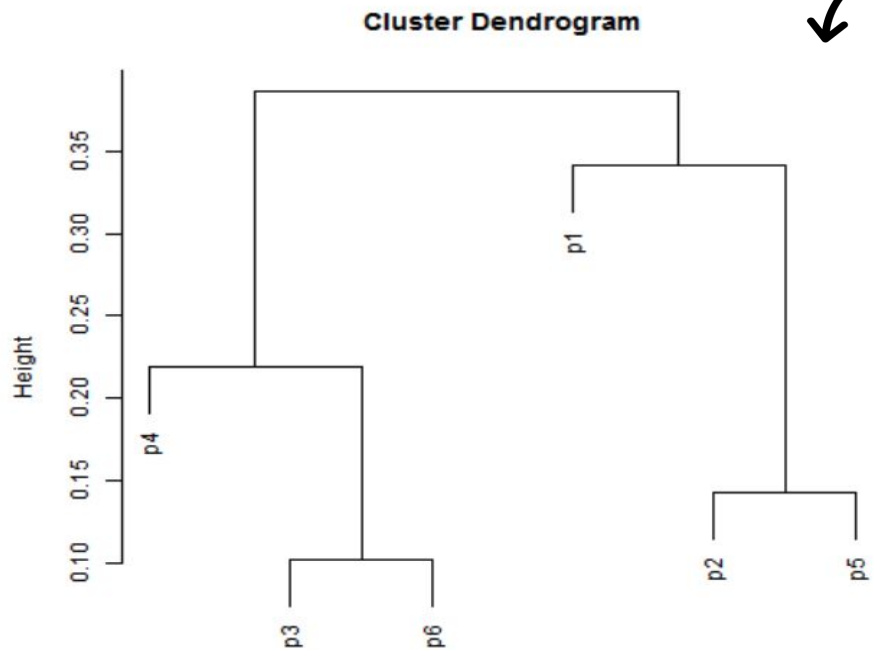
$\text{dist}(p1, p2) = 0,23$

$\text{dist}(p1, \{p3, p6\})$

...

- Elegimos la mínima para fusionar y eso nos da agrupar p2 y p5 entonces nos quedan los grupos: **{p1, {p2, p5}, {p3, p6}, p4}**

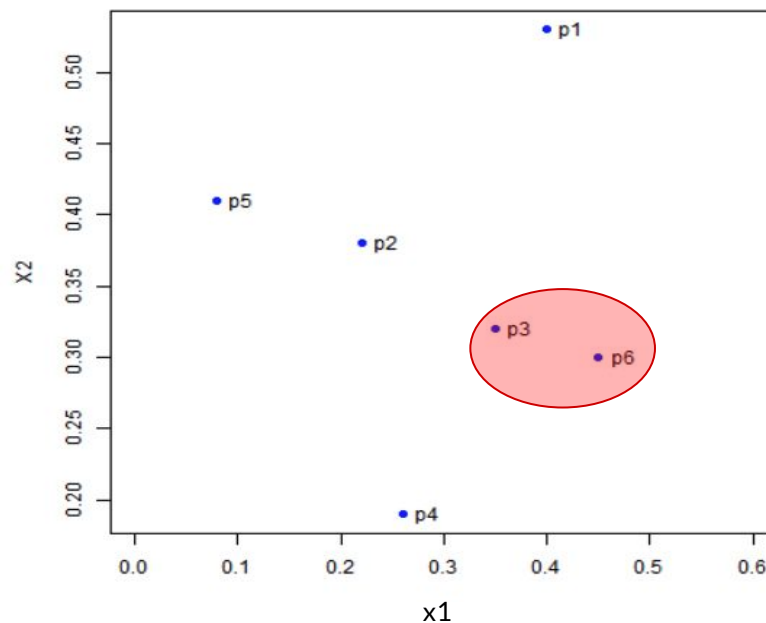
Resultado



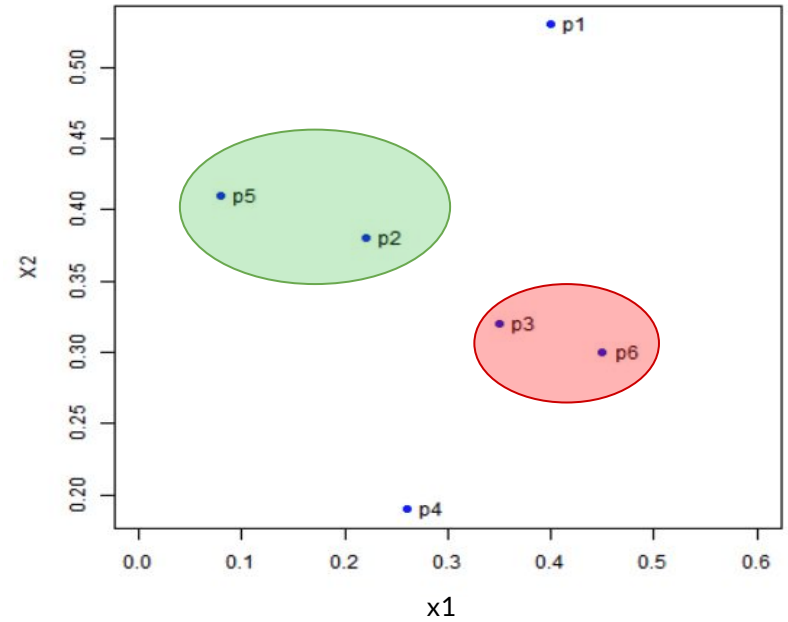
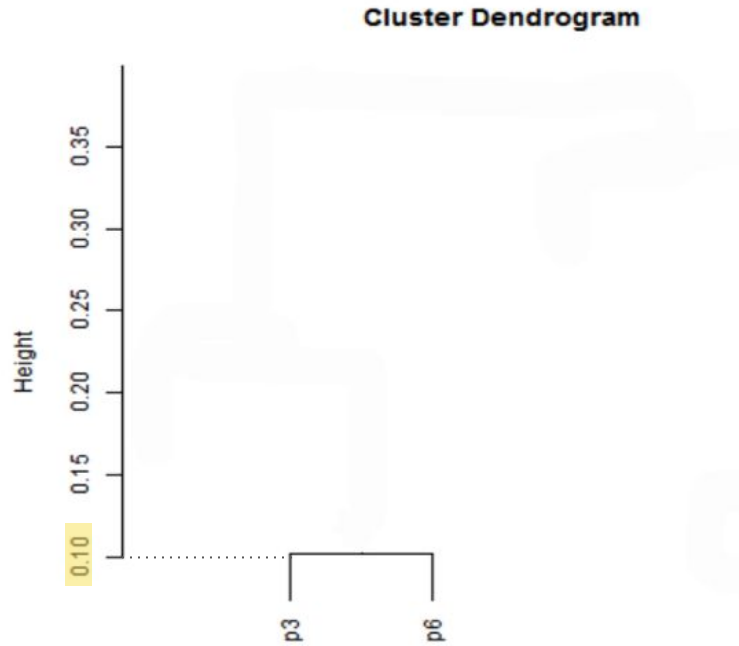
Resultado

Matriz de distancias

	p1	p2	p3	p4	p5	p6
p1	0	0,23	0,22	0,37	0,34	0,24
p2	0,23	0	0,14	0,19	0,14	0,24
p3	0,22	0,14	0	0,16	0,28	0,10
p4	0,37	0,19	0,16	0	0,28	0,22
p5	0,34	0,14	0,28	0,28	0	0,39
p6	0,24	0,24	0,10	0,22	0,39	0



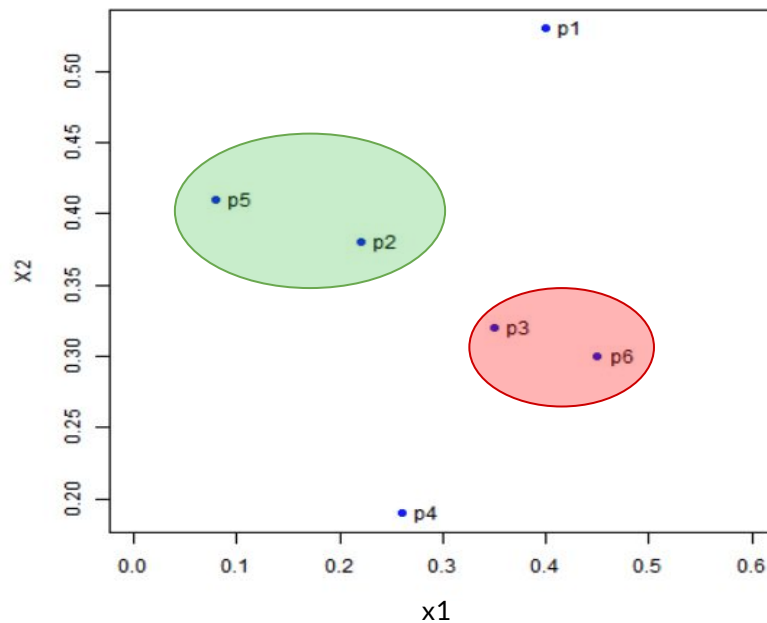
Resultado



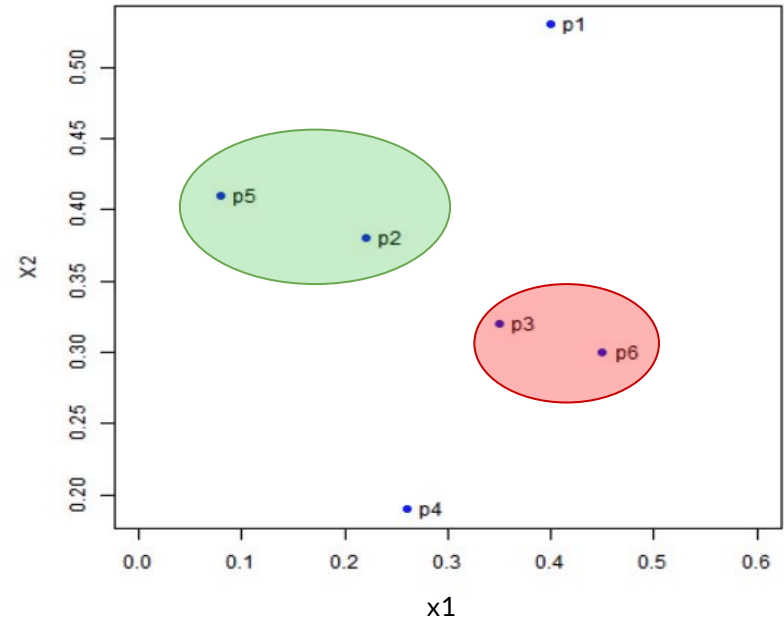
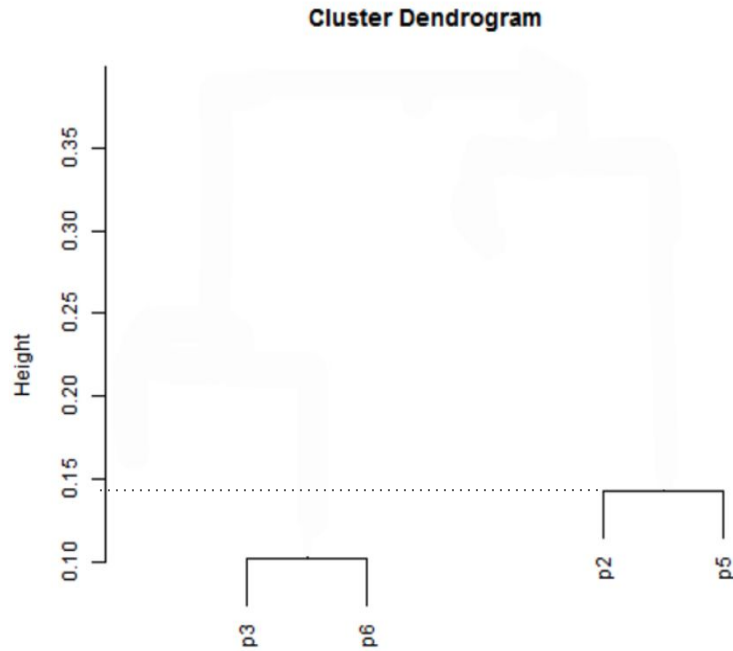
Resultado

Matriz de distancias

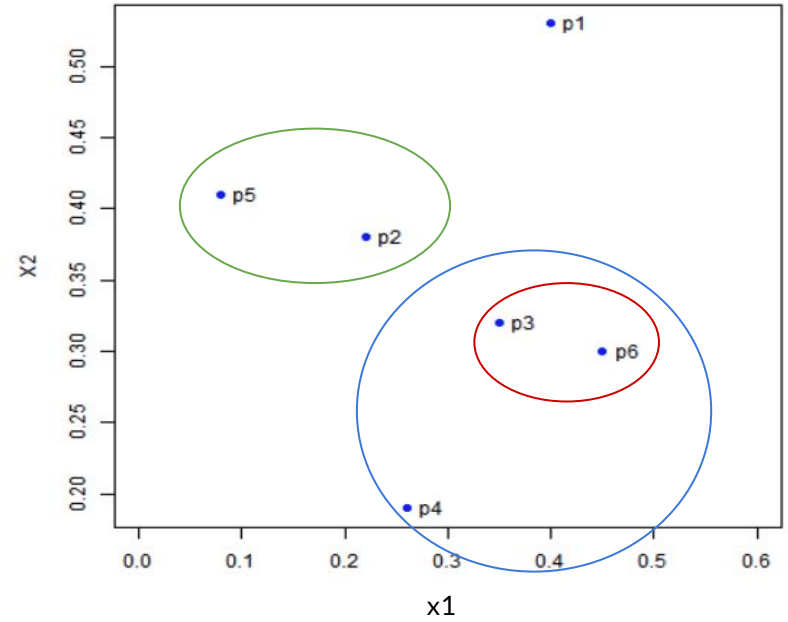
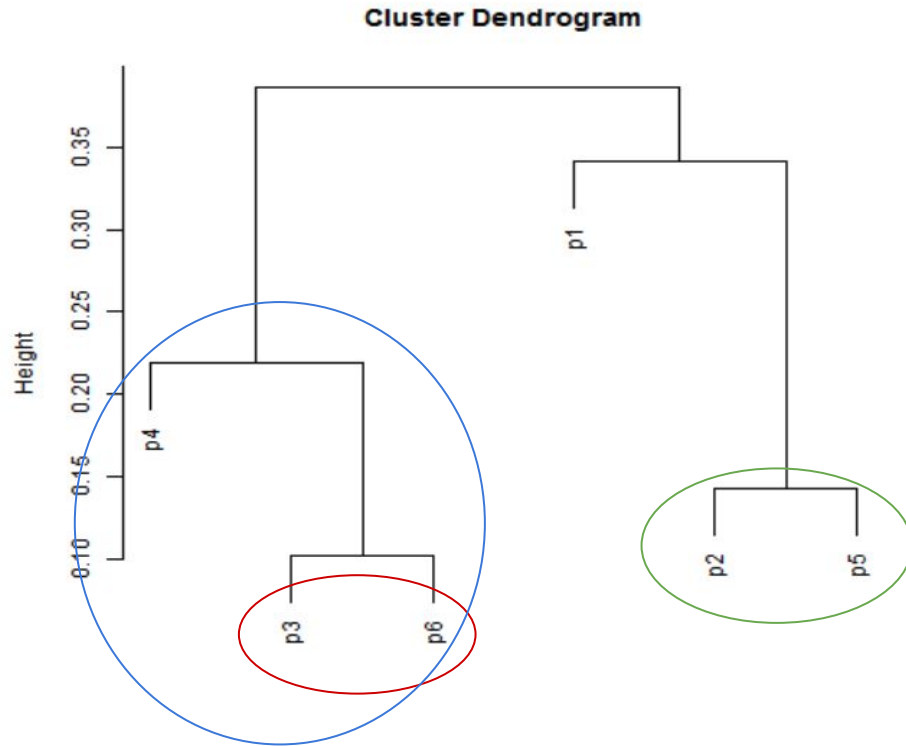
	p1	p2	p3	p4	p5	p6
p1	0	0,23	0,22	0,37	0,34	0,24
p2	0,23	0	0,14	0,19	0,14	0,24
p3	0,22	0,14	0	0,16	0,28	0,10
p4	0,37	0,19	0,16	0	0,28	0,22
p5	0,34	0,14	0,28	0,28	0	0,39
p6	0,24	0,24	0,10	0,22	0,39	0



Resultado



Resultado

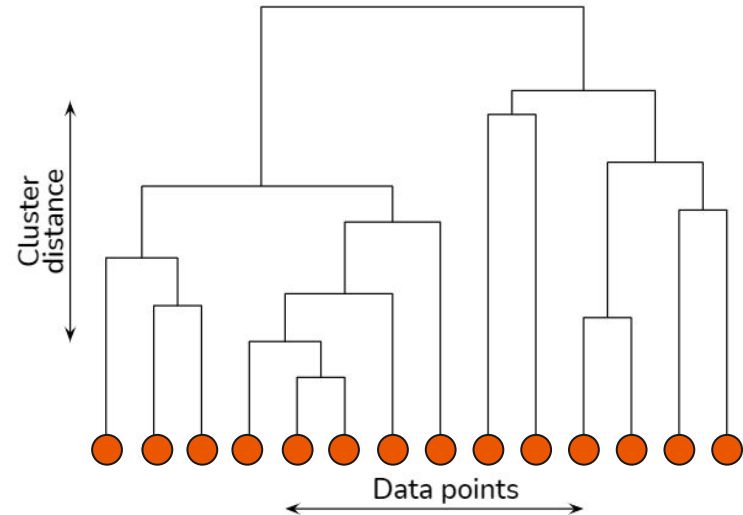


Dendrograma

Interpretación

De abajo hacia arriba

- Las **hojas** representan las observaciones.
- **Fusión entre hojas:**
 - corresponde a las observaciones similares.
 - más abajo en el dendrograma significa mayor similitud.



Agrupamiento Jerárquico

- El agrupamiento jerárquico tiene dos variantes: **aglomerativo** y **divisivo**.



Similitud en las observaciones

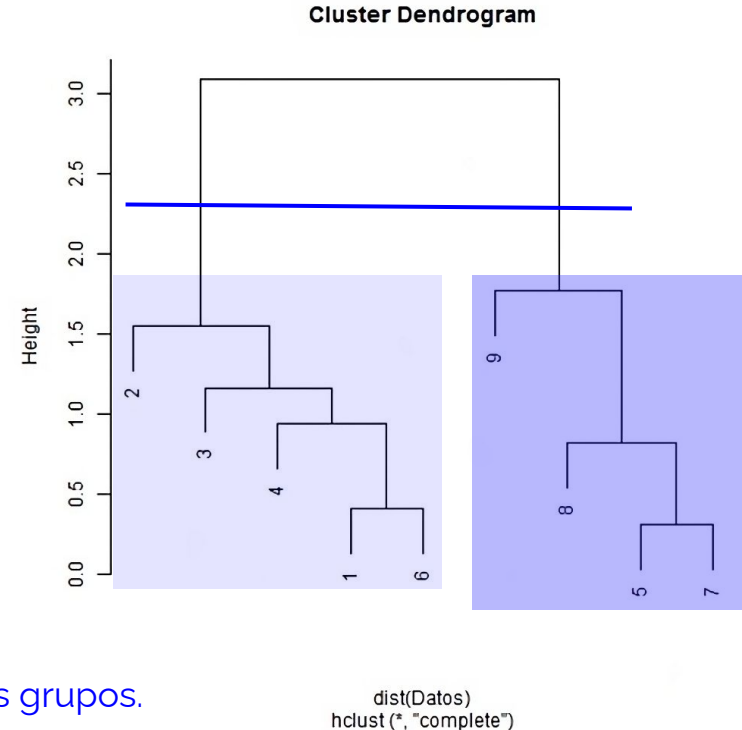


Para cualquier par de observaciones miramos el punto en el dendrograma donde las ramas se fusionan por primera vez, **la altura de esta fusión, medida en el eje Y** indica la **disimilitud** entre las observaciones.

¿Cómo obtenemos la cantidad de clusters?

A partir de un **umbral de similitud o distancia**, que lo podemos representar con una línea.

La cantidad de grupos se va a corresponder con la cantidad de veces que cruza esta línea con el dendrograma.

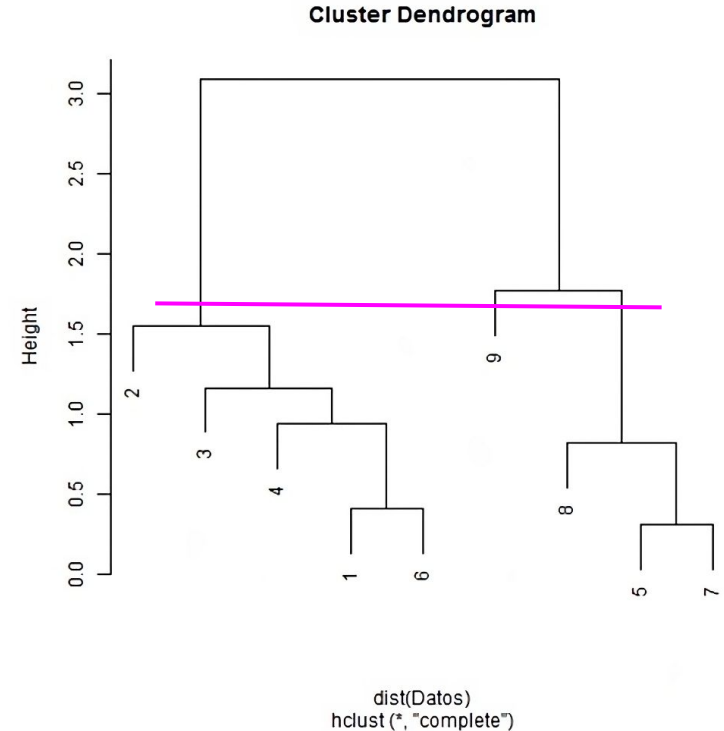


Línea azul -> cruza dos veces el dendrograma -> dos grupos.

Cantidad de clusters - Ejemplo

¿Cuántos grupos se obtienen?

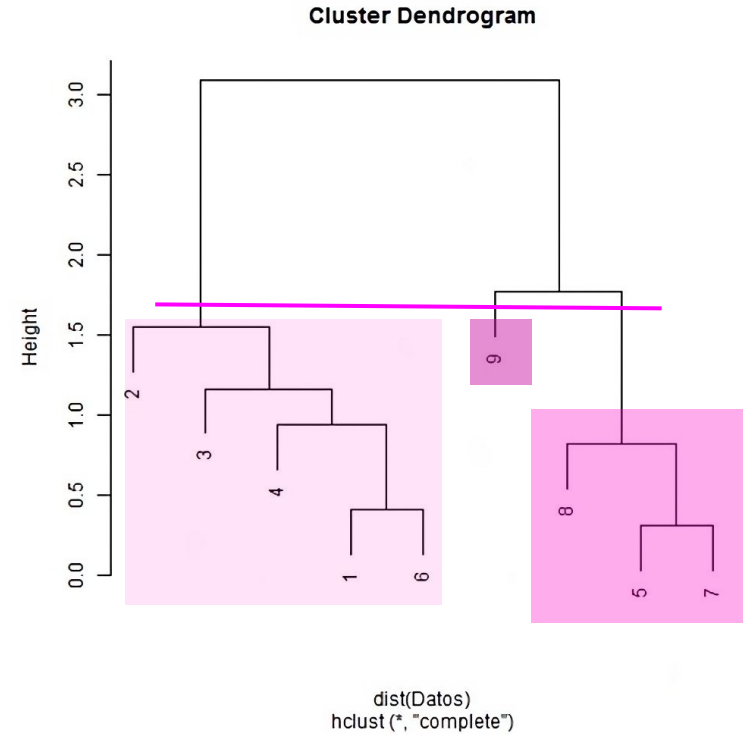
Línea rosa: ___ grupos



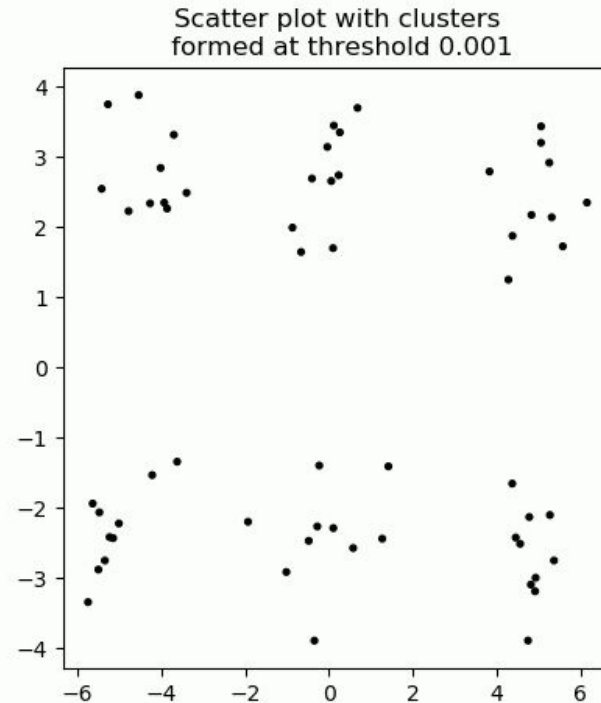
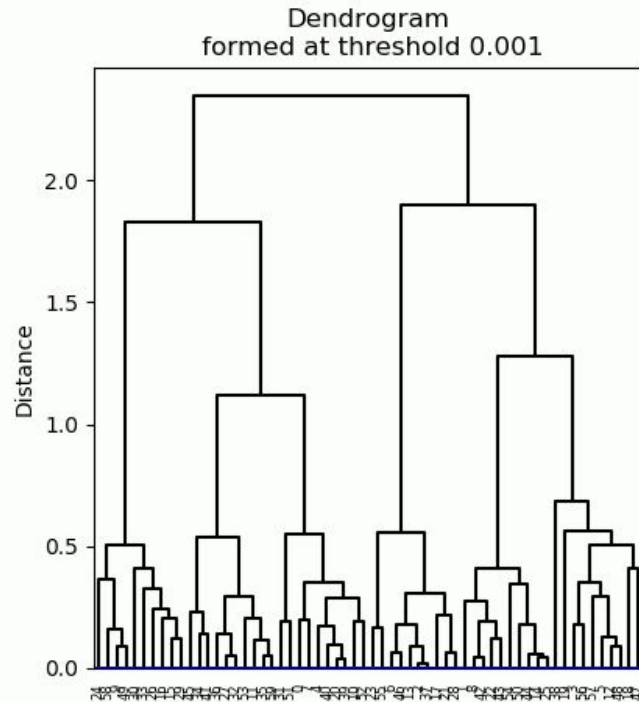
Cantidad de clusters - Ejemplo



Línea rosa: tres grupos



Cantidad de clusters



Similitud en las observaciones



Ventajas

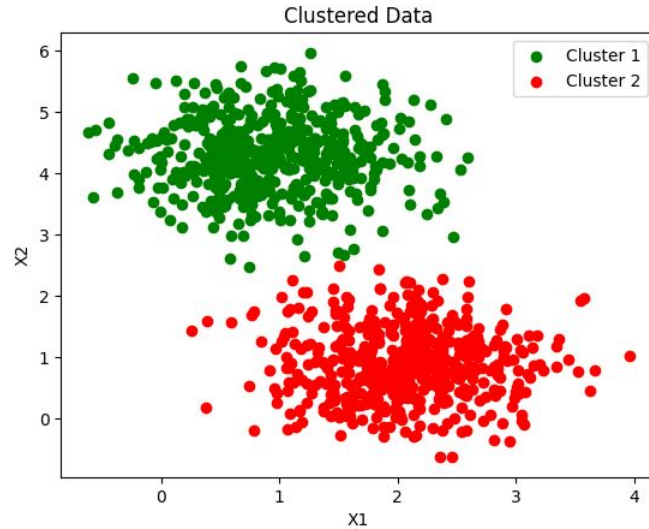
- Con un solo dendrograma es posible agrupar en la cantidad de grupos que se desee.

Desventajas

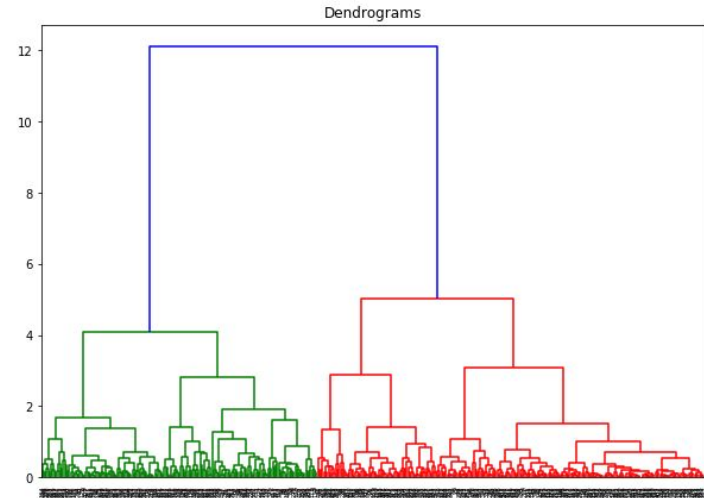
- El término jerárquico se refiere a que las clases están anidadas, pero esto no siempre ocurre.

K-medias vs Agrupamiento jerárquico

K-means




Agrupamiento jerárquico




K-medias vs Agrupamiento jerárquico



K-means

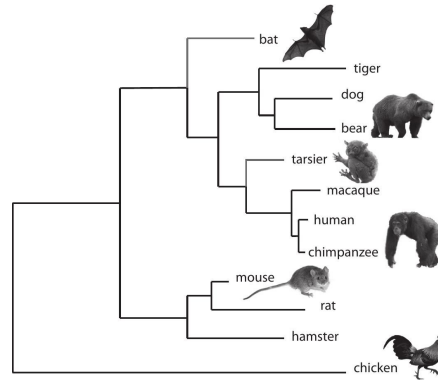
- Requiere que se especifique la cantidad de grupos (**k**).
 - Agrupamiento **particional**.
 - **Resultados variables** (dependen de inicialización de centroide).
 - **Menor costo computacional**.
- 

Agrupamiento jerárquico

- **No** requiere que se especifique la cantidad de grupos.
 - Agrupamiento **jerárquico**.
 - **Resultados reproducibles** (a igual medida de similitud).
 - Puede requerir **mayor costo computacional** (grandes datasets)
 - Se puede utilizar inicialmente de manera exploratoria (por ejemplo: **buscar k**)
- 

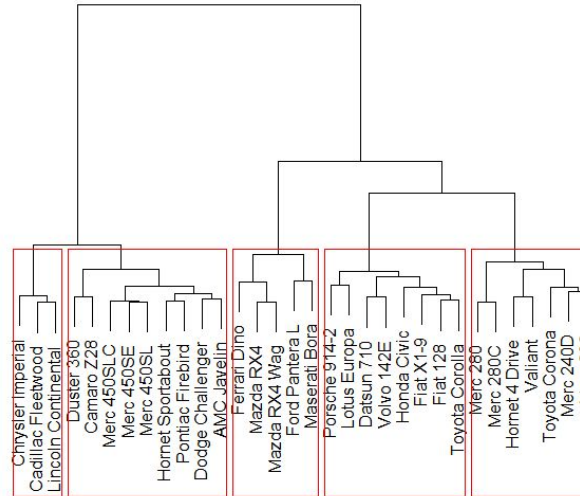
Usos de Agrupamiento Jerárquico

- **Biología molecular, Genética y Bioinformática:**
 - Clasificación de secuencias de ADN y proteínas para identificar familias genéticas o proteicas.
 - Estudio de relaciones filogenéticas entre especies.



Usos de Agrupamiento Jerárquico

- **Marketing y Segmentación de Clientes** → recomendaciones personalizadas



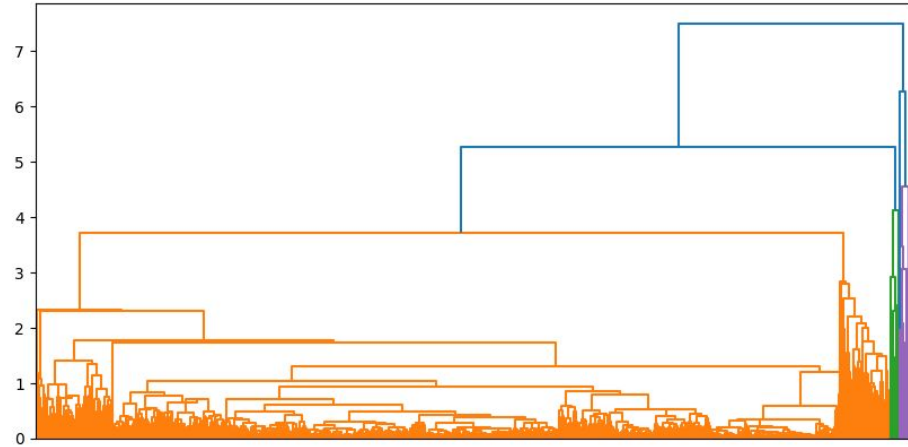
Dendograma

Consideraciones para el TP

Librerías

```
scipy.cluster.hierarchy  
plotly.figure_factory.create_dendrogram
```

Para hacer el dendograma a partir de la matriz de distancias



Ejercicio

Realizar el ejercicio utilizando otras distancias.

Puntos en el plano

$p1 = (0,40 \ 0,53)$

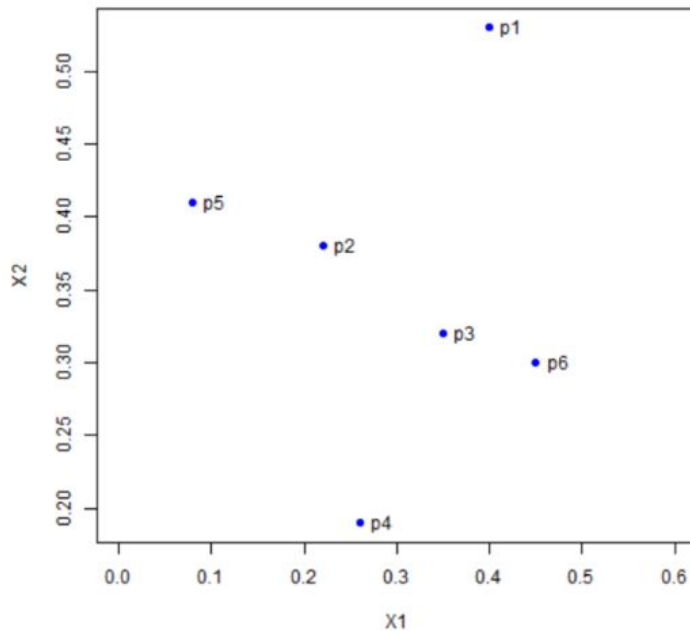
$p2 = (0,22 \ 0,38)$

$p3 = (0,35 \ 0,32)$

$p4 = (0,26 \ 0,19)$

$p5 = (0,08 \ 0,41)$

$p6 = (0,45 \ 0,30)$





iFin! ¿Alguna pregunta?