TOULOUSE
LAUTREC

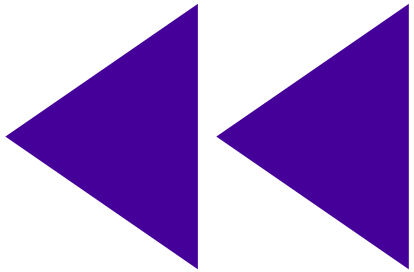
CLUSTERING

Objetivo de la clase

Al final de esta clase, el estudiante podrá desarrollar modelos de aprendizaje no supervisado en Python y hacer despliegues del modelo entrenado en internet.

Palabras clave: aprendizaje no supervisado, clustering, Hugging Face

Recordando conceptos ...



Recordando algunos conceptos ...

1. **Aprendizaje supervisado** es cuando **se tienen** bien **definidos los predictores y la respuesta**.

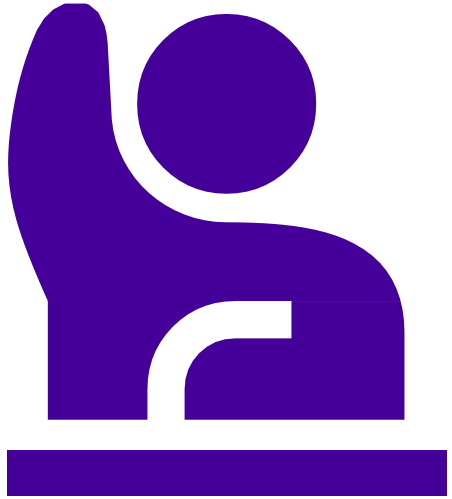
2. **Aprendizaje NO supervisado** es cuando **se tienen** bien **definidos los predictores**, pero **no está** bien **definida la respuesta**. Permite comprender la variación y estructura de agrupación de conjunto de datos.

3. No existe un gold standard (target) ni un objetivo único (precisión del dataset de test).

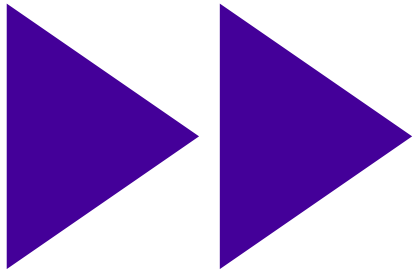
4. **Clustering** es una técnica de Machine Learning para **dividir datos en grupos similares**. **Cada instancia**/observación **debe ser similar y** al mismo tiempo **diferente entre los grupos**. Se utiliza una medida de similitud entre elementos.

Para participar:

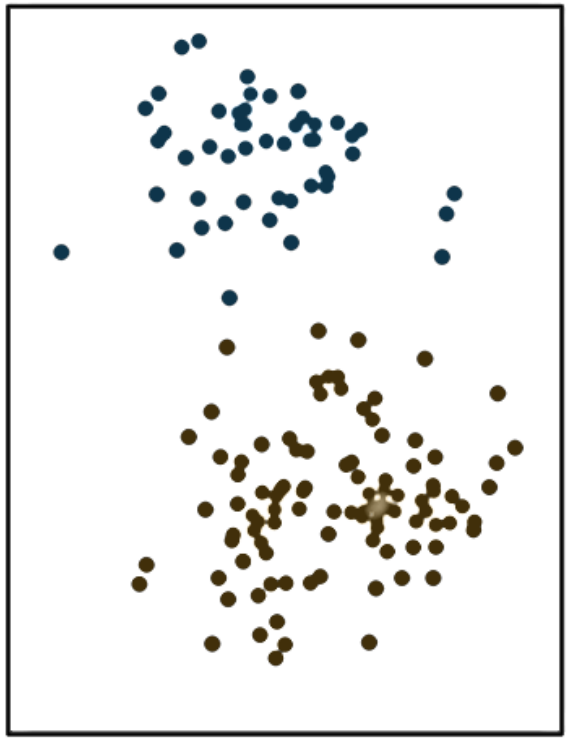
¿Qué algoritmos de aprendizaje no supervisado conoces?



Continuamos ...



K-means / K-medias

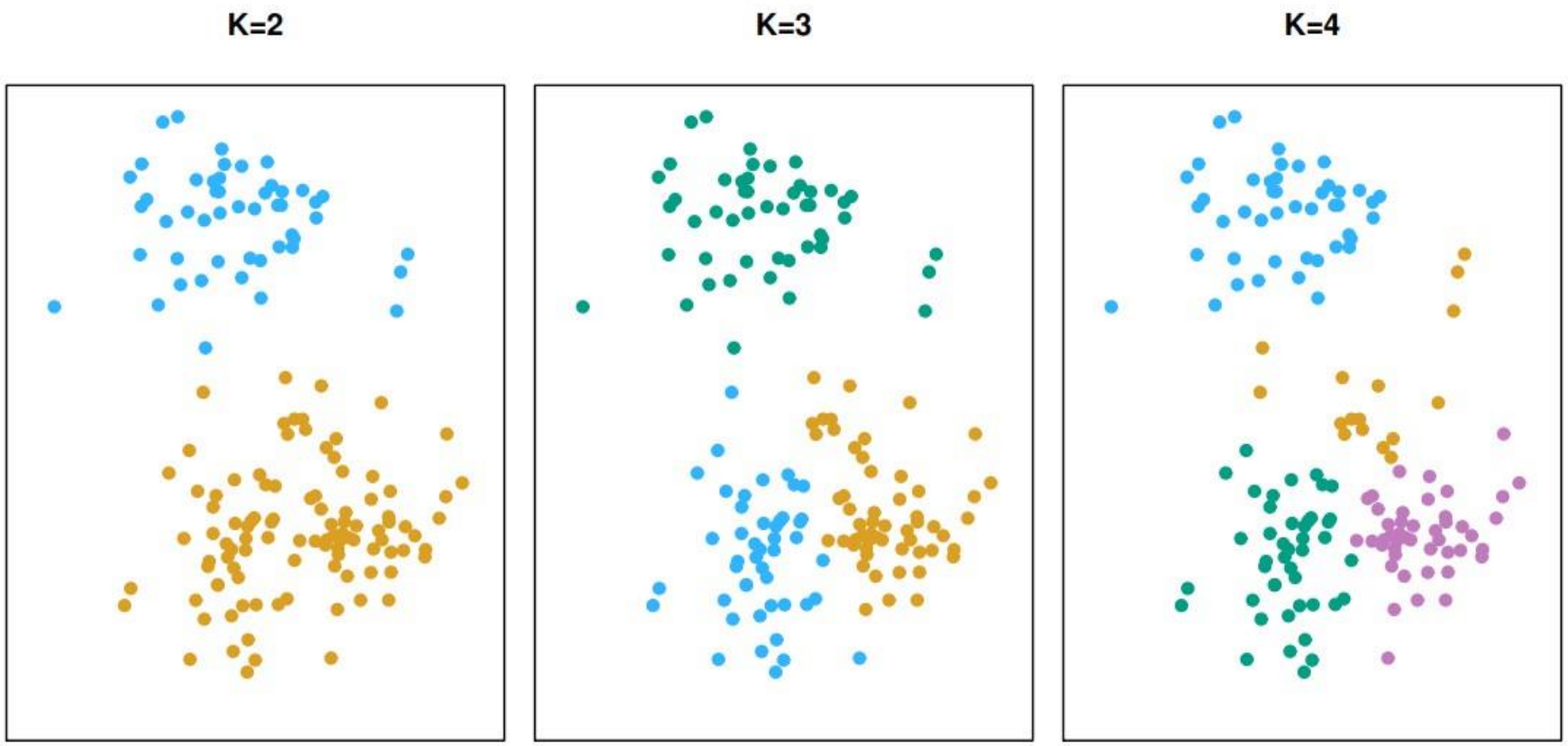


El algoritmo consiste en asignar cada uno de los N ejemplos a uno de los K clusters, donde K es un número definido previamente.

El objetivo es minimizar las diferencias entre los grupos de cada cluster y maximizar las diferencias entre clusters.

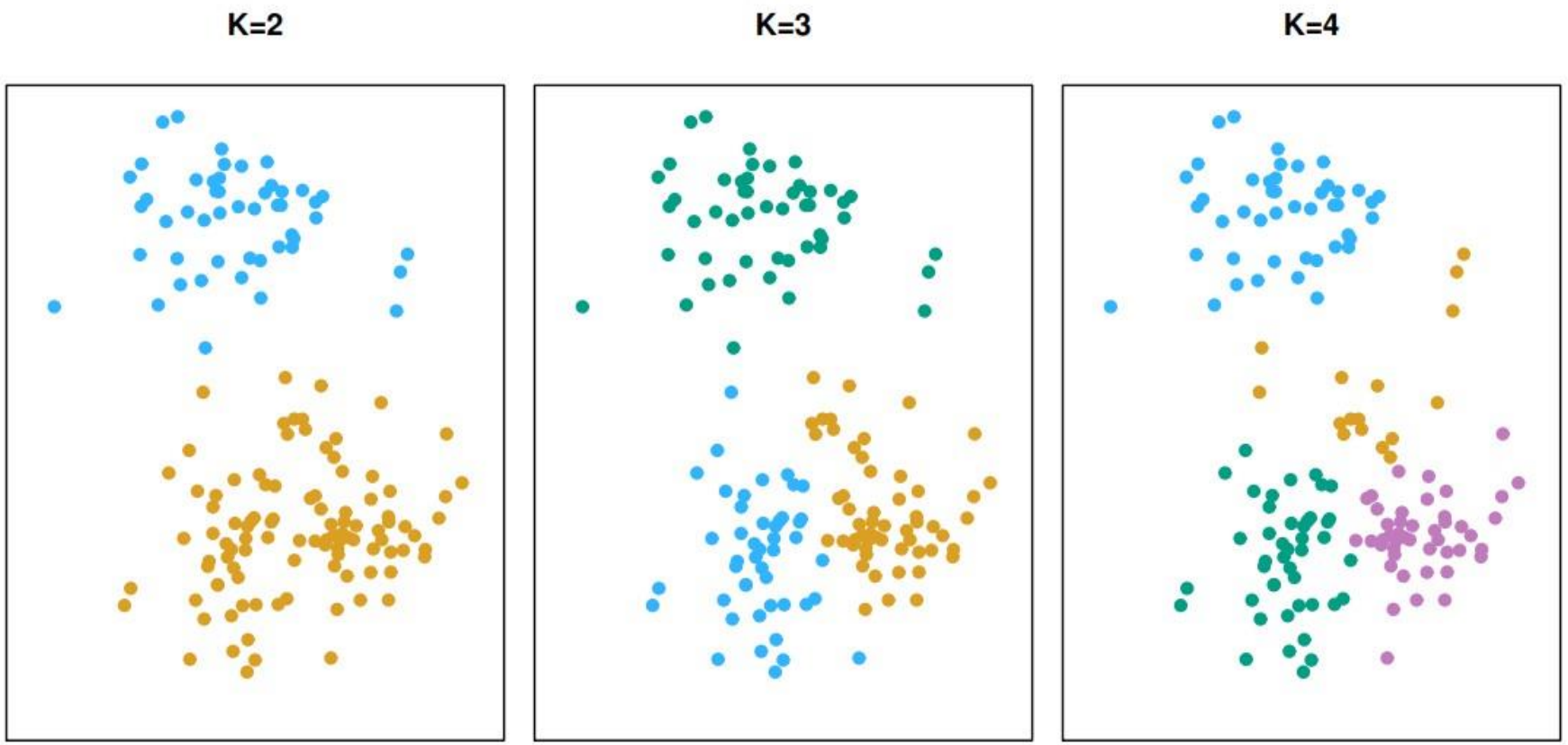
150 observaciones

K-means / K-medias



150 observaciones

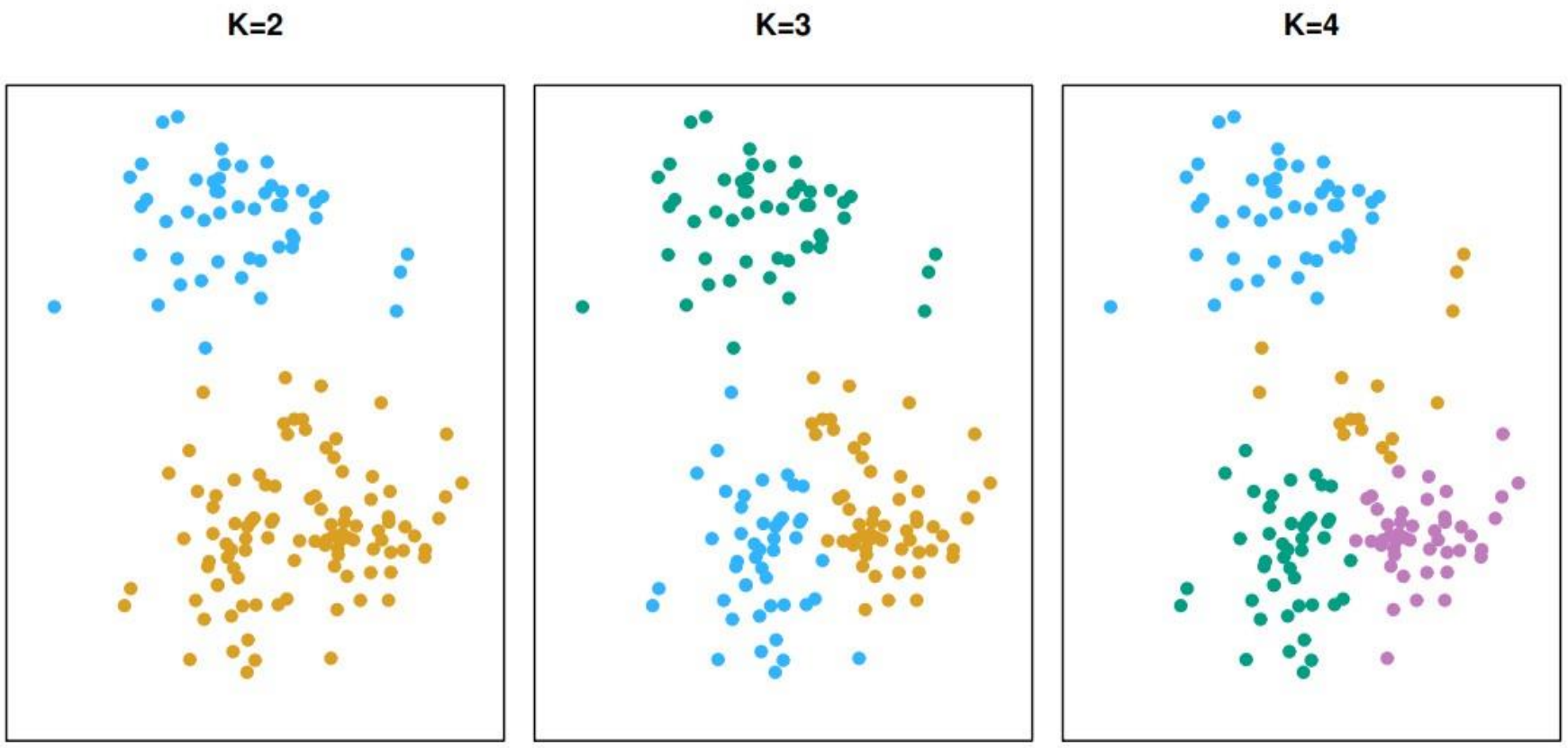
K-means / K-medias – Elegir el valor de K



$$K = \sqrt{\frac{N}{2}}$$

Donde:
K: Cantidad de clusters
N: Cantidad de observaciones

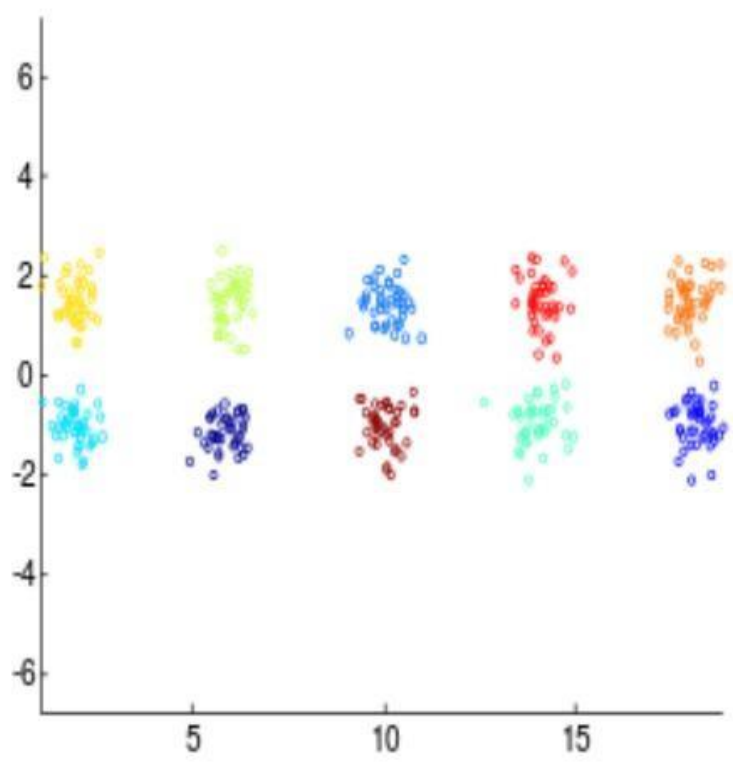
K-means / K-medias – Elegir el valor de K

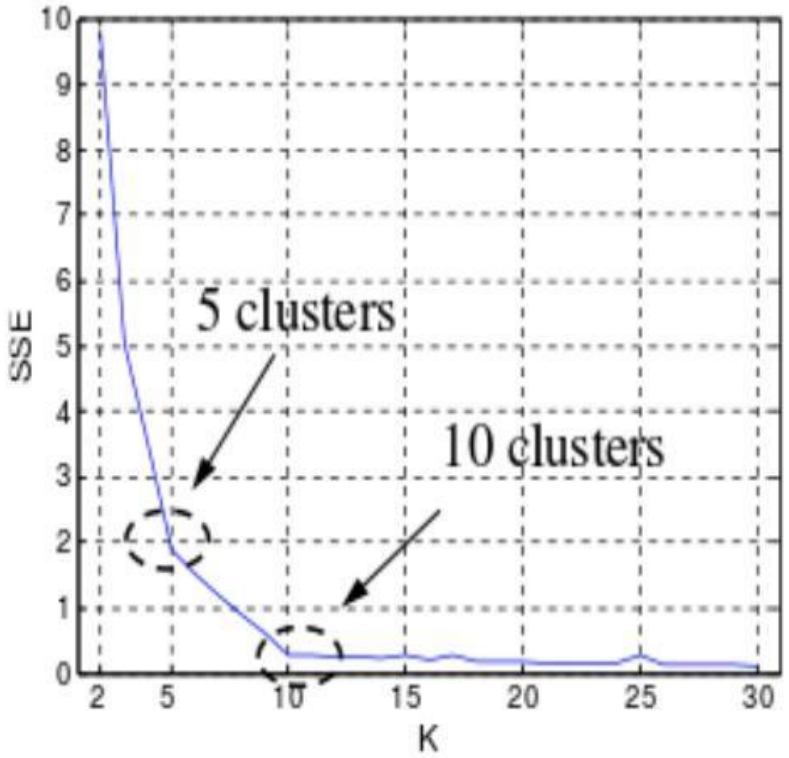


$$K = \sqrt{\frac{N}{2}} = \sqrt{\frac{150}{2}} = 8.66$$

K-means / K-medias – Elegir el valor de K

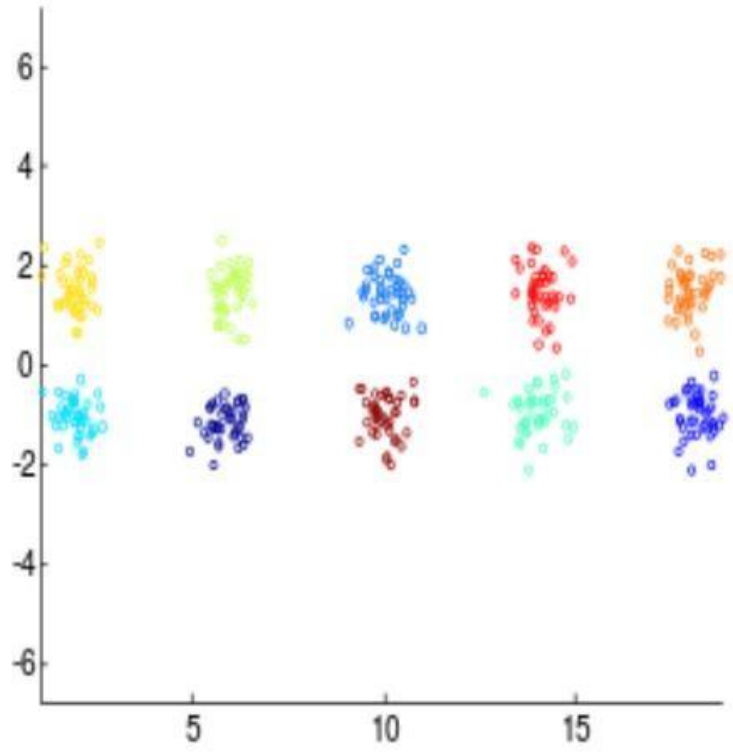
Método del codo (elbow method) para encontrar el número óptimo de clústeres.

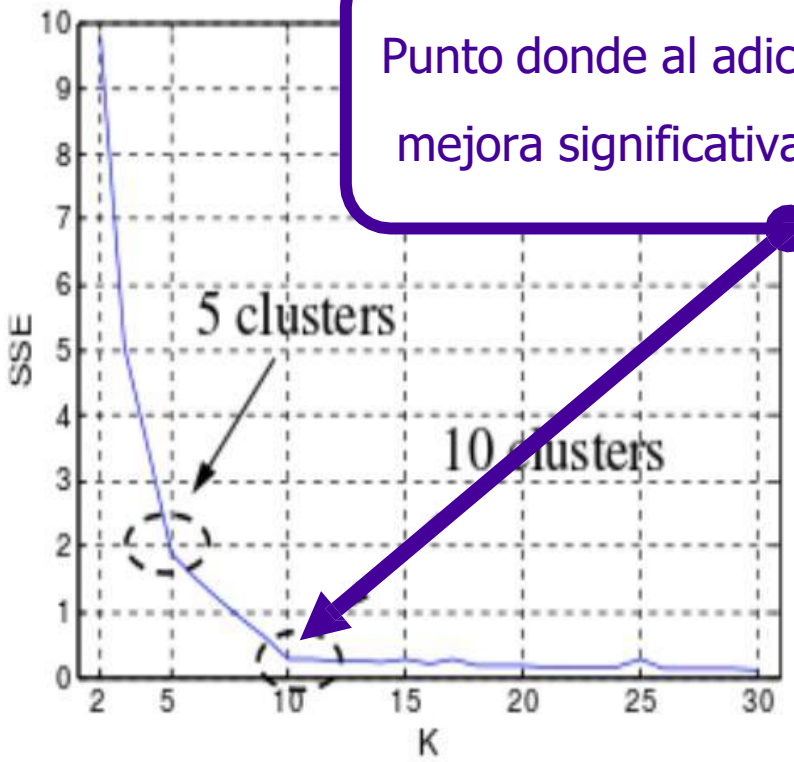




K-means / K-medias – Elegir el valor de K

Método del codo (elbow method) para encontrar el número óptimo de clústeres.





Punto donde al adicionar un cluster no mejora significativamente la varianza

Aplicaciones

Conocer de qué están hablando los clientes para saber qué nuevo producto/servicio ofrecer/retirar.

Conocer el grado de satisfacción/insatisfacción del cliente cuando no tengo una escala de Likert.

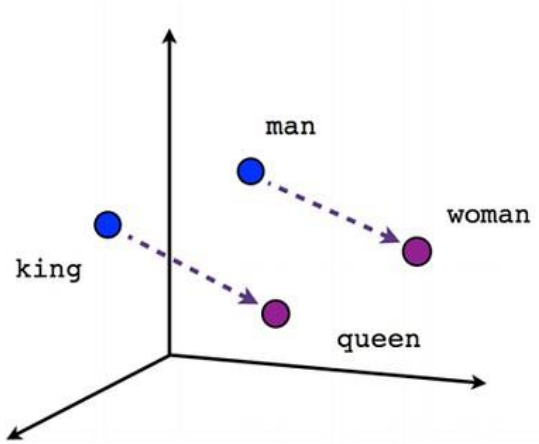
Caso N° 1 – Análisis sobre comida saludable

Partimos de un dataset donde hay texto.

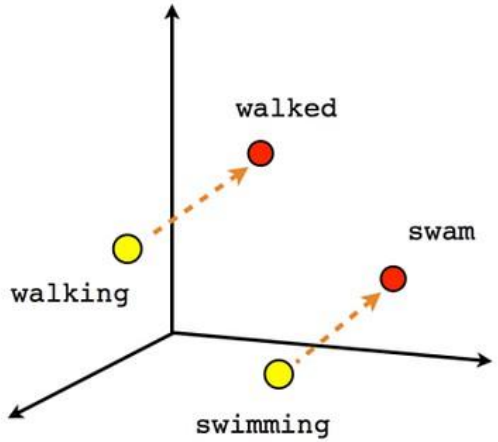
Fecha	Usuario	Comentario	is_retweet
28/11/2023 13:45:40	Usuario_1	Las legumbres son un ingrediente esencial de las dietas saludables.	True
28/11/2023 13:50:00	Usuario_2	Las legumbres son una gran fuente de proteína vegetal, con baja huella de carbono e hídrica y además:\n\n🚫 🌾 No tienen gluten, son ideales para …	True
28/11/2023 14:20:00	Usuario_3	¿#SabíasQue las legumbres tienen hierro y proteína vegetal, lo que te ayuda a estar más fuerte?	True
28/11/2023 17:12:10	Usuario_4	Nooooo de hecho también hable de la proteína vegetal !!!! Creo que todos debemos migrar hacia allá ! Pero en medios tenemos la responsabilidad de hablar a todos!!!	True
28/11/2023 21:01:20	Usuario_5	Por supuesto que conozco a @Tierradanimales desde sus inicios, y nos seguimos mutuamente. Incluso participe en varias pláticas de veganismo …	True
28/11/2023 23:59:50	Usuario_6	No, con el vegetarianismo imposible ya que sigue siendo una forma de explotación animal. Con el veganismo sí.	True

No existe una etiqueta que nos indique de qué están hablando los clientes.

Caso N° 1 – Análisis sobre comida saludable



Male-Female



Verb tense

El lenguaje humano es traducido a un lenguaje que pueda ser entendido por la computadora.

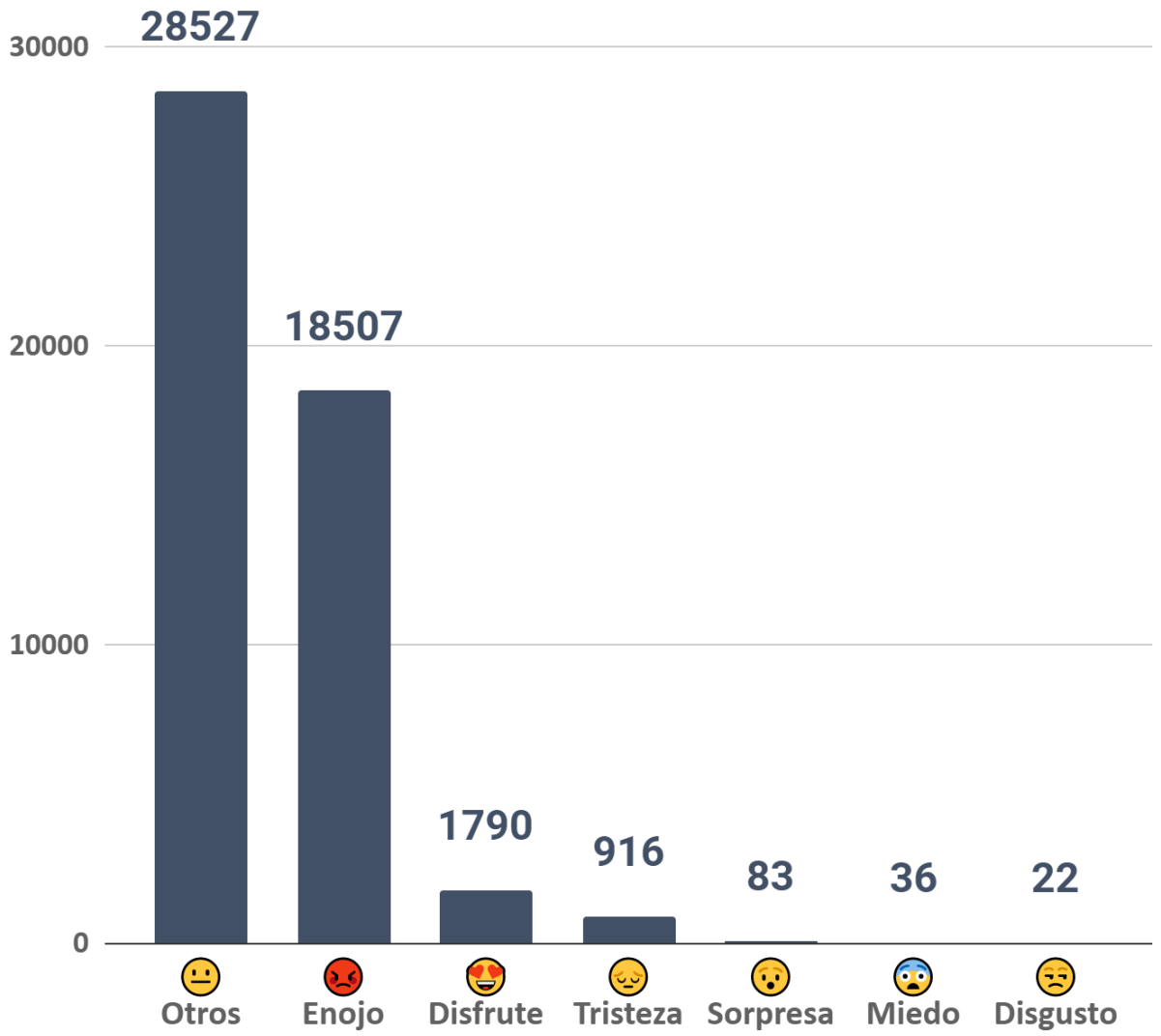
Nos apoyamos de Word embedding.

Word embedding es el texto representado de forma numérica (vector de múltiples dimensiones).

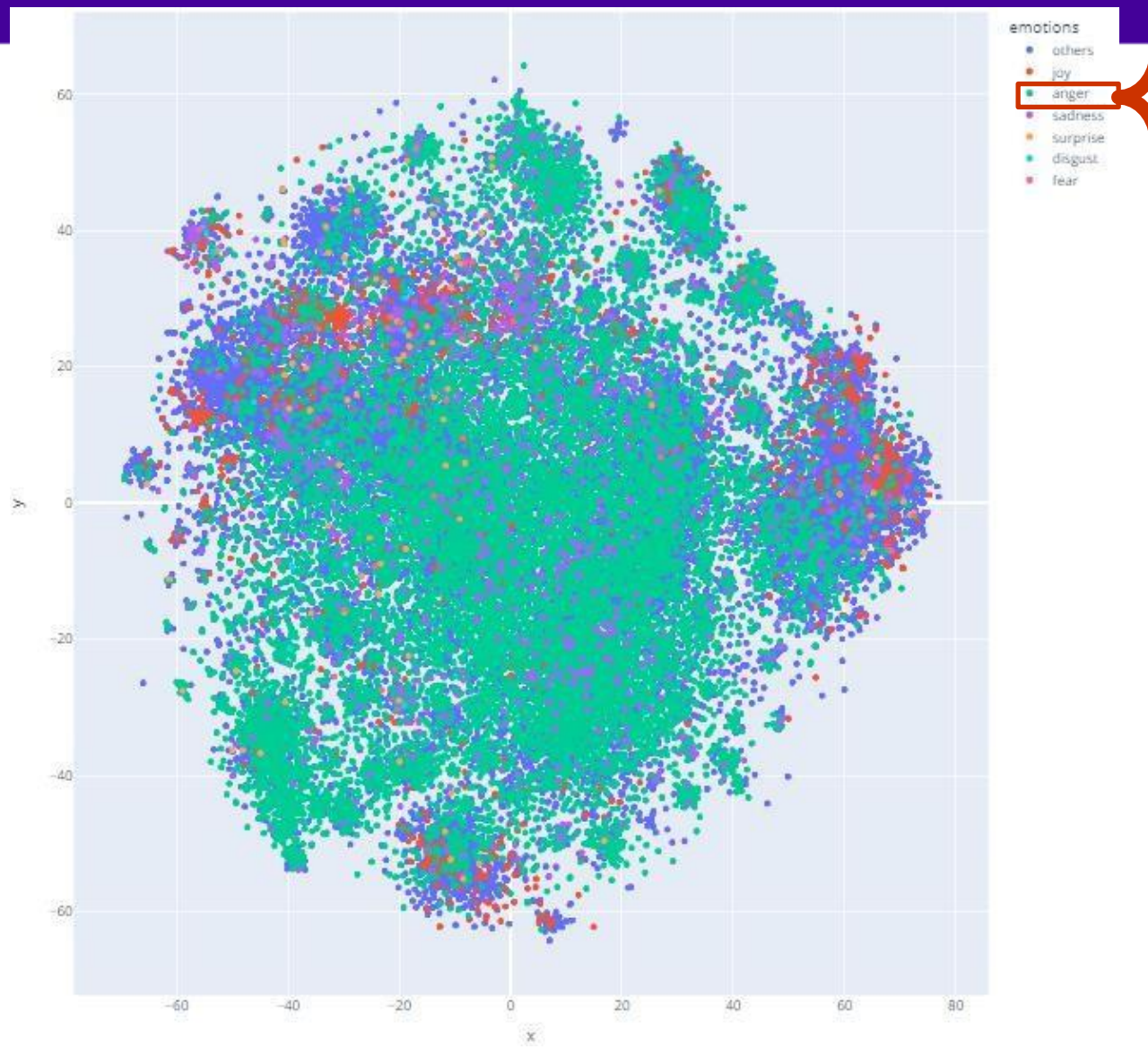
Rey - Hombre + Mujer = Reyna

Caminando - Caminó + Nadó = Nadando

Caso N° 1 – Análisis sobre comida saludable

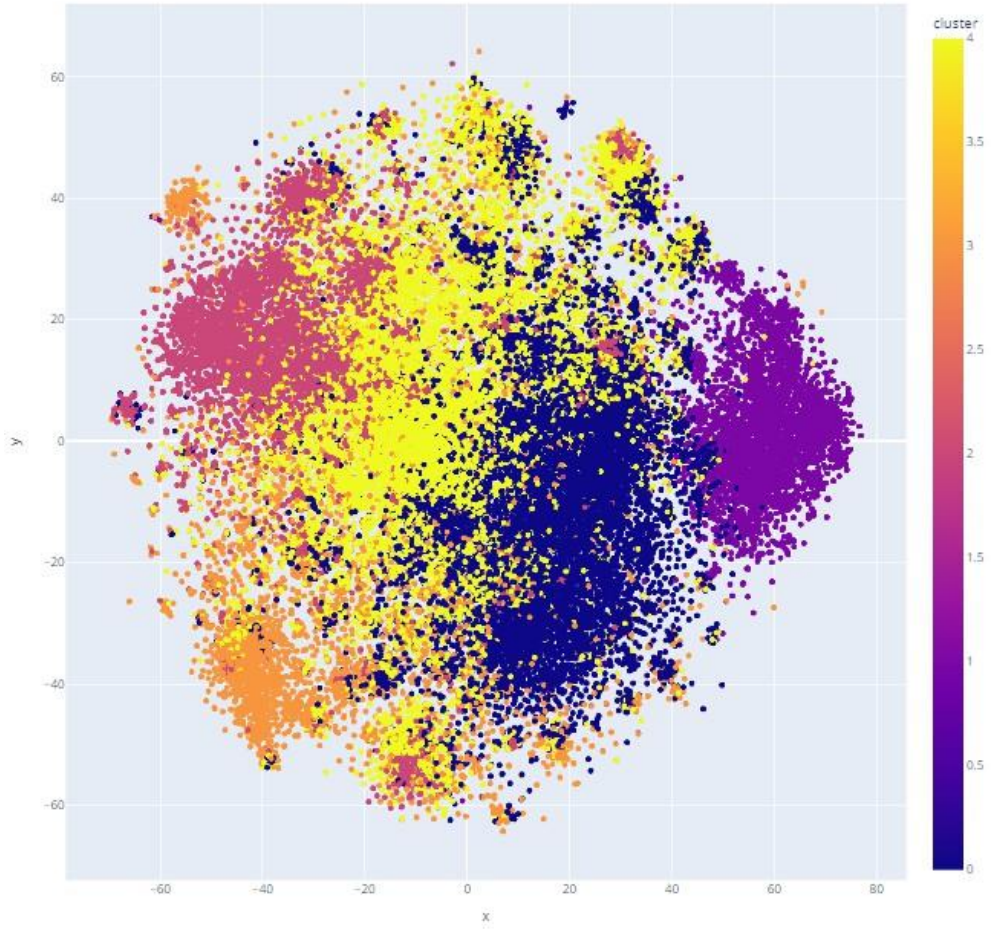


Caso N° 1 – Análisis sobre comida saludable



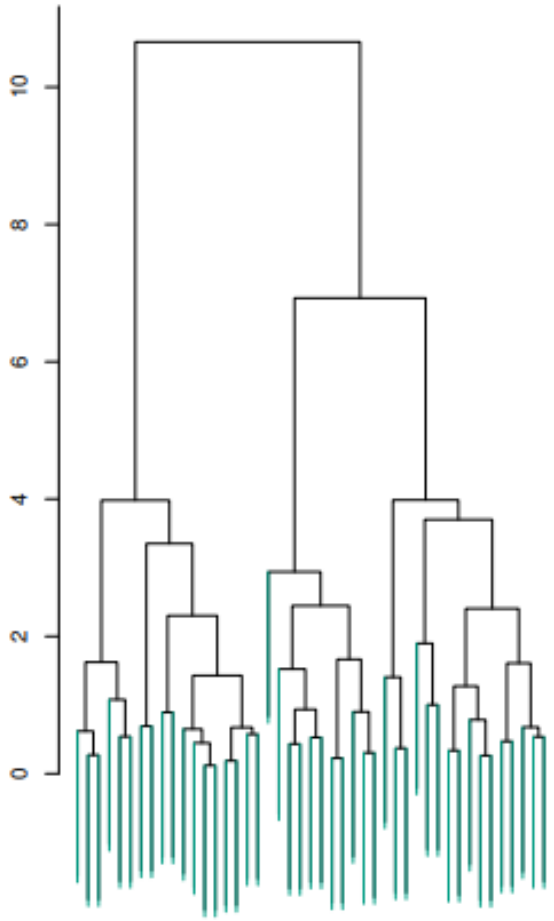
Caso N° 1 – Análisis sobre comida saludable

Como resultado se identificaron 5 grandes grupos:



- Consideran el veganismo como una postura ética.
- Comentan sobre lo deliciosas y saludables que son las hamburguesas vegetarianas y veganas.
- Comentarios sarcásticos sobre el veganismo.
- Personas totalmente en contra del veganismo.
- Consideran que hay personas que son veganas sólo por moda.

Clustering Jerárquico



No conocemos de antemano cuantos clusters queremos.

Obtenemos una representación basada en árboles llamada dendograma que nos permite visualizar los grupos obtenidos para cada posible número de clusters de 1 a N.

Clustering Jerárquico

Agglomerative

Es un método "bottom-up" (de abajo hacia arriba) cada observación empieza en un cluster y los pares de clusters se combinan cuando se avanza hacia arriba en la jerarquía.

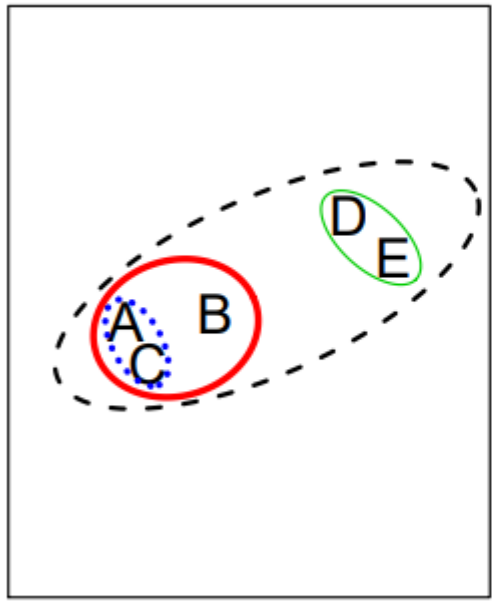
Inicia con tantos cluster como observaciones existan.

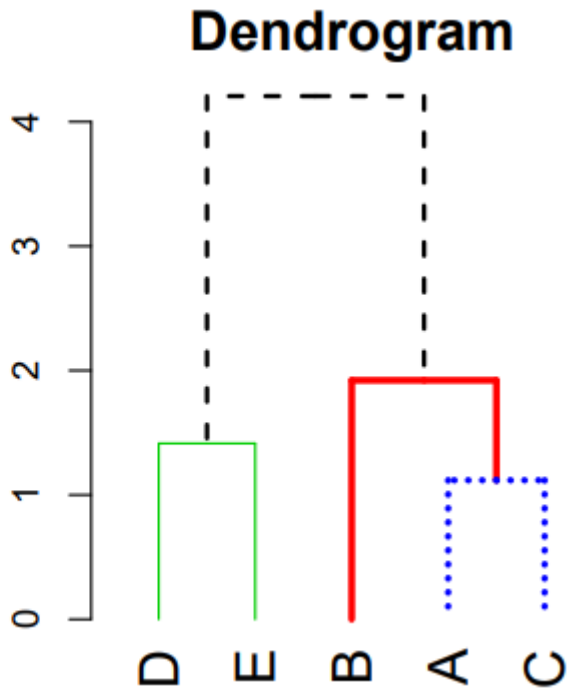
Divisive

Es un método "top-down" (de arriba hacia abajo) en donde todas las observaciones empiezan en un cluster y se van haciendo divisiones hacia abajo.

Inicia con un único cluster que agrupa a todas las observaciones.

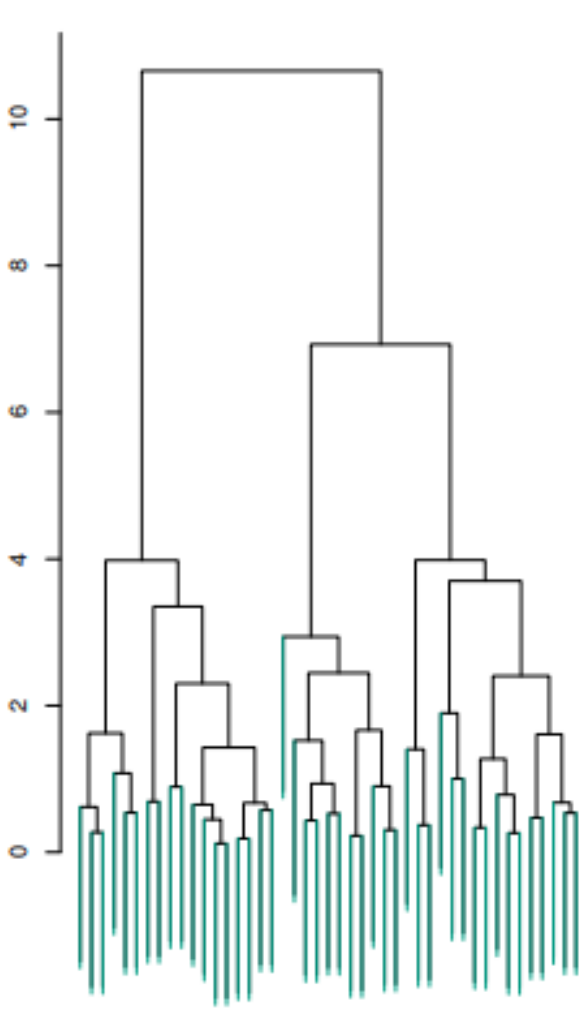
Clustering Jerárquico

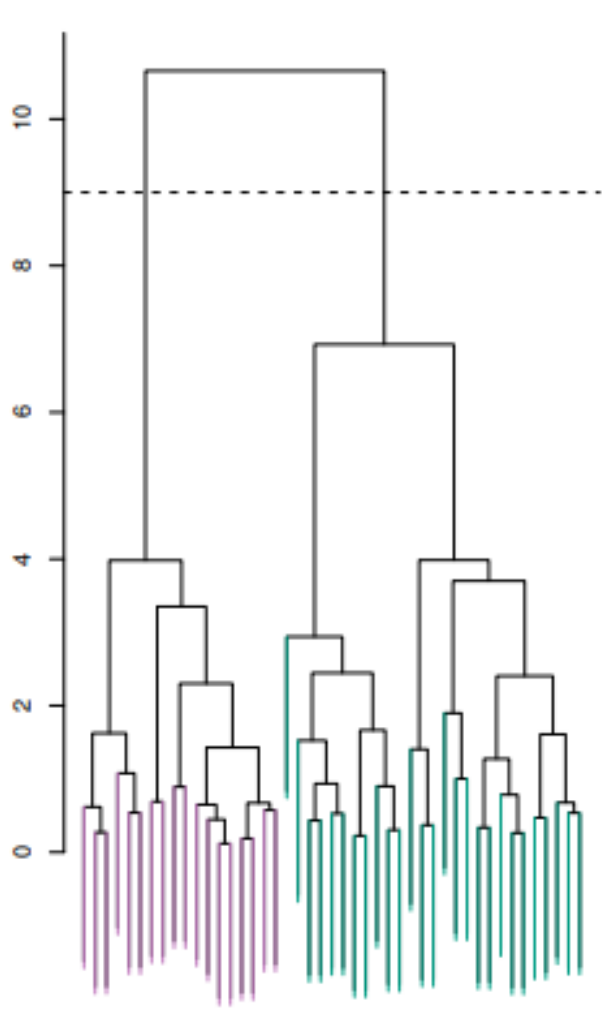


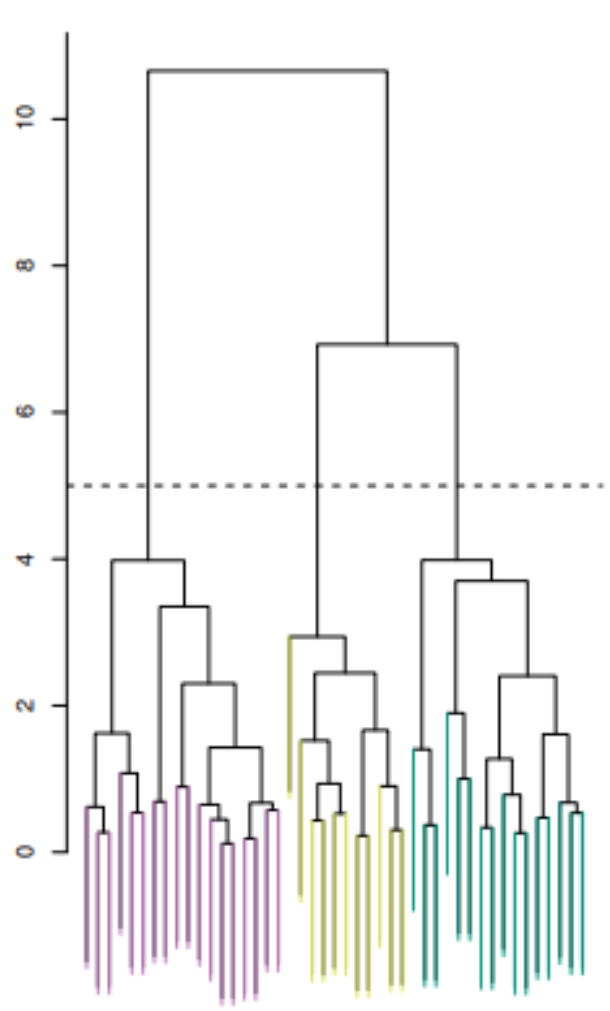


1. Iniciar con cada observación de su propio cluster.
2. Identificar los dos puntos más cercanos y los relaciona.
3. Repetir el procedimiento.
4. Terminar cuando todas las observaciones están en un único cluster.

Clustering Jerárquico







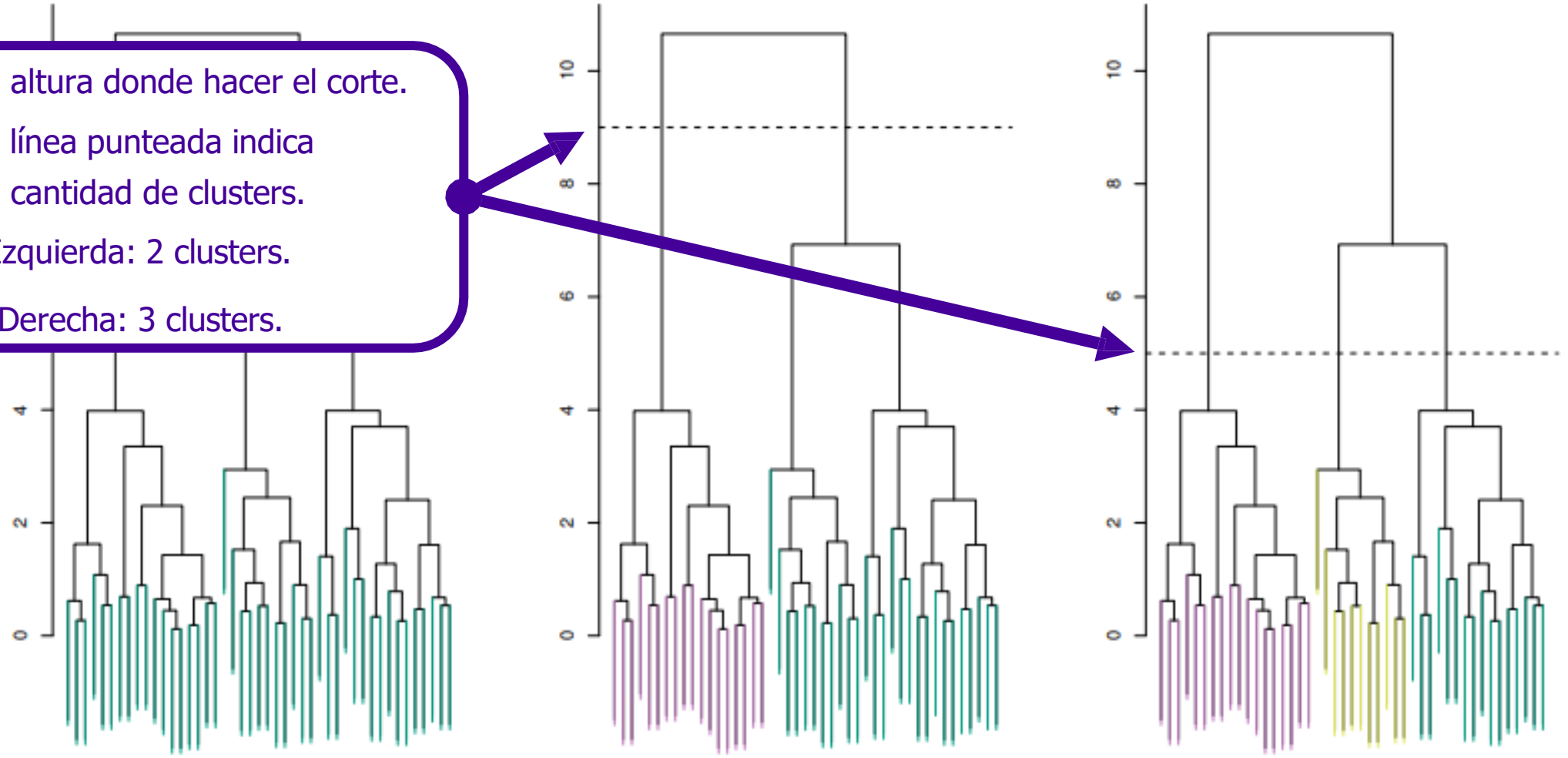
Clustering Jerárquico

Defino la altura donde hacer el corte.

La línea punteada indica la cantidad de clusters.

Izquierda: 2 clusters.

Derecha: 3 clusters.



Pausa de 5 minutos

Antes de finalizar
¿hay alguna pregunta respecto a la clase?

¿Qué he aprendido el día de hoy en clase?

TOULOUSE
LAUTREC