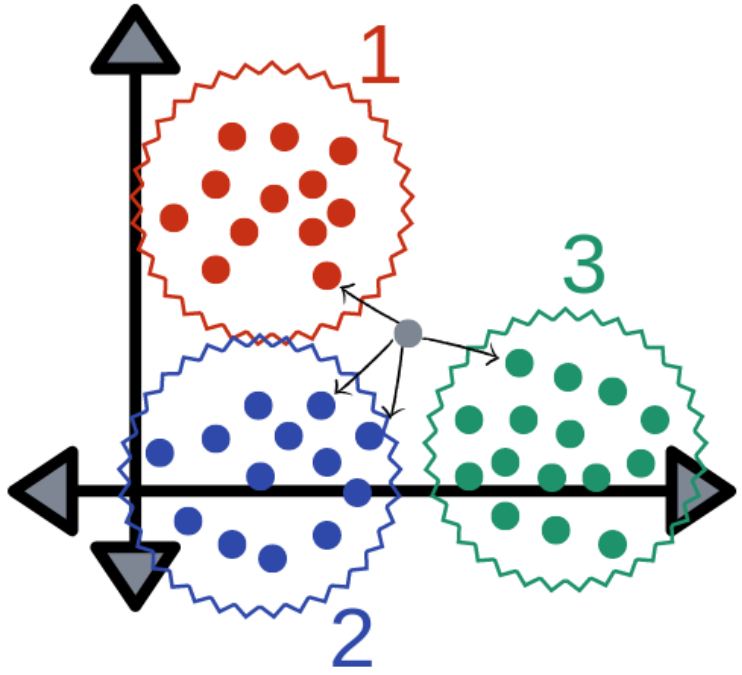
KNN: K vecinos más cercanos



K –Nearest Neighbors

> "En contraste con otros algoritmos de aprendizaje supervisado, **K-NN** no genera un modelo fruto del aprendizaje con datos de entrenamiento, sino que el aprendizaje sucede en el mismo momento en el que se prueban los datos de test."

Características

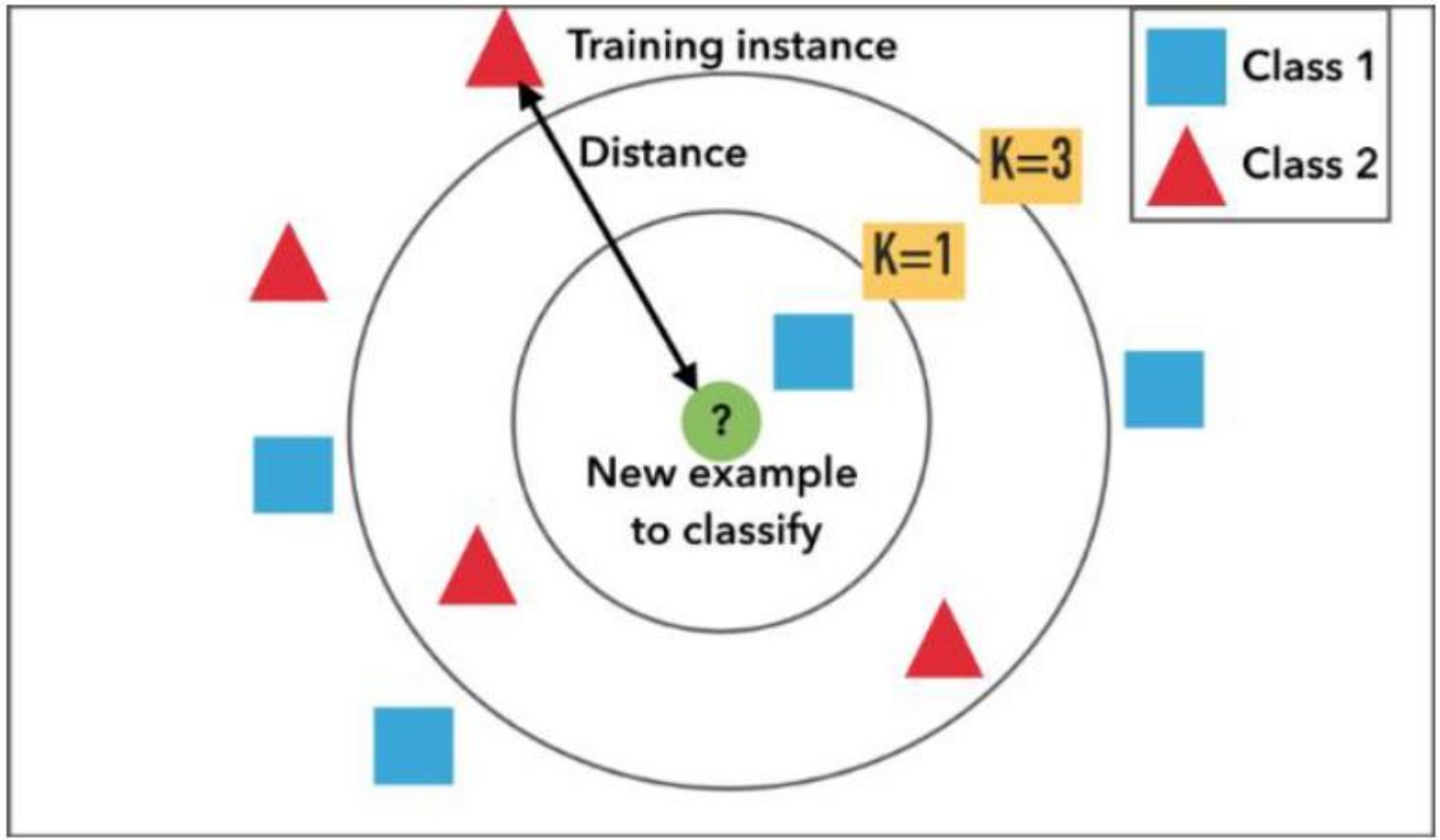
- ✓ Non – parametric: Significa que no hace ninguna suposición sobre la distribución de los datos.
- ✓ Lazy learning: No utiliza la data de entrenamiento para hacer alguna suposición.

Recomendación

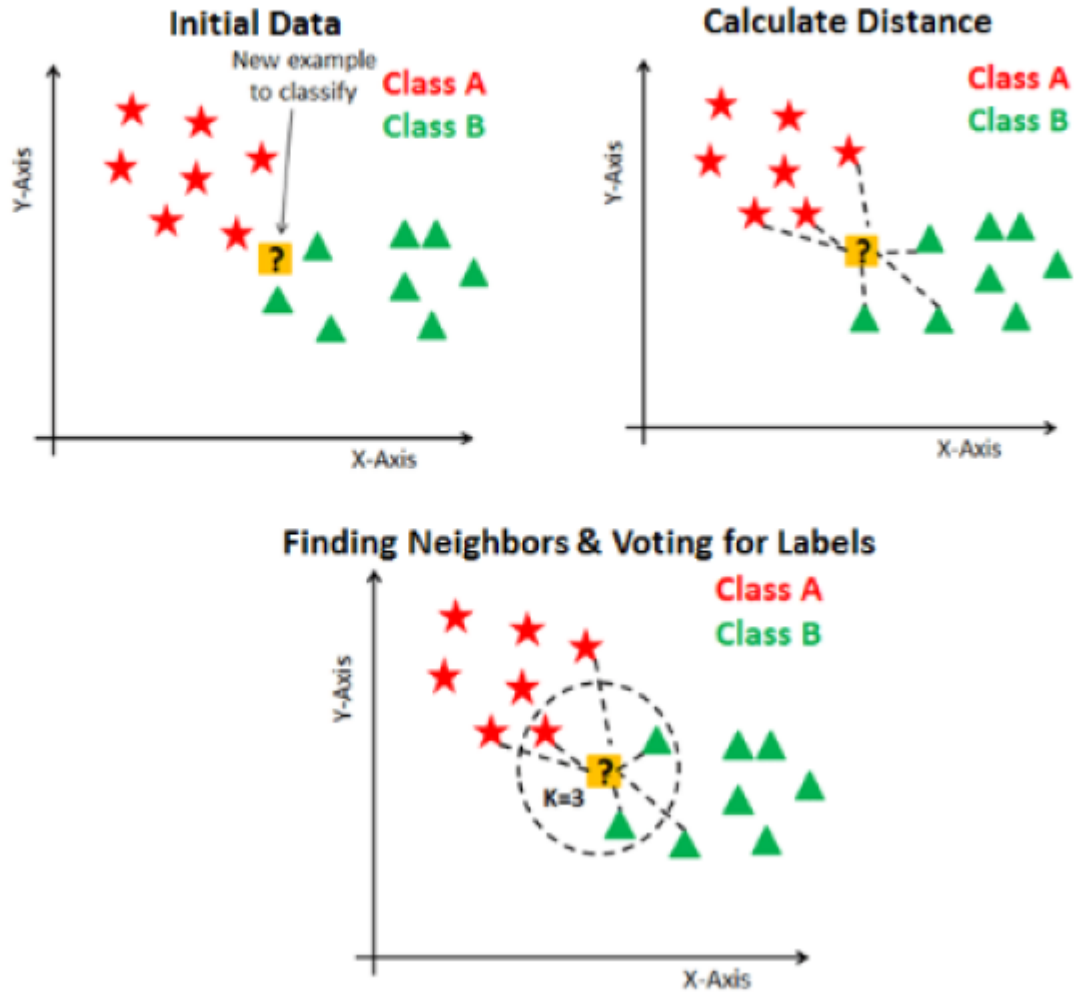
KNN puede y debe ser una de las primeras opciones a usar cuando se conoce poco o nada de la distribución de los datos.

K –Nearest Neighbors

"KNN es basado en calcular la similitud entre los elementos"



K –Nearest Neighbors



K –Nearest Neighbors

Pros

- ✓ No hipótesis acerca de la data.
- ✓ Algoritmo simple.
- ✓ Buena exactidud.
- ✓ Útil para clasificación y regresión.

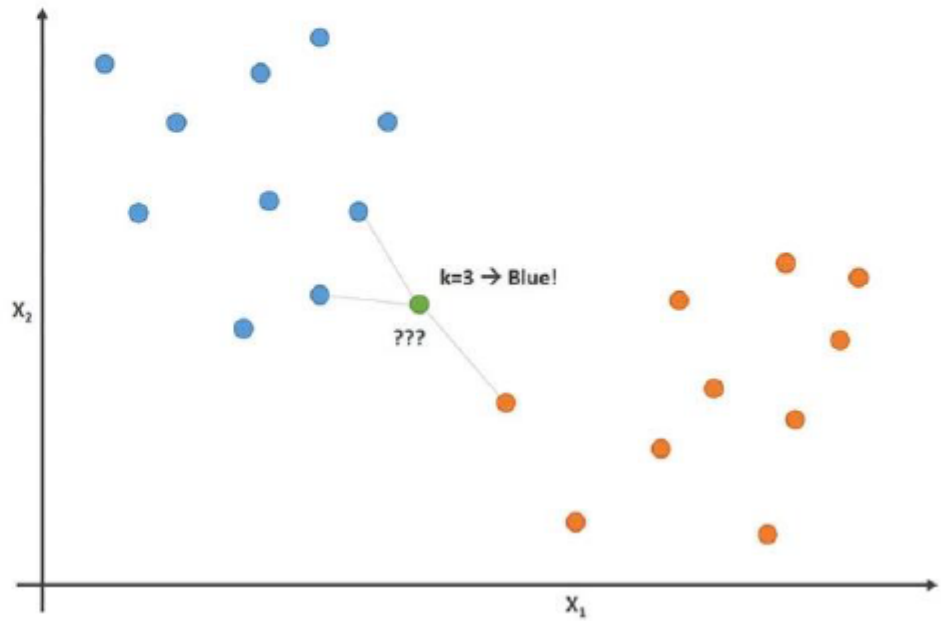
Contras

- ○ Computacionalmente caro.
- ○ Requerimiento de mucha memoria.
- ○ Almacenar toda o casi toda la data.
- ○ Etapas de predicción puede ser lenta.
- ○ Sensible a irrelevantes variables y la escala de la data.

K –Nearest Neighbors

Pseudo - código

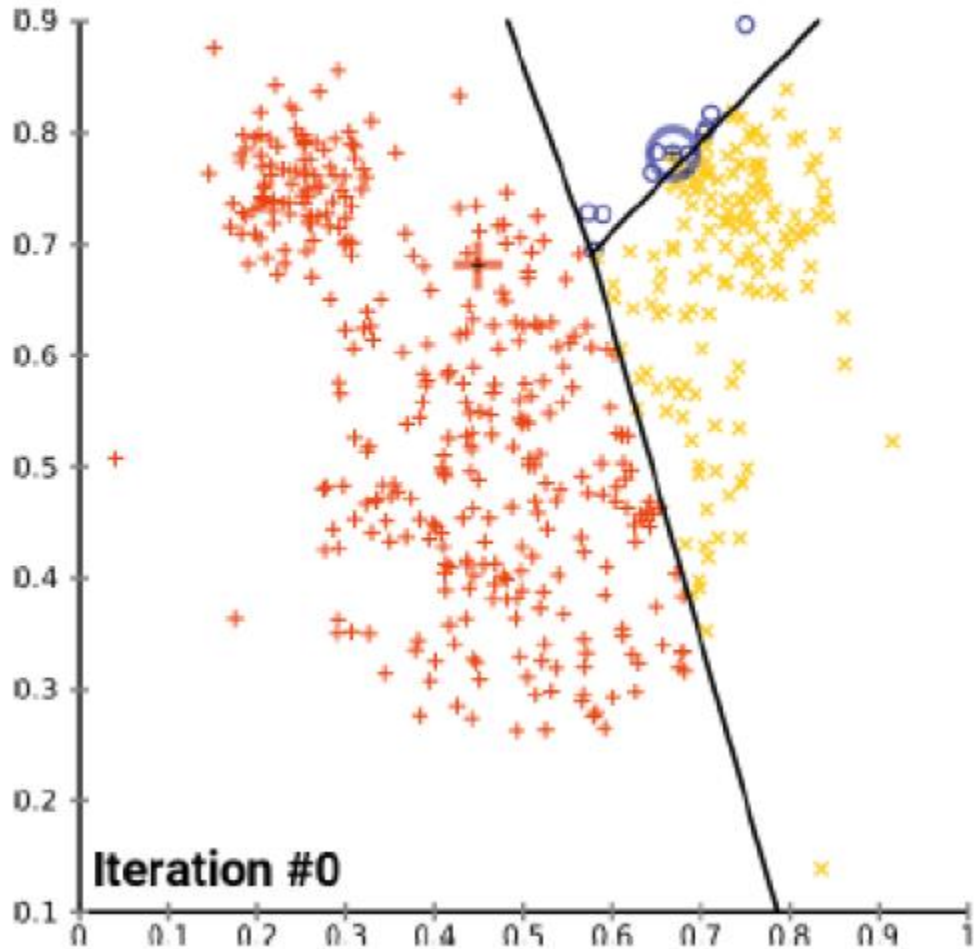
- Un número entero es definido (K).
- Seleccionamos los K elementos más parecido al elemento a clasificar.
- Encontramos la clase más común de los elementos.
- Asignamos la clase más común al elemento nuevo.



Métodos No Supervisados

Ing. Agustín Ullón Ramírez

Aprendizaje No Supervisado



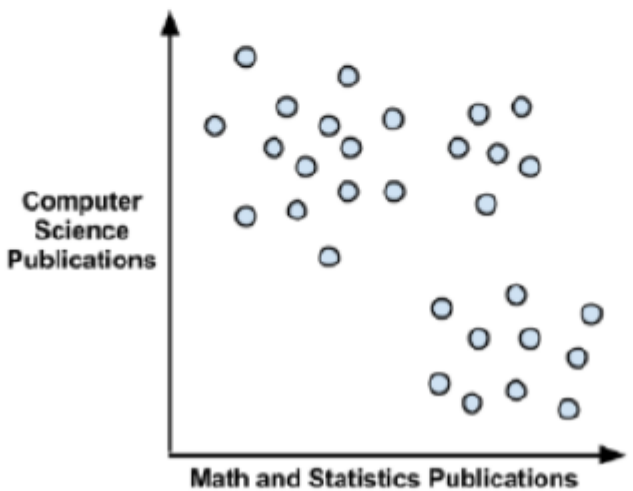
Aprendizaje No Supervisado

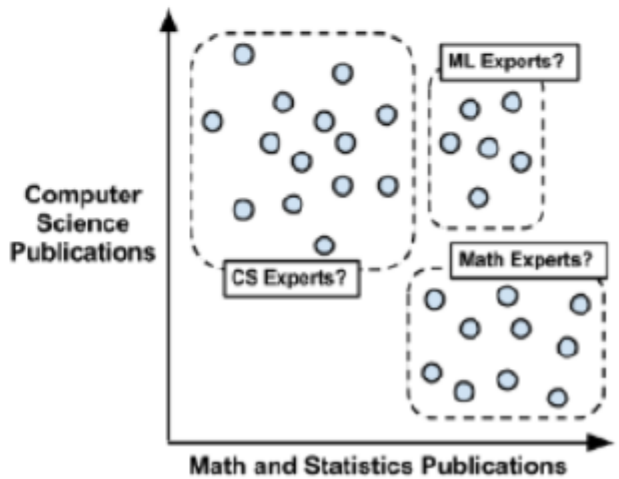
Los algoritmos de **aprendizaje no supervisado** son aquellos que aprenden de un dataset sin etiquetas, es decir, que no hay ejemplos con los que comparar durante el aprendizaje.

El objetivo del aprendizaje no supervisado consiste en extraer patrones o inferir propiedades de la distribución de los datos en el espacio de *features*, en función de algún criterio estadístico.

Se utilizan comúnmente para:

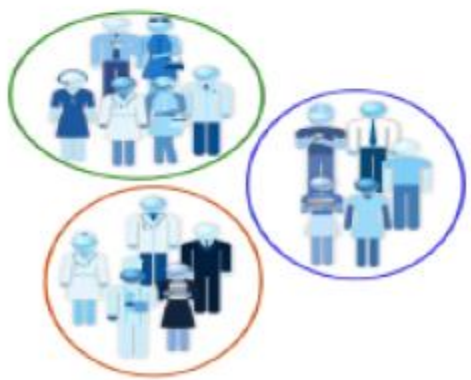
- Segmentación.
- Reducción de dimensionalidad.
- Detection de anomalias.
- Otros.





Aprendizaje No Supervisado

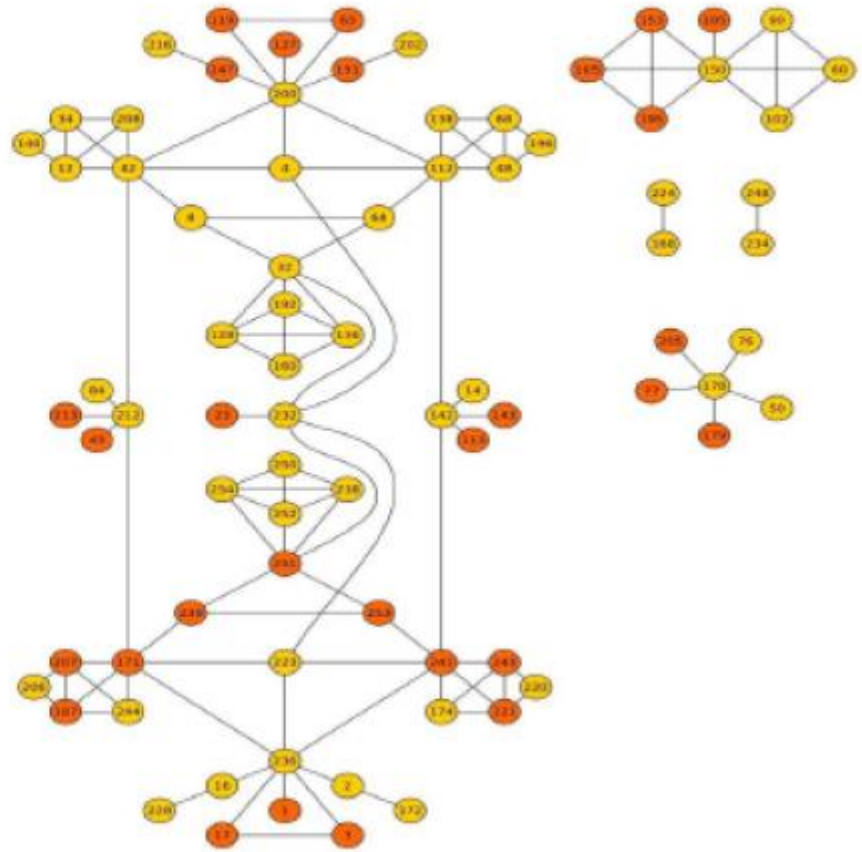
- En un problema de aprendizaje no supervisado, el objetivo principal es descubrir la estructura subyacente de los datos en función de ciertas características o variables.
- En el aprendizaje no supervisado, los datos de entrenamiento **no están etiquetados**, es decir, no existe variable objetivo.
- Los modelos de aprendizaje supervisado se pueden emplear para realizar análisis exploratorio de los datos (EDA).
- El tipo de problema más común dentro del marco de aprendizaje no supervisado es el **clustering**. Consiste la agrupación de los datos de entrada en función de ciertos criterios.



Clustering

- También conocido como segmentación, se refiere a dividir un conjunto de datos en grupos o clusters (no definidos previamente) de tal forma que cada grupo contenga muestras similares entre sí, basándose en alguna medida de similitud entre los miembros de cada grupo.

- En general la estrategia consiste en dividir los datos de tal forma que los miembros de los grupos sean **"muy similares"** entre sí, y **"muy diferentes"** entre miembros de otros grupos.



Clustering

Por ejemplo:

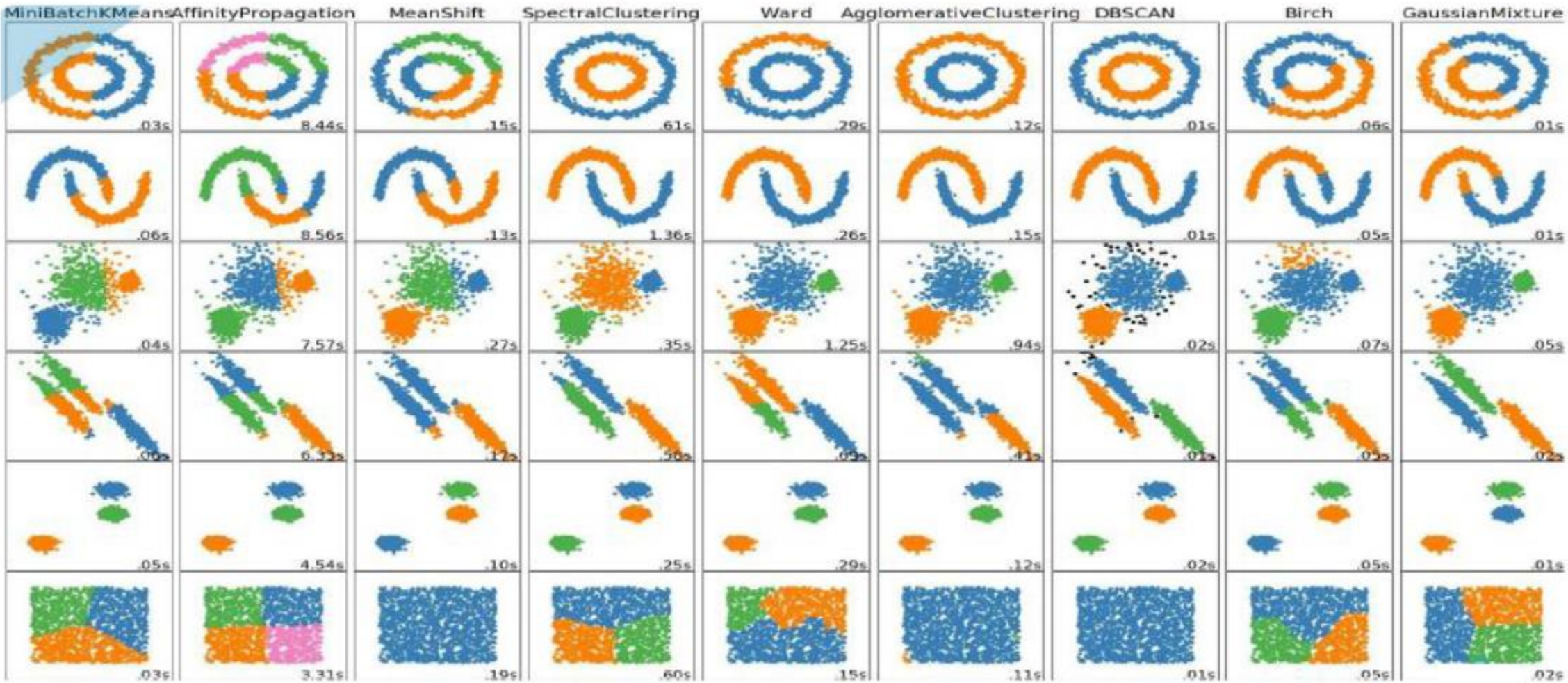
Supongamos que tenemos un dataset con datos de nuestros clientes. Los datos pueden ser un grupo de variables socio-demográficas (sexo, edad, grupo socio-económico, nivel educativo, etc) y un grupo de variables de consumo (compras realizadas por cada cliente, cantidad de dinero gastado en nuestros productos, cantidad de veces que visitan nuestra página web, tipos de websites que visitan, etc).

Aplicando clustering podríamos construir grupos de clientes, donde en cada grupo queden los clientes que se parecen socio-demográficamente y que tienen patrones de consumo similar.

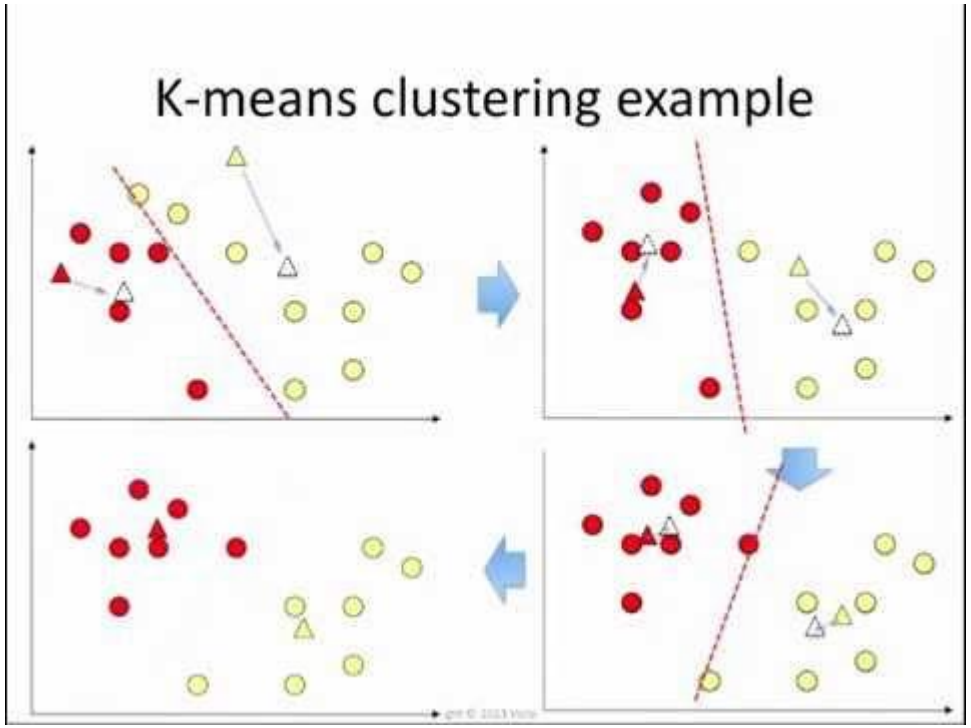
Otras aplicaciones comunes de las técnicas de clustering son:

- Categorizar productos similares en función del uso que les dan los clientes
- Análisis de redes sociales. Por ejemplo: agrupar seguidores en twitter en *lovers* o *haters* en función de las palabras que usan cuando mencionan a una marca

Clustering


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.03s
8.44s
.15s
.61s
.29s
.12s
.01s
.06s
.01s
.06s
8.56s
.13s
1.36s
.26s
.15s
.01s
.05s
.01s
.04s
7.57s
.27s
.35s
1.25s
.94s
.02s
.07s
.05s
.05s
.12s
.50s
.41s
.01s
.05s
.02s
.05s
4.54s
.10s
.25s
.29s
.12s
.02s
.05s
.01s
.03s
3.31s
.19s
.60s
.15s
.11s
.01s
.05s
.02s

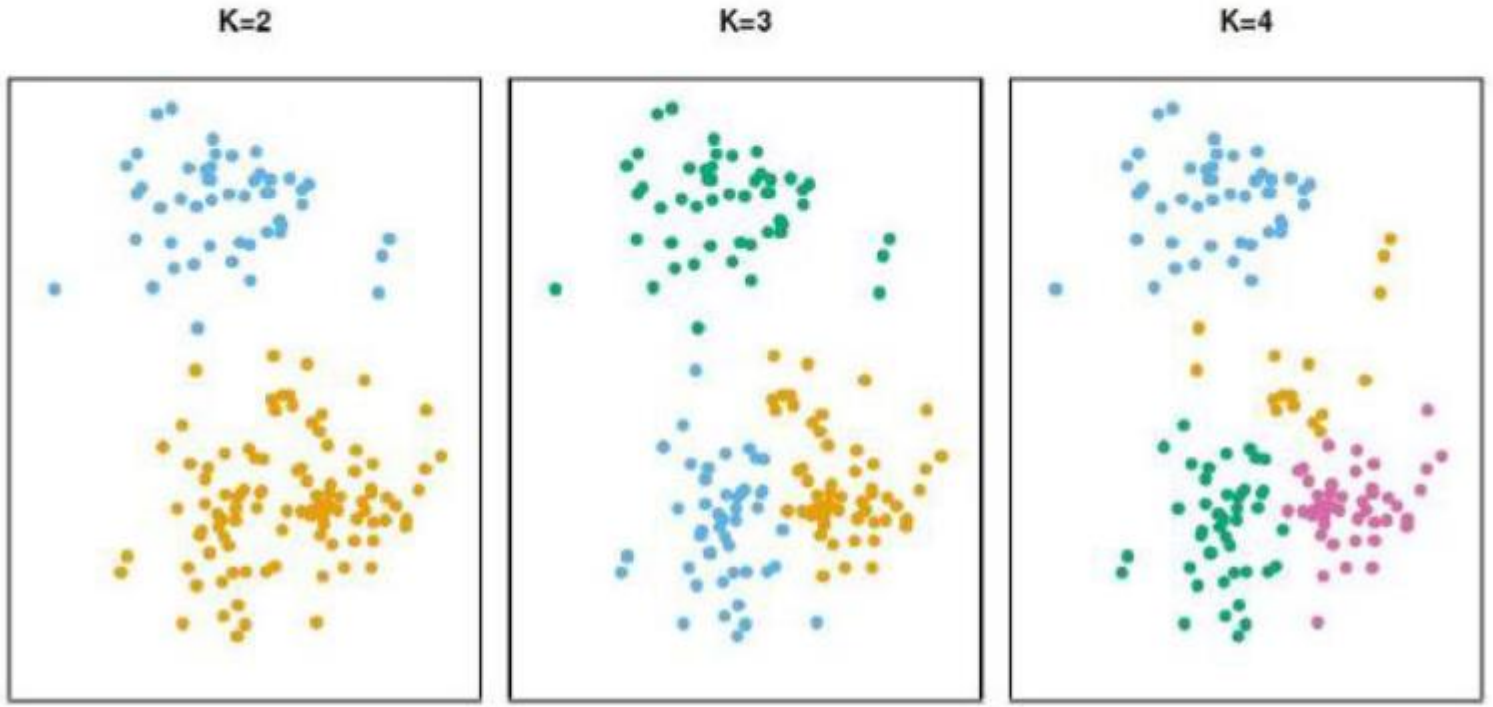
K means



K-means

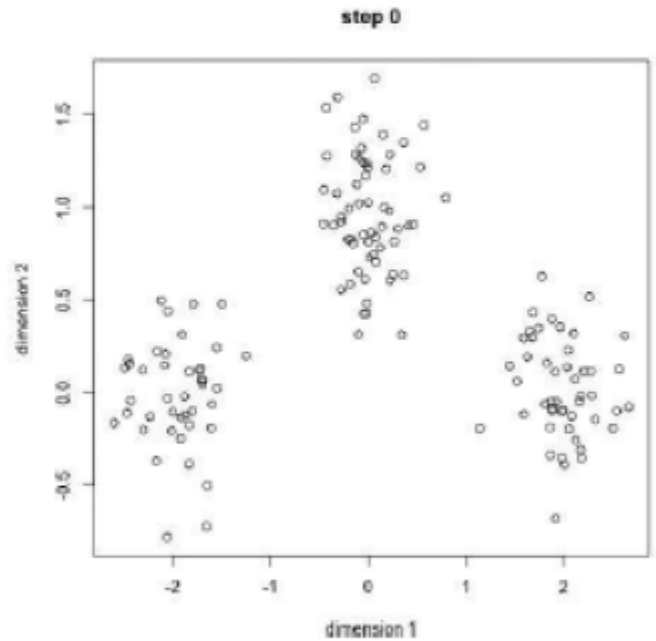
- En K-means el objetivo es agrupar las observaciones de un dataset en un número K de clusters.
- El número K es hiperparámetro que hay que darle al algoritmo.

- El siguiente gráfico muestra un ejemplo de realizar K-means en un dataset con dos predictores para distintos valores de K



K-means

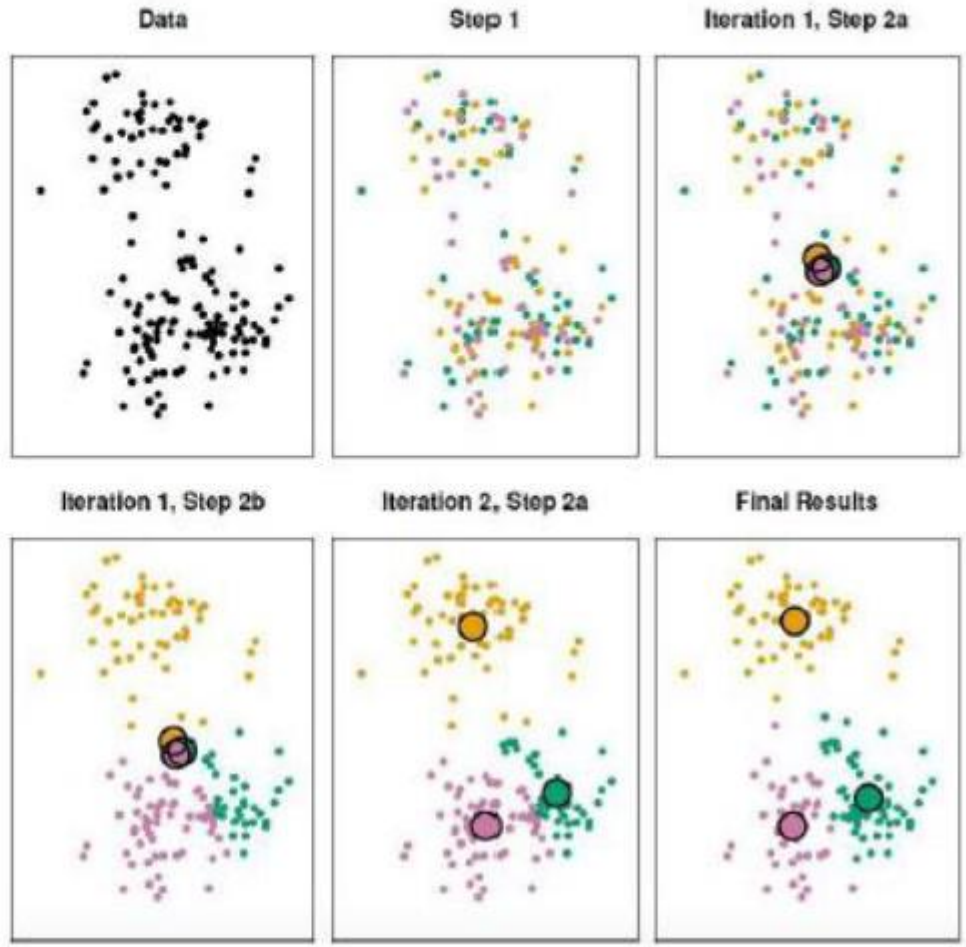
1. Seleccionar el número de clusters **K** (hiperparámetro)
2. Generación aleatoria de **K** centroides
3. Para cada iteración, o hasta convergencia:
   a. Calcular la distancia de cada muestra a cada centroide
   b. Asignar cada muestra al centroide más cercano
   c. Recalcular los **centroides**, como el vector de medias de los *D* predictores de las observaciones de cada clúster.
   d. Chequear convergencia



Al final, cada registro estará asignado a un centroide concreto (el más cercano), formando así los diferentes clústers.

K-means

El siguiente gráfico ilustra el algoritmo de K-means:



K-means

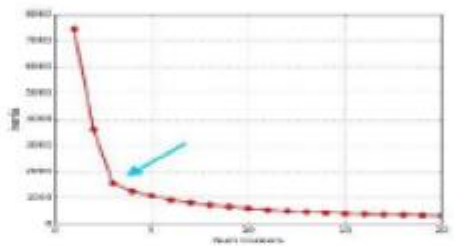

Union Square
San Francisco
Painted Ladies
SOUTH OF MARKET
WESTERN ADDITION
JAPANTOWN
Fillmore

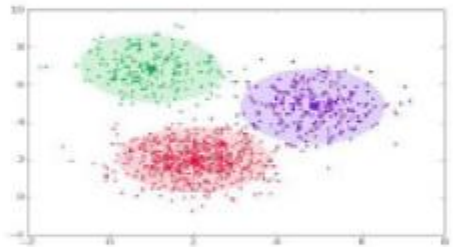
K-means

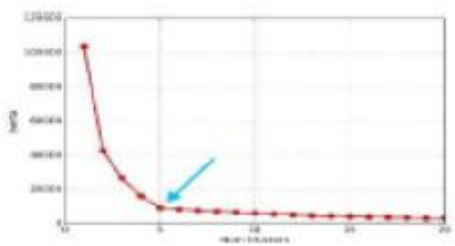
¿Cómo seleccionamos el número de clusters k?

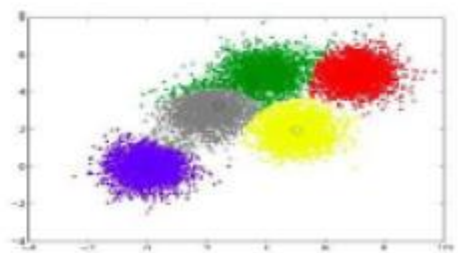
El "elbow method" consiste en medir la suma promedio de los cuadrados de la distancia de las observaciones a los centroides. El objetivo es minimizar esta suma total, tratando de escoger el mínimo número de clusters donde esta suma se minimiza:

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$





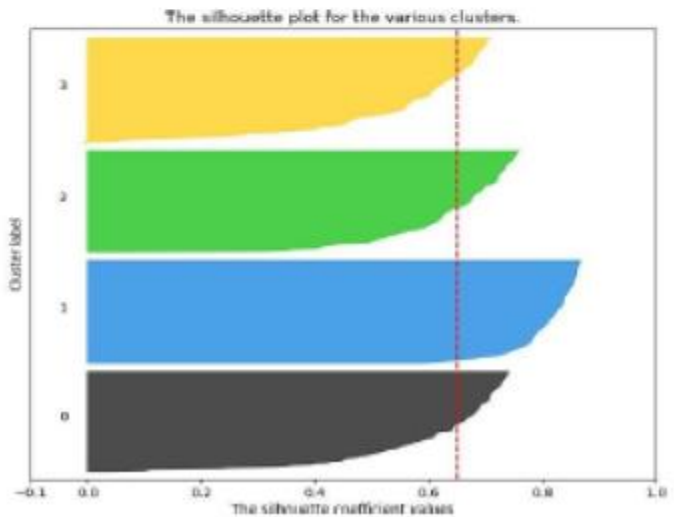


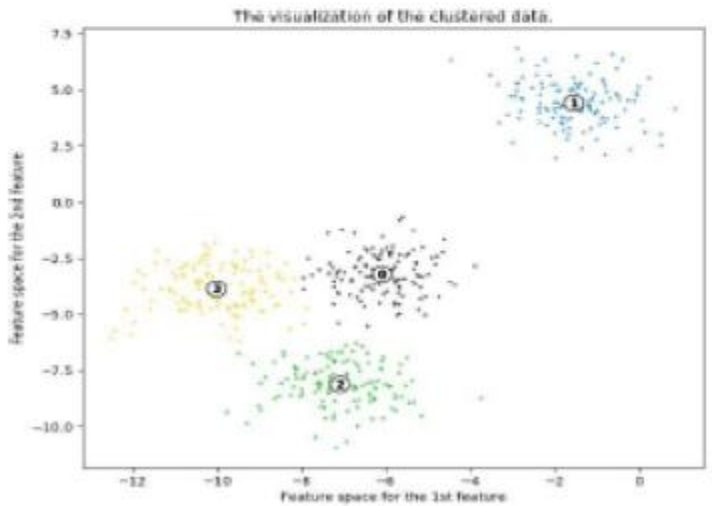


K-means

¿Cómo seleccionamos el número de clusters k?

Método de la silueta: la silueta es una medida de ajuste de cada muestra en su cluster. Se basa en la distancia intra cluster y la distancia inter-cluster. Cuando la silueta es igual a 1, la observación está en el cluster correcto, y cuando es -1 implica que está en el cluster incorrecto.

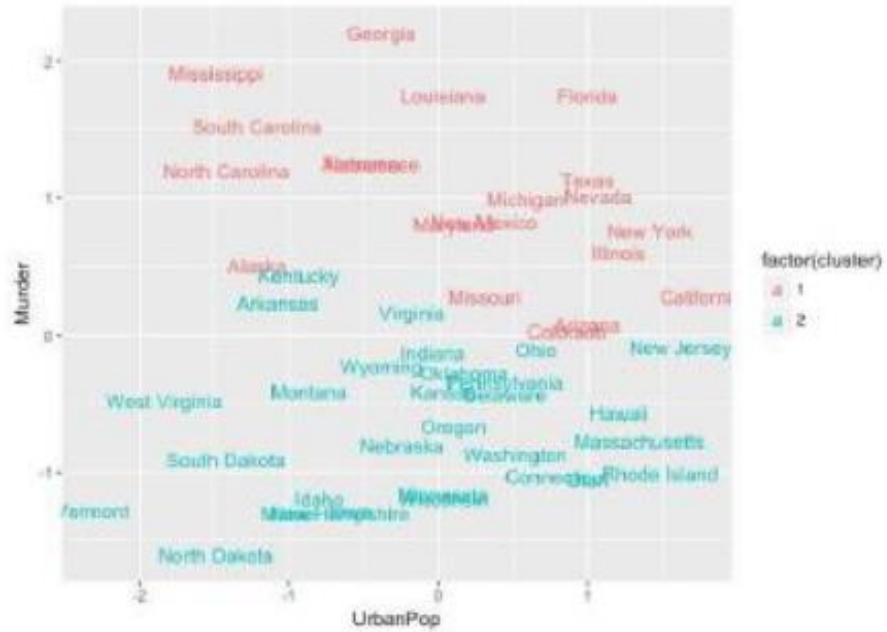




Presentación de clusters

Pair-plots

Representación gráfica de pares de features con el objetivo de observar relaciones entre variables y clusters. No es recomendable en problemas con alta dimensionalidad.

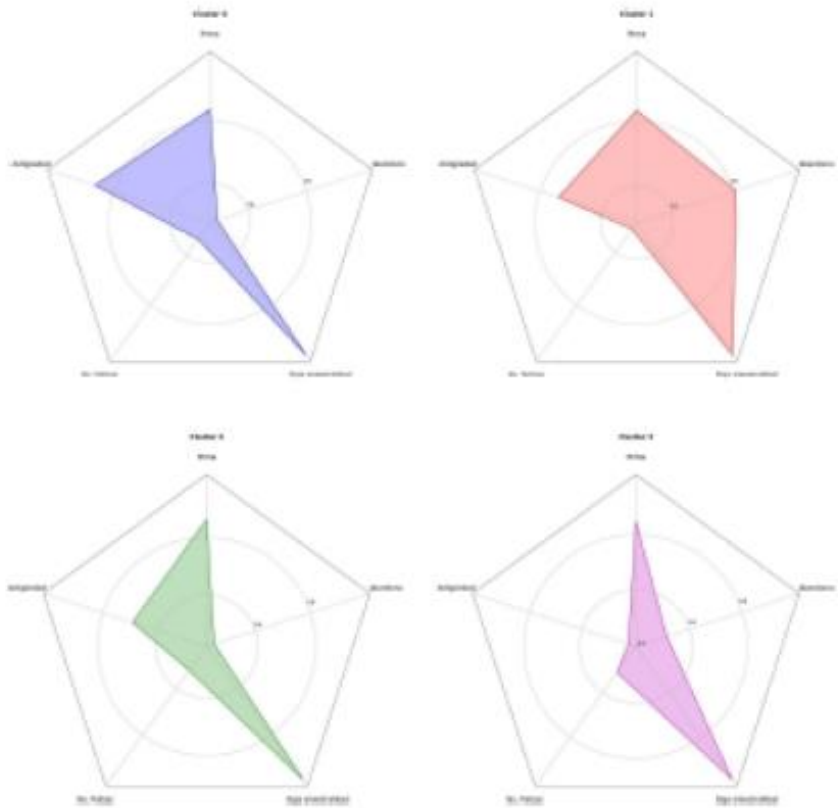


Source: https://uc-r.github.io/kmeans_clustering#optimal

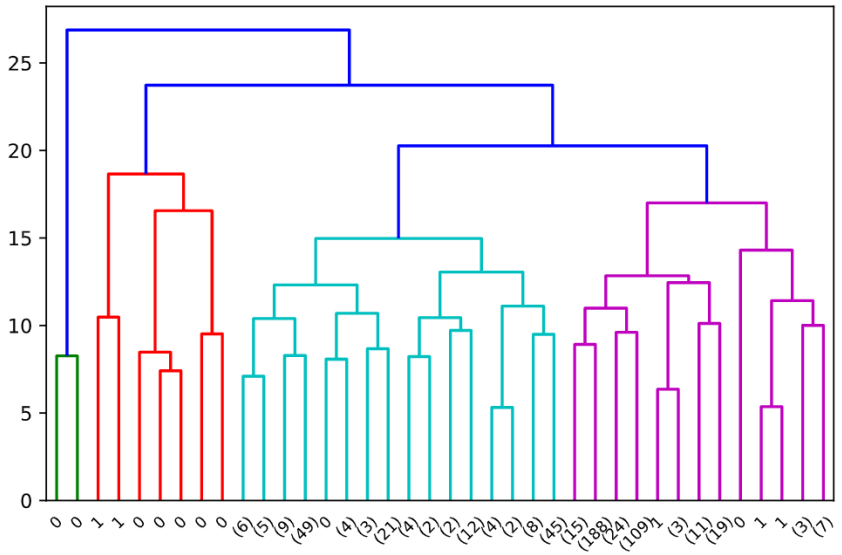
Presentación de clusters

Gráficos de araña

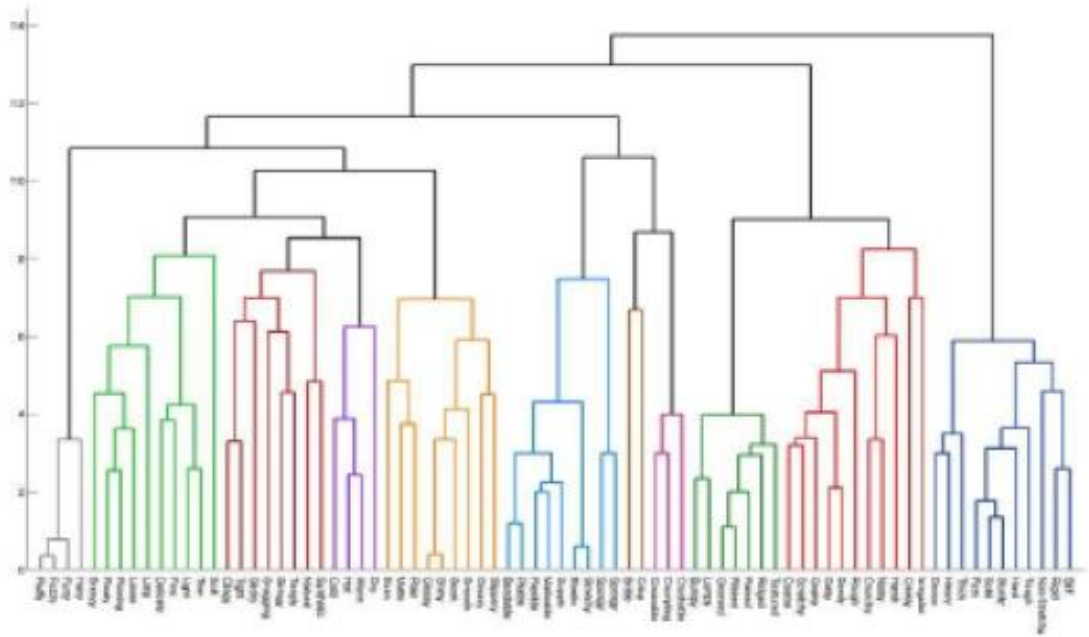
Es un método gráfico que permite representar datos multivariables en dos dimensiones. Permite perfilar y comparar clusters usando los valores medios de cada una de las variables a nivel de cluster.



Clustering Jerárquico



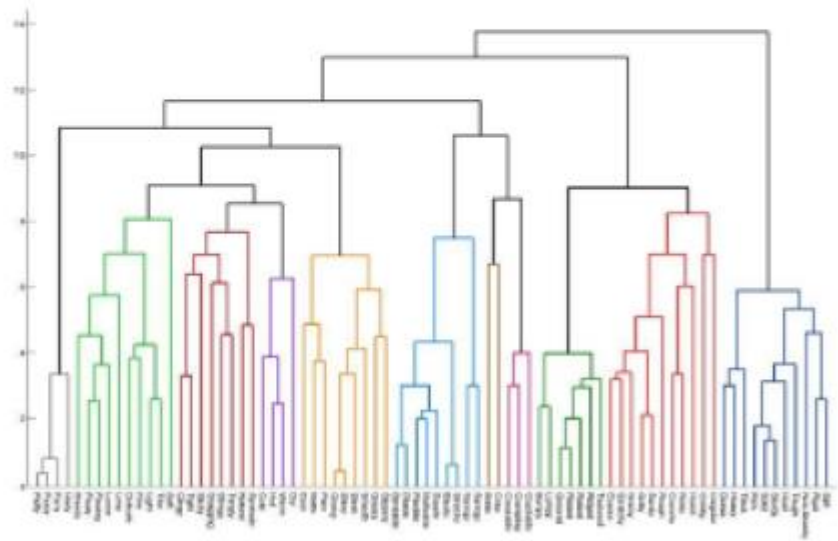
Clustering Jerárquico



Son métodos que organizan los datos (objetos) en una estructura jerárquica. De tal forma que en el extremo superior de la estructura de clustering hay un único cluster que contiene todos los datos y en el otro extremo hay un cluster por cada objeto.

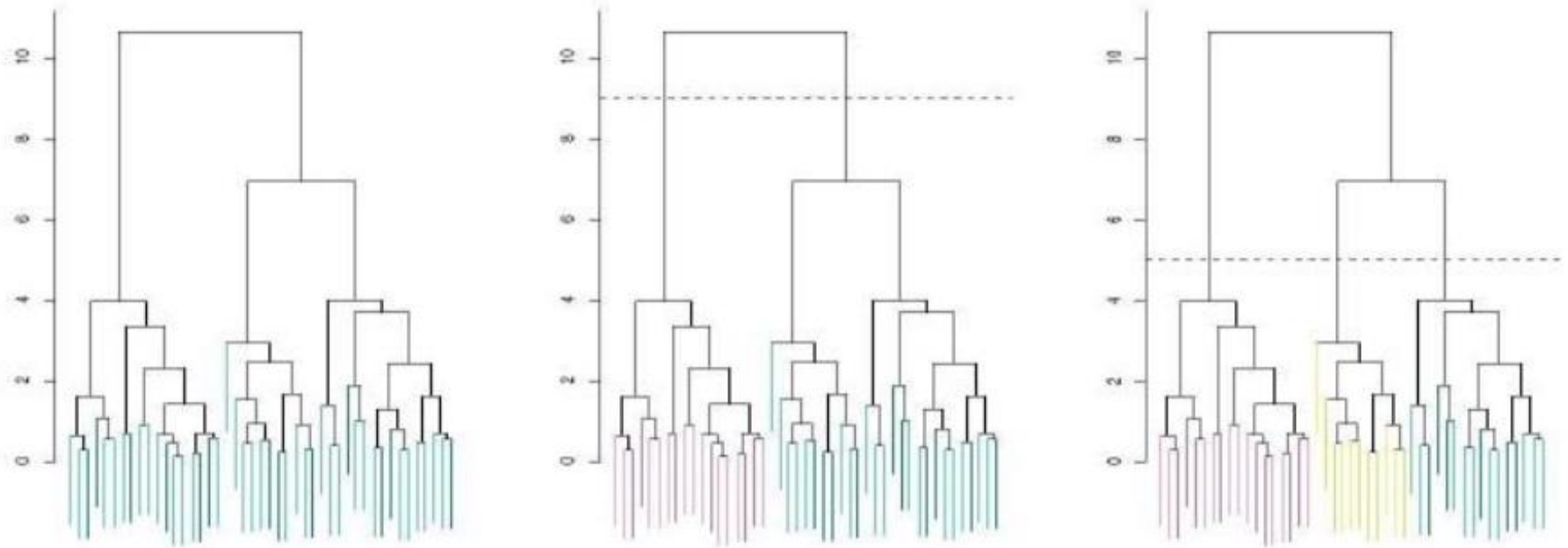
Clustering Jerárquico

- El clustering jerárquico puede construirse en base a dos estrategias:
  - **Aglomerativo**: De abajo hacia arriba. Se van construyendo los clusters agrupando muestras cercanas de forma iterativa.
  - **Divisivo**: De arriba hacia abajo. Partiendo de un único cluster, este se va dividiendo hasta conseguir grupos disímiles entre sí
- Los clusters son representados gráficamente en un **dendograma**.



Clustering Jerárquico

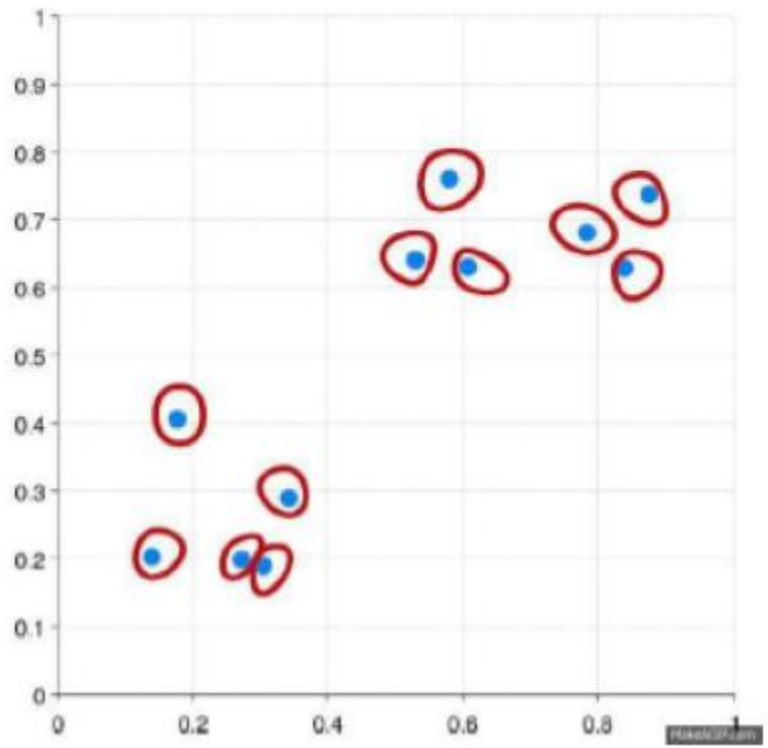
- El eje Y contiene la medida de similitud entre observaciones.
- Mientras más arriba en el eje Y, más diferentes son dos observaciones.
- La altura a la que cortemos el **dendograma** nos da el número de clusters.
- Un mismo dendograma nos sirve para probar con distintos números de clusters.



Clustering Jerárquico

Algoritmo de Clustering Jerárquico (Aglomerativo)

1. Comenzar con n observaciones y una medida de distancia entre todos los pares de observaciones (hay n(n-1)/2 posibles combinaciones, esto es la combinatoria de n en 2)
2. Para i = n, n-1, n-2…., 2:
    a. Examinar todas las dis-similitudes entre clusters e identificar las menos dis-similares. Esas dos se fusionan y forman un nuevo cluster.
    b. La dis-similitud entre estas dos observaciones o clusters se representan en el eje Y del dendograma.
    c. Calcular de nuevo todas las dis-similitudes entre todos los restantes i-1 clusters y repetir el paso 2.



Clustering Jerárquico

Medida de Dis-Similitud (linkage)

- Para determinar qué tanto se diferencian dos observaciones, se establece una medida de "dis-similitud" a la que se le llama **linkage**.
- La distancia entre dos observaciones puede ser distinta a distancia Euclidea.
- El **linkage** puede ser uno de los siguientes:
  - **Completo**: calcular todas las dis-similitudes entre las observaciones del cluster A y el cluster B, y quedarse con la máxima de estas medidas.
  - **Simple**: calcular todas las dis-similitudes entre las observaciones de los clusters A y B y quedarse con la mínima de estas medidas.
  - **Promedio**: mismo procedimiento, y calcular el promedio de todas las medidas.
  - **Centroide**: calcular la dis-similitud entre los centroides del cluster A y el B.

GRACIAS…