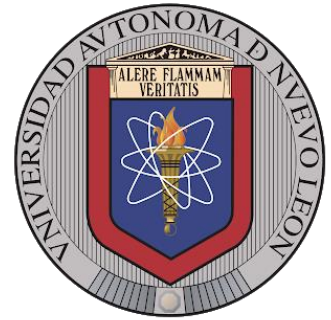


UANL

Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas

FCFM

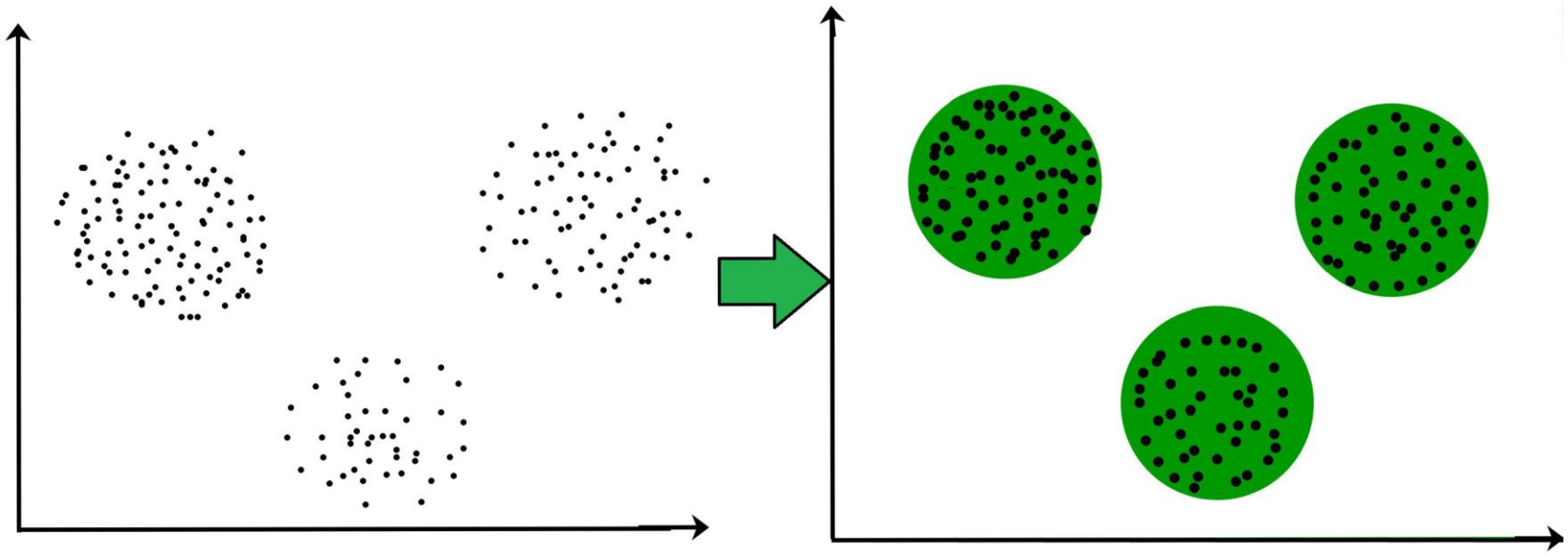
Clustering

Minería de Datos

Alan Zamarron Medrano 1625654
Carla Mayela De la Garza Fernández 1729734

Clustering

Es una técnica dentro de la disciplina de Inteligencia Artificial, identifica de manera automática agrupaciones (o clústeres de elementos) de acuerdo a una medida de similitud entre ellos.

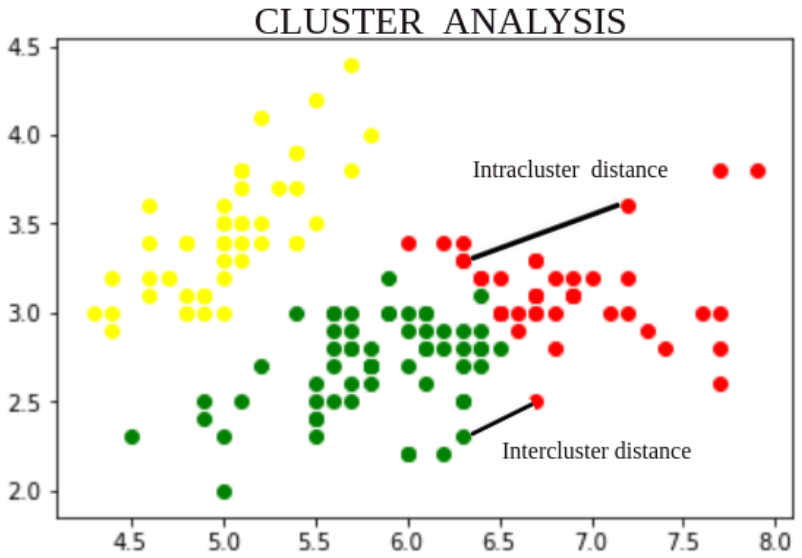


Objetivo

Su objetivo fundamental es identificar clústeres de elementos tal que:

La similitud media entre elementos del mismo clúster sea alta. (**Similitud intra-clúster alta**.)

La similitud media entre elementos de distintos clústeres sea baja. (**Similitud inter-clúster baja**.)



Aplicaciones del análisis de clustering

- **Marketing:** ayuda a los especialistas en marketing a encontrar grupos distintivos entre sus clientes, y así mejorar sus programas de marketing específicos.
- **Biología:** ayuda a definir clasificaciones de plantas y animales o a identificar genes con funcionalidades similares.
- **Descubrimiento web:** útil para clasificar documentos en la web para el descubrimiento de información.
- **Detección de fraudes:** útil en aplicaciones de detección de outliers, como la detección de fraudes con tarjetas de crédito.

Características deseadas con el clustering

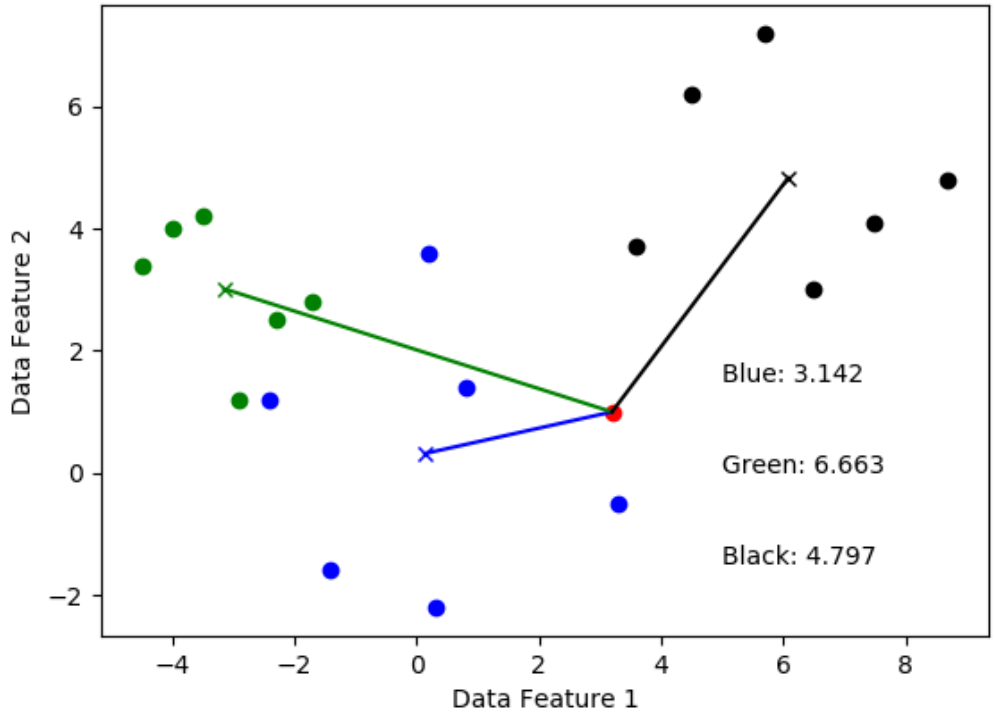
- **Escalabilidad**: los algoritmos de clustering deben poder manejar conjuntos de datos pequeños así como también grandes sin problemas
- Capacidad para **manejar diferentes tipos de atributos**, como datos binarios, cualitativos y numéricos
- **Independiente del orden de entrada de los datos**: los resultados del clustering no deben depender del orden de los datos de entrada
- **Interpretabilidad**: los resultados de los algoritmos de agrupación deben ser interpretables, lógicos y utilizables.

Métricas de distancia

Una métrica de distancia es una función $d(x, y)$ que especifica la distancia entre elementos de un conjunto de números reales no negativos.

Dos elementos son iguales bajo una métrica particular si la distancia entre ellos es cero.

Las funciones de distancia representan un método para calcular la cercanía entre dos elementos.

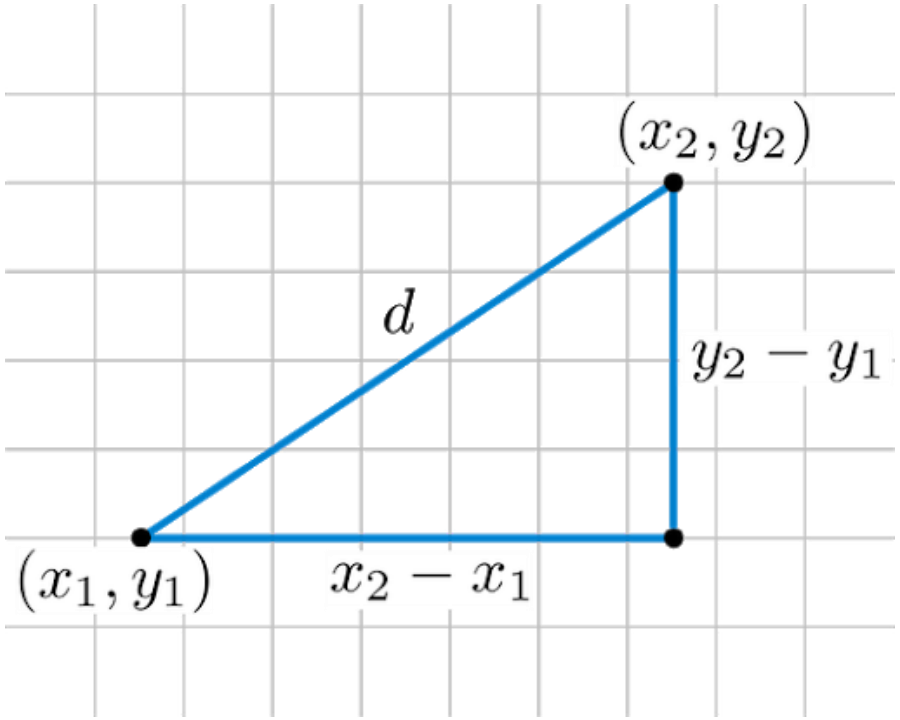


Distancia euclídea

Este tipo de distancia es usada principalmente para calcular distancias.

La distancia entre dos puntos en el plano con coordinadas (x, y) y (a, b) según la fórmula de la distancia euclidiana viene dada por:

$$\text{Euclidean dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$



Distancia euclídea

Usando la formula de la distancia euclídea podemos calcular la distancia de similitud entre personas.

	Variable 1	Variable 2
Persona 1	30	70
Persona 2	40	54
Persona 3	80	50

Los cálculos para la distancia entre persona 1 y 2 es:

$$Euclidean\ dist((30,70),(40,54)) = \sqrt{((30-40)^2 + (70-54)^2)} = \mathbf{18.86}$$

Los cálculos para la distancia entre persona 1 y 3 es:

$$Euclidean\ dist\big((30,70),(80,50)\big) = \sqrt{(30-80)^2 + (70-50)^2} = \mathbf{53.85}$$

Distancia euclídea

Los cálculos para la distancia entre persona 2 y 3 es:

$$Euclidean\ dist\big((40{,}54),(80,50)\big) = \sqrt{(40-80)^2+(54-50)^2}$$

$$= 40.19$$

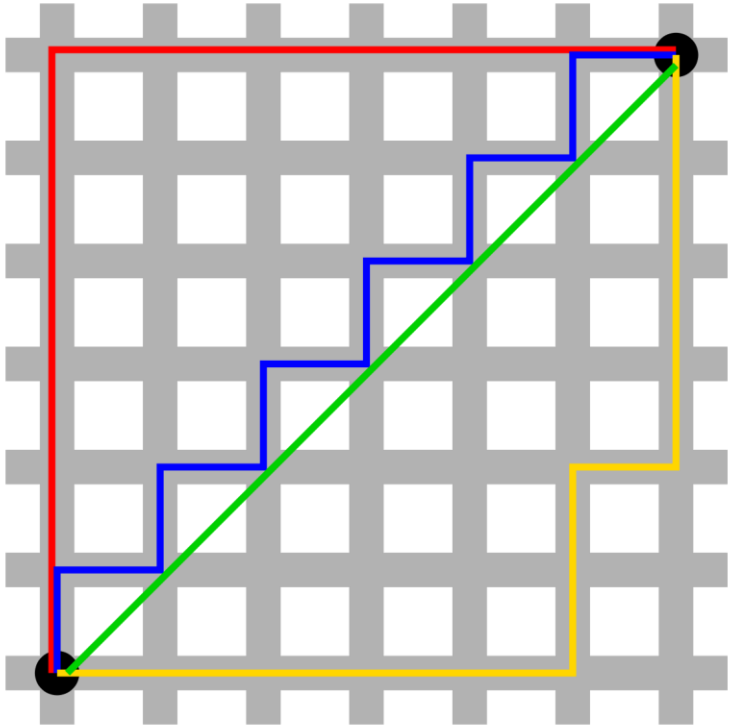
Esto indica que las personas 1 y 2 son las más similares entre sí mientras que las personas 1 y 3 son las más diferentes.

Distancia de Manhattan

Este tipo de distancia es definida como la suma de las longitudes de las proyecciones del segmento de línea entre los dos puntos en los ejes de coordenadas.

La formula está dada por:

$$\text{Manhattan dist}((x, y), (a, b)) = |x - a| + |y - b|$$



Distancia de Manhattan

Usando esta formula también podemos calcular la distancia de similitud entre las personas.

	Variable 1	Variable 2
Persona 1	30	70
Persona 2	40	54
Persona 3	80	50

Los cálculos para la distancia entre la persona 1 y 2 son:

$$Manhattan\ dist\big((30,70),(40,54)\big) = |30 - 40| + |70 - 54|$$

$$= \mathbf{26}$$

Los cálculos para la distancia entre la persona 1 y 3 son:

$$Manhattan\ dist\ \big((30,70),(80,50)\big) = |30 - 80| + |70 - 50|$$

$$= \mathbf{70}$$

Distancia de Manhattan

Y finalmente entre la persona 2 y 3:

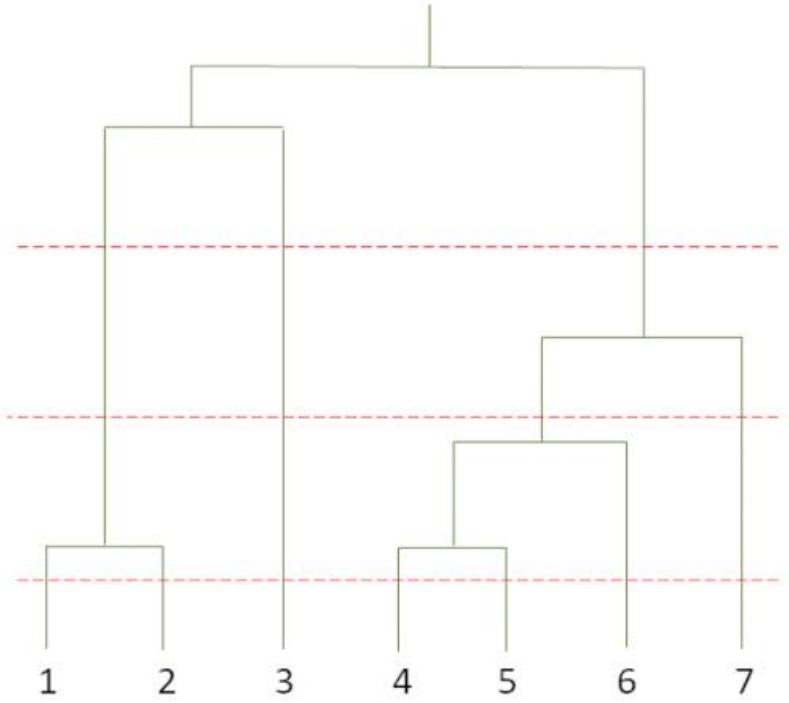
$$Manhattan\ dist((40, 54), (80, 50)) = |40 - 80| + |54 - 50|$$

$$= \mathbf{44}$$

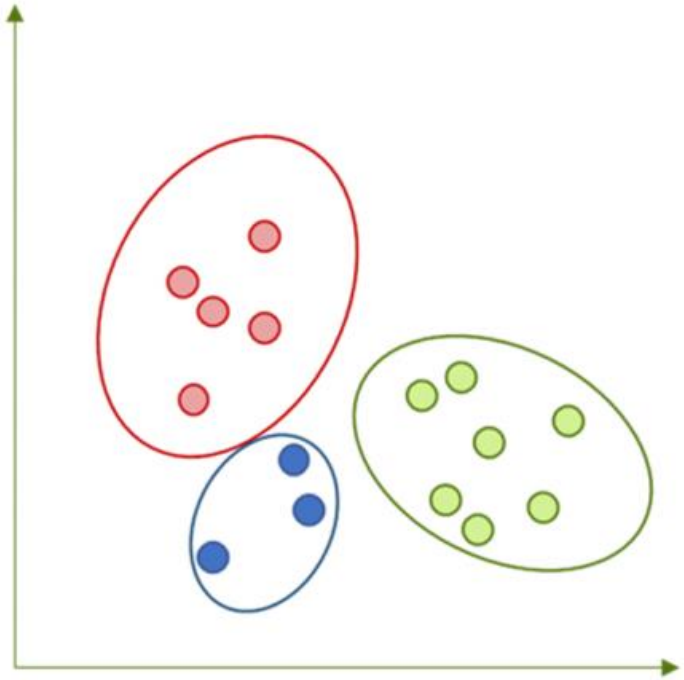
Esto nos indica que las personas 1 y 2 son las más similares y las personas 1 y 3 son las más diferentes, lo cual genera la misma conclusión que con la distancia euclídea.

Tipos de Clustering

Jerárquico

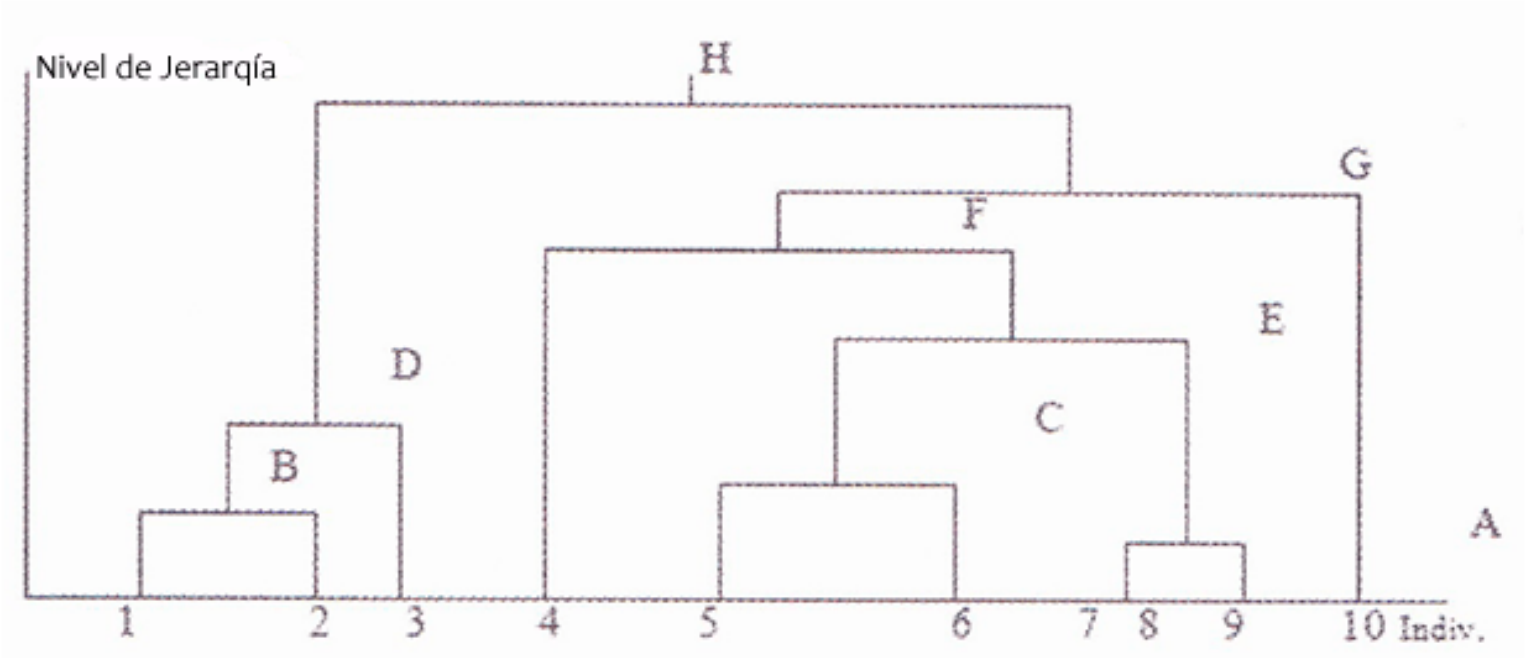


De partición



Clustering jerárquico

Uno de los métodos más utilizados, debido a la visualización práctica en forma de árbol que se obtiene. El clustering jerárquico puede realizarse tanto en forma divisiva o aglomerativa, y permite analizar alternativas para distintos números de grupos.



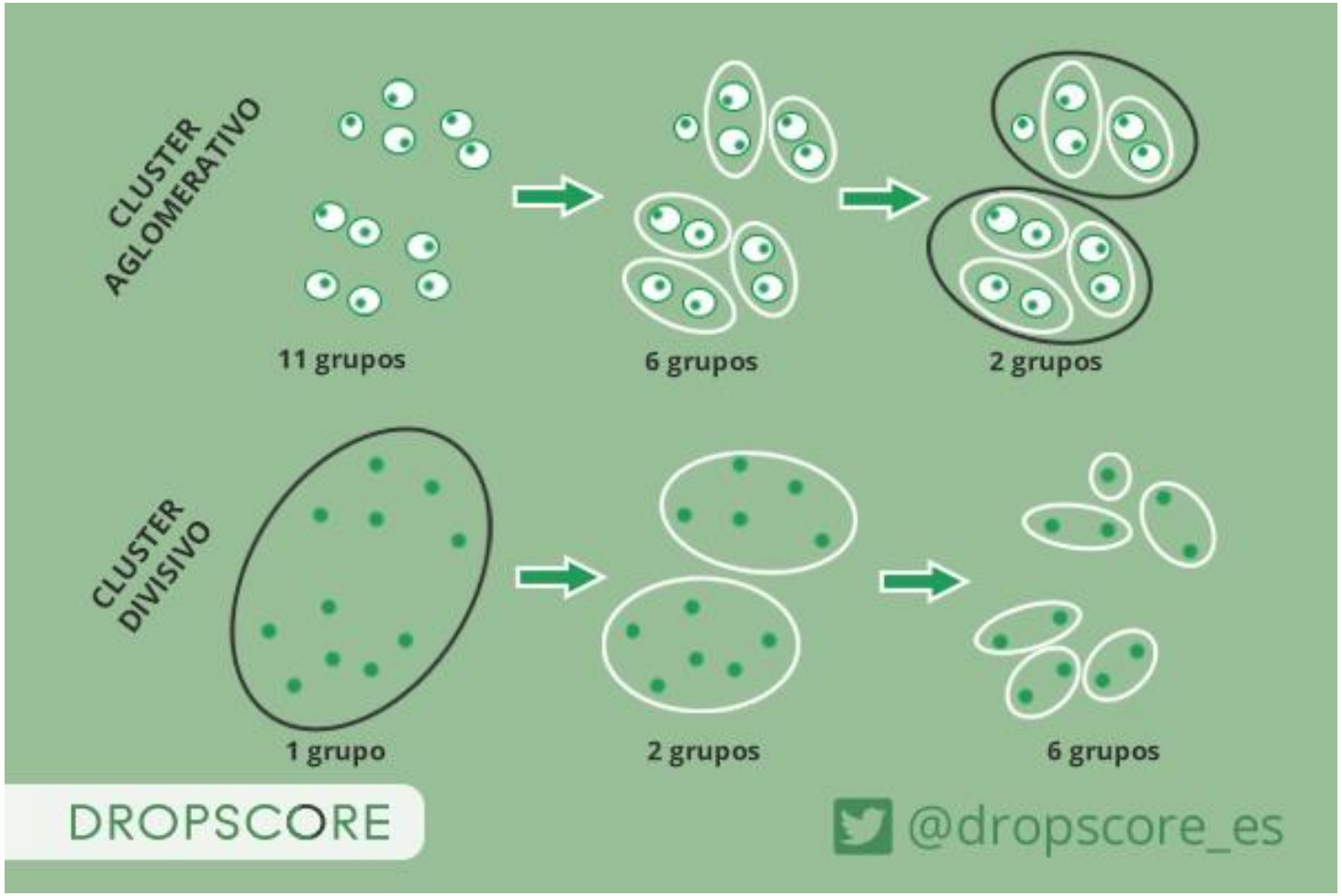
Tipos de clustering jerárquico

- **Clustering jerárquico aglomerativo**

  Se comienza con tantos clústeres como individuos y consiste en ir formando grupos según su similitud.

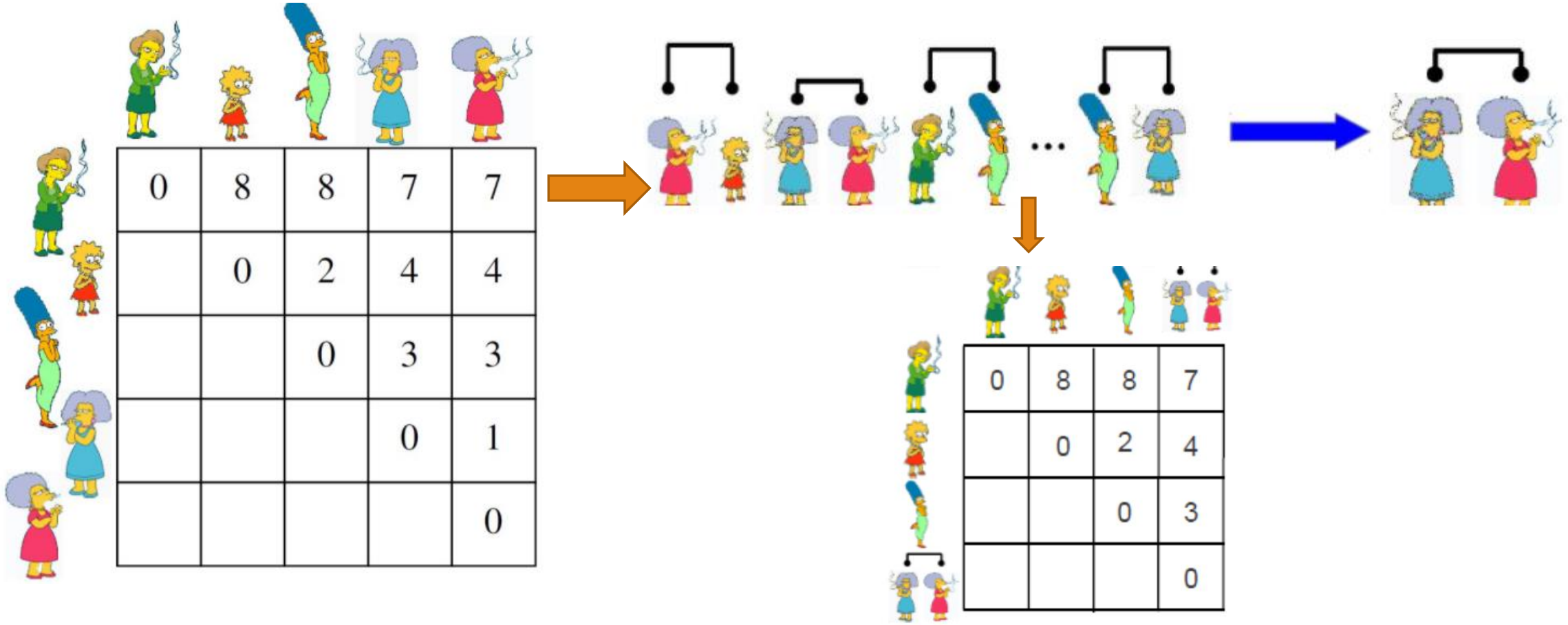
- **Clustering jerárquico de división**

  Se comienza con un único clúster y consiste en ir dividiendo clústeres según la disimilitud entre sus componentes.



Tipos de clustering jerárquico

Ejemplo



Clustering de partición

- Se encarga de dividir un conjunto de datos en una pequeña cantidad de agrupaciones o particiones, basado en sus atributos.

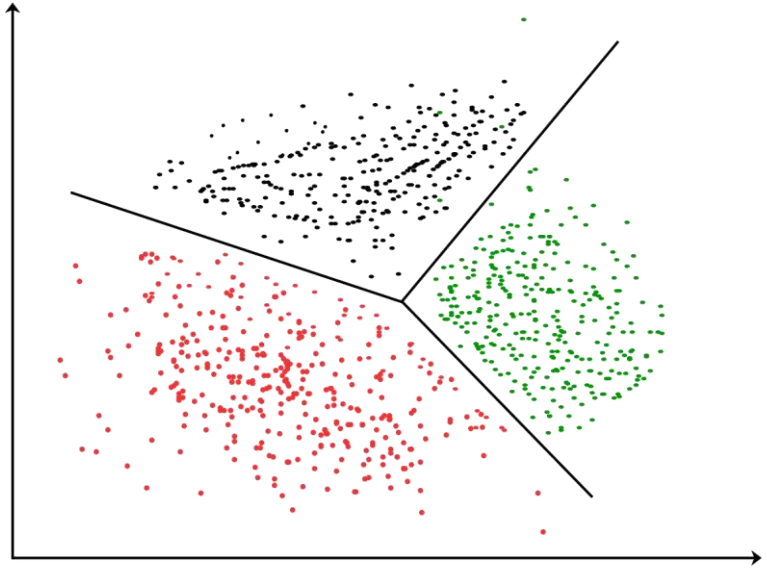
Partición cualitativa

Partición cuantitativa

Clustering de partición

La técnica de clustering de partición entorno a centroides realiza una distribución de los elementos entre un número prefijado de grupos. Esta técnica recibe como dato de entrada el número de clústers a formar además de los elementos a clasificar y la matriz de similitudes.

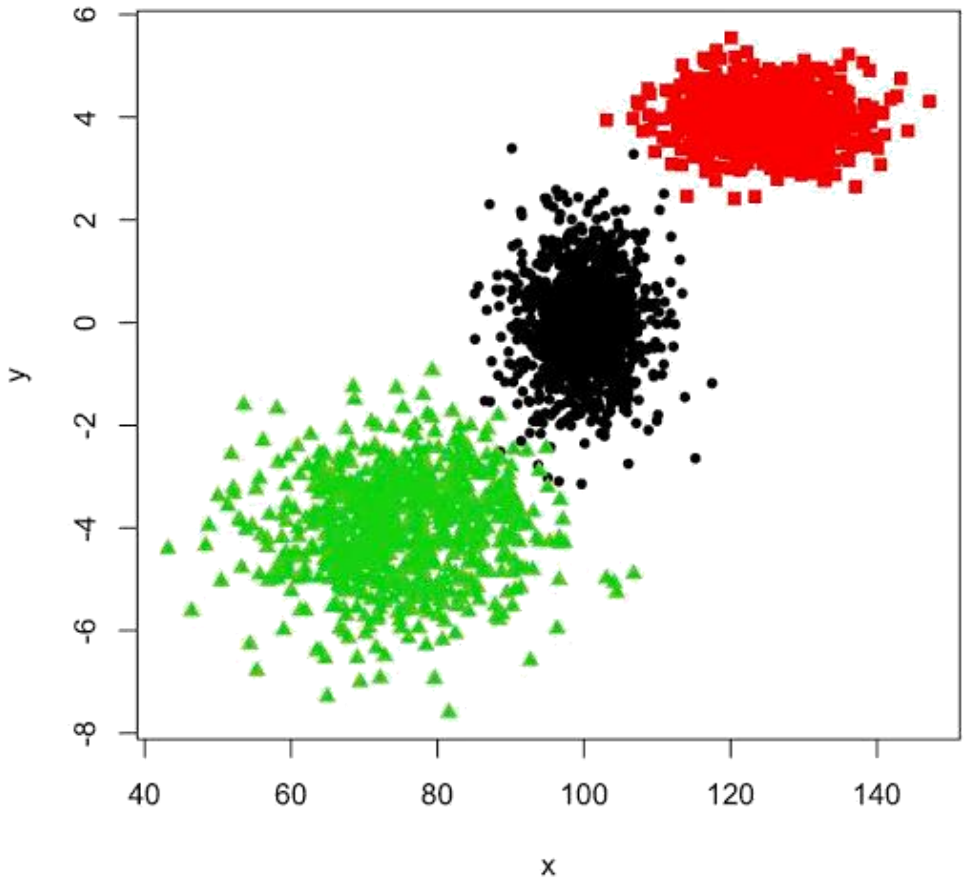
Consiste en agrupar los elementos entorno a elementos centrales llamados centroides a cada clúster. Definimos el centroide de un clúster como aquel elemento que minimiza la suma de las similitudes al resto de los elementos del clúster.



Algoritmo k-means

En el algoritmo k-means, n objetos se agrupan en k agrupaciones en función de características, donde k < n y k es un número entero positivo.

La agrupación de objetos se realiza minimizando la suma de cuadrados de distancias, es decir, una distancia euclidiana entre los datos y el centroide del grupo correspondiente.

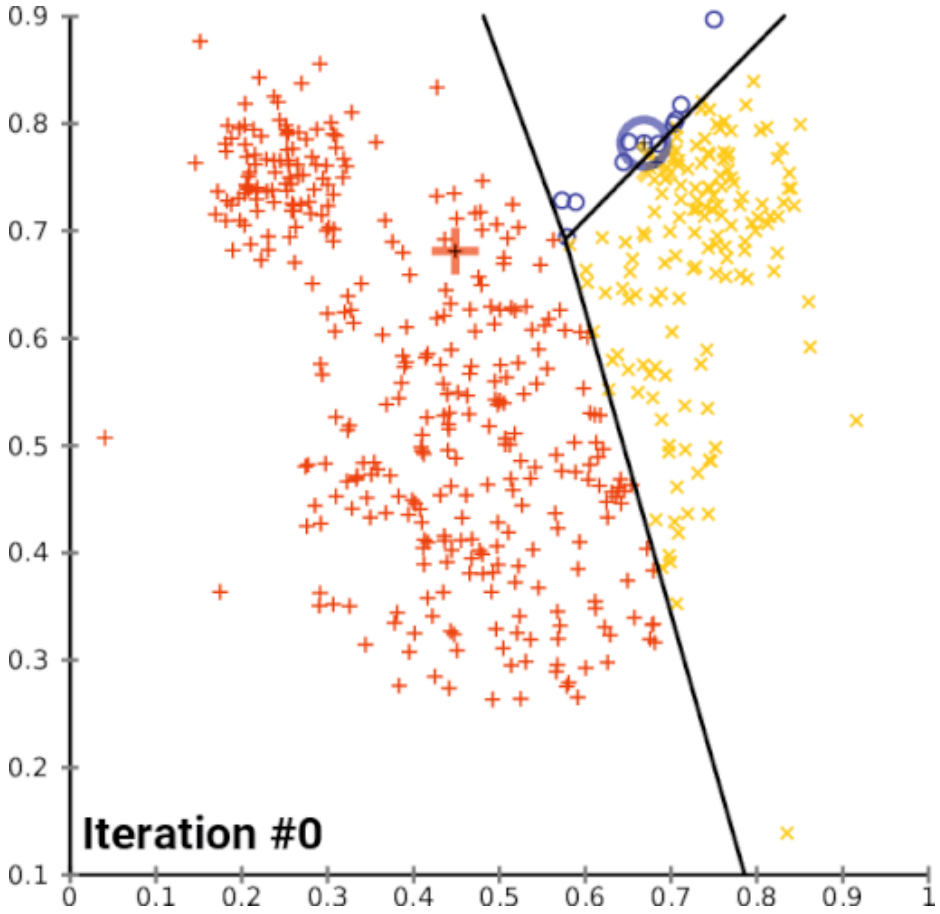


Características del k-means

- Los resultados del algoritmo de k-medias dependen intensamente de las suposiciones iniciales.
- El algoritmo de k-medias es sensible a outliers. Por lo tanto, puede dar malos resultados si se selecciona un outlier como valor inicial.
- El algoritmo de k-means no tiene en cuenta el tamaño de los clústeres. Los grupos pueden ser grandes o pequeños.
- No funciona bien con grupos de diferentes tamaño.

Metodología del k-means

1. Seleccionar el número k de clusters aleatoriamente.
2. Escoger aleatoriamente k centroides.
3. Calcular las distancias de todos los puntos a los **k** centroides.
4. Formar k grupos, asignando cada punto al centroide más cercano.
5. Recalcular los nuevos centroides.
6. Se repiten los pasos 3, 4 y 5 hasta que los centroides no se muevan.



Ejemplo de k-means

Clustering de las calificaciones de estudiantes utilizando el algoritmo de k-means.

Estudiante	Parcial 1	Parcial 2	Parcial 3	Parcial 4
S1	8	20	6	45
S2	6	18	7	42
S3	5	15	6	35
S4	4	13	5	25
S5	9	21	8	48
S6	7	20	9	44
S7	9	17	8	49
S8	8	19	7	39
S9	3	14	4	22
S10	6	15	7	32

Ejemplo de k-means

Paso 1 y 2: Seleccionar el número de clusters aleatoriamente y también los centroides.

k = 3

Estudiante	Parcial 1	Parcial 2	Parcial 3	Parcial 4
S1	8	20	6	45
S2	6	18	7	42
S3	5	15	6	35

Paso 3 y 4: Para calcular las distancias de todos los puntos a los centroides utilizaremos la distancia euclídea.

Estudiante	Parcial 1	Parcial 2	Parcial 3	Parcial 4	C1	C2	C3	Cluster asignado
S1	8	20	6	45	0	4.24	11.57	**C1**
S2	6	18	7	42	4.24	0	7.74	**C2**
S3	5	15	6	35	11.57	7.74	0	**C3**
S4	4	13	5	25	21.59	17.94	10.29	**C3**
S5	9	21	8	48	3.87	7.42	15	**C1**
S6	7	20	9	44	3.32	3.61	10.91	**C1**
S7	9	17	8	49	5.48	7.75	16.73	**C1**
S8	8	19	7	39	6.17	3.74	6.48	**C2**
S9	3	14	4	22	24.37	21.23	13.34	**C3**
S10	6	15	7	32	14.11	10.44	3.32	**C3**

Distancia euclídea entre S1 y C2 =

$$\sqrt{(8-8)^2 + (20-20)^2 + (6-6)^2 + (45-45)^2} = \mathbf{0}$$

Distancia euclídea entre S1 y C2 =

$$\sqrt{(8-6)^2 + (20-18)^2 + (6-7)^2 + (45-42)^2} = \mathbf{4.24}$$

Distancia euclídea entre S1 y C3 =

$$\sqrt{(8-5)^2 + (20-15)^2 + (6-6)^2 + (45-35)^2} = \mathbf{11.57}$$

Distancia euclídea entre S2 y C1 =

$$\sqrt{(6-8)^2 + (18-20)^2 + (7-6)^2 + (42-45)^2} = \mathbf{4.24}$$

Distancia euclídea entre S2 y C2 =

$$\sqrt{(6-6)^2 + (18-18)^2 + (7-7)^2 + (42-42)^2} = \mathbf{0}$$

Los clusters quedan como a continuación:

Cluster 1: S1, S5, S6, S7
Cluster 2: S2, S8
Cluster 3: S3, S4, S9, S10

Paso 5: Recalcular los nuevos centroides

	Parcial 1	Parcial 2	Parcial 3	Parcial 4
C1	8.25 Avg(8,9,7,9)	19.5 Avg(20,21,20,17)	7.75 Avg(6,8,9,7)	46.5 Avg(45,48,44,39)
C2	7 Avg(6,8)	18.5 Avg(18,19)	7 Avg(7,7)	40.5 Avg(42,39)
C3	4.5 Avg(5,4,3,6)	14.25 Avg(15,13,14,15)	5.5 Avg(6,5,4,7)	28.5 Avg(35,25,22,32)

Iteración 2

Estudiante	Parcial 1	Parcial 2	Parcial 3	Parcial 4
S1	8	20	6	45
S2	6	18	7	42
S3	5	15	6	35
S4	4	13	5	25
S5	9	21	8	48
S6	7	20	9	44
S7	9	17	8	49
S8	8	19	7	39
S9	3	14	4	22
S10	6	15	7	32

C1	C2	C3	Cluster asignado
2.37	4.94	17.83	C1
5.11	1.87	14.17	C2
12.89	6.89	6.58	C3
23.02	16.84	3.78	C3
2.26	8.22	21.27	C1
3.10	4.30	17.08	C1
3.62	8.92	21.31	C1
7.55	1.87	12.04	C2
25.92	19.69	6.86	C3
15.37	9.25	4.12	C3

Conclusión del ejemplo

De nuevo, obtenemos los mismos clusters, lo que nos indica que no hubo ningún cambio en los clusters y el algoritmo para.

Los clusters finales se muestra a continuación:

Cluster 1: S1, S5, S6, S7
Cluster 2: S2, S8
Cluster 3: S3, S4, S9, S10

Otros tipos de clustering

- **Método Density-based**

Se basa en funciones de conectividad y densidad. Se trata de los algoritmos de clustering que agrupan objetos según un criterio de densidad más que de proximidad.

- **Método Grid-based**

Se basa en una estructura de granularidad de múltiples niveles. Particionan el espacio y buscan los objetos que pertenecen a cada celda resultante de la partición.

Referencias

https://www.cs.us.es/~fran/curso_unia/clustering.html#:~:text=Clustering%20es%20una%20t%C3%A9cnica%20de,del%20mismo%20cl%C3%BAster%20sea%20alta.

https://medium.com/@jaywrkr/miner%C3%ADa-de-datos-4-3bc03e7fc234

https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/

http://blog.dropscore.com/metodos-de-clustering-mas-utilizados/

Tan, P. et al. (2006). Introduction to Data Mining. Pearson Addison-Wesley.

Bhatia, P. (2019). Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press.