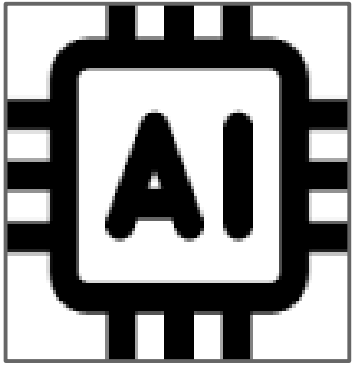
Introducción a la Inteligencia Artificial



Agrupamiento o Clustering - K-means

Contacto: ia@frh.utn.edu.ar

En esta Presentación

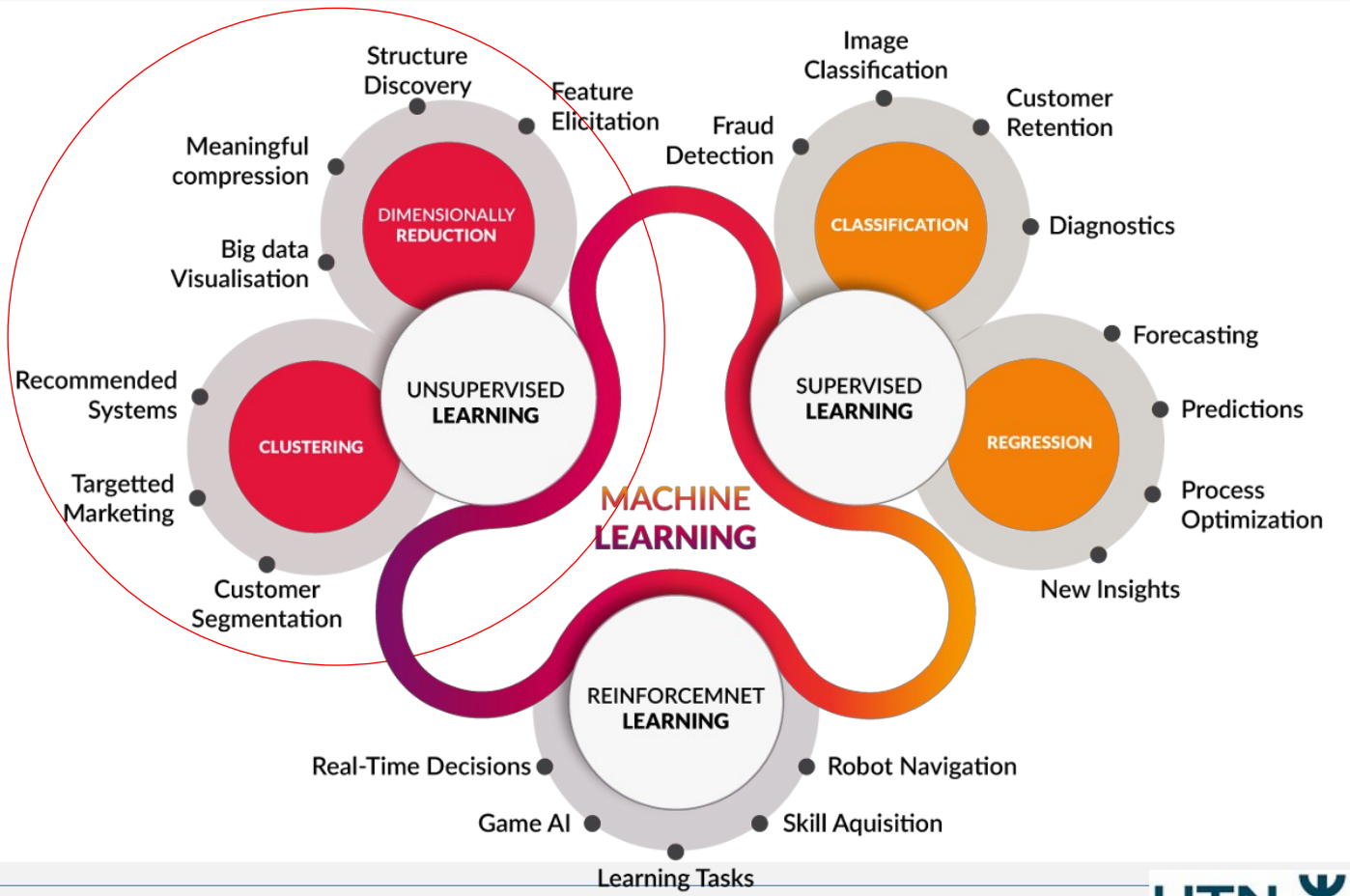
1. Introducción a Clustering

- Nociones de Clustering

2. El algoritmo k-means

- ¿Qué es k-means?
- Principios de funcionamiento
- Casos de Uso

Introducción a Clustering

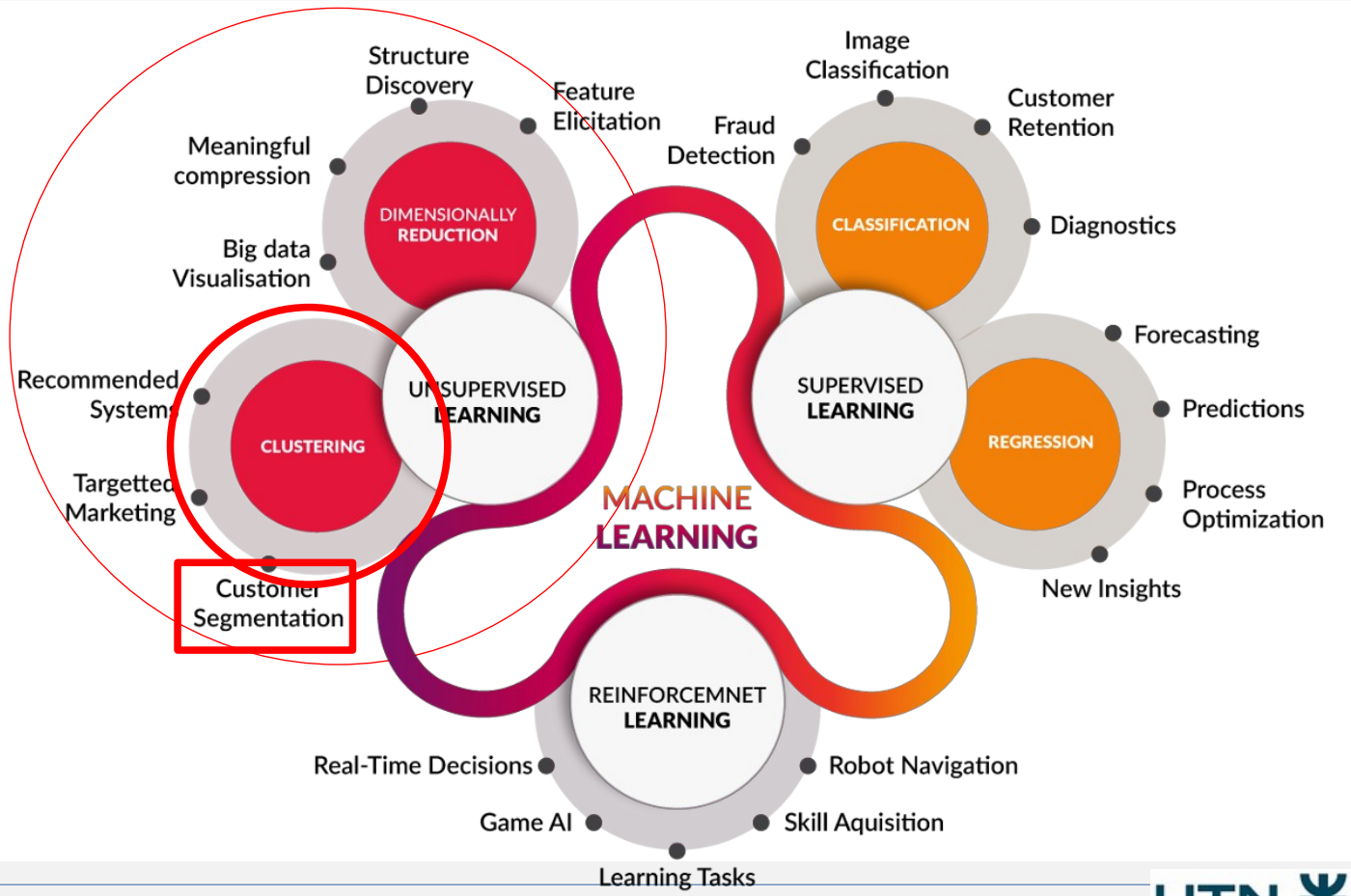


Aprendizaje no Supervisado

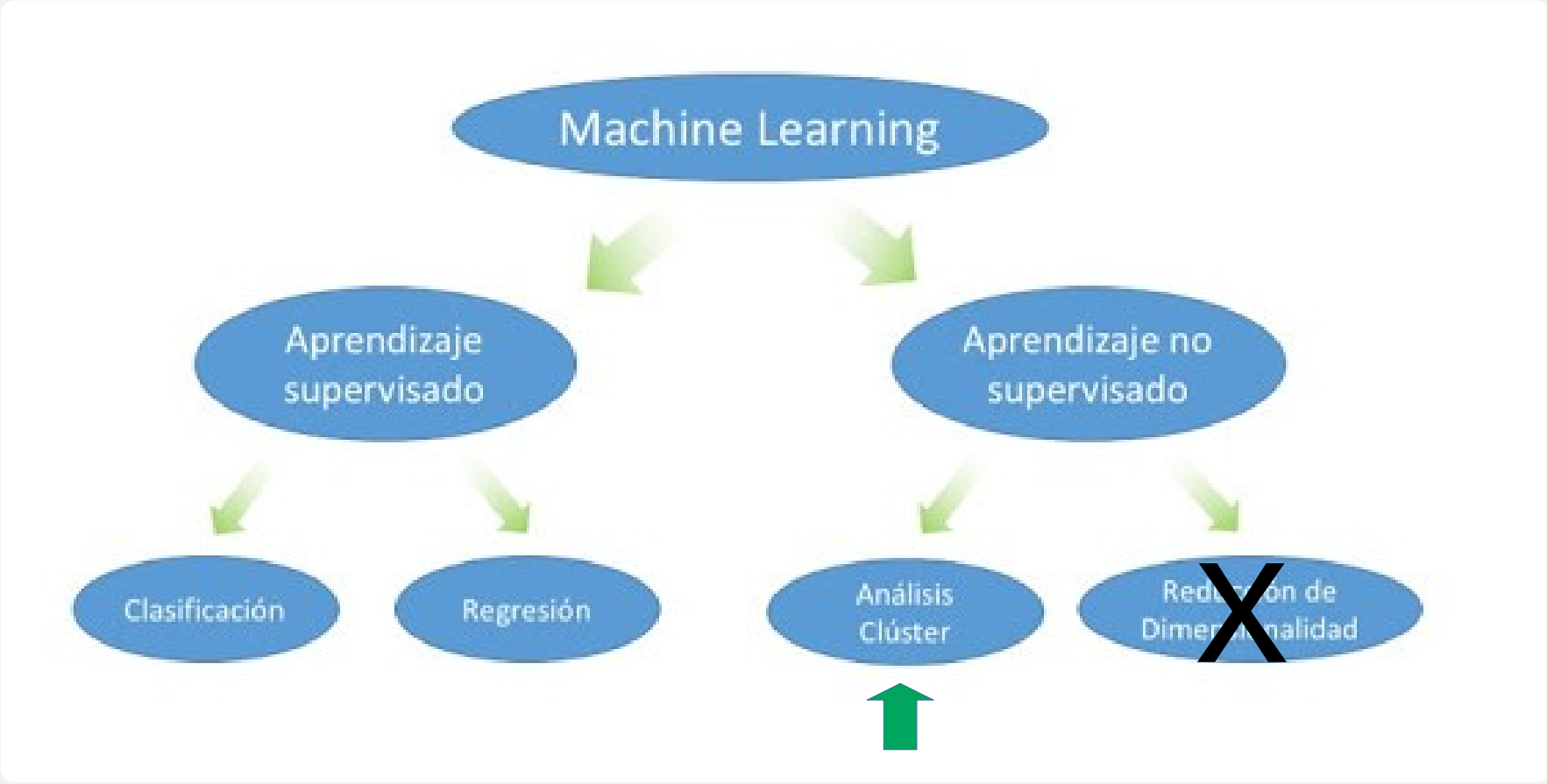
Enfoque:

En Aprendizaje no Supervisado, los datos no tienen etiquetas de clase y queremos identificarlos de algún modo.

Introducción a Clustering



Introducción a Clustering



¿Qué es Clustering?

Se agrupa una serie de vectores (registros, datos) de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud.

➔Se logra una separación de elementos en diferentes conjuntos que poseen ciertas características en común.

Ejemplo de aplicación: Banca

- Se desea ofrecer tarjetas de crédito a sus clientes.
- *Las campañas generales rara vez son efectivas*. Se requiere algún tipo del conocimiento del cliente.
- Resulta inviable analizar manualmente a cada cliente.
- Entonces se decide segmentar a sus clientes en grupos
- Criterio: segmentación por ingresos.

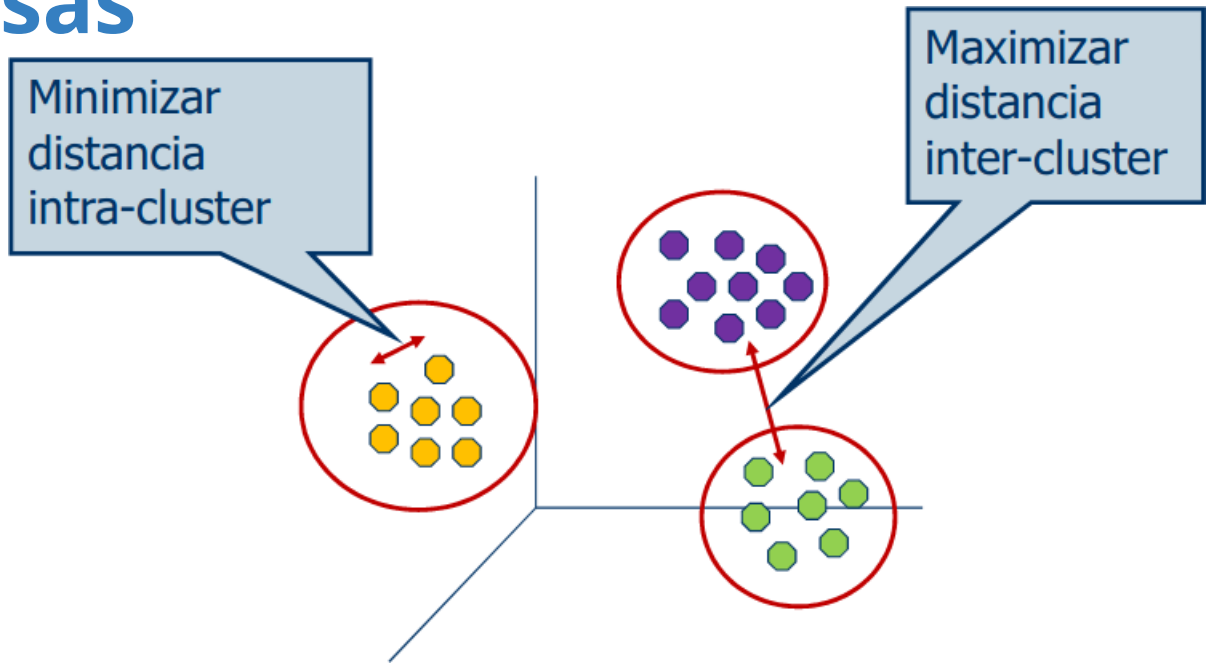
Premisas

1)Los elementos de un mismo clúster son todos similares entre sí (distancia intra cluster).

2)Un elemento de un clúster es diferente a todos los elementos de otros clústeres (distancia inter cluster).

Fuente:
https://aprendeia.com/metodo-de-agrupamiento-o-clustering-aprendizaje-no-supervisado/

Premisas



Fuente: Berzal, F. “Clustering”. Universidad de Granada, 2005.
Link: https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

Algunas distancias

Distancia de Manhattan (taxi)

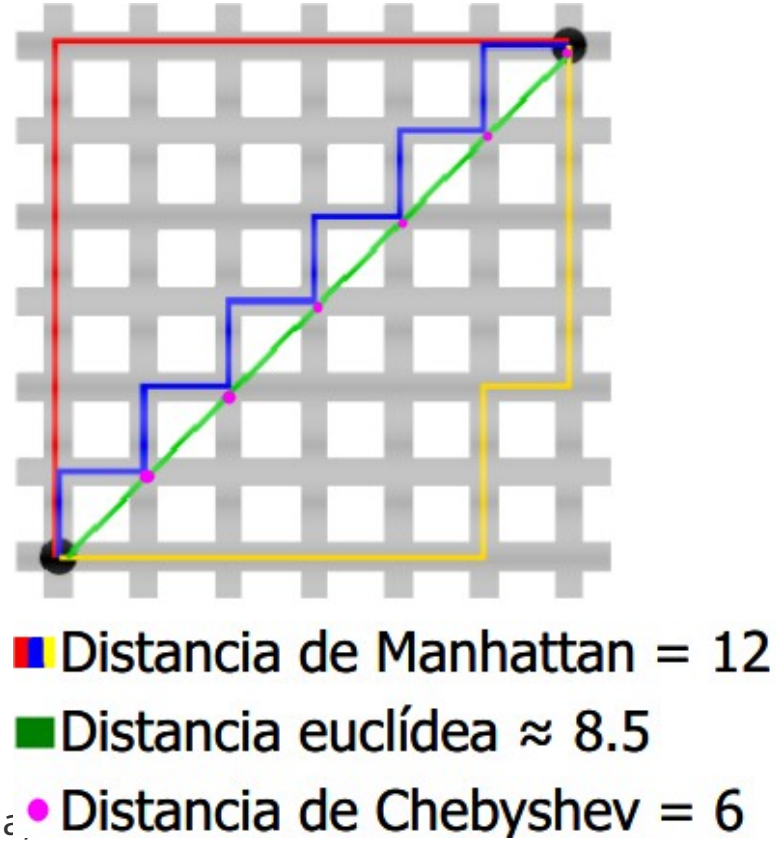
$$d_1(x, y) = \sum_{j=1}^{J} |x_j - y_j|$$

Distancia Euclídea

$$d_2(x, y) = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

Distancia de Chebyshev (Rey)

$$d_\infty(x, y) = \max_{j=1..J} |x_j - y_j|$$



Fuente: Berzal, F. “Clustering”. Universidad de Granada

Link: https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

Algoritmos de Clustering

Clustering por particiones:

- Suele fijarse k (O(xn): si se fija k)
- Ejemplos: K-means, Clarans

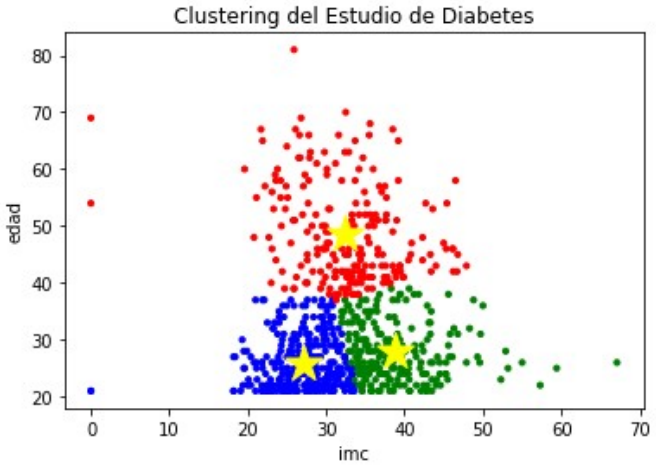
Clustering jerárquico:

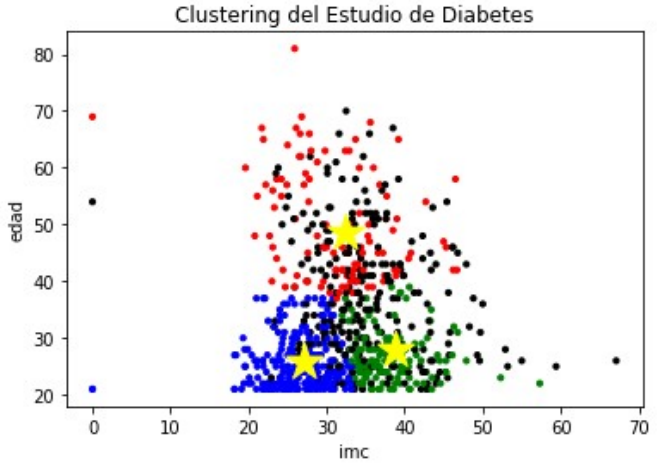
- K no se fija ($O(n^2)$)
- Ejemplos: Birch, Rock, Chamaleon

Fuente: Berzal, F. “Clustering”. Universidad de Granada, 2005.
Link: https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

Qué es k-means

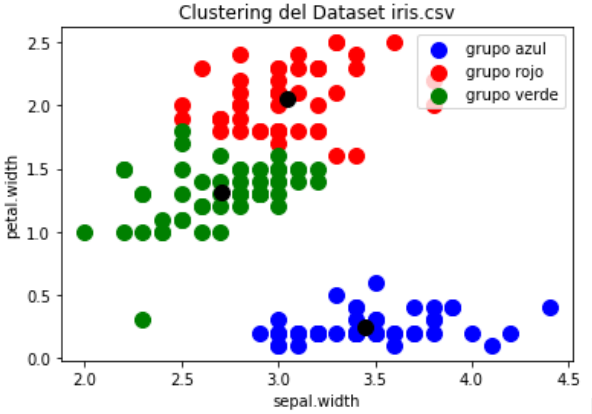
- Es uno de los algoritmos de clustering más usados.
- Presenta la forma más eficiente de dividir al conjunto de datos en “k” clusters o grupos.





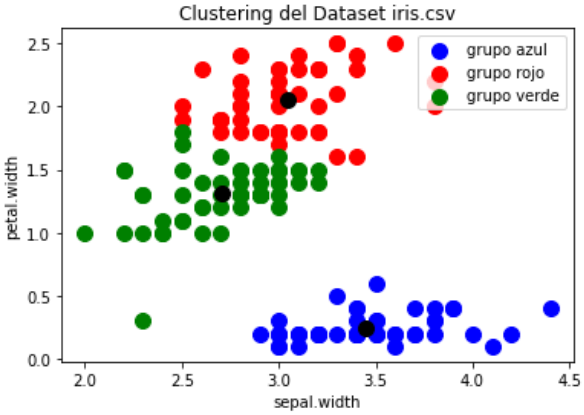
Qué es k-means

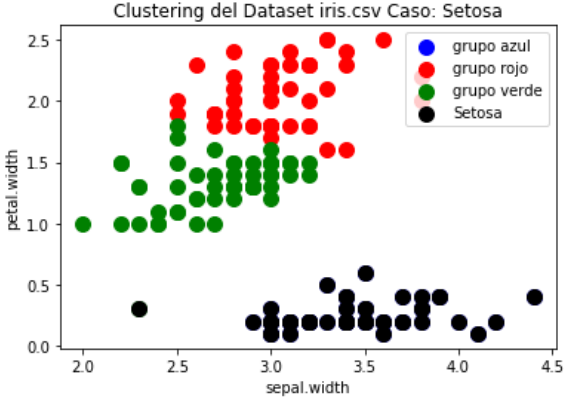
- El agrupamiento se realiza entre datos cuya distancia cuadrática a un punto llamado centroide se minimiza.
- Es decir, definidos los k centroides, cada dato tendrá un centroide al cual su distancia es menor.



Qué es k-means

- El agrupamiento se realiza entre datos cuya distancia cuadrática a un punto llamado centroide se minimiza.
- Es decir, definidos los k centroides, cada dato tendrá un centroide al cual su distancia es menor.





Qué es k-means

- **Inconveniente:** Es necesario definir k
- **Posible soluciones:**
  - Analizar la “curva Elbow” donde la pendiente cambia abruptamente.
  - Iniciar con k=2, variar su valor y evaluar los resultados: elegir k para el menor SSE (suma de errores cuadrados).

Principio de funcionamiento

Determinación de k: La curva Elbow

Se calcula la inercia obtenida tras aplicar K-means a diferente número de Clusters (desde 1 a N).

$$Inercia = \sum_{i=1}^{n} |x_i - \mu|^2$$

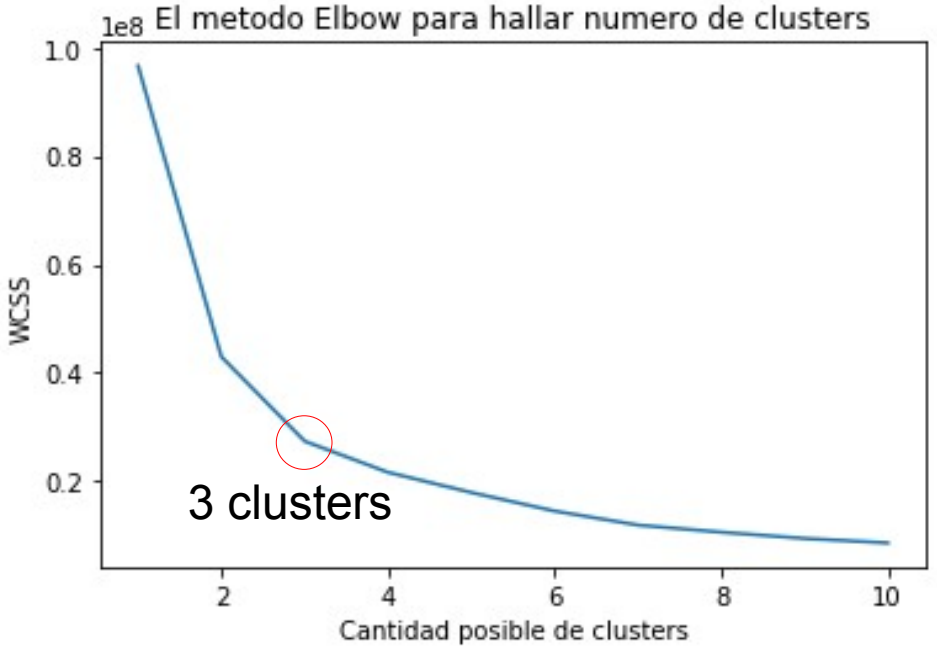
Dada la gráfica lineal de la inercia respecto del número de Clusters, se debe de apreciar un cambio brusco en la evolución de la inercia → k: número óptimo de Clusters.

Fuente: https://jarroba.com/seleccion-del-numero-optimo-clusters/

Principio de funcionamiento

Determinación de k: La curva Elbow

Se busca el valor de k donde la pendiente de la curva elbow tiene -45°.



Principio de funcionamiento

Determinación de k: Evaluación de SSE

Usar un valor de k, ver los resultados del SSE y ajustar.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d^2(m_i, x)$$

Probamos hasta hallar el valor de k que minimiza globalmente el SSE

Fuente: Berzal, F. “Clustering”. Universidad de Granada, 2005.
Link: https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

Principio de funcionamiento

Inicialización:

- K: Asignación de número k de clusters.
- Centroides Iniciales: Se ubican k centroides aleatorios.
- Primer Agrupamiento: Se asigna cada valor de la muestra al centroide cuya distancia sea la menor.

Fuente: Duk2, “K-Means” Link: https://estrategiastrading.com/k-means/

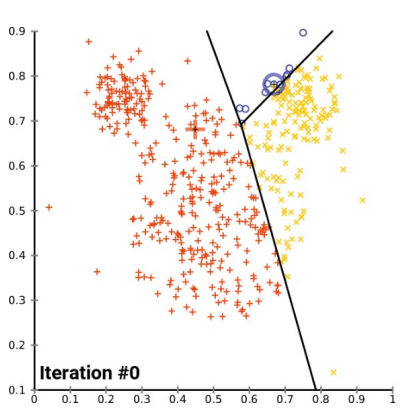
Principio de funcionamiento

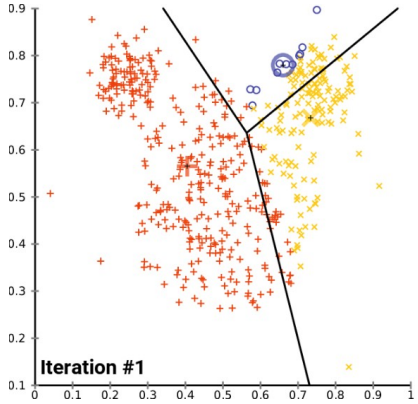
Fase iterativa:

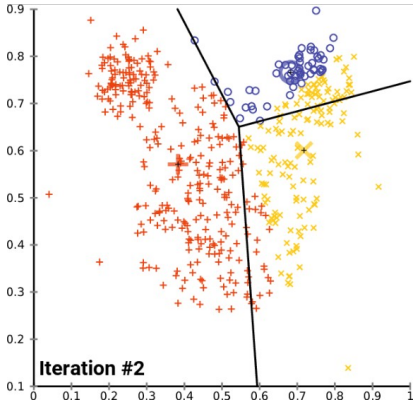
- Nuevos Centroides: cada centroide se desplaza a la ubicación media de las muestras de su grupo.
- Nuevos Agrupamientos: Se recalculan los agrupamientos al nuevo centroide de menor distancia.
  - *Se repite mientras se maximiza la distancia inter-clusters y se minimiza la distancia intra-clusters.*

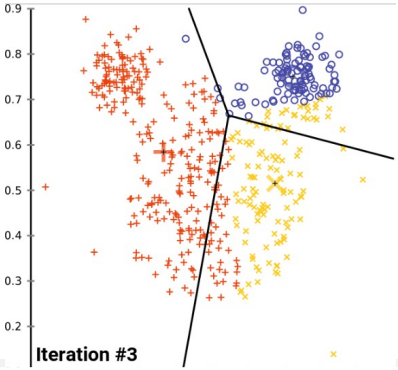
Principio de funcionamiento

Ejemplo:

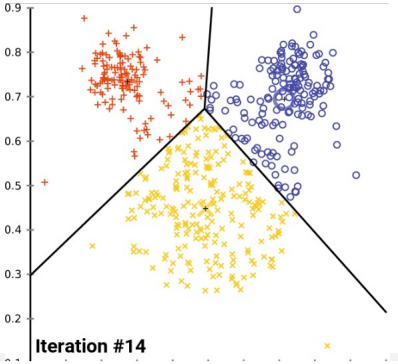








...



Fuente del Ejemplo: https://upload.wikimedia.org/wikipedia/commons/e/ea/K-means_convergence.gif

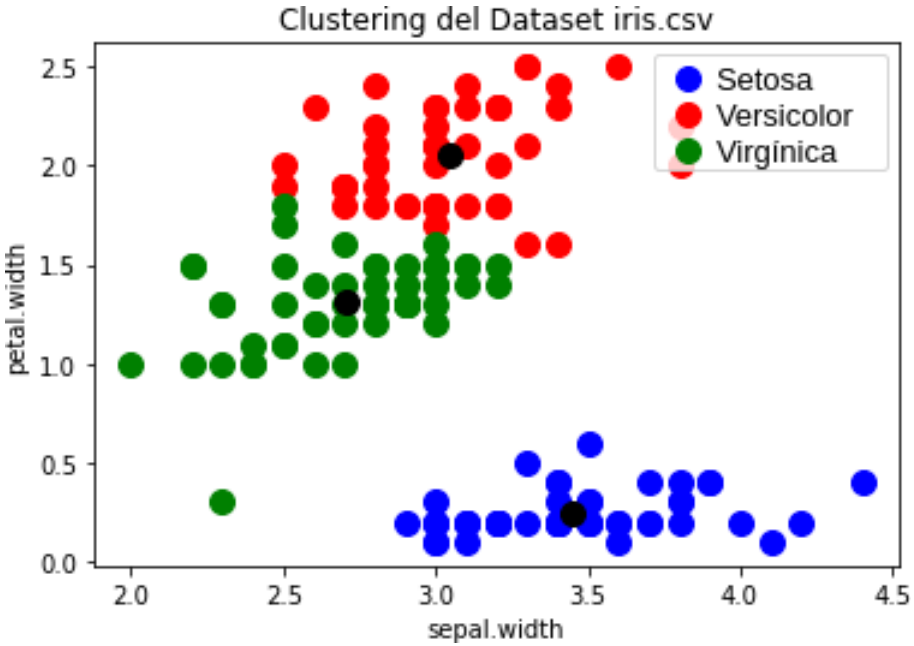
Casos de uso

- En biología para clasificar animales y plantas.
- En medicina para identificar enfermedades.
- En marketing para identificar personas con hábitos de compras similares.
- En teoría de la señal pueden servir para eliminar ruidos.

Fuente: Link https://es.wikipedia.org/wiki/Algoritmo_de_agrupamiento#Algoritmos

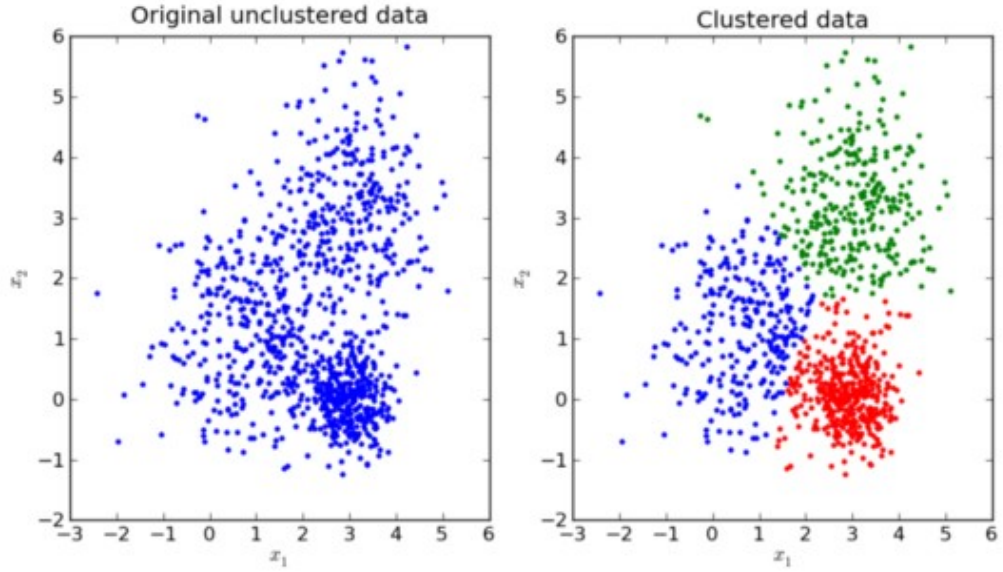
Casos de uso

- En biología para clasificar animales y plantas.



Casos de uso

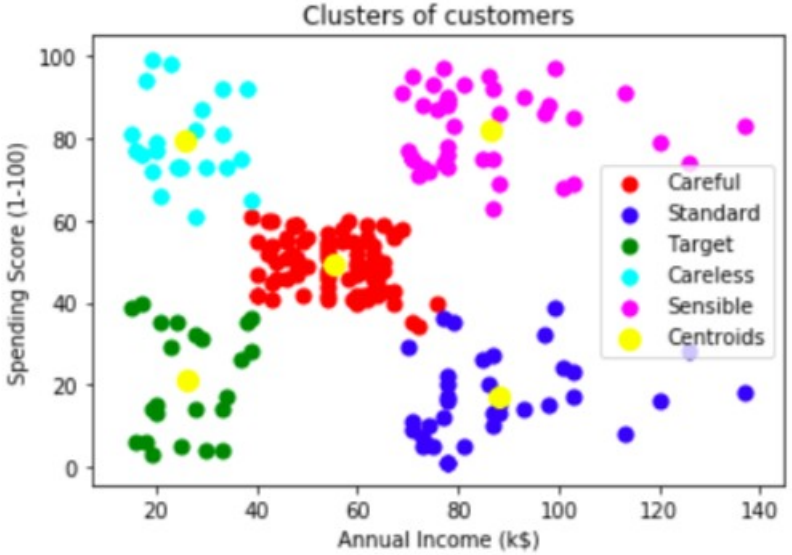
- En medicina para identificar enfermedades.



Fuente: Sitio medium.com. Link corto: https://bit.ly/35JYz0Z

Casos de uso

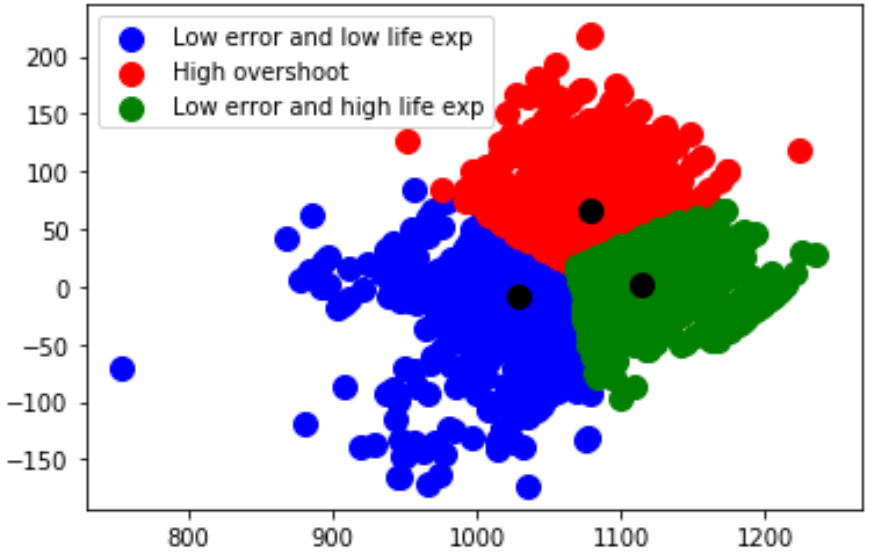
- En marketing para identificar personas con hábitos de compras similares.



Fuente: Sitio kdnuggets.com. Link corto: https://bit.ly/3xJ06jT

Casos de uso

- En Seguros para identificar y clasificar a los asegurados en grupos según siniestralidad.



Casos de uso

- En teoría de la señal pueden servir para eliminar ruidos.

Fuente: Ronchetti, Franco. “Curso Aprendizaje automático”, UNLP, 2017

¿Preguntas?