



Aprendizaje No Supervisado

Clase sincrónica

Implementar modelos de aprendizaje automático por medio de técnicas estadísticas, adecuando los diferentes algoritmos debidamente a la situación y requerimientos necesarios

- Unidad 1: Introducción al Machine Learning
- Unidad 2: Aprendizaje Supervisado y No Supervisado
(Parte I: No supervisado)
(Parte II: Clasificación)
(Parte III: Clasificación)
(Parte IV: Regresión)
(Parte V: Series de tiempo)
- Unidad 3: Aplicando lo aprendido
(Parte I: Preprocesamiento de datos)
(Parte II: Modelamiento)



Te encuentras aquí



¿Qué aprenderás en esta sesión?

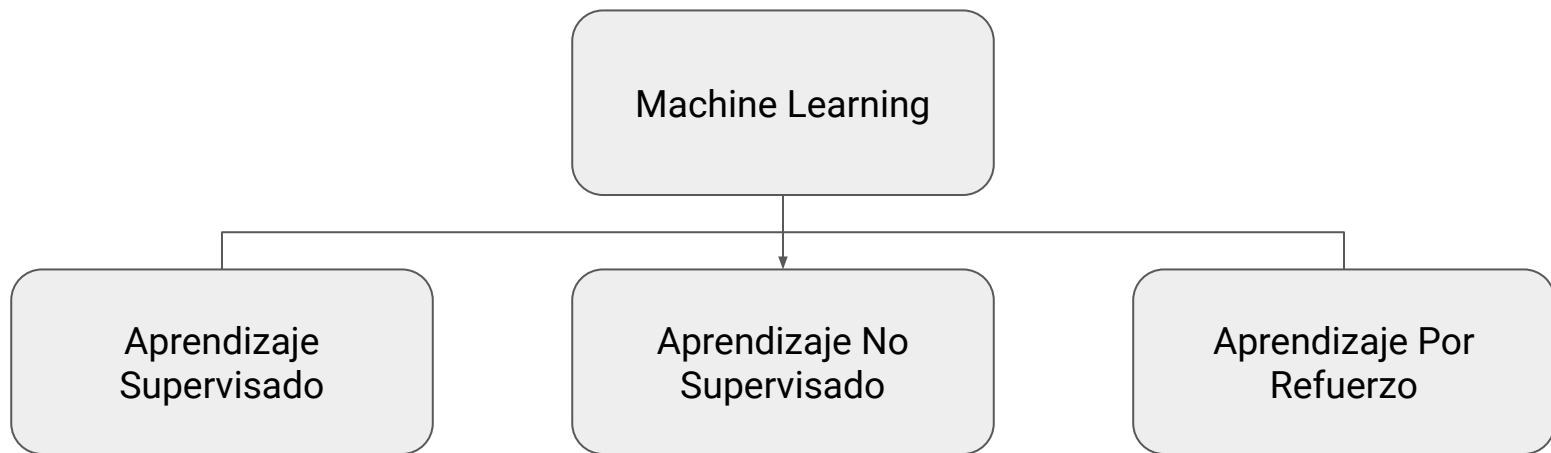
A utilizar algoritmos de clustering, cuáles son, su utilidad, implementación y evaluación.

¿Qué se entiende por
Machine Learning?
¿Qué tipos de
aprendizaje existen?



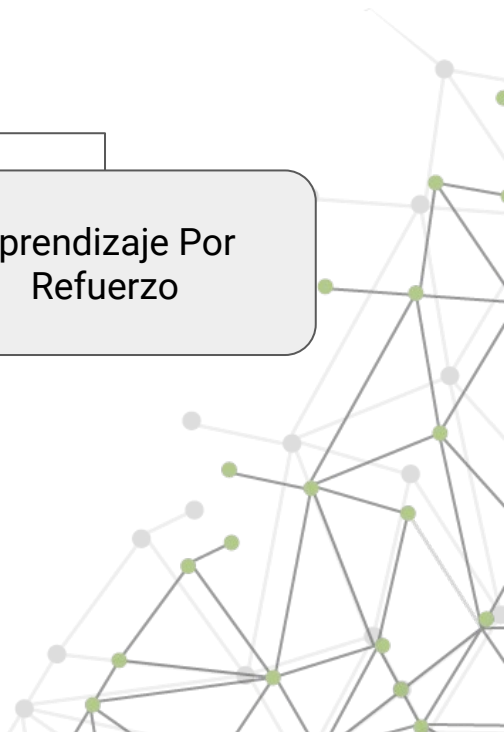
Repaso

Módulo Introducción Machine Learning



Descubrir patrones, estructuras o relaciones ocultas en un conjunto de datos sin la necesidad de tener etiquetas o variables objetivo.

Reducción de dimensionalidad y Clustering.



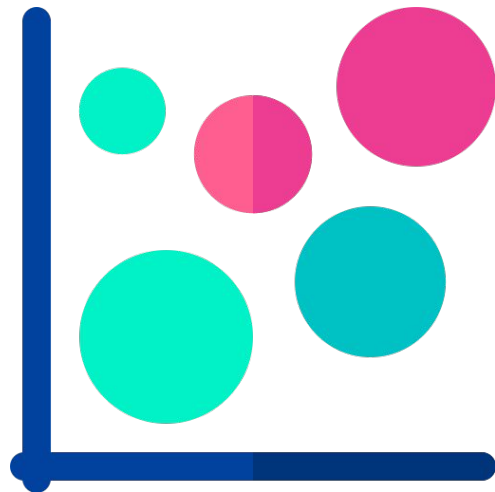
/* Clustering */

Clustering

¿Qué es?

El **clustering** es un conjunto de técnicas de aprendizaje no supervisado que buscan **agrupar objetos similares entre sí** a partir de los datos con los que se intenta describir los objetos a agrupar. Así, la definición de los grupos va a depender de qué variables se utilizan para generar los clusters.

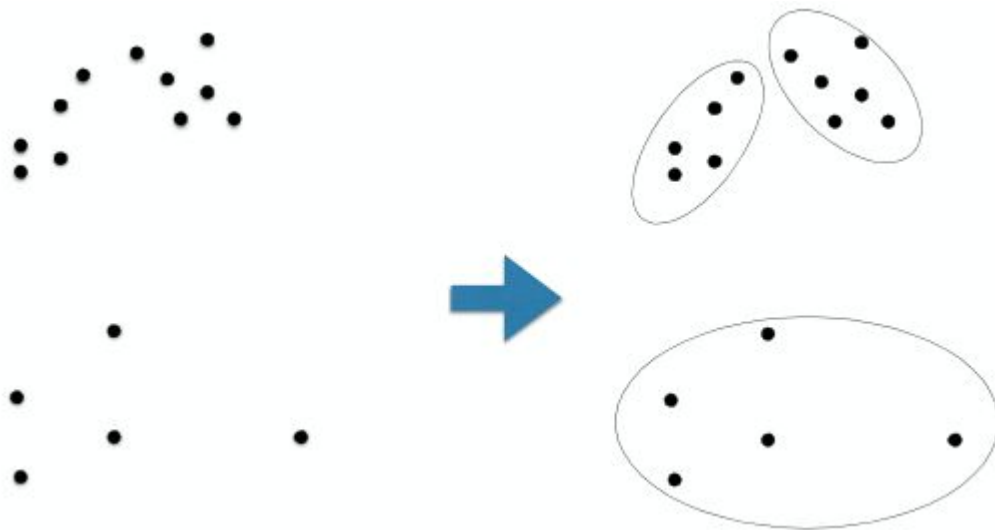
Un **cluster** es una colección o conjunto de objetos que comparten características similares entre sí y difieren de otros objetos en el conjunto de datos. Estos son los resultantes de aplicar un algoritmo de clustering a un set de datos.



Clustering

¿Qué es?

Dado un conjunto de observaciones (puntos), se busca organizarlos en clusters (grupos o clases).



Ejercicio:
Comente ejemplos de
clustering que imaginan o
conocen



Ejercicio

Ejemplos de aplicaciones de clustering en diferentes industrias

Retail

- Segmentación de clientes según su comportamiento transaccional
- Agrupación de productos por características, precio, popularidad. (Gestión de inventario, planificación de promociones)

Telecomunicaciones

- Cluster por patrones de uso de servicio
- Agrupación de torres celulares para planificación de red

Minería

- Clusters de datos geológicos para identificar áreas con características similares

Banca

- Segmentación de clientes para planes de fidelización y campañas de marketing

Salud

- Agrupación de genes y proteínas según su expresión, funciones e interacción



/* Métodos de evaluación de clusters */

Validación de un modelo de clustering

Dificultades

Validar un modelo de clustering puede ser desafiante debido a la falta de etiquetas. En el aprendizaje no supervisado no se conocen las categorías o grupos a priori, lo que dificulta la validación directa del modelo.

Ausencia de
etiquetas




Evaluación
intrínseca

Subjetividad de los
resultados

Validación de un modelo de clustering

Tipos

A pesar de lo anterior existen formas de medir la calidad de los modelos, donde están los siguientes grandes grupos:



Validación
cuantitativa

Validación de
Expertos

Validación cuantitativa de clustering

Métricas

Para realizar esta validación se van a utilizar métricas de calidad de los clusters generados, para lo cual las más utilizadas son:

Distancia Intra
Cluster (SSE)

Coeficiente de
Silhouette

Índice de
Davies-Bouldin

Validación cuantitativa de clustering

Distancia Intra Cluster (SSE)

La distancia intracluster mide la coherencia de los objetos dentro de un mismo grupo o cluster. Cuanto menor sea la distancia intracluster, mayor será la cohesión dentro del cluster. Se busca que los objetos dentro del mismo cluster estén lo más cerca posible unos de otros.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} (x - m_i)^2$$

m_i : corresponde al promedio del cluster i .

C_i : corresponde al conjunto de muestras del cluster i

Validación cuantitativa de clustering

Coeficiente de Silhouette

Mide la cohesión y separación de los clusters. Proporciona un valor entre -1 y 1, donde un valor cercano a 1 indica una buena separación entre los clusters y una cohesión interna alta.

Se puede calcular para puntos individuales, como también para clusters (promedios)

$$s = \frac{b-a}{\max(a,b)}$$

a: distancia promedio de i a los puntos de su propio cluster

b: mínimo distancia promedio de i a puntos de otro cluster

Validación cuantitativa de clustering

Índice de Davies-Bouldin

Mide la similitud media entre los clusters y la distancia entre los centroides de los clusters. Un valor más bajo indica una mejor separación y cohesión de los clusters.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij}$$

s_i : promedio de distancia de cada punto del cluster i al centroide del cluster

d_{ij} : distancia entre el centroide del cluster i y j

Validación de Expertos

La validación por expertos implica la participación de personas con conocimiento y experiencia en el dominio de los datos para evaluar y validar los resultados del clustering.

- Revisar y evaluar visualmente los clusters generados en función de su conocimiento y experiencia.
- Interpretar y asignar etiquetas o categorías a los clusters identificados.
- Evaluar la coherencia y relevancia de los grupos generados en relación con los objetivos del análisis.



¿Cuál es la importancia de la validación por expertos en la evaluación de la calidad de los clusters generados por un modelo?



/* Algoritmos de Clustering */

Algoritmos de clustering

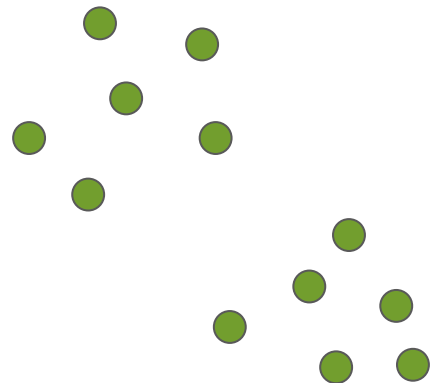
KMeans

1. **Paso de inicialización:** escogemos un número K de clusters deseados e iniciamos los centroides aleatoriamente
2. **Paso de asignación:** Asigna cada punto de datos al cluster cuyo centroide esté más cerca. La distancia se mide según alguna métrica (Euclidiana, por ejemplo).
3. **Paso de actualización:** Calcula los nuevos centroides de los clusters, con los puntos asignados a él
4. **Iteración:** Repite los pasos de asignación y actualización hasta que se cumpla algún criterio de convergencia.
5. **Detención:** Los centroides finales representan los puntos centrales de los clusters. Los puntos de datos se agrupan en K clusters basados en la asignación final de los centroides.

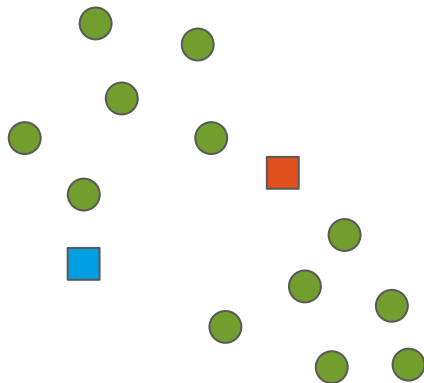
Algoritmos de clustering

KMeans

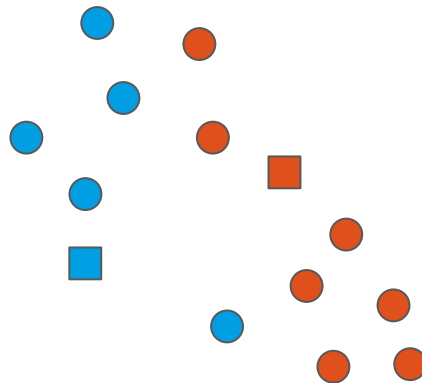
K = 2



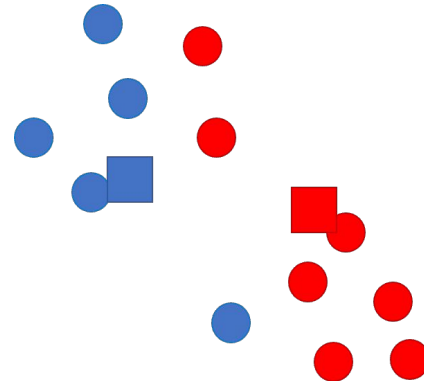
Centroides



Agrupamos



Recalculamos



Kmeans

Ventajas y desventajas

Ventajas

1. **Simplicidad:** K-means es un algoritmo simple y fácil de implementar.
2. **Eficiencia:** Es computacionalmente eficiente y puede manejar grandes conjuntos de datos.
3. **Escalabilidad:** Es adecuado para conjuntos de datos con un número significativo de dimensiones y características.



Kmeans

Ventajas y desventajas

Desventajas

1. **Sensibilidad a la inicialización:** Los resultados del clustering pueden variar dependiendo de la inicialización de los centroides.
2. **Requiere especificar el número de clusters (K) de antemano:** Es necesario conocer o estimar el número óptimo de clusters antes de aplicar el algoritmo.
3. **Sensible a los datos atípicos y ruido:** Los puntos de datos anómalos o ruidosos pueden afectar la formación de clusters y la asignación incorrecta de puntos.



¿Cómo podríamos estimar el número de cluster para entregar como input al algoritmo?

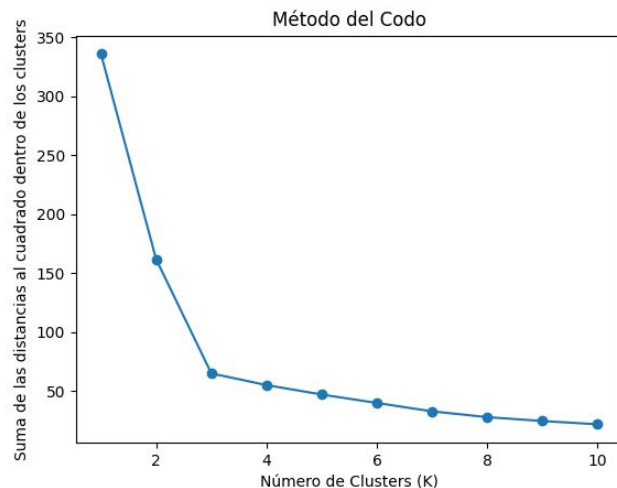


`/* Método del codo*/`

Método del codo

Se utiliza para determinar el número óptimo de clusters en el algoritmo de clustering. La idea principal es identificar el valor de K para el que se produce un cambio significativo en la variabilidad explicada por los clusters.

La idea es ejecutar el algoritmo para diferentes número de clusters, y graficar cómo varía la distancia intracluster. En la figura, se observa que para 3 clusters se produce un “codo”.



**Manos a la obra
¡A generar clusters!**



Ejercicio

¡A generar clusters!

Veremos ahora cómo podemos utilizar Python para aplicar KMeans. Para ello, observa los pasos que te mostrará tu profesor en Jupyter Notebook. Puedes abrir tú también un archivo para reproducir los pasos.

Aprenderemos a:

1. Determinar la cantidad óptima de clusters por el método del codo
2. Determinar los clusters usando KMeans
3. Validar el método con indicadores numéricos



/* Fuzzy C-means */

Fuzzy c means

¿En qué consiste?

El algoritmo Fuzzy C-means (FCM) es una técnica de clustering que permite asignar **grados de pertenencia** a los puntos de datos en lugar de asignarlos a clusters de forma binaria como en el algoritmo K-means. Así, un punto puede pertenecer a múltiples clusters con diferentes grados de pertenencia.

Este algoritmo requiere la especificación de un **parámetro de fuzziness**, que controla la difusión o borrosidad de las asignaciones de membresía. Valores más altos de este parámetro hacen que las asignaciones sean más difusas, permitiendo que los puntos de datos tengan grados de pertenencia más equilibrados entre los clusters.

Fuzzy C-means

Algoritmo

1. **Inicialización:** Se selecciona el número de clusters (K) y se inicializan aleatoriamente los centroides y los grados de pertenencia para cada punto de datos. (valores entre 0 y 1)
2. **Cálculo de los centroides:** Se calculan los centroides ponderando con los grados de pertenencia.
3. **Actualización de los grados de pertenencia:** Se actualizan los grados de pertenencia de cada punto de datos utilizando una función de pertenencia basada en la distancia a los centroides.
4. **Iteración:** se iteran los pasos anteriores, como con KMean
5. **Resultados:** Al finalizar las iteraciones, se obtienen los centroides finales y los grados de pertenencia de los puntos de datos.

Fuzzy C-means

Pertenencia

La fórmula de actualización de grados de pertenencia es

$$u_{ij} = \left(\sum_{k=1}^c \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}} \right)^{-1}$$

donde:

- u_{ij} es el grado de pertenencia del punto i al cluster j
- d_{ij} es la distancia del punto i al cluster j
- m es el **parámetro de fuzziness**. Para él se tiene que:
 - Si $m = 1$ la pertenencia es binaria (KMeans)
 - Si $m > 1$, la pertenencia es difusa. Cuanto mayor sea el valor de m , más difusas serán las asignaciones.

Fuzzy c means

Ventajas y desventajas

Ventajas

1. **Flexibilidad en asignación de pertenencia:** FCM permite asignar grados de pertenencia difusos a los puntos de datos, lo que refleja la incertidumbre en la asignación a clusters
2. **Tolerancia al ruido y atipicidad:** FCM es menos sensible a los valores atípicos y al ruido en los datos, ya que utiliza grados de pertenencia en lugar de asignaciones binarias.
3. **Mayor información sobre similitudes:** FCM puede revelar información sobre la similitud relativa entre los puntos de datos.

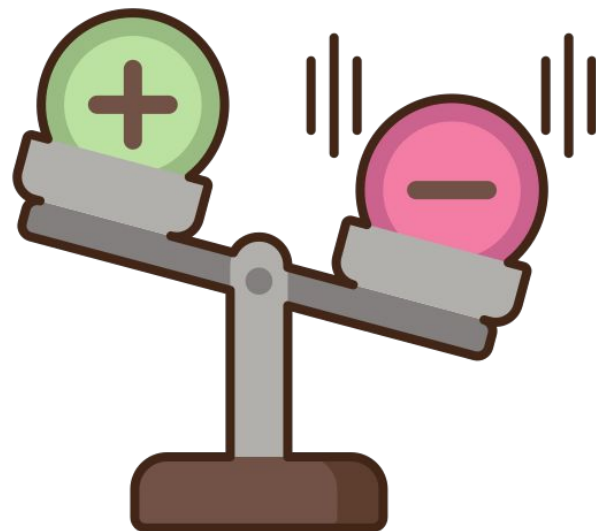


Fuzzy c means

Ventajas y desventajas

Desventajas

1. **Sensibilidad a la inicialización:** El rendimiento del algoritmo FCM puede depender en gran medida de la inicialización de los centroides y los grados de pertenencia.
2. **Mayor complejidad computacional:** El algoritmo FCM es más computacionalmente costoso que los algoritmos de clustering tradicionales, como K-means, debido a la necesidad de calcular y actualizar los grados de pertenencia en cada iteración.



/* Cluster Jerárquicos */

Cluster Jerárquicos

¿Qué son?

Corresponden a un método de agrupamiento que crea una jerarquía de clusters de manera recursiva.

Existen 2 grandes tipos y son:

1. **Aglomerativos:** Comienzan con cada punto de los datos como un cluster individual, y se van fusionando los cluster hasta tener un solo cluster con toda la data.
2. **Divisivo:** Comienzan con todos los puntos en un solo cluster, y van paso a paso dividiendo en cluster pequeños hasta que cada punto sea un cluster.

Cluster Jerárquicos

Criterios de enlace

1. **Simple Linkage:** La menor distancia entre 2 puntos en los grupos.

$$d_{single}(G, H) = \min_{i \in G, j \in H} (d_{ij})$$

2. **Complete Linkage:** La mayor distancia entre 2 puntos en los grupos.

$$d_{complete}(G, H) = \max_{i \in G, j \in H} (d_{ij})$$

3. **Average Linkage:** La distancia promedio entre todos los puntos del grupo opuesto.

$$d_{average}(G, H) = \frac{1}{n_G n_H} \sum_{i \in G, j \in H} d_{ij}$$

4. **Enlace de Ward (Ward's linkage):** Minimiza la suma de las diferencias cuadradas dentro del cluster al fusionar dos clusters.

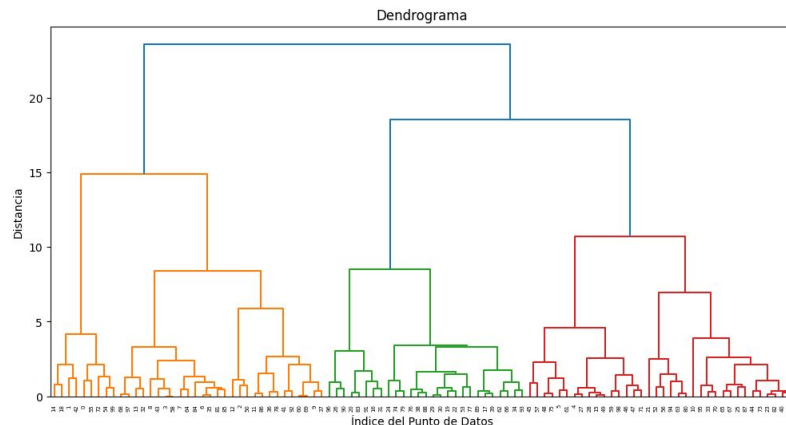
$$d_{Ward}(G, H) = |G - H|^2$$

Cluster Jerárquicos

Dendogramas

Un dendrograma es una representación visual de la estructura jerárquica de los clusters.

Los puntos de datos se muestran en la parte inferior del dendrograma y se agrupan en clusters a medida que se mueven hacia arriba.



Cluster Jerárquicos

Ventajas y desventajas

Ventajas

1. No se necesita especificar el número de clusters de antemano.
2. Proporciona una visualización natural de la estructura de datos con los dendrogramas.
3. Permite identificar diferentes niveles de clusters.
4. No depende de suposiciones sobre la forma de los clusters.

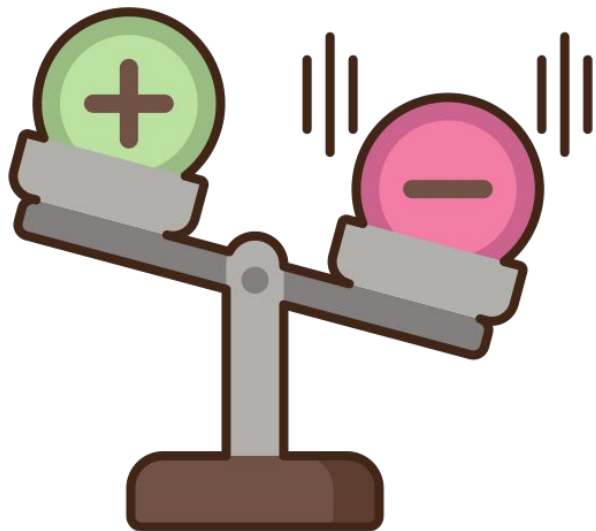


Cluster Jerárquicos

Ventajas y desventajas

Desventajas

1. Requiere más tiempo de cómputo en comparación con otros algoritmos de clustering, especialmente cuando se trata de conjuntos de datos grandes.
2. La complejidad del dendrograma puede dificultar la interpretación, especialmente en conjuntos de datos con muchos puntos.



Desafío - Segmentación de clientes



Desafío

“Segmentación de clientes”

- Descarga el archivo “Desafío”.
- Tiempo de desarrollo asincrónico: desde 4 horas.
- Tipo de desafío: individual.

¡AHORA TE TOCA A TI! 💪



Ideas fuerza



Clustering es una **técnica de análisis de datos** en el aprendizaje no supervisado que tiene como objetivo **agrupar objetos similares** en conjuntos llamados **clusters**.



Algunos algoritmos de clustering son **K Means**, **Fuzzy C Means** y **Cluster Jerárquicos**. Cada uno tiene sus ventajas y desventajas propias.



Existen diferentes **métodos de validación** de algoritmos de clustering, a través de **métricas de calidad** y **criterios cualitativos** a utilizar.

¿Qué importancia tienen los algoritmos de clustering en las diferentes industrias?



Recursos asincrónicos

¡No olvides revisarlos!

Para esta semana deberás revisar:

- Guía de estudio
- Desafío “Segmentación de clientes”





Próxima sesión...

- *Aprendizaje Supervisado*
- *Técnicas de clasificación*
- *Evaluación de modelos de clasificación.*

{desafío}
latam_

*Academia de
talentos digitales*

