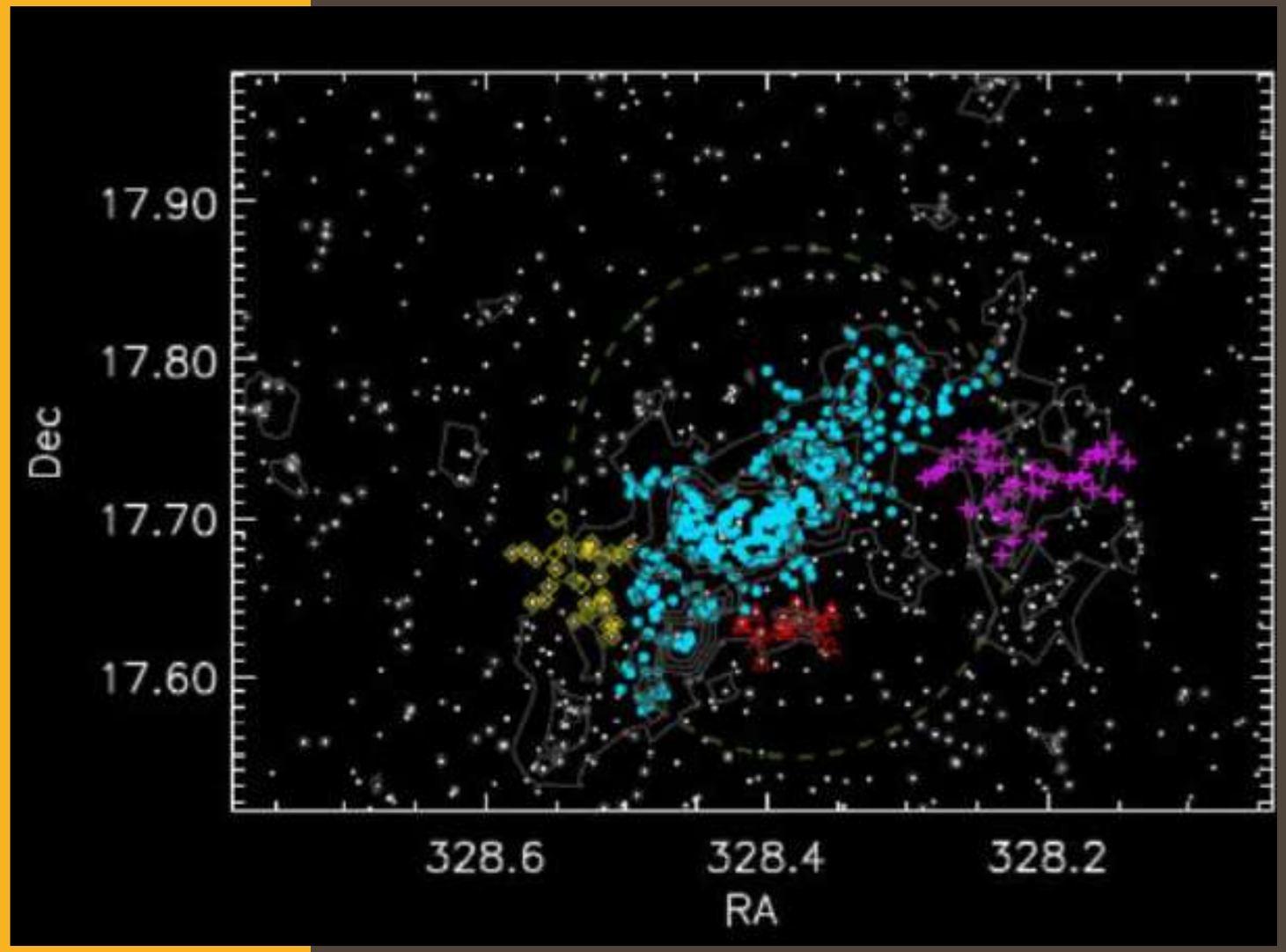
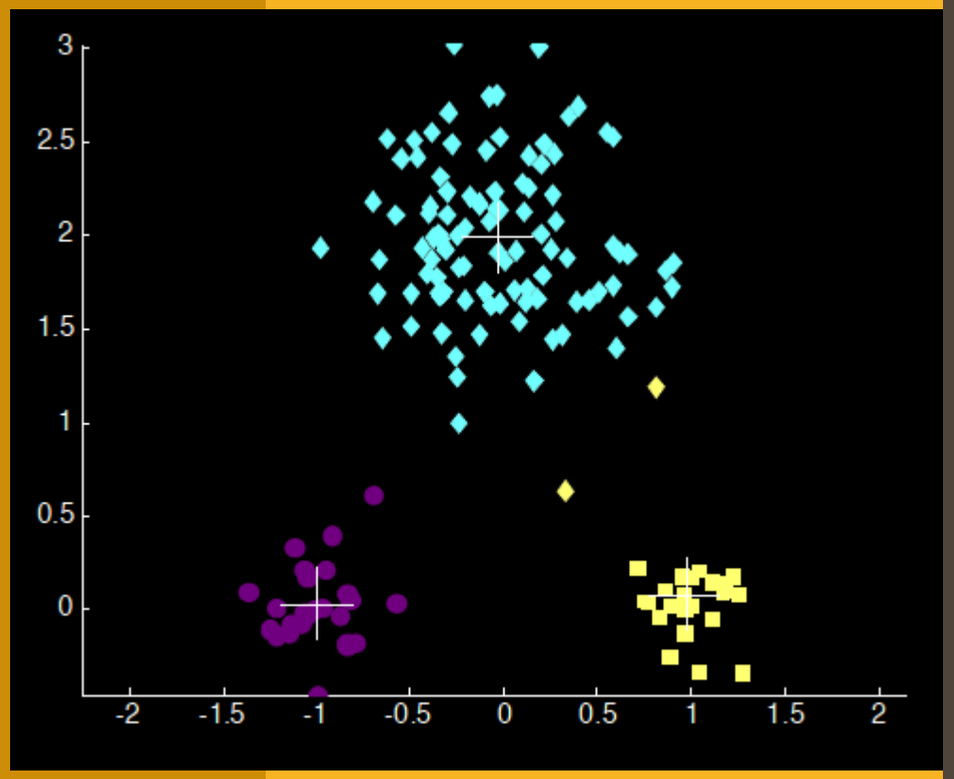
TÉCNICA DE ANÁLISIS EXPLORATORIO Y NO SUPERVISADO

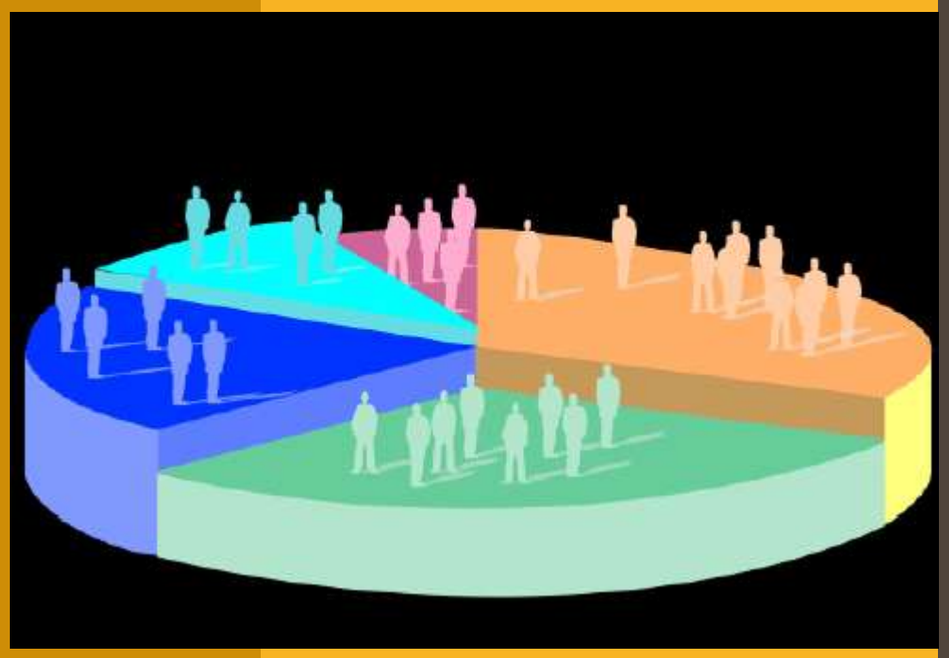


"CLUSTERING"

MUNDO NO SUPERVISADO

INTRODUCCIÓN





- El término clustering hace referencia a un amplio abanico de técnicas unsupervised cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones. Las particiones se establecen de forma que las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método unsupervised ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable). Esta característica diferencia al clustering de las técnicas estadísticas conocidas como análisis discriminante, que emplean un set de entrenamiento en el que se conoce la verdadera clasificación.

¿QUÉ SON LAS TÉCNICAS DE SEGMENTACIÓN?

Son técnicas que permiten resumir los datos. Se puede realizar un resumen global o específico de ciertas variables.

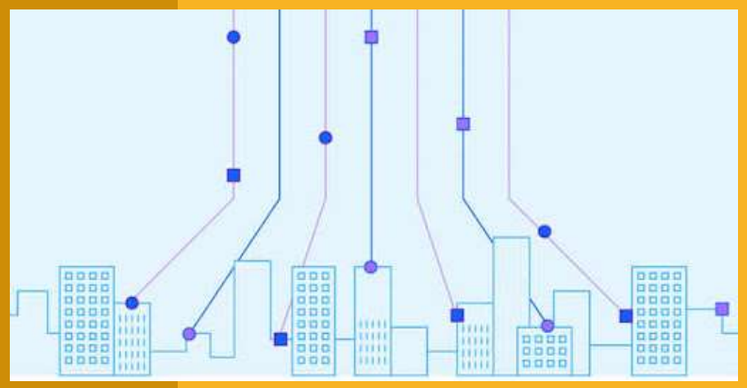
En general existen tres tipos de enfoques:

- Estimación de densidad: determina una representación compacta de la distribución de probabilidad sobre todos los datos $P(X)=P(X1,X2,\ldots,Xp)$.

- Búsqueda de patrones: busca una asociación descriptiva entre las variables.

- Clustering: separa las instancias en grupos de datos con características similares.

¿PARA QUE CLUSTERIZAR?



- BI - Hacer inteligencia de negocio dando uso inteligente de TODA la información disponible.
- Generar grupos o Clusters de Sujetos X idénticos o similares entre ellos (“gemelos estadísticos”), para compararlos para distintos fines.
- Caracterización.
- Minería de datos.
- Para Transitar desde el Mundo No Supervisado al Mundo Supervisado.

¿PARA QUE CLUSTERIZAR?

DEJAR QUE LOS DATOS HABLEN

ENTENDER

NO SESGAR

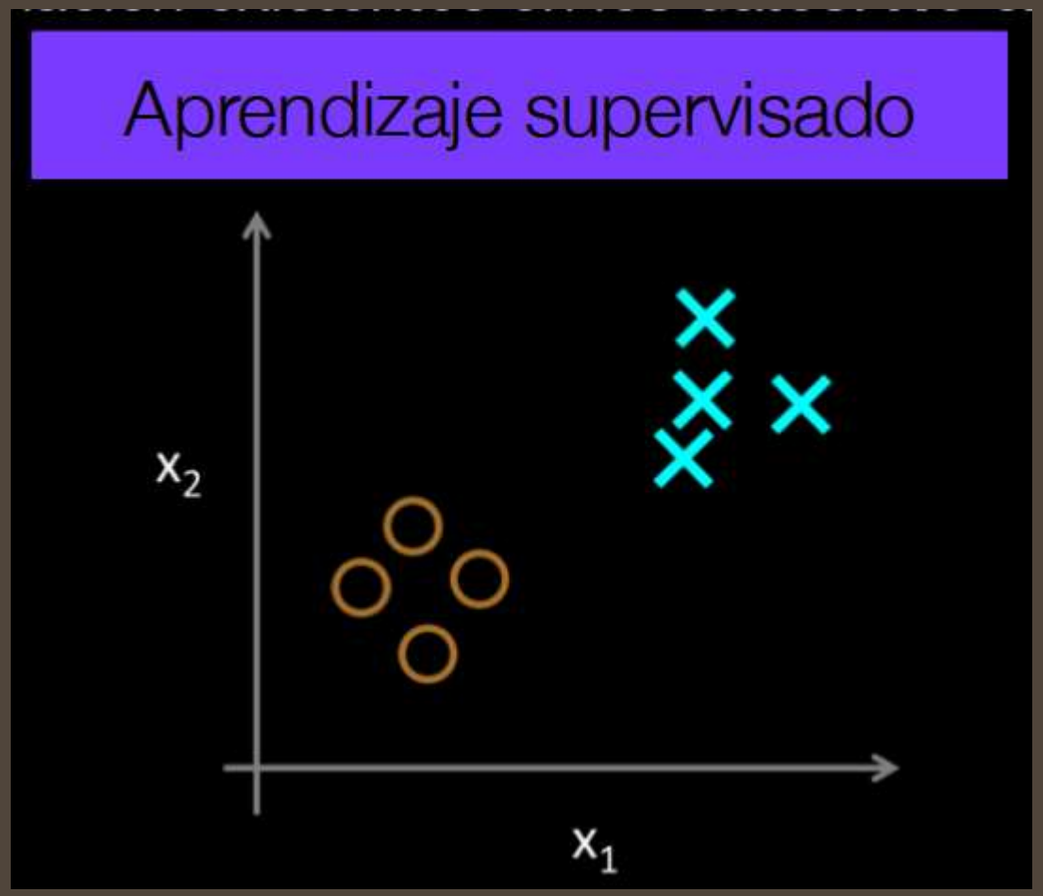
SESGOS COGNITIVOS

NO SESGAR

En el aprendizaje supervisado existe una variable objetivo a predecir.

• Si el objetivo son clases, se llama problema de clasificación o análisis discriminate.

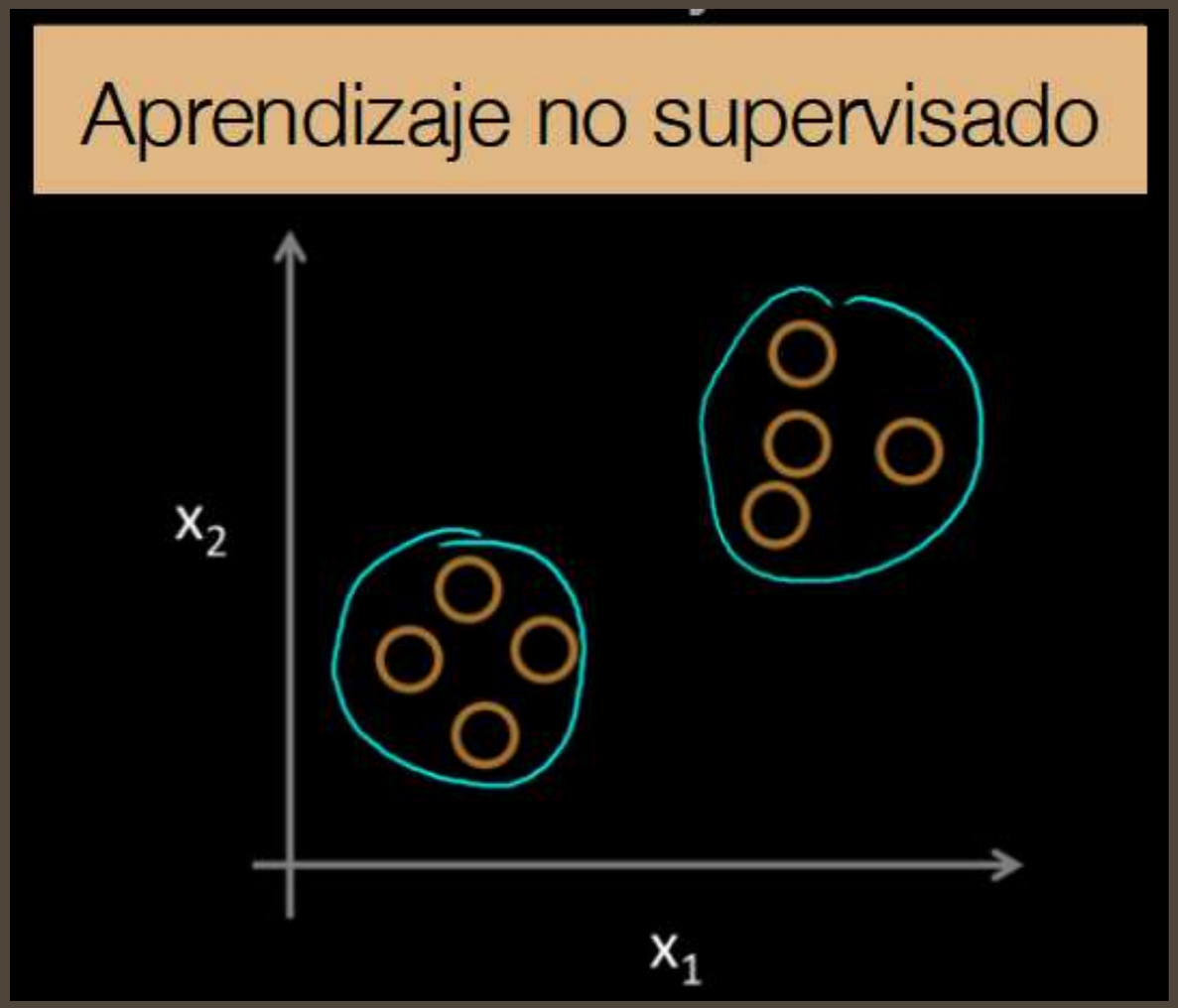
• Si los datos son continuos, se denomina problema de regresión.



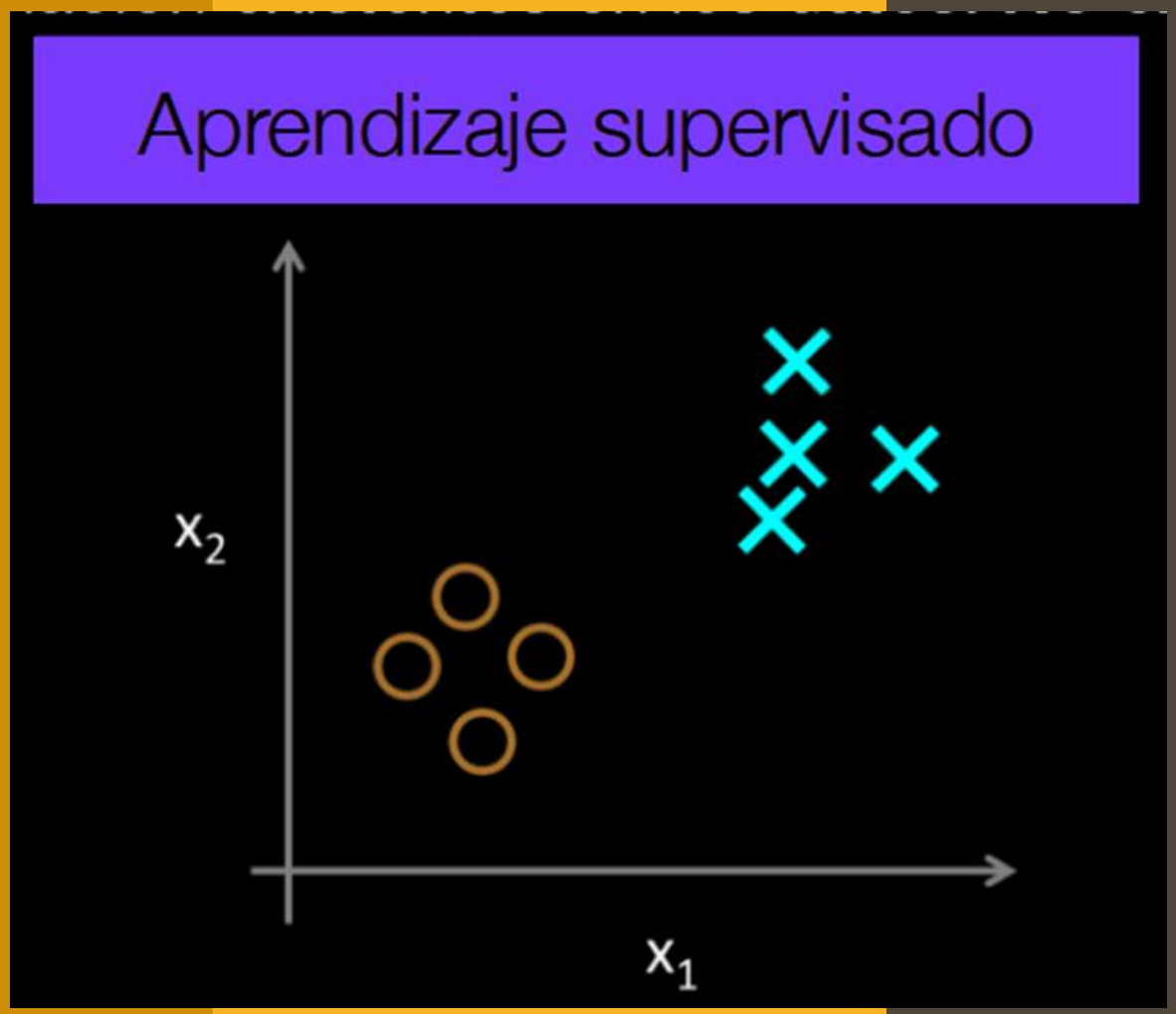
NO SESGAR

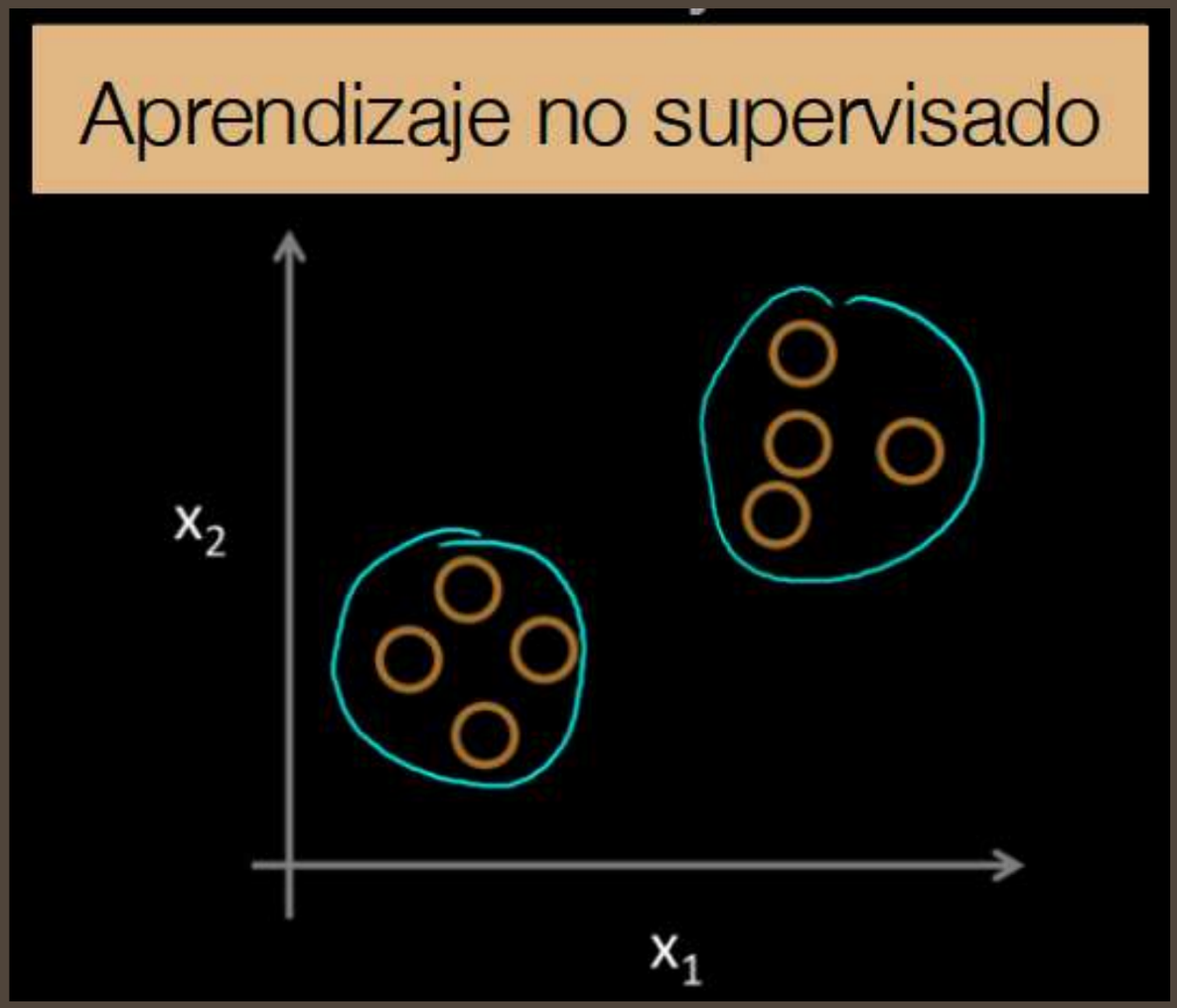
En el aprendizaje NO supervisado se busca observar, describir y también aprender la estructura o relación existentes en los datos.

No existe una variable objetivo.



NO SESGAR


Aprendizaje supervisado
x_2
x_1


Aprendizaje no supervisado
x_2
x_1

TIPOS DE PROBLEMAS

El tipo de problema o enfoque S o NS esta basado en el objetivo de la persona que analiza la tarea. Ejemplos:

• De un set de datos con etiqueta, cree un modelo capaz de predecir si un corredor de bolsa realizará algún fraude en el futuro cercano.

• Dado un set de datos sin etiqueta, agrupe a los corredores de bolsas en

grupos de personas homogéneas basado en su información demográfica

IDEA INTUITIVA

QUE BUSCAN LAS DISTINTAS TÉNICAS,MODELOS Y/O ALGORITMOSDE CLUSTERING?

MEDIDAS DE DISTANCIA

Todos los métodos de clustering tienen una denominador común; para poder llevar a cabo las agrupaciones necesitan definir “una Medida de Distancia”, y luego cuantificarla, con el objetivo de cuantificar la similitud o diferencia entre las observaciones (puntos en el plano cartesiano).

MEDIDAS DE DISTANCIA

El término distancia se emplea entonces dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión.

¿Podemos Utilizar Variables Dummy?

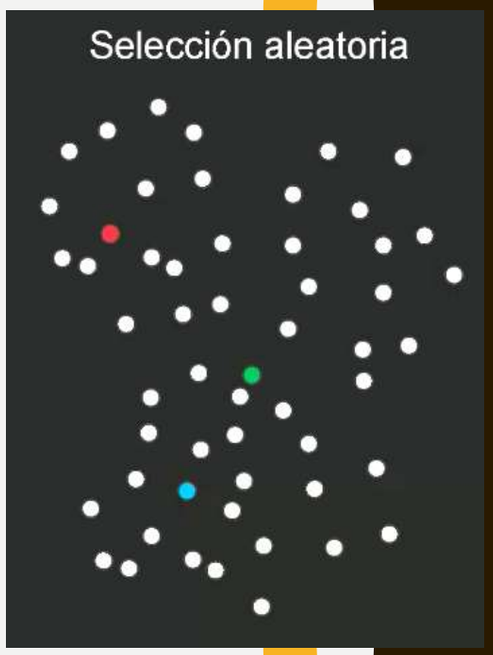
OBJETIVO DE TODO MODELO DE Clustering

MEDIDAS DE DISTANCIA

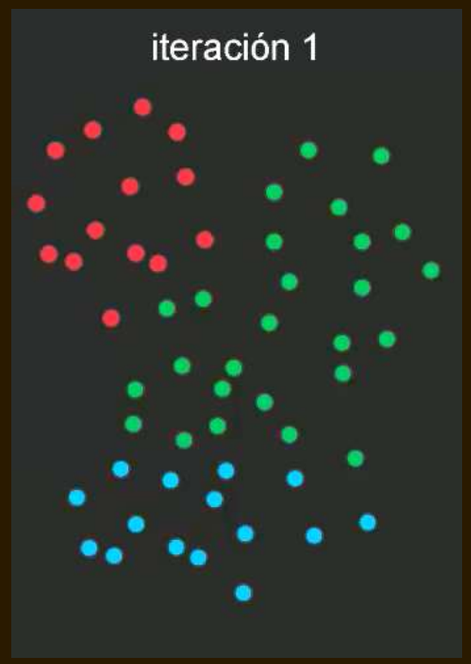
El objetivo del algoritmo es minimizar la distancia de los puntos dentro de cada cluster y maximizar la distancia entre Clusters.

Gráficamente

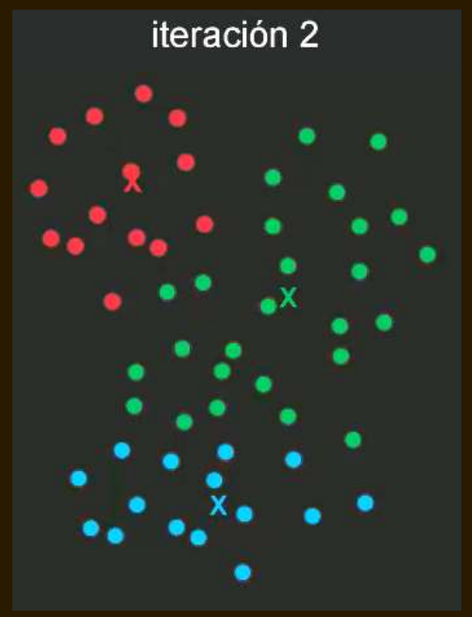
1. Si mi óptimo de Cluster (K) fue 3, el algoritmo seleccionará de manera aleatoria 3 observaciones.



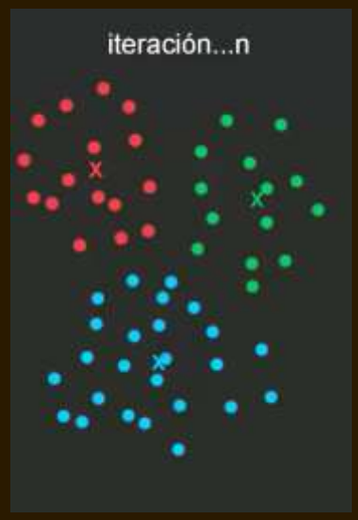
2. Luego, en una primera iteración, tomará las siguientes observaciones y las asignará al punto más cercano de las 3 observaciones anteriores (proto centroides).



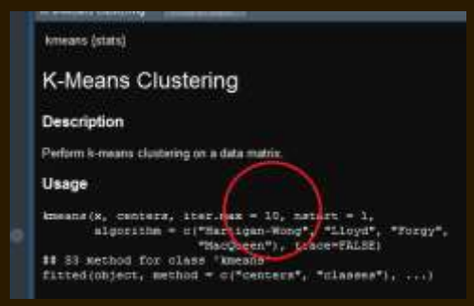
3. En una segunda iteración vuelve a calculará los centroides definitivos.



4. En una tercera iteración, vuelve a asignar las observaciones a los centroides definidos en el paso 3.



5. El algoritmo en Rstudio itera 10 veces.



ALGORITMOS

Dada la popularidad del clustering en disciplinas muy distintas (genómica, marketing...), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos. Pueden diferenciarse tres grupos:

- Partitioning Clustering: Este tipo de algoritmos requieren que el usuario especifique de antemano el número de clusters que se van a crear (K-means, K-medoids, CLARA).

- Hierarchical Clustering (Jerárquico): Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clusters. (agglomerative clustering, divisive clusterig).

- Métodos que combinan o modifican los anteriores (hierarchical K-means, fuzzy clustering, model based clustering y density based clustering).

CONSIDERACIONES PARA LA CORRELACION PARA

- Mundo Supervisado. 2 Caminos o 2 Objetivos:

  Predecir o

  Explicar

- Mundo No Supervisado. 2 Caminos o 2 Objetivos:

  Es Relevante que Influyan.

  No es Relevante que Influyan.

CONSIDERACIONES DE LA CORRELACION PARA…

MUNDO O UNIVERSO	CONSIDERACION	¿ES RELEVANTE LA CORRELACIÓN?
SUPERVISADO	Explicar	SI
SUPERVISADO	Predecir	SI
NO SUPERVISADO	Es Relevante que Influyan las Variables al objetivo del Análisis,	SI
NO SUPERVISADO	No es Relevante que Influyan las Variables al objetivo del Análisis: Muy Bien Justificado.	NO

TODO DEPENDERÁ DE LOS OBJETIVOS DEL ESTUDIO DE MINERÍA DE DATOS VERSUS LOS OBJETIVOS DE NEGOCIO

VENTAJAS Y DESVENTAJAS DE K-MEANS

K-means es uno de los métodos de clustering más utilizados. Destaca por la sencillez y velocidad de su algoritmo, sin embargo, presenta una serie de limitaciones que se deben tener en cuenta.

- Requiere que se indique de antemano el número de clusters que se van a crear. Esto puede ser complicado si no se dispone de información adicional sobre los datos con los que se trabaja. Una posible solución es aplicar el algoritmo para un rango de valores k y evaluar con cual se consiguen mejores resultados, por ejemplo, menor suma total de varianza interna.

- Presenta problemas de robustez frente a outliers. La única solución es excluirlos o recurrir a otros métodos de clustering más robustos como K-medoids (PAM).

- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. Para minimizar este problema se recomienda repetir el proceso de clustering entre 20 - 50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna. Aun así, no se garantiza que para un mismo set de datos los resultados sean exactamente iguales.

VENTAJAS Y DESVENTAJAS DE K-MEDOIS (PAM)

K-medoids es un método de clustering muy similar a K-means en cuanto a que ambos agrupan las observaciones en K clusters, donde K es un valor preestablecido por el analista. La diferencia es que en K-medoids cada cluster está representado por una observación presente en el cluster (medoid), mientras que en K-means cada cluster está representado por su centroide, que se corresponde con el promedio de todas las observaciones del cluster pero con ninguna en particular.

Una definición más exacta del término medoid es: elemento dentro de un cluster cuya distancia (diferencia) promedio entre él y todos los demás elementos del mismo cluster es lo menor posible. Se corresponde con el elemento más central del cluster y por lo tanto puede considerarse como el más representativo. El hecho de utilizar medoids en lugar de centroides hace de K-medoids un método más robusto que K-means, viéndose menos afectado por outliers o ruido. A modo de idea intuitiva puede considerarse como la analogía entre media y mediana.

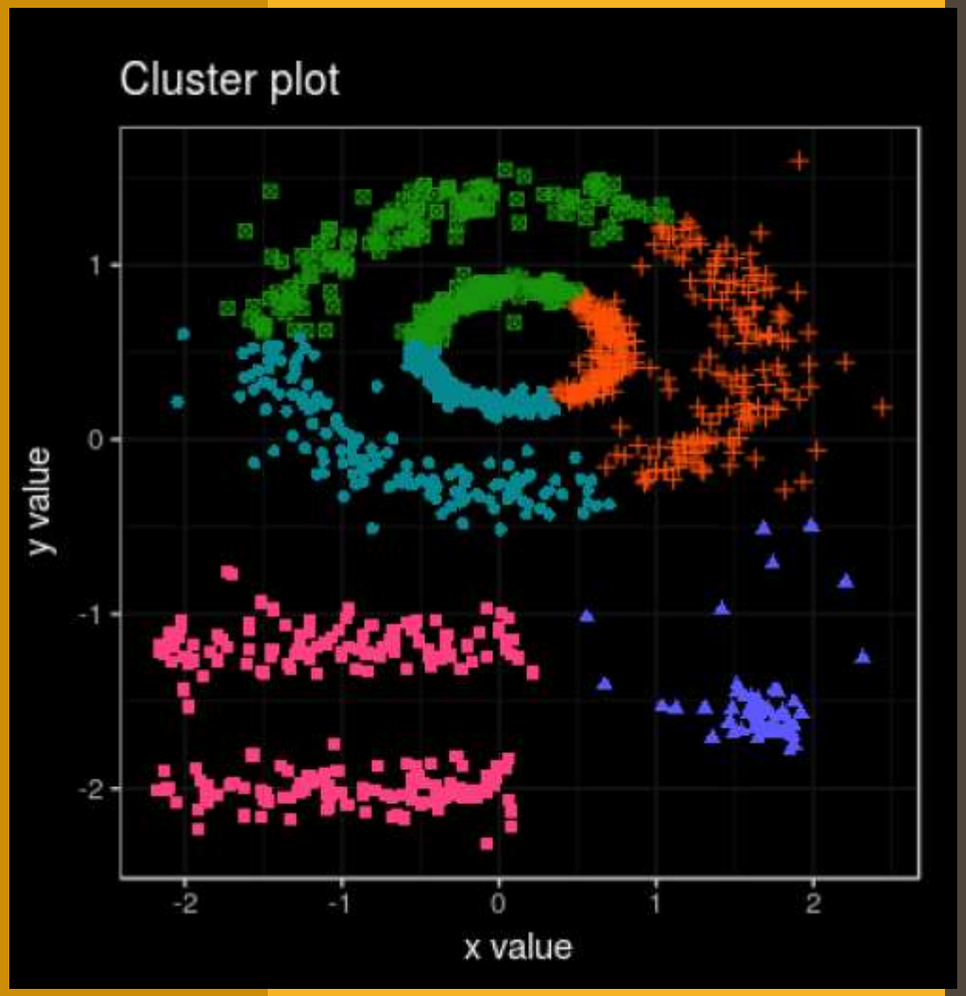
VENTAJAS Y DESVENTAJAS DE K-MEDOIS (PAM)

K-medoids es un método de clustering más robusto que K-means, por lo que es más adecuado cuando el set de datos contiene outliers o ruido.

- Al igual que K-means, necesita que se especifique de antemano el número de clusters que se van a crear. Esto puede ser complicado de determinar si no se dispone de información adicional sobre los datos.

- Para sets de datos grandes necesita muchos recursos computacionales. En tal situación se recomienda aplicar el método CLARA.

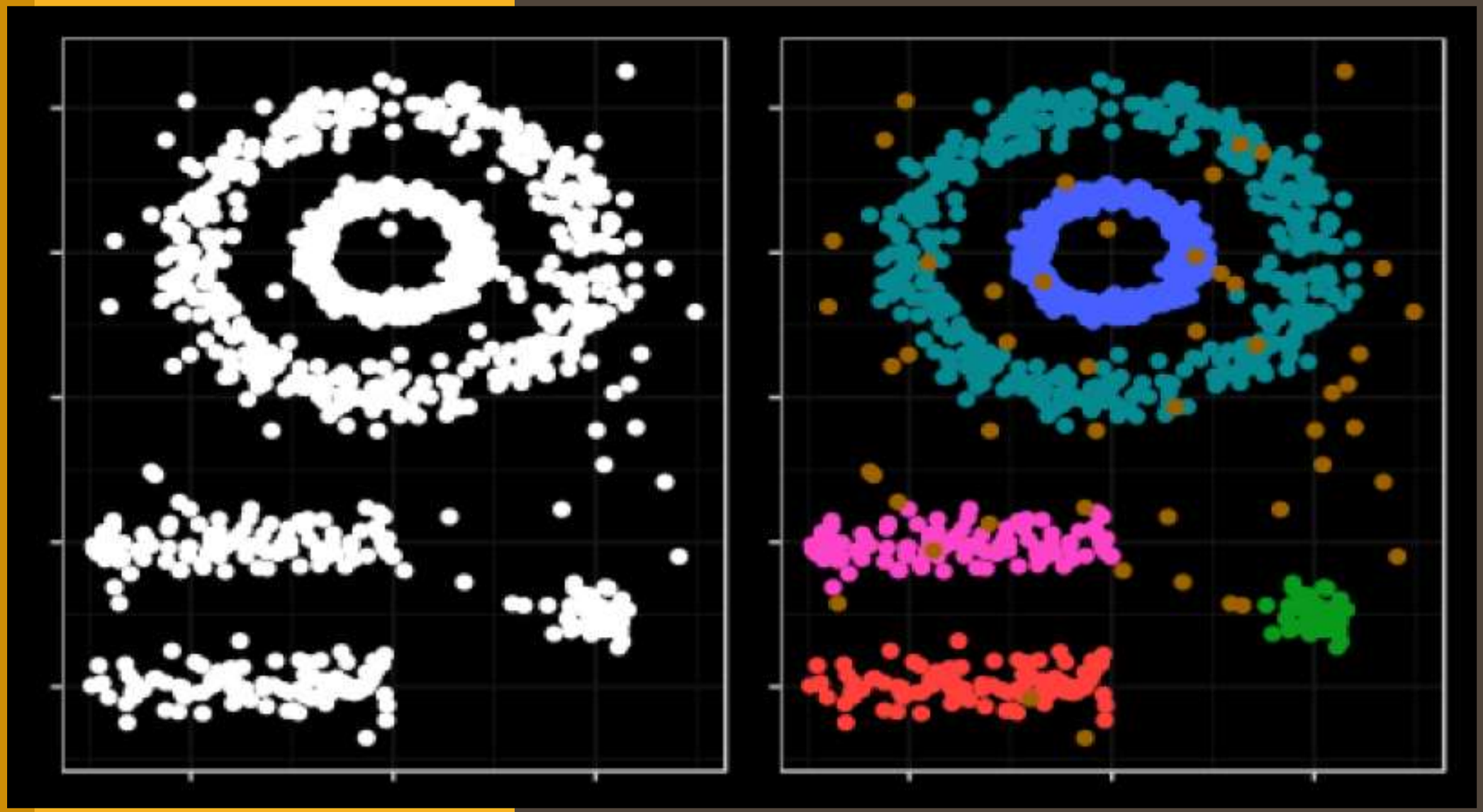
DENSITY BASED CLUSTERING (DBSCAN)



Density-based spatial clustering of applications with noise (DBSCAN) fue presentado en 1996 por Ester et al. como una forma de identificar clusters siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad.

DENSITY BASED CLUSTERING (DBSCAN)

El cerebro humano identifica fácilmente 5 agrupaciones y algunas observaciones aisladas (ruido).



VENTAJAS Y DESVENTAJAS DE BASED CLUSTERING (DBSCAN)

Ventajas de DBSCAN

- No requiere que el usuario especifique el número de clusters.
- Es independiente de la forma que tengan los clusters, no tienen por qué ser circulares.
- Puede identificar outliers, por lo que los clusters generados no se influenciados por ellos.

Desventajas de DBSCAN

- No es un método totalmente determinístico: los border points que son alcanzables desde más de un cluster pueden asignarse a uno u otro dependiendo del orden en el que se procesen los datos.
- No genera buenos resultados cuando la densidad de los grupos es muy distinta, ya que no es posible encontrar los parámetros ϵ y minPts que sirvan para todos a la vez.