FUNDACIÓN YPF

Módulo VI - Aprendizaje NO supervisado.

Clase 21: Clustering DBSCAN

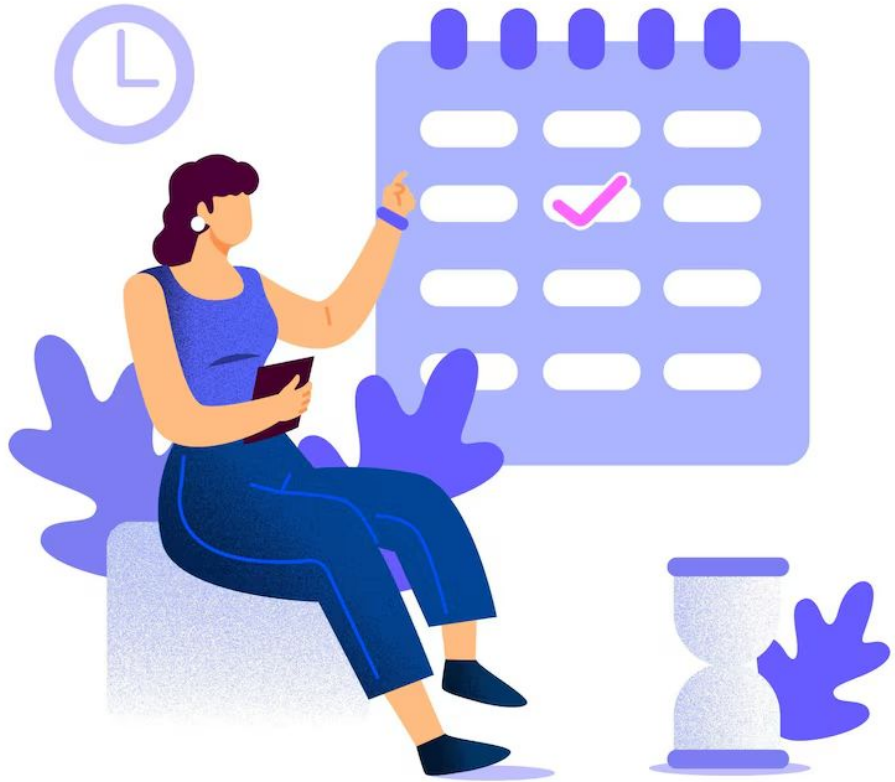
FUNDACIÓN YPF

¿Ponemos a grabar el taller?

FUNDACIÓN YPF

¿QUÉ VAMOS A VER HOY?

- CLUSTERING: DBSCAN



FUNDACIÓN YPF

REPASEMOS

FUNDACIÓN YPF

Eligiendo algoritmo

Tarea

Definir de forma clara el objetivo

1

Información

Con qué datos se cuenta para lograr el objetivo

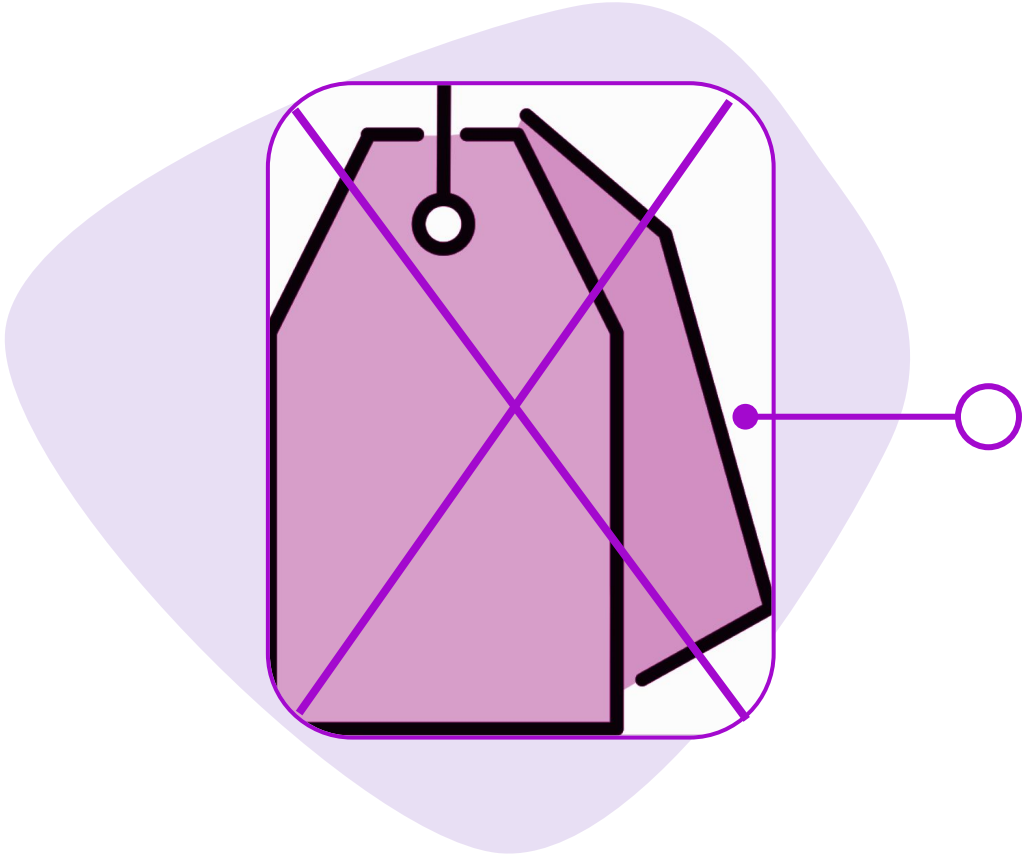
2

APRENDIZAJE **SUPERVISADO**

APRENDIZAJE **NO SUPERVISADO**

FUNDACIÓN YPF

Aprendizaje No Supervisado



El algoritmo infiere patrones de un conjunto de datos que, a diferencia del aprendizaje supervisado, **no** están **etiquetados**. Puede utilizarse para descubrir la estructura subyacente de los datos

FUNDACIÓN YPF

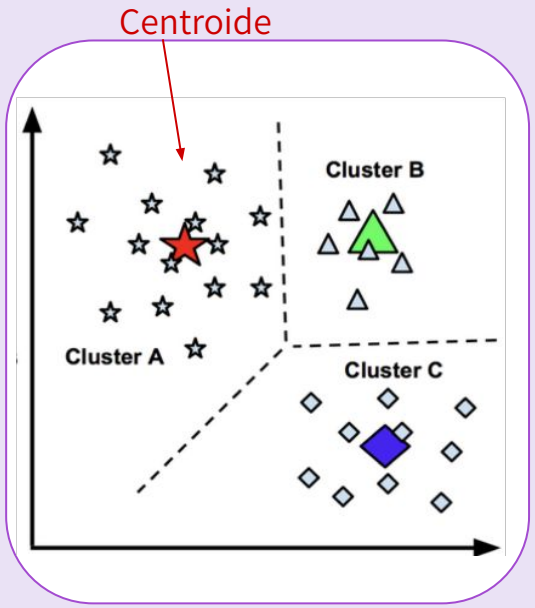
Clustering

El **objetivo** del clustering o agrupamiento es encontrar grupos (**clusters**) en los cuales las instancias pertenecientes sean parecidas.

- **Aplicaciones**
  - Investigación de mercado
  - Sistemas de recomendación
  - Medicina y Biología
- **Algoritmos**
  - K-means
  - DBSCAN
  - Hierarchical Clustering
  - Fuzzy C-Means
  - Gaussian Mixture Models

FUNDACIÓN YPF

K-Means



Encuentra un número **k** de centroides, uno por cada cluster, tal que la distancia entre los centroides y los datos más cercanos sea la **mínima posible**.

A continuación, cada instancia se identifica en el grupo del centroide más cercano

FUNDACIÓN YPF

DBSCAN

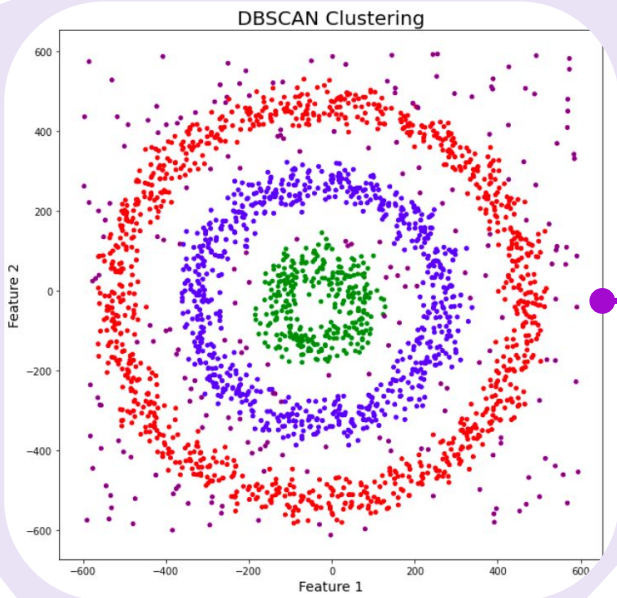
FUNDACIÓN YPF

DBSCAN

DBSCAN significa **Density-Based Spatial Clustering of Applications with Noise (agrupamiento espacial basado en densidad de aplicaciones con ruido).**
El algoritmo tiene como objetivo identificar un **número arbitrario** de clusters. Los clusters se definen por densidad de puntos. Puede haber puntos que no pertenezcan a ningún cluster.

FUNDACIÓN YPF

DBSCAN



Este algoritmo recorre todo el dataset e identifica aquellas zonas de puntos densamente pobladas como pertenecientes a un **mismo cluster**.

Los puntos aislados y que no pertenecen a **ningún cluster** serán considerados ruido u **outlier**.

FUNDACIÓN YPF

DBSCAN

A Se define una distancia **epsilon** como la vecindad de un punto. Se elige un número de puntos mínimos (**minPoints**) para considerar un cluster.

B Para cada punto del dataset:

- i Se selecciona un punto no visitado **random**. Se identifica si el punto tiene minPoints en su vecindario (punto core). Si no tiene, se lo llama noise. Se marca como visitado.
- ii Si es un punto core, se le asigna un nuevo cluster y todos los puntos de su vecindario se consideran dentro de su cluster. Si alguno de estos puntos también son cores, este proceso se repite. A los puntos asignados a un cluster que no son core, se los llama border. Todos se marcan como visitados.
- iii El proceso se repite hasta que todos los puntos hayan sido visitados

FUNDACIÓN YPF

Kmean VS. DBSCAN

Ventajas:

- Rápido
- No tiene parámetros
- Fácil asignar nuevas instancias

Desventajas

- Hay que definir el número de clusters
- Solo tiene buen desempeño con clusters tipo esferas
- Sensible a outliers

Ventajas:

- No hay que elegir el número de clusters
- Detecta cualquier forma de clusters
- Determina automáticamente datos outliers

Desventajas

- Hay que elegir bien los parámetros
- No tiene buen desempeño con clusters de diferentes densidades
- Es computacionalmente más costoso (

FUNDACIÓN YPF

Práctica:

Trabajamos con DBSCAN en la Notebook 21

FUNDACIÓN YPF

Trabajamos en salas

FUNDACIÓN YPF

Trabajamos en salas de zoom

DBSCAN

Trabajaremos con la Notebook 21

En los grupos establecidos, ejercitamos como se desarrolla un modelo de DBSCAN para Clustering

50 minutos de actividad

FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Repasamos dudas

Trabajamos con la Notebook 21

Revisamos los conceptos y el código trabajados en la notebook 21

FUNDACIÓN YPF

Desafío 15
(continuación)

Para la siguiente repasar la notebook 21

FUNDACIÓN YPF

¿Repaso?

FUNDACIÓN YPF

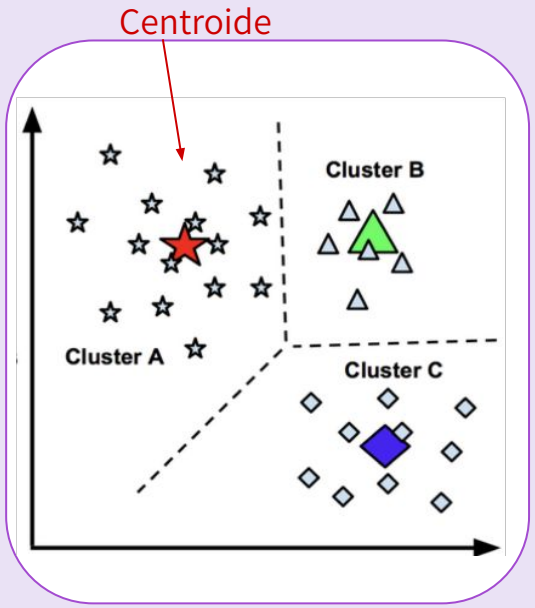
Clustering

El **objetivo** del clustering o agrupamiento es encontrar grupos (**clusters**) en los cuales las instancias pertenecientes sean parecidas.

- **Aplicaciones**
  - Investigación de mercado
  - Sistemas de recomendación
  - Medicina y Biología
- **Algoritmos**
  - K-means
  - DBSCAN
  - Hierarchical Clustering
  - Fuzzy C-Means
  - Gaussian Mixture Models

FUNDACIÓN YPF

K-Means

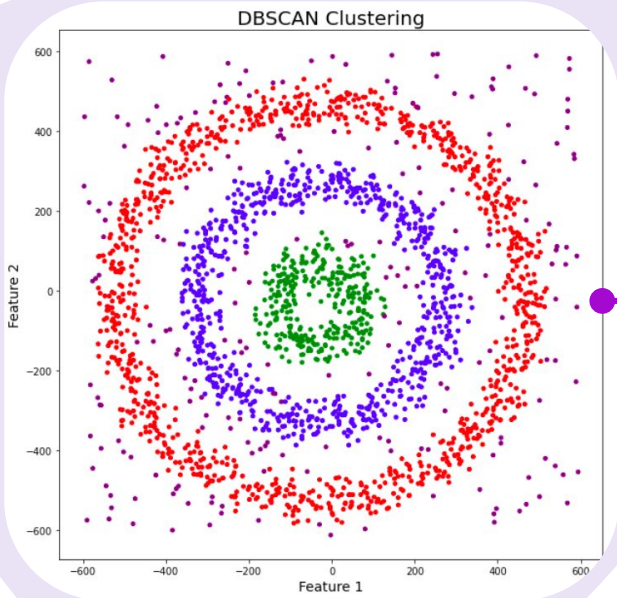


Encuentra un número **k** de centroides, uno por cada cluster, tal que la distancia entre los centroides y los datos más cercanos sea la **mínima posible**.

A continuación, cada instancia se identifica en el grupo del centroide más cercano

FUNDACIÓN YPF

DBSCAN



Este algoritmo recorre todo el dataset e identifica aquellas zonas de puntos densamente pobladas como pertenecientes a un **mismo cluster**.

Los puntos aislados y que no pertenecen a **ningún cluster** serán considerados ruido u **outlier**.

FUNDACIÓN YPF

¿Dudas?

FUNDACIÓN YPF

¡Muchas gracias!