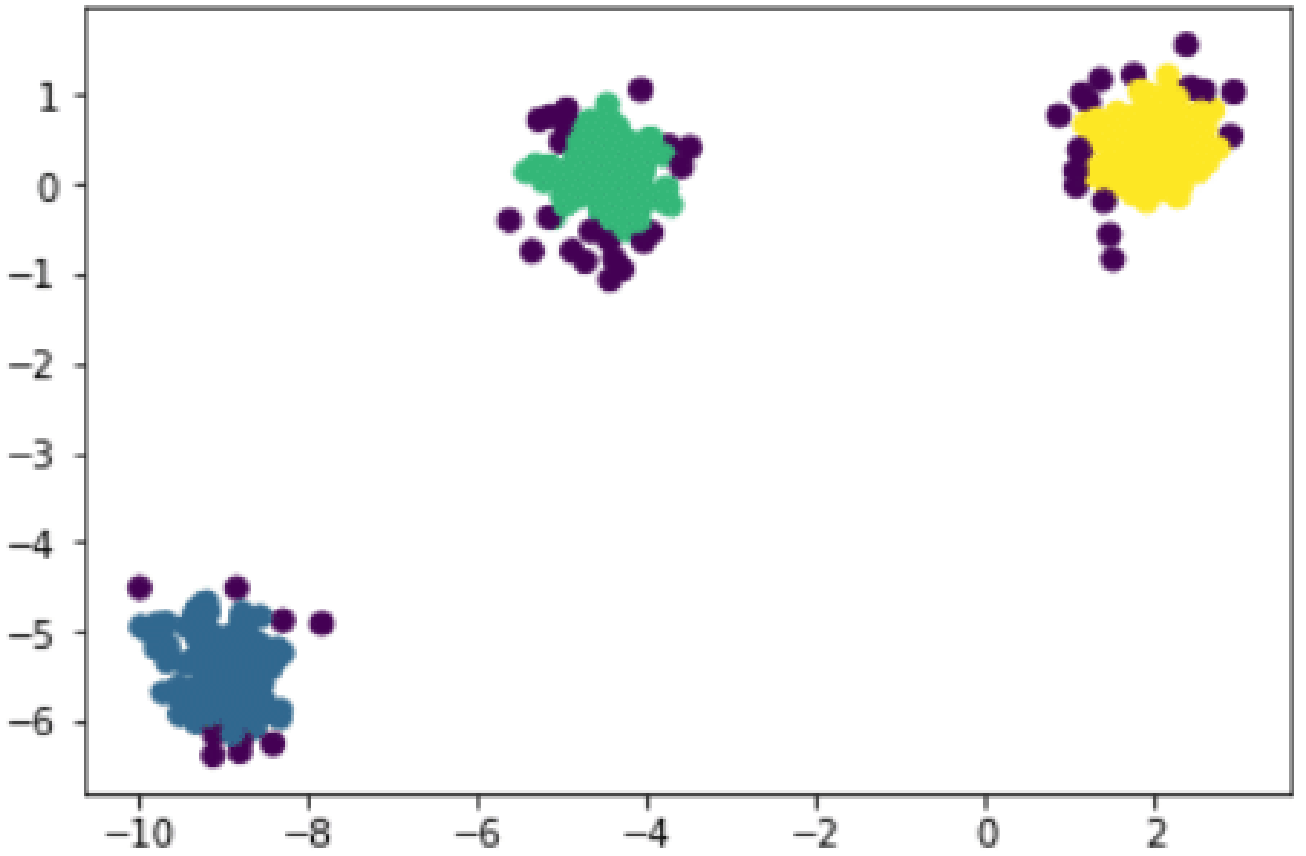


Cluster DBSCAN

PhD(e). Jonnatan Arias Garcia – jonnatan.arias@utp.edu.co – jariasg@uniquindio.edu.co

PhD. David Cardenas peña - dcardenasp@utp.edu.co

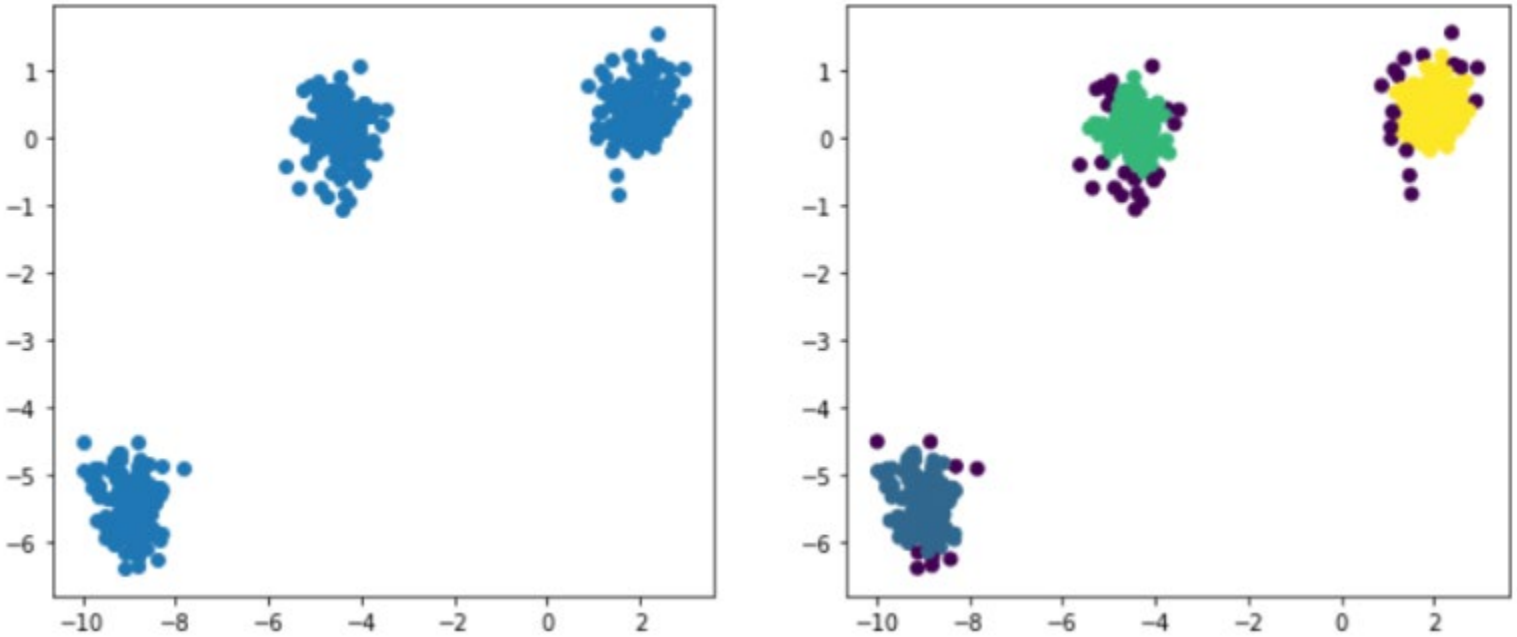
PhD. Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

DBSCAN

- Fue presentado en 1996 por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiawei Xu.
- Sus campos de aplicación son diversos: análisis cartográfico, análisis de datos, segmentación de una imagen, etc.

DBSCAN: Start

A partir de unos puntos y un número entero k, el algoritmo pretende **dividir los puntos en k grupos**, llamados **clústeres**, homogéneos y compactos.



DBSCAN: Start

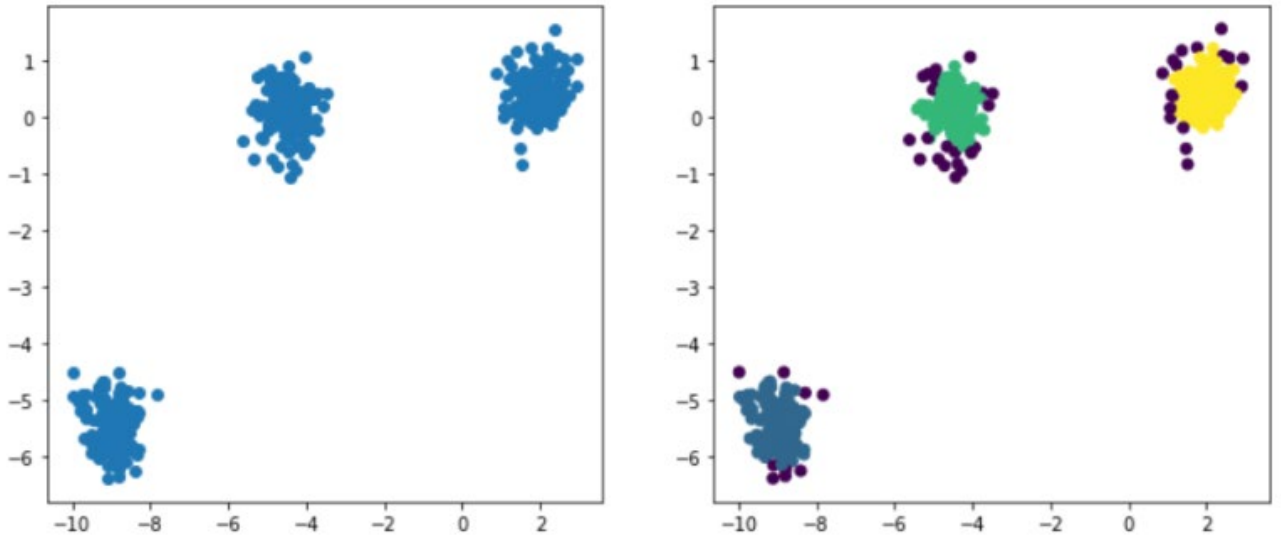
Es un algoritmo sencillo y se basa en la **estimación de la densidad local**.

1. Para cada observación miramos el número de puntos a una distancia máxima ε de ella. Esta zona se denomina **ε-vecindad de la observación**.
2. Si una observación tiene al menos un cierto número de vecinos, incluida ella misma, se considera una **observación central**. En este caso, se ha detectado una **observación de alta densidad**.
3. Todas las observaciones en la vecindad de una observación central pertenecen al mismo clúster. Puede haber observaciones centrales cercanas entre sí. **Por lo tanto, de un paso a otro, se obtiene una larga secuencia de observaciones centrales que constituyen un único clúster.**
4. Cualquier observación que no sea una observación central y que no tenga ninguna observación central en su vecindad se considera una **anomalía**.

- ¿Qué distancia **ε** hay que determinar para cada observación la **ε**-vecindad?
- ¿Cuál es el **número mínimo de vecinos necesario** para considerar una observación como una **observación central**?

No es necesario definir de antemano el número de clústeres, lo que hace que el algoritmo sea menos rígido.

DBSCAN es que también permite **gestionar los valores atípicos o las anomalías**.



El algoritmo ha determinado 3 clústeres principales: azul, verde y amarillo. Los puntos de color morado son anomalías detectadas por el DBSCAN. Obviamente, dependiendo del valor de **ε** y del número de vecinos mínimos, la partición puede variar.

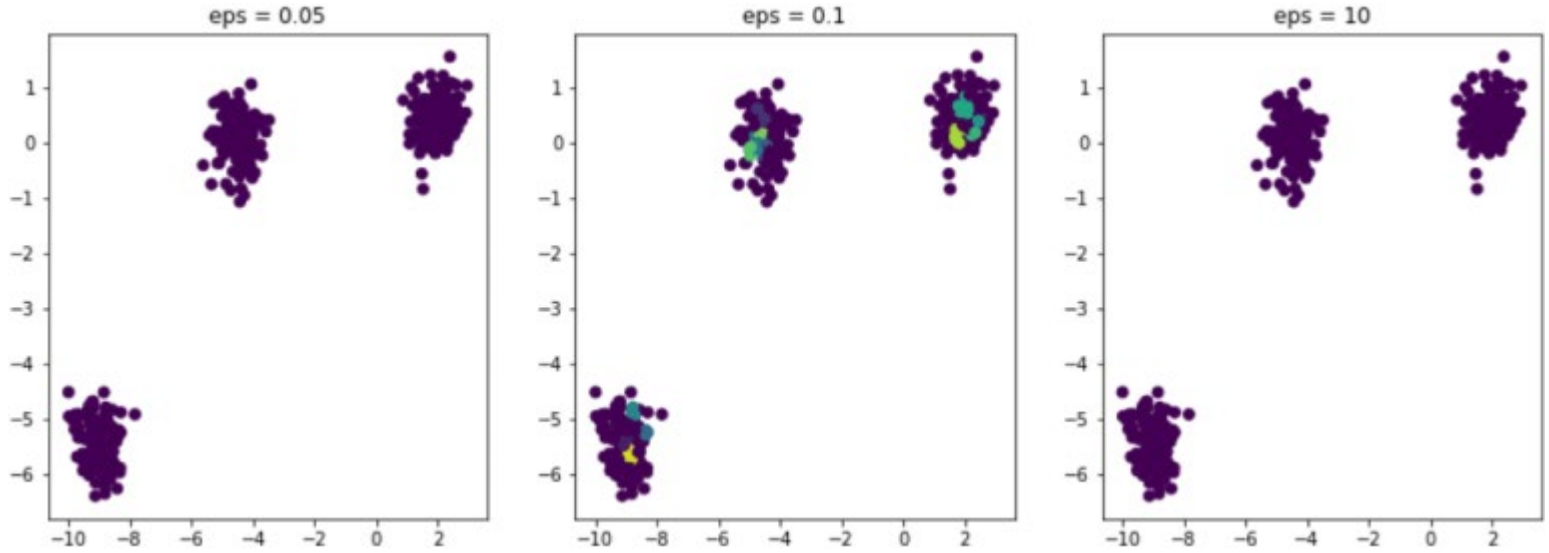
Distancia y **ε**

¿Cuál es la métrica utilizada para evaluar la distancia entre una observación y sus vecinos? ¿Cuál es la ε?

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Teniendo como máximo una distancia ε, calculamos la distancia euclidiana entre el vecino y la observación, despues comprobamos si es inferior a ε.

Distancia y **ε**



El numero de vecino fijado es 5.

si **ε** es demasiado pequeña, la **ε**-vecindad se considera como anomalías

si **ε** es muy grande, la **ε**-vecindad contendría todas las observaciones.

Para optimizar la **ε** se busca la distancia al vecino más cercano para cada observación.

Se fija **ε** que permita que una proporción grande de las observaciones. Entendemos el 90-95 % de las observaciones que deben tener al menos un vecino en su **ε**-vecindad.