Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos

# Density-based spatial clustering of applications with noise

*Marco Antonio Obregón Flores*

Profesor:
*José Alberto Benavides Vazquez*

18 de febrero de 2023

# 1 NearestNeighbors

El código utiliza la clase NearestNeighbors de scikit-learn para encontrar los k vecinos más cercanos para cada punto en el conjunto de datos. Luego, ordena las distancias de forma ascendente y calcula la curva de k-distancia. La curva de k-distancia muestra la distancia media al k-ésimo vecino más cercano para cada punto, ordenados de forma creciente.

El código busca la posición del "codo" en la curva de k-distancia, que corresponde a un punto donde el cambio en la distancia media comienza a disminuir drásticamente. Este punto indica el número de *minsamples* apropiado para los datos.

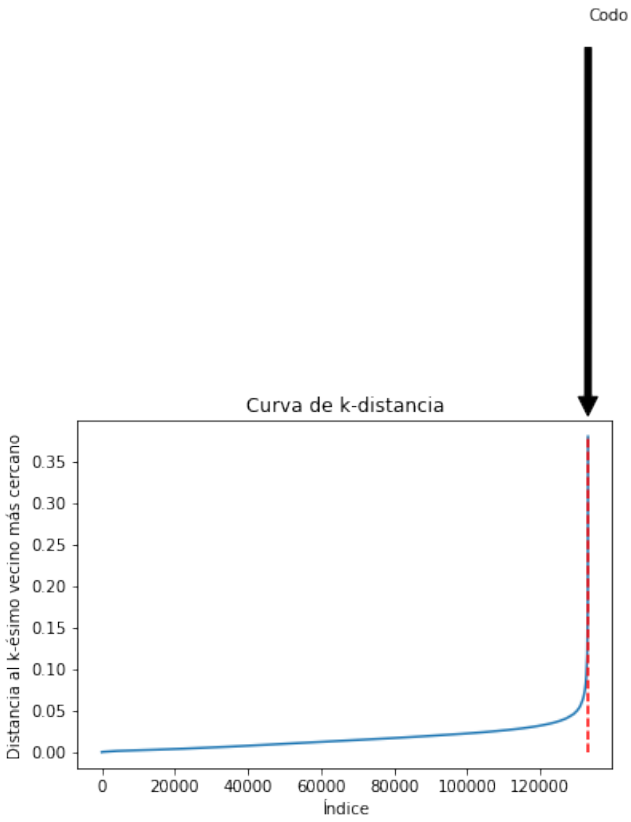


Figure 1: Representación gráfica de la curva de k-distancia

# 2 DBSCAN

Usando los resultados de NearestNeighbors, se aplicó el algoritmo DBSCAN con un valor de $\epsilon = 0.15$ y un valor de $min\_samples = 1$. El resultado fue la siguiente agrupación:

- Número de grupos encontrados: 26
- Número de puntos considerados como ruido: 0

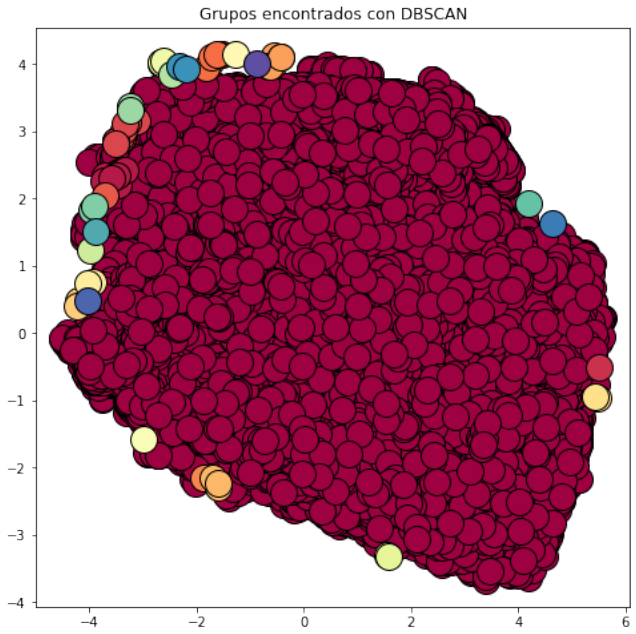


Figure 2: Representación gráfica de DBSCAN

De lo anterior, podemos observar que no todos los grupos encontrados son necesariamente buenos o útiles. En algunos casos, los grupos pueden ser un artefacto del ruido o de las fluctuaciones aleatorias en los datos, o pueden ser demasiado pequeños o poco significativos para ser útiles.

Es importante mencionar que al reducir la cantidad de datos y la dimensionalidad de los mismos, con el proposito de evitar el consumo excesivo de memoria, puede perderse información valiosa.

# 3 Conclusiones

En este trabajo se ha presentado un análisis detallado de los datos de un conjunto de imágenes médicas utilizando técnicas de aprendizaje automático. Se ha utilizado la técnica de reducción de dimensionalidad PCA para simplificar los datos y la técnica de clustering DBSCAN para identificar patrones en los datos.

Los resultados obtenidos muestran que la técnica de PCA es efectiva para reducir la dimensionalidad del conjunto de datos, lo que permite una mejor visualización y análisis. Además, se ha encontrado que la técnica de clustering DBSCAN es capaz de identificar patrones en los datos que no son visibles a simple vista.