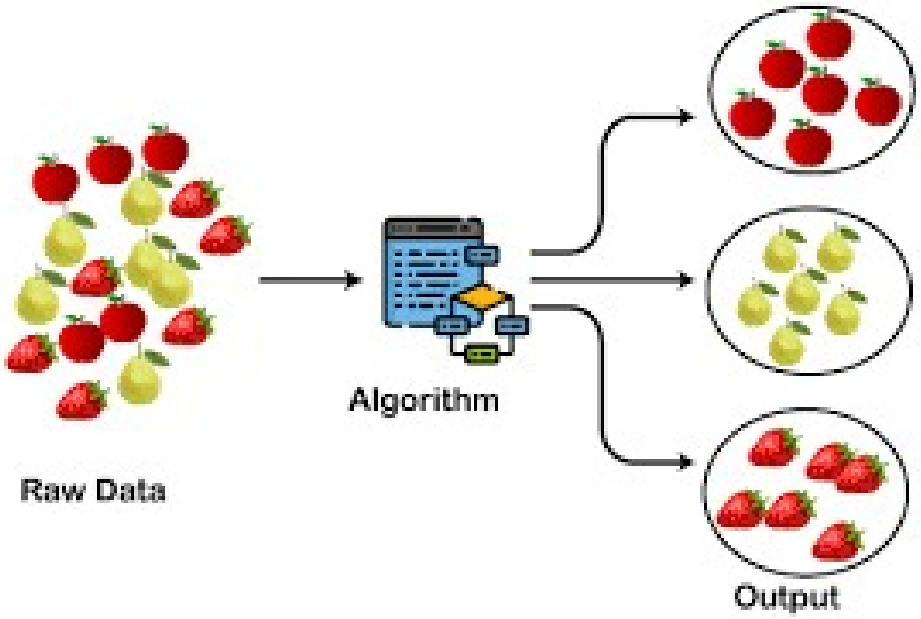
Clustering

Clustering

Es una técnica fundamental en el análisis de datos (aprendizaje no supervisado)



- Agrupa datos en subconjuntos similares
- Se basa en la similitud entre los datos

Aplicaciones Comunes del Clustering

- **Segmentación de Clientes**: Identificar grupos de clientes con comportamientos similares.
- **Detección de Anomalías**: Identificar puntos de datos que no pertenecen a ningún clúster.
- **Análisis de Imágenes**: Agrupar píxeles similares para segmentar imágenes.
- **Bioinformática**: Agrupar genes con funciones similares.

Particional

Divide los datos en **K** grupos distintos, donde **K** es un número predefinido de clústeres.

- **K-Means**: Es uno de los más populares. Asigna cada punto de datos al clúster cuyo centroide (media) está más cercano.

- **K-Medoids**: Utiliza puntos reales del conjunto de datos como representativos (medoids) en lugar de la media.

Jerárquico

Construye una jerarquía de clústeres, que puede ser :

- **Aglomerativo**: Comienza considerando cada punto de datos como un clúster individual y luego fusiona los más cercanos hasta formar una única estructura jerárquica.

- **Divisivo**: Parte de un único clúster que contiene todos los puntos y divide sucesivamente en clústeres más pequeños.

Basado en Densidad

Agrupa puntos que están en regiones densas del espacio de datos, separándolos de regiones menos densas.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Identifica clústeres de cualquier forma basándose en la densidad de puntos, y puede detectar outliers (puntos atípicos).

- **OPTICS (Ordering Points To Identify the Clustering Structure)**: Similar a DBSCAN pero puede manejar clústeres de diferentes densidades.

Comparación

	K-Means	Aglomerativo	DBSCAN	HDBSCAN
Tipo de clustering	Particional	Jerárquico	Basado en densidad	Basado en densidad (jerárquico)
Número de clusters	Debe especificarse (k)	No es necesario	No es necesario	No es necesario
Forma de los clusters	Esféricos	Flexible	Arbitraria	Arbitraria
Manejo de outliers	No maneja outliers	No maneja outliers	Identifica y elimina outliers	Identifica y elimina outliers
Complejidad computacional	Baja (O(n * k * i * d))	Alta ($O(n^2)$ o $O(n^3)$)	Media-Alta (depende den)	Media-Alta
Escalabilidad	Escala bien con grandes datasets	No escala bien	Escala moderadamente	Escala moderadamente
Parámetros principales	k	Métrica de enlace	eps,minPts	Sin parámetros críticos
Casos de uso ideal	Datos esféricos y uniformes	Datos jerárquicos	Datos con ruido y outliers	Datos con densidades variables

Métricas

- **Regla del codo:**

Indica que la reducción de la inercia (varianza intra-cluster) ya no es significativa después de k

- **Silhouette score**

Mide qué tan bien se agrupan los puntos dentro de un cluster y qué tan separados están de otros clusters.

Davies-Bouldin Index (DBI)

- Mide la compactación y separación de los clusters.

Métricas

- **Regla del codo:**

Indica que la reducci… …er) ya no es significativa desp…

- **Silhouette**

Mide qué tan … y qué tan separados est…

Davies-Bouldin Inde…

- Mide la compactación y separación de los clusters.

También sirven para elegir el número óptimo de clusters (cuando aplique)

Conclusiones

- Elegir el tipo de clustering adecuado depende de la naturaleza de los datos, el objetivo del análisis y las características específicas del problema a resolver.
- No hay un único mejor método
- Depende de la estructura de los datos y del objetivo
- Experimentar con diferentes enfoques es clave para buenos resultados