Aprendizaje no Supervisado

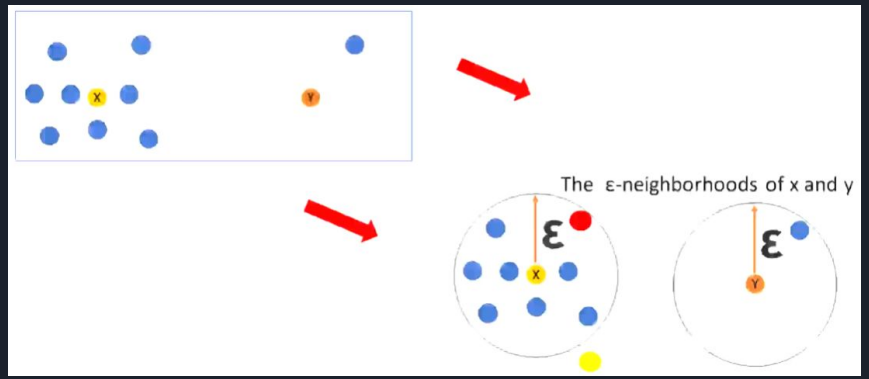
DBSCAN

Modelizado de Minería de Datos - Q22025

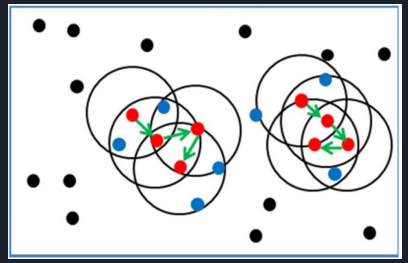
DBSCAN

DBSCAN significa Density Based Spatial Clustering of Application with Noise Fue propuesta por Martín Ester en 1996. DBSCAN es un algoritmo de agrupamiento basado en la densidad que funciona asumiendo que los agrupamientos son regiones densas en el espacio separadas por regiones de menor densidad.

- No requiere un “K”
- Devuelve el “K”



DBSCAN

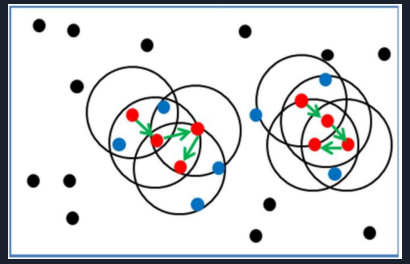


- A diferencia de K-means, DBSCAN no requiere que se especifique el número de clústeres de antemano
- Es capaz de descubrir clústeres de formas arbitrarias.

Conceptos clave

- **Puntos centrales**: Son puntos que tienen al menos un número mínimo de otros puntos (MinPts) dentro de una distancia determinada (ε o épsilon).
- **Puntos de borde**: Son puntos que están dentro de la distancia ε de un punto central, pero que no tienen suficientes vecinos MinPts para ser considerados puntos centrales.
- **Puntos de ruido**: Son puntos que no son ni puntos centrales ni puntos de borde. No están lo suficientemente cerca de ningún clúster como para ser incluidos.

DBSCAN



Hiper parámetros

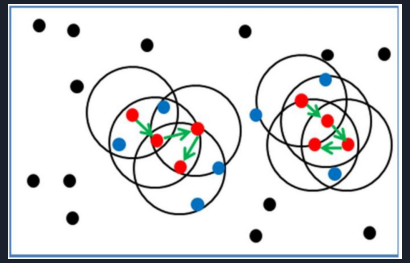
min_samples:

- Especifica el número mínimo de puntos necesarios para formar una región densa (un punto central).

eps:

- Este hiperparámetro define la distancia máxima entre dos puntos para que se consideren vecinos.

DBSCAN



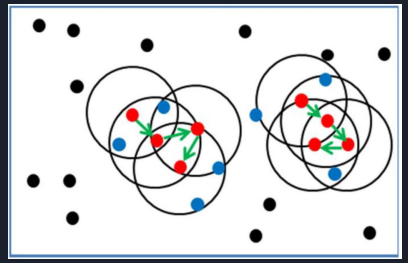
Ventajas:

- No requiere especificar el número de clústeres de antemano.
- Es capaz de descubrir clústeres de formas arbitrarias.
- Es robusto al ruido y a los valores atípicos.

Desventajas:

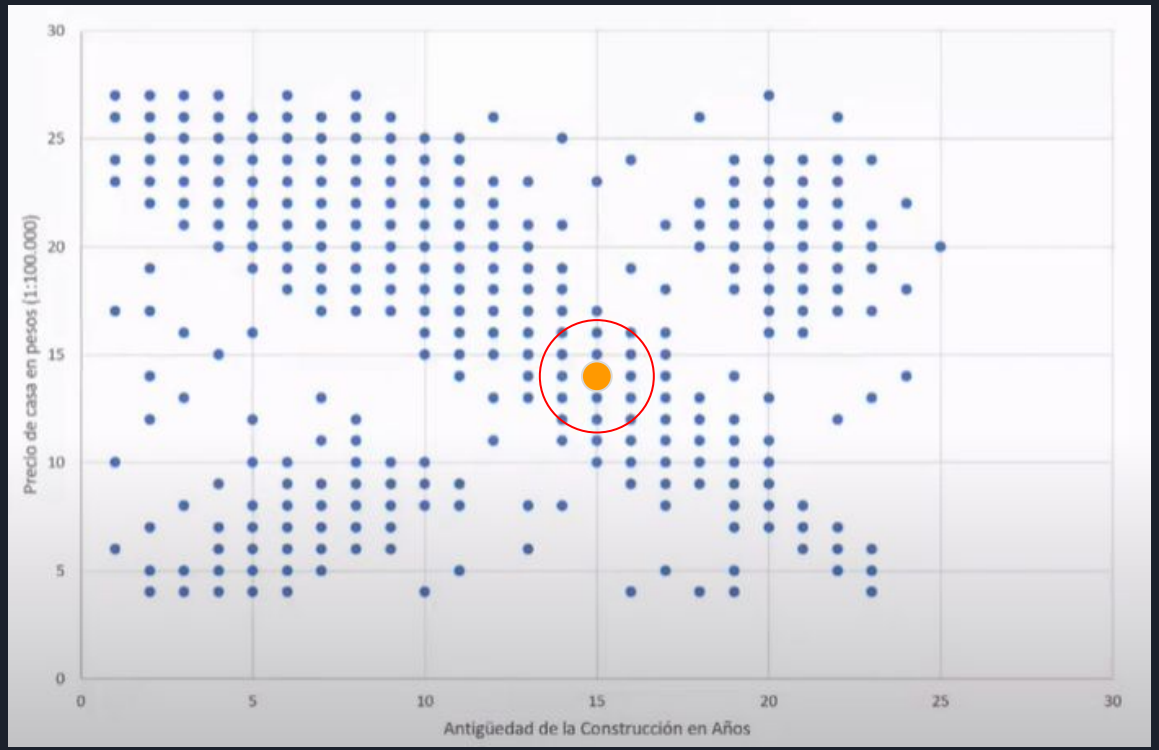
- Es sensible a los parámetros ε y MinPts.
- Puede tener dificultades para identificar clústeres con densidades variables.
- Puede tener problemas con Datasets de muy alta dimensionalidad.

K-means vs. DBSCAN



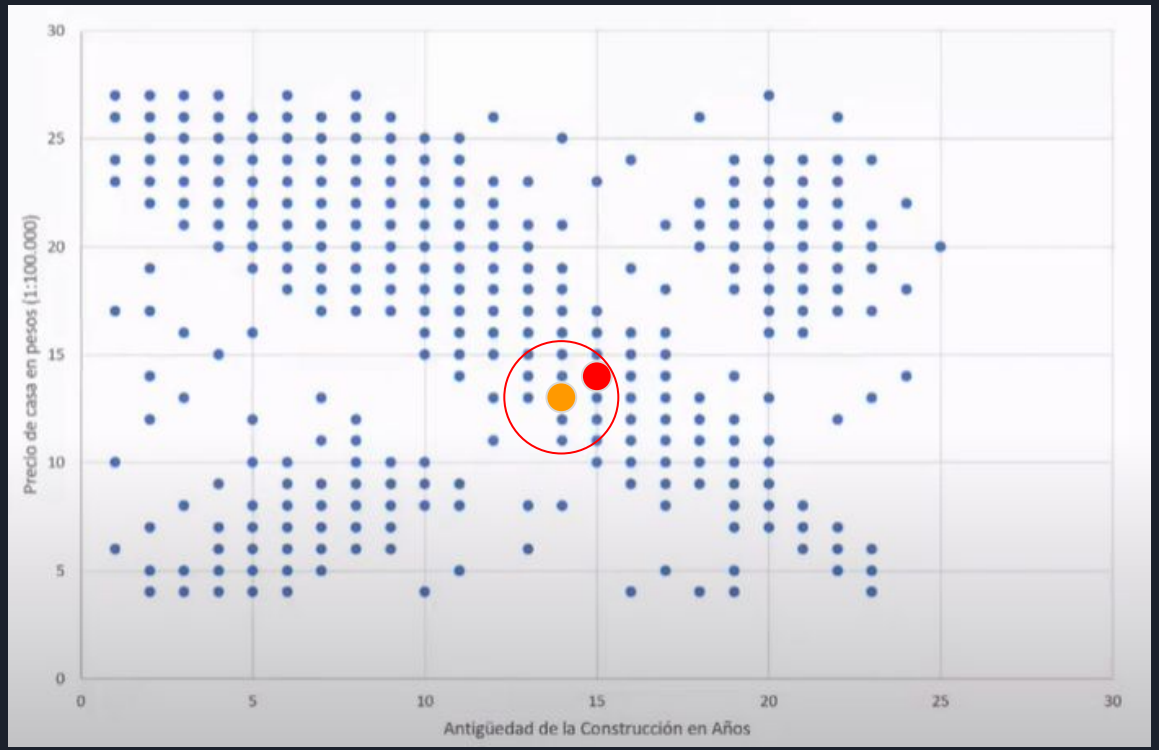
K-means	DBSCAN/HBSCAN
La agrupación del K means es sensible a la cantidad de agrupaciones definidas	No es necesario especificar el número de conglomerados
Su agrupación en Clústeres no funciona bien con valores atípicos y conjuntos de datos con ruido	Su agrupación en Clústeres maneja de manera eficiente los outliers y los datos ruidosos
Las densidades variables de los puntos de datos no afectan el algoritmo K means	La agrupación basada en clústeres no funciona muy bien para conjuntos dispersos o para puntos con densidad variable
La agrupación en clústeres de K-means es más eficiente para grandes conjuntos de datos	No puede manejar de manera eficiente conjuntos de datos con alta dimensión

Repaso funcionamiento



- 1. Selección aleatoria del primer punto
- 2. Inicia medición ϵ y min_samples
  - ϵ = 0.2
  - min_samples = 10
- 3. Si cumple, se lo considera como punto central

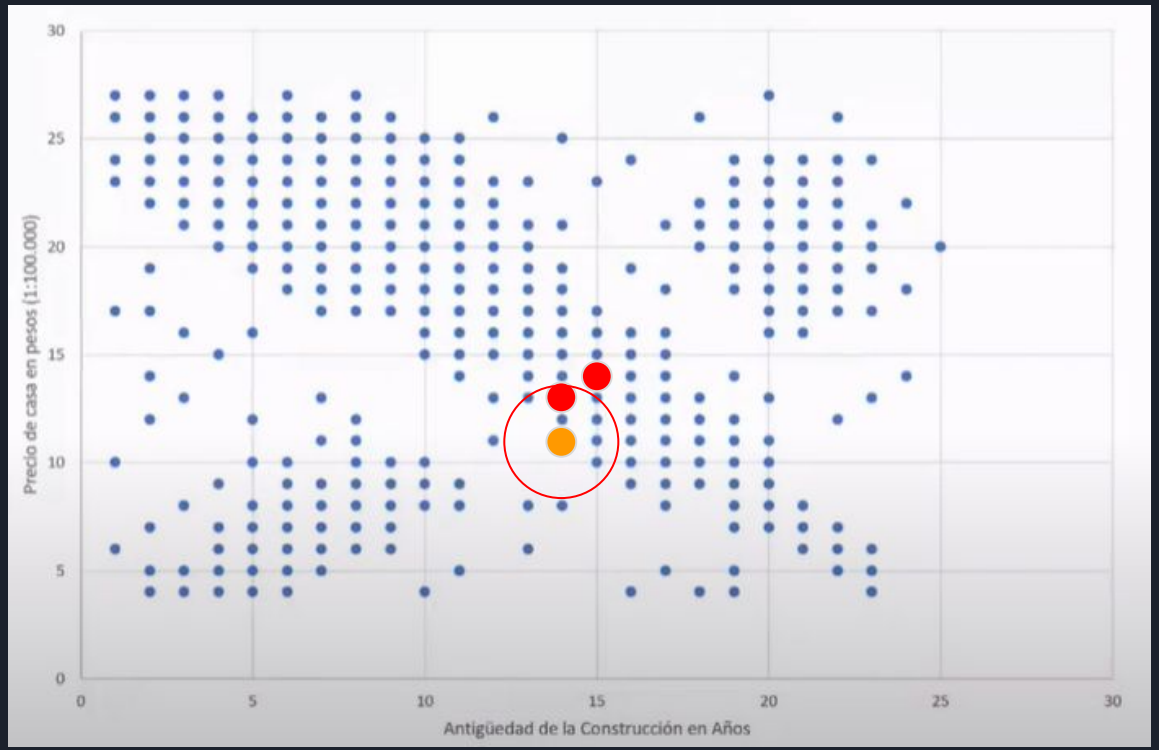
Repaso funcionamiento



Expansión

- 4. Se traslada a otro vecino del mismo grupo y vuelve a aplicar los criterios de medición

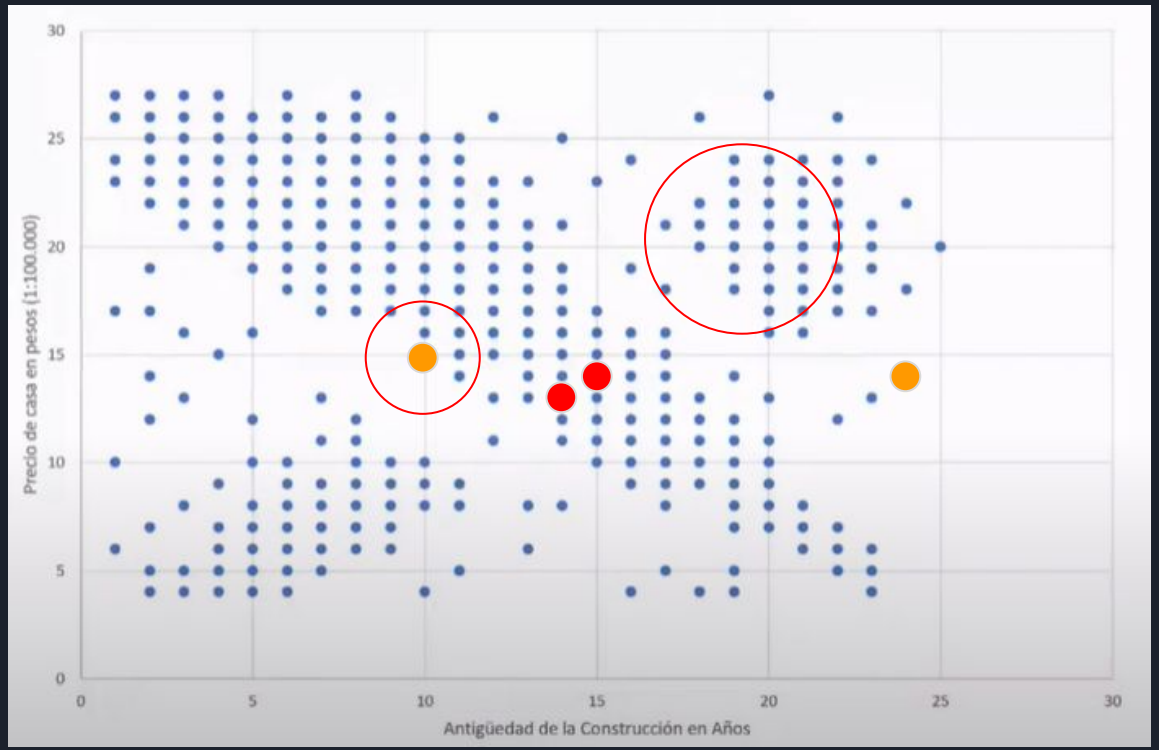
Repaso funcionamiento



Expansión

- 5. Continúa hasta que ya no puede expandirse

Repaso funcionamiento



Expansión

- 6. Finalizada la expansión vuelve a selecciona un punto aleatorio e intenta aplicar los criterios de medición.
- 7. Si no cumple, se lo etiqueta como “ruido” (-1)