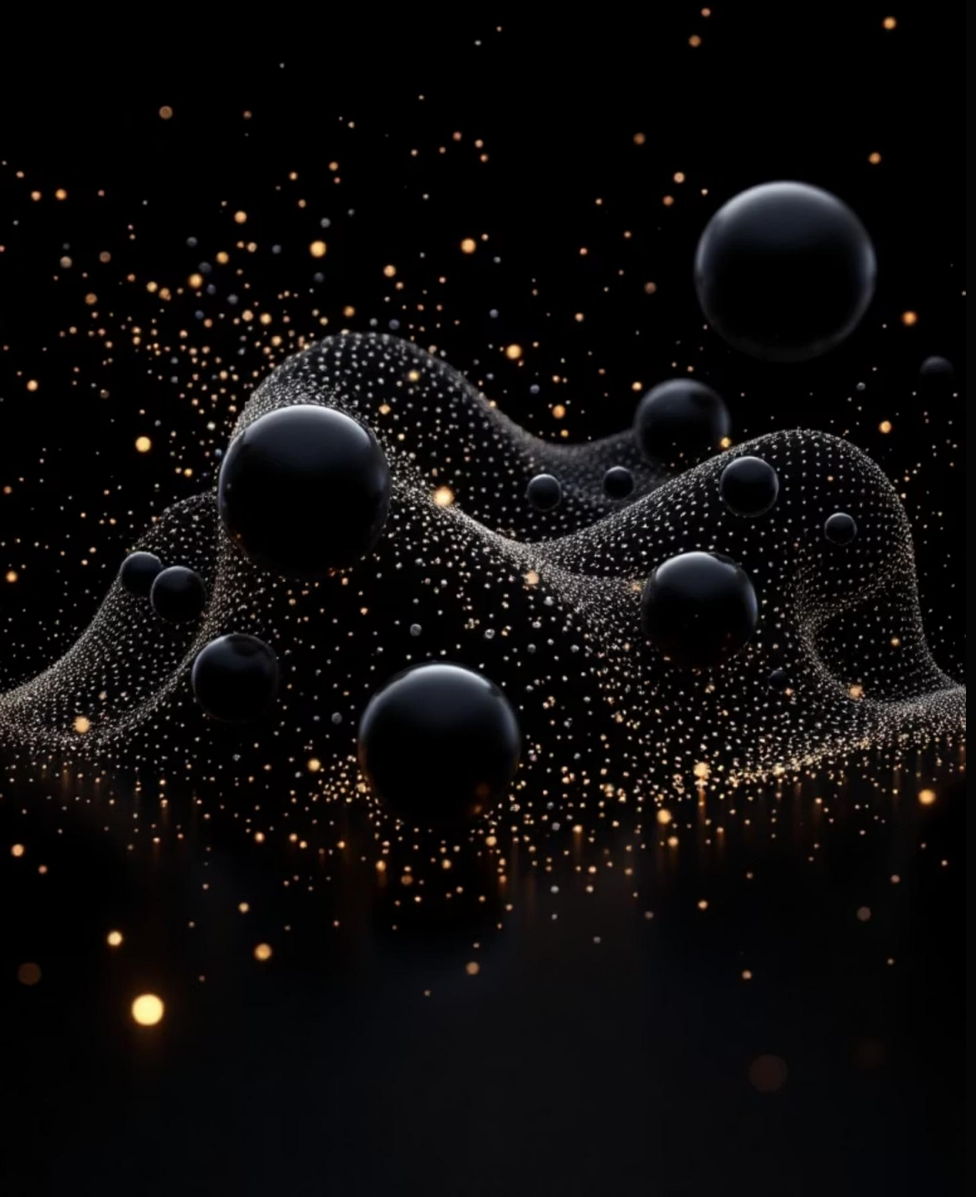


Aprendizaje no supervisado

- Sin conocimiento de una clase o valor objetivo
- Los datos no están etiquetados
- Meta: descubrir patrones, estructura, factores no observados o una representación mas simple

Aprendizaje supervisado

Edad	Ingresos	Tiene carro?
24	1’200.000	NO
23	4’500.000	SI
45	1’250.000	SI
32	1’100.000	NO

Datos **etiquetados**: “Respuestas correctas” disponibles

Factores/atributos/variables independientes, predictores, explicativos

Dependiente, objetivo, respuesta, salida

34	3’500.000

?

¿Cuál es el valor predicho para una instancia dada?

Aprendizaje no supervisado

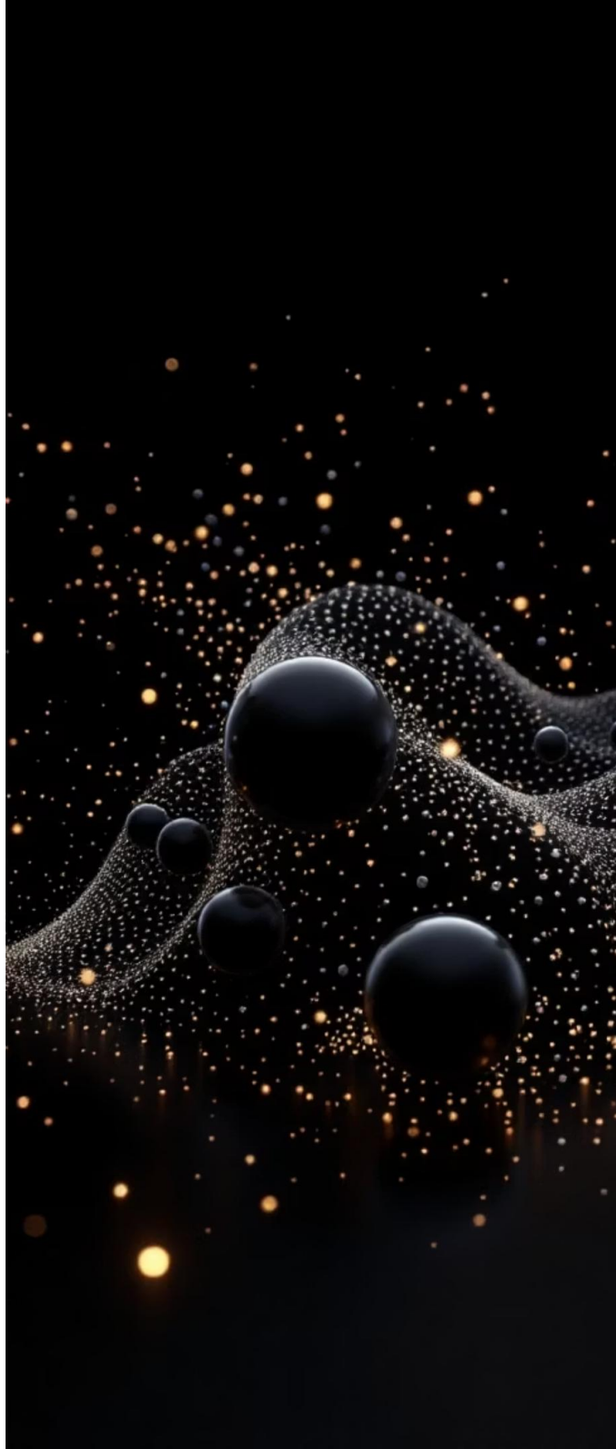
Edad	Ingresos
24	1’200.000
23	4’500.000
45	1’250.000
32	1’100.000

Factores/atributos/variables

¿Se puede encontrar alguna estructura en los datos?

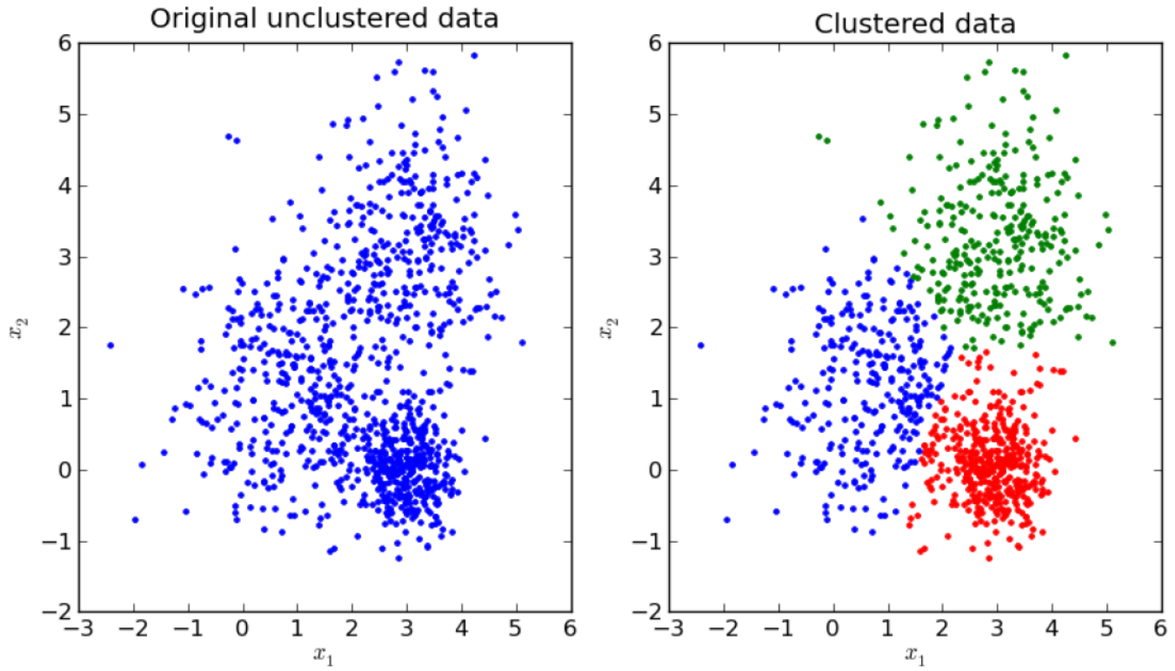
Aprendizaje no supervisado

- No se interesa por la predicción sino por encontrar una estructura, un nuevo punto de vista, una simplificación o un resumen de los datos
- Usualmente se incluye en la fase exploratoria de datos
- Tipos de tareas:
  - Segmentación (clustering)
  - Cambio de representación (e.g. reducción de dimensiones, selección de factores)
  - Reglas de asociación
  - Detección de anomalías (i.e. excepciones)
- Difícil de validar los resultados, ya que no se cuenta con un "gold standard"



Clustering

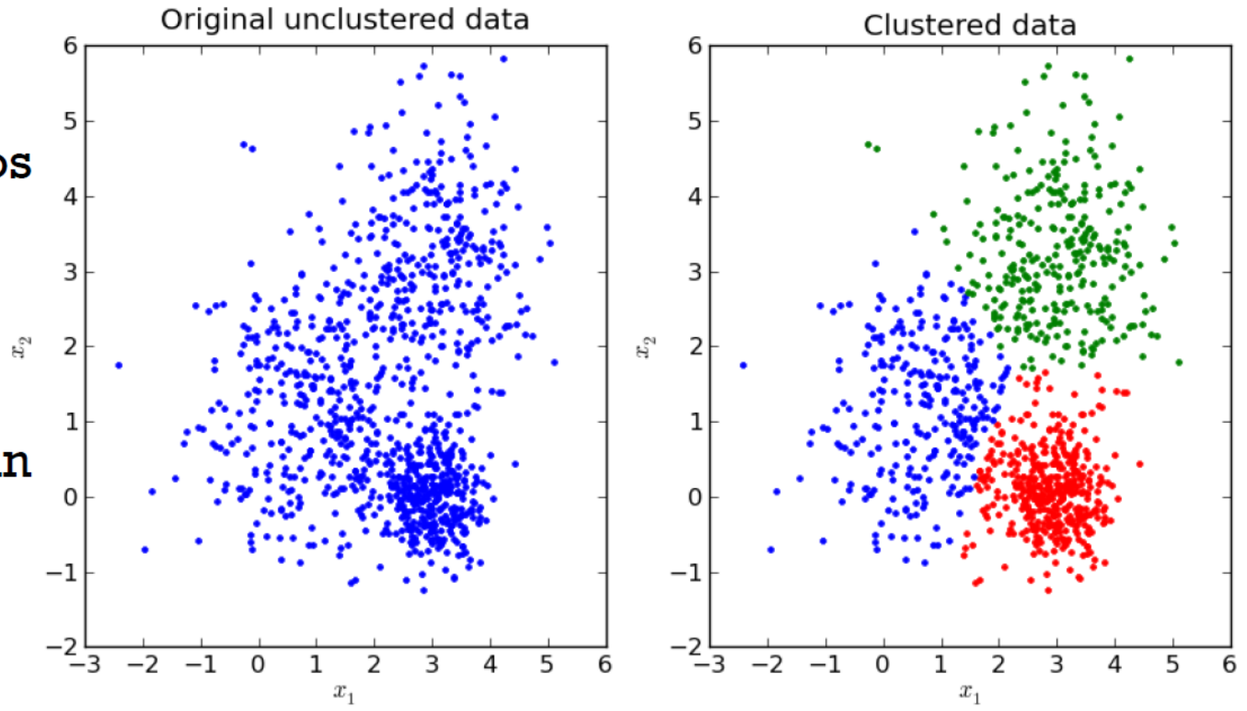
- No se tiene una variable objetivo
- Se busca agrupar los datos similares para encontrar patrones globales de los datos
- Agrupamiento por **similitud**, **proximidad**, **densidad**
- Particionar un conjunto en grupos, de forma tal que elementos en un mismo grupo sean similares entre sí y tan diferentes como sea posible de elementos en otros grupos.



http://pypr.sourceforge.net/kmeans.html

Clustering por distancia

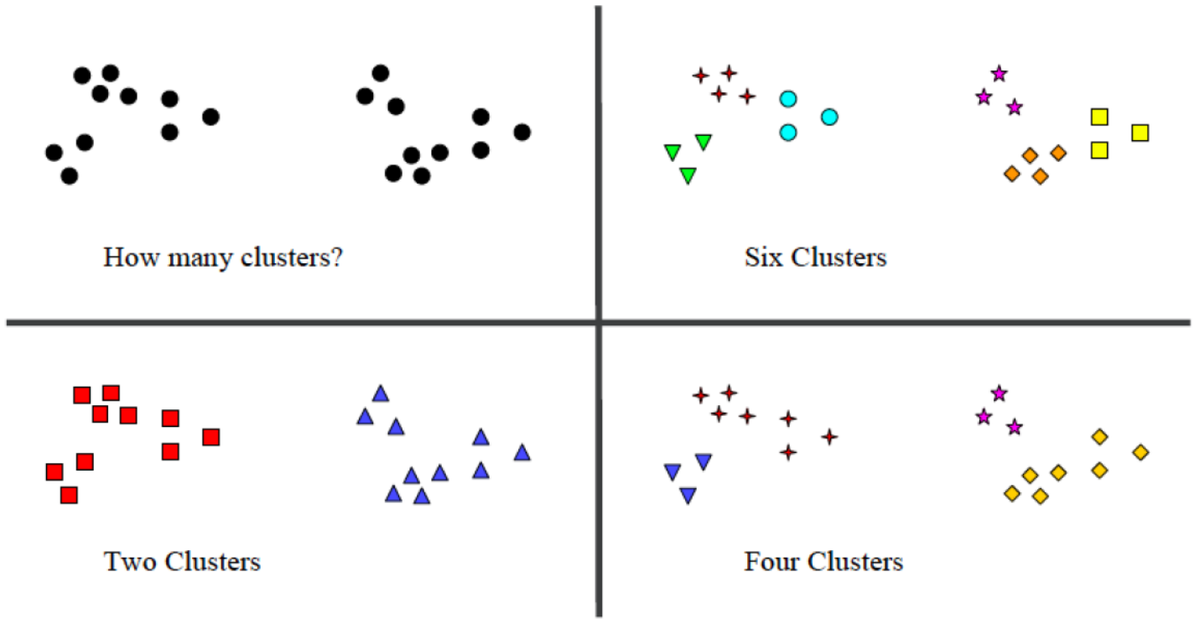
- Objetivo: descubrir **k grupos** o **segmentos** desconocidos que
  - Minimicen la distancia dentro de los grupos
  - Maximicen la distancia por fuera de los grupos
- Se basan en una noción de **distancia**
  - Definición de la medida a utilizar
  - Unidades de los atributos tienen gran influencia
    - → **Normalizar**
    - → **Estandarizar**



http://pypr.sourceforge.net/kmeans.html

Clustering por distancia

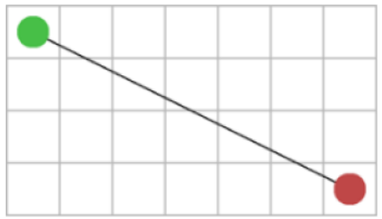
- Se pueden buscar segmentos de observaciones o de atributos
- No existe un método universal absoluto para establecer **k**, solo heurísticas
- Requiere juicio humano, más difícil de automatizar
- La interpretación de los resultados no se debe de hacer de manera absoluta, sino como un punto de partida para el análisis
- Los datos pueden no tener estructura, por lo cual su segmentación podría no tener sentido



http://governingstochastic.weebly.com/blog/category/clustering

Clustering por distancia

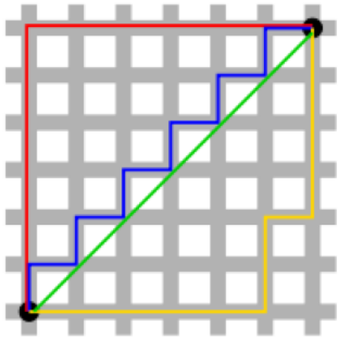
- Medidas de **similitud** o **distancia**:
  - **Euclidiana**: tamaño del segmento linear que une las dos instancias comparadas.



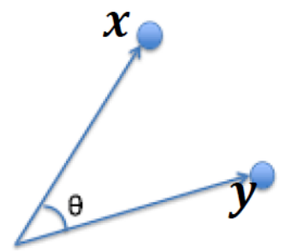
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

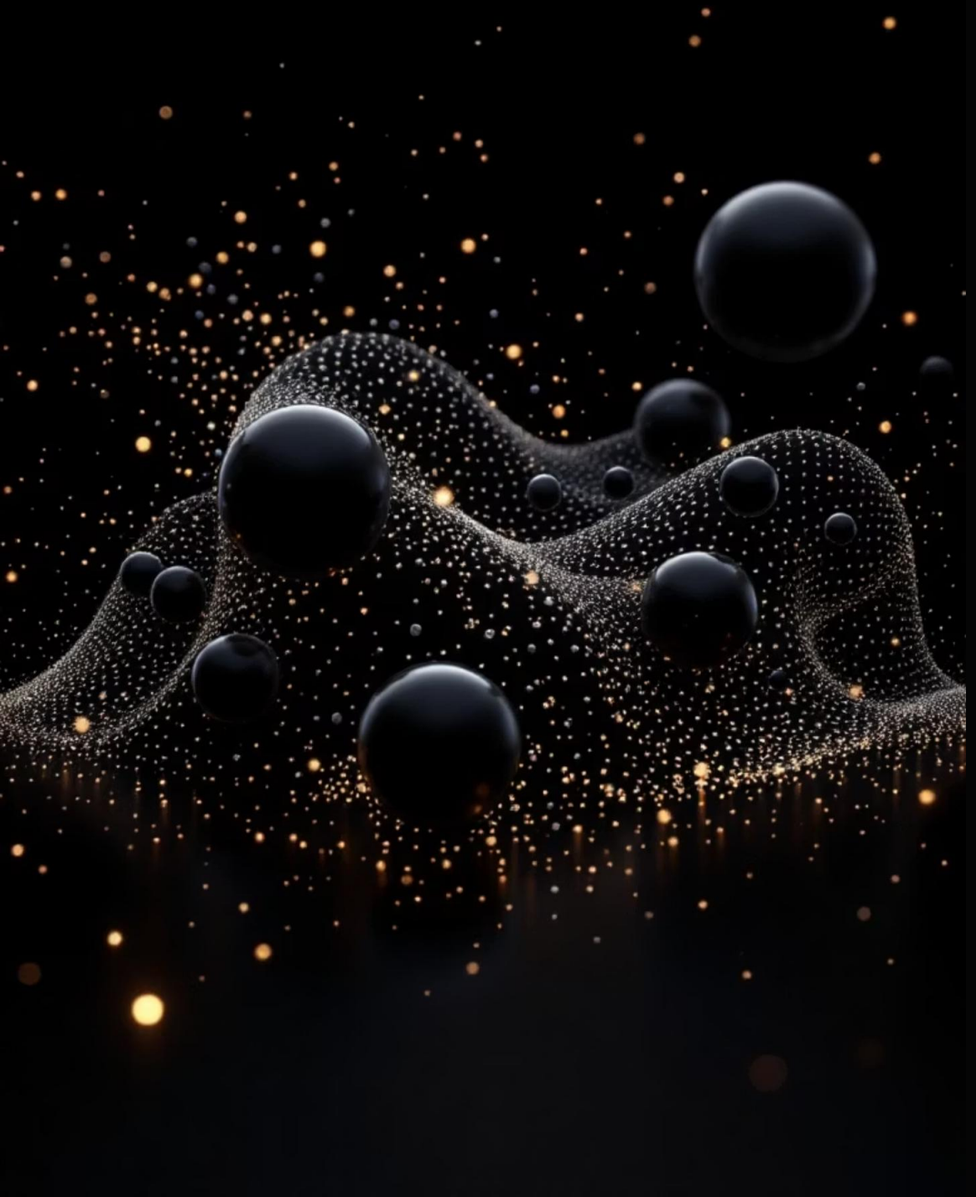
  - **Manhattan**: basada en una organización en bloques rectilíneos



  - **Coseno**: coseno del ángulo entre las dos instancias comparadas → Alta dimensionalidad y **big data**



$$sim(\boldsymbol{x},\boldsymbol{y}) = \cos(\theta_{x,y}) = \frac{\boldsymbol{x}.\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\sum_i x_i * y_i}{\sqrt{\left(\sum_i x_i * x_i\right) * \sum_i y_i * y_i}}$$



Segmentación de Datos con K-Means

Aprende a aplicar el algoritmo K-Means para agrupar datos sin etiquetas y descubrir patrones ocultos en conjuntos de información complejos.

¿Por qué necesitamos segmentación?

La capacidad de identificar grupos naturales en los datos es fundamental para la toma de decisiones estratégicas en múltiples sectores.

Aplicaciones Militares

Segmentar señales de radar para identificar y clasificar diferentes tipos de objetos y amenazas potenciales.

Estrategias Comerciales

Agrupar clientes según sus patrones de comportamiento de compra para crear campañas de marketing más efectivas.

Medicina Personalizada

Clasificar pacientes por patrones de síntomas para desarrollar tratamientos más precisos y personalizados.

Fundamentos del Aprendizaje No Supervisado

Características Clave

- Trabajamos con datos crudos sin etiquetas predefinidas
- El objetivo es **descubrir estructuras ocultas** en forma de clústeres
- K-Means se posiciona como uno de los algoritmos más populares y efectivos

La belleza del aprendizaje no supervisado radica en su capacidad de revelar patrones que no son evidentes a simple vista.

Algoritmo K-Means: Paso a Paso

01

Selección de K

Determina el número deseado de clústeres que mejor represente tu conjunto de datos.

02

Inicialización de Centroides

Coloca aleatoriamente los centroides iniciales en el espacio de características.

03

Asignación de Puntos

Cada punto de datos se asigna al centroide más cercano utilizando distancia euclidiana.

04

Recálculo de Centroides

Los centroides se actualizan calculando el promedio de todos los puntos asignados a cada clúster.

05

Iteración hasta Convergencia

El proceso se repite hasta que los centroides no cambien significativamente entre iteraciones.

Algoritmo K-Means: Paso a Paso

- Algoritmo:
  1. Inicializar los K centroides
  2. Asignar cada instancia al cluster del centroide más cercano
  3. Re calcular los centroides de cada cluster (el baricentro/promedio)
  4. Repetir pasos 2 y 3 hasta convergencia (hasta que los centroides permanezcan estáticos)
- Cada observación se asigna a un solo cluster, de manera absoluta
- Los clusters no se sobrelapan
- Objetivo: minimizar la variación dentro de los clusters (Within Sum of Squares - WSS):

$$WSS = \sum_{i=1}^{\#instancias} distancia(\boldsymbol{x_i} - centroide(\boldsymbol{x_i}))^2$$

Visualización Intuitiva del Proceso

Imagina datos de pasajeros plotear edad versus tarifa pagada. K-Means identifica automáticamente grupos naturales basados en la proximidad de estos puntos en el espacio bidimensional.

Clúster Azul

Pasajeros jóvenes con tarifas económicas

Clúster Verde

Adultos de mediana edad con tarifas moderadas

Clúster Rojo

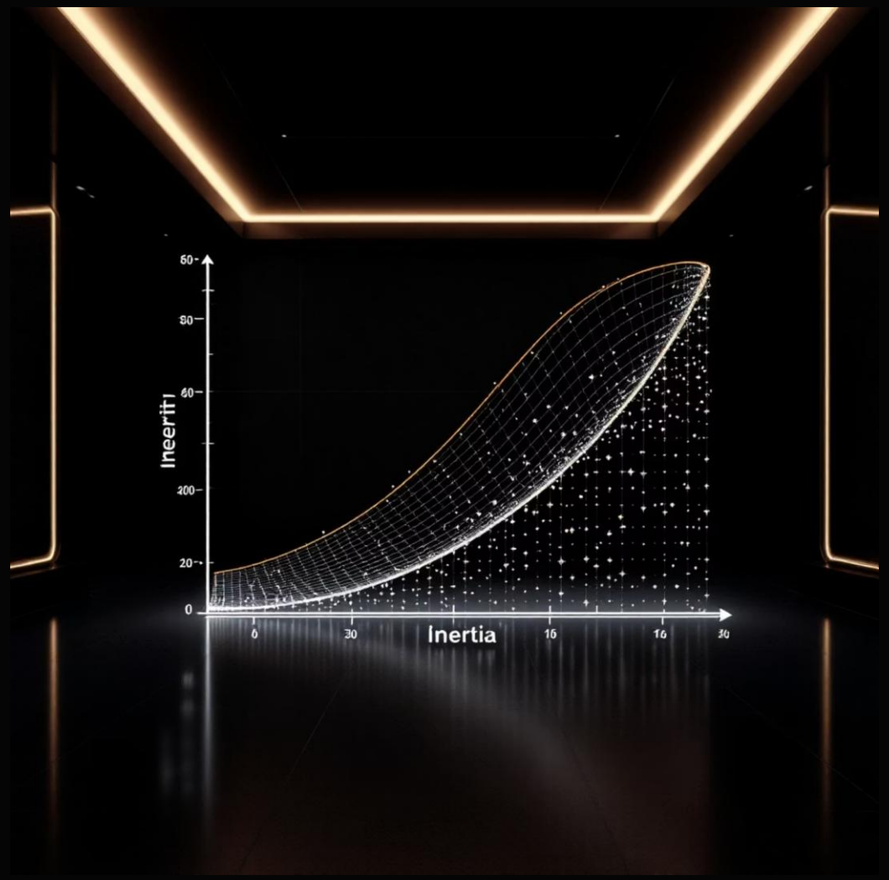
Pasajeros de alto valor con tarifas premium

Determinando el Número Óptimo de Clústeres

- Consideraciones:
  - ¿Cómo estimar el número de clusters (K)?
    - Mardia (1979): $\sqrt{n/2}$
    - Método "del codo"
    - Método Silhouette
    - Medida de CH
  - ¿Cómo inicializar los centroides de los clusters?
    - Escoger centros completamente aleatorios
    - Escoger puntos existentes aleatoriamente
    - Escoger los centroides utilizando K-Means ++

Determinando el Número Óptimo de Clústeres

Método del Codo



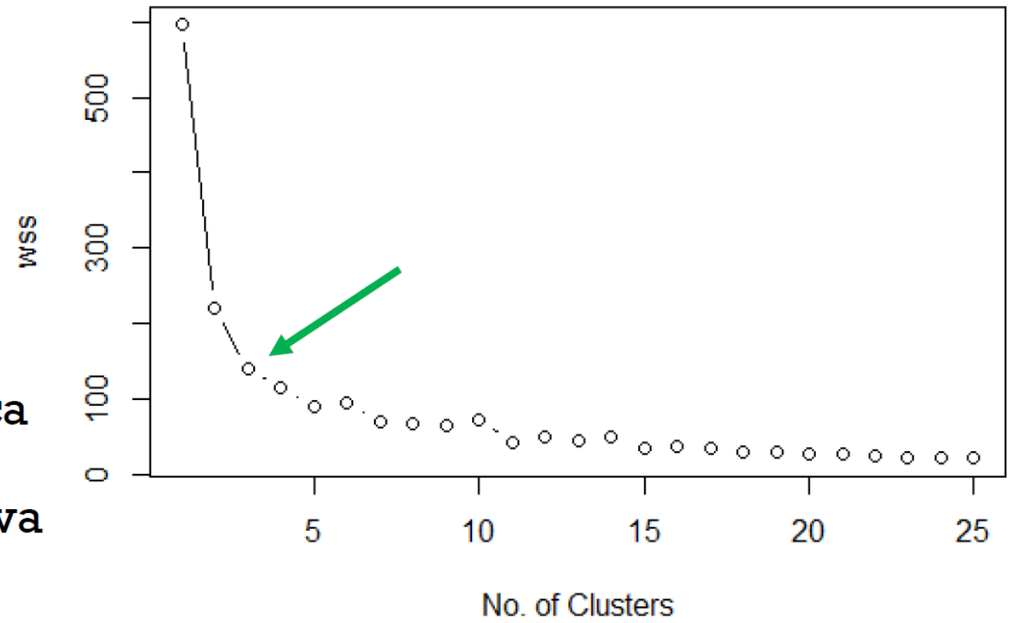
Analiza la suma de distancias dentro de los clústeres (inercia). El punto donde la curva se estabiliza indica el K óptimo.

Silhouette Score

Mide qué tan bien están separados los clústeres. Valores cercanos a 1 indican una excelente separación entre grupos.

Método del Codo

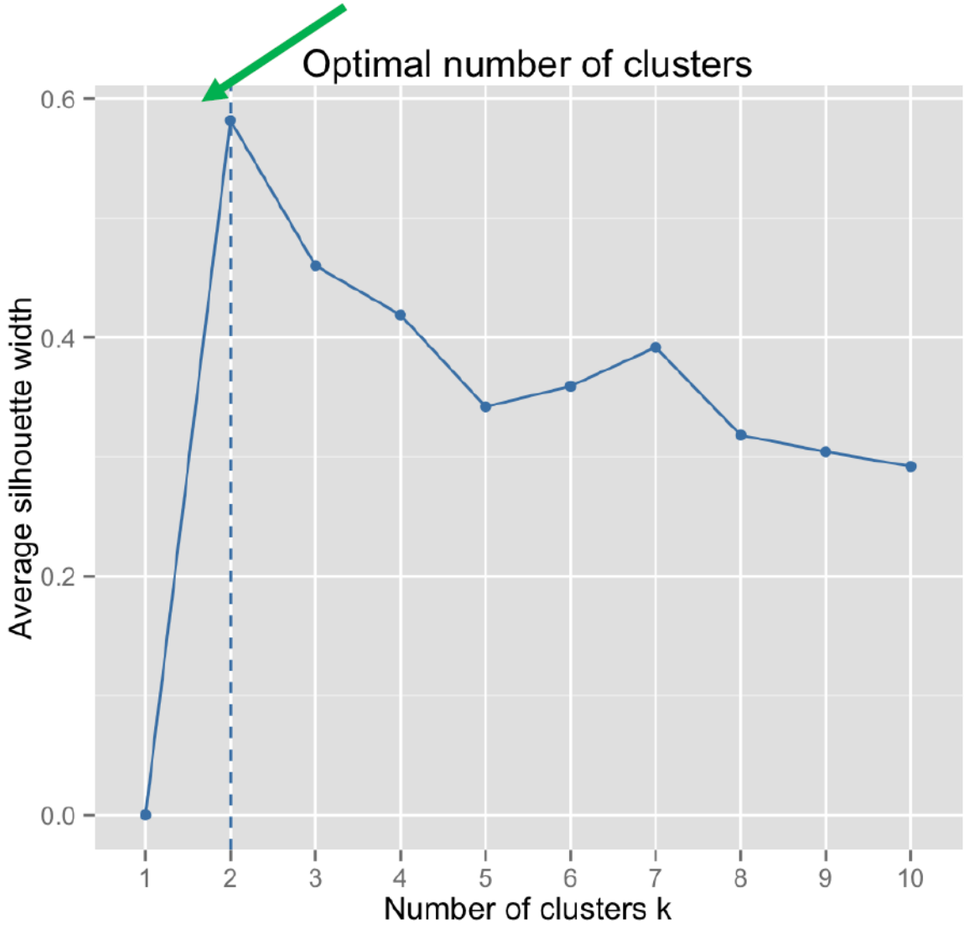
- Heurísticos:
  - No hay un método absoluto
  - Dependen del juicio del analista, se requiere conocimiento del negocio
- Método "del codo":
  - Plotear WSS para cada valor de K
  - Escoger el último valor de K que implica una reducción "considerable" del WSS del clustering resultante, cuando la curva se vuelve aproximadamente lineal



$$WSS = \sum_{i=1}^{\#instancias} distancia(\boldsymbol{x_i} - centroide(\boldsymbol{x_i}))^2$$

Silhouette Score

- Método Silhouette
  - Se busca el K que maximice la **separación** entre clusters, con clusters lo más **compactos** posibles
  - Analizar el ajuste de cada instancia al cluster al que fue asignado
  - Qué tan cerca está cada observación de las demás de su propio cluster
    - 0,7-1,0: el cluster es fuertemente robusto
    - 0,5-0,7: el cluster es razonablemente robusto
    - 0,25-0,5: el cluster puede ser artificial y puede no denotar una noción de estructura necesariamente
    - Inferior a 0,25: el cluster debería descartarse, no indica estructura
  - Se busca la maximización del valor Silhouette promedio de los clusters



Análisis Crítico: Ventajas vs Limitaciones

Simplicidad y Eficiencia

Algoritmo computacionalmente eficiente, ideal para conjuntos de datos grandes con implementación directa.

Interpretabilidad Clara

Los resultados son fáciles de visualizar y explicar, facilitando la comunicación de insights.

Dependencia del Parámetro K

Requiere definir previamente el número de clústeres, lo que puede ser desafiante sin conocimiento del dominio.

Sensibilidad a Escalas

Variables con diferentes escalas pueden sesgar los resultados. Los outliers también afectan significativamente la agrupación.

Limitación Geométrica

Solo identifica clústeres de forma aproximadamente esférica, fallando con formas más complejas.

Caso de Estudio: Dataset Titanic

Aplicamos K-Means a las variables **Edad** y **Tarifa** para segmentar pasajeros y descubrir patrones socioeconómicos interesantes.

Grupo Económico

Pasajeros jóvenes con tarifas bajas, principalmente en tercera clase con recursos limitados.

Clase Ejecutiva

Adultos profesionales con tarifas altas, representando la clase media-alta de la época.

Segmento Senior

Pasajeros mayores con tarifas intermedias, posiblemente jubilados con ahorros moderados.

Aplicaciones Estratégicas por Sector

Sector Defensa

Segmentación de patrones de comunicación enemiga, análisis de trayectorias de drones y clasificación automática de amenazas en tiempo real.

Inteligencia de Negocios

Segmentación avanzada de clientes para campañas de marketing ultra-personalizadas y optimización de estrategias de retención.

Sector Educativo

Agrupación de estudiantes según patrones de desempeño académico para desarrollar programas de apoyo personalizados y metodologías adaptativas.

Reflexiones Finales

Descubrimiento de Patrones Ocultos

K-Means revela estructuras latentes en datos sin etiquetas, abriendo nuevas perspectivas de análisis que no son evidentes inicialmente.

Preparación Crítica de Datos

El escalado y limpieza de datos son pasos fundamentales que determinan la calidad y confiabilidad de los resultados obtenidos.

Decisión Estratégica del Número de Clústeres

La elección correcta de K requiere una combinación de métodos estadísticos y conocimiento experto del dominio de aplicación.

Plataforma para Análisis Avanzados

K-Means sirve como punto de partida sólido para técnicas de segmentación más sofisticadas y análisis predictivos complejos.