IV.2. Métricas de evaluación de modelos

Myriam Raquel Almuina Orozco

Introducción

Objetivo del ejercicio:

- Evaluar modelos de agrupación y reducción de dimensionalidad usando métricas formales.
- Aplicar K-Means y PCA sobre un dataset con 4 atributos numéricos.
- Interpretar los resultados mediante tablas y visualizaciones.

Contexto:
El clustering permite agrupar datos sin etiquetas, mientras que la reducción de dimensionalidad ayuda a visualizar y simplificar los datos manteniendo la información esencial.

Métrica 1 de agrupación: *Silhouette Score*

Mide qué tan separado está cada punto de otros clústeres y qué tan compacto es dentro del propio clúster.

Fórmula:
s = (b - a) / max(a, b)

a = distancia promedio al clúster propio
b = distancia promedio al clúster más cercano

Interpretación:

- 1 → agrupación excelente
- 0 → clústeres solapados
- Negativo → asignación incorrecta

Métrica 2: *Davies–Bouldin Index*

Mide la relación entre la distancia entre clústeres y su dispersión interna.

Interpretación:

- Valores bajos = mejor clustering
- Ideal: cercano a 0

Ventajas: rápido de calcular
Limitaciones: sesgado a clústeres esféricos

Métrica 3: *Calinski–Harabasz*

Evaluación basada en la relación entre separación de clústeres e intra-dispersión.

Interpretación:

- Valores altos = mejor separación

Ventajas: muy estable
Limitaciones: favorece clústeres del mismo tamaño

Métricas de reducción de dimensionalidad

A) Varianza explicada

Definición: proporción de información capturada por los componentes principales.

Interpretación:

- Alto valor → PCA conserva bien la información.

B) Trustworthiness

Definición: mide qué tan bien la estructura local del espacio original se mantiene en el espacio reducido.

Interpretación:

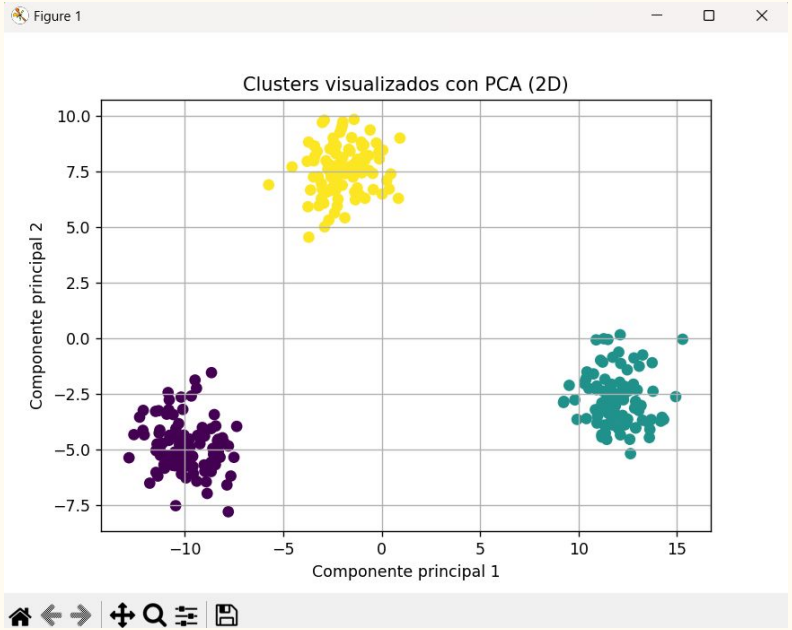
- 1.0 → estructura perfectamente conservada
- 0.9 → muy bueno

Descripción del dataset

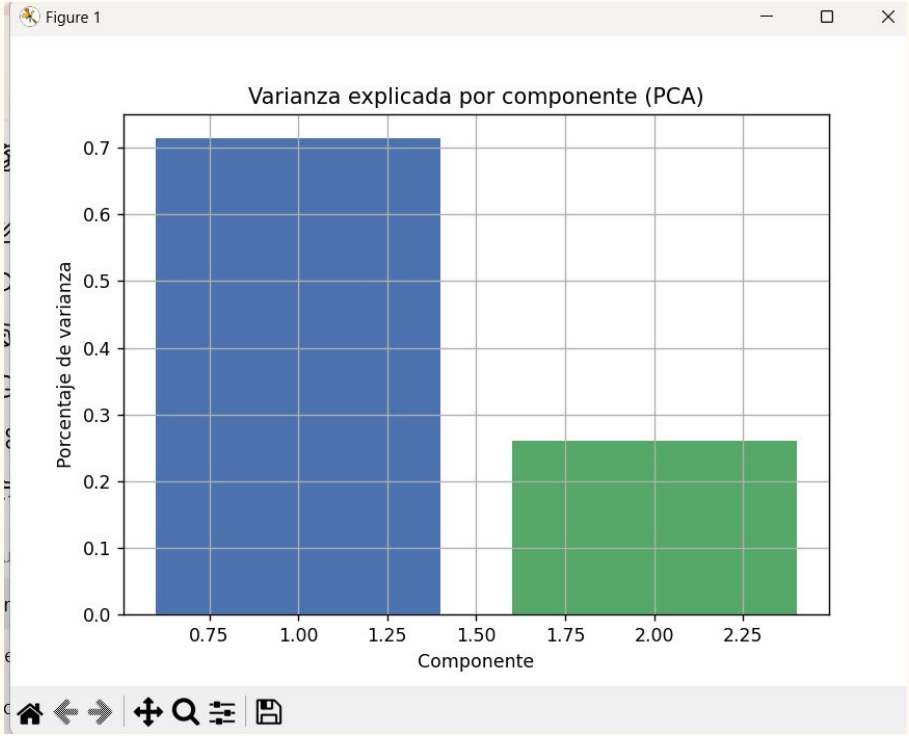
Dataset sintético generado con `make_blobs`:

- 300 muestras
- 4 atributos numéricos
- 3 clústeres
- Varianza moderada
- Ideal para evaluar K-Means y PCA

Resultados del clustering



Resultados de reducción



Comparativa y conclusiones

Comparativa

- *Clustering:* evalúa la calidad de las agrupaciones.
- *Reducción:* evalúa cuánta información se preserva al representar los datos en menos dimensiones.

Conclusiones

- K-Means funcionó excelente con este dataset.
- PCA mantuvo el 97% de la varianza, permitiendo visualizar los clústeres.
- Las métricas coinciden en que los grupos son claros y bien definidos.
- El trustworthiness muestra que, aunque se perdió algo de estructura local, la reducción es aceptable.