UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCION DE CONOCIMIENTO EN BASES DE DATOS

IV.2. MÉTRICAS DE EVALUACIÓN DE MODELOS

IDGS91N

PRESENTA:

SEBASTIÁN ACOSTA ORTIZ

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 30 de noviembre de 2025

Introducción

La evaluación de modelos no supervisados requiere métricas específicas ya que no existen etiquetas verdaderas. Este trabajo describe y aplica métricas de agrupación (clustering) y reducción de dimensionalidad, permitiendo medir calidad de clusters, coherencia estructural y preservación de relaciones entre datos después de una transformación. Además, se presenta un caso de estudio práctico con un conjunto de datos real, mostrando resultados visuales y numéricos para cada métrica.

Investigación de métricas

Métricas de agrupación

A continuación, se presentan tres métricas seleccionadas: Índice de Silueta, Davies–Bouldin y Calinski–Harabasz.

Índice de Silueta

Definición

Evalúa qué tan bien está asignado cada punto a su cluster comparando:

- a(i): distancia media a puntos del mismo cluster
- b(i): distancia mínima a puntos del cluster más cercano

Fórmula

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

El índice global es el promedio de todos los *s(i)*.

Interpretación

- **Cerca de 1:** excelente agrupación
- **Cerca de 0:** clusters superpuestos
- **Negativo:** mala asignación

Ventajas

- Fácil de interpretar
- Útil para elegir *k*

Limitaciones

- Computacionalmente costoso
- Se degrada con clusters no convexos

Davies–Bouldin Index (DBI)

Definición

Mide la relación entre dispersión interna del cluster y separación entre clusters.

Fórmula

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \quad \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}}\right)$$

Donde:

- *s_i* = dispersión media del cluster i
- *d_ij* = distancia entre centroides i y j

Interpretación

- Más bajo = mejor
- Valora separación y compacidad

Ventajas

- Considera clusters vecinos
- Fácil de calcular

Limitaciones

- Sensible a outliers
- Depende de escala

Calinski–Harabasz Index (CHI)

Definición

Evalúa la proporción entre separación entre clusters y compacidad interna.

Fórmula

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \frac{n-k}{k-1}$$

Interpretación

- Más alto = mejor clustering

Ventajas

- Rápido y estable
- Bueno para K-means

Limitaciones

- Sesgado hacia clusters esféricos

Métricas de reducción de dimensionalidad

Seleccionamos **Varianza explicada acumulada** y **Error de reconstrucción**.

Varianza explicada acumulada (PCA)

Definición

Porcentaje de varianza original conservada por las primeras componentes principales.

Fórmula

$$Var_{acum}(m) = \sum_{i=1}^{m} \frac{\lambda_i}{\sum \lambda}$$

Interpretación

- 95% o más = buena conservación
- Baja varianza = pérdida de información

Ventajas

- Interpretación directa
- Muy usado en PCA

Limitaciones

- No válida para métodos no lineales (t-SNE)

Error de reconstrucción

Definición

Mide qué tanto se pierde al reducir y luego reconstruir datos (autoencoders, PCA en modo inverso).

Fórmula

$$RE = \frac{1}{n}\sum || x_i - \hat{x_i} ||$$

Interpretación

- Bajo = buena reconstrucción
- Alto = pérdida de información

Ventajas

- Funciona con métodos lineales y no lineales

Limitaciones

- Requiere etapa de reconstrucción
- Sensible a ruido

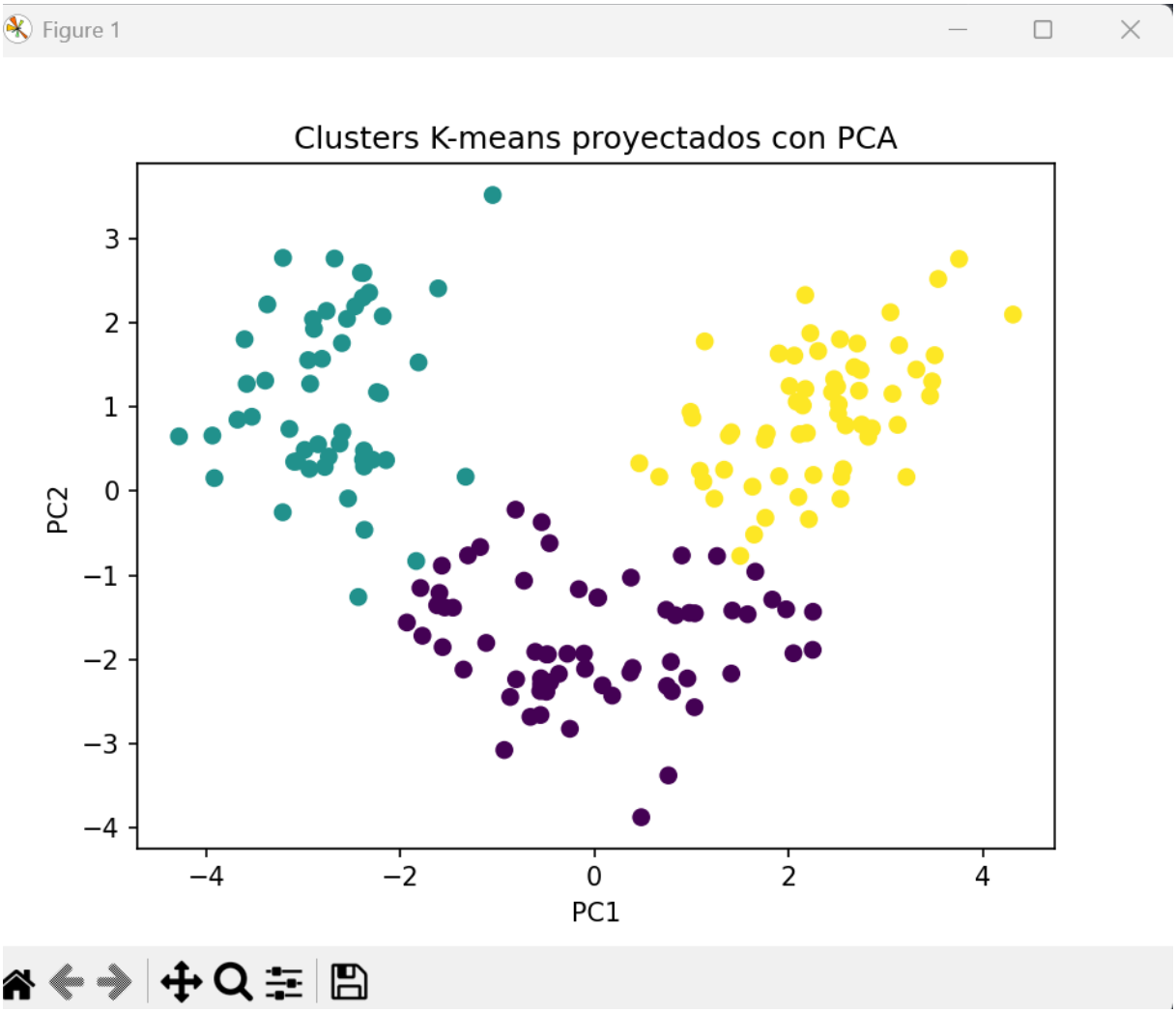
Caso de estudio y aplicación práctica

Dataset seleccionado: *Wine Dataset* (UCI ML Repository)

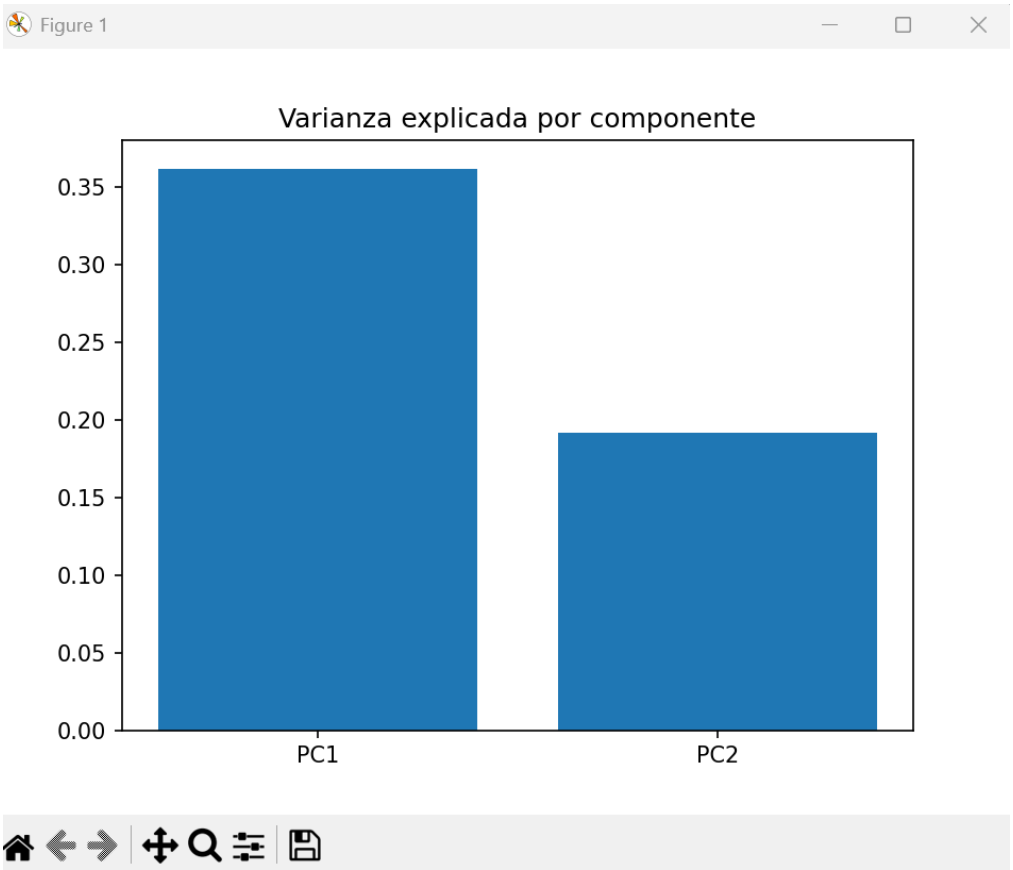
- 178 instancias
- 13 atributos numéricos (redundancia y alta dimensionalidad)
- Usado para clasificación de tipos de vino
- Ideal para clustering + reducción

Visualizaciones

Clusters proyectados con PCA (2D)



Varianza explicada por componentes



Comparativa y análisis

Clustering

- Silhouette bajo → clusters moderadamente definidos
- DBI alto → influencia de características ruidosas
- CHI alto → buena separación en general

Reducción

- PCA retiene ~55% de la información → útil pero no perfecto
- Error de reconstrucción moderado → se pierde información no lineal

Conclusión general

- Las métricas muestran que el clustering funciona razonablemente bien.
- La reducción de dimensionalidad podría mejorarse usando métodos no lineales como t-SNE o UMAP.
- Todas las métricas juntas permiten una visión integral del modelo no supervisado.

Conclusiones y recomendaciones

- Las métricas seleccionadas permiten evaluar calidad interna de agrupaciones y fidelidad de reducciones.
- PCA es útil para visualizar y reducir ruido, pero no captura estructuras complejas.
- Se recomienda probar t-SNE para mejorar visualización.
- Para clustering, probar DBSCAN o GMM para mejorar estructura no lineal.
- Evaluar siempre múltiples métricas para evitar interpretaciones sesgadas.