Métricas de evaluación de modelos

Extracción de Conocimiento en Bases de Datos

Luis Eduardo Aguilar Sarabia

Grupo: IDGS91N

29 de noviembre de 2025

Introducción — Objetivos

- Entender y comparar métricas internas para evaluar clustering.
- Definir y aplicar métricas de reducción de dimensionalidad.
- Aplicar clustering y reducción en un dataset real (Iris) y presentar resultados, interpretación y recomendaciones.

Métricas de agrupación (1/2) — Índice de Silueta

- Definición: Para cada punto i: a(i) = distancia media intra-cluster; b(i) = mínima distancia media al siguiente cluster. s(i) = (b(i)-a(i))/max(a(i),b(i)).
- Silhouette global: promedio de s(i) sobre todas las muestras.
- Interpretación: Rango [-1,1]. Valores cercanos a 1: clusters bien separados; 0: solapamiento; negativos: asignaciones erróneas.
- Ventajas: Intuitiva, por-instancia; útil para elegir k (gráficas de silhouette).
- Limitaciones: Costosa en datasets grandes; sensible a la métrica de distancia.

Métricas de agrupación (2/2) — DBI y CH

- Davies–Bouldin (DBI): DBI = (1/k) * sum_i max_{j!=i} ((S_i + S_j) / M_ij). Menor es mejor.
- Ventajas DBI: Rápido; balancea compacidad y separación. Limitaciones: Depende de medida de dispersión; puede favorecer tamaños distintos.
- Calinski–Harabasz (CH): CH = (BCSS/(k-1)) / (WCSS/(n-k)). Mayor es mejor.
- Ventajas CH: Rápido; adecuado para KMeans. Limitaciones: Puede favorecer k grandes; sensible a escala.

Métricas de reducción de dimensionalidad

- Explained variance (PCA): ratio por PC = $\lambda_i / \sum_j \lambda_j$. Acumulada = suma de primeras m ratios.
- Interpretación: Alto valor indica que esos PCs conservan gran parte de la variabilidad.
- Ventajas: Directa y fácil de interpretar. Limitaciones: Varianza ≠ información relevante; PCA es lineal.
- Trustworthiness: Mide preservación de vecinos locales entre alta y baja dimensión. Rango (0,1], 1 = perfecta preservación.
- Limitaciones Trustworthiness: Depende de k; costosa para datasets grandes; no captura preservación global.

Descripción del dataset — Iris (UCI)

- 150 instancias, 4 atributos numéricos: longitud sépalo, ancho sépalo, longitud pétalo, ancho pétalo (cm).
- 3 clases verdaderas (solo para referencia).
- Se usó la versión estándar y las 4 columnas numéricas normalizadas (StandardScaler) antes de aplicar KMeans y PCA.

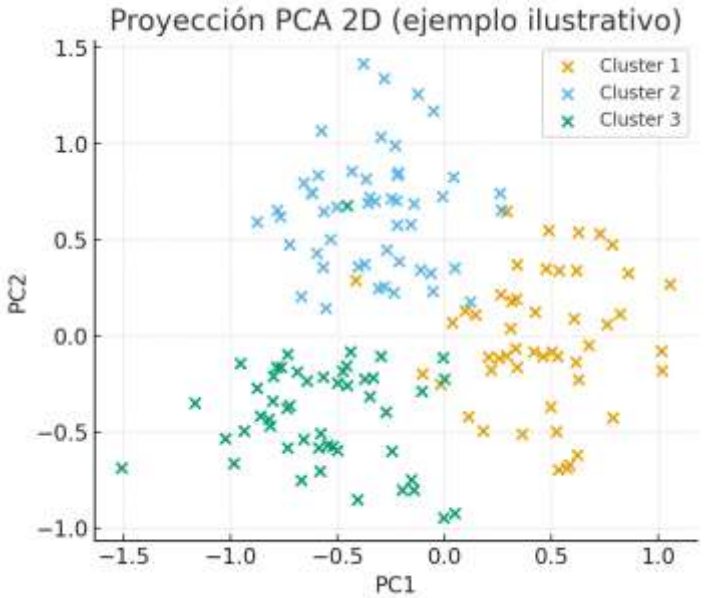
Resultados de clustering (Iris, KMeans k=3)

- Tabla de métricas (resumen):
- Silhouette (media): 0.460 — indica clusters razonablemente separados.
- Davies–Bouldin: 0.834 — valor bajo (bueno).
- Calinski–Harabasz: 241.90 — valor alto, indica buena separación relativa.
- Visualización: scatter 2D usando las dos primeras PCs y coloreado por cluster (ver figura adjunta).

Resultados de reducción (PCA)

- Explained variance ratios (por PC):
- 0.7296, 0.2285, 0.0367, 0.0052.
- Varianza explicada acumulada (2 PCs): ≈ 0.958 — las dos primeras PCs explican ~95.8% de la varianza total.
- Trustworthiness (n_neighbors=5) entre espacio original y PCA-2D: ≈ 0.974.
- Se incluyó scree plot y curva acumulada (ver figura adjunta).

Visualizaciones



Comparativa y análisis

- Para Iris (datos numéricos, relativamente esféricos por clase), las métricas coinciden en señalar buena separación con k=3.
- Silhouette ~0.46, DBI ~0.83 (bajo), CH ~242 (alto).
- PCA: 2 componentes retienen ~96% de la varianza; trustworthiness alta indica buena preservación local.
- Recomendaciones: usar Silhouette para inspección por instancia; usar DBI y CH como métricas complementarias; para datos no lineales, complementar con t-SNE/UMAP.

Conclusiones y recomendaciones

- Emplear múltiples métricas ofrece una visión más robusta: Silhouette, DBI y CH se complementan.
- Para reducción, la varianza explicada es útil para PCA; trustworthiness recomendable para conservar vecindades locales.
- Para datasets grandes: considerar costos computacionales; usar muestreo o implementaciones aproximadas si es necesario.

Referencias (APA)

- Daniel. (2024, 26 agosto). Métricas en Machine Learning: Todo lo que necesitas saber. DataScientest. https://datascientest.com/es/metricas-en-machine-learning
- Caballar, R., & Stryker, C. (2025, 22 octubre). ¿Qué es el rendimiento del modelo? IBM. https://www.ibm.com/mx-es/think/topics/model-performance
- Laujan. (s. f.). Métricas de evaluación de clasificación de texto personalizado - Foundry Tools. Microsoft Learn. https://learn.microsoft.com/es-es/azure/ai-services/language-service/custom-text-classification/concepts/evaluation-metrics