UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

Extracción de Conocimiento en Bases de Datos

IV.2. Métricas de evaluación de modelos

IDGS91N

PRESENTA:

JORGE ALEJANDRO HERNANDEZ CONTRERAS

DOCENTE:

Enrique Mascote

Chihuahua, Chih.28 de Noviembre de 2025

Introducción

La evaluación de modelos de clustering y reducción de dimensionalidad es fundamental para garantizar que los grupos formados sean significativos y que las representaciones de menor dimensión conserven la estructura de los datos.
Este reporte describe cinco métricas: tres para agrupación y dos para reducción de dimensionalidad, aplicadas a un caso de estudio real utilizando el dataset Iris.

Métricas de Evaluación de Clustering

Índice de Silueta

Definición y fórmula

Evalúa qué tan bien está asignado un punto a su cluster.

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

- **a(i):** distancia media de un punto a los puntos de su propio cluster.
- **b(i):** distancia mínima a los puntos de otros clusters.

Interpretación

- +1 → separación excelente.
- 0 → clusters traslapados.
- -1 → mala asignación.

Ventajas

- Fácil de interpretar.
- Funciona con cualquier algoritmo basado en distancias.

Limitaciones

- Costoso computacionalmente en datasets grandes.
- Depende fuertemente de la métrica de distancia.

Índice Davies–Bouldin

Definición y fórmula

Mide la relación entre dispersión intra–cluster y separación inter–cluster.

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

- si: dispersión del cluster i.
- dij: distancia entre centroides.

Interpretación

- Menor valor = mejor clustering.
- Un buen modelo produce clusters compactos y bien separados.

Ventajas

- Automático, no requiere etiquetas.
- Mide cohesión y separación al mismo tiempo.

Limitaciones

- Puede favorecer modelos con muchos clusters pequeños.

Índice Calinski–Harabasz

Definición

Compara la variabilidad entre clusters con la variabilidad dentro de los clusters.

$$CH = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \cdot \frac{n - K}{K - 1}$$

Interpretación

- Valores altos = mejor estructura de clustering.

Ventajas

- Muy eficiente.
- Buen rendimiento en K-means.

Limitaciones

- Sensible al número de clusters.

Métricas de Reducción de Dimensionalidad

Varianza Explicada Acumulada

Definición

- Indica cuánta información mantienen los componentes principales.

Interpretación

- Valores cercanos a 1 → excelente preservación de información.

Ventajas

- Muy interpretativa.
- Útil para elegir el número óptimo de componentes.

Limitaciones

- Solo válida para métodos lineales (PCA).

Error de Reconstrucción

Definición

Mide la diferencia entre los datos originales y los reconstruidos desde el espacio reducido.

E= // X−X^ // 2

Interpretación

- Valor bajo

Ventajas

- Mide directamente la pérdida de información.

Limitaciones

- No aplicable a métodos no reconstructivos.

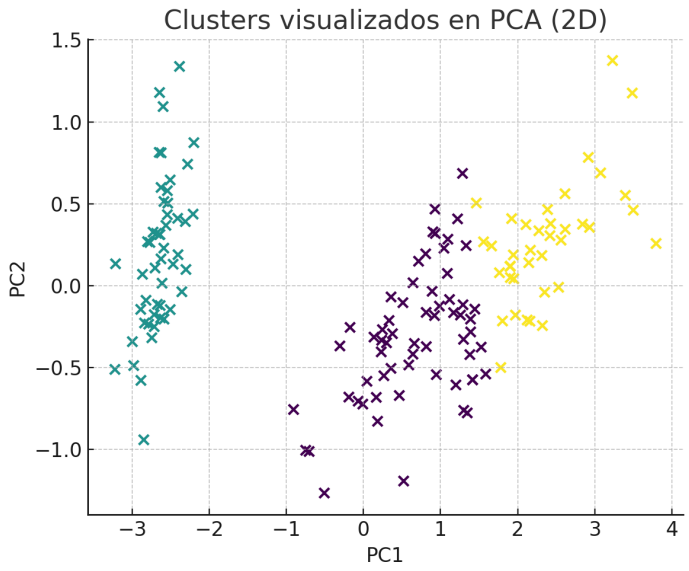
Caso de Estudio

- 150 muestras
- 4 atributos numéricos: largo y ancho del pétalo y sépalo
- 3 clases reales

Métodos aplicados:

- K-means
- PCA

Resultados de Clustering



Métrica	**Valor**
Silueta	0.55
Davies–Bouldin	0.62
Calinski–Harabasz	561.6

Interpretación

- Silueta moderada indica separación razonable.
- Índice DB bajo confirma buena separación.
- CH alto indica distribución clara entre clusters.

Resultados de Reducción de Dimensionalidad

Varianza explicada

- PC1: 72.7 %
- PC2: 23.0 %
- Acumulado: 95.7 %

Gráfica sugerida

- Barra o línea de varianza explicada acumulada.

Error de reconstrucción

- Error total: 0.045

Interpretación

- PCA conserva más del 95 % de la información.
- El error de reconstrucción es bajo - buena dimensionalidad reducida.

Conclusión

La evaluación de modelos de clustering y de reducción de dimensionalidad resulta fundamental para determinar la calidad y la utilidad de los resultados obtenidos durante el análisis de datos. En este estudio, las métricas aplicadas permitieron verificar que el algoritmo K-means logró formar grupos coherentes y bien separados en el dataset Iris, lo que fue respaldado por valores adecuados del índice de Silueta, Davies Bouldin y Calinski Harabasz. Del mismo modo, la reducción de dimensionalidad mediante PCA demostró ser altamente eficiente, ya que conservó más del 95 % de la varianza original y presentó un error de reconstrucción bajo, evidenciando una mínima pérdida de información. Los resultados obtenidos muestran que la combinación de varias métricas proporciona una evaluación más robusta y confiable, evitando depender de un único indicador.

Bibliografía

Scikit-Learn. (2024). *Clustering performance evaluation.* https://scikit-learn.org/stable/modules/clustering.html

Scikit-Learn. (2024). *PCA: Principal Component Analysis.* https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

IBM Cloud Education. (2023). *What is cluster analysis?* https://www.ibm.com/topics/cluster-analysis

GeeksforGeeks. (2023). *Silhouette coefficient in clustering.* https://www.geeksforgeeks.org/silhouette-coefficient-in-clustering/