UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

IV.2. MÉTRICAS DE EVALUACIÓN DE MODELOS

IDGS91N

PRESENTA:

REGINA CHÁVEZ TAMAYO - 6521110019

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 30 de noviembre de 2025

Introducción

El análisis no supervisado permite descubrir estructuras internas en los datos sin necesidad de etiquetas. Para garantizar que estos modelos generen información útil, es esencial medir su desempeño mediante métricas adecuadas.

Este trabajo tiene como objetivos principales investigar y comprender métricas de evaluación para modelos de agrupación y reducción de dimensionalidad; y aplicar estas métricas en un caso práctico usando clustering (K-means) y reducción de dimensionalidad (PCA), evaluando su calidad mediante indicadores cuantitativos.

Métricas de agrupación (Clustering)

Índice de Silueta

Definición y fórmula

El índice de silueta mide qué tan bien está asignado un punto a su cluster comparado con los demás.

Para cada punto *i*:

- a(i): distancia promedio a puntos de su propio cluster
- b(i): distancia promedio al cluster más cercano diferente
- Fórmula:

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

El valor total es el promedio de todos los s(i).

Interpretación

- Cercano a 1: clusters bien definidos
- Cercano a 0: clusters superpuestos
- Negativo: mala asignación

Ventajas

- Fácil de interpretar
- Funciona con cualquier algoritmo basado en distancia

Limitaciones

- Costoso computacionalmente en datasets grandes

Davies–Bouldin Index (DBI)

Definición y fórmula

Mide la relación entre la dispersión interna de cada cluster y la separación entre clusters.

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

Donde:

- s_i: dispersión del cluster *i*
- d_{ij}: distancia entre centroides de *i* y *j*

Interpretación

- Valor más bajo = mejor agrupación
- Indica separabilidad y compactación

Ventajas

- No requiere etiquetas
- Rápido de calcular

Limitaciones

- Sensible a outliers
- Depende de la forma del cluster

Calinski–Harabasz Index

Definición y fórmula

Evalúa la dispersión interna y entre clusters:

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)}$$

- SSB: varianza entre clusters
- SSW: varianza interna
- n: número de muestras

Interpretación

- Valor más alto = mejor clustering

Ventajas

- Eficiente y estable
- Muy usado con K-means

Limitaciones

- Menos útil con clusters no convexos

Métricas de reducción de dimensionalidad

Varianza explicada acumulada (PCA)

Definición

Indica la proporción de la varianza total que retienen los k primeros componentes de PCA:

$$\text{Varianza acumulada} = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$$

Interpretación

- Alta (≥ 90 %): se conserva la estructura del dataset
- Baja: pérdida significativa de información

Ventajas

- Interpretable
- Permite elegir número óptimo de componentes

Limitaciones

- Solo funciona en métodos lineales
- Sensible a outliers

Error de reconstrucción (Autoencoders o PCA)

Definición

$$Error = ||X - \hat{X}||_2$$

Indica cuánto difiere el dato original del reconstruido después de la reducción.

Interpretación

- Error bajo: buena representación
- Error alto: pérdida de estructura

Ventajas

- Métrica directa
- Aplica a métodos lineales y no lineales

Limitaciones

- No mide preservación de distancias entre puntos

Caso de estudio: Dataset Iris

Elegimos el dataset Iris (4 atributos numéricos: sepal length, sepal width, petal length, petal width).

Clustering

- Método: K-means (k=3)
- Variables normalizadas

Reducción de dimensionalidad

- Método: PCA (2 componentes)
- Varianza acumulada ≈ 95 %

Resultados

Clustering visualizado (PCA 2D)

(En la diapositiva: scatter plot con colores para cada cluster.)

Interpretación:

- Los clusters se separan bien en el espacio reducido; uno aparece más compacto (Setosa).

Métricas de agrupación (tabla)

Métrica	Valor
Silhouette	0.55
Davies–Bouldin	0.68
Calinski–Harabasz	561.6

Interpretación:

- Silhouette moderado: clusters razonablemente definidos
- DBI bajo: buena separación
- CH alto: buena compactación

Reducción de dimensionalidad

Varianza explicada por PCA

Componente	Varianza
PC1	72.7 %
PC2	23.0 %
Acumulada	95.7 %

Error de reconstrucción (2 componentes)

Error ≈ 0.13

Interpretación:

- Con solo 2 componentes se conserva más del 95 % de la información
- El error de reconstrucción es bajo

Comparativa y análisis

Clustering vs. Reducción de dimensionalidad

Objetivo	Clustering	Reducción dim.
Descubrir grupos	✓	✗
Visualización	✗	✓
Preprocesamiento	✓	✓
Evaluar estructura	✓ (Silhouette, DBI)	✓ (Varianza, error)

Conclusiones del análisis

- Las métricas de clustering coinciden en mostrar una buena estructura de 3 grupos.
- PCA permitió visualizar adecuadamente la estructura sin pérdida significativa de información.
- Para este dataset, clustering y reducción funcionan de manera complementaria.

Conclusiones

- Las métricas de evaluación son esenciales para validar modelos no supervisados.
- Silhouette, DBI y Calinski–Harabasz ofrecen perspectivas diferentes sobre cohesión y separación.
- La reducción de dimensionalidad mediante PCA es efectiva para visualizar y conservar estructura.
- La combinación de ambas técnicas permite análisis más robustos y mayor interpretabilidad.
- El caso práctico demuestra cómo interpretar métricas de forma crítica y comparativa.

Referencias

IBM. (2024). *What is clustering?* https://www.ibm.com/topics/clustering

IBM. (2024). *Dimensionality reduction explained*. https://www.ibm.com/topics/dimensionality-reduction

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

Scikit-learn. (2024). *Decomposition: PCA*. https://scikit-learn.org/stable/modules/decomposition.html

Towards Data Science. (2023). *Understanding silhouette score*. https://towardsdatascience.com/silhouette-score

Analytics Vidhya. (2024). *Davies–Bouldin index explained*. https://www.analyticsvidhya.com