UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Desarrollo de software

Extracción de Conocimiento en Bases de Datos

IV.2. Métricas de evaluación de modelos (50%)

IDGS91N

PRESENTA:

Juan Carlos Medina Sánchez

DOCENTE:

Enrique Mascote

Chihuahua, Chih., 29 de noviembre de 2025

1. Introducción

En la minería de datos y el aprendizaje automático, los modelos no supervisados desempeñan un papel fundamental para explorar estructuras ocultas dentro de un conjunto de datos. Entre estos modelos destacan los algoritmos de agrupación (clustering), que permiten encontrar grupos naturales dentro de los datos, y los algoritmos de reducción de dimensionalidad, que permiten simplificar conjuntos con muchos atributos conservando la mayor cantidad de información posible.

Para evaluar la calidad de estos modelos se emplean métricas internas que no requieren etiquetas verdaderas. En este documento se explican tres métricas de evaluación para modelos de agrupación y dos métricas para reducción de dimensionalidad, además de aplicar dichas métricas en un caso práctico utilizando el dataset Iris, el algoritmo K-means y la técnica PCA.

2. Investigación de métricas

2.1. Métricas de evaluación para agrupación

En los modelos de clustering no existen etiquetas reales, por lo que se emplean métricas internas basadas en distancias y coherencia de grupos. En este trabajo se seleccionan:

- Índice de silueta
- Índice de Davies–Bouldin
- Índice de Calinski–Harabasz

2.1.1. Índice de silueta

Definición y fórmula

El índice de silueta mide simultáneamente la cohesión del clúster y la separación respecto a otros clústeres.
Para cada punto i:

- $a(i)$: distancia promedio a los puntos de su mismo clúster.
- $b(i)$: distancia promedio al clúster más cercano al que no pertenece.

La silueta individual se define como:

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

El valor global es el promedio de $s(i)$para todos los puntos.

Interpretación

- Valores cercanos a **1** → clústeres bien definidos.
- Cercano a **0** → puntos en frontera.
- Negativo → puntos mal asignados.

Ventajas

- Evalúa cohesión y separación simultáneamente.
- Intuitivo y ampliamente usado.

Limitaciones

- Costoso en datasets grandes.
- Depende de la métrica de distancia utilizada.

2.1.2. Índice de Davies–Bouldin (DB)

Definición y fórmula

El índice Davies–Bouldin mide la relación entre la dispersión dentro del clúster y la distancia entre clústeres:

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

donde:

- s_i: dispersión del clúster i.
- d_{ij}: distancia entre los centroides de los clústeres i y j.
- k: número de clústeres.

Interpretación

- **Bajo** = mejor (clústeres compactos y bien separados).
- **Alto** = mala separación o solapamiento.

Ventajas

- Sencillo y rápido de calcular.
- Útil para comparar diferentes configuraciones de clustering.

Limitaciones

- Sensible al ruido y a la forma de los clústeres.
- No siempre refleja bien estructuras complejas.

2.1.3. Índice de Calinski–Harabasz (CH)

Definición y fórmula

Evalúa la razón entre la varianza entre clústeres y la varianza dentro de ellos.

$$CH = \frac{\text{tr}(B_k)/(k-1)}{\text{tr}(W_k)/(n-k)}$$

donde:

- B_k: dispersión entre clústeres.
- W_k: dispersión intra-clúster.
- n: número de observaciones.

Interpretación

- Valores **altos** → clústeres bien separados y compactos.
- Valores **bajos** → clústeres poco definidos.

Ventajas

- Muy utilizado para elegir el número óptimo de clústeres.
- Computacionalmente eficiente.

Limitaciones

- Sensible a outliers.
- Menos útil si los clústeres no tienen forma esférica.

2.2. Métricas para reducción de dimensionalidad

En métodos como PCA o t-SNE se busca medir cuánta información se conserva al reducir dimensiones.

Se seleccionan:

- Varianza explicada acumulada
- Error de reconstrucción

2.2.1. Varianza explicada acumulada

Definición

En PCA, cada componente principal tiene asociada una varianza.
La varianza explicada acumulada después de $m$componentes es:

$$VE_{acum}(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Interpretación

- Cerca de **1** → se conserva casi toda la información.
- Baja → se pierde mucha variabilidad.

Ventajas

- Simple e intuitiva.
- Útil para decidir cuántos componentes usar.

Limitaciones

- Solo válida para métodos lineales como PCA.
- No necesariamente preserva la estructura de clústeres.

2.2.2. Error de reconstrucción

Definición

Mide qué tan bien se recuperan los datos originales desde la representación reducida:

$$E = \frac{1}{n} \sum_{i=1}^{n} \| x_i - \hat{x}_i \|^2$$

Interpretación

- **Bajo** → la reducción conserva la información.

- **Alto** → se pierde información importante.

Ventajas

- Mide pérdida de información de forma directa.
- Ideal para comparar distintas configuraciones.

Limitaciones

- Depende de la escala.
- No siempre se correlaciona con la utilidad práctica.

3. Caso de estudio: Clustering + PCA con dataset Iris

3.1. Descripción del dataset

Se utilizó el conjunto de datos **Iris**, el cual contiene:

- 150 muestras de flores
- 4 atributos numéricos (largo y ancho de sépalos y pétalos)
- 3 clases reales (no usadas en el clustering)

Los datos se estandarizaron antes del análisis.

3.2. Resultados de Clustering (K-means)

Se aplicó K-means con **k = 3** clústeres.

Resultados de métricas

Métrica	**Valor**
Índice de silueta	**0.4599**
Davies–Bouldin	**0.8336**
Calinski–Harabasz	**241.90**

Interpretación

- Silueta ≈ 0.46: separación moderada entre clústeres.
- DB ≈ 0.83: buena compactación y separación.
- CH ≈ 242: clústeres relativamente compactos.

3.3. Resultados de reducción de dimensionalidad (PCA)

Varianza explicada por componente

- PC1: 0.7296
- PC2: 0.2285
- PC3: 0.0367
- PC4: 0.0052

Varianza explicada acumulada

- 1 componente: 0.73
- **2 componentes: 0.96**
- 3 componentes: 0.995
- 4 componentes: 1.00

Con dos componentes es posible visualizar los datos preservando la mayor parte de la información.

Error de reconstrucción (2 componentes)

- Error promedio: 0.0419

Interpretación:
La pérdida de información es mínima; PCA con dos componentes es adecuado para visualización.

4. Conclusiones

- Las métricas internas permiten evaluar la calidad del clustering sin usar etiquetas.
- En el dataset Iris, K-means con tres clústeres genera resultados razonables según silueta, Davies–Bouldin y Calinski–Harabasz.
- PCA permite reducir dimensiones preservando el 96 % de la varianza con solo dos componentes.
- El error de reconstrucción confirma que la pérdida de información es baja.
- La combinación clustering + PCA es útil para exploración de datos y visualización.

5. Referencias (APA)

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.