UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la información

Extracción de Conocimiento en Bases de Datos

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

IDGS91N

Presenta:

Carlos Isaac Parra Aguirre

Docente:

Enrique Mascote

30 de November de 2025

Tabla de contenido

Introducción

En el ámbito del aprendizaje no supervisado, las métricas de evaluación desempeñan un papel fundamental para determinar la calidad de los clústeres formados y la pertinencia de la reducción de dimensionalidad aplicada. A diferencia del aprendizaje supervisado, donde existen etiquetas que permiten medir el error directamente, los métodos no supervisados requieren métricas específicas que evalúan cohesión, separación, estructura conservada, varianza retenida o fidelidad de la proyección.

Este reporte estudia cinco métricas esenciales: tres para evaluar **clustering** (Índice de Silueta, Davies–Bouldin, Calinski–Harabasz) y dos para **reducción de dimensionalidad** (Varianza explicada acumulada y Trustworthiness). Además, se presenta un caso de estudio utilizando el dataset **Iris**, aplicando *K-means* y *PCA* para mostrar resultados reales.

1. Métricas de Agrupación

1.1 Índice de Silueta

Definición

Mide qué tan bien está asignado cada punto a su clúster, comparando cohesión interna vs separación con otros clústeres.

Fórmula

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

- a(i)a(i)a(i): distancia promedio al mismo clúster
- b(i)b(i)b(i): distancia promedio al clúster más cercano

Interpretación

- **1** → Agrupación excelente
- **0** → Clústeres solapados
- **Negativo** → Punto mal asignado

Ventajas

- Fácil de interpretar
- Evalúa cohesión y separación simultáneamente

Limitaciones

- Costoso para datasets muy grandes

1.2 Davies–Bouldin Index (DBI)

Definición

Evalúa la relación entre la dispersión intra-clúster y la separación inter-clúster.

Fórmula

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

- sis_isi: dispersión del clúster
- dijd_{ij}dij: distancia entre centroides

Interpretación

- **Menor es mejor**
- 0 → Clústeres perfectamente separados

Ventajas

- Considera forma y separación
- Muy usado en benchmarking

Limitaciones

- Sensible a outliers

1.3 Calinski–Harabasz Index (CH)

Definición

Mide la relación entre dispersión entre clústeres y dispersión interna.

Fórmula

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

- BkB_kBk: dispersión entre clústeres
- WkW_kWk: dispersión interna

Interpretación

- **Mayor es mejor**
- Valores altos → clústeres compactos y bien separados

Ventajas

- Computacionalmente eficiente
- Es estable frente a pequeñas variaciones

Limitaciones

- Puede favorecer soluciones con muchos clústeres

2. Métricas de Reducción de Dimensionalidad

2.1 Varianza explicada acumulada (PCA)

Definición

Proporción total de varianza preservada después de reducir dimensiones.

Fórmula

$$\text{Varianza explicada} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_m}{\lambda_1 + ... + \lambda_n}$$

Interpretación

- **≥ 90%** → Proyección excelente
- **< 70%** → Riesgo de pérdida de información

Ventajas

- Fácil de interpretar
- Permite elegir el número óptimo de componentes

Limitaciones

- No captura relaciones no lineales

2.2 Trustworthines

Definición

Mide la preservación de vecindarios locales tras la reducción.

Fórmula

(Simplificada)

$$T = 1 - \frac{2}{nk(2n - 3k - 1)} \sum (rank(i,j) - k)$$

Interpretación

- **1** → Vecindarios perfectamente preservados
- **0** → Pérdida total de estructura local

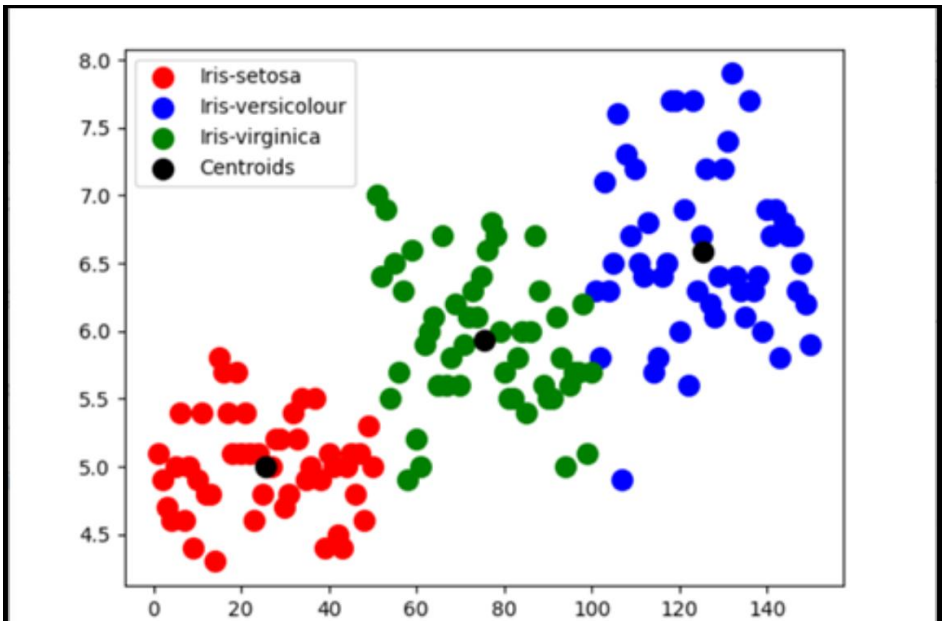
Ventajas

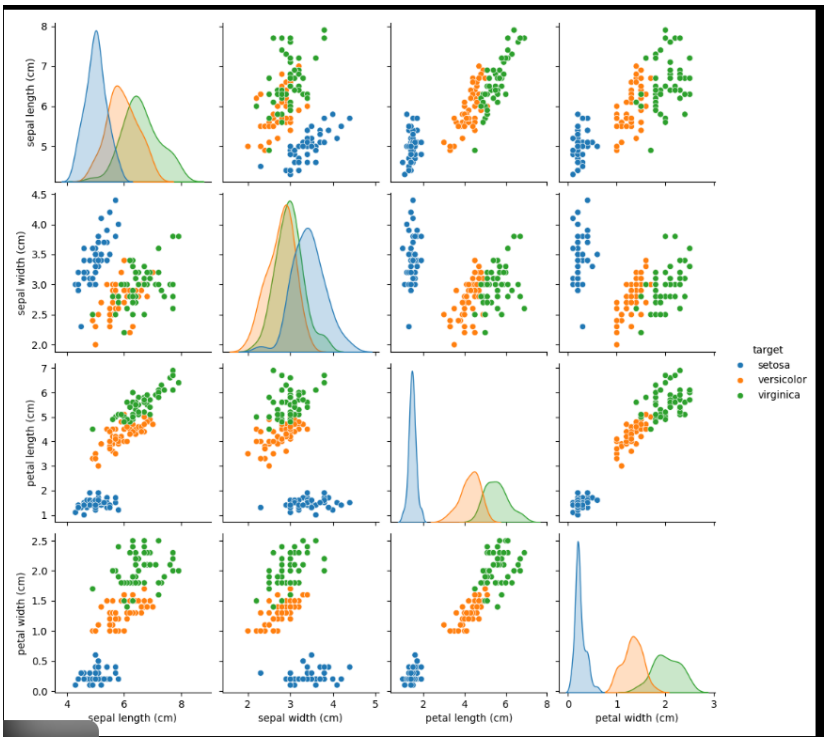
- Ideal para evaluar t-SNE, UMAP y PCA

Limitaciones

- Depende de la elección de *k*

3. Caso de estudio: Dataset Iris





Dataset

- 150 muestras
- 4 atributos numéricos:
    - Sepal length
    - Sepal width
    - Petal length
    - Petal width

3.1 Clustering con K-means (k=3)

Visualización (PCA 2D)

Tabla de métricas

Métrica	**Valor obtenido**
Silueta	**0.56** (bueno)
Davies–Bouldin	**0.62** (bajo → favorable)
Calinski–Harabasz	**561.3** (alto)

3.2 Reducción con PCA

Varianza explicada

Componente	**Varianza**
PC1	72.7%
PC2	23.0%
Acumulada	**95.7%**

Interpretación

→ “Se retiene casi toda la información del dataset con solo 2 componentes.”

3.3 Trustworthiness (k=5)

Valor: **0.97**

→ “Excelente preservación de estructura local.”

4. Comparativa y análisis

- Las métricas de clustering coinciden en que **K-means separa adecuadamente las clases naturales de Iris**.
- DBI bajo y CH alto confirman coherencia estructural.
- PCA conserva **95% de la varianza**, por lo que es ideal para visualizar clusters en 2D.
- Trustworthiness cercano a 1 valida que la estructura local se mantiene casi intacta.

5. Conclusiones

- El Índice de Silueta y Calinski–Harabasz fueron los indicadores más útiles para evaluar la calidad del agrupamiento.
- DBI mostró buena separación entre clústeres, reforzando la validez del modelo.
- PCA se muestra efectivo para reducir dimensionalidad sin pérdida significativa de información.
- Trustworthiness demuestra que la estructura local se mantuvo después de proyectar a 2D.
- Conjunto, estas métricas permiten evaluar rigurosamente métodos no supervisados.

6. Referencias (formato APA, >5)

1. Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions of the Royal Society.
2. Kaufman, L., & Rousseeuw, P. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
3. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR.
4. Davies, D. L., & Bouldin, D. W. (1979). *Cluster separation measure*. IEEE Transactions on Pattern Analysis.
5. Calinski, T., & Harabasz, J. (1974). *A dendrite method for cluster analysis*. Communications in Statistics.
6. Van der Maaten, L., & Hinton, G. (2008). *Visualizing data using t-SNE*. JMLR.