# UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

## Desarrollo y Gestión de Software


## Extracción de Conocimiento en Bases de Datos

### Métricas de evaluación de modelos

## IDGS 91N

PRESENTA:

Chaparro Estrada Hugo Uriel

Jimenez Carrera Carlos Abraham

Jaramillo Flores Brissa Idaly

Armendariz Rodriguez Luis Santiago

DOCENTE:

Enrique Mascote

Chihuahua, Chih., 20 oct 2025

# Introducción

En el contexto del aprendizaje no supervisado, el análisis de datos se orienta a descubrir estructuras o patrones ocultos sin la ayuda de etiquetas predefinidas. A diferencia del aprendizaje supervisado, donde las métricas de desempeño como precisión, recall o F1-score tienen interpretaciones directas asociadas a las etiquetas reales, en el análisis no supervisado no existen "verdades" con las cuales comparar los resultados. Por ello, surge la necesidad de utilizar métricas específicas que evalúen la calidad interna del agrupamiento o la fidelidad estructural de una representación reducida de los datos.

En este sentido, las métricas de evaluación para modelos de agrupación y reducción de dimensionalidad permiten validar si los resultados obtenidos son coherentes, significativos y útiles para propósitos posteriores como visualización, compresión, o análisis exploratorio. Este trabajo presenta una revisión profunda y la aplicación práctica de cinco métricas clave: tres de agrupación y dos de reducción de dimensionalidad, utilizando un conjunto de datos real.

# Métricas de Evaluación para Modelos de Agrupación (Clustering)

Los algoritmos de agrupamiento buscan organizar los datos en grupos donde los elementos dentro de un mismo grupo son similares entre sí, y diferentes respecto a los de otros grupos. Para cuantificar la calidad de estos agrupamientos sin conocimiento previo, se utilizan métricas como el índice de silueta, Davies–Bouldin y Calinski–Harabasz, que combinan conceptos de **cohesión** (compacidad dentro del grupo) y **separación** (distancia entre grupos).

## 1. Índice de Silueta

**Definición:**
El índice de silueta es una métrica que evalúa para cada punto cuán apropiadamente ha sido asignado a su clúster. Considera dos factores: la **distancia media entre el punto y los demás puntos del mismo clúster** (cohesión) y la **distancia mínima del punto a los puntos de cualquier otro clúster** (separación). Formalmente, se define como:

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

donde:

- $a(i)$ a(i): distancia promedio entre el punto $i$ i y todos los otros puntos en su mismo clúster.
- $b(i)$ b(i): mínima distancia promedio entre el punto $i$ i y los puntos de los otros clústeres.

**Interpretación:**

- $s(i) \approx 1$ s(i)≈1: la muestra está bien agrupada.
- $s(i) \approx 0$ s(i)≈0: la muestra está en el límite entre dos clústeres.
- $s(i) < 0$ s(i)<0: el punto está mal asignado.

**Ventajas:**

- Proporciona una interpretación intuitiva a nivel individual y global.

- No requiere etiquetas ni conocimiento previo de clases.
- Permite identificar outliers o puntos ambiguos.

**Limitaciones:**

- Costoso computacionalmente, especialmente con grandes conjuntos de datos.
- Supone clústeres convexos, lo cual puede no ser cierto en todos los casos.

## 2. Índice Davies–Bouldin (DBI)

**Definición:**

El índice Davies–Bouldin evalúa la relación entre la **dispersión interna** de cada clúster y su **distancia respecto a los otros clústeres**. Un valor más bajo indica clústeres más compactos y mejor separados.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}}\right)$$ DB=k1i=1∑kj=imax(dijsi+sj)

donde:

- $s_i$ si: desviación promedio de los puntos del clúster $i$ i respecto a su centro.
- $d_{ij}$ dij: distancia entre los centroides de los clústeres $i$ i y $j$ j.
- k: número total de clústeres.

**Interpretación:**

- **Valores cercanos a 0** indican clústeres compactos y bien separados.
- **Valores altos** indican solapamiento y poca cohesión.

**Ventajas:**

- Fácil de calcular.
- Proporciona una visión general del modelo de agrupación.
- No necesita etiquetas reales.

**Limitaciones:**

- Sensible a la escala de los datos.
- Puede ser influenciado por outliers o clústeres de tamaños dispares.

## 3. Índice Calinski–Harabasz (CH)

**Definición:**
También llamado "criterio de razón de varianza", este índice se basa en la relación entre la varianza entre clústeres y la varianza dentro de los clústeres. Se calcula como:


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

donde:

- $Tr(B_k)$: traza de la matriz de dispersión entre clústeres.
- $Tr(W_k)$: traza de la matriz de dispersión dentro de los clústeres.
- $n$: número de muestras, $k$: número de clústeres.

**Interpretación:**

- **Valores altos** indican una separación clara entre clústeres y baja variabilidad interna.
- **Valores bajos** indican solapamiento entre clústeres o mala agrupación.

**Ventajas:**

- Muy eficiente computacionalmente.
- No requiere etiquetas de clase.
- Ampliamente adoptado por su simplicidad.

**Limitaciones:**

- Tiende a favorecer configuraciones con más clústeres.
- Asume homogeneidad y esfericidad en los datos.

# Métricas de Evaluación para Modelos de Reducción de Dimensionalidad

Los algoritmos de reducción de dimensionalidad buscan representar los datos en un espacio de menor dimensión preservando su estructura y relaciones. La evaluación se enfoca en cuánta **información se conserva** (varianza) y **qué tan bien se preservan las relaciones locales** (vecindarios).

### 1. Varianza Explicada Acumulada

**Definición:**

En métodos como PCA, la varianza explicada acumulada es la proporción total de la varianza original que se conserva al proyectar los datos en un número reducido de componentes principales.


$$\text{Varianza acumulada} = \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{i=1}^{D} \lambda_i}$$

donde:

- $\lambda_i$: valor propio del componente $i$.
- $D$: número total de componentes originales, $d$: componentes seleccionados.

**Interpretación:**

- Si la varianza explicada acumulada es ≥ 90%, significa que casi toda la información se conserva.
- Si es baja, hay pérdida significativa de datos.

**Ventajas:**

- Muy útil para determinar el número óptimo de dimensiones.
- Ofrece una métrica directa sobre la calidad de la representación.

**Limitaciones:**

- Solo aplicable a métodos lineales como PCA.
- No refleja la conservación de la estructura local ni la geometría de los datos.

## 2. Trustworthiness

**Definición:**

La métrica trustworthiness evalúa qué tan fiel es el espacio reducido respecto a los vecindarios del espacio original. Mide cuántos vecinos “falsos” aparecen en la reducción.


$$T = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_k(i)} (r(i,j) - k)$$

donde:

- Uk(i): conjunto de vecinos que aparecen en el espacio reducido, pero no en el original.
- r(i,j): rango de j respecto a i en el conjunto original.
- k: número de vecinos más cercanos considerados.

**Interpretación:**

- **T ≈ 1** indica buena conservación local.
- **T < 0.9** puede reflejar distorsión.

**Ventajas:**

- Muy adecuado para evaluar t-SNE, Isomap y UMAP.
- Se enfoca en la preservación de relaciones locales.

**Limitaciones:**

- Depende del parámetro k, que debe seleccionarse cuidadosamente.
- No evalúa la estructura global ni la varianza.

## Caso de Estudio: Iris Dataset

### Descripción del conjunto de datos:

- Dataset clásico disponible en UCI Machine Learning Repository.
- 150 observaciones de flores de 3 especies.
- 4 atributos numéricos: largo y ancho de pétalos y sépalos.

### Proceso aplicado:

- **Clustering:** K-means con k=3
- **Reducción de dimensionalidad:** PCA (2 componentes), t-SNE (2D)
- **Evaluación:** cálculo de las cinco métricas anteriores

# Resultados Obtenidos

### Evaluación de Agrupamiento

| Métrica | Valor | Interpretación |
|---|---|---|
| Índice de Silueta | 0.55 | Buena cohesión y separación |
| Davies–Bouldin | 0.68 | Clústeres separados y compactos |

| | | |
|---|---|---|
| Calinski–Harabas z | Separación | clústeres |
| 561.6 3 | significativa entre | |

**Evaluación de Reducción de Dimensionalidad**

| Métrica | Valor | Interpretación |
|---|---|---|
| Varianza explicada | 95.8 % | Alta conservación de información |
| Trustworthiness (t-SNE) | 0.96 | Buena preservación de vecindarios |

**Gráficos de apoyo:**

- Gráfico 2D de clústeres tras PCA y t-SNE
- Scree plot con porcentaje de varianza explicada
- Visualización de vecinos para trustworthiness

# Comparativa y Análisis Crítico

- **Silueta y CH** reflejan correctamente la calidad del clustering sobre datos bien definidos como Iris.
- **DBI** aporta una visión balanceada entre cohesión y separación.
- **PCA** es excelente para visualizar estructura lineal y conserva gran parte de la información.
- **Trustworthiness** es fundamental cuando se prioriza la fidelidad local sobre la global.

**Recomendación:** Usar múltiples métricas en conjunto permite una evaluación más robusta y adaptada a los objetivos del análisis.

## Conclusiones

- La evaluación de modelos no supervisados debe considerar diferentes perspectivas: cohesión interna, separación externa y fidelidad estructural.
- Ninguna métrica por sí sola es suficiente; la combinación de indicadores complementarios es esencial.
- La correcta selección de métricas depende del algoritmo y del objetivo final: compresión, exploración, clasificación posterior o visualización.
- Para asegurar interpretabilidad y robustez, se recomienda normalizar los datos, validar con distintas inicializaciones y complementar con visualizaciones.

## Referencias


1. Tan, P. N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.
2. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
3. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579-2605.
4. Scikit-learn documentation. (2024). *Clustering performance evaluation*. https://scikit-learn.org/stable/modules/clustering.html
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
6. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.