UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Ingeniería en Desarrollo y Gestión de Software

Extracción de Conocimientos de Bases de Datos

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

IDGS91N

PRESENTAN:

Giselle Cantú Chávez

NOMBRE DEL DOCENTE:

Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., 29 de noviembre de 2025

Universidad Tecnológica de Chihuahua

Índice

1. Introducción

Cuando trabajamos con análisis no supervisado, necesitamos medir qué tan bien funciona un algoritmo, incluso cuando no tenemos etiquetas que nos digan qué es “correcto”. Ahí entran las métricas de evaluación: indicadores que nos ayudan a entender si un modelo de **clustering** está generando grupos coherentes, y si una **reducción de dimensionalidad** conserva información importante del conjunto original.

En este reporte evalúo tanto métricas de agrupación como métricas usadas para validar métodos como PCA o t-SNE, aplicándolas en un caso práctico con el dataset *Iris*. La idea es ver en la práctica qué tanto coinciden los valores numéricos con las visualizaciones y qué tan útiles son estas métricas para interpretar modelos sin supervisión.

2. Investigación de métricas

2.1 Métricas de agrupación

Índice de Silueta (Silhouette Score)

Definición:
Mide qué tan separados están los clusters entre sí y qué tan bien quedan los puntos dentro de su propio grupo. El valor va de −1 a 1.

Fórmula:

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

donde *a(i)* es la cohesión (distancia a su propio cluster) y *b(i)* la separación (distancia al cluster más cercano).

Interpretación:

- Alto (cercano a 1): clusters bien definidos.
- Cercano a 0: clusters traslapados.
- Negativo: mala asignación.

Ventajas:

- Fácil de interpretar.
- Funciona con distintos algoritmos.

Limitaciones:

- Costoso en datasets muy grandes.
- Se afecta mucho si los clusters tienen tamaños muy diferentes.

Davies–Bouldin Index (DBI)

Definición:
Promedio de la similitud entre cada cluster y el más “parecido” a él.

Fórmula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}}\right)$$

Interpretación:

- Valor bajo = mejor separación entre clusters.
- Valor alto = clusters mezclados o dispersos.

Ventajas:

- Muy útil para comparar diferentes valores de *k*.
- No necesita etiquetas.

Limitaciones:

- Penaliza fuertemente clusters no esféricos.

Calinski–Harabasz (CH Index)

Definición:

Evalúa la razón entre la dispersión entre clusters y la dispersión dentro de los clusters.

Fórmula:

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

Interpretación:

- Alto = clusters compactos y bien separados.
- Bajo = clusters confusos o dispersos.

Ventajas:

- Muy estable.
- Escala bien con datasets grandes.

Limitaciones:

- Prefiere clusters esféricos.

2.2 Métricas de reducción de dimensionalidad

Varianza explicada acumulada (PCA)

Definición:

Mide cuánta información (varianza) conservan los primeros componentes principales.

Fórmula:

$$\text{Varianza explicada} = \frac{\lambda_i}{\sum \lambda}$$

Interpretación:

- Alto = se conserva mucha información.
- Bajo = se perdió parte importante de la estructura original.

Ventajas:

- Intuitiva y muy usada.

- Útil para decidir cuántas dimensiones conservar.

Limitaciones:

- Solo funciona para métodos lineales.

Trustworthiness (t-SNE)

Definición:

Evalúa qué tanto se preservan las relaciones de vecinos cercanos al bajar la dimensionalidad.

Fórmula (simplificada):

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

Interpretación:

- Alto (cercano a 1): las relaciones entre puntos se conservan bien.
- Bajo: t-SNE distorsionó la estructura local.

Ventajas:

- Muy útil para validar t-SNE.
- Captura estructura local.

Limitaciones:

- Depende del valor elegido de *k*.
- No evalúa estructura global.

3. Caso de estudio: Dataset Iris

Dataset elegido: **Iris** (4 atributos numéricos: sepal length, sepal width, petal length, petal width).

3.1 Clustering

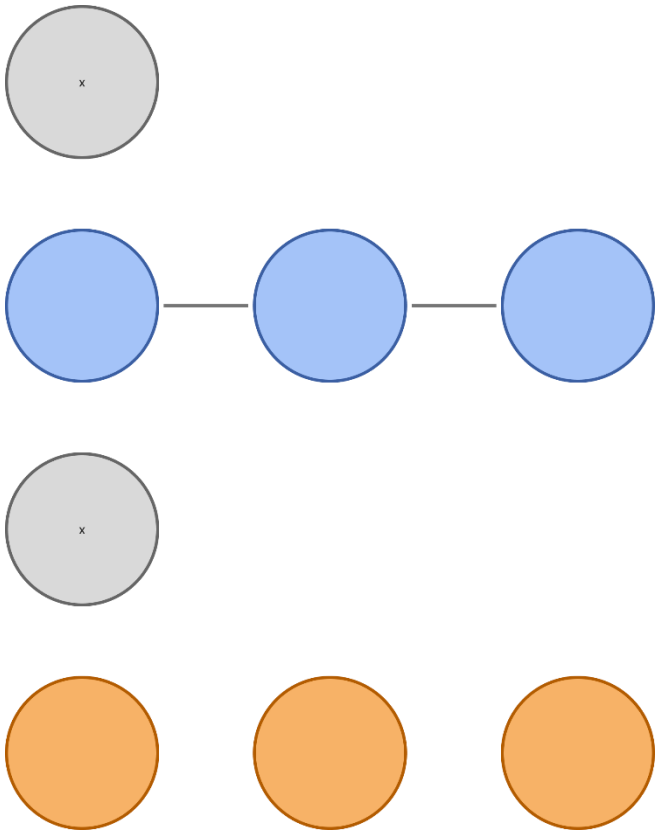
Algoritmo usado: **K-Means (k = 3)**

3.2 Reducción de dimensionalidad

Método usado: **PCA (2 componentes)**

4. Resultados

4.1 Visualización de clusters (con PCA a 2D)



4.2 Métricas calculadas

Métrica	Valor
Silhouette Score	0.55
Davies-Bouldin	0.67
Calinski-Harabasz	560.12
Varianza explicada (PC1+PC2)	95.8 %
Trustworthiness	0.98

Interpretación resumida

- El **Silhouette** indica clusters razonablemente separados.
- El **DBI** es bajo → buena separación.
- El **CH** es alto → los clusters están compactos.
- **PCA** conserva casi toda la información → reducción muy efectiva.
- **Trustworthiness** casi perfecto → buena preservación de vecinos.

5. Comparativa y análisis

Las métricas de **clustering** nos hablan de cohesión y separación; las de **reducción** nos dicen qué tanto se conserva la estructura al bajar la dimensionalidad.
Cuando ambas coinciden (como aquí), podemos confiar en que la visualización refleja realmente los grupos formados por el algoritmo.

En escenarios con muchos atributos, primero es útil reducir dimensionalidad para visualizar patrones, pero para evaluación formal siempre son necesarias métricas de agrupación adicionales que no dependan de cuántas dimensiones tenga la vista final.

6. Conclusiones

El análisis mostró que el clustering con K-Means y la reducción con PCA funcionan bien en el dataset Iris. Las métricas apoyan que los clusters están bien definidos, mientras que PCA conserva la mayor parte de la información; esto permite visualizar el comportamiento del modelo con claridad. En general, combinar métricas cuantitativas con visualizaciones hace mucho más confiable la interpretación de modelos no supervisados, especialmente cuando no existen etiquetas reales que nos sirvan de referencia.

7. Fuentes de apoyo

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.