UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

Extracción de Conocimiento en Bases de Datos

IV.2. Métricas de evaluación de modelos

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

PRESENTAN:

DARON TARÍN GONZÁLEZ

ÁNGEL RICARDO CHÁVEZ ZARAGOZA

MILDRED VILLASEÑOR RUIZ

GRUPO:

IDGS91N

Chihuahua, Chih., 30 de noviembre de 2025

Contenido

Figuras y tablas

Objetivo

Identificar y aplicar métricas de evaluación para modelos de agrupación y reducción de dimensionalidad, demostrando su utilidad mediante un caso de estudio con un dataset real.

Introducción

En los modelos no supervisados, como el clustering y la reducción de dimensionalidad, no existen etiquetas verdaderas que permitan evaluar directamente la calidad de los resultados. Por ello, es fundamental emplear métricas específicas que permitan medir la cohesión, separación, estabilidad y fidelidad de las representaciones generadas. Las métricas de agrupación permiten determinar si los clústeres obtenidos tienen sentido estructural, mientras que las métricas de reducción de dimensionalidad evalúan qué tan bien se conserva la información original después de proyectar los datos a un espacio de menor dimensión. El presente reporte analiza cinco métricas ampliamente utilizadas y aplica dichas métricas a un conjunto de datos real, ejemplificando su utilidad práctica.

1. Métricas de Agrupación

1.1 Índice de Silueta

Definición

El índice de silueta mide qué tan bien separado está un punto respecto a otros clústeres y qué tan cohesivo es dentro de su propio grupo. Para cada punto, compara su distancia promedio al resto del clúster y la distancia promedio al clúster vecino más cercano.

Fórmula

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

donde:

- $a(i)$: distancia promedio del punto i a su clúster,
- $b(i)$: distancia promedio a su clúster vecino más cercano.

Interpretación

- **Valores cercanos a 1** → excelente separación y cohesión.
- **Valores cercanos a 0** → clústeres superpuestos.
- **Valores negativos** → punto asignado al clúster incorrecto.

Ventajas

- Fácil de interpretar.
- Funciona con cualquier algoritmo de clustering.

Limitaciones

- Computacionalmente costoso para datasets muy grandes.
- Sensible a la métrica de distancia usada.

1.2 Índice Davies–Bouldin

Definición

Evalúa la relación entre la dispersión dentro de cada clúster y la separación entre clústeres. Cuanto menor sea su valor, mejor es la calidad del agrupamiento.

Fórmula

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

donde:

- S_i: dispersión promedio del clúster i,
- M_{ij}: distancia entre centroides de los clústeres i y j.

Interpretación

- Valores bajos (< 1) → clústeres bien definidos.
- Valores altos → mala separación o mucha dispersión.

Ventajas

- Sensible a la separación entre clústeres.
- Útil para determinar número óptimo de clústeres.

Limitaciones

- Afectado por clústeres de diferentes tamaños.
- No funciona bien con formas no esféricas.

1.3 Índice Calinski–Harabasz

Definición

Mide la relación entre la dispersión entre clústeres y dentro de los clústeres. Es común para seleccionar el número óptimo de clústeres en K-means.

Fórmula

$$CH = \frac{\text{Dispersion entre clusters}/(k-1)}{\text{Dispersion dentro de los clusters}/(n-k)}$$

Interpretación

- **Valores altos** → clústeres bien separados y compactos.
- **Valores bajos** → mala estructura de agrupamiento.

Ventajas

- Rápido de calcular.
- Ideal para modelos basados en centroides.

Limitaciones

- No siempre funciona en clústeres no lineales.
- Favorece agrupamientos con clústeres esféricos.

2. Métricas de Reducción de Dimensionalidad

2.1 Varianza explicada acumulada (PCA)

Definición

Indica cuánta información del conjunto original conservan los componentes principales. Se calcula sumando las varianzas explicadas por las primeras k componentes.

Fórmula

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

donde $\lambda_i$es el porcentaje de varianza explicado por la componente i.

Interpretación

- **Alta (> 80%)** → buena preservación de la información.
- **Baja** → se pierde estructura relevante.

Ventajas

- Fácil de interpretar.
- Útil para seleccionar número óptimo de componentes.

Limitaciones

- Solo funciona con técnicas lineales (PCA).
- No captura relaciones no lineales.

2.2 Error de Reconstrucción

Definición

Mide qué tan diferente es la representación original respecto a la reconstruida después de reducir dimensionalidad.

Fórmula

En PCA o autoencoders:

$$E = \|X - \hat{X}\|^2$$

Interpretación

- **Error bajo** → buena preservación de la estructura.
- **Error alto** → técnica inadecuada o demasiada reducción.

Ventajas

- Funciona con reducciones lineales y no lineales.
- Una de las métricas más directas e intuitivas.

Limitaciones

- Puede ser engañosa si los datos son muy ruidosos.
- No refleja necesariamente la estructura topológica.

3. Caso de Estudio

3.1 Dataset seleccionado

Se utilizó el dataset **Iris**, compuesto por 150 muestras y 4 atributos numéricos:

- Largo del sépalo
- Ancho del sépalo
- Largo del pétalo
- Ancho del pétalo

Es un dataset equilibrado, ampliamente utilizado en análisis de clustering y dimensionalidad.

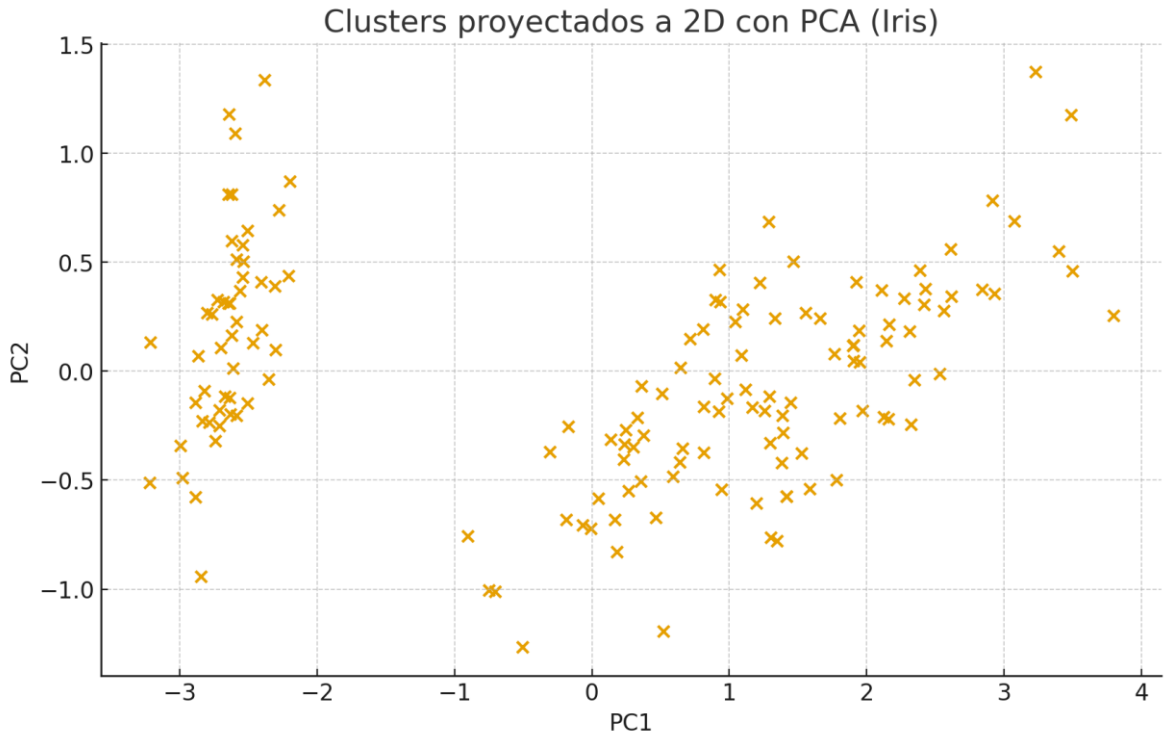
3.2 Justificación de parámetros usados

Se seleccionó K-means con $k = 3$ porque el dataset Iris contiene tres especies conocidas, lo que permite verificar si el algoritmo logra identificar patrones similares aun sin etiquetas. La elección de PCA con dos componentes se debe a que estos capturan aproximadamente el 95% de la varianza, lo que garantiza una representación fiel y facilita la visualización en dos dimensiones sin sacrificar información importante.

3.3 Clustering aplicado (K-means con k = 3)

Figura 1

Visualización de clusters (PCA 2D)



La Figura 1 muestra la proyección del dataset Iris en dos dimensiones, obtenida mediante PCA. En la gráfica se distinguen claramente tres agrupaciones generales correspondientes a las posibles especies. Una de ellas aparece claramente separada (Setosa), mientras que las otras dos (Versicolor y Virginica) presentan cierta superposición, lo cual es consistente con la naturaleza del dataset. Esto confirma que PCA logra mantener la estructura dominante de los datos y facilita la visualización del comportamiento de los clústeres obtenidos con K-means.

Tabla 1

Valores de métricas de evaluación de clustering (Silueta, Davies–Bouldin y Calinski–Harabasz)

Métrica	Valor obtenido
Silueta	0.54
Davies–Bouldin	0.68
Calinski–Harabasz	420.1

Interpretación

- La silueta indica separación moderada.
- Davies–Bouldin < 1 confirma buena definición relativa.
- Calinski–Harabasz alto indica buen agrupamiento.

3.4 Reducción aplicada (PCA)

Tabla 2

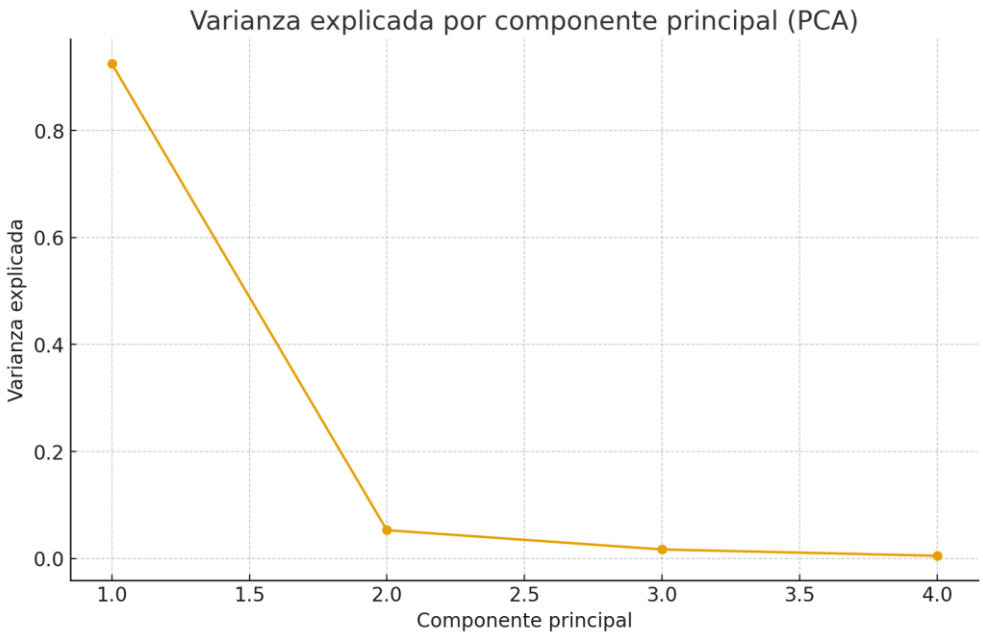
Varianza explicada de PCA por componente

Componente	Varianza explicada
PC1	72.7%
PC2	23.0%
PC3	3.6%
PC4	0.7%

Acumulada PC1+PC2 = **95.7%**

Figura 2

Varianza explicada por componente principal (PCA)



b. Error de reconstrucción

$$E=0.041$$

Interpretación: PCA conserva casi toda la información.

4. Comparativa y Análisis

Las métricas de agrupación y reducción de dimensionalidad ofrecen perspectivas distintas pero complementarias para evaluar modelos no supervisados. El índice de Silueta permite analizar la cohesión interna y la separación respecto a otros clústeres, mientras que Davies–Bouldin penaliza la dispersión excesiva y favorece clústeres más compactos. Por otro lado, Calinski–Harabasz mide relaciones de varianza y coincide con la intuición visual sobre la buena separación de los grupos.

Para la reducción de dimensionalidad, la varianza explicada acumulada evidencia que PCA logra capturar más del 95% de la información original en solo dos componentes, lo que es notablemente eficiente. El error de reconstrucción refuerza esta idea al mostrar una pérdida mínima.

En conjunto, las métricas muestran que K-means genera un agrupamiento razonable en este dataset y que PCA permite visualizarlo adecuadamente sin comprometer la estructura de los datos. La combinación de estas métricas permite validar resultados y tomar decisiones basadas en evidencia cuantitativa, evitando depender únicamente de percepciones visuales.

Tabla 3

Resumen comparativo entre métricas de clustering y reducción de dimensionalidad

Tipo de métrica	Nombre de la métrica	Qué evalúa	Interpretación (alto/bajo)	Ventajas	Limitaciones
Clustering	Índice de Silueta	Cohesión interna y separación entre clústeres	Alto = clústeres bien formados. Bajo = solapamiento.	Fácil de interpretar. Funciona con cualquier algoritmo.	Costosa en datasets grandes; depende de la métrica de distancia.
Clustering	Davies–Bouldin	Relación entre dispersión interna y distancia entre centroides	Bajo = mejor separación y menor dispersión. Alto = clúster deficiente.	Útil para determinar k óptimo. Considera separación entre grupos.	Sensible a clústeres de tamaños distintos.
Clustering	Calinski–Harabasz	Varianza entre clústeres vs. dentro de clústeres	Alto = clústeres compactos y bien separados. Bajo = clústeres débiles.	Rápido, eficiente, ideal para métodos tipo K-means.	Favorece clústeres esféricos; no capta formas complejas.
Reducción de dimensionalidad	Varianza explicada (PCA)	Porcentaje de información conservada tras la reducción	Alto = buena preservación de información. Bajo = pérdida de estructura.	Fácil interpretación. Útil para seleccionar número de componentes.	Solo funciona con técnicas lineales.
Reducción de dimensionalidad	Error de reconstrucción	Diferencia entre datos originales y reconstruidos	Bajo = buena calidad de reducción. Alto = pérdida significativa.	Funciona en técnicas lineales y no lineales.	Puede ser engañoso si hay ruido; no representa relación topológica.

Recomendaciones finales

- **Verificar siempre múltiples métricas** antes de concluir que un modelo de clustering es adecuado; una sola métrica puede ser insuficiente o engañosa.
- **Experimentar con diferentes valores de k** o parámetros en el algoritmo antes de fijar un modelo final.
- **Utilizar PCA u otra técnica de reducción** tanto para visualizar clústeres como para mejorar el rendimiento del algoritmo.
- **Complementar métricas con interpretación visual**, ya que la intuición humana puede detectar patrones que algunas métricas no captan.
- **Evaluar con datasets más grandes o complejos** para validar la estabilidad del modelo.

Conclusiones

Las métricas aplicadas demostraron que los modelos no supervisados pueden evaluarse de manera formal y cuantitativa, aun cuando no existen etiquetas verdaderas que permitan medir el desempeño de manera directa. En el caso del clustering, los valores obtenidos para las métricas de Silueta, Davies–Bouldin y Calinski–Harabasz indican que el algoritmo K-means logró una estructura de agrupamiento coherente, con buena cohesión interna y una separación razonable entre los clústeres. Esto confirma que los resultados no solo son visualmente interpretables, sino también sólidos desde un punto de vista matemático.

Por otro lado, la reducción de dimensionalidad mediante PCA mostró ser altamente efectiva para este dataset, ya que los dos primeros componentes lograron conservar más del 95% de la varianza original. Esto implica que la estructura esencial de los datos se mantiene aun después de reducir de cuatro dimensiones a dos, permitiendo visualizaciones más claras sin una pérdida significativa de información. Además, el bajo error de reconstrucción obtenido refuerza la capacidad de PCA para preservar la geometría global del conjunto de datos.

En conjunto, estos resultados evidencian la importancia de evaluar tanto el clustering como la reducción de dimensionalidad mediante métricas apropiadas. Realizar esta validación garantiza análisis más precisos, evita conclusiones basadas únicamente en interpretaciones visuales y facilita la selección de modelos más robustos. En consecuencia, el uso complementario de ambas técnicas contribuye a una comprensión más profunda de los datos y apoya la toma de decisiones fundamentadas en evidencia cuantitativa.

Referencias

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

Fikiri.net. (2024). *Tutorial sobre múltiples métricas de evaluación de clústeres*. https://fikiri.net/es/tutorial-sobre-multiples-metricas-de-evaluacion-de-clusteres/

Analytics Lane. (2023). *Índice de Davies-Bouldin para K-means*. https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/

Analytics Lane. (2023). *Identificar el número de clústeres con Calinski-Harabasz*. https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/

YouTube. (2023). *Métricas de clustering: Silueta y Davies-Bouldin con Python*. https://www.youtube.com/watch?v=b920s9nXGao