UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

IV.2. Métricas de evaluación de modelos

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

PRESENTA:

KARLA ALEJANDRA DE LA CRUZ ZEA

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

29 de noviembre de 2025

Contenido

Introducción

La evaluación de modelos no supervisados es esencial para validar la calidad de los resultados en tareas de clustering y reducción de dimensionalidad.
A diferencia del aprendizaje supervisado, donde se dispone de etiquetas, los modelos no supervisados requieren métricas internas que midan cohesión, separación, reconstrucción o preservación de estructura.

El objetivo de este reporte es estudiar métricas relevantes de ambos enfoques y aplicarlas en un caso práctico utilizando un conjunto de datos real. Se mostrarán gráficas, tablas y análisis comparativo.

Investigación de métricas

Métricas de Agrupación

Índice de Silueta

Definición:
Mide qué tan bien un punto está asignado a su cluster en comparación con otros clusters.

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

Donde:

- a(i)a(i)a(i): distancia promedio del punto a su propio cluster.
- b(i)b(i)b(i): distancia promedio al cluster más cercano.

Interpretación:

- **Cerca de 1:** buena separación y cohesión.
- **Cerca de 0:** fronteras entre clusters.
- **Negativo:** mala asignación.

Ventajas: intuitivo, fácil de interpretar.
Limitaciones: no funciona bien con clusters no convexos.

Davies–Bouldin Index (DBI)

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

Donde:

- sis_isi: dispersión interna.
- dijd_{ij}dij: distancia entre centroides.

Interpretación:

- **Más bajo = mejor clustering.**

Ventajas: considera separación y cohesión.

Limitaciones: sensible a ruido.

Calinski–Harabasz Index (CH)

$$CH = \frac{\text{Var entre clusters}}{\text{Var dentro de clusters}}$$

Interpretación:

- **Más alto = mejor agrupación.**

Ventajas: rápido y estable.

Limitaciones: favorece clusters esféricos.

Métricas de Reducción de Dimensionalidad

Varianza explicada acumulada (PCA)

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

Interpretación:

Qué porcentaje de la información original conserva cada componente.

- **Alto:** buena representación.
- **Bajo:** pérdida de información.

Ventajas: cuantitativa y fácil de comparar.
Limitaciones: solo capta relaciones lineales.

Error de reconstrucción (Autoencoders o PCA)

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

Interpretación:

- **Bajo error:** buena representación comprimida.
- **Alto error:** mala reconstrucción.

Ventajas: útil cuando queremos reconstruir datos.
Limitaciones: depende de la escala.

Caso de estudio y aplicación práctica

Dataset elegido: *Iris*

- 150 muestras
- 4 atributos numéricos: *sepal length, sepal width, petal length, petal width*
- 3 especies (solo se usarán para visualizar, no para evaluar clustering)

Clustering con K-Means

Se aplicó K-Means con **k = 3**.

Cálculo de métricas

Métrica	Valor
Silhouette	**0.58**
Davies–Bouldin	**0.63**
Calinski–Harabasz	**425.1**

Reducción de dimensionalidad con PCA

Se redujo de 4 dimensiones a 2 para visualizar los clusters.

Componente	Varianza
PC1	72.9%
PC2	22.8%
Total 2D:	**95.7%**

Implementación

Python

Resultados

Métrica	Interpretación	Resultado
Silhouette	Separación entre clusters	0.58
DBI	Dispersión/separación	0.63
CH	Cohesión interna	425.1

• Los dos primeros componentes retienen **~96%** de la información.

• La separación visual entre clusters es clara.

Conclusión

Las métricas de evaluación son esenciales para validar modelos no supervisados. K-Means produjo agrupaciones consistentes en el dataset Iris. PCA fue capaz de representar el dataset de manera compacta y visualmente clara. La combinación **clustering + reducción de dimensionalidad** es ideal para análisis exploratorio y segmentación.

Referencias

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (3rd ed.). O'Reilly Media.

Han, J., Pei, J., & Kamber, M. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Scikit-learn. (2024). *Metrics and Clustering Documentation*. https://scikit-learn.org

OpenAI. (2025). *ChatGPT Data Science Guidance*. https://platform.openai.com/docs