

UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA
TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN



IV.2. MÉTRICAS DE EVALUACIÓN DE MODELOS

MATERIA: Extracción de Conocimiento en Bases de Datos

MAESTR@: Enrique Mascote

ALUMNO: Carlos Adrián Mata Nevárez

Matricula: 1123250015

FECHA: 29/11/2025

ÍNDICE

INTRODUCCIÓN	1
INVESTIGACIÓN DE MÉTRICAS	2
Índice de Silueta (Silhouette Score).....	2
Índice de Davies-Bouldin	2
Índice de Calinski-Harabasz	2
REDUCCIÓN DE DIMENSIONALIDAD	3
Varianza Explicada Acumulada (VEE).....	3
Error de Reconstrucción	3
CASO DE ESTUDIO Y APLICACIÓN PRÁCTICA	4
Descripción del Conjunto de Datos: Iris	4
Aplicación de Clustering: K-means (k=3).....	4
Aplicación de Reducción de Dimensionalidad: PCA.....	6
Resultados de Reducción: Gráfica de Varianza Explicada	6
Conclusiones del Caso de Estudio	8
REFERENCIAS	9

INTRODUCCIÓN

El Análisis de Datos No Supervisado constituye un pilar fundamental en la extracción de conocimiento, abordando la identificación de estructuras y patrones en conjuntos de datos que carecen de etiquetas previas. Dentro de este campo, los Algoritmos de Agrupación (Clustering) y los métodos de Reducción de Dimensionalidad son herramientas esenciales.

El Clustering se enfoca en dividir un conjunto de datos en subgrupos coherentes (clusters), maximizando la similitud entre los elementos dentro de un mismo grupo y la disimilitud entre grupos diferentes. Esto permite la segmentación de clientes, la detección de anomalías y el descubrimiento de categorías naturales en los datos. Por otro lado, la Reducción de Dimensionalidad tiene como objetivo simplificar los datos eliminando características redundantes o poco informativas, mitigando la "maldición de la dimensionalidad". Esto mejora la eficiencia computacional, reduce el ruido y facilita la visualización de datos complejos.

Este reporte tiene como objetivo principal identificar y describir los algoritmos más representativos en ambas áreas (como K-means, DBSCAN, PCA y t-SNE). Para cada uno, se detallará su principio de funcionamiento, sus parámetros clave, sus ventajas, sus limitaciones y un ejemplo de aplicación conceptual simple. Finalmente, se presentará una comparativa crítica para establecer los criterios de elección entre las técnicas de agrupación y reducción, ofreciendo una visión integral de su aplicación en la ciencia de datos moderna.

INVESTIGACIÓN DE MÉTRICAS

Índice de Silueta (Silhouette Score)

El Índice de Silueta mide cohesión intra-cluster ($a(i)$: distancia media a otros puntos del mismo cluster) versus separación al cluster vecino más cercano ($b(i)$), con $S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ en rango $[-1,1]$. Valores altos (~ 1) indican clusters densos y bien separados; ~ 0 sugiere límites ambiguos; negativos implican mala asignación; promedio global evalúa clustering completo (wikipedia, 2025).

Índice de Davies-Bouldin

Mide similitud promedio entre cada cluster y su más similar: $DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{s_i + s_j}{d_{ij}}$, donde s es dispersión intra-cluster y d_{ij} distancia entre centroides. Valores bajos indican mejor clustering (clusters compactos y separados); útil para seleccionar k óptimo vía mínimo global (scikit learn, 2025).

Índice de Calinski-Harabasz

Calcula ratio varianza entre-clusters (SS_B) sobre intra-clusters (SS_W): $CH = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1}$, donde N es número de puntos y k clusters. Valores altos favorecen particiones con dispersión inter alta y compacta intra; ideal para k-means y encontrar k óptimo maximizando CH (mathworks, 2025).

REDUCCIÓN DE DIMENSIONALIDAD

Varianza Explicada Acumulada (VEE)

La VEE mide la proporción de varianza total retenida por los k componentes principales: $\sum_{i=1}^k \frac{\lambda_i}{\sum_{j=1}^p \lambda_j}$, donde λ_i son autovalores de la matriz de covarianza.

Valores cercanos a 1 (95-99%) indican preservación efectiva de información; se usa para seleccionar dimensionalidad óptima vía gráfica acumulativa o umbral fijo (Rodríguez, 2025).

Ventajas: Exacta para PCA, guía selección de k , computacionalmente eficiente.

Limitaciones: Solo lineal; ignora estructuras no lineales capturadas por t-SNE.

Error de Reconstrucción

Calcula MSE entre datos originales x_i y reconstruidos \hat{x}_i : $E_{rec} = \frac{1}{N} \sum \|x_i - \hat{x}_i\|^2$, proyectando a baja dimensión y regresando vía pseudoinversa (PCA) o decodificador (autoencoders). Bajo E_{rec} (~ 0) señala fidelidad alta; alto indica pérdida significativa.

Ventajas: Versátil para lineal/no lineal; mide pérdida directa de información.

Limitaciones: Escala-dependiente (requiere normalización); enfocado en euclidiana, no locales (fastercapital, 2025).

CASO DE ESTUDIO Y APLICACIÓN PRÁCTICA

Descripción del Conjunto de Datos: Iris

El Dataset Iris es un conjunto de datos clásico y ampliamente utilizado en el aprendizaje automático.

- Objetivo Original: Clasificar flores de lirio en tres especies.
- Instancias: 150 observaciones (50 por cada una de las tres especies: Setosa, Versicolor y Virginica).
- Atributos Numéricos (4): El dataset cumple con el requisito de tener al menos cuatro atributos numéricos:
 1. Largo del Sépalo (cm)
 2. Ancho del Sépalo (cm)
 3. Largo del Pétalo (cm)
 4. Ancho del Pétalo (cm)

En el caso de estudio buscamos aplicar técnicas no supervisadas sobre estos cuatro atributos y evaluar los resultados utilizando las métricas definidas previamente.

Aplicación de Clustering: K-means (k=3)

Se aplicó el algoritmo K-means asumiendo $k=3$ (el número de especies conocidas, simulando un resultado de segmentación). El algoritmo se ejecutó sobre las cuatro dimensiones originales.

Resultados de Clustering: Tabla de Métricas

Los valores de las métricas de agrupación fueron calculados para evaluar la calidad intrínseca de los $k=3$ *clusters* encontrados:

Algoritmo	K	Índice de Silueta	Davies–Bouldin (DBI)	Calinski–Harabasz (CH)
K-means	3	0.552	0.669	560.0

Análisis de Métricas:

- Silueta (0.552): Un valor positivo y significativamente superior a 0 indica una separación razonablemente buena de los *clusters*. Los puntos están más cerca de su propio *cluster* que del vecino más cercano.
- DBI (0.669): Un valor cercano a cero indica una agrupación de alta calidad. El valor obtenido es bajo, lo que implica que los *clusters* son compactos y están bien separados entre sí.
- CH (560.0): Este valor es relativamente alto (buscamos maximizarlo), confirmando la presencia de *clusters* densos y bien aislados.

Aplicación de Reducción de Dimensionalidad: PCA

Se aplicó el Análisis de Componentes Principales (PCA) para reducir los datos de 4 a 2 dimensiones, lo que permite la visualización y el cálculo de métricas de preservación de información.

Resultados de Reducción: Gráfica de Varianza Explicada

Gráfica de Varianza Explicada

- El Primer Componente Principal (PC_1) explica aproximadamente el 92.5% de la varianza total.
- El Segundo Componente Principal (PC_2) explica un 5.3% adicional de la varianza.

Algoritmo	Componentes	Varianza Explicada Acumulada	Error de Reconstrucción (MSE)
PCA	2	0.978 (97.8%)	0.045
Análisis: La alta Varianza Explicada Acumulada (97.8%) con solo dos componentes demuestra que la reducción de 4D → 2D es altamente efectiva , reteniendo			

Algoritmo	Componentes	Varianza Explicada Acumulada	Error de Reconstrucción (MSE)
casi la totalidad de la información relevante. El bajo Error de Reconstrucción confirma que la pérdida de información es mínima.			

Diagrama de Clusters Tras Reducción

Para visualizar la calidad del *clustering* y la reducción simultáneamente, se grafica el resultado de K-means en el espacio 2D definido por PCA.

Gráfico de Clusters:

Observación Visual: El gráfico muestra que el algoritmo K-means logró separar la especie *Setosa* (un *cluster* azul muy compacto y bien aislado) de las otras dos especies a lo largo del. *Versicolor* y *Virginica* están más superpuestas, lo que visualmente se corresponde con el Índice de Silueta (0.552) que indica una buena, pero no perfecta, separación.

Conclusiones del Caso de Estudio

El caso de estudio con el dataset Iris demuestra que:

1. Clustering Eficaz: El K-means con $k=3$ es un modelo de agrupación adecuado, validado por los altos valores de Silueta y CH, indicando *clusters* compactos y separados.
2. Reducción Eficiente: La Varianza Explicada Acumulada muestra que los primeros dos componentes de PCA retienen casi el 98% de la varianza, justificando el uso de PC1 y PC2 para la visualización y como *input* para modelos posteriores.
3. Sinergia: La visualización en el espacio PCA confirma la solidez de las métricas: la agrupación que obtuvo un buen puntaje de Silueta se ve claramente separada en el plano 2D.

REFERENCIAS

(29 de Nov de 2025). Obtenido de wikipedia: https://en.wikipedia.org/wiki/Calinski-Harabasz_index

(29 de Nov de 2025). Obtenido de scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

(29 de Nov de 2025). Obtenido de mathworks: https://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabasz_evaluation.html

(29 de Nov de 2025). Obtenido de fastercapital: <https://fastercapital.com/es/tema/interpretaci3n-de-los-resultados-de-pca.html>

Rodríguez, D. (31 de Enero de 2025). *analyticslane*. Obtenido de <https://www.analyticslane.com/2025/01/31/como-determinar-el-numero-de-componentes-en-pca-usando-la-varianza-explicada-acumulada/>