Universidad Tecnológica de Chihuahua

# EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

## IV.2. Métricas de evaluación de modelos

Enrique Mascote
RICARDO ALONSO RIOS MONRREAL

## Introducción

La evaluación de modelos en el aprendizaje no supervisado presenta un reto único en comparación con el aprendizaje supervisado: la ausencia de etiquetas de verdad ("ground truth") contra las cuales comparar las predicciones. En este contexto, la validación de los resultados depende de métricas intrínsecas que evalúan la estructura de los datos resultantes.

Este reporte tiene como objetivo investigar y aplicar métricas clave para algoritmos de agrupación (*clustering*) y reducción de dimensionalidad. Se analizarán indicadores de cohesión y separación para el primero, y de preservación de información para el segundo. Posteriormente, se implementará un caso de estudio utilizando el conjunto de datos "Wine" para demostrar la aplicación práctica de estas métricas en un entorno de análisis de datos real.

## 2. Investigación de métricas

A continuación se definen las métricas seleccionadas para evaluar la calidad de los modelos.

### 2.1 Métricas de Agrupación (Clustering)

### A) Índice de Silueta (Silhouette Coefficient)

- Definición y fórmula: Mide qué tan similar es un objeto a su propio clúster (cohesión) en comparación con otros clústeres (separación). Para un punto $i$, se define como:

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

Donde a(i) es la distancia media intra-clúster y b(i) es la distancia media al clúster más cercano.

- Interpretación: El valor varía de -1 a 1. Un valor cercano a 1 indica que el punto está bien clasificado y lejos de los clústeres vecinos. Un valor cercano a 0 indica solapamiento.
- Ventajas y limitaciones: Es muy visual e intuitivo, pero puede ser costoso computacionalmente en datasets masivos.

### B) Índice Davies-Bouldin

- Definición y fórmula: Evalúa la "similitud" promedio entre cada clúster y su clúster más similar. Se basa en la relación entre la dispersión dentro del clúster y la distancia entre los centroides de los clústeres.

- Interpretación: A diferencia de la mayoría de las métricas, aquí menor es mejor. Un valor bajo indica que los clústeres son compactos y están bien separados entre sí.
- Ventajas y limitaciones: Es más simple de calcular que la Silueta, pero se limita a distancias euclidianas y asume clústeres convexos.

**C) Índice Calinski-Harabasz**

- Definición y fórmula: Conocido como el criterio de radio de varianza. Es la relación entre la suma de la dispersión entre grupos y la dispersión dentro de los grupos.
- Interpretación: Un valor alto indica clústeres densos y bien definidos.
- Ventajas y limitaciones: Es rápido de calcular, pero tiende a favorecer clústeres convexos y de densidad uniforme.

**2.2 Métricas de Reducción de Dimensionalidad**

**A) Varianza Explicada Acumulada**

- Definición: En métodos como PCA, mide el porcentaje de la varianza total del dataset original que es retenida por los componentes principales seleccionados.
- Interpretación: Se busca un valor alto (generalmente >80% o >90%). Si el valor es bajo, la reducción ha eliminado información crítica.
- Ventajas: Es la métrica estándar para decidir cuántas dimensiones mantener.

**B) Error de Reconstrucción (MSE)**

- Definición: Calcula la diferencia promedio (Error Cuadrático Medio) entre los datos originales y los datos reconstruidos tras proyectarlos al espacio reducido y devolverlos al espacio original.
- Interpretación: Un error bajo indica una alta fidelidad en la compresión de los datos.

**3. Caso de estudio y aplicación práctica**

Descripción del Dataset:

Se utilizó el Wine Dataset disponible en la librería Scikit-Learn. Este conjunto contiene resultados de un análisis químico de vinos cultivados en una región de Italia.

- Instancias: 178 muestras.
- Atributos: 13 variables continuas (Alcohol, Ácido málico, Ceniza, Alcalinidad, Magnesio, Fenoles, etc.).

- Propósito: El objetivo es agrupar los vinos por similitud química y reducir la complejidad de las 13 variables para su visualización.

**Metodología:**

1. Preprocesamiento: Estandarización de datos (StandardScaler) para normalizar las escalas de las variables (ej. el Magnesio tiene valores de >100 mientras que los Fenoles son <5).
2. Clustering: Aplicación del algoritmo K-Means con $k=3$ (asumiendo tres perfiles de vino).
3. Reducción: Aplicación de PCA (Análisis de Componentes Principales) para reducir de 13 dimensiones a 2.

## 4. Resultados

Tras la ejecución del script, se obtuvieron los siguientes valores:

### 4.1 Evaluación del Clustering (K-Means, k=3)

| **Métrica** | **Valor Obtenido** | **Análisis** |
|---|---|---|
| **Silhouette Score** | 0.2849 | El valor positivo indica que hay estructura, pero al ser bajo (<0.5), sugiere que los límites entre los grupos de vinos no son claros y existe solapamiento significativo. |
| **Davies-Bouldin** | 1.3892 | Un valor moderado. Confirma que la separación entre los tipos de vino no es perfecta usando solo distancias euclidianas en 13 dimensiones. |
| **Calinski-Harabasz** | 70.94 | Indica una varianza inter-clúster aceptable, validando estadísticamente la partición en 3 grupos. |

### 4.2 Evaluación de la Reducción (PCA, 13D to 2D)

- **Varianza Explicada Acumulada: 55.41%**
    - *Interpretación:* Al reducir el dataset a solo 2 dimensiones para graficarlo, estamos perdiendo casi la mitad de la información química del vino. Esto explica por qué visualmente los clústeres pueden parecer mezclados.
- **Error de Reconstrucción (MSE): 0.4459**
    - *Interpretación:* Existe una pérdida de información considerable. Para un análisis más riguroso, sería necesario aumentar el número de componentes a 3 o 4 para capturar al menos el 80% de la varianza.

## 5. Comparativa y conclusiones

Análisis de Métricas:

En este ejercicio, el Índice de Silueta demostró ser la métrica más honesta sobre la calidad del agrupamiento, revelando que aunque K-Means forzó 3 grupos, la separación natural no es nítida. Por otro lado, la Varianza Explicada en PCA fue crítica para entender que una visualización 2D es una sobresimplificación de la complejidad química del vino.

Conclusiones Generales:

El uso combinado de agrupación y reducción de dimensionalidad es poderoso pero requiere precaución. Reducir dimensiones facilita la visualización, pero métricas como la Varianza Explicada nos advierten sobre la pérdida de datos. Del mismo modo, métricas como Davies-Bouldin nos ayudan a no confiar ciegamente en los grupos que "encuentra" un algoritmo, obligándonos a validar matemáticamente su compacidad.

Recomendaciones:

Para futuros trabajos con este dataset, se recomienda utilizar técnicas de reducción no lineales como t-SNE para la visualización, ya que podrían separar mejor los grupos complejos que PCA mezcla, mejorando potencialmente la interpretación visual de los clústeres.

## Anexos: Código de Implementación

```python
import matplotlib.pyplot as plt
import numpy as np
from sklearn.datasets import load_wine
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score, mean_squared_error

# 1. Carga y Preparación
data = load_wine()
X = data.data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 2. Clustering (K-Means)
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
labels = kmeans.fit_predict(X_scaled)

# Cálculo de métricas de agrupación
sil = silhouette_score(X_scaled, labels)
db = davies_bouldin_score(X_scaled, labels)
ch = calinski_harabasz_score(X_scaled, labels)

# 3. Reducción de Dimensionalidad (PCA)
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Cálculo de métricas de reducción
var_acum = np.sum(pca.explained_variance_ratio_)
X_recon = pca.inverse_transform(X_pca)
mse = mean_squared_error(X_scaled, X_recon)

# 4. Imprimir Resultados
print(f"--- Métricas de Clustering ---")
print(f"Silhouette Score: {sil:.4f}")
print(f"Davies-Bouldin: {db:.4f}")
print(f"Calinski-Harabasz: {ch:.4f}")

print(f"\n--- Métricas de Reducción (PCA 2D) ---")
print(f"Varianza Explicada Acumulada: {var_acum:.2%}")
print(f"Error de Reconstrucción (MSE): {mse:.4f}")
```

**Referencias**

*silhouette_score*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Rodríguez, D. (2023, June 18). *El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python*. Analytics Lane. https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/

Rodríguez, D. (2023, June 11). *Identificar el número de clústeres con Calinski-Harabasz en k-means e implementación en Python*. Analytics Lane. https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/

Rodríguez, D. (2025, January 25). *Cómo determinar el número de componentes en PCA usando la varianza explicada acumulada*. Analytics Lane. https://www.analyticslane.com/2025/01/31/como-determinar-el-numero-de-componentes-en-pca-usando-la-varianza-explicada-acumulada/

*Vista de Aplicación de Algoritmos de Estimación de Imágenes con Modelización Bayesiana | Ciencia Latina Revista Científica Multidisciplinar*. (n.d.). https://ciencialatina.org/index.php/cienciala/article/view/7856/11902