UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la información

Extracción de Conocimiento en Bases de Datos

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

Docente

ING. LUIS ENRIQUE MASCOTE CANO

Alumno

Erick Eduardo Maffiodo Delgado

IDGS 91N

Domingo, 30 de Noviembre del 2025

Índice

Introducción

En la minería de datos y el aprendizaje automático es fundamental evaluar la calidad de los resultados de modelos no supervisados como el agrupamiento (clustering) y la reducción de dimensionalidad. Este ejercicio tiene como objetivo identificar y aplicar distintas métricas de evaluación para estos modelos, entendiendo su definición, interpretación y limitaciones, y demostrar su uso en un caso de estudio práctico.

Objetivos del ejercicio:

- Definir tres métricas internas para evaluar agrupaciones (clústers) y dos métricas para reducciones de dimensionalidad, incluyendo sus fórmulas, interpretación (valores altos/bajos) y sus ventajas y limitaciones.

- Aplicar un algoritmo de **clustering** (ej. *K-means*) a un conjunto de datos con ≥4 atributos numéricos y calcular las métricas de agrupación seleccionadas.

- Aplicar un método de **reducción de dimensionalidad** (ej. *PCA* o *t-SNE*) sobre el mismo conjunto de datos y calcular las métricas de reducción seleccionadas.

- Presentar los resultados: descripción del dataset, visualizaciones de los clústers (en 2D/3D tras la reducción dimensional), tabla de valores de métricas de agrupación, gráfica de varianza explicada u otra métrica de reducción, seguido de un análisis comparativo y conclusiones.

Métricas de evaluación de agrupación (clustering)

Para evaluar la calidad de los clusters obtenidos, existen métricas **internas** (no requieren etiquetas verdaderas) que miden qué tan compactos y separados están los grupos. A continuación, se describen tres métricas internas seleccionadas: **Índice de Silueta**, **Índice Davies–Bouldin** y **Índice Calinski–Harabasz**. Cada métrica se define con su fórmula, se interpreta en términos de valores altos/bajos, y se discuten sus ventajas y limitaciones.

Índice de Silueta

Definición y fórmula: El *coeficiente de silueta* cuantifica cuán similar es cada objeto a los elementos de su mismo clúster (cohesión intra-clúster) en comparación con los de otros clústeres (separación inter-clúster). Para cada punto *i*, se calcula:

- **a(i)** = distancia media de *i* a los demás puntos de su mismo clúster (cohesión).

- **b(i)** = distancia media mínima de *i* a todos los puntos de cualquier otro clúster (la menor distancia a un clúster vecino, es decir, la mejor separación).

El índice de silueta para el punto *i* se define como:

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

Este coeficiente oscila entre **-1** y **+1**. Se puede calcular con cualquier métrica de distancia (euclidiana, Manhattan, etc.). El valor promedio de *s(i)* sobre todos los puntos entrega la **silueta promedio** del clustering.

Interpretación: Valores altos (cercanos a +1) indican que el objeto está bien asignado a un clúster (alta cohesión interna y buena separación de otros grupos). Un valor cercano a 0 sugiere que el punto está en el límite entre dos clústeres, pudiendo asignarse a uno u otro. Valores negativos significan una mala asignación: el punto está más cerca de un clúster distinto que del propio, indicando que podría haberse agrupado incorrectamente. En general, **silueta promedio** alta (por ejemplo >0.5) implica una estructura de clústeres bien definida, mientras que valores bajos (cercanos a 0) o negativos sugieren que la configuración de clústeres puede no ser apropiada (quizás demasiados o muy pocos clústeres).

Ventajas:

- Proporciona una medida intuitiva y visualizable de la calidad de agrupamiento, tanto a nivel global (promedio) como a nivel de cada punto (perfil de silueta).

- Combina en una sola métrica la **cohesión** (distancias intra-grupo) y la **separación** (distancia al siguiente clúster más cercano) de los clústeres.

- Permite comparar diferentes *k* (número de clústeres) para ayudar a seleccionar el número óptimo de clústeres maximizando la silueta promedio.

- Es independiente de la escala de los datos si se usa una distancia adecuada (generalmente se estandarizan los datos antes de calcularla).

Limitaciones:

- Su cálculo requiere computar distancias entre pares de puntos para cada punto y su clúster, lo cual es computacionalmente costoso $O(n^2)$ para datasets muy grandes.

- Asume clusters bien separados; si la estructura de los datos no tiene fronteras definidas (ej. clústeres con formas arbitrarias o solapados), la silueta puede arrojar valores bajos incluso si los clústeres son significativos en contexto.

- Puede favorecer soluciones con menos clústeres: por ejemplo, si dos grupos están cercanos, la silueta podría sugerir agruparlos en uno solo (porque *b(i)* sería pequeño), penalizando particiones con clústeres adyacentes aunque sean distintos.

- No está definida para casos extremos como $k=1$ (un solo clúster) – requiere al menos 2 clústeres para calcular *b(i)*.

Índice Davies–Bouldin (DBI)

Definición y fórmula: El índice *Davies–Bouldin* (DBI) es una métrica interna introducida por D.L. Davies y D.W. Bouldin (1979) para evaluar la calidad de una agrupación. Se basa en la idea de que una buena agrupación debe producir clústeres **compactos** (baja dispersión interna) y **bien separados** entre sí. El DBI cuantifica la relación entre la dispersión intra-clúster y la separación inter-clúster para todos los pares de grupos. Típicamente se define como:

DBI=1k∑i=1kmaxj≠i{Si+Sjd(ci,cj)},\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left\{\frac{S_i + S_j}{d(c_i, c_j)}\right\},DBI=k1∑i=1kmaxj=i{d(ci,cj)Si+Sj},

donde kkk es el número de clústeres, SiS_iSi es la *dispersión intra-clúster* del grupo iii (ej. el promedio de las distancias de cada punto del clúster *i* a su centroide) y d(ci,cj)d(c_i, c_j)d(ci,cj) es la *separación inter-clúster* medida como la distancia entre los centroides cic_ici y cjc_jcj. Para cada clúster *i*, se calcula su "similaridad" con otro clúster *j* como (Si+Sj)/d(ci,cj)(S_i + S_j)/d(c_i, c_j)(Si+Sj)/d(ci,cj); luego Ri=maxj≠i{(Si+Sj)/d(ci,cj)}R_i = \max_{j \neq i}\{(S_i+S_j)/d(c_i,c_j)\}Ri=maxj=i{(Si+Sj)/d(ci,cj)} representa el peor caso (el clúster *j* más "parecido" al *i*). El DBI final es el promedio de estos peores casos RiR_iRi sobre todos los clústeres.

Interpretación: A diferencia de la silueta, **valores más bajos de DBI** indican una mejor agrupación. Un DBI = 0 implicaría clústeres perfectamente separados (muy ideal y raro en la práctica), mientras valores crecientes indican que al menos algún par de clústeres está relativamente cercano o disperso. En general: **clústeres compactos y bien separados dan un DBI bajo**, mientras que clústeres solapados o muy dispersos aumentan el DBI. No existe un rango fijo superior (puede ir de 0 al infinito), por lo que su utilidad es principalmente *comparativa*: se calcula el DBI para distintas particiones y se prefiere aquella con el menor valor. Por ejemplo, al variar *k*, el número óptimo de clústeres puede estimarse donde el DBI se minimiza (es común graficar DBI vs. k).

Ventajas:

- Es sencillo de calcular a partir de los centroides y las dispersiones intraclúster, lo que suele ser computacionalmente eficiente (*O(k·n)* para calcular S_i y distancias de

centroides, comparado a métricas punto a punto más pesadas).

- Combina tanto la **cohesión interna** (S_i) como la **separación entre grupos** en una sola cifra. Captura la intuición de "clústeres compactos y separados" de forma cuantitativa.

- No requiere conocer etiquetas verdaderas ni parámetros más que la definición de distancia; es una métrica interna general aplicable a cualquier método de clustering que defina centroides (o se puede generalizar usando distancias mediana-mediana para algoritmos sin centroide explícito).

- Útil para **comparar modelos de clustering**: por ejemplo, para decidir entre distintas configuraciones de *k* (clústeres) o incluso comparar diferentes algoritmos en el mismo dataset (el de menor DBI es preferible, dado el mismo *k*).

Limitaciones:

- Al igual que otras métricas internas, **no tiene significado absoluto**: un valor de DBI por sí solo no "dice" si los clústeres son buenos sin un contexto. Solo es útil relativo a otras particiones del mismo conjunto de datos.

- Tiende a favorecer configuraciones con menor número de clústeres. Porque al aumentar *k*, aunque la cohesión S_i suele mejorar (clústeres más pequeños y compactos), la separación inter-centroides *d(c_i,c_j)* puede disminuir poco; esto puede dar como resultado DBI más altos para *k* grandes (penalizando sobresegmentación). En consecuencia, el mínimo de DBI a veces sugiere un *k* algo menor que otras métricas como la silueta.

- Depende de la noción de centroide y distancia media. En clústeres no esféricos o de tamaños muy desiguales, la fórmula (que usa promedios y distancia centroidal) puede no reflejar bien la calidad: p.ej., un clúster alargado con dos subgrupos separados podría tener gran dispersión pero igualmente alejado de otro clúster, y DBI lo penalizaría

fuertemente.

- Es **sensible a outliers**: la dispersión SiS_iSi puede aumentar por la presencia de valores atípicos dentro de un clúster, degradando el DBI. Una estrategia suele ser limpiar outliers o usar medianas en vez de medias para calcular dispersiones en esos casos.

Índice Calinski–Harabasz (CH)

Definición y fórmula: El *índice Calinski–Harabasz* (también llamado **criterio de la razón de varianza**, *Variance Ratio Criterion*) es otra métrica interna clásica para evaluar clústeres. Fue propuesto por T. Caliński y J. Harabasz (1974). Este índice se define como la razón entre la varianza inter-clúster y la varianza intra-clúster, normalizadas por sus grados de libertad. En forma de fórmula:

CH=SSB/(k−1)SSW/(N−k),\text{CH} = \frac{\displaystyle \text{SSB} / (k-1)}{\displaystyle \text{SSW} / (N - k)},CH=SSW/(N−k)SSB/(k−1),

donde:

- **SSB** (*Sum of Squares Between*) es la suma de cuadrados entre grupos (varianza **entre clústeres**): básicamente, cuantifica la separación global entre los clústeres, calculando la distancia de cada centroide de clúster al centroide global (media total), ponderada por el número de puntos en cada clúster.

- **SSW** (*Sum of Squares Within*) es la suma de cuadrados dentro del grupo (varianza **intra-clúster**): suma de las distancias cuadráticas de cada punto al centroide de su clúster, acumulada sobre todos los clústeres (mide la compacidad interna total).

- *k* es el número de clústeres, y *N* el número total de observaciones.

El factor (k−1)(k-1)(k−1) en SSB y (N−k)(N-k)(N−k) en SSW sirven para **normalizar** por los grados de libertad (análogamente a un estadístico *F* de ANOVA).

Interpretación: Un valor **CH alto** indica clústeres bien definidos: alta separación intergrupal (SSB grande) y baja dispersión intragrupal (SSW pequeña). Es decir, cuanto mayor sea este ratio, más consistentes y distintos son los clústeres entre sí. A diferencia del DBI, aquí "más grande es mejor". No hay un máximo teórico fijo, pero en la práctica se compara CH para diferentes particiones: el número de clústeres óptimo suele corresponder al primer pico o máximo local de CH cuando se grafica CH vs. *k*. Un CH muy bajo significaría o que todos los puntos están esencialmente en un solo grupo sin diferencias (SSB muy bajo), o que los clústeres formados son muy difusos internamente (SSW alto). Por construcción, CH no está definido para *k=1* (denominador N−kN-kN−k sería 0); requiere al menos 2 clústeres para evaluarse.

Ventajas:

- Es **fácil de calcular** usando estadísticas de dispersión: muchos programas calculan SSB y SSW durante el clustering, por lo que CH viene "de gratis" sin sobrecarga computacional significativa.

- Relacionado con la interpretación ANOVA de clustering: esencialmente es proporcional a un estadístico *F*, lo que brinda una intuición estadística (maximizar separación entre grupos relativo a variabilidad dentro de grupos).

- Tiende a funcionar bien cuando los datos realmente se estructuran en clústeres esféricos de varianzas similares. En esos casos, suele presentar un máximo claro en el número "correcto" de clústeres. Es útil en conjunto con otros métodos (ej. método del codo) para decidir *k*.

- Al ser un valor global, es menos sensible a pequeñas perturbaciones: agrega la info de todos los puntos. Esto puede hacerlo más **estable** ante ligeros cambios de datos (a diferencia de la silueta, donde unos pocos puntos con mala asignación bajan el promedio notablemente, CH diluye ese efecto).

Limitaciones:

- **Escala no acotada:** como mencionamos, CH puede crecer con *k*. De hecho, agregar más clústeres siempre reduce SSW (ya que cada clúster tiene menos puntos), y aunque SSB también cambia, a veces CH simplemente aumenta monotonamente con *k* (especialmente si el dataset tiene estructura jerárquica continua). En tales casos encontrar un pico es difícil – podría sugerir un *k* grande que quizás sobreajusta. Se recomienda buscar *elbow* o primer máximo local, pero no siempre es obvio.

- **Supone varianza**: CH usa distancias cuadráticas internas (varianza) y centroides, lo cual **favorece clústeres de forma hiperesférica** en espacios métricos euclídeos. Si los clústeres reales tienen formas complejas (no convexas) o distribuciones muy desiguales en tamaño/densidad, CH podría ser engañoso. Por ejemplo, un clúster densísimo y otro disperso: el disperso domina SSW y CH sale bajo aunque visualmente haya dos grupos.

- **Sensibilidad a escala**: al igual que otras métricas de distancia, es importante escalar las variables antes de clustering; de lo contrario, la SSW/SSB puede estar dominada por la variable de mayor varianza numérica, afectando CH. (Esto se aplica a cualquier métrica basada en distancias euclidianas).

- **Comparación entre algoritmos**: CH está pensado para comparar particiones del mismo conjunto de datos variando *k*. No es directamente comparable entre datasets distintos (depende de N, variaciones globales, etc.). Tampoco contempla ruido o outliers explícitamente: puntos alejados influyen en SSW y SSB, a veces inflando SSB (centroides más lejanos) y SSW a la vez.

(Otras métricas de cohesión/separación existen, como el índice de Dunn, coeficiente de separación, índice de Dunn, etc., pero Silueta, DBI y CH son de las más usadas y cubren el espectro de evaluaciones internas.)

Métricas de evaluación de reducción de dimensionalidad

Al reducir la dimensionalidad de datos (por ejemplo, de muchas variables a solo 2 componentes para visualización), necesitamos medir **qué tanta información conserva** la proyección de menor dimensión respecto de los datos originales. Las métricas de evaluación de reducción buscan cuantificar la calidad de una representación reducida. Aquí seleccionamos dos métricas: **Varianza explicada acumulada** (adecuada para métodos lineales tipo PCA) y **preservación del vecindario local (*Trustworthiness*)** (útil para métodos no lineales tipo *t-SNE/UMAP*). Se describen a continuación con sus definiciones, interpretaciones y pros/contras.

Varianza explicada acumulada

Definición: En técnicas como el **Análisis de Componentes Principales (PCA)**, cada componente principal captura cierta porción de la varianza total del conjunto de datos. La *varianza explicada* por un componente cuantifica cuánta información (dispersión) de los datos originales retiene ese componente. La **varianza explicada acumulada** tras seleccionar m componentes es simplemente la suma de las varianzas explicadas por los primeros m componentes, expresada como fracción o porcentaje del total. Por ejemplo, si el primer componente explica 40% de la variabilidad y el segundo 20%, la varianza explicada acumulada con 2 componentes sería 60%. Formalmente, si λ1,λ2,...,λp\lambda_1, \lambda_2, ..., \lambda_pλ1,λ2,...,λp son los valores propios (varianzas) de cada componente de PCA (con ppp igual al número original de variables), entonces la varianza explicada acumulada por las primeras m dimensiones es:

Vacum(m)=∑i=1mλi∑i=1pλi,V_{\text{acum}}(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i},Vacum(m)=∑i=1pλi∑i=1mλi,

frecuentemente expresado en porcentaje. Este valor indica qué proporción de la información total del conjunto de datos se conserva usando m componentes.

Interpretación: Se busca generalmente un **porcentaje acumulado alto** con el menor número de dimensiones posible. Por ejemplo, si con 3 componentes se logra un 95% de varianza explicada acumulada, significa que casi toda la estructura de los datos (95% de su dispersión original) está contenida en ese subespacio de 3D, lo cual es excelente. Valores acumulados bajos (ej. 50-60%) implican que la proyección reducida está perdiendo casi la mitad de la variabilidad original –

potencialmente información importante. Es común fijar umbrales como **90% o 95%** de varianza explicada acumulada como criterio para decidir cuántos componentes mantener. También se utiliza el llamado gráfico de *codo* o *scree plot*, donde se grafica Vacum(m)V_{\text{acum}}(m)Vacum(m) vs. *m*: al inicio la varianza acumulada crece rápido y luego se aplana; el punto donde se aplane (diminishing returns) sugiere el número adecuado de dimensiones. En resumen, un valor alto de varianza acumulada indica que la reducción de dimensionalidad retiene **la mayor parte de la información** estadística de los datos originales.

Ventajas:

- Es una medida **global e intuitiva** de información retenida. Tener el, por ejemplo, 95% de varianza retenida da confianza de que la estructura global de los datos se conserva en la proyección reducida.

- Fácil de calcular y de interpretar en PCA: la mayoría de softwares da directamente el % de varianza explicada por cada componente.

- Útil para **selección de dimensionalidad**: permite determinar objetivamente cuántos componentes son necesarios para lograr cierto nivel de fidelidad (conservar X% de la varianza). Esto ayuda a evitar tanto quedarnos cortos (perder información clave) como a reducir en exceso (mantener dimensiones irrelevantes).

- Aplicable a cualquier método lineal de reducción (PCA, SVD, etc.) donde la noción de varianza se mantiene. También se puede aplicar a autoencoders en términos de error de reconstrucción complementario (100% - varianza explicada = fracción del error).

Limitaciones:

- **Sólo captura varianza lineal.** Si se usan métodos no lineales (t-SNE, UMAP) la "varianza explicada" no es un concepto definido de la misma forma. Estos métodos no garantizan maximizar varianza, por lo que esta métrica no sirve directamente para

evaluarlos.

- Un porcentaje alto de varianza explicada **no siempre equivale a mejor desempeño en tareas** finales. Por ejemplo, PCA podría retener 90% de la varianza pero si el 10% perdida correspondía justamente a la señal útil para una cierta tarea (ej. clasificar una minoría), entonces podría fallar pese al alto porcentaje.

- **Distribución de la varianza**: Dos datos diferentes pueden tener misma varianza acumulada pero con significados distintos. Ejemplo: 90% con 1 componente vs 90% con 5 componentes – en ambos casos se retiene mucho, pero en el primer caso la estructura es básicamente unidimensional, en el segundo es más compleja. Siempre conviene mirar también la varianza explicada *individual* de cada componente (el gráfico de barras), no solo el acumulado.

- Depende de la **escalabilidad** de variables: PCA requiere estandarizar variables previamente. Si no se hace, una variable de gran varianza dominará el porcentaje explicado. Pero esto es un paso reconocido; la métrica asume que ya se ha hecho una buena preparación de datos.

Preservación del vecindario local (*Trustworthiness*)

Definición: Muchas técnicas de reducción de dimensionalidad, en especial las no lineales (*manifold learning* como t-SNE, UMAP), buscan **preservar las relaciones de proximidad** entre puntos. La métrica de *Trustworthiness* (confiabilidad o fidelidad de la proyección) cuantifica hasta qué punto la estructura de vecinos cercanos en el espacio original se conserva en el espacio reducido. Intuitivamente, penaliza las "falsas vecindades" que aparecen en la proyección. Se basa en el ranking de vecinos: para cada punto i, se consideran sus k vecinos más próximos en la proyección de baja dimensión, y se verifica cuán cerca estaban esos vecinos en el espacio original. La fórmula formal es:

T(k)=1−2Nk(2N−3k−1)∑i=1N∑j∈Uik(r(i,j)−k),T(k) = 1 - \frac{2}{N k (2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_i^k} (r(i,j) - k),T(k)=1−Nk(2N−3k−1)2∑i=1N∑j∈Uik(r(i,j)−k),

donde NNN es el número de puntos, UikU_i^kUik es el conjunto de puntos que **son vecinos de iii en la proyección (espacio reducido)** pero **no** lo eran entre sus *k* vecinos en el espacio original, y r(i,j)r(i,j)r(i,j) es la posición (ranking) de la distancia del punto jjj respecto a iii en el espacio original. En esencia, para cada "vecino inesperado" jjj que apareció cerca de iii tras la proyección, se calcula cuán lejos estaba jjj de iii originalmente (qué tan grande es r(i,j)r(i,j)r(i,j) comparado con kkk), y se acumula un castigo proporcional a esa diferencia. El factor de normalización asegura que T(k)T(k)T(k) varíe entre **0 y 1**.

Interpretación: El índice de *trustworthiness* toma valor **1.0 en el mejor de los casos**, cuando Nik\mathcal{N}^k_iNik (los vecinos de iii en la proyección) coinciden exactamente con sus *k* vecinos reales en el espacio original (no hay intrusos falsos). Si la proyección introduce vecinos que no correspondían, *T* disminuye; valores cercanos a 0 significarían que la estructura local se distorsionó severamente. En general, un **trustworthiness alto (≥0.9)** indica que la geometría local (hasta *k*-ésimos vecinos) está bien preservada por la proyección, mientras que un valor bajo indica que la proyección es engañosa en términos de proximidad (muchos puntos que parecen cercanos en la visualización no lo estaban en los datos originales). Es importante notar que *T(k)* depende del parámetro *k* (el tamaño de vecindario considerado); a menudo se examina para varios *k* o se elige un *k* típico (ej. 5 o 10) para evaluar la fidelidad local. *Trustworthiness* enfoca el error de tipo "intrusión" (puntos extraños que se meten como vecinos) y complementariamente existe la métrica de **continuidad** que enfoca el error de "exclusión" (puntos que eran vecinos en alta dimensión pero quedaron separados en la proyección). Ambas suelen reportarse juntas para una evaluación más completa de la proyección.

Ventajas:

- Es especialmente relevante para **visualización y análisis exploratorio**: asegura que los patrones locales observados (clústeres cercanos, puntos agrupados) en la baja dimensión no sean artefactos. Un trustworthiness alto da confianza de que "lo que ves es lo que hay" localmente en los datos originales.

- Proporciona una medida cuantitativa para comparar métodos de reducción: por ejemplo, se puede decir "t-SNE obtuvo T=0.95 vs PCA T=0.85" indicando que t-SNE preservó mejor las relaciones locales.

- Detecta **distorsiones locales** que la varianza explicada no capta. Un método podría explicar mucha varianza global pero embarajar vecinos (baja T), mientras otro mantiene vecinos (alta T) a costa de varianza. Según la aplicación (mantener estructura de clusters vs. maximizar varianza), T puede ser más pertinente que la varianza.

- Es general: se aplica a cualquier reducción donde podamos medir distancias en original y proyectado. No requiere supuestos fuertes, solo una métrica de distancia (usualmente euclídea) y un k.

Limitaciones:

- **Parámetro k:** La elección de k afecta el valor. T evalúa la preservación de vecinos hasta cierto radio. Un método podría optimizar la estructura muy local (k pequeño) pero fallar en una escala un poco mayor (k grande). Conviene analizar varios k, lo que complica resumir en un solo número.

- **Sólo evalúa estructura local (intrusiones):** No captura si la estructura global (relaciones de largo alcance) se respeta o no. Un método podría tener T perfecto pero quizás reordenó completamente clústeres entre sí. Por eso se complementa con **continuity** (exclusiones) y otras métricas globales (ej. correlación de distancias, estrés, etc.).

- Computacionalmente, requiere calcular vecinos en datos originales y proyectados, lo cual es $O(N^2)$ si se hace ingenuamente para todos los puntos (aunque con estructuras de vecinos o limitando N se maneja). Para datasets muy grandes, puede ser costoso evaluar T.

- No tiene una interpretación inmediata para usuarios finales en términos de "% de info retenida" como la varianza. Es una métrica técnica cuyo valor debe contextualizarse (por ejemplo, T=0.9 es bueno, pero cuánto se "pierde" con 0.9 vs 1.0 no es tan tangible).

Descripción del dataset de caso de estudio

Para demostrar estas métricas en la práctica, se utilizó el clásico conjunto de datos **Iris** de Fisher. Este dataset contiene 150 muestras de flores de iris, con 4 atributos numéricos cada una (largo y ancho de sépalo, largo y ancho de pétalo) y una etiqueta de especie (Setosa, Versicolor, Virginica) para validación externa. En nuestro análisis, consideramos solo los 4 atributos numéricos sin usar la etiqueta de especie durante el clustering. Este conjunto fue elegido por su tamaño manejable y porque se sabe que aproximadamente forma 3 grupos naturales (correspondientes a las especies) aunque con cierto solapamiento entre dos de ellas.

Antes de aplicar los algoritmos, los datos fueron **estandarizados** (restando la media y dividiendo por la desviación estándar de cada atributo) para asegurar que todas las variables contribuyan equitativamente a las distancias.

Para la fase de **agrupamiento**, empleamos el algoritmo *K-means* con k=3k=3k=3 clústeres (anticipando tres grupos posibles). Para la fase de **reducción de dimensionalidad**, usamos *PCA* (Análisis de Componentes Principales) para proyectar los datos originales de 4D a 2D, facilitando la visualización de los clústeres encontrados. Adicionalmente, calculamos las métricas de evaluación seleccionadas en cada caso: silueta, Calinski-Harabasz y Davies-Bouldin para el clustering, y varianza explicada acumulada (y trustworthiness) para la reducción dimensional.

Resultados de clustering: visualización y métricas

Figura 1: Visualización de los datos Iris agrupados mediante K-means (k=3), proyectados en 2D usando PCA. En la gráfica, cada punto representa una muestra de flor, proyectada sobre los dos primeros componentes principales. Los colores indican los **clústeres** encontrados por K-means. Se observa que el clúster azul está claramente separado (corresponde a la especie Setosa)

mientras que los clústeres naranja y verde aparecen más próximos y con algo de solapamiento (estas dos agrupaciones corresponden principalmente a Versicolor y Virginica, conocidas por ser más similares entre sí en este espacio de atributos). La proyección PCA captura **95.8% de la varianza** de los datos en estas dos dimensiones, por lo que la separación observada es representativa de la estructura real. Podemos apreciar un buen aislamiento del grupo Setosa (izquierda de la figura) y una cercanía entre Versicolor y Virginica (derecha, parcialmente traslapados), consistente con lo que se espera de este dataset.

A continuación, se presentan las **métricas de validación interna** calculadas para esta agrupación en 3 clústeres:

Métrica (Clustering)	**Valor (k=3)**
Silueta promedio	0.46 (aprox.)
Calinski–Harabasz (CH)	241.43
Davies–Bouldin (DBI)	0.832

Tabla 1: Valores de métricas de agrupación para la solución K-means con k=3 en el dataset Iris.

Interpretación de resultados: La **silueta promedio = 0.46** indica una calidad de clustering **moderada**. En la escala de silueta, ~0.46 sugiere que aunque existe cierta estructura (especialmente el clúster de Setosa que tenía siluetas individuales altas ~0.7), muchos puntos de los otros dos clústeres están cerca del límite entre grupos (algunos silueta < 0.4), reflejando el solapamiento Versicolor/Virginica. De hecho, calculando las siluetas individuales, se verifica que prácticamente todas las Setosa tienen silueta >0.7 (muy bien agrupadas), mientras muchas Versicolor/Virginica están alrededor de 0.3 e incluso algunas negativas, indicando confusión

entre esos dos clústeres – esto concuerda con la visualización 2D y la biología, donde esas dos especies tienen características similares.

El índice **Calinski–Harabasz = 241.4** es un número relativamente alto para 3 clústeres en 150 muestras, lo que sugiere que la separación entre clústeres (especialmente gracias al grupo Setosa) es significativa comparada con la dispersión interna. Por comparación, para k=2 clústeres en estos mismos datos CH alcanzó ~251, ligeramente mayor, mientras para k=4 bajó a ~206. Esto indicaría que estadísticamente quizás **2 clústeres** optimizan más la varianza (agrupando Versicolor y Virginica juntas). Sin embargo, la diferencia no es drástica entre 2 y 3 (251 vs 241), y dado el conocimiento previo de 3 especies, es razonable elegir k=3. En general, un CH elevado como 241 apoya que hay estructura apreciable diferenciando grupos en los datos.

El **Davies–Bouldin = 0.832** (recordemos, menor es mejor) está por debajo de 1.0, lo cual indica una calidad aceptable de los clústeres (clústeres relativamente compactos y separados). Para referencia, con k=2 clústeres en Iris el DBI era ~0.59 (mejor), y con k=4 subía a ~0.92 (peor). El valor 0.83 aquí refleja precisamente que uno de los clústeres (Setosa) está muy separado, pero los otros dos están más cercanos y dispersos, elevando el promedio de las peores ratios intra/inter en la fórmula. Aun así, 0.83 es un valor que sugiere que la agrupación tiene una estructura razonable (no aleatoria): clústeres ideales suelen dar DBI en el rango 0.3–0.7, mientras >1 indica clústeres bastante pobres. Vemos que pasar de 2 clústeres a 3 empeoró el DBI, consistente con la silueta: separar Versicolor y Virginica introdujo clústeres más “difusos”.

Resumen: Las tres métricas concuerdan en que existe una estructura clusterizable en los datos Iris, destacando un grupo muy bien definido y otros dos más tenues. La silueta y DBI sugieren que quizás la partición óptima en términos puramente geométricos sería 2 clústeres (porque Versicolor y Virginica en realidad forman un continuo), mientras que con 3 clústeres obtenemos valores algo menos óptimos pero aún indicadores de clusters discernibles. En un escenario real sin conocimiento previo, uno observaría que la silueta promedio máxima ocurre en k=2 (0.58) y empieza a disminuir en k=3 (0.46), y similarmente DBI mínimo en k=2. Sin embargo, si el objetivo es identificar subestructuras más finas, se podría sacrificar algo de silueta para obtener 3 grupos. Estas métricas nos ayudaron a **validar** que los clústeres encontrados (especialmente Setosa) son significativos y no aleatorios, y nos alertaron sobre la cercanía de los otros grupos.

Resultados de reducción de dimensionalidad: varianza y confianza

Figura 2: Varianza explicada por componente principal en el dataset Iris, y varianza explicada acumulada. El gráfico de barras (azul) muestra el porcentaje de varianza explicado por cada uno de los 4 componentes principales de PCA, y la línea anaranjada muestra la varianza acumulada. Podemos observar que el **primer componente** captura ~72.96% de la varianza total, el **segundo** ~22.0% adicional (llevando el acumulado a 95.8%), el tercero ya marginal ~3.7% (acumulado 99.5%) y el cuarto apenas ~0.5% restante (hasta 100%) – estos números se reflejan en la gráfica: la curva acumulada se aplana rápidamente después del segundo componente. Esto indica que **con solo 2 dimensiones PCA logra representar ~95.8% de la información** del dataset【60†】, lo cual es extremadamente bueno. De hecho, la dispersión de puntos en la figura 1 de clusters es una proyección casi completa de la estructura original (poca información se perdió al proyectar de 4D a 2D).

La elección de reducir a 2 dimensiones está justificada porque pasando de 2 a 3 componentes la ganancia de varianza es pequeña (de 95.8% a 99.5%, <4% extra) y para visualización 2D es preferible. Si quisiéramos *cero* pérdida de información, necesitaríamos las 4 dimensiones (100% varianza), pero en la práctica 95–99% suele considerarse suficiente. En este caso, incluso 2 comp. bastaron para recuperar la mayor parte de la estructura (los pétalos de iris ya determinan bien las especies en un plano).

Adicionalmente, calculamos la métrica de **trustworthiness** para cuantificar la preservación local de la proyección PCA 2D. Tomando $k = 5$ vecinos, obtuvimos $T(5) \approx \mathbf{0.97}$, es decir, 97% de las relaciones de vecindad de 5 más cercanos se mantuvieron. Un valor tan alto sugiere que la proyección PCA no distorsionó apreciablemente las distancias locales: los puntos que eran cercanos en 4D casi siempre permanecen cercanos en 2D. Esto tiene sentido ya que el dataset original es de baja dimensión (4) y la mayoría de su varianza es planar; PCA fue suficiente para "desenredar" la estructura linealmente. Para contrastar, si hubiésemos usado un método no lineal como *t-SNE*, posiblemente también obtendríamos un trustworthiness alto ($\gg 0.9$) porque su objetivo es precisamente mantener vecinos locales – la diferencia es que t-SNE podría incluso mejorar ligeramente la separación de los dos grupos solapados (Versicolor/Virginica) a expensas de distorsionar distancias globales, pero PCA ya era bastante adecuada aquí.

En resumen, la **varianza explicada acumulada** nos confirmó cuantitativamente que reducir de 4D a 2D fue apropiado (muy poca información perdida). Asimismo, la métrica de **trustworthiness** indicó que la estructura local (agrupamientos de puntos similares) se preservó en la visualización, validando que no introdujimos artefactos visuales. Ambas métricas se complementan: la primera mira la fidelidad global de datos, la segunda la fidelidad de la *estructura local*. En este caso de Iris, ambas salen altas debido a la naturaleza relativamente simple del conjunto de datos.

Comparativa y análisis de las métricas

Métricas de clustering: En este ejercicio, el **Índice de Silueta** mostró ser muy informativo para evaluar la calidad de los clústeres, ofreciendo una interpretación directa de cada punto y un criterio global para comparar diferentes números de clústeres. Vimos que la silueta promedio distinguió claramente cuando pasamos de 2 a 3 clústeres: bajó de 0.58 a 0.46, alertándonos del costo de separar clústeres que quizás no estaban bien aislados. Su ventaja es esa intuición inmediata (positivo bueno, negativo malo) y la posibilidad de examinar distribuciones de siluetas individuales para diagnosticar problemas (ej. qué porcentaje de puntos tiene silueta negativa, etc.). En problemas prácticos, **la silueta suele ser la métrica preferida** para validación interna por su equilibrio entre considerar cohesión y separación. Sin embargo, computacionalmente puede ser pesada en grandes datasets (distancias punto a punto). Ahí es donde **Calinski–Harabasz** y **Davies–Bouldin** destacan: ambos se calculan fácilmente a partir de sumas de cuadrados o centroides, escalando mejor a grandes N. En nuestro caso, CH y DBI corroboraron la historia que contaba la silueta (máximo CH y mínimo DBI en k=2, luego algo peor en k=3, etc.), lo cual aumenta la confianza en las conclusiones. Si alguna métrica hubiera discrepado fuertemente, habría que analizar por qué. Por ejemplo, CH tiende a favorecer clústeres de tamaño más equilibrado; en otro dataset con tamaños muy disparejos, puede preferir dividir un gran clúster en varios pequeños (subiendo SSB bastante) mientras la silueta podría penalizar por cohesión baja de los pequeños. **No existe una métrica "única" mejor en absolutamente todos los casos**, por eso es buena práctica observar varias.

En general, **Silhouette** es muy útil para *interpretación cualitativa* y selección de *k*, **CH** es excelente para *comparación estadística global* (especialmente en contextos parecidos a esferas

gausianas), y **DBI** ofrece un criterio sencillo inverso a optimizar. En nuestro análisis, la silueta fue la más reveladora (dándonos insight punto a punto), mientras que CH/DBI confirmaron cuantitativamente la estructura. Las tres métricas funcionaron "mejor" para diferentes propósitos: *Silueta* para diagnosticar solapamiento de clústeres, *CH* para respaldar el número de clústeres con un valor máximo, y *DBI* para contrastar compactación/separación promedio. Si tuviéramos que elegir, la **Silueta** probablemente sería la métrica más completa debido a su interpretabilidad y amplio uso, pero conviene no depender solo de una.

Métricas de reducción de dimensionalidad: Aquí también es útil combinar enfoques. La **varianza explicada** nos dio una visión global de cuánta información retenemos al proyectar. En un contexto como PCA o cualquier técnica lineal, es la métrica primaria para decidir la dimensionalidad óptima (buscando ese 90-95% típico de umbral). Funciona "mejor" cuando la estructura relevante de los datos coincide con la varianza (lo cual en tareas exploratorias suele ser cierto: componentes de mayor varianza suelen contener patrones interesantes). Sin embargo, la varianza no sabe nada de la tarea final o de la estructura de interés específica (por ejemplo, si nos interesa conservar la distancia entre clusters, la varianza podría no correlacionar con eso). Ahí es donde métricas como **trustworthiness** entran: evalúan *qué tan bien la reducción preserva la geometría de los datos*. En nuestro caso, trustworthiness fue casi perfecto para PCA 2D, lo cual esperábamos dado el alto porcentaje de varianza retenida. Pero imaginemos datos más complejos (ej. un conjunto con estructura en forma de "S" donde PCA falla): PCA podría conservar 80% de varianza pero aplanando la forma y mezclando vecinos (trustworthiness bajo). Un método no lineal como t-SNE podría tener varianza explicada formalmente baja (no se puede medir igual, pero ciertamente no maximiza varianza) y aun así *trustworthiness* alto manteniendo la forma en S. Por tanto, **¿qué métrica funciona mejor? Depende del objetivo**: si queremos asegurar mínima pérdida de información general -> varianza explicada; si queremos preservar relaciones locales o de clúster -> trustworthiness/continuity. En prácticas de visualización de datos, trustworthiness es crucial porque garantiza que los grupos vistos en 2D son confiables. En compresión de datos para reconstrucción, el **error de reconstrucción** (complemento de varianza explicada) es otra métrica clave: en PCA el error cuadrático promedio de reconstruir los datos desde los componentes descartados está directamente ligado a la varianza no explicada. Nosotros podríamos haber presentado el error de reconstrucción, pero es básicamente 1 - varianza

acumulada (en proporción). Para Iris con 2 comp., el error de reconstrucción sería ~4.2% de la varianza total (muy bajo).

Con respecto a **continuity** (continuidad), que no calculamos explícitamente, es el análogo inverso de trustworthiness: verifica que ningún vecino original se "perdió" en la proyección. En una buena proyección ambos *T* y *C* estarán altos. Si trustworthiness ≈ continuity ≈ 0.9+, significa que tanto intrusiones como exclusiones son pocas, y la proyección es globalmente fiable. Si hay disparidad (ej. T alta, C baja), querría decir que la proyección introduce pocos falsos vecinos pero algunos vecinos originales quedaron alejados (posible "estiramiento" de escalas locales); al revés (T baja, C alta) implicaría la proyección junta puntos que no deberían estar juntos, aunque mantuvo algunos grupos originales cercanos. Es interesante analizar estos detalles en aplicaciones donde la topología de los datos importe (manifold learning).

Resumen comparativo: En nuestro caso de estudio, todas las métricas seleccionadas desempeñaron bien su rol: las de clustering nos ayudaron a identificar la estructura adecuada (sabiendo que 3 clústeres era una decisión razonable pero con solapamiento observado), y las de reducción confirmaron que proyectar en 2D no comprometió seriamente la integridad de los datos. Podemos concluir que **la elección de métricas adecuadas proporciona una validación robusta** de resultados no supervisados. No se trata de cuál es "la mejor" métrica en términos absolutos, sino de entender qué aspecto mide cada una y usar ese conocimiento para evaluar distintos facetas de nuestros modelos. En prácticas reales, se recomienda usar **varias métricas en conjunto**: si todas apuntan a la misma conclusión, aumentamos la confianza en el hallazgo (por ejemplo, aquí silueta, CH y DBI concordaron en general). Si difieren, debemos investigar más a fondo la naturaleza de los clústeres o de la proyección para entender las discrepancias.

Conclusiones y recomendaciones

- **Evaluar modelos de clustering** con métricas internas es esencial para validar si los patrones encontrados son significativos. En este ejercicio vimos que la silueta, el índice Davies-Bouldin y el índice Calinski-Harabasz proporcionan perspectivas complementarias sobre la calidad de los clústeres sin requerir verdad terreno. La silueta resultó especialmente útil para interpretar la asignación de puntos a clústeres y detectar solapamientos, mientras que CH y DBI aportaron criterios globales para comparar

configuraciones de clústeres (elegir el número *k* apropiado buscando máximo CH o mínimo DBI). **Recomendación:** siempre calcular al menos una métrica de cohesión-separación (silueta o DBI) al hacer clustering, y utilizarla para afinar parámetros (ej. decidir *k* o eps en DBSCAN). Si es factible, complementar con una segunda métrica para corroborar resultados.

- **Evaluar reducciones de dimensionalidad** es igualmente importante, aunque a veces se omite. Aquí demostramos que medir la varianza explicada acumulada ayuda a determinar cuánta dimensionalidad podemos recortar con pérdidas aceptables. En el ejemplo, 2 componentes bastaron para retener >95% de la información de Iris, lo que validó nuestra elección. Asimismo, medir la confianza en la proyección (*trustworthiness*) nos aseguró que la estructura local no se distorsionó, algo crucial si usamos la visualización para extraer conclusiones. **Recomendación:** al aplicar PCA, usar la curva de varianza explicada para elegir el número de componentes (con un umbral de varianza retenida adecuado al caso de uso, e.j. 95%). Para métodos de proyección no lineales destinados a visualizar o preprocesar datos, calcular métricas de preservación de estructura como trustworthiness/continuity o error de reconstrucción (si aplica) para cuantificar la fidelidad de la representación reducida.

- **Balancear métricas con conocimiento de dominio:** Las métricas internas no reemplazan al conocimiento del problema. En nuestro caso, estrictamente DBI y silueta sugerían 2 clústeres óptimos, pero sabíamos que conceptualmente existen 3 clases de iris. Decidimos usar k=3, y las métricas nos mostraron la consecuencia (clústeres más cercanos). En escenarios reales, siempre interpretemos las métricas a la luz del contexto: por ejemplo, un silueta bajo no siempre significa que el clustering es inútil – quizá los datos en sí no tienen fronteras duras sino gradientes continuos. O una varianza explicada de 70% podría ser suficiente si las primeras componentes capturan la señal relevante y el resto es ruido.

- **Herramientas computacionales:** Muchas librerías (scikit-learn, R, MATLAB) ya implementan estas métricas, lo que facilita incorporarlas en el flujo de análisis.

Recomendaríamos automatizar la evaluación: por ejemplo, al probar varios *k* en K-means, generar una tabla/gráfico de silueta, CH, DBI vs. k para decidir óptimo (como se hizo conceptualmente con Iris). Igualmente, para PCA, graficar la varianza acumulada vs componentes (como nuestro scree plot) es una práctica estándar para comunicar la eficacia de la reducción.

En conclusión, **la evaluación rigurosa de modelos no supervisados mediante métricas cuantitativas** nos protege de interpretar patrones aleatorios como si fueran significativos. Las métricas de clustering aseguran que los grupos tengan sentido interno, y las de dimensionalidad aseguran que no comprimimos/visualizamos los datos de forma engañosa. Combinar varias métricas proporciona una imagen más completa y confiable. En nuestro caso de estudio, confirmamos que la agrupación hallada era razonable (aunque no perfecta, capturó bien 1 grupo y medianamente otros 2) y que la proyección 2D era fiel al espacio original. Estas conclusiones sirven para afianzar confianza en el análisis y guiar pasos siguientes (por ejemplo, podríamos decidir usar t-SNE para separar mejor los dos grupos solapados tras ver las métricas, o podríamos reportar que Setosa es fácilmente separable pero las otras dos requieren más atención).

Recomendación final: siempre incluir en informes y presentaciones de análisis no supervisado un apartado de *validación de modelos*, mostrando métricas como las discutidas, para respaldar con evidencia cuantitativa la calidad de los clústeres encontrados o de las reducciones dimensionales utilizadas.

Referencias (formato APA)

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

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

3. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

4. Venna, J., & Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. En *Proceedings of ICANN 2001* (pp. 485–491). London: Springer.