UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la Información: Desarrollo y Gestión de Software

IV. Métricas de evaluación de modelos

IDGS91N - Kevin Iván Aguirre Silva
Extracción de Conocimiento en Bases de Datos - Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., 30 de noviembre de 2025

Índice

1. Introducción

Este documento presenta un análisis de las métricas de evaluación aplicadas a un conjunto de datos de descuentos de Nike, en el contexto de la materia Extracción de Conocimiento en Bases de Datos de la Universidad Tecnológica de Chihuahua. Se utilizaron métricas de agrupamiento (K-means), como el Índice de Silueta, el Índice Davies-Bouldin y el Índice Calinski-Harabasz, para determinar el número óptimo de clústeres. Adicionalmente, se aplicaron métricas de reducción de dimensionalidad (PCA), incluyendo la Varianza Explicada Acumulada y el Error de Reconstrucción, para simplificar la representación de las variables de precio. Los resultados de ambos análisis se comparan para ofrecer conclusiones sólidas y recomendaciones prácticas, como la reducción de los datos a dos dimensiones y la segmentación del catálogo en dos clústeres principales.

2. Métricas de agrupación

Índice de Silueta

- Definición: Mide la calidad de la agrupación comparando la cohesión interna y la separación entre los clusters.
- Formula: Para cada punto i, la silueta es Si = bi – ai / max/(ai, b1), donde ai es la distancia promedio a otros puntos del mismo cluster y bi es la distancia promedio al cluster más cercano distinto.
- Interpretación: Valores cercanos a 1 indican clusters bien separados y cohesionados; valores cerca de 0 indican clusters superpuestos; valores negativos indican posible mala asignación de puntos.
- Ventajas: Proporciona una medida intuitiva y visualizable del ajuste de clusters.
- Limitaciones: Puede ser computacionalmente costoso en datasets grandes y menos útil si los clusters no son esféricos o tienen densidades variadas (Nabi).

Índice Davies–Bouldin

- Definición: Evalúa la calidad del clustering basándose en la relación entre dispersión intra-cluster y distancia inter-cluster.

- Fórmula: Se calcula como el promedio, para cada cluster, del máximo valor de la suma de las dispersión de dos clusters dividida por la separación entre ellos.
- Interpretación: Un valor bajo indica clusters compactos y bien separados; un valor alto indica clusters con poca separación o muy dispersos.
- Ventajas: Es simple y rápido de calcular.
- Limitaciones: Asume clusters esféricos y puede no funcionar bien con clusters de formas irregulares o tamaños muy desiguales (Rodríguez, 2023).

Índice Calinski–Harabasz

- Definición: Mide la relación entre la varianza entre clusters y la varianza dentro de clusters.
- Fórmula: CH = tb(Bk)/(k-1) / tr(Wk) / (n – k), donde tr(Bk) es la traza de la matriz de dispersión entre cluster, tr(Wk) la de dentro de clusters, k el número de cluster y n el número total de puntos.
- Interpretación: Valores más altos indican una mejor formación de clusters (más separados y compactos).
- Ventajas: Es eficaz para distinguir entre diferentes soluciones de clustering.
- Limitaciones: Sensible al número de clusters y no siempre claro cuál es el valor óptimo para detenerse (Help Alteryx, s.f.).

3. Métricas de reducción

Varianza Explicada Acumulada

- Definición: Indica la proporción de la varianza total del dataset capturada por las primeras componentes principales.
- Fórmula: Suma acumulativa de la varianza explicada por cada componente principal. Sigma k i = 1 varianza / varianza total
- Interpretación: Un valor alto indica que las primeras dimensiones capturan la mayor parte de la información original, mientras que un valor bajo indica que se pierde demasiada información al reducir dimensiones.

- Ventajas: Ayuda a elegir el número óptimo de componentes para mantener la mayor información posible.
- Limitaciones: Solo mide varianza, ignorando otros aspectos importantes, y puede ser engañosa en datos no lineales (IBM, 2025).

Error de Reconstrucción

- Definición: Mide la diferencia entre los datos originales y los datos reconstruidos después de la reducción y recuperación dimensional (por ejemplo, en autoencoders).
- Fórmula: Usualmente se calcula como la suma o promedio del error cuadrático medio entre las entradas originales y las reconstruidas.
- Interpretación: Un error bajo indica que la reducción conserva bien la información; un error alto sugiere pérdida significativa de información.
- Ventajas: Evalúa directamente la pérdida de información en términos de datos originales.
- Limitaciones: Puede ser costoso de calcular y no siempre refleja la calidad para tareas específicas como clasificación (IBM, 2025).

4. Descripción del Dataset

El dataset nike_discounts.json contiene 346 registros de productos en oferta de Nike.

Contenido Clave

- Cada registro proporciona detalles del producto e información del descuento:
- Identificación del producto: Nombre (title), subtítulo (subtitle), color (color_description), código (product_code), y URL.
- Información de precios y descuentos: Precio original (original_price), precio actual (current_price), monto del descuento (discount_amount), y porcentaje de descuento (discount_percent).
- Promociones: La mayoría de los artículos mencionan una promoción adicional de "Extra 25% w/ BFRIDAY".

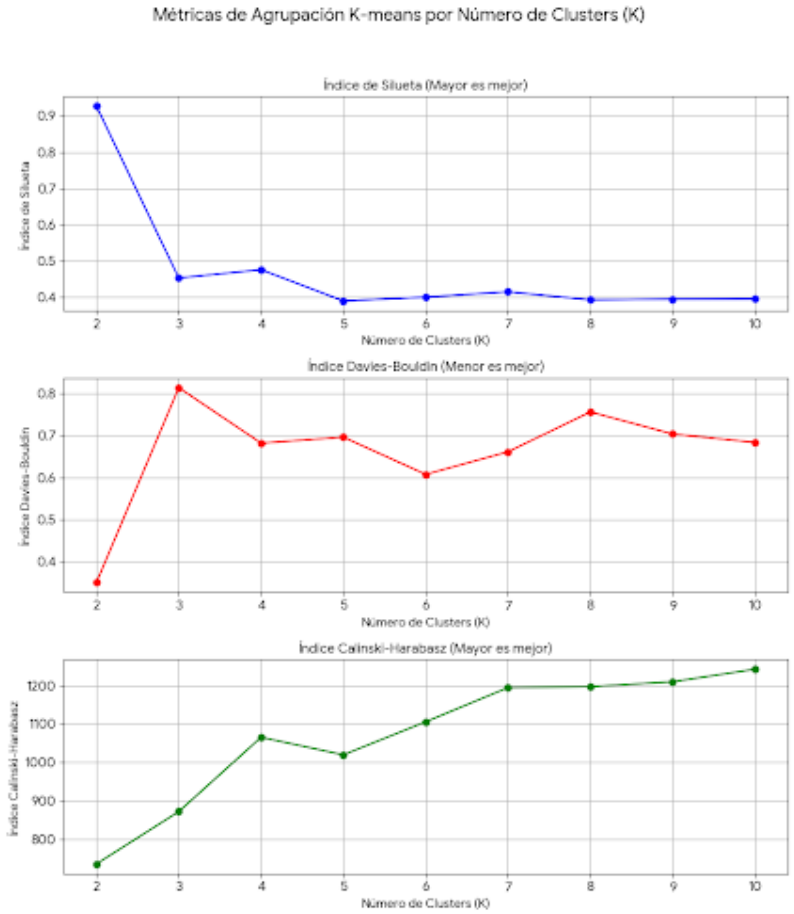
Enfoque del Catálogo

El catálogo se enfoca principalmente en ropa y accesorios para bebés, niños pequeños (Toddler) y niños grandes (Big Kids'), incluyendo conjuntos de 2 o 3 piezas, camisetas, pantalones cortos y chaquetas.

Aspectos Destacados de Precios

- El precio más alto es de $1099.
- El descuento más alto es del 57% (en un conjunto de shorts y camiseta Jordan para niños).

5. Resultados de clustering



Resultados de Métricas de Agrupación K-means

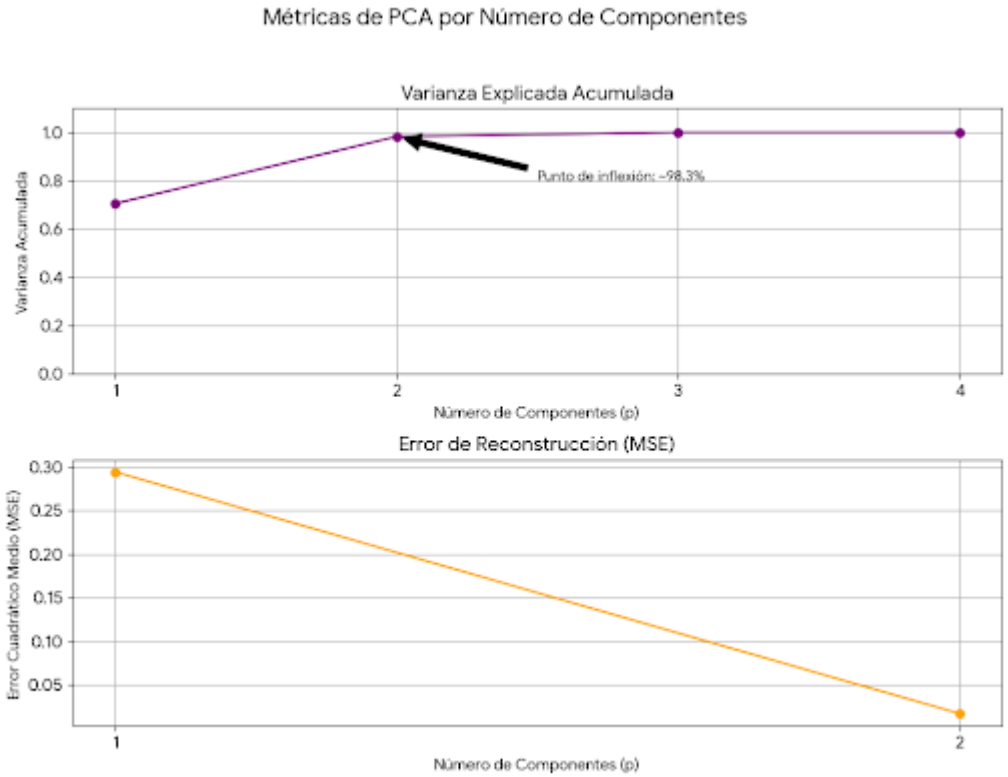
La siguiente tabla resume el rendimiento de las métricas para cada número de clústeres (K):

K	Índice de Silueta	Índice Davies-Bouldin	Índice Calinski-Harabasz
2	0.926599	0.350054	736.746
3	0.453444	0.814343	872.66
4	0.476279	0.682145	1065.53
5	0.390325	0.696362	1020
6	0.400788	0.607663	1105.77
7	0.415408	0.661494	1194.6
8	0.393278	0.756558	1196.99
9	0.395290	0.703867	1209.64
10	0.395724	0.683800	1242.35

Interpretación de las Métricas

Métrica	Objetivo	Mejor Resultado
Índice de Silueta	Mayor es mejor (cercano a 1.0)	**K=2** (0.926599)
Índice Davies-Bouldin	Menor es mejor (cercano a 0.0)	**K=2** (0.350054)
Índice Calinski-Harabasz	Mayor es mejor	**K=10** (1242.35)

6. Resultados de reducción



A continuación, se presentan los resultados de la Varianza Explicada Acumulada y el Error de Reconstrucción (MSE) para diferentes números de componentes:

Número de Componentes (p)	Varianza Explicada Acumulada	Error de Reconstrucción (MSE)
1	70.54%	0.294598
2	98.30%	0.016968
3	100.00%	= 0
4	100.00%	= 0

Visualización de Métricas de PCA

Varianza Explicada Acumulada

La Varianza Explicada Acumulada indica la cantidad total de información (varianza) retenida al mantener un número determinado de componentes principales.

- Con 1 componente se conserva aproximadamente el 70.54% de la varianza total de los datos originales.
- Al aumentar a 2 componentes, la varianza conservada salta a un 98.30%. Esto significa que solo se pierde alrededor del 1.7% de la información original al reducir el espacio de 4 a 2 dimensiones.
- A partir de 3 componentes, se conserva la totalidad de la varianza (100%), ya que las 4 variables originales están casi perfectamente correlacionadas.

Error de Reconstrucción (MSE)

El Error de Reconstrucción (MSE) mide la diferencia promedio cuadrática entre los datos originales estandarizados y los datos reconstruidos a partir del subespacio de menor dimensión. Un valor más bajo indica una mejor aproximación del conjunto de datos original con menos componentes.

- El MSE cae drásticamente de 0.2946 con 1 componente a 0.0170 con 2 componentes.
- Con 3 y 4 componentes, el error es prácticamente cero, lo que confirma que las 4 variables originales son casi linealmente dependientes.

7. Comparativa y análisis

¿Qué métricas funcionan mejor y por qué?

El análisis de las métricas para la agrupación K-means y el análisis de componentes principales (PCA) en el conjunto de datos de descuentos de Nike sugieren una estructura interna fuerte y simple en los datos. Las métricas de agrupación, el Índice de Silueta (con un valor muy alto de 0.926 en K = 2) y el Índice Davies-Bouldin (con el valor más bajo de 0.350 en K = 2), indican de manera robusta que dos clústeres son la solución óptima, ya que capturan la distinción más significativa en los datos. De manera similar, en el PCA, las métricas de Varianza Explicada Acumulada (que alcanza el 98.30%) y el Error de Reconstrucción (que cae a 0.0170) muestran un "codo" claro en dos componentes principales (p = 2), confirmando que la estructura de precios y descuentos se puede simplificar de 4 a 2 dimensiones sin una pérdida de información significativa.

8. Conclusiones y recomendaciones

El análisis de la estructura interna del conjunto de datos de descuentos de Nike revela una alta redundancia en la información de precios, ya que tanto la agrupación K-means como el análisis PCA apuntan fuertemente a una solución óptima de dos componentes o clústeres. Específicamente, métricas como el Índice de Silueta (0.926 en K=2) y la Varianza Explicada Acumulada (98.30% en p=2) sugieren que el dataset está dominado por dos grupos de precios o niveles de descuento principales, y que es posible reducir la complejidad del modelo de 4 a 2 dimensiones sin perder fidelidad. Por lo tanto, se recomienda la reducción de dimensionalidad a dos componentes principales para cualquier análisis posterior o modelado de datos, y considerar dos clústeres para segmentar los productos en futuras estrategias de marketing o inventario.

9. Referencias

Help Alteryx. (s.f.). *K-Centroids Diagnostics Tool Icon Herramienta Diagnóstico de centroides k*. Obtenido de Help Alteryx: https://help.alteryx.com/current/es/designer/tools/predictive-grouping/k-centroids-diagnostics-tool.html

IBM. (6 de Junio de 2025). *Varianza total explicada*. Obtenido de IBM: https://www.ibm.com/docs/es/spss-statistics/31.0.0?topic=detection-total-variance-explained

Nabi, I. (s.f.). *CONSTRUCCIÓN E INTERPRETACIÓN DEL COEFICIENTE SILUETA EN.*

Rodríguez, D. (30 de Junio de 2023). *El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python*. Obtenido de 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/