UNIVERSIDAD DE ESPECIALIDADES ESPÍRITU SANTO

MAESTRÍA EN INTELIGENCIA ARTIFICIAL

ASIGNATURA:
APRENDIZAJE AUTOMÁTICO

TÍTULO:

SEGMENTACIÓN DE USUARIOS MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

AUTORES:
SÁNCHEZ ARÉVALO ANÍBAL ANDRÉS

CABRERA PARRALES JOEL DAVID

MARIA VICTORIA MALDONADO PERALTA

MOYA VERA CARLOS ALBERTO

TUTOR:
MGTR. GLADYS VILLEGAS R.

SAMBORONDÓN, ECUADOR
NOVIEMBRE 2025

Introducción

En el presente trabajo se implementaron y analizaron modelos de aprendizaje no supervisado con el objetivo de segmentar perfiles de usuario y cliente a partir de datos históricos de comportamiento. La información disponible incluye variables de demanda, precio y atributos temporales derivados de la fecha, junto con una etiqueta de tendencia de demanda utilizada como parte de la información disponible.

En lugar de asumir grupos predefinidos, se aplicaron técnicas de clustering y reducción de dimensionalidad para descubrir patrones latentes en los datos:

- K-means y DBSCAN para formar clusters.
- PCA y t-SNE para representar los datos en 2D y facilitar la interpretación visual.

A partir de estos resultados, se buscó identificar perfiles de comportamiento diferenciados que puedan apoyar decisiones de marketing, planificación de inventarios y diseño de estrategias personalizadas.

Métodos Utilizados

1. K-means:

K-means es un algoritmo de clustering particional que requiere definir K (el número de clusters), inicializar K centroides y asignar cada observación al centroide más cercano. Luego actualiza los centroides como el promedio de los puntos de cada cluster y repite hasta converger (Géron, 2023). En este trabajo se utilizó K-means con inicialización k-means++ y se evaluaron diferentes valores de K mediante:

- Método del codo (Elbow): analiza la inercia (suma de distancias al centroide).
- Coeficiente de silueta: mide qué tan bien separados y compactos están los grupos.

Con base en estos criterios se seleccionó un valor de K representativo para el conjunto de datos.

2. DBSCAN:

El Density-Based Spatial Clustering of Applications with Noise es un algoritmo de clustering basado en densidad que no requiere especificar el número de clusters y define grupos como regiones de alta densidad separadas por áreas de baja densidad (Géron, 2023). Tiene dos parámetros principales:

- *eps*: radio de vecindad.
- *min_samples*: mínimo de puntos para considerar una región como densa.

Los puntos que no pertenecen a ninguna región densa se etiquetan como ruido. En el trabajo se usó Nearest Neighbors para explorar la curva de distancias (k-distance plot) y seleccionar un valor adecuado de *eps*, ajustando también *min_samples* en función de la dimensionalidad de los datos.

3. PCA

Principal Component Analysis es una técnica de reducción de dimensionalidad lineal que encuentra combinaciones lineales de las variables originales (componentes principales), ordena estos componentes según la varianza explicada y permite proyectar los datos en menos dimensiones conservando la mayor parte de la información (Géron, 2023).

En el análisis se utilizó PCA con 2 componentes, tanto para medir cuánta varianza explican estos componentes, como para visualizar los clusters de K-means y DBSCAN en un plano 2D.

4. t-SNE

t-distributed Stochastic Neighbor Embedding es una técnica de reducción de dimensionalidad no lineal orientada a visualización que intenta preservar las relaciones de vecindad (puntos cercanos en el espacio original deben seguir siendo cercanos en 2D). Es especialmente útil para detectar estructuras complejas que PCA (lineal) podría no capturar (Géron, 2023).

Debido a su costo computacional, se aplicó t-SNE a una muestra de los datos. Los resultados se usaron solo para visualización y comprensión cualitativa de los clusters producidos por K-means y DBSCAN.

Objetivos del análisis

Los principales objetivos del análisis fueron:

- Comparar el comportamiento y las características de los algoritmos K-Means y DBSCAN, evaluando su capacidad para segmentar el dataset y apoyándose en métricas como el coeficiente de silueta y en visualizaciones proyectadas con PCA y t-SNE.
- Aplicar técnicas de reducción de dimensionalidad (PCA y t-SNE) para facilitar la visualización de los clusters en 2D y verificar si las separaciones identificadas por los modelos tienen coherencia desde una perspectiva lineal y no lineal.
- Identificar patrones generales en los clusters formados por K-Means, describiendo diferencias en términos de variables relevantes del negocio (Sales Quantity, Price,

Promotions, factores estacionales y externos), con el fin de comprender comportamientos distintivos dentro del dataset.

- Derivar conclusiones y recomendaciones que permitan orientar acciones de marketing, segmentación y planeación operativa a partir de la estructura interna revelada por los datos.

Justificación del Análisis

El uso de aprendizaje no supervisado se justifica porque:

- Inicialmente, no se dispone de una segmentación óptima de usuarios o patrones de demanda. En este contexto, los algoritmos de clustering permiten descubrir grupos de forma exploratoria sin necesidad de etiquetas previas.
- La base de datos contiene múltiples variables numéricas, categóricas y temporales que dificultan la interpretación directa. El clustering permite sintetizar esta complejidad en un número reducido de perfiles que representan patrones generales del comportamiento.
- El clustering permite identificar posibles diferencias en variables relevantes del negocio, como volumen de ventas, precios, promociones o estacionalidad, lo que abre la puerta a diseñar estrategias específicas por segmento (por ejemplo, promociones diferenciadas, ajustes de precio o priorización de inventarios).
- Técnicas de reducción de dimensionalidad como PCA y t-SNE permiten validar visualmente si los clusters formados tienen coherencia, ayudando a evaluar si la segmentación refleja verdaderos patrones presentes en los datos.

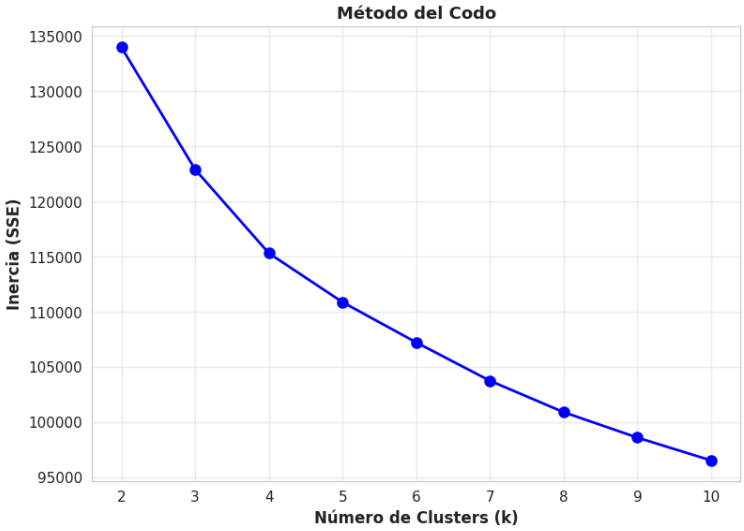
En resumen, el análisis no supervisado ofrece una perspectiva más rica que depender únicamente de métricas individuales o análisis univariado, ya que permite comprender la estructura interna del dataset y detectar patrones que no serían visibles por métodos tradicionales.

Interpretación de los Datos

1. Selección del número óptimo de clusters

El gráfico del método del codo muestra una disminución pronunciada de la inercia entre K = 2 y K = 4, lo que indica que en este rango el algoritmo logra reducir de forma significativa la variabilidad interna de los clusters. A partir de K = 5, la curva comienza a perder pendiente y las mejoras en la inercia se vuelven progresivamente menores, reflejando un rendimiento decreciente al incrementar el número de grupos.

Aunque la curva no presenta un "codo" extremadamente marcado, el punto donde la reducción deja de ser sustancial se ubica alrededor de K = 2, lo que sugiere que este valor ofrece un equilibrio adecuado entre calidad de segmentación y simplicidad del modelo. Por esta razón, K = 2 se considera una elección razonable para el modelo K-means, evitando la sobresegmentación sin sacrificar la capacidad de representar patrones relevantes en los datos.



Resultados numéricos que respaldan la segmentación

- Coeficiente de silueta (K=2): ~0.53: este valor indica una separación adecuada entre los clusters y un nivel razonable de cohesión interna. Aunque no representa una segmentación perfecta, sí sugiere que los grupos poseen fronteras identificables y una estructura interna consistente..
- Varianza explicada por PCA (2 componentes): ~71 %: esto implica que la proyección bidimensional conserva una parte significativa de la información del dataset, permitiendo visualizar la estructura general de los clusters sin perder patrones importantes.
- Proporción de ruido detectada por DBSCAN: prácticamente nula.
- Número de clusters detectados por DBSCAN: un único cluster (Cluster 0).
- DBSCAN no logró separar grupos debido a la densidad homogénea del dataset.

2. Segmentación obtenida con K-means

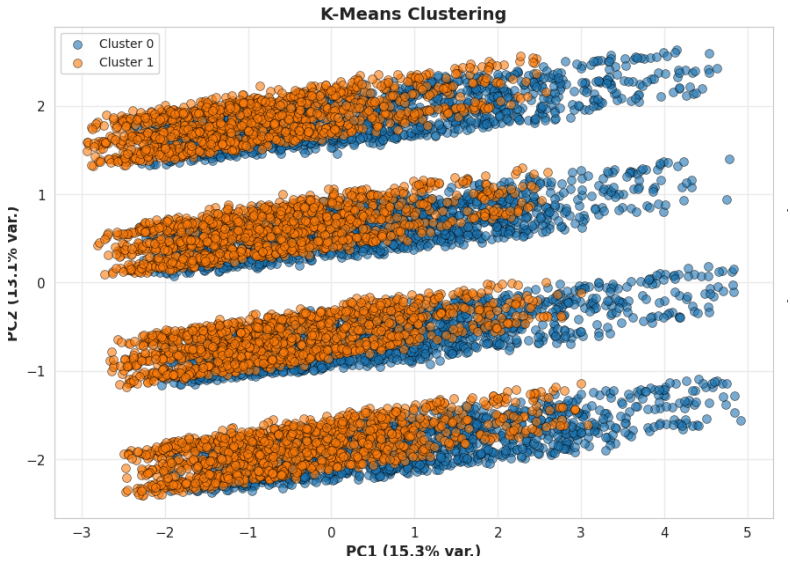
En el gráfico de dispersión proyectado con PCA se observa la segmentación obtenida por K-means con K = 2, donde los puntos se distribuyen en dos grupos diferenciados a lo largo de las bandas paralelas formadas por la combinación de los dos componentes principales.

Aunque los clusters no forman "agrupaciones esféricas" tradicionales, sí presentan patrones de distribución distintos:

- Cluster 0 (azul) tiende a concentrarse hacia la parte superior de cada banda.
- Cluster 1 (naranja) ocupa predominantemente las zonas inferiores de esas mismas bandas.

Este comportamiento sugiere que la separación entre clusters está influenciada por relaciones lineales y graduales entre las variables numéricas, como Sales Quantity, Price y las variables derivadas de la fecha, lo que provoca que la segmentación se alinee con las estructuras paralelas capturadas por PCA.

La consistencia del patrón en todas las bandas indica que los clusters representan formas diferentes de comportamiento latente, donde cada grupo mantiene una tendencia estable dentro de cada franja del espacio. Esto, junto con el coeficiente de silueta satisfactorio, confirma que K-means logra una segmentación válida y útil para interpretar diferencias sistemáticas dentro del dataset.



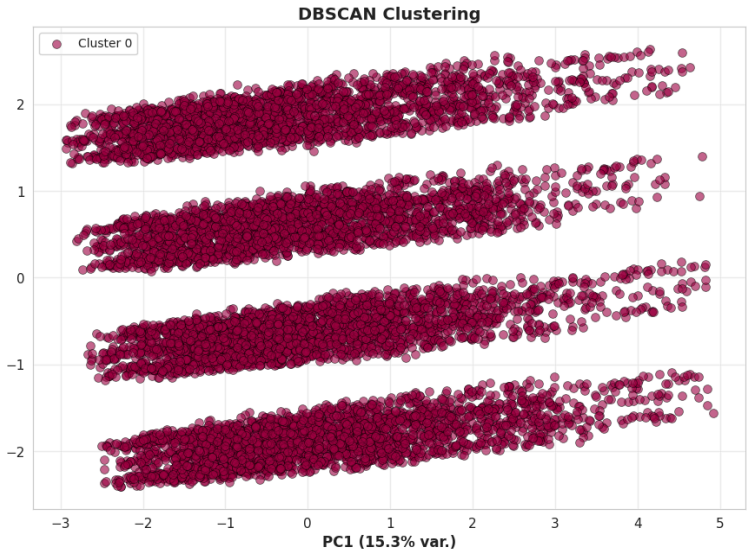
3. Segmentación mediante DBSCAN

El gráfico resultante de DBSCAN muestra un comportamiento muy diferente al observado con K-means. En este caso, el algoritmo no identifica múltiples clusters, sino que agrupa prácticamente todas las observaciones en un solo cluster (Cluster 0), sin distinguir estructuras internas dentro del dataset.

Este resultado indica que las densidades presentes en los datos son demasiado uniformes para permitir que DBSCAN diferencie regiones densas de regiones dispersas. Las bandas

paralelas que aparecen en la proyección PCA mantienen una distribución relativamente homogénea, lo que impide que el algoritmo detecte fronteras naturales basadas en densidad. Como consecuencia, DBSCAN no separa grupos ni identifica outliers relevantes, clasificando casi todos los puntos dentro de una misma región.

En términos prácticos, esto sugiere que DBSCAN no es adecuado para este dataset, ya que la estructura de los datos no presenta variaciones de densidad marcadas, un requisito fundamental para el funcionamiento del algoritmo. En contraste, K-Means sí logra capturar diferencias sistemáticas entre observaciones, por lo que se posiciona como el método más apropiado para generar una segmentación útil en este caso.



4. Visualización PCA

La reducción a dos componentes principales explica aproximadamente el 71 % de la varianza total, lo que indica que PCA logra condensar gran parte de la información relevante del dataset en un espacio bidimensional.

En esta proyección, los datos adoptan una estructura característica en forma de bandas paralelas, resultado de relaciones lineales entre variables como Sales Quantity, Price y factores temporales. Esta estructura es precisamente la que se observa en la gráfica de K-means proyectada sobre los componentes de PCA, donde los dos clusters se distribuyen de manera consistente a lo largo de estas bandas:

- el Cluster 0 predominando en las zonas superiores,
- y el Cluster 1 ubicándose principalmente en las zonas inferiores.

La presencia de estas bandas confirma que las variables originales contienen patrones sistemáticos que PCA logra capturar y representar de manera clara. Además, el comportamiento diferenciado de los clusters dentro de esta proyección respalda que K-means sí está identificando variaciones significativas en la combinación lineal de las características del dataset.

En conjunto, PCA funciona como una herramienta útil para comprender la estructura global de los datos y validar visualmente la segmentación obtenida por K-means.

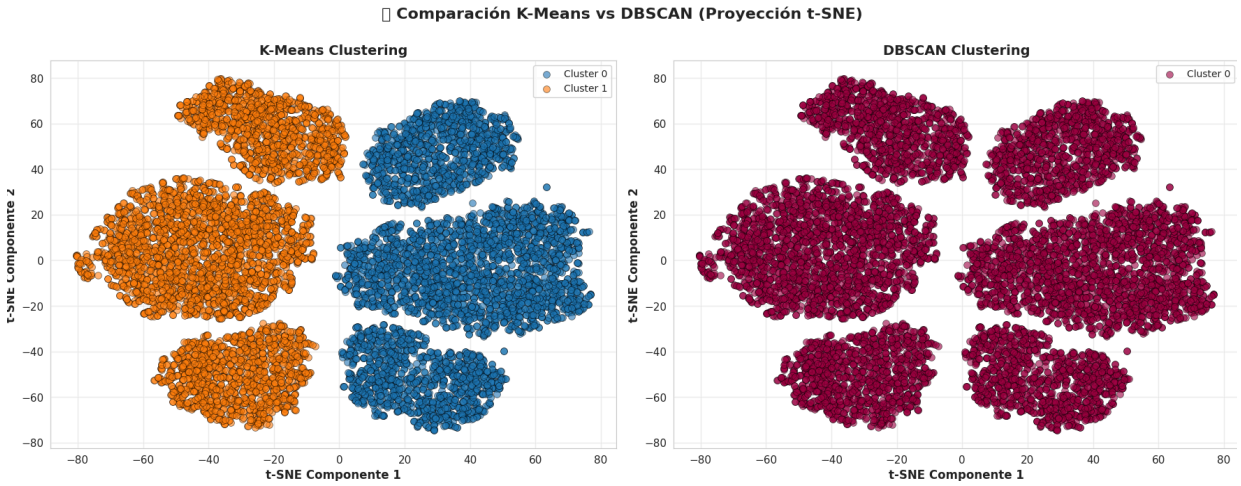
5. Visualización t-SNE

La proyección t-SNE permite observar la estructura no lineal del dataset y comparar visualmente el comportamiento de los algoritmos de clustering.

En el panel izquierdo, donde los puntos están coloreados según los clusters generados por K-means, se aprecia una separación clara entre dos grupos bien definidos. Cada cluster forma una región compacta y diferenciada, lo que indica que K-means logró capturar patrones sistemáticos en la relación entre las variables del dataset, incluso en un espacio no lineal.

En contraste, el panel derecho muestra los resultados de DBSCAN bajo la misma proyección. En este caso, todos los puntos aparecen coloreados como un único cluster, lo que confirma que DBSCAN no identificó una estructura interna significativa en los datos. A pesar de que t-SNE revela varias formaciones visuales, el algoritmo no las reconoce como regiones de distinta densidad y, por tanto, no genera clusters separados.

En conjunto, la visualización t-SNE refuerza la conclusión de que K-means es el método que mejor logra segmentar este dataset, mientras que DBSCAN no encuentra variaciones de densidad suficientes para formar grupos útiles.



Con base en todos los gráficos obtenidos, se pueden destacar cuatro conclusiones principales:

- K-means identifica dos clusters bien diferenciados, tanto en la proyección PCA como en la visualización t-SNE. En ambas representaciones, los grupos se mantienen separados de forma consistente, lo que indica que existen patrones estructurales que el algoritmo logra capturar de manera efectiva.
- DBSCAN no detecta estructura interna en el dataset: agrupa prácticamente todos los puntos en un solo cluster. Esto ocurre porque los datos presentan una distribución relativamente homogénea en densidad, por lo que el algoritmo no identifica regiones suficientemente densas como para formar grupos separados.
- La combinación de PCA y t-SNE demuestra que sí existe cierta separación natural entre observaciones, pero esta separación no está basada en densidades, sino en variaciones sistemáticas capturadas por K-means. Por tanto, K-means es el método que mejor refleja la estructura global del dataset.
- Cada cluster representa un patrón de comportamiento distinto, observable en cómo se distribuyen las observaciones dentro de las bandas proyectadas por PCA y en los grupos bien definidos mostrados por t-SNE. Estos patrones pueden interpretarse posteriormente en función de variables del negocio, como precio, promociones, factores estacionales y tendencia de demanda.

Reflexión y Comunicación

¿Qué tipo de perfiles se pueden identificar?

A partir de la segmentación obtenida con K-means (K = 2) y su proyección en PCA y t-SNE, se identifican dos perfiles principales de usuarios, cada uno asociado a patrones distintos en las variables del dataset:

1. Perfil 1: Usuarios con comportamiento más estable y consistente (Cluster 0)

Este grupo aparece predominantemente en la zona superior de las bandas proyectadas por PCA y forma una región compacta en t-SNE. Su menor dispersión indica:

- Patrones más homogéneos en variables como Sales Quantity y Price.
- Menor sensibilidad a factores externos, ya que su comportamiento se agrupa de forma más estable.
- Comportamientos posiblemente vinculados a segmentos definidos del dataset como Regular o Stable Demand Trend.

Representa a los usuarios cuya demanda tiende a ser más predecible y menos afectada por variaciones de promociones, estacionalidad o factores externos.

2. Perfil 2: Usuarios con mayor variabilidad y sensibilidad a condiciones externas (Cluster 1)

Este grupo domina la parte inferior de las bandas en PCA y se presenta más extendido en la visualización t-SNE. Su mayor dispersión sugiere:

- Mayor heterogeneidad en los patrones de compra, reflejada en diferencias marcadas de Sales Quantity y Price.
- Posible respuesta más fuerte a promociones y seasonality (por ejemplo, categorías impulsivas o más sensibles al precio).
- Más coincidencias con segmentos como Premium, Promotion-driven o Increasing/Decreasing Demand Trend.

Este cluster agrupa a los usuarios más influenciados por factores comerciales y temporales, con patrones menos estables y potencialmente más dinámicos.

En conjunto, los dos clusters revelan:

- Un grupo más estable y homogéneo (Cluster 0)
- Un grupo más volátil y sensible a variaciones externas (Cluster 1)

La estructura clara mostrada en t-SNE confirma que estas diferencias no son aleatorias, sino reflejo de patrones reales en las variables comerciales del dataset.

¿Qué diferencias clave surgieron entre los modelos?

El comportamiento de K-means y DBSCAN fue notablemente diferente, lo que permitió identificar fortalezas y limitaciones específicas de cada método en este dataset:

1. K-Means logró identificar una estructura clara de dos clusters

K-means encontró dos grupos bien diferenciados, visibles tanto en la proyección PCA como en la visualización t-SNE. Los clusters muestran fronteras definidas, patrones consistentes y separación estable en diferentes proyecciones. Esto indica que las variables del dataset contienen variaciones sistemáticas que el modelo pudo capturar de manera efectiva.

2. DBSCAN no identificó estructura: agrupó todo en un solo cluster

DBSCAN clasificó prácticamente todas las observaciones dentro de un único cluster (Cluster 0) y no detectó grupos separados ni ruido relevante. Esto se debe a que el dataset presenta densidad relativamente homogénea, patrones distribuidos en bandas paralelas lineales (capturadas por PCA) y ausencia de regiones densas y dispersas diferenciadas.

DBSCAN, al depender estrictamente de densidad, no encuentra fronteras naturales en este tipo de distribución.

3. PCA y t-SNE revelan la misma conclusión: separación lineal sí, separación por densidad no

Ambas técnicas de reducción dimensional muestran dos grupos bien separados cuando se colorean según K-means y una sola masa uniforme cuando se colorean según DBSCAN. Esto confirma que la estructura del dataset no está basada en densidad, sino en variaciones lineales/multivariadas capturadas por K-means.

¿Qué limitaciones encontraron y cómo las abordarían?

El análisis presentó varias limitaciones a considerar:

Limitación	Cómo abordarlo
Homogeneidad de densidad en los datos (afecta directamente a DBSCAN) El dataset presenta una distribución relativamente uniforme en forma de bandas paralelas, lo que impide que DBSCAN identifique regiones densas y dispersas de manera diferenciada. Agrupa casi todos los puntos en un solo cluster, volviéndose inútil para segmentación.	- Probar algoritmos basados en densidad más flexibles, como HDBSCAN. - Realizar ingeniería de características para generar variables que sí reflejen cambios de densidad (ej. ratios, variaciones temporales, interacciones).
Dependencia fuerte de los hiperparámetros (K en K-Means y eps/min_samples en DBSCAN) El resultado del clustering varía mucho según los valores elegidos. K-Means requiere validar distintos valores de K, y DBSCAN es extremadamente sensible a *eps*.	- Utilizar métricas sistemáticas como silhouette, elbow, Davies-Bouldin. - Aplicar búsqueda más robusta: grid search + validación cruzada no supervisada. - Justificar los parámetros con análisis visual y estadístico.
Limitación interpretativa al usar PCA y t-SNE Las proyecciones reducen la dimensionalidad, pero PCA solo captura relaciones lineales, Y t-SNE no preserva distancias reales ni escalas. Las	- Complementar con análisis de las medias y distribuciones por cluster (Sales Quantity, Price, Promotions, etc.).

visualizaciones ayudan, pero no pueden usarse para interpretar directamente el comportamiento comercial sin revisar las variables originales.	- Utilizar alternativas como UMAP, que combina linealidad y no linealidad.
Falta de interpretación profunda sin análisis de características por cluster Aunque K-Means separa dos grupos, la interpretación de negocio queda limitada si no se examinan las características internas de cada cluster. Los perfiles obtenidos se basan principalmente en patrones geométricos, no en análisis descriptivo de las variables del negocio.	- Calcular promedios, medianas y proporciones por cluster (Sales Quantity, Price, % Promotions, etc). - Crear tablas y gráficos comparativos entre clusters.
Naturaleza sintética del dataset El dataset es generado artificialmente (no es un dataset real de una empresa), lo que implica que las relaciones entre variables pueden no reflejar el comportamiento real de los consumidores, y algunos patrones (como las bandas paralelas en PCA) son consecuencia directa de cómo se generaron los datos.	- Utilizar datos reales cuando sea posible. - Validar resultados con expertos de negocio. - Ajustar modelos según necesidades del caso real.

Conclusiones

El uso combinado de K-means y DBSCAN permitió analizar diferentes perspectivas de segmentación dentro del dataset. K-Means logró identificar dos clusters bien diferenciados, visibles tanto en la proyección PCA como en la visualización t-SNE, lo que evidencia que las variables del dataset contienen patrones estructurales que el algoritmo pudo capturar de manera consistente. En contraste, DBSCAN no identificó una estructura interna útil, ya que la distribución relativamente homogénea de densidad llevó a que todas las observaciones fueran agrupadas en un único cluster, sin presencia significativa de ruido. Esto confirma que K-means es el método más adecuado para segmentar este tipo de datos.

La reducción de dimensionalidad con PCA y t-SNE fue fundamental para validar visualmente los resultados. PCA mostró la presencia de bandas paralelas que explican la

separación obtenida por K-means, mientras que t-SNE reforzó la existencia de dos grupos no lineales claramente diferenciados. Ambas técnicas aportaron evidencia de que la estructura captada por K-means es estable y consistente desde distintos enfoques.

Si bien el análisis permitió distinguir dos patrones generales de comportamiento, la interpretación profunda de los clusters se ve limitada por la necesidad de revisar directamente las características del negocio (Sales Quantity, Price, Promotions, Seasonality, External Factors, etc.) dentro de cada grupo. Este paso adicional permitiría identificar diferencias concretas entre los segmentos y traducir los resultados en acciones operativas o comerciales.

En conjunto, el análisis no supervisado aporta una comprensión sólida de la estructura subyacente del dataset y constituye un complemento valioso para cualquier enfoque supervisado. Los resultados permiten sentar las bases para futuras estrategias de segmentación, optimización de precios, análisis de demanda o personalización, y demuestran la utilidad de combinar técnicas de clustering con métodos de reducción de dimensionalidad para obtener una visión más completa del comportamiento de los datos.

Referencias Bibliográficas

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (3rd ed.). O'Reilly Media.