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 en la parte supervisada del análisis.

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 desempeño y las características de K-means y DBSCAN como métodos de clustering, apoyándose en visualizaciones y métricas como el coeficiente de silueta.
- Reducir la dimensionalidad del conjunto de datos usando PCA y t-SNE para facilitar la visualización de los clusters en 2D y verificar si las separaciones encontradas por los algoritmos tienen sentido en el espacio de características.
- Interpretar los perfiles de cada cluster en términos de variables de negocio (niveles de demanda, comportamiento temporal, etc.) y relacionarlos con la tendencia de demanda (*Demand Trend*).

- Derivar conclusiones y recomendaciones que puedan orientar acciones de marketing, segmentación y planeación operativa.

Justificación del Análisis

El uso de aprendizaje no supervisado se justifica porque:

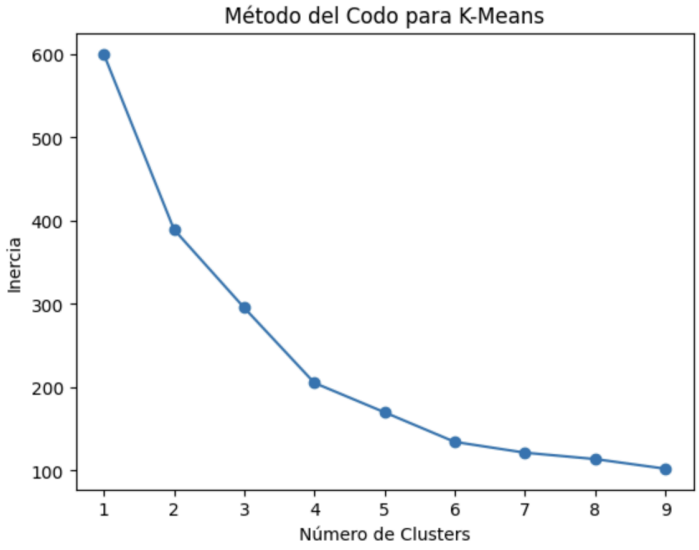
- No se dispone, a priori, de una segmentación óptima de usuarios o patrones de demanda. Es necesario descubrir grupos de forma exploratoria.
- La base de datos contiene múltiples variables numéricas y temporales que dificultan la interpretación directa; los algoritmos de clustering permiten sintetizar esta complejidad en un número reducido de perfiles.
- Complementar el análisis supervisado de clasificación (predicción de *Demand Trend*) con una visión de segmentos naturales en los datos ayuda a:
- Entender si existen grupos con mayor probabilidad de demanda alta o baja.
- Diseñar campañas o estrategias específicas para cada tipo de perfil.
- Técnicas como PCA y t-SNE permiten verificar visualmente si los clusters tienen sentido, evitando decisiones basadas solo en métricas numéricas.

En resumen, este enfoque aporta una perspectiva más rica que utilizar únicamente modelos supervisados, ya que explora la estructura interna de los datos sin requerir etiquetas.

Interpretación de los Datos

1. Selección del número óptimo de clusters

El gráfico del método del codo mostró que la inercia disminuye de forma pronunciada entre K=1 y K=3 y luego, a partir de K=4, la mejora se vuelve marginal. Ese punto de inflexión indica el número óptimo de clusters, ya que a partir de ahí aumentar K no aporta una ganancia significativa. Este comportamiento justificó la elección de K para el modelo K-means.



Coeficiente de silueta (K=4): ~0.53

Este valor indica una separación adecuada entre los clusters y una cohesión interna razonable.

Varianza explicada por PCA: ~71 %

Esto significa que la proyección en 2D conserva buena parte de la información relevante del dataset.

Proporción de ruido detectada por DBSCAN: ~28 %

Este porcentaje elevado confirma que los datos no presentan densidades homogéneas.

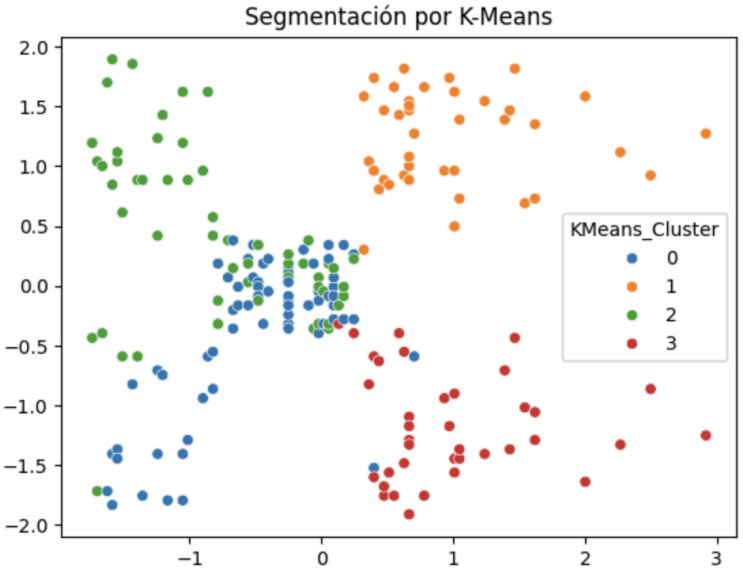
Número de clusters detectados por DBSCAN: 2 clusters densos + ruido

Esto refuerza que K-means es el método que mejor captura la estructura del dataset.

2. Segmentación obtenida con K-means

En el gráfico de dispersión de K-means (segunda imagen) se aprecia que los cuatro clusters están distribuidos en regiones distintas del espacio. Los grupos presentan formas relativamente compactas, que es lo esperado en K-means porque forma clusters esféricos, y existe una separación clara entre los clusters rojo, verde, azul y naranja. Esto indica que las variables utilizadas sí contienen patrones diferenciales.

K-means logra una partición bien definida y equilibrada, con grupos de tamaño relativamente similar, lo que lo convierte en un buen método de segmentación para estos datos.

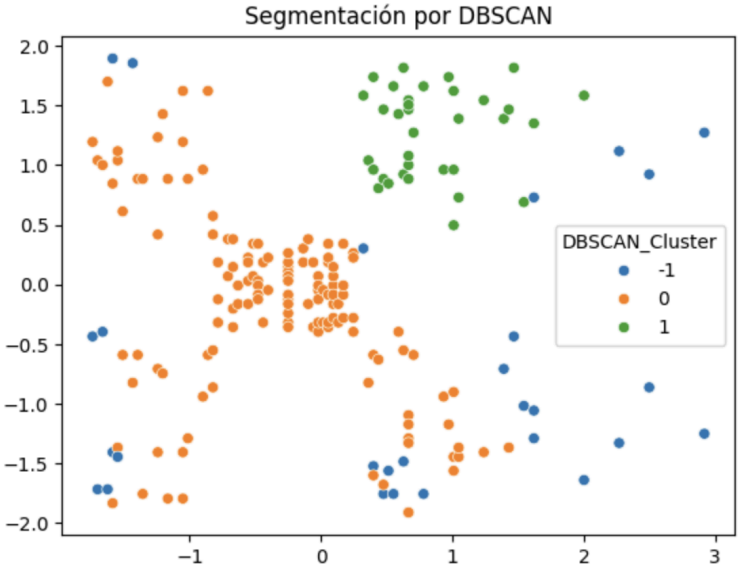


3. Segmentación mediante DBSCAN

El gráfico de DBSCAN muestra resultados distintos a K-means, ya que identifica dos clusters principales (0 y 1) y además detecta un grupo considerable de puntos como ruido (etiqueta -1). Los clusters encontrados son más "orgánicos", ya que DBSCAN se basa en densidad, no en formas esféricas.

DBSCAN identifica regiones densas, pero considera muchas observaciones como ruido. Esto sugiere que los datos no tienen densidades homogéneas y están esparcidos en el espacio. Comparativamente, K-means identifica 4 clusters bien formados, mientras que DBSCAN detecta solo 2 clusters densos y un conjunto grande de outliers.

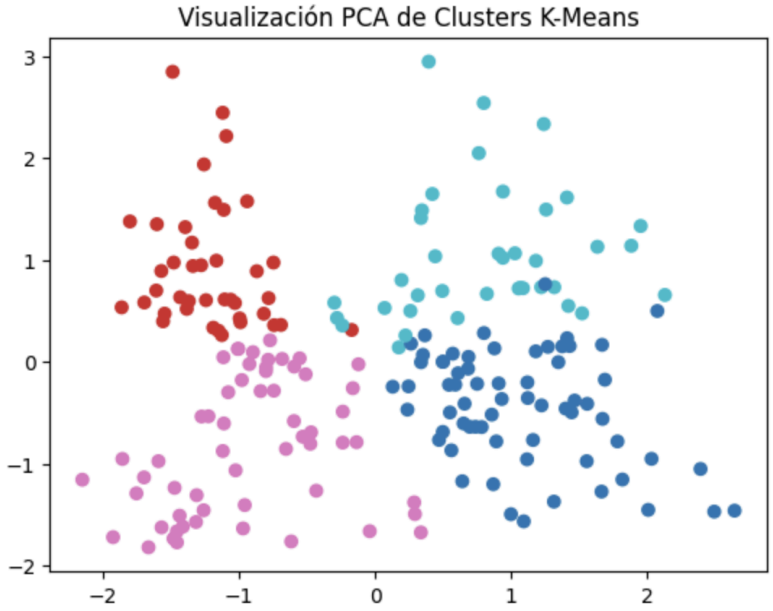
Esto indicaría que K-means funciona mejor para este dataset porque su estructura es más compatible con clusters compactos y equidistantes. Por otro lado, DBSCAN es útil para detectar outliers, pero no como método principal de segmentación.



4. Visualización PCA

El gráfico de PCA muestra los clusters de K-means proyectados en dos componentes principales. Los cuatro clusters se visualizan bien separados en el espacio 2D, especialmente:

- Un cluster (rojo) claramente separado hacia la izquierda.
- Otro cluster (azul) hacia la derecha.
- Dos clusters (rosado y celeste) en zonas más intermedias.



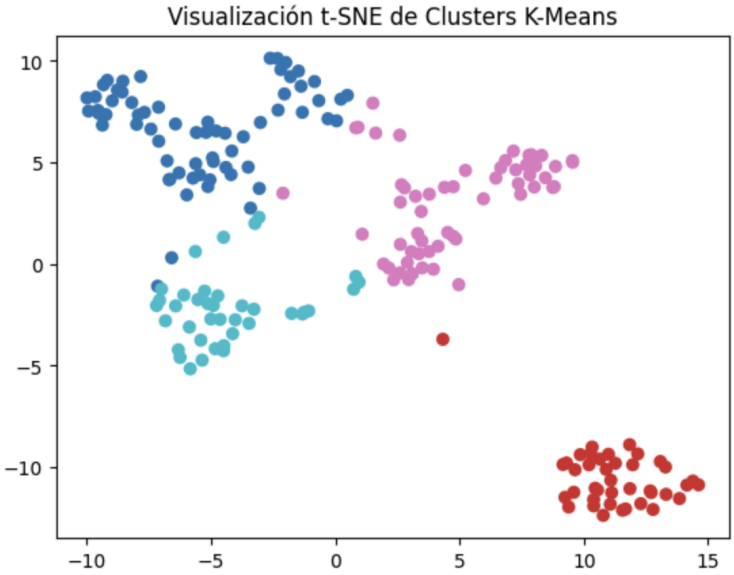
La separación sugiere que las variables originales sí contienen información relevante para segmentar. De esta forma, PCA confirma que la estructura encontrada por K-means tiene coherencia, y los clusters parecen responder a comportamientos distintos en los datos (por ejemplo: niveles de demanda, fluctuaciones, patrones temporales, etc.).

5. Visualización t-SNE

La representación con t-SNE, que capta relaciones no lineales, refuerza los hallazgos:

- Los cuatro clusters vuelven a agruparse de manera definida y clara.
- t-SNE revela formas más irregulares y patrones más complejos:
  - Un cluster (rojo) aparece totalmente separado y más denso.
  - Otros clusters (rosa, azul y celeste) mantienen separación relativa pero con estructuras más intrincadas.

t-SNE confirma que los clusters no sólo se separan linealmente, sino también en el espacio no lineal, lo que valida aún más la segmentación de K-means.



Con base en todos los gráficos tenemos cuatro puntos distintivos:

1. K-means encuentra 4 perfiles de clientes/usuarios bien diferenciados: los grupos tienen fronteras claras en PCA y t-SNE.
2. DBSCAN detecta solo 2 grupos densos y muchos outliers: lo cual es típico cuando los datos no presentan una densidad homogénea.
3. La combinación de PCA y t-SNE confirma que sí existe una estructura natural en los datos: y que esta estructura es mejor capturada por K-means.

4. Cada cluster representa un patrón de comportamiento distinto: y que después este debe interpretarse, por ejemplo: alto volumen, precios distintos, ciclos temporales, etc.

Conclusiones

El uso combinado de K-means y DBSCAN permitió explorar diferentes perspectivas de segmentación. K-means ofreció una partición clara en un número fijo de grupos, mientras que DBSCAN permitió detectar estructuras de densidad y puntos atípicos que no encajan en ningún cluster.

La reducción de dimensionalidad con PCA y t-SNE fue clave para visualizar la calidad de las particiones y confirmar si la elección de K en K-means era razonable. Esto permitió así mismo entender mejor la forma real de los clusters en un espacio de menor dimensión.

El análisis de las características medias por cluster mostró que los grupos difieren en variables relevantes para el negocio (como niveles de demanda y comportamiento temporal), lo que respalda el uso de estos perfiles para definir estrategias diferenciadas.

La relación entre los clusters y la tendencia de demanda (*Demand Trend*) sugiere que ciertos perfiles están más asociados a demanda alta, mientras que otros concentran demanda baja o media. Esto abre la puerta a:

- Planificar acciones específicas por segmento (promociones, cambios de precio, estrategias de retención).
- Priorizar recursos en aquellos clusters con mayor potencial de crecimiento.

En conjunto, el análisis no supervisado complementa la parte supervisada de clasificación y proporciona una comprensión más profunda de la estructura de los datos, facilitando la toma de decisiones informadas en marketing y operaciones.

Referencias Bibliográficas

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