# Descubrimiento de patrones de comportamiento del consumidor a través del aprendizaje no supervisado

Luis David Huante García
Tecnologías para la Información en Ciencias
UNAM ENES Morelia
luisdhuante@gmail.com

**Figure 1: La segmentación de clientes permite a las empresas identificar y abordar las necesidades específicas de diversos grupos, mejorando así la eficacia de sus estrategias de mercado.**

## ABSTRACT

Este proyecto explora la aplicación de técnicas de aprendizaje no supervisado para segmentar eficazmente el comportamiento del consumidor en el ámbito minorista. Se identificaron dos segmentos principales de clientes, destacando diferencias significativas en sus patrones de compra y respuestas a las campañas de marketing.

El primer segmento incluye consumidores de alto ingreso con preferencias por productos de lujo y una tendencia a comprar en tiendas y por catálogo. El segundo segmento comprende consumidores de ingresos más bajos, más inclinados a realizar compras con descuentos y frecuentes visitas al sitio web. Los resultados ilustran cómo el aprendizaje no supervisado puede ayudar en el diseño de estrategias de marketing dirigidas y efectivas.

## KEYWORDS

aprendizaje no supervisado, segmentación de clientes, comportamiento del consumidor, clustering jerárquico, minería de datos, marketing minorista, patrones de compra, análisis de mercado, promociones cruzadas, estrategias de marketing

**ACM Reference Format:**
Luis David Huante García. 2024. Descubrimiento de patrones de comportamiento del consumidor a través del aprendizaje no supervisado. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCCIÓN

En la era actual, las empresas enfrentan el desafío constante de comprender y prever el comportamiento del consumidor. La capacidad de segmentar eficazmente a los clientes según sus patrones de compra y preferencias puede proporcionar una ventaja competitiva significativa. En este contexto, el aprendizaje no supervisado es

una herramienta muy útil, capaz de descubrir estructuras ocultas y segmentos de clientes que pasan desapercibidos a simple vista.

Este proyecto se enfoca en la aplicación de técnicas de aprendizaje no supervisado para analizar el comportamiento del consumidor dentro de un entorno minorista. A través del uso de KMeans, Clustering jerárquico, Análisis exploratorio de datos y demás herramientas [sci [n. d.]], se busca identificar segmentos distintos de consumidores, basados en sus interacciones de compra y respuestas a las campañas de marketing. Este enfoque permite revelar cómo diferentes grupos de clientes interactúan con la gama de productos y promociones ofrecidas.

## 1.1 Descripción del Dataset

El conjunto de datos utilizado en este proyecto incluye 2240 instancias con varios atributos que se clasifican en categorías como datos personales, gastos en productos, promociones y comportamientos de compra en diferentes plataformas.

*1.1.1 Datos personales.* Incluyen ID del cliente, año de nacimiento, nivel educativo, estado civil, ingresos anuales del hogar, número de niños y adolescentes en el hogar, fecha de inscripción con la empresa, y la recencia de la última compra. Además, se incluye un atributo que indica si el cliente ha presentado quejas en los últimos dos años.

*1.1.2 Gastos en productos.* Información sobre los montos gastados en diferentes categorías de productos en los últimos dos años, incluyendo vinos, frutas, productos cárnicos, pescados, dulces y productos de oro.

*1.1.3 Promociones.* Datos sobre la participación del cliente en las ofertas promocionales de cinco campañas distintas, así como el número de compras realizadas con descuento.

*1.1.4 Comportamientos de compra.* Número de compras realizadas a través del sitio web de la empresa, compras realizadas utilizando un catálogo, compras directas en tiendas, y el número de visitas al sitio web en el último mes.

Los datos para este análisis están disponibles en Kaggle bajo el nombre de "Customer Personality Analysis". [Romero-Hernandez. 2021]

# 2 ANÁLISIS EXPLORATORIO DE DATOS (EDA)

En el marco de este proyecto, se llevó a cabo un análisis exploratorio de datos (EDA) para profundizar en la comprensión del conjunto de datos y asegurar los análisis de clustering subsecuentes. El EDA comenzó con la limpieza de datos, donde se identificaron y trataron valores faltantes, se eliminaron duplicados, y se verificaron inconsistencias en los datos. Esto permitió identificar tendencias, patrones y posibles outliers que podrían influir en los resultados del análisis de clustering.

Además, se realizaron visualizaciones detalladas que incluyeron histogramas, diagramas de caja y gráficos de dispersión para explorar las relaciones entre las variables. Esto no solo facilitó la detección de relaciones entre diferentes características de los consumidores, sino que también ayudó a comprender cómo las distintas variables influían en las respuestas a las campañas de marketing. A continuación se explicarán a fondo los hallazgos encontrados por el análisis:

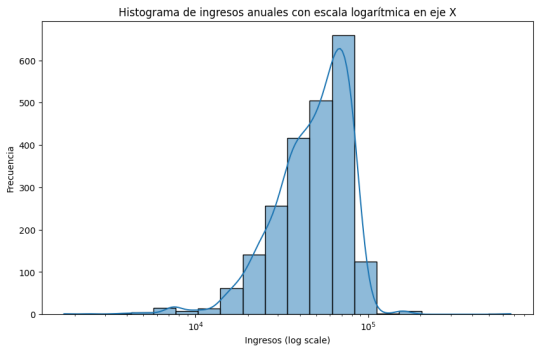


**Figure 2: Histograma de ingresos anuales con escala logarítmica en eje X**

En el histograma, observamos la distribución de ingresos anuales representada en una escala logarítmica en el eje X, lo que facilita la visualización de un rango amplio de ingresos. La forma del histograma, aproximadamente de una distribución log-normal, sugiere que la mayoría de los individuos gana menos, con un pico en el rango de ingresos medios, mientras que unos pocos ganan mucho más, evidenciado por la cola larga hacia la izquierda del gráfico. El uso de la escala logarítmica permite discernir con mayor claridad las diferencias en los rangos de ingresos bajos y medios, áreas que en escalas lineales convencionales pueden aparecer comprimidas. A pesar de que el gráfico ya revela información, podemos confirmarla con más detalle con una gráfica de caja.

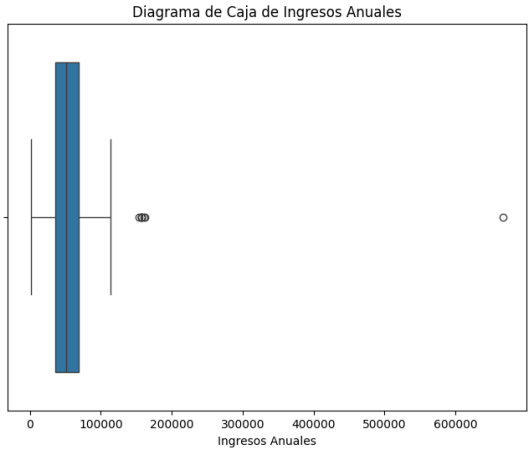


**Figure 3: Diagrama de caja de ingresos anuales**

El diagrama de caja de ingresos anuales muestra que los ingresos rondan alrededor de los $50,000, sin embargo algunos puntos quedan fuera de estos límites, por lo que es importante eliminarlos para tener más certeza sobre la información de los ingresos.

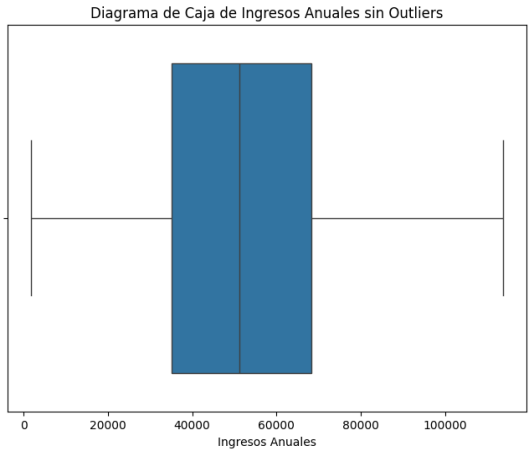


**Figure 4: Diagrama de caja de ingresos anuales sin outliers**

La eliminación de los valores atípicos ha permitido que la escala del eje x se ajuste a un rango más estrecho de ingresos, de 0 a 100,000, mejorando la claridad y permitiendo una interpretación más precisa de la distribución general de los ingresos.

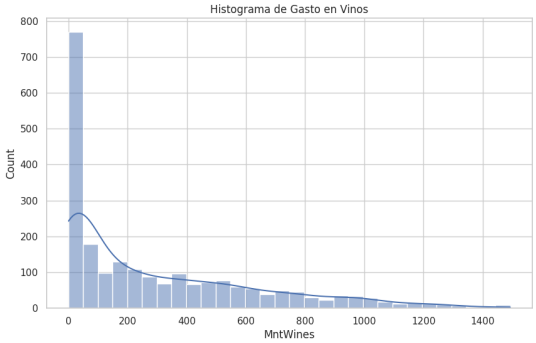


**Figure 5: Histograma de gasto en vinos**

La gráfica muestra el histograma de gasto en vinos, etiquetado como "MntWines" en el eje x, que representa el monto gastado. La distribución de los gastos en vinos es marcadamente sesgada hacia la derecha, lo que indica que la mayoría de los clientes gastan menos en vinos, con una concentración significativa de datos en el extremo inferior del rango de gasto. Observamos que la mayor frecuencia de gastos se encuentra en el rango de 0 a 200, con un pico muy pronunciado cerca de 0, sugiriendo que un gran número de clientes gasta poco o nada en vinos. A medida que el monto del gasto en vinos aumenta, la frecuencia de clientes disminuye de manera consistente. Los clientes que gastan más de 800 son notablemente menos, indicando que altos gastos en vinos son menos comunes entre la población estudiada.

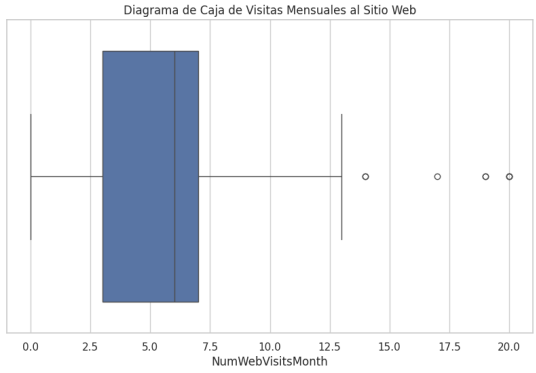


**Figure 6: Diagrama de caja de tráfico en página web**

La gráfica muestra un diagrama de caja de las visitas mensuales a un sitio web, ilustrando la mediana y el rango intercuartílico (IQR) de las visitas. La mayoría de las visitas se concentra entre aproximadamente 5 y 7.5 por mes, indicando un nivel de interacción moderado de los usuarios con el sitio. Además, se observan varios outliers, con valores alrededor de 15 y 20 visitas mensuales, lo que sugiere que algunos usuarios tienen una interacción significativamente mayor con el sitio. Estos puntos atípicos podrían señalar áreas de interés especial para estrategias de marketing o ajustes en la gestión del contenido web.

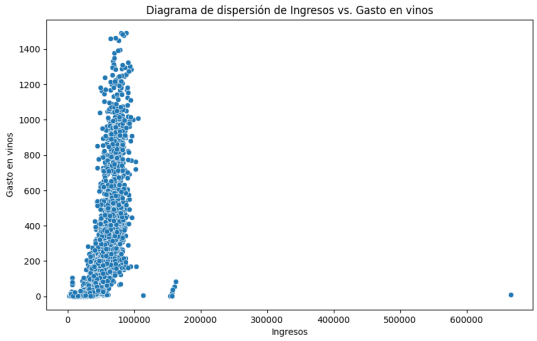


**Figure 7: Diagrama de dispersión de Ingresos vs. Gasto en vinos**

El anterior es un diagrama de dispersión que representa la relación entre los ingresos y el gasto en vinos. Observemos que la mayoría de los puntos se concentran en el rango bajo de ingresos. A medida que los ingresos aumentan, el gasto en vinos no muestra un incremento proporcional significativo, lo que sugiere que no hay una correlación fuerte entre mayores ingresos y un aumento en el gasto en vinos.

Notablemente, hay muy pocos puntos en el rango más alto de ingresos, y algunos de estos puntos indican gastos bajos en vinos, lo cual es inusual comparado con la tendencia general.

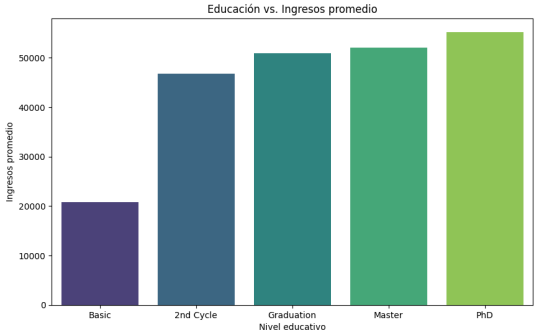


**Figure 8: Gráfica de Educación vs Ingresos promedio**

La anterior gráfica muestra la correlación entre el nivel educativo y los ingresos promedio, resaltando cómo los ingresos tienden a aumentar con niveles de educación más altos. Comenzando con el nivel básico, los ingresos son los más bajos. A medida que se asciende a "2nd Cycle", se observa un incremento significativo en los ingresos.

Al avanzar a niveles de educación superior, como "Graduation" y "Master", los ingresos se estabilizan, mostrando poca variación entre estos dos niveles, pero manteniéndose significativamente más altos que los niveles de educación inferior.

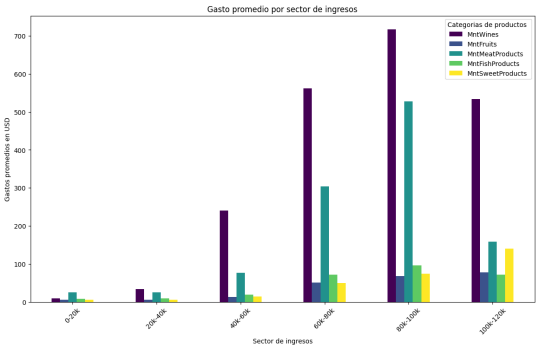


**Figure 9: Gasto promedio por sector**

El gráfico anterior ilustra el gasto promedio en diferentes categorías de productos segmentado por niveles de ingreso. Observamos que la categoría de vinos (MntWines) domina el gasto en todos los segmentos de ingresos, pero es especialmente prominente en los rangos de 60k-80k y 100k-120k USD. Este patrón sugiere que los consumidores con ingresos más altos tienden a gastar más en vinos, posiblemente reflejando un estilo de vida lujoso o una preferencia por productos premium, osbervación que confirmaremos más adelante en el análisis.

Por otro lado, las categorías de productos de carne (MntMeatProducts) y productos de pescado (MntFishProducts) también muestran un aumento en el gasto conforme aumentan los ingresos, aunque no tan marcado como en la categoría de vinos. Las categorías de frutas (MntFruits) y productos dulces (MntSweetProducts) muestran un gasto considerablemente menor y no presentan grandes aumentos con los ingresos más elevados.

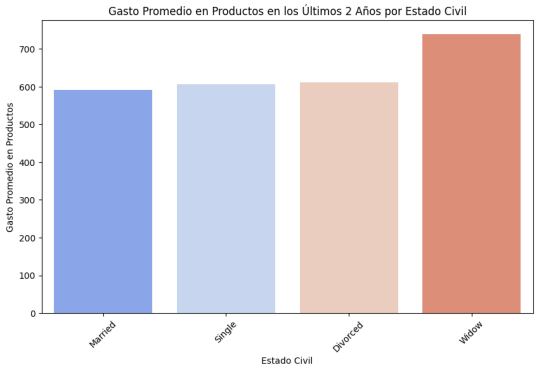


**Figure 10: Gasto promedio por estado civil**

Se observa que las personas viudas presentan el gasto promedio más alto en productos. Los divorciados, aunque gastan menos que los viudos, también muestran un gasto significativo, posiblemente debido a la transición hacia un nuevo estilo de vida. Los casados y solteros tienen un gasto similar y considerablemente más bajo, lo que refleja quizás una estabilidad económica o patrones de consumo más regulados.

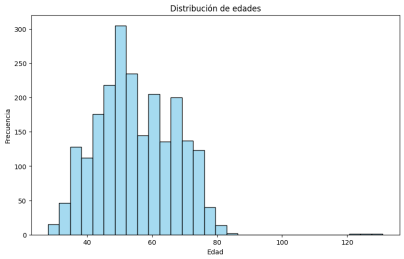


**Figure 11: Distribucón de edades**

En esta gráfica se hizo un histograma que muestra la frecuencia de edades. Observamos que la distribución es bimodal, con dos picos prominentes: uno alrededor de los 50 años y otro menor cerca de los 70 años. Las barras indican que hay una concentración significativa de individuos en estos dos rangos de edad, mientras que las edades extremadamente jóvenes y mayores son menos frecuentes.

## 3 ANÁLISIS NO SUPERVISADO

La siguiente sección detalla el proceso y las metodologías empleadas en el aprendizaje no supervisado. Este enfoque comenzó con el preprocesamiento de datos, que incluyó la normalización de las variables y eliminación de datos atípicos y nulos. Posteriormente, se aplicaron métodos avanzados de clustering para identificar patrones y agrupar efectivamente a los consumidores según sus características y comportamientos. Además, se exploraron técnicas de reducción de dimensionalidad para mejorar la eficiencia y la efectividad del modelado.

### 3.1 Preprocesamiento de datos

En el preprocesamiento de los datos, se seleccionaron las columnas numéricas relevantes, excluyendo aquellas que no son útiles para

el análisis posterior. Posteriormente, se llenan los valores faltantes con la mediana de cada columna.

## 3.2 Principal Component Analysis (PCA)

El Análisis de Componentes Principales (PCA) sirve para reducir la dimensionalidad de un conjunto de datos numéricos, facilitando así la visualización y el análisis posterior. Antes de aplicar PCA, es crucial estandarizar los datos, dado que PCA es sensible a las variaciones en la escala de las variables. Para esto, se emplea 'StandardScaler' de 'scikit-learn', que ajusta y transforma los datos para que cada característica tenga una media de cero y una desviación estándar de uno. Este paso asegura que todas las características contribuyan equitativamente al análisis sin sesgar los resultados hacia variables con mayor magnitud.

Una vez estandarizados los datos, se aplica PCA, especificando que se desea reducir la dimensionalidad a dos componentes principales. Este método funciona identificando los ejes que maximizan la varianza de los datos, proporcionando así las direcciones principales en las cuales los datos están más dispersos. Los dos primeros componentes principales capturan la mayor parte de la variabilidad en los datos, mientras que reducen la complejidad del espacio original, lo que es ideal para identificar patrones.

Finalmente, se evalúa la varianza explicada por cada componente principal para entender cuánta información de los datos originales se conserva en los componentes seleccionados. Este análisis es crucial, ya que proporciona una medida de la eficacia del PCA en la reducción de dimensionalidad y ayuda a determinar si los componentes seleccionados son suficientes para representar las estructuras clave de los datos.

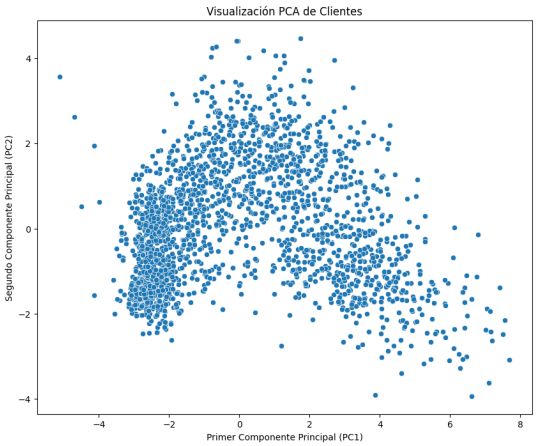


**Figure 12: Gráfico de PCA con 2 componentes**

La gráfica muestra la distribución de los clientes en un espacio bidimensional generado por el PCA, representando los dos primeros componentes principales (PC1 y PC2). En este espacio, cada punto representa un cliente, donde las coordenadas del punto se derivan de los dos principales componentes que resumen la mayor parte de la variabilidad en los datos.

## 3.3 Análisis de la dispersión:

La gráfica indica que la mayoría de los clientes están agrupados cerca del centro, aunque hay una dispersión considerable hacia la derecha a lo largo del primer componente principal (PC1). Esto sugiere que PC1 podría estar capturando una variable o un conjunto de variables que diferencian significativamente entre los clientes, como podría ser el ingreso, el gasto total o un comportamiento de compra específico. La dispersión a lo largo del segundo componente principal (PC2), aunque notable, es menos pronunciada en comparación con PC1, indicando que este componente podría estar capturando otra dimensión de variabilidad en los datos, posiblemente relacionada con otro tipo de comportamiento o característica demográfica de los clientes.

## 3.4 Interpretación de la varianza explicada:

Los valores de la varianza explicada por los dos componentes son 0.2833 (28.33%) para PC1 y 0.0880 (8.80%) para PC2. Esto significa que el primer componente principal solo explica alrededor del 28.33% de la variabilidad total en los datos, mientras que el segundo componente explica cerca del 8.80%. La suma de estos valores muestra que estos dos componentes juntos explican aproximadamente el 37.12% de la varianza total. Aunque esta cifra no es extremadamente alta, es típica en conjuntos de datos con muchas variables, donde cada componente adicional aporta incrementos más pequeños a la varianza explicada total.

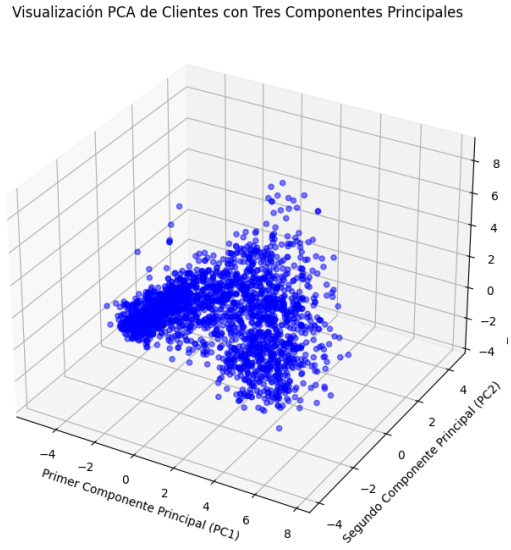


**Figure 13: Gráfico de PCA con 3 componentes**

Al aplicar PCA con tres componentes, aumenta un poco la explicación de la variabilidad en los datos. El primer componente principal explica el 28.33% de la variabilidad, destacando como el factor más influyente. Al agregar el segundo y tercer componente, que explican un 8.80% y 8.24% respectivamente, la varianza explicada acumulativa alcanza el 45.36%.

## 3.5 Clustering jerarquico

El clustering jerárquico es una técnica de análisis de clústeres utilizada para agrupar un conjunto de datos de manera que los objetos dentro del mismo grupo (llamado clúster) sean más similares entre sí que con los de otros grupos. Es especialmente útil en situaciones donde se necesita entender la relación jerárquica entre los objetos. Esta técnica no requiere que se especifique el número de clústeres de antemano, lo que la hace adecuada para aplicaciones donde la estructura de los grupos no es conocida a priori.

La visualización de esta técnica se realiza a través de un dendrograma, un tipo de diagrama que ilustra la disposición de los clústeres formados por el análisis jerárquico. En el dendrograma, cada hoja representa un objeto individual del conjunto de datos, y cada unión (o nodo) representa el punto donde dos clústeres se han fusionado. La altura de la unión indica la distancia o disimilitud entre los clústeres que se están fusionando. Cuanto mayor sea la altura, mayor será la disimilitud, lo que sugiere que una fusión a una altura mayor en el dendrograma señala diferencias más significativas entre los grupos.

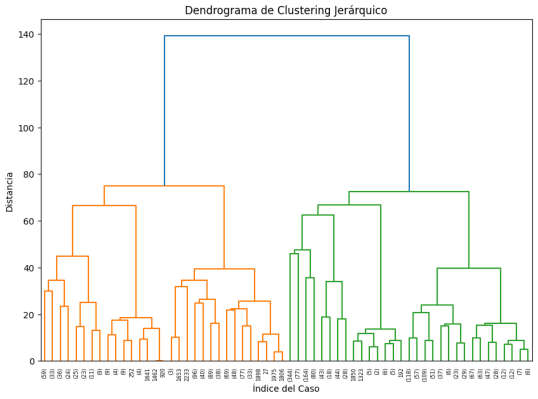


**Figure 14: Dendrograma de 5 niveles**

Al observar el dendrograma, notamos que al principio (parte inferior), las fusiones ocurren a distancias relativamente cortas, lo que indica que los casos son muy similares. A medida que avanzamos hacia la parte superior del dendrograma, las alturas de las fusiones aumentan, lo que refleja la combinación de clústeres cada vez menos similares. La línea azul vertical que cruza el dendrograma a una altura significativa es particularmente notable porque sugiere una distinción principal en el conjunto de datos, dividendo los casos en dos grupos grandes y fundamentalmente distintos.

Del lado izquierdo (coloreado en naranja), el grupo parece ser heterogéneo, con muchas fusiones a diferentes alturas que podrían indicar la presencia de subgrupos variados dentro de un grupo más amplio. Por otro lado, el grupo del lado derecho (coloreado en verde) muestra un patrón de fusión más uniforme, lo que podría sugerir una mayor cohesión interna entre los casos que forman estos clústeres.

La interpretación de este dendrograma es crucial para tomar decisiones informadas sobre el número de clústeres que podrían ser apropiados para análisis más detallados o aplicaciones prácticas. Un corte en la altura de la línea azul, por ejemplo, justificaría la división del conjunto de datos en dos grandes clústeres.

## 3.6 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering que se basa en la densidad de los datos para formar clústeres. A diferencia de otros métodos de clustering que requieren que el número de clústeres se especifique de antemano, DBSCAN puede identificar un número variable de clústeres basándose en las características espaciales de los datos. Esto lo hace particularmente útil en aplicaciones donde los clústeres pueden ser de diferentes tamaños y formas, y donde puede haber ruido o puntos atípicos en los datos.

El algoritmo DBSCAN funciona identificando "puntos centrales" que tienen al menos un número mínimo de otros puntos (minPts) dentro de un radio especificado (eps). Estos puntos centrales se consideran lo suficientemente densos como para formar el núcleo de un clúster. A partir de estos puntos centrales, el algoritmo luego intenta expandir el clúster agregando todos los puntos conectados que están dentro del radio eps. Este proceso continúa hasta que no se pueden agregar más puntos al clúster, y luego el algoritmo busca nuevos puntos centrales para formar clústeres adicionales. Los puntos que no forman parte de ningún clúster se consideran ruido o atípicos.La elección de los parámetros eps y minPts es crítica y puede influir significativamente en la calidad de los clústeres formados, lo que requiere una cuidadosa calibración basada en el conocimiento del dominio y la naturaleza de los datos.

Antes de seleccionar los hiperparámetros para el algoritmo DBSCAN, es esencial realizar un análisis preliminar para determinar un valor adecuado para el parámetro eps, que define la distancia máxima entre dos puntos para que sean considerados vecinos. Una técnica efectiva para estimar un buen valor de eps es analizar la distribución de las distancias al k-ésimo vecino más cercano de cada punto en el conjunto de datos.

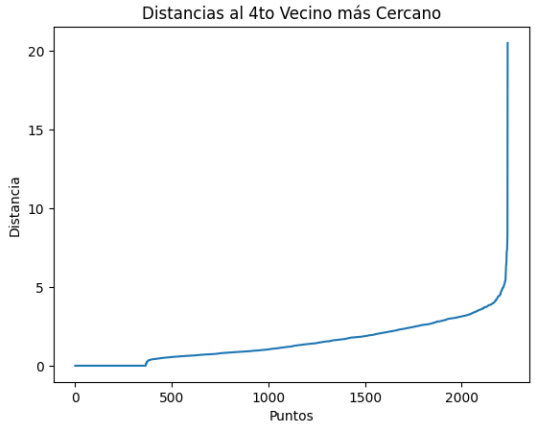


**Figure 15: Distancias al 4to NN**

En este caso, se utilizó NearestNeighbors de scikit-learn para encontrar los cinco vecinos más cercanos de cada punto en el conjunto de datos escalado. El parámetro neighbors se configuró en 5, que es igual a minsamples-1. Después de ajustar el modelo a los

datos, se obtienen las distancias a estos cinco vecinos más cercanos para cada punto.

Observando la gráfica, las distancias empiezan a aumentar de manera más significativa hacia el final. El "punto de codo", donde la curva comienza a aumentar drásticamente, es donde generalmente se recomienda establecer el valor de eps. Este punto representa un cambio en la densidad de los puntos, indicando un buen umbral para considerar si un punto debe ser considerado parte de un clúster o no.

En el gráfico, este cambio ocurre cerca del final de la gráfica, alrededor del punto 2300 en el eje x. La distancia correspondiente a este punto de codo parece estar alrededor de 5. Por lo tanto, un valor de eps alrededor de 5 podría ser un buen punto de partida para DBSCAN. Dado que el gráfico se basa en las distancias al cuarto vecino más cercano, un punto de partida para minsamples podría ser 5. Esto asegura que cada núcleo de un clúster debe tener al menos 5 puntos densamente agrupados. La aplicación de DBSCAN resultó

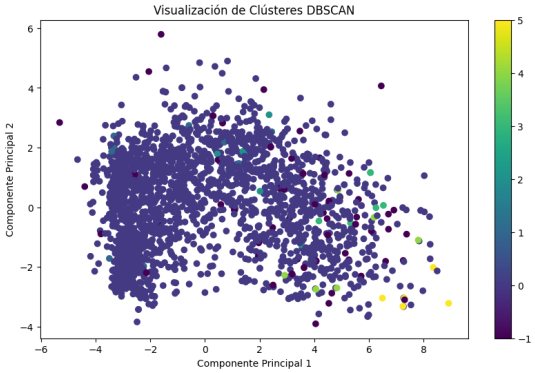


**Figure 16: Clústeres de DBSCAN**

en la formación de un clúster grande con muchos puntos dispersos clasificados como ruido, indicando que DBSCAN podría no ser el algoritmo más adecuado para esta tarea. La causa principal de este fenómeno puede ser la homogeneidad en la distribución de los puntos y la ausencia de variaciones significativas en la densidad a través del conjunto de datos. DBSCAN, que depende fuertemente de las diferencias de densidad para formar clústeres, tiende a agrupar en un único clúster a los puntos cuando estos están distribuidos de manera uniforme, mientras que los puntos en áreas de baja densidad, similares en extensión al radio eps utilizado, son frecuentemente identificados como ruido.

Dado este comportamiento, explorar otros métodos de clustering que no basen su criterio de agrupamiento únicamente en la densidad podría ser más beneficioso para este conjunto de datos.

## 3.7 K-Means

K-means es un algoritmo de clustering ampliamente utilizado debido a su simplicidad y eficacia para agrupar un conjunto de datos en un número específico de clústeres. El algoritmo opera asignando cada punto de datos al clúster cuyo centroide (la media de los puntos del clúster) está más cercano, luego recalculando los centroides de los clústeres y repitiendo este proceso hasta que las asignaciones de los puntos ya no cambien significativamente. El objetivo de K-means es minimizar la suma de las distancias cuadradas de cada punto al centroide de su clúster, lo que resulta en clústeres que son lo más compactos y distintos posibles.

Para optimizar los resultados de K-means, es crucial determinar el número adecuado de clústeres, conocido como $k$. Dos métodos comunes para ayudar a decidir el valor óptimo de $k$ son el método del codo y la puntuación de Silhouette.

El método del codo implica ejecutar el algoritmo K-means para una gama de valores de $k$ y calcular la suma de las distancias cuadradas dentro del clúster (inertia) para cada $k$. Al graficar estos valores, se busca un punto donde la tasa de disminución de la inertia se reduce significativamente, formando un "codo". Este punto de inflexión sugiere un número adecuado de clústeres, ya que indica un equilibrio entre la complejidad del modelo y la calidad del clustering.

Por otro lado, la puntuación de Silohuette evalúa qué tan similar es cada punto de un clúster a los puntos de su propio clúster en comparación con los puntos de otros clústeres. La puntuación de Silohuette varía de -1 a 1, donde valores cercanos a 1 indican que los puntos están bien agrupados, mientras que valores cercanos a -1 sugieren que los puntos podrían estar asignados al clúster incorrecto. Al calcular la puntuación de Silhouette para diferentes valores de $k$, se puede identificar el número de clústeres que maximiza esta puntuación, indicando una estructura de clústeres bien definida y claramente separada.

Estos métodos combinados determinan el número óptimo de clústeres en K-means, garantizando que los resultados sean tanto efectivos como interpretables:

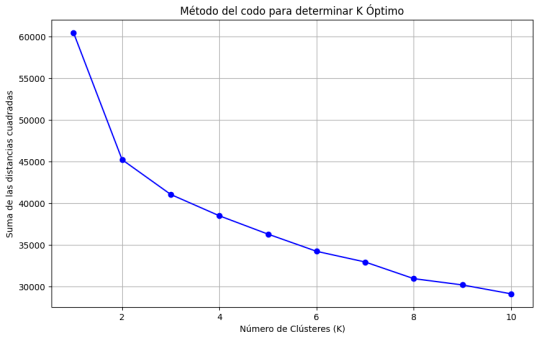


**Figure 17: Método del codo 2-10 K**

La imagen muestra un gráfico que utiliza el método del codo para determinar el número óptimo de clústeres (K).

El gráfico traza la curva de la suma de las distancias cuadradas para diferentes valores de K, desde 1 hasta 10. Al observar la curva, se puede identificar un punto donde la tasa de disminución de la inertia empieza a ser menos pronunciada. En este caso, el codo parece situarse alrededor de K=2. Después del codo, la reducción en la suma de las distancias cuadradas se desacelera, lo que sugiere que agregar más clústeres proporciona no nos da ningún beneficio.

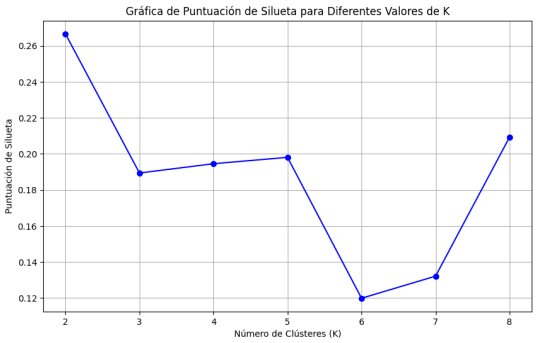


**Figure 18: Gráfica de puntuación Silhouette**

*3.7.1 Silhouette Score.* La imagen muestra el gráfico de puntuación de Silhouette para diferentes valores de K. La puntuación de Silhouette varía entre -1 y 1, donde valores cercanos a 1 indican que los puntos están bien agrupados dentro de sus clústeres y bien separados de otros clústeres. Valores cercanos a -1 sugieren que los puntos pueden estar asignados incorrectamente a los clústeres. El gráfico revela cómo varía esta puntuación al cambiar el número de clústeres de 2 a 8.

Observando el gráfico, se puede ver que la puntuación de Silhouette es más alta para K=2, lo que indica que con dos clústeres los datos están mejor agrupados y separados. Sin embargo, hay una caída significativa en la puntuación al aumentar K a 3 y 4, con una ligera recuperación para K=4 y 5. La puntuación de Silhouette disminuye notablemente al aumentar K a 6, y luego muestra una tendencia ascendente nuevamente hacia K=8.

En conclusión, el análisis de la puntuación de Silhouette sugiere que el mejor número de clústeres para este conjunto de datos podría ser 2, dado que proporciona la mayor puntuación.

*3.7.2 K-Means con k = 2.* Con la información obtenida a partir del método del codo y la puntuación de Silhouette, podemos tomar una decisión informada sobre el número óptimo de clústeres para aplicar el algoritmo K-means. El método del codo indica claramente que el número óptimo de clústeres es $k = 2$, ya que en este punto se observa un codo significativo en la curva, lo que sugiere una buena relación entre la complejidad del modelo y la calidad del clustering.

Además, la puntuación de Silhouette, que evalúa la cohesión interna y la separación entre clústeres, también respalda esta elección, mostrando que $k = 2$ proporciona una alta calidad de agrupamiento. Por lo tanto, con ambos métodos coincidiendo en que $k = 2$ es el valor óptimo, podemos proceder a aplicar el algoritmo K-means con $k = 2$ para segmentar el conjunto de datos de manera óptima y obtener insights valiosos sobre los diferentes grupos de clientes presentes en los datos.

La gráfica muestra los resultados del clustering K-means aplicado con $k = 2$ clústeres. Cada punto en el gráfico representa un cliente, visualizado en un espacio bidimensional utilizando los dos primeros componentes principales (PC1 y PC2) obtenidos a través del Análisis de Componentes Principales (PCA). Los puntos están coloreados y marcados según el clúster al que pertenecen: el clúster 0 está representado por círculos azules y el clúster 1 por cruces verdes.

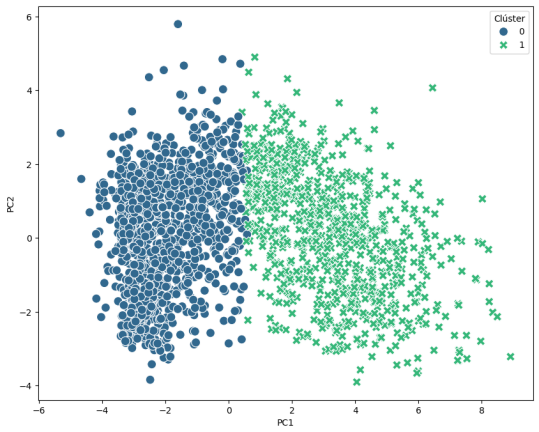


**Figure 19: Clústeres formados por KMeans**

La separación clara entre los dos clústeres sugiere que K-means ha logrado identificar dos grupos distintos dentro del conjunto de datos. El clúster 0 (azul) parece concentrarse en el lado izquierdo del gráfico, mientras que el clúster 1 (verde) se distribuye principalmente en el lado derecho. Esta separación indica diferencias significativas en las características de los clientes que componen cada clúster.

El resultado de K-means con $k = 2$ refleja una segmentación efectiva, donde los clientes dentro de cada clúster comparten similitudes que los distinguen de los clientes en el otro clúster. Esta información puede ser utilizada para diseñar estrategias de marketing personalizadas y mejorar la toma de decisiones basada en el comportamiento y las características específicas de cada grupo de clientes. La clara distinción entre los dos clústeres también valida la elección de $k = 2$ como el número óptimo de clústeres, respaldado por el método del codo y la puntuación de Silhouette.

*3.7.3 Análisis de las características de los clústeres.* Para comprender mejor los dos clústeres identificados mediante el algoritmo K-means, se realizó un análisis de las medias de las variables de cada grupo de clientes. Este análisis incluye la evaluación de variables significativas como ingresos, gastos en diferentes categorías de productos y otros comportamientos de compra. Las variables seleccionadas para este análisis incluyen el ingreso anual, el gasto en productos específicos (como vinos y carnes), y las distintas formas de interacción con la compañía (como compras en tienda, compras por catálogo y visitas al sitio web)

- **Ingreso anual (Income):** Representa los ingresos medios anuales de los clientes.
- **Gasto en productos específicos:**
  - **Vinos (MntWines):** Gasto en vinos durante los últimos dos años.
  - **Productos de carne (MntMeatProducts):** Gasto en productos de carne durante los últimos dos años.
- **Comportamiento de compra:**
  - **Visitas Web mensuales (NumWebVisitsMonth):** Número de visitas al sitio web de la compañía en el último mes.
  - **Compras con descuento (NumDealsPurchases):** Número de compras realizadas con descuentos.

- **Compras Web (NumWebPurchases):** Número de compras realizadas a través del sitio web.
- **Compras por catálogo (NumCatalogPurchases):** Número de compras realizadas mediante catálogo.
- **Compras en tienda (NumStorePurchases):** Número de compras realizadas directamente en tiendas físicas.

Posteriormente, se calcularon las medias de estas variables para cada clúster, permitiendo así una comparación clara de las características promedio entre los dos grupos. Este enfoque proporciona una visión detallada de los patrones de consumo y las preferencias de los clientes en cada clúster, lo cual es fundamental para diseñar estrategias de marketing personalizadas.

**Resumen de clústeres**

**Clúster 0:**

- **Ingreso Medio:** Aproximadamente $72,075
- **Gasto Medio en Vinos:** $610
- **Gasto Medio en Productos de Carne:** $362
- **Visitas Web Mensuales:** 3.7
- **Compras con Descuento:** 2.0
- **Compras Web:** 5.8
- **Compras por Catálogo:** 5.3
- **Compras en Tienda:** 8.6

**Clúster 1:**

- **Ingreso Medio:** Aproximadamente $39,063
- **Gasto Medio en Vinos:** $101
- **Gasto Medio en Productos de Carne:** $37
- **Visitas Web Mensuales:** 6.4
- **Compras con Descuento:** 2.5
- **Compras Web:** 3.0
- **Compras por Catálogo:** 0.9
- **Compras en Tienda:** 4.0

*3.7.4 Descripción de clústeres.* **Clúster 0** parece representar a clientes de mayor ingreso que gastan significativamente más en productos de lujo como vinos y carnes. Estos clientes también muestran una mayor frecuencia de compras en la tienda y por catálogo, lo que indica una preferencia por las compras tradicionales. La media de gasto en vinos y productos de carne es notablemente alta en comparación con el otro clúster, sugiriendo que estos clientes valoran productos de mayor calidad y están dispuestos a pagar un precio premium por ellos. Además, el menor número de visitas web mensuales podría indicar que estos clientes prefieren la experiencia de compra física, donde pueden interactuar con los productos directamente, o el uso de catálogos, que ofrecen una manera más cómoda y menos tecnológica de hacer sus compras. Este comportamiento puede estar relacionado con una mayor fidelidad a los métodos tradicionales de compra y una menor adopción de plataformas digitales para sus transacciones.

**Clúster 1**, por otro lado, incluye clientes con ingresos más bajos y un menor gasto en productos. Este grupo muestra una tendencia a realizar más compras con descuento, lo que sugiere que son más sensibles al precio y buscan activamente ofertas y promociones. La mayor frecuencia de visitas al sitio web sugiere una familiaridad y comodidad con las compras en línea, lo que podría ser una señal de que estos clientes son más receptivos a las campañas de marketing digital y promociones en línea. El menor gasto en productos como vinos y carnes indica que estos clientes pueden estar más enfocados en compras básicas y necesarias, en lugar de productos de lujo. La combinación de ingresos más bajos y una alta sensibilidad a las ofertas y promociones en línea sugiere que las estrategias de marketing para este clúster deben centrarse en descuentos, promociones atractivas y campañas digitales diseñadas para atraer a estos consumidores a través de plataformas en línea.

Estos insights pueden ser muy útiles para diseñar estrategias de marketing dirigidas y personalizadas. Por ejemplo, podríamos enfocarnos en ofrecer productos de lujo y promociones exclusivas en catálogo al Clúster 0, mientras que para el Clúster 1 podríamos enfocar las campañas digitales y ofertas en línea.

## 3.8 Reglas de Asociación

Las reglas de asociación son una técnica de minería de datos utilizada para identificar relaciones interesantes entre variables en grandes bases de datos. Estas reglas son especialmente útiles para descubrir patrones y tendencias en datos transaccionales, como los datos de ventas al por menor. Una regla de asociación se representa típicamente en la forma $A \rightarrow B$, donde $A$ es el conjunto de antecedentes y $B$ es el conjunto de consecuentes. El significado de esta regla es que si ocurre $A$, es probable que también ocurra $B$.

Dos métricas importantes utilizadas para evaluar la relevancia y la fuerza de las reglas de asociación son:

- **Soporte (support):** La proporción de transacciones en la base de datos en las que aparecen tanto $A$ como $B$. Esta métrica ayuda a identificar cuán frecuente es la regla en el conjunto de datos.
- **Confianza (confidence):** La proporción de transacciones que contienen $A$ y también contienen $B$. Esta métrica mide la fiabilidad de la inferencia hecha por la regla.

**Descripción de las Reglas**

(1) **(NumCatalogPurchases) => (NumWebPurchases)**
- **Soporte:** 0.733036
- **Confianza:** 0.992745
- **Descripción:** El 73.30% de las transacciones incluyen tanto compras por catálogo como compras web. Casi todos los clientes que compran por catálogo también compran por web.

(2) **(NumWebPurchases) => (NumCatalogPurchases)**
- **Soporte:** 0.733036
- **Confianza:** 0.749429
- **Descripción:** El 73.30% de las transacciones incluyen tanto compras web como por catálogo. El 74.94% de los clientes que compran por web también compran por catálogo.

(3) **(NumStorePurchases) => (NumWebPurchases)**
- **Soporte:** 0.975893
- **Confianza:** 0.982472
- **Descripción:** El 97.59% de las transacciones incluyen tanto compras en tienda como por web. El 98.25% de los clientes que compran en tienda también compran por web.

(4) **(NumWebPurchases) => (NumStorePurchases)**
- **Soporte:** 0.975893

- **Confianza:** 0.997718
- **Descripción:** El 97.59% de las transacciones incluyen tanto compras web como en tienda. El 99.77% de los clientes que compran por web también compran en tienda.

(5) **(NumDealsPurchases) => (NumWebPurchases)**
- **Soporte:** 0.961161
- **Confianza:** 0.981313
- **Descripción:** El 96.12% de las transacciones incluyen tanto compras con descuento como por web. El 98.13% de los clientes que compran con descuento también compran por web.

## 4 CONCLUSIÓN

En este trabajo, se exploró el análisis de datos de clientes utilizando diversas técnicas de aprendizaje no supervisado, incluyendo el análisis de componentes principales (PCA), clustering jerárquico, DBSCAN y K-means. Cada una de estas técnicas ha proporcionado perspectivas sobre la estructura y las características del conjunto de datos, permitiendo una comprensión más detallada del comportamiento del consumidor.

El PCA fue una herramienta utilizada para reducir la dimensionalidad de los datos, facilitando la visualización y la identificación de patrones subyacentes. La utilización de PCA nos permitió condensar la información en componentes principales que capturan la mayor parte de la variabilidad en los datos. Esto no solo simplificó el análisis posterior, sino que también destacó las principales diferencias y similitudes entre los clientes, proporcionando una base sólida para el clustering.

El clustering jerárquico y DBSCAN ofrecieron enfoques diferentes para la agrupación de datos. Mientras que el clustering jerárquico reveló la estructura jerárquica y las relaciones entre los clientes, DBSCAN demostró ser menos adecuado debido a la distribución homogénea de los datos, lo que resultó en un gran clúster dominante y varios puntos clasificados como ruido. Esto resalta bastante dada la importancia de seleccionar la técnica de clustering apropiada según la naturaleza de los datos.

El algoritmo K-means, optimizado mediante el método del codo y la puntuación de Silhouette, resultó ser el más eficaz a la hora de segmentar el conjunto de datos en dos clústeres principales. Los análisis de las características de estos clústeres revelaron diferencias significativas en los comportamientos de compra y preferencias de los clientes, proporcionando insights valiosos para estrategias de marketing personalizadas.

En conjunto, este trabajo demuestra cómo el uso de técnicas de aprendizaje no supervisado puede transformar datos complejos en insights prácticos y accionables, mejorando significativamente la toma de decisiones en el ámbito comercial.

## REFERENCES

[n. d.]. scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. https://scikit-learn.org/

Omar Romero-Hernandez. 2021. Customer Personality analysis. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data