Segmentación de Clientes en un Ecommerce para aplicar en campañas de Marketing

Nombre de los autores: Zhuyin Cruz Sandoya y Melanie Tacuri Galarza

Nombre de la Facultad: ESAI Business School

Nombre de la Maestría: Inteligencia de Negocios y Ciencia de Datos

Fecha de culminación: Noviembre de 2024

Palabras clave: Segmentación, Marketing, Ecommerce, Componentes Principales, Clusters.

A. Resumen Ejecutivo

El crecimiento del Ecommerce y la competencia en el sector han obligado a las empresas a buscar estrategias de marketing personalizadas para captar y retener clientes. Este proyecto aborda la necesidad de segmentar la base de clientes de un Ecommerce para identificar grupos con características y comportamientos de compra similares. Esto permitirá diseñar campañas de marketing personalizadas, aumentando la retención y la satisfacción del cliente.

Para identificar los segmentos clave, se utilizaron técnicas de análisis de datos como Análisis de Componentes Principales (PCA) y Cluster con K-medias. El PCA permitió reducir la dimensionalidad de los datos, conservando el 50% de la variabilidad en los dos primeros componentes principales, lo que facilitó la agrupación y visualización de clientes en un espacio reducido. Aplicando K-medias, se definieron cuatro segmentos principales: Cazadores de Ofertas, Compradores Económicos, Compradores Frecuentes y Compradores Premium. Cada grupo se distinguió por patrones específicos en hábitos de compra, preferencias de precios y categorías de

productos adquiridos, que fueron evaluados con métricas como el método del codo. Los resultados mostraron una distribución equilibrada de los segmentos a través de categorías y países, con ligeras variaciones que pueden ser utilizadas para estrategias personalizadas, como campañas dirigidas a Compradores Premium en mercados de alto valor (EE.UU. y Reino Unido) o promociones para Cazadores de Ofertas en segmentos más sensibles al precio. En resumen, el proyecto proporciona una base sólida para optimizar campañas de marketing y maximizar el impacto de las estrategias comerciales.

B. Resumen Ejecutivo Gráfico

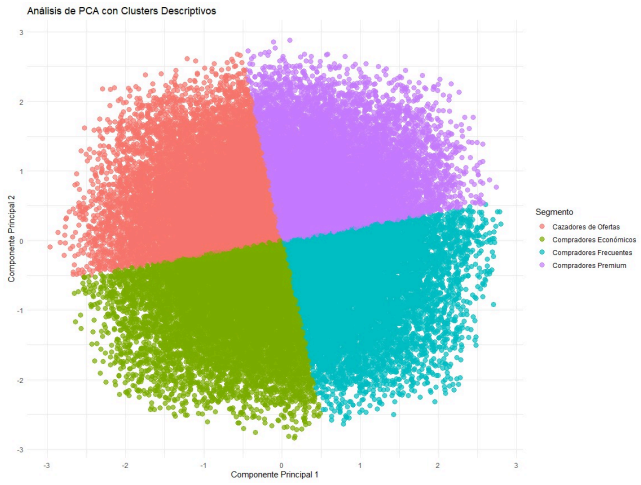

Gráfico 1: *"Segmentación de Clientes basada en PCA y Clusters"*

El **gráfico 1** muestra la segmentación de clientes en cuatro grupos principales, basados en su comportamiento de compra. La técnica de PCA permitió simplificar los datos al concentrar el 50% de la variabilidad en estas dos dimensiones, facilitando la separación de los segmentos. Los clusters identificados corresponden a perfiles específicos: Cazadores de Ofertas, Compradores Económicos, Compradores Frecuentes, y Compradores Premium, brindando información clave para personalizar estrategias de marketing y entender patrones de comportamiento.

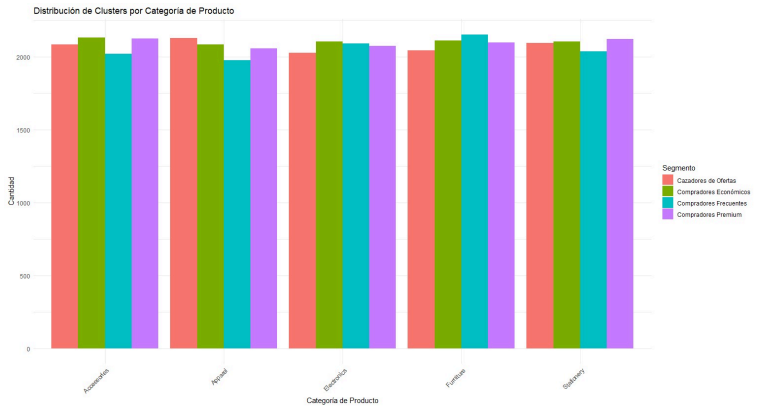

Gráfico 2: *"Distribución de clusters por categoría de producto"*

En el **gráfico 2** muestra la distribución de los clusters por categoría de producto obtenidos, aplicada tras reducir las dimensiones con PCA. Se observa que los cuatro segmentos tienen una participación similar en todas las categorías de producto. Esto indica que los segmentos son consistentes en su comportamiento de compra independientemente de la categoría. La técnica de PCA permitió simplificar el análisis al identificar los patrones clave en los datos, mientras que K-means clasificó a los clientes en grupos distintos.

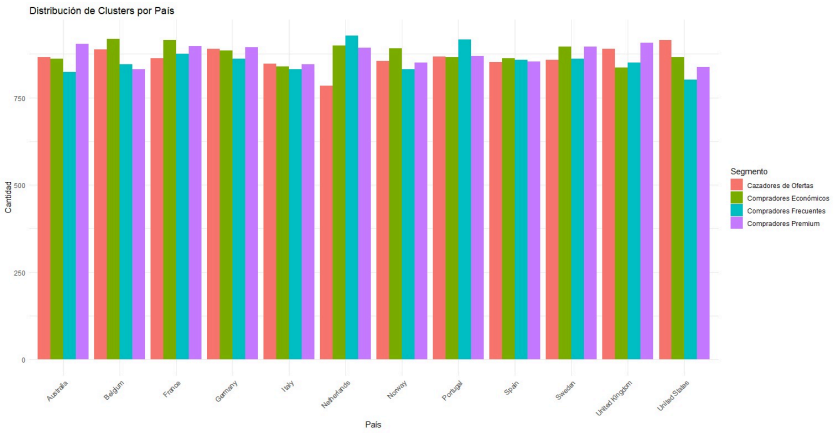

Gráfico 3: *"Distribución de clusters por país"*

El **gráfico 3** representa la distribución de los clusters por país. El resultado muestra una participación bastante equilibrada entre los segmentos en la mayoría de los países, con ligeras variaciones. Con la técnica de PCA, se simplificaron los datos al reducir las dimensiones a patrones

clave, y con K-means clustering, se agruparon los clientes en segmentos claros, permitiendo identificar diferencias regionales.

C. Detalle Técnico

1. Fuente de Datos

La base de datos utilizada proviene de Kaggle, incluye datos anónimos sobre transacciones de ventas en línea, que capturan diversos aspectos de las compras de productos, detalles del cliente y características de los pedidos. Esta base de datos recopila información desde el 2020 hasta el 2024 y el análisis fue realizado con la herramienta R Studio. Las variables a utilizar para el proyecto son:

- Cantidad: Representa la cantidad de productos comprados
- Precio Unitario: Indica el precio unitario de los productos
- Descuento: Refleja los descuentos aplicados.
- Costo de envío: Representa el costo de envío.
- Categoría: La categoría de producto.
- País: País dónde se compran los productos.

Continuando con el análisis exploratorio, este permite identificar hallazgos fundamentales para establecer criterios de segmentación basados en intereses y comportamientos de compra. La información obtenida no solo facilita la priorización de categorías clave en futuras campañas, sino que también proporciona un punto de partida para analizar cómo estos patrones de consumo se relacionan con los segmentos previamente identificados.

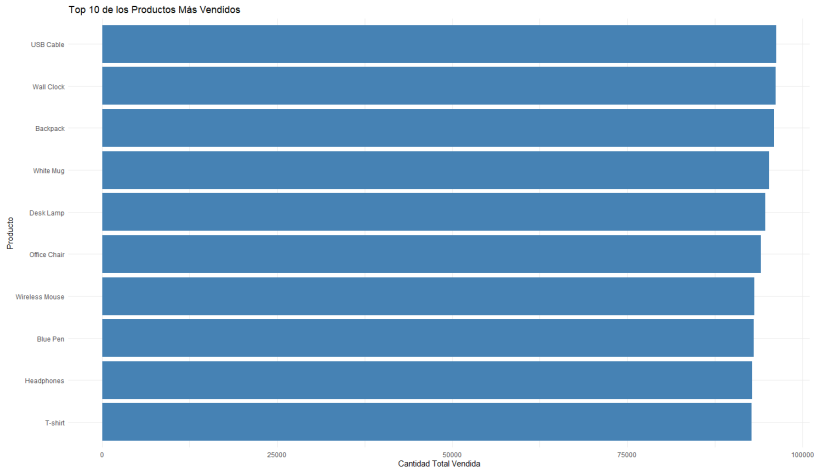

Gráfico 4: *"Top 10 de los productos más vendidos"*

En el gráfico 4 de los 10 productos más vendidos, permite identificar patrones de compra y preferencias de los clientes, lo que es clave para segmentarlos de manera efectiva. Los productos más populares, como "USB Cable" y "Wall Clock", pueden ser utilizados como criterio para agrupar clientes interesados en accesorios tecnológicos o artículos del hogar.

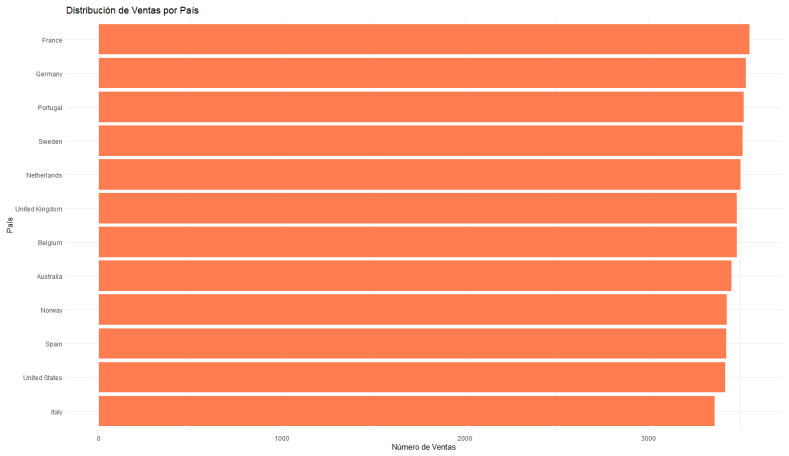

Gráfico 5: *"Distribución de Ventas por País"*

El **gráfico 5** muestra la distribución del número de ventas por país, destacando que algunos países, como Francia y Alemania, tienen un mayor número de ventas, mientras que países como Italia y Estados Unidos tienen menos. Sin embargo, la diferencia entre los países no es extrema, lo que sugiere que las ventas están relativamente bien distribuidas a nivel geográfico. Con esta información el PCA puede identificar patrones globales en los datos, mientras que los clusters permiten diferenciar segmentos específicos según comportamientos regionales.

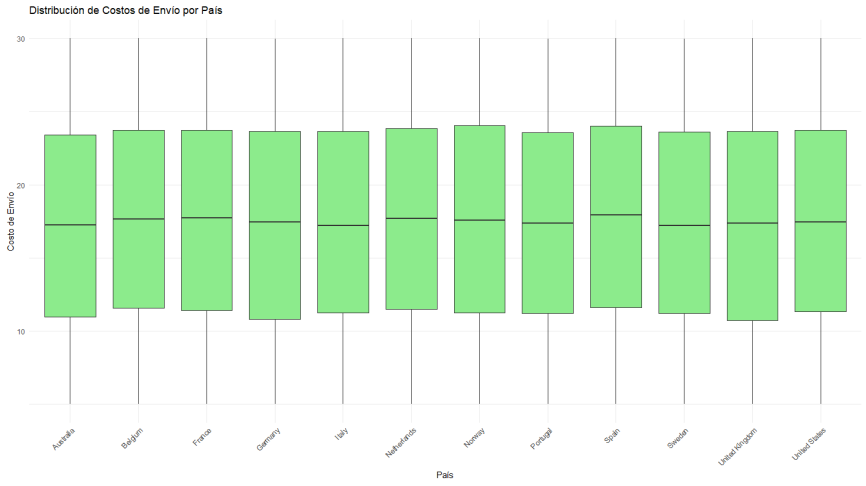

Gráfico 6: *"Distribución de Costos de envío por País"*

El **gráfico 6** muestra la distribución de los costos de envío por país, donde se observa que la mediana de los costos es similar entre los países, indicando una política de precios uniforme o condiciones logísticas estandarizadas que mantienen los costos equilibrados, independientemente del país.

2. Técnica

Para lograr una segmentación precisa, se emplearon dos técnicas, primero el análisis de Componentes Principales (PCA), esta técnica permitió reducir la dimensionalidad del conjunto de datos, destacando las variables que explican la mayor parte de la variabilidad en el comportamiento de compra de los clientes. Esto facilitó una visión clara de los factores más influyentes. Este enfoque también se utiliza para eliminar redundancias y simplificar el análisis de grandes conjuntos de datos (Abdulhafedh, 2021). La segunda técnica es el análisis de Clusters, se utilizó para agrupar a los clientes en segmentos significativos basados en su comportamiento y preferencias. Esta técnica permite descubrir grupos de clientes con patrones de compra similares, lo cual es esencial para el diseño de campañas personalizadas de marketing (Nikum Yulisasih et al., 2024).

El rendimiento del análisis de componentes principales (PCA) se evalúa determinando la proporción de la variabilidad general dentro del conjunto de datos que está representada por los

componentes principales iniciales. Esta evaluación se lleva a cabo analizando la varianza explicada por cada componente, así como su efecto acumulativo. En este análisis, la varianza elucidada por el PCA se asigna de manera uniforme entre los cuatro primeros componentes principales (Alvarez et al., 2021). Cada componente representa aproximadamente el 25% de la variabilidad (Anexo, gráfico 7) total y abarca de forma acumulativa toda la información contenida en el conjunto de datos. Esto sugiere que no hay componente dominante, lo que indica que todas las dimensiones desempeñan un papel crucial a la hora de dilucidar los patrones significativos de los datos. Además, al emplear únicamente los dos primeros componentes (PC1 y PC2), se puede visualizar casi el 50% de la varianza (Anexo, gráfico 7), lo que facilita la simplificación del análisis y, al mismo tiempo, conserva una cantidad sustancial de información (Gewers et al., 2018).

Dentro del proyecto el PCA no debe haber sobre ajuste (Anexo, tabla 1), ya que el objetivo del PCA es simplificar los datos al reducir las dimensiones mientras se preserva la máxima información relevante. Seleccionar un número excesivo de componentes (por ejemplo, explicando el 100% de la varianza) puede incluir ruido y redundancia, perdiendo su propósito de simplificación (Shlens, 2014).

Los clusters se obtuvieron aplicando el algoritmo de K-medias sobre los primeros componentes principales generados por el PCA, lo que permitió trabajar en un espacio relevante, eliminando redundancias. Para determinar el número óptimo de clusters, se utilizó el Método del Codo, analizando la suma de cuadrados intra-cluster (WCSS) y seleccionando el punto donde la reducción de WCSS se vuelve marginal (el "codo"). Esto garantiza que los clusters sean compactos (puntos dentro de un cluster cercanos entre sí) y bien separados entre sí, optimizando la segmentación del dataset (Moraguez, 2024).

La eficiencia del modelo se mide evaluando tanto el rendimiento del PCA como la calidad de los clusters. Para el PCA, se verifica que las primeras componentes principales explican una alta proporción de la varianza total (al menos el 50% con Dim1 y Dim2) (Anexo, gráfico 9), asegurando

que los datos han sido correctamente resumidos. Para los clusters, se mide la compacidad (puntos dentro de un cluster cercanos entre sí) y la separación (distancia entre clusters), utilizando métricas como la suma de cuadrados intra-cluster, que confirma qué tan bien definidos están los grupos. (Amat, 2020)

3. Resultados

En el proyecto se clasificó a los clientes en cuatro segmentos claves: Cazadores de ofertas, compradores económicos, compradores frecuentes y compradores premium, según sus patrones de compra, logrados mediante la aplicación del análisis de componentes principales y las técnicas de agrupamiento de K-medias. El conjunto de datos se simplificó utilizando la PCA, preservando así la mayor parte de la información importante. La implementación de la agrupación en clústeres de K-medias facilitó la clasificación de los clientes en segmentos explícitos basados en patrones de comportamiento compartidos. Según los resultados los segmentos muestran una distribución coherente entre diversas categorías de productos y regiones geográficas, con pequeñas discrepancias que pueden aprovecharse para diseñar las estrategias de marketing de forma eficaz. Por ejemplo, los compradores premium se encuentran predominantemente en países como los Estados Unidos y el Reino Unido, mientras que los compradores económicos son más frecuentes en Alemania. En conjunto, este análisis proporciona un marco sólido para la formulación de campañas de marketing específicas, la optimización de las ofertas promocionales y una mejor comprensión de las preferencias de los consumidores.

Anexo

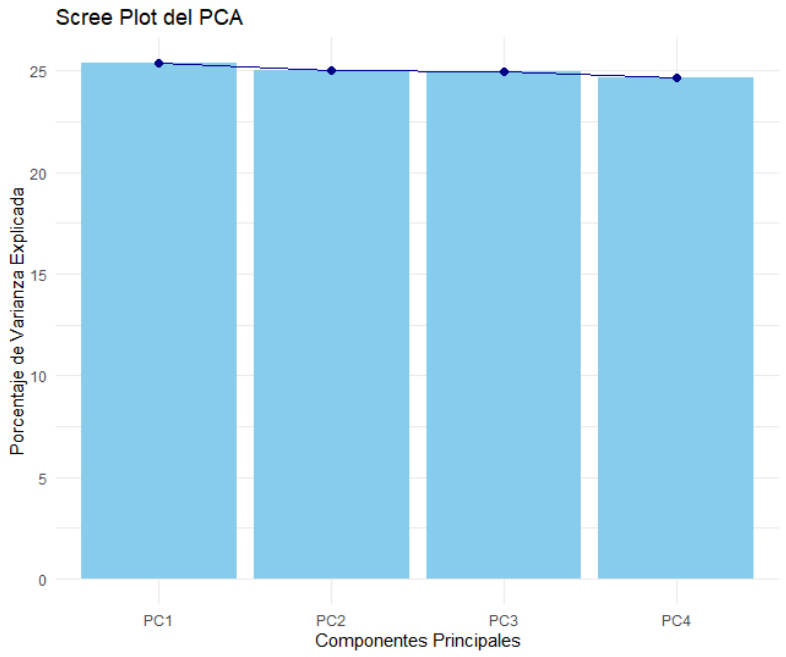

Gráfico 7: *"Gráfico de varianza explicada acumulativa"*

	Componente	Varianza_Entrenamiento	Varianza_Prueba
PC1	PC1	0.25373	0.2528271
PC2	PC2	0.25023	0.2470213
PC3	PC3	0.25008	0.2511126
PC4	PC4	0.24596	0.2490391

Tabla 1: *"Tabla de entrenamiento y prueba PCA"*

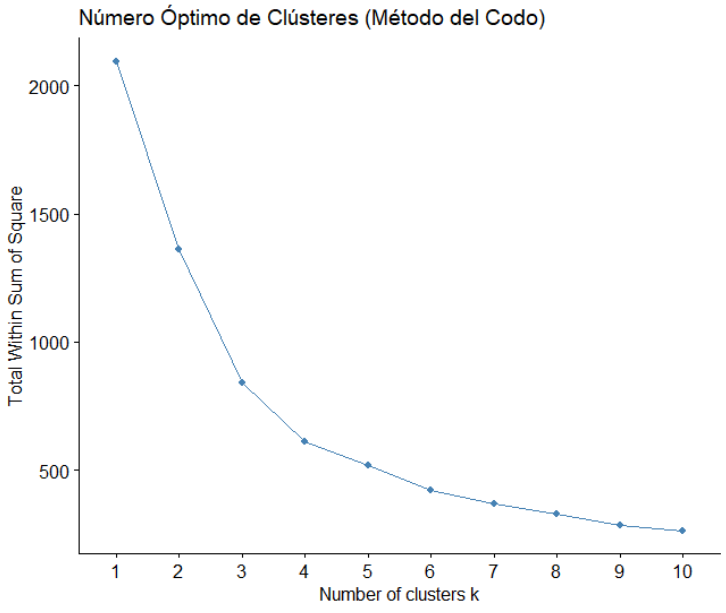

Gráfico 8: *"Número óptimo de Clústeres (Método del codo)"*

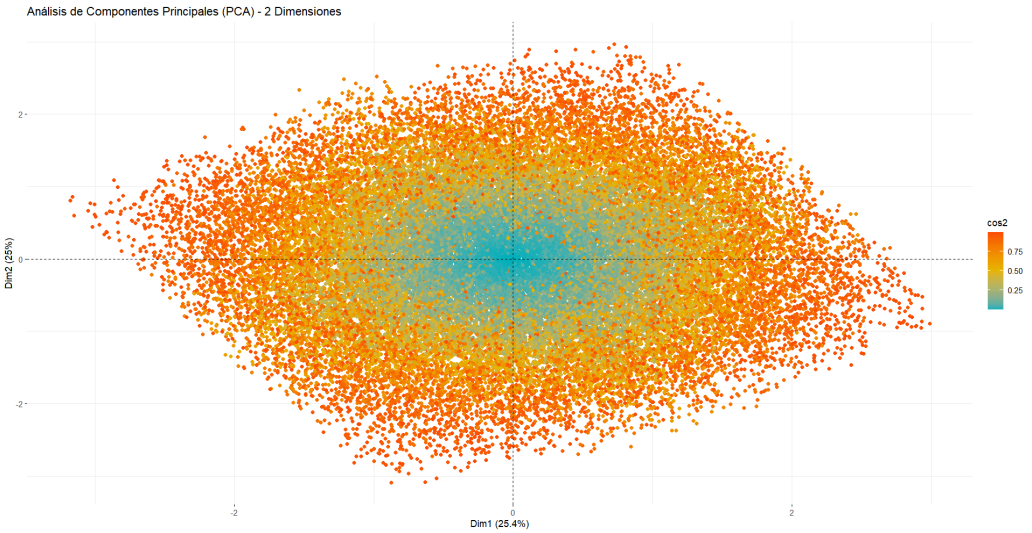

Gráfico 9: *"Análisis de Componentes Principales - 2 Dimensiones"*

E. Referencias Bibliográficas

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. https://www.researchgate.net/publication/349094412_Incorporating_K-means_Hierarchical_Clustering_and_PCA_in_Customer_Segmentation

Alvarez, C., Collazos, B., & Macuace, V. (2021). ANÁLISIS DE COMPONENTES PRINCIPALES (PCA). *Amazonaws.com*. https://rstudio-pubs-static.s3.amazonaws.com/813586_24a2a66d91ca430f99ac3d9d2449f452.html

Amat, J. (2020, Diciembre). PCA con Python. *Cienciadedatos.net*. https://cienciadedatos.net/documentos/py19-pca-python.html

Delikkaya, Y. (2024). *Online Sales Dataset*. https://www.kaggle.com/datasets/yusufdelikkaya/online-sales-dataset/data

Gewers, F. L., Ferreira, G. R., de Arruda, H. F., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. d. F. (2018). *Principal Component Analysis: A Natural Approach to Data Exploration*. https://arxiv.org/pdf/1804.02502

Moraguez, E. R. (2024, Octubre). K-Means Clustering: Guía Práctica para Segmentación de Datos. *LovTechnology*. https://lovtechnology.com/k-means-clustering-guia-practica-para-segmentacion-de-datos/

Munar, P. (2024, Octubre 8). Clustering en Marketing Digital: qué es y cómo segmentar a tu audiencia eficazmente. *Cyberclick*. https://www.cyberclick.es/numerical-blog/clustering-en-marketing-digital-que-es-y-como-segmentar-a-tu-audiencia-eficazmente#:~:text=El%20clustering%20es%20una%20t%C3%A9cnica,diferentes%20personas%20de%20forma%20autom%C3%A1tica

Nikum Yulisasih, B., Hernan, & Sunardi. (2024). View of K-Means Clustering Method For Customer Segmentation Based On Potential Purchases. *Poliban.ac.id*. https://eltikom.poliban.ac.id/index.php/eltikom/article/view/1137/134

Shlens, J. (2014). A Tutorial on Principal Component Analysis. https://arxiv.org/pdf/1404.1100