Segmentación de Clientes en Centros Comerciales utilizando Aprendizaje No Supervisado

Integrantes: Alejandra Maria Jerez Pardo, Camilo Alejandro Grande Sanchez, Johan Sebastián Morales Caro y Mateo Grisales Hurtado

Resumen

En un entorno competitivo, comprender el comportamiento y las preferencias de los clientes es fundamental para diseñar estrategias de marketing efectivas. Este proyecto aborda el desafío de segmentar a los clientes de un centro comercial en grupos homogéneos basados en sus características demográficas y patrones de gasto, utilizando técnicas de aprendizaje no supervisado, específicamente Modelos de Mezcla Gaussiana (GMM). El objetivo es identificar perfiles de clientes para personalizar las estrategias de marketing, optimizar promociones y mejorar la experiencia del cliente. La propuesta busca superar las limitaciones de segmentaciones genéricas, proporcionando un enfoque más eficiente y dirigido que facilita la toma de decisiones comerciales. Al integrar GMM con técnicas de reducción de dimensionalidad, este trabajo ofrecerá una visión detallada y accionable para diseñar campañas de marketing que respondan a las necesidades específicas de cada segmento, mejorando así la efectividad y satisfacción en el centro comercial. La contribución principal es una herramienta avanzada y adaptada al contexto único de los centros comerciales, que permitirá maximizar el impacto de las campañas y fortalecer la relación con los clientes.

Introducción

En el competitivo entorno del sector minorista, la personalización de estrategias de marketing se ha vuelto esencial para captar y retener clientes. La segmentación de clientes, una técnica clave dentro del marketing analítico, permite a las empresas entender mejor el comportamiento de compra de sus consumidores y ajustar sus ofertas a las necesidades específicas de cada grupo (John et al., 2023). Sin embargo, muchos centros comerciales aún utilizan estrategias genéricas que no aprovechan plenamente el potencial de los datos disponibles, resultando en campañas de marketing menos efectivas y una menor satisfacción del cliente.

El problema central de este proyecto es la falta de segmentación efectiva de los clientes en un centro comercial, lo que impide a los administradores y minoristas personalizar sus campañas de marketing de manera óptima. Los administradores de centros comerciales y las empresas minoristas, como clientes potenciales de este proyecto, enfrentan la necesidad de maximizar el impacto de sus campañas y mejorar la experiencia del cliente mediante la personalización de sus ofertas. Estudios recientes han demostrado que el uso de algoritmos de clustering, como K-means y modelos de mezcla gaussiana (GMM), facilita la agrupación de clientes en segmentos con comportamientos de compra similares, proporcionando información valiosa para la toma de decisiones estratégicas (John et al., 2023).

Este proyecto se enmarca en el área del aprendizaje no supervisado, donde el objetivo es descubrir patrones y estructuras ocultas en los datos sin necesidad de etiquetas predefinidas. Las técnicas de clustering, como las mencionadas anteriormente, no solo ayudan a identificar perfiles específicos de clientes, sino que también permiten mejorar la eficacia de las campañas de marketing y el servicio al cliente al centrarse en sus preferencias y comportamientos (Gomes, M. A., & Meisen, T., 2023). Con la aplicación adecuada de estas técnicas, los centros comerciales pueden transformar datos sin procesar

en estrategias de marketing personalizadas y efectivas, logrando así un mayor rendimiento y satisfacción del cliente.

Revisión preliminar de la literatura

La segmentación de clientes ha sido un área de estudio prolífica, especialmente en sectores como el retail, el comercio electrónico y los servicios financieros. Los métodos de clustering han sido comúnmente empleados para identificar patrones en el comportamiento de los consumidores y optimizar estrategias de marketing. A continuación, se presentan algunos estudios destacados en la literatura que abordan problemas y métodos similares al enfoque propuesto en este proyecto.

Kumar y Shah (2018) utilizaron el algoritmo K-means para segmentar clientes en un supermercado, logrando identificar patrones de gasto que ayudaron a la creación de programas de lealtad personalizados. Aunque este enfoque fue efectivo para captar patrones básicos de comportamiento, K-means asume que los clusters son esféricos y de tamaño similar, lo cual no siempre refleja la complejidad real de los datos de los consumidores.

Otro estudio relevante es el de García y López (2020), quienes aplicaron clustering jerárquico para segmentar clientes de una plataforma de comercio electrónico. Este método facilitó la personalización de recomendaciones de productos, mejorando significativamente la satisfacción del cliente. Sin embargo, el clustering jerárquico mostró dificultades para manejar grandes volúmenes de datos y definir el número óptimo de clusters, limitando su escalabilidad en contextos más complejos.

En el ámbito internacional, Jhon. et al. (2023) compararon varios métodos, incluidos K-means, Modelos de Mezcla Gaussiana (GMM), y DBSCAN, en un contexto de retail. Su estudio destacó que GMM ofreció la mejor precisión en la segmentación, con un puntaje de Silhouette de 0.80, superando a otros enfoques debido a su capacidad para modelar clusters de formas variadas y capturar la complejidad de los datos de manera más efectiva. Este hallazgo resalta la adaptabilidad de GMM en escenarios donde los comportamientos de los clientes no son claramente diferenciables.

La propuesta de este proyecto difiere de los enfoques presentados en la literatura en varios aspectos clave. En primer lugar, mientras que los estudios revisados han aplicado GMM en contextos de retail y plataformas de comercio electrónico, este proyecto se enfoca específicamente en un centro comercial, un entorno caracterizado por la diversidad de tiendas y productos que influyen en el comportamiento del cliente de maneras únicas. Además, la combinación de GMM con técnicas de reducción de dimensionalidad, como PCA, para optimizar la segmentación y facilitar la interpretación de los resultados, representa una innovación que no ha sido ampliamente explorada en estudios previos.

En conclusión, aunque los estudios existentes proporcionan un marco sólido sobre la aplicación de técnicas de clustering en la segmentación de clientes, el enfoque propuesto en este proyecto se distingue por su adaptación al contexto específico de un centro comercial y la integración de métodos avanzados para mejorar la precisión y aplicabilidad de los resultados en campañas de marketing personalizadas.

Descripción de los datos

Los datos utilizados en este proyecto provienen del **Mall Customers Dataset**, disponible en Kaggle, que contiene información sobre los clientes de un centro comercial. El dataset incluye 200 registros con variables clave como el ID del cliente, género, edad, ingreso anual, y puntaje de gasto. Estas variables

permiten realizar un análisis detallado de los patrones de comportamiento de los clientes, facilitando la segmentación en grupos homogéneos para diseñar estrategias de marketing personalizadas.

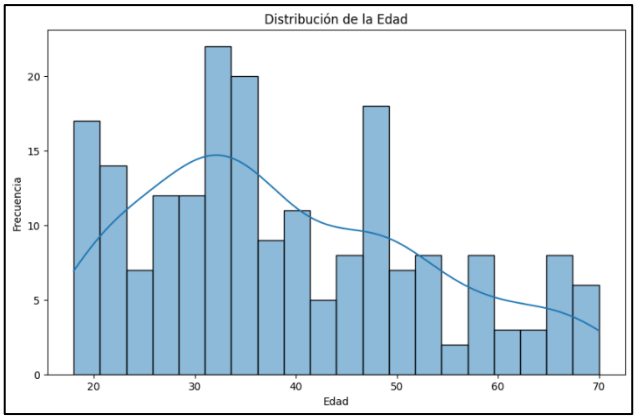
Variables y Tipos:

- ID_Cliente: Identificador único del cliente (tipo int, convertido a str para análisis categórico).
- Género: Género del cliente, categorizado como "Femenino" o "Masculino" (tipo object, convertido a str para el analisis).
- Edad: Edad del cliente en años (tipo int).
- Ingreso_Anual_(k$): Ingreso anual del cliente expresado en miles de dólares (tipo int).
- Puntaje_Gasto_(1-100): Puntaje de gasto del cliente en una escala de 1 a 100, donde valores más altos indican un mayor nivel de gasto (tipo int).

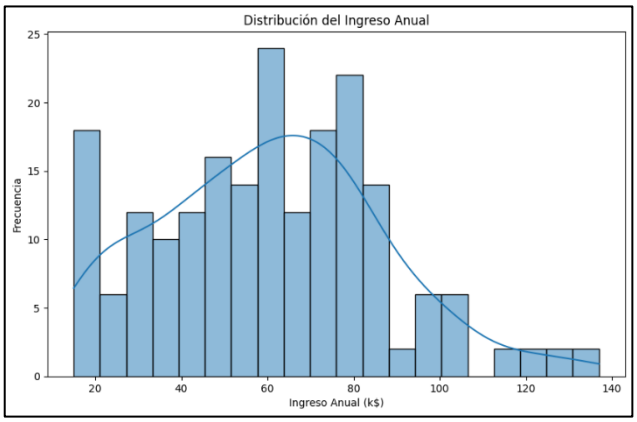
Estadísticas Descriptivas:

Las siguientes estadísticas describen las distribuciones de las variables numéricas en el dataset:

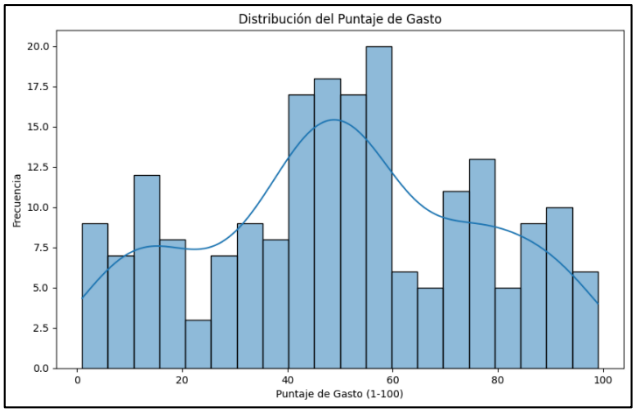
- Edad: La distribución de la edad muestra que los clientes tienen entre 18 y 70 años, con una media de 38.85 años y una desviación estándar de 13.97 años. La gráfica muestra una distribución que se concentra principalmente entre los 20 y 40 años, lo que indica una mayor presencia de clientes jóvenes y de mediana edad en el centro comercial.



- Ingreso_Anual_(k$): Los ingresos anuales varían de 15k$ a 137k$, con una media de 60.56k$ y una desviación estándar de 26.26k$. La gráfica indica que la mayoría de los clientes tienen ingresos entre 40k$ y 80k$, con una menor frecuencia de clientes de ingresos extremadamente bajos o altos, lo que sugiere una base de clientes predominantemente de clase media.

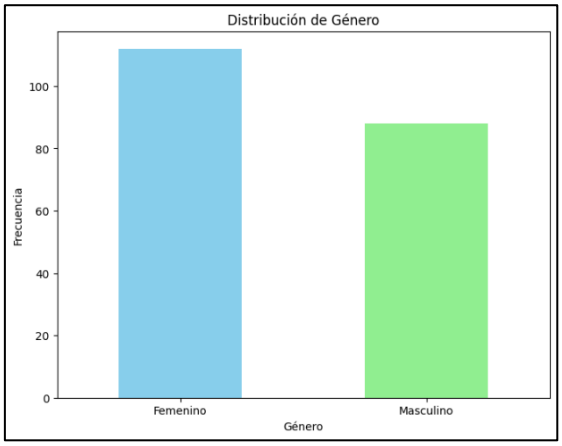


- Puntaje_Gasto_(1-100): El puntaje de gasto varía de 1 a 99, con una media de 50.2 y una desviación estándar de 25.82. La gráfica muestra una distribución casi uniforme, lo que implica que hay una buena variedad de clientes en términos de su nivel de gasto, desde clientes con bajo hasta alto gasto.



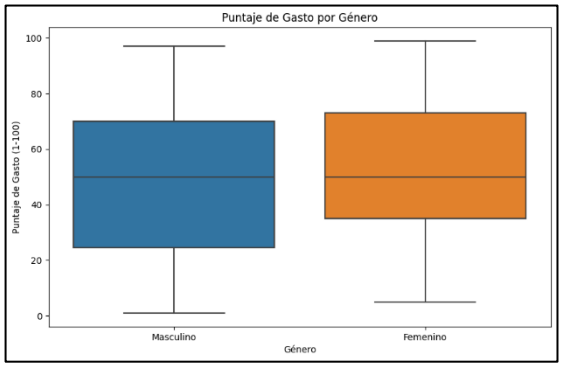
Distribución de Género:

El análisis de género revela que el 56% de los clientes son mujeres y el 44% son hombres. La distribución gráfica muestra una mayor representación femenina, lo que podría influir en la planificación de estrategias de marketing dirigidas.



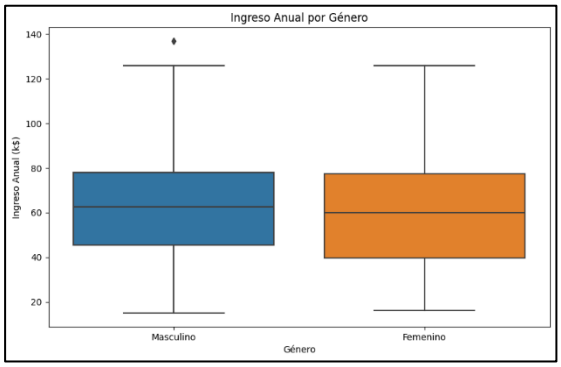
Análisis Comparativo por Género:

Los gráficos de caja muestran que tanto los ingresos anuales como los puntajes de gasto son similares entre géneros, aunque se observa una ligera tendencia a mayores ingresos entre los hombres. No se identifican diferencias significativas en los patrones de gasto entre hombres y mujeres, lo que sugiere que las estrategias de marketing no necesariamente deben diferenciarse por género en términos de incentivos de gasto.



Conclusiones de la Descripción de Datos:

El análisis de las variables demuestra una diversidad considerable en las características de los clientes, como la edad, los ingresos y los hábitos de gasto. No se observan correlaciones extremadamente fuertes entre las variables numéricas, lo que indica que cada variable aporta información única al análisis de segmentación. Esta diversidad de datos es fundamental para construir segmentos diferenciados y personalizados, que puedan ser utilizados para diseñar estrategias de marketing más efectivas en el centro comercial.



Propuesta metodológica

Para abordar la segmentación de clientes en un centro comercial, se ha seleccionado el **Modelo de Mezcla Gaussiana (GMM)** como la técnica principal de análisis. GMM es un modelo probabilístico que asume que los datos son generados por una combinación de múltiples distribuciones gaussianas, permitiendo capturar la complejidad de los comportamientos de los clientes y modelar clusters de

diversas formas y tamaños. A diferencia de métodos como K-means, que asumen que los clusters son esféricos y no permiten solapamientos, GMM ofrece una asignación más flexible y realista de los clientes a los diferentes segmentos, lo que es crucial en un entorno tan variado como el de un centro comercial, donde los patrones de compra pueden ser mixtos y no lineales.

Además de GMM, se consideran otras técnicas como K-means, clustering jerárquico, y DBSCAN. K-means y K-medoides son métodos populares por su simplicidad y rapidez, pero su limitación para capturar formas complejas de clusters los hace menos adecuados en contextos donde los datos no se ajustan a una estructura rígida. El clustering jerárquico y DBSCAN ofrecen ventajas en la detección de estructuras complejas y valores atípicos, pero presentan limitaciones en la escalabilidad y la parametrización. Por ello, GMM se destaca como el método más adecuado, especialmente cuando se combina con técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA), que ayudará a simplificar los datos y mejorar la precisión del modelo.

La elección de GMM se justifica no solo por su capacidad para modelar segmentos complejos y superpuestos, sino también por su adaptabilidad a los datos reales del centro comercial, que incluyen múltiples variables de comportamiento de los clientes. Esta flexibilidad permite un análisis más profundo y detallado de los patrones de compra, facilitando la creación de estrategias de marketing personalizadas y efectivas. Al integrar GMM con técnicas de reducción de dimensionalidad, se busca optimizar tanto la precisión como la interpretabilidad de la segmentación, permitiendo a los administradores del centro comercial tomar decisiones estratégicas basadas en datos y mejorar la experiencia del cliente.

Bibliografía

- John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. Analytics, 2(4), 809-823. https://doi.org/10.3390/analytics2040042
- Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21(3), 527-570. https://doi.org/10.1007/s10257-023-00640-4
- García, M., & López, F. (2020). Clustering jerárquico aplicado a la segmentación de clientes en plataformas de comercio electrónico. *Journal of Marketing Analytics*, 12(3), 234-245.
- Kumar, R., & Shah, A. (2018). Segmentación de clientes mediante K-means en un supermercado: Un enfoque para programas de lealtad. *International Journal of Retail & Distribution Management*, 46(8), 749-764.
- Mall Customers Dataset. Kaggle. Recuperado de: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Repositorio de GitHub

https://github.com/jsmoralesc/Proyecto-ANS