# 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.

***Materiales y Métodos.***

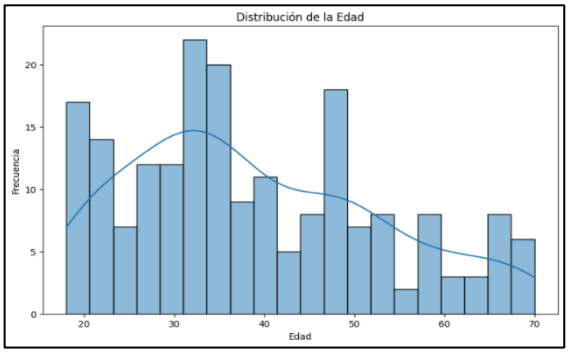
Los datos utilizados en este proyecto provienen del **Mall Customers Dataset**, disponible en Kaggle, que contiene información sobre los clientes de un centro comercial. El dataset incluye 200 registros con variables

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.

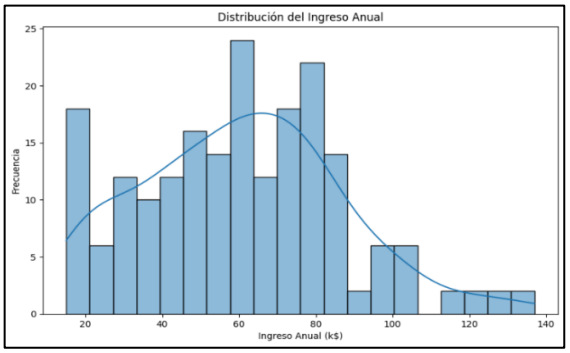
Estadísticas Descriptivas:

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

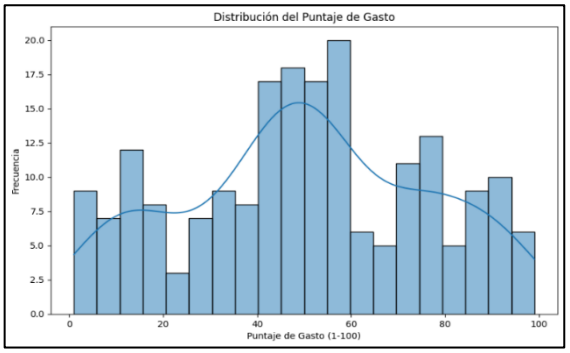
- Edad: La distribución de la edad muestra que los clientes tienen entre 18 y 70 años, con una media de 38.85 años y una desviación estándar de 13.97 años. La gráfica muestra una distribución que se concentra principalmente entre los 20 y 40 años, lo que indica una mayor presencia de clientes jóvenes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

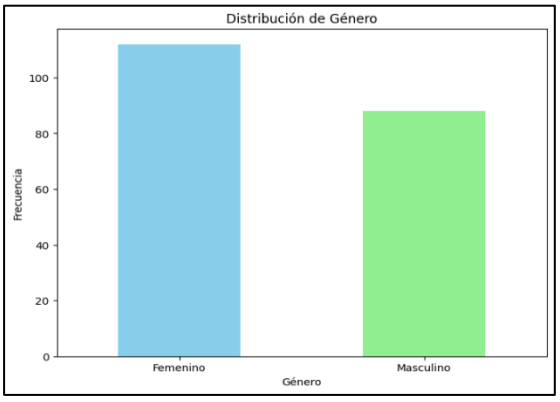



- Puntaje_Gasto_(1-100): El puntaje de gasto varía de 1 a 99, con una media de 50.2 y una desviación estándar de 25.82. La gráfica muestra una distribución casi uniforme, lo que implica que hay una buena variedad de clientes en términos de su nivel de gasto, desde clientes con bajo hasta alto gasto.



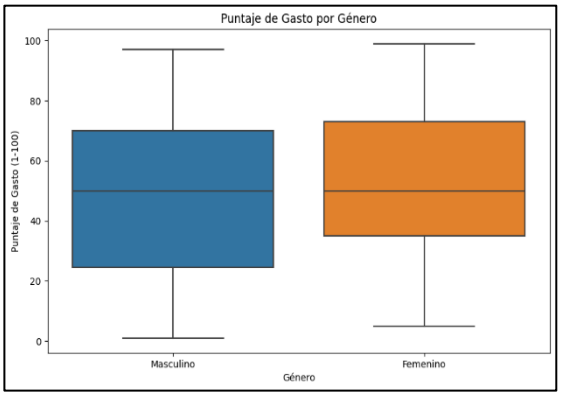

Distribución de Género:

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.



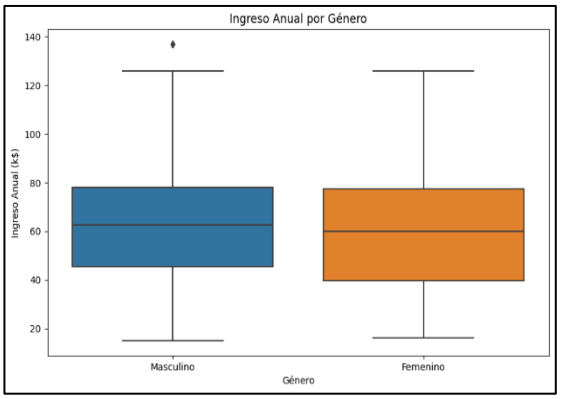
Análisis Comparativo por Género:

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.




Proceso de Limpieza y Preparación de Datos:

No hay valores faltantes ni duplicados, lo que facilita un análisis limpio sin necesidad de imputar o eliminar datos. Para manejar la variabilidad en las escalas de las variables y facilitar una comparación justa, los datos fueron estandarizados. Esto implicó ajustar cada característica para que tuviera una media de cero y una desviación estándar de uno, utilizando la transformación de puntuación Z.

- **Creación de Grupos de Edad y Conversión de Género:** En este estudio, se definieron grupos etarios específicos para facilitar el análisis de los patrones de ingresos y gastos según la edad. Las edades fueron categorizadas en intervalos. Este enfoque permite evaluar cómo las preferencias y capacidades financieras cambian a lo largo de las diferentes etapas de la vida. Adicionalmente, la variable 'Género' fue convertida a valores binarios, asignando 1 para 'Masculino' y 0 para 'Femenino'. Esta transformación binaria es esencial para su inclusión en modelos de aprendizaje automático y análisis estadístico, donde se prefieren variables categóricas simplificadas para una manipulación más eficiente.
- **Análisis de Ingresos y Gastos por Grupos de Edad y Género:** Se realizó un análisis detallado de los ingresos y gastos agrupando los datos tanto por grupos de edad como por género. Para cada grupo, se calcularon estadísticas descriptivas como la media y la desviación estándar de las variables 'Ingreso Anual' y 'Puntaje de Gasto'. Este nivel de detalle es crucial para identificar segmentos de mercado específicos que podrían beneficiarse de estrategias de marketing diferenciadas, ajustadas tanto por edad como por género, maximizando así la efectividad de las campañas dirigidas.

Para abordar la segmentación de clientes en un centro comercial de manera efectiva, se ha optado por el Modelo de Mezcla Gaussiana (GMM) como la técnica principal de análisis. Esta técnica permite identificar clústeres que no necesariamente son esféricos y puede adaptarse a la forma y tamaño de diferentes distribuciones de datos, lo cual es ideal para capturar la diversidad en los patrones de comportamiento de los clientes.

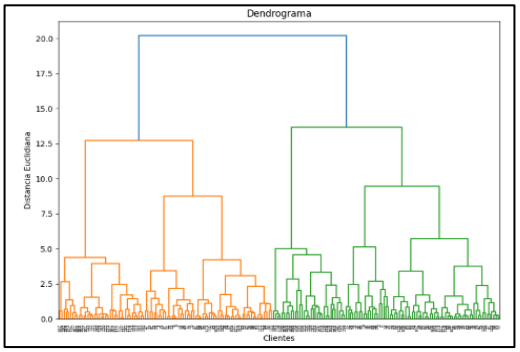
Además del GMM, se han implementado varios algoritmos de clustering para realizar una comparación exhaustiva de las técnicas. Estos incluyen el clustering jerárquico, que organiza los datos en una estructura de árbol basada en la similitud, permitiendo visualizar de manera intuitiva cómo se agrupan los datos en diferentes niveles de granularidad. También se ha utilizado DBSCAN, que es eficaz para identificar clústeres de forma arbitraria y manejar puntos atípicos, proporcionando así robustez frente a anomalías en los datos.

Las técnicas de K-means y K-medoids también se han aplicado para comparar sus rendimientos con el GMM. K-means es conocido por su simplicidad y eficiencia en grandes conjuntos de datos, centrando cada clúster en el promedio de los puntos que lo componen, mientras que K-medoids mejora la robustez del modelo utilizando puntos reales de los datos como centros de clúster, lo que es útil para reducir la influencia de valores atípicos.

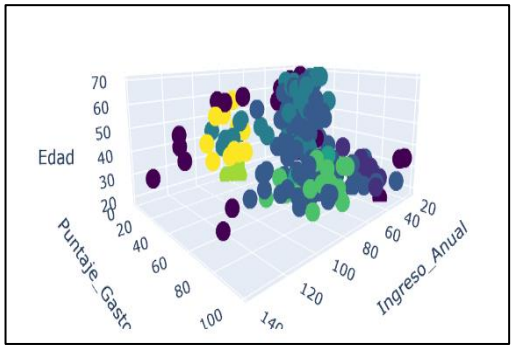
***Resultados y Discusión***

La implementación de varios algoritmos de clustering ha revelado patrones diferenciados y segmentos de clientes dentro del centro comercial. Aquí, se discuten los resultados clave de cada método y las estrategias de marketing propuestas en función de estos hallazgos.

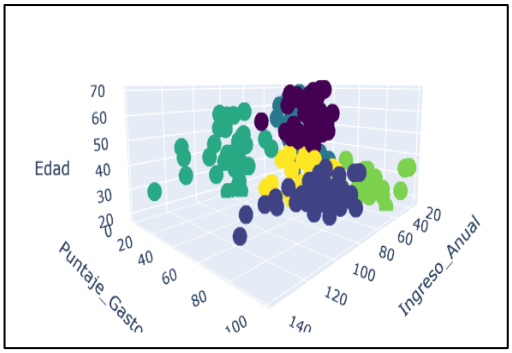
- Clustering Jerárquico: El clustering jerárquico ha identificado grupos con variados niveles de ingresos y gastos. Específicamente, el Cluster 3 sobresale como el segmento de mayor poder adquisitivo y gasto elevado, destacando como el público objetivo para productos y servicios premium. En contraste, el Cluster 1 y el Cluster 5, a pesar de sus ingresos modestos, muestran diferencias en el comporta miento de gasto, ofreciendo oportunidades para dirigir estrategias hacia bienes de consumo rápido y accesible.



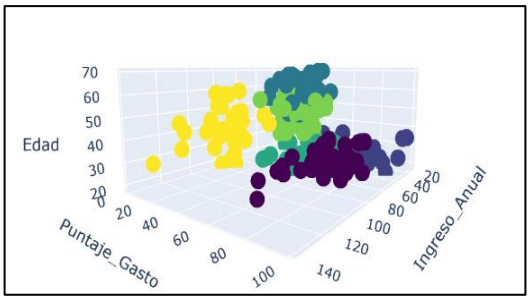

- DBSCAN: El análisis DBSCAN destaca por su habilidad para identificar out liers o ruido (Cluster -1), así como clusters definidos por patrones de ingresos y gastos altos (Clusters 0 y 4). Estos resultados sugieren la existencia de segmentos de mercado no convencionales, desde consumidores frugales que gastan más de lo esperado hasta aquellos con alta capacidad económica que también realizan grandes desembolsos, cada uno requiriendo estrategias promocionales específicas.



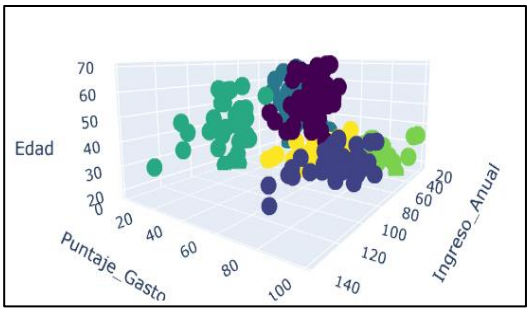

- K-Means: El método K-Means, con su enfoque en min imizar la varianza dentro del cluster, ha proporcionado una visión clara de la segmentación basada en el comportamiento económico. Los Clusters 1 y 3, ambos con altos ingresos pero variados niveles de gasto, ilustran la necesidad de enfoques de marketing diferenciados que potencialmente podrían incrementar la conversión de visitas en ventas, especialmente en el Cluster 3 donde el gasto es bajo.



- K-Medoids: Similar a K-Means, pero más robusto a los outliers, K-Medoids ha demostrado ser eficaz en agrupar consumidores en categorías claras de ingr esos y gastos. Los Clusters 0 y 5 son particularmente notables por su disparidad en el comportamiento de gasto, sugir iendo enfoques personalizados para cada grupo.




- Modelo de Mezcla Gaussiana (GMM): El GMM ha sido efectivo en identificar clusters sutiles, incluyendo aquellos con consumidores de bajos ingresos pero altos gastos (Cluster 4), y de altos ingresos pero bajos gast os (Cluster 3). Esta información es crucial para desarrollar campañas que aumenten la eficacia del gasto publicitario y la satisfacción del cliente.




*Estrategias de Marketing Sugeridas*

- Clusters de Alto Gasto: Ofrecer productos exclusivos y experiencias premium puede atraer más a este grupo, maximizando su contribución al volumen de ventas.
- Clusters de Ingreso Modesto pero Alto Gasto: Implementar descuentos y promociones atractivas puede ser una estrategia eficaz para mantener y aumentar la lealtad y la frecuencia de compra.
- Clusters de Alto Ingreso pero Bajo Gasto: Introducir productos y servicios que resalten el valor y exclusividad, o eventos que atraigan a este segmento a gastar más, puede transformar su potencial en ingresos efectivos para el centro comercial.

*Limitaciones y Futuras Direcciones*

Cada algoritmo tiene limitaciones, como la sensibilidad a los parámetros iniciales o la dificultad en manejar tipos de datos heterogéneos. Futuras investigaciones podrían explorar modelos híbridos o técnicas de machine learning avanzadas para mejorar la segmentación y personalización aún más. Además, profundizar en el análisis del comportamiento en tiempo real y la integración de datos online y offline podría proporcionar una comprensión más holística del cliente. Este enfoque integral no solo mejora nuestra comprensión de las dinámicas del consumidor, sino que también refina las estrategias de engagement y ventas, asegurando que el centro comercial pueda responder de manera efectiva a las necesidades cambiantes de su clientela.

***Conclusión***

Este proyecto ha evidenciado la utilidad del Modelo de Mezcla Gaussiana (GMM) y otras técnicas de clustering para una segmentación efectiva de clientes en un centro comercial, destacando cómo se pueden personalizar estrategias de marketing para mejorar tanto la satisfacción del cliente como el retorno de inversión. Al analizar detalladamente y aplicar múltiples métodos de clustering, hemos obtenido insights valiosos que permiten una toma de decisiones más precisa y centrada en el cliente para los administradores del centro comercial. A pesar de algunas limitaciones, este enfoque integrado sienta las bases para futuras investigaciones que podrían explorar modelos híbridos y técnicas avanzadas, mejorando aún más la precisión y efectividad de la segmentación de clientes.

***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