Proyecto Final de Aprendizaje No Supervisado

Segmentación de Clientes en el Mercado Automotriz Mediante Algoritmos de Clustering

Resumen

Este proyecto tiene como objetivo determinar la agrupación óptima de nuevos clientes en clústeres para una empresa automotriz que busca ingresar a un nuevo mercado basándose en atributos sociodemográficos y de consumo, con el fin de diseñar estrategias efectivas de mercadeo para cada grupo, y maximizar las ventas de automóviles. Se utilizaron datos públicos de la plataforma Kaggle, que incluyen información sobre clientes potenciales. A partir de esta información, se realizó un análisis detallado y un procesamiento de las variables y posteriormente y se aplicó el algoritmo de KMedoids con 4 clústeres. Por último, se analizó la relación entre la segmentación original y los cluster obtenidos y se evaluó la calidad de estos empleando el índice de Rand Ajustado. El resultado principal de este análisis es identificar clústeres de clientes, lo que facilitará la personalización de las campañas de marketing y permitirá a la empresa optimizar sus recursos.

Introducción

Actualmente, las empresas emplean diversas estrategias de mercadeo para captar la mayor cantidad posible de clientes en entornos de alta competitividad, estas estrategias implican inversión en recursos y logística para poder llegar a los clientes objetivo de las empresas, por lo cual es fundamental determinar el mercado óptimo que se quiere abordar de acuerdo con los objetivos empresariales de cada organización, minimizando costos y maximizando las ventas. Dada la creciente disponibilidad de datos cada día sobre el comportamiento de los clientes y las tendencias de los mercados se puede optimizar estas estrategias de mercadeo con algoritmos de aprendizaje no supervisado, como el clustering. Estos algoritmos pertenecen al campo de la minería de datos y son una técnica fundamental dentro del aprendizaje automático, específicamente en el área de segmentación de datos, la cual se enfoca en identificar patrones en los datos sin tener etiquetas establecidas o supervisión explicita.

El problema a solucionar en este proyecto se enmarca en el contexto de una empresa automotriz que planea ingresar a un nuevo mercado con sus productos actuales. La empresa cuenta con una base de datos de sus clientes potenciales, con información sociodemográficas y de consumo. En este sentido, se establece la siguiente pregunta problema: ¿Cuál es la agrupación óptima de clientes de la empresa de venta de automóviles en un nuevo mercado que permita maximizar las ventas de sus productos, mediante estrategias efectivas y personalizadas de mercadeo para cada clúster identificado?

Para abordar este problema, se revisó la literatura nacional e internacional sobre la segmentación de mercados utilizando algoritmos de clustering. A nivel nacional, estudios recientes han empleado modelos como DBSCAN y k-means en sectores como el consumo masivo y las bebidas alcohólicas, mientras que, a nivel internacional estos enfoques han sido aplicados en industrias bancarias, de telecomunicaciones y automotriz. En el sector automotriz, gracias a los algoritmos implementados, se identificaron clústeres clave basados en las preferencias de compra y satisfacción del cliente, lo que ha permitido desarrollar estrategias de marketing específicas y efectivas.

Ahora bien, en este proyecto se busca identificar los clústeres más relevantes en un nuevo mercado automotriz, utilizando la base de datos disponible de clientes potenciales. Estos clústeres se definirán con base en características sociodemográficas y de consumo, lo que permitirá generar un perfil claro de los distintos segmentos. Con base en los patrones identificados, la empresa podrá diseñar e implementar estrategias de marketing personalizadas para cada grupo, optimizando las campañas promocionales y maximizando el impacto el cada segmento de mercado.

Descripción detallada de los datos

Exploración de los datos:

El conjunto de datos utilizado proviene de la plataforma Kaggle, donde se especifica que los datos son públicos y están disponibles para el acceso. El objetivo de utilizar este conjunto de datos es simular una situación, en la que una empresa automovilística se enfrenta a un problema de segmentación y resolverlo aplicando los conocimientos adquiridos en aprendizaje no supervisado. El conjunto de datos incluye información sobre los clientes potenciales del nuevo mercado.

A continuación, se presenta el enlace de donde se obtuvieron los datos, junto con su análisis descriptivo. Fuente de Datos Kaggle

Variables:

Variables numéricas de tipo entero:

Nombre Variable	*Tiene Nulos*	*Descripción*
Edad	No	La edad que tiene el cliente en el momento actual
Experiencia Laboral	Si	La cantidad de años de experiencia laboral que el cliente ha acumulado a lo largo de su vida
Tamaño Familia	Si	El total de personas que conforman la familia del cliente en el momento actual

Variables categóricas:

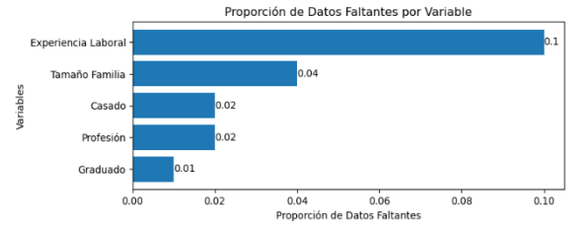
Nombre Variable	*Tiene Nulos*	*Descripción*
ID	No	Número de registro único del cliente en la empresa
Genero	No	Género del cliente registrado en su documento de identidad, especificado como Masculino o Femenino
Casado	Si	Indicación de si el cliente está casado, con opciones de respuesta Sí o No
Graduado	Si	Indicación de si el cliente ha completado un grado universitario, con opciones de respuesta Sí o No
Profesión	Si	Profesión que el cliente desempeña actualmente
Categoría de Gasto	No	Clasificación del cliente basada en sus patrones de consumo en el mercado, determinada por la empresa.

Dimensión de los Datos

En total, el conjunto de datos cuenta con 8,068 registros, cada uno correspondiente a un cliente específico, garantizando que no existen duplicados en los ID. Además, se dispone de 9 columnas que se utilizarán para realizar una segmentación adecuada de los clientes, lo que permitirá lanzar campañas focalizadas.

Es importante resaltar que algunas variables contienen valores nulos y esto nos permite evidenciar la necesidad de aplicar métodos para reemplazar los datos faltantes, ya que eliminar los registros implicaría no clasificar a 1,403 clientes. A continuación, se detalla la proporción de datos faltantes para cada una de las variables:

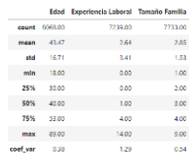
Se puede observar que la variable con la mayor cantidad de datos faltantes es la de años de experiencia laboral, con un 10%. Las otras cuatro variables presentan un porcentaje inferior al 5%.



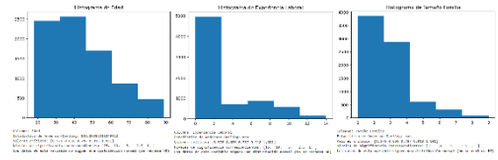
Para abordar los datos faltantes, se realizó una imputación utilizando la mediana para las variables numéricas y la moda para las variables categóricas. En el caso de la variable “Profesión”, que presentaba un 2% de valores nulos, se optó por eliminar esos registros, dado que su proporción era baja y no afectaba de manera significativa la calidad del análisis. Este enfoque permitió preservar la estructura de los datos, ya que la correlación entre las variables se mantuvo. Además, la distribución de porcentajes en las diferentes categorías de las variables categóricas no cambió de manera significativa, confirmando que la estructura de datos no se alteró sustancialmente.

Estadísticas descriptivas y Gráficos:

Variables Numéricas:



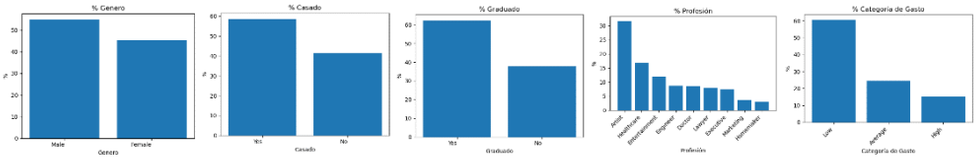
	Edad	Experiencia Laboral	Tamaño Familia
count	8068.00	7239.00	7733.00
mean	43.47	2.64	2.85
std	16.71	3.41	1.53
min	18.00	0.00	1.00
25%	30.00	0.00	2.00
50%	40.00	1.00	3.00
75%	53.00	4.00	4.00
max	89.00	14.00	9.00
coef_var	0.38	1.29	0.54



Observamos que ninguna de las distribuciones presentadas muestra una distribución normal. Tanto la edad, la experiencia laboral, como el tamaño de la familia presentan una concentración predominante hacia la izquierda. Además, al analizar la tabla con las medidas de tendencia central, se pueden resaltar los siguientes puntos:

- **Edad:** El 75% de los clientes tienen 53 años o menos, lo que indica que la mayoría se encuentra en el rango de 18 a 53 años. A medida que aumenta la edad, la cantidad de clientes disminuye, se observa que hay 2,000 clientes de 45 años, pero solo 500 de 85 años.
- **Experiencia Laboral:** La experiencia laboral promedio es de aproximadamente 2 años y medio. Es importante resaltar que un 25% de las personas no tiene experiencia laboral, mientras que la persona con más experiencia cuenta con 14 años. El coeficiente de variación de 1.29 sugiere una alta dispersión en esta variable en comparación con otras y la presencia de valores atípicos.
- **Tamaño de la Familia:** El 75% de las familias tiene entre 1 y 4 integrantes, además, se observa que algunos clientes viven solos, mientras que en el extremo opuesto hay familias de hasta 9 integrantes, siendo este el tamaño máximo registrado.

Variables Categóricas:

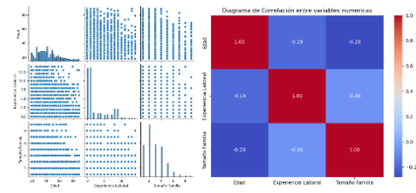


- **Género:** La empresa tiene más clientes hombres, aunque la diferencia con respecto a las mujeres es de solo un 9.5%.
- **Casado:** La mayoría de los clientes están casados, representando el 59% del total.
- **Graduado:** El 62% de los clientes se graduaron de la universidad, lo que muestra una proporción significativa de clientes con estudios superiores.
- **Profesión:** El 61% de los clientes trabaja en profesiones relacionadas con el arte, la salud o el entretenimiento, siendo artista la profesión más común con un 32%. Además, un 3% de los clientes se dedica a las tareas del hogar.
- **Categoría de Gasto:** La mayoría de los clientes de la empresa tienen un hábito de consumo bajo, representando aproximadamente un 61% del total.

Relación Entre variables:

Relación entre Variables Numéricas:

Al analizar el gráfico de dispersión que combina las diferentes variables, se observa que no existe una correlación lineal evidente entre ellas. Esto se confirma al calcular la correlación de Pearson, que muestra que las correlaciones entre las variables son bajas y poco significativas.

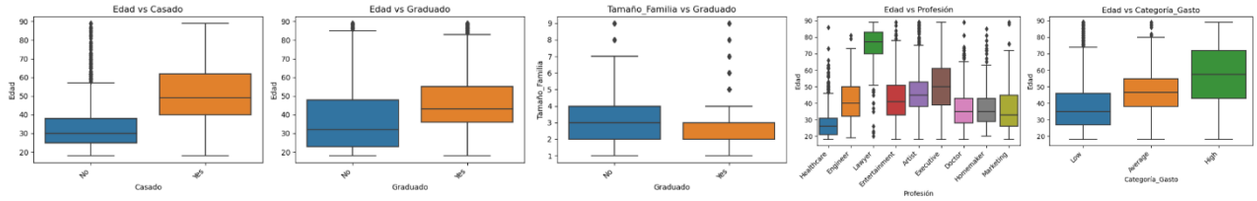


Relación entre Variables Categóricas:

Al realizar las tablas de contingencia entre variables categóricas, aplicar la prueba de chi-cuadrado y calcular el V de Cramer para medir la fuerza de la relación, encontramos un V de Cramer superior a 0.5 en las siguientes relaciones:

- **Casado y Profesión (V de Cramer igual a 0.52):** Observamos que, si el cliente está casado, es más probable que pertenezca a ciertas profesiones en comparación con otras, pues los clientes casados tienen una mayor probabilidad de ser Artistas, Ejecutivos o Abogados.
- **Casado y Categoría de Gasto (V de Cramer igual a 0.67):** Esto indica que los clientes que no están casados tienen exclusivamente una categoría de gasto baja, mientras que los clientes casados se distribuyen entre las diferentes categorías de gasto.

Relación entre Variables Numéricas y Categóricas:



Al realizar el análisis ANOVA para determinar la relación entre las variables categóricas y numéricas, y evaluar la significancia estadística de estas relaciones, hemos identificado las siguientes relaciones, donde el valor de F indica la fuerza de la relación:

- **Edad y Casado (F igual a 3,758):** Los clientes casados suelen tener mayor edad, mientras que los no casados tienden a tener edades más bajas.
- **Edad y Graduado (F igual a 477):** La mayoría de los clientes graduados se concentran entre los 35 y 55 años, mientras que los no graduados abarcan un rango más amplio, de 22 a 50 años.
- **Tamaño de Familia y Graduado (F igual a 432):** Los clientes graduados tienen familias más pequeñas, con un percentil 75 de 3 integrantes, en contraste con los no graduados, cuyo percentil 75 es de 4. Además, los graduados presentan más casos atípicos en el tamaño de sus familias.
- **Edad y Profesión (F igual a 977):** La edad varía según la profesión, un dato a resaltar es que el 75% de los clientes en profesiones relacionadas con leyes tienen más de 70 años.
- **Edad y Categoría de Gasto (F igual a 846):** A mayor edad, los clientes se clasifican en la categoría de gasto alta y los clientes más jóvenes, entre 25 y 45 años, tienden a tener una categoría de gasto baja.

Metodología:

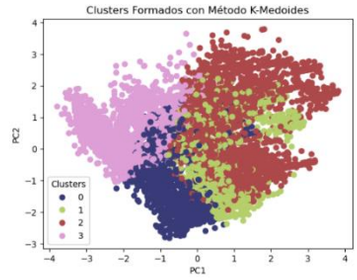
- ***Diseño del proyecto:*** Este proyecto se enfoca en la segmentación de clientes de una empresa automotriz por medio del algoritmo K-Medoides para determinar los clústeres que diferencian de forma óptima los clientes y diseñar estrategias de mercado efectivas en función de algunas características sociodemográficas y de consumo.
- ***Target y muestra:*** La base de datos contiene información sociodemográfica y de consumo de clientes de la empresa automotriz, que incluye datos sobre la edad, experiencia laboral, tamaño familiar, género, estado civil, nivel educativo, profesión y categoría de gasto. La muestra contiene 8068 registros de clientes.
- ***Fuente y preprocesamiento de los datos:*** Los datos son públicos y disponibles para el acceso y se obtuvieron de la plataforma Kaggle. El preprocesamiento incluye imputación de datos faltantes con el valor más frecuente para variables categóricas e imputación con la mediana para variables numéricas, codificación de variables categóricas como “Casado”, “Graduado” “Categoria_Gasto” y “Segmentation”, PCA para reducir la base de variables dummies y escalado de los datos a utilizar en modelo para que estos tengan una media de 0 y una desviación estándar de 1.
- ***Procedimiento y modelo:*** Se analizaron los resultados de diferentes modelos propuestos, algunos con reducción de dimensionalidad y otros seleccionando solo algunas de las variables disponibles y se seleccionó aquel que presentaba una mejor agrupación. Como primer paso se realizaron algunos ajustes a la base de datos, los cuales incluyen agrupación de las profesiones en categorías más generales, transformación de variables categóricas en dummy y escalado de los datos. Posterior a esto, se realizó la reducción de dimensionalidad mediante un PCA recuperando el 95 % de la variabilidad de los datos. Sobre esta base de datos ajustada se aplicó el algoritmo de KMedoids con 4 clusters. Por último, se analizó la relación entre la segmentación original y los clústeres obtenidos y se evaluó la calidad de estos empleando el índice de Rand Ajustado.
- ***Análisis de datos:*** Se generaron estadísticas descriptivas de cada clúster obtenido con el modelo, las cuales se presentan en la siguiente tabla resumen:

Clúster	Edad Promedio	Familia (Integrantes)	Experiencia Laboral (Años)	% Hombres	% Casados	% Graduados	Profesión Predominante		Categoría de Gasto
1	38	2	4	57%	97%	81%	Arte y entretenimiento	74%	Baja
2	56	2	2	61%	98%	73%	Arte y entretenimiento	55%	Baja (99,9%)
3	51	3	2	58%	100%	73%	Arte y entretenimiento	53%	Media (60%), Alta (38%)
4	28	4	2	53%	86%	70%	Salud	82%	Baja (95%)

Resultados y Discusión

Después de aplicar el modelo K-Medoides para agrupar los datos en 4 clústeres, se realizó un análisis detallado de las variables en cada uno, identificando diferencias clave que nos ayudarán a definir el público objetivo ideal para aplicar las estrategias de marketing. Estas estrategias estarán orientadas a atraer nuevos clientes interesados en la compra de

vehículos. A continuación, se muestran los clústeres en dos dimensiones, junto con una descripción detallada de cada uno de ellos:



Descripción Clúster 0 (1.423 Personas)

- ***Variables Numéricas:*** Las personas de este clúster tienen una edad promedio de 38 años, con familias de 2 integrantes y 4 años de experiencia laboral en promedio. Sin embargo, el 50% posee tiene 1 año de experiencia laboral.
- ***Variables Categóricas:*** En este clúster el 57% son hombres, el 97% están casados, el 81% son graduados, y la profesión predominante es arte y entretenimiento, representando al 74% de las personas, además se resalta que aquí todos tienen categoría de gasto baja.

Descripción Clúster 1 (1.289 Personas)

- ***Principales Diferencias Clúster 1 vs Clúster 0:***

    - Mayor edad promedio (la más alta en comparación con los otros clústeres).
    - Menos años de experiencia laboral.
    - Menor porcentaje de graduados.
    - Aunque la profesión de Arte y Entretenimiento sigue siendo la predominante, su proporción disminuye, y la profesión legal se posiciona como la segunda más importante.

- ***Variables Numéricas:*** Las personas en este clúster tienen una edad promedio de 56 años, con familias de 2 integrantes y 2 años de experiencia laboral, siendo este valor igual tanto en la media como en la mediana.
- ***Variables Categóricas:*** En este clúster, el 61% son hombres, el 98% están casados, el 73% están graduados y el 99,9% tienen categoría de gasto baja. La profesión predominante es Arte y Entretenimiento, representando al 55% de las personas, mientras que la profesión legal ocupa el segundo lugar con un 19%.

Descripción Clúster 2 (3.103 Personas)

- ***Principales Diferencias Clúster 2 vs Clúster 0 y 1:***

    - Aumento del número de integrantes en la familia a 3.
    - Todos los integrantes están casados.
    - La segunda profesión más relevante es Negocios y Marketing.
    - La mayoría de las personas tienen una categoría de gasto media.
    - Es el único clúster con la mayor proporción de personas en la categoría de gasto alta.
    - Es el clúster con mayor número de personas.

- ***Variables Numéricas:*** En este clúster, las personas tienen una edad promedio de 51 años, con familias de 3 integrantes y 2 años de experiencia laboral.
- ***Variables Categóricas:*** El 58% de las personas en este clúster son hombres, el 100% están casados y el 73% son graduados. El 60% tiene una categoría de gasto media, mientras que el 38% pertenece a la categoría de gasto alta. La profesión predominante es Arte y Entretenimiento, con un 53%, seguida de Negocios y Marketing, con un 17%.

Descripción Clúster 3 (2.129 Personas)

- ***Principales Diferencias Clúster 3 vs Clúster 0, 1 y 2:***

- Es el clúster con la menor edad promedio y el mayor número de integrantes en la familia.
- Mayor proporción de mujeres.
- Menor porcentaje de personas casadas.
- La profesión principal es el sector salud.

- ***Variables Numéricas:*** Las personas en este clúster tienen una edad promedio de 28 años, viven en familias de 4 integrantes y cuentan con 2 años de experiencia laboral.
- ***Variables Categóricas:*** El 53% de los integrantes de este clúster son hombres, el 86% están casados y el 70% son graduados. Además, el 95% pertenece a la categoría de gasto baja, y la profesión predominante es salud, representando al 82% de las personas.

Limitaciones del enfoque

Si bien esta metodología permitió segmentar a los clientes en 4 grupos diferenciados por sus características sociodemográficas, se observa que por la complejidad de las interacciones sociales este agrupamiento no es totalmente excluyente y que se solapan algunas de las observaciones. Una de las principales limitaciones del enfoque empleado es que requiere especificar el número de clúster a priori, lo que puede llegar a condicionar los resultados. Por otra parte, en espacios de alta dimensión como el del presente estudio, la noción de cercanía puede volverse menos intuitiva, y las medidas de disimilitud pueden ser menos efectivas.

Conclusión y Recomendación final:

Dado que el objetivo es maximizar las ventas de automóviles mediante la identificación de un grupo objetivo que permita atraer nuevos clientes a la empresa, y que esto implica diseñar estrategias de mercadeo para generar una respuesta de compra en el consumidor final, se recomienda enfocar las campañas de marketing en el clúster 2.

Este clúster destaca entre los cuatro analizados por ser el que reúne al mayor número de personas, además de contar con una mayoría de integrantes que presentan una categoría de gasto media o alta, lo que lo convierte en un objetivo clave para las estrategias comerciales. Con una experiencia laboral promedio de 2 años, el grupo podría reflejar estabilidad financiera. Las familias en este clúster, con un promedio de 3 integrantes, lo hacen especialmente atractivo para promover la compra de automóviles, ya que ofrecen una solución cómoda para el transporte familiar, a diferencia de otros vehículos como las motocicletas.

Un aspecto clave es la diversidad en la distribución de profesiones dentro del clúster 2, lo que permite adaptar las campañas a diferentes públicos. El automóvil no solo sería útil en el ámbito familiar, sino también para satisfacer necesidades de movilidad laboral. Por ejemplo, personas que trabajan en arte y entretenimiento, abogados, ingenieros, o quienes requieren un vehículo para sus actividades profesionales, encontrarán en el automóvil una solución versátil tanto para la vida personal como laboral.

Bibliografía.

Abadía, F. A. P., & Patiño, N. A. P. (2020). Segmentacion De Clientes De Una Empresa Comercializadora De Productos De Consumo Masivo En La Ciudad De Popayán Soportado En Machine Learning Y Analisis RFM (Recency, Frecuency y Money). Tesis pregrado. Fundación Universitaria de Popayán. Programa de ingeniería de sistemas. Popayán.

Chambi Condori, P. P. (2023). Segmentación de mercado: Machine Learning en marketing en contextos de covid-19. *Industrial Data*, *26*(1), 275-301.

Farhan, M., & Heikal, J. (2024). Used car customer segmentation using K-Means clustering model with SPSS: Case study Caroline.id. *Journal of Indonesian Social Science*, *5*(3), 543-558. https://www.researchgate.net/publication/379443160_Used_Car_Customer_Segmentation_Using_K-Means_Clustering_Model_With_SPSS_Program_Case_Study_CarolineId

IBM. (s.f.). ¿Qué es el aprendizaje no supervisado? Obtenido de https://www.ibm.com/mx-es/topics/unsupervised-learning

Hartoyo, H., Manalu, E., Sumarwan, U., & Nurhayati, P. (2023). Driving success: A segmentation of customer admiration in automotive industry. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(2), 100031.

Hassan, M. M. T. M., & Tabasum, M. (2018). Customer profiling and segmentation in retail banks using data mining techniques. *International journal of advanced research in computer science*, *9*(4), 24-29.

Lewaaelhamd, I. (2024). Customer segmentation using machine learning model: an application of RFM analysis. *Journal of Data Science and Intelligent Systems*, *2*(1), 29-36.

Mariño Santos, C. (2023). Análisis de clustering para la segmentación del mercado: un caso de estudio de una aplicación de una bebida alcohólica en las principales ciudades de Colombia. Universidad El Bosque. Facultad de Ciencias, Departamento de Matemáticas. Bogotá D.C, Colombia.

Rajesh, M. (2021). Customer segmentation using machine learning. Recent Trends in Intensive Computing, 39, 239.

Tsai, C. F., Hu, Y. H., & Lu, Y. H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, *32*(1), 65-76.