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 de venta de automóviles basándose en algunos atributos sociodemográficos y de consumo, con el fin de diseñar estrategias efectivas de mercadeo para cada grupo de individuos con características similares y maximizar las ventas de automóviles en un nuevo mercado. Por lo cual, inicialmente se hace una revisión preliminar de antecedentes en la literatura relacionada con algoritmos de clustering, para identificar en el contexto del problema a solucionar algunas de las investigaciones y métodos relevantes relacionados que permitan generar un estado del arte. Posteriormente, se realiza una exploración inicial de los datos para identificar las características de cada variable y la aplicabilidad de cada una en la propuesta metodológica establecida finalmente, con la cual se busca responder la pregunta problema por medio del análisis de la información generada a partir de los datos e identificar elementos que se pueden optimizar en el modelo propuesto para generar mayor precisión en los resultados.

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 y dispone de una base de datos de sus clientes con algunas características 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?

Revisión preliminar de antecedentes en la literatura.

La segmentación de mercados empleando herramientas de machine learning es un campo ampliamente estudiado tanto a nivel nacional como internacional, la revisión de la literatura al respecto muestra que los investigadores exploran estas técnicas desde diferentes perspectivas. En Colombia, se desarrolló un estudio empleando el uso de DBSCAN en combinación con técnicas de reducción de dimensionalidad para la segmentación de mercado en la industria de bebidas alcohólicas, el objetivo era identificar patrones entre los consumidores de alcohol para generar información que permitiera desarrollar estrategias de mercadeo efectivas, se analizaron cinco modelos de clustering y como métricas de evaluación se emplearon el índice de Calinski-Harabasz, el coeficiente de silueta y Davies-Bouldin y criterios como la simplicidad y claridad en la diferenciación de clusters buscando un modelo eficaz e interpretable, el modelo con mejor desempeño se obtuvo relizado reducción de la dimensionalidad con UMAP (Uniform Manifold Approximation and Projection) y posteriormente DBSCAN con hiperparametros optimizados (Mariño, 2023). También se realizó un estudio de aplicación de la técnica de clustering por medio del modelo RFM y del algoritmo de k-means para conocer la distribución de clientes en una empresa comercializadora de productos de consumo masivo, con el objetivo de lograr la fidelización de estos, como resultado se identificación 4 clusters aplicando k-means, los cuales se clasificaron como: Clientes VIP, Clientes Buenos, Clientes Regulares y Clientes de Poco Aporte (Abadía & Patiño, 2020).

En Perú, Chambi (2023) realizó una segmentación de consumidores sobre sus percepciones al comprar durante el COVID-19 usando k-means, identificando 4 grupos. Lewaaelhamd (2024) utilizó K-means, DBSCAN y RFM para segmentar el mercado y predecir el abandono de clientes, encontrando que DBSCAN produjo 6 clusters mejor interpretables. A nivel internacional, Hassan y Tabasum (2018) aplicaron Naive Bayes y BIRCH para segmentar y perfilar clientes bancarios, mientras que Rajesh (2021) segmentó clientes en telecomunicaciones con K-means, identificando 5 clusters y destacando la importancia de los gastos totales e ingresos anuales.

En el sector automotriz, Tsai et al. (2015) compararon k-means y EM EM (expectaction Maximization) para segmentar clientes de una concesionaria taiwanesa, basándose en información de satisfacción y ventas. Identificaron cuatro grupos: clientes fieles, potenciales, VIP y perdidos, permitiendo desarrollar estrategias de marketing específicas. Hartoyo et al. (2023) realizaron un análisis de clúster basado en un cuestionario en línea que recolectó información demográfica y sobre automóviles, segmentando a los clientes en tres grupos: "soñadores", "trabajando por el éxito" y "orientados a la familia", proporcionando una herramienta para comprender mejor el comportamiento del consumidor y crear estrategias de marketing específicas.

Otro ejemplo en el sector automovilístico es el estudio “Used Car Customer Segmentation Using K-Means Clustering Model With SPSS”, que utiliza el algoritmo K-Means para segmentar a los clientes del mercado de autos usados de Caroline.id en grupos con características similares. Este análisis permitió identificar dos clústeres principales, destacando uno conformado principalmente por millenials de 31 a 40 años, que prefieren autos de la marca Toyota con transmisión automática y que generalmente optan por pagos en efectivo. Estos resultados fueron clave para que la compañía pudiera diseñar estrategias de marketing personalizadas y más efectivas. Un ejemplo de ello fue promocionar vehículos específicos que son preferidos por el segmento más grande, como los Toyota de modelos recientes con transmisión automática. Además, se optimizaron las campañas promocionales en plataformas relevantes para este grupo, como OLX y redes sociales, y se implementaron estrategias de precios y facilidades de pago que reflejas las preferencias identificadas.

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. 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:

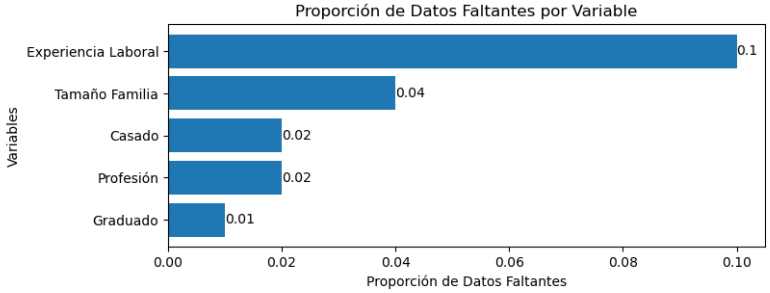
Nombre Variable	***Tiene Nulos***	***Descripción***
ID	No	Número de registro único del cliente en la empresa
Genero	No	Género del cliente registrado en su documento de identidad, especificado como Masculino o Femenino
Casado	Si	Indicación de si el cliente está casado, con opciones de respuesta Sí o No

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:

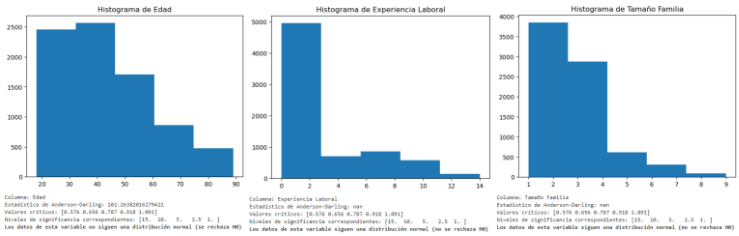


Se puede observar que la variable con la mayor cantidad de datos faltantes es la de años de experiencia laboral, con un 10%. Las otras cuatro variables presentan un porcentaje inferior al 5%.

Estadísticas descriptivas y Gráficos:

Variables Numéricas:

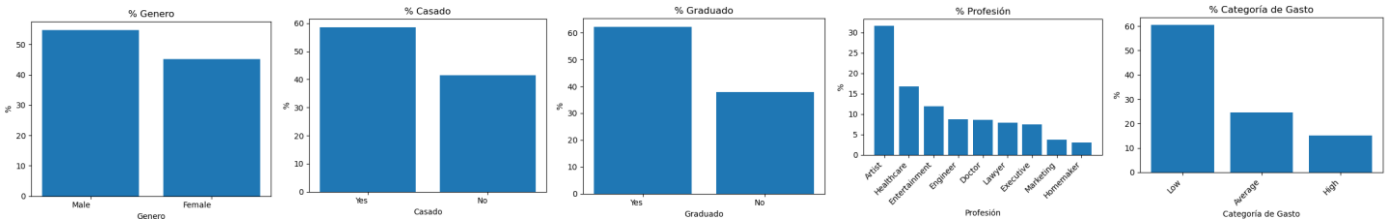
	Edad	Experiencia Laboral	Tamaño Familia
count	8068.00	7239.00	7733.00
mean	43.47	2.64	2.85
std	16.71	3.41	1.53
min	18.00	0.00	1.00
25%	30.00	0.00	2.00
50%	40.00	1.00	3.00
75%	53.00	4.00	4.00
max	89.00	14.00	9.00
coef_var	0.38	1.29	0.54



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.

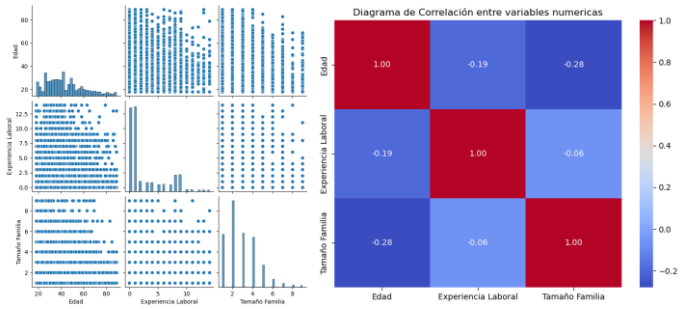
Variables Categóricas:



- **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:

```
Casado - Profesión

Tabla de contingencia:
 Profesión  Artist  Doctor  Engineer  Entertainment  Executive  Healthcare  \
Casado
No             713     366       267            358         37        1153
Yes           1774     311       415            579        550         145

Profesión  Homemaker  Lawyer  Marketing
Casado
No               112      40        189
Yes              128     575         96

Estadístico de chi-cuadrado: 2101.7281875939775
Valor p: 0.0
Grados de libertad: 8
```

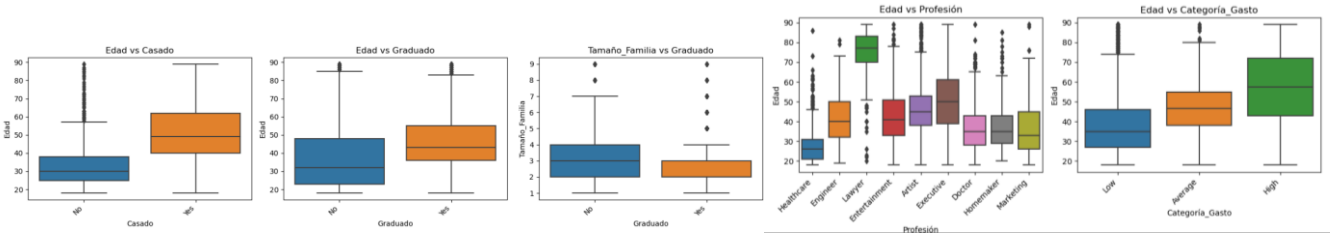
```
Casado - Categoría de Gasto

Tabla de contingencia:
 Categoría de Gasto  Average  High   Low
Casado
No                        0     0  3285
Yes                    1937  1176  1530

Estadístico de chi-cuadrado: 3626.4622494721225
Valor p: 0.0
Grados de libertad: 2
V de Cramer: 0.6763317252705825
```

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.

Propuesta metodológica

Como se describió anteriormente, en este proyecto se va a trabajar la segmentación de clientes de una compañía automotriz. Este tema incorpora el aprendizaje no supervisado, dado que no se tienen etiquetas predefinidas y se pretende identificar los distintos grupos de clientes basados en sus características demográficas. Cabe aclarar que los modelos de aprendizaje no supervisado se utilizan para tres tareas principales: agrupamiento, asociación y reducción de dimensionalidad. En este caso, se realizará una agrupación, que es una técnica de minería de datos que agrupa datos sin etiquetar en función a sus similitudes o diferencias. (IBM, s.f.)

Para este proyecto, se considerará el uso del algoritmo de K-Means para realizar la segmentación de clientes. Este método es adecuado para agrupar datos basados en su similitud. La idea es encontrar grupos dentro de los datos de los nuevos clientes que puedan corresponder con las categorías de segmentación previamente identificadas (A, B, C, D).

Antes de implementar el algoritmo, se realizará el análisis exploratorio y preprocesamiento de datos que incluye la limpieza y preparación de estos. Este proceso es esencial para asegurar que los datos sean consistentes, completos y adecuados para el análisis. Así mismo, se considerará realizar una reducción de dimensionalidad para de conservar la información más relevante y facilitar la identificación de los clústeres.

Posteriormente, se implementará el algoritmo, siguiendo los pasos pertinentes para determinar el número ideal de segmentos en los que se agruparán los clientes. Para esto, se utilizará el método del codo o el coeficiente de Silhouette, los cuales sirven de guía para determinar el número de clústeres.

Una vez determinado el número óptimo de clústeres, se ejecutará el algoritmo y se realizará un análisis de los resultados obtenidos. Este análisis permitirá interpretar y comparar los clústeres generados con las categorías de segmentación previamente identificadas (A, B, C, D). Dependiendo de los resultados, se evaluará la necesidad de ajustar los parámetros del algoritmo o considerar otras técnicas de clustering que puedan mejorar la segmentación.

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.