Instituto Tecnológico
y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Segmentación de Clientes para un *E-commerce* soportado en Shopify

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Presenta: **DANIEL LAGUNAS BARBA**

Director **DR. JAIME EMMANUEL ALCALÁ TEMORES**

Tlaquepaque, Jalisco, febrero de 2024.

AGRADECIMIENTOS

El autor desea dar las gracias al conjunto de profesores que le impartieron materia a lo largo de la maestría, ya que sus conocimientos moldearon su experiencia en el mundo de la Ciencia de Datos.
Así mismo a la institución (ITESO), por el apoyo económico que se le confirió mediante una beca.
También a su familia, amigos y pareja por siempre alentarlo con sus palabras y buenos deseos, así como siempre reafirmar sus capacidades intelectuales para surcar los retos que se le presentaron.
Y, por último, al dueño del *E-commerce* por compartir su confianza y sus datos para la elaboración y desarrollo de este trabajo de obtención de grado.

RESUMEN

La problemática que se planteó resolver en este trabajo fue la de generar perfiles de compra con base a los históricos de consumo de los compradores en la tienda para así se pudieran desarrollar estrategias de marketing que generen un aumento en las ventas. El nombre de la tienda se mantiene en anónimo durante todo el documento y funciona como una tienda de abarrotes o supermercado a domicilio a través de *Shopify*. Como una primera aproximación, se comenzó por usar modelos de *clustering* dado que se trataba de un problema de aprendizaje no supervisado. Ya que los modelos de *clustering* no dieron los resultados esperados, se desarrolló un sistema de recomendación para darle sugerencias a los clientes e incentivarlos a probar productos nuevos con base al comportamiento colectivo de todos los compradores.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

Resumen: En este capítulo se presenta brevemente algunos antecedentes de los análisis sobre el comportamiento de compra de clientes en tiendas de supermercado o tiendas de conveniencia, la justificación de implementar este tipo de análisis y modelado para una tienda *E-commerce*, la definición del problema o situación actual en la que se encuentra dicha tienda, y por último los objetivos generales y específicos que se trabajarán en este documento.

1.1. Contexto

El *E-commerce*, o comercio electrónico en español, consiste en el marketing y venta de productos o servicios a través de Internet [1]. Para esta actividad, la segmentación de clientes es importante, pues permite atender y vender productos adecuados para cada uno de los grupos de clientes. En el presente trabajo se usarán los datos de un *E-commerce* que opera a través de Shopify [2]. El *E-commerce* no cuenta con ningún modelo para la segmentación de clientes, aunque ya cuenta con metodologías enfocadas en los productos para la elaboración de sus promociones.

Como primera aproximación, se probó perfilar a los compradores usando modelos de *clustering* (*K-means* y aglomerativo), sin embargo, como se verá en la sección de Resultados y Discusión, no dieron los resultados esperados. Debido a esto, se desarrollaron las bases para un sistema de recomendación que gestione sugerencias para los productos que los clientes no han consumido aún dentro de la tienda para que sirvan de apoyo para el momento de la compra.

La idea de que los datos pueden arrojar relaciones o interacciones útiles surge después de la lectura del libro "Ciencia de los Datos" por Herbert Jones [3], donde en uno de sus capítulos menciona un caso de éxito donde la compañía Wal-Mart en los años 90's encontró una fuerte correlación entre las cervezas y los pañales, y gracias a las tarjetas de fidelidad que manejaban pudieron encontrar que el rango de edad de quienes compraban estos artículos rondaba entre los 25-35 años, la mayoría hombres, lo que les permitió mejorar su estrategia de ventas para este tipo de cliente. Este mismo caso se encuentra recopilado en una nota de FORBES "*Diaper-beer syndrome*" [4].

Para este trabajo se utilizó como base uno de los reportes que Shopify permite descargar con el detalle de las transacciones hechas del *E-commerce*.

1.2. Justificación

La principal fuente de ingresos que maneja la tienda, estilo tienda de abarrotes en línea, es a través de la venta de productos de consumo; por lo tanto, incentivar la venta de éstos con base en promociones que generen impacto positivo en esa venta, es su principal estrategia de marketing ya que las promociones incentivan la fidelidad de los clientes existentes y la atracción de nuevos clientes.

Se espera que al segmentar a los clientes se puedan encontrar hallazgos relevantes que permitan entender mejor de qué manera estructurar las promociones.
Así mismo el sistema de recomendación podrá hacer que los clientes agreguen más productos a sus órdenes y, esperando que la recomendación sea de su agrado, vuelvan a comprar nuevamente gracias a una experiencia positiva de compra.

1.3. Problema

El principal problema a resolver es la falta de una segmentación de clientes basada en los comportamientos de sus compras en el mismo *E-commerce*. Se espera que esto pueda ayudar a aumentar los márgenes de ganancia, tener o crear promociones efectivas que permitan aumentar el ticket promedio, rotación de inventarios próximos a caducar y principalmente aumentar las ventas.

Como se mencionó en la sección de justificación, actualmente se tienen implementadas estrategias de marketing y promociones, pero se espera poder mejorar e implementar nuevas estrategias mediante un mayor conocimiento de los clientes a los que da servicio el *E-commerce*.

La parte novedosa que aborda la intención de aumentar las ventas es el sistema de recomendación que, después de tiempo, se espera un mayor conocimiento de los clientes según las recomendaciones que se generen.

1.4. Objetivos

1.4.1. Objetivo General:

Crear un sistema que dé recomendaciones especializadas a los clientes con base el clúster en que se encuentren y/o con base en el comportamiento de consumo de todos los clientes de la tienda.

1.4.2. Objetivos Específicos:

1. Limpieza de datos.
2. Tratamiento de datos faltantes.
3. Preparación de datos para los Modelos de *Clustering.*
4. Implementación de los Modelos de *Clustering*.
5. Pruebas o medidas internas para validar.
6. Preparación de datos para el Sistema de Recomendación.
7. Implementación del Sistema de Recomendación.
8. Envío de recomendaciones con encuesta de opinión para validar el Sistema de Recomendación.

2. METODOLOGÍA

Resumen: En este capítulo se presenta en detalle la descripción de los datos trabajados, el proceso de exploración de los mismo, la descripción de los modelos propuestos, la descripción de las métricas usadas en los modelos y descripción de experimentos o simulaciones.

2.1. Descripción de los datos

Los datos provienen de un reporte a nivel detalle por producto descargado a través del sitio del *E-commerce* alojado en Shopify, estos comprenden desde el 1 de enero del 2020 hasta el 31 de diciembre del 2022.

Se anonimizaron los clientes generando un ID único mediante una lista de nombres únicos (*Billing Name*) y se aplicó un reset_index() de la biblioteca Pandas en python.
Posteriormente, se omitió información sensible se eliminando las siguientes columnas:

- Nombres (*Billing Name* y *Shipping Name*).
- Datos de contacto (*Email*, *Billing Phone*, *Shipping Phone* y *Phone*).
- Datos de dirección, ciudad, provincia, estado y código postal (*Billing Street*, *Billing Address1*, *Billing Address2*, *Billing Company*, *Billing City*, *Billing Zip*, *Billing Province*, *Billing Province Name*, *Billing Country*, *Shipping Street*, *Shipping Address1*, *Shipping Address2*, *Shipping Company*, *Shipping City*, *Shipping Zip*, *Shipping Province*, *Shipping Province Name* y *Shipping Country*).

En los datos de dirección se distingue tanto quién paga (*Billing*) el pedido o lo realiza, como el que recibe (*Shipping*) el pedido.

También se eliminaron columnas con información que no es de utilidad para el propósito de este trabajo, dichas columnas son:

- Fecha en que se realiza el pedido (*Created at*).
- Fecha en que se prepara el pedido (*Fulfilled at*).
- Estado del pedido (*Financial Status* y *Fulfillment Status*), ya que se queda únicamente con los pedidos pagados (*Paid*) y preparados (Fulfilled), mismo caso para los productos preparados (*Lineitem fulfillment status - Fulfilled*).
- Divisa (*Currency*), ya que todos los pedidos son en moneda mexicana (MXN).
- Costos de envío (*Shipping* y *Shipping Method*).
- Impuestos, ya que se encuentran vacías (*Taxes*, *Tax 1 Name*, *Tax 1 Value*, *Tax 2 Name*, *Tax 2 Value*, *Tax 3 Name*, *Tax 3 Value*, *Tax 4 Name*, *Tax 4 Value*, *Tax 5 Name* y *Tax 5 Value*).
- Fecha de cancelación (*Cancelled at*).
- Método de pago y/o reembolsos (*Payment Method, Payment Reference* y *Refund Amount*).
- Montos o códigos de descuento (*Discount Amount* y *Discount Code*).
- Notas internas y empleado que llena el pedido (*Tags*, *Notes*, *Note Attributes* y *Employee*).

- Si el cliente acepta mercadotecnia (*Accepts Marketing*).
- Ciertos datos del producto (*Lineitem sku*, *Lineitem discount*, *Lineitem requires shipping*, *Lineitem taxable* y *Lineitem compare at price*).
- Columnas que crea Shopify para funcionamiento interno o vienen vacías (*Oustanding Balance*, *Device ID*, *Id*, *Risk Level*, *Source*, *Receipt Number*, *Duties*, *Payment ID*, *Payment Terms Name*, *Next Payment Due At* y *Payment References*).
- Total y Subtotal de la orden, debido a que se recalcularon sumando los productos entregados.

Como la base de datos viene por orden y un nivel de detalle por producto, el total se muestra únicamente en el primer renglón donde aparezca la orden y los demás renglones, con el mismo ID de orden, muestran un renglón por producto y las columnas referentes a la información de la orden vacías. Otro problema acerca de los datos dentro del reporte es que no todos los productos cuentan SKU (*Stock-keeping unit*) para identificarse, por lo que hay productos que se repitan con un nombre distinto (un ejemplo es “manzana” y “manzan 1 pieza”).

Inicialmente se tenían 297,486 renglones con 79 columnas, siendo el histórico total de órdenes de la tienda hasta el día en que se descargó el reporte. Se encontró que el 98.40% de las órdenes eran pagadas, por lo que eliminar las órdenes canceladas o devueltas no afecta en gran medida al número de datos que tendríamos. En los registros había ítems con estatus *unfulfilled* (sin llenar o incompleto) debido a que estos ya no se encontraban en *stock* y la tienda no envío el producto, por lo que los renglones de estos productos se eliminaron de la base para tener únicamente órdenes y productos completos.

Para el problema de la duplicidad en los nombres de los productos, se categorizó los productos con base al catálogo de categorías que maneja internamente el *E-commerce*, pero otro problema es que sólo contenía los productos actuales, por lo que para productos en órdenes anteriores que ya no estuvieran en el catálogo se procesaron manualmente en Excel para encontrar los nombres más similares e ir categorizando producto por producto.

Después de que los valores faltantes fueran tratados, de categorizar los productos y de que se filtraran los registros que no cumplían con los requisitos anteriores, se terminó con un conjunto de datos de 264,645 renglones y 9 columnas como se muestra en la Tabla 1:

Tabla 1. Columnas del conjunto de datos, sus tipos de dato y breve descripción.

Columna	Tipo de Dato	Descripción
ID Order	Texto	Id de la orden.
Paid at	Fecha	Fecha de pago.
Lineitem quantity	Entero	Cantidad de productos en la orden.
Lineitem name	Texto	Nombre de producto.
Lineitem price	Flotante	Precio del producto.
Location	Texto	Sí la orden fue a domicilio o la persona acudió a la matriz (únicamente 2.35% son en la matriz, los cuales son casos excepcionales).
Vendor	Texto	Distribuidor del producto.
ID Name	Entero	ID anonimizado del cliente.
Category	Texto	Categoría del producto.

2.2. Análisis exploratorio

En la Tabla 2 se muestran, de las columnas de tipo texto, sus valores únicos y valores nulos; se observa que en el periodo comprendido hubo 44,663 clientes distintos, 4,471 productos diferentes vendidos y 16 categorías.

Tabla 2. Columnas del conjunto de datos, valores únicos y valores nulos.

Columna	Valores únicos	Valores Nulos
ID Order	44,663	0
Paid at	15,016	741
Lineitem name	4,471	0
Category	16	0
ID Name	5,869	0

Y en la Tabla 3 se encuentra que no se tienen datos faltantes para el precio y cantidad de los productos; pero un alto sesgo positivo por la curtosis mayor a 3, debido a que la mayoría de las órdenes tienden a tener baja cantidad de piezas por productos o productos de bajo precio unitario, como se observa también en la Figura 1.

Posteriormente, se crea una matriz resumida por la suma del precio de los productos por categoría y se divide a nivel orden para representar la proporción del gasto por categoría, por lo que la se suma todas las categorías de la orden da 100%.

Tabla 3. Columnas del conjunto de datos de tipo numérico y algunos estadísticos.

Columna	Valores Nulos	Media	Desviación Estándar	Mínimo	Máximo	Curtosis
Lineitem quantity	0	1.1318	1.1373	1	72	378.19
Lineitem price	0	27.6320	28.4900	0	1,175	63.60

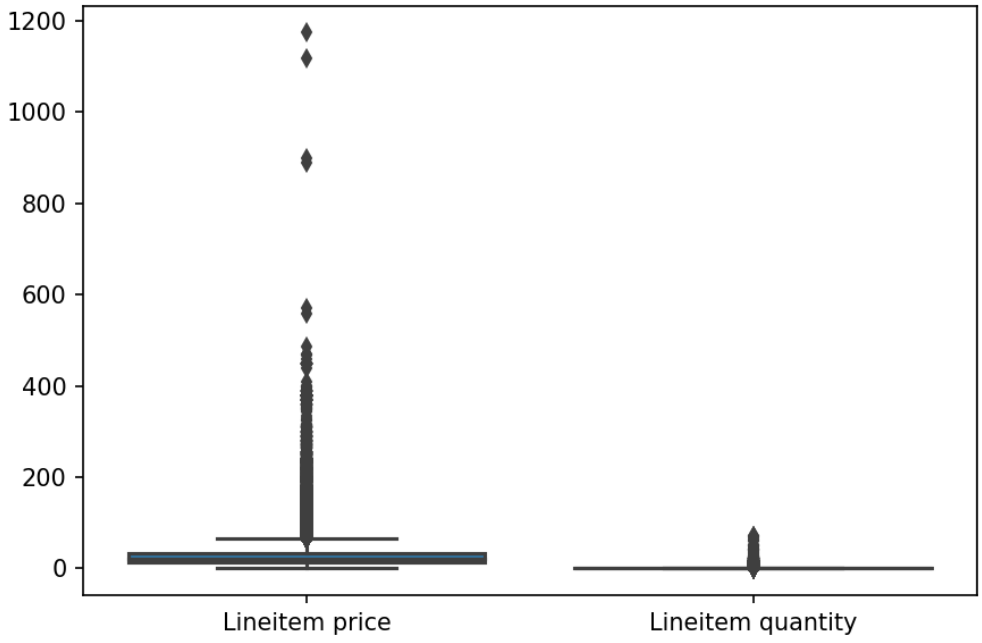

Figura 1. Boxplot de las variables precio y cantidad.

Con base a la matriz anterior se generó una matriz de correlaciones entre las categorías y no se encontraron categorías correlacionadas entre sí, la mayoría está en un rango entre -0.25 y 0.06 como se muestra en la Figura 2.

Y por último se visualiza un diagrama de dispersión, Figura 3, entre el monto de la orden y la cantidad de artículos de la orden. Del diagrama se observa una correlación del 70.67%, lo que indica que a mayor cantidad de artículos tenga la orden mayor será el monto pagado.

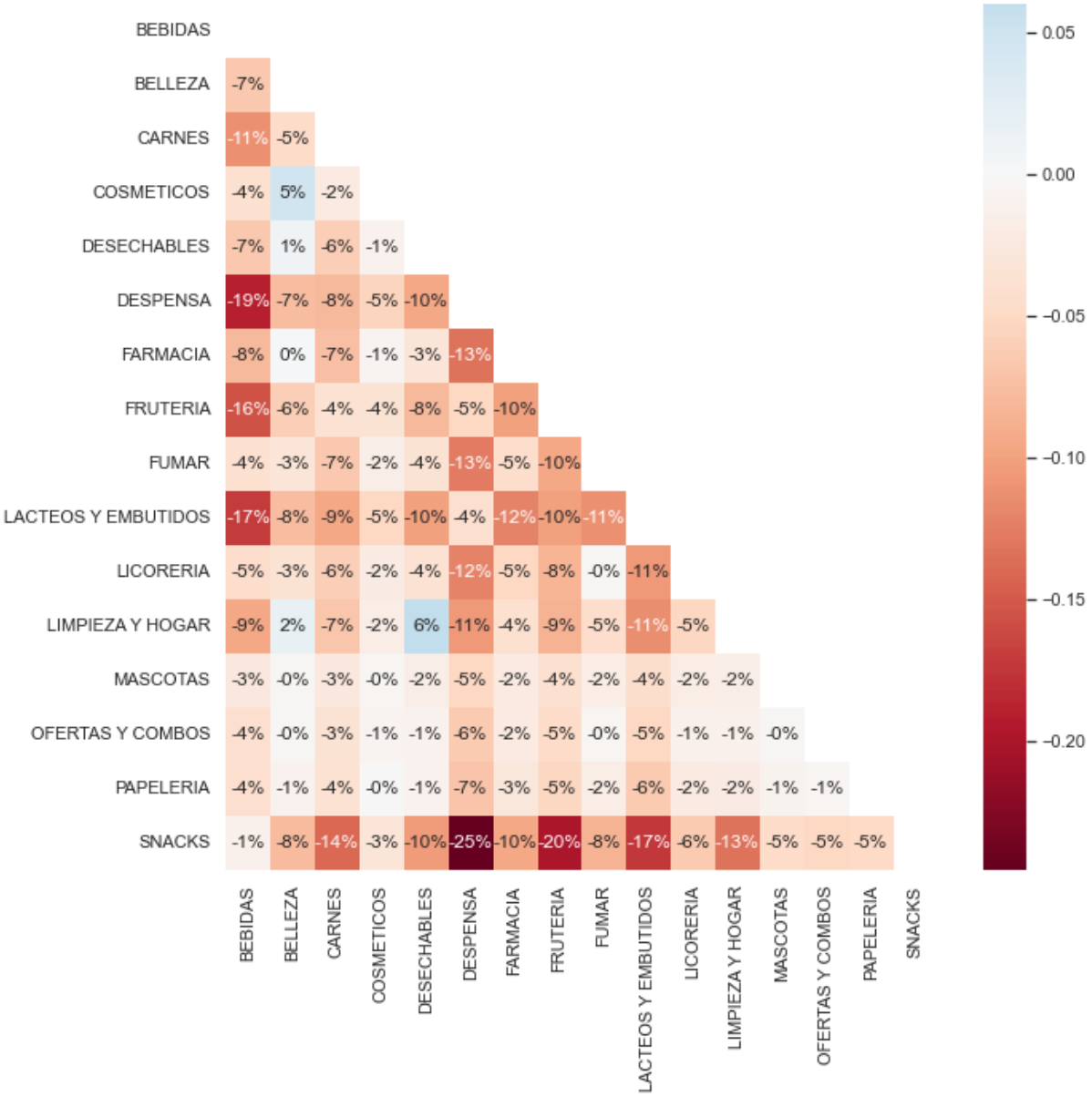

Figura 2. Matriz de correlaciones entre las categorías.

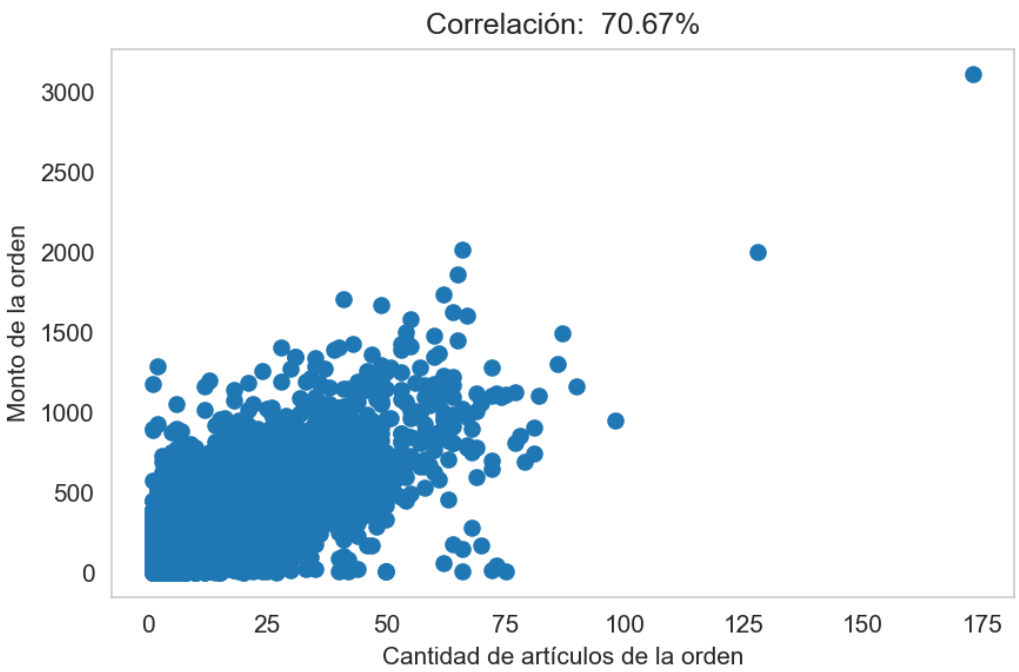

Figura 3. Diagrama de dispersión por monto de la orden y cantidad de artículos de la orden.

2.3. Descripción de los modelos

Los modelos a continuación son de *clustering*, optados por su capacidad para encontrar patrones ocultos dentro de un conjunto de datos, que en aplicaciones reales se han utilizado para segmentar poblaciones en conglomerados desconocidos; por lo cual los siguientes modelos son de gran utilidad para la segmentación de clientes.

Clustering Agglomerative: el *clustering* jerárquico es una familia general de algoritmos de *clustering* que construyen clústeres anidados fusionándolos o dividiéndolos sucesivamente. Esta jerarquía de conglomerados se representa en forma de árbol (o dendrograma). La raíz del árbol es el clúster único que reúne todas las muestras, siendo las hojas los clústeres con una sola muestra.

El objeto *AgglomerativeClustering* de la paquetería de *Scikit-learn* realiza un *clustering* jerárquico utilizando un enfoque ascendente: cada observación comienza en su propio clúster, y los clústeres se fusionan sucesivamente. El criterio de vinculación determina la métrica utilizada para la estrategia de fusión:

- ***Ward*:** minimiza la suma de las diferencias al cuadrado dentro de todos los conglomerados. Es un enfoque de minimización de la varianza y, en este sentido, es similar a la función objetivo de *k-means*, pero abordada con un enfoque jerárquico aglomerativo.

- La **vinculación máxima** o **completa** minimiza la distancia máxima entre observaciones de pares de conglomerados.
- La **vinculación media** minimiza la media de las distancias entre todas las observaciones de pares de conglomerados.
- La **vinculación simple** minimiza la distancia entre las observaciones más cercanas de los pares de conglomerados.

El *AgglomerativeClustering* también puede escalar a un gran número de muestras cuando se utiliza conjuntamente con una matriz de conectividad, pero es costoso computacionalmente cuando no se añaden restricciones de conectividad entre las muestras: considera en cada paso todas las fusiones posibles [5].

K-means: el algoritmo agrupa los datos intentando separar las muestras en n grupos de igual varianza, minimizando un criterio conocido como suma de inercia o suma de cuadrados dentro de un clúster (véase en la ecuación más adelante). Este algoritmo requiere que se especifique el número de conglomerados. Se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

El algoritmo k-means divide un conjunto de **N** muestras **X** en **K** conglomerados disjuntos **C**, cada uno descrito por la media $\boldsymbol{\mu}_j$ de las muestras del conglomerado. Las medias se denominan comúnmente "centroides" de los clústeres; nótese que, en general, no son puntos de **X** aunque vivan en el mismo espacio [6].

El algoritmo *K-means* tiene como objetivo elegir los centroides que minimicen la inercia, o el criterio de suma de cuadrados dentro del clúster:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left| \left| x_i - \mu_j \right| \right|^2 \right) \quad (1)$$

2.4. Descripción de las métricas

2.4.1. Modelos de c*lustering*

Las siguientes métricas fueron útiles para la definición del número de clústeres con los cuales entrenar los modelos, así como la combinación de hiper parámetros que maximice la separación entre grupos y la homogeneidad de los mismos.

Índice de Calinski-Harabasz: también conocido como criterio de relación de varianza, es una medida de similitud entre un objeto y el clúster al que pertenece (cohesión) en comparación con otros clústeres (separación). La cohesión se estima en función de la distancia de los puntos de datos del clúster hasta su centroide de clúster y la separación se basa en la distancia de los centroides del clúster desde el centroide global.

Un valor más alto del índice CH significa que los racimos son densos y están bien separados, aunque no existe un valor de corte «aceptable». Necesitamos elegir aquella solución que dé un pico o al menos un codo abrupto en el gráfico de líneas de los índices CH. Por otro lado, si la línea es suave (horizontal o ascendente o descendente) entonces no hay tal razón para preferir una solución sobre otras [7].

Índice Davies-Bouldin: esquema de evaluación interna en función de la relación entre la dispersión dentro del grupo y la separación entre grupos, un valor más bajo significará que la agrupación es mejor. Resulta ser la similitud promedio entre cada grupo y su más similar, promediada entre todos los grupos. Esto afirma la idea de que ningún grupo tiene que ser similar a otro y, por lo tanto, el mejor esquema de agrupamiento minimiza esencialmente el índice de Davies-Bouldin [8].

Índice de *Silhouette*: método de interpretación y validación de la consistencia dentro de grupos de datos. El valor de la silueta es una medida de cuán similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación).
La técnica de validación de Silhouette calcula el índice de silueta para cada muestra, el índice de silueta promedio para cada grupo y el índice de silueta promedio general para un conjunto de datos. Usando el enfoque, cada grupo podría representarse mediante el índice de silueta, que se basa en la comparación de su estanqueidad y separación. Si el valor del índice de silueta es alto, el objeto se corresponde bien con su propio grupo y no se corresponde con los grupos vecinos [9].

2.4.2. Sistema de recomendación

Se optó por la distancia coseno como medida de similitud ya que, incluso cuando los vectores tienen una distancia euclidiana muy grande, estos podrían tener un ángulo pequeño entre ellos. Cuanto menor es el ángulo, mayor es la similitud entre vectores.

Distancia Coseno: es una medida que se utiliza para calcular la similitud entre dos o más vectores, no es más que el coseno del ángulo entre vectores. Los vectores suelen ser distintos a cero y están dentro de un espacio de producto escalar.

En términos matemáticos, la similitud del coseno se describe como la división entre el producto escalar de los vectores y el producto de las normas euclidianas o la magnitud de cada vector. La similitud entre vectores está dada por la fórmula matemática del producto escalar:

$$\overline{u} \cdot \overline{v} = |u||v|cos\theta \quad (2)$$

Podríamos escribir la ecuación 2 de forma más intuitiva y simple como:

$$\frac{\overline{u} \cdot \overline{v}}{|u||v|} = \text{Similitud} \quad (3)$$

Con la ecuación 3 se puede calcular cómo de similares son dos vectores. Lo único que ha cambiado es que los términos |u||v| se han movido para ser los divisores de la ecuación, mientras que el coseno de θ ha pasado a ser ese término nuevo que estamos introduciendo, que es la similitud.

La parte de arriba de la ecuación representa el producto escalar y la parte de abajo es el producto del módulo de los vectores [10].

Puntaje de la recomendación: Para medir que tan frecuente es la compra de 2 artículos en conjunto se utiliza la siguiente ecuación:

$$\frac{\overline{h}\cdot\overline{s}}{\sum_{i=0}^{n} s_i} \quad (4)$$

Donde:

- $\overline{h}$ representa el vector binario de consumo del cliente ordenado para que los productos coincidan en mismo orden con los del vector $\overline{s}$.
- $\overline{s}$ representa el vector de distancias, del producto analizado con respecto a cada uno de los productos, y ordenado de manera descendente.
- s_i el i-ésimo elemento del vector s.

De esta manera si un producto que el cliente no ha consumido (producto analizado) tiene alta similitud con otros productos consumidos por el cliente, se verá reflejado en un valor más cercano a 1.

2.5. Descripción de los experimentos / simulaciones

2.5.1. Modelos de clustering

Se ajustaron los modelos de *clustering* con el *dataframe* por proporción de compra por orden, cada fila representa una orden y cada columna representa las categorías, así como el total de artículos y monto pagado de la orden. Se utilizaron únicamente los datos del 2022 debido a un problema de los recursos de la memoria para los modelos de *clustering*.
Se iteraron los hiper parámetros de los 2 modelos (Aglomerativo y *K-means*) para encontrar la mejor configuración que optimice las métricas mencionadas en la sección anterior.

El modelo aglomerativo tuvo los siguientes hiper parámetros [12]:

- ***Linkage*****:** *ward*, *complete*, *average*, *single*.
- ***Metric*****:** *euclidean* (también conocida como norma l2), *manhattan* (también conocida como norma l1), *cosine*.

Cabe resaltar, el tipo de unión (*linkage*) *ward* únicamente puede utilizar la métrica de distancia euclidiana.

Cómo se observa en la Figura 4, la mejor configuración resultó con el tipo de unión *ward* y la métrica de distancia euclidiana con 3 clústeres, donde se maximiza la métrica de Silhouette junto con Calinski-Harabasz, y se minimiza la métrica de Davies-Bouldin.

Las otras iteraciones arrojaron valores más bajos para la métrica Calinski-Harabasz que al graficar los clústeres se mostraba una gran agrupación de puntos en 1 sólo clúster y los clústeres restantes tenían 1 o un pequeño grupo de puntos de datos.

El modelo de *K-means* tuvo únicamente el siguiente hiper parámetro:

- ***Algorithm*:**
    - *lloyd*: el objetivo del algoritmo es encontrar centroides para los clústeres de manera que la suma de las distancias cuadradas entre los puntos de datos y sus centroides asignados sea mínima.
    - *elkan*: es una versión más eficiente del algoritmo *K-means* que utiliza desigualdades triangulares para reducir la cantidad de cálculos de distancia. Puede ser más rápido en conjuntos de datos grandes, pero sólo es compatible con la métrica euclidiana para calcular distancias.
    - *full*: utiliza el algoritmo de *Lloyd* con cálculos completos de distancias para asignar puntos a centroides en cada iteración. Es adecuado para conjuntos de datos más pequeños.

En la Figura 5 se muestra una de las iteraciones ya que no hubo diferencias significativas entre los resultados de los 4 tipos de algoritmos, además de que se tomaron 3 clústeres porque la métrica de Davies-Bouldin queda por debajo que con 4 clústeres.

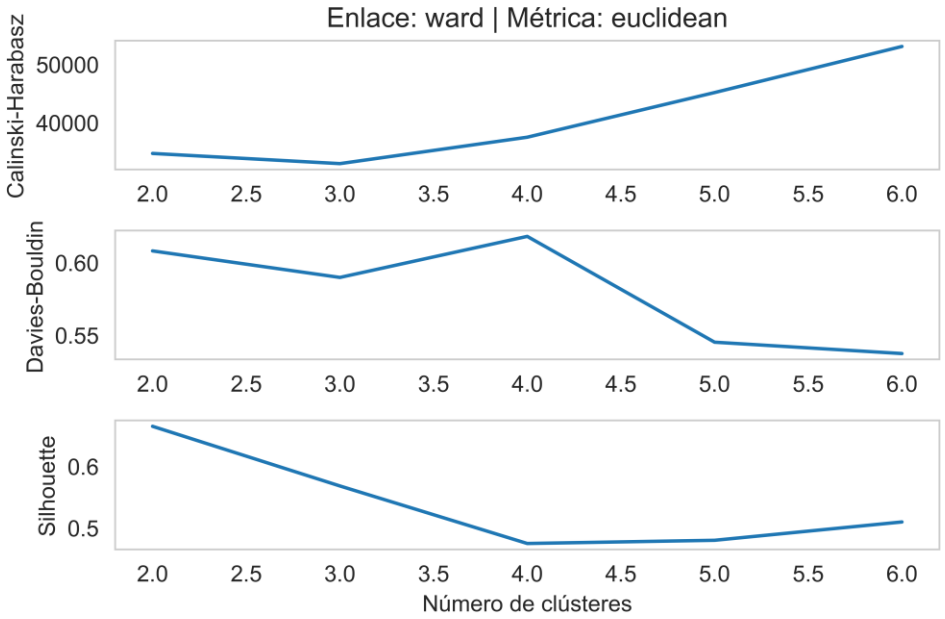

Figura 4. Métricas de evaluación para el modelo aglomerativo con *linkage*: *ward* y *metric*: *eucidean*.

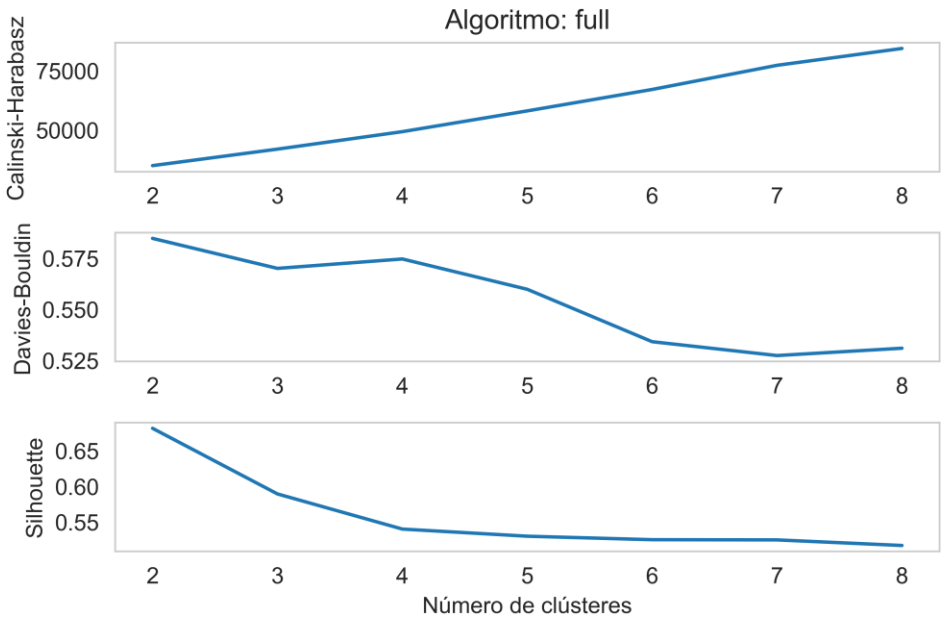

Figura 5. Métricas de evaluación para el modelo *K-means* con *algorithm*: *full*.

2.5.2. Sistema de recomendación

Finalmente, como se explica en la sección de Resultados, dado que los modelos de *clustering* no tuvieron los resultados esperados, se trabajó el sistema de recomendación que se menciona en la sección de Contexto en el cual se utilizó una matriz que refleja la similitud entre todos los productos del catálogo.

Para la creación de dicha matriz se utilizó la distancia coseno para calcular las similitudes entre los productos mediante una tabla con la siguiente estructura:

- **Renglones:** todas las órdenes en el periodo analizado, tamaño n .
- **Columnas:** todos los productos existentes a lo largo de nuestro periodo analizado, tamaño m .
- **Intersecciones o relleno de la tabla:** una codificación binaria que indica si la orden tiene el producto en dicha columna como 1 y si no lo tenía entonces 0, tamaño $n \times m$.

El resultado fue una matriz de tamaño $m \times m$ (los encabezados de renglones y columnas son los productos) con los resultados de la distancia coseno para cada combinación de productos, siendo la intersección entre el mismo artículo el valor de 1, similar a una matriz de correlaciones (ver la sección anterior para más detalles sobre el cómputo de esta variable).

Para la generación de las recomendaciones se codifica el vector de consumo del cliente de tamaño m , así como un vector vacío del mismo tamaño que se llena con el puntaje de la recomendación. Se itera para cada una de las columnas y si el producto ha sido consumido por el cliente (codificado como 1) se le asigna un valor de –1 como puntaje ya que no se

recomiendan productos consumidos, para el resto de los productos se asigna el resultado del puntaje de recomendación.

Se tomaron el top 5 de recomendaciones con base a los puntajes más altos y se enviaron vía mensaje de WhatsApp junto con una liga para que los clientes llenaran una encuesta de opinión sobre dichas recomendaciones, para la cual se tomó una muestra de 40 personas a conveniencia del dueño del *E-commerce*.

Las opciones listadas en la encuesta fueron:

1. Sí tomaría la recomendación.
2. Ya consumo el producto, pero lo compro a través de otros medios.
3. Sí he probado el producto, pero ya no he vuelto a comprar.
4. No me gustó la recomendación.
5. Otros.

Siendo las primeras 3 consideradas recomendaciones exitosas, pero siendo la 2 y la 3 un poco menos eficientes ya que el consumo seguiría siendo fuera del *E-commerce*.
La respuesta 4 sería un desacierto por parte del sistema de recomendación con el cual se pueda validar la eficiencia del mismo, y por último una opción abierta para que el cliente pueda dar otro tipo de retroalimentación en caso de no corresponder las otras opciones.

3. RESULTADOS Y DISCUSIÓN

Resumen: En este capítulo se presentan los resultados obtenidos de la implementación y validación de los modelos de *clustering* y el sistema de recomendación, así como una discusión sobre la utilidad que tendrían finalmente para el impacto del *E-commerce*.

3.1. Resultados

3.1.1. Modelos de clustering

En la Figura 6, semejante a lo que se vio en la Figura 3, se muestran los datos de 2022 a través de las métricas de monto total de la orden y la cantidad de artículos de la orden antes de la agrupación.

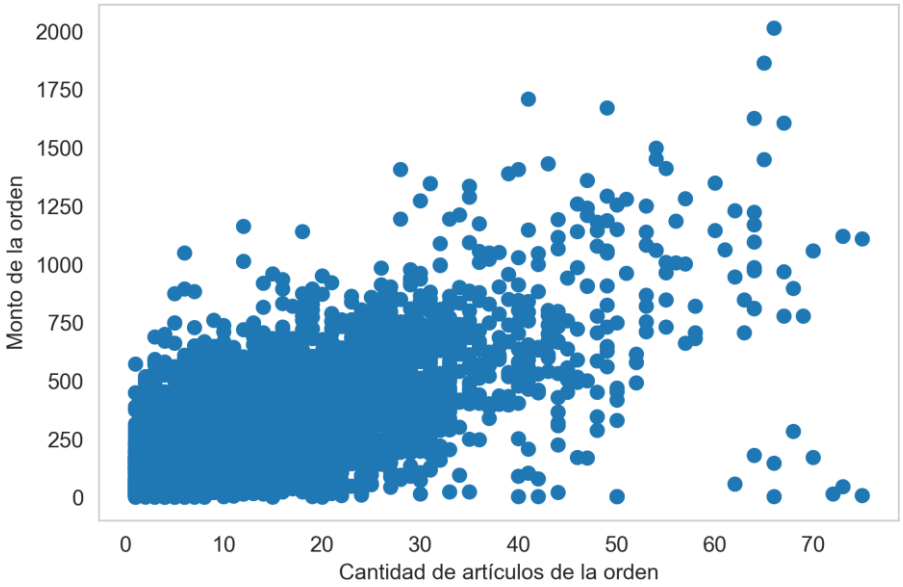

Figura 6. Diagrama de dispersión por monto de la orden y cantidad de artículos de la orden, únicamente órdenes de 2022.

En las Figuras 7 y 8 se muestran los mismos datos con distinción de los clústeres mediante el modelo aglomerativo y *K-means* respectivamente.

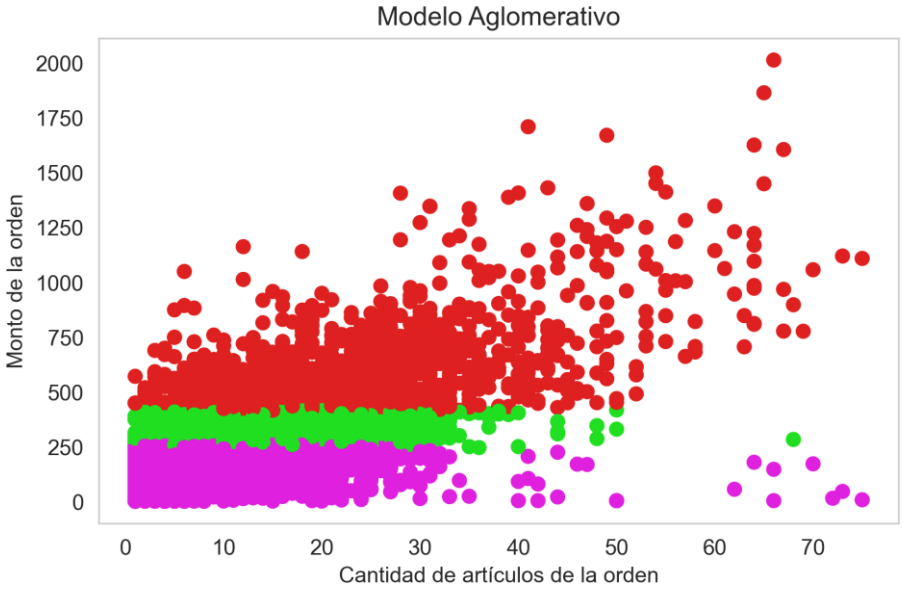

Figura 7. Resultado del modelo aglomerativo con 3 clústeres.

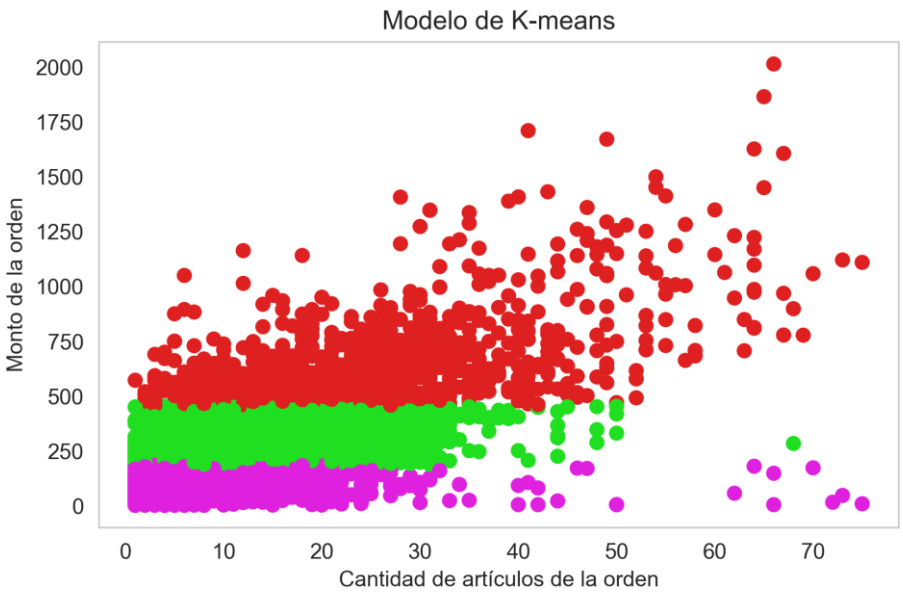

Figura 8. Resultado del modelo *K-means* con 3 clústeres.

3.1.2. Sistema de recomendación

Para el sistema de recomendación se generaron recomendaciones para 39 personas (ya que 1 cliente venía duplicado entre los listados compartido por el dueño del *E-commerce*), pero únicamente 15 llenaron la encuesta de opinión, así que en la Figura 9 se observan por recomendación las respuestas de esos 15 clientes.

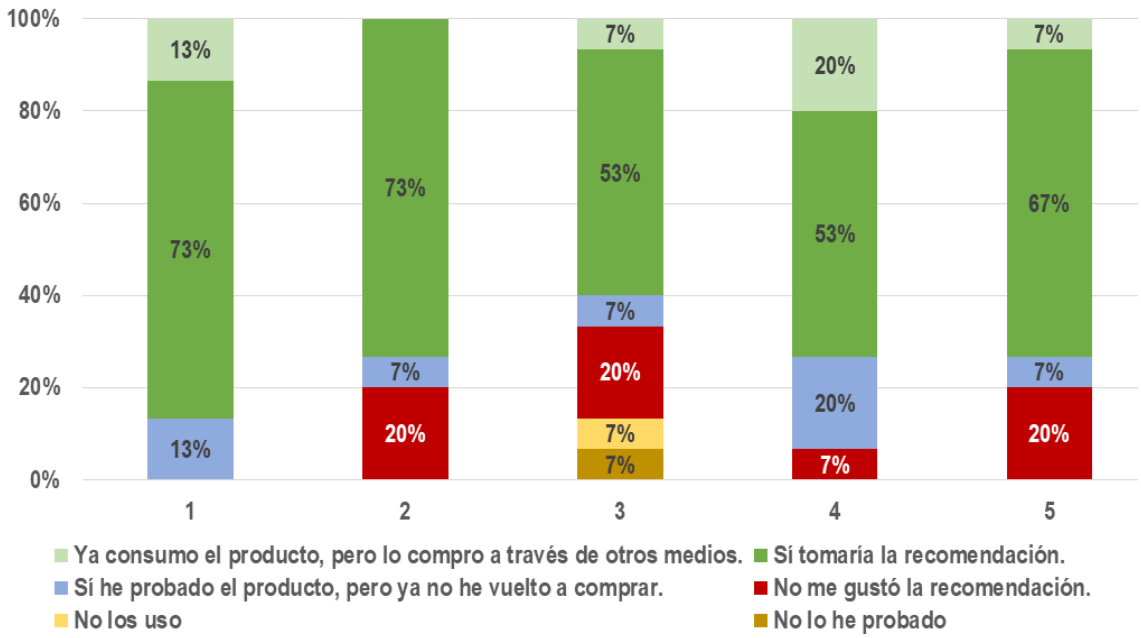

Figura 9. Resultados de la encuesta de opinión para las recomendaciones de los clientes.

Para medir el sistema se le asignó un valor a cada opción de respuesta que tenían los clientes, la codificación es la siguiente:

- **Valor de 1:** sí tomaría la recomendación.
- **Valor de 0.5:** ya consumo el producto o sí he probado el producto.
- **Valor de –1:** no me gustó la recomendación.
- **Neutro o 0:** para las otras respuestas no listadas.

Y, con base a lo anterior, se muestra una tabla en la Figura 10 que resume los resultados de las recomendaciones con la codificación.

Recomendación	Positivo	Neutro	Negativo	Final
1	87%	0%	0%	87%
2	77%	0%	-20%	57%
3	60%	13%	-20%	40%
4	73%	0%	-7%	67%
5	73%	0%	-20%	53%
Total	74%	3%	-13%	61%

Figura 10. Resultados de codificar las opiniones sobre las recomendaciones.

Y en la Figura 11 se encuentra representado las opiniones, pero a nivel categoría del producto recomendado.

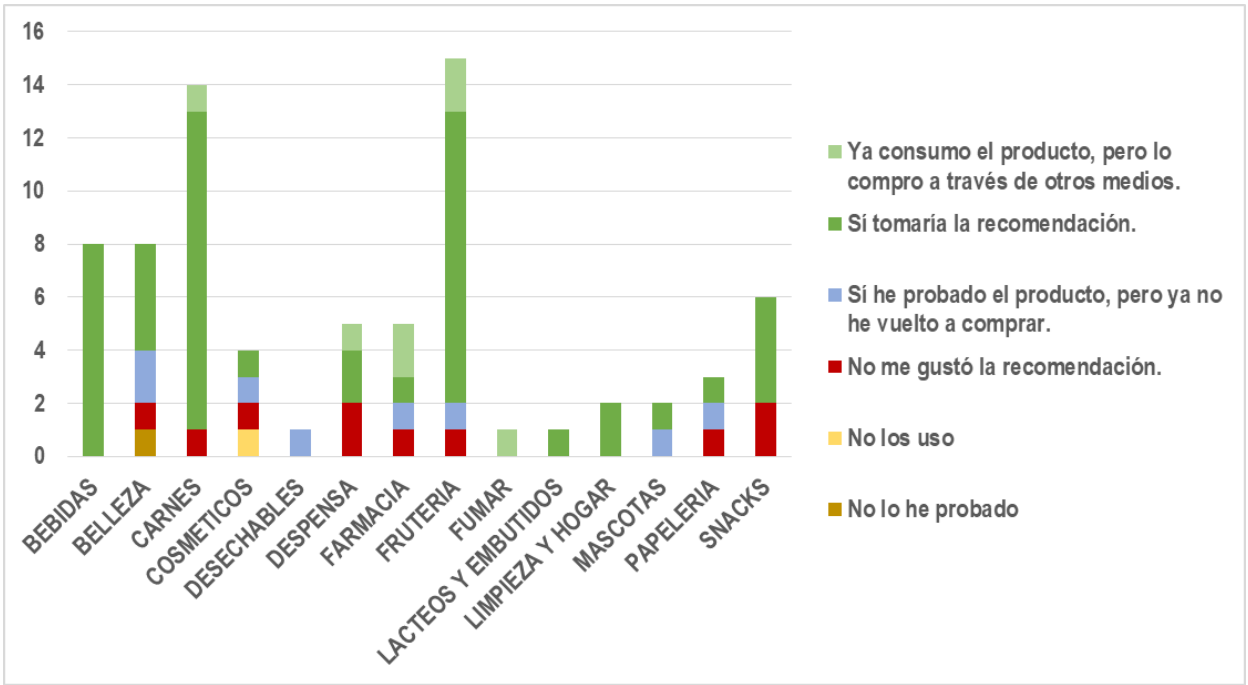

Figura 11. Resultados de la encuesta de opinión a nivel categoría.

3.2. Discusión

Como se observa en las Figuras 7 y 8, el modelo de *clustering* no encontró agrupaciones lo suficientemente homogéneas para distinguirlas entre ellas, lo cual parece indicar que el modelo sólo forzó una separación a través del monto y pareciera seguir una distribución uniforme o como cortes tipo percentiles.

Así vemos, en la Figura 9, que de las recomendaciones dadas la primera resulta ser la mejor recibida, considerando que las recomendaciones se generan en orden; pero contrario a lo que se esperaba, las recomendaciones 4 y 5 aún tuvieron relevancia para los clientes.

Dando un poco más de detalle a las respuestas en color amarillo/ocre, fueron 2 clientes de sexo femenino a las que se les recomendó un tinte (que el cliente posiblemente no se pinte el cabello ya que su respuesta fue “No los uso”) y un desodorante (que el cliente posiblemente no use ese tipo de desodorante al ser de la marca AXE y pudiera ser para hombre, por eso su respuesta “No lo he probado”).

En general, las recomendaciones tuvieron un promedio de 61 % de efectividad, como se muestra en la Figura 10, trabajando para mejorar las recomendaciones, aunque sólo se diera la primera recomendación a un mayor número de clientes se podría confiar que se tiene una tasa cercana al 85 %.

Por último, observamos en la Figura 11 como las categorías que más aparecen en las recomendaciones son las carnes y frutería con también una gran proporción de opinión tipo “Sí tomaría la recomendación”.

Aunque no hay una tendencia clara aún sobre si alguna categoría tuviera mayormente algún tipo de opinión, sí se debiera buscar otros datos que puedan nutrir o sustentar la razón por la que los clientes dieron dichas opiniones.

4. CONCLUSIONES

Resumen: En este capítulo se presentan las conclusiones derivadas de los resultados de este trabajo y los futuros con los que se pueda seguir lo desarrollado en este documento.

4.1. Conclusiones

En conclusión, a los objetivos generales, el modelo de *clustering* para los datos no resultó en una agrupación útil, ya que parece sólo hacer cortes a través del monto de la orden como si de "percentiles" se tratase.
Por otro lado, el sistema de recomendación tuvo resultados positivos reflejados en las encuestas de opinión que llenaron los clientes con los que se probó el sistema, siendo este último una herramienta aplicable al día a día del *E-commerce* y que se debería seguir puliendo para mejorar las recomendaciones, evitando recomendar productos que son para un sexo en específico cuando el cliente no es de ese sexo, productos sensibles que el cliente ha buscado no consumir como pueden ser el caso de vegetarianos o veganos, entre otros.

4.2. Trabajo Futuro

Algo que podría mejorar la propuesta sería la limpieza y tratamiento de los datos como se vio en la sección de Metodología se encontró que no se puede trazar relaciones con los productos directamente ya que se identifican a través de nombres y no un ID único que se mantenga a través del tiempo.
Mismo caso para los clientes que, al momento de hacer sus órdenes, no llenan todos los campos o los llenan con pequeñas variaciones (como acentos, letras repetidas, abreviaciones, mismo correo y teléfono, pero con el nombre de otra persona, etc.).

Analizar que otras métricas se pudieran crear para mejorar el desempeño de los modelos de *clustering*, como lo pueden ser:

- Frecuencia de compra de los clientes.
- Ticket promedio de los últimos 30, 90, 180 días.
- Método RFM (*Recency*, *Frequency* & *Monetary*).

Incluso se podría intentar conseguir una mayor muestra para el sistema de recomendación que, debido a que, por falta de comunicación cercana con los clientes por parte del dueño y falta de tiempo, sólo se consiguieron al menos 40 encuestas.

Y derivado de una mayor muestra, se podría analizar si existe alguna correlación entre las respuestas de los clientes para encontrar si contestaron todas las preguntas con la misma respuesta, o si hay alguna tendencia donde los clientes con mayor cartera de productos probados tengan más respuestas negativas al sistema de recomendación.

BIBLIOGRAFÍA

[1] "¿Qué es un Ecommerce? - Instituto Europeo de Posgrado," Aug. 13, 2018. https://iep.edu.es/que-es-el-ecommerce/https://iep.edu.es/que-es-el-ecommerce/ (accessed Sep. 05, 2023).

[2] Shopify, "¿Qué es Shopify y cómo funciona? Guía completa en español," Shopify, Dec. 22, 2022. https://www.shopify.com/es/blog/tutorial-shopify

[3] H. Jones, Ciencia de los datos: la guía definitiva sobre análisis de datos, minería de datos, almacenamiento de datos, visualización de datos, Big Data para empresas y aprendizaje automático para principiantes. United States: Independently Published, 2019.

[4] "Diaper-beer syndrome," Forbes. https://www.forbes.com/forbes/1998/0406/6107128a.html?sh=6ce13c66260f (accessed Feb. 25, 2023).

[5] "2.3. Clustering — scikit-learn 0.24.1 documentation," Scikit-learn.org. https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering.

[6] "2.3. Clustering — scikit-learn 0.20.3 documentation," Scikit-learn.org, 2010. https://scikit-learn.org/stable/modules/clustering.html#k-means.

[7] "Índice de Calinski-Harabasz: índices de validez de conglomerados | conjunto 3 – Barcelona Geeks," barcelonageeks.com. https://barcelonageeks.com/indice-de-calinski-harabasz-indices-de-validez-de-conglomerados-conjunto-3/ (accessed Sep. 18, 2023).

[8] "Índice de Davies-Bouldin - gaz.wiki," gaz.wiki. https://gaz.wiki/wiki/es/Davies%E2%80%93Bouldin_index#:~:text=El%20%C3%ADndice%20de%20Davies-Bouldin%20%28DBI%29%2C%20introducido%20por%20David (accessed Sep. 18, 2023).

[9] "Índice de silueta: índice de validez de clúster | conjunto 2 – Barcelona Geeks," barcelonageeks.com. https://barcelonageeks.com/silhouette-index-indice-de-validez-de-clusteres-conjunto-2/#:~:text=%C3%8Dndice%20de%20silueta%20%E2%80%93%20El%20an%C3%A1lisis%20de%20silueta (accessed Sep. 18, 2023).

[10] "sklearn.cluster.AgglomerativeClustering," scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[11] R. KeepCoding, “Similitud entre vectores o cosine similarity,” keepcoding.io, Jan. 12, 2023. https://keepcoding.io/blog/similitud-entre-vectores-o-cosine-similarity/ (accessed Nov. 09, 2023).

[12] scikit-learn, “sklearn.cluster.KMeans — scikit-learn 0.21.3 documentation,” Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html