CONOCE A TU CLIENTE

Marketing estratégico basado en datos y machine learning: segmentación de clientes y recomendación personalizada en e-commerce

Nerea López Ziluaga
Data Science Full Time - The Bridge Bilbao

1. Introducción

El panorama del comercio electrónico ha experimentado una transformación radical en las últimas décadas, pasando de ser una alternativa de compra a convertirse en el **canal principal para millones de consumidores en todo el mundo**. Paralelamente, el auge de los servicios de entrega a domicilio, catalizado por cambios en los estilos de vida y eventos globales recientes, ha redefinido las expectativas de conveniencia y accesibilidad. En este entorno altamente competitivo, la capacidad de **entender al cliente a un nivel profundo y ofrecer experiencias personalizadas** se ha vuelto fundamental. Ya no basta con tener un amplio catálogo de productos o precios competitivos; las empresas que sobresalen son aquellas que logran **anticipar las necesidades** de sus usuarios, **facilitar su proceso de compra** y construir una **relación sólida y duradera**.

1.1. El caso Instacart y la importancia de los datos

Instacart es una **plataforma líder en EEUU y Canadá** que conecta a los clientes con sus supermercados locales, ofreciendo compra y entrega de comestibles a domicilio o recogida en tienda. Su modelo de negocio genera una enorme cantidad de datos transaccionales detallados: qué compran los usuarios, con qué frecuencia, en qué momentos del día o semana, de qué tiendas, etc. Este volumen de datos constituye un activo invaluable que, si se analiza correctamente, puede proporcionar **insights accionables** para mejorar la operativa, optimizar el inventario y, crucialmente, personalizar la experiencia del cliente. Ignorar estos datos significa perder oportunidades significativas para aumentar la satisfacción del cliente, mejorar la eficiencia del marketing y, en última instancia, impulsar el crecimiento y la rentabilidad.

1.2. Objetivos del proyecto

Este proyecto se centra en **aprovechar el potencial de los datos de Instacart** para abordar dos aspectos clave del marketing estratégico: la comprensión profunda de la base de clientes y la personalización de la oferta de productos. Para ello, se plantean los siguientes objetivos principales:

- **Realizar una segmentación detallada de la base de clientes:** identificar grupos de usuarios con comportamientos de compra y características demográficas simuladas similares. Esta segmentación permitirá diseñar estrategias de marketing más dirigidas y efectivas, adaptando la comunicación, las promociones y las ofertas a las necesidades específicas de cada grupo.
- **Desarrollar un sistema de recomendación colaborativo:** crear una herramienta que sugiera productos de forma personalizada a cada usuario, basándose en los patrones de compra de usuarios similares. El objetivo es mejorar la experiencia de usuario facilitando el descubrimiento de nuevos productos y aumentando el tamaño del carrito de compra.

La combinación de la segmentación y el sistema de recomendación personalizada representa un enfoque dual poderoso. La segmentación proporciona el "quién" (los diferentes tipos de clientes), mientras que el sistema de recomendación ofrece el "qué" (los productos más relevantes para cada uno). Este enfoque

integrado permite **pasar de campañas de marketing masivas a estrategias hiper-personalizadas, maximizando el retorno de la inversión en marketing y fortaleciendo la relación con el cliente.**

2. Descripción y preparación de datos

Un análisis de datos robusto y un sistema de machine learning efectivo dependen críticamente de la calidad y comprensión de los datos de entrada. Esta sección describe el dataset original de Instacart utilizado en el proyecto, presenta los hallazgos clave del análisis exploratorio de datos (EDA) y detalla el proceso de simulación de datos demográficos para enriquecer el análisis.

El dataset utilizado proviene de un concurso de Kaggle y representa una muestra significativa de las operaciones de Instacart y contiene información detallada sobre el historial de compras de una gran base de usuarios.

El dataset se compone de las siguientes tablas, interconectadas por identificadores únicos:

- **aisles.csv:** contiene la lista de pasillos donde se agrupan los productos.
- **departments.csv: i**ncluye los distintos departamentos a los que pertenecen los productos.
- **products.csv: c**ontiene información sobre todos los productos, incluyendo a qué pasillo y departamento pertenecen.
- **orders.csv: r**egistra los pedidos realizados por los usuarios, incluyendo información temporal y de clasificación en conjuntos (prior, train, test).
- **order_products__prior.csv:** incluye los productos de los pedidos anteriores utilizados para entrenamiento.

2.1. Análisis Exploratorio de Datos (EDA)

El EDA fue una fase crucial para comprender las características subyacentes de los datos y los patrones de comportamiento de compra. El notebook **EDA_Proyecto_ML_Nerea.ipynb** detalla este análisis.

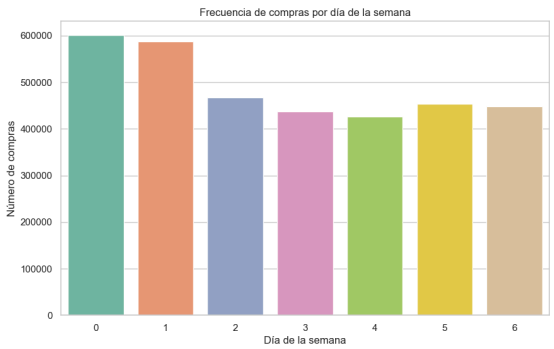
2.1.1. Descripción General

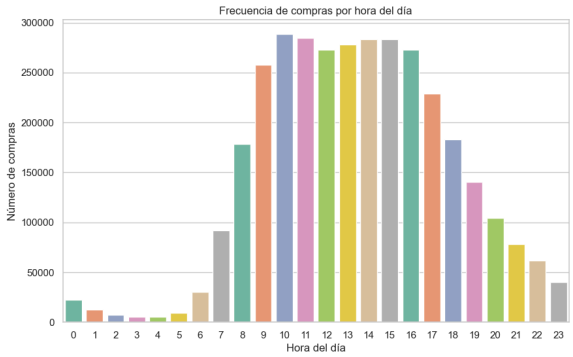
Inicialmente, se realizó una revisión de la estructura de cada tabla, tipos de datos, valores nulos y estadísticas descriptivas básicas (contar de usuarios únicos, productos, pedidos). Esto confirmó la magnitud del dataset y la necesidad de técnicas eficientes para su procesamiento.

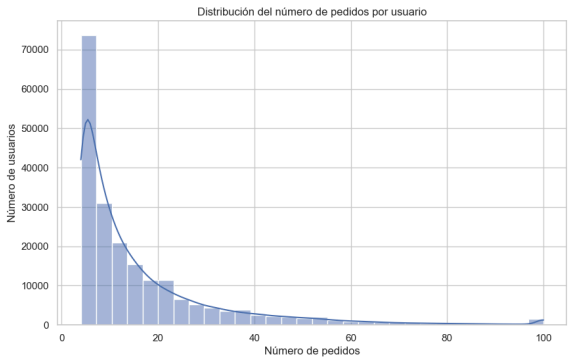
2.1.2. Patrones de Compra

Se analizaron patrones temporales en los pedidos. Se observó una clara estacionalidad semanal y diaria:

- Los **días de la semana** con mayor volumen de pedidos suelen ser los lunes y martes.
- Las **horas del día** de mayor actividad se concentran típicamente entre 10 AM y 4 PM.
- La **frecuencia de pedido** varía significativamente entre usuarios, con una distribución que muestra muchos usuarios realizando pocos pedidos y un grupo menor de usuarios muy activos que realizan pedidos frecuentes (esto se relaciona con la variable days_since_prior_order y la base para el cálculo de Recencia y Frecuencia en RFM).





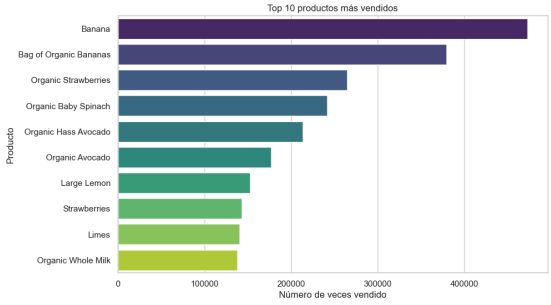


2.1.3. Productos más comprados

Se identificaron los productos, pasillos y departamentos más populares en la plataforma.

- Los **productos** más comprados suelen ser artículos básicos de alta rotación (por ejemplo: plátanos y bolsas de leche orgánica).].

El análisis de reorden (reordered) mostró que una proporción significativa de los productos comprados son ítems que los usuarios ya han comprado anteriormente, lo que subraya la importancia de la fidelización y la conveniencia en las compras de comestibles online.



2.3. Simulación y enriquecimiento con datos demográficos

El dataset original de Instacart carece de información demográfica directa como edad, género, ingresos o ubicación geográfica precisa. Si bien el análisis de comportamiento de compra es muy valioso, la

inclusión de variables demográficas permite realizar segmentaciones más ricas y contextualizadas, así como **diseñar estrategias de marketing más dirigidas**. Dada esta limitación, se optó por simular datos demográficos para cada *user_id* presente en el dataset. Se puede consultar en el notebook **Clustering_trabajado.ipynb.**

2.3.1. Justificación de la simulación

La simulación de datos demográficos se realizó con el propósito de:

- **Enriquecer la segmentación:** permitir la exploración de si los patrones de comportamiento de compra se correlacionan con perfiles demográficos específicos.
- **Crear perfiles de cliente más completos:** humanizar los segmentos identificados, asociándolos a características típicas que los equipos de marketing puedan reconocer y utilizar.
- **Facilitar la definición de estrategias:** adaptar las tácticas de marketing y comunicación a las características de cada segmento.

Es crucial destacar que estos datos son **simulados** y no reflejan la demografía real de los usuarios de Instacart. Por lo tanto, cualquier conclusión basada en estas variables simuladas debe interpretarse con precaución y considerarse como una hipótesis para futuras investigaciones o la aplicación en datasets con información demográfica real.

2.3.2. Variables demográficas simuladas

Se simularon las siguientes variables para cada usuario:

- **Edad:** generada a partir de una distribución normal con una media de aproximadamente 35 años y una desviación estándar de 12 años, limitada a un rango realista (ej: 18 a 75 años). Esto busca reflejar una población adulta con una concentración en edades activas laboralmente.
- **Género:** asignado aleatoriamente con una probabilidad del 50% para masculino y 50% para femenino, para asegurar una muestra balanceada.
- **Renta mensual:** categorizada en tres niveles (Bajo, Medio, Alto). La asignación se realizó con probabilidades que podrían reflejar una distribución de ingresos urbana típica o ajustada para introducir correlaciones plausibles con otras variables simuladas (mayor probabilidad de renta 'Alto' en ciertas zonas simuladas o rangos de edad).
- **Tamaño del hogar:** categorizado (Pequeño, Medio, Grande) o como un número, simulado con una distribución que podría favorecer tamaños medianos para reflejar familias.
- **Estado civil:** categorizado (Soltero/a, Casado/a, Divorciado/a, Viudo/a). Las probabilidades se ajustaron para ser plausibles con las edades simuladas.
- **Zona geográfica :** Categorizada en Urbana, Suburbana, Rural. La asignación se basó en probabilidades o intentos de correlacionarla con otras variables simuladas (mayor probabilidad de hogares grandes en zonas suburbanas).

2.3.3. Proceso de generación y unión de datos simulados

La simulación se implementó utilizando librerías como NumPy y Pandas en Python (como se refleja en el cuaderno **Clustering_trabajado.ipynb**). Para cada *user_id* único en el dataset de *orders*, se generaron

valores para las variables demográficas siguiendo las distribuciones y reglas heurísticas definidas. Se creó un nuevo dataframe con el *user_id* y las columnas demográficas simuladas.

Posteriormente, este dataframe de datos demográficos simulados se unió al dataframe principal que contenía las métricas de comportamiento de compra (que se calcularían posteriormente, como las RFM), utilizando el *user_id* como clave. Esta unión **permitió tener un conjunto de datos enriquecido para el análisis de segmentación**, donde cada usuario estaba caracterizado tanto por su comportamiento de compra histórico como por su perfil demográfico simulado.

3. Segmentación de clientes mediante clustering

La segmentación de clientes divide una base de clientes en grupos distintos y homogéneos, donde los miembros de cada grupo comparten características y comportamientos similares. El objetivo es poder dirigir esfuerzos de marketing específicos y personalizados a cada segmento, aumentando la relevancia de los mensajes, optimizando los recursos y mejorando la satisfacción y fidelidad del cliente. En este proyecto, **se aplicó el algoritmo de clustering no supervisado K-Means** para segmentar a los usuarios de Instacart basándose en una combinación de su comportamiento de compra histórico y las variables demográficas simuladas.

3.1. Feature engineering

Para capturar el comportamiento de compra de los usuarios de manera concisa y significativa, se construyeron las métricas RFM (Recency, Frequency, Monetary). Este es un modelo de segmentación de clientes basado en el análisis del historial de transacciones de un cliente. Las métricas RFM se calcularon a partir del dataset de *orders* y *order_products__prior*.

- **Recency (R):** mide cuántos días han pasado desde el último pedido de un cliente. Un valor bajo indica que el cliente ha comprado recientemente y es más probable que esté activo y receptivo a las comunicaciones. Se calculó identificando el último pedido de cada usuario y determinando los días transcurridos hasta una fecha de referencia.
- **Frequency (F):** mide la cantidad total de pedidos realizados por un cliente. Un valor alto indica un cliente fiel y comprometido. Se calculó contando el número único de pedidos para cada usuario.
- **Monetary (M):** mide el valor monetario total gastado por un cliente. Dado que el dataset de Instacart no incluye precios, se utilizó una aproximación: el número total de productos comprados por un cliente a lo largo de todos sus pedidos. Se calculó sumando el número de ítems en todos los pedidos de un usuario.

Estas tres métricas se calcularon para cada *user_id* , formando una **matriz RFM** donde cada fila representaba un usuario y las columnas sus valores de Recencia, Frecuencia y Monetario.

3.2. Selección del algoritmo de clustering (K-Means)

Se seleccionó el algoritmo K-Means para realizar la segmentación de clientes. K-Means es un algoritmo de clustering particional no supervisado que agrupa los datos en **k clústeres**, donde k es predefinido.

La elección de K-Means se justifica por su:

- **Simplicidad e interpretabilidad:** el concepto de centroides y la asignación a clústeres cercanos es intuitivo.
- **Eficiencia computacional:** es relativamente rápido para grandes datasets comparado con otros algoritmos de clustering, lo cual es importante dado el volumen de datos de Instacart.
- **Escalabilidad:** es adecuado para aplicar a la base completa de usuarios.

Las variables utilizadas como características de entrada para el algoritmo K-Means fueron las métricas RFM (Recencia, Frecuencia, Monetario) calculadas previamente, combinadas con las variables demográficas simuladas (Edad, Género, Renta Mensual, Tamaño del Hogar, Estado Civil, Zona Geográfica).

3.2.1. Preprocesamiento para K-Means

Antes de aplicar K-Means, fue necesario realizar pasos de preprocesamiento sobre las características de usuario para adecuarlas al algoritmo:

- **Variables categóricas (One-Hot Encoding):** las variables demográficas simuladas como Género, Renta mensual, Tamaño del hogar, Estado civil y Zona geográfica son de naturaleza categórica. Los algoritmos basados en distancia como K-Means no pueden trabajar directamente con etiquetas de texto. Por ello, se aplicó la técnica de **One-Hot Encoding** utilizando *pandas.get_dummies*. Esto transformó cada variable categórica en múltiples columnas binarias (0 o 1), donde cada nueva columna representa una categoría única original. Esto permite que las distancias calculadas por K-Means tengan sentido para estas características.
- **Escalado de Variables:** K-Means es sensible a la escala de las características de entrada. Las variables con rangos de valores mayores (como Frecuencia o Monetario, que pueden ser números grandes) podrían dominar el cálculo de distancias sobre variables con rangos menores (como Recencia en días o las variables binarias creadas por *get_dummies*), sesgando así los resultados del clustering. Para evitar esto, se aplicó un **escalado de las características** utilizando *StandardScaler* de *sklearn.preprocessing*.

El clustering se aplicó sobre estos datos escalados (*scaled_data*).

3.4. Determinación del número óptimo de clústeres (k=9)

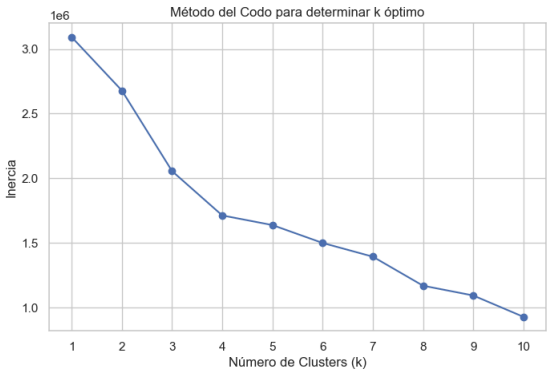
Uno de los desafíos clave al usar K-Means es seleccionar el número apropiado de clústeres (k). Para abordar este desafío y asegurar una segmentación robusta y accionable, se siguió un enfoque metodológico en varias fases, fundamental dado el gran volumen de nuestro dataset de 200.000 usuarios:

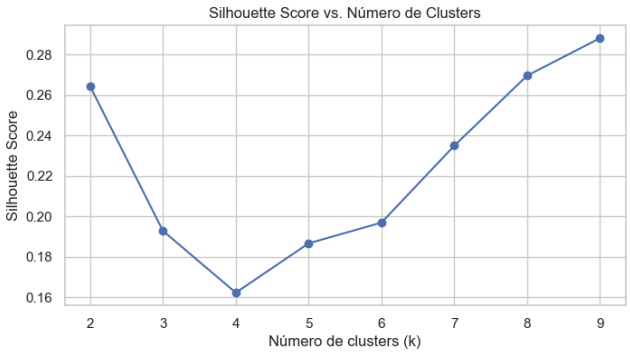
1. **Exploración Inicial (Método del Codo y Silhouette Score):** Primero, se aplicaron métodos cuantitativos como el **método del codo** (analizando la inercia) y el **Silhouette score** sobre los datos, explorando un rango de posibles valores para k. Esto nos dio una **idea preliminar** de dónde podría encontrarse el número óptimo de clústeres, observándose indicios de un buen agrupamiento alrededor de k=8 o k=9. (Ver gráficos abajo)

2. **Refinamiento eficiente con MiniBatchKMeans:** Con esa indicación inicial, se procedió a una fase de **refinamiento y validación más rápida** utilizando **MiniBatchKMeans**. Esta variante del algoritmo es ideal para grandes datasets, ya que aprende de **subconjuntos aleatorios (batches)** de los datos (*batch_size=2048* en nuestro caso), permitiendo probar y evaluar de forma **ágil y eficiente** los valores de k cercanos a nuestra estimación inicial (especialmente k=8 y k=9). Esta fase sirvió para **confirmar la tendencia** observada en el análisis inicial y validar la estabilidad de las soluciones de clustering en un entorno de exploración rápida.

3. **Confirmación final con K-Means estándar:** una vez que k=9 se reveló como el valor más consistente y prometedor a través de las fases exploratorias, se realizó la ejecución final del algoritmo **K-Means estándar sobre la totalidad del dataset** ya preprocesado (con *get_dummies* y *StandardScaler*). Esto aseguró la máxima precisión y representatividad de la segmentación final, basándose en el análisis completo de todos los puntos de datos.

4. **Evaluación cualitativa:** además de las métricas cuantitativas, la selección final de k=9 se basó en la interpretabilidad y utilidad de los clústeres desde una perspectiva de negocio. Al analizar las características promedio de los clústeres para varios valores de k, se determinó que con 9 clústeres se obtenía un equilibrio adecuado: **represetaban perfiles distintivos en términos de RFM y características demográficas simuladas, permitiendo definir estrategias claras y accionables.** Un número menor de clústeres fusionaba perfiles que parecían distintos, mientras que un número mayor resultaba en subdivisiones que no aportaban un valor estratégico adicional significativo.





4. Análisis y perfiles de los clústeres identificados

# Cluster (ID)	Nombre cluster	Edad Media	Renta Mensual Media	Recencia Media	Frecuencia Media	Monetario Medio	Perfil Clave (Resumen)	# Número de Usuarios	Estrategia de Marketing para Instacart	Notas
0	Hogares Urbanos (Hombres)	36.95	2495.92	153.75	11.3	96.34	Hombres Casados, Urbanos. RFM e Ingresos moderados.	25.303	Enfatizar conveniencia y ahorro de tiempo. Promociones en básicos del hogar, packs de "compra semanal". Destacar reorden fácil. Segmento base estable de tamaño considerable que justifica inversión continua.	Segmento base grande y estable
1	VIP Durmientes	44.46	3209.35	316.85	46.22	478.07	Edad media, Ingresos Muy Altos, Hogares Grandes, Casados, Suburbanos. RFM Muy Alto, pero Recencia Muy Alta.	35.253	PRIORIDAD MÁXIMA: Segmento de altísimo valor histórico. Campañas de reactivación premium y personalizadas. Ofrecer descuentos de alto valor en su próxima compra grande. Mostrar novedades y servicios exclusivos. Es uno de los segmentos más grandes y de mayor potencial; asignar recursos significativos.	Aunque inactivos, son los que más gastan y compran con mayor frecuencia históricamente. Reactivarlos es una prioridad máxima.
2	Mujeres Urbanas Casadas Activas	37.03	2497.13	153.1	11.21	95.36	Mujeres Casadas, Urbanas. RFM e Ingresos moderados. (Similar a C0).	27.559	Foco en necesidades familiares y productos frescos. Contenido (recetas, planificación de comidas). Promociones en categorías de frescos, lácteos. Segmento base muy grande e importante por ser tomadoras de decisión; alta prioridad.	Segmento base grande y crucial por su rol de responsable en la compra del hogar.
3	Solteros Urbanos (Hombres)	35.26	2350.37	141.6	10.42	87.44	Hombres Solteros, Urbanos/Suburbanos. RFM e Ingresos ligeramente bajos.	32.261	Promocionar conveniencia para cestas pequeñas/medias. Ofertas en productos de consumo individual (snacks, bebidas, comidas preparadas). Comunicación ágil vía app. Segmento grande con potencial de crecimiento a través de la adaptación a su estilo de vida.	Otro segmento muy grande con hábitos similares a C8.
4	Clientes Esenciales Rurales	53.74	1557.04	97.14	5.68	73.29	Edad Mayor, Ingresos Bajos, Hogar muy pequeño, Rurales. Recencia Baja, RFM Bajo. Alta prop. Viudos/Solteros.	7.207	Enfatizar fiabilidad de la entrega en zonas rurales y simplicidad de la app/web. Foco en básicos y productos esenciales a buen precio. Segmento de tamaño moderado, requiere una estrategia adaptada a su ubicación y perfil de bajo gasto.	Segmento base muy grande e importante que se responsabiliza de las compras del hogar
5	Veteranos con Valor	68.10	1455.71	91.05	5.16	57.98	Edad Más Alta, Ingresos Más Bajos, Hogar más pequeño. Predominantemente Viudos. Rurales/Suburbanos. RFM Mínimo, Recencia Baja.	5.446	Segmento de menor tamaño pero con necesidades muy específicas de experiencia de usuario y accesibilidad. Simplificar la compra al máximo. Ofertas en formatos individuales. Importante para la inclusión y servicio a un grupo vulnerable, aunque de bajo valor transaccional. Requiere una estrategia focalizada pero no masiva.	Un segmento más pequeño, pero importante para asegurar que la plataforma sea accesible y ofrezca un servicio adecuado a usuarios de mayor edad con necesidades de simplicidad.
6	Hogares Suburbanos Balanceados	38.37	2474.01	152.05	11.34	97.55	Balanceado Género, Casados, Suburbanos. RFM e Ingresos moderados. (Similar a C0/C2, diferenciado por zona/género).	37.081	Similar a C0/C2 pero adaptado a vida suburbana: promos en productos de volumen, artículos para el hogar/jardín. Destacar comodidad para familias. El segmento más grande; alta prioridad estratégica y justificación de inversión significativa.	El segmento más grande. Clave para el volumen general del negocio.
7	Compradores de Calidad Maduros	64.59	2007.70	122.34	7.68	133.75	Edad Mayor, Divorciados/as, Hogar pequeño, Rurales/Suburbanos. Baja Frecuencia, Alto Gasto Monetario.	1.608	Foco en productos premium, especialidades, campañas de calidad, apelando al disfrute. Alto potencial de margen por compra. Segmento nicho de alto valor unitario; justifica estrategias de marketing selectivas y de alta calidad.	Aunque pequeño, este nicho gasta más por pedido, ofreciendo oportunidades para productos de mayor margen.
8	Solteras Urbanas (Mujeres)	35.23	2356.24	140.74	10.39	87.51	Mujeres Solteras, Urbanos/Suburbanos. RFM e Ingresos ligeramente bajos. (Similar a C3, diferenciado por género).	34.491	Similar a C3 pero adaptado: foco en productos saludables, opciones rápidas/sanas, belleza, redes sociales. Segmento muy grande con potencial similar a C3.	Un segmento muy grande con hábitos de compra ágiles.

4.1. Conclusión de la Segmentación

La segmentación en 9 clústeres ha proporcionado una **visión granular y accionable** de la base de usuarios de Instacart. Al comprender las características demográficas simuladas, los patrones de compra RFM y el tamaño de cada segmento, Instacart puede pasar de un enfoque de marketing masivo a uno más segmentado y personalizado, mejorando la relevancia para el cliente y la eficiencia de las campañas.

5. Sistema de recomendación colaborativo

Además de entender los segmentos de clientes, es crucial ofrecer una experiencia personalizada a nivel individual. Un sistema de recomendación efectivo logra esto sugiriendo productos que son relevantes y atractivos para cada usuario. Esta sección describe el desarrollo de un sistema de recomendación colaborativo basado en K-Nearest Neighbors (KNN) con una capa de visualización. Ver notebook **Sistema_recomendacion_colaborativo_v2**.

5.1. Introducción a los sistemas de recomendación

Los sistemas de recomendación son herramientas esenciales en plataformas de e-commerce para ayudar a los usuarios a descubrir productos que probablemente les interesen. Esto mejora la experiencia de compra, aumenta el tiempo de permanencia en la plataforma, incrementa las ventas (tamaño del carrito y frecuencia de compra) y fomenta la fidelidad del cliente. Existen varios enfoques para construir sistemas de recomendación, incluyendo:

- **Basados en Contenido:** Recomiendan ítems similares a los que el usuario ha consumido o disfrutado previamente, basándose en las características de los ítems.
- **Filtrado Colaborativo:** Recomiendan ítems basándose en las preferencias de usuarios similares (filtrado colaborativo basado en usuario) o en ítems que son frecuentemente consumidos juntos (filtrado colaborativo basado en ítem).

Este proyecto se centra en el filtrado colaborativo basado en usuario debido a la naturaleza transaccional de los datos de Instacart, donde los patrones de compra compartidos son un fuerte indicador de preferencias similares.

5.2. Metodología: filtrado colaborativo basado en KNN

El sistema de recomendación desarrollado utiliza el enfoque de filtrado colaborativo basado en usuario, implementado con el algoritmo K-Nearest Neighbors (KNN). La idea central es que si dos usuarios tienen comportamientos de compra similares en el pasado, es probable que tengan preferencias similares en el futuro.

5.2.1. Creación de la Matriz usuario-producto

El primer paso fue **transformar los datos transaccionales en una representación matricial donde las filas son usuarios y las columnas son productos**. Cada celda de la matriz (i,j) representa la interacción

del usuario *i* con el producto *j.* En este caso, se utilizó una matriz dispersa (sparse matrix) para representar la información, ya que la mayoría de los usuarios solo han comprado un pequeño subconjunto del catálogo total de productos, resultando en una matriz con muchos valores cero. El valor en la celda es binario (1 si el usuario compró el producto, 0 si no).

La **matriz usuario-producto** fue construida a partir de las tablas orders, order_products__prior y *products*, uniendo la información para mapear cada compra al usuario y producto correspondiente.

5.2.2. Métrica de similitud (Similitud del Coseno)

Para **encontrar usuarios con patrones de compra similares**, se necesita una métrica para cuantificar la "distancia" o "similitud" entre los vectores de usuario en la matriz usuario-producto. Se seleccionó la similitud del coseno. La similitud del coseno mide el coseno del ángulo entre dos vectores no nulos en un espacio multidimensional. Es una métrica adecuada para datos dispersos como nuestra matriz usuario-producto, ya que se enfoca en la dirección de los vectores (es decir, qué ítems fueron comprados juntos o por usuarios similares) más que en su magnitud. **Un valor de similitud del coseno cercano a 1 indica alta similitud, mientras que un valor cercano a 0 o negativo indica baja similitud**.

5.2.3. Identificación de Vecinos Cercanos (KNN)

Una vez calculada la matriz de similitud entre pares de usuarios, el algoritmo KNN para filtrado colaborativo identifica los k usuarios más similares ("vecinos") al usuario objetivo. En este proyecto, **se definió k=10, buscando los 10 usuarios con la similitud del coseno más alta con el usuario** para el cual se generan las recomendaciones.

5.3. Proceso de Generación de Recomendaciones

Una vez identificados los 10 vecinos más cercanos para un usuario objetivo (por ejemplo, *user_ID* = 24877), el proceso para generar las recomendaciones sigue estos pasos:

1. **Recopilar todos los productos** comprados por los 10 vecinos más cercanos.
2. Filtrar los **productos** que el usuario objetivo **ya ha comprado** en su historial, ya que no tiene sentido recomendarle algo que ya conoce y compra.
3. **Contar la frecuencia** con la que cada producto restante aparece en el historial de compras de los 10 vecinos. Actúa como un indicador de relevancia del producto entre usuarios similares.
4. **Ordenar los productos** restantes de forma descendente según su frecuencia de aparición entre los vecinos.
5. **Seleccionar los N productos mejor clasificados** (en este caso, N=10 productos) como las recomendaciones para el usuario objetivo.

5.4. Implementación de la Visualización

Para mejorar la experiencia del usuario, las recomendaciones generadas no se presentaron simplemente como una lista de nombres de productos. Se desarrolló una capa visual atractiva que imita la presentación de productos en una tienda online.

Dado que el dataset de Instacart no incluye imágenes de productos, se optó por asociar una imagen representativa a cada **categoría** de producto. Esto se realizó mapeando los *department_id* a URLs de imágenes que representaran la categoría.

6. Limitaciones del estudio

- **Datos Demográficos Simulados:** Si bien permiten explorar la relación entre demografía y comportamiento de compra y enriquecer la segmentación, las conclusiones sobre los perfiles demográficos de los clústeres son hipotéticas y deben validarse con datos reales si estuvieran disponibles.
- **Ausencia de Valor Monetario Real:** La métrica "Monetario" se aproximó usando el número de productos. La disponibilidad de datos de precios permitiría calcular el valor monetario real, lo que refinaría el análisis RFM y la definición de clústeres basados en el gasto real.
- **Evaluación Limitada del Sistema de Recomendación:** La falta de una evaluación cuantitativa rigurosa limita la capacidad de comparar este sistema con otros enfoques o de medir su impacto real en métricas de negocio (ej: aumento del tamaño del carrito, tasa de conversión).
- **Modelo de Recomendación Simplificado:** El modelo KNN, si bien efectivo, es un punto de partida. Otros algoritmos de filtrado colaborativo (ej: Factorización de Matrices como SVD) o enfoques híbridos podrían potencialmente ofrecer mejores recomendaciones.

7. Conclusiones

- Se realizó una **segmentación de clientes efectiva** en 9 grupos distintivos, basada en métricas RFM y perfiles demográficos simulados. Esta segmentación proporciona a Instacart un mapa de su base de clientes, permitiendo una comprensión más profunda de sus diversos segmentos.
- Se desarrolló un **sistema de recomendación colaborativo basado en KNN** capaz de generar sugerencias de productos personalizadas para cada usuario.
- Se demostró cómo la **combinación de segmentación y recomendación** puede potenciar las estrategias de marketing, permitiendo enfoques más dirigidos y personalizados.

Este trabajo sirve como prueba de concepto y prototipo, mostrando el valor potencial de aplicar técnicas de análisis de datos y machine learning para optimizar las operaciones de marketing en una plataforma de e-commerce como Instacart.