PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN SISTEMA DE INFORMACIÓN,
MENCIÓN DATA SCIENCE

TEMA

"Creación de segmentos de clientes, basados en datos de órdenes de un e-commerce."

AUTOR: Henry Fabricio Caraguay Ordoñez, Ing.

DIRECTOR: Eduardo José Montero Bermúdez, Ing., MSc.

Loja - 2025

DEDICATORIA

A mi familia, por su constante apoyo a lo largo de todos estos años de formación, en especial a mis padres que estuvieron presentes y supieron darme su guía y consejos de forma continua cuando más lo necesitaba, a ellos que siguen guiando mis pasos desde el cielo, cuyas enseñanzas nunca olvidaré y su ejemplo seguirá vivo siempre en mí. A mis hermanos, por ser mi motivación a seguir. Y a mis amigos, que han sabido acompañarme en mis momentos de alegría y también en los de tristeza. Por ello, este logro no es solo mío, sino de todos nosotros.

AGRADECIMIENTO

Deseo en estas breves líneas expresar mi más profundo y sincero agradecimiento a todas las personas que me permitieron culminar exitosamente este grado de maestría. Primeramente, a mi familia, que son quienes han estado siempre presentes. A mis profesores y asesores, que me han brindado su conocimiento y consejos cuando lo he necesitado. A mis compañeros, con quienes compartí este viaje, y han sido una fuente de ayuda invaluable. Y a todas las personas que me han brindado su ayuda, y de alguna manera han contribuido a mi crecimiento tanto personal como profesional. Estoy, y les estaré eternamente agradecido.

Con gratitud,

Henry Caraguay O.

Resumen

El presente trabajo de titulación pretende analizar los datos de órdenes de un *e-commerce* con el fin de establecer segmentos de clientes, creando perfiles de clientes para cada segmento identificado, basados en características demográficas, comportamentales, uso de dispositivos y preferencias de productos.

Para lograrlo se planea utilizar algoritmos de análisis y clusterización de datos como *DBScan* y/o *K-means*.

Palabras clave: clientes, segmentación, análisis técnico, perfil, clusterización

Abstract

This thesis aims to analyze the order data of an e-commerce site in order to establish customer segments, creating customer profiles for each identified segment, based on demographic and behavioral characteristics, device usage, and product preferences.

To achieve this, it is planned to use data analysis and clustering algorithms such as DBScan and/or K-means.

Keywords: customers, segmentation, technical analysis, profile, clustering.

Capítulo 1

Introducción

El comercio electrónico ha experimentado un crecimiento sostenido en los últimos años en Latinoamérica, y Ecuador no ha sido la excepción. Con una economía evolucionando a la digitalización, el país ha visto un aumento en la adopción de plataformas de comercio electrónico y soluciones de pago en línea. La facilidad de acceso a Internet, la proliferación de dispositivos móviles y el uso creciente de redes sociales como canales de marketing han impulsado este fenómeno. De acuerdo con datos recientes, se proyecta que el sector del e-commerce en Ecuador cierre el año 2024 con un crecimiento del 14%, reflejando un aumento significativo en la confianza de los consumidores en las compras digitales.

En este contexto, la presente investigación se centra en el análisis de los clientes de una plataforma de comercio electrónico especializada principalmente en el sector de la comida. A diferencia de mercados más diversificados que incluyen tecnología y ropa, el segmento de alimentos presenta un interés particular debido a su volumen de transacciones y la frecuencia con la que los clientes realizan compras. Esta dinámica proporciona una gran cantidad de datos que pueden ser utilizados para optimizar estrategias de segmentación y personalización de la experiencia del usuario.

Las grandes plataformas de *e-commerce* a nivel mundial, como Amazon y eBay, han demostrado la importancia del análisis de datos para potenciar sus estrategias de marketing y retención de clientes. Inspirándose en estas tendencias de analítica de datos, esta investigación busca aplicar herramientas de análisis de datos para identificar patrones de comportamiento de los consumidores ecuatorianos en el sector de la comida.

El análisis de segmentación de clientes no solo permitirá una mejor comprensión del perfil de los consumidores, sino que también brindará información valiosa para diseñar estrategias de marketing más eficaces y dirigidas. Al identificar diferentes segmentos de clientes con características y necesidades específicas, las empresas podrán personalizar sus ofertas, mejorar la retención y aumentar la conversión de ventas.

Este estudio cobra relevancia en un momento en el que la competitividad en el sector digital es cada vez mayor y las plataformas de comercio electrónico buscan diferenciarse a través de la optimización de sus estrategias de negocio. En este sentido, los resultados de esta investigación podrán servir como una guía para la toma de decisiones de empresas ecuatorianas que deseen fortalecer su presencia en el mercado digital y mejorar la experiencia de compra de sus clientes.

1.1 Contextualización del tema u objeto

En este caso se abordó el análisis de los clientes de una plataforma de *e-commerce* que trabaja con el sector de comida, ropa, tecnología ya que esta plataforma ofrece servicios de pagos en línea a varias empresas dedicadas al comercio en estos rubros. Se tiene, entonces, datos de dos sectores principalmente, el sector de comida y el sector de retail (que engloba ropa, tecnología y artículos varios).

Este trabajo de investigación surge al notar que la mayoría de plataformas *e-commerce* gigantes (*Ebay*, *Amazon*, etc) utilizan los datos de sus usuarios para promover el uso de sus plataformas de forma regular, ayudando a tomar decisiones estratégicas para rentabilizar sus recursos. (Salcedo et al. , 2019).

Para abordar la segmentación de clientes, se utilizó características demográficas de los clientes (provincia/estado y ciudad de origen), patrones de compra, frecuencia de compra, y también de uso de dispositivos (pc, teléfono, tablet, etc). Utilizando herramientas de análisis de datos, la plataforma *e-commerce* buscó mejorar la segmentación de sus clientes para optimizar las campañas de marketing y aumentar la lealtad de sus clientes.

1.2 Justificación

Ecuador proyectó un cierre del año 2024 con un incremento del 14% tanto en *e-commerce* como en pagos en línea. Siendo las redes sociales como *Facebook* e *Instagram* los principales medios de publicidad y redirección hacia los *e-commerce*. Entre los más visitados tenemos MercadoLibre Ecuador, Fybeca, Deprati y Novicompu. (El Universo, 2024).

En base a estos datos, podemos ver que hay oportunidad de un estudio de segmentación de clientes para determinar las personas que más compran en la plataforma de comercio electrónico, clasificadas por edad, ciudad, canal, dispositivo, tratando de preservar su anonimato, ya que con estos análisis los comercios pueden depurar sus estrategias de marketing y aumentar sus transacciones.

1.3 Planteamiento del problema

El negocio del *e-commerce* presenta desafíos continuamente, desde la retención de clientes, a la fidelización de los mismos, se estima que la tasa de conversión de un *e-commerce* no alcanza el 3% de la tasa de tráfico total de visitantes. (Moro et al., 2020). Por ende se vuelve imperativo buscar nuevas metodologías y procesos que nos permitan mejorar estas cifras, en aras de aumentar la rentabilidad de la plataforma *e-commerce*, uno de estas metodologías puede ser en efecto establecer segmentos de clientes.

1.4 Objetivos

1.4.1 Objetivo General

Realizar una segmentación de clientes sobre los datos de órdenes del *e-commerce.*

1.4.2 Objetivos Específicos

- Analizar los datos demográficos, comportamentales y transaccionales de los clientes actuales del *e-commerce*.
- Desarrollar perfiles de clientes para cada segmento identificado, incluyendo características demográficas, comportamentales y preferencias de productos.
- Proponer estrategias de marketing personalizadas para cada segmento de clientes identificado, con el fin de aumentar la conversión y la retención.

Capítulo 2

Marco teórico

Este capítulo establece el marco teórico y contextual de la investigación. Se abordarán conceptos esenciales en aprendizaje automático y *e-commerce*.

2.1 Ciencia de Datos

La ciencia de datos es el estudio de datos con el fin de extraer información significativa para una empresa. Es un enfoque que combina muchas disciplinas, especialmente con principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de la computación para el análisis de un volumen grande de datos. Permite clasificar, predecir y analizar una gran cantidad de datos. (*Amazon Web Services*, s.f.).

2.2 Inteligencia Artificial

Es una rama de las ciencias computacionales preocupada por la automatización de la conducta inteligente, siendo una de las ramas que más interés ha despertado en la actualidad, debido a su enorme campo de aplicación. Al igual que la inteligencia natural no tiene un concepto definido, de igual forma la inteligencia artificial tiene un amplio rango de conceptos, pero podemos resumir que es la ciencia encargada de tratar que las máquinas sean capaces de percibir, razonar y actuar. (Ponce et al. , 2014)

2.3 Machine Learning

El *machine learning* o aprendizaje automático, es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del ambiente circundante. Técnicas basadas en el machine learning han sido aplicadas exitosamente en muchos campos desde el reconocimiento de patrones, visión por computadora, ingeniería aeroespacial, finanzas, entretenimiento y biología computacional. (Rebala et al., 2019)

2.3.1 Principales paradigmas del machine learning

El *machine learning* se clasifica en cinco tipos: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo, *deep learning* y *deep learning* por refuerzo. Se examinará los dos primeros más a detalle. Con respecto al aprendizaje por refuerzo, es aquel donde un agente aprende

a comportarse en un ambiente, realizando acciones y viendo los resultados. El *deep learning*, es el tipo de aprendizaje automático, inspirado por la estructura del cerebro humano, que involucra el uso de redes neuronales que selecciona *features* sin intervención humana.

2.3.2 Tipos de aprendizaje en *machine learning*

2.3.2.1 Aprendizaje Supervisado

En este tipo de aprendizaje, la máquina es entrenada usando data etiquetada. La data etiquetada actúa como un supervisor, y tanto entradas como salidas son provistos al modelo como fuente de aprendizaje. Después, el modelo predecirá un nuevo resultado para un nuevo par de datos. Hay dos tipos de aprendizaje supervisado: clasificación y regresión. Los algoritmos de regresión se usan principalmente cuando se quiere modelar o predecir una variable continua a partir de una o más variables independientes. Ejemplos de aplicación de regresión, los vemos en predicción de variables continuas como predicción de clima, o tendencias de mercado.

Algoritmos de clasificación son usados cuando la variable de salida es categórica, lo que significa que hay más de una clase (puede ser binario o multiclase). Ejemplos de aplicación de clasificación, los vemos en reconocimiento de voz, reconocimiento de escritura a mano, clasificación de documentos, entre otros.

2.3.2.2 Aprendizaje No Supervisado

En el aprendizaje no supervisado, la máquina es entrenada con datos no etiquetados sin ninguna guía. A diferencia del aprendizaje supervisado, aquí no es necesario ninguna clase, ya que no le fueron dadas etiquetas al algoritmo de aprendizaje, este encuentra estructura en la data de entrada por su cuenta.

Este aprendizaje puede ser categorizado en dos tipos: clusterización y asociación. La clusterización es el método de agrupar objetos en clusters tal que los objetos con mayores similitudes permanezcan en un solo grupo, y otros con menos o no similitudes permanezcan en otros grupos. Las aplicaciones de clusterización incluyen segmentación de mercado, análisis de datos estadísticos, análisis de redes sociales, segmentación de imágenes y detección de anomalías.

La asociación es usada para encontrar relaciones entre las variables de un dataset muy grande. Entre las aplicaciones de asociación están el diagnóstico médico, secuenciación de proteínas, diseño de catálogos entre otros. (Shyam et al., 2021).

2.4 Conceptos de marketing y ventas

2.4.1 Segmentación de clientes

La segmentación de clientes es una técnica fundamental en marketing que permite dividir a una base de clientes en grupos más pequeños y homogéneos con características similares. En el contexto del comercio electrónico, este proceso cobra especial relevancia debido a la gran cantidad de datos generados por cada interacción en la plataforma, como visitas, compras, métodos de pago y preferencias de productos. (Cuadros et al., 2017) Los datos de órdenes proporcionan información detallada sobre el comportamiento de los clientes, lo que facilita la identificación de patrones y la toma de decisiones estratégicas, orientadas a aumentar el nivel de transaccionalidad de la compañía o negocio.

Tradicionalmente, las empresas han utilizado métodos de segmentación basados en criterios demográficos o geográficos. Sin embargo, con el auge de las tecnologías de análisis de datos, ahora es posible emplear enfoques más sofisticados que consideran el comportamiento de compra y el valor de vida del cliente (CLV). Este tipo de segmentación no solo mejora la precisión de las campañas de marketing, sino que también optimiza la personalización de las experiencias de los usuarios en un entorno de alta competencia como el *e-commerce*.

2.4.2 Análisis RFM

El proceso de segmentación de clientes se sustenta en varios modelos teóricos y técnicas de análisis de datos. Uno de los enfoques más utilizados es el análisis RFM (Recencia, Frecuencia, Valor Monetario), que clasifica a los clientes según la fecha de su última compra, la frecuencia con la que compran y el valor total de sus compras. Fue creado hace aproximadamente 80 años, para satisfacer la necesidad de los negocios de aumentar sus ganancias, y fue muy popular para los pioneros del marketing en base al uso de datos como Stan Rapp, Tom Collins, David Pastor, entre otros. (Mas , 2016). Este método permite priorizar segmentos de clientes que tienen mayor potencial de generar ingresos y diseñar estrategias diferenciadas para cada grupo.

2.4.3 Customer Lifetime Value

La creación de segmentos de clientes en e-commerce se basa en un conjunto de conceptos clave relacionados con la analítica de datos. El *Customer Lifetime Value* (CLV), por ejemplo, es una métrica crítica que estima el valor económico que un cliente aportará a lo largo de su relación con la empresa. Este concepto es esencial para identificar a los clientes más valiosos y enfocar los esfuerzos de marketing en su retención. (Berger et al. , 1998)

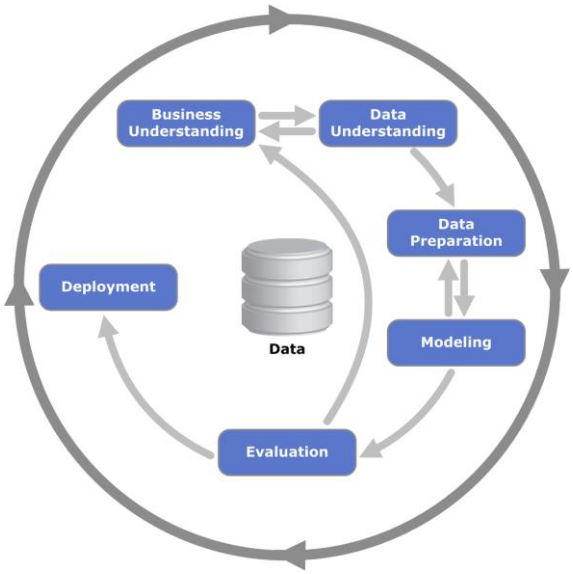
Por otro lado, el concepto de Personalización hace referencia a la capacidad de adaptar la oferta de productos y la comunicación a las necesidades específicas de cada segmento. Otro concepto importante es el Comportamiento de Compra, que incluye el análisis de patrones como el ciclo de vida de la compra, la frecuencia de pedidos, el valor medio del carrito, y las preferencias de productos. Todos estos factores contribuyen a una segmentación precisa, permitiendo a las empresas de e-commerce ajustar su propuesta de valor y optimizar la experiencia del cliente. (De Maya, 2006). Todos estos conceptos son fundamentales para desarrollar estrategias basadas en datos que mejoren la fidelización y maximicen los ingresos.

2.5 Metodología y Técnicas

2.5.1 Tipo de investigación

En esta investigación se usó un tipo de investigación cuantitativa. Este estudio se enfocó en el análisis de datos numéricos para identificar patrones y segmentar a los clientes del e-commerce en grupos homogéneos. En cuanto al diseño de la investigación, contempló un diseño no experimental y descriptivo. La investigación se basará en el análisis de datos históricos del comportamiento de los clientes, sin manipular las variables.

Fig. 1. Ciclo de la Metodología CRISP-DM



Fuente: AnalyStats, 2019.

2.5.2 Metodología CRISP-DM

En cuanto a metodologías, contamos con CRISP-DM (Cross-Industry Standard Process for Data Mining), que es una metodología estructurada que proporciona un marco detallado para llevar a cabo proyectos de minería de datos. Esta metodología se compone de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Su flexibilidad permite su aplicación en una variedad de industrias y se ha consolidado como uno de los estándares más utilizados en el análisis de datos. CRISP-DM promueve un enfoque iterativo, en el cual las fases no necesariamente deben seguirse en un orden secuencial estricto, lo que facilita la adaptación a los cambios en los datos o en los objetivos del negocio. (Larose et al., 2015)

Los pasos clave de CRISP-DM serían:

2.5.2.1 Comprensión del negocio.

Definir objetivos de negocio (segmentación de clientes) y planificar cómo los datos y el análisis apoyarán esos objetivos.

2.5.2.2 Comprensión de los datos.

Evaluar y familiarizarse con los datos históricos del e-commerce (transacciones, demografía, etc.). Esto incluye la recolección de datos, que en su mayoría serán datos transaccionales, datos de navegación en el sitio web, historial de compras, demografía, y otras variables relevantes que ya están disponibles en la base de datos del *e-commerce*.

2.5.2.3 Preparación de los datos.

Involucra algunos pasos como limpieza de datos; selección de variables: frecuencia de compra, valor promedio de compra, categorías de productos comprados; y escalado de datos.

2.5.2.4 Modelado.

Incluye análisis de clusters: *K-means*, clustering jerárquico, o *DBSCAN*; análisis de componentes principales (PCA): para reducir el número de variables; y Segmentación RFM de ser necesaria.

2.5.2.5 Evaluación.

Para evaluar la correcta ejecución de los algoritmos se utilizan medidas como la inercia , el índice de Silhouette o la distancia intra-clúster.

2.5.2.6 Despliegue.

Involucra interpretar los resultados y presentar resultados como las recomendaciones para la estrategia de marketing o ventas basada en los segmentos identificados.

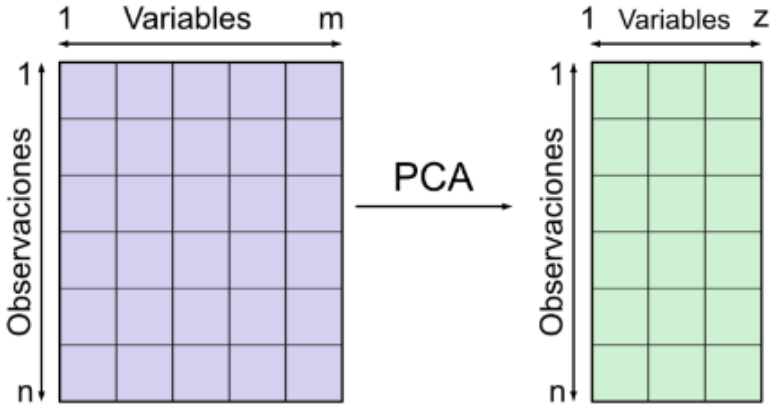
2.5.3 Técnicas de Modelado

2.5.3.1 Análisis PCA

El PCA es un proceso de reducción de dimensionalidad de los datos, supongamos que tenemos una matriz de datos de tamaño *n x m*, donde *n* representa el número de observaciones (filas) y *m* el número de variables (columnas). Tras aplicar PCA, la matriz se transforma en una nueva matriz de

tamaño *n x z*, donde *z* es un número reducido de variables resultantes, que capturan la mayor variabilidad de los datos originales. Además, cabe destacar que estas nuevas variables son independientes entre sí. (Calderón et al. ,2024).

Fig. 2. Proceso de reducción de dimensionalidad usando PCA



PCA se aplica sobre variables numéricas, ya que matemáticamente hablando, PCA se basa en la descomposición de la matriz de covarianza de las variables involucradas, esta matriz sirve para determinar la relación entre dos o más variables aleatorias.

Si deseamos usar PCA sobre variables categóricas, se deben primero convertir a un formato numérico, dependiendo de dichas variables se puede aplicar diferentes métodos, siendo el *One-Hot encoding* uno de los más populares, que crea una nueva columna binaria por cada valor único en la columna original. Si la columna es del tipo categórico, pero tienen un orden natural, la opción recomendada es la codificación por etiquetas (*Label Encoding*). Otro método recomendado para codificar las variables categóricas es la codificación de frecuencia (*Frecuency Encoding*), que asigna a cada categoría un valor numérico basado en la frecuencia con que aparece esa categoría en el conjunto de datos. (Harris, 2020).

2.5.3.2 K - Means

El algoritmo *K-Means* es una técnica popular de agrupamiento usada en aprendizaje automático no supervisado. Este divide un dataset en distintos clústeres K (grupos) con características similares, donde cada clúster está representado por su centroide (la media de los puntos dentro de ese cluster). Este algoritmo funciona iterativamente para asignar a cada punto de los datos a su centroide más cercano, y luego recalcula este centroide basado en las asignaciones existentes. El algoritmo continúa

hasta que las asignaciones ya no cambian o un número preestablecido de iteraciones es alcanzado. *K-means* es un algoritmo eficiente, pero tiene sus limitaciones, tales como una sensibilidad a la elección de los centroides iniciales y una tendencia a estancarse en mínimos locales. (MacQueen, 1967)

Determinar el número exacto de centroides (K) en el algoritmo de agrupamiento *K-means*, es un paso crucial, ya que esta cantidad puede afectar la calidad de los clusters. No existe un método o fórmula directa para calcular el número exacto de centroides, pero hay varios métodos y heurísticas que ayudan a determinar una buena aproximación para dicha cifra. Entre ellos tenemos:

- El método del codo, involucra graficar la inercia para diferentes valores de K y buscar el punto de inflexión en la gráfica, aquel punto donde la inercia empieza a decrecer más lentamente. La inercia mide la coherencia interna de los clústeres, a menor inercia mejor clustering.

- El *Silhouette Score*, este indicador numérico mide qué tan similar es cada punto a su clúster asignado comparado al resto de clústeres, provee una estimación de que tan bien el punto encaja dentro de su clúster asignado. Este valor varía entre 1 a -1, donde 1 indica que el punto está bien clusterizado y -1 indica que el punto pudo haber sido asignado al cluster equivocado.

- *Cross-Validation*, o validación cruzada, si bien es una técnica usada frecuentemente en aprendizaje supervisado, puede ser adaptada para clusterización, ya que mide que tan bien diferentes valores de K generalizan los datos.

- Conocimiento del dominio, conocer el contexto del problema, nos puede ayudar a determinar un número razonable de clusters basados en lo que sabemos del conjunto de datos, o del contexto del problema. (Lloyd, 1982)

Capítulo 3

Metodología

Este capítulo describe la implementación de un algoritmo de aprendizaje no supervisado para crear segmentos de clientes con la data de órdenes de un e-commerce.

Se usó la metodología CRISP-DM, con la cual se llevó a cabo la recolección y análisis de la información. El propósito final fue la creación de un modelo que permita identificar grupos de clientes diferenciados y proponer estrategias de marketing para ellos.

Se elegió la metodología CRISP-DM, ya que es un estándar reconocido y utilizado ampliamente en el campo de la minería de datos, que establece un marco bien definido de pasos y es altamente eficiente si es bien aplicado.

3.1 Entendimiento del negocio

3.1.1 Definición de Objetivos del Proyecto

El objetivo principal de este proyecto fue realizar una segmentación de clientes sobre los datos de órdenes de un e-commerce, lo que conllevó analizar los datos de los clientes actuales, desarrollar perfiles de clientes para los segmentos encontrados, y proponer estrategias de marketing personalizadas con el fin de aumentar métricas como la conversión y la retención de clientes.

3.1.2 Relevancia de la segmentación de clientes

La segmentación de clientes cobra relevancia hoy más que nunca, al encontrarse oportunidades de monetización en la información que los clientes proporcionan digitalmente a los comercios. Según estudios, se ha visto que proporcionar ofertas de productos y servicios personalizados a los clientes, aumenta la frecuencia de compra de los clientes, lo que se traduce en un aumento de las ganancias a las empresas que invierten en el estudio de las tendencias de compra.

3.1.3 Objetivos y Criterios de Éxito de Negocio

El éxito de este proyecto se medirá en base a la claridad de los segmentos de clientes identificados, que dichos segmentos tengan sentido desde una perspectiva de negocio, que los clientes

dentro de un segmento tengan patrones de comportamiento similares, y que se puedan recomendar estrategias de marketing o productos específicos para cada clúster.

3.1.4 Evaluación de la Situación

Se dispone de una base de datos que contiene la información de órdenes de un e-commerce que abarca el periodo de un año, lo que suma un aproximado de un millón cuatrocientos mil registros, en donde cada registro representa la compra de un cliente en diversos comercios afiliados al e-commerce. Para contextualizar más la situación, el *e-commerce* que proporcionó la información da el servicio de checkout y pagos en línea a diversos comercios a través de sus páginas web, todos estos comercios están situados en Ecuador y se dedican a diversos rubros, pero principalmente a la venta de comida, artículos de tecnología y de ropa.

Entre la información que encontramos, tenemos datos de dirección, ciudad, estado, código zip, email, nombres, apellidos, teléfonos, fecha de compra, monto de la orden, y datos de dispositivo como *deviceid*/*useragent*.

3.2 Comprensión de los datos

En esta fase, se examina la base de datos proporcionada para entender su estructura, revisar que campos nos serán útiles en nuestro estudio, y poder proponer hipótesis iniciales, de acuerdo a los lineamientos que nos sugiere la metodología CRISP-DM.

3.2.1 Recolección de datos iniciales.

En este proyecto, como se ha mencionado anteriormente, se utilizaron los datos específicos de las órdenes de un e-commerce. Estos datos principalmente contienen información del usuario, del dispositivo que se usó para la compra, de la orden y de los ítems que se usaron en la compra.

Datos del usuario: nombre, apellido, teléfono, código zip, email, datos de dirección.

Datos de la orden: fecha de compra y monto de orden.

Datos del dispositivo: *device id* y *user agent.*

Datos de ítems: nombre, cantidad y monto total de ítems por orden.

Estos datos vienen de distintas tablas en la base de datos del e-commerce, y se relacionan mediante un identificador único (order_id). Para propósitos de análisis se pivoteo la data de los ítems y se agregó una columna con la información del número de ítems comprados por orden por merchant.

3.2.2 Descripción de los datos

En esta sección describiremos la estructura y contenido de la base de datos disponible. Esta data contiene información de las órdenes de un e-commerce y principalmente contiene información de usuarios, items, órdenes y dispositivos.

La data proporcionada contiene un conjunto de 41 variables, que describen la información de una orden, es decir el usuario que realizó la compra, los items que contenía dicha compra, la fecha y monto de la orden, e información del dispositivo donde se hizo la compra. El conjunto de datos comprende la data recolectada en un año de funcionamiento del e-commerce en un país determinado (Ecuador) y contiene un total de 1 428 441 registros.

De este conjunto de datos seleccionaremos los datos más relevantes, en especial los datos de tipo numérico que nos servirán para crear nuestro modelo de aprendizaje no supervisado.

Información de atributos:

Tabla 1. Campos de la data proporcionada para el estudio

Orden	Variable	Descripción
0	address_description	Descripción de la dirección
1	address_1	Dirección del usuario (dada por gmaps)
2	address_2	Dirección del usuario (dada por el usuario)
3	city	Ciudad del usuario
4	country_code	Código del país
5	email_has	Hash del email del usuario
6	state_name	Estado del usuario
7	zip_code	Código zip del usuario
8	updated_at	Fecha del registro de la orden

9	order_id	Orden id (identificador de la orden)
10	amount	Monto total de la orden
11	id	Código único asociado al registro
12 - 34	Nro. items per merchant and order	Columnas conteniendo el número de ítems por orden.
35	device_id	Id del dispositivo.
36	referring_domain	Url del dominio (url de la página del comercio afiliado).
37	browser	Navegador usado para la compra.
38	browser_version	Versión del navegador usado.
39	os	Sistema operativo usado para la compra.
40	os_version	Versión del sistema operativo usado para la compra.
41	device	Nombre del dispositivo usado para la compra.

3.2.3 Exploración de los Datos

En esta fase de exploración de datos, se ha llevado a cabo un análisis de las variables numéricas para entender cómo se distribuye la data de órdenes recolectada.

Por propósitos de privacidad de la data, se han renombrado los nombres originales de los comercios con un alias. Cómo manejamos dos grupos de comercios, para poder hacer un mejor análisis hemos decidido clasificarlos en dos categorías: comida y retail.

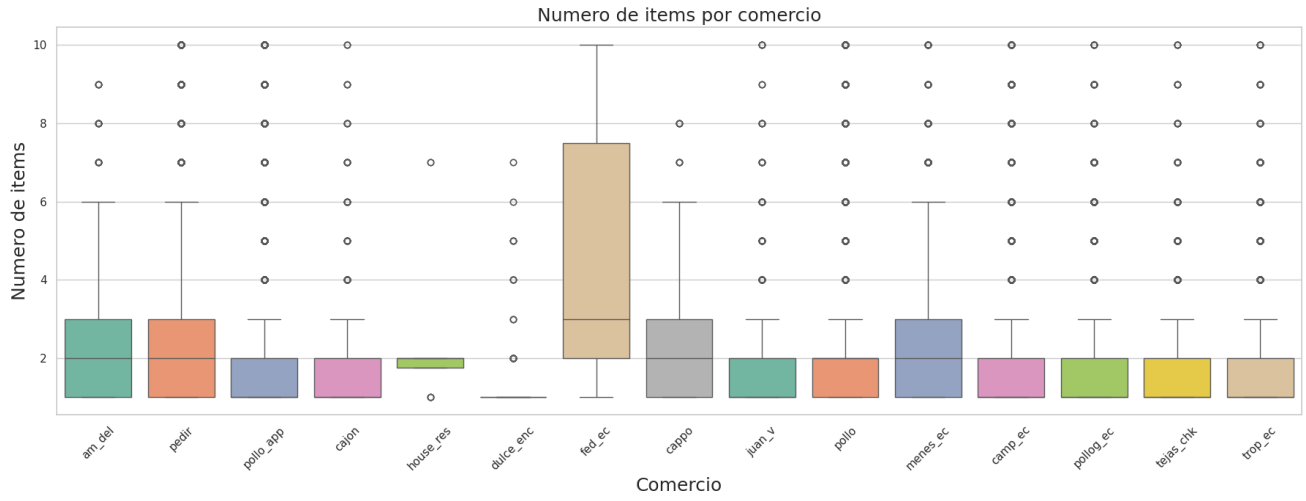
Tabla 2. Comercios involucrados en el estudio

Grupo Comida: comercios que venden comida.	Grupo Retail: comercios que venden tecnología, ropa y artículos varios.
am_del	chev_ec
pedir	gana_ec
pollo_app	imp_nov

cajon	bahia
house_res	novi_pc
dulce_enc	out_ec
fed_ec	store_ec
cappo	siete_ec
juan_v	
pollo	

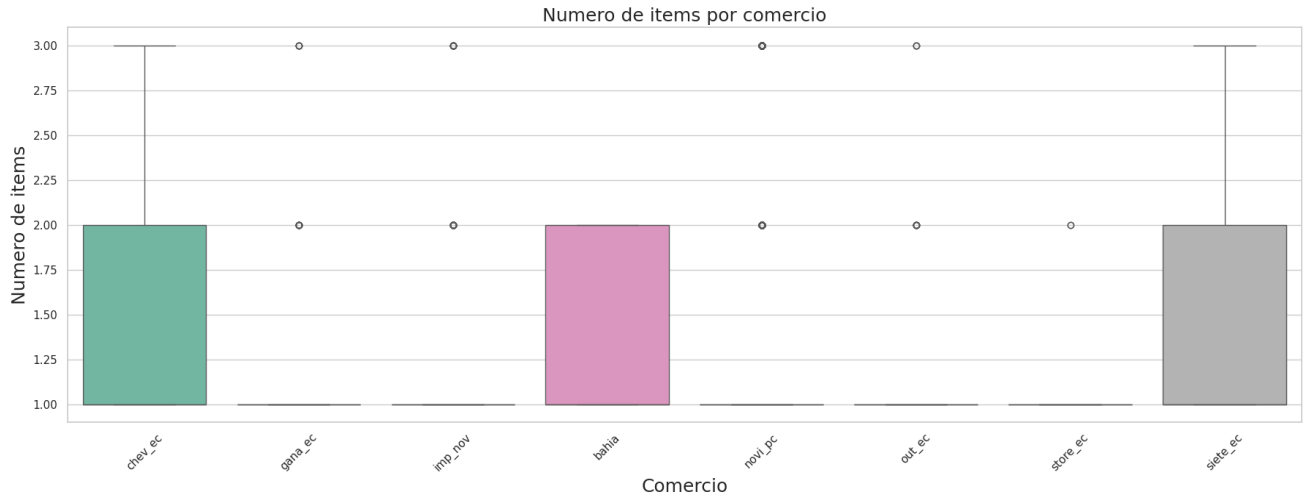
Para entender las diferencias en los patrones de compra entre comercios, se ha realizado un estudio entre la cantidad de ítems que tienen las compras, por ello se puede ver en la Fig. 3 que la mayoría de comercios que se dedican a la venta de comida tienen órdenes con dos o tres ítems en su mayoría, con valores atípicos que van hasta los 10 ítems por orden. El comercio *fed_ec* es el que mayor diferencia presenta, pero se atribuye esta diferencia a que es uno de los comercios con menor número de órdenes registrados.

Fig. 3. Número de **ítems** por orden en comercios de comida



En cuanto a los comercios que se dedican a la venta de items retail, podemos ver en la Fig. 4 que la mitad tienen 2 ítems por orden en su mayoría, y la otra mitad tienen 1 ítem por orden. Con la mayoría de valores atípicos extendiéndose hasta 3 ítems por orden. Lo que significa, que en este grupo la mayoría de clientes compraba un solo ítem, lo que tiene sentido al ser items costosos en su mayoría.

Fig. 4. Número de **ítems** por orden en comercios de retail



En cuanto al número de clientes por comercio, se puede ver en la Fig. 5, que para los comercios de comida, la gran mayoría de órdenes vienen del comercio *pollo_app*, con más de 900k registros, y se tiene comercios con menos de 100 registros como son el caso de fed_ec y house_res.

Fig. 5. Número de clientes por comercio del tipo comida

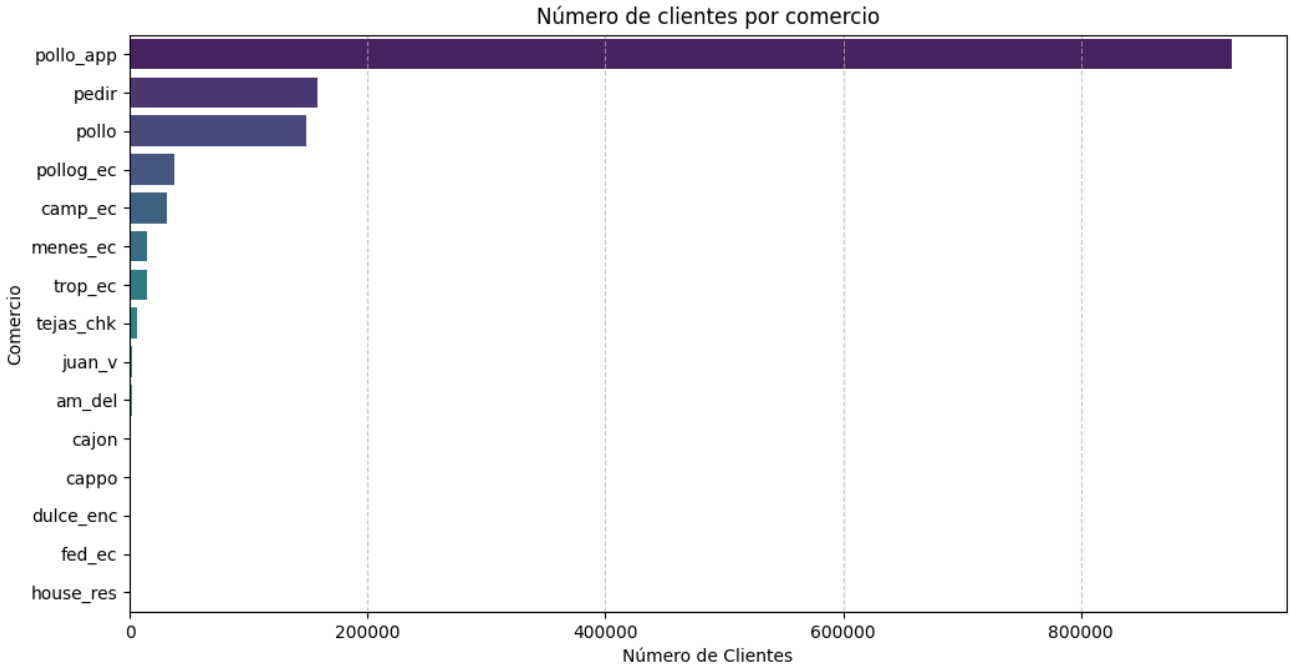


Tabla 3. Número de clientes por comercio del tipo comida

Comercio	Número clientes
pollo_app	926590
pedir	158314
pollo	148458
pollog_ec	37556
camp_ec	31771
menes_ec	14756
trop_ec	14203
tejas_chk	5993
juan_v	1932
am_del	1729

cajon	1453
cappo	463
dulce_enc	409
fed_ec	48
house_res	20

En cuanto a los comercios retail, se puede ver en la Fig. 6 que el mayor número de órdenes vienen del comercio *store_ec* con un poco más de 71k registros, y se tiene 2 comercios con un número de órdenes pequeño como es el caso de *out_ec* y *bahía*.

Fig. 6. Número de clientes por comercio del tipo retail

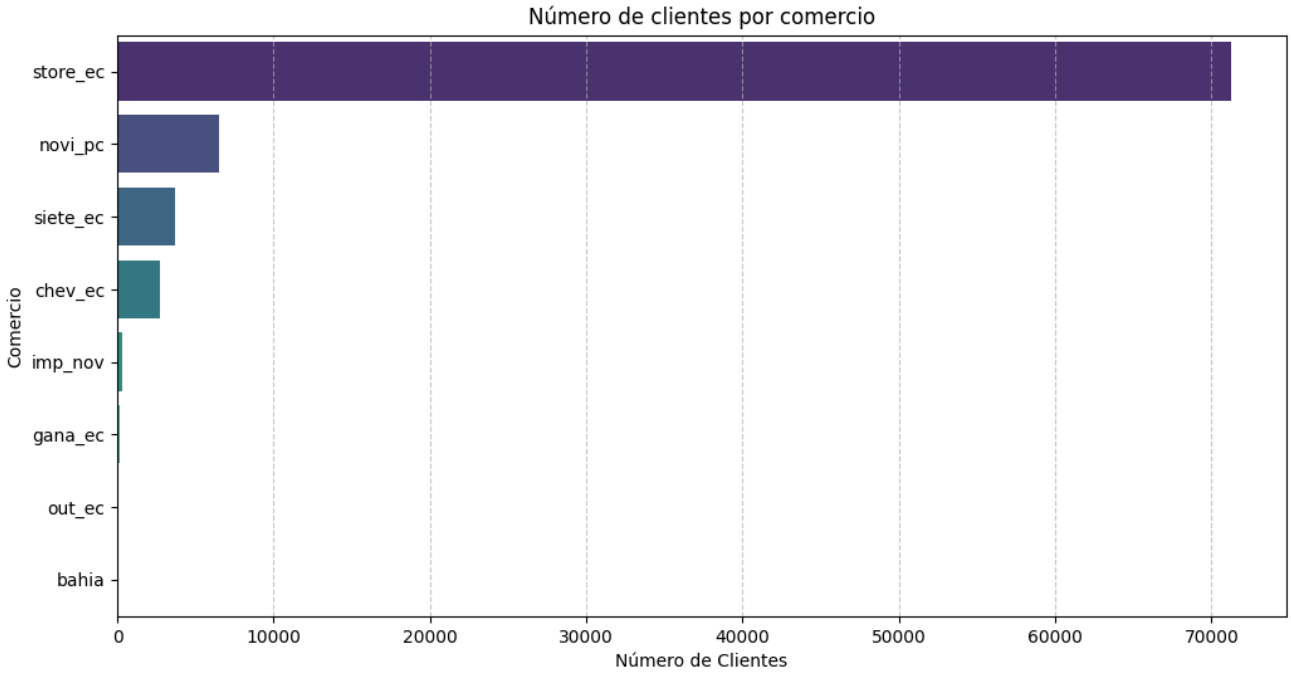
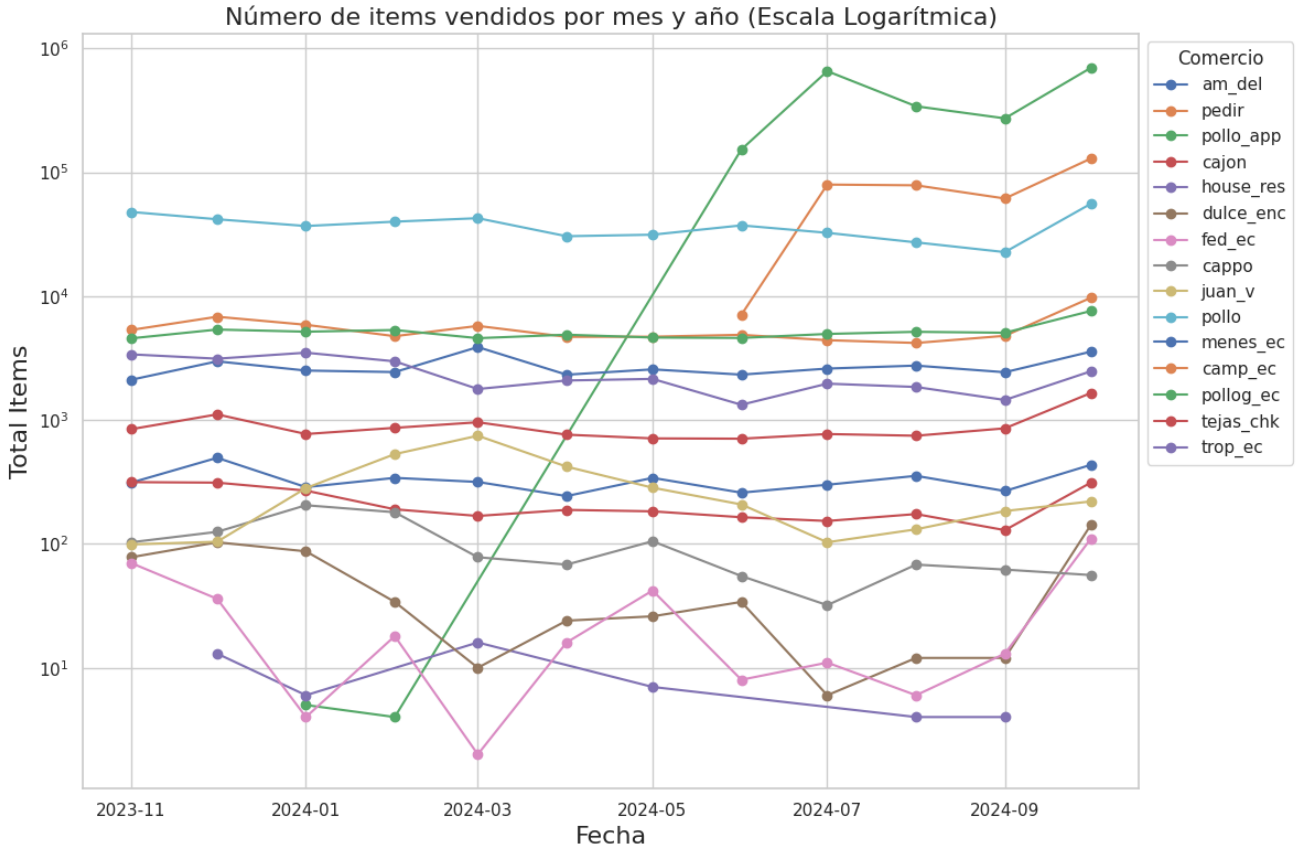


Tabla 4. Número de clientes por comercio del tipo retail

Comercio	Número clientes
store_ec	71237
novi_pc	6503
siete_ec	3702
chev_ec	2729
imp_nov	288
gana_ec	171
out_ec	73
bahia	38

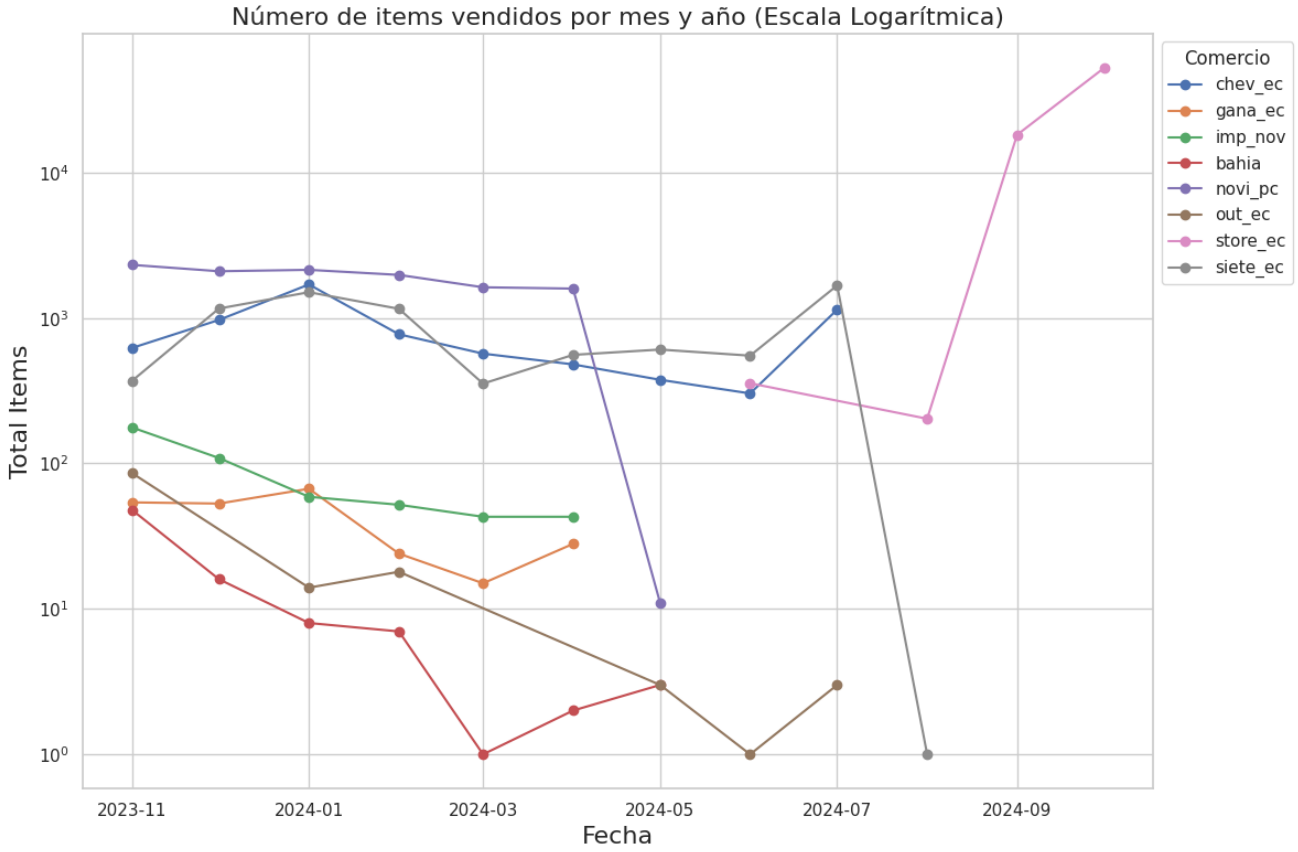
En cuanto a la evolución en tendencia de compra de los merchants de tipo comida a lo largo del año de data, se puede ver en la Fig. 7 que para la mayoría de comercios esta se mantuvo estable, exceptuando el caso de *pollo_app* que vio un incremento bastante grande en el número de ítems vendidos.

Fig. 7. Número de ítems vendidos a lo largo del tiempo - comercios de comida



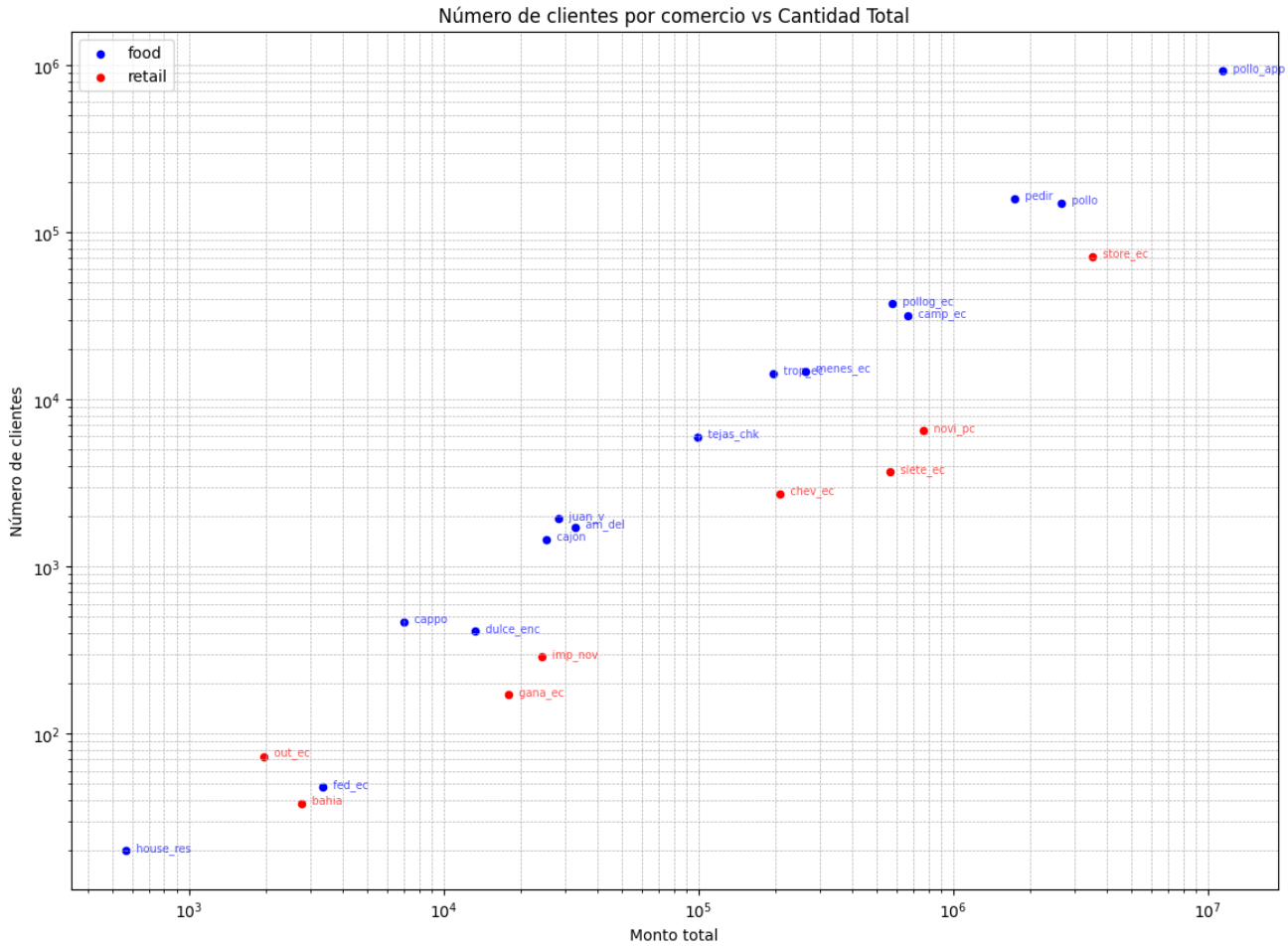
En cuanto a los merchants de tipo retail, se puede ver en la Fig. 8, que la mayoría de ellos tienen información de ventas al inicio del periodo analizado, pero no al final, esto se debe a que algunos comercios retail dejaron de trabajar con el e-commerce durante el periodo de tiempo analizado, y otros en cambio, como *store_ec* empezaron a trabajar a mitad del periodo analizado.

Fig. 8. Número de ítems vendidos a lo largo del tiempo - comercios de retail



Para entender el comportamiento del número de clientes por comercio y el monto total de las órdenes (medido en USD), se puede ver en la Fig. 9 una comparativa de los comercios de comida (en color azul) vs los comercios retail (en color rojo). Se puede ver que los comercios retail tienen un monto de orden superior a los de comida, lo que se evidencia en la gráfica donde a pesar que el número de registros (órdenes) de esta categoría es notablemente inferior, tienen números similares en cuanto al monto de compra.

Fig. 9. Número de clientes por merchant vs monto de compra en dólares



3.3 Preparación de los datos

Esta etapa consiste en limpiar y transformar los datos brutos antes del análisis e interpretación, involucra la selección de los datos, su limpieza, la construcción de datos , integración de datos y el formateo de datos de ser necesario.

3.3.1 Selección de datos

Del dataset original de datos, se seleccionó un subconjunto de campos, ya que haciendo un análisis de valores se ha podido ver que hay valores nulos en gran cantidad para ciertas columnas, los cuales hacen que usar esa información no sea viable. En la Tabla 5 tenemos un desglose de los campos que tienen una cantidad de registros nulos mayor a cero, todos aquellos campos con un porcentaje de nulos mayor a 60 serán descartados, revisando la causa de esta alta tasa de nulos, se ha determinado

que estos datos fueron recopilados mediante el frontend del servicio del *e-commerce*, no obstante debido a los bloqueadores de anuncios esta data a veces no es capturada correctamente.

Tabla 5. Columnas con un porcentaje de nulos mayor a 0.

Columna	Número de registros nulos	Porcentaje de nulos
deviceid	1 415 726	99.109869
referring_domain	1 243 687	87.066039
address_description	1 239 070	86.742820
address2	1 193 852	83.577271
os_version	1 193 807	83.574120
browser_version	1 193 100	83.524626
os	1 192 655	83.493473
browser	1 192 655	83.493473
device	1 192 655	83.493473
zip_code	1 161 144	81.287502
state_name	1 133 769	79.371077
country_code	1 087 434	76.127330
city	974 066	68.190846
address1	44 804	3.136566

Además de los campos del usuario, por motivos de privacidad he decidido eliminar los campos que hacen referencia a identificativos como en el caso del usuario:

- *first_name*
- *last_name*

- *identity_document*
- *email*
- *phone*

Se ha decidido persistir solo una columna *email_has* que es el valor original del email del usuario, como identificativo para poder reconocer clientes únicos en futuras etapas del proceso. Luego de este proceso de selección, en la Tabla 6 podemos ver los campos que nos servirán para nuestro estudio.

Tabla 6. Campos seleccionada de los datos

Orden	Variable	Descripción
1	address_1	Dirección del usuario (dada por gmaps)
2	email_has	Hash del email del usuario
3	updated_at	Fecha del registro de la orden
4	order_id	Orden id (identificador de la orden)
5	amount	Monto total de la orden
12 - 34	nro items per merchant and order	Columnas conteniendo el número de ítems por orden por comercio, en total 23 columnas

3.3.2 Limpieza de datos

En cuanto a limpieza de datos, en los campos que quedaron luego de la selección de datos no se ha notado la presencia de valores nulos. Por lo cual esta subetapa no conlleva acción de nuestra parte.

3.3.3 Construcción de datos

Se ha necesitado hacer algunas transformaciones a nuestra data, primero se ha decidido agrupar la información a nivel de usuario, usando como punto de agrupación el campo *email_has*, recordando que es nuestro identificativo a nivel de usuario. En este caso se sumó el número de ítems vendidos para las columnas de merchants, se sumó el monto de las órdenes para obtener un total a nivel de usuario, y se capturó la fecha de la orden más antigua y más reciente asociada a un usuario.

Al obtener el valor más antiguo y más reciente del campo *updated_at* se puede calcular dos variables útiles como lo indica la Fig. 10:

- *tenure*: antigüedad del cliente medida en días desde la primera compra hasta la fecha actual.
- *recency*: número de días desde la última compra a la fecha actual, es la métrica opuesta al tenure.

Fig. 10. Cálculo de tenure/recency

```
# Getting tenure and recency
current_date = pd.Timestamp.today()
result['updated_at_min'] = pd.to_datetime(result['updated_at_min'])
result['tenure'] = (current_date - result['updated_at_min']).dt.days
result['updated_at_max'] = pd.to_datetime(result['updated_at_max'])
result['recency'] = (current_date - result['updated_at_max']).dt.days
result
```

3.3.4 Integrar datos

La data proporcionada ya es el resultado de integrar cuatro fuentes de datos distintas: usuarios, órdenes, ítems y dispositivos, por lo cual no fue necesario de nuestro lado esta actividad.

3.3.5 Formato de los datos

En cuanto al formato final de los datos, vemos en la Tabla 7 los tipos de datos de los campos finales que se usarán en nuestro proyecto.

Tabla 7. Formato de los campos finales de la data

Orden	Variable	Tipo
1	email_has	Texto
2	recency	Numérico discreto
3	tenure	Numérico discreto
4	amount	Numérico continuo
5	order_count	Numérico discreto
12 - 34	nro items per merchant and order	Numérico discreto

3.4 Modelado

En esta fase de la metodología CRISP-DM, el modelado empezó con un análisis de correlación para seleccionar las variables más relevantes y ver si es posible aplicar un Análisis de Componentes (PCA) para reducir la dimensionalidad de los datos. Luego se aplicó el algoritmo *K-Means* para agrupar clientes en clústeres, donde se registraron asociaciones entre clientes con características similares.

3.4.1. Selección de técnicas de modelo

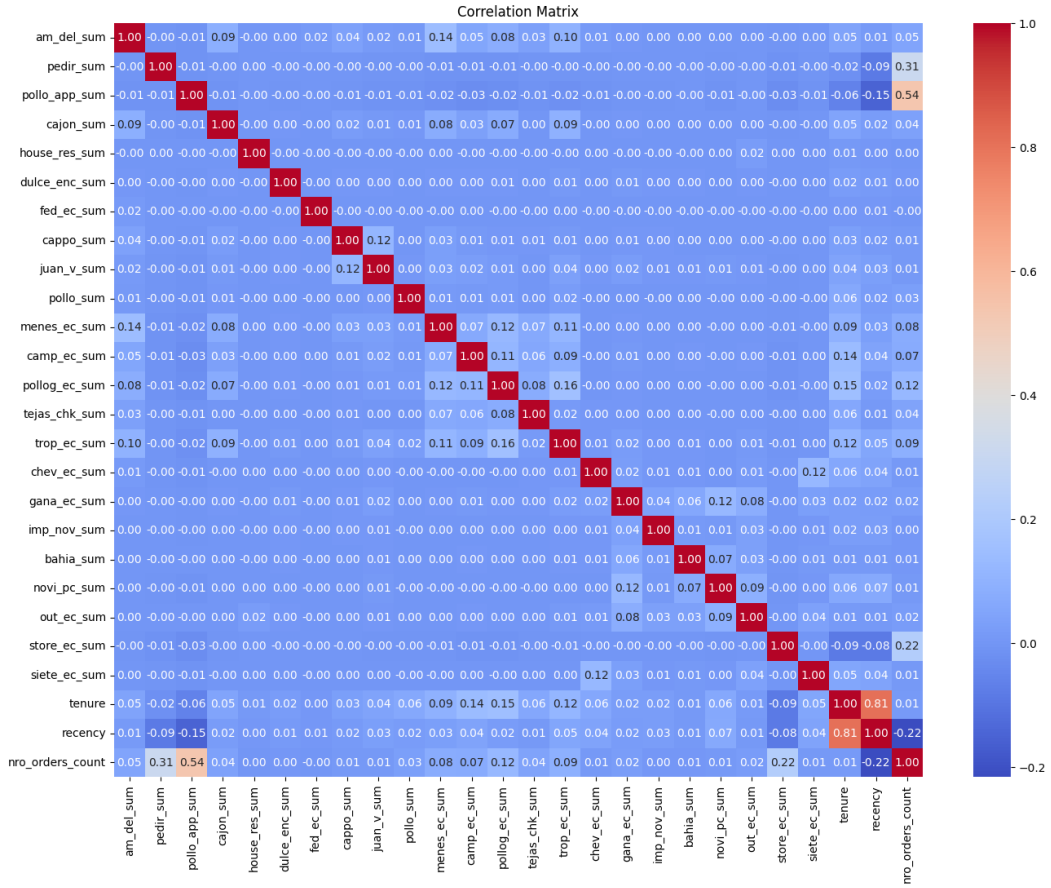
Inicialmente, se realizó un análisis de correlación para identificar y seleccionar las variables con mayor peso para el modelo. Este paso, es importante ya que permitió saber si tiene sentido aplicar componentes principales PCA.

Una vez concluido este paso, se procedió con la aplicación del algoritmo *K-means* que nos permitió la creación de clústeres. Una vez se dispuso de estos clústeres, se procedió a describir y analizar cada uno de ellos para obtener información relevante de cada uno, y así poder establecer segmentos de clientes de forma clara y eficiente.

3.4.1.1 Escoger la técnica de modelado

Antes de decidir aplicar o no Análisis de Componentes Principales, se realizó una análisis de correlación de variables.

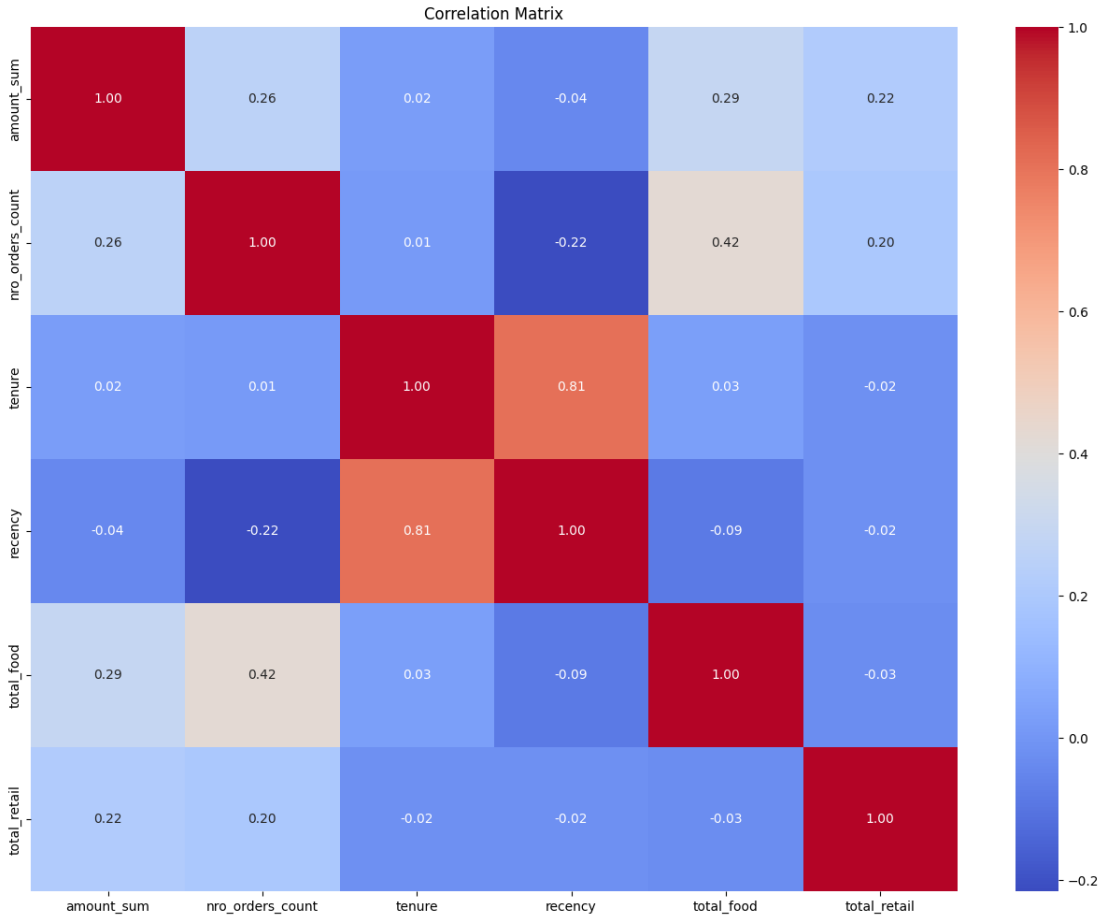
Fig. 11. Matriz de correlación original



No obstante, al observar los valores de correlación entre variables, se descartó la idea de aplicar PCA, debido a que existe un número muy bajo de variables con correlaciones altas. Por tanto, se decidió hacer un agrupamiento de la información en base al tipo de comercio, para tener un total de los ítems comprados por tipo de comercio, quedando nuestra data reducida a 7 variables, el *email_has* como identificativo y seis variables numéricas: *amount_sum* (total de las órdenes), *nro_orders_count* (número de órdenes por cliente), *tenure* (días transcurridos desde la orden más antigua a la fecha actual), *recency* (días transcurridos desde la orden más reciente a la fecha actual), *total_food* (número de ítems comprados en comercios del tipo comida) y *total_retail* (número de ítems comprados en comercios del tipo retail).

Al aplicar un análisis de correlaciones sobre esas variables obtuvimos el siguiente resultado:

Fig. 12. Matriz de correlación agrupando por tipo de comercio



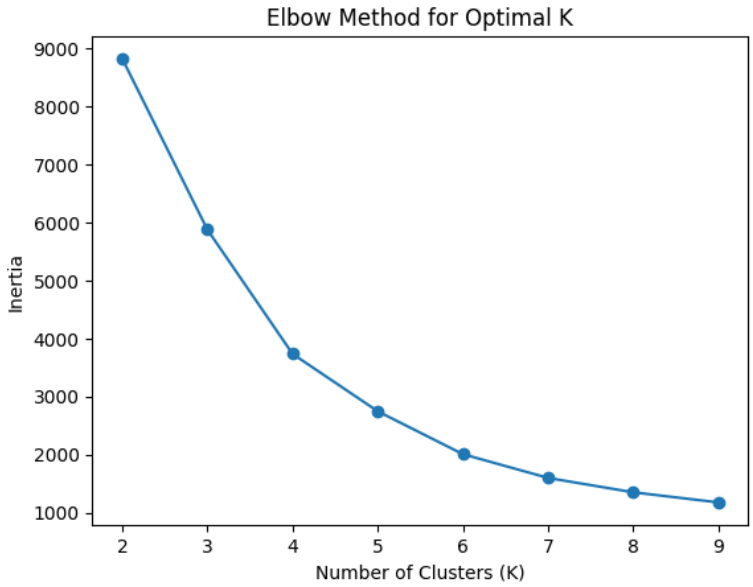
Sobre esta data se aplicó el algoritmo *K-means*, que crea clusters basados en las características similares de las variables del dataset.

3.4.1.2 Generación de modelos

En esta sección se detallan los parámetros e hiper parámetros empleados en el ajuste y modelado de las técnicas elegidas para la presente investigación.

En este caso, el principal hiper parámetro seleccionado es el número de clústeres, que fue seleccionado aplicando el método del codo, para lo cual se calculó la inercia para diferentes *k's* usando un rango de 2 a 9.

Fig. 13 Método del codo para determinar el *K* óptimo



Se puede apreciar claramente que $k = 4$ sería el número óptimo de clústeres en nuestro estudio.

3.5 Evaluación del Modelo

Para probar la calidad y validez de los clústeres, se usó un enfoque adaptado de evaluación, debido a la naturaleza no supervisada del análisis. Para evaluar la calidad de nuestros clústeres se usó Silhouette Score, que sirve para medir la calidad del agrupamiento, indicando que tan separados se encuentran los grupos.

Este cálculo se basa en dos medidas principalmente:

Cohesión: Se refiere a qué tan compactos están los puntos dentro de un mismo clúster. Mayor cohesión significa que los elementos de un clúster están muy cercanos entre sí.

Separación: Se refiere a qué tan distintos son los clústeres entre sí. Una buena segmentación logra una alta separación entre grupos.

Al ser el Silhoutte Score una medida sobre un punto, se calculó el promedio de todos los puntos para de esta forma determinar que tan buena fue la clusterización. El valor promedio resultante fue 0.8631 que nos indica que los datos dentro de cada clúster están bien agrupados, los clústeres están bien separados entre sí y que, en general, el modelo de clustering está funcionando correctamente.

Capítulo 4

Resultados

En este capítulo se abordan dos aspectos principales: la aplicación de análisis de clústeres generados para establecer segmentos de clientes, y sugerencias de marketing para cada segmento identificado. Para esto, se mostrará visualmente cómo está construido cada cluster tomando en cuenta sus variables.

4.1 Caracterización Clúster

Variables importantes:

Tabla 8. Variables consideradas en caracterización

Variable	**Descripción**
Amount sum	Suma del total de las órdenes del cliente.
Nro orders count	Número total de las órdenes del cliente.
Tenure	Número de días desde la compra más antigua del cliente a la fecha actual.
Recency	Número de días desde la compra más reciente del cliente a la fecha actual.
Total food	Número de ítems comprados por tipo de comercio de comida.
Total retail	Número de ítems comprados por tipo de comercio retail.

4.1.1 Clusterización K-means.

4.1.1.1 Clusterización y visualización de clústers formados.

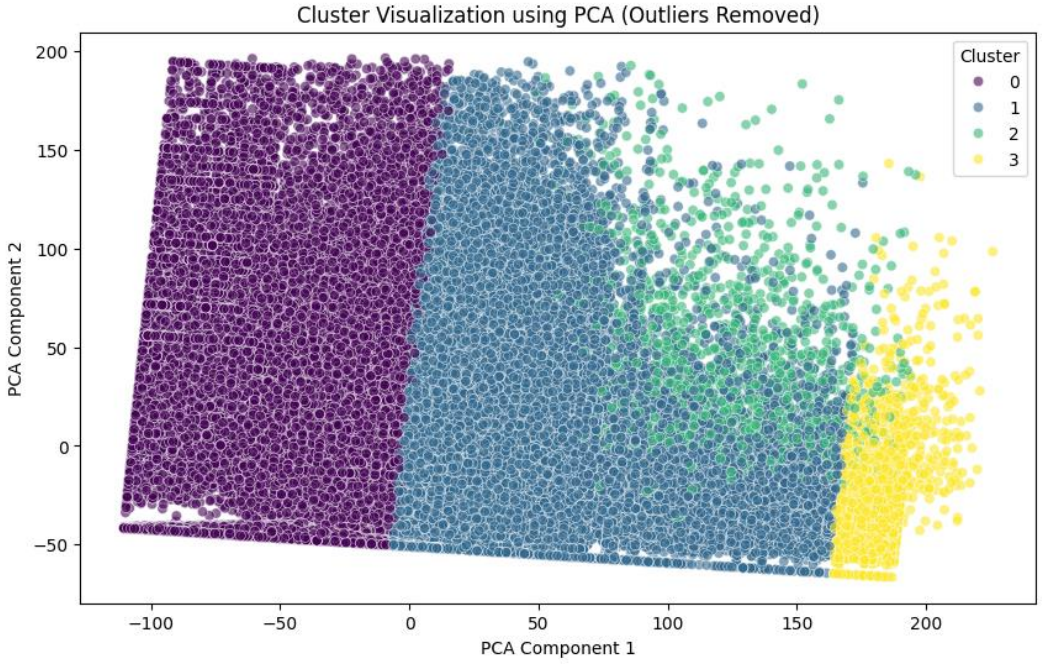
Se obtuvo un total de cuatro clústeres:

Tabla 9. Distribución de clústers formados

Clúster	Nro Registros	Porcentaje
0	109623	47%
1	77012	33%
3	32084	14%
2	11133	4%

Usando PCA para visualizar los clusters formados:

Fig. 14. Visualización de clusters usando PCA (outliers removidos)



4.1.1.2 Varianza explicada.

La PC1 explica 59.19% de la varianza total, mientras que la PC2 explica 34.99% de la varianza total. En conjunto, estos dos componentes explican 94.19% de la variabilidad en los datos, lo que indica que capturan la mayoría de la estructura de los datos originales.

4.1.1.3 Interpretación de los componentes principales.

Cada componente es una combinación lineal de las variables originales, con coeficientes (loadings) que indican la importancia relativa de cada variable.

PC1 (59.19% de la varianza)

Está fuertemente influenciado por variable *tenure* (0.7289) con tendencia positiva y dominante, y por la variable *recency* (0.6800), también con tendencia positiva y fuerte. Estas variables están relacionadas con la antigüedad y la frecuencia de interacción del cliente, lo que sugiere que PC1 representa la longevidad y actividad del cliente en el e-commerce.

La variable *amount_sum* (0.0792) también contribuye, pero en menor medida. Y la variable *nro_orders_count* (-0.0025) y *total_food* (-0.0008) apenas tienen impacto.

PC2 (34.99% de la varianza)

Está dominado por *amount_sum* (0.9889), lo que indica que este componente representa principalmente el monto total gastado por el cliente. Las variables *total_food* (0.0493) y *nro_orders_count* (0.0397) tienen un efecto menor pero positivo. Mientras la variable *recency* (-0.1330) tiene un peso negativo, lo que sugiere que los clientes con compras recientes pueden diferenciarse en este eje.

4.1.2 Caracterización de segmentos utilizando las variables principales

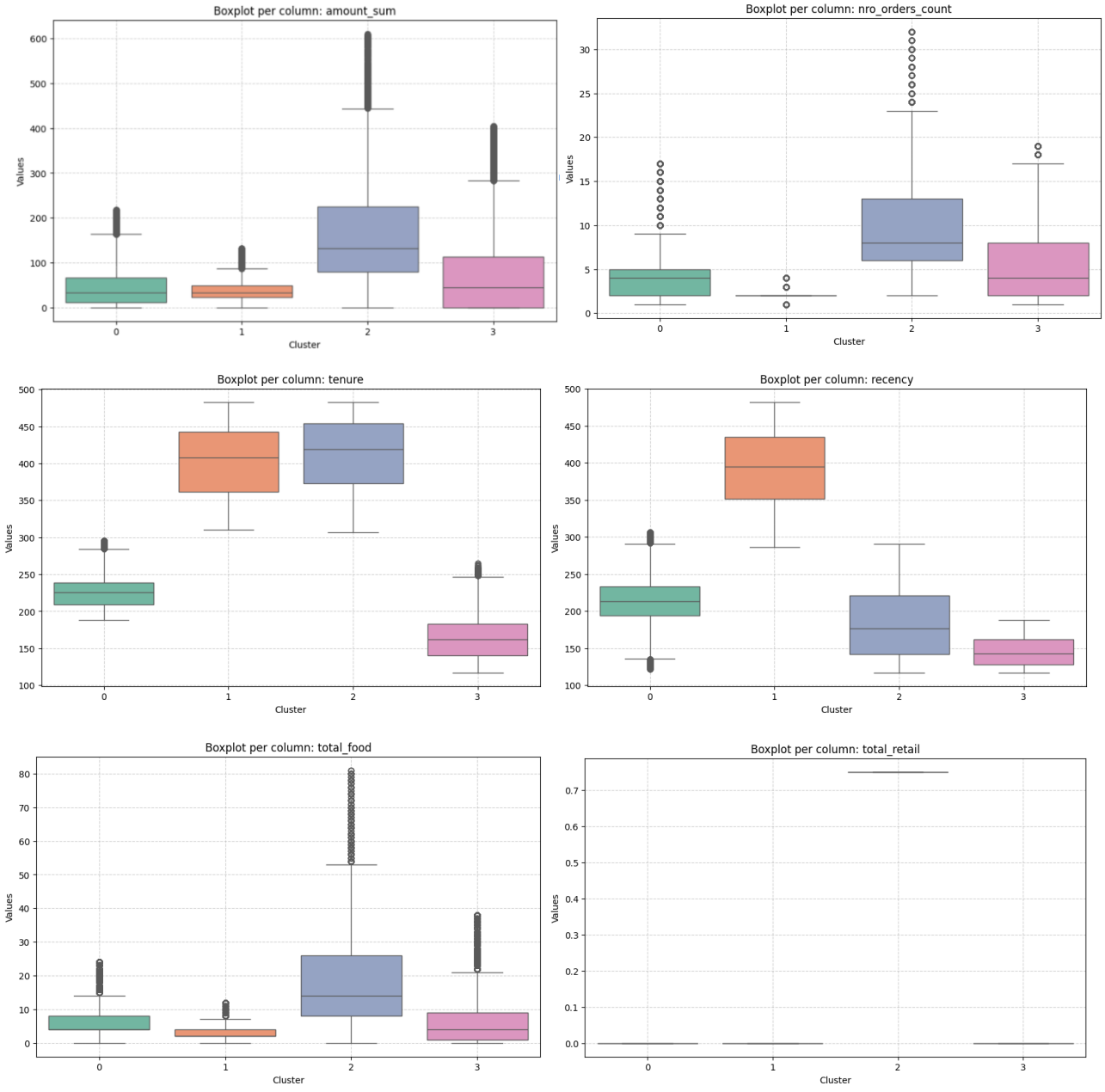
Para la caracterización de los segmentos se utilizó las variables principales implicadas en su construcción, para hacer evidente las diferencias entre los cuatro grupos identificados. Realizar una correcta diferenciación es clave para entender cómo está representado cada grupo de clientes, ya que nos permite establecer una base sólida para el desarrollo de estrategias de marketing.

Con este propósito en mente, se aplicó un rango intercuartílico, que se entiende como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de la información involucrada, se usa para identificar

la amplitud de la mitad central de los datos, es decir aquellos valores que se encuentra entre el 25% y el 75% de los datos ordenados. A través del rango intercuartílico, se puede identificar los valores atípicos, y excluirlos de nuestras gráficas con el fin de mostrar la información de manera óptima.

Se puede observar cómo se distribuye cada variable dentro de cada cluster, luego de aplicado un rango intercuartílico para excluir los outliers:

Fig. 15 Distribución de variables en los clústeres



También es relevante analizar los valores promedio de cada variable dentro de cada clúster, se adjuntan los resultados del cálculo, uno con los datos en su estado original y otro luego de excluidos los valores atípicos.

Valores promedio de variables utilizadas para construir los clústers

Tabla 10. Análisis Preliminar de clústeres sin tratamiento de Outliers

Cluster	amount_sum (USD)	nro_orders	tenure	recency	total_food	total_retail
0	74.47	5.43	228.28	213.86	12.09	0.058
1	79.78	2.53	403.14	393.41	6.21	0.59
2	211.47	12.09	412.07	183.74	29.49	0.88
3	112.19	7.30	166.00	145.65	15.29	0.68

Tabla 11. Análisis Preliminar de clústeres con tratamiento de Outliers - rango intercuartílico

Cluster	amount_sum (USD)	nro_orders	tenure	recency	total_food	total_retail
0	29.65	4.45	227.14	215.95	6.75	0.0
1	45.13	2.66	367.57	351.49	5.04	0.0
2	114	7.30	406.87	195.60	11.58	0.75
3	55.01	4.81	164.18	149.64	7.10	0.0

Para el cluster 0, se puede ver que el monto promedio de compra por cliente varía de 74.47 a 29.65 USD, y que el total de ítems comprados por comercio tipo comida varía de 12.09 a 6.75, el resto de variables varía ligeramente.

Para el cluster 1, se puede notar que el monto promedio de compra sufre una disminución significativa pasando de 79.78 a 45.13 USD, y que los valores de tenure y recency varían unas 40 unidades aproximadamente, el resto de variables varía ligeramente.

Para el cluster 2, se evidencia que el monto promedio de compra se reduce aproximadamente a la mitad sin outliers (pasando de 211.47 a a 114 USD), el número de órdenes también se reduce de

12.09 a 7,30, y que el total de ítems comprados por comercio tipo comida varía de 29.49 a 11.58, el resto de variables varía ligeramente.

Para el cluster 3, se puede notar que el monto promedio de compra también se reduce un 50% aproximadamente luego del tratamiento de outliers pasando de 112.19 a 55.01 USD, lo mismo ocurre con el número de órdenes promedio por cliente y el total de ítems comprados por comercio tipo comida, el resto de variables sufre una variación ligera.

Se puede apreciar que luego de hacer el tratamiento de outliers, a través de la aplicación de un rango intercuartílico, nos da una visión más precisa de cómo están constituidos cada clúster, lo que también a su vez facilita su caracterización.

4.1.3 Etiquetas de los clústeres

Descripción de los Clusters

Cluster 0 - Compradores Recientes de Bajo Gasto

Son clientes que gastan en promedio 29.65 USD por orden y realizan en promedio 4.45 órdenes. Considerando sus valores de tenure, recency y la última fecha de la data (31 de Octubre de 2024) se puede afirmar que este segmento de clientes compró durante un periodo de 15 días aproximadamente y la última fecha de compra fue 3 meses antes de la última fecha de compra registrada. Compran más productos de comida (6.75 ítems) y casi nada en retail.

Perfil: Clientes recientes con compras recurrentes de menor valor, especialmente en comida.

Cluster 1 - Compradores Antiguos de Bajo Gasto

Son clientes que gastan 45.13 USD por orden y realizan en promedio 2.66 órdenes. Considerando sus valores de tenure, recency y la última fecha de la data (31 de Octubre de 2024) se puede afirmar que este segmento de clientes compró durante un periodo de 20 días aproximadamente y la última fecha de compra fue 11 meses antes de la última fecha de compra registrada. Compran menos cantidad de productos de comida (5.04 ítems) y casi nada en retail.

Perfil: Clientes antiguos pero inactivos, con compras poco frecuentes y de bajo valor.

Cluster 2 - Compradores Premium de Alto Gasto

Son clientes que gastan 114 USD por orden y realizan en promedio 7.3 órdenes. Considerando sus valores de tenure, recency y la última fecha de la data (31 de Octubre de 2024) se puede afirmar que este segmento de clientes compró durante un periodo de 7 meses aproximadamente y la última fecha de compra fue 2 meses antes de la última fecha de compra registrada. Compran significativamente más productos de comida (11.58 ítems) y un ítem en promedio en retail.

Perfil: Compradores de alto valor y fidelidad, aunque con cierta inactividad reciente.

Cluster 3 - Compradores Recientes de Gasto Moderado

Son clientes que gastan 55 USD por orden y realizan en promedio 4.81 órdenes. Considerando sus valores de tenure, recency y la última fecha de la data (31 de Octubre de 2024) se puede afirmar que este segmento de clientes compró durante un periodo de 1 mes aproximadamente y la última fecha de compra fue 1 meses antes de la última fecha de compra registrada. Este es el segmento con clientes más recientes de la data. Compran en promedio 7.1 ítems de comida y nada en retail.

Perfil: Clientes nuevos o en crecimiento con hábitos de compra regulares.

4.2 Estrategias de marketing para los segmentos identificados.

4.2.1 Compradores Recientes de Bajo Gasto

Problema: Clientes relativamente nuevos con cierta frecuencia de compra, pero con un ticket promedio bajo.

Objetivo: Incrementar su ticket de compra y fidelizarlos a largo plazo.

Posibles Estrategias:

Descuentos por compra mínima: dar un descuento si supera cierto umbral en su monto de compra, para lograr aumentar el monto por orden, ya que este grupo posee el valor promedio de compra más bajo con un valor de 29.65 USD.

Cross-selling y up-selling: Sugerir productos complementarios en comida para que añadan más ítems al carrito, o crear promociones del tipo "Compra 3 y llévate el 4to gratis" para aumentar el número de ítems comprados. Esto aumentaría el número de ítems promedio comprados que dentro de este grupo es de 7 en la categoría de comercios tipo comida y 0 para el tipo retail.

Programa de recompensas rápido: Dar beneficios inmediatos, como puntos por cada compra para canjear en su siguiente orden.

Promociones de retención: Envío de recordatorios a través de medios como email y SMS con descuentos si pasan más de cierto tiempo en meses sin comprar.

Personalización en campañas de marketing: Envío de correos con recomendaciones basadas en sus compras recientes y/o encuestas para conocer mejor sus preferencias y ofrecerles ofertas personalizadas.

4.2.2 Compradores Antiguos de Bajo Gasto

Problema: Son clientes inactivos desde hace casi un año, con pocas compras y bajo gasto.

Objetivo: Reactivar su interés y aumentar su frecuencia de compra.

Posibles Estrategias:

Campañas de reactivación: Envío de emails/SMS personalizados con un descuento exclusivo, haciendo énfasis en que no han realizado compras recientes, además se puede incluir ofertas flash en productos de comida, ya que es su categoría preferida. Esto debido a que su recencia es la peor de todos los grupos, con casi un año a la fecha del estudio de su última compra.

Programas de recompensa: Incentivar compras con puntos de fidelidad acumulables por cada orden y otorgar bonos de bienvenida por la primera compra después de un periodo de inactividad.

Segmentación en redes sociales: Anuncios con productos que solían comprar y recordatorios de su última compra con recomendaciones alineadas a la misma.

Reactivación con envíos gratis o promociones: Para el caso de delivery, ofrecer envío gratis con su compra es una forma efectiva de lograr la reactivación de clientes, lo que puede ayudar a aumentar el monto promedio de compra que para este grupo es el segundo más bajo, con un valor de 45.13 USD.

Encuestas y feedback: Enviar cuestionarios cortos para entender por qué dejaron de comprar, se puede complementar estos cuestionarios con incentivos tales como cupones de descuento y similares.

4.2.3 Compradores Premium de Alto Gasto

Problema: Clientes de alto valor, pero con cierta inactividad reciente.

Objetivo: Mantener su lealtad y reactivar su compra frecuente.

Posibles Estrategias:

Membresías o suscripciones: Ofrecer una suscripción VIP con beneficios exclusivos como descuentos recurrentes o envíos gratis.

Acceso prioritario a productos o lanzamientos: Notificación sobre nuevos productos antes que al resto de los clientes. Además de acceso a ofertas personalizadas en sus categorías favoritas (comida y retail), ya que son el grupo con el número de ítems promedio más alto en ambas categorías (12 y 1 respectivamente).

Bonos por compras recurrentes: Implementar incentivos como promociones por compras en meses consecutivos, lo que aumentará la frecuencia del cliente.

Descuentos progresivos: otorgar descuentos especiales al superar cierto umbral de compra, lo que es factible dado el volumen de venta que se maneja en este segmento, siendo el grupo con el promedio de compra más grande, con un valor de 114 USD.

4.2.4 Compradores Recientes de Gasto Moderado

Problema: Son clientes bastante recientes, aún no completamente fidelizados.

Objetivo: Convertirlos en clientes recurrentes y aumentar su ticket promedio.

Posibles Estrategias:

Onboarding con incentivos: Correos de bienvenida con un descuento especial en su segunda compra, y la creación o promoción de programas de fidelización o suscripciones, ya que son el primer grupo con compras más recientes, con una recencia de 150 días aproximadamente a la fecha del estudio.

Cross-selling y up-selling: Sugerir productos relacionados a los de su carrito de compras, además de sugerir items comprados por clientes con perfiles de compra similares, lo que es factible al tener este grupo el segundo número más alto de ítems por compra promedio, específicamente para comercios tipo comida, con 7 ítems.

Gamificación: retos enviados a través de canales como email y SMS, con promociones del estilo “Compra n veces este mes y recibe un cupón de $10 para tu siguiente compra”

Recordatorios personalizados: envío de alertas cuando sus productos recientemente comprados están en descuento, y notificaciones de recompra al transcurrir cierto tiempo desde su última compra.

Ofertas para compras frecuentes: implementar descuentos si compran en intervalos regulares.

Capítulo 5

Conclusiones y Recomendaciones

5.1 Conclusiones

- La aplicación de algoritmos de aprendizaje no supervisado, como el método *K-means* para la clusterización, permitió segmentar el banco de datos en cuatro grupos de clientes, partiendo de una base de 229 852 clientes distintos. Al tener estos segmentos claramente identificados, permitió sugerir estrategias de marketing personalizadas para cada uno de ellos, lo que permitirá aumentar la conversión y retención de los clientes involucrados. Además, se realizó una validación de la clusterización usando métricas como el *Silhouette Score*, que corroboró la precisión del modelo planteado, asegurando la debida diferenciación y cohesión entre los puntos dentro de los clusters formados.

- A partir de una base de 229 852 clientes distintos, se pudo identificar cuatro segmentos clave. El primero etiquetado como *Compradores Recientes de Bajo Gasto* contempla aquellos clientes relativamente nuevos, pero con un ticket promedio bajo, con un total de 109 623 registros. El segundo segmento *Compradores Antiguos de Bajo Gasto* son aquellos clientes inactivos desde hace casi un año, con pocas compras y bajo gasto, que probablemente probaron el servicio, pero no quisieron volver a usarlo, contabilizando un total de 77 012 registros. El tercer segmento *Compradores Premium de Alto Gasto,* son aquellos clientes de alto valor, pero con cierta inactividad reciente, con un total de 11 133 registros. Y por último, el cuarto segmento *Compradores Recientes de Gasto Moderado* son aquellos clientes bastante recientes, aún no completamente fidelizados, que incluye un total de 32 084 registros.

- Se logró verificar la eficiencia y utilidad de las técnicas de Machine Learning aplicadas al ámbito del comercio electrónico, ya que permite estudiar en masa la data recolectada durante un periodo de tiempo determinado (en este caso un año), y poder capturar similitudes en la data que no son visibles de forma sencilla, lo que permite sacarle provecho a toda la información disponible, aumentado las ganancias para la empresa al fidelizar e incrementar el consumo de sus clientes.

- Se pudieron crear recomendaciones de marketing personalizadas para los diversos segmentos, tomando en cuenta el volumen de compra (monto y cantidad de órdenes), el periodo de tiempo que el cliente ha estado usando el servicio, y el tipo de comercio que más frecuenta (comida o retail).

5.2 Recomendaciones

- Los algoritmos de aprendizaje no supervisado, como *K-means*, han demostrado ser efectivos en la identificación de patrones y similitudes en la data permitiendo la creación de segmentos de clientes. Por lo tanto, se recomienda aplicar otros algoritmos del mismo tipo como DBSCAN, Clustering Jerárquico, entre otros, para poder realizar una comparación de los resultados con los del presente estudio, y así poder complementar los segmentos identificados con la información arrojada por estos modelos.

- Se recomienda usar las estrategias de marketing proporcionadas en esta investigación como como base para su implementación en un entorno productivo real, con lo cual se podría medir el impacto de las mismas, y de ser necesario iterar en el estudio de la data, ya con los resultados registrados, y con nueva data más reciente.

- Se recomienda integrar más información al modelo planteado, por ejemplo, la variable de género sería útil para identificar tendencias de compra y frecuencia ligadas a hombres/ mujeres lo que permitiría sugerir productos de forma más personalizada y de esta forma aumentar la efectividad de nuestras estrategias de marketing. Esto se podría implementar en la pantalla de registro del usuario, donde se constató que dicha información no está siendo recolectada. Otras variables como la edad también podrían resultar relevantes.

BIBLIOGRAFÍA

El Universo. (2024, July 24). ¿Cuáles son los sitios web más visitados para comprar en Ecuador? El Universo. https://www.eluniverso.com/noticias/informes/cuales-son-los-sitios-web-mas-visitados-para-comprar-en-ecuador-nota/

Moro, M. S., & Fernández, J. C. (2020). Marketing digital y dirección de e-commerce: Integración de las estrategias digitales. Esic Editorial.

Salcedo, R. A., & López, M. A. (2019). Big data: Aplicaciones de la gestión del dato en las distintas etapas del funnel de conversión. Revista de Marketing y Publicidad, 39-68.

Cuadros López, Á. J., Gonzales Caicedo, C., & Jiménez Oviedo, P. C. (2017). Análisis multivariado para segmentación de clientes basada en RFM. Tecnura, 21(54), 41-51.

Mas Díaz, R. M. (2016). Análisis del modelo RFM según el método convencional y el método de las 2-Tuplas.

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of interactive marketing, 12(1), 17-30.

De Maya, S. R., & Esteban, I. G. (Eds.). (2006). Comportamientos de compra del consumidor. ESIC Editorial.

Larose, D. T., & Larose, C. D. (2015). Descubriendo el conocimiento en los datos: Introducción a la minería de datos. Editorial Cengage Learning.

Amazon Web Services. (n.d.). ¿Qué es la ciencia de datos? AWS. Retrieved February 17, 2025, from https://aws.amazon.com/es/what-is/data-science/

Ponce Gallegos, J. C., Torres Soto, A., Quezada Aguilera, F. S., Silva Sprock, A., Martínez Flor, E. U., Casali, A., ... & Pedreño, O. (2014). Inteligencia artificial. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn).

Rebala, G., Ravi, A., Churiwala, S., Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine learning definition and basics. An introduction to machine learning, 1-17.

Shyam, R., & Chakraborty, R. (2021). Machine learning and its dominant paradigms. Journal of Advancements in Robotics, 8(2), 1-10p.

Harris, A. (2020). One-Hot Encoding vs Label Encoding in Machine Learning. [URL de la fuente, si es un artículo web].

Calderón, C. A. A., & Poveda, E. M. B. (2024). Implementación del Método de Análisis de Componentes Principales (PCA) para la reducción de la dimensionalidad en los datos inmobiliarios de la ciudad de Riobamba. Dominio de las Ciencias, 10(3), 2032-2051.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297. https://doi.org/10.1007/BF01812717

Lloyd, S. P. (1982). *Least squares quantization in PCM*. IEEE Transactions on Information Theory, 28(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489