UNIVERSIDAD NACIONAL AGRARIA DE LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"SEGMENTACIÓN DE CLIENTES DIGITALES DEL ECOMMERCE DE UNA EMPRESA DEL SECTOR RETAIL CON ALGORITMOS DE ANÁLISIS CLUSTER"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

LUCILA NOEMÍ CAYCHO HUAMANÍ

LIMA- PERÚ

2024

La UNALM es titular de los derechos patrimoniales de la presente investigación

(Art.24 – Reglamento de Propiedad Intelectual)

Monografía_TSP_LucilaCaycho VF2.docx

INFORME DE ORIGINALIDAD

17%	14%	1%	12%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.lamolina.edu.pe Fuente de Internet	3%
2	Submitted to Universidad Nacional Agraria La Molina Trabajo del estudiante	1%
3	Submitted to Universidad TecMilenio Trabajo del estudiante	1%
4	maplink.global Fuente de Internet	1%
5	Submitted to Universidad Carlos III de Madrid - EUR Trabajo del estudiante	1%
6	Submitted to ITESM: Instituto Tecnologico y de Estudios Superiores de Monterrey Trabajo del estudiante	1%
7	rstudio-pubs-static.s3.amazonaws.com Fuente de Internet	1%
8	Submitted to Universidad Anahuac México Sur	1%

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"SEGMENTACIÓN DE CLIENTES DIGITALES DEL ECOMMERCE DE UNA EMPRESA DEL SECTOR RETAIL CON ALGORITMOS DE ANÁLISIS CLUSTER"

PRESENTADO POR

LUCILA NOEMÍ CAYCHO HUAMANÍ

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

SUSTENTADO Y APROBADO ANTE EL SIGUIENTE JURADO:

……	……..
Dr. Fernando René Rosas Villena	Dr. Iván Dennys Soto Rodríguez
PRESIDENTE	ASESOR
……	……..
Dr. Rino Nicanor Sotomayor Ruiz	Dr. Raphael Félix Valencia Chacón
MIEMBRO	MIEMBRO

Lima – Perú

2023

A mis padres y hermanos por apoyarme a concluir mis estudios universitarios. Porque a pesar de las dificultades logramos salir adelante. A mi hija Lina, mi motor, motivo, mi luz del día a día.

AGRADECIMIENTOS

En primer lugar, a Dios por todas sus bendiciones, por acompañarme y por siempre brindarme claras respuestas de aquello que suele angustiarme.

A mi madre Rosa Huamaní, por ser el mejor ejemplo a seguir en mi vida. Gracias por motivarme a ser una mujer fuerte y luchadora.

A mi padre Freddy Caycho, por ser el mejor abuelo y padre que Lina pueda tener.

A mi hermano José, por dejarme siempre la meta más alta de superación profesional que cualquier estadística pueda tener.

A mi hermana Carol, por apoyarme, escucharme y aconsejarme en las decisiones que tomo.

A mi hija Lina, por entender a tu corta edad que todo el tiempo invertido en mis estudios darán frutos que nos permitirán cumplir nuestros sueños.

A mi compañero Kevin, gracias por permitirme ser parte de este hermoso equipo que somos.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Las empresas del sector retail cuentan no solo con canales de venta física sino también con canales de venta digital que permiten llegar a cualquier usuario con conexión a internet de forma sencilla, segura y directa para cubrir alguna necesidad, ya sea de consumo, de educación, de entretenimiento, de salud, etc. La empresa del sector retail en la cual me baso en el presente documento no fue la excepción, ya que en el período posterior al COVID-19, tuvo que repotenciar su canal de venta digital (ecommerce) de forma ágil, para continuar siendo una de las empresas con mayor presencia en el mercado peruano.

En la gerencia de Ecommerce, dentro de la dirección de Marketing, tenemos como objetivo principal brindar una excelente experiencia omnicanal de nuestros clientes, mediante la explotación de la información y experimentación constante. Nosotros consolidamos toda la información que recabamos de las plataformas de publicidad (como Facebook Ads, Google Ads, etc.), las plataformas de analítica digital (cómo Google Analytics, Hotjar, Google Optimize), las plataformas administradoras de activos digitales (cómo VTEX,), para convertirlos en insights que ayuden a la toma de decisiones.

En el presente trabajo se presentará un proyecto que tiene como objetivo armar una estrategia de CRM en base a la identificación de segmentos de los clientes digitales de productos de consumo masivo mediante el uso de algoritmos estadísticos. En una primera fase se consideró todo el proceso ETL que permite disponibilizar los datos de manera limpia y ordenada para luego realizar el análisis de clusterización mediante el algoritmo K-Means en el software gratuito R. Se identificaron 3 segmentos de clientes, a los cuáles se les comunicaron beneficios y ofertas afines a sus preferencias, logrando aumentar las redenciones, así como las interacciones con los canales de comunicación.

Palabras clave: Internet, Ecommerce, CRM, K-Means, Clusterización, Google Analytics.

ABSTRACT

Companies in the retail sector have not only physical sales channels but also digital sales channels that allow them to reach any internet-connected user easily, securely, and directly to fulfill various needs, whether they be related to consumption, education, entertainment, health, and more. The retail company that I am focusing on in this document was no exception. In the period following COVID-19, it had to swiftly enhance its digital sales channel (e-commerce) to continue being one of the most prominent players in the Peruvian market.

In the Ecommerce management under the Marketing department, our primary goal is to provide an excellent omnichannel experience for our customers by leveraging information and continuous experimentation. We consolidate all the data we gather from advertising platforms (such as Facebook Ads, Google Ads, etc.), digital analytics platforms (like Google Analytics, Hotjar, Google Optimize), and digital asset management platforms (such as VTEX) to transform them into insights that aid in decision-making.

This paper will present a project that aims to create a CRM strategy based on the identification of segments within the digital customers of fast-moving consumer goods, using statistical algorithms. In the initial phase, the entire ETL process was considered, which enables data to be made available in a clean and organized manner for subsequent cluster analysis using the k-means algorithm in the free software R.Four customer segments were identified, to whom relevant benefits and offers were communicated based on their preferences, resulting in increased redemptions as well as interactions through communication channels.

Keywords: internet, ecommerce, CRM, K-Means, clustering, google analytics

I. INTRODUCCIÓN

1.1 Problema Principal

En el segundo trimestre del 2022, los indicadores que permiten medir el desempeño de las campañas de CRM en el ecommerce tomaron una tendencia decreciente en comparación a lo alcanzado en el primer semestre. Se encontró que:

- El porcentaje de redención de beneficios (que se define al número total de clientes que utilizan el beneficio que se le asigna en base al segmento que pertenece sobre el total de clientes del segmento) disminuyó en -3 puntos porcentuales.

- El ROAS por segmentos (que es el ratio que resulta de la relación entre las ventas netas con los costos generados en la comunicación y en la redención de los beneficios) disminuyó en 20%.

Estos hallazgos generaron interés de redefinir las reglas de segmentación de los clientes cómo también los beneficios que se les brindaba a los clientes.

La segmentación que contaba la empresa se basaba en el comportamiento de compra de los clientes dentro del comercio electrónico mediante el análisis de RFM, el cual se construye en base a tres indicadores que conforman sus siglas: recencia (es el tiempo en días que pasan desde su último día de compra respecto a la fecha de corte del análisis), frecuencia (los días de diferencia entre su última compra con su penúltima compra) y monto (ticket promedio que el cliente maneja respecto a los últimos 6 meses previos a la fecha de corte del análisis). La clasificación tenía el objetivo de fidelizar a los clientes del canal digital, pero en primeramente conocer el comportamiento y las necesidades de nuestros clientes.

Cómo resultados de este análisis obtuvimos 4 clústers que se denominaron de la siguiente manera:

- Fidelizados: Aquellos clientes que compran de forma frecuente y que su última compra ha sido en el último mes del análisis.
- Esporádicos: Aquellos clientes que en el período de análisis sólo realizaron una compra, pero no la realizó en el último mes de análisis.
- Recuperación: Aquellos clientes que compraban de forma frecuente pero que no tienen alguna compra en el último mes de análisis.
- Nuevos: Aquellos clientes que tienen alta frecuencia de compra en el último mes del análisis.

  En la Figura 1 se mostrará la participación de los clientes respecto al total de ellos.

Figura. 1

Participación de clientes respecto al total.

Nota: Contenido adicional para representar a los tipos de clientes.

Esta clasificación dió buenos resultados en la época de pandemia, pues la presencia de los canales de venta digital daba indicadores altos de frecuencias que de cierta manera aseguraban la recompra en el sitio web. Sin embargo, ante el retorno de la presencialidad laboral y escolar, se observó que algunos clientes retornaban a la compra en tiendas presenciales. Es por esta razón que el equipo se retó a encontrar aquellos criterios de segmentación que no solo brinde luces de las frecuencias de compra sino también del interés en las categorías de productos para lograr mayor personalización en los beneficios que se les brinde.

1.2 Objetivos

1.2.1 Objetivo general

Segmentar a los clientes digitales que adquieren productos de consumo masivo en un ecommerce del sector retail, mediante la aplicación de un algoritmo de análisis clúster.

1.2.2 Objetivos específicos

- Determinar con qué análisis de clúster obtendremos segmentos que tengan relación con el negocio
- Identificar las categorías de producto de mayor consumo y preferencia en cada uno de los segmentos de clientes.
- Redefinir los beneficios que se les brindarán a cada uno de los segmentos de los clientes en las campañas de CRM.
- Mejorar los ratios de redención de los beneficios de cada uno de los segmentos de clientes hallados.

II. REVISIÓN DE LITERATURA

2.1. Comercio Electrónico

El comercio electrónico es el canal de venta que las empresas suelen implementar en Internet. TechTarget menciona lo siguiente en un artículo de Market Business News, “El e-commerce o comercio electrónico es la compra y venta de bienes y servicios o la transmisión de datos a través de una red electrónica, principalmente Internet”.

El comercio electrónico se fundamenta principalmente en la difusión y promoción de productos, artículos y servicios mediante estrategias de marketing digital. No obstante, existen distintas modalidades de comercio electrónico:

- **E-commerce B2B:** Esta modalidad implica transacciones entre empresas, es decir, Business to Business (B2B). Su enfoque no está dirigido al consumidor final.

- **E-commerce B2C:** Aquí, el comercio electrónico se orienta hacia el consumidor, es decir, Business to Consumer (B2C). El objetivo principal es llegar al usuario final.

- **E-commerce C2C:** Los consumidores son los actores principales, interactuando entre sí en un modelo conocido como Consumer to Consumer (C2C). Esta dinámica implica la traslación de actividades tradicionales, como ventas de garaje, intercambios entre vecinos o transacciones directas entre individuos, hacia el ámbito digital.

2.2. VTEX

Según Rafael Gallegos, “VTEX es una plataforma de comercio electrónico que ayuda a las empresas a vender sus productos en línea”. A través de esta esta plataforma las empresas pueden gestionar los productos, pueden recibir sus pedidos o las órdenes de compra, pueden realizan los cobros de las órdenes y pueden consolidar toda la información de sus usuarios o clientes.

VTEX ofrece una serie de plantillas prediseñadas en HTML y CSS que permite a las empresas reutilizarlas en un menor tiempo y con menor dependencia de equipos de desarrollo para su puesta en producción.

Así también, VTEX es una solución completa no solo para las áreas de desarrollo sino también para las áreas de marketing pues tiene integración directa con varias herramientas que permiten mejorar sus campañas de publicidad, las campañas de CRM, el posicionamiento ante búsquedas realizadas en Google (SEO).

2.3. CRM

Un sistema CRM, que significa Gestión de Relaciones con el Cliente (Customer Relationship Management, por sus siglas en inglés), se refiere a la administración de las interacciones de los clientes con nuestra marca. Se trata de un proceso que integra la estrategia empresarial con tecnología para supervisar todo el ciclo de vida del cliente. Al recopilar información detallada sobre clientes actuales y potenciales, proporciona una visión completa de cada interacción que tienen con la empresa.

Según Kate Legget de la revista Forrester, "Más del 27% de retención de clientes puede ser mejorado con un CRM potente". La implementación de un sistema CRM no se limita únicamente a una solución tecnológica; constituye el fundamento del marketing relacional contemporáneo, que se basa en la premisa de que el cliente es el activo más preciado de la empresa. Por ello, resulta crucial registrar toda la información relevante del cliente, permitiendo a las empresas ofrecer un servicio personalizado y adaptado a sus necesidades.

Una estrategia de CRM tiene como objetivo principal ofrecer a la empresa todo el material que pueda necesitar de cualquier cliente con el fin de obtener diferentes métricas de mercado para permitir la mejora de la estrategia comercial del mismo.

Algunas de las herramientas de CRM más conocidas son Salesforce, Pegasystems, Oracle, Microsoft. Tal como las podemos observar en la Figura.2 que muestra el top de herramientas según Gartner.

Figura. 2

Cuadrante Mágico para CRM

Nota: Elaborado en base a Gartner (June2021)

2.4. Salesforce Marketing Cloud

Es una plataforma de Salesforce que las pymes y grandes empresas utilizan para optimizar sus inversiones en estrategias de marketing digital a nivel profesional. Salesforce fue considerado como una de las mejores herramientas de automatización en el 2022, según los cuadrantes de Gartner.

Las compañías elaboran, adaptan y perfeccionan la trayectoria de sus clientes al comprenderlos mejor, evaluando sus resultados y optimizando su presupuesto de marketing. La tecnología que respalda esta herramienta facilita una gestión del marketing más eficiente y eficaz. Los equipos disponen de una solución que les permite crear experiencias multicanal, contactando al cliente en el momento y canal adecuados (correo electrónico, SMS, notificaciones push, redes sociales, anuncios, etc.), lo que conlleva a un aumento en la captación de clientes y las ventas. Con

Marketing Cloud, planificar, monitorear, analizar y tomar decisiones en tiempo real es más accesible que nunca. (Pklein,2019)

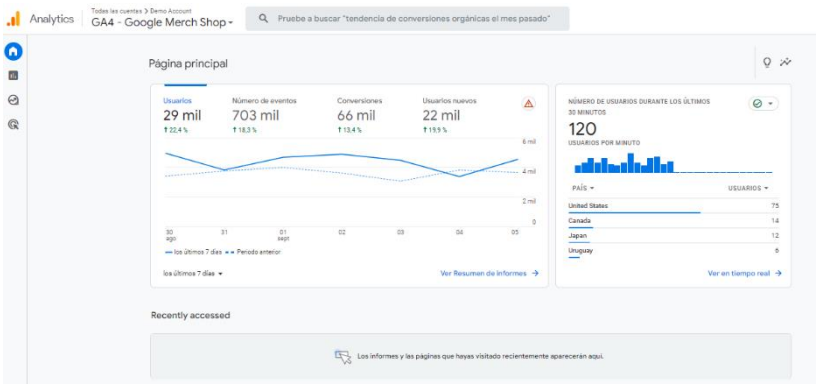
2.5. Google Analytics

Google Analytics es una plataforma que permite medir las interacciones de los usuarios dentro de una página web. Para lograr recoger esta información, se deben añadir códigos JavaScript en cada página o elemento de la web que se desee tener visibilidad.

Estos códigos no solo registran la actividad de los usuarios, sino que también recopilan información sobre su navegador, incluyendo detalles como el idioma configurado, el tipo de navegador (por ejemplo, Chrome o Safari), el dispositivo utilizado y el sistema operativo. Además, el código de seguimiento puede identificar la fuente de tráfico que llevó a los usuarios al sitio, ya sea a través de un motor de búsqueda, un anuncio publicitario al que hicieron clic o una campaña de marketing por correo electrónico.

Figura. 3

Vista Previa de Google Analytics (Ahora GA4)



Nota: Contenido elaborado por Analytics para principiantes (Google,2023)

Después de que Analytics procesa la información, esta se guarda en una base de datos que permanece inalterable.

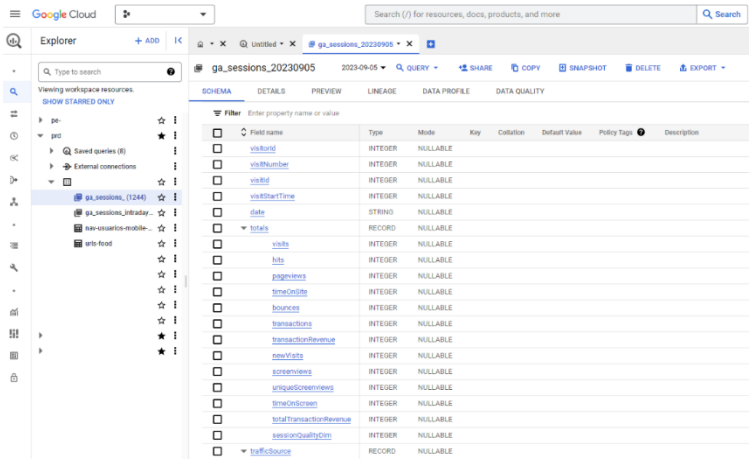
Algunas de las dimensiones y métricas de mayor uso son:

- Fuente/Medio: Permite identificar de qué medio y fuente el usuario está realizando su visita. Valores de ejemplo: Facebook/Pagado, Google/Pagado, Mail/Propio, etc.
- Páginas y Páginas de Destino: Nombre de las páginas o secciones que visitan en nuestra web. Ejemplo: “/home”, “/electrohogar/”, etc.
- Ubicación geográfica: País, Región, Ciudad, Distrito, etc.
- Categoría de dispositivo: Mobile, Desktop, Tablet
- Navegador: Navegador de ingreso del usuario. Ejemplo: Chrome, Firefox, etc.
- Sesiones: Cantidad de visitas que se tiene en determinado tiempo por cada navegador.
- Cantidad de páginas vistas: Cantidad de páginas vistas en el total de visitas en determinado tiempo.

Se pueden acceder a la información desde la misma herramienta o desde un repositorio situado en el módulo de Bigquery en Google Cloud Platform. Este último es muy usado en análisis más robustos de las áreas de marketing, ya que al estar en un repositorio en servicios cloud permite cruzar información con otras fuentes con volúmenes alto de información.

Figura. 4

Esquema de repositorio de Google Analytics en Bigquery (GCP)



Nota: Contenido adicional para ilustrar la vista previa de Google Analaytics en Bigquery.

2.6. Google Cloud Platform

Es una plataforma de recursos de computación en la nube que se utilizan para desarrollar, implementar y operar aplicaciones web. Algunas de las ventajas de usar computación en la nube son: son precios bajos, corto tiempo de procesamiento y consulta, colaboración simplificada entre usuarios para proyectos, entre otras. La plataforma de Google es una infraestructura completa para las empresas, en dónde les permite tener una gestión más eficiente y una seguridad digital completa.

La lista de funciones de GCP es extensa, así también cuenta con más de 100 productos y servicios para nuevas posibilidades y soluciones empresariales. Conoce algunos de ellos:

- Máquinas Virtuales
- Alojamiento de páginas web.
- Almacenamiento de archivos
- Servidores VPS
- Compartir y procesar datos

Estas y otras características que ofrece la plataforma son las que ayudan a elevar el potencial de escalabilidad de las empresas que dependen de la computación en la nube de Google Cloud Platform. Uno de las herramientas más usadas dentro de GCP es:

- Google BigQuery

Google Bigquery ofrece diversas capacidades centradas en la gestión de datos en múltiples entornos de nube, la recepción y análisis de información de manera ágil, y la creación de informes y paneles con total seguridad y confianza. A partir de esta herramienta podemos mejorar la toma de decisiones empresariales. “El análisis predictivo, por ejemplo, ayuda a proyectar los resultados comerciales con la integración de machine learning”. (Víctor Trafaniuc,2021)

2.7. Análisis Clúster

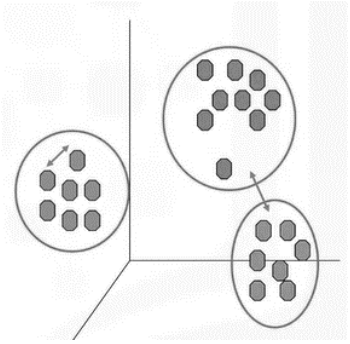
El análisis Clúster es una técnica multivariante cuyo objetivo es agrupar objetos formando conglomerados (clusters) de objetos con un alto grado de homogeneidad interna y heterogeneidad externa.

La similitud (homogeneidad) se determina usando una métrica de distancia, que mide la separación que tienen dos registros u observaciones entre sí. La distancia más popular entre dos vectores es la distancia euclidiana. Para calcular dicha distancia se debe restar uno del otro, elevarlo al cuadrado las diferencias, las sumamos y extraemos la raíz cuadrada (Peter Bruce,2022):

Figura. 5

Representación de Distancia Euclidiana

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2}$$



Nota: Elaborado por K-Means:Clustering. Medium (2020)

El Análisis Clúster tiene una importante presencia en muchas áreas de investigación, por ejemplo, en Astronomía (Agrupación de galaxias), Marketing (Segmentación de mercados, investigación de mercados), Ciencias Ambientales (Clasificación de ríos para establecer tipologías según la calidad de las aguas). (Gallardo,2012). Sin embargo, junto con los beneficios existen algunos inconvenientes pues es una técnica descriptiva, ateórica y no inferencial.

Los algoritmos de formación de conglomerados tienen dos categorías:

No Jerárquicos: Son conocidos como partitivos o de optimización, su objetivo es realizar solo una partición de los individuos en K grupos. Ello implica que el estadístico debe especificar a priori los grupos que deben ser formados, siendo ésta, posiblemente, la principal diferencia respecto a los métodos jerárquicos. Otra diferencia es que reside en que trabajan con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes.

Algunos ejemplos de estos algoritmos son:

- El método K-Means
- El Quick-Clúster análisis
- El método de Forgy
- Método Taxmap, etc.

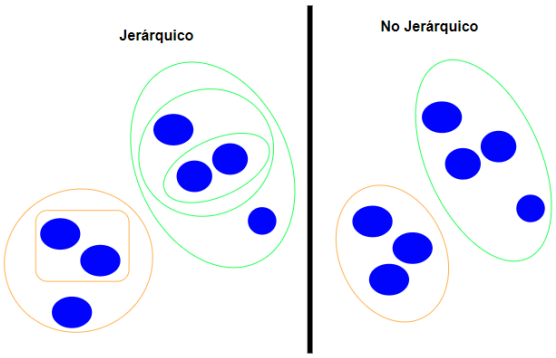
Jerárquicos: Estos métodos tienen por objetivo agrupar clusters para formar un nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función de distancia o bien se maximice alguna medida de similitud.

Independientemente del proceso de agrupación, hay diversos criterios para ir formando clusters, todos estos criterios se basan en una matriz de distancias o similitudes. Algunos que se destacan son:

- Método de amalgamiento simple
- Método del amalgamiento completo
- Método de centroide, etc.

Figura. 6

Análisis Clúster Jerárquico vs No Jerárquico



Nota: Gráfico adaptado de www.diegocalvo.es/cluster.

2.8. K – Means

El objetivo de K-Means es agrupar los datos similares para descubrir patrones que a simple vista se desconocen. Para hallarlos, busca un número fijo de (K) clusters en la base de datos.

Este término (K) representa el número de centroides (centro de clúster) que queremos encontrar en la base de datos (Jin, X., Han,2011). Para saber qué K tenemos que seleccionar, se puede utilizar alguna de las siguientes técnicas:

Tabla 1

Criterios para seleccionar K Óptimo.

Criterios	Descripción
Método de Codo	Consiste en encontrar el valor óptimo de k. Para hacerlo se grafican los valores de k junto con la suma de los errores cuadrados (SSE) para cada valor de k. El SSE es la suma de distancias de cada punto al centroide del clúster. A medida que aumenta k, el SSE disminuye, pero a un ritmo cada vez menor. El punto en el que SSE disminuye más lentamente se conoce como el codo y es el punto óptimo para elegir K. (Sanz,2023)
Método de la Silueta	Este método estima la distancia media entre agrupaciones, así como mide qué tan buena es la agrupación de cada observación. Este método permite gráficamente evaluar que tan buena puede ser la cantidad de agrupaciones en base a algunos parámetros.
Dendograma	Es un tipo de demostración gráfica en forma de árbol que organiza y agrupa los datos en subcategorías según similitud; dada por alguna medida de distancia. (Moya,2016)

Nota: Contenido elaborado para resumir criterios de elección de k-óptimos.

Este tipo de algoritmos se emplean en diversas áreas:

- **Segmentación de clientes:** Se utilizan para dividir a los clientes en grupos distintos según sus características o comportamientos. Esto resulta útil para realizar campañas de marketing más personalizadas o para tomar decisiones comerciales.
- **Clasificación de texto:** Se emplean que para categorizar documentos o artículos en diferentes grupos según su contenido.
- **Detección de anomalías:** Son útiles para identificar patrones irregulares dentro de un conjunto de datos y señalar posibles problemas o errores.

“El algoritmo de K-Means es una herramienta útil para la agrupación y descripción de datos. Es una de las opciones más populares de uso ante problemas de clasificación”. (Ramírez, 2023).

2.9. K-Prototype

Según Audhi Aprilliant, ante la presencia variables cualitativas en una base de datos que también presenta variables cuantitativas, recomienda usar el algoritmo K-Prototype para encontrar grupos de unidades que se relacionen entre sí. A diferencia del algoritmo de K-Means tradicional, que funciona mejor con datos numéricos, el algoritmo k-prototipo permite trabajar con datos mixtos, lo que lo hace más adecuado para situaciones en las que las variables categóricas también desempeñan un papel importante en la segmentación de los datos.

En el algoritmo k-prototipo, se busca agrupar los datos en k clústeres, donde k es un número predefinido. El algoritmo busca optimizar la asignación de datos a clústeres de manera que minimice la distancia entre los puntos dentro de un clúster y los centroides del clúster, tanto para las variables numéricas como para las categóricas.

El proceso general del algoritmo k-prototype es similar al del K-Means:

- Inicialización: Se eligen k centroides iniciales, que pueden ser tanto datos reales como aleatorios.
- Asignación: Cada punto de datos se asigna al clúster cuyo centroide (que incluye tanto valores numéricos como categóricos) es el más cercano en términos de distancia combinada.
- Actualización: Los centroides de cada clúster se recalculan utilizando la media para las variables numéricas y la moda para las variables categóricas de los puntos asignados al clúster.
- Reasignación y Actualización: Los pasos de asignación y actualización se repiten iterativamente hasta que los centroides convergen y los puntos ya no cambian de clúster de manera significativa.

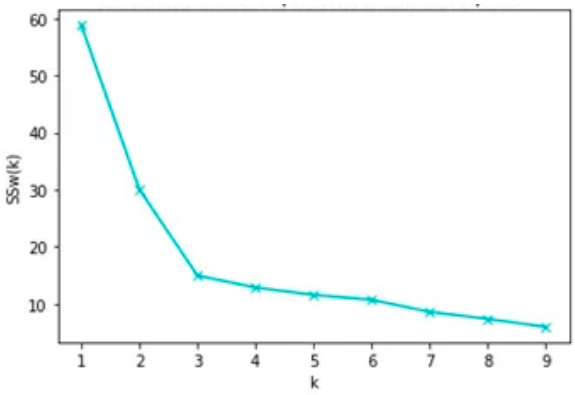
El algoritmo k-prototipo es útil cuando se trabaja con conjuntos de datos que contienen múltiples tipos de variables y se desea realizar una segmentación significativa basada en las características tanto numéricas como categóricas de los datos.

2.10. Representaciones gráficas de métodos para obtener el K óptimo

2.10.1. Método de codo

Figura. 7

Ejemplo del gráfico del método del Codo

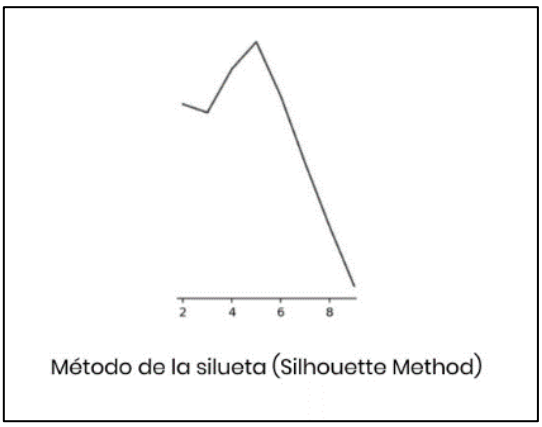


Nota: Contenido adicional adaptado de Medium.com (Ramirez,2018)

2.10.2. Método de la Silueta

Figura. 8.

Ejemplo de gráfico del método de la Silueta

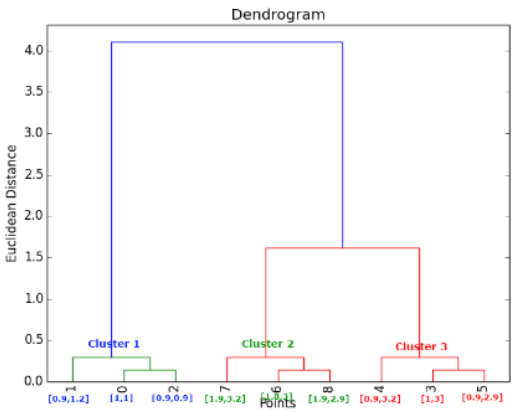


Nota: Contenido adicional adaptado de Medium.com (Ramirez,2018)

2.10.3. Dendograma

Figura. 9

Ejemplo de gráfico de un dendograma.



Fuente: Selección del número óptimo de Clusters (Moya,2016)

III. DESARROLLO DEL TRABAJO

En base al objetivo principal propuesto, este trabajo busca redefinir los criterios de clasificación de los clientes digitales de consumo masivo mediante un algoritmo de análisis de clúster en base a la información de las transacciones que realizan estos clientes en el ecommerce, así como también con información de indicadores de comportamiento digital. Se consideró como período de análisis toda la información de junio 2022 a julio del 2023.

Una gran ventaja para el trabajo es que todas las fuentes de información se encuentran en la nube de Google (Google Cloud Platform), específicamente en la herramienta de Bigquery. Bigquery es el repositorio principal de información de la empresa. Aquí es dónde con consultas SQL armamos y ordenamos aquellas fuentes de información que mediante la llave de unión "email" logremos poblar la mayor cantidad de información del cliente. Las fuentes de información finales fueron:

- Fuente de información transaccional (Origen Vtex): En dónde encontramos el detalle de todas las órdenes de compras de los clientes del período anteriormente mencionado. Aquí podemos encontrar variables cómo el producto de compra, la categoría a la que pertenece, la cantidad de transacciones que el cliente realiza, el ticket de su compra por transacción, el método de pago preferente, entre otras.

- Fuente de información de comportamiento digital (Origen Google Analytics): En dónde encontramos métricas como cantidad de visitas al ecommerce, cantidad de páginas vistas, cantidad de visitas que realiza desde su celular, entre otras.

El conjunto de datos final contiene información agrupada de 245,428 clientes que inicialmente contaban con un total de 24 variables, terminaron agrupándose por reglas de negocio en 14 variables, las cuáles se detallaran en la siguiente Tabla 2.

Tabla 2.

Descripción de variables

Variable	Tipo de variable	Descripción
documento	cualitativo	Número de documento de identidad.
meses_compra	cuantitativa	Cantidad de meses en los que ha realizado alguna compra en el ecommerce en el período de análisis.
pageviews	cuantitativa	Cantidad total de páginas vistas en el ecommerce en el período de análisis.
sesiones_mobile	cuantitativa	Visitas totales en el ecommerce desde un dispositivo mobile en el período de análisis.
sesiones_desktop	cuantitativa	Visitas totales en el ecommerce desde un dispositivo desktop en el período de análisis.
cuidadopersonal_belleza	cuantitativa	Monto total de compra (S/.) de las categorías Cuidado Personal, Belleza y Accesorios.
frescos	cuantitativa	Monto total de compra (S/.) de las categorías Frutas, Verduras, Carnes y Congelados.
abarrotes_packs	cuantitativa	Monto total de compra (S/.) de las categorías Abarrotes y Packs de Abarrotes.
lacteos_fiambres	cuantitativa	Monto total de compra (S/.) de las categorías Lácteos, Huevos y Desayunos.
comida_preparada	cuantitativa	Monto total de compra (S/.) de las categorías Pollo Rostizado, Pastelería y Panadería.
mundo_infantil	cuantitativa	Monto total de compra (S/.) de las categorías Bebé e Infantil, Juguetes, Juegos y Librería.
mascotas	cuantitativa	Monto total de compra (S/.) de la categoría Mascotas
localidad preferente	cualitativa	Localidad preferente (Lima o Provincia)
tipoentrega_preferente	cualitativa	Tipo de entrega preferente en el período de análisis: Envío Express (DE), Envío a Domicilio (DD) o Retiro en tienda (RT).
mediopago_preferente	cualitativa	Método de pago preferente Tarjeta de crédito, Billetera virtual, Tarjeta de débito.

Nota: Contenido adicional que muestra las variables utilizadas.

En un primer análisis descriptivo del conjunto de datos mencionado anteriormente, se observó que aprox. el 70% de los clientes suelen realizar una única compra en los últimos 12 meses de corte (meses_compra=1). Por lo cual se asumió a priori que ellos pertenecerán a un clúster inicial el cuál excluimos del análisis clúster y al cuál denominamos “Esporádicos”.
El nuevo universo de clientes por segmentar tuvo como estructura final lo que se muestra en la Tabla 3.

Tabla 3.

Conjunto de datos final para el análisis de clusters

documento	meses_compra	frescos	mascotas	…	flag_tarjetacredito
48267449	2	S/.1200	S/100		1
45345402	5	S/10000	S/23000		0
00769843	10	S/23000	S/.1200		1

Nota: Contenido adicional que muestra del conjunto de datos final.

En un siguiente análisis descriptivo de las variables se detectó la presencia de valores nulos para las variables pageviews, sesionesmobile y sesionesdesktop. A las cuáles al presentar el 2% de valores nulos se les imputó el promedio de aquellos que si cuentan con la información. Este criterio se realizó teniendo en cuenta las recomendaciones que se suele brindar en las áreas de analíticas de las empresas. Las cuáles se resumen a continuación en la tabla 4.

Tabla 4.

Criterios de tratamiento de valores nulos

Tipo de Variables		**<=10%**	**>10%**
Variables Continuas		Promedio	Dicotomizar
Variables Categóricas	Ordinal	Mediana	
	Nominal	Moda	

Nota: Tabla Adaptada de Medium.com

Al tener variables continuas y variables categóricas, se optó por realizar dos análisis clúster de tal forma que nos permita tener mayores opciones en la toma de decisiones.
Decidimos ejecutar el algoritmo K-Means mediante el software R y el algoritmo K-Medoides mediante Python. Esto debido a que el software R presentaba algunos limitantes (memoria RAM, versión de R, etc.) con las librerías que permiten obtener el k-optimo del algoritmo K-Medoides.

3.1 Algoritmo K-Means

Empezamos usando el algoritmo K-Means, en dónde se ingresaron las variables numéricas y las variables categóricas (convertidas a dummies), ya estandarizados (mediante la función scale en el software R.
Luego de ello, era necesario definir el número de grupos óptimos para poder identificar las agrupaciones de clientes. En este paso utilizamos la función NbClust del paquete NbClust del software R. Esta función brinda al usuario el K-óptimo de agrupaciones en base al índice D que busca el pico significativo de diversas simulaciones de cálculos de una variedad de medidas de distancias y de la aplicación de métodos de validación como son el método de la silueta, codo, etc.

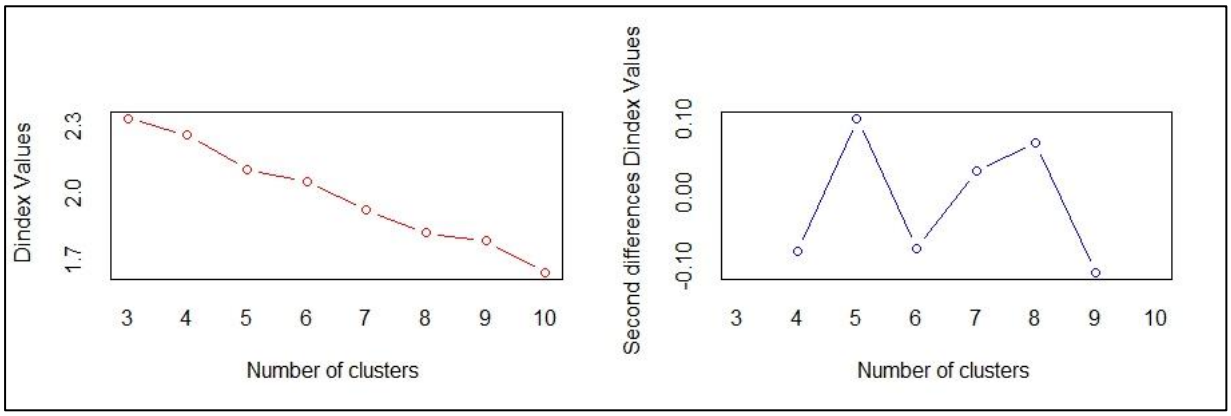
Comando en R:

```
library(NbClust)

nc <- NbClust(datos_reg2, min.nc=3, max.nc=10, method="kmeans")
```

El output de dicha función se muestra en la siguiente Figura.10

Figura. 10

Gráfico de índice D de la función NbClust del software R



Nota: Resultados con el software R, 2023

La Figura. 11 mostrará el resultado de la función en la consola de R, en dónde se visualiza que entre todas las simulaciones se determina que 3 es el número óptimo de agrupaciones para el proyecto.

Figura. 11

Resultado de la función NbClust en la consola de R.

```
*******************************************************************
* Among all indices:
* 8 proposed 3 as the best number of clusters
* 6 proposed 4 as the best number of clusters
* 7 proposed 7 as the best number of clusters
* 2 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

Nota: Resultados con el software R, 2023

3.2 Algoritmo K-Prototype

En Python mediante el paquete K-Prototype, se logró analizar el conjunto de datos completo (Variables Categorías y Variables Continuas) para complementar y poder determinar el número de agrupamientos de los clientes final. Al igual que el algoritmo K-Means es necesario determinar cómo paso inicial el número de K óptimo de agrupamiento. Por lo que en Python se armó una función que usa el método "Elbow" y la simulación de la aplicación del algoritmo para k=2,3,4,5,6,7,8,9, dónde determinamos qué k es el ideal.

Figura. 12

Script en Python para k óptimo del algoritmo K-Prototype

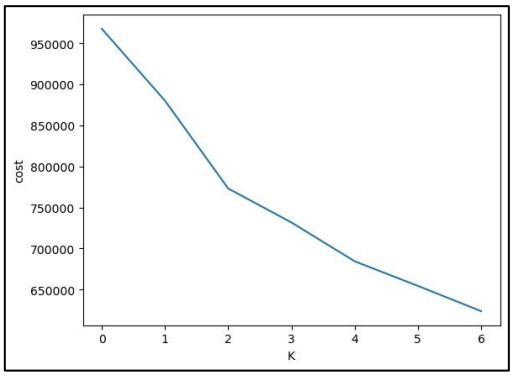
```
cost = []

for num_clusters in list(range(2,9)):
    kproto = KPrototypes(n_clusters=num_clusters, init='Huang', random_state=123)
    kproto.fit_predict(df2, categorical=catColumnsPos)
    cost.append(kproto.cost_)

plt.plot(cost)
plt.xticks = range(2,9)
plt.xlabel('K')
plt.ylabel('cost')
plt.show
```

Nota: Vista previa de resultados de jupyter Python (2023)

La función cost nos da la figura en dónde podemos intuir que el k óptimo es 4. Sin embargo, al consultar por el valor específico este nos da que 4 es el número de agrupamiento óptimo.

Figura. 13

Gráfico K óptimo en K-prototype



Nota: Gráfico adicional para representar k-prototype con jupyter, 2023

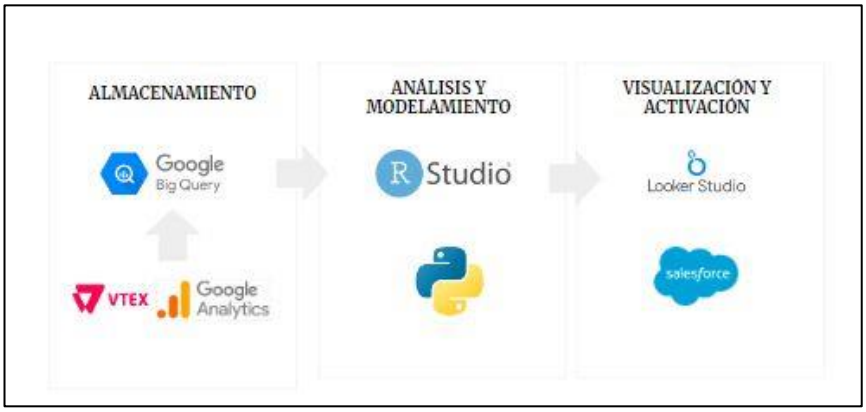
El resultado de los análisis obtenidos en R y Python, se enviaron nuevamente a Bigquery para su lectura y visualización de los clusters hallados. Esta acción nos permitió quedarnos con el algoritmo K-Means ya que vimos una mejor distribución de clientes por los 3 agrupamientos recomendados.

Posterior a ello, se enviarán las bases de los nuevos segmentos a Salesforce Marketing Cloud para activar las campañas de marketing digital en los canales de mail, sms, WhatsApp, por Facebook o por el buscador de google.

La arquitectura de herramientas propuestas para este proyecto se resume en la **Figura 14.**

Figura. 14.

Arquitectura de herramientas del proyecto.



Nota: Gráfico adicional de las herramientas utilizadas (2023)

IV. RESULTADOS Y DISCUSIÓN

Basándonos en la discusión previa, se evaluaron dos algoritmos de segmentación debido a la naturaleza de las variables recopiladas. No obstante, se notó que la distribución de los grupos utilizando la técnica k-Prototype no presentaba la misma simetría que la obtenida mediante K-Means (véase la Tabla 5.). Dado que la interpretación de los grupos de clientes resultaba complicada con el enfoque k-prototype, el equipo optó por seleccionar el algoritmo K-Means para la segmentación de los clientes, es decir, quedarnos con 3 grupos de clientes.

Tabla 5.

Comparación de Agrupaciones K-Means vs. K-Prototype

Grupo	K-Means		K-Prototype	
1	31,680	42.5%	32,477	43.6%
2	18,583	24.9%	10,155	13.6%
3	24,218	32.5%	31,255	42%
4	-		594	0.8%

Nota: Cuadro adicional de los resultados obtenidos (2023)

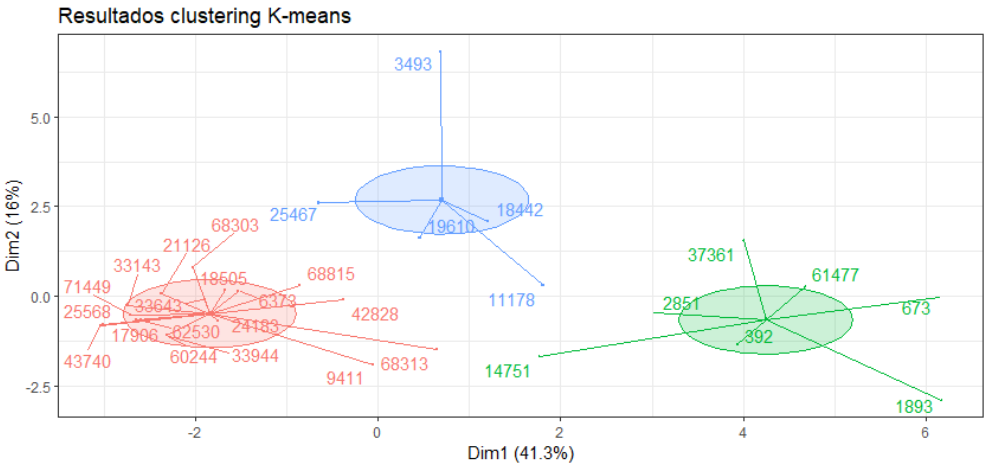
Para visualizar el agrupamiento de los datos, se optó por escoger una muestra de ellos que nos permita evidenciar las diferencias entre grupos y las similitudes en los mismos grupos. Esto se graficó mediante el siguiente script en R.

Comando en R:

```
Library(factoextra)

library(cluster)

km_clusters <- kmeans(x = d2f, centers = 3, nstart = 50)
fviz_cluster(object = km_clusters, data = df, show.clust.cent = TRUE,

          ellipse.type = "euclid", star.plot = TRUE, repel = TRUE,

          pointsize=0.5,outlier.color="darkred") +

 labs(title = "Resultados clustering K-means") +

 theme_bw() +  theme(legend.position = "none")
```

Figura. 15

Representación de resultados



Nota: Gráfico de representación de grupos de clientes (2023)

Al momento de describir los clústers se evidenciaron mayores diferencias entre las características, algunas como las que veremos en la Tabla 6.

Tabla 6.

Descriptivo parte 1 por segmentos de clientes obtenidos por K-Means

Grupo	Clientes	%Clientes	Promedio de meses de compra	% de sesiones desde Mobile	Promedio de páginas vistas
1	31,680	42.5%	4.5	57%	597.8
2	18,583	24.9%	2.7	65%	311.5
3	24,218	32.5%	3.7	73%	715.5

Nota: Cuadro adicional de los resultados obtenidos (2023)

Tal como se puede apreciar, la distribución de los clientes en los diferentes grupos muestra una casi perfecta simetría. El primer grupo alberga a la mayor proporción de clientes (42.5%), mientras que el segundo grupo tiene la menor cantidad de clientes (24.9%). Además de esto, se pudieron identificar las siguientes observaciones:

- La agrupación 1 (la cual contiene la mayor cantidad de clientes) contiene a aquellos clientes que tienen mayor presencia de compra (aprox. 4.5 meses de compra promedio en los últimos 12 meses de análisis). Luego, si bien vemos que las agrupaciones 2 y 3 tienen promedios de meses de compra similares, la agrupación 3 tendría una mayor frecuencia de compra que la agrupación 2.
- Los clientes de las tres agrupaciones realizan sus visitas al ecommerce desde su dispositivo Mobile como en desktop. Sin embargo, se puede ver cierta preferencia de uso de Mobile con mayor notoriedad en la agrupación 3 ya que el 73% de sus visitas son desde Mobile.

- Coincidentemente la agrupación 3 es también la que presenta mayor cantidad de vistas de páginas promedio respecto a los demás grupos. Esto nos puede dar luces que este tipo de cliente suele tener mayor exploración dentro del ecommerce, es decir, consume mayor cantidad del contenido que se muestra versus los demás grupos de clientes. En cambio, se observa que los clientes de la agrupación 2 tiene menor promedio de páginas vistas promedio, es decir, suele navegar menos que los demás grupos de clientes.

De la misma manera, se analizó la información de las variables que nos ayudan a definir con mayor precisión el interés de los clientes por ciertas categorías de productos. Esto también se muestra en la Tabla 7.

Tabla 7.

Porcentaje de consumo de categorías de productos por grupo de clientes.

Categoría de Productos	Grupo 1	Grupo 2	Grupo 3
Cuidado Personal y Accesorios	7.7%	8.6%	3.5%
Frescos	19.7%	9.2%	3.8%
Abarrotes y Packs	30.9%	37.4%	49.5%
Lácteos y Fiambres	28.5%	22.7%	28.4%
Comida Preparada	6.4%	8.1%	6.0%
Mundo Infantil	4.1%	10.4%	7.6%
Mascotas	2.9%	3.72%	1.0%

Nota: Cuadro adicional de los resultados obtenidos (2023)

En base al cuadro resumen, se analizaron los grupos encontrando lo siguiente:

- En los tres grupos predomina el consumo de productos de categorías de abarrotes y de packs de productos (en su mayoría de abarrotes) así como también los productos de las categorías Lácteos, Huevos y Fiambres. Ocupando casi el 60% de su monto de compra promedio.

- En el grupo 1 se puede observar que la categoría que gana protagonismo adicional a las mencionadas es la categoría Frescos. Aquí encontramos productos como frutas, verduras, congelados, entre otras. Con esta categoría hemos cubierto aproximadamente el 80% del consumo promedio de este grupo de clientes.

- En el grupo 2, las categorías Mundo Infantil y Cuidado Personal presentan también un porcentaje atractivo de participación en el consumo promedio de los clientes.

- En el grupo 3, con las categorías de Comidas Preparadas y Mundo infantil se describió aproximadamente el 92% del consumo promedio de esta agrupación.

- La categoría de Mascotas, en dónde se encuentran los alimentos y accesorios de las mascotas, tiene menor participación en el monto de consumo promedio de los grupos. Sin embargo, se observa que el grupo 2 tiene mayor participación de consumo en comparación a los demás grupos.

Continuando con el análisis del proyecto, observaremos la distribución de los clientes en cada grupo en relación de las variables categóricas Localidad Preferente, Tipo de Entrega Preferente y Método de Pago Preferente en la siguiente tabla:

Tabla 8.

Porcentaje de clientes según variables categóricas del conjunto de datos.

Variables Categóricas		Grupo 1	Grupo 2	Grupo 3
Localidad Preferente	Lima	85%	77%	67%
	Provincia	15%	23%	33%
Tipo de Entrega Preferente	Envío a domicilio	97%	2%	1%
	Retiro en tienda	1%	78%	90%
	Envío Express(90min)	2%	20%	9%
Método de Pago Preferente	Billetera Virtual	0.1%	12%	0.1%
	Tarjeta de Crédito	42%	0.1%	98%
	Tarjeta de Débito	57.9%	87.9%	1.9%

Nota: Cuadro adicional de los resultados obtenidos (2023)

En base a esta información podemos tener mayor detalle de las principales similitudes y diferencias que presentan los grupos. Algunos son:

- Los tres grupos tienen una gran participación de clientes que realizan sus pedidos para Lima Metropolitana. Sin embargo, se observa que en el grupo 3, el 33% de clientes realizan sus pedidos para provincia.

- El grupo 1 tiene mayor preferencia de solicitar el envío de sus productos a sus domicilios.

- Se observa que el grupo 2 y el grupo 3 prefiere retirar sus pedidos en las tiendas físicas. También se puede observar que se ve una participación interesante del 20% de clientes del grupo 2 que usan el envío express. Generalmente este tipo de entrega se da en aplicaciones móviles.

- Respecto al método de pago preferente, se puede evidencia que en el grupo 1 y en el grupo 2, los clientes suelen pagar con tarjetas de débito.

- El grupo 3, a diferencia de los otros grupos, tiene una preferencia elevada de utilizar la tarjeta de crédito de la empresa para realizar el pago de su pedido.

Luego de conocer un poco más de los grupos hallados. Se definieron a nuestros tres grupos bajo los nombres de Exploradores Digitales, Familias y Prácticos. En la Tabla 9, se dará mayor detalle de ello.

Tabla 9.

Definición de Grupos de Clientes

Grupo	Nombre	Características
1	Familias Prácticas	• Son clientes con mayor frecuencia de compra en el año. • Se evidencia que tienen facilidad de usar plataformas digitales ya que realizan compras desde sus dispositivos móviles y suelen revisar de forma regular el contenido del ecommerce (Por la cantidad de páginas que visualiza). • Las categorías de consumo conforman una canasta básica de alimentación común. (Abarrotes, Lácteos, Fiambres, Frescos y Congelados). • Piden que sus pedidos sean entregados en Lima Metropolitana y suelen pagar con tarjeta de débito.

2	Familias Tradicionales	• Son clientes con menor frecuencia de compra en el año. Se puede intuir que usan el ecommerce para realizar compras puntuales de hasta 3 meses en un año. • Manejan con facilidad plataformas digitales pues ingresan en la mayoría de veces por sus dispositivos móviles (65%) y al ver menor cantidad de páginas dentro del ecommerce se puede asumir que ya conocen la ruta con mayor precisión para realizar una compra. • Estos clientes complementan las categorías de consumo básico de supermercado con categorías de cuidado personal y con productos relacionados a bebés y niños. Así como también compran (en menor proporción) productos para mascotas. • Estos clientes prefieren retirar sus pedidos en las tiendas más cercanas a ellos.
3	Exploradores Digitales	• Estos clientes tienen una presencia regular en el ecommerce (3.7 meses de compra). • Son los que más usan sus dispositivos móviles para realizar una compra (73%) y los que más visitan el contenido de la web. • Los clientes de este grupo compran adicional a los productos de la canasta común de supermercado (Abarrotes y Frescos) productos relacionados a bebés y niños. Sin embargo y a diferencia del grupo 2, estos clientes suelen comprar también comidas preparadas (Por ejemplo, pollo rostizado, panes, pasteles, etc.). • Tienen como método de entrega preferente el Retiro a Tienda y suelen pagar sus pedidos con tarjeta de crédito de la empresa.

Nota: Cuadro adicional de los resultados obtenidos (2023)

V. CONCLUSIONES

El proyecto nos permitió identificar tres grupos de clientes mediante el uso del análisis clúster de “K-Means” mediante el software R. Esto gracias a la integración de fuentes de información de comportamiento del cliente en el ecommerce con la fuente de información transaccional mediante el uso de plataformas cloud, que aceleran toda la etapa de procesamiento a un menor costo.

Si bien tuvimos como reto utilizar otro tipo de análisis clúster debido a la presencia de variables categóricas, se identificó que aún esta información no es del todo representativa o diferente para lograr diferencias entre los clientes.

Por otro lado, el uso de las variables referentes a las categorías de productos (cuyo origen es la plataforma VTEX), nos permitió entender con mayor precisión los intereses de los clientes, por lo que podremos brindarles beneficios personalizados. Las variables de comportamiento digital (cuyo origen es la herramienta Google Analytics) nos brindaron insights valiosos, tales cómo entender qué tan amigables los clientes a las plataformas digitales son, qué tanto suelen revisar el contenido de la página web, qué tan frecuente es su presencia a lo largo de 12 últimos meses. También nos permitió conocer que el retorno presencial a los centros educativos y laborales generó un cambio en la preferencia de entrega del pedido, ya que vimos que dos de los tres grupos prefieren recoger en una tienda sus pedidos.

Finalmente, los grupos de clientes diferentes se nombraron en base a las características encontradas como “Exploradores Digitales”, “Familias Prácticas” y “Familias Tradicionales”. Cada grupo será enviado a Salesforce Marketing Cloud para activar campañas que promocionen ofertas comerciales a fines a su comportamiento e interés.

VI. RECOMENDACIONES

En los primeros análisis descriptivos de los clientes se registraron que casi el 70% de ellos, compran dentro de un solo mes de los últimos doce meses de análisis. Es probable que estos clientes solo realicen compras por el ecommerce por alguna campaña masiva de las plataformas digitales como un "Cyber", en dónde los precios suelen ser bastantes atractivos para el público en general. Se recomienda analizar el comportamiento de estos clientes para identificar oportunidades de recompra en la web.

Para tener una segunda etapa de segmentación, es necesario contar con información sociodemográfica del cliente. En el ecommerce, los clientes no suelen ingresar correctamente el número de su documento de identidad y este valor es necesario para lograr enriquecer su perfilamiento usando como recurso complementario los datos de las tiendas físicas. Por lo que es necesario que Vtex permita establecer como identificador único del cliente, al documento de identidad en lugar del correo electrónico.

Por lineamientos de la empresa y del área de seguridad de la información, es recomendable utilizar Python en lugar de R para el análisis de datos en la nube de Google Cloud Platform. Por esta razón y porque se identificó que en R algunas de las librerías con mayor potencial para análisis cluster de variables mixtas, están desfasadas o te piden una versión de R diferente a la actual se sugiere migrar los scripts a Python mediante el uso de plantillas de jupyter.

Finalmente, se recomienda que las comunicaciones hacia los clientes por medios propios sean de acuerdo a sus intereses de compra y no bajo supuestos que en su mayoría son sugeridos por los equipos comerciales de la empresa.

VII. REFERENCIAS BIBLIOGRÁFICAS

- Aprilliant, A. (2023, 7 enero). The K-prototype as clustering algorithm for mixed data type (Categorical and Numerical). *Medium*. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

- Ramirez, J. (2021, 7 diciembre). K-means: Elbow method and silhouette - Jonathan Ramirez - medium. *Medium*. https://medium.com/@jonathanrmzg/k-means-elbow-method-and-silhouette-e565d7ab87aa

- Cruz, V. (2021). What is E-commerce? Definition and examples. Market Business News. https://marketbusinessnews.com/financial-glossary/e-commerce/

- Jin, X., Han, J. (2011). K - Clustering de Medoides. En: Sammut, C., Webb, GI (eds) Enciclopedia de aprendizaje automático. Springer, Boston, MA.

- Leggett, K. (2020b). Top CRM Trends For 2020. Forrester. https://www.forrester.com/blogs/top-crm-trends-for-2020/

- ¿Qué es VTEX y cómo funciona? (s. f.). Gluo. https://www.gluo.mx/blog/que-es-vtex-y-como-funciona

- Pklein. (2019). ¿Qué es Salesforce Marketing Cloud? ¡Respondemos a tus dudas! www.wearemarketing.com. https://www.wearemarketing.com/es/blog/que-son-salesforce-marketing-cloud-assets.html

- Carter, R. (2022). Gartner Magic Quadrant for Sales Force Automation Platforms 2022. CX Today. https://www.cxtoday.com/crm/gartner-magic-quadrant-for-sales-force-automation-platforms-2022/

- Trafaniuc, V. (2022, June 28). Descubre qué es Google Cloud Platform y sus ventajas. Maplink Blog. https://maplink.global/blog/es/que-es-google-cloud/

- Práctica 8 | Estadística. (n.d.). http://wpd.ugr.es/~bioestad/guia-spss/practica-8/

- Calvo, D. (2018). Clúster jerárquicos y no jerárquicos. Diego Calvo. https://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/

- Sanz, F. (2023). Algoritmo K-Means Clustering – aplicaciones y desventajas. The Machine Learners. https://www.themachinelearners.com/k-means/

- Ramírez, L. (2023, January 5). Algoritmo k-means: ¿Qué es y cómo funciona? Thinking for Innovation. https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/

ANEXO I

K-MEANS: CORRIDAS EN R

```
install.packages("bigrquery")

library(bigrquery)

library(DBI)


con <- dbConnect(

  bigquery(),

  project = "project",

  dataset = "dataset",

  billing = "project"

)


sql<-"SELECT * FROM `project`"

my_results<-dbGetQuery(con,sql)

##Lectura de datos

data<-as.data.frame(my_results)

summary(data)
```

```
# Imputación de valores nulos (Menor de 10%)

promedio_pageviews<-mean(data$PAGEVIEWS, na.rm=TRUE)

promedio_sesionesmobile<-mean(data$SESIONES_MOBILE, na.rm=TRUE)

promedio_sesionesdesktop<-mean(data$SESIONES_DESKTOP, na.rm=TRUE)

data$PAGEVIEWS[is.na(data$PAGEVIEWS)] <- promedio_pageviews

data$SESIONES_MOBILE[is.na(data$SESIONES_MOBILE)] <-
promedio_sesionesmobile

data$SESIONES_DESKTOP[is.na(data$SESIONES_DESKTOP)] <-
promedio_sesionesdesktop

#Excluimos clientes que solo compran 1 vez en el año y nos quedamos con clientes de la categoría de supermercado

data <- subset(data, !(meses_compra == 1))

dataF <- subset(data, (tipo_prod == 'Food'))

dataF_index<-data[, c(1,2)]

dataF_num<-data[, c(3,7,8,9,10,12,13,14,15,18,19,22,23,24,25,26)]

dataF_cat<-data[, c(27,28,29)]

datacompletaF<-cbind(dataF_index,dataF_num,dataF_cat)

#Análisis Exploratorio de Valores

table(data$Moda)

table(data$Tecnologia)

 table(data$Deportes)
```

```
 table(data$Mundo_Infantil)

 table(data$Mascotas)

#Normalizamos el data.frame

df <- scale(dataF_num)

df<-as.data.frame(df)

#Hallamos el K-Óptimo

memory.limit(size=90000)

n<-dim(df)

n

set.seed(12345)

index<-sample(1:n,10000)

datos_reg2<-df[index,]

library(NbClust)

set.seed(123)

nc <- NbClust(datos_reg2, min.nc=3, max.nc=10, method="kmeans")

#Resulta k<-4

k_optimal <- 4

# Ajustar KMeans con el valor óptimo de K

kmeans_model <- kmeans(df, centers = k_optimal)
```

```
# Obtener las asignaciones de clusters para cada muestra

cluster_assignments <- kmeans_model$cluster


# Agregar la columna de asignaciones al DataFrame original

df$cluster <- cluster_assignments


# Mostrar un resumen de las asignaciones

table(cluster_assignments)


data_num$cluster<-cluster_assignments
```