UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL
INGENIERÍA EN GESTIÓN AMBIENTAL

Propuesta de segmentación de clientes aplicando técnicas de Machine Learning para mejorar la experiencia de compra mediante un sistema de recomendación de productos de Tottus

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial,

Obtener el título profesional de Ingeniero en Gestión Ambiental

AUTORES
Atencio Manyari Stefany Anyela
De la Rosa Flores Harold
Hilario Maravi Sayuri
Navarro Huarcaya Margareth
Rosas Vivanco Dianaluz Milagros

ASESOR
Junior John Fabián Arteaga
ORCID Nº 000-0001-9804-7795

Octubre, 2022

Resumen

Actualmente, el constante cambio en los factores externos como la tecnología, el mercado, y ahora la pandemia global están obligando a las empresas del sector *retail* a buscar diferentes estrategias de venta para mejorar la experiencia de compra de sus clientes y así obtener mejores beneficios. Por ello, este trabajo busca segmentar a los clientes a través de la aplicación de técnicas de *Machine Learning* para crear un sistema de recomendación de productos personalizados de acuerdo con las características a la cual pertenece cada cliente y así mejorar la experiencia de compra agilizando y facilitando el proceso desde el aplicativo móvil de la empresa. La propuesta de segmentación se realizó aplicando para el pre-procesamiento de los datos el método estadístico de PCA y se modeló mediante tres técnicas de aprendizaje no supervisado*: K-means, K-medoids y Clustering Jerárquico*. Estas técnicas se evaluaron de forma teórica considerando el método del codo y el dendograma los cuales resultaron en K grupos óptimos. Finalmente, para validarlo de forma práctica, se solicitó la evaluación de un experto de la empresa quien mediante una entrevista comparó los resultados de las técnicas y escogió a *K-medoids* como la segmentación más adecuada para el negocio.

Palabras Clave: Machine Learning, Clustering, PCA, K-means, K-medoids, Clustering Jerárquico.

Abstract

Currently, the constant change in external factors such as technology, the market, and now the global pandemic are forcing companies in the retail sector to seek different sales strategies to improve the shopping experience of their customers and thus obtain better benefits. Therefore, this work seeks to segment customers through the application of Machine Learning techniques to create a personalized product recommendation system according to the characteristics to which each customer belongs and thus improve the shopping experience by streamlining and facilitating the process from the company's mobile application. The segmentation proposal was carried out by applying the PCA statistical method for data pre-processing and modeled using three unsupervised learning techniques: K-means, K-medoids and Hierarchical Clustering. These techniques were evaluated theoretically considering the elbow method and the dendogram which resulted in K optimal clusters. Finally, to validate it in a practical way, the evaluation of an expert of the company was requested, who through an interview compared the results of the techniques and chose K-medoids as the most appropriate segmentation for the business.

Keywords: Machine Learning, Clustering, PCA, K-means, K-medoids, Hierarchical Clustering.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Introducción

Debido a la pandemia del Covid-19, los diferentes sectores empresariales han tenido que adaptarse a los diversos cambios que trajo consigo la pandemia. Por ello, particularmente las empresas B2C han dirigido sus esfuerzos al uso de nuevas tecnologías orientándose a mejorar sus procesos logísticos y operacionales, optimizar los procesos de ventas a través de sus canales digitales, ya que se puede evidenciar una mayor interconectividad a internet y a los dispositivos móviles. Esto se podrá alcanzar haciendo uso de la Inteligencia Artificial (IA) empleando modelos de *Machine Learning,* el cual proporciona beneficios que contribuyen en lograr los objetivos de las empresas y facilitar al cliente mejores herramientas de interacción, procesando y analizando una data amplia para desarrollar algoritmos de acuerdo a sus requerimientos. Asimismo, permite analizar los perfiles de compra de los clientes, reducir el tiempo de entrega e impulsar la venta de marcas propias.

El consumidor actual se encuentra en un escenario mucho más activo e interactúa con marcas de su preferencia por distintos canales. Según el Reporte de la Industria del *E-commerce* en el Perú 2021-22 de *BlackSip* (agencia especializada en comercio electrónico), Perú ha tenido la mayor tasa de crecimiento en *e-commerce* y actualmente es el 6to país en ventas online. También, se encontró que "el 77% de peruanos elige entre dos ofertas guiándose de los descuentos y promociones que publican los negocios." Verano, P. (2019b), y "el 50% de compradores on-line en Perú se ve atraído por bajos precios." Verano, P. (2019a).

El presente trabajo de tesis se desarrollará en la empresa Tottus, empresa del sector *retail,* donde actualmente la competencia en el sector se centra en las ventas online y el uso de la tecnología para satisfacer las necesidades de los consumidores cada vez más exigentes y brindarles una mejor experiencia de compra. Debido a que no se cuenta con una segmentación de clientes para el medio omnicanal Fazil, este estudio se orientará a desarrollar un modelo de *Machine Learning* para la segmentación de clientes de acuerdo a su perfil de compra para potenciar dicho canal, esto proporcionará a la empresa agrupar a sus clientes de acuerdo a los patrones de comportamiento de compra para desarrollar comunicaciones efectivas para campañas y promociones; y así crear relaciones más cercanas entre consumidor y empresa. De esta forma, Tottus podrá mejorar la experiencia de compra de sus clientes.

Después de todo lo mencionado anteriormente, nuestro estudio se desplegará en seis capítulos. El primer capítulo presenta el planteamiento del problema donde se describe la realidad problemática, justificación y delimitación de la investigación. La segunda parte, se centra en el marco teórico donde se ahondará los antecedentes de la investigación y las bases teóricas. Como tercer capítulo, se describe el entorno empresarial, modelo de negocio y mapa de procesos. Luego, en la cuarta parte se desarrolla el diseño de la investigación, metodología de implementación de la solución, los resultados de estos, el cronograma de actividades y presupuesto. En el capítulo cinco, se describe a detalle la propuesta y medición de la investigación. Finalmente, en el capítulo seis se presentan las conclusiones y recomendaciones del trabajo de investigación.

Capítulo I: Planteamiento del Problema

1.1 Descripción de la Realidad Problemática

En el contexto tecnológico y social que nos encontramos, el entorno cada vez tiende a ser más volátil, donde surge la necesidad de adaptarse con mayor rapidez a estos cambios tanto las personas como las empresas.

Según Hootsuite (2021), (citado en *Digital Commerce Partners,* 2021):

> "La transformación digital está en un escenario de constante crecimiento respecto a la conectividad a internet, el uso de las redes sociales y la navegación móvil donde el usuario digital es móvil, el cual interactúa con las marcas por diferentes canales. Con ello se pueden crear nuevas oportunidades y atraer nuevos usuarios (p.4)."

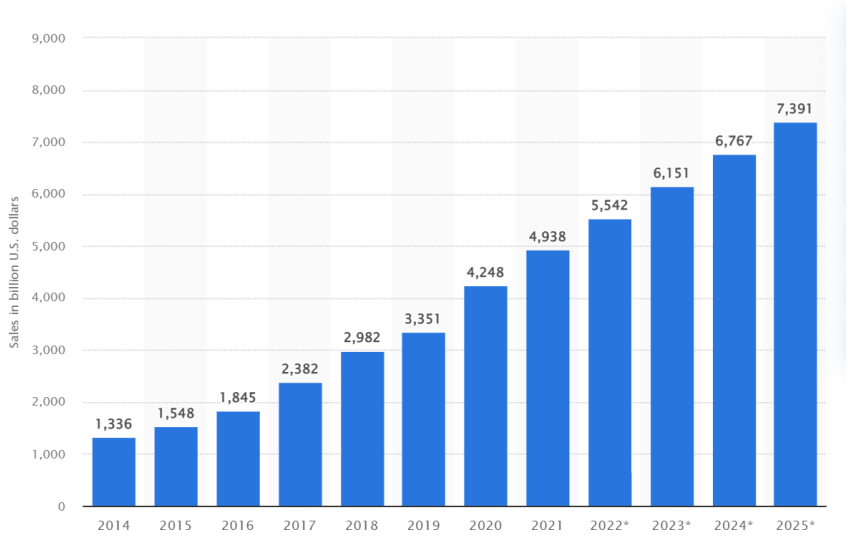
De acuerdo a *Digital Commerce Partners* (2021), "en el mundo, el 60.9% de personas están conectadas a internet y hacen uso del mismo durante 3 horas y 36 minutos al día en promedio desde un teléfono móvil" (p. 20).

En la era de la Industria 4.0, las empresas necesitan hacer uso de las tecnologías como *Machine Learning* para mantenerse en la competitividad del mercado, de lo contrario hasta podrían ir en tendencia al fracaso. Esta tecnología modela el conocimiento a partir del aprendizaje de acuerdo a los patrones que la data proporcionada. Actualmente, dicha tecnología se viene aplicando en el comercio electrónico, industrias manufactureras, entre otras.

Respecto a la crisis mundial de salud, ocasionado por el Covid-19, que hemos afrontado durante estos tres últimos años y que aún continuamos, ha tenido un impacto favorable para el comercio minorista, ya que los consumidores cada vez más exigentes requieren que las empresas brinden una mejor experiencia de compra. Por lo tanto, las empresas han requerido hacer uso de iniciativas digitales con mayor interés y frecuencia. Entre las principales predomina la movilidad, principalmente el desarrollo de aplicativos móviles y las actividades de marketing básicas, donde predomina la omnicanalidad.

Con el crecimiento en el uso de las diferentes herramientas digitales se puede identificar que las ventas minoristas de *e-commerce* ha ido en crecimiento año tras año y se tiene como proyección de crecimiento 7,391 billones de dólares al 2025, de acuerdo a *Statista Research Department* (2022).

*Figura 1:*Ventas minoristas de comercio electrónico en todo el mundo desde 2014 hasta 2025



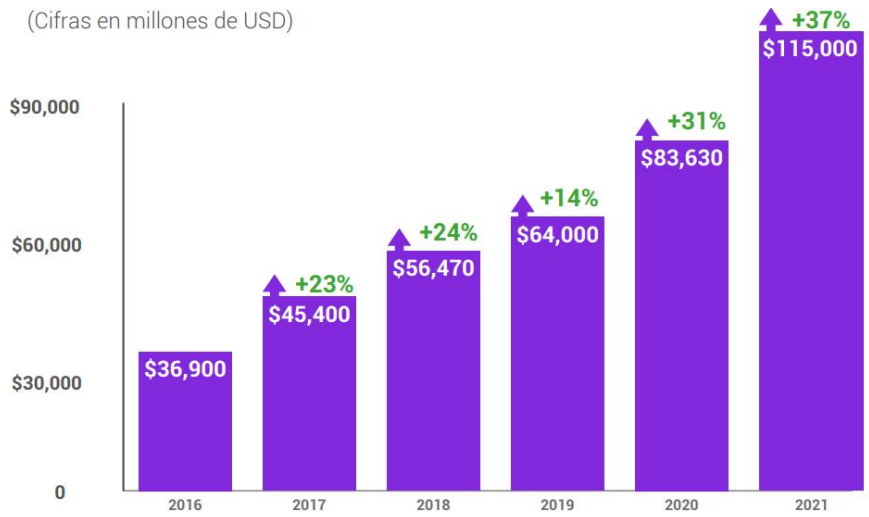
Nota. Se detallan las ventas minoristas del comercio electrónico en todo el mundo en billones de dólares estadounidenses. Obtenida de *Statista*, 2022

También, la crisis sanitaria por la Covid 19 trajo consigo uno de los retos más importantes para la industria y que sigue siéndolo, prevalecer la experiencia del cliente y enriquecer el contenido *e-commerce* poniendo en práctica los principios de usabilidad, ya que según Bazaarvoice (citado en *Digital Commerce Partners*, 2021), "el 54% de los consumidores disfrutan más de la experiencia de buscar en línea que ir a tiendas físicas y tienen más claro los factores de búsqueda como promociones, precios, reseñas, anuncios e *influencers*."

De acuerdo a *Statista Research Department* (2021), "para Latinoamérica el 2020 fue un año con el más rápido crecimiento del comercio electrónico minorista en el mundo con 36.7% respecto al año anterior", ya que tras la pandemia de la Covid-19 el confinamiento de las personas y el cierre de las tiendas físicas como medidas de protección para controlar los niveles de contagio perjudicaron los resultados financieros del año mencionado.

Debido a este hecho, se promovió el uso de las tecnologías digitales por los retailers para obtener resultados favorables en el 2021. Logrando vender $115 mil millones de dólares en dicho año.

*Figura 2:*Ventas minoristas de comercio electrónico en Latinoamérica desde 2016 hasta 2021

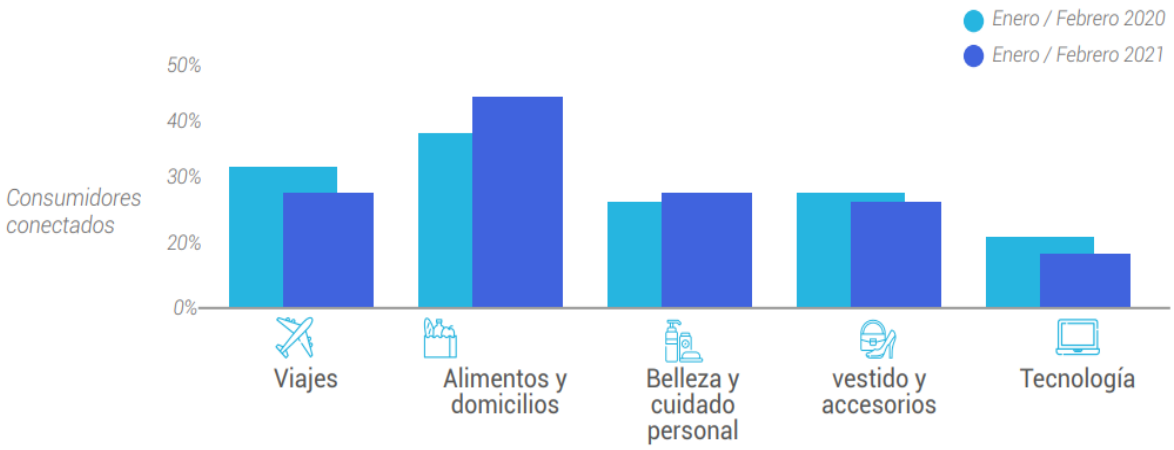


Nota. Ventas minoristas de comercio electrónico en Latinoamérica en millones de dólares estadounidenses. Obtenida de R*eporte industria e-commerce Perú,* 2021 - 2022

La región de América del Sur tiene dos grandes retos que afrontar para mantener el crecimiento dentro del *e-commerce*. El primero es innovar en los medios de pago y la promoción en el uso de las billeteras digitales para alcanzar su masificación como medio aceptado. El segundo reto es mejorar las condiciones que permitan tener mayor beneficio con respecto a los precios de envíos y los costos que se incurrirán en la entrega de los productos, según Euromonitor.

Otra consideración a tener en cuenta es el uso masificado de los celulares y su relación con el crecimiento de ventas a través de este medio, el cual se puede denominar *m-commerce*. Los números para América Latina se estimaron en un crecimiento aproximado de 35 % respecto al año anterior (64 mil millones de dólares estadounidenses) y se pronostica para el 2025 alrededor de 107 mil millones de dólares para la región, según *Digital Commerce Partners* (2021). La siguiente Figura muestra las preferencias de compra a través de dispositivos móviles y los productos con mayor índice de venta son los alimentos y productos de cuidado personal, con mayor crecimiento en los últimos años respecto a las otras categorías.

Figura 3: Compra a través de móviles por categoría en Latinoamérica



Nota. Se detalla la compra a través de celulares según categorías de los meses de enero y febrero del 2020 y 2021. Obtenida de R*eporte industria e-commerce Perú*, 2021 - 2022

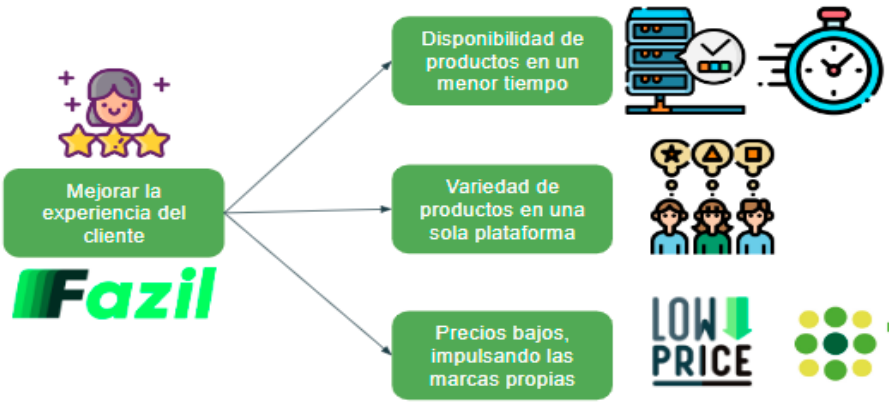
La experiencia y la atención de compra representa actualmente uno de los desafíos más retadores para las tiendas *retail*. Por tal motivo, las compañías están enfocadas en desarrollos de mejora de la exploración digital para guiar al cliente a la búsqueda de sus verdaderas necesidades. Por lo que los sistemas de recomendación de productos son usados específicamente para superar el problema de la falta de información completa en la descripción de los productos y la comunicación directa en las tiendas online que vienen potenciadas con la personalización de las recomendaciones lo que permite tener más información al alcance del cliente disminuyendo el tiempo invertido en el proceso de la compra.

Son casos de éxito para estos sistemas, tiendas consolidadas en el panorama físico como Amazon y Netflix y Best Buy. Por un lado, Best Buy cuando estaba al borde del fracaso en 2015, implementó un sistema de recomendación de productos para migrar y enfocarse en el crecimiento de la venta en línea por lo que la empresa reportó el aumento de sus ventas en 23.7%. Por otro lado, Amazon, reconocido minorista en línea decidió utilizar sus sistemas de recomendación para generar sugerencias en sus usuarios y así lograr mejores resultados en las ventas del canal *e-commerce*. De esta manera permitió a Amazon atribuir el 35% de sus ventas al sistema de recomendación que utiliza.

Tal como ya se pudo observar estas aplicaciones de delivery promocionan ofertas mediante notificaciones en los smartphones de los clientes intentando de esta manera generar mayor cantidad de visitas y conversión en compra de productos. Estas recomendaciones no necesariamente están asociadas a las preferencias de cada tipo de clientes, ni a su información específica.

Dada la necesidad de compra con mayor rapidez nace Fazil, un aplicativo móvil de delivery que se alinea a los tres objetivos asociados a *Tottus.* Los objetivos intrínsecamente relacionados son la experiencia del cliente: mediante la disponibilidad de productos en un menor tiempo; que sus clientes puedan encontrar la variedad de productos en una sola plataforma y que los clientes encuentren precios bajos impulsando la adquisición de marcas propias de *Tottus*.

Figura 4: Problemática de la empresa



Nota. Elaboración propia

Dando una revisión a los aplicativos presentes en el mercado y haciendo un comparativo con Fazil, todos cuentan con un sistema de campañas que promocionan productos según las estrategias planteadas o sugeridas por su gestión de Marketing. Sin embargo, las recomendaciones actuales no están segmentadas por tipo de cliente sino son promociones generales. Debido a lo mencionado, Tottus puede mejorar este sistema de campañas al implementar un sistema de recomendación de productos personalizados para agilizar y facilitar el proceso de compra que se realiza mediante el aplicativo móvil de la empresa. Esta personalización alineada a los tres objetivos específicos podría reducir los costos asociados a la generación de campañas de productos, así como el incremento de las ventas de Tottus.

1.2 Justificación de la Investigación

1.2.1. Justificación Teórica

Para esta investigación la aplicación de *Machine Learning,* específicamente aplicando PCA y técnicas de aprendizaje no supervisado, busca generar una segmentación de clientes del canal *e-commerce* según sus características, basándonos en los datos de las transacciones de venta, para así crear nuevas estrategias de ventas personalizadas según su perfil de compra

1.2.2. Justificación Práctica

Esta investigación tiene la finalidad de aportar una propuesta de segmentación de clientes, mediante el uso de *Machine Learning*, para que con esta base de información se pueda generar un sistema de recomendaciones personalizadas de productos según las características de los clientes. Esto con la finalidad de mejorar la experiencia de compra, agilizando y facilitando el proceso de compra mediante el aplicativo y a través de ello mejorar el cumplimiento de los objetivos estratégicos de la empresa.

1.2.3. Justificación Metodológica.

En la investigación se aplicará *Machine Learning,* mediante técnicas en dos etapas, para el pre-procesamiento de los datos se usará PCA y para el modelado se aplicará las técnicas de aprendizaje no supervisado: *K-means, K-medoids y Clustering Jerárquico*. Los datos que se usarán serán de las transacciones mensuales en conjunto con los datos de los clientes para así crear una segmentación ideal según las características o perfil de compra que se ajuste a la necesidad de la empresa.

1.3 Delimitación de la Investigación

1.3.1 Delimitación espacial

Para la investigación se ha realizado la evaluación de las transacciones realizadas por el aplicativo Fazil. Estas transacciones fueron realizadas de manera digital en Lima y provincias, pero para este estudio se tomará en cuenta solo las ventas generadas en Lima. Asimismo, también se cuenta con el dato de los distritos en donde ocurrió la venta, pero esto no será determinante en la descripción de los clientes.

1.3.2 Delimitación temporal

La presente investigación tomará la información proporcionados por la empresa Fazil; el periodo específico que se toma para el modelo se considera solo del mes de abril del 2022, esta data nos dará información sobre las transacciones por el aplicativo Fazil, lo que permitirá pronosticar las preferencias de los clientes y así poder clasificarlos considerando los datos de su compra.

1.3.3 Delimitación conceptual

Esta investigación tiene como finalidad contribuir a la creación de un sistema de recomendación de productos personalizados para grupos determinados de clientes. Para esta mejora se requiere la implementación de *Machine Learning*. Con el uso de *Machine Learning* se podrá lograr la personalización de recomendación de productos de acuerdo con las preferencias de sus clientes, datos geográficos, entre otros datos del cliente. El tipo de aprendizaje que se realizará es un aprendizaje no supervisado, ya que las variables de estudio, al no tener una variable objetivo, no hay una correlación de causa y efecto entre variables dependientes e independientes por lo que a partir de los datos seleccionados se realizará el análisis de las variables encontradas y permitirá analizar el comportamiento del consumidor y así agruparlos en grupos diferenciados. Para ello, se evaluarán 3 técnicas de aprendizaje no supervisado: *K-means, K-medoids , Clustering Jerárquico* y para el preprocesamiento de los datos se usará la técnica de PCA, de manera que se escogerá la técnica que genere *clusters* o agrupamientos con una mejor descripción de grupos de clientes. Asimismo, estos modelos serán apoyados mediante la opinión de un experto, para poder validarlo y escoger el número de *clusters* apropiado que describan mejor a los clientes de *Tottus*.

Capítulo II: Marco Teórico

2.1 Antecedentes de la Investigación

2.1.1. Tesis relacionadas

Cisterna Mollocco C. (2021). "Segmentación de clientes activos de una entidad financiera empleando el algoritmo de K-means y árbol de decisió*n*"

- **Problema**

La entidad financiera confidencial de la investigación, una de las cuatro entidades principales del Perú, generaba campañas comerciales para sus clientes que no tenían los resultados esperados ya que el perfil de los clientes es totalmente diferente y no estaba diferenciado o segmentado. Identificado este problema, se optó a realizar una segmentación sobre sus clientes activos, considerando cliente activo a aquel que hizo operaciones monetarias o no monetarias en sus canales físicos como digitales dentro de los últimos seis meses, con el objetivo de conocer de forma precisa su perfil para finalmente crear indicadores claves en el desarrollo de las campañas y/o acciones comerciales de la empresa que serán enfocadas por tipo de cliente.

- **Metodología**

La investigación busca segmentar clientes activos de la empresa según los datos registrados en sus operaciones monetarias o no monetarias dentro de sus canales físicos como digitales dentro de los últimos seis meses: de diciembre 2020 a mayo 2021. Los datos se obtuvieron de las diferentes bases de datos correspondientes a los clientes activos considerando: consumos realizados con tarjetas de crédito y débito en diferentes rubros, operaciones ejecutadas por los canales físicos y digitales, datos del cliente del sistema financiero, entre otros. De esta información se construyeron inicialmente alrededor de 250 variables.

La metodología empleada es la Crisp - DM.:

1. Como primer paso, se identificó la necesidad de tener una segmentación de clientes

2. Luego se recopiló de sus bases de datos la información de los clientes activos para identificar variables y generar promedios, desviaciones entre otros.
3. En el siguiente paso se procede a la limpieza de variables e identificar a las variables que se encuentren no correlacionadas para poder ser empleadas en el algoritmo de *K-Means.*
4. Posterior a ello se valida si la segmentación es estable en periodos previos evaluando la Silueta e Inercia.
5. Luego de obtener el número de clúster se realizará un perfilamiento en cada clúster para obtener una segmentación detallada, está segmentación se muestra a los responsables de negocio para aprobación.
6. Posterior a la reunión con los responsables, se elaboró un árbol de decisión, el cual se entrenó con la información del último mes. Esto se realiza ya que el algoritmo de *K–means* al modelar con nuevos clientes asigna diferentes marcas de clúster, pero mantiene los perfiles de los clientes. Por ello se emplea el árbol de decisión multi clase para evitar este inconveniente.

- **Solución:**

Luego de filtrar los datos se realiza un análisis de correlación entre las variables que están disponibles, sólo se considerarán variables que tengan una correlación máxima de 0.30. Estas variables se utilizaron para encontrar las combinaciones de variables más adecuadas que luego crearán los clústeres mediante el algoritmo de *K – means* con los datos del mes de mayo del 2021. Este modelo será evaluado mediante los indicadores de inercia y silueta en los meses de enero a mayo del 2021 para conocer si esta segmentación es estable en periodos anteriores, esto a fin de usar la segmentación en próximas campañas con nuevos públicos objetivos.

- **Resultados:**

Del análisis y de las combinaciones de las variables se encontró tres variables con correlación menor a 0.3 que contribuyen al modelado de la segmentación:

- V1: Número de operaciones totales realizadas en el aplicativo de la entidad financiera

- V2: Monto promedio en consumos de plataformas digitales en los últimos seis meses.
- V3: Monto promedio de ingreso a las cuentas de los clientes en los últimos seis meses.

Estas variables se usaron para el modelo con el algoritmo K – *means* con los datos del mes de mayo del 2021 obteniendo 5 números de clusters con una inercia de 12, 77 y silueta de 0.56. Luego de ello se procede con la validación del modelo para diferentes periodos como se muestra en la Tabla 1, en la que resulta en que los valores de inercia y silueta son muy parecidos por lo que se considera una propuesta estable.

*Tabla 1:*Validación de la segmentación en diferentes periodos

Cosecha	# Clu	Inercia	Silueta	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
Ene-21	5	12,534,640	0.59	34%	21%	18%	15%	11%
Feb-21	5	12,852,590	0.57	34%	21%	19%	15%	11%
Mar-21	5	12,777,675	0.52	34%	22%	18%	15%	11%
Abr-21	5	12,946,375	0.56	34%	22%	17%	15%	11%
May-21	5	12,777,820	0.56	34%	22%	18%	15%	11%

Nota. Tabla resumen extraida de Cisterna Mollocco, C. (2021)

Finalmente se procede a perfilar los clusters para presentarlos a los especialistas y luego de ello generar el árbol de decisión multi clase el cual tiene un accuracy de 0.99 lo que significa que el algoritmo del árbol de decisión tiene un buen desempeño, el cual se podrá emplear para nuevos públicos objetivos de manera mensual.

Palacios, F y Pastor, N. (2020). "Segmentación de clientes de una empresa comercializadora de productos de consumo masivo en la ciudad de Popayán soportado en Machine Learning y Análisis RFM (Recency, Frecuency y Money)"

- **Problema:**

  Se busca establecer los beneficios que se puede obtener para la empresa comercializadora de productos de consumo masivo en Popayán a partir de la segmentación de los clientes utilizando *Machine Learning*.

- **Metodología**

De acuerdo con lo mencionado en el objetivo, la investigación desarrolló 2 modelos de segmentación de clientes que a continuación se detallan:

- Modelo RFM:

  Se basa en tres variables ligadas con la interacción comercial de los clientes con la empresa, con las siglas RFM en inglés y que describen el modelo, se detallan a continuación:

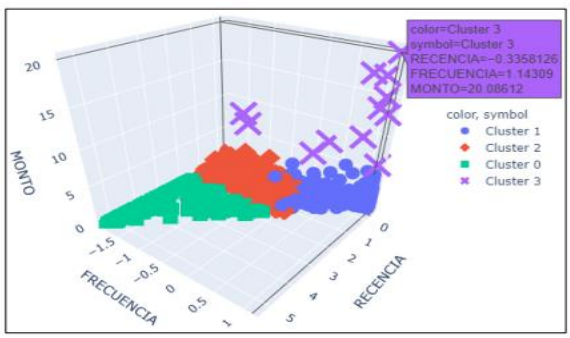
1. Recencia: Definido como el tiempo transcurrido entre la fecha actual y la fecha de la transacción más reciente.
2. Frecuencia: Son las transacciones que un socio ha realizado en un periodo de tiempo específico.
3. Monetario: Son las transacciones traducidas en términos monetarios.

Con este modelo se espera responder lo siguiente: ¿Qué valor tienen nuestros clientes?

- Modelo Clustering: K-Means.

  El modelo RFM nos brinda los datos preprocesados, por tal motivo, al implementar el algoritmo de clustering no será necesario, ya que se definieron con las variables de frecuencia, frecuencia y monto, cliente y fecha. La distancia Euclidiana sirvió para realizar el entrenamiento de 4 *clusters*. La Figura 5 presenta los resultados del modelo por medio de una gráfica de dispersión, donde cada cluster está representado por una forma y color.

Figura 5: Asignación de Clustering en Python



Nota. Cada cluster está clasificado por un color distinto. Obtenido de Palacios, F y Pastor, N, 2020

Tras obtener los resultados de *clusters* se almacenan los archivos csv que contienen los valores de Recencia, Frecuencia, Monto y Cluster respecto a cada cliente, a fin de analizar los outputs del proceso de clustering e identificar el grupo poblacional que representa cada cluster.

- **Resultados:**

  - Modelo RFM:

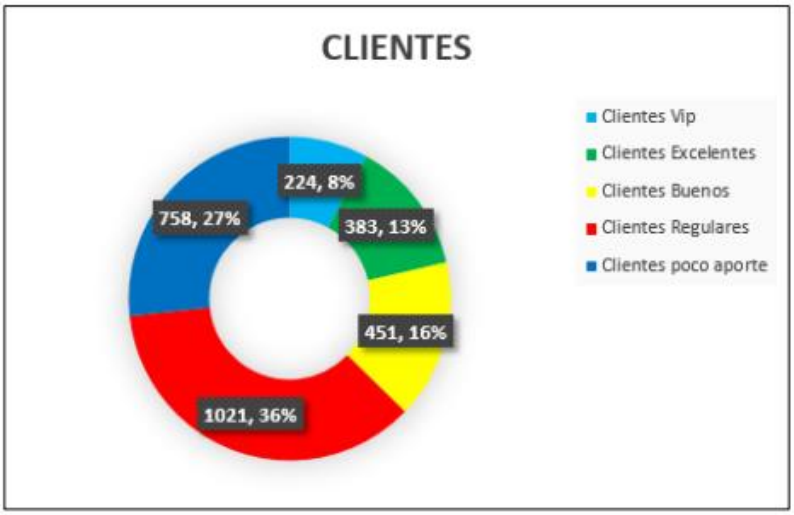
    A continuación, se muestran los resultados obtenidos a partir de la ejecución de la metodología de RFM.

*Tabla 2:*Resultado del análisis RFM

SEGM X VALOR	CLIENTES	% CTES	VENTAS $	% VENTAS	VTA X CLIENTE
Clientes Vip	224	8%	$ 3,416,250,880	42%	$ 15,251,120
Clientes Excelentes	383	14%	$ 1,952,670,083	24%	$ 5,098,355
Clientes Buenos	451	16%	$ 1,376,944,558	17%	$ 3,053,092
Clientes Regulares	1021	36%	$ 1,200,952,163	15%	$ 1,176,251
Clientes poco aporte	758	27%	$ 264,748,879	3%	$ 349,273

Nota. Tabla resumen de los cluster obtenidos por el modelo RFM. Obtenido de Palacios, F y Pastor, N, 2020

Figura 6: Porcentaje de distribución de clientes según RFM



Nota. Gráfico de pastel que diferencia a los *clusters* y % de participación de los clientes. Obtenido de Palacios, F y Pastor, N, 2020

- Modelo Clustering K-Means:

A continuación, se muestra la caracterización de clientes y los resultados de implementar el algoritmo de *K-Means* para segmentar los clientes.

*Tabla 3:*Resultados segmentación con K- Means

Clúster	Clientes	Porcentaje	Monto T	Monto T%	Prom(R)	Prom(F)	Prom(M)	Clasificacion
0	187	7%	46.900.369,00 COP	1%	276	5	250.804,00 COP	Clientes Poco Aporte
1	1717	61%	6.354.262.864,00 COP	77%	42	42	3.700.793,00 COP	Clientes Buenos
2	12	0%	1.054.201.070,00 COP	13%	50	39	87.850.089,00 COP	Clientes VIP
3	921	32%	756.220.060,00 COP	9%	59	15	821.066,00 COP	Clientes Regulares

Nota. Cuadro resumen de clasificación. Obtenido de Palacios, F y Pastor, N, 2020.

Calad Noreña. F. (2015). "Segmentación de clientes automatizada a partir de minería de datos k-means clustering"

- **Problema**

Tiendacol SA actualmente cuenta con un ERP donde está almacenada toda la información transaccional de la empresa. A pesar de poseer toda la información, se identifica que no se está explotando todo su potencial pues no se hace ningún análisis de fondo sobre la misma.

Dada la situación planteada de Tiendacol S.A, con el fin de poder enfocar mejor los esfuerzos comerciales y de servicio al cliente de la empresa, se busca cómo crear una segmentación de clientes automatizada, mediante el empleo de la técnica de minería de datos *K.means clustering* y las herramientas gratuitas R o *RapidMiner*, que permite realizar una clasificación de los mismos de acuerdo a sus características demográficas y patrones de compra.

- **Metodología**

Por ello, para mejorar el servicio al cliente se creará una segmentación automatizada usando *K-means* y *RapidMiner*. Para identificar los números de segmentación se usaron los datos de venta desde el 2011 al 2014. De los datos, se identificaron nueve variables: como tipo de compra ya sean al contado y al crédito, total de compras y el promedio de compras por año. La metodología que se usó fue *cross-industry*.

Como primer paso se identificaron los datos de las ventas y de los clientes, luego se identificaron las variables y se estandarizaron. En el tercer paso se diseñó el modelo para luego aplicar el algoritmo *K-means* y finalmente aplicar un árbol de decisión.

- **Solución:**

Una vez el modelo ejecutado y analizado se identifican seis categorías en 3 tipos de escenarios. Los *clusters* se caracterizaron de la siguiente manera:

Cluster 1: Son los clientes menos significativos. Este grupo incluye el mayor porcentaje de clientes (63.761 -> 45.69%).

Cluster 2: Representa a los clientes que más compran e invierten más dinero en las tiendas, agrupa menos población. (872 -> 0.62%).

Clusters 4 y 6: Representa a los clientes que compran con poca frecuencia y por montos bajos (61,433 -> 44.02%).

Además, se modeló los *clusters* por años según la tabla 4.

- **Resultados:**

Tras el desarrollo del modelo se permite obtener una segmentación que explica los 6 *clusters* en 3 tipos de escenarios donde se obtiene un buen porcentaje del total de clientes.

*Tabla 4:*Asignación de Clustering por año

	Mejores	Regulares	Peores
***Clusters* 2011**	1 y 3	5	2, 4 y 6
***Clusters* 2012**	1 y 3	5 y 6	2 y 4
***Clusters* 2013**	5 y 6	1	2, 3 y 4
***Clusters* 2014**	4 y 5	3	1, 2 y 6

Nota. Cuadro resumen del modelado de clusters por año. Obtenido de Calad Noreña, F. (2015)

2.1.2. Artículos relacionados

Cam, C. (2021). "Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa".

- **Problema:**

El autor del presente artículo busca identificar hallazgos relevantes en los datos, para que a través de estos permitan proponer acciones comerciales. Por lo tanto, los retos que plantea este problema son estos:

- Entender los intereses del consumidor.
- Proponer una segmentación de consumidores.
- Crear una infraestructura de *Big Data.*
- Proponer un sistema de recomendaciones que permita personalizar ofertas para incrementar la facturación y fidelizar la cartera de clientes

- **Metodología:**

La investigación tiene como objetivo segmentar a los clientes de una cadena de supermercados. Para ello se utilizó la metodología TDSP (*Team Data Science Process*) el cual tienes las siguientes etapas:

- Comprensión del negocio: En esta etapa se identificó que se cuenta con dos canales de venta presencial y virtual.
- Captura de datos: El cual tiene tres pasos adicionales: Ingesta y preprocesamiento de datos, en donde se obtuvieron de fuentes internas de la empresa y de externas como tuits de los clientes, INEI y de Nutriscore. Y la configuración de una arquitectura de *Big Data* para lo cual se usó en Amazon Web Services el EMR, EC2 y S3.
- Modelado: En esta etapa previamente se analizaron las variables por lo que se procedió a aplicar una primera segmentación encontrando seis *clusters.*

- **Resultados:**

Para la segmentación se excluyeron los registros con valores nulos en la columna edad. Se llevaron a cabo tres iteraciones. En la primera, con todas las variables, se obtuvo superposición de los puntos de varios *clusters*, pero no se logró una agrupación limpia; por lo tanto, se desechó esta opción.

La segunda iteración, se decidió eliminar la edad, pues excluye registros y no contribuye a la clusterización. En este caso, se observa que la superposición ha mejorado; sin embargo, se obtiene una agrupación que no es muy limpia y se puede ver que las variables sintéticas de los grupos de mercancías tampoco aportan al agrupamiento; por lo tanto, también se desecha esta opción. En la tercera iteración, se elimina la variable edad y las variables sintéticas de los grupos de mercancías.

En la tercera iteración sí se obtuvieron seis *clusters* claramente definidos con baja superposición (solo en dos grupos) y con los centroides bien ubicados. Por lo tanto, se considera a la tercera opción como el mejor resultado de la segmentación.

Tabla 5: Resultado de aplicación K-means

Clúster	Color	Cantidad	Antigüedad (días)	Promedio de visitas por mes	Gasto promedio por mes	Descuento promedio por mes	Ratio gasto/ renta
0	Rojo	367	627,77	77,34	135,68	3,19	4 %
1	Verde	336	151,16	144,6	281,82	10,37	1 %
2	Azul	1796	131,92	51,3	99,48	2,51	3 %
3	Amarillo	36	1 254,16	71,48	120,33	4,25	4 %
4	Morado	639	297,7	71,69	124,74	2,51	5 %

Nota. Explicación del resultado de la segmentación Citado por Cam, C. (2021).

Tripathi, S., Bhardwaj, A. y Poovammal. (2018). "Approaches to clustering in Customer Segmentation". International journal of engineering & technology, 7(3.12), 802.

- **Problema**

  Los modelos de agrupamiento disponibles para el análisis empresarial en el contexto de la segmentación de clientes generan algunos conflictos para que se tenga mucha cantidad de data disponible para bajo un enfoque analítico podamos evaluar correctamente la información del cliente y el análisis del valor de los clientes para una mejor comprensión del cliente.

  Actualmente con el diseño de soluciones de tecnología de la información (TI) se tiene algunos modelos los cuales tienen ventajas y desventajas tal como el modelo de *K-Means* y agrupamiento jerárquico por lo que se ve la necesidad de generar un modelo híbrido que pueda superar a los modelos individuales.

- **Metodología**

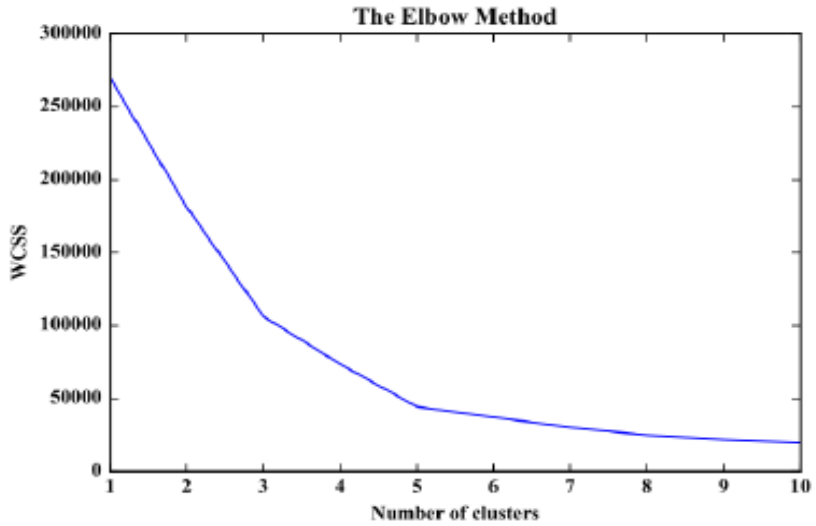
  El estudio de investigación emplea *K-means* y clustering jerárquico para la segmentación de clientes, con el objetivo de obtener una solución híbrida superando las técnicas de forma individual.

- **Solución**

  El trabajo plantea la elaboración de agrupaciones para la data histórica que se maneja

  - Agrupamiento de K-Means: Con el uso del método de codo llega a la conclusión de que con un K=5 se obtiene una buena segmentación de sus clientes.

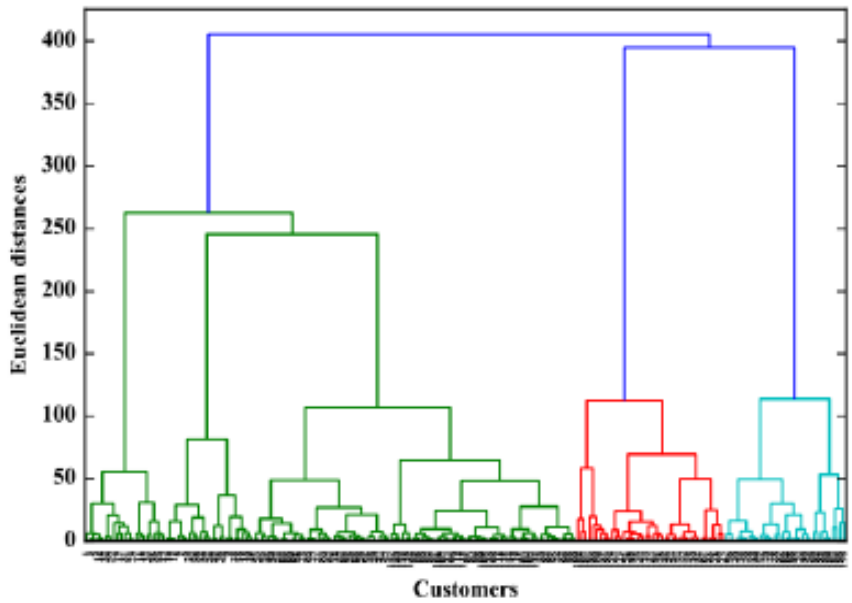
Figura 7: Validación de los clusters K-means



Nota. Obtenido de Tripathi, S., Bhardwaj, A. y Poovammal. (2018).

- Agrupación jerárquica: este se mide en comparación de modelo aglomerativo y decisivo. Para obtener resultados satisfactorios, podemos elegir cinco grupos (K=5) tal como se puede observar en la Figura 8

Figura 8: Validación de los clusters jerárquico



Nota. Obtenido de Tripathi, S., Bhardwaj, A. y Poovammal. (2018).

La agrupación jerárquica, ya que no requiere ningún centro de agrupación como entrada, nos brinda mejores opciones de elegir los grupos de conglomerados, así como su número. El agrupamiento jerárquico también brinda mejores resultados en comparación con *K-Means* cuando se usa un conjunto de datos aleatorio.

La salida o los resultados obtenidos cuando se usa el agrupamiento jerárquico tienen la forma de dendogramas, pero la salida de *K-Means* consta de agrupaciones de estructura plana que pueden ser difíciles de analizar. A medida que aumenta el valor de k, la calidad (precisión) del agrupamiento jerárquico mejora en comparación con el agrupamiento de *K-Means*.

- **Resultados**

Los modelos de agrupamiento deben poseer la capacidad de procesar esta enorme cantidad de datos de manera efectiva.

Cada uno de los algoritmos de agrupamiento discutidos anteriormente viene con su propio conjunto de ventajas y desventajas. La velocidad computacional del algoritmo de agrupamiento *K-Means* es relativamente mejor en comparación con los algoritmos de agrupamiento jerárquico, ya que estos últimos requieren el cálculo de la matriz de proximidad completa después de cada iteración. El agrupamiento de *K-Means* brinda un mejor rendimiento para una gran cantidad de observaciones, mientras que el agrupamiento jerárquico tiene la capacidad de manejar menos puntos de datos.

Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022). Research on segmenting E-commerce customer through an improved K-medoids clustering algorithm. Computational Intelligence and Neuroscience, 2022, 9930613.

- **Problema**

El artículo plantea que, en la segmentación de clientes, el RFM es el modelo más clásico, propuesto por Hughes y sobre este modelo muchos académicos desarrollaron técnicas de análisis de agrupamiento para segmentar clientes. Sin embargo, todavía existen algunos vacíos en la literatura existente.

Por un lado, cuando se seleccionan las características se centra en el uso de los datos históricos de pedidos de los clientes, que no pueden reflejar completamente las preferencias de comportamiento y los hábitos de consumo de los diferentes grupos de clientes. Por otro lado, en términos de selección del algoritmo de conglomerados, el algoritmo de conglomerado de K-medias propuesto por la literatura existente no consideró la eficiencia de la operación del algoritmo para mejorar el rendimiento de la segmentación de clientes de comercio. Debido a ello se plantea encontrar un modelo que permita establecer artificialmente valores en el algoritmo *K-Medoids* e introducir el CH como índice de evaluación de calidad de agrupamiento para determinar el mejor k-valores.

- **Metodología**

El estudio de investigación emplea la metodología de mejorar el K con el uso del algoritmo *K-Medoids* desde dos aspectos:

- Usa el índice de evaluación CH para determinar el número óptimo de grupos en el algoritmo *K-Medoids*.

- Plantea el uso del algoritmo *K-Means* al seleccionar los centros de agrupación iniciales.

Adicionalmente su metodología sigue los siguientes pasos:

- Limpieza de datos: Se procesan los datos con valores perdidos y anómalos, como los datos con gasto cero, los datos con la fecha de compra como valor inactivo y los datos con un gasto obviamente erróneo

- Extracción y Normalización de Indicadores: Los indicadores individuales en el modelo RFMCV se explican en

R: Representa el intervalo de tiempo entre la última compra del cliente en el período de observación y el 31 de diciembre de 2017.

F: Frecuencia de compra del cliente en el período de observación.

Metro: Representa la cantidad monetaria gastada por el cliente en el período de observación.

C: Frecuencia del cliente que agregó el producto al carrito en el período de observación.

V: Frecuencia del cliente que agregó el producto a favoritos en el período de observación.

-Análisis de Resultados Empíricos.

Basado en el modelo RFMCV, se ejecuta el algoritmo *K-medoids* mejorado. Los resultados muestran que todos los clientes se dividen en 4 grupos, denominados Tipo A, Tipo B, Tipo C y Tipo D.

- **Solución**

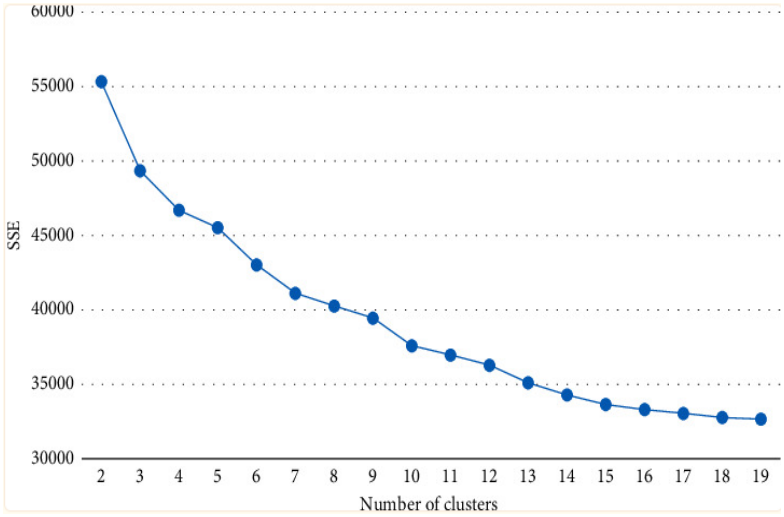
Esta investigación permite analizar a los clientes de comercio electrónico cuyos comportamientos de consumo se basan en la plataforma de Internet. Es necesario agregar más nuevas características y patrones de consumo en línea.

Por lo tanto, integramos dos características del comportamiento de consumo en línea en el modelo RFM, que incluyen agregar al carrito (C) y agregar favoritos (V).

En segundo lugar, para resolver los problemas de establecer artificialmente de k-valores y sensibilidad a los centros de agrupamiento iniciales, mejoramos el algoritmo de agrupamiento de *K-medoids* existente mediante la introducción del índice de evaluación de la calidad del agrupamiento CH y la idea del algoritmo *K-Means*.

Con el análisis un análisis empírico realizado para una muestra de 37.376 clientes de una plataforma de comercio electrónico se llega a un K=4 y a su vez se diferencian métodos obteniendo mejor resultado con el *K-Medoids*.

Figura 9: Validación del valor k obtenido

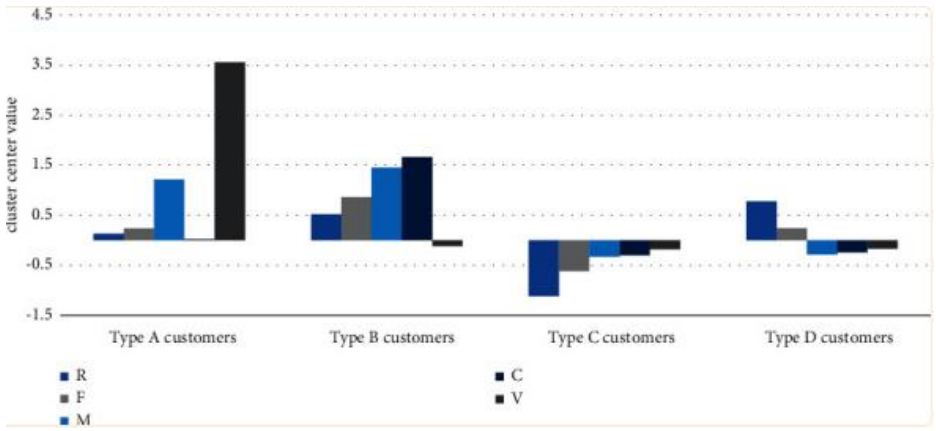


Nota. Obtenido de Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022).

- **Resultados**

Con el artículo se llega a complementar el modelo RFM tradicional integrando el comportamiento de consumo de los clientes. En segundo lugar, se introduce el índice CH para determinar el mejor valor de K. En tercer lugar, al combinarse con el algoritmo *K-Means*, el algoritmo *K-Medoids* se mejora seleccionando de manera óptima el centro de agrupación inicial. Finalmente, se llega a una segmentación de clientes en 4 *clusters* con el método propuesto que definen mejor a sus clientes tal como se muestra en la Figura 10.

Figura 10: Resultado de la segunda segmentación



Nota. Obtenido de Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022).

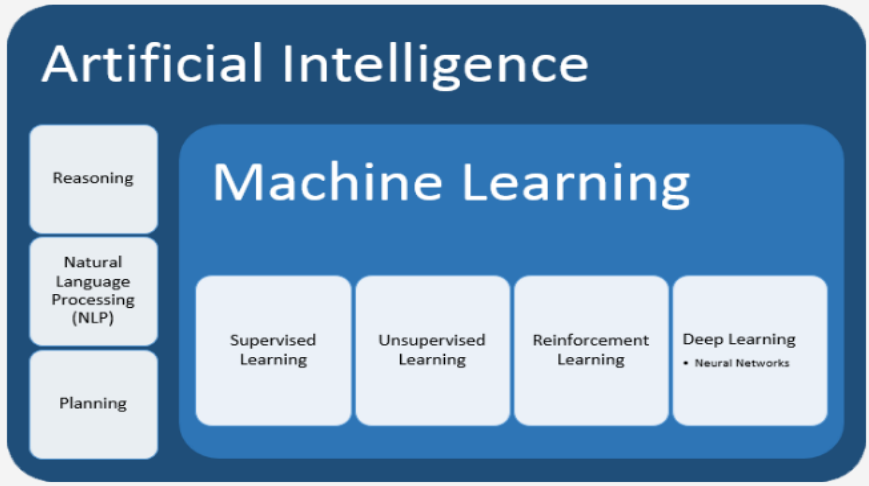
2.2 Marco Teórico

2.2.1 Inteligencia Artificial:

La definición de la inteligencia artificial, según diversos expertos no se debe considerar estática ya que con el avance de las tecnologías el concepto trae consigo nuevos aspectos e implicancias a lo largo de los años y del desarrollo de la misma.

Dentro de las definiciones de IA "[para] Cáceres (2021), (..) no es hasta 1956 donde John Mc-Carthy, Marvin Minsky y Claude Shannon, tres científicos destacados de la época, acuñaron el término "Inteligencia Artificial" durante la Conferencia de Dartmouth como "la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cálculo inteligentes". (Morales, 2021, p. 44)

Figura 11: Inteligencia Artificial



Nota. Representación gráfica de Machine Learning. Obtenido de página web IBM (2020).

Se debe considerar que la IA engloba dos conceptos macro: *Machine Learning* y *Deep Learning*. Así, se tienen los siguientes conceptos:

> "El ´*Machine Learning*´ o ´aprendizaje automático´, es la capacidad que tienen las máquinas de recibir un conjunto de datos y aprender por sí mismas, cambiando y ajustando los algoritmos a medida que procesan información y conocen el entorno". (Morales, 2021, p. 47)

"El "*Deep Learning*" o "aprendizaje profundo", por su parte, es una rama del *Machine Learning* que se ocupa de emular el enfoque de aprendizaje que los seres humanos utilizan para obtener ciertos tipos de conocimiento. (...). Los modelos computacionales de *Deep Learning* imitan las características arquitecturales del sistema nervioso". (Morales, C., 2021, p. 48).

2.2.2 Machine Learning:

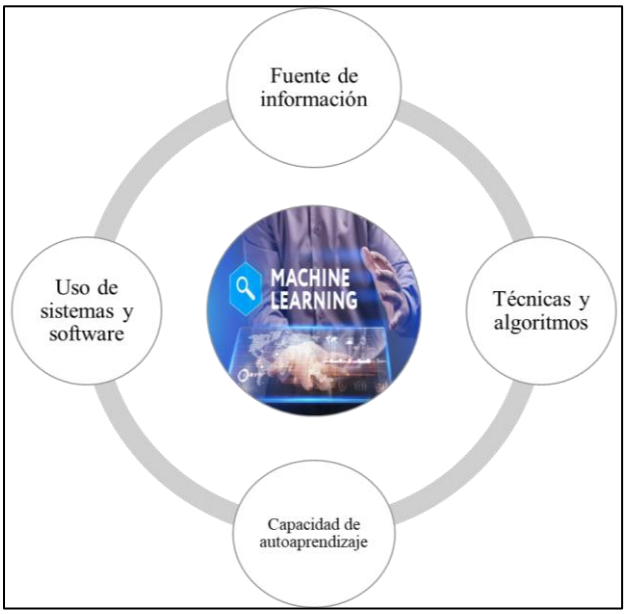
"Es una extensión de la Inteligencia Artificial encargada de desarrollar algoritmos que tienen capacidades de aprender y no tener que programarlos de manera explícita" (Sandoval, 2018).

Por otro lado, la página web *Management Solutions* (2018) sostiene que el *Machine Learning* son un conjunto de algoritmos que pueden hallar patrones en los datos automáticamente.

2.2.2.1 Componentes del Machine Learning.

A continuación, se desarrollarán los componentes del *Machine Learning*:

Figura 12: Componentes del Machine Learning



Nota. Resumen gráfico de los componentes de Machine Learning.

- **Las fuentes de información**, comprende datos, los cuales pueden ser:
  - Estructurados: Base de datos estructurada y organizada a partir de un reporte.
  - No estructurados: Son aquellos que se encuentran dispersos en diversas fuentes de datos como mails, entre otros.
- **Las técnicas y algoritmos**, relacionados con las tareas a realizarlas:
  - Técnicas de tratamiento de datos no estructuradas: parsing, mapas auto-organizativos, etc.
  - Modelos de *Machine Learning*, supervisados y no supervisados: modelos de clasificación, regresión, estocásticos etc.
- **La capacidad de autoaprendizaje,** que mejora las medidas de desempeño para permitir el reentrenamiento automático a partir de nueva información para combinar modelos y ponderación/calibración.
- **El uso de sistemas y software** para la visualización de la información y la programación:
  - Visualización: Power BI, QlikView, QlikView, SAS Visual Analytics, TIBCO Spotfire, Tableau.
  - Programación: Java, Scala, Ruby, SAS, Matlab, C, Python, Azure, R, SQL,

2.2.2.2 Etapas de proceso de Machine Learnnig:

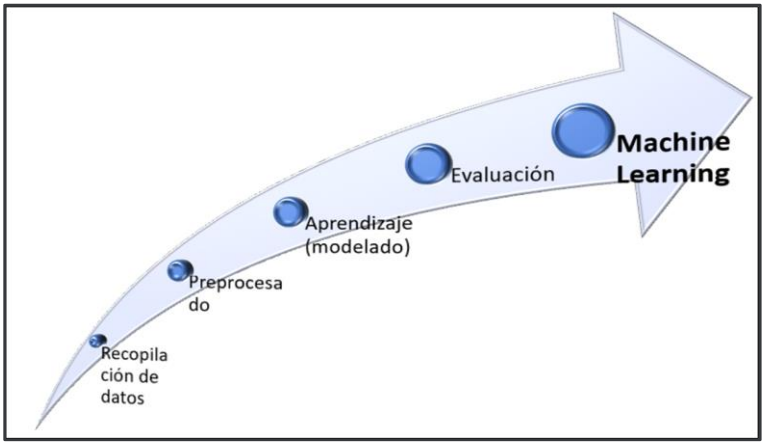
A. Recopilación de datos: Obtener la data puede ser una tarea difícil, pero existen formas automatizadas de realizarlas, buscar organizarlas es algo que se debe ejecutar en esta primera etapa.

B. Preprocesado: Aquí se debe preparar la data obtenida para esto es necesario detectar, depurar o corregir datos que presentan las siguientes características:
   - Datos ausentes: Presentan algún dígito errado porque se digitó manualmente o porque viene sin datos asociados la cual tiene un alto riesgo de error.
   - Datos espurios: Los espurios son valores atípicos.

C. Aprendizaje: Aplicando el algoritmo que más se adecue al modelo de la diversidad de algoritmos a los datos preparados en las fases anteriores, en esta etapa se halla la posible solución o modelo predictivo del objeto de estudio.

D. Evaluación: Es necesario evaluar la posible solución, en esta etapa se efectuará ello a través una técnica muy utilizada: Evaluación a través de conjunto de test, la cual consiste en tener dos conjuntos de datos, uno para el aprendizaje del modelo y otro para el testeo, teniendo en cuenta la variable de salida "Y" la cual se conoce.

Figura 13: Etapas de proceso de Machine Learning

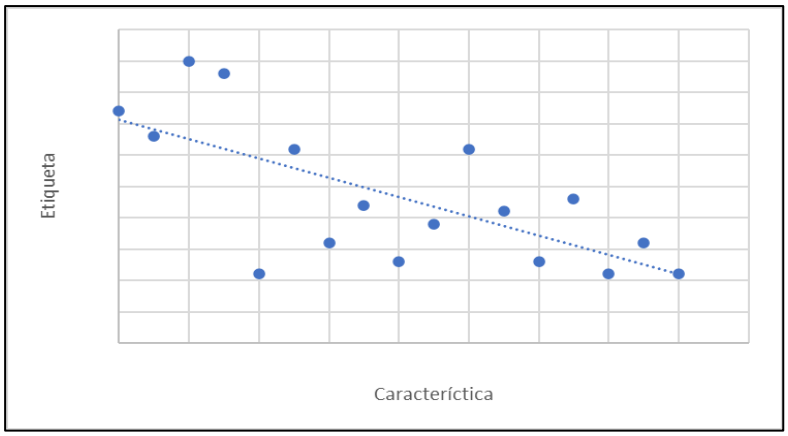


Nota. Resumen de los componentes de Machine Learning. Obtenido de González, A. y Alba, N. (2017).

2.2.2.3 Tipos de aprendizaje:

a. **Supervisado**: La finalidad del presente aprendizaje es entrenar una aplicación de un conjunto de variables independientes "x" en una variable *output* "y", esto último se es sabido de antemano. Además, si la variable *output* "Y" es continua se habla de un problema de regresión, mientras que cuando es nominal o discreta (si o no, moroso o no moroso, etc), se habla de un problema de clasificación.
   - Regresión: La variable de salida es continua (temperatura, ventas, etc) las cuales no se pueden organizar en grupos, cada uno de ellos son distintos entre sí.

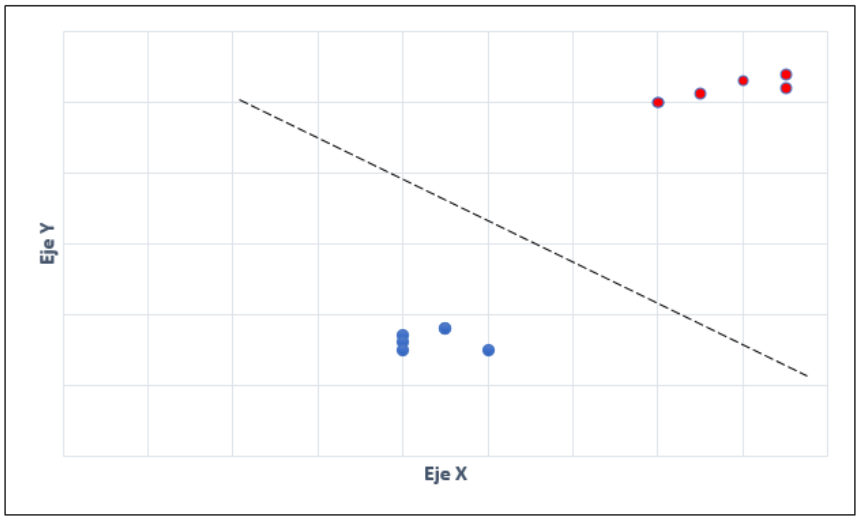
Figura 14: Gráfico Algoritmo de Regresión



Nota. Elaboración propia

- Algoritmo de Clasificación: El modelo busca predecir a qué grupo pertenece el ingreso de un nuevo dato, esto es posible una vez que se haya determinado las características que comparten cierto grupo de datos, el grupo ya se es una variable que se conoce, con lo cual se espera que el algoritmo nos diga a qué grupo pertenece el elemento en estudio. La variable por predecir es un conjunto de estados discretos o categóricos. Pueden ser: Binaria: sí o no, azul o rojo, etc. Múltiple: comprará productos: 1, 2 o 3. Ordenada: Riesgo-Bajo, Medio o Alto, entre otras.

Figura 15: Gráfico Algoritmo de Clasificación



Nota. Elaboración propia

2.2.3 Aprendizaje no supervisado

Según Ethem Mining (2020), el aprendizaje no supervisado no posee una clasificación o datos que se encuentren etiquetados. Solo está comprendido de datos de entrada, por lo que se encarga de agrupar características comunes y crear patrón de comportamiento basado en los datos que encuentre. Este método tiene la capacidad de interpretar características de datos ayudando a obtener una mayor comprensión a los analistas de diferentes áreas. Principalmente este tipo de aprendizaje se enfoca en tareas de agrupamiento. A continuación, se describirán las principales técnicas del aprendizaje no supervisado:

2.2.3.1 Clustering

Técnica de aprendizaje no supervisado basado en la clasificación por grupos de datos por medio de algoritmos.

Gonzáles, H. y Ticona, U. (2019), menciona que el objetivo de esta técnica es encontrar *clusters* (grupos) de acuerdo al comportamiento del conjunto de datos según la similitud de dicha información, es decir agruparlos de acuerdo a las características o similitudes que comparten dentro del mismo.

De acuerdo con Wiskott L. (2014): "el aprendizaje por Clustering es una técnica donde no se tiene en cuenta a qué grupo pertenece verdaderamente cada observación. Esta particularidad es lo que diferencia al clustering de otros métodos de clasificación (p.2)."

Los algoritmos de clustering pueden ser clasificados Partitioning, donde se requiere que especifique previamente el número de *clusters* a crearse y todos los grupos estén en el mismo nivel, Por otro lado, está el Hierarchical Clustering, donde no se requiere especificar previamente el número de conglomerados.

2.2.3.2 Algoritmo K - means

El *K-means* es un algoritmo partitioning (particional).

"Este algoritmo agrupa los datos en un número predefinido de K *clusters* de forma que, la suma de las varianzas internas de los *clusters*, sea lo menor posible". MacQueen (1967) citado en Torres, P, Gonzáles, J., López, V. y Vaca, S. (2020).

Con el modelo se espera subdividir un conjunto de datos, en K *clusters*, permitiendo de esta manera que se tenga la media más cercana a un punto central. Por lo que del total de conjunto de datos X, donde cada dato representa un valor real de dimensiones, *K-means*, se realiza las agrupaciones que permiten minimizar la suma de los cuadrados entre los puntos de datos y los puntos centrales dentro de cada grupo, obteniéndose μ que es la media de los puntos y minimizando las desviaciones cuadradas por pares de puntos en el mismo *cluster* o grupo.

Se tiene la siguiente fórmula:

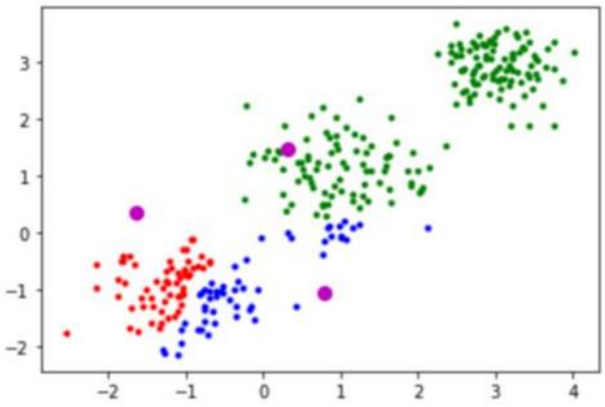
$$\sum_{x \in S_i} ||X - \mu||^2 = \sum_{X \neq y \in S_i} (X - \mu_i)(\mu_{i-y})$$

Para obtener el desempeño de *K-means* se tiene que seguir los siguientes pasos:

- Primer paso: Para determinar el número de K *clusters* a crear y especificar de forma aleatoria las K observaciones que serán los centroides iniciales del modelo (x,y).

- Segundo paso: Identificar las observaciones al centroide con el cual esté más cercano, teniendo en cuenta el valor medio de las observaciones.

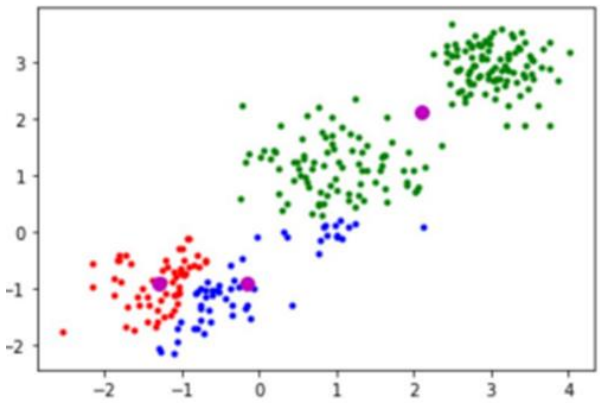
Figura 16: Asignación de puntos a los centroides (K-means)



Nota. Se visualiza los puntos de los centroides en color morado de cada cluster (puntos verde, rojo y azul). Obtenida de *Unioviedo,* 2020

- Tercer paso: Para cada K grupos recalcular los centroides o puntos medios para luego actualizar la información de acuerdo recálculo.

Figura 17: Reubicación de centroides (K-means)



Nota. Se señalan los nuevos puntos de los centroides (color morado) que toma el promedio de la distancia del dato en su cluster (puntos verde, rojo y azul). Obtenida de Unioviedo (2020)

- Cuarto paso: Se deben repetir los pasos dos y tres hasta que el algoritmo *K-means* llegue a converger, es decir, cuando las asignaciones ya no cambian. Se puede implementar este algoritmo tanto en el lenguaje de Python, R, entre otros.

2.2.3.3 Algoritmo K-Medoids

El algoritmo *K-Medoids* se usa para encontrar *Medoids* en un grupo que es el punto ubicado en el centro de un grupo. *K-Medoids* es más robusto en comparación con *K-Means*, ya que en *K-Medoids* encontramos "k" como objeto representativo para minimizar la suma de las diferencias de los objetos de datos, mientras que *K-Means* usa la suma de las distancias euclidianas al cuadrado para los objetos de datos. Y esta métrica de distancia reduce el ruido y los valores atípicos. (Arora et al., 2016)

Inconvenientes del algoritmo *K-Means*:

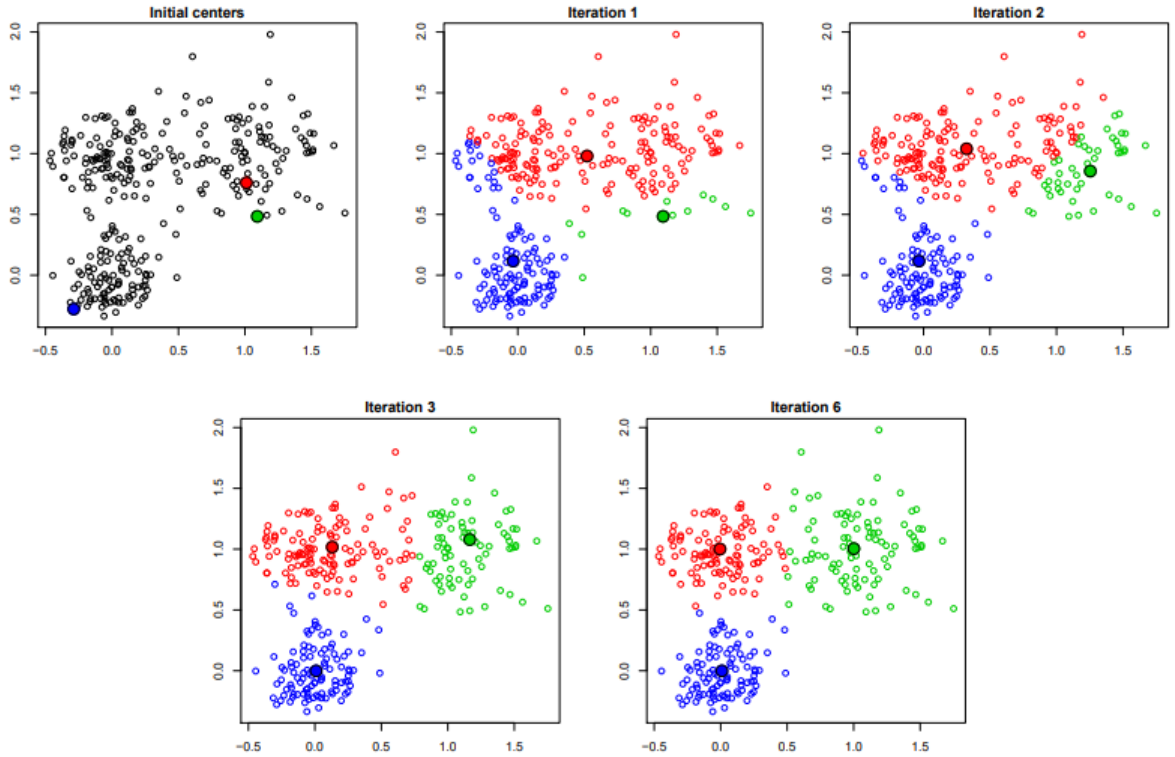
- Encontrar el valor K es una tarea difícil.
- No es efectivo cuando se usa con un cluster global.
- Si se han seleccionado particiones iniciales diferentes, puede variar el resultado para los *clusters.*
- El algoritmo no maneja un cluster de diferente tamaño y diferente densidad. (Arora et al., 2016)

El algoritmo *K-Medoids* comparte las propiedades de *K-Means*: cada iteración disminuye el criterio:

- El algoritmo siempre converge.
- Diferentes comienzos dan diferentes respuestas finales.
- No alcanza el mínimo global.

K-Medoids generalmente devuelve un valor más alto que *K-Means*. Además, *K-Medoids* es computacionalmente más difícil que *K-Means* porque calcular el *Medoid* es más difícil que calcular el promedio. (Tibshirani, 2013)

Figura 18: Obtención de clusters (K-medoids)



Nota. Proceso de obtención de clusters a través de *K-Medoids*. Obtenido de Tibshirani (2013).

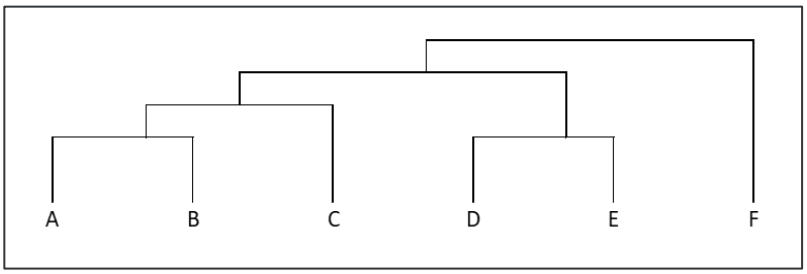
Por último, *K-Medoids* es mejor en términos de tiempo de ejecución, no es sensible a los valores atípicos y reduce el ruido en comparación con *K-Means*, ya que minimiza la suma de las diferencias de los objetos de datos. (Arora et al., 2016)

2.2.3.4 Clustering Jerárquico

El *Clustering* jerárquico desarrolla una estructura de elementos que se organizan o agrupan en subconjuntos cada vez más grandes hasta el punto en que todos los elementos pertenecen a uno mismo.

La representación gráfica de este tipo de *Clustering* se realiza a través del Dendrograma el cual presenta el orden en que se han unido los *clusters* y cuál es el nivel de aproximación que existe en los *clusters* que se han agrupado.

Figura 19: Ejemplo Dendograma.



Nota. Representación gráfica de un dendograma.

Existen 2 tipos de métodos jerárquicos:

- *Clustering* Jerárquico Aglomerativo: Se parte de *clusters* individuales en los cuales cada *cluster* alberga un solo componente. Luego de cada iteración se agrupan aquellos *clusters* más próximos. Este proceso continúa hasta obtener un único cluster.
- *Clustering* Jerárquico Divisivo: Inicia colocando todos los elementos en un único *cluster* y luego se van subdividiendo en partes más pequeñas. Este tipo de métodos son aplicados con muy poca frecuencia ya que es muy difícil encontrar reglas efectivas que permitan dividir en casos y es costosa su implementación. (Tan, Steinbach y Kumar, 2006)

Los algoritmos más usados para el clustering jerárquico son:

- Enlace simple: Consiste en determinar la distancia mínima entre dos componentes dentro de diferentes *clusters*, la norma para las siguientes combinaciones establece que los *clusters* que se unirán serán aquellos que tengan la menor distancia.

- Enlace completo: La proximidad entre *clusters* será entre elementos que posean máximas distancias, dichos componentes se encontrarán en diferentes *clusters*.
- Promedio del grupo: Se define de similar manera a las anteriores pero la distancia entre *clusters* está determinada por la proximidad promedio entre todos los pares de elementos que pertenecen a *clusters* diferentes. (Tan, Steinbach y Kumar, 2006)

2.2.3.5 Principal Component Analysis (PCA)

Según Amat (2017) el PCA es una metodología estadística la cual simplifica la complejidad en las distancias muestrales con muchas dimensiones. Para explicar mejor este concepto se plantea lo siguiente: una muestra con n cantidad de componentes donde se tiene un espacio muestral de "p" dimensiones ($X_1, X_2, \dots, X_p$). El PCA posibilita hallar un número de factores (z<p) que definen aproximadamente lo mismo que las "p" variables originales. Antes se requería p valores para definir a cada componente, ahora se redujo a "z" valores. A cada una de estas "z" nuevas variables se le denomina componente principal. Por lo cual, el PCA permite sintetizar la información otorgada por muchas variables en solo unos pocos componentes. Esto no significa desechar las informaciones de las otras variables.

A continuación, se explican 2 términos aplicados en el PCA:

- Eigen vectors: Se presenta la siguiente operación:

Figura 20: Multiplicación Matriz y un vector

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4\, x \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Nota. El resultado de la multiplicación es un múltiplo entero del vector original. Obtenido de Amat (2017).

"Los *eigenvectors* de una matriz son aquellos vectores que, al multiplicarlos por la matriz en mención, dan como resultado en el mismo vector o en un múltiplo entero del mismo" (Amat, 2017).

- *Eigenvalue*:

  "...Al multiplicar una matriz por alguno de sus eigenvectores da como resultado un múltiplo del vector original, es decir, el mismo vector multiplicado por un número. Al número que multiplica al eigenvector resultante se le denomina como eigenvalue. Todo eigenvector le corresponde un eigenvalue y viceversa" (Amat, 2017)

Por último, el análisis de PCA permite, principalmente, "sintetizar" la información proporcionada por un conjunto de variables en pocos componentes. Por lo que es un método de gran utilidad al aplicarlo antes de la implementación de otras técnicas estadísticas como regresión o *clustering*. (Amat, 2017)

2.2.4 Sistema de Recomendadores

Todas las personas realizamos actividades de compra por lo que estamos inmersos en tomar decisiones frente a una gran cantidad de opciones entre las que elegir, por lo que existen opciones de recomendaciones para filtrar dichas posibilidades permitiéndonos ganar tiempo y mejorar la calidad de la decisión tomada.

Se entiende por sistema de recomendadores a un proceso de asistencia social de decisión basada en conocimiento de acuerdo al perfil del usuario, mediante el uso de algoritmos de inferencia que permiten identificar el grado de correlación entre sus preferencias y los productos, servicios o contenidos existentes (Aggarwal, 2016; Carrer-Neto et al., 2012).

Algunas técnicas mediante las que estos sistemas calculan, evalúan, construyen y proporcionan sus resultados:

- Recomendación colaborativa: A los agentes inmersos en el proceso de selección se le dispondrán aquellos elementos elegidos anteriormente por otras personas con gustos similares.

Se indica que se tienen inmersos algunos estereotipos de por medio como mecanismo para construir modelos de usuarios basándose en una cantidad limitada de información sobre cada usuario.

Este modelo es el más utilizado actualmente. Algunos algoritmos colaborativos se basan en modelos cuyo fundamento es:

- Redes bayesianas y derivados.
- *Clustering* para filtrado colaborativo.
- Análisis iterativo de la componente principal.
- Recomendación basados en contenido: Al ente inmerso en el proceso se le recomendarán ítems parecidos a aquellos que eligió anteriormente; basados en ítems contenedores de información textual. Se usan perfiles con información relativa a los usuarios, sus gustos, preferencias y necesidades.
- Recomendación demográfica: A los entes se los clasifica de acuerdo con los datos demográficos los cuales agrupan atributos personales de una base de datos ya recolectada desde la cual se proporcionan recomendaciones potencialmente interesantes para cualquier persona perteneciente a dicho grupo demográfico.
- Recomendación basados en conocimiento: Se genera a partir de una red de asociación en la que se aplica un razonamiento que indica qué producto cumple los requerimientos del usuario. De esta manera se deja de lado valoraciones personales.
- Recomendación basados en utilidad: La recomendación se basa en el cálculo de la utilidad de cada objeto en particular con respecto a los rasgos identificados previamente para cada usuario.

Capítulo III: Entorno Empresarial

3.1 Descripción de la empresa

3.1.1 Reseña histórica y actividad económica

Hipermercados Tottus se especializa en la comercialización al minoreo de alimentos (perecibles y no perecibles) y no alimentación (aseo individual, vestuario, electrodomésticos, mejoramiento del hogar, entre otros) dentro del área *retail*. Nace de la extensión de la empresa Saga Falabella, la cual inicia operaciones en Perú en el rubro de Hipermercados a través de Hipermercados Tottus en el 2002. La cadena abrió su primer local en el centro comercial MegaPlaza en el distrito de Independencia (Lima). Más adelante, aperturaron las tiendas en San Isidro, San Miguel, San Juan de Miraflores y Puente Piedra. Hoy en día cuenta con más de 55 locales en todo el Perú, además cuenta con Hiperbodega Precio Uno que ya suma más de 10 tiendas. Actualmente, la cadena de supermercados tiene presencia a nivel nacional en las ciudades de Lima, Áncash, Arequipa, Cajamarca, Callao, Huánuco, Ica, La Libertad, Lambayeque, Piura, Junín, Cusco y Ucayali.

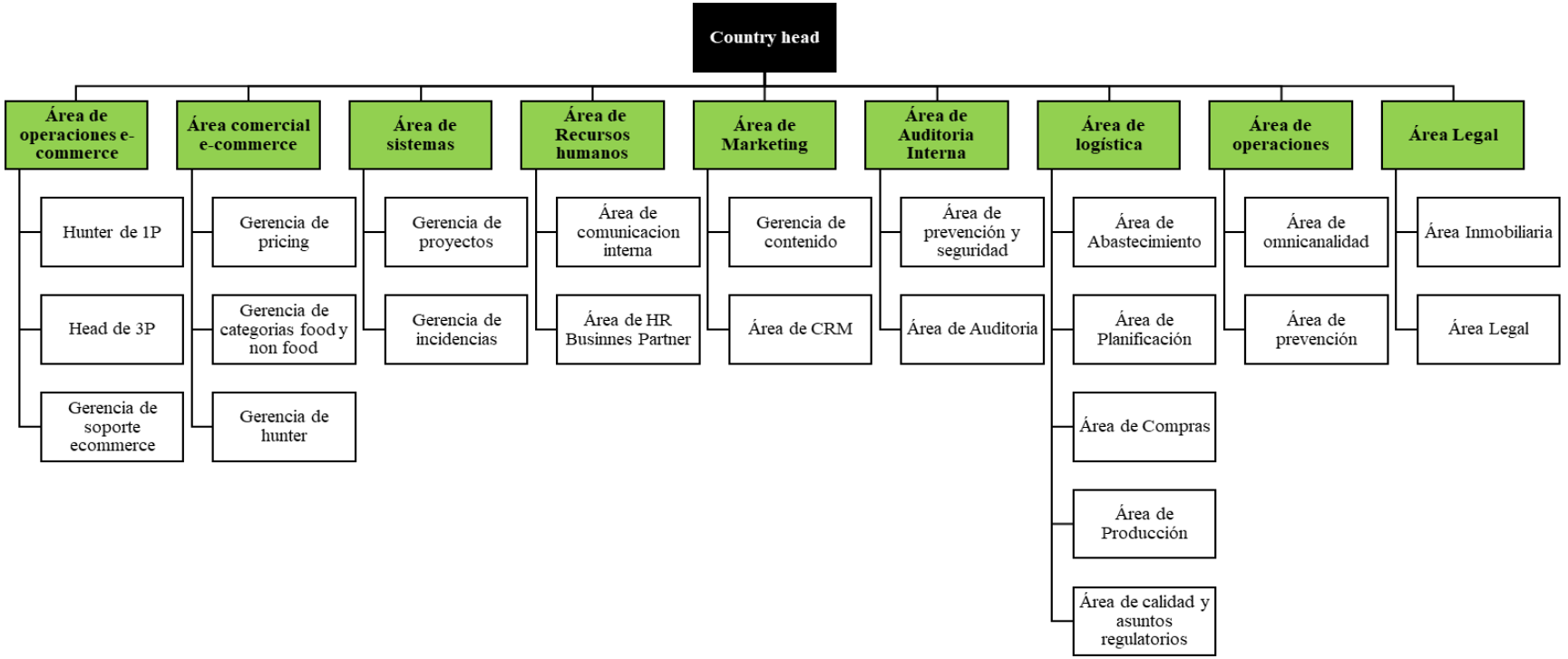
Tottus ha podido desarrollar marcas propias en su portafolio de menestras, aceite, arroz, productos de limpieza, entre otras. Por lo que esta categoría representa alrededor del 10% de las ventas totales de la cadena en el mercado peruano.

Cuenta con centros de producción para la línea de alimentos y tiendas grises (almacén de productos electrónicos) y desarrolló sus propios Centros de Producción en Chile y Perú, asegurando la selección, control y producción con los más altos estándares de inocuidad. También les exigen a todos sus proveedores las certificaciones y cumplimiento de normativas, con especial énfasis en sus marcas Tottus. En Perú, nuestra moderna Planta de Producción de Alimentos de Huachipa se encuentra ubicada en el mismo complejo donde operan el Centro de Distribución Frescos y el Centro de Distribución Secos. (Lucidez, 2018)

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

Figura 21: Organigrama de Tottus



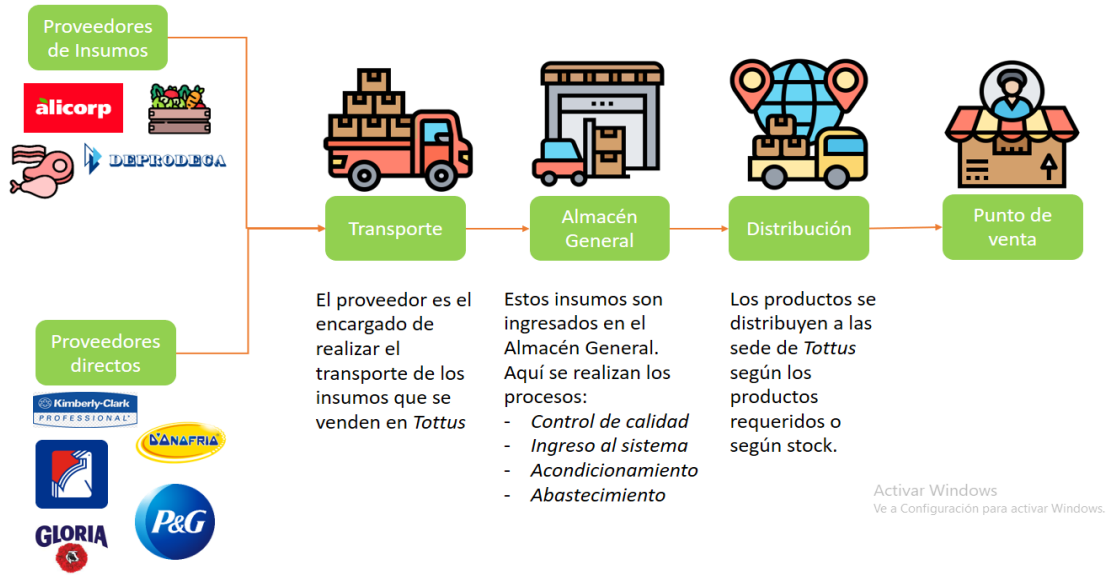
Nota. Elaboración propia

3.1.2.2 Cadena de suministros

La cadena de suministro para la empresa comienza con la obtención de insumos y productos finales de parte de sus proveedores, luego de que Tottus realiza la solicitud. Para la obtención de los insumos, *Alicorp* y *Deprodeca SAC* proveen a la empresa los productos que formarán parte de su oferta, así como los productos de marca propia. En el caso de productos finales que serán destinados a la comercialización, se tiene como proveedores a empresas como *P&G*, *Donofrio*, *San Fernando*, *Gloria*, entre otras marcas.

Seguidamente el proceso continúa con el transporte de los productos de las instalaciones de los proveedores hasta el almacén general ubicado en el distrito de Huachipa, una vez en el lugar se realizan los procesos de: control de calidad, ingresos de las entradas al sistema, acondicionamiento y almacenado. Luego de ubicados se revisa la guía de requerimientos de las tiendas según la solicitud ya sea por reposición o quiebre de stock correspondiente. En el punto de venta que comúnmente son las tiendas físicas, los reponedores se encargan de la disponibilidad los productos en góndolas, estos reponedores pueden ser de la marca proveedora (reponedores externo), como también los reponedores que trabajan en la planilla directa de *Tottus*. Los reponedores son los encargados de la disponibilidad los productos para que los clientes finales puedan verlos y comprarlos.

Figura 22: Cadena de suministros de Tottus



Nota. Elaboración propia

3.1.3 Datos generales estratégicos de la empresa

3.1.3.1 Visión, misión y valores o principios

Visión

Somos líderes en cada mercado donde competimos para ofrecer el lugar preferido para comprar y trabajar.

Misión

Ahorrarles dinero a las familias para que vivan mejor.

Los tres **valores** de la empresa son:

- **Integridad**: Actuar con respeto, honestidad y compromiso. Ser íntegro es: Ser coherente entre lo que digo y lo que hago.
- **Innovación:** Tomar la iniciativa, encontrar nuevas formas de impresionar a nuestros clientes. Ir más allá de las expectativas de mi cliente.
- **Excelencia:** Tener pasión por los productos ganadores. Ser un campeón en el servicio. Trabajar como un gran equipo. Hacer que las cosas mejoren. Tener actitud positiva. Pasión por ser los mejores en lo que hacemos.

3.1.3.2 Objetivos estratégicos

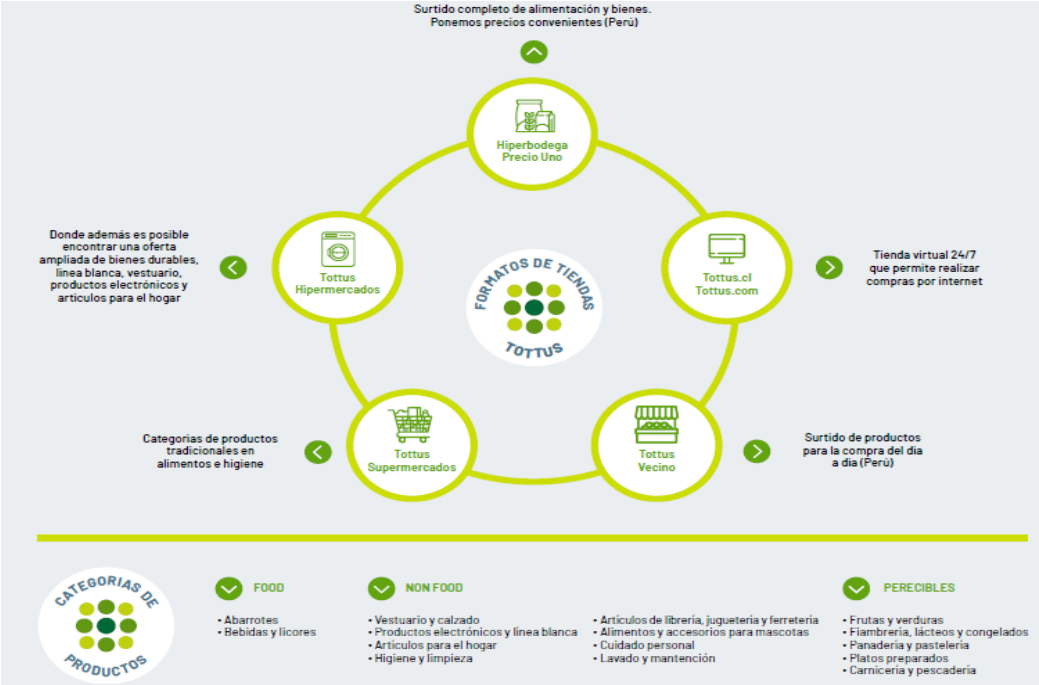
La empresa Tottus considera los siguientes objetivos:

- Aumento de eficiencias de procesos y optimización de costos y autoatención.
- Trabajo colaborativo basado en los programas de cumplimiento del grupo y fortalecimiento de la cultura organizacional.
- Crecimiento de *e-commerce* y robustecimiento de la estrategia omnicanal para cliente digital (mix, físico y digital).
- Aumento en tiendas de la capacidad de autoservicio y procesos de pago.
- Lanzamiento de los productos de la marca propia.

Con respecto a la estrategia omnicanal la empresa cuenta con un modelo de negocio el cual se centra en brindar comodidad, accesibilidad y solución a las necesidades de los clientes.

Por ello, han desarrollado una propuesta de valor omnicanal que cuenta con cuatro formatos de tiendas físicas, más una tienda virtual en internet (ver Figura 23) en donde ofrecen variedad de productos de marcas nacionales e internacionales de excelente calidad, clasificados en tres grandes categorías: *Food*, *Non Food* y Perecibles.

Figura 23: Modelo de negocio de la estrategia omnicanal

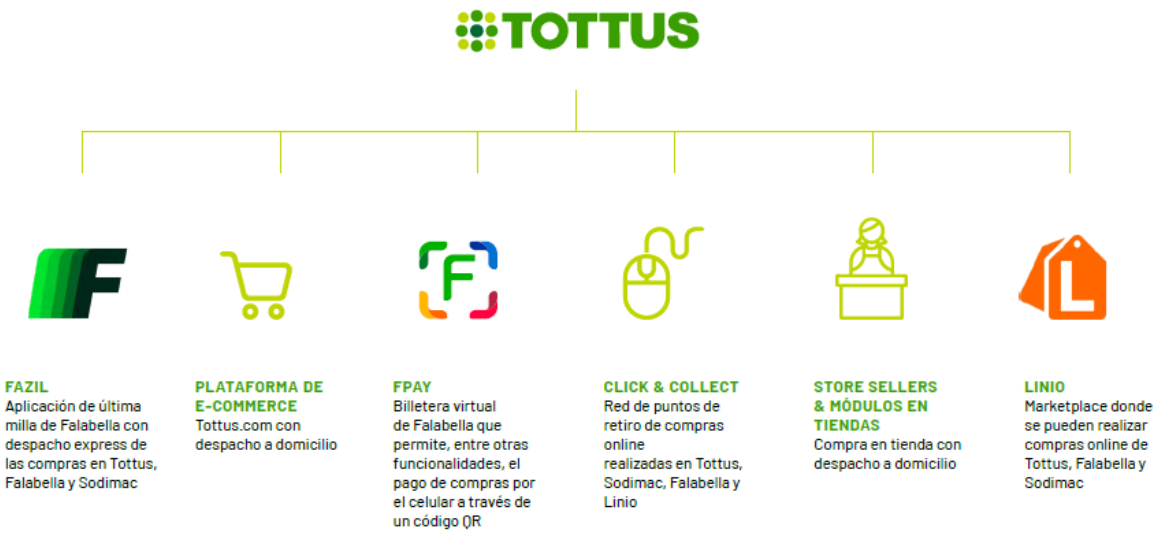


Nota. Diagrama que explica los cinco modelos de estrategia omnicanal. Obtenido de: Reporte de Sostenibilidad Tottus 2020

Siguiendo con el desarrollo de la estrategia omnicanal, Tottus cuenta a la fecha con seis soluciones de omnicanalidad (Ver Figura 24). En abril del 2020 lanzaron “Fazil”, un aplicativo de última milla con despacho *express*, para compras de productos en Tottus, Sodimac y Falabella.

Esta plataforma digital está orientada a satisfacer las necesidades de los clientes basados en: rapidez de la entrega, una robusta red logística y del respaldo de un equipo dedicado a gestionar la experiencia de compra en forma efectiva.

Figura 24: Soluciones de negocio de la estrategia omnicanal



Nota. De Reporte de Sostenibilidad Tottus 2020

3.1.3.3 Evaluación interna y externa

En la siguiente tabla se procede a definir las Fortalezas, Debilidades, Oportunidades y Amenazas de la empresa Tottus:

*Tabla 6:*Matriz FODA de Tottus

FORTALEZA	
F1	Variedad de productos
F2	Productos a menor precio
F3	Locales ubicados estratégicamente
F4	Cuenta con página web y app para realizar compras online

DEBILIDADES	
D1	Percepción negativa de peruanos por empresas chilenas
D2	Baja presencia en los demás departamentos del Perú
D3	Alta rotación de sus colaboradores
D4	Baja cartera de productos propios de la marca

OPORTUNIDADES	
F1	Incremento de compras online en el sector retail por la pandemia COVID-19
F2	Oportunidad de expansión en provincias debido a la disminución de precios en terrenos
F3	Las marcas propias son valoradas por los consumidores actuales
F4	Incremento del Mercado laboral capacitado

AMENAZAS	
D1	Incremento del tipo de cambio (dólar)
D2	Huelgas debido a la baja aceptación de las políticas de gobierno
D3	Incremento de la inseguridad ciudadana
D4	Incremento de la canasta básica familiar

Nota. Elaboración propia

Según Cancino (2012), el FODA se desarrolló a partir de un análisis detallado y diagnóstico interno (fortalezas y debilidades) y el entorno externo (oportunidades y amenazas) con la finalidad de tener una visión 360° del entorno empresarial actual que afronta la empresa.

Bajo esta premisa se procedió a desarrollar la Matriz FODA cuantitativa enfrentando cada una de las Fortalezas y Debilidades con las Oportunidades y Amenazas utilizando una puntuación del 1 al 7:

- Siendo 7 la mayor valorización que obtendrá una fortaleza que aproveche una oportunidad, también, se considerará lo mismo en caso de que la fortaleza ayude a afrontar de la mejor manera la amenaza con la que se cruce.
- La puntuación más alta, 7, para el cruce de debilidades y oportunidades significa que la primera no permite tomar ventaja de la oportunidad. Por otro lado, la misma puntuación se dará en caso una debilidad lleve a la activación de una amenaza.

Para el presente trabajo de investigación cada integrante del grupo ejecutó un análisis por separado que se incluyó en los anexos. La siguiente tabla muestra el resultado de consolidar los 5 análisis.

*Tabla 7:*Matriz FODA Cuantitativo (Promedio)

		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	6	5	6	2	4,7	3	3	2	4	2,9
	F2	6	5	6	2	4,7	4	2	2	5	3,0
	F3	2	7	4	3	4,2	2	3	4	2	2,8
	F4	7	4	4	2	4,3	1	2	4	2	2,3
	PROMEDIO	5,2	5,1	5,1	2,4		2,5	2,5	2,9	3,2	
DEBILIDADES	D1	3	2	3	1	2,4	3	3	1	2	2,1
	D2	4	6	2	1	3,6	1	4	2	5	2,7
	D3	1	2	1	4	2,2	2	3	3	3	2,6
	D4	2	1	4	1	2,1	2	1	2	3	1,9
	PROMEDIO	2,6	3,1	2,8	1,8		1,8	2,7	1,9	3,0	

Nota. Elaboración propia

Según lo observado en la Matriz FODA consolidada, se infiere lo siguiente:

Las Fortalezas 1 y 2: Variedad de productos y Productos a menor precio son las que ayudan a aprovechar mejor las oportunidades detalladas líneas arriba. Por otro lado, la Fortaleza 3: Locales ubicados estratégicamente no tiene la importancia necesaria para explotar las oportunidades descritas. La Fortaleza 2: Productos a menor precio es la que permite afrontar de la mejor manera las amenazas. Por lo tanto, debemos desarrollar aún más las F1 y F2 debido a que nos ayudan a aprovechar mucho mejor las oportunidades y afrontar de la mejor manera las amenazas. Por el lado de las debilidades la D2: Baja presencia en los demás departamentos del Perú es la que representa el principal obstáculo en el aprovechamiento de las oportunidades y también podría impulsar la activación de algunas amenazas. Esta debilidad debe ser foco de mejora en la empresa para que así no interfiera en el aprovechamiento de las oportunidades y activen las amenazas del entorno.

3.2 Modelo de negocio actual

"El modelo de negocio cuenta con actividades estratégicas y la relación entre ellas permite afrontar de la mejor manera los cambios de su entorno y plantear alternativas de mejoras para adaptarse a los rápidos cambios del mercado". El modelo de negocio de la empresa se encuentra detallado en la Figura 25.

El supermercado tiene como propuesta de valor brindar solución a las necesidades de las familias peruanas a través de su servicio omnicanal, por lo que viene realizando ajustes a sus operaciones, aumentando las capacidades en sus Centros de distribución y su red logística de *e-commerce* a través de un ecosistema físico y digital. Asimismo, está fortaleciendo el compromiso con sus clientes de ofrecer productos a precios bajos impulsando marcas propias sin perder de vista la experiencia de compra.

A través de su aplicación móvil brindan al consumidor entregas a un menor tiempo o mucho más rápidas alcanzando una cobertura de 27 tiendas en Perú e impulsaron las ventas de los alimentos alcanzando resultados considerables.

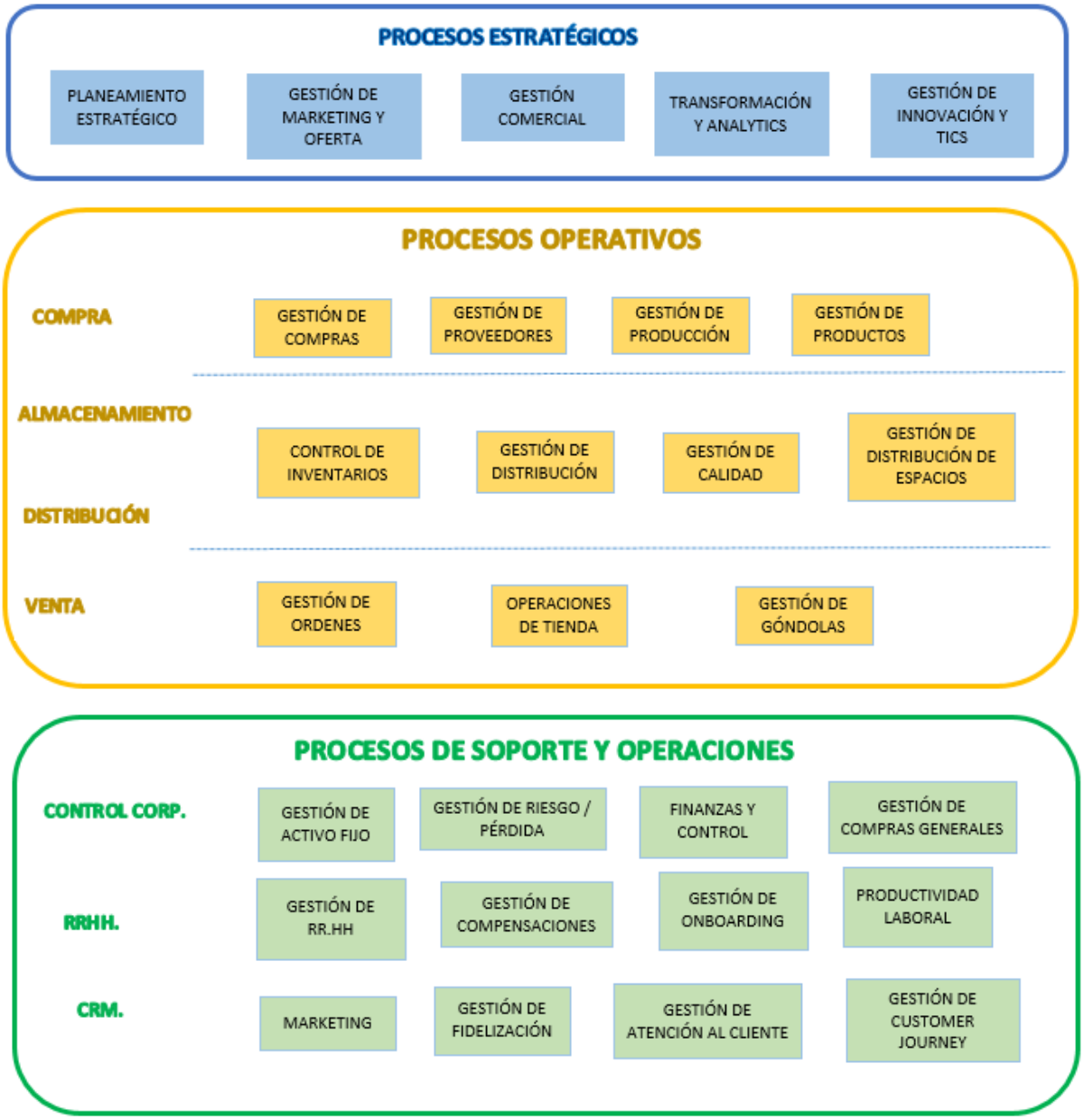
Figura 25: Modelo de negocio actual

Nota. Elaboración propia

3.3 Mapa de procesos actual

El mapa de procesos de la empresa que se presenta en la Figura 26, cuenta con 28 procesos agrupados dentro de las tres partes fundamentales para lograr un desempeño eficiente del negocio: Procesos Estratégicos, Operativos, Soporte y Operaciones.

Figura 26: Mapa de procesos



Nota. Elaboración propia

En los procesos de negocio a nivel estratégico se tiene cinco áreas que interactúan entre sí para conseguir mejores resultados tanto en el corto como en el largo plazo. Forman parte de ellas las áreas de planeamiento estratégico, gestión de marketing y oferta, gestión comercial, transformación y *Analytics*, gestión de innovación y TIC's. Por lo cual la empresa Tottus cuenta con objetivos establecidos, visión y misión direccionada para el desarrollo de planes y estrategias de mejora para la organización.

El proceso operativo consta de 11 áreas los cuales están agrupados en cuatro *Core Business* como son Compra, Almacenamiento, Distribución y Venta los cuales están estrechamente relacionados entre sí para permitir conseguir satisfacer la demanda con la misma calidad de siempre.

Los procesos de apoyo están formados parte de 12 áreas las cuales se agrupan en tres grandes áreas como es control corporativo, recursos humanos y CRM. Estos involucran la gestión de recursos humanos tiene como finalidad afianzar y desarrollar a los trabajadores de la organización los cuales realizan buenas estrategias para poder retener y fidelizar a los clientes.

El proceso que estaremos abarcando con el desarrollo de esta investigación está dentro de los procesos estratégicos como la Gestión de Innovación y TIC en conjunto con el área de marketing y oferta ya que con el modelo se busca mejorar las ofertas dirigidas a los clientes con el uso de las TICS.

Capítulo IV: Metodología De La Investigación

4.1 Diseño de la Investigación

4.1.1 Enfoque de la Investigación

Este estudio es de enfoque cuantitativo, ya que se propone desarrollar un modelo de recomendación de productos a los clientes de acuerdo con su comportamiento de compra, por lo cual se realizará una segmentación de clientes mediante el Análisis de Componentes Principales o PCA y posteriormente la aplicación de tres tipos de técnicas de aprendizaje no supervisados: *Clustering*: *K-Means*, K-*Medoids y Clustering* Jerárquico. Finalmente, se escogerá la segmentación ideal a través de indicadores estadísticos como la *Inertia* y con la validación de un experto.

4.1.2 Alcance de la Investigación

El estudio proporciona un alcance correlacional, debido a que se requiere encontrar grupos de perfiles de compra de los clientes para recomendar productos de acuerdo a sus preferencias y entender la relación entre las variables de la investigación.

4.1.3 Diseño o tipo de la investigación

Estudiaremos las variables a través de las técnicas de *Machine Learning*, por lo que el diseño de estudio es experimental. Del mismo modo, de acuerdo con la metodología planteada se realizará una limpieza de datos en el preprocesamiento, por lo cual las variables serán manipuladas.

4.1.4 Población y Muestra

- Población: Base de datos mensual de las transacciones generadas en Lima a través del aplicativo Fazil de la empresa
- Muestra: Base de datos de las transacciones generadas en abril del 2022 en Lima a través del aplicativo Fazil de la empresa.

4.1.5 Instrumentos de medida

- Instrumento: Técnicas de aprendizaje no supervisado: *K-Means*, *K-Medoids y* Clustering Jerárquico. Mediante la aplicación de las técnicas al conjunto de datos del negocio se busca subdividir en K grupos idóneos, para que a cada dato se le clasifique a un grupo o *cluster* según características en común.
- Responsable: Profesional experto en Marketing, quien comprobará y validará mediante una entrevista el resultado de las técnicas y escogerá la clasificación más conveniente para el negocio.

4.1.6 Operacionalización de variables

El trabajo de investigación tiene como variable independiente las técnicas de *Machine Learning* y como variable dependiente la mejora de la experiencia de compra de los clientes, en la siguiente tabla se describen las variables.

*Tabla 8:*Descripción de variables

Variable	Descripción	Método de evaluación
Independiente: Técnicas de Machine Learning	- Mediante tres algoritmos de aprendizaje no supervisado: *K-means, K-medoids y Clustering* Jerárquico, se hallarán los posibles k grupos según características similares de los datos. - Estos grupos serán comparados por un experto según las características y preferencias de los clientes	- Algoritmos: *K-means*: Método del codo *K-medoids*: Método del codo Clustering Jerárquico: Método del dendograma - Experto: El analista de adquisición de clientes realizará una comparación de los resultados mediante una entrevista en la cual escogerá la técnica que más conveniente para el negocio según las características de los clientes.
Dependiente: Mejora de la experiencia de compra de los clientes	La finalidad de obtener agrupaciones de clientes es generar estrategias de marketing (descuentos, promociones) que estén dirigidas a las preferencias de productos según el grupo al que pertenecen. Es decir, enviar promociones personalizadas a los grupos de clientes.	Para está investigación, no se evaluará la mejora en la experiencia ya que se tendría que implementar las estrategias de marketing según la agrupación de clientes y evaluar en un periodo de tiempo si está genera o no mejoras en la experiencia de compra por el aplicativo de Tottus.

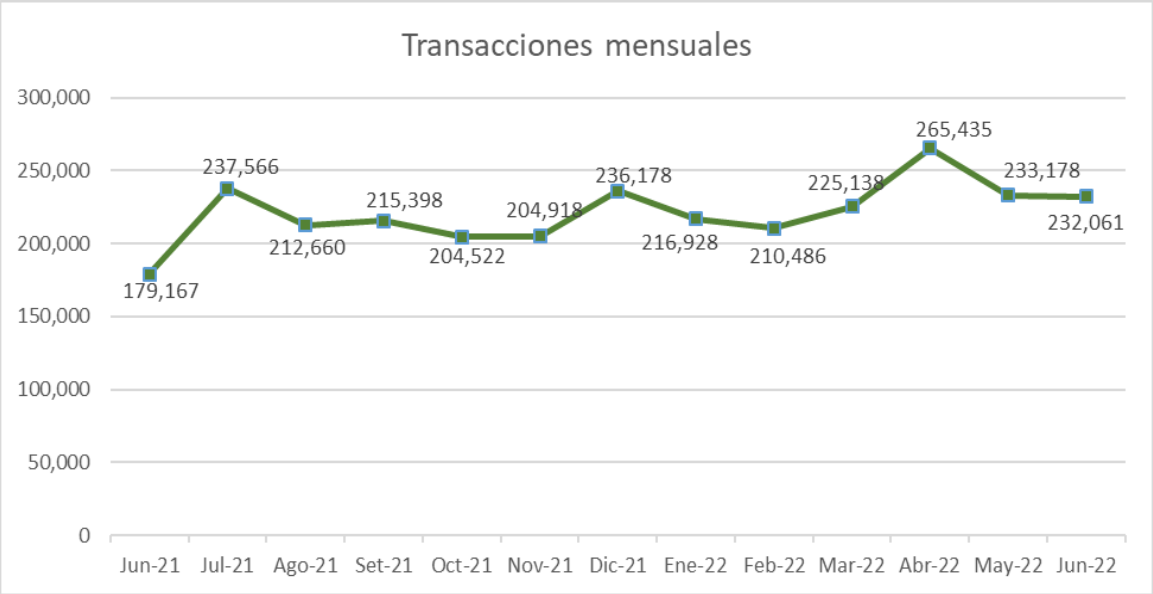
Nota. Elaboración propia

4.2 Metodología de implementación de la solución

- Recopilación de base de datos

Teniendo en cuenta las transacciones generadas desde el mes de junio del 2021 a junio del 2022 se evidencia que la mayor cantidad de transacciones realizadas fueron en los meses de julio 2021, diciembre 2021 y abril de 2022. Como sabemos que los meses de julio y diciembre están asociados a campañas donde se puede producir mayores transacciones, por ello tomamos como muestra el mes de abril del año 2022 con el fin de evitar datos de transacciones inflados de meses atípicos. Por lo tanto, se recopiló los 265,435 registros de transacciones de ventas del mes de abril del 2022.

Figura 27: Transacciones mensuales del aplicativo Fazil desde junio 2021 a junio 2022



Nota. Elaboración propia

- Presentación de variables

De la base de datos se encontraron 66 variables las cuales se indican en la Tabla 10 y serán evaluadas en el preprocesamiento.

- Preprocesamiento

  Antes de aplicar las tres técnicas se validará que los datos de cada variable estén completos, las variables sean características de los clientes, que no exista duplicidad de variables o datos, se transformaran variables categóricas a numéricas y se normalizaron los datos esto último para que todas las variables tengan una misma escala numérica. Luego de ello, mediante la aplicación de PCA se logrará, por la cantidad de datos de la muestra, aplicar más eficientemente las tres técnicas de *Clustering*, ya que reducirá las dimensiones de los datos y por ende será menos pesada para el modelado.

- Modelado

  Una vez limpios los datos se aplicarán tres técnicas de *Clustering* para buscar el K óptimo. En primer lugar, se aplicará *K-Means* y *K-Medoids*, para ello se realizarán diferentes interacciones considerando los valores de K entre dos a 11 y en base a los resultados de la inercia se evaluará el K óptimo. Finalmente, se aplicará *Clustering* Jerárquico para contrastar que coincidan los números óptimos de K escogidos en las dos anteriores técnicas.

- Evaluación de modelos

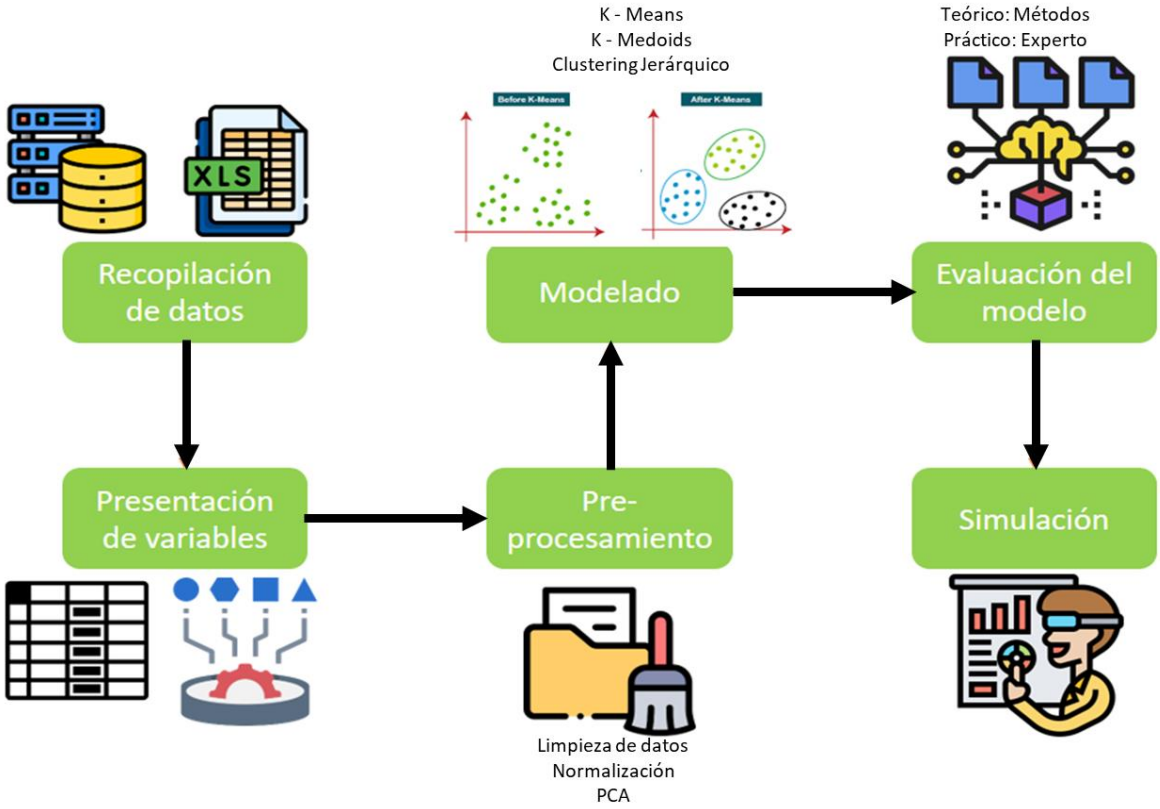
  Se considerará una evaluación teórica mediante las técnicas y una evaluación práctica mediante el experto. La evaluación teórica para la primera técnica (*K-means*) y segunda técnica (*K-Medoids*) se realizará contrastando la *Inertia* con el número de agrupamientos (K) mediante el Método del Codo. Finalmente se aplicará *Clustering* Jerárquico, el cual se evaluará mediante el método gráfico del Dendograma. La evaluación práctica la realizará el experto, quien será el encargado de escoger la técnica más conveniente según los grupos por cada técnica.

- Simulación

  Luego de la evaluación teórica y práctica se procederá a dar más detalle de las características generadas por cada grupo según la técnica escogida por el experto.

En esta etapa se tendría que realizar una prueba por un periodo de tiempo en la que se use la información de cada grupo de clientes para enviar estrategias personalizadas de productos según las características del cliente, mediante está prueba y su evaluación (encuesta a los clientes) se podrá validar si la implementación del agrupamiento de clientes genera o no mejoras en la experiencia de compra.

Figura 28: Metodología de la investigación



Nota. Elaboración propia

4.3 Metodología para la medición de resultados de la implementación

Según lo nombrado líneas arriba, después del desarrollo de los modelos propuestos de recomendación de productos a los clientes de la empresa según las transacciones de ventas de abril 2022, se debe corroborar si la cantidad de *clusters* o grupos (K) seleccionados y agrupados es el correcto. Para ello, se sugiere realizar dos validaciones una teórica y otra práctica:

- **Teórico:**

  Para el algoritmo *K-Means* se basa en el uso del método *Inertia*, la cual nos indica cuán congruente son los *clusters* internamente. Esto se mide como la suma de la distancia al cuadrado de cada punto hasta su centro más próximo o cercano de su agrupación asignada. Se tiene la siguiente fórmula:

$$\sum_{i=0}^{N} ||Xi - \mu|^2$$

  Donde "u" representa el centroide del cluster asignado y Xi es los valores evaluados.
  Lo que se requiere al aplicar el modelo de *K-Means* es seleccionar los centroides que minimicen la *Inertia* o el criterio de suma de cuadrados dentro del grupo.

  Para el algoritmo de *K-Medoids*, la obtención de la cantidad de *clusters* óptimos se realizará con la *Inertia*. Es similar al de *K-Means* pero tiene diferencias teóricas, ya que requiere que los centroides del conglomerado sean miembros del conjunto de entrada, la función objetivo está determinada por:

$$G_{\text{K-medoid}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots, \mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2.$$

Por último, se desarrollará el algoritmo de *Clustering* Jerárquico la cual consiste en la agrupación de elementos en subconjuntos cada vez más grandes hasta que todos estos pertenezcan a uno mismo. El Dendograma representará gráficamente la cantidad de agrupamientos resultantes.

Para ello se considerará lo siguiente:

- Definir cada punto de datos como un grupo: $Cx_n = \{Xkn\}$.
- Hallar los dos conglomerados, Cx_1 y Cy_2, que están más cerca uno del otro y se procederá a dibujar líneas verticales en el gráfico en la parte superior de cada grupo hasta la distancia de estos dos más cercanos.
- Fusionar los dos *clusters* más cercanos en uno, $Cz_1 = Cx_1$, enlazado a Cy_2, es decir, definir un nuevo cluster. Luego se reorganizará los grupos en la abscisa de modo que los dos nuevos más cercanos se conviertan en vecinos, este proceso se repetirá hasta obtener el *cluster* objetivo.

- **Práctico:**

  Tras la aplicación teórica de los 3 algoritmos se obtendrán una serie de resultados óptimos de *clusters*. A fin de determinar la mejor segmentación, se expondrá los resultados obtenidos del modelo a un experto de marketing de la empresa con el objetivo de que pueda evaluar la más idónea y conveniente, además de brindar una retroalimentación respecto la segmentación obtenida por los modelos.

4.4. Cronograma de actividades y presupuesto

A continuación, en la tabla 8 se presenta el cronograma de actividades de nuestra investigación, que se desarrolla desde la última semana de junio de 2022 hasta la última semana de octubre. Asimismo, en la tabla 9 se presenta el presupuesto estimado de los gastos realizados para llevarlo a cabo por todos los participantes de este estudio de la implementación de la propuesta de un sistema de recomendación de productos usando técnicas de *Machine Learning* para mejorar la experiencia de compra de los clientes de Tottus.

*Tabla 9:*Cronograma de Actividades

ACTIVIDADES	JUNIO	JULIO				AGOSTO				SETIEMBRE				OCTUBRE			
	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
CAPITULOI: PLANTEAMIENTO DEL PROBLEMA																	
1.1.Descripción de la realidad problemática																	
1.2. Justificación de la investigación																	
1.3.Demilitación de la investigación																	
CAPITULO II:MARCO TEÓRICO																	
2.1. Antecedentes de la investigación																	
2.2. Bases Teóricas																	
CAPITULO III: ENTORNO EMPRESARIAL																	
3.1. Descripción de la empresa																	
3.2. Modelo de negocio actual(CANVAS)																	
3.3. Mapa de procesos actual																	
CAPITULO IV: METODOLOGÍA DE LA INVESTIGACIÓN																	
4.1. Diseño de la investigación																	
4.2. Metodología de implementación de la solución																	
4.3. Metodología para la medición de resultados de la																	
4.4. Cronograma de actividades y presupuesto																	
CAPITULO V: DESARROLLO DE LA SOLUCIÓN																	
5.1. Propuesta solución																	
5.2. Mediciones de la solución																	
CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES																	
6.1. Conclusiones y recomendaciones																	
ENTREGA FINAL																	
Entrega de trabajo y Sustentación																	

Nota. Elaboración propia

*Tabla 10:*Presupuesto de investigación

Partidas y subpartidas	Unidad de medida	Cantidad	Costo Unitario	Total (S/.)
	Materiales de oficina			
Lapiceros	Unidad	2	S/5.00	S/10.00
Cuadernos	Unidad	4	S/7.00	S/28.00
Post-it	Unidad	3	S/4.00	S/12.00
Resaltador	Unidad	2	S/2.50	S/5.00
Lápiz	Unidad	2	S/1.50	S/3.00
Borrador	Unidad	2	S/1.50	S/3.00
Cinta de embalaje	Unidad	2	S/5.00	S/10.00
			Subtotal	S/71.00
	Materiales impresos/digitales			
Libros	Unidad	7	S/0.00	S/0.00
Articulos	Unidad	15	S/0.00	S/0.00
			Subtotal	0
	Maquinarias y equipos			
Laptops	Unidad	5	S/2,700.00	S/13,500.00
Mouse	Unidad	5	S/20.00	S/100.00
Teclado	Unidad	5	S/30.00	S/150.00
Celulares	Unidad	5	S/1,200.00	S/6,000.00
			Subtotal	19750
	Mano de obra			
Personal	Horas	500	S/5.00	S/2,500.00
			Subtotal	S/2,500.00
	Servicios Básicos			
Internet	Mes	5	S/100.00	S/500.00
Agua	Mes	5	S/35.00	S/175.00
Luz	Mes	5	S/100.00	S/500.00
			Subtotal	S/500.00
			Total	**S/22,821.00**

Nota. Elaboración propia

Capítulo V: Desarrollo de la Solución

5.1 Propuesta solución

5.1.1 Planteamiento y descripción de actividades

Para la metodología a utilizar y el cronograma presentado, se plantea analizar los 3 algoritmos no supervisados mencionados en puntos anteriores en los siguientes pasos:

- Recopilación de datos: La información fue descargada y almacenada en un archivo excel. Esta información son las transacciones mensuales a través del aplicativo *Fazil* del mes de abril del año 2022.
- Presentación de variables: Se definirán las variables para analizar las más importantes y representativas para el modelo.
- Pre-procesamiento de datos: La data será recopilada con el fin de que esta sea óptima Se buscarán valores nulos y vacíos. Asimismo, se realizará el Análisis de Componentes Principales (PCA).
- Modelado: Se aplicará la técnica *K-Means*, Clustering jerárquico y *K-Medoids* para simular los escenarios posibles, analizando los posibles k-óptimos.
- Evaluación del modelo: Se comparan los resultados obtenidos de cada modelo para establecer un algoritmo óptimo.
- Simulación: Se realizará el análisis de los resultados e interpretación de estos.

5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución.

5.1.2.1. Recopilación de datos

La información fue descargada y almacenada en un archivo excel. Esta información son las transacciones realizadas a través del aplicativo *Fazil* del mes de abril del año 2022.

5.1.2.2. Presentación de variables

Se obtuvieron las transacciones realizadas por el aplicativo *Fazil* del mes de abril del año 2022, teniendo así 265.435 registros y 66 variables, de las cuales solo se escogerán las relevantes.

*Tabla 11:*Descripción de variables

Nro	Variable	Descripción
1	Código Orden	Código de orden de compra del cliente
2	Nro. Orden	Número de orden compra del cliente
3	Estado	Estado en el que se encuentra la compra (En Despacho, En Embarque)
4	Sub Estado	Sub Estado en el que se encuentra el pago de la compra
5	Tipo OC	Tipo de Orden de Compra
6	Tipo Despacho	Tipo de despacho realizado en la compra
7	Cliente	Nombre del cliente
8	DNI	DNI del Cliente
9	Sexo	Sexo del cliente
10	Edad	Edad del cliente
11	Celular	Celular del cliente
12	Correo	Correo del cliente
13	Medio Pago	Medio de pago del cliente (CRÉDITO, VISA)
14	Código Reserva	Código de reserva del cliente
15	Distrito	Distrito de la vivienda del cliente
16	Latitud	Latitud del distrito
17	Longitud	Longitud del distrito
18	Dirección	Dirección de la vivienda del cliente
19	Punto Recojo	Punto de recojo de la compra
20	Fecha y Hora Compra	Fecha y hora de compra
21	Fecha y Hora Despacho	Fecha y hora de despacho
22	Fecha Compra	Fecha de compra
23	Hora Compra	Hora de compra
24	Fecha Despacho	Fecha de despacho
25	Hora Despacho	Hora de despacho

26	Código Picking Center	Código del centro de distribución para la compra
27	Picking Center	Nombre del centro de distribución para la compra
28	Código Local	Código del local de procedencia de la compra
29	Local Despacho	Nombre del local de procedencia de la compra
30	SKU Compra	SKU del producto comprado
31	EAN Compra	EAN del producto comprado
32	SKU Pickeado	SKU del producto pickeado
33	EAN Pickeado	EAN del producto pickeado
34	Marca	Marca del producto comprado
35	Producto	Nombre del producto comprado
36	Código División	Código de División (Macro categoría productos - Nivel 1)
37	División	Nombre de División (Macro categoría productos - Nivel 1)
38	Código Departamento	Código de Departamento (Nivel 2 Categoría)
39	Departamento	Nombre de Departamento (Nivel 2 Categoría)
40	Código Sub Departamento	Código Subdepartamento (Nivel 3 Categoría)
41	Subdepartamento	Nombre Subdepartamento (Nivel 3 Categoría)
42	Código Clase	Código de clase (Nivel 4 Categoría)
43	Clase	Nombre de clase (Nivel 4 Categoría)
44	Código Subclase	Código Sub Clase (Nivel 5 Categoría)
45	Subclase	Nombre Subclase (Nivel 5 Categoría)
46	Folio	Expediente de compra
47	Fase Folio	Fase de expediente de compra
48	Estado Folio	Estado de expediente de compra
49	Sub Estado Folio	Sub estado de expediente de compra
50	Estado de Creación Folio	Estado de creación de expediente de compra
51	Precio Unitario	Precio unitario del producto adquirido
52	Precio Regular	Precio regular del producto adquirido

53	Cantidad Solicitada	Cantidad de productos adquiridos
54	Unidad	Unidad de medida de productos adquiridos (UN, KG)
55	Precio Compra Total	Precio de compra total de productos adquiridos
56	Cantidad Pickeada	Cantidad de productos pickeados
57	Precio Venta Despacho	Precio de venta total de despacho
58	Promoción	Promoción realizada a la compra
59	Flete	Costo de flete por compra
60	Fecha Picking	Fecha de picking
61	Fecha Facturación	Fecha de facturación
62	Fecha Envío	Fecha de envío de compra
63	Fecha Entrega	Fecha de entrega de compra
64	Canal	Canal de venta
65	Tipo Orden	Tipo de Orden de Compra
66	Tipo SubOrden	Tipo de SUbOrden de compra

Nota. Elaboración propia

Figura 29: Vista previa de la base de datos en Python

	Codigo Orden	Nro. Orden	Estado	Sub Estado	Tipo OC	Tipo Despacho	Cliente	DNI	Sexo	Edad	...	Precio Venta Despacho	Promoción	Flete	Fecha Picking	Fecha Facturación
0	3341805	604409609	En Despacho	Enviado a POS	SINFOLIO	Despacho Express	Gabriel Palomino	45585502	M	33	...	6.80	NaN	5.0	2022-04-20 19:47:00	2022-04-20 21:42:00
1	3393249	604481411	En Despacho	Registrado en POS	SINFOLIO	Despacho Express	Alessandra Arredondo	76510415	F	32	...	2.31	NaN	5.0	2022-04-26 18:24:00	2022-04-26 18:31:00
2	3393289	604481185	En Despacho	Registrado en POS	SINFOLIO	Despacho Express	solange bast	5644323	F	38	...	9.40	NaN	5.0	2022-04-26 18:31:00	2022-04-26 21:28:00
3	3393282	604481395	En Despacho	Registrado en POS	SINFOLIO	Despacho Express	Alexis Davila Mundo	75696857	M	30	...	2.61	NaN	0.0	2022-04-26 18:30:00	2022-04-26 18:39:00
4	3393249	604481411	En Despacho	Registrado en POS	SINFOLIO	Despacho Express	Alessandra Arredondo	76510415	F	31	...	7.94	NaN	5.0	2022-04-26 18:24:00	NaN

5 rows × 66 columns

Nota. Elaboración propia

5.1.2.3. Pre-procesamiento de datos

En esta etapa se evaluaron los datos de cada variable y por eso se eliminaron los valores nulos, se transformaron las variables categóricas a numéricas y también se normalizaron los datos.

- Selección de variables: Al visualizar las 66 variables de la base de datos, se observó que hay variables que son identificadores y que no deben ser consideradas en el modelo, ya que no explican el comportamiento de la data por ser un ID único. Dentro de esas variables están: "Código de Orden", "Nro. Orden", "Codigo Reserva", "Código Picking Center", "SKU Compra", "EAN Compra", "SKU Pickeado", "EAN Pickeado" y "Folio".

  Asimismo, se encontraron variables que no contienen datos como: "Latitud", "Longitud", "Punto de recojo", "Fase folio", "Estado folio", "Subestado folio", "Estado creación folio", "Promoción", "Fecha envío", "Fecha entrega", "Tipo SubOrden".

  Por otro lado, se eliminaron las variables que no influyen en el modelo: "Estado", "Sub Estado", "Tipo OC", "Cliente", "Celular", "Correo", "Dirección", "Fecha y Hora Compra", "Fecha y Hora Despacho", "Fecha compra", "Hora compra", "Fecha despacho", "Picking Center", "Local Despacho", "Producto", "Division", "Departamento", "Sub Departamento", "Codigo Clase", "Clase", "Codigo Sub Clase", "Sub Clase", "Precio Unitario", "Precio Regular", "Cantidad Solicitada", "Unidad", "Precio Compra Total", "Cantidad Pickeada", "Precio Venta Despacho", "Flete", "Fecha Picking", "Fecha Facturación", "Canal", "Tipo Orden".

  Finalmente, se eliminaron las variables que tiene correlación con otras o son muy dispersas: "Codigo de División", "Codigo de subdepartamento", "Distrito", "Codigo local". Por lo tanto, escogieron las 5 variables más importantes: "Sexo", "Edad", "Medio de pago", "Marca", "Codigo de Departamento".

Figura 30: Eliminación de variables

```
datos_limpios=data.drop(['Codigo Orden','Nro. Orden','Estado','Sub Estado','Tipo Despacho','Cliente',
                         'Celular','Correo','Codigo Reserva','Latitud','Longitud','Direccion',
                        'Punto Recojo','Fecha y Hora Compra','Fecha y Hora Despacho',
                        'Hora Compra','Fecha Despacho','Hora Despacho','Código Picking Center',
                        'Picking Center','Local Despacho','EAN Compra','SKU Pickeado','EAN Pickeado',
                        'Division','Departamento','Sub Departamento','Codigo Clase','Clase',
                        'Codigo Sub Clase','Sub Clase','Folio','Fase Folio','Estado Folio','Sub Estado Folio',
                        'Estado de Creacion Folio','Unidad','Precio Unitario','Precio Regular',
                        'Precio Compra Total','Cantidad Pickeada','Precio Venta Despacho','Promoción',
                        'Flete','Fecha Picking','Fecha Facturación','Fecha Envío','Fecha Entrega',
                        'Canal','Tipo Orden','Tipo SubOrden','Tipo OC','Cantidad Solicitada','Producto',
                        'DNI','SKU Compra','Fecha Compra'],axis=1)
```

Nota. Elaboración propia

Figura 31: Variables seleccionadas

	Sexo	Edad	Medio Pago	Marca	Codigo Departamento
0	M	33	CREDITO	TOTTUS	J0101
1	F	32	visa	TOTTUS	J0401
2	F	38	CREDITO	TOTTUS	J0201
3	M	30	CREDITO	OTROS	J0101
4	F	31	visa	TOTTUS	J0502

Nota. Elaboración propia

- Eliminación de valores nulos: No se detectaron valores nulos.

Figura 32: Eliminación de valores nulos

```
datos_limpios.isna().sum().sum()
0
```

Nota. Elaboración propia

- Transformación de variables: Debido a que la mayoría de las variables son categóricas, estas deben ser transformadas a variables numéricas. Para ello, primero se modifica la variable decimal "Código Local" a una variable numérica. Asimismo, se procedió a transformar todas las variables restantes, excepto "Edad", a variables numéricas a través de la librería *sklearn.preprocessing* (Herramienta: *LabelEncoder)*.

Figura 33: Transformación de variables categóricas a numéricas

```
from sklearn.preprocessing import LabelEncoder
```

```
encoder = LabelEncoder()
datos_nonan['Medio Pago'] = encoder.fit_transform(datos_nonan['Medio Pago'])

encoder = LabelEncoder()
datos_nonan['Marca'] = encoder.fit_transform(datos_nonan['Marca'])

encoder = LabelEncoder()
datos_nonan['Codigo Departamento'] = encoder.fit_transform(datos_nonan['Codigo Departamento'])

encoder = LabelEncoder()
datos_nonan['Sexo'] = encoder.fit_transform(datos_nonan['Sexo'])
```

Nota. Elaboración propia

Figura 34: Variables transformadas

```
datos_nonan.head()
```

	Sexo	Edad	Medio Pago	Marca	Codigo Departamento
0	1	33	0	2	0
1	0	32	1	2	4
2	0	38	0	2	2
3	1	30	0	0	0
4	0	31	1	2	5

Nota. Elaboración propia

- Normalización de variables: Debido a que las variables tienen diferentes escalas numéricas, es necesario normalizar dichas variables a una misma escala. Se descargó la herramienta *StandarScaler* de la librería *sklearn.preprocessing*. A continuación, se muestran los resultados obtenidos.

Figura 35: Variables normalizadas

```
from sklearn.preprocessing import StandardScaler
```

```
normalizador = StandardScaler()
X_norm = normalizador.fit_transform(datos_nonan)
```

```
X_norm
```

```
array([[ 1.54211222,  0.09980339, -0.76430779,  1.10282807, -1.03674035],
       [-0.64846124, -0.09887153,  1.30837343,  1.10282807,  1.08581906],
       [-0.64846124,  1.09317799, -0.76430779,  1.10282807,  0.02453936],
       ...,
       [ 1.54211222, -0.49622137, -0.76430779,  1.10282807,  1.08581906],
       [-0.64846124, -0.09887153,  1.30837343,  1.10282807, -1.03674035],
       [ 1.54211222, -0.29754645, -0.76430779, -0.83828243, -1.03674035]])
```

Nota. Elaboración propia

- Análisis de Componentes Principales (PCA): Para este paso, se procedió a utilizar la herramienta PCA de la librería *sklearn.decomposition* y luego se procedió a la reducción de 3 dimensiones, ya que estos explican el 70% de la data analizada. A continuación, se muestran los resultados obtenidos:

Figura 36: Aplicación de PCA

```
pca = PCA(n_components=3)
pca.fit(X_norm)
```

```
PCA(n_components=3)
```

```
sum(pca.explained_variance_ratio_)
```

```
0.7089183348335666
```

```
X_norm_pca = pca.transform(X_norm)
```

```
X_norm_pca
```

```
array([[-0.27784567,  1.54082594,  0.48899651],
       [ 1.67333671, -0.44002368, -0.16625097],
       [ 1.05125229, -0.06916511, -0.77663443],
       ...,
       [ 0.44302762,  2.21204493, -0.50812709],
       [ 0.6401421 , -1.06281209,  0.65462973],
       [-1.62887668,  1.0978063 ,  0.80074471]])
```

Nota. Elaboración propia

5.1.2.4. Modelado

En esta etapa se realizará el modelado de los 3 algoritmos mencionados previamente.

- Aplicación de *K-Means:* Se realizaron diferentes iteraciones del valor de *K*, hasta obtener el valor o rango de valores de K apropiados para el modelo propuesto. Esta iteración de K está dada entre los valores de 2 a 11. Para la ejecución del *K*-óptimo se utilizó la herramienta *K-Means* de la librería *sklearn.cluster*, mediante el siguiente script:

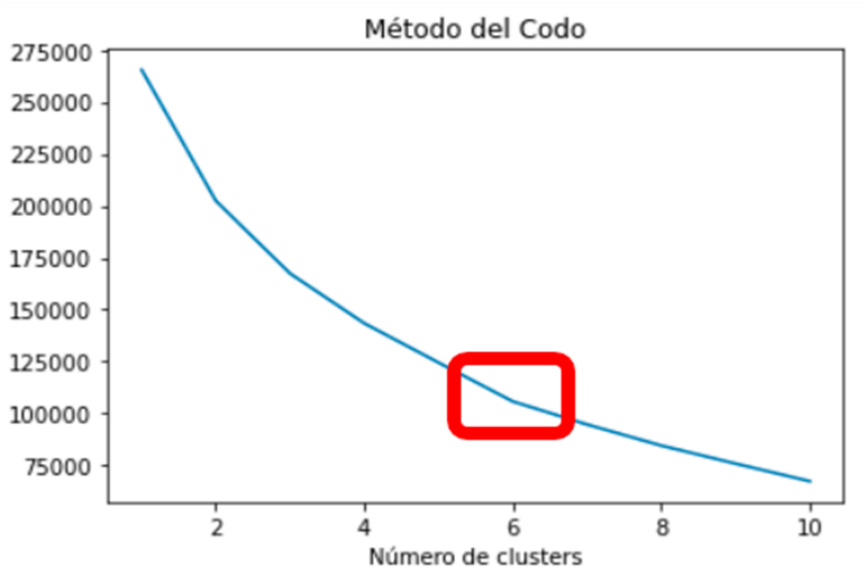
Figura 37: Aplicación de K-Means para K = [2;11]

```
from sklearn.cluster import KMeans
```

```
valores=[]
for k in range(2,11):
    datos_NS = KMeans(n_clusters=k)
    datos_NS.fit(X_norm_pca)
    valores.append(datos_NS.inertia_)
```

Nota. Elaboración propia

Figura 38: Inertia por Cluster (K-Means)



Nota. Elaboración propia

- Aplicación de K-Medoids: Se realizaron diferentes iteraciones del valor de *K*, hasta obtener el valor o rango de valores de K apropiados para el modelo propuesto. Esta iteración de *K* está dada entre los valores de 2 a 11. Para la ejecución del *K*-óptimo se utilizó la herramienta *K-Medoids* de la librería *sklearn_extra.cluster*, mediante el siguiente script:

Figura 39: Aplicación de K-Medoids para K = [2;11]

```
from sklearn_extra.cluster import KMedoids
```

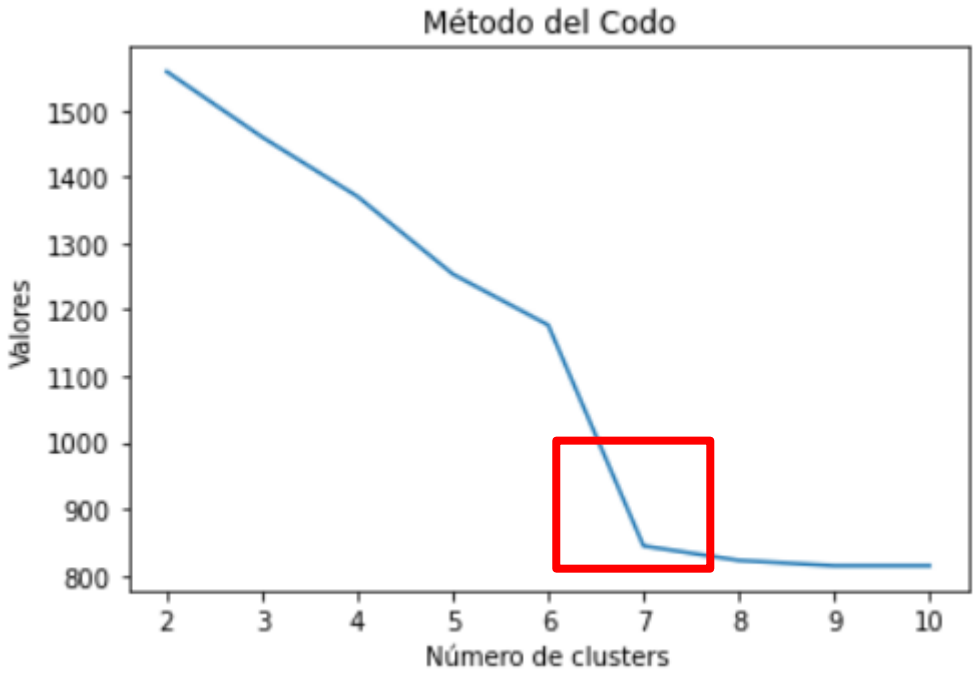
```
X_norm_pca[:1000,:].shape
```

```
(1000, 3)
```

```
valores=[]
for k in range(2,11):
    datos_NS = KMedoids(n_clusters=k)
    datos_NS.fit(X_norm_pca[:1000,:])
    valores.append(datos_NS.inertia_)
```

Nota. Elaboración propia

Figura 40: Inertia por Cluster (K-Medoids)



Nota. Elaboración propia

- Aplicación de *Clustering* Jerárquico*:* Se utilizó el dendograma para poder definir cuál es el número óptimo de *clusters*. Se observa que el número óptimo podría estar a partir de 5 *clusters*.

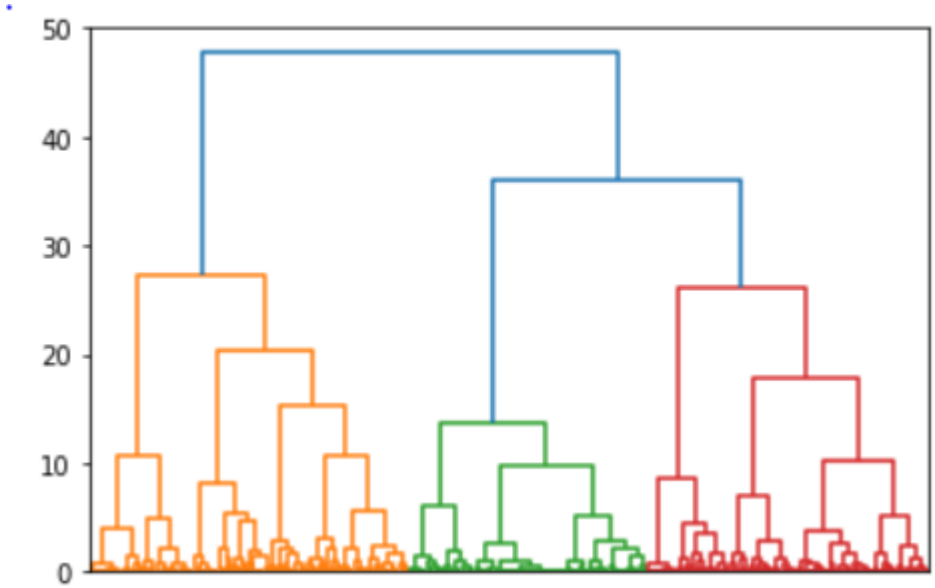
Figura 41: Aplicación del Dendograma

```
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as sch
```

```
dendrogram = sch.dendrogram(sch.linkage(X_norm_pca[:1000,:], method='ward'))
```

Nota. Elaboración propia

Figura 42: Dendograma (Clustering Jerárquico)



Nota. Elaboración propia

5.1.2.5. Evaluación del modelo

Previamente a la ejecución de la transformación y normalización de datos, se realizó una copia de la tabla inicial con las variables seleccionadas. Esto debido a que se quiere agrupar las columnas de los k-óptimo según cada modelo.

Figura 43: Copia de la tabla original

```
datos_nonan_copy = datos_nonan.copy()
```

Nota. Elaboración propia

Para la evaluación del modelo óptimo se generaron 10 modelos de K-Means y K-Medoids respectivamente y a partir de estos se analizaron y evaluaron las métricas de los resultados obtenidos junto con la opinión de un experto, para poder escoger el modelo óptimo y la cantidad de cluster adecuada.

Figura 44: Data etiquetada para evaluar los modelos K-Means y K-Medoids

	Sexo	Edad	Medio Pago	Marca	Codigo Departamento	Cluster Kmeans	Cluster Kmedoids
0	M	33	CREDITO	TOTTUS	J0101	5	5
1	F	32	visa	TOTTUS	J0401	4	1
2	F	38	CREDITO	TOTTUS	J0201	0	4
3	M	30	CREDITO	OTROS	J0101	2	5
4	F	31	visa	TOTTUS	J0502	4	1

Nota. Elaboración propia

Para escoger el K–óptimo, se escogieron dos propuestas presentadas al experto, la primera propuesta es del modelo de *K-Means* con 6 cluster y la segunda con el modelo de *K-Medoids* con 7 cluster. A continuación, se presentan las dos propuestas. Para más detalles de la evaluación y entrevista revisar el Anexo 6.

*Tabla 12:*Propuesta 1 (Modelo K-Means)

		Cluster					
		0	1	2	3	4	5
Variable	Edad	33-37 años	28-37 años	28-37 años	28-37 años	18-27 años	37-43 años
	Sexo	Femenino en su mayoría	Femenino	Femenino	Masculina	Femenino en su mayoría	Femenino en su mayoría
	Medio de pago	Credito: CMR	Visa: Otros medios de pago	Credito: CMR	Credito: CMR en su mayoría	Credito: CMR	Visa: Otros medios de pago
	Marca	Totttus en su mayoría	Otras marcas en su mayoría	Otras marcas en su mayoría	Otras marcas en su mayoría	Otras marcas	Otras marcas en su mayoría
	Departamento	Frutas y verduras	Abarrotes, lavado y mantenimiento	Abarrotes, lavado y mantenimiento	Abarrotes, lavado y mantenimiento	Líquidos y perfumería	Frutas y verduras, lácteos

Nota. Elaboración propia

*Tabla 13:*Propuesta 2 (Modelo K-Medoids)

		Cluster						
		0	1	2	3	4	5	6
Variable	Edad	28-32 años	33-37 años	23-27 años	28-32 años	33-37 años	28-32 años	33-37 años
	Sexo	Femenino	Femenino en su mayoría	Femenino	Femenino en su mayoría	Femenino en su mayoría	Masculino	Femenino
	Medio de pago	Credito: CMR en su mayoría	Visa: Otros medios de pago en su mayoría	Credito: CMR	Visa: Otros medios de pago en su mayoría	Credito: CMR	Credito: CMR en su mayoría	Credito: CMR en su mayoría
	Marca	Otras marcas en su mayoría	Tottus en su mayoría	Otras marcas en su mayoría	Otras marcas en su mayoría	Tottus en su mayoría	Otras marcas (64%) y Tottus (36%)	Otras marcas (51%) y Tottus (49%)
	Departamento	Abarrotes, Liquidos	Frutas y verduras, Lácteos, Lavado y Mantenimiento	Liquidos y Perfumería	Abarrotes, Lavado y mantenimiento	Frutas y verduras, Lacteos	Abarrotes, Frutas y verduras, Liquidos	Abarrotes, Lacteos, Lavado y mantenimiento

Nota. Elaboración propia

5.1.2.6. Simulación de resultados

Con el *K*=7 escogido (Algoritmo *K-Medoids*), se analizó los resultados encontrados, con el objetivo de resaltar los datos más importantes por cluster y así categorizar a cada uno de estos grupos, de manera que sea un impacto positivo para la empresa. Por ello, se analizó la variable "Edad" y se obtuvieron los siguientes resultados:

*Tabla 14:*Cuadro de doble entrada (cluster vs Edad)

Edad						
Cluster	**18-22**	**23-27**	**28-32**	**33-37**	**38-43**	**Total**
0	1215	1525	19920	9320	3100	**35080**
1		10	6235	22995	8950	**38190**
2	6430	15625	670			**22725**
3		25	16825	10510	8200	**35560**
4	15	1285	6345	15790	6180	**29615**
5	4975	9625	24945	23030	8525	**71100**
6	100	145	12690	15060	5170	**33165**
Total	**12735**	**28240**	**87630**	**96705**	**40125**	**265435**

Nota. Elaboración propia

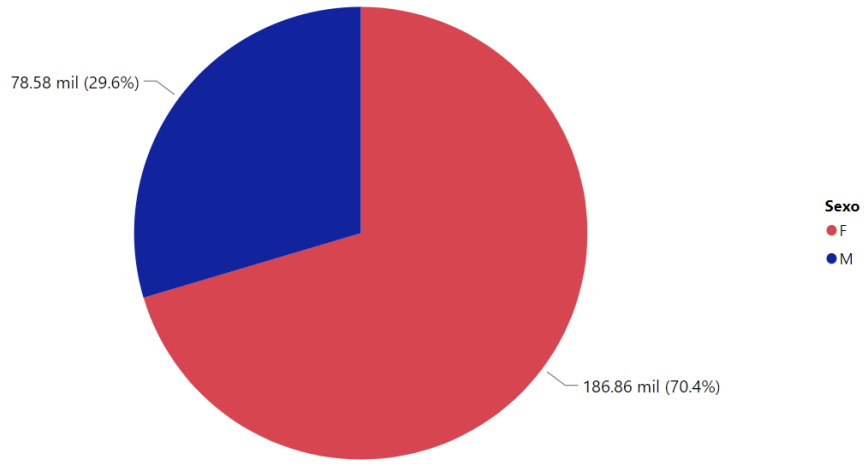
Para segmentar de manera idónea esta variable, se busca tener el % más representativo por *cluster* (aproximadamente mayor al 50%):

De los resultados obtenidos, se deduce lo siguiente:

- Para el *cluster* "0", el rango de edades entre 28 y 32 representa el 56,78% de los datos en este grupo.
- Para el *cluster* "1", el rango de edades entre 33 y 37 representa el 65,55% de los datos en este grupo.
- Para el *cluster* "2", el rango de edades entre 23 y 27 representa el 44,54% de los datos en este grupo.
- Para el *cluster* "3", el rango de edades entre 28 y 32 representa el 47,96% de los datos en este grupo.
- Para el *cluster* "4", el rango de edades entre 33 y 37 representa el 45,01% de los datos en este grupo.
- Para el *cluster* "5", el rango de edades entre 28 y 32 representa el 71,11% de los datos en este grupo.
- Para el *cluster* "6", el rango de edades entre 33 y 37 representa el 42,93% de los datos en este grupo.

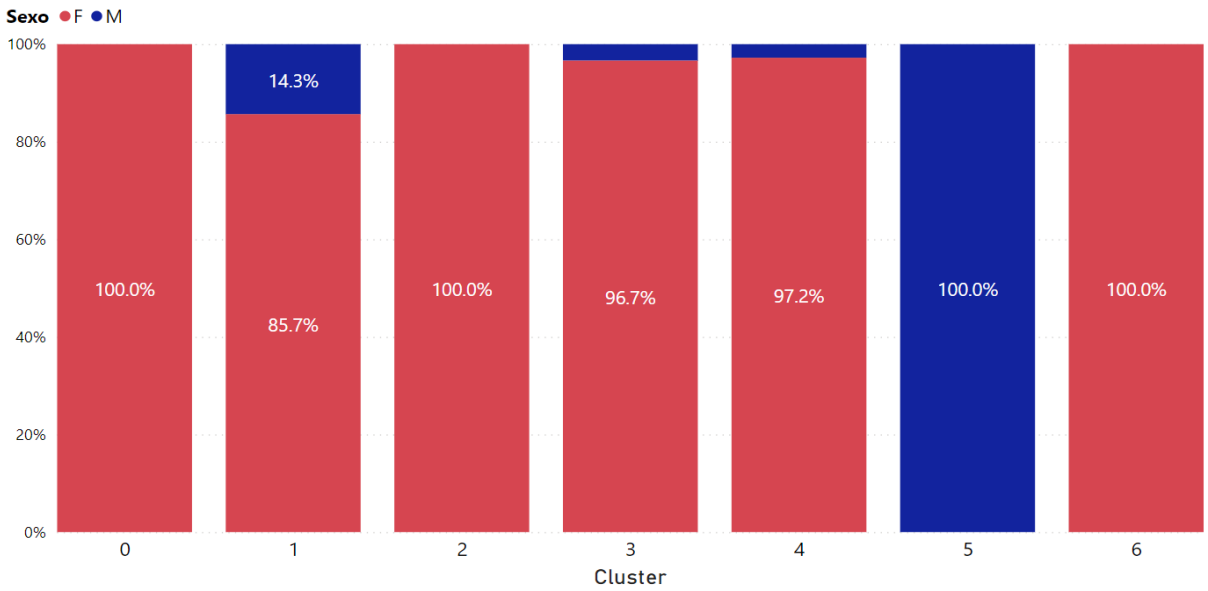
La siguiente variable analizada fue "Sexo" de los clientes para saber la proporción por *cluster*:

Figura 45: Segmentación de clientes por la variable “Sexo”



Nota. Elaboración propia

Figura 46: Segmentación de *clusters* por la variable “Sexo”



Nota. Elaboración propia

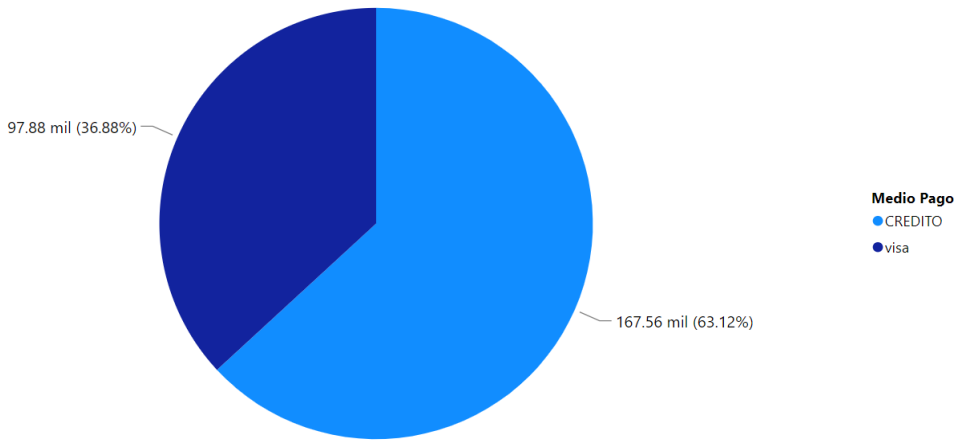
Como se puede visualizar en la Figura 45, en la base de datos general, la proporción de sexo de los clientes es de un 70% mujeres aproximadamente y un 30% hombre aproximadamente. Esto es un dato importante en la toma de decisiones comerciales.

Asimismo, como se visualiza en la Figura 46, se analizó la misma variable por *cluster* y los resultados fueron, en su mayoría, del sexo femenino. Los resultados son los siguientes:

- El *cluster* "0" tiene toda su población del sexo femenino.
- El *cluster* "1" tiene la proporción de 85,7% del sexo femenino y 14,3% del sexo masculino.
- El *cluster* "2" tiene toda su población del sexo femenino.
- El *cluster* "3" tiene la proporción de 96,7% del sexo femenino y 3,3% del sexo masculino.
- El *cluster* "4" tiene la proporción del 97,2% sexo femenino y 2,77% sexo masculino aproximadamente.
- El *cluster* "5" tiene toda su población del sexo masculino.
- El *cluster* "6" tiene toda su población del sexo femenino.

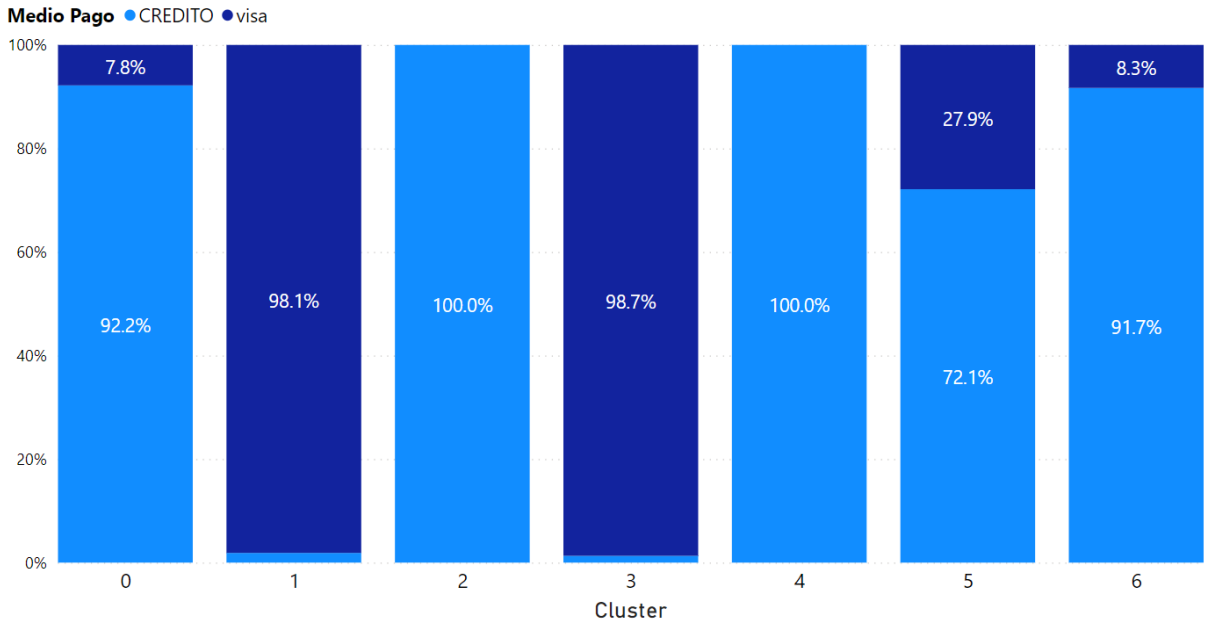
La siguiente variable analizada fue el medio de pago. Se debe tomar en consideración que en esta variable se cuenta con dos posibles resultados: "Crédito" y "Visa". "Crédito" hace referencia a la tarjeta CMR, la tarjeta oficial del grupo Falabella, mientras que Visa hace referencia a otro medio de pago que no sea la tarjeta CMR. Esta variable es importante, debido a que, siguiendo la estrategia comercial, se busca el uso de la tarjeta CMR, como medio de fidelización del cliente con el grupo Falabella.

Figura 47: Segmentación de clientes por la variable "Medio de pago"



Nota. Elaboración propia

Figura 48: Segmentación de clusters por la variable "Medio de pago"



Nota. Elaboración propia

Como se puede visualizar en la Figura 47, en la base de datos general, la proporción de "Medio de pago" de los clientes es de un 63% "Crédito" aproximadamente y un 37% "Visa" aproximadamente. Esto es un dato importante para la toma de decisiones relacionada a la fidelización de clientes.

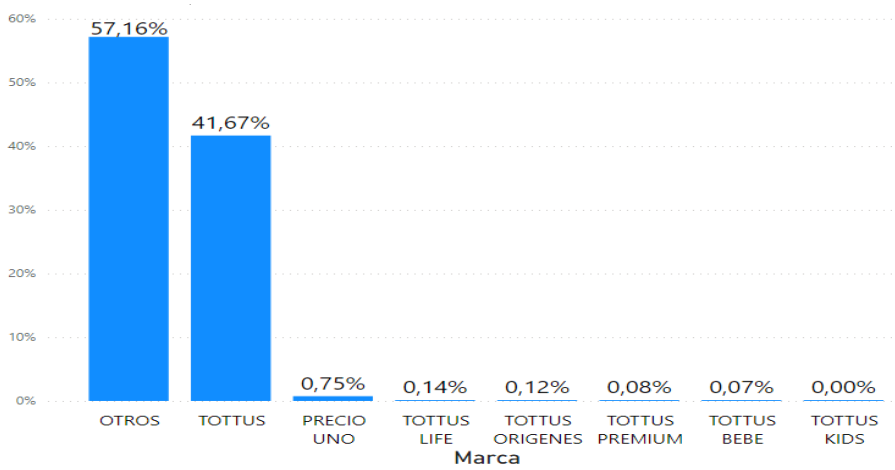
Asimismo, como se visualiza en la Figura 48, se analizó la misma variable por *cluster* y los resultados fueron, en su mayoría, de la categoría de "Crédito", lo cual es un resultado muy favorable. Los resultados por *cluster* son los siguientes:

- El *cluster* "0" tiene la proporción del 92% "Crédito" y 8% "Visa" aproximadamente.
- El *cluster* "1" tiene la proporción del 2% "Crédito" y 98% "Visa" aproximadamente.
- El *cluster* "2" tiene toda su población haciendo uso de la tarjeta CMR.
- El *cluster* "3" tiene la proporción del 1% "Crédito" y 99% "Visa" aproximadamente.
- El *cluster* "4" tiene toda su población haciendo uso de la tarjeta CMR.
- El *cluster* "5" tiene la proporción del 72% "Crédito" y 28% "Visa" aproximadamente.

- El *cluster* “6” tiene la proporción del 92% “Crédito” y 8% “Visa” aproximadamente.

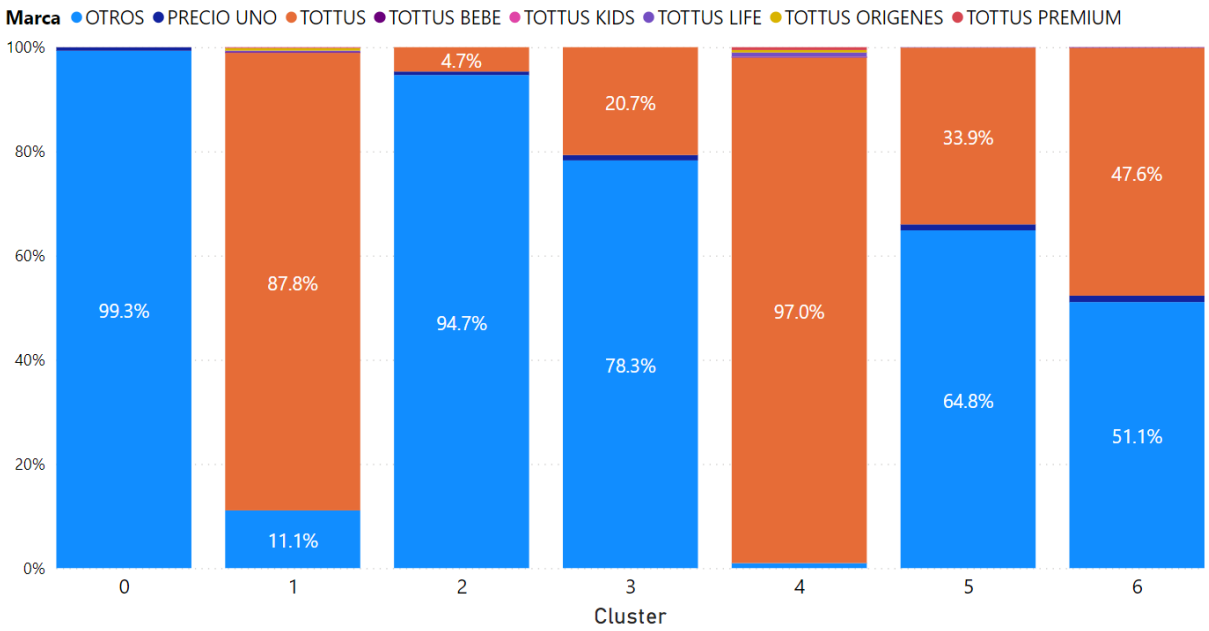
La siguiente variable analizada fue la marca adquirida por el cliente. Esta variable es importante, debido a que, siguiendo la estrategia de mejora de experiencia del cliente, se busca fomentar la compra de marcas propias por parte de los clientes.

Figura 49: Proporción de marcas adquiridas por cluster



Nota. Elaboración propia

Figura 50: Proporción de marcas adquiridas por cluster



Nota. Elaboración propia

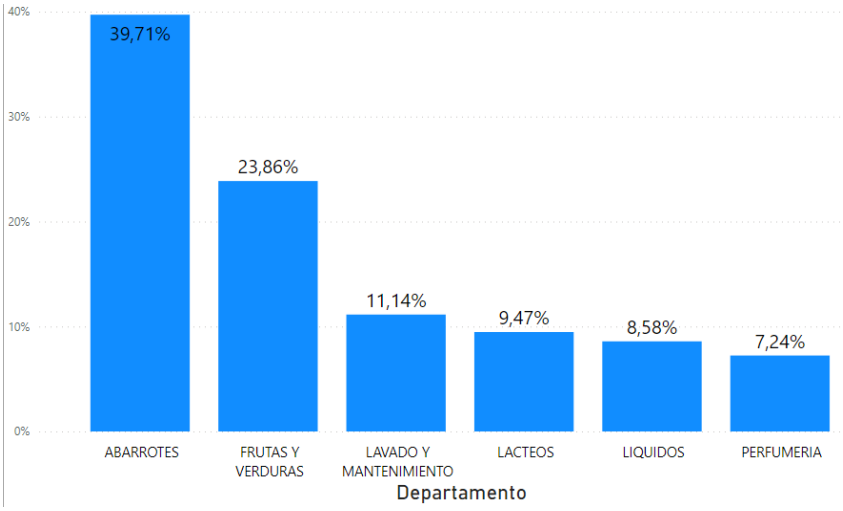
Como se puede visualizar en la Figura 49, en la base de datos general, la proporción de "Marca" adquirida por los clientes es de un 57% "Otros" aproximadamente, 41,67% "Tottus", 0,75% "Precio Uno", 0,14 % "Tottus Life", 0,12% "Tottus Orígenes", 0,08% "Tottus Premium", 0,07% "Tottus Bebe" y un % muy pequeño de "Tottus Kids". Por motivos de representatividad, las categorías a tomar en cuenta son "Otros" y "Tottus".

Asimismo, como se visualiza en la Figura 50, se analizó la misma variable por *cluster* y los resultados fueron los siguientes:

- El *cluster* "0" tiene una proporción de "Tottus" de un 0,67% y un 99,3% de "Otros".
- El *cluster* "1" tiene una proporción de "Tottus" de un 89% y un 11% de "Otros".
- El *cluster* "2" tiene una proporción de "Tottus" de un 5% y un 95% de "Otros".
- El *cluster* "3" tiene una proporción de "Tottus" de un 22% y un 78% de "Otros".
- El *cluster* "4" tiene una proporción de "Tottus" de un 97% y un 3% de "Otros".
- El *cluster* "5" tiene una proporción de "Tottus" de un 35% y un 65% de "Otros".
- El *cluster* "6" tiene una proporción de "Tottus" de un 49% y un 51% de "Otros".

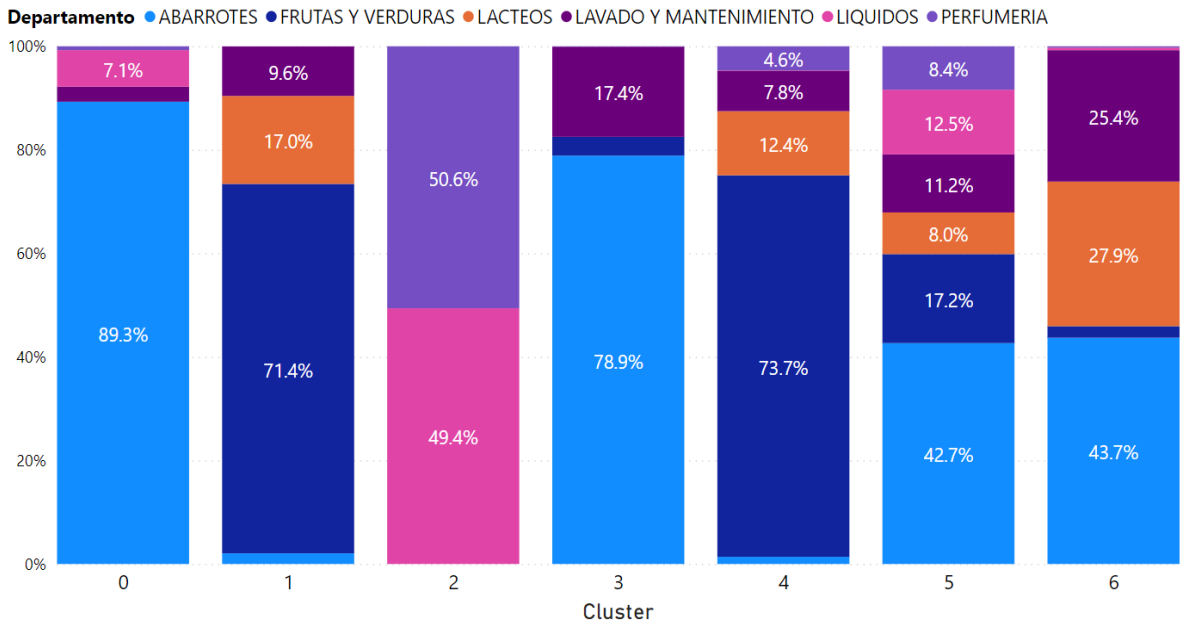
Finalmente, la última variable analizada fue el "Departamento", que como ya se definió en la presentación de variables, esta describe una subcategoría de productos luego de la variable "División". No se usó la variable "División" dentro del modelo, debido a que eran categorías muy amplias y no se usó la variable "Sub-Departamento", debido a que tenía muchas categorías específicas a analizar.

Figura 51: Proporción de ventas por “Departamento”



Nota. Elaboración propia

Figura 52: Proporción de “Departamento” por cluster



Nota. Elaboración propia

Como se puede visualizar en la Figura 51, en la base de datos general, la proporción de “Departamento” es de un 39,71% en “Abarrotes” aproximadamente, 23,86% “Frutas y Verduras”, 11,14% “Lavado y Mantenimiento”, 9,47 % “Lácteos”, 8,58% “Líquidos” y 7,24% “Perfumería”.

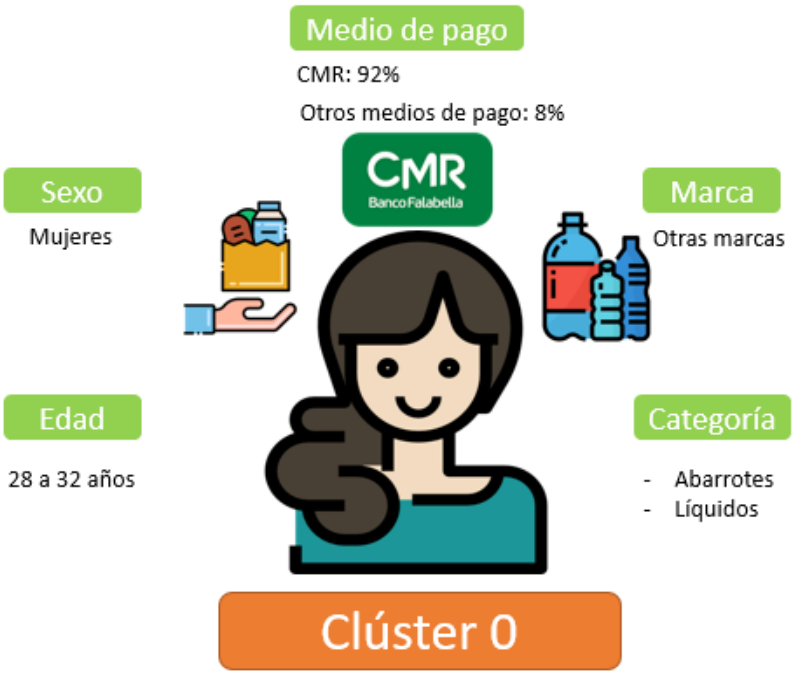
Asimismo, como se visualiza en la Figura 52, se analizó la misma variable por *cluster* y los resultados fueron los siguientes:

- El *cluster* "0" tiene una proporción de "Abarrotes" de un 89%. Se tiene otras tres categorías que no son representativas: 7,1% de "Líquidos", 2,89% de "Lavado y Mantenimiento" y 0,74% de "Perfumería" .
- El *cluster* "1" tiene una proporción de "Frutas y Verduras" de un 71,4% y un 17% de "Lácteos". Se tiene otras dos categorías que no son representativas: 9,6% de "Lavado y Mantenimiento" y 2,06% de "Abarrotes".
- El *cluster* "2" tiene una proporción de "Perfumería" de un 50,6% y un 49,4% de "Líquidos".
- El *cluster* "3" tiene una proporción de "Abarrotes" de un 78,9% y un 17,4% de "Lavado y Mantenimiento". Se tiene otra categoría que no es representativa: 3,63% de "Frutas y Verduras".
- El *cluster* "4" tiene una proporción de "Frutas y Verduras" de un 73,7% y un 12,4% de "Lácteos". Se tiene otras tres categorías que no son representativas: 7,8% de "Lavado y Mantenimiento", 4,6% de "Perfumería" y 0,07% de "Líquidos".
- El *cluster* "5" tiene una proporción de "Abarrotes" de un 42,7%, 17,2% de "Frutas y Verduras", 12,5% de "Líquidos", 11,2% de "Lavado y Mantenimiento", 8,38% de "Perfumería" y 8% de "Lácteos".
- El *cluster* "6" tiene una proporción de "Abarrotes" de un 43,7%, 27,9% de "Lácteos" y un 25,4% de "Lavado y Mantenimiento". Se tiene otras dos categorías que no son representativas: "Frutas y Verduras" y "Líquidos".

En base a los resultados presentados anteriormente, se propone la categorización de los siguientes *clusters*:

- *Cluster* "0": Este tipo de cliente, cuya edad está en el rango de 28 a 32 años, son mujeres consideradas como personas que prefieren comprar abarrotes y líquidos mediante el aplicativo *Fazil*. Asimismo, prefieren comprar otras marcas diferentes a la de Tottus haciendo uso de la tarjeta CMR en su mayoría, aunque existe un pequeño porcentaje que prefiere usar otro tipo de tarjetas (8%). Es por ello que se podría incentivar el uso de la tarjeta CMR para este grupo y además promocionar los productos de marcas propias, ya que esta categoría es predominante.

Figura 53: Descripción de Cluster "0"



Nota. Elaboración propia

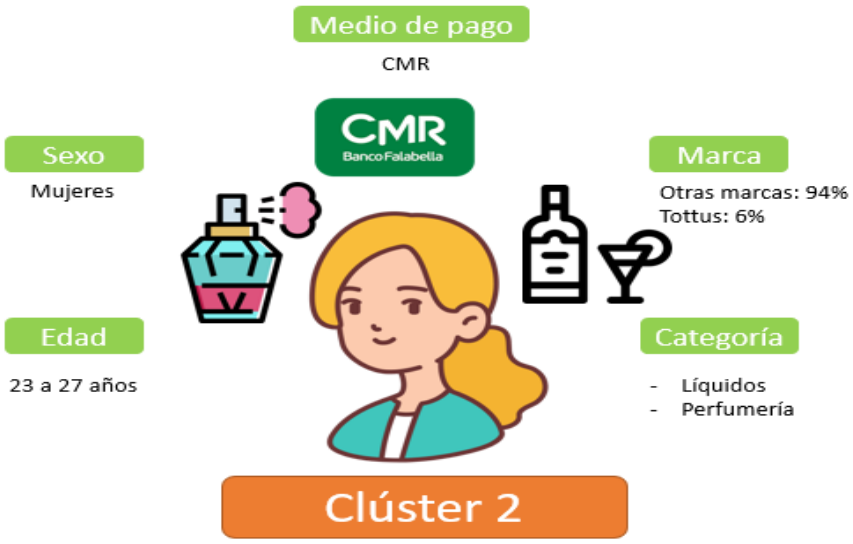
- *Cluster* "1": Estas clientas son mujeres en su mayoría, cuya edad está en el rango de 33 a 37 años, y son consideradas como personas que prefieren comprar frutas y verduras, lácteos y productos relacionados a la limpieza y cuidado del hogar. Estos productos son en su mayoría de marcas propias y el medio de pago que predomina en este grupo es otros medios de pago, en su mayoría, que no son la tarjeta CMR.

Figura 54: Descripción de Cluster "1"

Nota. Elaboración propia

- *Cluster* "2": Estas clientas son mujeres, cuya edad está en el rango de 23 a 27 años, y son consideradas como personas que prefieren comprar bebidas alcohólicas y productos relacionados a perfumería. Estos productos son en su mayoría de otras marcas y el medio de pago que predomina en este grupo es la tarjeta CMR.

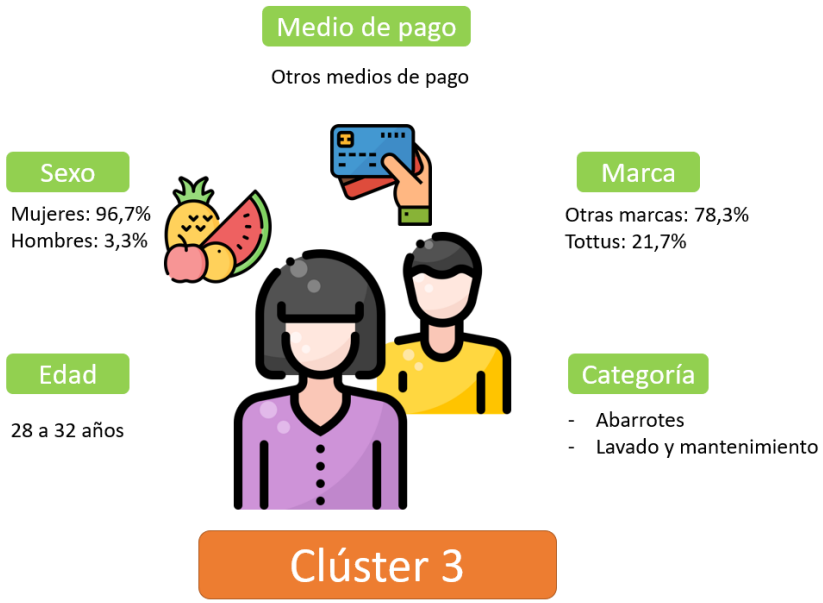
Figura 55: Descripción de Cluster "2"



Nota. Elaboración propia

- *Cluster* "3": Estas clientas son mujeres en su mayoría, cuya edad está en el rango de 28 a 32 años, y son consideradas como personas que prefieren comprar abarrotes y productos relacionados a la limpieza y cuidado del hogar. Estos productos son en su mayoría de otras marcas y el medio de pago que predomina en este grupo es diferente al de la tarjeta CMR.

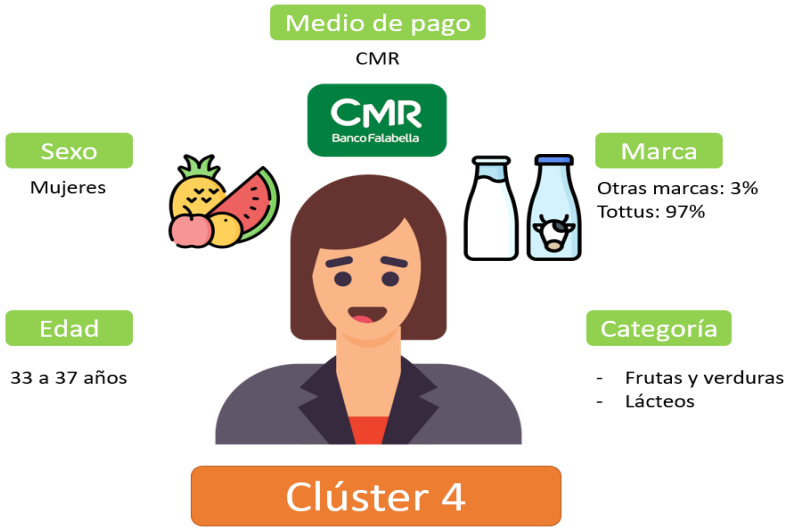
Figura 56: Descripción de Cluster "3"



Nota. Elaboración propia

- *Cluster* "4": Este tipo de clientes, cuya edad está en el rango de 33 a 37 años y que son en su mayoría mujeres, son consideradas como personas que prefieren una vida saludable comprando productos como frutas y verduras y también lácteos. Asimismo, la marca de productos adquirida en este grupo es en su mayoría pertenecientes a la marca Tottus y realizan compras con tarjeta CMR.

Figura 57: Descripción de Cluster "4"



Nota. Elaboración propia

- *Cluster* "5": Este tipo de clientes, cuya edad está en el rango de 28 a 32 años y que son hombres, son consideradas como personas que prefieren productos como "Abarrotes", "Frutas y verduras" y "Líquidos". Asimismo, este grupo de clientes utiliza, en su mayoría, la tarjeta CMR, pero también utiliza otros medios de pago. Respecto a la marca, compra productos de otras marcas, pero también marcas propias, por lo que hay una oportunidad importante para poder incentivar la compra de productos de marcas propias.

Figura 58: Descripción de Cluster "5"

Nota. Elaboración propia

- *Cluster* "6": Este tipo de clientes, cuya edad está en el rango de 28 a 32 años y que son hombres, son consideradas como personas que prefieren productos como "Abarrotes", "Frutas y verduras" y "Líquidos". Asimismo, este grupo de clientes utiliza, en su mayoría, la tarjeta CMR, pero también utiliza otros medios de pago. Respecto a la marca, compra productos de otras marcas, pero también marcas propias, por lo que hay una oportunidad importante para poder incentivar la compra de productos de marcas propias.

Figura 59: Descripción de Cluster "6"

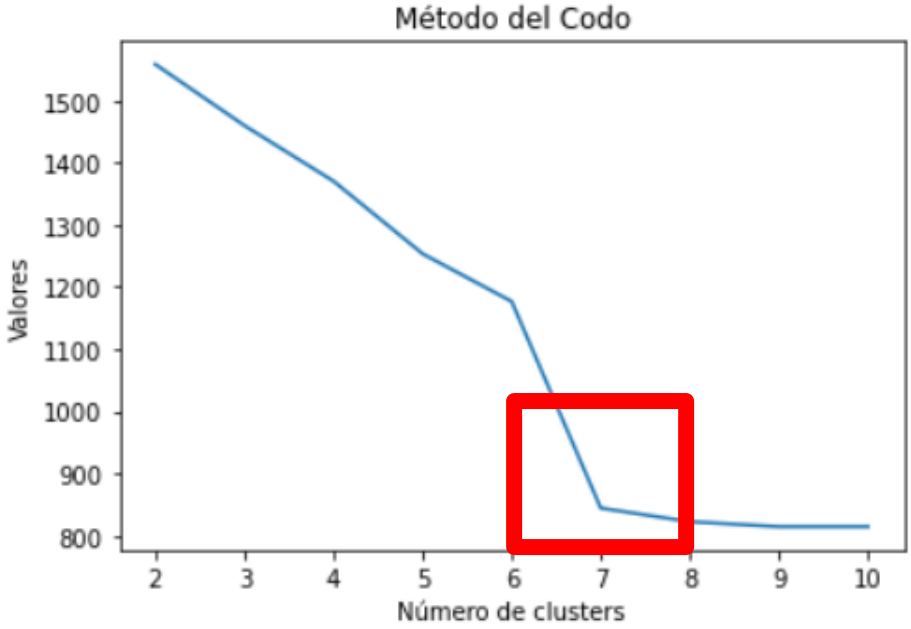
Nota. Elaboración propia

5.2 Medición de la solución

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo

En la evaluación del presente modelo, usando la técnica de *K-Medoids* (k = [2;11]), se tuvo la validación teórica del indicador "Inercia". Este indicador brinda al modelo construido, un puntaje basado en la cantidad de K utilizados. "Inercia" calcula la distancia de cada punto (dato) con el *cluster* que se obtuvo al final de realizada la evaluación de *K's*. Debido a esta premisa, se considera que si el modelo tiene una inercia menor es porque los datos son cercanos de cada *cluster*. Los resultados de la evaluación de Inercia se muestran en el siguiente gráfico:

Figura 60: Inertia por Cluster (K-Medoids)



Nota. Elaboración propia

Analizando los resultados obtenidos por cada k, se puede observar que empieza a horizontal la línea de tendencia y es en el *cluster* 7 en donde ocurre este cambio y es por ello, que se escogió este *cluster* como un K-óptimo, junto a la opinión de un experto que también validó los resultados obtenidos.

5.2.2 Simulación de solución

A continuación, se presenta una tabla resumen para realizar la comparativa de los *cluster* encontrados luego de realizar el análisis de las variables:

*Tabla 15:*Comparativo de cluster (K-Medoids)

		Cluster						
		0	1	2	3	4	5	6
Variable	Edad	28-32 años	33-37 años	23-27 años	28-32 años	33-37 años	28-32 años	33-37 años
	Sexo	Femenino	Femenino en su mayoría	Femenino	Femenino en su mayoría	Femenino en su mayoría	Masculino	Femenino
	Medio de pago	Credito: CMR en su mayoría	Visa: Otros medios de pago en su mayoría	Credito: CMR	Visa: Otros medios de pago en su mayoría	Credito: CMR	Credito: CMR en su mayoría	Credito: CMR en su mayoría
	Marca	Otras marcas en su mayoría	Tottus en su mayoría	Otras marcas en su mayoría	Otras marcas en su mayoría	Tottus en su mayoría	Otras marcas (64%) y Tottus (36%)	Otras marcas (51%) y Tottus (49%)
	Departamento	Abarrotes, Liquidos	Frutas y verduras, Lácteos, Lavado y Mantenimiento	Liquidos y Perfumería	Abarrotes, Lavado y mantenimiento	Frutas y verduras, Lacteos	Abarrotes, Frutas y verduras, Liquidos	Abarrotes, Lacteos, Lavado y mantenimiento

Nota. Elaboración propia

- *Cluster* "0": Son mujeres cuya edad está en el rango de 28 a 32 años consideradas como personas que prefieren comprar abarrotes y líquidos. Asimismo, prefieren comprar otras marcas diferentes a la de Tottus haciendo uso de la tarjeta CMR en su mayoría, aunque existe un pequeño porcentaje que prefiere usar otro tipo de tarjetas (8%).

- *Cluster* "1": Son mujeres en su mayoría, cuya edad está en el rango de 33 a 37 años, y prefieren comprar frutas y verduras, lácteos y productos relacionados a la limpieza y cuidado del hogar. Estos productos son en su mayoría de marcas propias y el medio de pago que predomina en este grupo es otros medios de pago, en su mayoría, que no son la tarjeta CMR.

- *Cluster* "2": Son mujeres cuya edad está en el rango de 23 a 27 años, y son consideradas como personas que prefieren comprar bebidas y productos relacionados a perfumería. Estos productos son en su mayoría de otras marcas y el medio de pago que predomina en este grupo es la tarjeta CMR.

- *Cluster* "3": Son mujeres en su mayoría, cuya edad está en el rango de 28 a 32 años, y son consideradas como personas que prefieren comprar abarrotes y productos relacionados a la limpieza y cuidado del hogar. Estos productos son en su mayoría de otras marcas y el medio de pago que predomina en este grupo es diferente al de la tarjeta CMR.

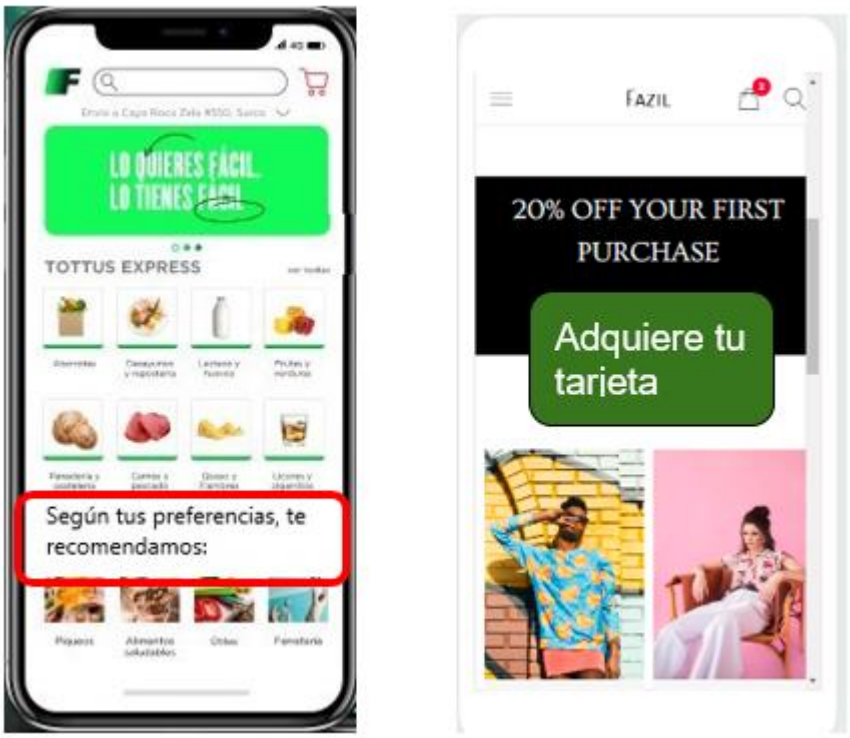
- *Cluster* "4": Este tipo de clientes, cuya edad está en el rango de 33 a 37 años y que son en su mayoría mujeres, son consideradas como personas que prefieren una vida saludable comprando productos como frutas y verduras y también lácteos. Asimismo, la marca de productos adquirida en este grupo es en su mayoría pertenecientes a la marca Tottus y realizan compras con tarjeta CMR.

- *Cluster* "5": Este tipo de clientes, cuya edad está en el rango de 28 a 32 años y que son hombres, son consideradas como personas que prefieren productos como "Abarrotes", "Frutas y verduras" y "Líquidos". Asimismo, este grupo de clientes utiliza, en su mayoría, la tarjeta CMR, pero también utiliza otros medios de pago. Respecto a la marca, compra productos de otras marcas, pero también marcas propias.

- *Cluster* "6": Este tipo de clientes, cuya edad está en el rango de 28 a 32 años y que son hombres, son consideradas como personas que prefieren productos como "Abarrotes", "Frutas y verduras" y "Líquidos". Asimismo, este grupo de clientes utiliza, en su mayoría, la tarjeta CMR, pero también utiliza otros medios de pago. Respecto a la marca, compra productos de otras marcas, pero también marcas propias.

Se puede observar que existen similitudes, pero también diferencias que pueden ser aprovechadas en cada *cluster* de cliente. A continuación, se muestra una plantilla de un ejemplo de sistema de recomendación en base a las preferencias de cada *cluster*:

Figura 61: Fazil: Sistema de recomendación



Nota. Elaboración propia

Capítulo VI: Conclusiones y Recomendaciones

6.1. Conclusiones

Actualmente, las empresas *retail* tienen un incremento en las ventas en sus canales del *e-commerce* por lo que deben afrontar el reto de mejorar la experiencia de los clientes y debido a que cuentan con una gran cantidad de datos ven la necesidad de usar *Machine Learning* usando sus herramientas para tener mejores resultados.

En este trabajo, se propone un modelo de segmentación de clientes mediante el uso de técnicas de *Machine Learning* de manera que se pueda generar un sistema de recomendación de productos para los clientes del canal *e-commerce* de Tottus partiendo del análisis de datos se toma en cuenta la cantidad de transacciones históricas y se evalúa tomar como muestra el mes de abril del 2022.

Para la elaboración de este estudio, se ha trabajado con una base de datos de 265,435 registros y 66 variables que representan las transacciones totales del mes de abril (mes de muestra), la cual se obtuvo directamente de las transacciones registradas en la data de ventas para el canal *e-commerce,* específicamente mediante su aplicativo móvil.

Como segunda etapa se preprocesaron lo datos, para ello se identificaron las variables que estaban asociadas a los datos de clientes, detalles de producto y otros datos que no aportaban en la explicación del comportamiento de compra por lo que se redujo de 66 a 5 variables, además se eliminaron lo valores nulos, duplicados y se normalizó los datos. Posteriormente se utilizó la técnica de PCA para reducir las dimensiones de los datos y hacerlo menos pasado para el modelado, lo que resultó en la reducción a 3 dimensiones que explican el 70% de los datos analizados.

Para el modelamiento se aplicó 3 técnicas las cuales nos permitieron segmentar a los clientes. En este trabajo se usó la técnica de *K-means, K-medoids* y el Agrupamiento Jerárquico las cuales fueron desarrolladas en el lenguaje de programación Python evaluando distintos escenarios.

Con la técnica de *K-means*, mediante el método del codo el mejor agrupamiento fue k óptimo en 6. Por otro lado, con la técnica de *K-medoids* y el método del codo, gráficamente resulta en siete *clusters* o grupos de clientes, es decir, un k optimo igual a 7. Estos valores óptimos de K fueron validados con la técnica del Agrupamiento Jerárquico donde en el gráfico del dendograma muestra que ambos métodos están dentro del rango del óptimo que explica los grupos. Posteriormente, se procedió a desarrollar las características de los segmentos según la edad, el medio de pago, sexo, marca y departamento de productos para los resultados de la técnica de *K-means y K-medoids.*

Finalmente, como se trata de técnicas de aprendizaje no supervisado fue necesario realizar una validación práctica por el experto del área de *Growth Marketing* de la empresa quien nos indicó que la mejor segmentación de clientes es la propuesta con la técnica *K-medoids* de donde se tiene el k óptimo de siete, esto debido a que los productos tienen mayor segmentación y diferencia entre los grupos.

6.2. Recomendaciones

Mediante la aplicación de las técnicas de *Machine Learning* desarrolladas en la investigación se logró crear una segmentación de clientes considerando sus preferencias de compra. Pero por el tiempo de desarrollo de este trabajo no se logró el despliegue del sistema de recomendación y demostrar cuantitativamente los beneficios para el negocio de esta segmentación.

Por lo tanto, desarrollar el despliegue de las estrategias de venta en el aplicativo de Tottus considerando las preferencias de los siete grupos propuestos mediante un sistema de recomendación de productos permitirá mejorar la experiencia de los clientes en un periodo de tiempo de prueba permitirá poder encontrar mejoras al modelo. Lograr realizar la medición del despliegue mediante encuestas de satisfacción permitirá validar si el modelo de recomendación generado generará una mejora en la experiencia de los clientes.

Además, según lo recomendado por el experto durante la entrevista, se podría añadir más variables que contemplen el perfil de compra de los clientes para que nos permitan tener mayor detalle de la compra como, por ejemplo: distrito de compra, distancia de la compra al local de despacho, horario de compra, etc.

Referencia Bibliografía

Amat, J. (2017). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y T-SNE.* Recuperado de : https://www.cienciadedatos.net/documentos/35_principal_component_analysis

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-means and K-medoids algorithms for big data. Procedia computer science, 78, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Aggarwal, C. C. (2018). Recommender Systems: The Textbook. Springer International Publishing.

Blacksip Digital Commerce Partners. (2021) Reporte de industria: El e-commerce en el Perú 2021 - 2022. Recuperado de: https://content.blacksip.com/reporte-del-e-commerce-en-peru-2021

Calad, F. (2015). "Segmentación de clientes automatizada a partir de minería de datos k-means clustering" . Escuela de Ingenieria de Antioquia. Recuperado de https://repository.eia.edu.co/bitstream/handle/11190/2288/CaladFelipe_2015_SegmentacionClientesAutomatizada.pdf?sequence=1&isAllowed=y

Cam, C. (2021). *"Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa".* Lima. Universidad de Lima. Recuperado de: https://revistas.ulima.edu.pe/index.php/Ingenieria_industrial/article/view/5808/5632

Cancino, C (2012). Matriz de análisis FODA cuantitativo. Recuperado de: https://christiancancino.cl/wp-content/uploads/2016/09/MATRIZ-DCS-FODA-CUANTITATIVA.pdf

Cisterna, C. *(2021). Segmentación de clientes activos de una entidad financiera empleando el algoritmo de K-means y árbol de decisión.* Universidad Nacional Mayor de San Marcos. Recuperado de: https://cybertesis.unmsm.edu.pe/handle/20.500.12672/17359

Gonzáles, H. y Ticona, U. (2019). Clustering, meditarraneidad y comercio internacional: aplicación empírica de los algoritmos Partioning Around Medoids y K-means. Revista Latinoamericana de Desarrollo Económico (32). Recuperado de : http://www.scielo.org.bo/scielo.php?pid=S2074-47062019000200005&script=sci_arttext

González, A. y Alba, F. (2017). *"Machine learning en la industria: El caso de la siderurgia".* Universidad Internacional de Rioja. Recuperado de: https://www.mincotur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/405/GONZALEZ%20MARCOS%20Y%20ALBA%20EL%C3%8DAS.pdf

Hipermercados Tottus. (s/f). Com.pe. Recuperado el 16 de julio de 2022, de https://www.greatplacetowork.com.pe/hipermercados-tottus

IBM (s.f). *¿Qué es Machine Learning?*. Recuperado de https://www.ibm.com/pe-es/analytics/machine-learning

Lee, Y., & Dubinsky, A. *(2017). Consumers desire to interact with a salesperson during eshopping: development of a scale. International Journal of Retail & Distribution Management, 20-39. Recuperado de : https://www.researchgate.net/publication/312155550_Consumers'_desire_to_interact_with_a_salesperson_during_e-shopping_development_of_a_scale*

Lucidez. (2018). *Así fue el crecimiento de Tottus en Perú*. Recuperado de https://lucidez.pe/asi-fue-el-crecimiento-de-tottus-en-peru/

Management Solutions (2018). "*Machine learning, una pieza clave en la transformación de los modelos".* España. Recuperado de: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Mining E. (2020).Machine Learning for beginners. A complete and phased beginners guide to learning and understanding Machine Learning and Artificial Intelligence. Recuperado de: https://es.3lib.net/book/5635209/091098

Morales, A., (2021). *El impacto de la inteligencia artificial en el Derecho. Advocatus,* 039, pp. 39-71. Recuperado de https://doi.org/10.26439/advocatus2021.n39.5117

Palacios, F y Pastor, N. (2020). *"Segmentación de clientes de una empresa comercializadora de productos de consumo masivo en la ciudad de Popayán soportado en Machine Learning y Análisis RFM (Recency, Frecuency y Money)"*. Fundación Universitaria de Popayán. Recuperado de: http://unividafup.edu.co/repositorio/files/original/58784efa51bf4609763d30e2e6f70bea.pdf

Reporte de Sostenibilidad. Tottus. Recuperado de http://q4live.s22.clientfiles.s3-website-us-east-1.amazonaws.com/351912490/files/doc_downloads/sustainability/2021/2021-Tottus.pdf

Reporte de Sostenibilidad. Tottus. Recuperado de https://downloads.ctfassets.net/dfhnfm93fvnr/3W728CcKVaRofWXWk8Utlf/b1edfeb8d949b1cf7da8d93f81f15443/RS_Tottus_22.pdf

Reporte de la Industria de E-commerce 2021- 2022 BlackSip. Recuperado de: https://www.peru-retail.com/el-52-de-los-peruanos-ya-compra-de-manera-online-segun-estudio/

Sandoval, L. (2018). *"Algoritmos de aprendizaje automático para análisis y predicción de datos".* El Salvador. Escuela Especializada en Ingeniería ITCA-FEPADE. Recuperado de: http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Statista Research Department (2022). América Latina: ventas de comercio electrónico 2021 - 2025, por país. Recuperado de: https://es.statista.com/estadisticas/1075464/america-latina-e-commerce-ventas/

Tan, P.-N., Steinbach, M., y Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison-Wesley.

Torres, P, Gonzáles, J., López, V. y Vaca, S. (2020). *Aprendizaje automático aplicado al análisis del consumo de alcohol y su relación con el estrés percibido. RISTI – Revista Ibérica de Sistemas y Tecnologías de Información, p. 483-395.*

Tottus. *¿Quiénes somos?*. Recuperado el 04 de agosto de 2022, de https://www.tottus.com.pe/quienes-somos

Tibshirani, R. (2013). Data Mining. Cmu.edu. Recuperado de https://www.stat.cmu.edu/~ryantibs/datamining/lectures/04-clus1.pdf

Tripathi, S., Bhardwaj, A. y Poovammal. (2018). "Approaches to clustering in Customer Segmentation". International journal of engineering & technology, 7(3.12), 802. https://doi.org/10.14419/ijet.v7i3.12.16505

Unioviedo (2020). El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. Recuperado de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

Verano, P. (30 de enero de 2019). El 77% prioriza descuentos online para elegir lugar de compra. Gestión. https://gestion.pe/economia/77-prioriza-descuentos-online-elegir-lugar-compra-257256-noticia/

Verano, P. (27 de febrero 2019). El 50% de compradores online en Perú se ve atraído por bajos precios. Gestión. https://gestion.pe/economia/50-compradores-on-line-peru-ve-atraido-bajos-precios-259849-noticia/

Wiskott, L. (2014). *Lecture Notes on Clustering.* Alemania. Institut f¨ur Neuroinformatik Ruhr-Universit¨at Bochum. Recuperado de: https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/Clustering-LectureNotesPublicVideoAnnotated.pdf

Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022). Research on segmenting E-commerce customer through an improved K-medoids clustering algorithm. Computational Intelligence and Neuroscience, 2022, 9930613. https://doi.org/10.1155/2022/9930613

Anexos

Anexo 1: Matriz FODA cuantitativo 1

		OPORTUNIDADES				AMENAZAS			
		O1	O2	O3	O4	A1	A2	A3	A4
FORTALEZAS	F1	5	5	7	2	1	1	1	4
	F2	4	3	7	2	7	1	1	5
	F3	1	7	4	2	1	1	4	1
	F4	7	3	5	1	1	1	1	1
DEBILIDADES	D1	1	1	1	1	5	5	1	1
	D2	7	7	1	1	1	1	1	1
	D3	1	1	1	1	4	1	1	1
	D4	1	1	1	1	3	1	1	1

Anexo 2: Matriz FODA cuantitativo 2

		OPORTUNIDADES				AMENAZAS			
		O1	O2	O3	O4	A1	A2	A3	A4
FORTALEZAS	F1	6	5	4	2	4	4	3	4
	F2	7	6	5	2	4	2	4	4
	F3	4	5	5	4	3	3	4	4
	F4	7	5	4	3	2	2	5	5
DEBILIDADES	D1	4	4	3	1	2	3	1	2
	D2	6	6	2	2	1	5	1	6
	D3	1	2	1	4	1	3	5	4
	D4	1	2	3	1	2	1	1	4

Anexo 3: Matriz FODA cuantitativo 3

		OPORTUNIDADES				AMENAZAS			
		O1	O2	O3	O4	A1	A2	A3	A4
FORTALEZAS	F1	6	7	5	3	5	1	3	5
	F2	5	5	7	3	1	1	1	4
	F3	4	7	5	6	1	1	3	1
	F4	6	4	5	4	1	1	1	1
DEBILIDADES	D1	3	1	1	1	3	4	1	3
	D2	3	5	1	1	1	5	3	5
	D3	1	1	1	4	1	5	1	2
	D4	1	1	4	1	1	2	3	1

Anexo 4: Matriz FODA cuantitativo 4

		OPORTUNIDADES				AMENAZAS			
		O1	O2	O3	O4	A1	A2	A3	A4
FORTALEZAS	F1	6	4	7	2	1	5	1	1
	F2	5	6	6	1	2	4	1	5
	F3	1	7	6	2	2	5	4	1
	F4	7	5	5	1	1	4	7	1
DEBILIDADES	D1	1	2	5	1	2	2	1	1
	D2	3	7	6	2	1	2	1	5
	D3	2	5	2	7	1	2	3	4
	D4	2	2	7	1	2	1	1	6

Anexo 5: Matriz FODA cuantitativo 5

		OPORTUNIDADES				AMENAZAS			
		O1	O2	O3	O4	A1	A2	A3	A4
FORTALEZAS	F1	7	3	6	1	3	2	2	7
	F2	7	5	5	2	4	1	1	7
	F3	2	7	1	3	3	6	5	2
	F4	7	3	3	1	2	3	5	1
DEBILIDADES	D1	7	3	6	1	1	2	1	1
	D2	2	7	1	1	1	5	2	6
	D3	1	2	2	3	1	2	6	4
	D4	3	1	7	1	1	1	2	2

Anexo 6: Entrevista con analista de Customer Acquisition de Tottus.

Duración:	17 minutos.
Fecha:	Miércoles 28 de setiembre 2022 a las 20:00 hrs
Plataforma virtual:	Google Meets

Entrevistador: Hola Srta. Camila, buenas noches. Actualmente estamos llevando un curso para titularnos para ello, en nuestra tesis, se está proponiendo una segmentación de clientes a través de *Machine Learning* y aplicación de algoritmos de cluster a fin de que la empresa Fazil pueda recomendar sus productos de acuerdo al comportamiento de los pedidos que han tenido últimamente sus clientes. Se han aplicado dos técnicas de segmentación: K-means y K-medoids, cada una con un resultado propio de la técnica, el primero da como resultado 6 cluster´s y el segundo 7, respectivamente.

Se han utilizado 5 variables: sexo, edad, medio de pago, marca y departamente que hace referencia a la categoría de productos. ¿Consideras que las variables utilizadas son representativas para segmentar a sus clientes?

Entrevistado: Si, de hecho en mi área se utilizan modelos para segmentar a los clientes y muchos de estos se alimentan de cómo estos clientes se comportan. La edad, el sexo y tipo medio de pago son variables ustilizadas para segmentarlos. Tengo una duda: ¿en la variable marcas solo se está considerando la marca tottus?

Entrevistador: No, solo es una etiqueta que permite caracterizar a todas las marcas propias de la empresa Tottus y marcas como Gloria, Nestlé, entre otras, que también son vendidas a través del aplicativo Fazil, están nombradas como otras marcas. La idea es identificar marcas propias de la empresa Tottus y ver la forma de recomendarlas a sus clientes e impulsarlas.

Entrevistado: Entiendo, está perfecto que para resumir se separe entre marcas de Tottus y otras marcas (que no son propias de tottus), de hecho el mayor porcentaje de pedidos vienen de Tottus.

Entrevistador: Correcto, pasando a explicarte la propuesta 1 (K-means, 6 *clusters*).

En cuanto a la variable edad, se concluye que clientes entre 18-27 años y 37-43 años son los que más destacan, la mayoría de estos pertenecen al sexo femenino y en casi todos los cluster´s utilizan la tarjeta CRM (en 4 de 6 cluster´s).

Por otro lado, de acuerdo a la muestra de pedidos proporcionada, solo en 1 *cluster* predomina la compra de marcas Tottus y en los otros, otras marcas. Por último, en cuanto al departamento (categoría de producto) se tiene que en 3 cluster´s se destacan la compra de abarrotes, lavado y mantenimiento que están en rangos de edades superior a los 28 años, además, es importante destacar que un *cluster* se compran más líquidos y perfumería siendo clientes en el rango de edad 18-27 años quienes son los más jóvenes.

Respecto a la propuesta 2 (*K-medoids*, 7 clusters).

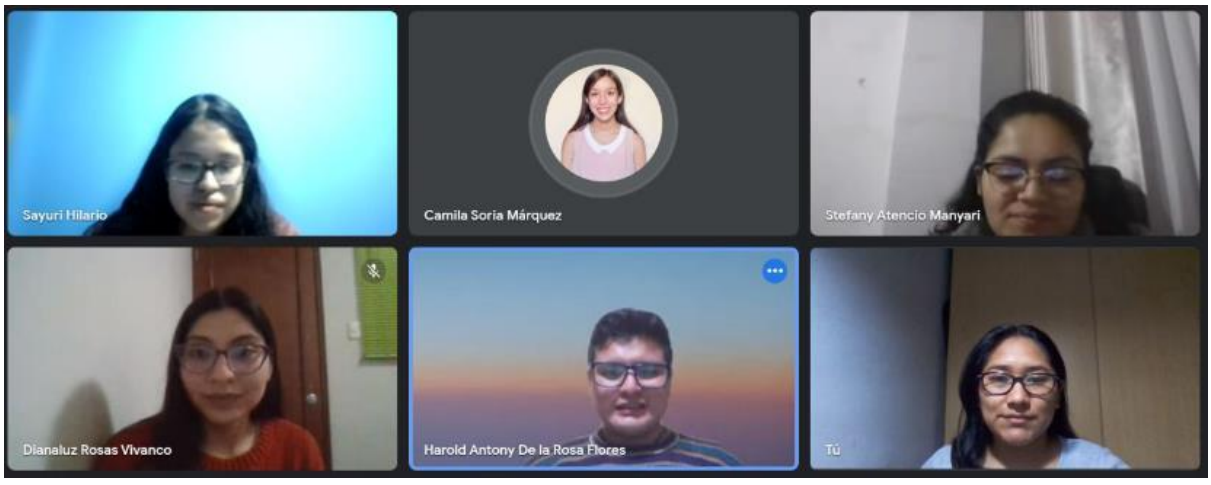
Las diferencias versus el anterior se centran en: los rangos de edades están más delimitadas, en cuanto a la marca se incremente el número de *clusters* donde predominan las marcas Tottus y por último en cuanto al departamento se suman a la compra de abarrotes, lavado y mantenimiento las categorías frutas, lácteos y verduras.

Entrevistado: Considero que la propuesta 2 es la más adecuada ya que se tiene un mayor rango de categorías de productos y así se puede armar estrategias de comunicación para impulsar estas categorías, además, me parece muy interesante en general ambas propuestas y las variables hacen sentido, así como los agrupamientos obtenidos en cada modelo. Son muy útiles para enviar las ofertas de ciertas categorías que se tienen en sus propuestas, considerando las variables que se han incluido.

Entrevistador: De acuerdo. Muchas gracias por tus comentarios, recomendaciones y por tu tiempo, nos ayudaron mucho para poder concluir con los análisis de los modelos.

Entrevistado: De nada, gracias a ustedes por mostrarme estos resultados que serán muy útiles para nuestra estrategia de comunicación.

Foto de Reunión:



Anexo 7: Perfil del experto de Tottus