UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL

Segmentación de clientes para mejorar la experiencia de compra de productos electrónicos en Falabella

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para obtener el título profesional de Ingeniero(a) Industrial y Comercial

AUTORES

Aragón Gallegos, Angela Del Carmen

Cerquin Silva Sabina Isabel

Escurra Yactayo, Renzo Omar

Roncalla Viena, Andrea Liliana

ASESOR

Fabian Arteaga, Junior John

ORCID Nº 0000-0001-9804-7795

Marzo, 2023

Grupo 06 - Presentación TSP

ORIGINALITY REPORT

22%	20%	2%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Student Paper	5%
2	dspace.unl.edu.ec Internet Source	3%
3	repositorio.esan.edu.pe Internet Source	2%
4	hdl.handle.net Internet Source	2%
5	www.researchgate.net Internet Source	1%
6	rstudio-pubs-static.s3.amazonaws.com Internet Source	1%
7	univida fup.edu.co Internet Source	1%
8	artyco.com Internet Source	1%
9	inaoe.repositorioinstitucional.mx Internet Source	1%

RESUMEN

En la presente investigación se pretende encontrar perfiles de consumidores de la empresa Saga Falabella y para esto analizamos las ventas del sector electro de la empresa entre los meses de noviembre del 2022 y enero del 2023, tomando en cuenta campos como el género de los consumidores, marcas de preferencia, categoría de equipos, métodos de pago y unidades vendidas, así como también si las compras fueron efectuadas por internet o en los diferentes locales que esta empresa posee a nivel nacional. Mediante la aplicación de métodos de aprendizaje no supervisado como: clustering jerárquico, K-Means y K-Medoids, se limpió, normalizó y procesó la data, de esta forma se consiguió obtener segmentos de consumidores bien definidos. Se obtuvieron cinco grupos de clientes con diferentes características y preferencias, esto ayudaría a Saga Falabella a enfocar mejor sus estrategias de marketing y de retención de clientes, favoreciendo el aumento de sus ventas y la preferencia de los consumidores por encima de otras empresas del mismo rubro.

Palabras Clave: Clustering, Machine Learning, Cluster Jerárquico, K-Medoids, K-Means.

ABSTRACT

In the present investigation it is intended to find profiles of consumers of the company Saga Falabella and for this we analyze the sales of the electro sector of the company between the months of November 2022 and January 2023, taking into account fields such as the gender of consumers, preferred brands, category of equipment, payment methods and units sold, as well as whether the purchases were made online or in the different locations that this company owns nationwide. Through the application of unsupervised learning techniques such as: hierarchical clustering, k means and k methods, the data was cleaned, normalized and processed, in this way it was possible to obtain well-defined consumer segments. Five groups of customers with different characteristics and preferences were obtained, this would help Saga Falabella to better focus its marketing and customer retention strategies, in order to increase sales and consumer preference over other companies in the same industry.

Keywords: Clustering, Machine Learning, Hierarchical Clustering, K-Medoids, K-Means.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Introducción

Debido al crecimiento del sector retail en estos últimos años, las organizaciones se han tenido que adaptar a los constantes cambios y a la alta demanda que generan los clientes. Después de los sucesos de la pandemia por el covid-19, muchos clientes han optado por realizar compras por web; actualmente el mercado de tiendas físicas también se ha visto potenciado por las reaperturas y reactivación de los negocios, por lo que las organizaciones buscan orientar sus objetivos en mejorar los procesos de venta en sus distintos canales. En el portal web Thefoodtech (2022) se comenta del entorno del sector retail posterior a la pandemia "Se encontraron algunos aspectos clave que conducirán a una industria más enfocada en la comodidad del cliente basado en la omnicanalidad, la personalización de las ofertas de acuerdo a cada comprador y la sustentabilidad."

La agrupación y segmentación de los consumidores o clientes, así como también el pronóstico de venta, se consideran primordiales para la planificación empresarial. Hoy en día, las organizaciones, principalmente privadas, realizan esfuerzos notables para desarrollar modelos de predicción mediante de métodos de machine learning, con la finalidad de aumentar la efectividad de las ventas, mejorar el proceso de compra del cliente, evaluar características particulares de los perfiles de consumidores, entre otros beneficios resultantes de la aplicación de técnicas de Inteligencia Artificial (IA).

El presente estudio propone analizar las ventas realizadas a través de los distintos canales de comercialización de productos electrónicos de la empresa Saga Falabella, mediante la aplicación de tres modelos de algoritmos de las técnicas de machine learning de aprendizaje no supervisado como el K-Medoids, K-Means y Cluster Jerárquico, para realizar la comparación e identificar el método óptimo para la agrupación de clientes de acuerdo a los atributos seleccionados. En base a esta información resultante, se espera poder tener la segmentación adecuada según su perfil de consumo, para lograr mejora en la experiencia de compra a través de desarrollo de campañas, publicidad y promociones efectivas de acuerdo al comportamiento de compra de cada segmento.

Con lo descrito anteriormente, el presente estudio se basará en desarrollar seis capítulos. Empezando desde el planteamiento de la realidad problemática, las justificaciones y delimitaciones que posee la investigación. Para la segunda parte, se desarrolla las bases

conceptuales y el marco teórico, que será de mucha importancia para entender los fundamentos de los métodos de estudio. En el tercer capítulo se comenta sobre el entorno del retail Saga Falabella, su estructura, modelo de negocio y su mapa de procesos. Para el capítulo cuatro se comenta las etapas metodológicas, el diseño de la investigación, cronograma de las actividades realizados y el presupuesto que se ha necesitado para realizar este estudio. Luego, en el capítulo cinco se analizan los resultados de la investigación, se evalúan los cluster obtenidos y se generan los perfiles del consumidor. Para terminar, el capítulo seis explica las conclusiones obtenidas y se determinan las recomendaciones que hemos recopilado durante la ejecución de este estudio.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la Realidad Problemática

En la actualidad, el entorno sector retail está en constante cambios y crecimiento, esto va de la mano con las variaciones en la demanda, cambios en las preferencias y requerimiento del cliente, por lo que las empresas buscan adaptarse a estas fluctuaciones.

Estos cambios se han visto reflejados a nivel mundial, en donde el sector E-commerce del sector retail ha experimentado un crecimiento continuo, el portal Stackscale (2022) ha recopilado información estadística durante el 2022 donde se determina que en promedio el 85% de los clientes a nivel mundial realizaron compras online. También se resalta a tres países que son los que generan mayores ventas en el mercado E-commerce, estos son China (+ 2,400 mil millones de euros), Estados Unidos (+ 770 mil millones de euros) y Reino Unido (+ 150 mil millones de euros).

Región	Porcentaje de usuarios que compraron online en 2020
Asia	86 %
Australia	79 %
Europa	83 %
Norteamérica	78 %
Sudamérica	86 %
Otras regiones	85 %

Figura 1: *Promedio de clientes que realizan compras en E-commerce.*
Fuente: *Stacksale (2022)*

Según la información recopilada por KPMG (2022), se hace un comparativo de crecimiento del comercio en línea durante el periodo de 2021 entre américa latina y el resto de los continentes, en donde se destaca lo siguiente:

El comercio en línea de Latinoamérica fue una de las que más creció durante este periodo, logrando un crecimiento alrededor del 37%, principalmente en los países de Argentina (79%), Brasil (35%) y México (27%), seguida por Norteamérica (32%), Europa (29%) y Asia Pacífico (26%).

Para el portal Kantar (2022) hace énfasis en el sector retail en Latinoamérica sobre el crecimiento que se tuvo frente al 2021 mencionando lo siguiente:

El gasto en bienes de consumo de rápido movimiento de los hogares de América Latina aumentó en +26% (2021 vs 2019). El ritmo de incremento de los canales muestra matices: nuevos se destacan, como el repinte y desarrollo de las farmacias y el grupo de tiendas especialistas".

Según el informe de BBVA Research (2019):

Las ventas en el sector retail en el 2020 incrementarán, a un ritmo equivalente a la de este año, consistente con el escenario base de proyecciones macroeconómicas. La mayor capacidad adquisitiva de los hogares, expansión de la clase media y el bono demográfico apuntan a que las ventas del sector retail tienen oportunidad para seguir creciendo. (p.2)

Figura 2: *Crecimiento estimado del sector Retail durante el 2020*

Fuente: *BBVA Research (2019)*

Si bien se proyectaba un mayor crecimiento durante el 2020 surgieron grandes factores que pusieron en riesgo el entorno del mercado, esto fue la aparición del virus COVID-19, como medida preventiva varios países entraron en pandemia y por ende muchas empresas tuvieron que cerrar o limitar el acceso a sus cadenas de tiendas para ajustarse a las medidas de seguridad impuestas por cada país.

Esto presenta grandes retos para poder adaptarse a estos cambios y mantener la competitividad en el mercado cumpliendo con las expectativas del cliente. Según Chicoma Daniel (2020) en el portal web Conexión ESAN menciona:

> Después de una restricción de salida, se migró a compra en mercados y supermercados. Estos últimos promocionaron las ventas por sus canales online, sin embargo, no tenían proyectada la enorme demanda, lo cual hizo notar que el sistema de distribución no era consistente y no podría abastecer lo solicitado. (p.3)

Debido al entorno varias empresas se tuvieron que adaptar, según lo mencionado en el medio digital Revista Gerencia PWC (2020), “El impacto que está generando en la adopción de canales online, la aceleración de las estrategias omnicanal, cambios en el comportamiento de consumo y la forma en que vino a agilizar el proceso de transformación de la industria” (p.5)

También se comenta en E-commerce en Perú (2021) que, “El boom del comercio electrónico se generó recién en junio, con un crecimiento del 86%, llegando a su nivel histórico en julio con 160%. Lo cual indica que conforme se ha ido eliminando las barreras en el E-commerce, esta industria ha venido creciendo”. (p.12)

	Pre Cuarentena (Enero 2020)	Durante (Julio 2020)	Total Cierre 2020
Penetración del Ecommerce en el consumo a través de tarjeta	12.5%	45%	35%
Crecimiento del ecommerce (YTY)	43%	160%	50%
Compradores Online	6 millones	8.9 millones	11.8 millones
Ticket promedio	S/171	S/231	S/141
Penetración del ecommerce sobre el total del comercio	1.5%	3.5%	5%
N° de negocios que venden online	65,800	131,600	263,200
Penetración Ecommerce sobre el retail	2.8%	6%	8%

Figura 3: *Cifras del E-commerce en el 2022.*
Fuente: *BBVA Research (2019)*

Con todos sucesos durante pandemia, los hábitos del consumidor pudieron haber cambiado, esto implicaría saber si desea comprar el producto de manera en línea por las plataformas web, si prefiere ver el producto y sentir la experiencia de compra en los canales directos en tiendas, conocer cómo han cambiado las tendencias de compra dentro de cada canal.

Estas dudas surgen de manera similar al analizar el mercado de Falabella Retail, desde que inició pandemia las cadenas de Falabella tuvieron que cerrar durante meses por medidas preventivas de contagio, en este periodo la empresa se adaptó a los cambios y preferencias de los clientes, fortaleció su plataforma web y mejoró la experiencia de compra virtual. Ya cuando se reabrieron las tiendas, los hábitos de los consumidores habían cambiado, desde la preferencia del método de pago hasta el tipo de compra que se realiza, si era física u online. Este trabajo de investigación pretende segmentar las preferencias del consumidor para así fortalecer las estrategias de venta y de marketing en base a estos resultados.

1.2 Justificación de la Investigación

1.2.1 Justificación Teórica

La investigación emplea herramientas de machine learning a través del aprendizaje no supervisado con el propósito de encontrar grupos o segmentos de los clientes de Saga Falabella, según patrones de compra y perfiles del consumidor, en base a esta información poder establecer o fortalecer nuevas estrategias comerciales según perfil de compra. Consideramos importante establecer estos patrones de compra posterior a la pandemia, pues el consumidor ha cambiado el hábito de compra.

La técnica del aprendizaje no supervisado consiste en que, dentro del conjunto de datos a analizar, no se conoce la variable que se busca predecir con la aplicación de la técnica. El Clustering y Asociación son las principales ramas del aprendizaje no supervisado. La técnica por emplear en el presente trabajo es la de clustering, la cual se basa en agrupar objetos o personas, en segmentos o conjuntos en los cuales los miembros tengan características comunes entre sí mismas, logrando diferenciarse del resto.

1.2.2 Justificación Práctica

Actualmente, el E-commerce es una gran ventana de oportunidades para las diferentes compañías de aumentar sus ventas, fidelizar clientes y disminuir costos. La mayoría de los consumidores ya se encuentran acostumbrados a comprar por internet porque es una forma cómoda de realizar compras, además, ya lo consideran como un medio seguro para realizar sus pagos.

También es un desafío para estas, las compras por internet generan una gran cantidad de transacciones que se convierte en una gran base de datos , de la cual se puede obtener información valiosa con las correctas herramientas de procesamiento y mediante esta las empresas puedan establecer estrategias que les permitan mejorar la experiencia al consumidor por ejemplo, realizar sugerencias personalizadas al consumidor de acuerdo a sus patrones de compra, de esta manera este reduce sus tiempos de compra.

De acuerdo a lo explicado previamente, el presente estudio tiene como objetivo brindar a las empresas una herramienta para el procesamiento de esta gran base de datos que obtienen por el comercio de sus productos electrónicos en sus diversos canales de venta. En este caso, se brinda una propuesta para la segmentación de clientes utilizando machine learning, mediante este análisis podrán clasificar sus clientes por grupos de preferencias y partir de las características de cada grupo proponer estrategias.

1.2.3. Justificación Metodológica

Esta investigación se realiza con la finalidad de detectar oportunidades de incrementar e incentivar la experiencia de compra de productos electrónicos de los consumidores de Saga Falabella utilizando técnicas de machine learning para clasificación según perfil de compra. El presente estudio se enfoca en la implementación de un modelo basándonos en técnicas de machine learning para la agrupación de clientes de aparatos tecnológicos para lo cual se

emplearán métodos de aprendizaje no supervisado como el K-Means, K-Medoids y Clúster Jerárquico.

Se recolectará la base de datos requerida, posteriormente se limpiará la data y se realizará el preprocesamiento de la data para que los resultados tengan el menor porcentaje de sesgos, finalmente se desarrollará el modelo de clasificación y se presentará el resultado de los métodos evaluados a fin de obtener el perfil de compra del cliente y ver alternativas de mejora que se ajusten a la necesidad de la empresa.

1.3 Delimitación de la Investigación

1.3.1. Delimitación espacial

La investigación considera todas las compras solo de productos electrónicos realizadas en Perú en el Retail Saga Falabella, de las cuales se seleccionan las variables más relevantes o determinantes para la segmentación a realizar.

1.3.2. Delimitación temporal

La información considerada para realizar el trabajo de investigación propuesto será enmarcada dentro del periodo de noviembre 2022 a enero 2023 considerando las compras de equipos electrónicos realizados a la empresa Saga Falabella, el cual será obtenido a través de fuentes primarias de la compañía, lo que permitirá realizar una segmentación de clientes según preferencia de compra a fin de orientar las estrategias de venta y marketing personalizadas.

1.3.3 Delimitación conceptual

Esta investigación considera técnicas de machine learning bajo el lenguaje no supervisado, específicamente la técnica de clustering debido a que las variables de estudio no tienen una correlación directa de causa-efecto o dependencia entre variables, por lo que se espera que a partir de la información analizada se determine grupos de consumidores según el patrón de compra y características específicas del cliente.

Se consideran tres técnicas a evaluar, K-Means, K-Medoids y Clúster jerárquico para encontrar grupos según la información proporcionada por el sistema, asimismo se utilizan métodos específicos como el “método del codo” para definir el número de grupos representados por las variables en estudio dentro del modelo K-Means.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes de la investigación

Palacios,F. & Pastor, N. "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, Frequency y Money)"

- **Problema**

El actual mundo dinámico de los negocios obliga a las empresas a identificar grupos o segmentos de clientes en base a ciertos criterios: características, comportamientos y necesidades para así brindar soluciones orientadas a ellas y poder sobresalir de la competencia.

Esto genera la necesidad de aplicar diferentes métodos para que las empresas puedan acceder a la información de forma precisa y clara y de esta forma lograr una toma de decisiones más rápida en base a los segmentos encontrados; por ejemplo, realizar campañas de fidelización y retención más efectivas y con objetivos claros.

La empresa comercializadora de lácteos de consumo masivo del estudio genera una gran base de datos, sin embargo, esta no es procesada debido a su dimensionalidad para convertirla en información útil. Esta empresa registra la información de sus clientes y sus ventas en excel, en base a esta realizan pronósticos de inventario, compras, etc; sin embargo, no usan estan información para estrategias de marketing, lo que le genera una desventaja en el mercado. Además, hay poco conocimiento del personal en tecnologías para el procesamiento de la información.

- **Base de Datos**

La fuente de datos usada para este artículo son 49 archivos de Excel del 2019 con 85538 registros, 112 variables y de 2837 clientes.

- **Metodología**

Modelo RFM

Al tener los 49 archivos de Excel separados usaron una macro en excel para poder unirlos. Aplicaron el modelo RFM que se basa en tres variables, las que escogieron fueron las siguientes: Fecha, monto_neto y codigo_cliente.

Después, buscaron los posibles registros vacíos o nulos, esto es primordial para que se garantice la no existencia de un sesgo de información en los datos y asegurarse que no se afecten los resultados del modelo que se está desarrollando. Luego de limpiar la data, procedieron a ordenarla con filtros avanzados para poder obtener montos totales por cliente y la última fecha de compra. En base a esto, calcularon la recencia; es decir, los días que han transcurrido desde la última compra de los clientes, luego, calcularon la cantidad de compras que realizaron los clientes en un periodo de tiempo establecido (# de transacciones total por cliente).

En base a todos los datos calculados anteriormente, procedieron a crear la matriz RFM logrando así obtener los rangos de las variables. Primero, antes de calcular los rangos de Recencia, Frecuencia y Monto, calcularon los valores mínimos, máximos, diferencia entre el valor y mínimo y máximo, número de segmentos de clientes para clasificación RFM y amplitud. Posteriormente, con todos estos datos calcularon los rangos y armaron la matriz.

Finalmente, después de obtenida la matriz, se calculan los puntajes del modelo RFM mediante el ponderado de los rangos para Recencia, Frecuencia y Monto, se le asigna un peso ponderado a cada una de las variables (monto , frecuencia y recencia) y se multiplica dicho valor con los rangos obtenidos anteriormente. Se obtienen como resultados 5 segmentos: Clientes VIP, excelentes, buenos, regulares y de poco aporte.

Clustering

El proceso fue llevado a cabo mediante el lenguaje de programación de Python. Después de revisar literatura referente al tema, escogen el modelo K-means para desarrollar el trabajo, ya que este modelo posee una grande capacidad para procesar volúmenes de datos ocupando poco espacio.

El procesamiento de los datos se realizó en la fase del modelo RFM, por ello ya no sería necesario volver a realizarlo para implementar clustering. Entonces, trabajaremos con 2837 registros del 2019, 5 variables (clientes, monto, frecuencia, recencia y fecha), 2837 números de clientes.

En python, primero se carga la data haciendo uso de las librería pandas, mediante la función “datos.describe()” se arma la tabla con los parámetros estadísticos de las variables: desviación estándar, la distribución de los datos en cuartiles y valores máximos y mínimos

Hicieron uso de la gráfica “Mapa de calor” para obtener el grado de correlación que mide la relación lineal entre variables cuantitativas. Estos valores nos permiten notar las variables con un alto grado de correlación entre sí, aportando información relevante a nuestro modelo

Mediante este mapa de calor se concluyó que hay una correlación negativa parcial entre la recencia y la frecuencia. Por otro lado, en el caso de la recencia y el monto la correlación es moderada y, por último, entre la frecuencia y el monto la correlación es casi nula.

Posterior a ello, normalizaron los datos y se empezó con el análisis para escoger el número óptimo de clústeres, para ello se usó el análisis de error de inercia, la gráfica que arroja es similar a un brazo y su codo, en donde este último representa un cambio de forma, obteniendo la siguiente gráfica.

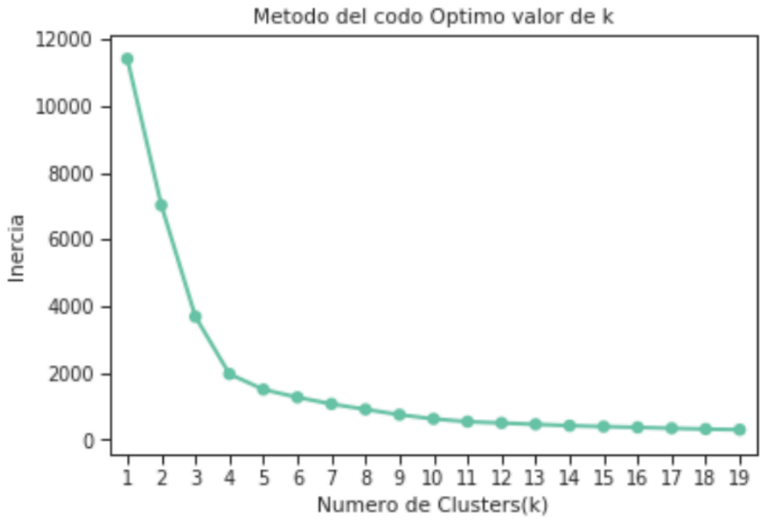

Figura 4: *Evaluación de la Inercia vs el número de clúster*
Fuente: *Palacios,F. & Pastor, N. (2020)*

Entonces, en base a esto, se desarrollará el modelo con un k = 4 clústeres. Elaboraron una gráfica en la que se observaba la dispersión y los clúster concretamente.

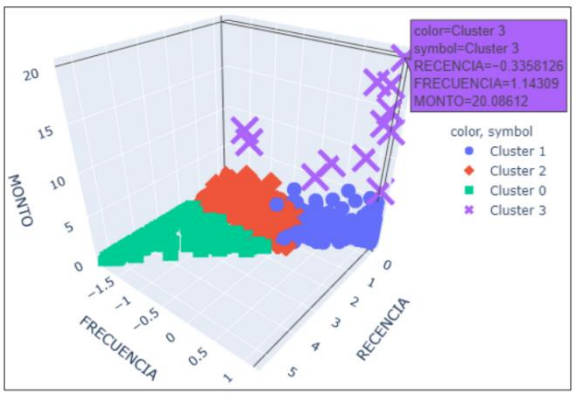

Figura 5: *Asignación de Clustering*
Fuente: *Palacios,F. & Pastor, N. (2020)*

Una vez que los realizaron los dos modelos se van a validar mediante validación interna: Coeficiente de Silueta, estudia la separación entre los segmentos. El valor de 0 indica que la muestra está lejos del demás clúster y cuando se obtiene un valor negativo refiere que la segmentación podría estar incorrecta.

Comparan los gráficos de silueta desde k=2 obteniéndose el mejor gráfico y con mayor coeficiente.

- **Resultados**

Mediante el análisis RFM se segmenta a los clientes en 5 grupos: Clientes Vip, clientes excelentes, clientes buenos, clientes regulares y clientes con poco aporte. En el siguiente gráfico se muestran los resultados completos:

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

Figura 6: *Resultados del análisis RFM*
Fuente: *Palacios,F. & Pastor, N. (2020)*

Con los resultados, concluyeron que el segmento o grupo de clientes de poco aporte de la empresa representa tan solo el 3% de las ventas totales, el segmento vip de la empresa al 42%. Además, identificaron que 117 clientes compraron y no volvieron por lo que indica que serían clientes perdidos y disminuye la tasa de retención de clientes de la empresa.

Mediante el modelo de clustering se obtuvieron 4 clúster: En el clúster 0, están los clientes de poco aporte, compraron hace mucho; el clúster 1, representa el 61% de los clientes y se les atribuye el 77% de los ingresos y son los denominados clientes buenos; clúster 2, conformado por 12 clientes y son los "clientes VIP" y por último, clúster 3 está conformado por 921 clientes y tienen el título de "clientes regulares.

Valdés , D., Reyes , R., Jaime, A,. Figueroa, E., Suárez , J. (2021) " Método de Clustering Jerárquico para la asignación del financiamiento a Proyectos de Desarrollo Local"

- **Problema:**

En Cuba, se considera al desarrollo local como un proceso direccionado al desarrollo económico y social ya que este eleva el estatus de vida de la población y favorece la explotación del potencial de las localidades.

De acuerdo con Valdés, D. (2021), el desarrollo local presenta las siguientes etapas: Análisis Estratégico, Proyección a programa, Financiamiento de Proyectos y Evaluación y Monitoreo.

La investigación realiza mayor foco en la fase de Financiamiento de Proyectos, ya que mediante esta los gobiernos municipales podrían desarrollar sus programas. Sin embargo, es un desafío para las entidades gubernamentales de cada territorio la administración eficiente del financiamiento y los recursos con los que cuenta para la ejecución de sus proyectos. En base a esto, los autores plantean como objetivo la implementación de técnicas de decisión multicriterio para contribuir a facilitar la elección de proyectos que se financiarán y así evitar gastos que no son necesarios.

- **Metodología**

En esta investigación los autores proponen una metodología de 8 pasos, lo cuales se detallan a continuación:

1. Definición de los Programas por sectores

En el estudio definen los siguientes programas:

- Sector Agropecuario

-Sector Servicios

- Sector Industrial

- Desarrollo Social

- Dimensión Natural

Cabe resaltar que cada uno de estos tiene como representante a una entidad estatal o no estatal, cada una de ellas tiene la responsabilidad de presentar proyectos para satisfacer el programa que tiene asignado. Asimismo, cada departamento tiene una "cartera de oportunidades" la cual está a cargo de organizar y reservar los proyectos, lo cuales son representados por las diferentes entidades mencionadas.

2. Definición de criterios de comparación

En enlistado de criterios será definido por los expertos y ellos mismos evaluarán cada proyecto en base a estos criterios, de tal manera que así se puedan comparar entre ellos. Además, la lista de criterios sirve para convertir la evaluación cualitativa en cuantitativa.

La calificación será del número 1 al 9, el número siendo el número1= igualmente preferible, 2= igualmente y moderadamente preferible, y así sucesivamente incrementando hasta el número 9= extremadamente preferible.

3. Creación de la Matriz de Comparaciones Pareadas

Los autores construyen una Matriz de Comparaciones Pareadas con la lista de criterios y con la Escala de preferencias de Saaty.

4. Proceso de Síntesis

En este paso, se obtiene a partir de análisis matemáticos de los coeficientes de prioridad de los criterios; es decir se identifica al criterio más importante. El proceso de síntesis consta de tres

etapas: en primer lugar, se suma cada columna de la matriz, segundo, se divide los criterios con el total de la columna correspondiente y como último paso, se promedia cada elemento y así se obtiene el criterio prioritario.

5. Creación de la Matriz de Prioridades por Pares de Proyectos

Se forma una matriz con cada uno de la lista de criterios.

6. Creación de la Matriz de Prioridades de los Proyectos

Con las matrices anteriormente elaboradas se aplica el procedimiento de síntesis y se genera un Vector de Prioridad de Proyectos.

7. Obtención de la Prioridad Global

Se obtiene de la multiplicación de la Matriz de Prioridades de los Proyectos con el Vector de Prioridad de los Criterios.

Posteriormente, se define los mecanismos de financiamiento que se van a aplicar para cada proyecto.

8. Aplicación del Clustering Jerárquico

Mediante el método de clustering jerárquico se elabora un árbol en el cual se muestra las relaciones de semejanza entre los diferentes proyectos. En este método, se inicia con algunos clústeres individualmente que después se van aglomerando por similitud.

- **Resultados**

Para la evaluación de los proyectos se utilizaron los siguientes criterios: equidad e inclusión social, población y género, impactos medioambientales e impactos sociales.

Mediante el desarrollo de la matriz de Comparaciones Pareadas y el proceso de Síntesis se concluye que el criterio más importante es el siguiente: Impactos Medioambientales.

Entonces se escogen 3 proyectos medioambientales que denominan: Proyecto Medioambiental 1, 2 y 3, según pares de proyectos por cada criterio se ejecuta la matriz de comparación mostrada:

	C1	C2	C3	C4
A	0.17	0.21	0.25	0.60
B	0.44	0.30	0.36	0.14
C	0.39	0.49	0.39	0.26

Figura 7: *Matriz de Comparación por pares de Proyecto*

Fuente: *Valdés, D. (2021)*

Por último, los autores obtienen los coeficientes de prioridad de cada proyecto: Proyecto Medioambiental 1= 0.22, Proyecto Medioambiental 2 = 0.30 y Proyecto Medioambiental 3 = 0.37.

Para la aplicación del método de clustering jerárquico se establecieron 4 categorías como ejemplo: Líneas, Articulación, volumen de inversión y componente monetario y se construyó la siguiente Matriz de Correlación entre los Proyectos y las fuentes de financiamiento.

	FF1	FF2	FF3	FF4
P1	0	**50 %**	45 %	0
P2	25 %	30 %	0	**10 %**
P3	**35 %**	0	5 %	0

Figura 8: *Matriz de correlación entre los proyectos y las fuentes de financiamiento*

Fuente: *Valdés, D. (2021))*

De esta figura se puede concluir que el P1 se financiará con la fuente 2, en caso , después de su ejecución quede presupuesto se le asigna a otro proyecto.

Billadoni, M. (2021) "Clustering de clientes en un grupo de e-Marketplaces del Perú"

- **Problema**

Los e-Marketplace son la industria #1 en crecimiento de los últimos años; sin embargo, las estrategias que se tiene para el envío de publicidad y promociones son masivas y sin mayor análisis de los clientes a los que van dirigidos y sus preferencias, causando en muchos casos la incomodidad de los clientes.

Por lo tanto, esta tesis se plantea como objetivo realizar un análisis para conocer los tipos de consumidores que tienen los e-Marketplace y qué estrategias emplear con cada uno de ellos.

- **Objetivo general:**

- Obtener segmentos de clientes en base a sus preferencias y características.

- **Objetivos específicos**.

- Cantidad de segmentos armados
- Describir a los tipos de consumidores por segmento
- Generar estrategias comerciales y de marketing para cada segmento encontrado

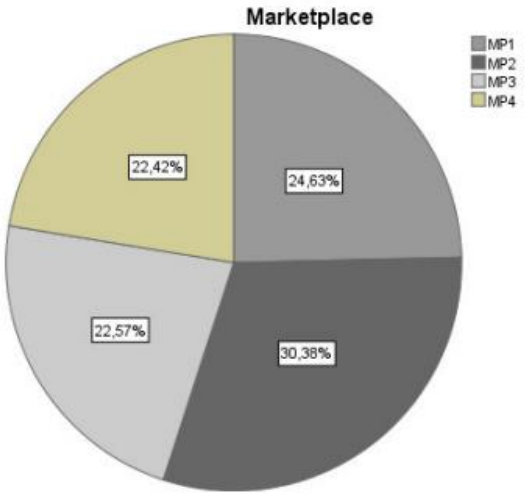

Figura 9: *Gráfico de sectores de compras por e-Marketplace*

Fuente: *Billadoni M.(2021)*

- **Metodología**

Adicionalmente se usará clúster bietápico (2 fases), está fundamentado por el algoritmo de clusterización BIRCH que sirve como descriptor de segmentos y su representación jerárquica, estas dos formas ayudan al modelo alcanzar velocidades de procesamiento y manejo de grandes grupos de datos y todavía en tiempo real Jiawei et al. (2012). que cuenta con bondades diferentes a otros métodos:

Medida de similaridad: para la agrupación de observaciones es necesario tener indicadores que nos señalan cuán parecidas son las variables entre sí. Para este caso se usó la distancia de log - verosimilitud

Algoritmo de agrupación: luego de formar grupos en base a sus distancias, se forman grupos de información en base a enfoques jerárquicos y no jerárquicos.

Pese a la identificación de valores fuera de los parámetros normales, los investigadores deciden no modificar la base de datos y mantenerla en su forma original, debido a que el método de clusterización bietápico con el Software IBM SPSS Statistics v.21 excluye los valores atípicos automáticamente para poder brindar resultados finales más acertados.

- **Resultados**

La solución de la investigación da como resultado que las preferencias de compras por página son homogéneas, teniendo un aproximado del 25% de cada cliente en las páginas. De igual forma cada segmento de cliente tiene una frecuencia de entrada a cada una de las páginas de entre el 22% al 35%. - Adicional el segmento o clúster que se obtuvo fue considerado como válido su desviación fue mayor a 0.2

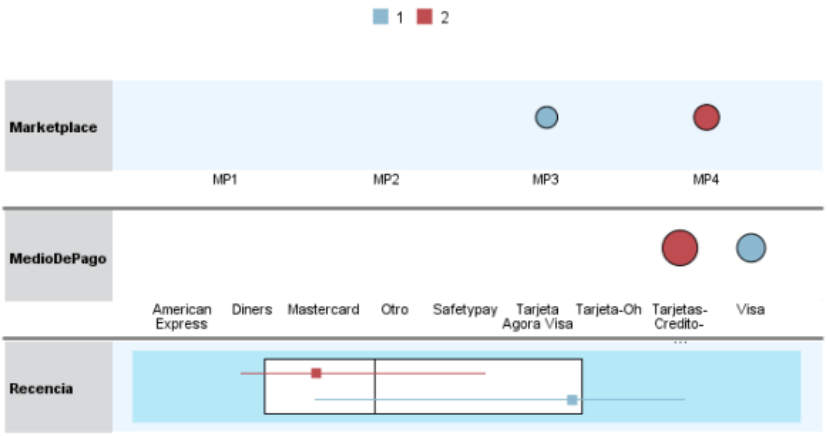

Figura 10: *Comparación de los clústeres*

Fuente: *Billadoni M.(2021)*

Chamba, S (2015) Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC

- **Problema**

Las compañías generan y almacenan mucha información por día, sin embargo, esta data no suele proporcionar beneficios directos para la empresa. Uno de los principales problemas para las compañías al elaborar estrategias para retener clientes, es que estas pueden ser conscientes de los distintos tipos de clientes que poseen, pero les resulta difícil identificarlos para llegar a ellos de forma efectiva y con cierto grado de personalización.

A pesar de tener un gran volumen de datos sobre las compras y clientes, les resulta difícil crear valor con esta información si no se cuenta con los métodos y procedimientos adecuados, sin embargo, hoy en día la segmentación de clientes es usada como una herramienta de diferenciación en el marketing, que facilita a las organizaciones entender a sus clientes y construir estrategias diferenciadas y personalizadas por grupo de clientes

Bajo lo indicado el estudio tiene como objetivo obtener la segmentación de clientes de la empresa tecnológica Master PC aplicando técnicas de minería de datos que permitan elaborar diferentes estrategias de marketing en base a los grupos de clientes que se obtengan.

- **Base de datos**

La fuente de información utilizada para este estudio corresponde a los registros de clientes y ventas realizadas desde el año 2010 hasta el año 2014, considerando 31,662 clientes, 85,272 transacciones y 06 variables.

- **Metodología**

CRISPDM (Cross-Industry Standard Process for Data Mining) fue la metodología usada para la Minería de datos, esta consta de cinco fases:

- Sample (Muestreo)
- Explore (Exploración)
- Modify (Modificación), Model (Modelado)
- Asess (Valoración)

Cada una de ellas abarca un conjunto de actividades, que se deben seguir para conseguir resultados de calidad.

En base a la literatura previa, el autor del artículo tomó 05 algoritmos para ser analizados: jerárquico, k-Means, Self-Organizing Maps (SOM), k-Medoids, Two-step, en base a la comparativa realizada por el autor se consideró los modelos más adecuados para obtener resultados más precisos para el estudio presentado.

Se tomará en cuenta la Recencia, frecuencia y Monto (RFM) para segmentar los clientes de Master PC. (RFM), y se aplicarán lossiguientes métodos: k-Means, k-Medoids, y Self-Organizing Maps (SOM).

- **Modelado**

El estudio aplicó los algoritmos seleccionados sobre las variables RFM (Recencia, Frecuencia y Monto)**.**

i) K-Means

Antes de la segmentación se realizó una evaluación a los grupos a fin de escoger el número de segmentos en que deben dividirse los datos, el autor aplicó dos medidas de evaluación para el número de clúster: el índice de la silueta y el método de curva de distorsión (método del codo), se aplicó estas técnicas de validación con distintos número de grupos

comprendidos entre 2 y 10 grupos. En base al análisis de estos resultados, se obtuvo que el número óptimo de clúster es de n = 5.

ii) K-Medoids

En este método, el autor indica que es importante calcular la distancia al punto cero ya que con esta se determina el valor de cada uno de los segmentos, los clientes con una distancia más alta son los más leales y los clientes con distancia más baja son los menos leales.

Para determinar el número de clúster, se empleó el método del índice de la silueta, se aplicó el modelado para grupos entre 2 y 10 grupos, de lo cual se obtuvo que el número de clúster óptimo es de n = 4.

iii) Self-Organizing Maps (SOM)

Para establecer el tamaño de la red, el autor aplicó una regla basada en la literatura, con la que el tamaño del mapa fue calculado en base a la siguiente regla $\mathbf{5} * \sqrt{\boldsymbol{N}}$, donde N es el número de la muestra, el valor obtenido fue 900 nodos, en base a ello, el tamaño del mapa establecido fue de 30 x 30.

Además, se seleccionó el número de iteraciones del algoritmo que finalmente fue 100, sin embargo se observa que fue un proceso iterativo de prueba y error. Asimismo, el tamaño de la red utilizada fue de 300 neuronas, con dimensión: 30x30**.**

Sin embargo, el autor detectó que los nodos generados por la red neuronal son demasiado extensos para poder ser interpretados. Ante esto, aplicó la técnica de clustering jerárquico para poder segmentar los nodos de la red y convertirlos en grupos de fácil procesamiento. mediante el método de Ward y lo observado a lo largo de la investigación se concluye que este método a través la construcción de un dendograma presenta mayor precisión en la agrupación de datos jerárquicos.

- **Solución**

En base a los datos de registros de clientes y ventas realizadas desde el año 2010 hasta el año 2014, el autor planteó simular varios modelos a fin de obtener el número óptimo de clúster con el mejor nivel de precisión, obteniendo la tabla de comparación según figura 11.

Bajo este análisis, el autor determinó que la clasificación de grupos generado por el método K-Medoids es el más adecuado pues proporciona una mayor precisión según la

comparativa realizada. Además, el número de clúster que pueden identificarse para definir la lealtad de clientes dentro de la empresa Master PC es 4.

Métodos	Precisión
K-means (5 grupos)	0.99991
K-medoids (4 grupos)	0.99999
SOM (5 grupos)	0.99992

Figura 11: *Comparación de resultados para los algoritmos KMEANS, K-MEDOIDS Y SOM*
Fuente: *Chamba, S (2015)*

- **Resultados**

El autor selecciona como método más adecuado para el estudio el k-Medoids con 04 clúster que etiqueta según nivel de lealtad: grupo 1 - Alto, grupo 2 - Bajo, grupo 3 - Medio, grupo 4: Muy Bajo. Con estos resultados el autor plantea sugerir a la empresa establecer estrategias de retención hacia sus clientes de acuerdo con los segmentos encontrados. Los resultados se muestran en la figura 12.

Es importante mencionar que, para el estudio, se encontró que entre los niveles de lealtad Bajo y Muy Bajo se reparten más del 50% de los clientes. El estudio concluye que, en base a la similitud del comportamiento entre clientes, le permitirá a la empresa en estudio elaborar estrategias de promoción y de recomendación de productos hacia sus clientes en los diferentes niveles de lealtad.

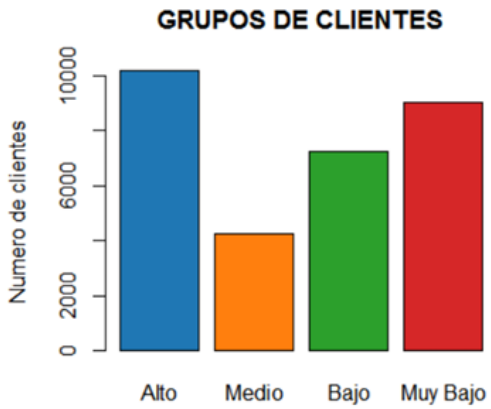

Figura 12 : *Distribución de número de clientes para cada grupo de lealtad la empresa Master PC*
Fuente: *Chamba, S (2015)*

Amazon (2021). Publicis Media automatiza la segmentación de la audiencia mediante el machine learning en AWS.

- Problema

Publics Media es una corporación multinacional de Francia del rubro de publicidad y relaciones públicas que busca generar valor para sus clientes mediante las marcas de agencia globales con las que trabaja y las capacidades de inversión, estrategia, información y análisis. El director Patrick Houlihan se enfoca en canalizar los clientes y ver la forma de utilizar la información recolectada, para ellos se hizo uso de las soluciones que brinda Amazon en donde se basa en implementar una herramienta de Amazon Web Services (AWS) que funciona a través del machine learning.

- Metodología y solución

La empresa Publics Media creó un machine learning mediante un modelo de petabytes que recopila datos de audiencia llamado Decisión Sciences Framework. Esta solución desarrollada utiliza los servicios de Amazon como el Elastic Map Reduce (EMR) que ayuda a generar clústeres para procesar la información de manera paralela de una gran cantidad de datos; también hace uso de Redshift y SparkML para generar atributos, segmentos y análisis personalizados.

Figura 13: *Representación del Elastic Map Reduce*
Fuente: *Amazon WebSite (2022*)

Para el director estos cambios representan una gran ventaja ya que estos sistemas funcionan como solucionadores y una herramienta para tomar decisiones ya que se generan

clustering en base a la calificación de perfiles, audiencias, proveedores, palabras claves, impresiones en los anuncios y otras características.

- Conclusión

Para Publics Media, la aplicación de machine learning enfocado a clustering ha sido de un impacto positivo, a que mencionan que el valor agregado al final del día a día serán toda la información recolectada y los algoritmos que se generan para dar como resultado una audiencia donde tendremos los atributos que generarán mayor impacto y en donde se tendrá que realizar los enfoques en las decisiones.

2.2 Bases Teóricas

2.2.1. Machine Learning

El Machine Learning es la ciencia o arte de programación de computadoras las cuales pueden aprender a partir de la data (Géron, 2019).

Según Hurwitz y Kirsch (2018):

El ML posee dos enfoques: descriptivo y predictivo. El primero se basa en comprender la situación actual y el segundo dispone del conocimiento recogido en el análisis descriptivo para anticiparse al futuro mediante la predicción de escenarios y resultados utilizando algoritmos de ML. (p.5)

Hurwitz y Kirsch (2018) proponen tres clasificaciones para los modelos de ML, los que se describen a continuación:

- Aprendizaje supervisado: En este se “enseña” al algoritmo a obtener una salida en base a información conocida y existente. Se suele usar cuando la data histórica es una buena referencia para la predicción de eventos futuros (Manrique, 2020).
- Aprendizaje no supervisado: El algoritmo no tiene información clara ni conocida e intenta aprender didácticamente (Géron, 2019). Analiza los datos no clasificados para obtener segmentos o clasificaciones. (Manrique, 2020)

- Aprendizaje reforzado: También llamado agente, analiza el medio, selecciona y define acciones (Géron, 2019). Presenta un comportamiento ligeramente familiar el aprendizaje no supervisado; sin embargo, este no usa datos de muestras para su aprendizaje, sino que usa el método de prueba y error.

Redes neuronales y aprendizaje profundo: se basan en las redes neuronales que posee el cerebro humano y suelen usarse cuando se posee data no estructurada, es decir, no organizada. Los modelos de redes neuronales cuentan con capas de entrada y salida de datos, y de modificación de estos en función a nodos interconectados. El aprendizaje profundo se puede definir como una versión más compleja de las redes neuronales que integra tanto el aprendizaje supervisado, como el no supervisado. (Hurwitz y Kirsch, 2018)

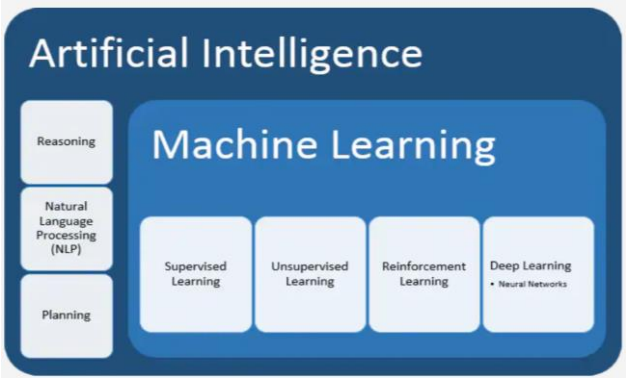

Figura 14: *Machine Learning*
Fuente: *IBM Limited Edition (2018)***.**

Asimismo, se describen los algoritmos de ML más utilizados según Hurwitz, J. y Kirsch, D. (2018):

- Bayesiano: el algoritmo Bayesiano permite codificar entendimientos o conocimientos previos de cómo debería verse el modelo.
- Clustering (agrupamiento): son algoritmos de fácil ejecución que permiten encontrar elementos dentro del conjunto de datos que están relacionados.

- Árboles de decisión: se estructuran a modo de ramificación para mostrar los resultados que se pueden generar en diferentes situaciones. Para este algoritmo se requieren de nodos que contienen las probabilidades de que ocurra un evento u otro.
- Redes neuronales y aprendizaje profundo: busca representar la forma en que un cerebro funciona a través de nodos interconectados.
- Regresión lineal: estos algoritmos son usados a menudo para análisis estadísticos y miden la correlación entre las variables del conjunto de datos.

2.2.2. Aprendizaje no supervisado

Según Ethem M. (2020), "el aprendizaje no supervisado no contiene una clasificación o data etiqueta, este se emplea cuando el problema demanda una cantidad fuerte de datos sin etiquetar. Asimismo, este método lleva a cabo un proceso iterativo, analizando los datos sin intervención humana."

Según, Gema V. (2022):

> El uso del aprendizaje no supervisado se vuelve muy común en las compañías, debido a que la mayoría de las veces nos encontramos con grupos de datos que no tienen las etiquetas que esperamos en nuestro análisis.
>
> Computacionalmente, el aprendizaje no supervisado es mucho más complejo que el supervisado, además, el porcentaje de precisión y confiabilidad de resultados del lenguaje no supervisado es menor que el supervisado. Esto no resta la valoración del aprendizaje no supervisado en el camino de la Inteligencia Artificial.
>
> En el Aprendizaje no supervisado se cuenta con diversas técnicas, entre las principales tenemos: El agrupamiento de datos, uso del ACP para reducir dimensionalidad, identificación de valores atípicos e identificación de novedades.

2.2.2.1 Clustering

Es un método de inteligencia artificial de aprendizaje no supervisado, el cual se basa en la segmentación de grupos por medio de algoritmos. La finalidad de esta técnica es encontrar grupos de acuerdo a la similitud que presentan en comportamiento las variables estudiadas. De acuerdo con Wiskott, L. (2014) “el aprendizaje por Clustering es una técnica donde no se considera a qué grupo pertenece verdaderamente cada observación. Esta particularidad es lo que diferencia al clustering de otros métodos de clasificación.” Wiskott L. (2014) (p.2).

Para el análisis de clúster requiere variables numéricas ya que este trabaja con igualdades y distancias, por dicha razón es necesario que todas aquellas variables que son cualitativas o categóricas pasen por una fase de pre- procesamiento de data para ser convertido en variables numéricas. Las diferentes distancias que se le exija al análisis nos darán diferentes y mayor cantidad de agrupaciones.

Algunas aplicaciones prácticas que tiene este tipo de aprendizaje es elaborar tipos A consumidores, siendo las distancias características ya sean físicas, demográficas o preferencias de compras.

Se cuenta con varias opciones de métodos de clustering, habiendo dos más conocidas:

• K-Means clustering: “Es un método no jerárquico para asociar objetos que agrupa el conjunto de datos en K clústeres diferentes, lo que representa que ninguna observación puede pertenecer a más de un clúster. El número de clústeres o subgrupos requeridos se establece al inicio”. Gil, C. (2018).

• K-Nearest Neighbours (k-NN): “Es un clasificador de aprendizaje supervisado no paramétrico, se usa la proximidad para hacer clasificación o predicción sobre la agrupación de un punto de datos individual. No obstante, puede utilizarse para problemas de regresión y clasificación, usualmente se emplea como algoritmo de clasificación, teniendo como que se puede encontrar puntos similares cerca unos de otros”. Artyco (2019)

• Mean-Shift Clustering: “Este es un algoritmo basado en una ventana deslizante que intenta encontrar áreas densas de puntos de datos. Está basado en el centroide, lo que implica que el objetivo es encontrar los puntos centrales de cada grupo/clase. Funciona actualizando los candidatos para que los puntos centrales sean la media de los puntos dentro de la ventana

deslizante. Estas ventanas candidatas luego se filtran en una etapa de post-procesamiento para eliminar casi todos los duplicados, formando el grupo final de puntos centrales y sus conjuntos correspondientes". Artyco (2019)

2.2.3 Inteligencia artificial -python

Según, Rouhiainen, L., (2018):

La IA es como la habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana, ósea, es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones, como lo haría un humano. (p.7)

Es importante destacar que el margen de error en alguna actividad o tarea es menor que cuando es realizada por los humanos. Las tecnologías que funcionan con inteligencia artificial ayudan a lograr mayor eficiencia.

Los campos en los cuales se pueden aplicar la IA son los siguientes: Reconocimiento de imágenes estáticas, clasificación y etiquetado, mejoras en el desempeño de la estrategia algoritmia comercial, procesamiento escalable de datos de pacientes, mantenimiento predictivo, detección y clasificación de objetivos, distribución de contenido en redes sociales y la protección contra amenazas de seguridad cibernética. (Rouhiainen, L. 2018)

También se menciona que la Inteligencia artificial actualmente genera un gran impacto en nuestras vidas, para Rouhiainen, L (2018) define la IA actualmente como:

> Las tecnologías de IA han comenzado a desarrollar como nunca antes la capacidad de ver (visión artificial), oír (reconocimiento de voz) y entender (procesamiento del lenguaje natural). Antes, estas habilidades pertenecían únicamente a los seres humanos, pero en el futuro próximo las máquinas y los robots las podrían desarrollar gracias a la IA. (pág. 13)

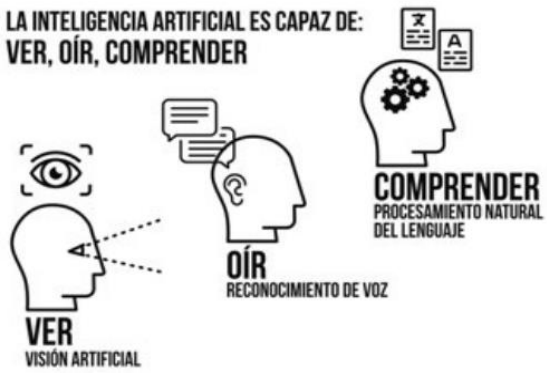

Figura 15: *Inteligencia Artificial*
Fuente: *Rouhiainen, L., (2018)*

Python es uno de los lenguajes de programación usados para crear modelos o sistemas de inteligencia artificial. Se pueden desarrollar tanto modelo con aprendizaje supervisado y no supervisado. Para Bellido, F. (2022) menciona que:

Python puede usarse para programación web, desarrollo de interfaces gráficas, librerías matemáticas, desarrollo de software y muchas más opciones. Actualmente es el lenguaje de programación con mayor uso en el ámbito de la IA con librerías para computación científica. computación avanzada o aprendizaje automático" (p.10)

2.2.4 Análisis de componentes principales (ACP)

Según Marín, J. (2020) indica que:

"El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de datos, aunque si esto último no se cumple se puede dar una interpretación más profunda de dichos componentes" (p.42)

Zapotitla, J. (2019) nos menciona que "El análisis de los Componentes Principales propone la transformación a un nuevo conjunto sintético de variables (los componentes principales), que no están relacionados y se encuentran ordenados de tal forma que los primeros conservan la mayor parte de la variación presente en todas las variables originales" (p.1). Es importante mencionar que las nuevas variables son independientes a las originales.

Se aplican dos conceptos matemáticos para su aplicación: eigenvectors, multiplicación entre una matriz y un vector da como resultante un múltiplo de vector original multiplicado por número, aquí nace el segundo concepto, eigenvalue, que es el número que multiplica al vector resultante.

Para Pamela, R. (2018) el ACP es:

"El algoritmo de Componentes Principales que consiste en una técnica de selección de características concretas que utiliza una transformación ortogonal para convertir un grupo de observaciones de variables, posiblemente correlacionadas, en un conjunto más reducido de variables que ya no guardan correlación y que se conocen como componentes principales." (p.30)

El ACP brinda una técnica en la cual la dimensionalidad de los datos se reduce sin afectar validez ni precisión de los mismos. Es una de las técnicas más usadas para el proceso de la aplicación de Machine Learning - Aprendizaje no Supervisado.

Según, Valenzuela, G. (2022), "El método emplea una serie de transformaciones lineales que apunta a proyectar sobre las direcciones de mayor varianza que muestran nuestros datos." (p.3), como se observa en la figura. Para Valenzuela, G. (2022), "El vector que tiende a apuntar más al eje de las abscisas es el componente principal y este apunta a los datos con mayor variabilidad y el segundo apunta a los datos con mayor varianza y menor correlación con el primer vector" (p.4)

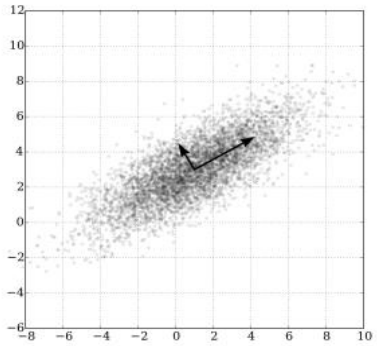

Figura 16: *Representación Algebraica ACP*

Fuente: *Valenzuela, G. (2022)*

2.2.5 Algoritmo K-Means.

El método K-Means consiste en la agrupación de un conjunto de datos numéricos. Para Franco, A. (2021) menciona que “Su objetivo es encontrar una partición de ese conjunto, la cual consiste en una serie de grupos que son representados por un centro.” Por lo que se puede definir como el procedimiento para poder determinar la posición de los centros de cada data evaluada, con la finalidad de encontrar grupos que tengan atributos similares entre ellos.

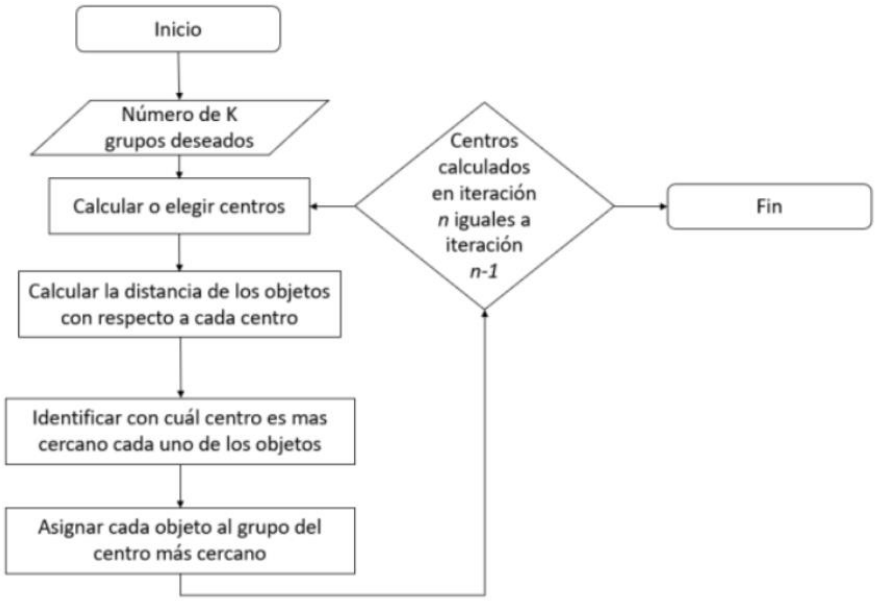

Figura 17: *Descripción del algoritmo K-Means*
Fuente: *Franco, A. (2021*)

López, S. (2007), también define esta metodología k-Means con los siguiente:

> El algoritmo k-Means fue propuesto hace poco más de tres décadas y es uno de los algoritmos de agrupamiento más usados en una amplia variedad de áreas. K-Means es un algoritmo de agrupamiento restringido, por lo que recibe como parámetro el número de agrupamientos a formar y se encuentra definido sobre datos continuos, es decir, únicamente permite trabajar con objetos descritos por medio de un conjunto de atributos numéricos. (p.19)

El procedimiento para esta metodología explicada por López, S. (2007) se basa en los siguiente:

- Seleccionar aleatoriamente los centros iniciales.
- Asignar cada objeto al agrupamiento cuya distancia con su centro sea mínima.
- Recalcular los centros.
- Repetir los pasos 2 y 3 hasta que no haya cambios en los centros para dos iteraciones consecutivas.

El autor Ordoñez, H. (2019) menciona en su artículo que esta metodología también presenta algunos inconvenientes:

> El algoritmo K-Means tiene algunos inconvenientes, como por ejemplo la sensibilidad en la inicialización de los centroides, que impacta directamente en la creación óptima de los grupos; y la necesidad de tener que definir previamente un número (k) de grupos a crear. Para obtener resultados apropiados, se deben evaluar distintos valores de k y se debe ejecutar varias veces el algoritmo partiendo de diferentes centroides aleatoriamente inicializados, por lo tanto, se incrementa el tiempo de ejecución. (p.10)

2.2.6 Algoritmo K-Medoids

Según Kaufman, L. y Rousseeuw, P. (1990), "El K-Medoids es un algoritmo de agrupamiento particional que se cambia ligeramente del algoritmo k-Means. El algoritmo k-Means elige la media como los centroides pero en El K-Medoids, se eligen puntos de datos originales para ser los medoides" (P.25)

De acuerdo con Jin, X. y J. Han, J. (2010):

> "Un medoide se puede definir como aquel objeto de un grupo, cuyo promedio de disimilitud a todos los objetos en el clúster es mínima. Cada objeto restante es agrupado con el medoid más cercano e iterativamente estos algoritmos realizan todos los intercambios posibles entre los objetos representativos y los que no lo son, hasta que se minimice una medida de disimilitud entre los k-Medoids y los vectores de observaciones que forman los conglomerados."

2.2.7 Clustering Jerárquico

El clúster jerárquico o hierarchical clustering es un enfoque en el cual no es necesario una agrupación inicial. Además de que tiene la posibilidad de tener representaciones, en este caso denominados árboles, las observaciones que son conocidas como dendrogramas.

El clúster Aglomerativo es el más común dentro de la segmentación jerárquica, y se refiere al hecho de que el dendograma se crea empezando por las hojas, luego combinando subgrupos hasta el "tronco".

Los tipos de clustering son:

-Dendograma: es una representación que ilustra una organización de jerarquía entre los componentes, se puede representar de forma vertical u horizontal. Cada una de las hojas de este gráfico representa un elemento, cada que se va formando el árbol algunas de las hojas se fusionan en ramas, ya que se va observando la similitud de unas con otras. Mientras más va creciendo el árbol, las ramas se van uniendo con más ramas u hojas, según similitud. Gil C. (2019)

Las hojas que se van uniendo al inicio del árbol son aquellos componentes que tienen mayor similitud, mientras que las que están más en la parte alta son las más diferentes entre sí. Por esto, la forma de leer un dendograma es fijándose donde se forman las ramas o la unión de las hojas; ya que las conclusiones irán saliendo de acuerdo con donde se vayan juntando los elementos. Gil C. (2019)

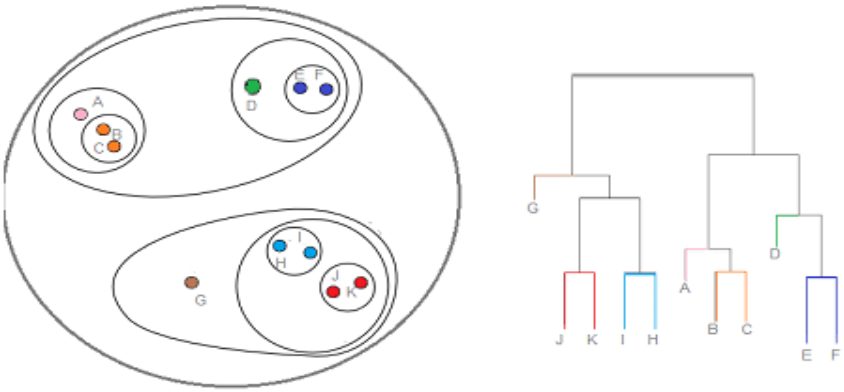

Figura 18: *Dendograma representando clústeres jerárquicos anidados (2017)*
Fuente: *Gil C. (2019)*

De acuerdo al autor Gil C. (2019), comenta que “Como primer paso es necesario establecer la medida de disimilitud a utilizar entre cada par de observaciones. Por otro lado, se encuentra la disimilitud entre pares de grupos de observaciones, donde aparece el concepto de método de unión o linkage, que mide esta disimilitud.”

Los tipos de linkage son:

+ Complete: Distancia máxima entre clústeres
+ Average: Distancia media entre clústeres
+ Single: Distancia mínima entre clústeres
+ Centroid: Distancia entre centros

2.2.8 Aprendizaje supervisado

Según Judith Sandoval (2018):

“El aprendizaje supervisado es cuando entrenamos un algoritmo de Machine Learning dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro el algoritmo hace una predicción conociendo las características. En este tipo de aprendizaje hay dos algoritmos (entrenamientos): el de clasificación y el de regresión.” (p.2)

El entrenamiento de clasificación indica a qué conjunto corresponde el elemento en cuestión. El algoritmo encuentra patrones en la data y la clasifica en grupos y mediante una comparación ubica al nuevo elemento.

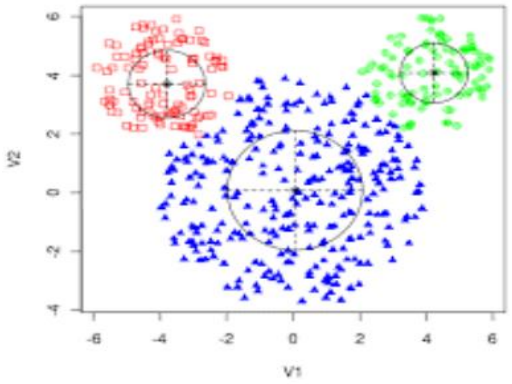

Figura 19: *Gráfico de un algoritmo de clasificación*
Fuente: *Sandoval J. (2018)*

En cuanto al algoritmo de regresión, en esta técnica se obtiene un número como resultado. Es decir que en base a unas características dadas arroja el valor de la etiqueta que le corresponde. Según Eliseo, P. (2019): “La regresión lineal es el modelo más sencillo empleado en el aprendizaje supervisado. Su función se trata de una combinación lineal de las características del ejemplo de entrada” (p.21)

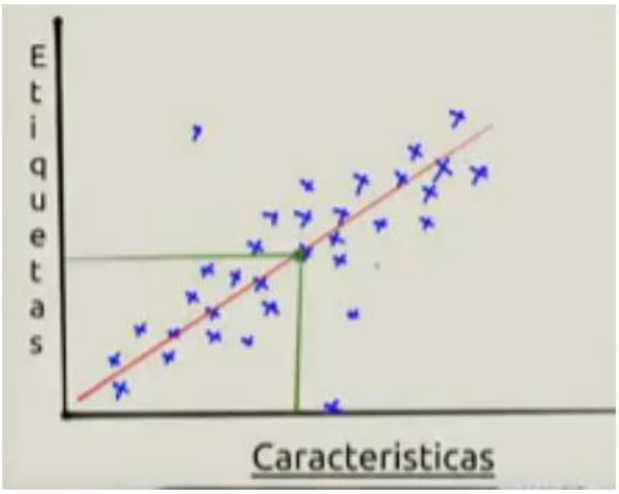

Figura 20: *Gráfico de un algoritmo de regresión*
Fuente: *Eliseo, P. (2019)*

CAPÍTULO III: ENTORNO EMPRESARIAL

3.1 Descripción de la empresa

3.1.1 Reseña histórica y actividad económica

Falabella da sus inicios hace más de 130 años, Salvatore Falabella, un inmigrante italiano, fue quien fundó en 1889 una sastrería pequeña ubicada en Chile, la cual se convirtió en unos años en una de las empresas más grandes del país.

Arnaldo Falabella, hijo de Salvatore Falabella, fue quien impulsó el nombre del negocio y junto a Alberto Magnasco ampliaron el taller familiar, incorporando nuevos productos y puntos de venta, entre estos productos se encontraban prendas femeninas, masculinas, artículos para el hogar.

Para 1980, Reinaldo Solari, director del grupo Falabella, ya había consolidado el nombre de la marca y comenzaron a implementar su propia tarjeta de crédito CMR Falabella. En este punto la compañía empezó con el proceso de expansión fuera de Chile, actualmente Falabella tiene mayor importancia en Sudamérica por tiendas por departamento como en Perú, Colombia, Brasil, Argentina, Uruguay y México.

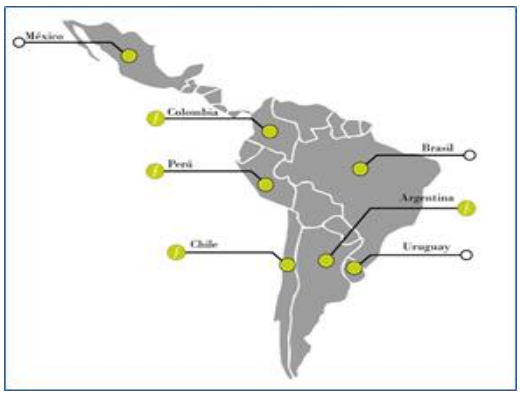

Figura 21: *Centro de Negocio*
Fuente: *Grupo Falabella (2018)*

Actualmente en Falabella Perú, hay varias cadenas asociadas que trabajan de forma independiente en cada tipo se sector, por ejemplo Sodimac/Maestro, que tienen su negocio enfocado al sector de materiales e insumos para la construcción y el hogar, Banco Falabella, que opera como entidad financiera, Mall Plaza/Open Plaza, enfocados en el sector de tiendas comerciales, Seguros Falabella, como proveedor de seguros y soporte, y están las tiendas por departamento y supermercado como Falabella Retail, Tottus y Linio.

Figura 22: *Empresas asociadas*
Fuente: *Grupo Falabella (2018)*

Falabella Retail, está dentro del mercado como tienda por departamento, en este sector el negocio se basa en la comercialización de productos y/o servicios en grandes cantidades hacía los clientes, esto implica que estas tiendas por departamento posean grandes almacenes y amplios centros de puntos de venta donde se pueda ofrecer al cliente una variedad de productos como textil, accesorios, decoración, muebles, tecnología, etc.

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

Falabella Retail Perú presenta la siguiente estructura organizacional. Figura 05

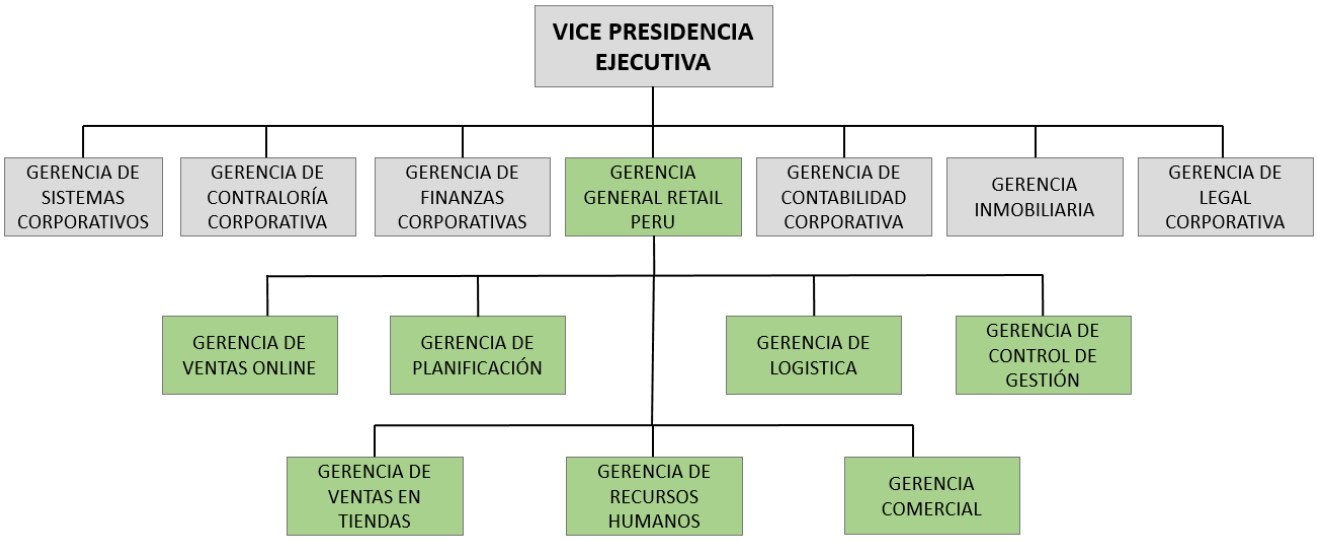

Figura 23: *Organigrama Falabella Perú*
Fuente: *Grupo Falabella (2018)*

3.1.2.2 Cadena de suministros

El proceso de la cadena de suministro inicia con la compra de la mercadería, esta puede ser compra de proveedor nacional o de importada, si es nacional el producto se recibe directamente en el centro de distribución (CD) si es importada primero llega a puerto para gestionar toda la documentación y luego recién es llevada al CD.

Cuando ingresa la orden de compra al CD, estas se pueden diferenciar en dos procesos, el primero es Cross Dock, la cual la OC ya se encuentra pre distribuida, es decir apenas ingresa el producto está ya está automatizada para enviarse a las tiendas según el detalle de la orden de compra. Si no es pre-distribuida, estos se empujan a tienda con previa evaluación del equipo de planeamiento según el nivel de inventario y rotación por tienda.

Estos productos luego pasan al muelle de tiendas, donde se cargan las unidades a las unidades de transporte y se realiza el proceso de transferencias a tienda (TRX), finalmente, esta mercadería llega a las tiendas según el cronograma de envíos y son puestas en el almacén de tienda para que en la brevedad esta esté exhibida en los puntos de venta al cliente final.

3.1.3 Datos generales estratégicos de la empresa

3.1.3.1 Visión, misión y valores o principios

- Visión:

Contribuir al mejoramiento de la calidad de vida de nuestros clientes en cada una de las comunidades en las cuales nos insertamos

- Misión:

Liderar el comercio Latinoamericano, entregando la mejor experiencia de compra omnicanal.

- Valores:

a) Superamos las expectativas de los clientes: El cliente es el centro de de nuestras decisiones

b) Hacemos que las cosas pasen: Nos hacemos cargo de nuestras acciones, tomamos riesgos y construimos oportunidades.

c) Crecemos por nuestros logros: Trabajamos en equipo e invertimos nuestras capacidades en atraer y formar a los mejores talentos como una ventaja competitiva.

d) Actuamos con sentido: Nos mueve hacer lo correcto para entregar lo mejor de nosotros a los consumidores de América Latina.

3.1.3.2 Objetivos estratégicos

a) Estabilidad operacional: A pesar del menor crecimiento de la economía peruana registrado en los últimos años, del menor dinamismo del consumo y la mayor competencia existente, Saga Falabella ha mostrado resultados operacionales relativamente estables gracias a las fortalezas desarrolladas.

b) Indicadores de endeudamiento estables: De acuerdo al último informe anual de Saga Falabella posterior a la pandemia, se observó una mejora respecto a años previos.

c) Líder de mercado: Las tiendas Saga Falabella mantienen el liderazgo del mercado con una participación de alrededor del 50%. Asimismo, cuentan con un sólido posicionamiento de la marca.

d) Adecuada capacidad de generación: enfocada en la repartición de dividendos que tienen como objetivo financiar expansión de otros negocios.

e) Sinergias con los diversos negocios del grupo: La alianza comercial con el Banco Falabella le otorga ventajas a la Empresa, al impulsar el consumo en sus tiendas a través del crédito.

3.1.3.3 Evaluación interna y externa. FODA cuantitativo

A continuación, se detalla en el gráfico las Fortalezas, Oportunidades, Debilidades y Amenazas identificadas para la empresa Saga Falabella. Asimismo, se desarrolló el FODA cuantitativo para la empresa, en cual se detalla en el siguiente gráfico.

FACTORES INTERNOS		Ponderación	Clasificación
FORTALEZAS			
1	Sólido equipo profesional y experiencia de accionistas	10%	2
2	Integración horizontal con otros negocios de retail	5%	3
3	Eficiencia logística con sistemas automatizados y técnología de vanguardia	10%	4
4	Gran posicionamiento de marca	15%	4
5	Capacidad financiera para mayor expansión	15%	3
DEBILIDADES			
1	Se percibe como "marca masiva"	10%	3
2	Baja interacción y comunicación con el cliente	10%	2
3	No tiene gran diferenciación respecto a la competencia	15%	4
4	Proceso de devolución lento, generando mala experiencia con el cliente	5%	2
5	Falta de conexión en la experiencia de Falabella.com	5%	3
		100%	3.15

FACTORES INTERNOS		Ponderación	Clasificación
OPORTUNIDADES			
1	Expansión y ampliación de operaciones, mayor presencia en provincias	10%	3
2	Aprovechamiento de locación virtual para captar más clientes	15%	2
3	Crecimiento del sector retail y capacidad adquisitiva de la población	15%	3
4	Auge de centros comerciales	10%	2
AMENAZAS			
1	Incremento de competencia, tiendas especializadas y nuevas marcas	15%	4
2	Tendencia de protección medioambiental	10%	4
3	Aumento del cybercrimen	5%	4
4	Posibilidad de incremento de impuestos para importación	10%	3
5	Inestabilidad política en Perú	10%	3
		100%	3.05

Tabla 1 : *FODA Saga Falabella Perú*
Fuente: *Elaboración propia (2023)*

En base al análisis realizado al FODA cuantitativo, se puede indicar que la empresa Saga Falabella se encuentra en el cuadrante I, es decir que la compañía debería tomar estrategias con la finalidad de continuar el crecimiento y/o construcción de su negocio.

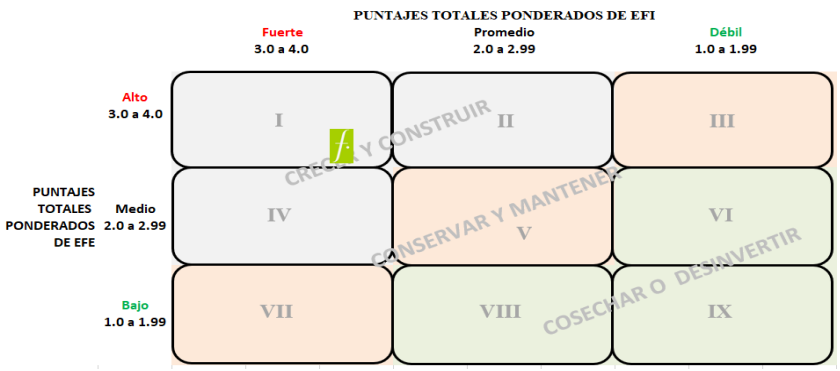

Figura 24: *Matriz EFI-EFE*
Fuente: *Elaboración propia (2023)*

3.2 Modelo de negocio actual (CANVAS)

El modelo de negocio actual de Falabella Retail consiste en ejecutar actividades estratégicas que permitan mantener la mejor calidad de relación con sus proveedores y clientes. Falabella basa sus estrategias en evaluar el mercado y analizar tendencias que le permitan ser competitivo frente al resto.

Ahora post pandemia su canal online se ha vuelto más visible por lo que siempre busca que el cliente tenga la mejor experiencia de compra, trata de fortalecer su propuesta de valor que indica simplificar y disfrutar más la vida tanto para los trabajadores y para clientes.

EMPRESA: Falabella Retail **MODELO DE NEGOCIO CANVAS**

SOCIOS CLAVES	ACTIVIDADES CLAVES	PROPUESTA DE VALOR	RELACIÓN CON LOS CLIENTES	SEGMENTO DE CLIENTES
-Proveedores ya sean importados o nacionales, que brinden distintos productos como textil, muebles, juguetes, tecnología. -Proveedores de servicios como los de instalaciones electrónicas, seguridad. -Falabella.com -MarketPlace.	-Venta de productos en tiendas por departamento y venta online. -Logística de distribución. -Soporte post-venta. **RECURSOS CLAVES** -Soporte y sistemas tecnológicos. -Base de datos de ventas. -Centro de distribución automatizado.	-Simplificar y disfrutar más la vida al cliente. -Ofrecer mejor servicio de calidad. -Brindar la mejor experiencia de compra. -Vender productos de calidad y garantizar despachos rápidos.	-Descuentos o promociones exclusivas con el uso de tarjetas CMR. -Servicio al cliente a través de redes sociales. **CANALES** -Tiendas físicas. -Plataforma de venta web. -Redes sociales. -Avisos publicitarios de marketing en TV.	-Familias del sector económico A, B y C. -Tiene plataformas de pago interactuadas en tiendas para las personas de tercera edad. -Zonas interactivas para probar el producto, segmentado a clientes jóvenes.

ESTRUCTURA DE COSTES		FLUJO DE INGRESOS
Costos Fijos	Costos Variables	-Venta de productos a través de sus canales. -Ingresos por acuerdos comerciales con proveedores.
-Sueldos a empleados. .Alquileres de espacios -Gastos de luz, agua, internet.	-Costo por volumen de compra. -Publicidad. -Servicios de transporte..	

Figura 25: *Modelo Canvas Falabella Retail.*
Fuente: *Elaboración propia (2023)*

3.3 Mapa de procesos actual

A continuación, se presenta el mapa de procesos dividido en 03 categorías donde se muestran las principales áreas o equipos de trabajo según la función dentro del diagrama o estrategia planteada por Saga Falabella.

En el proceso estratégico cuenta con 03 principales áreas que son planificación estratégica, contraloría corporativa y gestión de riesgos.

Respecto al proceso clave u operativo, se tiene las áreas de gestión comercial y marketing, gestión de la calidad, gestión de operaciones e investigación de mercado.

El proceso de soporte está formado por 05 áreas que son TI, recursos humanos, compras y abastecimiento, gestión de infraestructura y gestión financiera.

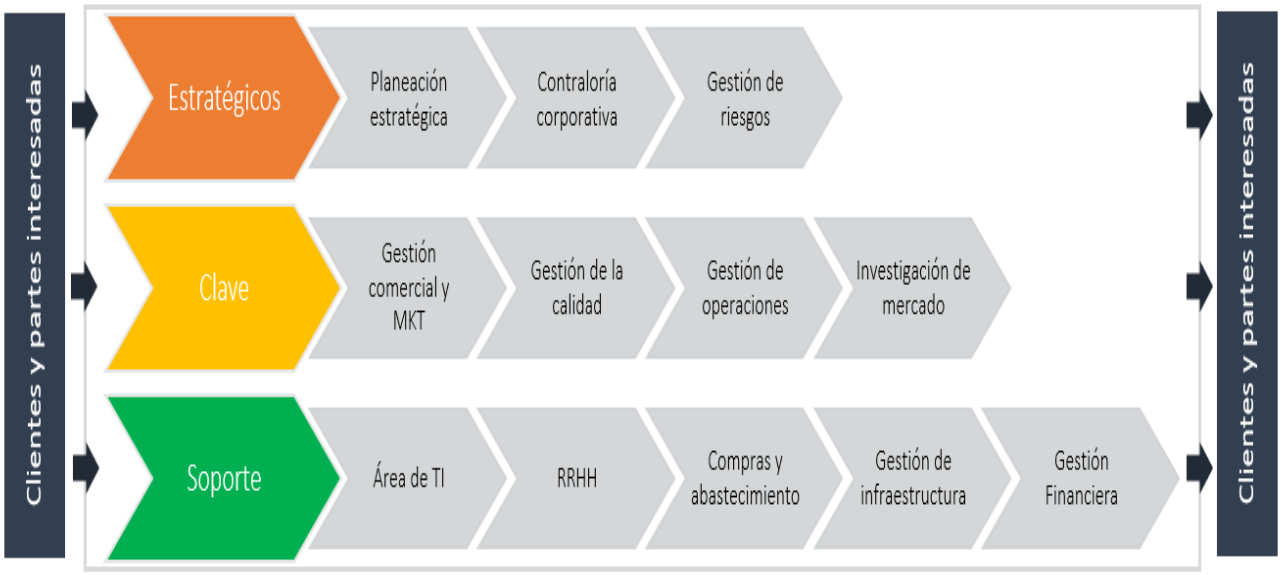

Figura 26: *Mapa de procesos Saga Falabella*
Fuente: *Elaboración propia (2023)*

CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

4.1 Diseño de la Investigación.

4.1.1. Enfoque de la investigación

El presente trabajo de investigación será desarrollado bajo un enfoque cuantitativo, se plantea desarrollar un modelo que permita segmentar a los clientes según comportamiento de compra e información clave del consumidor con la finalidad de desarrollar estrategias de ventas y marketing.

Se utilizará la técnica de aprendizaje no supervisado bajo los métodos de K-Means, K-Medoids y Cluster Jerárquico, asimismo se utilizará el método del codo para que sea una guía al elegir la cantidad de segmentos a utilizar.

4.1.2. Alcance de la investigación

Para esta investigación, el alcance es correlacional dado que se analizará las variables seleccionadas relacionadas al comportamiento de compra e información clave del consumidor para definir perfiles de compra y comprender la relación entre las variables en estudio.

4.1.3. Tipo de investigación

El estudio tiene un tipo de diseño no experimental ya que las variables no serán variadas intencionalmente sino observadas y analizadas en su contexto natural.

El diseño de la investigación es transeccional o transversal, ya que no se analizará la variable a lo largo del tiempo sino los datos serán recolectados en un solo momento determinado y poder caracterizar nuestras variables.

4.1.4. Población y muestra

- Población

Base de datos de las compras de equipos electrónicos totales realizadas a nivel país de la empresa Saga Falabella.

- Muestra

Base de datos de las compras de equipos electrónicos realizadas noviembre 2022 a enero 2023 de la empresa Saga Falabella.

4.2 Metodología de implementación de la solución.

A continuación, se detallan las etapas del proceso de implementación enfocado en la metodología de machine learning del tipo aprendizaje no supervisado-clasificación:

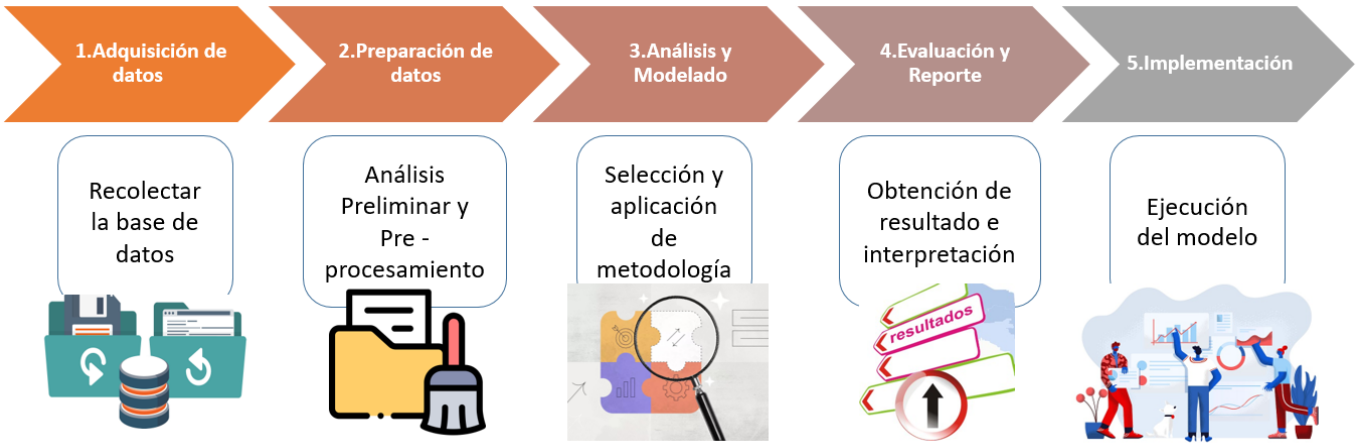

Figura 27: *Metodología de implementación*
Fuente: *Elaboración propia (2023)*

1. Adquisición de datos.

Es la recolección de los datos que serán analizados a lo largo del análisis, estos serán obtenidos del Data Ware House (DWH) de Falabella. Para este trabajo, nos enfocaremos son los siguientes: sexo, marca, sublínea, local, marca, transacción, venta en unidades.

2. Preparación de datos.

Con los datos recolectados se realiza un análisis preliminar, el objetivo es entender el comportamiento de estas informaciones y realizar un preprocesamiento, lo cual consiste en limpiar, integrar y consolidar los datos.

En la empresa Falabella se han presentado casos donde la información recolectada tiene data incorrectamente codificada o aparecen sin valores, por lo que al analizar bajo esta metodología se tendría que limpiar la base de datos para tener resultados más acertados.

3. Análisis y modelado.

El análisis se hará con el método de Machine Learning no supervisado por clustering, la construcción del modelo se realizará en “Jupyter” con el lenguaje Python.

Mediante la construcción del modelo se pretende que con la segmentación de los datos se pueda obtener la información para poder proponer estrategias de marketing para incrementar las ventas online y mejorar la experiencia de compra de los clientes.

4. Evaluación

En esta etapa se evaluará los resultados obtenidos en base al análisis del dataset estudiado, lo que se espera de la predicción bajo la aplicación de técnica de clustering - machine learning es obtener la segmentación de clientes en base a sus preferencias de compra, ya sea en productos y marcas.

Se deberá medir y validar el nivel de confiabilidad de la segmentación. Tengamos en cuenta que es complicado definir cuándo el agrupamiento es aceptable, para ello existe la validación interna y externa: la primera mide la efectividad del clustering con la información propia del algoritmo, la validación externa, hace uso de información adicional para su verificación.

En el presente trabajo se evaluará el modelo con una validación interna: Cohesión y Separación

5. Implementación

Con los resultados obtenidos de la modelación aplicando clustering - machine learning, se elaborarán estrategias enfocadas a la mejora de las ventas de Falabella y en la optimización de los procesos de compra de sus clientes en base a sus preferencias de acuerdo a la segmentación que realizó nuestro modelo. Además, Saga cuenta con una elaborada plataforma digital que daría soporte a nuestras propuestas.

4.3 Metodología para la medición de resultados de implementación.

Para el estudio es importante definir si la cantidad de grupos (K) o clúster seleccionado es el más adecuado para el desarrollo del modelo. Para ellos se utilizan tres métodos de clúster y un análisis que soportaría los resultados obtenidos (Método del codo)

- K-Means / K-Medoids / Cluster Jerárquico

Estos métodos miden las distancias de cada punto de los datos recopilados y los toman de referencia para generar agrupaciones. Se considera que un punto está en un grupo en particular si está más cerca a otro conjunto de puntos del mismo grupo.

- Método del codo

El método del codo ayuda a elegir el valor óptimo de 'k' (número de grupos) ajustando el modelo con un rango de valores de 'k'.

4.4 Cronograma de actividades y presupuesto.

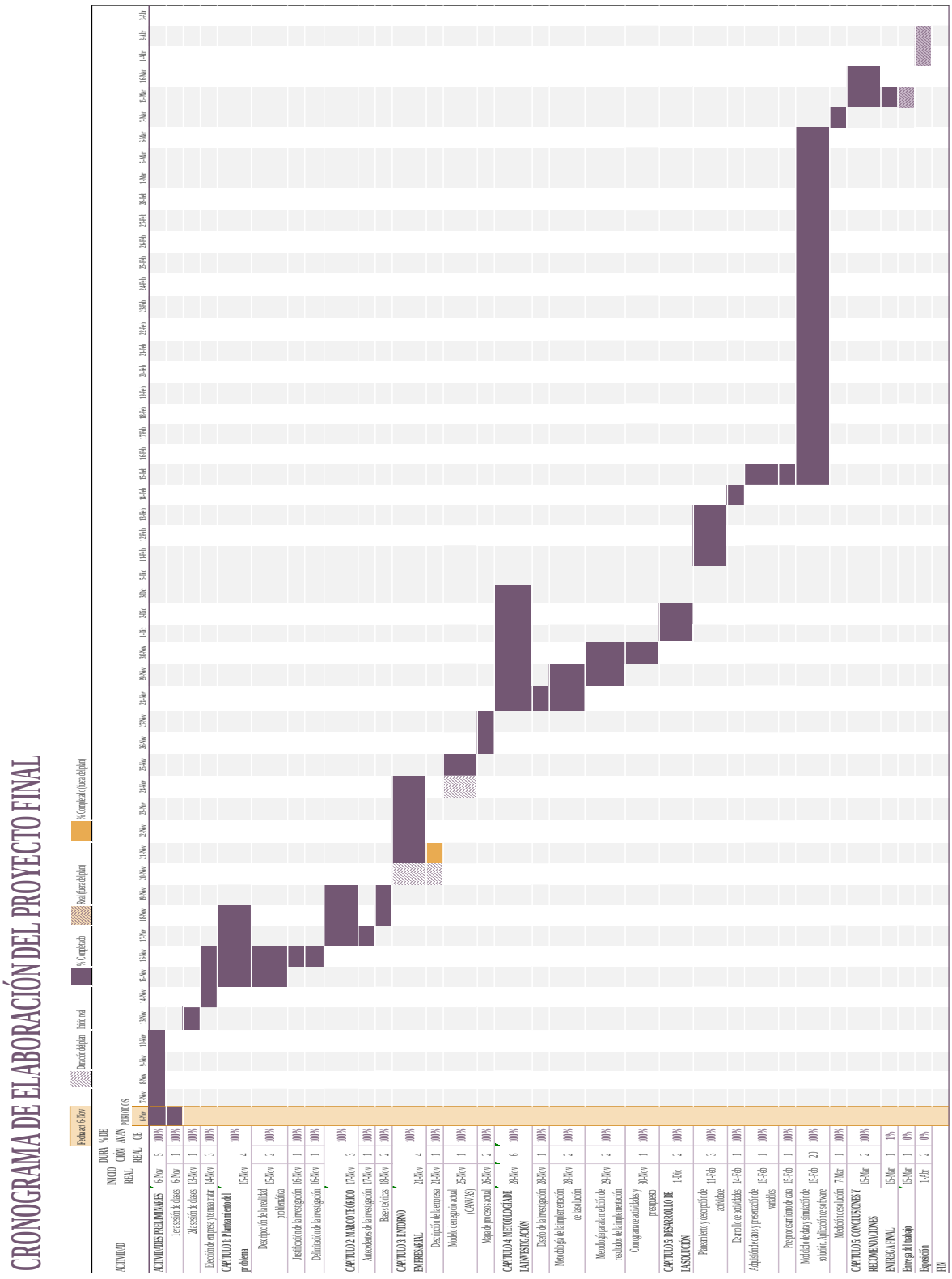

Figura 28: *Cronograma de actividades*

Fuente: *Elaboración propia (2023)*

Item	Cantidad	Importe por unidad (soles)	Mes	Total (soles)
Maquinaria y equipos				
Laptops	4	S/ 240.00	2	S/ 1,920.00
Mouse	4	S/ 100.00	2	S/ 800.00
EPP´s	4	S/ 103.50	2	S/ 828.00
Sub total				**S/ 3,548.00**
Servicios Complementarios				
Usuario WMS	4	S/ 280.00	2	S/ 2,240.00
Oficce 365	1	S/ 48.00	2	S/ 96.00
Utiles de oficina	4	S/ 57.00	2	S/ 456.00
Sub total				**S/ 2,792.00**
Mano de obra				
Bachilleres	4	S/ 2,135	2	S/ 17,083.36
Sub total				**S/ 17,083.36**
Servicios Básicos				
Luz	4	S/ 50.00	2	S/ 400.00
Internet	4	S/ 50.00	2	S/ 400.00
Sub total				**S/ 800.00**
Total				**S/ 24,223.36**

Tabla 2: *Presupuesto*
Fuente: *Elaboración propia (2023)*

CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

5.1 Propuesta solución.

5.1.1 Planteamiento y descripción de Actividades

En este trabajo de investigación se plantea iniciar las etapas para la ejecución de machine learning aprendizaje no supervisado, mediante la interfaz de "Jupyter" y el uso del lenguaje python, apoyándonos en las evaluaciones de los métodos de K-Means, K-Medoids y Jerárquico.

a) Adquisición de datos: Se recopila la información que se usará para el modelo de clustering.

b) Presentación de variables: Se definen las variables a utilizar para el análisis.

c) Proceso del Preprocesamiento de datos: La base de datos recopilados pasarán por una etapa de limpieza de información, que consiste en eliminar valores que alteren los resultados finales.

Etapa de Modelado: Se aplicará el método del codo para determinar el K idóneo, para luego aplicar las técnicas K-Means, K-Medoids y Jerárquico, para analizar las segmentaciones.

d) Simulación y análisis de resultados: Se simulan los resultados y se analizan los resultados obtenidos.

5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución.

5.1.2.1. Adquisición de datos

La base de datos fue recopilada mediante la plataforma Data Ware House (DWH) de Falabella, esta fue descargada y almacenada en un documento de excel para su posterior uso. La data cuenta con el número total de transacciones desde el mes de noviembre 2022 a enero 2023 de la categoría de Electro en Falabella Retail.

5.1.2.2. Presentación de variables

Con la base de datos definida se ha obtenido un total de 327,920 transacciones a nivel nacional, en donde presentamos 22 variables que se definen de la siguiente manera:

N°	ATRIBUTO	DESCRIPCIÓN
1	Local	Código del local de compra
2	LocalDESC	Descripción del local de compra
3	Mes	Mes de compra
4	Dia Comer	Fecha de compra
5	POS	Código de secuencia de caja
6	Sexo	Género del cliente
7	SubLinea	Código de línea del producto
8	SubLneaDESC	Descripción de línea del producto
9	Clase	Código de línea de la subclase del producto
10	ClaseDESC	Descripción de la clase del producto
11	SubClase	Código de subclase del producto
12	SubClaseDESC	Descripción de la subclase del producto
13	Modelo	Nombre del modelo del producto
14	Marca	Marca del producto
15	SKU	Código único del producto
16	SKUDESC	Descripción del producto
17	Proveedor	Nombre del proveedor del producto
18	Tipo de Producto	Cósigo del producto
19	TipoProductoDESC	Tipo del producto o servicio
20	Transaccion	Código de medio de pago
21	TransaccionDESC	Descripción del medio de pago
22	VentaUnidades	Total de unidades vendidas

Tabla 3: *Variables de estudio*
Fuente: *Elaboración propia (2023)*

	Local	LocalDESC	Mes	Dia Comer	POS	Sexo	SubLinea	SubLineaDESC	Clase	ClaseDESC	...	Modelo	Marca	
0	101	San Isidro	Noviembre 2022	2022-11-01	10	Hombre	J1102	AUDIO	J110213	AUDIFONOS	...	S2DMW-P740	SKULLCANDY	
1	101	San Isidro	Noviembre 2022	2022-11-01	30	Mujer	J1109	ELECTRODOMESTICOS	J110902	CUIDADO PERSONAL	...	SG-3049C02	SIEGEN	
2	101	San Isidro	Noviembre 2022	2022-11-01	34	Mujer	J1109	ELECTRODOMESTICOS	J110902	CUIDADO PERSONAL	...	RVDR5222LA2	REVLON	
3	101	San Isidro	Noviembre 2022	2022-11-01	35	Hombre	J1102	AUDIO	J110214	PARLANTE PORTATIL	...	DD-BREEZEK	DDESIGN	8
4	101	San Isidro	Noviembre 2022	2022-11-01	35	Mujer	J1105	TELEFONIA	J110503	EQUIPOS	...	CPPG 1001	CLARO	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
327915	420	SF Huanuco	Enero 2023	2023-01-31	44	Mujer	J1109	ELECTRODOMESTICOS	J110901	ELECT.MENORES	...	RRC-9	RECCO	8
327916	420	SF Huanuco	Enero 2023	2023-01-31	44	Mujer	J1109	ELECTRODOMESTICOS	J110901	ELECT.MENORES	...	RSJ-1052	RECCO	8
327917	420	SF Huanuco	Enero 2023	2023-01-31	52	Mujer	J1105	TELEFONIA	J110503	EQUIPOS	...	CPPG 1004	CLARO	
327918	420	SF Huanuco	Enero 2023	2023-01-31	52	Mujer	J1105	TELEFONIA	J110503	EQUIPOS	...	CPPG 1007	CLARO	
327919	420	SF Huanuco	Enero 2023	2023-01-31	52	Mujer	J1105	TELEFONIA	J110503	EQUIPOS	...	CPPG 1462	CLARO	

327920 rows × 22 columns

Figura 29: *Base de datos Falabella en Python*
Fuente: *Elaboración propia (2023)*

5.1.2.3. Proceso del Preprocesamiento de datos

Para esta primera etapa del preprocesamiento de datos, hemos seleccionado las variables que usaremos para este trabajo de investigación, de las 22 columnas, tenemos variables que no explican el comportamiento de compra del cliente por lo que no influye al modelo a aplicar, estas son: “LocalDESC”, “Dia Comer”, “POS”, “Sublínea”, “Clase”, ”Modelo”, “SKU”, “SKUDESC”, “Proveedor”, , “Tipo de Producto”, “TipoProductoDESC”, “Transaccion”.

Las variables que sí están relacionadas al comportamiento de compra del consumidor son: “Local”, “Mes”, “Sexo”,” SublineaDESC”, ”Marca”,”TransaccionDESC”, “VentaUnidades”.

Por lo que filtraremos estas columnas para el análisis, seguidamente usaremos el código “Drop” para eliminar la columna Mes, ya que se segmentará usando toda la base de datos desde noviembre a enero.

```
In [19]: 1 BDfala = falabella[['Local','Mes','Sexo','SubLineaDESC','Marca','TransaccionDESC','VentaUnidades']]

In [20]: 1 BD_drop= BDfala.drop(columns =['Mes'])
         2 BD_drop
```

Figura 30: *Reducción de variables*
Fuente: *Elaboración propia (2023)*

	Local	Sexo	SubLineaDESC	Marca	TransaccionDESC	VentaUnidades
0	101	Hombre	AUDIO	SKULLCANDY	Venta Credito Visa	1
1	101	Mujer	ELECTRODOMESTICOS	SIEGEN	Venta CMR Debito	1
2	101	Mujer	ELECTRODOMESTICOS	REVLON	Venta CMR Debito	1
3	101	Hombre	AUDIO	DDESIGN	Venta Credito Visa	2
4	101	Mujer	TELEFONIA	CLARO	Venta Contado	1
...	...	...	...	...	...	...
327915	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Contado	1
327916	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Credito Visa	2
327917	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1
327918	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1
327919	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1

327920 rows × 6 columns

Figura 31: *Selección de Variables*
Fuente: *Elaboración propia (2023)*

Ahora analizaremos si los datos de cada variable presenta ruido, como por ejemplo valores vacíos o nulos que pueden alterar el resultado del análisis, mediante el código "BD_drop.isnull().sum()" podemos ver qué atributos presentan valores nulos o vacíos. En este trabajo no se presentaron valores nulos.

```
In [29]: 1 BD_drop.isnull().sum()

Out[29]: Local             0
         Sexo              0
         SubLneaDESC       0
         Marca             0
         TransaccionDESC   0
         VentaUnidades     0
         dtype: int64
```

Figura 32: *Exclusión de valores nulos*
Fuente: *Elaboración propia (2023)*

Ya filtrando la data que se usará, tendremos que transformar las siguientes variables categóricas “Local, Sexo, SubLineaDESC, Marca, TransaccionDESC” a variables numéricas, en este caso usaremos la herramienta LabelEncoder.

```
In [26]:
1 from sklearn.preprocessing import LabelEncoder
2 cols = ['Local','Sexo','SubLineaDESC','Marca','TransaccionDESC']
3 BD_drop[cols] = BD_drop[cols].apply(LabelEncoder().fit_transform)
```

Figura 33: *Código para convertir variables categóricas*
Fuente: *Elaboración propia (2023)*

Luego de ejecutarlo nuestra nueva tabla denominada BD_drop presenta la siguiente información numérica:

	Local	Sexo	SubLineaDESC	Marca	TransaccionDESC	VentaUnidades
0	0	0	1	265	5	1
1	0	1	5	262	3	1
2	0	1	5	248	3	1
3	0	0	1	69	5	2
4	0	1	10	56	4	1
...	...	...	...	...	...	...
327915	29	1	5	243	4	1
327916	29	1	5	243	5	2
327917	29	1	10	56	5	1
327918	29	1	10	56	5	1
327919	29	1	10	56	5	1

327920 rows × 6 columns

Figura 34: *Reajuste de variables categóricas a numéricas*
Fuente: *Elaboración propia (2023)*

Con la data cuantificada aún no podemos realizar la metodología del K-Means, lo que necesitamos es estandarizar la información que tenemos. Con la función “describe()” podemos ver cómo es la distribución de la base de datos, se puede observar que los rangos de

cada fila son muy dispersos, por lo que la data no presenta una estandarización. Para este caso se aplica la función StandardScaler()

	Local	Sexo	SubLineaDESC	Marca	TransaccionDESC	VentaUnidades
count	327920.000000	327920.000000	327920.000000	327920.000000	327920.000000	327920.000000
mean	17.149500	0.519419	5.384847	189.799329	3.680678	1.372841
std	12.313432	0.499624	3.552847	87.006667	1.222346	3.358717
min	0.000000	0.000000	0.000000	0.000000	0.000000	-45.000000
25%	7.000000	0.000000	3.000000	124.000000	3.000000	1.000000
50%	15.000000	1.000000	5.000000	218.000000	4.000000	1.000000
75%	25.000000	1.000000	10.000000	257.000000	5.000000	1.000000
max	39.000000	1.000000	12.000000	339.000000	7.000000	424.000000

Figura 35: *Estandarización de datos*
Fuente: *Elaboración propia (2023)*

```
1 scaler= StandardScaler()
2 segmentation_std = scaler.fit_transform(BD_drop)
```

```
1 segmentation_std
```

```
array([[-1.39274947, -1.03962315, -1.23418034,  0.86431061,  1.07933783,
        -0.11100712],
       [-1.39274947,  0.96188701, -0.10832088,  0.82983044, -0.55686319,
        -0.11100712],
       [-1.39274947,  0.96188701, -0.10832088,  0.66892299, -0.55686319,
        -0.11100712],
       ...,
       [ 0.96240578,  0.96188701,  1.29900344, -1.53780782,  1.07933783,
        -0.11100712],
       [ 0.96240578,  0.96188701,  1.29900344, -1.53780782,  1.07933783,
        -0.11100712],
       [ 0.96240578,  0.96188701,  1.29900344, -1.53780782,  1.07933783,
        -0.11100712]])
```

Figura 36: *Variables normalizadas*
Fuente: *Elaboración propia (2023)*

Con la data estandarizada procedemos a realizar el Análisis de Componentes Principales (PCA), que consiste en analizar el número de componentes a definir y que estos explican el 80% de la base de datos. La función que se usa es “from sklearn.preprocessing import StandardScaler”

Con la función “import matplotlib.pyplot as plt”, podremos graficar los datos obtenidos de la varianzas (eje Y) relacionado con los componentes que incluyen (eje X).

```
pca=PCA()
pca.fit(segmentation_std)
```

```
PCA()
```

```
plt.figure(figsize = (10,8))
plt.plot(range(1,7), pca.explained_variance_ratio_.cumsum(), marker='o', linestyle = '--')
plt.title('Varianza explicada por componentes')
plt.xlabel('Numero de componentes')
plt.ylabel('Varianza explicada acumulada')
```

```
Text(0, 0.5, 'Varianza explicada acumulada')
```

Figura 37: *Aplicación de PCA*
Fuente: *Elaboración propia (2023)*

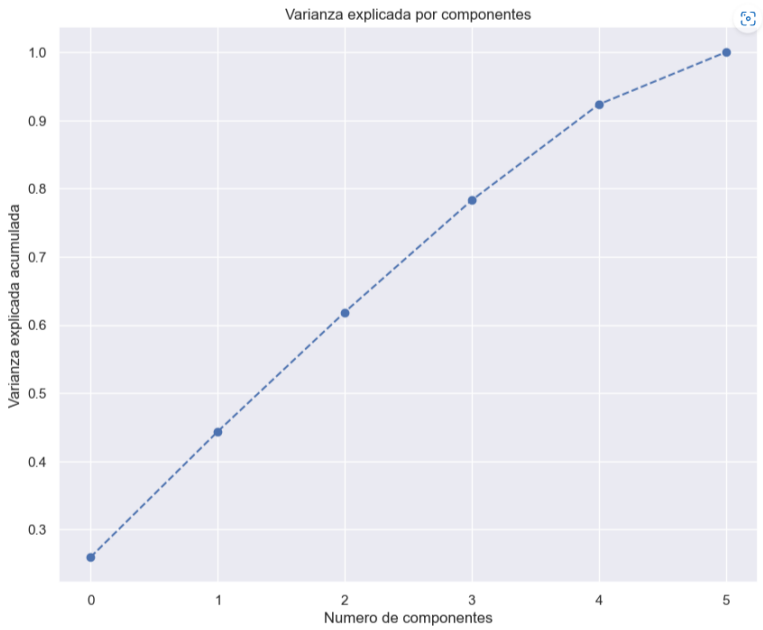

Figura 38: *Varianza datos*
Fuente: *Elaboración propia (2023)*

Del gráfico PCA podemos observar que a partir de la cantidad de tres componentes la cantidad de datos analizados son más representativo logrando tener 78% de la información, para el desarrollo de este trabajo se han considerado 4 componentes que nos daría mayor exactitud de análisis, para lo cual los hemos nombrado “Componente_1”, “Componente_2”, “Componente_3”, “Componente_4”.

```
PCA(n_components=4)
```

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

```
0.7801243564468319
```

```
from sklearn.decomposition import PCA

pca = PCA(n_components=4)
pca_fala = pca.fit_transform(BD_drop)
pca_fala_df= pd.DataFrame( data = pca_fala, columns = ['Componente_1' , 'Componente_2','Componente_3','Componente_4'])
pca_fala_df
```

	Componente_1	Componente_2	Componente_3	Componente_4
0	-75.269021	-16.923577	-4.172679	0.957649
1	-72.257840	-16.851027	-0.223485	0.425872
2	-58.257974	-16.897041	-0.263040	0.432093
3	120.729201	-17.524050	-4.587826	2.034040
4	133.754346	-17.607151	4.154266	-0.186470
...	...	...	...	...
327915	-53.162347	11.967002	-0.275642	-0.875252
327916	-53.162327	11.958750	-0.128209	0.101875
327917	133.850020	11.273325	4.155791	-1.496036
327918	133.850020	11.273325	4.155791	-1.496036
327919	133.850020	11.273325	4.155791	-1.496036

327920 rows × 4 columns

Figura 39: *Etiquetado de componentes*
Fuente: *Elaboración propia (2023)*

5.1.2.4. Etapa de Modelado

Esta etapa del desarrollo de la investigación, modelaremos 3 distintos métodos para la obtención del clúster, estos son K-Means. Clúster Jerárquico y K-Medoids

- Modelo K-Means

Para este proceso de modelado nos basaremos en la aplicación del K-Means, en la cual consiste en realizar varias iteraciones del valor i dentro de un rango limitado (1, 11), donde mediante el gráfico podemos ver el K óptimo según el método del codo, el cual consiste en la medición de distancias entre centroides. En el gráfico se puede interpretar que mientras mayor es el número de clúster “K”, la varianza entre clúster tiende a disminuir, el clúster idóneo para nuestro trabajo de investigación es el K=5.

```
In [59]:
distorsion = []
from sklearn.cluster import KMeans
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, max_iter= 300)
    kmeans.fit(segmentation_std)
    distorsion.append(kmeans.inertia_)
```

```
In [60]:
import matplotlib.pyplot as plt
plt.plot(range(1, 11), distorsion)
plt.title("Método del Codo - Encontrar el K óptimo")
plt.xlabel('Números de Clusters')
plt.ylabel('distorsion')
plt.show()
```

Figura 40: *Aplicación de K-Means*
Fuente: *Elaboración propia (2023)*

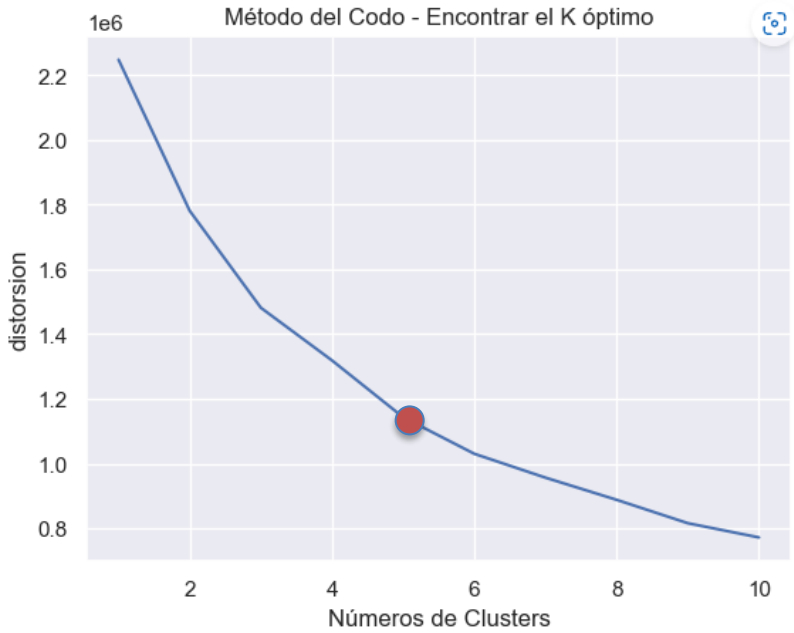

Figura 41: *Aplicación método del codo*
Fuente: *Elaboración propia (2023)*

Finalmente consolidamos los resultados obtenidos en la tabla inicial para evaluar los resultados, según lo mostrado en la siguiente figura

```
BD_copy['Cluster_KMeans'] = clusterfala.labels_
BD_copy
```

	Local	Sexo	SubLineaDESC	Marca	TransaccionDESC	VentaUnidades	Cluster_KMeans
0	101	Hombre	AUDIO	SKULLCANDY	Venta Credito Visa	1	3
1	101	Mujer	ELECTRODOMESTICOS	SIEGEN	Venta CMR Debito	1	2
2	101	Mujer	ELECTRODOMESTICOS	REVLON	Venta CMR Debito	1	2
3	101	Hombre	AUDIO	DDESIGN	Venta Credito Visa	2	3
4	101	Mujer	TELEFONIA	CLARO	Venta Contado	1	0
...	...	...	...	...	...	...	...
327915	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Contado	1	2
327916	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Credito Visa	2	2
327917	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0
327918	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0
327919	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0

327920 rows × 7 columns

Figura 42: *Resultados por cluster*
Fuente: *Elaboración propia (2023)*

- Clúster Jerárquico- Aglomerativo

Para el desarrollo de este clúster se está considerando la data estandarizada nombrada "segmentation_std", con esta información necesitamos realizar un gráfico conocido como Dendrograma en donde cada agrupamiento de datos es representado por una línea horizontal; desde el punto de la coordenada y la línea horizontal se define como un conjunto de datos agrupados en clúster que tienen similitudes.

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

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

Figura 43: *Código de dendrograma-python*
Fuente: *Elaboración propia (2023)*

Al tener la gráfica generada con 1000 iteraciones, se necesitará evaluar el corte horizontal, cada punto intersecado en el gráfico representa un clúster a usar, para este trabajo de investigación se ha considerado el punto Y=6 ya que las separaciones de los datos verticales son más pronunciadas que el resto de las ramas.

```
dendrogram = sch.dendrogram(sch.linkage(segmentation_std[:1000,:], method='ward', metric= 'euclidean'))
plt.axhline(y = 19, color = 'r', linestyle = '-')
```

```
<matplotlib.lines.Line2D at 0x153e6e235e0>
```

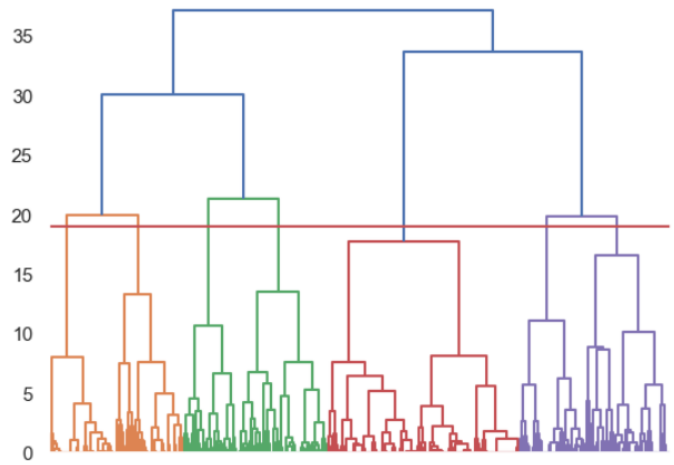

Figura 44: *Gráfico dendrograma*
Fuente: *Elaboración propia (2023)*

Se puede observar que al realizar el corte en y=19, los puntos encontrados con de K= 7, esto quiere decir que el número de clúster a usar por el método del clúster jerárquico es de 7 segmentaciones.

- Modelo K-Medoids

Para este método se usan los puntos de datos elegidos como centros conglomerados, es decir busca minimizar las diferencias entre los puntos de un grupo y los puntos considerados como centros de ese mismo grupo, para este proceso se ha realizo un rango de iteraciones K que va del 2 al 8. Se hará uso de la librería “sklearn_extra.cluster”:

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

```
segmentation_std[:500,:].shape
```

```
(500, 6)
```

```
valores=[]
for k in range(2,8):
    BD_medoids = KMedoids(n_clusters=k)
    BD_medoids.fit(segmentation_std[:500,:])
    valores.append(BD_medoids.inertia_)
```

Figura 45: *Código K-medoids*
Fuente: *Elaboración propia (2023)*

Elaboramos el gráfico del método del codo para determinar el número de clústeres que recomienda la simulación.

```
from matplotlib import pyplot
pyplot.plot(range(2,8),valores)
pyplot.title("Método del Codo")
pyplot.xlabel("Número de clusters")
pyplot.ylabel("Rango de Valores")
pyplot.show()
```

Figura 46: *Método del codo K-medoids*
Fuente: *Elaboración propia (2023)*

El número de clúster que recomienda el método de clúster K-Medoids es el de K = 5

5.1.2.5. Análisis de resultados

Comparación de Métodos Clustering

Para poder determinar el modelo óptimo debemos evaluar cada agrupación de clúster de los modelos de K-Means y K-Medoids, para el caso del modelo Clúster Jerárquico, no se podrá incluir en el análisis de la base de datos original ya que es demasiado pesado para la simulación por la cantidad de datos que se requiere.

	Local	Sexo	SubLineaDESC	Marca	TransaccionDESC	VentaUnidades	Cluster_KMeans	Cluster Kmedoids
0	101	Hombre	AUDIO	SKULLCANDY	Venta Credito Visa	1	3	0
1	101	Mujer	ELECTRODOMESTICOS	SIEGEN	Venta CMR Debito	1	2	2
2	101	Mujer	ELECTRODOMESTICOS	REVLON	Venta CMR Debito	1	2	2
3	101	Hombre	AUDIO	DDESIGN	Venta Credito Visa	2	3	1
4	101	Mujer	TELEFONIA	CLARO	Venta Contado	1	0	1
...	...	...	...	...	...	...	...	...
327915	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Contado	1	2	3
327916	420	Mujer	ELECTRODOMESTICOS	RECCO	Venta Credito Visa	2	2	4
327917	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0	1
327918	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0	1
327919	420	Mujer	TELEFONIA	CLARO	Venta Credito Visa	1	0	1

327920 rows × 8 columns

Figura 47: *Agrupación por Clústeres (K Means/K Medoids)*
Fuente: *Elaboración propia (2023)*

- Resultado y análisis Clúster K-Means:

	Cluster K-Means				
	0	**1**	**2**	**3**	**4**
Sexo	Hombres y mujeres	Hombres y mujeres	Mujeres	Hombres	Hombres en su mayoria
Local	Lima Norte,Chiclayo Mall y Mall de Sur	Internet	Jockey Plaza, San Miguel	Jockey Plaza, San Miguel	Internet
Método de Pago	Venta Credito Visa, Venta Contado y CMR Debido	Venta CMR Crédito	Venta Credito Visa	Venta Credito Visa y Venta CMR Débito	Venta CMR Crédito
Categoría	Telefonía y Video	Electrodomésticos	Electrodomésticos y Accesorios	Electrodomésticos y Audio	Audio y Video
Marca	Movistar, Claro, Entel, Samsung	Samsung y Oster	Recco y Oster	Skullcandy, Recco y Ddesign	Samsung y Skullcandy
Venta de Unidades	19.47%	31.32%	22.54%	21.14%	5.53%

Tabla 4: *Resultados K-Means*
Fuente: *Elaboración propia (2023)*

- Análisis de atributos K-Means:
  - Atributo “Sexo”:

SEXO	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total general
Hombre	38,116	30,984		88,288	204	157,592
Mujer	41,014	32,622	96,556		136	170,328
Total general	79,130	63,606	96,556	88,288	340	327,920

Tabla 5: *Resultados Clúster Sexo K-Means*
Fuente: *Elaboración propia (2023)*

Se está segmentando las variables que representen mayor porcentaje, a continuación, se detallará el peso por clúster:

- Clúster 0: los atributos son de hombre y mujer ya que representan el 51% y 48% de los datos en esta segmentación.
- Clúster 1: el peso para ambos atributos es muy cercano, por lo cual se están considerando ambos, mujer el 51% y hombres 48%.
- Clúster 2: esta segmentación solo ha agrupado al grupo de mujeres.
- Clúster 3: esta segmentación solo ha agrupado al grupo de hombres.
- Clúster 4: el atributo hombre es el más representativo con un 60% de los datos.

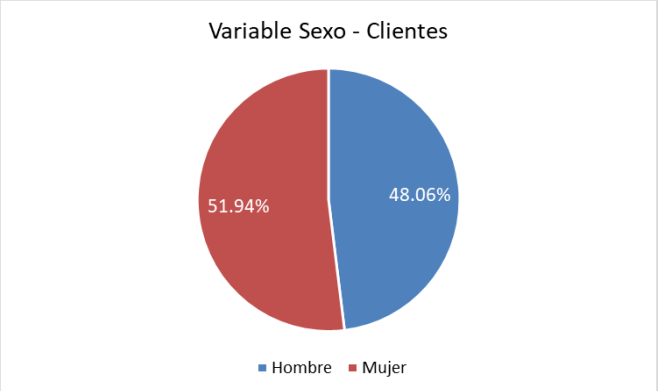

Figura 48: *Resultado Segmentación Sexo (K Means)*
Fuente: *Elaboración propia (2023)*

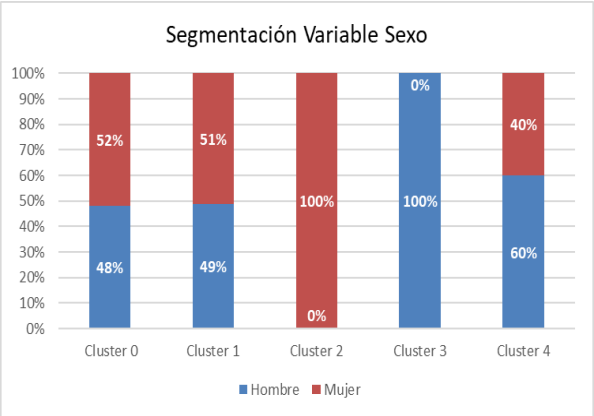

Figura 49: *Resultado Segmentación Sexo por Clúster (K Means)*
Fuente: *Elaboración propia (2023)*

Según lo mostrado en la Figura 48, el total de datos de la variable Sexo tiene la proporción de 48.06% hombres y 51.94% mujeres, esta primera variable brindará un primer indicio del tipo de cliente por segmento.

- Atributo “Local”:

LOCAL_ID	LOCAL_DESC	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
101	San Isidro	1,337	4	2,915	2,808	1
102	Jockey Plaza	3,914	30	8,185	7,259	1
106	Miraflores	1,045	36	4,292	3,823	
201	Lima Centro	1,390	14	1,000	1,101	
202	San Miguel	3,856	82	6,985	6,370	
204	Megaplaza	4,672	75	5,391	4,711	2
205	BELLAVISTA	1,774	41	2,382	2,198	
206	Atocongo	3,792	78	3,512	3,718	5
207	Angamos	1,961	66	3,440	3,313	1
208	Santa Anita	2,537	47	2,355	2,365	
209	Lima Norte	6,089	202	5,464	4,964	
210	Salaverry	1,179	221	3,665	3,356	
211	Centro Civico	3,612	165	3,578	3,928	
212	Mall del Sur	5,479	236	4,658	4,770	
213	Puruchuco	1,458	138	2,384	2,422	
214	COMAS	2,066	170	3,083	2,587	
303	Piura Centro	1,875	100	1,859	1,590	
304	Arequipa Cayma	2,515	140	3,003	2,378	
307	Chiclayo Mall	5,554	233	5,223	4,445	
309	Chimbote	2,391	85	1,866	1,689	
312	Trujillo Mall	3,296	234	3,896	3,484	
320	Piura Mall	2,334	222	3,461	2,870	
321	Arequipa Porongoche	2,066	158	3,107	2,721	
322	Cajamarca Quinde	2,055	130	1,570	1,423	
323	Ica Mall	1,715	124	1,938	1,740	
324	Huancayo	2,101	117	1,663	1,518	
325	Cusco	1,399	245	1,868	1,682	
400	SF Iquitos	1,182	137	838	622	
419	SF Pucallpa	858	228	628	538	
420	SF Huanuco	1,559	350	810	587	
421	Canete	1,089	543	479	469	
455	Express Huaraz	393	105	327	207	
456	Express Tacna	336	167	274	235	
621	Express San Isidro	2	4	53	55	
625	Express Mall del Sur			28	25	
662	Express Delivery Bellavista		9	13	19	
663	Express Delivery Plaza Norte		10	12	11	
664	Express Delivery Comas	1	11	1	6	
800	INTERNET	65	58,201	346	260	328
801	FONOCOMPRAS	183	448	4	21	2
	Total general	79130	63606	96556	88288	340

Tabla 6: *Resultados Clúster Sexo K-Means*
Fuente: *Elaboración propia (2023)*

Para este análisis se han seleccionado los locales que tienen el mayor número de transacciones, esto será relevante para determinar el comportamiento del cliente por cluster.

- Clúster 0: Las tiendas más representativas son Lima Norte, Mall de Sur y Chiclayo Mall.
- Clúster 1: La tienda más representativa es el canal de Internet.
- Clúster 2: Las tiendas más representativas son Jockey Plaza y San Miguel.
- Clúster 3: Las tiendas más representativas son Jockey Plaza y San Miguel.
- Clúster 4: La tienda más representativa es el canal de Internet.

- Atributo “Método de Pago”:

Transacción	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Fpay	117	1,634	40	100	
Venta American Express	868	1,821	1,040	1,060	
Venta CMR Credito	566	57,618	1,222	1,080	328
Venta CMR Debito	23,314	2,054	28,317	26438	4
Venta con solo cheque			3	3	3
Venta Contado	25,322	479	20,972	18,379	2
Venta Credito Visa	27,972		43,555	39,795	3
Venta Debito Visa	971		1,407	1,433	
Total general	79,130	63,606	96,556	88,288	340

Tabla 7: *Resultados Clúster Método de Pago K-Means*
Fuente: *Elaboración propia (2023)*

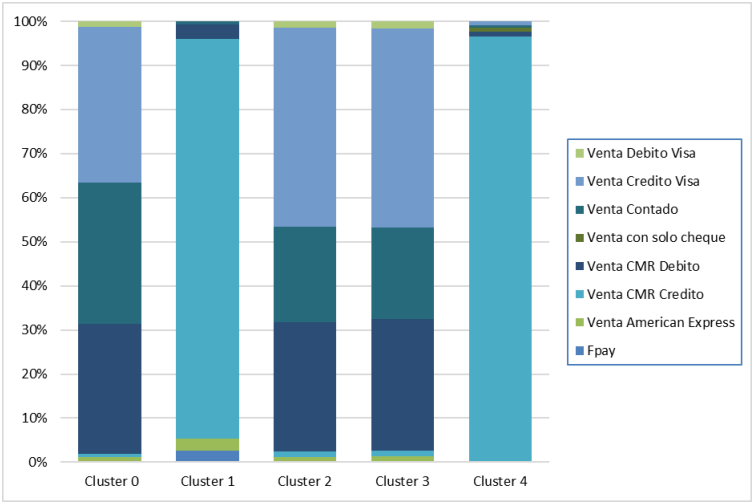

Figura 50: *Resultado Segmentación Transacción por Clúster (K-Means)*
Fuente: *Elaboración propia (2023)*

Para este análisis se han seleccionado los medios de pago que tienen el mayor número de transacciones, como se puede observar en la figura, hay ciertos tipos de transacciones que son más recurrentes al realizar la compra.

- Clúster 0: Los clientes de este clúster suelen usar los medios de pago de Venta Crédito Visa, Venta Contado y Venta CMR Debido.
- Clúster 1: Los clientes de este grupo recurren al método de pago Venta CMR Débito.
- Clúster 2: Los clientes de este grupo recurren al método de pago Venta Crédito Visa.
- Clúster 3: Los clientes de este grupo recurren al método de pago Venta Crédito Visa y Venta CMR Débito
- Clúster 4: Los clientes optan por usar el medio de pago CMR Crédito.

- Atributo “Sublínea”:

SublineaDESC	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ACCESORIOS		7,409	11,018	15,088	20
AUDIO		7,018	7,904	24,276	107
CLIMATIZACION		1,368	2,074	2,297	
COCINA		3,772	2,683	1,373	3
COMPUTACION Y HOGAR		5,736	4,976	10,267	22
ELECTRODOMESTICOS	28	22,361	63,167	31,645	8
FOTOGRAFIA	83	505	41	75	
LAVADO	83	2,860	2,306	2,204	7
REFRIGERACION	2,852	3,894	1,042	1,063	7
SERVICIOS E INSTALACIONES		27			
TELEFONIA	51,706	3,820	1,264		73
VIDEO	22,853	4,336	81		93
VIDEO JUEGOS	1,525	500			
Total general	79,130	63,606	96,556	88,288	340

Tabla 8: *Resultados clúster Sublínea K-Means*
Fuente: *Elaboración propia (2023)*

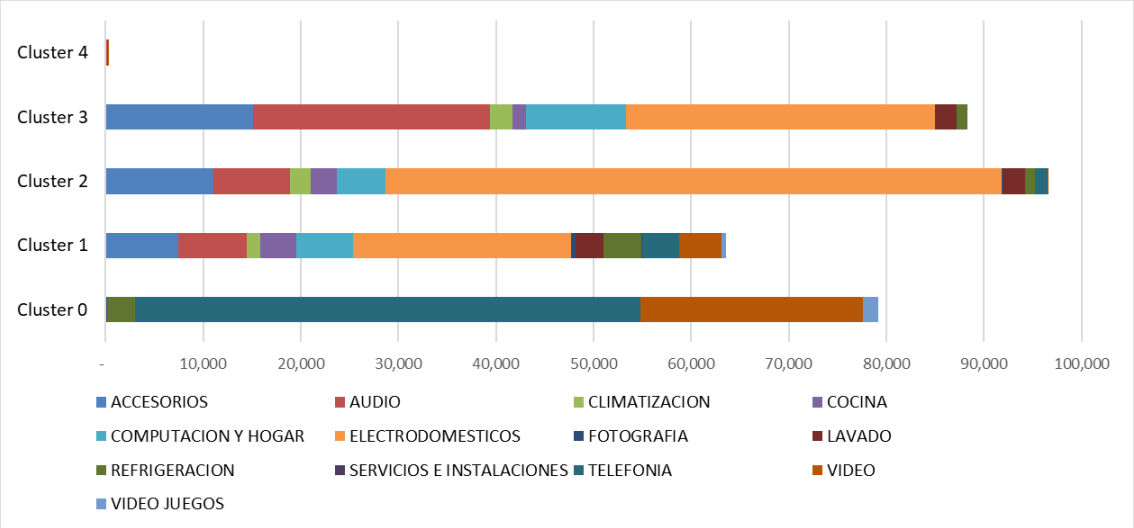

Figura 51: *Resultado Segmentación Transacción por SublíneaDESC (K-Means)*
Fuente: *Elaboración propia (2023)*

De la figura 51 se puede observar el número de transacciones que hay por sublínea o categorías de sector electro, en esta se ha seleccionado las que han generado mayor volumen de venta por cada clúster.

- Clúster 0: Los clientes optan por compras en la categoría de teléfonos y video.
- Clúster 1: Los clientes de este grupo realizan muchas compras de electrodomésticos.
- Clúster 2: Los clientes de este grupo suelen comprar accesorios y electrodomésticos.
- Clúster 3: Los clientes de este grupo compran electrodomésticos y productos de audio.
- Clúster 4: Los clientes optan por comprar productos de audio y video.

- Atributo “Marca”:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Marca	Movistar, Claro, Entel, Samsung	Samsung y Oster	Recco y Oster	Skullcandy, Recco y Ddesign	Samsung y Skullcandy

Tabla 9: *Resultados clúster por Marca K-Means*
Fuente: *Elaboración propia (2023)*

De la tabla se puede observar las marcas que mayor transacción de compra han realizado los clientes, a continuación, se describe cada clúster y marca seleccionada:

- Clúster 0: Las compras realizadas fueron realizadas en las marcas Movistar, Claro, Entel y Samsung.
- Clúster 1: Los clientes de este grupo prefieren productos de la marca Samsung y Oster.
- Clúster 2: Las marcas más compradas son Recco y Oster.
- Clúster 3: Los clientes de este grupo compran productos de las marcas Skullcandy, Recco y Ddesign.
- Clúster 4: Los clientes compran las marcas Samsung y Skullcandy.

- Atributo “Unidades Vendidas”:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Venta de Unidades	19.47%	31.32%	22.54%	21.14%	5.53%

Tabla 10: *Resultados clúster por Unidades Vendidas K-Means*
Fuente: *Elaboración propia (2023)*

En la tabla se ha considerado los pesos de las unidades vendidas de cada clúster de la base de datos recolectado

- Clúster 0: Representa el 19.47% del total de las unidades vendidas de todas las transacciones.
- Clúster 1: Representa el 31.32% del total de las unidades vendidas de todas las transacciones, este clúster es el más peso en volumen de ventas.
- Clúster 2: Representa el 22.54% del total de las unidades vendidas de todas las transacciones.
- Clúster 3: Representa el 21.14% del total de las unidades vendidas de todas las transacciones.
- Clúster 4: Este clúster representa el menor volumen de ventas con 5.53% comparado con el resto.

- Resultado y análisis Clúster K-Medoids

	Cluster K-Medoids				
	0	1	2	3	4
Sexo	Hombres	Hombres y mujeres	Mujeres en su mayoría	Mujeres	Mujeres
Local	Internet	Mall del Sur, Jockey Plaza, San Miguel, Chiclayo Mall y Lima Norte	Internet	Internet	Jockey Plaza, Miraflores, Lima Norte y San Miguel
Método de Pago	Venta Credito Visa, Venta Contado, Venta CMR Débito y Venta CMR Crédito	Venta Credito Visa	Venta CMR Debito y Venta CMR Crédito	Venta Contado y Venta CMR Debito	Venta Credito Visa
Categoría	Electrodomésticos, Audio y Telefonía	Telefonía	Electrodomésticos	Electrodomésticos y Telefonía	Electrodomésticos
Marca	Samsung, Sullcandy y Recco	Claro	Recco, Oster y Samsung	Movistar y Recco	Recco, Oster y Wurden
Venta de Unidades	46.20%	10.71%	21,43%	12,56%	9.11%

Tabla 11: *Resultados K-Medoids*
Fuente: *Elaboración propia (2023)*

- Análisis de atributos K-Medoids:
  - Atributo “Sexo”:

SEXO	Clúster 0	Clúster 1	Clúster 2	Cluster 3	Cluster 4	Total general
Hombre	144,962	11,791	839			157,592
Mujer		31,314	56,339	44,384	38,291	170,328
Total general	**144,962**	**43,105**	**57,178**	**44,384**	**38,291**	**327,920**

Tabla 12: *Resultados Clúster Sexo K-Medoids*
Fuente: *Elaboración propia (2023)*

Se está segmentando las variables que representen mayor porcentaje, a continuación, se detallará el peso por clúster:

- Clúster 0: esta segmentación solo ha agrupado al grupo de hombres.
- Clúster 1: el peso para ambos atributos es considerable, por lo cual se están seleccionados ambos, mujer el 73% y hombres 27%.
- Clúster 2: el atributo hombre representa el 99% de los datos en esta segmentación.
- Clúster 3: esta segmentación solo ha agrupado al grupo de mujeres.
- Clúster 4: esta segmentación solo ha agrupado al grupo de mujeres.

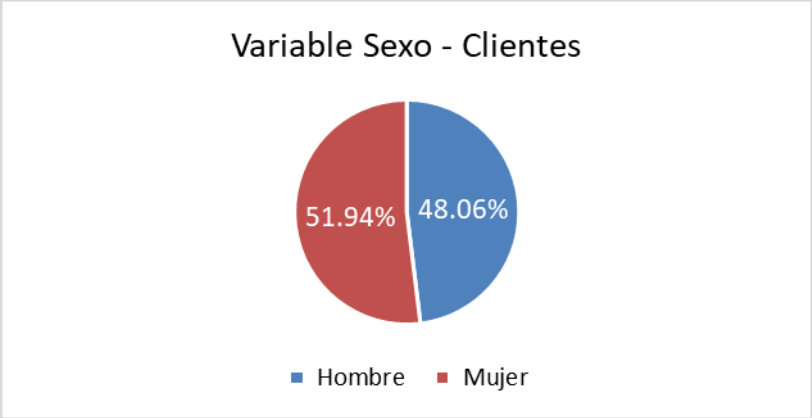

Figura 52: *Resultado Segmentación Sexo (K-Medoids)*
Fuente: *Elaboración propia (2023)*

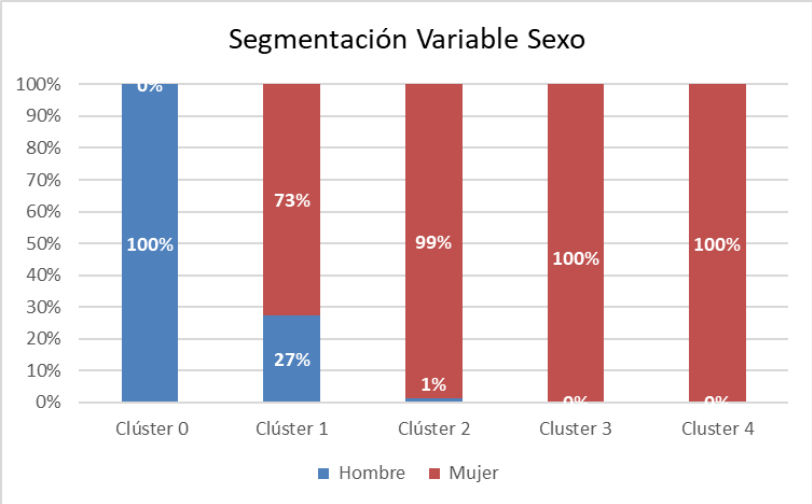

Figura 53: *Resultado Segmentación Sexo por Clúster (K-Medoids)*
Fuente: *Elaboración propia (2023)*

Según lo mostrado en la Figura 52, el total de datos de la variable Sexo tiene la proporción de 48.06% hombres y 51.94% mujeres, esta primera variable brindará un primer indicio del tipo de cliente por segmento.

- Atributo “Local”:

LOCAL_ID	LOCAL_DESC	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
207	Angamos	4,007	1,201	1,206	829	1,538
304	Arequipa Cayma	3,425	994	1,341	1,152	1,124
321	Arequipa Porongoche	3,582	819	1,421	1,064	1,166
206	Atocongo	5,030	2,102	1,146	1,564	1,263
205	BELLAVISTA	2,789	981	918	826	881
322	Cajamarca Quinde	2,219	804	704	943	508
421	Canete	1,089	453	337	484	217
211	Centro Civico	5,085	2,083	1,229	1,468	1,418
307	Chiclayo Mall	6,479	2,603	1,971	2,613	1,789
309	Chimbote	2,629	954	784	1,184	480
214	COMAS	3,298	1,326	1,052	1,065	1,165
325	Cusco	2,145	900	520	720	909
662	Express Delivery Bellavista	21	4	2	13	1
664	Express Delivery Comas	10	3	2	4	
663	Express Delivery Plaza Nort	16	4	2	9	2
455	Express Huaraz	367	212	68	235	150
625	Express Mall del Sur	25	2	4	16	6
621	Express San Isidro	52	10	7	26	19
456	Express Tacna	415	189	88	192	128
801	FONOCOMPRAS	244	193	14	186	21
324	Huancayo	2,313	889	699	868	630
323	Ica Mall	2,442	761	646	1,074	594
800	INTERNET	28,835	412	23,039	6,631	283
102	Jockey Plaza	8,441	2,936	2,376	1,902	3,734
201	Lima Centro	1,535	651	420	576	323
209	Lima Norte	7,139	2,767	1,992	2,616	2,205
212	Mall del Sur	6,711	3,038	1,648	2,127	1,619
204	Megaplaza	6,385	2,456	2,028	2,042	1,940
106	Miraflores	3,824	1,377	997	924	2,074
303	Piura Centro	2,201	851	746	1,061	565
320	Piura Mall	3,703	1,151	1,401	1,381	1,251
213	Puruchuco	2,853	985	782	761	1,021
210	Salaverry	3,758	842	1,102	734	1,985
101	San Isidro	3,092	971	958	759	1,285
202	San Miguel	7,384	2,715	2,460	1,925	2,809
208	Santa Anita	3,316	1,103	999	976	910
420	SF Huanuco	1,342	580	285	721	378
400	SF Iquitos	1,133	539	209	568	330
419	SF Pucallpa	982	406	146	506	212
312	Trujillo Mall	4,646	1,838	1,429	1,639	1,358

Tabla 13: *Resultados Clúster por Local K-Medoids*
Fuente: *Elaboración propia (2023)*

Para este análisis se han seleccionado los locales que tienen el mayor número de transacciones, esto será relevante para determinar el comportamiento del cliente por clúster.

- Clúster 0: Las tiendas más representativas es el canal de Internet
- Clúster 1: Las tiendas más representativas son Mall del Sur, Jockey Plaza, San Miguel, Chiclayo Mall y Lima Norte.
- Clúster 2: Las tiendas más representativas es el canal de Internet.
- Clúster 3: Las tiendas más representativas es el canal de Internet.

- Clúster 4: Las tiendas más representativas son Jockey Plaza, Miraflores, Lima Norte y San Miguel.

- Atributo “Método de Pago”:

Transacción	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Fpay	76		1,815		
Venta American Express	2,332		2,326	131	
Venta CMR Credito	29,854	270	23,783	6,907	
Venta CMR Debito	38,784	4,087	26,747	10,509	
Venta con solo cheque	4	3			2
Venta Contado	28,377	11,834	2,507	18,808	3,628
Venta Credito Visa	44,098	25,720		8,029	33,478
Venta Debito Visa	1,437	1,191			1,183
Total general	144,962	43,105	57,178	44,384	38,291

Tabla 14: *Resultados Clúster Método de Pago K-Medoids*
Fuente: *Elaboración propia (2023)*

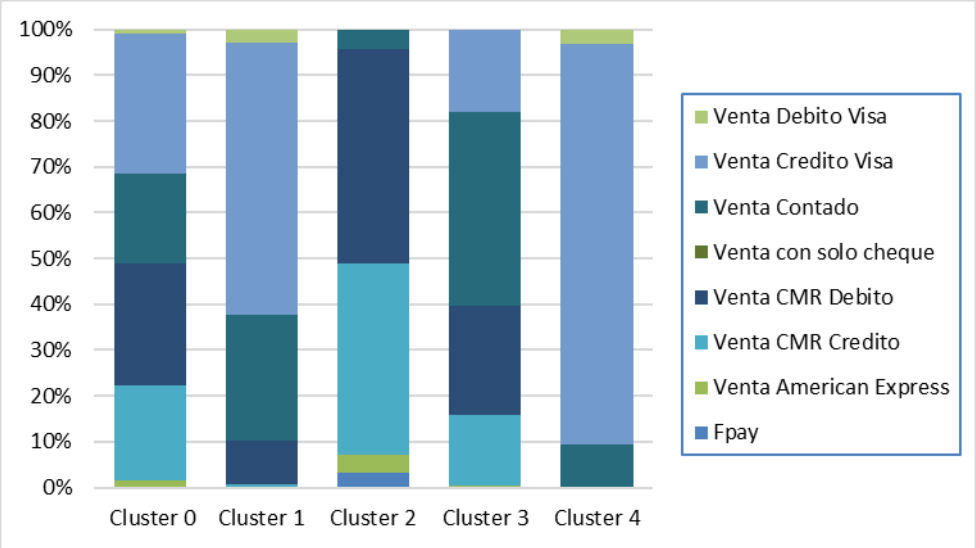

Figura 54: *Resultado Segmentación Transacción por Clúster (K-Medoids)*
Fuente: *Elaboración propia (2023)*

Para este análisis se han seleccionado los medios de pago que tienen el mayor número de transacciones, como se puede observar en la figura, hay ciertos tipos de transacciones que son más recurrentes al realizar la compra.

- Clúster 0: Los clientes de este clúster suelen usar los medios de pago de Venta Crédito Visa, Venta contado, Venta CMR Débito y Venta CMR Crédito.

Clúster 1: Los clientes de este grupo recurren al método de pago Venta crédito Visa.

- Clúster 2: Los clientes de este grupo recurren al método de pago Venta CMR Débito y Venta CMR Crédito.
- Clúster 3: Los clientes de este grupo recurren al método de pago Venta Contado y Venta CMR Débito.
- Clúster 4: Los clientes optan por usar el medio de pago Crédito Visa.

- Atributo "Sublínea":

sublinea_DESC	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ACCESORIOS	15,456	8,513	3,181	5,237	1,148
AUDIO	28,017	2,782	4,158	1,833	2,515
CLIMATIZACION	2,926	36	1,265	492	1,020
COCINA	2,361	1,434	2,231	1,463	342
COMPUTACION Y HOGAR	11,874	4,670	1,294	3,071	92
ELECTRODOMESTICOS	38,282	4,466	31,620	14,276	28,565
FOTOGRAFIA	293	155	21	234	1
LAVADO	3,666	70	1,893	1,214	617
REFRIGERACION	5,246	390	1,603	1,315	304
SERVICIOS E INSTALACIONE	16		11		
TELEFONIA	20,618	16,883	6,205	11,169	1,988
VIDEO	14,923	3,631	3,286	4,016	1,507
VIDEO JUEGOS	1,284	75	410	64	192
Total general	144962	43105	57178	44384	38291

Tabla 15: *Resultados Clúster Sublínea K-Medoids*
Fuente: *Elaboración propia (2023)*

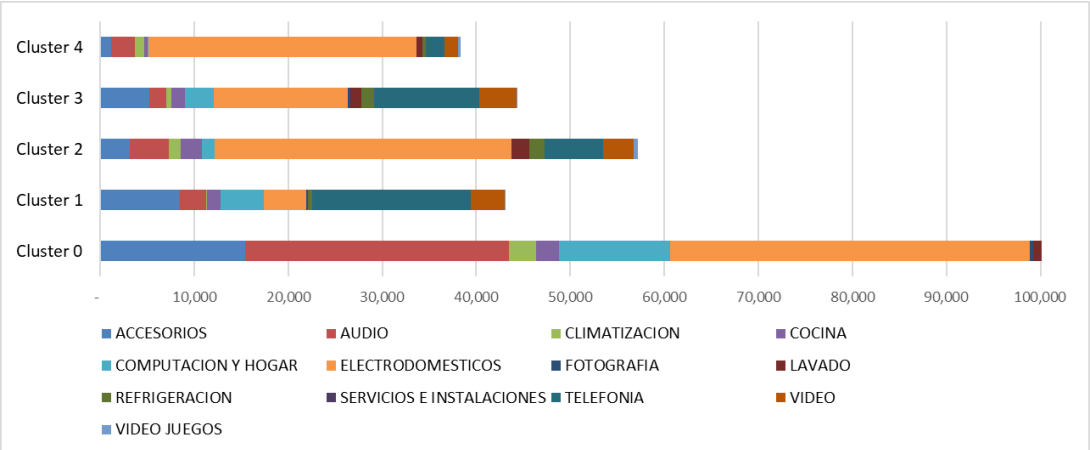

Figura 55: *Resultado Segmentación Transacción por Clúster (K-Medoids)*
Fuente: *Elaboración propia (2023)*

De la figura 55 se puede observar el número de transacciones que hay por sublíneas o categorías de sector electro, en esta se ha seleccionado las que han generado mayor volumen de venta por cada clúster.

- Clúster 0: Los clientes optan por compras en la categoría de Electrodomésticos, Audio y Telefonía.
- Clúster 1: Los clientes de este grupo realizan muchas compras de Telefonía.
- Clúster 2: Los clientes de este grupo suelen comprar accesorios y Electrodomésticos.
- Clúster 3: Los clientes de este grupo compran Electrodomésticos y Telefonía.
- Clúster 4: Los clientes optan por comprar Electrodomésticos.

- Atributo “Marca”:

	0	1	2	3	4
Marca	Samsung, Sullcandy y Recco	Claro	Recco, Oster y Samsung	Movistar y Recco	Recco, Oster y Wurden

Tabla 16: *Resultados Clúster por Marca K-Medoids*

Fuente: *Elaboración propia (2023)*

De la tabla se puede observar el número de transacciones que hay por marca, en esta se ha seleccionado las que han generado mayor volumen de venta por cada clúster.

- Clúster 0: Los clientes optan por compras en la marca de Samsung, Skullcandy y Recco.
- Clúster 1: Los clientes de este grupo realizan muchas compras de las marcas Claro, Ddesign.
- Clúster 2: Los clientes de este grupo suelen comprar las marcas Recco, Oster y Samsung.
- Clúster 3: Los clientes de este grupo compran Movistar y Recco.
- Clúster 4: Los clientes optan por comprar Recco, Oster y Wurden.

- Atributo “Unidades Vendidas”:

	0	1	2	3	4
Venta de Unidades	46.20%	10.71%	21.43%	12.56%	9.11%

Tabla 17: *Resultados Clúster por Unidades Vendidas K-Medoids*
Fuente: *Elaboración propia (2023)*

En la tabla se ha considerado los pesos de las unidades vendidas de cada clúster de la base de datos recolectado

- Clúster 0: Representa el mayor volumen de unidades vendidas representado con un 46.2%.
- Clúster 1: Representa el 10.71% del total de las unidades vendidas.
- Clúster 2: Representa el 21.43% del total de las unidades vendidas de todas las transacciones.
- Clúster 3: Representa el 12.56% del total de las unidades vendidas de todas las transacciones.
- Clúster 4: Este clúster representa el menor volumen de ventas con 9.11% comparado con el resto.

5.2 Medición de la solución.

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo.

En la valoración de los presentes modelos, usando la técnica de K-Means y K-Medoids, se tuvo la confirmación teórica del indicador “Inercia”. Esto significa que los modelos construidos nos brindarán un K óptimo a usar, esto se calcula con la distancia de cada punto con el tipo de clúster obtenido durante la simulación. Si al analizar los resultados se obtiene que la inercia es menor significa que hay proximidad en cada clúster. Los resultados de la evaluación de Inercia para los modelos de K-Means y K-Medoids se presentan a continuación:

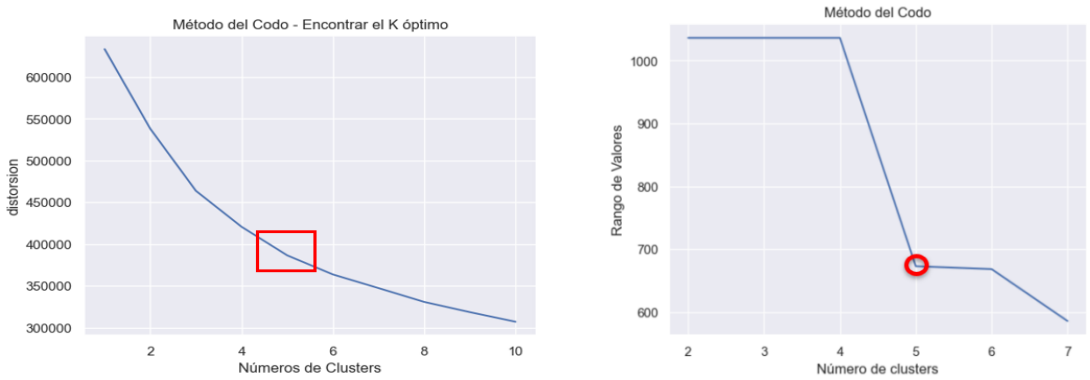

Figura 56: *Inercia por clúster K-Means y K-Medoids.*
Fuente: *Elaboración propia (2023)*

Con los resultados obtenidos por cada k, se observa que empieza a general una línea de tendencia y para ambos casos el punto de quiebre es en el clúster K = 5 en donde ocurre este cambio y es por ello, que se escogió este clúster como un K-óptimo.

5.2.2 Simulación de solución. Aplicación de Software

A continuación, se presentan la información consolidada según los clústers revisados, se evaluarán los perfiles del tipo de cliente.

- **Resultado Clúster K-Means:**

	Cluster K-Means				
	0	**1**	**2**	**3**	**4**
Sexo	Hombres y mujeres	Hombres y mujeres	Mujeres	Hombres	Hombres en su mayoria
Local	Lima Norte,Chiclayo Mall y Mall de Sur	Internet	Jockey Plaza, San Miguel	Jockey Plaza, San Miguel	Internet
Método de Pago	Venta Credito Visa, Venta Contado y CMR Debido	Venta CMR Crédito	Venta Credito Visa	Venta Credito Visa y Venta CMR Débito	Venta CMR Crédito
Categoría	Telefonía y Video	Electrodomésticos	Electrodomésticos y Accesorios	Electrodomésticos y Audio	Audio y Video
Marca	Movistar, Claro, Entel, Samsung	Samsung y Oster	Recco y Oster	Skullcandy, Recco y Ddesign	Samsung y Skullcandy
Venta de Unidades	19.47%	31.32%	22.54%	21.14%	5.53%

Tabla 18: *Resultados K-Means*
Fuente: *Elaboración propia (2023)*

Clúster 00: En este clúster se clasifica como hombres y mujeres que realizan compras en Lima Norte, Chiclayo Mall y Mall del Sur, y que adquieren principalmente el tipo de productos de la categoría telefonía y video preferentemente de la marca Movistar, Claro, Entel y Samsung, asimismo realizan el pago de sus compras a través de tarjeta CMR débito, Crédito Visa y al contado.

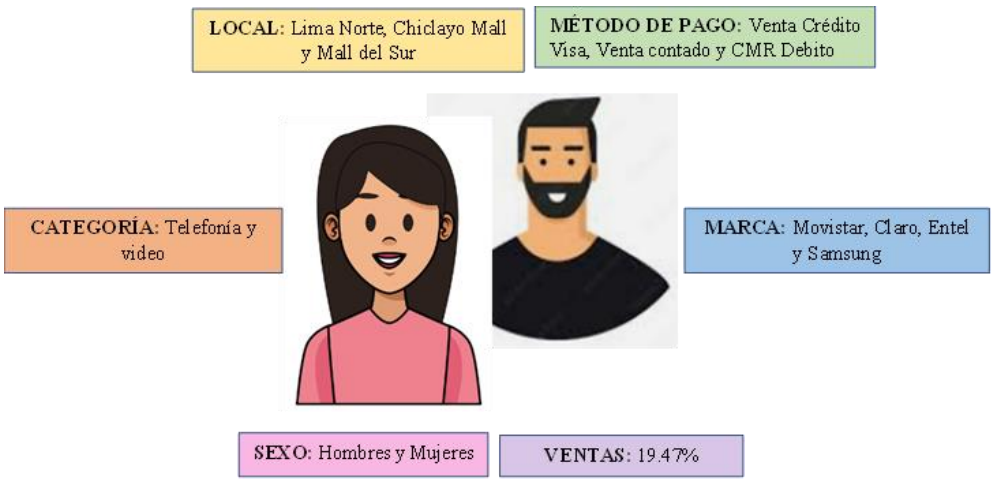

Figura 57: Descripción Clúster 00 K-Means.
Fuente: *Elaboración propia (2023)*

Clúster 01: Esta agrupación incluye hombres y mujeres en igual proporción que realizan sus compras por internet, cuya categoría de preferencia es electrodoméstica de la marca Oster y Samsung, adicionalmente predomina el pago a través de tarjeta CMR crédito.

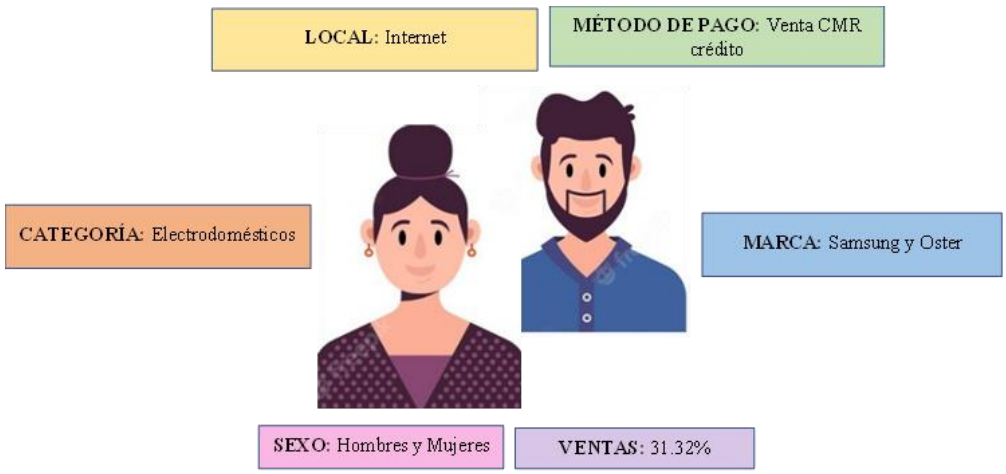

Figura 58: Descripción Clúster 01 K-Means.
Fuente: *Elaboración propia (2023)*

Clúster 02: En este grupo se describe el comportamiento de mujeres que prefieren usar el método de compra a través de crédito Visa, las tiendas más recurrentes son el Jockey Plaza

y San Miguel. La categoría de productos más comprados en este clúster son los electrodomésticos y accesorios, de los cuales se tiene como marcas más compradas Oster y Recco.

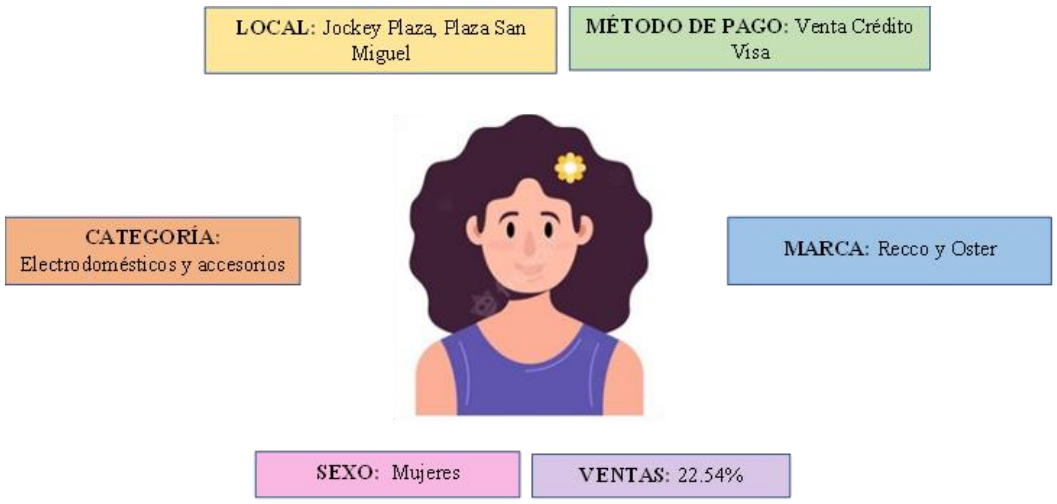

Figura 59: Descripción Clúster 02 K-Means.
Fuente: *Elaboración propia (2023)*

Clúster 03: Este clúster se clasifica como hombres que usan los métodos de pago crédito Visa y CMR débito, cuya categoría que suelen comprar es electrodomésticos y audio, dónde compran a marcas como Skullcandy, Recco y Ddesign. Las tiendas recurrentes en este sector son Jockey Plaza y San Miguel.

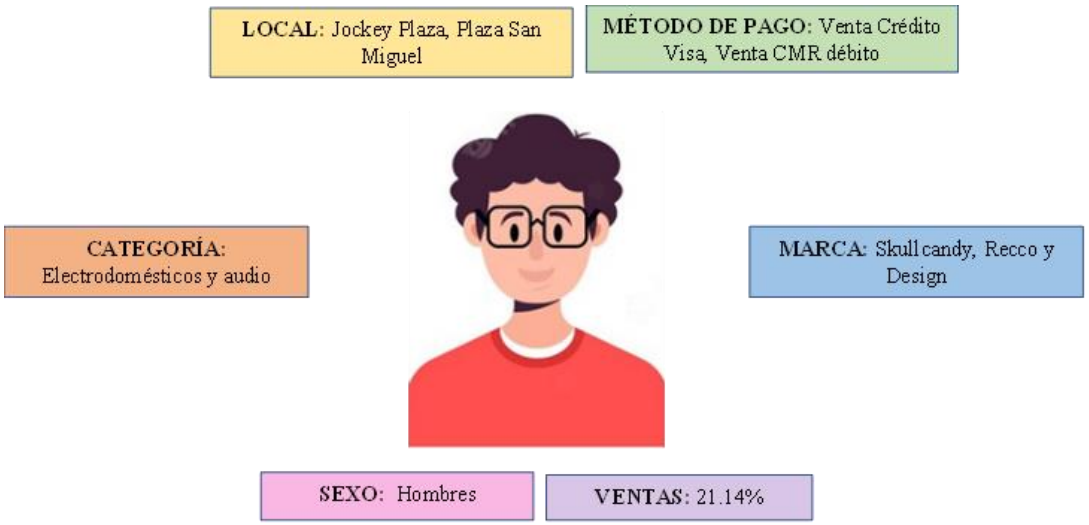

Figura 60: Descripción Clúster 03 K-Means.

Fuente: *Elaboración propia (2023)*

Clúster 04: Esta clasificación agrupa a hombres que prefieren realizar compras por internet, donde prima la categoría audio y video de las marcas Skullcandy y Samsung, en este grupo predomina el pago a través de tarjeta CMR crédito.

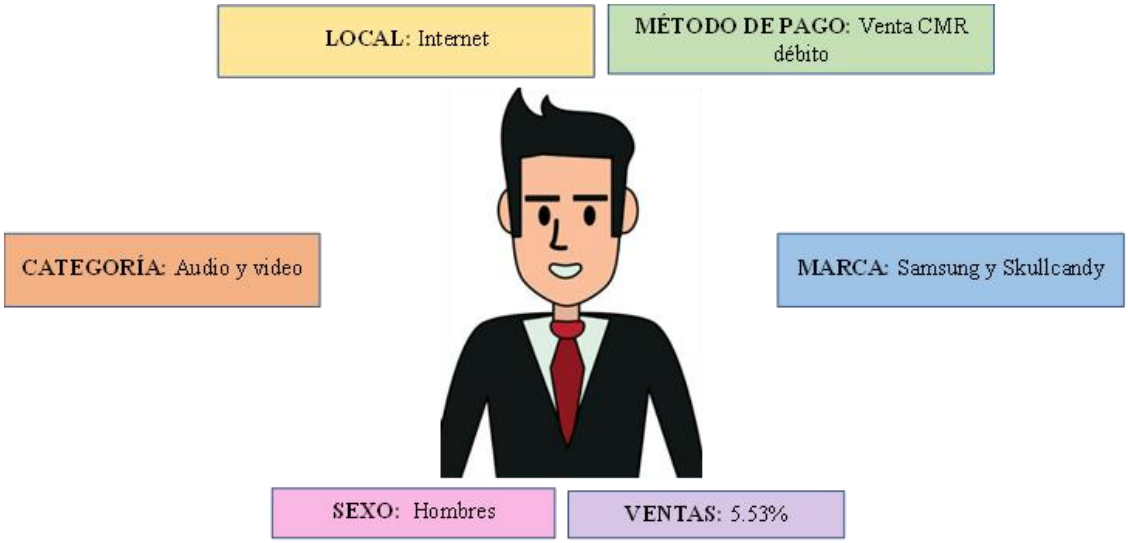

Figura 61: Descripción Clúster 04 K-Means.
Fuente: *Elaboración propia (2023)*

- **Resultado Clúster K-Medoids**

	Cluster K-Medoids				
	0	1	2	3	4
Sexo	Hombres	Hombres y Mujeres	Mujeres	Mujeres	Mujeres
Local	Internet	Mall del Sur, Jockey Plaza, San Miguel, Chiclayo Mall y Lima Norte	Internet	Internet	Jockey Plaza, Miraflores, Lima Norte y San Miguel
Método de Pago	Venta Credito Visa, Venta contado, Venta CMR Débito y Venta CMR Crédito	Venta Credito Visa	Venta CMR Débito y Venta CMR Crédito	Venta Contado y Vneta CMR Débito	Venta Crédito Visa
Categoría	Electrodomésticos, Audio y Telefonía	Telefonía	Electrodomésticos	Electrodomésticos y Telefonía	Electrodomésticos
Marca	Samsung Skullcandy y Recco	Claro, Ddesing	Recco, Oster y Samsung	Movistar y Recco	Recco, Oster y Wurden
Venta de Unidades	46.20%	10.71%	21.43%	12.56%	9.11%

Tabla 19: *Resultados K-Medoids*
Fuente: *Elaboración propia (2023)*

Clúster 00: En este clúster se clasifica como hombres que realizan compras en internet, y que adquieren principalmente el tipo de categoría electrodomésticos, audio y telefonía preferentemente de las marcas Samsung, Skullcandy y Recco, asimismo realizan el pago de sus compras a través de tarjeta CMR débito, CMR crédito, al contado y Crédito Visa.

Figura 62: Descripción Clúster 00 K-Medoids.
Fuente: *Elaboración propia (2023)*

Clúster 01: Esta agrupación incluye hombres y mujeres en igual proporción que realizan sus compras en Mall del Sur, Jockey Plaza, San Miguel, Chiclayo Mall y Lima Norte, cuya categoría de preferencia telefonía preferentemente de las marcas Claro y Ddesing, asimismo realizan el pago de sus compras a través de tarjeta Crédito Visa.

Figura 63: Descripción Clúster 01 K-Medoids.
Fuente: *Elaboración propia (2023)*

Clúster 02: En este grupo se describe el comportamiento de mujeres que prefieren usar el método de compra a través de crédito y débito CMR, vía Internet. La categoría de productos más comprados en este clúster son los electrodomésticos, de los cuales se tiene como marcas más compradas Samsung, Oster y Recco.

Figura 64: Descripción Clúster 02 K-Medoids.
Fuente: *Elaboración propia (2023)*

Clúster 03: Este clúster se clasifica como mujeres que usan los métodos de pago débito CMR y pago al contado, cuyas categorías recurrentes de compra es electrodomésticos y telefonía dónde las marcas de preferencia son Movistar y Recco. Las compras son por vía Internet.

Figura 65: Descripción Clúster 03 K-Medoids.
Fuente: *Elaboración propia (2023)*

Clúster 04: Esta clasificación agrupa a mujeres que prefieren realizar compras en las tiendas de Jockey Plaza, Miraflores, Lima Norte y San Miguel, donde prima la categoría electrodomésticos de las marcas Recco, Oster y Wurden, en este grupo predomina el pago a través de tarjeta Crédito Visa.

Figura 66: Descripción Clúster 04 K-Medoids.
Fuente: *Elaboración propia (2023)*

A continuación, se presentan la comparación consolidada según los métodos revisados:

	Clúster K-Means					Clúter K-Medoids				
	0	1	2	3	4	0	1	2	3	4
SEXO	Hombres y Mujeres	Hombres y mujeres	Mujeres	Hombres	Hombres	Hombres	Hombres y Mujeres	Mujeres	Mujeres	Mujeres
LOCAL	Lima Norte,Chiclayo Mall y Mall de Sur	Internet	Jockey Plaza, San Miguel	Jockey Plaza, San Miguel	Internet	Internet	Mall del Sur, Jockey Plaza, San Miguel, Chiclayo Mall y Lima Norte	Internet	Internet	Jockey Plaza, Miraflores, Lima Norte y San Miguel
METODO DE PAGO	Venta Credito Visa, Venta Contado y CMR Debido	Venta CMR Crédito	Venta Credito Visa	Venta Credito Visa y Venta CMR Débito	Venta CMR Crédito	Venta Credito Visa, Venta contado, Venta CMR Débito y Venta CMR Crédito	Venta Credito Visa	Venta CMR Débito y Venta CMR Crédito	Venta Contado y Vneta CMR Débito	Venta Crédito Visa
CATEGORÍA	Telefonía y Video	Electrodomésticos	Electrodomésticos y Accesorios	Electrodomésticos y Audio	Audio y Video	Electrodomésticos, Audio y Telefonía	Telefonía	Electrodomésticos	Electrodomésticos y Telefonía	Electrodomésticos
MARCA	Movistar, Claro, Entel, Samsung	Samsung y Oster	Recco y Oster	Skullcandy, Recco y Ddesign	Samsung y Skullcandy	Samsung Skullcandy y Recco	Claro, Ddesing	Recco, Oster y Samsung	Movistar y Recco	Recco, Oster y Wurden

Tabla 20: *Tabla comparativa K-Means VS K-Medoids*
Fuente: *Elaboración propia (2023)*

Clúster 00 K-Means y Clúster 03 K-Medoids: En ambos clústeres se clasifica a mujeres con preferencias en métodos de pago al contado y CRM débito, en categoría de compra Telefonía de la marca Movistar.

Clúster 01 K-Means y Clúster 02 K-Medoids: En ambas agrupaciones se tiene como público a hombres y mujeres que realizan compras por internet de electrodomésticos con método de pago CMR crédito de las marcas Samsung y Oster.

Clúster 02 K-Means y Clúster 04 K-Medoids: Ambos métodos nos muestran a mujeres con preferencia de visita a las tiendas de Jockey Plaza y San Miguel y compras de electrodomésticos de las marcas Recco y Oster usando como tipo de pago las tarjetas Crédito Visa.

Clúster 03 K-Means y Clúster 01K-Medoids: Estas clasificaciones agrupan a hombres que compran en el Jockey Plaza y San Miguel, la marca Ddesing con tarjetas de Crédito Visa.

Clúster 04 K-Means y Clúster 00 K-Medoids: En estos dos grupos se encuentra a hombres que compran por internet con preferencia en la categoría Audio con CMR crédito de las marcas Samsung y Skullcandy.

Con estos análisis de clústeres podemos tener una visión de estrategias de marketing a implementar ya sea para clientes que opten por hacer compras de manera electrónica o los que prefieran compran de manera presencial en tiendas, este tipo de comunicaciones se basaría al perfil de cada cluster obtenido.

Este desarrollo se planificaría de la mano con el área comercial, el área de marketing y el equipo de visual de tiendas. A continuación, mostraremos unos ejemplos de los comunicados que realiza Falabella, actualmente estos se anuncios se realizan de manera masiva a clientes registrados en la página web mas no cuentan con segmentación de clientes, por lo que es una oportunidad de focalizar las estrategias de venta.

Figura 67: Ejemplo de notificación a smartphones por cluster
Fuente: *Notificación Falabella (2023)*

Figura 68: Ejemplo de mailing por cluster
Fuente: *Mail Falabella (2023)*

Figura 69: Ejemplo de mensajes de texto por cluster
Fuente: *Elaboración Propia (2023)*

Para el caso de las tiendas físicas se implementarían corners o seasonals, que son espacios disponibles en una tienda donde hay módulos promocionando una marca y mostrando sus productos, el beneficio de este tipo de implementación es que es muy visual y el cliente puede tocar y usar el producto.

Figura 70: Ejemplo de corner en tienda por cluster
Fuente: *Mail Falabella (2023)*

CAPÍTULO VI: Conclusiones y recomendaciones

6.1. Conclusiones:

Debido al dinamismo de las empresas retail, estas empresas generan grandes cantidades de transacciones diarias, en esta investigación se ha demostrado que mediante un adecuado procesamiento de datos basado en machine learning se puede lograr procesar la información para conocer los criterios de compra del cliente y poder segmentarlos.

La data adquirida de la plataforma Data Ware House de Saga Falabella tiene un total de 327,920 transacciones con 22 variables o atributos recopilados en un lapso de tres meses desde noviembre 2022 a enero 2023. Esto significa que tenemos una base de datos muy amplia que nos ayudará a evaluar los atributos que lograrán determinar los clústeres.

En la etapa de pre-procesamiento de datos, se realizó una limpieza de la base de datos evitando cualquier ruido o valor nulo que altere el resultado final por ejemplo de los 22 atributos iniciales se redujeron a 6 los cuales son: Local, Sexo, descripción de la sub línea, marca, transacción y cantidad de unidades vendidas; siendo estas las más representativas para el análisis. Posteriormente se usaron técnicas como el PCA logrando reducir la cantidad necesaria de atributos, obteniendo el 78% de datos analizados.

Durante el modelado de las técnicas se evaluaron simular tres principales, K-Means, K-Medoids y Cluster Jerárquico, usando como lenguaje de programación Python a través de la plataforma Jupiter. Para la técnica de K-Means se usó el método del codo, logrando obtener un número de cluster K=5; para la segunda técnica de K-Medoids, el método del codo dio un resulta de K=5, por lo que validando ambos métodos se tiene la certeza que el K óptimo a usar es de cinco. Para la técnica de Cluster Jerárquico el K óptimo fue de siete, un valor alejado de las otras técnicas.

Finalmente, los resultados obtenidos basados en las dos evaluaciones nos brindaron segmentación de perfiles del cliente los cuales describen el comportamiento de compra. En nuestro análisis de ambas técnicas hemos encontrado que hay clústeres que comparten algunos atributos en común. Lo cual nos asegura una mejor toma de decisiones para elaborar estrategias de marketing asociadas a estos segmentos.

6.2. Recomendaciones:

Se recomienda evaluar otros métodos para determinar el número de clústeres óptimos, esto ayudará a tener una segmentación más exacta y a definir mejor las estrategias de mejora. También comentar que las técnicas aplicadas en el presente trabajo (K-Means, K-Medoids y Cluster Jerárquico) pueden ser implementada en cualquier otra empresa con registros parecidos a los del presente trabajo, lo importante es conocer cómo trabajas esta base de datos empleando un adecuado pre-procesamiento para que la simulación sea lo más acertada posible. Además, esto brindará un número de clúster (K) óptimo para realizar el análisis.

Adicionalmente, se debería consultar con expertos en el sector Retail para obtener opiniones sobre la base de datos utilizada y para corroboren los resultados obtenidos. Finalmente es recomendable extender el alcance de la simulación de machine learning aplicado en la empresa ya que existen ventajas adicionales de segmentación que podrían ser usados para analizar las campañas, cambios de temporadas, entre otros, permitiendo desarrollar estrategias adicionales que agreguen valor a la empresa.

Referencias Bibliográficas

Bellido, F. (2022). Narrativa digital con Inteligencia Artificial en Python. Universidad de Alicante, Escuela Politécnica Superior. Recuperado de: file:///C:/Users/roescurray/Downloads/Narrativa_digital_con_Inteligencia_Artificial_en_Py_Bellido_Delgado_Francesc.pdf

Cómo la pandemia transformó a las empresas de Retail y consumo. (3 de diciembre del 2020). Portal web PWC, recuperado de:https://www.pwc.com/cl/es/prensa/prensa/2020/Como-la-pandemia-transformo-a-las-empresas-de-retail-y-consumo.html

Chamba, Sairy (2015). Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC. Recuperado de https://dspace.unl.edu.ec/jspui/handle/123456789/10462

Crecimiento y estadísticas del E-commerce- Comercio electrónico, estadística y crecimiento 2022. (26 de abril del 2022). Portal Stackscale, recuperado de: https://www.stackscale.com/es/blog/crecimiento-estadisticas-ecommerce/

Cuál es el panorama para la industria del retail en 2023 (12 de diciembre del 2022). Portal web TheFoodTech, Recuperado de: https://thefoodtech.com/tendencias-de-consumo/cuales-es-el-panorama-para-la-industria-del-retail-en-2023/

Franco, A; Sobrevilla, V.; Gutierrez, M. García, L.; Suarez, A. & Rueda, E.(2021) Sistema de enseñanza para la técnica de agrupamiento K-means. Universidad Autónoma del Estado Hidalgo, Pachuca, Mexico. Recuperado de:https://repository.uaeh.edu.mx/revistas/index.php/icbi/article/view/7384/8278

Géron, Aurélien (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems [Aprendizaje automático práctico con Scikit-Learn, Keras y TensorFlow: conceptos, herramientas y técnicas para construir sistemas inteligentes], Sebastopol: O'Reilly Media, 2nd Edition.

Hurwitz, J., & Kirsch, D. (2018). Machine learning for dummies. USA: John Wiley & Sons, Inc. [recurso electrónico]. Recuperado de: https://www.ibm.com/downloads/cas/GB8ZMQZ3

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

La reformulación del Retail en Latinoamérica.(11 de mayo del 2022). Portal Kantar, recuperado de: https://www.kantar.com/latin-america/inspiracion/retail/2022-la-reformulacion-del-retail-en-latinoamerica

Lopéz, S.(2007). Algoritmos de Agrupamiento Global para datos Mezclados. Instituto Nacional de Astrofísica, óptica y electrónica. Puebla, Mexico. Recuperado de:https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/628/1/LopezES.pdf

Manrique, Esperanza. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para Desarrollo, Revista Ibérica de Sistemas e Tecnologías de Información; Lousada N.º E28, 586-599

Palacios,F. & 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, Frequency y Money)

Perú, Sector Retail 2019. (12 de diciembre del 2019). Portal BBVA Research, recuperado de:https://www.bbvaresearch.com/publicaciones/peru-sector-retail/

Reporte oficial de la industria Ecommerce en Perú, impacto del COVID-19 en el comercio electrónico en Perú y perspectivas al 2021. Elaborado por la Cámara Peruana de Comercio Electrónico (2021). Recuperado de: https://www.capece.org.pe/wp-content/uploads/2021/03/Observatorio-Ecommerce-Peru-2020-2021.pdf

Retail y comercio en el Perú:¿Cómo ha impactado la COVID.19?.(22 de mayo del 2020). Portal Web ConexionEsan, recuperado de:https://www.esan.edu.pe/conexion-esan/retail-y-comercio-en-el-peru-como-ha-impactado-la-covid-19

Rouhiaianen, L. .(2018). Inteligencia Artificial, 101 cosas que debes saber hoy sobre nuestro futuro.Editorial Planeta. Recuperado de:https://static0planetadelibroscom.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

Tendencias 2022 para la industria de Consumo y Retail en América del Sur.(Abril del 2022). Portal KPMG, recuperado de:https://assets.kpmg.com/content/dam/kpmg/pe/pdf/cr-tendencias-2022-en-america-del-sur.pdf

Yudith Sandoval (2019) Algoritmos de aprendizaje automático para análisis y predicción de datos. http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Zapotitla,R., (2019) Análisis de componentes principales. Recuperado de http://www.ptolomeo.unam.mx:8080/jspui/bitstream/132.248.52.100/139/7/A7.pdf