UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA EN GESTIÓN AMBIENTAL
INGENIERÍA INDUSTRIAL Y COMERCIAL
INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

"Implementación de técnicas de Machine Learning para la segmentación de clientes en una empresa del sector farmacéutico"

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los

requerimientos para:

Obtener el título profesional de Ingeniero en Gestión Ambiental,

Obtener el título profesional de Ingeniero Industrial y Comercial,

Obtener el título profesional de Ingeniero de Tecnologías de Información y Sistemas

AUTORES

Inga Llacza, Fabricio Gustavo
Miranda Manrique, Kevin Miguel Angel
Quispe Zuñiga, Dennys
Reyna Torres, July Mabel
Turriate Naveda, Santiago

ASESOR

Fabián Arteaga, Junior John

ORCID Nº 000-0001-9804-7795

Noviembre, 2023

TSP_Grupo 06

INFORME DE ORIGINALIDAD

15%	14%	3%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	3%
2	repositorio.unal.edu.co Fuente de Internet	2%
3	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Trabajo del estudiante	1%
4	repositorio.esan.edu.pe Fuente de Internet	1%
5	repositorio.uchile.cl Fuente de Internet	<1%
6	www.coursehero.com Fuente de Internet	<1%
7	www.analyticslane.com Fuente de Internet	<1%
8	Submitted to Massey University Trabajo del estudiante	<1%
9	www.scielo.org.co Fuente de Internet	

RESUMEN

La presente tesis se enfocó en la investigación e implementación de técnicas de Machine Learning para una empresa del sector farmacéutico, utilizando un conjunto de datos con más de 30 mil transacciones comerciales del período de junio a agosto de 2023. Esta investigación abarcó la recopilación, procesamiento, modelado y evaluación de los datos proporcionados por la empresa, para lo cual se emplearon técnicas de aprendizaje no supervisado como el modelo K-Means y Jerárquico, lo que llevó a la exitosa identificación de cuatro segmentos distintos de clientes. Estos hallazgos resaltan la efectividad de Machine Learning en la segmentación de clientes, lo que permitió poder identificar grupos con similitudes en sus características y patrones de comportamientos. Asimismo, se llevaron a cabo evaluaciones comparativas entre diferentes técnicas para determinar cuál se adaptaba mejor a las necesidades de la empresa. Tras un análisis detallado, se concluyó que el modelo K-Means era el más adecuado en este contexto, debido a que las agrupaciones se ajustaban más a la realidad del negocio. En consecuencia, se formularon estrategias personalizadas para aumentar la retención y satisfacción del cliente, con lo cual se tendrá mayor certeza en la toma de decisiones estratégicas y análisis de datos comerciales.

Palabras clave: Machine learning, segmentación, modelo K-Means, clúster jerárquico, industria farmacéutica.

ABSTRACT

This thesis focused on the research and implementation of Machine Learning techniques for a pharmaceutical sector company, utilizing a dataset with over 30,000 commercial transactions from the period of June to August 2023. The research encompassed the collection, processing, modeling, and evaluation of data provided by the company, using unsupervised learning techniques such as the K-Means and Hierarchical models. This led to the successful identification of four distinct customer segments. These findings emphasize the effectiveness of Machine Learning in customer segmentation, enabling the identification of groups with similar characteristics and behavior patterns. Comparative evaluations were also conducted among various techniques to determine the best fit for the company's needs. After a detailed analysis, the K-Means model was found to be the most suitable in this context, as it aligned more closely with the reality of the business. Consequently, personalized strategies were developed to increase customer retention and satisfaction, enhancing the certainty in strategic decision-making and commercial data analysis.

Key words: Machine learning, segmentation, K-Means model, hierarchical cluster, pharmaceutical industry.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

Actualmente, los avances tecnológicos han provocado cambios significativos en el comportamiento de los individuos y han revolucionado la toma de decisiones en las organizaciones. La accesibilidad a enormes cantidades de datos y el constante desarrollo de algoritmos de Machine Learning han proporcionado a las empresas las herramientas necesarias para abordar con éxito los desafíos que enfrentan y aprovechar las oportunidades emergentes.

En el contexto peruano, la industria farmacéutica opera en un entorno marcado por una diversidad demográfica y socioeconómica, así como por un crecimiento constante en el sector de la salud. Estos factores han dado lugar a una competencia intensa y a la necesidad urgente de estrategias de marketing altamente efectivas para atraer y retener clientes en un mercado saturado. En este contexto, las técnicas de Machine Learning desempeñan un papel fundamental en varios sectores, incluido el farmacéutico, al permitir la optimización de procesos internos, mejorar la toma de decisiones y, sobre todo, realizar una segmentación precisa de los clientes.

El presente estudio se direcciona a la implementación de técnicas de Machine Learning, particularmente el aprendizaje no supervisado, para analizar patrones de compra en la clientela de la empresa en estudio. La segmentación de clientes se vuelve esencial en respuesta al incremento de la competencia y a las cambiantes dinámicas del mercado, especialmente tras la pandemia.

Para abordar este desafío, se explorarán técnicas de Machine Learning como K-Means, K-Medoids y Clúster jerárquico, con el objetivo de identificar patrones de compra y segmentos específicos de clientes. Esto permitirá a la empresa ajustar su estrategia de mercado de manera precisa. A través de la metodología CRISP-DM, se garantizará un análisis estructurado que incluye la creación y validación de un modelo de segmentación, tanto desde una perspectiva técnica como mediante la validación de expertos del negocio.

La estructura de esta tesis se dividirá en varios capítulos que guiarán la investigación. En el Capítulo I, se planteará el problema, describirá la situación problemática, justificará la relevancia y se delimitará el estudio. Esto sentará las bases para comprender la importancia de la investigación que se llevará a cabo en los capítulos posteriores.

En el Capítulo II, se enfocará en el marco teórico, donde se analizarán los antecedentes y se presentarán los conceptos fundamentales del estudio. Esta sección proporcionará el contexto necesario para entender las teorías y conocimientos previos que respaldan la investigación.

El Capítulo III abordará el entorno empresarial, incluyendo un análisis FODA, el modelo de negocio CANVAS y un mapa de procesos. Esto permitirá una comprensión profunda del contexto en el que se desenvuelve la empresa objeto de estudio y cómo se relaciona con el problema planteado en el Capítulo I.

En el Capítulo IV, se centrará en la metodología de investigación, abordando el diseño del estudio, la estrategia de medición de resultados, el cronograma de actividades y el presupuesto. Aquí se explicará cómo se llevará a cabo la investigación y se detallarán los pasos a seguir para obtener los datos y resultados necesarios.

El Capítulo V presentará la propuesta de solución desarrollada a partir de la investigación. Se mostrarán los hallazgos y las recomendaciones derivadas del análisis de datos y la aplicación de técnicas de Machine Learning descritas anteriormente.

Finalmente, en el Capítulo VI, se concluirá el trabajo, proporcionando las conclusiones y recomendaciones basadas en los resultados. Además, se orientarán las futuras acciones en el contexto de la industria farmacéutica peruana, con la finalidad de promover el crecimiento y la innovación en el sector, cerrando así el ciclo de investigación y delineando una visión clara de cómo este proyecto puede hacer una contribución significativa y sostenible al desarrollo de la industria farmacéutica en el Perú.

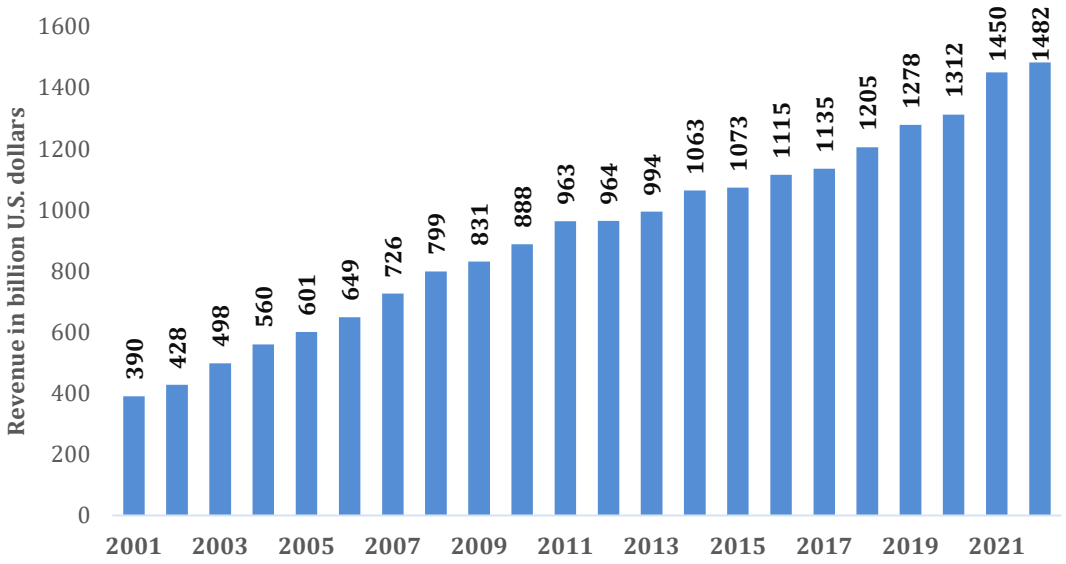
CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la Realidad Problemática

En la última década, la industria farmacéutica global ha experimentado cambios drásticos, evidenciados tanto en los patrones de consumo como en la distribución de mercados. Con el surgimiento de enfermedades crónicas y el envejecimiento de la población en muchas regiones, la demanda de medicamentos y tratamientos avanzados ha crecido exponencialmente. En 2021, el mercado farmacéutico global alcanzó un valor estimado de 1,5 billones de dólares, marcando un incremento significativo en comparación con años anteriores (Statista, 2023).

Figura 1

Ingresos del mercado farmacéutico mundial



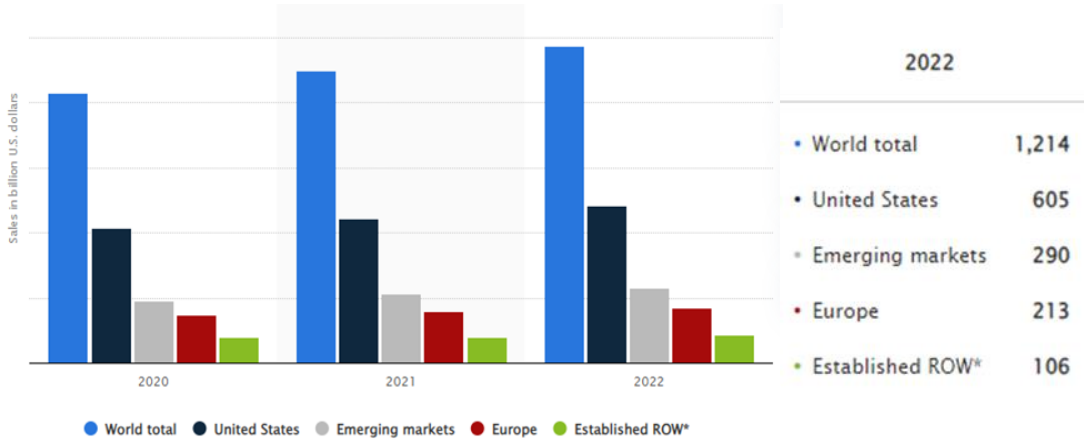
Nota. De *Evolución anual del volumen de ingresos de la industria farmacéutica a nivel mundial de 2001 a 2022 (en miles de millones de dólares estadounidenses)*, por Statista, 2023 (https://n9.cl/v3myl).

Continuando con esta trayectoria, en 2022, Estados Unidos consolidó su posición como el mercado farmacéutico líder, generando ingresos superiores a los 600 mil millones de dólares estadounidenses. A su vez, Europa no quedó atrás, contribuyendo con cerca de 213 mil millones de dólares estadounidenses.

Sin embargo, no se puede ignorar la influencia de los mercados en desarrollo como China, Rusia, Brasil e India. Estos, junto con regiones como América Latina y el subcontinente indio, destacan por sus tasas de crecimiento acelerado en ventas farmacéuticas, proyectándose como líderes hasta 2026.

Figura 2

Ventas farmacéuticas globales de 2020 a 2022 por región



Nota. De *Ventas farmacéuticas mundiales de 2020 a 2022 por región*, por Statista, 2023 (https://n9.cl/espji).

Sin embargo, este aumento en la demanda ha conllevado desafíos logísticos y operativos relacionados con la distribución y comercialización eficiente de los productos. El acceso a medicamentos sigue siendo un problema crítico en muchas áreas del mundo, con estimaciones que sugieren que aproximadamente una parte significativa de la población global todavía no cuenta con acceso regular a medicamentos esenciales (World Health Organization, 2019).

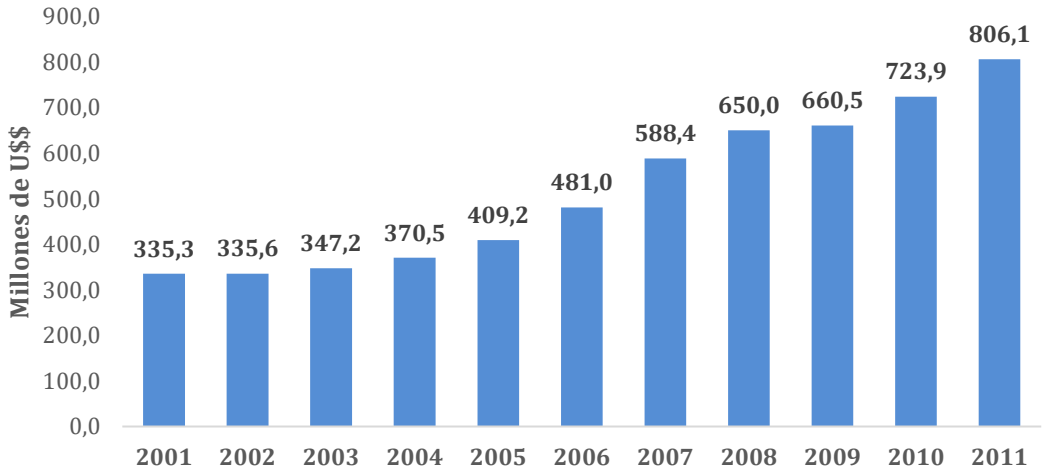
Junto con los desafíos logísticos, las empresas farmacéuticas enfrentan la ardua tarea de retener y gestionar a su base de clientes en un mercado saturado y competitivo. La diversidad de productos disponibles y la rapidez con la que emergen nuevas alternativas terapéuticas han hecho que la fidelización de clientes sea más esencial que nunca. Según la consultora McKinsey & Company (2020), la eficiencia en las estrategias de marketing y la correcta segmentación de clientes pueden incrementar las ventas en un 10-20%.

El impacto de la COVID-19 en la industria farmacéutica ha sido dual. Mientras que ha impulsado la demanda de compra de ciertos productos, como antivirales y medicamentos de soporte, también ha resaltado la fragilidad de las cadenas de suministro y la necesidad de estrategias de comercialización más resilientes. La competencia posterior a la pandemia ha intensificado la lucha por el mercado, con una avalancha de nuevos competidores y productos que amenazan la estabilidad de actores ya establecidos en la industria. Este ambiente competitivo, combinado con la necesidad de adaptarse rápidamente a los patrones cambiantes de consumo, ha complicado aún más la retención y atracción de clientes (Harvard Business Review, 2020).

Perú, inmerso en el contexto de crecimiento y transformación de la industria farmacéutica global, ha experimentado su propia oleada de desafíos y oportunidades. Tal como se refleja en tendencias globales, el país ha presenciado un notable auge en la demanda de medicamentos y servicios de salud. La Cámara de Comercio de Lima (2020) destaca que el mercado farmacéutico peruano ha mantenido un crecimiento anual cercano al 8%, una cifra que evidencia la robustez y potencial del sector en el país. Asimismo, se evidencia una evolución al alza con respecto a las tendencias de la venta de medicamentos en el sector.

Figura 3

Evolución del Mercado farmacéutico peruano



Nota. De *Evolución Anual de las Ventas de Medicamentos del Sector Retail (millones de US$)*, por Pan American Health Organization, 2018 (https://n9.cl/296e0).

Sin embargo, este panorama positivo en términos de crecimiento también ha traído consigo varios escenarios específicos para las empresas del sector en el país. La diversidad demográfica y socioeconómica de Perú implica una serie de necesidades y preferencias heterogéneas en términos de salud, lo que complica la tarea de las farmacéuticas para segmentar adecuadamente a su público objetivo (Ministerio de Salud de Perú, 2018).

Con el surgimiento de nuevas cadenas de farmacias y la aparición de distribuidores locales, la competencia en el sector ha ido en aumento. Esta dinámica competitiva lleva a menudo a las empresas a una lucha continua donde la diferenciación se destaca como un aspecto esencial para sobresalir (Procapitales, 2019). En este contexto tan variado y competitivo, es importante un conocimiento profundo del consumidor, así como la creación de ofertas de valor para asegurar y expandir la cuota de mercado.

Esta adaptabilidad y enfoque al cliente, tal como se refleja en las tendencias globales, resalta la importancia de la correcta segmentación en el mercado farmacéutico peruano, permitiendo a las empresas ofrecer soluciones que se alineen con las demandas particulares de su público.

Debido a temas de confidencialidad comercial, en este contexto, se hará referencia a la firma bajo el nombre de “empresa del sector farmacéutico”. Asimismo, es importante destacar que se adjunta el documento de conformidad proporcionado por el Coordinador General (Ver el anexo 1.), el cual autoriza y respalda el uso de los datos necesarios para llevar a cabo este proyecto de investigación.

Con un sólido posicionamiento en el mercado de salud, la empresa del sector farmacéutico en estudio enfrenta, al igual que muchas otras del mismo rubro, las consecuencias de las turbulencias y transformaciones en el ámbito comercial. Siendo un referente en la venta y distribución de productos farmacéuticos, la firma ha observado con preocupación los patrones de compra fluctuantes de su extensa clientela. Pese a contar con un portafolio diversificado, que abarca desde medicamentos genéricos hasta especializados, ha surgido la interrogante sobre cómo dirigir sus

estrategias comerciales y de marketing para satisfacer las demandas emergentes de sus clientes.

La globalización, reflejada en la entrada de competidores internacionales y en la adopción de patrones de consumo globales, ha influido en la manera en que los consumidores peruanos eligen y compran medicamentos. Además, a nivel nacional, la diversidad socioeconómica y cultural del Perú ha intensificado la segmentación del mercado, requiriendo soluciones específicas y ofertas personalizadas para cada segmento.

En un esfuerzo por adaptarse y atender con eficacia a las demandas variables de sus clientes, la empresa del sector farmacéutico ha decidido integrar técnicas de Machine Learning en su operación. A través de este enfoque, no solo busca obtener un panorama más amplio del perfil y comportamiento de sus clientes, sino también a identificar sus patrones de compra. Al aplicar Machine Learning, la empresa podrá ofrecer ofertas más ajustadas y pertinentes a las expectativas de sus clientes, enriqueciendo la experiencia de compra y afianzando su lealtad hacia la marca. Esta adaptación estratégica resalta su enfoque progresivo, demostrando su dedicación a la innovación y a mantenerse relevante en el dinámico mercado farmacéutico.

Con base en lo anterior, y comprendiendo la importancia estratégica de esta adaptación, en este estudio buscamos analizar la información transaccional de la empresa del sector farmacéutico seleccionada. Nuestro objetivo es aplicar técnicas avanzadas de Machine Learning para segmentar de forma óptima sus distintos grupos de clientes. Esta investigación no solo complementa el esfuerzo innovador de la firma, sino que también sienta las bases para abordar desafíos futuros en el dinámico mercado farmacéutico, garantizando que la empresa continúe siendo líder y referente en su sector.

1.2 Justificación de la Investigación

1.2.1. Justificación teórica

La aplicación de técnicas de Machine Learning, específicamente del aprendizaje no supervisado, ha adquirido relevancia en el análisis de datos en múltiples sectores, y la industria farmacéutica no es la excepción. Este estudio busca emplear dicho aprendizaje para identificar patrones de compra dentro de los datos relacionados con los clientes de la empresa del sector farmacéutico seleccionada para la investigación, en un contexto donde la segmentación de clientes se vuelve imperativa ante la creciente competencia y las cambiantes dinámicas del mercado. Posterior a la pandemia, estos patrones han experimentado transformaciones significativas, lo que resalta aún más la necesidad de reevaluar y comprender el comportamiento del consumidor en este sector.

Dentro de este contexto, la segmentación facilitará la identificación y definición precisa de grupos específicos de clientes. Esto brindará una comprensión más profunda de los consumidores, proporcionando a la empresa en estudio las herramientas necesarias para adaptar y fortalecer sus estrategias comerciales en respuesta a estas nuevas dinámicas post-pandemia.

1.2.2. Justificación metodológica

Para asegurar un análisis de datos coherente y efectivo, este estudio adoptará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es ampliamente reconocida por brindar un marco integral de trabajo, que guía durante las distintas fases del proceso, desde el entendimiento inicial de la operación hasta la ejecución y valoración de las soluciones.

Elegir una metodología apropiada es crucial para asegurar que la investigación aborde de manera precisa las necesidades de la problemática planteada. Dada la variabilidad en el comportamiento de compra en el sector farmacéutico, particularmente tras retos recientes como la pandemia, es esencial un enfoque que favorezca una segmentación basada en hechos. Por esta razón, el estudio opta por las técnicas de K-Means, K-Medoids y Clúster jerárquico, enfocándose en el aprendizaje no supervisado, ideal para interpretar y entender los patrones de compra de los clientes.

Utilizando dicho enfoque, es viable segmentar a los clientes en clústeres definidos por comportamientos y características afines. Estos grupos, respaldados por información actual y veraz, permiten no solo un análisis detenido de cada segmento, sino también la concepción de estrategias afinadas a cada conjunto. Atendiendo a la situación particular de la empresa del sector farmacéutico del proyecto, que aspira a proponer soluciones concordantes con los cambiantes patrones de comportamiento, la combinación de estas técnicas con la metodología CRISP-DM ofrece un análisis informado y flexible, facilitando la ágil adaptación de la empresa a las dinámicas del mercado farmacéutico.

1.2.3. Justificación práctica

Desde un enfoque empresarial, especialmente en el contexto de la industria farmacéutica, comprender y anticipar las necesidades de los clientes se ha vuelto más imperativo que nunca. La segmentación adecuada de clientes, lejos de ser solamente teórica, en la actualidad es una prioridad estratégica. Nos encontramos en una era donde el comportamiento de compra ha experimentado cambios notables, impulsados por factores globales y nacionales después de la pandemia y el incremento en la demanda de medicamentos. Es esencial que empresas del sector farmacéutico no solo reconozcan, sino que también actúen en base a estos cambios. Las empresas que logran identificar y comprender las especificidades de cada segmento pueden diseñar ofertas de valor que conecten directamente con las necesidades reales de sus clientes.

El uso de técnicas de Machine Learning para identificar patrones de comportamiento brinda a la empresa del sector farmacéutico en estudio un marco estructurado para adaptar sus estrategias comerciales. Al implementar técnicas avanzadas de segmentación, la empresa tiene la capacidad de alinear sus ofertas y servicios de manera precisa con las expectativas y necesidades cambiantes de cada segmento de cliente. Este enfoque no solo potencia la eficiencia en áreas como marketing y ventas, sino que, en la práctica, también busca estrechar la relación con los clientes. Al proporcionar productos y soluciones que se alinean con las necesidades de sus clientes, la firma puede fomentar una mayor lealtad, asegurando la retención y fortaleciendo su posición en un mercado farmacéutico cada vez más competitivo y dinámico.

1.3 Delimitación de la Investigación

1.3.1. Delimitación espacial

El alcance de esta investigación se enfoca en los datos comerciales de una empresa dentro del sector farmacéutico. Estos datos se obtienen de la compañía en cuestión, la cual ha otorgado su conformidad para el uso de la información, sin embargo, por razones de privacidad, el nombre de la empresa no se menciona directamente (Ver anexo 1).

En este contexto, se llevará a cabo una cuidadosa selección de las variables más relevantes y determinantes para el estudio, las cuales serán la base fundamental para la posterior segmentación de clientes. Este enfoque garantizará una precisión óptima en el proceso de agrupación y proporcionará información valiosa sobre los diversos segmentos identificados.

1.3.2. Delimitación temporal

La presente tesis se realizará durante los meses de junio hasta agosto de 2023. Es fundamental destacar que la data utilizada proviene de fuentes primarias de la empresa del sector farmacéutico seleccionada. Este intervalo temporal ha sido seleccionado para asegurar un análisis profundo de la información, permitiendo una segmentación efectiva de los clientes y, en consecuencia, ofreciendo una base sólida para el desarrollo de futuras estrategias comerciales acorde a los hallazgos.

1.3.3. Delimitación conceptual

Este estudio pone énfasis en la aplicación de Machine Learning, con un enfoque particular en el aprendizaje no supervisado. Utilizaremos técnicas como el K-Means, K-Medoids y Clúster jerárquico para identificar y definir segmentos de clientes. El principal objetivo es descubrir patrones de compra y, con base en esta información, determinar grupos específicos de clientes, analizando las dinámicas y características propias de cada uno de estos grupos. Para estructurar todo el proceso de análisis y garantizar una aplicación sistemática y efectiva de estas técnicas, implementaremos la metodología CRISP-DM, la cual proporciona una estructura integral que guía cada etapa del proceso, desde el inicio al fin.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes de la Investigación

2.1.1. Antecedente 1

Purnomo, M.; Azzam, A. & Khasanah, A. (2020). Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique. Journal of Physics: Conference Series. 1471. 012023. 10.1088/1742-6596/1471/1/012023.

El problema central de esta tesis busca abordar la relación entre la creciente competencia del mercado, que se ha intensificado debido a la rápida evolución de las tendencias de productos, y la complejidad del comportamiento del cliente. En este contexto, las estrategias de marketing tradicionales, que no consideran esta complejidad, han perdido efectividad. Esto es especialmente problemático porque el marketing es una actividad costosa y su ineficacia puede llevar a precios de productos más altos sin beneficio adicional, lo que disminuye la competitividad de los productos. La digitalización de los datos empresariales a través de tecnologías de la información ofrece oportunidades para comprender el comportamiento del cliente. La agrupación de clientes basada en datos de transacciones, utilizando parámetros como Recencia, Frecuencia y Monetario (RFM), se presenta como una herramienta efectiva para clasificar a los clientes según su comportamiento. Sin embargo, existe una necesidad de investigar más a fondo la efectividad y las implicaciones de esta técnica en el entorno del marketing y las ventas de productos. Con la investigación, se abordará el problema y se proporcionará perspectivas que puedan mejorar las estrategias de marketing y generar ventajas competitivas.

Metodología:

Se realizó un análisis en una empresa de muebles que opera tanto en entornos en línea como en tiendas físicas y se identificaron dos conjuntos de datos, uno relacionado con la información de los clientes y otro relacionado con los productos disponibles. Además, se recopiló una base de datos de transacciones, específicamente datos de ventas que registran las transacciones de los clientes que adquieren productos de la empresa. El estudio utilizó la información de 192 clientes registrados y abarcó un

catálogo de 37 tipos diferentes de productos de muebles. Para llevar a cabo el análisis, se extrajeron los registros de ventas de los años 2016 y 2018, lo que resultó en un conjunto de datos con un total de 30,240 registros. Estos datos extraídos se procesaron y almacenaron en una tabla temporal de RFM, la cual consta de 8 campos esenciales como el id del cliente, el nombre del producto, la cantidad, entre otros.

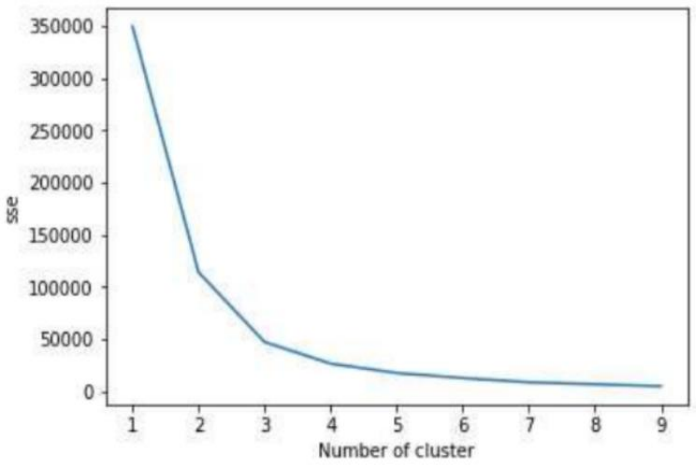
En este proceso, se calcularon los parámetros de Recencia, Frecuencia y Monetario utilizando fórmulas específicas. La Recencia se determinó restando la fecha actual a la última fecha registrada en la que un cliente realizó una compra. La Frecuencia se obtuvo contabilizando el número de transacciones realizadas por cada cliente. Finalmente, el parámetro Monetario se calculó sumando el valor total de los productos comprados por cada cliente.

Resultados:

El presente estudio utilizó la técnica de agrupación K-Means, que busca separar las muestras en clústeres con varianzas iguales y minimizar la inercia, evaluada mediante el valor del Error Cuadrático Sumado (SSE). Los resultados muestran que, al utilizar 4 clústeres, el valor del SSE deja de cambiar de manera significativa, por lo que, se optó por seleccionar un $n = 4$ clústeres para agrupar a los clientes.

Figura 4

Clúster óptimo



Nota. De *Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique* (p.4), por Purnomo, M.; Azzam, A. & Khasanah, A., 2020 (https://n9.cl/nh40u).

Además, se destaca que la media de cada clúster es significativamente diferente, lo que sugiere que los clientes en un clúster presentan valores de recencia (cuán recientemente un cliente ha realizado una transacción) considerablemente distintos de los clientes en otros clústeres. Además, se enfatiza que el cálculo de la frecuencia de compra se basó en el número de transacciones por cliente en lugar del número de facturas, lo que proporciona una evaluación más equitativa, teniendo en cuenta que una factura puede incluir múltiples transacciones.

Tabla 1

Resumen por Clúster

Recency_Clúster	count	mean	std	min	25%	50%	75%	max
0	11	167.1818	23.2198	132	152	163	185	201
1	33	88.4848	16.1305	68	75	83	99	126
2	53	45.8301	10.4839	30	37	44	54	65
3	95	13.0526	8.5307	0	6	12	20	29

Nota. De *Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique* (p. 4), por Purnomo, M.; Azzam, A. & Khasanah, A., 2020.

A lo largo de este capítulo, se explica que el análisis de los resultados revela que los clientes exhiben comportamientos diversos en términos de frecuencia de compra y lealtad monetaria. Por ejemplo, se muestra que la brecha en la frecuencia de compra de los clientes y su desviación estándar es relativamente alta, lo que indica que los patrones de compra varían significativamente. La agrupación de clientes según la frecuencia de compra se realizó utilizando el método de agrupación K-Means, similar al enfoque utilizado para la recencia. Además, se calculó el aspecto monetario lo cual refleja el gasto total realizado por los clientes, al multiplicar el número de productos adquiridos por el cliente por su precio correspondiente.

La agrupación de clientes según el aspecto monetario se realizó utilizando el método K-Means, y nuevamente se seleccionaron cuatro clústeres como el número óptimo. Esto permitió segmentar a los clientes en grupos con perfiles de lealtad monetaria similares.

Finalmente, se observa que los clientes pueden ser agrupados en 10 clústeres, pero desde una visión comercial, implementar 10 estrategias diferentes resultaría costoso y en ocasiones poco efectivo. Por lo tanto, se decidió comprimir los 10 clústeres de clientes en 3 grupos, que se pueden etiquetar como 'clientes poco leales', con una puntuación final de 0 a 2, 'clientes ligeramente leales', con una puntuación final de 3 a 6 y 'clientes leales', con una puntuación final de 7 a 9. En este sentido, se presenta una tabla con el perfil de los grupos de clientes y las alternativas de estrategias de marketing correspondientes:

Tabla 2

Perfil de los grupos de clientes y las alternativas de estrategias de marketing

Clúster de Clientes	Perfil del Clúster	Estrategias de Marketing Alternativas
1	Última visita hace más de 3 meses, frecuencia de visitas y gasto medios.	Introducir nuevos productos de precio bajo a medio para incentivar a los clientes a volver.
2	Última visita hace más de 1 mes, frecuencia de visitas media y gasto alto.	Introducir nuevos productos de precio medio y alto.
3	Última visita la semana pasada, frecuencia de visitas alta y gasto muy alto.	Introducir nuevos productos de precio alto y brindar un servicio cálido y personalizado a los clientes.

Nota. De *Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique* (p. 4), por Purnomo, M.; Azzam, A. & Khasanah, A., 2020.

Conclusiones:

A partir del estudio, se determina que los patrones de compra de los consumidores no son consistentes, y, por lo tanto, la estrategia de marketing debe diseñarse teniendo en cuenta estos patrones de compra. El modelo RFM se puede emplear como una herramienta de referencia para categorizar a los clientes y comprender mejor sus hábitos de compra. En cuanto a la agrupación de clientes, la técnica de K-Means, como método de agrupación no supervisado, puede utilizarse para agrupar el comportamiento del cliente al determinar previamente el número de clústeres. La combinación de RFM y la agrupación K-Means puede utilizarse como técnica de Machine Learning para entrenar a una computadora a realizar una minería de datos basada en el razonamiento de expertos humanos. Asimismo, se recomienda emplear técnicas de agrupación no supervisadas, es decir que los grupos de clientes se formarían de manera automática sin información previa, de manera que los miembros de un grupo tengan comportamientos de compra similares entre sí. Esto permitiría que las estrategias de marketing sean más efectivas, ya que podrían adaptarse de manera más precisa a las características y preferencias de cada grupo de clientes.

2.1.2. Antecedente 2

Madariaga, S. (2021) Machine Learning para predecir volúmenes operacionales de las líneas de negocio de Cenabast. Tesis para el título de magister. Universidad de Chile.

La investigación se realizó en base a las líneas de productos de Cenebast, organización vinculada al Ministerio de Salud de Chile, cuya finalidad consiste en asegurar el suministro de medicamentos y productos de salud a la red médica pública.

El estudio busca validar la capacidad de utilizar modelos de Machine Learning para realizar una predicción sobre la demanda de los bienes que ofrece Cenabast, tomando en cuenta variables clave que influyen en dicha demanda. De igual manera, se hace referencia a que la investigación posibilitará la obtención de acuerdos más beneficiosos al negociar mayores volúmenes de productos y conseguir economías en el ámbito gubernamental. La presente tesis implementará algoritmos de Machine Learning que pronostiquen la cantidad de futuras operaciones a fin de aportar información relevante durante la toma de decisiones estrategias de optimización.

Metodología:

Para llevar a cabo la proyección de la demanda de los productos farmacéuticos distribuidos por Cenabast, se emplea una base de datos que contiene información detallada acerca de los contratos vigentes con los proveedores de medicamentos. Esta información abarca aspectos como la cantidad de productos, los precios de venta y otros datos relevantes relacionados con la operación de la empresa. Los métodos que se aplicarán para anticipar la actividad en las diversas áreas de negocio son:

- Regresión Lineal
- Regresión de Árbol de Decisiones
- Regresión de Bosque Aleatorio
- Regularización
    - Regresión Ridge
    - Regresión Lasso
- Regresión Elastic Net

La base de datos utilizada para la tesis fue información histórica de suministros de Cenabast, que abarca las adquisiciones desde el año 2011 hasta el 2021, la cual contiene 130,105 registros con 50 columnas que pueden ser agrupadas en cinco categorías principales:

- Información de documentos, que incluye número de documento, tipo de compra, y método de compra, entre otros.
- Detalles de medicamentos, como código y nombre del medicamento, etc.
- Datos de los proveedores, incluyendo el RUT y nombre del proveedor, etc.
- Información de las compras, que abarca cantidad y precio de compra, así como detalles de facturación, etc.
- Seguimiento de las compras, que incluye la cantidad solicitada, cantidad entregada y fecha de entrega, entre otros aspectos.

En el proceso de modelado de Machine Learning, se desarrollan estos seis modelos, y se identifica la variable independiente se compone del año y el producto, mientras que la variable que se toma como referencia es la cantidad de producto requerida. En un principio, se emplean los modelos con los valores predeterminados, y luego se aplica el algoritmo de búsqueda en cuadrícula (Grid Search) para identificar

los parámetros óptimos de modelado para cada modelo, a partir de una combinación de valores preestablecidos. Se utiliza un grupo de datos de entrenamiento que abarca desde el año 2011 hasta el 2018, lo que equivale a 3,342 registros (74.57%), mientras que los datos de prueba corresponden al período comprendido entre 2018 y 2021, con un total de 1,140 observaciones (25.43%).

Resultados:

En cuanto a la información de los resultados de los seis modelos mencionados con sus parámetros predeterminados, se evaluaron las métricas de rendimiento. Se observó que el mejor modelo en términos de reducir el error entre la cantidad real y la predicción es la Regresión Lineal. Este modelo también exhibe el coeficiente de determinación más alto entre los seis modelos de Machine Learning, alcanzando un valor del 72.76%. Por otro lado, se identificó que el modelo con el peor desempeño en la estimación de la demanda es la Regresión Elastic Net, con un coeficiente de variación de tan solo -0.002%.

Tabla 3

Resultados de los modelos de Machine Learning

Modelo	**ECM**	**RECM**	**EAM**	**R2**
Regresión Lineal	21,560,773,792	146,836	65,971	72.76
Regresión de Árbol de Decisiones	28,438,521,470	168,637	35,962	64.07
Regresión de Bosque Aleatorio	23,741,388,020	154,082	35,503	70.00
Regresión Ridge	23,353,734,349	152,819	70,985	70.49
Regresión Lasso	21,574,245,878	146,882	66,066	72.74
Regresión ElasticNet	79,148,472,024	281,333	119,647	0.00
ECM: Error Cuadrático Medio				
RECM: Raíz del Error Cuadrático Medio				
EAM: Error Absoluto Medio				
R2: Coeficiente de Determinación				
(*)- Redondeado a 0.002 al decimal más cercano				

Nota. Adaptado de *Machine Learning para predecir volúmenes operacionales de las líneas de negocio de Cenabast*. (p.37), por Torres, M. & Mauricio, S., 2021 (https://n9.cl/04hqf).

Conclusiones:

A través de la metodología CRISP-DM, fue posible identificar el modelo de Machine Learning más apropiado para estimar la demanda. Se identificó que el más efectivo con sus parámetros predeterminados fue la Regresión Lineal, con un coeficiente de determinación del 72.76%. En el caso de utilizar un enfoque de ajuste de parámetros de modelado, el modelo de Regresión de Árbol de Decisión resultó ser el más eficaz en comparación a otros algoritmos que fueron aplicados. La aplicación de estas herramientas tiene el potencial de reducir pérdidas y facilitar la negociación de contratos más favorables con los proveedores.

2.1.3. Antecedente 3

Gil, C. (2023). El Machine learning en la predictibilidad de la logística de compra de los clientes de suplementos deportivos en cuatro distritos de Arequipa. Dataismo. 2. 10.53673/data.v2i7.103.

El objetivo central radica en analizar la utilidad de Machine Learning en la predicción de la compra de clientes de productos de suplementos deportivos en 4 ciudades del departamento de Arequipa. Para lograr esto, se pretende identificar las variables que tienen un impacto significativo en la capacidad predictiva de la logística de compra y describir cómo estas variables influyen en la predicción, lo cual permitirá facilitar el planteamiento de estrategias comerciales.

La metodología de adquisición de datos empleada en este estudio se realizó a través de una encuesta de opinión a 383 clientes de suplementos deportivos, se empleó para analizar la frecuencia de adquisición de suplementos deportivos. Los datos de la encuesta sentaron las bases para el desarrollo de la solución a través de la aplicación de Machine Learning. Para este propósito, se empleó la técnica de regresión logística.

La hipótesis principal de esta investigación sostiene que la aplicación de Machine Learning contribuye a establecer la capacidad de anticipar el flujo logístico relacionado con la adquisición de suplementos deportivos por parte de los clientes en cuatro distritos de Arequipa. Asimismo, las hipótesis específicas se formulan de la siguiente manera: La predictibilidad de la logística de compra se ve influenciada por la edad y el género como variables. Asimismo, la frecuencia de compra de suplementos

alimenticios está condicionada por las variables consideradas en la predictibilidad impactando en la estrategia de la compañía.

Metodología:

Para la construcción del modelo de Machine Learning, se ha determinado que las variables bajo estudio incluyen la edad, el género, el distrito y el tipo de suplementos. Estas variables se consideran como independientes, mientras que la frecuencia de compra se considera como la variable dependiente.

Los datos recopilados de las encuestas indican que, el 50.70% son mujeres y el 49.30% son hombres. Asimismo, se destaca que la frecuencia de compra predominante de los suplementos deportivos es mensual, con un 54.6%.

Tabla 4

Resumen del procesamiento de los datos

	Cantidad	Porcentaje
Frecuencia Semanal	3	3.8%
Frecuencia Quincenal	42	11.0%
Frecuencia Mensual	209	54.6%
Frecuencia Cada Tres Meses	129	33.7%
Sexo Masculino	189	49.3%
Sexo Femenino	194	50.7%
Válidos	383	100.0%
Perdidos	0	0
Total	383	

Nota. Adaptado de *El Machine learning en la predictibilidad de la logística de compra de los clientes de suplementos deportivos en cuatro distritos de Arequipa* (p.31), por Gil, C. L., 2022 (https://n9.cl/4naz0).

En la metodología de esta investigación, se emplea el método de eliminación hacia atrás como parte del análisis de datos. Inicialmente, se construye un modelo que abarca todas las variables y efectos potenciales especificados o considerados relevantes. Luego, se evalúa la importancia de cada efecto mediante un test de chi-cuadrado basada en la razón de verosimilitudes, determinando si su inclusión mejora la efectividad del modelo con respecto a la dispersión de los datos. Si un efecto no contribuye significativamente, se elimina de manera iterativa hasta obtener un modelo óptimo que refleje la relación entre las variables y los objetivos de la tesis.

Resultados:

Los resultados evidencian una fuerte asociación entre las variables género y tipo de suplemento, con un nivel de significancia superior al 52.70%, lo que señala una robusta capacidad predictiva del género en el modelo de Machine Learning. La bondad de ajuste, con un valor inferior a 0.05, respalda la aceptación de la hipótesis de que las predicciones del modelo tienen un impacto significativo en la estimación de la logística de compra, demostrando una adecuada adaptación del modelo final.

Tabla 5

Resultados del antecedente 3

Efectos	**Criterio de ajuste del modelo**	**Contrastes de selección de efectos**		
	log verosimilitud del modelo reducido	**Chi-cuadrado**	**gl**	**Significancia**
Intersección	306.263	0	0	-
Suplementos	322.58	16.317	3	0.001
Edad	319.274	13.011	3	0.005
Sexo	458.405	152.142	3	0
Distrito	333.402	27.138	3	0
Edad * Suplementos	317.009	10.746	3	0.013
Distrito * Suplementos	320.799	14.536	3	0.002

Nota. De *El Machine learning en la predictibilidad de la logística de compra de los clientes de suplementos deportivos en cuatro distritos de Arequipa* (p.32), por Gil, C. L., 2022 (https://n9.cl/4naz0).

De la misma manera, se obtuvo que el Valor de Nagelkerke logró explicar el 55.90% de la variación en la variable dependiente en relación con las variables independientes como: sexo, edad, el distrito y suplementos. El estadístico de chi-cuadrado, que compara el modelo final con uno más simple, indica que no se observa un aumento significativo en los grados de libertad al omitir un efecto, respaldando la hipótesis nula de que los parámetros de dicho efecto son insignificantes. Esto sugiere que el modelo final es adecuado y coherente para analizar la capacidad de predecir la logística de compra.

Conclusiones:

Los resultados del análisis de regresión logística respaldan la hipótesis central que establece que la implementación de Machine Learning tiene un impacto en la capacidad predictiva de la logística de compra de estos productos. Con relación a la primera hipótesis específica, se concluye que la edad no desempeña un papel significativo en la predicción del modelo, mientras que el sexo se revela como una variable relevante que debe ser considerada en la etapa de planteamiento de las estrategias con la oferta de suplementos deportivos. Respecto a la segunda hipótesis específica, se identifica que las variables más influyentes en la predictibilidad y toma de decisiones son el sexo y el distrito.

2.1.4. Antecedente 4

Mosquera González, D. (2022). Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación. Universidad Nacional de Colombia.

Durante la presente investigación se explica que la mayoría de los estudios se centran en el clúster de clientes utilizando enfoques tradicionales como RFM y métodos convencionales para predecir su valor monetario. A pesar de los avances en el uso de Machine Learning e IA para pronosticar el valor del cliente, rara vez se investiga cómo integrar estas predicciones en la segmentación de clientes. En este sentido, la pregunta principal que la tesis plantea es: ¿De qué manera se pueden potenciar los resultados en el clúster de consumidores al incluir la predicción del valor económico del cliente como un factor en el proceso de segmentación? En ese sentido, se delimitan los siguientes objetivos específicos:

- Llevar a cabo una minuciosa descripción de los datos de transacciones históricas de los clientes, utilizando técnicas que involucran la depuración, el análisis exploratorio y la descripción en detalle de los datos.
- Elegir enfoques, que pueden ser tanto basados en parámetros como en aprendizaje automático, que sean apropiados para predecir el valor financiero de los clientes.
- Idear un procedimiento metodológico que facilite el agrupamiento de los clientes, con una eficiente incorporación de la estimación del valor monetario de los mismos como un componente fundamental en el proceso de segmentación.
- Llevar a cabo una validación exhaustiva de los resultados de la clasificación de los clientes, donde se considera la predicción del valor monetario como un criterio para agruparlos.

Asimismo, se lleva a cabo un contraste entre los resultados obtenidos de la segmentación de clientes que no tiene en cuenta la anticipación del valor económico y la segmentación de clientes que sí la considera. Es decir, se lleva a cabo la programación correspondiente, teniendo en cuenta las transacciones o compras como variables independientes en un enfoque de aprendizaje no supervisado. No obstante, también se desarrolla un modelo en Python que incluye el valor monetario del cliente calculado como variable dependiente, con el propósito de contrastar los resultados obtenidos por cada modelo.

Tabla 6

Enfoques utilizados en el Antecedente 4

Enfoque de Aprendizaje	**Variables Utilizadas**	
	Independientes	**Dependientes**
No Supervisado	Información de transacciones	-
Supervisado	Información de transacciones	Valor monetario del cliente calculado

Nota. Elaboración propia, 2023.

Metodología:

En relación con la metodología empleada, se utilizó una base de datos de clientes que cumplía con criterios como la disponibilidad de un historial de compras de al menos un año, la presencia de registros transaccionales de compras con una identificación única para cada cliente, la existencia de fechas registradas para cada transacción y la inclusión de valores monetarios asociados a cada compra.

Esta base de datos está compuesta de registros de ventas de una empresa minorista en línea con sede en el Reino Unido. La empresa opera principalmente a través de Amazon y cuenta con una plantilla de 80 empleados, y se especializa en la comercialización de artículos de regalo para diversas ocasiones y atiende tanto a clientes minoristas como mayoristas, sin que exista una diferenciación explícita en los datos comerciales. El período de tiempo comprendido por los datos se extiende desde el 12 de diciembre de 2009 hasta el 9 de diciembre de 2011. En total, la base de datos contiene 1,067,371 registros transaccionales, que involucran 8 variables distintas, y abarca un conjunto de 5,833 clientes.

Tal como se mencionó anteriormente, en la tesis en mención se realiza la programación del aprendizaje supervisado y no supervisado en dos instancias. Las técnicas que se utilizaron fueron las siguientes:

Tabla 7

Resumen de las técnicas de Machine Learning utilizadas en el Antecedente 4

Tipo de Aprendizaje	Aplicación	Técnicas
Aprendizaje Supervisado	Análisis solo con la información de las variables de las transacciones	Regresión Lineal Regresión Ridge Regresión Lasso K-Nearest Neighbors Árbol de Decisión Bosque Aleatorio Máquina de Soporte Vectorial

Nota. Elaboración propia, 2023.

Tipo de Aprendizaje	Aplicación	Técnicas
Aprendizaje No Supervisado	Análisis incluyendo las variables de las transacciones y la predicción del valor monetario	K-Means Agrupamiento Aglomerativo Mezcla Gaussiana Birch

Nota. Elaboración propia, 2023.

Resultados:

En el contexto de la investigación, se busca predecir el valor monetario de los clientes y utilizar esta predicción en el proceso de segmentación. Se lleva a cabo la segmentación primero sin considerar la predicción del valor monetario y luego incorporándola. Para lograr esto, se aplican varios algoritmos de segmentación, incluyendo KMeans, *Agglomerative Clustering, GaussianMixture* y *Birch.* La incorporación de la predicción del valor monetario ayuda a reducir posibles sesgos en los datos y permite una segmentación más precisa y efectiva de los clientes.

- Selección de métodos para la estimación del valor monetario:

En este capítulo de la presente investigación, se realizó un proceso de estandarización de los datos y se experimentó con cuatro enfoques distintos con el objetivo de identificar el más eficaz. La selección del método de estandarización óptimo se basó en su impacto en el rendimiento de tres algoritmos de aprendizaje: LinearRegression, KNeighborsRegressor y RandomForest.

La evaluación de la eficiencia de estos algoritmos se basó en la métrica de Error Absoluto Medio (MAE). Los resultados señalaron que el método de estandarización más efectivo fue el StandardScaler, ya que arrojó el valor más bajo de MAE, que se situó en 509.566 en el caso de KNeighborsRegressor.

Posteriormente, se procedió a realizar pronósticos del valor monetario en un periodo futuro utilizando un conjunto de datos dividido en un 75% para entrenamiento y un 25% para pruebas, con una semilla aleatoria (random_state) establecida en 42.

Se aplicaron 7 técnicas de aprendizaje automático y se evaluaron a través de 3 métricas de error con el propósito de identificar el modelo más idóneo.

Tabla 8

Resultados implementación de modelos de Aprendizaje Supervisado

Modelo	MAE	RMSE	R2
Regresión Lineal	592.371	1255.323	0.468
Regresión Ridge	591.67	1253.757	0.47
Regresión Lasso	590.456	1252.966	0.47
K-Nearest Neighbors	510.397	1093.824	0.596
Árbol de Decisión	716.776	1690.701	0.036
Bosque Aleatorio	534.615	1371.611	0.365
Máquina de Vectores de Soporte	544.004	1355.345	0.38

Nota. Adaptado de *Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación*. (p.64), por Mosquera González, D., 2022 (https://n9.cl/jl5fv).

El autor Mosquera González, D. (2022), explica que el método Nearest Neighbors muestra un desempeño sobresaliente, ya que registra las métricas de error más favorables en las tres variables evaluadas. Presenta el MAE más bajo, con un valor de 510.397, el RMSE más bajo, con un valor de 1093.824, y el R_2 más alto, con un valor de 0.59. Por lo tanto, se opta por seleccionar este modelo como el más adecuado. (p.64)

- Clúster de clientes sin incorporar la predicción del valor monetario:

La tabla 9 presenta los resultados de la segmentación de los clientes de un grupo de datos que contiene 1.302 clientes identificados como datos normales. En la observación de los resultados, se puede notar que no existe un método claramente superior en las tres métricas. Birch demuestra un mejor rendimiento en las métricas de Davies_Bouldin y Silhouette, mientras que K-Means muestra resultados superiores en la métrica de Calinski_Harabasz.

Tabla 9

Resultados de la segmentación de clientes en conjunto de datos normales

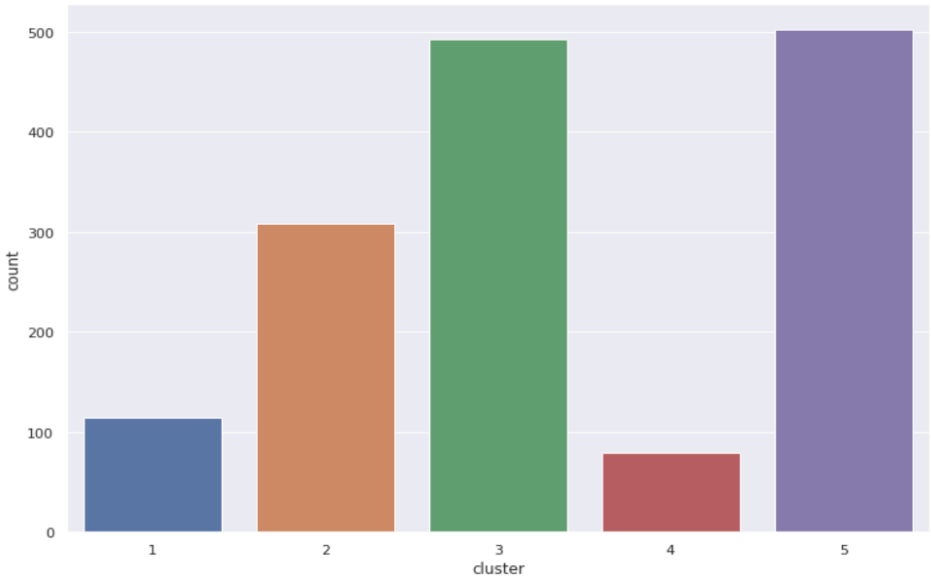
	Cantidad	Método de agrupamiento	K
Base de clientes	**1,495**		
Datos atípicos	193	K-Means	2
Datos normales	1,302	K-Means	3

Nota. Elaboración propia.

Además, los gráficos de barras que muestran el tamaño de cada uno de los conjuntos formados también se presentan. Estos gráficos revelan que el conjunto 5 es el más numeroso, con un total de 502 clientes, seguido por el conjunto 3 con 492 clientes. El tercer lugar lo ocupa el conjunto 2 con 308 clientes, seguido por el conjunto 1 con 114 clientes, y finalmente, el conjunto 4, el más pequeño, consta de 79 clientes.

Figura 5

Gráfico de barras total de clientes por clúster sin predicción de valor monetario



Nota. De *Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación*. (p.73), por Mosquera González, D., 2022 (https://n9.cl/jl5fv).

Finalmente, se proporciona un resumen del resultado final de los cinco clústeres formados, considerando las tres variables analizadas. Siguiendo la regla de que, cuanto menor sea el valor de "Recencia", mejor, y cuanto mayor sean las "Frecuencia" y "Valor_monetario", mejor, se podría concluir que el Clúster 1 se destaca como el grupo más destacado en este análisis.

Tabla 10

Resumen final segmentación sin incorporar predicción valor

		Recencia		Frecuencia		Valor_monetario	
Clúster	**Conteo**	**Media**	**Máx**	**Media**	**Máx**	**Media**	**Máx**
1	114	65.2	240	19.0	146	809.8	3219.9
2	308	41.8	193	8.5	20	451.2	846.5
3	492	53.6	145	3.1	10	239.2	550.5
4	79	402.3	544	2.3	18	308.5	1149.1
5	502	223.5	399	2.2	9	269.2	811.1

Nota. Adaptado de *Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación*. (p.74), por Mosquera González, D., 2022 (https://n9.cl/jl5fv).

- Segmentación de clientes incorporando la predicción del valor monetario:

En este capítulo, se señala que, a través del análisis del conjunto de datos, se detecta una dispersión significativa en las variables de "Valor_monetario" y "Frecuencia", y se observa nuevamente la existencia de datos atípicos o valores extremos. Para identificar estos datos atípicos, se emplea el algoritmo de Isolation Forest con un factor de contaminación automático, que equivale al 11.8% de los datos. Además, se procede a realizar la estandarización de los datos utilizando MaxAbsScaler (Pedregosa et al., 2011). Como resultado de este proceso, se identifican 177 clientes con datos atípicos y 1,318 clientes con datos considerados normales.

Tabla 11

Resultados métodos de agrupamiento para datos normales incorporando predicción del valor monetario

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
Kmeans	1.1	804.203	0.337	[3, 3, 3]
Agglomerative *Clustering*	1.229	643.002	0.304	[3, 3, 3]
GaussianMixture	1.562	469.024	0.241	[3, 3, 3]
Birch	1.044	667.991	0.335	[3, 3, 3]

Nota. De *Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación.* (p.78), por Mosquera González, D., 2022 (https://n9.cl/jl5fv).

Por último, en la tabla 12 se presenta un resumen de los conjuntos formados, y se destaca que el Clúster 1 sobresale como el grupo principal, basándose en la menor "Recencia", la mayor "Frecuencia", el mayor "Valor_monetario" y la predicción más alta del "Valor_monetario".

Tabla 12

Resultado final de segmentación incorporando predicción de valor monetario

		Recencia			Frecuencia			Valor monetario			Predicción valor monetario		
Clúster	Conteo	Mín	Media	Máx	Mín	Media	Máx	Mín	Media	Máx	Mín	Media	Máx
1	121	0	47,421	225	1	18,950	146	162,143	796,199	3219,905	44,976	2785,809	12709,860
2	247	0	40,939	176	1	8,672	22	135,338	453,025	805,573	286,999	1156,554	2463,184
3	543	0	54,871	146	1	3,569	14	8,500	250,811	604,000	0,000	370,518	1249,368
4	56	228	390,982	544	1	2,643	18	13,500	444,446	1581,520	0,000	176,510	745,335
5	528	138	233,341	436	1	2,237	12	10,080	263,691	814,200	0,000	182,315	778,602

Nota. De *Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación.* (p.78), por Mosquera González, D., 2022 (https://n9.cl/jl5fv).

Conclusiones:

Esta investigación aporta al ámbito de la estimación del valor del cliente y la segmentación de clientes. Se ha demostrado que las técnicas de aprendizaje automático, especialmente el algoritmo de Nearest Neighbors, superan a los modelos paramétricos convencionales en la predicción del valor del cliente, lo que respalda de manera empírica la eficacia de estas técnicas en este contexto específico. En lo que respecta a la segmentación de clientes, se destaca el sólido rendimiento del algoritmo K-Means en comparación con otras opciones. Además, se enfatiza la importancia de no excluir a los clientes con comportamientos potencialmente atípicos, sino más bien identificarlos y segmentarlos para obtener una comprensión más completa de su valor.

Asimismo, se concluye que, al validar la principal hipótesis de la investigación, se confirma que la segmentación de clientes que incorpora la predicción del valor económico resulta en un aumento sustancial de los ingresos y un valor promedio por cliente más elevado en comparación con la segmentación que no la incluye. Estos resultados sugieren que este enfoque puede mejorar la efectividad de las estrategias de marketing dirigidas a clientes de alto valor y ofrece una valiosa perspectiva para las decisiones comerciales y de marketing en el futuro.

2.1.5. Antecedente 5

Alkhayrat, M., Aljnidi, M. & Aljoumaa, K. (2020) A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data 7(9)

El objetivo de la presente tesis radica en abordar el desafío de la identificación de patrones en el comportamiento de clientes a partir de una base datos con alta dispersión en el contexto de una empresa de telecomunicaciones. Se busca evaluar el impacto de diferentes técnicas en la segmentación de clientes. Para lograrlo, se llevaron a cabo tres escenarios de comparación: el primero emplea el conjunto de datos sin reducción de características, el segundo se basa en la reducción de características utilizando únicamente PCA, y el tercero utiliza Autoencoders para la reducción de datos, lo que implica un aprendizaje de representación para mejorar el proceso de segmentación de clientes.

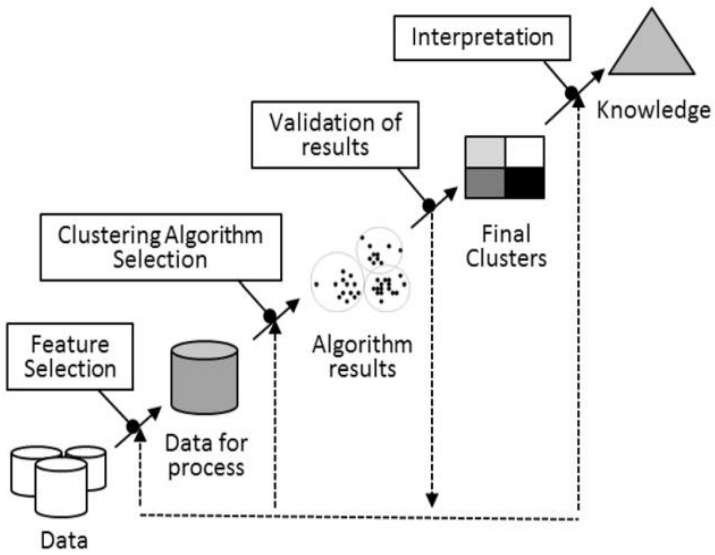
En esta investigación se utilizó un conjunto de datos de 220 características pertenecientes a 100,000 clientes de la empresa de telecomunicaciones, por lo que, la reducción de dimensionalidad es una fase relevante en el proceso de preprocesamiento de datos de minería de datos. Para ello, fue necesario aplicar la técnica de Análisis de Componentes Principales (PCA) y un enfoque basado en una Red Neuronal Autoencoder, realizando de esta manera una reducción de dimensionalidad de los datos originales. Luego, se aplicó el algoritmo de Agrupamiento K-Means tanto en el conjunto de data original como en el conjunto de datos reducido.

Metodología:

En este capítulo, los autores describen cómo se realizó el estudio y cómo se abordaron los diferentes aspectos de la investigación. Se describe el conjunto de datos utilizado, así como los pasos del proceso de agrupación de datos de alta dimensionalidad, que incluyen por ejemplo la preparación de datos de entrada, los cuales se basan en las transacciones de los usuarios móviles, las cuatro fuentes principales de datos fueron: Registros de Detalles de Llamadas (CDR), Base de Datos de Torres de Celdas, Servicios del Cliente e Información de Contratos de Clientes.

Figura 6

Pasos del proceso de clúster a partir de datos de alta dimensión



Nota. De *A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA*. J Big Data. (p.6), por Alkhayrat, M., Aljnidi, M. & Aljoumaa, K., 2020 (https://n9.cl/jmen2).

De igual manera, se destaca que la calidad de la segmentación depende en gran medida de la selección de características relacionadas con el comportamiento del cliente, los servicios suscritos y los datos demográficos, así como de la técnica de segmentación utilizada. Para lograr esto, se llevó a cabo la limpieza de datos para abordar problemas como valores faltantes y datos ruidosos, además de estandarizar los mismos utilizando el método *Z-score* y la escala *Min-Max* como parte de este proceso.

Con respecto a la reducción de la dimensionalidad de los datos, se utilizaron los enfoques de PCA y Autoencoder:

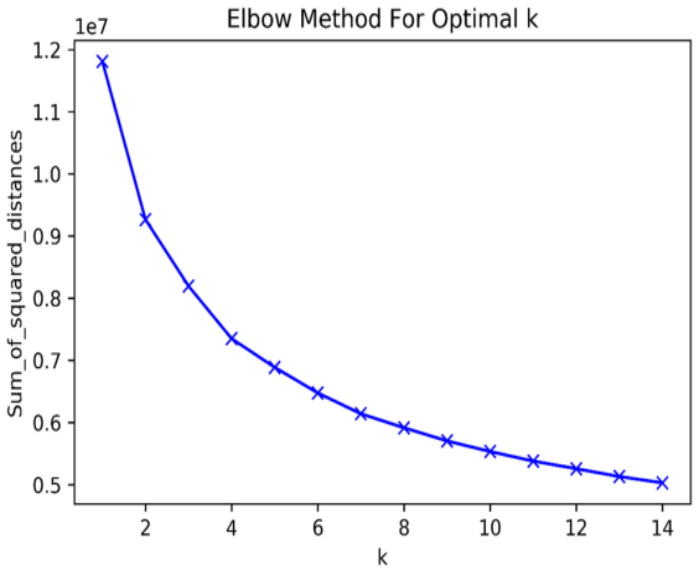
- Segmentación de clientes basada en PCA

Esta técnica busca encontrar direcciones de proyección que minimicen los errores cuadrados en la reconstrucción de los datos originales y maximicen la varianza de los datos proyectados. Se estandarizó la matriz de datos para que todas las características tengan una escala similar, y luego se calculó la matriz de covarianza de esta matriz estandarizada. Esto significa que se aplica el análisis de componentes principales (PCA) al conjunto de datos original a fin de transformar los datos en nuevas características que representan combinaciones de las características originales en un espacio completamente nuevo.

Para este caso, se encontró que 3 componentes principales explicaron el 72% de la variación total. Estos componentes se utilizaron para comprimir los datos y, finalmente, se aplicó el algoritmo de K-Means al conjunto de datos reducido para poder asignar a los clientes a grupos basados en similitudes en sus características. En este sentido, se aplicó el método del codo para identificar el número óptimo de clústeres, la figura 7 muestra un gráfico de codo con el SSE después de ejecutar el agrupamiento K-Means para un *k* que va desde 1 hasta 15, donde se observa un codo bastante claro en *k* = 4, lo que indica que 4 es el mejor número de clústeres.

Figura 7

Número óptimo de clústeres para los datos reducidos por PCA del antecedente 5



Nota. De *A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data*. (p.14), por Alkhayrat, M., Aljnidi, M. & Aljoumaa, K., 2020 (https://n9.cl/jmen2).

- Segmentación de clientes basada en Autoencoder:

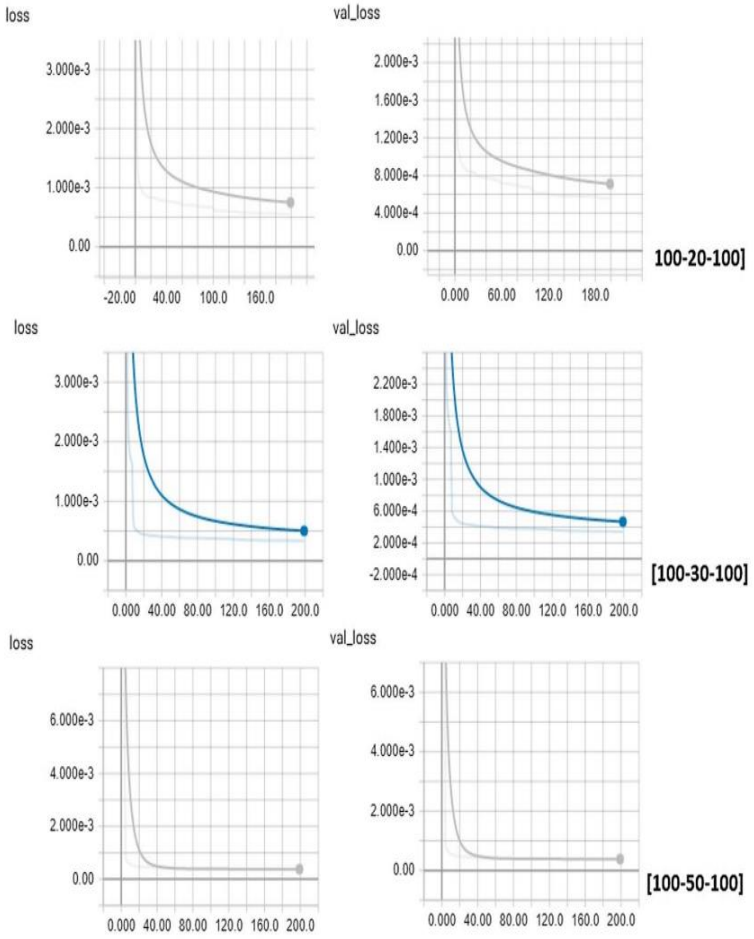
Además de examinar el rendimiento del modelo mediante la aplicación de PCA, se considera la reducción de la dimensionalidad basada en Redes Neuronales Artificiales (ANN), para el presente estudio se utilizó el autoencoder que permite comprimir los datos y luego reconstruirlos a su forma original.

Con respecto a este punto, los autores Alkhayrat, M. et al. (2020) explican que los autoencoders son un tipo de red neuronal artificial utilizada para aprender una representación o codificación de un conjunto de datos no etiquetados como un primer paso hacia la reducción de dimensionalidad mediante la compresión (codificación) de la matriz de datos de entrada a un conjunto reducido de dimensiones y luego reconstruyendo (decodificación) los datos comprimidos de vuelta a su forma original. Luego, los datos comprimidos se pueden utilizar en diferentes aplicaciones, como la segmentación. (p.12)

Una vez que se realizó el entrenamiento del autoencoder utilizando la biblioteca Keras de TensorFlow y compilando la red neuronal con una pérdida de error cuadrático medio y el optimizador Adam, se utilizó para reconstruir los datos originales mediante la codificación y decodificación de muestras del conjunto de validación, este proceso se puede visualizar a través de la interfaz web de TensorBoard, la cual permite representar gráficamente el error de reconstrucción (MSE) o función de pérdida, como se muestra en la figura 8:

Figura 8

Error de Reconstrucción (MSE) derivado del proceso del autoencoder



Nota. De *A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data.* (p.15), por Alkhayrat, M., Aljnidi, M. & Aljoumaa, K., 2020 (https://n9.cl/jmen2).

Resultados:

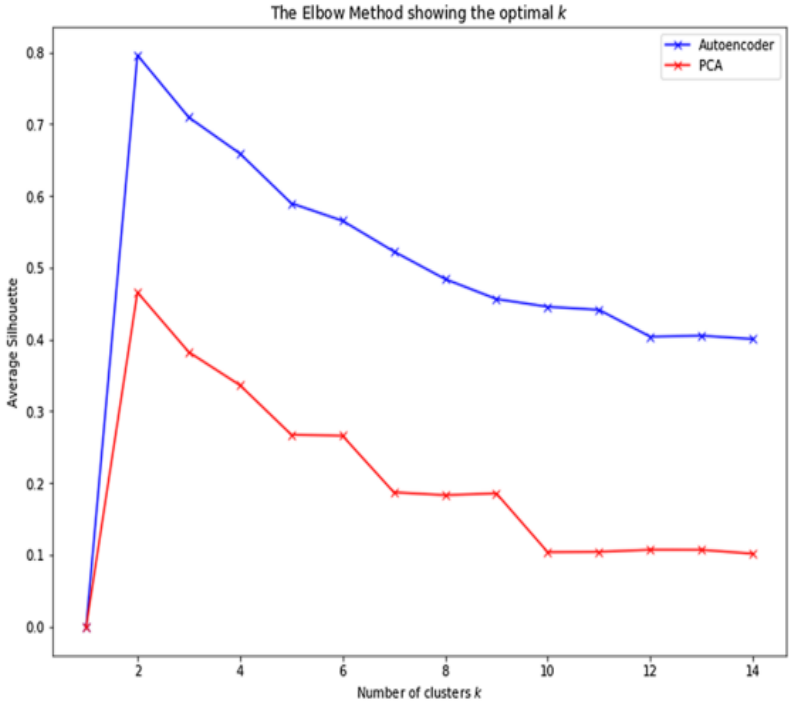
En este capítulo, se presentaron los resultados de los clústeres obtenidos para el conjunto de datos reducido con transformación PCA y la Red Neuronal Autoencoder. Después de reducir la dimensionalidad, se llevó a cabo el Clúster K-Means en los datos con un conjunto reducido de características. Dado que se trabajó con datos no etiquetados, se emplearon medidas internas para analizar la calidad de los resultados del clúster, incluyendo el coeficiente de silueta y el índice Davies–Bouldin (DBI).

Se realizaron experimentos configurando diferentes parámetros para el modelo y se evaluó su influencia en los resultados. Las arquitecturas probadas de la red Autoencoder para el conjunto de datos lograron minimizar el error de reconstrucción con 200 épocas y un tamaño de lote de 512 utilizando el optimizador Adam. Además, se aplicó el método K-Means en el conjunto de datos reducido a diferentes números de dimensiones, como 20, 30 y 50 características, y se evaluó el rendimiento sin reducción de dimensionalidad. Las métricas mencionadas se utilizaron para evaluar el agrupamiento en el conjunto de datos original y el conjunto de datos reducido.

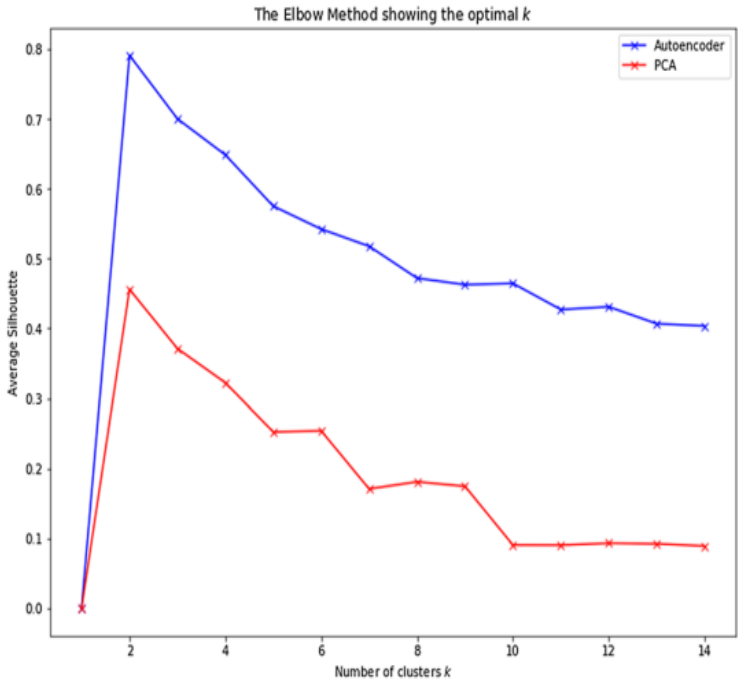
Los resultados experimentales indicaron que el método empleado fue efectivo en la tarea de segmentación de clientes y ofrecieron una idea de por qué fue necesario pasar de PCA a autoencoder. Se destacó que las puntuaciones más altas de silueta promedio se obtuvieron con el autoencoder para la reducción de dimensionalidad y el algoritmo de clúster K-Means. Las puntuaciones de silueta alcanzaron 0.682 con 3 clústeres y 0.571 con 5 clústeres, mientras que la puntuación obtenida en los datos originales con 220 dimensiones fue de 0.581.

Figura 9

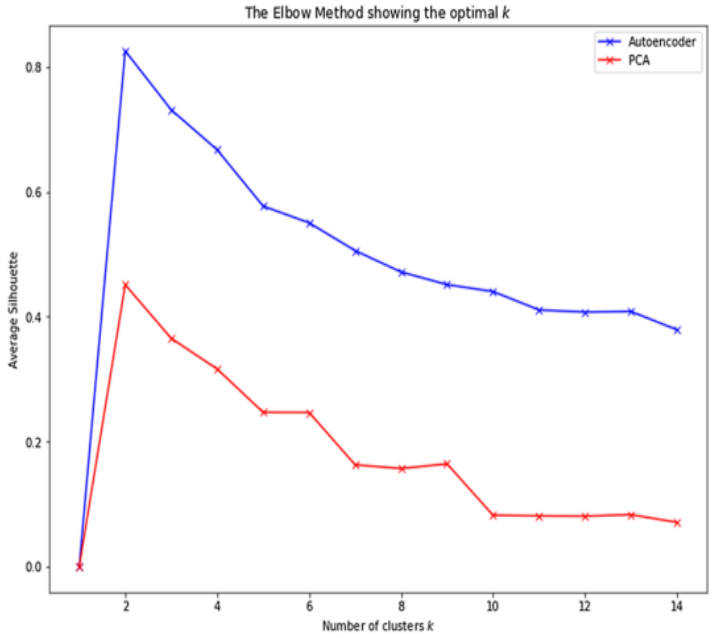
Método de la Silueta para diferentes dimensiones



Average Silhouette for PC= 20



Average Silhouette for PC= 30



Average Silhouette for PC= 50

Nota. De *A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA*. J Big Data. (p.18), por Alkhayrat, M., Aljnidi, M. & Aljoumaa, K., 2020 (https://n9.cl/jmen2).

Conclusiones:

La investigación demuestra que la reducción de dimensiones y los clústeres son enfoques efectivos para analizar grandes conjuntos de datos de clientes en la industria de las telecomunicaciones, ya que se observó que la reducción de dimensiones mediante el uso de Autoencoder resultó ser más eficaz que el PCA para abordar el desafío de segmentar clientes en un conjunto de datos caracterizado por su alta dimensionalidad.

Las métricas de evaluación mostraron que un $k = 3$ era el más adecuado para este conjunto de datos, lo que sugiere la existencia de tres segmentos distintos de clientes en función de su comportamiento, a partir de esta segmentación se pueden elaborar estrategias de marketing más personalizadas y efectivas.

De igual manera, se concluyó que la reducción de dimensiones tuvo un impacto positivo en el rendimiento del clúster. Por lo tanto, este enfoque presenta aplicaciones potenciales en entornos empresariales, ya que facilita una segmentación de clientes más efectiva.

2.2 Bases Teóricas

2.2.1. Machine Learning

La aplicación del Machine Learning ha experimentado un crecimiento significativo debido al interés de las organizaciones en aplicar esta tecnología para mejorar sus operaciones y tomar decisiones más informadas. Por lo que, la bibliografía sobre este tema es cada vez más amplia. El autor Zelcer, M. (2022) detalla que el Machine Learning es una rama de la IA que tiene como finalidad principal el desarrollo de técnicas que permitan a las computadoras aprender y mejorar a través de la experiencia. En este contexto, el término aprender se refiere a la capacidad de un programa o una aplicación de perfeccionarse a medida que adquiere experiencia. La mejora se evalúa mediante el cumplimiento de parámetros o reglas establecidas y la obtención de retroalimentación positiva de los usuarios del sistema. (p.18)

Asimismo, el Machine Learning puede dividirse en dos enfoques principales:

- Aprendizaje supervisado:

Con respecto a este tipo de aprendizaje, el autor Zelcer, M. (2022) menciona que los algoritmos en este enfoque se destacan por establecer una relación entre las entradas y las salidas deseadas, siendo comunes en problemas de clasificación, donde el sistema busca etiquetar elementos en categorías específicas. Generalmente, el aprendizaje supervisado se inicia con datos que ya están etiquetados, proporcionando al sistema múltiples ejemplos que incluyen tanto las entradas como las salidas deseadas. En este contexto, el proceso de aprendizaje busca identificar patrones dentro de estos datos que puedan ser utilizados para futuros análisis. (p.19)

De igual manera, el autor Brusil, C. (2020), explica que, dentro de la minería de datos, el aprendizaje supervisado se divide en dos aplicaciones principales: clasificación y regresión. En el caso de la clasificación, el algoritmo se entrena para etiquetar nuevas observaciones, utilizando como referencia datos previamente etiquetados. Esto puede implicar una clasificación binaria, como la distinción entre manzanas rojas y verdes (0 o 1), o una clasificación multiclase, como el reconocimiento de dígitos escritos a mano (0 a 9). En contraste, la regresión hace uso de múltiples

variables de predicción y una variable de respuesta continua para identificar una relación que permita estimar resultados numéricos. Un ejemplo típico de regresión lineal involucra ajustar una línea recta a los datos (como *X1* y *X2*) con el propósito de predecir resultados basados en la pendiente aprendida a partir de los datos de entrenamiento. (p.5)

- Aprendizaje no supervisado:

De igual manera, el autor Zelcer, M. (2022) considera que los algoritmos de aprendizaje no supervisado operan utilizando únicamente las entradas sin información previa sobre las categorías a las que pertenecen. En este enfoque, el sistema debe ser capaz de identificar patrones por sí mismo para agrupar o etiquetar las nuevas entradas, ya que no se le proporciona información sobre cómo deben clasificarse en categorías específicas. (p.19)

Con respecto a este concepto, el autor Brusil, C. (2020) define que, en situaciones de aprendizaje no supervisado, la información acerca de las observaciones en un conjunto de datos es limitada, ya que carecen de etiquetas. No obstante, es factible identificar similitudes entre grupos de observaciones y agruparlas en categorías adecuadas. El aprendizaje no supervisado se clasifica en dos categorías principales: agrupamiento (clúster) y reducción de dimensionalidad. (p.8)

2.2.2. Aprendizaje No Supervisado

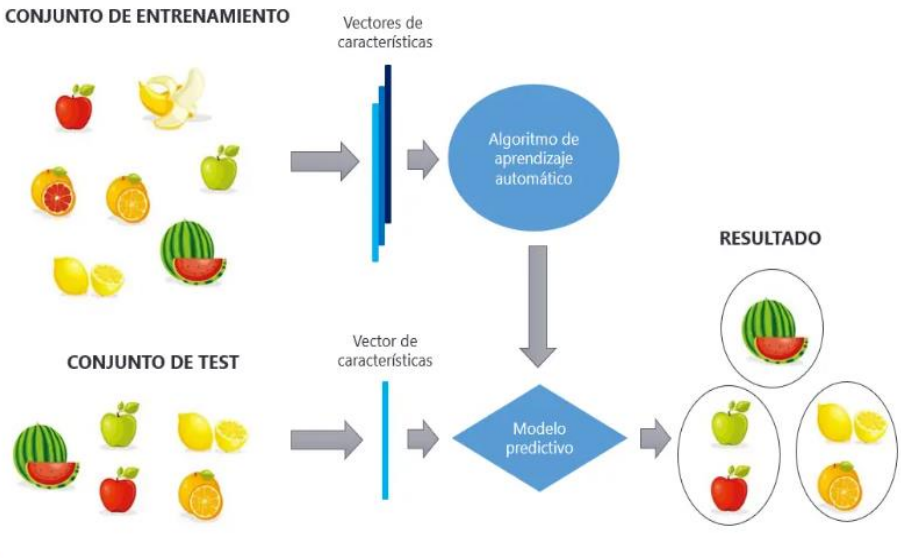
Los algoritmos de aprendizaje supervisado se capacitan para relacionar las entradas con las salidas y posteriormente se emplean para realizar predicciones o tomar decisiones en función de datos nuevos, y, por otro lado, los algoritmos de aprendizaje no supervisado se entrenan en un grupo de datos sin etiquetas ni salidas esperadas, a fin de descubrir patrones ocultos, estructuras o agrupaciones dentro de los datos.

Asimismo, Calvo, D. (2019) explica que las técnicas de aprendizaje no supervisado pueden aplicarse sin necesidad de contar con datos previamente etiquetados para el proceso de entrenamiento. Dado que solo se cuenta con datos de

entrada y no de datos de salida, el enfoque se concentra en descubrir algún tipo de patrón que simplifique el análisis. (párr.3)

Figura 10

Aprendizaje no supervisado



Nota. De *Aprendizaje no supervisado*. (párr. 5), por Calvo, D., 2019 (https://n9.cl/z9f9r).

Algunos ejemplos de aplicaciones del aprendizaje no supervisado son la identificación de posibles fraudes o la determinación de grupos de clientes con un comportamiento similar a fin de desarrollar estrategias de marketing. El aprendizaje no supervisado abarca varios enfoques:

- Clúster: Método que divide datos en grupos basados en similitud entre elementos, utilizando algoritmos como k-medias, agrupamiento jerárquico y agrupamiento espectral.
- Reducción de la dimensionalidad: El cual se centra en reducir variables o dimensiones manteniendo información relevante. PCA y t-SNE son ejemplos de técnicas que logran esto.
- Detección de anomalías: Identifica patrones inusuales en los datos, se aplica en la detección de fraudes, problemas de calidad, entre otros.

- Generación de datos: Se crean datos similares a los de entrenamiento, ejemplificado por Generative Adversarial Networks (GANs), empleados para generar imágenes, texto y otros datos sintéticos.
- Redes neuronales autoencoder: Estas redes aprenden una representación compacta de los datos y luego los reconstruyen, empleándose en compresión de imágenes, eliminación de ruido y más.

2.2.3. Clúster - Agrupamiento

Con respecto a este concepto, los autores Valles, M. et al. (2022) detallan que es una metodología de aprendizaje no supervisado que busca identificar grupos específicos en los datos. Se fundamenta en la idea de que un clúster es una región dentro de un espacio de datos contiguo con una alta densidad de elementos, separada de otros clústeres similares por zonas contiguas de baja densidad. Desde un punto de vista procedimental, los diversos métodos de agrupación intentan dividir los datos en *k* clústeres, minimizando las diferencias dentro de cada clúster y maximizando las diferencias entre grupos. Esto se logra mediante medidas de disimilitud dentro y entre clústeres utilizando una función de distancia. (p.3)

En este sentido, la definición de clúster se considera como la técnica de análisis de datos no supervisada que se base en encontrar la similitud entre los datos y se ejecuta mediante la aplicación de diversos algoritmos de agrupación con el fin de descubrir patrones dentro de los datos sin necesidad de etiquetarlos de antemano.

Según lo indicado en los estudios revisados, el proceso de clúster consta de varias etapas, que incluyen:

- Selección de atributos: Se eligen las características relevantes para el análisis y se preparan los datos.
- Elección del algoritmo: Se selecciona el algoritmo de clúster óptimo para el conjunto de datos y los objetivos del análisis.
- Definición de similitud o distancia: Se establece una medida de similitud o distancia entre los elementos del conjunto de datos, lo que determina cómo se agruparán.

- Ejecución del algoritmo: Se aplica el algoritmo elegido para agrupar los datos en clústeres.
- Evaluación de resultados: Se analiza la calidad de los clústeres obtenidos mediante métricas específicas, como el índice Davies-Bouldin o la silueta.
- Interpretación de resultados: Se analizan los clústeres resultantes para extraer conocimiento y tomar decisiones informadas.

2.2.3.1. K-Means

Tomando como referencia el artículo “Técnicas de agrupamiento para métricas en difractogramas”, el autor Castiblanco, J. (2020) menciona que K-Means es uno de los métodos más ampliamente empleados para la agrupación de datos mediante técnicas de aprendizaje no supervisado, debido a su aplicación simple y a los resultados efectivos que proporciona. El algoritmo tiene como finalidad clasificar un conjunto de datos compuesto por "*n*" elementos en "*k*" clústeres distintos. Para asignar cada elemento a un clúster específico, se definen centroides que se inicializan de manera aleatoria en el espacio de datos. Luego, se procede a minimizar la suma de los cuadrados de las distancias dentro de cada grupo. Esto se logra a través de una ecuación en la media y cada iteración se ajusta a la posición de los centroides hasta que se encuentre un conjunto de clústeres que minimice la distancia entre ellos. (p.7)

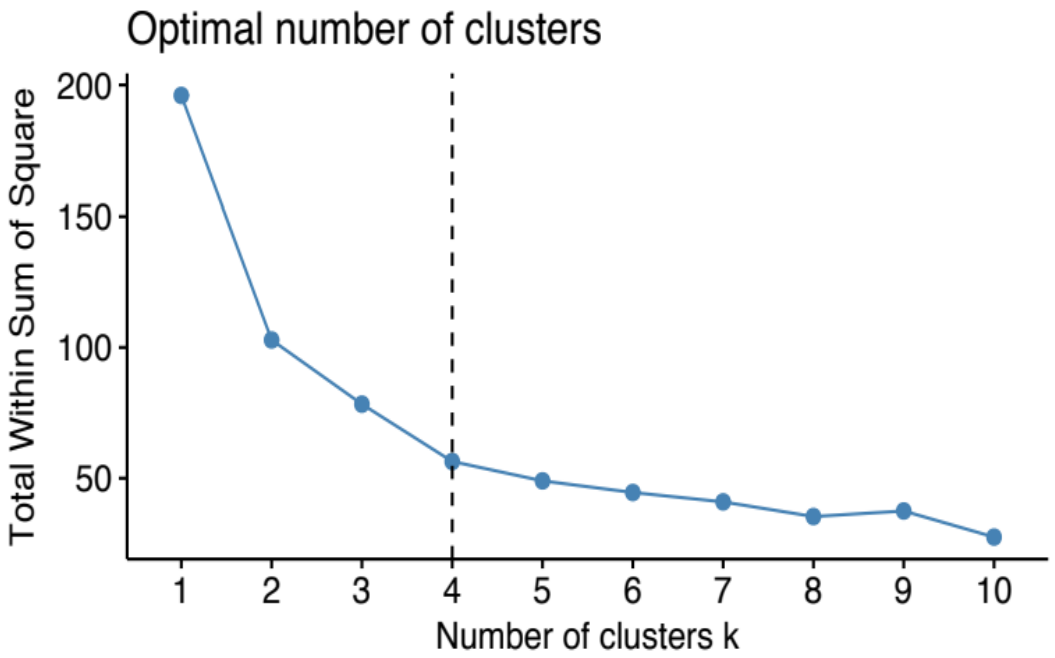
Asimismo, el autor Kassambara, A. (2017) menciona que a fin de estimar el número óptimo de clústeres se pueden utilizar diferentes valores de *k* y a su vez, trazar la suma de cuadrados dentro del clúster (*wss*) en función del número de clústeres.

Con respecto a este punto, la visualización de una curva en el gráfico se considera un buen indicador del número de clústeres. Por ejemplo, en la figura 11 se muestra que la varianza disminuye a medida que k aumenta, pero se puede observar un quiebre (o "codo") en $k = 4$. Este quiebre indica que los clústeres adicionales más allá del cuarto tienen poco valor. (p.40)

Figura 11

Clúster K-Means

```
library(factoextra)
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)
```



Nota. De *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning* (1st ed.). STHDA. (p.40), por Kassambara, A., 2017.

Con el propósito de llevar a cabo la implementación de K-Means en Python mediante la biblioteca *Sklearn*, se utiliza conjunto de datos como ilustración que contiene características de la personalidad de usuarios de Twitter.

El primer paso, consiste en la importación de las bibliotecas necesarias para llevar a cabo el algoritmo, así como la carga del archivo CSV, a fin de visualizar los primeros cinco registros del archivo en formato tabular:

Figura 12

Importación de las bibliotecas necesarias

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min

%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
```

```python
dataframe = pd.read_csv(r"analisis.csv")
dataframe.head()
```

	usuario	op	co	ex	ag	ne	wordcount	categoria
0	3gerardpique	34.297953	28.148819	41.948819	29.370315	9.841575	37.0945	7
1	aguerosergiokun	44.986842	20.525865	37.938947	24.279098	10.362406	78.7970	7
2	albertochicote	41.733854	13.745417	38.999896	34.645521	8.836979	49.2604	4
3	AlejandroSanz	40.377154	15.377462	52.337538	31.082154	5.032231	80.4538	2
4	alfredocasero1	36.664677	19.642258	48.530806	31.138871	7.305968	47.0645	4

Nota. De *K-Means en Python paso a paso*, por Aprende Machine Learning., 2018 (https://n9.cl/9bs8j).

Luego de establecer la estructura de datos que servirá como entrada para el algoritmo, se observa que se han cargado únicamente las columnas correspondientes a "*op*," "*ex*," y "*ag*" en la variable *X*.

$$X = np.array(dataframe[[\text{op, ex, ag}]])$$
$$y = np.array(dataframe['categoria'])$$
$$X.shape$$

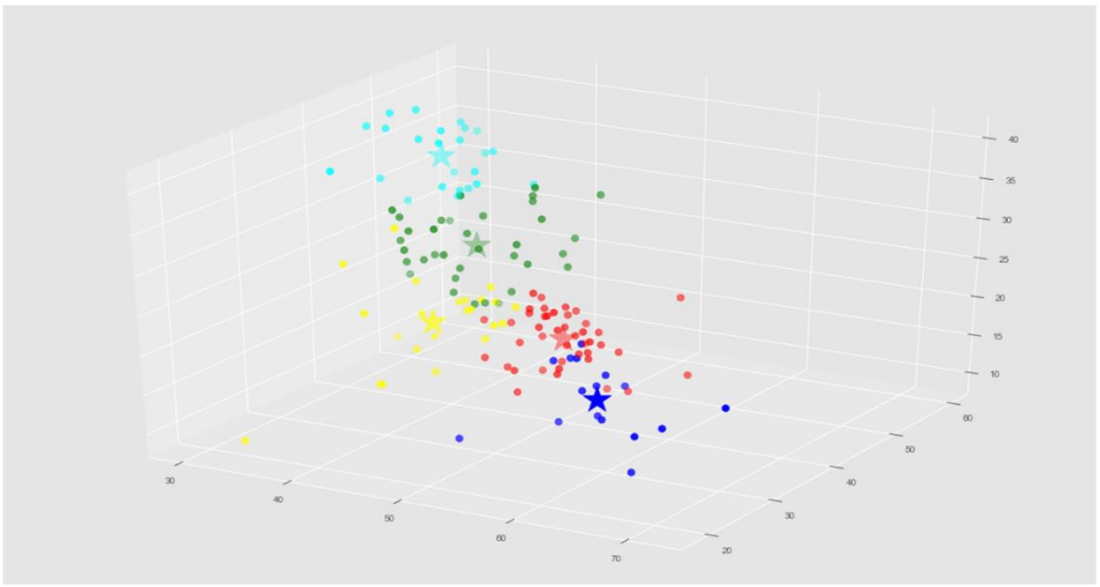
El resultado muestra que *X* tiene una forma de (140, 3), lo que indica que consta de 140 filas y 3 columnas. Después de realizar el método del codo para identificar un $k = 5$, ejecutamos el algoritmo para obtenemos las etiquetas y los centroides:

$$k - Means = K - Means(n\_clusters = 5).fit(X)$$
$$centroids = k - Means.cluster\_centers\_$$
$$print(centroids)$$

Figura 13

Obtención de las etiquetas y los centroides

```python
# Predicting the clusters
labels = kmeans.predict(X)
# Getting the cluster centers
C = kmeans.cluster_centers_
colores=['red','green','blue','cyan','yellow']
asignar=[]
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar,s=60)
ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c=colores, s=1000)
```



Nota. De *K-Means en Python paso a paso*, por Aprende Machine Learning., 2018 (https://n9.cl/9bs8j)..

2.2.3.2. K-Medoids

Kassambara, A. (2017) explica que se trata de un enfoque de agrupación diseñado para clasificar un grupo de datos en *k* grupos, donde cada clúster se representa mediante uno de sus puntos de datos, conocido como el "medoide" del grupo. El concepto de "medoide" se refiere a un elemento dentro de un grupo que tiene la menor diferencia promedio con respecto a todos los demás elementos del mismo grupo. Este elemento se encuentra en el centro del grupo y puede considerarse como un ejemplo típico de los miembros de ese grupo, lo que puede ser beneficioso en varias circunstancias. A diferencia de K-Means, donde el centro de un clúster se calcula como

el valor promedio de todos los datos, K-Medoids utiliza medoides como centros de clúster, lo que lo hace más robusto ante ruido y valores atípicos.

Este algoritmo requiere que se identifique el número de clústeres a generar, el cual suele hallarse a través del método de la silueta que es el más utilizado en este enfoque, asimismo resalta que uno de los métodos de K-Medoids más comunes es el algoritmo PAM (Partitioning Around Medoids). (p.48)

Asimismo, Das, A. (2022) expone que PAM (Partición Alrededor de Medoides) es un método de agrupación que transforma cada paso del proceso en un problema de estimación estadística, logrando así reducir la complejidad computacional de un grupo de datos de tamaño *n*. Los principios fundamentales de PAM abarcan los siguientes aspectos:

- Encontrar un conjunto de K-Medoides, donde *K* representa el número de clústeres y *M* corresponde a la colección de medoides, a partir de un conjunto de datos con n registros.
- Emplear una métrica de distancia, que puede ser la euclidiana, manhattan, u otras, para identificar medoides que minimicen la distancia total de los puntos de datos con respecto a su Medoide más próximo.
- Posteriormente, se lleva a cabo el intercambio de pares formados por medoides y no medoides, con el propósito de reducir la función de pérdida *L* entre todas las combinaciones posibles de pares k (n – k).

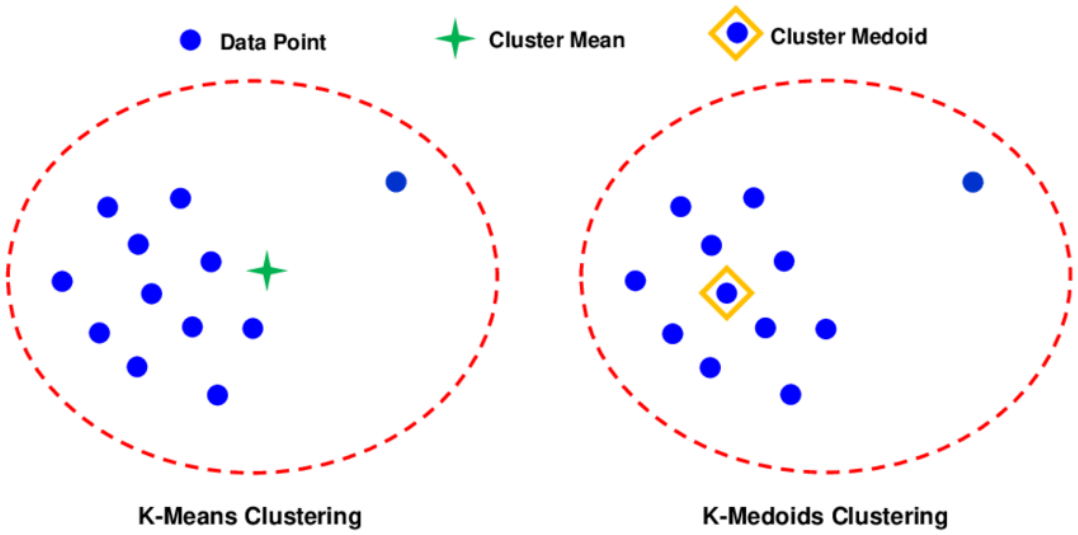
En el caso del algoritmo K-Means, en lugar de calcular simplemente la media de todos los puntos en el clúster, el algoritmo PAM adopta un enfoque diferente para actualizar el centroide. Si hay *m* puntos en un clúster, el centroide anterior se reemplaza por cada uno de los otros (*m-1*) puntos de manera sucesiva, y se selecciona el punto que minimiza la pérdida como el nuevo centroide del clúster. De esta manera, el autor comenta que la pérdida mínima se calcula mediante la siguiente función de costo:

$$L(M) = \sum^{n} i = 1 \min_{mE} M\, d(m, x_i)$$

En este sentido, la principal diferencia entre los métodos K-Means y K-Medoids radica en cómo determinan el punto central o representativo de un clúster. Ya que para K-Means, el punto central de un clúster se calcula como la media de todos los registros dentro del clúster. Mientras que, en K-Medoids, para identificar el punto central de un clúster se selecciona uno de los puntos de datos reales dentro del clúster como el representante, y este punto se denomina "Medoide" del clúster, para el cual la suma de las distancias entre él y todos los demás puntos del clúster es la mínima.

Figura 14

Representación gráfica de la diferencia entre los métodos de agrupamiento K-Means y K-Medoids



Nota. De *An innovative hybrid strategy for structural health monitoring by modal flexibility and Clustering methods*. Journal of Civil Structural Health Monitoring, por Entezami, A.; Sarmadi, H. & Razavi, B., 2020.

A fin de implementar K-Medoids en Python, se necesita utilizar la biblioteca *scikit-learn-extra*, el módulo específico a importar se realiza a través del siguiente código “from sklearn_extra.Clúster import K-Medoids”. Esta información se detalla en el portal Eduative(s.f) donde menciona que a través del algoritmo K-Medoids de *scikit-learn-extra*, es posible ajustar los puntos de datos proporcionados:

Figura 15

Algoritmo K-Medoids de scikit-learn-extra

```
!pip install scikit-learn-extra
from sklearn_extra.cluster import KMedoids
import numpy as np
```

```
np.asarray([[6, 9], [4, 10], [4, 4], [8, 5], [8, 3], [5, 2], [5, 8], [6, 4], [4, 8], [3, 9]])
```

```
k=2
kmedoids = KMedoids(n_clusters=k).fit(data)
```

Nota. De *Educative Answers - trusted answers to developer questions*, por Educative., (s.f.).

De igual manera, presenta los clústeres finales y sus medoides, así como los medoides finales:

Figura 16

Clústeres finales y sus medoides

```
clusters = kmedoids.cluster_centers_
medoids_final = data[kmedoids.medoid_indices_]

print(medoids_final)
print(clusters)
```

Cluster 0:
$[[6, 9], [8, 5], [8, 3], [6, 4], [3, 9]]$

Cluster 1:
$[[4, 10], [4, 4], [5, 2], [5, 8], [4, 8]]$

- The final medoids will be $(6, 9)$ and $(4, 4)$.

Nota. De *Educative Answers - trusted answers to developer questions*, por Educative., (s.f.).

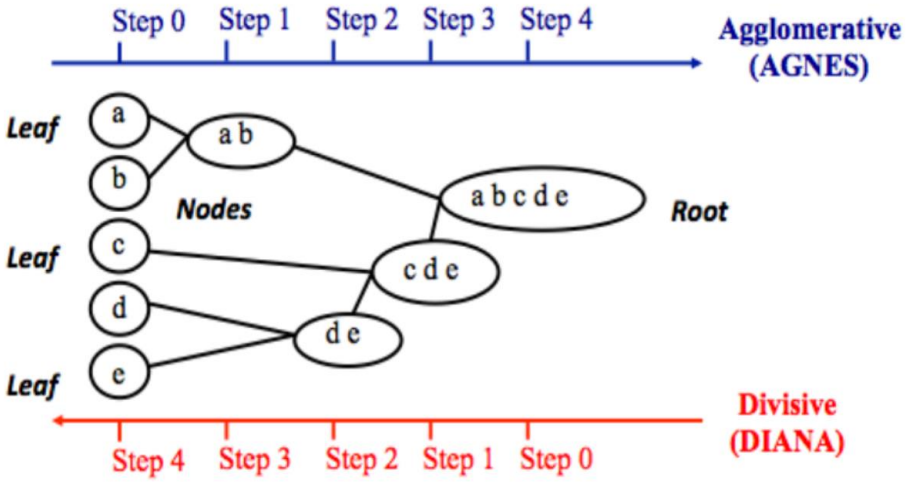
2.2.3.3. Clúster Jerárquico

Según Alam, A. et al. (2023) consideran al clúster Jerárquico como una técnica de análisis de grupos que reúne objetos similares en clústeres, siendo cada uno distinto de los demás. Hay dos tipos principales de Clúster Jerárquico: Aglomerativo (AGNES) y el Divisivo (DIANA). El primero es especialmente eficaz para identificar clústeres pequeños, mientras que el segundo es más adecuado para identificar clústeres más grandes.

En el enfoque aglomerativo, el proceso empieza tratando a cada objeto como un clúster individual y luego fusiona sucesivamente los clústeres más similares para formar grupos más grandes, hasta que todos los objetos sean miembros de un solo clúster grande. Por otro lado, el divisivo sigue un enfoque de arriba hacia abajo, comenzando con un solo clúster grande que se divide sucesivamente en clústeres más pequeños, hasta que cada objeto esté en su propio clúster individual. A fin de poder explicar lo antes mencionado, se presenta la figura17.

Figura 17

Clúster Jerárquico



Nota. *De International Journal of Creative Research Thoughts (IJCRT). Data Clustering: Prospects & Challenges 11*. 2320-2882, por Alam, A.; Esa, H.; Nazrul, K. & Hossain, A., 2023.

En este sentido, los autores Alam, A. et al. (2023) explican que el clúster jerárquico aglomerativo (AGNES) comienza considerando cada punto de datos como un clúster individual y gradualmente fusiona clústeres hasta formar un único grupo que contiene todos los puntos de datos. Los pasos del algoritmo son los siguientes:

- Primero, se considera cada punto de datos como un clúster individual, teniendo inicialmente tantos clústeres como puntos de datos.

- Luego, se realiza un cálculo sobre las distancias entre pares de clústeres o puntos de datos en el conjunto de datos. Se utilizan métricas comunes como la distancia euclidiana, la distancia de Manhattan u otras medidas apropiadas según el tipo.
- Se identifican los dos clústeres (o puntos de datos) con la distancia más pequeña entre ellos y los fusiona en uno solo. Esto reduce el número total de clústeres. Posteriormente, se recalculan las distancias entre el nuevo clúster formado y todos los demás clústeres restantes.
- Se repiten los pasos anteriores de manera iterativa hasta que quede un solo clúster que contenga todos los puntos de datos y se construye una estructura jerárquica en forma de árbol llamada dendrograma, que representa visualmente el proceso de fusión. (p.531)

Camacho, J. (2021) explica los pasos para implementar un algoritmo de clúster jerárquico, para este caso, utiliza a manera de ejemplo una base de 200 datos con las siguientes variables género, edad e ingreso anual; el nombre del archivo es "Clientes_Tienda.csv". En este sentido, se importan las librerías de la siguiente manera:

Figura 18

Importación de bibliotecas

```
# Clustering Jerárquico
# Importacion de librerias
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Carga del conjunto de datos
dataset =
pd.read_csv('Clientes_Tienda.csv')
X = dataset.iloc[:, [3, 4]].values
```

Nota. De *Clustering Jerárquico con Python | JacobSoft. JacobSoft*, por Camacho, J., 2021.

Asimismo, se presenta los comandos para crear el dendograma con la finalidad de encontrar el número óptimo de clústeres y gráfico resultante.

Figura 19

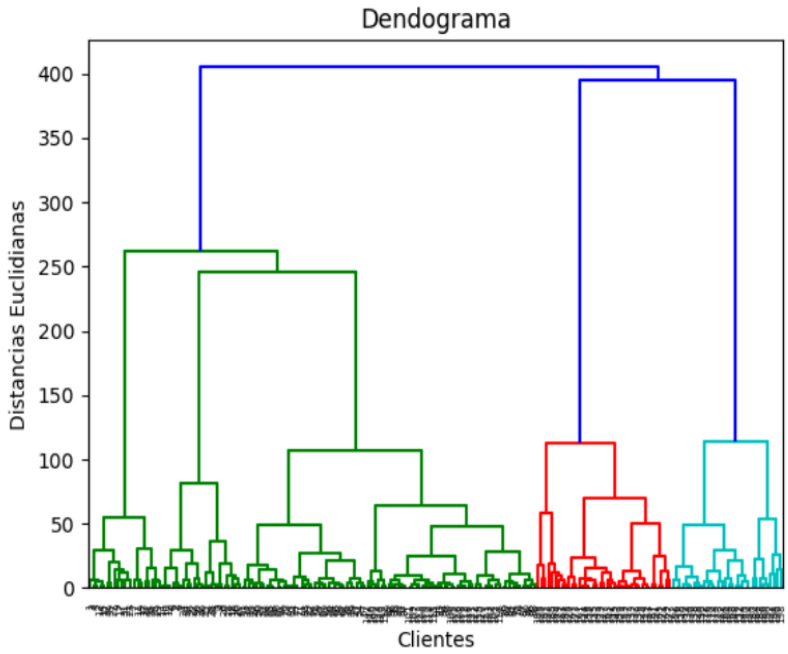
Comando de Dendograma

```
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(X,
method = 'ward'))
plt.title('Dendograma')
plt.xlabel('Clientes')
plt.ylabel('Distancias Euclidianas')
plt.show()
```

Nota. De *Clustering Jerárquico con Python | JacobSoft. JacobSoft*, por Camacho, J., 2021.

Figura 20

Gráfico de Dendograma



Nota. De *Clustering Jerárquico con Python | JacobSoft. JacobSoft*, por Camacho, J., 2021.

Con respecto al análisis a partir del dendrograma, se identificaron 5 clústeres marcados con números. Estos grupos se generaron utilizando el método aglomerante con la clase *AgglomerativeClustering* de la biblioteca *sklearn.Cluster*.

Figura 21

Comando para análisis de clúster

```
# Ajustando Clustering Jerárquico al conjunto de datos
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters=5,
                             affinity='euclidean',
                             linkage='ward')
y_hc = hc.fit_predict(X)
```

Nota. Elaboración propia, 2023.

Además, Camacho, J. (2021) indica que cada cliente o fila del conjunto de datos se asignó a uno de los 5 grupos, y esta asignación se almacenó en la variable "*y_hc*". Posteriormente, se visualizó la asignación de los 200 clientes a los 5 grupos mediante un gráfico de dispersión. Cada grupo se distinguió con un color y una etiqueta. (párr.12)

Figura 22

Agrupación de los clientes según sus características de ingresos y gastos

```
# Visualising the clusters
plt.scatter(X[y_hc == 0, 0], X[y_hc == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_hc == 1, 0], X[y_hc == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_hc == 2, 0], X[y_hc == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_hc == 3, 0], X[y_hc == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_hc == 4, 0], X[y_hc == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

Nota. De *Clustering Jerárquico con Python | JacobSoft. JacobSoft*, por Camacho, J., 2021.

2.2.3.4. Principal Component Analysis (PCA)

De acuerdo con De Silva, C. (2017), expone que el Análisis de Componentes Principales (PCA) es significativamente útil cuando se trabaja con datos complejos. PCA se emplea para reducir la cantidad de atributos o variables en un conjunto de datos que originalmente tiene muchas variables interrelacionadas.

El objetivo principal es conservar la mayor parte de la variabilidad de los datos originales mientras se reduce su dimensionalidad. PCA es un método no paramétrico que resulta útil para obtener información relevante de conjuntos de datos complejos. Este proceso se lleva a cabo mediante la transformación lineal del conjunto original de atributos en un conjunto más pequeño de atributos llamados componentes principales (PC). (p.3)

De igual manera, los autores Aguilar, L. & Vasquez Y. (2017) explican que el propósito fundamental de PCA radica en proyectar características de alta dimensionalidad (d) en un subconjunto de menor tamaño (k), al mismo tiempo que se conserva la máxima varianza de los registros obtenidos.

Se utiliza en dos aplicaciones principales: la compresión de datos, donde se representa la información en un espacio de dimensiones más reducido para minimizar la distorsión, y la extracción de características, que persigue disminuir la complejidad de los datos manteniendo la información más importante y relevante. Tanto en la compresión de datos como en la extracción de características, PCA es valioso para simplificar la representación de los datos sin sacrificar información crítica. (p. 31)

Por otro lado, Silva, C. (2017) describe los 5 pasos que se deben seguir para desarrollar el PCA:

Tabla 13

Pasos PCA

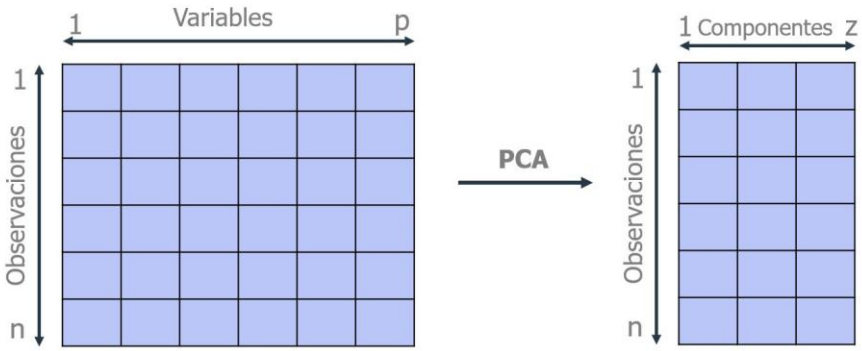
Paso	Descripción
1. Analizar y construir la matriz de datos	Construir una matriz *p* x *q*, donde *p* es el número de atributos y *q* es el número de muestras.
2. Estandarizar los datos y calcular la matriz de correlación	Procesar los datos a fin de encontrar un promedio igual a 0 y una desviación estándar equivalente a 1, luego calcular la matriz de correlación.
3. Calcular autovalores y autovectores de la matriz de correlación	Determinar los valores y autovectores de la matriz de correlación.
4. Examinar e interpretar los autovalores y autovectores	Analizar y entender el significado de los autovalores y autovectores.
5. Calcular los valores transformados y realizar análisis adicional	Obtener los valores transformados y llevar a cabo análisis adicionales, como gráficos o análisis posteriores.

Nota. De *Principal component analysis (PCA) as a statistical tool for identifying key indicators of nuclear power plant cable insulation degradation*. Iowa State University. (p.3), por De Silva, C., 2017.

Según Amat, J. (2020), para la muestra de un conjunto de datos que contenga "*n*" individuos, cada uno de ellos estará asociado a "*p*" variables (X_1, X_2, ..., X_p), lo que implica que el espacio de datos tiene p dimensiones; de esta manera, la técnica PCA identifica un conjunto menor de factores subyacentes "*z*" (donde $z < p$) que explican aproximadamente la misma variabilidad que las *p* variables originales. En lugar de requerir *p* valores para describir a cada individuo, solo sería necesario "z" valores, estas nuevas variables se denominan componentes principales. (párr. 2)

Figura 23

Representación gráfica del Análisis de Componentes Principales



Nota. De *Análisis de componentes principales PCA con Python*. Ciencia de datos, por Amat, J., 2020.

De igual manera, a fin de implementar el Análisis de Componentes Principales en Python, el portal Geek For Geeks (2023) específica que como primer paso es necesario importar las bibliotecas requeridas, que incluyen *NumPy, Matplotlib y Pandas*. Luego, se importa el conjunto de datos y se divide en componentes *X* e *Y* para su posterior análisis.

Figura 24

Importación de bibliotecas

```
# importing required libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
# importing or loading the dataset
dataset = pd.read_csv('wine.csv')

# distributing the dataset into two components X and Y
X = dataset.iloc[:, 0:13].values
y = dataset.iloc[:, 13].values
```

Nota. De *Principal Component Analysis with Python*. GeeksforGeeks, por GeeksforGeeks., 2023.

En el tercer paso, se procede a separar los datos en un conjunto de entrenamiento y un conjunto de prueba, luego, se realiza el escalado de características en ambos conjuntos, aplicando una técnica estándar a fin de que las características tengan la misma escala y no afecten negativamente el análisis. Finalmente, se aplica la función de Análisis de Componentes Principales (PCA) para ambos conjuntos para poder realizar un análisis detallado de los datos y reducir su dimensionalidad:

Figura 25

Análisis de Componentes Principales (PCA)

```
# Splitting the X and Y into the
# Training set and Testing set
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

```
# performing preprocessing part
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()

X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
# Applying PCA function on training
# and testing set of X component
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)

X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)

explained_variance = pca.explained_variance_ratio_
```

Nota. De *Principal Component Analysis with Python*. GeeksforGeeks, por GeeksforGeeks., 2023.

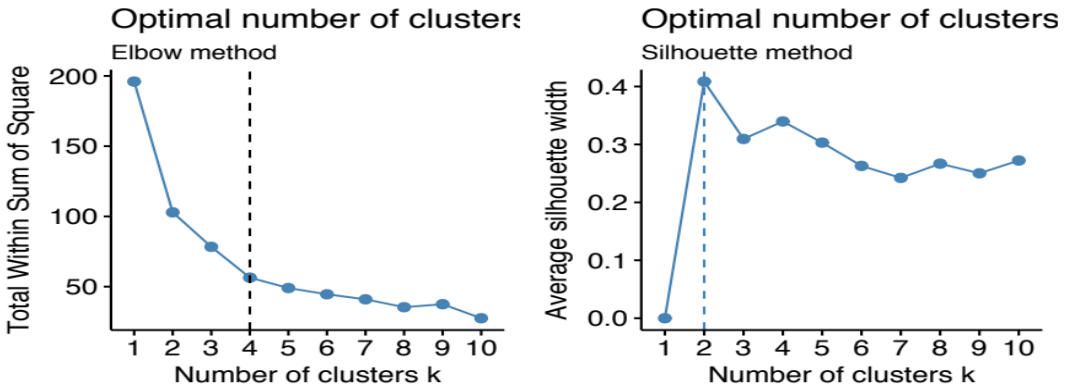
2.2.3.5. Selección del número óptimo de clústeres

Kassambara, A. (2017) explica que, para identificar la cantidad ideal de clústeres en un conjunto de datos, lo cual es un problema fundamental en su elección, se debe inspeccionar un dendograma. Este gráfico representa una solución simple y popular, producida mediante el clúster jerárquico para ver si sugiere un número específico de clústeres. En este contexto, se señala que se emplean varios enfoques para determinar el número adecuado de clústeres en algoritmos como K-Means, K-Medoids (PAM) y Clúster jerárquico. Estos métodos se dividen en dos categorías principales:

métodos directos, que optimizan criterios específicos como la reducción de la suma de cuadrados entre clústeres o la maximización del promedio de la silueta, y métodos de prueba estadística, que implican la comparación de evidencia frente a una hipótesis nula. Un ejemplo de este último enfoque es la estadística de brecha *(gap statistic).* (p.128)

Figura 26

Selección del número óptimo de Clústers



Nota. De *Practical Guide to Clúster Analysis in R: Unsupervised Machine Learning* (1st ed.). STHDA. (p.134), por Kassambara, A., 2017.

2.2.3.5.1. Método del codo

Kassambara, A. (2017) detalla que la razón para definir clústeres es a fin de minimizar la variación intraclúster total [o suma total de cuadrados es el clúster (WSS)]. Por lo que, el método del codo examina el WSS total como una función del número de clústeres. Deberíamos elegir un número de manera que agregar otro clúster no mejore significativamente el WSS total, este número se puede definir de la siguiente manera:

1. Realizar el cálculo del algoritmo de agrupamiento (por ejemplo, K-Means) para distintos valores de *k*, abarcando un rango de 1 a 10 clústeres.
2. Calcular la suma total de las distancias al cuadrado dentro de cada clúster (WSS) para cada valor de *k*.
3. Crear un gráfico que muestre la relación entre la curva de *WSS* y el número de clústeres *k*.
4. La ubicación de un punto de inflexión (o codo) en el gráfico generalmente se considera como un indicador del número óptimo de clústeres. (p.129)

En cuanto a la aplicación del método del codo en Python, según lo mencionado por el autor Rodríguez, D. (2023), se realiza mediante un proceso en el que se utiliza un bucle para iterar a lo largo de diferentes números de clústeres. En cada iteración, se entrena un modelo de K-Means con el fin de calcular la variación intra-Clúster. Posteriormente, se procede a representar los resultados de manera gráfica para identificar la ubicación del "codo". Este proceso se puede llevar a cabo por medio de la siguiente función:

Figura 27

Función del método del codo

```
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

def elbow_method(data, max_clusters=10):
    """
    Aplica el método del codo para seleccionar el número óptimo de
clusters en k-means.

    Parámetros:
    ----------
    data : matriz o matriz dispersa, forma (n_samples, n_features)
        Los datos de entrada.

    max_clusters : int, opcional (por defecto=10)
        El número máximo de clusters a considerar.
    """

    sum_of_squared_distances = []

    for k in range(1, max_clusters+1):
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(data)
        sum_of_squared_distances.append(kmeans.inertia_)
    # Graficar la curva del codo
    plt.plot(range(1, max_clusters+1), sum_of_squared_distances, 'bx-')
    plt.xlabel('Número de Clusters (k)')
    plt.ylabel('Suma de las Distancias al Cuadrado')
    plt.title('Curva del Codo')
    plt.show()
```

Nota. De *Método del codo (Elbow method) para seleccionar el número óptimo de clústeres en K-Means*. Analytics Lane. (párr.5), por Rodriguez, D., 2023.

Dentro de esta función, se realiza un bucle que varía el valor de "*k*" en el rango de 1 a *"max_Clústers"* con el propósito de entrenar un modelo y adquirir el valor del parámetro "*inertia_"*. Este parámetro es utilizado por K-Means para almacenar la suma de las distancias euclidianas al cuadrado que se generan durante el proceso de entrenamiento. Este procedimiento se puede comprobar mediante la utilización de un conjunto de datos generado de manera aleatoria.

Figura 28

Entrenamiento del modelo

```
from sklearn.datasets import make_blobs

# Genera un conjunto de datos de ejemplo con 4 clusters
data, _ = make_blobs(n_samples=500, centers=4, n_features=2,
random_state=42)

# Aplica el método del codo
elbow_method(data)
```

Nota. De *Método del codo (Elbow method) para seleccionar el número óptimo de clústeres en K-Means*. Analytics Lane. (párr.7), por Rodriguez, D., 2023.

2.2.3.5.2. Método de la silueta

Kassambara, A. (2017), evalúa la calidad de un agrupamiento al medir cuán bien cada elemento se ajusta a su respectivo grupo. Un alto elevado del ancho promedio de la silueta sugiere una alta calidad en la agrupación. El enfoque del ancho promedio de la silueta implica calcular este valor promedio para múltiples valores de *k*. El valor más adecuado de *k* es aquel que maximiza este promedio dentro de un rango de opciones posibles para *k*. Este enfoque es análogo al método del codo y se calcula de manera semejante. A fin de poder aplicar este algoritmo, se puede calcular de la siguiente manera:

- Calcular el algoritmo de clústeres para distintos valores de *k*, variando *k*, por ejemplo, de 1 a 10.
- Con respecto a cada valor de *k*, calcular el promedio de la silueta de las observaciones *(avg.sil).*
- Representar gráficamente la curva de *avg.sil* en función de *k*.

- La ubicación del valor máximo se considera como el número apropiado de Clústeres. (p.130)

En relación con la implementación de este método en Python, Ramírez, J (2021) explica a través de un ejemplo que primero se importan las bibliotecas necesarias y luego se crea un array / conjunto de datos ficticios, tal como muestra la figura 29.

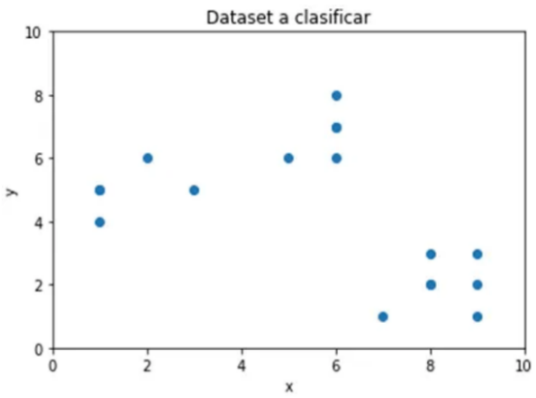
Figura 29

Importación de bibliotecas y modelado de array

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
from sklearn import metrics
from scipy.spatial.distance import cdist
from sklearn.metrics import silhouette_samples, silhouette_score
```

```
x1 = np.array([3,1,1,2,1,6,6,6,5,6,7,8,9,8,9,9,8])
x2 = np.array([5,4,5,6,5,8,6,7,6,7,1,2,1,2,3,2,3])
X = np.array(list(zip(x1,x2))).reshape(len(x1), 2)
```

```
1 plt.plot()
2 plt.xlim([0, 10])
3 plt.ylim([0, 10])
4 plt.title("Dataset a clasificar")
5 plt.xlabel("x")
6 plt.ylabel("y")
7 plt.scatter(x1, x2)
8 plt.show()
```

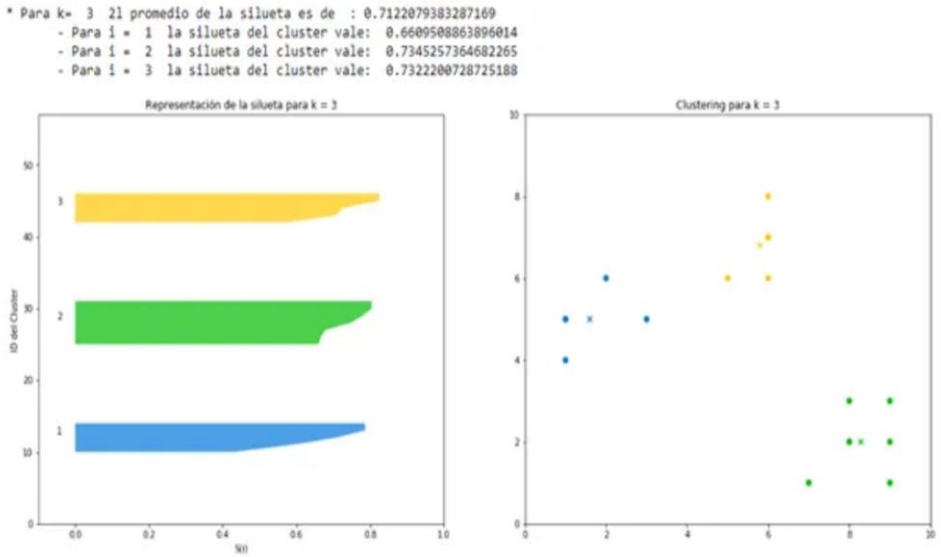


Nota. De *K-Means: Elbow method and silhouette*. Medium. (párr. 8), por Ramirez, J., 2021.

Posteriormente, se analizan los resultados conseguidos al aplicar el coeficiente de silueta. En la que desarrolla el algoritmo para un *k* igual a 2, 3 y 4. Para el ejemplo, se observa que cuando se establece $k = 3$, los coeficientes para cada grupo son muy similares entre sí, lo que lleva a la elección del número 3 como la cantidad óptima de clústeres.

Figura 30

Análisis de resultados con el K



Nota. De *K-Means: Elbow method and silhouette.* Medium. (párr. 12), por Ramirez, J., 2021.

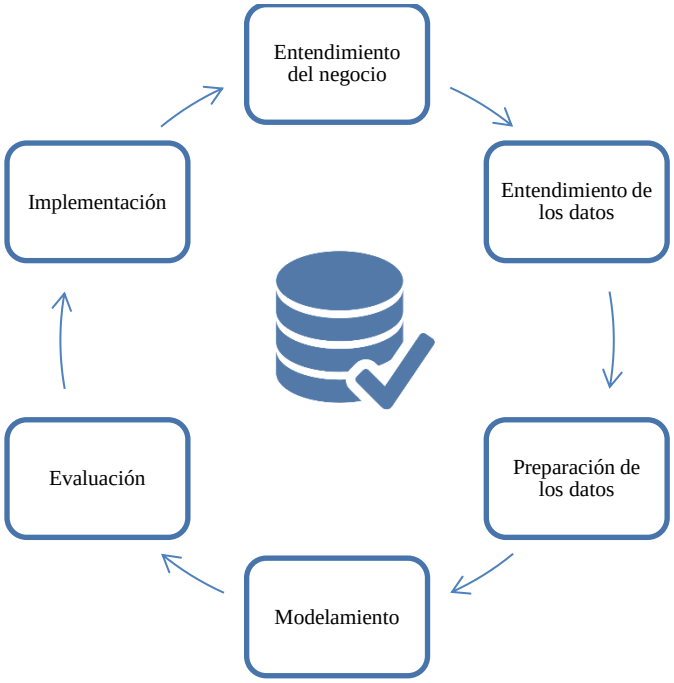
2.2.4. Metodología CRISP DM

La metodología CRISP-DM, cuyas iniciales en español significan “Proceso Estándar Transversal para la Minería de Datos” es un método utilizado en la minería de datos para organizar y sistematizar las actividades propias del análisis de datos, guiando la toma de decisiones informada y la generación de modelos predictivos o descriptivos.

Galán (2015) menciona que este método tiene su origen en las experiencias propias y también de los procedimientos estandarizados. Además, la metodología CRISP-DM comprende una guía estructurada de 6 etapas, ver figura 31, algunas de estas son reversibles, lo que significa que es posible retroceder a una fase anterior para su revisión, lo que implica que la secuencia de fases.

Figura 31

CRISP - DM



Nota. De *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*, por Galán, V., 2015.

La metodología CRISP-DM consta de las siguientes seis fases:

i. Entendimiento del negocio: En esta fase se determinan los objetivos del proyecto y se comprenden las necesidades y los objetivos del negocio. Es fundamental tener una comprensión completa del problema a resolver, así como de cómo la minería de datos puede generar valor.
ii. Entendimiento de los datos: En esta fase los datos disponibles se recopilan y analizan, se analiza la calidad de los datos y su relevancia para el proyecto. Esta etapa requiere una investigación detallada para comprender la estructura y el carácter de los datos.
iii. Preparación de los datos: Durante esta fase los datos se preparan para el modelado. Esto incluye la limpieza de datos, la transformación y la selección de características. Los datos se organizan de tal manera que los algoritmos de minería de datos puedan usarse.

iv. Modelamiento: Para crear modelos predictivos o descriptivos, en esta etapa se seleccionan y aplican técnicas de modelado. Se estudian y ajustan varios algoritmos para encontrar el mejor rendimiento. Para desarrollar un modelo efectivo basado en datos preparados, es necesaria la fase de modelado.

v. Evaluación: Los modelos creados se evalúan en función de los objetivos del proyecto. Se utilizan métricas específicas para evaluar la eficacia y la precisión del modelo. Para determinar si cumplen con las expectativas, los resultados se comparan con los requisitos del negocio.

vi. Implantación: En esta fase final los modelos evaluados y aprobados se implementan en el entorno de producción. Para su uso continuo, se integran en los sistemas existentes. Esta etapa garantiza que los resultados de la minería de datos se transformen en hechos y beneficios para el negocio.

2.2.5. Segmentación de clientes

Cuadros, A. et al. (2017) explican que el uso de técnicas de segmentación de clientes resulta importante para identificar aquellos clientes verdaderamente rentables. Esto permite enfocar recursos de manera más eficaz en estos clientes, maximizando su valor, y al mismo tiempo, utilizar los recursos de manera óptima en términos de adquisición, retención o recuperación de estos. Por lo tanto, las empresas deben segmentar a sus clientes según su capacidad de compra y otros factores como su solvencia y confiabilidad, ya que la identificación del valor y la rentabilidad del cliente posibilitan la implementación de estrategias de marketing más específicas y personalizadas. (p.42)

El autor Moreno, F. (2022) detalla que, con respecto a la estrategia de la segmentación del público objetivo, se puede establecer cuatro enfoques fundamentales:

- Vinculación: Las empresas trabajan para fortalecer sus relaciones comerciales con posibles clientes mediante estrategias que incluyen el aumento de las ventas y la prestación de servicios de alta calidad.
- Fidelización: La segmentación se utiliza para identificar a los clientes adecuados con los que se pueden establecer relaciones productivas y que brinde resultados.

- Mantenimiento: Implica mantener relaciones comerciales a lo largo del tiempo con aquellos clientes que contribuyen al equilibrio de la empresa y generan resultados positivos.
- Atracción: Es una de las principales estrategias comerciales, que consiste en atraer nuevos clientes para impulsar el crecimiento empresarial, pero esto solo es factible si se comprende a quiénes podría interesar el mensaje de la empresa. (párr. 9)

Esto brinda a las compañías una mayor comprensión de su audiencia y les facilita la adaptación de sus estrategias comerciales para satisfacer de manera precisa las necesidades individuales de cada grupo.

2.2.6. Análisis RFM

Los autores Cuadros, A. et al. (2017) describen que las estrategias de gestión de bases de datos han venido transformándose a partir de modelos sencillos como RFM (*Recency, Frequency, Monetary*), los cuales se apoyan en la recencia de las compras, la frecuencia de estas y el monto gastado. Khajvand et al. (2011) explican que es un enfoque que otorga ponderaciones a las variables *R, F y M*, teniendo en cuenta las particularidades de la industria, e incorporaron también la variable que indica la cantidad de artículos adquiridos para segmentar a los clientes. El análisis RFM se sustenta en observaciones básicas que han demostrado su relevancia en distintos sectores, como la noción de que los clientes que han realizado compras recientemente tienen una mayor propensión a responder favorablemente a las estrategias de marketing y a efectuar compras adicionales en comparación con aquellos que han permanecido inactivos durante un largo período. (p.42)

De igual manera, Santander, C. (2021) explica que la segmentación RFM (*Recency, Frequency, and Money*) es una estrategia que permite a los especialistas en marketing dirigirse a grupos específicos de clientes con comunicaciones altamente relevantes basadas en su comportamiento.

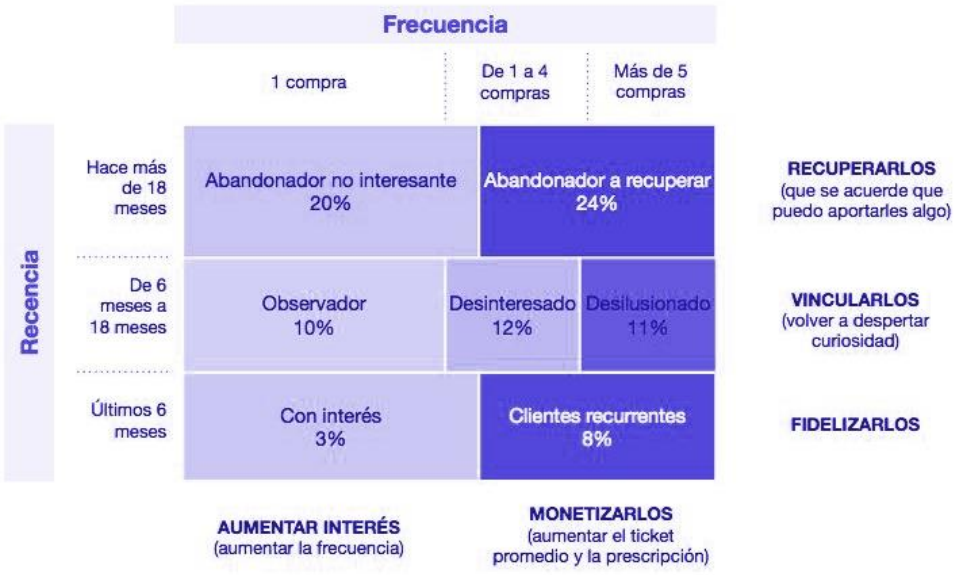
Esto conduce a tasas de respuesta más altas, una mayor lealtad del cliente y un aumento en el valor de vida del cliente. La metodología RFM se centra en tres factores cuantificables:

- *Recency* (Recencia): Se refiere a cuánto tiempo ha pasado desde la última interacción de un cliente con la empresa. Cuanto más reciente sea la actividad, más probable es que el cliente responda a aluna comunicación de la marca.
- *Frequency* (Frecuencia): Indica con qué frecuencia un consumidor interactúa o transacciona con la empresa en un período específico. Los clientes frecuentes tienden a ser más comprometidos y leales.
- *Money* (Gasto o Valor Monetario): Representa cuánto ha gastado un cliente con la marca durante un período determinado. Los clientes que realizan compras más grandes suelen recibir un tratamiento diferenciado.

El análisis RFM es popular debido a su simplicidad y facilidad de interpretación. Permite a los especialistas en marketing identificar grupos de clientes para una atención especial, utilizando datos como historial de compras, navegación web y datos demográficos. Esta segmentación se basa en escalas numéricas objetivas, lo que la hace concisa y efectiva para crear estrategias de marketing personalizadas.

Figura 32

Análisis RFM



Nota. De *Análisis RFM para segmentación de clientes (Recency, Frecuency, Money). Linkedin*, por Santander. C., 2021.

En este sentido, la técnica RFM es una estrategia de segmentación de clientes respaldada por datos que se emplea con el propósito de reconocer a los clientes más destacados según sus patrones de consumo. Esta metodología capacita a los profesionales del marketing para enfocarse en conjuntos específicos de clientes mediante mensajes adaptados a su comportamiento individual, lo que, en consecuencia, conlleva a tasas de respuesta superiores, así como a una fidelización y valor a largo plazo del cliente más sustanciales.

2.2.7. Estrategias de Crosselling y Upselling

Martín, S. (2023) explica de manera sencilla que el *upselling* es una estrategia de marketing y ventas que implica ofrecer al cliente un producto similar, pero de mayor calidad o precio al que originalmente tenía la intención de comprar. En general, esta técnica se basa en proponer un artículo más costoso que el deseado inicialmente. Por ejemplo, cuando un cliente planea comprar tres botellas de litro de Coca-Cola y, al ver una oferta de "Pack de 6 botellas al precio de 5", decide adquirir el doble de cantidad. Por otro lado, el cross-selling consiste en ofrecer al cliente productos complementarios a los que desea comprar o ya ha comprado. Por ejemplo, un vendedor sugiere a un cliente que está comprando un traje que adquiera una corbata, un cinturón o una camisa que complementen el conjunto. (párr. 10)

Figura 33

Cross-selling y Up-selling

Nota. De *Up selling y cross selling: qué son y por qué tienes que utilizarlos.* Metricool, por Martín, S., 2023.

Asimismo, los autores Kubiak, B, & Weichbroth, P. (2010) explican que el *cross-selling* y el *up-selling* son métodos ampliamente reconocidos en marketing que buscan aumentar el valor de una sola transacción de venta, mejorar la confianza del cliente y minimizar el riesgo de que la competencia capture al cliente. El *cross-selling* se enfoca en la venta de artículos relacionados o que se pueden integrar con el producto principal, mientras que el *up-selling* implica ofrecer a los clientes productos y servicios de calidad superior.

El *cross-selling* se considera una fuente de ventaja competitiva y sinergias en los negocios, especialmente en el contexto de adquisiciones. Esta estrategia brinda a los clientes existentes la oportunidad de adquirir otros artículos ofrecidos por el vendedor, a menudo complementando la compra original de alguna manera. El objetivo del *cross-selling* es capturar una mayor parte del mercado al satisfacer más necesidades y deseos de cada cliente individual. (p.3)

Además, Arévalo, B., & Aguilar, D. (2021), comparten las siguientes técnicas:

Tabla 14

Técnicas de Cross-Selling

Tipo de Estrategia	Descripción
Empaquetamiento	Esta técnica es comúnmente empleada en supermercados, donde se agrupan envases de productos complementarios, alentando a los clientes a adquirirlos en conjunto. A menudo se incluyen descuentos u ofertas en el envase para motivar la compra.
Agrupación en el Espacio	En esta estrategia, los productos relacionados se ubican físicamente cerca unos de otros en una tienda. Los clientes a menudo terminan comprando ambos productos de manera inconsciente debido a su cercanía.
Recomendaciones automáticas	Se utilizan mensajes como "Si te gustó este artículo, es posible que también te interese este otro" o "Otros productos que podrían resultarles interesantes". Cuanto mejor se conozca al cliente o se analice a través de algoritmos, más precisas serán las sugerencias.

Nota. De *Análisis de la estrategia cross–selling para mejora del comercio del Mercado Las Manuelas – Durán. Universidad de Guayaquil.* (p.14), por Arévalo, B., & Aguilar, D., 2021.

Tipo de Estrategia	Descripción
Marketing por correo electrónico	Esta técnica involucra el envío de correos electrónicos ofreciendo productos relacionados después de una compra importante.
Oferta de servicios complementarios	Ofrecer servicios como instalación, mantenimiento o entrega junto con la compra principal.
Retargeting	Segmentar el mercado utilizando data con respecto a las preferencias y datos sociodemográficos de los usuarios en sus perfiles de redes sociales. Estas estrategias de Cross-Selling se utilizan ampliamente para aumentar las ventas y mejorar la satisfacción de los mismos.

Nota. De *Análisis de la estrategia cross–selling para mejora del comercio del Mercado Las Manuelas – Durán. Universidad de Guayaquil.* (p.14), por Arévalo, B., & Aguilar, D., 2021.

Por otro lado, los autores Kubiak, B, & Weichbroth, P. (2010) detallan que el up-selling consiste en persuadir a los clientes para que consideren comprar una versión más cara del producto que ya están interesados, como optar por un televisor de 32 pulgadas en lugar de uno de 24 pulgadas. Este método se utiliza después de que el usuario ha seleccionado el artículo, pero antes de finalizar la compra. (p.3)

CAPÍTULO III: ENTORNO EMPRESARIAL

3.1 Descripción de la empresa

3.1.1. Reseña histórica y actividad económica

La compañía farmacéutica en cuestión es una organización especializada en la comercialización y distribución de medicamentos en todo el territorio peruano desde su fundación en 2008. Tiene una asociación directa con INDUFAR CISA, una empresa líder en la industria farmacéutica en Paraguay. Cuenta con un equipo de profesionales altamente capacitados y sus instalaciones están equipadas con tecnología de alta calidad. Gracias a estos estándares, la empresa ha obtenido la certificación de Buenas Prácticas de Almacenamiento otorgada por DIGEMID, lo que garantiza la calidad de sus productos.

Ofrecen una amplia gama de productos destinados a diversas especialidades médicas, como ginecología y hormonas, relajantes musculares, cardiovasculares, dermatológicos, digestivos, musculoesqueléticos, nerviosos, oftalmológicos y respiratorios. Gracias a su asociación con INDUFAR, pueden garantizar que sus productos lleguen en las mejores condiciones a cada punto de venta en todo el Perú. Con sedes en Lima, Arequipa, Trujillo y recientemente con un almacén en la zona del centro (Huancayo), para así gestionar la distribución de medicamentos a nivel nacional.

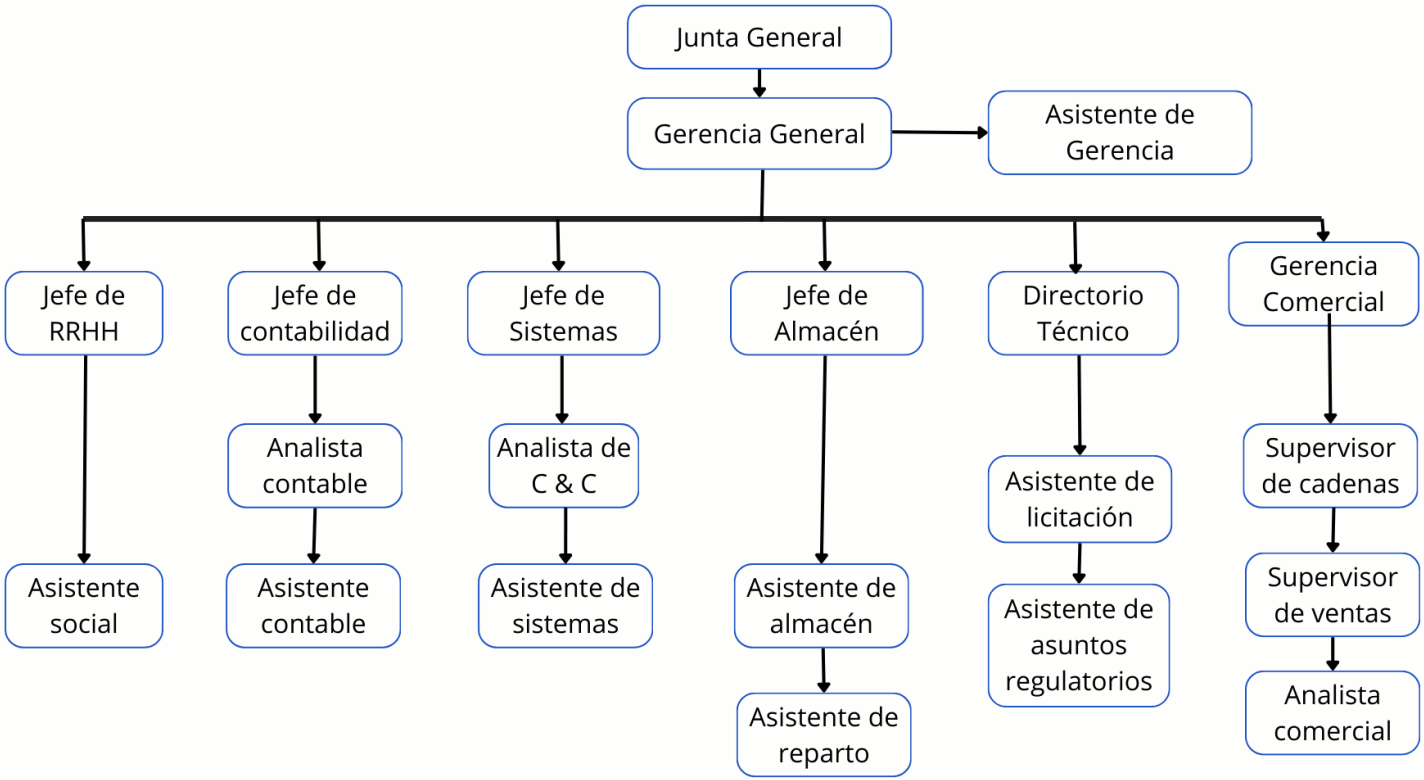
La empresa se dedica principalmente a la investigación, desarrollo y venta de productos farmacéuticos, como medicamentos y otros productos relacionados, destinados al tratamiento y prevención de enfermedades en seres humanos. Además, desempeña un rol importante en el cuidado de la salud pública y hace una importante contribución a la industria médica y científica.

3.1.2. Descripción de la organización

3.1.2.1 Organigrama

Figura 34

Organigrama de empresa del sector farmacéutico en estudio



Nota. De Empresa del sector farmacéutico, 2023.

3.1.2.1 Cadena de suministros

La empresa farmacéutica gestiona una compleja cadena de suministro que implica diversas etapas y participantes con el objetivo de garantizar la producción, distribución y entrega de productos farmacéuticos que sean tanto de alta calidad como seguros. A continuación, se ofrece una visión general de las etapas principales:

Una vez que se ha desarrollado una nueva droga o producto, comienza el proceso de fabricación a gran escala. Esto implica la producción de ingredientes activos farmacéuticos (API), formulación de medicamentos, pruebas de calidad y envasado. La fabricación debe cumplir con rigurosos estándares de calidad y regulaciones gubernamentales. El tipo de transporte (marítimo-aire) de los productos desde las instalaciones de almacenamiento (fabricación) hasta la empresa dependerá mucho de qué producto es, la cantidad y la importancia inmediata que tiene en el destino.

Después de llegar a la empresa, los productos farmacéuticos se almacenan en instalaciones especializadas que cumplen con requisitos específicos de temperatura y humedad para preservar su eficacia y seguridad. La distribución implica el transporte de estos productos desde los almacenes de la empresa hasta los puntos de venta, como farmacias y hospitales. Es importante destacar que la logística juega un papel fundamental en la cadena de suministro farmacéutica. Esto implica la gestión de inventarios, la planificación de rutas de transporte eficientes, la optimización del embalaje y la coordinación de las actividades de transporte. Todo esto se hace con el fin de asegurar la entrega a tiempo y segura de los productos farmacéuticos.

Después de que los productos llegan al mercado, la empresa farmacéutica sigue monitoreando la seguridad y eficacia de sus productos. Esto considera la recopilación y evaluación de informes de eventos adversos y la realización de estudios de seguimiento para garantizar que sus productos sigan siendo seguros para su uso a largo plazo. (Farmacovigilancia y Control de Calidad Post-Mercado).

La gestión eficiente de todas estas etapas asegura que los productos farmacéuticos sean seguros, efectivos y estén disponibles para los pacientes que los requieren. La tecnología desempeña un rol cada vez más relevante en la optimización

de estas operaciones, ya que proporciona una mayor visibilidad y control en toda la cadena de suministro farmacéutica.

Figura 35

Cadena de suministro de la empresa farmacéutica

Nota. De Empresa del sector farmacéutico, 2023.

3.1.3. Datos generales estratégicos de la empresa

La empresa ha establecido una misión, una visión y un conjunto de valores que han sido fundamentales en su identidad desde el principio. A continuación, se detallan estos elementos, así como sus metas estratégicas correspondientes.

3.1.3.1 Visión, misión y valores o principios

Visión:

Nuestra meta es convertirnos en una empresa líder destacada en Perú en la próxima década, especializada en la comercialización y distribución de productos farmacéuticos. También aspiramos a establecer una presencia sólida en diversas especialidades médicas a las que nos dirigimos en el mercado.

Misión:

Nos dedicamos a proporcionar opciones innovadoras y de alta calidad para los profesionales médicos, contribuyendo así al restablecimiento de la salud y el bienestar

de las personas. Nuestro compromiso también incluye mantener precios competitivos en el mercado.

Valores:

Calidad, Compromiso, Responsabilidad, Lealtad, Puntualidad y Confidencialidad.

3.1.3.2 Objetivos estratégicos

OE 1: Dar a conocer a nuestro mercado objetivo de la variedad de moléculas en toda la línea que ofrecemos, a través del marketing tradicional y digital.

OE 2: Ampliar nuestra presencia en boticas estratégicas, buscando una cobertura significativa del mercado nacional.

OE 3: Aprovechar los certificados de calidad para poder mejorar las entregas de los productos a través de alianzas estratégicas con empresas de distribución.

OE 4: Asignar recursos financieros para mejorar el proceso de abastecimiento eficiente de los clientes.

OE 5: Desarrollar un modelo de segmentación de clientes mediante técnicas de clúster con machine learning.

OE 6: Implementar una mejora en el plan estratégico comercial para maximizar la eficiencia del presupuesto asignado.

OE 7: Realizar un estudio de producto, verificando el margen y penetración para que a futuro no se descontinúe la molécula.

OE 8: Implementar técnicas de machine learning en el plan estratégico comercial.

OE 9: Incorporar la necesidad de los clientes de un abastecimiento eficiente en el plan estratégico comercial.

OE 10: Utilizar las técnicas de machine learning para facilitar el proceso de toma de decisiones.

OE 11: Realizar un estudio de mercado previo de moléculas para realizar compras antes del incremento de precio o escasez.

OE 12: Contar con distintos proveedores de materia, con una estandarización de calidad minuciosa.

OE 13: Destinar recursos financieros a la mejora de procesos.

OE 14: Utilizar a favor la exclusividad de la línea para hacer frente a los laboratorios multinacionales.

OE 15: Trabajar en conjunto con aliados estratégicos para innovar los procesos.

OE 16: Ofrecer formación continua a nuestro equipo de ventas y diseñar planes de crecimiento que incentiven la retención del talento.

OE 17: Realizar una estrategia comercial y de mercado para poder traer moléculas genéricas ante el alza de precios.

OE 18: Mejorar los esfuerzos de marketing mediante promociones que mitiguen el impacto del incremento de los precios.

OE 19: Reducir la rotación de personal para evitar el incremento de las brechas organizacionales con los competidores.

OE 20: Renovar el plan estratégico comercial considerando los esfuerzos de la competencia (*benchmarking*).

3.1.3.3 Evaluación interna y externa. FODA

Evaluación de factores internos:

Un puntaje ponderado total superior a 2.5 indica que la empresa tiene una posición interna sólida, mientras que un puntaje inferior a 2.5 señala debilidades internas significativas en la organización. La puntuación ponderada total promedio de 2.99 sugiere que la empresa se encuentra en una posición interna sólida y tiene fortalezas en su estructura y operaciones.

Figura 36

Evaluación de factores internos

	EVALUACIÓN FACTORES INTERNOS (EFI)			
	FORTALEZAS	**PONDERACIÓN**	**CLASIFICACIÓN**	**PUNTUACIÓN**
1	Exlusividad de la línea en el Perú	0.100	3	0.30
2	Certificados de calidad y buenas prácticas.	0.080	3	0.24
3	Distribución de los productos a nivel nacional.	0.150	3	0.45
4	Disponibilidad de recursos financieros para la mejora de procesos.	0.100	3	0.30
5	Alianza con desarrolladores de técnicas de machine learning	0.150	4	0.60

Nota. Elaboración propia, 2023.

	EVALUACIÓN FACTORES INTERNOS (EFI)			
	DEBILIDADES	**PONDERACIÓN**	**CLASIFICACIÓN**	**PUNTUACIÓN**
1	El departamento de marketing recién está teniendo mejoras.	0.070	2	0.14
2	Grupo Familiar (Toma de decisiones).	0.060	2	0.12
3	Falta de un plan estratégico comercial.	0.160	4	0.64
4	Cambio constante de personal.	0.060	1	0.06
5	Algunas moléculas por su composición tienen un precio elevando en el mercado.	0.070	2	0.14
	Total EFI:	**1.00**		**2.99**

Nota. Elaboración propia, 2023.

Evaluación de factores externos:

La puntuación ponderada total varía en una escala de 4.0 (indicando una respuesta excepcional de la organización a las oportunidades y amenazas del sector) a 1.0 (señalando una respuesta deficiente de las estrategias de la empresa a las oportunidades y amenazas del sector). Con una puntuación ponderada total promedio de 2.91, la empresa se encuentra en una posición relativamente sólida.

Figura 37

Evaluación de factores externos

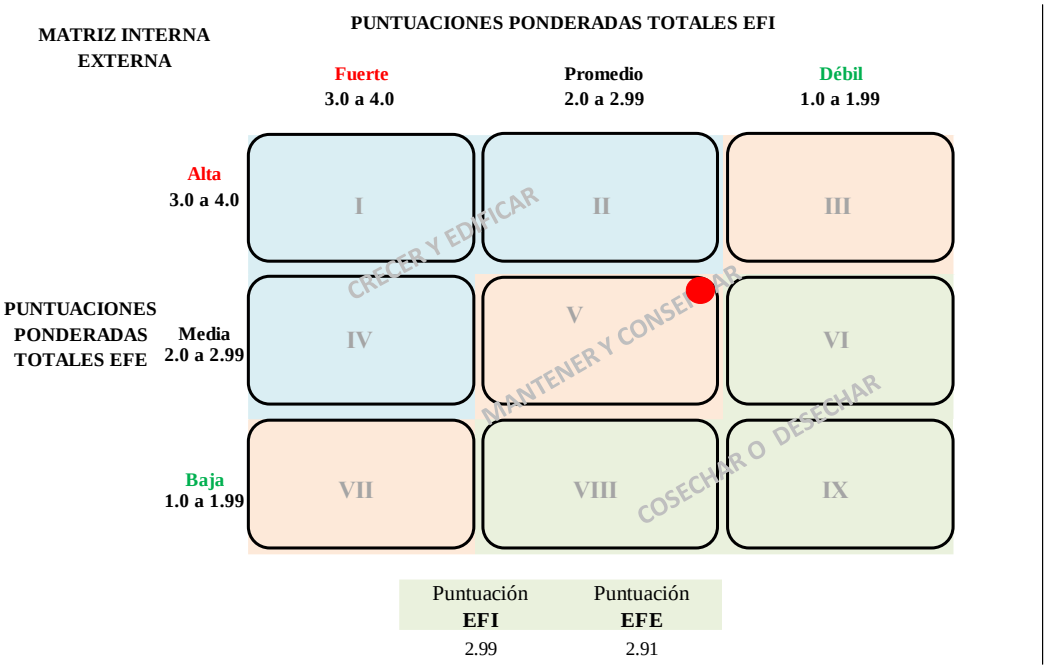
	EVALUACIÓN FACTORES EXTERNOS (EFE)			
	OPORTUNIDADES	**PONDERACIÓN**	**CLASIFICACIÓN**	**PUNTUACIÓN**
1	Crecimiento a nivel nacional.	0.150	3	0.45
2	Importación de moléculas altamente comerciales.	0.080	3	0.24
3	Productos de primera necesidad.	0.140	3	0.42
4	Clientes quieren un abastecimiento eficiente	0.050	3	0.15
5	Técnicas de machine learning avanzadas.	0.190	4	0.76
	AMENAZAS	**PONDERACIÓN**	**CLASIFICACIÓN**	**PUNTUACIÓN**
1	Incremento de precios de los productos.	0.080	2	0.16
2	Sanciones de los organismos fiscalizadores.	0.050	1	0.05
3	Competidores que buscan innovar en la gestión de sus procesos.	0.160	3	0.48
4	Situaciones imprevistas en el país de origen (problemas con la producción).	0.050	1	0.05
5	Laboratorios farmacéuticos multinacionales.	0.050	3	0.15
	Total EFE:	**1.00**		**2.91**

Nota. Elaboración propia, 2023.

Matriz Interna Externa (IE):

Figura 38

Matriz Interna Externa (IE)



Nota. Elaboración propia, 2023.

La puntuación ponderada total de EFI se ubica en el eje "*X*=2.99", mientras que la puntuación ponderada total de EFE se encuentra en el eje "*Y*=2.91", colocándose en el cuadrante 5 que representa la categoría de conservar y mantener. Los resultados brindan la oportunidad de adaptar estrategias para penetrar el mercado y desarrollar productos. Esta área, que busca mejoras mediante machine learning y datos cuantitativos, puede beneficiarse de análisis predictivos y patrones complejos para decisiones más informadas y estratégicas.

Matriz FODA:

La matriz FODA funciona como punto de partida para analizar la implementación de estrategias propuestas y evaluar sus costos y beneficios, lo que podría resultar en una ventaja competitiva. Es esencial comprender que esta matriz refleja la situación actual de la empresa y no debe considerarse un documento definitivo, ya que los factores internos y externos están en constante evolución.

Tabla 15

Matriz FODA

MATRIZ FODA				FORTALEZAS			DEBILIDADES
			1	Exlusividad de la línea en el Perú	1		El departamento de marketing recién está teniendo mejoras.
			2	Certificados de calidad y buenas prácticas.	2		Grupo Familiar (Toma de decisiones).
			3	Distribución de los productos a nivel nacional.	3		Falta de un plan estratégico comercial.
			4	Disponibilidad de recursos financieros para la mejora de procesos.	4		Cambio constante de personal.
			5	Alianza con desarrolladores de técnicas de machine learning	5		Algunas moléculas por su composición tienen un precio elevando en el mercado.
OPORTUNIDADES		**O**	**F**	*ESTRATEGIAS FO*	**O**	**D**	*ESTRATEGIAS DO*
1	Crecimiento a nivel nacional.	3	1	Dar a conocer a nuestro mercado objetivo de la variedad de moléculas en toda la línea que ofrecemos, a través del marketing tradicional y digital.	1	3	Implementar una mejora en el plan estratégico comercial para maximizar la eficiencia del presupuesto asignado.
2	Importación de moléculas altamente comerciales.	3	3	Ampliar nuestra presencia en boticas estratégicas, buscando una cobertura significativa del mercado nacional.	2	1	Realizar un estudio de producto, verificando el margen y penetración para que a futuro no se descontinúe la molécula.
3	Productos de primera necesidad.	4	2	Aprovechar los certificados de calidad para poder mejorar las entregas de los productos a través de alianzas estratégicas con empresas de distribución.	5	3	Implementar técnicas de machine learning en el plan estratégico comercial.
4	Clientes quieren un abastecimiento eficiente	4	4	Asignar recursos financieros para mejorar el proceso de abastecimiento eficiente de los clientes.	4	3	Incorporar la necesidad de los clientes de un abastecimiento eficiente en el plan estratégico comercial.
5	Técnicas de machine learning avanzadas.	5	5	Desarrollar un modelo de segmentación de clientes mediante técnicas de clustering con machine learning.	5	2	Utilizar las técnicas de machine learning para facilitar el proceso de toma de decisiones.

Nota. Elaboración propia, 2023.

MATRIZ FODA		FORTALEZAS			DEBILIDADES		
		1		Exlusividad de la línea en el Perú	1		El departamento de marketing recién está teniendo mejoras.
		2		Certificados de calidad y buenas prácticas.	2		Grupo Familiar (Toma de decisiones).
		3		Distribución de los productos a nivel nacional.	3		Falta de un plan estratégico comercial.
		4		Disponibilidad de recursos financieros para la mejora de procesos.	4		Cambio constante de personal.
		5		Alianza con desarrolladores de técnicas de machine learning	5		Algunas moléculas por su composición tienen un precio elevando en el mercado.
AMENAZAS		A	F	*ESTRATEGIAS FA*	A	D	*ESTRATEGIAS DA*
1	Incremento de precios de los productos.	1	3	Realizar un estudio de mercado previo de moléculas para realizar compras antes del incremento de precio o escasez.	5	4	Ofrecer formación continua a nuestro equipo de ventas y diseñar planes de crecimiento que incentiven la retención del talento.
2	Sanciones de los organismos fiscalizadores.	4	2	Contar con distintos proveedores de materia, con una estandarización de calidad minuciosa.	1	3	Realizar una estrategía comercial y de mercado para poder traer moléculas genéricas ante el alza de precios.
3	Competidores que buscan innovar en la gestión de sus procesos.	3	4	Destinar recursos financieros a la mejora de procesos.	1	1	Mejorar los esfuerzos de marketing mediante promociones que mitiguen el impacto del incremento de los precios.
4	Situaciones imprevistas en el país de origen (problemas con la producción).	5	1	Utilizar a favor la exclusividad de la línea para hacer frente a los laboratorios multinacionales.	3	4	Reducir la rotación de personal para evitar el incremento de las brechas organizacionales con los competidores.
5	Laboratorios farmacéuticos multinacionales.	3	5	Trabajar junto con los aliados estratégicos para innovar los procesos de gestión.	3	3	Renovar el plan estratégico comercial considerando los esfuerzos de la competencia (benchmarking).

Nota. Elaboración propia, 2023.

3.2 Modelo de negocio actual (CANVAS)

Figura 39

Modelo de negocio actual (CANVAS)

<table>
<tr>
<td rowspan="2">SOCIOS CLAVE

Proveedores:

INDUFAR CISA
GEMÜSELAB

Logistics

Bancos
Colegio médico
Cadenas famacéuticas
Red de obstetras</td>
<td>ACTIVIDADES CLAVE

Importaciones de productos.
Diseño del package

Ventas

Compras
Campaña médica</td>
<td rowspan="2">PROPUESTA DE VALOR

Nuestro objetivo es situar nuestros productos de manera óptima en el mercado a través de una estrategia comercial que refuerce nuestra ventaja competitiva. Lo que nos distingue de la competencia es la exclusividad que mantenemos con la línea INDUFAR en Perú. Además, ofrecemos precios competitivos gracias a una estrategia de negociación directa con los proveedores en el lugar de origen y con las agencias de aduanas en el destino, lo que nos permite reducir costos y ofrecer nuestros productos de manera asequible para nuestros clientes.</td>
<td>RELACION CLIENTE

Rapidez de entrega

Política de canje

Asesoramiento a los distintos segmentos</td>
<td rowspan="2">CLIENTES

Clientes principales:

Distribuidoras nacionales
Cadena de farmacias

Mayoristas Capón

Cliente final:
Población peruana.</td>
</tr>
<tr>
<td>RECURSOS CLAVE
SAP
Flota para envios locales
Recursos humanos</td>
<td>CANALES
Comunicación directa
Vía electrónica (email)
Teléfono

Web, blog y redes sociales</td>
</tr>
<tr>
<td colspan="3">ESTRUCTURA DE COSTOS
Costo de mantenimiento
Depreciación de equipos
Almacén
Costo financiero
Personal
Medicamentos</td>
<td colspan="2">FUENTES DE INGRESO

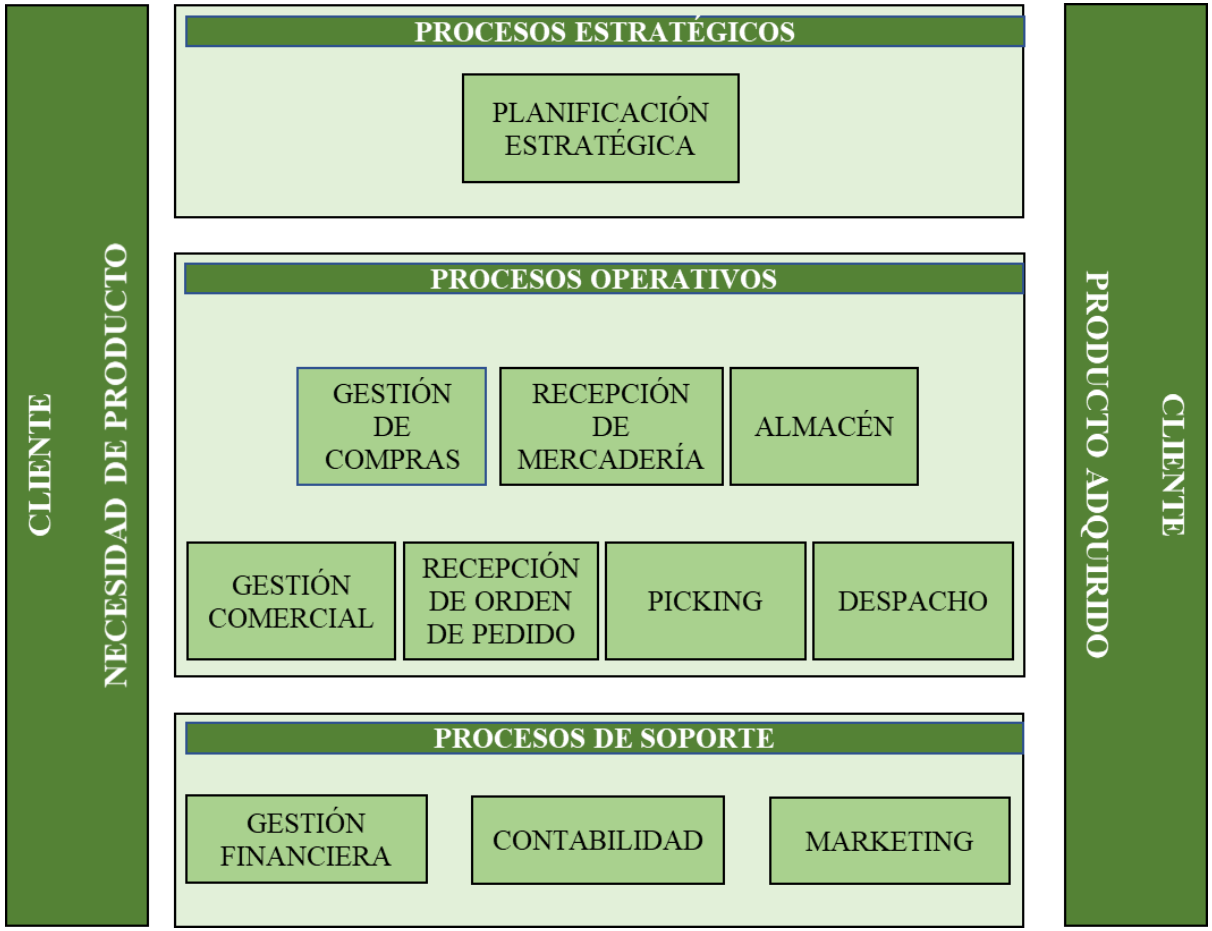
Ventas de productos finales (medicamentos).</td>
</tr>
</table>

Nota. Elaboración propia, 2023.

3.3 Mapa de procesos actual

Figura 40

Mapa de procesos actual



Nota. Elaboración propia, 2023.

Gestión de Compras:

En este punto, se establece la conexión entre la empresa y sus proveedores principales. Normalmente, este procedimiento comienza con la solicitud específica de un producto por parte de la empresa. Una vez que el pedido es recibido y aprobado por los proveedores, ellos envían la factura correspondiente. Una vez que la empresa realiza el pago, la mercancía se envía directamente al almacén de la empresa.

Recepción de mercadería:

Este procedimiento representa el cierre del acuerdo con un proveedor después de emitir una orden de compra. Se centra en los pasos para recibir la mercancía adquirida en el almacén. Las acciones esenciales incluyen descargar la mercancía del transporte y llevarla a la zona de recepción del almacén. Posteriormente, se verifica la mercancía recibida en cuanto a cantidad, calidad y documentación asociada, lo cual es de vital importancia.

Almacenamiento:

En esta operación, la mercancía sigue el camino desde la zona de recepción e inspección del almacén hasta su ubicación final de almacenamiento. Las tareas clave implican recoger la mercadería en la recepción, recibir instrucciones sobre su ubicación, como pasillos específicos, transportarla a la ubicación indicada y almacenarla en ese lugar asignado.

Gestión Comercial:

Este proceso es fundamental, ya que facilita la interacción de la empresa con el mercado. Se sitúa al final del proceso, donde los canales de venta suministran los productos al mercado y, a su vez, la empresa recibe recursos financieros derivados de estas transacciones comerciales.

Recepción de orden de pedido:

Este proceso comienza cuando el operario recibe la orden de venta y da su aprobación para proceder con la ubicación de los productos.

Picking:

En este procedimiento, se realizan las acciones de extracción y transporte de productos hacia la zona de expedición. Las actividades básicas comprenden la preparación de la orden de picking, el desplazamiento por el almacén hasta los lugares de extracción, la extracción de la mercancía de su ubicación y su transporte a la zona de expedición. La duración y distancia de estos desplazamientos varían según factores como el número de líneas de pedido, la diversidad de productos y la organización zonal del almacén.

Despacho:

Este proceso representa la salida efectiva de la mercadería del almacén. Implica cargar los productos en el medio de transporte que los llevará desde las instalaciones de la empresa hasta las del cliente. Se consolida las unidades de cada pedido para optimizar el espacio en el vehículo, basándose en clientes específicos o rutas del vehículo. Se emite la documentación necesaria, como el albarán de salida, la nota de entrega y el packing-list, que acompañará a la mercancía durante su transporte.

CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

4.1 Diseño de la Investigación

4.1.1. Enfoque de la investigación

Este estudio se realizará utilizando un enfoque cuantitativo con el propósito de desarrollar un modelo que simplifique la clasificación de clientes según sus hábitos de compra. Esto permitirá diseñar estrategias de venta y promoción más eficaces en el futuro.

Se emplearán métodos de Aprendizaje No Supervisado, tales como K-Means, K-Medoids y Clúster Jerárquico. Para identificar el número ideal de agrupaciones, se empleará el enfoque del punto de inflexión, y se confirmarán los resultados a través de la consulta de un especialista en la materia.

4.1.2. Alcance de la investigación

La investigación se restringe a un enfoque correlacional debido a la necesidad de reconocer los patrones de compra de los clientes. Este enfoque facilitará la recomendación de productos según las necesidades individuales y también ayudará a comprender la relación entre las variables que se están analizando.

4.1.3. Diseño de tipo de investigación

El estudio será de naturaleza no experimental, dado que no se realizarán intervenciones directas en las variables; en cambio, estas se observarán y examinarán en su entorno original.

Respecto al diseño de la investigación, se adoptará un enfoque transversal, lo que implica que no se realizará un seguimiento a lo largo del tiempo de las variables. En cambio, se recopilarán datos en un único momento específico con el fin de proporcionar una descripción de las características de dichas variables.

4.1.4. Población y muestra

- Población: Se hace referencia a la recopilación exhaustiva de datos sobre compras de productos farmacéuticos llevadas a cabo en todo el país por la empresa del sector farmacéutico que es el foco de la investigación.

- Muestra: Hace referencia al conjunto específico de datos relacionado con las compras de productos farmacéuticos efectuadas por la empresa del sector farmacéutico sujeta a estudio durante el período comprendido entre junio y agosto de 2023.

4.1.5. Instrumentos de medida

- Instrumento: En el conjunto de datos del negocio, se utilizarán técnicas de aprendizaje no supervisado, tales como K-Means y Clúster Jerárquico. El objetivo es segmentar los datos en *K* grupos apropiados, de manera que cada dato pueda ser asignado a un grupo o clúster en función de sus similitudes o características comunes.

- Responsable: La persona a cargo del negocio será responsable de verificar y validar los resultados obtenidos mediante las técnicas mencionadas, a través de una entrevista con el propósito de elegir la categorización que más se ajuste a los requisitos y metas de la empresa.

4.2 Metodología de implementación de la solución

En este capítulo, se proporciona una descripción a detalle del enfoque metodológico diseñado para realizar la segmentación de clientes en la empresa del sector farmacéutico que ha sido elegida. Este método se basará en el reconocido proceso de minería de datos CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual ha sido adaptado de manera específica para cumplir con las exigencias particulares de nuestro proyecto.

El principal propósito de esta investigación consistirá en identificar las características distintivas de cada grupo de clientes, lo que posibilitará el desarrollo de ofertas valiosas y estrategias comerciales eficaces. La utilización de técnicas avanzadas de Machine Learning en las etapas posteriores desempeñará un papel esencial para lograr este propósito

Figura 41

Metodología CRISP DM para el estudio

Nota. Elaboración propia, 2023.

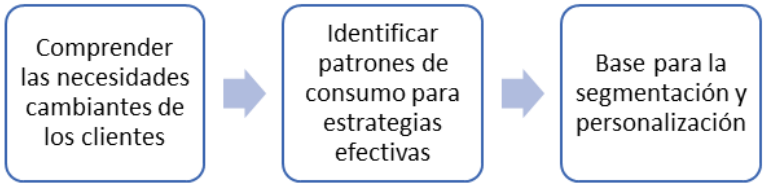
A continuación, se describirá en detalle cada fase del proceso de implementación, que se centra en la metodología de aprendizaje automático no supervisado para la clasificación:

1. Entendimiento del negocio

Durante esta fase, se llevará a cabo un examen minucioso de las necesidades de los clientes con el fin de comprender sus comportamientos de compra y las fluctuaciones en la demanda del mercado farmacéutico. Este análisis proporcionará el fundamento para la segmentación de clientes y para diseñar estrategias comerciales que sean eficaces al abordar las necesidades particulares de los clientes.

Figura 42

Entendimiento del negocio - CRIPS DM



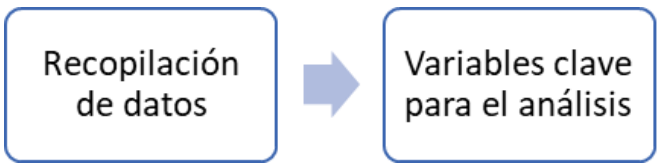
Nota. Elaboración propia, 2023.

2. Entendimiento de los datos

Durante esta etapa, se procederá a recopilar datos de la empresa del sector farmacéutico, abarcando el intervalo de junio a agosto de 2023. Para obtener información sobre las transacciones, se accederá al sistema SAP y se elegirán las variables esenciales para el análisis. Estos datos se utilizarán como entrada para aplicar las técnicas de Machine Learning.

Figura 43

Entendimiento de los datos - CRISP DM



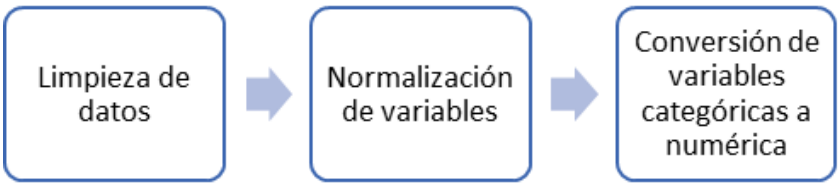
Nota. Elaboración propia, 2023.

3. Preparación de los datos

Previo a llevar a cabo la segmentación, será esencial preparar los datos de forma adecuada. Este proceso incluirá la depuración de datos para eliminar valores nulos y la normalización de variables, garantizando que todas estén en una escala numérica uniforme. Además, se verificará la integridad de los datos y se realizará la conversión de variables categóricas a formatos numéricos.

Figura 44

Preparación de los datos - CRISP DM



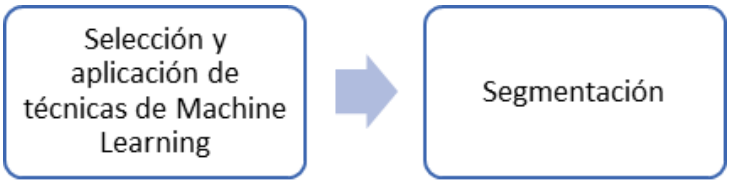
Nota. Elaboración propia, 2023.

4. Modelamiento

Durante esta fase, se elegirán y emplearán técnicas de Machine Learning como K-Means, K-Medoids y Clúster Jerárquico. Estas técnicas posibilitarán una segmentación precisa de los clientes en grupos homogéneos. El modelo obtenido proporcionará datos valiosos que serán fundamentales para desarrollar estrategias comerciales innovadoras y personalizar ofertas de valor.

Figura 45

Modelado - CRISP DM



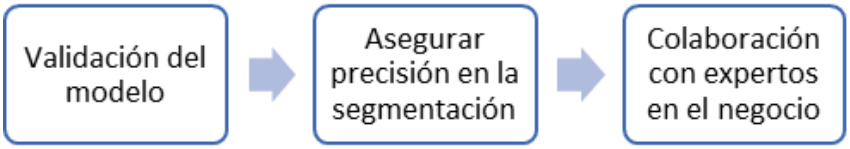
Nota. Elaboración propia, 2023.

5. Evaluación

En esta etapa, se llevará a cabo la validación del modelo de segmentación. Se examinarán minuciosamente los resultados obtenidos para garantizar la precisión en la clasificación de los clientes. Esta validación se llevará a cabo tanto internamente, evaluando la eficacia del algoritmo, como externamente, con la colaboración de expertos en el negocio que verificarán la calidad de los grupos formados.

Figura 46

Evaluación - CRISP DM



Nota. Elaboración propia, 2023.

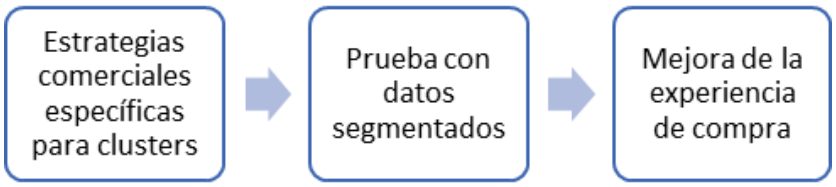
6. Implementación

Después de una evaluación profunda tanto desde un enfoque teórico como práctico, se realizará un análisis minucioso de las características distintivas de cada grupo, siguiendo el algoritmo seleccionado previamente por el experto. En esta etapa, se llevará a cabo una prueba que se extenderá a lo largo de un período específico. Durante este lapso, se utilizará la información de cada segmento de clientes para desarrollar estrategias comerciales de productos personalizadas, adaptadas a las características individuales de cada tipo.

Esta fase también implicará la realización de encuestas dirigidas a los clientes con el objetivo de evaluar si la segmentación efectivamente conduce a mejoras en la experiencia de compra, la satisfacción del cliente y, por último, el aumento de la retención de clientes.

Figura 47

Implementación - CRISP DM



Nota. Elaboración propia, 2023.

4.3 Metodología para la medición de resultados de la implementación

Una vez que se haya construido el modelo propuesto para la segmentación de clientes, es necesario verificar si el número de clústeres o grupos *K* seleccionados y formados es el correcto. Esta verificación se realizará mediante el "Método del codo (Elbow Method)" y se seguirá con una "Validación por expertos" del sector empresarial.

4.3.1. Método del Codo

Es una técnica crucial en el ámbito del aprendizaje automático para determinar la cantidad adecuada de grupos en algoritmos de clústeres como K-Means. Este es un algoritmo ampliamente utilizado en la agrupación de datos, tiene como objetivo clasificar un conjunto de puntos en "*K*" clústeres distintos. La cuestión clave es determinar cuál es el valor óptimo de "*K*".

La esencia del método del codo implica aplicar el algoritmo K-Means a varios valores de *K* dentro de un rango predefinido (por ejemplo, de 1 a 10). Para cada valor de *K*, se calcula la suma de los cuadrados del distanciamiento entre cada punto de datos y el centro de su respectivo clúster. Esta medida, también conocida como inercia o suma de cuadrados dentro del Clúster, cuantifica cuán dispersos están los puntos dentro de cada Clúster. Luego, estos valores de inercia se representan gráficamente en función del número de clústeres *K*.

La idea es observar el gráfico resultante de la inercia versus el número de clústeres. Cuando se representa, este gráfico a menudo muestra una curva que se asemeja a un codo. El punto en el cual esta curva comienza a aplanarse indica el número óptimo de agrupaciones. En otras palabras, este método tiene como objetivo descubrir el equilibrio óptimo entre la cantidad de clústeres y la cohesión interna de los datos. Identificar este "codo" es esencial para seleccionar un valor apropiado de *K*, permitiendo así una segmentación precisa y significativa de los datos en clústeres coherentes.

4.3.2. Validación de expertos

Es esencial disponer de herramientas confiables y validadas, y entre los diversos tipos de validación, se encuentra la "validación de expertos". Según Cabero y Llorente (2013), este método implica la evaluación por personas especializadas que ofrecen respuestas de calidad, profundizando en la valoración y demostrando facilidad de implementación. Este enfoque es particularmente valioso al evaluar conocimientos sobre temas complejos, novedosos o poco estudiados. Los expertos no solo proporcionan discernimiento sobre contenidos desafiantes, sino también informaciones valiosas relacionadas con el área de estudio. La validación de expertos se destaca como un procedimiento fundamental para garantizar la autenticidad y la relevancia de los instrumentos de evaluación, contribuyendo así a la fiabilidad de los resultados obtenidos en áreas de conocimiento complejas y especializadas.

4.4 Cronograma de actividades y presupuesto

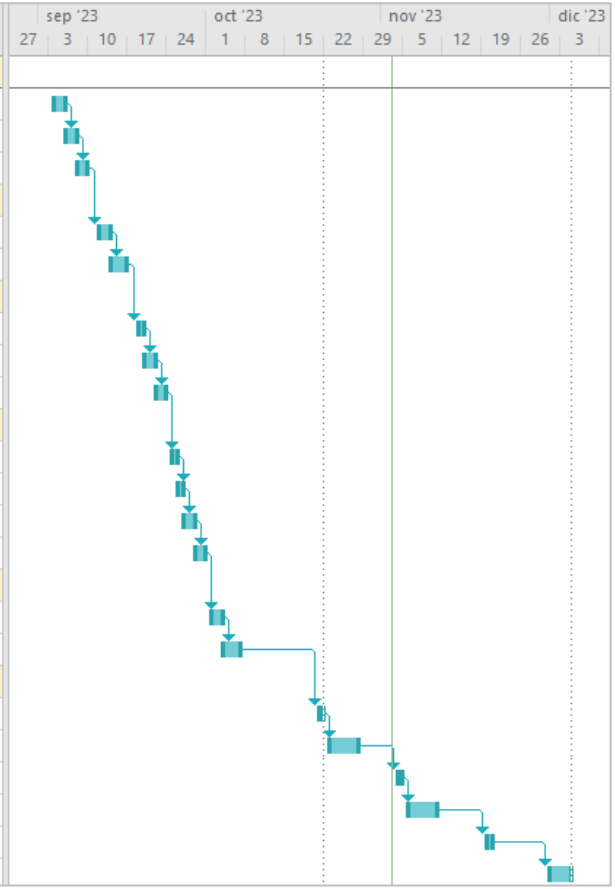
A continuación, se muestra el cronograma de actividades para nuestra investigación en la figura 48, abarcando desde la primera semana de septiembre de 2023 hasta la primera semana de diciembre. El cual considerará los 6 capítulos desarrollados:

1. Planteamiento del problema
2. Marco Teórico
3. Entorno empresarial
4. Metodología de la investigación
5. Desarrollo de la solución
6. Conclusiones

Figura 48

Cronograma de actividades

Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Capítulo I: Planteamiento del problema				
1.1 Descripción de la realidad problemática	2 días	lun 4/09/23	mar 5/09/23	
1.2 Justificación de la investigación	2 días	mié 6/09/23	jue 7/09/23	2
1.3 Delimitación de la investigación	2 días	vie 8/09/23	sáb 9/09/23	3
Capítulo II: Marco teótico				
2.1 Antecedentes de la investigación	2 días	mar 12/09/23	mié 13/09/23	4
2.2 Bases teóricas	3 días	jue 14/09/23	sáb 16/09/23	6
Capítulo III: Entorno empresarial				
3.1 Descripción de la empresa	1 día	mar 19/09/23	mar 19/09/23	7
3.2 Modelo de gestión actual (CANVAS)	2 días	mié 20/09/23	jue 21/09/23	9
3.3 Mapa de procesos actual	2 días	vie 22/09/23	sáb 23/09/23	10
Capítulo IV: Metodología de la investigación				
4.1 Diseño de la investigación	1 día	lun 25/09/23	lun 25/09/23	11
4.2 Metodología de la implementación de la solución	1 día	mar 26/09/23	mar 26/09/23	13
4.3 Metodología de la medición de resultados	2 días	mié 27/09/23	jue 28/09/23	14
4.4 Cronograma de actividades y presupuesto	2 días	vie 29/09/23	sáb 30/09/23	15
Capítulo V: Desarrollo de la solución				
5.1 Propuesta de solución	2 días	lun 2/10/23	mar 3/10/23	16
5.2 Medición de la solución	3 días	mié 4/10/23	vie 6/10/23	18
Trabajo de Suficiencia Profesional				
6.1 Asesoria	1 día	sáb 21/10/23	sáb 21/10/23	19
6.2 Levantamiento de observaciones	5 días	lun 23/10/23	vie 27/10/23	21
6.3 Asesoria	1 día	sáb 4/11/23	sáb 4/11/23	22
6.4 Levantamiento de observaciones	5 días	lun 6/11/23	vie 10/11/23	23
6.5 Entrega final	1 día	lun 20/11/23	lun 20/11/23	24
6.6 Sustentación	2 días	vie 1/12/23	lun 4/12/23	25



Nota. Elaboración propia, 2023.

Además, en la Tabla 16, se detallan los costos previstos para llevar a cabo el estudio, los cuales están divididos en 3 secciones.

- Equipo y útiles de escritorio: Incluye la compra de los equipos y los suministros de oficina fundamentales necesarios para llevar a cabo el proyecto. Asimismo, equipos electrónicos, suministros de oficina y cualquier otro recurso necesario para respaldar las actividades del equipo.
- Mano de obra: Comprende los costos asociados de los miembros del equipo que trabajarán en el proyecto.
- Servicios básicos: Comprende los gastos relacionados con los servicios requeridos para la ejecución del proyecto.

Tabla 16

Presupuesto de la investigación

Presupuesto del TSP			
Item	**Cantidad**	**Valor económico (S/.)**	**Total (S/.)**
Equipos y útiles de escritorio			
Laptops	5	500	S/ 2,500.00
Mouse	5	70	S/ 350.00
Cooler para laptop	5	200	S/ 1,000.00
Cuadernos	5	10	S/ 50.00
Lapiceros	5	5	S/ 25.00
Resaltadores	5	5	S/ 25.00
Subtotal			S/ 3,950.00
Mano de obra			
Bachilleres	5	2000	S/ 10,000.00
Subtotal			S/ 10,000.00
Servicios básicos			
Luz	5	50	S/ 250.00
Internet	5	70	S/ 350.00
Subtotal			S/ 600.00
Total			**S/ 14,550.00**

Nota. Elaboración propia, 2023.

CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

5.1. Propuesta de Solución

5.1.1. Planteamiento y descripción de Actividades

El propósito de la presente investigación es implementar técnicas de aprendizaje no supervisado en Machine Learning mediante el uso del entorno "Jupyter" y el lenguaje de programación Python. Nuestro enfoque se centrará en la evaluación de los métodos de K-Means, K-Medoids y Clúster Jerárquico. A continuación, se describen las etapas del proyecto:

- Entendimiento del negocio:
  Examen exhaustivo de las demandas y comportamientos de consumo en el mercado de la salud.
  Futura creación de ofertas altamente personalizadas basadas en datos y machine learning.

- Entendimiento de los datos:
  Recopilación de la información necesaria para el modelado de los clústeres.
  Validación inicial de la calidad de los datos adquiridos.
  Definición de las variables clave que serán utilizadas en el análisis.
  Descripción detallada de cada variable.

- Preparación de los datos:
  Etapa de depuración de datos y eliminación de valores que puedan influir en los resultados finales al aplicar la técnica.

- Modelado:
  Determinación del valor óptimo de *k* utilizando el método del codo para identificar el número ideal de clústeres.
  Aplicación de técnicas de Clúster implementando K-Means, K-Medoids y Jerárquico para obtener las segmentaciones deseadas.

- Evaluación: Simulación y Análisis de Resultados
  Evaluación detallada de los resultados obtenidos, analizando las características de cada clúster y comparando las técnicas utilizadas.
  Ejecución de simulaciones para validar el comportamiento de los clústeres.

- Implementación
  Enfoque en datos claves por clúster para una categorización efectiva y beneficiosa para la empresa.
  Búsqueda en la mejora de experiencia de compra, satisfacción y retención de clientes mediante la ejecución de propuestas comerciales personalizadas basadas en perfiles validados.

5.1.2. Desarrollo de actividades. Aplicación de herramientas de solución.

5.1.2.1. Entendimiento del negocio

En esta fase, se llevó a cabo un análisis profundo de las necesidades y expectativas de la empresa farmacéutica en estudio, considerando su posición actual y sus objetivos futuros en el sector de la salud, que es altamente dinámico y cambiante. Se buscó comprender en detalle los patrones de consumo de los clientes y cómo estas demandas evolucionaban en el mercado farmacéutico en constante transformación.

La información recopilada se convirtió en un recurso valioso para la empresa, permitiéndole anticipar y adaptarse proactivamente a las cambiantes preferencias de sus clientes. Esto sentó las bases para la segmentación de clientes, un enfoque que se convertiría en una herramienta clave para crear ofertas de valor altamente personalizadas.

La segmentación, impulsada por técnicas de machine learning, permitirá a la empresa desarrollar estrategias comerciales efectivas y específicas que se ajusten a las necesidades únicas de cada segmento.

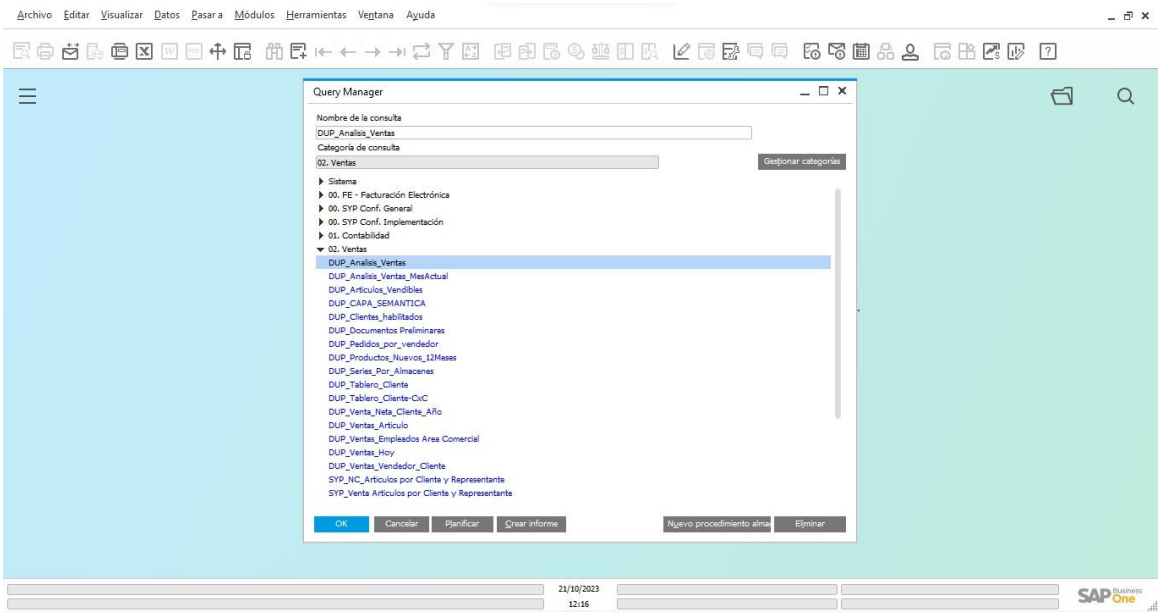
5.1.2.2. Entendimiento de los datos

La información empleada en este estudio fue recolectada a través del sistema SAP de la empresa de la empresa del sector farmacéutico en estudio. Dicha base de datos alberga transacciones comerciales específicas del periodo de junio a agosto de 2023. Para garantizar un manejo adecuado y facilitar su posterior análisis, los datos fueron extraídos y consolidados en un archivo Excel. Cabe indicar que esta data abarca diversas variables que reflejan el comportamiento comercial y transaccional de sus clientes durante esos meses, estableciendo una base sólida para el posterior proceso de segmentación y análisis. A continuación, se describen los pasos para recopilar la información mencionada:

i. En la biblioteca del Query Manager en SAP, que ofrece diversas plataformas para descargar información, se utiliza la del módulo de DUPVentas.

Figura 49

Biblioteca del Query Manager en SAP

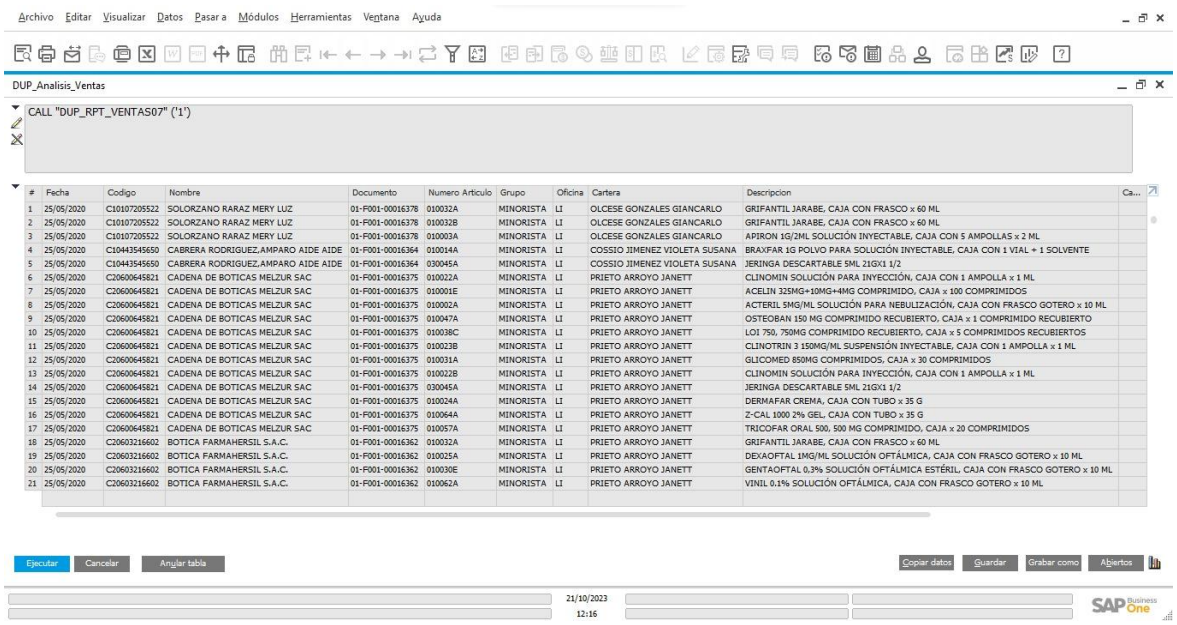


Nota. Elaboración propia, 2023.

ii. Una vez dentro de SAP, se accederá al conjunto de datos relacionados con las ventas.

Figura 50

Biblioteca del Query Manager en SAP

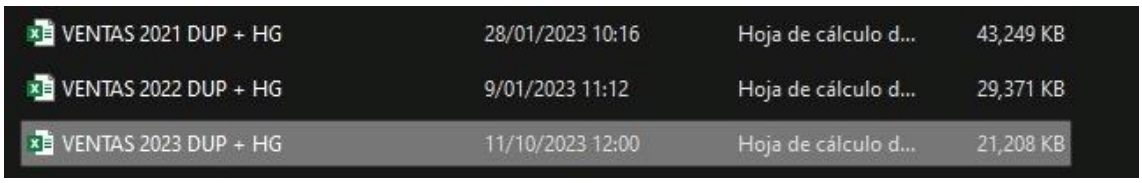


Nota. Elaboración propia, 2023.

iii. Se descarga la información desde SAP a Excel y se guarda en una ruta compartida para que los datos estén disponibles.

Figura 51

Carpeta contenedora de los dataset



Nota. Elaboración propia, 2023.

iv. En este sentido ya contamos con la información en un archivo excel para realizar las gestiones que correspondan.

Figura 52

Dataset

Vendedor	Oficina	RUC	Cliente	Grupo	Propiedad	NumReferenci	T.Documento	Fecha Contat	Codigo Articul	Articulo
RONDOY SAL	NO	C10402968341	BAUTISTA GARCIA ELIDA	MINORISTA	MINORISTA	01-F003-00023563		9/01/2023	030045A	JERINGA DESCARTABLE 5ML 21GX1 1/2
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010024A	DERMAFAR CREMA, CAJA CON TUBO x 35 G
CHAVARRY P	NO	C20532044147	DISTRIBUIDORA DROGUERI/	DISTRIBUIDC	DISTRIBUIDC	01-F003-00023564		9/01/2023	010032A	GRIFANTIL JARABE, CAJA CON FRASCO x 60 ML
RONDOY SAL	NO	C10402968341	BAUTISTA GARCIA ELIDA	MINORISTA	MINORISTA	01-F003-00023563		9/01/2023	010032A	GRIFANTIL JARABE, CAJA CON FRASCO x 60 ML
QUISPE HERN	LI	C20331066703	INRETAIL PHARMA S.A.	CADENA	CADENA	01-F001-00041212		9/01/2023	010011A	BANES FORTE 200MG/5ML SUSPENSIÓN ORAL, CA
CHAVARRY P	NO	C20532044147	DISTRIBUIDORA DROGUERI/	DISTRIBUIDC	DISTRIBUIDC	01-F003-00023564		9/01/2023	010094D	FRENAGRIP 500MG+10MG+4MG POLVO PARA SOI
RONDOY SAL	NO	C10402968341	BAUTISTA GARCIA ELIDA	MINORISTA	MINORISTA	01-F003-00023563		9/01/2023	010038E	LOI 750, 750MG COMPRIMIDO RECUBIERTO, CAJA
QUISPE HERN	LI	C20515346113	CORPORACION BOTICAS PE	CADENA	CADENA	01-F001-00041207		9/01/2023	010073A	MERIFAR 10MG/2ML SOLUCIÓN INYECTABLE, CAJ
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010003A	APIRON 1G/2ML SOLUCIÓN INYECTABLE, CAJA CC
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010032A	GRIFANTIL JARABE, CAJA CON FRASCO x 60 ML
CHAVARRY P	NO	C20532044147	DISTRIBUIDORA DROGUERI/	DISTRIBUIDC	DISTRIBUIDC	01-F003-00023564		9/01/2023	030045A	JERINGA DESCARTABLE 5ML 21GX1 1/2
DOMINGUEZ	LI	C20263805021	ESPECIALIDADES MEDICAS I	ESPECIAL	CIRCUITO CE	01-F001-00041215		9/01/2023	010048A	OTOMICIN SOLUCIÓN GOTAS ÓTICAS, CAJA CON
PALOMINO R	CE	C20486772078	DISTRIBUIDORA DROGUERI/	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041217		9/01/2023	010049A	PENBRONK 20MG COMPRIMIDO RECUBIERTO, CA
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010002A	ACTERIL 5MG/ML SOLUCIÓN PARA NEBULIZACIÓN
RONDOY SAL	NO	C10402968341	BAUTISTA GARCIA ELIDA	MINORISTA	MINORISTA	01-F003-00023563		9/01/2023	010038B	LOI 750, 750MG COMPRIMIDO RECUBIERTO, CAJA
DOMINGUEZ	LI	C20517738701	CLINICA JESUS DEL NORTE S	ESPECIAL	CIRCUITO CE	01-F001-00041205		9/01/2023	010002A	ACTERIL 5MG/ML SOLUCIÓN PARA NEBULIZACIÓN
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010023A	CLINOTRIN 3 150MG/ML SUSPENSIÓN INYECTABL
PALOMINO R	CE	C20486772078	DISTRIBUIDORA DROGUERI/	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041217		9/01/2023	010023A	CLINOTRIN 3 150MG/ML SUSPENSIÓN INYECTABL
PALOMINO R	CE	C20486213681	MIDHCO DISTRIBUCIONES S	DISTRIBUIDC	DISTRIBUIDC	01-F001-00041218		9/01/2023	010025A	DEXAOFTAL 1MG/ML SOLUCIÓN OFTÁLMICA, CAJ

Nota. Elaboración propia, 2023.

5.1.2.2.1. Presentación de variables

La base de datos configurada para este estudio incluye un total de 39,512 transacciones realizadas a nivel nacional. Este conjunto de datos abarca 24 variables, cada una de las cuales se describe en detalle para proporcionar una visión exhaustiva de la información recopilada.

Tabla 17

Variables de la base de datos

N°	ATRIBUTO DESCRIPCIÓN	DESCRIPCIÓN
1	DEPARTAMENTO	Indica la región geográfica en la que se realizó la transacción.
2	ZONA_ML	Especifica la zona o subregión dentro del departamento donde se ubica el punto de venta.
3	Vendedor	Nombre o identificación del representante de ventas responsable de la transacción.

Nota. Elaboración propia, 2023.

N°	ATRIBUTO DESCRIPCIÓN	DESCRIPCIÓN
4	Oficina	Ubicación específica o sucursal donde se efectuó la transacción.
5	RUC	Registro Único de Contribuyente del cliente.
6	Cliente	Nombre o identificación del cliente que realizó la compra.
7	Grupo	Tipo de cliente inicial asignado por la organización.
8	Propiedad	Identificación de canales.
9	NumReferencia	Número de referencia asociado a la transacción.
10	T.Documento	Tipo de documento emitido, como factura o boleta.
11	Fecha Contab	Fecha en que se contabilizó la transacción.
12	Codigo Articulo	Identificación única del artículo vendido.
N°	ATRIBUTO DESCRIPCIÓN	DESCRIPCIÓN
13	ESTATUS PRODUCTO	Estado actual del producto.
14	Articulo	Nombre o descripción del artículo vendido.
15	ARTICULO_2	Tipos de artículo.
16	Cantidad	Número de unidades vendidas en la transacción.
17	Monto	Valor total de la transacción sin IGV.
18	Lote	Número de lote asociado al producto vendido.
19	F.Venc Lote	Fecha de vencimiento del lote del producto.
20	MONTO CON IGV	Valor total de la transacción incluyendo el IGV.
21	MES	Mes en el que se realizó la transacción.
22	Año	Año en el que se realizó la transacción.
23	EMPRESA	Nombre o identificación de la empresa o marca del producto.
24	PRECIO	Precio unitario del producto vendido

Nota. Elaboración propia, 2023.

Figura 53

Base de datos en Jupyter-Python de la empresa del sector farmacéutico en estudio

```
dataset = archivo.parse("Detalle")
dataset
```

	DEPARTAMENTO	ZONA_ML	Vendedor	Oficina	RUC	Cliente	Grupo	Propiedad	NumReferencia	T.Documento
0	CALLAO	LIMA	ASTETE SAAVEDRA SANDRA DIANETT	LI	C10094266942	BEJAR ROCA DIODURA	MINORISTA	MINORISTA	01-F001-00045260	NaN
1	CALLAO	LIMA	ASTETE SAAVEDRA SANDRA DIANETT	LI	C10094266942	BEJAR ROCA DIODURA	MINORISTA	MINORISTA	01-F001-00045260	NaN
2	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	ESPECIAL	CIRCUITO CERRADO	01-F001-00045293	NaN
3	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	ESPECIAL	CIRCUITO CERRADO	01-F001-00045293	NaN
4	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	ESPECIAL	CIRCUITO CERRADO	01-F001-00045293	NaN
...	...	...	...	...	...	...	...	...	...	...
39507	Lima	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C097121297	Carazas Burgos,Juan Martin	MAYORISTA	MEDICO VISITA	03-B002-00004740	NaN
39508	LIMA	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C06671012	TORRES QUINTANA JOSE LUIS	MAYORISTA	MEDICO VISITA	03-B002-00004741	NaN
39509	Lima	LIMA	NUÑEZ CRUZ, ANA CECILIA	LI	C10624161	GUARDAMINO PAUCAR,DORIS NOEMI	MAYORISTA	OBSTETRA VISITA	03-B002-00004742	NaN
39510	Lima	LIMA	NUÑEZ CRUZ, ANA CECILIA	LI	C10624161	GUARDAMINO PAUCAR,DORIS NOEMI	MAYORISTA	OBSTETRA VISITA	03-B002-00004742	NaN
39511	Lima	LIMA	NUÑEZ CRUZ, ANA CECILIA	LI	C09824510	FRETEL ASCUE,DAVID	MAYORISTA	MEDICO VISITA	03-B002-00004744	NaN

39512 rows × 25 columns

Nota. Elaboración propia, 2023.

5.1.2.3. Preparación de los datos

En esta etapa, se realizó una meticulosa selección de las variables que serán pertinentes para nuestro estudio. Aunque contamos con un total de 24 columnas, identificamos que algunas de ellas, específicamente "DEPARTAMENTO", "Vendedor", "RUC", "Cliente", "Grupo", "NumReferencia", "T.Documento", "Fecha Contab", "Codigo Articulo", "Articulo", "Monto", "Lote", "F.Venc Lote", "MONTO CON IGV", "MES", "Año", "EMPRESA", no tienen un impacto significativo en nuestro modelo y, por lo tanto, no serán consideradas en el análisis.

En contraste, hay ciertas variables que están estrechamente relacionadas con el patrón de compra del cliente. Estas variables son "ZONA_ML", "Oficina", "Propiedad", "ESTATUS PRODUCTO", "ARTICULO_2", "Cantidad", "PRECIO". Dada su relevancia, nos centraremos en estas columnas para nuestro análisis.

Procederemos a filtrar la data para conservar únicamente las columnas mencionadas. Adicionalmente, a través del uso del comando "DROP", eliminaremos las columnas "Grupo" y "Mes". La columna "Grupo" será eliminada específicamente porque es la variable que se planea predecir.

Por otro lado, se tomará la decisión de descartar "Mes" ya que la segmentación se realizará utilizando la base de datos completa de junio a agosto, y esta columna específica no será necesaria para dicho propósito.

Figura 54

Eliminación de variable Grupo

```
df = df.drop('Grupo', axis = 1)
df
```

	DEPARTAMENTO	ZONA_ML	Vendedor	Oficina	RUC	Cliente	Propiedad	NumReferencia	T.Documento	Fecha Contab	...	ARTICULO_2
0	CALLAO	LIMA	ASTETE SAAVEDRA SANDRA DIANETT	LI	C10094266942	BEJAR ROCA DIODURA	MINORISTA	01-F001-00045260	NaN	2023-06-02	...	SUSPENSIÓN ORAL
1	CALLAO	LIMA	ASTETE SAAVEDRA SANDRA DIANETT	LI	C10094266942	BEJAR ROCA DIODURA	MINORISTA	01-F001-00045260	NaN	2023-06-02	...	SUSPENSIÓN ORAL
2	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	CIRCUITO CERRADO	01-F001-00045293	NaN	2023-06-03	...	INYECTABLES
3	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	CIRCUITO CERRADO	01-F001-00045293	NaN	2023-06-03	...	COMPRIMIDOS / TABLETAS
4	Callao	LIMA	CHAVEZ ARELLANO EVENLY PAMELA SUNLAY	LI	C20193336397	ASOCIACION EMMANUEL	CIRCUITO CERRADO	01-F001-00045293	NaN	2023-06-03	...	INYECTABLES
...	...	...	...	...	...	...	...	...	...	...	...	...
39507	Lima	LIMA	CHAVEZ ARELLANO EVENLY PAMELA	LI	C097121297	Carazas Burgos,Juan Martin	MEDICO VISITA	03-B002-00004740	NaN	2023-08-31	...	SUSPENSIÓN ORAL

Nota. Elaboración propia, 2023.

Figura 55

Eliminación de variable Mes

```
df = df.drop('MES', axis = 1)
df
```

	ZONA_ML	Oficina	Propiedad	ESTATUS PRODUCTO	ARTICULO_2	Cantidad	PRECIO
0	LIMA	LI	MINORISTA	REGULAR	SUSPENSIÓN ORAL	6	15.660567
1	LIMA	LI	MINORISTA	REGULAR	SUSPENSIÓN ORAL	6	59.979400
2	LIMA	LI	CIRCUITO CERRADO	REGULAR	INYECTABLES	15	21.594000
3	LIMA	LI	CIRCUITO CERRADO	REGULAR	COMPRIMIDOS / TABLETAS	10	39.801400
4	LIMA	LI	CIRCUITO CERRADO	REGULAR	INYECTABLES	20	17.959600
...	...	...	...	...	...	...	...
39507	LIMA	LI	MEDICO VISITA	REGULAR	SUSPENSIÓN ORAL	2	22.844800
39508	LIMA	LI	MEDICO VISITA	REGULAR	SUSPENSIÓN ORAL	50	5.841000
39509	LIMA	LI	OBSTETRA VISITA	REGULAR	INYECTABLES	12	10.065400
39510	LIMA	LI	OBSTETRA VISITA	REGULAR	INYECTABLES	6	15.871000
39511	LIMA	LI	MEDICO VISITA	REGULAR	COMPRIMIDOS / TABLETAS	10	11.339800

39512 rows × 7 columns

Nota. Elaboración propia, 2023.

Figura 56

Selección de variables de estudio

```
df = pd.DataFrame(dataset)
df
```

	ZONA_ML	Oficina	Propiedad	ESTATUS PRODUCTO	ARTICULO_2	Cantidad	PRECIO
0	LIMA	LI	MINORISTA	REGULAR	SUSPENSIÓN ORAL	6	15.660567
1	LIMA	LI	MINORISTA	REGULAR	SUSPENSIÓN ORAL	6	59.979400
2	LIMA	LI	CIRCUITO CERRADO	REGULAR	INYECTABLES	15	21.594000
3	LIMA	LI	CIRCUITO CERRADO	REGULAR	COMPRIMIDOS / TABLETAS	10	39.801400
4	LIMA	LI	CIRCUITO CERRADO	REGULAR	INYECTABLES	20	17.959600
...	...	...	...	...	...	...	...
39507	LIMA	LI	MEDICO VISITA	REGULAR	SUSPENSIÓN ORAL	2	22.844800
39508	LIMA	LI	MEDICO VISITA	REGULAR	SUSPENSIÓN ORAL	50	5.841000
39509	LIMA	LI	OBSTETRA VISITA	REGULAR	INYECTABLES	12	10.065400
39510	LIMA	LI	OBSTETRA VISITA	REGULAR	INYECTABLES	6	15.871000
39511	LIMA	LI	MEDICO VISITA	REGULAR	COMPRIMIDOS / TABLETAS	10	11.339800

39512 rows × 7 columns

Nota. Elaboración propia, 2023.

Con el fin de garantizar la integridad y exactitud de los datos, es esencial verificar la presencia de valores faltantes o nulos en las variables, ya que estos podrían influir en el resultado final del análisis. Para llevar a cabo esta revisión, se emplea el código *BD_drop.isnull().sum()*, que facilita la identificación de qué atributos contienen valores nulos o faltantes. En este estudio, se ha confirmado que no se encontraron tales valores en ninguna de las variables.

Figura 57

Verificación de valores nulos

```
print(df.isnull().sum())
```

```
ZONA_ML             0
Oficina             0
Propiedad           0
ESTATUS PRODUCTO    0
ARTICULO_2          0
Cantidad            0
PRECIO              0
dtype: int64
```

Nota. Elaboración propia, 2023.

Tras filtrar y validar la data, procederemos a convertir las variables categóricas, específicamente "'ESTATUS PRODUCTO', 'ARTICULO_2', 'ZONA_ML', 'Oficina'" y “Propiedad” en variables numéricas utilizando la herramienta *LabelEncoder*.

Figura 58

Conversión de Variables Categóricas a Numéricas con LabelEncoder

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
columnas_categoricas =  ['ESTATUS PRODUCTO', 'ARTICULO_2', 'ZONA_ML','Oficina','Propiedad']
df['ESTATUS PRODUCTO'] = le.fit_transform(df['ESTATUS PRODUCTO'])
df['ARTICULO_2'] = le.fit_transform(df['ARTICULO_2'])
df['ZONA_ML'] = le.fit_transform(df['ZONA_ML'])
df['Oficina'] = le.fit_transform(df['Oficina'])
df['Propiedad'] = le.fit_transform(df['Propiedad'])
```

Nota. Elaboración propia, 2023.

Tras llevar a cabo la transformación, se muestra la nueva tabla actualizada denominada *df* con la siguiente información:

Figura 59

Transformación de Variables Categóricas a Numéricas en la Tabla 'df'

	ZONA_ML	Oficina	Propiedad	ESTATUS PRODUCTO	ARTICULO_2	Cantidad	\
0	0	2	9	3	4	6	
1	0	2	9	3	4	6	
2	0	2	2	3	3	15	
3	0	2	2	3	0	10	
4	0	2	2	3	3	20	

	PRECIO
0	15.660567
1	59.979400
2	21.594000
3	39.801400
4	17.959600

Nota. Elaboración propia, 2023.

Luego de cuantificar los datos, no es viable proceder directamente con la metodología K-Means sin una previa estandarización de la información recolectada. Para este propósito, se procederá a normalizar los datos utilizando el método *StandarScaler().*

Figura 60

Normalización de Datos Utilizando StandarScaler()

```
# Normalizar los datos usando el método standardscaler
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)
df_scaled
```

```
array([[-1.23957305, -0.86476496,  0.74541445, ...,  1.06719125,
        -0.15863114,  0.33546792],
       [-1.23957305, -0.86476496,  0.74541445, ...,  1.06719125,
        -0.15863114,  4.38603634],
       [-1.23957305, -0.86476496, -1.65363731, ...,  0.47618855,
        -0.10981708,  0.87776057],
       ...,
       [-1.23957305, -0.86476496,  1.08813613, ...,  0.47618855,
        -0.12608844, -0.17590848],
       [-1.23957305, -0.86476496,  1.08813613, ...,  0.47618855,
        -0.15863114,  0.3547007 ],
       [-1.23957305, -0.86476496,  0.05997109, ..., -1.29681956,
        -0.13693601, -0.0594333 ]])
```

Nota. Elaboración propia, 2023.

Después de estandarizar la data, se aplicó el Análisis de Componentes Principales (PCA) con el objetivo de determinar cuántos componentes son necesarios para representar el 95% de la información contenida en el repositorio de información. Como se muestra en la figura 61, se observa que con 7 componentes se logra alcanzar dicho porcentaje de varianza explicada.

Figura 61

Código de Implementación del Análisis de Componentes Principales (PCA)

```
# Mostrar el número óptimo de componentes según el criterio del 95% de varianza explicada
import numpy as np
cumsum = np.cumsum(pca.explained_variance_ratio_)
n_components = np.argmax(cumsum >= 0.95) + 1
print(n_components)
```

```
7
```

```
# Crear y entrenar el modelo PCA con el número óptimo de componentes
pca = PCA(n_components=n_components)
pca.fit(df)
```

```
PCA(n_components=7)
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.
```

```
# Transformar los datos originales en los componentes principales
df_pca = pca.transform(df)
df_pca
```

```
array([[-2.92593179e+01,  3.59967881e+00,  2.10887923e+00, ...,
        -6.48547392e-01, -1.61286995e-01, -1.13223557e-01],
       [-2.94037230e+01,  4.78394247e+01,  2.68944695e+00, ...,
        -8.77928415e-01,  2.32345115e+00,  3.44412633e-02],
       [-2.02701917e+01,  9.62351820e+00, -4.76191080e+00, ...,
        -1.39177494e+00,  1.98482134e-01, -1.48569699e-01],
       ...,
       [-2.32426839e+01, -1.97862602e+00,  3.07103014e+00, ...,
        -1.13184505e+00, -7.38340792e-01, -1.09463968e-01],
       [-2.92615610e+01,  3.79738203e+00,  3.13911010e+00, ...,
        -1.16443745e+00, -4.16331818e-01, -8.99680569e-02],
       [-2.52439385e+01, -6.86598666e-01,  1.76527790e-01, ...,
        -2.83484195e+00, -1.25203305e+00, -9.55849945e-02]])
```

Nota. Elaboración propia, 2023.

5.1.2.4. Modelamiento

Durante esta fase del proceso de investigación, exploraremos los 3 métodos mencionados anteriormente.

En el caso del Modelo K-Means, el cual implica llevar a cabo múltiples iteraciones con valores en un rango específico (1, 15), utilizaremos el método del codo para determinar el valor óptimo de *K*. En la figura 63, observaremos que a medida que aumenta la cantidad de clústeres *K*, la sumatoria entre las distancias al cuadrado de los mismos tiende a disminuir. En nuestro estudio de investigación, hemos identificado que el valor óptimo de *K* es igual a 4.

Figura 62

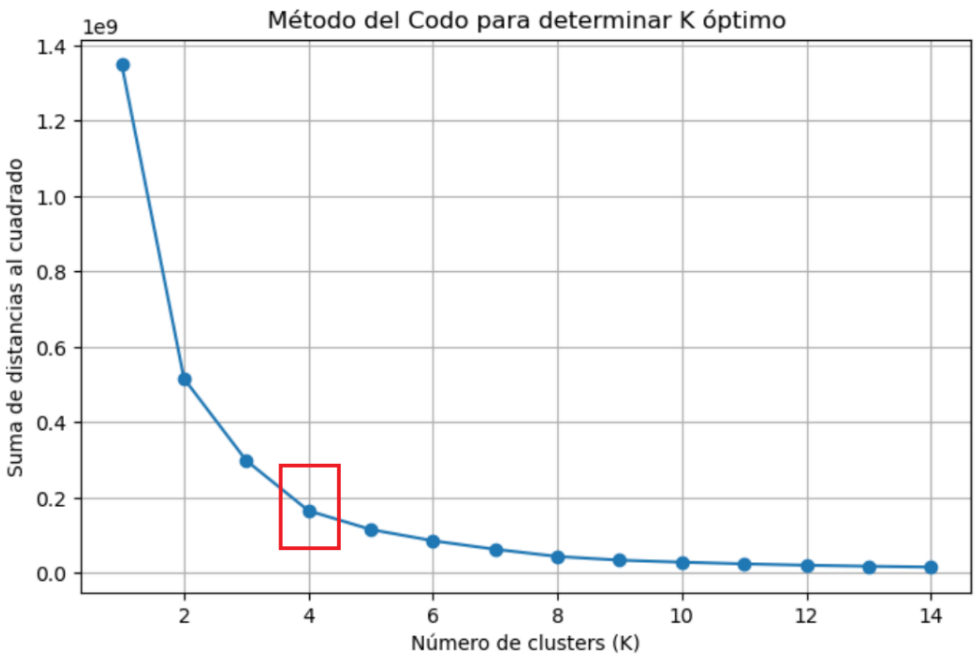
Aplicación de K-Means

```
import matplotlib.pyplot as plt
# Se determina el rango de valores de K a evaluar
# Por ejemplo, en este caso se evaluará entre 1 y 15.
rangos_k = range(1, 15)
# Luego, se calcula la suma de las distancias al cuadrado para cada valor de K
inertias = []
for k in rangos_k:
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(df_pca)
# Teniendo que en cuenta que la inercia del modelo es la suma de las distancias al cuadrado
    inertias.append(kmeans.inertia_)
```

Nota. Elaboración propia, 2023.

Figura 63

Aplicación del método del codo - Modelo KMeans



Nota. Elaboración propia, 2023.

Clúster Jerárquico Aglomerativo:

Para evaluar el clúster jerárquico y compararlo con los demás modelos, se emplea el coeficiente de silueta, que es una métrica utilizada para evaluar la calidad de la agrupación de datos. Proporciona una medida de cuán bien definidos están los clústeres en un conjunto de datos, es decir, si se agrupan de manera coherente.

Figura 64

Aplicación del Dendograma

```
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt
```

```
dendrogram = sch.dendrogram(sch.linkage(df_scaled[:1000,:], method = 'ward',metric = 'euclidean'))
plt.ylim(top=50)
plt.xlim(left=-500)
plt.axhline(y = 100,color = 'r', linestyle = '-')
```

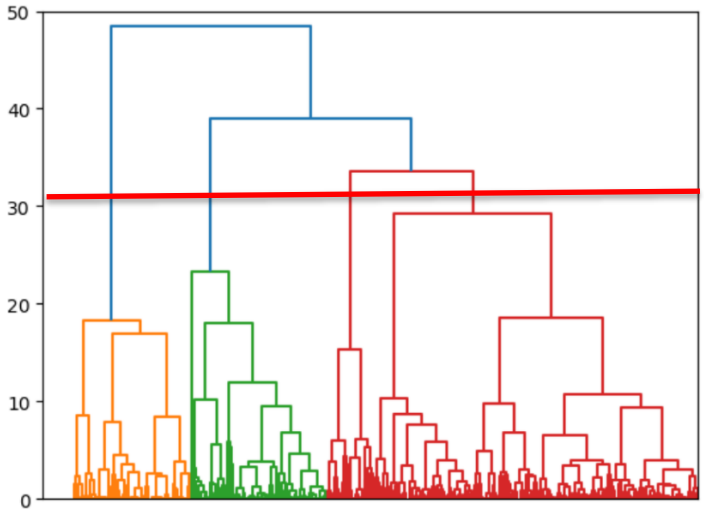
Nota. Elaboración propia, 2023.

Una vez que se haya finalizado el proceso de creación de la gráfica a través de 1000 iteraciones, se hace necesario realizar un análisis minucioso para determinar el punto de corte horizontal óptimo. Cada intersección en la figura 65 indica un clúster potencial para utilizar. En este estudio, se ha seleccionado el punto Y=4 como el más adecuado para el proyecto.

La elección de este valor se basa en la observación de que las divisiones verticales de los datos son más notables y distintivas en esta región del gráfico en comparación con otras áreas. Esta particularidad sugiere que la separación de los datos en cuatro clústeres en este punto no solo es significativa, sino que también se ajusta de manera óptima a la estructura subyacente de los datos.

Figura 65

Dendograma (Clúster Jerárquico)



Nota. Elaboración propia, 2023.

Clúster K-Medoids:

Este enfoque emplea puntos de datos específicos seleccionados como centros de clústeres con el propósito de reducir al mínimo las discrepancias entre los puntos dentro de una agrupación y los puntos designados como centros de la misma. Para este procedimiento, se ha realizado un conjunto de iteraciones en un rango que varía de 1 a 15. Esta técnica se llevará a cabo utilizando la biblioteca "sklearn_extra.Clúster".

Figura 66

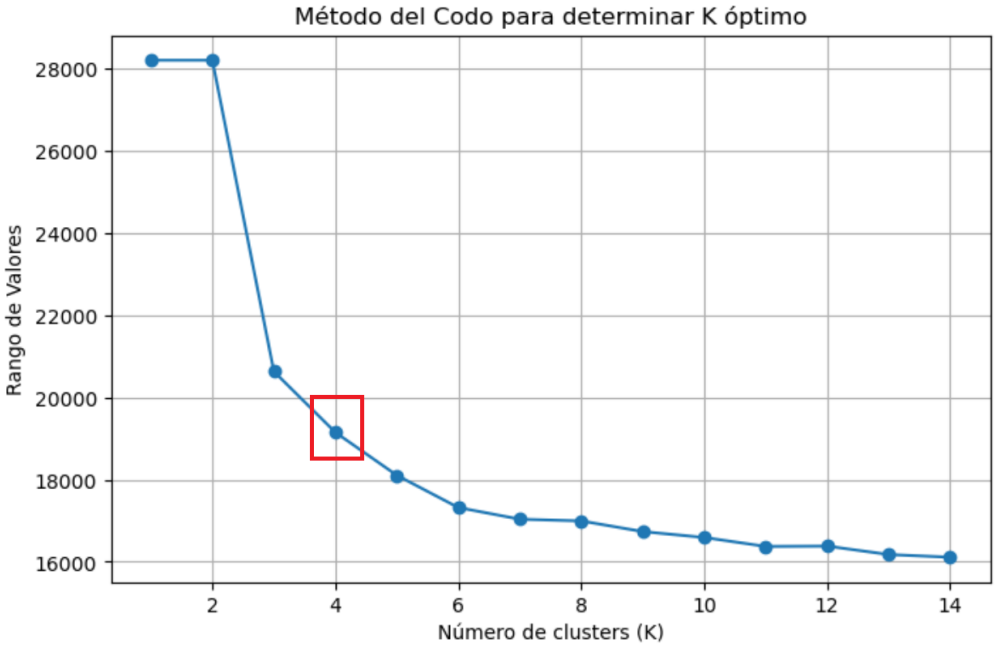
Código K-Medoids

```
from sklearn_extra.cluster import KMedoids
import matplotlib.pyplot as plt

rangos_k = range(1, 15)

inertias = []
for k in rangos_k:
    kmedoids = KMedoids(n_clusters=k)
    kmedoids.fit(df_pca[:1000,:])

    inertias.append(kmedoids.inertia_)
```

Nota. Elaboración propia, 2023.

Figura 67

Aplicación Método del Codo – Modelo K-Medoids



Nota. Elaboración propia, 2023.

El número de clúster que recomienda este método es el de $K = 4$.

5.1.2.5. Evaluación

Comparación de Métodos Clúster

En la evaluación del modelo K-Means, se empleó el indicador teórico de inercia para validar su eficacia. Esto implica que los modelos generados proporcionarán un valor óptimo de *K* para su aplicación, el cual se calcula teniendo en cuenta la sumatoria entre las distancias al cuadro de cada punto y el clúster al que se asigna durante la simulación. En cuanto al modelo K-Medoids, de igual forma lo validaremos con el mismo indicador inercia, el cual se calcula como la suma de distancias de cada punto al medoide de su clúster asignado.

Tener en consideración que cuando la inercia es menor, indica que existe mayor cercanía dentro de cada clúster. Los resultados de la evaluación de la Inercia para los modelos de K-Means y K-Medoids se describen a continuación:

Figura 68

Inercia por Clúster (K-Means)

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmeans = KMeans(n_clusters=4,n_init="auto")
kmeans.fit(df_pca)
#Modifica para determinar la efectividad del modelo KMeans
kmeans.inertia_
```

```
166347928.41986835
```

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmeans = KMeans(n_clusters=5,n_init="auto")
kmeans.fit(df_pca)
#Modifica para determinar la efectividad del modelo KMeans
kmeans.inertia_
```

```
132668869.79653475
```

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmeans = KMeans(n_clusters=6,n_init="auto")
kmeans.fit(df_pca)
#Modifica para determinar la efectividad del modelo KMeans
kmeans.inertia_
```

```
92916005.49247843
```

Nota. Elaboración propia, 2023.

Figura 69

Inercia por Clúster (K-Medoids)

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmedoids = KMedoids(n_clusters=5)
kmedoids.fit(df_pca[:1000,:])
#Modifica para determinar la efectividad del modelo KMedoids
kmedoids.inertia_
```

```
18104.82413188747
```

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmedoids = KMedoids(n_clusters=4)
kmedoids.fit(df_pca[:1000,:])
#Modifica para determinar la efectividad del modelo KMedoids
kmedoids.inertia_
```

```
19149.877427302417
```

```
#En base al gráfico, se visualiza que mi K óptimo es =
kmedoids = KMedoids(n_clusters=6)
kmedoids.fit(df_pca[:1000,:])
#Modifica para determinar la efectividad del modelo KMedoids
kmedoids.inertia_
```

```
17320.69758305813
```

Nota. Elaboración propia, 2023.

Otro método de validación para encontrar el mejor *K*, es el coeficiente de silueta, el cual es esencial para evaluar cada conjunto de clústeres de todos los modelos mencionados anteriormente. Para ello se realizó una comparación de los mismos.

Figura 70

Comparación de Métodos de Clúster

```
from sklearn.metrics import silhouette_score
from sklearn_extra.cluster import KMedoids
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import KMeans
for k in k_range:

    # Crear y entrenar el modelo KMeans con k clusters
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df_pca)
    # Calcular el coeficiente de silueta para el modelo KMeans
    kmeans_score = silhouette_score(df_pca, kmeans.labels_)
    # Añadir el coeficiente de silueta a la lista de resultados de KMeans
    kmeans_scores.append(kmeans_score)

    # Crear y entrenar el modelo KMedoids con k clusters
    kmedoids = KMedoids(n_clusters=k, random_state=42)
    kmedoids.fit(df_pca)
    # Calcular el coeficiente de silueta para el modelo KMedoids
    kmedoids_score = silhouette_score(df_pca, kmedoids.labels_)
    # Añadir el coeficiente de silueta a la lista de resultados de KMedoids
    kmedoids_scores.append(kmedoids_score)

    # Crear y entrenar el modelo Cluster jerárquico con k clusters y enlace ward
    hierarchical = AgglomerativeClustering(n_clusters=k, linkage="ward")
    hierarchical.fit(df_pca)
    # Calcular el coeficiente de silueta para el modelo Cluster jerárquico
    hierarchical_score = silhouette_score(df_pca, hierarchical.labels_)
    # Añadir el coeficiente de silueta a la lista de resultados de Cluster jerárquico
    hierarchical_scores.append(hierarchical_score)

    # Crear un dataframe con los resultados de cada método para cada k
results = pd.DataFrame({"k": k_range,
                        "KMeans": kmeans_scores,
                        "KMedoids": kmedoids_scores,
                        "Hierarchical": hierarchical_scores})
# Mostrar el dataframe con los resultados
print(results)
```

Nota. Elaboración propia, 2023.

Tabla 18

Resultados Coeficiente de Silueta

K	KMeans	KMedoids	Hierarchical
2	0.978924	0.019092	0.980780
3	0.945941	0.240009	0.940381
4	0.924280	0.322042	0.940730
5	0.889447	0.193165	0.811447
6	0.889464	0.266300	0.812593
7	0.847428	0.291774	0.812636
8	0.786528	0.240472	0.812748
9	0.669250	0.232818	0.745510
10	0.683925	0.255029	0.745962

Nota. Elaboración propia, 2023.

La tabla 18 proporciona los coeficientes de silueta para diferentes valores de *K* en tres técnicas de agrupación: K-Means, K-Medoids y Clúster Jerárquico. El coeficiente de silueta mide la calidad de las agrupaciones, evaluando la cohesión y la separación entre los clústeres. Un valor de silueta más alto indica una agrupación más efectiva, con una mayor similitud entre los puntos dentro de un clúster y una mayor distancia entre clústeres. Al analizar estos coeficientes para diversos valores de *K*, podemos determinar cuántos clústeres son óptimos para representar los patrones de datos de manera coherente en cada técnica.

Para K-Means y Clúster Jerárquico:

Se pueden observar que los valores del coeficiente de silueta están próximos a 1. Esto sugiere que existe una alta cohesión y separación efectiva entre los clústeres. Por lo tanto, usaremos estos métodos para la clasificación.

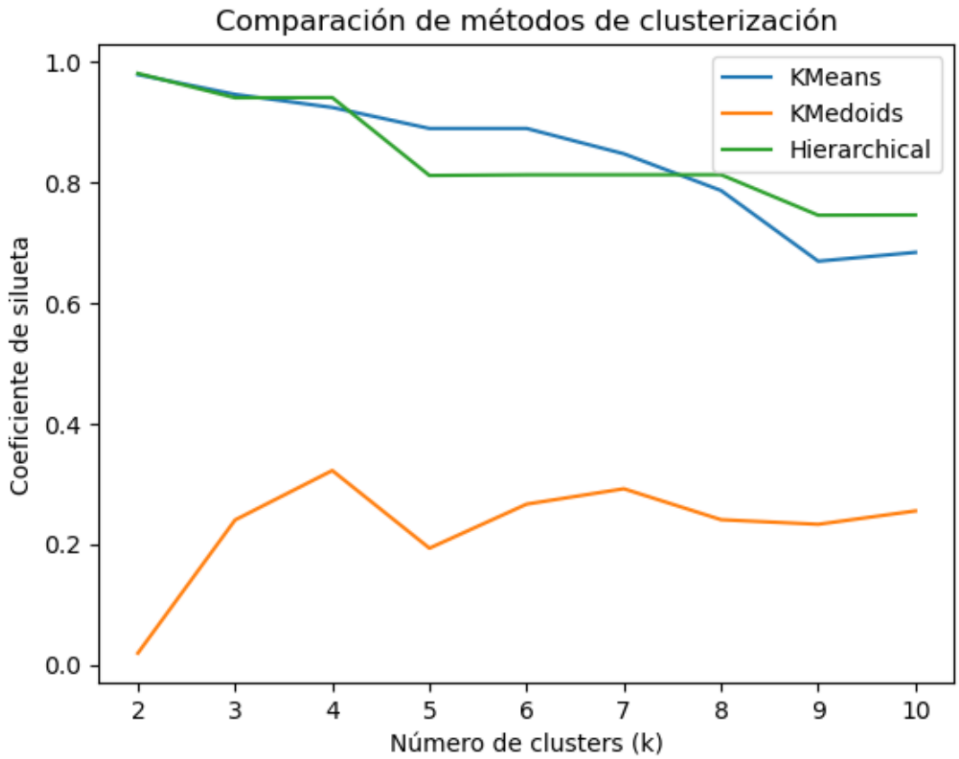
Para K-Medoids:

Se pueden observar que los valores del coeficiente de silueta están próximos a 0. Esto sugiere que existe una baja cohesión y separación inefectiva entre los clústeres. A raíz de esto descartaremos este método para la clasificación.

Figura 71

Gráfico de silueta

```
import matplotlib.pyplot as plt
# Crear un gráfico de línea con los resultados de cada método para cada k
plt.plot(k_range, kmeans_scores, label="KMeans")
plt.plot(k_range, kmedoids_scores, label="KMedoids")
plt.plot(k_range, hierarchical_scores, label="Hierarchical")
plt.xlabel("Número de clusters (k)")
plt.ylabel("Coeficiente de silueta")
plt.title("Comparación de métodos de clusterización")
plt.legend()
plt.show()
```



Nota. Elaboración propia, 2023.

Según la figura 71, podemos observar que en el caso del método K-Medoids nos da valores más cercanos a 0 lo que supone que hay superposición de los clústeres y por tanto esto no ayuda en la clasificación de los mismos. Por ello en conjunto con los resultados de la tabla 18 utilizaremos los métodos de K-Means y Clúster Jerárquico para nuestro modelo.

Figura 72

Agrupación por Clúster jerárquico y KMeans

	ZONA_ML	Oficina	Propiedad	ESTATUS PRODUCTO	ARTICULO_2	Cantidad	PRECIO	KMeans/cliente	Cluster jerarquico/cliente
0	0	2	9	3	4	6	15.660567	0	0
1	0	2	9	3	4	6	59.979400	0	0
2	0	2	2	3	3	15	21.594000	0	0
3	0	2	2	3	0	10	39.801400	0	0
4	0	2	2	3	3	20	17.959600	0	0
...	...	...	...	...	...	...	...	...	...
39507	0	2	7	3	4	2	22.844800	0	0
39508	0	2	7	3	4	50	5.841000	0	0
39509	0	2	10	3	3	12	10.065400	0	0
39510	0	2	10	3	3	6	15.871000	0	0
39511	0	2	7	3	0	10	11.339800	0	0

39512 rows × 9 columns

Nota. Elaboración propia, 2023.

Con el fin de validar que modelo es más efectivo, se presentaron dos propuestas al experto: la primera utilizaba el modelo de K-Means con 4 clústeres, mientras que la segunda empleaba el modelo de Clúster Jerárquico con 4 clústeres. A continuación, se describen ambas propuestas. Para obtener información adicional sobre la evaluación y la entrevista, ver Anexo 2.

Tabla 19

Propuesta 1 - Modelo K-Means

	Clúster K0	**Clúster K1**	**Clúster K2**	**Clúster K3**
ESTATUS PRODUCTO	BONIFICACION, REGULAR, CORTO VENCIMIENTO, PROMOCION, REGULAR	CORTO VENCIMIENTO, REGULAR	REGULAR	REGULAR, CORTO VENCIMIENTO, PROMOCION
ARTICULO_2	SUSPENSIÓN ORAL, COMPRIMIDOS / TABLETAS, INYECTABLES, CREMAS.	COMPRIMIDOS / TABLETAS, CREMAS.	INYECTABLES	COMPRIMIDOS / TABLETAS, CREMAS, SUSPENSIÓN ORAL

Nota. Elaboración propia, 2023.

	Clúster K0	Clúster K1	Clúster K2	Clúster K3
ZONA_ML	LIMA, NOR, SIERRA, SELVA Y SUR	LIMA, NOR, SELVA	LIMA	LIMA, NOR, SIERRA
OFICINA	CE, HG, LI, NO, SU	LI, NO, CE, SU	LI	LI, NO, SU
PROPIEDAD	MINORISTA, CALL CENTER, CAPON, DISTRIBUIDOR, CODISTIBUIDOR, CIRCUITO CERRADO, FARMACIA, MAYORISTA, MÉDICO VISITA, MINI CADENA	CAPON, CIRCUITO CERRADO, CODISTRIBUIDOR, MAYORISTA	MAYORISTA	CIRCUITO CERRADO, MAYORISTA, CODISTRIBUIDOR, MINI CADENA, MINORISTA

Nota. Elaboración propia, 2023.

Tabla 20

Propuesta 2 - Modelo Clúster Jerárquico

	Clúster J0	Clúster J1	Clúster J2	Clúster J3
ESTATUS PRODUCTO	REGULAR, BONIFICACION	REGULAR	REGULAR, CORTO VENCIMIENTO	REGULAR
ARTICULO_2	COMPRIMIDOS / TABLETAS, INYECTABLES, SUSPENSIÓN ORAL	CREMAS, INYECTABLES.	INYECTABLES, SUSPENSIÓN ORAL	INYECTABLES, CREMAS.
ZONA_ML	LIMA, SIERRA, SUR	LIMA, NOR	LIMA, SIERRA, NOR	LIMA
OFICINA	LI, NO, SU, HG, CE	LI, NO	LI, NO, SU, CE	LI, SU

Nota. Elaboración propia, 2023.

	Clúster J0	Clúster J1	Clúster J2	Clúster J3
PROPIEDAD	MINORISTA, CIRCUITO CERRADO, MINI CADENA, MAYORISTA, WEB, CODISTRIBUIDOR, FARMACIA HOSPITALARIA, DISTRIBUIDOR, OBSTETRA VISITA, CAPON, MÉDICO VISITA	MAYORISTA	MAYORISTA, CODISTRIBUIDOR, CAPON, CIRCUITO CERRADO, MINI CADENA	MAYORISTA, CODISTRIBUIDOR, CAPON, CIRCUITO CERRADO, WEB

Nota. Elaboración propia, 2023.

5.1.2.6. Implementación

Dado que el experto optó por una combinación de *K*=4 y el algoritmo K-Means, se procedió a realizar un análisis detallado de los resultados obtenidos. El objetivo de este análisis es identificar y resaltar los datos más relevantes para cada clúster, lo que facilitaría una categorización efectiva de estos grupos. Este enfoque busca generar un impacto beneficioso para la empresa objeto de estudio.

Tabla 21

Resultado y Evaluación del Clúster K-Means

	Clúster K0	Clúster K1	Clúster K2	Clúster K3
ESTATUS PRODUCTO	BONIFICACION, REGULAR, CORTO VENCIMIENTO, PROMOCION, REGULAR	CORTO VENCIMIENTO, REGULAR	REGULAR	REGULAR, CORTO VENCIMIENTO, PROMOCION
ARTICULO_2	SUSPENSIÓN ORAL, COMPRIMIDOS / TABLETAS, INYECTABLES, CREMAS.	COMPRIMIDOS / TABLETAS, CREMAS.	INYECTABLES	COMPRIMIDOS / TABLETAS, CREMAS, SUSPENSIÓN ORAL

Nota. Elaboración propia, 2023.

	Clúster K0	Clúster K1	Clúster K2	Clúster K3
ZONA_ML	LIMA, NOR, SIERRA, SELVA Y SUR	LIMA, NOR, SELVA	LIMA	LIMA, NOR, SIERRA
OFICINA	CE, HG, LI, NO, SU	LI, NO, CE, SU	LI	LI, NO, SU
PROPIEDAD	MINORISTA, CALL CENTER, CAPON, DISTRIBUIDOR, CODISTIBUIDOR, CIRCUITO CERRADO, FARMACIA, MAYORISTA, MÉDICO VISITA, MINI CADENA	CAPON, CIRCUITO CERRADO, CODISTRIBUIDOR, MAYORISTA	MAYORISTA	CIRCUITO CERRADO, MAYORISTA, CODISTRIBUIDOR, MINI CADENA, MINORISTA

Nota. Elaboración propia, 2023.

Análisis de atributos K-Means:

• Atributo “ESTATUS PRODUCTO”:

Tabla 22

ESTATUS_ PRODUCTO por K-Means

ESTATUS_PRODUCTO	Clúster 0	Clúster 1	Clúster 2	Clúster 3	TOTAL GENERAL
BONIFICACION	8217			5	8222
CORTO_VENCIMIENTO	1971	3	2	**17**	1993
PROMOCIÓN	1051			5	1056
REGULAR	**27643**	94	14	**490**	28241
TOTAL GENERAL	38882	97	16	517	39512

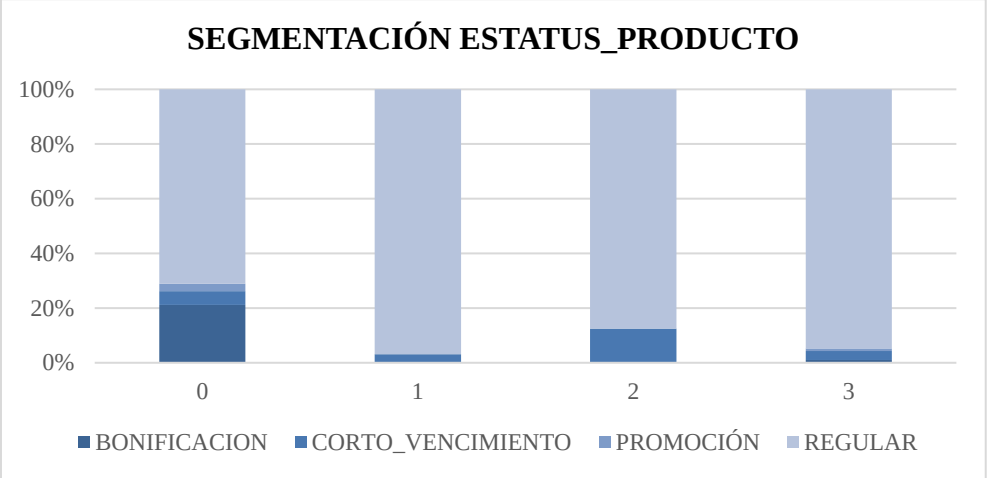
Nota. Elaboración propia, 2023.

Estamos dividiendo las variables que tienen un alto porcentaje de representación. A continuación, se proporcionará una descripción del peso para cada clúster.

- Clúster 0: El estatus de producto más representativo es el regular.
- Clúster 1: El atributo regular es el más representativo con 96.90%
- Clúster 2: Muestra un alto nivel de representación el estatus de producto regular.
- Clúster 3: El estatus de producto regular sigue siendo el más representativo.

Figura 73

Segmentación ESTATUS_PRODUCTO por K-Means



Nota. Elaboración propia, 2023.

• Atributo “ARTICULO_2”:

Tabla 23

ARTICULO_2 por K-Means

ARTICULO_2	Clúster 0	Clúster 1	Clúster 2	Clúster 3	TOTAL GENERAL
COMPRIMIDOS / TABLETAS	**12957**	13	3	64	13037
CREMAS	**2009**	24	6	47	2086
EFERVESCENTE	825				825
INYECTABLES	10975	51	5	**261**	11292
SUSPENSIÓN ORAL	**12116**	9	2	**145**	12272
TOTAL GENERAL	38882	97	16	517	39512

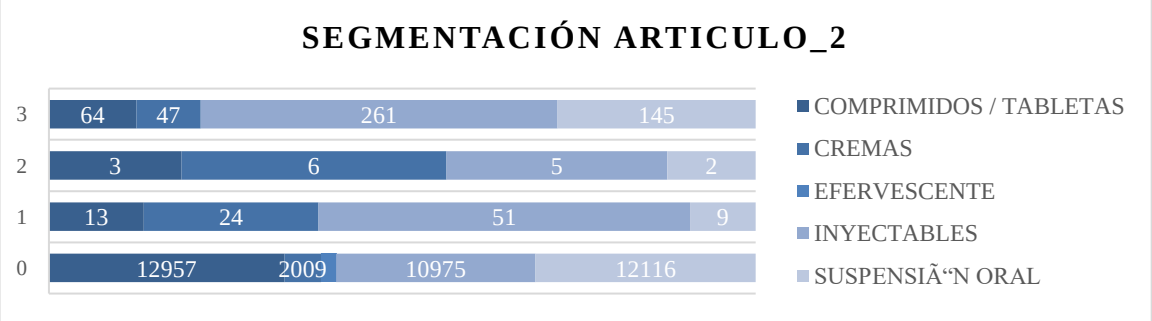
Nota. Elaboración propia, 2023.

Para llevar a cabo el análisis, se han elegido los artículos que registraron un mayor número de transacciones de compra. Esto se ha realizado con el fin de comprender en profundidad el comportamiento del cliente.

- Clúster 0: Los artículos comprimidos/tabletas, suspensión oral y efervescente son los más representativos.
- Clúster 1: Destaca por tener una alta representación del artículo "inyectables", con un 52.6% de incidencia.
- Clúster 2: Las cremas e inyectables tienen un peso muy cercano, 37.5 % y 31.3%, respectivamente; por lo cual se están considerando ambos.
- Clúster 3: Los artículos inyectables y suspensión oral son los más representativos.

Figura 74

Segmentación ARTICULO_2 por K-Means



Nota. Elaboración propia, 2023.

• Atributo “ZONA_ML”:

Tabla 24

ZONA_ML por K-Means

ZONA_ML	Clúster 0	Clúster 1	Clúster 2	Clúster 3	TOTAL GENERAL
LIMA	**10888**	71	11	**308**	11278
NOR	6824	6	4	**63**	6897
SELVA	3593			8	3601
SIERRA	**9577**	16	1	**120**	9714
SUR	**8000**	4		18	8022
TOTAL GENERAL	38882	97	16	517	39512

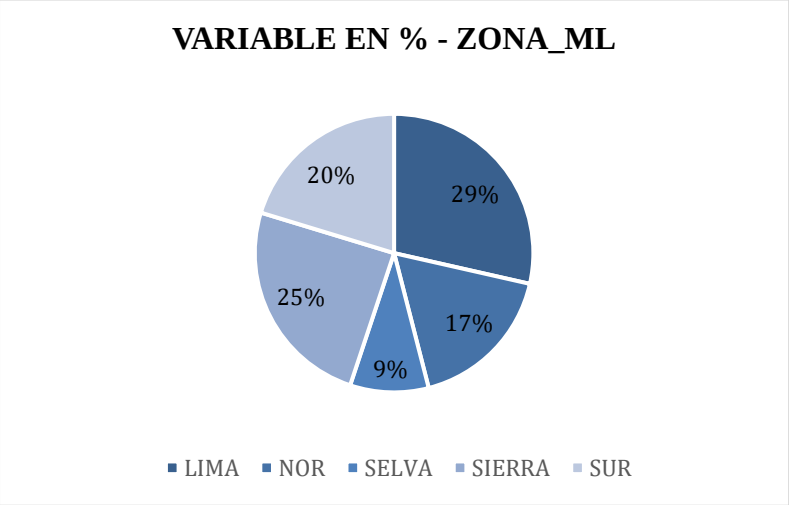
Nota. Elaboración propia, 2023.

Para el análisis, se han identificado las zonas con el mayor número de transacciones. A continuación, se presentan:

- Clúster 0: Las zonas más destacadas son Lima y el SUR.
- Clúster 1: Lima sobresale como la zona más representativa con un 73.2%.
- Clúster 2: Lima es la zona más predominante con un 68.8%.
- Clúster 3: Las zonas más relevantes son Lima y SIERRA.

Figura 75

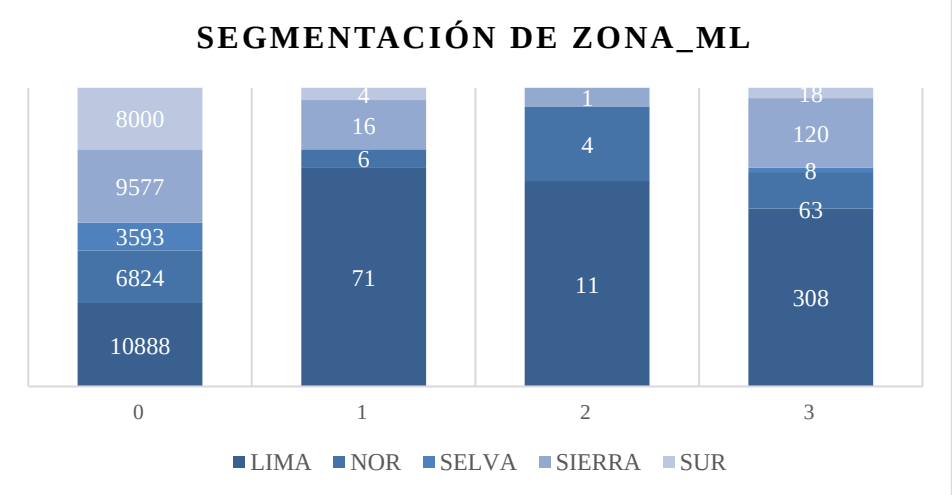
Porcentaje (%) Variable ZONA_ML por K-Means



Nota. Elaboración propia, 2023.

Figura 76

Segmentación ZONA_ML por K-Means



Nota. Elaboración propia, 2023.

• Atributo “Oficina”:

Tabla 25

OFICINA por K-Means

OFICINA	Clúster 0	Clúster 1	Clúster 2	Clúster 3	TOTAL GENERAL
CE	827	5	1	29	862
HG	961				961
LI	**13553**	71	11	316	13951
NO	**12160**	11	4	**106**	12281
SU	**11381**	10		66	11457
TOTAL GENERAL	38882	97	16	517	39512

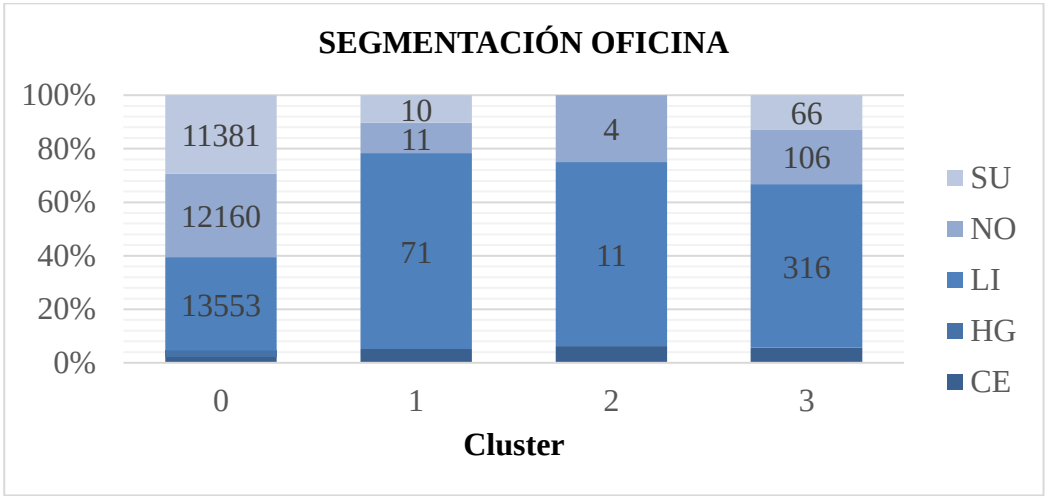
Nota. Elaboración propia, 2023.

En la figura 77, se pueden observar las oficinas con el mayor número de transacciones de compra.

- Clúster 0: Las oficinas más representativas se consideran Lima y Sur.
- Clúster 1: Lima destaca como el más representativo con un 73.2%.
- Clúster 2: Lima es el más representativo con un 68.8%.
- Clúster 3: Las oficinas más representativas son Lima y Norte.

Figura 77

Segmentación OFICINA por K-Means



Nota. Elaboración propia, 2023.

• Atributo “Propiedad”:

La tabla 26 muestra la segmentación de la variable propiedad basada en el mayor porcentaje de transacciones.

Tabla 26

PROPIEDAD por K-Means

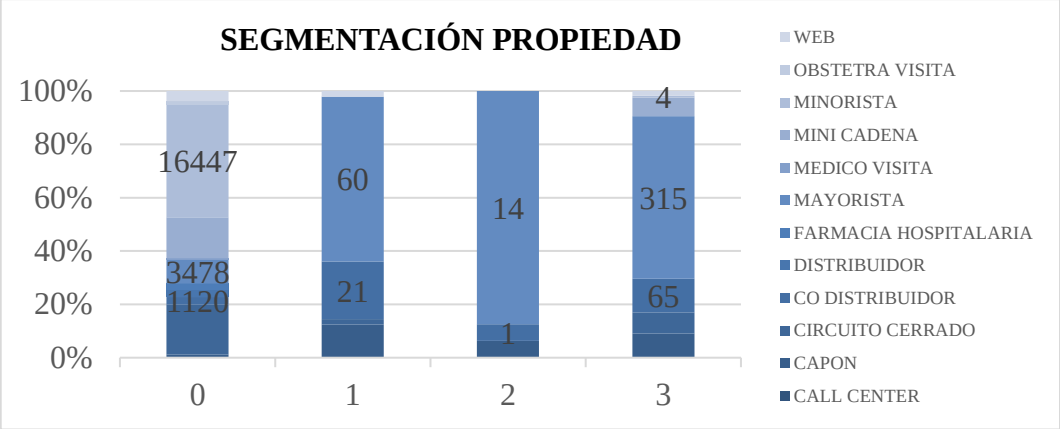
PROPIEDAD	CLÚSTER 0	CLÚSTER 1	CLÚSTER 2	CLÚSTER 3	TOTAL GENERAL
CALL CENTER	**13**				13
CAPON	**432**	12	1	47	492
CIRCUITO CERRADO	**7289**	2		41	7332
CO DISTRIBUIDOR	1120	21	1	**65**	1207
DISTRIBUIDOR	**961**				961
FARMACIA HOSPITALARIA	995				995
MAYORISTA	3478	60	14	**315**	3867
MEDICO VISITA	**204**				204
MINI CADENA	5970			36	6006
MINORISTA	16447			4	16451
OBSTETRA VISITA	487				487
WEB	1486	2		9	1497
TOTAL GENERAL	38882	97	16	517	39512

Nota. Elaboración propia, 2023.

- Clúster 0: Dentro de este grupo, la propiedad minorista se destaca como la más representativa.
- Clúster 1: La propiedad mayorista es la categoría predominante, representando un 61.9% del total.
- Clúster 2: La propiedad mayorista sobresale como la más representativa.
- Clúster 3: La propiedad mayorista representa un 60.9% del total en este grupo.

Figura 78

Segmentación PROPIEDAD por K-Means



Nota. Elaboración propia, 2023.

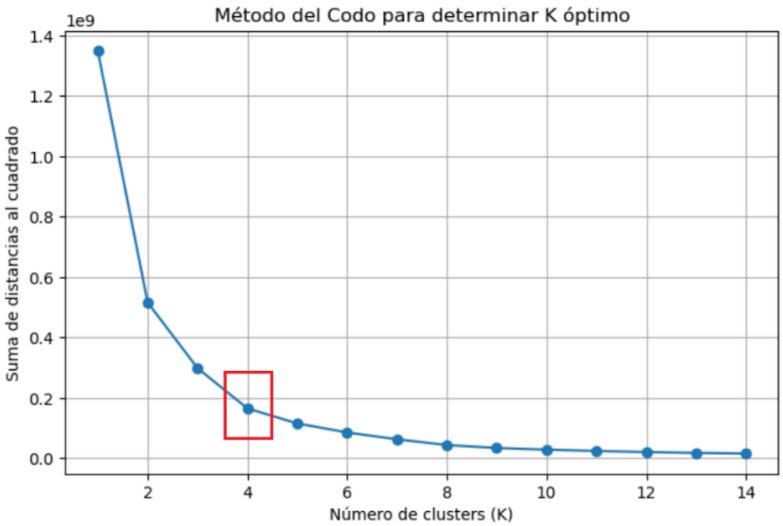
5.2 Medición de la solución.

5.2.1. Análisis de Indicadores cuantitativo y/o cualitativo

En la evaluación de los modelos actuales utilizando las técnicas K-Means y Clúster Jerárquico, se validó el indicador de inercia. Esta proporciona una medida de cuán bien separados están los clústeres en un conjunto de datos, la misma es útil para determinar si los datos se agrupan de manera coherente.

Figura 79

Método del codo



Nota. Elaboración propia, 2023.

Examinando los resultados para cada valor de *K*, se nota un punto de inflexión en el clúster número cuatro. Este cambio en la tendencia llevó a la selección de $K = 4$ como el número de clúster óptimo. Esta decisión también fue respaldada por la validación de un experto del negocio quien confirmó los resultados obtenidos.

5.2.2. Simulación de solución. Aplicación de Software

A continuación, se proporciona la información consolidada de acuerdo con los grupos revisados, permitiendo la evaluación de perfiles de clientes según su tipo.

Tabla 27

Resultado y Evaluación del Clúster K-Means

	Clúster K0	**Clúster K1**	**Clúster K2**	**Clúster K3**
ESTATUS PRODUCTO	BONIFICACION, REGULAR, CORTO VENCIMIENTO, PROMOCION, REGULAR	CORTO VENCIMIENTO, REGULAR	REGULAR	REGULAR, CORTO VENCIMIENTO, PROMOCION
ARTICULO_2	SUSPENSIÓN ORAL, COMPRIMIDOS / TABLETAS, INYECTABLES, CREMAS.	COMPRIMIDOS / TABLETAS, CREMAS.	INYECTABLES	COMPRIMIDOS / TABLETAS, CREMAS, SUSPENSIÓN ORAL
ZONA_ML	LIMA, NOR, SIERRA, SELVA Y SUR	LIMA, NOR, SELVA	LIMA	LIMA, NOR, SIERRA
OFICINA	CE, HG, LI, NO, SU	LI, NO, CE, SU	LI	LI, NO, SU
PROPIEDAD	MINORISTA, CALL CENTER, CAPON, DISTRIBUIDOR, CODISTIBUIDOR, CIRCUITO CERRADO, FARMACIA, MAYORISTA, MÉDICO VISITA, MINI CADENA	CAPON, CIRCUITO CERRADO, CODISTRIBUIDOR, MAYORISTA	MAYORISTA	CIRCUITO CERRADO, MAYORISTA, CODISTRIBUIDOR, MINI CADENA, MINORISTA

Nota. Elaboración propia, 2023.

Con el fin de lograr una comprensión más precisa de cada uno de los clústeres generados mediante el procesamiento del algoritmo K-Means, se han identificado y etiquetado con la letra *K* seguida de un número correlativo, como se muestra a continuación:

Clúster K0: Este Clúster se clasifica por los clientes que son minoristas, call center, distribuidor, co-distribuidor, etc. Los cuales realizan compras en Lima, Norte, Sierra, Selva y Sur del Perú y que adquieren productos de suspensión oral, comprimidos/tabletas, inyectables y cremas. Estos productos son suministrados por las oficinas de CE, HG, LI, NO Y SU.

Figura 80

Clúster K0

Nota. Elaboración propia, 2023.

Clúster K1: Este clúster se clasifica por los clientes que son de tipo capo, circuito cerrado, co-distribuidor y mayorista. Los cuales realizan compras en Lima, Norte y Selva del Perú y que adquieren productos de comprimidos/tabletas y cremas. Estos productos son suministrados por las oficinas de LI, NO, CE Y SU.

Figura 81

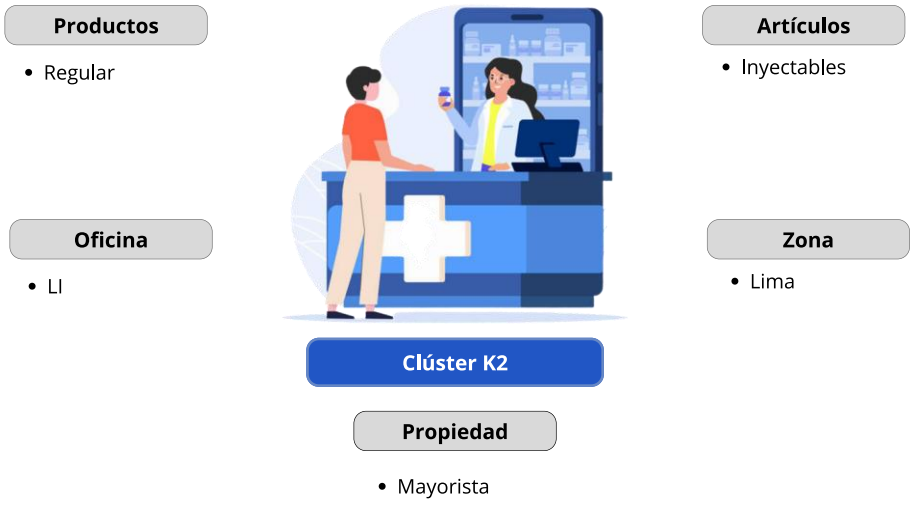
Clúster K1

Nota. Elaboración propia, 2023.

Clúster K2: Está segmentación comprende a clientes mayoristas que efectúan compras de forma regular en la ciudad de Lima, destacándose por su interés particular en la categoría de productos relacionados con inyectables. Estos artículos son abastecidos principalmente por las oficinas ubicadas en Lima.

Figura 82

Clúster K2



Nota. Elaboración propia, 2023.

Clúster K3: Este grupo engloba a clientes que pertenecen a diversas categorías, como circuito cerrado, mayorista, co-distribuidor, mini cadena y minorista. Estos clientes realizan sus compras en las regiones de Lima, Norte y Sierra del Perú, centrándose en la adquisición de productos que incluyen suspensión oral, comprimidos/tabletas y cremas. La distribución de estos productos se lleva a cabo a través de las oficinas ubicadas en Lima, Norte y Sierra.

Figura 83

Clúster K3

Nota. Elaboración propia, 2023.

A continuación, se presentan los detalles de cada clúster luego del análisis de variables:

- Clúster K0: Se clasifica un estatus de producto con código regular con preferencias de tipo de artículo inyectable y cremas en el canal Mayorista, Co Distribuidor, Capón y Circuito Cerrado. Cuya zona que más adquiere este tipo de producto es la región Lima, con oficina de atención en Lima y el Sur.

- Clúster K1: Se clasifica un estatus de producto con código regular con preferencias de tipo de artículo cremas en el canal mayorista y cuya zona que

más adquiere este tipo de producto es la región Lima y Norte, con oficina de atención también en Lima y Norte.

- Clúster K2: Se clasifica un estatus de producto con código regular con preferencias de tipo de artículo inyectable en el canal mayorista y cuya zona que más adquiere este tipo de producto es la región Lima, con oficina de atención también en Lima.

- Clúster K3: Se clasifica un estatus de producto con código regular y corto vencimiento con preferencias de tipo de artículo en suspensión oral en el canal circuito cerrado, mayorista, co distribuidor y mini cadena cuya zona que más adquiere este tipo de producto es la región Lima, Norte y Sierra, asimismo con oficina de atención en Lima, Norte y Sur.

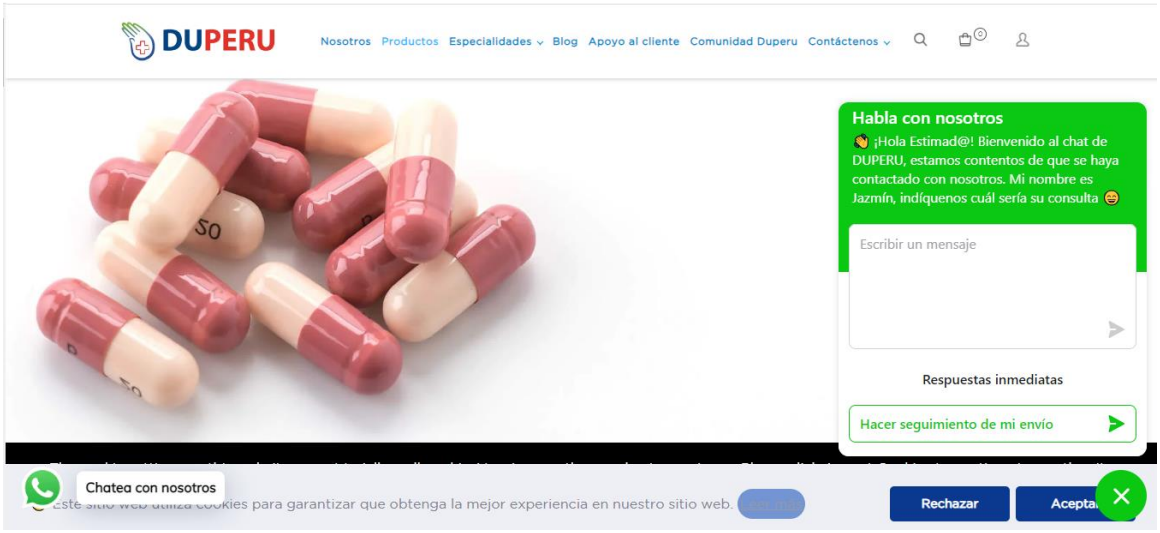
Mediante estos análisis de clústeres, se obtiene una visión más clara de las estrategias comerciales que la empresa pueda implementar, tanto para clientes que prefieren reponer medicamentos como inyectables, suspensión oral, cremas, tabletas, entre otros.

Además, para lograr un mayor acercamiento y, en consecuencia, aumentar las ventas en regiones específicas en función del tipo de producto demandado, se pueden desarrollar estrategias comerciales basadas en la zonificación. Esto permite obtener una comprensión más detallada de la penetración del producto en diferentes áreas geográficas. En este sentido, adaptamos las estrategias comerciales de acuerdo a cada clúster:

- Clúster K0:
    - Canal de Atención: Para este Clúster, se puede establecer un canal de atención telefónica dedicado con representantes de ventas altamente capacitados. Además, se podría desarrollar un portal en línea donde los clientes puedan realizar consultas y recibir asistencia en tiempo real.

Figura 84

Plugging de Chat del sitio web



Nota. Elaboración propia, 2023.

- Promociones: Las promociones pueden incluir descuentos atractivos en productos inyectables y cremas, especialmente para compras al por mayor. Para fomentar la fidelidad, se podrían ofrecer descuentos progresivos a medida que los clientes realicen compras recurrentes.

Figura 85

Promociones en el sitio web

Nota. Elaboración propia, 2023.

- Perfil de Cliente: Este grupo se compone de clientes que muestran una preferencia clara por productos regulares, inyectables y cremas. Se dirigen a los canales Mayorista, Co-distribuidor, Capón y Circuito Cerrado. Los clientes son principalmente de la región Lima.

Esta estrategia se centra en mantener y aumentar la satisfacción de los clientes regulares.

- Clúster K1:
    - Canal de Atención: Se recomienda un canal de atención telefónica y un sistema de tickets en línea para estos clústeres. Es esencial contar con un equipo de soporte que pueda responder a las consultas y necesidades de los clientes de manera eficiente.

    - Promociones: Con el fin de optimizar nuestras estrategias comerciales, contemplamos promociones estratégicas que abarcan descuentos en productos regulares y cremas, con un enfoque especializado en el canal Mayorista. Para este clúster en particular, proponemos acciones específicas de *up-selling*, aprovechando la naturaleza de las principales categorías de productos adquiridos por este grupo. Asimismo, contemplamos la opción de ofrecer charlas personalizadas dirigidas a nuestros clientes mayoristas.

      Estas sesiones no solo proporcionarán información detallada sobre nuestros productos, sino que también brindarán orientación sobre cómo ofrecer servicios personalizados a sus propios consumidores finales. Este enfoque no solo ofrecerá incentivos adicionales a nuestros clientes mayoristas, sino que también les dará la oportunidad de redistribuir las muestras gratuitas entre sus consumidores finales, generando un valor adicional y fortaleciendo nuestras relaciones comerciales, asimismo, se busca mejorar significativamente la fidelización y el nivel de satisfacción de nuestros clientes finales, contribuyendo a consolidar relaciones comerciales más sólidas y duraderas en el mercado.

Figura 86

Promociones de productos regulares y cremas

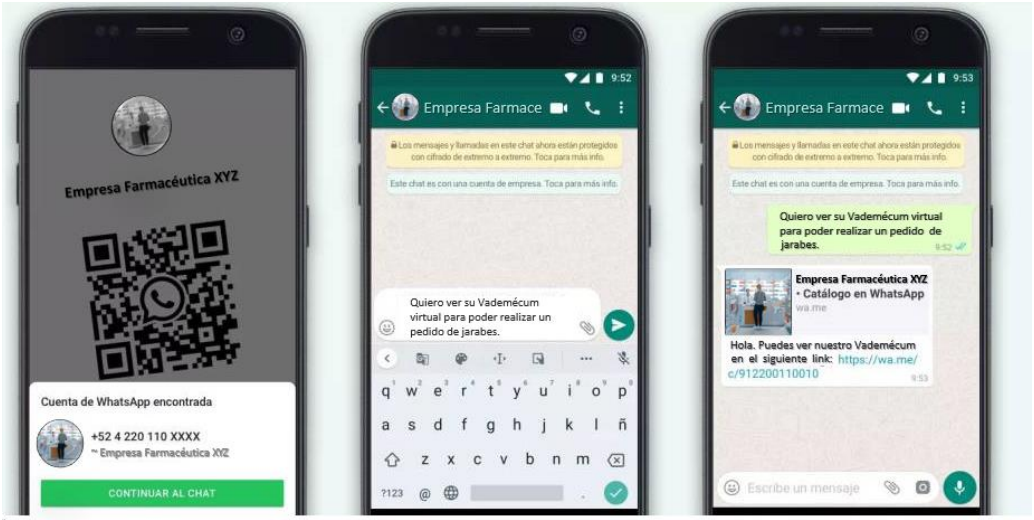
Nota. Elaboración propia, 2023.

- Perfil de Cliente: Este clúster se componen de clientes que prefieren productos regulares y cremas. La concentración principal de clientes está en la región Lima y Norte. Esta estrategia busca mantener la lealtad de los clientes y aumentar las compras en estas categorías.

- Clúster K2:
  - Canal de Atención: Para este clúster, se puede implementar un nuevo canal de atención por WhatsApp, con un equipo de asesores comerciales disponibles para brindar orientación y ofrecer productos a clientes y no clientes. Actualmente se cuenta con un Vademécum con el listado de productos que se ofrecen, por lo que, se puede disponibilizar de manera digital y tener el contacto directo para asesorar y concretar una venta.

Figura 87

Canal de Whatsapp



Nota. Elaboración propia, 2023.

- Promociones: Las promociones pueden incluir descuentos atractivos en productos regulares e inyectables, con un enfoque en el canal Mayorista. Con el fin de promover compras en grandes cantidades, es posible proporcionar descuentos escalonados en función de la cantidad adquirida.

- Perfil de Cliente: Este grupo se centra en clientes que adquieren productos regulares e inyectables a través del canal Mayorista. La mayoría de estos clientes se encuentran en la región Lima. La estrategia busca aumentar la participación en el mercado y mantener la satisfacción de los clientes.

- Cross-selling: Con respecto a este clúster, proponemos diseñar una estrategia de *cross-selling* ya que, a partir del desarrollo del modelo de ML, se identificó que este grupo está conformado mayoritariamente por clientes con preferencias de compra de productos de la categoría Regular. En este sentido, proponemos que se realicen campañas personalizadas que ofrecerán descuentos exclusivos en productos

complementarios como los de la categoría de Corto vencimiento, cuya relación verificamos en otros clústeres encontrados. En este sentido, consideramos que una buena estrategia podría partir de proporcionar información detallada sobre las ventajas de combinar ciertos productos para maximizar los beneficios en cuanto a la salud, además que les permite a los clientes, mejorar su oferta de valor frente a sus consumidores finales.

- Clúster K3:
    - Canal de Atención: Se puede establecer una línea de atención telefónica y un chat en línea con representantes de ventas especializados. La atención debe estar disponible para las regiones Lima, Norte y Sierra.

    - Promociones: Las promociones pueden incluir descuentos en productos regulares con corto vencimiento y en suspensión oral, con un enfoque en el canal circuito cerrado, Mayorista, co distribuidor y minicadena. Se pueden ofrecer incentivos para compras recurrentes y acuerdos de suministro a largo plazo.

Tabla 28

Propuesta de porcentaje de descuento

Producto	Fecha de Vencimiento	Días para vencimiento (fecha hoy – fecha vencimiento)	% de descuento
		> a 1 días	5%
		> a 3 días	7%
		> a 5 días	10%
		> a 10 días	12%

Nota. Elaboración propia, 2023.

- Perfil de Cliente: Este grupo se compone de clientes que compran productos regulares con corto vencimiento y en suspensión oral en los canales mencionados. La demanda se concentra en las regiones Lima, Norte y Sierra. Esta estrategia busca aumentar la retención de clientes y fortalecer las relaciones comerciales.

Cada una de estas estrategias se adapta a las preferencias y necesidades específicas de los clientes en cada clúster, ya que como se ha explicado anteriormente el sector farmacéutico en el Perú es altamente competitivo y está sujeto a regulaciones estrictas, por lo que es necesario que las estrategias comerciales busquen ser lo más eficaces como sea posible.

En este sentido, consideremos que es crucial realizar un seguimiento constante, recopilando datos y comentarios de los clientes para ajustar las estrategias según sea necesario y garantizar un aumento sostenible en las ventas y la satisfacción del cliente. A manera de propuesta, se plantea que existen dos enfoques clave para poder tener este seguimiento:

Realizar encuestas a clientes, ya que permiten obtener información directa de quienes compran los productos farmacéuticos de la empresa. A través de cuestionarios bien diseñados, se pueden recopilar datos sobre la satisfacción del cliente, la efectividad de los mensajes promocionales y las preferencias del consumidor. Por ejemplo, se podrían encuestar a clientes de los clústeres K1 y K2 con respecto a la asesoría y servicio que se le brinda a través de los distintos canales, lo que puede indicar si existe alguna necesidad de ajustar la estrategia de comunicación para ese grupo.

La revisión del *feedback* de las áreas comerciales, ya que el equipo de ventas y el área de soporte al cliente están en contacto directo con los clientes y tienen información valiosa sobre las preocupaciones, dudas y comentarios de estos, es necesario evaluar sus opiniones sobre la eficacia de un canal en particular, y por tanto indicar la necesidad de mejorar algún aspecto de marketing relacionados a los clústeres que se vean impactados. Este desarrollo se planificaría de la mano con el área comercial, marketing y desarrollo de producto.

CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

En los últimos años, el sector farmacéutico en Perú ha experimentado un crecimiento significativo, impulsado por una demanda en constante expansión y una competencia cada vez más intensa. En este contexto, el desarrollo del proyecto de segmentación de clientes se convierte en una iniciativa estratégica para la empresa en estudio. La capacidad de comprender y adaptarse a las preferencias cambiantes de los clientes se ha convertido en un factor crítico para el éxito en este entorno altamente dinámico.

El desarrollo del proyecto se basó en la metodología CRISP-DM, una estructura sólida y coherente que abarcó seis etapas. En la etapa de entendimiento del negocio, se llevó a cabo un análisis detallado de las necesidades y objetivos de la organización. Este paso inicial proporcionó una base sólida para identificar la necesidad de segmentar a los clientes, mejorando así las estrategias de marketing y la satisfacción del cliente.

La etapa del entendimiento de los datos involucró la exploración exhaustiva de 39,515 registros de SAP Business One correspondientes al periodo de junio a agosto 2023. Durante esta fase, se identificaron 7 variables claves que influyen en el comportamiento de compra de los clientes tales como ZONA_ML", "Oficina", "Propiedad", "ESTATUS PRODUCTO", "ARTICULO_2", "Cantidad", "PRECIO". Esta comprensión profunda de los datos permitió avanzar hacia la preparación de los mismos, donde se llevaron a cabo tareas de limpieza, transformación, normalización y selección. El resultado fue la obtención de conjuntos de datos limpios y adecuados disponibles para su uso en la siguiente fase.

La etapa de modelado representó un componente crítico del proyecto, ya que involucró la aplicación de tres técnicas de agrupamiento: K-Means, K-Medoids y Jerárquico, las cuales desempeñaron un papel fundamental en la segmentación efectiva de los clientes en grupos con preferencias y comportamientos similares. Durante este proceso, se utilizaron métricas como el método del codo, coeficiente de inercia y el

coeficiente de silueta para determinar el número óptimo de clústeres, un aspecto de suma importancia para garantizar la precisión de la segmentación.

En la etapa de evaluación, se llevó a cabo una validación exhaustiva de los datos. A partir de los resultados obtenidos, se evidenció que los modelos K-Means y Clúster Jerárquico se ajustaban adecuadamente al conjunto de datos, mientras que el modelo K-Medoids no resultó apropiado debido a que en la validación con el coeficiente de silueta arrojó valores que indican que sus clústeres tienen baja cohesión y separación inefectiva, esto provocaba una superposición de los clústeres y no contribuía de manera correcta a la clasificación.

Para seleccionar el valor óptimo de *K*, se presentaron dos propuestas al experto, la primera utilizaba el modelo de K-Means con 4 clústeres, mientras que la segunda empleaba el modelo Jerárquico con 4 clústeres. Siguiendo la recomendación del experto, que optó por una combinación de K=4 y el algoritmo K-Means, ya que se ajustaba de manera más precisa a la realidad del negocio, se procedió a realizar un análisis minucioso de los resultados obtenidos, lo que permitió una categorización efectiva de los grupos identificados y fortaleció la confianza en los hallazgos alcanzados.

Finalmente, la etapa de implementación tradujo los resultados del modelado en estrategias concretas, adaptadas a las necesidades y preferencias específicas de cada clúster de clientes. Este enfoque integral y centrado en el negocio aseguró que las estrategias propuestas estuvieran alineadas perfectamente con los objetivos de la empresa y los comportamientos de los clientes. Sin embargo, es importante destacar que los patrones de compra y las preferencias de los clientes pueden cambiar con el tiempo, lo que subraya la necesidad de un monitoreo continuo y la adaptación constante de las estrategias para mantenerse al día con las tendencias del mercado y las evoluciones en el comportamiento del cliente. Asimismo, la colaboración con expertos y la formación de un equipo multidisciplinario deberían ser considerados para fortalecer las estrategias comerciales. La capacitación del personal de soporte y ventas es crucial para garantizar la efectiva implementación de las estrategias y brindar un servicio de alta calidad a los clientes. Establecer canales de retroalimentación con los clientes externos, como encuestas y llamadas de retorno, proporcionará información valiosa

sobre las experiencias y expectativas de los clientes, lo que permitirá realizar mejoras continuas y adaptar las estrategias para satisfacer sus necesidades en constante evolución.

6.2 Recomendaciones

Se recomienda que la empresa farmacéutica aproveche la segmentación de clientes en cuatro clústeres para personalizar las ofertas y promociones de acuerdo con las preferencias de compra de cada grupo. Esto implica la creación de estrategias de marketing específicas para cada clúster, destacando los productos que se ajustan a las preferencias de cada segmento. Por ejemplo, podrían implementarse ofertas especiales en inyectables para el clúster K0 o promociones en cremas para el clúster K1. Esta personalización puede aumentar la relevancia de las ofertas y la satisfacción del cliente.

Dado que los patrones de compra y las preferencias de los clientes pueden cambiar con el tiempo, es fundamental que la empresa establezca un sistema de monitoreo continuo. Esto implica seguir recopilando datos y aplicar técnicas de Machine Learning de forma constante para mantenerse actualizada con las tendencias del mercado y los cambios en el comportamiento del cliente. El monitoreo continuo permitirá ajustar las estrategias a medida que evoluciona el mercado.

De igual manera, consideramos que es importante continuar colaborando con expertos como los gerentes o supervisores de la empresa, como el que ha asistido en la validación de los resultados. Esta colaboración garantiza que los enfoques y algoritmos utilizados estén alineados con las mejores prácticas y las necesidades cambiantes del mercado.

La empresa debería considerar la formación de un equipo multidisciplinario que incluya especialistas en Machine Learning y análisis de datos a fin de robustecer las estrategias comerciales. Asimismo, a medida que se implementen nuevas estrategias, es importante capacitar al personal de soporte y de ventas, ya que permitirá que se generen sinergias para brindar un servicio más efectivo a los clientes.

Finalmente, recomendamos que la empresa establezca un canal de retroalimentación con los clientes externos, como encuestas o llamadas de callback, ya que proporcionaría información valiosa sobre las experiencias y expectativas de los clientes. Evaluar la satisfacción del cliente y recopilar sus opiniones ayudará a la empresa a realizar mejoras y a adaptar sus estrategias para satisfacer sus necesidades.

REFERENCIAS BIBLIOGRÁFICAS

Aguilar, L. & Vasquez Y. (2017) Principal Component Analysis (PCA) para mejorar la performance de aprendizaje de los algoritmos Support Vector Machine (SVM) y Red Neuronal Multicapa (MLNN). Universidad Privada Antenor Orrego. Recuperado de: https://hdl.handle.net/20.500.12759/3398

Alam, A.; Esa, H.; Nazrul, K. & Hossain, A. (2023). Data *Clustering*: Prospects & Challenges. International Journal of Creative Research Thoughts (IJCRT) 11. 2320-2882. Recuperado de: https://www.researchgate.net/publication/374294872_Data_*Clustering*_Prospects_Challenges

Amat, J. (2020) Análisis de componentes princiaples PCA con Python. Ciencia de datos. Recuperado de: https://cienciadedatos.net/documentos/py19-pca-python

Aprende Machine Learning (2018). K-Means en Python paso a paso. Recuperado de: https://www.aprendemachinelearning.com/k-Means-en-python-paso-a-paso/

Arévalo, B., & Aguilar, D. (2021). Análisis de la estrategia cross–selling para mejora del comercio del Mercado Las Manuelas – Durán. Universidad de Guayaquil. Recuperado de: http://repositorio.ug.edu.ec/bitstream/redug/57905/1/ICT-136-2021-T1-%20AR%C3%89VALO%20SALAZAR_AGUILAR%20JIM%C3%89NEZ.pdf

Camacho, J. (2021). *Clustering* Jerárquico con Python | JacobSoft. JacobSoft. Recuperado de: https://www.jacobsoft.com.mx/es_mx/*Clustering*-jerarquico-con-python/

Cámara de Comercio de Lima. (2020). Informe anual del mercado farmacéutico en Perú. Lima, Perú.

Cuadros, A.; Gonzales, C. & Jiménez, P.(2017). Análisis multivariado para segmentación de clientes basada en RFM. Tecnura, 21(54), 41-51. https://doi.org/10.14483/22487638.12957

Das, A. (2022). K-Medoid *Clustering* (PAM)Algorithm in Python - towards Data Science. Medium. Recuperado de: https://towardsdatascience.com/k-medoid-*Clustering*-pam-algorithm-in-python-with-solved-example-c0dcb35b3f46

De Silva, C. (2017) Principal component analysis (PCA) as a statistical tool for identifying key indicators of nuclear power plant cable insulation degradation. Iowa State University. https://lib.dr.iastate.edu/etd/16120

Educative. (s.f.). Educative Answers - trusted answers to developer questions. Recuperado de: https://www.educative.io/answers/what-is-the-k-medoids-algorithm

Entezami, A.; Sarmadi, H. & Razavi, B. (2020). An innovative hybrid strategy for structural health monitoring by modal flexibility and *Clustering* methods. Journal of Civil Structural Health Monitoring. Recuperado de: https://www.researchgate.net/publication/342871651_An_innovative_hybrid_strategy_for_structural_health_monitoring_by_modal_flexibility_and_*Clustering*_methods

Galán Cortina. (2015). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario. Universidad Carlos III de Madrid. Recuperado de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

GeeksforGeeks. (2023). Principal Component Analysis with Python. GeeksforGeeks. Recuperado de: https://www.geeksforgeeks.org/principal-component-analysis-with-python/

Gil, C. (2023). El Machine learning en la predictibilidad de la logística de compra de los clientes de suplementos deportivos en cuatro distritos de Arequipa. Dataismo. 2. 10.53673/data.v2i7.103. Recuperado de: https://www.researchgate.net/publication/369958348_El_Machine_learning_en_la_predictibilidad_de_la_logistica_de_compra_de_los_clientes_de_suplementos_deportivos_en_cuatro_distritos_de_Arequipa

Harvard Business Review. (2020). How the pharmaceutical industry is adapting to changes driven by COVID-19. Recuperado de https://hbr.org

IBM. (s.f.).¿Qué es el aprendizaje supervisado?. Recuperado de: https://www.ibm.com/mx-es/topics/supervised-learning

Kassambara, A. (2017). Practical Guide to Clúster Analysis in R: Unsupervised Machine Learning (1st ed.). STHDA. https://xsliulab.github.io/Workshop/2021/week10/r-Clúster-book.pdf

Kubiak, B, & Weichbroth, P. (2010). Cross- And Up-selling Techniques In E-Commerce Activities. Journal of Internet Banking and Commerce. 15. Recuperado de: https://www.researchgate.net/publication/290560244_Cross-_And_Up-selling_Techniques_In_E-Commerce_Activities

Madariaga, S. (2021) Machine Learning para predecir volúmenes operacionales de las líneas de negocio de Cenabast. Tesis para el título de magister. Universidad de Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/182317/Machine-Learning-para-predecir-volumenes-operacionales-de-las-lineas-de-negocio-de-Cenabast.pdf

Martín, S. (2023). Up selling y cross selling: qué son y por qué tienes que utilizarlos. Metricool. Recuperado de: https://metricool.com/es/cross-selling-upselling/

McKinsey & Company. (2020). Revitalizing pharmaceutical marketing in an era of disruption. Recuperado de https://www.mckinsey.com

Ministerio de Salud de Perú. (2018). Perfil farmacéutico del Perú. Recuperado de https://www.minsa.gob.pe.

Moreno, F. (2022). Qué es la segmentación del público objetivo y qué beneficios tiene. Thinking for Innovation. Recuperado de: https://www.iebschool.com/blog/segmentacion-publico-objetivo-beneficios-marketing-digital

Mosquera González, D. (2022). Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación. Universidad Nacional de Colombia. Recuperado de:

https://repositorio.unal.edu.co/bitstream/handle/unal/81631/1152704626.2022.pdf?sequence=3&isAllowed=y

Pan American Health Organization. (2018). Acceso y uso de medicamentos en el Perú. Recuperado de https://www.paho.org.

Procapitales. (2019). El mercado farmacéutico en el Perú: Desafíos y oportunidades. Recuperado de https://www.procapitales.org.

ProInversión. (2019). Oportunidades de inversión en el sector salud y farmacéutico en Perú. Recuperado de https://www.proinversion.gob.pe.

Purnomo, M.; Azzam, A. & Khasanah, A. (2020). Effective Marketing Strategy Determination Based on Customers *Clustering* Using Machine Learning Technique. Journal of Physics: Conference Series. 1471. 012023.0.1088/1742-6596/1471/1/012023.
Recuperado de:
https://www.researchgate.net/publication/339846574_Effective_Marketing_Strategy_Determination_Based_on_Customers_*Clustering*_Using_Machine_Learning_Technique

Ramirez, J. (2021, December 7). K-Means: Elbow method and silhouette. Medium. Recuperado de:
https://medium.com/@jonathanrmzg/k-Means-elbow-method-and-silhouette-e565d7ab87aa

Rodríguez, D.(2023). Método del codo (Elbow method) para seleccionar el número óptimo de clústeres en K-Means. Analytics Lane. Recuperado de: https://www.analyticslane.com/2023/06/09/metodo-del-codo-elbow-method-para-seleccionar-el-numero-optimo-de-Clústeres-en-k-Means/

Santander, C. (febrero de 2021) Análisis RFM para segmentación de clientes (Recency, Frecuency, Money). LinkedIn. Recuperado de:
https://es.linkedin.com/pulse/an%C3%A1lisis-rfm-para-segmentaci%C3%B3n-de-clientes-recency-money-santander

Statista. (2023). Global pharmaceutical market - statistics & facts. Recuperado de https://www.statista.com

Statista. (2023). Global pharmaceutical sales from 2020 to 2022, by region. Recuperado de https://www.statista.com.

Tibco. (s.f.) ¿Qué es el aprendizaje no supervisado? Recuperado de: https://www.tibco.com/es/reference-center/what-is-unsupervised-learning

Valles, M.; Salazar, L.; Injante, R.; Hernandez, E.; Juárez,J.; Navarro, J.; Pinedo, L. & Vidaurre, P. (2022) Density Based Unsupervised Learning Algorithm to Categorize College Students into Dropout Risk Levels. Data 2022,7,165. https://doi.org/10.3390/data7110165Academic

World Health Organization. (2019). Access to medicines. Recuperado de https://www.who.int

Zelcer, M. (2022) Machine learning y lógicas semióticas: el caso de la publicidad digital. La Trama de la Comunicación, vol. 26, núm. 2, Julio-diciembre 2022, pp. 15-31. Recuperado de:
https://www.redalyc.org/articulo.oa?id=323973878001

ANEXOS

Anexo 1

Conformidad por parte de la empresa para el uso de la data. El logo, nombre de la empresa y dirección no se visualizan por motivos de confidencialidad.

18 de octubre de 2023

Sr: Fernando José Miyagi Yamashiro

ASUNTO: Solicitud de Uso de Data para proyecto de investigación

Estimado Sr. Miyagi:

En relación a nuestra reciente conversación sobre mi proyecto de investigación de grado, solicito formalmente su autorización para el uso de la data del periodo junio - agosto 2023 de la empresa. Creo firmemente que los hallazgos y análisis derivados de este trabajo no solo aportarán a mi formación académica, sino que también pueden generar valor y proponer mejoras para la empresa.

Quiero reiterarte mi compromiso de garantizar la total confidencialidad de la información y asegurarte que será utilizada exclusivamente para fines académicos y de investigación en este proyecto.

Para confirmar su conformidad con esta solicitud, le agradeceríamos firmar esta carta.

Quedo atento a su respuesta y agradezco de antemano su colaboración y apertura.

Saludos cordiales,

Firma de autorización
Nombre: Fernando José Miyagi Yamashiro
Cargo: Coordinador General

FERNANDO MIYAGI
Coordinador General

Nota. Elaboración propia

Anexo 2

Entrevista con el experto de la empresa del sector farmacéutico

Fecha de la entrevista: miércoles 19 de octubre 2023 a las 20:00 horas
Duración: 20 minutos
Ubicación: Virtual - Google Meets

Entrevistador: Buenas noches Sr. Fernando, de antemano agradecemos su tiempo para esta reunión. Como parte de nuestro proyecto de tesis en nuestro curso de titulación, estamos desarrollando una propuesta para aplicar Machine Learning mediante la segmentación de clientes y la implementación de algoritmos de clúster. Nuestro objetivo principal es permitir que la empresa del sector farmacéutico pueda identificar patrones de compra entre sus clientes y, en base a esta información, categorizarlos en grupos específicos, con el fin de que ustedes puedan diseñar estrategias de venta y promoción más efectivas.

Entrevistado: Hola, con gusto. Adelante.

Entrevistador: Comencemos hablando de las variables que hemos utilizado en nuestro análisis. Hemos considerado 7 tales como "ZONA_ML", "Oficina", "Propiedad", "ESTATUS PRODUCTO", "ARTICULO_2", "Cantidad", "PRECIO". ¿Crees que estas variables son adecuadas y representativas para llevar a cabo la segmentación de clientes en el sector farmacéutico?

Entrevistado: Sí, en general, esas variables son cruciales para comprender el comportamiento de los clientes en el sector farmacéutico y, de hecho, son consistentes con las que solemos utilizar en mi área.

Entrevistador: Excelente. A continuación, me gustaría explicarte las dos técnicas que hemos aplicado en nuestro trabajo. Comencemos con la propuesta 1 (K-Means, 4 clústeres), Observamos que, en esta técnica, la variable "estatus producto" muestra una prevalencia de categorías como "regular" y "corto vencimiento", y estas son especialmente aplicables a artículos como comprimidos, cremas y suspensión oral. Además, hemos notado que las zonas donde se realizan más compras son Lima, Norte y Sierra, y las oficinas más predominantes son

LI, NO y SU. Por último, en la variable "propiedad", se destacan el circuito cerrado, mayorista y co-distribuidor en 3 de los clústeres.

Entrevistado: Entiendo. Los resultados de la propuesta 1 parecen muy adecuados y se ajustan a la realidad de nuestro negocio. Estos hallazgos relacionados con los productos y las ubicaciones son particularmente interesantes y podrían ayudarnos a definir estrategias comerciales específicas.

Entrevistador: Efectivamente. Continuemos con la propuesta 2 (Clúster Jerárquico, 4 clústeres). La diferencia más notable en esta técnica es la mayor segmentación de "artículo", con un aumento en la presencia de inyectables. Además, en la variable "oficina", hemos identificado la inclusión de la categoría "Capon". En base a lo indicado, ¿Cuál de las propuestas crees que sería la más adecuada y por qué?

Entrevistado: En mi perspectiva, la propuesta 1 parece ser la más adecuada y se ajusta bien a la realidad de nuestro negocio. Los resultados obtenidos se centran en productos y ubicaciones de manera precisa, lo que podría resultar muy valioso para nuestras estrategias comerciales. Tanto los resultados como las variables seleccionadas me parecen interesantes, por lo que me gustaría conocer más sobre las estrategias que han considerado.

Entrevistador: Claro, agradecemos tus comentarios y apreciamos tu tiempo. Serán de gran ayuda para finalizar nuestros análisis de ambas técnicas. Adicionalmente, te comentamos que ya nos reuniremos para poder proporcionarte más información acerca de las estrategias que hemos considerado en ambos modelos.

Entrevistado: De nada, gracias a ustedes por realizar este proyecto, que será de gran utilidad para nuestra área y nos permitirá generar valor mediante el desarrollo y ejecución de estrategias comerciales orientadas a nuestros clientes.

Anexo 3

Foto de la reunión con el experto de la empresa del sector farmacéutico

Anexo 4

Perfil del experto de la empresa del sector farmacéutico