Caracterización de clientes de la empresa New Stetic

David Andrés Quintero Patiño

Viviana Taborda Ospina

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor: PhD. Juan Sebastián Parra Sánchez

Universidad Católica Luis Amigó
Facultad de Ingenierías y
Arquitectura
Especialización en Big Data e Inteligencia de Negocios
Medellín, Colombia
2023

Dedicatoria

A nuestros padres por su esfuerzo, motivación y confianza, su creencia en nosotros ha sido el motor para seguir adelante.

A nuestros colegas especialistas por su apoyo motivación y enseñanzas en este año de compañía.

A las diferentes personas de la empresa New Stetic que de una u otra forma fueron parte esencial en el desarrollo de esta investigación.

“Siempre que escalamos una montaña, nos damos cuenta que hay otras por escalar”
Nelson Mandela

Agradecimientos

Principalmente a Dios por brindarnos las habilidades, competencias y los valores necesarios para lograr cada uno de los objetivos propuestos en nuestras vidas.

Seguidamente a nuestras familias, por su apoyo incondicional, confianza, disposición y comprensión incondicional.

A la empresa New Stetic, por brindarnos todas las herramientas e información necesaria, por el apoyo y las indicaciones dadas para llevar a cabo la investigación.

A nuestro asesor de trabajo de grado Juan Sebastián Parra Sánchez, por la constancia, dedicación, apoyo y paciencia con el desarrollo de la investigación, de igual forma por los conocimientos compartidos.

A los diferentes docentes que con sus enseñanzas compartieron sus conocimientos.

A todas las personas, compañeros y amigos que de una u otra forma motivaron y apoyaron la realización de este trabajo de grado.

Resumen

Esta investigación se plantea desde la necesidad de la empresa New Stetic por conocer la distribución de sus clientes para lograr la fidelización de sus marcas, lo cual representa un aumento en los ingresos que genera la venta de sus productos para el sector odontológico. Para darle solución a la necesidad de New Stetic, se ha tomado como muestra los datos de las transacciones de las ventas en los años comprendidos entre el 2017 y febrero de 2023, se contó con un total de 789 registros de clientes, donde se implementó la técnica de clustering por medio del algoritmo de K-Means realizado en Python. Para el cumplimiento de los objetivos, se utilizó las fases de la metodología CRISP-DM encontrando como resultado tres segmentos de clientes para la implementación de la estrategia comercial con la metodología *Design Thinking.*

Palabras clave: Segmentación, clustering, K-Means, Clientes, CRISP-DM y Design Thinking.

Tabla de contenido

Lista de figuras

Lista de tablas

1. Introducción

El trabajo se realiza con el objetivo de perfilar y caracterizar a los clientes de la empresa *New Stetic*, lo que permite planificar estratégicamente acciones comerciales y gestionar la interacción con cada uno de ellos. La segmentación de clientes ayuda a identificar características comunes dentro de la base de datos de clientes, lo que facilita el entendimiento de las necesidades de cada grupo.

El trabajo está pensado como una práctica de estrategias de análisis de clientes. Tradicionalmente, se realizaba de manera manual basándose en la observación y el conocimiento del negocio. Sin embargo, en la actualidad se pueden utilizar herramientas y procesos más avanzados para analizar grandes volúmenes de datos y descubrir patrones y variables comunes de manera más eficiente.

El método empleado en el trabajo consiste en la segmentación de clientes, que implica dividir a los clientes en grupos más pequeños y homogéneos en un mercado específico. Se buscó identificar las características de cada grupo para poder atenderlas de manera más precisa, ofreciendo productos y servicios adecuados a cada segmento.

Algunas limitaciones del trabajo pueden incluir la disponibilidad y calidad de los datos de clientes, así como la capacidad de análisis y recursos técnicos de la empresa. Además, es importante tener en cuenta que la segmentación de clientes es un proceso continuo y en constante evolución, ya que las necesidades y comportamientos de los clientes pueden cambiar con el tiempo. Por lo tanto, es necesario realizar actualizaciones periódicas y adaptar las estrategias según sea necesario.

2. Planteamiento del Problema

En un mundo cada vez más globalizado e interconectado, se han presentado grandes cambios principalmente en temas económicos, sociales, políticos y culturales. Así mismo, las empresas han cambiado la forma en la que administran sus recursos. Es aquí entonces donde es necesario perfilar o caracterizar sus clientes, con el propósito que desde su actividad económica se pueda planear estratégicamente acciones comerciales en los que se gestione la interacción con cada uno de estos.

La segmentación de clientes permite definir una estrategia y es un punto de partida para el análisis, detección de oportunidades e implementación de acciones personalizadas para los clientes. Con lo anterior, la segmentación de clientes le permitirá a la empresa New Stetic identificar características comunes dentro de su base de datos de clientes. Esta práctica puede ser habitual en estrategias de comunicación hechas manualmente, basándose en la observación de esos grupos y el conocimiento del negocio.

Por otra parte, buscar patrones y variables comunes es una tarea realizada por personal comercial, pero esa capacidad está limitada por el total de información (cantidad de clientes). Para lograr dicho propósito se implementa un proceso denominado segmentación de clientes, el cual es definido por BBVA como: “una tarea que consiste en dividir en pequeños grupos homogéneos de clientes en un mercado concreto” BBVA (2017). Su objetivo fundamental es el de poder determinar con precisión las necesidades de cada grupo, de tal manera que la empresa las pueda atender mejor, ofreciéndole a cada uno de ellos un producto o servicio adecuado.

3. Justificación

En la actualidad la empresa New Stetic no realiza análisis de machine learning en el área comercial de la empresa, lo cual hace que sea un área en el cual mediante la implementación de modelos y algoritmos se puedan lograr resultados favorables para la segmentación de clientes y mejorar el análisis de la información como se vienen concibiendo actualmente por los métodos tradicionales de administración.

Por lo cual a partir del análisis de la información que provee la empresa sobre sus clientes, ventas y productos que comercializa actualmente, permite iniciar el modelo de caracterización para el análisis del comportamiento de sus clientes, con el fin de comprender los patrones y propensión de compra y crear en tiempo real ofertas personalizadas para los clientes de la compañía.

De esta manera, el análisis que se estará desarrollando en este proyecto, permitirá retroalimentar positivamente al área comercial de New Stetic al darle una perspectiva diferente del cliente. Se revisarán factores de comportamiento, volumen y periodicidad de las compras, con el propósito de generar propuestas más productivas y mucho más direccionadas a estos, con respecto a qué productos ofrecer e inclusive los intervalos de los tiempos para ello, todo esto, a partir de los algoritmos que se empleen para el análisis de la información.

Con este proyecto se busca brindar herramientas al área comercial para que puedan realizar una mejor toma de decisiones y mejores estrategias comerciales con sus compradores y que esto a su vez, tenga un impacto positivo en otras áreas de la empresa

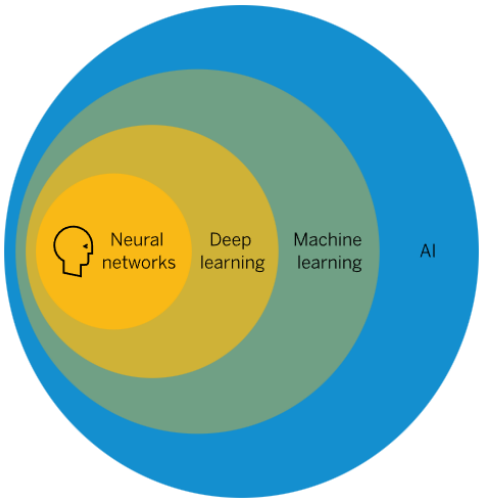
que se beneficien al ser más competentes en el momento de ofrecer sus productos, tener un mejor orden con respecto a la producción, al poder predecir más acertadamente en un nicho de clientes, donde se identifiquen los productos que requieren, en qué momento los pueden necesitar y la cantidad estimada de estos.

Actualmente el uso y manejo de datos genera una ventaja positiva en las organizaciones y en el mercado actual, por lo cual, es importante la implementación de nuestro proyecto para empresa New Stetic, debido a que ayudará de una forma más estratégica con las campañas comerciales.

4. Marco de Referencias

"Machine Learning, es una extensión de la Inteligencia Artificial encargada de desarrollar algoritmos que tienen capacidades de aprender y no tener que programarlos de manera explícita" (Sandoval., 2018, p.1). La definición de la inteligencia artificial, según diversos expertos no se debe considerar estática ya que con el avance de las tecnologías el concepto trae consigo nuevos aspectos e implicancias a lo largo de los años y del desarrollo de la misma. "la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cálculo inteligentes" (Corrales, 2021, p. 39).

Figura 1
Componentes de la Inteligencia artificial (IA)



***Nota*.** Gráfico de los componentes en la relación entre IA y machine learning. Tomada de SAP (2023)

4.1 Componentes del Machine Learning

- Las fuentes de información, comprenden datos, los cuales pueden ser: Estructurados (Bases de datos estructurada y organizada a partir de un reporte) y No estructurados (Aquellos que se encuentran dispersos en diversas fuentes de datos como mails, entre otros)
- Las técnicas y algoritmos, relacionados con las tareas a realizarlas: - Técnicas de tratamiento de datos no estructuradas: parsing, mapas autoorganizativos, etc. - Modelos de Machine Learning, supervisados y no supervisados: modelos de clasificación, regresión, estocásticos etc.
- La capacidad de autoaprendizaje, que mejora las medidas de desempeño para permitir el reentrenamiento automático a partir de nueva información para combinar modelos y ponderación/calibración.
- El uso de sistemas y software para la visualización de la información y la programación: - Visualización: Power BI, QlikView, QlikView, SAS Visual Analytics, TIBCO Spotfire, Tableau. - Programación: Java, Scala, Ruby, SAS, Matlab, C, Python, Azure, R, SQL (Anyela, 2022).

4.2 Tipo de aprendizaje – Machine Learning:

- **Aprendizaje supervisado:** Según (Gago, 2017, p.23) los datos en estos casos disponen de atributos adicionales que son los que se intentan predecir. Dentro de esta categoría destacan los algoritmos de clasificación, en los que las muestras están etiquetadas como como pertenecientes a dos o más clases y se requiere aprender a predecir la clase de datos sin etiquetar". Entre estos están: Algoritmo de Clasificación y Regresión.

- **Aprendizaje no supervisado:** los datos de entrenamiento consisten en un conjunto de vectores sin ningún valor o etiqueta correspondiente. El objetivo en estos casos puede ser descubrir grupos de ejemplos similares dentro de los datos". Entre este tipo tenemos Clustering, Algoritmo K – means, PCA (Análisis de componentes principales), Algoritmo K-Medoids (Palacios, 2020).
    - PCA (Análisis de componentes principales): es una de las técnicas más usadas para la reducción de dimensionalidad, mediante este algoritmo se busca reducir la cantidad de datos que se tiene a una menor escala pero que pueda representarlos de tal manera que se puedan evidenciar la correlación que existe entre estos. (Müller, 2016)
    - Algoritmo K – means: es un algoritmo de clustering o agrupamiento no supervisado que tiene como finalidad dividir un conjunto de datos en grupos o clusters homogéneos en función de la similitud entre ellos (Alpaydın, 2010). Adicional, uno de los objetivos principales es minimizar la suma de las distancias al cuadrado entre cada punto de datos y el centroide de su respectivo clúster. En otras palabras, el algoritmo busca encontrar los centroides que minimicen la varianza intra-cluster y maximicen la varianza inter-cluster. (Tan, 2006).

4.3 Estrategia comercial:

Según (Silva, 2020) la estrategia comercial es "conjunto de acciones que pone en práctica una empresa para dar a conocer un nuevo producto, para aumentar su cuota de venta o de participación de mercado".

Las empresas diariamente se enfrentan a la complicada tarea de generar valor para el cliente, adicional, se encuentran entornos en continuos cambios y un consumidor cada vez más exigente, deja claro que tener un buen producto no es suficiente para competir en el entorno actual.

a. **Creación de valor:** esto significa entender que es lo que desea el cliente, de forma que se pueda cubrir la necesidad y esperar satisfacerla. Son pocas las empresas que logran satisfacer esas necesidades y aún más complejo, estar en el primer lugar de recordación. Para generar valor se debe hacer estas preguntas: ¿Qué ofrece la empresa?, ¿Qué desea el cliente?, para la creación de valor. (Colomer, 2023)

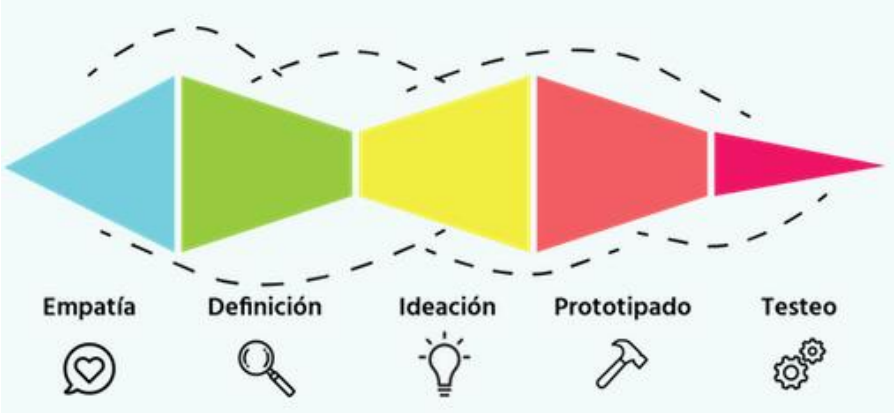
   De acuerdo a la necesidad para generar de valor, existen herramientas como *Design Thinking, Customer Jorney Map y Value* que brindan los recursos necesarios para la creación de una estrategia comercial acorde al dolor del cliente.

   Preguntas para hacer como negocio:

   - ¿cómo queremos que nos perciba el cliente?
   - ¿cómo podemos conocer qué desea el cliente?
   - ¿cómo materializamos esta necesidad?
   - ¿qué elementos aportan valor?

b. **Metodología Design Thinking (“Pensamiento de Diseño”):** es una metodología centrada en el usuario y orientada a la acción. El objetivo principal es generar soluciones de acuerdo a problemas detectados en un determinado marco de trabajo. (Guayara, 2021, p 25-26).

Figura 2
Metodología Design Thinking



Nota. Ilustración metodología Design Thinking. Tomada de Dinngo (2023)

Está metodología está enmarcada en tareas, esta a su vez se divide en múltiples pasos:

- **Seleccionar el usuario**: Encontrar un usuario para el proyecto de diseño.
- **Prepararse para realizar entrevistas:** Conocer una experiencia desde el punto de vista de tu usuario.
- **Selecciona un problema específico del usuario:** Identificar ideas de la entrevista del usuario para seleccionar un problema específico.
- **Generar ideas:** Evaluar las ideas de solución y seleccionar una para avanzar a la etapa de creación de prototipo
- **Prototipos:** Crear un modelo de tu solución para comunicar de manera efectiva la idea a tu usuario**.**

- **Conclusiones sobre el proceso de “Pensamiento de Diseño”:** Solicitar comentarios de los usuarios sobre el prototipo, los cuales ayudarán a iterar sobre el mismo y llegar a una solución sólida. Reflexionar sobre lo que aprendiste durante el proceso de diseño, con énfasis en la incorporación de los comentarios de los usuarios que contribuyen a una nueva iteración del prototipo. Dinngo (2023)

5. Antecedentes

Cuando se trata de mejorar el indicador de ventas en la empresa New Stetic, el factor cliente juega un papel importante a la hora de tomar decisiones, por tal motivo, en el marco de este trabajo de grado, cuyo objetivo es caracterizar los clientes de la empresa, haciendo uso de técnicas no supervisadas de machine learning, se abordaron desde la literatura algunos antecedentes relacionados con el objetivo de estudio.

Algunos autores han realizado investigaciones dando su punto de vista sobre la implementación de herramientas para la segmentación de clientes, permitiendo desarrollo de conocimientos gracias a la implementación de las metodologías existentes de Machine Learning.

En primer lugar, según Palacios (2020) se realizó una investigación que se plantea desde la necesidad de una empresa en la ciudad de Popayán, la cual genera sus ingresos con la venta de productos de consumo masivo y que a su vez quiere conocer la distribución de sus clientes para lograr la fidelización de su marca. Para darle solución a dicha problemática tomaron como muestra los datos de las transacciones del año 2019 de 2.837 clientes e implementaron la técnica de clustering por medio del modelo RFM en Excel y la implementación del algoritmo de K-Means realizado en Python y la herramienta Weka, utilizaron las fases de la metodología CRISP-DM dando como resultado 5 segmentos de los clientes de la empresa comercializadora de productos lácteos en el modelo RFM y 7 en K-Means, permitiéndole a la empresa el uso de esta información para generar estrategias de marketing (Palacios, 2020).

Por otra parte, en la ciudad de Bogotá se realizó la implementación de minería de datos para la segmentación de clientes para la empresa DPC Studio S.A.S. Para el análisis de los datos y poder lograr los objetivos de la investigación se utilizó un modelo no supervisado con la técnica de segmentación y la implementación del algoritmo K-Means que se encargó de clasificar los clientes a partir de un conglomerado de datos.

Esto dio como resultado tres tipos de clúster que son los más relevantes para la investigación teniendo en cuenta los productos del portafolio y los sectores más relevantes (Proaños, 2013) . Dando como resultado los siguientes beneficios para la empresa: aumento en el consumo de los portafolios, ingresos mensuales constantes, incremento en clientes fidelizados con la marca, un contante acompañamiento de expertos, el manejo de su presupuesto según la necesidad del consumidor y tarifas preferenciales entre otros.

En la ciudad de Bucaramanga, Rincón (2016), realizó el estudio del análisis del mercado objetivo de la empresa Madecentro Colombia S.A.S con sucursal en el Santander, por medio de la segmentación se conoció el perfil real de los clientes y los potenciales de la zona. Para esto se utilizaron las siguientes fuentes de información: a. Se realizaron encuestas personales a los clientes en los 3 puntos de venta de la zona Santander, estas encuestas se organizaron en las diferentes variables donde se agrupa el comportamiento de la compra y el uso del producto, b. para completar la información, se utilizaron las bases de datos suministradas por la empresa, específicamente en el área de mercadeo y Retail. Se obtuvo una base de datos con 1.000 registros de clientes a los que luego se aplicó la depuración y limpieza de la información, quedando como datos finales 242 registros de clientes. Se utilizó el software SPSS para el análisis del clustering jerárquico que permitió la segmentación de clientes por tipología, utilizando el método de agrupación de clúster Ward y una

media de distancia euclídea al cuadrado. Con esta investigación se pudo concluir que los clientes tienen una satisfacción favorable en el servicio, pero hay que mejorar en la calidad del producto y los precios.

Se destaca en los estudios anteriormente citados, una necesidad inherente de las compañías en realizar investigaciones basada en una solución inteligente de negocio, para apoyar la toma de decisiones en el área de ventas. Tener en cuenta el comportamiento de las ventas para realizar segmentos de clientes, permite crear estratégicas apoyadas en datos, gestión que facilita el aumento de las ventas con la implementación de nuevos portafolios, manejo del presupuesto según la necedad del cliente, clientes fidelizados, ingresos constantes en las ventas, entre otros. Adicionalmente, generan beneficios a los clientes como lo es un constante acompañamiento y tarifas preferenciales.

6. Objetivos

6.1 Objetivo General

Caracterizar los clientes de la empresa New Stetic, haciendo uso de técnicas no supervisadas de machine learning

6.2 Objetivos Específicos

- Realizar una revisión del estado del arte sobre Machine Learning para la segmentación de clientes.
- Aplicar un modelo óptimo para de segmentación de clientes de la empresa New Stetic soportado en machine learning.
- Realizar análisis y el despliegue de la información resultante del modelo de machine learning.

7. Viabilidad

Para la disponibilidad de recursos, se requiere el acceso a la base de datos con la información de las ventas de la empresa y equipos de cómputo que soporte el manejo de grandes volúmenes de datos (procesadores de 10ma o 11va generación, 16gb de RAM ddr4, puede ser portátil o pc de escritorio.

Instalar software de análisis de información (Microsoft SQL Server, Knime, Integración Services SSIS, Microsoft Excel)

El alcance se establece a partir de los datos obtenidos y el análisis de la información, hasta la caracterización de los clientes de la empresa New Stetic.

Las implicaciones iniciales están en la solicitud a la empresa New Stetic de la autorización del uso de la información, que garantice el cumplimiento del habeas data y el anonimato de la información de la empresa.

La consecuencia preponderante es poder identificar y caracterizar los clientes de la organización, donde se implementen estrategias de comunicación y promoción más eficaces, de manera que los análisis realizados permitan apalancar acciones personalizadas para los clientes.

8. Metodología

Para el trabajo de grado “caracterización de clientes de la empresa New Stetic" se empleará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) la cual proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. (Vallalta, 2023)

Figura 3
CRISP-DM: una metodología para minería de datos en salud

Nota. Ilustración metodología Crisp-DM. Tomada de (Vallalta, 2023)

El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea según los objetivos, el contexto, y el interés del usuario sobre los datos (Studer, 2021).

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él (Studer, 2021).

La metodología CRISP-DM establece un proyecto de minería de datos como una secuencia de fases, no obstante, cabe resaltar que estas fases si bien se desarrollan de forma consecutiva, existen partes del proceso en las que se realizan de forma iterativa, debido a que para obtener un modelo que responda a las necesidades del negocio se deben hacer constantes ajustes tanto en la fase de entendimiento del negocio y la comprensión de datos como en la preparación de datos y el modelamiento. A continuación, se profundizará en los pasos a desarrollados en cada una de las fases de la metodología CRISP – DM:

- Con respecto a la comprensión del negocio, se realizó el entendimiento del problema tal cual como se presentó en el objetivo general “Caracterización de clientes de la empresa New Stetic” y se validó con el objetivo general del negocio, de acuerdo a la información disponible para el proyecto, este se limitará a la caracterización de los clientes. Una vez acordados los objetivos de negocio con el proyecto a desarrollar estos se traducirán en el modelo de Machine Learning, que corresponde a una tarea no supervisada de clustering. Una vez definido esto, se establecerán, en conjunto con el negocio, los análisis del estudio y el despliegue del mismo.

- La siguiente fase corresponde a la comprensión de los datos, en esta se recopilaron las diferentes fuentes de información disponible para construir el set de datos de entrenamiento. La base de datos con la que se cuenta para el análisis es a partir del año 2017 hasta marzo del año 2023, con un total de 402.285 registros.

- La siguiente fase de la metodología corresponde a la preparación de datos, esto incluye tres pasos fundamentales: la limpieza de datos, construcción e integración de datos y finalmente la transformación y estandarización de variables.

- Para la fase de modelado se tomará como base el proceso K-medias, modelo no supervisado.

- Finalmente, para la parte de evaluación e implementación del modelo se realizará la caracterización de cada uno de los grupos resultantes de la clusterización para validar si estos aportaban información sobre los clientes. Para eso se caracterizarán los grupos de acuerdo con las variables que el negocio proporcionó como las determinantes en la evaluación de las ventas, estas incluyen: venta total, cantidades, fechas de transición de las compras, ciudad, departamento, medio de compra y el tipo de producto que compra.

- La secuencia de las fases no es rígida: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay que hacer después. Las flechas indican las dependencias más importantes y frecuentes.

9. Resultados

Con el propósito de cumplir con los objetivos propuestos, se cuenta con fuentes de información primaria y secundaria. La información primaria fue suministrada en una tabla de Excel, información que contiene datos de las transacciones de los clientes; con el fin de guardar la confidencialidad de estos datos, la información relevante fue sustraída y reemplazada, a cada cliente se le asignó un número único de identificación (Id_Cliente).

A continuación, se presenta el desarrollo de la etapa metodológica del presente estudio.

9.1 Pre- procesamiento:

Se describe paso a paso el proceso de manejo y recolección de la base de datos y procesamiento de la información.

a. **Información:** En este primer paso se obtuvo la base de datos del comportamiento en las ventas de la empresa New Stetic, periodos comprendidos entre el año 2017- 2023.

Tabla 1
Dataset ventas

CARACTERISTICAS DEL DATA SET	
Registros	402.285
Variables	13
Número de clientes	3247
Variables	Fecha de transacción, Zona, Ciudad, Grupo, Línea, Subgrupo, País, Subzona, Cantidad actual en pesos, Cantidad actual en inventario, Nombre del producto, Canal de distribución y Código del cliente

***Nota*.** Características de la dataset. Elaboración propia

En el cuadro anterior, se muestra una síntesis de las respectivas variables, lo cual da claridad de la información con la que se cuenta y de esta forma hacer correctamente los análisis.

b. Preparación de los datos: Con la información de la base de datos en una hoja de Excel, se da inicio al proceso de limpieza y depuración de los datos que no son útiles para el análisis.

Inicialmente se selecciona la información por la categoría Grupo y se decide trabajar con el grupo “Anestésico”.

Se depura las categorías País, Zona, Línea, Subgrupo, Subzona y Nombre del producto, información que no es relevante para el análisis.

Se consolida las cantidades de compra de cada uno de los clientes con un identificador único (Id_Cliente), se agrupa por los meses del año, se totaliza las ventas en cantidades “Total_general” y en pesos “Venta_Total”. Para continuar con el desarrollo del trabajo, se eligen las variables Id_Cliente, Canal_Distribución, Ciudad, “Venta_Total”, “Cantidad_mes” y “Total_general”

Tabla 2
Dataset Ventas_Final

CARACTERISTICAS DEL DATA SET final	
Registros	789
Variables	17
Número de clientes	789
Variables	Id_Cliente, Canal_Distribución, Ciudad, “Venta_Total”, “Cantidad_ene”, “Cantidad_feb”, “Cantidad_mar”, “Cantidad_abr”, “Cantidad_may”, “Cantidad_jun”, “Cantidad_jul”, “Cantidad_ago”, “Cantidad_sep”, “Cantidad_oct”, “Cantidad_nov”, “Cantidad_dic” y “Total_general”

***Nota*.** Características de la dataset final. Elaboración propia.

9.2 Procesamiento:

Siendo consecuentes con lo planteado en los objetivos y basados en el estudio realizado en la revisión bibliográfica, se determina que los algoritmos de K-means y DBSCAN son métodos más idóneos el ámbito de la segmentación de clientes por clusters. Se hace uso de la herramienta Google Colab.

a. Cargar y observar el conjunto de datos:

El primer paso que se realizó para la comprensión de los datos fue cargar y visualizar el dataset haciendo uso de la librería pandas, tal y como se muestra a continuación.

Figura 4

Características de la dataset.

	ID_Cliente	Canal_Distribución	Ciudad	Venta_Total	Cantidad_ene	Cantidad_feb	Cantidad_mar	Cantidad_abr	Cantidad_may	Cantidad_ jun	Cantidad_jul	Cantidad_ago	Cantidad_sep	Cantidad_oct	Cantidad_nov	Cantidad_dic	Total_general
0	788807	MAYORISTA	Bogotá, D.C.	14696600	NaN	60.0	30.0	80.0	200.0	30.0	126.0	80.0	45.0	NaN	NaN	NaN	651
1	889665	MAYORISTA	Ipiales	2112902	35.0	25.0	NaN	NaN	NaN	NaN	25.0	NaN	30.0	NaN	NaN	NaN	115
2	999999	CLIENTE DE CONTADO	Medellín	15099185	30.0	429.0	99.0	86.0	35.0	104.0	NaN	NaN	NaN	NaN	NaN	NaN	783
3	5766966	MAYORISTA	San Gil	1063120	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	40.0	40
4	5879797	MAYORISTA	Pasto	20351700	130.0	30.0	30.0	100.0	50.0	NaN	100.0	NaN	100.0	199.0	150.0	NaN	889
5	5999688	MAYORISTA	Pasto	580333400	1500.0	3150.0	2000.0	1360.0	1750.0	2500.0	1500.0	2520.0	2150.0	3650.0	1380.0	1600.0	25050
6	6778098	MAYORISTA	Pasto	193155600	900.0	550.0	450.0	531.0	900.0	1060.0	450.0	640.0	730.0	955.0	900.0	730.0	8796
7	6790869	MAYORISTA	Manizales	92578470	667.0	270.0	232.0	251.0	270.0	360.0	210.0	375.0	280.0	290.0	350.0	130.0	3675
8	6808708	MAYORISTA	Cartagena De Indias	42854212	NaN	100.0	300.0	50.0	NaN	135.0	154.0	350.0	200.0	200.0	100.0	100.0	1689
9	6866970	MAYORISTA	San Andrés	4563810	NaN	NaN	NaN	53.0	30.0	40.0	NaN	30.0	100.0	NaN	NaN	NaN	253

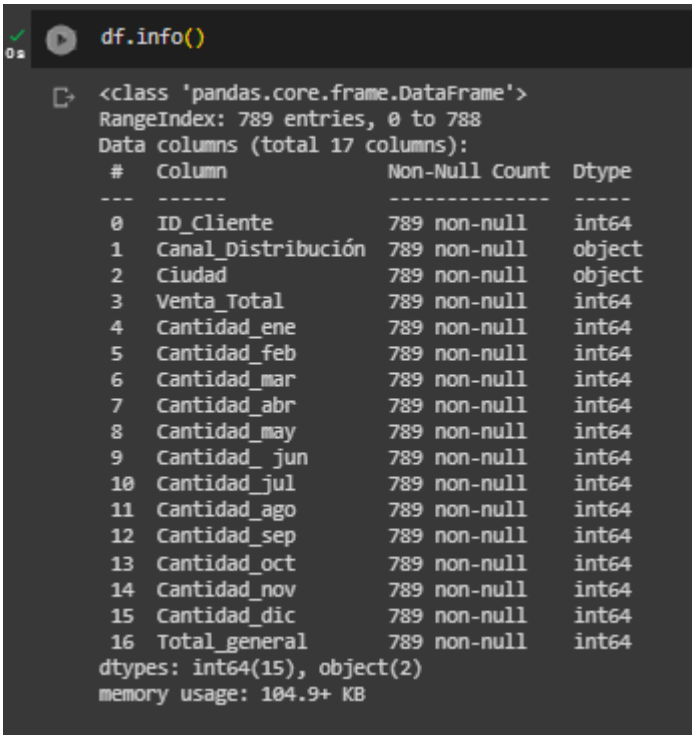
***Nota*.** Se detalla cada una de las variables con las que cuenta la dataset, vista previa en la herramienta Google Colab. Elaboración propia.

La visualización de los datos es necesaria para encontrar y tomar medidas adicionales, si faltan valores o están duplicados. Este proceso tiene como objetivo preparar nuestros datos para analizarlos y visualizarlos.

La visualización de la variable canal de distribución es importante, ya que permite identificar y representar gráficamente la participación en ventas por cada uno de los medios que dispone la empresa para llegar a sus clientes. De igual forma facilitará el plan para la estrategia de ventas.

Figura 5

Grafica identificación de las variables del dataframe



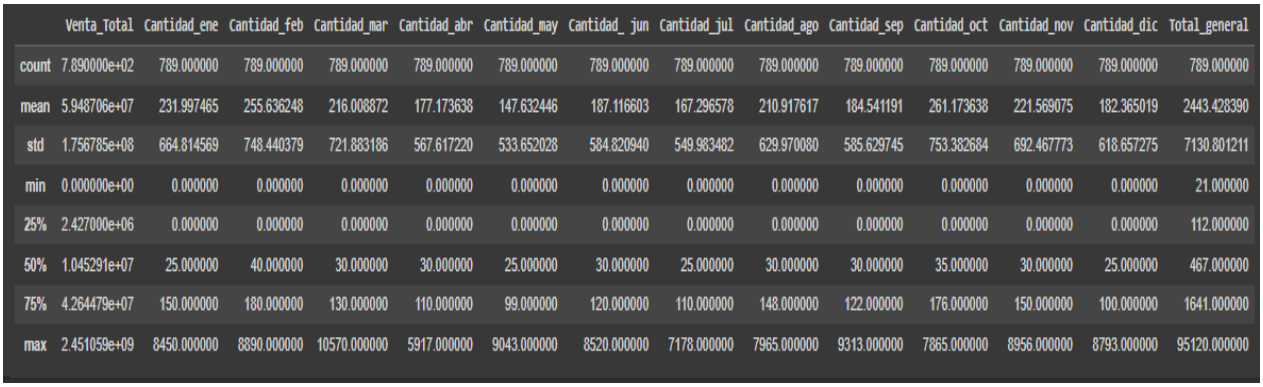
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 789 entries, 0 to 788
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID_Cliente          789 non-null    int64
 1   Canal_Distribución  789 non-null    object
 2   Ciudad              789 non-null    object
 3   Venta_Total         789 non-null    int64
 4   Cantidad_ene        789 non-null    int64
 5   Cantidad_feb        789 non-null    int64
 6   Cantidad_mar        789 non-null    int64
 7   Cantidad_abr        789 non-null    int64
 8   Cantidad_may        789 non-null    int64
 9   Cantidad_ jun       789 non-null    int64
 10  Cantidad_jul        789 non-null    int64
 11  Cantidad_ago        789 non-null    int64
 12  Cantidad_sep        789 non-null    int64
 13  Cantidad_oct        789 non-null    int64
 14  Cantidad_nov        789 non-null    int64
 15  Cantidad_dic        789 non-null    int64
 16  Total_general       789 non-null    int64
dtypes: int64(15), object(2)
memory usage: 104.9+ KB
```

Nota. Podemos ver que el conjunto de datos consta de 789 datos en 17 variables. Por lo tanto, parece que el marco de datos no tiene valores erróneos. Elaboración propia

Figura 6

Visualización estadística de datos

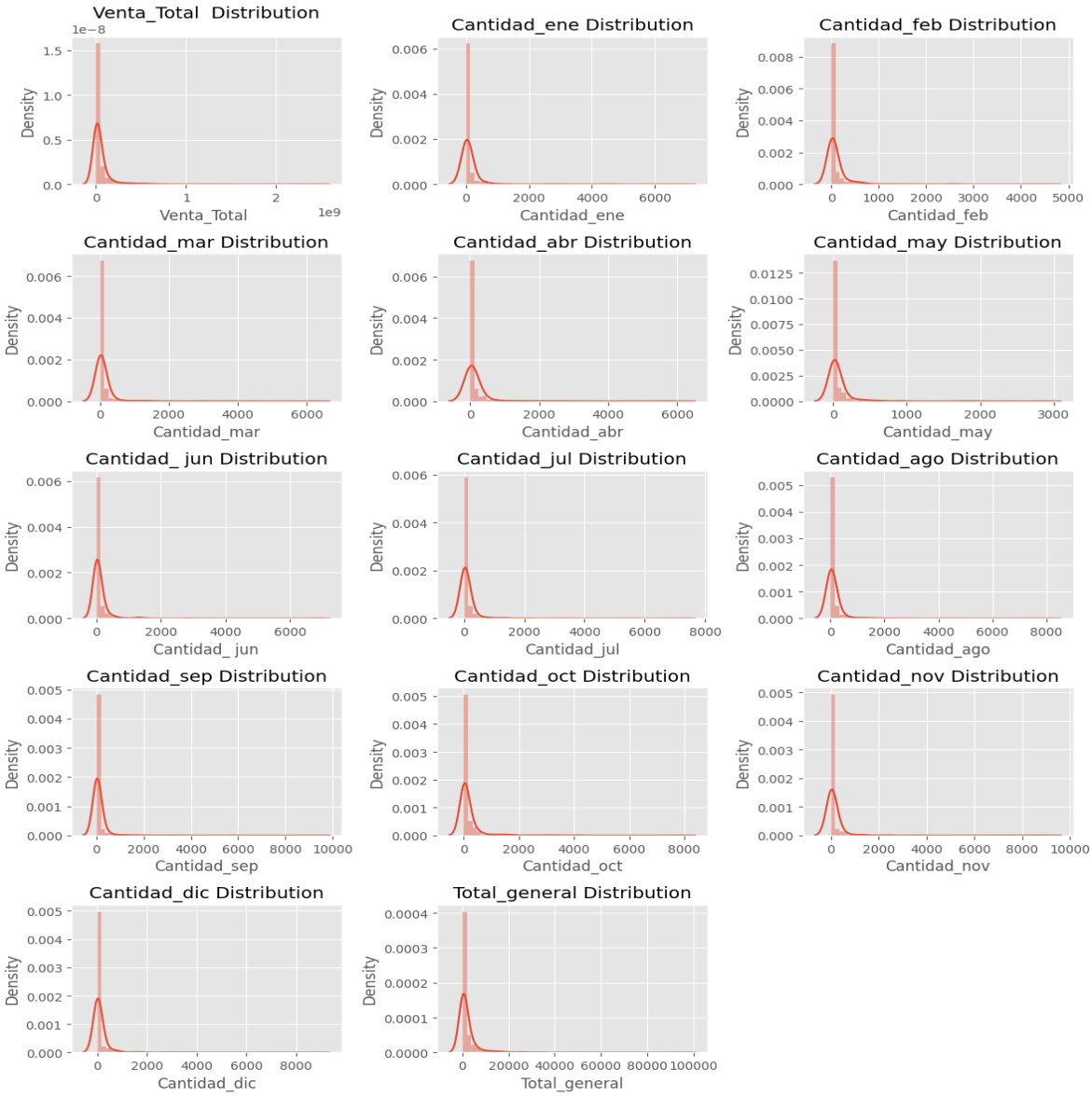


	Venta_Total	Cantidad_ene	Cantidad_feb	Cantidad_mar	Cantidad_abr	Cantidad_may	Cantidad_ jun	Cantidad_jul	Cantidad_ago	Cantidad_sep	Cantidad_oct	Cantidad_nov	Cantidad_dic	Total_general
count	7.890000e+02	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000	789.000000
mean	5.948706e+07	231.997465	255.636248	216.008872	177.173638	147.632446	187.116603	167.296578	210.917617	184.541191	261.173638	221.569075	182.365019	2443.428390
std	1.756785e+08	664.814569	748.440379	721.883186	567.617220	533.652028	584.820940	549.983482	629.970080	585.629745	753.382684	692.467773	618.657275	7130.801211
min	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	21.000000
25%	2.427000e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	112.000000
50%	1.045291e+07	25.000000	40.000000	30.000000	30.000000	25.000000	30.000000	25.000000	30.000000	30.000000	35.000000	30.000000	25.000000	467.000000
75%	4.264479e+07	150.000000	180.000000	130.000000	110.000000	99.000000	120.000000	110.000000	148.000000	122.000000	176.000000	150.000000	100.000000	1641.000000
max	2.451059e+09	8450.000000	8890.000000	10570.000000	5917.000000	9043.000000	8520.000000	7178.000000	7965.000000	9313.000000	7865.000000	8956.000000	8793.000000	95120.000000

Nota. Procedemos a visualizar las estadísticas descriptivas del conjunto de datos mediante “df.describe()” Para todos los datos numéricos, el índice del resultado incluirá recuento, media, estándar, mínimo y máximo, así como percentiles inferiores, medio (mediana) y superior. Elaboración propia

Figura 7

Grafica histograma de las variables



***Nota*.** Histograma de las variables, en esta imagen se observar el comportamiento individual de cada una de las variables las cuales presentan un sesgo a la derecha lo que puede indicar la presencia de valores atípicos o una distribución asimétrica en tus datos. Elaboración propia

Para confirmar la información que está mostrando el histograma, procedemos a visualizar las variables mediante un diagrama de caja para validar como es el comportamiento de los datos con respecto a la media, mediana y que comportamiento presentan los valores atípicos.

Figura 8

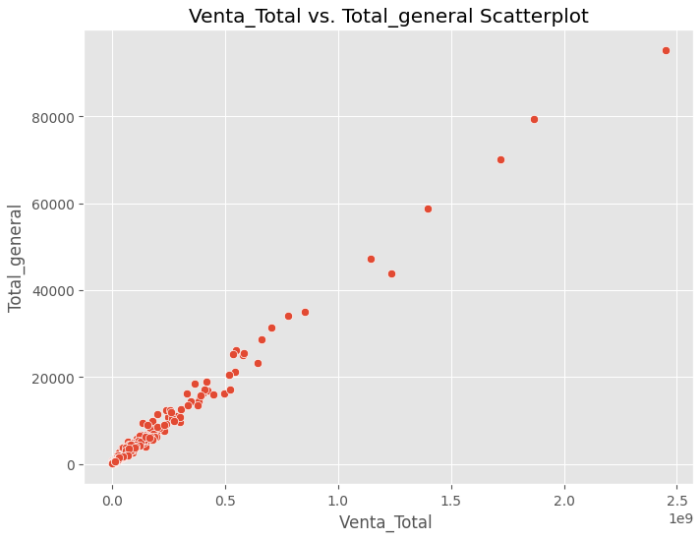
Diagrama de caja para cada variable numérica

Nota. Podemos evidenciar que el “Total_general” y “Venta_total” son las variables que presentan mayor variación en los datos atípicos por lo cual debemos evaluarlas para tomarlas como referencia para la agrupación. Elaboración propia.

Posteriormente, se procede a realizar un análisis bivariado para examinar la relación que presentan estas 2 variables y como se relacionan estas mediante un diagrama de dispersión.

Figura 9

Se genera un gráfico de dispersión entre la venta total y total general



Nota. Se genera un diagrama de dispersión de las variables numéricas “Total_general”, “Id_Cliente” y “Cantidad vendidas por mes”, existe una correlación positiva entre estas dos variables, es decir, a medida que aumentan las cantidades, aumentan las ventas (total general). Elaboración propia.

Figura 10

Correlación entre las dos variables

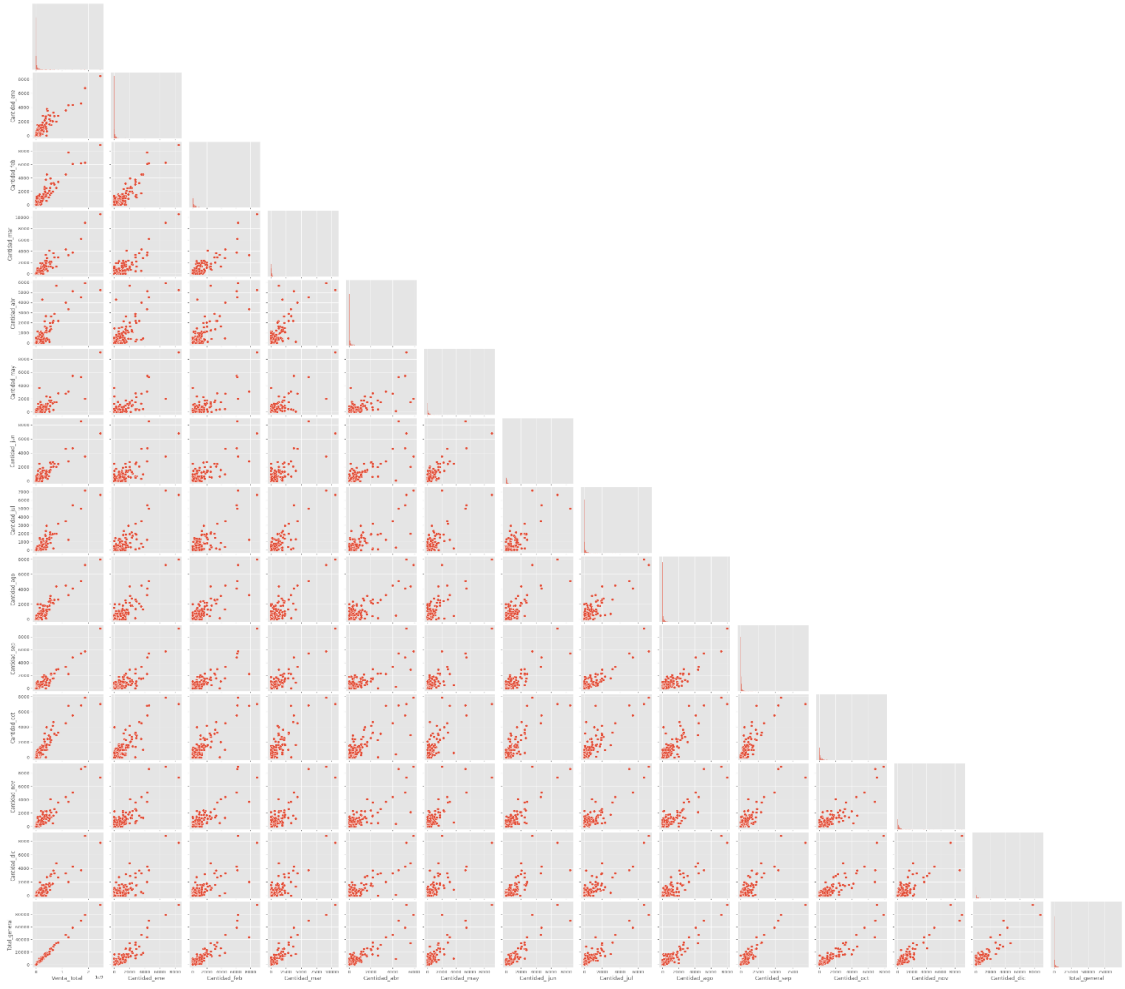
```
print('La correlación entre el Total_general y el Venta_Total es : {}'.format(round(df.corr()['Venta_Total ']['Total_general'],3)))
La correlación entre el Total_general y el Venta_Total es : 0.996
```

Nota. Se confirma la correlación existente entre las 2 variables con una mayor precisión podemos usar la función. “corr()” para mostrar el valor de correlación de Pearson (como método predeterminado) entre las dos variables. Elaboración propia.

Es importante destacar que una correlación alta no implica necesariamente causalidad. Aunque estas dos variables están altamente correlacionadas, es posible que existan otros factores o variables no considerados que puedan influir en esta relación. Es recomendable realizar un análisis más detallado y considerar el contexto y la teoría subyacente antes de sacar conclusiones definitivas, por lo cual consideramos realizar un análisis multivariante con el fin de analizar más de dos variables simultáneamente.

Además, este análisis nos ayudará a determinar la correlación entre todas las variables en el conjunto de datos.

Figura 11
Grafica Pair Plot

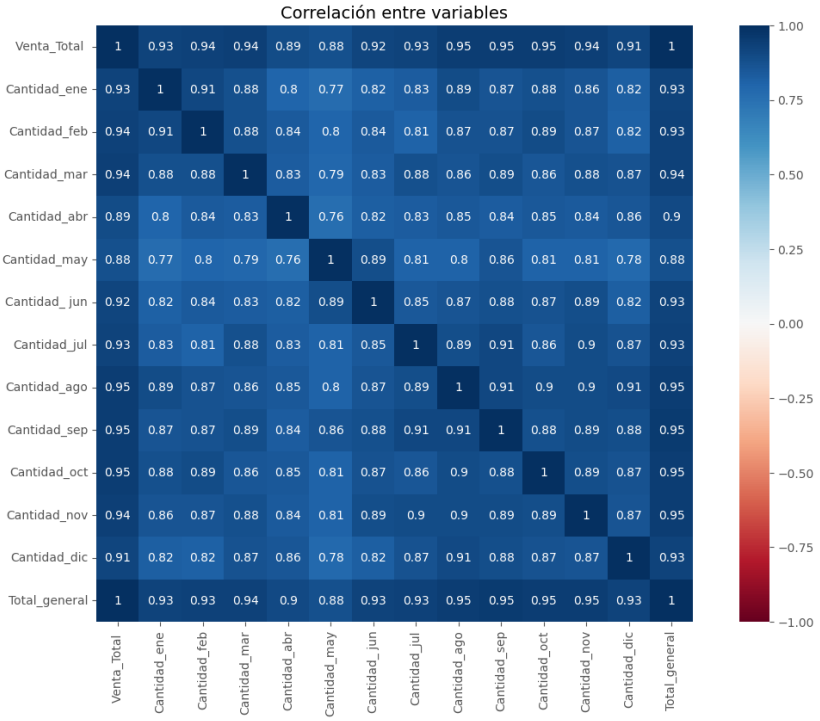


Nota. Mediante Pair Plot se pueden ver las relaciones entre pares de variables en nuestro conjunto de datos, Cada gráfico de dispersión muestra la relación entre dos variables, y la matriz completa de gráficos de dispersión proporciona una visión general de las relaciones entre todas las combinaciones de variables. Elaboración propia.

Podemos visualizar que las variables del dataset están estrechamente relacionadas positivamente entre ellas, también se pueden evidenciar los valores atípicos en algunas de estas por lo que debemos verificar mediante la matriz de correlación cual es relación de una manera más precisa.

Figura 12

Se genera la matriz de correlación con los pesos



Nota. Se genera la matriz de correlación para verificar la correlación entre las variables, es una forma de determinar cómo están conectadas las cosas. Los coeficientes de correlación se pueden interpretar entre más cercanas estén al 1, mayor correlación tienen estas, en el análisis podemos evidenciar una fuerte correlación entre todas las variables del dataset. Elaboración propia.

Se adjunta tabla 3 para analizar la correlación perfecta que existe entre las variables “Total_general” y “Cantidad vendidas por mes”.

Tabla 3

Interpretación de los coeficientes de correlación

± 0,00	± 0,09	Correlación nula
± 0,10	± 0,19	Correlación muy débil
± 0,20	± 0,49	Correlación débil
± 0,50	± 0,69	Correlación moderada
± 0,70	± 0,84	Correlación significativa
± 0,85	± 0,95	Correlación fuerte
± 0,96	± 1,0	Correlación perfecta

-	Relación negativa
+	Relación positiva

Nota. Tabla para la interpretación de los coeficientes de correlación entre las variables numéricas de la base de datos (“Total_general” y “Cantidad vendidas por mes”). Elaboración propia.

Para confirmar la información que está mostrando el histograma procedemos a visualizar las variables mediante un diagrama de caja para validar como es el comportamiento de los datos con respecto a la media, mediana y que comportamiento presentan los valores atípicos.

b. Limpieza de valores atípicos y normalización:

Para reducir la cantidad de clientes que seleccionamos, podemos filtrar los datos atípicos que presentan una venta total por debajo de las ventas medias para identificar de mejor manera el grupo de clientes que sobre los cuales debemos definir una mejor estrategia y sobre cuales debemos conservar o reforzar la estrategia que se viene aplicando actualmente. Se utiliza la mediana porque la distribución de datos de la variable de “Venta_Total” tiene una forma sesgada, por lo que es aconsejable utilizar el valor de la mediana en lugar de la media.

Figura 13
Normalización de los datos mediante la mediana

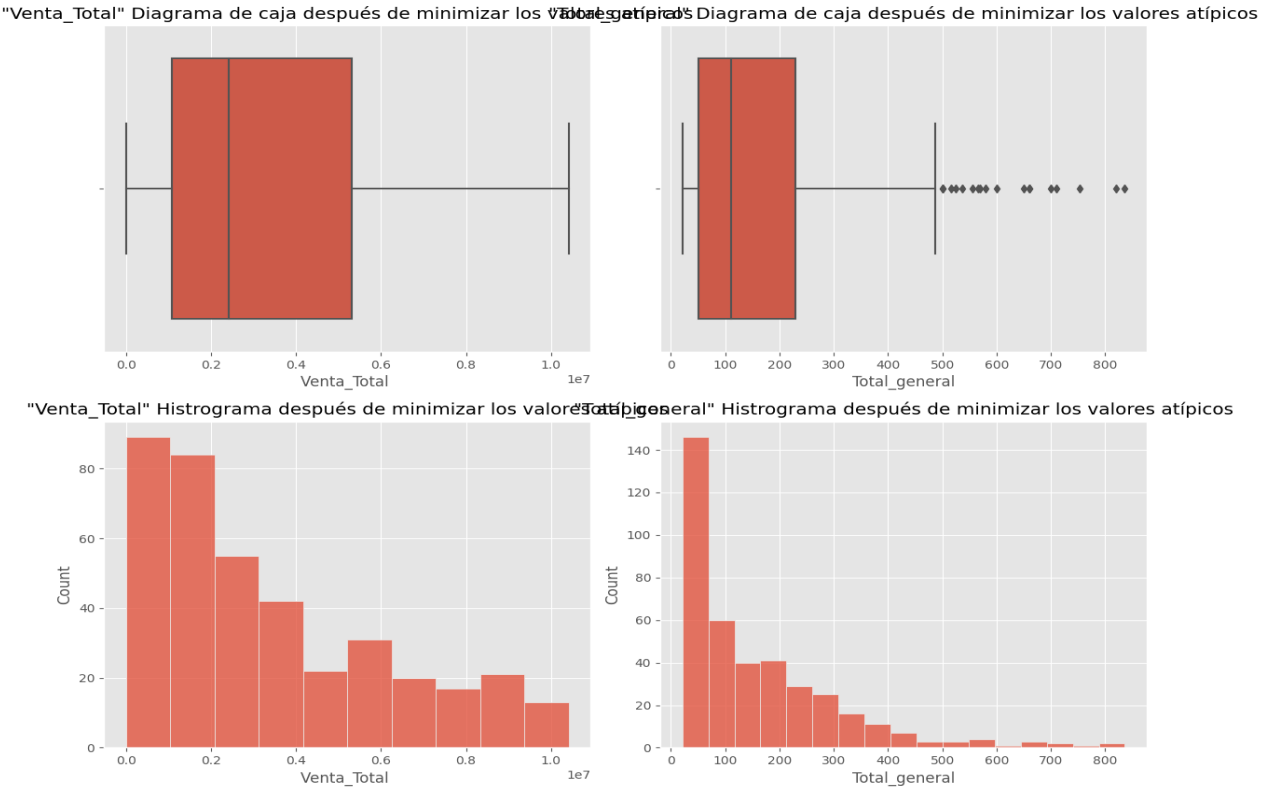
	ID_Cliente	Canal_Distribución	Ciudad	Venta_Total	Cantidad_ene	Cantidad_feb	Cantidad_mar	Cantidad_abr	Cantidad_may	Cantidad_ jun	Cantidad_jul	Cantidad_ago	Cantidad_sep	Cantidad_oct	Cantidad_nov	Cantidad_dic	Total_general
1	889565	MAYORISTA	Ipiales	2112902	35	25	0	0	0	0	25	0	30	0	0	0	115
3	5786965	MAYORISTA	San Gil	1063120	0	0	0	0	0	0	0	0	0	0	0	40	40
9	6866970	MAYORISTA	San Andrés	4553810	0	0	0	53	30	40	0	30	100	0	0	0	253
10	6907695	MAYORISTA	Bogotá, D.C.	628175	25	0	0	30	0	0	0	0	0	0	0	0	55
15	7709575	MAYORISTA	Neiva	470800	0	40	0	0	0	0	0	0	0	0	0	0	40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
775	9998878766	MAYORISTA	Turbo	2406570	0	30	0	0	0	0	0	0	0	0	0	30	60
776	9998865697	MAYORISTA	Montería	886016	0	24	0	0	0	0	0	0	0	0	0	0	24
777	56079078	MAYORISTA	Urrao	8999660	0	0	0	0	0	100	0	0	0	112	0	100	312
781	900078697	MAYORISTA	Cali	858159	0	0	0	0	0	30	0	0	0	0	0	0	30
784	900866080	MAYORISTA	Neiva	2344980	0	51	0	0	0	0	0	0	0	0	0	0	51

394 rows × 17 columns

Nota. Acá podemos evidenciar que con la aplicación de la mediana se redujeron los datos a la mitad, dejando los clientes sobre los cuales se debe hacer un mayor enfoque estratégico. Elaboración propia.

Ahora, con solo un simple filtrado anterior, podemos reducir nuestra selección en casi un 50%. Después de filtrar, procederemos a verificar el histograma y el diagrama de caja de la variable “Venta_Total” y “Total_general”.

Figura 14
Visualización limpieza de los datos

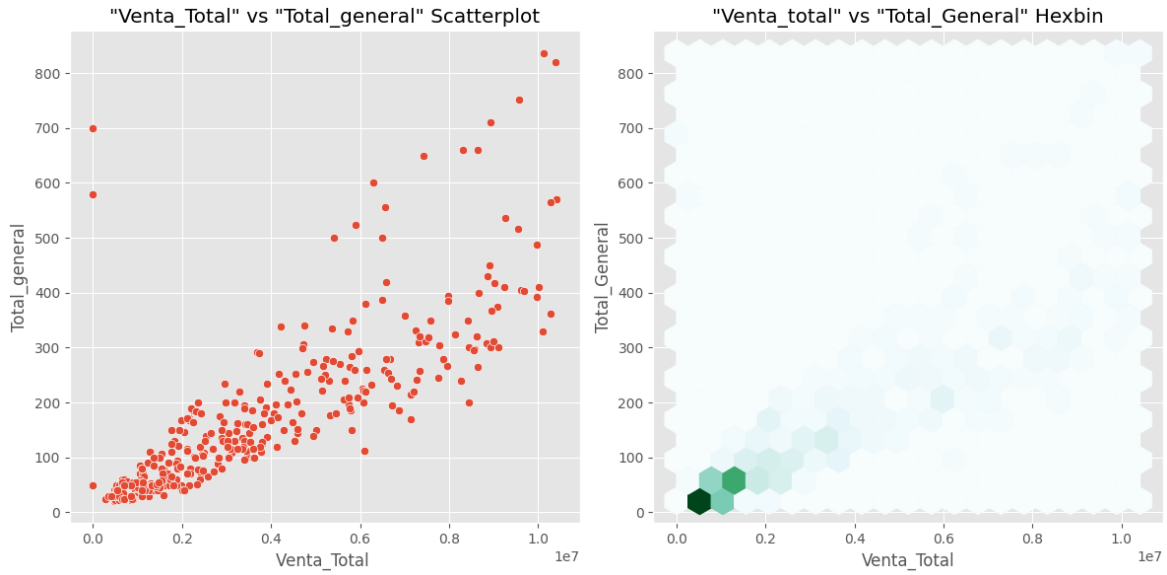


Nota. podemos evidenciar que hay una mayor limpieza en los datos y una disminución en los valores atípicos que teníamos en el análisis inicial que realizamos a estas 2 variables. Elaboración propia.

Con base en los gráficos anteriores, podemos proceder con confianza al siguiente paso para la agrupación porque los gráficos no muestran tantos valores atípicos como el grafico anterior, lo que indica que nuestro conjunto de datos ahora está más limpio que antes, antes de hacer el agrupamiento, debemos verificar la correlación de las variables que elegimos usando un gráfico de dispersión y la función. “hexbin()” de la biblioteca “matplotlib”.

Figura 15

Correlación de las variables posterior a la reducción de los valores atípicos



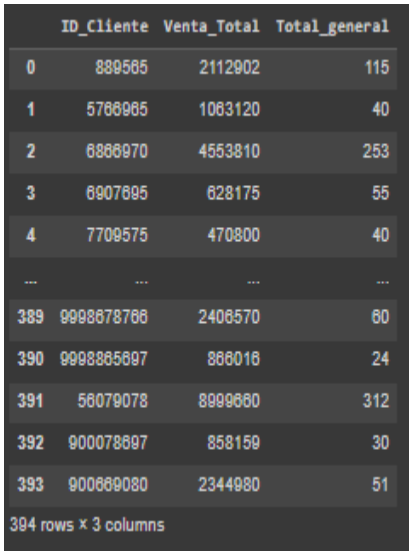
Nota: podemos evidenciar que continúa existiendo una correlación positiva entre ambas variables después de reducir los datos atípicos del dataset. Elaboración propia.

9.3 Modelado:

Para la caracterización de los clientes, se realizó una segmentación utilizando el método de k- means (técnica de clusterización) con el fin de identificar el comportamiento de las compras que realizan los clientes.

Antes de aplicar los algoritmos, debemos realizar el escalado de características, el cual consiste en alterar características numéricas para que tengan la misma escala. Dado que puede afectar significativamente el rendimiento de nuestro algoritmo, es un paso de preprocesamiento de datos crucial para la mayoría de los algoritmos de aprendizaje automático.

Figura 16
Nuevo dataset de los datos principales



	ID_Cliente	Venta_Total	Total_general
0	889565	2112902	115
1	5766965	1063120	40
2	6866970	4553810	253
3	6907695	628175	55
4	7709575	470800	40
...	...	...	...
389	9998678766	2406570	60
390	9998865697	866016	24
391	56079078	8999660	312
392	900078697	858159	30
393	900669080	2344980	51

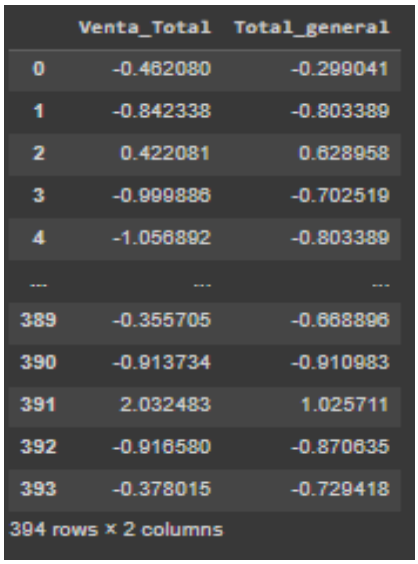
394 rows × 3 columns

Nota: Antes de realizar el escalado se crea un nuevo dataset que contenga los datos principales para el análisis, en este caso “Venta_Total”, “Total_general” y “ID_Cliente”. Elaboración propia.

Ahora, se puede comenzar a escalar usando StandardScaler () de la biblioteca scikit-learn

con el fin de estandarizar características numéricas en un conjunto de datos.

Figura 17
Estandarización de las características numéricas



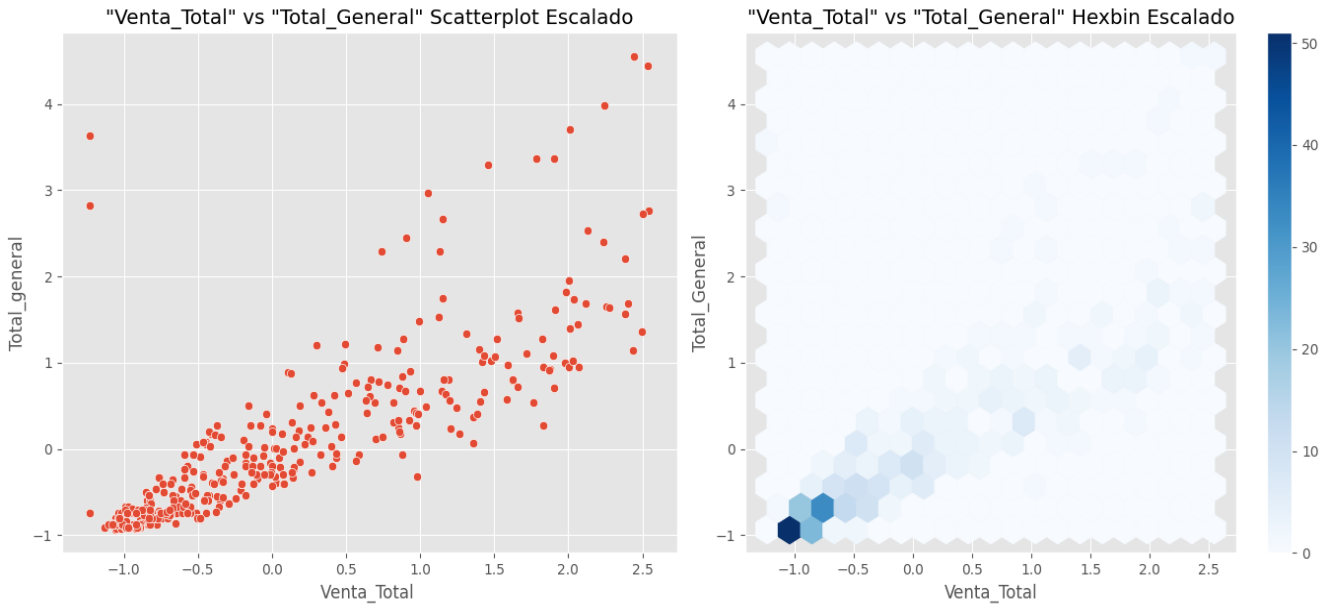
	Venta_Total	Total_general
0	-0.462080	-0.299041
1	-0.842338	-0.803389
2	0.422081	0.628958
3	-0.999886	-0.702519
4	-1.056892	-0.803389
...	...	...
389	-0.355705	-0.668896
390	-0.913734	-0.910983
391	2.032483	1.025711
392	-0.916580	-0.870635
393	-0.378015	-0.729418

394 rows × 2 columns

Nota. StandardScaler estandariza las características numéricas al ajustarlas para que tengan una media de cero y una desviación estándar de uno, lo que ayuda a garantizar que las características estén en la misma escala y facilita el análisis y entrenamiento de modelos de aprendizaje automático. Elaboración propia.

Una vez aplicada la estandarización, debemos verificar que la correlación de nuestras variables escaladas visualizándolas para asegurarnos de que hemos realizado el escalado correctamente.

Figura 18
Correlación de las variables después de escaladas



Nota. Se puede visualizar que posterior al escalado de las variables se sigue conservando la correlación positiva entre estas. Elaboración propia.

Al evidenciar que no hay diferencia en la correlación que se muestra antes de realizar el escalado, y después de escalar, podemos continuar con la selección de clústeres para aplicar K-Means.

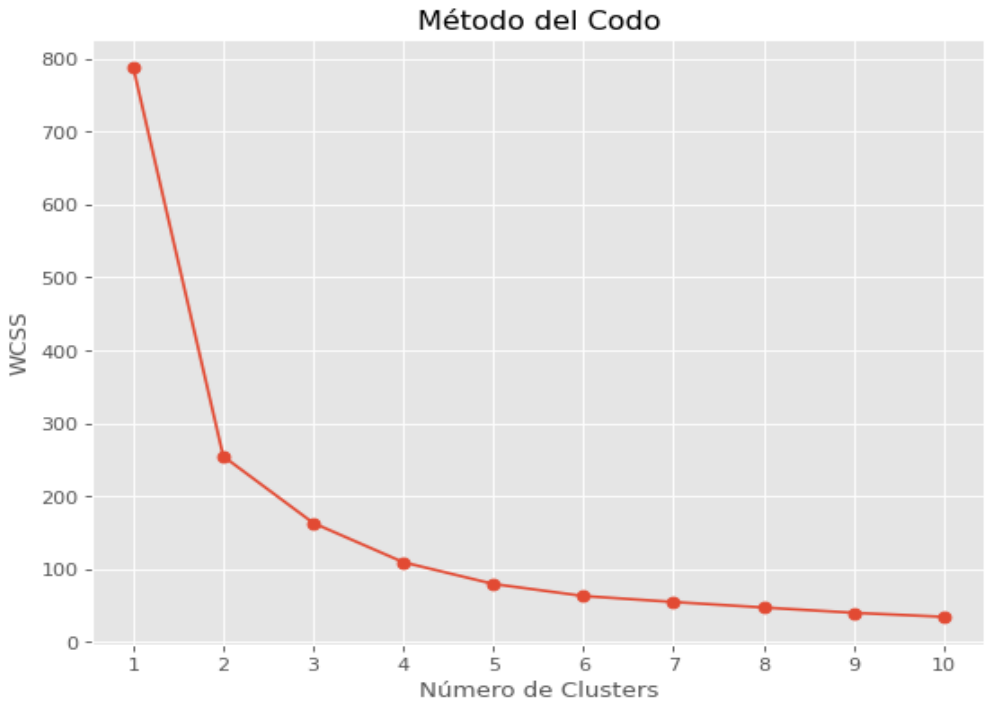
a. Aplicación de K- Means:

La agrupación en clústeres es parte de un algoritmo no supervisado en Machine Learning, el objetivo del agrupamiento es organizar los datos en grupos con características relacionadas y categorizarlos.

Se realizaron las iteraciones para obtener el número de clúster optimo

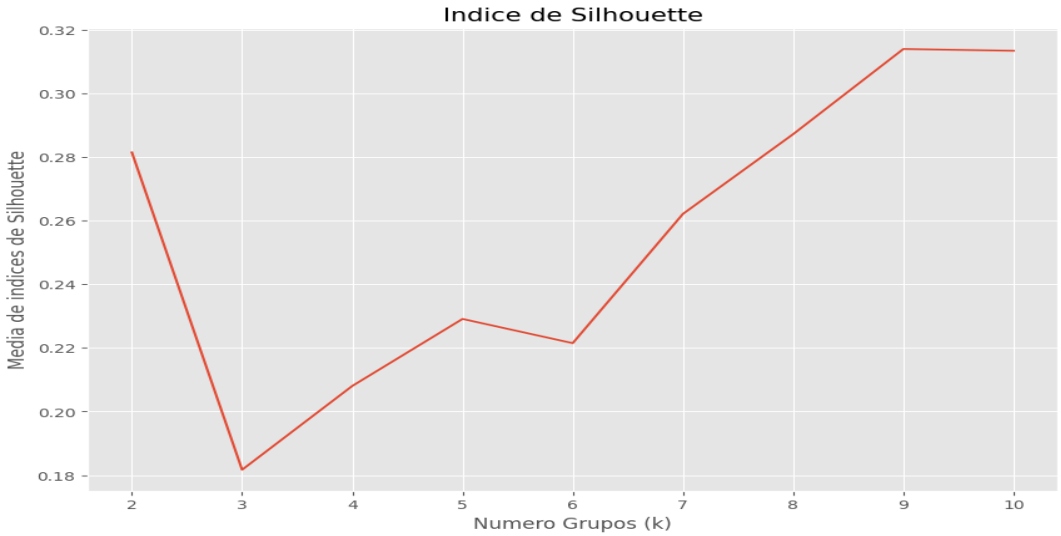
Figura 19

Método del codo



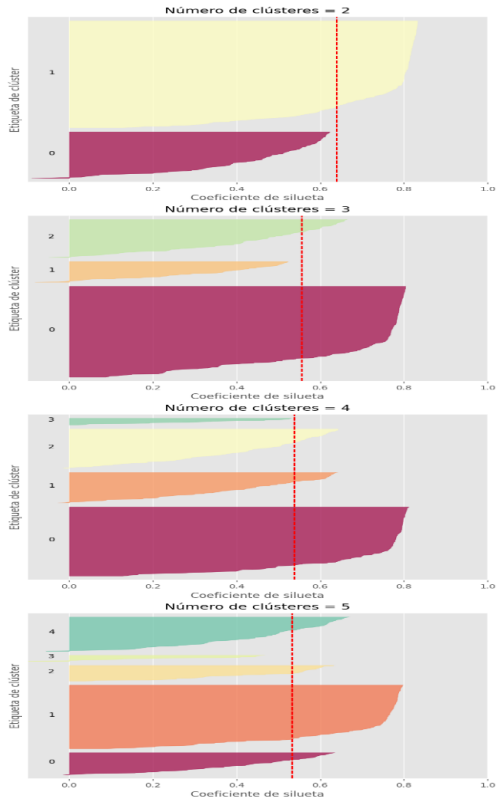
Nota. En la imagen se observa la implementación del método del codo, este nos ayuda a elegir un número apropiado de clusters para agrupar los datos, en este caso nos muestra que la mejor opción para continuar con el agrupamiento es 3 clusters óptimos. Elaboración propia

Figura 20
Método de Silhouette



Nota. En la imagen se observa la implementación del método de Silhouette, este método al igual que el método del codo nos ayuda a elegir un número apropiado de clusters para agrupar los datos. Elaboración propia.

Figura 21
Coeficiente de silueta

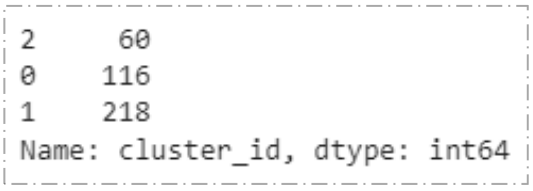


Nota. En la imagen se observa la implementación del coeficiente de Silueta, este método al igual que el método del codo, nos ayuda a elegir un número apropiado de clusters para agrupar los datos, por su distribución este método nos muestra 3 clusters óptimos. Elaboración propia.

El paso de obtener el número óptimo de clusters, nos ha dado un punto de partida para iniciar el entrenamiento del algoritmo. Una vez definidos los 3 clusters, se procede a visualizarlos.

Figura 22

Se visualiza la distribución del número de clientes por clusters



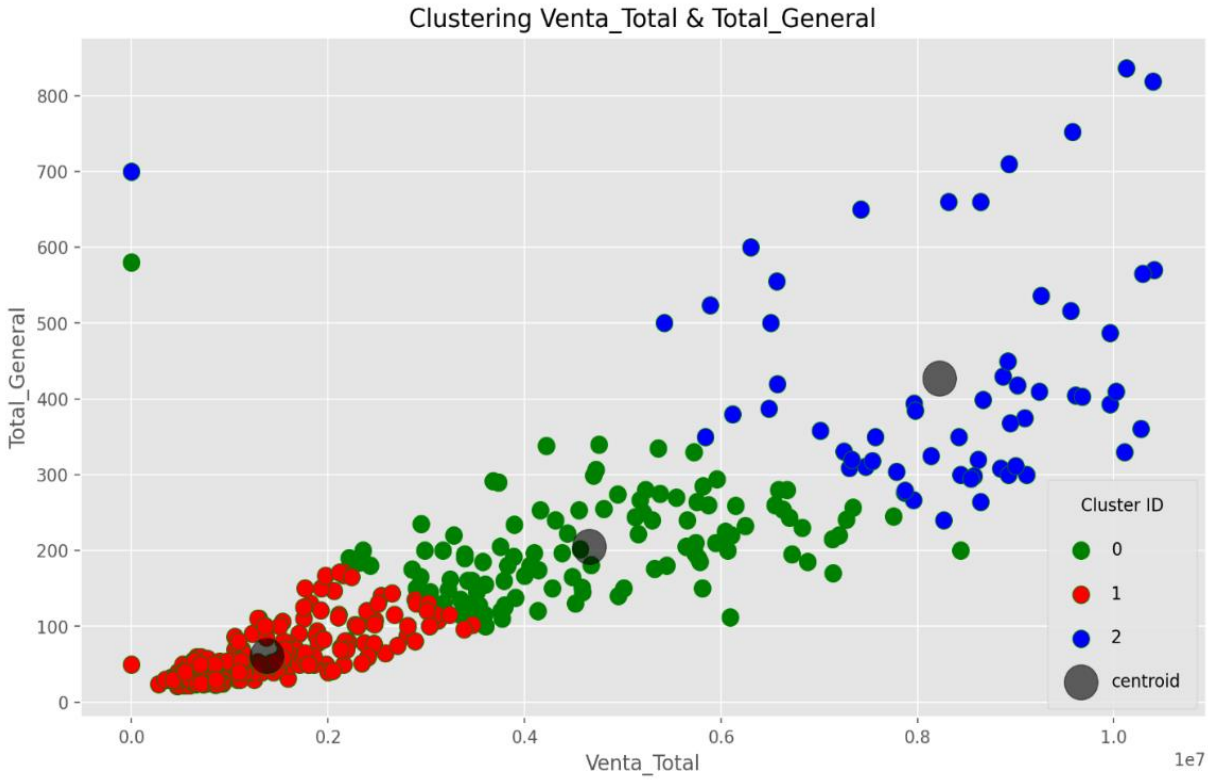
***Nota*.** El clúster 1 contiene 116 clientes, el clúster 2 contiene 218 clientes y el clúster 3 contiene 60 clientes, Elaboración propia

b. Visualización de los centroides:

Se realiza la visualización de los centroides de cada clúster con las cuales está más cercano - punto equidistante de los objetos pertenecientes a cada clúster, teniendo en cuenta el valor medio de las observaciones.

Figura 23

Se visualizan los centroides de cada clúster



Nota. Visualización de los 3 clúster (rojo – clúster 1, verde- clúster 0 y azul clúster 2. Se define por el método del Codo y el método de Silhouette que el número de clúster óptimo es 3. Elaboración propia.

Al extraer la información que nos da cada uno de los clusters con sus centroides, podemos visualizar que en el grupo1 se generan menos ventas totales y menores ingresos, el grupo0 genera ventas totales mediana y un ingreso medio, y el grupo2 genera mayores ventas y mayor volumen de unidades totales vendidas (Figura 23).

Figura 24

Visualización datos que hacen parte del clusters 0

	ID_Cliente	Venta_Total	Total_general	cluster_id
2	6866970	4553810	253	0
5	7780887	4056540	180	0
11	9957699	5325397	176	0
13	9999989	3226119	149	0
15	55959985	4285460	150	0
...	...	...	...	...
375	9958886695	5768590	190	0
376	9958999896	4704440	299	0
381	9980568879	3166177	200	0
382	9980785909	6721615	195	0
383	9989699679	6694940	243	0

116 rows × 4 columns

Nota. Visualización de algunos datos que hacen parte del clúster 0, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 0). Elaboración propia

Figura 13

Figura 25

Visualización datos que hacen parte del clusters 1

	ID_Cliente	Venta_Total	Total_general	cluster_id
0	889565	2112902	115	1
1	5766965	1063120	40	1
3	6907695	628175	55	1
4	7709575	470800	40	1
6	7967789	1755700	110	1
...	...	...	...	...
388	9996890676	2103892	112	1
389	9998678766	2406570	60	1
390	9998865697	866016	24	1
392	900078697	858159	30	1
393	900669080	2344980	51	1

218 rows × 4 columns

Nota. Visualización de algunos datos que hacen parte del clúster 1, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 1). Elaboración propia.

Figura 26

Visualización datos que hacen parte del clusters 2

```
cluster2 = df_cluster[df_cluster.cluster_id == 2]
display(cluster2)
```

	ID_Cliente	Venta_Total	Total_general	cluster_id
12	9978676	6506600	500	2
14	50968965	8665306	399	2
27	58865878	7304300	309	2
47	60856757	5892268	524	2
54	66997808	9261640	536	2
64	69580566	10411800	570	2
65	69696060	8439000	300	2
78	76759597	8642450	660	2
83	77085786	7867350	277	2
94	78998907	8619420	320	2
100	79678908	9108300	300	2

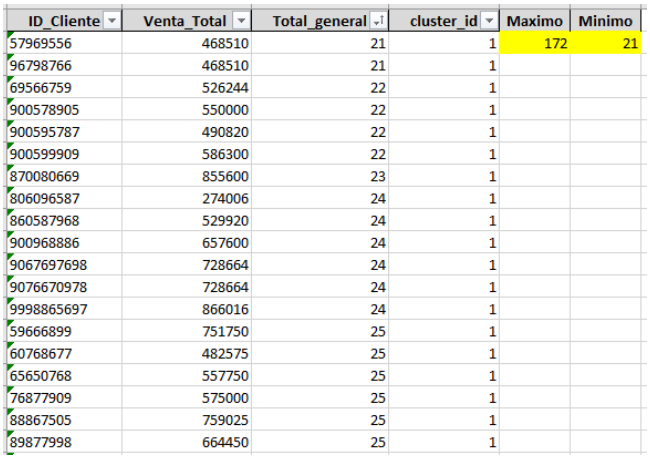
Nota. Visualización de algunos datos que hacen parte del clúster 2, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 2). Elaboración propia

Aplicando la técnica de agrupación a la base de datos, se encontraron grupos con factores reverentes, que permiten identificar perfiles de clientes. Se seleccionaron k=3, siendo esto los 3 tipos de clúster que representan perfiles e identidades de clientes para la investigación.

- **Clúster 1**: se caracteriza por ser un grupo representado por 218 clientes que han realizado transacciones en compras inferiores a las 172 unidades. Este segmento está representando a los clientes que tienen menor participación en las cantidades vendidas.

Figura 27

Descarga en Excel y visualización de los datos del clusters 1



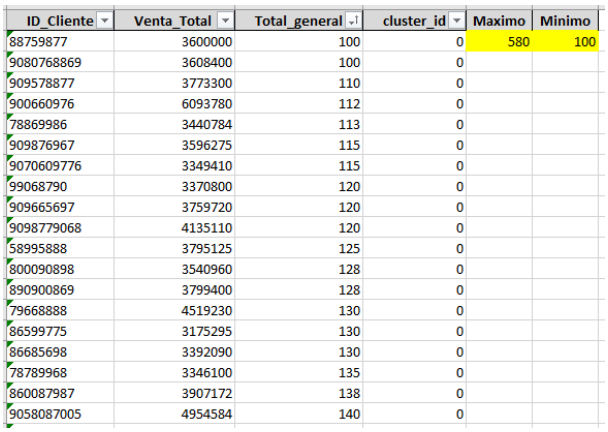
ID_Cliente	Venta_Total	Total_general	cluster_id	Maximo	Minimo
57969556	468510	21	1	172	21
96798766	468510	21	1		
69566759	526244	22	1		
900578905	550000	22	1		
900595787	490820	22	1		
900599909	586300	22	1		
870080669	855600	23	1		
806096587	274006	24	1		
860587968	529920	24	1		
900968886	657600	24	1		
9067697698	728664	24	1		
9076670978	728664	24	1		
9998865697	866016	24	1		
59666899	751750	25	1		
60768677	482575	25	1		
65650768	557750	25	1		
76877909	575000	25	1		
88867505	759025	25	1		
89877998	664450	25	1		

Nota. Se descargan y se visualizan en una hoja de Excel los datos que hacen parte del clúster 1, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 1). Elaboración propia

- **Clúster 0:** se caracteriza por ser un grupo representado por 116 clientes que han realizado transacciones en compras entre las 100 y 580 unidades. Este segmento representa a los clientes que tienen una participación promedio intermedio en las ventas de la compañía. De esta forma, se decide desarrollar la estrategia de fidelización de la marca con este grupo de clientes - “Clúster 0”.

Figura 28

Descarga en Excel y visualización de los datos del clusters 0



ID_Cliente	Venta_Total	Total_general	cluster_id	Maximo	Minimo
88759877	3600000	100	0	580	100
9080768869	3608400	100	0		
909578877	3773300	110	0		
900660976	6093780	112	0		
78869986	3440784	113	0		
909876967	3596275	115	0		
9070609776	3349410	115	0		
99068790	3370800	120	0		
909665697	3759720	120	0		
9098779068	4135110	120	0		
58995888	3795125	125	0		
800090898	3540960	128	0		
890900869	3799400	128	0		
79668888	4519230	130	0		
86599775	3175295	130	0		
86685698	3392090	130	0		
78789968	3346100	135	0		
860087987	3907172	138	0		
9058087005	4954584	140	0		

Nota. Se descargan y se visualizan en una hoja de Excel los datos que hacen parte del clúster 0, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 0). Elaboración propia

- Clúster 2: se caracteriza por ser un grupo representado por 60 clientes que han realizado transacciones en compras entre las 240 y 836 unidades. Este segmento representa a los clientes que tienen una participación mayoritaria en las ventas.

Figura 29

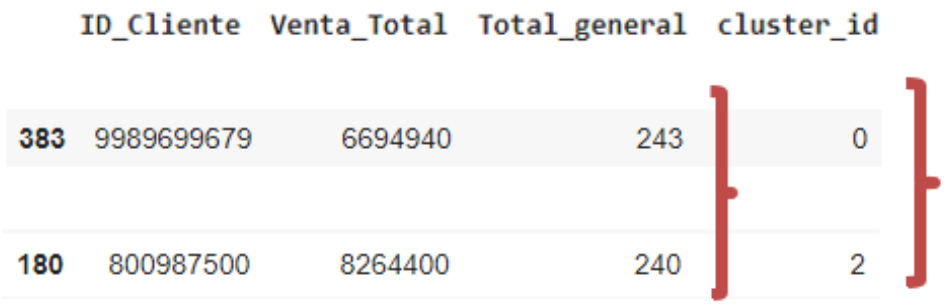
Descarga en Excel y visualización de los datos del clusters 2

ID_Cliente	Venta_Total	Total_general	cluster_id	Maximo	Minimo
800987500	8264400	240	2	836	240
900786688	8641700	264	2		
99577865	7957700	267	2		
77085786	7867350	277	2		
900779698	7872180	279	2		
900700700	8546185	295	2		
88975589	8576740	298	2		
69696060	8439000	300	2		
79678908	9108300	300	2		
909688659	8925200	300	2		
9989976689	7784800	304	2		
800968087	8845900	308	2		
58865878	7304300	309	2		
900708765	7470455	311	2		
56079078	8999660	312	2		
9057778005	7538879	318	2		
78998907	8619420	320	2		
9088669899	7333200	320	2		
909768887	8139270	325	2		

Nota. Se descargan y se visualizan en una hoja de Excel los datos que hacen parte del clúster 2, se puede observar el Id_Cliente con las cantidades totales (Total_general), la venta en pesos (Venta_Total) y el clúster al que pertenece (Clúster 2). Elaboración propia

Figura 30

Visualización de algunas "novedades" en los datos de los clusters 2 y 0



	ID_Cliente	Venta_Total	Total_general	cluster_id
383	9989699679	6694940	243	0
180	800987500	8264400	240	2

Nota. Existe una correlación entre las unidades vendidas y el valor de las ventas, sin embargo, por las promociones y/o descuentos realizados en las ventas, se pueden observar algunas cantidades similares presentes en diferentes clústeres, ejemplo el clúster 2 contienen cantidades a partir de 240 unidades y en el clúster 0 se observan cantidades hasta las 340 unidades. Elaboración propia

9.4 Implementación de la estrategia comercial:

El factor común para encontrar el éxito en las organizaciones, es la identificación de las tendencias del mercado y la definición de un modelo de negocio que enseñe y referencie la forma de generar valor comercial. Por lo anterior, se pretende enseñar un modelo sencillo pero integral cuyo propósito es determinar las palancas claves para saber llegar y comunicar de una mejor forma al cliente.

Teniendo en cuenta el resultado obtenido en la segmentación de los clientes por cada clúster, se decide aplicar la estrategia comercial a los clientes pertenecientes al clúster 0, puesto que son clientes potenciales para mejorar los niveles de ventas. Para el segmento de los clientes que hacen parte de los clusters 1 y 2 se decide continuar con la estrategia comercial actual.

Se despliegan los pasos para la implementación de la metodología Design Thinking:

- **Seleccionar el usuario**: Se tienen determinados 116 usuarios pertenecientes al clúster 0., clúster seleccionado para desarrollar la estrategia comercial.

- **Prepararse para realizar entrevistas:** Las entrevistas estarán enmarcadas en crear valor, escuchar el dolor del cliente – las necesidades. A partir de las preguntas de entrevista, obtener notas y conclusiones claves.

- **Selecciona un problema específico del usuario:** Como potenciar las ventas en los clientes objetivo agrupados en el clúster 0. Problema del usuario, expresado de manera concisa.

- **Generar ideas:** Imaginar y evaluar las ideas generadas a partir del dolor del cliente. Imágenes y descripción de la lluvia de ideas que se realice y la idea de solución que se selecciona. Sin mencionarle la idea, se debe entrevistar al cliente buscando todos los detalles que más se pueda. Evita hacer preguntas que reflejen un sesgo personal para obtener una respuesta genuina del cliente.

- **Prototipo:** prototipar la solución que resulta de las ideas y la descripción de iteraciones adicionales. Solicitar comentarios de los usuarios sobre el prototipo, los cuales ayudarán a iterar sobre el mismo y llegar a una solución sólida.

- **Conclusiones sobre el proceso de "Pensamiento de Diseño":** Mínimo 7 conclusiones. De acuerdo a las conclusiones se crea un plan de trabajo para el desarrollo y despliegue de la estrategia. Una estrategia comercial podrá ir acompañada de planes de comunicación (producción audiovisual, soluciones

web, piezas gráficas, entre otros), precios y un tiempo de permanencia de los planes de mercadeo.

10. Recomendaciones

Actualmente, las empresas presentan una mayor incertidumbre para la venta de sus productos, por lo que deben afrontar el reto de mejorar la experiencia de los clientes, sin embargo, cuentan con un gran volumen de datos e información importante, se recomienda el uso de técnicas de Machine Learning para tener mejores resultados en la toma de decisiones. Mediante la aplicación de la técnica de Machine Learning desarrollada en la investigación, y con el número y el valor de las compras por cliente, se logró crear una segmentación de clientes con la metodología de k-means, donde se obtuvo como resultado optimo la agrupación en 3 clusters.

En este trabajo, se propone un modelo de segmentación de clientes mediante el uso de técnicas de Machine Learning, de manera que se pueda generar un esquema para la estrategia comercial de ventas nacionales de la empresa New Stetic, para esto, se parte del análisis de datos donde se tiene en cuenta la cantidad de ventas históricas y se evalúa tomar como muestra los años comprendidos entre el 2018 y los dos primeros meses del año 2023.

Por tanto, se desarrolla el despliegue de la aplicación de Machine Learning en la empresa New Stetic, donde se consideraron principalmente las variables de cantidades y ventas totales para obtener como resultado tres grupos propuestos mediante un agrupamiento o segmentación de clientes.

De acuerdo con el gráfico de dispersión de la figura 15, grupo verde (Clúster 0) y según los resultados de la agrupación, es el segmento que más merece establecer un apalancamiento en la estrategia de ventas, por ser un grupo potencial para incrementar la participación en las ventas de la compañía. Donde las cantidades vendidas están entre las 100 y las 580 unidades.

Con el propósito de aumentar la participación de los productos en el mercado, mejorar la experiencia de los clientes e incrementar las ventas - facturación, la segmentación de clientes facilitará el desarrollo de una estrategia comercial enfocada en la fidelización de los clientes.

Adicional, este proceso será el punto de partida para el despliegue de planes comerciales a partir de datos confiables de fuentes internas.

Después de realizar un análisis del conjunto de datos de este trabajo y de acuerdo con el Core del negocio, se entiende que se pueden generar y tomar decisiones para las estrategias comerciales enfocadas en segmentos de clientes, se obtiene una lista final de clientes potenciales para fidelizar utilizando el método de agrupamiento de k-medias.

11. Referencias

Alpaydın, E. (2010). *Introduction to Machine Learning.* United States of America: Library of Congress Cataloging-in-Publication Information. Obtenido de https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf

Atencio Manyari Stefany Anyela, D. l. (2022). Propuesta de segmentación de clientes aplicando técnicas de Machine Learning para mejorar la experiencia de compra mediante un sistema de recomendación de productos de Tottus. 35-38. Obtenido de https://repositorio.esan.edu.pe/bitstream/handle/20.500.12640/3235/2022_IIC_22-2_03_TC.pdf?sequence=1&isAllowed=y

BBVA. (19 de septiembre de 2017). La importancia de la segmentación de mercado al desplegar una. Obtenido de https://www.bbva.es/finanzas-vistazo/ef/empresas/segmentacion-de-mercado.html

Colomer, J. V. (2023). *Universidad Politécnica de Cataluña.* Obtenido de https://upcommons.upc.edu/bitstream/handle/2117/16640/Viscarri_modelo_creacion_valor_cliente.pdf?sequence=1&isAllowed=y

Corrales, P. M. (2021). El impacto de la inteligencia artificial en el Derecho. *039*, 39-71. doi:https://doi.org/10.26439/advocatus2021.n39.5117

Dinngo . (2023). *Dinngo*. Obtenido de https://www.designthinking.es/inicio/

Gago, R. U. (2017). Uso de algoritmos de aprendizaje automático a base de datos genericos. 23-24. doi:http://hdl.handle.net/10609/65426

Guayara, Y. C. (2021). Propuesta De Consultoría Financiera Para Fortalecer La Toma De Decisiones Comerciales. (U. P. Colombia, Ed.) 25-26. Obtenido de http://repository.unipiloto.edu.co/handle/20.500.12277/10083

Javier Trujillano, A. S.-S. (2018).). Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardioAproximación a la metodología de clasificación y árboles de regresión. *Gaceta sanitaria*, 2-3. Obtenido de https://www.sciencedirect.com/science/article/pii/S0213911108712044

Müller, A. y. (2016). *Introducción al aprendizaje automático con Python: una guía para científicos de datos.* O'Reilly Media, Inc. Obtenido de https://books.google.es/books?hl=es&lr=&id=1-4lDQAAQBAJ&oi=fnd&pg=PP1&dq=M%C3%BCller,+AC+y+Guido,+S&ots=28oRHOKGVZ&sig=RSu9WX7bRsZj-eeyum_4LrpWLEA#v=onepage&q&f=false

Palacios, F. A. (2020). Segmentación de clientes de una empresa comercializadora de productos de consumo masivo en la ciudad de Popayán soportado en machine learning. 11-75. doi:http://unividafup.edu.co/repositorio/files/original/58784efa51bf4609763d30e2e6f70bea.pdf

Proaños, C. M. (2013). Aplicación de mineria de datos para la segmentación de clientes y desarrollo de estrategías de comunicación para la empresa DPC Studio S.A.S. *Universidad Piloto de Colombia*. doi:http://polux.unipiloto.edu.co:8080/00001084.pdf

Rincón, B. J. (2016). estudio del tanger objetivo de la empresa Madecentro Colombia S.A.S-Zona Santander. Obtenido de http://tangara.uis.edu.co/biblioweb/tesis/2016/164822.pdf

Roberth Chirinos, M. V. (mayo de 2017). BIG DATA para la segmentación de mercados en redes sociales en accesorios de moda emergente. 1-30. Obtenido de file:///C:/Users/vivia/Downloads/Dialnet-BigDataParaLaSegmentacionDeMercadosEnRedesSociales-7113491.pdf

Sandoval., L. J. (2018). Machine learning algorithms for data analysis and prediction. *11*, 2-5. Obtenido de http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

SAP AI. (2023). Obtenido de https://www.sap.com/latinamerica/insights/what-is-machine-learning.html#:~:text=El%20machine%20learning%20se%20compone,supervisado%2C%20semisupervisado%20o%20de%20refuerzo.

Silva, D. d. (2020). *Blog de Zendesk.* Obtenido de https://www.zendesk.com.mx/blog/estrategia-comercial/

Stefan Studer, T. B.-R. (2021). *A Machine Learning Process Model with Quality*. doi:https://doi.org/10.48550/arXiv.2003.05155

Tan, P. N. (2006). *Introduction to Data Mining.* Pearson Addison-Wesley. Obtenido de https://www.studywithus.net/sample/1712.pdf

Vallalta, J. F. (2023). *CRISP-DM: una metodología para minería de datos en salud*. Obtenido de Escuela de formación en inteligencia artificial en salud: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/