Pontificia Universidad Católica del Ecuador

Facultad de Hábitat, Infraestructura y Creatividad

Maestría en Sistemas de Información mención Data Science

Trabajo previo para la obtención del título de Magister en Sistemas de Información mención Data Science

Tema:

Caracterización de clientes mediante técnicas de clustering sobre patrones de consumo farmacéutico para estrategias de marketing.

Autor:

Carlos Junior Basurto Lara

Tutor:

Damián Aníbal Nicolalde Rodríguez Mg.

Quito - marzo 2026

Dedicatoria

A mi madre, cuya vida de esfuerzo sembró las oportunidades que han hecho posible esta meta.

A mis abuelos, raíz firme y silenciosa de todo lo que soy, cuya presencia sostuvo los cimientos de mi vida.

A Diego Peña y a Johana Mendoza, por creer en mí incluso en los momentos de incertidumbre y recordarme que la perseverancia encuentra su recompensa.

INDICE

1. CAPITULO 1: INTRODUCCIÓN

1.1 Introducción

El sector farmacéutico enfrenta en la actualidad desafíos comerciales sin precedentes derivados de la intensificación de la competencia global, la proliferación de canales de distribución y la presión constante por optimizar el retorno de las inversiones en marketing.

En este entorno la capacidad para identificar y priorizar clientes de acuerdo con su comportamiento de compra y valor económico se ha vuelto un criterio clave de competitividad y rentabilidad para las empresas farmacéuticas; sin embargo, muchas de ellas siguen utilizando criterios de segmentación tradicionales (por volumen de ventas, por tipos demográficos), y no aprovechan los datos analíticos contenidos en los masivos flujos de datos transaccionales generados al día. Esta falta de análisis de datos conlleva a diversos tipos de ineficiencias operativas directas que incluyen campañas de promociones con bajo retorno, la asignación incorrecta de recursos de la fuerza de ventas, la priorización incorrecta de clientes clave y la pérdida de oportunidades de negocios frente a la competencia.

Cuando combinamos modelos que miden el valor del cliente, como el clásico RFM (Recency, Frequency, Monetary) y su variante LRFM, que además considera el tiempo que el cliente lleva con la empresa, con técnicas de aprendizaje automático, obtenemos una forma muy sólida de transformar los datos de compras en segmentos realmente útiles para la operación. Estos modelos han demostrado ser muy efectivos para entender patrones de consumo y proyectar el valor futuro de los clientes en distintas industrias (Rahmadianti et al., 2020a) Y cuando los combinamos con algoritmos de clustering como K-means, Gaussian Mixture Models o Fuzzy C-Means, podemos agrupar a los clientes en perfiles con características internas muy parecidas, pero claramente diferenciados entre sí. Eso, a la hora de diseñar estrategias comerciales a medida para cada grupo, marca una gran diferencia (Zahrani Putri et al., 2023).

La evidencia empírica reciente respalda el impacto económico de estas metodologías. Investigaciones desarrolladas en el contexto del comercio minorista farmacéutico han documentado incrementos en la participación de ingresos de miembros de programas de fidelización desde 38.7% hasta 69.0% en períodos de tres años mediante la aplicación de segmentación basada en datos (Tan Jongprasert, 2023). De manera similar, proyectos de segmentación en distribución farmacéutica han reportado mejoras en tasas de crecimiento de ventas del 5% al 9% tras implementar acciones comerciales diferenciadas por segmento (Wu, Shi, Lin, Tsai, et al., 2020). En el ámbito de las relaciones entre empresas del sector farmacéutico, o lo que llamamos B2B, el uso combinado de modelos LRFM y técnicas de clustering ha resultado ser una estrategia muy práctica. Por ejemplo, ha permitido organizar bases de datos con más de mil farmacias o puntos de venta en grupos mucho más manejables y fáciles de priorizar. Esto se traduce directamente en una mejor asignación de las visitas comerciales y de los recursos destinados a promoción, justo donde más se necesita (Palupi & Fakhruzzaman, 2022).

Aunque el clustering ha sido exitoso en otros sectores, en el retail farmacéutico no puede aplicarse de forma automática. La regulación limita el uso de datos y las acciones promocionales, por lo que cualquier modelo de segmentación debe ajustarse a criterios legales y de manejo responsable de la información (Franjić, 2022). Adicionalmente, la heterogeneidad de canales de distribución que incluyen desde grandes cadenas de farmacias hasta puntos de venta independientes, pasando por distribuidores mayoristas y plataformas digitales demanda modelos capaces de integrar variables transaccionales, geográficas, firmográficas y de comportamiento multicanal (Kumar et al., 2024).

En términos metodológicos, la literatura más reciente indica que todavía no existe una posición común acerca de cuál es el algoritmo de agrupamiento más adecuado para manejar datos provenientes del sector farmacéutico. Por una parte, varios trabajos recalcan la velocidad y claridad de los grupos que K-means consigue (Wu, Shi, Lin, Tsai, et al., 2020); por otra parte, hay quienes aseguran que los modelos de mezcla gaussiana producen segmentaciones de mejor calidad si se mide con el coeficiente de silueta (Asmat et al., 2023). Incluso hay investigaciones que, al incorporar el tiempo entre compras al modelo RFM tradicional, han conseguido coeficientes de silueta de hasta 0.90, muy por encima de los 0.60–0.70 típicos (Katyayan et al., 2022a).

Diversos estudios coinciden en que no existe una técnica de segmentación que garantice resultados óptimos en todos los contextos. El desempeño de los algoritmos de clustering depende en gran medida de cómo estén organizados los datos y del entorno en el que se aplican(Abbasimehr et al., 2015). En este sentido, la elección de un método no debería basarse únicamente en su popularidad o en resultados obtenidos en otros sectores, sino en evaluaciones comparativas ajustadas a las características específicas del negocio.

Los trabajos que describen de forma completa cómo se implementan procesos de segmentación en el sector farmacéutico (Syaputra & others, 2020).En muchos casos, el análisis se concentra en indicadores técnicos o en la eficiencia del modelo, mientras que aspectos operativos como la integración con sistemas CRM, la definición práctica de los segmentos o la medición del impacto económico reciben menor atención (Y. Yoseph & Heikkilä, 2018a). Esta distancia entre el desarrollo metodológico y su aplicación ha limitado la adopción de estas herramientas, especialmente en empresas medianas.

Frente a este panorama, la presente investigación adopta técnicas de clustering, que incorpora un proceso que incluye la preparación de los datos, la comparación de modelos y la validación de los resultados, considerando además las particularidades regulatorias del sector (Abbasimehr et al., 2015). Esta situación cobra especial relevancia en el retail farmacéutico ecuatoriano, un sector caracterizado por la fragmentación del mercado y por un entorno regulatorio exigente que condiciona la forma en que se desarrollan las prácticas comerciales (García, 2022).

Por lo tanto, los propósitos de este estudio se enfocan hacia el análisis de los patrones de compra a través de variables LRFM generadas a partir de datos históricos, la comparación de algoritmos de agrupación en términos de estabilidad e interpretabilidad, la identificación de segmentos en función de su valor presente y potencial futuro, y la formulación de directrices

estratégicas acordes con cada perfil identificado (Syaputra et al., 2020; (Syaputra & others, 2020a; Y. Yoseph & Heikkilä, 2018a)

Por último, el documento se organiza en capítulos que recorren la revisión teórica, el diseño metodológico, la presentación y el análisis de los resultados, así como la discusión de sus implicaciones y limitaciones, con el fin de ofrecer una visión clara y actualizada de la segmentación de clientes farmacéuticos mediante técnicas de clustering.

1.2. Objetivos

Objetivo General:

Desarrollar un modelo de segmentación de clientes basado en técnicas de clustering y modelos de valor LRFM/CLV que optimice las estrategias de marketing en el sector farmacéutico ecuatoriano.

Objetivo Específicos:

- Identificar las variables transaccionales, comportamentales y contextuales que influyen en la segmentación de clientes del sector farmacéutico, incluyendo las categorías terapéuticas, patrones de compra cruzada y ubicación geográfica.
- Comparar distintos algoritmos de agrupamiento en relación con su capacidad para formar segmentos útiles, comprensibles y aplicables en decisiones comerciales.

1.3. Justificación del proyecto

Este estudio se basa en la necesidad de vincular lo que actualmente se debate en la teoría de datos con lo que realmente ocurre en el funcionamiento cotidiano de la industria farmacéutica. Para abordarlo, el trabajo se organiza alrededor de tres ejes: su utilidad práctica para el sector, su contribución metodológica a la ciencia de datos y su ayuda para llenar importantes vacíos en la literatura actual.

Desde el punto de vista práctico, la propuesta se basa en pruebas recientes que muestran que la utilización de modelos LRFM y clustering ha producido resultados tangibles: desde la reactivación de grupos de clientes inactivos hasta la modificación del momento en que se hacen las promociones o dar prioridad a los profesionales sanitarios y minoristas para visitas o programas formativos (Altuntas, 2023). Según estos casos, se puede traducir una investigación enfocada en la implementación en un uso más eficaz del presupuesto de marketing y en mejoras reales a nivel operativo (Celine & Kristiyanti, 2025a).

Por lo tanto, el trabajo tiene como objetivo ser útil para los expertos en la industria que requieren evidencia para respaldar sus decisiones de inversión en analítica avanzada y para los investigadores interesados en el desarrollo metodológico. Además, la investigación contribuye a la ciencia de datos aplicada al sistematizar y analizar de manera crítica técnicas de agrupamiento con regulaciones específicas proporcionando un marco referencial para investigaciones futuras en contextos parecidos (Fedorchenko et al., 2023).

Un punto central de esta investigación es precisamente responder a lo que todavía hace falta. Se requiere avanzar hacia marcos estandarizados de gobernanza de datos para el sector farmacéutico, así como generar estudios longitudinales que midan el impacto económico real de la segmentación bajo restricciones legales (Franjić, 2022; Tan Jongprasert, 2023).

1.4. Preguntas de Investigación

- ¿Cómo puede la aplicación de técnicas de clustering combinadas con modelos de valor del cliente (LRFM/CLV) mejorar la segmentación de clientes en el sector farmacéutico y optimizar las estrategias de marketing y asignación de recursos comerciales?
- ¿Qué variables de comportamiento comercial y LRFM contribuyen más a la creación de segmentos sólidos en clientes farmacéuticos?
- ¿Qué técnica de clustering produce segmentos más estables, interpretables y alineados con las necesidades del sector regulado?
- ¿Cómo varía el desempeño comercial (frecuencia de compra, ventas, CLV estimado) entre los diferentes segmentos identificados y qué estrategias pueden recomendarse para cada uno?

2. CAPITULO 2: DESARROLLO

2.1. Marco Teórico

2.1.1. Conceptos claves

Los conceptos desarrollados en esta sección constituyen el soporte teórico y analítico del estudio, y se organizan de manera progresiva desde los fundamentos de la segmentación y el análisis de datos, hasta su aplicación en el contexto específico del marketing farmacéutico. En conjunto, estos conceptos permiten comprender cómo los modelos de valor del cliente y las técnicas de clustering pueden integrarse en un marco metodológico orientado a la toma de decisiones comerciales, más allá de una descripción aislada de herramientas analíticas.

2.1.1.1. Segmentación de Clientes

La segmentación de clientes es el proceso sistemático y analítico mediante el cual una organización divide su base de clientes en grupos homogéneos que comparten características, comportamientos, necesidades o valores similares, con el propósito de diseñar e implementar estrategias de marketing diferenciadas y más efectivas (Barrera et al., 2023). Esta práctica permite a las empresas identificar la diversidad en su mercado, facilitando así la personalización de productos, servicios y comunicaciones.

Una segmentación debe satisfacer ciertos requisitos fundamentales para ser verdaderamente beneficiosa. No es suficiente con segmentar a los clientes; los segmentos tienen que ser medibles en lo que respecta a su tamaño y capacidad de compra, tener acceso por medio de canales viables, mostrar una relevancia económica adecuada, exhibir comportamientos diversos entre sí y posibilitar la realización de acciones específicas (Kotler & Armstrong, 2018a)

En la industria farmacéutica, donde las regulaciones limitan las promociones y los presupuestos de marketing a menudo son escasos, el empleo de modelos analíticos como LRFM o RFM presenta una ventaja evidente. Estos modelos hacen posible detectar con mayor exactitud a los clientes que contribuyen económicamente más y adecuar así las estrategias de manera más específica (Kotler & Armstrong, 2018b). A diferencia de los métodos tradicionales centrados únicamente en volumen de ventas o datos demográficos, el análisis basado en comportamiento transaccional proporciona una visión más estratégica del cliente (Syaputra & others, 2020b).

2.1.1.2. Clustering

El principio fundamental del clustering radica en maximizar la similitud intra-cluster o homogeneidad dentro de cada grupo y minimizar la heterogeneidad entre grupos diferentes.

Desde una perspectiva formal, el problema del clustering consiste en encontrar una función:

$$f: X \rightarrow \{1,2, \dots, K\}$$

Que asigna cada observación $x_i \in X$ a uno de los K clusters disponibles (Turkmen, 2022a)

Desde una perspectiva aplicada, el clustering se posiciona como la técnica central para llevar a la práctica la segmentación basada en datos, al permitir identificar grupos de clientes con comportamientos de consumo similares sin requerir categorías predefinidas. En este estudio, el clustering no se concibe únicamente como un ejercicio exploratorio, sino como una herramienta para generar segmentos estables, interpretables y directamente utilizables en decisiones de marketing farmacéutico, especialmente cuando se combina con variables de valor derivadas de modelos LRFM (Joga et al., 2023).

2.1.1.3. Aprendizaje Automático (Machine Learning)

El aprendizaje automático es una rama de la inteligencia artificial que permite a los algoritmos aprender de los datos y realizar predicciones o clasificaciones sin ser programados explícitamente para cada tarea (Chaudhary, 2023a; Wani et al., 2023a). Se divide en aprendizaje supervisado, no supervisado y por refuerzo. En la Tabla 1 se presenta una síntesis comparativa de estos enfoques, incluyendo sus características principales, métodos representativos y aplicaciones más comunes.

Tabla 1. Tipos de aprendizaje automático y sus características

Tipo de Aprendizaje	Descripción General	Métodos Representativos	Aplicación Principal	Referencias
Aprendizaje supervisado	El modelo aprende a partir de datos con etiquetas conocidas para predecir un resultado específico.	Regresión, clasificación	Predicción de valores numéricos y asignación de categorías	(Wani et al., 2023a)
Aprendizaje no supervisado	El modelo identifica patrones ocultos sin información previa sobre la variable objetivo.	Clustering, reducción de dimensionalidad	Agrupación de observaciones y simplificación de datos complejos	(Wani et al., 2023a)
Aprendizaje por refuerzo	El modelo aprende mediante interacción con un entorno, optimizando decisiones basadas en recompensas.	Q-learning, políticas óptimas	Optimización de secuencias de acciones e interacción dinámica	(Chaudhary, 2023a)

Para la segmentación de clientes, el aprendizaje no supervisado resulta fundamental para este estudio, dado que permite descubrir patrones latentes en los datos transaccionales farmacéuticos sin imponer supuestos previos sobre la estructura de los segmentos. Esta característica es especialmente relevante en mercados fragmentados y heterogéneos, donde la diversidad de comportamientos dificulta la aplicación de esquemas de segmentación rígidos.

2.1.1.4. Patrones y Dimensiones de Consumo

Los patrones de consumo farmacéutico son aquellas formas o comportamientos que se repiten cuando los clientes adquieren y usan medicamentos y/o productos relacionados, mismas que se pueden observar a través de distintas variables: clínicas, transaccionales, temporales y contextuales. Estas pautas reflejan tanto las necesidades terapéuticas de las personas como sus hábitos de consumo (F. Yoseph et al., 2020a). En la Tabla 2 se describen las principales dimensiones que se suelen considerar en este tipo de análisis, junto con las variables asociadas a cada una y por qué son relevantes cuando trabajamos con modelos LRFM o técnicas de clustering.

Tabla 2. Dimensiones analíticas de los patrones de consumo farmacéutico

Dimensión	Variables Asociadas	Relevancia para LRFM/Clustering
Transaccional	Recencia, Frecuencia, Monto, Ticket promedio, Diversidad de productos	Base del modelo RFM; mide valor monetario y nivel de interacción. Permite evaluar lealtad mediante compras repetidas y composición del ticket (Risky et al., 2024).
Temporal	Regularidad, Estacionalidad, Duración de la relación	Reconoce una estabilidad en el consumo y distingue entre clientes constantes e intermitentes. La segmentación toma en cuenta el factor tiempo (Abbasimehr et al., 2015).
Clínico-Terapéutica	Tipo de fármacos (agudos o crónicos), adherencia, politerapia	Extiende el análisis más allá de lo que es transaccional. Permite identificar a los pacientes crónicos de alta complejidad terapéutica y valor (Fedorovich, 2023)
Canal y Preferencia	Tipo de canal (digital o físico), sensibilidad con respecto al precio, inclinación hacia OTC o prescripción	Permite la segmentación por comportamiento multicanal y la adecuación a las restricciones regulatorias del sector (Kotler & Armstrong, 2018b)

El entendimiento claro estas dimensiones es fundamental para construir variables LRFM que realmente capturen no solo *cuánto* compran los clientes, sino también el *cómo*, el *cuándo* y el *qué*. De esta manera logramos enriquecer los datos de entrada del clustering y, con eso, los resultados que obtenemos.

2.1.1.5. Marketing Farmacéutico

El marketing farmacéutico engloba el conjunto de actividades estratégicas orientadas a satisfacer las necesidades de diferentes actores del sistema de salud (pacientes, prescriptores, farmacias y aseguradoras), en un entorno altamente regulado y éticamente exigente (Kevrekidis, Minarikova, Markos, & Souliotis, 2018).

El sector farmacéutico se caracteriza por una multiplicidad de actores en la decisión de compra, lo que complejiza las estrategias comerciales. A diferencia de lo que ocurre en otros mercados, aquí la decisión de compra no la toma una sola persona. Intervienen los médicos que recetan medicinas, farmacias que las dispensan, las aseguradoras que financian y, finalmente, los pacientes que consumen. Todo esto forma un ecosistema bastante más complejo e interdependiente (Kevrekidis, Minarikova, Markos, & Souliotis, 2018).

Por si fuera poco, la industria está sujeta a regulaciones muy estrictas que pueden llegar a controlar desde la publicidad y la promoción dirigida a profesionales de la salud, hasta las declaraciones sobre eficacia y las políticas de precios. Todo este círculo normativo busca proteger al consumidor y garantizar que los medicamentos se usen de forma adecuada (Barrera et al., 2023)

Todas estas particularidades hacen que la segmentación de clientes en el sector farmacéutico no pueda abordarse con una lógica puramente comercial, como sí se los hace con otros mercados.

2.1.1.6. Minería de Datos

La minería de datos se define como el proceso de descubrir patrones, relaciones, anomalías y conocimiento novedoso, válido y potencialmente útil a partir de grandes volúmenes de datos, mediante la aplicación sistemática de técnicas estadísticas, matemáticas y computacionales (Chaudhary, 2023a).

2.1.1.7. Proceso Estándar Intersectorial para la Minería de Datos

CRISP-DM es uno de los marcos más utilizados en minería de datos por su enfoque iterativo y su adaptación a sectores como retail y farmacia (Mariscal & others, 2024).

Su estructura de seis fases es:

- Entender el negocio establece objetivos y KPIs, como mejorar la segmentación LRFM (Mariscal & others, 2024).
- La comprensión de los datos examina la calidad y la estructura de la información que está a disposición (Husein et al., 2021)
- La preparación incorpora la depuración, unificación y creación de variables importantes (Ofori & others, 2025)
- El modelado pone a prueba algoritmos como DBSCAN y K-means (Bhardwaj & others, 2024).
- La evaluación combina análisis de utilidad comercial con métricas técnicas (Mariscal & others, 2024).
- El despliegue supone la puesta en marcha y el seguimiento del impacto sobre las ventas y la retención (Vasanthi & Moorthy, 2024).

2.1.1.8. Cadena de Suministro Farmacéutica

La segmentación de farmacias y distribuidores permite optimizar rutas, políticas de inventario y condiciones comerciales (Aslantaş, Rumelli, & Bakırlı, 2023).

2.1.1.9. Economía de la Salud

Los modelos de clustering aplicados a datos de prescripción y costos permiten clasificar poblaciones según riesgo y consumo, orientando decisiones actuariales y de política de salud (Aslantaş, Rumelli, & Bakırlı, 2023).

2.1.2. Síntesis del Marco Teórico

La revisión teórica desarrollada en este marco teórico evidencia que la segmentación de clientes basada en datos constituye un enfoque esencial para optimizar las estrategias de marketing en sectores regulados como el farmacéutico. Los modelos de valor del cliente, particularmente RFM y sus extensiones LRFM y CLV, proporcionan variables cuantitativas capaces de capturar el comportamiento transaccional, la duración de la relación y el valor económico esperado de los clientes. Al combinarlos con técnicas de clustering no supervisado, estos modelos nos permiten identificar grupos de clientes con características internas similares y, a su vez, diferenciados entre sí, que además son evaluables mediante métricas de validación que garantizan su calidad y estabilidad (Akande et al., 2024).

Asimismo, el análisis del contexto farmacéutico pone de manifiesto la necesidad de enfoques metodológicos que consideren simultáneamente las restricciones regulatorias, la heterogeneidad de los canales de distribución y la orientación a la toma de decisiones comerciales. Por ello, el marco teórico justifica la metodología adoptada en los siguientes capítulos, basada en la preparación rigurosa de los datos, así como la comparación empírica de algoritmos de *clustering* y la interpretación estratégica de los segmentos obtenidos, con el fin de generar resultados aplicables y replicables en el entorno estudiado. En la Figura 1 mostramos un esquema de este enfoque.

Figura 1. Articulación del enfoque de segmentación basado en LRFM y técnicas de clustering en el contexto del Retail farmacéutico.

En esta figura nos permite visualizar la lógica integradora del estudio, mostrando cómo los fundamentos analíticos, los enfoques de valor del cliente y las particularidades del sector farmacéutico se integran en un proceso de segmentación basado en datos. Esta articulación conceptual funciona como base para el diseño metodológico que se desarrolla en el siguiente capítulo, en el cual se detallan las etapas operativas, los algoritmos empleados y los criterios de validación utilizados.

2.2 Metodología

La presente investigación adopta un enfoque metodológico cuantitativo, basado en la minería de datos y el aprendizaje automático no supervisado, con el propósito de desarrollar un modelo de segmentación de clientes aplicable al contexto farmacéutico ecuatoriano. El diseño metodológico se alinea con el proceso estándar CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual garantiza un flujo de trabajo (Schröer et al., 2021). A continuación, se detallan las etapas centrales de este proceso.

2.2.1. Comprensión del Negocio

El objetivo fundamental de este estudio es desarrollar un modelo de segmentación de clientes basado en técnicas de clustering y modelos de valor LRFM que permita optimizar las estrategias de marketing en el sector farmacéutico ecuatoriano. Se busca superar las limitaciones de los esquemas tradicionales de segmentación (basados únicamente en variables demográficas o volumen de ventas) mediante un enfoque analítico que capture el valor real y el comportamiento de compra de los clientes (F. Yoseph et al., 2020b).

Para medir si el proyecto cumple con lo propuesto, se definen tres criterios de éxito:

1. Que los segmentos de clientes que obtengamos sean estadísticamente sólidos, validándolos con métricas internas de clustering.
2. Que los perfiles de cada segmento sean fáciles de interpretar y realmente útiles para tomar decisiones comerciales.
3. Derivar recomendaciones estratégicas concretas para cada grupo, de manera que ayuden a asignar mejor los recursos de marketing y a que las campañas promocionales sean más efectivas.

2.2.2. Comprensión de los Datos

Los datos utilizados en el presente estudio corresponden a transacciones reales de la empresa; en ninguna etapa del análisis se generaron ni emplearon datos sintéticos. La información proviene de un grupo empresarial dedicado a la comercialización de productos farmacéuticos y artículos de primera necesidad, que centraliza sus registros transaccionales dentro del canal farmacéutico ecuatoriano. La extracción se realizó directamente desde el Data Warehouse (DWH) y el sistema CRM de la organización, mediante consultas SQL ejecutadas en Apache Impala sobre una base de datos de más de 500 millones de registros transaccionales.

Para garantizar la representatividad y calidad del análisis, se definieron los siguientes criterios:

- **Período de análisis:** Se extrajeron todas las transacciones de clientes comprendidas entre el **1 de enero de 2023 y el 31 de diciembre de 2025** (fecha de corte).
- **Población objetivo:** Clientes del canal *retail* (venta directa al consumidor) con estado "activo" en el sistema a diciembre de 2025, excluyendo explícitamente los canales institucionales, mayorista y ventas entre empresas (Distribución).
- **Selección de la muestra:** Se aplicó un muestreo aleatorio estratificado sobre los clientes activos, seleccionando una muestra inicial de 20.000 clientes. Esta técnica asegura que la muestra final refleje la heterogeneidad de la base de clientes.
- **Criterios de inclusión:** Se incluyeron únicamente aquellos clientes con información demográfica completa (sexo y edad) y al menos una transacción válida dentro del período de estudio.
- **Depuración y exclusión:** Se excluyeron 330 clientes (equivalente al **1,65 % de la muestra inicial**) debido a:
    - Valores nulos en variables críticas para el cálculo de métricas LRFM.
    - Inconsistencias en los datos, como edades negativas (ej. -5 años) o superiores a 110 años, montos de venta no válidos (negativos o iguales a cero) y fechas de transacción fuera del rango de análisis.

Para garantizar la privacidad y el cumplimiento de la Ley Orgánica de Protección de Datos Personales (Asamblea Nacional del Ecuador, 2021), se implementó un riguroso proceso de anonimización. El identificador original cliente_id fue eliminado de los conjuntos de datos y reemplazado por un código único e irreversible (cliente_id_anon), generado mediante una función hash criptográfica (SHA-256). El mapeo entre el identificador original y el anonimizado se almacenó en un archivo independiente y protegido, fuera del entorno analítico, con el fin de evitar cualquier posibilidad de reidentificación.

En la Tabla 3 se presenta el diccionario de las variables finales empleadas en el estudio:

Tabla 3. Diccionario de variables del estudio

Variable	Descripción	Tipo	Rango/Valores	Unidad
cliente_id_anon	Identificador anonimizado del cliente	Categórica	-	-
sexo	Sexo del cliente	Categórica	FEMENINO, MASCULINO	-
edad	Edad del cliente al momento de la extracción	Numérica	8 - 106	Años
instruccion	Nivel de instrucción formal del cliente	Categórica	BACHILLER, BASICA, ELEMENTAL, ESPECIAL, INICIAL, NINGUNA, PRIMARIA,	-

			SECUNDARIA, SUPERIOR	
canton	Cantón de residencia del cliente	Categórica	203 cantones de Ecuador	-
unidad_negocio	Categoría del producto farmacéutico (agrupada)	Categórica	Belleza-Bienestar, Dermo-Cuid_Personal, Infantil-Minimarket, Medicinas-Otc, Vitamin-Suplem	-
fecha_transaccion	Fecha de la transacción de compra	Fecha	2023-01-01 a 2025-12-31	-
monto_total_usd	Monto total de la transacción en dólares	Numérica	> 0	USD
L (Length)	Días entre la primera y la última compra del cliente	Numérica	0 - 1095	Días
R (Recency)	Días desde la última compra hasta la fecha de corte (2025-12-31)	Numérica	0 - 939	Días
frecuencia	Número total de transacciones del cliente en el período	Numérica	1 - 1714	Conteo
monto_total	Suma del monto_total_usd por cliente en todo el período	Numérica	0.86 - 6552.65	USD
CLV (Customer Lifetime Value)	Valor de vida del cliente, calculado como el producto de frecuencia por monto_total. Es una métrica descriptiva *post-hoc* para evaluar el valor histórico de los segmentos, no se utilizó en el modelado.	Numérica	3.78 - 780,095.94	USD
[categoría]_pct	Porcentaje del gasto total del cliente en cada categoría de producto.	Numérica		

Las variables L, R, frecuencia y monto_total fueron transformadas mediante una función logarítmica (log1p) con el objetivo de reducir la asimetría en su distribución. Posteriormente, se aplicó a un escalado robusto (RobustScaler) para minimizar la influencia de valores

atípicos en el proceso de clustering. Las variables categóricas se mantuvieron en su forma original para el análisis descriptivo de los segmentos.

2.2.3. Preparación de los Datos

La preparación de datos incluyó limpieza, construcción de variables, transformación y anonimización.

En la limpieza, se corrigieron valores nulos, duplicados y errores (montos negativos, edades fuera de rango). Los registros sin información clave para calcular LRFM fueron excluidos. En variables demográficas secundarias se aplicó imputación simple o la categoría "No especificado".

Se construyeron las métricas LRFM a partir de los datos transaccionales por cliente:

- **L (Length):** días entre primera y última compra.
- **R (Recency):** días desde la última compra al 31/12/2025.
- **F (Frequency):** total de transacciones.
- **M (Monetary):** monto total facturado.

También se calcularon porcentajes de gasto por categoría para identificar patrones de consumo.

Dado el sesgo en L, R y F, se aplicó transformación logarítmica (log1p) y posteriormente RobustScaler para estandarizar las variables y reducir el impacto de valores atípicos (Hair et al., 2019).

Finalmente, toda la información fue anonimizada, eliminando identificadores directos y reemplazándolos por códigos únicos irreversibles para garantizar la privacidad.

2.2.4. Modelado

El núcleo analítico consistió en la aplicación y comparación de tres algoritmos de aprendizaje no supervisado sobre los datos transformados y proyectados en el espacio PCA.

- **K-means:** Seleccionado por su eficiencia computacional y amplia utilización en segmentación de clientes (Aslantaş, Rumelli, & Bakırlı, 2023). Se utilizó como modelo de línea base. La determinación del número óptimo de clusters (*k*) se realizó mediante el método del codo (elbow method) y el coeficiente de silueta, evaluando un rango de *k* = 2 a 15.
- **Clustering Jerárquico Aglomerativo:** se empleó para explorar estructuras naturales en los datos y también validar el número de clusters sugerido por K-means mediante el análisis de dendrogramas (Kaufman & Rousseeuw, 2009). Se utilizó el método de Ward y distancia euclidiana.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** se utilizó para identificar agrupaciones sin asumir formas esféricas y para detectar clientes atípicos dentro del conjunto de datos (Chaudhary, 2023b). Se realizó una

búsqueda sistemática de los parámetros eps y min_samples, utilizando el coeficiente de silueta como criterio para seleccionar la mejor configuración.

2.2.5. Evaluación

La calidad de los modelos de clustering se evaluó mediante un conjunto de métricas de validación interna que miden cohesión, separación y compacidad de los clusters (Turkmen, 2022b; Wani et al., 2023a),

- **Coeficiente de Silueta:** Este indicador mide qué tan similar es un objeto a su propio cluster en comparación con otros clusters. Cuando los valores se acercan a 1, significa que los clusters están bien definidos y claramente separados entre sí (Rousseeuw, 1987).
- **Índice Davies–Bouldin:** Evalúa la similitud promedio que existe entre cada cluster su vecino más cercano. En este caso, mientras más bajo es el valor, mejor separados están los grupos (Arbelaitz et al., 2023).
- **Índice Calinski–Harabasz:** También se conoce como índice de razón de varianza. Lo que hace es comparar la dispersión que hay entre los distintos clusters con la dispersión que existe dentro de cada uno. Si los valores son altos, quiere decir que los clusters están bien definidos y cohesionados (Liu et al., 2023).

Los resultados fueron los siguientes: K-means (silueta = 0.7719; Davies-Bouldin = 0.5957; Calinski-Harabasz = 14416.71), clustering jerárquico (silueta = 0.7249; Davies-Bouldin = 0.7018; Calinski-Harabasz = 13505.92) y DBSCAN (silueta = 0.8393; Davies-Bouldin = 0.5035; Calinski-Harabasz = 3912.18).

DBSCAN alcanzó los resultados más altos en la silueta y el índice de Davies-Bouldin. A pesar de que su índice Calinski-Harabasz fue inferior, esto se debe a la existencia de 78 observaciones clasificadas como ruido, las cuales no contribuyen a la varianza entre los grupos. Por esta razón, se eligió modelo final ya que identifica outliers y conservar una cohesión interna elevada.

Estos indicadores posibilitaron calcular el número apropiado de grupos en K-means, analizar la estabilidad y realizar una comparación objetiva del rendimiento de los tres algoritmos (Rahmadianti et al., 2020a).

DBSCAN logró los resultados más elevados en el índice de Davies-Bouldin y en la silueta es por ello que se eligió el modelo más adecuado en función de un análisis combinado de las tres métricas, poniendo como prioridad aquel algoritmo que lograra el equilibrio óptimo entre la separación inter-cluster y la cohesión intra-cluster, y que además produjera segmentos útiles e interpretables para la toma de decisiones comerciales.

2.2.6. Despliegue

Los resultados del modelado se traducen en recomendaciones estratégicas de marketing específicas para cada segmento de clientes identificado. Estas recomendaciones, detalladas en el Capítulo 4, incluyen: diseño de campañas de reactivación para clientes esporádicos, programas de fidelización para clientes de alto valor, y estrategias de bajo costo para clientes

inactivos. La implementación de estas estrategias permitirá al sector farmacéutico optimizar la asignación de recursos comerciales y mejorar el retorno de sus inversiones en marketing.

2.2.7. Herramientas de Software

Todo el proceso de análisis, desde la extracción y preparación de los datos hasta el modelado y la visualización de resultados, se implementó utilizando el lenguaje de programación Python (versión 3.9+) y su ecosistema de librerías científicas y de ciencia de datos. Las principales herramientas empleadas fueron:

- **Manipulación y análisis de datos:** Pandas, NumPy.
- **Preprocesamiento y machine learning:** Scikit-learn (Pedregosa et al., 2011), para la estandarización (RobustScaler), reducción de dimensionalidad (PCA), implementación de los algoritmos de clustering (K-Means, clustering jerárquico, DBSCAN) y cálculo de las métricas de validación (silhouette_score, davies_bouldin_score, calinski_harabasz_score).
- **Visualización:** Para las visualizaciones usamos Matplotlib y Seaborn. Con ellas generamos tanto los gráficos exploratorios, como histogramas y matrices de correlación, como los resultados finales, por ejemplo, los dendrogramas y la proyección de los clusters en el espacio PCA.
- **Entorno de desarrollo:** El análisis exploratorio lo hizo en Jupyter Notebooks, y para el resto del proceso usamos scripts modulares. Así buscamos que todo el estudio sea más fácil de replicar.

Con el fin de asegurar la reproducibilidad del estudio, el código fuente empleado en las etapas de análisis exploratorio, preprocesamiento, modelado y visualización se ha documentado y organizado en scripts modulares de Python (versión 3.9.13). Las principales librerías empleadas y sus versiones específicas fueron: pandas (1.5.3), numpy (1.24.3), scikit-learn (1.2.2), matplotlib (3.7.1) y seaborn (0.12.2). El proceso de extracción de datos se realizó mediante consultas SQL sobre Apache Impala.

Se fijó una semilla aleatoria (random_state = 42) en todos los algoritmos que lo permiten (como K-means y en la selección de la muestra para el dendrograma) para asegurar la estabilidad y reproducibilidad de los resultados. El flujo de ejecución seguido es:

- 01_extraccion_muestra.sql (sobre Impala) →
- 02_preparacion_datos.py →
- 03_analisis_exploratorio.py →
- 04_modelado_clustering.py →
- 05_evaluacion_resultados.py.

2.3. Revisión de Teorías y Modelos Existentes

En la segmentación de mercados no todos los clientes son iguales, y agruparlos según comportamientos y necesidades permite diseñar estrategias más acertadas. Diversos autores han señalado que los consumidores reaccionan de manera distinta ante las propuestas de

valor, lo que justifica dividirlos en segmentos relativamente homogéneos (Aslantaş, Rumelli, & Bakrl, 2023; Güney, 2019; Vayena & Blasimme, 2023). Desde los planteamientos iniciales de Wendell Smith hasta los criterios establecidos por Kotler mensurabilidad, accesibilidad, sustancialidad, diferenciabilidad y accionabilidad se ha afianzado un marco que ayuda a evaluar si una segmentación realmente es útil en la práctica.

En el sector farmacéutico, este enfoque ha avanzado hacia la aplicación de métodos de agrupamiento que posibilitan reconocer patrones distintos en las actitudes y preferencias, así como en la conducta de compra. Esto permite tomar decisiones comerciales más enfocadas y gestionar mejor los programas de fidelización (Kevrekidis, Minarikova, Markos, Malovecká, et al., 2018; Momahhed et al., 2023). Analizar dimensiones esenciales como la recencia, la frecuencia, el gasto y la duración de la relación se puede lograr al combinar algoritmos como K-means, métodos jerárquicos o DBSCAN con modelos de valor del cliente RFM y sus variaciones, tales como LRFM o LRFMS. Esto proporciona una visión más integral del cliente.

La colocación del Customer Lifetime Value (CLV) posibilita la clasificación de los segmentos no solo en función de su comportamiento anterior, sino también de su contribución financiera esperada. Esto guía las decisiones en función de criterios de rentabilidad y retención. Por otro lado, la metodología CRISP-DM proporciona estructura y coherencia en un ambiente regulado como el farmacéutico al organizar el proceso analítico desde que se define el problema hasta que se pone en marcha el modelo.

2.3.1 Modelos RFM

El modelo RFM (Recency, Frequency, Monetary) es uno de los marcos conceptuales más utilizados para analizar el comportamiento transaccional de los clientes (Akande et al., 2024; Aslantaş, Rumelli, & Bakrl, 2023). Su aplicación se basa en tres dimensiones:

Recencia (R): tiempo desde la última transacción, bajo la premisa de que clientes que compraron recientemente tienen mayor probabilidad de volver a comprar.

Frecuencia (F): Cuantifica el número de transacciones realizadas por el cliente en un período determinado, reflejando el nivel de engagement y lealtad hacia la empresa.

Valor Monetario (M): Representa el monto total de las transacciones, indicando el valor económico que el cliente aporta a la organización.

En el sector farmacéutico, el modelo RFM se utiliza para clasificar clientes, estimar el Customer Lifetime Value (CLV) y guiar las estrategias de marketing entre distribuidores y farmacias (Nikaein & Abedin, 2021; Rahmadianti et al., 2020a).

Las variables que se derivan del RFM se usan como punto de partida en algoritmos de clustering no supervisado para construir segmentos prácticos y poder priorizar clientes según el valor que representan o podrían llegar a tener (Aslantaş, Rumelli, & Bakrl, 2023).

2.3.2 Extensiones del modelo RFM

La evolución del modelo RFM ha generado extensiones más sofisticadas:

- *LRFM (Length-RFM):* incorpora la duración de la relación cliente-empresa como cuarta dimensión, reconociendo la antigüedad como predictor relevante de valor (Rahmadianti et al., 2020).
- *RFML:* incorpora métricas de lealtad, medida a través de indicadores como la concentración de compras o la resistencia a ofertas de la competencia (Nikaein & Abedin, 2021).
- *LRFMS:* integra análisis de series temporales multivariadas para capturar la evolución dinámica del comportamiento del cliente (Wang et al., 2024).

La Tabla 4 recoge una comparación de estos modelos, detallando sus dimensiones, el tipo de datos requeridos, cómo se aplican en el sector farmacéutico, y también sus principales ventajas y limitaciones.

Tabla 4. Comparación de extensiones del modelo RFM

Modelo	Dimensiones	Tipo de datos	Uso en pharma	Ventajas	Limitaciones
RFM	Recencia, Frecuencia, Valor monetario.	Datos de compras: fechas y montos	Segmentación básica de clientes según sus compras pasadas	Fácil de usar, conocido y replicable	No muestra cambios en el tiempo ni el contexto del cliente
LRFM	RFM + Duración de la relación	Tiempo que el cliente ha estado activo	Identificar clientes que se quedan más tiempo	Ayuda a ver la evolución del cliente y prever retención	Requiere historial largo y bien registrado
RFML	RFM + Lealtad (fidelidad)	Participación en compras, frecuencia relativa	Saber qué tan fiel es un cliente en la distribución de medicamentos	Útil para diseñar programas de fidelidad o beneficios	Puede ser difícil conseguir datos del mercado o competencia
LRFMS	LRFM + Datos en el tiempo (series)	Registros mensuales o por períodos	Detectar cambios en el comportamiento del cliente	Segmentación más actualizada y adaptable	Es más complejo y necesita herramientas avanzadas

2.3.3. Customer Lifetime Value (CLV)

El CLV representa el valor presente neto de los flujos futuros atribuibles a un cliente durante toda la relación comercial (Y. Yoseph & Heikkilä, 2018b). La formulación clásica del CLV se expresa como:

$$CLV = \sum_{t=1}^{T} \frac{(R_t - C_t)}{(1+d)^t}$$

Donde $R_t$son los ingresos generados en el período t, $C_t$los costos asociados, $d$la tasa de descuento y Tel horizonte temporal.

En la industria farmacéutica, las relaciones comerciales con clientes tienden a ser duraderas, lo que convierte el CLV en un indicador estratégico esencial para asignar recursos, diseñar programas de retención y orientar políticas de pricing (Nikaein & Abedin, 2021).

El CLV permite ordenar clusters en función de su valor esperado, generando tiers de clientes que guían estrategias específicas de CRM y marketing (Rahmadianti et al., 2020a).

2.3.4. Modelos de Difusión de Innovación

La teoría de difusión de Rogers describe las categorías de adopción de nuevas tecnologías: innovadores, adoptantes tempranos, mayoría temprana, mayoría tardía y rezagados. En pharma, permite identificar prescriptores y farmacias con distinta predisposición a adoptar nuevos medicamentos (Momahhed et al., 2023).El clustering de datos temporales revela segmentos alineados con estos perfiles, informando campañas de lanzamiento y detailing.

2.4. Algoritmos Principales para Clustering

2.4.1.K-means

K-means es el método más extendido debido a su simplicidad y eficiencia (Aslantaş, Rumelli, & Bakrl, 2023; Celine & Kristiyanti, 2025b).Minimiza la suma de distancias cuadradas dentro de cada cluster (WCSS):

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} || \, x - \mu_i \, ||^2$$

Donde el WCSS mide qué tan compactos son los clusters; K-means trata de que esa suma sea lo más pequeña posible.

- ***Ventajas:*** escalabilidad, eficiencia y claridad interpretativa de centroides.
- **Limitaciones:** necesidad de definir K, suposición de clusters esféricos, sensibilidad a outliers (Wu, Shi, Lin, & Tsai, 2020).

2.4.2. Clustering Jerárquico

Produce dendrogramas que revelan estructuras naturales en los datos y no requiere especificar K previamente.

- *Ventajas:* No requiere especificación previa del número de clusters; produce dendrogramas para exploración visual; permite identificar jerarquías naturales.
- *Limitaciones*: alta complejidad computacional para métodos básicos; irreversibilidad de las decisiones de fusión; sensibilidad al método de enlace seleccionado (Turkmen, 2022a).

2.4.3. DBSCAN

DBSCAN (Agrupamiento espacial de aplicaciones con ruido basado en la densidad) representa un enfoque basado en densidad que identifica clusters como regiones continuas

de alta densidad separadas por regiones de baja densidad (Katyayan et al., 2022b; Wani et al., 2023b).

- *Ventajas*: descubre clusters de forma arbitraria; identificación automática de outliers; no requiere especificación previa del número de clusters.
- *Limitaciones:* requiere ajuste cuidadoso de parámetros ε y MinPts; dificultad con clusters de densidades variables; (c) sensibilidad en espacios de alta dimensionalidad**.**

2.5. Métricas de Validación de Clustering

La calidad del clustering se evalúa mediante índices internos que miden cohesión, separación y compacidad (Turkmen, 2022a; Wani et al., 2023b).

- *Coeficiente de Silueta:* valores cercanos a 1 indican separación óptima.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

- *Índice Davies-Bouldin*: cuanto más bajo es el valor, mejor separados están los grupos entre sí.
- *Índice Calinski-Harabasz*: valores altos reflejan que los clusters están más cohesionados y definidos.

Estos índices permiten seleccionar el número óptimo de clusters y evaluar la estabilidad de los grupos generados (Rahmadianti et al., 2020a).

2.6. Técnicas de agrupamiento en la segmentación de clientes

- ***K-significa clustering*:** Esta técnica se usa ampliamente debido a su simplicidad y eficiencia para minimizar las distancias dentro de los clústeres, lo que permite formar distintos grupos de clientes para el análisis. Es particularmente eficaz cuando se combina con técnicas de preprocesamiento, como el escalado y la reducción de dimensionalidad, para mejorar los resultados de la agrupación (Panda et al., 2024).
- ***Agrupación jerárquica aglomerativa*:** este método crea una jerarquía de clústeres y proporciona vistas globales y locales de los segmentos de clientes. Es beneficioso para entender las relaciones entre los diferentes grupos de clientes (Panda et al., 2024).
- ***DBSCAN:*** este algoritmo identifica los clústeres en función de la densidad de los datos, lo que lo hace adecuado para gestionar formas arbitrarias y ruidos en los datos de los clientes (Abednego et al., 2023; Panda et al., 2024).
- ***Modelos de mezcla gaussianos*:** estos modelos capturan la distribución de datos con formas de clústeres flexibles, lo que permite una comprensión más matizada de los segmentos de clientes (Panda et al., 2024).

2.6.1. Aplicación en la industria farmacéutica

- ***Análisis de RFM*:** El modelo de actualidad, frecuencia y moneda (RFM) se suele utilizar junto con técnicas de agrupamiento para segmentar a los clientes en función de su historial de transacciones. Este modelo ayuda a cuantificar el comportamiento de los

clientes y a clasificarlos en distintos segmentos, lo cual es crucial para desarrollar estrategias de marketing específicas en el sector farmacéutico (Kasem et al., 2023).

- ***Segmentación basada en patrones de comportamiento:*** este enfoque se centra en identificar agrupaciones naturales en los datos de las clientes caracterizadas por patrones de comportamiento típicos. Es particularmente útil en la industria farmacéutica, donde comprender el comportamiento de los clientes puede conducir a estrategias de marketing más efectivas (Yang, 2009).

2.7. Beneficios de la segmentación de clientes

Cuando las empresas entienden bien las características y preferencias particulares de cada segmento de clientes, pueden ajustar sus estrategias de marketing para responder a necesidades concretas, y eso se nota en campañas más efectivas y en una mejor conexión con los clientes (PILLI SRI DURGA et al., 2023; Shchekoldin et al., 2023).

La segmentación también ayuda a usar los recursos de manera más inteligente: por ejemplo, permite concentrarse en los segmentos más rentables y reducir el gasto en marketing allí donde no es necesario (PILLI SRI DURGA et al., 2023; Shchekoldin et al., 2023).

Y hay otro punto importante: cuando las estrategias son personalizadas, los clientes se sienten más valorados, lo que fortalece la relación con la empresa y mejora la lealtad y la retención (Shchekoldin et al., 2023).

Dicho esto, aunque las técnicas de agrupamiento ofrecen ventajas importantes para segmentar clientes, también hay que tener en cuenta sus limitaciones y los desafíos que traen. Por ejemplo, los resultados pueden variar mucho según el algoritmo que se elija y la calidad de los datos con los que se trabaje. En el caso específico de la industria farmacéutica, además, hay retos adicionales como las restricciones regulatorias y la necesidad de contar con una demanda más o menos estable, aspectos que no se pueden ignorar al diseñar estrategias de segmentación (Greengrove, 2002; Shchekoldin et al., 2023). Por eso, lo más sensato es apostar por un enfoque integral, que combine el clustering con otras herramientas analíticas como el RFM o la segmentación basada en patrones de comportamiento, para construir un marco sólido que realmente ayude a optimizar las estrategias de marketing en el sector.

3. CAPITULO 3: DESARROLLO

3.1. Análisis Descriptivo de la Muestra

El análisis exploratorio de los datos constituye una etapa fundamental para comprender la estructura general del conjunto de observaciones y evaluar el comportamiento de las variables relevantes antes de aplicar técnicas de segmentación. La distribución de las variables asociadas al modelo LRFM, es decir, longitud de la relación, recencia, frecuencia y valor monetario, junto con la edad de los clientes, se presenta en la Figura 2.

Al analizar la variable de antigüedad, se observa una inclinación hacia los valores más altos del periodo estudiado, ya que muchos registros se ubican en el límite superior, lo cual indica que una parte importante de los clientes mantiene relaciones comerciales prolongadas y estables. En cuanto a la recencia, los datos se concentran en rangos bajos, lo que indica actividad reciente dentro de la cartera. La frecuencia de compra y el gasto total muestran distribuciones que alcanzan valores elevados, evidenciando un grupo reducido de clientes cuyo consumo supera el promedio. Por último, la edad indica una distribución más equilibrada, con mayor concentración entre los cuarenta y setenta años, este rango coincide con el perfil usual de consumidores de productos farmacéuticos.

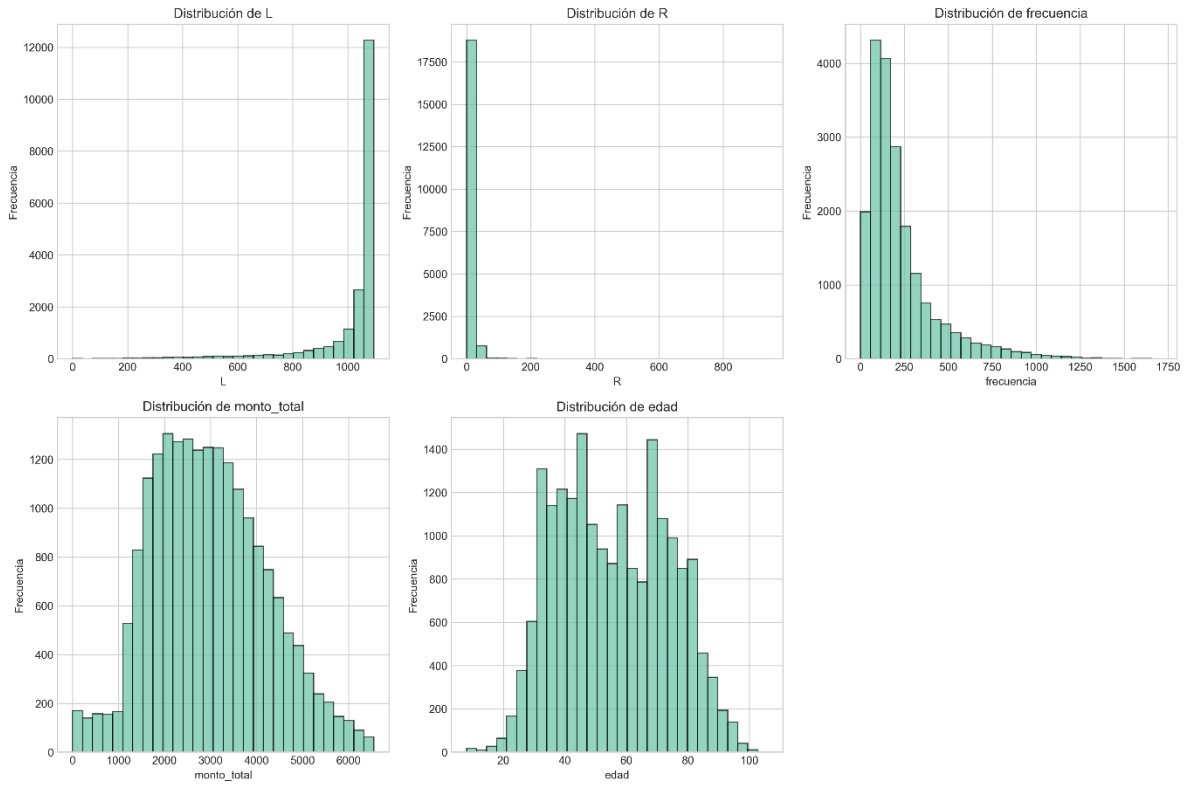

Figura 2. Distribución de variables LRFM y edad

La composición de la muestra según sexo se ilustra en la Figura 3.

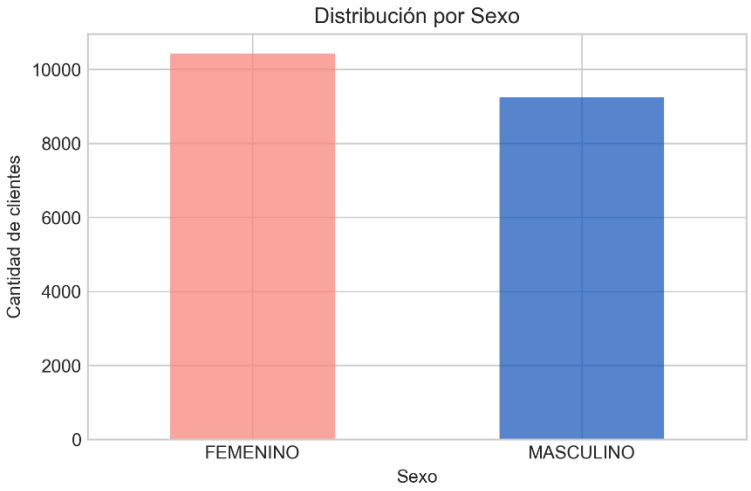

Figura 3. Distribución de clientes por sexo

Se observa una presencia ligeramente mayor de clientes mujeres, que representan un poco más de la mitad de la muestra. Aunque la diferencia con los hombres no es muy amplia, esta tendencia podría estar reflejando el rol más activo que podrían tener las mujeres en las decisiones de compra relacionadas con la salud y el bienestar, algo que se ha visto reflejado en investigaciones en el ámbito sanitario ya han documentado.

La distribución del nivel de instrucción de los clientes se presenta en la Figura 4.

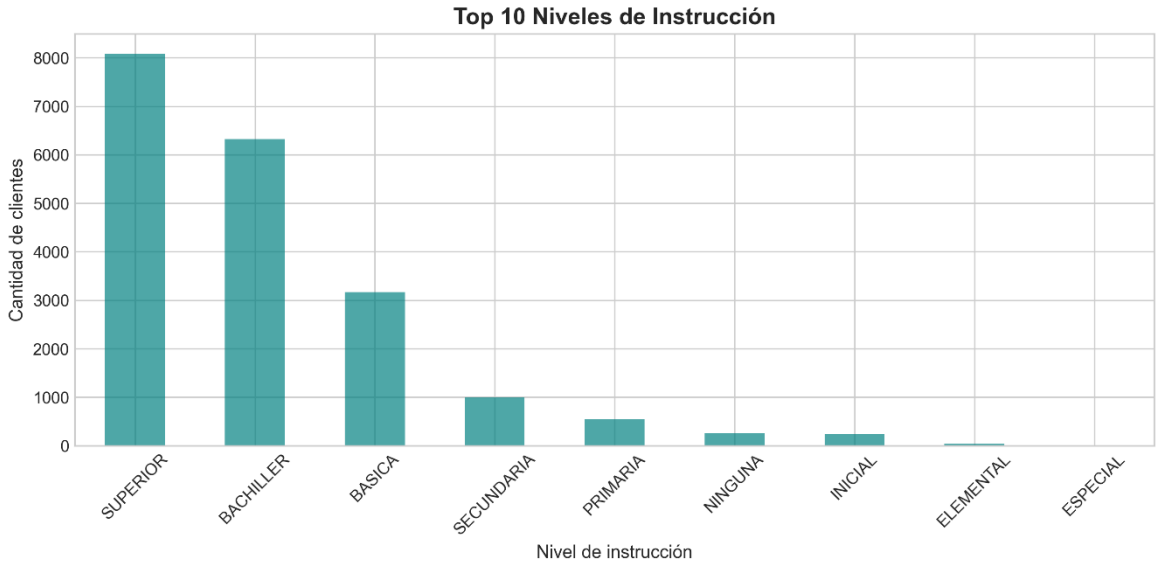

Figura 4. Distribución del nivel educativo de los clientes

Los resultados indican que la mayoría de los clientes se ubica en los niveles educativos de bachillerato y superior con un porcentaje mayor a los 2/3 de la muestra. En cambio, los niveles de escolaridad más bajos tienen una participación mucho menor. Por lo tanto, se puede decir que la población analizada forma un nivel educativo alto.

El análisis espacial de la muestra se presenta en la Figura 5, donde se ilustra la distribución de clientes según cantón de residencia.

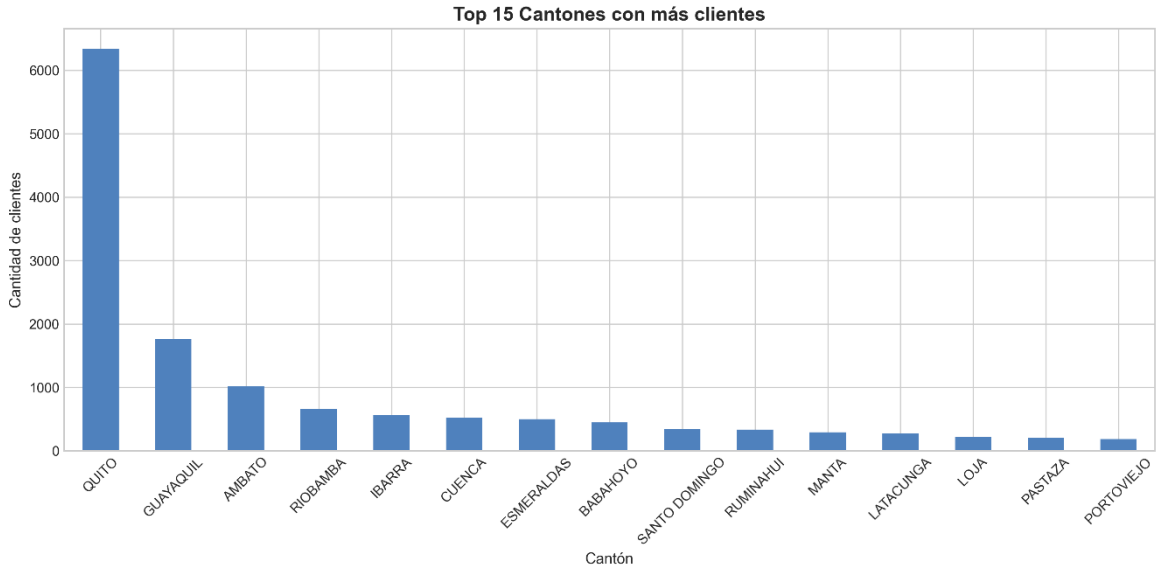

Figura 5. Distribución geográfica de clientes por cantón

Con base al análisis geográfico la mayor proporción de clientes se concentra en los cantones de Quito, el cual es el cantón con mayor representación dentro de la muestra, seguido por Guayaquil, Ambato y Riobamba, mientras que el resto de los cantones presenta una participación menor. Esta distribución va de la mano con la densidad poblacional y el movimiento económico que caracteriza a estas ciudades lo que indica que la cartera de clientes se encuentra, en su mayoría, vinculada a entornos urbanos con alta actividad comercial.

3.1.1. Perfil Demográfico

La muestra está integrada por 19 670 clientes del canal retail en Ecuador. Tal como se indicó anteriormente con las siguientes características:

- Ligera mayoría de clientes del sexo femenino.
- Predominio de clientes con nivel educativo de bachillerato y educación superior.
- Mayor concentración en cantones con alta densidad poblacional y actividad económica.

3.1.2. Comportamiento de Compra (LRFM)

Los estadísticos descriptivos presentados en la Tabla 5 permiten caracterizar de manera integral la estructura del comportamiento transaccional de los clientes durante el periodo de estudio, evidenciando patrones relevantes tanto en tendencia central como en dispersión.

Tabla 5. Estadísticas descriptivas de las variables LRFM

Variable	count	mean	std	Min	25%	50%	75%	max
L	19670	1015.26	146.65	0	1026	1072	1086	1095
R	19670	8.79	22.15	0	1	4	9	939
frecuencia	19670	225.82	200.73	1	98	164	275	1714
monto_total	19670	2967.61	1241.03	0.855	2036.232	2882.281	3796.374	6552.647
edad	19670	55.68	17.98	8	40	54	71	106
Belleza-Bienestar_pct	19670	0.04	0.04	0	0.0091	0.0245	0.0518	1
Dermo-Cuid_Personal_pct	19670	0.09	0.09	-7.91E-06	0.0229	0.0567	0.1138	1
Infantil-Minimarket_pct	19670	0.11	0.18	0	0.0071	0.0263	0.1153	1
Medicinas-Otc_pct	19670	0.69	0.22	0	0.5592	0.7503	0.8740	1
Vitamin-Suplem_pct	19670	0.08	0.08	-5.36E-06	0.0242	0.0536	0.1009	0.9222

La tabla reúne información de 19,670 registros y permite observar con claridad el comportamiento general de las variables del modelo. La variable L tiene un promedio de 1015.26 y una mediana de 1072. La cercanía entre estos valores es por tener una distribución relativamente homogénea, con ciertos casos atípicos. En el caso de R, tiene un promedio de 8.79 lo cual es considerablemente mayor que la mediana de 4. Esta diferencia indica que algunos valores elevados influyen en el cálculo del promedio.

En la frecuencia se puede observar una diferencia en los valores de la media y mediana junto con una desviación estándar alta esto se debe a que no todos los clientes compran con la misma intensidad. Existen usuarios con niveles de compra muy superiores al promedio, lo que incrementa la variabilidad.

El monto total mantiene una cierta estabilidad en el gasto típico. Sin embargo, el valor máximo de 6552.65 y la desviación estándar superior a 1200 muestran que hay clientes con un poder de compra significativamente mayor al resto.

En cuanto a la edad existe un rango amplio, aunque parece concentrarse en adultos de mediana edad.

Respecto a las categorías de consumo expresadas en proporciones, se observa que Medicinas OTC concentra la mayor participación indicando que el gasto se orienta hacia esta categoría. Sin embargo, Belleza-Bienestar, Dermocuidado Personal, Infantil-Minimarket y Vitaminas-Suplementos muestran participaciones medias menores indicando un gasto más fragmentado entre estas líneas.

3.1.3. Patrones de Consumo por Categoría

La distribución del gasto promedio por categoría de producto se presenta en la Figura 6, donde se observa la participación relativa de cada grupo dentro del total de compras realizadas por los clientes.

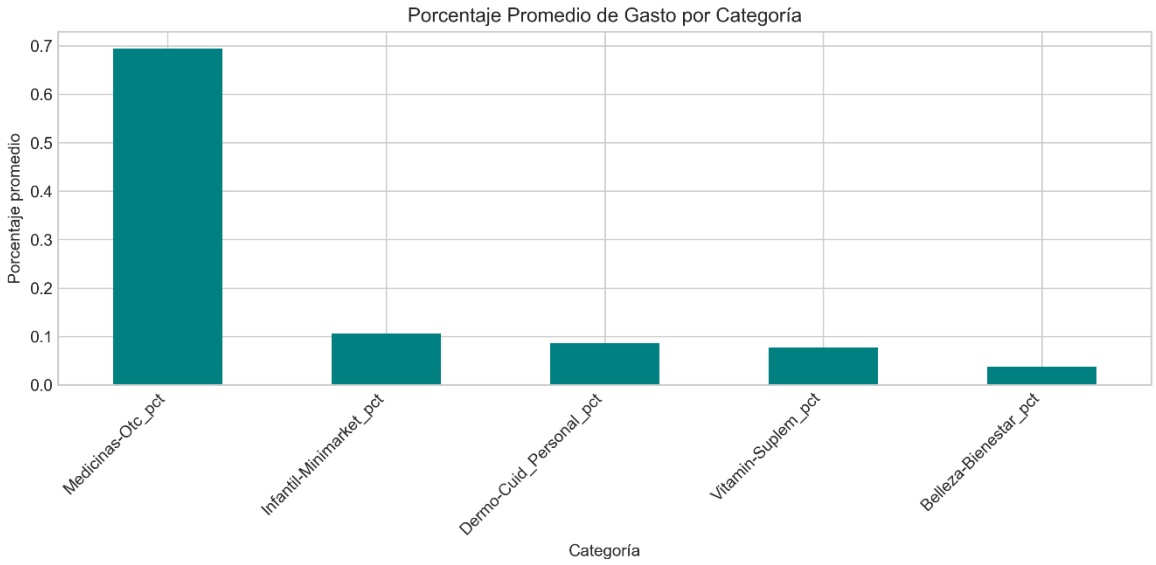

Figura 6. Porcentaje promedio de gasto por categoría

Los resultados muestran que el gasto se concentra principalmente en la categoría Medicinas-OTC, la cual absorbe cerca del 69 % del promedio total. Esta proporción trata del componente central del consumo, lo que es coherente con la naturaleza del negocio farmacéutico, donde los productos terapéuticos suelen liderar la demanda.

En un segundo nivel aparecen Infantil-Minimarket y Dermo-Cuidado Personal con 11 % y 9 %, respectivamente. Estos porcentajes forman muestran un parte habitual de la compra y actúan como categorías complementarias dentro del ticket promedio.

Por su parte, Vitaminas-Suplementos y Belleza-Bienestar presentan una participación más reducida, si bien su contribución es menor no dejan de representar nichos relevantes, asociados a necesidades específicas o a decisiones de compra más esporádicas.

3.1.4. Relación entre Variables

La relación estadística entre las variables analizadas se presenta en la Figura 7, donde se muestra la matriz de correlación que resume el grado de asociación lineal entre las dimensiones del modelo LRFM y las categorías de producto.

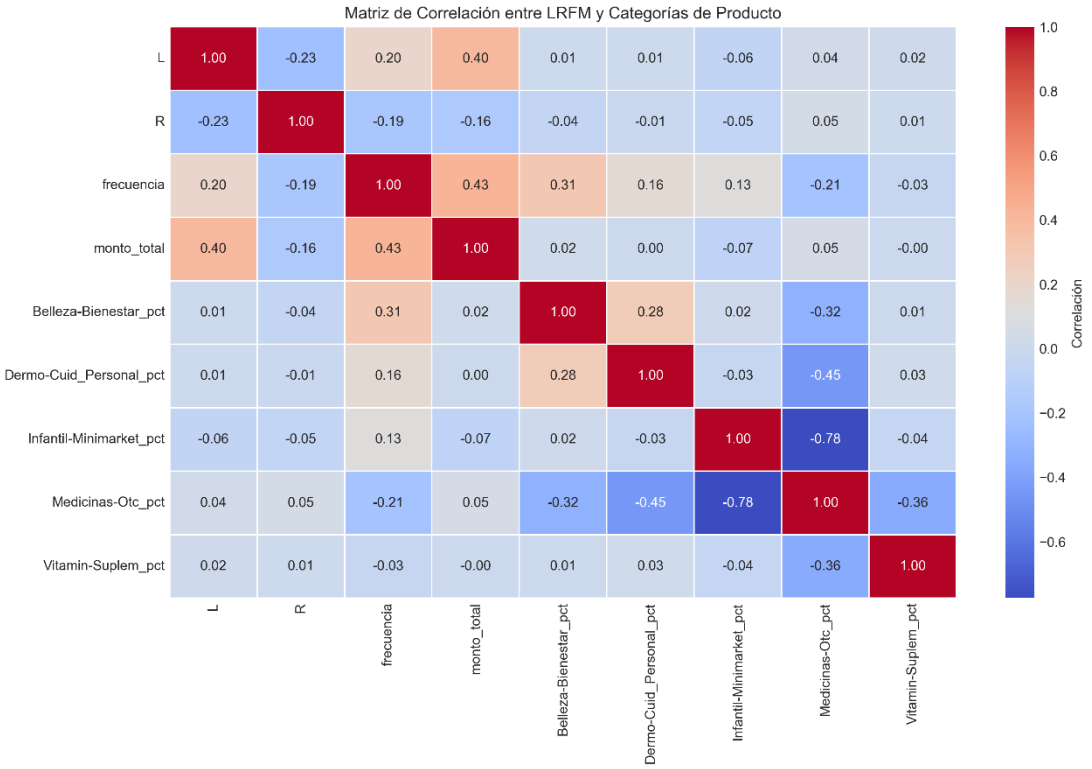

Figura 7. Matriz de correlación entre variables LRFM y categorías

En primer lugar, se observa una asociación positiva de magnitud moderada entre frecuencia y monto_total con un r = 0.43, a medida que aumenta el número de compras, también tiende a incrementarse el gasto acumulado. También, L mantiene una relación positiva con el monto total (r = 0.40) y, en menor medida, con la frecuencia (r = 0.20). Estos valores se deben a que los clientes con mayor permanencia muestran, en promedio, un comportamiento de compra más activo y un mayor nivel de desembolso.

Por el contrario, R presenta correlaciones negativas con frecuencia (r = −0.19), con monto_total (r = −0.16) y con L (r = −0.23). Esto resulta coherente puesto que cuanto más tiempo ha pasado desde la última compra, menor es la actividad del cliente y, en consecuencia, su contribución al gasto total.

Con una fuerte correlación negativa se muestran las categorías Medicinas-OTC e Infantil-Minimarket (r = −0.78) debido a que la proporción de gastos en medicamentos disminuye de manera considerable la participación relativa en productos infantiles y de minimarket. Por su parte, Medicinas-OTC también mantiene asociaciones negativas con Dermo-Cuidado Personal con un r = −0.45 y Belleza-Bienestar con un r = −0.32, esto indica que el gasto tiende a concentrarse en una categoría principal.

Por otra parte, Belleza-Bienestar tiene una correlación positiva con la frecuencia (r = 0.31) ya que este tipo de productos está más presente entre clientes que realizan compras con mayor regularidad.

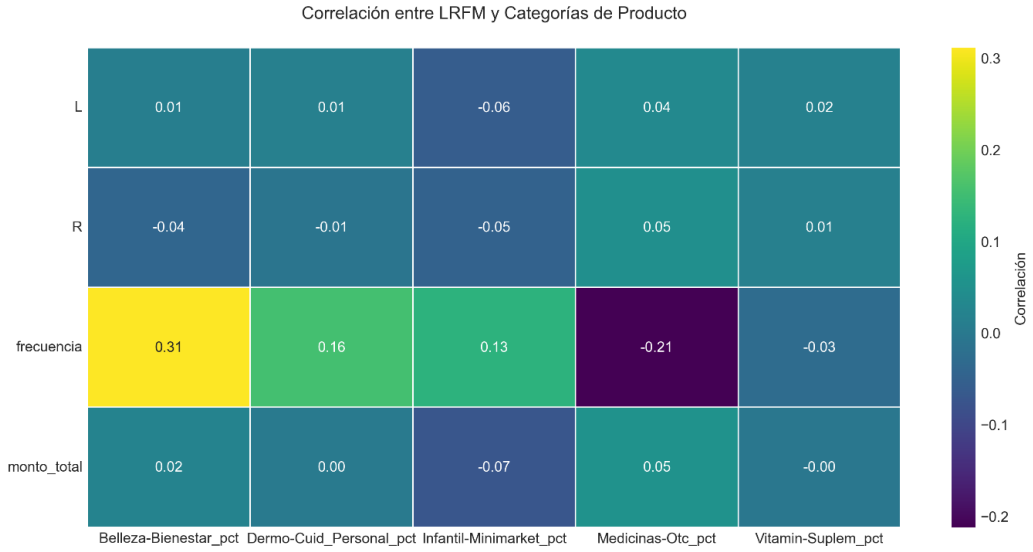

Figura 8. Correlaciones entre variables LRFM y categorías de producto

La figura 8, permite observar de manera específica la interacción entre las dimensiones LRFM y la composición del gasto por categoría, evidenciando que las asociaciones directas entre comportamiento transaccional y tipo de producto son, en general, de magnitud baja a moderada.

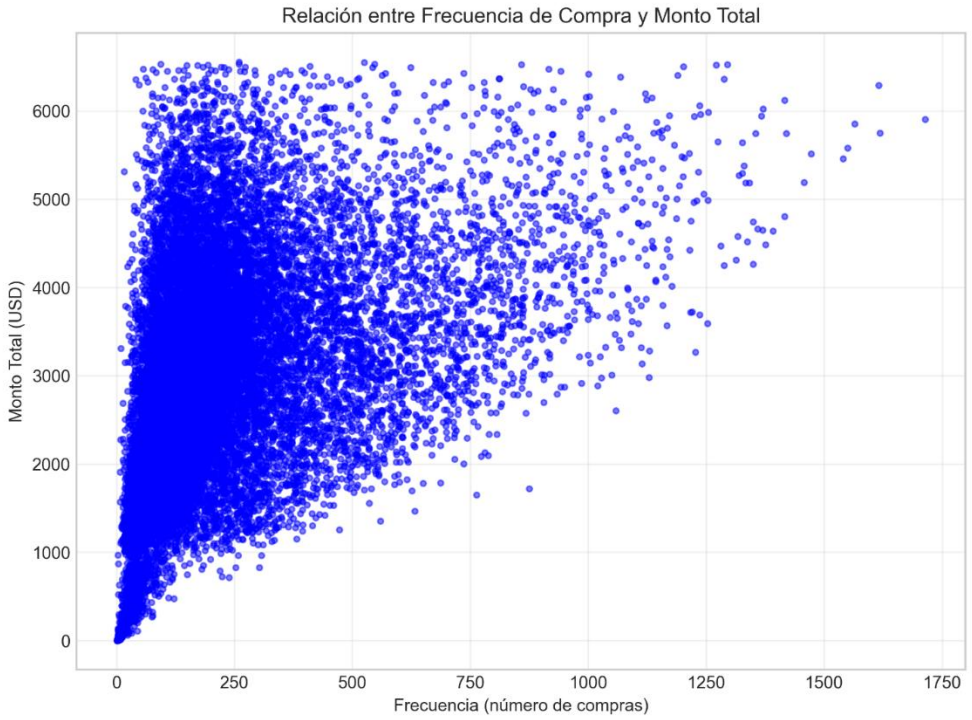

Figura 9. Relación entre frecuencia de compra y monto total

El gráfico refuerza lo que ya se había identificado en los coeficientes de correlación: existe una relación directa entre la frecuencia de compra y el gasto total. A medida que aumenta el número de transacciones, también se incrementa el monto acumulado.

Sin embargo, la nube de puntos muestra que esta relación no es perfectamente lineal. En los niveles más altos de frecuencia se aprecia una mayor dispersión en los valores de gasto indicando que no todos los clientes frecuentes desembolsan la misma cantidad. Algunos concentran montos significativamente superiores, mientras que otros mantienen niveles más moderados. Se puede decir que influyen otros elementos, como el tipo de productos adquiridos, la combinación de categorías dentro de cada ticket y el valor unitario de los artículos. En consecuencia, aunque la frecuencia constituye un factor relevante para explicar el monto total, no actúa de manera aislada dentro del patrón de consumo.

3.2. Reducción de Dimensionalidad con PCA

La varianza explicada acumulada por las componentes principales se presenta en la Figura 10, donde se observa la proporción de información retenida a medida que aumenta el número de componentes consideradas.

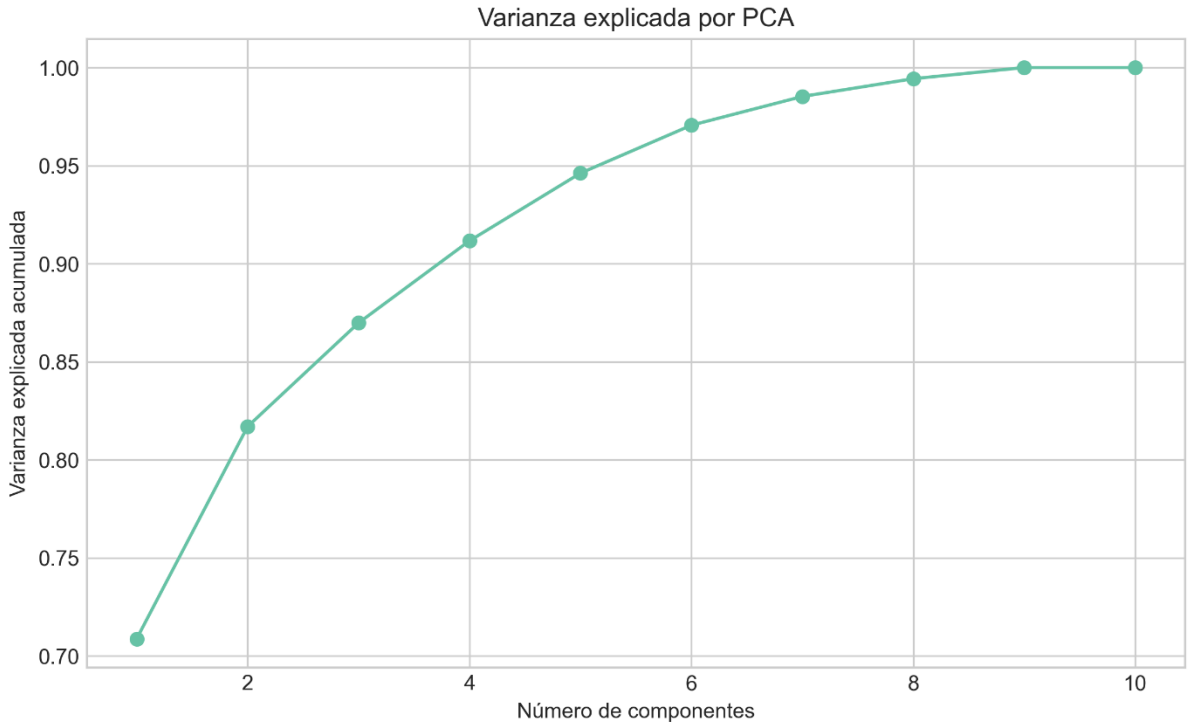

Figura 10. Varianza explicada acumulada del PCA

El análisis de componentes principales muestra que la primera componente concentra cerca del 71 % de la varianza total, lo que ya representa una parte importante de la información original. Al incorporar la segunda componente, la varianza acumulada asciende aproximadamente al 82 %, y con tres componentes se sitúa alrededor del 87 %. Con la cuarta, se supera el 90 %, alcanzando cerca del 91 %, lo que indica que la mayor parte de la estructura

del conjunto de datos puede resumirse en un espacio de menor dimensión sin pérdida relevante de información.

El aporte adicional va disminuyendo después de la quinta componente: con cinco se logra aproximadamente el 95 % y con seis se sobrepasa el 97 %. Esta conducta indica que, más allá de un determinado punto, cada nueva componente añade escasa información extra.

Se decidió trabajar con las primeras cuatro componentes principales porque, según estos resultados, posibilitan la conservación de más del 90 % de la varianza total. Esta disminución favorece la estabilidad de los algoritmos de agrupamiento que se emplean en la etapa siguiente, simplifica la estructura de los datos y hace más fácil el posterior análisis

3.3. Segmentación de Clientes

3.3.1. Determinación del Número Óptimo de Clusters (K-means)

Para el algoritmo K-means, se evaluó un rango de 2 a 15 clusters. El método del codo no mostró un punto de inflexión claro, por lo que la selección del número óptimo de clusters se basó principalmente en el Coeficiente de Silueta, que mide la cohesión y separación de los clusters. Los valores obtenidos se presentan en la Tabla 6.

Tabla 6. Coeficiente de Silueta para distintos valores de *k*

k	Coeficiente de Silueta
2	0.7719
3	0.7402
4	0.4930
5	0.5082
6	0.3440
7	0.3579
8	0.3763
9	0.3795
10	0.3780
11–15	0.2864 – 0.2962

El valor más alto de silueta se alcanzó con k=2 (0.7719), indicando que una partición en dos grandes grupos es la que mejor se ajusta a la estructura de los datos proyectados en el espacio PCA. Por lo tanto, se seleccionó k=2 como el número óptimo para el modelo K-means. La evidencia gráfica de este comportamiento se presenta en la Figura 11.

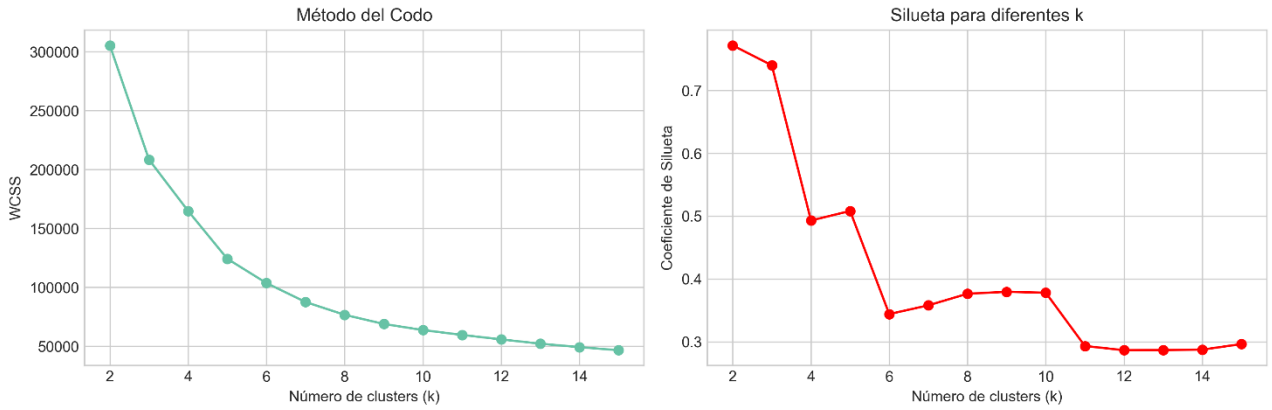

Figura 11. Evaluación del número óptimo de clusters mediante método del codo y coeficiente de silueta

3.3.2. Clustering Jerárquico

La estructura jerárquica de los datos se presenta en la Figura 12, donde se muestra el dendrograma obtenido mediante el método aglomerativo de Ward aplicado a una muestra de 1000 clientes.

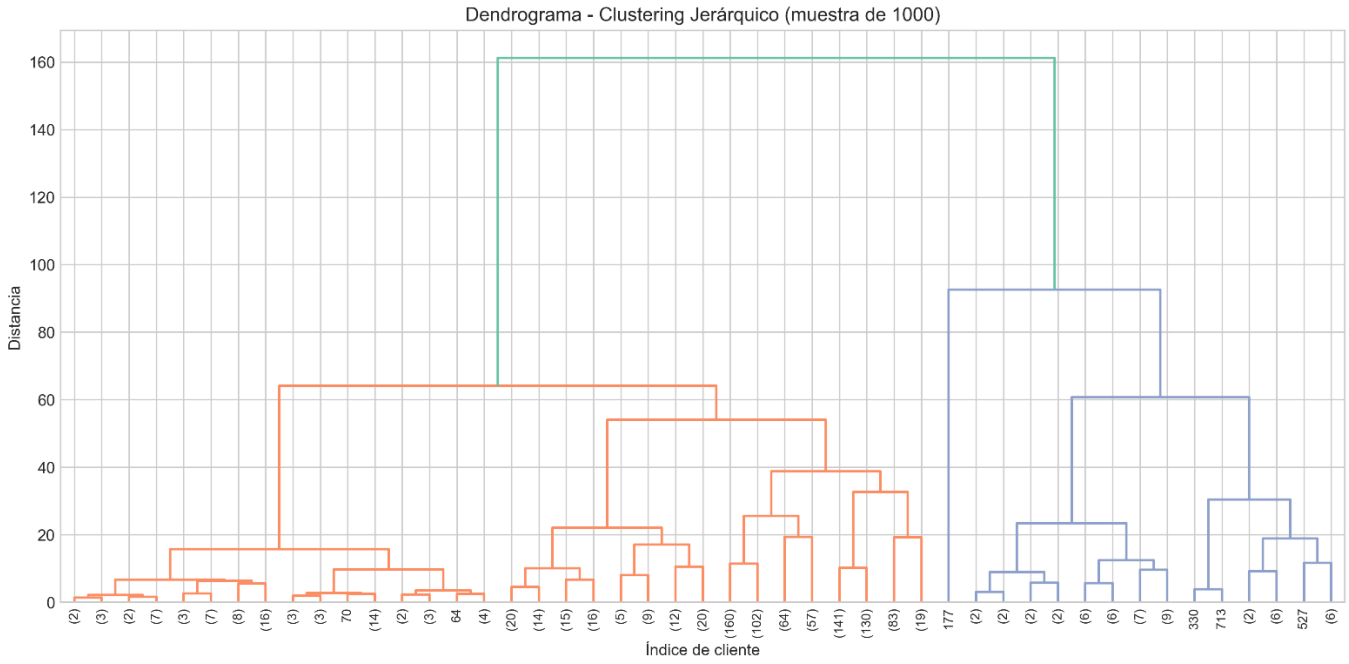

Figura 12. Dendrograma del clustering jerárquico

En el dendrograma se observa dos conglomerados principales que recién se unen a una distancia superior a 160 indicando una separación marcada entre ambos conjuntos. Dentro de cada grupo, las fusiones ocurren a distancias mucho menores a 60. La diferencia entre las alturas de unión muestra que los datos tienen una estructura natural bien definida en dos bloques.

En este caso se escogió dos conglomerados como una solución consistente y metodológicamente adecuada, ya que representa de manera clara la estructura jerárquica observada.

Si se optara por una solución con un mayor número de clusters, sería necesario realizar el corte en niveles más bajos del dendrograma, lo que implicaría subdividir grupos con diferencias menos pronunciadas.

3.3.3. Clustering DBSCAN

El algoritmo DBSCAN, basado en densidad, permite identificar agrupamientos naturales en los datos sin necesidad de especificar previamente el número de clusters. Con el objetivo de determinar la configuración óptima de parámetros, se realizó una búsqueda sistemática combinando distintos valores de eps y min_samples, evaluando cada resultado mediante el Coeficiente de Silueta como medida de calidad del agrupamiento.

Los resultados de esta exploración se presentan en la Tabla 7, donde se sintetiza el desempeño de cada combinación en términos del número de clusters generados, cantidad de observaciones clasificadas como ruido y valor del índice de silueta.

Tabla 7. Resultados de la búsqueda de parámetros para DBSCAN

eps	min_samples	Nº clusters	Nº ruido	Silueta
0.3	5	67	5561	-0.4178
0.3	10	17	7684	-0.3138
0.3	15	10	8858	-0.2302
0.3	20	3	9678	0.0108
0.5	5	36	2240	-0.1816
0.5	10	8	3234	-0.0936
0.5	15	5	3847	0.0791
0.5	20	2	4416	0.1633
0.7	5	21	1129	0.0721
0.7	10	10	1732	0.3925
0.7	15	2	2141	0.4489
0.7	20	3	2440	0.5277
1.0	5	7	475	0.4343
1.0	10	4	746	0.5595
1.0	15	3	921	0.6756
1.0	20	4	1115	0.6501
1.5	5	3	154	0.7994
1.5	10	3	256	0.6311
1.5	15	2	327	0.6343
1.5	20	1	383	N/A
2.0	**5**	**2**	**78**	**0.8393**
2.0	10	1	117	N/A
2.0	15	1	149	N/A
2.0	20	1	171	N/A

La selección del parámetro eps se fundamentó en el análisis del gráfico de k-distance, el cual representa la distancia de cada observación a su quinto vecino más cercano (considerando min_samples = 5). En este tipo de gráfico, el valor óptimo de eps suele identificarse en el punto de mayor curvatura, comúnmente denominado “codo”.

Como se aprecia en la Figura 13, dicho punto de inflexión se ubica alrededor de 2,0, lo que respalda la pertinencia del valor seleccionado durante la búsqueda sistemática de parámetros. A partir de valores superiores a 2,0, la distancia comienza a incrementarse de forma más pronunciada, lo que podría derivar en la unión de clústeres con densidades distintas y, en consecuencia, en una segmentación menos precisa.

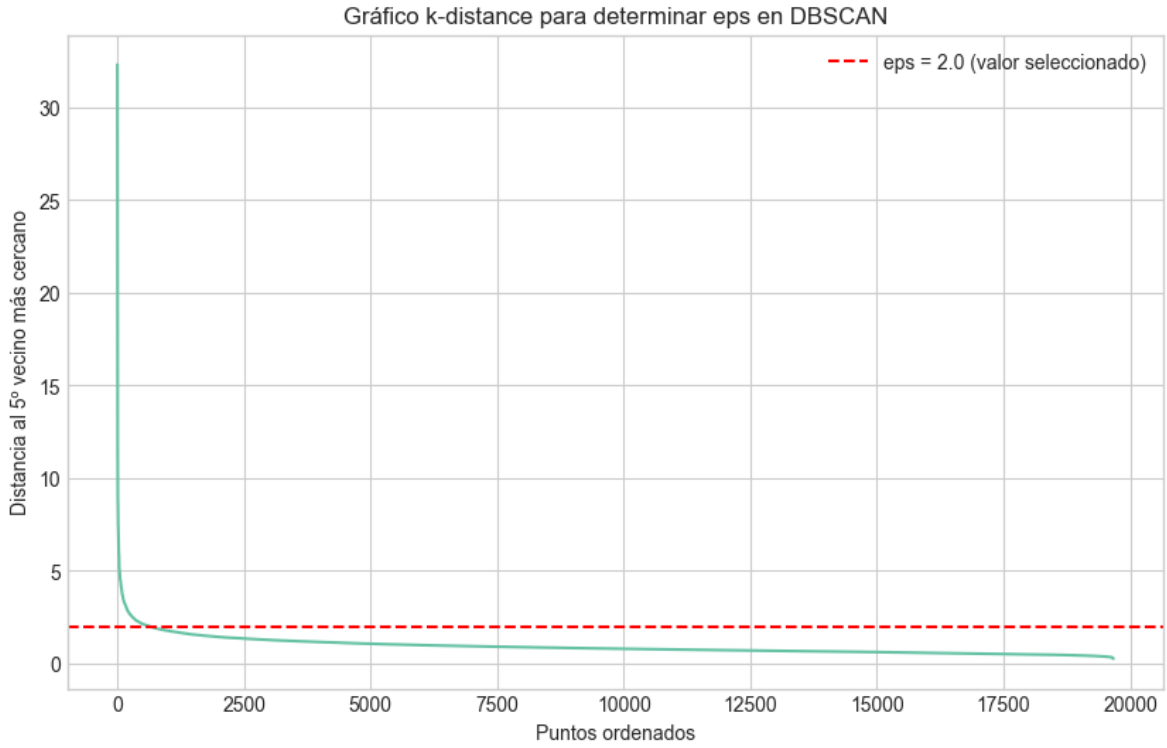

Figura 13. Gráfico k-distance para la determinación de eps en DBSCAN

El análisis comparativo muestra un patrón claro en el comportamiento del algoritmo frente a los cambios en los parámetros. Cuando se utilizan valores bajos de eps, se generan muchos clusters pequeños y una gran cantidad de puntos clasificados como ruido. En estos casos, los coeficientes de silueta son negativos o están próximos a cero; esto indica que las segmentaciones no son estables y se encuentran distantes de la estructura real de los datos.

Conforme el valor de eps se eleva, la cantidad de agrupaciones disminuye y la calidad del agrupamiento mejora, lo que se traduce en una distribución más coherente según la densidad. La mejor actuación fue la que se logró con eps = 2.0 y min_samples = 5. Con estos parámetros, se lograron dos grupos principales y únicamente 78 observaciones fueron

clasificadas como ruido; el coeficiente de silueta fue de 0.8393, el más elevado entre las combinaciones analizadas.

Además, se notó que si aumenta el número de min_samples mientras eps permanece constante, el modelo tiende a disminuir la cantidad de clusters y a incrementar los puntos considerados ruido. Esto es consistente con la lógica del método, pues un criterio de densidad más riguroso demanda una concentración superior de observaciones para crear un grupo.

En resumen, los hallazgos apoyan la selección de esta parametrización porque consigue retratar correctamente la estructura de densidad de los datos y proporciona una segmentación estable para el análisis ulterior.

3.3.4. Comparación de Modelos

Con el fin de evaluar el desempeño de los algoritmos de segmentación implementados, se realizó una comparación cuantitativa basada en dos métricas de validación interna: el Coeficiente de Silueta y el Índice Davies–Bouldin. Los resultados obtenidos se presentan en la Tabla 8.

Tabla 8. Comparación de métricas de desempeño entre modelos de clustering

Modelo	Coeficiente de Silueta	Índice Davies-Bouldin
K-means (k=2)	0.7719	0.5957
Jerárquico	0.7249	0.7018
DBSCAN (eps = 2.0, min_samples = 5)	0.8393	0.5035

La elección de DBSCAN como el mejor modelo no solo se basa en su superioridad numérica en las métricas de validación, sino también en su mayor coherencia interpretativa y estabilidad estructural. Mientras que K-means, al asumir clusters esféricos de tamaño similar, forzó una partición en dos grupos que, si es cierto que mostraban una silueta alta (0.7719), podrían no reflejar la complejidad real de los datos, como lo evidencian los 78 puntos clasificados como ruido por DBSCAN. El clustering jerárquico, aunque útil para explorar la estructura de los datos, resultó en una segmentación menos compacta y con una separación inter-cluster inferior (silueta 0.7249, Davies-Bouldin 0.7018), según reflejan sus métricas.

El modelo DBSCAN, al basarse en la densidad, identificó de forma natural los dos grandes conglomerados de comportamiento (clientes activos de alto valor e inactivos) y, además, aisló un conjunto de observaciones con patrones atípicos (ruido). Esta capacidad de detectar outliers de manera inherente es una ventaja crucial en un contexto de negocio, dado que permite identificar clientes con comportamientos excepcionales que podrían requerir estrategias de marketing específicas o, por el contrario, ser excluidos de análisis masivos para no distorsionar los patrones generales. La alta silueta (0.8393) indica que, una vez excluido

el ruido, los dos clusters principales están muy bien definidos y separados, lo que garantiza la estabilidad de la segmentación y la confianza en las estrategias derivadas de ella.

La representación gráfica de los clusters obtenidos se muestra en la Figura 14, donde se visualiza la proyección de los datos sobre las dos primeras componentes principales (PC1 y PC2).

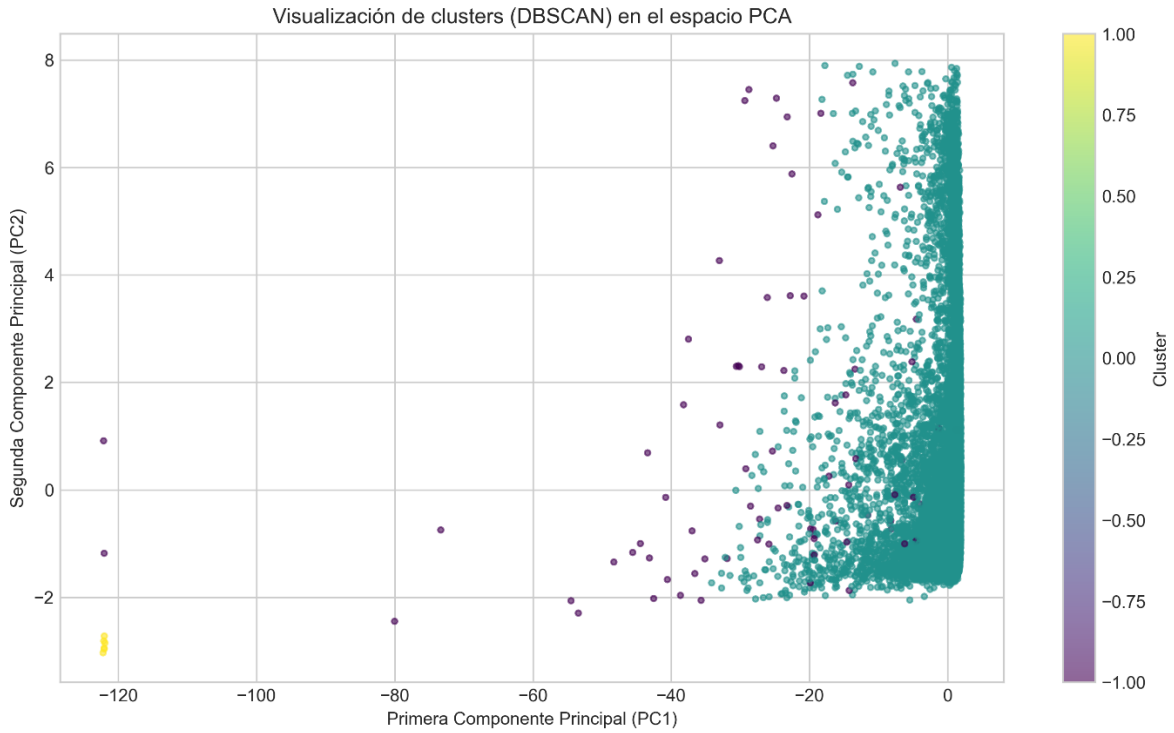

Figura 14. Distribución de clusters obtenidos con DBSCAN en el espacio PCA

3.4. Caracterización de los Segmentos Obtenidos (con DBSCAN)

La segmentación final se llevó a cabo mediante el algoritmo DBSCAN, el cual permitió detectar agrupaciones de manera automática, sin fijar previamente el número de clusters. Este método resultó ser el más adecuado ya que los datos que se presentan en formas irregulares basados en centroides suelen perder precisión.

En la Figura 15, se aprecia de forma general la diferenciación entre los grupos identificados, evidenciando patrones de comportamiento distintos dentro del conjunto analizado.

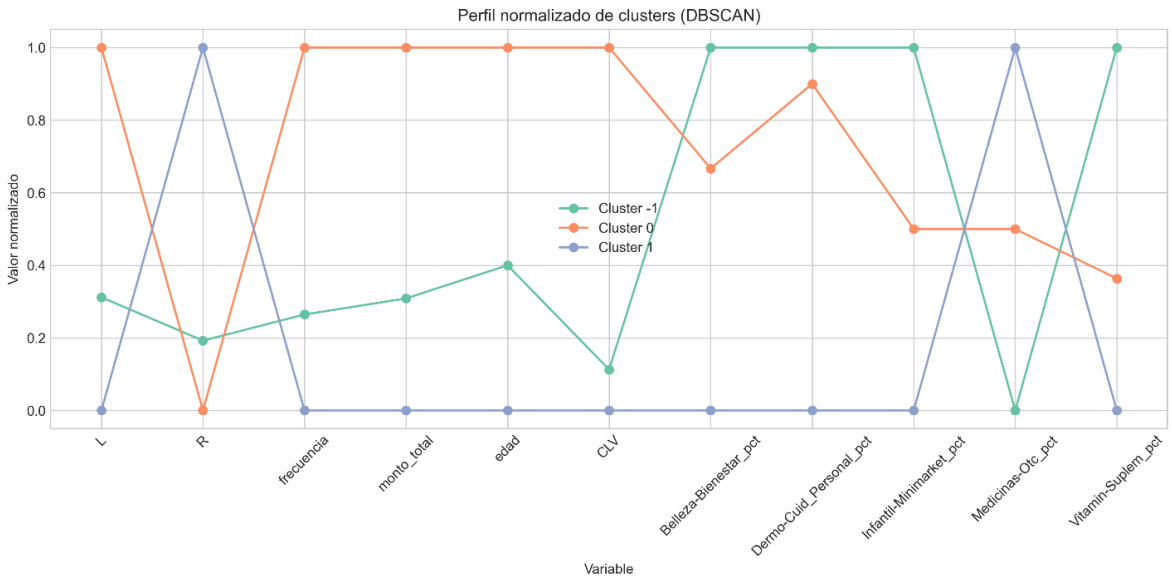

Figura 15. Perfil comparativo de los clusters obtenidos mediante DBSCAN

3.4.1. Perfil Numérico de los Clusters

Tras esta visualización inicial, se analizan con mayor detalle las características cuantitativas de cada segmento. La Tabla 9 presenta los valores promedio de las variables utilizadas en la segmentación: dimensiones LRFM, edad, proporción de compras en Medicinas-OTC y valor de vida del cliente. Esta síntesis permite comparar los clusters e interpretar su relevancia dentro del conjunto analizado.

Tabla 9. Perfil numérico promedio de los clusters identificados

cluster	L	R	frecuencia	monto_total	edad	Belleza-Bienestar_pct	Dermo-Cuid_Personal_pct	Infantil-Minimarket_pct	Medicinas-Otc_pct	Vitamin-Suplem_pct	CLV
-1	317.62	77.4	60.65	921.65	47.59	0.06	0.1	0.22	0.4	0.22	88054.65
0	1018.35	8.41	226.55	2976.66	55.72	0.04	0.09	0.11	0.7	0.08	780095.94
1	0	366.83	1	3.78	42.17	0	0	0	1	0	3.78

Nota metodológica con respecto a el CLV:

El CLV (Customer Lifetime Value) presentado en esta tabla es una métrica puramente descriptiva y retrospectiva, calculada post-hoc (después de la segmentación) como el producto de la frecuencia de compra por el monto total por cliente (CLV = frecuencia × monto_total). Su único propósito es cuantificar y comparar el valor económico histórico generado por cada segmento para facilitar su priorización estratégica. Es fundamental destacar que el CLV no fue utilizado como variable de entrada en el algoritmo DBSCAN ni en ningún otro modelo de clustering, evitando así cualquier tipo de circularidad o sesgo en la formación de los grupos. La segmentación se basa exclusivamente en las dimensiones LRFM y los porcentajes de gasto por categoría.

Para complementar el análisis de los promedios y visualizar la distribución de las variables clave dentro de cada cluster, se elaboraron diagramas de caja y bigotes (boxplots) para las dimensiones LRFM (L, R, Frecuencia y Monto Total). La Figura 16 presenta estos gráficos, donde se puede apreciar la mayor dispersión y los valores atípicos en el cluster -1 (ruido), la alta concentración y homogeneidad del cluster 0 (clientes de alto valor) y los valores mínimos del cluster 1 (clientes inactivos). Esta visualización refuerza la interpretación de los segmentos y la calidad de la separación lograda por DBSCAN.

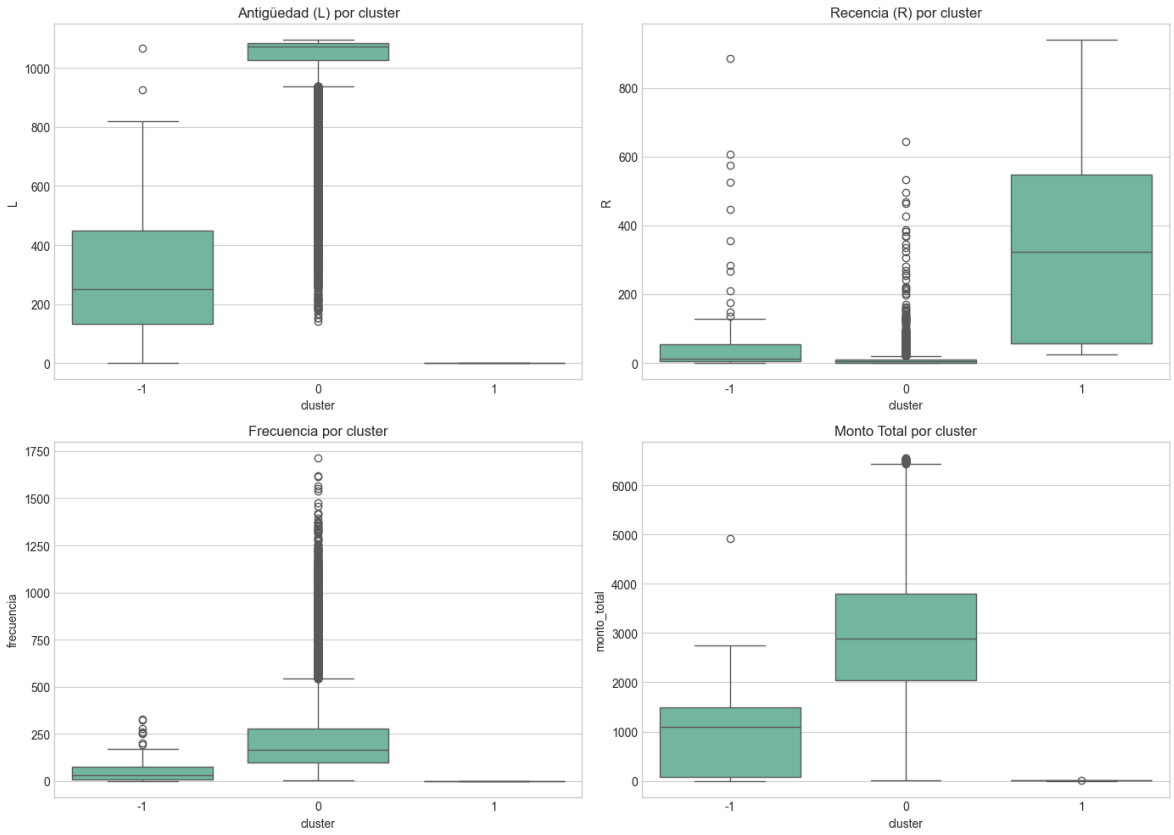

Figura 16. Distribución de variables LRFM por cluster (boxplots)

Conviene aclarar el rol del Customer Lifetime Value (CLV) dentro de esta caracterización. El CLV presentado en la Tabla 9 constituye una métrica retrospectiva y de carácter descriptivo, calculada post-hoc como el producto de la frecuencia de compra por el monto total (CLV = frecuencia * monto_total). Su propósito es estimar el valor económico histórico generado por cada segmento durante el período de análisis, facilitando así su jerarquización desde una perspectiva financiera. Es importante destacar que el CLV no fue utilizado como variable de entrada en el algoritmo DBSCAN; incorporación en el perfil de los clusters responde exclusivamente a fines de validación e interpretación de negocio, para cuantificar el impacto económico de cada segmento identificado.

El análisis permitió distinguir tres perfiles dentro de la base de clientes. El Cluster −1, identificado como ruido, representa cerca del 0.4 % de los registros. Agrupa clientes con baja

actividad y compras poco consistentes, aunque el CLV total es alto, se debe a algunos valores extremos indicando su carácter atípico.

El Cluster 0 tiene la mayoría de los clientes y es el más relevante en términos económicos ya que tiene mayor antigüedad, baja recencia y los niveles más altos de frecuencia y gasto. El consumo se enfoca principalmente en Medicinas-OTC con un patrón continuo. Además, registra el CLV más elevado, el cual registra más ingresos.

El Cluster 1 reúne a un grupo reducido con baja participación, este se caracteriza por alta recencia y mínimos niveles de frecuencia y gasto por sus compras aisladas. Su aporte económico es limitado.

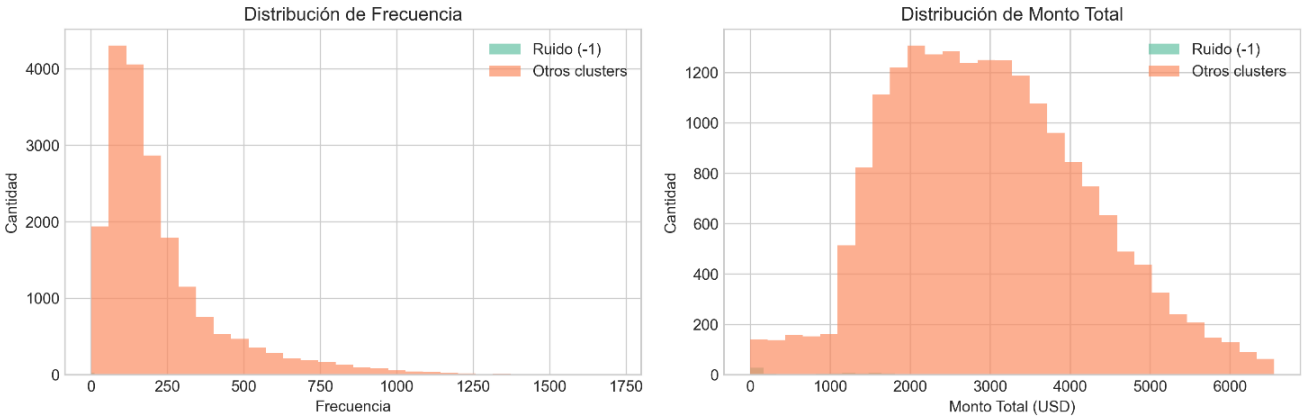

Figura 17. Distribución comparativa de frecuencia y monto total entre clusters

En la Figura 17 se puede observar estas diferencias, mostrando que el grupo de ruido se concentra en rangos bajos de frecuencia y monto, aunque incluye algunos casos extremos reforzando la capacidad del método para identificar comportamientos atípicos dentro del conjunto de datos.

3.4.2. Perfil Demográfico de los Clusters

Con el propósito de comprender con mayor profundidad las características de los clientes clasificados como ruido por el algoritmo DBSCAN, se llevó a cabo un análisis descriptivo de variables demográficas y de consumo relevantes. Este análisis permite examinar con mayor detalle la estructura interna de este segmento atípico y evaluar su posible significado dentro de la dinámica general de la base de clientes. Los resultados se presentan en las Tablas 10, 11 y 12 donde se sintetizan respectivamente la distribución por sexo, el nivel de instrucción y la composición del gasto por categorías de producto.

Tabla 10. Distribución por sexo del cluster de ruido

sexo	FEMENINO	MASCULINO
-1	0.5769	0.4231
0	0.5300	0.4700
1	0.6667	0.3333

Tabla 11. Nivel de instrucción del cluster de ruido

instrucción	bachiller	básica	elemental	especial	inicial	ninguna	primaria	secundaria	superior
-1	0.2692	0.1923	0.0000	0.0000	0.0256	0.0513	0.0513	0.0769	0.3333
0	0.3218	0.1610	0.0019	0.0007	0.0122	0.0128	0.0277	0.0506	0.4113
1	0.0000	0.3333	0.0000	0.0000	0.0000	0.0000	0.1667	0.0000	0.5000

Tabla 12. Distribución del gasto por categoría de producto en el cluster de ruido

cluster	Belleza-Bienestar_pct	Dermo-Cuid_Personal_pct	Infantil-Minimarket_pct	Medicinas-Otc_pct	Vitamin-Suplem_pct
-1	0.06	0.1	0.22	0.4	0.22

En el cluster de ruido se observa una ligera mayoría de mujeres con un 57.7 %, este porcentaje es muy parecido al de los otros grupos. Esto permite inferir que la clasificación como atípicos no guarda relación con el sexo, sino con la forma en que realizan sus compras.

En cuanto al nivel educativo, los clientes con estudios básicos y de bachillerato predominan, la presencia de personas con educación superior es menor frente al segmento principal, lo que podría estar asociado a diferencias en los hábitos de consumo, aunque no explica por sí solo el comportamiento irregular.

El gasto se distribuye de manera menos concentrada que en el cluster dominante por ejemplo, Medicinas-OTC representa el 40 %, pero categorías como Infantil-Minimarket y Vitaminas-Suplementos alcanzan cada una el 22 %, mostrando un patrón más variado indicando que las compras son ocasionales y no necesariamente continuas.

También aparecen algunos casos con montos superiores a 1,495 USD. A pesar de su baja frecuencia, realizan compras de alto valor, posiblemente vinculadas a necesidades puntuales.

En resumen, este grupo reúne perfiles diversos: clientes esporádicos y algunos compradores de alto gasto con poca recurrencia, esta identificación ayuda a entender comportamientos menos estables dentro de la base analizada.

3.5. Importancia de las Variables en la Segmentación

La Figura 18 muestra qué variables influyen más en la clasificación de los clientes dentro de los clusters obtenidos con DBSCAN. Para estimarlo, se entrenó un modelo Random Forest que permitió identificar qué factores pesan más en la asignación.

Se observa que L_log destaca claramente sobre las demás, es decir la antigüedad del cliente es el elemento que más influye en la formación de los grupos. En un segundo nivel aparecen Medicinas-Otc_pct, Vitamin-Suplem_pct, frecuencia_log y monto_total_log, reflejando que el nivel y tipo de gasto también marcan diferencias entre perfiles.

En cambio, Infantil-Minimarket_pct y R_log tienen menor incidencia, y variables como edad, Belleza-Bienestar_pct y Dermo-Cuid_Personal_pct muestran un peso pequeño en la segmentación quedando como limitado su influencia en la segmentación.

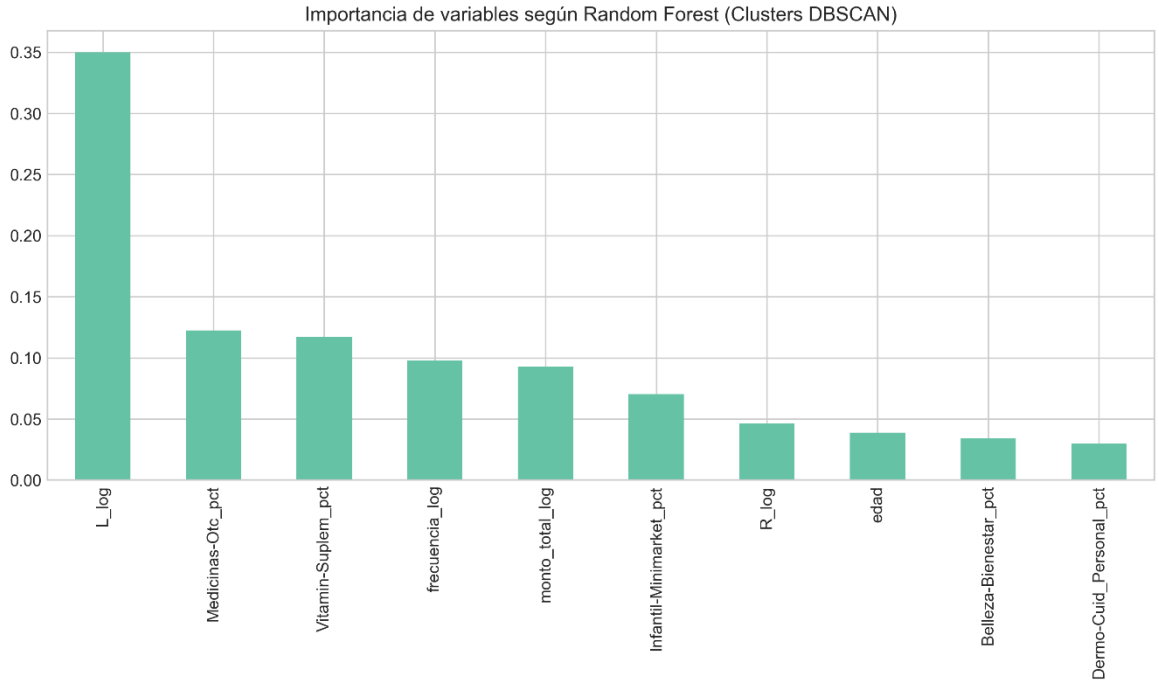

Figura 18. Relevancia predictiva de variables para la asignación a clusters

4. CAPITULO 4: DISCUSIÓN

4.1. Discusión

Los resultados obtenidos tienen coherencia con la revisión bibliográfica mostrada en el Capítulo 2 ya que la segmentación muestra que la antigüedad y el valor monetario son los factores que más diferencian a los clientes, lo que coincide con los planteamientos del modelo LRFM y con estudios previos que señalan estas variables como centrales para explicar el valor del cliente (Aslantaş, Rumelli, & Bakırlı, 2023; Rahmadianti et al., 2020b). En este caso, el tiempo de relación y el nivel de gasto terminan siendo los ejes que estructuran los perfiles.

Además, el análisis muestra que las variables como la frecuencia y la distribución del gasto tienen mayor peso en la segmentación que las características demográficas. En términos prácticos, las diferencias entre clientes se explican principalmente por su consumo y nivel de interacción comercial, más que por factores como la edad.

DBSCAN presentó resultados más fiables que K-means y el método jerárquico cuando se compararon los métodos. DBSCAN, al identificar agrupaciones en función de su densidad, se adaptó mejor a la forma real de los datos. Por el contrario, K-means necesita establecer antes cuántos grupos habrá y supone estructuras más o menos homogéneas. Esto fue particularmente importante en la investigación donde las pautas de compra no siguen patrones uniformes ni distribuciones consistentes. El modelo, además, no solo definió de manera precisa el segmento principal con un alto valor, sino que también separó los casos inusuales.

Asimismo, en este estudio actual la segmentación contribuye directamente a la administración de las actividades comerciales del sector farmacéutico minorista de Ecuador. Se encontró un grupo de clientes leales y valiosos que confirman que la duración en el tiempo y la repetición de las compras son fundamentales para mantener estables los ingresos.

Simultáneamente, la detección de clientes con conductas irregulares y de un grupo inactivo hace necesario gestionar los recursos con más criterio. No todos los perfiles producen el mismo rendimiento, por lo que tiene sentido modificar la cantidad de inversión en función del valor real y del potencial de cada segmento.

Para resumir, los resultados muestran que el enfoque LRFM sigue vigente y que las técnicas basadas en densidad logran captar con más exactitud la diversidad de comportamientos en la base de clientes. La contribución de esta investigación consiste en unir el soporte teórico, el análisis y la implementación prácticos en un ámbito donde la gestión del cliente es crucial.

4.2. Implicaciones Metodológicas y de Negocio

Al analizar los resultados, resulta revelador que el algoritmo DBSCAN haya ofrecido un mejor ajuste que los métodos tradicionales de particionamiento o jerárquicos. Este hecho sugiere que la estructura subyacente de los datos de compra en el retail farmacéutico no es

esférica ni de tamaños homogéneos, sino que se organiza en regiones de alta densidad (clientes con comportamiento "típico") separadas por regiones de menor densidad. Dicha característica es coherente con la existencia de segmentos mayoritarios con patrones de compra bien definidos (como el cluster 0) y de una minoría con comportamientos atípicos o poco representativos (cluster -1).

Desde la visión del negocio, esta segmentación basada en densidad permite una gestión de cartera más matizada y eficiente. No solo permite identificar los clientes más valiosos (cluster 0), sobre los cuales se deben concentrar los esfuerzos de retención y fidelización, sino que también ayuda a distinguir a aquellos con potencial de reactivación (algunos clientes en el cluster -1 con alto gasto) y aquellos que posiblemente nunca se convertirán en compradores recurrentes (cluster 1). De esta manera es posible optimizar la asignación de recursos destinados a marketing, evitando destinar inversión en segmentos de bajo o nulo retorno y diseñar estrategias de "reactivación" y/o "acompañamiento" para los casos atípicos que podrían generar valor.

Es importante enfatizar la estabilidad temporal de esta segmentación. Los segmentos identificados están basados en comportamientos de compra consolidados a lo largo de un período de análisis (L, F, M). Sin embargo, los patrones de consumo no son estáticos. Variaciones en el mercado, la introducción de nuevos productos, campañas de la competencia o incluso factores estacionales (como picos de enfermedades respiratorias) podrían alterar la densidad de los datos y, por ende, la composición de los clusters. Por lo tanto, se recomienda que este modelo de segmentación no sea estático, sino que se actualice periódicamente para así capturar la dinámica del mercado y ajustar las estrategias de marketing de forma proactiva. Un seguimiento longitudinal permitiría observar la migración de clientes entre clusters y evaluar el impacto real de las acciones comerciales implementadas.

4.3. Limitaciones del Estudio y Trabajos Futuros

Si bien el enfoque metodológico empleado demuestra solidez, el presente estudio presenta algunas de las limitaciones identificadas y que deben ser consideradas al interpretar sus resultados y que, al mismo tiempo, abren líneas interesantes para futuras investigaciones.

- **Dependencia del preprocesamiento:** La segmentación obtenida es sensible a las decisiones tomadas durante la preparación de los datos, tales como la transformación logarítmica para reducir el sesgo, el escalado robusto y, especialmente, la reducción de dimensionalidad mediante PCA. A pesar de que se utilizaron técnicas robustas, distintas elecciones en estas fases podrían generar variaciones en los resultados. En consecuencia, para futuros estudios se podría explorar el impacto de omitir la reducción de dimensionalidad o de utilizar otras técnicas de escalado en la estabilidad de los clusters.
- **Elección de parámetros en DBSCAN:** El desempeño de DBSCAN depende en gran medida de la adecuada selección de sus parámetros eps y min_samples. Aunque se llevó a cabo una búsqueda sistemática utilizando el coeficiente de silueta como criterio de evaluación, ello no garantiza que la combinación obtenida sea la óptima

en todos los escenarios. En investigaciones futuras se podrían explorar métodos más sofisticados para la optimización de estos hiperparámetros.

- **Validez externa y generalización:** El análisis se realizó a partir de datos correspondientes a un único retailer farmacéutico en Ecuador, abarcando un período de tres años (2023–2025). Por lo tanto, los resultados y las estrategias derivadas no son directamente aplicables a otras empresas de esta industria, regiones o períodos de tiempo sin un estudio de validación específico. La segmentación identificada responde refleja específicamente el comportamiento de la cartera de clientes de esta organización.
- **Ausencia de validación externa:** La calidad de los clusters fue evaluada únicamente mediante métricas de validación interna. Por ello, una validación externa, la cual consiste en contrastar los segmentos con una variable de interés no utilizada en el clustering (por ejemplo, la respuesta a una campaña de marketing previa, la tasa de abandono posterior, o datos de encuestas de satisfacción), permitiría obtener una evidencia mucho más sólida sobre la utilidad práctica de la segmentación. En los siguientes trabajos que se desarrollen, se podrían diseñar estudios para evaluar si los diferentes segmentos responden de manera diferenciada ante estímulos comerciales específicos, lo que confirmaría el valor predictivo de la segmentación.
- **CLV histórico y no predictivo:** El cálculo del CLV se realizó con un enfoque retrospectivo, basándose comportamiento de compra pasado de los clientes. Un enfoque más robusto consistiría en desarrollar modelos predictivos de CLV que estimen el valor futuro de cada cliente, o que facilitaría una priorización aún más estratégica y orientada a la toma de decisiones de largo plazo.

Como una línea de trabajo futura clave para abordar la limitación de la ausencia de validación externa, se propone diseñar un experimento controlado. Este podría consistir en lanzar una campaña de marketing diferenciada para cada segmento identificado (por ejemplo, una promoción específica para el cluster 0, un programa de reactivación para ciertos clientes del cluster -1 y una comunicación de bajo costo para el cluster 1). El impacto de estas intervenciones podría evaluarse a través de indicadores como la tasa de respuesta a la campaña y el aumento en la frecuencia de compra o en el CLV durante los meses posteriores. La comparación de los resultados entre los segmentos y con un grupo de control permitiría validar externamente la relevancia y el poder predictivo de la segmentación obtenida, cerrando así el ciclo entre el análisis de datos y la estrategia de negocio.

4.4. Conclusiones

Este trabajo desarrolló un modelo de segmentación para el retail farmacéutico ecuatoriano combinando clustering y variables LRFM con fines comerciales. Se analizaron 19.670 clientes, considerando antigüedad, recencia, frecuencia, gasto total y distribución del consumo por categorías.

En la comparación de métodos (K-means, jerárquico y DBSCAN), la mejor configuración fue DBSCAN con eps = 2.0 y min_samples = 5. Alcanzó la silueta más alta (0.8393) y el menor índice Davies-Bouldin (0.5035), además de identificar 78 clientes con

comportamientos atípicos. Aunque el índice Calinski-Harabasz fue menor, esto se relaciona con la exclusión de esos puntos extremos. La separación entre grupos y su cohesión respaldaron su elección.

La segmentación permitió distinguir tres perfiles. El cluster -1 agrupa clientes esporádicos (60 compras promedio, 921 USD de gasto y consumo diversificado). Representa 0.4 % de la base, aunque incluye algunos casos de alto valor. El cluster 0 concentra el mayor aporte: 227 transacciones en promedio, recencia de 8 días y 70 % del gasto en medicamentos. Con más de 1.000 días de relación y el CLV más alto, es el segmento clave para retención. El cluster 1 reúne clientes inactivos, con una sola compra (3.78 USD) y más de un año sin actividad, de impacto económico mínimo.

Estos resultados permiten orientar mejor los recursos: seguimiento selectivo al cluster -1, fortalecimiento de fidelización en el cluster 0 y baja inversión en el cluster 1.

Entre las limitaciones, los datos provienen de un solo retailer y cubren 2023–2025. El CLV se calculó con enfoque histórico y no predictivo. Además, la reducción mediante PCA pudo modificar parcialmente las distancias usadas por DBSCAN.

Como continuidad, sería útil incorporar modelos predictivos de valor futuro, análisis temporal del comportamiento y pruebas controladas que midan el efecto real de las acciones comerciales.

La selección de DBSCAN como modelo final no obedece únicamente a su mejor desempeño métrico, sino también a su coherencia con la estructura natural de los datos, la cual, como se discutió, no se ajusta a formas esféricas. La segmentación basada en densidad permitió una caracterización más precisa de los clientes, al aislar comportamientos atípicos y definiendo grupos con un alto nivel de homogeneidad interna.

Sin embargo, es importante reconocer que esta segmentación constituye una fotografía del comportamiento de los clientes en un período determinado. Los patrones de consumo evolucionan con el tiempo; por lo tanto, se recomienda la actualización periódica del modelo para reflejar la dinámica cambiante del mercado y ajustar las estrategias de marketing de forma proactiva. Un seguimiento longitudinal permitiría observar la migración de clientes entre clusters y evaluar con mayor precisión el impacto real de las acciones implementadas.

Como línea prioritaria para investigaciones futuras, se identifica la necesidad de realizar una validación externa de los segmentos. Contrastar estos clusters con variables no utilizadas en el modelo, como la respuesta a campañas de marketing o tasas de abandono posteriores, permitiría obtener una evidencia más sólida sobre su verdadero valor predictivo y su utilidad operativa. De este modo, se fortalecería el vínculo entre el análisis de datos y la toma de decisiones estratégicas en el sector farmacéutico.

5. BIBLIOGRAFÍA

Abbasimehr, H., Setak, M., & Tarokh, M. J. (2015). An analytical framework based on RFM model and time series clustering techniques for dynamic customer segmentation. *Journal of Retailing and Consumer Services*, *24*, 1–10. https://doi.org/10.1016/j.jretconser.2014.12.010

Abednego, L., Nugraheni, C. E., & Salsabina, A. (2023). Customer Segmentation: Transformation from Data to Marketing Strategy. *ADI International Conference Series*, *4*(1). https://doi.org/10.34306/conferenceseries.v4i1.645

Akande, O. N., Asani, E. O., & Dautare, B. (2024). Customer segmentation through RFM analysis and K-means clustering: Leveraging data-driven insights for effective marketing strategy. *Ceddi Journal of Information System and Technology*, *3*(1), 45–62. https://doi.org/10.56134/jst.v3i1.81

Altuntas, E. Y. (2023). COVID-19 Pandemic Learning: The Uprising of Remote Detailing in Pharmaceutical Sector Using Sales Force Automation and Its Sustainable Impact on Continuing Medical Education. *Sustainability*, *15*(11), 8955. https://doi.org/10.3390/su15118955

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2023). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/j.patcog.2022.107734

Asamblea Nacional del Ecuador. (2021). *LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES*. www.lexis.com.ec

Aslantaş, G., Rumelli, M., & Bakırlı, G. (2023). Customer segmentation using K-means clustering algorithm and RFM model. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *25*(74), 377–387. https://doi.org/10.21205/deufmd.2023257418

Aslantaş, G., Rumelli, M., & Bakrl, G. (2023). Customer segmentation using K-means clustering algorithm and RFM model. *Journal of Engineering Sciences*, *25*(74), 407–418. https://doi.org/10.21205/deufmd.2023257418

Asmat, F., Suryadi, K., & Govindaraju, R. (2023). Data mining framework for the identification of profitable customer based on recency, frequency, monetary (RFM). *AIP Conference Proceedings*, *2540*(1), 50008. https://doi.org/10.1063/5.0130290

Barrera, F., Segura, M., & Maroto, C. D. (2023). Multiple criteria decision support system for customer segmentation using a sorting outranking method. *Expert Systems with Applications*, *238*, 122310. https://doi.org/10.1016/j.eswa.2023.122310

Bhardwaj, A., & others. (2024). Domain Knowledge in CRISP-DM: An Application Case in Healthcare. *IFAC-PapersOnLine*, *57*(2), 123130. https://doi.org/10.1016/j.ifacol.2024.07.559

Celine, M., & Kristiyanti, D. A. (2025a). Promotion strategy optimization with sales data analysis: Comparison of K-Means, K-Medoids, and Gaussian Mixture Models for

Customer Segmentation. *Proceedings of the International Conference on Engineering Research and Applications*, 1–8. https://doi.org/10.1109/icera66156.2025.11087310

Celine, M., & Kristiyanti, D. A. (2025b). Promotion strategy optimization with sales data analysis: Comparison of K-Means, K-Medoids, and Gaussian Mixture Models for Customer Segmentation. *Proceedings of the International Conference on Engineering Research and Applications*, 1–8. https://doi.org/10.1109/icera66156.2025.11087310

Chaudhary, A. (2023a). Implementation & analysis of online retail dataset using clustering algorithms. *2023 International Conference on Inventive Computation Technologies (ICICT)*, 1456–1461. https://doi.org/10.1109/ICIEM59379.2023.10166552

Fedorchenko, A., Kulyk, A., & Ponomarenko, I. (2023). Features of the clusterization method application in marketing research of the pharmaceutical market of Ukraine. *Marketing and Digital Technologies*, *1*(7), 7–28. https://doi.org/10.15276/mdt.7.1.2023.1

Fedorovich, P. F. (2023). Comparative analysis of machine learning models for customer segmentation. In *Advances in Intelligent Systems and Computing* (Vol. 1449, pp. 61–72). Springer. https://doi.org/10.1007/978-3-031-35510-3_6

Franjić, S. (2022). The Production of Drugs Must Correspond to the Intended Purpose. *Journal of Clinical Images & Reports*, 1–5. https://doi.org/10.47363/JCIR/2022(1)101

García, A. A. (2022). *Modelo RFM y análisis de clúster en una empresa comercial en Ecuador* [Universidad Internacional de La Rioja]. https://reunir.unir.net/handle/123456789/13985

Greengrove, K. (2002). Needs-based segmentation: principles and practice. *International Journal of Market Research*, *44*(4), 405–421. https://doi.org/10.1177/147078530204400402

Güney, O. \. I. (2019). Consumption attributes and preferences on medicinal and aromatic plants: A consumer segmentation analysis. *Ciencia Rural*, *49*(4), e20180840. https://doi.org/10.1590/0103-8478CR20180840

Husein, A. M., Waruwu, F. K., Bara, Y. M. T. B., & others. (2021). Clustering algorithm for determining marketing targets based customer purchase patterns and behaviors. *SINKRON: Jurnal Dan Penelitian Teknik Informatika*, *6*(1), 98–106. https://doi.org/10.33395/sinkron.v6i1.11191

Joga, P., Harshini, B., & Sahay, R. (2023). *Comparative Analysis of Machine Learning Models for Customer Segmentation* (pp. 49–61). https://doi.org/10.1007/978-3-031-35510-3_6

Kasem, M. S., Hamada, M., & Taj-Eddin, I. A. T. F. (2023). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995–5005. https://doi.org/10.1007/s00521-023-09339-6

Katyayan, A., Bokhare, A., Gupta, R. K., & Rathore, P. S. (2022a). Analysis of unsupervised machine learning techniques for customer segmentation. In *Data Analytics and Artificial Intelligence for Inventory and Supply Chain Management* (pp. 535–549). Springer. https://doi.org/10.1007/978-981-16-7996-4_35

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data*. Wiley. https://doi.org/10.1002/9780470316801

Kevrekidis, D. P., Minarikova, D., Markos, A., Malovecká, I., & Minárik, P. (2018). Community pharmacy customer segmentation based on factors influencing their selection of pharmacy and over-the-counter medicines. *Saudi Pharmaceutical Journal*, *26*(1), 33–43. https://doi.org/10.1016/j.jsps.2017.11.002

Kevrekidis, D. P., Minarikova, D., Markos, A., & Souliotis, K. (2018). Community pharmacy customer segmentation based on factors influencing their selection of pharmacy and over-the-counter medicines. *Saudi Pharmaceutical Journal*, *26*(1), 33–43. https://doi.org/10.1016/j.jsps.2017.11.002

Kotler, P., & Armstrong, G. (2018a). *Principles of marketing* (17th ed.). Pearson. https://doi.org/10.4324/9781315660664

Kumar, R., Chakkerwar, K., Dhal, R., & Chandwani, V. (2024). A QUALITATIVE STUDY ON THE ADOPTION OF OMNICHANNEL MARKETING IN PHARMACEUTICAL INDUSTRY OF INDIA. *International Journal of Management, Economics and Commerce*, *1*(2), 39–53. https://doi.org/10.62737/nz8nvs43

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2023). Understanding and enhancement of internal clustering validation measures. *IEEE Transactions on Cybernetics*, *53*(5), 2716–2729. https://doi.org/10.1109/TCYB.2021.3129256

Mariscal, J. R., & others. (2024). The evolution of CRISP-DM for Data Science. *Revista Reviews*, *1*(1), 3757. https://doi.org/10.5753/reviews.2024.3757

Momahhed, S. S., Emamgholipour Sefiddashti, S., Minaei, B., & Shahali, Z. (2023). K-means clustering of outpatient prescription claims for health insureds in Iran. *BMC Public Health*, *23*, 815. https://doi.org/10.1186/s12889-023-15753-1

Nikaein, N., & Abedin, E. (2021). Customers' segmentation in pharmaceutical distribution industry based on the RFML model. *International Journal of Business Information Systems*, *36*(3), 315–334. https://doi.org/10.1504/ijbis.2021.115071

Ofori, E. A., & others. (2025). CRISP-DM-Based Mobile Application for Predicting High-Risk Areas. *Journal of Computer Science & Statistical Physics*, *2026*, 649–659. https://doi.org/10.3844/jcssp.2026.649.659

Palupi, G. S., & Fakhruzzaman, M. N. (2022). Indonesian pharmacy retailer segmentation using recency frequency monetary-location model and ant K-means algorithm. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(6), 6132. https://doi.org/10.11591/ijece.v12i6.pp6132-6139

Panda, A. R., Rout, P., & Gautam, A. (2024). Optimizing Marketing Strategy by Performing Customer Segmentation. *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, 1–6. https://doi.org/10.1109/ACROSET62108.2024.10743643

PILLI SRI DURGA, J. A. PAULSON, & MARRI SRINIVASAREDDY. (2023). Customer segmentation analysis for improving sales using clustering. *International Journal of Science and Research Archive*, *9*(2), 708–715. https://doi.org/10.30574/ijsra.2023.9.2.0663

Rahmadianti, R., Dhini, A., & Laoh, E. (2020a). Estimating customer lifetime value using LRFM model in pharmaceutical and medical device distribution company. *Proceedings of the 2020 International Conference on Information System and Technology (ICISS)*, 1–6. https://doi.org/10.1109/ICISS50791.2020.9307592

Risky, F., Hartanto, F., Azkia, N., & Heikal, J. (2024). Customer segmentation based on RFM analysis as the basis of marketing strategy in pharmaceutical companies. *Budgeting: Journal of Accounting and Public Sector*, *5*(2), 8994. https://doi.org/10.31539/budgeting.v5i2.8994

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shchekoldin, V., Tsoy, M., & Shevtsov, V. (2023). Identifying consumers of pharmaceutical products and developing their sustainable loyalty. *E3S Web of Conferences*, *451*, 05004. https://doi.org/10.1051/e3sconf/202345105004

Syaputra, A., & others. (2020a). Customer segmentation in the pharmaceutical industry using RFM-based clustering. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(5), 5210–5219. https://doi.org/10.11591/ijece.v10i5.pp5210-5219

Tan Jongprasert, T. (2023). Customer lifetime value prediction and customer segmentation analysis in a community pharmacy setting. *Research in Social and Administrative Pharmacy*, *19*(6), 892–901. https://doi.org/10.1016/j.sapharm.2023.02.015

Turkmen, B. S. (2022a). Customer segmentation with machine learning for online retail industry. *European Journal of Social & Behavioural Sciences*, *31*(2), 68–82. https://doi.org/10.15405/ejsbs.316

Vasanthi, N., & Moorthy, S. (2024). CRISP-DM in customer analytics: A retail case study. *International Journal of Data Mining*, *15*(4), 456–472. https://doi.org/10.1504/IJDMB.2024.10056789

Vayena, E., & Blasimme, A. (2023). Health research with big data: Time for systemic oversight. *Journal of Law, Medicine & Ethics*, *51*(2), 265–283. https://doi.org/10.1017/jme.2023.62

Wang, S., Lee, J., & Kim, H. (2024). A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. *Scientific Reports*, *14*, 17964. https://doi.org/10.1038/s41598-024-68621-2

Wani, A., Priyanka, M., & Prasath, R. (2023a). Unleashing customer insights: Segmentation through machine learning. *Proceedings of the 2023 World Conference on Communication & Computing (WCONF)*, 1–6. https://doi.org/10.1109/wconf58270.2023.10235136

Wu, J., Shi, L., Lin, W.-P., & Tsai, S.-B. (2020). An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, *2020*, 8884227. https://doi.org/10.1155/2020/8884227

Yang, Y. (2009). Behavioral Pattern-Based Customer Segmentation. In *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 140–145). IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch023

Yoseph, F., Malim, N. H. A. H., Heikkilä, M., & Ylä-Kujala, A. (2020a). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems*, *38*(6), 6391–6404. https://doi.org/10.3233/JIFS-179698

Zahrani Putri, A., Afdal, M., Monalisa, S., & Permana, I. (2023). Penerapan Algoritma Fuzzy C-Means Pada Segmentasi Pelanggan B2B dengan Model LRFM. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(3), 1423. https://doi.org/10.30865/mib.v7i3.6150

6. ANEXO

Anexo 1: Código SQL para la extracción de la muestra de datos

El siguiente script SQL, ejecutado en Apache Impala, fue empleado para la extracción de la muestra de clientes y sus respectivas transacciones. El proceso inició con la selección aleatoria de 20.000 clientes activos a diciembre de 2025, aplicando filtros por canales de retail (códigos de sucursal específicos) y por el período de facturación comprendido entre 2023 y 2025. Las uniones con las tablas de productos y personas permitieron enriquecer los datos con la unidad de negocio y las variables demográficas. Los códigos de sucursal correspondientes a canales no retail han sido omitidos por razones de confidencialidad.

```
WITH muestra_clientes AS (
   SELECT codcliente
   FROM (
      SELECT
         crm.codcliente,
         SUM(crm.monto) AS venta
      FROM externo.potencialescrm crm
      JOIN dwh.farmapersona fp
         ON fp.codpersona = crm.codcliente
      WHERE crm.aniomes = '202512'
       AND fp.codtipopersona = 'CI'
       AND crm.estado = 'activo'
       AND fp.sexo IS NOT NULL
       AND crm.codcliente NOT IN (
            '6','0800886194001','R54321','6666666666001',
            '99999999999999','9999999999999','3','T54321',
            '9999999999','89','G54321','U54321','UB54321',
            '000001','G1234','RT54321')
      GROUP BY crm.codcliente
   ) t
   ORDER BY RAND()
   LIMIT 20000
),
ventas_base AS (
   SELECT
      fc.codcliente,
      CASE
         WHEN fp.unegocioc IN ('MEDICINAS','OTC')
            THEN 'MEDICINAS & OTC'
         WHEN fp.unegocioc IN ('INFANTIL','MINIMARKET')
            THEN 'INFANTIL & MINIMARKET'
```

```
            WHEN fp.unegocioc IN ('VITAMINAS MINERALES','SUPLEMENTOS')
               THEN 'VITAMINAS & SUPLEMENTOS'
            WHEN fp.unegocioc IN ('DERMOCOSMETICA','CUIDADO PERSONAL')
               THEN 'DERMOCOSMETICA & CUIDADO PERSONAL'
            WHEN fp.unegocioc IN ('BIENESTAR','BELLEZA')
               THEN 'BIENESTAR & BELLEZA'
            ELSE fp.unegocioc
         END AS unidad_negocio,
         TO_DATE(fc.fechafactura) AS fecha_transaccion,
         fc.valorventa
      FROM dwh.farmacomercial fc
      JOIN muestra_clientes mc
         ON mc.codcliente = fc.codcliente
      JOIN dwh.farmaproductos fp
         ON fp.codproducto = fc.codproducto
      WHERE fc.anio IN (2023, 2024, 2025)
       AND fc.codsucursal NOT IN ('004','009','018','019') -- Filtro para canales de retail
       AND fp.unegocioc IN (
            'MEDICINAS','OTC',
            'INFANTIL','MINIMARKET',
            'VITAMINAS MINERALES','SUPLEMENTOS',
            'DERMOCOSMETICA','CUIDADO PERSONAL',
            'BIENESTAR','BELLEZA')
)
SELECT
   v.codcliente AS cliente_id,
   fp.sexo,
   fp.edad,
   fp.instruccion,
   fp.canton,
   v.unidad_negocio,
   v.fecha_transaccion,
   SUM(v.valorventa) AS monto_total_usd
FROM ventas_base v
JOIN dwh.farmapersona fp on fp.codpersona = v.codcliente
GROUP BY
   v.codcliente,
   fp.sexo,
   fp.edad,
   fp.instruccion,
   fp.canton,
   v.unidad_negocio,
   v.fecha_transaccion;
```