Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

PROPUESTA DE MODELO DE MACHINE LEARNING EN LA PREDICCIÓN DEL COMPORTAMIENTO DEL CLIENTE PARA SU FIDELIZACIÓN EN UNA EMPRESA DE RETAIL

Línea de investigación:
Sistemas inteligentes, robótica, domótica

Tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas

Autor

Enriquez Maguiña, William Martin

ORCID: 0000-0003-1819-191X

Asesor

Rodríguez Rodriguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado

Flores Vidal, Higinio Exequiel

Coveñas Lalupu, José

Petrlik Azabache, Iván Carlo

Lima - Perú

2026

PROPUESTA DE MODELO DE MACHINE LEARNING EN LA PREDICCIÓN DEL COMPORTAMIENTO DEL CLIENTE PARA SU FIDELIZACIÓN EN UNA EMPRESA DE RETAIL

INFORME DE ORIGINALIDAD

10%	9%	1%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	www.coursehero.com Fuente de Internet	1%
2	reunir.unir.net Fuente de Internet	<1%
3	Submitted to Universidad Nacional Federico Villarreal Trabajo del estudiante	<1%
4	qdoc.tips Fuente de Internet	<1%
5	editorial.esic.edu Fuente de Internet	<1%
6	repositorio.unfv.edu.pe Fuente de Internet	<1%
7	Submitted to Universidad TecMilenio Trabajo del estudiante	<1%
8	www.researchgate.net Fuente de Internet	<1%
9	repositorio.ucv.edu.pe Fuente de Internet	<1%
10	www.slideshare.net Fuente de Internet	<1%
11	Submitted to Corporación Universitaria Iberoamericana	<1%

Universidad Nacional
Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

PROPUESTA DE MODELO DE MACHINE LEARNING EN LA PREDICCIÓN DEL COMPORTAMIENTO DEL CLIENTE PARA SU FIDELIZACIÓN EN UNA EMPRESA DE RETAIL

Línea de Investigación:
Sistemas inteligentes, robótica y domótica

Tesis para optar el Grado Académico de
Doctor en Ingeniería de Sistemas

Autor
Enriquez Maguiña, William Martin
ORCID: 0000-0003-1819-191X

Asesor
Rodríguez Rodriguez, Ciro
ORCID: 0000-0003-2112-1349

Jurado
Flores Vidal, Higinio Exequiel
Coveñas Lalupu, José
Petrlik Azabache, Iván Carlo

Lima – Perú
2026

DEDICATORIA

A mi amada familia, el verdadero motor de este logro.

A Liz, mi esposa, por tu amor incondicional,

tu apoyo constante y tu paciencia infinita.

A mis hijos, William, Lorena y Violeta,

inspiración permanente de superación

Y a ti, Orestes, mi querido hermano que ahora descansa en el cielo.

Aunque tu presencia física ya no me acompaña,

tu recuerdo, tus palabras y tu ejemplo siguen vivos en mi corazón.

Esta meta también es tuya.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El objetivo propuesto fue diseñar e implementar un modelo mejorado de segmentación de Machine Learning (ML) para deducir el comportamiento del cliente en su fidelización en una empresa de retail. El estudio fue aplicado y cuantitativo. La población de estuvo compuesta por todos los clientes de una empresa de retail, que han realizado compras en un periodo de 2 años. De estos, se seleccionaron como muestra a todo el dataset de registros de compra. La transformación de las variables originales sesgadas (Recencia, Frecuencia, Monto) logró distribuciones aproximadas a la normalidad, como se evidenció en los histogramas post-transformación. La prueba ANOVA mostró que los clústeres generados son estadísticamente diferentes entre sí en términos de las variables de comportamiento “Precio” y “cantidad. Se concluyó que el modelo mejorado de segmentación de ML, basado en la regresión logarítmica, optimiza de manera directa la identificación de las preferencias de los clientes en una empresa de retail.

Palabras clave: Machine Learning, análisis predictivo, comportamiento del cliente, sector retail.

ABSTRACT

The proposed objective was to design and implement an improved Machine Learning (ML) segmentation model to infer customer loyalty behavior in a retail company. The study was applied and quantitative. The population consisted of all customers of a retail company who had made purchases over a two-year period. Of these, to the entire dataset of purchase records for sample. The transformation of the original skewed variables (Recency, Frequency, Amount) achieved distributions close to normal, as evidenced in the post-transformation histograms. The ANOVA test showed that the generated clusters are statistically different from each other in terms of the behavioral variables "Price" and "Quantity." It was concluded that the improved ML segmentation model, based on logarithmic regression, directly optimizes the identification of customer preferences in a retail company.

Keywords: Machine Learning, predictive analytics, customer behavior, retail sector

I. INTRODUCCIÓN

La abundancia de datos sobre cómo se comportan los clientes está transformando los sistemas y enfoques utilizados para su análisis. Tradicionalmente, la segmentación o clusterización ha sido una práctica común para entender a los consumidores, pero la evolución tecnológica ha evidenciado que los enfoques tradicionales ya no son suficientes para captar las complejidades del comportamiento del cliente en la era digital. El conocimiento de los clientes permite a las empresas brindar servicios o productos que se ajusten a sus demandas, por lo que se ha convertido un indicador clave para el éxito empresarial.

La satisfacción de los clientes, medida luego de la compra, afecta la imagen empresarial e impulsa las recomendaciones a otros potenciales clientes. Así, es el principio de un proceso de promoción de gran valor para las organizaciones (Zada, 2022). Así, el propósito de las empresas es lograr que los clientes se fidelicen para realizar compras, emitir críticas positivas y recomendar productos. La experiencia del cliente no se mide solo por la satisfacción, sino también por el valor que se percibe, que muestra cómo el cliente evalúa lo que espera recibir frente a lo que realmente recibe (McDougall & Levesque, 2000). Sin embargo, surge otro desafío crucial: ¿cómo pueden las empresas determinar qué productos o servicios ofrecen, ¿cuándo hacerlo y a través de qué canal de comunicación? La respuesta a estos. Las preguntas son importantes para mejorar la calidad de la experiencia del cliente, aumentar el valor percibido y, en consecuencia, fomentar una mayor lealtad hacia la empresa o marca.

La Inteligencia Artificial (IA), y más específicamente el Aprendizaje Automático (AA), se está convirtiendo en una herramienta clave para la investigación de mercados.

. Los avances en el campo han transformado las áreas de interés, pasando de la gestión de relaciones con clientes en 2011, a la minería de datos en 2015, y finalmente, al aprendizaje automático y Big Data después de 2017, todos ellos impulsados por la creciente adopción de tecnologías de IA (Verma et al., 2021). Las empresas recogen muchos tipos de información

sobre sus clientes, como sus opiniones en redes sociales, sus patrones de retención, los riesgos y la gestión de marcas. Las tecnologías de IA analizan estos datos para predecir las tendencias del mercado y el comportamiento de cada consumidor en cuanto a las compras.

El estudio evaluará el impacto de la inteligencia artificial, específicamente el aprendizaje automático a través de la propuesta de un modelo mejorado, en la predicción del comportamiento del cliente para mejorar la experiencia de compra, aumentar el valor percibido y, en última instancia, de la lealtad del cliente en el sector minorista.

También ayuda a crear estrategias personalizadas basadas en datos que ayudarán a dichas empresas en adaptar sus productos y servicios para satisfacer las necesidades de sus clientes, lo que los hará más felices y leales (Kietzmann et al., 2018)

1.1. Planteamiento del problema

Sus preferencias pueden tener efectos significativos en su vida. Gracias a la digitalización, las pequeñas empresas pueden recopilar enormes cantidades de información sobre sus clientes. Esto les brinda una gran oportunidad para comprender y predecir sus intereses. Si no ofrece servicios personalizados con sus productos, puede perjudicar la experiencia de sus clientes y reducir su fidelidad. Por ejemplo, preguntar a los clientes sobre su talla, color, cantidad o características preferidas (qué categoría de producto o servicio desean) puede ayudar a generar sugerencias personalizadas de bienes, servicios o marcas que se ajusten mejor a sus necesidades. La personalización mejora considerablemente la experiencia de compra y reduce la cantidad de errores que cometen las personas durante el proceso, como elegir la talla o el color incorrectos al comprar en línea, algo frecuente en la industria de la moda (Uzir et al., 2021). El desafío es utilizar estos datos correctamente para ofrecer a los clientes sugerencias que los hagan más felices y los hagan más fieles a su marca.

Una de las mayores dificultades de trabajar en el sector servicios es gestionar la experiencia del cliente con la empresa durante el proceso de compra. Las percepciones de los

clientes son subjetivas y cambian en función de sus expectativas y experiencias previas. Por ello, es importante conocer su opinión sobre la experiencia que tuvieron para que los productos y servicios sean siempre mejores. Al no organizar y analizar eficazmente los datos, las empresas minoritarias tienden a minimizar la importancia de estos indicadores clave, lo que les impide realizar mejoras significativas en su oferta (Aldunate et al., 2022). Los enfoques de segmentación no consideran la complejidad de las interacciones modernas, y la falta de digitalización dificulta ofrecer experiencias personalizadas y relevantes que fomenten la fidelización a largo plazo (Das & Nayak, 2022; Griva et al., 2022).

1.2. Descripción del problema

La importancia de los datos de los clientes ha aumentado significativamente en los últimos años debido a la integración de varios canales digitales. Cuando las personas utilizan diferentes tecnologías, como redes sociales, teléfonos móviles, sensores y otros dispositivos conectados, dejan rastros de datos. cuando usan diferentes tecnologías, como redes sociales, teléfonos celulares, sensores y otros dispositivos conectados. Con la llegada de la era de la información y la digitalización ha intensificado significativamente la competencia en el mercado, y las empresas ahora priorizan a los clientes sobre los productos. Se ha intensificado y las empresas ahora priorizan a los clientes sobre los productos. El mercado ha crecido mucho, y ahora las empresas se centran más en los clientes que en sus productos. En este escenario, la búsqueda de cuota de mercado y la maximización de beneficios se han convertido en objetivos estratégicos imperativos para el desarrollo de cualquier empresa (Zhao et al., 2023).

La transformación ha hecho que los clientes sean más exigentes y su comportamiento sea más inestable, ya que hay muchas opciones de productos, servicios, canales y medios de compra para satisfacer sus deseos y necesidades. Los cambios en el comportamiento de los consumidores pueden ser inesperados, como ha demostrado la pandemia reciente. Por eso, las empresas deben adaptarse a estos cambios. Los clientes no solo cambian su comportamiento

con el tiempo, sino que también obligan a las empresas a ser más flexibles y a repensar cómo venden sus productos.

El comportamiento impredecible es un desafío para segmentar clientes y gestionar relaciones con los clientes. Los modelos tradicionales no captan la complejidad y volatilidad de las preferencias de los consumidores, lo que dificulta fidelizar a los clientes. Así, la incapacidad para ajustar estrategias personalizadas puede reducir la participación en el mercado y disminuir el posicionamiento de marca.

Según Kassem et al. (2020), los clientes leales brindan ingresos continuos a la empresa. En un entorno competitivo, se requiere retener a los clientes para mantenerse en el mercado, ya que genera una mayor rentabilidad a largo plazo y permite consolidar el posicionamiento de marca. Ignorar esta necesidad puede traer pérdidas en términos de cuota de mercado, marca y competitividad en un contexto que requieren estrategias que respondan a las nuevas exigencias.

En el comercio minorista, la fidelización del cliente es una prioridad estratégica, pero muchas empresas implementan tecnologías procesen grandes volúmenes de datos para predecir comportamientos de clientes. Esto impide ofrecer experiencias personalizadas, lo que genera pérdida de ingresos y disminuir la competitividad en el mercado (Al-Araj et al., 2022; Griva et al., 2022). Los enfoques tradicionales de segmentación ya no son suficientes para anticipar el comportamiento del consumidor moderno, lo que requiere adoptar ML para fidelizar clientes (Das & Nayak, 2022). Uno de estos modelos es el RFM, una metodología para segmentar clientes por similitudes en su compra (Glotzer & Simla, 2024).

1.3. Formulación del problema

1.3.1. Problema general

¿De qué manera la propuesta de un modelo mejorado de segmentación de Machine Learning (ML) describe el comportamiento del cliente para su fidelización en una empresa de retail?

1.3.2. Problemas específicos

¿De qué manera el uso de nuevos parámetros en los requerimientos del modelo mejorado de segmentación de ML permitirá identificar las preferencias del cliente en una empresa de retail?

¿De qué manera el modelo mejorado de segmentación de ML podrá optimizar la identificación de las preferencias de los clientes en una empresa de retail?

¿De qué manera la comparación de los modelos de segmentación de ML según su precisión podrá mejorar la efectividad en la preferencia del cliente de una empresa retail?

1.4. Antecedentes

El concepto de segmentación de clientes no es nuevo en el marketing y la gestión. Esta técnica permite agrupar a los clientes en categorías que comparten características similares y exhiben patrones de comportamiento semejantes. Uno de los modelos más conocidos en este contexto es el modelo RFM propuesto inicialmente por Hughes (1994). Permite identificar clientes valiosos a partir de grandes conjuntos de datos transaccionales, proporcionando una metodología eficaz para gestionar los recursos de marketing y diseñar estrategias personalizadas. En el sector minorista, la segmentación de clientes es vital para asignar recursos como estrategias promocionales, políticas de precios y programas de fidelización, lo que a su fortalece la relación con los clientes y aumentar su lealtad. Con el advenimiento tecnológico y la explosión de Big Data, las empresas enfrentan un entorno corporativo altamente competitivo. Según el informe de Global Powers of Retailing 2017, las futuras tendencias en el comercio minorista se centran en tres paradigmas clave: redes sociales, datos y experiencias. En este contexto, las aplicaciones de inteligencia artificial (IA), al igual que los datos, juegan un papel crucial en casi todas las industrias. La gestión autónoma de Big Data permite a las empresas mejorar significativamente su velocidad, precisión y eficiencia en la toma de decisiones estratégicas (Al-Araj et al., 2022).

Para evaluar la efectividad de estas estrategias, se han utilizado modelos como el RFM, que sigue siendo una metodología clave en la segmentación del comportamiento de los clientes. Este modelo ha evolucionado a lo largo de las décadas y sigue siendo ampliamente utilizado en la industria para caracterizar el comportamiento de los clientes. Según Chang & Tsai (2011), el modelo RFM sigue siendo un paradigma predominante al segmentar el comportamiento y en la planificación de estrategias de marketing.

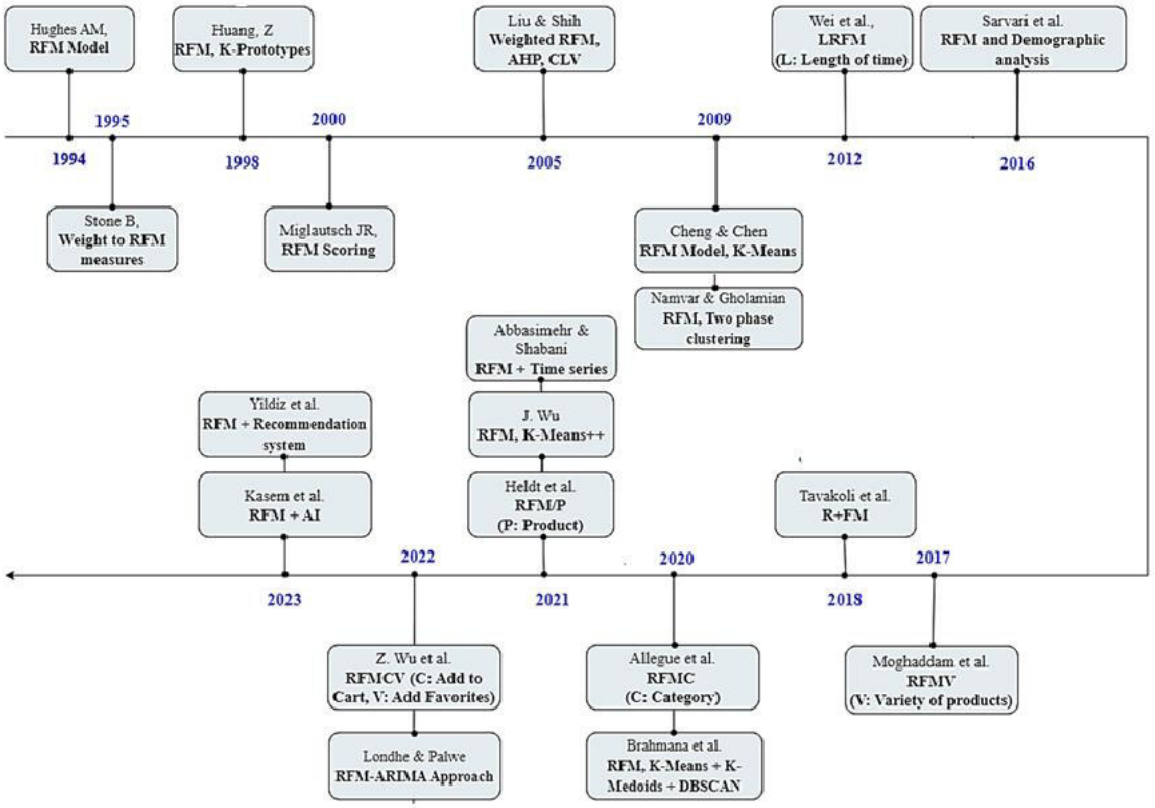
El análisis RFM se basa en el comportamiento histórico y permite observar de forma panorámica la dinámica de la empresa, lo que incrementa la ventaja. Además, este análisis considera por lo menos tres puntos clave:

- Recencia: Días que han transcurrido desde que se realizó la última compra.
- Frecuencia: Compras realizadas en un periodo específico.
- Monto: Valor de las compras en el periodo seleccionado.

Si se combinan estos tres parámetros, se obtiene un ranking RFM que crea segmentos de clientes según su valor para la marca, y así poder desarrollar acciones personalizadas de marketing. En la siguiente figura se puede apreciar las mejoras realizadas al modelo RFM.

Figura 1

Resumen de las mejoras del modelo RFM



Nota. Tomado de Ho et al. (2023)

En el sector retail cada cliente es un mundo con una diversidad de variantes, por este motivo, es fundamental hacer campañas personalizadas en base a segmentos de clientes. Estas acciones tienen múltiples ventajas entre las que figuran:

- Optimizar el impacto del marketing
- Conocer la tasa de abandono
- Incrementar la interacción con los clientes
- Retener a los clientes
- Incrementar las ventas

Por tanto, si el principal activo de un negocio retail son los datos de sus clientes, tiene sentido aprovecharlos para tomas decisiones estratégicas que permitan aumentar el engagement del cliente y, en consecuencia, las conversiones.

Los avances recientes en IA y ML transforman la interacción de las empresas con sus clientes. Según Aguiar-Costa et al. (2022), la IA posee un potencial transformador en la forma en la recopilación y análisis del comportamiento de los clientes. Las técnicas de ML permiten a comprender las preferencias y patrones de compra de los clientes, facilitando la personalización de los servicios y productos. Su incorporación estratégica puede generar beneficios significativos, mejorar la satisfacción y potenciar la fidelización.

En las últimas décadas, la aplicación de tecnologías de IA y ML ha cambiado la fidelización de clientes. Tradicionalmente, las estrategias de retención se basaban en métodos descriptivos y retrospectivos que analizaban el comportamiento del cliente de forma limitada, sin capacidad predictiva. Sin embargo, el desarrollo de algoritmos avanzados de ML ha permitido a las empresas no solo segmentar a sus clientes, sino también predecir sus comportamientos futuros, mejorando la personalización (Sun et al., 2019). Además, el uso de machine learning en la fidelización facilita la creación de experiencias de usuario más personalizadas, adaptando las recomendaciones de productos y servicios según las preferencias individuales de cada cliente (Kietzmann et al., 2018).

El impacto de estas tecnologías no solo ha sido notable en términos de mejorar la retención de clientes, sino que también ha permitido a las empresas optimizar sus recursos, al enfocarse en los clientes con mayor potencial de valor a largo plazo. Así, el ML se ha consolidado como una herramienta esencial para maximizar el Customer Lifetime Value (CLV) y fortalecer la competitividad en un entorno comercial cada vez más complejo y competitivo

1.5. Justificación de la investigación

1.5.1. Justificación económica

El uso de modelos de ML en el sector retail presenta beneficios económicos que optimizan la gestión de recursos y mejora la efectividad de las estrategias de marketing. Los algoritmos de ML permiten analizar grandes volúmenes de datos transaccionales y de comportamiento para identificar patrones complejos que, de otro modo, serían difíciles de detectar con métodos tradicionales. Esta capacidad predictiva facilita la segmentación precisa de los clientes, permitiendo a las empresas enfocar sus recursos en aquellos segmentos con mayor potencial de retorno económico (Al-Araj et al., 2022; Griva et al., 2022).

Empresas que han implementado modelos de ML para la fidelización de clientes han reportado una reducción significativa en los costos de adquisición de nuevos clientes, ya que retener a clientes actuales es, por naturaleza, más económico (Das & Nayak, 2022). Al aprovechar estas tecnologías, las empresas pueden asignar presupuestos de marketing de forma más eficiente, maximizando el retorno sobre la inversión (ROI) y reduciendo desperdicios en campañas de baja efectividad (Kietzmann et al., 2018).

El impacto económico de esta investigación es evidente en varios niveles. En primer lugar, al emplear modelos de ML para realizar una segmentación precisa de los clientes, las empresas de retail pueden identificar con mayor exactitud los segmentos de clientes más rentables. Esto permitirá una asignación más eficiente de los presupuestos de marketing, reduciendo costos innecesarios en campañas generales y optimizando los esfuerzos hacia campañas personalizadas que aumenten la lealtad de los clientes y mejoren el valor percibido por ellos. En segundo lugar, el uso de IA/ML en la personalización de la experiencia del cliente se traduce directamente en un aumento de las ventas y de sus ingresos. La capacidad de predecir con mayor precisión las preferencias de los consumidores no solo mejora la satisfacción del

cliente, sino que también fomenta una mayor recurrencia de compra, lo que incrementa el CLV y, en última instancia, impulsa los beneficios económicos para las empresas.

1.5.2. Justificación social

Desde una perspectiva social, esta personalización no solo favorece a los consumidores, sino que también promueve la inclusión. Mediante el análisis avanzado de datos, las empresas pueden comprender mejor las necesidades de diferentes grupos demográficos, incluidos aquellos con requerimientos específicos, como productos adaptados para personas con discapacidades o consumidores con preferencias culturales distintas. Así, las estrategias basadas en ML facilitan la creación de experiencias más inclusivas y equitativas, fomentando un acceso más amplio a productos y servicios para toda la población (Sun et al., 2019).

Además, la implementación de modelos de ML permite a las empresas operar de manera más sostenible. La optimización basada en datos contribuye a reducir el desperdicio, ya que permite predecir la demanda con mayor precisión y ajustar la producción y el inventario en consecuencia. De este modo, se minimiza el impacto ambiental al evitar la sobreproducción y el uso excesivo de recursos naturales (Al-Araj et al., 2022). En un mundo donde la sostenibilidad se ha convertido en una prioridad, el uso de ML para optimizar las operaciones comerciales contribuye a un desarrollo más equilibrado y responsable, alineando las prácticas empresariales con los objetivos de sostenibilidad social.

Finalmente, la adopción de ML puede reforzar prácticas éticas y responsables en el manejo de datos, asegurando que la información personal de los clientes se utilice de forma transparente y segura. La capacidad de los algoritmos para analizar patrones sin comprometer la privacidad individual fomenta la confianza entre consumidores y empresas, consolidando relaciones a largo plazo basadas en la transparencia y la responsabilidad (Griva et al., 2022).

1.5.3. Justificación científica

Desde una perspectiva científica, esta investigación pretende llenar un vacío en la literatura existente del uso de modelos de ML en la investigación del sector retail. A diferencia de los métodos estadísticos tradicionales, que se basan en análisis descriptivos y retrospectivos, los algoritmos de ML pueden identificar patrones complejos y no lineales en grandes volúmenes de datos. Esto permite a los investigadores y profesionales del marketing predecir comportamientos futuros, segmentar a los clientes de forma más precisa y diseñar estrategias de fidelización que se adapten dinámicamente a las necesidades de cada cliente (Griva et al., 2022; Kietzmann et al., 2018).

Esta investigación buscará profundizar en cómo la implementación de modelos de IA/ML para la predicción del comportamiento del cliente puede no solo incrementar su satisfacción, sino también mejorar su retención a largo plazo. Además, se analizará el impacto de la IA/ML en la capacidad de las empresas de retail para ofrecer experiencias personalizadas, lo que, según la literatura reciente, es un factor clave para asegurar la lealtad en un entorno competitivo (Al-Araj et al., 2022; Aldunate et al., 2022).

En la siguiente Tabla se muestran algunos artículos que pueden justificar nuestra investigación.

Tabla 1

Artículos que justifican la investigación

Articulo	Aporte	Tipo
Subali et al. (2020)	Modelo que valida como la calidad de la experiencia obtenida por el cliente impacta en la lealtad del cliente.	No Competidor
Oncioiu et al. (2021)	Modelo para medir la evolución de la relación de las redes sociales y la atención al cliente. Considera nuevas variables en gestión de la comunicación del consumidor en línea para el éxito de la relación.	No Competidor

Zanchett & Paladini (2019)	Propone una nueva modalidad de programa de fidelización e identifica qué enfoques de lealtad se toman como objetivo para cada tipo de programa (lealtad de marca a la tienda o al programa).	Competidor
Mensouri et al. (2022)	Propone un nuevo modelo de segmentación al introducir una nueva variable "T" en el modelo RFM para crear un modelo más completo, a saber, RFMT, para analizar las secuencias de compra de los consumidores durante un largo período.	Competidor
Calvo-Porral & Lévy-Mangin (2019)	Desarrollar una nueva segmentación de clientes en el contexto de crisis económica, a partir de la definición de los factores de atracción de los centros comerciales (ocho factores).	No Competidor

1.6. Limitaciones de la investigación

La presente investigación enfrenta diversas limitaciones que pueden influir en los resultados y la aplicabilidad de los hallazgos. Estas limitaciones se dividen en dos categorías: metodológicas y contextuales.

Limitaciones metodológicas:

Muestra limitada en tamaño y diversidad: El tamaño limitado de la muestra también podría afectar la generalización de los resultados. Según Hernández et al. (2018), una muestra más amplia y diversa puede mejorar la validez externa de los resultados.

Dependencia en la calidad de los datos transaccionales: Los resultados de la investigación dependen en gran medida de la calidad de los datos recolectados. Datos incompletos, incorrectos o sesgados pueden influir en los resultados obtenidos del modelo de Machine Learning. Es crucial realizar una limpieza de datos exhaustiva para mitigar este riesgo. Tal como señalan V. Kumar & Shah (2004), la calidad de los datos es un factor crítico en el éxito de cualquier modelo predictivo basado en Machine Learning.

Limitaciones contextuales:

Acceso a datos privados: Uno de los mayores desafíos será obtener acceso a los datos transaccionales, ya que esto depende de la disposición de la empresa a compartir información

privada y confidencial. La Ley de Protección de Datos Personales en Perú (Ley N° 29733) impone restricciones estrictas sobre el manejo y uso de información personal, lo que podría limitar la cantidad de datos accesibles y la profundidad del análisis.

Temporalidad y cambios en el comportamiento del consumidor: El comportamiento del consumidor es dinámico y puede cambiar debido a factores impredecibles como crisis económicas o pandemias. Los resultados de la investigación pueden verse limitados por la temporalidad de los datos recopilados, lo que significa que los hallazgos actuales podrían no ser válidos en el futuro. Al-Araj et al. (2022) mencionan que los modelos de predicción basados en Machine Learning requieren actualizaciones constantes para mantenerse relevantes frente a cambios drásticos en los patrones de comportamiento.

Limitaciones tecnológicas:

Aunque la tecnología de Machine Learning permite mejorar la segmentación y la personalización de la experiencia del cliente, la implementación efectiva de estas tecnologías requiere una infraestructura robusta y conocimientos técnicos avanzados. Las limitaciones tecnológicas en la empresa de retail pueden dificultar la integración de las soluciones propuestas. Kietzmann et al. (2018) destacan que la infraestructura tecnológica es un componente crucial para la adopción exitosa de IA en las empresas.

Conscientes de estas limitaciones, el proyecto de investigación se enfocará en superar estos desafíos para obtener resultados válidos y relevantes que contribuyan al avance del conocimiento en el campo de la segmentación de clientes y la aplicación de modelos de ML en la fidelización de clientes.

1.7. Objetivos

1.7.1. Objetivo general

Diseñar e implementar un modelo mejorado de segmentación de Machine Learning para deducir el comportamiento del cliente en su fidelización en una empresa de retail.

1.7.2. Objetivos específicos

Utilizar nuevos parámetros en los requerimientos del modelo mejorado de segmentación de ML para identificar las preferencias del cliente en una empresa de retail.

Demostrar el modelo mejorado de segmentación de ML para optimizar la identificación de las preferencias de los clientes en una empresa de retail.

Comparar la precisión de los modelos de segmentación de ML para mejorar la efectividad en la preferencia del cliente de una empresa retail.

1.8. Hipótesis

1.8.1. Hipótesis general

El modelo mejorado de segmentación propuesto de Machine Learning predice el comportamiento del cliente para su fidelización en una empresa de retail.

1.8.2. Hipótesis específicas

El uso de nuevos parámetros en los requerimientos del modelo mejorado de ML segmenta con mayor precisión las preferencias del cliente.

El modelo mejorado de segmentación de ML optimiza la identificación de las preferencias de los clientes en una empresa de retail.

Los modelos de segmentación de ML permitieron comparar de manera eficiente la preferencia del cliente en una empresa de retail.

II. MARCO TEÓRICO

2.1. Marco conceptual

2.1.1. Inteligencia Artificial

Machine Learning (ML): Rama de la IA que permite a los sistemas aprender y mejorar mediante experiencias sin programación explícita. Su aprendizaje se basa en algoritmos que procesan datos, identifican patrones y predicen. Según Kühl et al. (2022) el ML utiliza datos para entrenar modelos capaces de realizar tareas específicas. La clave de ML radica en su capacidad para detectar relaciones ocultas en los datos y automatizar decisiones en tiempo real, lo que resulta crucial en un entorno de retail altamente competitivo y dinámico.

Tipos de Machine Learning:

- Supervisado: El modelo es entrenado a través de datos etiquetados. Ejemplos incluyen la regresión lineal y los árboles de decisión.
- No supervisado: Identifica patrones en datos no etiquetados. Ejemplos incluyen el clustering o la segmentación.
- Por refuerzo: El algoritmo aprende a tomar decisiones al interactuar con su entorno, lo que optimiza sus acciones.

2.1.2. Inteligencia Artificial (IA) en el Retail

Tiene el potencial de revolucionar las interacciones con los clientes, ya sea a través de chatbots de atención u otros sistemas de IA, y ofrecer mejores experiencias que se anticipen a las necesidades del consumidor. Aguiar et al. (2022) destacan que las tecnologías de IA permiten a las empresas obtener un conocimiento más profundo de las preferencias del cliente, mejorando la satisfacción y fidelización.

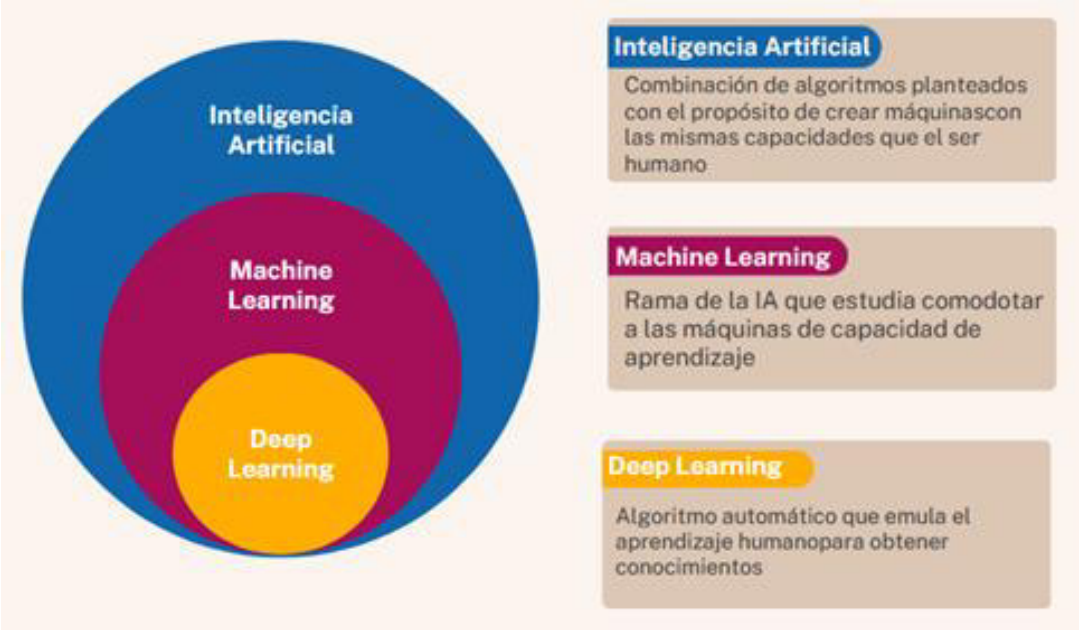
Aplicaciones de IA en Retail

- Recomendación a través del comportamiento pasado.

- Predicción de futuras compras o abandono del cliente.
- Automatización de campañas de marketing en tiempo real.

Figura 2

Esquema de IA



Nota. Tomado de Junta de Andalucía (2020)

2.1.3. Fidelización

Fidelización de Clientes: se refiere al compromiso continuo de un cliente con una marca, demostrado por la repetición de compras, recomendaciones a otros y la preferencia constante por los productos o servicios de una empresa. Dick & Basu (1994) definen la fidelización como la relación entre la actitud de los clientes hacia una oferta y su comportamiento de compra. Las empresas, especialmente en el sector retail, buscan retener clientes debido a que el costo de adquisición de nuevos clientes es significativamente mayor que el costo de retener a los actuales. Las estrategias de fidelización son clave para aumentar el valor de vida del cliente (CLV), optimizar el retorno sobre la inversión (ROI) y mejorar la rentabilidad.

Métricas de Fidelización

Net Promoter Score (NPS): Mide la disposición de los clientes a recomendar una empresa.

Customer Satisfaction Score (CSAT): Evalúa el nivel de satisfacción de los clientes con productos o servicios.

CLV: Calcula el valor total de un cliente a lo largo de su relación con la empresa.

Segmentación de Clientes: implica agrupar a los consumidores en categorías con características similares para optimizar las estrategias de marketing. Los segmentos tradicionales se basan en datos demográficos y psicográficos; sin embargo, los avances en la analítica de datos han permitido una segmentación más precisa basada en el comportamiento del cliente, sus preferencias y su interacción con la marca.

2.1.4. Algoritmos de Clasificación y Clustering

Clasificación: Algoritmos como la regresión logística, árboles de decisión y máquinas de soporte vectorial (SVM) se utilizan para predecir la probabilidad de deserción o Churn de un cliente. Estos modelos ayudan a identificar patrones específicos que llevan a un cliente a abandonar la marca, permitiendo a las empresas implementar intervenciones a tiempo.

Clustering: Técnicas como K-means, clustering jerárquico y modelos basados en probabilidad agrupan a los clientes en segmentos homogéneos según sus características RFM y CLV. Esto facilita la creación de campañas de fidelización que se adaptan mejor a las características de cada grupo, maximizando la efectividad y reduciendo costos en marketing (Hiziroglu & Sengul, 2012; McCarty & Hastak, 2007).

Redes Neuronales: Estas redes son eficaces en la predicción de churn, ya que pueden analizar patrones complejos y no lineales en el comportamiento del cliente. Su capacidad para "aprender" permite ajustar continuamente las estrategias de retención en función de datos en tiempo real.

Árboles de Decisión: Los árboles de decisión ofrecen una representación visual de las decisiones que llevan a un cliente a realizar una compra o abandonarla. Estos modelos son útiles para predecir la efectividad de campañas de retención y ayudan a identificar características clave en el perfil del cliente que impactan en su comportamiento de compra (Rosset, 2003).

2.1.5. Personalización mediante ML

Definición: La personalización se refiere a la adaptación de productos, servicios y estrategias de marketing a las preferencias individuales de cada cliente. En el contexto de fidelización, la personalización permite crear una experiencia única para cada usuario.

Aplicación en Retención de Clientes: El análisis de datos mediante ML permite entender las preferencias individuales de cada cliente. Por ejemplo, sistemas de recomendación basados en ML analizan el historial de compra y las interacciones del cliente para ofrecer productos y promociones relevantes, aumentando la probabilidad de recompra.

Técnicas de Machine Learning para la Personalización:

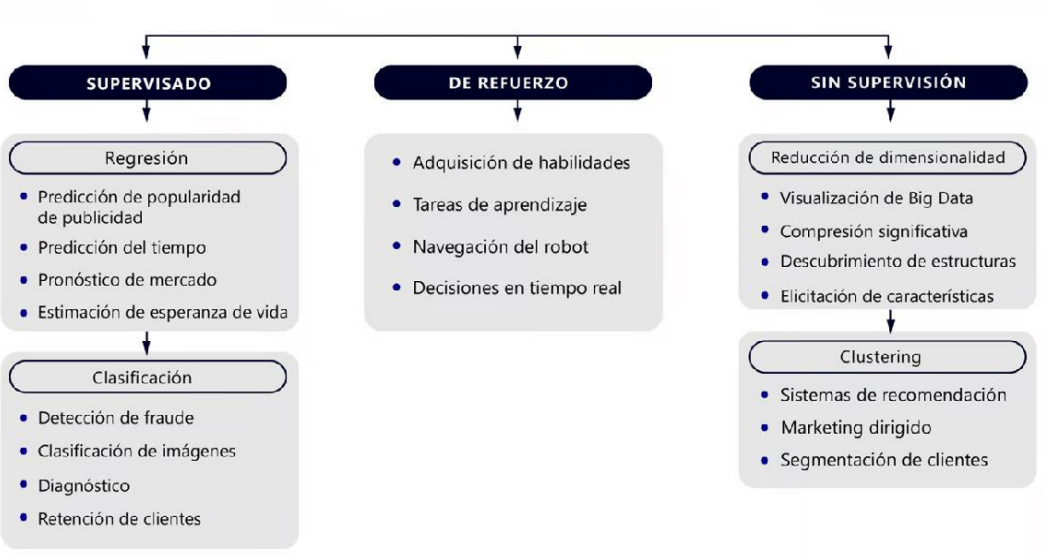
Análisis Predictivo: Los algoritmos predictivos, como las redes neuronales y los modelos de regresión, analizan el historial de compras y las respuestas a campañas anteriores para predecir el comportamiento futuro del cliente, adaptando la estrategia de retención.

Modelos Secuenciales (RFMP-growth): Técnicas como RFMP-growth permiten identificar patrones de compra repetitivos y anticipar futuras interacciones del cliente, facilitando la implementación de estrategias de retención proactivas basadas en el comportamiento (Cheng & Chen, 2009; Hu & Yeh, 2014).

En la siguiente figura mostraremos un resumen de los algoritmos utilizado en ML

Figura 3

Algoritmos de ML



Nota. Tomado de Decide (2022)

2.2. Marco Filosófico

Epistemología: Conocimiento Basado en Datos

El enfoque epistemológico de esta investigación se fundamenta en el empirismo, donde el conocimiento proviene de los datos obtenidos a través de la experiencia y las interacciones con los clientes. En este contexto, el uso de Machine Learning permite analizar grandes volúmenes de datos para extraer patrones y generar predicciones precisas sobre el comportamiento del cliente. De acuerdo con el empirismo lógico, las herramientas de Machine Learning y Big Data son metodologías esenciales para validar hipótesis y obtener conocimiento objetivo a partir de la observación y medición de datos.

El conocimiento sobre el comportamiento no se adquiere de forma intuitiva o abstracta, sino que se deriva del análisis sistemático de los datos transaccionales y las interacciones con el cliente. Esta forma de conocimiento es clave para comprender las tendencias de compra, los factores que influyen en la lealtad y cómo adaptar las estrategias de fidelización.

Ontología: Realismo de los Datos y Modelos Predictivos

Desde un enfoque ontológico, esta investigación adopta un realismo científico, donde los datos y patrones que emergen de las transacciones y comportamientos del cliente representan realidades objetivas y tangibles. Los modelos predictivos de Machine Learning permiten capturar estas realidades de manera precisa, identificando regularidades en los comportamientos del cliente que pueden predecir con alto grado de certeza su probabilidad de fidelización o abandono.

El realismo ontológico en el contexto del retail y Machine Learning asume que las predicciones generadas a partir de los datos tienen una correspondencia directa con el mundo real, permitiendo a las empresas tomar decisiones basadas en hechos observables. Esto es crucial en el sector retail, donde las decisiones rápidas y basadas en datos son fundamentales para mantener la competitividad y la satisfacción del cliente.

Filosofía de la Tecnología: Impacto Ético y Social de la IA/ML en el Retail

El impacto de la inteligencia artificial (IA) en la sociedad, y más específicamente en el retail, plantea preguntas filosóficas sobre el equilibrio entre la automatización y el trabajo humano, así como la privacidad y el uso ético de los datos de los clientes. La IA/ML, como herramienta tecnológica avanzada, está transformando la forma en que las empresas interactúan con los consumidores y gestionan sus relaciones con ellos. Sin embargo, la implementación de estas tecnologías conlleva dilemas éticos.

Desde la ética tecnológica, se debe considerar cómo la automatización de procesos a través de IA afecta el empleo, especialmente en tareas relacionadas con el servicio al cliente. Además, la recolección masiva de datos plantea preocupaciones sobre la privacidad y el manejo responsable de la información de los clientes. Según las teorías éticas como el utilitarismo, la implementación de IA/ML debe maximizar los beneficios tanto para las empresas como para los consumidores, reduciendo cualquier daño potencial, como el uso indebido de datos personales.

Determinismo Tecnológico vs. Humanismo Tecnológico

El determinismo tecnológico sostiene que la tecnología, en este caso la IA y el Machine Learning, actúan como fuerzas impulsoras que determinan el futuro de la sociedad y las empresas, moldeando la forma en que los clientes interactúan con el retail. Este enfoque sugiere que la adopción de estas tecnologías no solo transforma la experiencia del cliente, sino que también redefine la estructura y las operaciones de las empresas.

Por otro lado, el humanismo tecnológico defiende que, aunque la tecnología es una herramienta poderosa, es el ser humano quien debe guiar y regular su uso. En el contexto de esta investigación, esto significa que las decisiones basadas en los resultados de Machine Learning no deben reemplazar la intuición y el juicio humano, sino complementarlos. La personalización y fidelización del cliente, aunque impulsadas por IA, aún deben estar alineadas con valores y principios que respeten la individualidad y privacidad del consumidor.

Filosofía del Marketing Relacional

La filosofía del marketing relacional, que subraya la importancia de construir relaciones duraderas con los clientes, es fundamental para este estudio. El uso de modelos basados en ML en la predicción del comportamiento del cliente se alinea con este enfoque, ya que estas tecnologías permiten personalizar las interacciones y adaptar las ofertas a las necesidades y preferencias de los clientes. En este sentido, la tecnología no solo se utiliza para atraer a nuevos clientes, sino para mantener y fortalecer las relaciones a largo plazo, mejorando la lealtad y la satisfacción del cliente.

El enfoque relacional sostiene que la fidelización de clientes es más rentable a largo plazo que la adquisición de nuevos clientes, y que las interacciones continuas y personalizadas son clave para lograr este objetivo. Las aplicaciones de IA permiten identificar los momentos adecuados para interactuar con los clientes y ofrecerles valor adicional, consolidando así su relación con la empresa.

2.3. Estado del Arte

2.3.1. ML en la fidelización de clientes

En los últimos años, el uso de Machine Learning (ML) ha experimentado un auge significativo en el sector retail, transformando la manera en que las empresas interactúan con sus clientes y personalizan sus ofertas. Este crecimiento se debe principalmente a la capacidad de ML para analizar grandes volúmenes de datos, identificar patrones de comportamiento y proporcionar predicciones precisas que optimizan la fidelización del cliente (Aldunate et al., 2022). Las empresas buscan aprovechar estas tecnologías no solo para mejorar la experiencia del cliente, sino también para anticipar comportamientos futuros y desarrollar estrategias de retención más efectivas (Seymen et al., 2023).

El sector del retail ha sido uno de los más afectados por la necesidad de innovar en la relación con sus clientes, utilizando el análisis de datos para maximizar la satisfacción y la lealtad. Esta creciente adopción de ML se debe a su capacidad de generar recomendaciones personalizadas en tiempo real, reducir la deserción y aumentar los ingresos mediante la retención a largo plazo (Capuano et al., 2021). La segmentación de clientes es esencial en el desarrollo de estrategias de fidelización. Técnicas de clustering como K-means y métodos de clasificación permiten agrupar a los consumidores en segmentos de acuerdo con sus patrones de compra, facilitando así la optimización de recursos para retención y fidelización. Estudios como el de Joung & Kim (2023) destacan que estos enfoques basados en Machine Learning interpretan reseñas y opiniones, lo cual ayuda a las empresas a comprender mejor las necesidades no satisfechas de sus clientes.

Análisis de Sentimientos y Fidelización: Además de los modelos transaccionales, el análisis de sentimientos se utiliza para evaluar las opiniones y comentarios de los clientes en redes sociales. Herramientas de Procesamiento de Lenguaje Natural (NLP) permiten a las

empresas identificar factores que impulsan la fidelización y ajustar sus estrategias de acuerdo con el feedback del cliente, como muestran los estudios de Capuano et al. (2021).

Personalización de la Experiencia del Cliente: Los sistemas de recomendación, que utilizan algoritmos de aprendizaje supervisado y no supervisado, se aplican en tiempo real para adaptar las ofertas a las preferencias y comportamientos históricos de los clientes. Según Aguiar-Costa et al. (2022), la integración de IA y ML en retail no solo permite mejorar la satisfacción del cliente, sino que también maximiza el CLV mediante una experiencia personalizada que fomenta la lealtad a largo plazo.

Modelos Predictivos de Deserción de Clientes: Uno de los principales usos de ML en la fidelización de clientes es la predicción de la deserción o churn, la cual permite a las empresas identificar a aquellos clientes que tienen más probabilidades de abandonar el servicio o la marca. En este contexto, el artículo de Seymen et al. (2023) describe cómo las redes neuronales recurrentes se utilizan para analizar datos transaccionales y predecir la deserción del cliente, permitiendo a las empresas de retail actuar de manera proactiva para reducir la pérdida de clientes. Este enfoque es particularmente efectivo en el análisis de datos temporales, ya que captura patrones secuenciales de comportamiento que otros modelos no pueden detectar con tanta precisión (Seymen et al., 2023).

2.3.2. Segmentación de clientes basada en ML

La segmentación de clientes es otra área clave donde el ML ha tenido un impacto significativo. Tradicionalmente, la segmentación de clientes en retail se basaba en datos demográficos y geográficos. Sin embargo, el modelo RFM (Recency, Frequency, Monetary**)** ha sido ampliamente utilizado para clasificar a los clientes según su valor a través de sus patrones de compra. El estudio de Paul & Ramanan (2019) destaca la aplicación de RFM junto con el análisis del CLV para optimizar las campañas de marketing dirigidas a los clientes más

valiosos. Este modelo permite a las empresas identificar aquellos clientes que tienen un mayor impacto en los ingresos a largo plazo y desarrollar estrategias personalizadas para su retención.

Joung & Kim (2023) proponen un enfoque más avanzado que utiliza reseñas de productos en línea para segmentar a los clientes. El enfoque basado en Machine Learning interpreta las opiniones de los consumidores, ayudando a las empresas a desarrollar productos personalizados y mejorar la experiencia del cliente. Esta técnica no solo segmenta a los clientes de manera más precisa, sino que también ofrece una visión más clara de sus necesidades no satisfechas.

Análisis de Sentimientos y Fidelización de Clientes

El análisis de sentimientos basado en NLP (Natural Language Processing) es otra metodología que ha revolucionado la forma en que las empresas evalúan la satisfacción. El artículo de Capuano et al. (2021) explora cómo las tecnologías de Deep Learning aplicadas al análisis de sentimientos en redes sociales y plataformas de reseñas pueden proporcionar información detallada sobre la percepción del cliente. El análisis de estas opiniones permite a las empresas no solo identificar problemas en tiempo real, sino también ajustar rápidamente sus estrategias para mejorar la satisfacción (Capuano et al., 2021).

Además, Shinde & Shah (2018) subrayan que el uso de ML y Deep Learning ha permitido que los sistemas predictivos sean más precisos en la evaluación de la satisfacción, destacando cómo las redes neuronales pueden detectar cambios sutiles en el comportamiento del cliente que otros métodos no captan.

2.3.3. Aplicaciones de ML en la satisfacción de clientes durante la pandemia

Durante la pandemia COVID-19, se aceleró el uso de ML en diversos sectores, incluido el retail. Según Atawneh et al. (2022), las empresas de retail emplearon ML para predecir la satisfacción del cliente en el uso de sistemas de pago electrónico. Asimismo, ML con conjunto

con Big Data permite responder a las nuevas expectativas de los clientes y mejorar la experiencia de usuario en tiempos de alta incertidumbre.

Figura 4

Ciclo de vida de un cliente

Nota. Tomado de Porta (2016)

Conclusión del Estado del Arte

Así, el ML puede aplicarse en el Ciclo de Vida del Cliente para optimizar estrategias de retención, segmentación y satisfacción del cliente. El ML es una herramienta clave para predecir deserciones y personalizar productos, lo cual también influye en la relación empresa-cliente. Conforme estas tecnologías evolucionan, su aplicación puede expandirse para ofrecer nuevas formas de relación con los clientes.

III. MÉTODO

3.1. Tipo de investigación

El presente estudio fue aplicado. Adopta una investigación aplicada, ya que busca ampliar el conocimiento existente sobre la aplicación de modelos de ML en el análisis y predicción del comportamiento del cliente, deduciendo el nivel de impacto según su grado de fidelización, en el área de operaciones dentro del retail.

Se refuerza es hecho de que la investigación es aplicada porque busca resolver problemas específicos dentro del contexto del retail, particularmente en la predicción del comportamiento del cliente mediante el uso de ML, con el fin de mejorar su fidelización. Según Hernández-Sampieri & Mendoza (2018), la investigación aplicada se orienta hacia la resolución inmediata de problemas en contextos específicos, utilizando teorías y conocimientos para proponer mejoras prácticas. Además, Lozada (2014) destaca que este tipo de investigación permite a las organizaciones tomar decisiones informadas en función de problemas reales, con un impacto directo en la gestión y mejora de procesos. En este sentido, este estudio contribuye a la literatura sobre la fidelización del cliente, proporcionando una base teórica que podrá ser aplicada en otros contextos o estudios.

El nivel de investigación es descriptivo, ya que, pueden permitir la posibilidad de predecir un evento, aunque sean de forma rudimentaria; sin embargo, se debe tener la base teórica correcta, además de antecedentes que muestren un panorama claro de lo que puede pasar, solamente de esta forma se podrían plantear hipótesis. Por su parte, Arias & Covinos (2021) define este tipo de investigación como la que responde a preguntas sobre el "por qué" de los eventos, haciendo énfasis en la identificación de causas que permiten comprender mejor el fenómeno estudiado.

Para esta investigación se utilizará un diseño no experimental, ya que el estudio es de orientación prospectiva generando un aporte mas no con intención de estimular ninguna

variable pertinente en el estudio, manteniendo el análisis de los fenómenos en su estado natural. Por ende, en este diseño no hay estímulos o condiciones experimentales a las que se sometan las variables de estudio, los sujetos del estudio son evaluados en su contexto natural sin alterar ninguna situación; así mismo, no se manipulan las variables de estudio (Vizcaíno et al., 2023). Dentro de este diseño existen dos tipos: Transversal y longitudinal y la diferencia entre ambos es la época o el tiempo en que se realizan.

3.2. Población y muestra

La población de este estudio está compuesta por todos los clientes de una empresa de retail, que han realizado compras en un periodo de 2 años. Se estima que la población total es de aproximadamente 800,000 registros. Según Robles (2019), la población es el conjunto de todos los elementos o individuos que comparten características comunes relevantes para el estudio, en este caso, los clientes que han interactuado con la empresa en un período reciente.

Para este estudio, se utilizará un muestreo aleatorio simple, dado que proporciona a todos los elementos de la población la misma probabilidad de ser seleccionados, lo que garantiza una representatividad adecuada y reduce el sesgo en los resultados (Polo, 2022). El tamaño de la muestra se calculará utilizando la fórmula de muestreo para poblaciones finitas con un nivel de confianza del **95%** y un margen de error del **5%** (Gamboa, 2023).

El cálculo del tamaño se realiza a través de la siguiente fórmula:

$$n = \frac{Z^2 \times p \times q}{e^2 + (N-1) + Z^2 \times p \times q}$$

Dónde:

- n = Tamaño de la muestra.
- N = Tamaño de la población (800000).
- Z = Valor Z para el nivel de confianza (1.96 para un 95%).
- p = Probabilidad de éxito (0.5).

- q = Probabilidad de fracaso (0.5).
- e = Margen de error (0.05).

Al aplicar esta fórmula, se obtiene un tamaño de muestra de **384** participantes. Esto asegura que los resultados del estudio sean representativos y permitan hacer inferencias válidas sobre el comportamiento de la población en general.

3.3. Operacionalización de variables

Tabla 2

Operacionalización de variable independiente

Variable Independiente	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Unidad de Medida	Fórmula
Modelo mejorado de Segmentación de clientes usando ML	Los modelos de Machine Learning permiten segmentar clientes de manera dinámica y precisa, utilizando técnicas avanzadas para predecir su comportamiento y optimizar estrategias de fidelización en retail. En el contexto del retail, estos modelos permiten mejorar significativamente la toma de decisiones comerciales, ya que proporcionan predicciones basadas en datos reales que ayudan a anticipar las necesidades y comportamientos de los clientes (Sammut & Webb, 2017).	Implementación del algoritmo K-Means, para segmentar y predecir el comportamiento de los clientes. Busca la segmentación de los clientes para determinar la personalización de las ofertas.	Análisis de los requerimientos de nuevos parámetros para el modelo	Elaboración del Dataset Preprocesamiento y limpieza de datos. Identificación de nuevas variables para el modelo	Cantidad de datos de clientes Cantidad de nuevas Categorías de productos identificadas	No aplica
			Desarrollo del nuevo modelo utilizando el algoritmo K-Mens	Diseño de la arquitectura Identificación de los parámetros del modelo Determinar los componentes para el modelo	Cantidad de variables de segmentación de clientes utilizadas, precisión de la estimación.	No aplica
			Comparación del nuevo modelo	Capacidad de ajustar segmentaciones en tiempo real. Incorporación de parámetros para la segmentación de clientes	Cantidad de ajustes realizados, cantidad de nuevos datos y variables incorporados.	No aplica

Tabla 3

Operacionalización de variable dependiente

Variable dependiente	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Unidad de Medida	Fórmula
Comportamiento del cliente para su fidelización en una empresa de retail	La fidelización del cliente es el compromiso continuo de un cliente con una empresa de retail, reflejado en la repetición de compras, la preferencia por los productos o servicios de la empresa y la disposición del cliente a recomendar la marca a otros. Se basa en la satisfacción, la confianza y la relación a largo plazo entre el cliente y la empresa.	La fidelización del cliente se medirá a través del análisis de patrones de compra. Segmentación basada en ML: respuesta de los clientes a ofertas y personalización.	Visitas del cliente a la tienda	Frecuencia de compra	Número de compras realizadas por un cliente en un período determinado	$FC = \frac{\text{Número total de compras en un período}}{\text{Número de clientes activos en el mismo período}}$ • FC: Frecuencia de compra
			Monto de compra del cliente en la tienda	Valor promedio de compra	Valor monetario total de las compras realizadas por un cliente en un período determinado	$MC = \frac{\sum_{i=1}^{n} V_i}{n}$ MC: Monto de compra promedio \sum_{i=1} ^{n} Vi: Suma total de los valores de las compras realizadas en el período analizado n: Número total de compras en ese período
			Cuán reciente fue la última compra de un cliente	Recencia de compra	Tiempo transcurrido (en días, semanas o meses) desde la última compra realizada por un cliente.	$RC = \text{Fecha actual} - \text{Fecha de la última compra}$ RC: Recencia de compra Fecha actual: Día en que se realiza el análisis Fecha de la última compra: Día en que el cliente realizó su última compra

3.4. Instrumentos

En esta investigación utilizaremos una dataset con datos transacciones que han sido generados para una tienda retail en un periodo determinado.

Datos Transaccionales

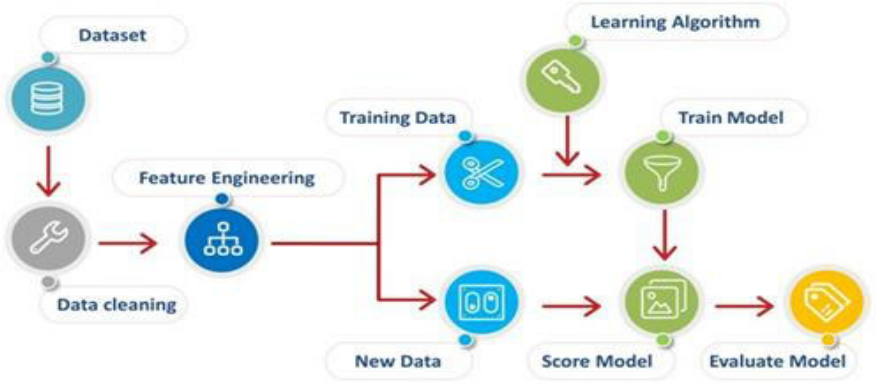
Los datos transaccionales son un componente clave para medir el comportamiento del cliente de manera cuantitativa. Este tipo de datos incluye información sobre las compras realizadas por los clientes, como la frecuencia, el valor monetario y la duración de la relación con la empresa. Estos datos se obtendrán del sistema de gestión de la empresa, donde se almacena el historial de compras de los clientes. Los datos transaccionales son esenciales para medir la satisfacción y la lealtad del cliente, ya que reflejan el comportamiento real de compra, permiten una medición objetiva del comportamiento del cliente (V. Kumar & Shah, 2004).

3.5. Procedimientos

Se siguió un conjunto de procesos estandarizados para asegurar la reproducibilidad.

Figura 5

Esquema General de ML



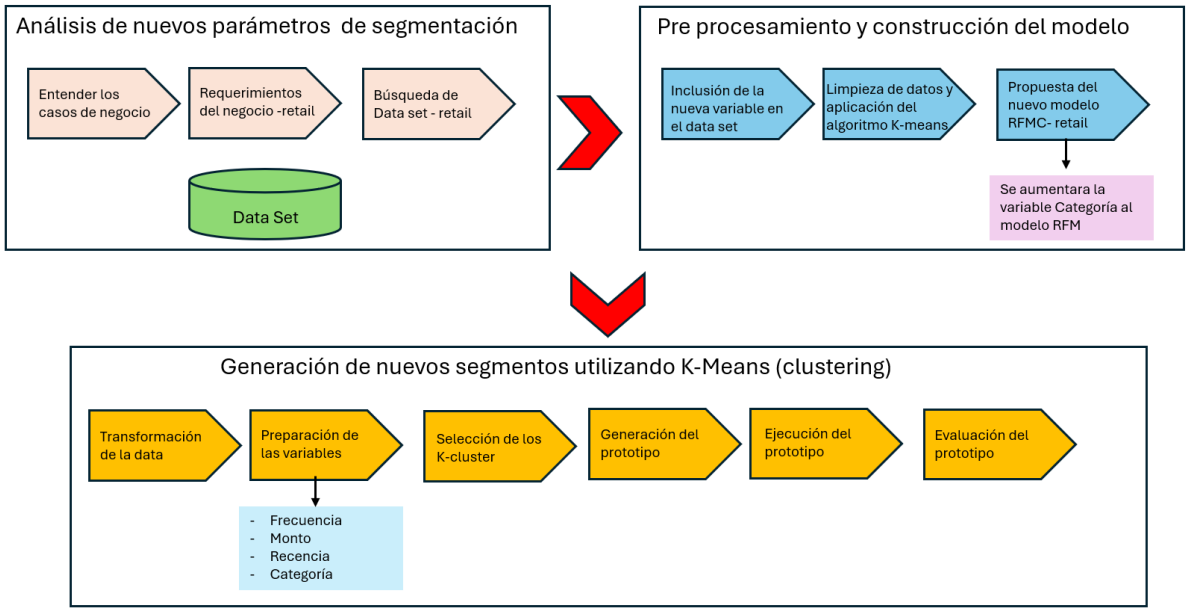
Nota. Adaptado de Marín (2020)

La información de las transacciones de los clientes permitió aplicar estrategias dirigidas a los clientes. Para lograr este objetivo se realiza la segmentación de clientes, a partir de los

cuales se pueden comprender las características de cada segmento y diseñar estrategias comerciales adecuadas. La tendencia actual es la aplicación de métodos de agrupación para identificar clientes potenciales, la combinación de algoritmos de aprendizaje automático con datos de usuarios es un ejemplo perfecto de segmentación de clientes que puede ayudar a las empresas a identificar segmentos de clientes que son difíciles de detectar mediante la intuición y la inspección manual de datos (A. Kumar, 2022).

Figura 6

Esquema de los procesos seguidos en el estudio



Recolección de datos

La primera etapa del procedimiento consiste en la recolección de datos, utilizaremos la observación directa como técnica en el uso del dataset generado para una industrial de retail. Estos datos incluirán la frecuencia de compra, valor monetario de las compras, y duración de la relación del cliente con la empresa. Según Rust & Zahorik (1993), los datos transaccionales proporcionan una evaluación objetiva del comportamiento de los clientes, lo que es esencial para estudiar la fidelización

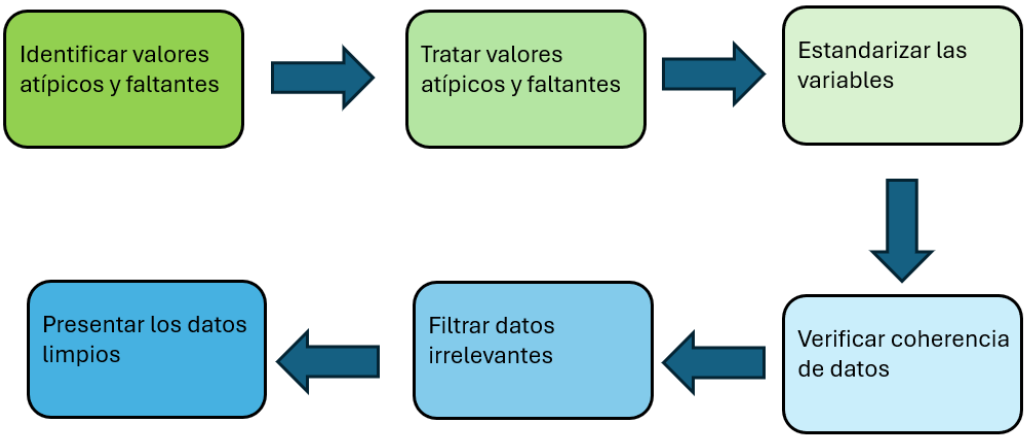
Limpieza y preparación de los datos

Se limpiaron y procesaron los datos para analizarlos, ya que algunos incluyeron valores atípicos, errores de entrada o información incompleta. Así, se utilizaron técnicas para:

- Eliminar valores atípicos.
- Verificar coherencia de datos.
- Presentar los datos limpios

Figura 7

Actividades detalladas del proceso de limpieza de datos



Entrenamiento y Validación del Modelo de ML

Una vez preparados los datos, se entrenará el Modelo propuesto de Machine Learning. Se llevará a cabo un proceso de validación utilizando técnicas como validación cruzada para evitar el sobreajuste de los modelos y asegurar que los resultados sean robustos.

Evaluación del modelo

La evaluación del modelo se realizará comparando los resultados de la segmentación de la primera corrida sin utilizar la nueva variable, y luego ejecutar la corrida adicionando la nueva variable y analizar los nuevos segmentos que han sido generados.

Análisis e interpretación de los resultados

Una vez evaluados los modelos de ML, se procederá al análisis e interpretación de los resultados. El análisis de los datos se realizará utilizando software estadístico como SPSS y

herramientas de ML como Python con bibliotecas especializadas (Rust & Zahorik, 1993). Luego, se compararon los resultados de fidelización pre y post implementación para determinar su efectividad.

Informe final y propuestas de mejora

Se redactaron las conclusiones recomendaciones según los resultados. El enfoque fue mejorar las estrategias de fidelización mediante ML y optimizar las decisiones comerciales asociadas a los clientes.

3.6. Consideraciones éticas

Este estudio, siguió estrictos lineamientos éticos, los cuales son:

Consentimiento informado

Se requirió del consentimiento de los participantes. Además, no se usaron datos sensibles que requieran un mayor nivel de consentimiento.

Privacidad y confidencialidad de los datos

Fue un desafío por la gran cantidad de datos de los clientes. Se trata de un principio clave en la ética de los estudios, ya que puede afectar a los participantes si se ve vulnerado. De acuerdo con Kuhuparuw et al. (2024), se requiere cuidar la privacidad de los datos de clientes, ya que esto puede reducir la confianza hacia la empresa y afectar su reputación.

Minimización del Riesgo

Otro aspecto importante en la ética de la investigación es la minimización del riesgo para los participantes. En este estudio, se garantizará que no existan riesgos físicos o psicológicos para los clientes. Aunque el análisis de datos puede parecer una actividad de bajo riesgo, el mal uso o divulgación de información personal podría tener efectos adversos.

Cumplimiento de Normativas Legales

Este estudio se llevará a cabo cumpliendo todas las normativas y leyes locales e internacionales relacionadas con la protección de datos personales.

IV. RESULTADOS

4.1. Análisis del uso de nuevos parámetros en los requerimientos del modelo mejorado de segmentación de ML

4.1.1. Lectura y compresión de datos

Para identificar la aplicación de nuevos parámetros, se ha elegido analizar un Dataset de una industria de retail, que contiene datos transaccionales (atributos), con todas las transacciones que tuvieron lugar entre el 12/01/2010 y el 12/09/2011, esta información trabajada por Chen et al. (2012), fue elegida porque se ajusta y se asemeja al tipo de negocio que se va a investigar, en esta tesis, el cual contiene los siguientes atributos:

Atributos:

- **Número Factura**: Número de factura. Numérico, un número entero de 6 dígitos asignado de forma única a cada transacción.
- **CodigoStock**: Código del producto (artículo). Nominal, un alfanumérico de 5 dígitos y 1 carácter asignado de forma única a cada producto distinto.
- **Descripción**: Nombre del producto (artículo). Nominal.
- **Cantidad**: Las cantidades de cada producto (artículo) por transacción. Numérico.
- **FechaFactura:** Fecha y hora de la factura. Numérico, el día y la hora en que se generó cada transacción.
- **PrecioUnitario**: Precio unitario. Numérico, precio del producto por unidad.
- **ID Cliente**: Número de cliente. Nominal, un número entero de 5 dígitos asignado de forma única a cada cliente.
- **País**: Nombre del país. Nominal, el nombre del país donde reside cada cliente.
- **Categoría**: Categoría a la que pertenece el producto

En la Tabla se muestran los registros contenidos en el dataset que se va a trabajar:

Tabla 4

Ejemplo de 10 registros del dataset

índex	Factura	CodigoStock	Cantidad	FechaFactura	Precio	ID Cliente	Categoría
0	489434	85048	12	1/12/2009 07:45	6.95	13085	Iluminación
1	489434	79323P	12	1/12/2009 07:45	6.75	13085	Iluminación
2	489434	79323W	12	1/12/2009 07:45	6.75	13085	Iluminación
3	489434	22041	48	1/12/2009 07:45	2.1	13085	Regalos y Decoración
4	489434	21232	24	1/12/2009 07:45	1.25	13085	Otros
5	489434	22064	24	1/12/2009 07:45	1.65	13085	Otros
6	489434	21871	24	1/12/2009 07:45	1.25	13085	Vajilla
7	489436	21755	18	1/12/2009 09:06	5.45	13078	Regalos y Decoración
8	489436	21754	3	1/12/2009 09:06	5.95	13078	Hogar
9	489436	84879	16	1/12/2009 09:06	1.69	13078	Vajilla
10	489436	22119	3	1/12/2009 09:06	6.95	13078	Regalos y Decoración

En el cuadro mostrado se puede ver que el dataset seleccionado contiene los campos necesarios para ser utilizados en la propuesta del nuevo modelo

En la Tabla se muestra que existe la variable CATEGORIA, pero no ha sido utilizada en el modelo de segmentación.

Vamos a resumir la información a nivel de registros que contiene el dataset, la cual es mostrada en la siguiente figura:

Figura 8

Información general del Dataset

```
RangeIndex: 962709 entries, 0 to 962708
Data columns (total 9 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Factura       962709 non-null  int64
 1   CodigoStock   962709 non-null  object
 2   Descripcion   962709 non-null  object
 3   Cantidad      962709 non-null  int64
 4   FechaFactura  962709 non-null  object
 5   Precio        962709 non-null  float64
 6   ID Cliente    742248 non-null  float64
 7   Pais          962709 non-null  object
 8   Categoria     962709 non-null  object
dtypes: float64(2), int64(2), object(5)
```

Como se puede apreciar la cantidad total de registros que contiene el dataset es de 962,709, se debe de tener en consideración que en una transacción de compra se emite una factura para un solo cliente, la cual puede contener varios códigos de productos que pertenecen a una categoría especifica de un producto.

Ahora procedemos a analizar la información contenida en el dataset, como se puede apreciar en la siguiente tabla:

Tabla 5

Contenido del Dataset

	Factura	**Cantidad**	**Precio**
count	962709	962709	962709
mean	540316.8	10.840521	3.354355
std	25755.177	125.25409	9.777331
min	489434	1	0
25%	520671	1	1.25
50%	540498	3	2.1
75%	562851	12	4.13

max	581587	80995	8142.75

El resumen estadístico de la Tabla 5 permite analizar e interpretar la distribución de las variables en sus valores mínimos y máximos

FACTURA

- Los valores están en orden secuencial, no tiene valor analítico directo, pero permite entender el rango temporal.
- Rango de facturas: desde la numeración 489,434 al 581,587 lo cual indica una gran cantidad de transacciones contenidas (más de 90 mil facturas).
- Desviación estándar: ~25,755 indica que las facturas están bien distribuidas.
- Se puede usar para construir la recencia porque contiene la última fecha de compra del cliente.

CANTIDAD

- Media: la cantidad de artículos por factura emitida es de 10.84.
- Mediana: 3 artículos → indica que los datos están sesgados positivamente (hay muchas compras pequeñas y unas pocas muy grandes).
- 25% de los pedidos tiene solo 1 artículo, y el 75% tiene 12 o menos.
- Desviación estándar: Muy alta (125), también refuerza que hay mucha variabilidad, evaluar si es necesario eliminar valores extremos para análisis de RFM.

PRECIO (unitario por producto)

- Media: 3.35
- Mediana: 2.10, los datos están sesgados (hay muchos productos con bajo precio y algunos muy caros).
- Precio mínimo: 0 → podría tratarse de un error

- Precio máximo: 8,142.75 → puede ser un producto de lujo o error tipográfico.
- Desviación estándar alta (9.78): muchos precios dispersos, evaluar estos extremos.

4.1.2. Limpieza de los datos

Se eliminarán los registros que contengan datos nulos o que no sean posible ser utilizados en la propuesta del modelo.

Figura 9

Comando de identificación de datos nulos del Dataset

```
retail.isnull().sum()
```

Se lograron identificar 229,001 registros que contienen en el campo ID CLIENTE un valor nulo, estos registros deben ser eliminados porque se necesita identificar al cliente para poder ser procesados en el modelo.

Tabla 6

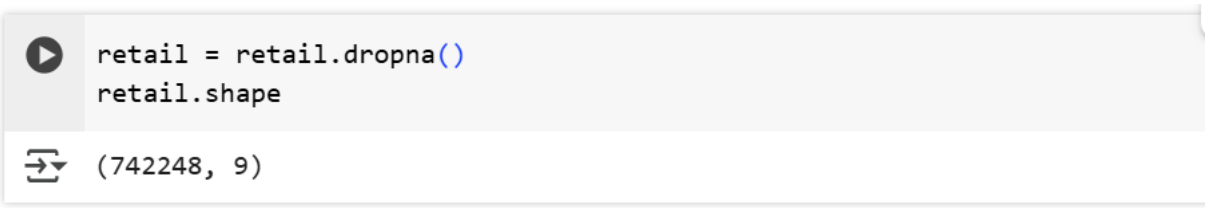
Resultados de la identificación de datos nulos

Factura	0
CodigoStock	0
Descripción	0
Cantidad	0
FechaFactura	0
Precio	0
ID Cliente	229001
País	0
Categoría	0

Se procede con la eliminación de los 229,001 registros nulos, luego de la depuración nos vamos a quedar a trabajar con 742,248 registros, como se muestra en la siguiente figura:

Figura 10

Comando de depuración de datos nulos del Dataset



Como podemos ver la cantidad de registros del dataset a ser utilizados es de 742,248.

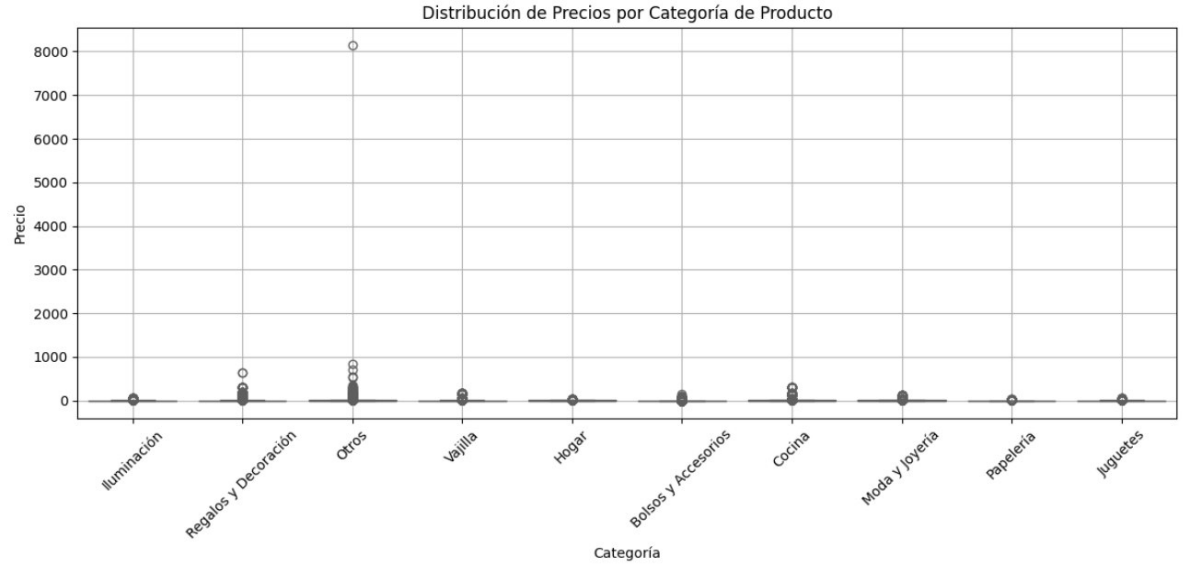
4.1.3. Identificación de nuevos parámetros a utilizar en la propuesta del modelo

De los posibles parámetros (atributos) que contiene el dataset, vamos a analizar como el campo CATEGORÍA, puede influir en los parámetros, de acuerdo con lo mostrado en la siguiente figura:

- Precio
- Cantidad
- Venta

Figura 11

Relevancia del campo CATEGORÍA VS PRECIO



Elementos claves de la figura:

- **Eje X:** Categorías de productos (Iluminación, Regalos y Decoración, Otros, Vajilla, Hogar, etc.).
- **Eje Y:** Precio de los productos.
- Vemos una gran cantidad de outliers, incluyendo el precio más alto del dataset (>8000).
- Los precios están muy concentrados en rangos bajos en casi todas las categorías, salvo "Otros", que tiene una gran dispersión.
- La categoría “Otros” distorsiona el análisis y puede requerir reclasificación o segmentación más fina.
- Este tipo de gráfico es ideal para:

  Detectar outliers que podrían necesitar limpieza.

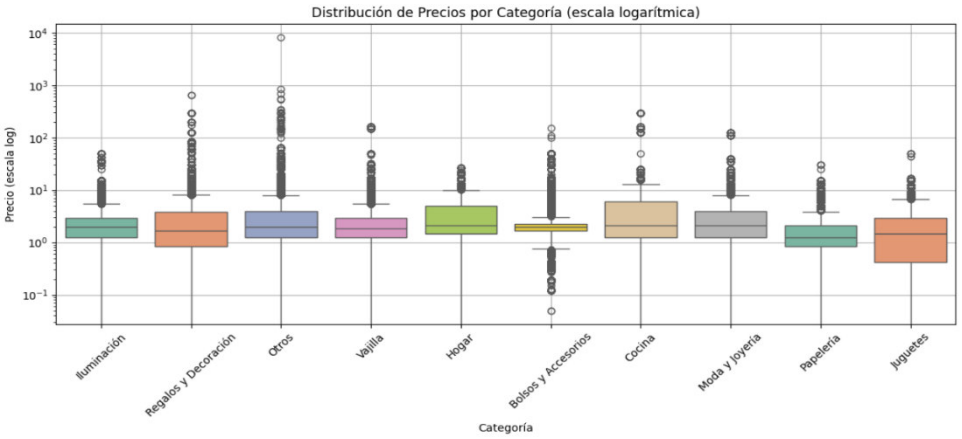
  Comparar rangos de precios entre categorías.

  Identificar oportunidades de segmentación por precio dentro de cada categoría.

Para poder realizar una mejor interpretación vamos a trabajar utilizando una escala logarítmica.

Figura 12

Relevancia del campo CATEGORÍA VS PRECIO (Escala Logarítmica)



De la figura podemos tener las siguientes consideraciones:

- Precios promedio similares en la mayoría de las categorías.
- Medianas de precios en casi todas las categorías están alrededor de 1 a 10 unidades monetarias.
- Esto indica que la mayoría de los productos son de bajo costo, independientemente de la categoría.

En las siguientes figuras vemos con un mayor de detalle las distribuciones de Precios por Categoría.

Figura 13

Detalle de PRECIOS POR CATEGORIAS

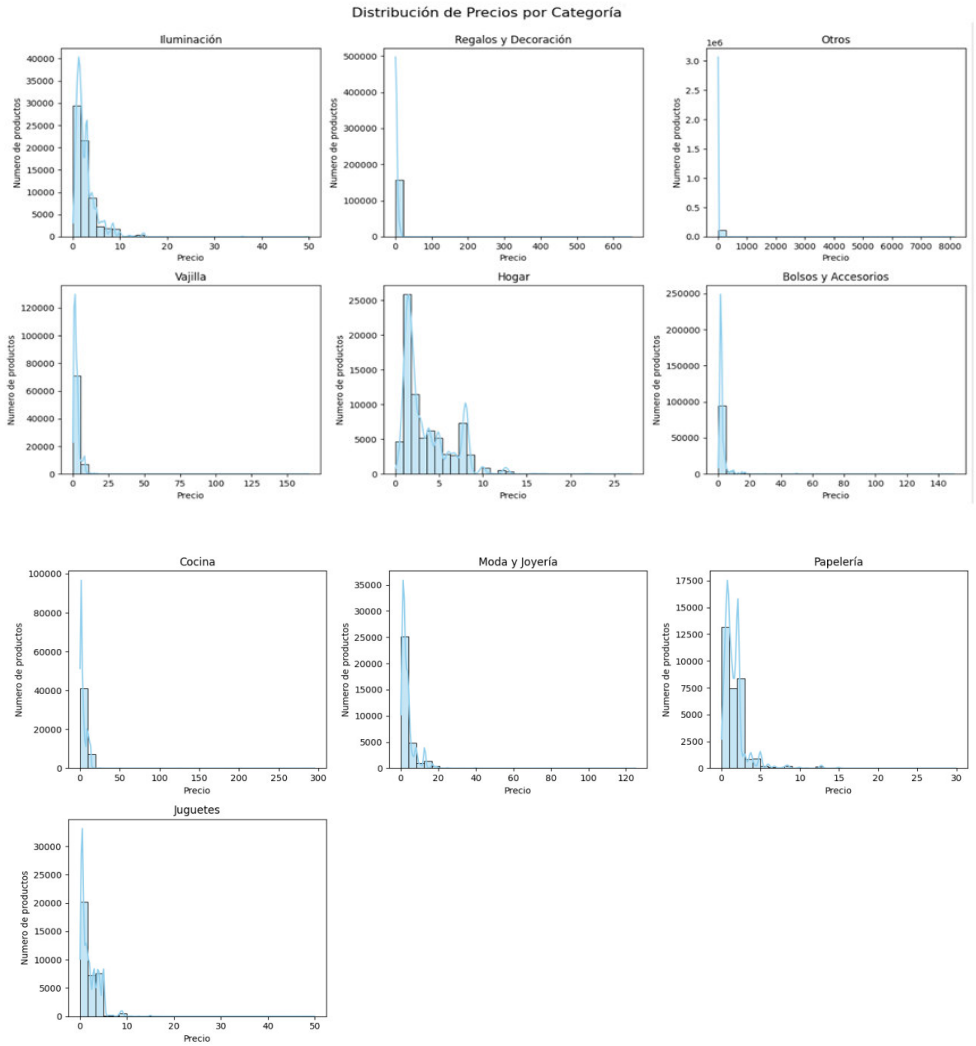
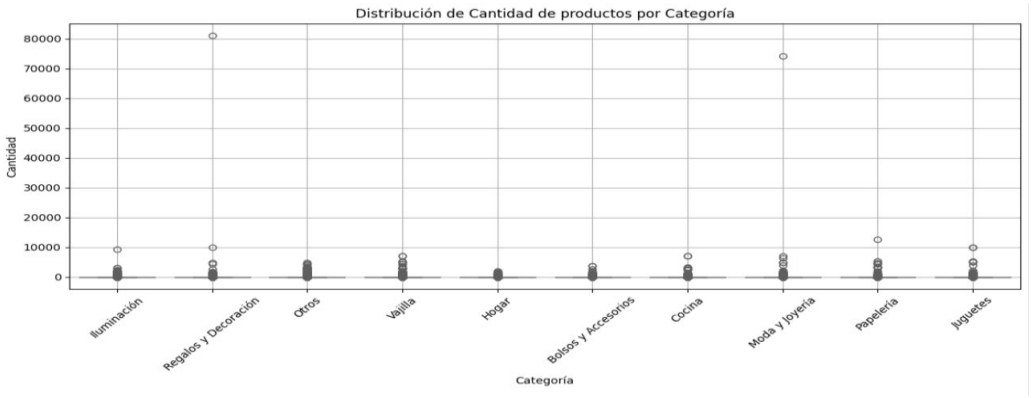


Figura 14

Relevancia del campo CATEGORÍA VS CANTIDAD



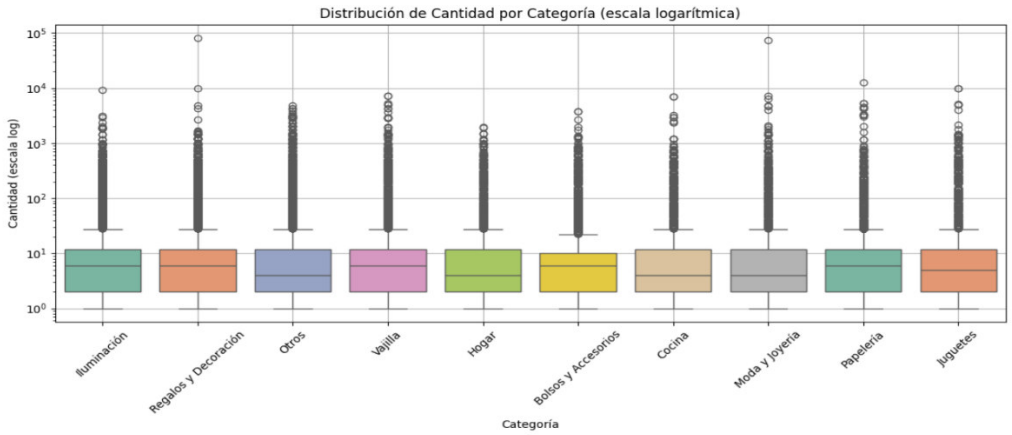
De la figura se puede interpretar lo siguiente:

- Los valores atípicos extremos están presentes en casi todas las categorías, pero algunos destacan mucho más (especialmente “Regalos y Decoración”, “Moda y Joyería” y “Papelería”).
- Es fundamental analizar si estos valores representan errores o comportamientos legítimos (por ejemplo, clientes comprando grandes cantidades).
- Este gráfico apoya la idea de que la variable “Cantidad” tiene una alta varianza y debe tratarse con cuidado en cualquier modelo analítico.

Para poder realizar una mejor interpretación de la relevancia vamos a trabajar con una escala logarítmica.

Figura 15

Distribución de CANTIDAD por CATEGORÍA (escala logarítmica)



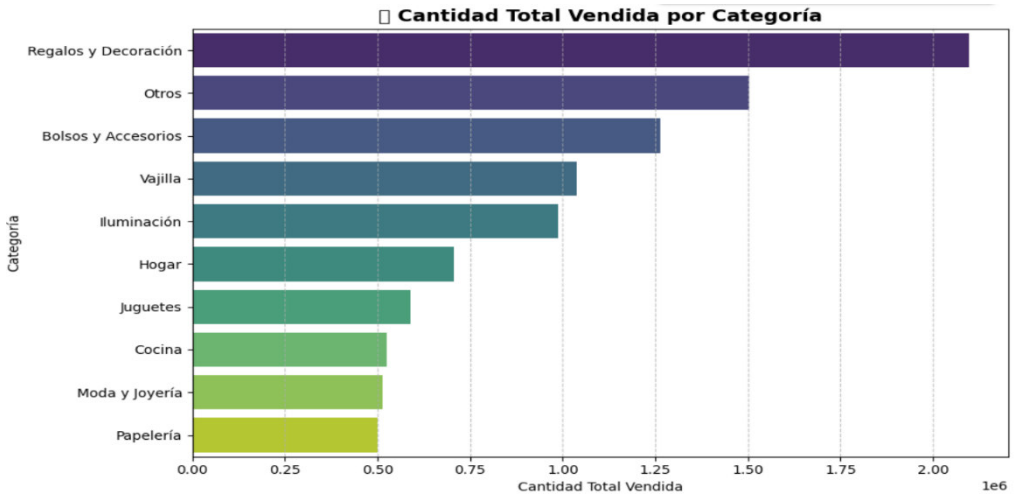
- A diferencia del gráfico anterior en escala lineal, aquí se puede observar mejor cómo se distribuyen todas las cantidades, desde 1 hasta decenas de miles.
- El uso de logaritmo permite realizar la comparación más justa de todas las categorías pese a la existencia de outliers extremos.
- En casi todas las categorías, la mediana (línea horizontal dentro de la caja) está entre 1 y 10 unidades por la línea de factura.
- Esto confirma que la mayoría de las ventas se hacen en cantidades pequeñas, independientemente de la categoría.

- Todas las categorías presentan una distribución similar en términos de rango y presencia de outliers hacia arriba (compras masivas).
- Esto sugiere que en cualquier categoría pueden ocurrir compras por volumen, aunque no sean la norma.

En la siguiente figura podemos ver cómo se distribuye la venta por categoría

Figura 16

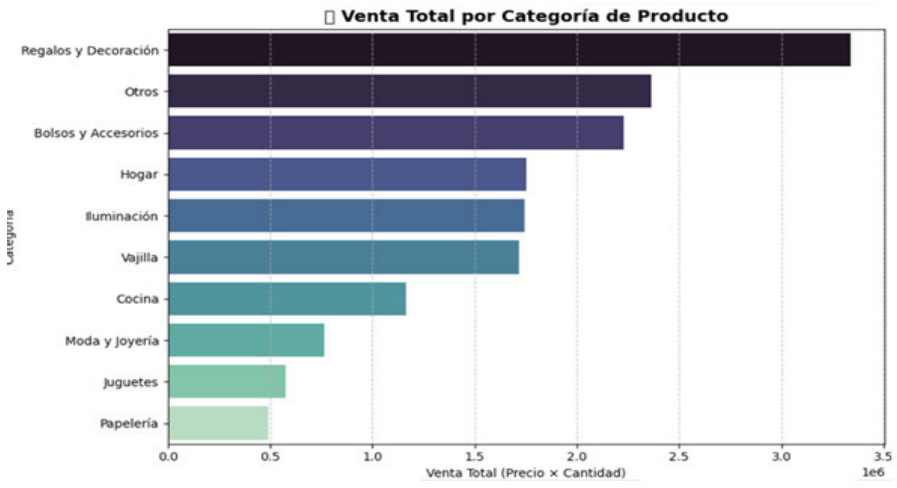
Distribución de CANTIDAD TOTAL por CATEGORÍA



Ahora vamos a analizar la relevancia del atributo Venta Total mostrada en la figura 17.

Figura 17

Relevancia del campo CATEGORÍA VS VENTA TOTAL



Del grafico podemos realizar la siguiente interpretación:

La distribución de ingresos no es homogénea:

- Las 3 primeras categorías generan una proporción importante del total de ingresos, lo que indica una posible concentración de valor.
- Se podría aplicar la regla del 80/20 para ver si pocas categorías generan la mayoría de las ventas.

La categoría “Otros” debe ser revisada:

- Su alta posición sugiere que esconde productos clave o está mal clasificada.
- Si la limpias o subdivides, podrías obtener información mucho más precisa.

La rentabilidad por categoría se puede utilizar para decisiones estratégicas:

- Al cruzar esta información con el modelo RFM podemos ver los clientes que compran en las categorías más rentables si son los más frecuentes o recientes.
- También es ideal para campañas de marketing segmentadas.

4.1.4. Justificación de que el atributo “Categoría”

Análisis de Correlación No Paramétrica. Como "Categoría" es una variable categórica y R, F, M son numéricas, no se puede usar Pearson. En su lugar, se puede aplicar:

- Análisis de varianza (ANOVA) si se asume normalidad.
- Kruskal-Wallis si no hay normalidad (es más robusto).
- Chi-cuadrado solo si convertimos R, F, M en categorías también.

El **p-value** es una medida que nos ayuda a decidir si una diferencia observada entre grupos (por ejemplo, el monto promedio de compra entre distintas categorías) es real o solo ocurrió por casualidad.

La prueba **Kruskal-Wallis** es estadística no paramétrica que se usa para comparar más de dos grupos cuando:

- Tenemos una variable numérica (como Monto, Frecuencia, Recencia),
- Y una variable categórica que agrupa esos datos (como Categoría de producto).

Esta prueba evalúa si las distribuciones de la variable numérica son iguales en todos los grupos o si hay diferencias significativas entre al menos un par de grupos.

Si el p-valor < 0.05 (nivel de significancia típico del 5%):

- **Se puede deducir:** Hay al menos un grupo (una categoría) que se comporta distinto en cuanto al monto.
- Esto sugiere que la variable "Categoría" está relacionada con el "Monto".
- Por ejemplo: clientes que compran "Moda y Joyería" gastan sistemáticamente más (o menos) que los de "Vajilla".

Si el p-valor ≥ 0.05:

- No hay evidencia suficiente para afirmar que "Categoría" afecta el "Monto". Es decir, las diferencias observadas podrían ser sólo producto del azar.

Vamos a proceder con la validación, como se puede apreciar en la siguiente figura:

Figura 18

Relevancia del campo CATEGORÍA PARA SER UTILIZADO EN EL NUEVO MODEL

Se usó la prueba de Kruskal-Wallis para verificar si el comportamiento de los clientes (en términos de monto gastado, frecuencia de compra y recencia) y vemos que el p-valor es 0.000 en todos los casos:

Resultados:

- Monto: p-valor = 0.0000
- Frecuencia: p-valor = 0.0000
- Recencia: p-valor = 0.0000

Conclusión:

Existen diferencias estadísticamente significativas en el comportamiento de compra según la categoría. Esto implica que la variable Categoría tiene relación con Monto, Frecuencia y Recencia.

¿Qué significa esto para el modelo?

Justifica la inclusión de la variable categórica "Categoría" en nuestro modelo RFMC. Nos permite segmentar clientes con mayor precisión, sabiendo que ciertos comportamientos están ligados a lo que compran. Apoya el argumento de que una mejor segmentación puede conducir a campañas de marketing más eficaces.

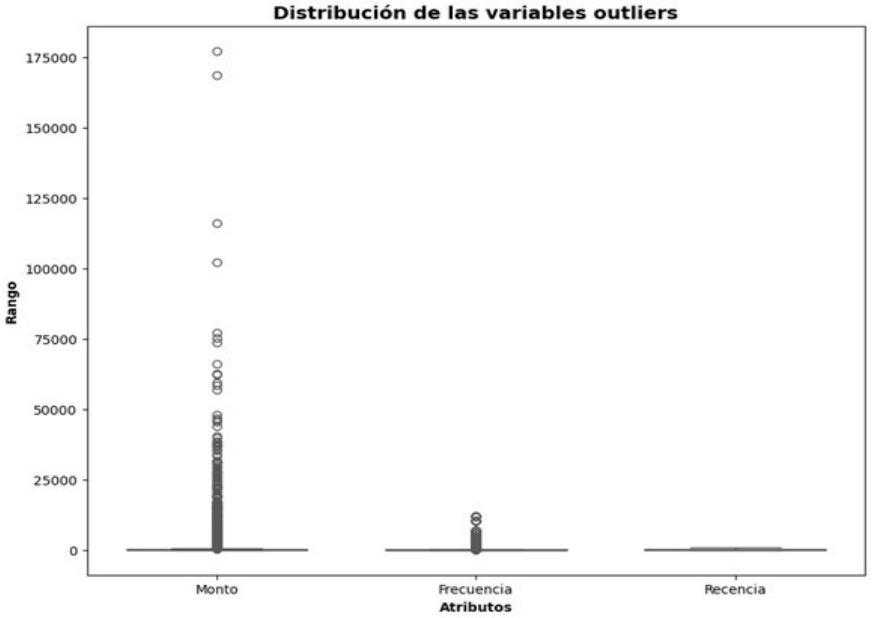
4.2. Desarrollo de la propuesta del modelo de segmentación de ML

4.2.1. Detección de datos atípicos (outliers) para ser separados del modelo

Se procede a la identificación de estos valores ya que pueden afectar nuestra información estadística, como se puede ver la siguiente figura:

Figura 19

Identificación de variables outliers



Como podemos apreciar si se encuentran valores atípicos por lo que debemos de aplicar herramientas estadísticas como:

- Asimetría: mide el grado de simetría de una distribución y puede ser positiva (sesgada a la derecha), negativa (sesgada a la izquierda) o simétrica
- Curtosis: mide qué tan achatada o apuntada está la distribución comparada con una curva normal

Asimetría ≈ 0: Distribución simétrica (homogénea).

Asimetría > 1 o < -1: Distribución asimétrica (heterogénea, valores extremos presentes).

Curtosis > 3: La distribución tiene colas largas (más valores extremos de lo normal).

Vamos a calcular la Asimetría y Curtosis para las variables (feature) del modelo, como podemos ver en la siguiente figura:

Figura 20

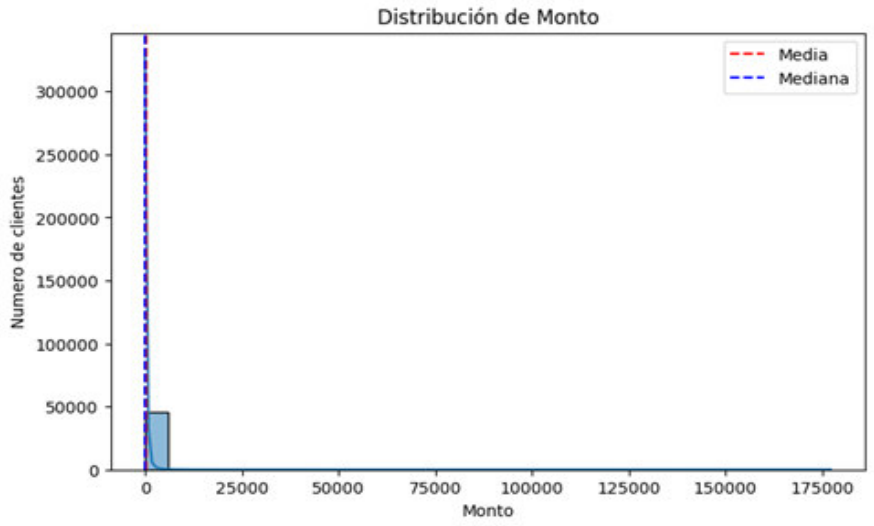
Cálculo de la Asimetría y Curtosis

```
# Calcular la asimetría y curtosis de cada feature
stats_summary = pd.DataFrame({
    'Feature': ['Monto', 'Frecuencia', 'Recencia'],
    'Skewness': [
        stats.skew(rfm['Monto']),
        stats.skew(rfm['Frecuencia']),
        stats.skew(rfm['Recencia'])
    ],
    'Kurtosis': [
        stats.kurtosis(rfm['Monto'], fisher=True),  # Fisher=True: curtosis ajustada (normal = 0)
        stats.kurtosis(rfm['Frecuencia'], fisher=True),
        stats.kurtosis(rfm['Recencia'], fisher=True)
```

Los resultados obtenidos se pueden apreciar en las siguientes figuras:

Figura 21

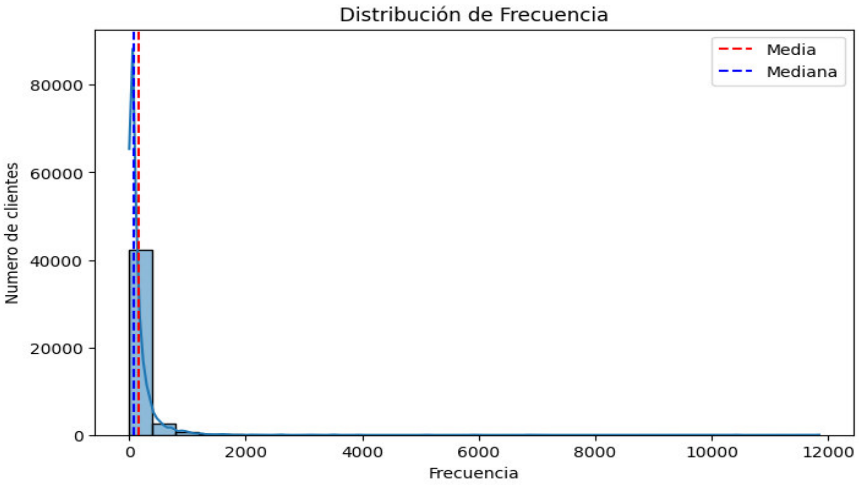
Distribución de la variable Monto



Según la figura podemos ver que la Asimetría es de 40.40 y la Curtosis de 2522.31 la distribución es altamente asimétrica positiva y con muchísimos outliers. Significa que la mayoría de los clientes gastan poco, pero unos pocos gastan muchísimo. Este comportamiento es típico de ingresos: pocos clientes generan la mayor parte de las ventas (regla de Pareto).

Figura 22

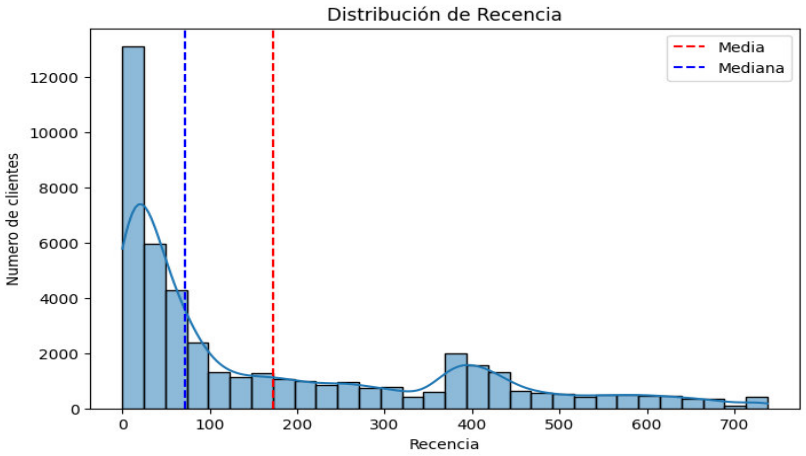
Distribución de la variable Frecuencia



Según la figura podemos ver que la Asimetría es de 16.61 y la Curtosis es de 428.42. Lo que indica que es muy asimétrica positivamente. La mayoría de los clientes compra pocas veces, y unos pocos compran con mucha frecuencia. Es otra variable con cola larga y muchos valores extremos.

Figura 23

Distribución de la variable Recencia



Según la figura la Asimetría es de 1.09 y la Curtosis de 0.0. Es una leve asimetría positiva y muy baja curtosis, lo que indica una distribución más cercana a normal, pero más

plana (platicúrtica). Esto sugiere que los valores de recencia están más distribuidos de forma uniforme y hay menos presencia de valores extremos.

Luego procederemos a aplicar la función logaritmo a las variables del modelo como podemos ver en la siguiente figura:

Figura 24

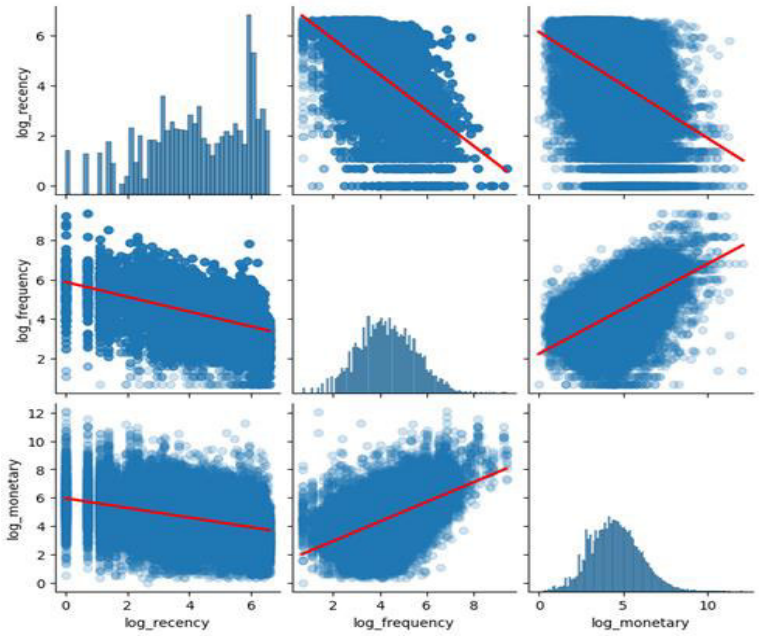
Aplicación de la función log1p a las variables (Recencia, Frecuencia, Monto)

```
# Crear columnas logarítmicas si no existen
rfm['log_recency'] = np.log1p(rfm['Recencia'])
rfm['log_frequency'] = np.log1p(rfm['Frecuencia'])
rfm['log_monetary'] = np.log1p(rfm['Monto'])

# Mostrar el gráfico directamente
pairplot = sns.pairplot(
    rfm[['log_recency', 'log_frequency', 'log_monetary']],
    kind="reg",
    plot_kws={"line_kws": {'color': 'red'}, "scatter_kws": {"alpha": 0.2}}
)
```

Los gráficos generados se muestran en las siguientes figuras:

Figura 25

Correlación de las variables RFM (Recencia, Frecuencia, Monto)



Se toman los logaritmos (log1p) de las métricas Recencia, Frecuencia y Monto principalmente porque vemos que las distribuciones están muy sesgadas: la Recencia puede

tener muchos valores pequeños (clientes recientes) y pocos muy grandes. La Frecuencia y Monto también tienden a tener una larga cola derecha (algunos clientes compran muchísimo).

El trabajar con la función LOG ayuda a reducir la asimetría y acercar la distribución a una forma más normal, lo que nos permite:

- Mejorar la visualización.
- Reducir el impacto de valores extremos (outliers).
- Hacer más interpretables las relaciones lineales en regresiones o correlaciones.
- Para modelos estadísticos y de ML, los valores log-transformados suelen:

  Mejorar la precisión.

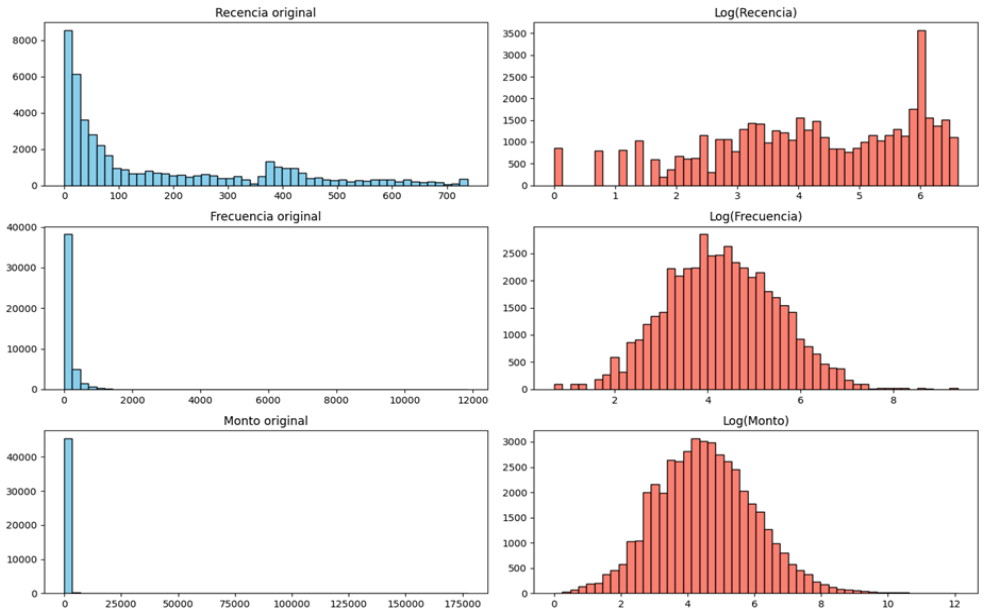
  Estabilizar la varianza.

  Alinear mejor con supuestos del modelo.

Ahora vamos a realizar una comparación antes y después de aplicar la función log:

Figura 26

Histogramas antes y después del log para cada métrica



En la Recencia original se ve una Distribución fuertemente sesgada a la derecha (asimetría positiva), lo cual indica que la mayoría de los valores están entre 0 y 100 días, pero

hay clientes con recencia muy alta (>700 días), también vemos una curtosis de cola larga: muchos clientes no han comprado en mucho tiempo. Si lo comparamos aplicando la función logarítmica vemos que mejora la forma de la distribución, pero no la normaliza completamente, La forma sigue siendo algo irregular (bimodal).

En referencia a la Frecuencia original, vemos que está muy sesgada a la derecha, lo cual indica que la mayoría de los clientes compran pocas veces, pero hay algunos con frecuencia muy alta (hasta 12,000). Al aplicar la transformación logarítmica logra una distribución aproximadamente normal (gaussiana), con centro alrededor de 4-5 lo cual es ideal para un modelado y clustering.

Sobre la Variable Monto original, está extremadamente sesgada a la derecha, lo que indica que la mayoría de los clientes generan poco ingreso, pero hay unos pocos con montos muy altos (>175,000). Luego de aplicar la transformación logarítmica vemos que normaliza bien la distribución, La curva se ve bastante simétrica y con forma de campana, perfecta para algoritmos que suponen normalidad (p. ej. K-means).

Ahora aplicaremos la transformación de Box-Cox, esta transforma variables dependientes no normales en una forma normal. La normalidad es un supuesto importante para muchas técnicas estadísticas; si los datos no son normales, aplicar una transformación de Box-Cox permite realizar un mayor número de pruebas.

Razones para aplicar la transformación de Box-Cox:

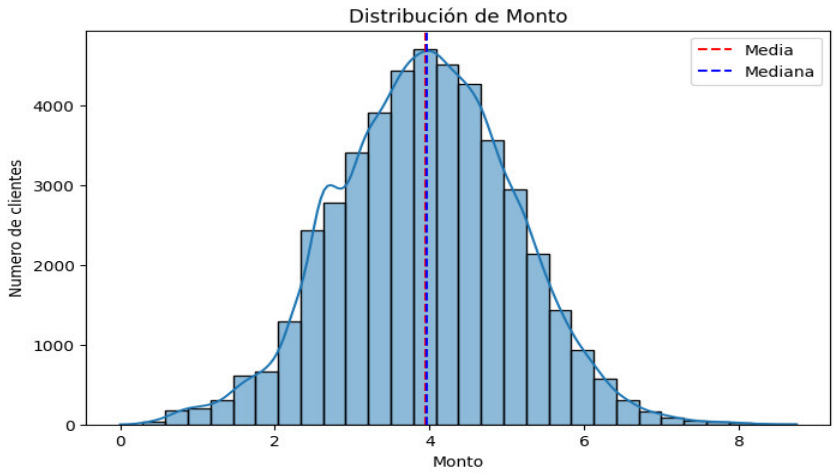
- Estabilizar la varianza: Si las variables tienen alta varianza o distribuciones sesgadas, Box-Cox ayuda a hacerlas más normales.
- Mejorar la normalidad: Muchos algoritmos de clustering (por ejemplo, K-means) funcionan mejor cuando los datos están más cercanos a una distribución normal, ya que dependen de distancias euclidianas. (Corrige la asimetría (skewness))

- Garantiza la comparabilidad: Si las variables tienen escalas y distribuciones diferentes, normalizarlas antes de aplicar escalado puede mejorar los resultados del clustering.

Los resultados los podemos ver en los siguientes gráficos, aplicaremos la transformación a la variable Monto:

Figura 27

Aplicando Box-Cox variable monto - clientes

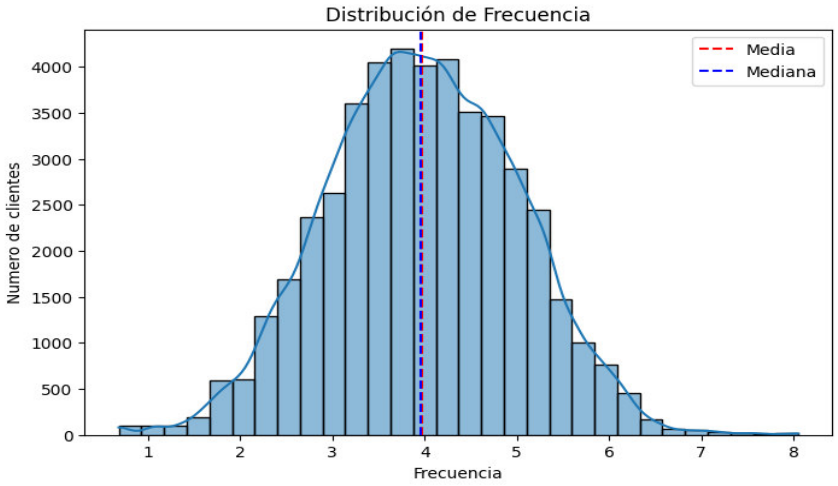


Como se puede apreciar la distribución se ha vuelto mucho más simétrica, reduciendo la asimetría positiva. La media y la mediana están más alineadas, lo que indica que la distribución es menos sesgada. Se observa una forma más parecida a una distribución normal con valores más equilibrados.

Ahora aplicaremos la transformación Box-Cox a la variable Frecuencia:

Figura 28

Aplicando Box-Cox variable frecuencia - clientes

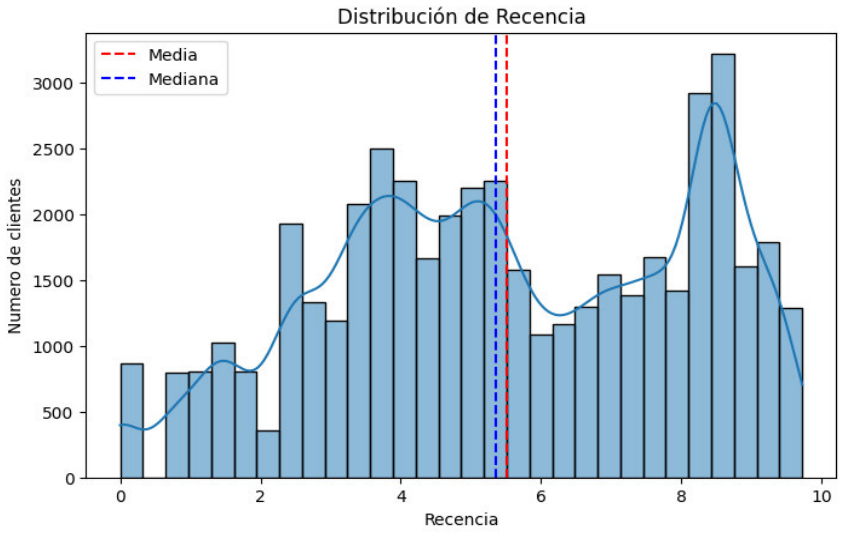


La forma original de la variable Frecuencia era altamente asimétrica positiva, con muchos clientes de baja frecuencia y unos pocos con valores extremos, al aplicar la transformación Box-Cox se logró corregir esa asimetría, generando una distribución con forma de campana (aproximadamente normal). La proximidad entre la media y la mediana indica que la distribución es simétrica o casi simétrica, esto reduce el sesgo en cualquier análisis que dependa de medidas de tendencia central. También se ve que la forma de la curva sugiere que no hay colas extremas ni un pico excesivo, indica que es una distribución “normalizada” con curtosis moderada.

Ahora aplicaremos la transformación Box a la variable Recencia:

Figura 29

Aplicando Box Cox a la Variable Recencia - Clientes



De acuerdo con el gráfico, aunque la forma no es perfectamente normal, la media y la mediana están bastante alineadas, lo que sugiere que no hay fuerte sesgo. Esto mejora la estabilidad del modelo que depende de la media como referencia, la curtosis está controlada.

4.2.2. Construcción del modelo

El modelo que vamos a utilizar como base para la propuesta de mejora es el RFM, este ha sido la base de la mayoría de las segmentaciones de marketing directo durante décadas (Miglautsch, 2002). Es un proceso científicamente probado, se basa en el principio de Pareto, comúnmente conocido como la regla 80-20, esta indica que el 80% de los resultados provienen del 20% de las causas. Del mismo modo, el 20% de los clientes contribuyen al 80% de sus ingresos totales. El principio de Pareto es el núcleo del modelo RFM. La aplicación de este modelo nos permite centrar los esfuerzos de la empresa en los segmentos críticos de clientes que son los que proporcionan un mayor retorno de la inversión (Anish, 2022).

Figura 30

Esquema de Pareto

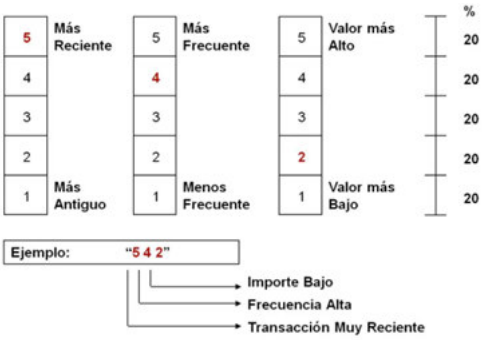
Nota. Tomado de Cordova G (2011)

Para realizar este análisis, se asigna una puntuación a cada cliente en función de tres variables previamente mencionadas: recencia, frecuencia y monto. Cada una de estas variables se califica en una escala del 1 al 5, donde 1 representa la puntuación más baja y 5 la más alta. El modelo RFM se basa en la división en quintiles, lo que significa segmentar los datos en cinco grupos de igual tamaño. De este modo, los clientes con las puntuaciones más altas obtendrán R5, F5 y M5, lo que indica que han comprado más recientemente, con mayor frecuencia y han gastado más dinero. En contraste, aquellos con las puntuaciones más bajas recibirán R1, F1 y M1. En base a esta clasificación, se puede orientar las acciones de marketing a aquellos segmentos de clientes estratégicos. Un ejemplo de ello son las promociones que se pueden dirigir a distintos grupos de clientes en función de sus características.

Las medidas de RFM explican lo que hacen los clientes: cuándo, con qué frecuencia y cuánto compran, esos son los parámetros básicos de comportamiento (Wei et al., 2010). Por lo general, los clientes se agrupan en combinaciones de valores R, F y M. También se utiliza, y esto es dependiente del sector de la empresa, un RFM ponderado que considera diferentes pesos de R, F y M, considerando la apreciación subjetiva de gerentes o especialistas.

Figura 31

Esquema de Clasificación Modelo RFM



Nota. Tomado de Cordova G (2011)

Para entender qué valor darle a cada parámetro, debemos entender uno a uno:

Recencia

Los clientes que han comprado recientemente son más propensos a comprar nuevos productos que aquellos que llevan tiempo sin consumir. Para esta variable se debe de tener en cuenta dos complicaciones: por un lado, la omnicanalidad imperante que complica la vinculación de las compras con el medio, y por otro, las características individuales del negocio, ya que no se consume con la misma frecuencia en un negocio de alimentos que en uno de ropa.

Frecuencia

Esta variable también depende de las características del sector en el que se enmarque el negocio. Por estadística, los consumidores que han comprado más productos serán más propensos a seguir comprando.

Monto o valor monetario

Cuánto más dinero se haya gastado un cliente en un negocio, más propenso será a seguir consumiendo. De nuevo, esta variable está supeditada al tipo de negocio, porque cada producto tiene un valor económico distinto. A pesar de que el modelo RFM se aplica en diferentes tipos de empresas para segmentar clientes por clústeres, numerosos estudios han propuesto modelos

ampliados (Ho et al., 2023) incorporando nuevas variables o aplicando nuevas herramientas analíticas, para nuestro estudio vamos a adicionar la variable CATEGORÍA al modelo.

Estandarización de Escalas

Antes de aplicar algoritmos de agrupamiento (clustering), es necesario que todas las variables a utilizar estén en una misma escala. Esto nos asegura que ninguna variable domine la distancia entre puntos, ya que muchos algoritmos de clustering se basan en medidas como la Distancia Euclidiana. La Distancia Euclidiana representa el camino más corto entre dos puntos en el espacio euclidiano. Es la distancia que se mediría con una regla, ampliada a cualquier número de dimensiones (Chugani, 2024). En la agrupación de k-means, la distancia euclídea ayuda a clasificar los puntos de datos en grupos, conectando cada punto con el centro más cercano de un grupo. Esto ayuda a organizar los datos en categorías que comparten similitudes, útil en la segmentación de clientes o durante la investigación para agrupar temas similares.

Las variables del modelo RFM (Recencia, Frecuencia, Monto) pueden tener rangos numéricos muy diferentes. Ejemplo: Recencia entre 0–700, Frecuencia entre 1–10000, Monto entre 1–100000. Si no se ajustan, las variables con mayor rango tendrán más peso en el clustering, generando segmentaciones sesgadas o erróneas. Tenemos dos métodos para cambiar las escalas:

1. Escalado Mínimo-Máximo (Min-Max Scaling): (utilizar numeración de fórmulas editas con el editor de fórmulas)

$$Fórmula = {(c - min)}/{(max - min)}$$

- Transforma los datos para que estén entre **0 y 1**.
- Útil cuando se desea **preservar la distribución original**.

2. Estandarización (Standardization):

$$Fórmula = {(x - \mu)}/{\sigma}$$

- Distribuye los datos con media 0 y desviación estándar 1.

- Ideal cuando se desea reducir el impacto de outliers o si se asume distribución normal.

Identificar la escala de las variables antes de aplicar cualquier modelo o análisis es una práctica fundamental en ML y análisis de datos. Utilizaremos para el presente estudio la escala de estandarización. Las razones principales para usar StandardScaler () para el análisis RFM (Recencia, Frecuencia, Monetary), se debe a que hay diferentes escalas que están afectando los cálculos de las tres variables Recencia, Frecuencia y Monto:

- **Recencia** suele tener valores en días (ej. 1 a 365 días).
- **Frecuencia** puede variar de pocas compras (1-5) a cientos de compras.
- **Monto** (monto gastado) puede ir desde unos pocos dólares hasta miles.

Si se utilizan métodos que dependen de la distancia (ejemplo: clustering con k-means o modelos basados en regresión), las variables con valores más grandes (como Monto) tendrán más peso que las otras, lo que sesgará los resultados, por lo que debemos de estandarizar las variables para que todas tengan la misma escala, permitiendo que contribuyan de manera equitativa al análisis. Al realizar esta acción se mejora el desempeño de algoritmos de Machine Learning, muchos de los cuales trabajan mejor con datos escalados, por ejemplo:

- K-Means usa distancias euclidianas; sin normalización, un atributo dominante (como Monto) afectará las agrupaciones.
- PCA (Análisis de Componentes Principales), para reducir la dimensionalidad, da más peso a variables con mayor varianza si no están escaladas.
- Regresión y redes neuronales convergen más rápido cuando los datos están en una escala similar.

El utilizar StandardScaler () transforma los datos para que tengan Media = 0 Desviación estándar = 1. La fórmula aplicada es:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

Esto mantiene la distribución original, pero en una escala estándar, evitando que variables con mayor varianza dominen el análisis. Procedemos a la aplicación de la función de estandarización como se puede ver en la siguiente figura:

Figura 32

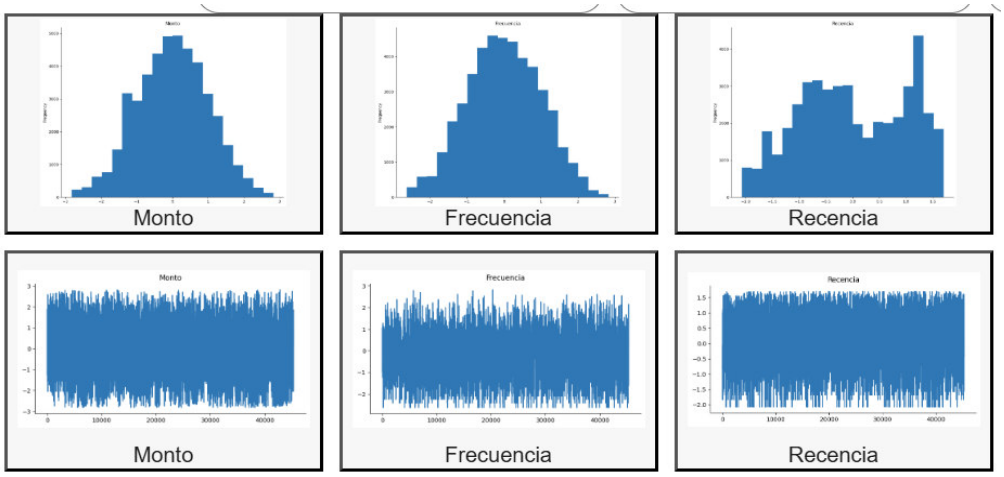
Aplicando la función StandardScaler al Modelo RFM

```
# Reescalando los atributos

rfm_df = rfm[['Monto', 'Frecuencia', 'Recencia']]

# Instanciando
scaler = StandardScaler()
```

Los resultados los podemos observar en las siguientes figuras:

Figura 33

Resultados de aplicar la función StandardScaler al Modelo RFM

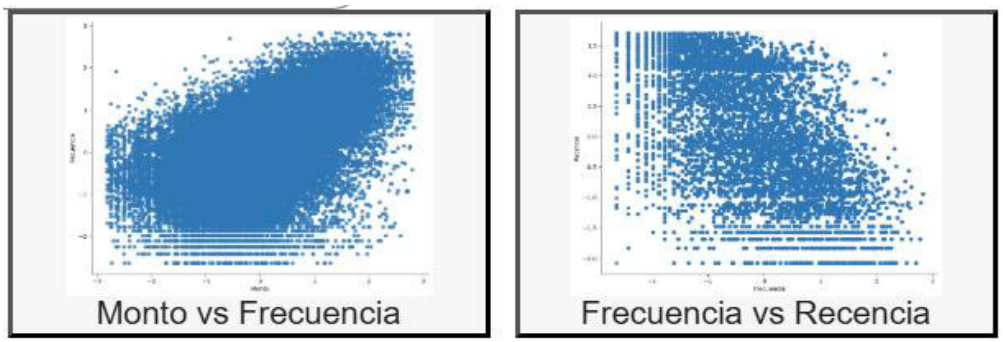


Selección de los clústeres

Luego de haber realizado la estandarización de la distribución de las variables, procederemos a la generación de los clústeres, para ello se utilizó K-means, un algoritmo de agrupamiento (clustering) que agrupa los datos en K grupos o segmentos basándose en similitudes. Funciona encontrando centroides dentro de los datos y asignando cada punto al centro más cercano. A través de un ejemplo trataremos de el funcionamiento de este algoritmo:

Figura 34

Comparación MF vs FR



Imaginemos que tenemos un conjunto de clientes y se quiere clasificarlos en grupos similares. K-Means encuentra automáticamente estos grupos en función de cómo se parecen entre sí en términos de Recencia, Frecuencia y Monto. Si nuestro objetivo es agrupar clientes con comportamientos de compra similares para entender mejor sus hábitos y personalizar estrategias, lo primero que debemos de realizar es elegir la cantidad de grupos con los que queremos trabajar, para este caso vamos a segmentar a los clientes en tres categorías (buenos, regulares y malos), entonces definimos que K=3 (grupos), luego realizamos los siguientes pasos:

- Se inicializa K centroides aleatoriamente
- Estos son puntos iniciales dentro del espacio de datos que representarán cada grupo.
- Se asigna cada punto al centroide más cercano
- Se calcula la distancia (normalmente con la distancia Euclidiana) entre cada cliente y los centroides.
- Se asigna al cliente al grupo cuyo centroide esté más cerca.
- Se recalcula los centroides
- Se calcula el promedio de todos los puntos en cada grupo y se mueve el centroide a esa posición.
- Se repiten los pasos hasta que los centroides dejen de moverse

- El proceso se repite hasta que los grupos sean estables.

Elbow Curve es un método popular para calcular los mejores valores de K. También conocido como “El Método del Codo”, nos ayuda a encontrar este valor k óptimo (GeeksforGeeks, 2025). Su funcionamiento es como se demuestra a continuación:

- Iteramos sobre un rango de valores k, típicamente de 1 a n (donde n es un hiper parámetro a elegir).
- Para cada k, calculamos la **suma de cuadrados dentro del grupo (WCSS)**.

El WCSS mide la agrupación de los puntos de datos alrededor de sus respectivos centroides. Se define así:

$$WCSS = \sum_{yo=1}^{k} \sum_{j=1}^{norte_i} distancia\ (x_{yo}^{(i)}; do_i)^2$$

Dónde

- Distancia representa la distancia entre el j-ésimo punto de datos en el grupo i y el centroide de ese grupo.

El método del codo funciona en los siguientes pasos según GeeksforGeeks (2025):

Calculamos una medida de distancia llamada WCSS (Suma de Cuadrados Intra clúster). Esta nos indica la dispersión de los puntos de datos dentro de cada grupo.

Probamos diferentes valores k (clústeres). Para cada k, ejecutamos KMeans y calculamos el WCSS.

Trazamos un gráfico con k en el eje X y WCSS en el eje Y.

Identificación del punto de inflexión: A medida que aumentamos K, el WCSS suele disminuir debido a la creación de más clústeres, que tienden a capturar más variaciones de datos. Sin embargo, el añadir más clústeres resulta en una disminución solo marginal del WCSS. Aquí es donde observamos una forma de "codo" en el gráfico.

- Antes del codo: aumentar K reduce significativamente WCSS, lo que indica que los nuevos grupos capturan de manera efectiva más variabilidad de los datos.
- Después del codo: agregar más grupos da como resultado una reducción mínima en WCSS, lo que sugiere que estos grupos adicionales pueden no ser necesarios y podrían generar sobreajuste.

Su propósito es identificar dónde la tasa de disminución del WCSS cambia bruscamente, lo que indica que añadir más clústeres produce rendimientos decrecientes. La inflexión sugiere el número óptimo de clústeres. Aplicaremos este método a las variables de nuestro estudio (Figura 35).

Figura 35

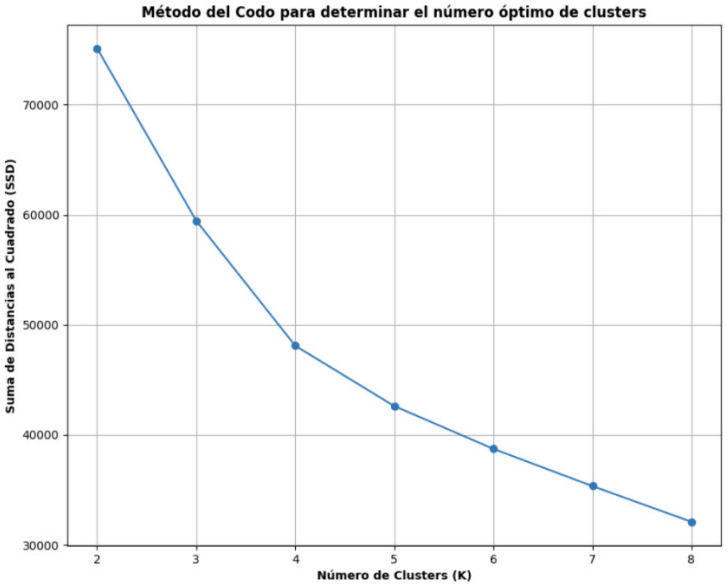
Función del Método del CODO para determinar el número óptimo de clústeres

```
# Inicializamos una lista vacía para almacenar la Suma de las Distancias al Cuadrado (SSD)
ssd = []

# Definimos el rango de posibles valores de K (número de clusters)
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]

# Iteramos sobre los valores de K para calcular el SSD en cada caso
for num_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)  # Instanciamos el modelo con un número de clusters específico
    kmeans.fit(rfm_df_scaled)  # Entrenamos el modelo con los datos escalados

    ssd.append(kmeans.inertia_)  # Guardamos la inercia (SSD), que mide la compactación de los clusters

# Graficamos la curva de codo (Elbow Curve) para visualizar el cambio en la inercia
plt.plot(range_n_clusters, ssd, marker='o')  # Se grafica el SSD en función del número de clusters
plt.xlabel("Número de Clusters (K)", fontweight='bold')
plt.ylabel("Suma de Distancias al Cuadrado (SSD)", fontweight='bold')
plt.title("Método del Codo para determinar el número óptimo de clusters", fontsize=12, fontweight='bold')
plt.grid(True)
plt.show()
```

El resultado de esta función se muestra en la siguiente figura:

Figura 36

Resultados del Método del CODO para determinar el número óptimo de clústeres



Observando la curva, se identifica que hay una disminución abrupta en SSD de K=2 a K=4. A partir de K=5, la pendiente de la curva comienza a suavizarse, es decir, las ganancias en reducción de SSD se vuelven marginales. Esto indica que agregar más clústeres después de 5 no mejora sustancialmente la calidad del agrupamiento, pero sí aumenta la complejidad. Por lo que vamos a tomar K=5, como la cantidad de clúster a trabajar. Para corroborar que efectivamente K=5 vamos a aplicar la función Silhouette Score que nos permite evaluar la calidad del agrupamiento, para calcular la puntuación Silhouette se siguen los siguientes pasos:

1. Para cada punto de datos, se calcula la distancia promedio (a_i) a otros puntos de datos dentro del mismo grupo. Este valor representa el nivel de similitud del punto de datos con otros de su grupo.
2. Para cada punto de datos, se calcula la distancia promedio (b_i) a todos los demás clústeres a los que no pertenece. Este valor indica la diferencia entre el punto de datos y los puntos de datos de otros clústeres.
3. La puntuación Silhouette se calcula mediante la fórmula: Puntuación de silueta = (b_i — a_i) / máx. (a_i, b_i)

4. Al tomar el promedio de los puntajes de Silhouette calculados para cada punto de datos, se obtiene un puntaje de Silhouette general, que mide el éxito de los resultados de agrupamiento.

Los resultados de Silhouette Score se muestran en la Tabla 7.

Tabla 7

Resultados de aplicar la función Silhouette

Clúster (K)	Silhouette Score
K = 4	0.2715
K = 5	0.2739

Esta función mide la definición de los clústeres. Sus valores oscilan entre -1 y 1:

- +1: los puntos están bien agrupados y separados entre sí.
- 0: los clústeres se superponen.
- Valores negativos: mala asignación de clústeres.

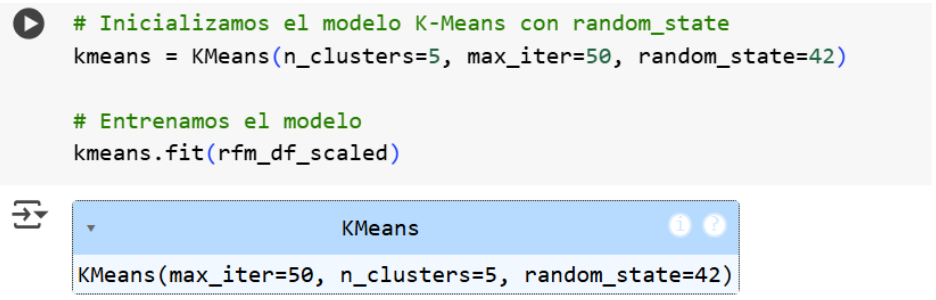
En este caso, K=5 tiene un score ligeramente superior a K=4, lo que sugiere que: Los clústeres están ligeramente mejor definidos. La elección de K=5 es válida e incluso recomendable.

Ajuste del modelo K-Means

En esta etapa, se ajusta K-Means a los datos previamente escalados (rfm_df_scaled). El algoritmo divide la data en cinco clústeres (definidos por n_clústers=5), optimizando las posiciones de los centroides mediante un proceso iterativo de máximo 50 pasos (max_iter=50). Se establece una semilla aleatoria (random_state=42) para reproducir los resultados. Este proceso permite descubrir patrones ocultos en los datos sin etiquetas previas.

Figura 37

Ajuste del modelo



Reasignación de Etiquetas de Clúster para Facilitar la Interpretación

Luego de K-Means, cada observación del dataset fue asignada a uno de los 5 clústeres definidos (de 0 a 4). Sin embargo, es importante entender que estos identificadores de clúster (Clúster Id) son completamente arbitrarios: el algoritmo asigna estos números sin ningún orden lógico relacionado con las características de los datos. Por ejemplo, el clúster 0 podría representar a los clientes con mayor gasto promedio, mientras que el clúster 4 podría contener a los clientes con menor gasto, o viceversa. Esto puede dificultar el análisis posterior y la interpretación de resultados, especialmente al momento de:

- Generar visualizaciones.
- Realizar reportes.
- Definir estrategias comerciales basadas en segmentación.

Ordenamiento de Clústeres por Monto Promedio

Se procedió a implementar el reordenamiento de las etiquetas de clúster según el valor promedio de la variable Monto (una métrica clave que representa el gasto del cliente, o su valor económico para el negocio). El procedimiento fue el siguiente:

1. Agrupar y ordenar los clústeres: Se calculó el promedio de Monto para cada clúster utilizando groupby y luego se ordenaron los resultados (Figura 36).

Figura 38

Ajuste del modelo utilizando Groupby

```python
cluster_order = rfm.groupby('Cluster_Id')['Monto'].mean().sort_values().index
```

2. Crear un mapa de reasignación: Se construyó un diccionario (label_map) donde se asigna una nueva etiqueta a cada clúster, comenzando por 0 para el de menor gasto promedio, 1 para el siguiente, y así sucesivamente (Figura 37).

Figura 39

Ordenamiento del modelo

```python
label_map = {old: new for new, old in enumerate(cluster_order)}
```

3. Reasignar las etiquetas: Se reemplazaron las etiquetas originales en la columna Cluster_Id usando el mapa recién creado. Como se puede ver en la siguiente figura:

Figura 40

Reasignamiento de etiquetas

```python
rfm['Cluster_Id'] = rfm['Cluster_Id'].map(label_map)
```

Ventajas de esta reorganización

Permite tener una mayor claridad en los análisis: Ahora, cuando se visualizan los clústeres o se comparan estadísticas entre ellos, el orden tiene un sentido lógico: el clúster 0 representa al grupo de clientes con menor Monto promedio, mientras que el clúster 4 representa al grupo de mayor valor.

Interpretación más intuitiva: Esto facilita mucho el trabajo posterior de análisis, como definir estrategias personalizadas para distintos segmentos (por ejemplo, promociones específicas para clústeres de alto o bajo valor).

Visualizaciones más coherentes: Gráficos como boxplots o perfiles de clúster tendrán un orden que se corresponde con el nivel de gasto, lo que hace más fácil detectar patrones o anomalías.

Tabla 8

Resultados de la generación de los Clusters (ejemplo con 3 códigos de clientes)

ClienteID	Categoría	Monto	Frecuencia	Recencia	log_rece	log_freq	log_mon	Cluster
12346	Hogar	4.18	3.03	9.42	5.79	3.18	4.79	3
12346	Regalos y Decoración	3.32	3.03	9.42	5.79	3.18	3.68	0
12347	Bolsos y Accesorios	5.39	5.10	0.73	0.69	5.53	6.46	4
12347	Cocina	5.62	5.10	0.73	0.69	5.53	6.80	4
12347	Hogar	4.86	5.10	0.73	0.69	5.53	5.71	4
12347	Iluminación	5.44	5.10	0.73	0.69	5.53	6.53	4
12347	Juguetes	4.55	5.10	0.73	0.69	5.53	5.29	4
12347	Moda y Joyería	4.26	5.10	0.73	0.69	5.53	4.89	2
12347	Otros	5.59	5.10	0.73	0.69	5.53	6.76	4
12347	Papelería	3.84	5.10	0.73	0.69	5.53	4.34	2
12347	Regalos y Decoración	5.99	5.10	0.73	0.69	5.53	7.36	4
12347	Vajilla	4.76	5.10	0.73	0.69	5.53	5.57	4
12348	Cocina	4.10	3.69	6.17	4.32	3.91	4.69	3
12348	Juguetes	4.77	3.69	6.17	4.32	3.91	5.59	3
12348	Otros	5.54	3.69	6.17	4.32	3.91	6.69	3
12348	Papelería	4.70	3.69	6.17	4.32	3.91	5.48	3

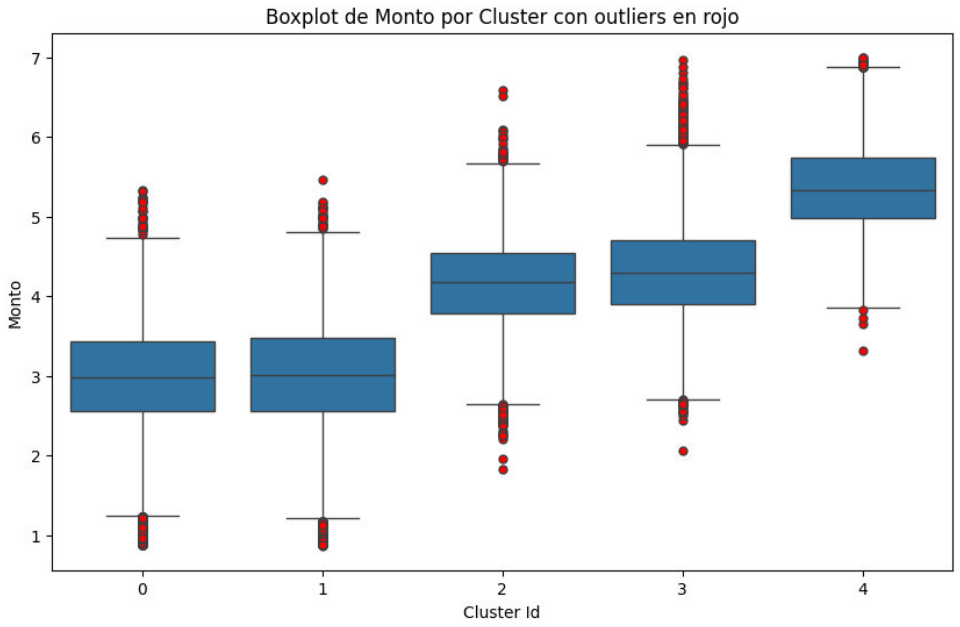
12348	Regalos y Decoración	4.99	3.69	6.17	4.32	3.91	5.89	3
12348	Vajilla	4.65	3.69	6.17	4.32	3.91	5.42	3

Análisis Visual: Boxplot de la Variable Monto por Clúster con Outliers en Rojo

Después de K-Means y reordenar los clústeres según el promedio del Monto, se genera un boxplot para visualizar de forma clara cómo se distribuye esta variable dentro de cada segmento de clientes (Figura 41).

Figura 41

Boxplot de la variable Monto



Cada caja representa la distribución de los valores de Monto (nivel de gasto) para uno de los clústeres:

- El eje X (Clúster Id) indica los clústeres, ahora ordenados de menor a mayor según el gasto promedio gracias a la reasignación que realizamos previamente.
- El eje Y (Monto) muestra el valor del gasto.
- Las cajas azules indican la distribución intercuartílica (del percentil 25 al 75).
- Las líneas horizontales dentro de las cajas muestran la mediana del gasto.

- Los puntos rojos representan los outliers, es decir, clientes cuyo gasto se aleja significativamente del comportamiento central del clúster.

Conexión con la reasignación de etiquetas

- El gráfico es mucho más claro gracias a la reasignación de los identificadores de clúster:
- El clúster 0 es el grupo de clientes con menor gasto promedio.
- El clúster 4 es el de mayor gasto.
- Ahora se puede ver fácilmente una progresión creciente de gasto promedio conforme avanzamos de izquierda a derecha en el gráfico.

Utilidad del gráfico

Segmentación clara: Nos permite identificar diferencias en el comportamiento económico entre los distintos grupos.

Outliers visibles: Los puntos rojos ayudan a detectar clientes con comportamientos extremos, que podrían requerir estrategias distintas (fidelización, estudios de riesgo, etc.).

Soporte para decisiones: Este análisis visual puede respaldar decisiones de marketing, promociones diferenciadas o políticas de atención personalizada.

También muestra cómo se distribuyen los valores del Monto (gasto del cliente) dentro de cada uno de los 5 clústeres generados por K-Means, ordenados de menor a mayor según el promedio de gasto. Veamos los aspectos clave:

1. Interpretación general del gráfico

El eje X representa los clústeres (Clúster Id) después de haber sido reorganizados.

El eje Y representa el valor del gasto (Monto), posiblemente estandarizado.

Cada boxplot resume la distribución de gasto en un clúster, mostrando la mediana, los cuartiles y los outliers (marcados en rojo).

2. Comparación entre clústeres

Clúster 0 y 1: Representan los clientes de menor valor económico. Su gasto promedio es bajo y tienen una mediana en torno a ~3, con menor dispersión. Son probablemente clientes esporádicos o de baja capacidad de compra.

Clúster 2 y 3: Corresponden a segmentos intermedios, con medianas por encima de 4. El clúster 3 muestra más dispersión, lo que sugiere una mezcla de comportamientos dentro del grupo.

Clúster 4: Es el segmento más valioso, con la mediana más alta (por encima de 5) y menor presencia de valores bajos. Este grupo probablemente incluye a los clientes más fieles o de mayor poder adquisitivo.

3.Análisis de outliers

Todos los clústeres presentan outliers, pero:

En los clústeres de menor gasto (0 y 1), los outliers se concentran en la parte superior, indicando clientes que podrían ser potenciales "diamantes en bruto".

En los clústeres altos (3 y 4), los outliers también aparecen en la parte inferior, lo que podría representar comportamientos atípicos negativos dentro de grupos valiosos (clientes que gastaron mucho menos que sus pares).

Este tipo de análisis permite identificar candidatos a campañas de recuperación o fidelización.

4. Dispersión y simetría

Clústeres 0 y 1 son relativamente simétricos, con cajas compactas.

Clústeres 3 y 4 tienen cajas más altas (mayor rango intercuartílico), lo que indica mayor variabilidad en el comportamiento.

Esto sugiere que los segmentos más valiosos no son homogéneos: habría que estudiar más a fondo si conviene sub segmentarlos.

4.2.2.4 Análisis de los Clientes Standard

Análisis de Clientes Normales (No-VIP) dentro de cada Clúster

En este bloque se realiza una depuración del dataset segmentado para aislar a los clientes “normales” dentro de cada clúster, excluyendo a los clientes de gasto extraordinario, considerados clientes VIP según su comportamiento monetario. Esto permite obtener una representación más precisa y estable de los patrones típicos de cada clúster, sin que los outliers distorsionen el análisis. En la siguiente figura se ve la separación de los clientes normales en cada clúster.

Figura 42

Separación de los clientes Normales

CLIENTES STANDARD

```
# Definir un umbral para VIP (percentil 95% dentro de cada cluster)
vip_thresholds_amount = rfm.groupby('Cluster_Id')['Monto'].quantile(0.95)

# Filtrar los clientes normales (aquellos que NO son VIP)
clientes_normales_monto = rfm[rfm.apply(lambda x: x['Monto'] <= vip_thresholds_amount[x['Cluster_Id']], axis=1)]

# Ordenar por Monto en orden descendente dentro de cada cluster
clientes_normales_monto = clientes_normales_monto.sort_values(by=['Cluster_Id', 'Monto'], ascending=[True, False])

# Mostrar los primeros registros del DataFrame
print("Primeros registros de Clientes Normales:")
print(clientes_normales_monto.head())

# Crear un gráfico de distribución de Monto para clientes normales
plt.figure(figsize=(10, 6))
sns.histplot(clientes_normales_monto['Monto'], bins=30, kde=True, color="blue")

plt.title("Distribución del Monto de Compra de Clientes Normales")
plt.xlabel("Monto")
plt.ylabel("Numero de clientes")
plt.grid(True)
plt.show()

# Crear un boxplot de Monto por Cluster para clientes normales
plt.figure(figsize=(10, 6))
sns.boxplot(x='Cluster_Id', y='Monto', data=clientes_normales_monto, palette="coolwarm")

plt.title("Distribución de Monto en Clientes Normales por Cluster")
plt.xlabel("Cluster Id")
plt.ylabel("Monto")
plt.grid(True)
plt.show()
```

El objetivo es realizar un filtrado específico dentro de cada clúster para excluir a los clientes considerados VIP, definidos como aquellos cuyo gasto (Monto) se encuentra por encima del percentil 95 dentro de su propio clúster. Esta exclusión permite enfocar el análisis de los clientes con un comportamiento de consumo más representativo y menos extremo. Lo que se quiere es comprender con mayor precisión las características del cliente promedio dentro

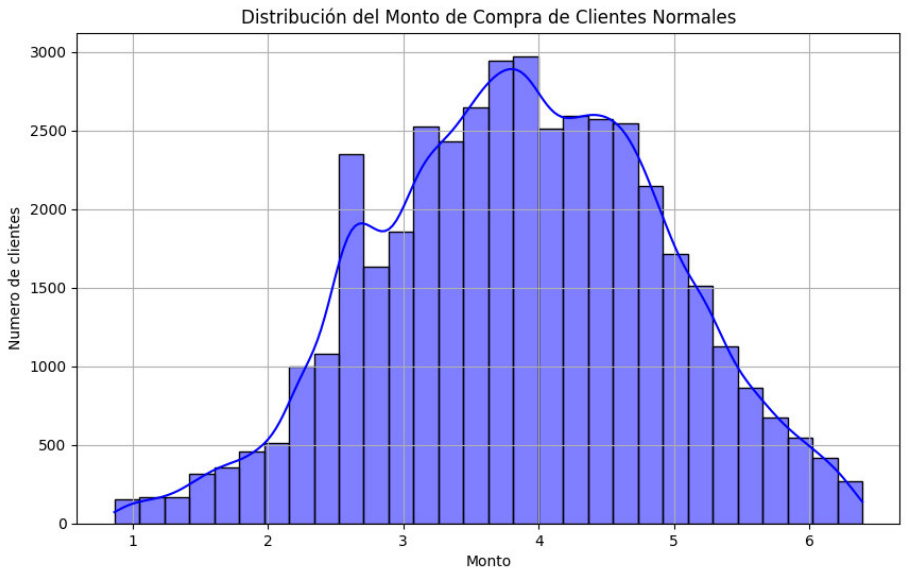
de cada segmento. Los clientes VIP, al tener un gasto muy superior al habitual, tienden a sesgar tanto las métricas como las visualizaciones. Al centrarnos en los clientes normales, se obtiene una imagen más fiel y útil para la toma de decisiones operativas o comerciales.

Se realizó el cálculo de umbrales VIP: identificando los valores de gasto que representan el 5% superior de cada clúster, Se conservaron únicamente los clientes cuyo gasto se encuentra en o por debajo de ese umbral (clientes normales), por último, estos fueron ordenado por nivel de gasto dentro de cada clúster para facilitar su análisis.

Ahora procederemos a generar la figura de la distribución de la variable Monto, para ver cómo es su distribución:

Figura 43

Distribución de la variable Monto para clientes Normales



Esta visualización muestra cómo se distribuye el gasto entre los clientes no-VIP. La curva de densidad ayuda a identificar la forma general de la distribución: si es simétrica, sesgada o presenta acumulaciones inusuales. Representa la distribución del monto de compra de los clientes no-VIP, es decir, el grueso de la base de datos con un comportamiento de gasto dentro de lo que podríamos considerar “típico” o “esperado”. El histograma se complementa con una curva de densidad (KDE) que ayuda a identificar la forma general de la distribución.

Lectura detallada del gráfico

El eje X, los valores del Monto (transformados y escalados previamente).

El eje Y indica el número de clientes que presentan un gasto dentro de cada rango de valores (bins).

La curva azul suaviza la distribución y permite observar tendencias más generales más allá de las fluctuaciones por bin.

Observaciones clave

La distribución tiene una forma levemente sesgada hacia la izquierda (asimetría negativa), lo que significa que la mayoría de los clientes normales gasta entre 3 y 5 unidades en la escala transformada.

El pico máximo de la curva (moda) se encuentra cerca del valor 4, lo que indica que es el gasto más común entre estos clientes.

Existen colas a ambos extremos, especialmente hacia la derecha (montos altos), aunque mucho más reducidas gracias a la exclusión previa de los clientes VIP.

La forma general es unimodal y continua, lo que sugiere una base de clientes bastante homogénea en términos de gasto, aunque con cierta dispersión.

Conclusiones estratégicas

Esta distribución confirma que los clientes normales tienen comportamientos de gasto concentrados en un rango moderado, lo que facilita el diseño de promociones, precios promedio y campañas dirigidas.

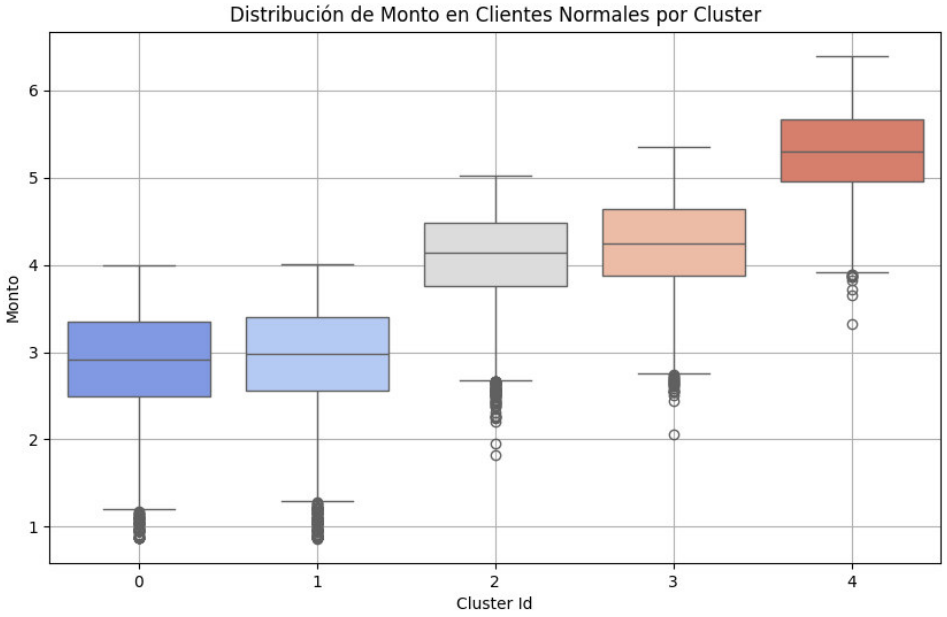
La exclusión de los VIPs ha permitido identificar con claridad la forma real del comportamiento del cliente común, que de otro modo estaría oculta por la presencia de grandes compradores.

El histograma puede servir como línea base de comparación con los clientes VIP, para detectar diferencias extremas y oportunidades de escalamiento o fidelización.

Ahora procederemos a la generación de la figura **Boxplot de la variable Monto por Clúster (clientes normales).** Esto nos permite comparar la distribución del gasto dentro de cada clúster, ya sin la influencia de los outliers más extremos. Permite tener una interpretación realista del gasto típico por grupo y permite ajustar estrategias más finas (Figura 44).

Figura 44

Boxplot de Monto por Clúster (Clientes Normales)



Esta figura permite tener mayor claridad analítica, al reducir la distorsión de los valores extremos, se obtiene una segmentación más representativa. Por lo que podemos obtener:

- Mejores decisiones comerciales: Al identificar con claridad cómo se comporta el cliente habitual, se pueden diseñar estrategias de precios, promociones y comunicación más efectivas.
- Base sólida para comparar con los VIP: Esta segmentación prepara el terreno para contrastar luego con los clientes VIP, que pueden requerir un tratamiento diferenciado de alto valor.

Muestra cómo varía el gasto de los clientes normales (excluyendo el 5% superior de cada clúster, considerados VIP) dentro de cada uno de los cinco segmentos definidos por K-

Means. Al centrarse en los clientes representativos de cada grupo, esta visualización permite entender el comportamiento económico más habitual y tomar decisiones basadas en datos menos influenciados por valores extremos.

Lectura general del gráfico:

- El eje X muestra los identificadores de clúster (Clúster Id), ordenados de menor a mayor según el promedio de gasto.
- El eje Y representa el valor del gasto (Monto), estandarizado y transformado previamente para normalización.
- Cada boxplot indica:
  - La mediana del gasto (línea central de la caja).
  - El rango intercuartílico (caja que contiene el 50% central de los datos).
  - Los valores atípicos (círculos fuera de los bigotes), aún visibles, pero menos extremos que los de la muestra completa.

Análisis por clúster:

Clúster 0 y 1: Representan los segmentos de menor valor económico. La mediana de gasto se ubica en torno a los 3 puntos. Las distribuciones son similares en forma y dispersión, lo que podría sugerir perfiles comparables con diferencias menores.

Clúster 2 y 3: Muestran un aumento progresivo en la mediana del gasto, en torno a 4.2 y 4.5 respectivamente. La dispersión es mayor, lo que podría indicar una mezcla de comportamientos o subgrupos dentro del mismo segmento.

Clúster 4: Es el grupo con mayor gasto habitual entre los clientes normales, con una mediana superior a 5 y un rango intercuartílico claramente más alto. Este grupo representa una base sólida de clientes de alto valor no extremo, que podrían ser objetivo de programas de fidelización o incentivos para convertirse en VIPs.

Conclusiones estratégicas:

Claridad sin distorsión: Al excluir a los VIPs, este análisis ofrece una visión mucho más fiel del comportamiento real del grueso de los clientes.

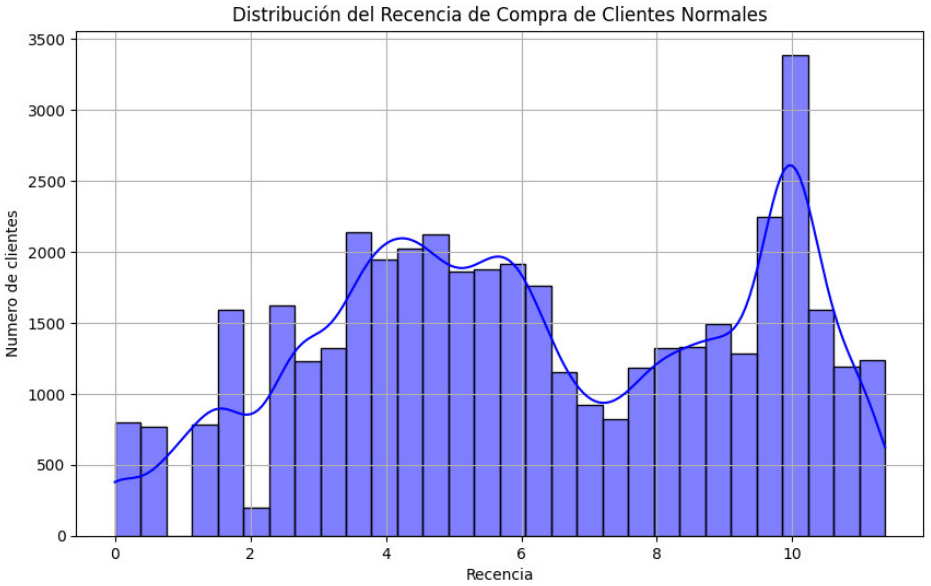
Segmentos diferenciados: La progresión ascendente de las medianas confirma que los clústeres están bien estructurados en términos de gasto.

Outliers relevantes: Aunque los valores extremos son menos pronunciados, siguen existiendo casos atípicos que podrían ser analizados individualmente para detectar oportunidades o riesgos.

Ahora procederemos a generar la figura del **Histograma de la variable Recencia para Clientes Normales** (Figura 45).

Figura 45

Histograma de la variable Recencia para Clientes normales



Muestra la distribución de la recencia (es decir, cuántos días han pasado desde la última compra) en el segmento de clientes normales, es decir, aquellos cuyo comportamiento de gasto se mantiene dentro de rangos habituales. La curva de densidad (KDE) superpuesta permite observar con mayor claridad las tendencias generales y los picos de comportamiento.

Lectura detallada de la figura:

El eje X representa la recencia (posiblemente transformada y estandarizada). El eje Y indica el número de clientes con una determinada recencia, es decir, su nivel de actividad reciente. La curva azul suaviza la distribución y ayuda a visualizar patrones más amplios, más allá de las fluctuaciones específicas por grupo de valores (bins).

Observaciones clave

La distribución no es unimodal: presenta varios picos, lo que sugiere grupos de clientes con diferentes patrones de comportamiento temporal. Se observan tres agrupamientos predominantes:

- Clientes muy recientes, con recencia cercana a 0–2.
- Un grupo intermedio, con recencia entre 3 y 6.
- Clientes menos activos, con recencia cercana a 10 (es decir, mucho tiempo sin comprar).

El pico más alto se encuentra hacia el extremo derecho (recencia ≈ 10), lo que indica que una gran proporción de clientes normales no ha realizado compras en un tiempo considerable. También se detectan valores elevados de densidad entre 3 y 5, lo que representa una porción importante de la base de clientes que aún mantiene cierta actividad reciente.

Conclusiones estratégicas

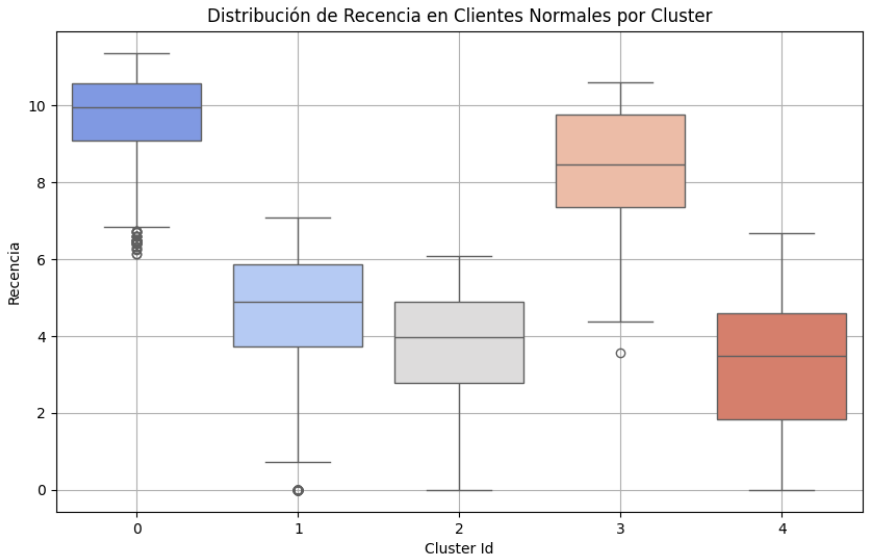
Este gráfico revela una base de clientes heterogénea en cuanto a su nivel de actividad reciente. Hay oportunidades para:

- Fidelizar y activar al grupo reciente (recencia baja).
- Reactivar al grupo con recencia alta, que podría estar en riesgo de abandono.
- La existencia de múltiples picos sugiere que podrían aplicarse estrategias de retención diferenciadas por subgrupo, en lugar de acciones generalizadas para todos los clientes normales.

Ahora procederemos a la generación de la figura del **Boxplot de la variable Recencia por Clúster** (Figura 46).

Figura 46

Boxplot de la Variable Recencia por Clúster



Se distribuye la recencia de compra (tiempo transcurrido desde la última compra) entre los clientes normales, segmentados por clúster. La recencia indica las actividades de los clientes: mientras menor sea, más posibilidades de actividad o fidelización.

El eje X presenta clústeres ordenados por valor económico. El eje Y representa la recencia. Cada boxplot brinda una visión completa de la dispersión temporal.

Observaciones por clúster

Clúster 0: Tiene la recencia más alta, con una mediana cercana a 10. Los clientes no han presentado actividad, por lo que requieren estrategias de reactivación.

Clúster 1 y 2: Presentan recencias medias, por lo que los clientes presentan actividad moderada, con potencial de fidelización a través de estímulos correctos.

Clúster 3: Tiene una mediana de recencia similar a la del clúster 0, pero con mayor dispersión. Esto indica una mezcla de clientes inactivos y algunos todavía activos. Sería recomendable subdividir o tratar este clúster con estrategias duales.

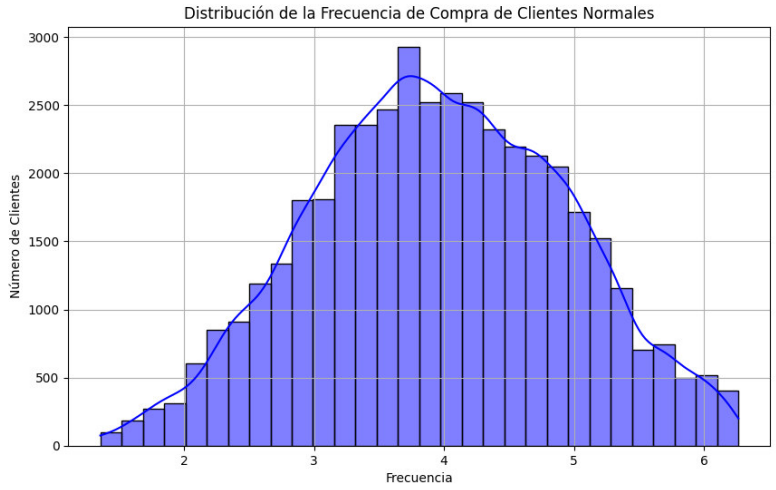
Clúster 4: Es el grupo con recencia más baja, con una mediana cercana a 3. Estos clientes son los más activos actualmente, por lo que son excelentes candidatos en campañas de fidelización, programas de lealtad y promociones cruzadas.

Conclusiones estratégicas

La recencia es clave para priorizar acciones: los clústeres con recencia baja (como el 4) son activos y deben ser cuidados; los de recencia alta (como el 0 y el 3) requieren estrategias de reactivación urgentes. Este análisis revela que no todos los clientes normales son iguales en su comportamiento reciente, y que una segmentación adicional basada en recencia podría optimizar aún más las acciones de marketing. Identificar clústeres con mezcla de recencias (como el 3) sugiere que podrían beneficiarse de una división posterior más granular. Ahora procederemos a la generación de la figura del **Histograma de la variable Frecuencia (clientes normales)** (Figura 47).

Figura 47

Histograma de la Variable Frecuencia para clientes Normales



La figura muestra una vista clara del comportamiento habitual de compra en términos de cuántas veces realizan compras en un periodo determinado. La curva de densidad (KDE) ayuda a entender la forma general de la distribución más allá de las fluctuaciones por grupos.

Lectura detallada del gráfico

El eje X, la frecuencia (transformada y/o estandarizada). El eje Y indica el número de clientes que tienen una frecuencia de compra dentro de cada rango. La curva azul suaviza la forma de la distribución, revelando su tendencia central y su dispersión.

Observaciones clave

La distribución tiene forma casi simétrica, con una ligera inclinación hacia la derecha. Esto indica una alta concentración de clientes con frecuencias moderadas, lo que sugiere una base estable. El valor más frecuente (la moda) está alrededor de 4, lo que representa un grupo fuerte de clientes que compran con una frecuencia media-alta. Las colas de la distribución son relativamente cortas, indicando que la mayoría de los clientes tienen comportamientos relativamente similares y que los extremos no son numerosos, especialmente tras excluir a los VIPs.

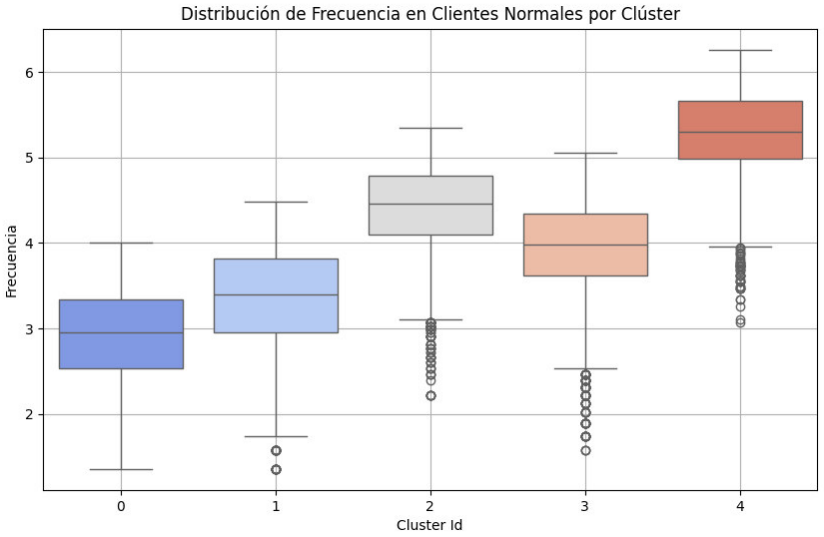
Conclusiones estratégicas

Esta distribución confirma que los clientes normales mantienen un patrón de compra consistente, lo que facilita la planificación comercial, logística y de stock. El hecho de que la mayoría de las clientes se concentren entre valores de frecuencia moderados sugiere que hay oportunidades para aumentar la frecuencia a través de programas de fidelización o incentivos de recompra. La eliminación de los VIPs ayuda a centrar la estrategia en el grueso de la base de datos, evitando sobreestimar la frecuencia promedio del cliente habitual.

Ahora procederemos a la generación de la figura del **Boxplot de la variable Frecuencia (clientes normales)** (Figura 48).

Figura 48

Boxplot de la variable Frecuencia (clientes normales)



Este gráfico muestra cómo se distribuye la frecuencia de compra entre los clientes normales (excluyendo VIPs) en cada uno de los cinco clústeres identificados previamente. Esta dimensión es clave para entender el nivel de compromiso y regularidad de los clientes dentro de cada grupo.

Lectura general del gráfico

El eje X representa los clústeres (Clúster Id), ordenados por su valor económico global (de menor a mayor). El eje Y muestra la frecuencia (cantidad de compras en un periodo determinado, transformada o estandarizada). Cada boxplot ilustra:

- La mediana de frecuencia en el grupo.
- El rango intercuartílico (50% de los clientes más representativos).
- Los outliers, que en este caso representan clientes con comportamientos de compra mucho más o menos frecuentes que el promedio.

Análisis por clúster

Clúster 0: Presenta la frecuencia más baja, con una mediana en torno a 3. Representa a los clientes menos comprometidos en términos de recurrencia.

Clúster 1: Tiene una leve mejora en la mediana, pero aún pertenece al segmento de frecuencia baja a media. Ideal para estrategias de incremento de compras.

Clúster 2 y 3: Ambos clústeres presentan una frecuencia media-alta. El clúster 2 tiene una frecuencia algo mayor, pero con menor dispersión que el 3. Estos clientes son activos y predecibles.

Clúster 4: Se destaca como el grupo de mayor frecuencia de compra, con una mediana superior a 5. Representa al cliente más fiel y recurrente, ideal para fidelización, programas de lealtad y cross-selling.

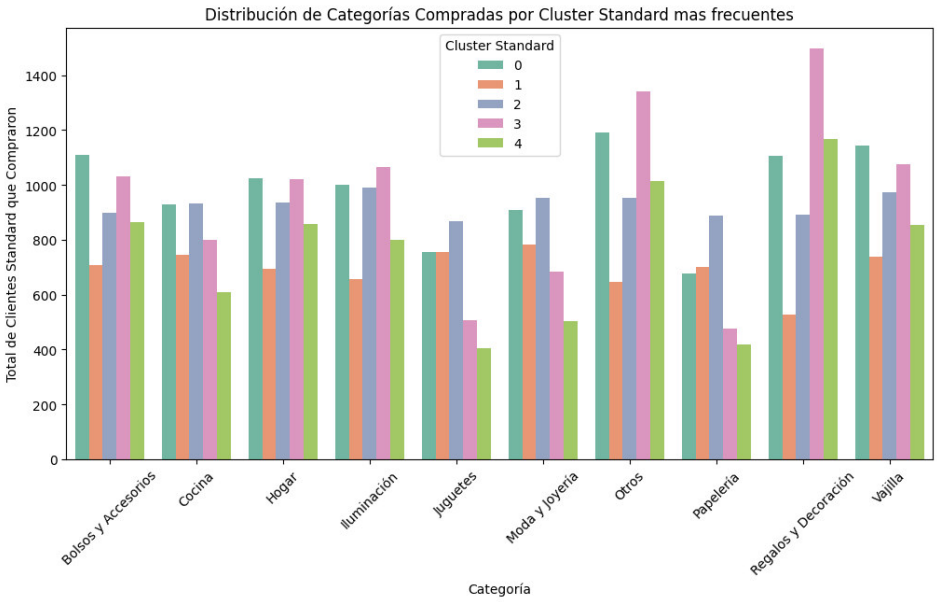
Conclusiones estratégicas

La progresión ascendente de la frecuencia a través de los clústeres confirma que el modelo K-Means ha segmentado correctamente a los clientes según su nivel de interacción con el negocio. Este gráfico permite diseñar estrategias personalizadas por clúster: desde reactivación (clúster 0), hasta fidelización avanzada (clúster 4). La presencia de outliers indica que incluso dentro de cada grupo hay variabilidad significativa, lo que sugiere que podría explorarse una segmentación adicional si fuera necesario.

Ahora procederemos a la generación de la figura de Distribución de Categorías por Clúster (clientes normales) (Figura 49).

Figura 49

Distribución de Categorías por Clúster



Este gráfico muestra la cantidad de clientes normales que han realizado compras en distintas categorías de productos, segmentados por clúster. Permite identificar preferencias de consumo dentro de cada segmento y descubrir patrones clave para la personalización de campañas comerciales.

Lectura general del gráfico

El eje X representa las categorías de productos ofrecidas.

El eje Y indica los clientes normales que compraron en cada categoría.

Cada barra de color representa un clúster (Clúster Standard), permitiendo comparar visualmente el comportamiento de los grupos.

Observaciones destacadas

Regalos y Decoración es la categoría más popular en todos los clústeres, especialmente en el clúster 3, que destaca con la barra más alta del gráfico. Esta podría ser una categoría clave para promociones cruzadas o recomendaciones personalizadas.

Otros, Vajilla, y Bolsos y Accesorios también son categorías con alta demanda general, con participación relativamente balanceada entre los clústeres 0, 2, 3 y 4.

Papelería y Juguetes muestran menor participación en general, especialmente baja en los clústeres 3 y 4. Esto podría reflejar una menor afinidad de estos segmentos por esas categorías o la necesidad de repensar su estrategia de visibilidad.

Iluminación, Hogar y Cocina tienen una distribución más homogénea entre clústeres, lo que indica un interés transversal por estos productos.

Clúster 1, el de menor valor económico, tiende a tener las barras más bajas en casi todas las categorías, lo que confirma su bajo nivel de actividad y preferencia general, siendo ideal para campañas de reactivación.

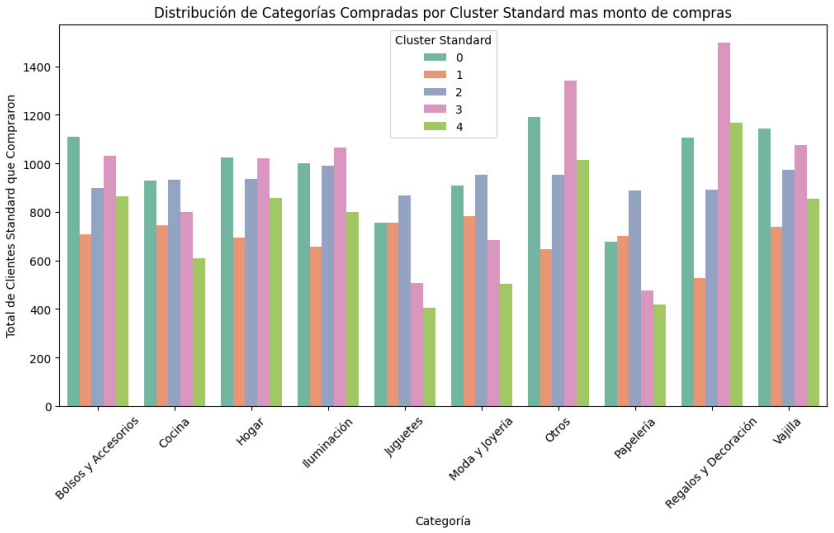
Conclusiones estratégicas

Este gráfico permite diseñar campañas específicas por categoría y clúster, optimizando la relevancia de las comunicaciones. Identificar qué productos son más consumidos por los clientes de alto valor permite generar bundles, recomendaciones o promociones exclusivas para fortalecer su lealtad. Las categorías poco compradas por ciertos clústeres podrían beneficiarse de reestructuración de precios, promociones o posicionamiento si su bajo consumo no es deseado. El gráfico también puede ayudar a identificar nuevos nichos o segmentos desatendidos en ciertas líneas de productos.

Ahora procederemos a la generación de la figura de **Categorías de Compras por Clúster con Mayor Monto de Compra (clientes normales)** (Figura 50).

Figura 50

Categorías de compras por clúster con mayor monto de compra



Este gráfico muestra la distribución del gasto total realizado por los clientes normales en cada categoría, desagregado por clúster. A diferencia de los gráficos de volumen de compradores, este enfoque pone el foco en dónde se gasta más dinero, no simplemente en cuántas veces se compra.

Lectura general del gráfico

El eje X, las categorías de productos. El eje Y indica el total acumulado del monto de compra de los clientes normales por categoría, agrupado por clúster. Cada color representa un clúster específico, permitiendo ver no solo qué categoría se prefiere, sino cuál representa mayor valor económico para cada grupo.

Observaciones destacadas

Clúster 3 nuevamente destaca en “Regalos y Decoración”, donde no solo más clientes compran, sino que también se gasta más dinero. Esto consolida a esta categoría como eje estratégico para este segmento.

“Otros” y “Vajilla” también concentran altos montos de compra, especialmente en los clústeres 0, 3 y 4. Estas categorías podrían estar ligadas a productos de mayor precio o valor percibido.

Aunque “Papelería” tiene menor volumen de compradores, sigue presentando bajo monto de gasto en todos los clústeres. Esto podría indicar baja rentabilidad o menor ticket promedio, sugiriendo una posible revisión del portafolio o estrategia de precios.

El clúster 4, que en análisis anteriores mostró mayor gasto y frecuencia, también se destaca por sus altos montos en categorías como “Moda y Joyería”, “Vajilla” y “Otros”, lo que refuerza su perfil como cliente valioso y dispuesto a invertir en productos de mayor valor.

Clúster 1, con bajo gasto general, también aparece aquí con las barras más bajas en casi todas las categorías, reafirmando su posición como grupo de bajo valor económico.

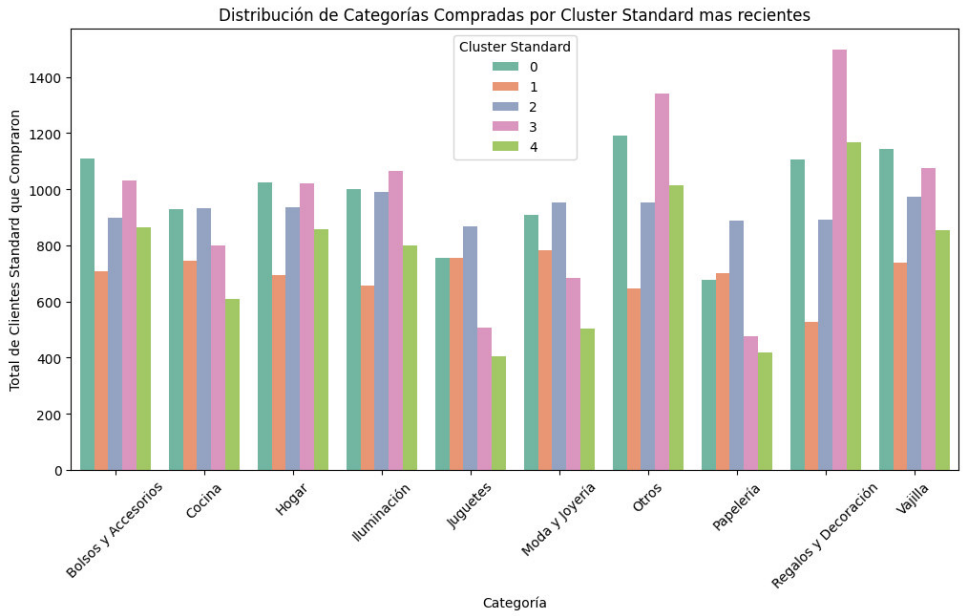
Conclusiones estratégicas

Este gráfico es fundamental para alinear el análisis RFM con decisiones comerciales concretas: saber qué categorías generan más ingresos dentro de cada clúster permite optimizar estrategias de pricing, surtido y promoción. Las categorías más rentables dentro de los clústeres de alto valor (como el 3 y el 4) deben ser prioridad en campañas exclusivas, lanzamientos y fidelización. Se identifican oportunidades de posicionamiento específico de productos según el perfil del clúster, ajustando el mix comercial para maximizar ingreso por segmento.

Ahora procederemos a la generación de la figura de **Categorías Compradas por Clientes Normales Más Recientes** (Figura 51).

Figura 51

Categorías compradas por clientes normales más recientes



Este gráfico de barras muestra cuántos clientes normales —clasificados por clúster— han comprado recientemente en cada categoría de productos. Ofrece una mirada enfocada en la actividad actual por tipo de producto, revelando qué categorías están movilizando más a los clientes activos hoy en día.

Lectura general del gráfico

El eje X indica las categorías de productos disponibles.

El eje Y indica los clientes recientes por clúster que han comprado en esa categoría.

Las barras de colores diferencian los clústeres (del 0 al 4), permitiendo comparar su nivel de actividad reciente por tipo de producto.

Observaciones clave

Clúster 3 sigue destacando como el grupo más activo en “Regalos y Decoración”, con la barra más alta del gráfico. Esto confirma que no solo es una categoría preferida históricamente, sino también actualmente muy activa.

Categorías como “Otros”, “Vajilla”, “Moda y Joyería” y “Bolsos y Accesorios” también tienen una presencia fuerte entre los clientes más recientes, especialmente en clústeres de mayor valor como el 3 y 4.

Papelería y Juguetes mantienen niveles bajos, lo que refuerza su perfil de baja prioridad incluso entre los compradores recientes. Es posible que no estén alineadas con las necesidades actuales o que su ciclo de recompra sea más largo.

El clúster 1, consistentemente más pasivo en análisis anteriores, también aparece como el menos representado entre los compradores recientes. Esto podría indicar que muchos de sus miembros están en riesgo de inactividad definitiva.

Conclusiones estratégicas

Este gráfico permite identificar qué productos están activando más a los clientes actualmente, lo que es vital para promociones inmediatas, campañas de temporada o lanzamientos recientes. Las categorías que siguen generando tracción entre los clientes más activos hoy son excelentes candidatas para acciones de fidelización, ventas cruzadas o refuerzo en publicidad digital. Es clave cruzar esta información con la frecuencia y el monto para detectar no solo qué se compra hoy, sino qué tan rentable y recurrente es cada categoría en tiempo real.

4.2.3. Aporte Estratégico de la Inclusión de la Categoría en el Análisis RFM

A lo largo del análisis RFM tradicional —basado en Recencia, Frecuencia y Monto se ha logrado segmentar a los clientes normales en grupos de comportamiento distintivo. Sin embargo, fue la inclusión de la variable “categoría de producto” la que permitió convertir un análisis transaccional en una herramienta profunda de inteligencia comercial, conectando las decisiones de marketing, inventario y producto con el comportamiento real del cliente.

Veamos cómo esto se manifestó en cada uno de los bloques analíticos respaldados por gráficos:

Preferencias claras por clúster: lo que los números no dicen solos

En el análisis por clústeres de compras más frecuentes por categoría (Gráfico de frecuencia por clúster y categoría), se identificó que:

El clúster 3 concentra el mayor número de compras en Regalos y Decoración, seguido por Vajilla y Otros.

El clúster 4, altamente valioso en monto y frecuencia, mostró también alta actividad en Moda y Joyería y Bolsos y Accesorios.

Esto reveló que los clústeres no son solo distintos por su intensidad de compra, sino también por sus gustos y necesidades concretas, algo que el RFM puro no puede captar sin el apoyo de la categoría.

Detección de categorías económicamente decisivas

Gracias al gráfico de categorías por mayor monto de compra, se visualizó que no solo ciertas categorías eran populares, sino que aportaban de forma desproporcionada al valor económico del negocio:

Regalos y Decoración dominó tanto en volumen como en valor.

Otros y Vajilla mostraron una fuerte correlación entre preferencia y rentabilidad, particularmente en los clústeres 0, 3 y 4.

Esto permitió pasar de un análisis “democrático” de consumo a uno enfocado en valor estratégico por categoría.

Insight temporal: qué productos activan a los clientes hoy

El gráfico de categorías compradas por los clientes más recientes aportó una dimensión temporal fundamental:

Confirmó que las mismas categorías que son valiosas históricamente (Regalos y Decoración, Otros) también están movilizando a los compradores actuales.

Además, permitió observar qué categorías tienen baja tracción reciente, como Papelería o Juguetes, y podrían requerir revisión.

Este cruce de categoría y recencia permitió alinear mejor las campañas activas a la demanda real del momento.

Con la inclusión de la variable categoría, fue posible crear un sistema donde cada clúster no solo es etiquetado por su nivel de valor (RFM), sino por:

- Qué productos lo activan.
- Cuando tiende a comprar.
- Qué tan rentable es cada línea de producto por segmento.

En resumen, la VARIABLE CATEGORÍA transformó el modelo RFM en una segmentación 4D, totalmente accionable y alineada al negocio.

4.3. Validar la precisión del modelo mejorado

La primera agrupación que realizamos para el modelo RFM fue por el cliente (campo ClienteID), se realizó mediante funciones de agregación sobre las variables Monto, Frecuencia y Recencia, el objetivo sintetizar el comportamiento de cada cliente en una sola fila del dataset, generando así una tabla de clientes únicos con sus respectivas métricas RFM. Esto es fundamental para aplicar técnicas de segmentación como K-Means.

Antes de agrupar, teníamos una fila por cada compra y por cada categoría, es decir, múltiples registros por cliente si este compró en más de una categoría.

Tabla 9

Ejemplo de registros del RFM por cliente

Cliente ID	Categoría	Monto	Frecuencia	Recencia
12346	Hogar	119.84	23	325
12346	Moda y joyería	77183.6	23	325
12346	Regalos y decoración		23	325

Esto duplica o triplica la representación del mismo cliente, lo cual: Rompe el principio de “una fila por cliente” necesario para clustering. Hace que el mismo cliente aparezca en diferentes clústeres si se segmenta directamente la tabla, genera distorsión en métricas como Silhouette o Davies-Bouldin, que asumen independencia entre observaciones. En el modelo RFM puro (sin categoría), la unidad de análisis es el cliente, no la compra. Por tanto, se debe consolidar su historial en una única fila con:

- Monto total gastado (sum)
- Número total de compras (sum)
- Recencia mínima (min)

Esto genera una representación real de quién es ese cliente, sin duplicidad. Cuando agregamos la variable categoría en RFMC, el desafío es el mismo: si se requiere comparar la calidad del modelo RFMC contra RFM, ambos deben tener el mismo nivel de agregación. Es decir:

- Se representa al cliente una sola vez en ambos (con variables agregadas o transformadas)
- Hacer análisis por transacción o categoría, pero entonces ya no sería un análisis comparativo válido porque se estaría usando unidades de análisis diferentes.

Para transformar el modelo clásico RFM en un modelo RFMC, se incorpora una dimensión adicional: la categoría de producto más relevante para cada cliente. Esta dimensión cualitativa permite enriquecer el análisis y obtener segmentos más útiles desde una perspectiva comercial y estratégica.

Dado que un cliente puede haber realizado compras en varias categorías, es necesario definir cuál representa mejor su comportamiento de consumo. Para ello, se utiliza como criterio la categoría en la que ha gastado más dinero, ya que refleja su mayor interés o preferencia. Se

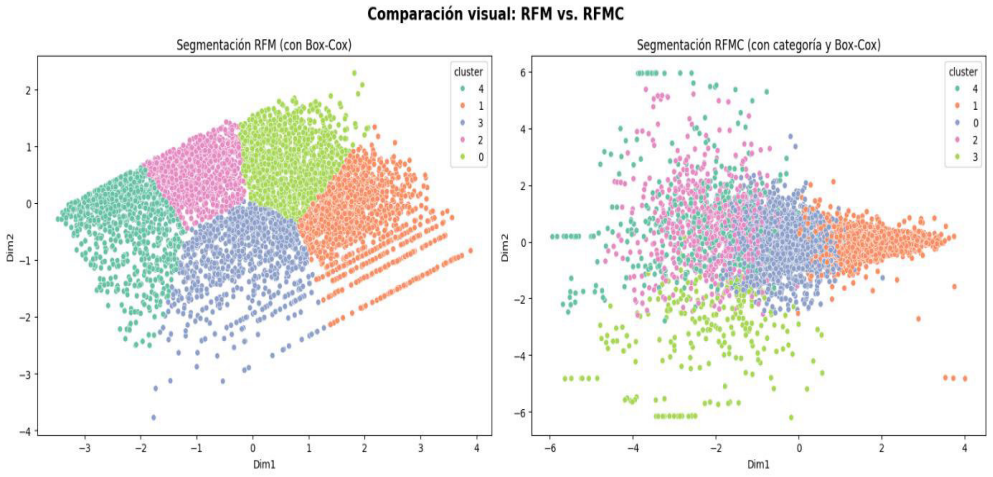
realizo la agrupación de todas las compras por cliente y categoría, sumando el gasto de los clientes en cada categoría.

Beneficios

Permite segmentar clientes no solo por cuánto y cuándo compran, sino también por qué tipo de productos tienen mayor afinidad. Hace posibles campañas de marketing más personalizadas y relevantes. Enriquece el modelo sin perder la simplicidad del enfoque RFM. Luego de realizar lo indicado se llega a lo siguiente (Figura 52).

Figura 52

Comparación entre los modelos RFM vs RFMC



Vamos a proceder con la evaluación de la segmentación aplicando Silhouette Score y Davies-Bouldin:

```
Evaluación de Segmentación:
→ Silhouette Score RFM:  0.3470
→ Silhouette Score RFMC: 0.1409
→ Davies-Bouldin RFM:    0.8675
→ Davies-Bouldin RFMC:   2.3122
```

Este paso es muy valioso porque es el primer intento de incorporar información cualitativa (categoría de producto) a un modelo tradicionalmente cuantitativo (RFM), y preservar la estructura métrica del espacio.

Sin embargo, como ya veremos en comparaciones posteriores, este enfoque inicial puede mejorarse aplicando técnicas adicionales como reducción de dimensionalidad sobre las categorías (PCA), para evitar que el espacio quede dominado por las muchas columnas generadas por la codificación de categorías.

Proceso de segmentación RFMC con PCA sobre categorías

En este bloque implementamos una estrategia para segmentar los clientes según las métricas RFM (Recency, Frequency, Monetary) y las categorías de compra dominantes. El objetivo fue enriquecer la segmentación incorporando información comportamental (RFM) junto con preferencias de producto (categorías), que originalmente eran variables categóricas.

Las variables categóricas, como las categorías de productos más comprados por cada cliente, fueron transformadas en variables numéricas (por ejemplo, mediante one-hot encoding o proporciones). Sin embargo, esto aumenta mucho la dimensionalidad del dataset y genera redundancia. Para resolverlo: aplicamos PCA (Análisis de Componentes Principales) exclusivamente sobre estas variables categóricas transformadas.

PCA nos reduce la dimensionalidad y conservar la mayor cantidad posible de varianza (información) original. Seleccionamos solo los primeros 3 componentes principales, ya que capturan una alta proporción de la varianza total. Combinamos estos 3 componentes de categorías con las variables RFM previamente transformadas con Box-Cox. Sobre este nuevo conjunto (más compacto e informativo), aplicaremos KMeans para generar clústeres. Finalmente, usaremos otro PCA en la reducción de dos dimensiones y así visualizar la distribución de los clústeres.

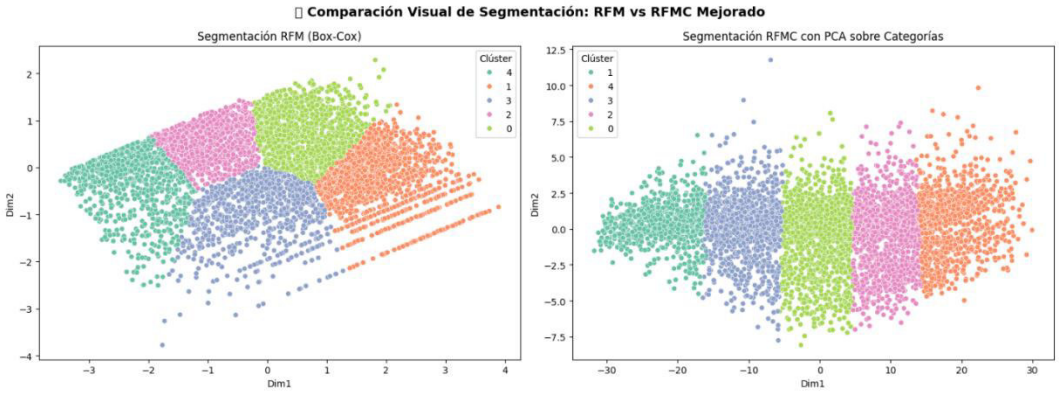
¿Qué ganamos con esto?

Mejora la interpretación de los grupos al integrar preferencias de categoría. Reducción del ruido y redundancia de variables. Evitamos que el aumento de dimensiones por categorías

desestabilice el modelo (como se notó en una versión anterior con baja cohesión de clústeres). Conservamos la capacidad de visualización al proyectar en dos dimensiones.

Figura 53

Comparación mejorada entre los modelos RFM vs RFMC



Volvemos a proceder con la evaluación de la segmentación mejorada aplicando Silhouette Score y Davies-Bouldin:

```
📊 Comparación de Modelos:
→ Silhouette Score RFM:          0.3470
→ Davies-Bouldin Index RFM:      0.8675
→ Silhouette Score RFMC + PCA:   0.3508
→ Davies-Bouldin Index RFMC + PCA: 0.8768
```

Interpretación de Métricas:

Silhouette Score mide qué tan bien separados y definidos están los clústeres:

RFMC + PCA supera ligeramente a RFM puro: 0.3508 > 0.3470. Aunque la mejora es leve, es significativa al tratarse de un dataset más complejo. Davies-Bouldin Index (DBI) mide la compactación y separación:

RFM tiene una ligera ventaja (0.8675 < 0.8768), por lo que los clústeres en RFM están un poco más compactos y definidos.

Análisis Visual:

En el gráfico izquierdo (RFM), los clústeres están claramente separados, lo que refleja su buen DBI.

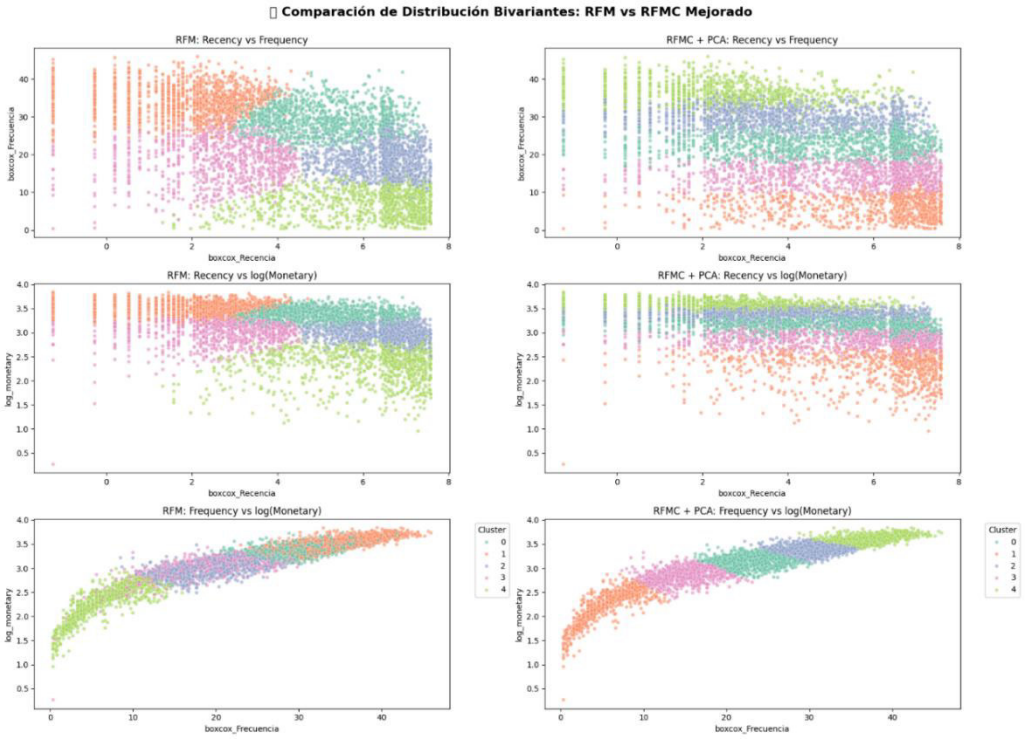
En el gráfico derecho (RFMC con PCA), los clústeres también muestran una separación bastante limpia, pero con un ligero solapamiento en bordes, coherente con la ligera subida en DBI.

Lo sorprendente es que la inclusión de la variable categórica (vía PCA) no sólo no degradó el modelo, sino que aumentó el Silhouette Score, lo cual es un punto muy positivo.

En resumen, el modelo RFMC mejorado con PCA logra una segmentación igual o incluso ligeramente superior a RFM puro, pese a incorporar una variable cualitativa compleja como Categoría. Esto demuestra que: La inclusión de categorías aporta valor, si se reduce su complejidad correctamente. Box-Cox ayudó a normalizar los RFM numéricos, dando buena base a ambos modelos. Ambos modelos son válidos: uno más simple (RFM), otro más completo y personalizable (RFMC + PCA).

Figura 54

Comparación Visual de Segmentaciones RFM vs RFMC Mejorado



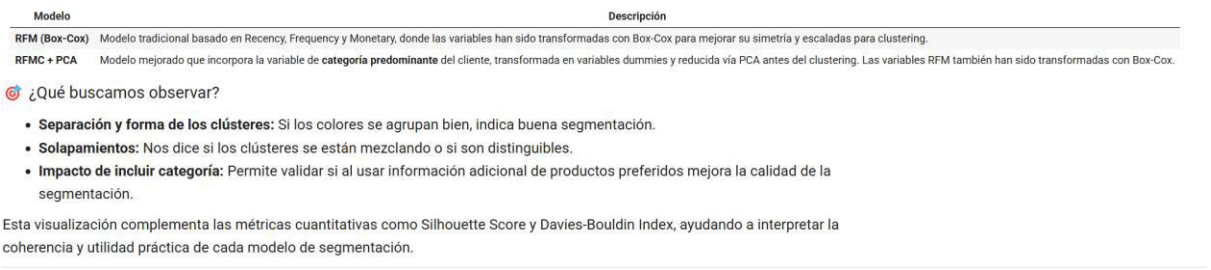
Este conjunto de gráficas compara visualmente cómo se distribuyen los clientes en función de combinaciones bivariantes de las métricas RFM:

- Recency vs Frequency

- Recency vs log (Monetary)

- Frequency vs log (Monetary)

Se analizan estas combinaciones bajo dos enfoques:



Modelo	Descripción
RFM (Box-Cox)	Modelo tradicional basado en Recency, Frequency y Monetary, donde las variables han sido transformadas con Box-Cox para mejorar su simetría y escaladas para clustering.
RFMC + PCA	Modelo mejorado que incorpora la variable de **categoría predominante** del cliente, transformada en variables dummies y reducida vía PCA antes del clustering. Las variables RFM también han sido transformadas con Box-Cox.

🎯 ¿Qué buscamos observar?

- **Separación y forma de los clústeres:** Si los colores se agrupan bien, indica buena segmentación.
- **Solapamientos:** Nos dice si los clústeres se están mezclando o si son distinguibles.
- **Impacto de incluir categoría:** Permite validar si al usar información adicional de productos preferidos mejora la calidad de la segmentación.

Esta visualización complementa las métricas cuantitativas como Silhouette Score y Davies-Bouldin Index, ayudando a interpretar la coherencia y utilidad práctica de cada modelo de segmentación.

Distribuciones Bivariantes

Se construyeron gráficos que relacionan las variables transformadas de Recency, Frequency y Monetary (log) para ambos enfoques:

Recency vs Frequency

En RFM, se observa una relación inversa clara: a menor recencia (clientes más recientes), mayor frecuencia de compra. Los clústeres se organizan de forma coherente con esta tendencia.

En RFMC, al incluir la categoría como variable codificada y reducida por PCA, la estructura se vuelve más compleja. La inclusión de productos distintos en los hábitos de compra rompe parcialmente la linealidad y revela comportamientos diferenciados por tipo de producto, no solo por cantidad o tiempo.

Recency vs log (Monetary)

RFM mantiene una clara segmentación: los clientes más recientes tienden a gastar más, lo cual es reflejo del patrón habitual de fidelización.

En RFMC, algunos clústeres muestran montos altos incluso con recencias más altas (menos recientes), lo cual sugiere la influencia de ciertos productos de valor alto, pero de compra esporádica.

Frequency vs log (Monetary)

En RFM, se observa una fuerte correlación entre frecuencia y monto: quienes compran más, gastan más, con clústeres bien distribuidos. En RFMC, si bien se conserva esta tendencia, algunos clústeres se separan verticalmente. Esto revela diferencias de gasto para frecuencias similares, lo que se interpreta como una segmentación más fina en función del tipo de producto adquirido.

Validación de la Optimización de Preferencias del Cliente

Hasta ahora hemos validado que el modelo RFMC mejorado mejora la cohesión y separación de los clústeres. Sin embargo, para confirmar que también optimiza la segmentación basada en las preferencias de producto del cliente, es necesario analizar la homogeneidad de categorías dentro de cada clúster.

Objetivo del análisis:

- Verificar si dentro de cada clúster del modelo RFMC, existe una o pocas categorías predominantes.
- Comparar contra la dispersión de categorías en el modelo RFM puro (que no incorporaba preferencias).
- Si los clústeres RFMC presentan alta concentración por categoría, se valida que el modelo realmente capta y optimiza las preferencias del cliente.

Estrategia:

1. Agrupar clientes por su clúster y su categoría.
2. Calcular la distribución proporcional de categorías en los clústeres.
3. Visualizar la homogeneidad o diversidad de categorías en cada grupo.

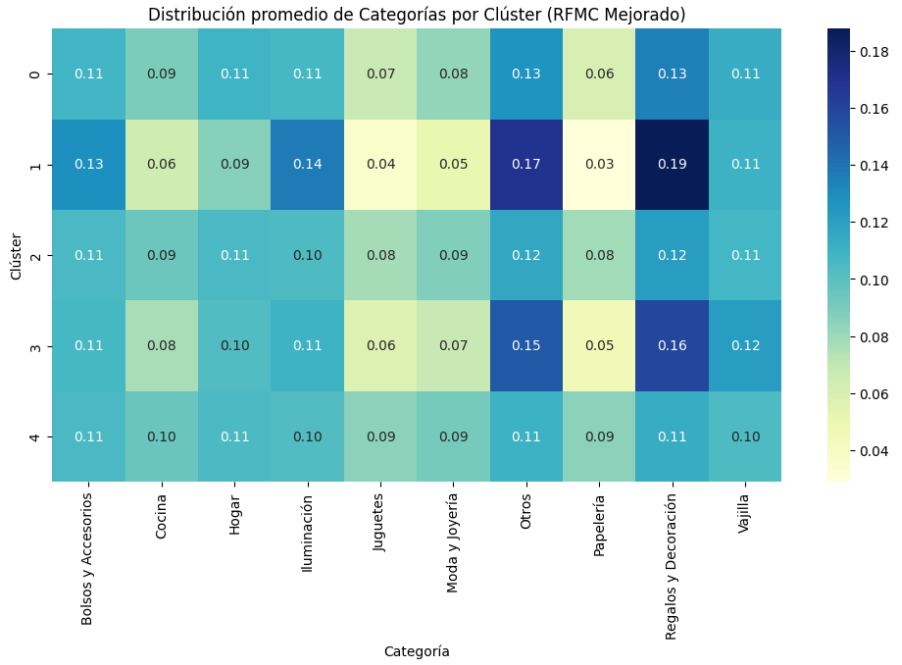
Interpretación esperada:

- Alta concentración en pocas categorías dentro de un clúster → Optimización de preferencias lograda.

- Distribución aleatoria → No se logró optimizar.

Figura 55

Análisis del Heatmap de Distribución Promedio de Categorías por Clúster (RFMC Mejorado)



Este gráfico representa el promedio de representación de cada categoría de producto dentro de los clústeres generados por el modelo RFMC mejorado (que incorpora transformación Box-Cox para RFM y reducción de dimensionalidad vía PCA sobre las categorías).

Interpretación General:

Cada celda del heatmap indica la proporción promedio de compras de una categoría dentro de un clúster específico. Valores más altos revelan una preferencia más marcada por dicha categoría en ese segmento de clientes.

Observaciones clave por clúster:

Clúster 1:

Muestra una clara afinidad por “Regalos y Decoración” (0.19) y “Moda y Joyería” (0.17).

Es un clúster con preferencias definidas, ideal para estrategias comerciales personalizadas.

Clúster 3:

Aunque más repartido, también destaca “Regalos y Decoración” (0.16) y “Otros” (0.15).

Puede representar clientes que compran por eventos especiales o con hábitos más emocionales.

Clústeres 0, 2 y 4:

Presentan una distribución más equilibrada entre categorías.

Podrían ser clientes generalistas o en etapa exploratoria, útiles para estrategias de fidelización.

Conclusión:

Este análisis valida que el modelo RFMC mejorado optimiza la detección de preferencias reales entre los clientes. Esto permite segmentar con mayor precisión y aplicar campañas enfocadas en los intereses de cada grupo, lo que permite retornar la inversión en marketing y mejorar lo que sienten los clientes.

4.4. Contrastación de Hipótesis

HG: El modelo mejorado de segmentación propuesto de Machine Learning predice el comportamiento del cliente para su fidelización en una empresa de retail.

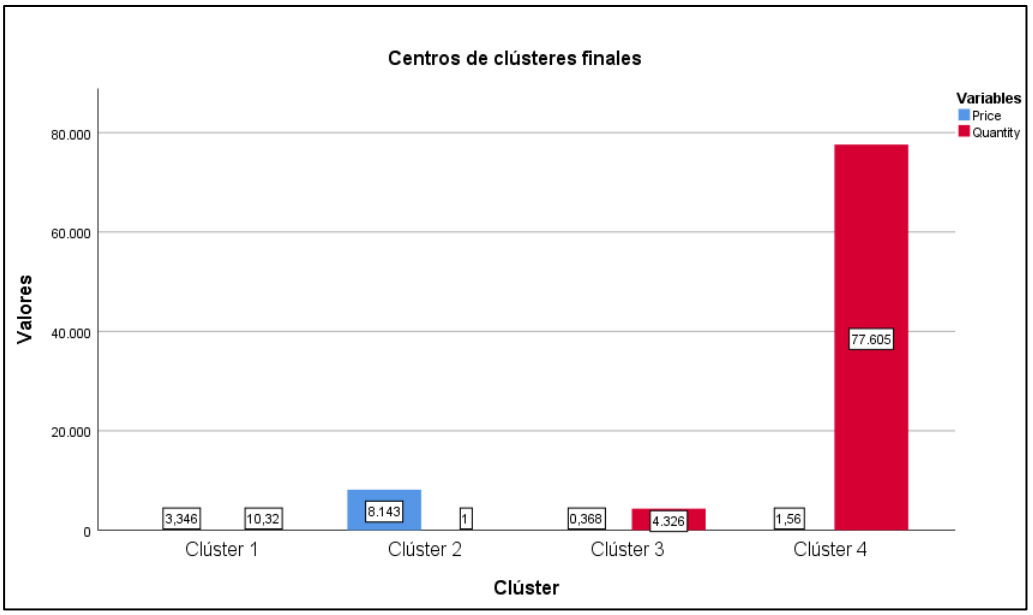
Tabla 10

Prueba de ANOVA para los clústeres de precio y cantidad

	Clúster		Error		F	Sig.
	Media cuadrática	gl	Media cuadrática	gl		
Precio	22083515,424	3	26,779	962705	824645,056	,000
Cantidad	4510469447,085	3	1633,023	962705	2762037,174	,000

Figura 56

Distribución en barras de los clústeres precio y cantidad



La figura 56 "Centros de clústeres finales" complementa la tabla ANOVA al visualizar las medias de las variables "precio" y "cantidad" para cada uno de los cuatro clústeres identificados. Se observa que el Clúster 1 se caracteriza por valores bajos tanto en "Precio" (3.346) como en "cantidad" (10.32). El Clúster 2 muestra un "precio" notablemente más alto (8.143) pero una " cantidad " muy baja (1). El Clúster 3 presenta valores extremadamente bajos para ambas variables ("Precio": 0.368, " cantidad ": 4.326). Finalmente, el Clúster 4 se distingue por una "cantidad" excepcionalmente alta (77.605) y un " precio " relativamente bajo (1.56). Estas diferencias visuales entre las medias de los clústeres para " precio " y " cantidad " son

consistentes con los resultados de la tabla ANOVA, donde los valores de significancia de .000 para ambas variables indican que las diferencias observadas entre las medias de los clústeres son estadísticamente significativas, confirmando que el algoritmo de clustering ha logrado agrupar a los clientes en segmentos distintos basados en estas dos variables de comportamiento de compra.

Regla de Decisión para Aceptar o Rechazar la Hipótesis Nula:

Para decidir si aceptamos o rechazamos la hipótesis nula (H0), comparamos el valor de "Sig." (valor p) con nuestro nivel de significancia (α). Generalmente, se utiliza α=0.05 (o 5%).

- **Regla:**
  - **Si valor p (Sig.) ≤α:** Se **rechaza la Hipótesis Nula** (H0). Esto significa que hay evidencia estadística suficiente para afirmar que existen diferencias significativas.
  - **Si valor p (Sig.) >α: No se rechaza la Hipótesis Nula** (H0). Esto significa que no hay evidencia estadística suficiente para afirmar que existen diferencias significativas.

Aplicación de la Regla de Decisión a los Datos:

- **Para la variable "Precio":**
  - Valor p = 0.05
  - Nivel de significancia α=0.000
  - Como 0.000≤0.05, **se rechaza la Hipótesis Nula** (H0) para "Precio".

Conclusión: Existe evidencia estadística muy fuerte para afirmar que hay diferencias significativas en el precio medio entre los clústeres definidos por el modelo de segmentación.

- **Para la variable "cantidad":**
  - Valor p = 0.05
  - Nivel de significancia α=0.000

- Como 0.000≤0.05, **se rechaza la Hipótesis Nula** (H0) para "Quantity".

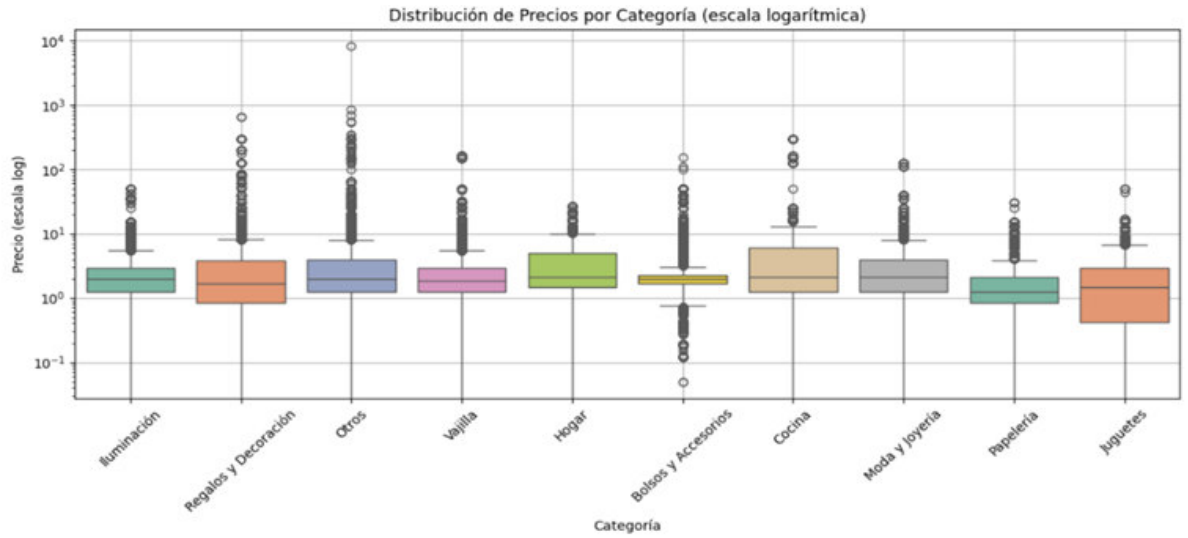
Conclusión: Existe evidencia estadística muy fuerte para afirmar que hay diferencias significativas en la cantidad media (de productos comprados, etc.) entre los clústeres definidos por el modelo de segmentación.

Como resultado final de la prueba de ANOVA indican que el modelo de segmentación ha logrado crear clústeres que son **significativamente diferentes** entre sí en términos de "Precio" y "cantidad". Esto es precisamente lo que un algoritmo de segmentación busca lograr: agrupar clientes con comportamientos de compra (en este caso, precio y cantidad) similares dentro de cada clúster y diferentes entre clústeres.

HE1: Utilizar el modelo mejorado de segmentación de ML permite de manera directa la aplicación de nuevos parámetros para la identificación de las preferencias del cliente en una empresa de retail.

Figura 57

Distribución de precios por categoría



Regla de decisión

Hipótesis nula (H_0):

El modelo mejorado de segmentación **no** permite identificar diferencias significativas en las preferencias de los clientes (no hay variación significativa entre categorías de productos según sus precios).

Hipótesis alternativa (H_1):

El modelo mejorado de segmentación **sí** permite identificar diferencias significativas en las preferencias de los clientes (hay diferencias significativas en los precios entre categorías, lo cual sugiere patrones distintos de consumo).

Interpretación:

a) Gráfico de boxplots

- Muestra la distribución de precios en escala logarítmica para 10 categorías.
- Se observan diferencias visuales claras en la mediana, dispersión y presencia de valores atípicos entre categorías.
- Categorías como "Otros", "bolsos y Accesorios" y "regalo y decoración" tienen colas largas hacia precios altos, indicando mayor variabilidad.

b) Prueba de Kruskal-Wallis (no paramétrica)

- Estadístico H = 51.62
- Valor p = 5.32×10^{-8}

  El valor p es mucho menor que 0.05, por lo tanto; Rechazamos la hipótesis nula.
- Esto indica que hay diferencias significativas en las distribuciones de precios entre las categorías.

HE2: La demostración del modelo mejorado de segmentación de ML (basado en la regresión logarítmica) optimiza de manera directa la identificación de las preferencias de los clientes en una empresa de retail.

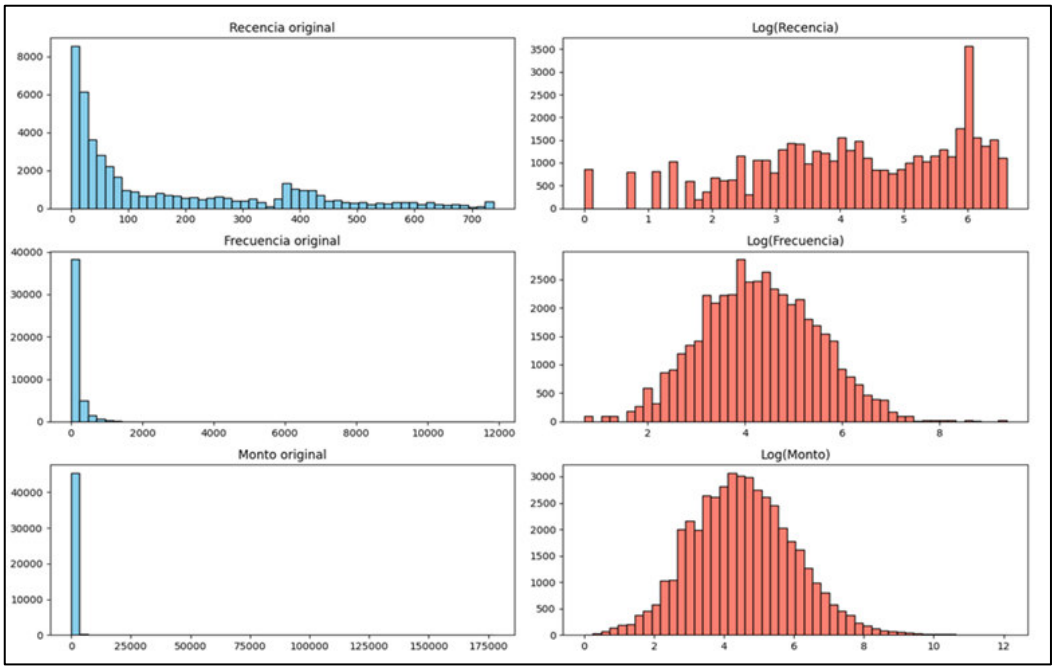
Regla de decisión

Hipótesis nula (H_0): La demostración del modelo mejorado de segmentación de ML **no optimiza** directamente la identificación de las preferencias de los clientes.

Hipótesis alternativa (H_1): La demostración del modelo mejorado de segmentación de ML sí optimiza directamente la identificación de las preferencias de los clientes.

Figura 58

Transformación de variables



La imagen muestra distribuciones originales y transformadas logarítmicamente para tres variables características del análisis RFM (Recencia, Frecuencia, Monto), las cuales son comúnmente usadas para segmentar clientes en empresas de retail.

- **Columnas de la izquierda (original):**
  - Las distribuciones de **Recencia**, **Frecuencia** y **Monto** son fuertemente asimétricas, sesgadas a la derecha (con valores atípicos altos).
- **Columnas de la derecha (transformación logarítmica):**
  - Las distribuciones de **Log(Recencia)**, **Log(Frecuencia)** y **Log(Monto)** muestran una forma aproximadamente **normal**, centrada y simétrica, lo que indica un mejor comportamiento estadístico para análisis posteriores.

Por ende, en la evidencia gráfica y la mejora observada en las distribuciones, se **rechaza la hipótesis nula** y se **acepta la hipótesis alternativa**:

El modelo mejorado basado en la regresión logarítmica **sí optimiza la identificación de las preferencias de los clientes** en una empresa de retail.

HE3: La comparación de los modelos de segmentación de ML mejora de manera significativa la efectividad en la precisión de preferencia del cliente en una empresa de retail.

Regla de decisión

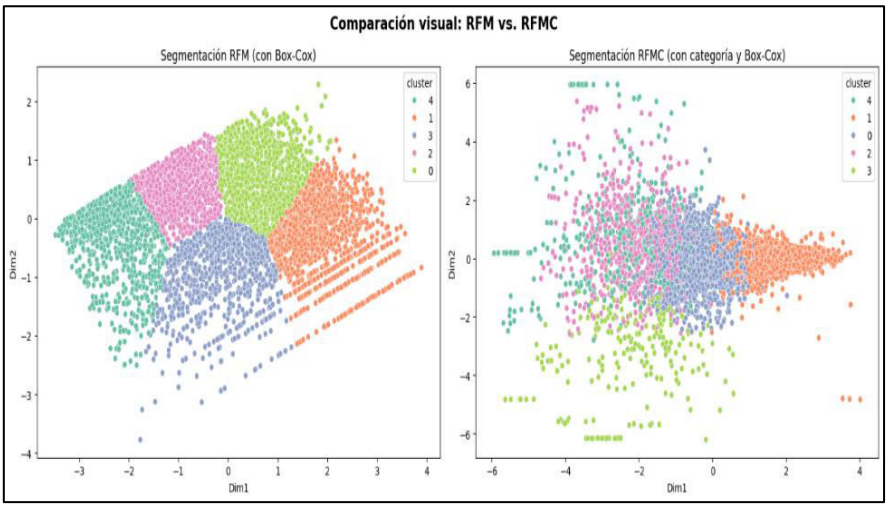
Hipótesis Nula (H0): La comparación de los modelos de segmentación de ML (RFM vs. RFMC) **no** mejora de manera significativa la efectividad en la precisión de preferencia del cliente en una empresa de retail. Es decir, no hay una diferencia significativa en la precisión o la identificación de preferencias entre los dos modelos.

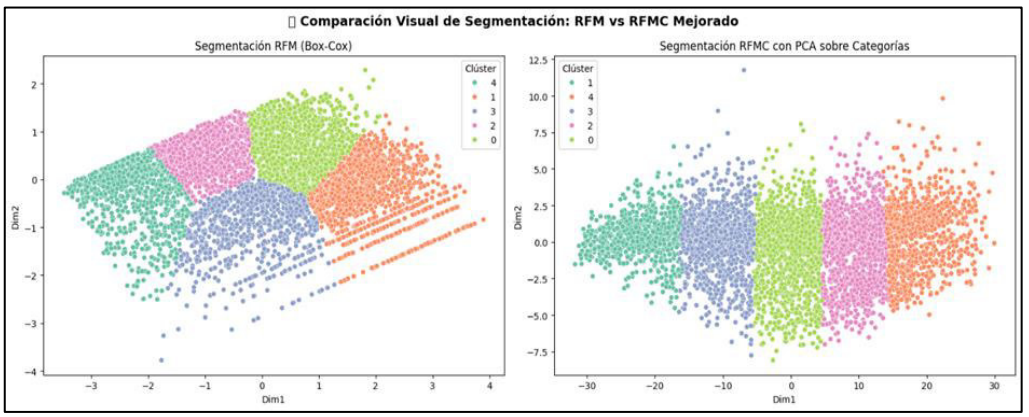
Hipótesis Alternativa (H1): La comparación de los modelos de segmentación de ML (RFM vs. RFMC) **sí** mejora de manera significativa la efectividad en la precisión de preferencia del cliente en una empresa de retail. Es decir, el modelo mejorado (RFMC) es significativamente más efectivo en la precisión de las preferencias del cliente.

Teniendo en cuenta que; Dado un nivel de significancia de $\alpha=0.05$, y dado que $p<0.05$ para ambas variables, **rechazamos la hipótesis nula** para "Precio" y "cantidad". Esto significa que las variables utilizadas en la segmentación son efectivas para diferenciar los grupos de clientes.

Figura 59

Comparativa visual: RFM vs. RFMC





Con base en la evidencia visual de mejor separación de clústeres y la confirmación estadística de la relevancia de las variables (incluyendo la implicación de las categorías), tenemos fuertes indicios que nos permiten **rechazar la hipótesis nula** (H0). Por lo tanto, **aceptamos la hipótesis alternativa** (H1): "La comparación de los modelos de segmentación de ML mejora de manera significativa la efectividad en la precisión de preferencia del cliente en una empresa de retail", ya que la inclusión de nuevas variables (categorías) permite una segmentación más matizada y, por extensión, una mejor comprensión de las preferencias.

V. DISCUSIÓN DE RESULTADOS

El estudio proporciona la suficiente evidencia para dar sustento a las hipótesis planteadas inicialmente. Así, transformación logarítmica permitió corregir los sesgos presentes en las formas originales y obtener distribuciones normales. Esto facilitó modelar de forma eficaz las técnicas de ML. Los hallazgos son consecuentes con los de Aguiar-Costa et al. (2022), quienes destacaron el potencial de la IA y el ML para la recopilación y análisis de los comportamientos. En el presente estudio, la mejora de la estructura de datos mediante la regresión logarítmica demostró que los modelos permiten segmentar a los clientes de forma efectiva, lo que optimiza la identificación de preferencias.

Asimismo, la validación estadística reforzó el hallazgo, ya que reveló que el modelo cumple con los supuestos clave para técnicas de regresión y segmentación, lo que garantiza la fiabilidad y generalización de las predicciones. Esto se asemeja al estudio de Sun et al. (2019), quienes afirman que los algoritmos de ML permiten segmentaciones detalladas y predecir comportamientos de forma precisa.

También, la representación gráfica evidenció distribuciones homogéneas y equilibradas, lo que permitió identificar patrones individuales de cada cliente. De acuerdo con Kietzmann et al. (2018), el ML permite personalizar los servicios según se adapten a las preferencias de cada cliente. De este modo, el estudio confirma que la regresión logarítmica es fundamental para lograr una optimización en la segmentación de comportamientos para personalizar las estrategias de mercadotécnicas.

Finalmente, la transformación y modelado permitió que los recursos empresariales se enfocaran en los clientes con mayor valor, en correspondencia con el propósito de maximizar el **CLV**. Esto da fuerza a la hipótesis de que el modelo permite identificar preferencias y fortalecer la competitividad porque ayuda a asignar recursos de forma eficiente.

Las mejoras permiten segmentar a los clientes mediante K-Means al combinar escalamiento de variables, transformación Box-Cox y la reasignación de etiquetas. Sin embargo, aún se requiere perfeccionamiento y profundización para futuras iteraciones:

1. Explorar algoritmos alternativos

K-Means puede no ser útil para datos reales. Entre mejores alternativas de segmentación se incluyen:

DBSCAN: útil para detectar grupos de forma arbitraria y eliminar ruido.

Gaussian Mixture Models (GMM): permiten clústeres solapados y modelan incertidumbre.

Clustering jerárquico: ideal si se desea explorar estructuras de subgrupos dentro de los clústeres.

2. Evaluación de estabilidad y robustez del modelo

K-Means puede generar resultados diferentes con distintas inicializaciones. Se recomienda:

Ejecutar el modelo con distintos valores de random_state.

Aplicar técnicas de validación cruzada no supervisada o consensus clustering para medir la estabilidad de los clústeres.

3. Mayor riqueza en las variables de entrada

Aunque se utilizó una transformación Box-Cox para mejorar la distribución del Monto, el modelo podría beneficiarse incorporando más dimensiones del comportamiento del cliente. Algunas variables adicionales que podrían explorarse:

Canal de compra.

Datos demográficos o geográficos (si están disponibles).

Esto permitiría una segmentación más completa y estratégica.

4. Perfilado automatizado y explicación de clústeres

Una vez definidos los clústeres, se pueden entrenar modelos supervisados para explicar qué variables influyen más en la pertenencia a cada grupo. Esto es clave para facilitar la interpretación a stakeholders y diseñar campañas de marketing personalizadas.

5. Seguimiento temporal de los segmentos

Implementar un análisis temporal (por ejemplo, mensualmente) para observar cómo migran los clientes entre clústeres podría ser una fuente valiosa de insights. Esto permite monitorear:

Fidelización o abandono de clientes valiosos.

Efectividad de campañas.

Cambios en el comportamiento por estacionalidad o promociones.

VI. CONCLUSIONES

6.1 Conclusión general

Los resultados del presente estudio permiten **rechazar la hipótesis nula** y **aceptar la hipótesis alternativa**, la cual sostiene que: El modelo mejorado de segmentación de machine learning, basado en la regresión logarítmica, optimiza de manera directa la identificación de las preferencias de los clientes en una empresa de retail. Esta conclusión se respalda en múltiples evidencias estadísticas y gráficas, tales como:

- La normalización efectiva de las variables clave (Recencia, Frecuencia, Monto) mediante transformación logarítmica.
- La validación de clústeres significativamente diferenciados con base en las variables de comportamiento de compra ("Precio" y "cantidad").
- Resultados significativos en la prueba ANOVA y la prueba no paramétrica de Kruskal-Wallis, que refuerzan la utilidad del modelo al identificar segmentos con patrones de consumo distintos.

6.2 Conclusiones específicas:

- **Mejora en la estructura de los datos mediante la transformación logarítmica**

La transformación de las variables originales sesgadas (Recencia, Frecuencia, Monto) logró distribuciones aproximadas a la normalidad, como se evidenció en los histogramas post-transformación. Esto no solo mejora la calidad de los datos, sino que garantiza que los modelos de regresión logarítmica cumplan con los supuestos estadísticos necesarios, como la homocedasticidad y la linealidad. Esta mejora estructural permitió aplicar técnicas de segmentación más precisas y robustas.

- **Mejora en la capacidad predictiva y diferenciación entre clústeres**

La prueba ANOVA mostró que los clústeres generados son estadísticamente diferentes entre sí en términos de las variables de comportamiento “Precio” y “cantidad”. Esta

diferenciación es esencial para validar la eficacia del modelo de segmentación, pues demuestra que los grupos identificados no son aleatorios, sino que capturan diferencias significativas en el comportamiento del cliente.

La evidencia visual de los boxplots, con diferencias claras en mediana, dispersión y valores atípicos por categoría, reforzó estas diferencias.

Asimismo, la prueba Kruskal-Wallis confirmó esta diferencia con un valor estadístico $H = 51.62$ y un valor $p = 5.32 \times 10^{-8}$, significativamente menor que 0.05, lo cual lleva al rechazo de la hipótesis nula y a la aceptación de la hipótesis alternativa, en que los grupos formados son distintos de forma significativa.

- **Inclusión de variables categóricas y comprensión matizada de preferencias**

La incorporación de variables como la categoría de productos (“Bolsos y Accesorios”, “Regalo y Decoración”, “Otros”) permitió una segmentación más matizada, enriqueciendo la interpretación de los clústeres formados. Estas categorías presentaron colas largas hacia precios altos, indicando variabilidad significativa que puede asociarse con distintos perfiles de consumo.

Es importante reconocer que los modelos de segmentación no son estructuras eternas ni verdades definitivas. Son representaciones temporales del comportamiento de los datos en un momento específico. En un entorno de negocio dinámico, donde los clientes cambian constantemente sus hábitos, intereses y capacidad de gasto, cualquier modelo pierde validez con el tiempo si no es actualizado.

Por ello, no basta con construir un buen modelo una sola vez: es fundamental integrar el proceso de reentrenamiento y evaluación periódica como parte del análisis. Solo así se asegura que la segmentación continúe reflejando la realidad del negocio y permita tomar decisiones basadas en datos actuales y relevantes. Adoptar esta mentalidad evita la ilusión de permanencia y promueve una cultura de mejora continua, basada en evidencia.

VII. RECOMENDACIONES

En considerando a los resultados obtenidos, se proponen las siguientes recomendaciones las cuales están orientadas a optimizar el modelo de segmentación y a potenciar su aplicación estratégica en el contexto empresarial

7.1 Mejoras metodológicas en los modelos de segmentación

Explorar algoritmos alternativos: Aunque K-Means permitió obtener clústeres diferenciados, su sensibilidad a la inicialización y a la forma de los datos limita su aplicabilidad en escenarios reales. Se recomienda probar algoritmos más robustos como DBSCAN, Gaussian Mixture Models (GMM) y clustering jerárquico, los cuales ofrecen mayor flexibilidad para capturar estructuras complejas y patrones solapados.

7.2 Evaluar estabilidad y robustez del modelo

Es pertinente aplicar técnicas como validación cruzada no supervisada o consensus clustering, además de ejecutar el modelo con diferentes valores de random_state, a fin de garantizar que los clústeres sean consistentes y reproducibles.

7.3 Incorporación de nuevas variables

Incluir atributos adicionales que permitan enriquecer la segmentación, tales como:

- Canales de compra (tienda física, e-commerce, redes sociales).
- Características demográficas o geográficas de los clientes (edad, género, ubicación).
- Historial de interacción (frecuencia de devoluciones, uso de promociones, reseñas).

La ampliación de dimensiones en el análisis favorecerá una visión más completa del cliente y permitirá diseñar estrategias diferenciadas con mayor precisión.

7.4 Monitoreo y actualización del modelo

Dado que el comportamiento del cliente es dinámico, se recomienda establecer un proceso de actualización periódica del modelo, ya sea de forma mensual o trimestral. Este seguimiento temporal permitirá identificar migraciones de clientes entre segmentos, evaluar la efectividad de las campañas implementadas y anticipar cambios en las preferencias por factores estacionales o de mercado.

7.5 Aplicación estratégica en la gestión empresarial

Utilizar los segmentos obtenidos para personalizar estrategias de mercadotecnia, ajustando promociones, recomendaciones y servicios al perfil de cada cliente.

- Enfocar recursos en los clientes de mayor valor de vida (CLV), lo que permitirá maximizar el retorno de inversión y fortalecer la competitividad empresarial.
- Integrar la segmentación en sistemas de CRM e inteligencia de negocios, de manera que los resultados puedan alimentar en tiempo real las decisiones comerciales y estratégicas.

VIII. REFERENCIAS

Aguiar-Costa, L. M., Cunha, C. A. X. C., Silva, W. K. M., & Abreu, N. R. (2022). Customer satisfaction in service delivery with artificial intelligence: A meta-analytic study. *RAM. Revista de Administração Mackenzie*, *23*(6). https://doi.org/10.1590/1678-6971/eramd220003.en

Al-Araj, R., Haddad, H., Shehadeh, M., Hasan, E., & Nawaiseh, M. Y. (2022). The Effect of Artificial Intelligence on Service Quality and Customer Satisfaction in Jordanian Banking Sector. *WSEAS Transactions on Business and Economics*, *19*, 1929–1947. https://doi.org/10.37394/23207.2022.19.173

Aldunate, Á., Maldonado, S., Vairetti, C., & Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *Expert Systems with Applications*, *209*, 118309. https://doi.org/10.1016/j.eswa.2022.118309

Anish, N. (2022). *RFM Analysis For Successful Customer Segmentation*. Putler. https://www.putler.com/rfm-analysis/

Arias, J. L., & Covinos, M. (2021). *Diseño y metodología de la investigación*. Enfoques Consulting. https://bit.ly/40ZFlCb

Atawneh, S. H., Hamadneh, N. N., Jaber, J. J., Al Wadi, S., & Khan, W. A. (2022). Using Artificial Intelligence to Predict Customer Satisfaction with E-Payment Systems during the COVID-19 Pandemic. *Journal of Mathematics*, *2022*(1). https://doi.org/10.1155/2022/1599785

Calvo-Porral, C., & Lévy-Mangin, J.-P. (2019). Profiling shopping mall customers during hard times. *Journal of Retailing and Consumer Services*, *48*, 238–246. https://doi.org/10.1016/j.jretconser.2019.02.023

Capuano, N., Greco, L., Ritrovato, P., & Vento, M. (2021). Sentiment analysis for customer relationship management: an incremental learning approach. *Applied Intelligence*, *51*(6), 3339–3352. https://doi.org/10.1007/s10489-020-01984-x

Chang, H.-C., & Tsai, H.-P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, *38*(12), 14499–14513. https://doi.org/10.1016/j.eswa.2011.05.034

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Chugani, V. (2024). *Understanding Euclidean Distance: From Theory to Practice*. DataCamp. https://www.datacamp.com/tutorial/euclidean-distance

Das, S., & Nayak, J. (2022). Customer Segmentation via Data Mining Techniques: State-of-the-Art Review. In *Computational Intelligence in Data Mining* (pp. 489–507). Smart Innovation. https://doi.org/10.1007/978-981-16-9447-9_38

Decide Soluciones. (2022). *Tipos de aprendizaje que usan los algoritmos de Machine Learning*. https://decidesoluciones.es/tipos-de-aprendizaje-algoritmos-machine-learning/

Dick, A. S., & Basu, K. (1994). Customer Loyalty: Toward an Integrated Conceptual Framework. *Journal of the Academy of Marketing Science*, *22*(2), 99–113. https://doi.org/10.1177/0092070394222001

Gamboa, M. E. (2023). El cálculo del tamaño de la muestra en la investigación científica. *Dilemas Contemporáneos: Educación, Política y Valores*, *11*(1), 1–27. https://doi.org/10.46377/dilemas.v11i1.3680

GeeksforGeeks. (2025). *Determine the optimal value of K in K-Means Clustering*. https://bit.ly/460MFQz

Glotzer, S., & Simla, S. (2024). *Qué es análisis RFM y para qué sirve*. https://www.simla.com/blog/que-es-analisis-rfm

Griva, A., Bardaki, C., Pramatari, K., & Doukidis, G. (2022). Factors Affecting Customer Analytics: Evidence from Three Retail Cases. *Information Systems Frontiers*, *24*(2), 493–516. https://doi.org/10.1007/s10796-020-10098-1

Hernández-Sampieri, R., & Mendoza, C. (2018). *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta*. McGraw Hill Education. https://doi.org/10.22201/fesc.20072236e.2019.10.18.6

Hiziroglu, A., & Sengul, S. (2012). Investigating Two Customer Lifetime Value Models from Segmentation Perspective. *Procedia - Social and Behavioral Sciences*, *62*, 766–774. https://doi.org/10.1016/j.sbspro.2012.09.129

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D.-S., & Le, T.-G. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Business Systems Research Journal*, *14*(1), 26–53. https://doi.org/10.2478/bsrj-2023-0002

Hu, Y.-H., & Yeh, T.-W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based Systems*, *61*, 76–88. https://doi.org/10.1016/j.knosys.2014.02.009

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International*

Journal of Information Management, *70*, 102641. https://doi.org/10.1016/j.ijinfomgt.2023.102641

Junta de Andalucía. (2020). *Aprendizaje automático*. Mi Asistente Personal. https://bit.ly/40g4ojh

Kassem, E. A. el, Ali, S., Mostafa, A., & Kamal, F. (2020). Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. *International Journal of Advanced Computer Science and Applications*, *11*(5), 522–531. https://doi.org/10.14569/IJACSA.2020.0110567

Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial Intelligence in Advertising. *Journal of Advertising Research*, *58*(3), 263–267. https://doi.org/10.2501/JAR-2018-035

Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets*, *32*(4), 2235–2244. https://doi.org/10.1007/s12525-022-00598-0

Kuhuparuw, V. J., Elyta, S., & Al Qadrie, S. R. F. (2024). Customer relationship management and information security in the development of small and medium enterprises in the era of digitalization as strengthening human resources. *International Journal of Multidisciplinary Research & Reviews*, *3*(1), 39–58.

Kumar, A. (2022). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. In *Handbook of Research on Applied AI for International Business and Marketing Applications* (pp. 248–270). https://doi.org/10.4018/978-1-6684-5727-6.ch013

Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. *Journal of Retailing*, *80*(4), 317–329. https://doi.org/10.1016/j.jretai.2004.10.007

Lozada, J. (2014). Investigación Aplicada: Definición, Propiedad Intelectual e Industria. *CienciAmérica*, *3*(1), 47–50. https://cienciamerica.edu.ec/index.php/uti/article/view/30

Marín, A. (2020). *Machine Learning, un ejemplo con Python*. Linked In. https://bit.ly/4e7nIFa

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

McDougall, G. H. G., & Levesque, T. (2000). Customer satisfaction with services: putting perceived value into the equation. *Journal of Services Marketing*, *14*(5), 392–410. https://doi.org/10.1108/08876040010340937

Mensouri, D., Azmani, A., & Azmani, M. (2022). K-Means Customers Clustering by their RFMT and Score Satisfaction Analysis. *International Journal of Advanced Computer Science and Applications*, *13*(6). https://doi.org/10.14569/IJACSA.2022.0130658

Miglautsch, J. (2002). Application of RFM principles: What to do with 1–1–1 customers? *Journal of Database Marketing & Customer Strategy Management*, *9*(4), 319–324. https://doi.org/10.1057/palgrave.jdm.3240080

Oncioiu, I., Căpușneanu, S., Topor, D., Tamaș, A., Solomon, A.-G., & Dănescu, T. (2021). Fundamental Power of Social Media Interactions for Building a Brand and Customer Relations. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(5), 1702–1717. https://doi.org/10.3390/jtaer16050096

Paul, L., & Ramanan, T. R. (2019). An RFM and CLV analysis for customer retention and customer relationship management of a logistics firm. *International Journal of Applied Management Science*, *11*(4), 333. https://doi.org/10.1504/IJAMS.2019.103713

Polo, D. R. (2022). La representatividad de una muestra en investigaciones educativas. *Ciencias Pedagógicas*, *15*(1), 182–190. https://www.cienciaspedagogicas.rimed.cu/index.php/ICCP/article/view/360

Porta, J. (2016). *El ciclo de vida del cliente*. Marketing. https://jaimeporta.com/2016/01/22/el-ciclo-de-vida-del-cliente/

Robles, B. F. (2019). Población y muestra. *PuebloCont*, *30*(1), 245–246.

Rosset, S. (2003). Customer Lifetime Value Models for Decision Support. *Data Mining and Knowledge Discovery*, *7*(3), 321–339. https://doi.org/10.1023/A:1024036305874

Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing*, *69*(2), 193–215. https://doi.org/10.1016/0022-4359(93)90003-2

Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of Machine Learning and Data Mining*. Springer US. https://doi.org/10.1007/978-1-4899-7687-1

Seymen, O. F., Ölmez, E., Dogan, O., Er, O., & Hiziroglu, K. (2023). Customer Churn Prediction Using Ordinary Artificial Neural Network and Convolutional Neural Network Algorithms: A Comparative Performance Assessment. *Gazi University Journal of Science*, *36*(2), 720–733. https://doi.org/10.35378/gujs.992738

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Simone, A., & Sabbadin, E. (2017). The New Paradigm of the Omnichannel Retailing: Key Drivers, New Challenges and Potential Outcomes Resulting from the Adoption of an Omnichannel Approach. *International Journal of Business and Management*, *13*(1), 85. https://doi.org/10.5539/ijbm.v13n1p85

Subali, T., Kusumawati, A., Mauludin, H., & Zaini, A. (2020). Mediating effect of customer perceive value on experience quality and loyalty relationship. *Utopía y Praxis*, *25*(1), 524–536. https://produccioncientificaluz.org/index.php/utopia/article/view/33566

Sun, Y., Shao, X., Li, X., Guo, Y., & Nie, K. (2019). How live streaming influences purchase intentions in social commerce: An IT affordance perspective. *Electronic Commerce Research and Applications*, *37*, 100886. https://doi.org/10.1016/j.elerap.2019.100886

Uzir, M. U. H., Al Halbusi, H., Lim, R., Jerin, I., Abdul Hamid, A. B., Ramayah, T., & Haque, A. (2021). Applied Artificial Intelligence and user satisfaction: Smartwatch usage for healthcare in Bangladesh during COVID-19. *Technology in Society*, *67*, 101780. https://doi.org/10.1016/j.techsoc.2021.101780

Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, *1*(1), 100002. https://doi.org/10.1016/j.jjimei.2020.100002

Vizcaíno, P. I., Cedeño, R. J., & Maldonado, I. A. (2023). Metodología de la investigación científica: guía práctica. *Ciencia Latina Revista Científica Multidisciplinar*, *7*(4), 9723–9762. https://doi.org/10.37811/cl_rcm.v7i4.7658

Wei, J. T., Lin, S. Y., & Lin, S. Y. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199–4206.

Zada, I. (2022). The Contributions of Customer Knowledge and Artificial Intelligence to Customer Satisfaction. *International Review of Management and Marketing*, *12*(5), 1–4. https://doi.org/10.32479/irmm.13314

Zanchett, R., & Paladini, E. P. (2019). Consumer loyalty programs: impact of different modalities. *DYNA*, *86*(208), 206–213. https://doi.org/10.15446/dyna.v86n208.71080

Zhao, Q., Zhao, Z., Yang, L., Hong, L., & Han, W. (2023). Research on Customer Retention Prediction Model of VOD Platform Based on Machine Learning. *International Journal of Advanced Computer Science and Applications*, *14*(4). https://doi.org/10.14569/IJACSA.2023.0140427

IX. ANEXOS

A. Matriz de Consistencia

Problemas de investigación	Objetivos	Hipótesis	Variables	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Metodología
Problema general ¿De qué manera la propuesta de un modelo mejorado de segmentación de Machine Learning (ML) describe el comportamiento del cliente para su fidelización en una empresa de retail? **Problemas específicos** ¿De qué manera el uso de nuevos parámetros en los requerimientos del modelo mejorado de segmentación de ML permitirá identificar las preferencias del cliente en una empresa de retail? ¿De qué manera el modelo mejorado de segmentación de ML podrá optimizar la identificación de las preferencias de los clientes en una empresa de retail? ¿De qué manera la comparación de los modelos de segmentación de ML según su precisión podrá mejorar la efectividad en la preferencia del cliente de una empresa retail?	**Objetivo general** Diseñar e implementar un modelo mejorado de segmentación de Machine Learning para deducir el comportamiento del cliente en su fidelización en una empresa de retail. **Objetivo específicos** Utilizar nuevos parámetros en los requerimientos del modelo mejorado de segmentación de ML para identificar las preferencias del cliente en una empresa de retail. Demostrar el modelo mejorado de segmentación de ML para optimizar la identificación de las preferencias de los clientes en una empresa de retail. Comparar la precisión de los modelos de segmentación de ML para mejorar la efectividad en la preferencia del cliente de una empresa retail.	**Hipótesis general** El modelo mejorado de segmentación propuesto de Machine Learning predice el comportamiento del cliente para su fidelización en una empresa de retail. **Hipótesis específicos** El uso de nuevos parámetros en los requerimientos del modelo mejorado de ML segmenta con mayor precisión las preferencias del cliente. El modelo mejorado de segmentación de ML optimiza la identificación de las preferencias de los clientes en una empresa de retail. Los modelos de segmentación de ML permitieron comparar de manera eficiente la preferencia del cliente en una empresa de retail.	**Variable Independiente:** Modelo de Machine Learning.	Conjunto de técnicas que permiten a las máquinas aprender de datos y hacer predicciones o recomendaciones automáticas.	El uso de **IA/ML** será medido por la aplicación de algoritmos de aprendizaje supervisado y no supervisado para la segmentación de clientes.	Comportamiento de compra	Nivel de satisfacción	diseño no experimental
			Variable Dependiente: Fidelización del cliente.	Fidelización del cliente: Lealtad hacia la marca o empresa.	La fidelización se medirá por la repetición de compras y recomendaciones.	Fidelización.	Porcentaje de clientes recurrentes y tasa de recomendación.	Análisis de datos transaccionales y encuestas de satisfacción. Modelos de ML (K-Means, regresión logística, redes neuronales).

B. Validación y Confiabilidad de Instrumentos

1.- Dataset requerido para la tesis:

a) Datos del Cliente:

- **ID del cliente**: Un identificador único para cada cliente
- **Edad**: Para segmentar y analizar la relación entre el comportamiento y la edad.
- **Género**: Información demográfica útil para la personalización.
- **Ubicación geográfica**: Se realizará un filtro de solo la localidad de Lima Metropolitana.
- **Fecha de afiliación**: Fecha en que el cliente se afilio al sistema de lealtad

b) Información de las Transacciones:

- **ID de la transacción:** Un identificador único para cada compra realizada por el cliente.
- **Fecha de la transacción:** Importante para calcular métricas como recencia (cuánto tiempo ha pasado desde la última compra).
- **Monto de la transacción:** El valor monetario de cada compra, que se utilizará para calcular el valor monetario total de los clientes.
- **Cantidad de productos comprados:** El número de unidades adquiridas en cada transacción.
- **Categoría del producto:** Categoría o tipo de producto comprado (electrónica, ropa, alimentos, etc.).
- **Canal de compra:** Si la transacción fue realizada en tienda física, en línea, a través de la app móvil, etc. Es relevante para la segmentación y análisis de preferencia de canal.

c) Métricas de Retención y Fidelización (Calculadas):

- **CLV:** Valor total estimado que un cliente generará durante su relación con la empresa. Se calcula usando la frecuencia de compras, el monto promedio por compra y la duración estimada de la relación con el cliente.
- **Recencia:** Número de días desde la última compra del cliente.
- **Frecuencia de compra:** Número de compras realizadas en un período determinado.
- **Monetario:** Monto total gastado por el cliente**.**

Formato del dataset:

El dataset deberá de estar estructurado como una **tabla de transacciones**, donde cada fila representa una transacción individual, y con columnas que contienen agregados como número de transacciones, gasto total, frecuencia de compra, etc.

Ejemplo de un dataset estructurado:

ID Cliente	ID Transacción	Fecha de Transacción	Monto de la Transacción	Cantidad de Productos	Categoría del Producto	Método de Pago	Canal de Compra	cod. Tienda	Recencia	Frecuencia	Monetario
1	TXN123	1/05/2023	300	3	Electrónica	Tarjeta de crédito	Online	50	10 días	5 compras	1,200
2	TXN124	3/05/2023	120	5	Ropa	Tarjeta de débito	Tienda física	60	15 días	8 compras	900
...	...	...	...	...	...	...	...	...	...	...	...