Universidad Internacional de La Rioja
Facultad de Empresa, Comunicación y Marketing

Máster Universitario en Inteligencia de Negocio

Predicción de Abandono de Clientes en una Empresa de Comercialización de Productos de Consumo Masivo

Trabajo fin de estudio presentado por:	Ernesto Armando Rojas Mañón
Tipo de trabajo:	Proyecto de Inteligencia de Negocio
Modalidad:	Individual
Director/a:	Karla Barajas Portas
Fecha:	16/07/2024

Resumen

Este TFM aborda la problemática del abandono de clientes en una empresa de comercialización y distribución de productos de consumo masivo, destacando la crucial importancia de la retención de clientes y la necesidad de desarrollar modelos predictivos para prevenir la fuga. El objetivo principal del estudio es desarrollar un modelo de predicción del abandono de clientes en la empresa seleccionada. Para lograrlo, se identificarán los factores determinantes que influyen en el abandono, se analizará el comportamiento de compra de los clientes y se implementarán diversas técnicas de análisis de datos, incluyendo el aprendizaje supervisado y no supervisado. Los modelos predictivos sugeridos serán evaluados en términos de precisión y efectividad, y se propondrán estrategias específicas para anticipar, prevenir y minimizar el abandono de clientes, tales como la implementación de programas de lealtad, estrategias de upselling y cross-selling, entre otras. Asimismo, se explorarán herramientas tecnológicas específicas como el CRM y se sugerirá plataformas de automatización de marketing para optimizar la gestión de relaciones con los clientes y mejorar la eficiencia operativa de la empresa.

Palabras clave: Abandono de clientes, fidelización, retención, predicción, productos de consumo masivo.

Abstract

This master's thesis addresses the problem of customer churn in a FMCG commercialization and distribution company, highlighting the crucial importance of customer retention and the need to develop predictive models to prevent churn. The main objective of the study is to develop a customer churn prediction model for the selected company. To achieve this, the key factors influencing churn will be identified, customer buying behavior will be analyzed, and various data analysis techniques, including supervised and unsupervised learning, will be implemented. The suggested predictive models will be evaluated in terms of accuracy and effectiveness, and specific strategies will be proposed to anticipate, prevent, and minimize customer churn. These strategies include the implementation of loyalty programs, upselling and cross-selling strategies, among others. Likewise, specific technological tools such as CRM systems and marketing automation platforms will be explored to optimize customer relationship management and improve the company's operational efficiency.

Keywords: Customer churn, loyalty, retention, prediction, mass consumer products.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

En el actual entorno empresarial, caracterizado por una alta competencia, la retención de clientes emerge como una prioridad estratégica, especialmente en el sector de distribución y comercialización de productos de consumo masivo. En este contexto, el fenómeno del abandono de clientes plantea un desafío significativo que puede afectar adversamente tanto la rentabilidad como la reputación empresarial. Por consiguiente, resulta imperativo desarrollar herramientas efectivas para prevenir y gestionar este fenómeno. La capacidad de predecir el abandono de clientes se erige como una herramienta estratégica para la toma de decisiones, posibilitando la implementación de acciones proactivas que mitiguen la pérdida de clientes y potencien la fidelización de estos.

1.1. Planteamiento general: descripción y justificación del proyecto

El planteamiento del problema se articula a través de las siguientes interrogantes: ¿Qué variables inciden de manera significativa en la retención de clientes? ¿Cómo pueden ser modelados y anticipados los comportamientos de abandono? ¿Qué estrategias se pueden implementar para disminuir la tasa de deserción?

Este Trabajo de Fin de Máster se centra en la predicción del abandono de clientes mediante el empleo de técnicas avanzadas de análisis de datos y modelado predictivo. Para fundamentar esta investigación, se examinará la literatura académica pertinente, la cual incluye a autores destacados como Kotler y Keller (2012), quienes señalan que la retención de clientes resulta entre cinco y diez veces más económica que la adquisición de nuevos. Asimismo, Reichheld (1996) asegura que la tasa media de abandono de clientes en este sector alcanza aproximadamente el 30% anual y propone que la lealtad del cliente constituye un indicador crucial para la salud financiera a largo plazo de una empresa, abogando por la adopción de medidas que fomenten relaciones duraderas.

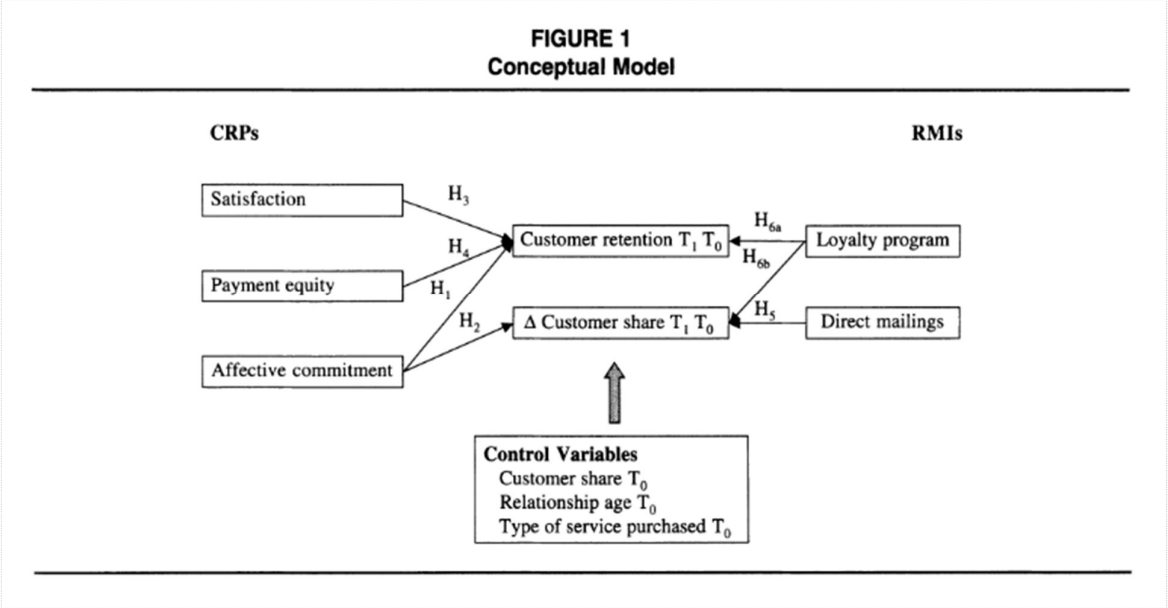
Es pertinente resaltar que diversos estudios e informes posteriores corroboran la elevada tasa de abandono en este ámbito:

- Un estudio de la consultora Forrester Research (2014)[1] reveló que la tasa promedio de abandono de clientes en empresas B2C es del 23% anual.

- Por su parte, un informe de la firma Gartner (2018)[2] indicó que las empresas pierden un promedio del 30% de sus clientes cada año debido a una mala experiencia de cliente.

De igual forma, Verhoef et al. (2003) abordan la gestión del ciclo de vida del cliente y sugieren en su modelo conceptual (*figuras 1 y 2*) que las estrategias centradas en la satisfacción y la fidelización del cliente pueden incrementar la rentabilidad a largo plazo de la empresa. También se tomarán en cuenta los trabajos de Han et al. (2001) y Zhang et al. (2006), quienes proponen modelos predictivos para la retención de clientes en entornos empresariales.

Figura 1. Modelo conceptual cliente-fidelización



Fuente: Verhoef, P. C. (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development (Vol. 67). Journal of Marketing. Pag 35

[1] Forrester Research (2014). The State Of Customer Experience In 2014. https://www.forrester.com/report/the-business-impact-of-customer-experience-2014/RES113421

[2] Gartner (2018). CRM Customer Experience Management Magic Quadrant. https://www.gartner.com/reviews/market/crm-customer-engagement-center

Figura 2. Resultado de prueba e hipótesis para la retención de clientes

TABLE 6
Summary of Hypothesis-Testing Results

	Customer Retention			Customer Share Development		
Antecedents	**Hypothesis (Sign)**	**Effect**	**Support**	**Hypothesis (Sign)**	**Effect**	**Support**
Affective commitment	H_1 (+)	+	Yes	H_2 (+)	+	Yes
Satisfaction	H_3 (+)	0; positively moderated by relationship age	No	No effect	0	Yes
Payment equity	H_4 (+)	0	No	No effect	0	Yes
Direct mailings	No effect	N.A.	N.A.	H_5 (+)	+	Yes
Loyalty program	H_{6a} (+)	+	Yes	H_{6b} (+)	+	Yes

Notes: N.A. = not available; this effect could not be estimated because of data limitations.

Fuente: Verhoef, P. C. (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development (Vol. 67). Journal of Marketing. Pag 41

1.2. Objetivos del TFE

1.2.1. Objetivo General:

El objetivo principal es desarrollar un modelo predictivo preciso y aplicable que permita anticipar el abandono de clientes en una empresa de distribución y comercialización de productos de consumo masivo.

1.2.2. Objetivos Específicos:

- Identificar los factores determinantes del abandono de clientes en la empresa seleccionada mediante un análisis exhaustivo.
- Recopilar y analizar datos pertinentes sobre el comportamiento de compra de los clientes, segmentándolos en función de su riesgo de abandono.
- Desarrollar y validar un modelo de predicción, haciendo uso de técnicas avanzadas de análisis de datos.
- Evaluar la eficacia del modelo de predicción implementado, proponer recomendaciones prácticas y fomentar la retención a largo plazo.

2. La Retención de Clientes: Pilar Fundamental para el Éxito

La retención de clientes es un proceso continuo que requiere un enfoque estratégico y proactivo. Es esencial en entornos altamente competitivos. Invertir en la fidelización de los clientes es una inversión en el futuro del negocio, asegurando su rentabilidad y sostenibilidad.

2.1. Importancia Estratégica

En un mercado cada vez más competitivo, la retención de clientes es crucial para la supervivencia y el éxito a largo plazo de las empresas. Captar nuevos clientes implica altos costos en inversión de tiempo, recursos y esfuerzo. En contraste, retener a los clientes existentes es una estrategia significativamente más eficiente y rentable.

2.2. Rentabilidad a Largo Plazo

Reichheld (1996), reconocido experto en fidelización, argumenta que la retención de clientes es fundamental para la rentabilidad a largo plazo de una empresa. Los clientes leales no solo realizan compras recurrentes, sino que también se convierten en defensores de la marca, generando un impacto positivo en la reputación y el crecimiento del negocio.

- Costos de Adquisición vs. Retención: Captar un nuevo cliente implica un considerable gasto en marketing, publicidad, comisiones de ventas y otros procesos. En contraste, mantener a un cliente existente es significativamente más económico. Según estudios de Harvard Business Review[3], la adquisición de un nuevo cliente puede llegar a ser hasta 5 veces más costosa que la retención de uno actual.

- Customer Lifetime Value (CLV): Los clientes leales no solo realizan compras recurrentes, sino que también tienden a aumentar su inversión a lo largo del tiempo. El CLV, que mide el valor total que un cliente aporta a la empresa durante su relación, es considerablemente mayor para clientes leales. Frederick Reichheld indica que un aumento del 5% en la retención de clientes puede generar un aumento del 25% al 95% en las ganancias.

[3] Harvard Business Review: The Importance of Customer Retention: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

2.3. Indicador de Salud Financiera

La lealtad del cliente, medida a través de la tasa de retención, se considera un indicador clave de la salud financiera de la empresa. Un alto índice de retención refleja una base de clientes sólida y comprometida, lo que se traduce en ingresos estables y predecibles.

2.4. Beneficios Adicionales

Además de generar ingresos recurrentes, la retención de clientes tiene impactos positivos adicionales en la empresa. Verhoef et al. (2003) destacan que los clientes leales no solo contribuyen a los ingresos continuos, sino que también están más inclinados a:

- Realizar compras de mayor valor: Los clientes leales tienden a aumentar su gasto con la empresa a lo largo del tiempo. A medida que desarrollan una relación sólida con la marca, están más dispuestos a invertir en productos o servicios adicionales ofrecidos.

- Recomendar la marca a sus amigos y familiares: La lealtad del cliente va de la mano con la defensa de la marca. Los clientes satisfechos y comprometidos tienen más probabilidades de compartir su experiencia positiva con otros, lo que puede generar un efecto de boca a boca positivo y atraer nuevos clientes potenciales.

- Ser más tolerantes ante errores o dificultades: Los clientes leales tienen una mayor disposición a perdonar errores o problemas ocasionales que puedan surgir en la interacción con la empresa. Esta tolerancia puede ayudar a mitigar los impactos negativos de situaciones adversas y mantener la relación a largo plazo.

- Brindar información valiosa sobre sus necesidades y preferencias: Los clientes leales suelen estar más dispuestos a proporcionar retroalimentación detallada sobre sus experiencias y expectativas. Esta información es invaluable para la empresa, ya que permite comprender mejor las necesidades del cliente y adaptar las estrategias de marketing y servicio al cliente para satisfacerlas de manera más efectiva.

3. Abandono de Clientes: Sus Implicaciones

Como ya hemos mencionado, el abandono de clientes es un problema que afecta a todas las empresas y tiene un impacto negativo en la rentabilidad. Implementar estrategias proactivas para prevenir el abandono de clientes es fundamental para asegurar el éxito a largo plazo de las empresas. Han et al. (2001) señalan que el abandono de clientes puede ser resultado de diversos factores, incluyendo insatisfacción con el producto o servicio, cambios en las circunstancias del cliente, o la aparición de competidores más atractivos.

3.1. Definición y Medición:

El abandono de clientes, también conocido como "Churn", se refiere al porcentaje de clientes que dejan de comprar o utilizar los productos o servicios de una empresa en un período determinado. La tasa de abandono o Churn Rate se calcula dividiendo el número de clientes perdidos por el número total de clientes al inicio del período, y multiplicando por 100.

$$\frac{\text{USERS AT BEGINNING OF PERIOD - USERS AT END OF PERIOD}}{\text{USERS AT BEGINNING OF PERIOD}} = \text{CHURN RATE}$$

3.2. Impacto en la Rentabilidad:

Perder clientes implica:

- Disminución en los ingresos: Los clientes que abandonan la empresa dejan de generar ingresos, lo que afecta directamente la facturación y las ganancias.
- Costos de adquisición: La empresa debe invertir recursos en la captación de nuevos clientes para reemplazar a los que se han ido.
- Pérdida de oportunidades: Los clientes que abandonan la empresa podrían haber realizado compras adicionales en el futuro.
- Daño a la reputación: El abandono de clientes puede afectar la imagen y la reputación de la empresa, lo que dificulta la atracción de nuevos clientes.

3.3. Causas del Abandono de Clientes:

Existen diversos factores que pueden incidir en el abandono de clientes, como:

3.3.1. Factores Internos:

- Insatisfacción con el producto o servicio: La calidad del producto o servicio es determinante en la satisfacción del cliente. Si el producto no cumple con sus expectativas o presenta fallas, es probable que busque alternativas en la competencia.
- Mala experiencia del cliente: Una experiencia negativa en el proceso de compra, atención al cliente, o uso del producto puede generar insatisfacción y llevar al abandono.
- Falta de valor percibido: Si el cliente no percibe que el producto o servicio ofrece un valor suficiente en comparación con la competencia, es probable que busque opciones más atractivas.
- Errores y fallos: Errores en los pedidos, problemas de facturación, o fallos en el servicio pueden generar una experiencia negativa y conducir al abandono.
- Problemas de comunicación: La falta de transparencia, información errónea o una comunicación deficiente con los clientes puede generar insatisfacción y abandono.

3.3.2. Factores Externos:

- Competencia: La presencia de competidores con ofertas más atractivas, precios más bajos, o una mejor experiencia del cliente puede provocar la fuga de clientes.
- Situaciones económicas: Las crisis económicas o cambios en el mercado pueden afectar la capacidad de los clientes para seguir comprando productos o servicios.
- Cambios en las tendencias: La aparición de nuevas tecnologías, modas o preferencias del público pueden hacer que un producto o servicio se vuelva obsoleto, llevando al abandono de los clientes.

3.3.3. Factores Personales:

- Cambios en las necesidades del cliente: Los clientes pueden modificar sus necesidades o preferencias con el tiempo, haciendo que un producto o servicio ya no sea adecuado para ellos.
- Mudanzas o cambios de estilo de vida: Mudanzas a otras ciudades, cambios de trabajo o la formación de una familia pueden afectar la capacidad del cliente para continuar utilizando un producto o servicio.

- Fallecimiento del cliente: En el caso de empresas que ofrecen servicios a personas, el fallecimiento del cliente es un factor inevitable que puede llevar al abandono.

3.3.4. Factores Tecnológicos:

- Dificultades en el uso del producto o servicio: Si los clientes encuentran el producto o servicio difícil de usar, es probable que lo abandonen.
- Problemas de accesibilidad: La falta de accesibilidad para ciertos dispositivos o plataformas puede limitar el uso del producto o servicio y llevar al abandono de algunos clientes.
- Fallos en la plataforma o sitio web: Errores técnicos, fallos en la plataforma o problemas de seguridad pueden generar una experiencia negativa, resultando en el abandono de los clientes.

3.4. Estrategias para Reducir el Abandono de Clientes:

Para prevenir el abandono de clientes, las empresas pueden implementar diversas estrategias, como:

3.4.1. Enfoque en la Satisfacción del Cliente:

- Producto o Servicio de Alta Calidad: Ofrecer un producto o servicio que cumpla o supere las expectativas del cliente es fundamental para la satisfacción. La calidad debe ser consistente en todos los aspectos, desde el diseño y la funcionalidad hasta la atención al cliente.
- Atención al Cliente Excepcional: Brindar una atención al cliente rápida, eficiente y personalizada es clave para fidelizar a los clientes. Es importante resolver las dudas, quejas y problemas de manera oportuna y eficaz.
- Personalización de la Experiencia: Adaptar la experiencia del cliente a sus necesidades y preferencias individuales crea una sensación de valor y atención. Esto se puede lograr a través de la segmentación, recomendaciones personalizadas, ofertas relevantes y comunicaciones individualizadas.

3.4.2. Implementación de Programas de Fidelización:

- Recompensas por la Lealtad: Ofrecer descuentos, beneficios exclusivos o experiencias personalizadas a los clientes leales es una forma efectiva de incentivar su permanencia. Los programas de fidelización pueden ser de puntos, niveles o membresías.

- Reconocimiento y Agradecimiento: Mostrar reconocimiento y agradecimiento a los clientes por su fidelidad es esencial. Se puede realizar a través de mensajes personalizados, regalos, eventos especiales o programas de embajadores.

- Programas de Referidos: Incentivar a los clientes a recomendar la empresa a sus amigos y familiares es una estrategia efectiva para captar nuevos clientes y aumentar la retención. Se puede ofrecer descuentos, beneficios o comisiones por cada nuevo cliente referido.

3.4.3. Comunicación Constante y Efectiva:

- Información Clara y Transparente: Mantener a los clientes informados sobre nuevos productos, ofertas, promociones, cambios en la empresa y cualquier otra información relevante es fundamental para mantener una relación sólida.

- Canales de Comunicación Diversos: Utilizar diversos canales de comunicación, como correo electrónico, SMS, redes sociales, teléfono o chat en vivo, permite llegar a los clientes de manera efectiva y personalizada.

- Recopilación de Feedback y Opiniones: Solicitar feedback y opiniones a los clientes sobre su experiencia con la empresa es crucial para identificar áreas de mejora y tomar decisiones estratégicas. Se pueden realizar encuestas, entrevistas o grupos focales.

3.4.4. Análisis de Datos y Segmentación:

- Monitorización del Comportamiento del Cliente: Monitorear el comportamiento del cliente, como sus compras, interacciones y uso del producto, permite identificar patrones y predecir el riesgo de abandono.

- Segmentación de la Base de Clientes: Segmentar la base de clientes en función de sus necesidades, preferencias y comportamiento permite implementar estrategias personalizadas y más efectivas.

- Análisis de la Competencia: Analizar las estrategias de la competencia en cuanto a precios, productos, atención al cliente y programas de fidelización permite identificar oportunidades de mejora y diferenciación.

3.4.5. Inversión en Tecnología y Capacitación:

- Implementación de Software CRM: Un software CRM permite gestionar la información de los clientes, sus interacciones y su historial de compras, lo que facilita la segmentación, la personalización y la atención al cliente.

- Capacitación del Personal: Capacitar al personal en técnicas de atención al cliente, resolución de problemas, comunicación efectiva y venta consultiva es fundamental para ofrecer una experiencia excepcional al cliente.

3.4.6. Cultura Centrada en el Cliente:

- Implementación de una cultura centrada en el cliente en toda la empresa es crucial para que la satisfacción del cliente sea una prioridad en todos los departamentos y niveles de la organización.

- Medición y Seguimiento del Abandono de Clientes: Es fundamental medir y seguir la tasa de abandono de clientes para identificar las causas del problema y evaluar la eficacia de las estrategias implementadas.

3.5. Matriz de Correlaciones:

La matriz de correlaciones es una herramienta estadística que permite medir la relación entre dos variables. Se utiliza para determinar si existe una correlación positiva, negativa o nula entre las variables. Se puede utilizar para identificar las variables que están relacionadas con el abandono de clientes.

Se puede calcular la correlación entre las siguientes variables:

- Satisfacción del cliente: Medida a través de encuestas o entrevistas.

- Tasa de quejas: Número de quejas recibidas por la empresa.

- Tiempo de respuesta al cliente: Tiempo que tarda la empresa en responder a las solicitudes de los clientes.

- Frecuencia de uso del producto o servicio: Número de veces que el cliente utiliza el producto o servicio.

- Valor de la vida útil del cliente: Ingresos totales que genera un cliente durante su relación con la empresa.

El análisis de la matriz de correlaciones puede ayudar a identificar las variables que tienen un mayor impacto en el abandono de clientes. Esta información se puede utilizar para desarrollar estrategias relacionadas a la problemática.

3.6. Up-selling/Cross-selling:

Up-selling consiste en ofrecer al cliente un producto o servicio de mayor valor que el que está comprando. Ejemplo: Ofrecer a un cliente que compra un teléfono móvil un plan de datos de mayor capacidad.

Mientras que Cross-selling consiste en ofrecer al cliente productos o servicios complementarios al que está comprando lo cual puede aumentar su satisfacción y su frecuencia de compra. Ejemplo: Ofrecer a un cliente que compra una cámara de fotos un trípode o una funda para la cámara.

4. Modelos Predictivos: Análisis Profundo y Detallado

Los modelos predictivos son herramientas matemáticas y analíticas que utilizan datos históricos para predecir eventos futuros, y pueden ayudar a las empresas a identificar a los clientes en riesgo de abandono y tomar medidas proactivas para prevenirlo al ofrecer una experiencia más personalizada. Zhang et al. (2006) proponen el uso de técnicas avanzadas como las redes neuronales para predecir el valor vitalicio del cliente, lo que permite a las empresas segmentar clientes y adaptar estrategias de retención personalizadas.

Podríamos también citar el trabajo de Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014) en su libro "Mining of Massive Datasets". Es una referencia ampliamente reconocida en el campo de la minería de datos que proporciona una sólida introducción al uso de árboles de decisión para el análisis predictivo.

A pesar de que los árboles de decisión son modelos intuitivos y fáciles de interpretar, presentan ciertas debilidades que limitan su eficacia. En particular, los árboles de decisión son propensos al sobreajuste, ya que pueden capturar ruidos y peculiaridades del conjunto de datos de entrenamiento, lo que reduce su capacidad de generalizar a nuevos datos. Además, un único árbol de decisión puede ser inestable, ya que pequeñas variaciones en los datos pueden conducir a árboles muy diferentes.

En contraste, el modelo Random Forest, desarrollado Breiman (2001), supera estas limitaciones mediante la construcción de múltiples árboles de decisión a partir de diferentes subconjuntos de datos y características. Este enfoque de conjunto reduce la variabilidad y el sobreajuste, lo que resulta en un modelo más robusto y preciso. El uso de Random Forest permite aprovechar la fuerza combinada de múltiples árboles para mejorar significativamente la precisión de las predicciones y la capacidad de generalización.

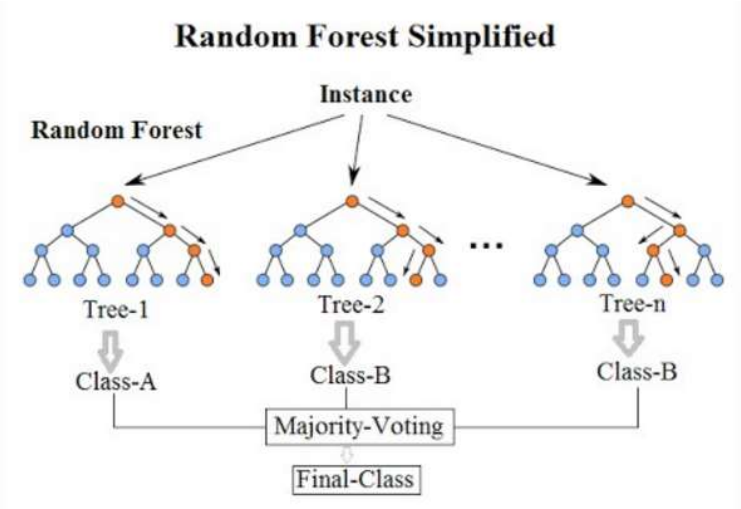
4.1. Tipos de Modelos Predictivos:

En este estudio se profundizará en los siguientes tipos de modelos predictivos utilizados para analizar el abandono de clientes:

- Modelos de Regresión Logística: Estos modelos analizan la relación entre variables independientes (como las características del cliente, su comportamiento de compra o su historial de interacciones) y la variable dependiente (abandono del cliente).

- Random Forest: Estos modelos clasifican a los clientes en diferentes grupos en función de sus características y comportamiento, permitiendo identificar los factores que más influyen en el abandono. Al combinar múltiples árboles de decisión, el Random Forest proporciona mayor robustez y precisión en las predicciones. *Ver figura 3.*

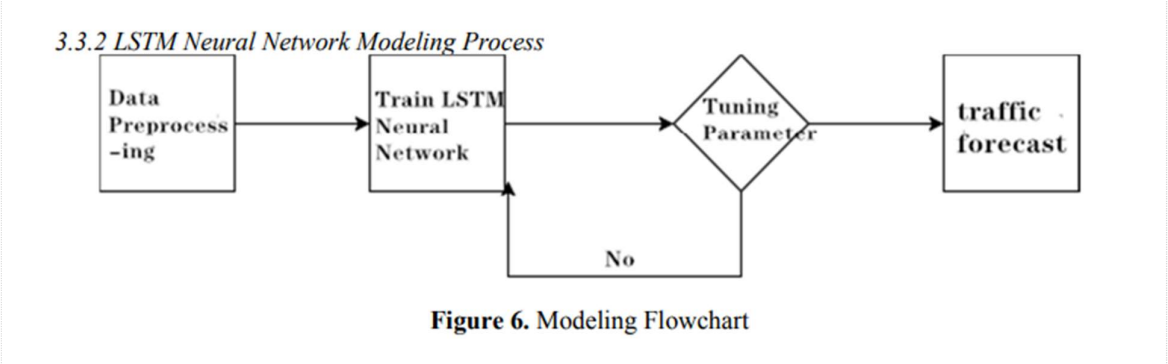
Figura 3. Diagrama para el Modelado de Random Forest



Fuente: Random forest architecture (Breiman, 2001). https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

- Redes Neuronales Artificiales: Estos modelos complejos pueden aprender patrones de datos y realizar predicciones más precisas que los modelos tradicionales. En el diagrama presentado por Zhang, Y. &. (*figura 4*) podemos entender el flujo de procesos que sugiere este modelo.

Figura 4. Diagrama para el Modelado de Red Neuronal



Fuente: Zhang, Y. &. (2006). Customer Lifetime Value Forecasting with Neural Network Approach (Vol. 3). International Conference on Machine Learning and Cybernetics. Pag 8

4.2. Implementación de Modelos Predictivos:

Para implementar con éxito un modelo predictivo de abandono de clientes, es esencial seguir un proceso estructurado que abarque diversas etapas clave. En primer lugar, se requiere la recopilación de datos históricos relacionados con los clientes, abarcando desde sus características demográficas hasta su comportamiento de compra y su historial de abandono. Posteriormente, estos datos deben someterse a una exhaustiva preparación, que incluya la limpieza, el formato adecuado y la organización para garantizar su compatibilidad con el modelo predictivo seleccionado.

Seguidamente, se procede a la selección del modelo más apropiado en función de las necesidades específicas de la empresa y la naturaleza de los datos disponibles. Una vez elegido el modelo, se lleva a cabo su entrenamiento utilizando los datos históricos, permitiendo que el modelo aprenda a identificar los patrones que predicen el abandono con precisión.

La fase de evaluación del modelo es crucial, donde se analiza su desempeño utilizando un conjunto de datos de prueba para verificar su exactitud y efectividad.

4.3. Beneficios de los Modelos Predictivos:

Como hemos mencionado antes, los modelos predictivos ofrecen múltiples beneficios para las empresas. Principalmente, permiten identificar de manera anticipada a los clientes en riesgo de abandono, posibilitando la implementación de medidas proactivas para retenerlos. Esta capacidad de prevención ayuda a reducir la tasa de abandono, incrementar los ingresos y disminuir los costos de adquisición, fortaleciendo la lealtad del cliente.

Además, optimizan el uso de recursos al enfocar los esfuerzos de marketing y retención en los clientes más propensos al abandono, maximizando así el retorno de la inversión. En términos de eficiencia operativa, estos modelos automatizan procesos críticos, como la identificación de clientes en riesgo y la implementación de medidas correctivas, liberando tiempo para actividades estratégicas y proporcionando información valiosa para la toma de decisiones.

Por último, facilitan la personalización de la experiencia del cliente. Al ofrecer mensajes y promociones personalizadas basadas en el riesgo de abandono, aumentan la satisfacción del cliente y fortalecen la relación con la empresa.

4.4. Desafíos de los Modelos Predictivos:

A pesar de sus ventajas, los modelos predictivos enfrentan desafíos que deben abordarse cuidadosamente. La disponibilidad de datos adecuados es fundamental, ya que la construcción de modelos confiables requiere conjuntos de datos extensos y representativos. La falta de datos puede provocar un sobreajuste del modelo y predicciones inexactas. Además, la elección del modelo adecuado es crucial y puede ser un desafío dada la variedad de opciones disponibles[4].

Es importante evaluar el rendimiento del modelo y seleccionar el más adecuado para el problema en cuestión. La interpretación de los resultados también puede ser un desafío, especialmente en modelos complejos como las redes neuronales, donde la comprensión de cómo se llega a las predicciones puede ser difícil[5].

Por último, el mantenimiento del modelo a lo largo del tiempo es esencial, ya que los datos y las relaciones entre variables pueden cambiar, lo que afecta su precisión. Es necesario actualizarlo periódicamente con nuevos datos y monitorear su rendimiento en producción para garantizar su eficacia continua[6].

[4] https://discoverthenew.ituser.es/predictive-analytics/2022/01/los-desafios-de-la-analitica-predictiva
[5] https://es.linkedin.com/pulse/interpretabilidad-en-modelos-de-machine-learning-antonio-soto
[6] https://es.ifixit.com/Device/Machine_Learning_Models

5. Análisis

Para nuestro análisis, nos centramos en una empresa de distribución de consumo masivo, que cuenta con una sólida red logística que abarca todo el país, respaldada por ocho centros de distribución estratégicamente ubicados. Los datos utilizados en nuestro estudio fueron recolectados durante el período comprendido entre 2021 y 2023 para productos del renglón bebida, excluyendo los quiebres de stock como motivo para la no venta. Por motivos de confidencialidad, optamos por no revelar el nombre de la empresa ni detalles específicos sobre sus productos y clientes de manera directa.

5.1. Extracción de los datos

Una vez que descargamos los datos del sistema ERP que utilizan, empleamos herramientas de machine learning para gestionar y analizar la gran cantidad de información recopilada. Nuestro enfoque se centra en examinar los resultados obtenidos mediante el uso de lenguajes de programación como Python y R.

Python:

```python
import pandas as pd
import os

# Obtiene la ruta del folder TFM
ruta_tfm = "/Users/erojas/                                                    "

# Lista para almacenar los DataFrames de cada archivo Excel
lista_dfs = []

# Recorre cada archivo en el folder TFM
for archivo in os.listdir(ruta_tfm):
    # Verifica si el archivo es un archivo Excel
    if archivo.endswith(".xlsx"):
        # Lee el archivo Excel y lo almacena en un DataFrame
        df = pd.read_excel(os.path.join(ruta_tfm, archivo))

        # Agrega el DataFrame a la lista
        lista_dfs.append(df)

# Combina los DataFrames en un solo DataFrame
df_consolidado = pd.concat(lista_dfs)

# Imprime el DataFrame consolidado
print(df_consolidado)
```

R:

```r
library(tidyverse)
library(readxl)
library(forecast)
library(writexl)
library(openxlsx)
library(stats)
library(scales)
library(cluster)

patron <- "*.xlsx"
StartDirectory <- "C:/Users/erojas/[redacted]/
setwd(StartDirectory)
directorio <- list.files(pattern = patron)

datos <- lapply(directorio, read_excel) %>%
  bind_rows() %>%
  rename(Centro = `Ce.`,
         Fecha = `FechaFact.`,
         Cantidad = `Cantidad U`,
         Monto = `Valor neto`) %>%
  mutate(Fecha = as.Date(Fecha, format = "%Y-%m-%d")) %>%
  mutate(año = factor(substr(Fecha, 1, 4)))

final <- datos %>%
  select(-MoPed, -GClt, -UMR, -Mun., -ClFac, -Impuesto, -Entrega, -Alm.,
  drop_na(Material)

# Función genérica para asignar valores numéricos a una columna única
asignar_valor_numerico <- function(columna) {
  unicos <- unique(as.character(columna))
  return(match(as.character(columna), unicos))
}

final$Canal_num <- asignar_valor_numerico(final$Canal)
```

Después de filtrar los datos por año y materiales relevantes para el análisis, eliminamos los valores faltantes (NA) y asignamos valores numéricos a las variables de texto para facilitar su procesamiento en los modelos predictivos. Finalmente, renombramos las variables y eliminamos los espacios vacíos para asegurar la coherencia en la estructura de los datos y facilitar su manipulación en los análisis subsiguientes.

Python:

```python
# Filtra los años 2021, 2022, 2023 de la columna Fecha y Materiales específicos
df_filtrado_fecha = df_consolidado[df_consolidado['FechaFact.'].dt.year.isin([2021, 2022,
2023])]
materiales = [7000, 7003, 7005, 7006, 7007, 7008, 7013, 7014, 7015, 7370, 7310, 7331]
df_filtrado_material = df_filtrado_fecha[df_filtrado_fecha['Material'].isin(materiales)]

# Elimina las filas con valores NaN en la columna 'Canal'
df_filtrado_material = df_filtrado_material.dropna(subset=['Canal'])

# Agrega una nueva columna llamada Canal_num con valores numéricos asignados a los datos de
la columna Canal
canales_uniq = df_filtrado_material['Canal'].unique()
canal_num_map = {canal: i for i, canal in enumerate(canales_uniq)}
df_filtrado_material['Canal_num'] = df_filtrado_material['Canal'].map(canal_num_map)

# Renombramos la tabla filtrada como db_filtrado e imprimimos tabla
db_filtrado = df_filtrado_material
```

```
print(db_filtrado)

# Renombrar la columna 'Valor neto' a 'Monto'y eliminando espacios
db_filtrado.rename(columns=lambda x: x.strip(), inplace=True)
db_filtrado.rename(columns={'Valor neto': 'Monto'}, inplace=True)
```

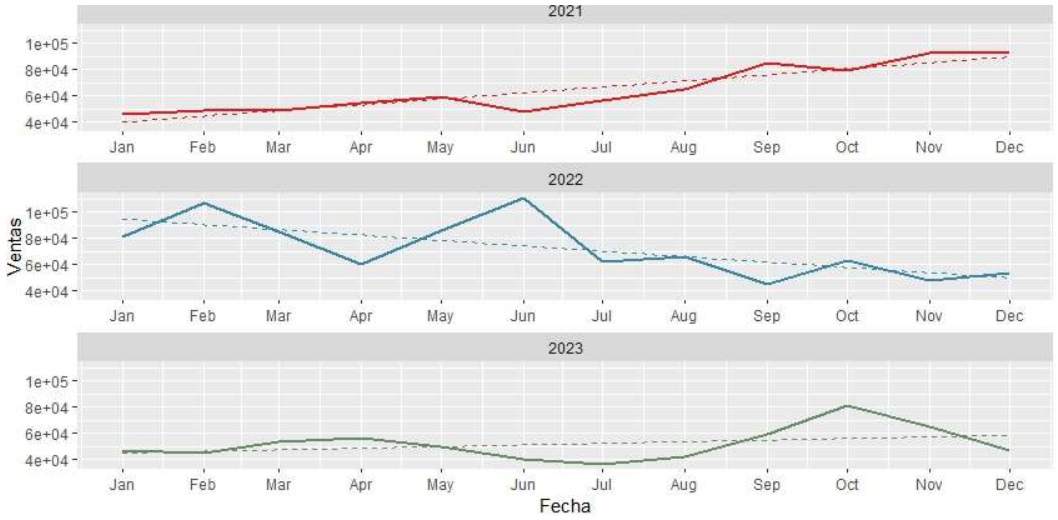
R:

```
# Filtrar los datos para los materiales especificados y los años 2021, 2
final_filtrado <- final %>%
  filter(Material %in% c(7000, 7003, 7005, 7006, 7007, 7008, 7013, 7014,
         año %in% c("2021", "2022", "2023"),
         !is.na(Canal))
```

5.2. Análisis exploratorio de datos

Se analizó el comportamiento de las ventas totales de la empresa a lo largo del tiempo, representado en una serie temporal en la *figura 5*, que muestra de forma comparativa la evolución mensual de las ventas para los años 2021, 2022 y 2023. Se observó un marcado descenso en el año 2022, aunque febrero y junio de dicho año destacaron como meses positivos en términos de ventas. La media más baja se registró en el año 2023.

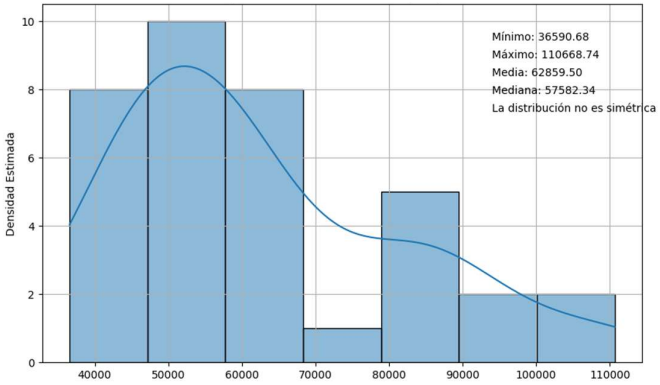
Figura 5. Evolución mensual de las ventas



Fuente: Elaboración propia

Es esencial analizar los estadísticos descriptivos como los mínimos, máximos, medias y medianas de los datos para comprender su distribución y características generales. Estos valores proporcionan información sobre la variabilidad y la centralidad de los datos. Asimismo, se emplearon histogramas (*ver figura 6*) y boxplot para visualizar la distribución de las ventas mensuales en cajas (*ver figura 7*).

Figura 6. Histograma de ventas



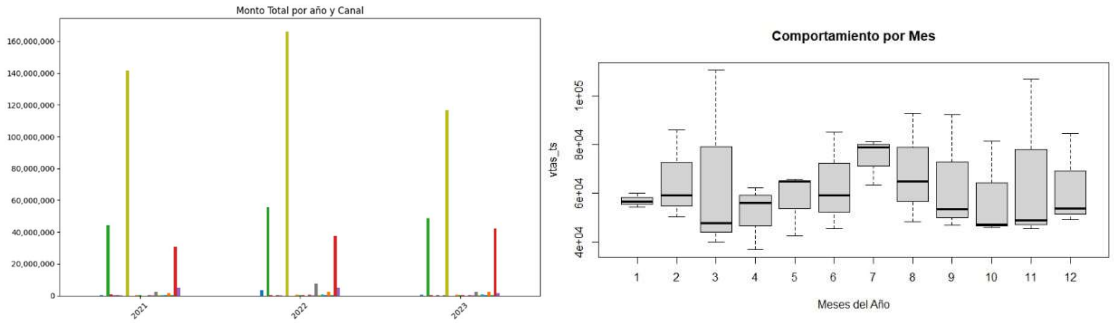
Fuente: Elaboración propia

El gráfico de histograma nos permitió identificar fácilmente que la distribución no es simétrica, mostrando un sesgo hacia la derecha debido a valores relativamente altos en la media (62,859 cajas) con relación a la mediana (57,582 cajas).

Los valores atípicos pueden influir en gran medida en los resultados de los análisis estadísticos, y la asimetría de los datos puede afectar su identificación y manejo. La detección de valores atípicos es crucial en el análisis predictivo.

El análisis se amplió a los canales atendidos, y aunque profundizaremos en este tema más adelante, se reveló que cada año el canal de Colmados se destaca por sus altos montos totales de ventas, superando los $120,000,000 anuales, alcanzando los $160,000,000 en 2022.

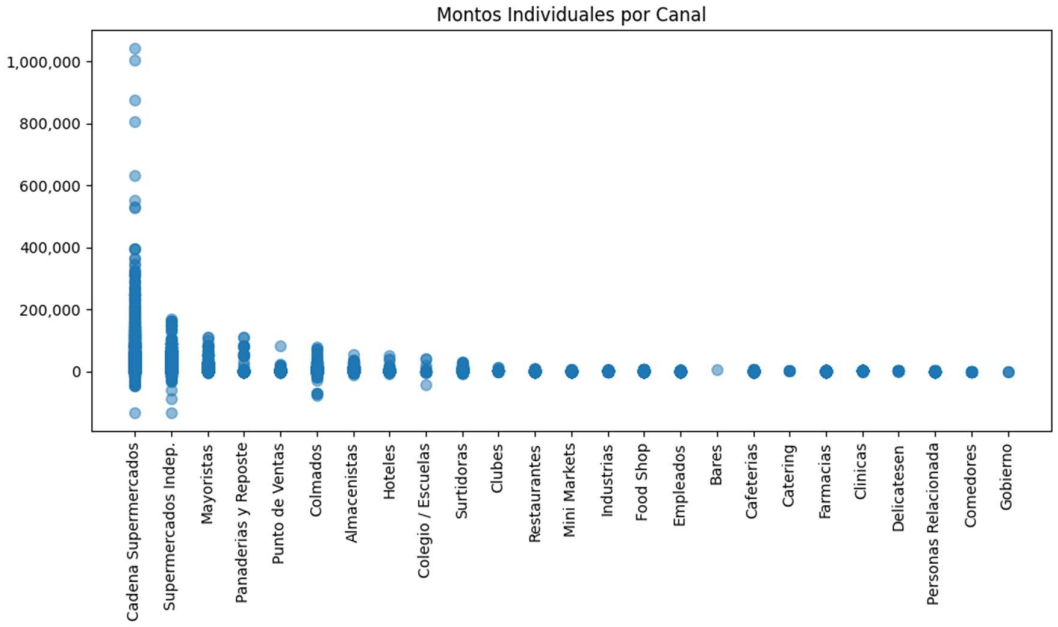
Figura 7. Comportamiento de Ventas



Fuente: Elaboración propia

Se revisaron las ventas individuales de los clientes, notando que la mayoría realizó compras por debajo de $50,000 en un periodo de tres años, con excepción de los clientes de los canales Supermercados Independientes y Cadena de Supermercados, donde varios clientes superaron los $100,000 en dicho período (*ver figura 8*). Esto indica que, a diferencia de otros canales, estos mantienen clientes con un impacto financiero significativo a nivel individual para la empresa, aunque no necesariamente gestionan una gran cantidad de clientes en comparación.

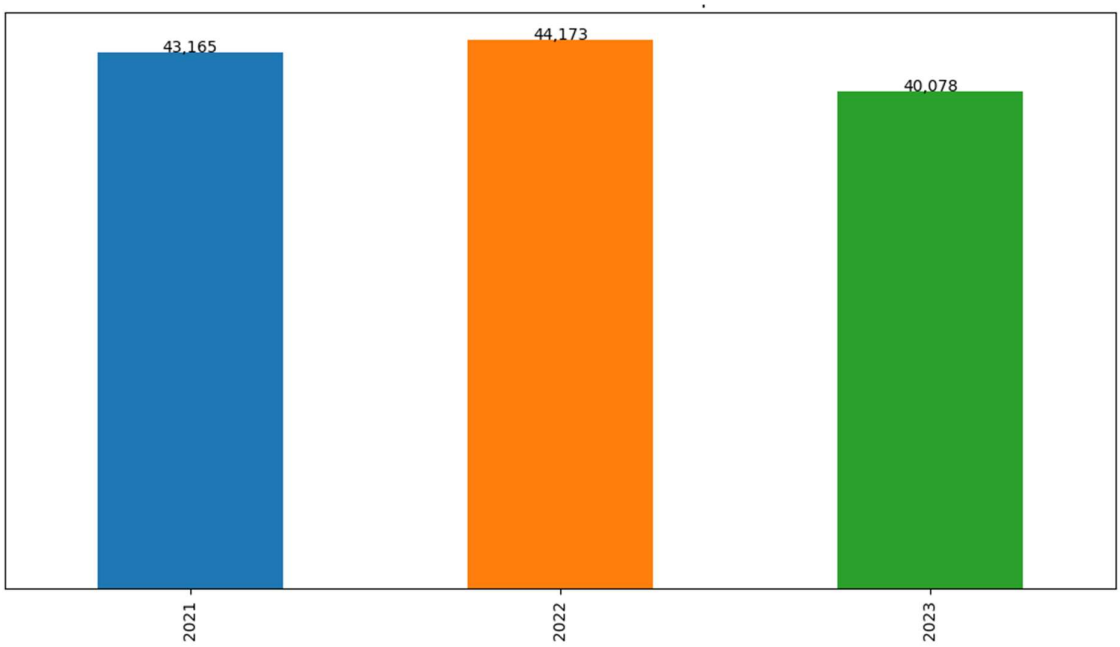
Figura 8. Montos individuales por Canal



Fuente: Elaboración propia

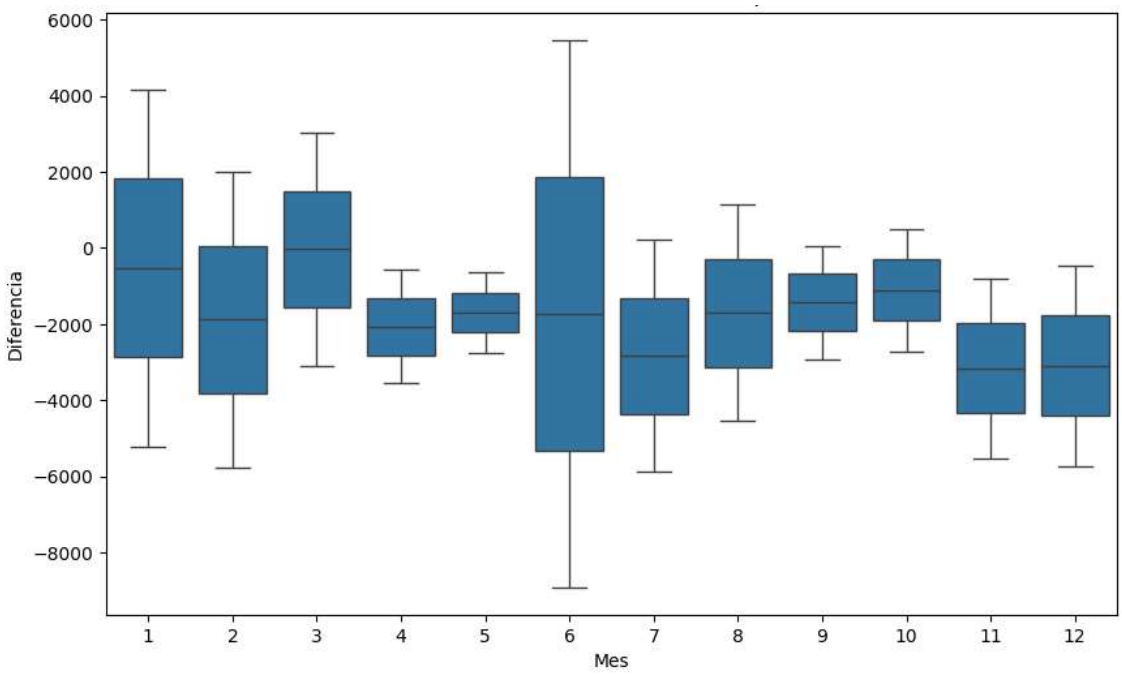
En adición, se analizó la cantidad de clientes atendidos por cada canal y año, notando en la *figura 9* un incremento del 2% en 2022 con respecto a 2021, seguido de una caída del 10% en 2023. Se identificaron los meses con mayor variación en la captación y abandono de clientes, destacando junio como el mes con mayor volatilidad (*ver figura 10*). Los meses de abril, mayo, noviembre y diciembre mostraron abandonos consecutivos de clientes durante el periodo 2021-2023.

Figura 9. Cantidad Total de Clientes atendidos por año



Fuente: Elaboración propia

Figura 10. Variación en el conteo de clientes por mes



Fuente: Elaboración propia

5.2.1. Análisis de Clientes y Canales

El desglose de la cantidad de clientes por canal y año permitió identificar las diferencias año tras año y visualizar con mayor claridad la cantidad de clientes que dejaron de comprar.

Tabla 1. Cantidad de clientes por canal. Diferencias por canal

	Cantidad Clientes x Canal			Diferencia x Canal			% Diferencia x Canal		
Canal	**2021**	**2022**	**2023**	**2021**	**2022**	**2023**	**2021**	**2022**	**2023**
Colmados	41,401	42,451	38,541	0	1,050	-3,910	0.0%	2.5%	-9.2%
Surtidoras	627	573	429	0	-54	-144	0.0%	-8.6%	-25.1%
Empleados	359	379	388	0	20	9	0.0%	5.6%	2.4%
Supermercados Indep.	274	309	282	0	35	-27	0.0%	12.8%	-8.7%
Cafeterias	95	101	71	0	6	-30	0.0%	6.3%	-29.7%
Mini Markets	79	77	67	0	-2	-10	0.0%	-2.5%	-13.0%
Farmacias	74	72	64	0	-2	-8	0.0%	-2.7%	-11.1%
Cadena Supermercados	51	61	61	0	10	0	0.0%	19.6%	0.0%
Personas Relacionada	74	60	57	0	-14	-3	0.0%	-18.9%	-5.0%
Almacenistas	84	76	56	0	-8	-20	0.0%	-9.5%	-26.3%
Mayoristas	28	33	36	0	5	3	0.0%	17.9%	9.1%
Restaurantes	5	9	10	0	4	1	0.0%	80.0%	11.1%
Food Shop	1	6	6	0	5	0	0.0%	500.0%	0.0%
Panaderias y Reposte	2	5	6	0	3	1	0.0%	150.0%	20.0%
Industrias	2	2	4	0	0	2	0.0%	0.0%	100.0%
Punto de Ventas	3	4	4	0	1	0	0.0%	33.3%	0.0%
Colegio / Escuelas	2	2	3	0	0	1	0.0%	0.0%	50.0%
Clinicas	1	2	1	0	1	-1	0.0%	100.0%	-50.0%
Delicatesen	1	1	1	0	0	0	0.0%	0.0%	0.0%
Hoteles	1	1	1	0	0	0	0.0%	0.0%	0.0%

Fuente: Elaboración propia

Se agruparon los clientes según el centro de distribución para identificar áreas geográficas que requieren mayor atención. Se observaron variaciones en la cantidad de clientes por año en cada centro, lo que proporcionó información relevante para la gestión de la distribución.

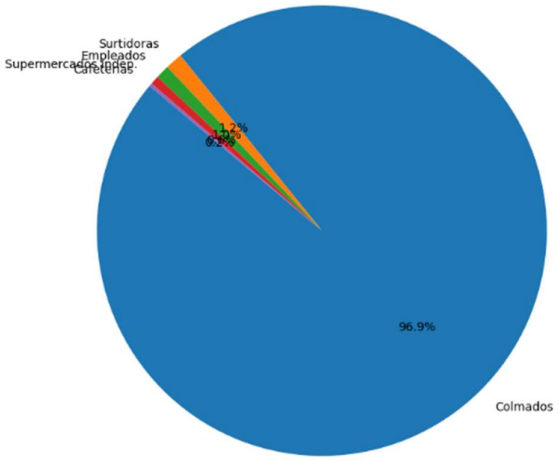
Tabla 2. Cantidad de clientes por centro de distribución. Diferencias por centro

	Cantidad Clientes x Centro Distrib			Diferencia x Centro			% Diferencia x Centro		
Centro	**2021**	**2022**	**2023**	**2021**	**2022**	**2023**	**2021**	**2022**	**2023**
BP02	27,382	28,769	24,691	0	1,387	-4,078	0.0%	5.1%	-14.2%
BP06	14,994	14,545	13,717	0	-449	-828	0.0%	-3.0%	-5.7%
BP15	0	0	1,893	0	0	1,893	0.0%	0.0%	########
BP11	434	439	442	0	5	3	0.0%	1.2%	0.7%
BP01	290	316	308	0	26	-8	0.0%	9.0%	-2.5%
BP03	96	105	91	0	9	-14	0.0%	9.4%	-13.3%
BP09	0	22	43	0	22	21	0.0%	21900.0%	95.5%
BP14	1	1	7	0	0	6	0.0%	0.0%	600.0%

Fuente: Elaboración propia

Al examinar los cinco principales canales que atienden a la mayoría de los clientes y su participación en la cartera total, se destaca en la *figura 11* que el 96.9% de los clientes pertenecen al canal de Colmados, seguido por Surtidoras con el 1.2%, Empleados con el 1%, Supermercados Independientes con el 0.6%, y Cafeterías con el 0.2%. Este análisis se presenta visualmente a través de un gráfico de pastel que muestra la distribución porcentual de los clientes entre estos cinco canales principales.

Figura 11. Top 5 Canales por cantidad de clientes



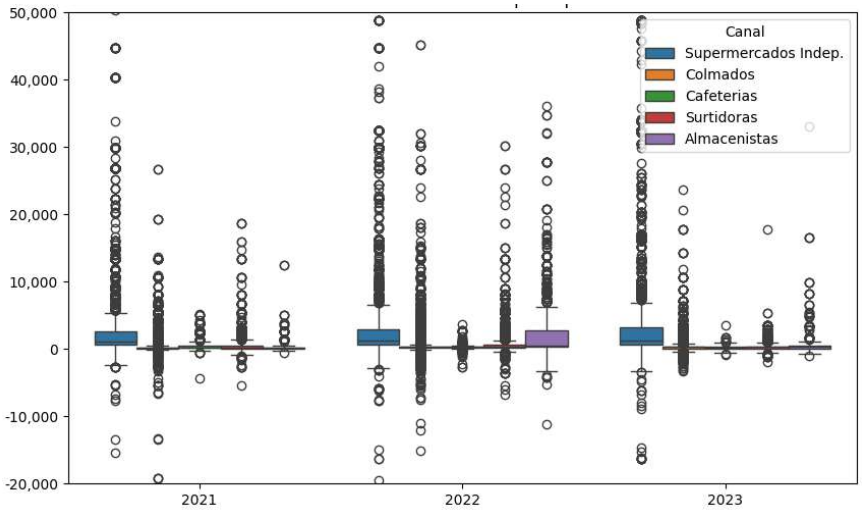
Fuente: Elaboración propia

5.2.1.1. Los Outliers (valores atípicos)

Los outliers son puntos de datos que se destacan notablemente del resto del conjunto, a menudo representados como puntos individuales fuera de los bigotes en un diagrama de caja o boxplot. Pueden surgir por diversas razones, como errores de medición, variabilidad natural en algunos fenómenos o eventos excepcionales. Interpretar la presencia de outliers es crucial para una comprensión precisa de los datos. Por un lado, su detección puede indicar errores en la recolección o almacenamiento de datos, lo que subraya la necesidad de una revisión exhaustiva de los datos originales para corregirlos. Por otro lado, la presencia de outliers puede señalar una variabilidad extrema dentro del conjunto de datos o representar eventos únicos que requieren un análisis detallado para entender su impacto en las conclusiones derivadas de los datos.

Se identificó que estos cinco principales canales muestran ventas individuales con numerosos valores atípicos, como se observa en el boxplot de la *figura 12*. La existencia de una cantidad considerable de valores atípicos puede distorsionar la interpretación de la tendencia central y la dispersión de los datos, especialmente cuando se utiliza la media como medida de tendencia central. En presencia de valores atípicos significativos, es recomendable considerar el uso de medidas de tendencia central más robustas, como la mediana o la moda.

Figura 12. Monto de ventas Individuales por top 5 Canales



Fuente: Elaboración propia

5.3. Análisis RFM

El análisis RFM (Recencia, Frecuencia, Monto) constituye una técnica fundamental en el ámbito del marketing, empleada para segmentar clientes según sus patrones de compra. Este enfoque permite a las empresas dirigir de manera más efectiva sus recursos, mejorar la retención de clientes y aumentar su rentabilidad.

5.3.1. Variables Clave del Análisis RFM

- Recencia (Recency): Indica el tiempo transcurrido desde la última compra del cliente, siendo los clientes más recientes considerados de mayor valor debido a su interacción actual con la marca.
- Frecuencia (Frequency): Representa el número promedio de compras realizadas por el cliente en un periodo determinado, siendo los clientes con mayor frecuencia de compra considerados más leales y comprometidos.
- Monto (Monetary): Refleja el valor total de las compras efectuadas por el cliente a lo largo del tiempo, siendo aquellos que gastan más considerados más rentables para la empresa.

5.3.2. Asignación de valores y segmentación de clientes

Cada cliente recibe una puntuación en cada dimensión (Recencia, Frecuencia, Monto) según su historial de compras. Estas puntuaciones pueden ser valores numéricos simples o categorizaciones que reflejen su nivel relativo en cada dimensión.

La combinación de estas puntuaciones permite crear segmentos de clientes basados en su perfil RFM. Los segmentos comunes incluyen:

- Clientes Top: Alta Recencia, alta Frecuencia y alto Monto. Requieren estrategias de retención y fidelización.
- Clientes Prometedores: Alta Recencia y alta Frecuencia, pero bajo Monto. Representan potenciales clientes VIP y requieren estrategias para estimular sus compras.
- Clientes en Riesgo: Baja Recencia, baja Frecuencia y bajo Monto. Necesitan estrategias de reactivación para recuperar su interés y participación.

Tabla 3. Valores RFM para cada cliente

Solic.	Canal_num	Monto	Recency	Frequenc
100039	1	1,562,538	17	64
100044	1	2,930,593	16	61
100045	1	143,482	24	42
100047	1	83,410	32	10
100052	1	331,571	2	43
100053	1	849,758	2	51
100054	1	514,885	51	18
100056	1	1,084,567	31	55
100057	1	1,176,899	16	66
100058	1	339,993	2	55
100075	2	1,391,493	943	21
100079	2	126,886	832	1
100082	2	134,031	832	1
100091	1	314,614	3	46
100095	3	41,562	564	6
100096	1	429,130	16	43

Fuente: Elaboración propia

5.3.3. Aplicaciones y Beneficios del Análisis RFM

- Personalización de campañas de marketing: Dirigir campañas de marketing específicas a cada segmento de clientes en función de su perfil RFM. Ofrecer descuentos, cupones o beneficios especiales a los clientes Top o Prometedores para incentivar su compra.
- Programas de fidelización adaptados. Diseñar programas de fidelización personalizados para recompensar a los clientes Top y fomentar su lealtad.
- Mejorar la rentabilidad: Identificar y prevenir el abandono de clientes, así como mejorar la rentabilidad de las campañas de marketing y programas de fidelización al concentrarse en los clientes más valiosos.

5.3.4. Limitaciones del Análisis RFM

- Dependencia del historial de compras: El análisis RFM se fundamenta únicamente en datos históricos de compras, sin considerar factores externos como preferencias cambiantes o tendencias del mercado.
- Posible simplicidad excesiva: Existe el riesgo de simplificar demasiado el comportamiento del cliente, sin tener en cuenta la complejidad de sus motivaciones y decisiones de compra.
- Necesidad de datos actualizados: Es vital contar con datos de compra precisos y actualizados para mantener una segmentación de clientes precisa y relevante.

5.3.5. Ranking RFM

El Ranking RFM representa un avance en la segmentación de clientes al combinar las variables RFM para asignar una puntuación individual a cada cliente. Este enfoque jerárquico implica la fusión de las dimensiones RFM con ponderaciones personalizadas, permitiendo a las empresas optimizar sus estrategias de marketing y centrarse en los clientes más valiosos.

Después de asignar las puntuaciones en Recency, Frequency y Monetary (Monto), se normalizan para asegurar una escala comparable entre clientes. Este proceso facilita la comparación y el ordenamiento de los clientes según su puntuación RFM, lo que permite una priorización precisa basada en el valor individual y el potencial de crecimiento.

El ranking RFM no solo optimiza la segmentación, sino que también ofrece información valiosa y detallada que las empresas pueden utilizar para tomar decisiones informadas y efectivas en marketing, ventas y fidelización de clientes. Este proceso se implementó utilizando machine

learning en Python, resultando en la creación de variables como Recency_Rank, Frequency_Rank y Monto_Rank en un nuevo conjunto de datos denominado db_rfm.

```
db_rfm
```
[8] ✓ 0.0s

	Solic.	Monto	Frequency	Recency	Canal_num	Recency_Rank	Frequency_Rank	Monto_Rank
0	939	14205.57	67	10	7	51994.5	55202.0	47488.0
1	1632	1671.53	9	158	7	27092.0	27726.0	17986.0
2	1763	101.05	1	172	7	26238.0	3217.0	1740.5
3	2416	315.47	2	877	7	2693.0	8816.5	4892.0
4	2897	5573.98	37	457	7	13531.5	50410.0	33978.0
...	...	...	...	...	...	...	...	...
56806	700362	2581.22	7	114	6	30067.0	23820.0	23041.0
56807	1200045	31456.50	4	146	2	27679.5	16273.5	54142.0
56808	POS01	554579.92	505	0	5	56810.5	56809.0	56696.0
56809	POS03	4402061.58	935	1	5	56461.5	56810.0	56808.0
56810	POS04	1354965.42	1003	1	5	56461.5	56811.0	56767.0

5.3.6. Normalización de Datos

La normalización de datos constituye un paso crucial en la preparación de datos para el análisis de clustering. Este proceso implica transformar los valores de cada columna a un rango específico, comúnmente entre 0 y 1. Al eliminar las disparidades de unidades y mejorar la comparabilidad de escalas, la normalización garantiza que los algoritmos de clustering funcionen de manera eficiente y encuentren agrupamientos más precisos y relevantes. Además, facilita la interpretación de los resultados y la visualización de los datos.

La normalización ofrece diversas ventajas para mejorar la efectividad del clustering:

1. Comparabilidad de Escalas:
    a. *Eliminación de Diferencias de Unidades*: La normalización elimina las disparidades de unidades entre las variables, permitiendo comparar valores de diferentes dimensiones sin que una variable predomine en el análisis debido a su mayor escala.
    b. *Interpretación Uniforme*: Facilita la interpretación de los resultados del clustering al mantener todas las variables en el mismo rango numérico.
2. Distancia Euclidiana Significativa:
    a. *Distancias Proporcionales*: La normalización asegura que las distancias euclidianas, ampliamente utilizadas en algoritmos de clustering como k-

means, reflejen la magnitud relativa de las diferencias entre los puntos de datos, no solo su magnitud absoluta.

b. *Mejora de la Convergencia*: Ayuda a que los algoritmos de clustering converjan más rápidamente y encuentren agrupamientos más estables, ya que las variables con mayor escala no dominan el cálculo de distancias.

3. Evitar la Dominancia de Variables:
    a. *Prevención de Sesgos*: Impide que variables con valores extremadamente altos o bajos dominen el análisis de clustering, lo que podría resultar en agrupamientos sesgados o distorsionados.
    b. *Representación Justa*: Garantiza que todas las variables tengan una representación equitativa en el análisis de clustering, sin que una variable influya desproporcionadamente en los resultados.

En nuestro análisis, agregamos las variables Recency_Normalized, Frequency_Normalized y Monto_Normalized a partir de Recency_Rank, Frequency_Rank y Monto_Rank.

```
db_rfm
[8]  ✓ 0.0s
```

	Solic.	Monto	Frequency	Recency	Canal_num	Recency_Rank	Frequency_Rank	Monto_Rank	Recency_Normalized	Frequency_Normalized	Monto_Normalized
0	939	14205.57	67	10	7	51994.5	55202.0	47488.0	0.915218	0.969978	0.835892
1	1632	1671.53	9	158	7	27092.0	27726.0	17986.0	0.476828	0.457309	0.316582
2	1763	101.05	1	172	7	26238.0	3217.0	1740.5	0.461794	0.000000	0.030620
3	2416	315.47	2	877	7	2693.0	8816.5	4892.0	0.047303	0.104480	0.086094
4	2897	5573.98	37	457	7	13531.5	50410.0	33978.0	0.238106	0.880565	0.598081
...	...	...	...	...	...	...	...	...	...	...	...
56806	700362	2581.22	7	114	6	30067.0	23820.0	23041.0	0.529201	0.384427	0.405562
56807	1200045	31456.50	4	146	2	27679.5	16273.5	54142.0	0.487171	0.243619	0.953019
56808	POS01	554579.92	505	0	5	56810.5	56809.0	56696.0	1.000000	0.999963	0.997976
56809	POS03	4402061.58	935	1	5	56461.5	56810.0	56808.0	0.993856	0.999981	0.999947
56810	POS04	1354965.42	1003	1	5	56461.5	56811.0	56767.0	0.993856	1.000000	0.999225

5.4. Customer Lifetime Value (CLV)

El CLV es una métrica crucial para las empresas que desean evaluar la rentabilidad a largo plazo de sus clientes. Representa el valor total que un cliente aporta a la empresa durante su relación comercial.

Un CLV alto indica que un cliente es más valioso para la empresa, generando mayores ingresos y manteniendo una relación comercial duradera. Por el contrario, un CLV bajo sugiere que un cliente es menos rentable y puede requerir estrategias de retención o reactivación.

5.4.1. Cálculo del CLV

Existen varios métodos para calcular el CLV, pero un enfoque común implica considerar tres factores clave:

- Valor Promedio de Compra (Average Purchase Value): Se calcula dividiendo el total de los ingresos generados por todas las compras entre el número total de facturas realizadas.
- Frecuencia de Compra (Purchase Frequency): Representa el número promedio de compras realizadas por el cliente en un período determinado.
- Duración de la Relación (Customer Lifetime): Se calcula restando la fecha de la primera compra de cada cliente a la fecha actual y luego calculando el promedio de todas esas diferencias.

5.4.2. Fórmula

*CLV = (Valor promedio de compra * Frecuencia de compra) * Duración de la relación*

En su libro sobre métricas de marketing "Marketing Metrics" (Farris, 2010), Paul Farris, Peter Van Ryckere y Robert Bell presentan una definición detallada del CLV y su fórmula básica. Otro libro relevante es "Customer Lifetime Value: Marketing for the Long Term" (Fader, 2016) de Peter S. Fader, que profundiza en el concepto del CLV y sus métodos de cálculo.

Wayne L. Winston en "Marketing Analytics: Using Data to Make Better Marketing Decisions" (Winston, 2019) también ofrece una introducción completa al análisis de marketing, incluyendo la medición del CLV y su aplicación estratégica.

Algunos artículos académicos importantes sobre el tema son:

- "The Value of Customer Lifetime"[7] de Andrew Ehrenberg y Richard T. Smith.
- "Customer Lifetime Value: A Multidisciplinary Review"[8] de Bernd H. Schmittberger and Christian H. Zentes.

[7] Ehrenberg, A. S. C., & Smith, R. T. (1988). The value of customer lifetime. Journal of Marketing Research, 25(4), 246-257.
[8] Schmittberger, B. H., & Zentes, C. H. (1998). Customer lifetime value: A multidisciplinary review. Journal of Marketing Research, 35(3), 373-389.

- "A Framework for Customer Lifetime Value Measurement"[9] de Venkatesh Narasimhan and Rajendra Srivastava.

Para nuestra tesis y análisis, adoptaremos una fórmula basada en los valores definidos según el módulo 6 del temario propuesto por la profesora Rocío González. La fórmula es:

$$CLV = 0.5(Monto\_Rank) + 0.25(Frequency\_Rank) + 0.25(Recency\_Rank)$$

Las ponderaciones en la fórmula reflejan una decisión estratégica sobre qué dimensiones del comportamiento del cliente tienen más peso en la evaluación de su valor a largo plazo para la empresa. El uso de 0.5 para Monto_Rank indica que el valor monetario de las compras es el factor más significativo, mientras que Frequency_Rank y Recency_Rank, con 0.25 cada uno, contribuyen de manera equitativa pero menos prominente al cálculo general del CLV.

Esta fórmula proporciona una aproximación al CLV, si bien es simple y no considera todos los factores específicos del negocio, como los márgenes de ganancia o los costos de adquisición de clientes. Las ponderaciones asignadas a cada dimensión son fijas y podrían necesitar ajustes según las particularidades de cada empresa o sector.

5.5. Clusterización de Datos

Clusterizar es una técnica de análisis de datos que consiste en agrupar puntos de datos en función de sus similitudes. Los grupos resultantes se denominan clústeres y comparten características comunes entre sí.

Emplearemos el método del codo y el coeficiente de Silhouette para determinar el número óptimo de clústeres en el conjunto de datos a partir de Recency_Normalized, Frequency_Normalized y Monto_Normalized encontradas en db_rfm.

5.5.1. Método del Codo y Coeficiente de Silhouette

El método del codo es una técnica gráfica popular que se basa en observar la curva que representa la suma total dentro del error cuadrático (SSE) en función del número de clústeres

[9] Narasimhan, C., & Srivastava, R. K. (1999). A framework for customer lifetime value measurement and management. Journal of Marketing Research, 36(1), 26-40.

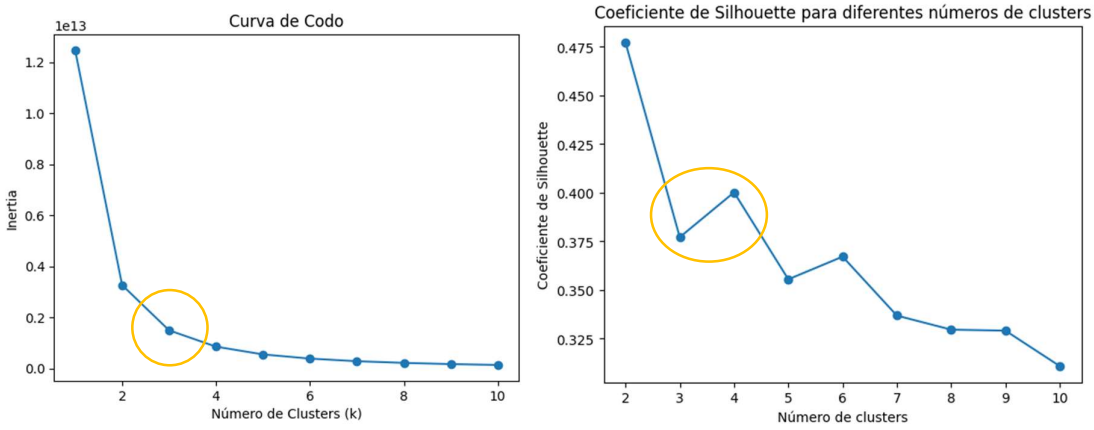
(k). Aunque es una herramienta simple y efectiva, es importante considerar sus limitaciones y complementar su uso con otras métricas y técnicas de evaluación.

Una vez identificado el número óptimo de clústeres, evaluaremos la calidad de los clústeres resultantes utilizando métricas como el coeficiente de Silhouette o la separación entre clústeres. El coeficiente de Silhouette sirve para determinar qué tan bien se agrupan los puntos de datos dentro de un clúster y qué tan separados están de los puntos de otros clústeres.

```
# Seleccionar las variables relevantes y normalizarlas
variables = ['Recency_Normalized', 'Frequency_Normalized', 'Monto_Normalized']
X = db_rfm[variables]

# Aplicar el método del codo para encontrar el k óptimo
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, n_init=10, random_state=42)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)

# Calcular el coeficiente de Silhouette
    silhouette_avg = silhouette_score(X, cluster_labels)
    silhouette_scores.append(silhouette_avg)
```

La curva del Codo indica que 3 es el número óptimo de clústeres. El coeficiente de Silhouette indica que pueden ser 3 ó 4. Escogemos entonces trabajar con k=3. *Ver figura 13.*

Figura 13. Curva del Codo y Coeficiente de Silhouette



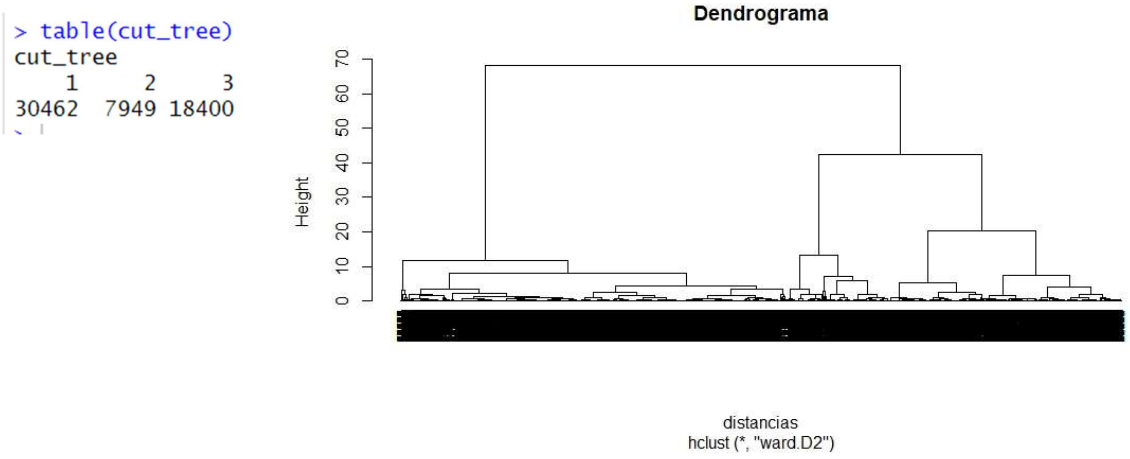
Fuente: Elaboración propia

5.5.1.1. Dendograma

Un dendrograma es una representación gráfica en forma de árbol que se utiliza para visualizar la jerarquía de clústeres resultante de un análisis de clusterización. En otras palabras, el dendrograma muestra cómo se han ido agrupando los datos paso a paso durante el proceso de clusterización, reflejando las similitudes y distancias entre los clústeres.

El dendrograma (*ver figura 14*) puede ayudar a determinar el número óptimo de clústeres para el análisis, observando los niveles de corte en los que se produce un cambio significativo en la distancia entre clústeres.

Figura 14. Dendograma



Fuente: Elaboración propia

5.5.2. Caracterización de los Clústeres

Para facilitar la identificación de cada clúster utilizamos la mediana de cada uno, que es una medida de tendencia central robusta ante los valores atípicos.

Tabla 4. Valores mediana RFM para cada clúster y conteo

Cluster	R	F	M
0	0.196613	0.10448	0.167242
1	0.790017	0.82985	0.830857
2	0.505576	0.457309	0.498733

Cluster	Conteo
0	18,628
1	18,459
2	19,724

Fuente: Elaboración propia

5.5.2.1. Descripción general de cada clúster:

Clúster 0:

Recencia (R): 0.196613 (bajo)	Frecuencia (F): 0.10448 (bajo)	Monetario (M): 0.167242 (bajo)

Este clúster se caracteriza por tener valores bajos en las tres variables. Esto podría indicar que los clientes en este grupo no han realizado compras recientes, las mismas han sido poco frecuentes y de menor valor monetario. Los llamaremos RIESGO.

Posibles características de estos clientes:

- Clientes nuevos o poco frecuentes.
- Compras de bajo valor.
- Posiblemente sensibles al precio.

Clúster 1:

Recencia (R): 0.790017 (alto)	Frecuencia (F): 0.82985 (alto)	Monetario (M): 0.830857 (alto)

Los clientes de este clúster se caracterizan por tener valores altos en las tres variables. Esto apunta a que los miembros de este grupo han realizado compras recientes, han sido frecuentes y de mayor valor monetario. Los llamaremos TOP.

Posibles características de los clientes:

- Clientes leales o frecuentes.
- Compras de alto valor.
- Posiblemente menos sensibles al precio.
- Clientes valiosos para la empresa.

Clúster 2:

Recencia (R): 0.505576 (medio)	Frecuencia (F): 0.457309 (medio)	Monetario (M): 0.498733 (medio)

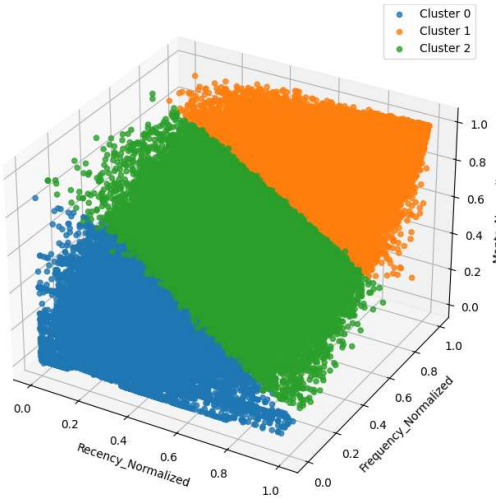
Se caracteriza este clúster por tener valores intermedios en las tres variables. Los participantes en este grupo tienen un comportamiento intermedio entre los clústeres 0 y 1. Los llamaremos PROMETEDORES.

Posibles características de los clientes:

- Clientes con un comportamiento de compra variable.
- Compras de valor medio.
- Posiblemente sensibles al precio en algunos casos.

Hemos observado adicionalmente, que la distribución de clientes por clúster muestra un equilibrio notable entre ellos. El grafico 3D de la *figura 15* nos permite apreciar el balance.

Figura 15. Clusterización 3D



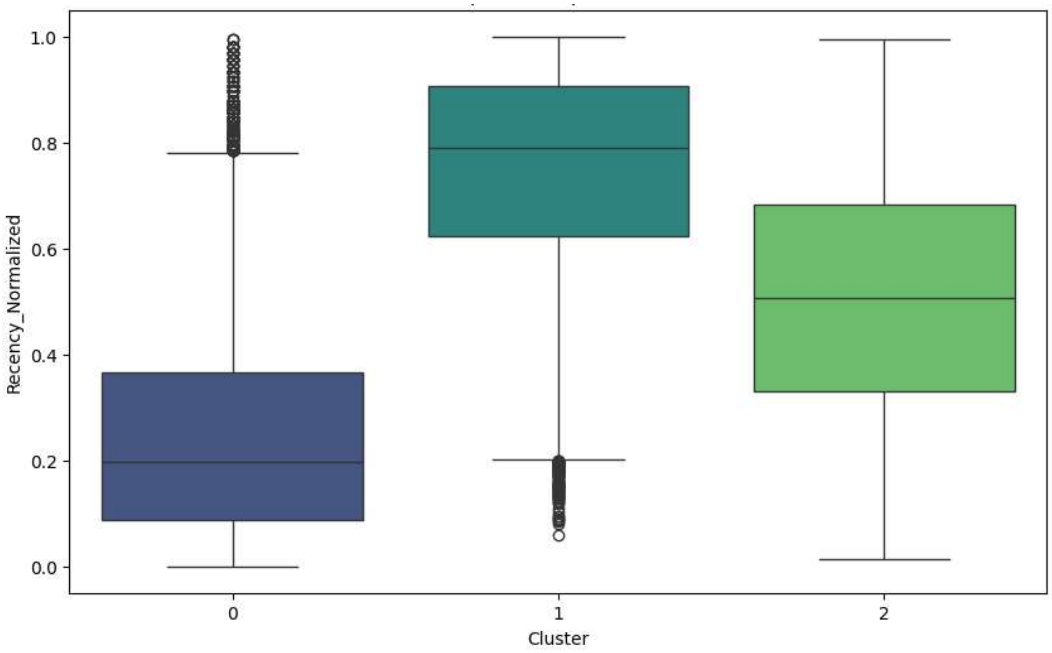
Fuente: Elaboración propia

En las *figuras 16,17,18*, analizamos el boxplot de cada clúster utilizando el Rango Intercuartílico (IQR) para identificar patrones y tendencias en los datos. Este gráfico facilita la comprensión de las diferencias entre cada clúster.

El IQR de los clientes en el clúster RIESGO está entre 0.08 y 0.38. Para los clientes TOP, el IQR varía de 0.62 a 0.91. En cuanto a los clientes PROMETEDORES, su IQR se sitúa entre 0.35 y 0.68.

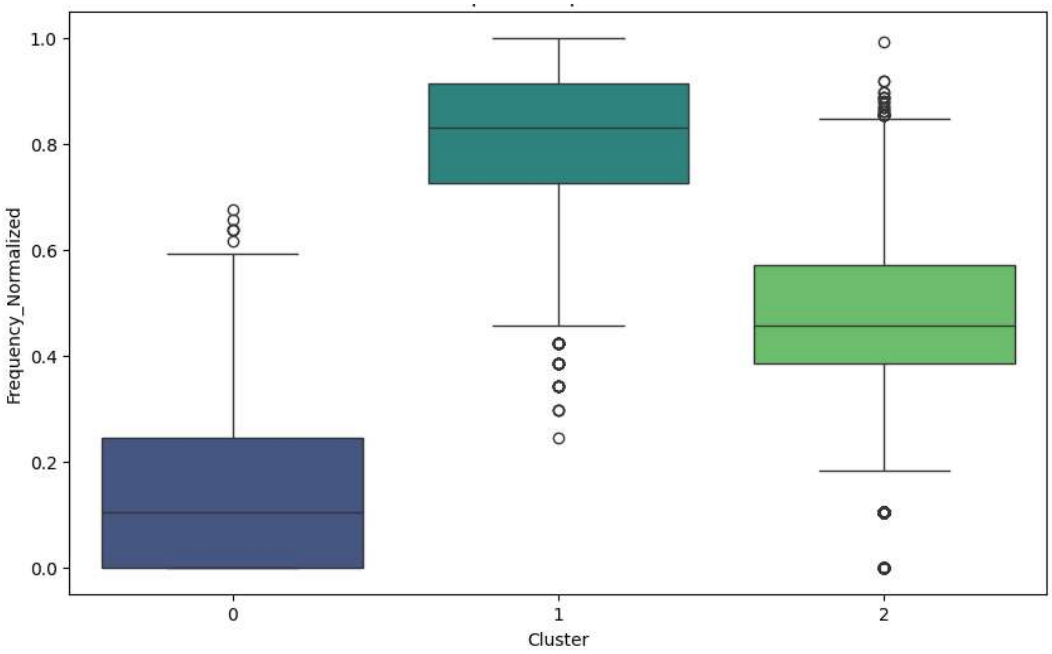
El IQR es una medida de dispersión utilizada en los boxplots para representar el rango del 50% central de los datos, excluyendo los valores atípicos. Esta herramienta nos ayuda significativamente a identificar patrones y tendencias dentro del conjunto de datos analizado.

Figura 16. Boxplot Recency por Cluster



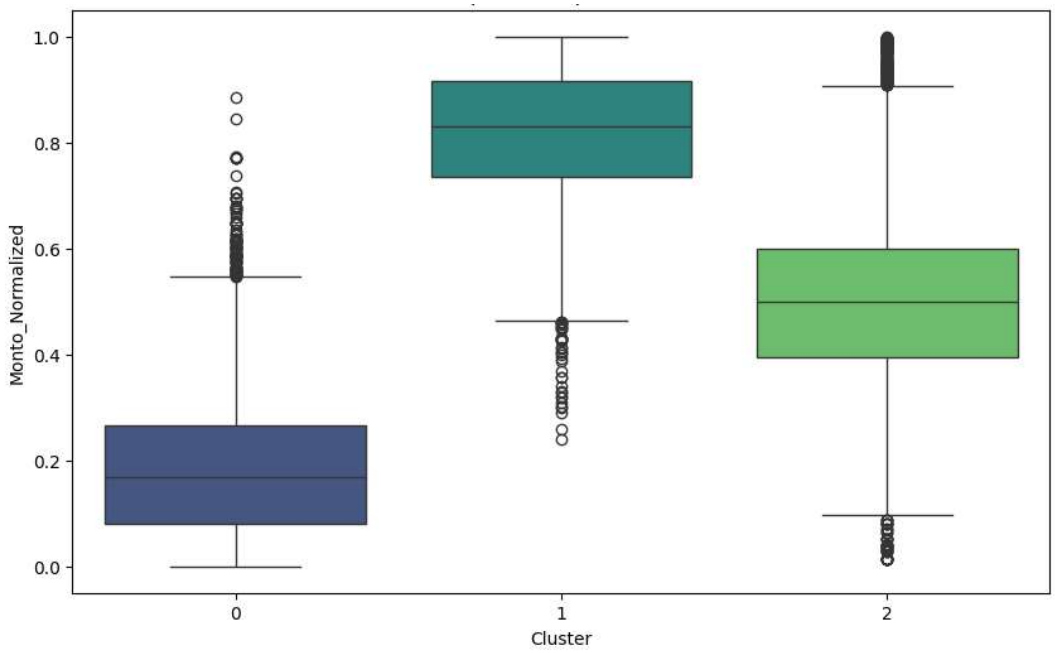
Fuente: Elaboración propia

Figura 17. Boxplot Frequency por Cluster



Fuente: Elaboración propia

Figura 18. Boxplot Monetary por Cluster



Fuente: Elaboración propia

5.5.3. Identificación de Zonas de Venta con Clientes en RIESGO

Asimismo, identificamos las zonas de venta que contienen a los clientes clasificados como RIESGO, permitiéndonos así focalizar estrategias específicas de retención y reactivación en esas áreas.

Tabla 5. Top 5 ZV con clientes en RIESGO

ZV	Conteo
SFM001	1,324
VGA001	1,274
SDN001	1,204
SDE001	699
SDO010	685

Fuente: Elaboración propia

5.5.4. Clusterización basada en CLV

Realizamos un análisis adicional de clusterización utilizando exclusivamente los valores de CLV. Observamos similitudes significativas con el análisis anterior basado en RFM, validando así las características comunes para cada clúster tanto a través de RFM como de CLV.

Tabla 6. Valores de la mediana RFM para cada clúster CLV

CLV_Cluster	R	F	M
0	0.206595	0.10448	0.167902
1	0.790017	0.82985	0.835821
2	0.503085	0.488348	0.505774

Fuente: Elaboración propia

5.5.5. Análisis de la Matriz ANOVA

La matriz ANOVA (Análisis de Varianza) es una técnica estadística muy útil para evaluar las diferencias entre las medias de dos o más grupos. Su aplicación común radica en la comparación de las medias de un factor categórico (llamado factor de grupo) para determinar si existe alguna diferencia significativa entre ellas. Este análisis descompone la variación total en los datos en diversas fuentes de variación, como la variación entre grupos y la variación dentro de los grupos.

5.5.5.1. Coeficiente de Correlación de Eta Cuadrado

En el contexto de la correlación entre variables no numéricas, como por ejemplo Canal y Zona de Venta, la matriz ANOVA se utiliza para calcular el coeficiente de correlación de eta cuadrado. Este coeficiente mide la fuerza de la asociación entre las categorías del factor de grupo (los diferentes canales) y los valores de la variable dependiente (por ejemplo, Frequency).

	Source	SS	DF	MS	F	p-unc	np2
0	Canal	1.273080e+10	24	5.304499e+08	349347.395877	0.0	0.838302
1	Within	2.455615e+09	1617236	1.518402e+03	NaN	NaN	NaN

Interpretación de los Resultados

- F (F-value): Es la relación entre la varianza entre grupos y la varianza dentro de los grupos. En este caso, el valor de F es muy alto, lo que indica que hay una diferencia significativa entre los diferentes canales en términos de su efecto en la variable "Frequency".
- p-unc (p-value): Es el valor p obtenido del análisis de ANOVA. Indica la probabilidad de obtener el valor de F observado si la hipótesis nula (que no hay diferencia entre los grupos) es verdadera. Un valor p bajo (generalmente <0.05) sugiere que la diferencia entre los grupos es estadísticamente significativa.
- np2 (Partial eta-squared): Es una medida de la fuerza de la asociación entre la variable independiente (Canal) y la variable dependiente (Frequency), que indica la proporción de la variabilidad total de Frequency explicada por la variable Canal. En este caso, el valor de np2 es alto (0.838), lo que sugiere que la variable "Canal" explica una gran parte de la variabilidad en "Frequency".

Al filtrar los datos únicamente para los clientes en RIESGO, observamos los siguientes resultados significativos:

- La estadística F es 79.83, lo que indica una razón alta de variabilidad entre los grupos respecto a la variabilidad dentro de los grupos. Este hallazgo sugiere una fuerte asociación entre "Canal" y "Frequency".
- El valor p es prácticamente cero (8.786005e-260), lo que respalda la idea de que la relación entre "Canal" y "Frequency" es estadísticamente significativa.
- El eta cuadrado parcial es 0.017, lo que indica que el 1.7% de la variabilidad en "Frequency" se atribuye a la variable "Canal". Esta asociación moderada resalta la influencia de los canales en la frecuencia de ventas.

	Source	SS	DF	MS	F	p-unc	np2
0	Canal	6060.676904	16	378.792307	79.825243	8.786005e-260	0.017197
1	Within	346361.980544	72991	4.745270	NaN	NaN	NaN

5.5.5.2. Consideraciones Adicionales:

Aunque una asociación moderada entre el canal y la frecuencia de ventas sugiere que aumentar la frecuencia de ventas puede no ser la solución única para abordar el abandono, es crucial considerar otros factores. La satisfacción del cliente, la calidad del producto o servicio, la competencia en el mercado y las estrategias de marketing son solo algunos de los elementos que también pueden influir en el comportamiento de los clientes.

Para abordar el abandono de clientes de manera efectiva, es fundamental realizar un análisis integral que contemple múltiples factores y emplee enfoques integrados. Esto implica acciones dirigidas tanto a mejorar la frecuencia de ventas como a abordar otros aspectos del ciclo de vida del cliente y la experiencia general del cliente.

5.6. Análisis Geoespacial y Geográfico

En esta sección del análisis, emplearemos Power BI para generar un informe y vincular nuestra tabla de datos con información geográfica sobre provincias y municipios. Esto nos permitirá obtener una visión detallada de la distribución espacial de los clientes en riesgo. Identificaremos áreas geográficas específicas con una mayor concentración de clientes en riesgo para comprender mejor las tendencias regionales.

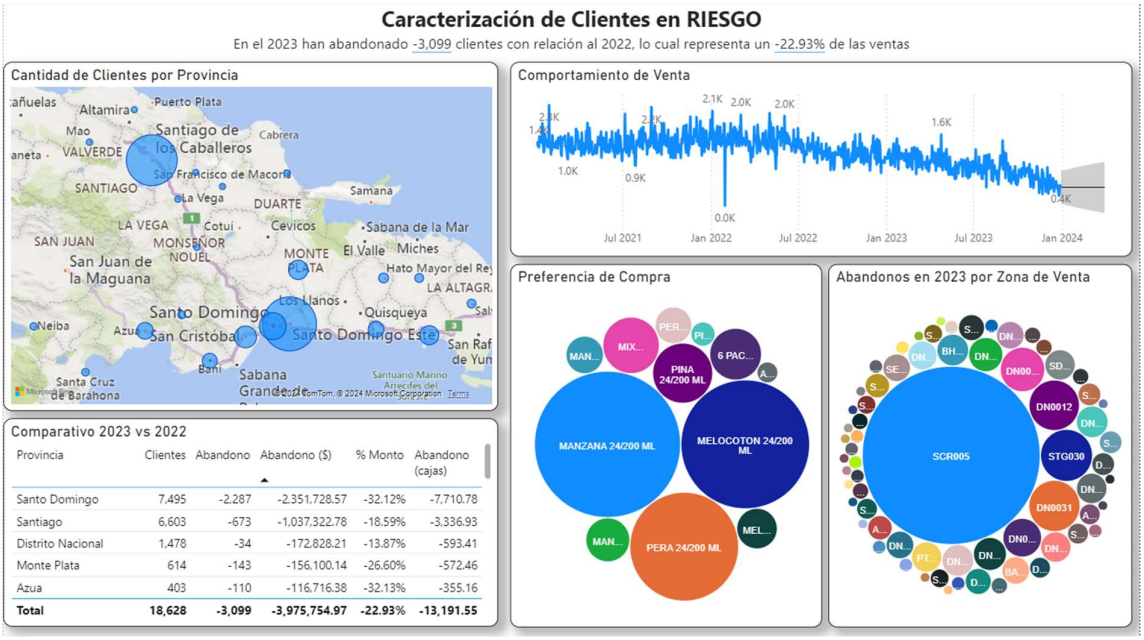
5.6.1. Cálculos y Análisis de Variación

Utilizando DAX (Data Analysis Expressions), realizaremos cálculos para determinar la variación en el monto de ventas entre 2022 y 2023 por provincia. Este análisis nos ayudará a identificar las provincias con la mayor pérdida en ingresos, así como la cantidad de clientes y cajas perdidos en cada una.

5.6.2. Hallazgos Clave

El dashboard presentado en la *figura 19* destaca a las provincias Santo Domingo, Santiago, Monte Plata, Azua y San Pedro de Macorís, las cuales sobresalen por la cantidad significativa de clientes perdidos y el mayor porcentaje de monto dejado de percibir. A nivel general, la empresa experimentó una pérdida considerable en comparación con el año anterior: dejaron de vender a 3,099 clientes, se redujeron las ventas en 13 mil cajas y perdieron 4 millones de pesos, lo que representa un 23% de pérdidas respecto a las ventas del año anterior.

Figura 19. Dashboard cliente en RIESGO



Fuente: Elaboración propia

5.6.3. Análisis de Materiales por Zona de Venta

Explorar los materiales por zona de venta ofrece una visión detallada sobre los productos que han dejado de venderse en grupos específicos de clientes y las áreas geográficas donde estos productos son populares. Esta información es crucial para desarrollar estrategias de marketing relevantes y enfocadas, que pueden ayudar a recuperar clientes perdidos y aumentar las ventas en áreas específicas.

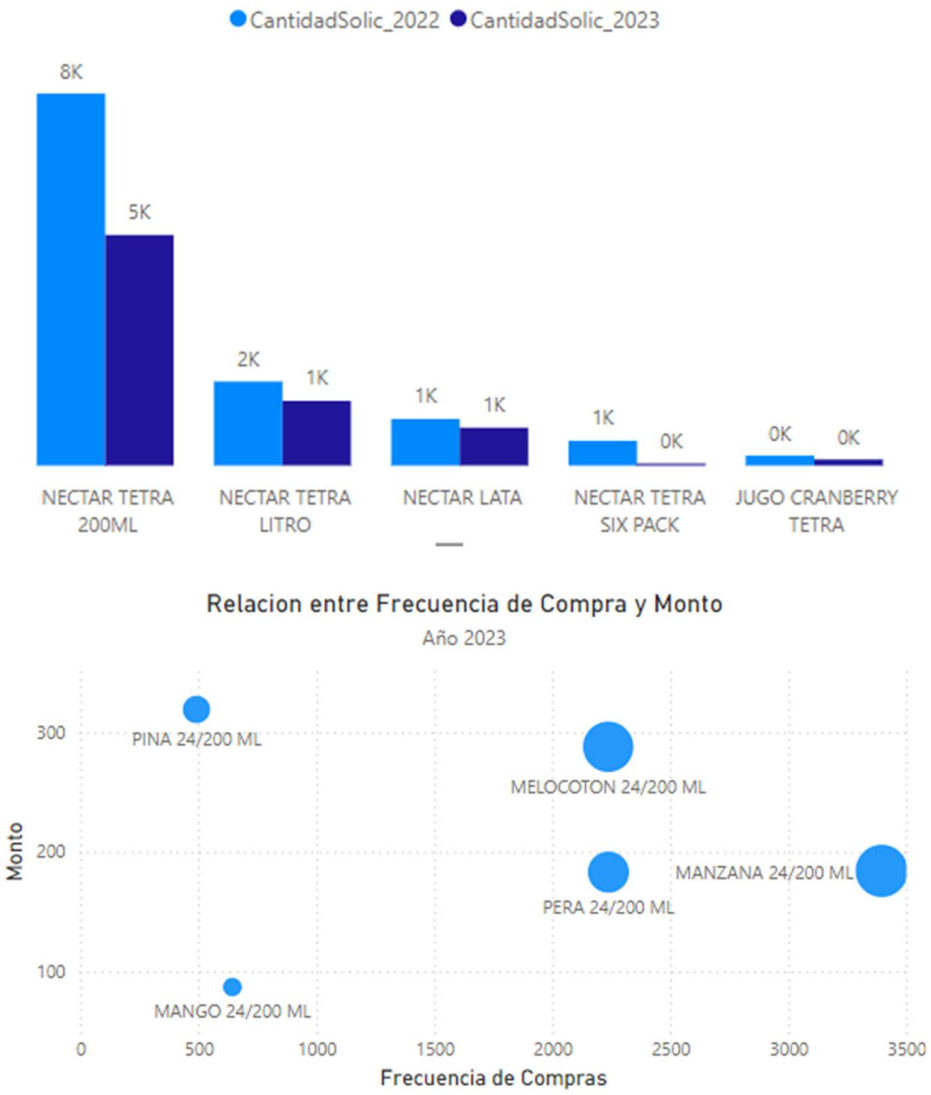
En nuestro análisis, el gráfico de burbujas apiladas muestra claramente que, en 2023, la zona de ventas SCR005 tiene el mayor número de clientes abandonados, lo cual indica que esta área debe ser prioritaria en las acciones a tomar para recuperación.

5.6.4. Análisis de Materiales por Categoría

En la *figura 20*, se observa un nivel más alto de abandono en la categoría de néctares Tetrapak de 200 ml. Esto es coherente con los datos, dado que estos productos son los más vendidos, y el mayor nivel de abandono está relacionado con esta categoría. Dentro de los néctares de 200 ml, el sabor de manzana es el más afectado, siendo el preferido por los clientes en riesgo.

Más de 3,300 clientes de este clúster compraron este producto, que tiene un precio de $184.19 situado entre el de piña ($319.11) y mango ($87.12).

Figura 20. Clientes RIESGO por Categoría de Productos



Fuente: Elaboración propia

Conocer y estudiar todas estas caracterizaciones nos permite tener una visión más clara del tipo de cliente al que debemos prestar atención. Además, nos permite cuantificar de manera tangible la magnitud que representan estos clientes dentro de la organización.

6. Modelos de Predicción

6.1. Modelo de Regresión Logística

El modelo de regresión, según Peña (2002), constituye una herramienta estadística fundamental para predecir el valor de una variable dependiente (Y) en función de una o más variables independientes (X). En este contexto, se busca establecer una relación matemática que explique o prediga la variable de interés (Y) a partir de otras variables predictoras (X).

En el interés de construir un modelo de predicción para el abandono de clientes, es imperativo contar con una variable dependiente adecuadamente etiquetada. Esta variable dependiente refleja el resultado que se desea predecir, como por ejemplo, si un cliente ha abandonado o no la relación comercial con la empresa. Sin esta variable, no es posible desarrollar ni entrenar de manera efectiva un modelo supervisado, tal como es el caso de la regresión logística.

Dado que nuestra base de datos no cuenta con esta variable etiquetada de abandono, hemos considerado varias estrategias para abordar esta limitación:

- Etiquetado Manual: Revisar los datos históricos y establecer criterios definidos para determinar cuándo un cliente se considera "abandonado". Por ejemplo, etiquetar como abandonado a aquellos clientes que no han realizado una compra en los últimos seis a doce meses.
- Segmentación de Clientes: Aplicar técnicas no supervisadas como el clustering para agrupar a los clientes en segmentos con comportamientos similares. Posteriormente, analizar los patrones de cada grupo para inferir cuáles representan clientes en riesgo de abandono.
- Análisis de Cohorte: Realizar un análisis de cohorte para estudiar el comportamiento de diferentes grupos de clientes a lo largo del tiempo. Este análisis permite identificar tendencias y patrones de abandono, lo que facilita la inferencia de una variable dependiente basada en estos comportamientos.
- Análisis de Supervivencia: Emplear técnicas de análisis de supervivencia para estimar el tiempo hasta que ocurra el evento de abandono. Este tipo de análisis es útil cuando se tienen datos censurados (clientes que aún no han abandonado en el período de estudio) y puede proporcionar información sobre la duración esperada de la relación con el cliente antes de que abandonen.

- Comportamiento de Compra: Utilizar variables derivadas del comportamiento de compra, como la frecuencia, la recencia y el monto (RFM), para inferir el riesgo de abandono. Por ejemplo, podrías etiquetar como en riesgo de abandono a los clientes que han disminuido significativamente su frecuencia de compra o que no han realizado compras recientemente.
- Modelo Supervisado con Proxies: Desarrollar modelos supervisados utilizando variables proxy que aproximen el abandono. Esto implica identificar indicadores sustitutos basados en datos de interacción del cliente, como una reducción en las interacciones o en el nivel de gasto, que puedan ser indicativos de un cliente que está considerando abandonar.

En nuestro caso particular, hemos optado por un enfoque híbrido que combina el etiquetado manual basado en el comportamiento de compra de los clientes utilizando la variable RFM. Específicamente, hemos definido el criterio de abandono basado en la variable Recency, donde un valor superior a 365 días indica que un cliente no ha realizado una compra en el último año. Esto se traduce en etiquetar como **Abandono (1)** a los clientes con este perfil, mientras que aquellos con una Recency inferior o igual a 365 días son etiquetados como **No Abandono (0)**.

Las características independientes seleccionadas para el modelo son: Frequency_Normalized, Monto_Normalized, ZV_num (número de la zona de venta) y Canal_num (número del canal de venta). Estas variables son fundamentales para predecir la probabilidad de abandono de un cliente en base a su comportamiento histórico de compra y contexto geográfico.

Construimos el modelo de regresión logística con esta secuencia de pasos:

1- Carga y preparación de datos:
   a. Definimos el criterio de abandono.
   b. Seleccionamos las características independientes y la variable objetivo.
2- División del conjunto de datos y entrenamiento del modelo:
   a. Dividimos los datos en conjuntos de entrenamiento y prueba usando train_test_split.
   b. Se entrena el modelo de regresión logística y luego se evalúa el rendimiento del modelo en el conjunto de prueba.
3- Evaluamos el modelo y visualizamos los resultados:

a. Usamos Matriz de Confusión y el Reporte de Clasificación que nos darán una idea clara de cómo está funcionando el modelo en términos de precisión, recall y f1-score para ambas clases.
b. Coeficientes del Modelo: Esto nos permitirá interpretar la importancia de cada variable en el modelo.

6.1.1. Medidas de bondad del modelado

Figura 21. Resultados Modelo Regresión Logística

```
[[430049   6422]
 [ 41174   7534]]
              precision    recall  f1-score   support

           0       0.91      0.99      0.95    436471
           1       0.54      0.15      0.24     48708

    accuracy                           0.90    485179
   macro avg       0.73      0.57      0.59    485179
weighted avg       0.88      0.90      0.88    485179

                     Coeficiente
Frequency_Normalized    -3.584199
Monto_Normalized        -1.330292
ZV_num                  -0.007532
Canal_num_x              0.009979
```

Fuente: Elaboración propia

La *figura 21* muestra las siguientes métricas claves:

- Verdaderos negativos (TN): 430049, falsos positivos (FP): 6422, falsos negativos (FN): 41174, verdaderos positivos (TP): 7534
- **No Abandono (0)**, con un nivel de precisión del 91%, un recall del 99%, F1-Score de 95% para un total de 436,471 casos de esta clase en el conjunto de prueba. El modelo es excelente para identificar correctamente los clientes que no abandonan. El recall indica que hay una cantidad significativa de falsos negativos, lo que indica que el modelo está perdiendo algunos casos de no abandono.
- **Abandono (1)**, con un nivel de precisión del 54%, un recall del 15%, F1-Score de 24% para un total de 48,708 casos de esta clase en el conjunto de prueba. Entendemos por estos valores que hay muchos falsos positivos. El modelo está etiquetando incorrectamente muchos clientes que no abandonan como si lo hicieran. El recall muestra que la mayoría de los clientes

que abandonan son incorrectamente identificados. Debido a la baja precisión, el F1-score es bastante bajo, lo que indica un desequilibrio entre precisión y recall para esta clase.

- La exactitud del modelo es del 90%, con un promedio macro de precisión del 73%, recall del 57% y F1-Score del 59%. La mayoría de las predicciones son correctas, pero esto se debe en gran parte a la alta precisión en la clase mayoritaria (no abandono). El weighted average está influenciado más por la clase **No Abandono (0)** debido a su mayor cantidad de instancias.

Los coeficientes proporcionados son:

- Frequency_Normalized: -3.584199
- Monto_Normalized: -1.330292
- ZV_num: -0.007532
- Canal_num_x: 0.009979

Estos coeficientes sugieren la relación entre cada característica y la probabilidad de pertenecer **Abandono (1)**. Un coeficiente negativo indica una relación inversa, mientras que un coeficiente positivo indica una relación directa.

- Frequency_Normalized: Un coeficiente negativo, lo que sugiere que a medida que aumenta la frecuencia normalizada, la probabilidad de pertenecer a **Abandono (1)** disminuye significativamente.
- Monto_Normalized: También negativo, indicando que un mayor monto normalizado reduce la probabilidad de pertenecer a la clase **Abandono (1)**.
- ZV_num: Negativo pero cercano a cero, sugiere una relación inversa muy débil.
- Canal_num_x: Positivo, indicando que un mayor valor de esta característica aumenta ligeramente la probabilidad de pertenecer a la clase **Abandono (1)**.

Se recomienda usar técnicas adicionales de re-muestreo, como SMOTE, o ajustar los umbrales de decisión del modelo para mejorar el recall de la clase minoritaria.

Volvimos a correr el modelo utilizando SMOTE (*ver figura 22*). La precisión para **No Abandono (0)** aumentó de 0.91 a 0.96, mientras que para **Abandono (1)** disminuyó de 0.54 a 0.26, indicando una mayor proporción de falsos positivos. El recall para la **Abandono (1)** aumentó de 0.15 a 0.69, mejorando la identificación de clientes que abandonan, aunque el recall para

No Abandono (0) disminuyó de 0.99 a 0.78. El F1-score para **Abandono (1)** mejoró de 0.24 a 0.38, pero disminuyó de 0.95 a 0.86 para **No Abandono (0)**.

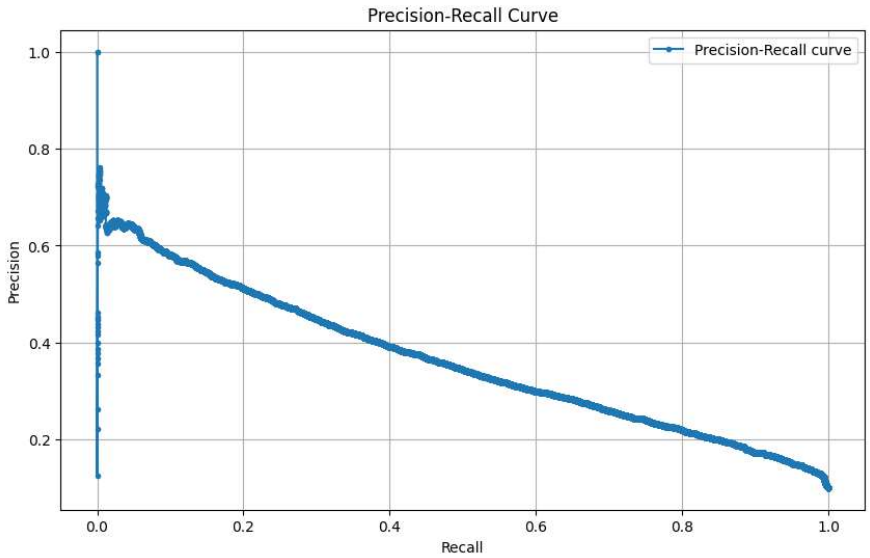
La exactitud general del modelo disminuyó de 0.90 a 0.77 con SMOTE. El uso de SMOTE ha mejorado significativamente la capacidad del modelo para identificar clientes que abandonan (recall), aunque ha reducido la precisión y la exactitud general, lo que es común al equilibrar clases desbalanceadas. La frecuencia de compra normalizada y el monto normalizado siguen siendo las características más influyentes.

Figura 22. Resultados Modelo Regresión usando SMOTE

```
[[342252  94219]
 [ 15072  33636]]
              precision    recall  f1-score   support

           0       0.96      0.78      0.86    436471
           1       0.26      0.69      0.38     48708

    accuracy                           0.77    485179
   macro avg       0.61      0.74      0.62    485179
weighted avg       0.89      0.77      0.81    485179

                     Coeficiente
Frequency_Normalized    -4.016997
Monto_Normalized        -1.545011
ZV_num                  -0.007262
Canal_num_x              0.005523
```

Fuente: Elaboración propia

Figura 23. Gráfica Precision-Recall para diferentes umbrales



Fuente: Elaboración propia

Observaciones de la Curva de Precisión-Recall (*ver figura 23*):

- Recalls Altos: Cuando el recall se aproxima a 1, la precisión tiende a ser baja. Este fenómeno indica que el modelo predice muchos resultados positivos, pero muchos de ellos son falsos positivos. Es decir, predice que los clientes abandonan cuando en realidad no lo hacen en todos los casos identificados
- Precisión Moderada: A medida que el recall disminuye, la precisión tiende a aumentar. Sin embargo, es crucial notar que la precisión no alcanza niveles muy altos antes de que el recall se reduzca significativamente. Esto sugiere un desafío constante en encontrar un equilibrio óptimo entre la precisión y el recall.

Al ajustar el umbral de decisión del modelo, notamos que incrementar este umbral mejora la precisión pero al costo de reducir ligeramente el recall. Por ejemplo, el umbral de 0.2 parece ofrecer un equilibrio mejorado generalmente, con una notable mejora en la precisión y un aumento en el F1-Score. No obstante, incluso en este punto, la precisión aún muestra valores relativamente bajos.

Dado el desafío persistente en lograr un equilibrio satisfactorio entre precisión y recall con el modelo actual y el balanceo aplicado, hemos decidido explorar la aplicación de otro modelo avanzado: Random Forest. Este enfoque podría ofrecer una mejora en la capacidad del modelo para manejar la complejidad de nuestros datos y mejorar el rendimiento general en la predicción del abandono de clientes.

6.2. Random Forest

Según Liaw y Wiener (2002), Random Forest es un método de aprendizaje automático que integra múltiples árboles de decisión para formar un modelo predictivo robusto y preciso. Este enfoque construye un conjunto de árboles de decisión, donde cada árbol se entrena con un subconjunto aleatorio de los datos y características del conjunto de entrenamiento. Posteriormente, para predecir el valor de una nueva instancia, cada árbol del bosque realiza una predicción y el resultado final se obtiene promediando las predicciones de todos los árboles.

Ventajas:

- Alta precisión: Los modelos Random Forest suelen demostrar mayor precisión en comparación con los árboles de decisión individuales, gracias a la combinación de múltiples árboles entrenados con diferentes subconjuntos de datos.
- Menor sobreajuste: A diferencia de los árboles de decisión individuales, los Random Forests son menos susceptibles al sobreajuste. Esto se debe a la naturaleza de combinación de modelos y la aleatorización durante el proceso de entrenamiento, lo cual reduce la dependencia de características específicas de los datos de entrenamiento.
- Robustez a los valores atípicos: Frente a valores atípicos en los datos Random Forest muestra robustez. Esto se debe a que el modelo considera múltiples árboles que pueden compensar efectivamente los efectos de datos anómalos en la predicción final.

Hacemos código en Python con esta secuencia de pasos:

1- Definición de la variable objetivo 'Abandono': Se crea una variable categórica basada en el criterio del tiempo desde la última compra que sea mayor a 1 año Si la condición se cumple, se asigna el valor 1, de lo contrario, se asigna el valor 0.
2- División de los datos: Se dividen los datos en características (X) y la variable objetivo (y). Las características incluyen 'Frequency_Normalized', 'Monto_Normalized', 'ZV_num' y 'Canal_num_x'.
3- División en conjuntos de entrenamiento y prueba: Se dividen los datos en conjuntos de entrenamiento (70%) y prueba (30%) utilizando la función 'train_test_split()' de scikit-learn.
4- Creación y entrenamiento del modelo Random Forest: Se crea un modelo de Random Forest utilizando la clase 'RandomForestClassifier' de scikit-learn con 100 estimadores. El modelo se entrena utilizando los datos de entrenamiento resampleados.
5- Predicciones y evaluación del modelo: Se realizan predicciones en el conjunto de prueba utilizando el modelo entrenado. Se evalúa el rendimiento del modelo utilizando una matriz de confusión y un informe de clasificación que incluye precision, recall y f1-score para cada clase.
6- Importancia de las características: Se calcula la importancia de cada característica en el modelo utilizando el atributo 'feature_importances_' del modelo Random Forest. Se

crea un DataFrame que muestra la importancia de cada característica en la predicción del modelo.

6.2.1. Medidas de bondad del modelado

Los resultados se presentan en la *figura 24*, evidenciando que el modelo Random Forest exhibe un desempeño destacado en términos de precisión y recall para ambas clases.

La Matriz de Confusión revela los siguientes resultados:

- Verdaderos Negativos (TN): 435,940 clientes no abandonados correctamente identificados.
- Falsos Positivos (FP): 531 clientes identificados como abandono, pero que en realidad no abandonaron.
- Falsos Negativos (FN): 438 clientes que abandonaron, pero que no fueron identificados como tal.
- Verdaderos Positivos (TP): 48,270 clientes abandonados correctamente identificados.

Figura 24. Resultados Random Forest

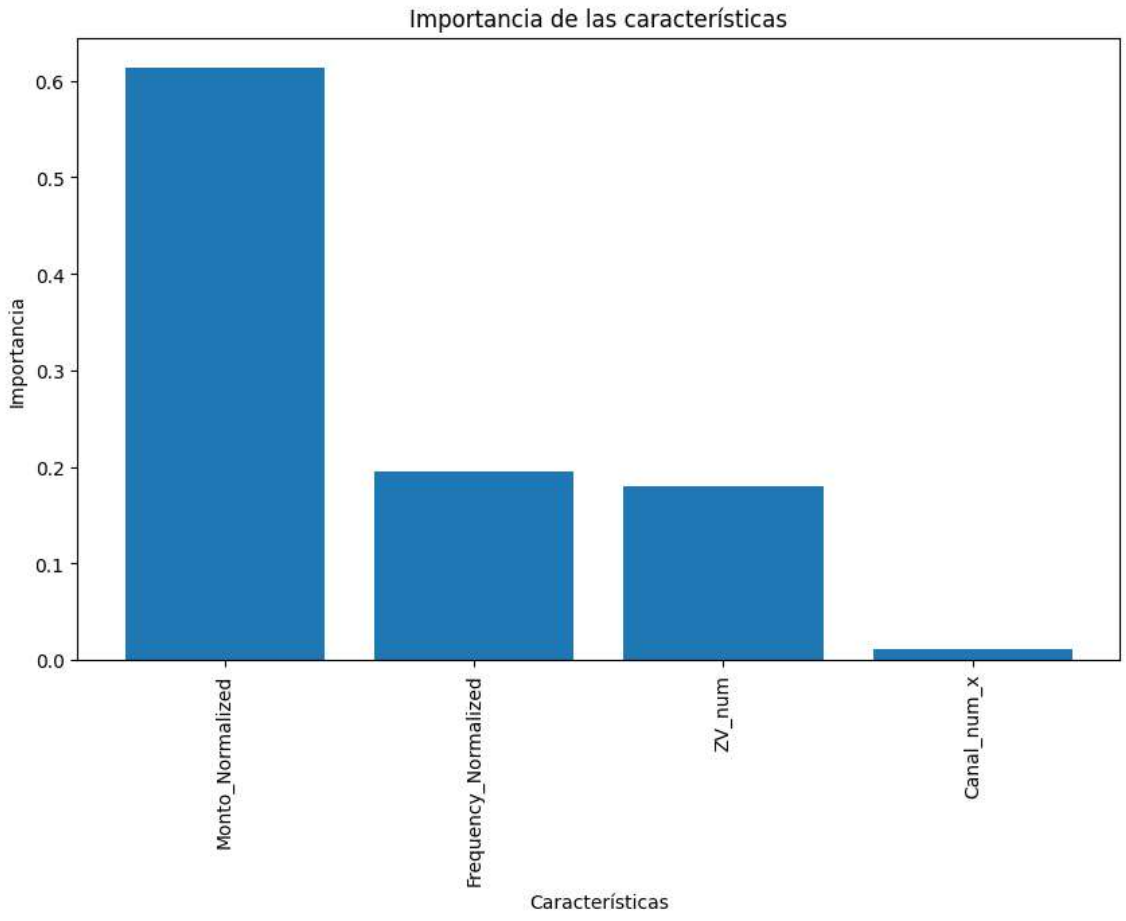
```
Confusion Matrix:
[[435940    531]
 [   438  48270]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    436471
           1       0.99      0.99      0.99     48708

    accuracy                           1.00    485179
   macro avg       0.99      0.99      0.99    485179
weighted avg       1.00      1.00      1.00    485179


Importancia de las características:
                     Importancia
Frequency_Normalized     0.195264
Monto_Normalized         0.613698
ZV_num                   0.179987
Canal_num_x              0.011051
```

Fuente: Elaboración propia

El informe de clasificación revela que el modelo Random Forest muestra un rendimiento excepcional en la predicción del abandono de clientes. Para la clase **No Abandono (0)**, la precisión es del 100%, lo que indica que todas las instancias predichas como no abandono realmente pertenecen a esa categoría y fueron clasificadas correctamente.

En cuanto a la clase **Abandono (1)**, la precisión es del 99%, lo que sugiere que casi todas las predicciones positivas son precisas. El recall también es del 99%, indicando que la mayoría de los casos positivos se identifican correctamente, aunque existe la posibilidad de que algunas instancias predichas como abandono sean falsos positivos. El F1-Score alcanza un valor muy alto del 99%, reflejando un excelente equilibrio entre precisión y recall.

Figura 25. Importancia de las características Random Forest



Fuente: Elaboración propia

La importancia de las características revela qué tanto influyen cada una de ellas en la predicción del modelo. La figura 25 facilita su interpretación, destacando que la característica más crucial es 'Monto_Normalized'. Esto sugiere que el monto de las transacciones desempeña un papel fundamental en la predicción del abandono. A continuación, con una importancia moderada, encontramos 'Frequency_Normalized' y 'ZV_num', indicando que la frecuencia de compra y la zona de venta también son factores significativos. Por otro lado, 'Canal_num_x' es la característica menos relevante según el modelo.

A pesar de la efectividad del modelo Random Forest en la predicción del abandono en este conjunto de datos, exploraremos un tercer modelo predictivo para enriquecer la investigación y asegurar una contrastación aún más robusta de los resultados.

6.3. Red Neuronal

Las redes neuronales artificiales (RNA) son modelos de aprendizaje automático inspirados en el cerebro humano, compuestos por unidades interconectadas llamadas neuronas artificiales. Estas redes pueden aprender patrones complejos a partir de datos sin necesidad de una programación explícita. Además, tienen la capacidad de generalizar lo aprendido a nuevos datos y son robustas frente al ruido en los datos, lo que las convierte en herramientas poderosas para modelar relaciones complejas entre variables.

El modelo de red neuronal entrenado en este estudio (*ver figura 26*) utilizó el algoritmo de descenso de gradiente a lo largo de 10 épocas. Cada época representa un ciclo completo a través de todos los datos de entrenamiento, ajustando los pesos de las conexiones entre las neuronas para mejorar la precisión del modelo. Durante el entrenamiento, se monitorea la precisión (accuracy) del modelo, que indica la proporción de muestras clasificadas correctamente, y la pérdida (loss), que representa el error del modelo en el conjunto de entrenamiento.

La matriz de confusión del modelo en el conjunto de prueba revela su rendimiento específico: 433,304 verdaderos negativos (TN), 5,595 verdaderos positivos (TP), 3,167 falsos positivos (FP) y 43,113 falsos negativos (FN). Este análisis proporciona una evaluación detallada de cómo el modelo de red neuronal clasifica los casos de abandono de clientes, destacando su capacidad para identificar tanto los verdaderos positivos como los negativos.

Figura 26. Resultados Red Neuronal

```
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 26s 707us/step - accuracy: 0.9018 - loss: 0.2581
Epoch 2/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 28s 776us/step - accuracy: 0.9031 - loss: 0.2493
Epoch 3/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 36s 1ms/step - accuracy: 0.9031 - loss: 0.2476
Epoch 4/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 39s 1ms/step - accuracy: 0.9036 - loss: 0.2456
Epoch 5/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 45s 1ms/step - accuracy: 0.9037 - loss: 0.2452
Epoch 6/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 49s 1ms/step - accuracy: 0.9040 - loss: 0.2446
Epoch 7/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 51s 1ms/step - accuracy: 0.9040 - loss: 0.2442
Epoch 8/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 51s 1ms/step - accuracy: 0.9042 - loss: 0.2428
Epoch 9/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 54s 2ms/step - accuracy: 0.9043 - loss: 0.2424
Epoch 10/10
35378/35378 ━━━━━━━━━━━━━━━━━━━━ 47s 1ms/step - accuracy: 0.9043 - loss: 0.2417
15162/15162 ━━━━━━━━━━━━━━━━━━━━ 14s 921us/step
Confusion Matrix:
[[433304   3167]
 [ 43113   5595]]

Classification Report:
...
              accuracy                           0.90    485179
             macro avg       0.77      0.55      0.57    485179
          weighted avg       0.88      0.90      0.87    485179
```

Fuente: Elaboración propia

6.3.1. Medidas de bondad del modelado

El modelo de red neuronal exhibe un rendimiento generalmente alto en términos de precisión y pérdida durante el entrenamiento, alcanzando una precisión final de aproximadamente el 90%. Sin embargo, al analizar la matriz de confusión y las métricas de clasificación, se observa que el modelo enfrenta desafíos significativos al predecir correctamente la clase minoritaria (**Abandono**). Esto se refleja en el alto número de falsos negativos y en un recall más bajo para esta clase.

Similar a la regresión logística, aunque el modelo de red neuronal muestra eficacia para la clase mayoritaria, demuestra ser menos efectivo para la clase minoritaria. Este hallazgo resalta la necesidad de considerar estrategias adicionales, como el ajuste de umbrales de

decisión o el uso de técnicas de balanceo de clases, para mejorar la capacidad del modelo para identificar correctamente los casos de abandono de clientes.

6.4. Aplicación del modelo Random Forest a datos nuevos

Una vez que los datos estén preparados, podemos utilizar el método *predict()* del modelo de Random Forest para hacer predicciones. Esto nos dará una predicción binaria para cada cliente: 0 si se predice que no abandonará y 1 si se predice que abandonará.

```
# Preparar los datos del cliente objetivo en un DataFrame
datos_cliente = pd.DataFrame({
    'Frequency_Normalized': [0.62],
    'Monto_Normalized': [0.21],
    'Canal_num_x': [0],
    'ZV_num': [1]})

# Verificar que las columnas sean las mismas que el modelo espera
expected_columns = rf_model.feature_names_in_  # feature_names_in_ es un atributo de los
modelos de sklearn
for col in expected_columns:
    if col not in datos_cliente.columns:
        datos_cliente[col] = 0

# Asegurar el mismo orden de columnas
datos_cliente = datos_cliente[expected_columns]

# Hacer la predicción utilizando el modelo de Random Forest
prediccion_cliente = rf_model.predict(datos_cliente)

# Imprimir la predicción
if prediccion_cliente[0] == 0:
    print("El cliente NO va a abandonar.")
else:
    print("El cliente SÍ va a abandonar.")
```

Resultado:

```
El cliente NO va a abandonar.
```

Si escogemos otro cliente con otros parámetros, obtenemos lo siguiente:

```
# Preparar los datos del cliente objetivo en un DataFrame
datos_cliente = pd.DataFrame({
    'Frequency_Normalized': [0.22],
    'Monto_Normalized': [0.81],
    'Canal_num_x': [6],
    'ZV_num': [12]})
```

El resultado es:

```
El cliente SÍ va a abandonar.
```

Para poner a prueba el modelo con casos reales y de los cuales conocemos el resultado final, utilizaremos los datos RFM de los clientes XXX, YYY y ZZZ según sus cualificaciones específicas. Esta vez nos interesa saber la predicción con un cliente identificado dentro de las Zonas de Venta de mayor riesgo vs un cliente en otra zona (ver capítulo 5.6).

Cliente XXX:

- 'Frequency_Normalized': [0.902022614],
- 'Monto_Normalized': [0.893451857],
- 'Canal_num_x': [1],
- 'ZV_num': [1]

Resultado:

```
El cliente NO va a abandonar.
```

Cliente YYY:

- 'Frequency_Normalized': [0.838498713],
- 'Monto_Normalized': [0.995352931],
- 'Canal_num_x': [8],
- 'ZV_num': [70]

Resultado:

```
El cliente SÍ va a abandonar.
```

Cliente ZZZ:

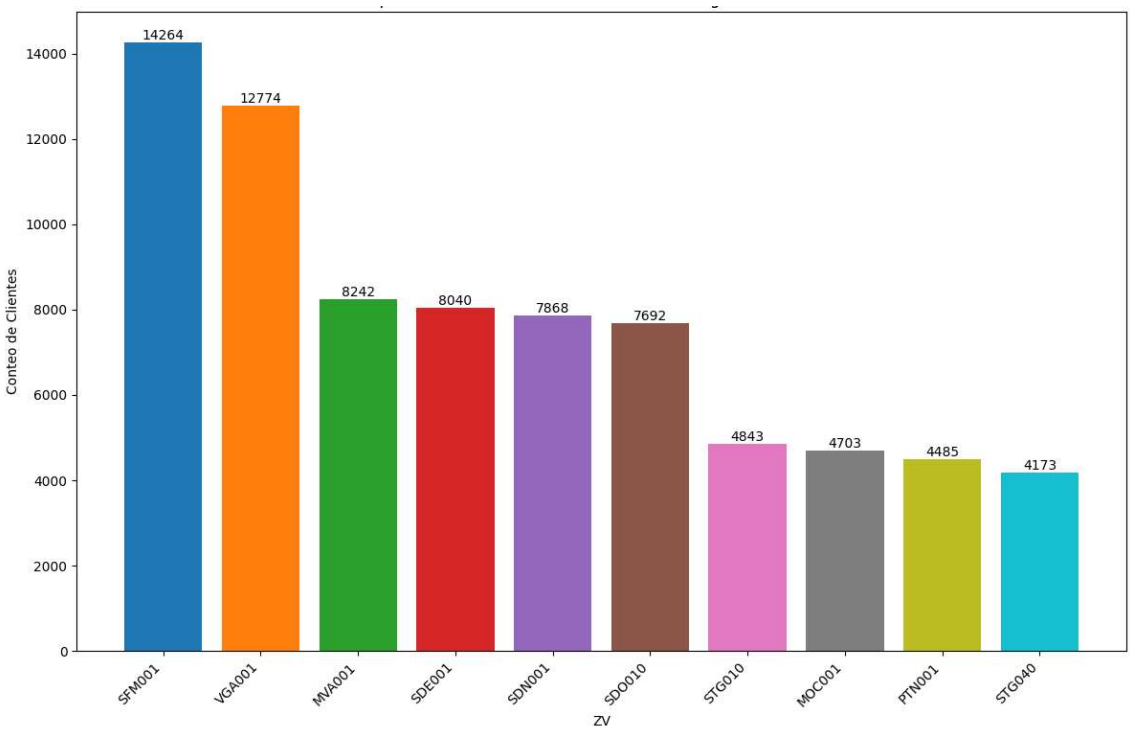
- 'Frequency_Normalized': [0.692558869],
- 'Monto_Normalized': [0.879140996],
- 'Canal_num_x': [12],
- 'ZV_num': [18]

Resultado:

```
El cliente SÍ va a abandonar.
```

Utilizando el mismo criterio de abandono, definido como clientes que no han realizado compras en más de 1 año, hemos identificado las 10 zonas de ventas con la mayor concentración de clientes en riesgo. Estas áreas representan la primera sugerencia para que el equipo comercial y de marketing concentre su atención:

Figura 27. Top 10 ZV no han comprado último año



Fuente: Elaboración propia

Nota: Refiérase al Anexo A para el listado completo de los valores de 'Canal_num_x' y 'ZV_num'.

7. Conclusiones

El desarrollo de estrategias efectivas para prevenir el abandono de clientes y fomentar la fidelización es crucial para el éxito a largo plazo de la empresa de comercialización de productos de consumo masivo. Al centrarnos en la satisfacción del cliente, la segmentación y personalización, y el uso de métricas como RFM y CLV, podemos conocer y fortalecer nuestras relaciones con los clientes, y garantizar su lealtad a lo largo del tiempo. Además, la implementación de herramientas de análisis avanzadas como R y Python, junto con Power BI para informes inteligentes, son elementos clave en este proceso, permitiendo una toma de decisiones ágil y basada en datos.

El modelo predictivo desarrollado en este TFM ofrece una herramienta poderosa para la empresa en cuestión. Al identificar de manera precisa a los clientes en riesgo de abandono y ofrecer recomendaciones prácticas para su retención, podemos mejorar significativamente la lealtad de los clientes y el rendimiento general. Este proyecto no solo cumple con los objetivos establecidos, sino que también proporciona una base sólida para futuras iniciativas de análisis de datos y mejora continua en la gestión de la relación con los clientes.

7.1. Apoyo en la Literatura

Este estudio es coherente con el trabajo de Verhoef, P. C. (2003) sobre la satisfacción y fidelización del cliente, destacando la importancia de comprender y gestionar las relaciones con los clientes para maximizar su valor a largo plazo.

7.2. Salida o Solución Ofrecida

En el marco del Trabajo de Fin de Máster, se ha desarrollado un modelo predictivo avanzado para anticipar el abandono de clientes. Esta solución nos permite identificar a los clientes con mayor riesgo de abandono, permitiéndonos tomar medidas proactivas para retener a estos clientes y mejorar la lealtad a largo plazo.

7.2.1. Objetivos Prometidos y Cumplidos

1- Desarrollar un modelo predictivo preciso y aplicable: Hemos completado con éxito el desarrollo del modelo predictivo utilizando técnicas avanzadas de análisis de datos, incluyendo Random Forest, regresión logística, red neuronal y otros algoritmos de machine learning. El modelo ha demostrado una alta precisión en la predicción del

abandono de clientes, cumpliendo así con nuestro objetivo general. Se valida su eficacia y aplicabilidad en un entorno empresarial real.

2- Identificar los factores determinantes del abandono de clientes: Nuestro análisis exhaustivo de los datos nos ha permitido identificar varios factores clave que influyen en el abandono de clientes. Entre estos factores, los más determinantes fueron la frecuencia de compra, el monto de las compras, la zona de ventas y el canal de compra utilizado por los clientes. Estos hallazgos permiten a la empresa entender mejor los comportamientos que llevan al abandono y ajustar sus estrategias en consecuencia.

3- Recopilar y analizar datos pertinentes sobre el comportamiento de compra de los clientes: Hemos recopilado y analizado datos detallados sobre las transacciones de los clientes, sus patrones de compra y sus interacciones con la empresa, permitiéndonos segmentar a los clientes según su riesgo de abandono en Top, Prometedores y en Riesgo.

4- Evaluar la eficacia del modelo de predicción e implementar recomendaciones prácticas: La eficacia del modelo fue evaluada continuamente a lo largo del proyecto, y se realizaron ajustes para mejorar su precisión y aplicabilidad. Con base en los resultados del modelo, se propusieron varias recomendaciones prácticas para la empresa, incluyendo estrategias de retención de clientes, programas de lealtad personalizados y mejoras en la experiencia del cliente. Estas recomendaciones tienen el potencial de reducir significativamente la tasa de abandono y fomentar la retención a largo plazo.

7.3. Estrategias para Prevenir el Abandono de Clientes y Fidelización

7.3.1. Segmentación y personalización

La segmentación de clientes en las categorías de Top, Prometedores y Riesgo permite una estrategia más precisa y dirigida para cada grupo. Para los clientes en Riesgo, se recomienda una atención especial para identificar y abordar sus preocupaciones y necesidades. Por otro lado, los clientes Prometedores pueden beneficiarse de estrategias para estimular sus compras, como descuentos exclusivos o programas de recompensas. Además, el uso de RFM y CLV nos permite identificar patrones de comportamiento de los clientes y orientar nuestras estrategias de marketing y retención.

7.3.2. Enfoque en la satisfacción del cliente

Es muy importante centrarse en la satisfacción del cliente y ofrecer una experiencia excepcional en todos los puntos de contacto con la empresa. Se deben implementar medidas para garantizar una experiencia positiva en todos los puntos de contacto con la empresa, desde la compra hasta el servicio postventa. La implementación de herramientas de análisis avanzadas como R y Python, para el manejo y análisis de datos, y la utilización de Power BI para informes inteligentes nos proporcionan una visión en tiempo real de las métricas clave y el rendimiento del modelo, facilitando así la toma de decisiones informadas y la identificación de áreas de mejora en la estrategia de retención de clientes.

7.4. Plan de Marketing para Clientes Top y Prometedores

Personalizar las comunicaciones y promociones dirigidas a cada segmento de clientes es fundamental. Los clientes Top pueden recibir ofertas exclusivas y programas de fidelización de alto nivel, mientras que los clientes Prometedores pueden ser incentivados con promociones especiales para aumentar su compromiso y lealtad.

7.4.1. Programas de lealtad

Estas son algunas de las acciones que podemos llevar a cabo para fomentar la retención y la fidelización de nuestros clientes:

- Diseñar programas de recompensas exclusivos y atractivos a los clientes Top que se adapten a sus necesidades y preferencias individuales, utilizando herramientas como **Annex Cloud**[10] o **Salesforce Loyalty Management**[11].
- Ofrecer servicios adicionales y asesoramiento personalizado.
- Lanzar promociones especiales y campañas segmentadas con **HubSpot CRM**[12].
- Desarrollar programas de lealtad personalizados basados en el comportamiento del cliente utilizando herramientas como **Microsoft Dynamics**[13]
- Implementar envíos prioritarios y estrategias de comunicación enfocadas.

[10] https://www.annexcloud.com/
[11] https://www.salesforce.com/ca/products/loyalty-management/overview/
[12] https://www.hubspot.es/
[13] https://www.microsoft.com/es-es/dynamics-365

Estas acciones pueden incluir descuentos por tiempo limitado, ofertas de compra cruzada o bonificaciones por compras recurrentes, utilizando las herramientas mencionadas para maximizar la efectividad de cada campaña.

7.4.2. Monitoreo y Evaluación Continua

Es fundamental realizar un seguimiento constante del desempeño del plan de marketing, evaluando su efectividad y realizando ajustes según sea necesario. Utilizar métricas clave como la tasa de conversión, la retención de clientes y el valor del ciclo de vida del cliente nos permite medir el impacto del plan y optimizar su rendimiento a lo largo del tiempo. Herramientas como **Qualtrics**[14] pueden ser muy útiles para este propósito.

7.4.3. Estrategias de Upselling y Cross-Selling:

Implementar estrategias de upselling y cross-selling para maximizar el valor de cada cliente es fundamental. Como mencionamos en el capítulo 3.6 el upselling implica ofrecer productos o servicios de mayor calidad o con características adicionales a los clientes prometedores mientras que el cross-selling se refiere a la venta de productos o servicios complementarios. Estas estrategias no solo aumentan los ingresos, sino que también mejoran la satisfacción del cliente al ofrecer soluciones más completas y adaptadas a sus necesidades.

Ambas estrategias requieren una comunicación efectiva y constante con los minoristas e intermediarios, además de construir relaciones sólidas y de confianza con ellos.

7.5. CRM

Tras la validación del modelo, se recomienda a la empresa la implementación de un sistema de gestión de la relación con los clientes (CRM). Este sistema, como menciona Esteban J. A. Leon et al. (2018), recopila y analiza datos del comportamiento del cliente, facilitando el desarrollo y la aplicación de modelos predictivos.

Motivamos a esta empresa utilizar herramientas tecnológicas como **Salesforce** o **HubSpot CRM** para gestionar estas relaciones, automatizar campañas y analizar los resultados.

[14] https://www.qualtrics.com/

La integración del modelo predictivo con las mencionadas opciones de plataforma CRM permite identificar de manera proactiva a los clientes en riesgo de abandono y tomar medidas preventivas para retenerlos, optimizando así la gestión de la relación con los clientes.

Un CRM puede ser una herramienta invaluable para una empresa de productos de consumo masivo, ofreciendo funcionalidades que optimizan la interacción con los clientes, mejoran la eficiencia operativa y potencian las estrategias de retención y fidelización.

En resumen, un buen software de CRM habilita a la empresa para:

- Centralizar los datos y segmentar clientes: Reunir toda la información relevante sobre los clientes en un solo lugar y que agrupe a los clientes según diversos criterios para personalizar las estrategias de marketing.
- Automatizar campañas de marketing: Enviar mensajes personalizados y relevantes a diferentes segmentos de clientes.
- Gestionar consultas y quejas: Facilitar la gestión de consultas, quejas y solicitudes de servicio, asegurando una respuesta rápida y eficiente.
- Analizar informes y métricas: Evaluar el rendimiento de las estrategias de marketing y ventas, la satisfacción del cliente y otras métricas clave. El acceso a análisis detallados y en tiempo real lleva a la empresa tomar decisiones estratégicas basadas en datos, mejorando la precisión y efectividad de sus estrategias de negocio.

Estas conclusiones reflejan el compromiso con la excelencia en la gestión de la relación con los clientes y proporcionan una guía clara para continuar mejorando las estrategias de retención y fidelización.

8. Limitaciones y prospectiva

8.1. Mejora de la Recopilación de Datos:

Aunque el modelo predictivo ha mostrado una precisión aceptable, la disponibilidad de datos adicionales podría mejorar su desempeño. Actualmente, no contamos con datos de quejas y feedback de los clientes, histórico de los planes de acción para evitar abandonos, interacciones en redes sociales, encuestas de satisfacción y otros indicadores similares. Integrar estos datos permitirá una comprensión más profunda del contexto y posibles motivos de abandono de los clientes. Es esencial que la empresa considere la recopilación de información adicional al solo pedido de compra de cada cliente.

Otro aspecto es la falta de variables categóricas específicas en los datos disponibles ha requerido asumir criterios basados en indicadores como la frecuencia de compra y la recencia. A pesar de los esfuerzos por crear modelos robustos, la inclusión de variables adicionales podría proporcionar una imagen más completa y precisa de los factores que influyen en el abandono de clientes.

Se recomienda explorar técnicas avanzadas de análisis de datos para identificar patrones sutiles y correlaciones ocultas, incluyendo técnicas de aprendizaje no supervisado para descubrir segmentos de clientes con comportamientos similares y análisis de texto para extraer información de interacciones en redes sociales y otras fuentes no estructuradas.

Optimizar la captura de datos y expandir las variables consideradas en el análisis no solo mejorará la precisión de los modelos predictivos, sino que también proporcionará a la empresa una ventaja competitiva al comprender mejor las necesidades y comportamientos cambiantes de sus clientes.

8.2. Limitaciones como Intermediarios en la Cadena de Suministro:

Es adecuado reconocer que, como empresa distribuidora de productos, la relación con los clientes es indirecta, ya que no se interactúa directamente con el consumidor final, sino que la entrega es a través de intermediarios. Esta dinámica expone a diversos factores externos que pueden influir en la retención de clientes por parte de los clientes minoristas.

Por lo tanto, se hace imperativo mantener una vigilancia constante sobre estos factores y trabajar en estrecha colaboración con los clientes minoristas para mitigar cualquier efecto

negativo, identificar y resolver cualquier problema que pueda surgir, y garantizar así una experiencia positiva para el consumidor final.

Además, es recomendable explorar estrategias integradas con los intermediarios para fortalecer la cadena de suministro, como la optimización de inventarios, la mejora de la logística y la implementación de sistemas de recolección de información eficientes que permitan una gestión más ágil y efectiva de los pedidos, las entregas, y la evaluación del servicio.

9. Referencias bibliográficas

Breiman, L. (2001). Random Forests. (R. E. Schapire, Ed.) *Machine Learning, 45*(1), 5-32.

Esteban J. A. León, M. C.-G.-C. (2018). CRM como herramienta para la identificación de clientes en riesgo de abandono. *Revista de Administração Contemporânea, 22*(4), 43-64.

Fader, P. S. (2016). *Customer lifetime value: Marketing for the long term.* Pearson Education Limited.

Farris, P. W. (2010). *Marketing Metrics* (5th ed. ed.). Prentice Hall.

Han, J. K. (2001). *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.

Jure Leskovec, A. R. (2014). *Mining of Massive Datasets.* Cambridge University Press.

Kotler, P. &. (2012). *Marketing management (14th ed.).* Pearson Education.

Liaw, A. &. (2002). Classification and regression by random forests. *R news, 3*(2), 18-22.

Peña, J. F. (2002). *Análisis de regresión.* Editorial Paraninfo.

Reichheld, F. F. (1996). *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value.* Harvard Business School Press.

Verhoef, P. C. (2003). *Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development* (Vol. 67). Journal of Marketing.

Winston, W. L. (2019). *Marketing analytics: Using data to make better marketing decisions* (8th ed. ed.). Cengage Learning.

Zhang, Y. &. (2006). *Customer Lifetime Value Forecasting with Neural Network Approach* (Vol. 3). International Conference on Machine Learning and Cybernetics.

ANEXO A. Leyenda de valores asignados para las variables no numéricas:

Canales:

Canal	Canal_num
Supermercados Indep.	0
Colmados	1
Mayoristas	2
Cafeterias	3
Surtidoras	4
Punto de Ventas	5
Empleados	6
Personas Relacionada	7
Cadena Supermercados	8
Clubes	9
Clinicas	10
Almacenistas	11
Farmacias	12
Mini Markets	13
Restaurantes	14
Comedores	15
Hoteles	16
Industrias	17
Food Shop	18
Panaderias y Reposte	19
Colegio / Escuelas	20
Delicatesen	21
Gobierno	22
Bares	23
Catering	24

Zonas de Ventas:

ZV	ZV_num
DE0014	0
SCR005	1
SCR010	2
SDO010	3
SDE001	4
DE0011	5
DE0004	6
SDO005	7
SCR001	8
SDN001	9
STG040	10
STG025	11
SDO001	12
STG001	13
STG015	14
STG005	15
MOC001	16
VGA001	17
MVA001	18
STG035	19
DNA015	20
BAN001	21
SFM001	22
STG010	23
SDE010	24
STG020	25
PTN001	26
COL001	27
SDE025	28
SPM003	29
SPM004	30
BAN005	31
DNA005	32
SCR015	33
DNA025	34
SDO025	35
SO001	36
SDE050	37
SDE020	38
SP001	39
DE0003	40
STG030	41
AZ001	42
SDN005	43
SDE035	44
HM001	45
DNA020	46
SDE015	47
PTS005	48
SPM005	49
DNA001	50
SDN015	51
SE001	52
SDO015	53
MP005	54
MP001	55
SDE030	56
SDE012	57
SDO029	58
SDE005	59
SPM009	60
SPM002	61
BLANK()	62
DNA010	63
SDN003	64
SDE045	65
EMPLMA	66
DE0005	67
DE0001	68
DE0008	69
DE0009	70
DE0007	71
DE0012	72
DE0013	73
DE0015	74
DE0002	75
DE0010	76
DN0017	77
SDN009	78
SDN010	79
MP010	80
SDN020	81
SDO012	82
SDN019	83
SDN018	84
SDO020	85
SDN004	86
SDN007	87
DN0014	88
SDN023	89
SDN008	90

SDO004	91
DN0016	92
DN0026	93
SDN022	94
DN0032	95
AZU001	96
SDO019	97
DN0013	98
DN0019	99
HTM001	100
STG045	101
DN0001	102
SDE054	103
DN0011	104
SDN013	105
SDN014	106
SDN006	107
SDN012	108
SDE027	109
SDE022	110
DN0031	111
DN0015	112
DN0024	113
DN0005	114
SDO028	115
SDN011	116
SDE040	117
SDN017	118
SDN016	119
SDO017	120
SDN002	121
MPE004	122
SDO023	123

SDE007	124
SDO027	125
SDO013	126
AZU010	127
DN0006	128
SDE044	129
SDO016	130
DN0021	131
DN0012	132
DN0018	133
SDE052	134
COL010	135
SDE032	136
DN0010	137
AZU008	138
SDO009	139
SDN021	140
DN0028	141
MPE003	142
SPM015	143
SDE038	144
SDO022	145
SPM006	146
SDO026	147
SDE053	148
SDE051	149
SDE046	150
SDE033	151
SDE034	152
SDO018	153
SDO024	154
SDO006	155
SDO014	156

SDO002	157
SDO030	158
SDO008	159
SDO011	160
SDO007	161
SDO021	162
DN0023	163
SDO003	164
DN0002	165
SDE011	166
SDE028	167
MP015	168
AZU004	169
MPE008	170
DN0022	171
DN0027	172
SDE017	173
DN0030	174
SDE002	175
SDE014	176
SDE043	177
DN0029	178
DN0020	179
DN0003	180
HTM003	181
COL005	182
DN0025	183
AZU003	184
MPE005	185
SDE013	186
SDE003	187
BHR002	188
SDE026	189

SDE021	190
SDE029	191
DN0009	192
SDE009	193
SDE016	194
DN0007	195
SDE004	196
BAH004	197
DN0008	198
SDE041	199
SDE036	200
DN0004	201
SPM001	202
SDE049	203
SDE008	204
SDE006	205
SDE042	206
MPE002	207
DE0020	208
MPE001	209
AZU005	210
SDE031	211
SDE037	212
SDE047	213
SDE019	214
SDE024	215
SDE039	216
SDE018	217
SDE023	218
AZU002	219
AZU007	220
SDE048	221
SDE058	222