Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Modelo de análisis predictivo para la identificación de clientes con tendencia a la deserción.

Presentado por:

Víctor Santiago Patarroyo Velasco

Bogotá, D.C. 18 de julio de 2023

Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Modelo de análisis predictivo para la identificación de clientes con tendencia a la deserción.

Presentado por:

Víctor Santiago Patarroyo Velasco

Bajo la dirección de:

Daniel Leonardo Cruz Castro

Bogotá, D.C., Colombia, 18 de julio de 2023

Tabla de Contenido

Agradecimientos

Agradezco a mi director de proyecto Daniel Cruz, quien me guío en la construcción de este proyecto empresarial, el cual fue un desafío en todo momento. Agradezco a la Universidad del Rosario, por aportar a través de sus docentes los conocimientos necesarios para culminar este proceso en mi vida académica y agradezco a Dios por la vida, y todas las bendiciones que llegan a mi vida.

Santiago Patarroyo Velasco.

Dedicatoria

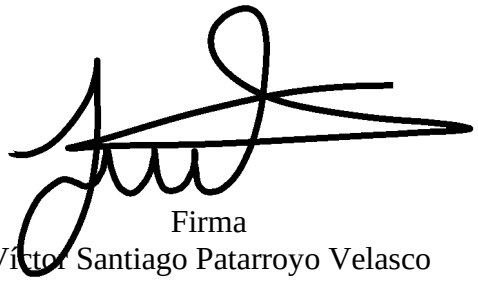
Dedico este trabajo a mi hijo, mi esposa y mi madre y sobre todo a Dios por darme la sabiduría, la templanza y la capacidad de poder lograr un reto más y por no dejarme desfallecer a lo largo de esta maestría.

Santiago Patarroyo Velasco.

Declaración de originalidad y autonomía.

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi(nuestra) propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

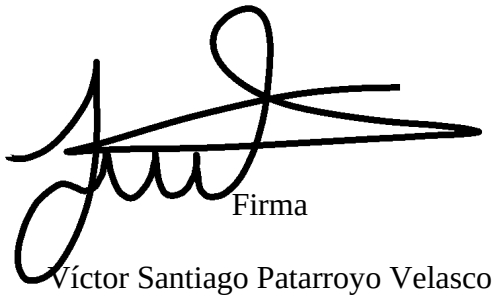


Firma
Víctor Santiago Patarroyo Velasco

Firmado en Bogotá, D.C. el 18 de julio de 2023

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Firma

Víctor Santiago Patarroyo Velasco

Firmado en Bogotá, D.C. el 18 de julio de 2023

Lista de figuras

Lista de tablas

Glosario

Modelo RFM: “El modelo Recency, Frequency, & Monetary (RFM) es una herramienta clásica de análisis y segmentación para identificar a sus mejores clientes”(Sharma, 2019).

Crisp DM: “Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos”. (IBM, 2021)

Resumen Ejecutivo

Modelo de análisis predictivo para la identificación de clientes con tendencia a la fuga.

La industria colombiana requiere de una alta oferta de envases y empaques para cubrir las necesidades del sector industrial, por esta razón, en el país existen numerosas empresas que fabrican o comercializan este tipo de productos, lo que genera un mercado altamente dinámico y competitivo. Sin embargo, debido a esta situación, resulta difícil asegurar la fidelidad y continuidad en las compras por parte de los clientes. Lo que lleva a la empresa a perder al año en promedio un 19% de los clientes anualmente. A pesar de esta perdida, también se observa una constante afluencia de nuevos clientes que, en cierta medida, compensan a aquellos que han dejado de comprar.

Teniendo en cuenta este contexto, se pretende utilizar la base de datos suministrada por la compañía para llevar acabo un análisis que permita identificar el comportamiento de compra, segmentar a los clientes potenciales y desarrollar un modelo que sirva como herramienta para que el equipo comercial pueda detectar de identificar de manera oportuna posibles fugas de clientes. Con estas herramientas la compañía podrá generar estrategias de fidelización y retención de clientes con el fin de crear relaciones comerciales de largo plazo. En conclusión, lo que se busca es evitar la fuga de clientes, perdida de negocios potenciales y tener que incrementar los costos de ventas para realizar el proceso de recuperación y reactivación.

Palabras clave

Industria colombiana, abastecimiento, envases, empaques, mercado dinámico, fidelización de clientes, retención de clientes, pasos analíticos, comportamiento de compra, clientes potenciales, modelo, equipo comercial, desgaste de clientes, relaciones comerciales, churn de clientes, pérdida de ingresos, costos de ventas, recuperación, proceso de reactivación.

Abstract

Predictive analytics model for the identification of customers with a tendency to churn.

The Colombian industry requires a substantial supply of containers and packaging to meet the needs of the industrial sector. Consequently, there is a significant number of companies in the country involved in the manufacturing or marketing of such products, contributing to a highly dynamic and competitive market. Maintaining customer loyalty and ensuring consistent purchasing patterns can be challenging as a result. On average, the company loses 19% of its customers annually. However, there is a constant influx of new customers that partially offset the ones who have discontinued their purchases.

Considering this context, the intention is to utilize the database provided by the company to conduct analytical steps in order to identify purchasing behavior, segment potential customers, and develop a model that serves as a tool for the commercial team to promptly identify possible customer attrition. Subsequently, the goal is to formulate customer loyalty and retention strategies, enabling the establishment of long-term business relationships. Ultimately, the aim is to prevent customer churn, potentiDEserciónal revenue loss, and the need for increased sales costs associated with recovery and reactivation processes.

Keywords: Colombian industry, supply, containers, packaging, dynamic market, customer loyalty, customer retention, analytical steps, purchasing behavior, potential customers,

model, commercial team, customer attrition, business relationships, customer churn, revenue loss, sales costs, recovery, reactivation process.

1. Introducción

La industria en Colombia ha venido creciendo de manera gradual gracias a la creación de nuevos formatos comerciales y de distribución de productos de consumo masivo de bajo costo, esto ha permitido que pequeñas y medianas empresas del sector de alimentos, aseo, cosméticos y otros productos hayan tenido la oportunidad de crecer, dar a conocer sus productos e incrementar sus ventas, generando así una mayor dinámica y demanda por los envases y empaques fabricados en plástico y vidrio. En el país existe una amplia oferta de este tipo de productos tanto de fabricación nacional como importados, dando la posibilidad a los empresarios de escoger sus proveedores por precio, calidad y cumplimiento como principales factores.

Es por ello, que la competencia en este sector tiene una fuerte presencia y siempre busca retener y atraer a los clientes con algunos de los factores anteriormente mencionados, en especial con el precio. Por estas razones no existe fidelidad ni continuidad de compra de los clientes y los lleva a que estén adquiriendo sus envases con un proveedor u otro con el fin de encontrar un beneficio y/o reducción en su estructura de costos para mantener los márgenes de utilidad que necesitan y así poder vender sus productos o mantenerse en los establecimientos de bajo costo donde los comercializan.

El país cuenta con un alto número de pequeñas y medianas empresas en los sectores mencionados anteriormente. Además, se suman a estas cifras todas aquellas personas naturales que tienen negocios menos formarles pero que contribuyen a la gran demanda de los envases, representando aproximadamente un 40% de los clientes de la compañía. Este tipo de cliente

tienden a ser más propensos a abandonar o buscar otros proveedores debido a problemas con el servicio y la calidad. Es posible pensar que el factor precio no es una variable tan relevante para ellos, ya que su infraestructura no es tan sólida y no requieren de tantos costos para operar, lo que les permite vender sus productos a precios competitivos.

Otra razón por la cual los clientes muestran poca fidelidad es la falta en variedad de diseños o modelos de envases y empaques en Colombia, en consecuencia, los clientes tienen la posibilidad de comprar a cualquier proveedor, ya que la mayoría ofrece productos similares en su catálogo. Especialmente en el caso de los envases de vidrio. En cuanto a los envases de plástico hay más opciones de diseños disponibles siguen siendo muy similares entre sí. Por lo tanto, para los clientes informales, cambiar de envase no representa un inconveniente significativo, sin embargo, para las Pymes, esto puede afectarles, por lo que tienden a buscar dos o tres proveedores para abastecerse de uno u otro en caso de que uno de ellos no tenga inventario disponible.

Basándonos en todo lo expuesto, se llevó a cabo un análisis preliminar en colaboración con el área comercial de la compañía a la que se aplicará el proyecto. Durante este análisis, se identificó que existe una fuga de clientes que probablemente no es fácilmente detectada por la fuerza comercial debido al alto flujo de clientes y al elevado número de ventas, especialmente en el caso de aquellos clientes pequeños con volúmenes de compra no tan altos, pero con una frecuencia de compra constante. Por esta razón, el objetivo es desarrollar un modelo predictivo que permita identificar de manera oportuna la tendencia a la fuga, lo cual facilitará la creación de estrategias de fidelización y retención. Estas estrategias, a su vez, contribuirán a reducir la

pérdida de clientes y aprovechar los recursos para mejorar la calidad de servicio de la compañía de manera efectiva.

2. Objetivos

Objetivo General

Desarrollar una metodología que permita identificar de manera oportuna la fuga de clientes para diseñar e implementar estrategias de retención y fidelización.

Objetivos Específicos

- Identificar la fuga de clientes y tasas de fuga de forma preliminar con análisis descriptivos.
- Diseñar una segmentación acertada de clientes que generen valor a la compañía para aplicar las políticas de fidelización.
- Identificar clientes potenciales con tendencia a la fuga usando un modelo predictivo.
- Identificar los clientes valiosos y prioritarios para la empresa.

3. Alcance

Desde la comprensión y el conocimiento del negocio hasta la aplicación e implementación de diversas herramientas de Business Analytics, el objetivo es identificar la fuga o pérdida de clientes que impacta los presupuestos de ventas y las estrategias organizacionales. Con el fin de lograrlo, se seguirán los pasos correspondientes para llevar a cabo un análisis efectivo. Se espera desarrollar un modelo que permita a la compañía detectar de manera anticipada la fuga de clientes, lo cual servirá como herramienta para el diseño de estrategias de marketing, segmentación de clientes y planes de acción dentro de las áreas correspondientes.

4. Metodología

El proyecto empresarial será trabajado y desarrollado con las herramientas más utilizadas en el Business Analytics la cuales se enuncian a continuación:

- Metodologías agiles: usando esta metodología previamente se ha realizado el cronograma de trabajo, registrado los sprint y respectivos entregables. Adicional determinaremos factores importantes de la organización a través de las historias de usuario y su priorización, a su vez se realizará la matriz DOFA que nos dará a conocer más sobre el contexto del negocio y su problemática.

- Metodología crisp -dm: Con esta metodología realizaremos todo el proceso de entendimiento del negocio, comprensión de los datos, formulación de hipótesis, preparación de datos, modelado, la evaluación del resultado de los modelos desarrollados e implementación de este y respectivos ajustes que sean necesarios.

- Se realizarán un análisis RFM para generar una segmentación de los clientes e identificar nivel de fidelidad a través de su frecuencia y recurrencia de compra.

- El modelo predictivo se desarrollará con un modelo de clasificación.

5. Cronograma

Tabla 1
Cronograma desarrollo del proyecto

CRONOGRAMA				SCRUM BOARD
Tareas	Sprint	Inicio	Fin	Finalizada
Definición del cronograma	Ante Proyecto	01/07/2022	05/07/2022	*Finalizado*
Definición de alcance, objetivos, introducción		01/07/2022	20/07/2022	*Finalizado*
Identificación Fuentes de Información		20/07/2022	30/07/2022	*Finalizado*
Diseño de la Metodología por aplicar		01/08/2022	11/08/2022	*Finalizado*
Descripción de situación organizacional y Problemática	Desarrollo Matriz DOFA, historias de usuario e identificación de factores estratégicos y problemática	15/08/2022	10/09/2022	*Finalizado*
Aplicación Metodología CRISP-DM	Comprensión de los datos, Análisis en la calidad y dimensión de los datos, exploración de los datos en Python.	13/09/2022	10/10/2022	*Finalizado*
Análisis y de depuración de las bases de datos	Limpieza de datos	12/10/2022	30/10/2022	*Finalizado*
Generación de tablero	Visualización de datos	01/11/2022	12/11/2022	*Finalizado*
Retroalimentación	Revisar oportunidades de mejora	12/11/2022	15/11/2022	*Finalizado*
Actualización base de datos	Actualización nuevos registros y nueva variable	18/11/2022	22/11/2022	*Finalizado*
Revisión avances trabajo con director	Revisión y ajustes	26/11/2022	26/11/2022	*Finalizado*
Análisis descriptivo de la organización	Creación de tableros y análisis	16/11/2022	20/11/2022	*Finalizado*
Realización Modelo análisis RFM	Creación de variables y puntajes	21/11/2022	25/11/2022	*Finalizado*
Identificación y análisis de clientes Pareto	Clasificar clientes Pareto	27/11/2022	29/11/2022	*Finalizado*

Generación de clasificación por monto y tiempo de compra	Segmentación y análisis	30/11/2022	02/12/2022	*Finalizado*
Revisión avances trabajo con director	Preliminar de resultado	03/12/2022	03/12/2022	*Finalizado*
Ajuste y preparación de segundo entregable y presentación	segunda entrega proyecto	02/12/2022	06/12/2022	*Finalizado*
Presentación (Entregable)	Exposición avances y hallazgos preliminares	10/12/2022	10/12/2022	*Finalizado*
Recolección de datos adicionales de ser necesario para la identificación de causas de fuga	Recolección de datos adicionales para base	20/02/2023	22/02/2023	*Finalizado*
A partir de los datos transaccionales se debe identificar la tasa de fuga y probabilidad de fuga	datos de fuga y probabilidad de fuga	01/03/2023	30/03/2023	*Finalizado*
Corrida de modelo con base original	Modelado y análisis de resultados	05/04/2023	10/04/2023	*Finalizado*
Creación de nuevas variables para mejorar resultado del modelo		20/04/2023	05/05/2023	*Finalizado*
Probar con diferentes modelos de clasificación y realizar evaluación estadística de resultado obtenido		06/05/2023	15/05/2023	*Finalizado*
Tomar la decisión de cual modelo es el mejor para realizar las conclusiones		16/05/2023	20/05/2023	*Finalizado*
Realización de conclusiones y recomendaciones	Hallazgos y recomendaciones área comercial	21/05/2023	26/05/2023	*Finalizado*

Nota. Cronograma correspondiente al desarrollo del proyecto especificando tareas y tiempos de entrega. Fuente: Elaboración propia.

6. Descripción de la Situación organizacional donde se realizará el proyecto

Entendimiento del negocio:

La compañía objeto de estudio se dedica a la comercialización de envases de vidrio desde hace más de 60 años y se ha convertido en el principal distribuidor en el país. A lo largo de su trayectoria, ha ampliado su oferta incluyendo nuevos productos en su portafolio. Hace 30 años, incorporó la línea de productos químicos como un complemento a los envases de vidrio. Además, hace 12 años incursionó en la fabricación y comercialización de envases plásticos, aprovechando el crecimiento del mercado.

El nicho de mercado de la empresa abarca empresas pequeñas, medianas y grandes que fabrican productos cosméticos, alimentos, farmacéuticos y de aseo, todos ellos ubicados en el territorio nacional. Además, también atiende a personas naturales que fabrican productos artesanales para la comercialización informal. El negocio no se segmenta por estratos, ya que se trata de un sector industrial presente en diversas zonas del país, como parques industriales, barrios de la ciudad y zonas francas. Sin embargo, es posible categorizarlo según el sector al que pertenezca y el tipo de producto que se adquiera para su proceso.

De acuerdo con una investigación realizada por Asociación Colombiana de la industria de la comunicación grafica [ANDIGRAF] (2021), la industria colombiana ha experimentado un cambio en la tendencia de uso y consumo de los diferentes envases y empaques requeridos en los diferentes sectores de producción. Se han observado cambios en los materiales utilizados, pasando del vidrio al plástico PET, así como diseños más innovadores e incluso una disminución

en la capacidad o tamaño de los envases, pasando de los tradicionales 235 ml y 300 ml. Estos cambios responden a las transformaciones en los hábitos de consumo de los usuarios finales.

> Entre los años 2013 y 2018, las ventas de envases y empaques experimentaron un crecimiento del 0.9% en Colombia. En este periodo, tanto los envases de vidrio como los de plástico representaron el 18.6% de las ventas cada uno, sumando un total general del 37.2%. En cuanto a los sectores, la industria de alimentos y bebidas consumió el 93% de la producción total de Colombia, mientras que el resto se distribuyó en los demás sectores productivos. (Asociación Colombiana de la industria de la comunicación grafica, 2021, pp. 16–21)

Estos datos nos permiten observar que se trata de un sector con numerosas oportunidades de crecimiento y negocios, y que es fundamental generar valor para los clientes en este contexto.

7. Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución

Para iniciar con el desarrollo del análisis conoceremos inicialmente los objetivos estratégicos organizacionales para identificar su direccionamiento y metas propuestas a través de sus diferentes KPI´s de negocio.

Objetivos Estratégicos Organizacionales:

Para la organización, su prioridad y principal objetivo es ofrecer un servicio integral respaldado por diferentes canales de atención, siempre contando con el compromiso de sus equipos de trabajo.

Las acciones clave que se implementan para lograr este objetivo son:

1. Brindar productos con altos estándares de calidad y precios competitivos.
2. Trabajar constantemente en la mejora continua, fomentando el esfuerzo constante de los equipos de trabajo para ofrecer el mejor servicio posible a los clientes.
3. Mantener la fidelidad a los principios y valores organizacionales, transmitiendo confianza y tranquilidad a los clientes durante toda la relación comercial.
4. Mantener un enfoque constante y alineado con las necesidades y requerimientos de los clientes.

Estas acciones son fundamentales para asegurar que la organización satisfaga las expectativas de sus clientes y se mantenga en línea con sus valores organizacionales.

Adicionalmente, la compañía tiene una serie de indicadores relacionados a los objetivos que le permiten tener un seguimiento y control continuo sobre la ejecución de la estrategia.

KPI´s de negocio relacionados con los objetivos.

- Seguimiento continuo del mercado y de los diferentes cambios en precios relacionados a los factores económicos nacionales e internacionales.
- Seguimiento constante a los clientes para la identificación de nuevas necesidades y oportunidades de negocio.
- Nivel de reconocimiento y posicionamiento de la marca en el mercado.
- Medir el compromiso y trabajo en equipo al interior de la organización para mantener siempre la calidad de servicio.
- Conocer y entender los requerimientos de los clientes para ofrecer soluciones que agreguen valor a su negocio
- Incremento la tasa de retención de clientes

Factores Estratégicos Críticos:

Identificación de los factores externos e internos que están relacionados con el desarrollo del negocio y que de manera directa e indirecta impacta en el problema. Esta identificación lo haremos haciendo uso de la matriz DOFA aplicado a la compañía en estudio, este análisis nos dará un conocimiento preliminar del negocio y su posición en el mercado

Tabla 2

Matriz Dofa

FORTALEZAS	DEBILIDADES
1. Trayectoria y reconocimiento en el mercado. 2. Presencia comercial a nivel nacional con diferentes sucursales en ciudades principales. 3. Respaldo y acompañamiento en todos los procesos de nuevos desarrollos. 4. Equipo humano calificado y orientado a prestar el mejor servicio. 5. Proveedores reconocidos a nivel nacional e internacional. 6. Portafolio amplio en las diferentes líneas de productos. 7. Ofrecimiento de diferentes canales de venta facilitando la experiencia de compra al cliente. 8. Atención oportuna a las reclamaciones de calidad.	1. No existen políticas de seguimiento a clientes. 2. desaprovechamiento de las herramientas tecnológicas por parte del área comercial. 3. Incumplimiento a los compromisos de entregas adquiridos con anterioridad. 4. No se dan respuestas oportunas a clientes medianos y pequeños frente a reclamaciones de calidad. 5. Tiempos muy largos para el desarrollo de productos exclusivos y propios. 6. La no identificación oportuna de la pérdida de clientes. 7. Los precios actuales no son tan competitivos frente a la competencia.
OPORTUNIDADES	AMENAZAS
1. Creación de nuevos emprendimientos en los diferentes sectores del mercado que generan oportunidades de negocio para la compañía. 2. Reactivación del mercado después de la crisis de la pandemia. 3. La apertura continua de establecimientos de venta de productos de consumo masivo a bajo costo. 4. El desarrollo de nuevos productos y negocios de los clientes actuales	1. La devaluación del peso frente al alto incremento del dólar. 2. El incremento de las materias primas a nivel mundial. 3. El alto costo de los fletes marítimos que afecta la importación de productos 4. Bajos precios de la competencia aun con la situación económica. 5. Aumento de la Inflación y el precio del producto final. 6. La entrada de nuevos fabricantes y distribuidores al mercado.

Nota. Matriz Dofa realizada a empresa en estudio. Fuente: Elaboración propia

Problema o necesidad identificada:

Identificación Problema del Negocio.

Antes de abordar la problemática, es importante comprender que este proyecto tiene como objetivo mejorar el Marketing Industrial.

El Marketing Industrial se refiere a un tipo de negocio B2B (Business to Business) o marketing de empresa a empresa, que se centra en establecer relaciones sólidas y duraderas con los clientes.(Quiroa, 2022).

Una vez que se han identificado los principales factores críticos que influyen en la causa raíz del problema utilizando la matriz DOFA, se puede inferir de manera preliminar que los bajos precios ofrecidos por la competencia, la entrada de nuevos actores en el mercado, los factores internacionales y de suministro pueden ser factores influyentes en la decisión de los clientes al elegir a su proveedor, lo que los lleva a buscar diferentes alternativas en el mercado. Además, se puede afirmar que la falta de seguimiento por parte del área comercial podría influir en la falta de identificación oportuna de la fuga de clientes.

Problemas analíticos que resolver

a) Identificar las posibles causas de fuga de clientes mediante el conocimiento del negocio y el mercado.

b) Evaluar el nivel de fidelidad de los clientes según su frecuencia y repetición de compra utilizando el análisis RFM.

c) Utilizar modelos de clasificación para predecir de manera anticipada la tendencia de fuga de clientes y proponer estrategias de fidelización y retención.

Con base en lo anterior, se toman en cuenta los objetivos organizacionales y los factores estratégicos para elaborar las historias de usuario, asignando prioridad y estableciendo los criterios de aceptación correspondientes.

Historias de usuarios.

Tabla 3

Historias de usuario

Historia de usuario	Prioridad	Criterios de aceptación
1. Conocer las principales causas de fuga o fuga del cliente.	1. Alta	1 describir las posibles causas a partir del conocimiento de negocio del área comercial.
		2. Determinar el impacto que genera estas causas.
2. Tener de primera mano los reportes de ventas de los últimos 3 años para identificar la tasa de fuga.	2. Media	1. Tener un reporte del comportamiento de los clientes
		2. tener definido si las variables obtenidas son suficientes para el análisis de la causa.
3. La organización debe conocer sus clientes y respectiva segmentación para el enfoque de las estrategias.	3. Alta	1.Obteniendo un análisis RFM para identificar la fidelidad de los clientes.
		2. Identificar los clientes más potenciales y con probabilidad de fuga.
4. después de obtener todos los resultados de análisis descriptiva y de diagnóstico la organización espera los análisis predictivos para la toma de decisiones	4. Media	1. Identificando si las variables estudio son importantes y tienen relación con el problema.
		2. Desarrollando y Aplicando un modelo predictivo.

Nota. Relación Historias de usuario con respectivos criterios de aceptación y Prioridad. Fuente: Elaboración propia

Solución propuesta a la compañía en estudio.

Se propone a la organización comenzar con una segmentación e identificación adecuada de los clientes, para ubicarlos en un grupo prioritario para la atención y fidelización. De esta

manera, el área comercial podrá concentrar sus esfuerzos y recursos en este segmento. A continuación, se desarrollará un modelo predictivo que permita anticipar la fuga de clientes, y este modelo servirá como base para el diseño de estrategias de fidelización y retención, convirtiéndose así en la solución a la fuga de clientes.

Entre las características principales del producto a entregar se encuentra la generación de valor para el área comercial. A través de los resultados obtenidos, se fortalecerá el seguimiento continuo a los clientes, lo que ayudará a reducir el desgaste comercial causado por la gestión de recuperación. Además, esta herramienta será útil para la gerencia al tomar decisiones de carácter comercial y estratégico.

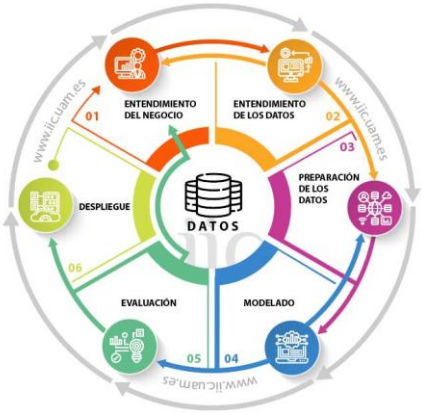
Análisis y limpieza de datos

El Paso por seguir, realizar el tratamiento de los datos por medio del método CRISP-DM.

A continuación, la ilustración del método y respectivo proceso.

Figura 1

Esquema Metodología CRISP-DM



Nota. Método para realizar Minería de Datos. Fuente: Instituto de ingeniería del conocimiento (2023)

Teniendo en cuenta que líneas atrás se puso en contexto todo lo relacionado con el entendimiento del negocio se procede a realizar todo el tratamiento y comprensión de la data suministrada por la organización.

Descripción de los datos:

El conjunto de datos consta de 916.173 registros y 7 variables, la cual se entregó completamente anonimizada para preservar la confidencialidad de la información.

A continuación, se realizará una revisión para determinar la calidad de los datos suministrados, aplicando validación de dimensiones en cuanto a su completitud, validez y precisión. Con los resultados obtenidos, se realizará el tratamiento correspondiente.

Tabla 4

Descripción Base de datos

	DESCRIPCIÓN VARIABLES
Id_Cliente	Código Asignado al cliente
Fecha_Contabilización	Fecha correspondiente a la transacción
Doc_Num	Numero de factura o documento asociado a la transacción
Referencia	Numero asignado a cada una de las referencias
Venta_Unidades	Unidades vendidas a cada cliente
Venta_Dinero	Monto en dinero Vendido a cada cliente
Municipio	Lugar de ubicación del cliente

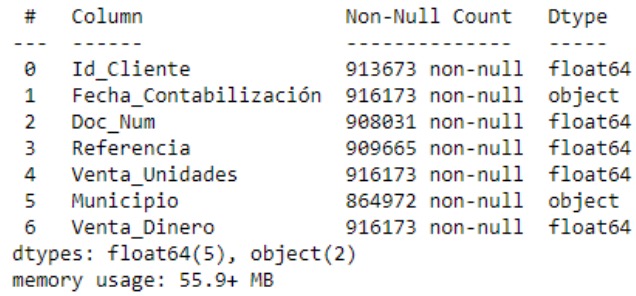
Nota. Descripción de cada una de las variables. Fuente: Elaboración propia

Dimensión de Completitud:

Realizando esta validación se espera identificar que tan completa es la información suministrada, con cuantos registros reales se cuenta y el tipo de variable. Para ello se realiza una exploración en Python para obtener un resultado a través de esta herramienta.

Figura 2

Dimensión Variables



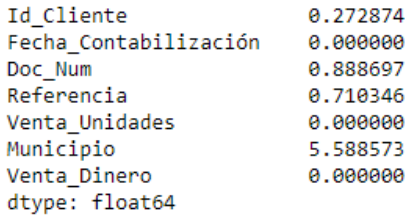
```
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Id_Cliente            913673 non-null  float64
 1   Fecha_Contabilización  916173 non-null  object
 2   Doc_Num               908031 non-null  float64
 3   Referencia            909665 non-null  float64
 4   Venta_Unidades        916173 non-null  float64
 5   Municipio             864972 non-null  object
 6   Venta_Dinero          916173 non-null  float64
dtypes: float64(5), object(2)
memory usage: 55.9+ MB
```

Nota. Resultado de la validación de completitud en Python. Fuente: Elaboración propia

De acuerdo con el resultado de la **Figura 2** se puede denotar que la base no cuenta con registros completos en 3 de las 7 variables.

Figura 3

Porcentaje de faltantes



```
Id_Cliente              0.272874
Fecha_Contabilización   0.000000
Doc_Num                 0.888697
Referencia              0.710346
Venta_Unidades          0.000000
Municipio               5.588573
Venta_Dinero            0.000000
dtype: float64
```

Nota. Valores porcentuales en los faltantes de cada una de las variables. Fuente: Elaboración propia

La variable Municipio es la que presenta mayor faltante de datos por lo que será la de mayor revisión y tratamiento. Dicho resultado es tal como se denota en la **Figura 3.**

Con los anteriores resultados se procede a realizar la tabla con las dimensiones de completitud y sus respectivos porcentajes

Tabla 5

Completitud de los datos

VARIABLES	ATRIBUTOS REQUERIDOS	REGISTROS TOTALES	COMPLETITUD
Id_Cliente	916,713	913,673	99.7%
Fecha_Contabilización	916,713	916,713	100%
Doc_Num	916,713	908,031	99,1%
Referencia	916,713	909,665	99,3%
Venta_Unidades	916,713	916,713	100%
Venta_Dinero	916,713	916,713	100%
Municipio	916,713	864,972	94,4%

Nota. Identificación de la completitud de los datos. Fuente: Elaboración propia

Usando la herramienta Python se logró identificar que los datos suministrados por la compañía no cuentan con los registros completos en la mayoría de las variables, tal y como lo podemos evidenciar en la anterior tabla.

Dimensión de Validez:

Tabla 6

Validez de los datos

VARIABLES	NOMBRE VALIDO	REGISTROS TOTALES	VALIDEZ
Id_Cliente	916,713	916,713	100%
Fecha_Contabilización	916,713	916,713	100%
Doc_Num	916,713	916,713	100%
Referencia	916,713	916,713	100%

VARIABLES	NOMBRE VALIDO	REGISTROS TOTALES	VALIDEZ
Venta_Unidades	916,713	916,713	100%
Venta_Dinero	916,713	916,713	100%
Municipio	410,822	916,713	44.8%

Nota. Identificación de validez en los datos. Fuente. Elaboración propia

En cuanto a la validez de los datos se encontró que la variable Municipio tan solo el 44.8% de los datos están correctos y bien diligenciados los nombres de cada uno de las ciudades o municipios, al restante es necesario realizar un ajuste y corrección.

Dimensión de Precisión:

No pueden tener valores negativos o en 0 las variables Venta_Unidades y Ventas_Dinero.

Tabla 7

Precisión de los datos

VARIABLES	REGISTROS PRECISOS	REGISTROS TOTALES	PRECISIÓN
Id_Cliente	916,713	916,713	100%
Fecha_Contabilización	916,713	916,713	100%
Doc_Num	916,713	916,713	100%
Referencia	916,713	916,713	100%
Venta_Unidades	885,629	916,713	97%
Venta_Dinero	882,995	916,713	96%
Municipio	916,713	916,713	100%

Nota. Se identifica datos negativos o con 0. Fuente: Elaboración propia, (2022)

Se identificó que las variables Venta_Unidades y Ventas_Dinero contienen valores negativos dentro de los registros transaccionales, por lo tanto, es necesario tratar y corregir.

Las dimensiones Puntualidad, Unicidad y Consistencia no pueden ser valoradas porque al no tener la base datos maestra no tenemos como validar la información original.

Recopilación de los datos:

Los datos obtenidos fueron recopilados durante una serie de tiempo que comprende del 1 de enero del 2.020 hasta el 18 de noviembre del 2022. Dichos datos requieren de información adicional para poder realizar el modelo de clasificación en Python.

Exploración de los datos:

Se realiza una exploración de los datos usando Python para identificar información relevante.

Figura 4

Exploración datos

Id_Cliente	Fecha_Contabilización	Doc_Num	Referencia	Venta_Unidades	Municipio	Venta_Dinero
20.0	01/01/2020	1.0	43.0	40000.0	NaN	3.080000e+06
20.0	01/01/2020	1.0	42.0	40000.0	NaN	5.080000e+06
33783.0	01/01/2020	2.0	314.0	45500.0	BOGOTA	1.092000e+07
9873.0	02/01/2020	3.0	117.0	26400.0	FUNZA	6.072000e+06
33813.0	07/01/2020	96.0	160.0	72.0	CALI	1.072800e+04
...	...	...	...	...	...	...
1120.0	18/11/2022	NaN	939.0	-8.0	ITAGUI	-7.637840e+03
1120.0	18/11/2022	NaN	28.0	-1.0	ITAGUI	-7.075400e+04
1120.0	18/11/2022	NaN	1258.0	-2.0	ITAGUI	-3.612666e+04
1120.0	18/11/2022	NaN	232.0	-1.0	ITAGUI	-1.902701e+04
1120.0	18/11/2022	NaN	55.0	-1.0	ITAGUI	-4.159656e+04

916173 rows x 6 columns

Nota. Exploración preliminar de datos en Python. Fuente: Elaboración propia

En esta primera descripción identificamos que la información no viene limpia y contiene datos nulos, erroneos, existen Outliers etc, por lo tanto nos lleva a realizar un tratamiento y limpieza de la base, el tratamiento se realizara directamente en Excel.

Figura 5

Datos nulos

Id_Cliente	Fecha_Contabilización	Doc_Num	Referencia	Venta_Unidades	Municipio	Venta_Dinero
20.0	False	False	False	False	True	False
20.0	False	False	False	False	True	False
33783.0	False	False	False	False	False	False
9873.0	False	False	False	False	False	False
33813.0	False	False	False	False	False	False
...	...	...	...	...	...	...
1120.0	False	True	False	False	False	False
1120.0	False	True	False	False	False	False
1120.0	False	True	False	False	False	False
1120.0	False	True	False	False	False	False
1120.0	False	True	False	False	False	False

916173 rows x 6 columns

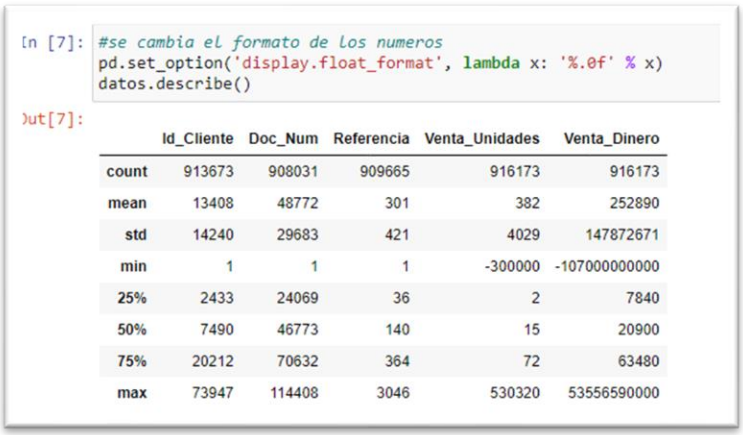
Nota. Identificación datos Nulos en la base. Fuente: Elaboración propia

Se logra identificar que en las variables Doc_num y Municipio contienen datos nulos los cuales se deben verificar y tomar las respectivas medidas de acuerdo con la importancia de los datos que contengan las otras variables.

Ahora, realizamos una descripción de los datos para identificar máximos, mínimos, media etc.

Figura 6

Descripción datos



```
In [7]: #se cambia el formato de los numeros
        pd.set_option('display.float_format', lambda x: '%.0f' % x)
        datos.describe()
```

Out[7]:

	Id_Cliente	Doc_Num	Referencia	Venta_Unidades	Venta_Dinero
count	913673	908031	909665	916173	916173
mean	13408	48772	301	382	252890
std	14240	29683	421	4029	147872671
min	1	1	1	-300000	-107000000000
25%	2433	24069	36	2	7840
50%	7490	46773	140	15	20900
75%	20212	70632	364	72	63480
max	73947	114408	3046	530320	53556590000

Nota. Se Identifica distribución y descripción estadística de los datos. Fuente: Elaboración propia

Por ser una base de datos anonimizada se le da un número de identificación a cada cliente para remplazar su Nit o cedula, dicha identificación inicia desde cliente 1 hasta el cliente 73947.

También se valida que la información cuenta con valores negativos, esta es información que no va a aportar al análisis por lo que también se le debe dar un tratamiento para que no altere los resultados. Por ultimo Las ventas están concentradas en 2.537 referencias.

Figura 7

Número de clientes

```
337.0      3056
416.0      2834
375.0      2699
1813.0     2411
4355.0     1853
           ...
23304.0       1
24403.0       1
11669.0       1
35041.0       1
37703.0       1
Name: Id_Cliente, Length: 53538, dtype: int64
```

Nota. Se identifica total real de clientes. Fuente: Elaboración propia

Por último, con el fin de tener claridad en el total de clientes que se va a trabajar, se evidencia en la **Figura 7** que las transacciones corresponden a las compras de 53.538 clientes.

Figura 8

Variable Municipio

```
In [12]: datos[["Municipio"]].value_counts(sort=False)

Out[12]: Municipio
         0                          18
         1                           6
         11 DE NOVIEMBRE            23
         11001                      18
         3205901714                  4
                                    ..
         santander                   3
         tenjo                       4
         via, Zipaquira - Briceno   15
         zipaquira                   5
         ÁBREGO                      2
         Length: 1376, dtype: int64
```

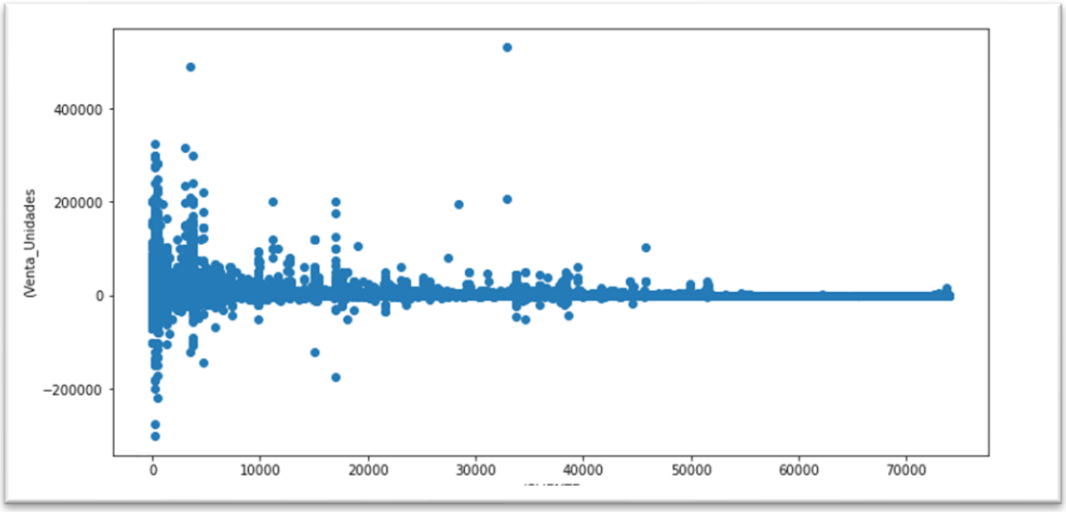
Nota. Revisión variable Municipio Fuente: Elaboración propia

Como se mencionó anteriormente, la variable Municipio es la que más contiene datos nulos, es por ello por lo que se procede a validar de manera más completa. De acuerdo con la **Figura 8** se identifica que adicional tiene información mal diligenciada y errónea por lo que se hace indispensable limpiar esta variable, realizar una depuración o corrección según sea el caso.

Para hallar datos atípicos o outliers se realizó un gráfico de dispersión en Python tomando como punto de análisis la variable Venta_Unidades y el código de Cliente.

Figura 9

Identificación datos atípicos



Nota. Gráfico de dispersión para identificar datos atípicos. Fuente: Elaboración propia

Se identifica en el conjunto de datos 25.448 valores negativos Los cuales representan un 2.77% del total de los datos. Adicional, también encontramos dos outliers los cuales se encuentra ubicados hacia la parte superior derecha e izquierda tal como lo muestra la **Figura 9.**

Con respecto a los clientes con consumos muy altos se considera relevante dejarlos para que sean parte del análisis porque es importante identificar su frecuencia de compras y si están activos actualmente.

Verificación de la calidad de los datos:

Una vez realizado todo el entendimiento de los datos y respectiva descripción de estos se logró obtener un resultado más claro respecto a la mala calidad, por lo que se hace necesario realizar una limpieza a toda la base para eliminar los datos atípicos, revisar los nulos y determinar que tanto afectaría al total de datos en caso de que se decida eliminarlos.

Por otro lado, sería la primera recomendación que se debe realizar a la empresa en estudio para mejorar en la recopilación y almacenamiento de los datos y que para futuros análisis se pueda obtener mejores resultados.

Preparación de los datos

Identificado que la base de datos contiene información que debe ser depurada o corregida, a continuación, se detalla el paso a paso para la respectiva limpieza de datos, este proceso se realizara directamente en la base de datos de Excel donde se realizara Inclusión o exclusión de datos según sea el caso o la necesidad.

Limpieza de datos:

En primer lugar, se va a tratar los datos nulos, donde se identificará cual o cuales variables contienen más datos nulos y se definirá si se eliminan o se les da algún valor en caso de que llegara afectar los datos su eliminación.

Para este caso se cogió primero la variable **Municipio;** se encontró que tienen 51.741 registros nulos los cuales corresponden a un 5,58%, pero al revisar las variables transaccionales notamos que tenía datos muy importantes en cuanto a número de clientes y se determinó que se va a remplazar estos espacios nulos por la categoría OTRO.

Con ello buscamos dentro del reporte que se presente resaltar la importancia de diligenciar correctamente la información y de no afectar el análisis. Por último, con los datos mal diligenciados o con errores, se procedió a corregirlos para unificarlos de acuerdo con las ciudades, departamentos o municipios que forman correcta.

Figura 10

Limpieza variable Municipio

Id_Clien	Fecha_Contabilizaci	Doc_Nu	Referenc	Venta_Unidad	Municipio	Venta_Dine
20	01/01/2020	1	43	40000	OTRO	3080000
20	01/01/2020	1	42	40000	OTRO	5080000
8659	07/01/2020	77	21	1	OTRO	14540
8709	07/01/2020	48	35	2	OTRO	40030
8709	07/01/2020	48	13	24	OTRO	4992
8709	07/01/2020	48	306	24	OTRO	4992
8659	07/01/2020	77	146	24	OTRO	4224
8659	07/01/2020	77	168	150	OTRO	6150
17452	07/01/2020	78	190	24	OTRO	4152
17452	07/01/2020	78	21	1	OTRO	14540
9526	07/01/2020	6	21	1	OTRO	13230
9526	07/01/2020	6	17	2	OTRO	40200
9526	07/01/2020	6	30	48	OTRO	9504
9526	07/01/2020	6	1993	24	OTRO	3960
19513	07/01/2020	50	1375	1	OTRO	38126
19513	07/01/2020	50	177	64	OTRO	8576
33788	07/01/2020	20	11	4	OTRO	45924
33788	07/01/2020	20	10	96	OTRO	10944
21014	07/01/2020	18	4	1	OTRO	17671
21014	07/01/2020	18	69	12	OTRO	504

Nota. Validación de corrección de la variable Municipio. Fuente: Elaboración propia

La Variable **Referencia** tiene 6.508 registros nulos y al validar con las demás variables reporta valores en 0 y otros nulos. Por lo que se toma la decisión de eliminar esos registros. Se puede evidenciar tal como muestra el ejemplo de la **Figura 11**.

Figura 11

Identificación Variable Referencia

Id_Clien	Fecha_Contabilizaci	Doc_Nu	Referenc	Venta_Unidad	Venta_Dinero	Municipio
	10/12/2020			0	-$ 10,187	#N/D
50369	27/10/2020			0	-$ 918,918	BOGOTA
50164	26/10/2020			0	-$ 30,375	BOGOTA
45687	11/08/2020			0	-$ 22,448	BARRANQUILLA
42872	03/07/2020			0	-$ 47,124	INDEFINIDO
40323	01/07/2020			0	-$ 136,370	RIONEGRO
39024	07/05/2020			0	-$ 8,100	CALI
39023	07/05/2020			0	-$ 3,896	CALI
38726	30/06/2020			0	-$ 10,662	ARAUCA
38356	22/04/2020			0	-$ 4,422	CALI
38045	04/09/2020			0	-$ 1,500,000	INDEFINIDO
35657	12/02/2020			0	-$ 9,103	CALI
35657	29/07/2020			0	-$ 4,304	CALI
35657	24/09/2020			0	-$ 11,688	CALI
34628	26/02/2020			0	-$ 343,400	ESPINAL
34628	16/06/2020			0	-$ 147,015	ESPINAL

Nota. Se valida que la variable contiene datos nulos. Fuente: Elaboración propia

La Variable **Doc_Num** tiene 7.395 registros nulos, y al validar con las otras variables se evidencia que no afectaría el eliminarlas ya que se detalla valores en 0 o negativos y estos no afectan nuestro análisis. Tal como muestra el ejemplo de **Figura 12**.

Figura 12

Limpieza variable Doc_Num

Id_Clien	Fecha_Contabilizaci	Doc_Nu	Referenc	Venta_Unidad	Municipio	Venta_Dine
10345	07/01/2020		102	-24	BOGOTA	-3960
8151	09/01/2020			0	MEDELLIN	-12000
8483	09/01/2020			0	CALI	-31570
400	10/01/2020		741	-336	CAJICA	-31248
3363	10/01/2020			0	CALI	-717466
934	13/01/2020		1119	-3	CALI	-67674
9030	13/01/2020		1177	-1	HUILA	-10101.791
9030	13/01/2020		1177	-7	HUILA	-73787
9030	13/01/2020		38	-48	HUILA	-7200
9030	13/01/2020		148	-49	HUILA	-7350
9030	13/01/2020		104	-50	HUILA	-7500
9030	13/01/2020		78	-50	HUILA	-7500
5469	14/01/2020			0	CALI	-6401
3828	14/01/2020		1213	-10	BUENAVENTURA	-213550
33937	15/01/2020		98	-3875		-155000
250	15/01/2020		104	-32343	BARRANQUILLA	-3644409.24
611	16/01/2020			0	CALI	-24337
774	16/01/2020			0	CALI	-15494
24551	16/01/2020			0	BOGOTA	-406350
943	17/01/2020			0	BOGOTA	-79605

Nota. Se demuestra que la variable contiene los datos nulos. Fuente: Elaboración propia

Como se evidenció en el gráfico de dispersión tenemos muchos datos negativos, y al validar en el conjunto de datos se identifica que estos ascienden a 25.448 registros, por lo tanto, se procede a eliminarlos para dejar solo valores positivos.

Por último, se evidencio que la variable Cliente tenía 2.500 datos nulos. Se contrastaron con las otras variables y al analizarlos se encontró que no había forma de reemplazarlos, promediarlos o realizar otro ajuste se decide eliminarlos.

- **Informe limpieza de datos**

Con el objetivo de tener las transacciones de los tres últimos años de manera completa se incluye los registros hasta el 29 de diciembre del 2022. A continuación, se presentará el total de los registros con los que se va a trabajar.

Figura 13

Identificación total datos

```
<class 'pandas.core.frame.DataFrame'>
Index: 939550 entries, 01/01/2020 to 29/12/2022
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   DocNum             930846 non-null  float64
 1   Referencia         932627 non-null  float64
 2   SUM(Cantidad)      939550 non-null  float64
 3   Municipio          887330 non-null  object
 4   Cliente            937050 non-null  float64
 5   SUM(Precio_total)  939550 non-null  float64
dtypes: float64(5), object(1)
memory usage: 50.2+ MB
```

Nota. La figura muestra el total de registros con el que se realiza el análisis y la fecha de recolección de estos. Fuente: Elaboración propia

Luego de la adición, se obtiene un total de 939.550 registros los cuales corresponden a las ventas desde el 01/01/2020 hasta el 29/12/2022. A estos datos se le realizo la respectiva limpieza tal y como se hizo anteriormente, siguiendo los mismos pasos.

Tras analizar los datos y las variables existentes, se ha determinado que es necesario incluir más variables para enriquecer los modelos y obtener mejores resultados. Por lo tanto, se solicita complementar la base de datos con esta información adicional. Una vez que se haya suministrado la información, se procederá a incorporarla al proyecto. Además, se crearán variables adicionales mediante una segmentación y clasificación que se describirá más adelante.

En este sentido, a continuación, se proporciona una descripción detallada de cada una de las variables disponibles actualmente.

Tabla 8

Descripciones variables

	DESCRIPCIÓN VARIABLE
Id_Cliente	Código Asignado al cliente
Fecha_Contabilización	Fecha correspondiente a la transacción
Doc_Num	Numero de factura o documento asociado a la transacción
Referencia	Numero asignado a cada una de las referencias
Venta_Unidades	Unidades vendidas a cada cliente
Venta_Dinero	Monto en dinero Vendido a cada cliente
Municipio	Lugar de donde realiza la compra el cliente
Vidrio	Línea de envases de vidrio
Plástico	Línea envases plásticos
Químico	Línea de Ingredientes químicos
Cristar	Línea de productos de cristalería
Complementos	Línea de tapas, válvulas y tapones

	DESCRIPCIÓN VARIABLE
Tiene_Credito	Identifica al cliente que tiene o no crédito
Clasificación_Cliente	Clasificación según su comportamiento de compra
Recency	Días trascurridos desde la última compra
Frequency	Frecuencia de compra en un tiempo determinado
Monetary	Valor total de las compras en dinero
RFM	Puntaje obtenido de la calificación de (Recency, Frequency, Monetary)
Fuga	Variable dependiente para el modelo predictivo

Nota. Tabla con total de variables a trabajar. Fuente: Elaboración propia

A los nuevos datos se les realizó un análisis y revisión en Python para determinar su calidad y completitud. Como podemos observar en la **Figura 14** y **Figura 15** ya no tienen datos incompletos, nulos y se puede observar que se aumentó a 18 variables. Una vez validado lo anterior, se procede a reducir la base de los 939.550 registros a transacciones únicas por cliente llegando así a obtener un total de 54.157 filas o clientes con registros únicos, se realiza este proceso con el propósito de tener un conjunto de datos consolidado y que ofrezca un mejor análisis al momento de realizar los pasos para el modelado.

Figura 14

Nuevos registros

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 54157 entries, 1 to 74730
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Municipios            54157 non-null  int64
 1   Ventas_Dinero2020     54157 non-null  int64
 2   Ventas_Unidades2020   54157 non-null  int64
 3   Ventas_Dinero2021     54157 non-null  int64
 4   Ventas_Unidades2021   54157 non-null  int64
 5   Ventas_Dinero2022     54157 non-null  int64
 6   Ventas_Unidades2022   54157 non-null  int64
 7   Vidrio                54157 non-null  object
 8   Plastico              54157 non-null  object
 9   Quimico               54157 non-null  object
 10  Cristal               54157 non-null  object
 11  Complementos          54157 non-null  object
 12  Tiene_Credito         54157 non-null  object
 13  Clasificacion_Cliente 54157 non-null  object
 14  RFM                   54157 non-null  int64
 15  Recency               54157 non-null  int64
 16  frecuency             54157 non-null  int64
 17  Monetary              54157 non-null  int64
 18  Fuga_Dependiente      54157 non-null  int64
dtypes: int64(12), object(7)
memory usage: 8.3+ MB
```

Nota. La figura muestra el total de registros y número de clientes. Fuente: Elaboración propia

Figura 15

Datos no nulos

	Id_Cliente	Municipios	Ventas_Dinero2020	Ventas_Unidades2020	Ventas_Dinero2021	Ventas_Unidades2021	Ventas_Dinero2022	Ventas_Unidades2022
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
54153	False	False	False	False	False	False	False	False
54154	False	False	False	False	False	False	False	False
54155	False	False	False	False	False	False	False	False
54156	False	False	False	False	False	False	False	False
54157	False	False	False	False	False	False	False	False

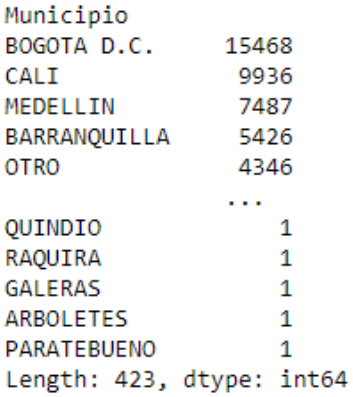
54157 rows × 20 columns

Nota. En la figura se observa que la data esta sin nulos. Fuente: Elaboración propia

Para terminar con la revisión y validación de los nuevos datos, se verifica la corrección realizada la variable Municipio, corrección que se puede evidenciar en la **Figura 16** donde los datos ya quedaron bien diligenciados y sin errores.

Figura 16

Corrección Variable Municipio



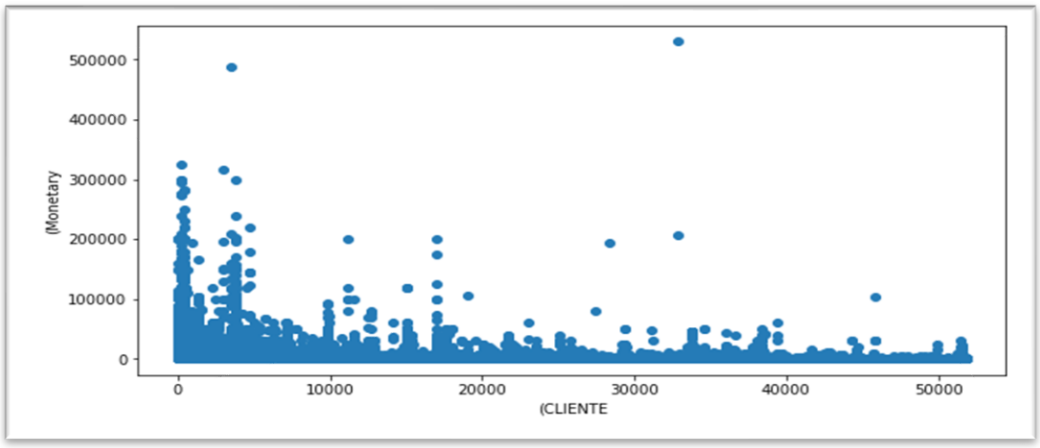
```
Municipio
BOGOTA D.C.      15468
CALI              9936
MEDELLIN          7487
BARRANQUILLA      5426
OTRO              4346
                  ...
QUINDIO              1
RAQUIRA              1
GALERAS              1
ARBOLETES            1
PARATEBUENO          1
Length: 423, dtype: int64
```

Nota. Figura que evidencia corrección en la variable Municipio. Fuente: Elaboración propia

Se eliminaron todos los datos negativos que nos arrojaba en la primera descripción. Aun continua un par de outliers los cuales vamos a mantener para el análisis. En caso de ver que no son tan relevantes a medida que se vaya realizando el proyecto se realizara su eliminación.

Figura 17

Distribución datos



Nota. El grafico de dispersión muestra cómo queda la distribución después de eliminar valores negativos Fuente: Elaboración propia

8. Análisis Descriptivo de la Organización e identificación de fuga y probabilidades de fuga.

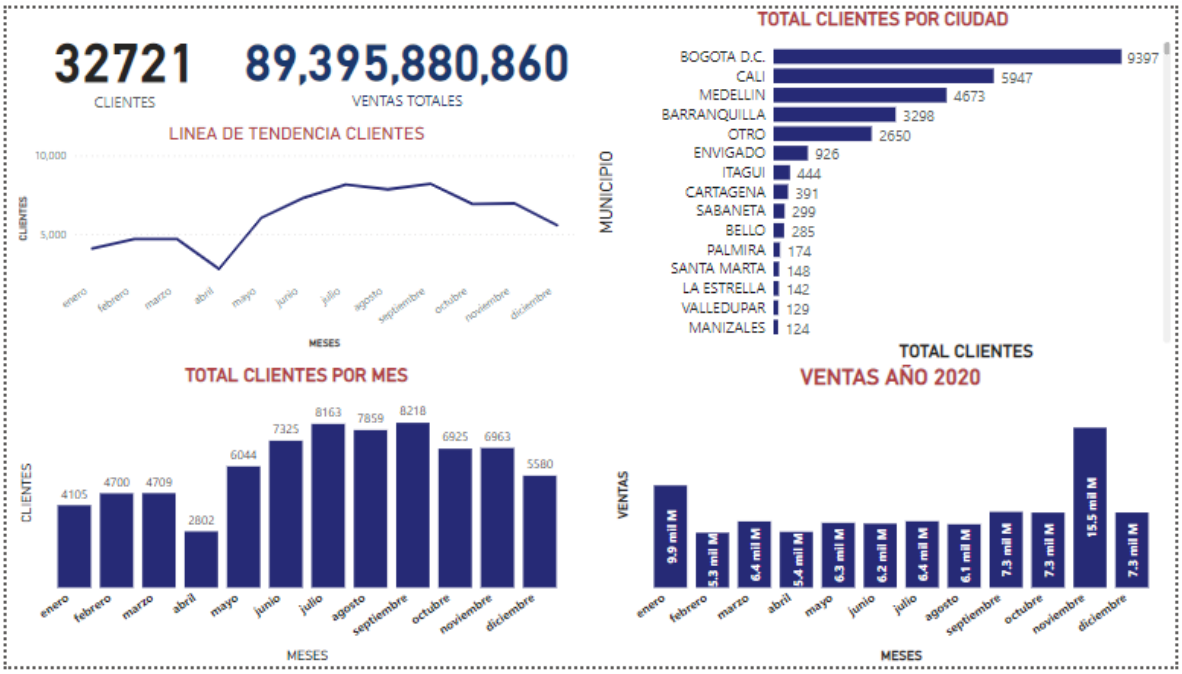
Una vez que se ha completado la limpieza y preparación de la base de datos, es momento de proceder con el análisis y el proceso. El siguiente paso consiste en realizar un análisis exhaustivo de la organización en términos del comportamiento de compra, la frecuencia y la antigüedad de los clientes.

Para llevar a cabo este paso, se tomarán en cuenta las transacciones realizadas desde enero de 2020 hasta diciembre de 2022. Se realizará un análisis año tras año de las ventas y del número de clientes mes a mes. Este análisis proporcionará una visión preliminar de la situación de la compañía en lo que respecta a sus clientes y ventas.

La compañía en estudio registra ventas anuales que superan los ochenta mil millones de pesos. Para el año 2020 el número de clientes que realizaron las compras fue de 32.721, los cuales presentan un comportamiento mes a mes donde podemos denotar que tiene meses con más participación que otros, tal como muestra la **Figura 18**. Entrando en detalle, se identifica que el mes de abril tuvo un decrecimiento importante de clientes respecto al primer trimestre del año, pero también cabe resaltar que, aunque los meses posteriores tuvieron una mayor participación de clientes se evidencia que en el último trimestre nuevamente hay disminución de clientes frente al tercer trimestre

Figura 18

Visualización Año 2.020



Nota. La figura nos muestra como fue el comportamiento de las ventas y los clientes año 2020. Fuente: Elaboración propia

Aunque la compañía realiza ventas a nivel nacional en más de 420 municipios, la mayoría de sus clientes están ubicados o realizan compras en las principales ciudades como Bogotá, Medellín, Cali y Barranquilla. Esto se puede atribuir a la presencia de sucursales en cada una de estas ciudades.

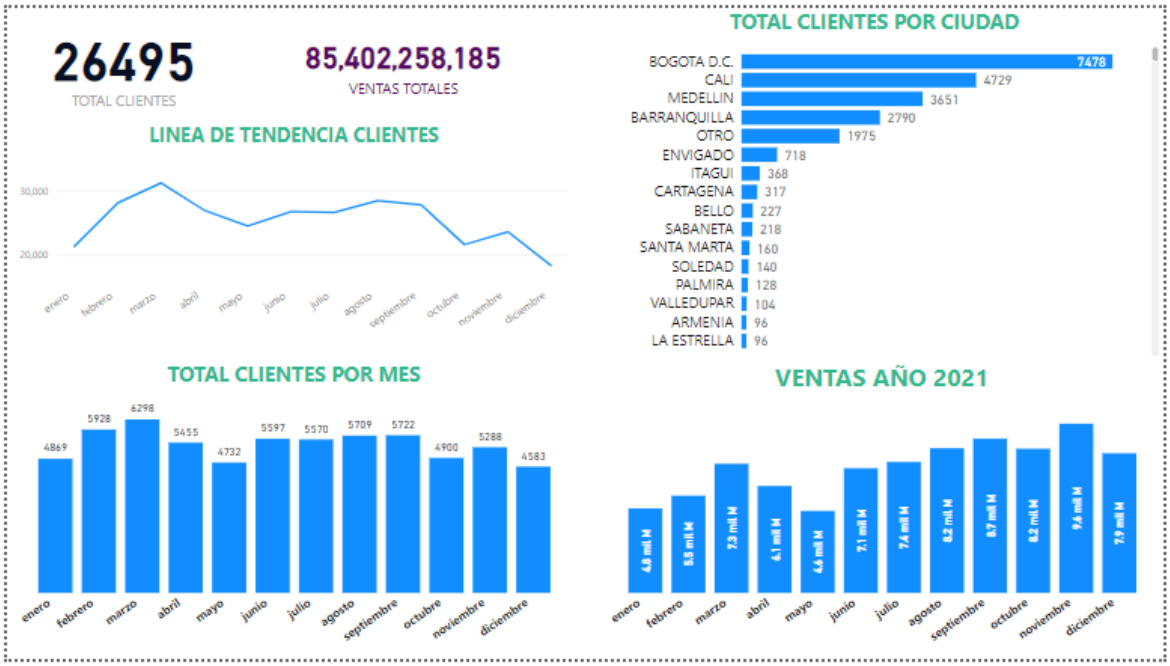
Por otro lado, en el año 2021 se observa una disminución significativa en el número de clientes en comparación con el año 2020, como se muestra en la Figura 19. Durante este año, las ventas se realizaron a 26.495 clientes.

Ante la reducción en el número de clientes, se puede suponer que existe una fuga de clientes. Sin embargo, es importante determinar si estos clientes dejaron de comprar de forma permanente o simplemente han ajustado su frecuencia de compra.

Con el análisis del año 2022, podremos confirmar la existencia del problema de fuga de clientes y determinar la tasa de fuga y el porcentaje correspondiente. Será crucial abordar este problema de forma inmediata para su pronta solución.

Figura 19

Visualización Año 2.021



Nota. La figura nos muestra como fue el comportamiento de la organización en ventas y clientes año 2021. Fuente: Elaboración propia

De acuerdo con la **Figura 20**, se observa una continua disminución en el número de clientes en comparación con el año anterior. Para el año 2022, se registraron 20.833 clientes.

Es importante aclarar que, al revisar la visualización, se nota que el promedio mensual de clientes que realizan compras es de aproximadamente 4.250. Si sumamos estos valores mes a mes, obtendríamos más de 51.000 clientes en total. Sin embargo, esto no es indicativo de la cantidad real de clientes, ya que las compras pueden variar en frecuencia y ser esporádicas, dependiendo de las necesidades y demandas individuales de cada cliente.

En este análisis descriptivo de la organización, se confirma la existencia de una fuga de clientes. Más adelante, se procederá a identificar tanto la fuga real como la probabilidad de fuga. Utilizando el número de clientes por año y aplicando la fórmula correspondiente, podremos determinar la tasa de fuga anual con base en los datos anteriores.

$$\frac{\text{\# Clientes inicio periodo - \# Clientes final de periodo}}{\text{\# Clientes inicio periodo}} \times 100$$

Tabla 9

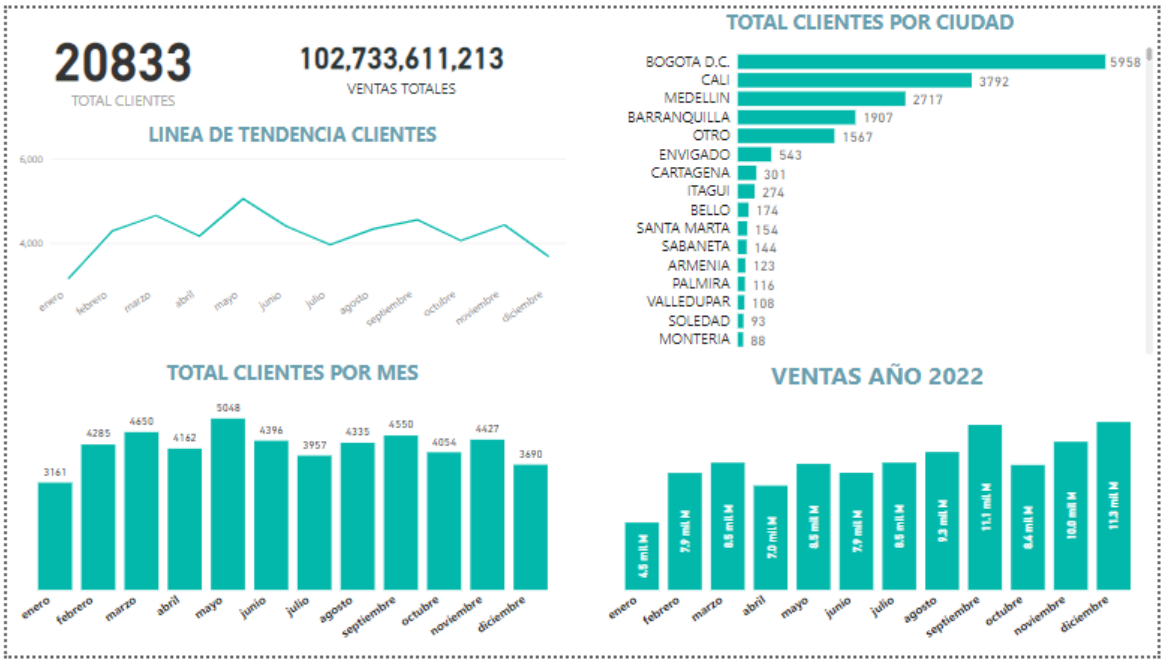
Tasa Fuga clientes por año

Año	Clientes inicio periodo	Clientes final de periodo	Tasa de Fuga
2021	32.721	26.495	19%
2022	26.495	20.833	21%

Nota. Se identifica la tasa de fuga anual basado en los datos descriptivos. Fuente. Elaboración propia

Figura 20

Visualización Año 2022



Nota. Comportamiento de la organización en ventas y clientes año 2022. Fuente: Elaboración propia

Tasas de fuga y probabilidad fuga total.

Como se menciona en la introducción, el sector industrial es altamente competitivo y carece de la implementación de herramientas tecnológicas para la detección oportuna de la fuga de clientes, convirtiendo la pérdida de clientes en un problema empresarial. Por lo tanto, retener a los clientes requiere comprender y entender cómo deben ser construidas e implementadas todas las estrategias para fomentar la lealtad del cliente y establecer relaciones comerciales a largo plazo.

Es evidente que, para las empresas, su principal activo es el cliente, por lo tanto, es imperativo conocer a fondo los diferentes comportamientos de consumo para desarrollar políticas y estrategias que prolonguen y aprovechen el potencial comercial. (García et al., 2017)

La minería de datos y la construcción de modelos predictivos de fuga de clientes aún es un tema desconocido en el sector industrial. Por esta razón, se pretende desarrollar este proyecto con el objetivo de dar los primeros pasos para encontrar las herramientas más adecuadas y realizar los cambios necesarios para obtener resultados que permitan crear estrategias más eficientes y generar valor tanto para la empresa como para el cliente.

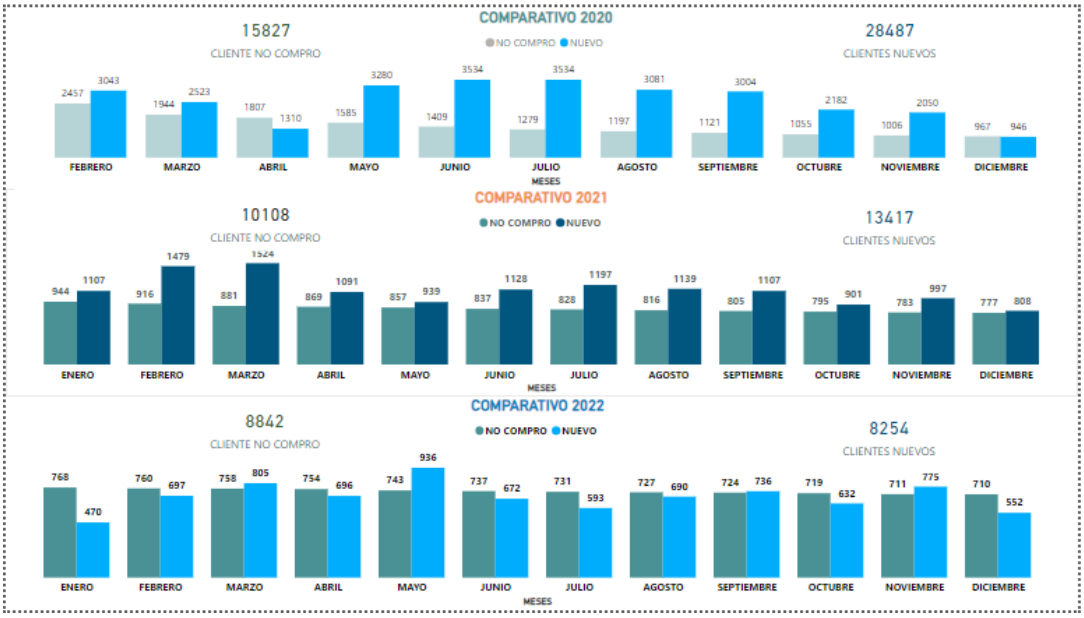
Dentro de la compañía objeto de análisis, no existe una política clara para identificar a los nuevos clientes, aquellos que dejaron de comprar e incluso los que se han fugado. Del mismo modo, no se cuenta con un modelo de segmentación basado en la frecuencia de compra, el tiempo en que el cliente dejó de comprar, entre otros aspectos relevantes. Estos factores son importantes para determinar los clientes potenciales a los cuales se deben implementar políticas de retención y fidelización.

Para iniciar el análisis del comportamiento de los clientes, se procede a identificar el número de nuevos clientes y los que dejaron de comprar durante los tres años de análisis, realizando una comparación mes a mes y separando los datos por cada uno de los años.

Como se mencionó en el capítulo anterior, este mercado es altamente dinámico y presenta una rotación de clientes mes a mes, como se muestra en la **Figura 21**. Esta visualización tiene como objetivo comprender el comportamiento mensual de los clientes en términos numéricos y determinar si hay más clientes que no realizan compras o si hay más clientes nuevos.

Figura 21

Visualización clientes nuevos y que no compraron



Nota. La visualización muestra como es la dinámica entre clientes nuevos y que no compraron en una comparación mes a mes. Fuente: Elaboración propia

En el análisis, se observa que para el año 2020, el número de clientes nuevos fue significativamente mayor en comparación con los clientes que dejaron de comprar en ese mismo año. Esto se traduce en una diferencia positiva del 79,9%. Sin embargo, es importante destacar que durante ese periodo el país estaba experimentando una pandemia, lo cual influyó en ese aumento debido a que muchos clientes estaban adquiriendo envases y complementos para productos como gel antibacterial y alcohol. Por esta razón, debemos tener cuidado al interpretar estos datos, ya que podrían sesgar nuestra comprensión y no ser un punto de referencia relevante.

En el año 2021, se observa que el número de clientes nuevos supera a los clientes fugados, con una diferencia positiva del 32,7%. A partir de este resultado, se puede deducir que no hubo un impacto significativo en las ventas. Sin embargo, en el año 2022, ocurrió lo

contrario, ya que, por una diferencia mínima, hubo más clientes fugados que clientes nuevos, con una diferencia negativa del -6,65%.

En la **Tabla 10** y la **Figura 22** se puede apreciar de manera general la diferencia año tras año en la comparación realizada anteriormente.

Tabla 10

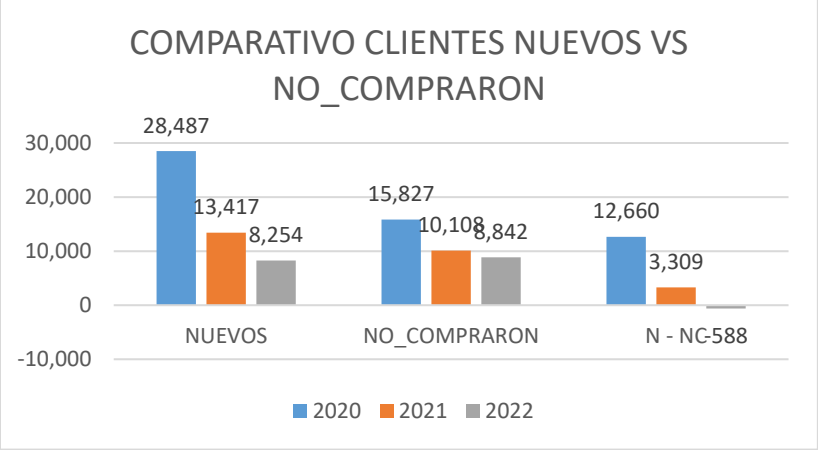
Diferencia clientes Nuevos – No compraron

AÑO	NUEVOS	NO_COMPRARON	N - NC
2020	28,487	15,827	12,660
2021	13,417	10,108	3,309
2022	8,254	8,842	-588

Nota. La tabla muestra la diferencia entre clientes que no compraron y los nuevos Fuente: Elaboración propia

Figura 22

Gráficos nuevos vs No compraron



Nota. La grafica muestra el comportamiento comparado año a año de los clientes nuevos y los que no compraron. Fuente: Elaboración propia

Seguido a la identificación de número de clientes que no compraron vs los nuevos el paso a seguir es determinar cómo es el comportamiento de la variable Recency y analizar los estadísticos de la variable para tener una visual de los tiempos en que los clientes dejan de comprar.

Después de identificar el número de clientes que no realizaron compras en comparación con los nuevos clientes, el siguiente paso es determinar el comportamiento de la variable "Recency" y analizar las estadísticas de dicha variable para obtener una visualización de los períodos en los que los clientes dejan de comprar.

Para iniciar el análisis, se establece el 1 de enero de 2023 como el día cero y se retrocede en el tiempo para determinar los días de recencia. Los resultados se muestran en la **Figura 23**, donde se observa lo siguiente: el período más largo en el que un cliente ha dejado de comprar es de 1.090 días, lo que sugiere que algunos clientes realizaron su última compra al comienzo del período de estudio. Por otro lado, el período más corto de recencia es de 3 días. La media se sitúa en 504 días, con una desviación estándar de 333 días.

Figura 23
Estadísticos variables Recency

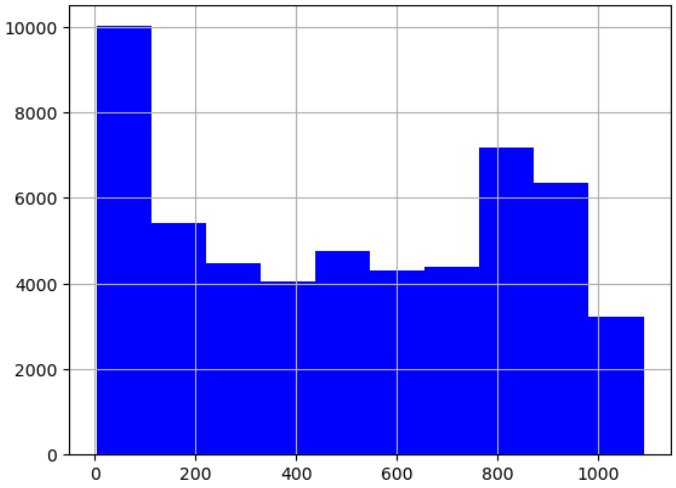
```
count   54157
mean      504
std       333
min         3
25%       185
50%       508
75%       821
max      1090
Name: Recency, dtype: float64
```

Nota, Resultado de los estadísticos de la variable Recency, Fuente: Elaboración propia

Para comprender la distribución de los datos y cómo se concentra la información en términos de días y clientes, se genera un histograma que se muestra en la **Figura 24**. En dicho histograma, se observa que el centro de los datos se encuentra alrededor de los 600 días, lo que corresponde a aproximadamente 4.100 clientes. En cuanto a los clientes con menor recencia, se identifica que más de 10.000 clientes se encuentran en el rango de 0 a 100 días, siendo este el grupo más numeroso. Además, aproximadamente 7.000 clientes se ubican en el rango de 780 a 980 días. El resto de los clientes se distribuye entre los 200 y 700 días.

Figura 24

Histograma Recency



Nota. El Histograma nos da a conocer la distribución de los clientes según los días de recencia. Fuente: Elaboración propia

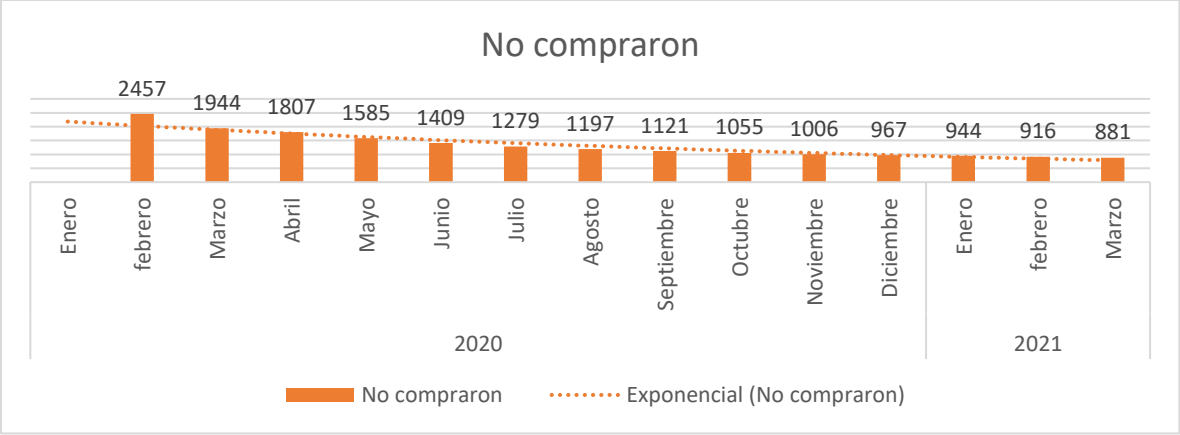
A continuación, se describirá el proceso para determinar la tasa de fuga. Para este análisis, se tomará como punto de partida los clientes que realizaron compras en enero de 2020, lo cual corresponde a un total de 4.113 clientes activos. Este mes será considerado como el mes 0, y a partir de aquí se identificará el número de clientes que se han fugado durante los siguientes 14 meses.

En primer lugar, es importante contextualizar cómo se llevó a cabo el ejercicio para identificar la fuga. Se seleccionaron los 4.113 clientes que realizaron compras en enero de 2020. Luego, se revisó cuáles de estos clientes no realizaron compras en febrero. Este primer paso reveló que 2.457 clientes no realizaron compras en dicho mes. A continuación, se analizaron aquellos que no compraron ni en febrero ni en marzo, lo cual resultó en 1.944 clientes que no realizaron compras en esos dos períodos. Posteriormente, se identificaron aquellos que no compraron en febrero, marzo y abril, dando como resultado 1.807 clientes en este punto, correspondiente al tercer mes de fuga. Este proceso se repitió sucesivamente hasta llegar al mes 14.

En conclusión, mediante este ejercicio se determinó que 881 clientes no volvieron a realizar compras desde enero de 2020 hasta marzo de 2021, lo cual indica que estos clientes se han fugado durante el período analizado. Esto genera una tasa de fuga del 21%. El resultado se puede observar de manera gráfica en la **Figura 25**.

Figura 25

Clientes que no compraron mes a mes



Nota, Con este grafico se muestra el número de clientes que no compraron mes a mes hasta identificar la fuga real. Fuente: Elaboración propia

Una vez que se determina la tasa de fuga es importante resaltar dentro de las recomendaciones que se darán a la empresa la importancia de dar atención a la tasa de clientes que dejan de comprar mes a mes, esto representa en términos de monetarios un impacto directo los ingresos de la compañía. Este resultado se puede observar de manera representativa en la **Tabla 11.**

Tabla 11

Identificación tasa de fuga

	NO COMPRARON	TASA DE NO COMPRA	SI COMPRARON	TASA DE SI COMPRA
MES 0			4113	
MES 1	2457	60%	1656	40%
MES 2	1944	47%	2169	53%
MES 3	1807	44%	2306	56%
MES 4	1585	39%	2528	61%
MES 5	1409	34%	2704	66%
MES 6	1279	31%	2834	69%
MES 7	1197	29%	2916	71%
MES 8	1121	27%	2992	73%
MES 9	1055	26%	3058	74%
MES 10	1006	24%	3107	76%
MES 11	967	24%	3146	76%
MES 12	944	23%	3169	77%
MES 13	916	22%	3197	78%
MES 14	881	21%	3232	79%

Nota. La tabla nos muestra comportamiento en la tasa de fuga mes a mes. Fuente: Elaboración propia

Ahora, se va a determinar cuál es la probabilidad de fuga total, para ello se realizará un proceso iterativo durante un periodo de 14 meses donde se identificará el número de clientes que han dejado de comprar durante ese periodo y seguido se hallaran las diferentes tasas en cada uno de los meses de no compra. Tomamos nuevamente como mes cero a enero del año 2.023 hasta llegar a marzo del 2.021. tal como muestra la **Tabla 12.**

Tabla 12

Tasa probabilidad de fuga total

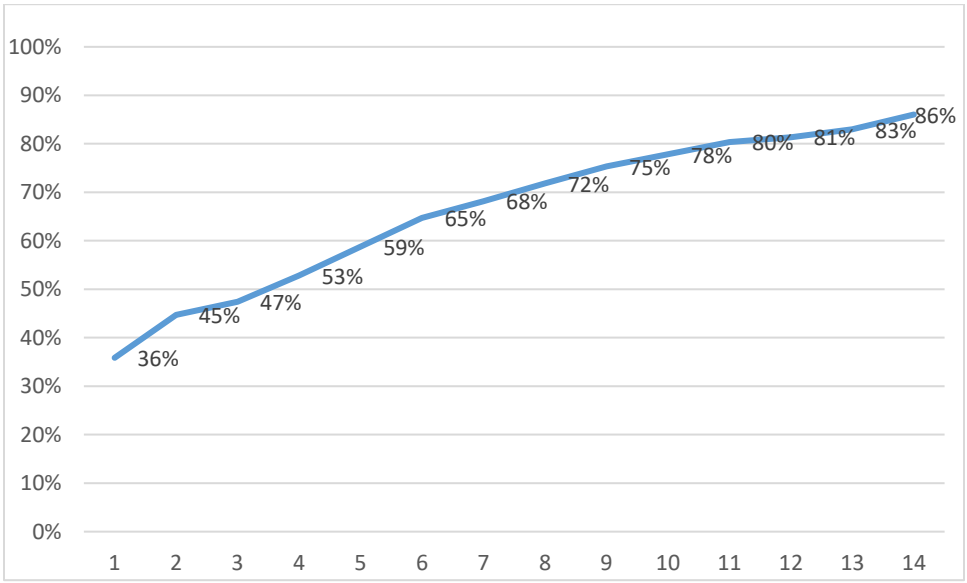
Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12	Mes 13	Mes 14
2457	1944	1807	1585	1409	1279	1197	1121	1055	1006	967	944	916	881
36%	45%	47%	53%	59%	65%	68%	72%	75%	78%	80%	81%	83%	86%

Nota. Muestra tasa de probabilidad desde el mes 1 al 14. Fuente Elaboración propia

Dado el resultado, se define que el mes numero 10 será el punto de partida para determinar la fuga total de un cliente. Cuando se refiere a fuga total, se hace referencia al momento en que un cliente no vuelve a comprar a la empresa y decide adquirir sus productos a la competencia de manera definitiva. Para este caso, tal como muestra la F**igura 26**, en el mes 10 la probabilidad de pérdida total de un cliente es del 75%.

Figura 26

Probabilidad de fuga



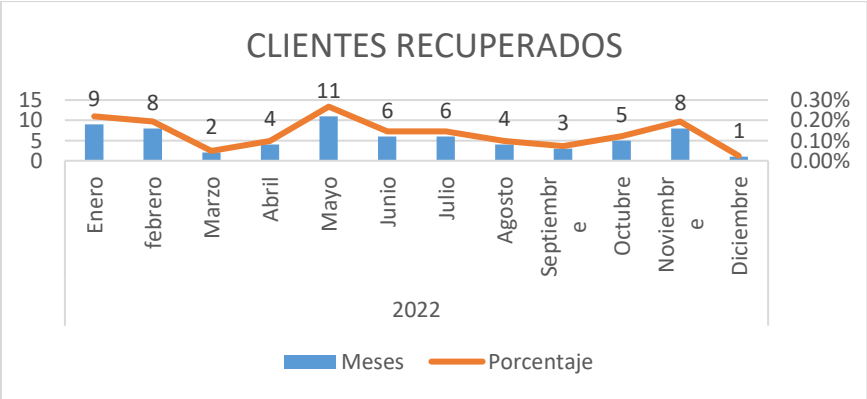
Nota. La grafica muestra la probabilidad de fuga en diferentes tiempos. En el eje “X” tenemos el número de meses que se definió para el análisis, partimos del mes 1 al mes 14 y en el eje “Y” la probabilidad de fuga. Fuente: Elaboración propia

De acuerdo con este resultado se resalta la importancia de realizar el seguimiento y gestión de continuidad del cliente maximizando los vínculos comerciales los cuales permitan identificar los posibles síntomas de una posible fuga a la competencia anticipando así la pérdida total. Este seguimiento permitirá que el área comercial identifique las necesidades que surgen en el día a día mejorando el nivel de satisfacción de los clientes con relación a la calidad de servicio.(García et al., 2017)

A parte de identificar fuga de clientes se considera importante revisar y evaluar el número de clientes recuperados durante el año 2.022. Para la identificación de estos clientes se tomó como punto de partida enero del año 2020 y se tuvo en cuenta los clientes que no habían vuelto a comprar durante dos años. Este ejercicio nos da un resultado de 118 clientes recuperados durante el año evaluado correspondiendo así a un al 2.87%. este resultado lleva a realizar la recomendación al área comercial de crear estrategias de recuperación con el objetivo de aumentar esa tasa de recuperación de clientes.

Figura 27

Clientes Recuperados



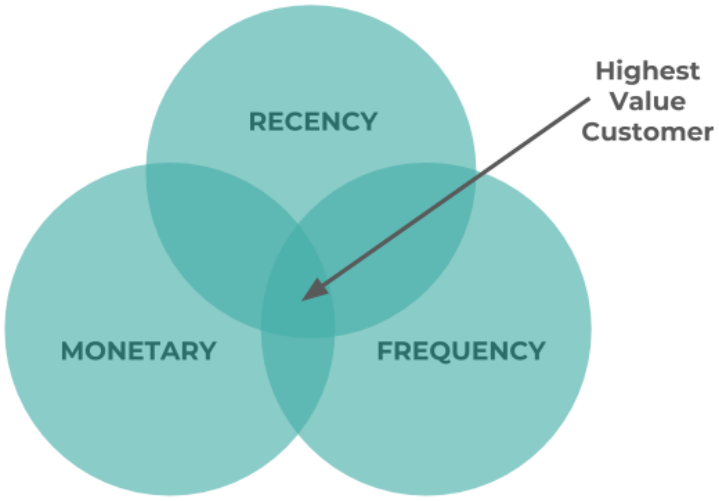
Nota. La grafica muestra el umero de clientes recuperados mes a mes en el año 2022. Fuente: Elaboración propia

9. Segmentación e identificación de clientes potenciales

A continuación, procederemos a realizar el análisis RFM para generar la segmentación de clientes según su comportamiento de compra. Antes de llevar a cabo todo el proceso, proporcionaremos un contexto sobre en qué consiste y cómo se realiza el análisis RFM.

El análisis RFM es una técnica de marketing utilizada para determinar de manera cuantitativa cuáles son los mejores clientes, examinando la recencia de sus compras, la frecuencia con la que compran y el monto que gastan. Este método se emplea para analizar el comportamiento y definir segmentos de mercado.(Sharma, 2019)

Figura 28
Ilustración RFM



Nota. Proceso RFM. Fuente: Sharma, (2019)

Para realizar este análisis, se toman en cuenta tres variables clave:

- RECENCY (Recencia): Determina la cantidad de días transcurridos desde la última compra realizada por el cliente.
- FREQUENCY (Frecuencia): Permite conocer la frecuencia de compra durante un período determinado.
- MONETARY (Monetario): Representa el valor monetario total de las compras realizadas por el cliente durante el período analizado.

> Los pasos del análisis RFM consisten en asignar una escala a los clientes en función de cada factor RFM por separado. La segmentación comienza con RECENCY, luego continúa con FREQUENCY y finalmente se evalúa el valor MONETARY. Se comienza clasificando a los clientes según su recencia, es decir, el tiempo transcurrido desde su última compra, en orden ascendente (los compradores más recientes se sitúan en la parte superior). Luego, los clientes se dividen en quintiles (grupos iguales), asignando una puntuación de 5 al primer 20%, una puntuación de 4 al siguiente 20% y así sucesivamente. Este proceso se repite para la variable FREQUENCY, clasificando a los clientes de mayor a menor frecuencia y asignando una puntuación de 5 al 20% superior, y puntuaciones decrecientes (4, 3, 2 y 1) a los quintiles menos frecuentes. El mismo proceso se aplica a la variable MONETARY. Finalmente, todos los clientes se clasifican concatenando los valores R, F y M (Birant, 2011, p. 92)

El análisis RFM asigna puntuaciones de valor a cada cliente en función de su comportamiento previo. Utilizando el sistema de quintiles explicado anteriormente, se

> pueden asignar hasta 125 puntuaciones diferentes (5x5x5). Estas puntuaciones varían en tamaño entre sí. La puntuación de un cliente puede oscilar entre 555, que es la más alta, y 111, que es la más baja. Los mejores clientes se encuentran en el quintil 5 de cada factor (555), lo que indica que han comprado más recientemente, con mayor frecuencia y han gastado más dinero. (Birant, 2011, p. 92)

A continuación, se presenta un ejemplo ilustrativo en las siguientes tablas utilizando una base de datos para mostrar cómo se clasificarían y calificarían los clientes. Al reducir la base de datos a un solo registro por cliente, se obtiene un total de 54,157 clientes y la tabla de las tres variables quedaría de la siguiente manera:

Tabla 13

Variables RFM

Id_ Cliente	RECENCY	FREQUENCY	MONETARY
1	202	45	$1,203,624,756
2	114	7	$373,582
3	363	1	$200,214
4	95	18	$1,014,330
5	3	302	$50,478,127
6	5	68	$20,235,326
7	6	144	$2,320,461,039
8	11	204	$3,161,250,447
9	136	21	$1,769,170
10	23	43	$240,761,266

Nota. Se crean las tres variables RFM Fuente: Elaboración propia,

Como se puede observar en la **Tabla 13**, se ha tomado como ejemplo a los primeros 10 clientes y se ha consolidado el total de compras que realizaron durante el período comprendido entre el 1 de enero de 2020 y el 31 de diciembre de 2022. A continuación, se ha determinado en días cuánto tiempo ha pasado desde su última compra, tomando como fecha de análisis el 1 de enero de 2023. Tomemos el cliente número 1 como ejemplo: sus compras durante todo el período suman $1,203,624,756, su última compra fue hace 202 días y su frecuencia es de 45 compras durante todo el período analizado.

A continuación, procederemos a asignar puntajes y calificar las variables RFM. Tomemos como ejemplo los primeros 15 clientes, como se muestra en la **Tabla 14**.

Tabla 14

Puntuación RFM

Id_ Cliente	RECENCY	R	FREQUENCY	F	MONETARY	M	RFM
1	202	4	45	5	$1,203,624,756	5	455
2	114	5	7	5	$373,582	4	554
3	363	4	1	3	$200,214	4	434
4	95	5	18	5	$1,014,330	5	555
5	3	5	302	5	$50,478,127	5	555
6	5	5	68	5	$20,235,326	5	555
7	6	5	144	5	$2,320,461,039	5	555
8	11	5	204	5	$3,161,250,447	5	555
9	136	4	21	5	$1,769,170	5	455
10	23	5	43	5	$240,761,266	5	555
11	4	5	118	5	$1,499,673,954	5	555
12	74	5	12	5	$795,107	5	555
13	11	5	329	5	$22,377,711	5	555
14	215	4	22	5	$2,083,095	5	455
15	46	5	14	5	$1,435,268	5	555

Nota. Se asigna puntajes de acuerdo con criterios de calificación Fuente: Elaboración propia

Una vez asignados los puntajes de acuerdo con los parámetros establecidos previamente, podemos realizar la segmentación y análisis correspondientes.

En la columna RFM se ha concatenado los puntajes de cada una de las variables. Siguiendo con el ejemplo del cliente número 1, podemos observar que su última compra fue hace 202 días con respecto a la fecha de referencia cero (1 de enero de 2023), lo cual le otorga un puntaje de 4. Además, su frecuencia de compra le otorga un puntaje de 5, al igual que el monto total de compras realizado durante todo el período. Al concatenar estos puntajes, obtenemos un resultado final de 455 puntos.

Con base en esta puntuación, podemos determinar que este cliente es de alto valor, ya que presenta un monto de compra significativo y una frecuencia alta. Sin embargo, también muestra un riesgo elevado de fuga, dado que ha pasado más de 5 meses sin realizar una compra. De acuerdo con el análisis realizado en el capítulo anterior, donde se determinó la probabilidad de fuga definitiva, este cliente se encuentra en un 50% de probabilidad de no volver a comprar. Por lo tanto, es crucial comenzar a implementar estrategias de retención para fidelizarlo.

Tabla 15

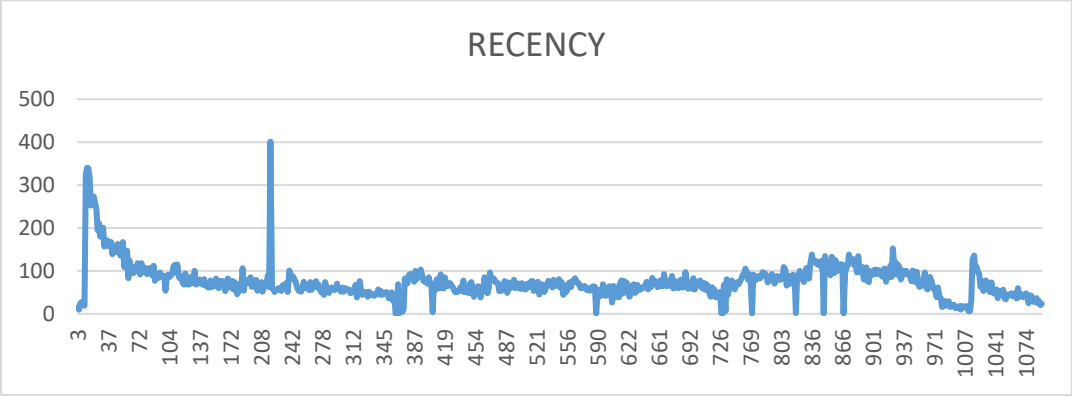
Estadísticos Variable Recency

MODA	220
MEDIANA	508
PROMEDIO	504
MAXIMO	1090
MINIMO	3

Nota. Se obtiene los estadísticos de la variable Fuente: Elaboración propia

Figura 29

Grafica Estadística Variable Recency



Nota. La grafica muestra en eje X los días de Recencia y Eje Y Cantidad de clientes. Fuente: Elaboración propia, (2022)

Realizando un análisis estadístico a la variable Recency se identifica que el promedio de días en los que un cliente ha dejado de comprar corresponde a los 504 días y la moda está concentrada en los 220 días, También se puede observar en la **Figura 29** que dicha Moda está concentrada sobre unos 400 clientes siendo el grupo con mayor número de clientes en ese tiempo de recencia. Dentro del análisis de identifica un máximo de Recencia de 1090 días. Importante resaltar que hay un grupo importante de clientes con una Recencia entre los 3 y 30 días denotando una buena continuidad de compra.

Tabla 16

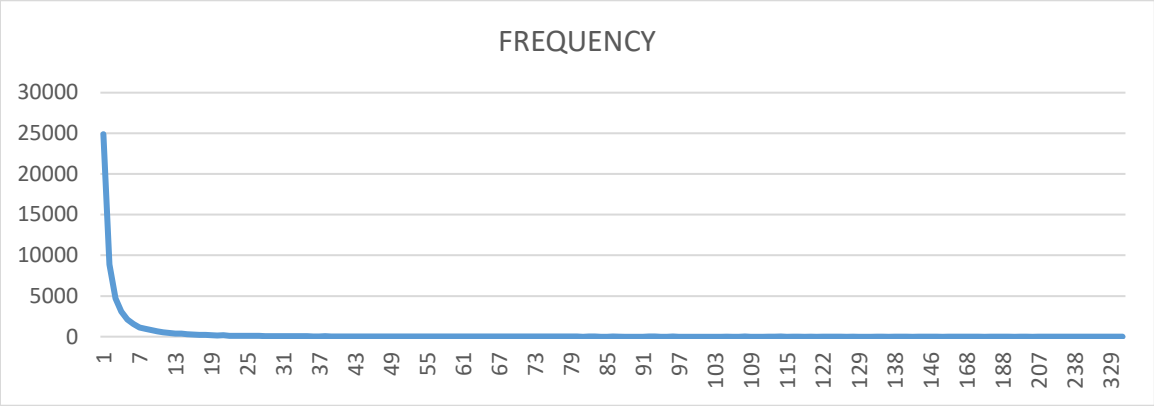
Estadísticos Variable Frequency

FREQUENCY	
MODA	1
MEDIANA	2
PROMEDIO	4.87
MAXIMO	425
MINIMO	1

Nota. Se obtiene los estadísticos de la variable Frequency. Fuente Elaboración propia

Figura 30

Gráfico Estadísticos Frequency



Nota. Figura nos muestra en Eje X Frecuencia de compra y Eje Y total de clientes Fuente: Elaboración propia

Con la variable Frequency se identifica que el promedio en la frecuencia de compra de los clientes es de 4.87, pero se observa de forma atípica que gran parte de los clientes únicamente han realizado una compra durante todo el periodo analizado, esto se puede observar en el valor mínimo y en la moda, y de acuerdo como se ve en la **Figura 29** está concentrado en cerca de 25.000 clientes, los cuales obedece casi que la mitad de los clientes registrados en la base de datos. Y en el máximo se ubica una frecuencia de compra de 425, la cual se hace importante identificar que clientes son y cuáles son sus montos de compra.

- **Identificación de Clientes Pareto**

Para realizar la identificación de los clientes Pareto se toma todos los registros transaccionales desde el 1 de enero del 2.020 hasta el 31 de diciembre del 2022, por lo tanto, se obtienen los registros de la variable MONETARY y realizamos una identificación del Pareto segmentándolo en 5 grupos y con ello buscamos determinar cuál es el número de clientes que

más representan valor monetario para la organización y a los cuales se deben direccionar las estrategias de retención.

Tabla 17

Clasificación clientes Pareto

PARETO	TOTAL, CLIENTES	VALOR
0.16%	86	$120,971,301,923.61
1.69%	917	$84,679,911,346.53
14.75%	7990	$29,033,112,461.67
33.25%	18006	$6,532,450,303.88
50.15%	27158	$725,827,811.54

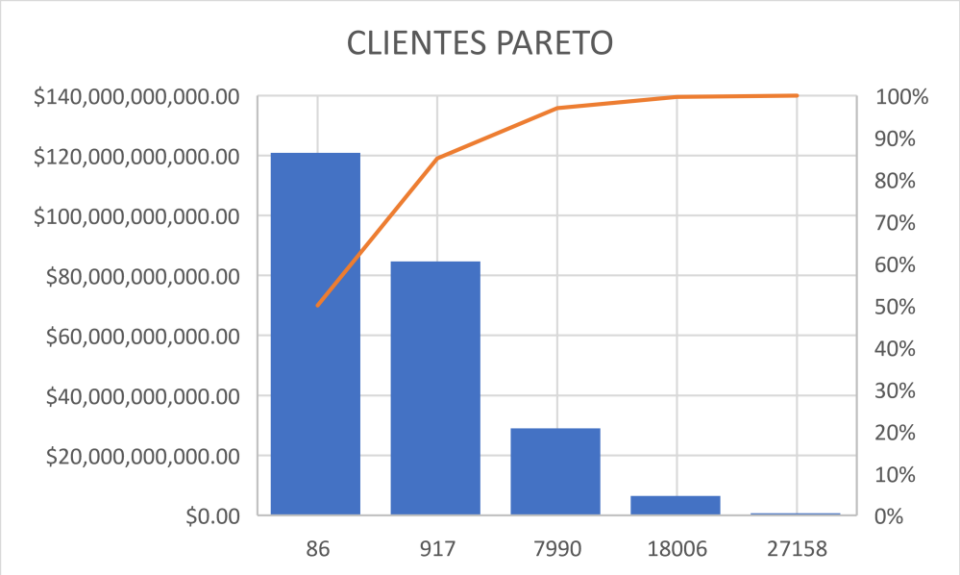
Nota. La tabla muestra la distribución de clientes e identificación del Pareto. Fuente. Elaboración propia

Se identifica que el 50% de las ventas de la compañía están representadas en 86 clientes que equivale tan solo al 0.16% del total de los clientes del total que realizaron compras durante todo el periodo analizado. El segundo grupo, está conformado por 917 clientes y corresponden al 1.69% del total, con una participación en las ventas del 30% esto para completar un 80% del total de las ventas de la empresa. a este segundo grupo de clientes hay que prestarle atención ya que se puede convertir en clientes más importantes y con posibilidad de crecimiento. Los últimos 3 grupos representan el otro 20% de las ventas repartidas en el 98.15% de los clientes.

Con el anterior análisis se determina que la compañía depende en un alto porcentaje de sus ventas en un número muy reducido de clientes. Con esto se determinar que es un riesgo muy alto ya que en caso de que un cliente de estos deserte o se fuge generara un déficit importante en las ventas, especialmente si ocurre con alguno de los 86 clientes.

Figura 31

Gráfico clientes Pareto



Nota. Gráfico Ilustrativo de la distribución del Pareto. Fuente: Elaboración propia

- **Clasificación y Segmentación con modelo RFM**

Después de haber determinado el Pareto de clientes, se procede a realizar la clasificación y segmentación haciendo uso del modelo RFM.

Para este proceso vamos a tener en cuenta las variables Frequency y Monetary, en segundo plano la variable Recency. Lo primero es realizar la tabla de clasificación y luego se procede a realizar la matriz de clasificación.

En este método de clasificación de clientes segmentamos por los siguientes criterios.

- Clientes con Mayor frecuencia y Mayor monto de compra.
- Clientes con Mayor monto de compra y Menor frecuencia

- Clientes con Mayor frecuencia y Menor monto de compra
- Clientes con Menor frecuencia y Menor monto de compra.

Se realiza esta segmentación con el objetivo de identificar los clientes con la mejor frecuencia y el mejor ticket de compra, con el resultado se presentará el listado de clientes para que la compañía inicie con el desarrollo de las estrategias de retención en caso de presentar posibilidad de fuga o de fidelización en caso contrario.

Tabla 18

Clasificación por criterio de Frecuencia y Monto

F - M	# Clientes	Porcentaje	Monetary%	< Recency	< Recency	> Recency	> Recency
>M - >F	16804	31.0%	95.96%	11654	21.52%	5150	9.51%
>M - <F	4858	9.0%	3.11%	1419	2.62%	3439	6.35%
>F - <M	4858	9.0%	0.24%	2394	4.42%	2464	4.55%
<F - <M	27637	51.0%	0.70%	6195	11.44%	21442	39.59%

Nota Resultado de la clasificación entre Monto y frecuencia. Fuente: Elaboración propia

En el primer segmento, se toma los clientes con una mayor Frecuencia de compra y Montos altos de compra, de aquí vamos a identificar los clientes del alto valor. Para esta clasificación se tomó los clientes que tenían 5 puntos en Frequency y 5 puntos en Monetary la cual dio como resultado 16.804 clientes, siendo el 31% del total de clientes y un 95,96% de participación en las ventas de la compañía.

Ahora para definir los clientes frecuentes, tomamos los de mayor frecuencia y bajo monto, en este segmento nos da un total de 4.858 clientes, siendo 8.97% del total de clientes y

tan solo un 0,23% de participación en las ventas. estos clientes son un buen objetivo para hacerlos incrementar sus compras, sean ventas cruzadas o aumento en sus compras habituales.

Por último, se obtienen los clientes de menor Frecuencia y bajo Monto. En este segmento encontramos que está representado por 27.637 clientes siendo un 51% del total de clientes. Seguramente esto grupo de clientes son los que compran esporádicamente y lo hacen directamente en los mostradores o por los diferentes canales de venta atendidos por el área de servicio al cliente. Para entender mejor la segmentación se ha creado la matriz **Figura 32** en el cual está dividido los diferentes grupos de acuerdo con los criterios anteriormente mencionados.

Tabla 19

Segmentación por Recencia

F – M	# Clientes	**< Recency**	**# Clientes**	**> Recency**
>M - >F	11654	21.52%	5150	9.51%
>M - <F	1419	2.62%	3439	6.35%
>F - <M	2394	4.42%	2464	4.55%
<F - <M	6195	11.44%	21442	39.59%

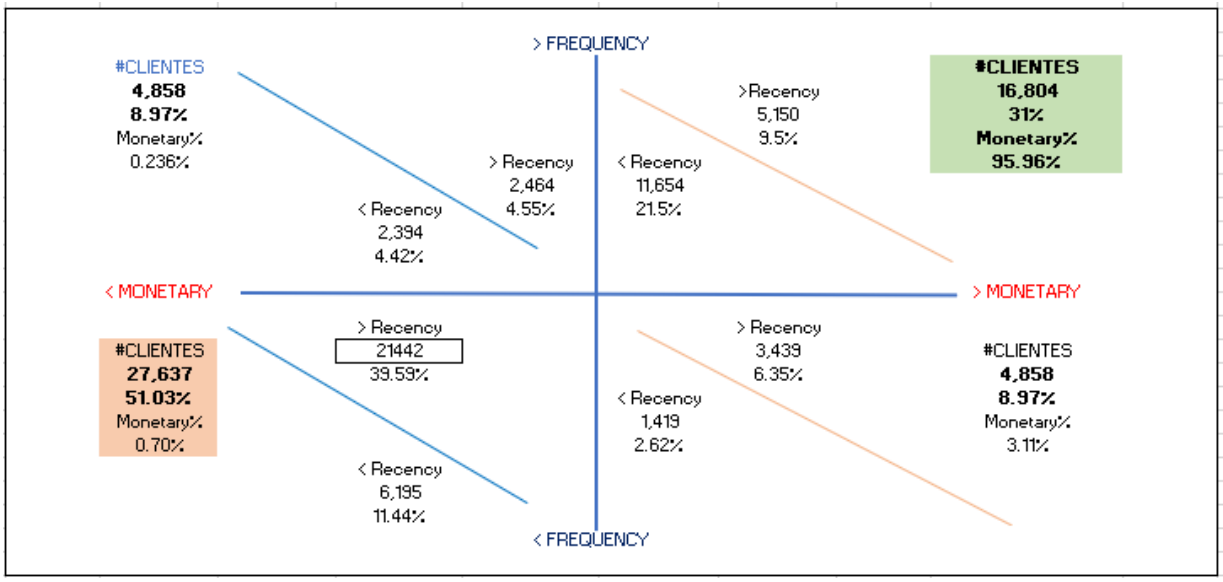
Nota. La tabla muestra como quedan los grupos segmentados por Recencia Fuente: Elaboración propia

Con el fin de complementar el anterior análisis, se toma cada uno de los grupos y se dividen por mayor o menor Recencia, de tal forma que se puede realizar una mejor clasificación de clientes y así orientar mejor las estrategias. De acuerdo con la **Tabla 19,** del grupo de mayor Monto y mayor Frecuencia 11.564 clientes presentan una baja Recencia y 5.150 clientes una mayor recencia. Para este primer segmento los de menor recencia se convierten en el grupo de mayor atención y priorización para implementar las estrategias de fidelización y a los de mayor

recencia a los que se deben aplicar las estrategias de retención. En los otros grupos se deben desarrollar las políticas de seguimiento para validar su nivel de importancia.

Figura 32

Matriz de segmentación de clientes



Nota. La figura muestra el resultado de la clasificación según los criterios de evaluación. Fuente: Elaboración propia

10. Aplicación modelos de clasificación

- **Preliminares del modelado:**

Con el objetivo de obtener resultados más precisos y una identificación más acertada de la tendencia a la fuga, se utilizarán tres algoritmos de clasificación: árboles de clasificación, random forest y regresión logística. Esto permitirá un mejor ajuste a los datos y obtendremos resultados más confiables en cuanto a la predicción de la fuga. A continuación un breve contexto de cada uno de los modelos.

Regresión Logística: Es un modelo estadístico utilizado para predecir la probabilidad de ocurrencias de un evento binario, es decir, cuando la variable de respuesta solo puede tener dos valores posibles, como "sí/no", "verdadero/falso", "éxito/ fracaso", etc. En lugar de predecir directamente los valores de la variable de respuesta, la regresión logística estima las probabilidades de que un evento ocurra utilizando una función logística. Esta función transforma una variable lineal ponderada por coeficientes en un rango de 0 a 1, lo que permite interpretar el resultado como una probabilidad. Es un modelo relativamente simple y fácil de entender y de implementar. Su principal desventaja es que puede sufrir de multicolinealidad si las variables independientes están altamente correlacionadas. (Torres Valverde & Padilla Rivadeneira, 2013)

Árbol de Clasificación: Los árboles de clasificación son modelos de aprendizaje automático que utilizan una estructura de árbol para tomar decisiones y clasificar instancias en diferentes categorías desde un enfoque de clasificación supervisada. Cada nodo del árbol

representa una característica o atributo, y las ramas del nodo representan las posibles combinaciones de valores para esa característica. Las hojas del árbol representan las etiquetas de clasificación o los resultados finales. Los árboles de clasificación son modelos de aprendizaje automático versátiles y fáciles de interpretar que se utilizan para resolver problemas de clasificación. Aunque tienen algunas limitaciones, su simplicidad y capacidad para manejar diferentes tipos de variables los hacen ampliamente utilizados y aplicables en diversas áreas.(Medina Merino & Ñique Chacón, 2017).

Random Forest: Es un algoritmo de aprendizaje automático que se basa en la combinación de múltiples árboles de decisión, formando un conjunto o "bosque". Cada árbol en el bosque se entrena con una muestra aleatoria de datos y toma decisiones de clasificación o regresión basadas en las características de entrada. Es altamente preciso y eficaz en una amplia gama de problemas de clasificación y regresión. (Medina Merino & Ñique Chacón, 2017)

“La generalización de error para los bosques converge a un límite en cuanto el número de árboles en el bosque sea grande”. (2017, p. 170)

Utilizando el conjunto de datos con los que hemos estado trabajando, se hizo necesario generar variables comportamentales adicionales con el fin de complementar y agregar más criterios de validación al ejecutar los modelos

Generar hipótesis:

De acuerdo con el conocimiento de negocio se realizarán las respectivas hipótesis del resultado esperado de los modelos y su aplicación en la compañía.

1. Creemos que una correcta aplicación de los modelos permitirá identificar los clientes más propensos a desertar y se logra si las variables predictoras se ajustan correctamente al algoritmo.

2. Creemos que un buen resultado de los modelos permitirá que el área comercial tenga una buena herramienta para diseñar las estrategias y se logra con la elección del mejor resultado.

3. Creemos que los diferentes modelos permitirán identificar los errores y ajustes que se debe realizar a la base de datos y se logra con el correcto entendimiento de los resultados.

Medición de asociaciones:

Para realizar la Medición de asociaciones se utilizará La matriz de Correlación, (Ver **Figura 33**) la cual proporciona información sobre la relación lineal entre las diferentes variables del conjunto de datos. Cada valor de conexiones se encuentra en el rango de -1 a 1, donde -1 indica una conexión negativa perfecta, 0 indica una falta de conexión y 1 indica una conexión positiva perfecta.

El resultado proporcionado por la Matriz de Correlación muestra que entre las variables Frecuencia_compra y Facturas_Generadas: El valor de correlación es 0.98. Esta correlación positiva cercana a 1 indica una fuerte relación lineal positiva entre las variables. Esto significa que a medida que aumenta la frecuencia de compra, también aumenta el número de facturas, y viceversa.

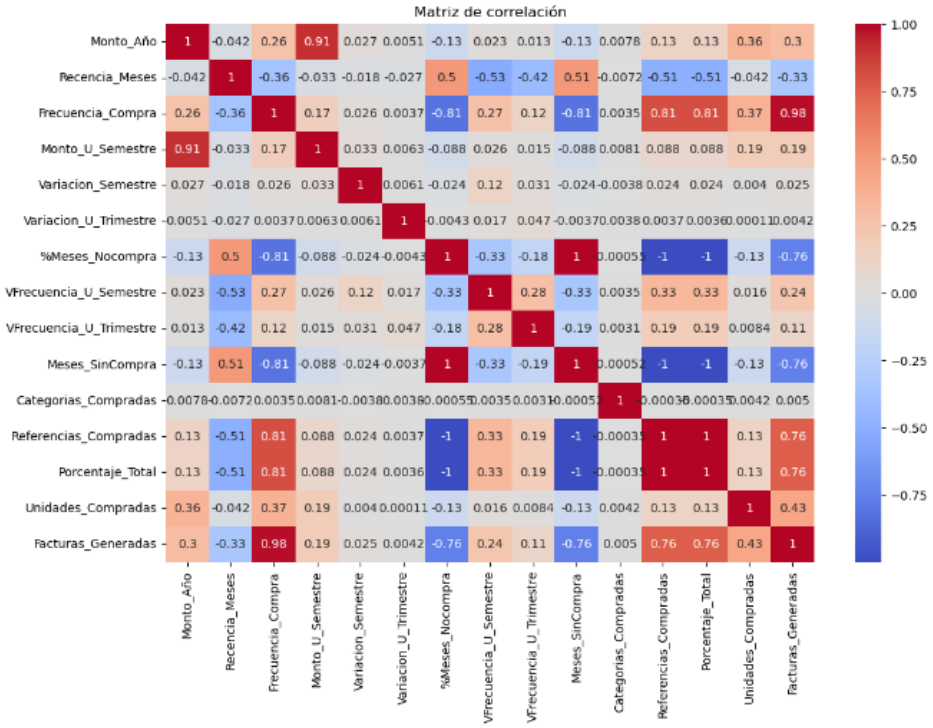
En una segunda validación se evidencia una correlación positiva muy fuerte entre las variables Monto_Año y Monto_U_Semestre. El valor de correlación de 0.91, que está cerca de 1, esto sugiere una relación lineal positiva casi perfecta entre estas dos variables.

Al identificar que estas variables están altamente relacionadas como es el caso de Frecuencia_compra y Facturas_Generadas la cual el valor de correlación es 0.9789, puede indicar la presencia de Multicolinealidad.

La alta correlación entre variables puede indicar problemas de multicolinealidad, lo que puede afectar la interpretación y estabilidad de los modelos. Es importante tomar medidas para abordar este problema y garantizar resultados más confiables.

Figura 33

Matriz de correlación



Nota. La figura muestra la principal correlación entre variables. Fuente: Elaboración propia

Plan de modelamiento

Para ejecutar los modelos, se llevaron a cabo los siguientes pasos:

- Se crearon tres ventanas de tiempo y se definió la variable objetivo (Target).
- Se creó una cuarta ventana de tiempo para validar el modelo y determinar la tendencia de fuga.
- Se determinó la fuga de clientes en cada una de las ventanas para crear la variable dependiente.
- La variable Tiene_Credito al ser una variable categórica (SI / NO) fue necesario convertirla en Variable Dummies para no afectar el resultado de los modelos.
- Se evidencio en el conjunto de datos la presencia de variables con valores muy altos y otras con valores bajos, por lo tanto, se realizó un proceso de estandarización o normalización con el objetivo de eliminar las diferencias en la escala y rango de las variables lo cual podría afectar negativamente el resultado de los modelos.
- Se llevo a cabo un aumento de la clase minoritaria mediante sobremuestreo para equilibrar el conjunto de datos. Este proceso se realizó con el objetivo de mejorar el rendimiento de los modelos de aprendizaje automático al garantizar que se disponga de suficiente información sobre la clase minoritaria y realizar predicciones precisas.
- Se estableció una semilla para generar números aleatorios de manera reproducible.
- Se ajustaron los pesos de clase para tratar con datos desbalanceados utilizando la técnica de oversampler. El ajuste se realiza en la base de entrenamiento dividiendo 1 sobre la proporción de clientes fugados y 1 sobre la proporción clientes No fugados.

Tabla 20

Resultado balanceo de datos

VARIABLE Fuga_Cliente	CLIENTES	PROPORCION	PESO CLASE	PESO CLASE
NO FUGADOS	38,728	40.472%	1/40.476%	2.47
FUGADOS	56,964	59.528%	1/59.524%	1.68
TOTAL VARIABLE Fuga_Cliente	95,692	100%		

Nota. Tabla peso de clases. Fuente: Elaboración propia

- Para abordar el impacto de la multicolinealidad encontrada en la matriz de correlación se aplicó la regularización a cada uno de los modelos con el fin de evitar el sobreajuste.

  En el caso de la regresión logística se aplicó la regularización L2 (Ridge). Con el objetivo de penalizar los coeficientes grandes y favorecer los coeficientes más pequeños y así controlar la complejidad del modelo.

  Para el árbol de clasificación y el random forest se empleó el parámetro ccp_alpha para aplicar regularización mediante la poda del árbol. Controlando la adición de nuevos nodos al árbol.
- Se llevó a cabo la validación cruzada en cada uno de los modelos con el objetivo de afinar y ajustar los hiperparametros, identificar posibles problemas de sobreajustes, y obtener la curva ROC mediante la validación cruzada.
- En el caso de la regresión logística, se imprimió la salida del modelo incluyendo los coeficientes correspondientes para comprender y utilizar las predicciones derivadas por el modelo. Además, se imprimió el resumen del modelo que presento sus errores estándar, los cuales son útiles para interpretar el modelo y comprender la influencia de las variables independientes en la variable dependiente. Estos errores estándar también pueden ser utilizados para realizar pruebas de hipótesis o intervalos de confianza. Por último, se

calcularon los puntos de corte para interpretar y utilizar las predicciones del modelo en un problema de clasificación. En resumen, la salida del modelo, los coeficientes y sus errores, así como los puntos de corte, tienen información esencial para interpretar, utilizar y evaluar los modelos de aprendizaje automático estándar en función de los objetivos establecidos.

- Se calcula la precisión de modelo tanto en el conjunto de entrenamiento como el conjunto de prueba para determinar y comparar su exactitud.
- Antes de generar la curva ROC se determina el Accuracy promediado con su desviación estándar con el objetivo de resumir e imprimir de manera más completa la precisión del modelo de clasificación. Esta métrica permite medir la proporción de predicciones correctas realizadas por el modelo.
- A través de validación cruzada se obtiene la curva ROC tanto en el conjunto de entrenamiento como el conjunto de prueba, esto con el fin de realizar una evaluación más completa y confiable del rendimiento del modelo en términos de su capacidad para clasificar correctamente los datos positivos y negativos, y su capacidad de generalización a nuevos datos.
- Se muestran las métricas tanto para el conjunto de entrenamiento como para el conjunto de prueba, con el objetivo de obtener una idea de si el modelo está sobreajustando los datos de entrenamiento. Si el modelo muestra un rendimiento significativamente mejor en los datos de entrenamiento en comparación con los datos de prueba, es probable que esté sobreajustado.

Resultados del modelo

Se presenta los resultados obtenidos de los 3 modelos de clasificación en el cual se entrega un resumen de los diferentes evaluadores que permiten definir cuál de ellos es el mejor con respecto a la variable objetivo que corresponde a los clientes fugados y no fugados.

Modelo de regresión logística

Se realizó el cálculo del VIF (Factor de Inflación de la Varianza) para verificar la correlación entre las variables predictoras, y se encontró una alta correlación significativa entre ellas. Esto confirma la presencia de multicolinealidad en las diferentes variables del conjunto de datos. Por lo anterior se realizó el proceso de regularización L2(Ridge) para este modelo.

En la salida del modelo los coeficientes indicados corresponden a los pesos asignados a cada variable predictora en el modelo de regresión logística. El coeficiente representa el cambio esperado en la variable de respuesta (en escala logarítmica) por cada unidad de cambio en la variable predictora correspondiente, manteniendo constantes las demás variables. Por ejemplo, si tomamos el primer coeficiente (3.54659636) y la primera variable predictora, significa que un incremento de una unidad en esa variable está asociado con un aumento esperado de aproximadamente 3.54659636 en la probabilidad logarítmica del evento de interés, manteniendo constantes las demás variables.

El intercepto (0.11776067) representa el valor esperado de la variable de respuesta cuando todas las variables predictoras son iguales a cero.

Se calculó la precisión del modelo ajustado y arrojo el siguiente resultado:

Precisión (conjunto de entrenamiento): 0.7437338160243356

Precisión (conjunto de prueba): 0.7642817333147555

En este caso, la precisión en el conjunto de entrenamiento es del 74,37%. Esto significa que el modelo clasifica correctamente el 74,37% de las instancias del conjunto de entrenamiento.

La precisión en el conjunto de prueba es del 76,43%. Esto significa que el modelo clasifica aceptablemente el 76,43% de las instancias durante el conjunto de prueba,
En general, una precisión alta indica que el modelo tiene una buena capacidad para realizar predicciones precisas.

Se calculan las métricas de evaluación promedio y su desviación estándar arrojando el siguiente resultado: Accuracy promedio: 0.7437 desviación estándar: 0.0038. El promedio de precisión es de 0.7437, este valor indica que, en promedio, el modelo clasifica correctamente el 74.37% de las instancias en el conjunto de datos evaluados. Con esta medida se evalúa el rendimiento general del modelo.

La desviación estándar es de 0.0038. La desviación estándar es una medida de la dispersión de los valores individuales con respecto a los medios. En este caso, la desviación estándar baja indica que los valores de precisión en diferentes evaluaciones del modelo están cercanos al promedio, lo que sugiere una consistencia en el rendimiento del modelo.

Se calcula el punto de corte arrojando los siguientes resultados:
Precisión con punto de corte 0.5 (conjunto de entrenamiento): 0.7437338160243356,
Precisión con punto de corte 0.5 (conjunto de prueba): 0.7642817333147555.

Cuando se utiliza un punto de corte de 0.5, se considera que una instancia pertenece a la clase positiva si la probabilidad estimada por el modelo está por encima de 0.5, y la clase negativa está por debajo de 0.5. En este caso, la precisión con un punto de corte de 0,5 en el conjunto de entrenamiento es del 74,37%. Esto significa que, al utilizar el punto de corte de 0.5 para clasificar las instancias en el conjunto de entrenamiento, el 74.37% de las instancias se clasificaron correctamente.

La precisión con un punto de corte de 0,5 en el conjunto de prueba es del 76,43%. Esto significa que, al utilizar el punto de corte de 0.5 para clasificar las instancias en el conjunto de prueba, el 76.43% de las instancias se clasificaron correctamente.

En resumen, al utilizar un punto de corte de 0.5, el modelo logra una precisión similar tanto en el conjunto de entrenamiento como en el conjunto de prueba. Esto sugiere que el modelo generaliza bien y mantiene un buen rendimiento.

Se realiza validación cruzada para hallar la curva ROC y se obtiene los siguientes resultados en las métricas de clasificación:
ROC-AUC en entrenamiento: 0.7437338160243356
ROC-AUC en prueba: 0.7435749882253008

En este caso, el valor del ROC-AUC en el conjunto de entrenamiento es de 0.7437. Esto significa que el modelo logra una buena capacidad para discriminar entre las clases positiva y negativa en el conjunto de entrenamiento, con un rendimiento del 74,37%. Y el valor en el conjunto de prueba es de 0.7436. Esto indica que el modelo también tiene una buena capacidad de discriminación en datos no vistos durante el entrenamiento, con un rendimiento del 74,36%.

El modelo muestra un rendimiento consistente en la capacidad de discriminación entre las clases positiva y negativa tanto en el conjunto de entrenamiento como en el conjunto de prueba, según la métrica ROC-AUC.

Modelo Árbol de Clasificación

El árbol de clasificación arroja los siguientes resultados al momento de calcular el Accuracy promediado y la desviación estándar:

Accuracy promedio: 0.7742

Desviación estándar: 0.0020

El resultado de evaluación del árbol de clasificación indica una precisión promedio de 0.7742 y una desviación estándar de 0.0020. Estos proporcionan información sobre la precisión y la consistencia de las predicciones realizadas por el modelo.

El promedio de precisión, también conocido como tasa de acierto, es la proporción de instancias clasificadas correctamente sobre el total de instancias evaluadas. En este caso, una

precisión promedio de 0.7742 indica que el modelo ha clasificado correctamente en un 77.42% de las instancias. la desviación estándar de 0.0020 sugiere que los resultados tienden a ser consistentes, con poca diferencia entre las predicciones realizadas por el árbol de clasificación.

Se calculó a través de validación cruzada la curva ROC-AUC para el conjunto de entrenamiento y el conjunto de prueba. Estas métricas evalúan el rendimiento del modelo de clasificación en términos de su capacidad para clasificar correctamente las clases positivas y negativas.(Wikipedia, 2022)

En este caso, se obtuvo un ROC-AUC de 0,8519 para el conjunto de entrenamiento y un ROC-AUC de 0,8488 para el conjunto de prueba. Estos indican que el modelo tiene una capacidad de discriminación relativamente alta, ya que se acerca al valor máximo de 1.

Modelo Random Forest

Para el caso del modelo Random Forest se obtuvo resultados muy similares al árbol de clasificación. En primer lugar, se muestra el Accuracy promediado con su desviación estándar,

Accuracy promedio: 0.7709

Desviación estándar: 0.0026

El resultado de 0.7709 indica que, en promedio, el modelo de Random Forest clasificó correctamente el 77.09% de las instancias en el conjunto de datos utilizados para la evaluación. Este valor se encuentra entre 0 y 1, donde 1 representa una precisión del 100%.

La desviación estándar de 0.0026 indica que es bastante baja, lo que sugiere que los resultados del modelo fueron bastante consistentes.

En el cálculo de la curva ROC usando validación cruzada se obtiene los siguientes resultados de las métricas de clasificación:
ROC-AUC en entrenamiento: 0.8511619138198334
ROC-AUC en prueba: 0.8483740208589073

Se obtuvo la métrica ROC-AUC (Área bajo la curva ROC) para evaluar el rendimiento del modelo. El valor de 0.8511619138198334 para ROC-AUC en entrenamiento indica que el modelo obtuvo un buen rendimiento al clasificar las instancias del conjunto de entrenamiento.

El valor de 0.8483740208589073 para ROC-AUC en prueba indica el rendimiento del modelo en el conjunto de prueba, que contiene datos no vistos durante el entrenamiento. Este valor también indica un rendimiento bastante bueno.

Finalmente, se evalúan las métricas que complementan la definición de los resultados. En esta validación, se consideran el F1 score, Recall, Precisión, el AUC y la Accuracy. Para este propósito, se unifican los resultados en la siguiente tabla con el objetivo de determinar cuál fue el modelo más destacado. Luego, se analizarán las métricas de cada modelo para tomar la decisión final sobre cuál modelo se utilizará para realizar nuevas predicciones en el futuro.

Tabla 21
Evaluadores del modelo

	RESULTADO BASE ENTRENAMIENTO			RESULTADO BASE PRUEBA		
	REGRESION LOGISTICA	ARBOL DE CLASIFICACION	RANDOM FOREST	REGRESION LOGISTICA	ARBOL DE CLASIFICACION	RANDOM FOREST
AUC	0,74	0.85	0.85	0.74	0.85	0.85
ACURACCY	0.74	0.78	0.78	0.76	0.79	0.8
RECALL (FUGADOS)	0.85	0.85	0.86	0.85	0.84	0.86
RECALL (NO FUGADO)	0.64	0.72	0.7	0.64	0.71	0.7
PRECISSION (FUGADOS)	0.7	0.75	0.74	0.78	0.81	0.81
PRECISSION(NO FUADOS)	0.81	0.82	0.84	0.74	0.75	0.77
F1 SCORE (FUGADOS)	0.77	0.79	0.8	0.81	0.83	0.83
F1 SCORE (NO FUGADOS)	0.71	0.77	0.76	0.69	0.73	0.73

Nota Resultados de los modelos de clasificación. Fuente: Elaboración propia

Regresión logística: se evalúan los resultados obtenidos en la base de entrenamiento y se comparan con la base de prueba, en la **Figura 34** encontramos la información suministrada por el modelo. Con este resultado se espera validar si tenemos problemas de sobreajuste.

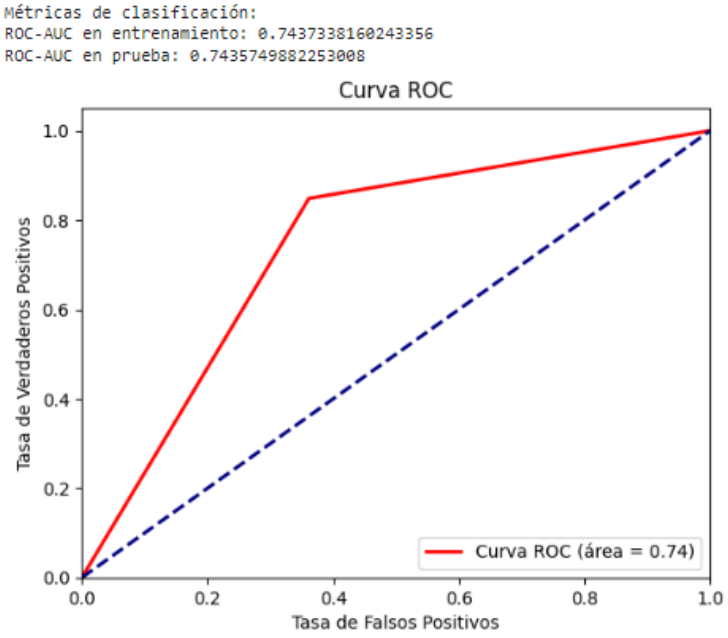
Figura 34
Métricas Modelo Regresión logística

```
Métricas de entrenamiento:
              precision    recall  f1-score   support

           0       0.81      0.64      0.71     39777
           1       0.70      0.85      0.77     39777

    accuracy                           0.74     79554
   macro avg       0.75      0.74      0.74     79554
weighted avg       0.75      0.74      0.74     79554

Métricas de prueba:
              precision    recall  f1-score   support

           0       0.74      0.64      0.69     11521
           1       0.78      0.85      0.81     17187

    accuracy                           0.76     28708
   macro avg       0.76      0.74      0.75     28708
weighted avg       0.76      0.76      0.76     28708
```

Nota. Resultados de las diferentes métricas usadas en los dos conjuntos de datos del modelo de regresión logística. Fuente: Elaboración propia

AUC (Area Under the Curve): Tanto en la base de entrenamiento como en la base de prueba, el AUC es de 0.74, lo cual indica que el modelo tiene un desempeño similar en ambas bases y es capaz de clasificar correctamente en un 74% de los casos.

Figura 35
Curva ROC Regresión logística



Nota. Resultado de la Curva ROC de la base entrenamiento y base prueba de la regresión Logística. Fuente: Elaboración propia

Accuracy : El Accuracy mide la proporción de predicciones correctas en relación con todas las predicciones realizadas. En la base de entrenamiento, la precisión es del 74%, mientras que en la base de prueba es del 75%. Esto sugiere que el modelo generaliza de manera aceptable, ya que la precisión en la base de prueba es ligeramente mejor que en la de entrenamiento.

Recall: El recall, también conocido como sensibilidad o tasa de verdaderos positivos, mide la proporción de casos positivos correctamente identificados. Tanto en la base de

entrenamiento como en la base de prueba, el recall para la clase de "fugados" es del 85%, lo cual indica que el modelo tiene un buen rendimiento al detectar a los fugados en ambos conjuntos de datos. Sin embargo, el recall para la clase "no fugado" es del 64% en ambos conjuntos de datos, lo que sugiere que el modelo tiene dificultades para identificar correctamente a los no fugados.

Precision: La precisión mide la proporción de casos positivos correctamente identificados en relación con todos los casos clasificados como positivos. En la base de entrenamiento, la precisión para la clase de "fugados" es del 70%, mientras que en la base de prueba es del 64%. Esto indica que el modelo tiende a clasificar más casos como "fugados" en la base de prueba, lo cual podría ser un indicio de sobreajuste.

F1 Score: El F1 Score es una medida combinada de precisión y recall que proporciona una medida equilibrada del desempeño del modelo. En la base de entrenamiento, el F1 Score para la clase "fugados" es del 0.77, mientras que en la base de prueba es del 0.81, Esto indica que el modelo logra una mejor combinación de precisión y recall para la clasificación de los casos "fugados" en la base de prueba en comparación con la base de entrenamiento. el modelo tiene una capacidad similar para identificar correctamente a los casos "no fugados" en ambos conjuntos.

Árbol de clasificación: se evalúan los resultados obtenidos en la base de entrenamiento y se comparan con la base de prueba, en la **Figura 36** encontramos la información suministrada por el modelo. Con este resultado se espera validar si tenemos problemas de sobreajuste.

Figura 36
Métricas Modelo Árbol de Clasificación

Métricas de entrenamiento:

	precision	recall	f1-score	support
0	0.82	0.72	0.77	39777
1	0.75	0.85	0.79	39777
accuracy			0.78	79554
macro avg	0.79	0.78	0.78	79554
weighted avg	0.79	0.78	0.78	79554

Métricas de prueba:

	precision	recall	f1-score	support
0	0.75	0.71	0.73	11521
1	0.81	0.84	0.83	17187
accuracy			0.79	28708
macro avg	0.78	0.78	0.78	28708
weighted avg	0.79	0.79	0.79	28708

Nota. Resultados de las diferentes métricas usadas en los dos conjuntos de datos con el modelo de árbol de clasificación Fuente: Elaboración propia

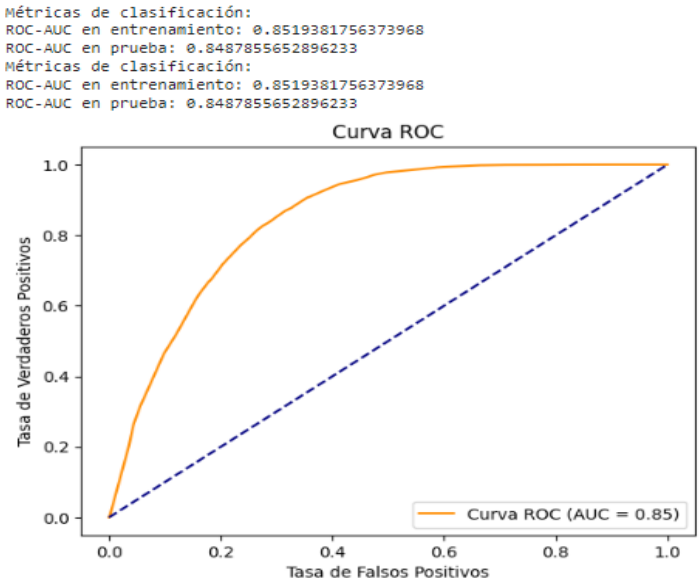
AUC (Area Under the Curve): Tanto en la base de entrenamiento como en la base de prueba, el AUC es de 0.85. Esto indica que el modelo tiene un buen desempeño para distinguir entre las clases en ambos conjuntos de datos y no hay diferencias significativas entre ellos.

Accuracy: El Accuracy en la base de entrenamiento es del 78%, mientras que en la base de prueba es del 79%. Estos valores son similares, lo que sugiere que el modelo generaliza bien y no hay evidencia de sobreajuste.

Recall: En la base de entrenamiento, el recall para la clase "fugados" es del 85%, mientras que en la base de prueba es del 84%. Estos valores son cercanos y sugieren que el modelo tiene un buen desempeño al detectar los casos de "fugados" en ambos conjuntos de datos. Sin embargo, el recall para la clase "no fugado" es del 72% en la base de entrenamiento y del 71% en la base de prueba. Estos valores también son similares, lo que indica que el modelo

tiene un desempeño similar al identificar correctamente los casos "no fugados" en ambos conjuntos.

Figura 37
Curva ROC Árbol de Clasificación



Nota. Resultado de la Curva ROC de la base entrenamiento y base prueba del Árbol de clasificación. Fuente: Elaboración propia

Precision: En la base de entrenamiento, la precisión para la clase "fugados" es del 75%, mientras que en la base de prueba es del 81%. Esto sugiere que el modelo clasifica más casos correctamente como "fugados" en la base de prueba en comparación con la base de entrenamiento. Para la clase "no fugado", la precisión es del 82% en la base de entrenamiento y del 75% en la base de prueba. Estos valores indican que el modelo tiende a clasificar más casos correctamente como "no fugados" en la base de entrenamiento en comparación con la base de prueba.

F1 Score: El F1 Score combina la precisión y el recall en una sola métrica. En la base de entrenamiento, el F1 Score para la clase "fugados" es del 0.79, mientras que en la base de prueba es del 0.83. Esto indica un mejor desempeño en la base de prueba para la clasificación de los casos "fugados". Para la clase "no fugados", el F1 Score es del 0.77 en la base de entrenamiento y del 0.73 en la base de prueba. Estos valores sugieren un desempeño similar en ambas bases de datos para la clasificación de los casos "no fugados".

En general, no parece haber evidencia clara de sobreajuste en el modelo de árbol de clasificación. Las métricas son consistentes entre la base de entrenamiento y la base de prueba, y no hay grandes diferencias en el rendimiento del modelo. Sin embargo, se observa una ligera mejora en la precisión, el F1 Score y el recall para la clase "fugados" en la base de prueba en comparación con la base de entrenamiento. Esto puede indicar una generalización ligeramente mejor del modelo en la base de prueba.

Random Forest: se evalúan los resultados obtenidos en la base de entrenamiento y se comparan con la base de prueba, en la **Figura** 38 encontramos la información suministrada por el modelo. Con este resultado se espera validar si tenemos problemas de sobreajuste.

AUC (Area Under the Curve): Tanto en la base de entrenamiento como en la base de prueba, el AUC es de 0.85. Esto indica que el modelo tiene un buen desempeño para distinguir entre las clases en ambos conjuntos de datos y no hay diferencias significativas entre ellos.

Figura 38
Metricas modelo Random Forest

Métricas de entrenamiento:

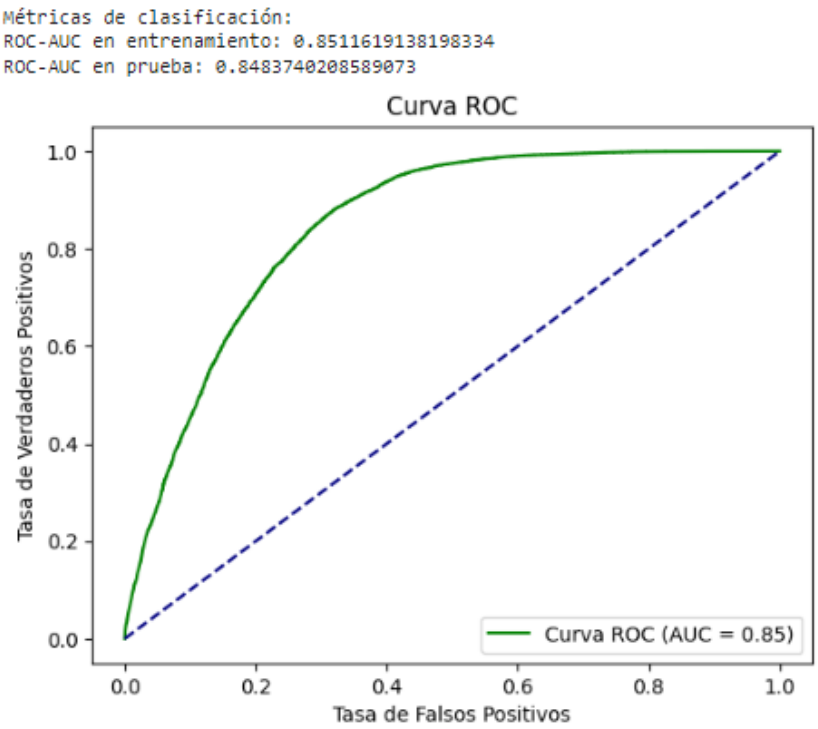
	precision	recall	f1-score	support
0	0.84	0.70	0.76	39777
1	0.74	0.86	0.80	39777
accuracy			0.78	79554
macro avg	0.79	0.78	0.78	79554
weighted avg	0.79	0.78	0.78	79554

Métricas de prueba:

	precision	recall	f1-score	support
0	0.77	0.70	0.73	11521
1	0.81	0.86	0.83	17187
accuracy			0.80	28708
macro avg	0.79	0.78	0.78	28708
weighted avg	0.79	0.80	0.79	28708

Nota. Resultados de las diferentes métricas usadas en los dos conjuntos de datos con el modelo de Random Forest Fuente: Elaboración propia

Figura 39
Curva ROC Random Forest



Nota: Resultado de la Curva ROC de la base entrenamiento y base prueba Random Forest. Fuente: Elaboración propia

Accuracy: La precisión en la base de entrenamiento es del 78%, mientras que en la base de prueba es del 80%. Estos valores son similares, lo que sugiere que el modelo generaliza bien y no hay evidencia de sobreajuste.

Recall: En la base de entrenamiento, el recall para la clase "fugados" es del 86%, mientras que en la base de prueba es del 86%. Estos valores son cercanos y sugieren que el modelo tiene un buen desempeño al detectar los casos de "fugados" en ambos conjuntos de datos. El recall para la clase "no fugado" es del 70% en la base de entrenamiento y del 70% en la base de prueba. Estos valores también son similares, lo que indica que el modelo tiene un desempeño similar al identificar correctamente los casos "no fugados" en ambos conjuntos.

Precisión: En la base de entrenamiento, la precisión para la clase "fugados" es del 74%, mientras que en la base de prueba es del 81%. Esto sugiere que el modelo clasifica más casos correctamente como "fugados" en la base de prueba en comparación con la base de entrenamiento. Para la clase "no fugado", la precisión es del 84% en la base de entrenamiento y del 77% en la base de prueba. Estos valores indican que el modelo tiende a clasificar más casos correctamente como "no fugados" en la base de entrenamiento en comparación con la base de prueba.

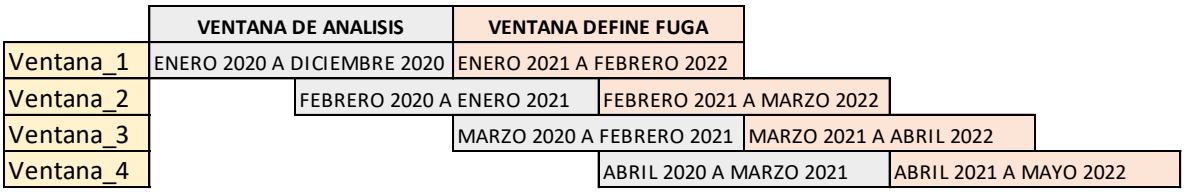
F1 Score: En la base de entrenamiento, el F1 Score para la clase "fugados" es del 0.80, mientras que en la base de prueba es del 0.83. Esto indica un mejor desempeño en la base de prueba para la clasificación de los casos "fugados". Para la clase "no fugados", el F1 Score es del

0.76 en la base de entrenamiento y del 0.73 en la base de prueba. Estos valores sugieren un desempeño similar en ambas bases de datos para la clasificación de los casos "no fugados".

En conclusión, no parece haber evidencia clara de sobreajuste en el modelo de Random Forest. Las métricas son consistentes entre la base de entrenamiento y la base de prueba, y no hay grandes diferencias en el rendimiento del modelo.

Es importante contextualizar cómo se crearon las ventanas de tiempo y cómo se determinó la fuga en cada una de ellas. En primer lugar, se decidió utilizar un periodo de 12 meses para el análisis de las variables predictoras y para determinar la fuga se estableció que se tomarían los 14 meses siguientes al periodo anteriormente mencionado. Por ejemplo, en la primera ventana, se considera el período de analisis desde enero de 2020 hasta diciembre de 2020, y a partir de enero del año 2.021 se identifican los clientes que no realizaron compras durante los 14 meses siguientes. De esta manera, se establece la situación de fuga que abarca desde enero de 2021 hasta febrero de 2022. Este procedimiento se repite en las dos ventanas siguientes. Y por último la ventana de tiempo número 4 será la base de prueba para validar la precisión de los modelos.

Figura 40
Ventanas de tiempo



Nota. Ilustración del diseño de las ventanas de tiempo Fuente: Elaboración propia.

11. Conclusiones y recomendaciones

Se desarrolló una metodología que utiliza la analítica de datos para identificar la propensión de los clientes a abandonar la compañía. Esta metodología incluye el cálculo de la tasa real de fuga y la segmentación de los clientes utilizando el modelo RFM. Además, se desarrolló modelos predictivos basados en árboles de clasificación, regresión logística y Random Forest.

Como resultado de este estudio, se encontró que la tasa de fuga real de los clientes es del 21% anual, lo cual se considera un porcentaje alto en comparación con el total de clientes de la compañía.

La segmentación de los clientes mediante el modelo RFM permitió dividirlos en cuatro grupos. Entre ellos, se identificaron 11,654 clientes de alto valor para la compañía, ya que presentaban una alta frecuencia de compra, un monto de compra elevado y una recencia baja.

Para determinar el mejor modelo para aplicar en el análisis predictivo de identificación de la tendencia de fuga de clientes, fue necesario considerar las métricas relevantes y la interpretación de los resultados de los tres modelos (regresión logística, árbol de clasificación y Random Forest).

Si nos centramos en el F1 Score para la clase "fugados", el modelo de Random Forest obtuvo el mejor desempeño en la base de prueba, con un valor de 0.83. Esto indica una buena

combinación de precisión y recall para la clasificación de los casos "fugados". Además, el modelo de Random Forest también tuvo un buen rendimiento en otras métricas, como el AUC, la precisión y el recall.

Sin embargo, es importante considerar que el modelo de regresión logística y el árbol de clasificación también tuvieron un desempeño bueno en términos de métricas. El modelo de regresión logística tuvo un F1 Score de 0.81 para la clase "fugados" en la base de prueba, mientras que el árbol de clasificación obtuvo un F1 Score de 0.83. Ambos modelos también tuvieron valores similares en otras métricas clave, como el AUC, la precisión y el recall. Aunque todos los modelos lograron un rendimiento aceptable al identificar los casos de fuga de clientes (clase "fugados"). Esto se evidencia por los valores de recall, precisión y F1 Score relativamente altos para esta clase en los conjuntos de entrenamiento y prueba se recomienda aplicar el modelo Random Forest para futuros análisis toda vez que este modelo ofrece ventajas en términos de precisión, capacidad para manejar diferentes tipos de variables y datos faltantes, resistencia al sobreajuste, robustez ante valores atípicos y ruido, evaluación de la importancia de características y eficiencia en grandes conjuntos de datos. Estas características lo convierten en una opción atractiva como modelo escogido para el análisis predictivo de la tendencia de fuga de clientes.

las variables más importantes para el modelo, en orden descendente de importancia, son las siguientes:

1. Monto_Año: 0.15084966275734948
2. Recencia_Meses: 0.12924914201752646

3. Variacion_U_Trimestre: 0.12930571059339355
4. Unidades_Compradas: 0.11825564042695597
5. Monto_U_Semestre: 0.08771870907464113

Estas variables tienen las importancias más altas en el modelo y, por lo tanto, se consideran las más influyentes para las predicciones.

En cuanto a la tendencia de fuga, es importante analizar las variables en función de su impacto en la retención o abandono de los clientes. En este caso, se pueden identificar dos variables relevantes relacionadas con la fuga:

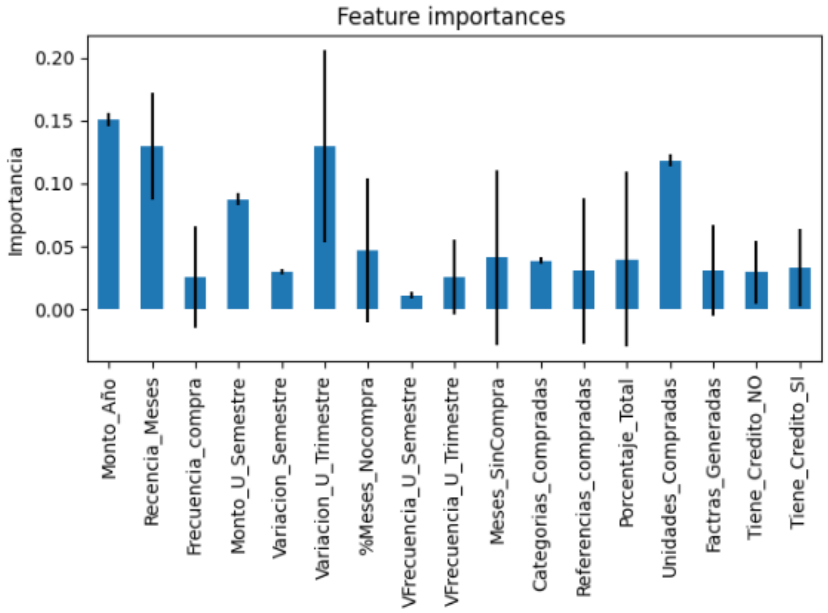
Recencia_Meses: Con una importancia de 0.12924914201752646, indica que la variable que representa la cantidad de meses desde la última compra es un factor significativo para predecir la fuga de los clientes. Cuanto más tiempo haya pasado desde la última compra, mayor será la probabilidad de que un cliente haya abandonado.

Variacion_U_Trimestre: Esta variable, con una importancia de 0.12930571059339355, también está relacionada con la tendencia de fuga. Indica que su comportamiento de compra durante el último trimestre es mínimo o no ha comprado puede ser un indicio de fuga. Estas dos variables indican la tendencia de fuga y son importantes para detectar clientes.

Se sugiere al área comercial comenzar identificando a aquellos clientes que presenten comportamientos de compra relacionados con las variables mencionadas anteriormente. Por ejemplo, se deben identificar aquellos clientes que han dejado de comprar durante un periodo

determinado o aquellos que han experimentado una disminución en sus compras. Esta segmentación permitirá enfocar los esfuerzos en retener a estos clientes en riesgo de fuga, implementando estrategias específicas y personalizadas para reactivar su interés y recuperar su nivel de compromiso con la empresa.

Figura 41
Gráfico variables importantes



Nota. Identificación de las variables que más impactan en el modelo para predecir la fuga en un modelo de Random Forest. Fuente: Elaboración Propia

El conjunto de datos proporcionado por la compañía resultó limitado, ya que solo contenía 6 variables iniciales. Como resultado, fue necesario crear variables adicionales basadas en el comportamiento, frecuencia, monto y recencia de compra. Se sugiere a la compañía que amplíe la captación de información en sus bases de datos, especialmente en el proceso de

creación de clientes y facturación, para aumentar el número de variables y mejorar el entrenamiento del modelo de predicción.

Es importante que la compañía brinde capacitación al equipo comercial sobre la importancia de la completitud y precisión de los datos registrados en los sistemas de información.

Se recomienda incluir variables demográficas en la recopilación de datos de clientes naturales, con el fin de comprender mejor los comportamientos de compra de este grupo y realizar predicciones más precisas en análisis futuros.

Dado que todos los modelos lograron un rendimiento aceptable al identificar los casos de fuga de clientes se sugiere implementar de inmediato políticas de seguimiento a los clientes y es fundamental determinar las principales razones de abandono, adicional es necesario diseñar y aplicar diversas estrategias de fidelización para los clientes con propensión a abandonar.

Por último, se recomienda clasificar a los clientes de alto valor identificados mediante el modelo RFM en función de su importancia para la compañía, y luego desarrollar estrategias específicas para cada grupo.

Además, sería interesante complementar y aplicar este enfoque analítico en otros sectores distintos al financiero y de telecomunicaciones, ya que hay pocos proyectos de analítica

orientados al sector industrial. Esto permitiría obtener mejores resultados y adentrar a este grupo de empresas en el mundo de la analítica.

12. Referencias Bibliográficas

Asociación Colombiana de la industria de la comunicación grafica. (2021). *Perfil envases y empaques*. https://andigraf.com.co/wp-content/uploads/2021/03/Envases-y-Empaques-2021.pdf

Birant, D. (2011). Data Mining Using RFM Analysis. En K. Funatsu (Ed.), *Knowledge-Oriented Applications in Data Mining* (pp. 91–108). InTech. https://doi.org/10.5772/13683

García, D. L., Nebot, À., & Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: A survey. *Knowledge and Information Systems*, *51*(3), 719–774. https://doi.org/10.1007/s10115-016-0995-z

IBM. (2021, agosto 17). https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Medina Merino, R. F., & Ñique Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*, *10*, 165–189.

Quiroa, M. (2022, abril 1). *Marketing industrial*. Economipedia. https://economipedia.com/definiciones/marketing-industrial.html

Sharma, P. P. M. (2019, noviembre 8). *Customer Segmentation using RFM Analysis*. data analytics edge. https://dataanalyticsedge.com/2019/11/08/customer-segmentation-using-rfm-analysis-using-r/

Torres Valverde, E. P., & Padilla Rivadeneira, G. S. (2013). *Medición de la intención de compra con base en un modelo de regresión logística de productos de consumo masivo.* [BachelorThesis]. http://dspace.ups.edu.ec/handle/123456789/5772

Wikipedia. (2022). Curva ROC. En *Wikipedia, la enciclopedia libre*.

https://es.wikipedia.org/w/index.php?title=Curva_ROC&oldid=145650383